Contents

Figures ix

Tables xi

Boxes xiii

Preface xv

1 What is data science? 1
Data, information, knowledge, wisdom 1
Data everywhere 3
The data deserts 4
Data science 5
The potential of data science 8
From research data services to data science in libraries 10
Programming in libraries 14
Programming in this book 15
The structure of this book 16

2 Little data, big data 17
Big data 17
Data formats 19
Standalone files 20
Application programming interfaces 21
Unstructured data 25
Data sources 27
Data licences 37

3 The process of data science 39
Modelling the data science process 39
Frame the problem 41
Collect data 44
Transform and clean data 46
Analyse data 48

Visualise and communicate data 50
Frame a new problem 54

4 Tools for data analysis 55
Finding tools 55
Software for data science 56
Programming for data science 69

5 Clustering and social network analysis 77
Network graphs 77
Graph terminology 79
Network matrix 80
Visualisation 82
Network analysis 85

6 Predictions and forecasts 97
Predictions and forecasts beyond data science 97
Predictions in a world of (limited) data 99
Predicting and forecasting for information professionals 101
Statistical methodologies 102

7 Text analysis and mining 113
Text analysis and mining, and information professionals 113
Natural language processing 115
Keywords and n-grams 125

8 The future of data science and information professionals 133
Eight challenges to data science 134
Ten steps to data science librarianship 139
The final word: play 144

References 147

Appendix – Programming concepts for data science 165
Variables, data types and other classes 165
Import libraries 167
Functions and methods 168
Loops and conditionals 170
Final words of advice 171
Further reading 172

Index 173