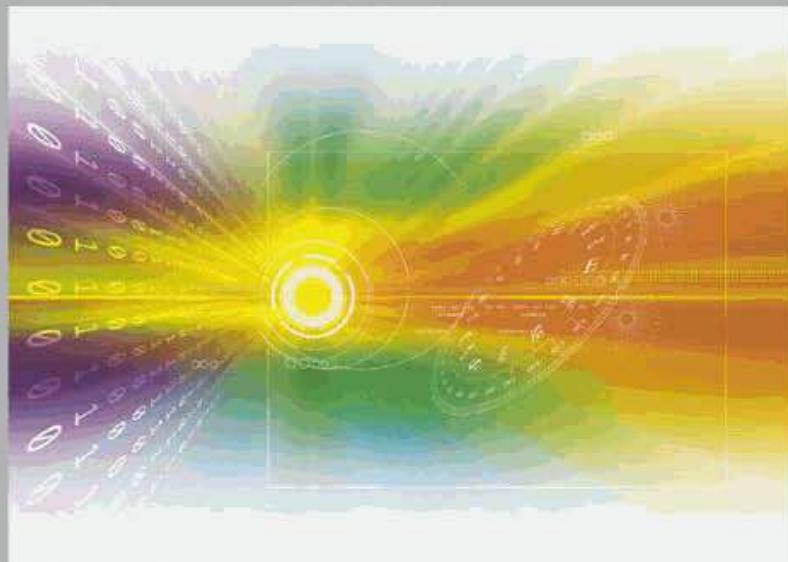


Guide to Intelligent Data Analysis

How to Intelligently Make
Sense of Real Data



Michael R. Berthold
Christian Borgelt
Frank Höppner · Frank Klawonn

Texts in Computer Science

Editors

David Gries

Fred B. Schneider

For further volumes:

<http://www.springer.com/series/3191>

Michael R. Berthold · Christian Borgelt ·
Frank Höppner · Frank Klawonn

Guide to Intelligent Data Analysis

How to Intelligently Make
Sense of Real Data



Prof. Dr. Michael R. Berthold
FB Informatik und
Informationswissenschaft
Universität Konstanz
78457 Konstanz
Germany
Michael.Berthold@uni-konstanz.de

Dr. Christian Borgelt
Intelligent Data Analysis &
Graphical Models Research Unit
European Centre for Soft Computing
C/ Gonzalo Gutiérrez Quirós s/n
Edificio Científico-Technológico
Campus Mieres, 3^a Planta
33600 Mieres, Asturias
Spain
christian.borgelt@softcomputing.es

Prof. Dr. Frank Höppner
FB Wirtschaft
Ostfalia University of Applied Sciences
Robert-Koch-Platz 10-14
38440 Wolfsburg
Germany
f.hoepner@ostfalia.de

Prof. Dr. Frank Klawonn
FB Informatik
Ostfalia University of Applied Sciences
Salzdahlumer Str. 46/48
38302 Wolfenbüttel
Germany
f.klawonn@ostfalia.de

Series Editors

David Gries
Department of Computer Science
Upson Hall
Cornell University
Ithaca, NY 14853-7501, USA

ISSN 1868-0941
ISBN 978-1-84882-259-7
DOI 10.1007/978-1-84882-260-3
Springer London Dordrecht Heidelberg New York

Fred B. Schneider
Department of Computer Science
Upson Hall
Cornell University
Ithaca, NY 14853-7501, USA

e-ISSN 1868-095X
e-ISBN 978-1-84882-260-3

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2010930517

© Springer-Verlag London Limited 2010

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licenses issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

The use of registered names, trademarks, etc., in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Cover design: VTeX, Vilnius

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The main motivation to write this book came from all our problems to find suitable material for a textbook that would really help us to teach the practical aspects of data analysis together with the needed theoretical underpinnings. Many books out there tackle either one or the other of these aspects (and, especially for the latter, there are some fantastic text books out there), but a book providing a good combination was nowhere to be found.

The idea to write our own book to address this shortcoming arose in two different places at the same time—when one of the authors was asked to review the book proposal of the others, we quickly realized that it would be much better to join forces instead of independently pursuing our individual projects.

We hope that this book helps others to learn what kind of challenges data analysts face in the real world and at the same time provides them with solid knowledge about the processes, algorithms, and theories to successfully tackle these problems. We have put a lot of effort into balancing the practical aspects of applying and using data analysis techniques while making sure at the same time that we did not forget to also explain the statistical and mathematical underpinnings behind the algorithms beneath all of this.

There are many people to be thanked, and we will not attempt to list them all. However, we do want to single out Iris Adä who has been a tremendous help with the generation of the data sets used in this book. She and Martin Horn also deserve our thanks for an intense last minute round of proof reading.

Konstanz, Germany
Oviedo, Spain
Braunschweig, Germany

Michael R. Berthold
Christian Borgelt
Frank Höppner and Frank Klawonn

Contents

1	Introduction	1
1.1	Motivation	1
1.1.1	Data and Knowledge	2
1.1.2	Tycho Brahe and Johannes Kepler	4
1.1.3	Intelligent Data Analysis	6
1.2	The Data Analysis Process	7
1.3	Methods, Tasks, and Tools	11
1.4	How to Read This Book	13
	References	14
2	Practical Data Analysis: An Example	15
2.1	The Setup	15
2.2	Data Understanding and Pattern Finding	16
2.3	Explanation Finding	20
2.4	Predicting the Future	21
2.5	Concluding Remarks	23
3	Project Understanding	25
3.1	Determine the Project Objective	26
3.2	Assess the Situation	28
3.3	Determine Analysis Goals	30
3.4	Further Reading	31
	References	32
4	Data Understanding	33
4.1	Attribute Understanding	34
4.2	Data Quality	37
4.3	Data Visualization	40
4.3.1	Methods for One and Two Attributes	40
4.3.2	Methods for Higher-Dimensional Data	48
4.4	Correlation Analysis	59

4.5	Outlier Detection	62
4.5.1	Outlier Detection for Single Attributes	63
4.5.2	Outlier Detection for Multidimensional Data	64
4.6	Missing Values	65
4.7	A Checklist for Data Understanding	68
4.8	Data Understanding in Practice	69
4.8.1	Data Understanding in KNIME	70
4.8.2	Data Understanding in R	73
	References	78
5	Principles of Modeling	81
5.1	Model Classes	82
5.2	Fitting Criteria and Score Functions	85
5.2.1	Error Functions for Classification Problems	87
5.2.2	Measures of Interestingness	89
5.3	Algorithms for Model Fitting	89
5.3.1	Closed Form Solutions	89
5.3.2	Gradient Method	90
5.3.3	Combinatorial Optimization	92
5.3.4	Random Search, Greedy Strategies, and Other Heuristics .	92
5.4	Types of Errors	93
5.4.1	Experimental Error	94
5.4.2	Sample Error	99
5.4.3	Model Error	100
5.4.4	Algorithmic Error	101
5.4.5	Machine Learning Bias and Variance	101
5.4.6	Learning Without Bias?	102
5.5	Model Validation	102
5.5.1	Training and Test Data	102
5.5.2	Cross-Validation	103
5.5.3	Bootstrapping	104
5.5.4	Measures for Model Complexity	105
5.6	Model Errors and Validation in Practice	111
5.6.1	Errors and Validation in KNIME	111
5.6.2	Validation in R	111
5.7	Further Reading	113
	References	113
6	Data Preparation	115
6.1	Select Data	115
6.1.1	Feature Selection	116
6.1.2	Dimensionality Reduction	121
6.1.3	Record Selection	121
6.2	Clean Data	123
6.2.1	Improve Data Quality	123

6.2.2	Missing Values	124
6.3	Construct Data	127
6.3.1	Provide Operability	127
6.3.2	Assure Impartiality	129
6.3.3	Maximize Efficiency	131
6.4	Complex Data Types	134
6.5	Data Integration	135
6.5.1	Vertical Data Integration	136
6.5.2	Horizontal Data Integration	136
6.6	Data Preparation in Practice	138
6.6.1	Data Preparation in KNIME	139
6.6.2	Data Preparation in R	141
	References	142
7	Finding Patterns	145
7.1	Hierarchical Clustering	147
7.1.1	Overview	148
7.1.2	Construction	150
7.1.3	Variations and Issues	152
7.2	Notion of (Dis-)Similarity	155
7.3	Prototype- and Model-Based Clustering	162
7.3.1	Overview	162
7.3.2	Construction	164
7.3.3	Variations and Issues	167
7.4	Density-Based Clustering	169
7.4.1	Overview	170
7.4.2	Construction	171
7.4.3	Variations and Issues	173
7.5	Self-organizing Maps	175
7.5.1	Overview	175
7.5.2	Construction	176
7.6	Frequent Pattern Mining and Association Rules	179
7.6.1	Overview	179
7.6.2	Construction	181
7.6.3	Variations and Issues	187
7.7	Deviation Analysis	194
7.7.1	Overview	194
7.7.2	Construction	195
7.7.3	Variations and Issues	197
7.8	Finding Patterns in Practice	198
7.8.1	Finding Patterns with KNIME	199
7.8.2	Finding Patterns in R	201
7.9	Further Reading	203
	References	204

8 Finding Explanations	207
8.1 Decision Trees	208
8.1.1 Overview	209
8.1.2 Construction	210
8.1.3 Variations and Issues	213
8.2 Bayes Classifiers	218
8.2.1 Overview	218
8.2.2 Construction	220
8.2.3 Variations and Issues	224
8.3 Regression	229
8.3.1 Overview	230
8.3.2 Construction	231
8.3.3 Variations and Issues	234
8.3.4 Two Class Problems	242
8.4 Rule learning	244
8.4.1 Propositional Rules	245
8.4.2 Inductive Logic Programming or First-Order Rules	251
8.5 Finding Explanations in Practice	253
8.5.1 Finding Explanations with KNIME	253
8.5.2 Using Explanations with R	255
8.6 Further Reading	257
References	258
9 Finding Predictors	259
9.1 Nearest-Neighbor Predictors	261
9.1.1 Overview	261
9.1.2 Construction	263
9.1.3 Variations and Issues	265
9.2 Artifical Neural Networks	269
9.2.1 Overview	269
9.2.2 Construction	272
9.2.3 Variations and Issues	276
9.3 Support Vector Machines	277
9.3.1 Overview	278
9.3.2 Construction	282
9.3.3 Variations and Issues	283
9.4 Ensemble Methods	284
9.4.1 Overview	284
9.4.2 Construction	286
9.4.3 Further Reading	289
9.5 Finding Predictors in Practice	290
9.5.1 Finding Predictors with KNIME	290
9.5.2 Using Predictors in R	292
References	294

10 Evaluation and Deployment	297
10.1 Evaluation	297
10.2 Deployment and Monitoring	299
References	301
A Statistics	303
A.1 Terms and Notation	304
A.2 Descriptive Statistics	305
A.2.1 Tabular Representations	305
A.2.2 Graphical Representations	306
A.2.3 Characteristic Measures for One-Dimensional Data	309
A.2.4 Characteristic Measures for Multidimensional Data	316
A.2.5 Principal Component Analysis	318
A.3 Probability Theory	323
A.3.1 Probability	323
A.3.2 Basic Methods and Theorems	327
A.3.3 Random Variables	333
A.3.4 Characteristic Measures of Random Variables	339
A.3.5 Some Special Distributions	343
A.4 Inferential Statistics	349
A.4.1 Random Samples	350
A.4.2 Parameter Estimation	351
A.4.3 Hypothesis Testing	361
B The R Project	369
B.1 Installation and Overview	369
B.2 Reading Files and R Objects	370
B.3 R Functions and Commands	372
B.4 Libraries/Packages	373
B.5 R Workspace	373
B.6 Finding Help	374
B.7 Further Reading	374
C KNIME	375
C.1 Installation and Overview	375
C.2 Building Workflows	377
C.3 Example Flow	378
C.4 R Integration	380
References	383
Appendix A	383
Appendix B	383
Index	385

