

---

# Texts in Computer Science

## Series Editors

David Gries

Orit Hazzan

Titles in this series now included in the Thomson Reuters Book Citation Index! ‘Texts in Computer Science’ (TCS) delivers high-quality instructional content for undergraduates and graduates in all areas of computing and information science, with a strong emphasis on core foundational and theoretical material but inclusive of some prominent applications-related content. TCS books should be reasonably self-contained and aim to provide students with modern and clear accounts of topics ranging across the computing curriculum. As a result, the books are ideal for semester courses or for individual self-study in cases where people need to expand their knowledge. All texts are authored by established experts in their fields, reviewed internally and by the series editors, and provide numerous examples, problems, and other pedagogical tools; many contain fully worked solutions. The TCS series is comprised of high-quality, self-contained books that have broad and comprehensive coverage and are generally in hardback format and sometimes contain color. For undergraduate textbooks that are likely to be more brief and modular in their approach, require only black and white, and are under 275 pages, Springer offers the flexibly designed Undergraduate Topics in Computer Science series, to which we refer potential authors.

More information about this series at  
[www.springer.com/series/3191](http://www.springer.com/series/3191)

---

Michael R. Berthold • Christian Borgelt •  
Frank Höppner • Frank Klawonn •  
Rosaria Silipo

# Guide to Intelligent Data Science

How to Intelligently Make Use  
of Real Data

Second Edition

 Springer

Michael R. Berthold  
Department of Computer and Information  
Science  
University of Konstanz  
Konstanz, Germany

Frank Höppner  
Department of Computer Science  
Ostfalia University of Applied Sciences  
Wolfenbüttel, Germany

Frank Klawonn  
Department of Computer Science  
Ostfalia University of Applied Sciences  
Wolfenbüttel, Germany

Christian Borgelt  
Department of Computer Sciences  
University of Salzburg  
Salzburg, Austria

Rosaria Silipo  
KNIME AG  
Zurich, Switzerland

*Series Editors*

David Gries  
Department of Computer Science  
Cornell University  
Ithaca, NY, USA

Orit Hazzan  
Faculty of Education in Technology and  
Science  
Technion – Israel Institute of Technology  
Haifa, Israel

ISSN 1868-0941  
Texts in Computer Science  
ISBN 978-3-030-45573-6  
<https://doi.org/10.1007/978-3-030-45574-3>

ISSN 1868-095X (electronic)  
ISBN 978-3-030-45574-3 (eBook)

© Springer Nature Switzerland AG 2010, 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

---

## Preface

The interest in making sense of data has become central to pretty much every business and project in the past decade. When we wrote the first edition of this volume, it was aimed at the people whose task it was to analyze real world data—often in a side department, prepared to solve problems given to them. In the meantime, data science use has spread from those isolated groups to the entire organization. This fits our goal from the first edition even more: providing different levels of detail for each section of the book. Readers more interested in the concepts and how to apply methods, focus more on the first half of each chapter. Readers, also interested in the underpinnings of the presented methods and algorithms, find the theory in the middle. And for both audiences, the end of most chapters shows practical examples of how to apply these to real data.

Still, given the advance in research we decided it worth revising, updating, and expanding the content to also cover select new methods and tools that have been invented in the meantime. We also adjusted to the new way of thinking about data analysis and how many organizations now see this in the bigger context of Data Science, as it is visible by the many universities adding Data Science curriculum or even entire Data Science departments. As the original volume covered the entire process from data ingestion to deployment and management (which we expanded substantially in this edition), we decided to also follow this trend and use Data Science as the overarching umbrella term. This also made it obvious that having someone else with more real world data science experience would be beneficial so we were glad when Rosaria agreed to share the (re)writing duty and add decades of industrial data science experience to the team.

And finally we decided to put the focus of the practical exercises more on KNIME Analytics Platform—not because there is only one tool out there but because visual workflows are easier to explain and therefore lend themselves to be included in printed material to illustrate how the methods are being used in practice. On the book’s website: [www.datascienceguide.org](http://www.datascienceguide.org) you can still find the R examples from the first edition and we have also added examples in Python as well as all of the KNIME workflows described in the book. We are also providing teaching material and will continuously update the site in the coming years.

