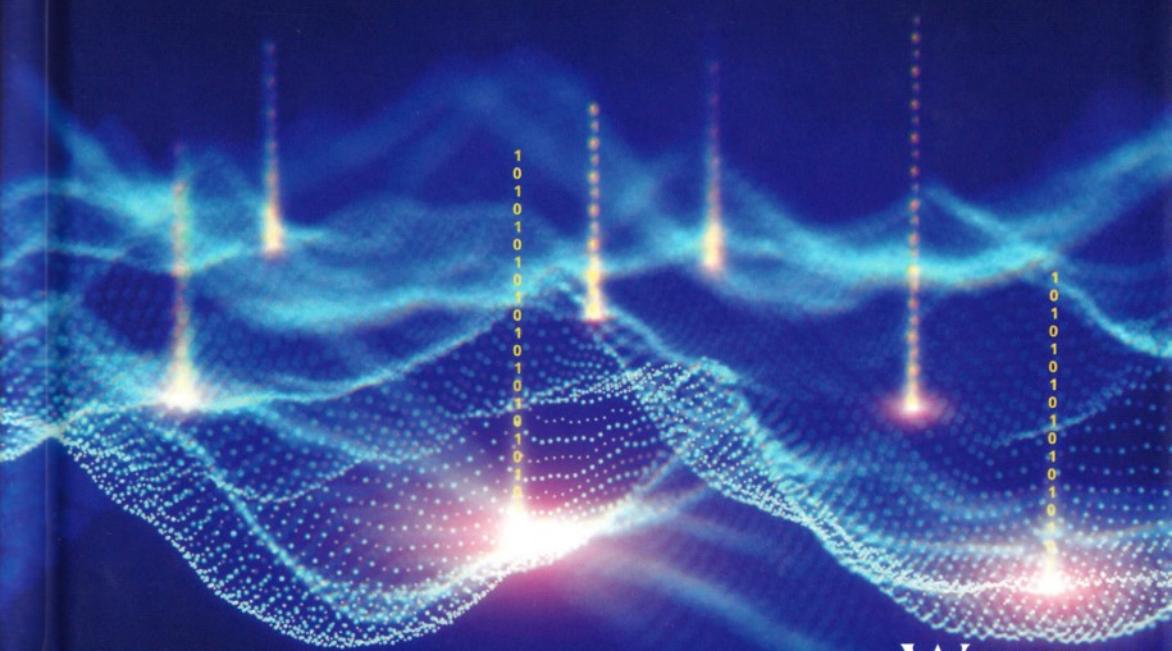


MARIA CRISTINA MARIANI, PhD | OSEI KOFI TWENEBOAH, PhD  
MARIA PIA BECCAR-VARELA, PhD

# DATA SCIENCE IN THEORY AND PRACTICE

TECHNIQUES FOR BIG DATA ANALYTICS  
AND COMPLEX DATA SETS



WILEY

## Contents

**List of Figures** *xvii*

**List of Tables** *xxi*

**Preface** *xxiii*

- 1 Background of Data Science 1**
  - 1.1 Introduction 1
  - 1.2 Origin of Data Science 2
  - 1.3 Who is a Data Scientist? 2
  - 1.4 Big Data 3
    - 1.4.1 Characteristics of Big Data 4
    - 1.4.2 Big Data Architectures 5
  
- 2 Matrix Algebra and Random Vectors 7**
  - 2.1 Introduction 7
  - 2.2 Some Basics of Matrix Algebra 7
    - 2.2.1 Vectors 7
    - 2.2.2 Matrices 8
  - 2.3 Random Variables and Distribution Functions 12
    - 2.3.1 The Dirichlet Distribution 15
    - 2.3.2 Multinomial Distribution 17
    - 2.3.3 Multivariate Normal Distribution 18
  - 2.4 Problems 19
  
