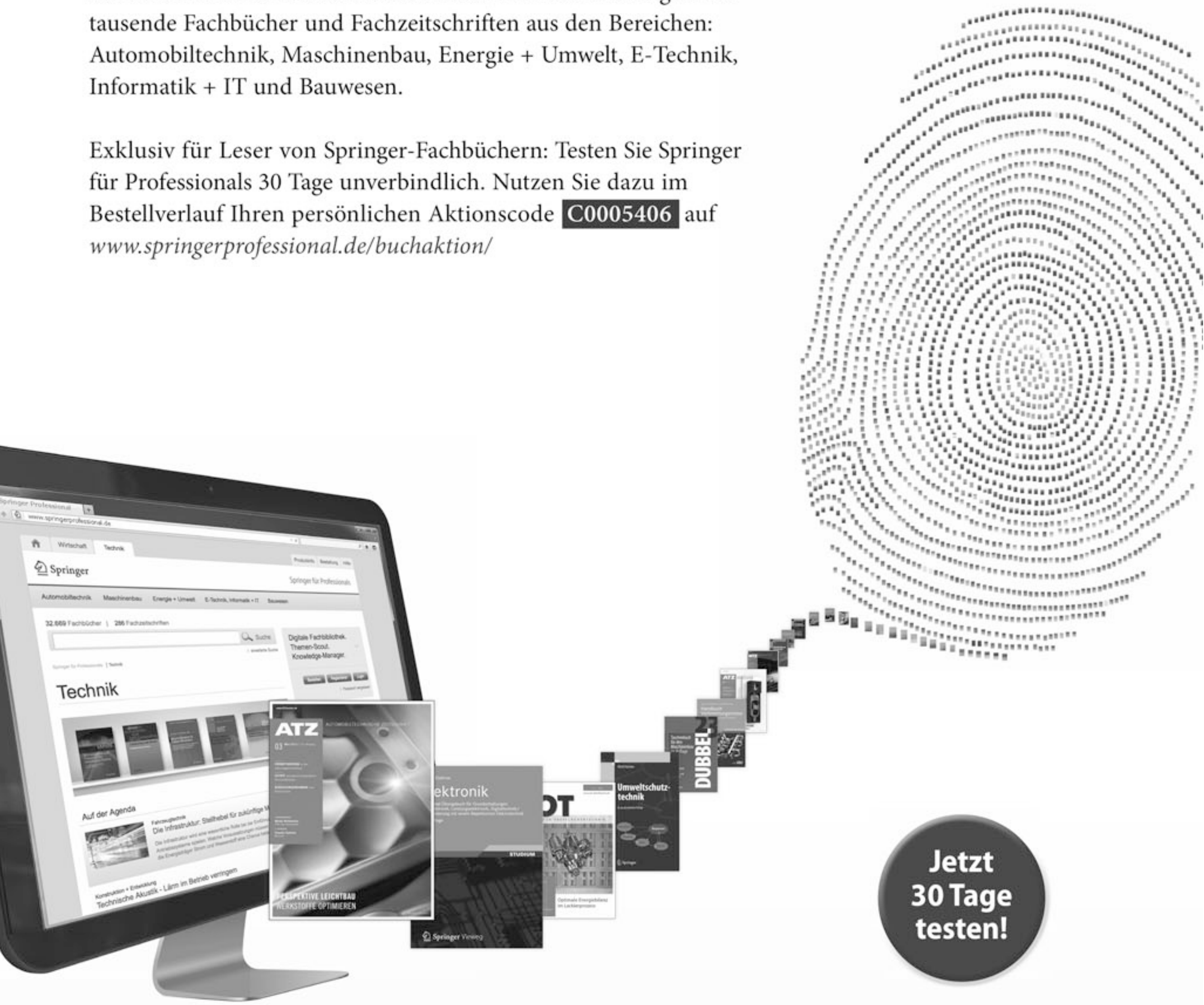
# Data Analytics


Springer
Technik
Auf der Agenda
ATZ
Umweltschutz-technik
DUBBEL
Jetzt 30 Tage testen!

Thomas A. Runkler

# Data Analytics

## Models and Algorithms for Intelligent Data Analysis

2nd Edition

Springer Vieweg

Thomas A. Runkler
Siemens AG
München, Germany

ISBN: 978-3-658-14074-8 ISBN: 978-3-658-14075-5 (eBook)
DOI 10.1007/978-3-658-14075-5

Library of Congress Control Number: 2016942272

Springer Vieweg

Printed on acid-free paper

Springer Vieweg is a brand of Springer Nature
The registered company is Springer Fachmedien Wiesbaden GmbH

# Preface

The information in the world doubles every 20 months. Important data sources are business and industrial processes, text and structured databases, images and videos, and physical and biomedical data. Data analytics allows to find relevant information, structures, and patterns, to gain new insights, to identify causes and effects, to predict future developments, or to suggest optimal decisions. We need models and algorithms to collect, preprocess, analyze, and evaluate data, from various fields such as statistics, machine learning, pattern recognition, system theory, operations research, or artificial intelligence. With this book, you will learn about the most important methods and algorithms for data analytics. You will be able to choose appropriate methods for specific tasks and apply these in your own data analytics projects. You will understand the basic concepts of the growing field of data analytics, which will allow you to keep pace and to actively contribute to the advancement of the field.