Symbols

A, A_i	attribute, variable [e.g., $A_1 = \text{color}$, $A_2 = \text{price}$, $A_3 = \text{category}$]
ω	a possible value of an attribute [e.g., $\omega = \text{red}$]
$\Omega, \text{dom}(\cdot)$	set of possible values of an attribute [e.g., $\Omega_1 = \Omega_{\text{color}} = \text{dom}(A_i) = \{\text{red}, \text{blue}, \text{green}\}$]
\mathcal{A}	set of all attributes [e.g., $\mathcal{A} = \{\text{color}, \text{price}, \text{category}\}$]
m	number of considered attributes [e.g., 3]
x	a specific value of an attribute [e.g., $x_2 = x_{\text{price}} = 4000$]
\mathcal{X}	space of possible data records [e.g., $\mathcal{X} = \Omega_{A_1} \times \dots \times \Omega_{A_m}$]
\mathcal{D}	set of all records, data set, $\mathcal{D} \subseteq \mathcal{X}$ [e.g., $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$]
n	number of records in data set
\mathbf{x}	record in database [e.g., $\mathbf{x} = (x_1, x_2, x_3) = (\text{red}, 4000, \text{luxury})$]
\mathbf{x}_A	attribute A of record \mathbf{x} [e.g., $\mathbf{x}_{\text{price}} = 4000$]
$\mathbf{x}_{2,A}$	attribute A of record \mathbf{x}_2
$\mathcal{D}_{A=v}$	set of all records $\mathbf{x} \in \mathcal{D}$ with $\mathbf{x}_A = v$
C	a selected categorical target attribute [e.g., $C = A_3 = \text{category}$]
Ω_C	set of all possible classes [e.g., $\Omega_C = \{\text{quits}, \text{stays}, \text{unknown}\}$]
Y	a selected continuous target attribute [e.g., $Y = A_2 = \text{price}$]
\mathcal{C}	cluster (set of associated data objects) [e.g., $\mathcal{C} \subseteq \mathcal{D}$]
c	number of clusters
\mathcal{P}	partition, set of clusters $\{\mathcal{C}_1, \dots, \mathcal{C}_c\}$
$p_{i j}$	membership degree of data #j to cluster #i
$[p_{i j}]$	membership matrix
d	distance function, metric (d_E : Euclidean)
$[d_{i,j}]$	distance matrix

Chapter 1

Introduction

In this introductory chapter we provide a brief overview over some core ideas of intelligent data analysis and their motivation. In a first step we carefully distinguish between “data” and “knowledge” in order to obtain clear notions that help us to work out why it is usually not enough to simply collect data and why we have to strive to turn them into knowledge. As an illustration, we consider a well-known example from the history of science. In a second step we characterize the data analysis process, also often referred to as the knowledge discovery process, in which so-called “data mining” is one important step. We characterize standard data analysis tasks and provide a brief catalog of methods and tools to tackle them.

1.1 Motivation

Every year that passes brings us more powerful computers, faster and cheaper storage media, and higher bandwidth data connections. Due to these groundbreaking technological advancements, it is possible nowadays to collect and store enormous amounts of data with amazingly little effort and at impressively low costs. As a consequence, more and more companies, research centers, and governmental institutions create huge archives of tables, documents, images, and sounds in electronic form. Since for centuries lack of data has been a core hindrance to scientific and economic progress, we feel compelled to think that we can solve—at least in principle—basically any problem we are faced with if only we have enough data.

However, a closer examination of the matter reveals that this is an illusion. Data alone, regardless of how voluminous they are, are not enough. Even though large databases allow us to retrieve many different single pieces of information and to compute (simple) aggregations (like average monthly sales in Berlin), general patterns, structures, and regularities often go undetected. We may say that in the vast amount of data stored in some databases we cannot see the wood (the patterns) for the trees (the individual data records). However, it is most often exactly these patterns, regularities, and trends that are particularly valuable if one desires, for example, to increase the turnover of a supermarket. Suppose, for instance, that a



Purchasing
Behavior



1

Benefits
Sought



2

Customer
Journey Stage



3



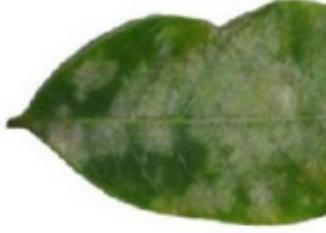
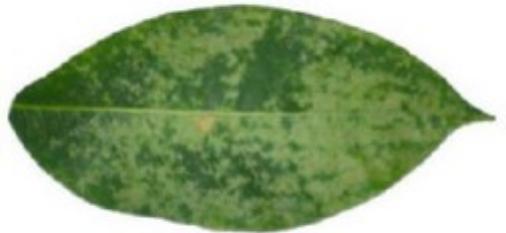
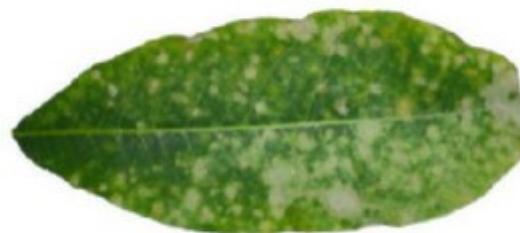
Customer
Satisfaction



Customer
Loyalty



Interest
Based



Eng

supermarket manager discovers, by analyzing the sales and customer records, that certain products are frequently bought together. In such a case sales can sometimes be stimulated by cleverly arranging these products on the shelves of the market (they may, for example, be placed close to each other, or may be offered as a bundle, in order to invite even more customers to buy them together).

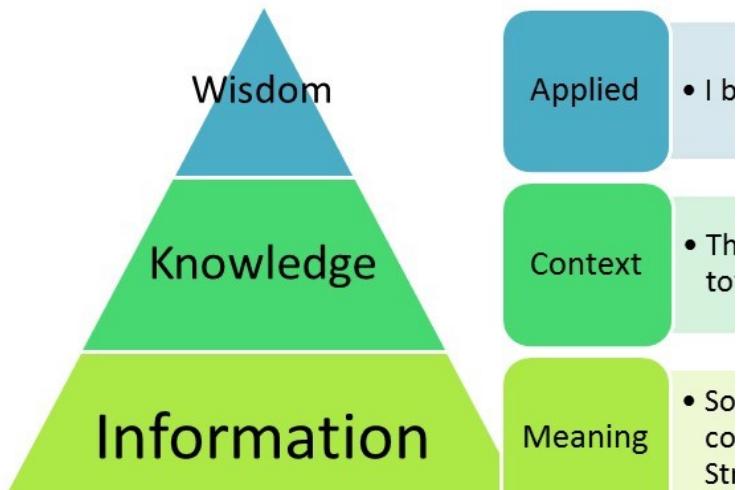
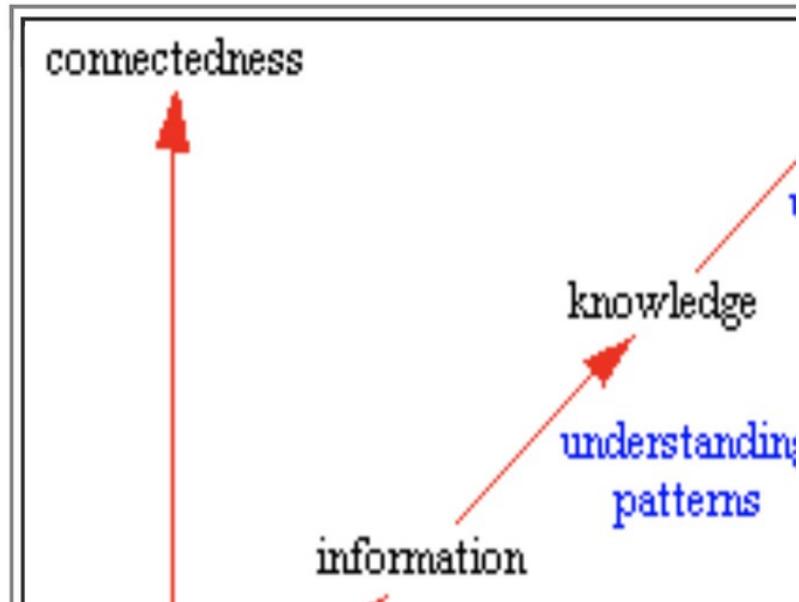
Unfortunately, it turns out to be harder than may be expected at first sight to actually discover such patterns and regularities and thus to exploit a larger part of the information that is contained in the available data. In contrast to the overwhelming flood of data there was, at least at the beginning, a lack of tools by which raw data could be transformed into useful information. Almost fifteen years ago John Naisbett aptly characterized the situation by saying [3]: “We are drowning in information, but starving for knowledge.” As a consequence, a new research area has been developed, which has become known under the name of *data mining*. The goal of this area was to meet the challenge to develop tools that can help humans to find potentially useful patterns in their data and to solve the problems they are facing by making better use of the data they have. Today, about fifteen years later, a lot of progress has been made, and a considerable number of methods and implementations of these techniques in software tools have been developed. Still it is not the tools alone, but the *intelligent composition* of human intuition with the computational power, of sound background knowledge with computer-aided modeling, of critical reflection with convenient automatic model construction, that leads *intelligent data analysis* projects to success [1]. In this book we try to provide a hands-on approach to many basic data analysis techniques and how they are used to solve data analysis problems if relevant data is available.

1.1.1 Data and Knowledge

In this book we distinguish carefully between *data* and *knowledge*. Statements like “Columbus discovered America in 1492” or “Mister Smith owns a VW Beetle” are **data**. Note that we ignore whether we already know these statements or whether we have any concrete use for them at the moment. The essential property of these statements we focus on here is that they refer to single events, objects, people, points in time, etc. That is, they generally refer to single instances or individual cases. As a consequence, their domain of application and thus their utility is necessarily limited.

In contrast to this, **knowledge** consists of statements like “All masses attract each other” or “Every day at 7:30 AM a flight with destination New York departs from Frankfurt Airport.” Again, we neglect the relevance of these statements for our current situation and whether we already know them. Rather, we focus on the essential property that they do *not* refer to single instances or individual cases but are general rules or (physical) laws. Hence, if they are true, they have a large domain of application. Even more importantly, though, they allow us to make predictions and are thus highly useful (at least if they are relevant to us).

We have to admit, though, that in daily life we also call statements like “Columbus discovered America in 1492” knowledge (actually, this particular statement is



used as a kind of prototypical example of knowledge). However, we neglect here this vernacular and rather fuzzy use of the notion “knowledge” and express our regrets that it is not possible to find a terminology that is completely consistent with everyday speech. Neither single statements about individual cases nor collections of such statements qualify, in our use of the term, as knowledge.

Summarizing, we can characterize data and knowledge as follows:

data

- refer to single instances
 - (single objects, people, events, points in time, etc.)
 - describe individual properties
 - are often available in large amounts
 - (databases, archives)
 - are often easy to collect or to obtain
 - (e.g., scanner cashiers in supermarkets, Internet)
 - do not allow us to make predictions or forecasts
-

knowledge

- refers to *classes* of instances
 - (*sets* of objects, people, events, points in time, etc.)
 - describes general patterns, structures, laws, principles, etc.
 - consists of as few statements as possible
 - (this is actually an explicit goal, see below)
 - is often difficult and time-consuming to find or to obtain
 - (e.g., natural laws, education)
 - allows us to make predictions and forecasts
-

These characterizations make it very clear that generally knowledge is much more valuable than (raw) data. Its generality and the possibility to make predictions about the properties of new cases are the main reasons for this superiority.

