Chapter 14

Bibliography

[Abe13] Andrew Abela. *Advanced Presentations by Design: Creating Communication that Drives Action.* Pfeiffer, 2nd edition, 2013.

[Ans73] Francis J Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973.

[Bab11] Charles Babbage. *Passages from the Life of a Philosopher.* Cambridge University Press, 2011.

[Bar10] James Barron. Apple's new device looks like a winner. From 1988. *The New York Times*, January 28, 2010.

[Ben12] Edward A Bender. *An Introduction to Mathematical Modeling.* Courier Corporation, 2012.

[Bis07] Christopher Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics.* Springer, New York, 2007.

[BK07] Robert M. Bell and Yehuda Koren. Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

[BK16] Benjamin Bengfort and Jenny Kim. *Data Analytics with Hadoop: An Introduction for Data Scientists.* O'Reilly Media, Inc., 2016.

[BP98] Serge Brin and Larry Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. 7th Int. Conf. on World Wide Web (WWW)*, pages 107–117, 1998.

[Bra99] Ronald Bracewell. *The Fourier Transform and its Applications.* McGraw-Hill, 3rd edition, 1999.

[Bri88] E. Oran Brigham. *The Fast Fourier Transform.* Prentice Hall, Englewood Cliffs NJ, facimile edition, 1988.

[BSC$^+$11] M. Borjesson, L. Serratosa, F. Carre, D. Corrado, J. Drezner, D. Dugmore, H. Heidbuchel, K. Mellwig, N. Panhuyzen-Goedkoop, M. Papadakis, H. Rasmusen, S. Sharma, E. Solberg, F. van Buuren, and A. Pelliccia. Consensus document regarding cardiovascular safety at sports arenas. *European Heart Journal*, 32:2119–2124, 2011.

[BT08] Dimitri Bertsekas and John Tsitsklis. *Introduction to Probability.* Athena Scientific, 2nd edition, 2008.

© The Author(s) 2017

S.S. Skiena, *The Data Science Design Manual,*
Texts in Computer Science, https://doi.org/10.1007/978-3-319-55444-0

[BVS08] Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. International Sentiment Analysis for News and Blogs. In *Proceedings of the International Conference on Weblogs and Social Media*, Seattle, WA, April 2008.

[BWPS10] Mikhail Bautin, Charles B Ward, Akshay Patil, and Steven S Skiena. Access: news and blog analysis for the social sciences. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1229–1232. ACM, 2010.

[CPS$^+$08] J Robert Coleman, Dimitris Papamichail, Steven Skiena, Bruce Futcher, Eckard Wimmer, and Steffen Mueller. Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320(5884):1784–1787, 2008.

[CPS15] Yanqing Chen, Bryan Perozzi, and Steven Skiena. Vector-based similarity measurements for historical figures. In *International Conference on Similarity Search and Applications*, pages 179–190. Springer, 2015.

[dBvKOS00] Mark de Berg, Mark van Kreveld, Mark Overmars, and Otfried Schwarzkopf. *Computational Geometry: Algorithms and Applications.* Springer, 2nd edition, 2000.

[DDKN11] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning future media environments*, pages 9–15. ACM, 2011.

[Don15] David Donoho. 50 years of data science. Tukey Centennial Workshop, Princeton NJ, 2015.

[DP12] Thomas H Davenport and DJ Patil. Data scientist. *Harvard Business Review*, 90(5):70–76, 2012.

[EK10] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a highly connected world.* Cambridge University Press, 2010.

[ELLS11] Brian Everitt, Sabine Landau, Mmorven Leese, and Daniel Stahl. *Cluster Analysis.* Wiley, 5th edition, 2011.

[ELS93] Peter Eades, X. Lin, and William F. Smyth. A fast and effective heuristic for the feedback arc set problem. *Information Processing Letters*, 47:319–323, 1993.

