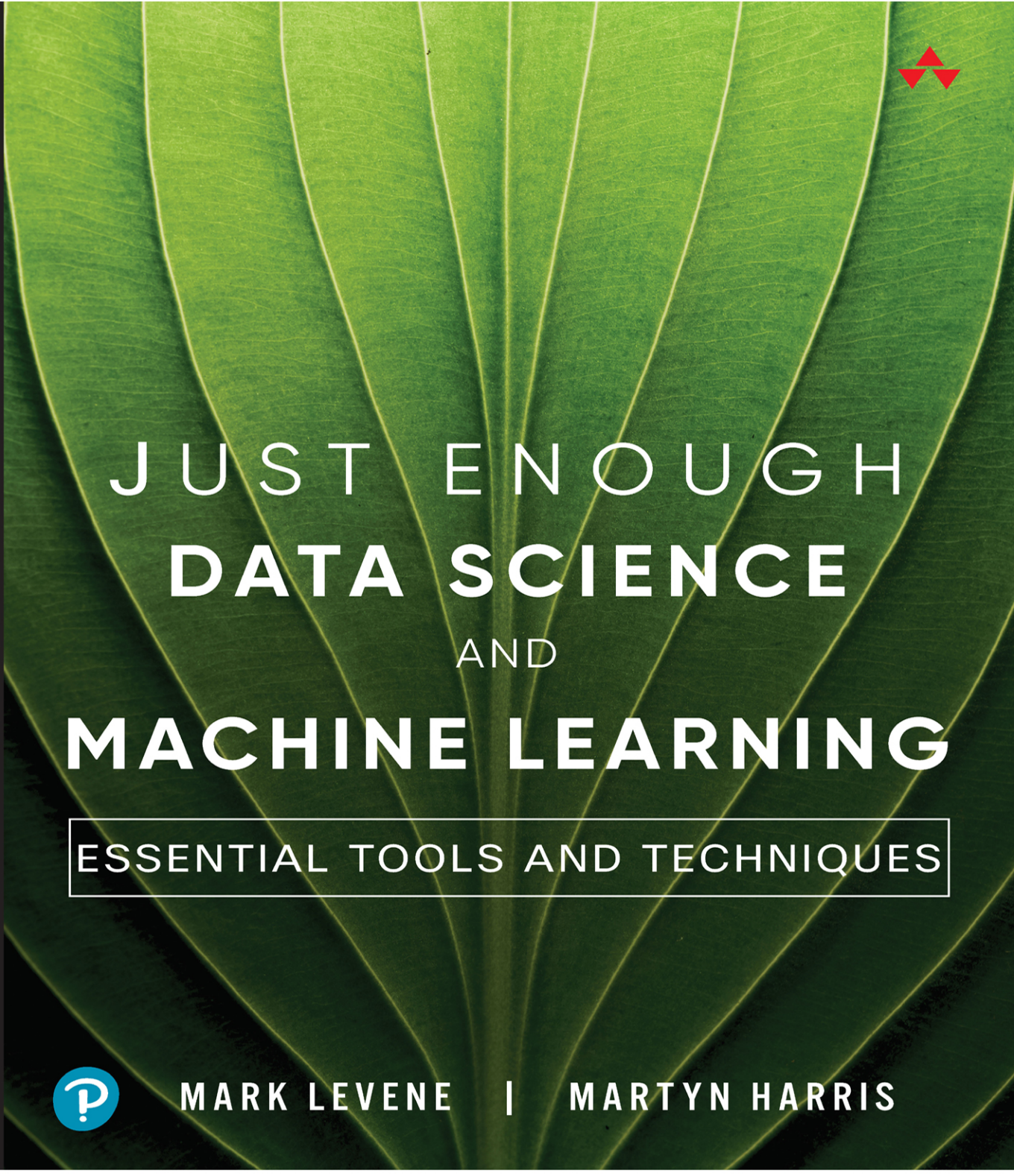

FREE SAMPLE CHAPTER |

Just Enough Data Science and Machine Learning

This page intentionally left blank

Just Enough Data Science and Machine Learning

Essential Tools and Techniques

Mark Levene
Martyn Harris

Addison-Wesley

Figure 3.7: Papers with Code
Cover Image: tabaca/Shutterstock

Visit us on the Web: informit.com/aw

Library of Congress Control Number: 2024947279

Hoboken, NJ

ISBN-13: 978-0-13-834074-2
ISBN-10: 0-13-834074-9

$PrintCode

Contents

List of Figures ix

Preface xvii

About the Authors xix

1 What Is Data Science? 1

2 Basic Statistics 3
- 2.1 Introductory Statistical Notions 3
- 2.2 Expectation 17
- 2.3 Variance 21
- 2.4 Correlation 26
- 2.5 Regression 28
- 2.6 Chapter Summary 32

3 Types of Data 33
- 3.1 Tabular Data 33
- 3.2 Textual Data 38
- 3.3 Image, Video, and Audio Data 40
- 3.4 Time Series Data 41
- 3.5 Geographical Data 42
- 3.6 Social Network Data 44
- 3.7 Transforming Data 46
- 3.8 Chapter Summary 51

4 Machine Learning Tools 52
- 4.1 What Is Machine Learning? 52
- 4.2 Evaluation 57
  - 4.2.1 Evaluation for Supervised Models 58
  - 4.2.2 Evaluation for Unsupervised Models 62
- 4.3 Supervised Methods 68
  - 4.3.1 K-Nearest Neighbours 68
  - 4.3.2 Naive Bayes 71

4.3.3 Support Vector Machines 77
4.3.4 Decision Trees and Random Forests 85
4.3.5 Neural Networks and Deep Learning 93
4.4 Unsupervised Methods 105
4.4.1 K-Means 106
4.4.2 Hierarchical Clustering 110
4.4.3 Principal Components Analysis 113
4.4.4 Topic Modelling 117
4.4.5 DBSCAN 121
4.5 Semi-Supervised Methods 125
4.6 Chapter Summary 129

