

Index

A page number in **bold** indicates a box or table while *italics* denotes a figure.

- ABC programming language, **76**
- abstraction, 20, 21, 36
- Adams, S., 270
- Advanced Step in Innovative Mobility (ASIMO), **290**
- age, 133, *134*, 216, 330, 332
- agglomerative clustering, 271, 336
 - bottom-up approach of building clusters, 271
 - common problem, 272
 - computation of distance matrix, **272**
 - data points, 271
 - dendrogram (maximum distance), 273
 - dendrogram (minimum distance), 273
 - distance calculation formula, 272
 - hands-on examples, 271–275
 - number of clusters when threshold distance set at nine, 273
 - number of clusters when threshold distance set at six, 274
 - plotting result, 275
- Akaike information criterion (AIC), 279, 282, 287
 - comparing models, 321, 322
 - glossary, 296
- algorithmic bias, **101**, 288
- algorithms, 100, 187, **197**, 198
 - ML techniques (application), 203–205
- alternative hypothesis, **91**, 91, 139, 143
 - glossary, 146
- Amazon, 186, **198**
 - value of each user, 31
- Amazon Elastic Compute Cloud (EC2), 169, 176
- Amazon Linux AMI, 169, 171
- Amazon Web Services (AWS), xvi
 - cloud computing, 169–176
 - management console, 169
 - using Python (hands-on example), 172
 - virtual machine (creation), *170*
 - virtual machine (launch), *170*
 - working through Cloud 9, *173*
 - working with ~, *173–175*
- Amazon Web Services (AWS) Cloud9, 98, 173, 176
 - avoidance of charges, 95
 - browser-based IDE, *172*
 - building ML model, 175
 - “can be free,” 169
 - command line capabilities, 168
 - Python installation, 175
 - working with AWS, *173*
- Amazon Web Services (AWS) EC2 instance
 - connected to SSH session, *172*
 - connecting from PC using PuTTY, *171*
- Anaconda Navigator, 113, 114, 344, 365
 - installation, 363
 - screenshot, *113*
 - snapshot, *364*
- ANAEROB dataset, 91
- analysis of variance (ANOVA), 143–144, 149
 - extension of t-test (*qv*), 143
 - hands-on example, 143–144
 - three-step process, 143–144
 - try it yourself, 144
- analytic dashboard, 92
- analytics: types, 15
- analyzing data
 - try it yourself, 29
- analyzing data (hands-on examples), 26–29
- anomaly detection
 - glossary, 261
 - machine learning, **260**
- Apple Watch, 9
- application programming interfaces (APIs), 47, 301, **355–356**
 - glossary, 68, **356**
 - Reddit, 336–340
 - YouTube, 342–346

- area under curve (AUC), 219, 220
- Argonne National Laboratory, 11
- arithmetic operators, 115
- Arminger, G., 267
- artificial intelligence (AI), xi, 77
 - “broader than ML or data mining,” 189
 - data science, 18–20
 - definition, 36
- Association of College and Research Libraries, 12
- association rules, 236–238
 - accuracy (confidence), 236
 - support, 236
 - weather data (decision rules, decision tree), 237
 - weather data (training dataset), **237**
- auto insurance, 39
- automated decision-making (ADM), **101**
- balloons datasets, 90, 238
 - decision tree (final version), 234
 - how decision tree works, **231**
- batch gradient descent, 198
 - glossary, 103
- Bayes’ theorem, 246, 248, 361
- Bayesian information criterion (BIC), 279, 282, 287, 290
 - comparing models, 322
 - glossary, 296
- Bayesian network: anomaly detection, **260**
- Belsley, D. A., 296
- Berners-Lee, T., 43
- best practices, 32, 34–35
- bias, 30–35, **287–289**
 - algorithmic, **101**
 - ML, 203
 - personal daily data consumption, 288
 - sources, 288
 - versus fairness, 289
- Bieraugel, M., 12
- Big Brother, 316, **355**
- big data, 11, 12, 156, 165, **294**
 - Yelp reviews and ratings, 349–354
- Biometrika* (journal), 142
- black-box approach, 245
- Bogalusa Heart Study, 268
- bookshelf (sorting), 21–22
- Boolean values, 117
- bootstrap sampling, 241
- Boston: Office of New Urban Mechanics, 11
- boxplots, 89
 - interquartile range, 89
- brainstorming, 80, 312
- brand loyalty, 101, **188**
- business analytics
 - definition, 15, 36
 - relation to data science, 15
- Cambridge Analytica, 8, 31
- cancer, 81, 203
- canonical hyperplane, 254, 255
- carbon emissions, 10, 98, 179
- Carnegie Mellon University, 70
- categorical variables, 80, 83, 217
- Catlett, C., 11
- causal analysis
 - glossary, 103
 - same as “diagnostic analytics” (*qv*), 91
- causation
 - definition, 102
 - example, 93
 - glossary, 102
 - versus correlation, **92–93**
- census data, 77, 97
 - population map (US, 2021), 79
- central processing units (CPUs), 14, 154, 169
- centroids, 83, 87, **102, 278, 353**
- chance, 90, 139
- Char, D. S., **203**
- Chicago, 7, 46, 89
- classification, 81, 245, 260, 319, 320
 - kNN, 225–230
 - random forest, 334–335
- classification rule, 236
- Clickworker, **312**
- climate change, 98
 - data science domain, 10
- clinical data
 - classification with random forest, 334–335
 - clustering with *k*-means, 335
 - distribution of age per disease, 333
 - loading, 330–332
 - online appendices, **331**
 - problems (five-step process), 330–335
 - regression with gradient descent, 332–334
 - try it yourself, 354
 - visual exploration, 332
- Clinton, H., 8, 13

- cloud computing, 151–179, **295**
 - advantages, 151–152
 - Amazon Web Services, 169–176
 - certification (not miracle for job-hunting), **97**
 - further reading, 179
 - GCP, 152–161
 - hands-on problems, 178–179
 - key terms, 177–178
 - Microsoft Azure, 161–169
 - moving between platforms, 176–177
 - practical worth, **172**
 - summary, 177
 - used at your job, **100**
 - vampire charges, **163**
- cloud platforms, xii, xvi, 151, 175
- cloud services, 97, 151, 152, 161, 173, 175, 177
- cloud: definition, 151
- clustering, 271, 353
 - glossary, 295
 - hands-on problems, 296–297
 - k*-means, 335
 - try it yourself, 279
- Coagmento, 314, 327
- Coase, R., 132
- codes of conduct, 32, 34–35
- cognitive bias, 288
- collaborative filtering (CF), 88
 - glossary, 103
 - ML-based, **187**
- comma-separated values (CSV), 27, 30, 49–50, 132, 314
- complete data case, 279
- computational social science, 15–16
- computational thinking, xii, 20, 29
 - definition, 20, 36
 - hands-on examples, 21–22
 - three-stage process, 20
- Compute Engine, 152
- computer science, xi–xiv, 1, 18
 - overlap with data science, 14
- Computer Statistics (CompStat), **137**
- conceptual introductions, xii, 1–106
 - data, 43–73
 - data science, 3–40
 - techniques, 75–106
- confidence interval, 137, 141
- confusion matrix, 244, 251
- constant comparison, 318
 - glossary, 326
- constants, 360
- construction, 14, 15
- context, 18
- continuous data, 86, 102
 - glossary, 68
- contrastive language-image pre-training (CLIP), 19
- control structures
 - Python, 120–121
 - try it yourself, 121
- conversational agents: usage with decision trees, 234–235
- convolutional neural networks (CNN or ConvNet), **295**
- correlation, 138, 189
 - definition, 92
 - diagnostic analytics, 92–94
 - example, 93
 - glossary, 102
 - hands-on example, 93–94, 138
 - spurious, **94**
 - statistical inference, 102
 - try it yourself, 94, 138
 - versus causation, **92–93**
- correlation analysis, xii, 1, 58, 98, 145, 351
- cost function, 85, 86, 199, 202
- “counter” (variable), 122
- COVID-19 pandemic, 10, 98
- credit card, 87, 152
- cross-validation technique, 205, 324
 - glossary, 327
 - holdout method, 82
 - k*-fold method, 82
 - leave-one-out method, 324
- crowdsourcing services, **91**
- Cuban, M., 210
- cube (function), 123, 124
- customer data, xv, 44, 56, 132
- customer relationship management, 96
- customer segmentation, **252**
- data, 1, 17, 43–73
 - amount and nature, 357
 - bias, **54**
 - definition, 36
 - ethics and privacy, 354–355
 - further reading and resources, 73