[FCH$^+$14] Matthew Faulkner, Robert Clayton, Thomas Heaton, K Mani Chandy, Monica Kohler, Julian Bunn, Richard Guy, Annie Liu, Michael Olson, MingHei Cheng, et al. Community sense and response systems: Your phone as quake detector. *Communications of the ACM*, 57(7):66–75, 2014.

[Few09] Stephen Few. *Now You See It: simple visualization techniques for quantitative analysis.* Analytics Press, Oakland CA, 2009.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning.* Springer, 2001.

[FPP07] David Freedman, Robert Pisani, and Roger Purves. *Statistics.* WW Norton & Co, New York, 2007.

[Gay14] C. Gayomali. NYC taxi data blunder reveals which celebs don't tip and who frequents strip clubs. http://www.fastcompany.com/3036573/, October 2. 2014.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016.

[GFH13] Frank R. Giordano, William P. Fox, and Steven B. Horton. *A First Course in Mathematical Modeling.* Nelson Education, 2013.

[Gle96] James Gleick. A bug and a crash: sometimes a bug is more than a nuisance. *The New York Times Magazine*, December 1, 1996.

[GMP$^+$09] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[Gol16] David Goldenberg. The biggest dinosaur in history may never have existed. FiveThirtyEight, http://fivethirtyeight.com/features/the-biggest-dinosaur-in-history-may-never-have-existed/, January 11, 2016.

[Gru15] Joel Grus. *Data Science from Scratch: First principles with Python.* O'Reilly Media, Inc., 2015.

[GSS07] Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. *Int. Conf. Weblogs and Social Media*, 7:21, 2007.

[HC88] Diane F Halpern and Stanley Coren. Do right-handers live longer? *Nature*, 333:213, 1988.

[HC91] Diane F Halpern and Stanley Coren. Handedness and life span. *N Engl J Med*, 324(14):998–998, 1991.

[HS10] Yancheng Hong and Steven Skiena. The wisdom of bookies? sentiment analysis vs. the NFL point spread. In *Int. Conf. on Weblogs and Social Media*, 2010.

[Huf10] Darrell Huff. *How to Lie with Statistics.* WW Norton & Company, 2010.

[Ind04] Piotr Indyk. Nearest neighbors in high-dimensional spaces. In J. Goodman and J. O'Rourke, editors, *Handbook of Discrete and Computational Geometry*, pages 877–892. CRC Press, 2004.

[Jam10] Bill James. *The New Bill James Historical Baseball Abstract.* Simon and Schuster, 2010.

[JLSI99] Vic Jennings, Bill Lloyd-Smith, and Duncan Ironmonger. Household size and the Poisson distribution. *J. Australian Population Association*, 16:65–84, 1999.

[Joa02] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142. ACM, 2002.

[Joh07] Steven Johnson. *The Ghost Map: The story of London's most terrifying epidemic – and how it changed science, cities, and the modern world.* Riverhead Books, 2007.

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning.* Springer-Verlag, sixth edition, 2013.

[Kap12] Karl M Kapp. *The Gamification of Learning and Instruction: Game-based methods and strategies for training and education.* Wiley, 2012.

[KARPS15] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. ACM, 2015.

[KCS08] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456. ACM, 2008.

[KKK$^+$10] Slava Kisilevich, Milos Krstajic, Daniel Keim, Natalia Andrienko, and Gennady Andrienko. Event-based analysis of people's activities and behavior using flickr and panoramio geotagged photo collections. In *2010 14th International Conference Information Visualisation*, pages 289–296. IEEE, 2010.

[Kle13] Phillip Klein. *Coding the Matrix: Linear Algebra through Computer Science Applications.* Newtonian Press, 2013.

[KSG13] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proc. National Academy of Sciences*, 110(15):5802–5805, 2013.

[KTDS17] Vivek Kulkarni, Yingtao Tian, Parth Dandiwala, and Steven Skiena. Dating documents: A domain independent approach to predict year of authorship. Submitted for publication, 2017.

[Lei07] David J. Leinweber. Stupid data miner tricks: overfitting the S&P 500. *The Journal of Investing*, 16(1):15–22, 2007.

