

Contents

1	“Data Science” and “Big Data”	1
1.1	Transport Data Science Examples	1
1.1.1	Origin-Destination Analysis on the London Orbital Motorway	1
1.1.2	Airline Pricing and Arbitrage	4
1.1.3	Pothole Monitoring	4
1.1.4	Foursquare	5
1.1.5	Self-driving Cars	5
1.1.6	Taxi Services	5
1.2	The Claim	5
1.3	Definitions	7
1.4	Relationship with Other Fields	7
1.5	Ethics	10
1.6	Cynical Views	10
1.7	Exercise: <i>itsleeds</i> Virtual Desktop Setup	11
1.8	Further Reading	13
1.9	Appendix: Native Installation	14
2	Python for Data Science Primer	15
2.1	Programming Skills Check	15
2.2	Programming Languages	17
2.3	Programming Environment	17
2.4	Core Language	18
2.4.1	Lists	19
2.4.2	Dictionaries	20
2.4.3	Control Structures	20
2.4.4	Files	20
2.4.5	Functions	21
2.5	Libraries	21
2.5.1	Modules	21
2.5.2	Mathematics	22
2.5.3	Plotting	23
2.5.4	Data Frames	23
2.5.5	Debugging	24
2.6	Further Reading	24

3	Database Design	27
3.1	Before the Relational Model	27
3.2	Picturing the World	28
3.2.1	Ontology	29
3.2.2	Philosophical Ontology	29
3.2.3	Data Ontology	34
3.2.4	Structured Query Language (SQL)	37
3.3	Exercises	37
3.3.1	Setting up PostgreSQL	37
3.3.2	SQL Creation Language	37
3.3.3	SQL Query Language	38
3.3.4	SQL Python Binding	40
3.3.5	Importing Vehicle Bluetooth Data	40
3.4	Further Reading	41
4	Data Preparation	43
4.1	Obtaining Data	43
4.2	Basic Text Processing	44
4.3	Formal Grammar: The Chomsky Hierarchy	44
4.3.1	Regular Languages (Type 3)	45
4.3.2	Context-Free Languages (Type 2)	46
4.3.3	Beyond CFGs (Types 1 and 0)	47
4.4	Special Types	48
4.4.1	Strings and Numbers	48
4.4.2	Dates and Times	48
4.4.3	National Marine Electronics Association (NMEA) Format	50
4.5	Common Formats	50
4.6	Cleaning	51
4.7	B + Tree Implementation	52
4.8	Exercises	53
4.8.1	Reading the Database with Pandas	53
4.8.2	<i>printf</i> Notation	53
4.8.3	DateTimes	54
4.8.4	Time Alignment and Differencing	54
4.8.5	Parsing	55
4.8.6	Vehicle Bluetooth Munging	55
4.9	Further Reading	55
5	Spatial Data	57
5.1	Geodesy	57
5.2	Global Navigation Satellite System (GNSS)	59
5.3	Geographic Information Systems (GIS)	61
5.3.1	Role of GIS System	61
5.3.2	Spatial Ontology	62
5.3.3	Spatial Data Structures	63

5.4	Implementations	65
5.4.1	Spatial Files	65
5.4.2	Spatial Data Sources	65
5.4.3	Spatial Databases	66
5.4.4	Spatial Data Frames	66
5.5	Exercises	66
5.5.1	GPS Projections	66
5.5.2	PostGIS	66
5.5.3	GeoPandas	67
5.5.4	QGIS Road Maps	68
5.5.5	Plotting Open Street Map (OSM) Roads	69
5.5.6	Obtaining OSM Data	70
5.5.7	Bluetooth Traffic Sensor Sites	70
5.6	Further Reading	71
6	Bayesian Inference	75
6.1	Bayesian Inference Versus “Statistics”	75
6.2	Motorway Journey Times	76
6.3	Bayesian Inference	78
6.3.1	Bayes’ Theorem	78
6.3.2	Legal Inference: A Pedestrian Hit-and-Run Incident	78
6.3.3	Priors and Posteriors	79
6.3.4	Road User Tracking	80
6.4	Bayesian Networks	80
6.4.1	Bayesian Network for Traffic Lights	81
6.4.2	Bayesian Network for Traffic Accidents	82
6.4.3	Reporting	83
6.4.4	Car Insurance	84
6.5	Priors and Prejudice	85
6.6	Causality	86
6.7	Model Comparison and Combination	89
6.8	Exercises	90
6.8.1	Inferring Traffic Lights with PyMC3	90
6.8.2	Inferring Accident Road State Change with PyMC3	90
6.8.3	Switching Poisson Journey Times	91
6.9	Further Reading	91
7	Machine Learning	93
7.1	Generative <i>Versus</i> Discriminative Vehicle Emissions	93
7.2	Simple Classifiers	95
7.2.1	Linear Discriminant Analysis (LDA)	95
7.2.2	Nearest Neighbor	95
7.2.3	Template Matching	96
7.2.4	Naïve Bayes Classification	96
7.2.5	Decision Trees	97