5 Data Science Topics 130
5.1 Searching, Ranking, and Rating 130
5.1.1 The Vector Space Model 134
5.1.2 Ranking with PageRank 138
5.1.3 Rating with the Elo System 143
5.1.4 Recommender Systems and Collaborative Filtering 145
5.2 Social Networks 150
5.2.1 The Basics of Social Networks 151
5.2.2 Centrality Measures 154
5.2.3 Power Laws and the 80–20 Rule 157
5.2.4 SIS and SIR Models for the Spread of Disease 164
5.3 Three Natural Language Processing Topics 171
5.3.1 Sentiment Analysis 171
5.3.2 Named Entity Recognition 174
5.3.3 Word Embeddings 177
5.4 Chapter Summary 183

6 Selected Additional Topics 184
6.1 Neuro-Symbolic AI 184

6.2 Conversational AI 185
6.3 Generative Neural Networks 185
6.4 Trustworthy AI 186
6.5 Large Language Models 187
6.6 Epilogue 187

7 Further Reading 189
7.1 Basic Statistics 189
7.2 Data Science 189
7.3 Machine Learning 190
7.4 Deep Learning 191
7.5 Research Papers 191
7.6 Python 191

Bibliography 192

Index 195

This page intentionally left blank

List of Figures

2.1 Histogram (top left) and corresponding PDF (top right), ECDF (bottom left) and CDF (bottom right). 6

2.2 The histogram for the flipper length of penguins from the Palmer Archipelago (Antarctica) data set and the corresponding kde plot (overlaid in red). 7

2.3 Histogram (left) and PMF (right) for the proportion of candidate votes; a Bernoulli distribution with $p = 0.52$ is fitted to the data, that is, $X \sim \text{Bern}(0.52)$. 9

2.4 Histogram and PMF (overlaid in red) for the proportion of heads; a binomial distribution with $p = 0.5$ and $n = 100$ is fitted to the data, that is, $X \sim \text{Bin}(100, 0.5)$. 10

2.5 Histogram and PMF (overlaid in red) for the number of births within a period of an hour; a Poisson distribution with $\lambda = 2.01$ is fitted to the data, that is, $X \sim \text{Pois}(2.01)$. 11

2.6 Histogram and PDF (overlaid in red) of the distribution of the heights of women; a normal distribution with parameters $\mu = 63.709$ and $\sigma = 2.696$ is fitted to the data, $X \sim \text{N}(63.709, 2.696)$. 12

2.7 Histograms and PDFs (overlaid in red) of heart surgery survival rates (left), and HackerNews posts (right); an exponential distribution with $\lambda = 223.28$ is fitted to the heart data, that is, $X \sim \text{Exp}(222.28)$, while an exponential distribution with $\lambda = 0.99$ is fitted to the HackerNews data, that is, $X \sim \text{Exp}(0.99)$. 14

2.8 Histogram plot of FIFA 2019 player salaries (left), and a further plot of the top 1,500 words and their frequency counts in the Brown corpus (right); a Pareto distribution with parameters $x_m = 2.46$ and $\alpha = 0.13$ is fitted to the football salary data, that is, $X \sim \text{Pareto}(2.46, 0.13)$, while a Pareto distribution with $x_m = 72.192$ and $\alpha = 1.0435$ is fitted to the word count data, that is, $X \sim \text{Pareto}(72.192, 1.0435)$, where X is a random variable indicating the count of a word in the corpus. 16

2.9 Boxplot of the salary and qualification data set. 19
2.10 PDF for the bootstrap computation of 95% confidence intervals for the mean (left), median (middle), and mode (right) of the heights of women in the telephone survey data set. 20
2.11 Two line plots with the mean number of passengers (left) and median number of passengers (right) on international flights leaving the US. The band represents the confidence intervals of the 2.5th and 97.5th percentiles computed from the monthly passenger count for each year. 22
2.12 A histogram and PDF (overlaid in red) for the sale price of properties, demonstrating that the data set is positively skewed. 24
2.13 A histogram and PDF (overlaid in red) for age at death of notable individuals in 2016, demonstrating that the data set is negatively skewed. 25
2.14 The histogram and PDF of the maximum likelihood fit of the normal distribution (overlaid in red) for the first-order differences of the value of Bitcoin, demonstrating that the excess kurtosis present in the data set is high. 26
2.15 A scatter plot showing the correlation between two random variables X and Y on a synthetic data set with regression lines denoting the linear fit between X and Y. The reported Pearson correlation coefficient ρ appears in each subplot. The subplots in the top row illustrate positive correlation, while those in the bottom row illustrate negative correlation. 27
2.16 A scatter plot (left) of the price plotted against the average number of rooms in a property overlaid with a red line denoting the least squared linear fit and the corresponding residual errors (right). 30

2.17 A logistic regression plot showing the classification of the data points into those with diabetes (top blue data points), and those without (bottom blue data points). A regression curve has been fit to the data with a 95% confidence interval computed from 10,000 bootstrap resamples. 31

3.1 A stacked bar chart displaying the monthly temperature and humidity recorded in London in 2015 by the *UK Met Office*. 34

3.2 A heatmap of the pairwise Pearson correlation scores generated from the matrix representation for the breakfast cereal data set. (The legend on the left of the heatmap maps colour intensity to values.) 36

3.3 Histogram depicting the distribution of recorded magnitudes for earthquakes worldwide in July 2021. 37

3.4 A scatter plot of the salary for 30 individuals against their years of experience, together with a linear regression line fitted to the data. The slope of the fitted line is 9,345, its intercept is 26,816, and the coefficient of determination of $R^2 = 0.97$. 38

3.5 A word cloud summarising the most frequent words captured from Reddit posts published in 2021 on the topic of the coronavirus pandemic. 39

3.6 A rank-order distribution (left) and Zipf plot (right) of the 500 most frequent words captured in the Reuters corpus. 40

3.7 A sample of a collection of over three million images of cats and dogs for use as CAPTCHAs. 41

3.8 The count of monthly sunspots visualised as the raw time series (top), a moving average of the time series with a window size of 12 months representing a year of observations (middle), and a correlogram of the time series with a 95% confidence interval about 0, which represents no correlation (bottom). 43

3.9 A variogram for the copper deposits with the range=864.08, the sill=574.90, and the nugget=163.25. 45

3.10 A visualisation of the Karate Club social network. 46

3.11 A histogram of the number of economically active adults in the UK recorded in a Labour survey conducted by the Department for BIS 2010. 50
3.12 A bar chart of the top ten artists ranked by the number of streams recorded on a popular music digital service on 1 January 2017. 51