It is obvious, though, that not all kinds of knowledge are equally valuable as any other. Not all general statements are equally important, equally substantial, equally significant, or equally useful. Therefore knowledge has to be assessed, so that we do not drown in a sea of irrelevant knowledge. The following list (which we do not claim to be complete) lists some of the most important criteria:

criteria to assess knowledge

- correctness (probability, success in tests)
 - generality (domain and conditions of validity)
 - usefulness (relevance, predictive power)
 - comprehensibility (simplicity, clarity, parsimony)
 - novelty (previously unknown, unexpected)
-

In the domain of science, the focus is on correctness, generality, and simplicity (parsimony) are in the focus: one way of characterizing science is to say that it is the search for a minimal correct description of the world. In economy and industry, however, the emphasis is placed on usefulness, comprehensibility, and novelty: the main goal is to gain a competitive edge and thus to increase revenues. Nevertheless, neither of the two areas can afford to neglect the other criteria.

1.1.2 Tycho Brahe and Johannes Kepler

We illustrate the considerations of the previous section with an (at least partially) well-known example from the history of science. In the sixteenth century studying the stars and the planetary motions was one of the core areas of research. Among its proponents was Tycho Brahe (1546–1601), a Danish nobleman and astronomer, who in 1576 and 1584, with the financial help of King Frederic II, built two observatories on the island of Ven, about 32 km north-east of Copenhagen. He had access to the best astronomical instruments of his time (but no telescopes, which were used only later by Galileo Galilei (1564–1642) and Johannes Kepler (see below) to observe celestial bodies), which he used to determine the positions of the sun, the moon, and the planets with a precision of less than one angle minute. With this precision he managed to surpass all measurements that had been carried out before and to actually reach the theoretical limit for observations with the unaided eye (that is, without the help of telescopes). Working carefully and persistently, he recorded the motions of the celestial bodies over several years.

Stated plainly, Tycho Brahe collected data about our planetary system, fairly large amounts of data, at least from the point of view of the sixteenth century. However, he failed to find a consistent scheme to combine them, could not discern a clear underlying pattern—partially because he stuck too closely to the geocentric system (the earth is in the center, and all planets, the sun, and the moon revolve around the earth). He could tell the precise location of Mars on any given day of the year 1582, but he could not connect its locations on different days by a clear and consistent theory. All hypotheses he tried did not fit his highly precise data. For example, he developed the so-called Tychonic planetary system (the earth is in the center, the sun and the moon revolve around the earth, and the other planets revolve around the sun on circular orbits). Although temporarily popular in the seventeenth century, this system did not stand the test of time. From a modern point of view we may say that Tycho Brahe had a “data analysis problem” (or “knowledge discovery problem”). He had obtained the necessary data but could not extract the hidden knowledge.

This problem was solved later by Johannes Kepler (1571–1630), a German astronomer and mathematician, who worked as an assistant of Tycho Brahe. Contrary to Brahe, he advocated the Copernican planetary system (the sun is in the center, the earth and all other planets revolve around the sun in circular orbits) and tried all his life to reveal the laws that govern the motions of the celestial bodies. His approach was almost radical for his time, because he strove to find a mathematical description. He started his investigations with the data Tycho Brahe had collected and which he extended in later years. After several fruitless trials and searches and long and cumbersome calculations (imagine: no pocket calculators), Kepler finally succeeded. He managed to combine Tycho Brahe’s data into three simple laws, which nowadays bear his name: **Kepler’s laws**. After having realized in 1604 already that the course of Mars is an ellipse, he published the first two of these laws in his work

“Astronomia Nova” in 1609 [6] and the third law ten years later in his magnum opus “Harmonices Mundi” [4, 7]:

1. The orbit of every planet (including the earth) is an ellipse, with the sun at a focal point.
2. A line from the sun to the planet sweeps out equal areas during equal intervals of time.
3. The squares of the orbital periods of any two planets relate to each other like the cubes of the semimajor axes of their respective orbits:
 $T_1^2/T_2^2 = a_1^3/a_2^3$, and therefore generally $T \sim a^{3/2}$.

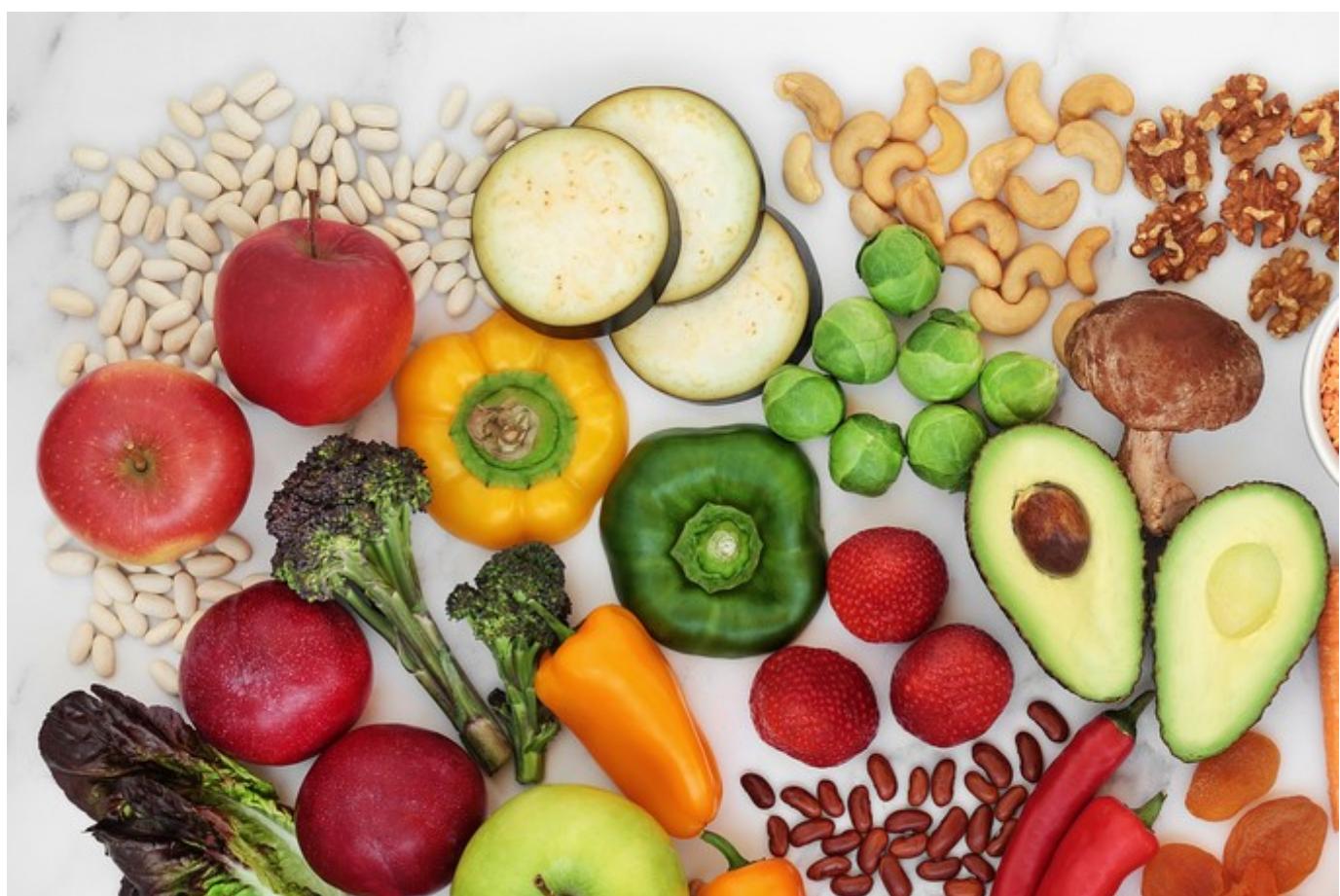
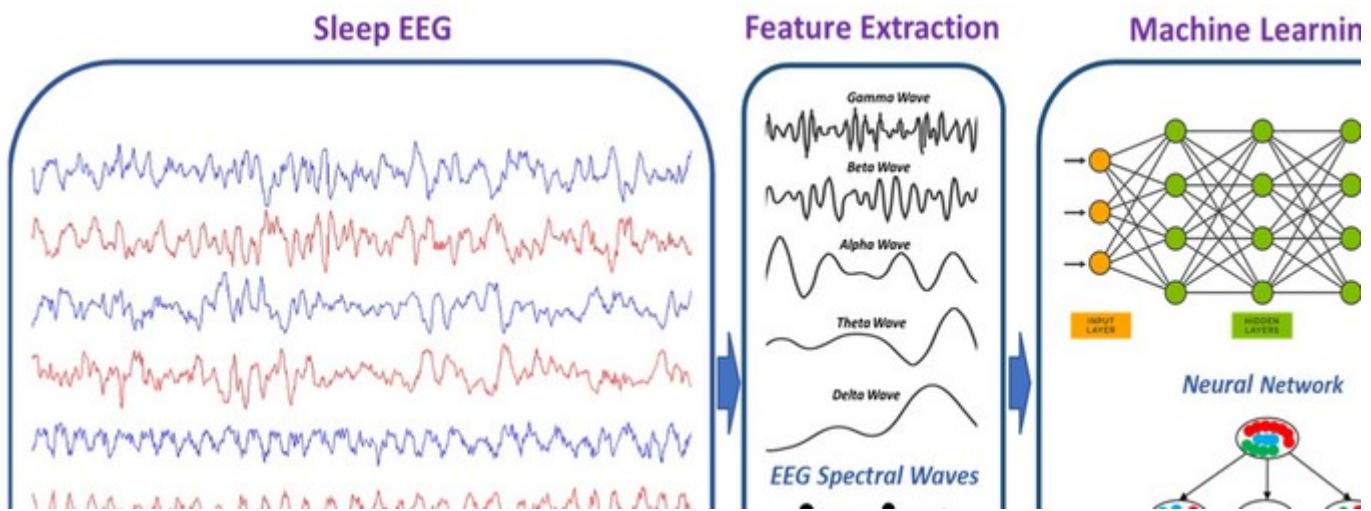
Tycho Brahe had collected a large amount of astronomical data, and Johannes Kepler found the underlying laws that can explain them. He discovered the hidden knowledge and thus became one of the most famous “data miners” in history.

Today the works of Tycho Brahe are almost forgotten—few have even heard his name. His catalogs of celestial data are merely of historical interest. No textbook on astronomy contains excerpts from his measurements—and this is only partially due to the better measurement technology we have available today. His observations and precise measurements are raw data and thus suffer from a decisive drawback: they do not provide any insight into the underlying mechanisms and thus do not allow us to make predictions. Kepler’s laws, on the other hand, are treated in basically all astronomy and physics textbooks, because they state the principles according to which planets and comets move. They combine all of Brahe’s observations and measurements in three simple statements. In addition, they permit us to make predictions: if we know the location and the speed of a planet relative to the sun at any given moment, we can compute its future course by drawing on Kepler’s laws.

How did Johannes Kepler find the simple astronomical laws that bear his name? How did he discover them in Tycho Brahe’s long tables and voluminous catalogs, thus revolutionizing astronomy? We know fairly little about his searches and efforts. He must have tried a large number of hypotheses, most of them failing. He must have carried out long and cumbersome computations, repeating some of them several times to eliminate errors. It is likely that exceptional mathematical talent, hard and tenacious work, and a significant amount of good luck finally led him to success. What we can be sure of is that he did not possess a universally applicable procedure or method to discover physical or astronomical laws.

Even today we are not much further: there is still no silver bullet to hit on the right solution. It is still much easier to collect data, with which we are virtually swamped in today’s “information society” (whatever this popular term actually means) than to discover knowledge. Automatic measurement instruments and scanners, digital cameras and computers, and an abundance of other automatic and semiautomatic devices have even relieved us of the burden of manual data collection. In addition, database and data warehouse technology allows us to store ever increasing amounts of data and to retrieve and to sample them easily. John Naisbett was perfectly right: “We are drowning in information, but starving for knowledge.”