- data (cont.)
 - key terms, 67–68
 - mass production, 19
 - “often dirty” in real world, **54–55**, 54
 - online appendix, 43
 - problems (hands-on solutions), 329–359
 - “raw material from which information obtained,” 43
 - summary, 67
 - terminology usage in this book, 4
 - volume (increase), 6
- data analysis, **77**, 301, 303
 - distinguished from “data analytics,” 76–77
 - glossary, 103
 - summary, 326
- data analysis techniques, 75–106, 315–319
 - classification, 77
 - descriptive analysis, 77–91
 - further reading and resources, 107
 - hands-on problems, 104–106
 - key terms, 102–103
 - mechanistic analysis, 98–99
 - mixed method studies, 318–319
 - qualitative methods, 317–318
 - quantitative methods, 316–317
 - summary, 100–101
- data analysts, 24–25
- data analytics, xiv, 1
 - bias and inclusion, 34
 - differences from ‘data analysis’, 76–77
 - fairness, **32**
 - glossary, 102
- data analytics firms
 - functions, **77**
 - revenue-generation activities, **77**
- data analytics techniques
 - diagnostic analytics, 91–94
 - exploratory analysis, 97–98
 - predictive analytics, 95–96
 - prescriptive analytics, 96–97
- data bias, 288
- data cleaning, 55–57, **59**, 61, **77**, 95, 348
 - data munging, 56
 - data wrangling, 61
 - handling missing data, 56–57, 61
 - smooth noisy data, 57, 61
- data collection, xiii, **77**, 95, 301
 - finding users, 312
- data collection methods, 304–314
 - heat map (eye-tracking data), 315
 - interviews and focus groups, 309–312
 - log and diary data, 312–313
 - surveys, 304–309
 - user studies in lab and field, 314
- data collections, 46–49
 - open data, 46–47
- data conversion error, **116**
- data cube
 - aggregation, 58
 - definition, 58
 - glossary, 68
- data discretization, 59
 - data pre-processing, 65
- data engineers, 12, 25, 380
- data evaluation, xiii, 85, 303
- data evaluation (comparing models), 319–324, 326
 - AIC, 321
 - BIC, 322
 - cross-validation, 324
 - F*-measure, 85
 - precision metric, 320
 - recall, 320–321
 - receiver operating characteristic curve, 321
 - testing variable A against variable B, 323–324
 - training and testing, 322–323
- data experimentation, xiii, 23, 301, 303
- data integration, 57–58, 61
- data literacy, xi
 - definition, 23
 - importance, **24**
- data mining, xii, 188–189
 - overlaps with ML, **188**
 - properties, **188**
- data munging, 56
- data overfitting, 76, 102, 204, 240, 260, 261, 322, 323, 335
 - glossary, 261
- data points, 76, 87, 90, 190, 198, 225, 250
 - agglomerative clustering, 271
 - without labels, 271
- data pre-processing, 1, 54–67, 348
 - assault and murder dataset (US), 68–70
 - data cleaning, 55–57, 61
 - data discretization, 59, 65

- data integration, 57–58, 61
- data reduction, 58–59, 64
- data transformation, 58, 63
- forms, 55
- hands-on example, 60–65
- hands-on problems, 68–73
- Pittsburgh bridges dataset, 70–72
- try it yourself, 66–67
- UNICEF child mortality dataset, 72–73
- wine consumption and mortality, 60–65
- data processing, xi, 29, 34, 67, 156, 162
- data reduction, 58–59
 - data pre-processing, 64
- data science, xi
 - AI, 18–20
 - best practices and codes of conduct, 34–35
 - bias and fairness, **287–289**
 - bias, ethics, privacy, 30–35
 - data, 43–73
 - datum, data, science (definitions), 4
 - definition, 5
 - definition (concise), 36
 - “fancy term for statistics,” 76
 - hand-on-problems, 37–40
 - hands-on approach, xiii
 - importance, 5
 - information schools (iSchools), 18
 - introduction, 3–40
 - key terms, 36
 - machine learning, 270–298
 - mathematics, **197**
 - nature (definitions and notions), 3–6
 - nature, applications, context, xii, 1
 - real-life problems, xiii
 - relationship with information science, 16–18
 - roles, 26
 - skills, 23–24
 - summary, 35–36, 325
 - techniques, 1, 75–106
- data science (domains), 6–12
 - climate change, 10
 - education, 11–12
 - finance, 6–7
 - healthcare, 9–10
 - libraries, 12
 - politics, 8–9
 - public policy, 7–8
 - urban planning, 10–11
- data science (relation to other fields), 13–16
 - business analytics, 15
 - computational social science, 15–16
 - computer science, 14
 - engineering, 14–15
 - social science, 15–16
 - statistics, 13–14
- Data Science for Social Good (DSSG), 8
- data science in practice, xiii
 - analytics firms, **77**
 - approaching data problems in field, 348–349
 - cloud computing (value to students), **172**
 - cloud platforms (used at your job), **100**
 - correlations (spurious), **94**
 - data cleaning, **59**
 - data types (power), **116**
 - data visualization to rescue, 137
 - debugging loops, **122**
 - decision trees and conversational agents, 234–235
 - finding users for data collection, 312
 - gradient descent in industry, **198–199**
 - hypothesis testing, **139–140**
 - job labels and skills, **12**
 - logistic regression in industry, 221
 - naïve Bayes, 251
 - Python at job, **127**
 - reinforcement learning for robots, **291**
 - social media data analysis (societal issues), 341–342
 - user testing, **314**
 - vampire charges, **163**
- data science jobs, 24–25, 380–384
 - categories, 380
 - corporate retail and sales, 381
 - data analysts, 24–25
 - data wrangling, 25
 - health and human services, 383–384
 - legal sector, 381–382
 - marketing, 381
- data science tools, 29–30
- data scientists, 23, 380
 - data-driven companies, 25
 - ethical concerns, 32
 - roles, 5
 - “we are data, data is us” companies, 25
- data sharing, 4
- data storage and presentation, 49–54