4.1 The confusion matrix. 59
4.2 The confusion matrix describing the model performance of a logistic regression model trained on the Pima Indian diabetes data set. 61
4.3 An elbow plot for the Iris data set, with the number of clusters along the x-axis and the SSE on the y-axis. 63
4.4 The silhouette plot for $k = 6$ clusters generated by *k*-means computed over the Iris data set. The vertical line represents the silhouette coefficient, which is 0.571 in this case (see Table 4.4). 66
4.5 The confusion matrix describing the model performance of a KNN model computed on the exoplanet data set. 70
4.6 The confusion matrix for the Bernoulli naive Bayes model trained on the SMS data set. 74
4.7 The confusion matrix for the multinomial naive Bayes model trained on the SMS data set. 75
4.8 The confusion matrix for the Gaussian naive Bayes model trained on the mushroom data set. 76
4.9 Example of an SVM hard margin. 79
4.10 Example of SVM hard and soft margins. 79
4.11 A scatter plot of the Iris data set for the two classes representing the species Iris setosa and Iris versicolour. The decision boundary was computed using hard margin SVM classification and shows the hard margin between the two classes. 80

4.12 A scatter plot of the Iris data set for the two classes representing the species Iris versicolour and Iris virginica. The decision boundary was computed using soft margin SVM classification and shows the soft margin between the two classes. 81
4.13 The confusion matrix for a hard margin SVM binary classifier trained on the Iris data with classes Iris versicolour and Iris setosa. 81
4.14 The confusion matrix for a soft margin SVM binary classifier, with $C = 100$, trained on the Iris data with classes Iris versicolour and Iris virginica. 82
4.15 Example of SVM nonlinear kernel, showing a dense circle of data points surrounded by a ring of data points (left plot) and its nonlinear kernel transformation (right plot). 83
4.16 Example of a simple decision tree based on heart attack data. 86
4.17 Node types in a tree data structure. 86
4.18 A path taken from the root node (labelled “Age”) through the internal node (labelled “Hypertension”) leading to the leaf node (labelled “No heart attack”). 87
4.19 Crucial step of the decision tree construction algorithm. 88
4.20 The confusion matrix for the decision tree classifier trained on the heart disease data set. 91
4.21 The confusion matrix for the random forest classifier trained on the heart disease data set. 92
4.22 Example of a feedforward NN with two hidden layers. 94
4.23 Example of a perceptron, that is, a feedforward NN with no hidden layers. 95
4.24 The confusion matrix for the perceptron trained on the *Titanic* data set. 99
4.25 The confusion matrix for a feedforward neural network with one hidden layer trained on the *Titanic* data set. 100
4.26 The confusion matrix for a feedforward neural network with two hidden layers trained on the *Titanic* data set. 101
4.27 Example of a convolutional neural network. 102

4.28 Example of a recurrent neural network. 103
4.29 Example of a long short-term memory NN. 104
4.30 Elbow plot for *k*-means applied to the heart disease data set, with the number of clusters along the x-axis and the SSE on the y-axis. 108
4.31 A dendrogram of HAC applied to the Iris data set using the single-link method. The x-axis is labelled with the data point ID (if it is an initial cluster) or the number of data points in the cluster (in brackets), and the y-axis represents the distance between two clusters being merged. 112
4.32 A dendrogram of HAC applied to the Iris data set using the average-link method. The x-axis is labelled with the data point ID (if it is an initial cluster) or the number of data points in the cluster (in brackets), and the y-axis represents the distance between two clusters being merged. 113
4.33 A dendrogram of HAC applied to the Iris data set using Ward's linkage method. The x-axis is labelled with the data point ID (if it is an initial cluster) or the number of data points in the cluster (in brackets), and the y-axis represents the distance between two clusters being merged. 114
4.34 Example of PCA in two dimensions. The x-axis is labelled by pc1, the first principal component, while the y-axis is labelled by pc2, the second principal component. 116
4.35 A scatter plot of the data points in the breast cancer data set associated with the two principal components. As in Figure 4.34 the x-axis is labelled by pc1, the first principal component, while the y-axis is labelled by pc2, the second principal component. 117
4.36 Graphical model plate notation for probabilistic latent semantic analysis. 119
4.37 A plot of the four topics from the ABC news group data set with the top-ten words per topic in descending order ordered by weight. 121

4.38 A plot of the average topic coherence scores for the NMF model, with the number of topics ranging from $k = 2$ to $k = 30$. 122
4.39 DBSCAN concepts. 123
4.40 A plot of the clusters obtained from DBSCAN applied to the Whooper Swan data set, representing the moulting sites of the swans. 125
4.41 The confusion matrix for a semi-supervised model trained on the breast cancer data set. 128

5.1 The architecture of a prototype search engine. 132
5.2 Basic crawler algorithm. 133
5.3 A small web graph with five web pages, used to illustrate PageRank; the actual PageRank values are shown beside the web pages. 140
5.4 The citation curve of the top-200 cited publications contained in Albert Einstein's citation vector, whose *h*-index is 165 (left plot) and, correspondingly, the citation curve of the top-200 cited publications contained in the citation vector of an anonymous theoretical physicist whose *h*-index is 63 (right plot). The index of publications is shown along the x-axis and the number of citations per publication is shown on the y-axis. 143
5.5 A visualisation of the Karate Club social network, with blue nodes representing followers of the administrator (node 1) and orange nodes representing followers of the instructor (node 34). 151
5.6 A two-dimensional grid with random shortcuts. 154
5.7 The inlinks degree distribution plot (left) and the log-log plot (right) of the 1999 crawl data set. 161
5.8 A Pareto chart for the 1999 crawl data set. 163
5.9 Transfer diagram for the SIS compartmental model. 165
5.10 Transfer diagram for the SIR compartmental model. 166
5.11 A plot showing the dynamics of the SIS model, where the *x*-axis is time in days and the *y*-axis is the number of individuals in a compartment; blue=susceptible and red=infected. 169

5.12 A plot showing the dynamics of the SIR model, where the *x*-axis is time in days and the *y*-axis is the number of individuals in a compartment; blue=susceptible, red=infected, and green=recovered. 170
5.13 The skip-gram word2vec neural network architecture. 179
5.14 The CBOW word2vec neural network architecture. 180
5.15 A visualisation of a small sample of word embeddings in two-dimensional vector space, which strives to maintain the cosine similarity between the embeddings. 182

Preface

The topic. Data science is an interdisciplinary field that has evolved from a synergy between computer science and statistics. While data science focuses on the analysis and interpretation of data, machine learning is its algorithmic part enabling the discovery of patterns from a statistical model formed from the data.