There are many people to be thanked, and we will not attempt to list them all. However, Iris Adä and Martin Horn deserve mentioning for all their help with the first edition. For the second round, we owe thanks to Satoru Hayasaka, Kathrin Melcher, and Emilio Silvestri who have spent many hours proof reading and updating/creating workflows.

Konstanz, Germany  
Salzburg, Austria  
Braunschweig, Germany  
Zurich, Switzerland

Michael R. Berthold  
Christian Borgelt  
Frank Höppner and Frank Klawonn  
Rosaria Silipo

---

# Contents

<b>1</b>	<b>Introduction</b>	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 Science	6
1.2	The Data Science Process	7
1.3	Methods, Tasks, and Tools	11
1.4	How to Read This Book	13
	References	14
<b>2</b>	<b>Practical Data Science: An Example</b>	15
2.1	The Setup	15
2.2	Data Understanding and Pattern Finding	16
2.3	Explanation Finding	19
2.4	Predicting the Future	21
2.5	Concluding Remarks	23
<b>3</b>	<b>Project Understanding</b>	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
<b>4</b>	<b>Data Understanding</b>	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	62
4.5	Outlier Detection	65

4.5.1	Outlier Detection for Single Attributes . . . . .	66
4.5.2	Outlier Detection for Multidimensional Data . . . . .	68
4.6	Missing Values . . . . .	69
4.7	A Checklist for Data Understanding . . . . .	72
4.8	Data Understanding in Practice . . . . .	73
4.8.1	Visualizing the Iris Data . . . . .	74
4.8.2	Visualizing a Three-Dimensional Data Set on a Two- Coordinate Plot . . . . .	82
	References . . . . .	82
<b>5</b>	<b>Principles of Modeling . . . . .</b>	<b>85</b>
5.1	Model Classes . . . . .	86
5.2	Fitting Criteria and Score Functions . . . . .	89
5.2.1	Error Functions for Classification Problems . . . . .	91
5.2.2	Measures of Interestingness . . . . .	93
5.3	Algorithms for Model Fitting . . . . .	93
5.3.1	Closed-Form Solutions . . . . .	93
5.3.2	Gradient Method . . . . .	94
5.3.3	Combinatorial Optimization . . . . .	96
5.3.4	Random Search, Greedy Strategies, and Other Heuristics . . . . .	96
5.4	Types of Errors . . . . .	100
5.4.1	Experimental Error . . . . .	102
5.4.2	Sample Error . . . . .	109
5.4.3	Model Error . . . . .	110
5.4.4	Algorithmic Error . . . . .	111
5.4.5	Machine Learning Bias and Variance . . . . .	111
5.4.6	Learning Without Bias? . . . . .	112
5.5	Model Validation . . . . .	112
5.5.1	Training and Test Data . . . . .	112
5.5.2	Cross-Validation . . . . .	114
5.5.3	Bootstrapping . . . . .	114
5.5.4	Measures for Model Complexity . . . . .	115
5.5.5	Coping with Unbalanced Data . . . . .	121
5.6	Model Errors and Validation in Practice . . . . .	121
5.6.1	Scoring Models for Classification . . . . .	122
5.6.2	Scoring Models for Numeric Predictions . . . . .	124
5.7	Further Reading . . . . .	125
	References . . . . .	125
<b>6</b>	<b>Data Preparation . . . . .</b>	<b>127</b>
6.1	Select Data . . . . .	127
6.1.1	Feature Selection . . . . .	128
6.1.2	Dimensionality Reduction . . . . .	133
6.1.3	Record Selection . . . . .	134
6.2	Clean Data . . . . .	136
6.2.1	Improve Data Quality . . . . .	136

