



Dipl.-Wirtsch.-Ing. Stefan Heldmann

Big data analytics for the volatile world

New methodology and proof of concept for sales forecasting in an
industrial case study

DOCTORAL THESIS

to achieve the university degree of
Doktor der technischen Wissenschaften

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dipl.-Ing. Dr.techn. Christian Ramsauer

Institute of Innovation and Industrial Management

Univ.-Prof. Dipl.-Ing. Dr.mont. Hubert Biedermann

Montanuniversität Leoben

Graz - March 2018

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

Date

Signature

Acknowledgements

My research has benefited from substantial support. I would like to express my gratitude to everyone involved in the journey from the first research idea to the research project to the final version of the present doctoral thesis. In the following, I focus on the major contributors to my work.

My supervisor Prof. Dr. Christian Ramsauer has been a great source of motivation. His entrepreneurial spirit provided the opportunity to engage in a new format of research at his Institute of Innovation and Industrial Management (IIM) at the Graz University of Technology. Whenever encountering a roadblock, he has put his doctoral student first. I am thankful for the trustful relationship that has emerged over the course of my doctoral studies. I would like to thank Prof. Dr. Hubert Biedermann for being a supportive co-supervisor. The open exchange of ideas with him and his team in Leoben was an important milestone in my journey. Prof. Dr. Stefan Vorbach is an academic dean that always makes himself available for his students – many thanks for taking the time to review my research approach and results. Prof. Dr. Viktor Mayer-Schönberger has provided valuable inspiration that helped to shape my research idea in the very early stage. I am grateful for his advice and feedback during our discussions. My thanks also go to Prof. Dr. Jeffrey Saltz for the extensive exchange of literature on big data processes.

I am grateful for having been part of a great institute with an outstanding team. I would like to mention Martin Kremsmayr and Christian Rabitsch for innumerable hours of discussion, an unrivaled work environment of trust and support, and their extensive review work on my dissertation. Alexander Pointner and Matthias Schurig have been remarkable guides from the very first day of my journey onwards. I would also like to thank the Agility Research Group including Dominik Luczak, Thomas Deubel, Andreas Hönl, Marco Wampula, and Sebastian Miller. Our book project is a memorable experience and my work has benefited from each working session.

This work would not have been possible without the pioneering spirit at the partnering industrial company. I am very grateful for the opportunity to conduct the research project sponsored by the CEO and the Head of Strategy & Business Development. I owe a great debt of gratitude to the project team, Roland, Andreas and Heinz, for many extra miles travelled next to their daily business. Thanks also to our cooperation partner, especially Stefan and Tiago.

Thanks to my godfather Mathias Ulrich for taking the time to review my thesis. Particular thanks go to my family. The unconditional support of my parents, Sieglinde and Eugen, is invaluable. Inga and Hannes, you make my engine run – thank you for your love, understanding & motivation. *I dedicate this work to our little nugget.*

Abstract

Today's business environment is characterized by increasing volatility and uncertainty. Companies therefore seek for new concepts like agility to better cope with this volatile world. However, companies have growing difficulties to anticipate future changes. The age of analytics - driven by big data volumes, cheap data storage & processing, and advanced analytics – emerges at the same time. Companies consider *big data analytics (BDA)* as new source of competitive advantage. It provides the opportunity to gain a better understanding of the volatile business environment. Sales forecasting represents an important BDA application because many business functions benefit from an improved understanding of sales behavior. Therefore, the *primary research objective* of this work is to develop a *methodology* to decide where to use BDA in order to gain a better understanding of the volatile business environment (*general part*) and to develop an application for sales forecasting (*specific part*). A glance at existing literature on sales forecasting based on big data analytics reveals the early stage of this research area and that only a minor fraction considers cases in business-to-business industries. Consequently, the *secondary research objective* is a *proof of concept* that the BDA approach for sales forecasting works in industrial practice.

The development of the new methodology follows a design science research approach. The applicable knowledge is built from a literature review of process models and methodologies (jointly referred to as *processes*) for development of (big data) analytics applications. The review comprises a total of 76 processes of which the 25 most relevant ones are examined in detail. Furthermore, additional studies on success factors and issues from process implementation in practice are considered. A solid understanding of process steps, tasks, methods and lifecycles derives from this basis. The literature evaluation results in 27 improvement levers compared to existing processes that are aggregated into six design requirements. Furthermore, *CRISP-DM (Cross Industry Standard Process for Data Mining)* is identified as most appropriate candidate for the basic design of the new methodology. Based on the design requirements and basic design, the new methodology is built and evaluated during a case study with an European manufacturer of printed circuit boards.

The resulting methodology consists of five consecutive steps comprising 17 major tasks and builds upon a specific team setup. The *team setup* covers all relevant skill areas by defined roles and assigns responsibilities based on a workflow model. The setup introduces the *BDA manager* as new role bridging the core skill areas and ensuring a collaborative process. The initial *business understanding* step combines the concept of agility with the idea of big data analytics. It enables identification and prioritization of use cases that are considered for development. The methodology introduces a completely new step dedicated to identification, assessment and selection of data sources. This *big data sources* step facilitates the deliberate selection of multiple internal and external data sources containing relevant structured and unstructured data. The remainder steps – *data understanding*, *data preparation*, and *modeling & evaluation* – are specifically designed towards the sales forecasting use case based on structured data. A major advancement regarding data understanding is the introduction of the *BDA book*. This tool is tailored to the BDA manager role and provides an integrated approach to management of metadata as well as sourcing and verification of the data. The data preparation step includes a novel set of methods that particularly aim for integration of domain knowledge into the generation and prioritization of time series while still considering a large volume of data. These time series are subsequently used for modeling the sales forecasting application. In total, the new methodology includes 26 methods in form of tools and techniques to answer the *'how to do it'*.

In the case study, the BDA manager supported by the BDA book has proven to be effective for project management and coordination of the multidisciplinary project team. The role has also enabled content-related communication with stakeholders leading to support by key functions such as IT. The business understanding step has identified eleven BDA use cases for the volatile world and has prioritized sales forecasting as one of the two most relevant use cases. Based on the novel big data sources step, a long list of 28 sources has been identified and successively reduced to eight sources representing the desired data mix. In the data understanding step, a total of

191 datasets have been selected, sourced, explored, and verified for the sales forecasting use case. The processed data volume on the Hadoop-based project cluster has added up to more than 320 gigabytes, therefore exceeding the typical data volume for sales forecasting in a business-to-business industry. Time-series generation has resulted in more than 4 million time series of which 1,360 have been prioritized as modeling input based on defined quality and relevance criteria. Both modeling approaches, classification and regression, show reasonable performance regarding medium-term forecasts of sales growth. Based on the big data input, support vector machines as best performing classifier achieve an accuracy of up to 85% and elastic net as selected regression model has outperformed the conventional forecasting approach.

In conclusion, the new methodology fulfills the primary objective of the presented research with only minor limitations regarding implementation of the design requirements, such as a lack of use of visualization tools. Expanding the methodology with additional specific parts to address other use cases, for example, technology monitoring, and integration of unstructured data in the specific part represent key refinements to be addressed by future work. The observed forecasting performance provides a positive indication regarding the desired proof of concept. However, further research on generalization of the model performance is required due to the limited number of observations for model validation. Oversampling strategies and advanced feature selection are identified as adequate measures for this purpose. In addition, model parameter optimization and ensemble models provide key refinement options for future research aiming to further improve forecasting performance.

Kurzfassung

Das heutige Geschäftsumfeld ist von zunehmender Volatilität und Unsicherheit geprägt. Unternehmen streben daher nach neuen Konzepten wie Agilität, um diese volatile Welt besser zu bewältigen. Unternehmen haben jedoch zunehmend Schwierigkeiten, künftige Veränderungen zu antizipieren. Das Zeitalter der Analytik - getrieben durch große Datenmengen, billige Datenspeicherung und -verarbeitung, sowie Advanced Analytics – entwickelt sich zur gleichen Zeit. Unternehmen sehen in *Big Data Analytics (BDA)* eine neue Chance für Wettbewerbsvorteile. BDA bietet insbesondere die Möglichkeit, das volatile Geschäftsumfeld besser zu verstehen. Die Absatzprognose stellt hierbei eine wichtige BDA-Anwendung dar, da viele Geschäftsfunktionen von einem besseren Verständnis des zukünftigen Absatzverhaltens profitieren. Das *primäre Forschungsziel* dieser Arbeit ist infolgedessen die Entwicklung einer *Methodik* zur Entscheidungsunterstützung, in welchen Anwendungsfällen BDA zum Zweck eines besseren Verständnisses des volatilen Geschäftsumfeldes eingesetzt werden soll (*allgemeiner Teil*), und zur Entwicklung einer Anwendung für den Fall der Absatzprognose (*spezifischer Teil*). Ein Blick in die vorhandene Literatur zur Absatzprognose auf Basis von Big Data Analytics offenbart das Anfangsstadium des Forschungsgebietes, und nur ein kleiner Teil beschäftigt sich mit Fällen in Business-to-Business-Branchen. Folglich ist das *sekundäre Forschungsziel* ein Proof of Concept um zu zeigen, dass der BDA-Ansatz für die Absatzprognose in der industriellen Praxis funktioniert.

Die Entwicklung der neuen Methodik folgt dem Ansatz des Design Science Research. Das anwendbare Wissen basiert auf einer Literaturrecherche von Prozessmodellen und Methodiken (gemeinsam als *Prozesse* bezeichnet) für die Entwicklung von (Big Data) Analytics-Anwendungen. Der Literaturüberblick umfasst insgesamt 76 Prozesse, von denen die 25 relevantesten detailliert untersucht werden. Darüber hinaus werden zusätzliche Studien zu Erfolgsfaktoren und Problemstellungen bei der praktischen Implementierung dieser Prozesse betrachtet. Auf Basis dieser Grundlage ergibt sich ein solides Verständnis von Prozessschritten, Aufgaben und Methoden. Die Auswertung der Literatur führt zu 27 Verbesserungsansätzen im Vergleich zu bestehenden Prozessen, welche in sechs Designanforderungen zusammengefasst werden. Darüber hinaus wird *CRISP-DM (Cross Industry Standard Process for Data Mining)* als der am besten geeignete Prozess für das grundlegende Design der neuen Methodik identifiziert. Basierend auf den Designanforderungen und dem Basisdesign wird die neue Methodik im Rahmen einer Fallstudie mit einem europäischen Hersteller von Leiterplatten entwickelt und evaluiert.

Die resultierende Methodik besteht aus fünf aufeinander folgenden Schritten, die 17 Hauptaufgaben umfassen und auf einem bestimmten *Team Setup* aufbauen. Das Team Setup deckt alle relevanten Kompetenzbereiche anhand definierter Rollen ab und ordnet über den gesamten Prozessverlauf allen Aufgaben Verantwortlichkeiten zu. Das Setup führt den *BDA-Manager* als neue Rolle ein, welche die kritischen Kompetenzbereiche überbrückt und einen kollaborativen Prozess gewährleistet. Der erste Methodik-Schritt (*Business Understanding*) kombiniert das Konzept der Agilität mit der Idee der Big-Data-Analyse. Dies ermöglicht die Identifizierung und Priorisierung von Anwendungsfällen, die für eine Entwicklung in Betracht gezogen werden. Die Methodik führt darauffolgend einen vollständig neuen Schritt zur Identifizierung, Bewertung und Auswahl von Datenquellen ein. Dieser Schritt (*Big Data Source(s)*) erleichtert die gezielte Auswahl interner und externer Datenquellen, welche relevante strukturierte und unstrukturierte Daten enthalten. Die übrigen Schritte - *Data Understanding*, *Data Preparation*, und *Modeling & Evaluation* - sind speziell auf den Anwendungsfall der Absatzprognose, basierend auf strukturierten Daten, ausgerichtet. Ein wesentlicher Fortschritt hinsichtlich des Data Understanding ist die Einführung des *BDA Books*. Dieses Tool ist auf die BDA-Manager-Rolle zugeschnitten und bietet einen integrierten Ansatz für die Verwaltung von Metadaten, sowie die Beschaffung und Verifizierung der Daten. Data Preparation basiert auf einem neuartigen Methodenset, das insbesondere unter Beibehaltung der großen Datenmengen auf die Integration von Domänenwissen in die Generierung und Priorisierung von Zeitreihen abzielt. Diese Zeitreihen werden anschließend zur Modellierung der Absatzprognosemodelle verwendet. Insgesamt enthält die neue Methodik 26 Tools und Techniken, sodass in jedem Prozessschritt die Frage „wie es zu tun ist“ beantwortet wird.

In der Fallstudie hat sich der vom BDA-Buch unterstützte BDA-Manager für das Projektmanagement und die Koordination des multidisziplinären Projektteams als effektiv erwiesen. Die Rolle hat auch die inhaltliche Kommunikation mit verschiedenen Stakeholdern ermöglicht, so dass die Unterstützung von Schlüsselfunktionen wie der IT gesichert wurde. Im Business Understanding wurden elf BDA-Anwendungsfälle für die volatile Welt identifiziert und die Absatzprognose als einer der zwei relevantesten Anwendungsfälle priorisiert. Basierend auf dem neuen Big Data Sources-Schritt wurde eine Long List von 28 Datenquellen identifiziert und sukzessive auf acht Quellen mit dem gewünschten Datenmix reduziert. Während des Schritts für Data Understanding wurden insgesamt 191 Datensätze für den Anwendungsfall Absatzprognose ausgewählt, beschafft, untersucht und verifiziert. Das auf dem Hadoop-basierten Projekt-Cluster verarbeitete Datenvolumen überstieg 320 Gigabyte und somit auch das in einer Business-to-Business-Branche typischerweise für die Absatzprognose genutzte Datenvolumen. Die Zeitreihengenerierung resultierte in mehr als 4 Millionen Zeitreihen, von denen mit Hilfe von festgelegten Qualitäts- und Relevanzkriterien 1.360 als Input für die Modellierung priorisiert wurden. Beide Modellierungsansätze, Klassifizierung und Regression, zeigten eine annehmbare Performance für mittelfristige Prognosen des Absatzwachstums. Basierend auf dem Big-Data-Input erreichten Support-Vektor-Maschinen als bester Klassifikator eine Genauigkeit von bis zu 85% und das Elastic Net als ausgewähltes Regressionsmodell zeigte eine bessere Prognosegüte im Vergleich zu einem konventionellen Prognoseansatz.

Zusammenfassend betrachtet erfüllt die neue Methodik das primäre Ziel der vorgestellten Forschung mit nur geringen Einschränkungen hinsichtlich der Implementierung der Designanforderungen, wie z. B. die fehlende Berücksichtigung von Visualisierungswerkzeugen. Die Erweiterung der Methodik um weitere spezifische Teile für andere Anwendungsfälle, z.B. ein Technologie-Monitoring, und die Integration von unstrukturierten Daten in den spezifischen Teil, stellen wesentliche Verbesserungspotenziale für die künftige Forschung dar. Die beobachtete Prognoseperformance ergibt eine positive Indikation für den anvisierten Proof of Concept. Aufgrund der begrenzten Anzahl von Beobachtungen zur Modellvalidierung sind jedoch weitere Untersuchungen zur Generalisierung der Performance erforderlich. Oversampling-Strategien und fortgeschrittene Ansätze von Feature Selection sind potentiell geeignete Maßnahmen für diesen Zweck. Darüber hinaus bieten Parameter-Optimierung der Analysemodelle sowie der Einsatz von Ensemble-Modellen vielversprechendes Optimierungspotenzial für weitere Forschung bezüglich der Prognosegenauigkeit.

Table of Contents

LIST OF FIGURES	IV
LIST OF TABLES	VII
LIST OF ABBREVIATIONS	IX
1 INTRODUCTION	1
1.1 RESEARCH MOTIVATION	2
1.1.1 Situation: Volatile business environment.....	2
1.1.2 Opportunity: Big data and analytics.....	3
1.1.3 Application: Sales forecasting	4
1.2 RESEARCH DESIGN.....	6
1.2.1 Research objectives.....	6
1.2.2 Research approach.....	7
1.2.3 Research structure.....	9
2 FUNDAMENTALS	13
2.1 AGILITY	14
2.1.1 Definition.....	14
2.1.2 Corporate agility system.....	15
2.2 BIG DATA	17
2.2.1 Definition.....	17
2.2.2 Data types.....	19
2.3 ANALYTICS	21
2.3.1 Definition.....	21
2.3.2 Types of analytics	23
2.3.3 Analytics models.....	24
2.4 ANALYTICS FOR SALES FORECASTING	26
2.4.1 Traditional approaches.....	26
2.4.2 Big data analytics models	27
2.5 BIG DATA ANALYTICS FOUNDATIONS.....	32
3 RELATED WORK.....	37
3.1 RELATED RESEARCH ON SALES FORECASTING	38
3.1.1 Overview.....	38
3.1.2 Sales forecasting based on big data analytics.....	38
3.1.3 Sales forecasting in the printed circuit board industry	43

3.1.4	Interim conclusion	45
3.2	PROCESSES FOR ANALYTICS APPLICATIONS	47
3.2.1	Overview	47
3.2.2	Traditional processes.....	48
3.2.3	Advanced processes	62
3.2.4	Big data analytics processes.....	70
3.2.5	Interim conclusion	88
4	NEW METHODOLOGY	91
4.1	CONCEPT	92
4.1.1	Design considerations.....	92
4.1.2	Methodology overview	93
4.1.3	Delimitations	96
4.2	TEAM SETUP.....	98
4.2.1	Project team roles	98
4.2.2	New team roles	99
4.2.3	Roles and responsibilities	102
4.2.4	External roles	103
4.3	BUSINESS UNDERSTANDING.....	105
4.3.1	Step overview: Determine, define & select	105
4.3.2	Determine business objectives	106
4.3.3	Define and select use cases	111
4.4	BIG DATA SOURCES.....	120
4.4.1	Step overview: Identify, assess & select	120
4.4.2	Identify potential data sources	121
4.4.3	Filter and pre-assessment	124
4.4.4	Final assessment and selection	126
4.5	DATA UNDERSTANDING.....	129
4.5.1	Step overview: Select, source & describe	129
4.5.2	Dataset selection & sourcing	130
4.5.3	Data exploration and verification	138
4.6	DATA PREPARATION	140
4.6.1	Step overview: Generate & prioritize	140
4.6.2	Time series generation.....	141
4.6.3	Time series prioritization	146
4.7	MODELING & EVALUATION	166
4.7.1	Step overview: Select, build & assess	166

4.7.2	Model selection	167
4.7.3	Test design	169
4.7.4	Model building, assessment and business evaluation	170
5	EVALUATION	177
5.1	PRE-STUDY	178
5.2	CASE STUDY INTRODUCTION	179
5.2.1	Industry background	179
5.2.2	Project setup.....	180
5.3	CASE STUDY RESULTS	181
5.3.1	Team setup.....	181
5.3.2	Business understanding	182
5.3.3	Big data sources.....	186
5.3.4	Data understanding.....	192
5.3.5	Data preparation.....	200
5.3.6	Modeling & evaluation	218
6	CONCLUSION.....	237
6.1	DISCUSSION	238
6.1.1	Research summary	238
6.1.2	Methodology evaluation	238
6.1.3	Proof of concept	242
6.2	FUTURE WORK	243
6.2.1	New methodology refinements.....	243
6.2.2	Sales forecasting application refinements	244
	BIBLIOGRAPHY.....	247
	APPENDIX	295
	A – EUROSTAT AND OECD.STAT DATASETS	296
	B – TARGET DATA FORMAT.....	305

List of Figures

- FIGURE 1 - DESIGN SCIENCE RESEARCH FRAMEWORK8
- FIGURE 2 - STRUCTURE OF THE WORK10
- FIGURE 3 - CORPORATE AGILITY SYSTEM.....16
- FIGURE 4 - RELATIONSHIPS BETWEEN COMMON CONCEPTS FOR ANALYTICAL USE OF DATA22
- FIGURE 5 - TYPES OF ANALYTICS23
- FIGURE 6 - BASIC LEARNING APPROACHES AND PREDOMINANT MODEL TYPES.....24
- FIGURE 7 - NOTATION FOR SEASONAL ARIMA MODELS.....27
- FIGURE 8 - BASIC PRINCIPLE OF SUPPORT VECTOR MACHINES28
- FIGURE 9 - DECISION TREE EXAMPLE30
- FIGURE 10 - OVERVIEW OF REVIEWED RESEARCH38
- FIGURE 11 - BIG DATA-RELATED PUBLICATIONS AND HITS SINCE 200039
- FIGURE 12 - SEARCH QUERY FOR SCOPUS DATABASE40
- FIGURE 13 - RESULTS OF REVIEW PROCESS41
- FIGURE 14 - OVERVIEW OF PROCESSES48
- FIGURE 15 - KDD PROCESS50
- FIGURE 16 - CRISP-DM STEPS, TASKS AND OUTPUTS51
- FIGURE 17 - DATA MINING PROCESS.....56
- FIGURE 18 - POLL ON PROCESSES USED IN PROJECTS.....59
- FIGURE 19 - DEPLOYMENT AND DEVELOPMENT CYCLES.....63
- FIGURE 20 - DATA MINING ENGINEERING64
- FIGURE 21 - MANAGEMENT PROCESS OF MoProPEI.....65
- FIGURE 22 ADAPTIVE SOFTWARE DEVELOPMENT FOR BUSINESS INTELLIGENCE68
- FIGURE 23 - SNAIL SHELL KNOWLEDGE DISCOVERY VIA DATA ANALYTICS.....74
- FIGURE 24 - AGILE BI DELIVERY FRAMEWORK76
- FIGURE 25 - DATA SCIENCE EDGE77
- FIGURE 26 - DATA ANALYTICS LIFECYCLE.....79
- FIGURE 27 - BIG DATA ANALYTICS LIFECYCLE.....81
- FIGURE 28 - BIG – DATA, ANALYTICS, AND DECISIONS FRAMEWORK82
- FIGURE 29 - FRAMEWORK FOR IMPLEMENTATION OF BIG DATA PROJECTS83
- FIGURE 30 - BIG DATA WORKFLOW84
- FIGURE 31 - BASIC DESIGN CONCEPT OF THE NEW METHODOLOGY92
- FIGURE 32 - OVERVIEW OF THE NEW METHODOLOGY93
- FIGURE 33 - MAJOR DELIMITATIONS OF THE NEW METHODOLOGY96
- FIGURE 34 - SKILL AREAS OF PROJECT TEAM ROLES 101

FIGURE 35 - BUSINESS UNDERSTANDING STEP	105
FIGURE 36 - STRATEGIC AND OPERATIONAL VALUE OF BIG DATA ANALYTICS	106
FIGURE 37 - ANALYTICS VIEW VERSUS CAUSALITY APPROACH	109
FIGURE 38 - ADVANCED CORPORATE AGILITY SYSTEM	109
FIGURE 39 - DETERMINATION OF NEED FOR ACTION.....	111
FIGURE 40 - PORTFOLIO MATRIX FOR BDA USE CASES.....	116
FIGURE 41 - DATA SOURCES FUNNEL.....	121
FIGURE 42 - DATA SOURCE SCORING MODEL FOR FINAL ASSESSMENT	126
FIGURE 43 - DATA MIX MATRIX.....	127
FIGURE 44 - OVERVIEW DATA UNDERSTANDING STEP.....	129
FIGURE 45 - DATA HIERARCHY TAXONOMY	130
FIGURE 46 - DATA SELECTION & SOURCING FOR SENSITIVE DATASETS.....	131
FIGURE 47 - BDA BOOK (OVERVIEW SHEET).....	135
FIGURE 48 - BDA BOOK (DATASET SHEET)	135
FIGURE 49 - BDA BOOK (EXTENDED DATASET SHEET)	139
FIGURE 50 - OVERVIEW DATA PREPARATION STEP	140
FIGURE 51 - TIME SERIES GENERATOR SHEET (BDA BOOK)	143
FIGURE 52 - TIME SERIES GENERATION (EXAMPLE).....	145
FIGURE 53 - TIME SERIES GENERATOR SHEET (SIMPLIFIED VERSION)	146
FIGURE 54 - OVERVIEW OF TIME SERIES PRIORITIZATION	149
FIGURE 55 - EVALUATION TOOL REPORT	150
FIGURE 56 - CONCEPT OF CROSS-CORRELATION FOR TIME SERIES	151
FIGURE 57 - AGGREGATION FILTERS OVERVIEW	153
FIGURE 58 - EVALUATION REPORT	155
FIGURE 59 - INDIVIDUAL SCORES BASED ON EVALUATION REPORT	156
FIGURE 60 - OVERALL SCORE IN SCORING MODEL	158
FIGURE 61 - GENERAL ASSESSMENT BASED ON SCORING MODEL.....	160
FIGURE 62 - GENERAL ASSESSMENT STEP 3 (EXAMPLE)	161
FIGURE 63 - GENERAL AND DETAILED ASSESSMENTS.....	162
FIGURE 64 - TIME SERIES DECISION TEMPLATE (EXAMPLE).....	165
FIGURE 65 - MODEL SELECTION TEMPLATE.....	168
FIGURE 66 - CONCEPT OF K-FOLD CROSS-VALIDATION METHOD	170
FIGURE 67 - CONFUSION MATRIX (CLASSIFICATION)	172
FIGURE 68 - TIMELINE OF PRE-STUDY AND CASE STUDY	180
FIGURE 69 - PROJECT TEAM SETUP	182
FIGURE 70 - OVERVIEW USE CASE DEFINITION AND SELECTION.....	183
FIGURE 71 - USE CASE ASSESSMENT TEMPLATE (EXEMPLARY EXCERPT)	184

FIGURE 72 - USE CASE PORTFOLIO MATRIX.....	185
FIGURE 73 - USE CASE DECISION TEMPLATE (EXCERPT).....	186
FIGURE 74 - DATA QUERY SCOPE.....	187
FIGURE 75 - DATA SOURCE PRE-ASSESSMENT WITH DECISIVE FACTORS	189
FIGURE 76 - FINAL ASSESSMENT OF DATA SOURCES.....	191
FIGURE 77 - DATA SOURCES FUNNEL AND DATA MIX MATRIX.....	191
FIGURE 78 - CUSTOMIZABLE STRUCTURE OF FINANCIAL DATABASE.....	194
FIGURE 79 - SAMPLE ENTRY FROM BDA BOOK	196
FIGURE 80 - QUERY AND DATASET SHEETS FOR STOCK MARKET DATA A.....	197
FIGURE 81 - CONCEPT OF COMPANY LIST DEFINITION.....	198
FIGURE 82 - DATA SOURCING OVERVIEW	199
FIGURE 83 - EXTENDED BDA BOOK IN DATA EXPLORATION & VERIFICATION	200
FIGURE 84 - GENERAL STRUCTURE OF COMBINATIONS AND FILTERS (FINANCIAL DATABASE)..	202
FIGURE 85 - GENERATION RULES FOR ANALYST RECOMMENDATIONS	202
FIGURE 86 - SUMMARY TIME SERIES GENERATION.....	205
FIGURE 87 - ASSESSMENT STEP 1 RESULTS (ALL DATA SOURCES)	208
FIGURE 88 - ASSESSMENT STEP 3 RESULTS (ALL DATA SOURCES)	211
FIGURE 89 - ASSESSMENT STEP 4: EXAMPLE FOR SPURIOUS CORRELATIONS	212
FIGURE 90 - ASSESSMENT STEP 4: EXAMPLE FOR INFERIOR FILTERS	213
FIGURE 91 - ASSESSMENT STEP 4: EXAMPLE FOR SPARSE VARIABLES.....	213
FIGURE 92 - ASSESSMENT STEP 5: EXAMPLES FOR SUBSTANTIATION OF INFERIOR FILTERS AND SPARSE VARIABLES.....	215
FIGURE 93 - ASSESSMENT STEP 5: EXAMPLES FOR REMOVAL OF REDUNDANT INFORMATION..	216
FIGURE 94 - ASSESSMENT STEP 5: EXAMPLE FOR WEAKLY CORRELATED TIME SERIES.....	217
FIGURE 95 - ASSESSMENT STEPS 4 & 5 RESULTS (ALL DATA SOURCES)	218
FIGURE 96 - DEFINITIONS AND DISTRIBUTIONS OF CLASSES.....	222
FIGURE 97 - DATA CONDITIONING FOR CLASSIFICATION.....	223
FIGURE 98 - DATA CONDITIONING FOR REGRESSION	228
FIGURE 99 - OVERALL REGRESSION PERFORMANCE.....	230
FIGURE 100 - ACTUAL VERSUS FORECAST VALUES (OVERVIEW).....	231
FIGURE 101 - TARGET DATA STRUCTURE	305

List of Tables

TABLE 1 - ANALYTICAL USE OF DATA IN BUSINESS	21
TABLE 2 - OVERVIEW OF KDDM PROCESSES	49
TABLE 3 - COMPARISON OF BASIC KDDM PROCESSES	57
TABLE 4 - COMPARISON OF PROCESS STEPS	58
TABLE 5 - IMPROVEMENT AREAS FOR KDDM PROCESSES	62
TABLE 6 - BIG DATA PROJECT SUCCESS FACTORS.....	71
TABLE 7 - IMPROVEMENT AREAS FOR BDA PROCESSES	73
TABLE 8 - EVALUATION OF BDA PROCESSES	86
TABLE 9 - DESIGN REQUIREMENTS FOR NEW METHODOLOGY	89
TABLE 10 - OVERVIEW OF METHODS ON TASK LEVEL	95
TABLE 11 - SETUPS FOR PROJECT TEAM ROLES	99
TABLE 12 - WORKFLOW SUMMARY BY PROCESS STEPS	103
TABLE 13 - EXAMPLE FOR INCREASED INFORMATION BASE.....	108
TABLE 14 - USE CASE IDENTIFICATION WORKSHOP (EXEMPLARY AGENDA).....	112
TABLE 15 - COMPANY EXPERIENCE DIMENSION OF IDEA GENERATION	113
TABLE 16 - USE CASE ASSESSMENT TEMPLATE.....	115
TABLE 17 - USE CASE DECISION TEMPLATE.....	117
TABLE 18 - DATASET SELECTION SHEET	132
TABLE 19 - OPERATIONS FOR DIMENSION SUBSETS	143
TABLE 20 - DEFINITIONS OF QUALITY AND RELEVANCE MEASURES	152
TABLE 21 - AGGREGATION FILTERS SETUP.....	154
TABLE 22 - METHODS FOR MODEL PERFORMANCE VALIDATION	169
TABLE 23 - CLASSIFICATION PERFORMANCE MEASURES	173
TABLE 24 - REGRESSION PERFORMANCE MEASURES.....	174
TABLE 25 - LONG LIST OF POTENTIAL DATA SOURCES	188
TABLE 26 - SELECTED DATASETS OF FINANCIAL DATABASE.....	194
TABLE 27 - SELECTED ERP DATASETS.....	195
TABLE 28 - HYPOTHESES PER DATASET FOR ERP DATA.....	204
TABLE 29 - DEFINED FILTER SETUPS FOR EVALUATION REPORTS	206
TABLE 30 - DEFINED WEIGHTS OF THE SCORING MODEL	208
TABLE 31 - SENSITIVITY-BASED RANKING (EUROSTAT & OECD.STAT)	209
TABLE 32 - ASSESSMENT STEP 3 RESULTS FOR EUROSTAT & OECD.STAT	209
TABLE 33 - ASSESSMENT STEP 3 RESULTS FOR ERP	210
TABLE 34 - MODEL SELECTION TEMPLATE (BUSINESS USER APPROVAL).....	219

TABLE 35 - NORMALIZATION APPROACHES FOR CLASSIFICATION	222
TABLE 36 - ACCURACY PERFORMANCE FOR 2 CLASSES (BDA INPUT)	224
TABLE 37 - OVERALL CLASSIFICATION PERFORMANCE (DOMAIN INPUT)	225
TABLE 38 - EXAMPLE FOR LABEL SHUFFLING.....	225
TABLE 39 - LABEL SHUFFLING RESULTS FOR SVM(RBF) WITH 3 CLASSES	226
TABLE 40 - ACCURACY PERFORMANCE COMPARISON FOR BDA INPUT AND DOMAIN INPUT	226
TABLE 41 - ARIMA PARAMETER GRID SEARCH.....	229
TABLE 42 - SELECTED EUROSTAT DATASETS.....	304
TABLE 43 - SELECTED OECD.STAT DATASETS	304

List of Abbreviations

A

ADF. Agile BI Delivery Framework
API. Application Programming Interface
ARIMA. Auto-Regressive Integrated Moving Average
ASD. Adaptive Software Development
ASD-BI. Adaptive Software Development for Business Intelligence
ASD-DM. Adaptive Software Development for Data Mining
ASUM. Analytics Solution Unified Method
ASUM-DM. Analytics Solutions Unified Method for Data Mining/Predictive Analytics

B

B2B. Business-to-Business
BDA. Big Data Analytics
BDA book. Big Data Analytics book
BDA manager. Big Data Analytics manager
B-DAD. Big – Data, Analytics, and Decisions Framework
BDAL. Big Data Analytics Lifecycle
BDAM. Big Data Analytics Methodology
BEC. Broad Economic Activities
BI. Business Intelligence
BPM5. Balance of Payments and International Investment Position Manual – fifth edition
BPM6. Balance of Payments and International Investment Position Manual – sixth edition

C

CEO. Chief Executive Officer
CFO. Chief Financial Officer
CPI. Consumer Price Index
CPU. Central Processing Unit
CRISP-DM. Cross Industry Standard Process for Data Mining
CRM. Customer Relationship Management
CSV. Comma-Separated Values

D

DA. Data Analytics
DAL. Data Analytics Lifecycle
DBDP. Doing a Big Data Project
DDDM. Domain Driven Data Mining
DFD. Design for Deployment
DM. Data Mining
DME. Data Mining Engineering
DMIE. Data Mining for Industrial Engineering
DS. Data Science
DSE. Data Science Edge

E

EA19. Euro Area
ECU. European Currency Unit
EDP-DM. Engineering Design Process for Data Mining
EFTA. European Free Trade Association
EMS. Electronics Manufacturing Services
EMU. Economic and Monetary Union
ERP. Enterprise Resource Planning
ESA. European System of Accounts
ETL. Extract, Transform and Load
EU. European Union
EU28. European Union with 28 member states

F

FIBD. Framework for Implementation of Big Data Projects

G

GQM. Goal Question Metrics

H

HDFS. Hadoop Distributed File System
HICP. Harmonised Indices of Consumer Prices

I

IIM. Institute of Innovation and Industrial Management, Institute of Innovation and Industrial Management
IKDDM. Integrated Knowledge Discovery and Data Mining
IoT. Internet of Things
IT. Information Technology

K

KD. Knowledge Discovery
KDAA. Snail Shell Knowledge Discovery via Data Analytics
KDD. Knowledge Discovery in Databases
KDDM. Knowledge Discovery and Data Mining
KDDS. Knowledge Discovery in Data Science
KM. Knowledge Management
kNN. k-Nearest Neighbors

L

LFS. Labour Force Survey

M

MAPE. Mean Absolute Percentage Error
MoProPEI. Modelo de Proceso de Proyectos de Explotación de Información (information mining project development process model)

N

NACE. Nomenclature statistique des activités économiques dans la Communauté européenne (statistical classification of economic activities in the European Community)

O

OECD. Organisation for Economic Co-operation and Development
OEM. Original Equipment Manufacturer
OLAP. Online Analytical Processing
OLS. Ordinary Least Squares
OSAT. Outsourced Semiconductor Assembly and Test

P

PCB. Printed Circuit Boards, Printed Circuit Board
PCC. Pearson Correlation Coefficient

R

RAMSYS. RAPid collaborative data Mining SYStem
rbf. radial basis function
RMSE. Root-Mean-Square Error

S

SCM. Supply Chain Management
SEMMA. Sample, Explore, Modify, Model, and Assess
SITC. Standard International Trade Classification
SMART. Specific, Measurable, Achievable, Relevant, Time-bounded
SQL. Structured Query Language
SRM. Supplier Relationship Management
SVM. Support Vector Machines

V

VFT. Value Focused Thinking

X

XML. Extensible Markup Language

1 Introduction

1.1 Research motivation

1.1.1 Situation: Volatile business environment

The business environment today is characterized by increasing volatility and uncertainty in markets (Abele, Reinhart 2011, p. 175; Biedermann 2010, p. 23), as well as threats by major events such as economic slumps or trade embargos (Abele, Reinhart 2011, p. 19). The influence factors of permanent change therefore not only stem from the immediate market environment but also include the overall economy including the financial system, social as well as political factors, and the ecosystem (Westkämper 2007, pp. 3–4). Kremsmayr (2017) describes four key drivers behind this environment. The increasing global economic *integration* breaks down local barriers such that companies are exposed to changes in the business environment across geographic and industry borders. Furthermore, a higher level of *disruption* can be observed, especially in the form of accelerated innovation cycles and rapid technology shifts. *Granularity* describes the effect of ever higher customization of product and service offers that results in fragmented and complex market structures. Finally, *digitization* represents a mega trend with multiple effects on business models, business processes, and customer behavior (Kremsmayr 2017, pp. 47–52). The effects of the volatile business environment are as diverse as its drivers. Demand and commodity price volatilities, supply chain risks, and capital cost uncertainty are a few examples (Manyika et al. 2012, p. 69). *Welcome to the volatile world.*¹

Comin, Mulani (2004) show a long-term trend of increasing volatility in company sales, and therefore indicate that the volatile world is not a short-term phenomenon. Furthermore, Aschenbrücker et al. (2014, p. 5) report results from a survey among *Chief Financial Officers (CFOs)* from 2012 where 88% of the respondents agreed on increasing volatilities in the business environment. As a consequence, adaptability is a basis for competitiveness and success (Spath et al. 2013, p. 21). Industrial companies require "[...] the ability to cope with unexpected changes, to survive unprecedented threats of [the] business environment, and to take advantage of changes as opportunities" (Sharifi, Zhang 1999, p. 9). Such an approach to cope with the volatile world is described by the concept of *agility* (Sharifi, Zhang 2001, p. 774). Studies indicate that companies with agile characteristics have a higher level of competitiveness (Ren et al. 2003) and that agile companies outperform lean companies in terms of business performance (Yusuf, Adeleye 2002). Furthermore, Williams et al. (2013) observe that most profitable companies share agility as common characteristic as they "[...] adapt to business change more quickly and reliably than their competitors [...]". Quantitative studies described by Deubel (2017) confirm this observation. They show that agile companies characterized by a lower break-even level and

¹ Quote taken from the report "Welcome to the volatile world" (McKinsey & Company 2010).

high adaptability of costs are more profitable compared to their peer group (Deubel 2017, pp. 106–111).

Growing difficulties of companies to anticipate or predict relevant changes in the future (Abele, Reinhart 2011, p. 19) are a major challenge in the volatile business environment. The multitude of influencing factors is one of the main causes for this (Möller et al. 2016, p. 509). In particular, it becomes increasingly harder to predict market trends (Friedli, Schuh 2012, p. 13), and more specifically to forecast demand (Chase 2013a, p. 32). For example, Wilson, Demers (2015, p. 5) present a survey result where "forecast accuracy and demand variability" is considered as the top obstacle to achieve supply chain goals.

1.1.2 Opportunity: Big data and analytics

At the same time, Henke et al. (2016) describe three trends as basis for *the age of analytics*: big data, greater computation and data storage capacities, and advances in analytics (Henke et al. 2016, pp. 22–25). The most prominent characteristic of *big data* is the increasing data volume. The *International Data Corporation (IDC)* regularly estimates the *Digital Universe* as "[...] the amount of digital data created annually" (IDC 2014b). IDC (2014a) represents the seventh study and estimates the Digital Universe in 2013 with 4.4 zettabyte². This represents a 36-fold increase compared to 2005 and a further 10-fold increase is expected by 2020. About two thirds of the Digital Universe is generated by consumers with the remainder originating from companies. However, companies get involved with 85% of the consumer data (IDC 2014a). The exponential data growth is driven by a wide range of data sources "[...] such as sensors, purchase transactions and social media networks" (Wang et al. 2016, p. 747). Companies across all industries generate and collect increasing volumes of data: more than 500 million tweets per day are published on the social media platform Twitter (internet live stats 2017), retail company Wal-Mart tracks more than 267 million transactions per day across all stores worldwide (Bryant et al. 2008), and industrial company General Electric collects terabytes of data from a single jet engine (GE Reports 2015). Another important dimension of big data is the variety of data types ranging from quantitative to text and audiovisual data (Hashem et al. 2015, p. 102). Dhar (2013, p. 66) reports that text and other unstructured data grows even faster than quantitative data and therefore represents about 90% of all archived data worldwide.

In parallel, the global *computation capacity* to process data multiplied 22-fold from 2000 to 2007 and the globally installed *data storage capacity* increased 6-fold in the same time period (Hilbert, López 2011, pp. 60–64). These trends continue and the increase in data processing speed is a key characteristic as it enables new applications for the data (Henke et al. 2016, p. 24). It is equally important that costs of data storage and processing exponentially dropped over the last 15 years or so (Fogelman-Soulié, Lu 2016, p. 143). Moreover, "[the] increasing availability of

² A zettabyte equals one billion terabytes.

data has fueled advances in analytical techniques and technologies" (Henke et al. 2016, p. 23). Russom (2011, p. 9) describes the basic motivation to bring big data and *advances in analytics* together: "Most tools designed for data mining or statistical analysis tend to be optimized for large data sets. In fact, the general rule is that the larger the data sample, the more accurate are the statistics and other products of the analysis." Lavalley et al. (2010) performed a study with more than 3,000 managers and analysts. The study ascertains that "[top] performing organizations were twice as likely to use analytics to guide day-to-day operations and future strategies as lower performers" (Lavalley et al. 2010, p. 4).

Big data in combination with analytics is a new source of competitive advantage as it is "[...] considered as a game changer enabling improved business efficiency and effectiveness because of its high operational and strategic potential" (Wamba et al. 2017, p. 357). Faster, better and proactive decision making is the key driver of the strategic value (Hagen et al. 2013, p. 4) while operational benefits are many and varied. They range from targeted marketing to fraud detection to manufacturing yield optimization (Russom 2011, p. 11). Based on a global cross-industry survey, Russom (2011, p. 11) furthermore reports "recognition of sales and market opportunities", "quantification of risks", "trending for market sentiments", "understanding of business change", and "better planning and forecasting" as expected benefits from big data analytics. Bange et al. (2015, p. 15) additionally identify "better understanding of the market and competition" as one of the top realized benefits. These studies underline the opportunity to gain a better understanding of the volatile business environment based on big data and analytics. However, big data analytics is "[...] still in its infancy [...]" (Shi-Nash, Hardoon 2017, p. 337) and 86% of surveyed companies have not started a big data initiative or are still in the stage of pilot projects (Bange et al. 2015, p. 12). The research company Gartner furthermore predicted that more than 85% of the world's largest companies will fail to generate a competitive advantage based on big data (Gartner 2011). As a consequence, the outlined opportunity represents a challenge at the same time.

1.1.3 Application: Sales forecasting

Sales forecasting "[...] is a critical function that influences companies worldwide across all industries" (Chase 2013a, p. 31). Chase (2013b, p. 28) reports that improvements of sales forecasting is the top priority for supply chain executives, for instance. According to Rey et al. (2012), it is not only supply chain management that benefits from superior forecasts. Long-term forecasts build the basis for strategic planning, medium-term forecasts enable resource and asset management, and short-term forecasts support marketing decisions. Furthermore, sales forecasts facilitate a better understanding of the market for managers. Forecasting capabilities can be seen as source of competitive advantage because of this wide area of application (Rey et al. 2012, p. 2). For instance, "[f]orecasting sales and demand over 6-24 month horizon is crucial [...]" for production planning in industries with complex processes such as the electronics industry (Sa-ngasoongsong et al. 2012, 875). The benefits from quality forecasting range from

increased revenues and efficiency to decreased costs to higher customer satisfaction (Chang et al. 2009, p. 344).

However, Chang, Lai (2005) observe that traditional forecasting approaches seem to be increasingly inadequate in the current business environment. It is more and more difficult to capture the relations between sales and its influencing factors. Furthermore, the constant change in the business environment poses a challenge to traditional approaches (Chang, Lai 2005, p. 948). As a consequence, the development of forecasting methods is seen as one of the top challenges to be addressed by big data (Bange et al. 2015, p. 13). Halper (2014, p. 7) reports survey results where companies state that predicting trends is the most important driver to employ big data analytics. Executives therefore plan to invest in new sales forecasting solutions because they "[...] believe big data to be a forecasting priority for the future" (Chase 2013b, p. 28). In summary, sales forecasting can be considered as important application of big data analytics for a better understanding of the volatile world.

1.2 Research design

1.2.1 Research objectives

Industrial companies face the challenges of the volatile world and need to react by concepts of adaptability, for instance, in the form of agility. The anticipation of changes is a specific challenge in the volatile business environment. At the same time, the age of analytics offers an opportunity to gain a better understanding of this environment. However, industrial companies are still at the beginning of utilizing big data and advances in analytics. In addition, sales forecasting is a diverse source of competitiveness for industrial companies. The anticipation of changes in sales is a specific challenge in the volatile world and traditional approaches are not adequate anymore. Companies seek for advances in forecasting capabilities and regard big data analytics as promising approach. Furthermore, the case study performed with an industrial company identifies sales forecasting as priority application of big data analytics in the volatile world. *Situation, opportunity and application* imply the *research hypothesis* of the present work:

Big data analytics, especially as application for sales forecasting, is a novel approach to improve the understanding of the volatile business environment.

The primary objective of this research is to provide a methodology to develop a big data analytics application for sales forecasting that enables a better understanding of the volatile world. However, the remarks on the current situation have shown that volatilities are not only relevant on the sales market. The methodology therefore should not assume sales forecasting as a given case of application, but rather provide support on linking the challenges of the volatile world with the opportunities of big data analytics. Furthermore, research on sales forecasting based on big data analytics is still at an early stage for industrial companies. For this reason, a proof of concept for this novel sales forecasting approach is the secondary objective of this research. The following research questions provide guidance for the research towards these objectives:

- *Research question 1:* How can industrial companies decide where to use big data analytics in order to gain a better understanding of the volatile business environment?
- *Research question 2:* How can industrial companies develop a big data analytics application for sales forecasting?
- *Research question 3:* Does a big data analytics approach for sales forecasting work in industrial practice?

1.2.2 Research approach

1.2.2.1 Selection of research approach

The research approach of this work follows the *design science* paradigm as introduced by Simon (1996). Design science "[...] supports a pragmatic research paradigm that calls for the creation of innovative artifacts to solve real-world problems" (Hevner, Chatterjee 2010, p. 9). The notion of design as "[...] act of creating an explicitly applicable solution to a problem [...]" is a widely accepted research approach in engineering disciplines (Peffers et al. 2007, p. 47). This work is specifically based on the design science research framework for information systems proposed by Hevner et al. (2004). In general, information systems aim to improve the effectiveness and efficiency of an organization (Hevner et al. 2004, p. 76). Research in information sciences typically includes multiple disciplines in order to "[...] solve problems at the intersection of *information technology (IT)* and organizations" (Peffers et al. 2007, p. 46). Hevner, Chatterjee (2010) emphasize the difference towards computer science and software engineering. While these two disciplines rather focus on software code and development, respectively, information systems are "[...] closer to deployment of information technology in an organization" (Hevner, Chatterjee 2010, p. 7). Guarino (1998) describes information systems as a combination of application programs, information resources and user interfaces. The integration of these components serves a specific business purpose (Guarino 1998, p. 10). The implementation of analytics (*application program*) on the basis of big data (*information resource*) for a better understanding of the volatile business environment (*business purpose*) therefore represents a specific information system³ considered in the presented research. Such an information system is referred to as *big data analytics application* in this work.

1.2.2.2 Description of research approach

The following describes the research framework based on the initial proposal by Hevner et al. (2004) whereof Figure 1 provides an overview. The *environment* describes the application domain including people, organizations, and technology. Objectives, issues, and opportunities that exist in the environment define business needs. Taking into consideration existing technologies, applications or capabilities, these needs provide the basis to determine a definite *business need* defining the research problem. "Design science addresses research through the building and evaluation of artifacts designed to meet the identified business need" (Hevner et al. 2004, pp. 79–80). *Relevance* is the key characteristic for the relation between the environment and research work. The focus on a business need underlines the emphasis on "[...] practical significance of the outputs of design science work" (Sharma 2008, p. 92).

³ The user interface component plays a subordinate role in this work as the deployment of the application (information system) is not in focus. More details regarding this delimitation are discussed in *Section 4.1*.

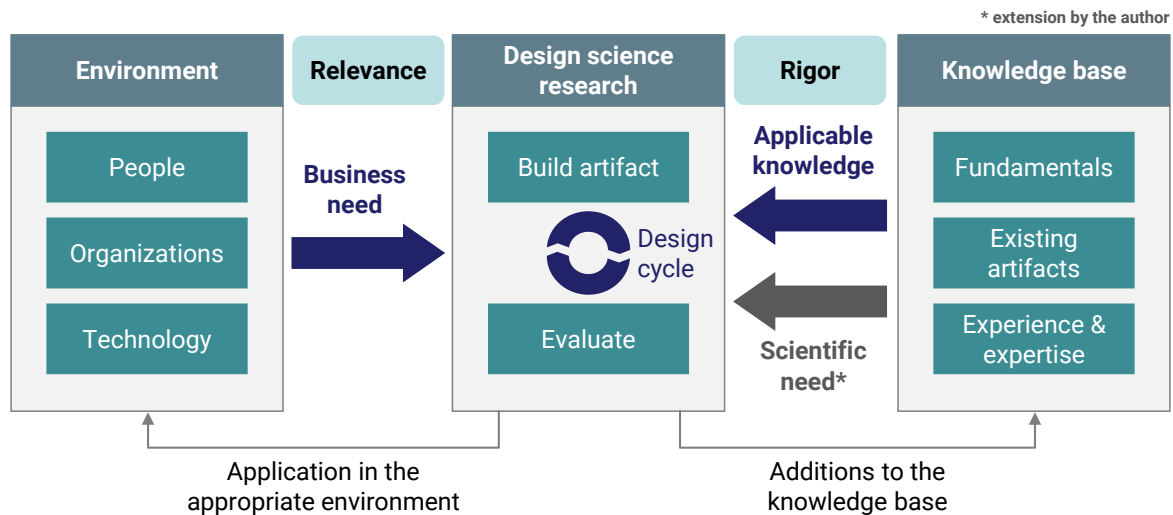


Figure 1 - Design science research framework
[based on (Hevner et al. 2004, p. 80; Hevner 2007, p. 88)]

Build and *evaluate* represent the core of the research process and they are performed in a *design cycle* that iterates "[...] between the construction of an artifact, its evaluation, and subsequent feedback to refine the design further" (Hevner 2007, p. 90). *Refinements* can be directly implemented in the design cycle, but generally are also described as future research directions (Hevner et al. 2004, p. 80). According to Hevner, Chatterjee (2010, pp. 17–18), the *knowledge base* comprises three types of *applicable knowledge*: (1) extensive fundamentals of existing scientific theories and methods, (2) existing artifacts in the research domain as well as (3) domain experience and expertise. *Rigor* describes the appropriate application of elements from the knowledge base during the research process (Hevner et al. 2004, p. 80). *Additions to the knowledge base* represent the result of design science research and includes advancements of theories and methods, new artifacts, and experiences from the research process itself (Hevner, Chatterjee 2010, p. 18). The implementation of the research approach in this work furthermore considers deficiencies identified during the review of the knowledge base as *scientific need*.

Hevner, Chatterjee (2010, p. 6) describe an *artifact* based on the work of Simon (1996) as "[...] something that is artificial, or constructed by humans, as opposed to something that occurs naturally." According to March, Smith (1995), four basic types of artifacts exist in design science research on information systems. *Constructs* establish specific vocabulary and shared knowledge in order to describe problems and solutions of a domain. Furthermore, a *model* provides a descriptive representation for relations between constructs. "A *method* is a set of steps (an algorithm or guideline) used to perform a task" (March, Smith 1995, p. 257). Finally, the realization of artifacts in the environment is called *instantiation*. It represents an operational implementation of a construct, model, or method (March, Smith 1995, pp. 256–258). In this work, the artifact type 'method' in form of a guideline is referred to as process model. Moreover,

a *methodology* rests on a process model and additionally operationalizes its steps.⁴ The artifact of the present research comprises an integrated methodology to decide on appropriate applications based on big data analytics for a better understanding of the volatile world and to develop a specific application for sales forecasting. The presented design science research approach has been previously applied for similar artifacts: an integrated analytics process model (Sharma 2008), a model-based method for analytics on process data in production processes (Wieland, Fischer 2013), a methodology for analytics with big data from social media (Asamoah, Sharda 2015), and a framework for integration of big data and analytics into the decision making process (Elgendy, Elragal 2016).

March, Smith (1995, p. 254) define *evaluation* as "[...] the process of determining how well the artifact performs." This research follows the case study approach as design evaluation method. A case study represents an in-depth study of the artifact in the business environment (Hevner et al. 2004, p. 86). The *case study* was performed with a manufacturer of printed circuit boards where the methodology was gradually built and evaluated. It therefore builds the basis for answers to *research questions 1 & 2*. Furthermore, the evaluation of the methodology results in a realization of an application for sales forecasting. This provides the basis for a *proof of concept* regarding the big data analytics approach to sales forecasting and thus enables an answer to *research question 3*. In addition to the case study, a *pre-study* with different industrial companies was performed beforehand in order to *validate the business need*. The author was also part of a research group on the topic of agility for industrial companies. This *Agility Research Group* comprised researchers from the *Institute of Innovation and Industrial Management (IIM)* at the Graz University of Technology and practitioners from industry. The research work performed between 2014 and 2017 resulted in a novel concept of agility published in Ramsauer et al. (2017). Focus of the author's research contribution was monitoring of the volatile business environment and integration of big data into the agility concept. Research results from the Agility Research Group are incorporated as substantiation of the knowledge base.

1.2.3 Research structure

The structure of this work, as shown in Figure 2, reflects the implementation of the design science research approach of the presented research. **Section 1.1** of this chapter provides the motivation of this research and determines the *business need* that can be split into two dimensions. The presented situation and opportunity define the general business need for a methodology to determine applications based on big data analytics that provide a better understanding of the volatile business environment. Sales forecasting represents a particular application in this context and therefore represents a specific business need. **Chapter 2** provides the relevant *fundamentals* as first part of *applicable knowledge* for this research. The presented agility concept

⁴ More details on the definition of *process model* and *methodology* are discussed in *Section 3.2.1*.

further substantiates the understanding of the situation of industrial companies in the volatile world and serves as design element of the new methodology. The remaining sections provide the required fundamentals from the big data and analytics domains. This also includes analytics for sales forecasting and general foundations for big data analytics with regard to technology, organization and culture. **Chapter 3** discusses related work on *existing artifacts* in order to determine the second part of *applicable knowledge*. The interim conclusions of this chapter furthermore substantiate the *scientific need* related to the research questions addressed by this work. **Section 3.1** examines research on sales forecasting related to this work. The section is divided into sales forecasting based on big data analytics and sales forecasting in the printed circuit board industry. Both sections provide an understanding of existing big data analytics applications and latest advancements in the industry of the case study, respectively. **Section 3.2** provides a comprehensive study on relevant process models and methodologies that are summarized under the term 'processes'. This study provides the basis for the methodology design and furthermore identifies design requirements to be addressed by the new methodology design. The *design cycle* as main part of this work is jointly represented by the subsequent two chapters.

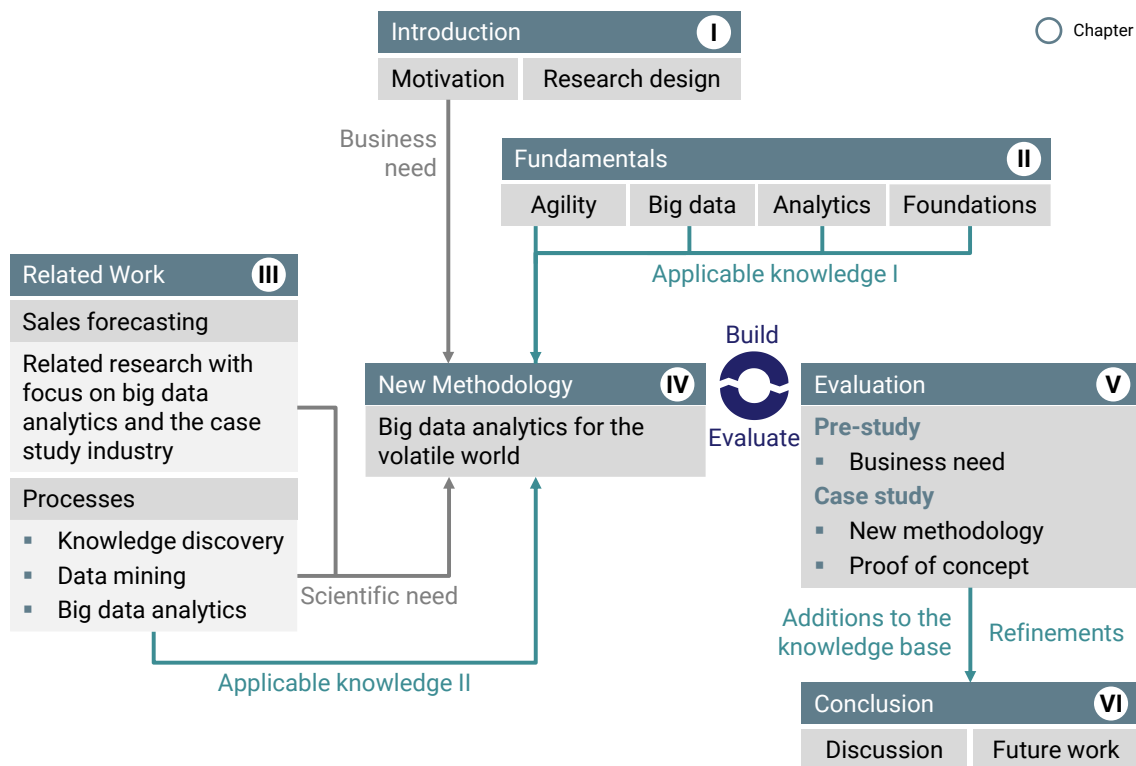


Figure 2 - Structure of the work

Chapter 4 describes the *built artifact* of the research. The chapter introduces the design concept and discusses delimitations of the new methodology. A detailed description of the methodology design is organized by individual sections for each major design element. **Chapter 5** begins with a brief summary of the pre-study including findings regarding the business need, followed by an introduction to the case study. The case study results *evaluate the artifact* in the same step-by-step

structure as the previous chapter. Besides this methodology evaluation, the case study results also include the proof of concept based on the realized big data application for sales forecasting. **Chapter 6** concludes this work with a discussion of the *additions to the knowledge base*. The discussion rests on the results from the methodology evaluation including the proof of concept. Furthermore, an outlook on future work for *refinements* of the research results is provided. The thesis follows the guideline for communication of design science research proposed by Hevner et al. (2004, p. 90):

"Design-science research must be presented both to technology-oriented as well as management-oriented audiences. Technology-oriented audiences need sufficient detail to enable the described artifact to be constructed (implemented) and used within an appropriate organizational context. [...] Management-oriented audiences need sufficient detail to determine if the organizational resources should be committed to constructing (or purchasing) and using the artifact within their specific organizational context."

2 Fundamentals

2.1 Agility

2.1.1 Definition

Industrial companies need to handle the volatile business environment by some form of adaptability or changeability (Wiendahl et al. 2007, pp. 783–785). Flexibility, transformability⁵ and agility are major concepts addressing this issue (Schurig et al. 2014, p. 957). For instance, Toni, Tonchia (1998, p. 1609) describe *strategic flexibility* as "[...] the firm's ability to successfully vary the mix of its competitive priorities or businesses [...]" and *operational flexibility* as "[...] the ability to positively react to the internal and external changes as these occur". Westkämper (1999) describes *transformability* in the context of manufacturing as variable structures and processes. Elements of the manufacturing system should be adaptable to a changing business environment in an anticipatory manner (Westkämper 1999, 131–133). In contrast to operational flexibility, transformability is not restricted to an a priori defined extent of necessary adaptations (Reinhart et al. 1999, p. 22) and therefore can be seen as a concept advancement. *Agility* is based on the concept of transformability with regard to manufacturing, however, it also includes a strategic component such that other business functions, for example, sales or purchasing, are also considered (Heinen et al. 2008, p. 25). These selected definitions illustrate that the concept of agility incorporates flexibility and transformability.⁶ Schurig (2016) and Rabitsch (2016) provide a comprehensive overview and comparison on existing concepts in scientific literature, and their research confirms this view. The following therefore presents a detailed definition of agility and an overview on one of the most recent agility concepts.

Rabitsch, Ramsauer (2015) describe three key characteristics of agility based on the understanding that flexibility and transformability are included. *Proactive preparation* implies that companies think ahead what changes in the business environment could occur. It also includes the preparation of alternative options for action in the identified scenarios of change. *Fast reaction* rests on rapid implementation of company reactions which includes straightforward decision making and processes. In addition, *optimized profitability* underlines that agility is not an end in itself. Following the principle of agility needs to serve superior objectives in alignment with the company's strategy. Objectives can range from increased profits to more stable cash flow to higher market share, for example (Rabitsch, Ramsauer 2015, pp. 2–3). It is important to note that agility puts equal focus on risks as well as opportunities arising from the volatile world (Heldmann et al. 2015, p. 35). Schurig (2017, p. 79) adds the prerequisite that agility requires

⁵ Transformability is mainly discussed in German-speaking literature and translates into "Wandlungsfähigkeit" (Rabitsch 2016, p. 23).

⁶ More holistic approaches are also discussed in relation to transformability. For example, Baumgartner et al. (2006) describe a *generic management* approach including strategic and cultural components.

consideration of external organizations beyond the boundaries of a company, such as suppliers. The definition of agility used in this work is based on this previous research and is as follows:

Agility describes proactive preparation for opportunities as well as risks in the volatile business environment and fast reaction to occurring changes. Agile companies strive to improve their long-term business success and consider all elements of their value chain while doing so.

2.1.2 Corporate agility system

The concept of agility requires a structured approach for implementation in business practice (Rabitsch et al. 2015, pp. 48–49). The Agility Research Group therefore developed the *corporate agility system* that is comprehensively described in Ramsauer et al. (2017). Heldmann et al. (2015) provide an overview on the corporate agility system based on its key building blocks: monitoring, control, and agility levers.

Monitoring represents the interface to the volatile business environment and aims for early detection of relevant changes. Insights from monitoring build the basis for strategic and operational control of the company. *Strategic control* comprises adaptations of strategy, targets, and budgets. For example, an industrial company adapts its product offering due to an innovative manufacturing technology and consequently adjusts its targets for market development as well as accommodates the budget to account for the product line expansion. These steps require fast implementation which is ensured by activating *agility levers*. Agility levers represent measures that help to increase the agility of the company. To put it differently, agility levers enable fast reaction to changes. Dynamic budgeting, based on short-term budget contracts and synchronization with financial planning, is an exemplary lever in the given scenario. The time horizon of *operational control* is clearly shorter such that agility levers are directly activated where necessary. In case of a jump in product demand, modular production equipment and a temporary shift of workforces allow for a short-term adaption of production capacity. Straightforward processes how to react towards indicated changes are required. Control therefore defines rules and competencies for decision making. Furthermore, relevant agility levers need to be identified and prepared prior to changes.⁷

Figure 3 provides an overview of the corporate agility system as described before. In addition, Rabitsch (2017) describes how companies can define an adequate level of agility based on its individual situation with regard to the volatile business environment. Wampula (2017) furthermore adds features of an agile organization and corporate culture, for instance, in the form of project-based resource management or an entrepreneurial mindset.

⁷ This paragraph is a free translation of prior work of the author (Heldmann et al. 2015, p. 36).

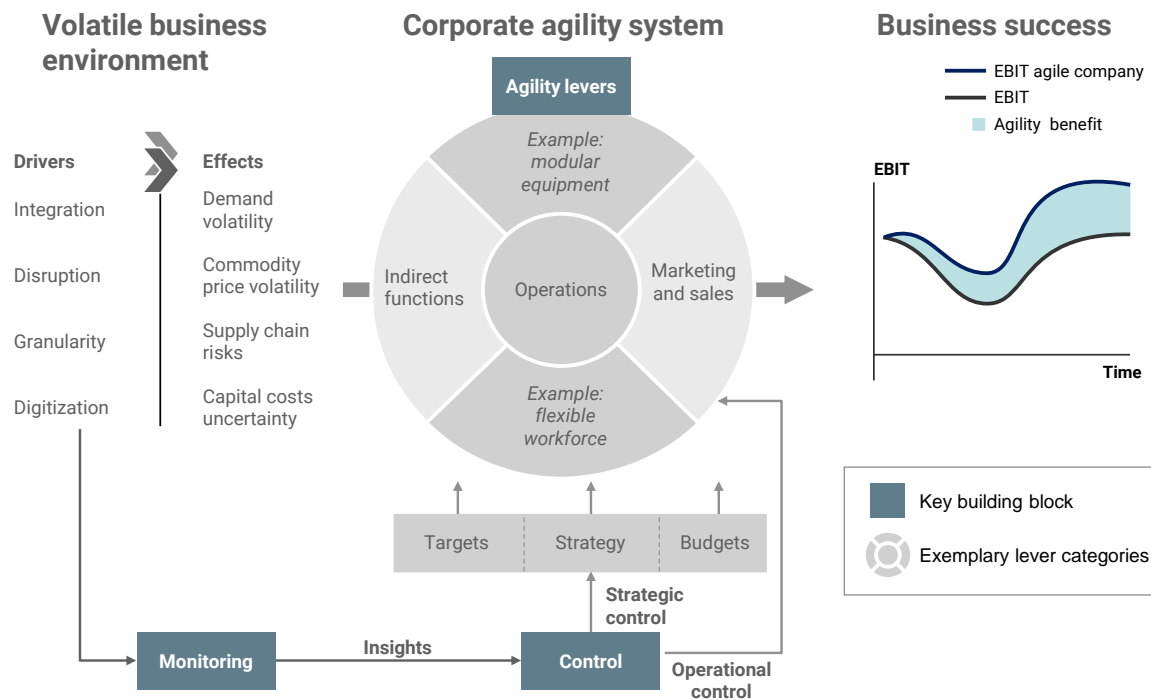


Figure 3 - Corporate agility system [based on (Heldmann et al. 2015, p. 36; Luczak 2017, p. 21)]

Heldmann (2017) underlines the essential role of monitoring in the corporate agility system because detection of changes is a necessary prerequisite for proactive reactions. Furthermore, the earlier a company detects changes, the faster it can react. The main functions of monitoring are collection and processing of information about the business environment. This ranges from a targeted search for qualitative information, that helps to identify innovative technologies, to utilization of available quantitative information as basis for sales forecasting, for example (Heldmann 2017, pp. 162–166). Monitoring also represents a field of application for big data and analytics because it benefits from an extended information base and advanced methods to process this information (Heldmann 2017, pp. 182–185). The remainder of this chapter provides the fundamentals for big data and analytics.

2.2 Big data

2.2.1 Definition

The origin of *big data* can be established in various ways according to a discussion of historical reviews by Wu et al. (2016, pp. 4–5). For example, Press (2013) provides a list of milestones that describe the notion of big data as increasing data volume. The list starts with an estimation that the size of university libraries regularly redoubles over a time period of sixteen years (Rider 1944). Another work in the discussion, places the start of big data at the U.S. Census in 1880 characterized by an information overload that lead to a processing time of eight years (Winshuttle 2017). Furthermore, other works establish the origin of big data at the time where the term 'big data' was used for the first time, and the work of Cox, Ellsworth (1997) represents a potential candidate for the introduction of 'big data' in accordance with its modern meaning (Wu et al. 2016, p. 5). Cox, Ellsworth (1997, p. 235) formulate the following with regard to the challenge of data visualization: "[...] data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*."

While the exact determination of the origin of big data is difficult due to the generic nature of the term (Lohr 2013), multiple definitions for big data exist today. Overviews of definitions can be found in NIST Big Data Public Working Group (2015, pp. 10–11), Wamba et al. (2015, p. 236) or Wu et al. (2016, p. 10). They range from highlighting challenges for handling and processing data (Fisher et al. 2012, p. 53) to the notion of a cultural shift in decision making (Dutcher 2014). Baars, Kemper (2015) provide a categorization of perspectives towards big data. The practice-oriented understanding focuses on the strategic usage of big data and does not relate to specific issues or technologies. This perspective raises awareness but does not provide a clear definition. A technology-oriented understanding delineates novel technologies for big data from those that are not adequate anymore. Finally, the problem-oriented understanding puts challenges related to the utilization of big data into the focus (Baars, Kemper 2015, pp. 223–224). A commonly used set of big data definitions stems from the problem-oriented perspective, which has its origin in the work of Laney (2001) that structures "[...] challenges along three dimensions: volume, velocity, and variety". This so-called $3V$ definition of big data is formulated by Gartner (2017) as follows:

"Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

McAfee, Brynjolfsson (2012) provide a description for each dimension from a business perspective. *Volume* describes the fact that data is created at an increasing rate today which gives companies the opportunity to collect and utilize large amounts of data. Furthermore, *velocity* refers to "[...] the speed of data creation [...]" (McAfee, Brynjolfsson 2012, p. 64) that enables real-time applications for companies. *Variety* represents the use of heterogeneous data types

from different data sources (McAfee, Brynjolfsson 2012, pp. 63–64). The 3V definition has been extended by "[...] other dimensions of big data [...]" (Gandomi, Haider 2015, p. 139). IBM (2017b) provides a 4V definition that additionally includes *veracity* referring to the "[...] quality or trustworthiness of the data." This dimension is typical for certain data sources, for example, "[...] customer sentiments in social media are uncertain in nature, since they entail human judgment" (Gandomi, Haider 2015, p. 139). Moreover, Oracle (2013) introduces *value* as additional dimension of big data which is motivated by varying economic value of diverse data. Oracle (2013, p. 4) additionally observes:

"Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is identifying what is valuable and then transforming and extracting that data for analysis."

Demchenko et al. (2013) summarize the previously discussed dimensions as a 5V definition of big data. According to Wu et al. (2016), there exist further definitions including up to eleven dimensions. For example, *visibility* "[...] emphasizes that you need to have a full picture of data in order to make informative decision[s]" (Wu et al. 2016, p. 9) or *variability* describes "[...] the variation in the data flow rates" (Gandomi, Haider 2015, p. 139).

The 3V definition is "[...] largely found in the literature [...]" (Kacfeh Emani et al. 2015, p. 71) and overviews on definitions often focus on the 5V definition (Shim et al. 2015, p. 799; Wang et al. 2016, p. 750) because it "[...] provides a straightforward and widely accepted definition related to what is (and what is not) a big-data-based problem, application, software, or framework" (Bello-Orgaz et al. 2016, p. 45). The 5V definition can also be expressed as "[...] dealing effectively with *Big Data* requires one to create *value* against the *volume*, *variety* and *veracity* of data while it is still in motion (*velocity*) [...]" (Kacfeh Emani et al. 2015, p. 72). This underlines the specific role of the value dimension, because it is typically understood as the primary objective of utilizing big data, which goes beyond the narrower definition by Oracle (2013). Kacfeh Emani et al. (2015, p. 72) describe value in "[...] two categories: analytical use (replacing/supporting human decision, discovering needs, segmenting populations to customize actions) and enabling new business models, products and services." Based on this understanding, value is dependent of some form of utilization. The definition of big data in this work consequently includes the 4V dimensions only, because they represent "[...] the primary aspects of Big Data" (Ohlhorst 2013, p. 3) and they relate to direct challenges for the application of big data (Dorschel 2015, p. 7). Moreover, Manyika et al. (2011) argue that the definition of big should be subjective in order to account for technological advancement and differences among industries. With regard to the volume dimensions, big data therefore "[...] refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al. 2011, p. 1).

In conclusion, the definition of big data used in this work extends the idea of subjectivity as follows:

Big data represents data with any or all of the following characteristics: volume, velocity, variety, and veracity, whereby each characteristic exceeds typical capabilities for a particular application in a specific industry.

2.2.2 Data types

Big data can be "[...] classified into different categories to better understand their characteristics" (Hashem et al. 2015, p. 100) and data sources as well as data structures are two key classifications (Hashem et al. 2015, p. 101). Data sources generally divide into internal and external sources. *Internal sources* include "[...] private or proprietary data that is collected and owned by the business where you control access" (Marr 2015, p. 62). Transactions are one of the most important types of internal business data and includes customer interactions such as orders or payments, for example (Baesens 2014, p. 14). Machines and sensors are further examples for sources of internal data (Hashem et al. 2015, p. 102), in particular for manufacturing companies. The increasing datafication in manufacturing due to the advent of Industry 4.0 underlines the importance of machine-generated data for applications such as predictive maintenance (Ramsauer 2013, p. 11) that can also benefit from other internal sources such as systems for production planning or quality management (Biedermann 2016a, pp. 13–14). Such *smart maintenance* systems can benefit from a big data approach to multiple data sources of a company (Biedermann 2016b, p. 134). *External sources* comprise "[...] the infinite array of information that exists outside your business" (Marr 2015, p. 63). Public and private are two subtypes of external sources (Ohlhorst 2013, p. 37) and Marr (2015, p. 63) describes them as follows:

"Public data is data that anyone can obtain – either by collecting it for free, paying a third party for it or getting a third party to collect it for you. Private data is usually something you would need to source and pay for from another business or third party data supplier."

Macroeconomic data represents commonly used public data, while financial data such as credit scores is typically sourced from specialized data suppliers (Baesens 2014, pp. 14–15). The internet is a vast source of external data including data from online searches, blogs, or forums (Chen et al. 2014, p. 179). Social media data from platforms such as Twitter is a special form of internet data (Baesens 2014, p. 15). Whether internet data is public or private depends on the individual terms of use for each source. Another nascent source of big data is the *Internet of Things (IoT)* (Chen et al. 2014, p. 179) that consists of a large number of various devices generating data as they "[...] sense, communicate, compute and potentially actuate" (Rao et al. 2012, p. 374).

The extent of big data types is substantial ranging from clickstream or interaction data from social media platforms to machine data in the form of sensor readings further to billing records as transaction data and "[...] human-generated data, including vast quantities of [...] data such as [...] voice recordings, emails, paper documents, surveys, and electronic medical records" (Shim et al. 2015, p. 800). EMC Education Services (2015) describe four types of data structures that help to classify different data types. *Structured data* has a well-defined format and a spreadsheet represents the simplest form. A common example from business is transactional data in traditional databases. *Semi-structured data* describes text data with a self-descriptive format and can be found in markup languages with a rule schema, for instance, in the case of *Extensible Markup Language (XML)*. The data type is called *quasi-structured* if additional effort is required to create a common format. For example, clickstream data from internet sources potentially have inconsistent formats. Finally, *unstructured data* provides no implicit structure and includes general text documents as well as audio, video, and image data (EMC Education Services 2015, pp. 5–7). In this work, all three types that show a lack of structure are summarized as unstructured data.

2.3 Analytics

2.3.1 Definition

In accordance with the previous discussion regarding the value dimension of big data, Franks (2012, p. 6) states that "[n]either the fact that big data is big nor the fact that it is data adds any inherent value. The value is in how you analyze and act upon the data to improve your business." Based on the value definition of Kacfeh Emani et al. (2015), the focus of this work lies on the analytical use of big data instead of creation of new business models or products based on big data. According to Davenport (2014), the idea to analyze data in order to achieve business improvements is not new, but has changed over time. It already started in the 1970s with the concept of decision support and recently entered the era of big data (Davenport 2014, p. 10). Table 1 summarizes this development.

Concept	Period	Meaning
Decision Support	1970-1985	Data analysis to support decision making
Executive Support	1980-1990	Focus on data analysis for decisions by management
Online Analytical Processing (OLAP)	1990-2000	Software for analysis of multidimensional structured data
Business Intelligence	1989-2005	Supportive systems for data-based decisions with focus on reporting
Analytics	2005-2010	Focus on statistical and mathematical analysis of data for decision making
Big Data	2010-today	Focus on data with 4V characteristics

Table 1 - Analytical use of data in business [with minor adaptations from Davenport (2014, p. 10)]

Although the concept of big data induces a shift in focus towards data, analytics are still required and earlier concepts such as business intelligence are not obsolete. Analytics can be generally defined as "[...] discovery of meaningful patterns in data [...]" (NIST Big Data Public Working Group 2015, p. 8) or more specifically as "[...] the scientific process of transforming data into insights for making better decisions" (informatics 2017). However, a clear definition for this work requires a systematic classification of different concepts that are discussed with regard to analytical use of data today. Dedic, Stanier (2017) provide a comprehensive framework shown in Figure 4, that describes the relationships between most common concepts.

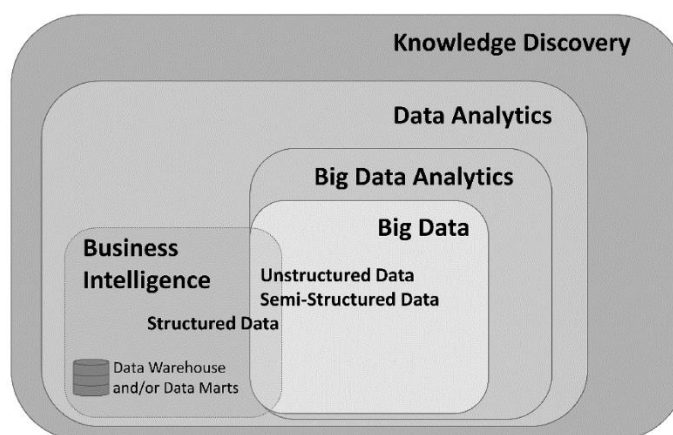


Figure 4 - Relationships between common concepts for analytical use of data
(Dedic, Stanier 2017, p. 115)

Knowledge Discovery (KD) builds on the concepts of *Knowledge Discovery in Databases (KDD)* and *Data Mining (DM)*. KDD describes the "[...] overall process of discovering useful knowledge from data" (Fayyad et al. 1996c, p. 28) whereby "[d]ata mining is a step in the KDD process that consists of applying data analysis and discovery algorithms [...]" (Fayyad et al. 1996a, p. 40). Other steps in the KDD process range from data selection to interpretation of results that are all required to create value from the data (Fayyad et al. 1996c, p. 28). In contrast to this notion, DM is alternatively used as synonym for KDD (Chen et al. 1996, p. 866), which is also reflected by the definition of *Knowledge Discovery and Data Mining (KDDM)* as "[...] series of activities to discover or identify knowledge of domain(s) from databases" (Barclay, Osei-Bryson 2015, p. 2). Another concept closely related to DM is *Data Science (DS)* which can be defined as "[...] a set of fundamental principles that support and guide the principled extraction of information and knowledge from data" (Provost, Fawcett 2013a, p. 52).

Data Analytics (DA) is defined as "[...] any activity that involves applying an analytical process to data to derive insight from the data" (Ridge 2015, p. 4). Runkler (2016, p. 2) describes DA in more detail:

"Data analytics is defined as the application of computer systems to the analysis of large data sets for the support of decisions. Data analytics is a very interdisciplinary field that has adopted aspects from other scientific disciplines such as statistics, machine learning, pattern recognition, system theory, operations research, or artificial intelligence."

In the business domain, DA is often referred to as *business analytics* (Chamoni, Gluchowski 2017, p. 9). KD is regarded "[...] as a higher entity encompassing DA, which is not exclusively related only to computer-based concepts" (Dedic, Stanier 2017, p. 118). *Big Data Analytics (BDA)* simply describes the application of specific DA techniques to big data (Elgendy, Elragal 2014, p. 215), and therefore builds a subset in the framework. The major difference of *Business Intelligence (BI)* is the focus on structured data from traditional data sources such as data warehouses or data marts (Dedic, Stanier 2017, p. 119), whereas BDA operates along the 4V dimensions and

leverages a variety of data sources. This work simply refers to DA and BDA as *analytics* independent from data characteristics and the domain.

2.3.2 Types of analytics

While a wide range of analytics approaches exists, they can be grouped by different types of analytics. As illustrated by Figure 5, Lanquillon, Mallow (2015a) provide a categorization based on the respective problem type. *Descriptive analytics* apply to problems addressing the question “*What happened?*” and reporting is a typical approach here. In case the question rather is “*What happens now?*”, *real-time analytics* can be employed for monitoring purposes. *Diagnostic analytics* provide answers to “*Why did it happen?*” in the form of root cause analysis where OLAP can be employed. Another type is *predictive analytics* and addresses the question “*What will happen?*”, for example, by predictions about possible future sales. At the highest level of decision making support, *prescriptive analytics* relates to the question “*What should be done?*”. This type of analytics is often based on optimization methods or simulations, in combination with descriptive and predictive analytics, to provide recommendations for action (Lanquillon, Mallow 2015a, pp. 56–57).

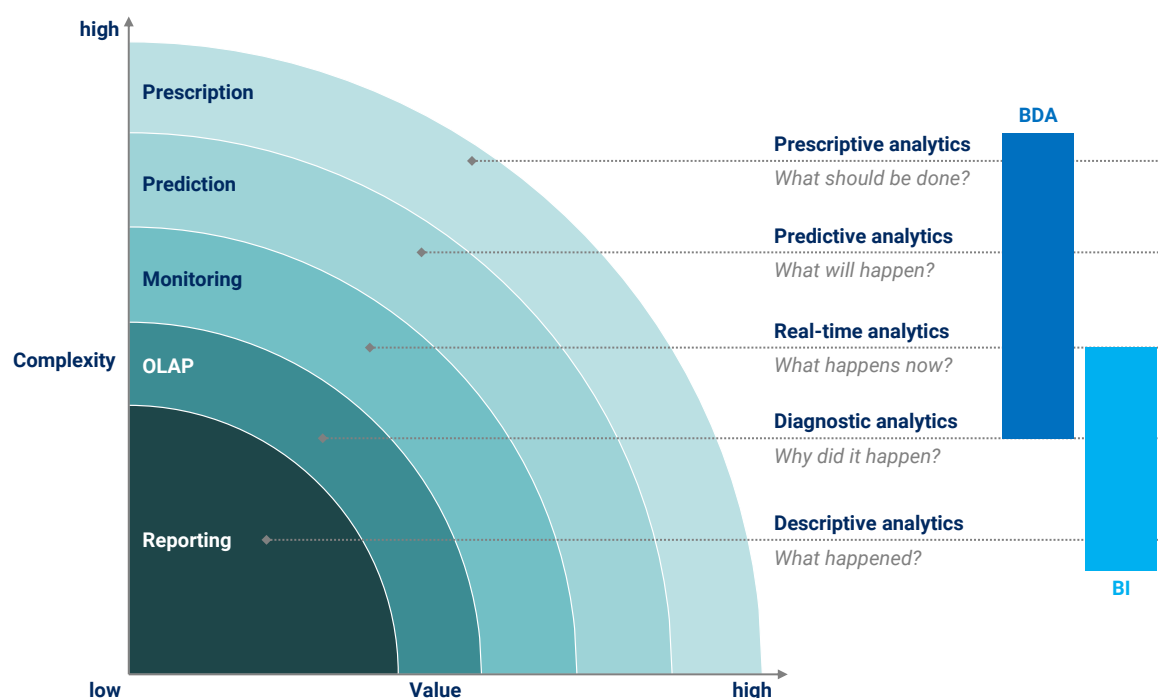


Figure 5 - Types of analytics

[based on Lanquillon, Mallow (2015a, p. 56) and Ereth, Kemper (2016, p. 459)]

These categories also reflect the distinction between concepts for analytical use of data. Descriptive analytics builds the traditional core of BI and BDA focuses on predictive as well as prescriptive analytics (Ereth, Kemper 2016, 469-460). Furthermore, the value of analytics gradually increases in parallel to its complexity when moving from descriptive to prescriptive analytics. Lanquillon, Mallow (2015a) also describes auxiliary types of analytics. *Exploratory*

analytics build a better understanding of the data in order to support the formulation of a problem, for example. *Visual analytics* or *visualization* offer a straightforward and interactive access to the data or analytics results (Lanquillon, Mallow 2015a, p. 58).

2.3.3 Analytics models

In general terms, "[...] a model is a simplified representation of reality created to serve a purpose" (Provost, Fawcett 2013b, p. 44). Analytics techniques are therefore also referred to as *analytics models*. For example, "[...] a predictive model is a formula for estimating the unknown value of interest: the target. The formula could be mathematical, or it could be a logical statement such as a rule" (Provost, Fawcett 2013b, 45). According to Sheikh (2013), there exists a difference between models and algorithms, because the latter implement the technique of a model based on the given data. This process is called *model learning* or *model building*. The resulting model is always specific to the given problem or data, respectively, while an algorithm is "[...] a general-purpose piece of software that doesn't change if the data set is changed [...]" (Sheikh 2013, p. 8). In this work, general analytics techniques as well as specific analytics models are simply referred to as *models*.

A vast number of different models exists (Finlay 2014, p. 104), however, most of them follow one of two basic approaches of model learning. On the one hand, *unsupervised learning* describes models without "[...] specific purpose or target [...]" (Provost, Fawcett 2013b, p. 24). This approach generally searches for interesting patterns in the data that are not known a priori (Murphy 2012, p. 2). On the other hand, models aim to learn a relation between some input data and a specific target in *supervised learning* (Murphy 2012, p. 2), and therefore this approach necessarily requires "[...] data on the target" (Provost, Fawcett 2013b, p. 24). Building supervised models with historical data is often referred to as *backtesting* (Baesens 2014, p. 134). Other less prominent approaches include *semi-supervised learning* as combination of the presented approaches, *active learning* utilizing human input in the learning process (Han et al. 2012, p. 25), and *reinforcement learning* including a notion of reward for learning (Murphy 2012, p. 2).

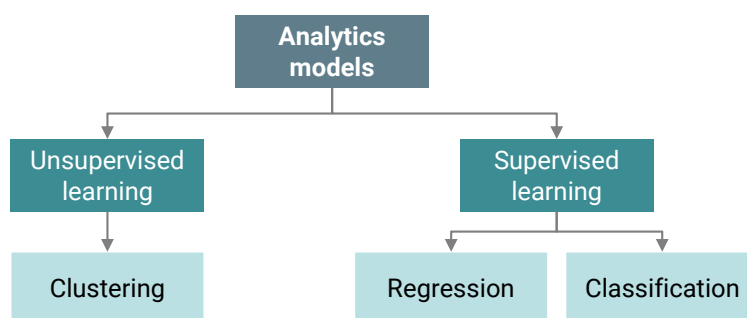


Figure 6 - Basic learning approaches and predominant model types

According to Murphy (2012, pp. 3–12), predominant model types for each basic learning approach exist: clustering for unsupervised learning and classification as well as regression for supervised learning. *Clustering* describes models that group "[...] data points into clusters based on their 'likeness' with one another" (Sheikh 2013, p. 11). Murphy (2012) further describes the use of unsupervised learning as a preprocessing step for other analytics. For instance, *factor analysis* aims to identify a reduced number of factors that describe the majority of variability in a large dataset. Furthermore, *graph structures* reveal the strength of correlation between variables and *matrix completion* is a technique to handle missing values in data (Murphy 2012, pp. 11–16). *Regression* provides quantitative targets as model output (Hastie et al. 2017, p. 10) while the output in the case of *classification* is represented in the form of distinguishable classes or concepts (Han et al. 2012, p. 18). Analytics models of the classification type are also referred to as *classifiers*. Figure 6 provides an overview of basic learning approaches and predominant model types. More details on specific models can be found in Mitchell (1997), Giudici (2003), Bishop (2006), Murphy (2012), Mohri et al. (2012), or Hastie et al. (2017). This work refers to models, that are difficult for lay people to understand due to their complexity and lack of transparency regarding relations between model input and output (Finlay 2014, p. 126; Biesdorf et al. 2013, p. 9), as *black box models*.

2.4 Analytics for sales forecasting

2.4.1 Traditional approaches

Sales forecasting can be defined as "[...] attempt to estimate the level of future sales through the use of previous and current information available about the phenomenon under study (sales)" (Belmokaddem et al. 2014, p. 21). A wide range of forecasting approaches exists in the business domain and Makridakis et al. (1980) provides a structured overview. Besides *informal approaches* such as ad-hoc or intuitive methods, *formal approaches* are grouped into qualitative and quantitative techniques. Technology-based forecasts represent one subgroup of *qualitative techniques*. They rest on studies of customers or competitors, exploration of current information, or future needs. In addition, subjective methods, as alternative subgroup of qualitative techniques, include assessments by the management team, aggregation of sales force estimates or techniques based on subjective probabilities for certain events. *Quantitative techniques* are divided into time series analysis and causal methods. The latter comprises simple and multiple regression as well as econometric models representing systems of multiple regressions. Time series analysis represents the largest subgroup of techniques. The naïve forecast follows simple rules such as 'forecast equals most recent sales volume'. Trend extrapolations utilize different continuous forms of past sales behavior and exponential smoothing rests on weighted averages of sales data. Furthermore, decomposition approaches consider time series features such as trend, seasonality, and random influences. Filtering builds forecasts by linear combination of past sales data and autoregressive techniques additionally incorporate occurred variances (Makridakis et al. 1980, pp. 42–52). Further overviews on forecasting approaches can be found in Armstrong (2002b), Brockwell, Davis (2002), Abraham, Ledolter (2005), Mertens, Rässler (2012), Kühnapfel (2013), or Gansser, Krol (2015).

Auto-Regressive Integrated Moving Average (ARIMA) models are an advanced form of time series analysis that are successfully used for forecasting problems (Chase 2013a, p. 85). They are among the two most commonly applied techniques for forecasting time series, such as company sales, besides exponential smoothing (Hyndman, Athanasopoulos 2017). The target of an ARIMA model is to forecast a "[...] time series that is modeled as a linear combination of its own past values and past values of an error series [...]" (SAS Institute 2017). According to Brockwell, Davis (2002), the $AR(p)$ part represents an autoregressive model based on past values of the time series where p denotes the order of the autoregressive process. The order value describes how many past time periods are considered in the model. The $MA(q)$ part is a weighted moving-average of past errors with q denoting the process order (Brockwell, Davis 2002, pp. 83–84). Abraham, Ledolter (2005) describe the need for the $I(d)$ part of the model that accounts for nonstationary time series where characteristics such as mean and variance are time-dependent. This behavior is commonly observed in economic and business time series. The integrated part transforms a nonstationary into a stationary time series by differencing (Abraham, Ledolter 2005, pp. 225–231) which "[...] compute[s] the differences between

consecutive observations" (Hyndman, Athanasopoulos 2017). The parameter d represents the order of differencing required to reach stationarity (Hyndman, Athanasopoulos 2017). Seasonality is another key characteristic of time series where values of the same season show high correlations (Abraham, Ledolter 2005, p. 281). Seasonal ARIMA models for a specific seasonality m account for this characteristic by an adapted form of differencing (Brockwell, Davis 2002, p. 203). These models are described by a set of seven parameters where uppercase letters describe the seasonality-related parts (Hyndman, Athanasopoulos 2017), as shown in Figure 7.

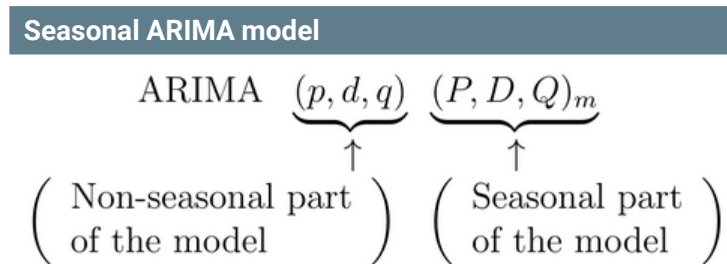


Figure 7 - Notation for seasonal ARIMA models (Hyndman, Athanasopoulos 2017)

2.4.2 Big data analytics models

Rey et al. (2012) discuss the conceptual difference between the traditional quantitative techniques and advanced analytics approaches to forecasting. The traditional approach develops a modeling of relations between the target and explanatory variables, especially based on statistical methods. The effect of individual variables is therefore mostly transparent. It generally follows the principle of causality. By contrast, an analytics approach starts with a large set of potentially explanatory variables. Strong dependencies among these variables can exist such that cause-and-effect relations are difficult to identify. The overall objective therefore is to find the variables that "[...] do the best job of forecasting [...]" (Rey et al. 2012, p. 5). Correlations among variables play a dominant role particularly in the context of big data (Anderson 2008). The focus on correlations rather than causality is even seen as major advantage of big data (Mayer-Schönberger, Cukier 2013, pp. 50–72). Theobald, Föhl (2015) argue that analytics models help to identify relevant correlations and therefore also reveal relations that would have been otherwise unknown. However, a strict focus on correlations also poses a risk and analytics users consequently need to carefully evaluate results (Theobald, Föhl 2015, p. 121). Traditional quantitative approaches are often not sufficient anymore due to the increasing complexity of the business environment (Chase 2013a, p. 32). On the other hand, it becomes increasingly straightforward to build analytics models today such that they should be deployed "[...] to provide the highest-quality forecasts possible" (Rey et al. 2012, p. 6), especially in combination with big data (Chase 2014). Models that are built on big data input are referred to as *BDA models* in this work.

Sales forecasting aims to answer a question of the type “*What will happen?*” and therefore falls into the category of predictive analytics. Sales represents the target variable and companies know their historic sales data such that models can be build following the approach of supervised learning. Baesens (2014), Dean (2014), or Finlay (2014) provide an overview of common predictive models, for example. The following introduction focuses on the most relevant models in this work including *k-nearest neighbors*, *support vector machines* and *decision trees* utilized as classification-type models as well as *elastic net regression* representing a regression model.

According to Hastie et al. (2017), *k-Nearest Neighbors (kNN)* classifiers are memory-based and therefore do not require to train a set of equations as model representation, for example. For any new observation, the classifier identifies k closest historic observations based on a distance measure. The majority class of these k neighbors defines the class predicted for the new observation. It represents a simple approach that has been proven to be successful in a large variety of classification problems (Hastie et al. 2017, pp. 463–465). The number of neighbors is the key parameter of the kNN classifier, and the performance does improve with increasing k only to a certain level as more distant neighbors are more likely members of a different class (Kubat 2015, p. 53). In the extreme case of k equals the number of existing observations, the classifier always predicts the majority class of the full dataset (Murphy 2012, p. 22).