[Lew04] Michael Lewis. *Moneyball: The art of winning an unfair game.* WW Norton & Company, 2004.

[LG14] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014.

[LKKV14] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[LKS05] Levon Lloyd, Dimitrios Kechagias, and Steven Skiena. Lydia: A system for large-scale news analysis. In *SPIRE*, pages 161–166, 2005.

[LLM15] David Lay, Steven Lay, and Judi McDonald. *Linear Algebra and its Applications.* Pearson, 5th edition, 2015.

[LM12] Amy Langville and Carl Meyer. *Who's #1? The Science of Rating and Ranking.* Princeton Univ. Press, 2012.

[LRU14] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of Massive Datasets.* Cambridge University Press, 2014.

[Mal99] Burton Gordon Malkiel. *A Random Walk Down Wall Street: Including a life-cycle guide to personal investing.* WW Norton & Company, 1999.

[MAV$^+$11] J. Michel, Y. Shen A. Aiden, A. Veres, M. Gray, Google Books Team, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. Nowak, and E. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331:176–182, 2011.

[MBL$^+$06] Andrew Mehler, Yunfan Bao, Xin Li, Yue Wang, and Steven Skiena. Spatial Analysis of News Sources. In *IEEE Trans. Vis. Comput. Graph.*, volume 12, pages 765–772, 2006.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[McK12] Wes McKinney. *Python for Data Analysis: Data wrangling with Pandas, NumPy, and IPython.* O'Reilly Media, Inc., 2012.

[McM04] Chris McManus. *Right Hand, Left Hand: The origins of asymmetry in brains, bodies, atoms and cultures.* Harvard University Press, 2004.

[MCP$^+$10] Steffen Mueller, J Robert Coleman, Dimitris Papamichail, Charles B Ward, Anjaruwee Nimnual, Bruce Futcher, Steven Skiena, and Eckard Wimmer. Live attenuated influenza virus vaccines by computer-aided rational design. *Nature Biotechnology*, 28(7):723–726, 2010.

[MOR$^+$88] Bartlett W Mel, Stephen M Omohundro, Arch D Robison, Steven S Skiena, Kurt H. Thearling, Luke T. Young, and Stephen Wolfram. Tablet: personal computer of the year 2000. *Communications of the ACM*, 31(6):638–648, 1988.

[NYC15] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 427–436. IEEE, 2015.

[O'N16] Cathy O'Neil. *Weapons of Math Destruction: How big data increases inequality and threatens democracy.* Crown Publishing Group, 2016.

[O'R01] Joseph O'Rourke. *Computational Geometry in C.* Cambridge University Press, New York, 2nd edition, 2001.

[Pad15] Sydney Padua. *The Thrilling Adventures of Lovelace and Babbage: The (mostly) true story of the first computer.* Penguin, 2015.

[PaRS14] Bryan Perozzi, Rami al Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710. ACM, 2014.

[PFTV07] William Press, Brian Flannery, Saul Teukolsky, and William T. Vetterling. *Numerical Recipes: The art of scientific computing.* Cambridge University Press, 3rd edition, 2007.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[RD01] Ed Reingold and Nachum Dershowitz. *Calendrical Calculations: The Millennium Edition.* Cambridge University Press, New York, 2001.

[RLOW15] Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills. *Advanced Analytics with Spark: Patterns for Learning from Data at Scale.* O'Reilly Media, Inc., 2015.

[Sam05] H. Samet. Multidimensional spatial data structures. In D. Mehta and S. Sahni, editors, *Handbook of Data Structures and Applications*, pages 16:1–16:29. Chapman and Hall/CRC, 2005.

[Sam06] Hanan Samet. *Foundations of Multidimensional and Metric Data Structures.* Morgan Kaufmann, 2006.

[SAMS97] George N Sazaklis, Esther M Arkin, Joseph SB Mitchell, and Steven S Skiena. Geometric decision trees for optical character recognition. In *Proceedings of the 13th Annual Symposium on Computational Geometry*, pages 394–396. ACM, 1997.