In recent years, the field of data science and its subfield machine learning have emerged as indispensable tools for extracting valuable insights and making predictions from potentially vast amounts of data. As these disciplines continue to shape our world, it becomes increasingly important, not only to a wide range of information technology (IT) professionals and researchers, but also to enthusiasts who wish to grasp their fundamental principles and techniques.

Motivation. Our aim in this book is that it will serve as an introductory guide to data science and machine learning. In our journey to explore the fundamentals of these fields, we focus on their applied side, giving practical examples of the concepts and methods. However, we do not shy away from presenting the fundamental theory of these subjects.

Thus our motivation for writing this book stems from a desire to provide a middle ground within the vast spectrum of literature on these topics. Our aim was to strike a balance between theoretical rigour and practical utility, offering readers an applied perspective enriched with algorithmic insights and essential statistical foundations.

Coded examples. Recognising the evolving nature of programming languages and the potential deprecation of code examples, within the book itself we have chosen to present the core concepts without listing the actual code, in order to ensure the longevity and relevance of the content for years to come. Nonetheless, all of the illustrative examples in the book are drawn from real-world data sets. The code that was used to produce data summaries and plots presented in each of the examples is hosted on a dedicated web site enabling readers to reproduce them, and ensuring a hands-on learning experience while future-proofing the material against evolving programming languages.

The audience. This book is tailored for anyone seeking a comprehensive, yet accessible, introduction to data science and machine learning. Whether you are a student venturing into these fields for the first time an IT professional or researcher looking to broaden your skill set, or an enthusiast eager to explore the potential of data-driven decision making, this book is designed to cater to your needs.

Prerequisites. We present an overview of data science with minimal mathematical prerequisites, ensuring that readers can grasp the fundamentals without the need for grasping complex equations. A basic understanding of mathematics and statistics will suffice to embark on this journey. We hope you will enjoy the book as much as we have enjoyed writing it.

Acknowledgments

We extend special thanks to our editor at Pearson, Kim Spenceley, who patiently guided us through the publication process. We would also like to thank the reviewers of the draft version of the book for their constructive comments.

Dedications

Mark Levene dedicates this book to his wife Sara and their three children Tamara, Joseph, and Oren. Martyn Harris dedicates this book to his wife Ilaria, their three children Zeno, Carolina, and Susanna, and to the memory of Angela and Nigel Harris.

Mark Levene and Martyn Harris

London, 2024

Register your copy of *Just Enough Data Science and Machine Learning* on the InformIT site for convenient access to updates and/or corrections as they become available. To start the registration process, go to informit.com/register and log in or create an account. Enter the product ISBN (**9780138340742**) and click Submit. A link to the code resources accompanying the book is also available on InformIT.

About the Authors

Mark Levene is emeritus professor of Computer Science at Birkbeck University of London. His main area of expertise encompasses Data Science and Machine Learning, including Applied Machine Learning, Trustworthy and Safe AI.

Martyn Harris is a lecturer and Programme Director at Birkbeck University of London. His areas of expertise include Data Science, Applied Machine Learning, and Natural Language Processing.

This page intentionally left blank

1

What Is Data Science?

I do not fear computers. I fear the lack of them.

Isaac Asimov, Science fiction writer

Any sufficiently advanced technology is indistinguishable from magic.

Arthur C. Clarke, Science fiction writer

Data science is inherently an interdisciplinary activity that has evolved from a synergy between computer science and statistics.

To do data science, we need data! So we have a data set; it could be a structured database of employee records with all their details, an unstructured collection of textual documents (say emails), a large collection of images of animals, a time series of financial data from the stock market, epidemiological data giving the number of infected individuals per day for a given region over a period of time, or geographical data pertaining to businesses in central London.

Having data is not enough; we need to have a problem we would like to solve or some questions we wish to answer using the data. For example, in an employee data set, we may wish to know the employee characteristics that determine their salary band or, in an epidemiological data set, we may wish to determine how fast a virus is spreading in the population. Now, in a broad sense, once the data is available and we have a well-defined problem to work on, several steps determine the tasks a data scientist should perform to tackle the problem at hand and find out what the data is telling us.

It is always sensible to start with an exploratory data analysis phase, which is carried out with the aid of visualisation tools. Exploring data will help us form some hypotheses about the data, which in turn allows us to build a statistical model of the data. Inevitably, we will use an algorithmic method, based on our model, whose output will assist us in verifying or refuting the hypotheses we have formed.

In a nutshell, this is what data science is about. The algorithmic method essentially enables the discovery of patterns in the data, which may be large in size and/or complex, according to the statistical model we have formed. This is often referred to as *machine learning*; however, the general process of pattern or knowledge discovery is known as *data mining*. In our exposition of data science, we prefer to use the term *machine learning* as the subfield of computer science responsible for the algorithmic part of data science.

Therefore, in a very broad sense, data science comprises the methods and algorithms used to analyse the data and present the findings from the ensuing analysis.

Taking this a step further, who are the stakeholders in this discipline and activity called data science? Computer scientists, such as the authors, are responsible for designing and implementing the algorithms in such a way that they scale to very large and potentially complex data sets. Then statisticians are responsible for model building, which is an essential part of data science. However, one could argue that the data scientist combines skills from these two disciplines of computer science and statistics, leaning toward one side or another depending on their background. Still, we have a third group of stakeholders who bring the data and the problems to the table: they may be social scientists, economists, epidemiologists, or any other professionals from any other discipline that would like to use data science to aid them in answering questions they have about the data they possess.

For successful data science to take place, more often than not, an interdisciplinary team needs to be working on the problem at hand. There is also a breed of data scientists who, from the start, build their expertise in this field rather than in the field of computer science or statistics. Moreover, others, such as the authors, started off in computer science or statistics and have moved their expertise to the middle ground of data science.

Ultimately, the question of what exactly is the relationship between data science, statistics, and computer science/machine learning will remain an ongoing debate. It is important from our perspective to appreciate that data science demands the application of expertise from both these disciplines to solve real-world problems emanating from data. Furthermore, our goal in this book is to provide a relatively brief technical introduction to this exciting field that can be understood by practitioners and researchers alike, coming from diverse backgrounds.