- 3 Multivariate Analysis 21**
  - 3.1 Introduction 21
  - 3.2 Multivariate Analysis: Overview 21
  - 3.3 Mean Vectors 22
  - 3.4 Variance–Covariance Matrices 24
  - 3.5 Correlation Matrices 26

3.6	Linear Combinations of Variables	28
3.6.1	Linear Combinations of Sample Means	29
3.6.2	Linear Combinations of Sample Variance and Covariance	29
3.6.3	Linear Combinations of Sample Correlation	30
3.7	Problems	31
<b>4</b>	<b>Time Series Forecasting</b>	<b>35</b>
4.1	Introduction	35
4.2	Terminologies	36
4.3	Components of Time Series	39
4.3.1	Seasonal	39
4.3.2	Trend	40
4.3.3	Cyclical	41
4.3.4	Random	42
4.4	Transformations to Achieve Stationarity	42
4.5	Elimination of Seasonality via Differencing	44
4.6	Additive and Multiplicative Models	44
4.7	Measuring Accuracy of Different Time Series Techniques	45
4.7.1	Mean Absolute Deviation	46
4.7.2	Mean Absolute Percent Error	46
4.7.3	Mean Square Error	47
4.7.4	Root Mean Square Error	48
4.8	Averaging and Exponential Smoothing Forecasting Methods	48
4.8.1	Averaging Methods	49
4.8.1.1	Simple Moving Averages	49
4.8.1.2	Weighted Moving Averages	51
4.8.2	Exponential Smoothing Methods	54
4.8.2.1	Simple Exponential Smoothing	54
4.8.2.2	Adjusted Exponential Smoothing	55
4.9	Problems	57
<b>5</b>	<b>Introduction to R</b>	<b>61</b>
5.1	Introduction	61
5.2	Basic Data Types	62
5.2.1	Numeric Data Type	62
5.2.2	Integer Data Type	62
5.2.3	Character	63
5.2.4	Complex Data Types	63
5.2.5	Logical Data Types	64
5.3	Simple Manipulations – Numbers and Vectors	64
5.3.1	Vectors and Assignment	64

5.3.2	Vector Arithmetic	65
5.3.3	Vector Index	66
5.3.4	Logical Vectors	67
5.3.5	Missing Values	68
5.3.6	Index Vectors	69
5.3.6.1	Indexing with Logicals	69
5.3.6.2	A Vector of Positive Integral Quantities	69
5.3.6.3	A Vector of Negative Integral Quantities	69
5.3.6.4	Named Indexing	69
5.3.7	Other Types of Objects	70
5.3.7.1	Matrices	70
5.3.7.2	List	72
5.3.7.3	Factor	73
5.3.7.4	Data Frames	75
5.3.8	Data Import	76
5.3.8.1	Excel File	76
5.3.8.2	CSV File	76
5.3.8.3	Table File	77
5.3.8.4	Minitab File	77
5.3.8.5	SPSS File	77
5.4	Problems	78
<b>6</b>	<b>Introduction to Python</b>	<b>81</b>
6.1	Introduction	81
6.2	Basic Data Types	82
6.2.1	Number Data Type	82
6.2.1.1	Integer	82
6.2.1.2	Floating-Point Numbers	83
6.2.1.3	Complex Numbers	84
6.2.2	Strings	84
6.2.3	Lists	85
6.2.4	Tuples	86
6.2.5	Dictionaries	86
6.3	Number Type Conversion	87
6.4	Python Conditions	87
6.4.1	If Statements	88
6.4.2	The Else and Elif Clauses	89
6.4.3	The While Loop	90
6.4.3.1	The Break Statement	91
6.4.3.2	The Continue Statement	91
6.4.4	For Loops	91



- 6.4.4.1 Nested Loops 92
- 6.5 Python File Handling: Open, Read, and Close 93
- 6.6 Python Functions 93
- 6.6.1 Calling a Function in Python 94
- 6.6.2 Scope and Lifetime of Variables 94
- 6.7 Problems 95