It took a distinguished researcher like Johannes Kepler several years (actually half a lifetime) to evaluate the data that Tycho Brahe had collected—data that from



a modern point of view are negligibly few and of which Kepler actually analyzed closely only those about the orbit of Mars. Given this, how can we hope today to cope with the enormous amounts of data we are faced with every day? “Manual” analyses (like Kepler’s) have long ceased to be feasible. Simple aids, like the visualization of data in charts and diagrams, even though highly useful and certainly a first and important step, quickly reach their limits. Thus, if we refuse to surrender to the flood of data, we are forced to develop and employ computer-aided techniques, with which data analysis can be simplified or even automated to some degree. These are the methods that have been and still are developed in the research areas of intelligent data analysis, knowledge discovery in databases and data mining. Even though these methods are far from replacing human beings like Johannes Kepler, especially since a mindless application can produce artifacts and misleading results, it is not entirely implausible to assume that Kepler, if he had been supported by these methods and tools, could have reached his goal a little earlier.

1.1.3 Intelligent Data Analysis

Many people associate any kind of data analysis with **statistics** (see also Appendix A, which provides a brief review). Statistics has a long history and originated from collecting and analyzing data about the population and the state in general.

Statistics can be divided into *descriptive* and *inferential statistics*. **Descriptive statistics** summarizes data without making specific assumptions about the data, often by characteristic values like the (empirical) mean or by diagrams like histograms. **Inferential statistics** provides more rigorous methods than descriptive statistics that are based on certain assumptions about the data generating random process. The conclusions drawn in inferential statistics are only valid if these assumptions are satisfied.

Typically, in statistics **the first step of the data analysis process is to design the experiment** that defines how data should be collected in order to be able to carry out a reliable analysis based on the obtained data. To capture this important issue, we distinguish between *experimental* and *observational studies*. In an **experimental study** one can control and manipulate the data generating process. For instance, if we are interested in the effects of certain diets on the health status of a person, we might ask different groups of people to stick to different diets. Thus we have a certain control over the data generating process. In this experimental study, we can decide which and how many people should be assigned to a certain diet.

In an **observational study** one cannot control the data generating process. For the same dietary study as above, we might simply ask people on the street what they normally eat. Then we have no control about which kinds of diets we get data and how many people we will have for each diet in our data.

No matter whether the study is experimental or observational, there are usually **independence assumptions involved, and the data we collect should be representative**. The main reason is that inferential statistics is often applied to *hypothesis testing* where, based on the collected data, we desire to either confirm or reject some

Null hypothesis



- *The number of working hours in a day doesn't affect morale.*
- *Scheduling more team meetings doesn't affect deliverables.*
- *Employees are more productive if the company takes one break every two hours, as opposed to one hour.*
- *Eating three balanced meals a day has a positive productivity, as opposed to eating small snacks.*

hypothesis about the considered domain. In this case representative data and certain independencies are required in order to ensure that the test decisions are valid.

In contrast to hypothesis testing, **exploratory data analysis** is concerned with **generating hypotheses** from the collected data. In exploratory data analysis there are no or at least considerably weaker model assumptions about the data generating process. Most of the methods presented in this book fall into this category, since they are mostly universal methods designed to achieve a certain goal but are not based on a rigorous model as in inferential statistics.

The typical situation we assume in this book is that we already have the data. They might not have been collected in the best way, or in the way we would have collected them had we been able to design the experiment in advance. Therefore, it is often difficult to make specific assumptions about the data generating process. We are also mostly goal-oriented—that is, we ask questions like “Which customers will yield the highest profit”?—and search for methods that can help us to answer such questions or to solve our problems.

The opportunity of analyzing large business databases that were initially collected for completely different purposes came with the availability of powerful tools and technologies that can process and analyze massive amounts of data, so-called **data mining** techniques. A few years ago some people seemed to believe that with just the right data mining tool at hand any kind of desired knowledge could be squeezed out of a given database *automatically* with no or only little human interference. However, practical experience demonstrates that every problem is different and a full automatization of the data analysis process is simply impossible. Today we understand by **knowledge discovery in databases** (KDD) an interactive “process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” [3]. This process consists of multiple phases, and the data mining or modeling step became just a single step in it. That is, after a period of time where powerful tools were (sometimes) naively applied to the data, the “intelligent analyst” is brought back into the loop. As a consequence, the KDD process differs not so much anymore from classical statistical data analysis (except where the lacking principled data acquisition takes its toll). To emphasize that every project is different and therefore intelligence is required to make the most out of the already gathered data, we use the term **intelligent data analysis**, which was coined by David Hand [1, 5] (and is used today almost synonymously with the KDD process).

In this book we strove to provide a comprehensive guide to intelligent data analysis, outlining the process and its phases, presenting methods and algorithms for various tasks and purposes, and illustrating them with two freely available software tools. In this way we hope to offer a good starting point for anyone who wishes to become more familiar with the area of intelligent data analysis.

1.2 The Data Analysis Process

There are at least two typical situations in which intelligent data analysis may help us to find solutions to certain problems or provide answers to questions that arise.

In the first case, the problem at hand is by no means new, but it is already solved as a matter of routine (e.g., approval of credit card applications, technical inspection during quality assurance, machine control by a plant operator, etc.). If data has been collected for the past cases together with the result that was finally achieved (such as poor customer performance, malfunction of parts, etc.), such historical data may be used to revise and optimize the presently used strategy to reach a decision. In the second case, a certain question arises for the first time, and only little experience is available, or the experience is not directly applicable to this new question (e.g., starting with a new product, preventing abuse of servers, evaluating a large experiment or survey). In such cases, it is supposed that data from related situations may be helpful to generalize the new problem or that unknown relationships can be discovered from the data to gain insights into this unfamiliar area.

What if we have no data at all? This situation does not occur literally in practice, since in most cases there is always *some* data. Especially in businesses huge amounts of data have been collected and stored for operational reasons in the past (e.g., billing, logistics, warranty claims) that may now be used to optimize various decisions or offer new options (e.g., predicting customer performance, reducing stock on hand, tracking causes of defects). So the right question should be: How do we know if we have enough *relevant* data? This question is not answered easily. If it actually turns out that the data is not sufficient, one option is to acquire new data to solve the problem. However, as already pointed out in the preceding section, the experimental design of data acquisition is beyond the scope of this book.

There are several proposals about what the intelligent data analysis process should look like, such as SEMMA (an acronym for *sample*, *explore*, *modify*, *model*, *assess* used by SAS Institute Inc.), CRISP-DM (an acronym for *Cross Industry Standard Process for Data Mining* as defined by the CRISP-DM consortium) [2], or the KDD-process [3] (see [8] for a detailed comparison). In this book, we are going to follow the CRISP-DM process, which has been developed by a consortium of large companies, such as NCR, Daimler, and SPSS, and appears to be the most widely used process model for intelligent data analysis today.

CRISP-DM consists of six phases as shown in Fig. 1.1. Most of these phases are usually executed more than once, and the most frequent phase transitions are shown by arrows. The main objective of the first **project understanding** step (see Chap. 3) is to identify the potential benefit as well as the risks and efforts of a successful project, such that a deliberate decision on conducting the full project can be made. The envisaged solution is also transferred from the project domain to a more technical, data-centered notion. This first phase is usually called business understanding, but we stick to the more general term *project understanding* to emphasize that our problem at hand may as well be purely technical in nature or a research project rather than economically motivated.

Next we need to make sure that we will have sufficient data at hand to tackle the problem. While we cannot know this for sure until the end of the project, we at least have to convince ourselves that there is enough relevant data. To achieve this, we proceed in the **data understanding** phase (see Chap. 4) with a review of the available databases and the information contained in the database fields, a visual

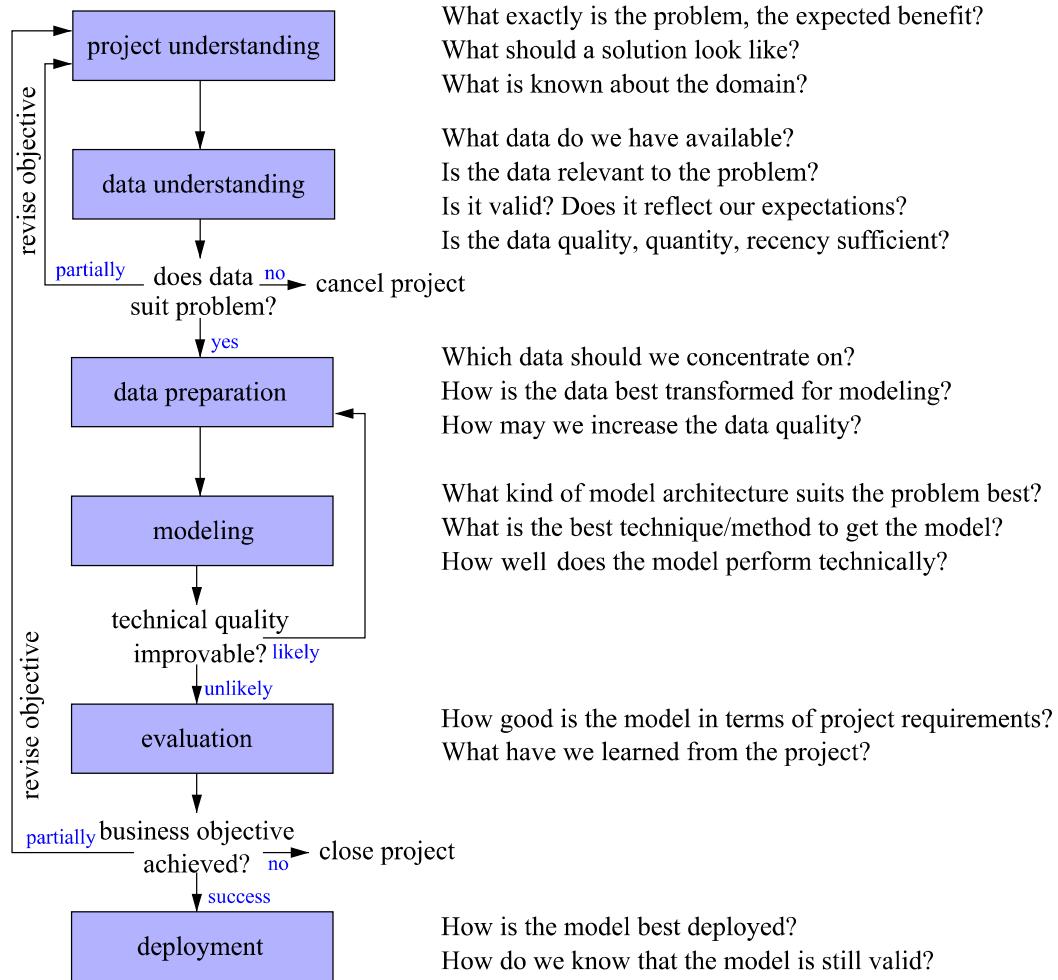


Fig. 1.1 Overview of the CRISP-DM process together with typical questions to be asked in the respective phases

assessment of the basic relationships between attributes, a data quality audit, an inspection of abnormal cases (outliers), etc. For instance, outliers appear to be abnormal in some sense and are often caused by faulty insertion, but sometimes they give surprising insights on closer inspection. Some techniques respond very sensitively to outliers, which is why they should be treated with special care. Another aspect is empty fields which may occur in the database for various reasons—ignoring them may introduce a systematic error in the results. By getting familiar with the data, typically first insights and hypotheses are gained. If we do not believe that the data suffices to solve the problem, it may be necessary to revise the project's objective.

So far, we have not changed any field of our database. However, this will be required to get the data into a shape that enables us to apply modeling tools. In the **data preparation** phase (Chap. 6) the data is selected, corrected, modified, even new attributes are generated, such that the prepared data set best suits the problem and the envisaged modeling technique. Basically all deficiencies that have been identified in the data understanding phase require special actions. Often the outliers

and missing values are replaced by estimated values or true values obtained from other sources. We may restrict the further analysis to certain variables and to a selection of the records from the full data set. Redundant and irrelevant data can give many techniques an unnecessarily hard time.