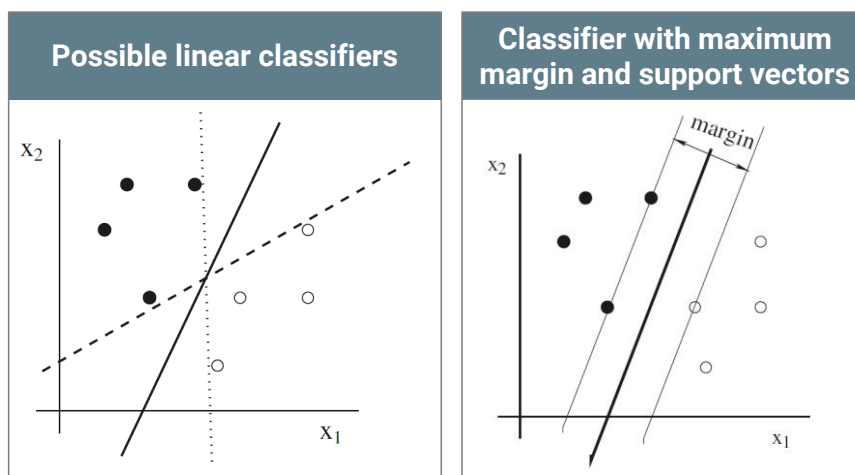


Figure 8 - Basic principle of support vector machines [based on Kubat (2015, pp. 84–85)]

Kubat (2015) describes the basic principle of *Support Vector Machines (SVM)* when used as a classifier. The left-hand side of Figure 8 illustrates a common issue of linear classification: the identification of the best classifier for predicting new observations among all possible classifiers for given observations. All three linear classifiers perfectly separate the two classes but it is unknown which classifier correctly predicts a new dot. In the given example, the dashed line classifiers are very close to the given observations of each class, that is to say their margin is small. However, “[...] the greater the margin, the higher the chances that the classifier will do well on future data” (Kubat 2015, p. 85). The *margin* can be defined by support vectors which are represented by the thin lines in the right-hand side of Figure 8. SVM determine those support vectors providing the maximum margin (Kubat 2015, pp. 84–85). The two classes of the

presented example are linearly separable. SVM can also handle cases that are not linearly separable⁸ by accounting for outliers, which lie on the wrong side of the separating line, in the definition of the margin (Mohri et al. 2012, pp. 71–72). Furthermore, SVM follow the same principle with large numbers of variables where the classes are separated by a multidimensional *hyperplane* (Hastie et al. 2017, pp. 417–419). This also holds true for *multi-class* problems where more than two classes need to be separated (Bishop 2006, pp. 338–339). Mohri et al. (2012) describe *kernel methods* as extension of models such as SVM, especially for non-linear classification. In that case the class-separating hyperplane is non-linear instead of linear. Kernel methods basically map the given input space of variables into "[...] a higher-dimensional space [...], where linear separation is possible" (Mohri et al. 2012, pp. 89–90). The *radial basis function (rbf) kernel* is a common method to handle non-linear relations in SVM classification (Hsu et al. 2016, p. 4). A *linear kernel* does not provide non-linear mapping but is generally useful for problems with large numbers of variables (Hsu et al. 2016, p. 12). scikit-learn (2017c) describes the major model parameters in case of a SVM with rbf kernel:

*"The parameter **C**, common to all SVM kernels, trades off misclassification of training examples against simplicity of the decision surface. A low **C** makes the decision surface smooth, while a high **C** aims at classifying all training examples correctly. **gamma** defines how much influence a single training example has. The larger **gamma** is, the closer other examples must be to be affected."*

Decision trees follow a simple concept but have proven as powerful analytics models (Hastie et al. 2017, p. 305). Mitchell (1997, pp. 52–53) illustratively describes the decision tree representation of an analytics problem:

*"Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some **attribute** of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute in the given example. This process is then repeated for the subtree rooted at the new node."*

This concept is best understood by the simple example in Figure 9 provided by Mitchell (1997). The example represents the two-class classification problem whether to play tennis or not. The root node tests the attribute of the weather outlook and branches to the humidity level as next decision node in case the attribute value equals 'Sunny', for example. From here, the leaf nodes of the defined classes, 'No' and 'Yes', are determined or predicted based on the attribute values 'High' and 'Normal', respectively (Mitchell 1997, p. 53). In contrast to parametric models that utilize the entire input dataset for learning, decision trees as nonparametric models do not assume class representations in form of mathematical functions and they do not rely on a

⁸ This case is also described as *overlapping classes* (Hastie et al. 2017, p. 417).

predefined tree structure (Alpaydin 2010, pp. 185–186). Furthermore, "[...] the tree grows, branches and leaves are added, during learning depending on the complexity of the problem inherent in the data" (Alpaydin 2010, p. 186). Decision tree models require a large set of model parameters including the maximum level of nodes or the maximum number of attributes considered at an individual node, for example (scikit-learn 2017a). Instead of the graphical representation shown for the example, a decision tree model can also be represented by "[...] a set of IF-THEN rules that are easily understandable" (Alpaydin 2010, p. 187). Moreover, decision trees are also suitable for predictive analytics with a continuous target variable and large data input (Ahlemeyer-Stubbe, Coleman 2014, pp. 129–133) such as a sales forecast based on big data.

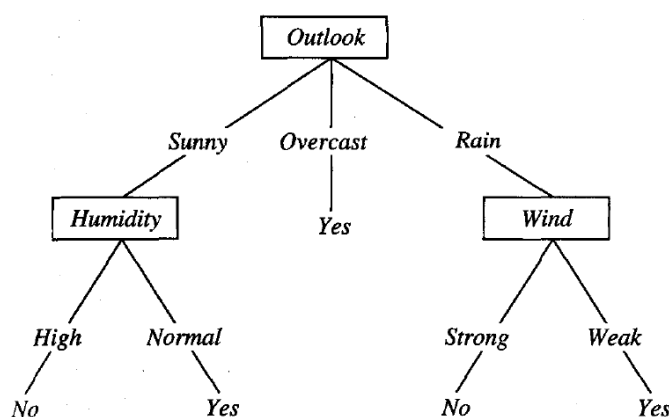


Figure 9 - Decision tree example (Mitchell 1997, p. 53)

Multiple linear regression is an alternative approach to predict continuous target variables and follows an easy to understand concept that generally provides high prediction accuracy (Ahlemeyer-Stubbe, Coleman 2014, p. 109). 'Multiple' refers to the fact that various explanatory variables are linearly combined with regression coefficients as weights in order to predict the target variable. Learning a multiple linear regression model requires to determine these coefficients, for example, using an *Ordinary Least Squares (OLS)* solution (Murphy 2012, pp. 219–220). However, with a large number of explanatory variables the solution is not unique and typically all coefficients will be nonzero complicating the interpretation of the resulting model (Hastie et al. 2015, p. 2). Furthermore, a reduction of coefficients or even removal of explanatory variables by setting selected coefficients to zero can increase prediction accuracy (Hastie et al. 2015, p. 7). Different *regularization* methods address these issues of regression coefficient determination. *Lasso* is a common regularization method where the total sum of absolute values of coefficients is limited to a maximum value (Hastie et al. 2015, p. 8). "The key property of Lasso [...] is that it leads to a sparse solution [for coefficients], that is one with few non-zero components" (Mohri et al. 2012, p. 257). Murphy (2012) describes *ridge* as an alternative regularization method. Ridge is motivated by the fact that solutions for regression coefficients aim to "[...] perfectly interpolate[s] the data" (Murphy 2012, p. 225) and therefore produce unstable models that are highly dependent on the given data. The method introduces

a complexity penalty for "[...] the sum of the magnitudes of the [...]" (Murphy 2012, p. 226) coefficients and therefore promotes coefficients to be small (Murphy 2012, pp. 225–227). Zou, Hastie (2005) introduce a new regularization method called *elastic net* that combines lasso and ridge in order to address their shortcomings and to merge their advantages. The elastic net "[...] selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge" (Hastie et al. 2017, p. 73). In this work, a multiple linear regression model with elastic net regularization is simply referred to as *elastic net*. The mix of lasso and ridge regularization is the main model parameter but the elastic net model also includes further parameters such as weights for variables, for example (Hastie, Qian 2014).

2.5 Big data analytics foundations

The application of big data analytics requires fundamental prerequisites regarding technology, organization, and culture in order to gain a competitive advantage (Biesdorf et al. 2013, p. 1; Court 2015). The following provides a brief overview on these foundations with focus on relevant dimensions for this work.

Technological foundations are required "[...] to aggregate, manipulate, manage, and analyze big data" (Manyika et al. 2011, p. 31). The main drivers for technology are the 4V characteristics of big data and intended analytics (Lanquillon, Mallow 2015b, p. 263). Loshin (2013a) describes the major dimensions of big data technology. Storage systems need to be scalable in order to handle large volume datasets. Data management potentially requires new concepts such as non-relational schemes capable to handle unstructured data, for example. Furthermore, computing technology must enable parallel processing and fast access to the data storage. Finally, a development framework comprises a programming environment and access to analytics models providing support for the analytics process and model building (Loshin 2013a, pp. 49–50). Data security and privacy are two further dimensions that also need to be addressed by the technological foundations (Fogelman-Soulié, Lu 2016, p. 157).

Hadoop is a key technological enabler for big data because the "[...] open-source platform for storage and processing of diverse data types [...] enables data-driven enterprises to rapidly derive the complete value from all their data" (Minelli et al. 2013, p. 61). Minelli et al. (2013) describe the *Hadoop Distributed File System (HDFS)* and *MapReduce* as the two major components of Hadoop. HDFS is the current standard for distributed data storage without constraints on data types or volume. The basic idea is to distribute fractions of data across multiple servers whereby each fraction is replicated on more than one server. Hadoop typically "[...] runs on clusters of commodity servers and each of those servers has local CPUs and disk storage that can be leveraged by the system" (Minelli et al. 2013, p. 62).⁹ The cluster architecture represents a cost-effective way to store and process data. Processing includes calculations and manipulation of the data and is managed by MapReduce as standard for a "[...] fault-tolerant parallel programming framework that was designed to harness distributed processing capabilities" (Minelli et al. 2013, p. 90). MapReduce distributes processing tasks across the servers in the Hadoop cluster and collates the results (Minelli et al. 2013, 61–91). *Spark* provides a development framework for Hadoop cluster that enables "[...] a fast in-memory data processing system that achieves high performance for applications [...]" (Tang et al. 2016, p. 165). Furthermore, Spark simplifies programming in the Hadoop environment and supports most of the programming languages for analytics (Wierse, Riedel 2017, pp. 316–318) including *Python* as one of the most common languages in this domain (Wierse, Riedel 2017, p. 345).

⁹ CPU stands for *central processing unit*.

In companies, *data warehouses* store various datasets for different purposes or topics and *data marts* host smaller datasets for a specific application or department (Davenport 2014, p. 114). It is important to note that these traditional database types are also capable of housing big data (Minelli et al. 2013, p. 89) such that Hadoop clusters should be seen as a complement and not a replacement (Lanquillon, Mallow 2015b, p. 276). Moreover, traditional data management rests on relational database management systems where structured data is organized in tables with rows and columns (Ameri 2016, pp. 139–141). The big data characteristics sometimes demand non-relational data management. New approaches are typically summarized as NoSQL¹⁰ solutions that provide new data models replacing the fixed tabular scheme and other advantageous features such as simplified interfaces (Ameri 2016, pp. 143–144). However, a "[...] proper database management type should be chosen dependent on the application requirements" (Ameri 2016, p. 144). The relational model is still a valid approach, especially for data stemming from traditional systems, and is also supported by most big data technologies (Loshin 2013a, p. 83).

Organizational foundations have two major dimensions: *talent* and *structure*. BDA utilization requires professionals with capabilities for "[...] different analytics disciplines, different types of data, and different tools [...]" (Franks 2014, pp. 209–216). Espinosa, Armour (2016, p. 1114) therefore describe big data analytics as "[...] a multi-discipline team-based activity that brings together various perspectives [...]". Companies can build up these required capabilities internally by hiring relevant talent or by providing training to existing staff (Manyika et al. 2011, p. 114). As an alternative, a company can also outsource BDA work or complement internal teams with external experts in case of lacking capabilities (Wierse, Riedel 2017, pp. 236–240). The *data scientist* represents a prominent form of talent that has major capabilities in analytics and technological foundations (Ohlhorst 2013, pp. 29–30) but is also capable to connect data and analytics with the application domain (NIST Big Data Public Working Group 2015, pp. 8–9). Internal capabilities require an organizational structure (Ohlhorst 2013, p. 34). According to Franks (2014), no standard structure exists and companies use many different forms. One possible structure is a center of excellence as centralized unit of internal BDA talent. Such a center is typically organized by different business areas or functions (Franks 2014, 218–220). Furthermore, the organizational structure should also include some form of governance to address security and privacy issues, for example, in the form of security clearances or privacy standards (Franks 2014, pp. 147–174).

The *cultural foundation* is vividly described by the *analytics culture* in Franks (2014). It addresses various issues of making analytics work in a company. The main part is a new mindset that acknowledges the value of BDA professionals, for example, and this mindset must be supported by top management. Furthermore, company policies need to support the new culture. An

¹⁰ NoSQL stands for "Not only SQL" (*Structured Query Language*) (Loshin 2013a, p. 83).

analytics culture facilitates success by promoting searches for unexpected value in data, for instance. It also makes room for experimentation and promotes acceptance of related failure (Franks 2014, pp. 237–262). Such a culture also urges employees "[...] to base decisions on hard facts" (Davenport 2006). The black box character of some BDA applications poses a challenge to this fact-based decision making. Examples show that solutions based on black box models can be rejected in practice (Biesdorf et al. 2013, p. 9).

The presented foundations address multiple challenges related to big data analytics as described by Sivarajah et al. (2017). They address *data challenges* related to big data characteristics and *management challenges* such as security and privacy. However, *process challenges* describe "[...] the group of challenges encountered while processing and analyzing the data that is from capturing the data to interpreting and presenting the end results" (Sivarajah et al. 2017, pp. 269–275). Although these process challenges also benefit from the presented foundations, there is a general need for a process itself (Chamoni, Gluchowski 2017, p. 12). While BDA applications for sales forecasting are the focus of *Section 3.1*, processes to develop BDA applications are discussed in *Section 3.2*.

3 Related work

3.1 Related research on sales forecasting

3.1.1 Overview

The research motivation reveals that sales forecasting is a key application for big data analytics. However, big data analytics is still a relatively new research area and many companies just start to make plans on future applications for sales forecasting (compare *Section 1.1*). This section examines the current state of research on sales forecasting based on big data analytics in order to gain an understanding of existing applications. The corresponding review rests on research that utilizes big data for sales forecasting in companies. In addition, a second review considers the industry of the case study company. The focus here is less restrictive because general research on sales forecasting in the printed circuit board industry is considered. This provides an understanding of the latest advancements of research regarding this industry. Finally, the interim conclusion summarizes findings as well as assesses the applicable knowledge and scientific need regarding the specific business need for BDA applications for sales forecasting. Figure 14 provides an overview of reviewed research and the structure of this section.

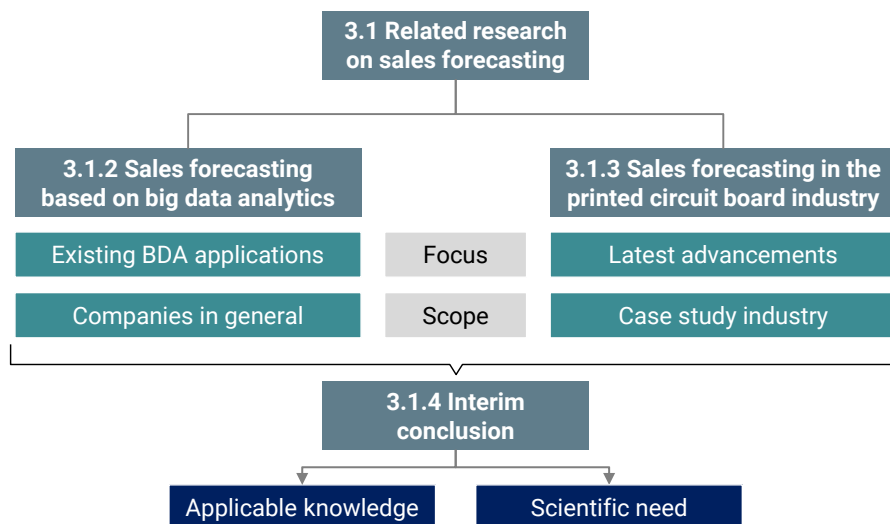


Figure 10 - Overview of reviewed research

3.1.2 Sales forecasting based on big data analytics

The research motivation presented in *Section 1.1* demonstrates that sales forecasting is crucial for industrial companies and that traditional forecasting methods are increasingly inadequate in today's business environment. In addition, big data analytics is considered to be a source of competitive advantage due to improved forecasting capabilities. Practitioners deem the development of predictive models based on big data as opportunity to reduce uncertainty about future developments. As a consequence, this subsection investigates the current state of research on big data-based approaches for sales forecasting. To put this discussion into perspective, Figure 11 provides an overview of the development of general big data-related research since

2000. The overview is based on similar search queries for 'big data' publications in any language and excluding patents. The queries were performed on the leading database for peer-reviewed literature *Scopus*¹¹ and the search engine for scientific literature *Google Scholar*¹². The results show that big data-related research is still a relatively young discipline. This becomes even clearer if one considers the share of most recent publications. Publications from the previous five years represent 99% and 94%, respectively, of all publications on this topic since 2000. Nevertheless, a sizeable body of big data-related literature exists due to the strong expansion in recent years.

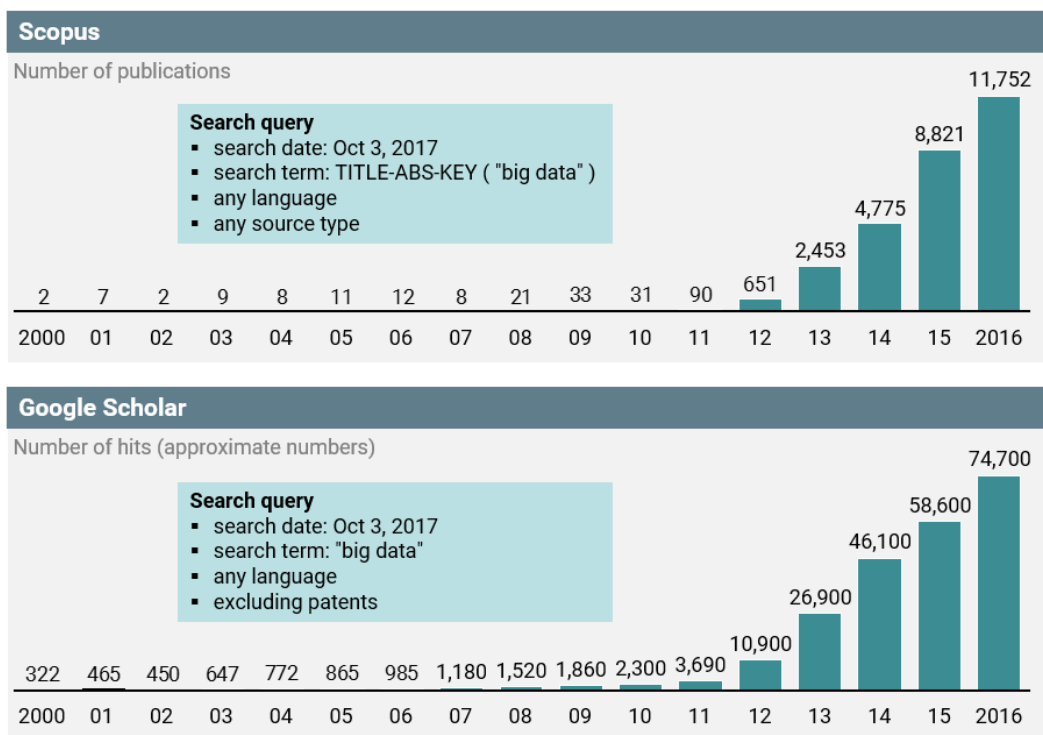


Figure 11 - Big data-related publications and hits since 2000

A structured literature review assesses the role of sales forecasting within the current state of research. Systematic review of existing research is a valid approach to describe the background of newly developed research and to "[...] identify any gaps in current research in order to suggest areas for further investigation" (Kitchenham, pp. 1–2). The literature review is based on a search query performed on the Scopus database¹³ and aims to capture big data-related research on sales forecasting. The term *analytics* is omitted for less restrictive search results and because *big data* is the key BDA characteristic. In order to account for synonyms and alternative spellings (Kitchenham, p. 8), the query includes *demand* and *revenues* as alternatives to *sales* as well as

¹¹ <https://www.scopus.com/> [last access date: 10/25/2017]

¹² <https://scholar.google.com/> [last access date: 10/25/2017]

¹³ The search query includes publications of any language as well as any source type and was performed on June 27, 2017.

prediction and *monitoring* as alternatives to *forecasting*. The query considers any sensible combination of these search terms. Furthermore, each combination allows for appearance of the search terms within five words and uses the word stem of each term in order to be less restrictive. The search query considers title, abstract, and keywords of publications over a period of 20 years and is presented in Figure 12.

```
Search query
TITLE-ABS-KEY (
( sale* W/5 forecast* ) OR ( sale* W/5 predict* ) OR ( sale* W/5 monitor* )
OR ( demand* W/5 forecast* ) OR ( demand* W/5 predict* ) OR ( demand* W/5 monitor* )
OR ( revenue* W/5 forecast* ) OR ( revenue* W/5 predict* ) OR ( revenue* W/5 monitor* )
AND ( "big data" ) )
AND PUBYEAR > 1996
```

Figure 12 - Search query for Scopus database

The search query results in 161 publications of which 13 represent complete proceedings that are excluded from the review. In order to identify relevant literature, the remaining publications are reviewed based on their abstract and full text if necessary. In doing so, relevant literature is defined as scientific work that discusses an application for sales forecasting of a company considering the aforementioned synonyms. This review process identifies three different groups of irrelevant literature:

- 1) *Energy demand related research*: A significant number of publications presents research related to energy demand. For example, Huang, Zhu (2016) propose a model for energy demand forecasting based on smart meter data, Coelho et al. (2016) introduce a deep learning forecasting model for household electricity demand, and Zhang, Grijalva (2015) apply big data analytics to smart meter data for electric vehicle charging demand. This group comprises 32 publications and thus demonstrates extensive research on big data in relation to energy demand.
- 2) *General demand forecasts*: This group of publications deals with demand on the level of countries or cities rather than companies. For example, Li et al. (2017) and Yahya et al. (2017) discuss forecast models for tourist volumes in Beijing, Singapore, and Indonesia, respectively. A neural network model to forecast bus transportation demand in the Seoul metropolitan area (Baek And, Sohn 2016) and the use of search engine data to predict personal credit demand in Turkey (Zeybek, Ugurlu 2015) are further examples.
- 3) *Others*: The remainder of identified literature represents scientific work that does not fit the focus on company-related sales forecasting for various reasons. The range of research goes from commodity price forecasting for mining companies (Ming et al. 2016) to predictions on cloud computing demand for web traffic during sports events (Baughman et al. 2016) to improved forecasting of optimal nursing staff in patient care (McNair 2015). It furthermore includes forecasting of elections (Huberty 2015) or local weather for better planning of renewable energy generation (Corne et al. 2014). Other

examples of this group include systematic approaches to include BDA into supply chain processes (Fukui 2016; Banica, Hagiú 2016), the introduction of a new forecasting model for big data in form of time series (Singh 2015), and an approach to predict demand for files on a Hadoop cluster (Kousiouris et al. 2013).

The review process, as shown in Figure 13, reveals 37 scientific works that describe BDA applications for sales forecasting of companies, of which 29 describe cases related to consumer goods and services. E-commerce and retail are the most common domains representing half of the consumer cases. For example, Chong et al. (2016) build a neural network model in order to identify the predictive power of consumer reviews and promotional activities such as discounts for product sales of a large online retailer. Watanabe et al. (2016) build linear regression and neural network models to forecast product demand of supermarkets based on daily weather and sales data including prices. Other consumer-related applications include global sales forecasting of an apparel and sports equipment company (Boldt et al. 2016), forecasting of food demand such as pizza sales (Lee, Kim 2015), predicting ratings of TV series with decision trees and linear regression based on Twitter data (Molteni, Ponce De Leon 2016). Further works include the utilization of data from online search traffic to forecast sales volumes of hybrid vehicles as case for new products (Jun et al. 2014), a study on predictability of taxi demand using a large set of spatio-temporal data (Zhao et al. 2016), and applications in the business domain of pharmacy products related to weather data (Lin, Tsai 2016) as well as short-term air passenger volumes at an airport based on data from online search engine queries (Kim, Shin 2016).

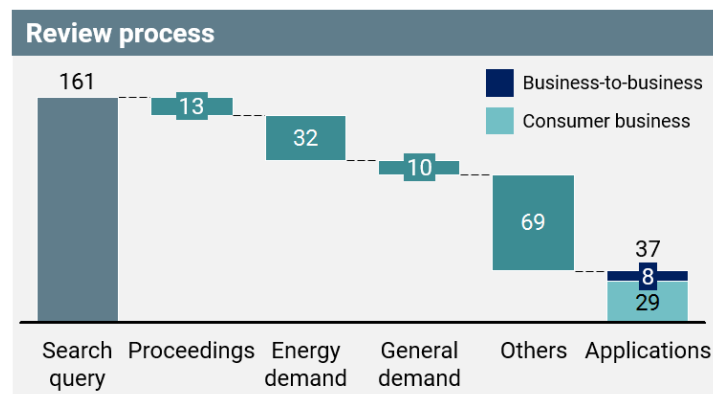


Figure 13 - Results of review process

Only eight publications¹⁴ represent BDA applications for *Business-to-Business* (B2B) companies, however, they are limited in terms of utilized data and applied analytics, respectively. Palanimalai, Paramasivam (2016) describe a simple heuristic that integrates historic sales data and the customer relationship management system of a pharmaceutical company in order to forecast sales of different product categories sold to various accounts, especially hospitals. The

¹⁴ Ji et al. (2015) is not discussed in detail here as the full text is published in Chinese only.

goal of their approach is to provide a dashboard in order to track sales trends and reveal sales opportunities, for instance (Palanimalai, Paramasivam 2016, pp. 212–218). Nita (2015) discusses a demand forecasting application for food manufacturers where they deploy unspecified machine learning. Data input is solely described by examples such as shipment data, sales data from retailers, weather data, or information on advertisement activities. Automated forecasts after creating the models based on historic data and regular reviews of forecast performance are two additional features of the approach (Nita 2015, pp. 92–93). Otsuka et al. (2015) present a similar application for demand forecasting of IT-related spare parts. Forecasting models, again unspecified, are built for different categories of parts based on data for shipments, products in use and usage time. Models can also be dependent on specific events, such as different weather, but no further details are provided here. Inventory reduction at the spare part provider is identified as the major advantage of improved forecasting (Otsuka et al. 2015, pp. 80–82). Ji et al. (2017a) and Ji et al. (2017b) describe the use of big data from the food supply chain for improved demand forecasting in case of food companies. The data input is only defined by generic categories of data sources for data on food consumption, for example, retailers or third-party brokers, and for data on food logistics such as transportation operators. A Bayesian network is proposed as forecast model because it represents causal relationships between variables. It therefore requires experts to identify influencing factors on demand before building the model (Ji et al. 2017a, pp. 2–8). The main goal of the application is to optimize the production setup with regard to identified demand (Ji et al. 2017b, p. 9). Qiu et al. (2016) propose a cloud manufacturing architecture for polymer material producers where cloud computing and big data technologies are utilized to integrate information along the entire supply chain. This architecture supports the key task of balancing production and market demands as it expands the scope of available data (Qiu et al. 2016, pp. 239–243). The work has a strong focus on technological foundations and therefore no details on specific data or models for demand forecasting are presented. Williams (2013) introduces a tool that combines data from enterprise resource planning systems with other sources of supply chain data in the electronics manufacturing industry. The purpose of the tool is to apply machine learning in order to increase accuracy of demand forecasting but also to optimize supply chains based on prescriptive analytics (Williams 2013, pp. 5–6). However, the brief conceptual introduction does not include any further details on big data or analytics.

Although the presented literature review makes no claim to be exhaustive, several conclusions can be drawn from it. The earliest publication in the results list dates back to 2012. Identified applications for sales forecasting therefore represent 0.1% of all big data-related publications in the Scopus database during the period from 2012-2016. This indicates the limited scope of research in that area as of today. Taking a closer look at the applications reveals that current research focuses on consumer products and services. Furthermore, the few B2B cases generally provide a low level of detail on big data and analytics utilized.

3.1.3 Sales forecasting in the printed circuit board industry

The company of the case study in this work is part of the *Printed Circuit Board (PCB)* industry which also represents a B2B industry. In order to gain a better understanding of the current research state on sales forecasting approaches for this industry, a review of relevant literature is presented here. Due to the limited research on big data-based sales forecasting, this review presents the latest research regardless of a big data characteristic. Chang et al. (2005b) initiated a series of research¹⁵ on sales forecasting in the PCB industry motivated by the emergence of new analytics models. These represent promising alternatives to traditional forecasting approaches in the industry such as statistic methods in the form of trend analysis or time series analysis (Chang et al. 2005b, pp. 83–84). The study is based on monthly PCB sales data from a Taiwanese electronics company over a period of five years and utilizes predefined data from three different sources as input (Chang et al. 2005b, pp. 84–85):

- 1) *Macroeconomic variables* including gross national product, unemployment rate and indices for consumer prices as well as import and export trade
- 2) *Market demand variables* for major PCB application areas including computers, notebooks, motherboards, monitors, televisions, and mobile phones
- 3) *Industry variables* described by manufacturing production, manufacturing sales, manufacturing production value, semiconductor production and PCB production value indices

Chang et al. (2005b) select one variable from each group as input for their analytics model by identifying the maximum influence on PCB sales based on a technique called grey relation analysis. Historic PCB sales data is also used as input, however, exponential smoothing is applied in order to capture seasonality and trend effects. An evolving neural network is selected as model which generates the weights between network nodes with a genetic algorithm. The reported results for monthly forecasts show that the new approach improves forecasting accuracy compared to a linear regression model utilizing the same variables as input (Chang et al. 2005b, pp. 85–91).

Chang, Wang (2006) discuss an alternative sales forecasting approach based on the same data for PCB sales and input variables as in the previously discussed work. A hybrid model is proposed that includes expert opinions into a neural network. For this purpose, experts from sales and production departments are surveyed in order to define different weights for input variables. The results show a similar forecasting performance compared to the evolving neural

¹⁵ The most recent and most cited (according to Scopus) publications are discussed here. Further research includes the following publications: Hicham et al. (2012a), Liu, Wang (2012), Hicham et al. (2012b), Wang et al. (2009), Chang et al. (2007a), Chang et al. (2006b), Chang et al. (2005a), and Chang, Lai (2005).

network (Chang, Wang 2006, pp. 717–725). Chang et al. (2006a) build upon the most influential variables and smoothed sales data as described by Chang et al. (2005b) and propose an alternative hybrid model that integrates neural networks, fuzzy rules, and genetic algorithms. The neural network serves as clustering for the input and sales data in order to generate homogeneous subsets. For each cluster, an individual set of fuzzy rules, that is more representative for this cluster as for the entire dataset, generates the forecast. Fuzzy rules represent a forecasting approach designed for nonlinear and ambiguous data whereby the hybrid model utilizes a genetic algorithm for optimizing these rules. This approach results in an improved forecasting performance in comparison to earlier models (Chang et al. 2006a, pp. 1258–1263). Chang et al. (2007b) present a similar approach that directly integrates a neural network and fuzzy theory as forecasting model. Again the same data as well as variable selection approach is employed and similar forecasting performance is reported (Chang et al. 2007b, pp. 88–95). Chang et al. (2009) also use a fuzzy neural network for sales forecasting and they integrate the clustering approach as discussed by Chang et al. (2006a). However, they use a k-means algorithm instead of a neural network approach for clustering (Chang et al. 2009, pp. 345–354). Again building on the same selected data input, Chang et al. (2008) combine fuzzy theory and case-based reasoning as forecasting model. Case-based reasoning compares a new set of input variables to known cases of variables including PCB sales. A weighted average of the known cases defines the forecast for the sales variable of the new case, whereby a measure of similarity between new and known cases defines the weights of known cases. The approach has worse forecast accuracy compared to previously presented ones (Chang et al. 2008, pp. 2052–2055).

Other researchers continued the research on sales forecasting for PCB manufacturers initiated by the Taiwanese research group. Hadavandi et al. (2011) present the first work that keeps the same PCB sales data and variables as described before. They introduce a new approach that combines k-means clustering of input data and a set of genetic fuzzy systems per cluster to forecast sales. The genetic fuzzy system is created by defining and tuning fuzzy rules based on unique genetic algorithms instead of plain definition of rules with a simple genetic algorithm as in the case of Chang et al. (2006a). The authors report improved forecasting accuracy in comparison to the results from the Taiwanese research group (Hadavandi et al. 2011, pp. 9394–9399). Hichama et al. (2013) propose a further advancement for sales forecasting by implementing a novel way of data clustering. Fuzzy clustering allows that data elements are potentially assigned to more than one cluster, which increases "[...] the number of elements of each cluster and consequently improve[s] the accuracy of the proposed forecasting system" (Hichama et al. 2013, p. 949). Furthermore, a novel form of genetic fuzzy system is created based on clustered data and this system feeds into an adaptive neural network. That is to say, there exists a neural network for each identified cluster and these jointly determine the sales forecast value (Hichama et al. 2013, pp. 951–960). Finally, Tavakkoli et al. (2015) present the latest forecasting approach to the PCB sales case that is markedly different to previous work. A

support vector machine with linear or radial basis function kernel is proposed as analytics model. The optimal set of parameters describing the model is defined with the help of a bio inspired algorithm called bat algorithm (Tavakkoli et al. 2015, pp. 197–203). The normalized variables and PCB sales data directly serve as input to the model and the forecasting accuracy (Tavakkoli et al. 2015, pp. 205–208) is comparable to the early work of Chang et al. (2005b).

This review reveals that the focus of current research lies on the advancement of forecasting models. There is a trend towards more complicated approaches in the form of hybrid models that integrate multiple types of analytics. Furthermore, there is a tendency to use more advanced models such as neural networks. Only the most recent work of Tavakkoli et al. (2015) returns to a simpler form based on a rather standard model. The big data dimension of BDA is not addressed in this research stream which builds on 15 variables from three data sources. The data sources solely provide structured economic data that is therefore not characterized by variety or veracity. Furthermore, the number of variables in combination with five years of historic data is not an extraordinary volume of data, and forecasting monthly sales at a monthly frequency does not represent a high velocity as well. The research discussed is 'data-driven' in a different way as data provided by a PCB manufacturer is the only relation to the company. In other words, research is performed following an 'outside-in' approach instead of an in-depth case study.

3.1.4 Interim conclusion

The key results from both reviews of related research on sales forecasting are summarized as follows:

- Big data-based sales forecasting is generally at an early stage
- Majority of big data-related research describes applications for consumer goods and services instead of B2B cases
- B2B cases are characterized by a low level of detail regarding both big data and analytics
- Research on sales forecasting in the PCB industry focuses on analytics in form of hybrid models rather than big data input
- PCB cases are characterized by outside-in research where data is the only input from the company

In summary, reviewed literature addresses the specific business need to leverage big data analytics for improved sales forecasting only to a limited extent. It is therefore a *scientific need* to add further evidence whether BDA sales forecasting works in industrial practice, especially in B2B cases. This scientific need relates to *research question 3* and underlines the value of a proof of concept for a BDA application for sales forecasting in B2B industries. Moreover, the

RELATED WORK

restriction that the company involved in PCB-related research provided data as the only research input endorses the case study approach while addressing this need.

The reviews furthermore reveal a lack of detail for BDA applications in industrial practice and a strong focus on specific analytics in case of the PCB industry. As a consequence, reviewed research does not provide substantial applicable knowledge for the objective of this work. The following *Section 3.2* examines applicable knowledge for the intended methodology, but with a broader focus on processes to develop general BDA applications because sales forecasting represents only one possible application to provide a better understanding of the volatile business environment.

3.2 Processes for analytics applications

3.2.1 Overview

Starting in the late 1980s, research began to discover opportunities provided by *Knowledge Discovery and Data Mining (KDDM)* (Barclay, Osei-Bryson 2015, p. 2). KDDM initiated "[...] a rush to develop [data mining] algorithms that were capable of solving all problems of searching for knowledge in data" (Marban et al. 2007, p. 578). One of the first attempts to formalize the approach to KDDM is the work *The process of knowledge discovery in databases: a first sketch* by Brachman, Anand (1994) (Anand et al. 2007, p. 22). *The KDD process for extracting useful knowledge from volumes of data* by Fayyad et al. (1996c) is regarded as seminal work (Barclay, Osei-Bryson 2015, p. 2). Process models for KDDM applications "[...] serve the purpose of a roadmap or guide, that provide prescriptive guidance towards how each task in the end-to-end process can be implemented" (Sharma et al. 2012, pp. 11335–11336). Cios et al. (2007) outline reasons why a process model is required. A structured approach ensures that results are practical and understandable from a user perspective. Moreover, it enables to explain the underlying KDDM mechanics to decision-makers and thus makes them comfortable to take responsibility for decisions based on KDDM insights. KDDM typically involves a project team such that a process model serves as framework for necessary project management. Finally, standardized methods and procedures lead to "[...] project performance that is faster, cheaper, more reliable, and more manageable" (Cios et al. 2007, pp. 9–10).

Process model, methodology, and lifecycle represent different levels of structured approaches towards KDDM applications and are similar to projects in other engineering disciplines (Mariscal et al. 2010, pp. 140–141). A *process model* describes a structure of tasks (Pressman 2010, pp. 30–37) required to perform a KDDM project and aims to make a project manageable and repeatable (Mariscal et al. 2010, p. 140). On the other hand, a *methodology* "[...] can be defined as a process model instance, in which not only tasks, inputs and outputs must be specified but also the way in which the tasks must be carried out" (Mariscal et al. 2010, p. 140). Techniques or tools are therefore required in order to execute required tasks (Pressman 2010, p. 50). A *lifecycle* defines the sequence of project activities (Lester 2014, pp. 47–48) and determines required outcomes to move from one step to the next (Mariscal et al. 2010, p. 141). The approaches discussed in this section are generally referred to as *processes* as they include process models and methodologies. All presented processes generally include a definition of the lifecycle represented by successive *steps* that break down into multiple *tasks*. *Methods* generally describe the execution of a task whereby *tools* require technical implementation, for instance in form of software. *Techniques* represent general methods, for example, in the form of simple procedures such as workshop formats. In general, analytics are supported by two types of processes where one supports the development of analytics applications and the other the continued use in practice (Heath, Hull 2015, p. 175). The focus here lies on the former type, and includes processes that stem from academic research as well as industry (Singh et al. 2011, pp. 279–280).

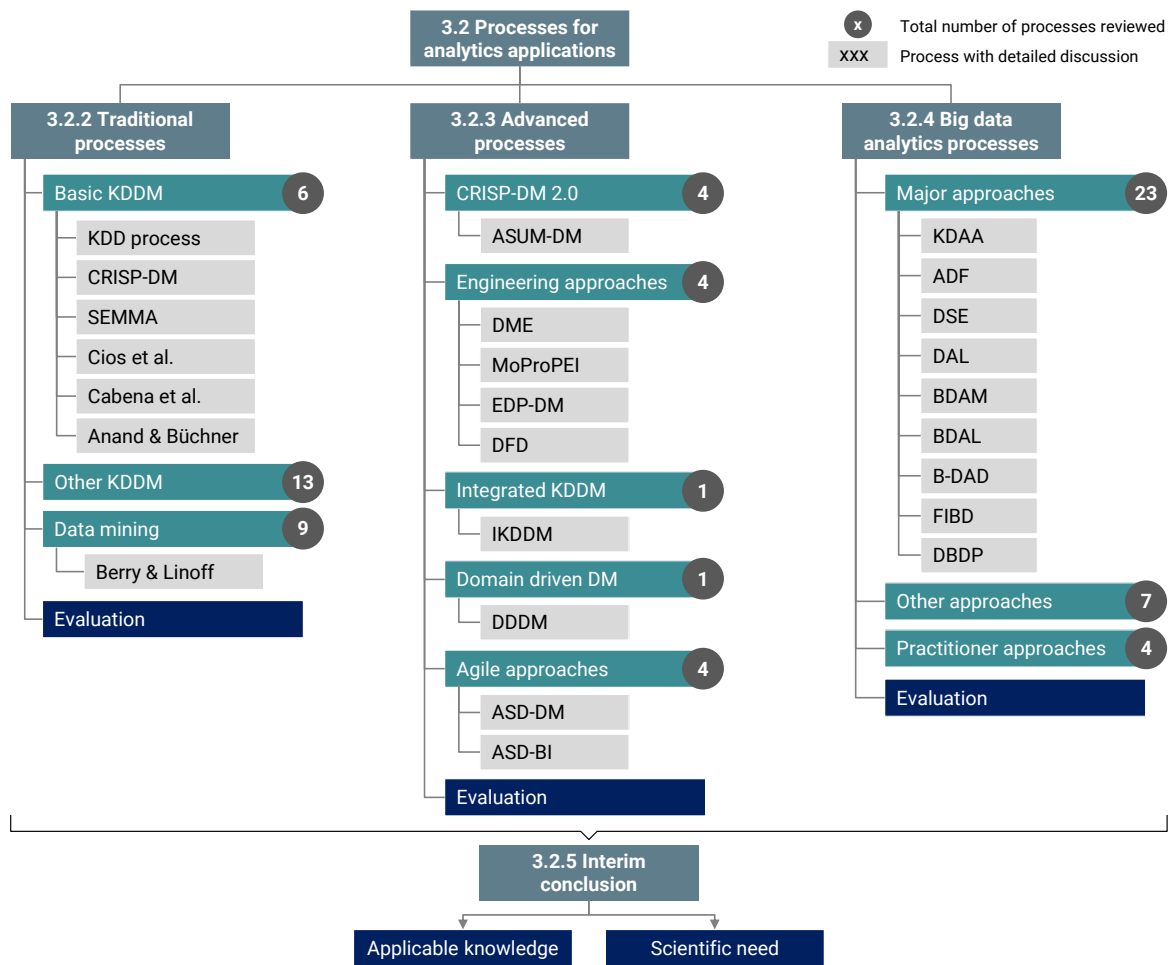


Figure 14 - Overview of processes

The following discussion on existing processes starts with approaches from traditional KDDM, and then moves on towards advanced KDDM processes as well as current approaches addressing the emergence of big data analytics. Figure 14 provides an overview of processes and the structure of this section. The review comprises 76 processes from which the 25 most relevant ones are discussed in detail. Each subsection includes an evaluation of the presented processes. These evaluations build the basis for the interim conclusion that summarizes findings as well as assesses applicable knowledge and scientific need regarding the new methodology.

3.2.2 Traditional processes

Existing literature provides various discussions of KDDM processes. Table 2 provides an overview of scientific works that discuss at least three different processes and collectively span a period of more than ten years into the past. This subsection starts with detailed descriptions of basic KDDM processes that consistently appear in academic discussions over time. Brief introductions to the other KDDM processes and approaches stemming from data mining literature follow.

Source	Basic KDDM processes						Other KDDM processes						
	KDD process	CRISP-DM	SEMMA	Cios et al.	Cabena et al.	Anand & B�chner	5 A's	KDD roadmap	Six sigma	Two Crows	DMIE	Human-centered	RAMSYS
Cios, Kurgan (2005)													
Kurgan, Musilek (2006)													
Marban et al. (2007)													
Azevedo, Santos (2008)													
Francois (2008)													
Sharma (2008)													
Marban et al. (2009a)													
Mariscal et al. (2010)													
Oprean (2011)													
Anand (2012)													
Sharma (2015)													
Rogalewicz, Sika (2016)													


 process discussed by source

Table 2 - Overview of KDDM processes

3.2.2.1 Basic KDDM processes

KDD process

Fayyad et al. (1996a) reason the development of a process for *Knowledge Discovery in Databases (KDD)* with the need for a structured overview of required activities and their interaction. This requirement was mainly driven by knowledge discovery changing substantially from a manual to a technology-enabled task with the emergence of the digital age. Increasing volumes of data require to assist or replace analysts with computers and software in order to gain valuable insights from this data. As this holds true across nearly all domains – ranging from science to healthcare to finance – a general approach to KDD was required (Fayyad et al. 1996a, pp. 37–38). Moreover, a general approach also reflects the increasingly multidisciplinary character of KDD (Fayyad et al. 1996b, p. 82).

Fayyad et al. (1996b, p. 83) define the KDD process as follows:

"KDD Process is the process of using the database along with any required selection, preprocessing, subsampling, and transformations of it; to apply data mining methods (algorithms) to enumerate patterns from it; and to evaluate the products of data [mining] to identify the subset of the enumerated patterns deemed 'knowledge'."

It should be noted that pattern is used as collective expression for extracted knowledge from data (Fayyad et al. 1996b, p. 83), including predictive models, for example. Figure 15 provides an overview of the KDD process and Fayyad et al. (1996b) explain its nine successive steps in detail. (1) Starting point is an understanding of the domain for KDDM application. There are two major tasks related to this including the collection of relevant knowledge and setting the objective of the project. (2-4) The following three steps are concerned with data input to the knowledge discovery. After selecting relevant data, it must be cleaned and preprocessed before it is prepared for analytics by data reduction or transformation. (5) The next step determines appropriate analytics methods (e.g., regression or classification) before (6) specific models (e.g., elastic net or SVM) are selected. Step (7) constitutes analytics termed as data mining and leads into (8) results preparation (e.g., visualization) and interpretation. (9) Finally, discovered knowledge is transferred to related areas of the domain. The process is of iterative nature,

whereby unsatisfactory results are a key trigger for reconsideration in previous steps (Fayyad et al. 1996b, p. 84).

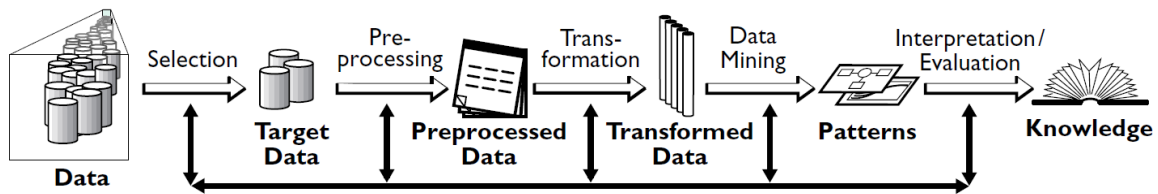


Figure 15: KDD process (Fayyad et al. 1996c, p. 29)

CRISP-DM

The lack of a standard process made KDDM projects in practice strongly dependent from the talent involved which obstructed a sustainable dissemination of analytics (Wirth, Hipp 2000, pp. 29–30). The missing standard also hindered the introduction of new users to analytics and providing a process applicable across various industries as well as organizations further motivated the introduction of a new process (Shearer 2000, p. 13, 2000, p. 19).

Chapman et al. (2000) provide a comprehensive documentation of the *Cross Industry Standard Process for Data Mining (CRISP-DM)*. The process has a multi-level structure including process steps, tasks and process instances. Tasks are divided into generic and specialized tasks, where the former are universally applicable and the latter reflect circumstances that require a specific implementation of a task. Process instances include actions, decisions and results as they occur in an individual analytics project. The reference model describes the major two levels of CRISP-DM and therefore provides its generic process design. In addition, a user guide provides more details on the implementation of this process (Chapman et al. 2000, pp. 9–10).

The CRISP-DM reference model, as shown in Figure 16, consists of six process steps that break down into 24 generic tasks that are described in detail by Chapman et al. (2000). (1) *Business understanding* initiates the process with focus on the objective of the process, the translation from the business perspective into an analytics task in consideration of existing circumstances, and the implementation of a project plan. (2) *Data understanding* starts with the collection of relevant data and additional tasks to become acquainted with this data by means of description and exploration. From this basis, issues with data quality can be identified and preliminary insights concerning the defined objective can be revealed. (3) *Data preparation* transforms original raw data into data input for the models applied during analytics. For this purpose, data is selected with regard to its relevance and quality issues are cleaned. Data construction and integration enable generation of more meaningful data input before formatting applies necessary syntactic adjustments required by the models. (4) *Modeling* represents the analytics core of the process. After selecting a set of appropriate models, each model is built, tested and assessed in order to achieve optimal performance. (5) The purpose of the *evaluation* step is twofold. On the one hand, comprehensive evaluation and a review of the development process secure robustness of working models from an analytics perspective. On the other hand, analytics results are

transferred back to the business context as they are evaluated against the initial objectives. Both perspectives enable decisions on next steps including transition to deployment. (6) The final step of *deployment* ensures application of the analytics results in business practice. This requires preparation of deployment and subsequent monitoring in daily operations. A final report and review of the project enable documentation of results and dissemination of lessons learned which frequently initiate related projects (Chapman et al. 2000, pp. 13–34).¹⁶

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p>Determine Business Objectives Background Business Objectives Business Success Criteria</p> <p>Assess Situation Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</p> <p>Determine Data Mining Goals Data Mining Goals Data Mining Success Criteria</p> <p>Produce Project Plan Project Plan Initial Assessment of Tools and Techniques</p>	<p>Collect Initial Data Initial Data Collection Report</p> <p>Describe Data Data Description Report</p> <p>Explore Data Data Exploration Report</p> <p>Verify Data Quality Data Quality Report</p>	<p><i>Data Set</i> Data Set Description</p> <p>Select Data Rationale for Inclusion / Exclusion</p> <p>Clean Data Data Cleaning Report</p> <p>Construct Data Derived Attributes Generated Records</p> <p>Integrate Data Merged Data</p> <p>Format Data Reformatted Data</p>	<p>Select Modeling Technique Modeling Technique Modeling Assumptions</p> <p>Generate Test Design Test Design</p> <p>Build Model Parameter Settings Models Model Description</p> <p>Assess Model Model Assessment Revised Parameter Settings</p>	<p>Evaluate Results Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</p> <p>Review Process Review of Process</p> <p>Determine Next Steps List of Possible Actions Decision</p>	<p>Plan Deployment Deployment Plan</p> <p>Plan Monitoring and Maintenance Monitoring and Maintenance Plan</p> <p>Produce Final Report Final Report Final Presentation</p> <p>Review Project Experience Documentation</p>

Figure 16 - CRISP-DM steps, tasks and outputs (Wirth, Hipp 2000, p. 34)

SEMMA

SAS Enterprise Miner is the analytics software for business users from SAS Institute. It provides access to various models and formalizes the analytics process in its user interface (SAS Institute Inc. 2017). *SEMMA* (*Sample, Explore, Modify, Model, and Assess*) represents the process integrated into the software solution with main emphasis on modeling (Mariscal et al. 2010, p. 144). The process is designed for ease of use and subdivides into five steps (Azevedo, Santos 2008, p. 183). Dean (2014) provides a description of the process that starts with (1) *sample* as an optional step. Taking a representative subset of available data enables exploration of interesting patterns with low processing time. Sample therefore enables idea generation for business-relevant analytics. (2) *Explore* aims to uncover unforeseen patterns in order to gain a better understanding of the

¹⁶ Mendes et al. 2012) provide a detailed example for an application of CRISP-DM in practice.

data and to further substantiate the formulation of ideas. For this purpose, SEMMA utilizes visualization and statistical methods. (3) *Modify* prepares the data input for analytics models and includes methods for data selection and construction. (4) *Model* represents the selection of analytics models where the software subsequently optimizes models for best performance. The final step (5) *assess* evaluates built models with regard to their business value and robustness (Dean 2014, pp. 61–64).

Cios et al.

Cios et al. (2000) base their process on CRISP-DM as field-proven approach (Cios, Kurgan 2005, pp. 5–6). Their advancement aims for a more generic process with stronger orientation towards research. Furthermore, they provide a higher level of detail for feedback loops and a modification of the final step enabling knowledge transfer across different domains (Cios et al. 2007, pp. 14–15). Cios, Kurgan (2005) describe the six-step process as follows: (1) *Understanding the problem domain* details out the issue to be solved with KDDM and rests on close cooperation with domain experts. Problem understanding includes relevant terminology, existing solutions and relevant constraints. The initial step sets objectives for the project that are translated into data mining goals and includes a preselection of analytics models. (2) *Understanding the data* starts with the definition of required data based on a review of sample data. Data exploration verifies practicality of data and domain knowledge prioritizes data by importance before data quality is assessed. The critical step of the process is (3) *preparation of the data* as it ultimately selects data input for data mining. Furthermore, it includes data cleaning, construction and transformation. (4) *Data mining* takes this data input and builds models according to selected training, testing and assessment methods. In (5) *evaluation of the discovered knowledge*, domain experts help to make sense of results and evaluate their impact. Only models approved in this step are considered for application and a review of the conducted process reveals lessons learned for future improvements. (6) *Using the discovered knowledge* determines the approach to apply new knowledge and includes a plan for monitoring implementation as well as a report documenting results. The final step also evaluates potential extension of the application scope to other domains (Cios, Kurgan 2005, pp. 5–9).

Cabena et al.

Cabena et al. (1997) introduce a five-step process based upon their experience with KDDM projects at IBM. They refer to the KDD process but have a focus on applications in business (Mariscal et al. 2010, p. 145). Sharma (2015) provides a summary description of the process. (1) *Determination of business objectives* identifies business issues to be resolved with KDDM. This first step also aims to clarify expectations and to ensure management support. (2) *Data preparation* comprises identification, preprocessing and transformation of data in order to provide input to the analytics models. The process aims to select relevant data from multiple sources, to ensure high data quality and to provide data input in an appropriate format to models. (3) *Data mining* feeds this input into selected models in order to develop insights regarding the business

objective. (4) *Analysis of results* is strongly interrelated with the previous step as it aims to make sense of the analytics results and to evaluate model performance. In case actionable knowledge is perceived, the process concludes with (5) *assimilation of knowledge*. Application in business operations requires a knowledge representation that promotes its transfer into business context and the development of methods that optimally utilize discovered knowledge (Sharma 2015, pp. 14–15).

Anand & Büchner

Anand, Büchner (1998) introduce a generic KDDM process that was further developed as web mining process based on marketing and sales applications covering customer lifecycles (Mariscal et al. 2010, pp. 145–146). Anand et al. (1998) and Büchner et al. (1999) describe the generally applicable eight-step process which assumes a given business issue as activation. (1) *Human resource identification* secures essential experts on the underlying domain, data and analytics. The expert team starts with (2) *problem specification* in order to gain a better understanding of the business issue and breaking it down into specific data mining tasks. Each of these tasks is associated with a data mining approach (e.g., classification) and a specific objective for its application. (3) *Data prospecting* evaluates data required for the data mining tasks. From a technical perspective, this step considers data access and data storage, while it also includes identification of relevant subsets and integrity of available data. During (4) *domain knowledge elicitation*, data mining and domain experts jointly identify domain knowledge to be incorporated in analytics. (5) *Methodology identification* defines appropriate analytics models for the specified data mining tasks (e.g., SVM for a classification task). (6) *Data preprocessing* addresses data quality issues such as outliers and missing values. Furthermore, this step prepares data for the analytics models by transforming and constructing data input. (7) *Pattern discovery* takes the processed data input and builds selected models in order to reveal patterns. In the final step, (8) *knowledge postprocessing*, identified patterns are examined for relevant insights which are then presented to users in business operations. However, further validation of results is required in order to ensure robustness before actual application in practice. The process explicitly points to necessary refinement iterations between pattern discovery and knowledge postprocessing but refinement can be generally required in any of the process steps (Anand et al. 1998, pp. 449–461; Büchner et al. 1999, pp. 13–22).

3.2.2.2 Other KDDM and data mining processes

Similar to SEMMA, there exist other KDDM processes connected to analytics software. According to Oprean (2011), the statistics software package SPSS introduced *5 A's* as a five-step process including assess, access, analyze, act and automate. The focus laid on automation of data mining tasks in order to enable novice users to perform analytics. However, the process did not include steps for business or data understanding and "[...] was abandoned in 1999" (Oprean 2011, pp. 8–9). Debuse et al. (2001) introduce a process with eight steps influenced by KDD process and CRISP-DM. Their *KDD roadmap* includes an inner feedback loop that

connects each step and also represents potential repetitions within the lifecycle, however, the major difference is an additional resourcing step after problem specification (Debusse et al. 2001, pp. 180–181). The process was used as blueprint for a KDDM software toolkit (Debusse et al. 2001, p. 195). The statistics and analytics software provider StatSoft included a *six sigma*-based process in their STATISTICA Data Miner beneficial for industrial applications. It follows the five steps of the standard DMAIC concept: define, measure, analyze, improve, and control (StatSoft 2013). Kudyba, Hoptroff (2001) introduce a process with nine steps and clear focus on the analytics part of KDDM, whereby they emphasize its strong reference to six sigma as well (Kudyba, Hoptroff 2001, pp. 45–57).

Two Crows Consulting describes a practitioner process with seven steps that includes building a project database as comprehensive step after definition of the business problem. The database builds the basis for all subsequent steps and the process also directs towards its maintenance (Two Crows Corporation 1999, pp. 22–33). The *Two Crows* process is otherwise very similar to the KDD process (Mariscal et al. 2010, p. 145). Solarte (2002) introduces a process called *Data Mining for Industrial Engineering (DMIE)* inspired by CRISP-DM. It leverages a systems analysis approach in order to provide a process suitable for the industrial engineering domain (Solarte 2002, pp. 27–70). The major advancement of DMIE is an independent step "[...] involving data backups, data maintenance, data mining model updates and software updates when needed" (Mariscal et al. 2010, pp. 151–152). Castellano et al. (2007) provide another KDDM process based on CRISP-DM. It merges the business and data understanding steps and expands the process by an additional step dedicated to maintenance of discovered knowledge after its deployment (Castellano et al. 2007, pp. 479–483).

Brachman, Anand (1996) refine the KDD process with a focus on essential process tasks and the introduction of a support environment based on their practical experience. In doing so, they emphasize the role of humans involved in the process. Their *human-centered* approach reflects complex interactions with the data and required support in order to better integrate humans in the process (Brachman, Anand 1994, pp. 2–10). As a consequence, the approach "[...] shows in a clearer way which decisions the user has to make" (Mariscal et al. 2010, p. 144). Gertosio, Dussauchoy (2004) build on this human-centered process for industrial applications. They further propose an economic evaluation for budget limitations and an industrial evaluation for assessment of analytics models during application (Gertosio, Dussauchoy 2004, p. 36). Haglin et al. (2005) also present a version of the KDD process with emphasis on the human role. They specifically focus on the role of a scientist as their focus lies on research work (Haglin et al. 2005, pp. 41–42).

According to Blockeel, Moyle (2002), the *RApid collaborative data Mining SYStem (RAMSYS)* extends CRISP-DM for collaborative work within dispersed teams. It rests on the idea that more data mining experts should promote better results of KDDM projects. RAMSYS builds on three pillars including the areas of communication, problem solving and knowledge sharing. The process requires every team involved to provide required skills and knowledge such that

process management focuses on information flow, workflows, and security instead of direct team control (Blockeel, Moyle 2002, p. 22). RAMSYS requires support by some groupware that manages project information such as metadata or code descriptions (Bohanec et al. 2001, p. 8). There also exist other proposals for collaborative KDDM processes that support involvement of multiple experts in the process (Horeis, Sick 2007; Diamantini et al. 2006).

Data mining is often seen as the analytics step within a KDDM process. However, there exists comprehensive data mining literature that generally includes a description of the *data mining process*, but with largely differing levels of detail. For example, Runkler (2010) and Han et al. (2012) limit their processes to the core activities (Runkler 2010, pp. 1–3; Han et al. 2012, pp. 6–8):

- Data collection and integration
- Data preprocessing (cleaning & transformation)
- Data selection
- Modeling
- Evaluation, interpretation and presentation of results

At an advanced level, elaboration of the underlying problem complements the process upstream. Domain-specific problem definitions in the form of initial hypotheses about knowledge in the data represent a typical approach here (Kantardzic 2011, pp. 6–9). Application of data mining results represents further downstream advancement of the process that feeds into decision making by its users (Vercellis 2009, pp. 84–90). Furthermore, processes ranging from problem definition to application add supplemental steps or tasks, for example, exploratory analysis for more directed data preprocessing (Giudici 2003, pp. 6–10) or capture of lessons learned for future improvements (Ahlemeyer-Stubbe, Coleman 2014, pp. 19–30). Moreover, some processes describe different roles in the team performing a data mining project (Hofmann, Tierney 2009, pp. 54–55). These advanced data mining processes become very similar to KDDM processes previously discussed and some directly reference to CRISP-DM (Nisbet et al. 2009, pp. 34–46).

Berry & Linoff

Berry, Linoff (2004) introduce a four-phase process with a detailed description on the analytics part. The first phase of their *virtuous cycle of data mining* aims to (1) *identify the business opportunity* in order to ensure that potential results will be useful and applied in the organization. Business experts need to assist this effort and specific data mining techniques should not be in focus yet. (2) *Transforming data into information* represents the analytics part of the process and (3) *taking actions* describes the transfer of analytics results into business operations. (4) *Measuring the outcome* closes the cycle and builds the basis for continuous improvement (Berry, Linoff 2004, pp. 26–32). The second phase represents the core of the overall process addressing various challenges

such that Berry, Linoff (2004) provide a detailed description of ten steps that are shown in Figure 17. Identified business problems are translated into analytics problems by selecting appropriate analytics approaches. Availability and relevance guide the selection of appropriate data which is subsequently explored and reduced to data input for modeling. Data is cleaned and transformed in order to reveal most meaningful information before models are built and assessed. Deployment transfers performing models from the analytics to the operations environment where they can be assessed during application. Finally, new ideas or issues raised throughout the process initiate a new loop of data mining (Berry, Linoff 2004, pp. 54–86).

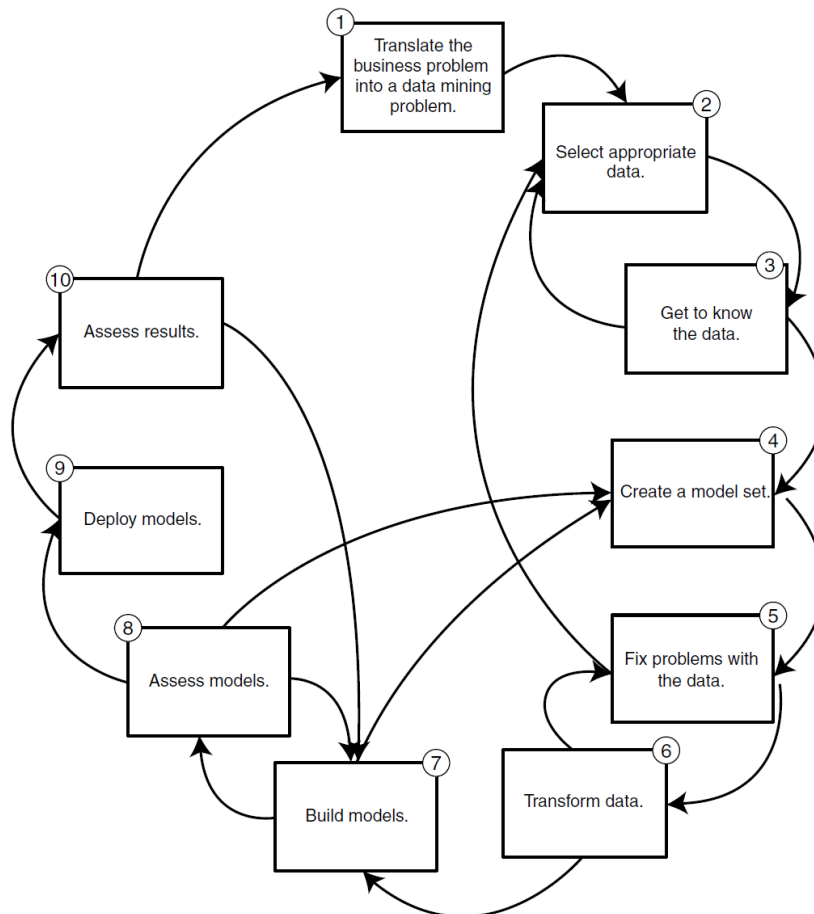


Figure 17 - Data mining process (Berry, Linoff 2004, p. 55)

3.2.2.3 Comparison of basic KDDM processes

Various comparative studies describe KDDM processes with regard to major characteristics and point towards advantages as well as disadvantages. Table 3 provides an overview for basic KDDM processes based on six different studies. Cios, Kurgan (2005) emphasize extensive practical application for CRISP-DM and Cios et al. They list model selection late in the process and the lack of a data understanding step as shortcomings of the KDD process and Cabena et al. (Cios, Kurgan 2005, pp. 9–10).

Basic KDDM processes							
Source	Comparison type	KDD process	CRISP-DM	SEMMA	Cios et al.	Cabena et al.	
Cios, Kurgan (2005)	Assessment	- Selection of appropriate models late in the process causes unnecessary iterations	+ Extensive validation in business applications		+ Application in various medical projects	- Lack of data understanding step	Anand & Buchner
Kurgan, Musilek (2006)	Assessment	+ Explicitly includes iterations - Limited details for iterations - Model choice after data preprocessing (unnecessary iterations)	+ Use of straightforward language + Comprehensive documentation + Hierarchical organization of steps and tasks - Limited details on iterations		+ Clear description of iterations and interactions - Orientation towards research applications	+ Accessible for novice users + Explicitly includes iterations - Limited details on iterations step	+ Clear description of iterations - Lack of step for application of results - Limited details on iterations
Francis (2008)	Application areas	Various	Various		Medical science, software	Business	Business
Marban et al. (2009a)	Software support	yes	yes	yes*	no	no	no
Mariscal et al. (2010)	Industry engagement	no	yes		no	yes	no
	Usability	low	high		medium	medium	low
	Assessment	+ High level of detail	+ Based on comprehensive feedback from practitioners	- Focus on analytics steps	- Lack of tasks to reduce data scope for modeling		
	Similarities	n/a	n/a	n/a	Based on CRISP-DM with focus on academic research	Based on KDD process and similar to Anand & Buchner	Based on KDD process and similar to Cabena et al.
	Assessment	+ Open process applicable in various environments	+ Neutral towards different domains and software toolkits	- Integration in commercial software (limits use in other environments) - Lack of step for business understanding - Lack of step for application of results	- Based on specific technologies (e.g., Extensible Markup Language (XML))	- No significant differences from KD process	- No significant differences from KD process
	Delta to CRISP-DM	- Lack of detailed explanations for business understanding step - Lack of detailed explanations for application step	n/a	- Lack of business understanding step - Lack of step for application of results	+ Institutionalized iteration steps	- Lack of documentation	- Model selection before data preprocessing
Rogalewicz, Sika (2016)	Assessment	- No focus on analytics methods	+ Strong feed of business understanding into data preprocessing	- Strong relation to commercial software			

* Information added based on SAS Institute Inc. (2017)

process not covered + advantage / - disadvantage

Table 3 - Comparison of basic KDDM processes

RELATED WORK

Kurgan, Musilek (2006) compare various characteristics in addition to an overall assessment. KDD process and CRISP-DM are the only processes with various application areas and software tool support. The latter is furthermore based on engagement by industry as is Cabena et al. as well. The study identifies CRISP-DM as the only process with high usability which builds on its straightforward language, comprehensive documentation, and hierarchical organization of steps and tasks. Cios et al. is the only process with comprehensive details on iterations and interactions throughout the process. Moreover, the study lists the lack of a step dedicated to application of analytics results as major shortcoming of Anand & Büchner (Kurgan, Musilek 2006, pp. 12–16). Francois (2008) includes SEMMA in the comparison and highlights its analytics focus that results in a lack for objective formulation and deployment of results. Furthermore, Cios et al. lacks tasks to reduce the scope of data for effective modeling (Francois 2008, p. 242). Marban et al. (2009a) investigate similarities among the processes and determine Cios et al. as derivative of CRISP-DM. Cabena et al. and Anand & Büchner derive from KDD process and show high similarity among each other (Marban et al. 2009a, pp. 2–5). Mariscal et al. (2010) add the limited use outside of the underlying software as additional disadvantage of SEMMA while CRISP-DM is neutral in this respect. Cios et al. builds on specific technologies which represents an important limitation of the process. Moreover, a direct comparison with CRISP-DM reveals a lack of critical steps for SEMMA, inadequate explanations of these steps for KDD process, and the absence of documentation procedures for Cabena et al. (Mariscal et al. 2010, pp. 144–157). Rogalewicz, Sika (2016) highlight the strong interrelation of business understanding and data preprocessing steps in the case of CRISP-DM. They furthermore note that KDD process does not focus on analytics methods (Rogalewicz, Sika 2016, pp. 100–101).

Basic KDDM processes						Data Mining
KDD process	CRISP-DM	SEMMA	Cios et al.	Cabena et al.	Anand & Büchner	Berry & Linoff*
9 steps	6 steps	5 steps	6 steps	5 steps	8 steps	10 steps
Learning the application domain	Business understanding		Understanding the problem domain	Determination of business objectives	Domain Knowledge Elicitation	
					Human resource identification	
					Problem specification	
Creating a target dataset	Data understanding	Sample	Understanding the data	Data preparation	Data prospecting	Select appropriate data
		Explore			Methodology identification	Get to know data
Data cleaning and preprocessing	Data preparation	Modify	Preparation of the data		Data preprocessing	Fix problems with data
Data reduction and projection						Transform data
Selecting analytics method	Modeling	Model	Data mining	Data mining	Pattern discovery	Build models
Selecting model						
Data mining	Evaluation	Assess	Evaluation of the discovered knowledge	Analysis of results	Knowledge postprocessing	Assess models
Interpretation						
Application of discovered knowledge	Deployment		Using the discovered knowledge	Assimilation of knowledge		Deploy models
						Assess results

step not covered by process

* based on Berry, Linoff (2004)

Table 4 - Comparison of process steps

[based on (Mariscal et al. 2010, p. 159; Kurgan, Musilek 2006, p. 6; Oprean 2011, p. 12)]

In order to assess the scope of basic KDDM processes, Table 4 provides a comparison of process steps taking into account associated tasks. The overview adapts the lifecycle of Anand & Büchner for better comparison and includes Berry & Linoff as an example for an advanced data mining process. The comparison clearly illustrates the disadvantages in form of lacking steps for SEMMA and Anand & Büchner. Furthermore, the remaining processes show a high level of consistency in terms of their scope. As representative example of data mining processes, Berry & Linoff confirm this conclusion because its scope is only partly reduced with regard to initial domain understanding. However, Berry & Linoff must be seen as part of the virtuous cycle of data mining whose first phase addresses this issue.

As a conclusion of the comparisons presented, CRISP-DM stands out as universally applicable process with industry support and comprehensive documentation. It generally offers the most advantages compared to the other processes. CRISP-DM has become "[...] the de facto standard for developing data mining and knowledge discovery projects" (Mariscal et al. 2010, p. 146). Polls on the main processes used for "[...] analytics, data mining, or data science projects" (Piatetsky 2014) confirm this statement. As shown in Figure 18, CRISP-DM is the most widely used process by a large margin for more than a decade. The initiators of CRISP-DM already noted that users of the process benefit from repeated application in project teams (Wirth, Hipp 2000, p. 38). Moreover, IBM recommends CRISP-DM when using its SPSS Modeler software for analytics (IBM 2017a, p. 25). The following discussion on evaluation and advancements of KDDM processes consequently puts CRISP-DM at its center. SEMMA as alternative process with a significant share in the poll has multiple shortcomings as explained before. Proprietary processes (*My own* or *My Organizations'*) are typically tailored for specific applications or an existing software infrastructure (Li et al. 2016b, p. 2) and they are usually not accessible for detailed investigation. Nevertheless, the recent increase for own and other processes indicates a need that might not be served by CRISP-DM.

Main process for analytics projects
Share of respondents [%]

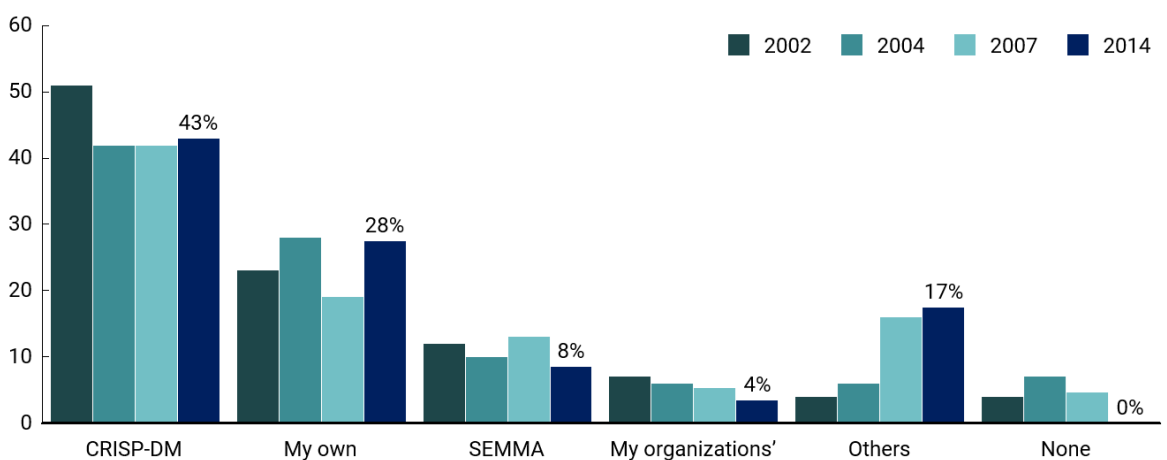


Figure 18 - Poll on processes used in projects
[data taken from: Piatetsky (2014), KDnuggets (2004) and KDnuggets (2002)]

3.2.2.4 Evaluation of basic KDDM processes

Despite its prolonged success, there exist weak points indicating that "[...] a replacement for unmaintained CRISP-DM is long overdue" (Piatetsky 2014). Marban et al. (2009b) state that the process lacks tasks in comparison to established software engineering approaches, especially regarding project management and the post-development phase. They claim the need for these tasks in industry projects (Marban et al. 2009b, pp. 93–105). Sharma (2008) contributes a comprehensive study on limitations of KDDM processes with focus on CRISP-DM. The study identifies the checklist approach as major shortcoming because it hinders process implementation (Sharma 2008, p. 32). To be more precise, a specific method is only provided for less than 10% of all tasks and these might not even meet requirements for efficient implementation (Sharma 2008, pp. 23–24). The observation that CRISP-DM mainly "[...] defines *what to do* and not *how to do*" (Mariscal et al. 2010, p. 139) underlines this issue. Fischer et al. (2014, pp. 169–170) confirm the need to substantiate the abstract process.

Sharma (2008) further reveals that the business understanding step is regularly implemented in provisional manner despite its key role in the process. Insufficient guidance by CRISP-DM for this step can potentially result in inefficient or unsuccessful projects (Sharma 2008, pp. 41–44). Interdependencies within the process are adequately covered only on step level while there is high fragmentation on the task level, therefore impeding automation of the process (Sharma 2008, pp. 33–37). Especially in real-time applications, CRISP-DM hinders automation due to its "[...] manual process of steps [...]" (Siriweera et al. 2015, p. 276), although availability of advanced technology for analytics in form of software and hardware indicate a shift towards automation of adequate tasks (Shahapurkar 2016, p. 36).

Sharma, Osei-Bryson (2009) identify another deficiency of CRISP-DM. It does not explicitly describe the role of humans for tasks and also does not "[...] describe the manner in which human intelligence could be leveraged" (Sharma, Osei-Bryson 2009, p. 53). In particular, they describe various tasks that would benefit from domain expert participation, for example, preparation of the project plan, selection and preparation of data input, or formulation of objectives (Sharma, Osei-Bryson 2009, pp. 54–60). Domain expertise requires regular communication within the project team which is especially important in cases where data scientists have no domain background (Ahangama, Poo 2015a, 6–7). Insufficient consideration of project management (Marban et al. 2009b, p. 94) and ineffective project organization (Mariscal et al. 2010, pp. 162–163) are further shortcomings of CRISP-DM.

Evaluation of KDDM processes in general confirm these findings. Users have to make various decisions based on multiple choices but existing processes show a "[...] lack of user guidance" (Oprean 2011, p. 7). Karunakaran (2013, p. 113) criticizes an inordinate focus on data-related tasks with technical focus that hinders adequate consideration of business aspects, in particular, a lack of business objectives adversely affects application of results. Furthermore, it requires domain experts in order to make sense of analytics results. Implementation of industrial

standards is a general request towards KDDM processes in order to facilitate automation (Kurgan, Musilek 2006, p. 19). As all KDDM processes involve different actors, there is "[...] a need for a collaborative process model that covers the full spectrum of actors and interactions involved" (Tuovinen 2016, p. 240). Anand (2012, pp. 14–16) points out that processes generally assume availability of relevant data that needs to be preprocessed for analytics and thus neglect data source issues in practice.

A review of data mining literature reveals the same issues, amongst others, and thus further verifies previous findings. Charest et al. (2006) criticize the strong focus of data mining research on technology that leads to insufficient consideration of methodological aspects. Their assessment points to the abstract character of KDDM processes and particularly identifies a need for "[...] explanations, heuristics and recommendations on how to effectively carry out the particular steps of the methodology" (Charest et al. 2006, p. 593). Another observation is the limited use of domain expertise that has adverse effects on the results (Charest et al. 2006, p. 593). There is no question about the explorative character of data mining, but a lack of understanding on the intended results contributes to project uncertainty (Singh et al. 2011, p. 280). It underlines the need for a thorough business understanding including formulation of objectives at the beginning of the process. Lavrac et al. (2004) recommend senior management support for data mining projects and also highlight the requirement of clear objectives for a successful project. They furthermore recognize the need to perform data mining in a team setting with adequate project management. In particular, the team lead should not be held by a data scientist but someone with business background (Lavrac et al. 2004, pp. 20–21). Automation of data mining processes is also seen as key topic because it can help to avoid errors and to increase project efficiency (Yang, Wu 2006, p. 602).

Another way to look at KDDM projects is to identify success factors. Barclay (2015) derives critical success factors from a project in practice. These factors include basic requirements regarding resources in form of personnel, software and budget. Moreover, many success factors relate to fundamentals of analytics including data access, data sourcing, data handling, data quality and use of an appropriate analytics model. Besides these basic factors, project team members need to possess capabilities with regard to KDDM processes and suitable technology. Business understanding including clear objectives is identified as success factor that benefits from domain expert input. Business objectives also need to be aligned with analytics objectives. In general, ongoing involvement of users is beneficial to project success. Commitment of key stakeholders represents a specific form of domain expert involvement and buy-in of data owners strongly builds on active consideration of confidentiality concerns. Finally, project leadership needs to be in charge of the KDDM process itself and must not limit itself to general project management activities (Barclay 2015, pp. 175–182). The study of Nemati, Barko (2003) reveals similar basics with regard to data, technology, capabilities, and general project management as success factors. KDDM proficiency of users is identified as critical success factor because it supports identification with the project. In addition, findings of the study propose outsourcing

RELATED WORK

of KDDM projects in case the organization lacks required capabilities (Nemati, Barko 2003, pp. 285–291). Hilbert (2005) confirms the crucial role of basic factors with the addition of senior management commitment. Sim (2003, pp. 81–84) proposes to put more focus on data accessibility, quality, complexity and volume as success factor of data mining projects.

The observations on process shortcomings and project success factors can be grouped into six improvement areas for KDDM processes. (I) *Project team* summarizes all improvement potentials with regard to an effective project organization which includes the team setup as well as the working mode. (II) *Domain knowledge* demands to involve domain experts and users along the entire lifecycle of a process. Furthermore, (III) *business understanding* refers to the need to build a thorough understanding of the domain, in particular for business applications. An increased focus on (IV) *data input* and provision of (V) *methods* that support implementation of a process are further improvement areas. Finally, (VI) *automation* is viewed as enabler for project efficiency and measure to avoid errors. Table 5 provides an overview on these improvement areas including specific dimensions from the previous discussion.

	(I) Project team	(II) Domain knowledge	(III) Business understanding	(IV) Data input	(V) Methods	(VI) Automation
Main objective	Enable effective project organization	Extensive involvement of domain experts and users throughout the lifecycle	Avoid provisional implementation in order to ensure thorough	Increased focus on data input	Address the "how to do" and not only "what to do"	Enable automation to avoid errors and to increase project efficiency
Dimensions	Cover full spectrum of roles involved	Include domain experts in objective formulation, data selection, data preparation and evaluation of results	Formulation and use of clear business objectives	Explicitly consider data sourcing as task (do not assume relevant data to be directly available)	Provide explanations for proposed methods	Consider dependencies on task level (integrated process)
	Provide required analytics capabilities		Alignment of business and analytics objectives	Identify relevant data sources and clarify access	Provide guidance on necessary choices	Leverage technology (esp. software tools)
	Consider outsourcing in case of lacking internal capabilities			Consider quality, complexity and volume of data	Enable regular collaboration including domain experts	Implementation of standards
	Secure stakeholder support, especially senior management and data owner					
	Team leader with business background					
Team leader with responsibility for analytics process (not only general project management)						

Table 5 - Improvement areas for KDDM processes

3.2.3 Advanced processes

This subsection provides an overview of advanced KDDM processes grouped by different approaches. The first group includes direct derivatives of CRISP-DM followed by processes stemming from the engineering domain. The remainder of the subsection presents approaches that address specific dimensions of a KDDM process.

3.2.3.1 CRISP-DM 2.0

In 2006, a working group was initiated in order to update CRISP-DM to CRISP-DM 2.0, however, the effort apparently discontinued as there is no more activity by this group (Wikipedia 2017). There still exist various minor advancements of CRISP-DM. Asamoah, Sharda (2015) describe a substantiation of the process for the case of unstructured data analytics in the healthcare domain. Their approach offers guidelines that help to specify the steps of data understanding and data preparation in this context. For example, they propose a keyword

approach and specific filtering methods for extraction and cleaning of social network data (Asamoah, Sharda 2015, 4–10). Wieland, Fischer (2013) substantiate CRISP-DM with focus on early steps in case of analytics in manufacturing processes. Their major advancement of CRISP-DM rests on holistic modeling of the manufacturing process as input-output system. Based on this, they derive a target data structure that guides identification of relevant data sources. Moreover, the model-based approach enables integration of domain experts (Wieland, Fischer 2013, 53–62). Mariscal et al. (2010) introduce a combination of CRISP-DM and KDD process including their derivatives as new process. This *Refined Data Mining Process* comprises 17 steps organized in three phases which mainly reduces inter-step dependencies. The increased level of detail aims for a better ease of use (Mariscal et al. 2010, pp. 159–162), but it remains unclear how to implement each step (Li et al. 2016b, p. 2). IBM Analytics Services proposes an advancement of CRISP-DM named *Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM)* that generally maintains the analytics part but adds additional tasks and methods (Haffar 2015) as well as modernizes terminology in use (Jensen 2017). The *Analytics Solution Unified Method (ASUM)* is part of IBM solutions for analytics (Roman 2016) and is leveraged to extend CRISP-DM (Brethenoux 2016). IBM (2016) describes key features of ASUM including the following: integration of agile principles, project management system for multiple projects based on industry standards, and adoption of industry standards for validation. With the new approach, IBM also aims to further emphasize the step of deployment (see Figure 19) as critical step for value capture in analytics projects and IBM integrates proprietary methods into the process (Brethenoux 2016). However, availability of the solution is basically limited to an IBM environment (IBM 2016, p. 3) and the main objective of ASUM-DM is to cover infrastructure and operational issues of a KDDM process which are both not covered in CRISP-DM (Wierse, Riedel 2017, p. 234).

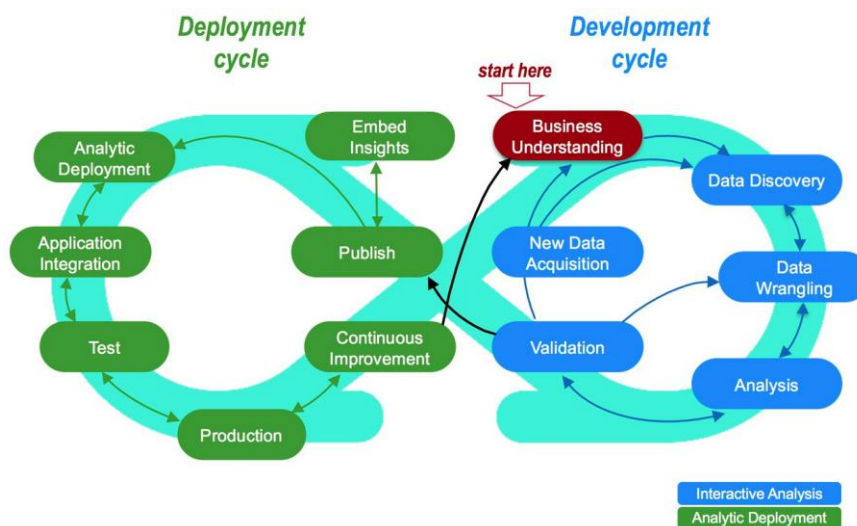


Figure 19: Deployment and development cycles (Brethenoux 2016)

Martins et al. (2016b) propose *MoProPEI*¹⁷ that also augments the development process with an additional management process and aims to overcome the incomplete character of DME. An adapted version of CRISP-DM serves as development process and the management process represents a transversal layer to development resting on methods developed by the research group of the authors (Martins et al. 2016b, pp. 505–508). Figure 21 provides an overview of the comprehensive management process.

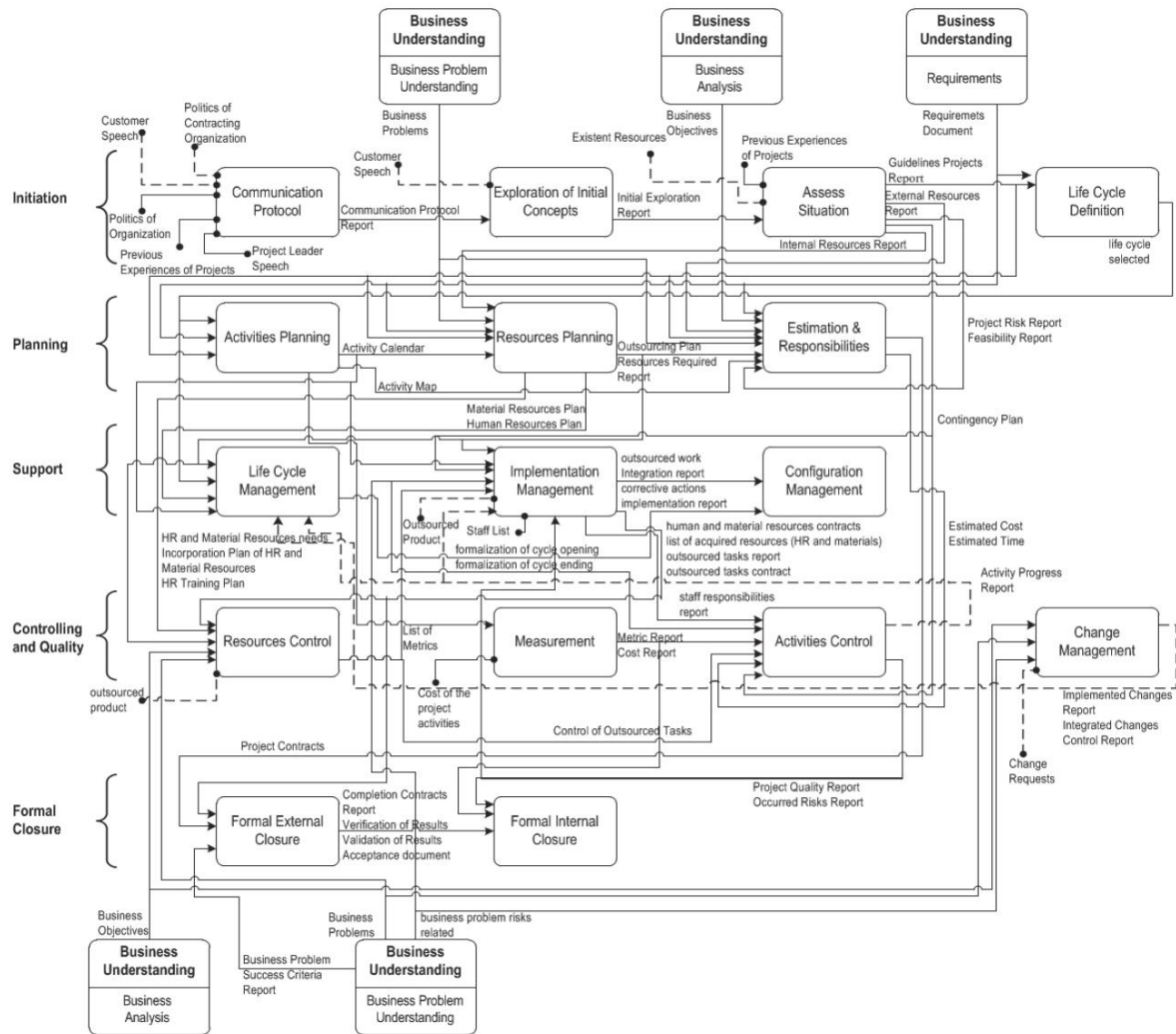


Figure 21: Management Process of MoProPEI (Martins et al. 2016b, p. 507)

Martins et al. (2016b) discuss various advantages of MoProPEI in comparison with CRISP-DM. A project feasibility analysis early in the process avoids initiation of projects with no prospect of success. Establishing metrics to plan project progress and outcome serves as basis for improvements of both. Monitoring of costs and timeliness of the project aims to keep them according to plan. Planning also covers personnel and material resources over the entire lifecycle

¹⁷ Abbreviation for *Modelo de Proceso de Proyectos de Explotación de Información* (Spanish) which can be translated as *information mining project development process model* (Martins et al. 2016a, p. 4).

which ensures project progress. A documentation management technique helps to coordinate teamwork, to iterate to previous stages and to transfer results to future projects. MoProPEI identifies outsourcing options and includes tasks to manage outsourced project activities. In addition, legal commitments define responsibilities in case multiple parties are involved in the project. Furthermore, an analysis of project characteristics helps to select an optimal development process and limits its scope to necessary tasks such that duration and costs are optimized (Martins et al. 2016b, pp. 507–509). In summary, MoProPEI particularly addresses shortcomings of CRISP-DM "[...] associated with the maturity level and success rate of projects" (Martins et al. 2016b, p. 508).

Rohanizadeha, Moghadama (2009) introduce an approach based on the engineering design process for application in industrial operations. Their process is referred to as *Engineering Design Process for Data Mining (EDP-DM)* and is organized into five steps including 15 tasks. EDP-DM proposes a framework of change factors, such as new materials or customers, as technique to formulate project objectives. Analytics tools are identified as a key resource and the process includes a comprehensive method to select an appropriate tool for the project. It takes price, performance, functionality, usability, support, and analytics to be performed into consideration. A decision matrix assesses tools along these dimensions in order to compare alternatives on the basis of an overall score. Another decision matrix serves as method for project valuation (Rohanizadeha, Moghadama 2009, pp. 43–49). Except for the three presented methods, EDP-DM does not provide any further improvement compared to CRISP-DM.

Shahapurkar (2016) describes the *Design for Deployment (DFD)* process by applying the systems engineering approach to CRISP-DM. The result is a lifecycle in form of the typical V-model comprising 15 steps with modeling at its center. DFD is motivated due to increasing importance of mission critical applications in analytics, for example, self-driving cars. While traditional KDDM processes primarily focus on creating performing models, DFD puts a major focus on reliability. The right branch (upwards) of the V-model therefore consists of validation steps that are coupled with steps on the same level of the downward branch. This structure represents the idea that each development step needs to think ahead of the validation (Shahapurkar 2016, 36–46). Overall, DFD has a strong focus on deployment and especially demonstrates its main advantages in ongoing analytics efforts rather than individual projects (Shahapurkar 2016, p. 152).

3.2.3.3 Integrated Knowledge Discovery and Data Mining

Sharma (2008) introduces the *Integrated Knowledge Discovery and Data Mining (IKDDM)* process. The design of the approach rests on a detailed study on dependencies of tasks within and across different steps. Furthermore, it follows the principle to support execution of tasks by "[...] semi-automating the dependency relationships [...] and through a set of [methods] [...]" (Sharma 2008, p. 385). Sharma, Osei-Bryson (2010) elucidate the motivation for IKDDM based on limitations of existing KDDM processes. They point towards the checklist character and

disregard of task dependencies, especially for CRISP-DM. The lack of explicating dependencies represents an impediment to semi-automation because execution of dependent tasks cannot benefit from previous results. For example, business objectives directly feed into analytics objectives and therefore determine the relevant scope of the latter. Furthermore, they state lacking methods, especially for the business understanding step, as motivation (Sharma, Osei-Bryson 2010, pp. 51–52). According to the process overview by Sharma, Osei-Bryson (2015b), IKDDM proposes nine different methods to support the business understanding step while the provision of methods is limited to generic tools such as analytics or spreadsheet software for data understanding and preparation steps. However, the process consistently lists sources of valuable input to perform each task of these initial steps, whereby domain experts are also included across all steps (Sharma, Osei-Bryson 2015b, pp. 33–35). For modeling and evaluation steps, the how-to character of the process mainly stems from clearly defined tasks instead of methods (Sharma 2008, pp. 247–289). A comparative study of CRISP-DM and IKDDM based on a survey indicates higher efficiency and effectiveness of the advanced process (Sharma et al. 2012, pp. 11341–11347).

3.2.3.4 Domain Driven Data Mining

According to Cao (2009), *Domain Driven Data Mining (DDDM)* is motivated by the fundamental goal of KDDM processes to provide actionable knowledge. Hence, different forms of intelligence need to be jointly integrated in the process in order to achieve this. Data and network intelligences basically refer to patterns in structured and unstructured data. Human intelligence has two dimensions and includes explicit involvement of humans in the form of empirical knowledge or beliefs, for example. On the other hand, it also refers to implicit involvement such as emotional intelligence or inspiration. Social intelligence describes interactions of human actors. Finally, domain intelligence "[...] refers to domain resources that not only wrap a problem and its target data but also assist in the understanding and problem-solving of the problem" (Cao 2009, p. 5). Cao (2010) provides more details on each intelligence type and related methods for implementation of DDDM. Domain intelligence builds upon formalization of domain knowledge and an interaction design enabling transfer of knowledge from domain experts into analytics. Semantic webs and ontological engineering are presented as exemplary methods hereto. Group decision making or adaptive interaction describe methods supporting the integration of human intelligence. Social intelligence comprises team interactions and swarm intelligence such that common project management does only address a small part of relevant interactions (Cao 2010, 758–761). Kumari (2011) states an orientation towards problem-solving abilities and deliverables as major differences compared to KDDM processes that are more technically focused and aim for automated analytics (Kumari 2011, p. 66).

3.2.3.5 Agile approaches

The *Manifesto for Agile Software Development* laid down key principles of agile software development (Agile Alliance 2017) at about the same time as CRISP-DM was established (Shahapurkar 2016, p. 36). Alnoukari et al. (2009) introduce a process following the agile software development paradigm. It leverages the fact that intangible analytics can be adapted at low costs. They recommend their approach for cases with uncertain requirements and leverage *Adaptive Software Development (ASD)* in order to replace "[...] the static Plan-Design-Build lifecycle, with the dynamic Speculate-Collaborate-Learn life cycle" (Alnoukari et al. 2009, p. 154). Their *Adaptive Software Development for Data Mining (ASD-DM)* process, combines business and data understanding in the preparatory *speculation* step. It is followed by an iterative cycle of *collaborative modeling* and *learning* due to evaluation and deployment. Modeling is collaborative as it depends on strong stakeholder involvement. Deployment is implemented in the form of testing such that experimentation is encouraged (Alnoukari et al. 2009, p. 154).