[SF12] Gail M. Sullivan and Richard Feinn. Using effect size: or why the p value is not enough. *J. Graduate Medical Education*, 4:279282, 2012.

[Sil12] Nate Silver. *The Signal and the Noise: Why so many predictions fail-but some don't.* Penguin, 2012.

[Ski01] S. Skiena. *Calculated Bets: Computers, Gambling, and Mathematical Modeling to Win.* Cambridge University Press, New York, 2001.

[Ski08] S. Skiena. *The Algorithm Design Manual.* Springer-Verlag, London, second edition, 2008.

[Ski12] Steven Skiena. Redesigning viral genomes. *Computer*, 45(3):0047–53, 2012.

[SMB$^+$99] Arthur G Stephenson, Daniel R Mulville, Frank H Bauer, Greg A Dukeman, Peter Norvig, Lia S LaPiana, Peter J Rutledge, David Folta, and Robert Sackheim. Mars climate orbiter mishap investigation board phase i report. *NASA, Washington, DC*, page 44, 1999.

[SRS$^+$14] Paolo Santi, Giovanni Resta, Michael Szell, Stanislav Sobolevsky, Steven H Strogatz, and Carlo Ratti. Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences*, 111(37):13290–13294, 2014.

[SS15] Oleksii Starov and Steven Skiena. GIS technology supports taxi tip prediction. Esri Map Book, 2014 User Conference, July 14-17, San Diego, 2015.

[Str11] Gilbert Strang. *Introduction to Linear Algebra.* Wellesley-Cambridge Press, 2011.

[Sur05] James Surowiecki. *The wisdom of crowds.* Anchor, 2005.

[SW13] Steven S. Skiena and Charles B. Ward. *Who's Bigger?: Where Historical Figures Really Rank.* Cambridge University Press, 2013.

[Tij12] Henk Tijms. *Understanding Probability.* Cambridge University Press, 2012.

[Tuc88] Alan Tucker. *A Unified Introduction to Linear Algebra: Models, methods, and theory.* Macmillan, 1988.

[Tuf83] Edward R Tufte. *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, CT, 1983.

[Tuf90] Edward R Tufte. *Envisioning Information.* Graphics Press, Cheshire, CT, 1990.

[Tuf97] Edward R Tufte. *Visual Explanations.* Graphics Press, Cheshire, CT, 1997.

[VAMM$^+$08] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[Vig15] Tyler Vigen. *Spurious Correlations.* Hatchette Books, 2015.

[Wat16] Thayer Watkins. Arrow's impossibility theorem for aggregating individual preferences into social preferences. http://www.sjsu.edu/faculty/watkins/arrow.htm, 2016.

[Wea82] Warren Weaver. *Lady Luck.* Dover Publications, 1982.

[Wei05] Sanford Weisberg. *Applied linear regression*, volume 528. Wiley, 2005.

[Wes00] Doug West. *Introduction to Graph Theory.* Prentice-Hall, Englewood Cliffs NJ, second edition, 2000.

[Whe13] Charles Wheelan. *Naked Statistics: Stripping the dread from the data.* WW Norton & Company, 2013.

[ZS09] Wenbin Zhang and Steven Skiena. Improving movie gross prediction through news analysis. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 301–304. IEEE Computer Society, 2009.

[ZS10] Wenbin Zhang and Steven Skiena. Trading strategies to exploit blog and news sentiment. In *Proc. Int. Conf. Weblogs and Social Media (ICWSM)*, 2010.