7.3	Neural Networks and “Deep Learning”	99
7.3.1	Parallel Computing Back-Propagation	102
7.4	Limitations and Extensions	103
7.5	Exercises	105
7.6	Further Reading	106
8	Spatial Analysis	107
8.1	Spatial Statistics	108
8.2	Bayesian Spatial Models	109
8.2.1	Markov Random Fields (MRF)	109
8.2.2	Gaussian Processes (Kriging)	113
8.3	Vehicle Routing	114
8.3.1	Link-breaking	115
8.4	Spatial Features	115
8.5	Exploratory Analysis	116
8.6	Scaling Issues	118
8.7	Exercises	119
8.7.1	Gaussian Processes in GPy	119
8.7.2	Gaussian Process Traffic Densities	122
8.7.3	Vehicle Routing with PostGIS	122
8.7.4	Finding Roadside Sensor Sites	122
8.8	Further Reading	123
9	Data Visualisation	125
9.1	Visual Perception	125
9.1.1	Colours	125
9.1.2	Visual Attention	126
9.2	Geographic Visualization (Maps)	131
9.2.1	Traffic Flow Maps	133
9.2.2	Slippy Maps	136
9.2.3	Info-Graphics	137
9.3	Exercises	141
9.3.1	Web Page Maps with Leaflet	141
9.3.2	Bluetooth Origin-Destination Flows	143
9.3.3	Large Project Suggestions	144
9.4	Further Reading	145
10	Big Data	147
10.1	Medium-Sized Data Speedups	147
10.2	Enterprise Data Scaling	149
10.3	CAP Theorem	151
10.4	Big Data Scaling	152
10.4.1	Data “Lakes”	152
10.4.2	Grid Computing	153
10.4.3	Map-Reduce and Cloud Computing	154
10.4.4	Hadoop Ecosystem	155
10.4.5	Non-relational Databases (“NoSQL”)	156
10.4.6	Distributed Relational Databases (“NewSQL”)	157

- 10.5 Exercises. 158
 - 10.5.1 Prolog AI Car Insurance Queries. 158
 - 10.5.2 Map-Reduce on Vehicle Bluetooth Data 158
 - 10.5.3 Setting up Hadoop and Spark 160
 - 10.5.4 Finding Vehicle Matches in Hadoop 160
 - 10.5.5 Traffic Flow Prediction with Spark 161
 - 10.5.6 Large Project Suggestions 164
- 10.6 Further Reading 164
- 11 Professional Issues 165
 - 11.1 Morals, Ethics, and Law 166
 - 11.2 Ethical Issues 167
 - 11.2.1 Privacy. 167
 - 11.2.2 De-anonymization (“Doxing”) 168
 - 11.2.3 Predictive Analytics. 170
 - 11.2.4 Social and Selfish Equilibria 170
 - 11.2.5 Monetization 171
 - 11.2.6 Ontological Bias 173
 - 11.2.7 *p*-hacking. 174
 - 11.2.8 Code Quality 176
 - 11.2.9 Agency Conflicts 177
 - 11.2.10 Server Jurisdiction 177
 - 11.2.11 Security Services 177
 - 11.3 UK Legal Framework 178
 - 11.3.1 Data Protection Act 1988. 178
 - 11.3.2 General Data Protection Regulation (GDPR) 179
 - 11.4 Role of the Data Scientist. 180
 - 11.5 Exercises. 181
 - 11.6 Further Reading 181
- Index 183