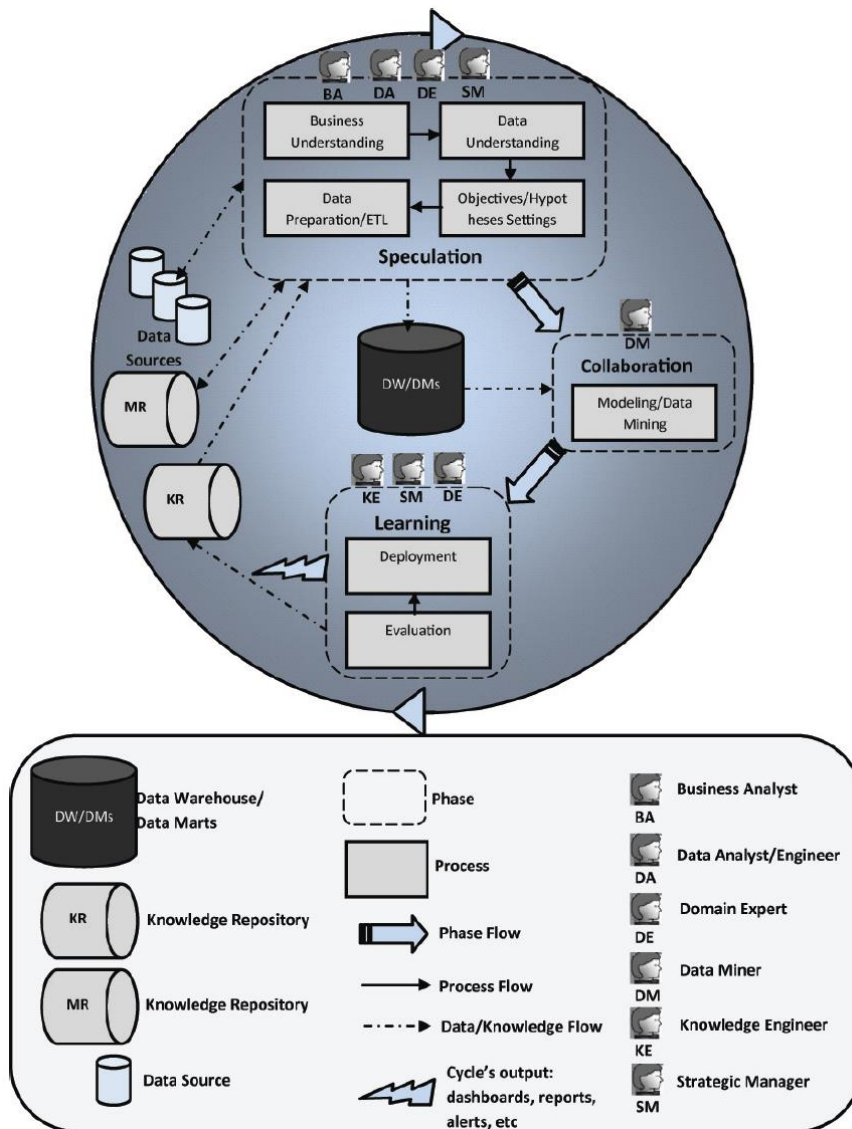


Figure 22: Adaptive Software Development for Business Intelligence (Alnoukari 2012, p. 189)

Alnoukari (2012) advances ASD-DM in various ways resulting in a new KDDM process referred to as *Adaptive Software Development for Business Intelligence (ASD-BI)*.¹⁸ Firstly, metadata and existing knowledge enrich the data sources before entering the Speculate-Collaborate-Learn lifecycle. Secondly, it adds collaborative workshops held in an iterative cycle as methodology in the speculation phase. This phase results in a data mart or warehouse that is ready for analytics. Thirdly, ASD-BI incorporates roles and responsibilities into the process including "[...] business analyst, data analyst/engineer, data miner, domain expert, knowledge engineer, and strategic manager" (Alnoukari 2012, p. 192). For instance, domain experts play a key role in the speculation phase while data miners are responsible for the modeling. Lastly, agile methodologies like small releases or planning games are integrated for project management (Alnoukari 2012, pp. 187–194). Figure 22 provides an overview of ASD-BI.

A further approach based on an alternative process for agile software development, called Open Unified Process, is presented by Nascimento, Oliveira (2012). It organizes the KDDM process into four phases with a total of 16 steps but is restricted to the description of this lifecycle (Nascimento, Oliveira 2012, pp. 61–63). *Agile Analytics* also proposes agile software principles for analytics work and related subjects such as data warehouse building (Collier 2012, p. 3). Despite providing a comprehensive overview of methods, Agile Analytics does not offer a holistic KDDM process.

3.2.3.6 Evaluation of advanced KDDM processes

The Refined Data Mining Process represents one of the major derivatives of CRISP-DM. As it mainly represents a combination with other KDDM processes, it does not address identified improvement areas. ASUM-DM as an alternative derivative has a focus on foundational infrastructure, deployment and project management. It implements industry standards which represent a key dimension for automation. ASUM-DM also provides methods but they are proprietary and therefore restricted to the use within IBM software. Most engineering approaches have major limitations. DME must be seen as blueprint rather than an applicable approach. The process advancement by EDP-DM is limited to a small number of methods and DFD provides a process for the special case of analytics with strict requirements regarding validation. MoProPEI provides major improvements regarding an effective project organization due to comprehensive project management and also provides methods in this area. However, the process also comes with a high level of complexity. The major advantage of IKDDM is the thorough design of the business understanding step where the process also contributes most of its methods. Based on the integrated design, the process also enables automation but mainly by labelling candidate tasks instead of actual implementation. DDDM extends project management

¹⁸ Although referring to business intelligence, the process aims to "[...] enhance the way of building business intelligence and data mining applications" (Alnoukari 2012, p. 183).

and emphasizes the role of humans in the project team. In doing so, the process emphasizes the key part domain knowledge plays in a project. Moreover, DDDM provides methods related to teamwork and inclusion of domain knowledge. ASD-BI as most advanced process among agile approaches substantiates the team setup with specific roles and responsibilities. It also includes selected methods for team collaboration and project management. In summary, none of the advanced KDDM processes exhaustively consider improvement areas because they typically focus on a specific area. They also do not match major advantages of CRISP-DM, especially its high usability and comprehensive documentation.

3.2.4 Big data analytics processes

The 4V characteristics cause new challenges with regard to big data projects (Grady et al. 2014, pp. 12–13). For instance, heterogeneous and unstructured data differentiates big data analytics even from closely related domains such as statistics (Dhar 2013, p. 64). Although technological foundations are imperative for successful BDA implementation, the innovative character of big data requires "[...] special attention to processes and people involved in Big Data projects" (Gao et al. 2015, p. 827). The specific requirements of BDA projects, especially based on its focus on data, underlines the need for different processes (Saltz et al. 2017b, p. 1015). Studies confirm that BDA capabilities of management and personnel (Wamba et al. 2017, pp. 362–363), organization and culture, and processes are at least as important as the underlying big data technology (Clark, Wiesenfeld 2017). However, there exists "[...] a distinct lack of established processes and methodologies [...]" for BDA projects as of today (Das et al. 2015, p. 2072) and there are doubts as to whether KDDM processes, such as CRISP-DM, are applicable in the big data environment (Dutta, Bose 2015, p. 294). Development of KDDM processes took place prior to the major rise of big data such that they require an update to account for the new requirements (Li et al. 2016b, p. 1). More than half of all big data projects fail according to a 2012 company survey (Kelly, Kaskade 2013). Although there exists a wide variety of reasons for project failure, it is to be expected that a solid process provides potential for improvement (Saltz et al. 2017a, pp. 183–184). This subsection studies success factors, current issues and surveys among practitioners for big data-related projects at first. This enhances the understanding of process design and builds the basis for evaluation of BDA processes of which the most relevant ones for this research are subsequently discussed.

3.2.4.1 Success factors and project issues

In line with evaluation of KDDM processes, there also exist success factors for big data projects. Saltz, Shamshurin (2016) present a comprehensive overview of success factors based on seven

different studies¹⁹. They collect 33 success factors and categorize them into six categories: data, governance, process, objectives, team, and tools (Saltz, Shamshurin 2016, pp. 2876–2877) that are listed in Table 6.

Saltz, Shamshurin (2016)		Mapping (by the author)		
Category	Success factor	Focus	Improvement area	KDDM dimensions
Team	Multidisciplinary team (i.e., across different departments)	Project	(I) Team setup	Cover full spectrum of roles involved
	Stakeholder coordination / shared understanding			Secure stakeholder support, especially senior management and data owner
	Data science, technology, business & management skills			
	People skills & ability to self-organize when needed			
Process	Flexibility and agility, with freedom for experimentation		(III) Business understanding	Formulation and use of clear business objectives
	Close collaboration between IT and business			
	Clarity of project deliverables (clear or ambiguous)			
Governance	Project management process defined		(I) Team setup	Team leader with responsibility for analytics process (not only general project management)
	Management priority / sponsorship / support			Secure stakeholder support, especially senior management and data owner
	Data protection and privacy by design			
	Big Data strategy alignment (with organization's vision)			
Objectives	Well defined scope – that is understood by the team	(III) Business understanding	Formulation and use of clear business objectives	
	Measurable project outcome		Formulation and use of clear business objectives	
	Focus on small projects and known questions			
	Specified business case			
	Feasibility study			
	Skill gap analysis			
Data	Data & data quality management / ownership	(IV) Data input	Consider outsourcing in case of lacking internal capabilities	
	Document collection/access to sources		Consider quality, complexity and volume of data	
	Representativeness of data		Identify relevant data sources and clarify access	
	Unstructured/structured data		Identify relevant data sources and clarify access	
Tools	Reporting and visualization technology	(V) Methods		
	Discovery technology			

Team	Development of skills / training	Foundations	expansion or addition
Process	Communication about the data and initiatives		
	Focus on change management		
Governance	Performance management		
	Well defined organizational structure		
	Culture of being data-driven		
Data	Data integration & security		
Tools	Investment in IT infrastructure, technology & tools		
	Investment in data sources & data storage		

Table 6 - Big data project success factors

The full list of success factors can be divided into two groups. On the one hand, there are success factors that need to be implemented in each individual project. On the other hand, some factors relate to BDA foundations that underlie all projects. Foundational factors include capability building, for example, in the form of skill training. Communication on BDA initiatives as well as change management are enablers for project initiations and practical application of project results, respectively. Furthermore, performance management ensures best allocation of BDA resources. Data integration and security represent specific technology dimensions that are typically addressed on an organizational level instead of individual projects. The success factors

¹⁹ Following studies are considered: Ahangama, Poo (2015b), Gao et al. (2015), Muller, Hart (2016), Cato et al. (2015), Brooks et al. (2015), Cosic et al. (2012), and Sicular (2012).

regarding organizational structure, culture and technology investments directly relate to the BDA foundations as described in *Section 2.5*.

Success factors for individual projects can be mapped against identified improvement areas of KDDM processes. The comparison shows that many factors directly relate to dimensions identified within the scope of KDDM processes. However, the project-related success factors also reveal additional dimensions or substantiate existing ones. The project team should not only include skills in the area of analytics (data science) but also for technology, business, and management. Furthermore, the project team should be self-organized and able to promote an agile working mode that allows for experimentation. The study of Saltz, Shamshurin (2016) also points towards the role of IT as key stakeholder in BDA projects. Data protection and privacy are additional project requirements that need to be covered by the project team. With regard to the improvement area of business understanding, the success factors indicate a focus on small projects that are specified in the form of a business case. This approach is referred to as *use cases* in this work and includes the requirement for a well-defined scope. Furthermore, project objectives need to be feasible which relates to exploration and communication of potential difficulties for implementing the project. The alignment of big data utilization with the company's strategy also needs to be addressed at project level because a BDA application should be aligned with overall objectives. The company mission statement as representation of strategic objectives can serve as guidance here. Another success factor that leads to an addition to the data input improvement area, is the use of structured and unstructured data. The combination of different data types is referred to as *data mix* in this work. Finally, utilization of analytics technology, especially knowledge discovery and visualization tools, represents a new dimension of methods for improved BDA processes.

While Saltz, Shamshurin (2016) derive their observations from a multitude of case studies, reports and literature from practice, Janssen et al. (2017) reveal success factors from one extensive big data project. Their findings confirm most success factors across all categories as discussed above, however, they also add further insights. Most interesting of all, they substantiate the project team requirements with the need of a team member that combines skills in big data, analytics, and business. Moreover, integration and standardization make the big data process less costly and thus the issue of automation is taken up. They also point towards the need of domain knowledge, at least during the analytics step (Janssen et al. 2017, pp. 341–343).

A look at current issues in big data projects complements the relevant characteristics defined by the review of success factors. Similar to KDDM processes, research is focused on which tasks need to be done instead of providing specific methods that answer the question how to do it (Saltz, Shamshurin 2015, p. 2098). Driven by the hype around big data, many companies jump into the collection of data with no clear analytics objectives defined. As a result, these companies are set up for failure and would benefit from thorough business understanding (Priebe, Markus 2015, p. 2063). Business understanding remains a key step of an analytics process in the era of big data, however, current approaches discussed in academia and practice remain focused on

data and analytics methods (Li et al. 2016a, p. 1248). Kelly, Kaskade (2013) confirm this issue as their survey identifies "'Lack of Business Context Around the Data' [...] and 'Lack of Expertise to Connect the Dots' [...] as the top reasons Big Data projects fail" (Kelly, Kaskade 2013). The latter top reason indicates a lack of capabilities in project teams. Big data projects typically require a broader set of skills in comparison with traditional analytics (Saltz et al. 2017a, pp. 186–187). Furthermore, Saltz (2015, p. 2066) points out that big data projects depend on teamwork that requires coordination to be effective. KDDM processes already include data collection as task but big data poses new challenges that require a better understanding of data input (Li et al. 2016b, p. 4). This does not only refer to challenges due to data volume or data types, for example. Deliberate selection of multiple internal and external data sources is key to benefit from big data (Barton, Court 2012, pp. 80–81).

Domain knowledge still plays a crucial role across various domains in times of big data. For instance, domain knowledge is used for big data analytics in areas as diverse as business retail (Bradlow et al. 2017), transportation (Anda et al. 2017), and medical diagnostics (Baechle et al. 2017). Current research also focuses on new methods to incorporate domain knowledge into big data analytics, for example, via visualization systems (Ruan, Zhang 2017). In view of the enormous scope of big data, manual processing of tasks can be very resource-intensive such that automation of the process remains a valid goal for BDA processes (Siriweera et al. 2015).

	(I) Project team	(II) Domain knowledge	(III) Business understanding	(IV) Data input	(V) Methods	(VI) Automation
Main objective	Enable effective project organization	Extensive involvement of domain experts and users throughout the lifecycle	Avoid provisional implementation in order to ensure thorough	Increased focus on data input	Address the "how to do" and not only "what to do"	Enable automation to avoid errors and to increase project efficiency
Dimensions	Cover full spectrum of roles involved	Include domain experts in objective formulation, data selection, data preparation and evaluation of results	Formulation and use of clear business objectives	Explicitly consider data sourcing as task (do not assume relevant data to be directly available)	Provide explanations for proposed methods	Consider dependencies on task level (integrated process)
	Provide required analytics, technology, business and management capabilities		Alignment of business and analytics objectives; overall alignment with company strategy (guidance by company mission)	Identify relevant data sources and clarify access	Provide guidance on necessary choices	Leverage technology (esp. software tools)
	Consider outsourcing in case of lacking internal capabilities		Use case approach (focus on small projects with specified business case and well-defined scope)	Consider quality, complexity and volume of data	Enable regular collaboration including domain experts	Implementation of standards
	Secure stakeholder support, especially senior management, IT and data		Feasibility study (including communication of project difficulties and limitations)	Data mix (structured & unstructured and internal & external)	Analytics tools (including visualization)	
	Team leader with business and BDA background for coordination of multidisciplinary team			Deliberate selection of data sources		
	Team leader with responsibility for analytics process (not only general)					
	Address data protection and privacy issues					
	Self-organized team with agile working mode (experimentation)					

expansion
addition

Table 7 - Improvement areas for BDA processes

Surveys among practitioners across different industries and countries reinforce identified issues. Russom (2011, p. 12) and Lavallo et al. (2010, p. 7) list lacks of management sponsorship, compelling business cases and skills as barriers to BDA adoption. Further barriers identified by Russom (2011, p. 12) mainly concern BDA foundations. Lavallo et al. (2010) describe more hurdles that relate to the project level. Firstly, ambiguity about the use of analytics and lack of

starting points relate to shortcomings of business understanding. Secondly, insufficient management capacity points towards inefficient project management which is in need of a strong team leader. Lastly, the barriers of data sourcing capabilities, data concerns and data ownership underline data input issues (Lavalle et al. 2010, p. 7). Table 7 provides an update of improvement areas that includes expansions and additions from success factors, issues and surveys related to big data. This represents the basis for evaluation of BDA processes discussed next.

3.2.4.2 Relevant BDA processes

Snail Shell Knowledge Discovery via Data Analytics (KDAA)

Li et al. (2016b) present a big data process based on the CRISP-DM lifecycle, however, the *Snail Shell Knowledge Discovery via Data Analytics (KDAA)* process is highly iterative as it does not assume a strict sequence of steps. Moreover, it adds two additional steps to the overall process. *Problem formulation* represents the starting point of the process that aims for formulation of clear business objectives against the background of increasing complexity and decreasing structure in analytics problems. On the other hand, *maintenance* complements the typical deployment step in order to address new requirements within the big data environment (Li et al. 2016b, p. 3). Figure 23 provides an overview of the KDAA process.

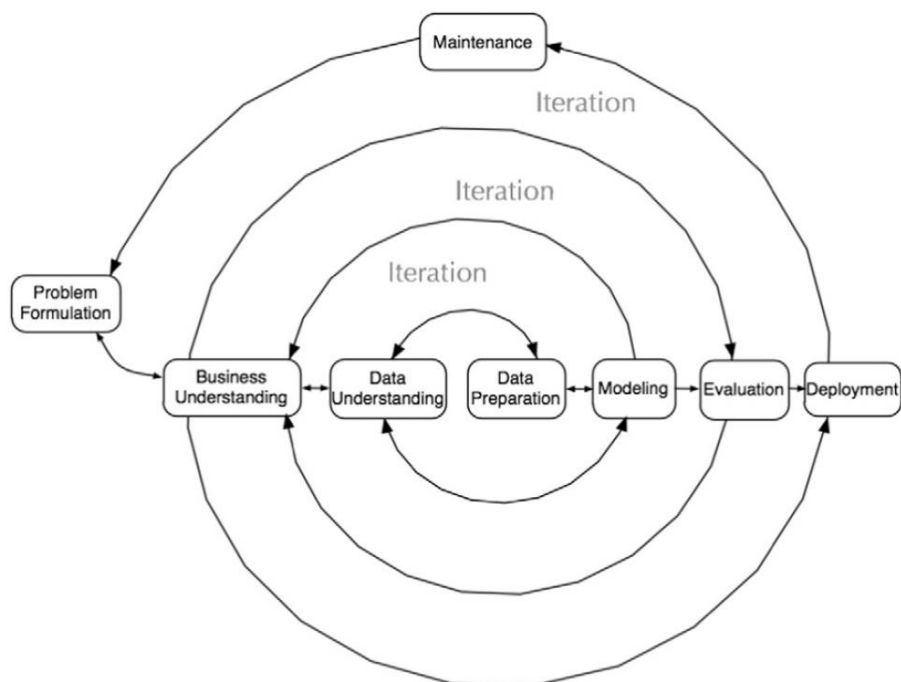


Figure 23: *Snail Shell Knowledge Discovery via Data Analytics* (Li et al. 2016b, p. 3)

Li et al. (2016b) describe various advances of KDDA compared to traditional KDDM processes. Problem formulation identifies business problems and translates them into analytics problems. KDAA proposes to deploy problem formulation strategies, such as modeling or decomposition, in order to "[...] focus on controllable components of a decision situation [...]"

(Li et al. 2016b, p. 4). In addition, the process points towards general methods of problem formulation including *Value Focused Thinking (VFT)*, *Goal Question Metrics (GQM)*, and *SMART (Specific, Measurable, Achievable, Relevant, Time-bounded)* criteria to guide determination of measurable business objectives. Furthermore, a strict definition of the business problem is required to answer key "[...] what, why, and how questions" (Li et al. 2016b, p. 4). KDAA also advances the business understanding step in various ways. It builds a comprehensive business case and seeks for alignment with senior management on this basis. The new task of enterprise knowledge acquisition systematically collects explicit domain knowledge, for example, from business process documentation, and implicit knowledge from experts. Furthermore, a software selection framework is proposed in order to select adequate tools for the business case. Business understanding of KDAA also includes an assessment of BDA foundations that is not typical for other processes. For data understanding, KDDA proposes the use of visualization methods and inclusion of business requirements in addition to analytics requirements. These requirements are also to be included during data preparation. In the modeling step, a BDA knowledge repository provides guidelines for building and assessing different types of models. KDDA complements evaluation with a field test under real-world conditions and a review of results with stakeholders including senior management. Data preparation and deployment steps are generally very similar to KDDM processes (Li et al. 2016b, pp. 3–7).

Other BDA processes also build on CRISP-DM in a variety of domains, but do not provide the same level of detail and advancement. For instance, Kalgotra et al. (2016) adapt data preparation of CRISP-DM for handling of streaming data in medical diagnosis. Heit et al. (2016) introduce an advancement with focus on the deployment step. Niño et al. (2015) discuss the business understanding step in case of a big data process at a manufacturing company. Also other KDDM processes are discussed in regard to the big data era, for example SEMMA (Dean 2014; Woodside 2016) or data mining processes (Chen et al. 2015). In general, they do not cover the full scope of CRISP-DM and do not provide substantial advancement towards the process design.

Agile BI Delivery Framework (ADF)

Larson, Chang (2016) take up agile principles for business intelligence and data science. Their *Agile BI Delivery Framework (ADF)* comprises *BI Delivery* and *Fast Analytics/Data Science* processes as shown in Figure 24. While BI Delivery represents the process to develop and implement BI solutions based on structured data, Fast Analytics/Data Science provides an iterative process for developing BDA models (Larson, Chang 2016, pp. 702–705).

According to Larson, Chang (2016), fast analytics refers to visual analytics and data science represents data mining based on big data. They define the *scope* of an analytics problem by the scope of data sources that can include structured and unstructured data. Furthermore, the BI program management provides problems to solve including objectives, restrictions, and expectations. The idea of *data acquisition/discovery* is to create a big data repository without

RELATED WORK

necessarily understanding the meaning of all data a priori. Potential value and use of data is then discovered with the help of visualization. The *analyze/visualize* step represents exploratory data analysis that results in the definition of data input for modeling or in a dashboard to be used as BI tool. *Model/design/develop, validate, and deployment* represent the steps to build analytics models, to optimize model performance, and the transfer to practical use. In doing so, ADF proposes to consider new data sources for optimization (Larson, Chang 2016, pp. 706–707). Fast Analytics/Data Science processes "[...] are inherently agile as each follows iterations, use small teams, and require collaboration between business subject matter experts and technical resources" based on co-located resources (Larson, Chang 2016, p. 707). ADF furthermore discusses valuable agile methods that can be used in the process (Larson, Chang 2016, pp. 707–708).

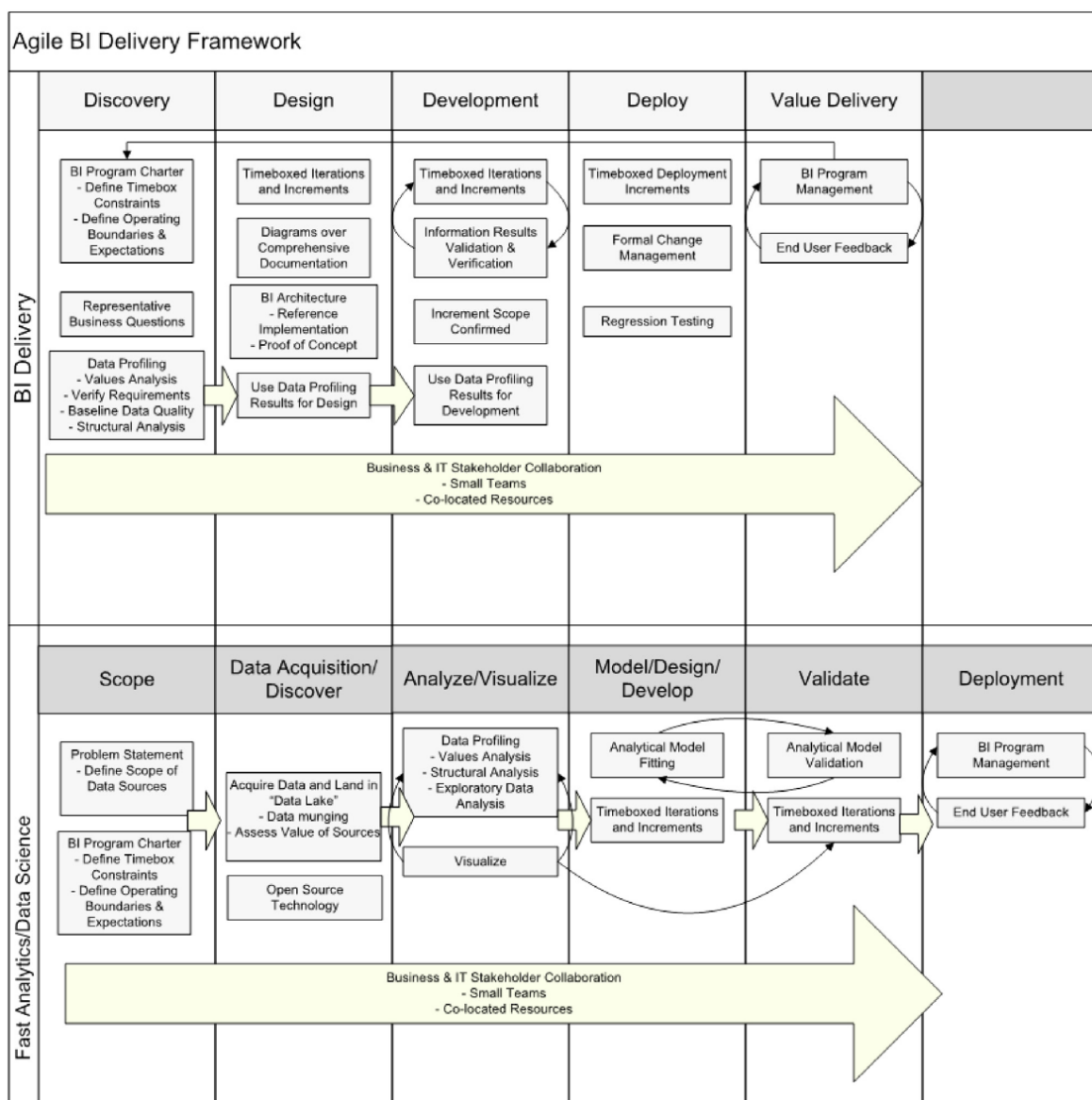


Figure 24 - Agile BI Delivery Framework (Larson, Chang 2016, p. 702)

Further research with regard to agile principles in the domain of big data is still limited (Larson, Chang 2016, p. 704). For instance, Journey (2014) provides agile methods for big data and explains how to use them and Earley (2014) offers guiding principles for agile analytics in the era of big data. Frankova et al. (2016) discuss management of big data projects through the lens of an agile approach and Chen et al. (2016) focus on the underlying technology architecture.

Data Science Edge (DSE)

Grady (2016) discusses the use of big data for *Knowledge Discovery in Data Science (KDD)* and proposes a new process called *Data Science Edge (DSE)*. The process is organized in five steps that represent different levels of data maturity and include plan, collect, curate, analyze, and act. Figure 25 provides an overview of DSE.

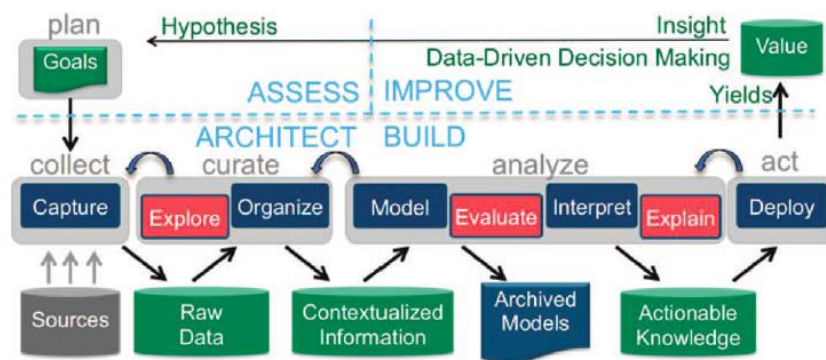


Figure 25 - Data Science Edge (Grady 2016, p. 1605)

According to Grady (2016), DSE reflects the steps of CRISP-DM with rearrangement of tasks and addition of new tasks that are explained in the following. In the *plan* step, definition of organization boundaries considers the role of external providers and justification of big data technology investments incorporates the idea of a business case. DSE also requires to take regulatory constraints for data usage into account. New tasks in the *collect* step include selection of an appropriate database technology, a data distribution strategy for high volume data, and inclusion of external data sources. The *curate* step involves visualization for data exploration, provision of metadata derived from domain expertise to ensure proper use of data, and special care regarding data privacy. Furthermore, a decision on a single data repository versus multiple distributed repositories as well as a strategy for handling data quality issues and data sampling are required. DSE includes several tasks related to technical implementation of analytics in the *analyze* step, for example, parallel execution of algorithms. This step also proposes three guidelines regarding conception of analytics: model results can be based on correlation rather than causation, hypotheses help to direct analytics, and simpler analytics approaches should be identified in case the intended approach is technically not feasible. Finally, the *act* step includes visualization as method for explaining results and activities for protecting the big data system (Grady 2016, pp. 1605–1607).

There exist other works based on the idea of KDDs. Schutt, O'Neil (2013) provide a high-level description of a data science process and they briefly describe the role of the data scientist. Guo (2013) introduces a process including all tasks relevant for the role of a data scientist. Priebe, Markus (2015) build upon both previously mentioned works and show that they do not adequately address the critical steps of business understanding and deployment. They use CRISP-DM in order to create an end-to-end process and propose a business information model containing metadata as method to support identification and sourcing of relevant data (Priebe, Markus 2015, pp. 2062–2064). Saltz, Shamshurin (2015) exemplarily describe a data science process from practice as observed in an advertising company.

Data Analytics Lifecycle (DAL)

EMC Education Services (2015) propose a six-step process that builds on CRISP-DM and other established approaches in order to address BDA challenges. The *Data Analytics Lifecycle (DAL)*, as shown in Figure 26, does not only describe the process but also includes a definition of key roles that are required to successfully perform the project. DAL introduces seven specific roles and defines general responsibilities as well as key project outcomes for each. Furthermore, the process relies on involvement of stakeholders who need to be identified at the outset of the project as well. In the initial *discovery* step, the team determines the extent of expert knowledge required for building analytics models in the underlying domain. This ensures the right balance of expertise within the team. Furthermore, required resources are assessed and gaps identified. This includes technological BDA resources such as systems and analytics tools, however, required skills in the project team and available data regarding the project objective are also considered here. With required resources secured, the next task is to frame the analytics problem. DAL proposes to state the problem in business terms and to include relevant stakeholders, especially the project sponsor, in this task. Stakeholder involvement helps to clearly define the objective and success criteria as well as to manage expectations. The process includes a comprehensive guideline for interviewing the project sponsor including a set of useful questions. Development of initial hypotheses by the project team, stakeholders or domain experts is another task in the discovery step and serves as basis for later analytics. The initial hypotheses also guide the identification of potential data sources which represents the last task of the initial process step. DAL provides a set of activities that are helpful to identify and explore potential data sources as well as additionally proposes validation of identified sources with domain experts. The second step, *data preparation*, starts with the set-up of a project-specific 'data warehouse' that provides an environment to explore and analyze structured and unstructured data. This so-called analytic sandbox should be designed in cooperation with IT and requires data sourcing as independent task to include data from internal and external sources. In order to manage the large extent of data, DAL introduces a dataset inventory as tool to track availability, accessibility and sourcing status of different datasets. Furthermore, data preparation represents typical preprocessing tasks which are extended by the use of data visualization in order to support a better understanding of data including identification of data quality issues.

DAL also includes a short list of common software tools that are useful for data preparation. *Model planning* as third step substantiates prior data exploration with focus on selecting relevant data and appropriate models for the underlying problem. DAL provides a brief overview on common tools that can be used during model planning. The subsequent *model building* step comprises all tasks required to build selected models. DAL points out that domain experts should be involved in evaluation of model results and lists common analytics tools for this step of the process. *Communicate results* and *operationalize* conclude the process by evaluation and implementation of analytics results. DAL highlights the key role of stakeholders also for evaluation of results because they are ultimately responsible for practical application (EMC Education Services 2015, pp. 26–53).

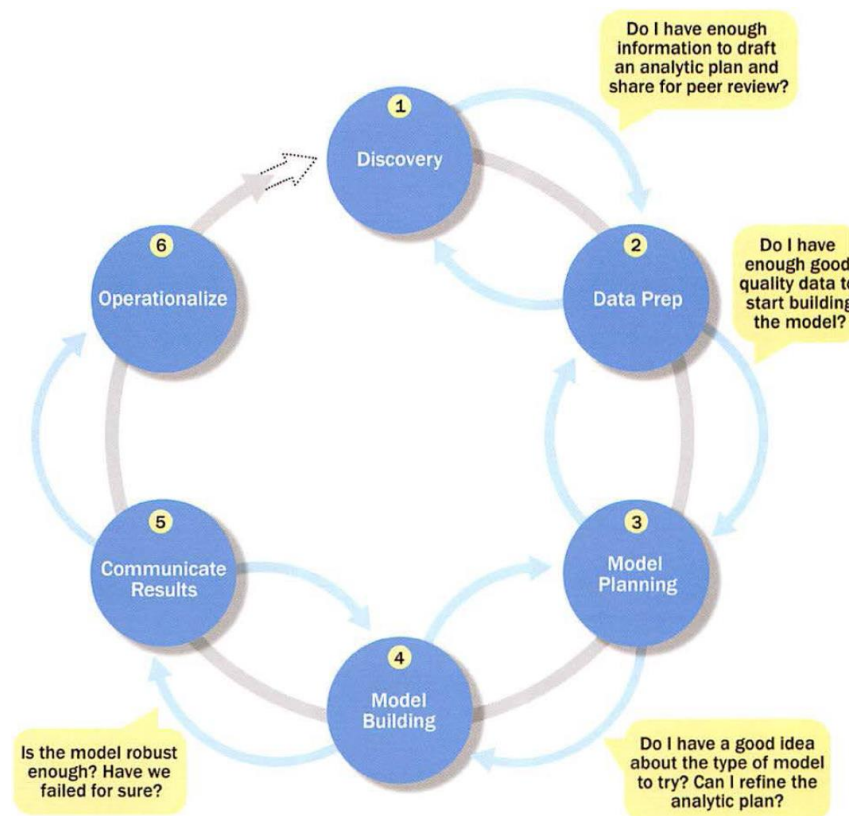


Figure 26: Data Analytics Lifecycle (EMC Education Services 2015, p. 29)

Schmarzo (2013, pp. 40–49) generally describes the same process as DAL but puts a stronger focus on project team roles, thereby formulates key responsibilities and provides a detailed process description focused on the data scientist role.

Big Data Analytics Methodology (BDAM)

Raghupathi, Raghupathi (2014) outline a process with strong practical character. Their *Big Data Analytics Methodology (BDAM)* organizes a list of more than 20 tasks into four steps and points towards the need for an interdisciplinary project team. The initial *concept design* step clarifies the motivation for a big data project including its importance for the organization. In the subsequent *proposal development* step, the concept is detailed out with regard to the problem addressed, the

motivation for the organization, and the chosen BDA approach. The proposal also requires background information on the underlying domain and related prior work in this domain (Raghupathi, Raghupathi 2014, pp. 60–62). BDAM proposes to evaluate the concept and proposal along four major dimensions (*4Cs*) (Raghupathi, Raghupathi 2014, p. 62):

- 1) *Completeness*: The concept design needs to be complete.
- 2) *Correctness*: The concept design needs to be technically feasible and to use correct terminology.
- 3) *Consistency*: The proposal needs to be consistent and to allow for continuity.
- 4) *Communicability*: The proposal requires professional preparation including use of simple understandable language.

Raghupathi, Raghupathi (2014) describe the next step as *implementation* of the BDA methodology as described by the concept and proposal. It is for the most part similar to traditional analytics but BDAM highlights evaluation and selection of appropriate analytics tools as key task. Furthermore, the process lists identification of data sources and data sourcing as individual tasks (Raghupathi, Raghupathi 2014, p. 62). In general, Raghupathi, Raghupathi (2014, pp. 53–54) assume structured as well as unstructured data from internal and external sources as basis for their process. The final step of *implementation* rests on an evaluation of analytics results including stakeholders and BDAM proposes a set of evaluation criteria, for example, robustness of results (Raghupathi, Raghupathi 2014, p. 63). Moreover, the process recommends to involve users during implementation (Raghupathi, Raghupathi 2014, p. 63).

Big Data Analytics Lifecycle (BDAL)

Erl et al. (2016) present the *Big Data Analytics Lifecycle (BDAL)* in order to address specific needs of analytics projects based on big data. As shown in Figure 27, their process consists of nine steps. (1) *Business case evaluation* clarifies the addressed business problem, required business resources as well as budget and objectives. BDAL refers to the SMART technique in order to provide measurable objectives that allow for evaluation results at a later stage. (2) *Data identification* rests on the idea to combine datasets from a variety of internal and external sources. (3) *Data acquisition & filtering* includes data sourcing as independent task and filtering represents the removal of low quality and irrelevant data. Furthermore, technical metadata is retrieved from data sources in order to maintain data provenance information. (4) *Data extraction* includes the extraction of information, for example, specific characteristics of text data, and transformation into a format appropriate for intended analytics. BDAL proposes to use redundant data to identify invalid data or to handle missing values and to use in-memory processing to perform (5) *data validation & cleansing* in case of real-time analytics. (6) *Data aggregation & representation* aim for integrating various datasets from different sources into a single data repository for analytics. According to BDAL, (7) *data analysis* can be used to confirm hypotheses or to discover previously unknown patterns. (8) *Data visualization* and simple statistics are used for

communication and interpretation of results with the help of business users. Finally, (9) *utilization of analysis results* determines how to use insights in the organization (Erl et al. 2016, pp. 55–70).

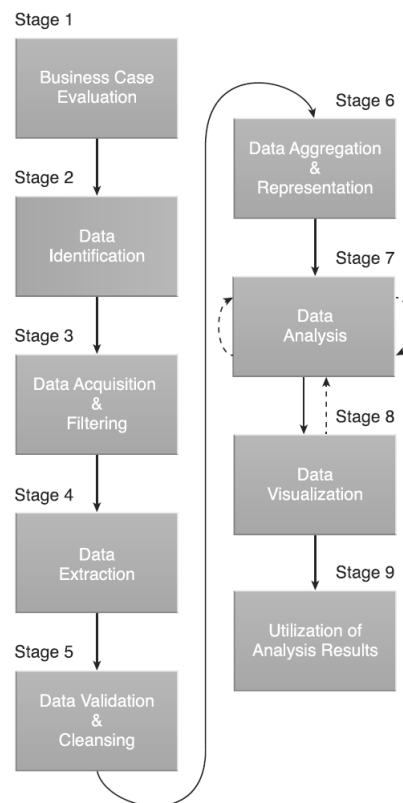


Figure 27: Big Data Analytics Lifecycle (Erl et al. 2016, p. 55)

Big – Data, Analytics, and Decisions Framework (B-DAD)

Elgendy, Elragal (2016) introduce the *Big – Data, Analytics, and Decisions Framework (B-DAD)* that maps "[...] big data tools, architectures, and analytics to the different decision making phases" (Elgendy, Elragal 2016, p. 1071). Figure 28 provides an overview of the framework with its four hierarchical phases. A major limitation of B-DAD exists, because "[...] the framework assumes that the decision domain is already known, and does not need to be explored first in order to extract a problem which needs to be solved, or a question which needs to be answered" (Elgendy, Elragal 2016, p. 1073). According to Elgendy, Elragal (2016), B-DAD mainly differs from KDDM processes in the *intelligence phase* due to consideration of structured as well as unstructured data from internal and external sources. After identification of data sources, data is sourced and stored before organization of data follows which is similar to preprocessing and transformation tasks of the KDD process. The framework lists many tools for storage, management and processing of big data. The *design phase* comprises model planning for selection of appropriate analytics models and data analytics for building these models. B-DAD lists potentially useful analytics tools also for this phase. The *choice phase* starts with evaluation of analytics results in order to identify and prioritize potentially useful results with the help of evaluation techniques such as simulation of potential solutions or what-if scenarios. In addition,

RELATED WORK

B-DAD lists useful tools for visualization of big data. The phase concludes with the final decision on the optimal course of action. *Implementation* is the final phase and adopts results monitoring and feedback as means for operationalizing results (Elgendy, Elragal 2016, pp. 1073–1075).

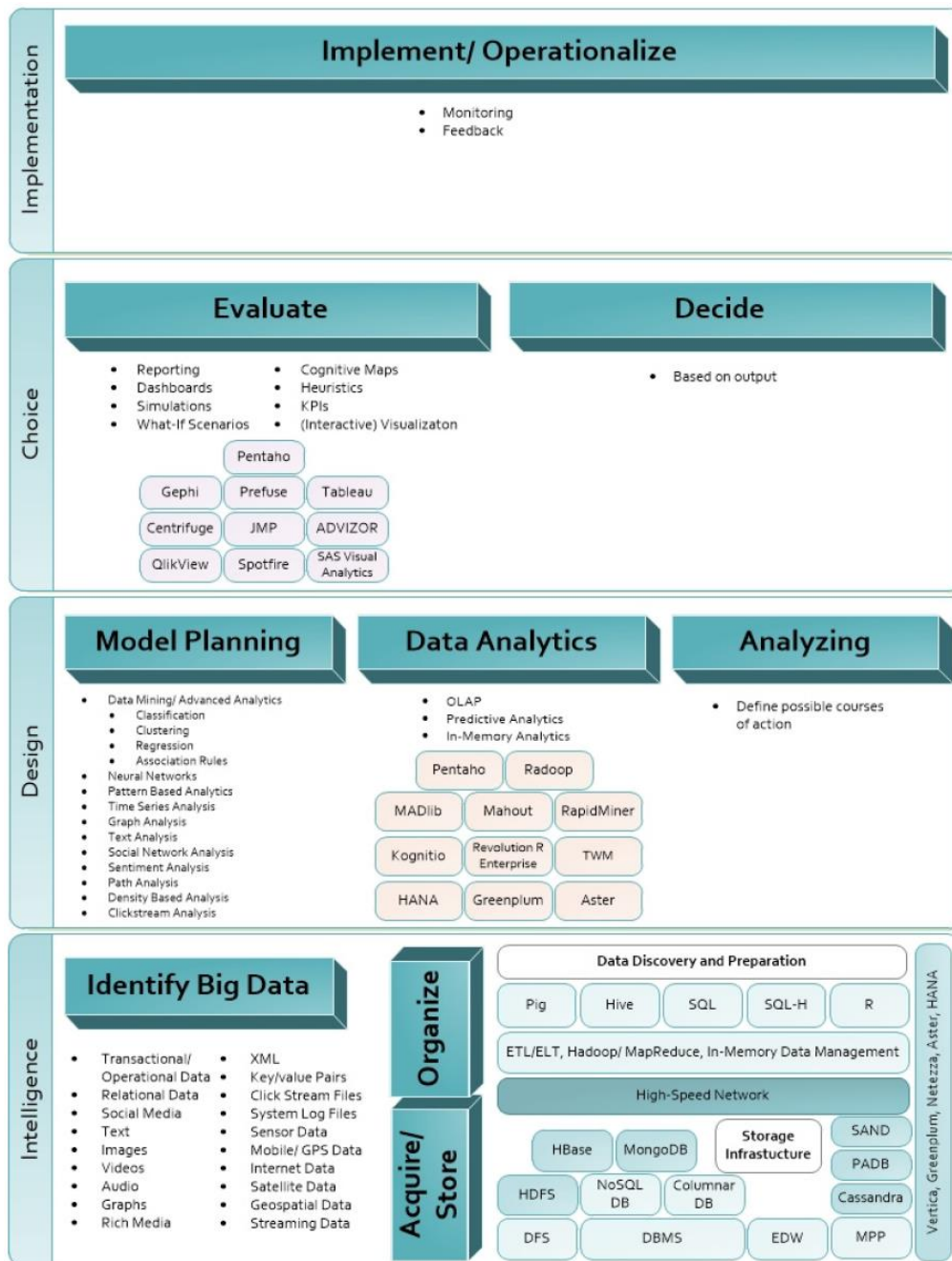


Figure 28: Big – Data, Analytics, and Decisions Framework (Elgendy, Elragal 2016, p. 1083)

Framework for Implementation of Big Data Projects (FIBD)

Dutta, Bose (2015) present a *Framework for Implementation of Big Data Projects (FIBD)* based on ten steps organized in three phases as shown in Figure 29. *Strategic groundwork* starts with the understanding of the *business problem* that should be addressed by big data analytics. FIBD

emphasizes the value of including senior management and stakeholders for scoping the problem and for setting adequate expectations for the project. The *research* step builds an understanding of existing solutions to the problem by other organizations and for available technology such as analytics tools. Furthermore, the groundwork requires *cross functional team formation* that includes business user, BDA experts, and IT experts next to the stakeholders and senior management. Business users are especially required to provide input for model building. The first phase of FIBD concludes with a *project roadmap* that defines project implementation according to the established groundwork. *Data collection & examination* initiates the *data analytics* phase. FIBD proposes to collect structured and unstructured data from internal as well as external sources. Furthermore, data examination is required to integrate data of various types. *Data analysis & modeling* represents the analytics core of the process that is augmented by *data visualization* in order to support the generation of insights from the data. *Insight generation* represents the transfer of analytics results to actionable insights from a business perspective. The *final implementation* phase describes deployment of results and puts a focus on the *integration with the existing IT system* as well as *training of people*. While the former addresses technical issues, the latter mainly aims for increasing the acceptance by users (Dutta, Bose 2015, pp. 294–296). The major advancement of FIBD is the combination of "[...] change management aspects of an IT project management framework with the data management aspects of an analytics framework" (Dutta, Bose 2015, p. 296).

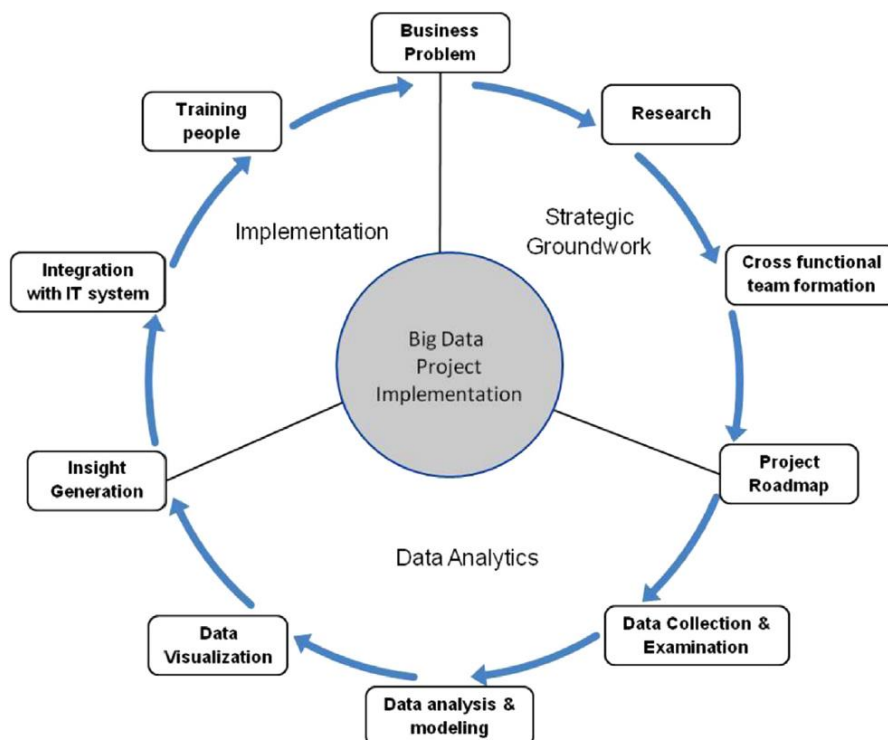


Figure 29: Framework for Implementation of Big Data Projects (Dutta, Bose 2015, p. 295)

Doing a Big Data Project (DBDP)

Feinleib (2014) describes how to perform a big data project under the term *Doing a Big Data Project (DBDP)* in two respects: general guidelines to set up a big data project and a five-step workflow for the analytics work. The first guideline asks for a definition of the desired outcome which needs to be measurable in business terms and to be as specific as possible. DBDP also requires to rigorously measure the business value in order to monitor performance of big data solutions over time. Identification of questions to be answered by the project is closely related to the previous two guidelines. Such a list of questions helps to direct the use of big data for intended purposes and reflects the explorative character of the process. Another guideline is the creation of data policies that set rules regarding data usage and security. Furthermore, DBDP demands the identification of required resources including human resources, data sources, and analytics tools. Relevant data for the project can include structured and unstructured data from both internal and external sources. Finally, DBDP proposes visualization as a mean for better understanding of results (Feinleib 2014, pp. 103–111).

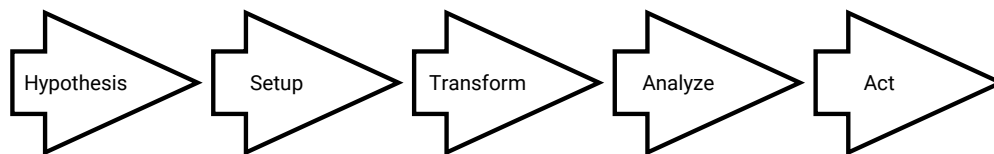


Figure 30: Big data workflow (Feinleib 2014, p. 112)

Figure 30 illustrates the big data workflow of DBDP as described by Feinleib (2014). The *creation of a hypothesis* is the starting point and defines the analytics to be performed. *Setup of the systems* describes the identification of relevant data sources and integration of the data in a project data warehouse or in the cloud as basis to manage big data input. *Transformation* prepares data for the subsequent task to *analyze the data* and data visualization is proposed as one tool. *Act on the data* uses analytics results to propose adequate actions for the company (Feinleib 2014, pp. 112–117).

Other BDA processes

Further processes exist to meet BDA challenges. In general, they do not provide the same level of advancement, structure or detail as previously discussed approaches such that they are only briefly discussed here. Berman (2013) describes a nine-step process that starts with formulation of the problem to be solved by analytics without preparing a deeper understanding of the underlying domain. The process furthermore ends at results evaluation and generally puts a focus on the motivation as well as exemplary description of each step (Berman 2013, pp. 157–165). Another analytics-based approach comes from Ridge (2015). *Guerilla Analytics* addresses challenges from dynamic project environments driven by changes in data, requirements or resources (Ridge 2015, pp. 9–10). Except for the business understanding step, it follows the same lifecycle as CRISP-DM. However, it is adapted to take care of the dynamic changes (Ridge 2015, pp. 16–18). *Guerilla Analytics* is strongly data-driven with focus on technical implementation of the analytics project, whereby a large number of practice tips are provided

(Ridge 2015). Fogelman-Soulié, Lu (2016) introduce another process for big data projects with strong focus on the analytics work. Ankam (2016), Davenport (2013), and Fisher et al. (2012) further provide brief discussions on processes for big data analytics projects with strong focus on the lifecycle. Franks (2014, pp. 177–179) underlines that lifecycles of different BDA processes are not fundamentally different and that they are still similar to traditional processes like CRISP-DM.

Practitioner Approaches

Practitioner approaches to the BDA process are often linked to products such as big data or analytics platforms. For example, Severtson et al. (2017) describe a process similar to CRISP-DM in relation to the cloud-based platform Microsoft Azure (Microsoft 2017). Other practitioners provide concise process descriptions that cover the same scope as CRISP-DM. Henke et al. (2016) propose to start the process with clear formulation of the issue addressed, envisioned business effect of the project, and consideration of BDA applications in the form of use cases. They also underline the value of integrating data from internal and external sources. Furthermore, deployment includes process redesign and change management in order to ensure integration into business operations and adoption by the organization (Henke et al. 2016, pp. 34–35). Almquist et al. (2015) reinforce these aspects as they claim to focus on "[...] a small number of high-value business problems [...]" (Almquist et al. 2015, p. 2) as a critical factor as well as addition of external data sources to be beneficial. Moreover, they propose to relate hypotheses to data sources in order to identify relevant data, to include multiple stakeholders in data understanding tasks, and to reduce adverse effects from the black box phenomenon. Finally, they also list adoption activities such as capability building for deployment (Almquist et al. 2015, p. 2). Hagen et al. (2013, pp. 14–16) further underscore the focus on high impact use cases, outlining of future states after deployment, and comprehensive description of objectives as key elements at the beginning of a BDA process. In general, these practitioner approaches provide a lower level of detail compared to previously discussed processes.

3.2.4.3 Evaluation of BDA processes

The previous subsection discusses nine processes for big data analytics because they represent the most detailed and advanced approaches among all identified processes. In order to evaluate these processes, their key features are mapped against the improvement areas for BDA processes (compare Table 7). Each process is evaluated for its level of advancement regarding the identified dimensions across all improvement areas. The result of this evaluation is summarized in Table 8Table 8.

RELATED WORK

	(I) Project team	(II) Domain knowledge	(III) Business understanding	(IV) Data input	(V) Methods	(VI) Automation
KDDA - Li et al. (2016b)						
Key features	- Business case and results review with stakeholders including senior management	- Enterprise knowledge acquisition (explicit from documentation and implicit from experts) - Business requirements for data understanding and preparation	- Use of problem formulation strategy - Business problem definition (including costs, requirements, constraints, and resources)	- Business and analytics requirements for data quality (guideline)	- General methods for problem formulation (proposals) - Software selection framework (not specified) - Data visualization (proposed tools) - BDA knowledge repository for modeling (not specified)	-
Evaluation	low	medium	high	low	medium	not
ADF - Larson, Chang (2016)						
Key features	- IT stakeholder collaboration - Small teams - Co-located resources	- Business stakeholder collaboration	- BI program management including problems, objectives, feasibility (externally given)	- Data sources define scope - Inclusion of structured and unstructured data - Accessible data repository - New data sources for model optimization	- Data visualization - Agile methods	-
Evaluation	medium	low	medium	medium	low	not
DSE - Grady (2016)						
Key features	- Definition of organizational boundaries - Consideration of regulatory constraints and data privacy issues	- Metadata from domain expertise	- Business case (focus on technology investments)	- Selection of appropriate database technology - Definition of data distribution strategy - Integration of external sources - Decision on single versus multiple data repositories - Strategy for data sampling and quality issues	- Data visualization (exploration and explanation) - Correlation-based analytics, hypothesis-led analytics, simple approaches (only guidelines for analytics work)	-
Evaluation	low	low	low	medium	low	not
DAL - EMC Education Services (2015)						
Key features	- Definition of 7 specific roles including general responsibilities and key project outcomes - Identification of capability gaps - Identification of stakeholders - Cooperation with IT	- Determination of required domain knowledge - Development of initial hypotheses as basis for analytics - Validation of potential data sources - Evaluation of model results	- Definition of required resources - Formulation of problem in business terms - Involvement of stakeholder and project sponsor in problem formulation (success criteria, expectations)	- Identification of potential data sources - Analytic sandbox for structured and unstructured data - Data sourcing from internal and external sources	- Guideline for project sponsor interviews - Set of useful activities for identification of potential data sources - Dataset inventory - Data visualization for data understanding and exploration - List of software tools for data preparation (brief overview) - List of software tools for model planning (brief overview) - List of analytics tools (brief overview)	- Proposal to use existing tools (software)
Evaluation	medium	high	medium	high	medium	low
BDAM - Raghupathi, Raghupathi (2014)						
Key features	- Need for interdisciplinary team - Evaluation of analytics results including stakeholders	- User involvement during implementation	- Concept and proposal as basis for analytics work - Evaluation framework (4Cs)	- Identification of data sources and data sourcing as individual tasks - Structured/unstructured data from internal/external sources	- Evaluation and selection of appropriate analytics tools - Set of evaluation criteria	-
Evaluation	low	low	medium	medium	low	not
BDAL - Eri et al. (2016)						
Key features	-	- Evaluation with business users	- Business case approach	- Variety of internal and external data sources - Independent data sourcing - Filtering for quality and relevance of data (no details) - Technical metadata for data provenance - Single data repository for analytics	- SMART technique for objectives - Use of data redundancy to handle quality issues - Data visualization and simple statistics for communication and interpretation of results	- In-memory processing for real-time data validation and cleaning
Evaluation	not	low	low	high	medium	low
B-DAD - Elgendy, Elragal (2016)						
Key features	-	-	Assumes domain including problem already known	- Structured/unstructured data from internal/external sources - Data sourcing as individual task	- List of tools for storage, management and processing of big data - List of analytics tools - List of big data visualization tools - List of techniques for analytics evaluation	- Proposal to use existing tools (software)
Evaluation	not	not	not	medium	medium	low
FIBD - Dutta, Bose (2015)						
Key features	- Involvement of senior management and stakeholders - Cross functional team	- Business user input for model building	- Setting adequate expectations for the project - Consider existing solutions to the problem	- Structured/unstructured data from internal/external sources	- Scan for available analytics tools - Data visualization	-
Evaluation	low	low	low	low	low	not
DBDP - Feinleib (2014)						
Key features	- Policies on data usage and security	- Development of initial hypotheses as basis for analytics	- Formulation of specific objectives measurable in business terms	- Structured/unstructured data from internal/external sources - Data warehouse or cloud to manage big data	- Use of analytics tools as resource - Visualization of data and results	-
Evaluation	low	low	low	medium	low	not
Evaluation score						
	not addressed	low advancement level	medium advancement level	high advancement level		

Table 8 - Evaluation of BDA processes

The following provides the key insights from the evaluation organized by improvement area:

- I. *Project team*: The design of an effective project organization is not in focus for most of the processes. ADF and DAL are two exceptions because the former rests on agile principles with a natural focus on effective development processes and the latter is the only approach including a concrete description of team roles. However, all dimensions of the project team improvement area are addressed across the totality of processes except for background and responsibility of the team leader.
- II. *Domain knowledge*: The majority of processes explicitly includes knowledge from experts only for specific steps or tasks. KDDA proposes a systematic approach to acquire relevant knowledge, however, a specific use of this knowledge is only considered in the form of business requirements during data understanding and preparation. DAL provides a higher level of advancement because the process identifies required knowledge and utilizes such expertise for key tasks. However, none of the processes consistently involves domain experts or users throughout the entire lifecycle.
- III. *Business understanding*: Considering the review of the business case with senior management as form of strategic alignment, KDDA addresses all dimensions for improved business understanding. ADF, DAL, and BDAM provide a medium level of advancement as they build on a predefined framework or lack selected dimensions. Despite its overall low level of advancement, FIBD suggests research on existing solutions by other companies to the identified business problem as valuable addition to the improvement area. None of the processes uses the company mission for strategic alignment or considers multiple alternative use cases.
- IV. *Data input*: The focus on data input is clearly increased across all processes and the level of advancement is medium or high except for KDDA and FIBD. It is remarkable that nearly all processes consider structured and unstructured data as input while deliberate selection of data sources is still neglected in most cases. Five of the processes concerned – ADF, DSE, DAL, BDAL, and DBDP – include the setup of a data repository for the project which is consequently considered as additional dimension for the data input improvement area.
- V. *Methods*: This area is dominated by the proposal to use analytics tools, especially in the form of data visualization. Apart from this, proposed tools and techniques cover selected tasks but are not consistently offered throughout the lifecycle. The question on how to implement the process therefore remains unanswered for a significant part such that advancement is restricted to a medium level here.

- VI. *Automation*: In contrast to data input, automation is not in focus of the processes. It is basically addressed by leveraging technology in form of software tools but process integration or standards are not explicitly discussed. One potential explanation is a stronger emphasis on experimentation in development of BDA applications. Integration and standards rather support a stringent process and should consequently only be considered where useful.

In summary, the processes provide a large variety of advancements across all improvement areas. This underlines the importance of previously identified dimensions. However, no individual process shows strong advancement considering the entire range of improvement potential. KDDA and BDAL demonstrate high advancement levels for selected areas but also low levels for others. DAL is the only process with medium advancement for all areas except for automation and also provides a relatively comprehensive documentation. Although many processes relate to CRISP-DM, the level of detail is clearly below the documentation of the de facto standard KDDM process. Moreover, all processes are applicable in the business domain but lack extensive validation in practice. This can also be seen by the fact that all processes were published within the previous three years of this work. In comparison to previously identified improvement dimensions, the evaluation of BDA processes reveals additional insights for the process design. Utilization of existing BDA solutions, setup of a project data repository, and the subordinate role of automation due to the experimental nature of developing BDA applications represent new design dimensions. They are consequently considered as substantiation of the improvement areas so far.

3.2.5 Interim conclusion

Basic KDDM and data mining processes represent the origin of formalized approaches to analytics work in the digital age. The comparison of these processes (see *Subsection 3.2.2.3*) reveals a high level of consistency regarding their scope. CRISP-DM stands out due to several advantages including universal applicability, comprehensive documentation, high usability, and validation in practice. Although being termed as the de facto standard KDDM process, CRISP-DM also shows shortcomings that, at the same time, represent critical success factors for KDDM projects. These shortcomings and success factors can be grouped into six improvement areas (see *Subsection 3.2.2.4*) and serve as evaluation basis for advanced KDDM processes (see *Subsection 3.2.3.6*). These advanced processes include derivatives of CRISP-DM, adaptations from engineering processes, and approaches with focus on a specific KDDM dimension. They address the identified improvement areas only to a limited extent while still not matching the major advantages of CRISP-DM. Moreover, analytics in the era of big data come with specific challenges that translate into new issues and success factors. The dimensions of the process improvement areas are therefore updated (see *Subsection 3.2.4.1*) and used for evaluating BDA processes (see *Subsection 3.2.4.3*). The evaluation shows a large variety of advancements and reveals strong consistency of identified improvement dimensions, and yet no single process

substantially covers all improvement areas. Furthermore, the aforementioned advantages of CRISP-DM still remain valid in comparison with these processes.

This presented review is based on a detailed discussion and evaluation of the 25 processes identified as being the most relevant. The extensive review of KDDM and BDA processes provides a solid understanding of process steps, tasks, methods and lifecycles that are required as *applicable knowledge* for the design of the new methodology. Furthermore, the review also reveals most advantageous processes as candidates for the basic design of the new methodology. In particular, CRISP-DM and DAL represent the most promising candidates from KDDM and BDA processes, respectively.

In general, all KDDM and BDA processes are designed for general applicability with some exceptions, for example, SEMMA for use in a specific software or DFD for mission critical applications. None of the presented processes directly focuses on BDA applications against the background of challenges from the volatile business environment, especially in the form of improved sales forecasting. Substantiation of these processes represents a *scientific need* in relation with *research questions 1 & 2*. As a consequence, *Chapter 4* introduces a new methodology that incorporates the applicable knowledge and addresses the business need to develop BDA applications in the volatile world with focus on sales forecasting. Moreover, the identified improvement areas represent a key result from the knowledge base review. Realization of the improvement dimensions is another *scientific need* to be addressed by the new methodology design and they are therefore considered as design requirements. Table 9 summarizes the final design requirements including updates from the evaluation of BDA processes.

	(I) Project team	(II) Domain knowledge	(III) Business understanding	(IV) Data input	(V) Methods	(VI) Automation
Main objective	Enable effective project organization	Extensive involvement of domain experts and users throughout the lifecycle	Avoid provisional implementation in order to ensure thorough understanding	Increased focus on data input	Address the "how to do" and not only "what to do"	Enable automation to avoid errors and to increase project efficiency (<i>without major restrictions to</i>
Dimensions	Cover full spectrum of roles involved	Include domain experts in objective formulation, data selection, data preparation and evaluation of results	Formulation and use of clear business objectives	Explicitly consider data sourcing as task (do not assume relevant data to be directly available)	Provide explanations for proposed methods	Leverage technology (esp. software tools)
	Provide required analytics, technology, business and management capabilities		Alignment of business and analytics objectives; overall alignment with company strategy (guidance by company mission)	Identify relevant data sources and clarify access	Provide guidance on necessary choices	Consider dependencies on task level (integrated process) - <i>only where useful</i> -
	Consider outsourcing in case of lacking internal capabilities		Use case approach (focus on small projects with specified business case and well-defined scope)	Consider quality, complexity and volume of data	Enable regular collaboration including domain experts	Implementation of standards - <i>only where useful</i> -
	Secure stakeholder support, especially senior management, IT and data		Feasibility study (including communication of project difficulties and limitations)	Data mix (structured & unstructured and internal & external)	Analytics tools (including visualization)	
	Team leader with business and BDA background for coordination of multidisciplinary team		Consider existing BDA solutions to the problem (esp. from other companies)	Deliberate selection of data sources		
	Team leader with responsibility for analytics process (not only general			Project-specific data repository		
	Address data protection and privacy issues					
	Self-organized team with agile working mode (experimentation)					

adjustment
addition

Table 9 - Design requirements for new methodology

4 New methodology

4.1 Concept

4.1.1 Design considerations

The previous chapter has shown that research on BDA applications, especially for sales forecasting, is limited for industrial companies as of today. Moreover, existing processes to develop BDA applications have shortcomings in various dimensions and do not directly address utilization of big data analytics for a better understanding of the volatile business environment such that a new methodology is introduced here. The CRISP-DM process serves as *basic design* for multiple reasons (compare *Subsection 3.2.5*). Firstly, although big data comes with novelties, it does not make established processes obsolete (Zhou et al. 2014, p. 64). Secondly, processes are very consistent across various types of data and analytics. Because of the significantly high consistency among methodologies developed in the past, there is no "[...] need to reinvent the wheel every time we need to apply analytics in a different fashion [...]" (Franks 2014, pp. 177–179). This observation is in line with the general consistency observed among the various processes and advancements of CRISP-DM as discussed in *Section 3.2*. Thirdly, the majority of analytics projects follow CRISP-DM (Shi-Nash, Hardoon 2017, p. 333) and many other processes in practice build upon the same principles (Eckerson 2007, p. 11). Finally, CRISP-DM is regarded as valid basis to build advanced methodologies (Shahapurkar 2016, p. 32; Wierse, Riedel 2017, p. 234). Chapman et al. (2000) provide the most comprehensive description of CRISP-DM and thus this work serves as the foundation of the basic design.

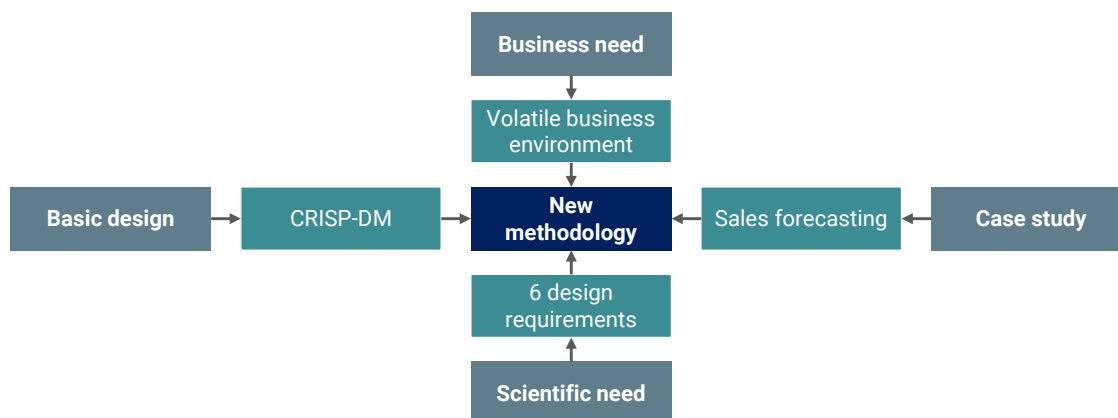


Figure 31 – Basic design concept of the new methodology

The new methodology represents a substantiation of CRISP-DM in a certain context including specific improvements. This follows the idea that new approaches can use CRISP-DM "[...] to augment and improve it in order to make it relevant to today's problems and challenges" (Shahapurkar 2016, p. 35). The *context* is given by the identified *business need* (compare *Section 1.1*) for a better understanding of the volatile business environment. The concept of agility is therefore integrated into the new methodology as it allows to bridge the gap between challenges of the volatile world and potential benefits of BDA applications in this context. Moreover, the *case study* (compare *Subsection 5.3.2*) further specifies the context of the methodology by defining

sales forecasting as primary use case for a BDA application. This is in line with the specific business need to provide improved sales forecasting in the volatile world (compare *Section 1.1 & 5.1*). The *scientific need* (compare *Subsection 3.2.5*) in form of design requirements provides the basis for *improvements*. The new methodology addresses all six identified design requirements. Figure 31 summarizes the basic design concept of the new methodology.

4.1.2 Methodology overview

The new methodology, as presented in Figure 32, consists of five consecutive steps comprising 17 major tasks and builds upon a specific *team setup*. Design requirements for project team (I) and domain knowledge (II) are fundamentally addressed by the team setup. It is designed to meet all coordination challenges within a multidisciplinary BDA team and covers all relevant skill areas including the business domain. The methodology individually describes responsibilities of defined team roles for all tasks. This ensures a collaborative process under the lead of a newly introduced process leader role that bridges most relevant skill areas. Moreover, the team setup incorporates comprehensive business domain knowledge on the basis of multiple roles. The responsibility assignment for these roles as well as the design of each task enables integration of domain knowledge where useful. The following sections provide details on responsibilities and utilization of domain knowledge for each task throughout the entire methodology.

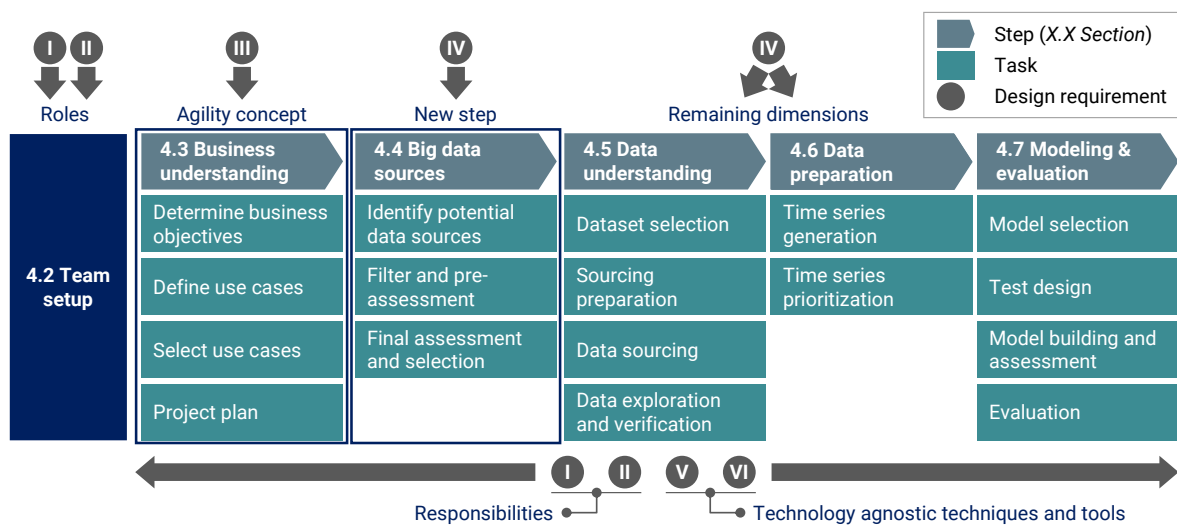


Figure 32 - Overview of the new methodology

The fundamental design on task level incorporates the design requirements regarding methods (V) and automation (VI). The methodology provides specific techniques and tools that give clear guidance on how to implement each task. All methods are described on a conceptual level such that they are technology agnostic. That is to say, they can be realized by different means, in particular, in form of software. This promotes the possibility to automate parts of the methodology. Design requirements with regard to improved business understanding (III) and

data input (IV) are addressed by complete steps. The initial *business understanding* step introduces an *advanced corporate agility system* based on big data analytics which provides business context and connects BDA applications with the challenges of a volatile business environment. Combined with a *six-step method* to determine effects of the volatile business environment, the agility concept guides formulation of *business objectives*. A use case approach underlies the remaining tasks in order to determine specific issues for BDA utilization within the scope of determined business objectives. The task to *define use cases* builds upon a *use case identification workshop* technique. A *use case assessment template* helps to prioritize uses cases before the *use case decision template* and *portfolio matrix for BDA use cases* build the basis to finally *select use cases*. The step concludes with a project plan that builds the framework for all subsequent steps. *Big data sources* represents a complete new step compared to CRISP-DM and other processes. It takes account of key design requirement dimensions for data input (IV) as it enables deliberate selection of multiple internal and external data sources with relevant structured and unstructured data. A structured *data query* serves as method to *identify potential data sources*. *Filter and pre-assessment techniques* extract a long list of data sources from this before a list of *pre-assessment criteria* results in a short list. *Final assessment* is premised on a *data source scoring model* and a *data mix matrix* as preparation for ultimate *selection* during a data source selection workshop.

Data understanding and data preparation address the remaining design requirement dimensions for data input (IV), especially regarding explicit sourcing, project-specific data repository, and data quality. *Dataset selection* initiates the *data understanding* step and defines relevant data within the scope of selected data sources. This task utilizes a simple *dataset selection sheet* as technique. *Sourcing preparation* and *data sourcing* are responsible for extraction of selected datasets from internal as well as external sources. These tasks also install a *project cluster* that stores data for further use including *data exploration and verification* which ensure the use of correct and complete data. The major tool of data understanding is an information repository named *Big Data Analytics book (BDA book)* which describes the project data and documents the required metadata. The BDA book also supports definition of the data sourcing structure and tracks progress of ten sub-tasks comprised in the step. Furthermore, the tool documents results from exploring and verifying data as decision guidance on excluding low quality data. *Data preparation* is organized in two major parts whereby *time series generation* represents the task to generate time-dependent data for use in predictive analytics models. The methodology provides two methods to generate time series from data previously sourced and processed. *Hypothesis-based generation* is the method of choice because it directly incorporates domain knowledge in order to generate most meaningful time series from available data. *Automated generation* serves as alternative option in case the project team lacks sufficient knowledge on a specific data source. The *knowledge & dimensionality test* provides a technique to decide which of the two methods should be applied. *Time series prioritization* as second major part of data preparation carries out the task of selecting time series that are used during modeling. It builds upon three tools specifically developed for this task. *Evaluation tool*, *evaluation report* and *scoring model* determine relevance and quality

characteristics of generated time series at different levels of aggregation. Time series prioritization utilizes this information in two steps. Firstly, *general assessment* reduces the scope of time series to a manageable extent that allows for thorough review of remaining time series. Secondly, *detailed assessment* represents the final definition of modeling input under the use of domain knowledge. The methodology describes both assessment subtasks by a particular technique based on five coordinated assessment steps.

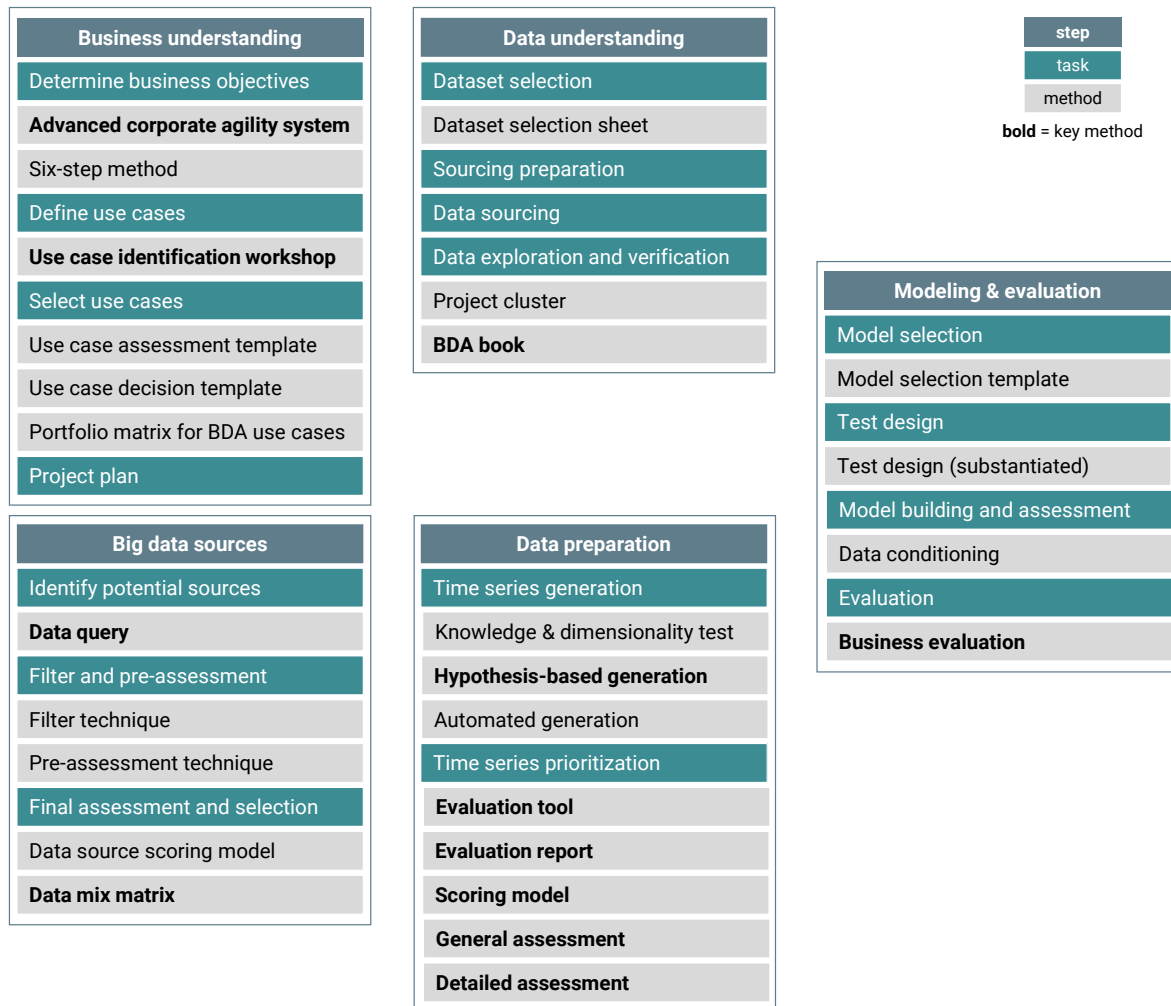


Table 10 - Overview of methods on task level

The final step of *modeling & evaluation* starts with definition of appropriate analytics models for the underlying use case. A *model selection template* facilitates *model selection* from the vast number of existing options. The *test design* represents a substantiated form of a standard method to assess model performance. *Model building and assessment* represent standard tasks to determine the best models for the objective of the use case. The methodology includes *data conditioning* as preparatory subtask here. It defines required data construction actions that transform modeling input into actual input to the model. *Business evaluation* represents a method to augment model assessment from a business perspective and results in a decision about practical application of the developed model as final outcome of the methodology. In summary, 26 methods in form of tools and techniques form a comprehensive methodology to develop BDA applications with

respect to the volatile business environment. Table 10 provides an overview of all methods employed and their affiliation on task level. It also highlights the key methods per step in terms of innovation level and scope. A discussion of differences and similarities towards CRISP-DM is provided in each introductory *step overview* of the following subsections on methodology steps.

4.1.3 Delimitations

In terms of CRISP-DM, the new methodology concludes when "[...] the resultant model appears to be satisfactory and appears to satisfy business needs" (Chapman et al. 2000, p. 58). Additional evaluation tasks, in particular a process review and determination of next steps, follow at this point according to CRISP-DM. These tasks represent preparatory work for the subsequent *deployment* step that covers tasks to implement a model in practical use (Chapman et al. 2000, pp. 58–62). As deployment "[...] involves larger groups of people and is technically less complex, it should be a separate and more strictly managed project" (Lavrac et al. 2004, p. 20). This step is consequently not considered in the new methodology. Moreover, validity of the methodology steps must be divided into a *general part* and a *specific part*. The proposed methodology generally aims for development of BDA applications to promote a better understanding of the business environment in volatile times. The specific part of the methodology substantiates this business need in form of selected use cases. Sales forecasting was selected as use case in the conducted case study that underlies the presented methodology. This use case is characterized by the use of structured data in the form of time series and the developed methods for data understanding, data preparation, and modeling & evaluation take this circumstance into account. The steps for business understanding and big data sources are still valid for the general part. Furthermore, the team setup represents a general advancement provided by the new methodology and is not limited to a methodology for BDA applications for the volatile world. Figure 33 summarizes the major delimitations of the new methodology.

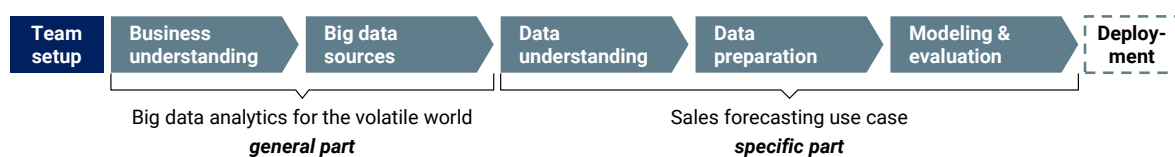


Figure 33 - Major delimitations of the new methodology

The overview on required competencies for big data analytics in *Section 2.5* revealed the need for technological, organizational and cultural foundations. While the new methodology addresses organizational issues regarding the project team, it does not take into account advanced concepts, for instance, in the form of a BDA center of excellence. The methodology reflects a development effort for specific use cases in a project setup. It is not necessary that all required human resources are provided by the company itself. On the contrary, the methodology explicitly allows for integration of external providers, especially for roles that require BDA capabilities, such that no requirements are set towards the internal organization in

this regard. Furthermore, the methodology generally assumes availability of technological foundations which can be supplied by an external provider as well. Cultural aspects in development and utilization of BDA applications are not considered by the methodology.

From data understanding onwards, implementation of tools in the form of software code are required. The methodology conceptually describes all tools in full extent ensuring practical applicability. However, it does not explain details on technical implementation which reflects the idea of technology agnostic methods. A more specific delimitation applies to the tasks of modeling because this step is restricted to the use of available analytics models. The methodology does not reflect any tasks required to build a customized model.

The underlying case study constitutes another delimitation with regard to the proof of concept. The case study was conducted at an industrial company offering technology products to other industrial companies representing a B2B setting. The selected datasets, generated time series, and prioritized model input tend to be specific to this type of business. The resulting BDA application including observed performance of the analytics model for sales forecasting must be regarded against the background of an industrial B2B company. Model performance is covered by model assessment and business evaluation of the new methodology. The proof of concept is therefore examined by these two parts of the methodology.

4.2 Team setup

4.2.1 Project team roles

The methodology follows a use case approach where BDA applications are developed in a project setup and therefore the initial step is to define the roles involved in such a project. The management of big data projects is crucial (Saltz et al. 2017a, p. 184), but processes generally neglect the role of human resources (Alnoukari 2012, p. 192). Big data analytics requires a broad set of skills across various areas including big data technologies, machine learning, software engineering, and data privacy among others (Shi-Nash, Hardoon 2017, p. 341). Moreover, not only technological skills are relevant because business skills need to be involved as well (Loshin 2013b, pp. 49–50). The role of the data scientist largely dominates the discussions on required skills (Saltz et al. 2016, pp. 2896–2897), however, "[...] no single person may be skilled in all these [relevant] areas [...]" (Grady 2016, p. 1604). In order to cover the extensive diversity of skills, projects require multidisciplinary teams (Gao et al. 2015, p. 827).

Hofmann, Tierney (2009) provide an overview of eight project team roles based on skill grouping. *Business analysts* have an understanding of business aspects and formulate project objectives as well as evaluates its results. *Data administrators* are familiar with requirements and designs of databases which they construct, while *data engineers* are able to extract knowledge from domain expertise in order to prepare data for analytics. *Domain experts* have deep subject matter expertise which they feed into dependent process activities. *Data miners* with a broad analytics skill set generate algorithms and models before *knowledge engineers* ensure the use of discovered knowledge. *Strategic managers* have extensive business knowledge which enables them to identify business problems for analytics, assess the strategic circumstances, and secure provision of data. *Project managers* run the project based on general project management skills (Hofmann, Tierney 2009, pp. 62–64). Alnoukari (2012, pp. 192–193) confirms these roles as relevant human resources in development of BDA applications. EMC Education Services (2015) describes a similar setup of the project team with only seven different roles because *business users* act as business analysts and domain experts here. Moreover, this work mainly updates role names and further specifies the *project sponsor* as initiator and funder of the project (EMC Education Services 2015, pp. 26–28). Table 11 compares both setups of project team roles including main responsibilities for each role. Collier (2012, pp. 64–65) describes three groups of similar roles that cover the same scope but adds *stakeholders* as another role to be considered in the *planners* group. As the need for stakeholder involvement highly depends on the underlying use case, they are generally not regarded as fixed role in the team setup. The roles of *knowledge engineer* and *BI analyst*, respectively, are not considered in the following as they are mainly involved in the evaluation and deployment steps.

Hofmann, Tierney (2009)		EMC Education Services (2015)		Collier (2012)
Role	Main responsibilities	Role	Main responsibilities	Role group
Business Analyst	Formulate objectives; evaluate analytics results	Business User	Provide domain knowledge; evaluate and operationalize results	Consumers
Domain Expert	Provide domain expertise			
Data Administrator	Construct databases	Database Administrator	Provide database environment	Doers
Data Engineer	Prepare data	Data Engineer	Manage, extract and ingest data	
Data Miner	Generate algorithms and models	Data Scientist	Design and execute models	
Knowledge Engineer	Ensure use of discovered knowledge	BI Analyst	Gain insights from analytics results; process results	
Strategic Manager	Identify issues; assess strategic circumstances; provide data	Project Sponsor	Initiate and fund project; define business problem; set priorities and clarify objectives	Planners (incl. stakeholders)
Project Manager	Manage the project	Project Manager	Manage objectives, time and quality	

role not considered for team setup

Table 11 - Setups for project team roles

4.2.2 New team roles

The evaluation of existing processes in *Section 3.2* revealed a need for improvement in project management and that team setup plays a crucial role here. In particular, the large number of tasks to be performed by a multidisciplinary team requires significant coordination (Espinosa, Armour 2016, p. 1112) in order to manage the interdependencies among tasks and roles (Malone, Crowston 1994). Espinosa, Armour (2016) break down coordination within BDA projects into three different types. *Technical coordination* relates to the infrastructure, tools and other methods such that it is provided by the foundations and process design. *Temporal coordination* describes adherence to timelines and *process coordination* stands for the diligent use of the process (Espinosa, Armour 2016, pp. 1114–1116). While technical coordination is universally required across BDA applications, the latter two types of coordination describe the scope of work for a traditional project manager of an individual use case. However, they do not explicitly address the challenge of coordinating a multidisciplinary team. The team setup needs to be adapted to account for management of differently skilled roles in addition to management of the project itself. This additional type of coordination is taken into account in the presented methodology and is referred to as *team coordination*. Data scientist and business user already play a central role in a project team and thus are candidates to fill in this gap. However, an effective team does not only require an adequate organization but also necessary skills (Phillips-Wren et al. 2015, p. 25). Although data scientists bridge the analytics and business domain (Zemmouri et al. 2011, p. 18), they usually show deficiencies in business understanding. An important trend is increasing specialization of data scientists due to the ever more variety of analytics (Henke et

al. 2016, p. 38), which shifts the skill focus of data scientists further away from the business domain. Business users have limited BDA knowledge (Collier 2012, pp. 16–17) and therefore lack skills to coordinate other roles such as data engineers. Strategic managers or project sponsors also show a lack of analytics skills (Ransbotham et al. 2015, p. 63) and database administrators as well as data engineers typically have specialized skills. None of these roles provides a balanced skill profile that is required for team coordination.

As the project manager is already responsible for coordination tasks, an obvious approach is to upgrade this role. The idea is to transform the project manager from a passive role that manages the project into an active role that also manages required skills across all disciplines. Manyika et al. (2011) separate the traditional data scientist role with very broad skill set into two different roles. On the one hand, *deep analytical talents* are capable of generating business insights from big data by using advanced analytics and therefore describe data scientists in a narrow sense. This definition of data scientist is used in the remainder of this work. On the other hand, *data-savvy managers* have a basic understanding of big data and analytics such that they can formulate BDA problems as well as evaluate their results (Manyika et al. 2011, p. 103). The latter provides the archetype to fill in the new project management role including team coordination but a more specific description of its skill set is still required. Ariker et al. (2014) describe general roles and their relation to skill areas within a BDA organization²⁰, whereby they put the head of analytics at the center of the three major functions of IT, business domain, and analytics. Taking into account the prominent role of big data, this view can be expanded in order to describe the roles within a BDA project team where the new role is at its center as well. Figure 34 describes the relevant skill areas underlying the team setup.

As the new role touches the major areas of big data, analytics, and business domain it is termed *Big Data Analytics manager (BDA manager)*.²¹ The BDA manager is responsible for team coordination with data scientists, data engineers, database administrators and all roles within the business domain. Although the role covers all key skill areas, it demands stronger skills in the intersection of analytics and business domain. This reflects the key role of business understanding in the process (compare *Section 3.2*) and the BDA manager as central role in the project should also be in charge of this step. The *database administrator* role lies at the intersection of big data and IT²² such that they connect to the BDA manager via data scientist or data engineer.

²⁰ A BDA organization can be seen as organizational unit that is responsible for multiple projects or use cases, respectively.

²¹ Ariker et al. (2014) use the term *translator* for someone “[...] bridging different functions within the organization [...]”, however, they limit the role to two complementary skill areas.

²² IT mainly represents the technological foundations as described in *Section 2.5*.

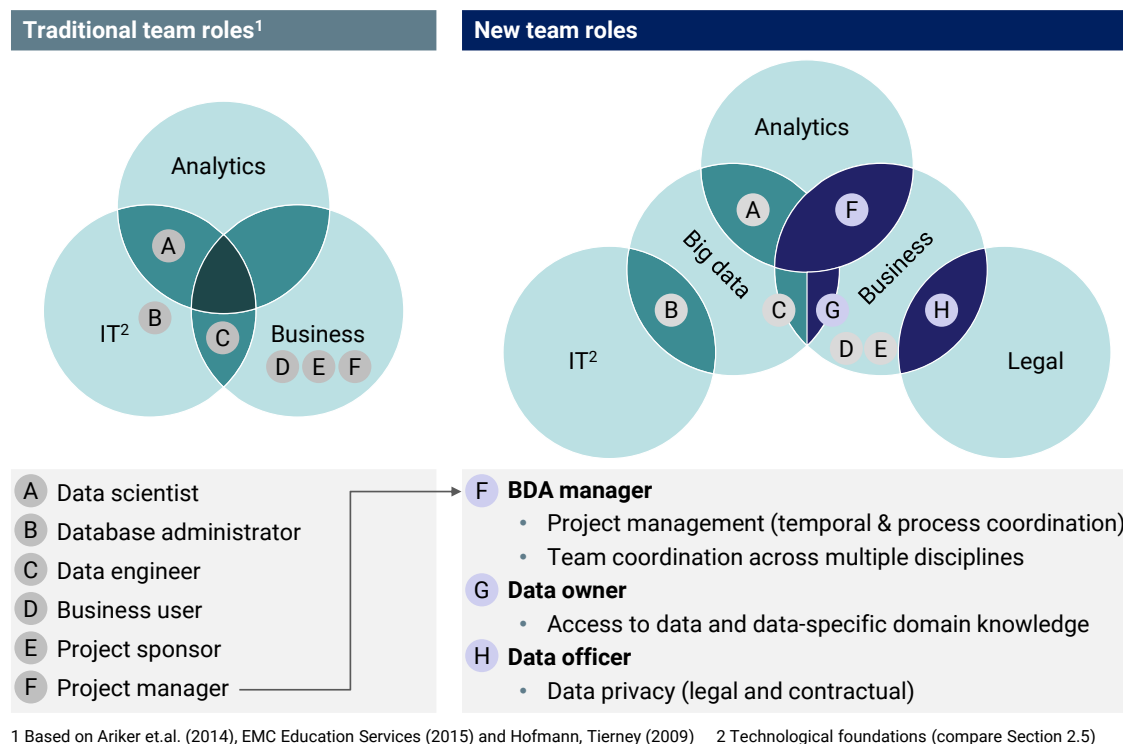


Figure 34 – Skill areas of project team roles

As the BDA manager is also responsible for temporal and process coordination, the skill set²³ is not limited to the presented areas. Candidates should have strong outcome orientation, eagerness to learn, propensity to experimentation and they should be promoters of advanced technologies (Viaene, van den Bunder 2011). Furthermore, they need to be able to assume responsibility for senior management interaction as critical practice in projects (Saltz, Shamshurin 2015, p. 2104). There is a positive effect of implementing the BDA manager for this task. It reduces the skill requirements for increasingly specialized data scientists or data engineers and increases quality of communication compared to a lead by a project manager with less BDA capabilities. Business knowledge is reported to be the most critical skill to perform predictive analytics (Halper 2014, p. 15), such that a business background is beneficial. Another quality of the BDA manager is the capability to cope with ambiguity and uncertainty. This is important as BDA projects are less structured compared to other project types, for example, in software development (Provost, Fawcett 2013b, pp. 34–35). The two previous qualities relate to the demand for presentation skills and creativity (Franks 2012, pp. 224–225). Adequate profiles that can combine BDA capabilities with business knowledge and communication skills are scarce (Janssen et al. 2017, p. 342). However, the variety of human resources involved in

²³ In their recent publication, Henke et al. (2018) describe the *analytics translator* as similar role with the following skill set: domain knowledge, technical fluency, project management, and entrepreneurship.

BDA activities is already very broad (Russom 2011, pp. 14–15) such that candidates can be drawn from various backgrounds.

The BDA manager plays a major role in the project team setup. In contrast to the traditional roles, two additional roles complement the setup: data owner and data officer. Both roles represent stakeholders that are regularly involved in BDA projects and their skill areas lie at intersections of the business domain, as depicted in Figure 34. Data ownership describes the possession and control over data (Techopedia 2017). *Data owners* within a company are therefore crucial to gain access to relevant data, and furthermore they are sources of domain expertise related to this data (Cios et al. 2007, p. 469). Their skill set lies at the intersection of the business and big data domains. Data owners help to keep control over big data as they can support to identify quality issues, validate consistency of data across multiple sources, interpret data, and enrich data if possible (Loshin 2013a, p. 43). For example, data owners of enterprise resource planning data can be found in controlling and owners of customer relationship management data in sales. Data owners also exist for external sources, for instance in the form of customer support in the case of professional data providers or user support of public databases. The BDA manager serves as liaison to data owners of external sources.

Data security and data privacy are two important issues in big data related activities (Ou et al. 2016, p. 285). Data security is mainly concerned with unauthorized access to data but also addresses other technical issues such as data corruptions and backups (Ohlhorst 2013, p. 63). Access control, encryption or other technologies are means to keep data secure (Ohlhorst 2013, pp. 69–70). Unauthorized use of personal data poses an increasing privacy concern in relation to big data (Xu et al. 2014, p. 1150). Certain types of data are protected by law, for example, personally identifiable information or sensitive information (Minelli et al. 2013, p. 159). Privacy concerns can also rise apart from legal protection, for instance, when customers have higher expectations towards privacy compared to the legal status quo or when a company and its supplier conclude an agreement on the use of shared information. As a consequence, the *data officer* provides a role that addresses the challenges of data privacy. The intersection of business and legal defines the skill set of the role. Business knowledge is required to provide an understanding of privacy concerns based on existing relations with a company's stakeholders, for example. Legal capabilities address issues rising from privacy laws or individual contracts. Data owner and data officer are also coordinated by the BDA manager.

4.2.3 Roles and responsibilities

Roles including their required skills and main responsibilities provide the blueprint for setting up the project team. For this team to successfully run the BDA project, clear guidance by the process is required in order to avoid neglected or uncoordinated tasks (Saltz 2015, p. 2067). The mapping of roles and tasks is called workflow model and can be organized by process steps within a BDA project (Tuovinen 2014, pp. 88–101). The presented methodology implicitly

integrates the workflow model in the process description such that each role is described for each task. Table 12 summarizes the workflow model across all roles and aggregated to step level.

	Traditional team roles					New team roles		
	Business user	Project sponsor	Data scientist	Data engineer	Database administrator	BDA manager	Data owner	Data officer
Business understanding								
Big data sources								
Data understanding								
Data preparation								
Modeling & evaluation								


 role with responsibility in the step

Table 12 - Workflow summary by process steps

The overview underlines the key roles of business user and data scientist as they provide comprehensive skills as well as knowledge in the key skill areas of business, big data, and analytics, respectively. As will become clear in the discussion of individual steps, their responsibilities in the presented methodology are clearly linked to their skill base. The BDA manager does not only hold the overall project responsibility but also answers for each single step in order to coordinate all other roles. Involvement of the remaining roles depends on specific needs of a step. For example, project sponsor involvement is required to formulate business objectives and to select use cases during business understanding as well as during business evaluation to assess final results after modeling. The workflow summary furthermore provides guidance for the assignment of team roles. The initial team setup comprises BDA manager, project sponsor, business user, and data scientist. The other team roles can be assigned in accordance with the workflow. This allows to consider specific features that are defined in the course of the project. Specification of selected data sources facilitates the assignment of an appropriate database administrator, for example. Furthermore, it is important to note that a team role can comprise multiple team members and vice versa.