Index

A/B testing, 86
Aaron Schwartz case, 68
AB testing, 137
academic data, 66
accuracy, 215, 228
activation function, 380
AdaBoost, 364
add-one discounting, 357
agglomerative cluster trees, 338
aggregation mechanisms, 83
Akaike information criterion, 289, 335
algorithm analysis, 397
Amazon Turk, 67, 84
  tasks assigned, 85
  Turkers, 84
American basketball players, 97
analogies, 312
anchoring, 82
angular distance, 310
Anscombe's Quartet, 159
AOL, 64
API, 65
Apple iPhone sales, 34
application program interfaces, 65
area under the ROC curve, 219
Aristotle, 326, 327
Arrow's impossibility theorem, 84, 114
artifacts, 69
Ascombe quartet, 272
asking interesting questions, 4
associativity, 244
autocorrelation, 46
average link, 340

Babbage, Charles, 57, 90
backpropagation, 382
Bacon, Kevin, 9
bag of words, 14
bagging, 362
balanced training classes, 295
bar charts, 179
  best practices, 181
  stacked, 181
Barzun, Jacques, 5
baseball encyclopedia, 5
baseline models, 210
  for classification, 210
  for value prediction, 212
Bayes' theorem, 150, 205, 299
Baysian information criteria, 289
bell-shaped, 123, 141
  distribution, 101
bias, 202, 417
  lexicographic, 405
  numerical, 405
  temporal, 405
bias–variance trade-off, 202
big data, 391
  algorithms, 397
  bad data, 392
  statistics, 392
big data engineer, 4
big oh analysis, 397
binary relations, 320
binary search, 398
binomial distribution, 123

© The Author(s) 2017
S.S. Skiena, *The Data Science Design Manual,*
Texts in Computer Science, https://doi.org/10.1007/978-3-319-55444-0

bioinformatician, 50
Blumenstock, Josh, 27
Body Mass Index, 96, 177
Bonferroni correction, 141
boosting, 363, 364
  algorithms, 364
bootstrapping, 374
Borda's method, 108
box plots, 175
Box, George, 201
box-and-whisker plots, 176
bubble plots, 179
bupkis, 391
Bush, George W., 326

C–language, 59
cache memory, 401
canonical representation, 400
canonization, 400
CAPTCHAs, 89
Carroll, Lewis, 423
cartograms, 189
Center for Disease Control, 204
center of mass, 332
centrality measures, 34
centroids, 331
character code unification, 74
characteristic equation, 256
characterizing distributions, 39
chart types, 170
  bar charts, 179
  data maps, 187
  dot and line plots, 174
  histograms, 183
  pie charts, 179
  scatter plots, 177
  tabular data, 170
cicada, 46
classification, 16, 210, 289
  binary, 290, 314
  multi-class, 297
  regression, 290
classification and regression trees, 357
classifiers
  balanced, 217
  evaluating, 213
  one-vs.-all, 298
  perfect, 218
Clinton, Bill, 326
Clinton, Bill], 326
closest pair of points, 398
cloud computing services, 410
cluster
  conductance, 343
  distance, 337
clustering, 327, 373
  agglomerative, 336
  applications, 328
  biological, 337
  cut-based, 341
  k-means, 330
  single link, 339
  visualization of, 337
clusters
  number of, 333
  organization of, 337
Clyde, 111, 112
coefficient vector, 270
Cohen's d, 136
collaborative filtering, 9
communication, 408, 416
commutativity, 243
company data, 64
computer scientist, 2
conditional probability, 30, 31
Condorcet jury theorem, 84
confusion matrix, 214, 220
connected components, 324
contingency table, 214
convex, 280
convex hull, 368
coordination, 408
correlation
  analysis, 40
  interpretation, 43
  significance, 45
correlation and causation, 45, 135
cosine similarity, 309
cross validation, 227
  advantages, 227
CrowdFlower, 67, 84, 86

crowdsourcing, 67, 80
  bad uses, 87
crowdsourcing services, 84
cryptographic hashing, 400
CSV files, 62
cumulative density function, 33, 132, 186
currency conversion, 75
cut, 343