- data storage and presentation (cont.)
 - comma-separated values, 49–50
 - extensible markup language, 50–51
 - JavaScript Object Notation, 53–54
 - really simple syndication, 51–53
 - tab-separated values, 50
- data structures
 - DataFrames, 118
 - definition, 117
 - dictionaries, 118
 - hands-on example, 118–119
 - lists, 117
 - Python, 117–119
 - sets, 118
 - try it yourself, 119
 - tuples, 118
- data supply chain, 33–34
- data transformation, 58
 - data pre-processing, 63
 - processes, 58
- data tuples, 157, 253
- data types, 44–46
 - customer data sample, **44**
 - definition, 117
 - power, **116**
 - structured data, 44–45
 - unstructured data (challenges), 45–46
- data users: concerns (bias, ethics, privacy), 31
- data visualization, 136, **137**, 145
- data warehouses, 4, 44
- data wrangling, 25, 56
 - data cleaning, 61
 - for recipe, **56**
- database (DB) systems, 14
- DataFrames, 119, 125, 129, 136, 138, 149, 199, 229, 376
 - data structure (Python), 118
 - glossary, 146
- datasets, xiii, 19. *See* online appendices
 - anomaly detection, **260**
- Davenport, T. H., 5
- debugging loops, **122**
- decision nodes, 231, 234
- decision rules, 235–236, 237
 - derivation, 235
- decision tree algorithm, 231, 232, 239, 323
 - “big problem,” 240
- decision trees, 230–239, 260, 336
 - algorithms that generate $\sim$, 82
 - anomaly detection, **260**
 - association rules, 236–238
 - based on entropy and information gain (three-step process), 234
 - classification rule, 236
 - decision nodes and leaf nodes, 231
 - decision rule, 235–236
 - dynamic adaptation, **235**
 - efficiency, **235**
 - final (for balloons dataset), 234
 - frequency tables, 95
 - guiding conversations, **234**
 - hands-on example, 238–239
 - hands-on problems, 265–266
 - plotting, 239
 - random forest data, 242
 - selection of attributes, **241**
 - try it yourself, 83
 - usage with conversational agents, 234–235
 - “used for classification problems,” 230
- decomposition, 20, 21, 36
- deep learning, 294–295
 - definition, **295**
 - “impressive results,” **295**
- demographic transition, 264
- dendrograms
 - agglomerative clustering, 275
 - agglomerative clustering (maximum distance), 273
 - agglomerative clustering (minimum distance), 273
 - glossary, 295
- dependent variables, 80, 188, 206, 224
 - glossary, 102
- descriptive analysis, 77–91
 - definition, 77
 - dispersion of distribution, 87–91
 - frequency distribution, 81–84
 - glossary, 103
 - key points, 77
 - measures of centrality, 85–87
 - variables, 78–81
- descriptive statistics, 145, 317
 - glossary, 146
- diagnostic analytics
 - correlations, 92–94
 - data analytics technique, 91–94
 - glossary, 103
- dictionaries, 119, 129
 - data structure (Python), 118

- differential calculus, *xiii*
 - formulas, 360–361
- dimensionality reduction, 59
- disease prediction, **252**
- disinformation: definition, **324**
- dispersion of distribution, 87–91
 - interquartile range, 88–89
 - range, 87
 - standard deviation, 90–91
 - variance, 90
- distance matrix: computation (agglomerative clustering), **272**
- distributed computing, **156, 295**
- distributions
 - comparison, **91**
 - dispersion (descriptive analysis), 87–91
 - try it yourself, 84
- divisive clustering, 271, 275–279
 - hands-on example, 276–278
 - “works in top-down mode,” 99
- Doll, R., 322
- Domingos, P., 183
- DS. *See* data science
- Eclipse, 107, 128, 363, 365
- e-commerce: gradient descent, **198**
- education: data science domain, 11–12
- educators, 11, **24, 24**
- Edwards Deming, W., 329
- Einstein, A., 316, 318
- electronic health records (EHRs), 9
- else-if (elif) sequences, 120
- email, 307
 - “unstructured data,” 45
- engineering: relation to data science, 14–15
- ensemble methods, 240
- ensemble model, 187, 260
 - glossary, 262
- entropy, 233
 - decision tree (three-step process), 234
 - definition, 231
 - depiction, 232
 - formula, 231
 - glossary, 261
 - using frequency tables (one attribute, two attributes), 233
- error function, 85
 - formula, 195
 - generalization, 197
- error surface, *196*
- error values, *195*
- ethics, *xiii, xv, 19, 35, 354–355*
 - data scientists, 32
 - data users, 31
 - ML bias, 203
- Euclidean distance, 229, 271, 277
- European Union, 316
- evaluation. *See* data evaluation
- Excel, 25, 82, 307
 - organization of survey results, **103**
- expectation maximization (EM), 85, 279–287, 336
 - classification plot, 285
 - density plot, 286
 - diabetes dataset, 280
 - hands-on example, 279–280
 - hands-on problems, 297–298
 - pairwise scatterplots showing classification, 281
 - plot showing number of components and corresponding BIC, 283
 - try it yourself, 287
 - visualizing BIC criterion, 290
 - visualizing integrated completed likelihood criterion, 290
- exploratory analysis, 97–98
 - application, 97
 - definition, 97
 - glossary, 103
- extensible markup language (XML), 47, 50–51, 52, 314, 366, 382
 - example of page, 50
- Facebook, 8, 31, 187, **355**
 - value of each user, 31
- Facebook Graph API, 47
- fact checker, 103, 320
- fairness, *xiii, xv, 32*
- Fairness, Accountability, and Transparency (FAccT), **32**
- fake news,
 - telltale signs, **324**
- false negative (FN), 320
- false positive (FP), 320
- false positive rate (FPR), 220, 321
 - glossary, 91
- family tree, 133
- feature (in ML), 188
 - glossary, 206
- feature creation (in ML), 188