4.2.4 External roles

A BDA project does not necessarily need to be conducted with only internal resources. Outsourcing part of the work or even an entire project is an option depending on the existing resources and the type of project, whereby a potential outsourcing scope can be defined by eligible tasks (Martins et al. 2016b, pp. 508–509). Wierse, Riedel (2017, pp. 236–240) provide some guidelines for partial and full outsourcing. The function of outsourcing regarding the presented methodology is represented by the definition of external roles that are not covered by the company itself. Data scientist, data engineer, and database administrator are typical candidates for outsourcing. As the methodology clearly defines each role and its responsibilities, it allows to fill in these roles with external human resources. Furthermore, the introduction of

NEW METHODOLOGY

the BDA manager further enables a team setup including external roles. The BDA manager remains the leading role in order to coordinate interactions across organizational boundaries and the role itself can be provided by an external source, for example, by a specialized consultant. In this case in conjunction with outsourcing of BDA related work to a specialized provider, three parties are involved in the project. As this results in two organizational interfaces from the BDA manager perspective, the project team setup can be extended with additional assistant project managers covering for straightforward project management in each respective organization. In the given case, potential candidates for these roles are a lead business user for the company and a lead data scientist for the external BDA provider, because both roles are involved throughout the entire workflow of the methodology.

4.3 Business understanding

4.3.1 Step overview: Determine, define & select

In this first step, the foundation for all subsequent steps is laid out. Business understanding produces a project plan coordinating all subsequent tasks, it provides the scope of potentially relevant data as well as applicable models (e.g., predictive analytics based on time series data), and it sets the framework for model evaluation (Sharma et al. 2012, p. 11338). Business understanding therefore is a key step in developing a BDA application, however, "business requirements analysis is often neglected" (Priebe, Markus 2015, p. 2056).

In general, there are two different approaches to determine objectives of analytics initiatives. Singh et al. (2011) describe the use of domain knowledge and data-driven approaches as the basic options. In the latter case, they differentiate between the mapping of existing data with modeling techniques that is still lead by some influence of domain knowledge and the pure search for interesting patterns in data (Singh et al. 2011, pp. 280–281). Both data-driven approaches are based on explorative data analysis, that should be part of the data understanding step while an underlying business objective needs to be defined beforehand (Lanquillon, Mallow 2015a, p. 75). In other words, there is a need for a methodical approach on where to apply big data replacing "[...] the far more open ended question of trying to find 'value in the data' [...]" (Saltz 2015, p. 2067). Vanauer et al. (2015, p. 910) furthermore state that data-driven approaches rather aim for the development of new business models and therefore differ from solving a given issue of the company. The volatile business environment potentially poses different issues and it is the major aim of business understanding to determine whether such challenges exist as well as to determine all relevant details in order to develop a solution based on big data analytics.

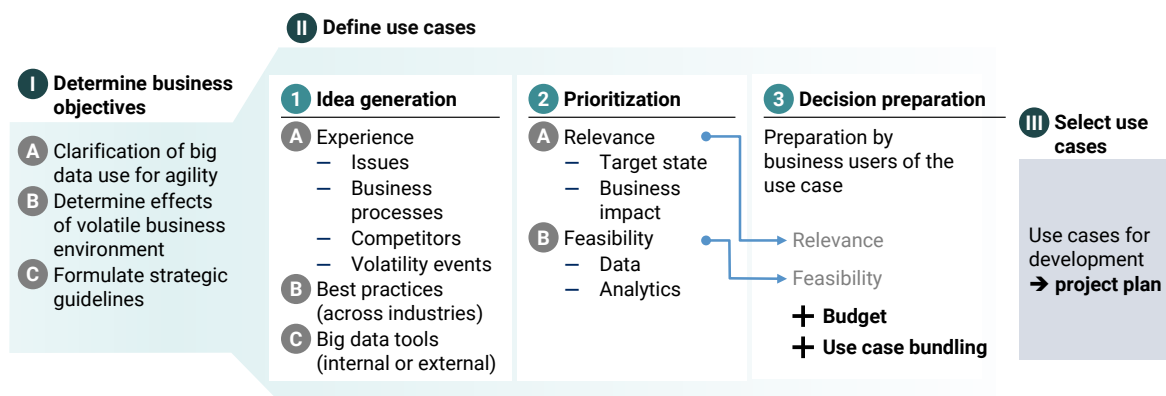


Figure 35 - Business understanding step [based on Heldmann et al. (2017)]

The major outcome of the business understanding step is scoping of BDA applications. In doing so, the focus should lie on known business issues that can be formulated as specific use cases and the required capabilities for each use case need to be considered (Cato et al. 2015, p. 137). Ideation of BDA applications and their assessment are the key tasks for this use case-based approach (Vanauer et al. 2015, p. 908). As inadequate business context is a key driver of big data

project failure (Kelly, Kaskade 2013) and BDA utilization is meant to improve handling of the volatile business environment, the concept of agility serves as framework for the initial step. The first task is to *determine business objectives* based on domain knowledge with the help of the agility framework. It results in a common understanding of the business problem areas and strategic guidelines for the subsequent tasks. These guidelines serve as the business objectives and business success criteria that BDA applications must submit to. The methodology therefore represents a substantiation of the initial task of CRISP-DM in the context of big data analytics for volatile times (Chapman et al. 2000, pp. 35–37). *Define and select use cases* cover the assessment of the situation and establishment of analytics goals in the form of individual use cases which then result in a project plan to develop BDA applications (Chapman et al. 2000, pp. 37–42). Figure 35 summarizes the overall design of the business understanding step that is based on the work of Heldmann et al. (2017).

4.3.2 Determine business objectives

4.3.2.1 Big data analytics in the agility framework

In general, two different ways exist how big data analytics creates value for a company. Strategic value benefits from faster as well as proactive decision-making and, on the other hand, operational value comes from increased efficiency, for example, due to optimized processes (Omri 2015, p. 104). Improved profitability can serve as simple framework for industrial companies that seek internal optimization of manufacturing processes (Heldmann et al. 2017, pp. 83–84). Isson, Harriott (2013, 41–42) discuss other operational business issues that can be addressed by big data analytics such as increase of customer retention or employee productivity.

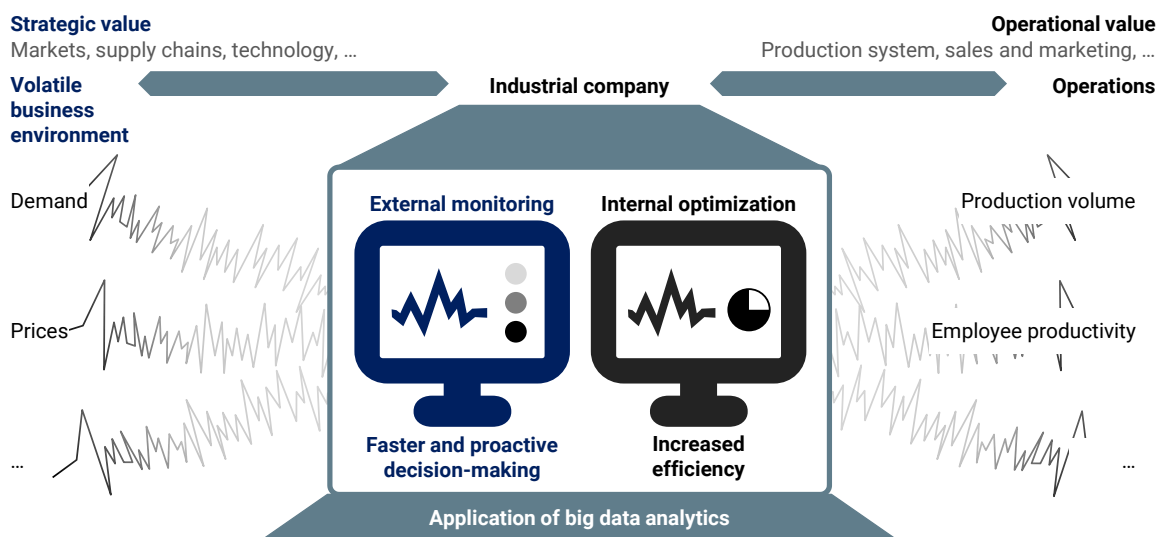


Figure 36 - Strategic and operational value of big data analytics
[based on (Heldmann et al. 2017, p. 80)]

The focus of the presented methodology lies on the strategic value of big data analytics against the background of a volatile business environment, for example, in the form of volatile demand or market prices, which is illustrated by Figure 36.

"Formulation of business objectives is the first step in any [...]" (Sharma, Osei-Bryson 2015a, p. 56) project and problem formulation methods exist that "[...] provide some structure towards formulating business problems in the ill-structured decision context of [...]" (Li et al. 2016b, p. 4) big data analytics. Sharma, Osei-Bryson (2015a) introduce a novel comprehensive method to formulate business objectives as part of their IKDDM process and Li et al. (2016a, pp. 1251–1253) extend their KDDA process with a similar method in the context of environmental risk management. Both methods are built upon specific techniques, especially value-focused thinking (Keeney 1992) and goal question metrics (Basili, Weiss 1984), that are not specifically designed for objective formulation in the given context of BDA utilization for the volatile world. As a consequence, a new method based on the agility concept is proposed here. It starts with the clarification of BDA use for agility which has two dimensions. On the one hand, the company needs to acknowledge the challenge of a volatile business environment and that striving for increased agility is aligned with the overall strategy. On the other hand, the company also needs to endorse big data analytics as valid approach for a better understanding of the volatile world.

The corporate agility system, as presented in *Section 2.1*, comprehensively describes how companies cope with the volatile business environment. It is not necessary that a company strives to strictly implement this holistic system as it still can serve as a descriptive framework that generally describes how companies cope with such an environment. Furthermore, the idea of big data analytics can be integrated into this framework and therefore bridges the gap between arising challenges and potential BDA applications. It has been shown that monitoring of the external business environment is a key building block of this concept. In order to integrate big data into the concept, the monitoring function can be extended. Heldmann et al. (2015) discuss that monitoring of the business environment has become increasingly difficult due to the variety of potential volatilities and their underlying drivers. As a consequence, it is more difficult to monitor volatility-driven changes, however, this challenge can be addressed by a BDA approach to monitoring (Heldmann et al. 2015, pp. 37–38). According to Heldmann (2017), the basic idea behind BDA utilization is two-fold: firstly, utilizing big data extends the available information base compared to small data (*data view*), and secondly application of analytics allows to gain valuable insights from this base (*analytics view*). The data view rests on the 4V definition of big data. Increasing volume and velocity of data input allows for in-depth information due to a higher level of granularity of information. Furthermore, variety and veracity of information create a broader information base as more data types from different sources are considered (Heldmann 2017, pp. 182–185). Table 13 provides an example for each element of the information base with regard to a better understanding of volatile sales behavior.

Element	Example
Volume	In order to gain a better understanding of sales behavior, a company can look into order backlog data of individual customers instead of typically aggregated order data for business segments or regions. This allows to identify order behavior of individual customers that might serve as early warning indicators, for example.
Velocity	The most extreme example of velocity would be real-time analytics of changes in orders. While this has limited practical relevance for industrial companies, an increased frequency of analyzing order data is already an improvement. Order backlog data is often only used on a semi-annual or even annual basis within a budgeting process to create sales outlooks. Continuous utilization of this data, for example, on a monthly basis, can enable faster decision-making.
Variety	The previous two examples illustrate how volume and velocity increase the depth of the information base even for data traditionally used in industrial companies. However, big data analytics enables to bring various data sources together. A company seeking to better understand sales dynamics could include market information in the form of industry news from the web. This unstructured data most likely includes different information than structured order data and thus increases the width of the information base.
Veracity	The use of web-based information also serves as example for veracity. In contrast to internal order data, this kind of data potentially comes with issues of credibility. However, advanced analytics techniques still enable the use of such information today.

Table 13 - Example for increased information base [based on (Heldmann 2017, pp. 183–184)]

As discussed in *Section 2.3*, utilization of big data requires analytics in the form of models. The analytics view as proposed by Heldmann (2017) provides a generalized description of this aspect, and builds upon the idea of correlation-based analytics. Correlation needs to be seen in a broader sense here and describes the use of big data analytics in order to find valuable patterns or insights which is in principle contrary to causality (Mayer-Schönberger, Cukier 2013, pp. 50–72). Causality requires to explicitly model relationships and is therefore limited by a priori knowledge as well as underlying complexity of the observed dynamics, while correlations can also uncover unknown relations and are not generally restricted by system complexity (Heldmann 2017, pp. 183–185). The major advantage of the latter is the ability to describe complex dynamics with the help of big data instead of small data input. However, it is also part of the analytics view that models have a black box character to a lesser or greater extent, which poses a challenge for acceptance in practical applications (Biesdorf et al. 2013, p. 9). It is therefore addressed during business evaluation of this methodology (compare *Section 4.7*). Figure 37 illustrates the correlation-based analytics view in contrast to a traditional causality approach.

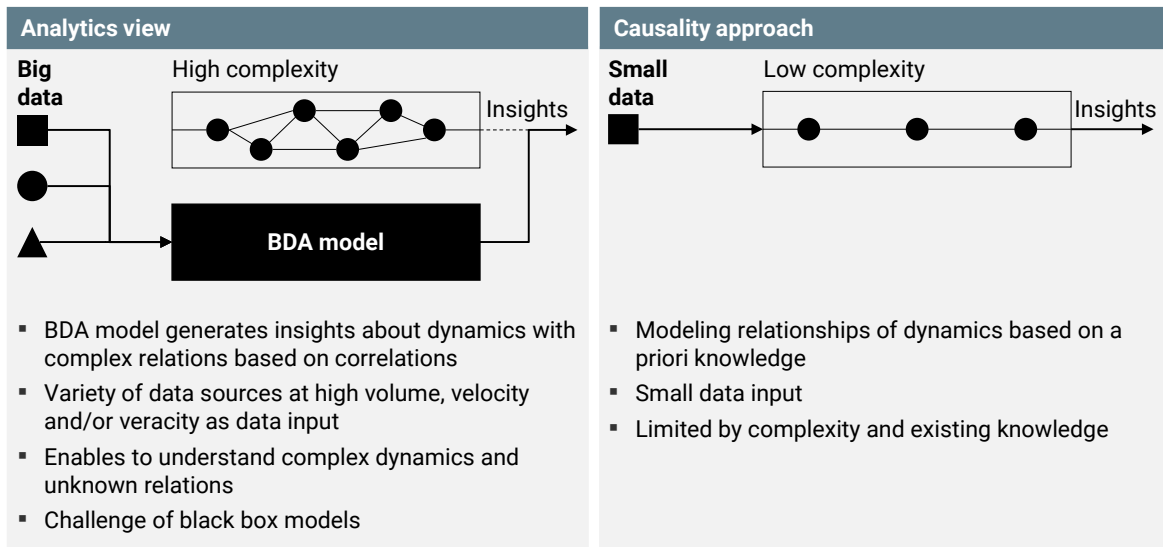


Figure 37 - Analytics view versus causality approach [based on (Heldmann 2017, p. 183)]

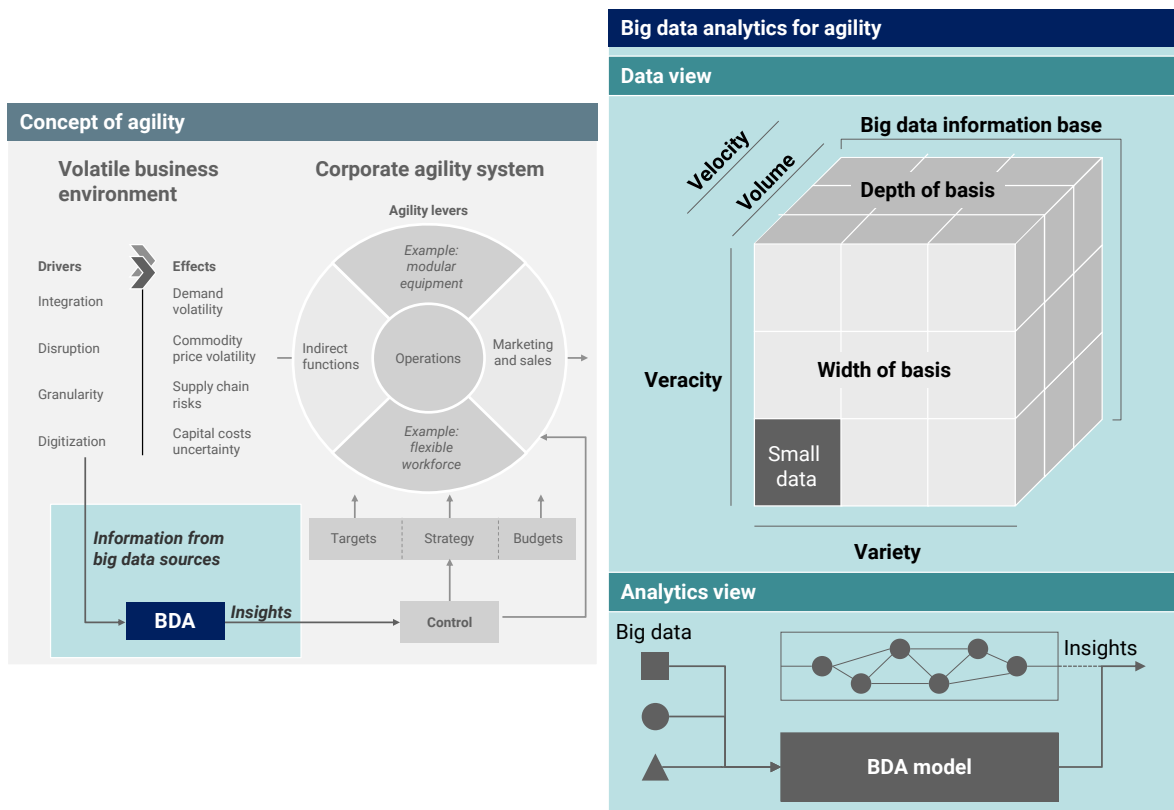


Figure 38 – Advanced corporate agility system [based on (Heldmann 2017, pp. 163–185)]

The data and analytics views allow to expand the corporate agility system as depicted in Figure 38. It serves as framework for the initiation workshop with the project sponsor and optionally business users. The workshop is led by the BDA manager who utilizes the framework to clarify the notion of agility and the role of big data analytics as part of the solution. Alignment between the agility framework and the company strategy can be further substantiated by appraisal of the mission statement. Isson, Harriott (2013, p. 42) state that "[u]nderstanding and validating business challenge priority within your organization and aligning with your business's mission will allow you to focus your analytics expertise on the most critical challenges [...]". The mission statement summarizes a company's value proposition or business model and serves as guidance for strategy formulation (Isson, Harriott 2013, pp. 36–37). With positive alignment and endorsement of BDA solutions by the project sponsor, determination of business objectives is substantiated by assessment of the business environment.

4.3.2.2 Volatile business environment and strategic guidelines

In order to focus on the most critical challenges within the agility framework, it is necessary to determine effects of the volatile business environment on the company. Kremsmayr (2017) introduces a six-step method that enables determination of key volatilities in a structured way. In the *first step*, a list of potential volatilities as complete as possible is compiled whereby identification of volatilities rests on three different views. The *macro view* scans the business environment with regard to economics, technology, politics as well as law, ecology, and social issues. The *micro view* assesses the situation within the industry based on a conventional market and competitive analysis. Potential volatilities at the operational level are identified by discussions with employees across different functions. This *internal view* is less relevant in the given methodology due to the focus on external volatilities and strategic value of big data analytics. The *second step* is to estimate the probability of occurrence for each identified volatility. A qualitative estimate is sufficient here, for example, ranging from *very rare* to *highly probable*. The *third step* assesses the company's responsiveness to identified volatilities. In order to evaluate responsiveness, the entire reaction process including identification of a necessity of acting, decision-making regarding specific reactions represented by agility levers in the corporate agility system, and their implementation. It is important to note that the speed of responsiveness is not the only criteria but effectiveness of agility levers also play a role here. Closely related is the impact assessment of the *fourth step*. Ideally, the financial impact is calculated to describe the effects of volatilities and associated reactions. This financial impact is translated into qualitative categories depending on the company situation which then enables to derive the need for action in the *fifth step*. Figure 39 provides an example how probability of occurrence and financial impact lead to different levels of need for action. In the *sixth step*, the most relevant volatilities are selected based on the prior assessments (Kremsmayr 2017, pp. 63–74).

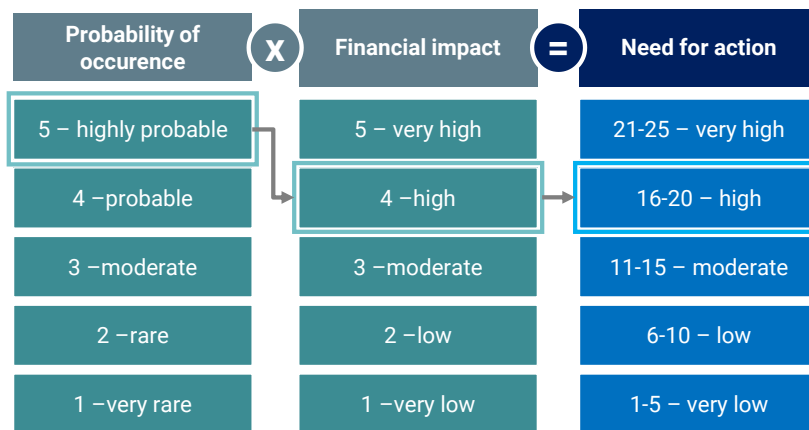


Figure 39 - Determination of need for action [based on (Kremsmayr 2017, p. 72)]

The presented six-step method represents a comprehensible way to determine most relevant volatilities. However, it also constitutes a significant effort that requires to involve experts for the various assessments whereby some expertise might lie outside of the project team. Dutta, Bose (2015, p. 294) state that "[d]irections from senior management, inputs from various business units who are stakeholders of the project help in understanding the scope of the [BDA] problem". That is to say that project sponsor and business user play an important role in identifying most relevant volatilities in the business environment. Senior managers typically have a good understanding of essential issues for their company. The initiation workshop can therefore serve as an alternative to the six-step method at minimal burden. Once a list of relevant volatilities is determined in the workshop they can be subsequently verified with business users related to them. In case this practical approach is not successful, the six-step method can still be employed. The list of most relevant volatilities serves as basis for formulating strategic guidelines, which "[...] are primarily identified to scope the search for use cases" (Vanauer et al. 2015, p. 911).

4.3.3 Define and select use cases

Once business objectives in form of strategic guidelines are formulated, the development of use cases for BDA utilization follows. The business objectives hereby act as guidelines for the following discussions. Main roles are again the BDA manager and business user as they define and assess use cases as well as prepare the project plan after selection of uses cases in alignment with the project sponsor. In addition, a data scientist supports the discussions in order to provide detailed knowledge on big data and analytics. This is crucial as development of use cases requires knowledge of BDA applications in business including "[...] what realistic expectations are from the various approaches [...]" (Shahapurkar 2016, pp. 40–41).

4.3.3.1 Idea generation

The identification of use cases is based on a structured workshop format that covers three different dimensions: company experience, best practices, and big data tools. The workshop is led by the BDA manager who prepares the discussion and consolidates results. An exemplary workshop agenda based on two and a half hours length of time is shown in Table 14.

Agenda topic	Duration
Introduction	20 mins
Idea generation	90 mins
<ul style="list-style-type: none"> • Experience • Best practices • Big data tools 	<i>(approx. 30 mins per dimension)</i>
Processing of idea generation results (grouping, aggregation, information collection)	30 mins
Wrap-up & next steps	10 mins

Table 14 - Use case identification workshop (exemplary agenda)

During the *introduction*, the BDA manager recaps the established business context and strategic guidelines as defined previously. It ensures that the project team has the same understanding of the guidelines before identifying specific use cases. Identification of use cases can be supported by creativity techniques that are tailored to the roles involved in the process (Vanauer et al. 2015, p. 912). In particular, brainstorming is a technique in order to identify ideas for "[...] where and how to leverage big data [...]" (Schmarzo 2013, p. 134). Brainstorming is an established technique that can also be used by inexperienced groups (McFadzean 1998, p. 137). Despite existing flaws of brainstorming, it is widely used in business practice because it allows idea generation in a group of different experts and as a democratic method it creates buy-in among participants (Chamorro-Premuzic 2015). Wilson (2013, pp. 1–41) provides general explanations on the procedure and rules as well as practical advice on performing brainstorming that go beyond the scope of this work. The use of focused brainstorming, which requires strict focus on predefined goals and uses deliberate constraints on the idea generation process, is proposed as specific techniques for diverse teams in business practice (Ulwick 2005, pp. 143–147). The following describes the idea generation process to identify use cases based on a tailored technique of focused brainstorming.

Viewpoint	Description	Guiding questions
Issues	Specific issues the company suffers from with regard to identified business objectives	<ul style="list-style-type: none"> • Where do you see issues that might benefit from BDA utilization and how? • Where do you have limited understanding of changes in the business environment?
Business processes	Business processes that relate to identified business objectives	<ul style="list-style-type: none"> • Which business processes are affected by identified business objectives? • What are important internal and external data sources related to these processes? • Are there untapped data sources that are potentially valuable to improve the
Competitors	Capabilities that provide a competitive edge with regard to identified business objectives	<ul style="list-style-type: none"> • Which BDA use cases of competitors do you know? • What data sources are used by competitors?
Volatility events	Specific changes in the business environment that affected the company in the areas of identified business objectives	<ul style="list-style-type: none"> • What changes in the business environment strongly affected the company's business in the past? • Where is the complexity of business environment so intense such that current approaches failed to provide understanding of past changes?

Table 15 - Company experience dimension of idea generation

Most important part of the *idea generation* is the discussion on company *experience*. It examines four different points of view in order to collect business user insight on potential use cases. The first viewpoint directly addresses *issues* that are present within the company and therefore builds on needs for action already known by business users. Another important point of view in identification of use cases are *business processes* (Vanauer et al. 2015, p. 912). Furthermore, the discussion takes a closer look at BDA applications of *competitors*. Benchmarking the company against competitors is a common way to learn from competition (Ulrich, Lake 1991, p. 90). This viewpoint is to some extent opportunistic, as competitors typically seek to keep their own learnings proprietary (Hemmatfar et al. 2010, p. 164). Finally, *volatility events* in the business environment that posed challenges to the company in the past are considered. This view can build on results from the six-step method if employed before but focuses on specific examples observed by the business user. The four views provide guidance to the idea generation discussion and each view is focused on the predefined business objectives. The discussion is furthermore supported by guiding questions for each point of view. Table 15 provides a summary of the company experience dimension including guiding questions to be used by the BDA manager.

The second dimension of the idea generation is guided by *best practices* that can be observed across industries. "The purpose [...] is to understand how similar problems have been addressed by other companies [...]" (Dutta, Bose 2015, p. 295), whereby similar problems relate to the identified business objectives again. This part of the structured workshop requires preparation of relevant best practices by the BDA manager and mainly builds upon available information in academic and practitioner literature. Each best practice is briefly described in order to facilitate a discussion about potential transfer to the company. The third dimension builds upon *big data tools* that are readily available for use. Big data tools represent analytics tools capable of big data input that are generally applicable, that is to say they are not specific to an application such as sales forecasting. On the one hand, these can be big data tools already deployed in the company. On the other hand, also externally available big data tools are considered in case the project includes external providers. Each big data tool is described with regard to its functionality in a way that is easy to understand by business users. The data scientist plays a central role in this discussion by evaluating potential use cases built on the big data tools identified.

Identification of use cases concludes with grouping and aggregation of generated ideas (Schmarzo 2013, pp. 136–137) in order to describe distinct use cases by title and short description. The processing of idea generation results also includes the collection of further information revealed during the focused brainstorming, in particular, data input and big data tools identified for a specific use case. The resulting long list of use cases provides the basis for subsequent assessment.

4.3.3.2 Prioritization

Assessment of the use case long list aims to generate a short list of most valuable use cases from business user perspective. This prioritization task can be seen as project portfolio selection, which is widely applied for activities including developing of new products, launching new production processes as well as implementing new information systems (Archer, Ghasemzadeh 1999, p. 207). The long list of identified use cases can be seen as new information systems that need to be selected for potential implementation. Archer, Ghasemzadeh (1999) provide an overview of approaches for ranking different options. *Ad hoc approaches* eliminate options that do not meet minimum requirements or use an iterative process based on multiple interactions with business users and project sponsor in order to identify the best option. *Comparative approaches* assess each option regarding their contribution to different criteria in comparison to the other options. As these criteria are weighted, the approach results in a ranking list. *Scoring models* determine an overall measure for each option by aggregating scores of multiple criterions. *Optimization models* aim to identify the option with the best benefit, for example, in form of the maximum net present value. *Portfolio matrices* assess options along two dimensions and use graphical representation to aid selection (Archer, Ghasemzadeh 1999, p. 210).

It is difficult to define specific minimum requirements for BDA use cases and project sponsors are typically top managers such that an iterative process is costly. Ad hoc approaches therefore

are not the preferred choice in the same manner as comparative approaches due to their high level of effort for comparing multiple options (Archer, Ghasemzadeh 1999, p. 210). Prioritization approaches for use cases are often based on quantitative assessments, for example in the form of scoring models or optimization models in form of financial evaluations such as discounted cash flow models (Williams 2016, pp. 91–95). Despite being conceptually simple and transparent, scoring models are time-consuming and scoring can be arbitrary (Milosevic 2003, pp. 30–31). This becomes clear when considering the issues of constructing a scoring model: definition of scoring criteria, relative importance of criteria, and quantitative measurement (Martinelli, Milosevic 2016, p. 35). The major drawback of optimization models is the need for extensive forward-looking data (Milosevic 2003, p. 45) that is typically not readily available at initiation of a BDA use case. As a consequence, the proposed technique is based on the idea of using two major assessment dimensions with qualitative assessment criteria as proposed by the portfolio matrix approach (Archer, Ghasemzadeh 1996, pp. 17–23). Its dimensions cover business impact and feasibility of implementation as well as adopts a workshop format for prioritization as proposed by Schmarzo (2013, pp. 138–139).

	Use case	Relevance		Feasibility			Rank
		Target state	Business impact	Data	Analytics	Assessment (optional)	
required information	short description	issue addressed, analytics output, functionality and user interface of operative tool	high, medium, low	relevant data input	applicable analytics models	high, medium, low	relative ranking
responsible roles	BDA manager	Business user	Business user	Business user, data scientist	Data scientist	Data scientist	Project team (workshop)

Table 16 - Use case assessment template

In order to determine business impact, a target state is formulated for each use case covering the following dimensions: specific business issue addressed, required analytics output as well as functionality and user interface requirements of an operative tool. The target state substantiates the use case idea into a specific outcome and therefore supports the qualitative assessment of business impact in categories low, medium, and high. Target state and business impact together describe the *relevance* of the use case and business users of the use case prepare required information. *Feasibility* assessment takes data and analytics into account (Schmarzo 2013, pp. 140–143) which both build upon findings from the previous brainstorming. Responsibility for preparing the overview of applicable analytics models lies with the data scientist and potential data input is jointly prepared with business users. The goal here is not to fully specify data sources and analytics but rather to get a better understanding of availability of these two major resources. The BDA manager aggregates information in the assessment template and can

also propose further input on each dimension. Table 16 shows the assessment template including required information and responsibilities.

Based on the information collected on relevance and feasibility, use cases are ranked in order to define the ones to be considered for implementation (Shahapurkar 2016, p. 41). The ranking is compiled in a workshop with the entire project team under the guidance of the BDA manager. The purpose of the ranking is to jointly agree on a short list of use cases for final selection by the project sponsor. The workshop provides an interdisciplinary discourse to ensure all dimensions are considered in the prioritization effort. Optionally, using a graphical representation of the portfolio matrix can provide further support. This requires an additional step for translating the feasibility dimension into a categorical assessment, in exchange for an information representation conducive for decision-making workshops (Archer, Ghasemzadeh 1996, p. 17). Another advantage is that use cases can be represented as groups of business objectives which they address. The coverage of business objectives is an additional information not covered by plain ranking. It allows to consider different objectives when compiling the short list. Figure 40 provides a conceptual example of a portfolio matrix for BDA use cases.

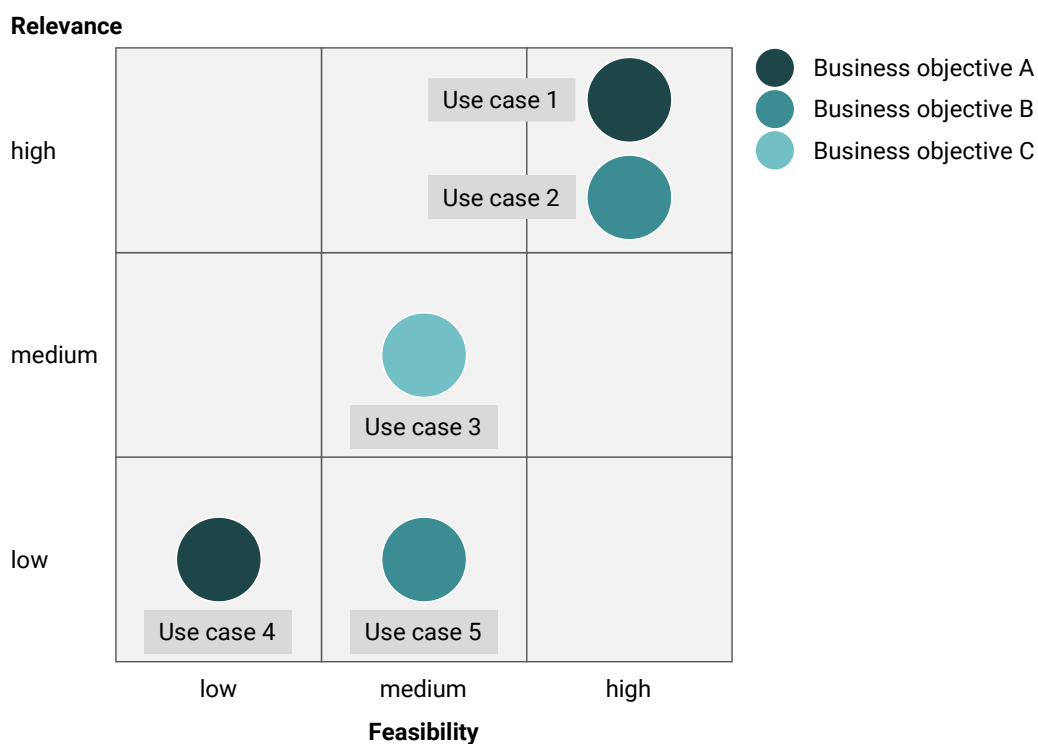


Figure 40 - Portfolio matrix for BDA use cases

4.3.3.3 Decision preparation and use case selection

The short list of use cases needs to be prepared for final selection. While relevance and feasibility information from prior assessment provide a solid foundation, additional information is required for decision making by the project sponsor. Furthermore, information needs to be presented at an adequate level of detail. The BDA manager therefore compiles a decision

template, shown in Table 17, which allows for a side-by-side comparison across shortlisted use cases.

Each use case description starts with an allocation to the underlying business objective and then further specifies the business issue or improvement potential addressed. *Objective & scope* furthermore describe the BDA approach in form of the analytics outcome from a business perspective. It therefore neglects technical details and puts focus on the analytics outcome in relation to the underlying business issue. Moreover, limitations to the scope, such as business segments or customer groups, are provided. The *project approach* describes whether the project builds on existing big data tools. It also clarifies whether the use case can be developed by an internal team or requires external partners. This is relevant as the use case potentially involves sensitive information such that selecting trustful partners is decision-relevant. The approach can be divided into different phases, for example, feasibility study, prototyping, and deployment. *Deliverables* describe the project outcome from the perspective of the project sponsor. In the case of different project phases, deliverables for each phase need to be defined. For example, proof of concept, working prototype, and operative tool with regard to previous mentioned phases.

	Use case 1	Use case n
Objective & scope	<ul style="list-style-type: none"> • <i>Business objective</i> • <i>Specific issue addressed</i> • <i>Big data analytics from business perspective</i> • <i>Scope limitations</i> 	...
Project approach	<ul style="list-style-type: none"> • <i>Use of existing big data tools</i> • <i>Use of external partners</i> • <i>Project phases</i> 	...
Deliverables	<ul style="list-style-type: none"> • <i>Final outcome of project</i> • <i>Intermediate outcomes (if applicable)</i> 	...
Project setup	<ul style="list-style-type: none"> • <i>Internal roles with active participation</i> • <i>Internal roles with passive participation</i> • <i>External roles</i> • <i>Effort estimate for active roles</i> 	...
Project budget	<ul style="list-style-type: none"> • <i>Estimate of project budget</i> 	...
	<i>Consideration of bundling</i>	
Comments	<ul style="list-style-type: none"> • <i>Information from use case assessment</i> • <i>Other decision-relevant information</i> 	...

Table 17 - Use case decision template

Project setup provides an overview of internal and external roles involved in the project. It is based on project team roles as defined in *Section 4.2*. Internal roles are divided into active and passive roles. While the former are involved in the project on an ongoing basis, the latter are only selectively required. An effort estimate for active roles indicates the workload on the organization posed by the project. The estimate of the financial *project budget* needs to cover all costs to conduct the project. Major cost items are required foundations, primarily technology in form of IT and software, as well as costs for external partners. In case of use of external partners including their foundations, their project offer defines the required budget. Use case bundling needs to be considered with regard to the project budget. Bundling represents the joint development of different use cases in a single project. The basic idea is to select use cases with similar characteristics in order to address multiple use cases at a lower total project budget. The following criteria provide a guideline for assessing similarity of use cases:

- 1) *Data*: Use cases build on similar types of data. In an ideal case, the data scope is the same or one use case utilizes a sub-group of data of the other.
- 2) *Project setup*: Use cases only slightly differ in required roles. They should ideally use the same active internal and external roles, if applicable. A common business objective of bundled use cases is a good measure here. In particular, the active role of business users is typically defined by the business objective.
- 3) *Analytics*: Use cases ideally only require different analytics models in case standard models can be applied. Joint use of existing big data tools is preferable.

The guideline provides criteria in descending order of importance. Data understanding and preparation typically account for the major share of analytics projects, up to more than 80% (Cios et al. 2007, p. 19), such that data congruence is the key criteria. It also has been shown that the team setup is crucial for project success and therefore the defined roles need to remain a good fit for the project. Assessing analytics provides a control criterion avoiding bundling of use cases where modeling represents a major share of the project. This is important as projects with similar effort for data and analytics work can be observed in practice as well (Hirji 2001, p. 92). The decision template provides all necessary information to decide about bundling of use cases.

Comments provide additional information for each use case. This can include selected information from use case assessment, for example, the overall ranking of the case. Other decision relevant information might point towards substantial changes regarding the presented use case descriptions. For instance, future availability of a new big data tool could indicate deferral of implementing affected use cases. After decision preparation concludes with the final decision template, use cases for development are selected. All roles involved in the preparation take part in the decision meeting with the project sponsor in order to answer detailed questions in their area of expertise. The final outcome is the decision about which use cases are developed. Information compiled in the decision template serves as basis for the project plan required for

each use case selected for development. Most important part next to the project team setup is a "[...] project roadmap with milestones and timelines" (Dutta, Bose 2015, p. 294). The project roadmap reflects the steps as well as tasks of the presented methodology, and includes identification of critical steps, decision points and review points (Chapman et al. 2000, pp. 41–42).

4.4 Big data sources

4.4.1 Step overview: Identify, assess & select

A fundamental idea behind the presented methodology is that "[...] organizations should start with a business problem first and then let the business problem lead to the right data" (Franks 2014, p. 35). Against the background of the big data hype, many organizations focus on collecting data instead of putting data to directed use (Franks 2014, p. 36). Shi-Nash, Hardoon (2017) caution companies to start big data analytics from existing data and solutions, but instead start with a specific objective. Moreover, companies need to find the right data, and if some data is not accessible or existent it is irrelevant as the objective drives the process (Shi-Nash, Hardoon 2017, pp. 339–340). Ohlhorst (2013, pp. 38–39) also proposes to use the objective of BDA utilization as starting point to find data and points out that external data sources are more difficult to identify. That is a reason for preferred selection of easily available data although relevance of data is more important than its availability (Rajpurohit 2013, p. 30). Based on the underlying business problem, useful data input can be formulated but usefulness of data varies by the type of problem as well as the underlying industry (Berry, Linoff 2004, pp. 60–61). Many analytics process models do not address the issue of data sources (Alnoukari 2012, pp. 187–188) and only 40% of managers report they have all required data according to a 2014 survey (Ransbotham et al. 2015, p. 64). As a consequence, there is a clear opportunity in providing support to find data (Fisher et al. 2012, p. 54). The challenges regarding data sources can be summarized by the following two questions (Phillips-Wren et al. 2015, p. 24):

"How should relevant data sources [...] for a given problem be identified before retrieval? What metrics should be used to identify relevant data sources for a problem?"

In order to address these challenges, the methodology dedicates one entire step to data sources as basis for big data input. Figure 41 provides an overview of the proposed data sources funnel. Input from the work on use case definitions during business understanding and a structured query for data sources result in a list of potential data sources. This input is gradually reduced into a short list based on filtering and pre-assessment. At last, a final assessment leads to the selected data sources as basis for big data input of the use case.

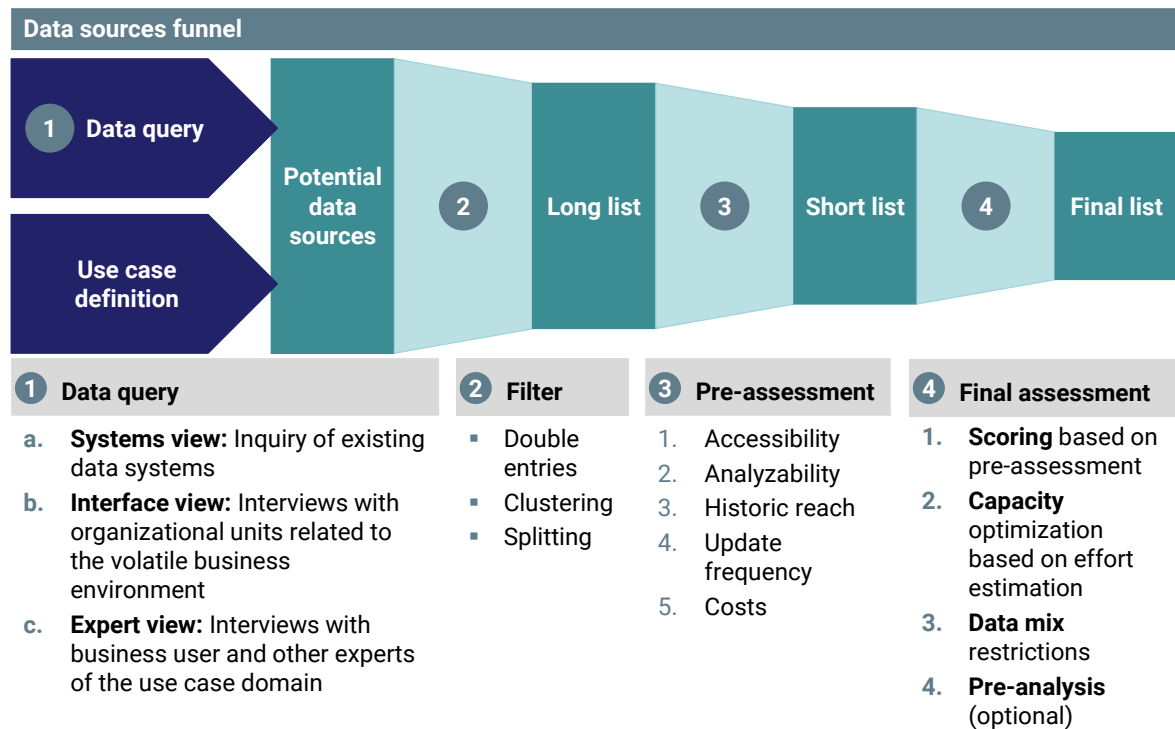


Figure 41 - Data sources funnel

4.4.2 Identify potential data sources

There exist three different types of data sources according to Pyle (2003). *External data* covers a broad range from meteorological data over financial markets to census information. *Existing internal data* is generated by the company and was collected for various business purposes. *Purpose-developed data* is specifically generated for the use in an analytics model and therefore is not directly available (Pyle 2003, pp. 221–225). As a consequence, the identification of data sources is limited to external and existing internal data.

Potential data input is already discussed during definition and assessment of use cases and hereby identified sources are fed into the list of potential data sources. However, a data query is used in order to assure deliberate identification of data sources. Due to the large variety of data sources (Davenport et al. 2012), this is essential in the case of big data. Approaches to data source identification are often limited to arbitrary lists of potential big data that is described in terms of data types but not sources (Elgendy, Elragal 2016, p. 1073; Berry, Linoff 2004, pp. 60–61). Missing to identify relevant data at the beginning of a BDA project can lead to inferior outcomes (Saltz, Shamshurin 2015, p. 2102) which underlines the need for a structured approach to identify data sources. Finding sources within a company is challenging, for instance, due to lack of documentation, such that it requires more than a single person to provide transparency (Berry, Linoff 2004, p. 62). The basic idea of the *data query* is to take three different perspectives in order to search for potential sources.

4.4.2.1 Systems view

The existing data sources of a company typically evolved historically and are characterized by isolated solutions which results in "[...] a lack of understanding what data exists and where it comes from" (Priebe, Markus 2015, p. 2056). The systems view seeks to provide transparency about the existing data landscape within the company. Therefore, a list of all existing systems that hold internal data is compiled and documentation that describes the content of the data system is collected where possible. Documentation allows to build an understanding of the data content of identified systems which is required to decide whether a system should be considered as data source for the project. Although each company has its individual data landscape, the following list provides a brief overview of typical systems:

- Enterprise Resource Planning (ERP) systems (Elragal 2014)
- Supply Chain Management (SCM) systems (Sun et al. 2015)
- Customer Relationship Management (CRM) systems (Sun et al. 2015)
- Supplier Relationship Management (SRM) systems
- Knowledge Management (KM) systems (Sun et al. 2015)
- Business Intelligence (BI) systems (Marin-Ortega et al. 2014)
- Communication systems (e.g., email)
- File sharing systems (e.g., Microsoft SharePoint)

Due to the individual character of the data landscape, support by the business user is required, especially in case of an external BDA manager. Furthermore, the creation of the systems view should be supported by organizational units familiar with the existing systems. Controlling, IT and finance departments are examples for typical system experts. There exist more internal data sources beyond the major systems. Data can be hidden in informal systems such as data collections on local hard drives or in small systems that are only known by specific users. For example, a member of the sales department might collect reports on customer visits and stores them on her computer. It is consequently not sufficient to focus on major systems when searching for potential data sources. Moreover, mainly driven by the increasing web access (Fraser 2017, pp. 356–357), external data is a valuable addition (Gentsch, Kulpa 2016, p. 36) for BDA applications. The following two views therefore focus on internal and external sources.

4.4.2.2 Interface view

The ultimate goal of BDA applications is to provide a better understanding of the volatile business environment. The second view thus puts focus on company interfaces with the volatile world. This outward oriented view also reflects the importance of external data sources. Interface view uses interviews with organizational units that are directly linked to the outside world. Relevant units again are company-specific and the BDA manager with aid of the business user needs to identify adequate interview partners. However, the following organizational units should provide helpful interview partners in most industrial companies: purchasing, supply chain management, sales, and to a lesser extent marketing & communications.

It is important to give an introduction about the project and its background to interview partners because they are not part of the regular project team. A project summary needs to be prepared by the BDA manager for this purpose. The summary is based on existing information from the business understanding step and should cover the following:

- Business context (1 page): Agility concept and derived business objectives
- Use case overview (1 page): Summary of use case definition and selection
- Use case details (1 page): Details on the use case covered by the current project based on decision template information
- Project plan (1 page): Overview of major steps and timeline
- Data source funnel (1-2 page): Procedure and motivation of the data source selection method

The discussions are set up in the style of semi-structured interviews and are performed by the BDA manager and business user. This form of interviews is focused on a specific topic and uses a list of questions in a flexible way that allows the interviewee to share own ideas (Edwards, Holland 2013, p. 29). The project summary provides the framework for this semi-structured approach and the following lists guiding questions:

- 1) What internal data systems do you use?
- 2) What data do you get from external business partners (e.g., customer, supplier)?
- 3) What data do you get from external providers (e.g., market research provider)?
- 4) What data would you like to use for the use case of this project disregarding current availability?

4.4.2.3 Expert view

The business user is an expert in the domain of the use case by definition. However, not all relevant experts are involved in the project such that their ideas on potential data sources are not considered during the business understanding step. The expert view of the data query fills in this gap and uses the same approach of interviews including project summary introduction as in the interface view. The following list summarizes guiding questions to be used in an expert view interview:

- 1) Who are additional experts for the use case of this project?
- 2) What data is currently used in the domain of the use case?
- 3) What data would you like to use for the use case of this project disregarding current availability?

The expert view has a second, less formal dimension. The business user is involved in all efforts of the different views and his or her participation can therefore be seen as ongoing brainstorming for data sources.

4.4.3 Filter and pre-assessment

4.4.3.1 Filter

The result of the data query is a list of potential data sources in an informal way. This list needs to be transferred into a structured long list of data source options. It is the responsibility of the BDA manager to perform this filtering task. The long list should describe each data source along multiple dimensions including a description of data content, source type (internal vs. external), and data type (structured vs. unstructured) (Marr 2015, p. 100). In addition, information on the data source collected during the data query is documented as well. This information serves as basis for the subsequent pre-assessment. Designation of data owners is also listed because their input is also required for the assessment. During processing of the data query outcome, the BDA manager removes *double entries* and integrates overlapping ideas for data sources into *clusters*. Overlap often occurs when data sources provide the same information. For example, different interviews result in the ideas to use revenue data of customer companies from an external financial Database A and analyst estimates from a different Database B. In the case of Database A including both data, the ideas are represented as one data source in the long list. However, filtering can also require to *split* a data source in case it contains data with strongly different characteristics which does not allow to assess the source as a whole. For example, a data source can contain structured and unstructured data, and if both are relevant, they are treated as individual sources in the long list.

4.4.3.2 Pre-assessment

In order to reduce the number of potential data sources the long list is reduced to a short list by applying a set of decisive factors. These criteria represent essential requirements and are used to eliminate alternatives in an assessment (Heinrich et al. 2014, p. 403). If the requirement of any decisive factor is not met, the data source is not considered for final selection. The following five factors are considered:

- 1) *Accessibility*: Technical accessibility of data sources must be ensured (Theobald, Föhl 2015, p. 118), however, is not sufficient. Actual access can also be limited due to entitled use of proprietary data sources as well as legal or contractual restrictions to the use of the data.
- 2) *Analyzability*: Available analytics capabilities must be sufficient to analyze the data provided by each source. This also includes minimum requirements to the structure of the data such that they are amenable for analysis (Theobald, Föhl 2015, p. 116).
- 3) *Historic reach*: Analytics commonly use historic data for modeling (Berry, Linoff 2004, p. 63) such that each data source must provide sufficient data history.
- 4) *Update frequency*: Velocity is a key characteristic of big data and can be interpreted as "[...] frequency of data generation [...]" (Russom 2011, p. 7). The nature of analytics in the context of the volatile world require constant provision of current data and therefore the data source must meet a minimum frequency for updating data.
- 5) *Costs*: Capturing data from data sources comes at a cost (Marr 2015, p. 100). Internal sources potentially require specialists to extract data from a system. In addition to sourcing costs, licensing costs might incur for external sources. These costs address all expenses beyond data capturing efforts covered by the project team.

Accessibility and analyzability are digital decisive factors. Access to a data source is given or not and its data can be utilized for analytics or not. The remaining three factors represent maximum or minimum criteria, respectively. The costs of data capturing must not exceed a certain budget entitled to data input. Each data source needs to provide a minimum historic reach and update frequency. The maximum and minimum levels for these factors need to be specified based on the individual use case. In addition to BDA manager and business user, the data scientist is required for conducting pre-assessment. Data scientist expertise is required to decide on the digital criteria as well as for setting the levels in case of historic reach and update frequency.

4.4.4 Final assessment and selection

4.4.4.1 Final assessment

Final assessment adds crucial information about the data sources of the short list in preparation of final selection. Scoring and effort estimation for shortlisted data sources provides such additional information. The aim of the *scoring* is to provide an indication about relative eligibility for the project. It uses a scoring model approach where the score of each data source is derived from an aggregation of multiple criteria (Archer, Ghasemzadeh 1999, p. 210). All three decisive factors with maximum or minimum level serve as criteria. Starting from these levels, ranges are defined that correspond with a certain score value. The overall score is calculated as weighted sum of these values. The ranges and weights are use case specific and therefore require definition by the BDA manager. However, historic reach is typically the most important factor due to its relevance for modeling. Similarly, update frequency can be assumed to have a higher weight than costs. Figure 42 provides an overview of this data source scoring model.

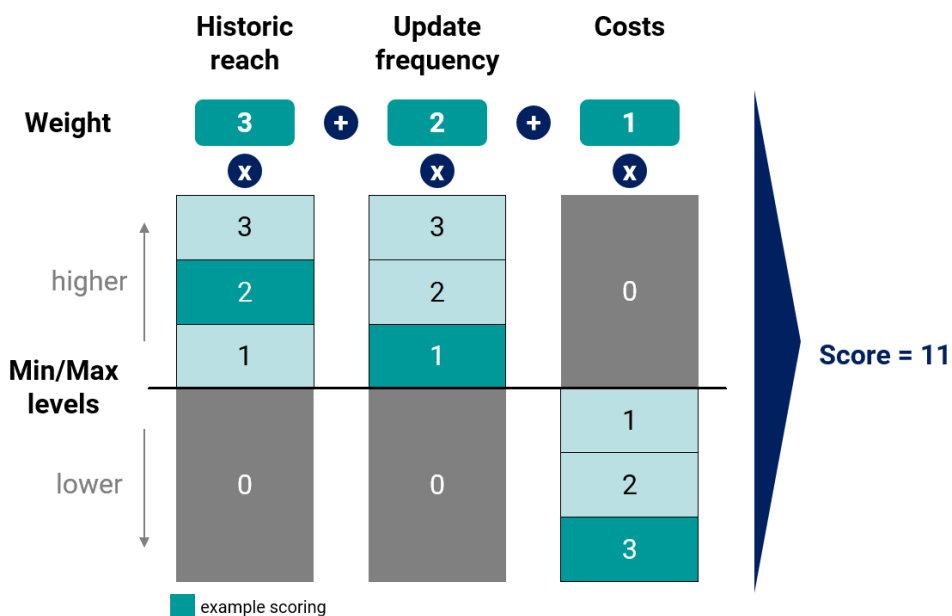


Figure 42 - Data source scoring model for final assessment

The project plan defines the available *capacity* of data scientist, database administrator, and data engineer. The final selection of data sources must adhere to this capacity. The data scientist therefore needs to estimate the effort for data sourcing, data preparation and modeling. The total effort to utilize the final list of data sources must not exceed the given capacity.

4.4.4.2 Data source selection

The project team represented by BDA manager, business user, and data scientist establishes the final list building upon the additional information from final assessment. While this information is helpful and necessary, the ultimate goal is to make a deliberate choice of data sources from a business perspective. Selected data sources must represent the most valuable data input for business users under present conditions such as effort limitations. The data source selection workshop is furthermore guided by the fact that big data analytics benefits from a data input mix (Franks 2012, p. 22). Integration of internal and external data is seen as improvement potential for various types of analytics (Nisbet et al. 2009, pp. 26–28; Mehanna et al. 2016, p. 506). The same applies to combining structured with unstructured data (Henke et al. 2016, p. 70; Davenport, Dyché 2013, p. 3). It is often assumed that around 80% of business information exists in the form of unstructured data (Grimes 2008). At the same time, surveys among companies indicate that they still dominantly use data from internal systems (Bange et al. 2015, p. 32; Gronau et al. 2016, p. 477) and that only a minority of companies already utilizes unstructured data in the case of predictive analytics (Halper 2014, p. 10). In order to ensure a balanced *data mix*, the portfolio matrix technique is used in the workshop for data source selection. The matrix represents both data mix dimensions and thus enables a balanced mix (Archer, Ghasemzadeh 1996, p. 17) of data sources. Effort estimations and scores can also be integrated in the form of bubble sizes and colors, respectively, to provide further graphical guidance. A conceptual version of the data mix matrix is shown in Figure 43.

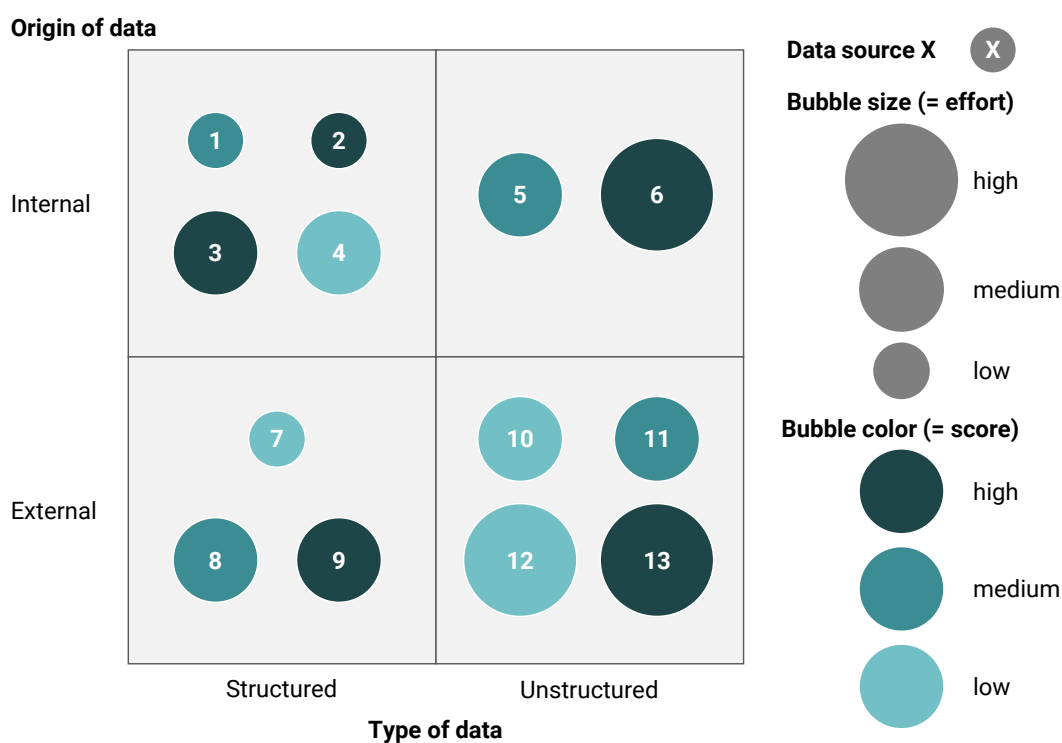


Figure 43 - Data mix matrix

Relevance of the data in a source is another potential factor to consider in the selection. It describes whether data is suitable for the desired analytics outcome (Lavrac et al. 2004, p. 21; Theobald, Föhl 2015, p. 115). A *pre-analysis* of the data can therefore provide additional guidance for the selection. However, there are various prerequisites in order to perform pre-analysis. Data must be available without additional sourcing efforts and should not require extensive data preparation. The type of analysis also needs to build on readily available analytics models. As a consequence, testing relevance of data is an optional step, especially since external data usually does not meet the prerequisites. The big data input step concludes with the final list of data sources that provides the basis for big data input.

4.5 Data understanding

4.5.1 Step overview: Select, source & describe

4.5.1.1 Tasks overview

The overall goal of the data understanding and data preparation steps is to provide processed *data input for modeling*. For that purpose, the initial task is to define the scope of data by selecting relevant datasets from the *big data sources*. These datasets subsequently are prepared for sourcing before the actual sourcing task results in a *project cluster*. The project cluster represents a 'big data warehouse' for the use case. It contains all relevant data and also builds the basis for data preparation. Description of data including clarification of open questions starts with sourcing preparation before data exploration and verification complement the understanding of data. The entire step is supported by a novel tool called *BDA book* which represents a repository of information about the data. It is the major outcome of the data understanding step, next to the project cluster, because it provides necessary information for subsequent tasks. Dataset selection, sourcing preparation, and data sourcing correspond to the *data collection* and *data description* tasks of CRISP-DM, while the remainder represents a substantiated form of *data exploration and verification* tasks (Chapman et al. 2000, pp. 43–47). Figure 44 provides an overview of the data understanding step including an outlook on data preparation. It is important to note that the methodology assumes structured data in the form of time series from this point onwards.

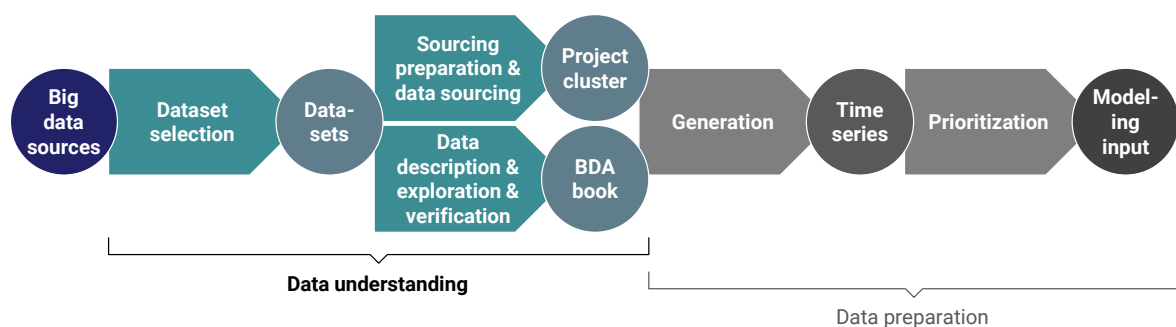


Figure 44 - Overview data understanding step

4.5.1.2 Data hierarchy

The proposed methodology operates at different levels of detail regarding 'data'. Identification and selection of data sources proceeds on the highest level of the data hierarchy. In order to facilitate the understanding of the following tasks, Figure 45 presents an overview of the data hierarchy applied by the methodology including relations to the tasks of the data understanding and data preparation steps.

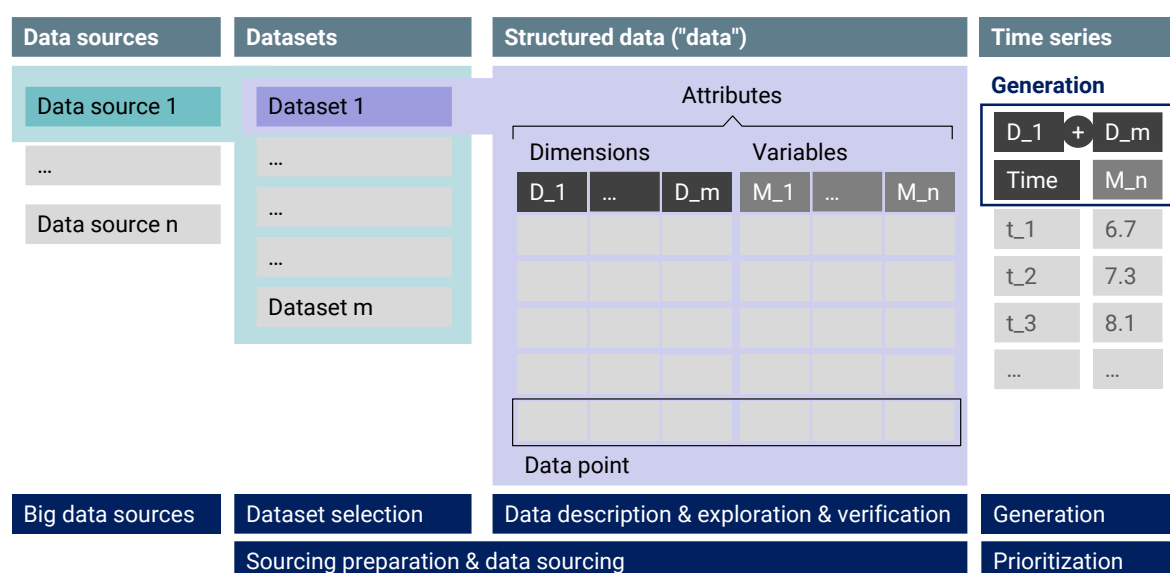


Figure 45 - Data hierarchy taxonomy

Each *data source* comprises multiple *datasets* that contain different types of information. In the case of *structured data*, information within a dataset is organized in tables (Müller, Lenz 2013, p. 78) where each row represents a *data point* and each column corresponds to an *attribute* (Han et al. 2012, p. 40). The totality of all data points build the dataset and its attributes describe different features (Han et al. 2012, p. 40). Within this work, two general types of attributes are defined. *Dimensions* represent nominal attributes which include information in form of "[...] some kind of category, code, or state, and so nominal attributes are also referred to as categorical" (Han et al. 2012, p. 41). A *variable* represents a numeric attribute that "[...] is a measurable quantity, represented in integer or real values" (Han et al. 2012, p. 43). *Time series* are the lowest level of the hierarchy and they are generated from the structured data of a dataset. The methods of time series generation and prioritization will be discussed in detail in *Section 4.6*.

4.5.2 Dataset selection & sourcing

4.5.2.1 Outline

In general, there are three types of datasets that require different approaches for dataset selection & sourcing. A data source typically contains only one type of dataset. *Fixed datasets* do not allow for major changes prior to sourcing and therefore require little additional sourcing preparation. The structure of data can be changed by deleting or adding dimensions and variables for *customizable datasets*. Furthermore, custom subsets can be defined by filtering for certain dimensions, for example. It can also be the case that a data source is organized in distinct datasets but does not provide sourcing of entire sets. As a consequence, definition of the structure is a prerequisite for selection and sourcing of datasets in this case. While internal data usually occurs in the form of customizable datasets, both types can be found for external data. In case internal data contains sensitive information, sourcing preparation potentially needs to

be extended by adequate measures. The approach of data selection & sourcing for such *sensitive datasets* is outlined in Figure 46. It also includes an overview of team role responsibilities because six out of eight roles are involved across a multitude of tasks and subtasks here. Data selection & sourcing for the other two dataset types can be simplified. For example, standard customizable datasets do not require involvement of a data officer and fixed datasets furthermore do not require to define a detailed sourcing structure. The remainder of this subsection describes the case of sensitive datasets.

■ Role with responsibilities

		BDA manager	Business user	Data owner	Data officer	Database administrator	Data scientist
Dataset selection	Preparation	■		■			
	Selection	■	■				
Sourcing preparation	Description & clarification	■	■	■	■		
	Sourcing structure	■	■			■	
	Test download					■	
	Clearance	■			■		■
Data sourcing	Extract					■	
	Transfer	■				■	
	Transform					■	
	Load					■	

Figure 46 - Data selection & sourcing for sensitive datasets

4.5.2.2 Dataset selection

Data sources selected for the use case usually comprise a large number of datasets. While more volume of data is a key characteristic of big data, having more data "[...] doesn't add any value by itself" (Franks 2012, p. 6). Davenport, Dyché (2013, p. 2) state that variety of data is more important than the ability to handle vast datasets. It is therefore important to select relevant datasets (Sparks et al. 2016, p. 36) but selection is also required due to "[...] technical constraints such as limits on data volume [...]" (Chapman et al. 2000, p. 24). Armstrong et al. (2015, p. 1719) underline the importance to focus on relevant data in the case of forecasting and propose to leverage experts in order to identify such data. As a consequence, relevant datasets are selected with the help of the business user in the project team.

In order to *prepare* selection, the BDA manager needs to compile an overview of available datasets. This overview can be drawn from relevant documentation about the data sources. Alternatively, this information can be provided by the data owner directly or indirectly by granting access to the data source for the BDA manager. Furthermore, information collected

during the big data sources step can be valuable here as well. The goal of this preparatory work is to establish distinct datasets and to describe each of them by its key dimensions and variables. During decision preparation, the BDA manager is also responsible to preselect datasets. Utilizing the business understanding of the BDA manager is especially important for data sources with a large number of datasets. Actual *selection* of datasets is based on a review of all remaining datasets per source based on the dataset selection sheet as shown in Table 18. The task for the business user is less a matter of deciding whether a dataset is relevant for the use case based on analytical measures, because this will be ultimately assessed during time series prioritization. It is more about eliminating datasets that seem not to be relevant from a business perspective.

Data source A			
Dataset	Key dimensions	Key variables	Selection
Dataset 1	<ul style="list-style-type: none"> • Dimension 1 • Dimension 2 • ... 	<ul style="list-style-type: none"> • Variable 1 • Variable 2 • ... 	yes/no

Table 18 - Dataset selection sheet

In case the selection decision is not obvious for the business user based on the brief information provided, the BDA manager should lead the discussion towards focusing on the idea of correlation. Drawing insights from data based on correlations is a fundamental BDA concept and in particular in the case of predictive analytics (Provost, Fawcett 2013a, pp. 56–57). Assuming that the dataset provides correlations, the business user needs to decide whether this information should be used or still be disregarded for the use case. In case the business user does not believe in the data input of the model, there is no reason to use this data. However, if the discussion is inconclusive the dataset should be included - *if in doubt, leave it in* - because datasets will be assessed with regard to correlations more rigorously during data preparation.

4.5.2.3 Sourcing preparation

With data selection concluded, the scope for data sourcing is finalized. Sourcing preparation describes required tasks before data is actually sourced and it begins with *description & clarification*. Description of datasets represents the collection of selected metadata. Metadata, in the sense of superordinate information about data, plays a crucial role since the start of information processing but is particularly important for large data repositories such as data warehouses (Inmon 2005, pp. 102–103). Hofmann, Tierney (2009, p. 67) define²⁴ metadata as information

²⁴ The definition is based on the work of Klösigen (2002).

"[...] on the semantic, structural, statistical, and physical level in order to support tasks such as data validation and imputation, selection and application of [analytics], and interpretation of the results". According to Kimball, Caserta (2004), there exist three categories of metadata relevant for sourcing (Kimball, Caserta 2004, pp. 380–381):

- 1) *Business metadata* describes data from a business point of view and includes data source information, attribute information, business definitions, and tracking of alterations of data during the sourcing process.
- 2) *Technical metadata* represents technical aspects, for example, data types.
- 3) *Process execution data* provides statistics about the sourcing process, for example, number of rows loaded.

There is a need to store this information in a metadata repository in order to facilitate navigation through the datasets and information sharing in the project team (Pant 2009). Dataset descriptions focus on business metadata under the lead of the BDA manager while the database administrator is responsible for documenting technical and process execution metadata during data sourcing. A repository for business metadata in case of data warehousing is typically called *data dictionary* (Kimball, Caserta 2004, pp. 361–363), however the term is also used for repositories restricted to business definitions (Soares 2011, pp. 47–54). Most importantly, the concept of a data dictionary is also proposed as supporting tool for big data analytics (Anand et al. 2007, pp. 41–45; Lanquillon, Mallow 2015a, p. 78). A new version of a data dictionary is introduced for the use within this methodology and that takes into account the presence of a BDA manager with business knowledge. As this tool will be extended in subsequent tasks and in order to avoid misunderstanding, it is termed *BDA book*. Preparation of the BDA book is a key responsibility of the BDA manager and follows three consecutive stages:

- 1) *BDA book setup*: The BDA manager determines the required structure of the BDA book and enters all existing information. This information mainly stems from collected documentation and preparatory work during the dataset selection task. The format of the BDA book is not restricted but it should allow access for all project team members. The BDA book contains an overview sheet for each data source and detailed sheets for all datasets within the source. While the structure can vary between individual use cases, Figure 47 and Figure 48 provide a general structure summarizing minimum requirements. The *overview sheet* provides an outline across all datasets of a single data source. It also includes a tracker to monitor progress of sourcing preparation and subsequent data sourcing. The use of a tracker is valuable not only because of the large number of datasets to source but also due to the multitude of roles to be coordinated in this stage. Moreover, the overview sheet lists all data owners required to cover all datasets. The *dataset sheet* comprises the business metadata and reflects the structure of the data including grouping of attributes into dimensions and variables. Figure 48 also

summarizes key information collected on attributes and provides explanations for each type of information.

- 2) *Identify clarification need*: Structure of the data as well as name and header usually do not require further clarification. Identification of timestamp dimensions can also be done independently by the BDA manager. Need for clarification regarding units exists in case documentation does not provide sufficient information. The most extensive clarification need typically exists for explanations. However, the BDA book must not be seen as documentation means but rather as tool for the BDA manager to successfully guide the development of an analytics application for the use case. That is to say not every single attribute needs an explanation in the BDA book such that the focus lies on attributes that cannot be explained by the BDA manager directly, for examples, towards the data scientist or data engineer. Finally, the BDA manager checks data against reasons for sensitivity, such as terms and conditions of an external data source or sensitive personal information within an internal data source (Terrizzano et al. 2015, p. 4), and flags potentially sensitive data.
- 3) *Establish & document clarification*: The BDA manager collects information from data owners and business users in order to address clarification needs for explanations and units. Due to the multitude of data owners across all datasets, the tracker of the BDA book also monitors open questions including the person responsible to provide clarification. In the dataset sheet, the BDA manager documents all provided answers. The status of sensitivity is determined with the help of the data officer. In case sensitivity is confirmed, the flag remains in the BDA book in order to ensure adequate measures. Furthermore, the BDA manager together with the data officer defines what measure needs to be taken, for example, deletion of sensitive personal information such as employee names or anonymization of customer company names.

A selected data source potentially contains a large number of datasets such that preparation of individual dataset sheets becomes burdensome. In this case, groups of datasets with the same basic attribute structure are identified. A *meta dataset sheet* describes the overall structure of the data source including all existing dimensions and variables. A *mapping sheet* describes the subset of attributes related to each group of similar datasets and documents any further specifics of each group. There is no change in information scope or preparation approach when using these two extensions of the BDA book. They represent an efficient way to manage a large number of datasets. The BDA manager is responsible to specifically build them for a given use case because the structure of both sheets strongly depends on the characteristics of the underlying data source.

Data source A						
Dataset description				Tracker		
Name	Description	Information	Data owner	Clarification		Open questions
<ul style="list-style-type: none"> Dataset label 	<ul style="list-style-type: none"> Brief summary of data 	<ul style="list-style-type: none"> Historic range Frequency Other relevant information on data 	<ul style="list-style-type: none"> Name Contact details 	Ready for test	■ ■ ■	<ul style="list-style-type: none"> List of open questions Including responsible for final clarification
				Original test download	■ ■ ■	
				Anonymized ¹ test download	■ ■ ■	
				Legal clearance	■ ■ ■	Tracker logic
				Analytics clearance	■ ■ ■	<ul style="list-style-type: none"> not started in progress finished
				Extraction	■ ■ ■	
				Data anonymization ¹	■ ■ ■	
				Data transfer	■ ■ ■	
				Transform & load	■ ■ ■	

Information for each dataset of the data source

¹ Anonymization as example measure for sensitive data

Figure 47 - BDA book (overview sheet)

Dataset A.1							
Information	Dimensions				Variables		
Name	<ul style="list-style-type: none"> Attribute name from original data source 						
Header	<ul style="list-style-type: none"> Translation of the attribute name into attribute label used on the project cluster Required as attribute names may not be unique and this allows translation into more meaningful names 						
Timestamp	<ul style="list-style-type: none"> Flag for time-related dimensions suitable for time series generation 						
Units	<ul style="list-style-type: none"> Unit information for variables (e.g., currencies) Only required for variables without unit information given in according dimensions 						
Explanations	<ul style="list-style-type: none"> Business information on attributes (only where required) Provided in form of brief definitions or examples 						
Sensitivity	<ul style="list-style-type: none"> Flag for sensitive data Measure to handle sensitive data (if applicable) 						

Focus of clarification

Figure 48 - BDA book (dataset sheet)

The established BDA book builds the basis for all further tasks in data understanding and data preparation. In order to extract selected datasets from the data sources, the next task is to define the *sourcing structure*. In general, extraction depends on the data source type and its means of access. A data source can provide data in different formats such as database tables or data files (Kabiri, Chiadmi 2013, p. 219). There also exist different forms of access to the data of a source

including bulk downloading a multitude of files or *Application Programming Interface (API)* access (Koumenides et al. 2010). While the database administrator is responsible for technical implementation of data extraction, the sourcing structure as conceptual representation is jointly defined with the BDA manager. It defines necessary characteristics of data extraction in order to define "[...] the correct subset of source data that has to be submitted to the [sourcing] workflow [...]" (Vassiliadis, Simitsis 2009, p. 1096) and includes the following information:

- 1) *Scope*: List of all required attributes according to the BDA book.
- 2) *Data history*: As not all data sources have scarce data histories the sourcing should be limited to a certain historic range. As a minimum requirement for a forecasting use case, data history should cover the range of the time series to be forecasted, plus the forecast horizon.
- 3) *Frequency*: In case a data source does not include continuous data but rather consecutive snapshots, extraction must be specified in order to generate a suitable frequency of data. A typical example here is order backlog that is recorded at regular intervals by the ERP system. The sourcing structure needs to define at which frequency these records are extracted.
- 4) *Timestamp*: For datasets that do not provide sufficient time information for generation of time series, this information needs to be added during extraction. In the example above, the time and date information from order backlog records extracted needs to be added to the actual order data sourced for each record.
- 5) *Attribute splits*: Attributes of a dataset can include multiple information that requires separation into new distinct attributes. Variables including unit information represent a major case where splitting is required to ensure easier processing in later stages.
- 6) *Attribute filter*: Datasets can be very extensive in case dimensions have a large number of instances. Following the same principle as for dataset selection, excluding dimensions the business user would not include in the analysis is beneficial. It is important to note that filtering is not about identifying best or relatively better instances of a dimension. Instances that are invaluable from a business perspective are excluded from datasets. The BDA manager must identify dimensions that are potential candidates for filtering. With the help of the business user, the filter is subsequently specified as part for the sourcing structure.

The BDA book also serves as documentation for the sourcing structure in case of high complexity. For example, if a data source requires to define API queries, the BDA book is extended by a *sourcing sheet* that explains the structure of these queries. The structure of sourcing sheets highly depends on the individual data source and must be defined by the BDA manager together with the database administrator. Once the sourcing structure is established, the database administrator can translate it into source-specific extraction tools. With the help of

these tools, a *test download* representing a brief sample of data can be provided. The test download is limited in its historic range to keep data volumes small and builds the basis for subsequent *clearance*. Data sourcing represents a substantial effort (Kabiri, Chiadmi 2013, p. 220) and "[...] creation of a valid database is the first and most important operation that must be carried out in order to obtain useful information [...]" (Giudici 2003, p. 20). *Functional clearance* by the database administrator therefore ensures correct working of sourcing in form of a test run. All measures for sensitive data that are codified in the BDA book are applied to the test downloads, which are then presented to the data officer in order to gain *legal clearance*. In *analytics clearance*, the data scientist reviews the test downloads with regard to the use of sourced data for analytics. With all clearances successfully completed, data sourcing can start but otherwise the sourcing structure or measures for sensitive data need to be revised. There is no need for all data sources to be cleared at the same time and each source can independently move into the next task once sourcing preparation is completed. In the same sense, test downloads and clearance can take place while clarification of the dataset is not fully finalized. This especially holds true for clarifications of explanations where answers can be collected in parallel to subsequent tasks. The status *ready for test* in the BDA book tracker (see Figure 47) supports this approach as it indicates finalization of the sourcing structure. Structuring dataset selection & sourcing in clearly defined task and subtasks including responsibilities allows the BDA manager to ensure efficiency to a best possible extent.

4.5.2.4 Data sourcing

Data sourcing follows the *Extract, Transform and Load (ETL)* principle that includes *extraction* from the data sources, *transformation* into the target data structure, and *loading* into the data repository (Vassiliadis, Simitsis 2009, p. 1095). The process design and implementation are responsibilities of the database administrator. This subsection provides a brief overview pointing towards particularities of the methodology. More detailed explanations including technical details and applicable tools can be found in Kimball, Caserta (2004), Inmon et al. (2010), Ponniah (2010), Krishnan (2013), and Reeves (2009).

Extraction is already formulated during sourcing preparation by the defined sourcing structure and its implementation as extraction tools. In contrast to test downloads, datasets are extracted without additional limitation during data sourcing. Extraction is usually divided into initial extraction as one-time population of the data warehouse and incremental extraction for updates of changed data (El-Sappagh et al. 2011, p. 93). The ETL process is restricted to initial extraction as it provides all data required for modeling. Implementation of incremental extraction is required for potential deployment of a developed solution and not in scope of this methodology. An introduction to extraction of changed data can be found in Kimball, Caserta (2004, pp. 105–112). Moreover, prescribed measures for sensitive data need to be applied where applicable. Owing to the circumstance that external big data infrastructure can be used, *transfer* as additional task between extraction and transformation is required. In order to minimize transfer efforts,

all external data should be extracted directly by the external provider outside of the company. The means of transfer for internal data are selected by the company and the BDA manager conducts a physical transfer in case no other option of secure transfer is suitable. Transformation brings data into the target data structure which is derived from the conceptual data structure as defined by the dataset sheet of the BDA book. The target data structure represents the view from the data repository perspective. Fundamental idea of loading in this methodology is to integrate all data in a data repository dedicated to the use case. This project data repository resembles a traditional data warehouse, in the sense that it provides the basis for information processing by providing integrated and granular data (Inmon et al. 2010, p. 7): integration refers to the holistic view of all available data and granularity describes the flexible use of data due to its high level of detail. A data warehouse usually serves more than one application and the data repository can be integrated into more general architectures such as corporate information factories (Kimball, Ross 2013, pp. 26–30). The focus lies more on BDA readiness than integration of different applications here. As conventional approaches do not meet requirements of storing and analyzing big data, Hadoop is proposed as framework for data warehouses (Kimball, Ross 2013, pp. 528–531; Krishnan 2013, p. 230). The methodology suggests to load data into a Hadoop-based cluster. Cluster refers to the fact that the hardware of the data repository can be built on commodity hardware and therefore enables sufficient performance in a cost-effective way (White 2015, p. 284).

4.5.3 Data exploration and verification

After completion of data sourcing, all data is available for further tasks. Data exploration and verification conclude data understanding and a detailed description can be found in Chapman et al. (2000). As the presented methodology is based on a use case approach with clear formulation of analytics need and also provides a comprehensive method for identifying interesting data subsets (see *Section 4.6*), data exploration can be reduced to a review of attribute characteristics in support of data verification (Chapman et al. 2000, p. 45). Also the handling of missing values is integrated into subsequent tasks, such that data verification is limited to assessment of data completeness and correctness (Chapman et al. 2000, pp. 46–47). Moreover, these tasks must not strictly be seen as preceding step to time series generation. Whenever there is a need to explore or verify data in order to assist time series generation, the process iterates back to these tasks. The BDA book is extended to support managing these tasks. The BDA manager can add *correctness checks* for selected attributes that allows the data engineer to verify data. Examples for correctness checks are non-negative values or simple relations of the form: $gross\ profit\ margin = (revenues - costs\ of\ goods\ sold) / revenues$. *Exploration information* documents the results of attribute reviews which can be checks for completeness, for example. Most importantly, the BDA book also indicates whether an attribute fails a check or review by setting the *quality flag*. These attributes are then no longer considered as modeling input. While the data engineer is responsible to carry out these tasks on the cluster, the BDA manager defines checks

and reviews as well as decides on the quality flag. Figure 49 summarizes the extension of the BDA book for data exploration and verification.

Dataset A.1							
Information	Dimensions				Variables		
Previous information	<ul style="list-style-type: none"> Name, Header, Timestamp Units, Explanations, Sensitivity 						
Correctness checks	<ul style="list-style-type: none"> Check logics for attributes based on business knowledge 						
Exploration information	<ul style="list-style-type: none"> Tracking of information on attributes gained during exploration 						
Issue flag	<ul style="list-style-type: none"> Flag for attributes to be excluded from analytics due to failed correctness check or indication by exploration information 						
Additions for exploration and verification							

Figure 49 - BDA book (extended dataset sheet)

4.6 Data preparation

4.6.1 Step overview: Generate & prioritize

4.6.1.1 Task overview

In order to provide model input for time series-based analytics, time series need to be generated from the datasets and prioritized afterwards. The proposed methods for time series generation and time series prioritization represent the step of data preparation as depicted in Figure 50.

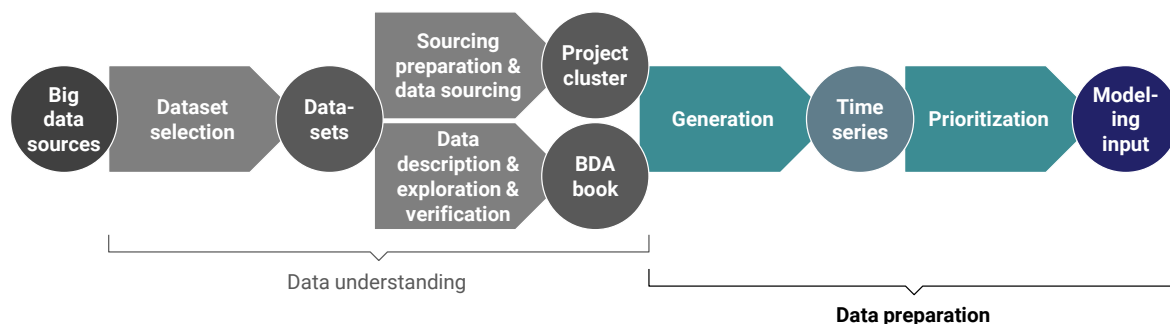


Figure 50 - Overview data preparation step

According to Chapman et al. (2000) the following tasks are involved in CRISP-DM. *Select data* ultimately defines which data is used for modeling based on relevance and quality. *Clean data* describes measures to ensure a data quality level adequate for modeling. *Construct data* includes the creation of derived attributes or changed values for existent attributes, whereas *integrate data* refers to the creation of new attributes or values by combining data from multiples datasets. *Format data* finally modifies data for use in the analytics model of choice while keeping its original meaning (Chapman et al. 2000, pp. 48–52). Data preparation in the proposed methodology consists of two major tasks. *Time series generation* and *time series prioritization* integrate four of the previously described preparation tasks. Time series generation represents data construction and additionally includes selection of data as well as integration of data where applicable. The primary goal of time series prioritization is to select data and the method includes the provision of cleaned data. Because data formatting is generally dependent of the model selected (Han et al. 2012, 112–119; Cleve, Lämmel 2014, pp. 209–215), this task is attributed to model building.

4.6.1.2 Time series data

Data in the project data repository exists in a structured format described by its dimensions and variables. Timestamp information as particular form of dimension allows to convert this data into multiple time series. Time series generally are a representation of data in time sequence that is still independent from the analytics model applied to them (Evans 2003, p. 30). The basic idea of time series generation is to build all possible combinations of dimensions for each variable in a dataset and to use timestamp information defining the time sequence. Figure 45 in *Subsection 4.5.1* conceptually depicts this approach and the relation to higher levels of the data hierarchy.

4.6.2 Time series generation

4.6.2.1 Knowledge & dimensionality test

The large number of dimensions available for time series generation potentially leads to high-dimensional data which poses a challenge for computational processing (Bolon-Canedo et al. 2015, p. 2). Despite the continuous increase of cost-effective computing technology this remains a challenge for analytics (Yang et al. 2015, p. 2) and in this case it is necessary to reduce dimensionality (Destrero et al. 2009, p. 26). Moreover, data construction, integration and selection provide an opportunity to leverage domain experts (Guyon, Elisseeff 2003, p. 1170; Peng, Kou 2008, p. 48). The proposed method for time series generation thus considers both the computational challenge and the opportunity to incorporate domain knowledge. It differentiates between two different approaches out of which *hypothesis-based generation* leverages domain knowledge while controlling dimensionality for each hypothesis. In contrast, *automated generation* uses a single default method to create dimension combinations and typically does not create computational issues due to high dimensionality for the entire data source.

The *dimensionality & knowledge test* helps to decide which approach should be applied in what form. Decision about the approach is taken on the data source level in order to ensure consistency for related datasets. The test consists of the following two questions:

- 1) *Knowledge*: Does domain knowledge in the project team allow for its use in time series generation?
- 2) *Dimensionality*: Which types of dimension combinations for time series generation lead to dimensionality that are too high for available data processing capabilities?

Answering the first question is the responsibility of the BDA manager and requires a qualitative assessment of the feasibility to select meaningful subsets of dimensions with reasonable effort. In case the condition is met, the hypothesis-based approach is employed and automated generation otherwise. The answer to the second question requires exploration of dimensions within each data source. It is difficult to define an exact limit for tolerable dimensionality as this strongly depends on the specific conditions of the individual project. The data engineer is responsible to provide an answer for each data source, which defines types of dimension combinations feasible for time series generation. For example, in case utilization of all possible combinations of available dimensions, the so-called power set, leads to an intolerable level of dimensionality, the time series generation approach must be restricted to less extensive types of dimension combinations.

4.6.2.2 Hypothesis-based generation

Selecting a subset of dimensions is also known as attribute selection and removing irrelevant attributes is an important issue in analytics practice according to Witten et al. (2011). Although advanced models strive to select best attributes themselves, practical experience indicates performance improvements due to preselection of attributes (Witten et al. 2011, p. 306). For that reason, it is not only preferable but imperative to reduce dimensions where possible. Applying domain knowledge for this purpose is seen as effective opportunity (Ahlemeyer-Stubbe, Coleman 2014, p. 96). In analytics, "[a] hypothesis is a proposed explanation whose validity can be tested by analyzing data" (Berry, Linoff 2004, p. 50). While hypotheses are often used to define valuable analytics for an organization (EMC Education Services 2015, p. 35), domain knowledge-based hypotheses can also be used for attribute selection (Kopanas et al. 2002, p. 293). This idea underlies hypothesis-based generation of time series that follows three basic steps:

- 1) *Formulate*: BDA manager and business user formulate hypotheses for each dataset based on their domain knowledge.
- 2) *Transfer*: The BDA manager transfers the hypothesis into the data structure as defined by the BDA book and compiles all relevant information for generation of time series representing the hypothesis.
- 3) *Generate*: The data engineer implements time series generation on the project cluster based on the given information.

In order to facilitate this process, the BDA book is extended with the addition of a *time series generator sheet* for each dataset. Documentation approaches of data transformations in addition to a conventional data dictionary exist (Anand et al. 2007, pp. 44–45) but are too general for the given use case with time series-based analytics. The time series generator sheet serves as tool for the transfer of hypotheses and documents all relevant information for subsequent generation. Figure 51 provides a conceptual overview of this BDA book extension.

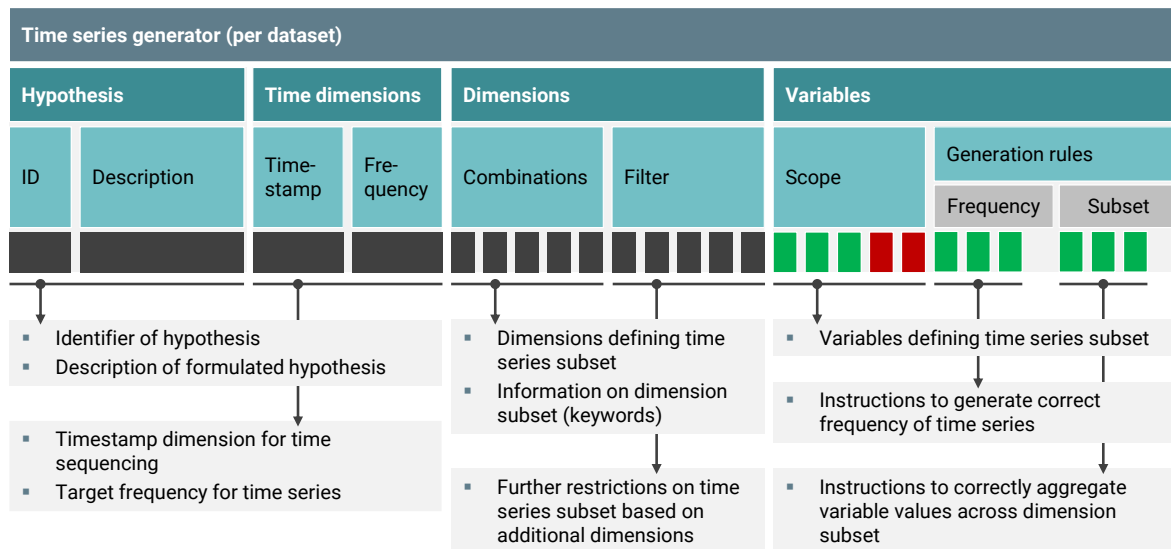


Figure 51 - Time series generator sheet (BDA book)

Keyword	Operation
all	power set ²⁵ excluding empty set
full	aggregation of all values
singles	all individual values
ex(value_1, ..., value_n)	singles excluding value_1, ..., value_n
group(value_1, ..., value_n)	subset including value_1, ..., value_n

Table 19 - Operations for dimension subsets

Each hypothesis is tracked by a unique *identifier* and the *description* provides an explanation in business terms. *Timestamp* explicitly defines which time dimensions is used for building the time sequence of the time series, which is relevant in case the dataset contains multiple dimensions applicable for sequencing. Furthermore, the *frequency* of the time series to be generated is defined. This is important in case it differs from the inherent frequency of the timestamp dimension as this requires to generate the defined frequency. *Combinations* describe the dimensions required to generate the subset of all possible time series in the dataset. Each dimension is composed of a set of values (Pawlak 1981, pp. 205–206) and therefore the relevant subset of these values is defined as well. A standardized set of operations is used for this definition. Table 19 shows an overview of these operations and associated keywords used for documentation in the BDA book. All combinations according to the selected operation serve

²⁵ Power set includes the full set, all possible subsets and the empty set (Halmos 1974, pp. 19–20).

as new dimension values of the time series. The list of feasible operations can be restricted as a result of the dimensionality test.

Other dimensions not selected for combinations can be used as *filter* that exclude certain data points from time series generation. For definition of filters the same operations as for combinations apply. The *scope* determines which variables of the dataset are considered as time series. For the ones selected, *generation rules* are specified. These rules are dependent on the characteristics of the variables, for example, different approaches are required for sales volumes in contrast to price information in various currencies. The BDA manager must ensure that the resulting time series are still meaningful from a business perspective. Frequency rules address this issue along the time dimension and subset rules for required aggregations due to selected dimension subsets. Generation rules optionally include data from other datasets and therefore represent integration of data. For example, currency information from a financial dataset can be used to construct time series measured in a single currency within a dataset containing variables originally measured in different currency units. Determination of combinations, filter, scope and generation rules represent data construction and forms part of data selection. There can be multiple hypotheses for each dataset and a hypothesis generates a number of time series. That is to say, hypothesis-based generation splits a dataset into different sets of time series. Figure 52 provides a simple example based on order data for further explanation of this hypothesis-based time series generation.

4.6.2.3 Automated generation

If the knowledge & dimensionality test indicates that the hypothesis-based generation is not feasible, time series are generated following the automated generation approach. The BDA manager prepares a simplified version of the time series generator sheet. It provides the same information on time dimensions and determines which dimensions and variables are used for time series generation. The default operation for domain subset definition is *singles* as all other operations require domain knowledge in order to check for reasonable subsets. For the same reason, automated generation does not use filters on dimensions. Generation rules are generally required and the BDA manager must provide them where applicable. However, the restriction to *singles*-based dimension subset allows more stringent automation. In case there are no hierarchical dependencies among dimensions, such as segment and sub-segment in the example from Figure 52, subset rules are not necessary.