## **7 Algorithms 97**

- 7.1 Introduction 97
- 7.2 Algorithm – Definition 97
- 7.3 How to Write an Algorithm 98
- 7.3.1 Algorithm Analysis 99
- 7.3.2 Algorithm Complexity 99
- 7.3.3 Space Complexity 100
- 7.3.4 Time Complexity 100
- 7.4 Asymptotic Analysis of an Algorithm 101
- 7.4.1 Asymptotic Notations 102
- 7.4.1.1 Big O Notation 102
- 7.4.1.2 The Omega Notation,  $\Omega$  102
- 7.4.1.3 The  $\Theta$  Notation 102
- 7.5 Examples of Algorithms 104
- 7.6 Flowchart 104
- 7.7 Problems 105

## **8 Data Preprocessing and Data Validations 109**

- 8.1 Introduction 109
- 8.2 Definition – Data Preprocessing 109
- 8.3 Data Cleaning 110
- 8.3.1 Handling Missing Data 110
- 8.3.2 Types of Missing Data 110
- 8.3.2.1 Missing Completely at Random 110
- 8.3.2.2 Missing at Random 110
- 8.3.2.3 Missing Not at Random 111
- 8.3.3 Techniques for Handling the Missing Data 111
- 8.3.3.1 Listwise Deletion 111
- 8.3.3.2 Pairwise Deletion 111
- 8.3.3.3 Mean Substitution 112
- 8.3.3.4 Regression Imputation 112
- 8.3.3.5 Multiple Imputation 112
- 8.3.4 Identifying Outliers and Noisy Data 113
- 8.3.4.1 Binning 113

- 8.3.4.2 Box and Whisker plot 113
- 8.4 Data Transformations 115
- 8.4.1 Min–Max Normalization 115
- 8.4.2 Z-score Normalization 115
- 8.5 Data Reduction 116
- 8.6 Data Validations 117
- 8.6.1 Methods for Data Validation 117
- 8.6.1.1 Simple Statistical Criterion 117
- 8.6.1.2 Fourier Series Modeling and SSC 118
- 8.6.1.3 Principal Component Analysis and SSC 118
- 8.7 Problems 119
  
- 9 Data Visualizations 121**
- 9.1 Introduction 121
- 9.2 Definition – Data Visualization 121
- 9.2.1 Scientific Visualization 123
- 9.2.2 Information Visualization 123
- 9.2.3 Visual Analytics 124
- 9.3 Data Visualization Techniques 126
- 9.3.1 Time Series Data 126
- 9.3.2 Statistical Distributions 127
- 9.3.2.1 Stem-and-Leaf Plots 127
- 9.3.2.2 Q–Q Plots 127
- 9.4 Data Visualization Tools 129
- 9.4.1 Tableau 129
- 9.4.2 Infogram 130
- 9.4.3 Google Charts 132
- 9.5 Problems 133
  
- 10 Binomial and Trinomial Trees 135**
- 10.1 Introduction 135
- 10.2 The Binomial Tree Method 135
- 10.2.1 One Step Binomial Tree 136
- 10.2.2 Using the Tree to Price a European Option 139
- 10.2.3 Using the Tree to Price an American Option 140
- 10.2.4 Using the Tree to Price Any Path Dependent Option 141
- 10.3 Binomial Discrete Model 141
- 10.3.1 One-Step Method 141
- 10.3.2 Multi-step Method 145
- 10.3.2.1 Example: European Call Option 146
- 10.4 Trinomial Tree Method 147

- 10.4.1 What is the Meaning of Little o and Big O? 148
- 10.5 Problems 148
  
- 11 Principal Component Analysis 151**
  - 11.1 Introduction 151
  - 11.2 Background of Principal Component Analysis 151
  - 11.3 Motivation 152
    - 11.3.1 Correlation and Redundancy 152
    - 11.3.2 Visualization 153
  - 11.4 The Mathematics of PCA 153
    - 11.4.1 The Eigenvalues and Eigenvectors 156
  - 11.5 How PCA Works 159
    - 11.5.1 Algorithm 160
  - 11.6 Application 161
  - 11.7 Problems 162
  