feature matrix, 80	generative AI, 19
feature space selection, 64	generative pre-trained transformer 3 (GPT-3), 19
glossary, 68	geopandas (package), 125
Fenves, S. J., 70	GitHub, 135, 375
FICO scores, 96	Glassdoor, 380
file system, 156, 157	glossaries, 36
financial data scientists, 6	cloud computing, 177–178
financial sector, 301	data analysis techniques, 102–103
data science domain, 6–7	data terms, 67–68
gradient descent, 198	ML, 206
Fisher, R., 263	Python for statistical analysis, 146
Fitbit, 9	supervised learning, 261–262
FiveThirtyEight, 13	unsupervised learning, 295–296
flake8 package, 92, 101	Gmail, 152
F-measure, 85	golf, 23, 82, 88, 210, 246, 297
focus groups, 309–312	Google, 355
procedure, 310–311	fairness, 32
pros and cons, 312	value of each user, 31
purpose, 310	Google BigQuery, 100, 379
“for” loop, 129, 130	Google Cloud Platform (GCP), xvi, 152–161
formulas, 360–361	creating new project, <i>153</i>
differential calculus, 360–361	creating virtual machine, <i>153</i>
further reading, 361	establishing connection to virtual instance using PuTTY, <i>156</i>
probability theory, 361	Hadoop, 155–157
Foster, T. A., 268	interface for notebooks, <i>158</i>
fraud, 33, 259	using Python, 157–161
frequency distribution, 81–84	using Python (try it yourself), 89
frequency tables, 233, 247	Google Colaboratory (Colab), 157–161
“of act and inflated,” 233	“always free,” 169
Friston, K., 48	creating Jupyter notebook, <i>159</i>
F-statistic, 143, 144	finding the app, <i>159</i>
functions, 129, 130	installing package, <i>127</i>
Python, 122–124	notebook style, 168
reusability of code, 76, 125	running Python code, <i>160</i>
try it yourself, 124	writing code, <i>160</i>
Galton, Sir Francis, 190	Google Docs, 307
Gartner, 96	Google Forms, 307, 308
Gates, B., 290	Google Sheets, 27, 50, 75, 82, 83
Gauss, C. F., 190	Gosset, W. S., 142
Gaussian mixture model (GMM), 274, 282, 287, 290	governments, 44, 46
Gaussian Naïve Bayes, 249, 250	gradient descent, 190, 195–203
Gelman, A., 13	finding best regression line, <i>202</i>
General Data Protection Regulation (GDPR), 316	glossary, 77, 206
generalizability, 322, 324	hands-on example, 199–203
generalization, 21	hands-on problems, 208
generative adversarial networks (GANs), 19	in industry (DS in practice), 198–199
	regression (clinical data), 332–334

- try it yourself, 203
- visualization of cost function, 202
- gradient descent algorithm, 99, 197, 199, 206, 208
 - regression lines, 201
- graphical processing units (GPUs), 14, 151, 154, **295**
- graphical user interface (GUI), 128, 367, 370, 371–373, 376
- graphics, 136
- grounded theory, 318
 - glossary, 326
- Gueaguen, N., 265
- Hadoop, 155–157, 165, 380, 383
 - advantage, 156
 - role, 156
 - “set of open-source programs,” 156
 - version (3.3.3), 163
- Hadoop modules, 156–157
 - distributed file system, 156
 - Hadoop Common, 157
 - MapReduce, 157
 - YARN, 157
- hands-on approach, xiii, 4, 276
- hands-on examples, xiii
 - agglomerative clustering, 271–275
 - analyzing data, 26–29
 - ANOVA, 143–144
 - computational thinking, 21–22
 - correlation, 93–94, 138
 - data pre-processing, 60–65
 - data structures, 118–119
 - decision tree, 238–239
 - divisive clustering, 276–278
 - expectation maximization, 279–280
 - exploring clinical data, 330–335
 - gradient descent, 199–203
 - histogram, 82
 - interquartile range, 88–89
 - kNN, 226–230
 - linear regression, 191–193
 - logistic regression, 214–220
 - MySQL, 376–377
 - naïve Bayes, 249–251
 - pie chart, 83
 - random forest, 242–244
 - Reddit, 338
 - regression, 99–100
 - reinforcement learning, 291–294
 - softmax regression, 221–225
 - support vector machine, 259
 - t-test (steps 1-3), 140–142
 - using Python with AWS, 172
 - using Python with Azure, 165–168
 - using Python with GCP, 157–161
 - Yelp, 348–354
 - YouTube, 342–344
- Hands-On Introduction to Data Science*
 - online appendices, 3
 - real-life expectations, xv
 - requirements and expectations, xi–xii
 - second edition (new elements), xvi
 - second edition (new feature), xiii
 - strengths and unique features, xiv–xvi
 - target readership, xi, xiv–xv
 - updates, xvi
 - use of book in teaching, xiv
 - website, xiii–xiv, xvi
- hands-on problems
 - cloud computing, 178–179
 - clustering, 296–297
 - data analysis techniques, 104–106
 - data pre-processing, 68–73
 - data science, 37–40
 - decision trees, 265–266
 - expectation maximization, 297–298
 - kNN, 264–265
 - logistic regressions, 262–263
 - ML, 207–208
 - naïve Bayes, 267–268
 - Python, 129–130
 - Python for statistical analysis, 147–149
 - random forests, 266–267
 - softmax regressions, 263–264
 - supervised learning, 262–268
 - support vector machine, 268
 - unsupervised learning, 296–298
- hands-on with solving data problems, 329–359
 - clinical data, 330–335
 - practice questions, 356–358
 - Reddit data (collection and analysis), 336–340
 - summary, **355–356**
 - Yelp reviews and ratings, 349–354
 - YouTube data (collection and analysis), 342–348
- hardware, 14, 50, 98, 151, 314

- Harris, J., 23, 24
Harvard Business Review, 23
 HDInsight. *See* Microsoft Azure HDInsight
 healthcare, **140**
 data science domain, 9–10
 gradient descent, **199**
 ML bias, 203
 heath and human services: data science jobs, 383–384
 height-IQ data, 45
 height-weight data, 26–29, 43, 44, **93**
 Hill, B., 322
 Hindi film industry, 40
 Hippocratic oath, 32
 histogram, 134
 action, 81
 hands-on example, 82
 productivity data, 83
 try it yourself, 83
 Hive, 97, 383
 Holmes, E., **354**
 Holmes, S., 3, 5
 Holtz, D., 24
 Honda Research Institute (HRI), California, **290**
 honesty, 308, 311, 312
 household income (USA): highly-skewed distribution, 86
 Huffington, A., 8
 human capabilities: augmentation and assistance, 19
 human intelligence task (HIT), **92**
 human resources, 34, 173
 hyperplanes, 235, 252, 253
 and their margins, 254
 definition, 252
 HyperText Markup Language (HTML), 43, 51
 hypotheses, **91**, 348
 definition, **91**, 139
 glossary, 146
 hypothesis function, 88, 197, 212, 214
 hypothesis testing, 137, 139, 145
 four-step process, 139
 glossary, 92
 in practice, **139–140**
 IBM, 149, **173**, **316**
 predictive analytics, 95
 ID3 (Quinlan) algorithm, 231, 232
 identity theft, 33, 355
 “if” statements, 81
 if-else formula, 121
 if-then expressions, 235, 236, 237
 IMDB, **40**, 208
 inclusion, 34
 incremental gradient descent, 198
 glossary, 103
 independent variable, 80, 224
 glossary, 102
 India, 61
 inference techniques, 57, 138
 infinite loop problem, **122**
 informatics, 10, 16
 information, **4**, 17, 36
 information gain (IG), 233
 decision tree (three-step process), 234
 glossary, 261
 mathematical measurement, 232
 information retrieval (IR), 319
 information schools (iSchools), 18
 information science
 definition, 36
 relationship with data science, 16–18
 users, 17–18
 Infrastructure as a Service (IaaS), 95
 integrated complete-data likelihood (ICL), 290
 integrated development environment (IDE), 112–114, 363
 glossary, 103
 interactions
 Python, 124
 try it yourself, 124
 Interactive Python (IPython) Notebook: now known as “Jupyter” (*qv*), 363
 intercept, 99, 192, 195, 203, 208, 224
 “interesting” insights, 97
 internet, 8, 175, 288
 internet connection, xii, 151
 internet of things (IoT), 47, 289
 interquartile range
 definition, 88
 dispersion of distribution, 88–89
 hands-on example, 88–89
 try it yourself, 89
 interval variable, 80
 glossary, 102