damping factor, 325
Darwin, Charles, 81
data, 237, 391
  quantitative vs. categorical, 15
  big vs. little, 15
  cleaning, 69
  collecting, 64
  compatibility, 72
  errors, 69
  for evaluation, 225
  for testing, 225
  for training, 225
  logging, 68
  properties, 14
  scraping, 67
  structured, 14
  types, 14
  unstructured, 14
  visualizing, 155
data analysis, 404
data centrism, 2
data cleaning, 376
data errors, 417
data formats, 61
data hygiene
  evaluation, 225
data munging, 57
data parallelism, 409
data partition, 404
data processing, 10
data reduction, 329
data science, 1
  languages, 57
  models, 210
data science television, 17
data scientist, 2
data sources, 64
data visualization, 155
data-driven, 156
  models, 207
de Méré, Chevalier, 29
decision boundaries, 291
decision tree classifiers, 299, 357
decision trees, 357
  advantages, 358
  construction, 359
  ensembles of, 362
deep learning, 202, 352, 377
  models, 209
  network, 378
DeepWalk, 385
degree of vertex, 325
depth, 378
derivative, 281
  partial, 282
  second, 281
descriptive statistics, 34
deterministic sampling algorithms, 404
developing scoring systems, 99
dictionary maintenance, 399
DiMaggio, 148
dimension reduction, 277, 376
dimensional egalitarianism, 308
dimensions, 384
dinosaur vertebra, 78
directed acyclic graph, 110
directed graph, 321
discounting, 209, 356
disk storage, 401
distance methods, 303
distance metrics, 304
  L_k, 305
  euclidean, 305
  manhattan distance, 305
  maximum component, 305
distances, 319
  measuring, 303
distributed file system, 396
distributed processing, 407
divide and conquer, 411
DNA sequences, 402

dot product, 241
duality, 268
duplicate removal, 400

E-step, 335
edge cuts, 324
edges, 319
effect size, 136
eigenvalues, 255
    computation, 256
    decomposition, 257
    properties of, 255
Elizabeth II, 326
Elo rankings, 104
embedded graph, 323
embedding, 321
Emoji Dick, 86
Engels, Friedrich, 391
ensemble learning, 363
entropy, 310
equations
    determined, 251
    underdetermined, 251
error, 202, 221
    absolute, 221
    detection, 155
    mean squared, 223, 331
    relative, 222
    residual, 269
    root mean squared, 223
    squared, 222
    statistics, 221
errors vs. artifacts, 69
ethical implications, 416
Euclidean metric, 303
evaluation
    environments, 224
    statistics, 214
event, 28
Excel, 59
exclusive or, 361
exercises, 23, 53, 90, 119, 151, 199, 234, 263, 301, 346, 388, 419
expectation maximization, 335
expected value, 28
experiment, 27
exploratory data analysis, 155, 156
    visualization, 160

F-score, 216
Facebook, 21
false negatives, 214
false positives, 214
fast Fourier transform, 47
fault tolerance, 408
feature engineering, 375
feature scaling, 274
    sublinear, 275
    z-scores, 275
features
    highly-correlated, 277
Fermat, Pierre de, 30
Feynman, Richard, 229
FFT, 47
filtering, 403
financial market, 126
financial unification, 75
fit and complexity, 288
FoldIt, 89
football, 111
    American players, 97
    game prediction, 111
forecasting
    time series, 212
formulation, 354
Freedom of Information Act, 12, 65
frequency counting, 400
frequency distributions, 184
furthest link, 340

Galton, Francis, 81
games with a purpose, 88
gamification, 88
garbage in, garbage out, 3, 69
Gates, Bill, 130
Gaussian
    elimination, 250
Gaussian distribution, 124
Gaussian noise, 125
General Sentiment, 20
genius, 19

geometric mean, 35
geometric point sets, 238
geometry, 240
Gini impurity, 360
Global Positioning System, 12
gold standards, 99
good scoring functions, 101
Goodhart's law, 303
Goodhart, Charles, 303
Google
  AlphaGo, 372
  TensorFlow, 377
Google Flu Trends, 394
Google News, 219
Google Ngrams, 10
Google Scholar, 66
government data, 65
gradient boosted decision trees, 359, 366
gradient descent search, 281
graph embeddings, 384
graph theory, 323
graphs, 238, 319, 321
  cuts, 342
  dense, 322
  directed, 321
  embedded, 323
  labeled, 323
  non-simple, 322
  simple, 322
  sparse, 322
  topological, 323
  undirected, 321
  unlabeled, 323
  unweighted, 322
  weighted, 320, 322
Gray, Dorian, 251
grid indexes, 315
grid search, 409