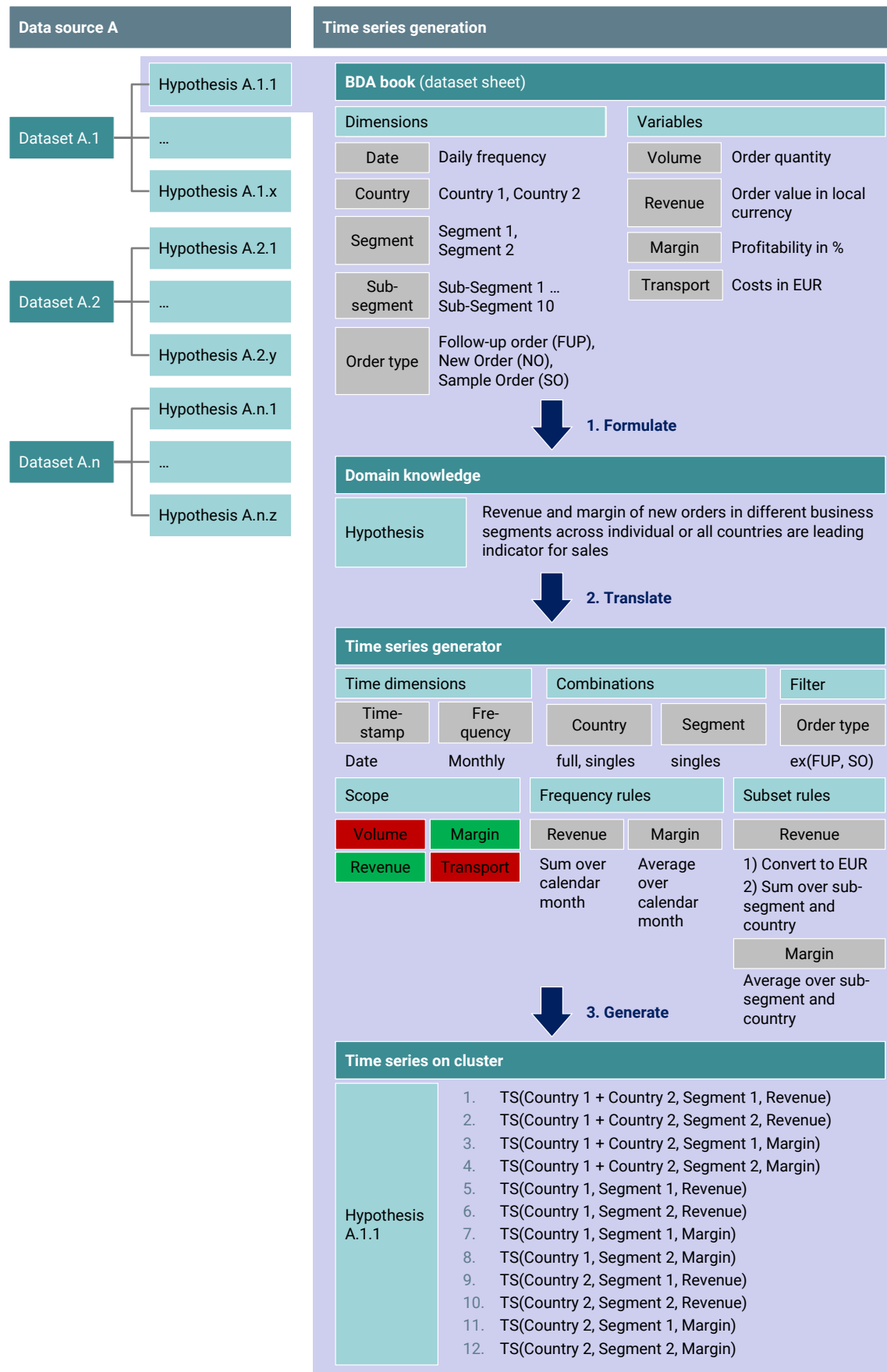


Figure 52 - Time series generation (example)

Especially public datasets often provide dimension values representing predefined subsets, for instance, different regions for a country dimension. These datasets also regularly offer their data at multiple frequencies such that the adequate frequency for time series can be selected without a need for frequency rules. These observations from practice underline validity of the automated approach including restriction to the singles operator. Automated generation results in a single set of time series per datasets which are not further split into subsets. Figure 53 summarizes the approach in form of the simplified time series generator sheet. The data engineer finally generates the time series based on this information.

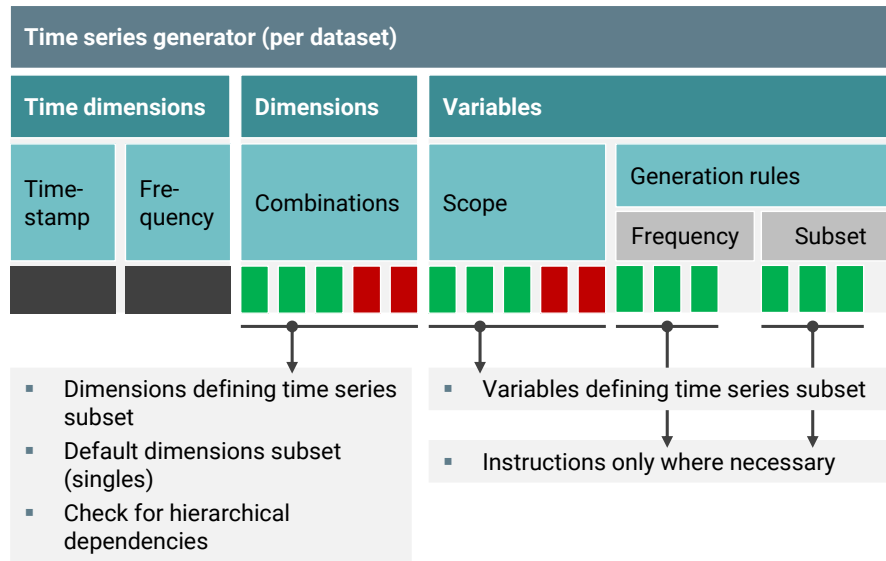


Figure 53 - Time series generator sheet (simplified version)

4.6.3 Time series prioritization

4.6.3.1 Outline

Generated time series represent the part of data in the project repository that is considered as modeling input. Each time series is considered as *feature* for the analytics model and the presented approach based on big data input usually results in a large number of features. Although time series generation addresses the issue of high dimensionality with regard to computational processing limitations, further issues remain for high-dimensional feature input. In particular, high dimensionality is a key challenge for effectively building models (Yang et al. 2015, p. 1; Miao, Niu 2016, p. 919). This challenge is often referred to as curse of dimensionality (Keogh, Mueen 2010, pp. 257–258) and generally requires reduction of dimensionality (Bolon-Canedo et al. 2015, p. 2; Larose, Larose 2015, pp. 92–93). Methods for dimension reduction that aim to keep as much valuable information as possible are typically referred to as "*feature selection or reduction*" (Theodoridis, Koutroumbas 2009, pp. 261–262). Chakrabarti et al. (2009)

provide an overview of these methods.²⁶ *Aggregation methods* combine features, *dimensionality reduction* transforms features into a compressed representation, and *numerosity reduction* aims to represent data by a reduced set of parameters or by a specific sample. *Subset selection* does not change features but defines an adequate selection of original features (Chakrabarti et al. 2009, pp. 84–98). In a narrower sense, the first three methods can be seen as feature reduction and the latter is to be understood as feature selection here. Aggregation methods require definition of aggregation rules which is infeasible given a very large number of features. The remaining feature reduction methods result in artificial representations of original data that is not directly accessible for interpretation anymore. For these reasons, the approach to reduce dimensionality is based on feature selection. Feature selection provides multiple benefits, besides improvement of model performance and comprehensibility (Yu, Liu 2003, 856), it also increases efficiency of model building and future data sourcing due to the reduced feature set (Guyon, Elisseeff 2006, pp. 4–5).

There exist three basic approaches to feature selection (Chandrashekar, Sahin 2014, p. 17). *Filter* methods represent a preprocessing task that is fully independent from the model used for analytics while *wrapper* methods leverage model performance during search for the best subset of features (Das 2001, pp. 74–75). *Embedded* methods integrate feature selection into model building such that they typically are specific to the selected model (Guyon, Elisseeff 2006, p. 5) and therefore not considered for time series prioritization. Kubat (2015) provides more detail on the wrapper and filter approaches. Wrapper methods generate a subset of features and test model performance based on this subset. Then they compare performance to alternative feature subsets and repeat this process until no more improvement is achieved. The idea of filtering is to assess some value of utility regarding the modeling problem for each feature. Based on this assessment, features are ranked and the best features are selected whereby there is no strict definition on the size of the resulting subset of features (Kubat 2015, p. 205). Although wrapper are powerful methods (Kubat 2015, p. 205), they are computationally expensive (Das 2001, p. 75) and have black box character as they rely on machine learning algorithms (Guyon, Elisseeff 2006, p. 5). In contrast, filter methods do not account for optimization of model performance (Guyon, Elisseeff 2006, p. 5), but provide better computational efficiency (Yu, Liu 2003, p. 856) and therefore are better suited for large numbers of features (Das 2001, p. 75; Bolon-Canedo et al. 2015, p. 16). Furthermore, filter methods enable utilization of "[...] general characteristics of the data [...]" (Shin et al. 2009, p. 60) and they are successfully implemented in practice (Chandrashekar, Sahin 2014, p. 17). Two general methods, *individual evaluation* and *subset evaluation* (Yu, Liu 2004, p. 1209; Nisbet et al. 2009, p. 78), are applicable for the filter approach (Yu, Liu 2003, p. 857). Evaluation of features assesses their relevance and redundancy which both represent their value of utility. Relevance describes whether a feature has valuable

²⁶ Similar overviews can be found in Han et al. (2012, pp. 99–111) and Cleve, Lämmel (2014, pp. 206–208).

information for the targeted analytics and redundant features provide the same or similar information (Bolon-Canedo et al. 2015, pp. 14–15). According to Yu, Liu (2004), subset evaluation is capable of handling both relevance and redundancy. However, the search for a feature subset is complex and thus typically not suitable for problems with features in the range of ten thousand or more features. Individual evaluation, in contrast, ranks features by their relevance but does not remove redundant features. Their advantage is their applicability to large numbers of features (Yu, Liu 2004, pp. 1209–1210). Furthermore, they can be implemented based on simple methods such as correlation coefficients (Nisbet et al. 2009, p. 78). A single best method for feature selection is not existent and thus the idea is to find "[...] a good method for a specific problem setting" (Bolon-Canedo et al. 2015, p. 16). As a consequence, time series prioritization as method for feature selection is fundamentally designed as a filter approach with individual evaluation based on correlation.

Another design specification for time series prioritization is the inclusion of domain knowledge. As previously discussed for time series generation, utilization of domain knowledge is beneficial when selecting data (Guyon, Elisseeff 2003, p. 1170; Peng, Kou 2008, p. 48; Ahlemeyer-Stubbe, Coleman 2014, p. 96). Following a filter approach that strictly avoids a black box character by utilizing intuitive evaluation methods provides the basis for integrating business user, including their domain knowledge, into prioritization. However, high dimensionality of features represents a crucial issue here. Direct inclusion of domain knowledge would result in prohibitive effort. Time series prioritization therefore follows a multi-step approach that allows to reduce relevant time series to a manageable number before leveraging domain knowledge.

Data quality plays an important role in analytics. "If poor data is inserted into data mining machines, then the results are as useless as the [...] data itself" (Chen 2015, p. 131). This circumstance is often referred to as *garbage in, garbage out* and data cleaning represents the approach to deal with it (Rahm, Do 2000, p. 3). Data cleaning methods "[...] attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data" (Han et al. 2012, p. 88). Missing values describe the fact that data points of a feature have no value and noise represents "[...] a random error or variance in a measured variable" (Chakrabarti et al. 2009, pp. 72–73). Inconsistencies violate integrity of data, for example, across multiple attributes or datasets, or stem from errors such as values out of range (Cleve, Lämmel 2014, pp. 199–205). Noise in business data typically appears in the form of inconsistencies (Bose, Mahapatra 2001, p. 215), such that noise, in its original meaning of the term, can be disregarded. Inconsistent data is difficult to identify as it requires detailed inspection and correction of errors cannot be automated without limitation to a very specific domain (Maletic, Marcus 2010, p. 23). However, one can leverage an outlier approach based on statistical measures in order to indirectly diagnose data for existing inconsistencies (Maletic, Marcus 2010, p. 23). In contrast, identification of missing values is rather easy (Berry, Linoff 2004, p. 590) and simple methods exist to automatically fill in the value by using variable characteristics such as mean or median (Han et al. 2012, pp. 88–89). Although data quality is crucial in high-

dimensional data, there is no general approach how to deal with quality issues (Yang et al. 2015, p. 2). As the methodology is designed for big data input leveraging external data, identification and correction of inconsistencies cannot rely on domain knowledge. Such an approach would clearly exceed domain knowledge by the business user as well as BDA manager, and the effort would not be reasonable even after reducing the number of time series to a few hundreds. The methodology is based on the idea that the use of big data, in particular large volume of data, increases the margin of error (Mayer-Schönberger, Cukier 2013, p. 35). Moreover, the goal is to provide data clean enough for effective analytics and is not intended to provide perfectly cleaned data (Franks 2012, p. 211). As a consequence, the third design input to time series prioritization is to expand filtering to quality dimensions regarding outliers as characterization of inconsistencies and missing values.

Time series prioritization					
Data source A	Evaluation tool	Evaluation report	Scoring model	General assessment	Detailed assessment
Dataset A.1	Evaluates each time series of a dataset along two dimensions:	Overall evaluation per dataset based on evaluation tool information	Calculation of score for each dataset based on evaluation report information	Selection of datasets based on scoring information	Selection of time series based on detailed review
...					
Dataset A.n					
Data source B	1) Quality <ul style="list-style-type: none"> Time range Missing values Outlier 	Summarization of individual time series evaluations based on aggregation filter and aggregation rules	Enables comparison of datasets along quality and correlation dimensions	3 assessment steps to ensure relevant data, robust scoring and data mix	Pre-selection by BDA manager to remove spurious correlations
Dataset B.1					
Dataset B.n					
Data source m	2) Relevance <ul style="list-style-type: none"> Pearson correlation Cross-correlation 	Result: Quality and relevance evaluations per dataset	Result: Ranking of datasets	Result: Top datasets per data source	Final selection by business user to select best time series and remove redundant information
Dataset m.1					
Dataset m.n					
Prioritize datasets / hypotheses					Prioritize time series
Focus on quality and relevance					Including redundancy
Quantitative information					Domain knowledge

Figure 54 - Overview of time series prioritization

Consideration of all three design specifications - feature selection, domain knowledge, data cleaning - results in an integrated approach as shown in Figure 54. Time series prioritization starts with an evaluation of each time series regarding quality and relevance dimensions, whereby relevance is measured as correlation in two different ways. Based on this information provided by *the evaluation tool*, an overall evaluation per dataset is compiled. Therefore, the *evaluation report* summarizes quality and relevance information of individual time series following defined aggregation filters and rules. The *scoring model* utilizes evaluation reports to provide an overall score in order to rank datasets within each data source. *General assessment* selects the best datasets, however, does not simply set a cut-off point per data source. Three assessment steps consider sensitivity analysis and a data mix requirement in addition to the overall score. The number of potential feature time series is substantially reduced after general assessment such that time

series prioritization shifts from primarily using quantitative information to utilizing domain knowledge. Moreover, *detailed assessment* looks at individual time series instead of datasets and additionally addresses the issue of redundant information. The assessment is divided into a pre-selection by the BDA manager and final selection by the business user which reflects different levels of detail in the domain knowledge applied.

4.6.3.2 Evaluation tool

Individual evaluation of time series, representing features of the analytics model, is implemented by the evaluation tool. It evaluates each time series of a hypothesis or dataset regarding relevance and quality. The value of relevance is based on correlation measures. Furthermore, missing values and outliers are considered for quality. Figure 55 provides an overview of the output resulting from the evaluation tool.

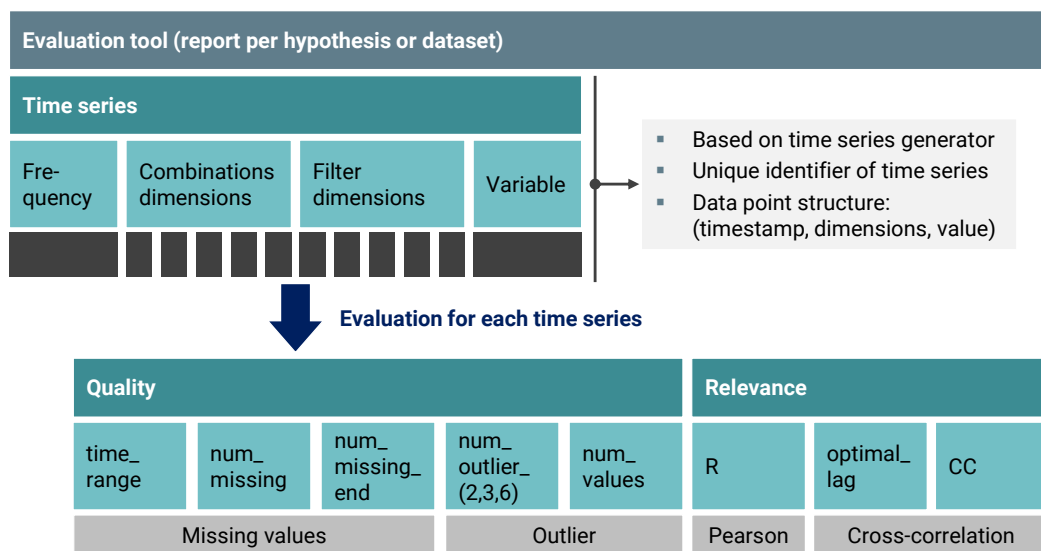


Figure 55 - Evaluation tool report

The evaluation tool considers *missing values* in three ways. Firstly, *time_range* disregards all missing values before the first non-missing value. It therefore provides a measure for the length of the time series. Secondly, *num_missing* counts the number of missing values within *time_range*, that is to say initial missing values are not counted. Thirdly, *num_missing_end* counts missing values at the end of the time series and thus checks whether most recent values are present. Outliers are defined as deviations from the remainder values of the time series²⁷ and its detection is based on exceeding a critical value of a statistical criterion (Grubbs 1969, pp. 1–3). Standard deviation from the mean is the criterion applied in the evaluation tool and the critical value is derived from the *three sigma rule*. Assuming normal distribution, the rule indicates that 68% of values lie

²⁷ In statistical terms, the time series represents a sample.

within one standard deviation, 95% within two, and 99.7%²⁸ within three standard deviations (Kriegel et al. 2009, p. 1650). As one standard deviation would represent a rather restrictive quality filter, it is not used as critical value. The number of outliers outside two standard deviations are counted as *num_outlier_2* and in case of three standard deviations as *num_outlier_3*. In order to compensate for the assumption of normally distributed data, a third outlier measure is introduced. Especially accounting for long-tailed or fat-tailed distributions, *num_outlier_6* counts outliers outside six standard deviations ('six sigma'). This approach to outliers is meant to be intuitive for business users in order to make the process of time series prioritization easy to understand. *Num_values* represents the count of non-missing values in the time series and serves as auxiliary measure for calculating the share of outliers.

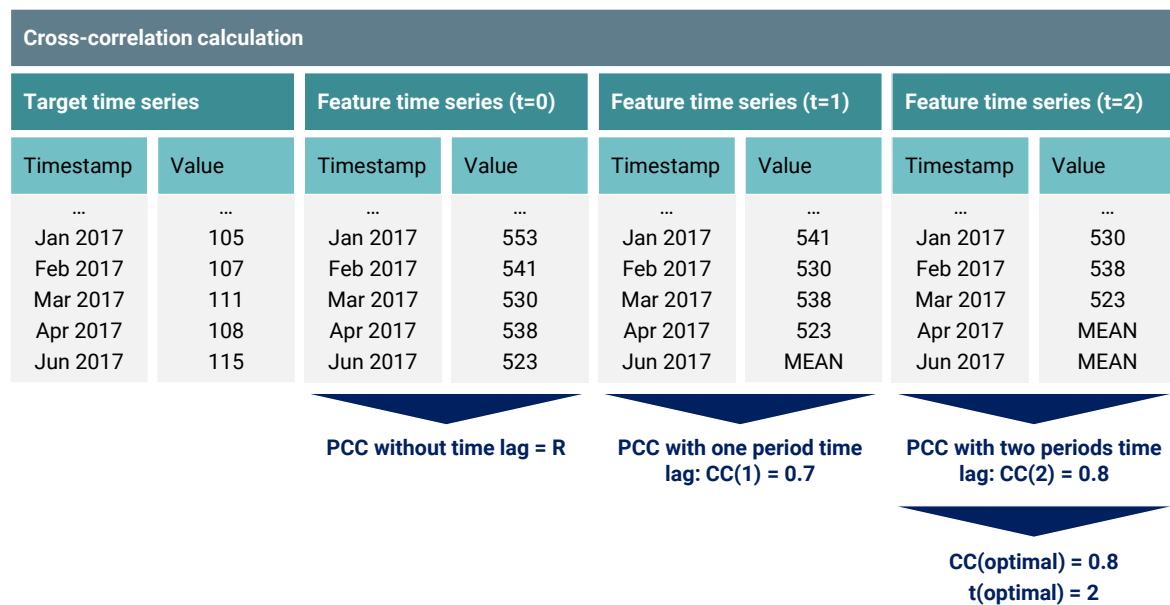


Figure 56 - Concept of cross-correlation for time series

Two different correlation approaches evaluate the relevance dimension, whereby correlation is always measured between the target time series and an individual feature time series. In a forecasting use case, the target time series describes the variable to be forecasted and feature time series represent the modeling input. The *Pearson Correlation Coefficient (PCC)* is a standard measure²⁹ for correlation and represents a measure for the linear relation of two variables (Shevlyakov, Oja 2016, pp. 12–13). PCC is selected for its "[low] computational and statistical complexity" (Guyon, Elisseeff 2006, p. 15). In addition, PCC demonstrated high effectiveness in a benchmark of feature selection methods despite neglecting feature redundancy (Guyon et al. 2006, pp. 237–238) and non-linearities. *R* provides the PCC value for correlation between the feature time series and the target time series. As time series prioritization selects data input

²⁸ The rule is also referred to as *68-95-99.7 rule* (Kriegel et al. 2009, p. 1650).

²⁹ Further details including calculation formula can be found in Shevlyakov, Oja (2016, pp. 12–24).

for a predictive model, the second correlation dimension reflects the idea to identify time series with characteristics of a leading indicator. These indicators provide forward-looking information and are sensitive to future changes such that they are long-established in economic forecasting (Klein, Moore 1983, pp. 119–120) or sales forecasting (Box et al. 2015, p. 468), for example. The goal is therefore to select feature time series that show such leading characteristics with regard to the target time series. Cross-correlation is a suitable approach, because it considers a time lag between variables when calculating PPC (Box et al. 2015, p. 431) and is an applicable method for feature selection (Wells, Rey 2015, p. 124). Figure 56 illustrates the calculation for cross-correlation values as implemented in the evaluation tool.

Correlation between target and feature time series without time lag equals PCC. Introducing a time lag implies to move the feature time series by one period at a time and calculating the new PCC value. The time lag creates a backwards shift of the time series in order to assess it for leading indicator characteristics. Furthermore, shifted values are filled in with the mean of the time series. As PCC calculation is based on the distance between value and mean, this represents a neutral approach regarding calculation. The time lag with the highest correlation value is tracked as *optimal_lag* and the corresponding correlation value as *CC*. Table 20 gives an overview of quality and relevance measures embodied in the evaluation tool.

Measure	Definition
Quality	
time_range	Length of time series starting from first non-missing value, measured in years.
num_missing	Number of missing values within time_range.
num_missing_end	Number of consecutive missing values counted from most recent timestamp.
num_outlier_(2,3,6)	Number of outlier based on 2/3/6 standard deviations from mean.
num_values	Number of non-missing values within time_range.
Relevance	
R	Pearson correlation coefficient (PCC) between feature time series and target time series.
optimal_lag	Time lag with maximum PCC between feature time series and target time series. Measured in periods (months, quarter, years).
CC	PCC value for optimal_lag.

Table 20 - Definitions of quality and relevance measures

4.6.3.3 Evaluation report

The evaluation tool provides the information base for the evaluation report. The report summarizes evaluation information from individual time series across datasets and hypotheses, respectively. This aggregated evaluation provides the basis for subsequent scoring of datasets

and hypotheses within each data source. There are two reasons why features are not strictly selected based on a ranking across all available time series. Firstly, this would not ensure to keep the data mix deliberately created during the big data input step. Making a sub-selection for each data source keeps the data mix structure in the overall selection of time series. Secondly, comparing datasets and hypotheses based on aggregated evaluation information simplifies the search for a cut-off point. The scope of a data source usually lies in the range between tens and hundreds of datasets or hypotheses instead of thousands to millions of time series.

The method to generate the evaluation report bases on a scheme of aggregation filter and aggregation rules that are applied to the information provided by the evaluation tool. Figure 57 presents an overview of aggregation filters including their relation to evaluation tool information.

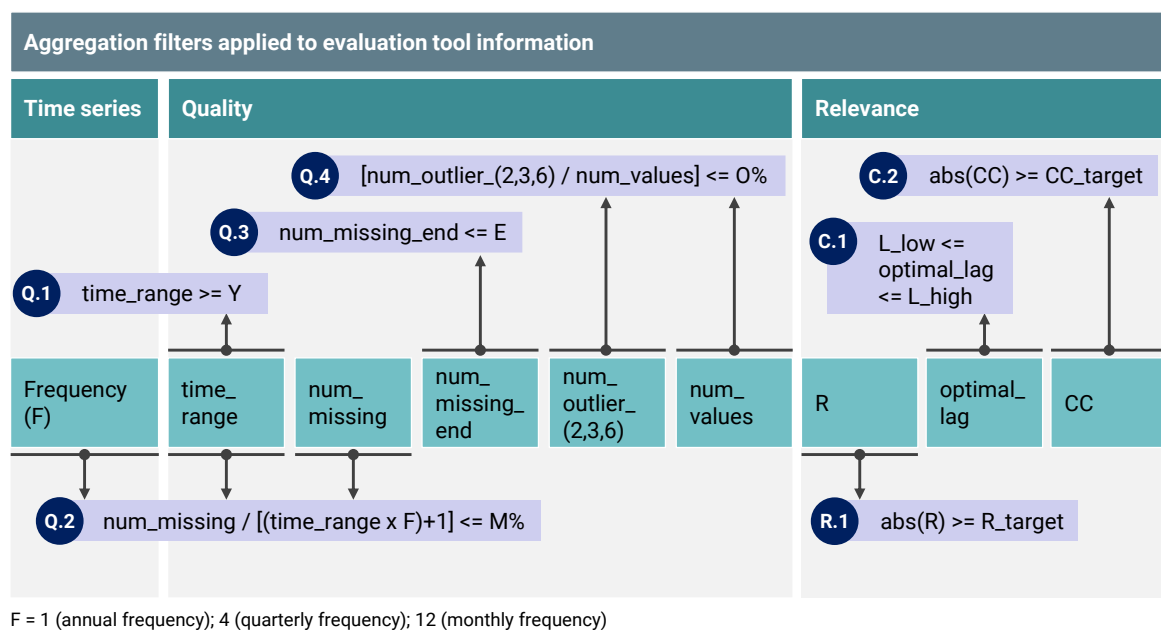


Figure 57 - Aggregation filters overview

In total, the evaluation report considers seven aggregation filters. Four of them address quality dimensions. *Filter Q.1* removes times series without a minimum length of Y years. Although data sources that provide a certain historic range are selected, individual time series still can have insufficient historic data for modeling. The requirement for length must not match the length of the target time series as this would penalize newly introduced data with a naturally short history. *Filter Q.2* excludes time series with an unacceptable high share of missing values. Calculation of the share requires translation of $time_range$ into number of observations which is frequency-dependent. Time series with sufficient historic range and low share of missing values can nonetheless be unsuitable in case most recent data is missing. A certain number of consecutive missing values at the end of a series indicates lacking timeliness of the data, and *filter Q.3* accounts for this issue. Outliers are also restricted to a maximum share within the available $time_range$. *Filter Q.4* is multidimensional as it tests outlier shares for the three different

standard deviation levels with a specific threshold $O\%$ each. Aggregation filters for relevance dimensions set requirements for the strength of correlation. The filters are independent of the direction of correlation because negative relations are considered as equally valuable. Thus, filters use absolute values for correlation coefficients. *Filter R.1* sets a target level R_{target} for the Pearson correlation. For cross-correlation, only those time series that show an optimal correlation within a certain range for their time lag are considered. This is motivated by the fact that forecasting naturally aims for a certain forecast horizon. In order to filter for time series with a meaningful leading indicator characteristic, the optimal_lag is restricted to a range around this horizon by applying L_{low} and L_{high} as thresholds. Time series with optimal cross-correlation in this range are furthermore restricted to a minimum correlation level of CC_{target} .

Filter	Threshold	Frequency-dependent	Basic setup	...	Setup n
Q.1	Y (in years)	no	<i>Use case specific definition of threshold values. Various setups allow for sensitivity analysis.</i>		
Q.2	M%	no			
Q.3	E (in observations)	yes			
Q.4	O_2%	no			
Q.4	O_3%	no			
Q.4	O_6%	no			
R.1	R_target	no			
C.1	L_low	yes			
C.1	L_high	yes			
C.2	CC_target	no			

Table 21 - Aggregation filters setup

It is important to note that only time series that pass all quality filters are considered for relevance filters, however, filters for Pearson correlation and cross-correlation are independent. Table 21 provides an overview of all filters including required thresholds. These thresholds should be defined specifically for the use case. BDA manager and data scientist are responsible to determine the filter setup. Thresholds for filter Q.3 and filter C.1 are frequency-dependent as their calculations are based on observations. A value for each existing frequency must be determined individually. Table 21 also indicates the possibility to introduce a sensitivity analysis by variation of threshold values. Making use of this option provides additional insight for subsequent general assessment.

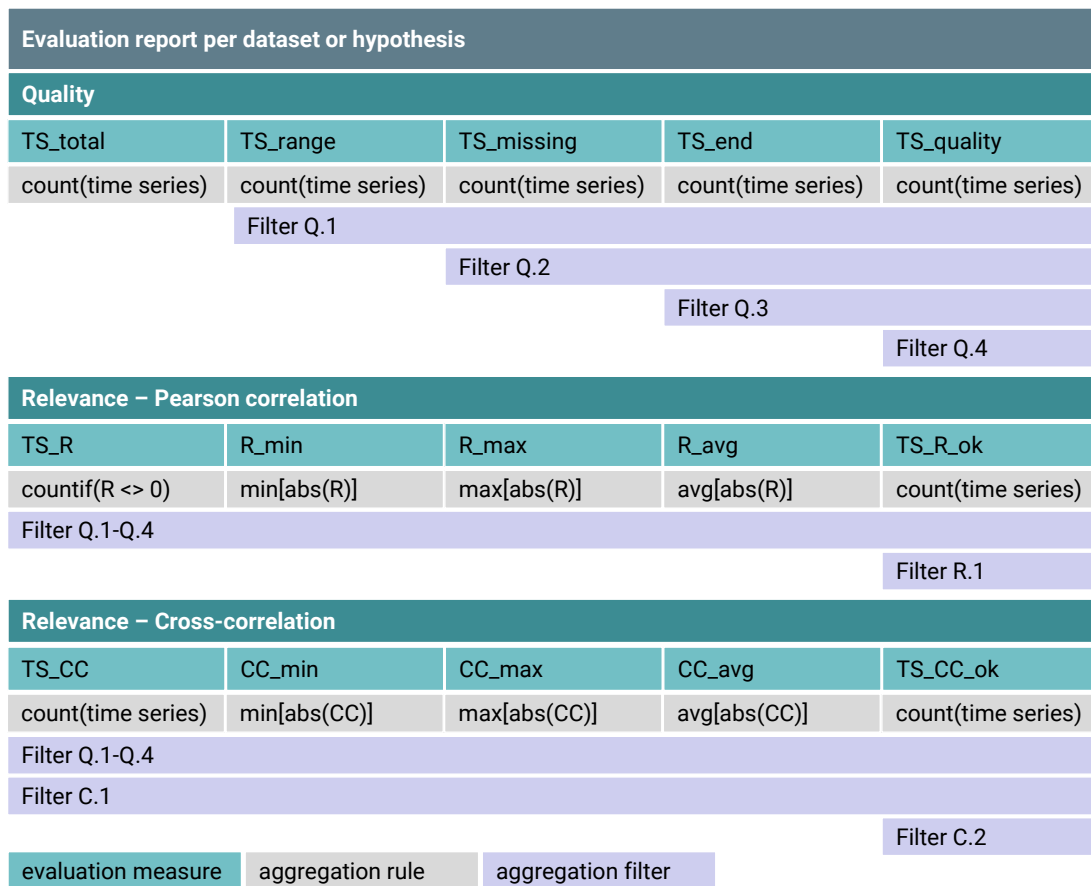


Figure 58 - Evaluation report

The structure of the evaluation report including relations between evaluation measures and aggregation rules as well as aggregation filters is presented in Figure 58. The report is divided into three sections representing quality measures and relevance measures for both correlation approaches. *TS_total* as first measure in the quality section counts all time series generated within the dataset or hypothesis and serves as baseline. The remaining quality measures reduce the count of time series by successive application of the four aggregation filters. Each measure describes the number of time series passing the according filter with *TS_quality* as final aggregate with sufficient quality level. This subset of time series is the basis for both considerations of correlation. *TS_R* shows how many of these time series show a correlation based on Pearson at all. The strength of Pearson correlation is described by the minimum, maximum and average value of PCC. *TS_R_ok* counts the time series with R exceeding the target level. The view on cross-correlation is structured in the same way. *TS_CC* describes the number of high quality time series that show an optimal cross-correlation in the prescribed range for time lag. *CC_min*, *CC_max*, and *CC_avg* provide an overview on strength of cross-correlation while *TS_CC_ok* additionally filters out time series with CC below the established threshold. It is the responsibility of the data scientist to implement and perform evaluations that lead to the presented report. The evaluation report represents an aggregated evaluation of time series for

each dataset or hypothesis. The following scoring model provides a method to compare datasets and hypothesis within a data source on this basis.

4.6.3.4 Scoring model

Identification of best features based on a score for each individual feature is the standard approach of filtering methods for feature selection (Forman 2003, p. 1291). Determination of a single score that represents the quality and both relevance dimensions is the underlying rationale behind the scoring model. It utilizes aggregated information from the evaluation report and thus creates a basis of comparison between datasets and hypothesis, respectively. The presented scoring model therefore extends the basic approach of filtering by combining multiple scores that are applied to predefined subsets of time series. The methodological concept is based on allocating an individual score to multiple measures that cover all relevant dimensions and combining them into an overall score taking into consideration different weights (Eisenführ, Weber 2013, pp. 111–120). Quality, Pearson correlation, and cross-correlation form the three dimensions of the scoring model and Figure 59 provides an overview of all scores applied including their relation to evaluation report information.

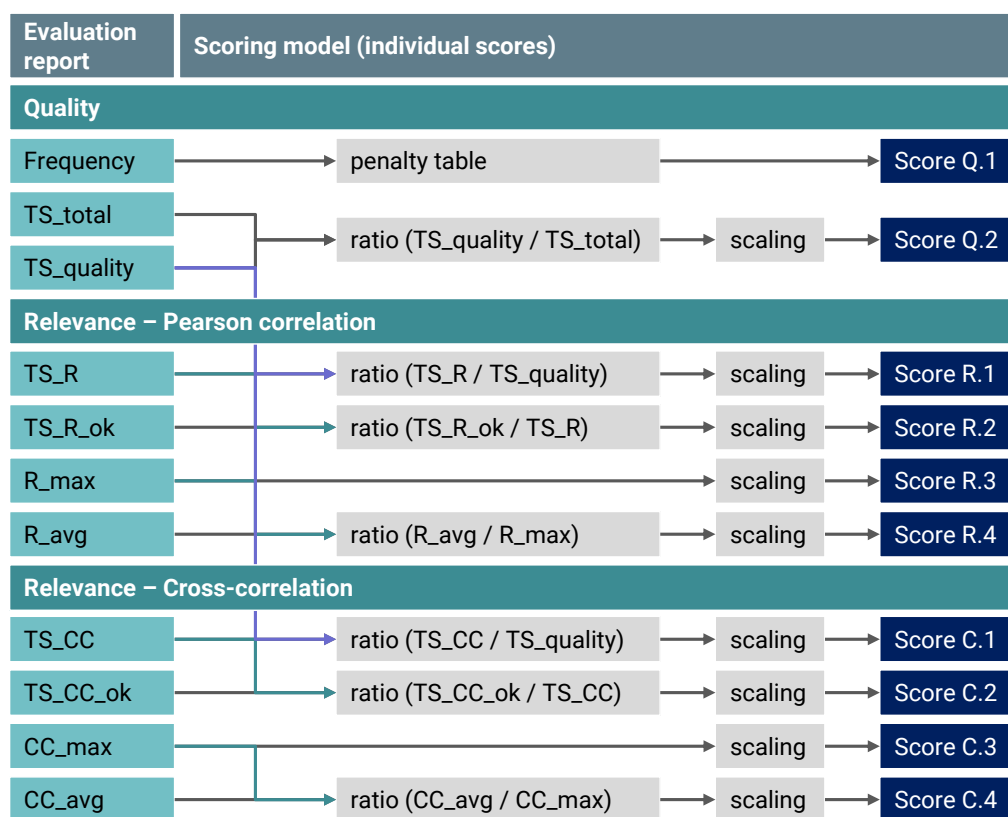


Figure 59 - Individual scores based on evaluation report

The predominant frequency of time series in a dataset or hypothesis is added as quality dimension in the scoring model. Frequencies lower than the frequency of the target time series are penalized for two reasons. On the one hand, lower frequencies restrict the potential update

rate of forecasts. In case of a monthly target time series and a quarterly frequency of a dataset, new information for an updated forecast are only available every quarter despite new observations each month. On the other hand, they potentially require additional preparation in the form of disaggregation to a higher frequency depending on the analytics model employed. A penalty table consequently reflects these circumstances. Specific penalty tables are determined for the individual use case but should adhere to the following guideline.³⁰ Datasets with the same frequency as the target time series get a value for *score Q.1* that allows for the maximum overall score. The score value is successively reduced for lower frequencies and the lowest frequency receives the lowest score or highest penalty, respectively. *Score Q.2* assesses the quality level of the dataset by calculating the ratio between time series passing all quality filters and all generated time series. This ratio is scaled to the range between the minimum and maximum ratio of all datasets evaluated in the data source. Scaling implies a value range of the score between zero and one. It is also applied to all further scores such that a value of one consistently represents the optimal score. Each of the relevance-related dimensions comprises four scores. They follow the same logic and are only distinguished by the information input which is based on Pearson correlation or cross-correlation, respectively. *Score R.1* and *score C.1* represent how many time series convey correlation-based information, in other words the ratio of time series with PCC value and all qualitatively acceptable ones. In case of cross-correlation, only time series with correlation within the specified range for time lags are considered in accordance with filter C.1 as presented before. *Score R.2* and *score C.2* refine this view by assessing how many out of the correlating time series fulfill the requirement of the predefined target level for correlation. Relevance scores 1 and 2 therefore describe which proportion of a dataset is valuable. They are implemented as ratios as the number of time series can greatly vary across datasets. The other two scores add a view on the strength of correlation-based relevance. *Score R.3* and *score C.3* simply take the maximum value of correlation found in a dataset. These scores resemble a traditional score for feature ranking the most. Although being very important, the maximum value is not representative for the entire dataset such that *score R.4* and *score C.4* provide an assessment about relevance consistency. In case the average correlation is much lower than the maximum, the resulting score is low because the maximum correlation is not very representative for the entire dataset. The opposite holds true for an average close to the maximum correlation and a higher score consequently reflects this consistency. Comparison of datasets based on a set of ten individual scores would be burdensome. The scoring model ultimately provides an overall score and Figure 60 presents the logic behind it.

³⁰ Due to frequency aggregation during time series generation, no dataset or hypothesis has higher frequency than the target time series.

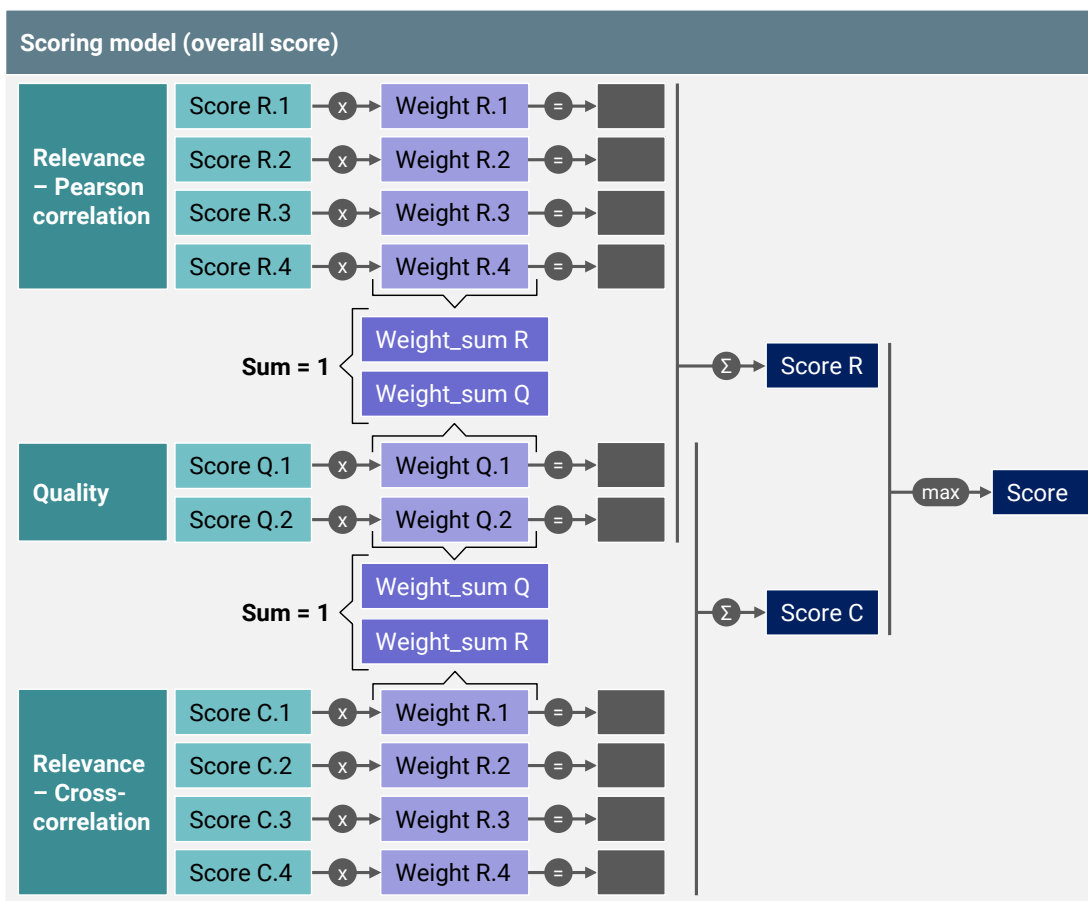


Figure 60 - Overall score in scoring model

The scoring model determines the overall score on two different paths. Each path combines the quality dimension scores with the relevance scores of Pearson correlation and cross-correlation, respectively. Due to equivalence of the individual scoring structure for both relevance dimensions, the logic is identical for the paths. Motivation for splitting the scoring is twofold. On the one hand, this provides an intermediate score for quality and Pearson correlation (*Score R*) as well as for quality and cross-correlation (*Score C*). The maximum score of both represents the overall score for each dataset (*Score*). As a consequence, datasets are compared based on their highest value for quality and relevance, irrespective of the origin of relevance value. Neither Pearson correlation nor cross-correlation is a perfect measure for relevance and there is no clear rationale to prefer one over the other. The source of relevance value should therefore not matter for the score. A simple example illustrates this motivation. Assuming two datasets with identical quality score but different relevance score. One has high scores based on Pearson correlation but low scores based on cross-correlation and the other has average scores for both. Building an overall score including both sets of correlation scores may result in a higher score for the average dataset, although the alternative dataset has a stronger indication for relevance value based on one of the correlation dimensions. The presented structure of scoring avoids such cases. On the other hand, separating scoring and replicating the scoring structure simplifies determination of weights. The set of weights is

reduced from ten to six and this approach ensures that none of the correlation dimensions is given an advantage due to weighting. Setting the weights is key for scoring and requires alignment between BDA manager, data scientist, and potentially business user. However, there exist some general guidelines to be followed:

- 1) The overall sum of weights ($Weight_sum\ R + Weight_sum\ Q$) should equal one as this ensures a maximum total score of one as well. This facilitates comparability between datasets.
- 2) Weighting for relevance ($Weight_sum\ R$) should be significantly higher than for quality ($Weight_sum\ Q$) because relevance scores are exclusively based on time series with sufficient quality.
- 3) Within the relevance dimensions, more weight should be given to the share of time series with correlation above target level ($Weight\ R.2$) and maximum correlation ($Weight\ R.3$). The former ultimately represents the subset of time series that pass all quality and relevance filter defined. Moreover, even if only a small share of time series have the maximum correlation within a dataset, this information should be crucial for comparing datasets. Prioritizing small subsets of highly correlated time series still leads to a sizeable model input when following a big data approach.

The data scientist is responsible for implementing and running the scoring model. This includes multiple runs of the model based on different evaluation reports previously created by alternative filter setups. As a result, scoring information for each data source is available in order to select datasets and hypothesis, respectively.

4.6.3.5 General assessment

General assessment initiates actual selection of data after preparatory work by the evaluation tool, evaluation report, and scoring model. In a strict sense, it is a pre-selection that reduces the scope for final selection during detailed assessment. Results from the scoring model serve as the primary decision-making basis. Figure 61 illustrates the report from the scoring model.

Report from scoring model								
Dataset/ hypothesis	Relevance source	Score				Sensitivity analysis		
		Basic	Setup 1	Setup 2	Setup 3	Setup 1	Setup 2	Setup 3
ID_84	C	0.98	Yellow	Yellow	Orange
ID_54	R	0.94	Orange	Orange	Red
ID_74	R	0.83	Green	Light Green	Green
ID_90	C	0.76	Light Green	Yellow	Light Green
ID_06	C	0.52	Red	Orange	Light Green
ID_23	Q	0.31	Light Green	Yellow	Yellow
ID_117	R	0.29	Yellow	Orange	Red
...

<ul style="list-style-type: none"> ▪ Pearson = R (Score = Score R) ▪ Cross-correlation = C (Score = Score C) ▪ Q = Quality only (Score only based on quality scores) 	<ul style="list-style-type: none"> ▪ Score values per dataset/hypothesis ▪ Ranking from highest to lowest score based on basic filter setup ▪ Sensitivity scores based on alternative filter setups as defined for evaluation report 	<ul style="list-style-type: none"> ■ Strong improvement ■ Moderate improvement ■ Neutral ■ Moderate decline ■ Strong decline
--	--	---

Assessment step 1	Assessment step 2
-------------------	-------------------

Figure 61 - General assessment based on scoring model

The report lists all datasets and hypotheses, respectively, for a data source and indicates whether its score is based on Pearson correlation or cross-correlation. An alternative scenario occurs if the score solely comes from quality scores. The datasets are ranked by their score with basic filter setup and the report also shows scores for the alternative filter setups. Moreover, a sensitivity analysis shows the changes in scoring value in relation to the basic setup. A simple categorization ranging from strong improvement to strong decline in scoring provides an easy-to-read overview. There exists no strict rationale to decide which datasets to keep or not, because filter approaches are not directly related to the model performance. However, general assessment follows a guideline with three steps in order to achieve a selection:

- 1) *Relevance requirement:* The importance of relevance clearly dominates quality. All datasets without relevance-based scoring (status *Q* in the scoring model report) are consequently removed in *assessment step 1*.
- 2) *Robustness requirement:* Different scores from the alternative filter setups should not substantially change the ranking of datasets. *Assessment step 2* therefore searches for a top list of datasets that is stable across all filter setups. This step is particularly helpful for extensive data sources in order to reduce a large number of datasets in regard to the

final step. This implies that this step is optional for datasets with a comprehensible number of datasets.³¹

- 3) *Data mix requirement:* Data sources were deliberately selected to represent a data mix. This idea is transferred to the intra source-level. Datasets remaining after previous assessment steps are consolidated in common groups. Typically, similar hypotheses exist for hypothesis-based generated data or the data source documentation provides superior hierarchies for datasets. Datasets of each group are compared based on their scores including sensitivities and clearly inferior datasets are removed. Illustrating scores as ranking for a group facilitates the comparison across multiples setups. Figure 62 shows a simple example for *assessment step 3*.

Dataset group 1								
Dataset/ hypothesis	Score				Ranking			
	Basic	Setup 1	Setup 2	Setup 3	Basic	Setup 1	Setup 2	Setup 3
ID_84	0.98	0.92	0.93	0.91	Rank 1	Rank 1	Rank 1	Rank 1
ID_27	0.92	0.83	0.81	0.80	Rank 2	Rank 2	Rank 3	Rank 2
ID_135	0.87	0.81	0.83	0.77	Rank 3	Rank 3	Rank 2	Rank 3

Dataset group 2								
Dataset/ hypothesis	Score				Ranking			
	Basic	Setup 1	Setup 2	Setup 3	Basic	Setup 1	Setup 2	Setup 3
ID_46	0.96	0.90	0.89	0.90	Rank 1	Rank 2	Rank 2	Rank 1
ID_50	0.91	0.92	0.91	0.89	Rank 2	Rank 1	Rank 1	Rank 2
ID_79	0.77	0.61	0.73	0.65	Rank 3	Rank 3	Rank 3	Rank 3

■ Selected dataset
 ■ Rank 1
 ■ Rank 2
 ■ Rank 3

Figure 62 - General assessment step 3 (example)

The BDA manager performs general assessment and presents conclusions to the business user for review and final approval. Assessment steps 1 to 3 lead to a substantial reduction of feature time series, therefore enabling a change of focus from datasets to time series in the detailed assessment. Instead of selecting datasets within a data source, time series within a dataset are selected. The BDA manager remains the key person in charge during assessment steps 4 and 5, however, there is a higher degree of involvement of the business user. In the final decision workshop, the business user actively takes selection decisions in contrast with approving decisions made by the BDA manager during general assessment. Figure 63 illustrates this two-

³¹ It is difficult to define a clear limit between extensive and comprehensible size of a data source as this depends on the complexity of the use case and available resources. The range between 30 and 50 datasets or hypotheses can serve as a rough indication.

stage approach of time series prioritization including focus of each stage and all five consecutive assessment steps.

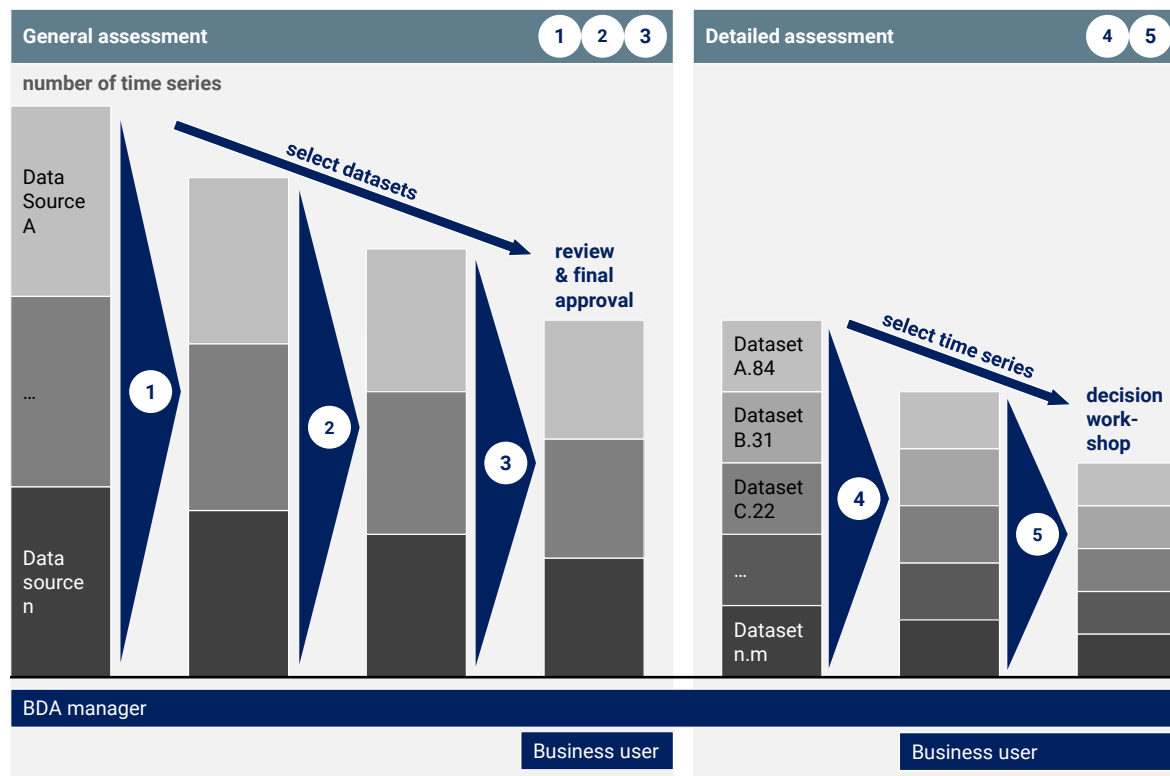


Figure 63 - General and detailed assessments

4.6.3.6 Detailed assessment

The overall goal of detailed assessment is to improve the selection of time series in order to create final modeling input. The assessment addresses the lack of filtering for redundancies during general assessment but is generally build on a broader set of criteria. Detailed assessment can be seen as a combination of additional sub-filters that advances the reduction of feature time series including removal of redundancies (Guyon et al. 2006, p. 238). It is organized in two assessment steps where the former requires general domain knowledge and the latter specific business knowledge. The information base for detailed assessment stems from the evaluation tool which provides an overview of all time series per dataset including quality and relevance evaluations.

Assessment step 4 is performed by the BDA manager and includes the following three selection tasks:

- 1) *Remove spurious correlations*: Spurious correlation is defined as correlation without underlying causation (Simon 1954, p. 467). It poses a specific challenge when using big data input (Gandomi, Haider 2015, p. 143), because high dimensionality leads to statistical correlation between variables without causal link (Fan et al. 2014, p. 298). Occurrence of spurious correlations is not an issue for hypothesis-based generated time

series as they are deliberately generated based on domain knowledge. However, automated generation of time series does not deliberately choose dimensions and variables such that spurious correlations are possible. In order to ensure an efficient search for spurious correlations when reviewing a dataset, the BDA manager should follow two instructions. Firstly, technical dimensions are excluded from the review. These include different calculation forms such as seasonal adjustments or different growth rates of variables, for example. Secondly, the review starts with the dimension having the smallest set of instances and gradually moves to more extensive dimensions. Removing spurious correlations in the former reduces the effort in the latter as the number of instances to be reviewed is potentially reduced.

- 2) *Remove inferior filters*: There exist two ways of using filter in hypothesis-based time series generation. They remove instances considered as irrelevant or they create specific subsets, for example, different material types for factor costs or product types for orders. These subsets are often hierarchical in the sense that one filter configuration represents a subgroup of another configuration. Such filter configurations are useful to test the data for best correlations, however, they can also result in multiple sets of time series with similar quality and relevance characteristics. In that case, the BDA manager needs to decide upon the best filter configuration. Time series resulting from alternative filter configurations are removed. This selection task represents a special form of redundancy reduction.
- 3) *Remove sparse variables*: Due to the way time series are generated, same combinations of dimensions apply to multiple variables. The dimension combinations can represent alternative views on the data but also an aggregate view. Financial data of companies in the business environment is an example for the latter case. Time series can be generated for financial variables grouped by different types of companies such as competitor, supplier, customer, and so on. A variable that shows relevance by correlation only for a small number of these groups is considered as sparse. It does not reflect the assumption that consistent behavior across multiple company groups indicates a credible correlation. These variables and thus the according time series are removed. While the first two selection tasks are dependent from the type of time series generation, removal of sparse variables applies to all datasets and hypotheses.

The BDA manager builds up detailed knowledge about the data and therefore prepares the decision workshop as basis for *assessment step 5*. The focus lies on four different types of selection tasks in the workshop:

- 1) *Substantiate spurious correlations, inferior filters, and sparse variables*: The BDA manager does not necessarily has sufficient domain knowledge in order to fully perform all selections tasks of assessment step 4. Therefore, all open selection decisions for spurious correlations, inferior filters, and sparse variables are prepared for the workshop.

- 2) *Remove redundant information:* Time series generation as well as general assessment do not address the issue of redundancy for data within a dataset. For that reason, the BDA manager identifies variables and dimensions that potentially represent redundant information. Because all time series passed the relevance and quality filters during general assessment, final decision-making concentrates on selecting the best attributes from a business perspective. Details on relevance and quality scores are not required here.
- 3) *Clarify anonymized dimensions:* In case a dataset includes anonymized dimensions, the BDA manager needs guidance from the business user which dimensions or subsets thereof are relevant for analytics.
- 4) *Add weakly correlated time series:* A structured assessment of time series requires to set target values for Pearson correlation and cross-correlation. Even though these target values are determined with a certain rationale involving the business user, they remain arbitrary to some extent. Although general assessment partly addresses this issue by its sensitivity analysis, the final selection task of assessment step 5 takes this circumstance into consideration. Based on the reduced scope of each dataset after applying selection tasks in assessment step 4, the BDA manager carefully reviews the datasets for weakly correlated time series. In this context, weak correlation refers to PCC values and time lags moderately off the target values as defined in the basic filter setup. Finding candidate time series for addition that show some kind of consistency serves as guideline for this review, because it is not meant to reduce previous selection efforts to absurdity. Candidates must show some relation to time series fulfilling correlation requirements or they must represent a group of similar time series that represent a new perspective of the dataset. The following two examples explain both cases. (1) A metric shows high correlation for a certain dimension, however, two out of ten instances of this dimensions miss the correlation targets by a small margin. Time series based on this metric and the two instances are valid candidates. (2) A dataset contains highly correlated time series based on different variables and dimensions. Another group of time series based on the same dimensions but a different variable does not fully meet the threshold. In case this variable does not provide redundant information with regard to the existing ones, they are valid candidates as well. For instance, the existing variables describe order volumes but the candidates are price variables. The BDA manager collects proposals for such candidates including the underlying rationale for addition.

The required types of selection tasks differ among datasets and they are applied on a strict need basis. The BDA manager prepares decision templates for each dataset. Each template explains the selection decisions required and furthermore provides explanatory information based on the BDA book, for instance, notes on variable definitions. Its design strongly depends on the selection types whereby different types can be combined for efficient decision-making. It is also a key task of the BDA manager to focus on decision-relevant information. Figure 64 provides

an example for the case with removing redundant information and adding weakly correlated time series.

Decision template – Dataset A.84																																													
Decision 1 – Order volume	Decision 2 – Pricing																																												
<ul style="list-style-type: none"> ▪ gross_quantity and net_quantity with strong correlation across the same set of business segments: select best variable ▪ segment_C and segment_I with weak correlation for both variables: decide on additional segments 	<ul style="list-style-type: none"> ▪ price_sale_to_min and price_sale_to_target with weak correlation for same segments as order volume variables: decide on additional variables ▪ price_sale_to_min = ratio of sales price to minimum price ▪ price_sale_to_target = ratio of sales price to target price 																																												
Select best variables	Select additional variables																																												
<table border="1"> <thead> <tr> <th>Segment</th> <th>Variable</th> <th></th> </tr> </thead> <tbody> <tr> <td>segment_A</td> <td>gross_quantity</td> <td rowspan="3">} ✓</td> </tr> <tr> <td>segment_D</td> <td>gross_quantity</td> </tr> <tr> <td>segment_F</td> <td>gross_quantity</td> </tr> <tr> <td>segment_A</td> <td>net_quantity</td> <td rowspan="3">} ✗</td> </tr> <tr> <td>segment_D</td> <td>net_quantity</td> </tr> <tr> <td>segment_F</td> <td>net_quantity</td> </tr> </tbody> </table>	Segment	Variable		segment_A	gross_quantity	} ✓	segment_D	gross_quantity	segment_F	gross_quantity	segment_A	net_quantity	} ✗	segment_D	net_quantity	segment_F	net_quantity	<table border="1"> <thead> <tr> <th>Segment</th> <th>Variable</th> <th></th> </tr> </thead> <tbody> <tr> <td>segment_A</td> <td>price_sale_to_min</td> <td rowspan="6">} ✗</td> </tr> <tr> <td>segment_C</td> <td>price_sale_to_min</td> </tr> <tr> <td>segment_D</td> <td>price_sale_to_min</td> </tr> <tr> <td>segment_F</td> <td>price_sale_to_min</td> </tr> <tr> <td>segment_I</td> <td>price_sale_to_min</td> </tr> <tr> <td>segment_I</td> <td>price_sale_to_min</td> </tr> <tr> <td>segment_A</td> <td>price_sale_to_target</td> <td rowspan="5">} ✓</td> </tr> <tr> <td>segment_C</td> <td>price_sale_to_target</td> </tr> <tr> <td>segment_D</td> <td>price_sale_to_target</td> </tr> <tr> <td>segment_F</td> <td>price_sale_to_target</td> </tr> <tr> <td>segment_I</td> <td>price_sale_to_target</td> </tr> </tbody> </table>	Segment	Variable		segment_A	price_sale_to_min	} ✗	segment_C	price_sale_to_min	segment_D	price_sale_to_min	segment_F	price_sale_to_min	segment_I	price_sale_to_min	segment_I	price_sale_to_min	segment_A	price_sale_to_target	} ✓	segment_C	price_sale_to_target	segment_D	price_sale_to_target	segment_F	price_sale_to_target	segment_I	price_sale_to_target
Segment	Variable																																												
segment_A	gross_quantity	} ✓																																											
segment_D	gross_quantity																																												
segment_F	gross_quantity																																												
segment_A	net_quantity	} ✗																																											
segment_D	net_quantity																																												
segment_F	net_quantity																																												
Segment	Variable																																												
segment_A	price_sale_to_min	} ✗																																											
segment_C	price_sale_to_min																																												
segment_D	price_sale_to_min																																												
segment_F	price_sale_to_min																																												
segment_I	price_sale_to_min																																												
segment_I	price_sale_to_min																																												
segment_A	price_sale_to_target	} ✓																																											
segment_C	price_sale_to_target																																												
segment_D	price_sale_to_target																																												
segment_F	price_sale_to_target																																												
segment_I	price_sale_to_target																																												
Select additional segments																																													
<table border="1"> <thead> <tr> <th>Segment</th> <th>Variables</th> <th></th> </tr> </thead> <tbody> <tr> <td>segment_C</td> <td>gross/net_quantity</td> <td>} ✓</td> </tr> <tr> <td>segment_I</td> <td>gross/net_quantity</td> <td>} ✓</td> </tr> </tbody> </table>	Segment	Variables		segment_C	gross/net_quantity	} ✓	segment_I	gross/net_quantity	} ✓																																				
Segment	Variables																																												
segment_C	gross/net_quantity	} ✓																																											
segment_I	gross/net_quantity	} ✓																																											

Figure 64 - Time series decision template (example)

In the decision workshop, BDA manager and business user go through all decision templates and make the final selection decisions on time series. The data preparation step concludes with this and the selected modeling input is transferred to the modeling & evaluation step.

4.7 Modeling & evaluation

4.7.1 Step overview: Select, build & assess

Modeling represents the transition from modeling input to desired insights by application of analytics models. CRISP-DM breaks down the modeling step into four separate tasks (Chapman et al. 2000, pp. 53–56):

- 1) *Model selection* determines appropriate analytics models for the use case.
- 2) *Test design* defines the approach "[...] to test the model's quality and validity" (Chapman et al. 2000, p. 54).
- 3) *Model building* includes setting parameters for selected models and running models based on model input data.
- 4) *Model assessment* describes model performance with adequate evaluation criteria.

Development of an analytics model is of iterative nature (Vercellis 2009, pp. 67–70), particularly between model building and assessment. In the present methodology, model building includes data formatting as this task is model-dependent and various models are typically utilized. Model parameter settings and formatting choices directly influence performance such that finding the best model requires to iterate between model building and assessment. The proposed methodology therefore consolidates both tasks. Moreover, model assessment represents "[...] a purely technical assessment based on the outcome of the modeling [...]" (Chapman et al. 2000, p. 56) such that additional *business evaluation* is required. Based on acceptable technical performance established during model building and assessment, evaluation appraises the resulting model from a business perspective with regard to the objectives of the use case (Wirth, Hipp 2000, p. 34). Chapman et al. (2000, pp. 58–59) propose evaluation as independent step including process review for quality assurance and determination of next steps. The latter two tasks direct towards deployment of the model and thus are not in scope of this work. Evaluation is consequently integrated into this step of the methodology. Modeling and evaluation are highly standardized and lie within the core competency of the data scientist. This section presents particularities for a time series-based forecasting use case in a business environment.

4.7.2 Model selection

Model selection³² generally depends on the underlying analytics problem and model input (Berry, Linoff 2004, p. 605). For each analytics problem, a vast number of potentially applicable models exist, however, selection can be based on "[...] well-defined categories of models [...]" (Hand et al. 2001, pp. 151–152). With a given forecasting use case, the search space is naturally restricted to the category of predictive models (Vercellis 2009, p. 70). Furthermore, two basic types of predictive models exist: *classification* models provide predictions in form of distinct classes and *regression* models predict a specific value (Two Crows Corporation 1999, p. 9). The goal of model selection is to provide a manageable number of model alternatives across the applicable categories that are subsequently built and assessed in order to determine which one works best. In order to do so, the data scientist narrows down the choice by identifying appropriate models based on five selection criteria according to Linoff, Berry (2011). These criteria include differentiation between supervised and unsupervised learning, form and quality of target or input data, ease of use for modeling, and importance of model explicability (Linoff, Berry 2011, pp. 95–98). The methodology aims to build predictive models following a backtesting approach with historical sales data such that there is no need for unsupervised learning. Furthermore, both target and input data are provided in structured numerical form such that categorical data does not limit model choice, for example. Also the issue of missing values and outliers is reduced due to quality evaluation during time series prioritization. Nevertheless, some models are more sensitive to data quality issues, for example, neural networks (Linoff, Berry 2011, p. 97).

The complete methodology grounds in the idea of deliberate and careful selection of data input. Berman (2013) reflects this by stating: "Pick better metrics, not better algorithms" (Berman 2013, p. 162). It is tempting to use most advanced models but they often provide less benefits than expected and it can be more difficult to get satisfactory results as they typically require more model parameters to control (Domingos 2012, p. 85). Moreover, it becomes more challenging to build generally valid models with increasing complexity such that it is advisable to start with more simple models (Lanquillon, Mallow 2015a, p. 80). Advanced approaches often represent black box models that are typically difficult to understand and to interpret for the business user. Furthermore, there exists no general evidence that complex models improve forecasting accuracy compared to simple models (Green, Armstrong 2015). As a consequence, the data scientist together with the BDA manager prepares two different sets of models. The first set represents more simple models and are the starting point for model building. More

³² Model selection often also refers to deciding which model to choose from given alternatives based on their individual performance (Murphy 2012, pp. 22–24). This understanding of model selection applies to model assessment here.

advanced models of the second set are only built in case first set models do not provide sufficient performance.

Model selection template		
	Classification	Regression
	Predictions in form of distinct classes	Predictions of specific values
Starting models	Model C.1	Model R.1
	Functional principle (comprehensible for analytics novices)	
	Rationale for selection (advantages compared to other models)	Model R.2
	Explanation of alternative designs (if applicable)	
	Model C.2	Model R.3
Advanced models	Model C.3	Model R.4
	Model C.4	Model R.5

Figure 65 - Model selection template

Analytics models are readily available by multiple software libraries or analytics tools and therefore are often applied in a black box manner where the inner workings are not known by the user (Rocha et al. 2012, p. 2). While expert knowledge of the data scientist guides model selection (Alonso et al. 2012, p. 7526), the business user still needs to agree with the proposed choice. Larose, Larose (2015) propose a white box approach to facilitate an understanding of analytics models applied. At its core, the approach answers why proposed models were selected and how they work. They additionally suggest to perform exercises with the model in order to strengthen understanding by the user (Larose, Larose 2015, p. xxiii). Following this idea, the BDA manager prepares a model selection template as illustrated by Figure 65. The template lists all models including alternative designs proposed by the data scientist grouped by category and level of advancement. It also provides brief explanations of the decision-making rationale and model workings. Business user, data scientist, and BDA manager review the model selection template in order to approve the proposed selection.

4.7.3 Test design

The test design for a forecasting use case comprises two major features. Firstly, model testing follows a *backtesting* approach "[...] that aims at comparing ex-ante made predictions with ex-post observed numbers" (Baesens 2014, p. 134). Secondly, the test design promotes *generalization* of the model. Generalization describes how well a model performs on observations not utilized for model building (Alpaydin 2010, p. 39). To put it differently, the expected forecast error for application on future data should be minimized (Murphy 2012, p. 23). Poor performance on unseen data, equal to a high generalization error, is often referred to as *overfitting* (Apte et al. 2003, p. 21). As future data is not available to test generalization of a model, historic data on-hand can be divided into disjoint sets called *training set* and *validation set*. The former is used to learn the model and the latter for estimating model performance (Runkler 2010, p. 77).

Method	Description	Suitability
Hold-out	(Random) division into training and validation sets	For abundant observations
Random subsampling	Hold-out repeated multiple times	Not recommended
k-fold cross-validation	Division into k "folds" and repeat learning k-times	Recommended for limited observations
Leave one out	Each fold with exactly one observation; special case of cross-validation where k equals the number of observations	Not recommended
Bootstrapping	Random selection of observations with replacement to create the training set; non-selected observations build validation set	For limited observations

Table 22 - Methods for model performance validation [based on (Shahapurkar 2016, pp. 74–78)]

Division into two sets is called *hold-out* method and represents the simplest form of creating training and validation sets (Cleve, Lämmel 2014, pp. 231–233). However, hold-out reduces the number of observations available for model training (Murphy 2012, pp. 23–24) while the largest possible number of observations is required for learning with high-dimensional data (Donoho 2000). Available data in terms of observations is typically scarce (Hastie et al. 2017, p. 241), therefore other methods are required (Witten et al. 2011, p. 147). Shahapurkar (2016) provides an overview of popular methods, as shown in Table 22.

Shahapurkar (2016, p. 77) recommends *k-fold cross-validation* and bootstrapping for cases of scarce data, with cross-validation being "[...] the method of choice in most practical limited-data situations" (Witten et al. 2011, p. 147). K-fold cross-validation divides the data into k parts of which one is used for validation and the remaining parts for training. Furthermore, this

process is repeated k times such that each fold serves as validation set once, and model performance is aggregated across all repetitions (Hastie et al. 2017, 241–242). Figure 66 illustrates the method for k=5. Dividing data bears the risk that specific instances are over- or underrepresented in the folds created (Witten et al. 2011, p. 152) which represents an issue for learning a generalizable model. *Stratification* ensures that each fold has the same distribution of observations (Shahapurkar 2016, p. 75). As a consequence, *stratified k-fold cross-validation* is proposed as method to estimate the generalization error in the test design.

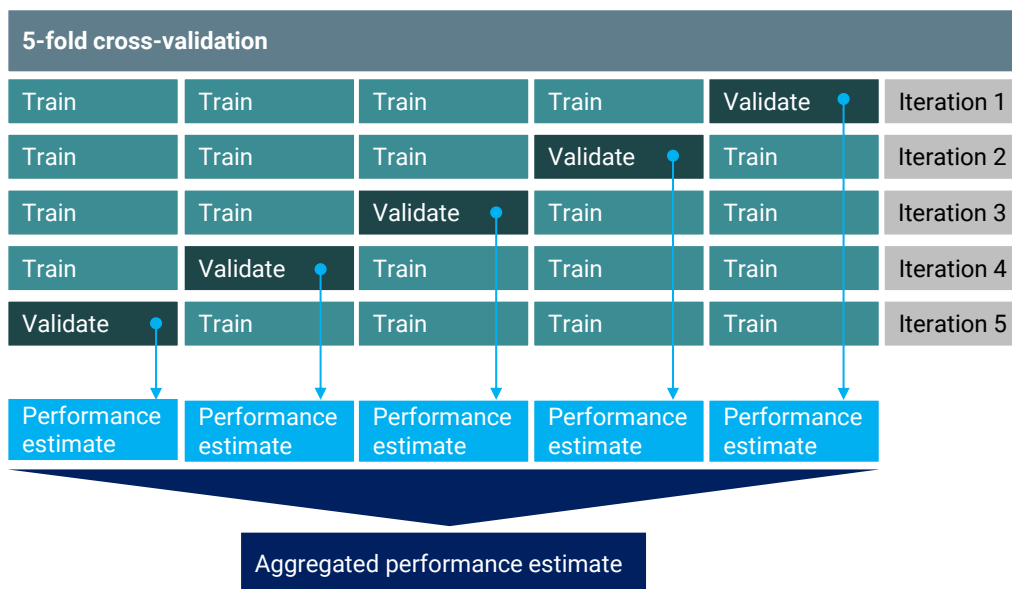


Figure 66 - Concept of k-fold cross-validation method

4.7.4 Model building, assessment and business evaluation

Data conditioning is the final task to transform modeling input into data effectively used by the models. It represents a special form of data construction. While time series generation applies individual construction rules to specific subsets of data, data conditioning refers to the entire modeling input with the same rules. For time series data generated and selected during data understanding and preparation, conditioning includes the following options:

- 1) *Normalization*: Roiger (2017) defines normalization as transformation of numeric values to a specific range. It is critical for performance of certain models, especially for distance-based ones, and four basic types exist. *Decimal scaling* is the simplest form and transforms each value by a power of ten. *Min-max normalization*³³ brings all values to the range between zero and one. *Z-scores* utilize the mean and standard deviation for normalization and are typically applied when maximum and minimum are unknown.

³³ The scoring model of time series prioritization applies min-max normalization to individual scores.

Logarithmic normalization determines the exponent of a logarithmic base that represents the value (Roiger 2017, pp. 208–209).

- 2) *Handling missing values*:³⁴ Time series prioritization generally ensures an acceptable level of missing values. However, some models strictly require a value for each data point. Three basic approaches to handle missing values exist according to Baesens (2014). *Impute* refers to substituting the missing value with a known value such as the mean. In case that missing values are not meaningful, a feature containing missing values can be *removed*. If the occurrence of a missing value contains valuable information, it is a valid option to *keep* it as additional model input (Baesens 2014, p. 19). Furthermore, missing values can also be kept in case the model is capable to deal with them during model building (Hastie et al. 2017, p. 333).
- 3) *Time windows*: The simplest form to represent a time series as feature for modeling is to take its value for any given point in time. This disregards information about dynamics of the time series that are potential features as well. These dynamics can be calculated as descriptive statistics for specified time windows. They can be summary statistics such as mean and standard deviation or time series-specific statistics such as the trend (Chatfield 2016, pp. 11–12).

This overview on data conditioning reveals the existence of alternative options that can be applied to the models. As each option generally has a different effect on model performance, the data scientist can apply different ones during model building in order to determine the best choice. Moreover, some conditioning techniques come at a substantial cost. Testing performance with a simple approach can be done first in order to decide whether a costly one is required at all to improve performance. Two more degrees of freedom to build models are available for the data scientist. Firstly, analytics models typically have freely selectable parameters (Mohri et al. 2012, p. 4). In a narrower sense, model building can be described as "[...] optimization of the model parameters [...]" (Alpaydin 2010, p. 210) in order to improve model performance. Finding optimal parameters is strongly model-dependent and a large variety of techniques exist that are beyond the scope of this work. Jimenez et al. (2009, pp. 2824–2825) provide a brief overview of several techniques and Chappelle et al. (2002) give a prominent example for support vector machines. Secondly, the models can be combined with wrapper or embedded methods for additional feature selection if appropriate. Application of other filter methods is another valid alternative. Combining different feature selection approaches is proposed especially for high dimensionality of data input (Liu, Motoda 1998, p. 87). Similarly to model parameters, deciding on further feature selection depends on the model and other prerequisites. Guyon, Elisseeff (2006), Liu, Motoda (2008) and Chandrashekar, Sahin (2014) provide a comprehensive overview and survey on various techniques, respectively. Time series

³⁴ Little, Rubin (2002) provide an extensive overview on analytics with missing values.

prioritization already represents a comprehensive feature selection method. It is designed to be easy to understand by business users in order to facilitate incorporation of domain knowledge and aims for selecting data input approved from a business perspective. Additional feature selection must not necessarily comply with these restrictions and their primary goal is to improve model performance.

Based on the choices for conditioning, model parameter optimization, and feature selection, the data scientist subsequently builds the models and tracks their performance following the test design. Model building refers to the implementation in an analytics tool (Chapman et al. 2000, p. 55) and fitting the model to the input data. Because the model learns from historic data, this activity is often also referred to as *model training* (Eckerson 2007, p. 7). Tracking of generalization performance requires model assessment which is defined by Hastie et al. (2017, p. 219) as follows:

"Assessment of this performance is extremely important in practice, since it guides the choice of learning method or model, and gives us a measure of the quality of the ultimately chosen model."

In simple terms, assessment answers the question whether a model works or not (Linoff, Berry 2011, p. 180) for the given use case. Model assessment requires measurement of performance which is dependent on the model category. In the case of predictive analytics, performance is generally measured by accuracy of predictions (Eckerson 2007, p. 11). Models in the classification category require different measures compared to regression models because their model output is not a metric value. Performance measures for classification are best explained in the case of two prediction classes using a confusion matrix (Witten et al. 2011, p. 164), which is shown in Figure 67. The confusion matrix summarizes all possible prediction outcomes as basis for calculation of performance measures.

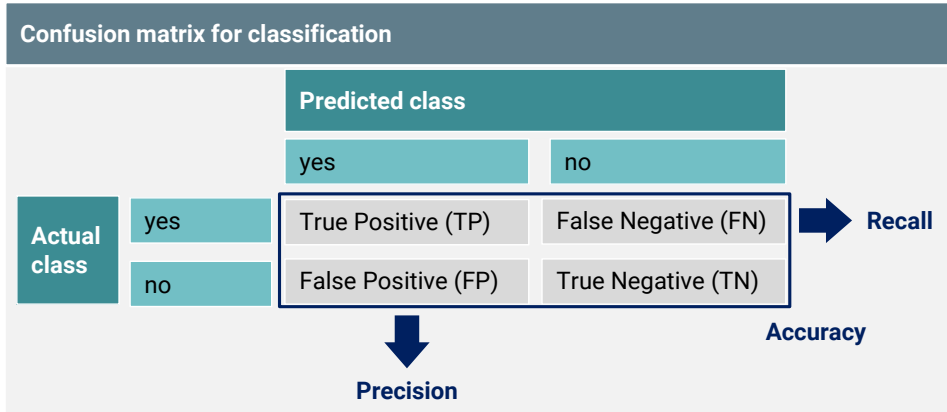


Figure 67 - Confusion matrix (classification) [based on (Witten et al. 2011, p. 164)]

There exist four major measures for classification according to Witten et al. (2011). *Accuracy* is the ratio between all correct classifications and the total number of classifications. *Recall* describes how many observations of a single class are covered by the corresponding class prediction and *precision* represents individual accuracy for each class. Recall and precision are

often combined into a single measure called *F1* (Witten et al. 2011, pp. 163–177). Table 23 provides a summary of classification measures including calculation for a two-class prediction.

Measure	Calculation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Recall	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
F1	$\frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}$

Table 23 - Classification performance measures
[based on (Witten et al. 2011, pp. 163–177; Silva, Ribeiro 2003, p. 1664)]

The same measures are also applicable for multi-class prediction (Sokolova, Lapalme 2009, p. 430). In that case, it becomes clear why accuracy should not be the only measure considered. Accuracy represents an average across all classes and can be misleading when the performance for individual classes is significantly lower (Ganesan 2014). This is particularly important in cases where "[...] classes are very imbalanced" (scikit-learn 2016). While accuracy for multi-class classification is calculated in the exact same way, two alternative ways for the other measures are presented by Özgür et al. (2005). *Micro-averaging* computes measures globally which means that values for TP, FP and FN are added up across all classes before applying the formula for recall or precision. In *macro-averaging* the measure is determined for each class individually and the average across all classes is calculated afterwards. As a result, classes with many observations dominate in micro-averaging and sparse classes in macro-averaging (Özgür et al. 2005, p. 611). Comparison of micro and macro values for the measures enables an assessment of model bias towards highly or lowly populated classes. If the macro value is significantly lower than the micro value, classification for lowly populated classes is worse and vice versa.

Building performance measures on the concept of different accuracy rates is not appropriate for models of the regression category, because errors "[...] are not simply present or absent; they come in different sizes" (Witten et al. 2011, p. 180). Regression models provide the actual (\mathbf{a}_t) and predicted (\mathbf{p}_t) value such that all measures are based on the prediction error represented by the difference of both values (Larose, Larose 2014, pp. 278–280). Numerous measures following this basic principle exist and Witten et al. (2011, pp. 180–182) provide an overview. Table 24 exemplarily shows *Root-Mean-Square Error (RMSE)* as most commonly used measure and *Mean Absolute Percentage Error (MAPE)* error as an alternative.

Measure	Calculation ³⁵
Root-mean-square error	$RMSE = \sqrt{\frac{\sum_{t=1}^n (p_t - a_t)^2}{n}}$
Mean absolute percentage error	$MAPE = \frac{100}{n} \sum_{t=1}^n \left \frac{a_t - p_t}{a_t} \right $

Table 24 - Regression performance measures [based on (Witten et al. 2011, p. 180)]

There exist other assessment criteria for models that either evaluate model complexity (Hastie et al. 2017, pp. 230–241) or investigate consistency of model performance. However, accuracy- and error-based measures are adequate for practical applications (Witten et al. 2011, p. 156). In particular, this holds true as the presented methodology builds upon existing library models. The use of so-called commodity models is a pragmatic approach and model assessment ends "[...] when something good enough is found" (Franks 2012, p. 157). The BDA manager and data scientist jointly review model assessments in order to decide whether starting models are sufficient or advanced models are required. After completion of assessment, they prepare business evaluation by aggregating information for the business user. The primary goal of model assessment is to identify working models while business evaluation aims to decide on the applicability of built models in practice. In order to facilitate the decision the following three activities are helpful:

- 1) *Baseline comparison*: Performance comparison of built models with baseline models provides additional insight for evaluation. Existing solutions currently utilized by the company in the domain of the use case are primary choice for setting a baseline performance. In case such solutions are not available, simple models not based on big data analytics are valuable as well.
- 2) *Practice test*: The optimal way to assess models suitable for practical application is to evaluate their performance on new data that was not utilized for model training or validation (Hastie et al. 2017, p. 222). As discussed previously, data is typically scarce and therefore the test design includes all available data. However, models with adequate performance based on assessment can be employed for a practice test. Testing forecasting models based on real time data is the optimal approach to determine their practicality (Armstrong 2002a, pp. 446–447). Applicability of the practice test depends on the update rate of the forecasting model and the urgency of implementation.
- 3) *Confidence & explicability*: Besides model performance, two more dimensions need to be addressed during business evaluation. The level of confidence for the forecasts and the

³⁵ t denominates the number of n observations

level of explicability for the model (Linoff, Berry 2011, p. 181) need to be sufficient from a business user perspective.

The methodology concludes with the business evaluation review including project sponsor, business user, BDA manager, and data scientist. The focus lies on model assessment and insights from aforementioned activities but also includes a brief review of the use case, big data sources, data understanding, and data preparation in order to set the stage. The outcome is a decision about application of the model and initiates deployment preparations in case of a positive vote. Preparation and implementation of deployment are out of scope for this methodology and thus it is referred to Chapman et al. (2000, pp. 58–62).

5 Evaluation

5.1 Pre-study

The pre-study was performed by the author in order to validate the business need to determine and develop BDA applications that provide a better understanding of the volatile world. The pre-study included multiple discussions with an automotive supplier, semiconductor manufacturer, and contract manufacturer in the automotive industry. The discussions were held with the following roles at these companies: Chief Executive Officer, Chief Operating Officer, Head of Supply Chain Innovation, Vice President Marketing & Sales, Director Business Continuity Planning, and Project Leader Industry 4.0 Strategy. Each discussion was based on an introduction to the situation of the volatile business environment and to the advanced corporate agility system including big data analytics as described in this work.

The pre-study findings underline the alignment between the business point of view on the volatile world and the understanding behind the agility framework. They furthermore confirm that industrial companies regard BDA applications as opportunity to gain a better understanding of their volatile business environment. The following summarizes key insights from the discussions with the three pre-study companies:

- Increasing volatility is observed across all business segments and markets: *this reinforces the need to address the challenges posed by a volatile business environment.*
- Agility includes responsiveness in form of agility levers but also alertness for changes: *this underlines the key role of monitoring in the corporate agility system.*
- Traditional approaches to understand volatilities disregard essential factors which poses a barrier to become an agile company: *this can be addressed by an increased information base (data view).*
- Information regarding volatilities is often collected or prepared by a central unit for the overall company and becomes available for business units or functions only at low frequencies: *this can be addressed by building BDA models that focus on specific needs of business units or functions as defined by use cases.*
- Demand volatility is a key challenge and has great potential for improvement in all companies: *this emphasizes the importance of sales forecasting.*

In summary, the pre-study findings confirm the relevance of the presented methodology. They furthermore provide evidence that the later defined use case of sales forecasting is relevant across different industries.

5.2 Case study introduction

5.2.1 Industry background

The presented methodology was built and evaluated during a case study with an European manufacturer of *Printed Circuit Boards (PCB)* with global business and operations. PCBs fundamentally serve as platform that carries and connects other electronic components such as semiconductors, and they have become increasingly complex representing a high technology product today (LaDou 2006, p. 211). PCB manufacturing as an intermediate industry supplies its products for use in many applications (Suarez et al. 1996, p. 226). Typical applications of PCBs include consumer electronics, telecommunications equipment, industrial electronics including power electronics, medical electronics as well as automotive and aerospace electronics (PCB Wizards 2006). PCB manufacturing therefore represents a classic B2B industry. *Original Equipment Manufacturer (OEM)* of the application domains represent typical customers of a PCB manufacturer. In 2016, the global market size was USD 58.2 billion and thus PCBs correspond to 13% of all *Electronics Manufacturing Services (EMS)* (IPC 2017). The worldwide industry outlook is solid with market growth to be higher than 3% annually in upcoming years which is mainly driven by the communications, computer and automotive industries (ReportLinker 2017). Due to the short lifespan and high innovation rate of application products, sales forecasting is crucial to manage material flows and cycle time at PCB manufacturers (Chang et al. 2009, p. 344). The high level of competition requires PCB manufacturers to put their decision-making on a solid information base of which forecasting is a key part (Chang et al. 2007b, p. 86). Moreover, forecasting is seen as source of competitiveness as it enables better capacity planning and inventory control (Chang, Wang 2006, p. 715).

5.2.2 Project setup

The case study was performed between April 2015 and December 2016 over a period of 20 months. After an initiation phase, a formal research project was set up in September 2015 with the PCB manufacturer, the author of this work as researcher from the *Institute of Innovation and Industrial Management (IIM)* at the Graz University of Technology and the *Know-Center* as external provider of BDA capabilities. The initiation phase covered the team setup and business understanding step while the remaining steps were developed and evaluated during the research project where the author was committed to full-time. Figure 68 presents the timeline of the pre-study and case study including respective focus regarding the new methodology.

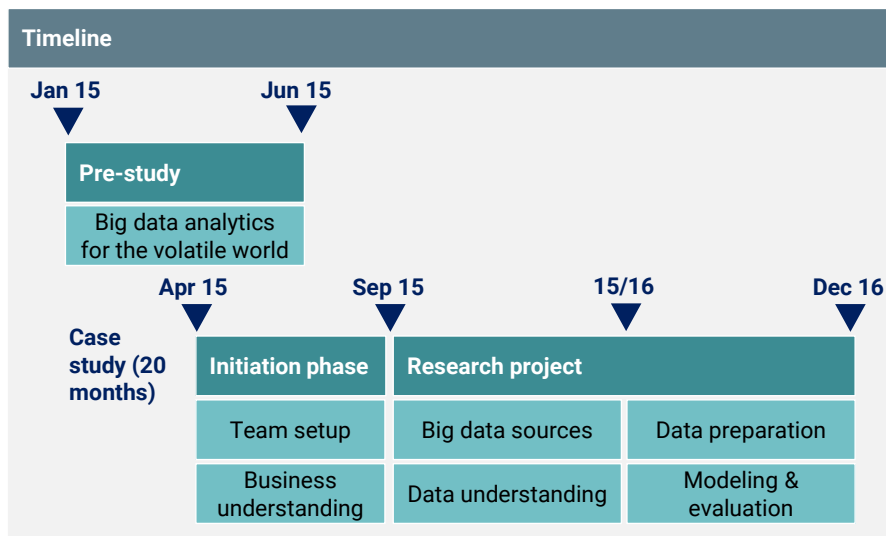


Figure 68 - Timeline of pre-study and case study

5.3 Case study results

5.3.1 Team setup

The roles of the multidisciplinary team setup are reflected by the project team of the case study. The methodology is designed around the central role of the *BDA manager* as this role covers project management and team coordination as basis for comprehensive involvement of domain knowledge. The author occupied the BDA manager role on a full-time basis and provides the required skillset. The academic background in business administration with mechanical engineering³⁶ serves as a solid basis for this role. Knowledge in the business domain stems from more than three years of prior work across different industries and functions including an initiative on advanced analytics in manufacturing industries. The author primarily worked as consultant and therefore collected substantial experience in project management. Moreover, BDA knowledge was previously acquired during research work as doctorate candidate at the IIM institute since September 2014.

All business-related roles were covered by the PCB manufacturer. The *business user* is the second key role in the team setup and was covered by the director and three (senior) analysts from the strategy and business development department. *Project sponsorship* was provided by the *Chief Executive Officer (CEO)* of the company. A legal counsel and risk manager as well as the data protection officer of the company served as *data officers* in the project. The role of *data owner* was split into two groups because internal and external data sources were utilized. For overall coordination and data access to internal sources, the manager of IT applications and the manager of the company data warehouse acted as lead data owners for internal sources. In total, seven data owners covered the range of different data selected for the project. Employees from controlling, purchasing, logistics, and IT were involved here. Customer service teams and IT helpdesks were utilized as data owners of external data sources. The three required BDA roles were covered by the external provider. Two employees from the knowledge discovery department took on the roles of *data scientist* and *data engineer* and three additional *database administrators* covered data extraction and transformation from four data source types and loading to the project cluster. As the setup involved three partners, the proposed assistant project managers were also installed. A lead business user was identified therefore on the company side and the deputy head of the knowledge discovery department took on this auxiliary role for the BDA provider. Figure 69 provides an overview of the team setup.

³⁶ This field of study is known as *Wirtschaftsingenieurwesen, Fachrichtung Maschinenbau* in German.

Team setup				
IIM institute	Company		BDA provider	External data sources
BDA manager	Business user	Project sponsor	Data scientist	Data owner
Researcher (author)	Director strategy & business development	Chief Executive Officer	Data engineer	Customer service
	Assistant project manager	Data officer	Deputy head of knowledge discovery	IT helpdesk
	Senior analyst	Legal counsel and risk manager	Senior knowledge discovery analyst	
	Analyst	Data protection officer	Junior knowledge discovery analyst	
	Data owner		Database administrator	
	Manager IT applications	Controlling	Data source type 1	
	Manager data warehouse	Purchasing	Data source type 2	
		Logistics	Data source type 3 & 4	
		IT		
Full-time	Part-time or as-needed basis			As-needed basis

Figure 69 - Project team setup

5.3.2 Business understanding

5.3.2.1 Agility-based business objectives

The project of the case study started with the initiation workshop including the project sponsor and BDA manager based on the following agenda:

- 1) *Overview on volatilities in industries (5 mins)*: Selected examples of observed volatilities in related industries served as introduction to the issue of a volatile business environment. The examples covered various areas including raw material price as well as sales fluctuations, technology shifts, and regulatory changes.
- 2) *Introduction to the idea of big data analytics (5 mins)*: The data and analytics views were presented against the background of the increasingly volatile and complex business world in order to create a common understanding of strategic value of big data analytics in this environment.
- 3) *Presentation of the advanced corporate agility system (10 mins)*: The agility-based framework was used to build a common understanding of the connection between volatility challenges and company reaction including the potential utilization of big data analytics.

- 4) *Alignment with mission statement (5 mins)*: The BDA manager reviewed the company's mission statement with regard to its business model and highlighted elements that represent agile characteristics as well as high-level benefits from BDA utilization in these areas.
- 5) *Discussion on business objectives (20 mins)*: This part covers the first part of the business understanding step and is based on directions from senior management such that employment of the six-step method to determine effects of the volatile business environment was not required.

The initiation workshop resulted in positive alignment of the company's business model with the agility framework. Perception of the volatile business environment as one of the overall key challenges as well as a need for improvement here reinforced the acceptance of agility as underlying principle. The project sponsor furthermore endorsed big data analytics as potential approach to gain a better understanding of volatilities. Finally, two strategic guidelines were formulated to restrict the scope of potential use cases to the most relevant areas: *technology change* and *sales fluctuations*. Fast detection of changing technological requirements or newly available technologies enables a company to offer state-of-the-art products and to occupy technology niches. A better understanding of sales fluctuations results in improved capacity utilization and product mix in production. These guidelines were verified with business users responsible for both areas in two individual interviews.

5.3.2.2 Use case definition & selection

Within the given guidelines of agility-based business objectives, examination of company issues, business processes, competitor benchmarking, and volatility events in the use case identification workshop generated a long list of eleven ideas. As shown in Figure 70, the number of use cases was reduced to a short list of four use cases based on assessment and prioritization that led to two use cases for final selection after bundling.

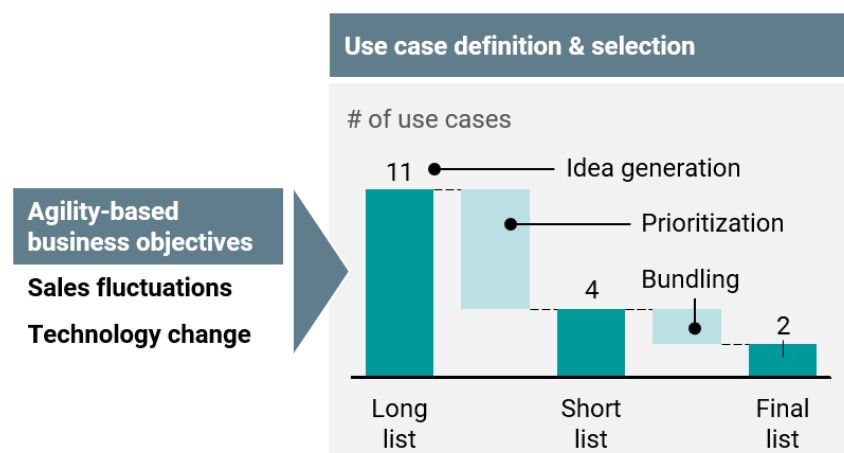


Figure 70 - Overview use case definition and selection

EVALUATION

The relevance and feasibility dimensions of the use case assessment template proved to be efficient and effective. A significant share of required information was already collected during the idea generation workshop and preparation of missing information was less than a working day effort for both business user and data scientist. The formulation of a target state per use case led to a clear qualitative assessment with high business impact for eight use cases, medium impact for one use case, and low impact for two use cases. Out of the high impact use cases, two were removed based on their feasibility assessment as no adequate analytics models were identified as available. The prioritization workshop further revealed a shortage of potential big data input for another high impact use case. To further differentiate the ranking, both feasibility dimensions were translated into a qualitative assessment in the same way as for business impact. This optional assessment revealed a differentiation between four top ranked use cases with high feasibility and one with medium feasibility only. As a consequence, the prioritization of use cases directly resulted from the assessment based ranking. Figure 71 provides an exemplary excerpt of the use case assessment template for the top ranked use cases and Figure 72 summarizes the results of assessment and prioritization in form of the use case portfolio matrix.