- interviews, 309–312
 - data analysis, 311
 - procedure, 310–311
 - pros and cons, 312
 - purpose, 309–310
- iris dataset, 149, 226, 228, 229
- Jackson (Michigan), 11
- Jarausch, K. H., 267
- Java, 29, 128, 157, 380
- JavaScript Object Notation (JSON), 53–54, 339, 378
 - definition, 53
 - receiving data, 53
 - sending data, 53
 - structures, 53
- job labels and skills: data science in practice, **12**
- Jupyter, 113, 117, 363, 364
 - console, 344
 - installing package, 126
 - previously known as ‘IPython’, 80
- Jupyter environment, 158, 166
- Jupyter notebook, 115, 157
 - created in Google Colab, 159
- k* nearest neighbor (kNN), 225–230, 260, 335
 - anomaly detection, **260**
 - hands-on examples, 226–230
 - hands-on problems, 264–265
 - plotting of iris data, 227
- Kasik, D., 76
- Katz, R., 208
- kernels, 92, 270, 298
- key pair, 169
- Klein, J., 11
- k*-means algorithm, 275–279, 353
 - A against B plotted in D₂ graph, 277
 - centroids, **102**, **277–278**
 - dataset, **277**
 - first step-through, **102**
 - initialization of two clusters, **277**
 - second step-through, **278**
- k*-means clustering
 - basic step, 276
 - clinical data, 335
 - convergence (three-step process), 276
 - purpose, 276
- Knowledge Discovery in Data (KDD), **188**
- kurtosis, 84, 84
- labels, 44, 45, 97, 188, 206, 212, 229, 230, 239, 251, 270, 353
- large language models (LLMs), 235
- leaf nodes, 81, 235
- learning, 184, 187
 - operational definition, 185
- learning rate, 107, 198, 200, 333
- legal sector: data science jobs, 381–382
- Legendre, A-M., **190**
- Lending Club, 7
- leptokurtic distribution, 84
- libraries, 81, 136, 175
 - data science domain, 12
 - glossaries, 128
- likelihood, 85, 95
- likelihood function, 98, 213, 282
- likelihood tables, 247
- Likert scales, 73, 318
 - glossary, 102
- linear model, 86, 183, 193, 199, 208, 333, 336
 - glossary, 206
- linear regression, 98, 200, 353
 - algorithms, 204
 - annual return versus excess return of stock, 191
 - error surface, error value, 196
 - glossary, 206
 - hands-on example, 191–193
 - hands-on problems, 207
 - independent variable (predictor) versus dependent variable (response), 80
 - predicted outcomes, 257
 - regression line plotted on scatterplots of *x* versus *y*, 194
 - scatterplot of regression.csv data, 193
 - SVM, 256
 - try it yourself, 195
- LinkedIn, 5, **312**, 380
- Linux, 76, 164, 173, 366, 380
- lists, 119, 129
 - data structure (Python), 117
- Lo, F., 5
- log and diary data, 312–313
 - generic activity log template, **313**
 - generic diary template, **313**
- log likelihood, 84, 279, 280, 281

- log likelihood (cont.)
 - glossary, 103
- logical operators, 115, 116
- logistic regression, 212–220, 260
 - hands-on problems, 262–263
 - in industry (data science in practice), 221
 - receiver operating curve, 220
 - sample of *Titanic* data, 215
 - try it yourself, 220
 - visualizing missing values, 217
- logistics, 10, 91, 104, 263
 - gradient descent, **199**
- longley.csv file, 136
- loops
 - control structures, 120–121
 - debugging, **122**
- Lowe, M., 355
- m* and *b* values, 83, 195
 - partial derivatives, 196
- Mac, 107, 112, 344, 366
- machine learning (ML), xi–xii, xiv, 19, 36, 109, 380
 - algorithms, 14
 - anomaly detection, **260**
 - Azure, *166*
 - bias and fairness, **287–289**
 - bias, ethics, healthcare, 203
 - data mining, 188–189
 - decision tree, 230–239
 - definition, 184
 - ensemble method, 240
 - fundamental issues, 184–188
 - further reading and resources, 83
 - glossary, 206
 - gradient descent, 195–203
 - hands-on problems, 207–208
 - key terms, 206
 - mathematics, **197**
 - online appendices, 183
 - regression, 189–195
 - summary, 205
- machine learning for data science, 270–298
 - introduction and regression, 183–207
 - supervised learning, 210–268
 - unsupervised learning, 270–298
- machine learning techniques (application), 203–205
 - accuracy, 204
 - choosing right estimator, 204–205
 - features, 204–205
 - linearity, 204
 - parameters, 204–205
 - training time, 90
- Madoff, B., **354**
- manufacturing, 15, **92**
- MapReduce, 102, 157, 165
- marketing: data science jobs, 381
- mathematical reasoning, 23
- mathematics, 23, 25, 76, 181, 230
 - data science and ML, **197**
- matplotlib, 98, 134, 136, 168, 191, 226, 238
- maximum likelihood estimation (MLE), 280, 287
 - glossary, 98
- maximum marginal hyperplane (MMH), 253, 254
- mean, 105, 133
 - comparison using t-test, 140
 - glossary, 102
 - measure of centrality, 85–87
- mean absolute test set error, 324
- measures of centrality, 85–87
 - mean, 85–87
 - median, 87
 - mode, 87
- Mechanical Turk (MTurk), **312**
- mechanistic analysis, 98–99
 - glossary, 103
 - regression, 98–99
- median, 105, 134
 - glossary, 102
 - measure of centrality, 87
- medical imaging, 48, 199
- metadata, 12, 47, 57
- Microsoft Azure, xvi, 161–169, 204
 - configuration menu, *167*
 - ML platform “never free,” 169
 - ML services, *166*
 - ML studio, *166*
 - portal (interface), *162*
 - service navigation, *165*
 - use of Python, 165–168
 - using Python (try it yourself), 94
 - working with ~, *167–168*
 - working with Python notebook, *167*
- Microsoft Azure HDInsight, *163*
 - cluster, 161, 165
 - cluster details (overview), *164*
 - configuration of storage options, *165*