Once the data is prepared, we select and apply modeling tools to extract knowledge out of the data in the form of a model (Chaps. 5 and 7–9). Depending on what we want to do with the model, we may choose techniques that are easily interpretable (to gain insights) or less demonstrative black-box models, which may perform better. If we are not pleased with the results but are confident that the model can be improved, we step back to the data preparation phase and, say, generate new attributes from the existing ones, to support the modeling technique or to apply different techniques. Background knowledge may provide hints on useful transformations that simplify the representation of the solution.

Compared to the modeling itself, which is typically supported by efficient tools and algorithms, the data understanding and preparation phases take considerable part of the overall project time as they require a close manual inspection of the data, investigations into the relationships between different data sources, often even the analysis of the process that generated the data. New insights promote new ideas for feature generation or alter the subset of selected data, in which case the data preparation and modeling phases are carried out multiple times. The number of steps is not predetermined but influenced by the process and findings itself.

When the technical benchmarks cannot be improved anymore, the obtained results are analyzed in the evaluation phase (Chap. 10) from the perspective of the problem owner. At this point, the project may stop due to unsatisfactory results, the objectives may be revised in order to succeed under a slightly different setting, or the found and optimized model may be deployed.

After deployment, which ranges from writing a report to the creation of a software system that applies the model automatically to aid or make decisions, the project is not necessarily finished. If the project results are used continuously over time, an additional monitoring phase is necessary: during the analysis, a number of assumptions will be made, and the correctness of the derived model (and the decisions that rely on the model) depends on them. So we better verify from time to time that these assumption still hold to prevent decision-making on outdated information.

In the literature one can find attempts to create cost models that estimate the costs associated with a data analysis project. Without going into the details, the major key factors that remained in a reduced cost model derived from 40 projects were [9]:

- the number of tables and attributes,
- the dispersion of the attributes (only a few vs. many values),
- the number of external data sources,
- the type of the model (prediction being the most expensive),
- the attribute type mixture (mixture of numeric and nonnumeric), and
- the familiarity of the staff with data analysis projects in general, the project domain in particular, and the software suites.

While there is not much we can do about the problem size, the goal of this book is to increase the familiarity with data analysis projects by going through each of the phases and providing first instructions to get along with the software suites.

1.3 Methods, Tasks, and Tools

Problem Categories Every data analysis problem is different. To avoid the effort of inventing a completely new solution for each problem, it is helpful to think of different problem categories and consider them as building blocks from which a solution may be composed. These categories also help to categorize the large number of different tools and algorithms that solve specific tasks. Over the years, the following set of method categories has been established [3]:

- **classification**

Predict the outcome of an experiment with a finite number of possible results (like *yes/no* or *unacceptable/acceptable/good/very good*). We may be interested in a prediction because the true result will emerge in the future or because it is expensive, difficult, or cumbersome to determine it.

Typical questions: *Is this customer credit-worthy? Will this customer respond to our mailing? Will the technical quality be acceptable?*

- **regression**

Regression is, just like classification, also a prediction task, but this time the value of interest is numerical in nature.

Typical questions: *How will the EUR/USD exchange rate develop? How much money will the customer spend for vacation next year? How much will the machine's temperature change within the next cycle?*

- **clustering, segmentation**

Summarize the data to get a better overview by forming groups of similar cases (called clusters or segments). Instead of examining a large number of similar records, we need to inspect the group summary only. We may also obtain some insight into the structure of the whole data set. Cases that do not belong to any group may be considered as abnormal or outliers.

Typical questions: *Do my customers divide into different groups? How many operating points does the machine have, and what do they look like?*

- **association analysis**

Find any correlations or associations to better understand or describe the inter-dependencies of all the attributes. The focus is on *relationships* between all attributes rather than focusing on a single target variable or the cases (full record).

Typical questions: *Which optional equipment of a car often goes together? How do the various qualities influence each other?*

- **deviation analysis**

Knowing already the major trends or structures, find any exceptional subgroup that behaves differently with respect to some target attribute.

Typical questions: *Under which circumstances does the system behave differently? Which properties do those customers share who do not follow the crowd?*

The most frequent categories are *classification* and *regression*, because decision making always becomes much easier if reliable predictions of the near future are available. When a completely new area or domain is explored, cluster analysis and association analysis may help to identify relationships among attributes or records. Once the major relationships are understood (e.g., by a domain expert), a deviation analysis can help to focus on *exceptional situations* that deviate from regularity.

Catalog of Methods There are various methods in each of these categories to find reliable answers to the questions raised above. However, there is no such thing as a *single gold method* that works perfectly for all problems. To convey some idea which method may be best suited for a given problem, we will discuss various methods in Chaps. 7–9. However, in order to organize these chapters, we did not rely on the problem categories collected above, as some methods can be used likewise for more than one problem type. We rather used the intended task of the data analysis as a grouping criterion:

- **finding patterns** (Chap. 7)

If the domain (and therefore the data) is new to us or if we expect to find interesting relationships, we explore the data for new, previously unknown patterns. We want to get a full picture and do not concentrate on a single target attribute yet. We may apply methods from, for instance, segmentation, clustering, association analysis, or deviation analysis.

- **finding explanations** (Chap. 8)

We have a special interest in some target variable and wonder why and how it varies from case to case. The primary goal is to gain new insights (knowledge) that may influence our decision making, but we do not necessarily intend automation. We may apply methods from, for instance, classification, regression, association analysis, or deviation analysis.

- **finding predictors** (Chap. 9)

We have a special interest in the prediction of some target variable, but it (possibly) represents only one building block of our full problem, so we do not really care about the *how* and *why* but are just interested in the best-possible prediction. We may apply methods from, for instance, classification or regression.

Available Tools As already mentioned, the key to success is often the proper combination of data preparation and modeling techniques. Data analysis software suites are of great help as they reduce data formatting efforts and ease method linking. There is a long list of commercial and free software suites and tools, including the following *classical* products:

- **IBM SPSS PASW Modeler** (formerly Clementine)

Clementine was the first commercial data mining workbench in 1994 and is a commercial product from SPSS, now IBM.

<http://www.spss.com/>

- **SAS Enterprise Miner**

A commercial data mining solution from SAS.

<http://www.sas.com/>

- The R-project

R is a free software environment for statistical computing and graphics.

<http://www.r-project.org/>

- Weka

Weka is a popular open-source collection of machine learning algorithms, initially developed by the University of Waikato, New Zealand.

<http://www.cs.waikato.ac.nz/ml/weka/>

For an up-to-date list of software suites see, for instance,

<http://www.kdnuggets.com/software/suites.html>

Although the choice of the software suite has considerable impact on the project time (usability) and can help to avoid errors (because some of them are easily spotted using powerful visualization capabilities), the suites cannot take over the full analysis process. They provide at best an initial starting point (by means of analysis templates or project wizards), but in most cases the key factor is the *intelligent combination* of tools and background knowledge (regarding the project domain and the utilized tools). The suites exhibit different strengths, some focus on supporting the human data analyst by sophisticated graphical user interfaces, graphical configuration and reporting, while others are better suited for batch processing and automation.

In this book, we will use R, which is particularly powerful in statistical techniques, and KNIME (the *Konstanz Information Miner*¹), which is an open-source data analysis tool that is growing in popularity due to its graphical workflow editor and its ability to integrate other well-known toolkits.

1.4 How to Read This Book

In the next chapter we will take a glimpse at the intelligent data analysis process by looking over the shoulder of Stan and Laura as they analyze their data (while only one of them actually follows CRISP-DM). The chapter is intended to give an impression of what will be discussed in much greater detail throughout the book. The subsequent chapters roughly follow the CRISP-DM stages: we analyze the problem first in Chap. 3 (project understanding) and then investigate whether the available data suits our purposes in terms of size, representativeness, and quality in Chap. 4 (data understanding). If we are confident that the data is worth carrying out the analysis, we discuss the data preparation (Chap. 6) as the last step before we enter the modeling phase (Chaps. 7–9). As already mentioned, data preparation is already tailored to the methods we are going to use for modeling; therefore, we have to introduce the principles of modeling already in Chap. 5. Deployment and monitoring is briefly addressed in Chap. 10. Readers who, over the years, have lost some of their statistical knowledge can (partially) recover it in Appendix A. The statistics

¹ Available for download at <http://www.knime.org/>.

appendix is not just a glossary of terms to quickly look up details but also serves as a book within the book for a few preparative lessons on statistics before delving into the chapters about intelligent data analysis.

Most chapters contain a section that equips the reader with the necessary information for some first hands-on experience using either R or KNIME. We have settled on R and KNIME because they can be seen as extremes on the range of possible software suites: R is a statistical tool, which is (mostly) command-line oriented and is particularly useful for scripting and automatization. KNIME, on the other hand, supports the composition of complex workflows in a graphical user interface.² Appendices B and C provide a brief introduction into both systems.

References

1. Berthold, M., Hand, D.: Intelligent Data Analysis. Springer, Berlin (2009)
2. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: Cross Industry Standard Process for Data Mining 1.0, Step-by-step Data Mining Guide. CRISP-DM consortium (2000)
3. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press, Menlo Park/Cambridge (1996)
4. Feynman, R.P., Leighton, R.B., Sands, M.: The Feynman Lectures on Physics. Mechanics, Radiation, and Heat, vol. 1. Addison-Wesley, Reading (1963)
5. Hand, D.: Intelligent data analysis: issues and opportunities. In: Proc. 2nd Int. Symp. on Advances in Intelligent Data Analysis, pp. 1–14. Springer, Berlin (1997)
6. Kepler, J.: Astronomia Nova, aitioletos seu physica coelestis, tradita commentariis de motibus stellae martis, ex observationibus Tychonis Brahe. (New Astronomy, Based upon Causes, or Celestial Physics, Treated by Means of Commentaries on the Motions of the Star Mars, from the Observations of Tycho Brahe) (1609); English edition: New Astronomy. Cambridge University Press, Cambridge (1992)
7. Kepler, J.: Harmonices Mundi (1619); English edition: The Harmony of the World. American Philosophical Society, Philadelphia (1997)
8. Kurgan, L.A., Musilek, P.: A survey of knowledge discovery and data mining process models. Knowl. Eng. Rev. **21**(1), 1–24 (2006)
9. Marban, O., Menasalvas, E., Fernandez-Baizan, C.: A cost model to estimate the effort of data mining process (DMCoMo). Inf. Syst. **33**, 133–150 (2008)

²The workflows discussed in this book are available for download at the book's website.

Chapter 2

Practical Data Analysis: An Example

Before talking about the full-fledged data analysis process and diving into the details of individual methods, this chapter demonstrates some typical pitfalls one encounters when analyzing real-world data. We start our journey through the data analysis process by looking over the shoulders of two (pseudo) data analysts, Stan and Laura, working on some hypothetical data analysis problems in a sales environment. Being differently skilled, they show how things should and should not be done. Throughout the chapter, a number of typical problems that data analysts meet in real work situations are demonstrated as well. We will skip algorithmic and other details here and only briefly mention the intention behind applying some of the processes and methods. They will be discussed in depth in subsequent chapters.

2.1 The Setup

Disclaimer The data and the application scenario used in this chapter are fictional. However, the underlying problems are motivated by actual problems which are encountered in real-world data analysis scenarios. Explaining particular applicational setups would have been entirely out of the scope of this book, since in order to understand the actual issue, a bit of domain knowledge is often helpful if not required. Please keep this in mind when reading the following. The goal of this chapter is to show (and sometimes slightly exaggerate) pitfalls encountered in real-world data analysis setups and not the reality in a supermarket chain. We are painfully aware that people familiar with this domain will find some of the encountered problems strange, to say the least. Have fun.