This text is designed for undergraduate and graduate courses on data analytics for engineering, computer science, and math students. It is also suitable for practitioners working on data analytics projects. The book is structured according to typical practical data analytics projects. Only basic mathematics is required. This material has been used for more than ten years in numerous courses at the Technical University of Munich, Germany, in short courses at several other universities, and in tutorials at international scientific conferences. Much of the content is based on the results of industrial research and development projects at Siemens.

History of the book versions:

- Data Analytics, second edition, 2016, English
- Data Mining, second edition, 2015, German
- Data Analytics, 2012, English
- Data Mining, 2010, German
- Information Mining, 2000, German

Thanks go to everybody who has contributed to this material, in particular to the reviewers and my students for suggesting improvements and pointing out errors and to the editorial and publisher team for their professional collaboration.

Munich, Germany Thomas A. Runkler
April 2016

# Contents

# List of Symbols

| | |
|---|---|
| $\forall x \in X$ | for each $x$ in $X$ |
| $\exists x \in X$ | there exists an $x$ in $X$ |
| $\Rightarrow$ | if . . . then . . . |
| $\Leftrightarrow$ | if and only if |
| $\int\limits_a^b f\,dx$ | integral of $f$ from $x = a$ to $x = b$ |
| $\frac{\partial f}{\partial x}$ | partial derivative of $f$ with respect to $x$ |
| $\wedge$ | conjunction |
| $\vee$ | disjunction |
| $\cap$ | intersection |
| $\cup$ | union |
| $\neg$ | complement |
| $\setminus$ | set difference |
| $\subset, \subseteq$ | inclusion |
| $\cdot$ | product, inner product |
| $\times$ | Cartesian product, vector product |
| $\{\}$ | empty set |
| $[x, y]$ | closed interval from $x$ to $y$ |
| $(x, y], [x, y)$ | half-bounded intervals from $x$ to $y$ |
| $(x, y)$ | open interval from $x$ to $y$ |
| $\lvert x \rvert$ | absolute value of $x$ |
| $\lvert X \rvert$ | cardinality of the set $X$ |
| $\lVert x \rVert$ | norm of vector $x$ |
| $\lfloor x \rfloor$ | smallest integer $a \geq x$ |
| $\lceil x \rceil$ | largest integer $a \leq x$ |
| $\binom{n}{m}$ | vector with the components $n$ and $m$, binomial coefficient |
| $\infty$ | infinity |
| $a \ll b$ | $a$ is much less than $b$ |
| $a \gg b$ | $a$ is much greater than $b$ |
| $\alpha(t)$ | time-variant learning rate |

| | |
|---|---|
| argmin $X$ | index of the minimum of $X$ |
| argmax $X$ | index of the maximum of $X$ |
| arctan $x$ | arctangent of $x$ |
| artanh $x$ | inverse hyperbolic tangent of $x$ |
| $c_{ij}$ | covariance between features $i$ and $j$ |
| $CE(U)$ | classification entropy of $U$ |
| cov $X$ | covariance matrix of $X$ |
| $d(a,b)$ | distance between $a$ and $b$ |
| eig $X$ | eigenvectors and eigenvalues of $X$ |
| $F_c$ | Fourier cosine transform |
| $F_s$ | Fourier sine transform |
| $h(X)$ | Hopkins index of $X$ |
| $H(a,b)$ | Hamming distance between $a$ and $b$ |
| $H(Z)$ | minimal hypercube or entropy of $Z$ |
| $H(Z \mid a)$ | entropy of $Z$ given $a$ |
| inf $X$ | infimum of $X$ |
| $\lambda$ | eigenvalue, Lagrange variable |
| $L(a,b)$ | edit distance between $a$ and $b$ |
| $\lim_{x\to y}$ | limit as $x$ approaches $y$ |
| $\log_b x$ | logarithm of $x$ to base $b$ |
| $\max X$ | maximum of $X$ |
| $\min X$ | minimum of $X$ |
| $a \bmod b$ | $a$ modulo $b$ |
| $N(\mu,\sigma)$ | Gaussian distribution with mean $\mu$ and standard deviation $\sigma$ |
| `NaN` | undefined (not a number) |
| $PC(U)$ | partition coefficient of $U$ |
| $\mathbb{R}$ | set of real numbers |
| $\mathbb{R}^+$ | set of positive real numbers |
| $r$ | radius |
| $s$ | standard deviation |
| $s_{ij}$ | correlation between features $i$ and $j$ |
| sup $X$ | supremum of $X$ |
| tanh $x$ | hyperbolic tangent of $x$ |
| $u_{ik}$ | membership of the $k$th vector in the $i$th cluster |
| $X$ | set or matrix $X$ |
| $\bar{x}$ | average of $X$ |
| $X^T$, $x^T$ | transpose of the matrix $X$, or the vector $x$ |
| $x_k$ | $k$th vector of $X$ |
| $x^{(i)}$ | $i$th component of $X$ |
| $x_k^{(i)}$ | $i$th component of the $k$th vector of $X$ |
| $x$ | scalar or vector $x$ |
| $x(t)$ | time signal |
| $x(j2\pi f)$ | spectrum |