---

6.2.2	Missing Values . . . . .	137
6.2.3	Remove Outliers . . . . .	139
6.3	Construct Data . . . . .	140
6.3.1	Provide Operability . . . . .	140
6.3.2	Assure Impartiality . . . . .	142
6.3.3	Maximize Efficiency . . . . .	144
6.4	Complex Data Types . . . . .	147
6.5	Data Integration . . . . .	148
6.5.1	Vertical Data Integration . . . . .	149
6.5.2	Horizontal Data Integration . . . . .	150
6.6	Data Preparation in Practice . . . . .	152
6.6.1	Removing Empty or Almost Empty Attributes and Records in a Data Set . . . . .	152
6.6.2	Normalization and Denormalization . . . . .	153
6.6.3	Backward Feature Elimination . . . . .	154
6.7	Further Reading . . . . .	155
	References . . . . .	155
<b>7</b>	<b>Finding Patterns . . . . .</b>	<b>157</b>
7.1	Hierarchical Clustering . . . . .	159
7.1.1	Overview . . . . .	160
7.1.2	Construction . . . . .	162
7.1.3	Variations and Issues . . . . .	164
7.2	Notion of (Dis-)Similarity . . . . .	167
7.3	Prototype- and Model-Based Clustering . . . . .	173
7.3.1	Overview . . . . .	174
7.3.2	Construction . . . . .	175
7.3.3	Variations and Issues . . . . .	178
7.4	Density-Based Clustering . . . . .	181
7.4.1	Overview . . . . .	181
7.4.2	Construction . . . . .	182
7.4.3	Variations and Issues . . . . .	184
7.5	Self-organizing Maps . . . . .	187
7.5.1	Overview . . . . .	187
7.5.2	Construction . . . . .	188
7.6	Frequent Pattern Mining and Association Rules . . . . .	189
7.6.1	Overview . . . . .	191
7.6.2	Construction . . . . .	192
7.6.3	Variations and Issues . . . . .	199
7.7	Deviation Analysis . . . . .	206
7.7.1	Overview . . . . .	206
7.7.2	Construction . . . . .	207
7.7.3	Variations and Issues . . . . .	210
7.8	Finding Patterns in Practice . . . . .	211
7.8.1	Hierarchical Clustering . . . . .	211



7.8.2	<i>k</i> -Means and DBSCAN . . . . .	211
7.8.3	Association Rule Mining . . . . .	214
7.9	Further Reading . . . . .	214
	References . . . . .	215
<b>8</b>	<b>Finding Explanations . . . . .</b>	<b>219</b>
8.1	Decision Trees . . . . .	220
8.1.1	Overview . . . . .	221
8.1.2	Construction . . . . .	222
8.1.3	Variations and Issues . . . . .	225
8.2	Bayes Classifiers . . . . .	230
8.2.1	Overview . . . . .	230
8.2.2	Construction . . . . .	231
8.2.3	Variations and Issues . . . . .	235
8.3	Regression . . . . .	241
8.3.1	Overview . . . . .	241
8.3.2	Construction . . . . .	243
8.3.3	Variations and Issues . . . . .	246
8.3.4	Two-Class Problems . . . . .	254
8.3.5	Regularization for Logistic Regression . . . . .	255
8.4	Rule learning . . . . .	258
8.4.1	Propositional Rules . . . . .	258
8.4.2	Inductive Logic Programming or First-Order Rules . . . . .	265
8.5	Finding Explanations in Practice . . . . .	267
8.5.1	Decision Trees . . . . .	267
8.5.2	Naïve Bayes . . . . .	268
8.5.3	Logistic Regression . . . . .	269
8.6	Further Reading . . . . .	270
	References . . . . .	271
<b>9</b>	<b>Finding Predictors . . . . .</b>	<b>273</b>
9.1	Nearest-Neighbor Predictors . . . . .	275
9.1.1	Overview . . . . .	275
9.1.2	Construction . . . . .	277
9.1.3	Variations and Issues . . . . .	279
9.2	Artificial Neural Networks . . . . .	282
9.2.1	Overview . . . . .	283
9.2.2	Construction . . . . .	286
9.2.3	Variations and Issues . . . . .	290
9.3	Deep Learning . . . . .	292
9.3.1	Recurrent Neural Networks and Long-Short Term Memory Units . . . . .	293
9.3.2	Convolutional Neural Networks . . . . .	295
9.3.3	More Deep Learning Networks: Generative-Adversarial Networks (GANs) . . . . .	296
9.4	Support Vector Machines . . . . .	297