The Data For the following examples, we will use an artificial set of data sources from a hypothetical supermarket chain. The data set consists of a few tables, which have already been extracted from an in-house database:¹

¹Often just getting the data is a problem of its own. Data analysis assumes that you have access to the data you need—an assumption which is, unfortunately, frequently not true.

- **Customers:** data about customers, stemming mostly from information collected when these customers signed up for frequent shopper cards.
- **Products:** A list of products with their categories and prices.
- **Purchases:** A list of products together with the date they were purchased and the customer card ID used during checkout.

The Analysts Stan and Laura are responsible for the analytics of the southern and northern parts, respectively, of a large supermarket chain. They were recently hired to help better understand customer groups and behavior and try to increase revenue in the local stores. As is unfortunately all too common, over the years the stores have already begun all sorts of data acquisition operations, but in recent years quite a lot of this data has been merged—however, still without a clear picture in mind. Many other stores had started to issue frequent shopping cards, so the directors of marketing of the southern and northern markets decided to launch a similar program. Lots of data have been recorded, and Stan and Laura now face the challenge to fit existing data to the questions posed. Together with their managers, they have sat down and defined three data analysis questions to be addressed in the following year:

- differentiate the different customer groups and their behavior to better understand their impact on the overall revenue,
- identify connections between products to allow for cross selling campaigns, and
- help design a marketing campaign to attract core customers to increase their purchases.

Stan is a representative of the typical self-taught data analysis newbie with little experience on the job and some more applied knowledge about the different techniques, whereas Laura has some training in statistics, data processing, and data analysis process planning.

2.2 Data Understanding and Pattern Finding

The first analysis task is a standard data analysis setup: customer segmentation—find out which types of customers exist in your database and try to link them to the revenue they create. This can be used later to care for clientele that are responsible for the largest revenue source or foster groups of customers who are under-represented. Grouping (or *clustering*) records in a database is the predominant method to find such customer segments: the data is partitioned into smaller subsets, each forming a more coherent group than the overall database contains. We will go into much more detail on this type of data analysis methods in Chap. 7. For now it suffices to know that some of the most prominent clustering methods return one typical example for each cluster. This essentially allows us to reduce a large data set to a small number of representative examples for the subgroups contained in the database.

Table 2.1 Stan's clustering result

Cluster-id	Age	Customer revenue
1	46.5	€ 1,922.07
2	39.4	€ 11,162.20
3	39.1	€ 7,279.59
4	46.3	€ 419.23
5	39.0	€ 4,459.30

The Naive Approach Stan quickly jumps onto the challenge, creates a dump of the database containing customer purchases and their birth date, and computes the age of the customers based on their birth date and the current day. He realizes that he is interested in customer clusters and therefore needs to somehow aggregate the individual purchases to their respective “owner.” He uses an aggregating operator in his database to compute the total price of the shopping baskets for each customer. Stan then applies a well-known clustering algorithm which results in five prototypical examples, as shown in Table 2.1.

Stan is puzzled—he was expecting the clustering algorithm to return reasonably meaningful groups, but this result looks as if all shoppers are around 40–50 years old but spend vastly different amount of money on products. He looks into some of the customers’ data in some of these clusters but cannot seem to find any interesting relations or any reason why some seem to buy substantially more than others. He changes some of the algorithm’s settings, such as the number of clusters created, but the results are similarly uninteresting.

The Sound Approach Laura takes a different approach. Routinely she first tries to understand the available data and validates that some basic assumptions are in fact true. She uses a basis data summarization tool to report the different values for the string attributes. The distribution of first names seems to match the frequencies she would expect. Names such as “Michael” and “Maria” are most frequent, and “Rosemarie” and “Anneliese” appear a lot less often. The frequencies of the occupations also roughly match her expectations: the majority of the customers are employees, while the second and third groups are students and freelancers, respectively. She proceeds to checking the attributes holding numbers. In order to check the age of the customers, she also computes the customers’ ages from their birth date and checks minimum and maximum. She spots a number of customers who obviously reported a wrong birthday, because they are unbelievably young. As a consequence, she decides to filter the data to only include people between the ages of 18 and 100. In order to explore the data more quickly, she reduces the overall customer data set to 5,000 records by random sampling and then plots a so-called histogram, which shows different ranges of the attribute *age* and how many customers fall into that range. Figure 2.1 shows the result of this analysis.

This view confirms Laura’s assumptions—the majority of shoppers is middle aged, and the number of shoppers continuously declines toward higher age groups.

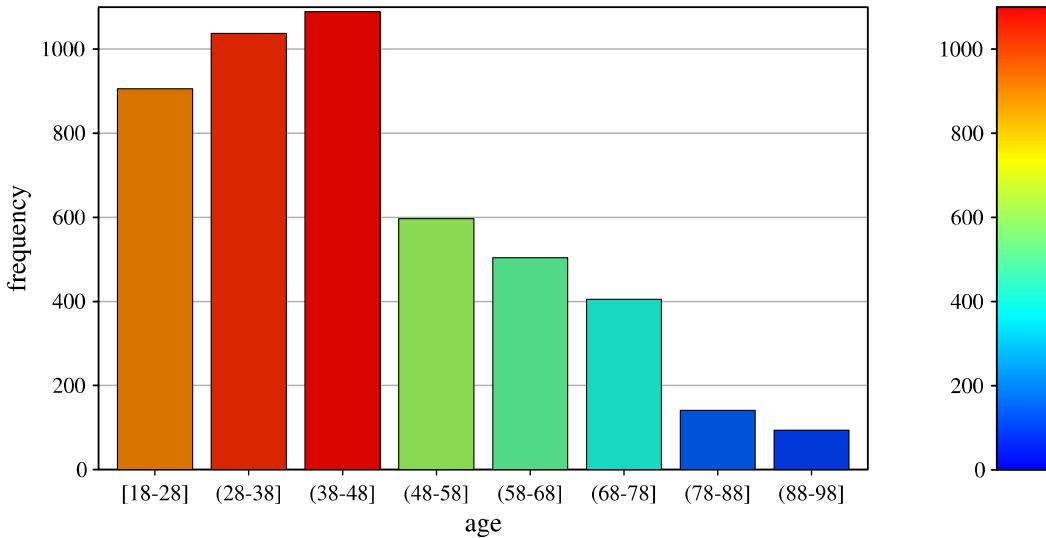


Fig. 2.1 A histogram for the distribution of the value of attribute *age* using 8 bins

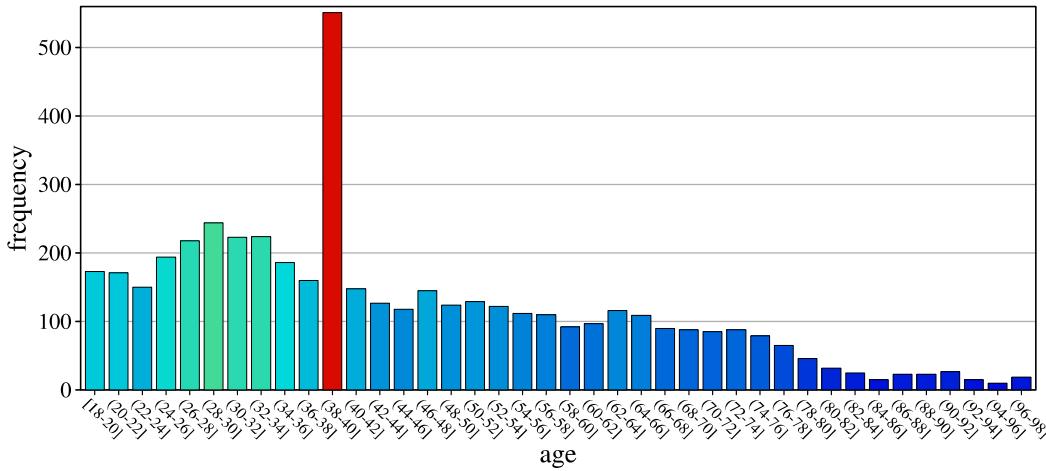


Fig. 2.2 A histogram for the distribution of the value of attribute *age* using 40 bins

She creates a second histogram to better inspect the subtle but strange cliff at around age 48 using finer setting for the bins. Figure 2.2 shows the result of this analysis.

Surprised, she notices the huge peak in the bin of ages 38–40. She discusses this observation with colleagues and the administrator of the shopping card database. They have no explanation for this odd concentration of 40-year-old people either. After a few other investigations, a colleague of the person who—before his retirement—designed the data entry forms suspects that this may have to do with the coding of missing birth dates. And, as it turns out, this is in fact the case: forms where people entered no or obviously nonsensical birth dates were entered into the form as zero values. For technical reasons, these zeros were then converted into the Java 0-date which turns out to be January 1, 1970. So these people all turn up with the same birth date in the customer database and in turn have the same age after the

Table 2.2 Laura's clustering result

Cluster	Age	Avg. cart price	Avg. purchases/month
1	75.3	€ 19.-	5.6
2	42.1	€ 78.-	7.8
3	38.1	€ 112.-	9.3
4	30.6	€ 16.-	4.8
5	44.7	€ 45.-	3.7

conversion Laura performed initially. Laura marks those entries in her database as “missing” in order to be able to distinguish them in future analyses.

Similarly, she inspects the shopping basket and product database and cleans up a number of other outliers and oddities. She then proceeds with the customer segmentation task. As in her previous data analysis projects, Laura first writes down her domain knowledge in form of a cognitive map, indicating relationships and dependencies between the attributes of her database. Having thus recalled the interactions between the variables of interest, she is well aware that the length of customer’s history and the number of overall shopping trips affect the overall basket price, and so she settles on the average basket price as a better estimator for the value of a particular customer. She considers also distinguishing the different product categories, realizing that those, of course, also potentially affect the average price. For the first step, she adds the average number of purchases per month, another indicator for the revenue a customer brings in. Data aggregation is now a bit more complex, but the modern data analysis tool she is using allows her to do the required joining and pivoting operations effortlessly. Laura knows that clustering algorithms are very sensitive to attributes with very different magnitudes, so she normalizes the three attributes to make sure they all three contribute equally to the clustering result. Running the same clustering algorithm that Stan was using, with the same setting for the number of clusters to be found, she gets the result shown in Table 2.2.

Obviously, there is a cluster (#1) of older customers who have a relatively small average basket price. There is also another group of customers (#4) which seems to correlate to younger shoppers, also purchasing smaller baskets. The middle-aged group varies wildly in price, however. Laura realizes that this matches her assumption about family status—people with families will likely buy more products and hence combine more products into more expensive baskets, which seems to explain the difference between clusters #2/#3 and cluster #5. The latter also seem to shop significantly less often. She goes back and validates some of these assumptions by looking at shopping frequency and average basket size as well and also determines the overall impact on store revenues for these different groups. She finally discusses these results with her marketing and campaign specialists to develop strategies to foster the customer groups which bring in the largest chunk of revenue and develop the ones which seem to be under-represented.

2.3 Explanation Finding

The second analysis goal is another standard shopping basket analysis problem: find product dependencies in order to better plan campaigns.

The Naive Approach Stan recently read in a book on practical data analysis how association rules can find arbitrary such connections in market basket data. He runs the association rule mining algorithm in his favorite data analysis tool with the default settings and inspects the results. Among the top-ranked generated rules, sorted by their confidence, Stan finds the following output:

```
'foie gras' (p1231) <- 'champagne Don Huberto' (p2149),
  'truffle oil de Rossini' (p578) [s=1E-5, c=75%]
'Tortellini De Cecco 500g' (p3456)
  <- 'De Cecco Sugo Siciliana' (p8764) [s=1E-5, c=60%]
```

He quickly infers that this representation must mean that foie gras is bought whenever champagne and truffle oil are bought together and similarly for the other rule. Stan knows that the confidence measure c is important, as it indicates the strength of the dependency (the first rule holds in 3 out of 4 cases). He considers the second measure of frequency s to be less important and deliberately ignores its fairly small value. The two rules shown above are followed by a set of other, similarly luxury/culinary product-oriented rules. Stan concludes that luxury products are clearly the most important products on the shelf and recommends to his marketing manager to launch a campaign to advertise some of the products on the right side of these rules (champagne, truffle oil) to increase the sales of the left side (foie gras). In parallel, he increases orders for these products, expecting a recognizable increase in sales. He proudly sends the results of his analysis to Laura.