- 12 Discriminant and Cluster Analysis 165**
  - 12.1 Introduction 165
  - 12.2 Distance 165
  - 12.3 Discriminant Analysis 166
    - 12.3.1 Kullback–Leibler Divergence 167
    - 12.3.2 Chernoff Distance 167
    - 12.3.3 Application – Seismic Time Series 169
    - 12.3.4 Application – Financial Time Series 171
  - 12.4 Cluster Analysis 173
    - 12.4.1 Partitioning Algorithms 174
      - 12.4.2  $k$ -Means Algorithm 174
      - 12.4.3  $k$ -Medoids Algorithm 175
    - 12.4.4 Application – Seismic Time Series 176
    - 12.4.5 Application – Financial Time Series 176
  - 12.5 Problems 177
  
- 13 Multidimensional Scaling 179**
  - 13.1 Introduction 179
  - 13.2 Motivation 180
  - 13.3 Number of Dimensions and Goodness of Fit 182
  - 13.4 Proximity Measures 183
  - 13.5 Metric Multidimensional Scaling 183
    - 13.5.1 The Classical Solution 184
  - 13.6 Nonmetric Multidimensional Scaling 186
    - 13.6.1 Shepard–Kruskal Algorithm 186
  - 13.7 Problems 187

<b>14</b>	<b>Classification and Tree-Based Methods</b>	<b>191</b>
14.1	Introduction	191
14.2	An Overview of Classification	191
14.2.1	The Classification Problem	192
14.2.2	Logistic Regression Model	192
14.2.2.1	$l_1$ Regularization	193
14.2.2.2	$l_2$ Regularization	194
14.3	Linear Discriminant Analysis	194
14.3.1	Optimal Classification and Estimation of Gaussian Distribution	195
14.4	Tree-Based Methods	197
14.4.1	One Single Decision Tree	197
14.4.2	Random Forest	198
14.5	Applications	200
14.6	Problems	202
<b>15</b>	<b>Association Rules</b>	<b>205</b>
15.1	Introduction	205
15.2	Market Basket Analysis	205
15.3	Terminologies	207
15.3.1	Itemset and Support Count	207
15.3.2	Frequent Itemset	207
15.3.3	Closed Frequent Itemset	207
15.3.4	Maximal Frequent Itemset	208
15.3.5	Association Rule	208
15.3.6	Rule Evaluation Metrics	208
15.4	The Apriori Algorithm	210
15.4.1	An example of the Apriori Algorithm	211
15.5	Applications	213
15.5.1	Confidence	214
15.5.2	Lift	215
15.5.3	Conviction	215
15.6	Problems	216
<b>16</b>	<b>Support Vector Machines</b>	<b>219</b>
16.1	Introduction	219
16.2	The Maximal Margin Classifier	219
16.3	Classification Using a Separating Hyperplane	223
16.4	Kernel Functions	225
16.5	Applications	225
16.6	Problems	227



<b>17</b>	<b>Neural Networks</b>	<b>231</b>
17.1	Introduction	231
17.2	Perceptrons	231
17.3	Feed Forward Neural Network	231
17.4	Recurrent Neural Networks	233
17.5	Long Short-Term Memory	234
17.5.1	Residual Connections	235
17.5.2	Loss Functions	236
17.5.3	Stochastic Gradient Descent	236
17.5.4	Regularization – Ensemble Learning	237
17.6	Application	237
17.6.1	Emergent and Developed Market	237
17.6.2	The Lehman Brothers Collapse	237
17.6.3	Methodology	238
17.6.4	Analyses of Data	238
17.6.4.1	Results of the Emergent Market Index	238
17.6.4.2	Results of the Developed Market Index	238
17.7	Significance of Study	239
17.8	Problems	240
<b>18</b>	<b>Fourier Analysis</b>	<b>245</b>
18.1	Introduction	245
18.2	Definition	245
18.3	Discrete Fourier Transform	246
18.4	The Fast Fourier Transform (FFT) Method	247
18.5	Dynamic Fourier Analysis	250
18.5.1	Tapering	251
18.5.2	Daniell Kernel Estimation	252
18.6	Applications of the Fourier Transform	253
18.6.1	Modeling Power Spectrum of Financial Returns Using Fourier Transforms	253
18.6.2	Image Compression	259
18.7	Problems	259
<b>19</b>	<b>Wavelets Analysis</b>	<b>261</b>
19.1	Introduction	261
19.1.1	Wavelets Transform	262
19.2	Discrete Wavelets Transforms	264
19.2.1	Haar Wavelets	265
19.2.1.1	Haar Functions	265
19.2.1.2	Haar Transform Matrix	266