Assessment template (excerpt)			
Use case	Target state	Data	Analytics
Sales forecasting	<ul style="list-style-type: none"> ▪ Sales forecast in the medium term ▪ Regular forecast updates ▪ Focus on business segment level ▪ Reduced reaction time to changes in sales 	<ul style="list-style-type: none"> ▪ Order data ▪ Pricing data ▪ Financial reports ▪ Customer & supplier data ▪ Economic & industry data ▪ News data 	<ul style="list-style-type: none"> ▪ Predictive Analytics
Sales monitoring	<ul style="list-style-type: none"> ▪ Early warning system for significant changes in sales ▪ Continuous monitoring ▪ Focus on individual customer ▪ Reduced reaction time to changes in sales 	<ul style="list-style-type: none"> ▪ Order data ▪ Pricing data ▪ Financial reports ▪ Customer & supplier data 	<ul style="list-style-type: none"> ▪ Predictive Analytics ▪ Outlier detection
New technology	<ul style="list-style-type: none"> ▪ Identification of new technologies ▪ Continuous monitoring ▪ Focus on unknown technologies ▪ Representation of dependencies ▪ Early recognition of relevant technologies 	<ul style="list-style-type: none"> ▪ Research data ▪ Patents ▪ Expert forums (e.g., exhibitions) ▪ Competitor data ▪ Startups data 	<ul style="list-style-type: none"> ▪ Text mining ▪ Semantic enrichment & contextualization ▪ High-dimensional time series analytics
Technology trend	<ul style="list-style-type: none"> ▪ Determination of technology maturity levels ▪ Continuous monitoring ▪ Focus on known technologies ▪ Better timing of market readiness 	<ul style="list-style-type: none"> ▪ Research data ▪ Patents ▪ Expert forums (e.g., exhibitions) ▪ Competitor data 	<ul style="list-style-type: none"> ▪ Text mining ▪ Semantic enrichment & contextualization ▪ High-dimensional time series analytics

Figure 71 - Use case assessment template (exemplary excerpt)

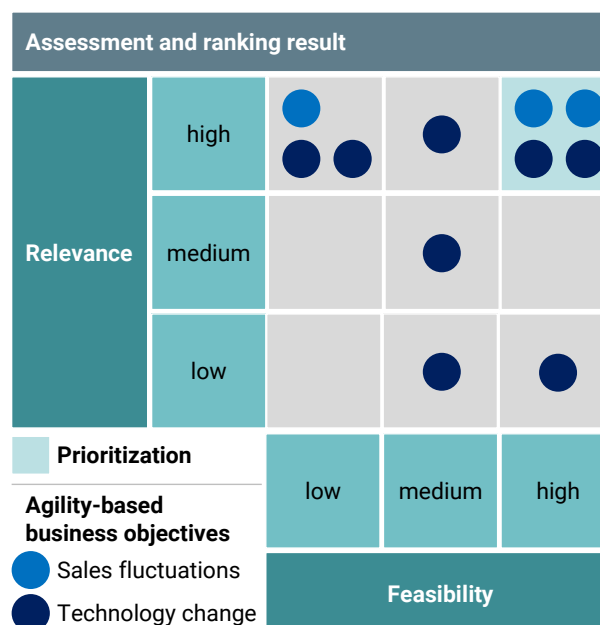


Figure 72 - Use case portfolio matrix

The prioritized set of use cases comprised two cases for each of the business objectives which makes them candidates for use case bundling. In the given company, sales fluctuations and technology change both lie in the area of responsibility of the strategy and business development department with additional involvement of the research and development team for the latter case. Therefore, each pair of use cases allowed for the same project setup regarding key roles. It also becomes clear from the assessment excerpt shown in Figure 71 that both feasibility dimensions show a high level of consistency. *Sales monitoring* uses a subset of data from *sales forecasting* and the same holds true the other way around for analytics. Both technology change related use cases are based on the same approach for analytics and *technology trend* utilizes the same data as *new technology* to a great extent. The resulting use case bundles were discussed with the project sponsor based on the use case decision template. Although both bundles were considered for continuation, the bundle on sales fluctuations³⁷ was selected as initial project. As the case study represented a pilot effort of BDA utilization outside of operations, the project was seen as field test for other projects to follow. This is also reflected by the project approach in two phases. A proof of concept was the deliverable of the first phase and covers all further steps of the methodology including modeling & evaluation. After successful completion, the second phase would then result in a working prototype. Such a prototype would work in the operational environment of the company with real-time instead of historic data. This approach mitigated project failure risk which is important in the studied case due to involvement of an external partner with according budget commitments. Figure 73 summarizes the decision template for sales forecasting as part of the selected use case bundle.

³⁷ A brief summary on use case results can also be found in Winkler et al. (2017) and Kern (2017).

Decision template (excerpt)	
	Use case "sales forecasting"
Objective & scope	<ul style="list-style-type: none"> ▪ Increase mid-term visibility on sales dynamics ▪ Sales growth forecast (measured in units) <ul style="list-style-type: none"> ▪ 12 months forecast period ▪ Monthly forecast updates ▪ Focus on most volatile business segment ▪ Classification or regression model both valid
Project approach	<ul style="list-style-type: none"> ▪ Use of standard predictive analytics models ▪ Cooperation with external BDA provider ▪ External big data analytics infrastructure (hardware & software) ▪ 2-phase approach: Feasibility study + Use case deployment
Deliverables	<ul style="list-style-type: none"> ▪ Phase I: Proof of concept based on built and evaluated models ▪ Phase II: Working prototype <ul style="list-style-type: none"> ▪ Transfer into company infrastructure ▪ Transfer from backtesting to real-time data
Comments	<ul style="list-style-type: none"> ▪ Highest potential impact among all use cases ▪ Project risk mitigated by 2-phase approach

Figure 73 - Use case decision template (excerpt)

The project setup is already discussed in *Subsection 5.3.1* and budget details are omitted for confidentiality reasons. However, the decision template already reflects substantiations compared to the stage of assessment which resulted from decision preparation. In particular, the sales forecasting use case focused on the most volatile business segment and should provide monthly updates on sales growth rates based on sales units over a forecast period of 12 months. Furthermore, the scope of analytics was restricted to established models but does not restrict the category of predictive analytics. Project analytics utilized existing BDA infrastructure of the external partner to avoid investments for investigation of the proof of concept. The remainder of the case study focuses on the first phase of the sales forecasting use case as this covers the entire lifecycle of the methodology. However, the following results on big data sources are directly applicable to the sales monitoring use case as well. The requirement of customer-specific data is the only restriction that requires a subset of data from the identified data sources for the monitoring use case.

5.3.3 Big data sources

5.3.3.1 Identification and filtering

The core activity to identify potential data sources is the data query. In the *systems view*, information on existing internal data sources was collected from ten different employees. The IT department played a key role in providing a comprehensive overview and documentation on major systems. However, other functions were also required to gain a full understanding. Purchasing uses a completely external system for process management, for example. In order to generate ideas beyond existing sources, six semi-structured interviews were performed as part

of the *interface view*. Interview partners at manager level were selected from sales, purchasing, SCM and communications departments following the idea to collect ideas from interfaces with the volatile business environment. The *expert view* involving three of the business users complemented the data query such that in total 19 individual views were collected. Figure 74 provides an overview of the performed data query. The following provides examples for observed adjustment needs during *filtering* of collected ideas:

- *Remove double entries*: Review of the purchasing process management system revealed that it is entirely based on data provided by the ERP system and thus does not provide any additional data. Similar situations were observed for accounting and quotation systems.
- *Clustering*: Different business data from various reports were proposed as potential data sources. Most of the reports were drawn from a single ERP data warehouse accessible via a single interface such that they were aggregated as one data source.
- *Splitting*: The company mail server was proposed as data source as well. Despite originating from the same technical system, email metadata such as time & date information was treated as separate source from email content. They represent structured and unstructured data and, on the other hand, they are treated differently in the system's archive which results in significantly different historic reach for both.

The data query plus ideas previously collected during use case definition resulted in a long list of 28 data sources after filtering which is presented in Table 25. The long list includes 16 internal versus 12 external sources as well as 16 structured versus 12 unstructured data sources.

Data query			
System view		Interface view	Expert view
IT (3)	Sales/engineering (2)	SCM (2)	Business user (3)
Controlling	Purchasing	Sales (2)	
Accounting	Communications	Purchasing	
Sales		Communications	
Σ=19 individual views			

Figure 74 - Data query scope

EVALUATION

Data Sources				
#	Source	Data	Source type	Data type
1	ERP data warehouse	Business data (e.g., orders)	internal	structured
2	ERP system	Business data (e.g., quotations)	internal	structured
3	Google Analytics	Clickstream data on company website	internal	structured
4	SCM file server	Customer forecasts	internal	structured
5	CRM system	Sales opportunities (e.g., volume and probabilities)	internal	structured
6	Treasury files server	Customer credit scores	internal	structured
7	Strategy & business development file server	Industry reports by market researches (quantitative data)	internal	structured
8	Visitor system	Visitor tracking (e.g., timing and frequency)	internal	structured
9	VoIP system*	Landline calls metadata (e.g., date & time or origination)	internal	structured
10	Mobile telecommunications provider	Mobile call metadata (e.g., date & time or origination)	internal	structured
11	Company mail server	Email metadata (e.g., date & time or origination)	internal	structured
12	Company mail server	Email texts	internal	unstructured
13	Sales file server	Customer visit reports	internal	unstructured
14	Strategy & business development file server	Industry reports by market researches (text)	internal	unstructured
15	Media monitoring provider	News (e.g., market trends, brands)	internal	unstructured
16	Strategy & business development file server	Customer satisfaction survey	internal	unstructured
17	Financial database	Financial data (e.g., revenues, earnings, share price)	external	structured
18	Public statistics database (Europe)	Economic and industry data (e.g., leading indicators)	external	structured
19	Public statistics database (global)	Economic and industry data (e.g., leading indicators)	external	structured
20	Google Trends	Search engine data	external	structured
21	Private statistics provider	Industry data	external	structured
22	Company websites	News and communication (e.g., competitor or customer)	external	unstructured
23	News websites	Industry news	external	unstructured
24	Facebook	Social media data	external	unstructured
25	Event websites	Trade fair and conference data (e.g., attendance rate)	external	unstructured
26	Technology blogs	Blog entries	external	unstructured
27	Industry newsletter (subscriptions)	Industry news	external	unstructured
28	Twitter	Social media data	external	unstructured

* VoIP = Voice over Internet Protocol

Table 25 - Long list of potential data sources

5.3.3.2 Assessment and selection

The selection process of data sources starts with the pre-assessment along five decisive factors of which three require specification for the individual use case. Ideally, the *historic reach* of a data source meets the data history of the target time series plus forecast period. As this would penalize newly established data sources, the minimum reach was set by practical values for predictive analytics. Data history is reported to be sufficient in the range between two and three years (Berry, Linoff 2004, p. 63; Armstrong et al. 2015, p. 1724) such that all sources with less than two years historic reach were removed from the long list. The *update frequency* describes the rate at which information of the data source is refreshed. Based on the use case objective to provide monthly forecast updates, data with a monthly update frequency is preferable. Again, this would generally be too restrictive because data with lower frequency still can contain highly valuable information that can be used for analytics. The minimum requirement was consequently set to annual frequency. Only licensing fees were considered regarding *costs* such that data sources incurring fees above a predefined cash budget were excluded. Figure 75 shows an exemplary list of data sources removed from the long list during pre-assessment.

Pre-assessment examples						
Data source		Decisive factors				
Source	Data	Accessibility [yes/no]	Analyzability [yes/no]	Historic reach [> 2 years]	Update frequency [> annual]	Costs [< budget]
CRM system	Sales opportunities	Green	Green	Red	Green	Green
Treasury files server	Customer credit scores	Green	Green	Green	Red	Green
Strategy file server	Customer satisfaction survey	Green	Red	Green	Green	Green
Event websites	Trade fair and conference data	Red	Green	Green	Green	Red
Twitter	Social media data	Green	Green	Red	Green	Red

Figure 75 - Data source pre-assessment with decisive factors

The CRM system was newly introduced at the company before the project start and therefore the minimum requirement for historic reach was not met. Credit scores of customers are updated occasion-related such that scores generally do not change in the short or medium term. As this does not provide consistent information updates with an annual or higher frequency the data source was dismissed. A review of customer satisfaction surveys revealed insufficient structure of the data. Customers define the structure of the feedback and decide on the timing. This does not allow to construct a clear time series of information across multiple customers. Event websites for trade fairs and industry conferences typically require a login in order to access valuable information. Furthermore, information content can be restricted to paying participants and thus coverage of a reasonable number of events ran against the budget restriction. Twitter

restricts access to its freely available data via rate limits (Twitter 2017a) which does not allow to accumulate a long history of data. The fee-based access to the archive with full history of Twitter data (Twitter 2017b) did not meet budget requirements for the present case. In total, 13 data sources did not meet the requirements defined by the decisive factors and consequently the short list for final assessment comprised 15 sources.

The project plan provided the available capacity for each step and each data source required a different level of effort. In order to guide the final selection, each data source was assessed regarding efforts for data sourcing and modeling as key differentiators of sources. Modeling also considers source-specific data preparation tasks such as preprocessing of text data. The assessment was based on capabilities and analytics tools of the provider. As pre-assessment of costs was simplified to screening for unacceptable licensing fees, actual cost information was not available for all data sources. The scoring scope was consequently reduced to historic reach and update frequency with the following setup:

- *Historic reach* (weight = 3)
 - 2-3 years: score = 1
 - 4-7 years: score = 2
 - >7 years: score = 3
- *Update frequency* (weight = 2):
 - Annual to quarterly: score = 1
 - Quarterly to monthly: score = 2
 - Monthly and higher: score = 3

The scoring on historic reach reflects the potential use of the time window approach when data is available for a longer history than the target time series. Characteristics such as a long-term trend of a time series can be utilized then. A higher update frequency reveals changes in the business environment earlier and therefore monthly or higher frequencies were scored highest. Figure 76 summarizes the final assessment with categorical effort estimates from low to high.

Data source short list						
Data source				Final assessment		
Type	Source	Data	Data sourcing	Modeling	Total score	
Internal	Structured	ERP data warehouse	Business data	Low	Low	12
		ERP system	Business data	Medium	Low	12
		Google Analytics	Clickstream data	Low	Low	9
		Strategy file server	Industry reports (numbers)	High	Medium	8
		Visitor system	Visitor tracking	Medium	Medium	9
	Unstructured	Company mail server	Email texts	Medium	High	12
		Sales file server	Customer visit reports	Medium	High	10
		Strategy file server	Industry reports (text)	High	High	8
External	Structured	Financial database	Financial data	Medium	Low	15
		Public database (Europe)	Economic & industry data	Medium	Low	13
		Public database (global)	Economic & industry data	Medium	Low	13
		Google Trends	Search engine data	Medium	Medium	15
	Unstructured	Company websites	Company news	Medium	High	12
		News websites	Industry news	Medium	High	12
		Facebook	Social media data	High	High	12

Effort estimates: ■ Low ■ Medium ■ High

Figure 76 - Final assessment of data sources

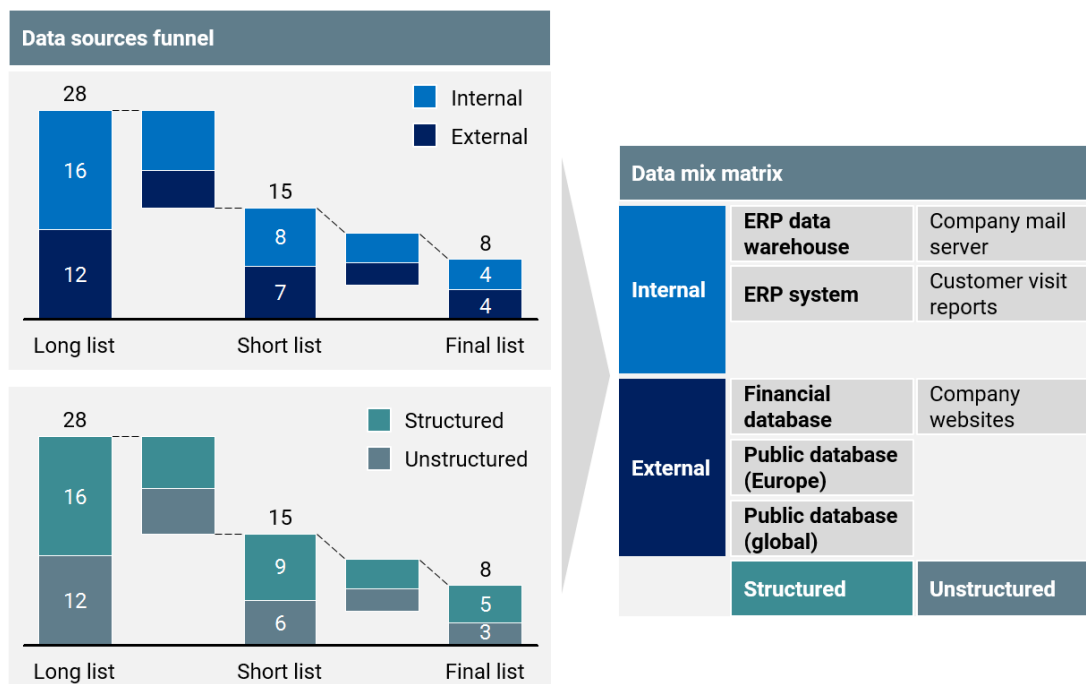


Figure 77 - Data sources funnel and data mix matrix

On the basis of final assessment and following the data mix requirement, the final list of data sources was established. Eight data sources covering the whole data mix landscape were selected in the discussion between business user, BDA manager, and data scientist. Figure 77 shows the resulting data mix matrix and provides an overview of the overall data sources funnel. The funnel clearly shows that the data query approach resulted in a comprehensive list of data sources while pre-assessment significantly reduced the scope to a manageable number for final assessment. Moreover, the data sources finally selected within each type category show the highest scores within this category with only one exception. This is not contradictory as the score only serves as a guideline in addition to effort estimation and value from business perspective. Google trends has a high score due to long data history at high frequency, but shows medium level of effort and competes with highly valuable data sources in the category of external structured data.

The methods of the big data sources step resulted in a balanced mix of internal and external as well as structured and unstructured sources. Consequently, they effectively represent the idea of big data along the 4V dimensions. The remainder of the case study, however, builds on structured data sources only. There are two reasons for this: (1) Mails and customer visit reports represent sensitive personal data that cannot be used for analytics even in anonymized form according to local data protection laws. Although utilization of this data is possible at other company locations, this approach was not pursued. (2) Focus of the use case was to provide a proof of concept for sales forecasting based on big data input. The remaining five data sources still represent an unprecedented volume and variety of data for this purpose. Furthermore, the focus towards monthly updates of forecasts and their underlying data demonstrates high velocity with regard to typical frequencies of industrial sales forecasts in the medium term. Thus, exclusion of company news did not inhibit the proof of concept despite not utilizing the full potential of big data. The focus on structured data furthermore enabled the development and implementation of new methods to understand and prepare the data as described in the following two subsections.

5.3.4 Data understanding

5.3.4.1 Dataset selection

The selected structured data sources included all three types of datasets and the discussion of data understanding is organized by groups of data sources with the same type. The public databases represent *fixed datasets* while the financial database is characterized by *customizable datasets* and the internal data sources contain *sensitive datasets*.

Fixed datasets: Eurostat & OECD.Stat

"Eurostat is the statistical office of the European Union [...]" (Eurostat 2017a) and its database provides access to public statistics organized by different themes (Eurostat 2017b). The database consists of "[...]" over 4 600 datasets containing more than 1.2 billion statistical data values [...]" (Eurostat 2017c). *Eurostat* served as data source for European economic and industry data. The *Organisation for Economic Co-operation and Development (OECD)* comprises 35 member states cooperating on economic and social challenges (OECD 2017a) and offers nearly 600 datasets across various topics (OECD 2017b). The publicly accessible *OECD.Stat* data warehouse represents the data source for global economic and industry data. Both public databases allow downloads of full datasets where each dataset covers a clearly defined topic. Although Eurostat and *OECD.Stat* technically facilitate to detail dataset structures, they are treated as fixed dataset sources for two major reasons. On the one hand, the topical structure of the databases is sufficient for selecting datasets and the given structure of datasets allows direct extraction from the source. On the other hand, the total number of potential datasets is particularly high which makes customization of datasets very costly. Moreover, the methodology, in particular in time series prioritization, assures that relevant data is used for modeling such that a data-driven approach is valid here. Selection of datasets was based on a full review of each data source based on the comprehensive documentation available.^{38,39} In total, 166 datasets were selected from Eurostat and *OECD.Stat* whereof *Appendix A* provides an overview.

Customizable datasets: Financial database

The proprietary database of a leading provider of financial information⁴⁰ represented the financial database source. The database offers multiple ways to access data of which API was selected for the project. API access is restricted in terms of data scope compared to alternative ways but is the preferred option for a potential deployment of an operative forecasting tool. Consequently, data not offered via API was not considered. The API offers more than 60 views on the database where each view represents a specific category of financial information. These views are not datasets with a fixed structure and therefore cannot be extracted as a whole. Each view is represented with a list of associated attributes that are defined along various characteristics as shown in Figure 78. Data extraction therefore requires formulation of customized queries, however, the views can be seen as datasets. Views were aggregated in case they shared related information and the same dimensions structure which allows to build a single

³⁸ Eurostat documentation can be found here:

<http://ec.europa.eu/eurostat/data/database> [last access date: 10/25/2017]

³⁹ OECD.Stat documentation can be found here:

<https://data.oecd.org/searchresults/?r=+f/type/datasets> [last access date: 10/25/2017]

⁴⁰ The vendor of the financial database is not specified for confidentiality reasons.

query type per aggregated view. On this basis, 12 different datasets were selected as scope of financial information as shown in Table 26. Stock market data was divided into two datasets because their structure requires two separate queries. Share prices and volatility measures are examples for variables in stock market data A and stock market data B, respectively.

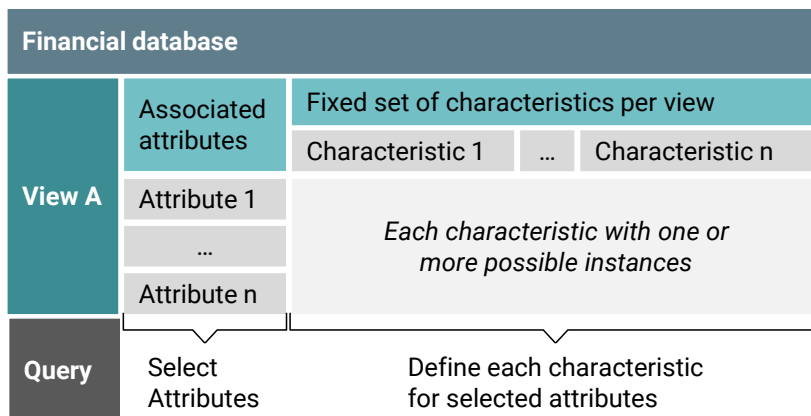


Figure 78 - Customizable structure of financial database

Selected datasets	
Income statement	Forward multiples
Balance sheet	Analyst recommendations
Cash flow statement	Analyst estimates
Financial & growth ratios	Surprises ⁴¹
Debt capital structure	Stock market data A
Trailing multiples	Stock market data B

Table 26 - Selected datasets of financial database

Sensitive datasets: ERP data warehouse ↔ ERP system

The ERP⁴² data warehouse is divided into various areas representing different functions of the company. Multiple data cubes, multidimensional representations of the contained data (Han et al. 2012, p. 187), represent different types of information for each functional area. Each data cube is described in the documentation of the data warehouse and provides a structured data model, however, subsets of the entire cube can also be defined. One type of subsets are reports that are compiled on a regular basis for specific data users. These reports bring two major advantages as they represent a verified and tested structure of available data and, on the other hand, their scope provides a strong indication on most relevant data of a data cube. The data

⁴¹ Surprises represent deviations from actually reported financials to prior estimates from analysts (Investopedia 2017).

⁴² The ERP vendor is not specified for confidentiality reasons.

cubes represent structured datasets that can be customized and 60 of these were reviewed in the case study. Dataset selection resulted in choice of 10 data cubes. Each data cube was associated with its major report as predefinition of the dataset structure. This structure was subsequently finalized during sourcing preparation. In case of data from the ERP system, datasets are directly defined by existing reports that are directly pulled from the system for individual functions or users. A list of potentially interesting reports was collected during the data query and sample reports were collected for dataset selection resulting in three selected datasets. The datasets of the ERP data warehouse and ERP system are still customizable as the underlying reports can be changed, for example, by adding more attributes available in the data cube. Table 27 provides an overview of all selected ERP datasets.

ERP data warehouse	ERP system
Sales	Debtor payment performance
Order backlog	Creditor payment performance
Quotations	Logistics costs
Purchasing prices	
Purchasing spent	
Total inventory	
Finished goods inventory	
Supplier rating	
Supplier risk assessment	
Supplier delivery performance	

Table 27 - Selected ERP datasets

5.3.4.2 Sourcing preparation

Four different BDA books were compiled for the project altogether, with datasets from the ERP data warehouse and ERP system jointly described in one information repository. Selection of high quality sources for external data was beneficial, as comprehensive documentation and qualified customer service enabled the BDA manager to prepare BDA books with the same level of detail as for internal sources with direct access to multiple data owners. Moreover, BDA books proved to be an effective tool for documentation as well as for collaboration. For example, information from nine internal data owners was collected for BDA book compilation. Maintaining open questions in the overview sheet enabled coordination between BDA manager and business user as clarification was partly undertaken by the latter. Moreover, the BDA book enabled remote collaboration between the BDA manager and external database administrators. It also reduced the required involvement of the BDA manager during data sourcing and therefore freed up capacity for parallel arrangement of the data preparation step. Figure 79 shows a sample entry in extracts from the BDA book of ERP data for overview (excluding

EVALUATION

tracker) and dataset sheets. This example shows that internal data contained some sensitive information. Sensitive information required for analytics was anonymized, as shown in the example. Information not required or not permitted for use, such as personal sensitive information, was deleted. Quotations data had the most extensive dataset sheet within the BDA book for ERP data containing 32 dimensions and 39 variables. Eurostat data with 161 datasets represented a special case. In order to keep a clear BDA book, a meta dataset sheet for all possible dimensions and variables within the selected data source scope was created and datasets were clustered into 69 groups of datasets with identical structure. The mapping sheet furthermore explained which dimensions and variables described by the meta dataset sheet are relevant for each group.

Excerpt: BDA book "ERP data"							
Overview sheet							
Dataset	Description	Information	Data owner	Open questions [responsible]			
OBL	<ul style="list-style-type: none"> Order backlog Binding orders, customer forecasts, material reservations Volume, revenue and related costs information 	<ul style="list-style-type: none"> Historic reach: 7 years Frequency: weekly data freezes Additional data freeze at first day of each month Freeze date information not included in standard report 	<ul style="list-style-type: none"> John Doe Controlling john.doe@company.com 012/345678 	<ul style="list-style-type: none"> @IT: addition of freeze date information [business user] @IT: change of currency [business user] @data owner: definition of order type abbreviations [BDA manager] 			
+ 12 other ERP datasets							
Dataset sheet "Order backlog"							
	Time dimensions		Dimensions				+ 25 other dimensions
Name	not included	Delivery date	Country of origin	Segment	Ordering party		
Header	freeze_date	dod_date	country	segment	customer		
Timestamp	yes	-					
Units							
Explanations	monthly data freeze (first day of month)	expected delivery	plant location	business segment (customer)	main customer company (no subsidiaries)		
Sensitivity	-	-	-	-	yes (anonymize)		
	Variables						+ 10 other variables
Name	Op Qty net	Exp. rev	Exp. Rev (a)	Exp. Rev (b)	Transportation		
Header	quantity_net	rev_total	rev_a	rev_b	costs_transport		
Timestamp							
Units	m ²	EUR	EUR	EUR	EUR		
Explanations	order size	expected total revenues	product revenues	one-time revenues	transportation costs (optional field in orders)		
Sensitivity	-	-	-	-	-		

Figure 79 - Sample entry from BDA book

Definition of the sourcing structure is most simple for fixed datasets as it only requires a list of all required datasets, which is documented in the BDA book, and a target data history. For Eurostat and OECD.Stat, the target value for data history was set to 20 years in order to provide a basis to calculate long-term averages, for example. Due to building datasets based on existing reports, the sourcing structure of ERP data also built on data extraction processes for these reports. All required changes were documented by the BDA book: its scope defined additions or removals of attributes compared to the report, split attributes such as variables and units were documented as separate columns, and the creation of data frequency and addition of missing timestamp information is described by the example of Figure 79. Order backlog is recorded at predefined data freeze dates and the frequency was created by selecting the freeze state of the first day of each month because the target time series was of monthly frequency as well. However, this information was not included in the order backlog report and therefore needed to be added as new time dimension of the dataset. Another advantage of building the dataset based on tested reports is the integrated attribute filter function and therefore no definition of additional filters was required.

Sourcing sheet (excerpt)						
Query type	Attributes	Companies	Characteristics			
			Begin date	End date	Frequency	Currency
<ul style="list-style-type: none"> ▪ Each view with different query types ▪ Defines query response and set of characteristics 	share_price	<ul style="list-style-type: none"> ▪ List of relevant companies ▪ Filter for datasets only available for public companies 	<ul style="list-style-type: none"> ▪ Definition of time range extracted by query ▪ 15 years selected for financial database data 		<ul style="list-style-type: none"> ▪ Daily frequency for market data 	trading
	trading_volume					trading
	volatility					n/a
	trading_date					n/a
Sourcing structure:		Scope Timestamp	Attribute filter	Data history	Frequency	
trading_date	company_ID	company_cat	share_price	trading_volume	Volatility	
Time dimension	Dimensions		Variables			

Dataset sheet (excerpt)

Figure 80 - Query and dataset sheets for stock market data A

The customizable datasets of the external financial database require full preparation of the sourcing structure. Although the BDA book defines the target structure of each dataset, adequate queries were required to source appropriate data via API from the database. As a consequence, the BDA book for the financial database was extended with a sourcing sheet for each dataset. Figure 80 shows an excerpt of the sourcing sheet for stock market dataset A and illustrates the connection with the data structure as defined per dataset sheet. Each view of the financial database offers various query types that define characteristics to be selected and the response of the API which is important information to the database administrator.

Furthermore, the sourcing sheet lists all attribute identifiers and therefore defines the scope of the dataset. It is most important to select an attribute that is suitable as timestamp like the trading day date in the example presented. For datasets representing information from financial reporting such as the balance sheet, the date of filing was selected as timestamp as this best represents the timing of when the data is actually available. As an example, first quarter earnings can be filed in June of the same year and therefore taking the end of the first quarter as timestamp would be misleading. All available variable attributes were selected for each view of the financial database. This is motivated by the large number of variables for most datasets, for example, the dataset on analyst estimates contained 237 variables. The idea is to source all available information and filter for the most relevant variables following the approach of time series prioritization. An individual review of a total of nearly 1,300 variables would be very costly for domain knowledge integration. The definition of characteristics represents the data history and frequency dimensions of the sourcing structure as well as provides unit information for variables such as the trading currency, for example.

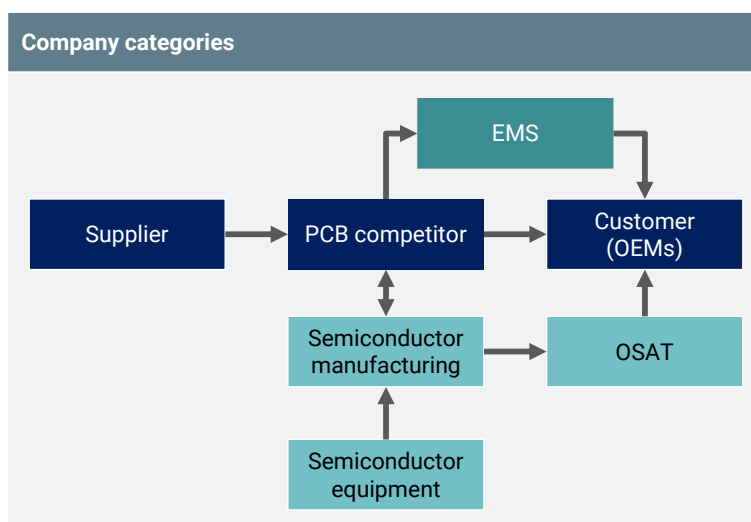


Figure 81 - Concept of company list definition

A specific feature of the financial database is the definition of a list with relevant companies. The database includes a very large number of companies across all industries worldwide and therefore a large share of these companies is irrelevant from a business perspective. As a consequence, the company list served as an attribute filter and was defined by a business-oriented concept as shown in Figure 81. Seven different categories of companies were defined that represent the industrial environment of a PCB manufacturer. At its core, the major supply chain comprises *suppliers*, *competitors*, and *customers*. PCB manufacturing is also indirectly linked to its customers via the *EMS industry* that represents another company category. Furthermore, *semiconductor manufacturing* represents a closely related industry and *semiconductor equipment* providers were also considered to move further upstream in the supply chain. Companies offering *Outsourced Semiconductor Assembly and Test (OSAT)* services represent an important industry segment of electronics manufacturing that is linked to both customers and

semiconductor manufacturers. As a result, the BDA manager and business user jointly defined a list of 412 companies grouped by these categories. The information on the company category was added as dimension to all datasets as illustrated with *company_cat* in the stock market data example. Moreover, the query for each dataset included a filter on the company list based on the company type. Some datasets only include data for publicly listed companies while the remainder are available also for private companies. The sourcing sheets built the basis for the database administrator to create actual queries in the API environment of the financial database.

5.3.4.3 Data sourcing, exploration & verification

After successful legal, analytics and functional clearances by data officer, data scientist and database administrator, respectively, data was sourced following a specific ETL process for each data source based on the sourcing structure as described before. Figure 82 summarizes commonalities and peculiarities across all processes.

Data sourcing process by type			
	Eurostat/OECD.Stat	ERP	Financial database
Sourcing structure	Fixed datasets	Customized reports	Customized queries
Extraction	API access	CSV file download	API access
		Anonymization	
	CSV files	CSV files (anonymized)	MySQL database
Transform	CSV format extraction and mapping to target format structure		MySQL format extraction and mapping to target format structure
	Conversion to target structure based on Apache Parquet data storage format		
Load	Storage in Hadoop cluster		
	Data integrity checks		

Figure 82 - Data sourcing overview

Data from Eurostat and OECD.Stat was extracted via API access that delivered *Comma-Separated Values (CSV)* files like the downloads of ERP data. In the latter case, data was extracted in more than 500 individual files. ERP data was internally anonymized by a business user and subsequently transferred to the external provider on a secure hard drive by the BDA manager. The customized queries extracted data from the financial database via API access and temporarily stored the data in a MySQL database. All extracted data had a source-specific format that needed to be converted into the target format structure of the project cluster. The database administrator defined this target structure based on BDA book information and *Appendix B* shows the resulting format. The data format from extraction was mapped to the target structure first, and then converted into the Apache Parquet format. Apache Parquet is a data storage

EVALUATION

format for the Hadoop ecosystem that is independent of specific data processing infrastructure or programming languages (Apache Software Foundation 2014). Finally, all data was stored in the Hadoop cluster and data integrity checks were performed based on Apache Spark, which is a big data processing engine for Hadoop that provides software libraries, database management and analytics (Apache Software Foundation 2017). In total, the sourced data on the project cluster summed up to more than 800 million data rows with an optimized storage volume of 41 gigabyte. The total processed data volume of the project accounted for more than 320 gigabytes. Following the data sourcing, data was explored and verified leveraging the BDA book again. Correctness checks were defined, where possible, and results as well as conclusions were documented as illustrated by the example in Figure 83. As a result, from data exploration and verification, 42 variables from ERP datasets and 161 variables from finance database datasets were flagged due to an identified issue and consequently excluded for further consideration. This corresponds to a 14% reduction of variables within both datasets. Moreover, 20 datasets were flagged for Eurostat and OECD.Stat data which represents 12% of sourced datasets here. As a result, the method for dataset selection proved as effective but data exploration and verification still added value.

Excerpt: Dataset sheet "OBL" (order backlog)					
	Variables				
Name	quantity_net	rev_total	rev_a	rev_b	costs_transport
Correctness checks	quantity_net > 0	rev_total = rev_a + rev_b	rev_a > rev_b	rev_b > 0	-
Exploration information	-	-	-	-	values only for <1% of rows
Issue flag	-	-	-	-	exclude

Figure 83 - Extended BDA book in data exploration & verification

5.3.5 Data preparation

5.3.5.1 Time series generation

The explored and verified data builds the basis for time series generation and the approach is defined by the dimensionality and knowledge test. ERP datasets showed characteristics of dimensionality too high for automated time series generation as the following analyses show. Assuming powerset combinations of the available dimensions, ERP data would have resulted in $2^{902} = 3.38 \times 10^{271}$ different time series. To put this number into perspective, the estimated number of atoms in the observable universe is of the order of 10^{78} to 10^{82} (Villanueva 2015). Even removing the two most populated dimensions still resulted in $2^{60} = 1.15 \times 10^{18}$ time series, which is still large compared to 4.3×10^{17} seconds passed since the beginning of the universe (Mastin 2009). Dimensionality to this extent is an insurmountable challenge even for modern

big data processing technology. On the other hand, extensive domain knowledge regarding business data was available due to the business user and BDA manager such that hypothesis-based generation was selected for ERP data. Data of the financial database had a considerably lower number of dimensions and the scope of the major dimension representing companies was reduced by the attribute filter, however, the level of dimensionality was still critical. Analyst estimates represent the dataset with highest potential dimensionality where all possible combinations for 412 companies including all subsets, 14 different estimate horizons, and 206 variables results in 3.05×10^{127} potential time series. Although the domain knowledge on financial data within the project team was lower compared to business data, still a sufficient level of expertise was available for hypothesis-based time series generation. This was not the case for Eurostat and OECD.Stat data. Even after comprehensive study of the data sources by the BDA manager during data understanding, knowledge-based definition of most relevant time series subsets across 146 different datasets was found not to be feasible. As a consequence, automated generation was selected for both data sources and processing feasibility was confirmed. For example, the number of dimension combinations following the automated generation approach with singles as default operation did not exceed 7×10^4 for OECD.Stat datasets.

The approach of hypothesis-based time series generation can be described as a general structure across all datasets of the financial database because the datasets share the same structure with only a few exceptions. Correlations of different types of financial information with future company sales was the fundamental hypothesis which was divided into three specific types. *Type I* was defined as correlations between individual companies and future sales while *type II* considered aggregate dynamics of the seven defined company categories. *Type III* represented an advancement of the category-based hypothesis that provided a higher level of consistency. For that type, a subset of companies which commonly appeared across multiple variables was considered. The definition of this subset was defined by the BDA manager following a simple heuristic. For each dataset, the companies with no missing values for each variable were identified and only those companies consistently appearing for the most populated variables were selected. Each hypothesis type was combined with an additional filter that further reduced the scope of companies. The filter optionally removes companies that were not considered as directly relevant to the business of the company by assessment of the business user. For example, PCB manufacturer serving other market segments are excluded as they do not represent direct competition.

Time series generator sheet (general structure)				
	Variables		Filter	
	Type I	company_ID = singles	fwd_period = singles	company_ID = singles; ex(list of indirectly relevant companies)
Type II	company_cat = singles	fwd_period = singles		
Type III	company_cat = singles	fwd_period = singles	company_ID = ex(list of inconsistent companies)	

optional for datasets with forward-looking data

Figure 84 - General structure of combinations and filters (financial database)

Figure 84 summarizes the general structure of combinations and filters applied for time series generation. The overview also includes another combinations dimension that is required for two exceptions. Forward multiples and analyst estimates represent forward-looking data, for example, price-earnings ratio can be calculated as the ratio between share price and earnings per share for the next quarter or the next business year. In order to find the best forward-looking period (*fwd_period*), multiples and analyst estimates were considered for all available periods. In total, 34 hypotheses were generated for the 12 datasets of the financial database because the heuristic of type III hypotheses did not provide a meaningful list of companies for analyst estimates and stock market data B datasets.

Time series generator sheet (analyst recommendations)				
	Variables			
	Group 1	Group 2	Group 3	Group 4
	Frequency rules			
Type I			variable / share_price [stock market data A]	average over month
Type II	average over month	average over month	a) variable / share_price [stock market data A]	a) scaling to [0;1] over time period
Type III			b) average over month	b) average over month
	Subset rules			
Type II			average per category	
Type III	average per category	sum per category		
	Group 1 EPS growth rate [%]	Group 2 'Buy' recommendations [#]	Group 3 Target share price [trading currency]	Group 4 Target share price standard deviation [trading currency]

Figure 85 - Generation rules for analyst recommendations

Another feature of the financial database is the high number of variables per dataset. The average number of variables for a dataset was 95 compared to 10 for ERP datasets. To ensure efficient implementation of time series generation by the data engineer, the BDA manager built groups of variables that require the same frequency and subset rules. Furthermore, many variables are recorded in different currencies, for example, the share price is denominated by the local trading currency⁴³ of the primary stock market for each company. When aggregating companies to a category, companies with trading currency in Japanese Yen would outweigh Euro-based prices by a factor of 130⁴⁴. All currency-based data was therefore scaled to a range between 0 and 1 over the observational time period. Figure 85 shows the second part of the time series generator sheet for analyst recommendations as representative example.

This dataset required frequency aggregation as the original data is recoded at daily frequency. Group 1 included variables such as *earnings per share (EPS)* growth rates estimations by analysts and group 2 included the number of analysts per recommendation category such as 'buy' or 'sell'. Both groups required the same frequency and subset rules for all hypothesis types. Growth rates are measured in percentage and therefore the average was calculated to create monthly frequency and the average represented each category of company in type II and type III. The sum of recommendations was considered as representative value for a company category. Type I did not require a subset rule as no aggregation along the company dimension is required. The estimated target share price of group 3 is an example for data construction with data integration. The hypothesis assumed that the target estimate is more meaningful in relation to the current share price and therefore this ratio was constructed as part of the frequency rule. Moreover, the information of the current price was integrated from another dataset (stock market data A). The hypothesis assumed the standard deviation of share price estimates, as example of group 4, to be an indicator for uncertainty about the future in the market and was therefore not put in relation to the current share price. However, the variable was measured in trading currencies and therefore scaled before frequency and subset aggregation as explained before. For group 3 and 4, category subsets were calculated as the average per company category.

Hypothesis-based generation for ERP data was less structured compared to the financial database case, because datasets do not generally share a common set of dimensions. Moreover, ERP datasets contained more than 14 dimensions per dataset compared to the two standard dimensions used in generation for financial data. The variability of dimensional scope was also high ranging from 1 to 38 dimensions per dataset that were suitable to build combinations and filters. This is reflected by the total number of 61 hypothesis for the 13 ERP datasets and a range between 2 and 8 hypotheses per dataset as shown by Table 28. The hypotheses utilized

⁴³ Data can be extracted in one common currency from the database. This option was not utilized as dynamics in exchange rates then overwrite trends in the underlying variable.

⁴⁴ Based on the current reference exchange rate from 28 Aug 2017 (ECB 2017).

EVALUATION

44% of available dimensions and 69% of available variables on average. This substantial reduction underlines the need for domain knowledge in formulation of hypotheses on ERP data. A simplified but representative example of hypotheses-based generation for ERP data is provided in *Subsection 4.6.2*.

Dataset	# of hypotheses
Sales	6
Orders	3
Quotations	5
Purchasing prices	4
Purchasing spent	7
Total inventory	7
Finished goods inventory	8
Supplier rating	2
Supplier risk assessment	6
Supplier delivery performance	6
Depitor payment performance	3
Creditor payment performance	2
Logistics costs	2
Total	61

Table 28 - Hypotheses per dataset for ERP data

The maximum frequency of sourced Eurostat and OECD.Stat data was monthly and therefore time series generation did not require frequency aggregation. Moreover, the automated generation was performed with the default use of the singles operator for defining dimension subsets. A review by the BDA manager did not reveal any hierarchical dependencies for any dataset such that also no generation rules for subsets were required. Consequently, time series generator sheets were reduced to the selection of relevant dimensions for time series generation. This selection was reduced to removal of dimensions that are technically not required for generation such as time format of the timestamp. There was a notable exception, namely the dimension for age and sex that are existent in datasets describing the labor market. A breakdown of this data into age groups or by sex was not considered as relevant time series. To simplify, the information on excluded dimensions was directly integrated into the dataset sheets of the BDA books.

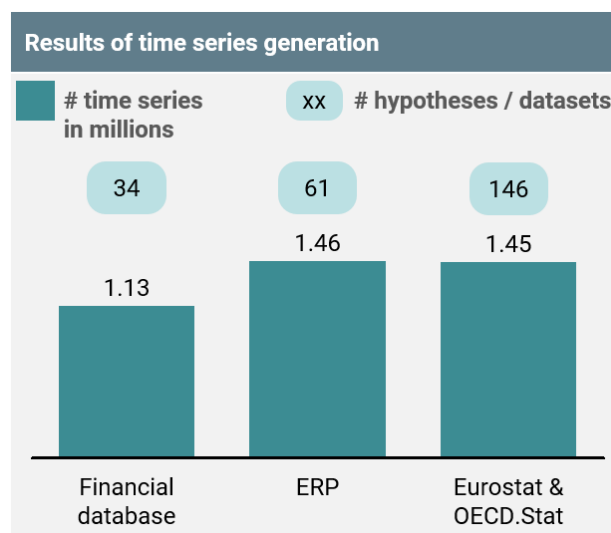


Figure 86 - Summary time series generation

Figure 86 summarizes time series generation that resulted in a total number of more than 4 million individual time series. Data from ERP data warehouse and ERP system were grouped for subsequent time series prioritization. The same holds true for Eurostat and OECD.Stat data. The groups not only represent same type information but they were also comparable in size regarding generated time series. The implementation of time series generation as well as the tools of time series prioritization was realized with Python and Scala programming languages⁴⁵ within the Apache Spark processing framework on the project cluster.

5.3.5.2 Time series prioritization

In the first part of time series prioritization the most relevant datasets and hypotheses with adequate quality characteristics are selected for each group of data sources. Evaluation tool, evaluation report, and scoring model build the basis for this general assessment. Next to the generated time series, the evaluation tool requires the target time series as input. As the use case objective was to provide a sales forecast for the most volatile business segment, the target time series was represented by sales of this segment. The data was readily available as part of the sales dataset from the ERP data warehouse, however, some preprocessing was required. Sales can be measured as volume or in monetary terms (revenues) and different ways exist how to measure sales. In order to be independent from specific characteristics of one sales figure, the following two target time series were defined for use in time series prioritization:

⁴⁵ More details on employed programming languages can be found here [last access date: 10/25/2017]: <https://www.python.org/> (Python) and <https://www.scala-lang.org/> (Scala)

EVALUATION

- (1) Sales volume measured in units
- (2) Product revenues measured in company currency (excluding one-time revenues and rebates)

These target time series were represented each by one variable of the sales dataset and were generated by sum aggregation for the considered business segment and additional filtering. Excluding specific cases such as internal orders, returns or claims ensured that the target time series represented actual sales.

Filter Setup ->		(I) Basic			(II) Short			(III) Long			(IV) High_Corr			(V) Outlier		
Frequency ->		M	Q	A	M	Q	A	M	Q	A	M	Q	A	M	Q	A
Filter	Condition															
Q.1	Y	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Q.2	M%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Q.3	E	2	1	0	2	1	0	2	1	0	2	1	0	2	1	0
Q.4	O_2 %	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	17%	17%	17%
Q.4	O_3 %	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	8%	8%	8%
Q.4	O_6 %	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	4%	4%	4%
R.1	R_target	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.9	0.9	0.9	0.7	0.7	0.7
C.1	L_low	9	3	1	9	3	1	15	5	2	9	3	1	9	3	1
C.1	L_high	21	7	2	15	5	1	21	7	2	21	7	2	21	7	2
C.2	CC_target	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.7	0.7	0.7	0.5	0.5	0.5

alternative setups

Table 29 - Defined filter setups for evaluation reports

In order to aggregate evaluation tool information into an evaluation per dataset or hypothesis, the threshold values need to be defined for the filter of the evaluation report. Table 29 provides an overview of the basic and alternative filter setups as defined in the case study. Based on filter Q.1, time series with a historic range of less than two years were removed. The quality threshold for data sources was also applied for individual time series because two years is considered as minimum requirement for meaningful analytics. Thresholds for filters on missing values and outliers were defined based on an acceptable level from business perspective. Assuming an annual time series, the acceptable rate for missing values and outliers was defined as one in ten years. This translates into a threshold value of 10% for filter Q.2 as well as filter Q.4 and implies six occurrences over a five-year period for monthly time series, for example. The basic filter setup only considered outliers based on six standard deviations as data exploration generally revealed long tail distributions for the data. To test for changes in the quality assessment based on more restrictive filtering for outliers, the alternative setup (V) was introduced. Again, threshold values were defined based on a discussion about the acceptable occurrence rate for different types of outliers. Outliers defined by values outside two standard deviations were accepted once every 6 months for a monthly time series, in case of three standard deviations once every 12 months and for six standard deviations once every 24 months. The threshold for missing values at the end of a time series was set at restrictive levels, for example, at a maximum of two for monthly data. This avoids considering time series that potentially are not maintained anymore and therefore would not be available in future. The remaining filter setup substantiates the relevance dimension of evaluation and builds upon the general idea that PCC values greater than 0.5 indicate moderate correlation and greater than 0.7 strong correlation (Rasli 2006, p. 29).

These values served as target levels for Pearson correlation and cross-correlation in the basic filter setup. The threshold was set at a lower level for cross-correlation as it additionally includes the time dimension covered by the predefined ranges for optimal lag in filter C.1. The range of optimal range was set to 9 months and 21 months, respectively, for monthly time series and to according values for other frequencies. Given a target forecast period of 12 months, this preferred time series with moderate cross-correlation at the long end. Three sensitivities of correlation-related thresholds were defined in order to account for the fact that the correlation dimension is crucial in the assessment of datasets and hypotheses. The filter setup (II) put focus on cross-correlation lags around the forecast period while filter setup (III) based evaluation on the long end. In addition, filter (III) raised the target levels to the strong level for cross-correlation and to a very strong level for Pearson correlation, which is represented by a PCC value of 0.9 or higher.

The evaluation report provides an aggregated evaluation for each dataset or hypotheses based on the defined filter setups. In order to get to the score for each dataset, the weights of the scoring model as well as the penalty table for low frequency data need to be defined. As the target time series frequency is monthly, all datasets with the same frequency were not penalized by assigning 1.0 as value for Score Q.1. From business perspective, updates on quarterly basis are still very valuable and therefore the score value was set to 0.5. Annual frequency was considered as critical update rate regarding trends in sales development such that a significant penalty was selected by a value of -1.0 . The scoring weights were selected in adherence to the guidelines. The total weight of quality dimensions was set to $\text{weight_sum Q} = 0.3$ with equal weight for frequency and quality level scores. Relevance dimensions were weighted significantly higher with $\text{weight_sum R} = 0.7$ which also ensured a maximum score = 1.0. The highest individual weight was assigned to the score of maximum correlation with $\text{weight R.3} = 0.3$ because evaluation of relevance ultimately seeks for best correlation. The weight value was equivalent to the total weight of quality dimensions. The share of time series conform to relevance target levels is second most important and therefore weighted with $\text{weight R.2} = 0.2$. The remaining relevance scores are equally weighted at one step lower with $\text{weight R.1} = \text{weight R.4} = 0.1$ as they provide additional value about the distribution of correlation within a dataset. Table 30 summarizes the chosen weights of the scoring model. As time series prioritization in the case study rested on two target time series, scores were calculated separately and averaged across both views.

EVALUATION

Dimensions	Individual score	Weight	Weight value	
Quality	Score Q.1	Weight Q.1	0.15	Weight_sum Q = 0.3
	Score Q.2	Weight Q.2	0.15	
Relevance	Score R.1; Score C.1	Weight R.1	0.10	Weight_sum R = 0.7
	Score R.2; Score C.2	Weight R.2	0.20	
	Score R.3; Score C.3	Weight R.3	0.30	
	Score R.4; Score C.4	Weight R.4	0.10	
Total weight = 1.0				

Table 30 - Defined weights of the scoring model

The scoring model provides the decision basis for general assessment that comprises three consecutive steps. Assessment step 1 removes all datasets or hypotheses without any significant correlation and the results for all three groups of data sources are presented in Figure 87. The relative reduction lies in a range between 9% for financial data and 25% for ERP data. These reasonable shares of non-correlating data indicate that dataset selection and formulation of hypotheses generally result in relevant data input. This becomes even more clear when looking at datasets instead of hypotheses for financial and ERP data. In the latter case, only finished goods inventory and supplier risk assessment datasets were excluded which represents 15% of all datasets. Stock market data B dataset is the only removed dataset for the financial database representing 8% of its entire scope.

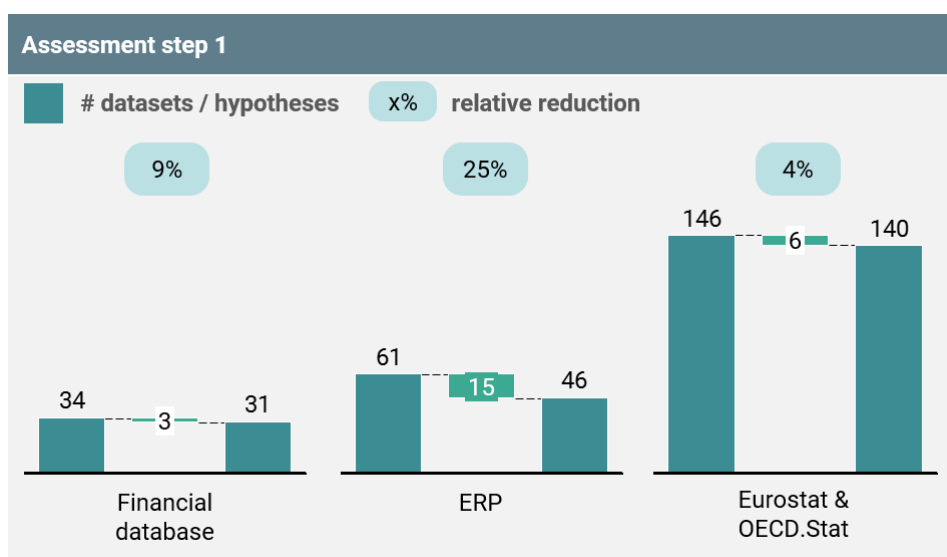


Figure 87 - Assessment step 1 results (all data sources)

As the results of assessment step 1 show, the number of remaining hypotheses for financial database and ERP was already within a manageable range. Consequently, the definition of a score-based top list in assessment step 2 was only required for Eurostat & OECD.Stat. As the top list should be robust across sensitivities, the ranking of datasets is based on the average score across all defined filter setups. This is exemplarily shown for the top 5 datasets in Table 31. This example also reveals a common pattern across all data sources. Sensitivity for filter setup (V), representing stricter thresholds on outliers, was generally low. On the one hand, this

sensitivity does not provide much additional information to prioritize datasets and was therefore disregarded for further assessment. On the other hand, it also states that outliers do not represent a substantial issue for the selected data of all data sources. To reach the same range as the other data sources, the scope for top ranked datasets of Eurostat & OECD.Stat was set to top 40. As the 41st (namq_10_a10_e) and 39th (namq_10_a10) dataset represented two different sets of economic indicators⁴⁶ based on the same industry breakdown, the former was also kept for further assessment such that 41 datasets remained after assessment step 2.

Rank	dataset_id	Relevance source	Score					Sensitivity changes				Ranking Score
			(I) Basic	(II) Short	(III) Long	(IV) High_Corr	(V) Outlier	(II) Short	(III) Long	(IV) High_Corr	(V) Outlier	
1	ei_isppi_q	Corr	0.82	0.77	0.80	0.70	0.82	-6%	-2%	-16%	0%	0.78
2	ei_isset_q	Corr	0.81	0.77	0.79	0.69	0.81	-5%	-2%	-17%	0%	0.78
3	ei_isse_q	Corr	0.81	0.78	0.74	0.70	0.81	-4%	-9%	-16%	0%	0.77
4	ei_isppe_q	Corr	0.79	0.78	0.72	0.69	0.79	-2%	-10%	-15%	0%	0.76
5	mei_m	R	0.72	0.72	0.72	0.72	0.72	0%	0%	0%	0%	0.72

Table 31 - Sensitivity-based ranking (Eurostat & OECD.Stat)

To reach a final selection, assessment step 3 compares datasets and hypotheses in consideration of the data mix requirement. This required to define common groups of datasets for Eurostat & OECD.Stat data. A common group is defined as datasets containing the same information that is represented by different views such as alternative ways of variable calculation. Within the set of 41 datasets, four common groups were identified and datasets were compared along their score-based ranking within each group as shown in Table 32. The assessment step resulted in an additional removal of seven inferior datasets.

Common group	Commonality	dataset_id	Score-based ranking				
			(I) Basic	(II) Short	(III) Long	(IV) High_Corr	TOTAL
1	Industry indicators as (a) indices and (b) growth rates	(a) ei_isin_m	2	2	2	2	2
		(b) ei_isir_m	1	1	1	1	1
2	Turnover in services as (a) growth rates or (b) indices	(a) ei_isse_q	1	1	2	1	2
		(b) ei_isset_q	2	2	1	2	1
3	Oil stocks measured in (a) tonnes or (b) days equivalent (emergency stocks)	(a) nrg_142m	2	2	2	2	2
		(b) nrg_143m	1	1	1	1	1
4	Harmonized index of consumer prices measured as (a) annual rates of change with constant taxes or (b) indices (constant taxes) or (c) annual rates of change with basis year 2015 or (d) indices (basis 2015) or (e) 12-month average rate of change (basis 2015)	(a) prc_hicp_cann	5	5	5	5	5
		(b) prc_hicp_cind	2	2	2	2	2
		(c) prc_hicp_manr	1	1	1	1	1
		(d) prc_hicp_midx	3	3	3	3	3
		(e) prc_hicp_mv12r	4	4	4	4	4

selected dataset

Table 32 - Assessment step 3 results for Eurostat & OECD.Stat

In case of ERP and financial data, the definition of common groups was generally given by the datasets, that is to say all remaining hypotheses of a dataset build such a group. It was therefore ensured that each dataset is represented in detailed assessment, except for those eliminated

⁴⁶ The dataset namq_10_a10 contains gross domestic product and income while namq_10_a10_e supplements this data with employment indicators.

EVALUATION

during assessment step 1. A high level of diversity was hereby maintained for the data mix within each source. Hypotheses of ERP datasets either represented alternatives based on similar dimensions and the same set of metrics or they were complementary by using completely different sets of dimensions or variables. As a consequence, alternative hypotheses were assessed as common groups within a dataset. Table 33 summarizes the result of assessment step 3 for ERP data. While the assessment of common groups identified 18 superior hypotheses, both best ranked hypotheses of orders data and the second hypotheses of logistics costs were kept during review & final approval with the business user.

Common group	Commonality	hypothesis_id	Score-based ranking				
			(I) Basic	(II) Short	(III) Long	(IV) High_Corr	TOTAL
sales_A	sales variables aggregated by business or customer subsets	sales_1	1	1	1	1	1
		sales_2	2	2	2	2	2
		sales_3	3	3	3	3	3
		sales_4	4	4	4	4	4
sales_B	selected sales variables aggregated by order types	sales_5	2	2	2	2	2
		sales_6	1	1	1	1	1
orders	no grouping required	orders_1	1	2	1	1	1
		orders_2	2	1	2	2	2
		orders_3	3	3	3	3	3
quotations_A	comprehensive variable set aggregated by different business or customer subsets	quotations_1	3	3	3	3	3
		quotations_2	1	1	1	1	1
		quotations_3	2	2	2	2	2
quotations_B	selected variables with additional filtering	quotations_4	2	2	2	2	2
		quotations_5	1	1	1	1	1
purchasing_prices_A	price variables aggregated by vendors or materials	prices_1	2	2	2	2	2
		prices_2	1	1	1	1	1
purchasing_prices_B	price variables aggregated by vendors or materials with additional filtering	prices_3	2	2	2	1	2
		prices_4	1	1	1	2	1
purchasing_spent_A	spent variables aggregated by vendors or materials	spent_1	3	3	2	2	2
		spent_2	1	1	3	3	3
		spent_3	2	2	1	1	1
purchasing_spent_B	spent variables aggregated by material types with additional filtering	spent_4	3	2	1	2	2
		spent_5	2	3	2	3	3
		spent_6	1	1	3	1	1
total_inventory_A	inventory variables aggregated by material types	inventory_1	3	3	3	3	3
		inventory_2	4	4	4	4	4
		inventory_3	1	1	1	2	1
		inventory_4	2	2	2	1	2
total_inventory_B	inventory variables aggregated by market or customer subsets	inventory_5	1	1	1	1	1
		inventory_6	3	3	3	3	3
		inventory_7	2	2	2	2	2
supplier_rating	no grouping required	rating_1	1	1	1	1	1
		rating_2	2	2	2	2	2
supplier_delivery_performance	no grouping required	delivery_1	4	5	4	5	5
		delivery_2	5	4	5	4	4
		delivery_3	1	1	1	1	1
		delivery_4	2	2	2	2	2
		delivery_5	3	3	3	3	3
		delivery_6	6	6	6	6	6
payment_performance_A	payment performance variables aggregated by business or customer subsets (joint assessment of creditors and debtors)	debtor_1	1	1	1	2	1
		debtor_2	3	3	3	1	3
		creditor_1	2	2	2	3	2
		creditor_1	4	4	4	4	4
payment_performance_B	payment terms variables for debtor segments only	debtor_3	1	1	1	1	1
logistics_costs	no grouping required	logistics_1	1	1	1	1	1
		logistics_2	2	2	2	2	2

 selected dataset

Table 33 - Assessment step 3 results for ERP

There was no need for additional definition of common groups for the datasets of the financial database because they shared a common set of hypothesis types. Assessment step 3 revealed

that aggregation of company categories generally results in higher scores. There were two exceptions to this general rule to prefer company categories over individual companies: trailing multiples and stock market data A. Their total ranking score for type I was slightly higher compared to type II. Moreover, type II and type III generally showed relative small scoring differences such that the former was chosen as default hypothesis because of its independence of an additional heuristic. Based on the findings in general assessment, the decision was taken to consistently utilize type II also for the two exceptions during review & final approval. Figure 88 summarizes the results of assessment step 3 for all groups of data sources. The higher rates of reduction for financial and ERP data results from hypothesis-based time series generation. Alternative hypotheses were formulated for a dataset and general assessment determined the best hypothesis based on the quality and relevance of time series of each hypothesis. Overall, general assessment reduced the scope of Eurostat & OECD.Stat from 146 to 34 datasets which represents a total reduction of 77%.

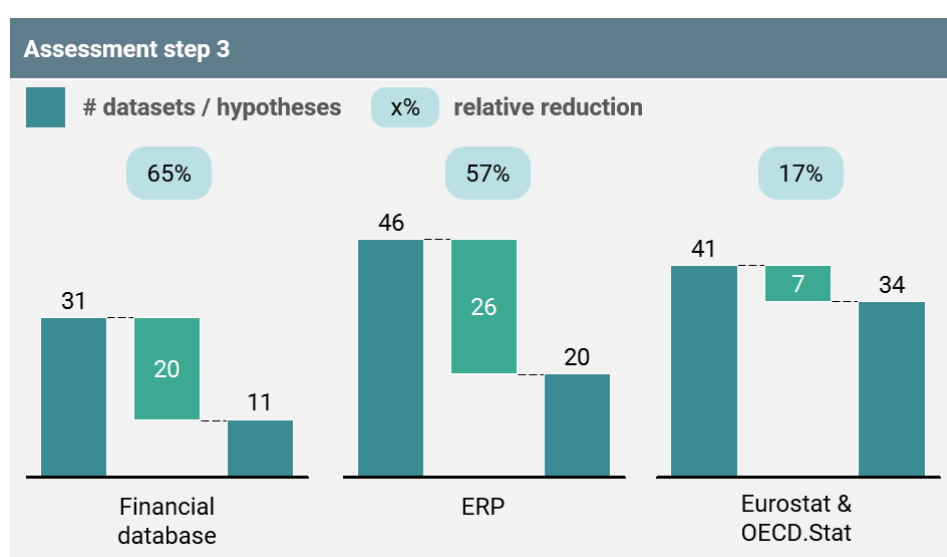


Figure 88 - Assessment step 3 results (all data sources)

All 65 datasets and hypotheses were reviewed during detailed assessment based on the evaluation tool output because this report provides details on individual time series. It is important to note that assessment step 4 generally focuses on time series that fulfill the threshold levels for quality and relevance as defined by the basic filter setup. Detailed assessment during the case study utilized time series with relevant correlation with both target time series.

The first task of assessment step 4 is to remove spurious correlations and Figure 89 provides an example for data on producer prices in industry (sts_inppd_m) from Eurostat. The dataset comprises four different dimensions that were used for automated generation of time series. After filtering for quality and relevance, 11 geographic instances (geo), 16 economic sectors (nace_r2), and two different forms of calculation (unit) represented a total of 18 time series with the business trend indicator for the domestic output price index (indic_bt) as the only variable. Unit represents a technical dimension, with index value or percentage change compared to same

EVALUATION

period in previous year (pch_sm) in the given example, and therefore should not be regarded for removal of spurious correlations. The same holds true for seasonal adjustments (s_adj) where all time series share the same adjustment type. For efficient assessment of spurious correlations, the review started with the dimension having the smallest number of instances. The analysis of geo showed that 8 instances represent a rather arbitrary mix of single economies, however, also includes 3 time series representing various subsets of the *European Union (EU)*. Consequently, the economic sectors for these time series were assessed. Each EU-based region showed correlation in the same sector, however, building of pleasure and sporting boats clearly represents a spurious correlation for the studied use case.

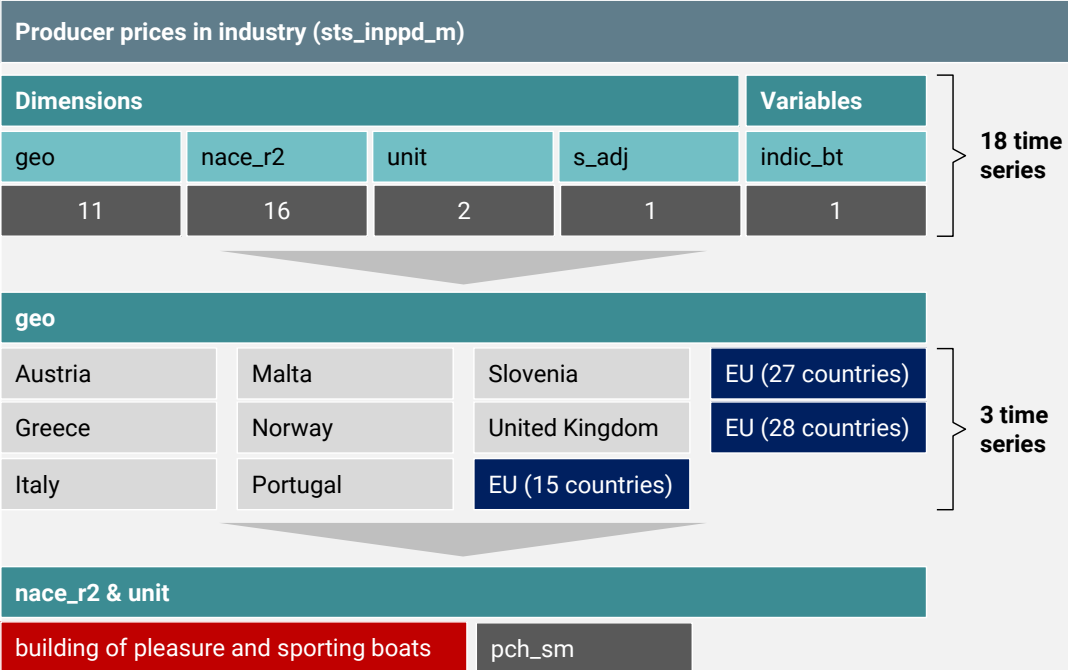


Figure 89 - Assessment step 4: example for spurious correlations

The second task of assessment step 4 is to identify and remove inferior filters and Figure 90 serves as an example. The hypothesis sales_1 generated time series for various sales variables, such as sales volume, sales value, or margins, aggregated by different business segments. Furthermore, the hypothesis considered two filters that removed irrelevant segments (segment_out) and customers (customer_out). The remaining two filters had two instances each that are interlinked. Both filters either represented actual revenues (a + a) or total revenues (b + b) that also includes internal sales, claims or pre-series, for example. These filters therefore represented alternative subsets of the data where business users could not define a priori which one is better. The hypothesis provided 50 time series of adequate quality and sufficient correlation of which only 2 were represented by the filter setup of total sales. Moreover, this setup reduced the dimensions covered from 6 to 1 and the variables from 11 to 2 while the setup for actual sales included all available for both. The filter setup (b + b) clearly is inferior and related time series were removed accordingly.

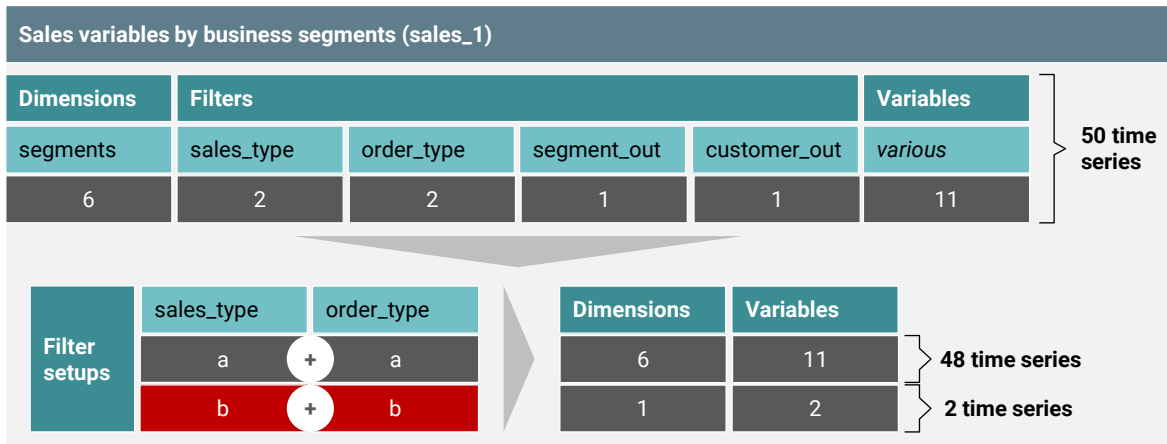


Figure 90 - Assessment step 4: example for inferior filters

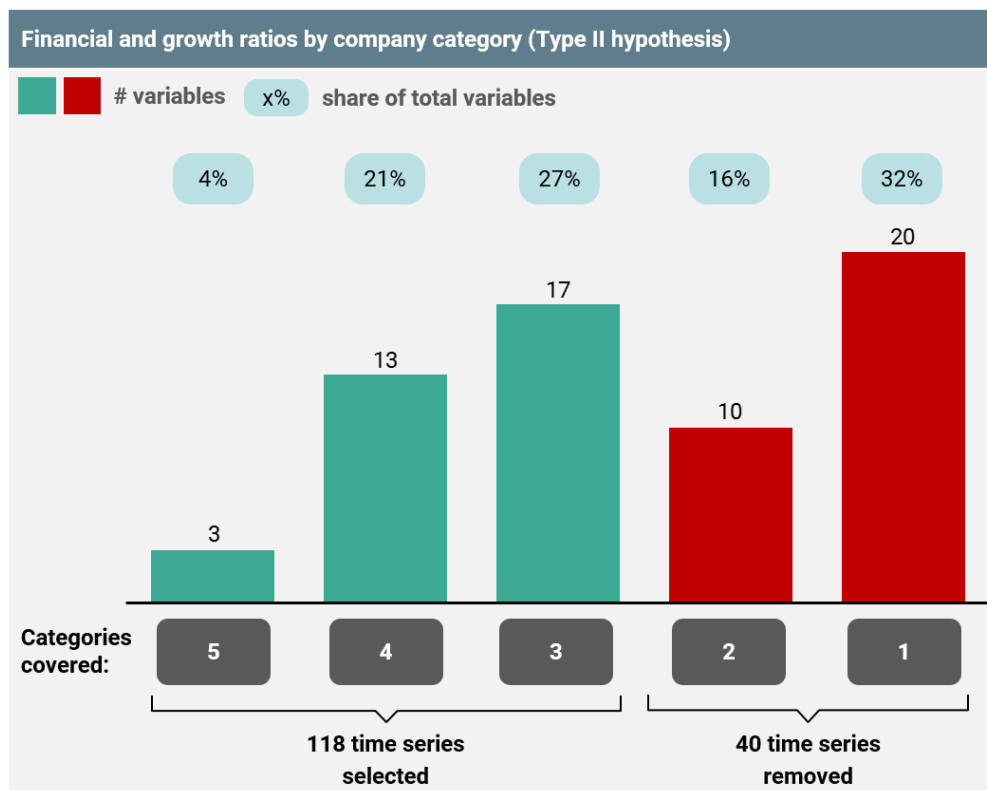


Figure 91 - Assessment step 4: example for sparse variables

Removal of sparse variables is the last task of assessment step 4 and Figure 91 provides an example based on the selected hypothesis for financial & growth ratios from the financial database⁴⁷. After assessment of the mandatory filter regarding relevant companies⁴⁸, the hypotheses contained 158 relevant time series distributed across 6 out of the 7 company categories and 63 variables. Ratio of total debt to equity, net income margin, or return on capital are examples for variables here. No variable occurred in all company categories, however, 33 variables covered three or more categories and therefore more than half of the available categories. The remainder variables were defined as sparse and thus removed from the hypothesis. Selected variables represented 52% of all variables considered but 75% of time series. This is a typical outcome of sparse variable removal as variables occurring across various dimensions are prioritized.

Substantiation of spurious correlations, inferior filters, or sparse variables is the first selection task of assessment step 5. Lack of clarity about spurious correlations is rather an exception because hypothesis-generated time series deliberately construct time series and the business background of the BDA manager usually enables detection for time series from automated generation. The example of Figure 92 consequently focuses on the remaining two cases. The hypothesis *inventory_3* generated time series for inventory variables aggregated by commodity groups as well as used different filter setups. General assessment resulted in two alternative sets of time series that represented the same commodity groups and inventory variables but differed in the filter setup for material flow. The first set did not restrict materials to a certain flow type and the second one only included inbound flow materials. As both sets covered the same commodity groups as well as variables, none was clearly inferior to the other and the filter setup was discussed with business users during the decision workshop of assessment step 5. In the given example, the restriction to inbound flow materials was disregarded with the help of business domain expertise. Figure 92 also shows an excerpt of the decision template for hypothesis *sales_1* which is already explained in the discussion about inferior filters above. After removing the inferior filter setup, the review by the BDA manager revealed one variable with sparse characteristics. It only appeared for two business segments while other variables typically showed up for five or more segments. However, there could be a business reason for this behavior of the variable unknown to the BDA manager and therefore required clarification with the business user. The workshop revealed that express delivery is the standard for the two segments and changes in express costs from other segments would only be seen as valuable

⁴⁷ Note: As sparse variables were a key issue for financial data, detailed assessment was performed on joint evaluation tool reports for hypotheses of type II and type III (if applicable) in order to increase the number of observable cases.

⁴⁸ The filter setup for removal of indirectly relevant companies was selected.

leading indicator for future sales. As a result of the decision workshop, `costs_express` was removed as sparse variable.

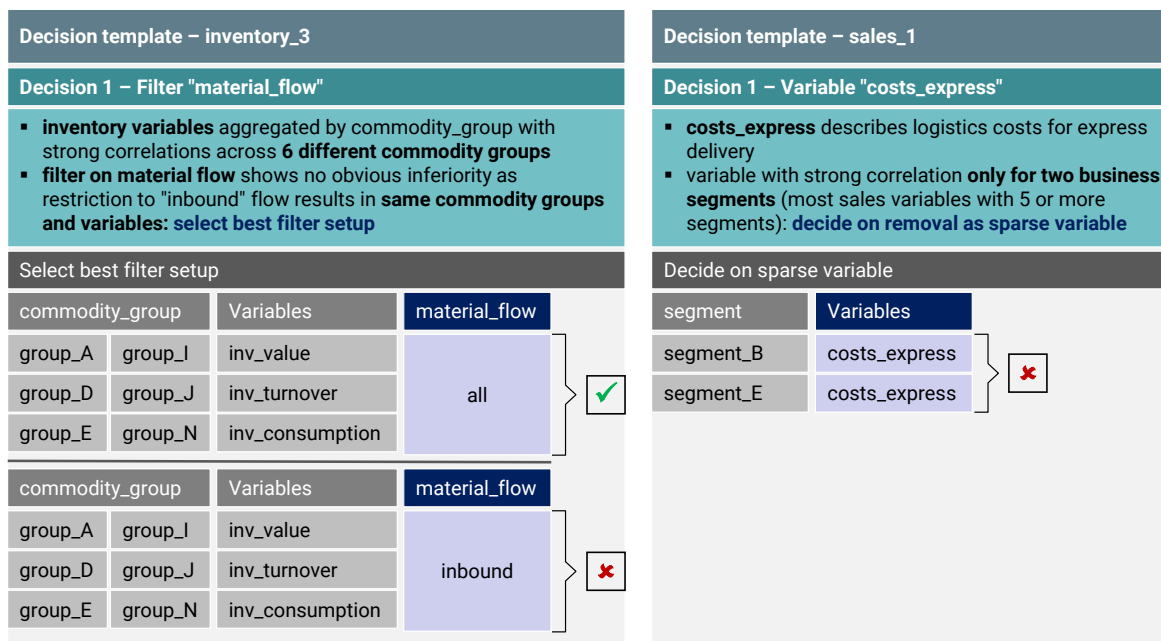


Figure 92 - Assessment step 5: examples for substantiation of inferior filters and sparse variables

The key selection task of the decision workshop in assessment step 5 is to remove redundant information because redundancy is not directly addressed in pervious assessments. Figure 93 provides an example⁴⁹ for the already known hypothesis sales_1. In addition to inferior filter setups and sparse variables, the BDA manager identified three groups of potentially redundant sales and profitability variables. The first group represented three different types of measurement for sales volume where joint observation was not assumed insightful from a business perspective such that *gross_quantity* was selected as the most frequently used variable. Total revenue (*rev_total*) and revenue excluding one-time effects (*rev_product*) also represented redundant information from a business perspective in the same way as the two different types of profitability margins (*margin_a* and *margin_b*). For both groups, the most common variables were selected as well. The decisions taken represent a removal of 19 time series out of a total 33 time series which equals a reduction of 58%. This example consequently underlines the importance of removing redundant information. Some additional clarifications might be required for anonymized data. For example, the hypothesis on debtor payment performance selected during general assessment created time series for individual customers that were anonymized. In order to decide whether the customers showing correlations are meaningful, the list of customers was reviewed by the business user. In this case, 4 out of 55 customers were

⁴⁹ The example is simplified as one of the dimensions for each variable group had characteristics of a weak correlation with one of the alternative variables. The actual decision template included the selection decision on weakly correlated time series as well.

EVALUATION

selected as relevant for the correlating payment performance metrics. The selected set of customers was provided to the BDA manager in form of anonymized customer IDs.

Decision template – sales_1						
Decision 2 – Redundant sales and profitability variables						
<ul style="list-style-type: none"> ▪ Different sales and profitability variables aggregated by business segments with strong correlations across 4-5 segments ▪ Variable groups for sales volume, sales revenues and profitability with potential redundancies: remove redundant variables 						
Remove redundant sales volume variables						
Segment	Variable	Segment	Variable	Segment	Variable	
segment_A	net_quantity	segment_A	gross_quantity	segment_A	pieces	
segment_C		segment_C		segment_C		
segment_D		segment_D		segment_D		
segment_F		segment_F		segment_F		
segment_I		segment_I		segment_I		
✘		✔		✘		
Remove redundant sales revenues variables						
Segment	Variable	Segment	Variable			
segment_A	rev_total	segment_A	rev_product			
segment_C		segment_C		segment_C		
segment_D		segment_D		segment_D		
segment_F		segment_F		segment_F		
segment_I		segment_I		segment_I		
✔		✘				
Remove redundant sales profitability variables						
Segment	Variable	Segment	Variable			
segment_A	margin_a	segment_A	margin_b			
segment_C		segment_C		segment_C		
segment_F		segment_F		segment_F		
segment_I		segment_I		segment_I		
✔		✘				

Figure 93 - Assessment step 5: examples for removal of redundant information

The Eurostat dataset ei_bsin_m_r2 represents monthly data of a business survey in industry. Based on automated generation, time series for different survey variables, with or without adjustments such as seasonality, and various aggregations for geographic regions or countries were evaluated. As a result, two time series representing the assessment of the current level of

stocks of finished products (*BS-ISFP*) by players of the manufacturing industry across the *European Union with 28 member states (EU28)* and the *Euro Area (EA19)* remained after general assessment. However, the review by the BDA manager revealed that further survey variables for the same geographic regions showed correlation slightly below the set target level of $R_Target > 0.7$ for Pearson correlation. The time series based on the industrial confidence indicator (*BS-ICI*) and the production expectations over the next quarter (*BS-IPE*) had PCC values of $R > 0.6$ and were therefore additionally selected as modeling input as shown in Figure 94. This more than doubles the initial set of time series for modeling which also underlines the significance of the last selection task.

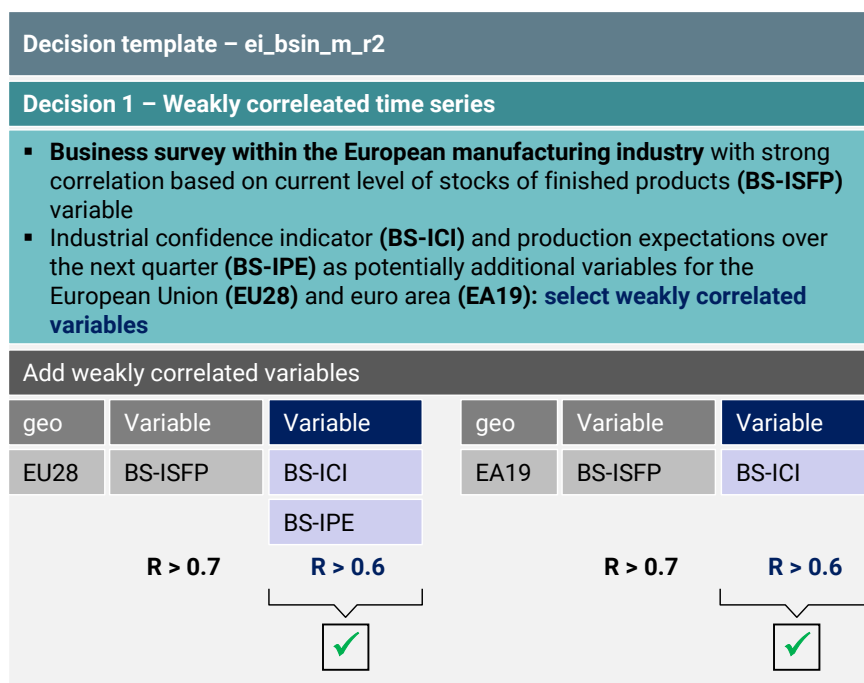


Figure 94 - Assessment step 5: example for weakly correlated time series

Completion of assessment step 5 constitutes the set of time series used during modeling. As shown in Figure 95, the modeling input consisted of 1,360 time series that were prioritized from a generated set of more than 4 million time series. This includes a reduction by 610 time series during detailed assessment going from assessment step 4 to 5. However, assessment step 5 did not reduce the number of time series for all data sources simultaneously. The majority of reduction stems from ERP data where many alternative filter setups and redundant variables were removed with the help of business domain knowledge. This is not surprising as the hypothesis-based time series generation deliberately created redundant sets of time series ensuring to find some with strong correlation. In case more than one alternative shows strong correlation, the help of the business user is required to make selection decisions. Time series from the financial database were also generated on the basis of hypotheses, however, they were built on a simpler structure and also without testing of alternative filter setups. Furthermore, no anonymization was applied such that no domain-specific knowledge was required here and the BDA manager did not discover many redundancies during the review of hypotheses. As a

consequence, assessment step 5 was skipped for this data source. The number of time series actually increased for Eurostat & OECD.Stat data during the final assessment step which is mainly driven by additions of weakly correlated time series. Strong presence of weak correlations can be partially attributed to automated time series generation. Automated generation does not benefit from construction or integration of data as in the case of hypothesis-based generation, and therefore information quality of time series is not enriched which potentially makes it more difficult to show strong correlations. The distribution of time series was very uneven after assessment step 4 ranging from 8% for Eurostat & OECD.Stat and 65% for ERP data. This changed after the final assessment step with a range between 28% for Eurostat & OECD.Stat and 38% for financial data. With strict application of assessment step 5 on the financial database, a more even distribution would be expected. This shows that the detailed assessment is also important to keep the balance in the data mix before entering modeling.

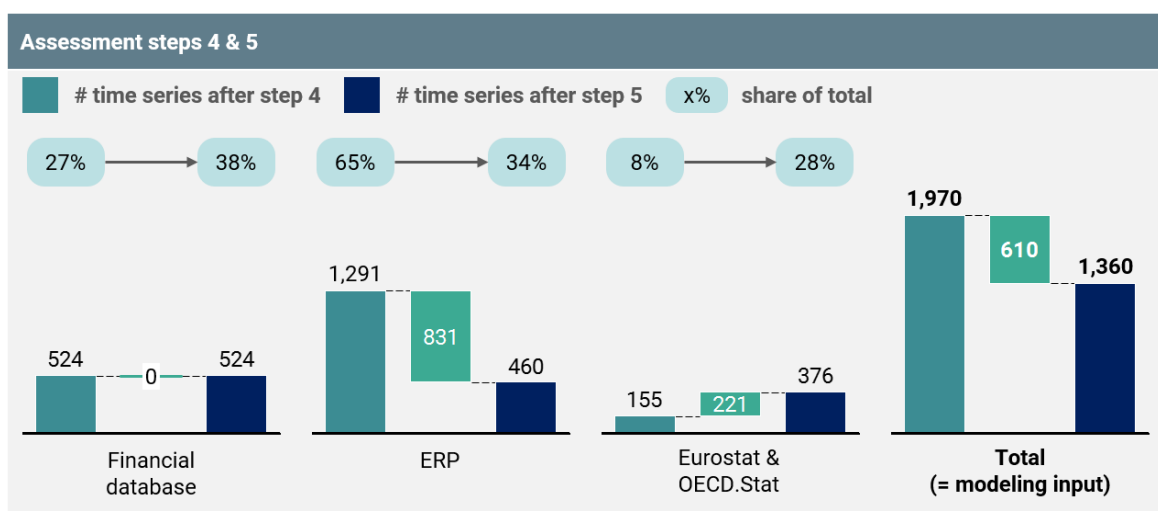


Figure 95 - Assessment steps 4 & 5 results (all data sources)

5.3.6 Modeling & evaluation

5.3.6.1 Model selection and test design

The project plan of the use case prescribed the use of standard models for predictive analytics from both classification and regression categories. The BDA manager together with the data scientist identified eleven appropriate models. Table 34 shows the segmentation into starting as well as advanced models after approval from the business user. The model selection template also shows a brief summary of functional principle⁵⁰, rationale for selecting the models, and the explanation for using different model designs in case of SVM and elastic net.

⁵⁰ Selected starting models are discussed in detail in *Section 2.4* and therefore the overview only includes a brief description of functional principles.

	Classification	Regression
Starting models	K-Nearest Neighbors (kNN)	Autoregressive integrated moving average (ARIMA)
	find a predefined number (k) of observations closest in distance to a new observation and predict the class from these closest points	model that only uses the time series to be forecasted as model input
	simplicity and configuration-less application	conventional approach for time series forecasting (baseline model)
	Support Vector Machines (SVM)	Elastic net
	- SVM model is a representation of observations as points in space, mapped so that the examples of the separate classes are divided by a clear gap that is as wide as possible - new observations are mapped into that same space and predicted to belong to a class based on which side of the gap they fall on	multiple linear regression with integrated regularization (integrated feature selection in order to avoid overfitting)
	performance in practical applications and extensibility to non-linearity	efficiency and interpretability based on embedded feature selection
	a) linear kernel: basic approach with less computational effort b) radial basis function kernel: non-linear approach that is at least as good as the linear kernel but comes with higher computational effort	a) standard: modeling input as defined during data preparation b) reduced: additional feature reduction filter
	Decision tree	
	- decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute (feature time series) - each branch represents the outcome of a test, and each leaf (or terminal) node belongs to a specific class	
	high interpretability and simplicity	
Advanced models	Ensemble classifiers (e.g., random forests)	State space models
	high interpretability and simplicity	high suitability for time series forecasting
	Sequence classifiers	Holt-Winters
	applicability on time dependent data	additional exponential smoothing of time series
	Artificial Neural Networks (ANN)	Kalman filter
	analyse „hidden“, perhaps non-linear, relationships between variables	additional reduction of outliers and estimation of missing values

model
functional principle
selection rationale
alternative model designs

Table 34 - Model selection template (business user approval)

The starting models for classification are based on the scikit-learn library that provides open source models based on Python (scikit-learn 2017d). The *k-Nearest Neighbors* (*k*NN) model was selected for its simplicity and configuration-less application as it requires a relatively small number of model parameters only. *Support Vector Machines* (*SVM*) have proven to be very effective in practical applications with high dimensional input data and sparse training data which both apply to the case study. They furthermore enable extension with kernel methods that allow for non-linear separation of different classes which is why the *radial basis function* (*rbf*) kernel was used as alternative model design. The linear kernel will typically provide performance less or equal to *rbf* kernel, however, it is computationally less costly. The underlying approach of *decision trees* is simple and the major selection rationale was the interpretability of its predictions which is premised on the possibility to visualize the model. Documentation of the applied library models is provided by the following list:⁵¹

- a) *k*NN: <http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>
- b) *SVM*(linear): <http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>
- c) *SVM*(*rbf*): <http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>
- d) *Decision tree*: <http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

The *Auto-Regressive Integrated Moving Average* (*ARIMA*) model is a traditional approach to sales forecasting and was selected as baseline model for the regression category. *Elastic net* was the starting regression model because it includes regularization as embedded feature reduction. It provides an efficient way to further reduce the feature time series in order to increase generalization performance against the limited number of training observations. Furthermore, the reduced number of features increases the interpretability of the underlying multiple linear regression model. The filter approach of data preparation mainly aims for prioritization of feature time series based on easy-to-understand quality and relevance metrics as well as with the aid of domain knowledge. While the resulting modeling input is approved by the business user, it still might not be optimal for modeling. Multicollinearity describes the phenomenon where features in a multiple regression can be constructed as a linear combination of other features. This results in high sensitivity of the model to changes in the input data and thus poses a challenge to generalization (Gujarati 2003, pp. 341–370). While regularization generally addresses the problem, it can be difficult for the algorithm to effectively remove such collinear feature time series. An alternative model design with additional feature selection filter (*reduced elastic net*) was therefore included in order to examine the effect on model performance. The

⁵¹ Last access date: 10/25/2017

standard elastic net utilized the model input as provided by data preparation and documentation of the model library used can be found here:

https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html.⁵²

As the assessment of starting classification and regression models will show, their performance was already sufficient as proof of concept. Consequently, the advanced models were not built and assessed such that a detailed discussion is dispensed here.

The test design is based on stratified k-fold cross-validation to estimate model performance. For the sales forecast models in the case study, k was selected as 10 which implies a separation of available data into 10 folds and 10 repetitions of model learning. 10-fold cross validation is a typical choice for model learning (Hastie et al. 2015, p. 13). While two different target time series were employed for time series prioritization, the model learning focused on sales volume measured in units. Furthermore, the annual growth rate of sales volume for each month is defined as forecast variable such that each model aimed to provide a monthly prediction of sales growth over the next 12 months. The available data history of the target time series was exactly seven years⁵³ with monthly frequency such that 84 observations were available for model training. An additional feature of the test design was the use of two different sets of modeling input data. Following the method of data preparation, the modeling input was defined with assessment step 5 at the end of time series prioritization. However, the resulting set of time series after assessment step 4 was used as alternative modeling input in order to evaluate the benefit of assessment step 5. This is motivated by the cost of including domain knowledge of the business user in addition to the previous review by the BDA manager. To facilitate the discussion of model performance, data input after step 4 is referred to as *BDA input* and after step 5 as *domain input*.

5.3.6.2 Model building and assessment: classification

The classification-based forecast requires to define classes for sales growth. As these classes represent the model output, they need to be meaningful from a business perspective. Figure 96 shows the three different classifications as defined by the business user⁵⁴ including the distribution of observations from the target time series.

⁵² Last access date: 10/25/2017

⁵³ Based on a total history of eight years, the annual growth on a monthly basis can be calculated for seven years.

⁵⁴ Exact growth ranges are omitted for confidentiality reasons.

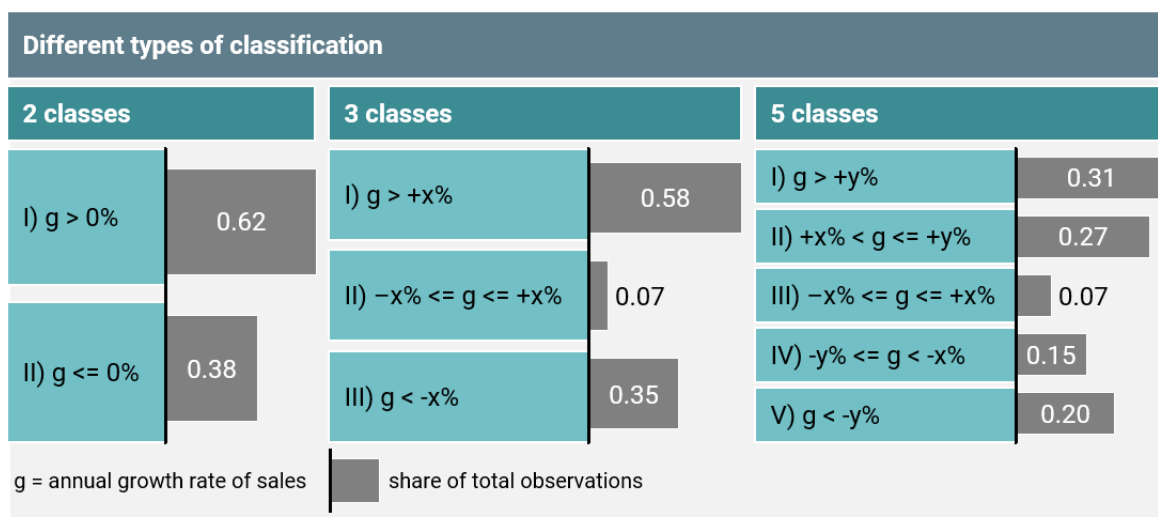


Figure 96 - Definitions and distributions of classes

The data conditioning for classification produces different model inputs. Normalization was required due to application of kNN and SVM which represent distance-based models. Different scales of the feature time series cause potentially unwanted weighting in favor of features with larger scales. Two alternative normalization approaches were tested and their calculations are described in Table 35. Z-score normalization transforms each feature to zero mean ($\mu = 0$) and unit variance ($\sigma = 1$) whereas min-max normalization brings features to the range [0,1].