- Microsoft Azure Migrate, 176
- Microsoft Azure Site Recovery (ASR), 95
- Microsoft Excel, 27, 50, 75, 82
- “Microsoft Office 365,” 307
- misinformation, 341
 - definition, **324**
- missing data, 56–57, 70, 71, 73
 - data cleaning, 61
- missing values, 214
 - visualization (logistic regressions), 217
- Mitchell, T., 184, 185
- mixed method studies, 318–319
- ML. *See* machine learning
- mode
 - definition, 87
 - glossary, 102
 - measure of centrality, 87
 - visualization, 87
- model, 184
 - ML glossary, 206
- Modigliani, F., 296
- modulus operator, 120
- MongoDB, 378
- monitoring, 9, 10, 15, 32, 47, 187, 259, 316
- monolithic models, 19
- Mosteller, F., 264
- multimedia data, 43, 48
- multimodal data, 47–48
- multinomial logistic regression, 221, 223
- multiple linear regression, xv, 98
- multiple variables, 80, 360
- MySQL, xiii, 67
 - connecting to server, 368
 - databases, 25
 - “most popular open-source database platform,” 366
 - Python, 366–377
- MySQL (accessed with Python), 375–377
 - hands-on example, 376–377
 - try it yourself, 377
- MySQL (creating and inserting records), 369–371
 - creating table, 370–371
 - importing data, 369–370
 - inserting records, 371
- MySQL (getting started), 366–369
 - logging in, 367–369
 - obtaining MySQL, 366–367
 - using command line, 367
 - using GUI client, 368–369
- MySQL (retrieving records), 371–373
 - reading details about tables, 92
 - retrieving information from tables, 372–373
 - running SQL query, 373
 - try it yourself, 373
- MySQL (searching), 374–375
 - full-text searching with indexing, 374–375
 - “like” and “match” approaches, **375**
 - searching within field values, 374
 - try it yourself, 375
- MySQL Workbench, 367, 368, 376
- naïve Bayes, 245–251, 260
 - frequency and likelihood tables, 247
 - hands-on example, 249–251
 - hands-on problems, 267–268
 - in practice, 251
 - independence assumption, 248, 261
 - try it yourself, 251
 - weather dataset, **246**
- NCAA March Madness tournament, 81
- Netflix, 103, **198**
- New York City, 7, **101**, **137**, 305, **354**
- New York Times*, 8, **54**
- Newton’s second law, 124
- NIST, 268
- noisy data, 70, 71
 - glossary, 68
- nominal data: glossary, 68
- nominal variable, 80
 - glossary, 102
- non-governmental organizations (NGOs), 8, 46
- non-linear (curve fitting) regression, 194
- normal distribution, 84, 86
 - example, 84
 - glossary, 102
- not only SQL (NoSQL), 378
- null hypothesis, 76, 80, 91, 139, 143
 - definition, **91**
 - glossary, 102
- numbers, 21–22
- numerical representation: advantage over words, 78
- NumPy (Numerical Python), 133, 134, 159, 228
- OA. *See* online appendices
- Obama, B. H., 8

- Olteanu, A., **355**
 Olympics, 176, 313
 online appendices, xiii, 3, 68, 70, 72, 75, 83, 89,
 91, 94, 100, 104, 105, 133, 136, 148, 207,
 214, 220, 221, 225
 abalone dataset, 266
 AnthroKids dataset, 91
 automated answer-rating CSV file, 263
 balloon datasets, 231, 238, 245
 bank marketing dataset, 242
 blues guitarists dataset, 267
 clinical data, **331**
 combined cycle dataset, 259
 concrete slump dataset, 297
 contact lenses dataset, 239, 251
 crash.csv dataset, 262
 data processing, 48
 datasets, 43
 golf dataset, 246, 249
 gradient descent, 203
 hands-on with solving data problems, 75
 hate-crime dataset, 358
 horseshoe crab dataset, 263
 hsbdemo dataset, 230
 immunotherapy dataset, 263
 iris dataset, 226, 263
 life-cycle savings dataset, 296
 LPGA dataset, 297
 mysqlsampledatabase.sql, 376
 Nazi dataset, 267
 NFL 2014 combine performance results
 dataset, 264
 Portuguese sea battles dataset, 268
 problematic behaviors of students dataset,
 265
 Project 16P5 dataset, 264
 regression.csv, 191, 199, 255
 SGEMM GPU kernel dataset, 298
 spam collection dataset, 267
 StoneFlakes dataset, 274
 supervised learning, 210–211
 techniques, 75
 test.csv data, 262
 train.csv data, 262
 unsupervised learning, 75
 user knowledge modeling dataset, 279, 287
 waiter jokes (effect on tipping) dataset, 265
 weather.csv, 264
 youtube_search.py script, 346
 online dataset (OD), xiii
 open data, 44, 46–47
 glossary, 68
 multimodal data, 47–48
 principles, 46–47
 social media data, 47
 synthetic data, 48–49
 Open Data Policy (USA), 46
 optical character recognition (OCR), 185
 optical character recognition (PCR), 186
 optimal hyperplane, 254, 255
 Oracle, 173, 382, 383
 ordinal data, 102
 glossary, 68
 ordinal variable, 80
 glossary, 102
 ordinary least squares regression (OLSR), 190
 Orwell, G., **355**
 outcome variables, 80, 212, 246
 glossary, 102, 206
 outlier detection, **259**
 machine learning, **260**
 outliers, 70, 86, 88
 glossary, 68
 out-of-bag error, 241
 overfitting. *See* data overfitting
 packages
 definition, 125
 glossary, 128
 pandas, 75, 159, 375
 Python package, 125
 pandas library, 118, 119, 136, 230
 parameter estimation process, 90
 parameter: glossary, 206
 partial derivatives, 214, 360
 Patil, D. J., 5
 Pearson's correlation, 92, 93, 99, 100, 106, 339,
 340
 role, 138
 petabytes (PB), 4
 pie chart: hands-on example, 83
 “pip” command, 125
 Pixelfed, **251**
 Platform as a Service (PaaS), 152, 379
 platykurtic distribution, 84, 86
 politics
 data science domain, 8–9
 definition, 8

- posterior probability, 246, 247, 248, 282
- poultry business, 37, 39
- precision
 - calculation, 320
 - glossary, 326
- predicted outcomes
 - linear regression, 257
 - SVM, 259
- predictive analytics, 95–96
 - applications, 96
 - glossary, 103
 - process, 95
 - stages, 95
- predictive modeling, **77**
- predictor attributes, 76, 87, 207, 279
- predictor variables, 80, 98, 207, **221**, 221, 240, 264, 266
 - glossary, 78, 102
- prescriptive analytics, 96–97
 - definition, 96
 - glossary, 103
- Priceonomics, 13
- prior probability, 246, 250
- privacy, xiii, xv, 8, 31, 32, 33, 34, 35, 354–355
- probabilistic models, 279
- probability, xiii, 141
- probability theory
 - conditional probability, 361
 - formulas, 361
 - independent events, 361
 - mutually exclusive events, 361
- probability value. *See* *p*-value
- programming languages, 29, 76, 115, 116, 117, 125, 127, 128
- programming tools, 29, 36, 76, 128
- Project Open Data (USA, 2013), 46
- Prolific, **312**
- public policy, 44
 - data science domain, 7–8
- PuTTY, 165, 171
 - connecting AWS EC2 instance from PC, *171*
 - establishing connection to virtual instance in GCP, *156*
 - generation of SSH key, *154*
- p*-value, 90, 91, 141, 144, 219, 340
- PyDev, 113
- Python, 29, 80, 109, 184, 347, 357
 - at job, **127**
 - basic examples, 115–117
 - basic operations (try it yourself), 117
 - control structures, 120–121
 - data structures, 117–119
 - decision tree-based classifier (implementation), 90
 - definition, 111
 - downloading and installation, 112–125
 - expectation maximization, 279–280
 - functions, 122–124
 - further reading and resources, 83
 - getting access, 111–114
 - hands-on problems, 129–130
 - interactivity, 124
 - MySQL, 366–377
 - naïve Bayes classification, 249
 - number one for programming languages, 127
 - origins, **114**
 - random forest, 91
 - reinforcement learning, 291–294
 - run through console, 112
 - running ~ code in Google Colab, *160*
 - “scripting language,” 29
 - summary, 127–128
 - SVM classification problem, 255
 - tool for data science, 111–131
 - use with Azure, 165–168
 - used through IDE, 112–114
 - used with GCP, 157–161
- Python console, 115, 124
- Python for statistical analysis, 132–150
 - comparing means using t-test, 140
 - data (importing), 136
 - data (plotting), 136
 - further reading and resources, 149–150
 - graphics and data visualization, 136
 - hands-on problems, 147–149
 - hypothesis testing, 139
 - key terms, 146
 - statistical inference, 137–144
 - summary, 145
- Python libraries, 125, 199, 228
- Python packages, 133, 363
 - installation and use, 125–126
- Python programming, xii, 30, 113, 114, 365
- qualitative data, 76, 78, 326
- qualitative methods, 316, 317–318
 - glossary, 326
 - weaknesses, 318