9.4.1	Overview . . . . .	298
9.4.2	Construction . . . . .	302
9.4.3	Variations and Issues . . . . .	303
9.5	Ensemble Methods . . . . .	304
9.5.1	Overview . . . . .	304
9.5.2	Construction . . . . .	306
9.5.3	Variations and Issues . . . . .	309
9.6	Finding Predictors in Practice . . . . .	312
9.6.1	$k$ Nearest Neighbor (kNN) . . . . .	312
9.6.2	Artificial Neural Networks and Deep Learning . . . . .	312
9.6.3	Support Vector Machine (SVM) . . . . .	313
9.6.4	Random Forest and Gradient Boosted Trees . . . . .	314
9.7	Further Reading . . . . .	315
	References . . . . .	315
<b>10</b>	<b>Deployment and Model Management . . . . .</b>	<b>319</b>
10.1	Model Deployment . . . . .	319
10.1.1	Interactive Applications . . . . .	320
10.1.2	Model Scoring as a Service . . . . .	320
10.1.3	Model Representation Standards . . . . .	320
10.1.4	Frequent Causes for Deployment Failures . . . . .	321
10.2	Model Management . . . . .	322
10.2.1	Model Updating and Retraining . . . . .	323
10.2.2	Model Factories . . . . .	324
10.3	Model Deployment and Management in Practice . . . . .	324
10.3.1	Deployment to a Dashboard . . . . .	325
10.3.2	Deployment as REST Service . . . . .	326
10.3.3	Integrated Deployment . . . . .	327
	References . . . . .	328
<b>A</b>	<b>Statistics . . . . .</b>	<b>329</b>
A.1	Terms and Notation . . . . .	330
A.2	Descriptive Statistics . . . . .	331
A.2.1	Tabular Representations . . . . .	331
A.2.2	Graphical Representations . . . . .	332
A.2.3	Characteristic Measures for One-Dimensional Data . . . . .	335
A.2.4	Characteristic Measures for Multidimensional Data . . . . .	342
A.2.5	Principal Component Analysis . . . . .	344
A.3	Probability Theory . . . . .	350
A.3.1	Probability . . . . .	350
A.3.2	Basic Methods and Theorems . . . . .	353
A.3.3	Random Variables . . . . .	359
A.3.4	Characteristic Measures of Random Variables . . . . .	365
A.3.5	Some Special Distributions . . . . .	369
A.4	Inferential Statistics . . . . .	375
A.4.1	Random Samples . . . . .	376

---

A.4.2	Parameter Estimation . . . . .	376
A.4.3	Hypothesis Testing . . . . .	388
<b>B</b>	<b>KNIME</b> . . . . .	395
B.1	Installation and Overview . . . . .	395
B.2	Building Workflows . . . . .	398
B.3	Example Workflow . . . . .	400
	<b>References</b> . . . . .	409
	<b>Index</b> . . . . .	411

# Symbols

$A, A_i$	attribute, variable [e.g., $A_1 = color, A_2 = price, A_3 = category$ ]
$\omega$	a possible value of an attribute [e.g., $\omega = red$ ]
$\Omega, \text{dom}(\cdot)$	set of possible values of an attribute [e.g., $\Omega_1 = \Omega_{\text{color}} = \text{dom}(A_i) = \{red, blue, green\}$ ]
$\mathcal{A}$	set of all attributes [e.g., $\mathcal{A} = \{color, price, category\}$ ]
$m$	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., $\mathbf{x} = (x_1, x_2, x_3) = (red, 4000, luxury)$ ]
$\mathbf{x}_A$	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., $C = A_3 = category$ ]
$\Omega_C$	set of all possible classes [e.g., $\Omega_C = \{\text{quits, stays, unknown}\}$ ]
$Y$	a selected continuous target attribute [e.g., $Y = A_2 = price$ ]
$\mathcal{C}$	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
$d_{\cdot}$	distance function, metric ( $d_E$ : Euclidean)
$[d_{i,j}]$	distance matrix