In Chapter 2 we introduce the basic statistical notions needed to become a data scientist. In Chapter 3 we introduce the fundamental data types that data scientists need to understand when going about their daily job. Chapter 4 is a machine learning crash course for budding data scientists. In Chapter 5 we examine several topics of the authors' choice in data science that will enhance data scientists' knowledge and give them insight into typical applications they may come across during their work. Finally, in Chapter 6 we summarise the material we have covered in this introduction.

This page intentionally left blank

Index

Note: The italic letter *f* following page numbers refers to figures.

A

accuracy, of algorithms, 59
actors, in social networks, 44
actual value, 54
acyclic networks, 94
additive smoothing, 73
adjacency matrix, 152, 155
agent, 57
AGI (artificial general intelligence), 184
AI (artificial intelligence)
  conversational AI, 185
  neuro-symbolic AI, 184, 185
  trustworthy AI (TAI) characteristics, 186
Albert Einstein's citation vector, 142, 143
algorithm pseudocode
  DBSCAN, 124
  DT (decision tree), 89
  HAC (hierarchical agglomerative clustering), 111
  k-means, 107
  k-medoids, 109
  KNN (*k*-nearest neighbours), 69
  LP (label propagation), 127
  naive Bayes, 73
  perceptron, 98
  random surfer model, 140
  RF (random forest), 93
  SVM (support vector machines), 84
All of Statistics: A Concise Course in Statistical Inference (Wasserman), 189
An Introduction to Search Engines and Web Navigation (Levene), 190
Artificial Intelligence: A Modern Approach (Russell and Norvig), 191
artificial neural networks. *See* neural networks (NNs)
ArXiv research archive, 191
ASI (artificial superintelligence), 184
Asimov, Isaac, 1
aspect-based sentiment analysis, 171
attention weights, 104, 105
audio data, 40, 41
autocorrelation, 22, 28
autocovariance, 22, 28
axiomatic properties, 156–157

B

backpropagation, 98–99, 103
backpropagation through time (BPTT), 103
bagging, 90
bag-of-words assumption, 38, 118, 134
bag-of-words model, 72
Barabási, A. -L., 189, 192
Bayes' theorem, 4
Bayesian neural networks (BNNs), 187
bell curve, 12
Bengio, Y., 191, 192
Bernoulli distribution, 7–9, 9*f*, 29
Bernoulli naive Bayes model, 72
betweenness centrality, 155
bibliometrics, 140, 157
bigram language model, 137
binary independence model, 72
binning, 49
binomial distribution, 8–10, 10*f*
Bishop, C. M., 190, 192
black box and white box models, 85, 93, 105

"black-box" syndrome, 186
BNNs (Bayesian neural networks), 187
bootstrapping
  aggregation, 90
  resampling, 19–21
Box-Cox transformation, 48
BPTT (backpropagation through time), 103

C

calibrated probabilities, 60, 62
CAPTCHAs, 41
CB (content-based recommender), 145–146
CBOW (continuous bag-of-words) model, 179–183, 180*f*
CDF (cumulative distribution function), 4
centrality measures, 154–157
CF (collaborative filtering) recommender
  item-based CF prediction, 147–148
  k-nearest neighbour notation in, 146
  rating problems of, 149–150
  user-based CF prediction, 147
citation curve, 143*f*
citation vector, 140–143, 143*f*
city block norm, 63
Clarke, Arthur C., 1
classifier algorithm, 53
click rate, 132
cluster analysis, 106
clustering coefficient, 154
clustering measures, 64–66
CNNs (convolutional neural networks), 101–103, 102*f*
coefficient of determination, 28, 29
compartments, in epidemic models, 164
complementary CDF, 5
conditional probability, 4
confidence intervals, 19–21
confusion matrix, 58–60, 128
confusion matrix measures
  Bernoulli naive Bayes model, 74*f*
  decision tree classifier, 91*f*
  feedforward neural network, 100*f*, 101*f*
  Gaussian naive Bayes model, 76*f*
  hard margin SVM (support vector machine) binary classifier, 81*f*
  k-nearest neighbours (KNN) model, 70*f*
  logistic regression model, 61*f*
  multinomial naive Bayes model, 75, 75*f*
  perceptron, 99
  random forest classifier, 92*f*
  a two-class problem, 59*f*
connected components, 152
connected group, 123
continuous random variables, 3
conversational AI, 185
copyright infringement, 187
corellogram, 42
corpus, 38, 39, 117
correlation
  in collaborative filtering, 146
  between random variables, 26, 28
cosine similarity, 134
cost function, 54
Courville, A., 191, 192
covariance, 21–22
covariance matrix, 114
crawler algorithm, 132, 133*f*
CRFs (conditional random fields), 174, 175
curse of dimensionality, 53, 106
cyclic networks, 94

D

dangling web page, 139
data mining, 1
Data Mining: Practical Machine Learning Tools and Techniques (Witten, Frank, Hall, and Pal), 190
data science
  field of, 1, 2, 187, 188
  resources, 189
Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking (Provost and Fawcett), 190
data set
  definition, 1
  features, 53
  labelled, 52–54
  unlabelled. *See* unlabelled data set
data types
  geographical, 42–44
  image, video, and audio, 40–41