- Qualtrics, 308
- quantitative analysis, 17, 317
- quantitative data, 78, 318, 319
- quantitative methods, 326
 - weaknesses, 318
- Quinlan, J. R., 231
- R, xiii, xvi, 29, 30, 82, 183, 184, 219, 336, 338, 342, 357, 380, 381, 383
- random forest, 239–245
 - advantages, 240, 245
 - algorithm, 240
 - bar plot depicting data points, 243
 - clinical data classification, 334–335
 - “does better job than individual decision trees,” 242
 - hands-on example, 242–244
 - hands-on problems, 266–267
 - mode of operation, 240
 - “not silver bullet,” 245
 - “panacea for all DS problems,” 244
 - try it yourself, 245
 - “uses large set of decision trees,” 260
- random numbers, 134, 135, 228
- range: definition, 87
- RapidMiner, 95
- ratio variable, 80
 - glossary, 102
- really simple syndication (RSS), 51–53
- recall, 320–321
 - calculation, 320
- receiver operating characteristic (ROC), 321
 - comparing models, 321
- receiver operating curve (ROC), 88, 220, 220
- Reddit, xv–xvi, **312**, 356
 - data collection and analysis, 336–340
 - hands-on example, 338
 - try it yourself, 340–354
- Reddit APIs, 336–340
 - interface for signing up, 337
 - interface showing client ID and client secret, 337
 - step 1 (signing up), 336
 - step 2 (create Reddit app and get API credentials), 86
 - step 3 (install required tools), 338
- regression, 78, 98–99, 189–195
 - glossary, 103
 - gradient descent (clinical data), 332–334
 - hands-on example, 99–100
 - lines produced using gradient descent algorithm, 201
 - try it yourself, 100
- regression algorithms, 190
- regression analysis, 91, 138
 - definition, 98
- regression history, **101**
- regression.csv file, 256
- Reich, Y., 70
- reinforcement learning (RL), 196, 289–294
 - glossary, 296
 - hands-on example, 291–294
 - model, 289
 - robotics, **291**
- relational database management system (RDBMS), 378
- relative risk, 224, 225
 - glossary, 261
- replicated subtree problem, 236
- Respondent, **314**
- response variable, 148
 - glossary, 102, 206
- retail and sales: data science jobs, 381
- robotics, 99, **291**
- Rooney rule, 289
- root mean square error (RMSE)
 - calculation in Python, 257
 - linear regression model, 258
 - SVR model, 258
- root nodes, 235, 236
- Rossum, G. van, **114**
- RStudio, **172**, 214
- Rutherford, E., 303
- Samuel, A., 184
- San Francisco, 13, 24
- Sanders, H., 37
- scatterplots, 136, 357
 - regression.csv, 193
 - x versus y (regression line), 194
- Scholastic Aptitude Test (SAT), 78
- science: definition, **4**, 36
- scipy, 141, 275, 339
- seaborn, 193, 216, 280, 349
- search engine result pages (SERP), 17
- secure shell (SSH), 112, 154, 165, 171, 367, 369
 - connected to AWS EC2 instance, 172

- glossary, 177
- tunneling approach, 93
- secure shell key
 - added to virtual machine, 155
 - generation by PuTTY, 154
- self-driving car, 185
 - ML technology, 186
- sentiment analysis, **251**
- separating hyperplane, 252, 253, 254
- sets, 119
 - data structure (Python), 118
- sigmoid function, 212, 213, 214, 221
 - formula, 85
- significance level, 139, 142
- Silver, N., 13
- simple linear regression, 98
- skewed distribution, 84
 - negative versus positive, 85
- skills for data science, 23–24
 - data literacy, 23–24
 - mathematical reasoning, 23
 - willingness to experiment, 23
- sklearn, 102, 126, 199, 226, 229, 238, 243, 249, 274, 280
- smartphones, 312
 - data plan, 4
- SmartSurvey, 307
- smoking, 17, 322
- Snap Surveys Ltd, 304
- Snow Corps, 11
- social good, xiii, 8, 301
- social media, 208, 301, 307, 308, **312**, 324, **355**
 - data analysis, 341–342
 - marketing campaigns, 92
 - Reddit, 336–340
 - YouTube, 342–348
- social media data, 47
- social science
 - problem-analysis types (excellent, good, weak), 101
 - relation to data science, 15–16
- societal issues (social media data analysis), 341–342
 - economic issues, 341
 - further reading, 342
 - health issues, 341
 - social issues, 341
 - socio-political issues, 341
- softmax regression, 221–225, 260
 - hands-on example, 221–225
 - hands-on problems, 263–264
 - try it yourself, 225
- software development, **127**, **173**
 - precision (critical importance), **116**
- software development kit (SDK): glossary, 99
- software engineering, 23, 25
- Sondergaard, P., 75
- spam detection, **78**
- Spearman's rank correlation, 138
- Spielkamp, M., **101**
- spreadsheet programs, 27, 50, **75**, 82, 83, 308, 371
- spreadsheets, 49, 50, 67
 - "structured data," 46
- spurious correlation, **94**
 - glossary, 103
- Spyder, 113, 117, 363
 - console, 115, 136, 344
 - IDE (screenshot), 114
 - installation through Anaconda Navigator, 113
 - installing package, 125
 - launch, 113
 - snapshot, 365
- Spyder Run Configuration, 347
- square roots, 90
 - "sqrt" function, 122
- SSH. *See* Secure Shell
- standard deviation, 99, 133
 - computation, 90
 - dispersion of distribution, 90–91
 - explanation, 90
 - formula, 91
 - try it yourself, 91
- StandardScaler, 274, 282, 289
- Stanford Medicine, 9
- Stanton, J. M., 12
- statistical analyses, 81, 109, 132, 139, 318
- statistical bias. *See also* bias
 - glossary, 103
- statistical inference, 137–144
 - glossary, 102
- statistical parametric mapping (SPM), 48
- statistical significance, 87, 140
- statistical techniques, 29, 36, 48, 77, 183, 184
- statistical tests, **91**, 348