19.2.2	Daubechies Wavelets	267
19.3	Applications of the Wavelets Transform	269
19.3.1	Discriminating Between Mining Explosions and Cluster of Earthquakes	269
19.3.1.1	Background of Data	269
19.3.1.2	Results	269
19.3.2	Finance	271
19.3.3	Damage Detection in Frame Structures	275
19.3.4	Image Compression	275
19.3.5	Seismic Signals	275
19.4	Problems	276
<b>20</b>	<b>Stochastic Analysis</b>	<b>279</b>
20.1	Introduction	279
20.2	Necessary Definitions from Probability Theory	279
20.3	Stochastic Processes	280
20.3.1	The Index Set $I$	281
20.3.2	The State Space $S$	281
20.3.3	Stationary and Independent Components	281
20.3.4	Stationary and Independent Increments	282
20.3.5	Filtration and Standard Filtration	283
20.4	Examples of Stochastic Processes	284
20.4.1	Markov Chains	285
20.4.1.1	Examples of Markov Processes	286
20.4.1.2	The Chapman–Kolmogorov Equation	287
20.4.1.3	Classification of States	289
20.4.1.4	Limiting Probabilities	290
20.4.1.5	Branching Processes	291
20.4.1.6	Time Homogeneous Chains	293
20.4.2	Martingales	294
20.4.3	Simple Random Walk	294
20.4.4	The Brownian Motion (Wiener Process)	294
20.5	Measurable Functions and Expectations	295
20.5.1	Radon–Nikodym Theorem and Conditional Expectation	296
20.6	Problems	299
<b>21</b>	<b>Fractal Analysis – Lévy, Hurst, DFA, DEA</b>	<b>301</b>
21.1	Introduction and Definitions	301
21.2	Lévy Processes	301
21.2.1	Examples of Lévy Processes	304
21.2.1.1	The Poisson Process (Jumps)	305
21.2.1.2	The Compound Poisson Process	305

21.2.1.3	Inverse Gaussian (IG) Process	306
21.2.1.4	The Gamma Process	307
21.2.2	Exponential Lévy Models	307
21.2.3	Subordination of Lévy Processes	308
21.2.4	Stable Distributions	309
21.3	Lévy Flight Models	311
21.4	Rescaled Range Analysis (Hurst Analysis)	312
21.5	Detrended Fluctuation Analysis (DFA)	315
21.6	Diffusion Entropy Analysis (DEA)	316
21.6.1	Estimation Procedure	317
21.6.1.1	The Shannon Entropy	317
21.6.2	The $H$ - $\alpha$ Relationship for the Truncated Lévy Flight	319
21.7	Application – Characterization of Volcanic Time Series	321
21.7.1	Background of Volcanic Data	321
21.7.2	Results	321
21.8	Problems	323
<b>22</b>	<b>Stochastic Differential Equations</b>	<b>325</b>
22.1	Introduction	325
22.2	Stochastic Differential Equations	325
22.2.1	Solution Methods of SDEs	326
22.3	Examples	335
22.3.1	Modeling Asset Prices	335
22.3.2	Modeling Magnitude of Earthquake Series	336
22.4	Multidimensional Stochastic Differential Equations	337
22.4.1	The multidimensional Ornstein–Uhlenbeck Processes	337
22.4.2	Solution of the Ornstein–Uhlenbeck Process	338
22.5	Simulation of Stochastic Differential Equations	340
22.5.1	Euler–Maruyama Scheme for Approximating Stochastic Differential Equations	340
22.5.2	Euler–Milstein Scheme for Approximating Stochastic Differential Equations	341
22.6	Problems	343
<b>23</b>	<b>Ethics: With Great Power Comes Great Responsibility</b>	<b>345</b>
23.1	Introduction	345
23.2	Data Science Ethical Principles	346
23.2.1	Enhance Value in Society	346
23.2.2	Avoiding Harm	346
23.2.3	Professional Competence	347
23.2.4	Increasing Trustworthiness	348