numeric, 34
social network nodes, 44–46
tabular, 33–38
textual data, 38–40
time series, 41–42
data visualisations
geostatistics time series, 43*f*
geostatistics variogram, 45*f*
image recognition data set, 41*f*
linear regression line fitted to the data, 38*f*
Pearson correlation scores heatmap, 36*f*
rank-order, 50*f*
rank-order distribution histogram, 51*f*
social network nodes, 46*f*
tabular data in a histogram, 37*f*
tabular data in a stacked bar chart, 34*f*
textual data rank-order distribution and Zipf plot, 40*f*
textual data word cloud, 39*f*
DBSCAN (density-based spatial clustering of applications with noise), 121–125, 123*f*, 125*f*
decision function, 53
deep fakes, 186
deep learning, 185, 191
Deep Learning (Goodfellow, Bengio, and Courville), 191
degree, of a node, 152
degree centrality, 154, 155
degree distribution, 159–162
delta rule, 97–98
dendrogram, 111
dendrogram of HAC, 114*f*
dependent variable, 28
destination links, 131
Digital Image Processing (Gonzalez and Woods), 190
dimensionality reduction, 53, 106, 113, 116
directed graph, 126, 152
discount factor, 156
discrete random variables, 3
discriminative models, 105
distance measures, 56–57, 105, 106
distributions
Bernoulli, 7–9, 9*f*, 29
binomial, 8, 10
degree, 159–162
exponential, 13–14, 14*f*, 158, 163
Gaussian, 47, 75, 76
nonparametric, 6
normal, 150. *See also* normal distributions
parameters, 5
parametric, 6
Pareto, 14–16
Poisson, 10–11
power law, 157–159
documents, 38, 117, 134
dot product, 77, 178
DT (decision tree)
algorithm construction, 88–90
algorithm pseudocode, 89
data structure, 85–92, 86*f*, 87*f*, 88*f*
node types in a tree data structure, 86*f*
Dunn Index, 65–66

E

ECDF (empirical cumulative distribution function), 5
EDF (empirical distribution function), 17, 162
eigenvectors, 115, 155–156
80-20 rule, 157, 164, 168
elbow heuristic, 107–108
elbow plot, 63, 64, 108*f*
Elo, Arpad, 143
Elo rating system, 143–145
empirical probabilities, 62
end-to-end machine learning, 57, 105
ensemble learning, 90
epidemic modelling
compartments, 164
parameters, 164–165
with SIR model, 170
with SIS model, 166–169
superspreaders, 168
threshold value, 167
epidemiology, mathematical, 164
error matrix. *See* confusion matrix
estimator, 17
Euclidean norm, 62
evaluation of algorithms
accuracy, 59
calibration, 60, 62

confusion matrix, 58, 59
discounted cumulative gain (DCG), 134
goal of, 57
mean average precision (MAP) scores, 134
precision, recall, and F1, 59–60
precision and recall, 133–134
for semi-supervised models, 127–128
for supervised models, 58–62
for unsupervised models, 62–68
events, defined, 3–4
expected value, 17–18
explainability (TAI characteristic), 186
explanatory variable, 28
exponential cutoff, 159
exponential distribution, 13–14

F

F1 score, 60
Facebook, 159
Fawcett, T., 190, 192
feature selection, 53
feedforward neural networks, 94, 94*f*, 100–102, 178
fine-tuning, 185, 187
first moment of *X*, 17
first-mover advantage, 163
first-rater problem, 149
Firth, John R., 171
fourth moment of *X*, 23
fractal behaviour, 159
Frank, E., 190, 193
Friedman, J., 189, 192
friendship paradox, 152
function, sigmoid, 97

G

GANs (generative adversarial networks), 186
Gaussian distribution, 11, 47, 75. *See also* normal distribution
Gaussian kernel, 84
Gaussian mixture model (GMM), 109
Gaussian naive Bayes, 73
general AI, 184
generative deep learning, 185
generative models, 105
generative neural networks
generative adversarial networks (GANs), 186
variational autoencoders (VAEs), 185
geographical data, 42
geometric potential gain, 155, 156
geostatistical analysis, 43–44
Gini impurity, 88, 89
GMM (Gaussian mixture model), 109
GNNs (graph neural networks), 105
Gonzalez, R. C., 190, 192
Goodfellow, I., 191, 192
graphical model, 118, 119, 119*f*
greedy algorithm, 153–154

H

Hall, M. A., 190, 193
Hanin, B., 191, 192
hard clustering methods, 109
hard margin, 78
harmonic mean, 60
Hastie, T., 189, 192
hat notation, 54
heatmap, 35
heavy-tailed, 158
Hebb, Donald, 105
hierarchical clustering
dendrogram of HAC, 113*f*
divisive hierarchical clustering, 111–112
hierarchical agglomerative clustering (HAC), 110–111
high bias, 56
high variance, 56
high-dimensional data sets, 113
histograms, defined, 5
Hopkins statistic, 67–68
hyperbolic tangent (tanh), 96
hyperlinks, 131
hyperparameters, of statistical models, 54
hyperplane, 77, 78

I

identity matrix, 34
image data, 40, 53, 101

independence, of events, 4
independent variable, 28
indexer, of web pages, 132
inductive learning, 126
inlinks degree distribution plot, 161*f*
internal node, 85, 87, 88
interval data, 34, 35
Introduction to Information Retrieval (Manning, Raghavan, and Schutze), 190
inversion attack, 187
IR (information retrieval) models
  citation vector, 140–143, 143*f*
  collaborative filtering recommender (CF), 146–150
  content-based recommender (CB), 145–146
  and conversational AI, 185
  Elo rating system, 143–145
  evaluation, 133–134
  statistical language models, 135, 137–138, 183
  vector space model, 134, 135
item-user matrix, 147
*i*th principal component, 115

J

joint probability, 4
Jurafsky, D., 190, 192

K

k = 6 clusters silhouette plot, 66*f*
k centroids, 62
k latent topics, 117
kde (kernel density estimation), 6
kernel trick, 84
K-factor, 144
k-fold cross-validation, 54, 58
k-means algorithm, 106–108
k-medoids algorithm, 108–109
k-nearest neighbours (KNN) algorithm, 68–71, 148
k-neighbourhood graph, 126
KNN (*k*-nearest neighbours) algorithm, 68–71, 148
knowledge graph semantic network, 175, 177
KS (Kolmogorov-Smirnov) statistic, 162
kurtosis, 5, 23

L

label propagation, 127–128
labelled data set, 52–54
lag, 22
large language models (LLMs), 187
layers, in neural networks, 94
LDA (latent Dirichlet allocation), 121
leaf nodes, 85–90, 86*f*, 87*f*
least squares, 28, 29
Levene, M., 190, 192
lexicon-based classifier, 172, 173
likelihood of events (Bayesian), 71
linear kernel, 84
linear relationship, 28
linkage criteria, 110
Liu, B., 190, 192
LLMs (large language models), 187
logarithmic binning, 160
logistic function, 29
logistic regression, 29, 31
logistic transformation, 48
logit function, 29
logit transformation, 48
log-log plots, 160, 161*f*
log-log transformation, 48
log-odds function, 29
long tails, 159
loss functions, 54–56
LSTM (long short-term memory) neural networks, 104, 104*f*
LSVM (linear support vector machines), 84