- statistics, xi–xii, 1, 76
 - bar graph (age distribution), 134
 - bar graph (thousand random numbers), 135
 - essentials, 133–135
 - relation to data science, 13–14
 - try it yourself, 134, 135
- statsmodels, 144, 192, 223
- stereotyping: definition, 193
- Sterling, A., 296
- stochastic gradient descent, 198
 - glossary, 103
- stop words, 375
- storage options: configuration in HDInsight, 165
- structured data, 44–45
 - definition, 44
 - glossary, 67
- structured query language (SQL), 29, 30, 67, 81, 157, 380, 383
- student grades: data analysis versus data analytics, 76
- Student's t-test. *See* t-test
- subreddit, 97, 339, 357
- subsets, 231, 232, 240, 271, 295, 324. *See also* clustering
- sum of squares (SS), 335
- supervised learning, 103, 196, 210–268
 - algorithms (role), 211
 - classification with kNN, 225–230
 - decision tree, 230–239
 - definition, 261
 - further reading and resources, 83
 - hands-on examples, 214–220
 - hands-on problems, 262–268
 - key terms, 261–262
 - logistic regression, 212–220
 - naïve Bayes, 245–251
 - online appendices, 210–211
 - random forest, 239–245
 - softmax regression, 221–225
 - summary, 260–261
 - support vector machine, 252–259
 - types, 211
- support vector machine (SVM), 252–259
 - anomaly detection, 260
 - “best stock classifier for ML tasks,” 252
 - classification problem, 255
 - from line to hyperplane, 253
 - hands-on example, 259
 - hands-on problems, 268
 - hyperplanes and their margins, 254
 - linearly separable data, 253
 - power and sophistication, 261
 - predicted outcomes, 257, 259
 - regression line fitted onto data, 257
 - regression.csv file, 256
 - try it yourself, 259
 - two-class problem, 252
- support vector regression (SVR), 258
- support vectors, 252, 254, 255
- survey question types, 304–306
 - dichotomous (closed), 306
 - Likert scales, 305
 - multiple-choice, 100
 - open-ended, 306
 - rank-order, 305
 - rating or open-ended, 305
- SurveyMonkey, 308
- surveys, 304–309
 - audience, 306–307
 - data analysis, 308
 - Likert-type question, 309
 - methods, 304
 - pros and cons, 308–309
 - results organized in Excel, 103
- synthetic data, 48–49
- system bias, 288
- Tableau, 25, 383
- tab-separated values (TSV), 49, 50, 132
- target attribute, 85, 135
- taxonomy, 80, 95
- techniques
 - data analysis, 75–106
 - hard to separate from ‘tools’, 75
- temperature, 80, 106, 129
- test statistic, 139, 141, 146
- testing datasets, 191, 213, 229, 322–323, 334
 - glossary, 261
- TestMate, 314
- Therac-25 machine, 122
- tools for data science, xii, 151–179
 - cloud computing, 151–179
 - hard to separate from “techniques,” 75
 - Python, 111–131
 - Python for statistical analysis, 132–150
- Torvalds, L., 111, 151

- Towards Data Science blog, **295**
- training datasets, 78, 229, 250, 252, 322–323, 334
 - glossary, 261
 - sampling, *241*
 - training instances and independent variables, **241**
- training tuples, 252
- true positive (TP), 320
- true positive rate (TPR), 220, 321
 - glossary, 91
- Trump, D., 8, 13
- try it yourself, xiii
 - analyzing data, 29
 - ANOVA, 144
 - basic operations, 117
 - basic statistics, 134, 135
 - clinical data, 354
 - clustering, 279
 - computational thinking, 22
 - control structures, 121
 - correlation, 94, 138
 - data pre-processing, 66–67
 - data structures, 119
 - decision tree, 83
 - distributions, 84
 - expectation maximization, 287
 - functions, 124
 - gradient descent, 203
 - installing package, 92
 - interactions, 124
 - interquartile range, 89
 - linear regression, 195
 - logistic regression, 220
 - MySQL (accessed with Python), 377
 - MySQL (searching), 375
 - naïve Bayes, 251
 - random forest, 245
 - Reddit, 340–354
 - regression, 100
 - retrieval, 373
 - softmax regression, 225
 - standard deviation, 91
 - support vector machine, 259
 - t-test, 142
 - using Python with AWS, 91
 - using Python with Azure, 94
 - using Python with GCP, 89
 - variables, 82
 - Yelp, 354
 - YouTube, 346–348
- TryMyUI, **314**
- t-test. *See also* analysis of variance
 - comparison of means, 140
 - definition, 140
 - hands-on example (steps 1-3), 140–142
 - real name of “Student,” **142**
 - try it yourself, 142
 - two-sample $\sim$, 140
- Tukey, J. W., 264
- tumor size, 81
- tuples, 119
 - data structure (Python), 118
- t-value, 139, 142
- Twitter, 31, 308, **324**, 357
 - now X, 8
- two-class classification, 225, 260
- UCI machine learning repository, 135
- United States, **316**
 - presidential election (2008), 13
 - presidential election (2016), 8, 13, 358
 - US government, 7, 46, 73
- University of Chicago, 11
- UNIX, 111, 154, 177, 344, 366
- unstructured data, 45–46
 - challenges, 45–46
 - definition, 44
 - glossary, 67
- unsupervised learning, 103, 196, 270–298
 - agglomerative clustering, 271–275
 - deep learning, 294–295
 - definition, 295
 - divisive clustering, 275–279
 - expectation maximization, 279–287
 - further reading and resources, 83
 - hands-on problems, 296–298
 - key terms, 295–296
 - online appendices, 75
 - reinforcement learning, 289–294
 - summary, 294
- Urban Center for Computation and Data (UrbanCCD), 11
- urban planning, 18
 - data science domain, 10–11
- user interface (UI), **314**, 314
- user testing, 314, **314**

validation data: glossary, 261	women, 9, 54, 219, 297
vampire charges, 163	height-weight data (analysis), 26–29
variables, 78–81	height-weight data (visualization), 28
faulty conversion, 116	Worthy, S. L., 265
try it yourself, 82	
x (input) versus y (output), 191	YARN, 165
variables of interest, 189, 221	“yet another resource negotiator,” 157
variance, 105, 133	Yau, N., 14
dispersion of distribution, 90	Yeh, I-C., 297
formulas, 90	Yelp, xv, 47, 189 , 357
vectors, 252. <i>See also</i> support vectors	business data (different states), 350
multiple feature versus single feature, 94	clustering result for individuals according to
velocity, volume, variety (3V) model, 5, 355	fans counts, 354
financial data, 6	hands-on example, 348–354
virtual instance, 98, 153	reviews and ratings, 349–354
establishing connection in GCP using	scatterplot showing number of reviews, 352
PuTTY, 156	try it yourself, 354
virtual machine (VM), 152, 164	YouTube, xv, 357
addition of SSH key, 155	data (collection and analysis), 342–348
AWS (creation), 170	hands-on example, 342–344
AWS (launch), 170	try it yourself, 346–348
connection to virtual instance, 155	YouTube APIs, 342–346
creation on GCP, 153	creating credentials, 345
glossary, 177	creating project in Google console, 343
Visual Studio Dev Essentials program,	enabled but not yet available for project, 344
161	enabling, 344
Wayfair, 5	getting credentials (API key), 346
West, D. M., 11, 12	library available for Google projects, 343
“while” loop, 105, 121, 122 , 129, 130	step 1 (signing up), 342
wildcards, 374	step 2 (selecting API and obtaining API
wine consumption and mortality: data pre-	key), 342
processing, 60–65	step 3 (installation of packages), 344
Wing, J., 20	zettabytes (ZB), 5