Normalization	Calculation
Z-score normalization	$value_{new} = \frac{value_{original} - \mu (mean)}{\sigma (standard\ deviation)}$
Min-max normalization	$value_{new} = \frac{value_{original} - min_{original}}{max_{original} - min_{original}}$

Table 35 - Normalization approaches for classification [based on (Roiger 2017, pp. 208–209)]

There was no further time window approach applied to the modeling input such that the value of each time series for a given point in time represented the model input for learning. In the given use case of 12-months sales growth forecasting, the feature input is represented by a vector built from the values of all time series 12 months prior to the forecast period. This model input vector potentially had 1,970 entries for BDA input and 1,360 for domain input. However, not all time series had monthly frequency and therefore did not provide a value for each monthly observation. In order to sustain the simple approach without additional time window approach⁵⁵ the modeling was restricted to monthly feature time series only. This restricted the model input

⁵⁵ Dynamics of time series with lower frequency could be calculated for a given time window in order to fill in the frequency-based gaps.

to 1,531 feature time series for BDA input and 879 for domain input. Within this reduced feature set, 396 time series had missing values in the former case and 212 in the latter case. Missing values of the selected time series from ERP, financial and economic data are not meaningful and therefore should not be regarded as model input. Two different measures were applied to handle missing values. They were substituted with the mean value of the feature time series (*impute*) or, on the other hand, features with missing values were disregarded (*remove*). Figure 97 summarizes data conditioning for classification.

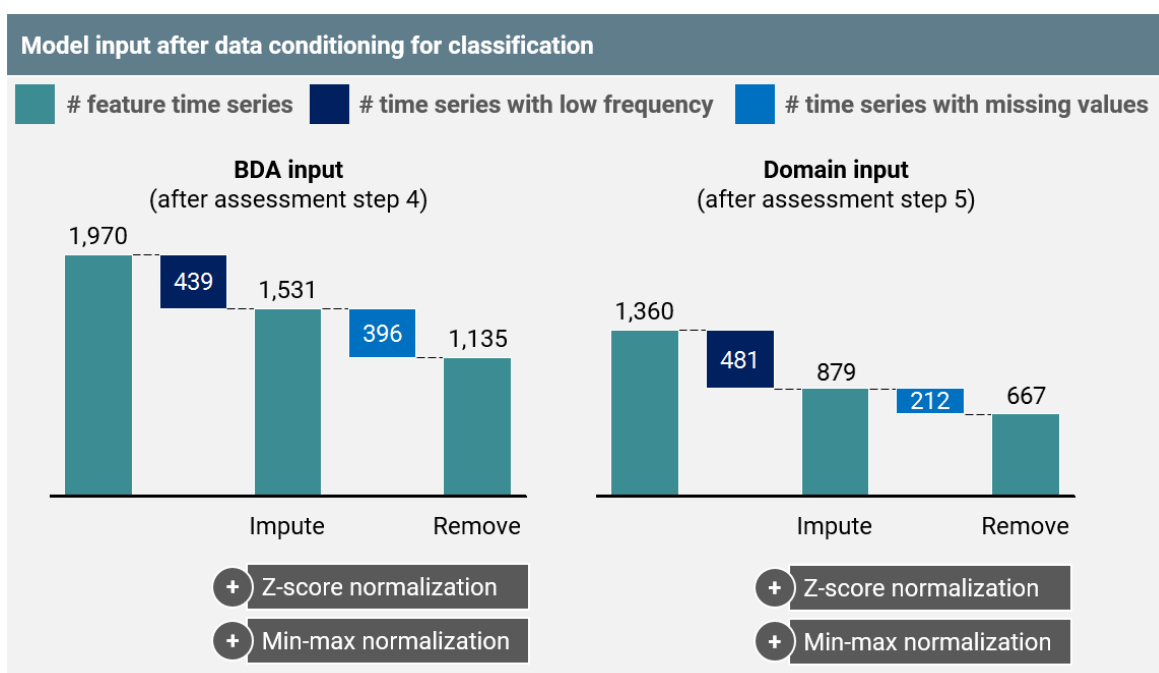


Figure 97 - Data conditioning for classification

Each classification was built with the selected models whereby each model was applied with its default settings for parameters as described by the documentation referenced before. Table 36 presents classification performance based on the accuracy measure for BDA input with different approaches for normalization and missing values handling in the case of 2 classes. The results clearly show that normalization is required to get to a performance level of 80% or higher accuracy. Furthermore, min-max normalization does only provide reasonable performance for kNN and SVM(linear) models while it does not create any advancement for SVM(rbf). All models benefit from Z-score normalization and even show their maximum performance here, except for kNN with highest accuracy for min-max normalization in combination with remove, however, with a small margin only. The effects of impute and remove are rather inconclusive as both are superior in different cases depending on the model and normalization measure. The margins between both missing value handling measures are also relatively low such that remove in combination with Z-score normalization was chosen as conditioning setting for further model assessment. This represents a conservative approach as the available model input was restricted to higher quality time series without missing values. It should be noted that the consistent performance of the decision tree reflects expectations. Decision trees do not

EVALUATION

necessarily require normalization for good performance and they are also good at handling missing values.

	Z-score & impute	Z-score & remove	Min-max & impute	Min-max & remove	Impute only	Remove only
kNN	0.820	0.845	0.833	0.870	0.640	0.640
SVM(linear)	0.885	0.808	0.838	0.848	0.738	0.703
SVM(rbf)	0.883	0.883	0.620	0.620	0.620	0.620
Decision tree	0.770	0.713	0.700	0.688	0.710	0.700

 accuracy > 80%

Table 36 - Accuracy performance for 2 classes (BDA input)

Table 37 summarizes the overall performance for all types of classification utilizing domain input based on the data conditioning as described above. In addition to accuracy all major classification measures are shown. Precision, recall, and F1 were investigated each using micro-averaging as well as macro-averaging and the latter is shown because of the effect due to the lowly populated center class for 3 and 5 classes. The major result is the strong performance of classification with 3 classes. The maximum accuracy across all models drops only from 88.3% in case with 2 classes to 85.3% whereby 3 classes provide much better insight into sales growth dynamic from a business perspective. Even more insightful is the case with 5 classes but maximum accuracy is significantly reduced to 62.0% which is not an adequate performance for business applications. SVM(rbf) provides maximum accuracy across all three cases and it also shows the best performance regarding the other measures. This is illustrated by the model ranking which is based on each individual performance measure. SVM(rbf) and kNN are consistently ranked first and second while they are also the only two models with performance values higher than 80% for 2 classes and 3 classes. SVM(linear) and decision tree alternatively rank in third and last position with a maximum accuracy of 71.0% for the former in case of 3 classes. Looking at the best two models in the most promising case of 3 classes in some more detail reveals that precision, recall, and F1 performance is reduced in case of macro-averaging. This indicates that performance for lowly populated classes is worse. In the given case, this is most likely a result from class II) which represents only 7% of all observations. Imbalanced data is a common issue for classification and Sun et al. (2009) provide an overview of possible remedies. For instance, they include random oversampling of the lowly populated class, adjusting the learning algorithm to apply a higher penalty for misclassification, or the use of ensemble classifiers that provide predictions based on different classification models (Sun et al. 2009, pp. 700–710). Such optimization strategies for better generalization performance were not applied in the case study. However, it holds potential for improvement considering the strong overall performance despite weak performance in the sparse class.

2 classes								
	Model performance				Model ranking			
	kNN	SVM (linear)	SVM (rbf)	Decision tree	kNN	SVM (linear)	SVM (rbf)	Decision tree
accuracy	0.870	0.725	0.883	0.783	2	4	1	3
precision_macro	0.885	0.710	0.903	0.784	2	4	1	3
recall_macro	0.847	0.703	0.857	0.792	2	4	1	3
f1_macro	0.851	0.690	0.863	0.760	2	4	1	3

3 classes								
	Model performance				Model ranking			
	kNN	SVM (linear)	SVM (rbf)	Decision tree	kNN	SVM (linear)	SVM (rbf)	Decision tree
accuracy	0.808	0.710	0.853	0.696	2	3	1	4
precision_macro	0.630	0.561	0.660	0.556	2	3	1	4
recall_macro	0.668	0.571	0.694	0.543	2	3	1	4
f1_macro	0.639	0.551	0.668	0.629	2	4	1	3

5 classes								
	Model performance				Model ranking			
	kNN	SVM (linear)	SVM (rbf)	Decision tree	kNN	SVM (linear)	SVM (rbf)	Decision tree
accuracy	0.610	0.545	0.620	0.523	2	3	1	4
precision_macro	0.500	0.485	0.509	0.457	2	3	1	4
recall_macro	0.543	0.499	0.546	0.427	2	3	1	4
f1_macro	0.498	0.470	0.503	0.412	2	3	1	4

Table 37 - Overall classification performance (Domain input)

In order to substantiate the assessment of the classification models, *label shuffling* was applied. Label shuffling randomly generates a series of observations while maintaining the distribution of classes from the original target time series. Table 38 provides a simplified example that explains how shuffling works. Shuffling was repeated 100 times for the sales target time series and each model was assessed based on the shuffled time series. The mean (μ) and standard deviation (σ) of performance measures are recorded across all repetitions.

Timestamp	Original target time series	Shuffling 1	Shuffling 2	Shuffling 3
Jan 17	-30%	-30%	0%	50%
Feb 17	-30%	0%	50%	-30%
Mrz 17	0%	-30%	50%	-30%
Apr 17	0%	50%	0%	50%
Mai 17	50%	0%	-30%	0%
Jun 17	50%	50%	-30%	0%

Table 38 - Example for label shuffling

It can be assumed that shuffling performance is normally distributed such that actual performance outside the range of $\mu + 2 \times \sigma$ is not random with a 95% confidence level. Table 39 presents the results from the test based on label shuffling for SVM(rbf) in the case of 3 classes and domain input. It shows that actual performance lies clearly outside of the confidence range and therefore is not random. To put it differently, the increase in performance from random shuffling to actual observations stems from information about the target time series behavior contained in the data input selected. This is also why the observed performance from the classification models can be interpreted as a lower bound for the information content of data input regarding the target time series. The random shuffling test was performed for all classification types (3), all models (4), all performance measures (7) and the BDA input as well

EVALUATION

as domain input (2). In total, the test was performed 168 times ($3 \times 4 \times 7 \times 2 = 168$) and not a single case failed which provides additional confidence in the overall model performance.

Performance measure	Actual performance	Shuffling performance - mean	Shuffling performance - standard deviation	Shuffling - 95% confidence range	Test result
accuracy	0.853	0.553	0.038	0.627	OK
precision_micro	0.853	0.553	0.039	0.630	OK
recall_micro	0.853	0.552	0.037	0.624	OK
f1_micro	0.853	0.554	0.037	0.626	OK
precision_macro	0.660	0.310	0.077	0.460	OK
recall_macro	0.694	0.417	0.040	0.495	OK
f1_macro	0.668	0.323	0.043	0.408	OK

Table 39 - Label shuffling results for SVM(rbf) with 3 classes

Table 40 provides a comparison of accuracy performance between BDA input and domain input for all classification types and models. The result indicates that model performance benefits from inclusion of domain knowledge by business users during time series prioritization. In 10 out of 12 cases the accuracy with domain input is higher or at least equal compared to BDA input. The increase of accuracy averages 6.6 percentage points across all cases. This finding generally holds true when looking at other performance measures as well. As a consequence, assessment step 5 does not only increase business user confidence in the modeling input but can also increase model performance.

		kNN	SVM (linear)	SVM (rbf)	Decision tree
2 classes	BDA input	0.845	0.808	0.883	0.713
	Domain input	0.870	0.725	0.883	0.783
3 classes	BDA input	0.761	0.272	0.840	0.694
	Domain input	0.808	0.710	0.853	0.696
5 classes	BDA input	0.648	0.309	0.614	0.447
	Domain input	0.610	0.545	0.620	0.523

Table 40 - Accuracy performance comparison for BDA input and domain input

The reported performance results for classification models are a positive indication for the proof of concept. The best model provides an accuracy considerably better than 80% for the cases with two and three classes. As the classes were defined by the business user, the sales growth forecasts represent a practical benefit for the company. Furthermore, it has been shown that this result is not random such that the selected big data input can be considered as useful for the forecast.

5.3.6.3 Model building and assessment: regression

Regression models do not require definition of classes as they predict annual sales growth as a singular numeric value on a monthly basis. The elastic net model was built as a multiple linear regression where the feature time series represent explanatory variables and the target time series equals the dependent variable. In addition, the target time series was also considered as

explanatory variables as it was used as model input for ARIMA as well. ARIMA does not use any further variables as input and therefore no further data conditioning was required. Z-score normalization was applied to the explanatory variables of the elastic net model, which is required as variables occur in different units (Hastie, Qian 2014). In accordance with classification, given values of feature time series were directly taken as explanatory variables without application of a time window approach and the set of variables was also restricted to monthly time series. Missing values were treated by imputing the mean of respective feature time series.

The use of a regularized regression is mainly motivated by the idea to build a model with a reduced number of feature time series that increases the possibility to interpret results. In order to account for possible issues stemming from multicollinearity, reduced elastic net as alternative model design included a 2-step filter method to further reduce the number of feature time series:

- 1) *Chi-squared test of independence*: This statistical test checks "[...] statistical independence of two discretely distributed [...] variables" (Pestman 2009, p. 188). Statistical significance tests such as chi-square test of independence are a means of deciding whether a feature should be added to the model (Domingos 2012, p. 82), and therefore it was utilized as additional filter evaluating relevance of feature time series. Each feature time series was paired with the target time series and the test evaluates whether a potential relationship between these two variables is significant or just by chance. Furthermore, the target time series was shifted by 12 months adopting the idea to find features with leading indicator characteristics. In order to transform each time series into a discretely distributed variable, min-max normalization was applied. Each time series observation therefore lied in the range $[0,1]$ and therefore represented ordinaly scaled categories. A contingency table with categories from both variables was built and the chi-square test statistic was calculated based on relative frequencies drawn from the table. In case the test statistic exceeds a certain level of significance, the common level of 0.05 was used here, the hypothesis that both variables are statistically independent needs to be withdrawn (Pestman 2009, p. 191). The result of the test is not a typical ranking of feature time series but all independent features were removed. The first step can be seen as advancement of relevance evaluation of time series prioritization.
- 2) *Covariance test*: Time series prioritization addresses the issue of redundant information during detailed assessment leveraging domain knowledge. The second step of the additional filter method further removes feature time series that show strong correlation among each other. It should be noted that correlating features do not necessarily represent redundancy (Guyon, Elisseeff 2006, p. 12), however, they still cause multicollinearity in the model input. Covariance is a measure for the monotone relation between variables. A large positive covariance implies that large (small) values of one variable correspond with large (small) values of the other, and the maximum value is defined by the product of the standard deviations of both variables which is equivalent to perfect correlation (Mittag 2016, pp. 126–128). Covariance was calculated among

EVALUATION

feature time series and the acceptable limit was set to 95% of the maximum or lower. In cases of covariance above this limit, the feature time series with the higher times series ID⁵⁶ was removed implementing a simple heuristic.

Figure 98 provides an overview of model input after data conditioning and additional feature selection for the elastic net model. Most interestingly, the number of time series in case of elastic net reduced is lower by 68% for BDA input and 89% for domain input in comparison with classification input based on removal of missing values.

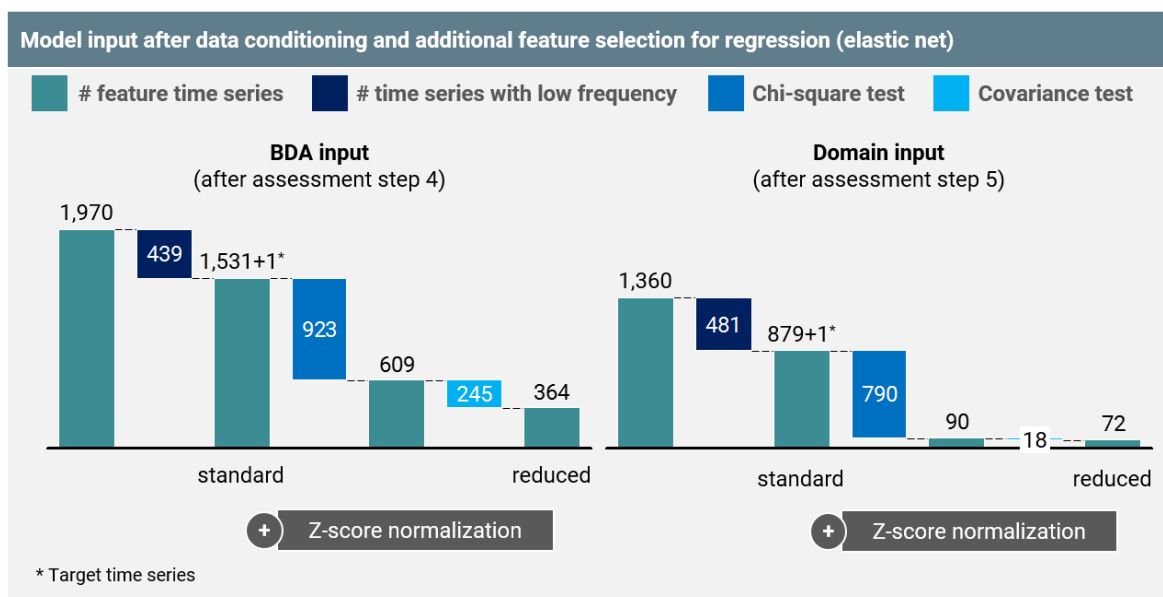


Figure 98 - Data conditioning for regression

Standard elastic net and reduced elastic net were built based on the model library as described before. They only differed in the final model input where the feature time series were significantly reduced, for instance, from 880 including the target time series to 72 in case of domain input. The model parameters were taken at default values except the mixing parameter α which controls the balance between lasso ($\alpha = 1$) and ridge ($\alpha = 0$) regularization. For parameter optimization, the mixing parameter was stepwise increased in 0.05 increments and the best model is selected after 100 runs with 10-fold cross-validation. A seasonal ARIMA model was used in combination with parameter optimization in order to provide the best possible baseline to compare with elastic net performance. The model can be described by seven parameters that were optimized based on a grid search. A grid search scans the space for parameters not optimized during model learning for the setup that provides the best model performance (scikit-learn 2017b). The grid search for ARIMA model parameters followed the algorithm proposed by Hyndman, Khandakar (2008, pp. 8–12), and the results are shown in Table 41 whereby m represents a given parameter.

⁵⁶ All time series on the project cluster have a unique numerical identifier (ID).

ARIMA (p,d,q)(P,D,Q)m		
Parameter	Non-seasonal part	Seasonal part
order of autoregressive part	p = 1	P = 0
degree of first differencing involved	d = 0	D = 1
order of the moving average part	q = 2	Q = 0
number of periods per season	-	m = 12

Table 41 - ARIMA parameter grid search

Assessment of ARIMA and elastic net models was based on RMSE and MAPE as two standard regression performance measures and the results are shown in Figure 99. The highest time period of the training set is always limited to 12 months prior to the forecast period such that available observations for model training are strongly limited for early periods and they were therefore disregarded for model assessment. The calculation of both error terms was consequently based on prediction values for the five most recent years and were represented by the mean of the 95% confidence level predictions. Elastic net models clearly outperform the ARIMA baseline which is most obvious when comparing against the reduced elastic net with domain input. MAPE is reduced by 60% and RMSE even by 70% which represents a significant increase in model performance. The worst performing elastic net model still reduces error measures by 19% and 40%, respectively. Another important observation is the increase in performance due to use of domain input instead of BDA input. Both error measures are reduced for standard and reduced elastic nets when using domain-selected feature time series. Both error measures are decreased by 31% on average when changing to domain input. The standard model with domain input even performs slightly better than the reduced model in combination with BDA input. This superior performance is consistent with the results from classification and thus underscores the value of domain knowledge for preparing modeling input. When looking at the alternative elastic net models based on the same data input, it can be seen that additional feature reduction pays off. Reduced elastic net models on average decrease MAPE by 25% and RMSE by 21%, respectively. In summary, the BDA approach provides better performance, especially when considering involvement of business users in detailed assessment of time series prioritization. Additional feature selection also provides a positive effect on performance.

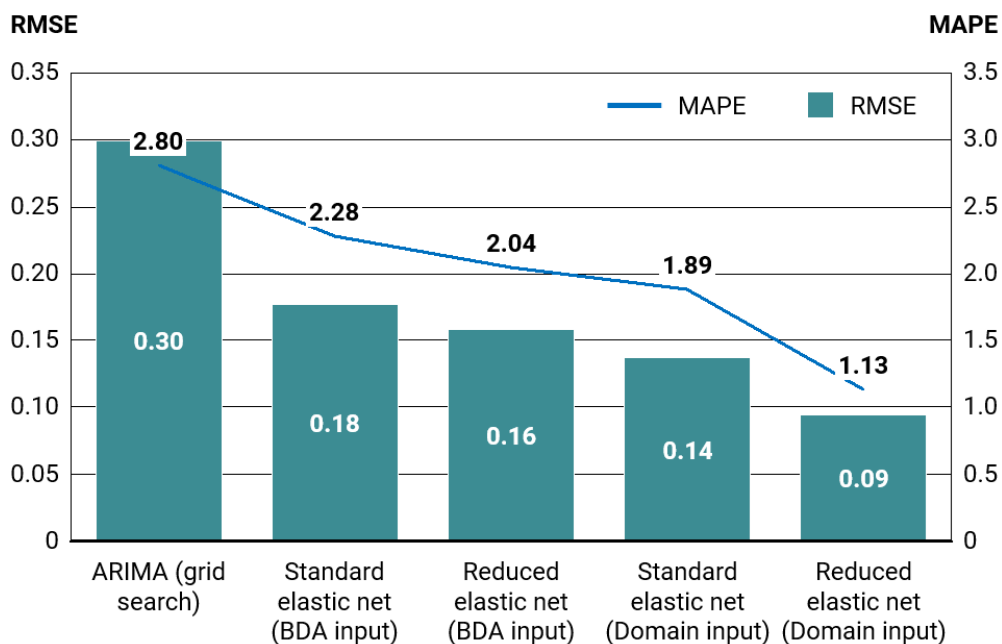


Figure 99 - Overall regression performance

The absolute values of performance measures are still moderate. Forecasting singular growth rates is very demanding with the relatively high volatility of the target time series. The results are therefore rather uncertain when considering to use the models outputs as forecasts for the specific sales growth over the next 12 months. However, the data scientist assessed the performance with $RMSE = 0.09$ as reasonable for the given target time series and utilizing models with default parameter settings. This holds true for understanding the sales growth trends disregarding major shocks. This can be best observed when looking at the comparison of actual and forecast values over time. Figure 100 presents the actual sales growth rate aligned with the forecast value from 12 months earlier over the 5 years period considered for performance assessment. The ARIMA model shows only very weak capability to follow the trend of sales growth while the standard elastic net with BDA input already starts to capture changing dynamics in sales behavior. However, it also becomes clear that this model has large deviations for selected periods which is also reflected by the relatively high error rates measured. This is obviously improved when moving to the reduced elastic net with domain input. Focusing on forecast periods with a minimum of 48 months historic observations available underlines the indication that the best regression model can serve as basis to forecast demand trends.



Figure 100 - Actual versus forecast values (overview)⁵⁷

It is important to note why the presented results for regression should be regarded as a positive indication in the sense of a proof of concept. Each forecast by elastic net models is based on an individual α -value and set of explanatory variables which principally poses a restriction to generalization. The mixing parameter for reduced elastic net with domain input takes on three different values ($\alpha = [0.25; 0.50; 0.75]$), however, model learning with lasso ($\alpha = 1$) and ridge ($\alpha = 0$) setups revealed only small differences in performance. The embedded feature selection functionality in form of regularization is also limited as the final model utilizes 60 feature time series out of the model input of 72 after the covariance test. The influence of the mixing parameter on performance can therefore be assumed to be controllable when building a fully generalized model. The use of individual variable sets also needs to be put into perspective. Out of the 60 overall used variables only 21 are used for any forecast on average and 18 individual variables are used in more than half the forecasts. Ten of the explanatory variables occur even in more than three quarters of the cases. This illustrates the moderate variance in the set of explanatory variables across all forecast periods. In summary, it can be expected that model optimization paired with further advancements in feature selection potentially can result in reasonable performance that is generally valid.

⁵⁷ Dimensions are omitted for sales growth values (vertical) and time periods (horizontal) for confidentiality reasons.

5.3.6.4 Business evaluation

Model assessment identified working models for classification and regression. In the light of utilizing library models with default settings, SVM(rbf) and kNN for classification as well as reduced elastic net provided reasonable performance. As a consequence, these results generally represent a positive indication for the proof of concept of utilizing big data analytics for sales forecasting, and therefore applicability in practice was discussed with the business user and project sponsor. Comparing model performance with already available solutions is a good way to evaluate practicability, however, no reasonable baseline was identified for medium-term sales forecasts. On the one hand, this means any reasonable forecast is useful for the company as it closes an existing gap. On the other hand, the missing point of reference is a burden for a discussion on model assessments based on performance measures not directly interpretable. While accuracy of a classification proved to be sufficient for a business user to decide whether model performance is useful, regression error measures were helpful only to decide which version of elastic net was best but not to determine whether performance is reasonable at all. The use of a standard forecast model (ARIMA) as auxiliary baseline was found to be helpful here, especially in combination with visualization of results, and was therefore used as basis for the business evaluation review. A practice test was not considered for two reasons. Firstly, the focus on the medium term requires a long period of observations to make a valid evaluation. Secondly, the use case was designed to provide a proof of concept and therefore did not result in a working prototype within the company infrastructure. Costly tasks such as data anonymization and transfer would still be required for a test.

Explicability of the models strongly differs between classification and regression. SVM(rbf) and kNN as presented operate with several hundred feature time series which gives them a black box character, where "[...] the relationship between joint values of the input variables and the resulting predicted response value [...]" (Hastie et al. 2017, pp. 351–352) is not accessible for interpretation. On the other hand, reduced elastic net is based on simple linear regression and uses 20 explanatory variables on average. The model provides coefficients that describe the influence of each variable on the forecast value in an intuitive way. A range of 20 variables facilitates the addition of a monitoring dashboard or reporting system as well. The picture on confidence is more balanced. Both forecast categories need to cope with the limited training set. Although there does not exist "[...] a general rule on how much training data is enough [...]" (Hastie et al. 2017, p. 222), the total number of 84 available observations can be seen as critical.

The issue has been addressed by 10-fold cross-validation but requires further investigation in order to ensure generalization fully reliable for business application. However, it also needs to be noted that recent research on advanced sales forecasting in the PCB industry utilizes only four years of monthly sales data as training set and measures performance on a deliberately selected hold-out set instead of cross-validation (compare *Subsection 3.1.3*). Classification models are especially prone to overfitting here, because their number of explanatory variables represent

a high level of model complexity (Hastie et al. 2017, p. 38). The random shuffling test still has shown that feature time series convey significant information about sales growth classes and measured performance is noticeably strong. Modeling with additional feature reduction is therefore an interesting advancement option. It should be noted that the issue of overfitting can be expected to be more severe for advanced models such as ANN as they typically represent an even higher level of model complexity.

The overall vote from the business evaluation review is positive and therefore confirms the indicative proof of concept that was already revealed based on model assessments. In consideration of the discussions on explicability and confidence, three major steps were identified in order to move from proof of concept towards potential deployment:

- 1) *Advanced feature selection*: The aim for classification models is to reduce the number of feature time series to a range that is manageable for monitoring via a dashboard or reporting in order to ensure explicability. At the same time, it also addresses the issue of potential overfitting. Feature selection in case of regression needs to be optimized to provide a stable set of explanatory variables.
- 2) *Increase of test set*: The frequency of the target time series can be raised, for example, from monthly to weekly, in order to increase the number of available observations for testing. As most feature time series have a monthly frequency at maximum, this poses a further restriction to data input or requires transformation of these time series. Another option is to artificially expand the training set. For instance, various oversampling strategies exist that additionally address imbalanced datasets (Chawla 2010, pp. 879–881), which was identified as additional challenge for classification. The increase of the test set addresses generalization performance of both model categories likewise and thus potentially promotes confidence. Transfer of the methodology and models to other business units of the company was also identified as increase of the test set in a broader sense. Although this does not directly affect generalization performance of the models, replication of the performance in a similar but different setting would strengthen confidence from the business user perspective.
- 3) *Model optimization*: Modeling can be improved along three dimensions. Firstly, data conditioning offers the option of a time window approach which was not utilized in the case study. Inherent dynamics might have even more explanatory power compared to the plain time series. A time window approach would also allow to incorporate data with lower frequencies. Secondly, model parameters can be tuned for better performance, for example, with the help of a grid search approach as used for the ARIMA model. Thirdly, an ensemble method combining different models can be used to improve performance. For example, SVM(rbf) and kNN are potential candidates to be combined in case their prediction errors are independent from each other. All optimization efforts

EVALUATION

generally apply to classification as well as regression and aim for higher confidence due to better model performance.

6 Conclusion

6.1 Discussion

6.1.1 Research summary

The situation of companies today (compare *Subsection 1.1.1*) is characterized by a volatile business environment. In order to adapt to this volatile world, anticipation of changes is required. The age of analytics (compare *Subsection 1.1.2*) offers big data, cheap data processing and storage, and advances in analytics as opportunity to provide a better understanding for the diverse volatilities in the business environment. In particular, sales forecasting is considered as priority BDA application and companies show readiness to invest into their development (compare *Subsection 1.1.3*). The consideration of big data analytics as opportunity, especially in form of sales forecasting, is confirmed by the pre-study (compare *Section 5.1*) and case study (compare *Subsection 5.3.2*) with four companies from three different industries. Consequently, the primary objective of this research is motivated by the business need to identify and select BDA applications that provide a better understanding of the volatile business environment as well as to specifically develop an application for sales forecasting (compare *Subsection 1.2.1*). An extensive review on processes for analytics (compare *Section 3.2*) provides the knowledge base for a new methodology as answer to this business need. CRISP-DM is identified as the most qualified process among all candidates and provides the basic design of this methodology. However, the review also reveals the scientific need to substantiate the process towards the primary research objective. This scientific need relates to the first two research question of this work. The review also includes process evaluations, comparisons and success factors that identify further scientific need for process improvements. These findings are categorized by six different improvement areas where each is described by detailed dimensions, and they are considered as design requirements for the new methodology. Moreover, the review of related research on sales forecasting (compare *Section 3.1*) reveals the scientific need to provide a proof of concept whether BDA sales forecasting works in practice, especially in B2B industries. This underlines the importance of the third research question. The new methodology is built and evaluated in a case study with a PCB manufacturer. The presented case study results (compare *Section 5.3*) provide the basis for the discussion of the methodology evaluation. As the evaluation of the methodology results in a BDA application for sales forecasting, the results furthermore allow to discuss the proof of concept as well.

6.1.2 Methodology evaluation

The case study results describe the methodology evaluation in detail and therefore enable a final conclusion. The design of the initial business understanding step of the new methodology is based on the agility concept. In particular, the advanced corporate agility system connects the challenges of the volatile business environment with the opportunities of big data analytics. This approach facilitated a common understanding in the project team and set strategic guidelines for a focused discussion on potential use cases (compare *Subsection 5.3.2.1*). The subsequent

tasks identified eleven use cases for BDA applications regarding the volatile world and prioritized the two most relevant use cases for each agility-based business objective (compare *Subsection 5.3.2.2*). The general part of the new methodology therefore successfully supports the decision where to use BDA applications for a better understanding of the volatile business environment and consequently provides an evaluated answer to research question 1. The second step of the general part, big data sources, extends the lifecycle of existing processes by a dedicated step to identify and select sources of big data input. The tasks and methods are not specific to a use case or certain types of data but also aim for an extended information base facilitating a better understanding of the volatile world. Based on the new methodology, a long list of 28 data sources was identified and successively reduced to a data mix of eight prioritized data sources covering all data types (compare *Subsection 5.3.3*).

The big data sources step provides the basis for the specific part of the methodology that aims to develop a BDA application for sales forecasting. In the data understanding step, a total of 191 datasets from considered data sources were selected, sourced, explored, and verified (compare *Subsection 5.3.4*). Time-series generation resulted in more than 4 million time series of which 1,360 were prioritized as modeling input based on quality and relevance criteria (compare *Subsection 5.3.5*). The final modeling & evaluation step considered five different models for sales forecasting. These models were built and their performance assessed. The business evaluation regarded SVM(rbf), kNN, and reduced elastic net as working models for sales forecasting based on the previously defined big data input (compare *Subsection 5.3.6*). As a consequence, the specific part of the new methodology provides an evaluated answer to research question 2 on how to develop a BDA application for sales forecasting.

The new methodology furthermore incorporates the six identified design requirements (compare *Subsection 3.2.5*) that are discussed in the following:

(I) Project team

The team setup underlying the methodology is generally independent from the business need of this work. It describes the full spectrum of roles required to implement the methodology. In combination with the integrated workflow model that consistently assigns responsibilities for each task throughout the entire lifecycle, the team roles proved to be effective in the case study as they ensured operationality (compare *Subsection 5.3.2 to 5.3.6*). The key design element of the new methodology with regard to the project team is the newly introduced BDA manager role. The new role reflects the need for a team leader with business and BDA background. In the case study, the BDA manager role proved to be effective for project management and coordination of the multidisciplinary team across the lifecycle of the analytics process (compare *Subsection 5.3.2 to 5.3.6*). Furthermore, the role enabled content-related instead of process-related communication with stakeholders which led to support by key functions such as IT. The introduction of the other new roles, data owner and data officer, allowed for implementation of new tasks such as providing data-specific domain knowledge (compare *Subsection 5.3.4.2*) or legal

CONCLUSION

clearance for data privacy reasons (compare *Subsection 5.3.4.3*), respectively. The lack of internal BDA capabilities was covered by external roles. Furthermore, the use of auxiliary project managers at the company and the external provider were beneficial to an efficient methodology implementation (compare *Subsection 5.3.1*). The new methodology considers all defined dimensions for an effective project organization except for elements from agile approaches, for example, an agile working mode.

(II) Domain knowledge

The role of the business user combines domain expertise and user perspective. In the new methodology, the business user is involved in many tasks across all steps and related methods are specifically designed to facilitate this involvement. An exemplary list of business user performed tasks provides evidence of successful domain knowledge integration during the case study (compare *Subsection 5.3.2 to 5.3.6*): verifying agility-based business objectives, preparing the use case assessment template, identifying data sources during the data query, providing metadata for the BDA book, formulating hypotheses for time series generation, prioritizing time series in the decision workshop, and evaluating model performance from a business perspective. The case study results furthermore show that integration of domain knowledge into the prioritization of data input has a positive effect on model performance (compare *Subsection 5.3.6.2 & 5.3.6.3*). The design requirement to involve domain expertise and user perspective throughout the lifecycle is fully met by the new methodology.

(III) Business understanding

The advanced corporate agility system represents a key substantiation of the CRISP-DM design towards BDA utilization in volatile times. At the same time, it also enables a thorough implementation of the business understanding step. The case study results show that this approach facilitated a common understanding for the BDA opportunity in the volatile world and the alignment of business and analytics objectives based on the company's mission statement (compare *Subsection 5.3.2.1*). Moreover, the use case approach determined possible BDA applications with clear scope and business rationale under consideration of project limitations such as the focus on a proof of concept (compare *Subsection 5.3.2.1*). The new methodology furthermore considers existent BDA solutions with regard to the identified business objectives, however, ideas generated from competitor benchmarking were limited in the case study. While the methodology design considers all dimensions for an improved business understanding, this represents the only design element without positive evaluation for applicability.

(IV) Data input

The new methodology dedicates a whole new step to the design requirement for an increased focus on data input. In the case study, the application of tasks and methods along the data sources funnel resulted in a deliberate selection of data sources covering all dimensions of the data mix matrix (compare *Subsection 5.3.3*). While this ensured the variety and veracity

dimensions of big data, the considered update frequency also ensured data at relatively high velocity in the form of monthly updates for annual sales growth rates. Data was subsequently sourced from three different types of data sources and integrated into a Hadoop-based project cluster that enabled processing of more than 320 gigabyte of data (compare *Subsection 5.3.4*), exceeding the typical data volume for sales forecasting of a business segment in the B2B industry. Handling complexity of data is mainly addressed by the BDA book and the case study proved it to be an effective tool for documentation of metadata as well as for collaboration between different team roles (compare *Subsection 5.3.4.2*). Furthermore, the BDA book enables data exploration and verification such that roughly 10% of the data could be excluded for quality reasons during data understanding (compare *Subsection 5.3.4.3*). However, data quality is mostly considered by time series prioritization and its methods, such as filter setups in the evaluation reports. Prioritization in the case study resulted in an effective removal of low quality time series (compare *Subsection 5.3.5.2*). As a result, the methodology effectively addresses all dimensions of the data input design requirement.

(V) Methods

The new methodology comprises 26 tools and techniques in order to describe 'how to' conduct each task. As the case study results across all steps show, these methods led to an effective implementation of all 17 tasks based on the detailed explanations and decision guidelines provided by the methodology (compare *Subsection 5.3.2 to 5.3.6*). The results also prove enablement of regular collaboration, for example for BDA manager, business user, data owner and database administrator via the BDA book during sourcing preparation and data sourcing (compare *Subsection 5.3.4.2 & 5.3.4.3*). Furthermore, also existing BDA tools were considered, especially in form of Apache Spark and the scikit-learn library (compare *Subsection 5.3.5 & 5.3.6*). The only dimension not addressed by the new methodology is visualization tools.

(VI) Automation

The consideration of automation is mainly reflected by the methodology design as highly integrated process. As the case study results show, there is a clear output-input relation for subsequent tasks and only limited iterations exist (compare *Subsection 5.3.2 to 5.3.6*) which resulted in a high level of usability. However, technical implementation of automation was restricted to individual subtasks, for example, those covered by the tools for time series generation and prioritization (compare *Subsection 5.3.5.1 & 5.3.5.2*). In these tasks, automated time series generation and consideration of data quality during general assessment represent major implementations automation.

In summary, the new methodology provides two major additions to the knowledge base. Firstly, it represents a substantiated form of CRISP-DM addressing the business need regarding identification and selection of BDA applications for the volatile world as well as development of an application for sales forecasting. Secondly, the design of the new methodology includes various advancements regarding improvement areas of KDDM and BDA processes. Although

CONCLUSION

the case study was performed with an industrial company of the PCB industry, no industry-specific assumptions were made that restrict the validity of the methodology in other industries. Furthermore, the improvements are generally not restricted to a methodology for the given business need. They can be considered for other needs as well, for example, methods of time series generation and prioritization are potentially valuable for any predictive analytics application based on time series data. The BDA book also represents a useful tool for general data understanding tasks.

6.1.3 Proof of concept

The case study ultimately results in BDA applications for sales forecasting. Implementing the new methodology, the subtasks of model assessment (compare *Subsection 5.3.6.2* & *5.3.6.3*) and business evaluation (compare *Subsection 5.3.6.4*) provide the basis for a proof of concept. Both approaches, classification and regression, show reasonable performance regarding medium-term forecasts of sales growth. SVM(rbf) represents the best performing classifier and the positive assessment rests on its absolute performance measured by an accuracy of 85% in case of three classes for sales growth. For the elastic net regression model, significant performance improvement compared to the ARIMA baseline model and reasonable reproduction of sales growth trends are the basis for the positive assessment. However, two major restrictions must be considered here. The number of time series in the model input is large, especially for classification, and the number of observations is limited due to the maximum history for sales data of seven years. This can be seen as critical regarding the generalization of model performance. On the other hand, various improvement potentials⁵⁸ for model performance were identified for both BDA applications. As a consequence, the case study results are considered as positive indication for the proof of concept that a BDA approach for sales forecasting works in industrial practice. This represents the answer to research question 3 but is limited to B2B industries. The selected data and the characteristics of sales are potentially specific for this type of industry and therefore general transferability of results should not simply be assumed. Furthermore, the BDA applications including their specifications represent another addition to the knowledge base that provides a higher level of detail compared to reviewed applications (compare *Section 3.1*).

⁵⁸ For details, see discussion on refinements in *Subsection 6.2.2*.

6.2 Future work

Further refinements of the research are another important result beyond the additions to the knowledge base. The following presents directions to future work on the methodology and BDA applications for sales forecasting based on the previous discussion on methodology evaluation and proof of concept. Further observations from the research work on methodology design elements are also included.

6.2.1 New methodology refinements

The research motivation describes a business need that generally applies to different industries. It is therefore an interesting research target to further validate the new methodology in other industries. Automotive and semiconductor industries, where the business need was explicitly validated by the pre-study, represent appropriate candidates for initial research objects. Moreover, BDA applications for sales forecasting represent a specific part of the overall business need. The volatile business environment comprises various challenges and the case study revealed other use cases, for example, identification of new technologies. Determining methodologies to develop BDA applications for alternative use cases and integration with the general part of the new methodology would be an advancement towards broader applicability. Furthermore, the business understanding step is designed to capture the strategic value of big data analytics in the volatile world. The case study has proven the validity of the agility concept for this purpose but approaches to capture the operational value also exist, for example, profitability-based concepts in process industries (Hammer et al. 2017). Further research is required to determine the dependencies at the interface between strategic and operational approaches (Heldmann et al. 2017, p. 84).

Data input for the specific part of the methodology is restricted to structured data, however, unstructured data is considered as valuable part of the information base and the big data sources step determines a data mix including both data types. An extension of the specific methodology part in order to include analytics based on unstructured data therefore represents a further refinement. For example, text mining as form of analytics on unstructured data strongly depends on preprocessing tasks such as natural language processing (Feldman, Sanger 2006, pp. 57–63). The data understanding step establishes a project cluster that integrates all data required by the selected use case in a single data warehouse. As discussed earlier, various use cases based on different data inputs potentially exist for a company. Data lakes represent an interesting approach to store and manage data from different data sources, including data warehouses, in order to utilize big data for multiple applications (Schmarzo 2016, pp. 133–151). Understanding the implications of implementing such an approach and necessary adaptations of the methodology represents an interesting research topic. In data preparation, values for filter thresholds and scoring weights are set in order to enable prioritization of datasets. While these values are carefully selected and the case study confirms the practicality of this design element, it would

be interesting to investigate whether optimal scoring and filter setups exist. Further research could assess the difference in model performance driven by changes in the data input due to alternative setups, for example. The new methodology concludes with the modeling & evaluation step and therefore does not include deployment of working BDA applications. The review in *Section 3.2* includes a large number of processes that include deployment. On this basis, it could be evaluated how the methodology can be integrated with existing deployment approaches. Furthermore, the framework of the corporate agility system could be integrated here. Integration of deployable BDA applications with the other key building blocks, control and agility levers, represents an interesting research direction for companies seeking to implement a corporate agility system. The new methodology builds upon the team setup that allows for external roles, especially covering for BDA capabilities. While the methodology defines the roles and directs towards the use of specialized providers, it remains open how to select external partners. Due to the shortage of analytics talent it is difficult for companies to staff teams only internally (Fogarty, Bell 2014), and therefore methods to evaluate potential partners represent a valuable addition for the methodology.

Another group of refinements stem from design requirements that are not fully met by the new methodology. Automation is mainly implemented for selected parts, especially for automated time series generation and general assessment during time series prioritization. There exist good reasons not to automate all tasks, for example, detailed assessment is specifically designed for integration of domain knowledge and therefore cannot be fully automated. However, it is a remaining research need to assess the potential for further automation of the methodology. Furthermore, the methodology incorporates the idea of experimentation principally through integration of modeling and evaluation as well as the task of data conditioning which both facilitate iterations. In addition, the methodology builds and assesses alternative models based on the same data input. Iterations between data preparation and modeling is considered a useful way to improve analytics (Domingos 2012, p. 84), and represents a starting point for refinements towards a higher level of experimentation that includes data input as well. Furthermore, methods for agile analytics work (Alnoukari 2012) represent a potential approach to strengthen the role of experimentation. Methods to support the identification of existing BDA solutions during business understanding are a specific refinement need based on observed case study results. Finally, it needs to be assessed how visualization tools can help to improve data understanding, data processing or model assessment. Despite the given refinement need, the new methodology also successfully addresses a long list of design requirements. The transfer to other existing or new processes is therefore also a need for future research.

6.2.2 Sales forecasting application refinements

The model assessment and business evaluation of the classification and regression models for sales forecasting reveal improvement potentials regarding their performance. The following lists refinement options for both BDA applications:

- *Oversampling strategies* extend the limited number of observations for model building.
- *Advanced feature selection* further reduces the number of time series considered as model input (classification) or to provide a consistent set of time series as model input for all individual forecasts (regression).
- *Time window approach* offers the opportunity to capture dynamics in model input time series.
- *Low frequency time series* (e.g., quarterly) are not considered as model input but could be utilized based on a time window approach as well.
- *Parameter optimization* potentially improves performance of the models with default parameter settings, for example, based on a grid search.

The model selection furthermore identified advanced models that are to be assessed for performance with the provided modeling input. For example, ensemble models are reported to potentially improve accuracy by 5-30% (Finlay 2014, p. 130) and the operational capability of artificial neural networks for sales forecasting is confirmed for trading companies (Crone 2010). These advanced models provide the potential to further increase the confidence in the proof of concept

Bibliography

Abele, Eberhard; Reinhart, Gunther (2011): *Zukunft der Produktion. Herausforderungen, Forschungsfelder, Chancen*. München: Carl Hanser Verlag.

Abraham, Bovas; Ledolter, Johannes (2005): *Statistical methods for forecasting*. Hoboken, NJ: Wiley (Wiley series in probability and statistics).

Agile Alliance (2017): *Manifesto for Agile Software Development*. Available online at <https://www.agilealliance.org/agile101/the-agile-manifesto/>, checked on 7/8/2017.

Ahangama, Supunmali; Poo, Chiang-Choon Danny (2015a): *Designing a Process Model for Health Analytic Projects*. In : PACIS 2015 Proceedings. 19th Pacific Asia Conference on Information Systems (PACIS 2015). Singapore, July 5-9, 2015: AIS Electronic Library, pp. 1–13.

Ahangama, Supunmali; Poo, Chiang-Choon Danny (2015b): *Improving health analytic process through project, communication and knowledge management*. 2015 International Conference on Information Systems (ICIS 2015). Available online at <http://aisel.aisnet.org/icis2015/proceedings/DecisionAnalytics/4/>, checked on 5/22/2017.

Ahlemeyer-Stubbe, Andrea; Coleman, Shirley (2014): *A practical guide to data mining for business and industry*. Chichester: John Wiley & Sons.

Almquist, Eric; Senior, John; Springer, Tom (2015): *Three promises and perils of Big Data*. Bain & Company.

Alnoukari, Mouhib (2012): *ASD-BI: A Knowledge Discovery Process Modeling Based on Adaptive Software Development Agile Methodology*. In Asim Abdel Rahman El Sheikh, Mouhib Alnoukari (Eds.): *Business intelligence and agile methodologies for knowledge-based organizations. Cross-disciplinary applications*. Hershey, PA: Business Science Reference, pp. 183–207.

Alnoukari, Mouhib; Alzoabi, Zaidoun; El Sheikh, Asim (2009): *Applying ASD-DM methodology on Business Intelligence solutions: A case study on building customer Care Data Mart*. In Ajith P. Abraham (Ed.): *Proceedings of Data Mining 2009. IADIS European Conference on Data Mining (MCCSIS 2009)*. Algarve, Portugal, 18-20 June 2009, pp. 153–157.

Alonso, Fernando; Martinez, Loic; Perez, Aurora; Valente, Juan P. (2012): *Cooperation between expert knowledge and data mining discovered knowledge. Lessons learned*. In *Expert Systems with Applications* 39 (8), pp. 7524–7535. DOI: 10.1016/j.eswa.2012.01.133.

Alpaydin, Ethem (2010): *Introduction to machine learning*. 2nd ed. Cambridge, MA: MIT Press (Adaptive computation and machine learning).

Ameri, P. (2016): *Database Techniques for Big Data*. In Rajkumar Buyya, Rodrigo N. Calheiros, Amir Vahid Dastjerdi (Eds.): *Big data. Principles and paradigms*. Cambridge, MA: Morgan Kaufmann - Elsevier (Science Direct e-books), pp. 139–159.

-
- Anand, Aneeth (2012): Expanding Data Mining Theory for Industrial Applications. Master Thesis. Arizona State University. ASU Graduate College. Available online at <https://repository.asu.edu/items/14560>, checked on 7/8/2017.
- Anand, Sarabjot S.; Büchner, Alex G. (1998): Decision support using data mining. London: Financial Times Management (Management Briefings, Information Technology).
- Anand, Sarabjot S.; Patrick, A. R.; Hughes, John G.; Bell, David A. (1998): A Data Mining methodology for cross-sales. In *Knowledge-Based Systems* 10 (7), pp. 449–461. DOI: 10.1016/S0950-7051(98)00035-5.
- Anand, Sarabjot Singh; Grobelnik, Marko; Herrmann, Frank; Hornick, Mark; Lingenfelder, Christoph; Rooney, Niall; Wettschereck, Dietrich (2007): Knowledge discovery standards. In *Artificial Intelligence Review* 27 (1), pp. 21–56. DOI: 10.1007/s10462-008-9067-4.
- Anda, Cuauhtemoc; Erath, Alexander; Fourie, Pieter Jacobus (2017): Transport modelling in the age of big data. In *International Journal of Urban Sciences* 21 (sup1), pp. 19–42. DOI: 10.1080/12265934.2017.1281150.
- Anderson, Chris (2008): The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Edited by WIRED. Available online at <https://www.wired.com/2008/06/pb-theory/>, checked on 6/19/2017.
- Ankam, Venkat (2016): Big data analytics. A handy reference guide for data analysts and data scientists to help obtain value from big data analytics using Spark on Hadoop clusters. Birmingham: Packt Publishing.
- Apache Software Foundation (2014): Apache Parquet. Available online at <http://parquet.apache.org/>, checked on 5/12/2017.
- Apache Software Foundation (2017): Apache Spark. Available online at <https://spark.apache.org/>, checked on 4/10/2017.
- Apte, C. V.; Hong, S. J.; Natarajan, R.; Pednault, E. P. D.; Tipu, F. A.; Weiss, S. M. (2003): Data-intensive analytics for predictive modeling. In *IBM Journal of Research and Development* 47 (1), pp. 17–23. DOI: 10.1147/rd.471.0017.
- Archer, Norm P.; Ghasemzadeh, Fereidoun (1996): Project portfolio selection techniques. A review and a suggested integrated approach. Innovation Research Working Group (Working Paper No. 46). Available online at: <https://macsphere.mcmaster.ca/bitstream/11375/5415/1/fulltext.pdf>, checked on 5/4/2017.
- Archer, Norm P.; Ghasemzadeh, Fereidoun (1999): An integrated framework for project portfolio selection. In *International Journal of Project Management* 17 (4), pp. 207–216. DOI: 10.1016/S0263-7863(98)00032-5.
-

Ariker, Matt; Breuer, Peter; McGuire, Tim (2014): How to get the most from big data. McKinsey & Company. Available online at <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/how-to-get-the-most-from-big-data>, checked on 7/9/2017.

Armstrong, J. Scott (2002a): Evaluating forecasting methods. In Jon Scott Armstrong (Ed.): Principles of forecasting. A handbook for researchers and practitioners. 2. ed. Boston, MA: Kluwer Academic (International series in operations research & management science), 443-472.

Armstrong, J. Scott; Green, Kesten C.; Graefe, Andreas (2015): Golden rule of forecasting. Be conservative. In *Journal of Business Research* 68 (8), pp. 1717–1731. DOI: 10.1016/j.jbusres.2015.03.031.

Armstrong, Jon Scott (Ed.) (2002b): Principles of forecasting. A handbook for researchers and practitioners. 2. ed. Boston, MA: Kluwer Academic (International series in operations research & management science).

Asamoah, Daniel Adomako; Sharda, Ramesh (2015): Adapting CRISP-DM Process for Social Network Analytics. In : 21st Americas Conference on Information Systems (AMCIS 2015). 21st Americas Conference on Information Systems (AMCIS 2015). Fajardo, Puerto Rico, 13-15 August 2015. Atlanta, Georgia: Association for Information Systems, Code 118635.

Aschenbrücker, Andreas; Horváth, Péter; Michel, Uwe (2014): Controlling im volatilen Umfeld. In *Controller Magazin* (Januar/Februar), pp. 4–11.

Azevedo, Ana; Santos, Manuel Filipe (2008): KDD, semma and CRISP-DM: A parallel overview. In Ajith Abraham (Ed.): Proceedings IADIS 2008. IADIS European Conference on Data Mining 2008. Amsterdam, The Netherlands, July 24-26, 2008. IADIS 2008, pp. 182–185.

Baars, Henning; Kemper, Hans-Georg (2015): Integration von Big Data-Komponenten in die Business Intelligence. In *Controlling - Zeitschrift für erfolgsorientierte Unternehmenssteuerung* 27 (4/5), pp. 222–228.

Baechle, Christopher; Agarwal, Ankur; Zhu, Xingquan (2017): Big data driven co-occurring evidence discovery in chronic obstructive pulmonary disease patients. In *Journal of Big Data* 4 (1), Article 9 (18 pages). DOI: 10.1186/s40537-017-0067-6.

Baek And, Junghan; Sohn, Keemin (2016): Deep-Learning Architectures to Forecast Bus Ridership at the Stop and Stop-To-Stop Levels for Dense and Crowded Bus Networks. In *Applied Artificial Intelligence* 30 (9), pp. 861–885. DOI: 10.1080/08839514.2016.1277291.

Baesens, Bart (2014): Analytics in a big data world. The essential guide to data science and its applications. Hoboken, NJ: John Wiley & Sons (Wiley & SAS business series).

Bange, Carsten; Grosser, Timm; Janoschek, Nikolai (2015): Big Data Use Cases. Getting real on data monetization. BARC Research Study. BARC. Available online at https://www.sas.com/content/dam/SAS/bp_de/doc/studie/ba-st-barc-bigdata-use-cases-de-2359583.pdf, checked on 5/27/2017.

- Banica, L.; Hagi, A. (2016): Using big data analytics to improve decision-making in apparel supply chains. In Tsan-Ming Choi (Ed.): *Information systems for the fashion and apparel industry*. Duxford: Woodhead (Woodhead publishing series in textiles, Number 179), pp. 63–95.
- Barclay, Corlane (2015): Critical Success Factors in Knowledge Discovery and Data Mining Projects. In Corlane Barclay, Kweku-Muata Osei-Bryson (Eds.): *Knowledge discovery process and methods to enhance organizational performance*. Boca Raton, FL: CRC Press, pp. 165–185.
- Barclay, Corlane; Osei-Bryson, Kweku-Muata (2015): Introduction. In Corlane Barclay, Kweku-Muata Osei-Bryson (Eds.): *Knowledge discovery process and methods to enhance organizational performance*. Boca Raton, FL: CRC Press, pp. 1–7.
- Barton, Dominic; Court, David (2012): Making advanced analytics work for you. A practical guide to capitalizing on big data. In *Harvard Business Review* 90 (10 (October)), pp. 78–83.
- Basili, Victor R.; Weiss, David M. (1984): A Methodology for Collecting Valid Software Engineering Data. In *IEEE Transactions on Software Engineering* SE-10 (6), pp. 728–738. DOI: 10.1109/TSE.1984.5010301.
- Baughman, Aaron K.; Bogdany, Richard; Harrison, Benjie; O’Connell, Brian; Pearthree, Herbie; Frankel, Brandon et al. (2016): IBM Predicts Cloud Computing Demand for Sports Tournaments. In *Interfaces* 46 (1), pp. 33–48. DOI: 10.1287/inte.2015.0820.
- Baumgartner, Rupert J.; Biedermann, Hubert; Klügl, Franz; Schneeberger, Thomas; Strohmeier, Georg; Zielowski, Christian (2006): *Generic Management. Unternehmensführung in einem komplexen und dynamischen Umfeld*. 1. Auflage. Wiesbaden: Deutscher Universitäts-Verlag.
- Bello-Orgaz, Gema; Jung, Jason J.; Camacho, David (2016): Social big data. Recent achievements and new challenges. In *Information Fusion* 28, pp. 45–59. DOI: 10.1016/j.inffus.2015.08.005.
- Belmokaddem, Mostefa; Benatek, Omar; Benameur, Abdelkrim; Mellal, Rabiaa (2014): Methods of sales forecasting and modeling of supply chains. Case study: Atlas Chimie Algeria. In *Romanian Journal of Economics* 39 (2 (48)), pp. 19–33.
- Berman, Jules J. (2013): *Principles of big data. Preparing, sharing, and analyzing complex information*. Amsterdam: Elsevier.
- Berry, Michael J. A.; Linoff, Gordon S. (2004): *Data mining techniques. For marketing, sales, and customer relationship management*. 2nd ed. Indianapolis (Ind.): Wiley.
- Biedermann, Hubert (2010): *Generic Management als umfassendes Konzept zur Sicherstellung der Wandlungsfähigkeit industrieller Produktion*. In Peter Nyhuis (Ed.): *Wandlungsfähige Produktionssysteme*. Berlin: GITO-Verlag, pp. 23–42.
- Biedermann, Hubert (2016a): *Lean Smart Maintenance*. In *WINGbusiness* (1), pp. 12–15.

Biedermann, Hubert (2016b): Lean Smart Maintenance. In Hubert Biedermann (Ed.): Industrial Engineering und Management. Beiträge des Techno-Ökonomie-Forums der TU Austria. Wiesbaden: Springer Gabler, pp. 119–141.

Biesdorf, Stefan; Court, David; Willmott, Paul (2013): Big data: What's your plan? In *McKinsey Quarterly* (March 2013). Available online at <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-whats-your-plan>, checked on 5/11/2015.

Bishop, Christopher M. (2006): Pattern recognition and machine learning. New York, NY: Springer (Information science and statistics).

Blockeel, Hendrik; Moyle, Steve (2002): Collaborative data mining needs entrained model evaluation. In *Proceedings of the ICML-2002 Workshop on Data Mining Lessons Learned*, pp. 21–28. Available online at <https://lirias.kuleuven.be/bitstream/123456789/133117/1/>, checked on 7/3/2017.

Bohanec, Marko; Moyle, Steve; Wettschereck, Dietrich; Miksovsky, Petr (2001): A Software Architecture for Data Pre-Processing using Data Mining and Decision Support Models. In C. Giraud-Carrier, N. Lavrac, Steve Moyle, B. Kavsek (Eds.): ECML/PKDD01 Workshop. ECML/PKDD01 Workshop Integrating Aspects of Data Mining, Decision Support and Meta-Learning (IDDM-2001). 2nd International Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning. Freiburg, Germany, September 2001.

Boldt, Linda Camilla; Vinayagamoorthy, Vinothan; Winder, Florian; Schnittger, Melanie; Ekran, Mats; Mukkamala, Raghava Rao et al. (2016): Forecasting Nike's Sales using Facebook Data. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data. Washington D.C., USA, Dec 05-Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 2447–2456.

Bolon-Canedo, Veronica; Sanchez-Marono, Noelia; Alonso-Betanzos, Amparo (2015): Feature Selection for High-Dimensional Data. Cham: Springer International Publishing (Artificial Intelligence: Foundations, Theory, and Algorithms).

Bose, Indranil; Mahapatra, Radha K. (2001): Business data mining — a machine learning perspective. In *Information & Management* 39 (3), pp. 211–225. DOI: 10.1016/S0378-7206(01)00091-X.

Box, George E. P.; Jenkins, Gwilym M.; Reinsel, Gregory C.; Ljung, Greta M. (2015): Time series analysis. Forecasting and control. Fifth edition. Hoboken, New Jersey: John Wiley & Sons (Wiley series in probability and statistics).

Brachman, Ronald J.; Anand, Tej (1994): The Process of Knowledge Discovery in Databases. A First Sketch. In Usama M. Fayyad, Ramasamy Uthurusamy (Eds.): Knowledge discovery in databases. Papers from the 1994 AAAI Workshop. AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94). Seattle, WA, USA, July 31-August 1, 1994. American

-
- Association for Artificial Intelligence. Menlo Park, CA: AAAI Press (Technical report, WS-94-03), pp. 1–12.
- Brachman, Ronald J.; Anand, Tej (1996): The process of knowledge discovery in databases. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy (Eds.): *Advances in knowledge discovery and data mining*. Menlo Park, CA: AAAI/MIT Press, pp. 37–57.
- Bradlow, Eric T.; Gangwar, Manish; Kopalle, Praveen; Voleti, Sudhir (2017): The Role of Big Data and Predictive Analytics in Retailing. In *Journal of Retailing* 93 (1), pp. 79–95. DOI: 10.1016/j.jretai.2016.12.004.
- Brethenoux, Erick (2016): An ode to the analytics grease monkeys (analytics deployment = ROI). IBM. IBM Data Science Experience. Available online at <http://datascience.ibm.com/blog/an-ode-to-the-analytics-grease-monkeys-analytics-deployment-roi/>, updated on 8/22/2016, checked on 7/4/2017.
- Brockwell, Peter J.; Davis, Richard A. (2002): *Introduction to time series and forecasting*. 2. ed. New York, NY: Springer (Springer texts in statistics).
- Brooks, Patti; El-Gayar, Omar; Sarnikar, Surendra (2015): A framework for developing a domain specific business intelligence maturity model. Application to healthcare. In *International Journal of Information Management* 35 (3), pp. 337–345. DOI: 10.1016/j.ijinfomgt.2015.01.011.
- Bryant, Randal E.; Katz, Randy H.; Lazowska, Edward D. (2008): *Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society*. Computing Community Consortium (CCC). Available online at http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf, checked on 8/4/2017.
- Büchner, Alex G.; Mulvenna, Maurice D.; Anand, Sarab S.; Hughes, John G. (1999): An Internet-enabled Knowledge Discovery Process. In *Proceeding of 9th International Database Conference* 9th International Database Conference (IDC'99), Hong Kong, 15-17 July 1999, pp. 13–27.
- Cabena, Peter; Hadjinian, Pablo; Stadler, Rolf; Verhees, Jaap; Zanasi, Alessandro (1997): *Discovering data mining. From concept to implementation*. Englewood Cliffs, N.J.: Prentice Hall.
- Cao, Longbing (2009): Introduction to Domain Driven Data Mining. In Longbing Cao, Philip S. Yu, Chengqi Zhang, Huaifeng Zhang (Eds.): *Data Mining for Business Applications*. New York, NY: Springer Science+Business Media, pp. 3–10.
- Cao, Longbing (2010): Domain-Driven Data Mining. Challenges and Prospects. In *IEEE Transactions on Knowledge and Data Engineering* 22 (6), pp. 755–769. DOI: 10.1109/TKDE.2010.32.
- Castellano, Marcello; Mastronardi, Giuseppe; Fiorino, Flaviano; Grecis, Giuliano Bellone de; Arcieri, Francesco; Summo, Valerio (2007): *Orchestrating the Knowledge Discovery Process*.

In Jie Lu, Da Ruan, Guangquan Zhang (Eds.): E-Service intelligence. Methodologies, technologies and applications, vol. 37. Berlin: Springer (Studies in Computational Intelligence, 37), pp. 477–496.

Cato, Patrick; Gölzer, Philipp; Demmelhuber, Walter (2015): An investigation into the implementation factors affecting the success of big data systems. In Leila Ismail (Ed.): Proceedings of the 2015 11th International Conference on Innovations in Information Technology (IIT) Innovations 2015. 2015 11th International Conference on Innovations in Information Technology (IIT). Dubai, United Arab Emirates, 1/11/2015 - 3/11/2015. Piscataway, NJ: IEEE, pp. 134–139.

Chakrabarti, Soumen; Nadeau, Thomas P.; Cox, Earl; Neapolitan, Richard E.; Frank, Eibe; Pyle, Dorian et al. (2009): Data mining. Know it all. Burlington: Morgan Kaufmann.

Chamoni, Peter; Gluchowski, Peter (2017): Business Analytics — State of the Art. In *Controlling & Management Review* 61 (4), pp. 8–17. DOI: 10.1007/s12176-017-0030-6.

Chamorro-Premuzic, Tomas (2015): Why Group Brainstorming Is a Waste of Time. Harvard Business Review. Available online at <https://hbr.org/2015/03/why-group-brainstorming-is-a-waste-of-time>, updated on 3/25/2015, checked on 5/4/2017.

Chandrashekar, Girish; Sahin, Ferat (2014): A survey on feature selection methods. In *Computers & Electrical Engineering* 40 (1), pp. 16–28. DOI: 10.1016/j.compeleceng.2013.11.024.

Chang, Pei-Chann; Lai, K. Robert (2005): Combining SOM and Fuzzy Rule Base for Sale Forecasting in Printed Circuit Board Industry. In Jun Wang (Ed.): Advances in neural networks - ISNN 2005. International Symposium on Neural Networks, Chongqing, China, May 30 - June 1, 2005; proceedings, vol. 3498. Berlin [u.a.]: Springer (Lecture Notes in Computer Science, 3498), pp. 947–954.

Chang, Pei-Chann; Liu, Chen-Hao; Fan, Chin-Yuan (2009): Data clustering and fuzzy neural network for sales forecasting. A case study in printed circuit board industry. In *Knowledge-Based Systems* 22 (5), pp. 344–355. DOI: 10.1016/j.knosys.2009.02.005.

Chang, Pei-Chann; Liu, Chen-Hao; Fan, Chin-Yuan; Chang, Hsiao-Ching (2007a): Data Clustering and Fuzzy Neural Network for Sales Forecasting in Printed Circuit Board Industry. In : IEEE Symposium on Computational Intelligence and Data Mining. CIDM 2007. Honolulu, HI, USA, April 1 - 5, 2007. Piscataway, NJ: IEEE, pp. 107–113.

Chang, Pei-Chann; Liu, Chen-Hao; Lai, Robert K. (2008): A fuzzy case-based reasoning model for sales forecasting in print circuit board industries. In *Expert Systems with Applications* 34 (3), pp. 2049–2058. DOI: 10.1016/j.eswa.2007.02.011.

Chang, Pei-Chann; Liu, Chen-Hao; Wang, Yen-Wen (2006a): A hybrid model by clustering and evolving fuzzy rules for sales decision supports in printed circuit board industry. In *Decision Support Systems* 42 (3), pp. 1254–1269. DOI: 10.1016/j.dss.2005.10.013.

- Chang, Pei-Chann; Liu, Chen-Hao; Yeh, Chia-Hsuan; Chen, Shih-Hsin (2006b): The development of a weighted evolving fuzzy neural network. In De-Shuang Huang, George William Irwin, Kang Li (Eds.): International Conference on Intelligent Computing. ICIC 2006. Kunming, China, August 16-19, 2006. Berlin Heidelberg: Springer-Verlag (Lecture Notes in Computer Science, 4114), pp. 212–221.
- Chang, Pei-Chann; Wang, Yen-Wen (2006): Fuzzy Delphi and back-propagation model for sales forecasting in PCB industry. In *Expert Systems with Applications* 30 (4), pp. 715–726. DOI: 10.1016/j.eswa.2005.07.031.
- Chang, Pei-Chann; Wang, Yen-Wen; Liu, Chen-Hao (2005a): Fuzzy Back-Propagation Network for PCB Sales Forecasting. In Ke Chen, Yew Soon Ong, Lipo Wang (Eds.): Advances in Natural Computation, vol. 3610. Berlin: Springer-Verlag (Lecture Notes in Computer Science, 3610), pp. 364–373.
- Chang, Pei-Chann; Wang, Yen-Wen; Liu, Chen-Hao (2007b): The development of a weighted evolving fuzzy neural network for PCB sales forecasting. In *Expert Systems with Applications* 32 (1), pp. 86–96. DOI: 10.1016/j.eswa.2005.11.021.
- Chang, Pei-Chann; Wang, Yen-Wen; Tsai, Chi-Yang (2005b): Evolving neural network for printed circuit board sales forecasting. In *Expert Systems with Applications* 29 (1), pp. 83–92. DOI: 10.1016/j.eswa.2005.01.012.
- Chapelle, Olivier; Vapnik, Vladimir; Bousquet, Olivier; Mukherjee, Sayan (2002): Choosing multiple parameters for support vector machines. In *Machine Learning* 41 (1), pp. 131–159.
- Chapman, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin; Wirth, Rüdiger (2000): CRISP-DM 1.0. Step-by-step data mining guide. CRISP-DM consortium. Available online at <http://www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISPWP-0800.pdf>, checked on 6/27/2017.
- Charest, M.; Delisle, S.; Cervantes, O.; Shen, Yanfen (2006): Intelligent Data Mining Assistance via CBR and Ontologies. In : 17th International Conference on Database and Expert Systems Applications. Proceedings DEXA 2006. 17th International Conference on Database and Expert Systems Applications (DEXA'06). Krakow, Poland, 04-08 Sept. 2006. Los Alamitos, CA: IEEE Computer Society, pp. 593–597.
- Chase, Charles W. (2013a): Demand-driven forecasting. A structured approach to forecasting. Second edition. Hoboken, NJ: Wiley (Wiley & SAS business series).
- Chase, Charles W. (2013b): Using Big Data to Enhance Demand-Driven Forecasting and Planning. In *Journal of Business Forecasting* (Summer 2013), pp. 27–32.
- Chase, Charles W. (2014): Innovations in Business Forecasting: Predictive Analytics. In *Journal of Business Forecasting*, pp. 26–32.

BIBLIOGRAPHY

- Chatfield, Chris (2016): *The Analysis of Time Series. An Introduction*. 6th ed. Hoboken, NJ: Chapman & Hall - CRC Press (Texts in Statistical Science).
- Chawla, Nitesh V. (2010): Data Mining for Imbalanced Datasets: An Overview. In Oded Maimon, Lior Rokach (Eds.): *Data Mining and Knowledge Discovery Handbook*. New York, NY: Springer, pp. 875–886.
- Chen, Edward (2015): Issues and Considerations in the Application of Data Mining in Business. In Corlane Barclay, Kweku-Muata Osei-Bryson (Eds.): *Knowledge discovery process and methods to enhance organizational performance*. Boca Raton, FL: CRC Press, pp. 123–141.
- Chen, Feng; Deng, Pan; Wan, Jiafu; Zhang, Daqiang; Vasilakos, Athanasios V.; Rong, Xiaohui (2015): Data Mining for the Internet of Things. Literature Review and Challenges. In *International Journal of Distributed Sensor Networks* 11 (8), Article ID 431047 (14 pages). DOI: 10.1155/2015/431047.
- Chen, Hong-Mei; Kazman, Rick; Haziyevev, Serge (2016): Agile Big Data Analytics for Web-Based Systems. An Architecture-Centric Approach. In *IEEE Transactions on Big Data* 2 (3), pp. 234–248. DOI: 10.1109/TBDDATA.2016.2564982.
- Chen, Min; Mao, Shiwen; Liu, Yunhao (2014): Big Data. A Survey. In *Mobile Networks and Applications* 19 (2), pp. 171–209. DOI: 10.1007/s11036-013-0489-0.
- Chen, Ming-Syan; Han, Jiawei; Yu, Philip S. (1996): Data mining. An overview from a database perspective. In *IEEE Transactions on Knowledge and Data Engineering* 8 (6), pp. 866–883. DOI: 10.1109/69.553155.
- Chong, Alain Yee Loong; Li, Boying; Ngai, Eric W.T.; Ch'ng, Eugene; Lee, Filbert (2016): Predicting online product sales via online reviews, sentiments, and promotion strategies. A big data architecture and neural network approach. In *International Journal of Operations & Production Management* 36 (4), pp. 358–383. DOI: 10.1108/IJOPM-03-2015-0151.
- Cios, K. J.; Teresinska, A.; Konieczna, S.; Potocka, J.; Sharma, S. (2000): A knowledge discovery approach to diagnosing myocardial perfusion. In *IEEE Engineering in Medicine and Biology Magazine* 19 (4), pp. 17–25. DOI: 10.1109/51.853478.
- Cios, Krzysztof J.; Kurgan, Lukasz A. (2005): Trends in Data Mining and Knowledge Discovery. In Nikhil R. Pal, Lakhmi Jain (Eds.): *Advanced Techniques in Knowledge Discovery and Data Mining*. London: Springer, pp. 1–26.
- Cios, Krzysztof J.; Pedrycz, Witold; Swiniarski, Roman W.; Kurgan, Lukasz A. (2007): *Data mining. A knowledge discovery approach*. New York, NY: Springer.
- Clark, Todd; Wiesenfeld, Dan (2017): 3 Things Are Holding Back Your Analytics, and Technology Isn't One of Them. In *Harvard Business Review*. Available online at <https://hbr.org/2017/06/3-things-are-holding-back-your-analytics-and-technology-isnt-one-of-them>, checked on 7/10/2017.

- Cleve, Jürgen; Lämmel, Uwe (2014): *Data Mining*. München: De Gruyter Oldenbourg.
- Coelho, Vitor N.; Coelho, Igor M.; Rios, Eyder; Filho, Alexandre S.T.; Reis, Agnaldo J.R.; Coelho, Bruno N. et al. (2016): A Hybrid Deep Learning Forecasting Model Using GPU Disaggregated Function Evaluations Applied for Household Electricity Demand Forecasting. In *Energy Procedia* 103, pp. 280–285. DOI: 10.1016/j.egypro.2016.11.286.
- Collier, Ken (2012): *Agile analytics. A value-driven approach to business intelligence and data warehousing*. Upper Saddle River, NJ: Addison-Wesley.
- Comin, Diego; Mulani, Sunil (2004): *Diverging Trends in Macro and Micro Volatility: Facts*. National Bureau of Economic Research. Available online at <http://www.nber.org/papers/w10922>, checked on 7/4/2017.
- Corne, David; Dissanayake, Manjula; Peacock, Andrew; Galloway, Stuart; Owens, Eddie (2014): Accurate localized short term weather prediction for renewables planning. In : 2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG). 2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG). Orlando, FL, USA, 9/12/2014 - 12/12/2014. Piscataway, NJ: IEEE, pp. 1–8.
- Cosic, Ranko; Shanks, Graeme; Maynard, Sean (2012): Towards a business analytics capability maturity model. In John Lamp (Ed.): *ACIS 2012 : Location, location, location : Proceedings of the 23rd Australasian Conference on Information Systems 2012*. 23rd Australasian Conference on Information Systems 2012. Geelong, Australia, 3-5 Dec. 2012: ACIS (Australasian Conference on Information Systems), pp. 1–11. Available online at <http://dro.deakin.edu.au/view/DU:30049067>, checked on 5/22/2017.
- Court, David (2015): Getting big impact from big data. In *McKinsey Quarterly* (January). Available online at <https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/getting-big-impact-from-big-data>, checked on 8/4/2017.
- Cox, Michael; Ellsworth, David (1997): Application-controlled demand paging for out-of-core visualization. In Roni Yagel (Ed.): *Proceedings of the 8th conference on Visualization '97*. VIS '97. Phoenix, AZ, October 18 - 24, 1997. Los Alamitos, CA: IEEE Computer Society Press, pp. 235–244.
- Crone, Sven F. (2010): *Neuronale Netze zur Prognose und Disposition im Handel*. 1. Aufl. Wiesbaden: Gabler (Gabler Research : Betriebliche Forschung zur Unternehmensführung).
- Das, Manirupa; Cui, Renhao; Campbell, David R.; Agrawal, Gagan; Ramnath, Rajiv (2015): Towards methods for systematic research on big data. In Howard Ho, Beng Chin Ooi, Mohammed J. Zaki, Xiaohua Hu, Laura Haas, Vipin Kumar et al. (Eds.): *2015 IEEE International Conference on Big Data*. 2015 IEEE International Conference on Big Data. Santa Clara, CA, USA, Oct 29-Nov 01, 2015: IEEE, pp. 2072–2081.

BIBLIOGRAPHY

Das, Sanmay (2001): Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection. In Carla E. Brodley, Andrea Pohoreckj Danyluk (Eds.): Machine learning. Proceedings of the Eighteenth International Conference (ICML 2001) : Williams College, June 28-July 1, 2001. Eighteenth International Conference on Machine Learning. Williams College, June 28-July 1, 2001. San Francisco, CA: Morgan Kaufmann Publishers, pp. 74–81.

Davenport, Thomas H. (2006): Competing on Analytics. Harvard Business Review. Available online at <https://hbr.org/2006/01/competing-on-analytics>, checked on 8/4/2017.

Davenport, Thomas H. (2013): Keep Up With Your Quants. An innumerate's guide to navigating big data. In *Harvard Business Review* 91 (7/8 (July-August)), pp. 120–123.

Davenport, Thomas H. (2014): big data @ work. Chancen erkenne, Risiken verstehen. 1. Aufl. München: Franz Vahlen.

Davenport, Thomas H.; Barth, Paul; Bean, Randy (2012): How 'Big Data' Is Different. MIT Sloan Management Review. Magazine: Fall 2012. Available online at <http://sloanreview.mit.edu/article/how-big-data-is-different/>, updated on 7/30/2012, checked on 12/10/2016.

Davenport, Thomas H.; Dyché, Jill (2013): Big Data in Big Companies. International Institute for Analytics. Available online at http://docs.media.bitpipe.com/io_10x/io_102267/item_725049/Big-Data-in-Big-Companies.pdf, checked on 10/12/2016.

Dean, Jared (2014): Big data, data mining, and machine learning. Value creation for business leaders and practitioners. Hoboken, New Jersey: Wiley (Wiley & SAS business series).

Debusse, J.C.W.; La Iglesia, B. de; Howard, C. M.; Rayward-Smith, V. J. (2001): Building the KDD Roadmap. In Rajkumar Roy (Ed.): Industrial knowledge management. A micro-level approach. London: Springer, pp. 179–196.

Dedic, Nedim; Stanier, Clare (2017): Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery. In Felix Piazzolo, Verena Geist, Lars Brehm, Rainer Schmidt (Eds.): Innovations in Enterprise Information Systems Management and Engineering. Revised Papers. 5th International Conference, ERP Future 2016 - Research. Hagenberg, Austria, November 14, 2016. Cham: Springer International Publishing (Lecture Notes in Business Information Processing, 285), pp. 114–122.

Demchenko, Yuri; Grosso, Paola; Laat, Cees de; Membrey, Peter (2013): Addressing big data issues in Scientific Data Infrastructure. In Waleed W. Smari (Ed.): International Conference on Collaboration Technologies and Systems (CTS), 2013. 2013 International Conference on Collaboration Technologies and Systems (CTS). San Diego, CA, USA, 20-24 May 2013. International Conference on Collaboration Technologies and Systems (CTS); International Symposium on Big Data and Data Analytics in Collaboration (BDDAC); International Workshop on Collaborative Mobile Systems and Sensors Networks (CMSSN); International

Workshop on E-Transactions Systems (ETS); International Symposium on Collaboration, Social Computing, New Media, and Networks (SoMNet); International Symposium on Security in Collaboration Technologies and Systems (SECOTS); International Workshop on Collaborative Robots and Human Robot Interaction (CR-HRI); International Workshop on Collaborations in Emergency Response and Disaster Management (ERDM); International Workshop on Collaboration and Gaming (CoGames); International Workshop on Collaboration Technologies and Systems in Healthcare and Biomedical Fields (CoHeB). Piscataway, NJ: IEEE, pp. 48–55.

Destrero, Augusto; Mosci, Sofia; Mol, Christine de; Verri, Alessandro; Odone, Francesca (2009): Feature selection for high-dimensional data. In *Computational Management Science* 6 (1), pp. 25–40. DOI: 10.1007/s10287-008-0070-7.

Deubel, Thomas (2017): Profitabel - Der Effekt von Agilität auf das Unternehmensergebnis. In Christian Ramsauer, Detlef Kayser, Christoph Schmitz (Eds.): *Erfolgsfaktor Agilität. Chancen für Unternehmen in einem volatilen Marktumfeld*. Weinheim: Wiley-VCH, pp. 101–125.

Dhar, Vasant (2013): Data science and prediction. In *Communications of the ACM* 56 (12), pp. 64–73. DOI: 10.1145/2500499.

Diamantini, Claudia; Potena, Domenico; Smari, Waleed W. (2006): Collaborative Knowledge Discovery in Databases: A Knowledge Exchange Perspective. In Vasant Honavar, Timothy W. Finin (Eds.): *Semantic Web for collaborative knowledge acquisition. Papers from the AAAI Fall Symposium*. Menlo Park, CA: AAAI Press (Technical report, FS-06-06), pp. 24–31. Available online at <http://www.aaai.org/Papers/Symposia/Fall/2006/FS-06-06/FS06-06-004.pdf>, checked on 7/4/2017.

Domingos, Pedro (2012): A few useful things to know about machine learning. In *Communications of the ACM* 55 (10), pp. 78–87. DOI: 10.1145/2347736.2347755.

Donoho, David L. (2000): High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. AMS Math Challenges Lecture. Available online at <http://statweb.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf>, updated on 8/9/2000, checked on 3/12/2017.

Dorschel, Joachim (2015): Einführung und Überblick. In Joachim Dorschel (Ed.): *Praxishandbuch Big Data. Wirtschaft - Recht - Technik*. Wiesbaden: Springer Gabler, pp. 5–13.

Dutcher, Jennifer (2014): What Is Big Data? Berkeley School of Information. datascience@berkeley blog. Available online at <https://datascience.berkeley.edu/what-is-big-data>, updated on 9/3/2014, checked on 7/3/2017.

Dutta, Debprotim; Bose, Indranil (2015): Managing a Big Data project. The case of Ramco Cements Limited. In *International Journal of Production Economics* 165, pp. 293–306. DOI: 10.1016/j.ijpe.2014.12.032.

BIBLIOGRAPHY

Earley, Seth (2014): Agile Analytics in the Age of Big Data. In *IT Professional* 16 (4), pp. 18–20. DOI: 10.1109/MITP.2014.44.

ECB (2017): ECB euro reference exchange rate: Japanese yen (JPY). European Central Bank. Available online at https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/eurofxref-graph-jpy.en.html, updated on 8/28/2017, checked on 8/28/2017.

Eckerson, Wayne W. (2007): Predictive Analytics. Extending the Value of Your Data Warehousing Investment. TDWI (TDWI Best Practices Report). Available online at https://tdwi.org/research/2007/01/bpr-1q-predictive-analytics/bpr_1q07_report.aspx?tc=assetpg, checked on 10/12/2016.

Edwards, Rosalind; Holland, Janet (2013): What is qualitative interviewing? London: Bloomsbury (What is? Research methods series).

Eisenführ, Franz; Weber, Martin (2013): Rationales Entscheiden. Zweite, verbesserte Auflage. Berlin: Springer (Springer-Lehrbuch).

Elgendy, Nada; Elragal, Ahmed (2014): Big Data Analytics: A Literature Review Paper. In Petra Perner (Ed.): Advances in Data Mining. Applications and Theoretical Aspects. 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014. Proceedings, vol. 8557. 14th Industrial Conference, ICDM 2014. St. Petersburg, Russia, July 16-20, 2014. Cham: Springer International Publishing; Imprint; Springer (Lecture Notes in Computer Science, 8557), pp. 214–227.

Elgendy, Nada; Elragal, Ahmed (2016): Big Data Analytics in Support of the Decision Making Process. In *Procedia Computer Science* 100, pp. 1071–1084. DOI: 10.1016/j.procs.2016.09.251.

Elragal, Ahmed (2014): ERP and Big Data. The Inept Couple. In *Procedia Technology* 16, pp. 242–249. DOI: 10.1016/j.protcy.2014.10.089.

El-Sappagh, Shaker H. Ali; Hendawi, Abdeltawab M. Ahmed; El Bastawissy, Ali Hamed (2011): A proposed model for data warehouse ETL processes. In *Journal of King Saud University - Computer and Information Sciences* 23 (2), pp. 91–104. DOI: 10.1016/j.jksuci.2011.05.005.

EMC Education Services (2015): Data science and big data analytics. Discovering, analyzing, visualizing and presenting data. Indianapolis, IN: Wiley.

Ereth, Julian; Kemper, Hans-Georg (2016): Business Analytics und Business Intelligence. In *Controlling* 28 (8-9), pp. 458–464. DOI: 10.15358/0935-0381-2016-8-9-458.

Erl, Thomas; Khattak, Wajid; Buhler, Paul (2016): Big data fundamentals. Concepts, drivers & techniques. Boston, MA: Prentice Hall; Service Tech Press (The Prentice Hall service technology series from Thomas Erl).

Espinosa, J. Alberto; Armour, Frank (2016): The Big Data Analytics Gold Rush: A Research Framework for Coordination and Governance. In Tung X. Bui, Ralph H. Sprague (Eds.):

- Proceedings of the 49th Annual Hawaii International Conference on System Sciences. 2016 49th Hawaii International Conference on System Sciences (HICSS). Koloa, HI, USA, 5/1/2016 - 8/1/2016. Hawaii International Conference on System Sciences; Annual Hawaii International Conference on System Sciences; HICSS. Piscataway, NJ: IEEE, pp. 1112–1121.
- Eurostat (2017a): About Eurostat: Overview. Available online at <http://ec.europa.eu/eurostat/about/overview>, checked on 5/12/2017.
- Eurostat (2017b): Data: Database. Available online at <http://ec.europa.eu/eurostat/data/database>, checked on 5/12/2017.
- Eurostat (2017c): Help: Database. Available online at <http://ec.europa.eu/eurostat/help/first-visit/database>, checked on 5/16/2017.
- Evans, Michael K. (2003): Practical business forecasting. Oxford: Blackwell Publishers.
- Fan, Jianqing; Han, Fang; Liu, Han (2014): Challenges of Big Data Analysis. In *National science review* 1 (2), pp. 293–314. DOI: 10.1093/nsr/nwt032.
- Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996a): From Data Mining to Knowledge Discovery in Databases. In *AI Magazine* 17 (3), pp. 37–54.
- Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996b): Knowledge Discovery and Data Mining: Towards a Unifying Framework. In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad (Eds.): Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Second International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, OR, USA, August 2–4, 1996. Menlo Park, CA: AAAI Press, pp. 82–88.
- Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996c): The KDD process for extracting useful knowledge from volumes of data. In *Communications of the ACM* 39 (11), pp. 27–34. DOI: 10.1145/240455.240464.
- Feinleib, David (2014): Big Data Bootcamp. What Managers Need to Know to Profit from the Big Data Revolution. Berkeley, CA: Apress.
- Feldman, Ronen; Sanger, James (2006): The text mining handbook. Advanced approaches in analyzing unstructured data. New York, NY: Cambridge University Press.
- Finlay, Steven (2014): Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods. Basingstoke: Palgrave Macmillan (Business in the Digital Economy).
- Fischer, Marco; Wieland, Uwe; Hilbert, Andreas (2014): Industrie 4.0 und Data-Mining-Projekte - interdisziplinäre Zusammenarbeit von Ingenieuren und Datenanalysten im Produktionsprozess. In Dennis Kundisch, Leena Suhl, Lars Beckmann (Eds.): MKWI 2014 - Multikonferenz Wirtschaftsinformatik. Tagungsband. Paderborn, Germany, 26. - 28. Februar 2014. Paderborn: Universität Paderborn, pp. 167–180.

- Fisher, Danyel; DeLine, Rob; Czerwinski, Mary; Drucker, Steven (2012): Interactions with big data analytics. In *interactions* 19 (3), pp. 50–59. DOI: 10.1145/2168931.2168943.
- Fogarty, David; Bell, Peter C. (2014): Should You Outsource Analytics? In *MIT Sloan Management Review* 55 (2), pp. 41–45.
- Fogelman-Soulié, Françoise; Lu, Wenhuan (2016): Implementing Big Data Analytics Projects in Business. In Nathalie Japkowicz, Jerzy Stefanowski (Eds.): *Big Data Analysis. New Algorithms for a New Society*. 1st ed. 2016. Cham: Springer International Publishing (Studies in Big Data, 16), pp. 141–158.
- Forman, George (2003): An Extensive Empirical Study of Feature Selection Metrics for Text Classification. In *Journal of Machine Learning Research* 2, pp. 1289–1305.
- Francois, Damien (2008): Methodology and standards for data analysis with machine learning tools. In Michel Verleysen (Ed.): *ESANN 2008 Proceedings. 16th European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. 16th European Symposium on Artificial Neural Networks*. Bruges, Belgium, April 23-24-25, 2008. Evere: d-side, pp. 239–246.
- Frankova, Patricia; Drahosova, Martina; Balco, Peter (2016): Agile Project Management Approach and its Use in Big Data Management. In *Procedia Computer Science* 83, pp. 576–583. DOI: 10.1016/j.procs.2016.04.272.
- Franks, Bill (2012): *Taming the big data tidal wave. Finding opportunities in huge data streams with advanced analytics*. Hoboken, NJ: John Wiley & Sons (Wiley et SAS business series).
- Franks, Bill (2014): *The analytics revolution. How to improve your business by making analytics operational in the big data era*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Fraser, Neil (2017): Strategy Development and Big Data Analytics. In Hwaiyu Geng (Ed.): *Internet of things and data analytics handbook*. Hoboken, NJ: John Wiley & Sons, Inc, pp. 347–364.
- Friedli, Thomas; Schuh, Günther (2012): *Wettbewerbsfähigkeit der Produktion an Hochlohnstandorten*. 2. Aufl. Berlin: Springer Vieweg.
- Fukui, Tomohiro (2016): A systems approach to big data technology applied to supply chain. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): *Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data*. Washington D.C., USA, Dec 05-Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 3732–3736.
- Gandomi, Amir; Haider, Murtaza (2015): Beyond the hype. Big data concepts, methods, and analytics. In *International Journal of Information Management* 35 (2), pp. 137–144. DOI: 10.1016/j.ijinfomgt.2014.10.007.

-
- Ganesan, Kavita (2014): Computing Precision and Recall for Multi-Class Classification Problems. Text Mining, Analytics & More. Available online at <http://text-analytics101.rxnlp.com/2014/10/computing-precision-and-recall-for.html>, updated on 10/20/2014, checked on 9/12/2016.
- Gansser, Oliver; Krol, Bianca (2015): Markt- und Absatzprognosen. Wiesbaden: Springer Fachmedien.
- Gao, Jing; Koronios, Andy; Selle, Sven (2015): Towards A Process View on Critical Success Factors in Big Data Analytics Projects. In : 21st Americas Conference on Information Systems (AMCIS 2015). 21st Americas Conference on Information Systems (AMCIS 2015). Fajardo, Puerto Rico, 13-15 August 2015. Atlanta, Georgia: Association for Information Systems, pp. 824–837.
- Gartner (2011): Gartner Reveals Top Predictions for IT Organizations and Users for 2012 and Beyond. Press Release. Available online at <https://www.gartner.com/newsroom/id/1862714>, updated on 12/1/2011, checked on 8/4/2017.
- Gartner (2017): Big Data. Gartner IT Glossary. Available online at <http://www.gartner.com/it-glossary/big-data/>, checked on 1/7/2017.
- GE Reports (2015): Data - an airline's most productive jet stream. Available online at <https://www.ge.com/reports/09-09-2015data-an-airline-s-most-productive-jet-stream/>, checked on 8/4/2017.
- Gentsch, Peter; Kulpa, Andreas (2016): Mit externen Big Data neue Möglichkeiten erschließen. In *Controlling & Management Review* (Sonderheft 1), pp. 32–38. DOI: 10.1007/978-3-658-13444-0_4.
- Gertosio, Christine; Dussauchoy, Alan (2004): Knowledge discovery from industrial databases. In *Journal of Intelligent Manufacturing* 15 (1), pp. 29–37. DOI: 10.1023/B:JIMS.0000010073.54241.e7.
- Giudici, Paolo (2003): Applied data mining. Statistical methods for business and industry. Chichester: John Wiley & Sons.
- Grady, Nancy W. (2016): Knowledge Discovery in Data Science. KDD meets Big Data. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data. Washington D.C., USA, Dec 05-Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 1603–1608.
- Grady, Nancy W.; Underwood, Mark; Roy, Arnab; Chang, Wo L. (2014): Big Data: Challenges, practices and technologies. NIST Big Data Public Working Group Workshop at IEEE Big Data 2014. In Jimmy Lin (Ed.): 2014 IEEE International Conference on Big Data. 2014 IEEE

BIBLIOGRAPHY

International Conference on Big Data (Big Data). Washington, DC, USA, 27-30 Oct. 2014. Piscataway, NJ: IEEE, pp. 11–15.

Green, Kesten C.; Armstrong, J. Scott (2015): Simple versus complex forecasting. The evidence. In *Journal of Business Research* 68 (8), pp. 1678–1685. DOI: 10.1016/j.jbusres.2015.03.026.

Grimes, Seth (2008): Unstructured Data and the 80 Percent Rule. Breakthrough Analysis. Available online at <https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/>, updated on 8/1/2008, checked on 10/5/2016.

Gronau, Norbert; Thim, Christof; Fohrholz, Corinna (2016): Business Analytics in der deutschen Praxis. In *Controlling* 28 (8-9), pp. 472–479. DOI: 10.15358/0935-0381-2016-8-9-472.

Grubbs, Frank E. (1969): Procedures for Detecting Outlying Observations in Samples. In *Technometrics* 11 (1), pp. 1–21. DOI: 10.2307/1266761.

Guarino, Nicola (1998): Formal Ontology and Information Systems. In Nicola Guarino (Ed.): *Formal ontology in information systems. Proceedings of the first international conference (FOIS'98)*, June 6-8, Trento, Italy. Amsterdam: IOS Press, pp. 3–15.

Gujarati, Damodar N. (2003): *Basic econometrics*. 4th edition. Boston, MA: McGraw Hill.

Guo, Philip (2013): Data Science Workflow: Overview and Challenges. Edited by blog@CACM. Communications of the ACM. Available online at <https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext>, updated on 10/30/2013, checked on 6/22/2017.

Guyon, Isabelle; Elisseeff, Andre (2003): An Introduction to Variable and Feature Selection. In *Journal of Machine Learning Research* 3, pp. 1157–1182.

Guyon, Isabelle; Elisseeff, Andre (2006): An Introduction to Feature Extraction. In Isabelle Guyon, Steve Gunn, Masoud Nikravesh, Lotfi A. Zadeh (Eds.): *Feature extraction. Foundations and applications*. Berlin: Springer (Studies in Fuzziness and Soft Computing, 207), pp. 1–25.

Guyon, Isabelle; Gunn, Steve; Hur, Asa Ben; Dror, Gideon (2006): Design and Analysis of the NIPS2003 Challenge. In Isabelle Guyon, Steve Gunn, Masoud Nikravesh, Lotfi A. Zadeh (Eds.): *Feature extraction. Foundations and applications*. Berlin: Springer (Studies in Fuzziness and Soft Computing, 207), pp. 237–263.

Hadavandi, Esmail; Shavandi, Hassan; Ghanbari, Arash (2011): An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering. Case study of printed circuit board. In *Expert Systems with Applications* 38 (8), pp. 9392–9399. DOI: 10.1016/j.eswa.2011.01.132.

Haffar, Jason (2015): Have you seen ASUM-DM? IBM. developerWorks. Available online at <https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/>, updated on 10/16/2015, checked on 7/4/2017.