M

machine learning
  algorithms. *See* algorithms
  and artificial intelligence (AI), 184, 185
  and data science, 52–54
  definition of, 1
  evaluation process in, 57–68
  resources for study, 190
  training phase in, 54–56

macro-averaged F1, 60
Manhattan distance, 56, 109, 154
Manhattan norm, 63
Manning, C. D., 190, 192
marginal probability, 4
Markov model, 137
Martin, J. H., 190, 192
matrix factorisation, 119
max pooling, 103
maximum likelihood
  estimation, 31
  fitting, 5
  in power law distributions, 160–162
maxmin method, 107
mean, 18
measures, inter-cluster, 64
median, 18
micro-averaged F1, 60
Milgram, Stanley, 153
mode, 18
Monte Carlo sampling, 17
multinomial naive Bayes model, 72
Murphy, K. P., 191, 192

N

naive Bayes, 71–77, 105
naive Bayes classifier, 137, 173
narrow AI, 184
negative sampling, in word2vec algorithm, 181
negatively skewed, 24
neighbours (data)
  and the friendship paradox, 152–153
  in KNN algorithm, 68, 69
  labelling, 126
NER (named entity recognition)
  conditional random fields (CRFs), 174–175
  and conversational AI, 185
  knowledge graph semantic network, 175–177
  named entities as class objects, 171
  named entity detection, 174
network data structure, 44–46
Network Science (Barabási and Pósfai), 189
neuro-symbolic AI, 184, 185
90-10 rule, 162
NLP (natural language processing)
  and conversational AI, 185
  named entity recognition (NER), 174, 177
  sentiment analysis, 171–174
NMF (non-negative matrix factorisation), 120, 148–149
NMF model, 122*f*
NNs (neural networks)
  activation functions, 96–97
  applications, 42, 93
  and deep learning, 95
  large language models (LLMs) as, 187
  and neuro-symbolic AI, 185
  perceptron model, 95–96, 95*f*
  structure, 94–95
node types, 85, 86
nodes
  in decision tree data structure, 86*f*, 87*f*
  in neural networks, 94
  root, 87, 111
  seed, 126
  in social networks, 44, 130, 151, 151*f*, 153
  in tree data structures, 85, 86, 90
  in a web graph, 131
noise, 121–125, 123*f*, 160
nominal data, 34, 35
nonlinear transformations, 48
non-negative matrix factorisation (NMF), 120, 148–149
normal distributions
  central limit theorem, 11–12
  in Gaussian naive Bayes algorithm, 73
  and the Gaussian kernel, 84
  and histograms, 5, 6*f*, 12*f*
  and kurtosis, 23–26
  and power law distributions, 150, 157–158, 164
normalisation of the data set, 47
norms, of vectors, 62–63, 114
Norvig, P., 191, 192
nugget, in a variogram, 44

O

Occam's razor, 56
odds ratio, 29

one-hot encoding vector, 178, 179
ordinal data, 35
outliers, 19, 28, 108
outlinks, 131
overfitting, 56

P

page popularity measures, 138
PageRank, 138, 139, 140, 140*f*, 155
paid recommendations, 150
Pal, C. J., 190, 193
parameters
  defined, 17
  of distributions, 5, 6
  in epidemic modelling, 164–165
  of a statistical model, 54
Pareto, Vilfredo, 158
Pareto chart, 163*f*
Pareto distribution, 14–16, 16*f*, 158–159, 164
parts-of speech (POS) tagging, 174
paths, between nodes, 85, 152, 154
Pattern Recognition and Machine Learning (Bishop), 190
PCA (principal components analysis), 113–117, 116*f*, 117*f*
PDF (probability density function), 4, 158, 160
Pearson's correlation coefficient, 27
penalty term, 56, 78, 79
perceptron neural network, 95–96, 95*f*, 97–100
plate notation, 118, 119*f*
PLSA (probabilistic latent semantic analysis), 118–121
PMF (probability mass function), 4
pointwise mutual information, 67
Poisson distribution, 10–11
population, 17
POS (parts-of speech) tagging, 174
Pósfai, M., 189, 192
posterior events (Bayesian), 71
power law distributions, 157–159
predictions, as statistical model outputs, 52
predictor variable, 28
preferential attachment, 162, 163
presentation position bias, 132
prior events (Bayesian), 71, 118
probabilistic latent semantic analysis (PLSA), 118–121
Probabilistic Machine Learning: An Introduction (Murphy), 191
probabilistic topic modelling. *See* topic modelling
probability, 4
probability distributions, 4
problem definition, 1
prompt engineering, 187
Provost, F., 190, 192
pseudo-rating, 146, 150
Python Jupyter, 191

Q

querying a search engine, 131–132

R

radial basis function kernel, 84
Raghavan, P., 190, 192
random forest algorithm, 90, 92–93
random processes, 3
random surfer algorithm, 138–140
random variables, denotation of, 3
range, in a variogram, 44
rank sink, 139
rank transformation, 39, 40, 49, 51
ranking, of web pages, 132
rank-order distribution, 159
rating matrix, 146
ratio data, 34, 35
rectifier function, 96
regression analysis, 28–32
regularisation, 56
reinforcement learning, 57
reliability diagram, 62
RELU function, 96–97
resampling, 19
research papers (for download), 191
residual errors, 28
RNNs (recurrent neural networks), 103, 103*f*
Roberts, D. A., 191, 192
Russell, S., 191, 192