23.2.5	Maintaining Accountability and Oversight	348
23.3	Data Science Code of Professional Conduct	348
23.4	Application	350
23.4.1	Project Planning	350
23.4.2	Data Preprocessing	350
23.4.3	Data Management	350
23.4.4	Analysis and Development	351
23.5	Problems	351

**Bibliography** 353

**Index** 359



# EXPLORE THE FOUNDATIONS OF DATA SCIENCE WITH THIS INSIGHTFUL NEW RESOURCE

*Data Science in Theory and Practice* delivers a comprehensive treatment of the mathematical and statistical models useful for analyzing data sets arising in various disciplines, like banking, finance, health care, bioinformatics, security, education, and social services. Written in five parts, the book examines some of the most commonly used and fundamental mathematical and statistical concepts that form the basis of data science. The authors go on to analyze various data transformation techniques useful for extracting information from raw data, long memory behavior, and predictive modeling.

The book offers readers a multitude of topics all relevant to the analysis of complex data sets. Along with a robust exploration of the theory underpinning data science, it contains numerous applications to specific and practical problems. The book also provides examples of code algorithms in R and Python and provides pseudo-algorithms to port the code to any other language.

Ideal for students and practitioners without a strong background in data science, readers will also learn from topics like:

- Analyses of foundational theoretical subjects, including the history of data science, matrix algebra and random vectors, and multivariate analysis
- A comprehensive examination of time series forecasting, including the different components of time series and transformations to achieve stationarity
- Introductions to both the R and Python programming languages, including basic data types and sample manipulations for both languages
- An exploration of algorithms, including how to write one and how to perform an asymptotic analysis
- A comprehensive discussion of several techniques for analyzing and predicting complex data sets

Perfect for advanced undergraduate and graduate students in Data Science, Business Analytics, and Statistics programs, *Data Science in Theory and Practice* will also earn a place in the libraries of practicing data scientists, data and business analysts, and statisticians in the private sector, government, and academia.

**MARIA CRISTINA MARIANI, PhD**, is Shigeko K. Chan Distinguished Professor and Chair in the Department of Mathematical Sciences at The University of Texas at El Paso. She currently focuses her research on Stochastic Analysis, Differential Equations and Machine Learning with applications to Big Data and Complex Data sets arising in Public Health, Geophysics, Finance and others. Dr. Mariani is co-author of other Wiley books including *Quantitative Finance*.

**OSEI KOFI TWENEBOAH, PhD**, is Assistant Professor of Data Science at Ramapo College of New Jersey. His main research is Stochastic Analysis, Machine Learning and Scientific Computing with applications to Finance, Health Sciences, and Geophysics.


**MARIA PIA BECCAR-VARELA, PhD**, is Associate Professor of Instruction in the Department of Mathematical Sciences at the University of Texas at El Paso. Her research interests include Differential Equations, Stochastic Differential Equations, Wavelet Analysis and Discriminant Analysis applied to Finance, Health Sciences, and Earthquake Studies.

Cover Design: Wiley

Cover Image: © nobeastsofierce/Shutterstock

[www.wiley.com](http://www.wiley.com)

**WILEY**

 Also available  
as an e-book