- Hagen, Christian; Evans, Hugo; Miller, Jason; Ciobo, Marco; Wall, Dan; Yadav, Ajay (2013): Big Data and the Creative Destruction of Today's Business Models. A.T. Kearney (Ideas and Insights). Available online at <https://www.atkearney.at/documents/10192/698536/Big+Data+and+the+Creative+Destruction+of+Todays+Business+Models.pdf/f05aed38-6c26-431d-8500-d75a2c384919>, checked on 6/27/2017.
- Haglin, D.; Roiger, R.; Hakkila, J.; Giblin, T. (2005): A tool for public analysis of scientific data. In *Data Science Journal* 4 (8), pp. 39–53. DOI: 10.2481/dsj.4.39.
- Halmos, Paul R. (1974): Naive set theory. New York, NY: Springer (Undergraduate Texts in Mathematics).
- Halper, Fern (2014): Predictive Analytics for Business Advantage. TDWI Best Practices Report. TDWI Research. Available online at <https://tdwi.org/research/2013/12/best-practices-report-predictive-analytics-for-business-advantage/asset.aspx?tc=assetpg>, checked on 4/20/2017.
- Hammer, Markus; Somers, Ken; Karre, Hugo; Ramsauer, Christian (2017): Profit Per Hour as a Target Process Control Parameter for Manufacturing Systems Enabled by Big Data Analytics and Industry 4.0 Infrastructure. In *Procedia CIRP* 63, pp. 715–720. DOI: 10.1016/j.procir.2017.03.094.
- Han, Jiawei; Kamber, Micheline; Pei, Jian (2012): Data mining. Concepts and techniques. 3rd ed. Waltham, MA: Morgan Kaufmann Publishers - Elsevier (The Morgan Kaufmann series in data management systems).
- Hand, David J.; Mannila, Heikki; Smyth, Padhraic (2001): Principles of data mining. Cambridge, MA: MIT Press (Adaptive computation and machine learning).
- Hashem, Ibrahim Abaker Targio; Yaqoob, Ibrar; Anuar, Nor Badrul; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015): The rise of “big data” on cloud computing. Review and open research issues. In *Information Systems* 47, pp. 98–115. DOI: 10.1016/j.is.2014.07.006.
- Hastie, Trevor; Qian, Junyang (2014): Glmnet Vignette. Stanford University. Available online at https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html#lin, updated on 6/26/2014, checked on 7/11/2016.
- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. (2017): The elements of statistical learning. Data mining, inference, and prediction. Second edition, corrected at 12th printing. New York, NY: Springer (Springer series in statistics).
- Hastie, Trevor J.; Tibshirani, Robert John; Wainwright, Martin J. (2015): Statistical learning with Sparsity. The lasso and generalizations. Boca Raton: CRC Press, Taylor & Francis Group (Monographs on statistics and applied probability, 143).
- Heath, Fenno F.; Hull, Richard (2015): Analytics Process Management: A New Challenge for the BPM Community. In Fabiana Fournier, Jan Mendling (Eds.): Business process management

workshops, vol. 202. BPM 2014 International Workshops. Eindhoven, The Netherlands, September 7–8, 2014. Cham: Springer (Lecture Notes in Business Information Processing, 202), pp. 175–185.

Heinen, Tobias; Rimpau, Christoph; Wörn, Arno (2008): Wandlungsfähigkeit als Ziel der Produktionssystemgestaltung. In Peter Nyhuis, Gunther Reinhart, Eberhard Abele (Eds.): Wandlungsfähige Produktionssysteme. Heute die Industrie von morgen gestalten, 19–32.

Heinrich, Lutz J.; Riedl, René; Stelzer, Dirk (2014): Informationsmanagement. Grundlagen, Aufgaben, Methoden. 11. Auflage. München: De Gruyter Oldenbourg.

Heit, Juergen; Liu, Jiayi; Shah, Mohak (2016): An Architecture for the Deployment of Statistical Models for the Big Data Era. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data. Washington D.C., USA, Dec 05–Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 1377–1384.

Heldmann, Stefan (2017): Informiert - Monitoring als Schnittstelle zum unsicheren Geschäftsumfeld. In Christian Ramsauer, Detlef Kayser, Christoph Schmitz (Eds.): Erfolgsfaktor Agilität. Chancen für Unternehmen in einem volatilen Marktumfeld. Weinheim: Wiley-VCH, pp. 161–200.

Heldmann, Stefan; Hammer, Markus; Ramsauer, Christian (2017): Eine strategische und operative Perspektive zur Anwendung von Big Data in der Industrie. In *ZWF* 112 (1-2), pp. 79–85.

Heldmann, Stefan; Rabitsch, Christian; Ramsauer, Christian (2015): Big Data-basiertes Monitoring. Ein neuer Ansatz für agile Industrieunternehmen in der volatilen Welt. In *Industrie 4.0 Management* 31 (5), pp. 35–39.

Hemmatfar, Mahmood; Salehi, Mahdi; Bayat, Marziyeh (2010): Competitive Advantages and Strategic Information Systems. In *International Journal of Business and Management* 5 (7), pp. 158–169. DOI: 10.5539/ijbm.v5n7p158.

Henke, Nicolaus; Bughin, Jacques; Chui, Michael; Manyika, James; Saleh, Tamim; Wiseman, Bill; Sethupathy, Guru (2016): The Age of Analytics. Competing in a data-driven world. McKinsey Global Institute. Available online at <http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world>, checked on 7/25/2017.

Henke, Nicolaus; Levine, Jordan; McInerney, Paul (2018): You Don't Have to Be a Data Scientist to Fill This Must-Have Analytics Role. In *Harvard Business Review*. Available online at <https://hbr.org/product/you-dont-have-to-be-a-data-scientist-to-fill-this-must-have-analytics-role/H045B7-PDF-ENG>, checked on 2/6/2018.

-
- Hevner, Alan; Chatterjee, Samir (2010): Design Research in Information Systems. Theory and Practice. 1. Aufl. New York: Springer (Integrated series in information systems, 22).
- Hevner, Alan R. (2007): A Three Cycle View of Design Science Research. In *Scandinavian Journal of Information Systems* 19 (2), pp. 87–92.
- Hevner, Alan R.; March, Salvatore T.; Park, Jinsoo; Ram, Sudha (2004): Design Science in Information Systems Research. In *MIS Quarterly* 28 (1), pp. 75–105.
- Hicham, Attarius; Mohamed, Bouhorma; Abdellah, El Fallahi (2012a): A model for sales forecasting based on fuzzy clustering and Back-propagation Neural Networks with adaptive learning rate. In Mohamed Essaaidi (Ed.): 2012 IEEE International Conference on Complex Systems. Agadir, Morocco, 5 - 6 Nov. 2012. ICCS. Agadir, Morocco, 5/11/2012 - 6/11/2012. Piscataway, NJ: IEEE, pp. 1–5.
- Hicham, Attarius; Mohamed, Bouhorma; Abdellah, El Fallahi (2012b): An improved approach based on fuzzy clustering and Back-Propagation Neural Networks with adaptive learning rate for sales forecasting: Case study of PCB industry. In *International Journal of Computer Science Issues* 9 (3), pp. 404–413.
- Hichama, Attarius; Mohameda, Bouhorma; Abdellahb, El Fallahi (2013): A novel approach based on genetic fuzzy clustering and adaptive neural networks for sales forecasting. In *Journal of Computer Science* 9 (8), pp. 949–966. DOI: 10.3844/jcssp.2013.949.966.
- Hilbert, Andreas (2005): Critical Success Factors for Data Mining Projects. In Daniel Baier, Reinhold Decker, Lars Schmidt-Thieme (Eds.): Data analysis and decision support. Berlin: Springer (Studies in classification, data analysis, and knowledge organization), 231–240.
- Hilbert, Martin; López, Priscila (2011): The world's technological capacity to store, communicate, and compute information. In *Science* 332 (6025), pp. 60–65. DOI: 10.1126/science.1200970.
- Hirji, Karim K. (2001): Exploring data mining implementation. In *Communications of the ACM* 44 (7), pp. 87–93. DOI: 10.1145/379300.379323.
- Hofmann, Markus; Tierney, Brendan (2009): Development of an Enhanced Generic Data Mining Life Cycle (DMLC). In *The ITB Journal* 10 (1), pp. 50–71.
- Horeis, Timo; Sick, Bernhard (2007): Collaborative Knowledge Discovery & Data Mining: From Knowledge to Experience. In : IEEE Symposium on Computational Intelligence and Data Mining. CIDM 2007. Honolulu, HI, USA, April 1 - 5, 2007. Piscataway, NJ: IEEE, pp. 421–428.
- Hsu, Chih-Wei; Chang, Chih-Chung; Lin, Chih-Jen (2016): A Practical Guide to Support Vector Classification. Department of Computer Science, National Taiwan University. Available online at <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>, checked on 8/4/2017.
-

BIBLIOGRAPHY

Huang, Zhichuan; Zhu, Ting (2016): Leveraging multi-granularity energy data for accurate energy demand forecast in smart grids. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data. Washington D.C., USA, Dec 05-Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 1182–1191.

Huberty, Mark (2015): Can we vote with our tweet? On the perennial difficulty of election forecasting with social media. In *International Journal of Forecasting* 31 (3), pp. 992–1007. DOI: 10.1016/j.ijforecast.2014.08.005.

Hyndman, Rob J.; Athanasopoulos, George (2017): ARIMA models. Forecasting: principles and practice. Available online at <https://www.otexts.org/fpp/8>, checked on 1/14/2017.

Hyndman, Rob J.; Khandakar, Yeasmin (2008): Automatic Time Series Forecasting. The forecast Package for R. In *Journal of Statistical Software* 27 (3), pp. 1–22. DOI: 10.18637/jss.v027.i03.

IBM (2016): Analytics Solutions Unified Method. Implementations with Agile principles. IBM. Analytics services - Datasheet. Available online at <ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf>, checked on 7/4/2017.

IBM (2017a): IBM SPSS Modeler 18.0 User's Guide. Available online at <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerUsersGuide.pdf>, checked on 7/3/2017.

IBM (2017b): What is big data? Available online at <https://www.ibm.com/analytics/us/en/big-data/>, checked on 2/15/2017.

IDC (2014a): The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. EMC Digital Universe with Research & Analysis by IDC. Available online at <https://www.emc.com/leadership/digital-universe/2014iview/index.htm>, checked on 8/4/2017.

IDC (2014b): The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Infobrief. EMC Digital Universe with Research & Analysis by IDC. Available online at <https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf>, checked on 8/4/2017.

informs (2017): Operations Research & Analytics. The Institute for Operations Research and the Management Sciences. Available online at <https://www.informs.org/Explore/Operations-Research-Analytics>, checked on 4/3/2017.

Inmon, William H. (2005): Building the Data Warehouse. Fourth edition. Indianapolis, IN: Wiley Publishing.

Inmon, William H.; Strauss, Derek; Neushloss, Genia (2010): DW 2.0. The architecture for the next generation of data warehousing. Amsterdam.: Elsevier, Morgan Kaufmann Publishers (The Morgan Kaufmann series in data management systems).

internet live stats (2017): Twitter Usage Statistics. Available online at <http://www.internetlivestats.com/twitter-statistics/>, checked on 10/18/2017.

Investopedia (2017): Earnings Surprise. Available online at <http://www.investopedia.com/terms/e/earnings Surprise.asp>, checked on 1/19/2016.

IPC (2017): Current industry trends. IPC — Association Connecting Electronics Industries. Available online at <https://www.ipc.org/ContentPage.aspx?pageid=Current-Industry-Trends>, checked on 3/12/2017.

Isson, Jean Paul; Harriott, Jesse (2013): Win with Advanced Business Analytics. Creating business value from your data. Hoboken, NJ: John Wiley & Sons.

Janssen, Marijn; van der Voort, Haiko; Wahyudi, Agung (2017): Factors influencing big data decision-making quality. In *Journal of Business Research* 70, pp. 338–345. DOI: 10.1016/j.jbusres.2016.08.007.

Jensen, Kenneth A. (2017): CRISP-DM. Cross-Industry Standard Process for Data Mining. Available online at <https://sites.google.com/view/kajensen/profile/project-methodology>, checked on 7/8/2017.

Ji, Guojun; Hu, Limei; Tan, Kim Hua (2017a): A study on decision-making of food supply chain based on big data. In *Journal of Systems Science and Systems Engineering* 26 (2), pp. 183–198. DOI: 10.1007/s11518-016-5320-6.

Ji, Guojun; Tan, KimHua; Zhao, L.; Xavior, A.; Cai, J.; You, L. (2017b): A Big Data Decision-making Mechanism for Food Supply Chain. In *MATEC Web of Conferences* 100, Article number 02048. DOI: 10.1051/mateconf/201710002048.

Ji, Xu; Zhong, Ganji; Yu, Yang; Li, Zhongming (2015): Key technologies and application for cloud manufacturing in polymer industry. In *Jisuanji Jicheng Zhibiao Xitong/Computer Integrated Manufacturing Systems* 21 (11), pp. 3072–3078.

Jimenez, Alvaro Barbero; Lazaro, Jorge Lopez; Dorronsoro, Jose R. (2009): Finding optimal model parameters by deterministic and annealed focused grid search. In *Neurocomputing* 72 (13–15), pp. 2824–2832. DOI: 10.1016/j.neucom.2008.09.024.

Jun, Seung-Pyo; Yeom, Jaeho; Son, Jong-Ku (2014): A study of the method using search traffic to analyze new technology adoption. In *Technological Forecasting and Social Change* 81 (1), pp. 82–95. DOI: 10.1016/j.techfore.2013.02.007.

Jurney, Russel (2014): Agile data science. Building data analytics applications with Hadoop. Sebastopol, CA: O'Reilly.

BIBLIOGRAPHY

- Kabiri, Ahmed; Chiadmi, Dalila (2013): Survey on ETL processes. In *Journal of Theoretical and Applied Information Technology* 54 (2), pp. 219–229.
- Kacfeh Emani, Cheikh; Cullot, Nadine; Nicolle, Christophe (2015): Understandable Big Data. A survey. In *Computer Science Review* 17, pp. 70–81. DOI: 10.1016/j.cosrev.2015.05.002.
- Kalgotra, Pankush; Sharda; Ramesh (2016): Progression Analysis of Signals: Extending CRISP-DM to Stream Analytics. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): *Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data*. Washington D.C., USA, Dec 05-Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 2880–2885.
- Kantardzic, Mehmed (2011): *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd Edition. Hoboken, NJ: John Wiley & Sons.
- Karunakaran, K. Priya (2013): Review of Domain Driven Data Mining. In *International Journal of Innovations in Engineering and Technology* 2 (3), pp. 112–116.
- KDnuggets (2002): What main methodology are you using for data mining? KDnuggets. Available online at <http://www.kdnuggets.com/polls/2002/methodology.htm>, updated on July 2002, checked on 7/5/2017.
- KDnuggets (2004): Data Mining Methodology. KDnuggets. Available online at http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm, updated on April 2004, checked on 7/5/2017.
- Keeney, Ralph L. (1992): *Value-focused thinking. A path to creative decisionmaking*. Cambridge, MA: Harvard University Press.
- Kelly, J.; Kaskade, J. (2013): CIOs & Big Data: What Your IT Team Wants You to Know. Infochimps. Available online at <http://www.infochimps.com/resources/report-cios-big-data-what-your-it-team-wants-you-to-know-6/>, checked on 12/20/2015.
- Keogh, Eamonn; Mueen, Abdullah (2010): Curse of Dimensionality. In Claude Sammut, Geoffrey I. Webb (Eds.): *Encyclopedia of machine learning*. New York, NY: Springer, pp. 257–258.
- Kern, Roman (2017): Demand Forecasting. In *TU Graz research* 17 (1), pp. 19–21.
- Kim, Seongdo; Shin, Do Hyung (2016): Forecasting short-term air passenger demand using big data from search engine queries. In *Automation in Construction* 70, pp. 98–108. DOI: 10.1016/j.autcon.2016.06.009.
- Kimball, Ralph; Caserta, Joe (2004): *The Data Warehouse ETL Toolkit. Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Indianapolis, IN: Wiley Publishing.
- Kimball, Ralph; Ross, Margy (2013): *The Data Warehouse Toolkit. The Definitive Guide to Dimensional Modeling*. Hoboken, NJ: John Wiley & Sons.

Kitchenham, Barbara: Procedures for Performing Systematic Reviews. Keele University & National ICT Australia Ltd. (Keele University Technical Report TR/SE-0401 & NICTA Technical Report 0400011T.1). Available online at <http://www.ifs.tuwien.ac.at/~weippl/systemicReviewsSoftwareEngineering.pdf>, checked on 10/16/2016.

Klein, Philip A.; Moore, Geoffrey H. (1983): The leading indicator approach to economic forecasting—retrospect and prospect. In *Journal of Forecasting* 2 (2), pp. 119–135. DOI: 10.1002/for.3980020204.

Klösgen, Willi (2002): Types and Forms of Data. In Willi Klösgen, Jan M. Zytow (Eds.): Handbook of data mining and knowledge discovery. Oxford: Oxford University Press, pp. 33–44.

Kopanas, Ioannis; Avouris, Nikolaos M.; Daskalaki, Sophia (2002): The Role of Domain Knowledge in a Large Scale Data Mining Project. In Ioannis P. Vlahavas, Constantine D. Spyropoulos (Eds.): Methods and applications of artificial intelligence. Proceedings : Second Hellenic Conference on AI, SETN 2002, vol. 2308. Thessaloniki, Greece, April 11 - 12, 2002. Hellenic Conference on AI. Berlin: Springer (Lecture notes in artificial intelligence, 2308), pp. 288–299.

Koumenides, Christos L.; Salvadores, Manuel; Alani, Harith; Shadbolt, Nigel R. (2010): Global Integration of Public Sector Information. Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 26-27th, 2010, Raleigh, NC, USA. Available online at <http://journal.webscience.org/303/>, updated on 10/25/2011, checked on 9/20/2016.

Kousiouris, George; Vafiadis, George; Varvarigou, Theodora (2013): Enabling Proactive Data Management in Virtualized Hadoop Clusters Based on Predicted Data Activity Patterns. In Fatos Xhafa, Leonard Barolli, Dritan Nace, Salvatore Vinticinqu, Alain Bui (Eds.): 2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC). 2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC). Compiegne, France, 28 - 30 Oct. 2013. Piscataway, NJ: IEEE, pp. 1–8.

Kremsmayr, Martin (2017): Unsicher - Auswirkungen einer veränderten Welt. In Christian Ramsauer, Detlef Kayser, Christoph Schmitz (Eds.): Erfolgsfaktor Agilität. Chancen für Unternehmen in einem volatilen Marktumfeld. Weinheim: Wiley-VCH, pp. 33–76.

Kriegel, Hans-Peter; Kröger, Peer; Schubert, Erich; Zimek, Arthur (2009): LoOP: Local Outlier Probabilities. In David Cheung, Il-Yeol Song, Wesley Chu, Xiaohua Hu, Jimmy Lin, Jiexun Li, Zhiyong Peng (Eds.): Proceedings of the 18th ACM conference on Information and knowledge management. CIKM '09 Conference on Information and Knowledge Management. Hong Kong, China, November 02-06, 2009. New York, NY: ACM, pp. 1649–1652.

Krishnan, Krish (2013): Data warehousing in the age of big data. Amsterdam: Morgan Kaufmann - Elsevier.

BIBLIOGRAPHY

- Kubat, Miroslav (2015): An introduction to machine learning. Cham: Springer.
- Kudyba, Stephan; Hoptroff, Richard (2001): Data mining and business intelligence. A guide to productivity. Hershey, Pa.: Idea Group Pub.
- Kühnapfel, Jörg B. (2013): Vertriebscontrolling. Methoden im praktischen Einsatz. Wiesbaden: Springer Gabler.
- Kumari, Mitu (2011): Data Driven Data Mining to Domain Driven Data Mining. In *Global Journal of Computer Science and Technology* 11 (23), pp. 65–68.
- Kurgan, Lukasz A.; Musilek, Petr (2006): A survey of Knowledge Discovery and Data Mining process models. In *Knowl. Eng. Rev.* 21 (1), pp. 1–24. DOI: 10.1017/S0269888906000737.
- LaDou, Joseph (2006): Printed circuit board industry. In *International journal of hygiene and environmental health* 209 (3), pp. 211–219. DOI: 10.1016/j.ijheh.2006.02.001.
- Laney, Doug (2001): 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group Research Note. Available online at <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>, checked on 12/11/2016.
- Lanquillon, Carsten; Mallow, Hauke (2015a): Advanced Analytics mit Big Data. In Joachim Dorschel (Ed.): *Praxishandbuch Big Data. Wirtschaft - Recht - Technik*. Wiesbaden: Springer Gabler, pp. 55–89.
- Lanquillon, Carsten; Mallow, Hauke (2015b): Big Data-Lösungen. In Joachim Dorschel (Ed.): *Praxishandbuch Big Data. Wirtschaft - Recht - Technik*. Wiesbaden: Springer Gabler, pp. 263–277.
- Larose, Daniel T.; Larose, Chantal D. (2014): *Discovering knowledge in data. An introduction to data mining*. 2nd ed. Hoboken, NJ: John Wiley & Sons (Wiley Series on methods and applications in data mining).
- Larose, Daniel T.; Larose, Chantal D. (2015): *Data Mining and Predictive Analytics*. 2. Auflage. New York, NY: John Wiley & Sons (Wiley Series on Methods and Applications).
- Larson, Deanne; Chang, Victor (2016): A review and future direction of agile, business intelligence, analytics and data science. In *International Journal of Information Management* 36 (5), pp. 700–710. DOI: 10.1016/j.ijinfomgt.2016.04.013.
- Lavalle, Steve; Hopkins, Michael S.; Lesser, Eric; Shockley, Rebecca; Kruschwitz, Nina (2010): *Analytics: The new path to value. How the smartest organizations are embedding analytics to transform insights into action*. MIT Sloan Management Review and the IBM Institute for Business Value (Research Report). Available online at http://cci.uncc.edu/sites/cci.uncc.edu/files/media/pdf_files/MIT-SMR-IBM-Analytics-The-New-Path-to-Value-Fall-2010.pdf, checked on 6/27/2017.

- Lavrac, Nada; Motoda, Hiroshi; Fawcett, Tom; Holte, Robert; Langley, Pat; Adriaans, Pieter (2004): Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving. In *Machine Learning* 57 (1/2), pp. 13–34. DOI: 10.1023/B:MACH.0000035516.74817.51.
- Lee, Daebum; Kim, Juntae (2015): Pizza sales forecasting using big data analysis. In *Information (Japan)* 18 (5), pp. 1577–1584.
- Lester, Albert (2014): Project management, planning and control. Managing engineering, construction and manufacturing projects to PMI, APM and BSI standards. Sixth edition. Amsterdam: Butterworth-Heinemann.
- Li, Xin; Pan, Bing; Law, Rob; Huang, Xiankai (2017): Forecasting tourism demand with composite search index. In *Tourism Management* 59, pp. 57–66. DOI: 10.1016/j.tourman.2016.07.005.
- Li, Yan; Thomas, Manoj; Osei-Bryson, Kwaku-Muata; Levy, Jason (2016a): Problem Formulation in Knowledge Discovery via Data Analytics (KDDA) for Environmental Risk Management. In *International Journal of Environmental Research and Public Health* 13 (12), pp. 1245–1261. DOI: 10.3390/ijerph13121245.
- Li, Yan; Thomas, Manoj A.; Osei-Bryson, Kwaku-Muata (2016b): A snail shell process model for knowledge discovery via data analytics. In *Decision Support Systems* 91, pp. 1–12. DOI: 10.1016/j.dss.2016.07.003.
- Lin, Kuo-Yi; Tsai, Jeffrey J. P. (2016): A Deep Learning-Based Customer Forecasting Tool. In : 2016 IEEE Second International Conference on Multimedia Big Data (BigMM 2016). 2016 IEEE Second International Conference on Multimedia Big Data (BigMM). Taipei, Taiwan, 20-22 April, 2016. Piscataway, NJ: IEEE, pp. 198–205.
- Linoff, Gordon S.; Berry, Michael J. A. (2011): Data Mining Techniques. For Marketing, Sales, and Customer Relationship Management. Third Edition. Indianapolis, IN: Wiley Publishing.
- Little, Roderick J. A.; Rubin, Donald B. (2002): Statistical analysis with missing data. 2nd ed. Hoboken, N.J.: Wiley (Wiley series in probability and statistics).
- Liu, Chen-Hao; Wang, Yen-Wen (2012): Establish A Cluster Based Evolutionary Adaptive Weighted Fuzzy CBR for PCB Sales Forecasting. In Kae Dal Kwack, Shigeo Kawata, Soonwook Hwang, Dongsoo Han, Franz Ko (Eds.): 2012 7th International Conference on Computing and Convergence Technology (ICCIT, ICEI and ICACT). ICCCT2012. Seoul, Korea, December 3 - 5, 2012. Piscataway, NJ: IEEE, pp. 1417–1422.
- Liu, Huan; Motoda, Hiroshi (1998): Feature Selection for Knowledge Discovery and Data Mining. New York, NY: Springer Science+Business Media.
- Liu, Huan; Motoda, Hiroshi (Eds.) (2008): Computational methods of feature selection. Boca Raton: Chapman & Hall/CRC.

- Lohr, Steve (2013): The Origins of 'Big Data': An Etymological Detective Story. The New York Times. Available online at https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/?_r=0, updated on 2/1/2013, checked on 8/5/2017.
- Loshin, David (2013a): Big data analytics. From strategic planning to enterprise integration with tools, techniques, Nosql, and Graph. Waltham, MA: Morgan Kaufmann - Elsevier.
- Loshin, David (2013b): Business intelligence. The savvy manager's guide. 2nd ed. Waltham, MA: Morgan Kaufmann - Elsevier (The savvy manager's guide).
- Luczak, Dominik (2017): Agil - Erfolgsfaktor agiles Unternehmenssystem. In Christian Ramsauer, Detlef Kayser, Christoph Schmitz (Eds.): Erfolgsfaktor Agilität. Chancen für Unternehmen in einem volatilen Marktumfeld. Weinheim: Wiley-VCH, pp. 17–32.
- Makridakis, Spyros; Reschke, Hasso; Wheelwright, Steven C. (1980): Prognosetechniken für Manager. Wiesbaden: Gabler Verlag.
- Maletic, Jonathan I.; Marcus, Andrian (2010): Data Cleansing: A Prelude to Knowledge Discovery. In Oded Maimon, Lior Rokach (Eds.): Data Mining and Knowledge Discovery Handbook. New York, NY: Springer, pp. 19–32.
- Malone, Thomas W.; Crowston, Kevin (1994): The interdisciplinary study of coordination. In *ACM Computing Surveys* 26 (1), pp. 87–119. DOI: 10.1145/174666.174668.
- Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles; Byers, Angela Hung (2011): Big data: The next frontier for innovation, competition and productivity. Edited by McKinsey Global Institute. Available online at <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>, checked on 7/25/2017.
- Manyika, James; Sinclair, Jeff; Dobbs, Richard; Strube, Gernot; Rassey, Louis; Mischke, Jan et al. (2012): Manufacturing the future: The next era of global growth and innovation. McKinsey & Company. Available online at <https://www.mckinsey.com/business-functions/operations/our-insights/the-future-of-manufacturing>, checked on 7/4/2017.
- Marban, Oscar; Mariscal, Gonzalo; Menasalvas, Ernestina; Segovia, Javier (2007): An Engineering Approach to Data Mining Projects. In Hujun Yin, Peter Tino, Emilio Corchado, Will Byrne, Xin Yao (Eds.): Intelligent data engineering and automated learning - IDEAL 2007. Proceedings of the 8th international conference, vol. 4881. 8th International Conference on Intelligent data engineering and automated learning - IDEAL 2007. Birmingham, UK, December 16-19, 2007. IDEAL <8, 2007, Birmingham>. Berlin: Springer (Lecture Notes in Computer Science, Vol. 4881), pp. 578–588.
- Marban, Oscar; Mariscal, Gonzalo; Segovi, Javier (2009a): A Data Mining & Knowledge Discovery Process Model. In Julio Ponce, Adem Karahoca (Eds.): Data Mining and Knowledge Discovery in Real Life Applications. Vienna: I-Tech, pp. 1–16.

- Marban, Oscar; Segovia, Javier; Menasalvas, Ernestina; Fernández-Baizán, Covadonga (2009b): Toward data mining engineering: A software engineering approach. In *Information Systems* 34 (1), pp. 87–107. DOI: 10.1016/j.is.2008.04.003.
- March, Salvatore T.; Smith, Gerald F. (1995): Design and natural science research on information technology. In *Decision Support Systems* 15 (4), pp. 251–266. DOI: 10.1016/0167-9236(94)00041-2.
- Marin-Ortega, Pablo Michel; Dmitriyev, Viktor; Abilov, Marat; Gomez, Jorge Marx (2014): ELTA. New Approach in Designing Business Intelligence Solutions in Era of Big Data. In *Procedia Technology* 16, pp. 667–674. DOI: 10.1016/j.protcy.2014.10.015.
- Mariscal, Gonzalo; Marban, Oscar; Fernandez, Covadonga (2010): A survey of data mining and knowledge discovery process models and methodologies. In *Knowl. Eng. Rev.* 25 (2), pp. 137–166. DOI: 10.1017/S0269888910000032.
- Marr, Bernard (2015): Big data. Using smart big data, analytics and metrics to make better decisions and improve performance. Chichester, West Sussex, United Kingdom: Wiley.
- Martinelli, Russ J.; Milosevic, Dragan Z. (2016): Project management toolbox. Second edition. Hoboken, NJ: John Wiley & Sons.
- Martins, Sebastian; Pesado, Patricia; Garcia-Martinez, Ramon (2016a): Intelligent Systems in Modeling Phase of Information Mining Development Process. In Hamido Fujita, Moonis Ali, Ali Selamat, Jun Sasaki, Masaki Kurematsu (Eds.): Trends in Applied Knowledge-Based Systems and Data Science. 29th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2016, Morioka, Japan, August 2-4, 2016, Proceedings. Cham: Springer International Publishing (SpringerLink : Bücher, 9799), pp. 3–15.
- Martins, Sebastian; Pesado, Patricia; García-Martínez, Ramón (2016b): Information Mining Projects Management Process. In KSI Research Inc. and Knowledge Systems Institute Graduate School, USA (Ed.): SEKE 2016. Proceedings of the Twenty-Eighth International Conference on Proceedings of the Twenty-Eighth International Conference on Software Engineering & Knowledge Engineering. The 28th International Conference on Software Engineering and Knowledge Engineering. Redwood City, CA, July 1-3, 2016. Pittsburgh, PA: KSI Research Inc. and Knowledge Systems Institute Graduate School (International Conferences on Software Engineering and Knowledge Engineering), pp. 504–509.
- Mastin, Luke (2009): The Universe by numbers. Available online at <http://www.physicsoftheuniverse.com/numbers.html>, checked on 1/16/2016.
- Mayer-Schönberger, V.; Cukier, K. (2013): Big Data. A Revolution that Will Transform how We Live, Work, and Think. New York, NY: Houghton Mifflin Harcourt. Available online at <https://books.google.ca/books?id=uy4lh-WEhhIC>.

BIBLIOGRAPHY

McAfee, Andrew; Brynjolfsson, Erik (2012): Big Data: The Management Revolution. In *Harvard Business Review* 90 (10), pp. 60–68.

McFadzean, Elspeth (1998): The Creativity Continuum. Towards a Classification of Creative Problem Solving Techniques. In *Creativity and Innovation Management* 7 (3), pp. 131–139. DOI: 10.1111/1467-8691.00101.

McKinsey & Company (2010): Welcome to the volatile world. Challenges for the German economy emerging from fundamental market changes. Available online at https://www.mckinsey.com/~media/mckinsey/global%20themes/europe/after%20the%20crisis%20refining%20germanys%20economic%20model/welcome_to_the_volatile_world.ashx, checked on 7/4/2017.

McNair, Douglas S. (2015): Enhancing Nursing Staffing Forecasting With Safety Stock Over Lead Time Modeling. In *Nursing administration quarterly* 39 (4), pp. 291–296. DOI: 10.1097/NAQ.0000000000000124.

Mehanna, Walid; Tatzel, Jan; Vogel, Philipp (2016): Business Analytics im Controlling - Fünf Anwendungsfelder. In *Controlling* 28 (8-9), pp. 502–508.

Mendes, Armando B.; Cavique, Luís; Santos, Jorge M.A. (2012): Data Mining Process Models. A Roadmap for Knowledge Discovery. In Luiz Moutinho, Kun-Huang Huarng (Eds.): Quantitative modelling in marketing and management. Singapore: World Scientific, pp. 405–433.

Mertens, Peter; Rässler, Susanne (Eds.) (2012): Prognoserechnung. 7th ed. Berlin Heidelberg: Springer Verlag.

Miao, Jianyu; Niu, Lingfeng (2016): A Survey on Feature Selection. In *Procedia Computer Science* 91, pp. 919–926. DOI: 10.1016/j.procs.2016.07.111.

Microsoft (2017): Was ist Azure? Available online at https://azure.microsoft.com/de-de/overview/what-is-azure/?&wt.mc_id=AID529462_SEM_, checked on 23-07.2017.

Milosevic, Dragan Z. (2003): Project management toolbox. Tools and techniques for the practicing project manager. Hoboken, NJ: J. Wiley & Sons.

Minelli, Michael; Chambers, Michele; Dhiraj, Ambiga (2013): Big data, big analytics. Emerging business intelligence and analytic trends for today's businesses. Hoboken, New Jersey: John Wiley & Sons, Inc.

Ming, Jian; Hu, Nailian; Sun, Jinhai (2016): Study of iron concentrate price forecasting models based on data mining. In : Proceedings of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA 2016). 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). Chengdu, China, 5/7/2016 - 7/7/2016. Piscataway, NJ: IEEE, pp. 140–145.

-
- Mitchell, Tom M. (1997): *Machine Learning*. New York: McGraw-Hill (McGraw-Hill series in computer science).
- Mittag, Hans-Joachim (2016): *Statistik. Eine Einführung mit interaktiven Elementen*. 4th edition. Berlin: Springer Spektrum (Springer-Lehrbuch).
- Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012): *Foundations of machine learning*. Cambridge, MA: MIT Press (Adaptive computation and machine learning series).
- Möller, Klaus; Federmann, Frank; Svenja Pi; Knezevic, Michael (2016): Predictive Analytics zur kurzfristigen Umsatzprognose. Entwicklung eines Prognosemodells auf Basis von Auftragseingängen bei der Infineon Technologies AG. In *Controlling* 28 (8-9), pp. 509–518. DOI: 10.15358/0935-0381-2016-8-9-509.
- Molteni, L.; Ponce De Leon, J. (2016): Forecasting with twitter data. An application to USA TV series audience. In *International Journal of Design & Nature and Ecodynamics* 11 (3), pp. 220–229. DOI: 10.2495/DNE-V11-N3-220-229.
- Muller, Louis; Hart, Mike (2016): Updating Business Intelligence and Analytics Maturity Models for New Developments. In Shaofeng Liu, Boris Delibasic, Festus Oderanti (Eds.): *Decision Support Systems VI - Addressing Sustainability and Societal Challenges*. 2nd International Conference (ICDSSST 2016). Plymouth, UK, May 23-25, 2016. Cham: Springer (Lecture Notes in Business Information Processing, 250), pp. 137–151.
- Müller, Roland M.; Lenz, Hans-Joachim (2013): *Business Intelligence*. Berlin: Springer Vieweg (eXamen.press).
- Murphy, Kevin P. (2012): *Machine learning. A probabilistic perspective*. Cambridge, MA: MIT Press.
- Nascimento, Givanildo Santana do; Oliveira, Adicinea Aparecida de (2012): An Agile Knowledge Discovery in Databases Software Process. In Yang Xiang, Mukaddim Pathan, Xiaohui Tao, Hua Wang (Eds.): *Data and Knowledge Engineering*. Third International Conference, ICDKE 2012, Wuyishan, Fujian, China, November 21-23, 2012. Proceedings, vol. 7696. Third International Conference, ICDKE 2012. Wuyishan, Fujian, China, November 21-23. Berlin: Springer (Lecture Notes in Computer Science, 7696), pp. 56–64.
- Nemati, Hamid R.; Barko, Christopher D. (2003): Key factors for achieving organizational data-mining success. In *Industrial Management & Data Systems* 103 (4), pp. 282–292. DOI: 10.1108/02635570310470692.
- Niño, Mikel; Blanco, José Miguel; Illarramendi, Arantza (2015): Business Understanding, Challenges and Issues of Big Data Analytics for the Servitization of a Capital Equipment Manufacturer. In Howard Ho, Beng Chin Ooi, Mohammed J. Zaki, Xiaohua Hu, Laura Haas, Vipin Kumar et al. (Eds.): *2015 IEEE International Conference on Big Data*. 2015 IEEE

BIBLIOGRAPHY

International Conference on Big Data. Santa Clara, CA, USA, Oct 29-Nov 01, 2015: IEEE, pp. 1368–1377.

Nisbet, Robert; Elder, John; Miner, Gary (2009): Handbook of statistical analysis and data mining applications. Amsterdam, London: Elsevier/Academic Press.

NIST Big Data Public Working Group (2015): NIST Big Data Interoperability Framework. Volume 1, Definitions: National Institute of Standards and Technology.

Nita, Souichirou (2015): Application of big data technology in support of food manufacturers' commodity demand forecasting. In *NEC Technical Journal* 10 (1), pp. 90–93.

OECD (2017a): About the OECD. Available online at <http://www.oecd.org/about/>, checked on 5/15/2017.

OECD (2017b): Catalogue of OECD databases. Available online at <https://data.oecd.org/searchresults/?r=+f/type/datasets>, checked on 5/15/2017.

Ohlhorst, Frank (2013): Big data analytics. Turning big data into big money. Hoboken, New Jersey: Wiley (Wiley & SAS business series).

Omri, Fouad (2015): Big Data-Analysen: Anwendungsszenarien und Trends. In Joachim Dorschel (Ed.): *Praxishandbuch Big Data. Wirtschaft - Recht - Technik*. Wiesbaden: Springer Gabler, pp. 104–112.

Oprean, Cristina (2011): Towards user assistance in Data Mining. Master Thesis. Telecom Bretagne. Available online at <https://dumas.ccsd.cnrs.fr/dumas-00636764>, checked on 7/3/2017.

Oracle (2013): Oracle: Big Data for the Enterprise. Oracle White Paper. Available online at <http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf>, checked on 4/15/2017.

Otsuka, Noriaki; Nishina, Mitsuki; Higashihara, Katsunori; Umezu, Keisuke; Nagai, Yoshi; Motohashi, Yousuke et al. (2015): Demand Forecasting Solution Contributing to Components Inventory Repair Optimization. In *NEC Technical Journal* 10 (1), pp. 79–82.

Ou, L.; Qin, Z.; Yin, H.; Li, K. (2016): Security and Privacy in Big Data. In Rajkumar Buyya, Rodrigo N. Calheiros, Amir Vahid Dastjerdi (Eds.): *Big data. Principles and paradigms*. Cambridge, MA: Morgan Kaufmann - Elsevier (Science Direct e-books), pp. 285–308.

Özgür, Arzucan; Özgür, Levent; Güngör, Tunga (2005): Text Categorization with Class-Based and Corpus-Based Keyword Selection. In pInar Yolum, Tunga Güngör, Fikret Gürgen, Can Özturan (Eds.): *Computer and information sciences - ISCIS 2005*. P20th International Symposium, Istanbul, Turkey, October 26–28, 2005, Proceedings, vol. 3733. 20th International Symposium on Computer and Information. Istanbul, Turkey, 26–28 October, 2005. Berlin: Springer (Lecture Notes in Computer Science, 3733), pp. 606–615.

- Palanimalai, Shanmugasundaram; Paramasivam, Ilango (2016): Big data analytics bring new insights and higher business value - an experiment carried out to divulge sales forecasting solutions. In *International Journal of Advanced Intelligence Paradigms* 8 (2), pp. 207–218. DOI: 10.1504/IJAIP.2016.075728.
- Pant, Prashant (2009): Essential Components of a Successful BI Strategy. Information Management. Available online at <https://www.information-management.com/news/essential-components-of-a-successful-bi-strategy>, updated on 7/28/2009, checked on 4/10/2016.
- Pawlak, Zdzislaw (1981): Information systems theoretical foundations. In *Information Systems* 6 (3), pp. 205–218. DOI: 10.1016/0306-4379(81)90023-5.
- PCB Wizards (2006): Printed Circuit Applications. Available online at <http://www.pcbwizards.com/applications.htm>, updated on 12/31/2006, checked on 3/15/2017.
- Peffer, Ken; Tuunanen, Tuure; Rothenberger, Marcus A.; Chatterjee, Samir (2007): A Design Science Research Methodology for Information Systems Research. In *Journal of Management Information Systems* 24 (3), pp. 45–77. DOI: 10.2753/MIS0742-1222240302.
- Peng, Yi; Kou, Gang (2008): A Domain Knowledge-Driven Framework for Multi-Criteria Optimization-Based Data Mining Methods. In Jinhwa Kim, Durson Delen, Jinsoo Park, Franz Ko, Yun Ji Na (Eds.): *Networked Computing and Advanced Information Management*, 2008. NCM '08. Fourth International Conference on. 2008 Fourth International Conference on Networked Computing and Advanced Information Management (NCM). Gyeongju, South Korea, 2/9/2008 - 4/9/2008, pp. 46–49.
- Pestman, Wiebe R. (2009): *Mathematical Statistics*. Berlin: Walter de Gruyter.
- Phillips-Wren, Gloria; Iyer, Lakshmi S.; Kulkarni, Uday; and Ariyachandra, Thilini (2015): Business Analytics in the Context of Big Data: A Roadmap for Research. In *Communications of the Association for Information Systems* 37, Article 23, checked on 6/27/2017.
- Piatetsky, Gregory (2014): CRISP-DM, still the top methodology for analytics, data mining, or data science projects. KDnuggets. Available online at <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>, updated on October 2014, checked on 6/21/2017.
- Ponniah, Paulraj (2010): *Data Warehousing. Fundamentals for IT Professionals*. Second edition. New York, NY: Wiley.
- Press, Gil (2013): A Very Short History Of Big Data. Forbes. Available online at <https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#62a3f0b765a1>, updated on 5/9/2013, checked on 8/12/2017.
- Pressman, Roger S. (2010): *Software engineering. A practitioner's approach*. 7th ed. New York, NY: McGraw-Hill Higher Education.

BIBLIOGRAPHY

Priebe, Torsten; Markus, Stefan (2015): Business Information Modeling: A Methodology for Data-Intensive Projects, Data Science and Big Data Governance. In Howard Ho, Beng Chin Ooi, Mohammed J. Zaki, Xiaohua Hu, Laura Haas, Vipin Kumar et al. (Eds.): 2015 IEEE International Conference on Big Data. 2015 IEEE International Conference on Big Data. Santa Clara, CA, USA, Oct 29-Nov 01, 2015: IEEE, pp. 2056–2065.

Provost, Foster; Fawcett, Tom (2013a): Data Science and its Relationship to Big Data and Data-Driven Decision Making. In *Big data* 1 (1), pp. 51–59. DOI: 10.1089/big.2013.1508.

Provost, Foster; Fawcett, Tom (2013b): Data Science for Business. What You Need to Know about Data Mining and Data-Analytic Thinking. Sebastopol, CA: O'Reilly Media.

Pyle, Dorian (2003): Business modeling and data mining. San Francisco, CA: Morgan Kaufmann Publishers.

Qiu, Xumeng; He, Ge; Ji, Xu (2016): Cloud manufacturing model in polymer material industry. In *International Journal of Advanced Manufacturing Technology* 84 (1-4), pp. 239–248. DOI: 10.1007/s00170-015-7580-6.

Rabitsch, Christian (2016): Methodology for Implementing Agility in Manufacturing Companies. Dissertation. Technische Universität Graz. Institute of Industrial Management and Innovation Research.

Rabitsch, Christian (2017): Strategisch - Das richtige Maß an Agilität. In Christian Ramsauer, Detlef Kayser, Christoph Schmitz (Eds.): Erfolgsfaktor Agilität. Chancen für Unternehmen in einem volatilen Marktumfeld. Weinheim: Wiley-VCH, pp. 127–159.

Rabitsch, Christian; Ramsauer, Christian (2015): Towards a management approach for implementing agility in the manufacturing industry. MOTSP 2015 - International Conference Management of Technology - Step to Sustainable Production - Brela, Croatia - 06/10/15-06/12/15. In *Proceedings International Conference Management of Technology – Step to Sustainable Production*, pp. 1–8.

Rabitsch, Christian; Schurig, Matthias; Ramsauer, Christian (2015): Operationalisierung der Agilität. Agilitätsdimensionen und Stellgrößen. In *Industrie 4.0 Management* 31 (4), pp. 48–52.

Raghupathi, Wullianallur; Raghupathi, Viju (2014): Big Data Analytics. Architectures, Implementation Methodology, and Tools. In Stephan Kudyba (Ed.): Big Data, Mining, and Analytics. Components of Strategic Decision Making. Boca Raton: CRC Press - Taylor & Francis Group, pp. 49–70.

Rahm, Erhard; Do, Hong Hai (2000): Data Cleaning: Problems and Current Approaches. In *Bulletin of the Technical Committee on Data Engineering* 23 (4), pp. 3–13.

Rajpurohit, Anmol (2013): Big data for business managers — Bridging the gap between potential and value. In Xiaohua Hu (Ed.): IEEE International Conference on Big Data, 2013. 6-9 Oct. 2013, Silicon Valley, California, USA ; proceedings. 2013 IEEE International

Conference on Big Data. Silicon Valley, CA, USA, 6/10/2013 - 9/10/2013. IEEE International Conference on Big Data; Big Data Conference. Piscataway, NJ: IEEE, pp. 29–31.

Ramsauer, Christian (2013): Industrie 4.0 – Die Produktion der Zukunft. In *WINGbusiness* (3), pp. 6–12.

Ramsauer, Christian; Kayser, Detlef; Schmitz, Christoph (Eds.) (2017): Erfolgsfaktor Agilität. Chancen für Unternehmen in einem volatilen Marktumfeld. Weinheim: Wiley-VCH.

Ransbotham, Sam; Kiron, David; Prentice, Pamela Kirk (2015): Minding the Analytics Gap. In *MIT Sloan Management Review* 56 (3), pp. 63–68.

Rao, B. B. Prahlada; Saluja, Payal; Sharma, Neetu; Mittal, Ankit; Sharma, Shivay Veer (2012): Cloud computing for Internet of Things & sensing based applications. In : 2012 Sixth International Conference on Sensing Technology. ICST 2012. Kolkata, India, 18 - 21 Dec. 2012. Piscataway, NJ: IEEE, pp. 374–380.

Rasli, Amran (2006): Data analysis and interpretation. A handbook for postgraduate social scientists. Skudai: Universiti Teknologi Malaysia.

Reeves, Laura (2009): A Manager's Guide to Data Warehousing. Indianapolis, IN: John Wiley & Sons.

Reinhart, Gunther; Dürrschmidt, Stephan; Hirschberg, Arnd; Selke, Carsten (1999): Reaktionsfähigkeit für Unternehmen. Eine Antwort auf turbulente Märkte. In *ZWF* 94 (1-2), pp. 21–24.

Ren, J.; Yusuf, Y. Y.; Burns, N. D. (2003): The effects of agile attributes on competitive priorities. A neural network approach. In *Integrated Manufacturing Systems* 14 (6), pp. 489–497. DOI: 10.1108/09576060310491351.

ReportLinker (2017): Growth Opportunities in the Global Printed Circuit Board Market. PR Newswire. Available online at <http://www.prnewswire.com/news-releases/growth-opportunities-in-the-global-printed-circuit-board-market-300433799.html>, updated on 4/3/2017, checked on 5/12/2017.

Rey, Tim; Kordon, Arthur; Wells, Chip (2012): Applied Data Minig for Forecasting Using SAS. Cary, NC: SAS Press.

Rider, Fremont (1944): The Scholar and the Future of the Research Library. A Problem and Its Solution. New York, NY: Hadham Press.

Ridge, Enda (2015): Guerrilla Analytics. A Practical Approach to Working with Data. Waltham, MA: Morgan Kaufmann - Elsevier.

Rocha, Anderson; Papa, Joao Paulo; Meira, Luis A. A. (2012): How far do we get using machine learning black-boxes? In *International Journal of Pattern Recognition and Artificial Intelligence* 26 (02), p. 1261001. DOI: 10.1142/S0218001412610010.

BIBLIOGRAPHY

Rogalewicz, Michal; Sika, Robert (2016): Methodologies of Knowledge Discovery from Data and Data Mining Methods in Mechanical Engineering. In *Management and Production Engineering Review* 7 (4), pp. 97–108. DOI: 10.1515/mpcr-2016-0040.

Rohanizadeha, Seyyed Soroush; Moghadama, Mohammad Bameni (2009): A proposed data mining methodology and its application to industrial procedures. In *Journal of Industrial Engineering* 4, pp. 37–50.

Roiger, Richard J. (2017): *Data Mining. A Tutorial-Based Primer. Second Edition.* Boca Raton, FL: CRC Press (Chapman & Hall/CRC data mining and knowledge discovery series).

Roman, Julio Villena (2016): CRISP-DM: The methodology to put some order into Data Science projects. [singular. data&analytics.](https://data.singular.team/en/art/40/crisp-dm-the-methodology-to-put-some-order-into-data-science-projects) Available online at <https://data.singular.team/en/art/40/crisp-dm-the-methodology-to-put-some-order-into-data-science-projects>, updated on 8/8/2016, checked on 7/4/2017.

Ruan, Guangchen; Zhang, Hui (2017): Closed-loop Big Data Analysis with Visualization and Scalable Computing. In *Big Data Research* 8, pp. 12–26. DOI: 10.1016/j.bdr.2017.01.002.

Runkler, Thomas A. (2010): *Data Mining. Methoden und Algorithmen intelligenter Datenanalyse.* Wiesbaden: Vieweg+Teubner (Studium).

Runkler, Thomas A. (2016): *Data Analytics. Models and Algorithms for Intelligent Data Analysis.* 2nd Edition. Wiesbaden: Springer Vieweg.

Russom, Philip (2011): *Big Data Analytics. TDWI Best Practices Report.* Edited by TDWI Research. Available online at <https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf>, checked on 6/16/2017.

Saltz, Jeffrey; Shamshurin, Ivan; Connors, Colin (2017a): A Framework for Describing Big Data Projects. In Witold Abramowicz, Rainer Alt, Bogdan Franczyk (Eds.): *Business Information Systems Workshops. Revised Papers. BIS 2016 International Workshops.* Leipzig, Germany, July 6-8, 2016. Cham: Springer International Publishing (Lecture Notes in Business Information Processing, 263), pp. 183–195.

Saltz, Jeffrey; Shamshurin, Ivan; Crowston, Kevin (2017b): Comparing Data Science Project Management Methodologies via a Controlled Experiment. In : *Proceedings of the 50th Hawaii International Conference on System Sciences. HICSS 2017 : 50th annual Hawaii International Conference on System Sciences.* Waikoloa, HI, USA, Jan 4, 2017 - Jan 7, 2017 (Proceedings of the Annual Hawaii International Conference on System Sciences), pp. 1013–1022.

Saltz, Jeffrey S. (2015): The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness. In Howard Ho, Beng Chin Ooi, Mohammed J. Zaki, Xiaohua Hu, Laura Haas, Vipin Kumar et al. (Eds.): *2015 IEEE International Conference on Big Data.* 2015 IEEE International Conference on Big Data. Santa Clara, CA, USA, Oct 29-Nov 01, 2015: IEEE, pp. 2066–2071.

- Saltz, Jeffrey S.; Shamshurin, Ivan (2015): Exploring the Process of Doing Data Science Via an Ethnographic Study of a Media Advertising Company. In Howard Ho, Beng Chin Ooi, Mohammed J. Zaki, Xiaohua Hu, Laura Haas, Vipin Kumar et al. (Eds.): 2015 IEEE International Conference on Big Data. 2015 IEEE International Conference on Big Data. Santa Clara, CA, USA, Oct 29-Nov 01, 2015: IEEE, pp. 2098–2105.
- Saltz, Jeffrey S.; Shamshurin, Ivan (2016): Big Data Team Process Methodologies: A Literature Review and the Identification of Key Factors for a Project's Success. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data. Washington D.C., USA, Dec 05-Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 2872–2879.
- Saltz, Jeffrey S.; Yilmazel, Sibel; Yilmazel, Ozgur (2016): Not All Software Engineers Can Become Good Data Engineers. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data. Washington D.C., USA, Dec 05-Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 2896–2901.
- Sa-ngasoongsong, Akkarapol; Bukkapatnam, Satish T.S.; Kim, Jaebeom; Iyer, Parameshwaran S.; Suresh, R. P. (2012): Multi-step sales forecasting in automotive industry based on structural relationship identification. In *International Journal of Production Economics* 140 (2), pp. 875–887. DOI: 10.1016/j.ijpe.2012.07.009.
- SAS Institute (2017): Notation for ARIMA Models. SAS/ETS(R) 9.3 User's Guide. Available online at https://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_tffordet_sect016.htm, checked on 3/23/2017.
- SAS Institute Inc. (2017): SAS® Enterprise Miner™ fact sheet. Available online at https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf, checked on 7/1/2017.
- Schmarzo, Bill (2016): Big data MBA. Driving business strategies with data science. Indianapolis, IN: John Wiley & Sons.
- Schmarzo, William D. (2013): Big data. Understanding how data powers big business. Hoboken, NJ: Wiley.
- Schurig, Matthias (2016): Methodology to evaluate the agility of a production network using a stress test approach. Dissertation. Technische Universität Graz. Institute of Industrial Management and Innovation Research.
- Schurig, Matthias (2017): Definiert - Was man unter Agilität versteht. In Christian Ramsauer, Detlef Kayser, Christoph Schmitz (Eds.): Erfolgsfaktor Agilität. Chancen für Unternehmen in einem volatilen Marktumfeld. Weinheim: Wiley-VCH, pp. 77–99.

Schurig, Matthias; Rabitsch, Christian; Ramsauer, Christian (2014): Agile Produktion. Ein Produktionskonzept für volatile Zeiten. In *ZWF* 109 (12), pp. 956–959.

Schutt, Rachel; O'Neil, Cathy (2013): Doing data science. Straight talk from the frontline. First edition. Sebastopol, CA: O'Reilly.

scikit-learn (2016): Precision-Recall. Available online at http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html, checked on 9/12/2016.

scikit-learn (2017a): Decision Tree Classifier. Available online at <http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>, checked on 8/4/2017.

scikit-learn (2017b): Grid Search: Searching for estimator parameters. Available online at http://scikit-learn.org/0.17/modules/grid_search.html#grid-search, checked on 6/12/2017.

scikit-learn (2017c): Parameters of the RBF Kernel. Available online at <http://scikit-learn.org/stable/modules/svm.html#parameters-of-the-rbf-kernel>, checked on 8/4/2017.

scikit-learn (2017d): scikit-learn: Machine Learning in Python. Available online at <http://scikit-learn.org/stable/index.html#>, checked on 6/12/2017.

Severtson, Brad; Ericson, Gary; Franks, Larry; Grondlund, C. J. (2017): Team Data Science Process lifecycle. Microsoft. Available online at <https://docs.microsoft.com/en-us/azure/machine-learning/data-science-process-overview>, updated on 2/8/2017, checked on 7/8/2017.

Shahapurkar, Som (2016): Crossing the Chasm: Deploying Machine Learning Analytics in Dynamic Real World Scenarios. Dissertation. Arizona State University. ASU Graduate College. Available online at <https://repository.asu.edu/items/40710>, checked on 7/8/2017.

Sharifi, H.; Zhang, Z. (1999): A methodology for achieving agility in manufacturing organisations: An introduction. In *International Journal of Production Economics* 62 (1-2), pp. 7–22.

Sharifi, H.; Zhang, Z. (2001): Agile manufacturing in practice - Application of a methodology. In *International Journal of Operations & Production Management* 21 (5/6), pp. 772–794.

Sharma, Sumana (2008): An Integrated Knowledge Discovery and Data Mining Process Model. Dissertation. Virginia Commonwealth University. Available online at <http://scholarscompass.vcu.edu/etd/1615/>, checked on 6/26/2017.

Sharma, Sumana (2015): Overview of Knowledge Discovery and Data Mining Process Models. In Corlane Barclay, Kweku-Muata Osei-Bryson (Eds.): Knowledge discovery process and methods to enhance organizational performance. Boca Raton, FL: CRC Press, pp. 11–23.

Sharma, Sumana; Osei-Bryson, Kweku-Muata (2009): Role of Human Intelligence in Domain Driven Data Mining. In Longbing Cao, Philip S. Yu, Chengqi Zhang, Huaifeng Zhang (Eds.):

Data Mining for Business Applications. New York, NY: Springer Science+Business Media, pp. 53–61.

Sharma, Sumana; Osei-Bryson, Kweku-Muata (2010): Toward an integrated knowledge discovery and data mining process model. In *The Knowledge Engineering Review* 25 (01), pp. 49–67. DOI: 10.1017/S0269888909990361.

Sharma, Sumana; Osei-Bryson, Kweku-Muata (2015a): A novel Method for Formulating the Business objectives of Data Mining Projects. In Corlane Barclay, Kweku-Muata Osei-Bryson (Eds.): Knowledge discovery process and methods to enhance organizational performance. Boca Raton, FL: CRC Press, pp. 55–81.

Sharma, Sumana; Osei-Bryson, Kweku-Muata (2015b): An integrated Knowledge Discovery and Data Mining Process Model. In Corlane Barclay, Kweku-Muata Osei-Bryson (Eds.): Knowledge discovery process and methods to enhance organizational performance. Boca Raton, FL: CRC Press, pp. 25–52.

Sharma, Sumana; Osei-Bryson, Kweku-Muata; Kasper, George M. (2012): Evaluation of an integrated Knowledge Discovery and Data Mining process model. In *Expert Systems with Applications* 39 (13), pp. 11335–11348. DOI: 10.1016/j.eswa.2012.02.044.

Shearer, Colin (2000): The CRISP-DM Model: The New Blueprint for Data Mining. In *Journal of Data Warehousing* 5 (4), pp. 13–22.

Sheikh, Nauman Mansoor (2013): Implementing analytics. A blueprint for design, development, and adoption. Amsterdam: Elsevier.

Shevlyakov, Georgy L.; Oja, Hannu (2016): Robust correlation. Theory and applications. Chichester: John Wiley & Sons (Wiley series in probability and statistics).

Shim, Jung P.; French, Aaron M.; Guo, Chengqi; Jablonski, Joey (2015): Big Data and Analytics: Issues, Solutions, and ROI. In *Communications of the Association for Information Systems* 37, 797-810, checked on 6/27/2017.

Shin, Sangmun; Le Yang; Park, Kyungjin; Choi, Yongsun (2009): Robust Data Mining: An Integrated Approach. In Julio Ponce, Adem Karahoca (Eds.): Data Mining and Knowledge Discovery in Real Life Applications. Vienna: I-Tech, pp. 59–74.

Shi-Nash, Amy; Hardoon, David R. (2017): Data Analytics and Predictive Analytics in the Era of Big Data. In Hwaiyu Geng (Ed.): Internet of things and data analytics handbook. Hoboken, NJ: John Wiley & Sons, Inc, pp. 329–345.

Sicular, Svetlana (2012): No Data Scientist Is an Island in the Ocean of Big Data. Gartner. Available online at <https://www.gartner.com/doc/2020415/data-scientist-island-ocean-big>, updated on 5/18/2012, checked on 5/22/2017.

BIBLIOGRAPHY

Silva, C.; Ribeiro, B. (2003): The importance of stop word removal on recall values in text categorization. In : Proceedings of the International Joint Conference on Neural Networks, 2003. 2003 International Joint Conference on Neural Networks. Portland, OR, USA, 20-24 July 2003: IEEE (Volume 3), pp. 1661–1666.

Sim, Jaesung (2003): Critical success factors in data mining projects. Dissertation. University of North Texas. Available online at <https://digital.library.unt.edu/ark%3A/67531/metadc4293/>, checked on 7/5/2017.

Simon, Herbert A. (1954): Spurious Correlation: A Causal Interpretation. In *Journal of the American Statistical Association* 49 (267), pp. 467–479.

Simon, Herbert Alexander (1996): The sciences of the artificial. 3. ed. Cambridge, MA: MIT Press.

Singh, Pritpal (2015): Big Data Time Series Forecasting Model: A Fuzzy-Neuro Hybridize Approach. In Satchidananda Dehuri, D. P. Achariya, Sugata Sanyal (Eds.): Computational intelligence for big data analysis. Frontier advances and applications, vol. 19. Cham: Springer (Adaptation, Learning, and Optimization, 19), pp. 55–72.

Singh, Saurabh; Solanki, A. K.; Trivedi, Nitin; Kumar, Manoj (2011): Data mining challenges and knowledge discovery in real life applications. In S. Arumuga Perumal (Ed.): ICECT 2011. 2011 3rd International Conference on Electronics Computer Technology. 2011 3rd International Conference on Electronics Computer Technology (ICECT). Kanyakumari, India, 8/4/2011 - 10/4/2011. Piscataway, NJ: IEEE, pp. 279–283.

Siriweera, T.H.A.S.; Paik, Incheon; Kumara, Banage T.G.S.; Koswatta, K.R.C. (2015): Intelligent Big Data Analysis Architecture Based on Automatic Service Composition. In Barbara Carminati (Ed.): 2015 IEEE International Congress on Big Data (BigData Congress). 2015 IEEE International Congress on Big Data (BigData Congress). New York City, NY, USA, June 27, 2015 - July 2, 2015. Piscataway, NJ: IEEE, pp. 276–280.

Sivarajah, Uthayasankar; Kamal, Muhammad Mustafa; Irani, Zahir; Weerakkody, Vishanth (2017): Critical analysis of Big Data challenges and analytical methods. In *Journal of Business Research* 70, pp. 263–286. DOI: 10.1016/j.jbusres.2016.08.001.

Soares, Sunil (2011): The IBM data governance unified process. Driving business value with IBM software and best practices. Ketchum, ID: McPress.

Sokolova, Marina; Lapalme, Guy (2009): A systematic analysis of performance measures for classification tasks. In *Information Processing & Management* 45 (4), pp. 427–437. DOI: 10.1016/j.ipm.2009.03.002.

Solarte, Jose (2002): A Proposed Data Mining Methodology and its Application to Industrial Engineering. Master Thesis. University of Tennessee, Knoxville. Available online at http://trace.tennessee.edu/utk_gradthes/2172/, checked on 6/26/2017.

- Sparks, Ross; Ickowicz, Adrien; Lenz, Hans J. (2016): An Insight on Big Data Analytics. In Nathalie Japkowicz, Jerzy Stefanowski (Eds.): *Big Data Analysis. New Algorithms for a New Society*. 1st ed. 2016. Cham: Springer International Publishing (Studies in Big Data, 16), pp. 33–48.
- Spath, Dieter; Ganschar, Oliver; Gerlach, Stefan; Hämmerle, Moritz; Krause, Tobias; Schlund, Sebastian (2013): *Produktionsarbeit der Zukunft – Industrie 4.0*. Stuttgart: Fraunhofer Verlag.
- StatSoft, Inc. (2013): *Electronic Statistics Textbook*. Statsoft. Tulsa, OK. Available online at <http://www.statsoft.com/textbook/>, checked on 7/3/2017.
- Suarez, Fernando F.; Cusumano, Michael A.; Fine, Charles F. (1996): An empirical study of manufacturing flexibility in printed circuit board assembly. In *Operations Research* 44 (1), pp. 223–240.
- Sun, Yanmin; Wong, Andrew K. C.; Kamel, Mohamed S. (2009): Classification of imbalanced data: A review. In *International Journal of Pattern Recognition and Artificial Intelligence* 23 (04), pp. 687–719. DOI: 10.1142/S0218001409007326.
- Sun, Zhaohao; Pambel, Francisca; Wang, Fangwei (2015): Incorporating Big Data Analytics into Enterprise Information Systems. In Ismail Khalil, Erich J. Neuhold, A. Min Tjoa, Li D. Xu, Ilsun You (Eds.): *Information and communication technology. Third IFIP TC 5/8 International Conference, ICT-EurAsia 2015, and 9th IFIP WG 8.9 Working Conference, CONFENIS 2015, held as part of WCC 2015, Daejeon, Korea, October 4-7, 2015, Proceedings. International Conference on Research and Practical Issues of Enterprise Information Systems*. Cham: Springer (Lecture Notes in Computer Science, 9357), pp. 300–309.
- Tang, S.; He, B.; Liu, H.; Lee, B.-S. (2016): Resource Management in Big Data Processing Systems. In Rajkumar Buyya, Rodrigo N. Calheiros, Amir Vahid Dastjerdi (Eds.): *Big data. Principles and paradigms*. Cambridge, MA: Morgan Kaufmann - Elsevier (Science Direct e-books), pp. 161–188.
- Tavakkoli, Amirmohammad; Rezaeenour, Jalal; Hadavandi, Esmail (2015): A Novel Forecasting Model Based on Support Vector Regression and Bat Meta-Heuristic (Bat-SVR). Case Study in Printed Circuit Board Industry. In *International Journal of Information Technology & Decision Making* 14 (1), pp. 195–215. DOI: 10.1142/S0219622014500849.
- Techopedia (2017): Data Ownership. Available online at <https://www.techopedia.com/definition/29059/data-ownership>, checked on 7/25/2017.
- Terrizzano, Ignacio; Schwarz, Peter; Roth, Mary; Colino, John E. (2015): Data Wrangling: The Challenging Journey from the Wild to the Lake. In *7th Biennial Conference on Innovative Data Systems Research (CIDR'15) January 4-7, 2015, Asilomar, California, USA*, pp. 1–9. Available online at http://cidrdb.org/cidr2015/Papers/CIDR15_Paper2.pdf, checked on 6/27/2017.

BIBLIOGRAPHY

- Theobald, Elke; Föhl, Ulrich (2015): Big Data wird zu Smart Data – Big Data in der Marktforschung. In Joachim Dorschel (Ed.): Praxishandbuch Big Data. Wirtschaft - Recht - Technik. Wiesbaden: Springer Gabler, pp. 112–123.
- Theodoridis, Sergios; Koutroumbas, Konstantinos (2009): Pattern recognition. 4. ed. Burlington, MA: Academic Press - Elsevier.
- Toni, A. de; Tonchia, S. (1998): Manufacturing flexibility. A literature review. In *International Journal of Production Research* 36 (6), pp. 1587–1617. DOI: 10.1080/002075498193183.
- Tuovinen, Lauri (2014): From machine learning to learning with machines. Remodeling the knowledge discovery process. Dissertation. University of Oulu. Available online at <http://jultika.oulu.fi/Record/isbn978-952-62-0524-3>, checked on 7/4/2017.
- Tuovinen, Lauri (2016): A conceptual model of actors and interactions for the knowledge discovery process. In Ana Fred, Jan Dietz, David Aveiro, Kecheng Liu, Jorge Bernardino, Joaquim Filipe (Eds.): IC3K 2016. Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. Porto, Portugal, November 9-11, 2016. Institute for Systems and Technologies of Information, Control and Communication; International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management; IC3K. Setúbal, Portugal: SCITEPRESS - Science and Technology Publications Lda, pp. 240–248.
- Twitter (2017a): API Rate Limits. Available online at <https://dev.twitter.com/rest/public/rate-limiting>, checked on 3/11/2017.
- Twitter (2017b): Full-Archive Search API. Available online at http://support.gnip.com/apis/search_full_archive_api/, checked on 3/12/2017.
- Two Crows Corporation (1999): Introduction to data mining and knowledge discovery. 3rd ed. Potomac, MD: Two Crows Corporation.
- Ulrich, Dave; Lake, Dale (1991): Organization capability. Creating competitive advantage. In *Academy of Management Executive* 5 (1), pp. 77–92. DOI: 10.5465/AME.1991.4274728.
- Ulwick, Anthony W. (2005): What customers want. Using outcome-driven innovation to create breakthrough products and services. New York, NY: McGraw-Hill.
- Vanauer, Martin; Bohle, Carsten; Hellingrath, Bernd (2015): Guiding the Introduction of Big Data in Organizations. A Methodology with Business- and Data-Driven Ideation and Enterprise Architecture Management-Based Implementation. In Tung X. Bui, Ralph H. Sprague, JR. (Eds.): Proceedings of the 48th Annual Hawaii International Conference on System Sciences. HICSS 2015. 2015 48th Hawaii International Conference on System Sciences (HICSS). Kauai, HI, USA, 05.01.2015 - 08.01.2015. Piscataway, NJ: IEEE, pp. 908–917.

- Vassiliadis, Panos; Simitsis, Alkis (2009): Extraction, Transformation, and Loading. In Ling Liu, M. Tamer Özsu (Eds.): *Encyclopedia of database systems*. New York, NY: Springer, pp. 1095–1101.
- Vercellis, Carlo (2009): *Business intelligence. Data mining and optimization for decision making*. Chichester: John Wiley & Sons.
- Viaene, Stijn; van den Bunder, Annabel (2011): The Secrets to Managing Business Analytics Projects. In *MIT Sloan Management Review* 53 (1), pp. 65–69.
- Villanueva, John Carl (2015): How many atoms are there in the universe? *Universe Today*. Available online at <https://www.universetoday.com/36302/atoms-in-the-universe/#gsc.tab=0>, updated on 12/24/2015, checked on 1/14/2016.
- Wamba, Samuel Fosso; Akter, Shahriar; Edwards, Andrew; Chopin, Geoffrey; Gnanzou, Denis (2015): How ‘big data’ can make big impact. Findings from a systematic review and a longitudinal case study. In *International Journal of Production Economics* 165, pp. 234–246. DOI: 10.1016/j.ijpe.2014.12.031.
- Wamba, Samuel Fosso; Gunasekaran, Angappa; Akter, Shahriar; Ren, Steven Ji-fan; Dubey, Rameshwar; Childe, Stephen J. (2017): Big data analytics and firm performance. Effects of dynamic capabilities. In *Journal of Business Research* 70, pp. 356–365. DOI: 10.1016/j.jbusres.2016.08.009.
- Wampula, Marco (2017): Verankert - Agile Organisation und Unternehmenskultur. In Christian Ramsauer, Detlef Kayser, Christoph Schmitz (Eds.): *Erfolgsfaktor Agilität. Chancen für Unternehmen in einem volatilen Marktumfeld*. Weinheim: Wiley-VCH, pp. 265–287.
- Wang, Hai; Xu, Zeshui; Fujita, Hamido; Liu, Shousheng (2016): Towards felicitous decision making. An overview on challenges and trends of Big Data. In *Information Sciences* 367-368, pp. 747–765. DOI: 10.1016/j.ins.2016.07.007.
- Wang, Yen-Wen; Liu, Chen-Hao; Fan, Chin-Yuan (2009): The Hybrid Model Development of Clustering and Back Propagation Network in Printed Circuit Board Sales Forecasting. In Janusz Kacprzyk, Been-Chian Chien, Tzung-Pei Hong (Eds.): *Opportunities and Challenges for Next-Generation Applied Intelligence*, vol. 214. Berlin: Springer (SpringerLink: Springer e-Books, 214), pp. 213–218.
- Watanabe, Takuya; Muroi, Hiroaki; Naruke, Motoki; Yono, Kyoto; Kobayashi, Gen; Yamasaki, Masanori (2016): Prediction of regional goods demand incorporating the effect of weather. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): *Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data*. Washington D.C., USA, Dec 05-Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 3785–3791.

BIBLIOGRAPHY

Wells, Chip; Rey, Tim (2015): Process and Methods for Data Mining for Forecasting. In Michael Gilliland, Len Tashman, Udo Sglavo (Eds.): Business Forecasting: Practical Problems and Solutions. Hoboken, NJ: John Wiley & Sons, pp. 120–126.

Westkämper, E. (1999): Die Wandlungsfähigkeit von Unternehmen. In *wt Werkstattstechnik online* 89 (4), pp. 131–140.

Westkämper, Engelbert (2007): Digital Manufacturing In The Global Era. In Pedro Filipe Cunha, Paul G. Maropoulos (Eds.): Digital Enterprise Technology. Boston, MA: Springer US, pp. 3–14.

White, Tom (2015): Hadoop. The definitive guide. 4th edition. Sebastopol, CA: O'Reilly Media.

Wieland, Uwe; Fischer, Marco (2013): Zur methodischen Vorbereitung von Data-Mining-Projekten unter Verwendung von CRISP-DM im Kontext diskreter Produktionsprozesse. In Henning Baars (Ed.): Workshop Business Intelligence 2013. Tagungsband des 5. Workshops "Business Intelligence" der GI-Fachgruppe Business Intelligence. 5. Workshops "Business Intelligence" der GI-Fachgruppe Business Intelligence. Freiberg, Germany, July 12, 2013: CEUR Workshop Proceedings, pp. 47–63.

Wiendahl, H.-P.; ElMaraghy, H. A.; Nyhuis, P.; Zäh, M. F.; Wiendahl, H.-H.; Duffie, N.; Brieke, M. (2007): Changeable Manufacturing - Classification, Design and Operation. In *CIRP Annals* 56 (2), pp. 783–809. DOI: 10.1016/j.cirp.2007.10.003.

Wierse, Andreas; Riedel, Till (2017): Smart Data Analytics. Mit Hilfe von Big Data Zusammenhänge erkennen und Potentiale nutzen. Berlin, Berlin: De Gruyter Oldenbourg (De Gruyter Praxishandbuch).

Wikipedia (2017): Cross Industry Standard Process for Data Mining. Available online at https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining, updated on 6/28/2017, checked on 7/1/2017.

Williams, Lamar (2013): Enhancing Visibility and Agility in the Electronics Manufacturing Supply Chain. Precogs. Available online at http://www.circuitnet.com/news/uploads/2/Enhancing_Visibility_2013.pdf, checked on 6/25/2017.

Williams, Steve (2016): Business intelligence strategy and big data analytics. A general management perspective. Cambridge, MA: Morgan Kaufmann - Elsevier.

Williams, Thomas; Worley, Christopher G.; Lawler, Edward E. [III] (2013): The Agility Factor. strategy+business. Available online at <https://www.strategy-business.com/article/00188?gko=6a0ba>, updated on 4/15/2013, checked on 6/7/2017.

Wilson, Chauncey (2013): Brainstorming and Beyond. A User-Centered Design Method. Burlington: Elsevier Science.

Wilson, Eric; Demers, Marc (2015): Revolutionary and Evolutionary Approaches to Leveraging Predictive Analytics. In *Journal of Business Forecasting* (Winter 2014-2015), pp. 4–10.

Winkler, Roland; Heldmann, Stefan; Heise, Matthias (2017): Bedeuten mehr Daten auch bessere Entscheidungen? Bitkom. Big Data Summit. Available online at <https://www.bitkom-bigdata.de/programm/bedeuten-mehr-daten-auch-bessere-entscheidungen>, checked on 4/15/2017.

Winshuttle (2017): Big Data and the History of Information Storage. Available online at <https://www.winshuttle.com/big-data-timeline/>, checked on 8/12/2017.

Wirth, Rüdiger; Hipp, Jochen (2000): CRISP-DM: Towards a standard process model for data mining. In Practical Application Company (Ed.): Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. Manchester, UK, 11th-13th April 2000. Blackpool, Lancashire, UK: Practical Application Company, pp. 29–39.

Witten, Ian H.; Frank, Eibe; Hall, Mark A. (2011): Data mining. Practical machine learning tools and techniques. 3. ed. Amsterdam [u.a.]: Elsevier, Morgan Kaufmann Publishers (The Morgan Kaufmann series in data management systems).

Woodside, Joseph M. (2016): BEMO. A Parsimonious Big Data Mining Methodology. In *Online Academic Journal of Information Technology (AJIT-e)* 7 (24), pp. 113–123. DOI: 10.5824/1309-1581.2016.3.007.x.

Wu, C.; Buyya, R.; Ramamohanarao, K. (2016): Big Data Analytics = Machine Learning + Cloud Computing. In Rajkumar Buyya, Rodrigo N. Calheiros, Amir Vahid Dastjerdi (Eds.): Big data. Principles and paradigms. Cambridge, MA: Morgan Kaufmann - Elsevier (Science Direct e-books), pp. 3–38.

Xu, Lei; Jiang, Chunxiao; Wang, Jian; Yuan, Jian; Ren, Yong (2014): Information Security in Big Data. Privacy and Data Mining. In *IEEE Access* 2, pp. 1149–1176. DOI: 10.1109/ACCESS.2014.2362522.

Yahya, Nurhaziyatul A.; Samsudin, Ruhaidah; Shabri, Ani (2017): Tourism forecasting using hybrid modified empirical mode decomposition and neural network. In *International Journal of Advances in Soft Computing and its Applications* 9 (1), pp. 14–31.

Yang, Qiang; Wu, Xindong (2006): 10 challenging problems in data mining research. In *International Journal of Information Technology & Decision Making* 5 (4), pp. 597–604. DOI: 10.1142/S0219622006002258.

Yang, Xin-She; Lee, Sanghyuk; Lee, Sangmin; Theera-Umpon, Nipon (2015): Information Analysis of High-Dimensional Data and Applications. In *Mathematical Problems in Engineering* Volume 2015, pp. 1–2. DOI: 10.1155/2015/126740.

- Yu, Lei; Liu, Huan (2003): Feature Selection for High-Dimensional Data. A Fast Correlation-Based Filter Solution. In Tom Fawcett, Nina Mishra (Eds.): Proceedings of the 20th International Conference on Machine Learning. 20th International Conference on Machine Learning. Washington, DC, August 21-24, 2003. Menlo Park, CA: AAAI Press, pp. 856–863.
- Yu, Lei; Liu, Huan (2004): Efficient Feature Selection via Analysis of Relevance and Redundancy. In *Journal of Machine Learning Research* 5, pp. 1205–1224.
- Yusuf, Y. Y.; Adeleye, E. O. (2002): A comparative study of lean and agile manufacturing with a related survey of current practices in the UK. In *International Journal of Production Research* 40 (17), pp. 4545–4562. DOI: 10.1080/00207540210157141.
- Zemmouri, E. L. Moukhtar; Behja, Hicham; Marzak, Abdelaziz (2011): Towards a knowledge model for multi-view KDD process. In Adel M. Alimi (Ed.): 2011 3rd International Conference on Next Generation Networks and Services (NGNS). 2011 3rd International Conference on Next Generation Networks and Services (NGNS). Hammamet, Tunisia, 18-20 Dec. 2011. International Conference on Next Generation Networks and Services; NGNS. Piscataway, NJ: IEEE, pp. 18–22.
- Zeybek, Ömer; Ugurlu, Erginbay (2015): Nowcasting credit demand in Turkey with Google trends data. In *Journal of Applied Economic Sciences* 10 (2), pp. 293–300.
- Zhang, Xiaochen; Grijalva, Santiago (2015): An advanced data driven model for residential electric vehicle charging demand. In : Power & Energy Society General Meeting, 2015 IEEE. 2015 IEEE Power & Energy Society General Meeting. Denver, CO, USA, 26-30 July 2015. IEEE Power & Energy Society. Piscataway, NJ: IEEE, pp. 1–5.
- Zhao, Kai; Khryashchev, Denis; Freire, Juliana; Silva, Claudio; Vo, Huy (2016): Predicting taxi demand at high spatial resolution: Approaching the limit of predictability. In James B. D. Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia et al. (Eds.): Proceedings 2016 IEEE International Conference on Big Data. 2016 IEEE International Conference on Big Data. Washington D.C., USA, Dec 05-Dec 08, 2016. IEEE Computer Society. Piscataway, NJ: IEEE, pp. 833–842.
- Zhou, Zhi-Hua; Chawla, Nitesh V.; Jin, Yaochu; Williams, Graham J. (2014): Big Data Opportunities and Challenges. Discussions from Data Analytics Perspectives. In *IEEE Computational Intelligence Magazine* 9 (4), pp. 62–74. DOI: 10.1109/MCI.2014.2350953.
- Zou, Hui; Hastie, Trevor (2005): Regularization and variable selection via the elastic net. In *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2), pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

Appendix

A – Eurostat and OECD.Stat datasets

Eurostat ⁵⁹	
dataset_id	dataset title
cpc_insts	Candidate countries and potential candidates: short-term business statistics
ei_bpca_m	Balance of payments - BPM5 - Current account - monthly data
ei_bpfa_m	Balance of payments - BPM5 - Financial account - monthly data
ei_bpii_q	Balance of payments - BPM5 - International investment position - quarterly data
ei_bpm6ca_m	Balance of payments - BPM6 - Current account - monthly data
ei_bpm6fa_m	Balance of payments - BPM6 - Financial account - monthly data
ei_bpm6iip_q	Balance of payments - BPM6 - International investment position - quarterly data
ei_bsbu_m	Business surveys - NACE Rev. 1.1 - Construction - monthly data
ei_bsbu_m_bc	Business surveys - back-cast - Construction - monthly data
ei_bsbu_m_r2	Business surveys - NACE Rev. 2 - Construction - monthly data
ei_bsci_m	Business surveys - NACE Rev. 1.1 - Euro-zone Business Climate Indicator - monthly data
ei_bsci_m_r2	Business surveys - NACE Rev. 2 - Euro-zone Business Climate Indicator - monthly data
ei_bsco_m	Consumer surveys - Consumers - monthly data
ei_bsfs_m	Business surveys - NACE Rev. 2 - Financial services - monthly data
ei_bsin_m	Business surveys - NACE Rev. 1.1 - Industry - monthly data
ei_bsin_m_bc	Business surveys - back-cast - Industry - monthly data
ei_bsin_m_r2	Business surveys - NACE Rev. 2 - Industry - monthly data
ei_bsrt_m	Business surveys - NACE Rev. 1.1 - Retail sale - monthly data

⁵⁹ Based on Eurostat bulk download documentation [access date: 12/22/2015]:

<http://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing>

ei_bsrt_m_bc	Business surveys - back-cast - Retail sale - monthly data
ei_bsrt_m_r2	Business surveys - NACE Rev. 2 - Retail sale - monthly data
ei_bsse_m	Business surveys - NACE Rev. 1.1 - Services - monthly data
ei_bsse_m_bc	Business surveys - back-cast - Services - monthly data
ei_bsse_m_r2	Business surveys - NACE Rev. 2 - Services - monthly data
ei_bssi_m	Business surveys - NACE Rev. 1.1 - Sentiment indicators - monthly data
ei_bssi_m_r2	Business surveys - NACE Rev. 2 - Sentiment indicators - monthly data
ei_cphi_m	Consumer prices - Harmonised indices - monthly data
ei_etea19_m	International trade - Euro area 19 international trade - monthly data
ei_eteu28_m	International trade - EU28 international trade - monthly data
ei_hppi_q	Housing price statistics - House price index (2010 = 100) - quarterly data
ei_isbr_m	Industry, trade and services - Construction - monthly data - growth rates (NACE Rev. 2)
ei_isbu_m	Industry, trade and services - Construction - monthly data (NACE Rev. 2)
ei_isen_m	Industry, trade and services - Energy - monthly data
ei_isin_m	Industry, trade and services - Industry - monthly data (NACE Rev. 2)
ei_isir_m	Industry, trade and services - Industry - monthly data - growth rates (NACE Rev. 2)
ei_isppe_q	Industry, trade and services - Service producer prices - quarterly data - growth rates
ei_isppi_q	Industry, trade and services - Service producer prices - quarterly data - index
ei_isrr_m	Industry, trade and services - Retail trade - monthly data - growth rates (NACE Rev. 2)
ei_isrt_m	Industry, trade and services - Retail trade - monthly data (NACE Rev. 2)
ei_isse_q	Industry, trade and services - Turnover in services - quarterly data - growth rates (NACE Rev.2)
ei_isset_q	Industry, trade and services - Turnover in services - quarterly data - index

APPENDIX

ei_lmhr_m	Labour market - Harmonised unemployment rates (%) - monthly data
ei_lmhu_m	Labour market - Harmonised unemployment (1 000) - monthly data
ei_lmju_q_r2	Labour market - Job vacancy rate
ei_lmhc_q	Labour market - Labour cost index, nominal value - quarterly data
ei_mfef_m	Monetary and financial indicators - Effective exchange rates indices - monthly data
ei_mfir_m	Monetary and financial indicators - Interest rates - monthly data
ei_mfrt_m	Monetary and financial indicators - Euro/Ecu exchange rates - monthly data
ei_naag_q	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 1.1 - quarterly data
ei_naag_q_r2	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - quarterly data - current prices
ei_naar_q	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 1.1 - quarterly data - growth rates
ei_naem_q_r2	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - quarterly data - employment
ei_naga_a	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - Government accounts - annual data
ei_nagd_q_r2	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - General government deficit (-) and surplus (+) - quarterly data
ei_nagt_q_r2	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - General government gross debt - quarterly data
ei_naia_q	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - Income aggregates - quarterly data
ei_nama_q	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - Main aggregates - quarterly data
ei_namr_q	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - Main aggregates - quarterly data - growth rates
ei_nanf_q	National accounts - ESA 1995 - Quarterly Sector Accounts - Main aggregates

ei_napc_q	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - Price and cost indices - quarterly data
ei_nasa_q	National accounts - ESA 1995 - Quarterly Sector Accounts - Headline indicators
ei_navq_r2	National accounts - ESA 1995 - Aggregates by branch - NACE Rev. 2 - quarterly data - volumes
enpr_insts	Industry, trade and services - ENP countries: short-term business statistics
ert_bil_eur_m	Bilateral exchange rates - Euro/ECU exchange rates - monthly data
ert_eff_ic_m	Effective exchange rate indices - Industrial countries' effective exchange rates including new Member States - monthly data
ext_st_28msbec	International trade short-term indicators - Member States (EU28) trade by BEC product group since 1999
ext_st_ea19bec	International trade short-term indicators - Euro area19 trade by BEC product group since 1999
ext_st_ea19sitc	International trade short-term indicators - Euro area19 trade by SITC product group since 1999
ext_st_eftacc	International trade short-term indicators - Macro series for EFTA and enlargement countries (raw data and growth rates)
ext_st_eu28bec	International trade short-term indicators - EU28 trade by BEC product group
ext_st_eu28sitc	International trade short-term indicators - EU28 trade by SITC product group
irt_euryld_m	Interest rates - Euro yield curves - monthly data
irt_h_cgby_m	Interest rates - Central government bond yields - monthly data
irt_h_ddmr_m	Interest rates - Day-to-day rates for euro area countries - monthly data
irt_h_ecu11_m	Interest rates - ECU interest rates and yields - monthly data
irt_h_eurcoe_d	Interest rates - Euro yields - Coefficients - daily data
irt_h_euryld_m	Interest rates - Euro yields - Euro yield curves - monthly data
irt_h_mr3_m	Interest rates - 3-month rates for euro area countries - monthly data
irt_lt_gby10_m	Interest rates - Government bond yields, 10 years' maturity - monthly data

APPENDIX

irt_lt_mcby_m	Interest rates - EMU convergence criterion series - monthly data
irt_st_m	Interest rates - Money market interest rates - monthly data
lc_lci_r1_cow	Labour costs - Labour costs index: historical data - NACE Rev. 1.1 - Country weights
lc_lci_r1_itw	Labour costs - Labour costs index: historical data - NACE Rev. 1.1 - Item weights
lc_lci_r1_q	Labour costs - Labour costs index: historical data - NACE Rev. 1.1 - Labour cost index, nominal value - quarterly data
lc_lci_r2_cow	Labour costs - Labour costs index - Country weights - NACE Rev.2
lc_lci_r2_itw	Labour costs - Labour costs index - Item weights - NACE Rev.2
lc_lci_r2_q	Labour costs - Labour cost index, nominal value - quarterly data (NACE Rev. 2)
ifsi_dwl_a	Labour Force Survey main indicators - Duration of working life - annual data
ifsi_emp_a	Labour Force Survey main indicators - Employment (main characteristics and rates) - annual averages
ifsi_emp_q	Labour Force Survey main indicators - Employment (main characteristics and rates) - quarterly data
ifsi_exi_a	Labour Force Survey main indicators - Average exit age from the labour force - annual data
ifsi_grt_q	Labour Force Survey main indicators - Employment growth and activity branches - quarterly data
ifsi_jhh_a	Labour Force Survey main indicators - Population in jobless households - annual data
ifsi_long_q	Labour market transitions - LFS longitudinal data - quarterly data
nama_10_gdp	National accounts - ESA 2010 - GDP and main components (output, expenditure and income)
namq_10_a10	National accounts - ESA 2010 - Gross value added and income A*10 industry breakdowns
namq_10_a10_e	National accounts - ESA 2010 - Employment A*10 industry breakdowns

namq_10_an6	National accounts - ESA 2010 - Gross fixed capital formation with AN_F6 asset breakdowns
namq_10_exi	National accounts - ESA 2010 - Exports and imports by Member States of the EU/third countries
namq_10_fcs	National accounts - ESA 2010 - Final consumption aggregates
namq_10_gdp	National accounts - ESA 2010 - GDP and main components (output, expenditure and income)
namq_10_lp_ulc	Quarterly national accounts - Labour productivity and unit labour costs
namq_10_pc	Quarterly national accounts - Main GDP aggregates per capita
nasq_10_f_bs	Quarterly sector accounts (ESA 2010) - Financial flows and stocks - Financial balance sheets
nasq_10_f_gl	Quarterly sector accounts (ESA 2010) - Financial flows and stocks - Revaluation account
nasq_10_f_oc	Quarterly sector accounts (ESA 2010) - Financial flows and stocks - Other changes in volume
nasq_10_f_tr	Quarterly sector accounts (ESA 2010) - Financial flows and stocks - Financial transactions
nasq_10_ki	Quarterly sector accounts (ESA 2010) - Key indicators
nasq_10_nf_tr	Quarterly sector accounts (ESA 2010) - Non-financial transactions
nrg_101m	Energy statistics - Supply and transformation of solid fuels - monthly data
nrg_102m	Energy statistics - Supply and transformation of oil - monthly data
nrg_103m	Energy statistics - Supply of gas - monthly data
nrg_104m	Energy statistics - Supply and transformation of nuclear energy - monthly data
nrg_105m	Energy statistics - Supply of electricity - monthly data
nrg_122m	Energy statistics - Imports - solid fuels - monthly data
nrg_123m	Energy statistics - Imports - oil - monthly data
nrg_124m	Energy statistics - Imports - gas - monthly data
nrg_125m	Energy statistics - Imports - electricity - monthly data

APPENDIX

nrg_132m	Energy statistics - Exports - solid fuels - monthly data
nrg_133m	Energy statistics - Exports - oil - monthly data
nrg_134m	Energy statistics - Exports - gas - monthly data
nrg_135m	Energy statistics - Exports - electricity - monthly data
nrg_142m	Energy statistics - Oil stocks - stocks held for other countries and stocks held abroad - monthly data
nrg_143m	Energy statistics - Oil stocks - emergency stocks in days equivalent - monthly data
nrg_ind_342m	Energy statistics - Supply electricity - short-term monthly data
nrg_ind_343m	Energy statistics - Supply natural gas - short-term monthly data
nrg_jodi	Energy statistics - Supply oil – short-term monthly data
prc_hicp_aind	Harmonised indices of consumer prices - HICP (2005 = 100) - annual data (average index and rate of change)
prc_hicp_cann	Harmonised indices of consumer prices - HICP at constant taxes - monthly data (annual rate of change)
prc_hicp_cind	Harmonised indices of consumer prices - HICP at constant taxes - monthly data (index)
prc_hicp_cmon	Harmonised indices of consumer prices - HICP at constant taxes - monthly data (monthly rate of change)
prc_hicp_cow	Harmonised indices of consumer prices - HICP - Country weights
prc_hicp_inw	Harmonised indices of consumer prices - HICP - Item weights
prc_hicp_manr	Harmonised indices of consumer prices - HICP (2005 = 100) - monthly data (annual rate of change)
prc_hicp_midx	Harmonised indices of consumer prices - HICP (2005 = 100) - monthly data (index)
prc_hicp_mmor	Harmonised indices of consumer prices - HICP (2005 = 100) - monthly data (monthly rate of change)
prc_hicp_mv12r	Harmonised indices of consumer prices - HICP (2005 = 100) - monthly data (12-month average rate of change)

prc_hpi_inw	House price index - Item weights
prc_hpi_q	House price index (2010 = 100) - quarterly data
prc_ipc_a	National consumer price indices - annual data
prc_ipc_g20	G20 CPI all-items - Group of Twenty - Consumer price index
sts_cobp_m	Short-term business statistics - Building permits - monthly data (2010 = 100)
sts_colb_m	Short-term business statistics - Labour input in construction - monthly data (2010 = 100)
sts_copi_m	Short-term business statistics - Construction cost (or producer prices), new residential buildings - monthly data (2010 = 100)
sts_copr_m	Short-term business statistics - Production in construction - monthly data (2010 = 100)
sts_inlb_m	Short-term business statistics - Labour input in industry - monthly data (2010 = 100)
sts_inpi_m	Short-term business statistics - Import prices in industry - monthly data (2010 = 100)
sts_inpp_m	Short-term business statistics - Producer prices in industry, total - monthly data (2010 = 100)
sts_inppd_m	Short-term business statistics - Producer prices in industry, domestic market - monthly data (2010 = 100)
sts_inppnd_m	Short-term business statistics - Producer prices in industry, non domestic market - monthly data (2010 = 100)
sts_inpr_m	Short-term business statistics - Production in industry - monthly data (2010 = 100)
sts_intv_m	Short-term business statistics - Turnover in industry, total - monthly data (2010 = 100)
sts_intvd_m	Short-term business statistics - Turnover in industry, domestic market - monthly data (2010 = 100)
sts_intvnd_m	Short-term business statistics - Turnover in industry, non domestic market - monthly data (2010 = 100)

APPENDIX

sts_selb_m	Short-term business statistics - Labour input in services - monthly data (2010 = 100)
sts_sepp_q	Short-term business statistics - Service producer prices - quarterly data (2010 = 100)
sts_setu_m	Short-term business statistics - Turnover in services - monthly data (2010 = 100)
sts_trlb_m	Short-term business statistics - Labour input in wholesale and retail trade - monthly data (2010 = 100)
sts_trtu_m	Short-term business statistics - Turnover and volume of sales in wholesale and retail trade - monthly data (2010 = 100)
une_ltu_q	Unemployment - LFS adjusted series - Long-term unemployment by sex - quarterly average, %
une_nb_m	Unemployment - LFS adjusted series - Unemployment by sex and age - monthly average, 1 000 persons
une_rt_m	Unemployment - LFS adjusted series - Unemployment rate by sex and age - monthly average, %

Table 42 - Selected Eurostat datasets

OECD.Stat ⁶⁰	
dataset_id	dataset title
eo_q	Economic Outlook
itf_y	ITF Transport Statistics - Goods transport
mei_m	Main Economic Indicators
qasa_q	Public Sector Debt
qna_q	Quarterly National Accounts

Table 43 - Selected OECD.Stat datasets

⁶⁰ Based on OECD Data documentation [access date: 12/22/2015];

<https://data.oecd.org/searchresults/?r=+f/type/datasets>

B – Target data format

ERP		
Date	Frequency	Unit
<ul style="list-style-type: none"> Type: String Format: 'YYYY-MM-DD' Content: Date 	<ul style="list-style-type: none"> Type: String Format: 'Y', 'Q', 'M' or 'D' Content: Year, Quarter, Month or Day 	<ul style="list-style-type: none"> Type: String Format: General Text Content: Unit name, e.g., 'EUR' or 'kWh'
Properties	Values	
<ul style="list-style-type: none"> Type: String Format: General Text Content: Multiple text columns like 'Country', 'Industry', etc. 	<ul style="list-style-type: none"> Type: Float Format: Decimal numbers Content: Multiple value columns like 'Revenue', 'Costs', etc. 	
Financial database / Eurostat & OECD.Stat		
Date	Frequency	Unit
<ul style="list-style-type: none"> Type: String Format: 'YYYY-MM-DD' Content: Date 	<ul style="list-style-type: none"> Type: String Format: 'Y', 'Q', 'M' or 'D' Content: Year, Quarter, Month or Day 	<ul style="list-style-type: none"> Type: String Format: General Text Content: Unit name, e.g., 'EUR' or 'kWh'
Properties	Value	
<ul style="list-style-type: none"> Type: String Format: General Text Content: Multiple text columns like 'Country', 'Industry', etc. 	<ul style="list-style-type: none"> Type: Float Format: Decimal numbers Content: One value column according to "Variable" 	
Variable		
<ul style="list-style-type: none"> Type: String Format: General Text Content: Indicator name, e.g., 'GDP', 'Energy' 		

Figure 101 - Target data structure