S

SA (sentiment analysis)
  aspect-based, 171
  lexicon-based classifier, 172, 173
  using a naive Bayes classifier, 137, 173
  using a vector space model, 173
sample, 17
scale-free distributions, 159
scaling values, 47, 146
scatter plot, 27
Schütze, H., 190, 192
search engine technology, 130–132, 132*f*, 187
second moment of *X*, 21
seed nodes, 126
self-similarity, 159
semi-log transformation, 48
semi-supervised algorithms
  graph-based learning, 126–127, 128
  label propagation (LP), 127–128
  self-learning, 126, 128
Sentiment Analysis: Mining Opinions, Sentiments, and Emotions (Liu), 190
sentiment analysis (SA)
  aspect-based, 171
  lexicon-based classifier, 172, 173
  using a naive Bayes classifier, 137, 173
  using a vector space model, 173
sentiment orientation distance measures, 172
sequence to sequence (seq2seq) models, 104–105
70-30 split, 58
shortcuts, 154, 154*f*
sigmoid function, 96, 97, 178
silhouette coefficient, 65–66
sill, in a variogram, 44
SIR (susceptible-infected-recovered) epidemic model, 166–168, 170*f*
SIS (susceptible-infected-susceptible) epidemic model, 165–166, 165*f*, 168, 169*f*
"six degrees of separation", 153
skewness, 5, 23–26
skip-gram neural network, 178–179, 179*f*
slack variables, 78
small-world problem, 153, 154
social network data, 44, 46
social networks
  degree distribution, 159, 160
  distribution of friendship, 150, 153
  epidemic modelling, 164, 170
  as greedy algorithm, 153, 154
  structure as undirected graph, 151–152
  visualisations, 47*f*, 151*f*
soft clustering, 109
soft clustering method, 109
soft margin, 78
source links, 131
spam counter-measures, 138
sparsity problem, 149
spatial data, 42, 44, 121
spatial random variables, 43
spectral radius, 156
Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Jurafsky and Martin), 190
speech recognition, 185
spider. *See* crawler algorithm
square matrix, 34
SSB (between-cluster sum of squared error), 64
SSE (sum of squared error), 63*f*, 64
SST (total sum of squared error), 64
SSW (within-cluster sum of squared error), 64
standard deviation, 21
standard error, 21
statistical models
  built by machine learning, 1, 2, 52
  as a decision function, 53
  language models, 135, 137–138, 183
  overfitting/underfitting, 56
statistical visualisations
  Bernoulli distribution fitted to data set, 9*f*
  binomial distribution fitted to the data set, 10*f*
  bootstrap computation, 20*f*
  boxplot of mean and median values, 19*f*
  confidence intervals, 20*f*
  excess kurtosis, 26*f*
  exponential distribution fitted to the data, 14*f*
  histogram, PDF, ECDF, and CDF, 6*f*
  kde plot overlay on histogram, 7*f*

least squared linear fit and residual errors, 30*f*
mean and median confidence intervals, 22*f*
negatively skewed data set, 25*f*
normal distribution fitted to the data set, 12*f*
Pareto distribution fitted to the data, 16*f*
positive and negative correlation, 27*f*
positively skewed data set, 24*f*
regression curve fitted to the data, 31*f*
statistics
defined, 17
resources for study, 189
Statistics (Witte and Witte), 189
stop words, 38
super AI, 184
superspreaders, 168
supervised algorithms
decision tree (DT), 89–93
k-nearest neighbours (KNN), 68–71
naive Bayes, 71–77
neural networks (NNs), 93–105
random forest model, 92–93
support vector machines (SVM), 77–84
supervision of learning algorithms
semi-supervised, 54
supervised, 53
unsupervised, 53
surfing (the web), 138
survival function, 5
SVM (support vector machines)
algorithm pseudocode, 84
applications, 77
evaluation, 81–83
hard and soft margins, 79*f*, 80*f*
nonlinear kernel, 83*f*
soft margin classification, 81*f*
structure, 77–78
symbolic reasoning, 185
symmetric matrix, 152

T

tabular data, 33–38
TAI (trustworthy AI) characteristics, 186–187
tanh (hyperbolic tangent), 96
teleportation, 139
textual data analysis, 38, 40
TF (term frequency), 134
TF-IDF (term frequency inverse document frequency), 135, 136–137
TF-IDF (term frequency-inverse document frequency), 137
The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Hastie, Tibshirani, and Friedman), 189
The Principles of Deep Learning Theory (Roberts, Yaida, and Hanin), 191
third moment of *X*, 23
threshold value, 167
Tibshirani, R., 189, 192
time series data, 22, 41, 42
Tobler, Walter, 43
topic coherence, 67, 120
topic modelling, 66–67, 117–121
total sum of squared error (SST), 64
train and test split, 58
training, in machine learning, 52, 54–56, 57, 58, 95, 96
transductive learning, 126
transfer diagram, 165*f*, 166*f*
transfer learning, 185, 187
transformation of data
binning, 49
Box-Cox, 48
linear, 47–48
nonlinear, 48
purpose of, 46
rank-order, 49–51
unlabelled to labelled, 62–63
tree data structure, 85
trigram model, 137
trustworthy AI (TAI) characteristics, 186–187
Twitter, 159

U

unbiased estimator, 18
uncertainty quantification, 186, 187
underfitting, 56
undirected graph, 126, 152
unigram language model, 137
unlabelled data set
role in algorithm training, 52–54, 125–127

transformation to labelled pairs, 62, 106
use in DBSCAN, 122
use in large language models (LLMs), 187
use in recommender systems, 146
use in unsupervised models, 105
unsupervised algorithms
DBSCAN, 121–125
hierarchical clustering, 110–113
k-means, 106–108
k-medoids, 108–109
principal components analysis (PCA), 113–117
topic modelling, 117–121
user-user correlation matrix, 148, 149

V

VAEs (variational autoencoders), 185, 186
validation, external, 63
vanishing gradient problem, 97, 103, 104
variance
autocovariance, 22, 28
covariance matrix, 114
definition of, 21
high variance, 56
in power law distributions, 158
in principal components analysis (PCA), 115–116
skewness, 23–26
Ward's minimum variance criterion, 110
variational autoencoders (VAEs), 185, 186
variogram, 44
vector space model, 134–135, 173
video data, 40

W

Ward's linkage criterion, 110
Wasserman, L., 189, 193
weak supervision methods, 128
web graph, 131
web pages, 131
weighted links in neural networks, 94
weighting predictions, 150, 152
Witte, J. S., 189, 193
Witte, R. S., 189, 193
Witten, I. H., 190, 193
Woods, R. E., 190, 192
word cloud, 38
word embeddings
construction of, 177, 178
in sentiment analysis, 172
visualisation of, 182*f*
word2vec algorithm, 178–183
word2vec algorithm
continuous bag-of-words (CBOW) model, 179–183
skip-gram neural network, 178–179
WWW (World Wide Web), 131

Y

Yaida, S., 191

Z

Zachary, Wayne, 151
Zipf, George Kingsley, 159
Zipf plot, 39
z-scores, 47