The Sound Approach Laura is puzzled by those nonintuitive results. She reruns the analysis and notices the support values of the rules extracted by Stan—some of the rules Stan extracted have indeed a remarkably high confidence, and some do almost forecast shopping behavior. However, they have very low support values, meaning that only a small number of shopping baskets containing the products were ever observed. The rules that Stan found are not representative at all for his customer base. To confirm this, she runs a quick query on her database and sees that, indeed, there is essentially no influence on the overall revenue.

She notices that the problem of low support is caused by the fact that Stan ran the analysis on product IDs, so in effect he was forcing the rules to differentiate between brands of champagne and truffle oil. She reruns the analysis based on the product categories instead, ranks them by a mix of support and confidence, and finds a number of association rules with substantially higher support:

```
tomatoes <- capers, pasta [s=0.007, c=32%]
tomatoes <- apples [s=0.013, c=22%]
```

Laura focuses on rules with a much higher support measure s than before and also realizes that the confidence measure c is significantly higher than one would expect

by chance. The first rule seems to be triggered by a recent fashion of Italian cooking, whereas the apple/tomato-rule is a known aspect.

However, she is still irritated by one of the rules discovered by Stan, which has a higher than suspected confidence despite a relatively low support. Are there some gourmets among the customers who prefer a very specific set of products? Rerunning this analysis on the shopping card owners yields almost the same results, so the (potential) gourmets appear among their regular customers. Just to be sure, she inspects how many different customers (resp. shopping cards) occur for baskets that support this rule. As she had conjectured, there is a very limited number of customers that seem to have a strong affection for these products. Those few customers have bought this combination frequently, thus inflating the overall support measure (which refers to shopping baskets). This means that the support in terms of the *number of customers* is even smaller than the support in terms of *number of shopping baskets*. The response to any kind of special promotion would fall even shorter than expected from Stan's rule.

Apparently the time period in which the analyzed data has been collected influences the results. Thinking about it, she develops an idea how to learn about changes in the customers shopping behavior: She identifies a few rules, some rather promising other well-known facts, and decides to monitor those combinations on a regular basis (say quarterly). She got to know that a chain of liquor stores will soon open a number of shops close to the own markets, so she picks some rules with beverages in their conclusion part to see if the opening has any impact on the established shopping patterns of the own customers. As she fears a loss of potential sales, she plans a comparison of rules obtained not only over time but also among markets in the vicinity of such stores versus the other markets. She wonders whether promoting the products in the rule's antecedent may help to bring back the customer and decides to discuss this with the marketing&sales team to determine if and where appropriate campaigns should be launched, once she has the results of her analysis.

2.4 Predicting the Future

The third and final analysis goal we consider in this brief overview is a forecasting or prediction problem. The idea is to find some relationship in our existing data that can help us to predict if and how customers will react to coupon mailings and how this will affect our future revenue.

The Naive Approach Stan believes that no detailed analysis is required for this problem and notices that it is fairly straightforward to monitor success. He has seen at a competitor how discount coupons attract customers to purchase additional products. So he suggests launching a coupon campaign that gives customers a discount of 10% if they purchase products for more than €50. This coupon is mailed to all customers on record. Throughout the course of the next month, he carefully monitors his database and is positively surprised when he sees that his campaign is obviously

working: the average price of shopping baskets is going up in comparison with previous months. However, at the end of the quarter he is shocked to see that overall revenues for the past quarter actually fell. His management is finally fed up with the lack of performance and fires Stan.

The Sound Approach Laura, who is promoted to head of analytics for the northern and southern super market chain first cancels Stan's campaign and looks into the underlying data. She quickly realizes that even though quite a number of customers did in fact use the coupons and increased their shopping baskets, their average number of baskets per month actually went down—so quite a number of people seem to have simply combined smaller shopping trips to be able to benefit from the discount offer. However, for some shoppers, the combined monthly shopping basket value did go up markedly, so there might be value here. Laura wonders how she can discriminate between those customers who simply use the coupons to discount their existing purchases and those who are actually enticed to purchase additional items. She notices that one of the earlier generated customer segments correlates better than others with the group of customers whose revenue went up—this fraction of customers is significantly higher than in the other groups. She considers using this very simple, manually designed predictor for a future campaign but wants to first make sure that she cannot do better with some smarter techniques. She decides that in the end it is not so important if she can actually understand the extracted model but only how well it performs.

To provide good starting points for the modeling technique, she decides to generate a few potentially informative attributes first. Models that rely on thousands of details typically perform poor, so providing how often every product has been bought by the customer in the last month is not an option for her. To get robust models, she wants to aggregate the tiny bits of information, but what kind of aggregation could be helpful? She returns to her cognitive map to review the dependencies. One aspect is the availability of competitors: She reckons that customers may have alternative (possibly specialized) markets nearby but have been attracted by the coupon this time, keeping them away from the competitors. She decides to aggregate the money spent by the customer per month for a number of product types (such as beverages, thinking of the chain of liquor stores again). She conjectures that customers that perform well on average, but underperform in a specific segment only, may be enticed by the coupon to buy products for the underperforming segment also. Providing the segment performance before and after Stan's campaign should help a predictor to detect such dependencies if they exist.

The cognitive map brings another idea into her mind: people who appreciate the full assortment but live somewhat further away from the own stores may see the coupon as a kind of travel compensation. So she adds a variable expressing a coarse estimation of the distance between the customer home and the nearest available market (which is only possible for the shopping card owners). She continues to use her cognitive map to address many different aspects and creates attributes that may help to verify her hypotheses. She then investigates the generated attributes visually and also technically by means of feature selection methods.

After selecting the most promising attributes, she trains a classifier to distinguish the groups. She uses part of the data to simulate an independent test scenario and thereby evaluates the expected impact of a campaign—are the costs created by sending coupons to customers who do not purchase additional products offset by customers buying additional items? After some additional model fine tuning, she reaches satisfactory performance. She discusses the results with the marketing&sales team and deploys the prediction system to control the coupon mailings for the next quarter. She keeps monitoring the performance of these coupon campaigns over future quarters and updates her model sporadically.

2.5 Concluding Remarks

In this chapter we have, very briefly and informally, touched upon a number of issues data analysts may encounter while making sense of real-world data. Many other problems can arise, and many more methods for data analysis exist in the academic literature and in real-world data analysis tools. We will attempt at covering the most prominent and most often used examples in the following chapters.

Note that one of the biggest problems data analysts very often have is that the data they get is not suited to answer the questions they are asked. For instance, if we were supposed to use the data in our customer database to find out how to differentiate Asian shopping behavior from European, we would have a very hard time. This data can only be used to distinguish between different types of European shoppers because it contains data from European markets only. Note also that we are (why ever) assuming that we used a nice, representative sample of all different types of European shoppers to generate the data—very often this is not the case, and the data itself is already biased and will bias our analysis results—in this example we could be heavily biased by the type of supermarket chain we used to record the data in the first place. An upscale delicatessen supermarket will have dramatically different shopping patterns than the low-scale discounter. We will be discussing these points later in more depth as well.

Chapter 3

Project Understanding

We are at the beginning of a series of interdependent steps, where the project understanding phase marks the first. In this initial phase of the data analysis project, we have to map a problem onto one or many data analysis tasks. In a nutshell, we conjecture that the nature of the problem at hand can be adequately captured by some data sets (that still have to be identified or constructed), that appropriate modeling techniques can successfully be applied to learn the relationships in the data, and finally that the gained insights or models can be transferred back to the real case and applied successfully. This endeavor relies on a number of assumptions and is threatened by several risks, so the goal of the project understanding phase is to assess the main objective, the potential benefit, as well as the constraints, assumptions, and risks. While the number of data analysis projects is rapidly expanding, the failure rate is still high, so this phase should be carried out seriously to rate the chances of success realistically. The project understanding phase should be carried out with care to keep the project on the right track.

We have already sketched the data analysis process (CRISP-DM in Sect. 1.2). There is a clear order in the steps in the sense that for a later step, all precedent steps must have been executed. However, this does not mean that we can run once through all steps to deterministically achieve the desired results. There are many options and decisions to be made. Most of them will rely on our (subjective and dynamic) understanding of the problem at hand. The line of argument will not always be from an earlier phase to a later one. For instance, if a regression problem has to be solved, the analyst may decide that a certain method seems to be a promising choice for the modeling phase. From the characteristics of this technique he knows that all input data have to be transformed into numerical data, which has to be carried out beforehand (data preparation phase). This requires a careful look at the multivalued ordinally scaled attributes already in the data understanding phase to see how the order of the values is best preserved. If it is not considered in time, it may happen that later, in the evaluation phase, it turns out that the project owner expected to gain insights into the input–output relationship rather than having a black-box model only. If the analyst had considered this requirement beforehand, he might have chosen a different method. Changing this decision at any point later than in this initial

Table 3.1 Problems faced in data analysis projects, excerpt from [1]

Problem source	Project owner perspective	Analyst perspective
Communication	Project owner does not understand the technical terms of the analyst	Analyst does not understand the terms of the domain of the project owner
Lack of understanding	Project owner was not sure what the analyst could do or achieve	Analyst found it hard to understand how to help the project owner
	Models of analyst were different from what the project owner envisioned	
Organization	Requirements had to be adopted in later stages as problems with the data became evident	Project owner was an unpredictable group (not so concerned with the project)

project understanding phase often renders some (if not most) of the earlier work in data understanding, data preparation, and modeling useless. While the time spent on project and data understanding compared to data preparation and modeling is small (20% : 80%), the importance to success is just the opposite (80% : 20%) [4].

3.1 Determine the Project Objective

As a first step, a *primary objective* (not a long list but one or two) and some *success criteria* in terms of the project domain have to be determined (who will decide which results are desired and whether the original project goal was achieved or not). This is much easier said than done, especially if the analysis is not carried out by the domain expert himself. In such cases the project owner and the analyst *speak different languages* which may cause misunderstandings and confusion. In the worst case, the communication problems lead to very soft project goals, just vague enough to allow every stakeholder seeing his own perspective somehow accounted for. At the end, all of a sudden, the stakeholders recognize that the results do not fit their expectations. The challenge here is usually not a matter of technical but of communicative competence.

Table 3.1 shows some typical problems occurring in such projects. To overcome language confusion, a glossary of terms, definition, acronyms, and abbreviations is inevitable. Knowing the terms still does not imply an understanding of the project domain, the objectives, constraints, and influencing factors. One interviewing technique that may help to get most out of the expert is to rephrase all of her statements, which often provokes additional relativizing statements. Another technique is to use explorative tools such as mind maps or cognitive maps to sketch beliefs, experiences, and known factors and how they influence each other.