Hadoop distributed file system, 414
Hamming, Richard W., 1
hash functions, 399
  applications, 399
hashing, 399
heatmaps, 178
hedge fund, 1
hierarchy, 298
higher dimensions, 307, 370
histograms, 32, 183, 222
  best practices, 186
  bin size, 184
Hitler, Adolf, 326
HTML, 67
hypothesis development, 328
hypothesis driven, 156, 392

Imagenet, 379
IMDb, 7
imputation
  by interpolation, 78
  by mean value, 77
  by nearest neighbor, 78
  by random value, 77
  heuristic-based, 77
independence, 30, 123, 354
inflation rates, 76
information gain, 360
information theoretic entropy, 360
infrastructure, 396
inner product, 243
institutional review board, 88
Internet Movie Database, 7
Internet of Things, 68
inverse transform sampling, 132
IPython, 61
IQ testing, 89

Jaccard distance, 341
Jaccard similarity, 341
Jackson, Michael [136], 262
Java, 59
Jesus, 326
JSON, 63

k-means clustering, 343, 409
k-mediods algorithm, 332
k-nearest neighbors, 313
Kaggle, viii
kd-trees, 316
kernels, 371
Kolmogorov-Smirnov test, 139

Kruskal's algorithm, 339
Kullback-Leibler divergence, 311

labeled graphs, 323
Lang, Andrew, 267
Laplace, 356
Laplacian, 343
large-scale question, 9
latent Dirichlet allocation, 373
learning rate, 283, 383
learning to rank, 119
least squares regression, 270
lie factor, 164
Lincoln, Abraham, 242
line charts, 174
  advantages, 174
  best practices, 175
line hatchings, 177
linear algebra, 237
  power of, 237
linear algebraic formulae
  interpretation, 238
linear equation, 238
linear programming, 369
linear regression, 212, 267
  error, 269
  solving, 270
linear support vector machines, 369
linear systems, 250
Linnaeus, Carl, 326
live data, 395
locality, 401
locality sensitive hashing, 317
logarithm, 47
logistic classification
  issues, 295
logistic function, 381
logistic regression, 366
logit, 381
logit function, 106, 292
loss function, 280, 294
LU decomposition, 254
lumpers, 329

M-step, 335
machine learning, 351
  algorithms, 375
  classifiers, 85
  models, 208
main memory, 401
major league baseball, 6
MapReduce, 407, 410
  programming, 412
MapReduce runtime system, 415
matchings, 324
Mathematica, 59, 61
Matlab, 58
matrix, 14, 237, 270, 373
  addition, 242
  adjacency, 246, 320
  covariance, 245, 257, 271
  determinant, 249, 254
  eigenvalues, 255
  eigenvectors, 255
  factoring, 252
  identity, 246, 248
  inversion, 248
  linear combinations of, 242
  multiplication, 243
  multiplicative inverse of, 249
  non-singular, 249
  permutation, 247
  rank, 251
  reasons for factoring, 252
  rotation, 248
  singular, 249
  transpose of, 242
  triangular, 254
  underdetermined, 256
matrix multiplication, 243, 398
  applications, 244
matrix operations
  visualizing, 241
maximum margin separator, 367
mean, 34, 83, 125, 132, 138, 212, 227, 403
  arithmetic, 35
  geometric, 35
measurement error, 125
median, 35, 83, 132, 212, 403
mergesort, 398
metadata, 7