An example of a **cognitive map** in the shopping domain considered in Sect. 2 is given in Fig. 3.1. Each node of this graph represents a property of the considered product or the customer. The variable of interest is placed in the center: how often

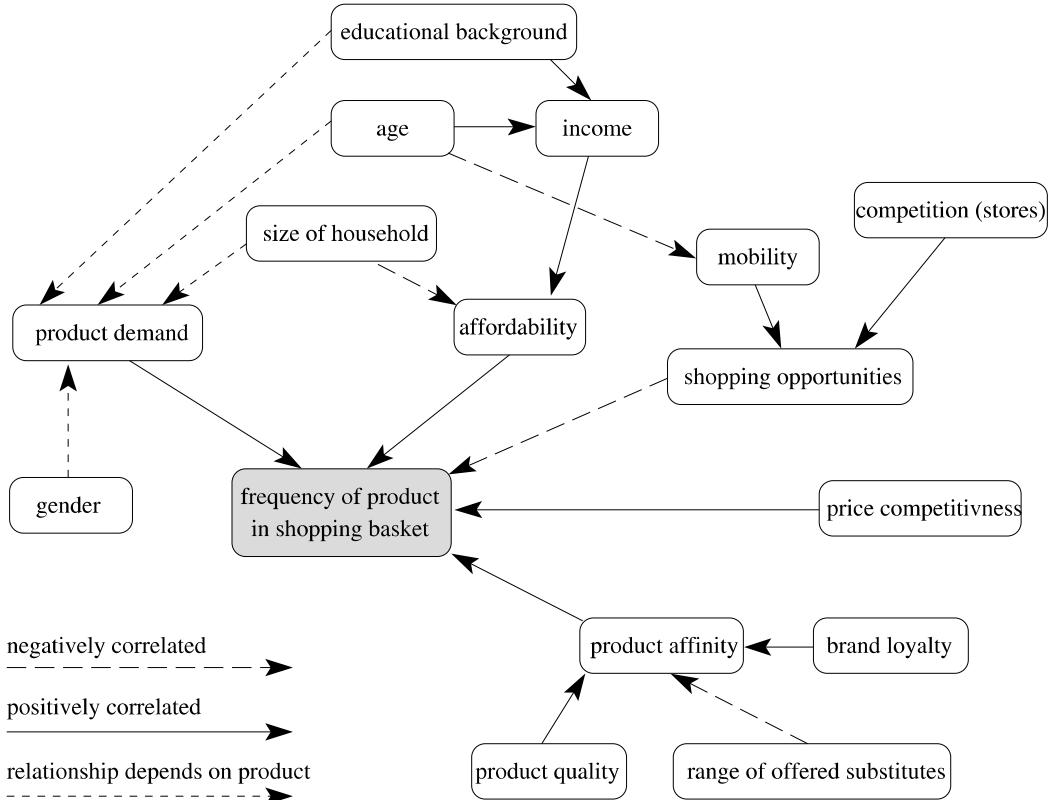


Fig. 3.1 A cognitive map for the shopping domain: How often will a certain product occur in a shopping basket of some customer? The positive correlation between *income* and *affordability* reads like *the higher the income, the higher the affordability*, whereas an example of a negative correlation reads like *the broader the range of offered substitutes, the lower the product affinity*

will a certain product be found in the shopping basket of the customer? This depends on various factors, which are placed around this node. The direction of influence is given by the arrows, and the line style indicates the way how the variables influence each other: The higher the customer's affinity to the product, the more often it will be found in the basket. The author of the cognitive map conjectures that the product affinity itself is positively influenced by a high product quality and the customer's brand loyalty (a loyal customer is less likely to buy substitute products). On the other hand, the broader the range of offered substitutes, the more likely a customer may try out a different product. Other relationships depend on the product itself: The higher the demand of a certain product, the more often it will be found in the shopping basket, but the demand itself may, depending on the product, vary with gender (e.g., razor blades, hairspray), age (e.g., denture cleaner), or family status (e.g., napkins, fruit juice). The development of such a map supports the domain understanding and adjustment of expectations.

While constructing a cognitive map, a few rules should be adhered to: First, to keep the map clear, only direct dependencies should be included in the graph. For instance, the size of the household influences the target variable, but only indirectly via the generated product demand and the affordability, and therefore there is no direct connection from *size of household* to *frequency of product in shopping basket*.

Table 3.2 Clarifying the primary objectives

<u>Objective:</u>	Increase revenues (per campaign and/or per customer) in direct mailing campaigns by personalized offer and individual customer selection
<u>Deliverable:</u>	Software that automatically selects a specified number of customers from the database to whom the mailing shall be sent, runtime max. half-day for database of current size
<u>Success criteria:</u>	Improve order rate by 5% or total revenues by 5%, measured within 4 weeks after mailing was sent, compared to rate of last three mailings

Secondly, the labels of the nodes should be chosen carefully, so that they are easily interpretable when plugged into the relationship templates such as *the higher ...*, *the higher ...*. As an example, the node *size of household* could have been named *family status*, but then it is not quite clear what *the more family status ...* actually means.

Once an understanding of the domain has been achieved, the problem and primary objective have to be identified (see Table 3.2). Again, it is often useful to discuss or model the current solution first, for instance, by using techniques from software engineering (business process modeling, UML use cases, etc.) [3]. When the current solution has been elaborated, its advantages and disadvantages can be explored and discussed. Often, the primary objective is assumed to be known beforehand, probably the project would not have been initiated without having identified a problem first. But as there are many different ways to attack a problem, the objective should be precise about the direction to follow. A general statement about the goal is easily made (“model the profitable customers to increase the sales”), but it is often not precise enough (how do we precisely identify a profitable customer?) and not actionable (how exactly shall this model help to increase the sales?). To render the objective more precise, it is necessary to sketch the target use already at this early stage. Thus it becomes clear what kind of result has to be delivered, which may range from a written technical report with interesting findings to a user-friendly software that uses the final model to automatize decisions.

From the perspective of the project owner some of these elaborate steps may appear unnecessary—they master their domain already, after all. However, these steps must be considered as a preparation of the closely linked data understanding phase (see next section). All the identified factors, situations, and relationships that are assumed to be relevant must be present and recognizable in the data. If they cannot be found in the data, either there is a misconception in the project understanding or (even worse) the data is not substantial or detailed enough to reflect the important relations of the real-world problem. In both cases, it would be fatal to miss this point and proceed unworried.

3.2 Assess the Situation

The next step is to estimate the chances of a successful data analysis project. This requires the review of available resources, requirements, and risks. The most impor-

tant resources are data and knowledge, that is, databases and experts who can provide background information (about the domain in general and about the databases in particular). Besides a plain listing of databases and personnel, it is important to clarify the access to both: if the data is stored in an operative system, mining the data may paralyze the applications using it. To become independent, it is advisable to provide a database dump. Experts are typically busy and difficult to grasp—but an inaccessible knowledge source is useless. A sufficiently large number of time slots for meetings should be arranged.

Based on the domain exploration (cognitive map, business process model, etc.), a list of explicit and implicit assumptions and risks is created to judge the chances of a successful project and guide the next steps. Data analysis lives on data. This list shall help to convince ourselves that the data is meaningful and relevant to the project. Why should we undertake this effort? We will see whether we can build a model from this data later anyway. Unfortunately, this is only half of the truth. After reviewing a number of reports in a data analysis competition, Charles Elkan noted that “when something surprising happens, rather than question the expectations, people typically believe that they should have done something slightly different” [2]. Expecting that the problem can be solved with the given data may lead to continuously changing and “optimizing” the model—rather than taking the possibility into account that the data is not appropriate for this problem. In order to avoid this pitfall, the conjectured relations and expert-proven connections can help us in verifying that the given data satisfy our needs—or to put forward good reasons why the project will probably fail. This is particularly important as in many projects the available data have not been collected to serve the purpose that is intended now. To prevent us from carrying out an expensive project having almost no prospect of success, we have to carefully track all assumptions and verify them as soon as possible. Typical requirements and assumptions include:

- requirements and constraints
 - *model requirements*,
e.g., model has to be explanatory (because decisions must be justified clearly)
 - *ethical, political, legal issues*,
e.g., variables such as gender, age, race must not be used
 - *technical constraints*,
e.g., applying the technical solution must not take more than n seconds
- assumptions
 - *representativeness*:
If conclusions about a specific target group are to be derived, a sufficiently large number of cases from this group must be contained in the database, and the sample in the database must be representative for the whole population.
 - *informativeness*:
To cover all aspects by the model, most of the influencing factors (identified in the cognitive map) should be represented by attributes in the database.
 - *good data quality*:
The relevant data must be of good quality (correct, complete, up-to-date) and unambiguous thanks to the available documentation.

- *presence of external factors:*

We may assume that the external world does not change constantly—for instance, in a marketing project we may assume that the competitors do not change their current strategy or product portfolio at all.

Every assumption inherently represents a risk (there might be other risks though). If possible, a contingency should be sketched in case the assumption turns out to be invalid, including options such as the acquisition of additional data sources.

3.3 Determine Analysis Goals

Finally, the primary objective must be transformed into a more technical data mining goal. An architecture for the envisaged solution has to be found, composed out of building blocks as discussed in Sect. 1.3 (data analysis tasks). For instance, this architecture might contain a component responsible for grouping the customers according to some readily available attributes first, another component finds interesting deviating subgroups in each of the groups, and a third component predicts some variable of interest based on the customer data and the membership to the respective groups and subgroups. The better this architecture fits the actual situation, the better the chances of finding a model class that will prove successful in practice. To achieve this analogy, the discussions about the project domain are of great help.

Again there is the danger of accepting a reasonable architecture quickly, underestimating or even ignoring the great impact on the overall effort. Suppose that a company wants to increase the sales of some high-end product by direct mailing. One approach is to develop a model that predicts who will buy this product using the company’s own customer database. Such a model might be interesting to interpret (useful for a report), but if it is used to decide to whom a mailing should be sent, most of the customers may have the product already (within the same customer database). Applying the model to people not being in the database is impossible as we lack the information about them that is needed by the model. The predictive model may also find out that customers buying the product were loyal customers for many years—but *artificially* increasing the duration of the customer relationship to support the purchase of the product is unfortunately impossible. If a foreseeable result is ignored or a misconception w.r.t. the desired use of the model is not recognized, considerable time may be wasted with building a correct model that turns out to be useless in the end.

For each of the building blocks, we can select a model class and technique to derive a model of this class automatically from data. There is nothing like *the unique best method for predictive tasks*, because they all have their individual weaknesses and strengths and it is impossible to combine all their properties or remove all biases (see Chap. 5). Although the final decision about the modeling technique will be made in the modeling phase, it should be clear already at this point of the analysis which properties the model should have and why. The methods and tools optimize the technical aspects of the model quality (such as accuracy, see also Chap. 5). Other

aspects are often difficult to formalize and thus to optimize (such as interestingness or interpretability), so that the choice of the model class has the greatest influence on these properties. Desirable properties may be, for instance:

- **Interpretability:**

If the goal of the analysis is a report that sketches possible explanations for a certain situation, the ultimate goal is to understand the delivered model. For some *black-box models*, it is hard to comprehend how the final decision is made, and their model lacks interpretability.

- **Reproducibility/stability:**

If the analysis is carried out more than once, we may achieve similar performance—but not necessarily similar models. This does no harm if the model is used as a black box, but hinders a direct comparison of subsequent models to investigate their differences.

- **Model flexibility/adequacy:**

A flexible model can adapt to more (complicated) situations than an inflexible model, which typically makes more assumptions about the real world and requires less parameters. If the problem domain is complex, the model learned from data must also be complex to be successful. However, with flexible models the risk of overfitting increases (will be discussed in Chap. 5).

- **Runtime:**

If restrictive runtime requirements are given (either for building or applying the model), this may exclude some computationally expensive approaches.

- **Interestingness and use of expert knowledge:**

The more an expert already knows, the more challenging it is to “surprise” him or her with new findings. Some techniques looking for associations (see Sect. 7.6) are known for their large number of findings, many of them redundant and thus uninteresting. So if there is a possibility of including any kind of previous knowledge, this may ease the search for the best model considerably on the one hand and may prevent us from rediscovering too many well-known artifacts.

When discussing the various modeling techniques in Chaps. 7–9, we will give hints which properties they possess. The final choice is then up to the analyst.

3.4 Further Reading

The books by Dorian Pyle [4, 5] offer many suggestions and constructive hints for carrying out the project understanding phase. [5] contains a step-by-step workflow for business understanding and data mining consisting of various *action boxes*. An organizationally grounded framework to formally implement the business understanding phase of data mining projects is presented in [6]. In [1] a template set for educating and documenting project requirements is proposed.