method centrism, 2
metric, 304
  identity, 304
  positivity, 304
  symmetry, 304
  triangle inequality, 304
minima
  global, 284
  local, 284
minimum spanning tree, 324, 339
missing values, 76, 376
mixture model, 333
Moby Dick, 86
mode, 36
model-driven, 417
modeling, 201, 328, 416
  philosophies of, 201
  principles for effectiveness, 203
models
  ad hoc, 208
  baseline, 210
  blackbox, 206
  data science, 210
  data-driven, 207
  deep learning, 209
  descriptive, 206
  deterministic, 208
  evaluating, 212
  first-principle, 207
  flat, 209
  Google's forecasting, 204
  hierarchical, 209
  linear, 206
  live, 204
  machine learning, 208
  neural network, 206
  non-linear, 206
  overfit, 203
  simplifying, 286
  simulation, 229
  stochastic, 208
  taxonomy of, 205
  underfit, 202
Moneyball, 5
monkey, 215
monotonic, 307
Monte Carlo
  sampling, 134
  simulations, 229
Mosteller, Frederick, 121
multiclass systems
  evaluating, 219
multiedge, 322
multinomial regression, 299
multiplying probabilities, 48

naive Bayes, 354, 363
name unification, 74
Napoleon, 326
NASA, 73
National Football League, 112
natural language processing, 20
nearest centroid, 340
nearest neighbor, 339, 397
nearest neighbor classification, 311
  advantages, 311
nearest neighbors
  finding, 315
negative class, 213
Netflix prize, 9
network
  depth, 379
network methods, 303
networks, 109, 238, 319, 378
  induced, 320
  learning, 379
neural networks, 377
new data set, 156
New York, 277
Nixon, Richard, 326
NLP-based system, 21
no free lunch theorem, 353
node
  bias of, 381
non-linear classifiers, 366
non-linear functions
  fitting, 273
non-linear support vector machines, 369
non-linearity, 358, 377, 380
norm, 287
normal, 125

normal distribution, 79, 109, 124, 141
  implications, 126
normality testing, 141
normalization, 103, 376
normalizing skewed distribution, 49
norms, 309
NoSQL databases, 415
notebook environments, 59
numerical conversions, 73

Obama, Barack, 326
Obama, Barack [91], 327
Obama, Barack], 326
Occam's razor, 201, 211, 286
Occam, William of, 202
Oh G-d, 177
optimization
  local, 284
Oracle of Bacon, 9
outlier, 118
  detection, 78
outlier detection, 329
outliers
  removing, 272
overfitting, 202, 296
overlap percentage, 137
ownership, 417

p-values, 145
packing data, 402
PageRank, 100, 111, 325
pairwise correlations, 158
parallel processing, 407
parallelism, 406
parameter fitting, 279
parameter spaces
  convex, 280
partition function, 299
Pascal's triangle, 123
Pascal, Blaise, 30
paths, 246
Pearson correlation coefficient, 41, 136
penalty function, 293
performance of models, 39
periodic table, 188
Perl, 58
permutation, 246
  randomly generating, 147
  tests, 145
personal wealth, 130
pie charts, 179
  bad examples, 183
  best practices, 181
point spread, 112
points
  rotating, 248
points vs. vectors, 309
Poisson distribution, 127
position evaluation function, 372
positive class, 213
power law distribution, 129
power law function, 276
precision, 3, 215, 221
prefetching, 401
principal components, 260
  analysis, 260
prior probability, 354
privacy, 418
probabilistic, 203
probability, 27, 29, 354
probability density function, 32, 132, 186
probability distribution, 32
probability of an event, 28
probability of an outcome, 28
probability vs. statistics, 29
program flow graph, 322
programming languages, 57
protocol buffers, 63
proxies, 99
psychologists, 89
Pubmed, 70
pure partition, 360
Pythagorean theorem, 306
Python, 58, 67

Quant Shop, 17

R, 58
Rand index, 341

random access machine, 397
random sampling, 403, 406
random variable, 28
ranking systems
  class rank, 100
  search results, 100
  top sports teams, 100
  university rankings, 100
rankings, 95
  digraph-based, 109
  historical, 117
  merging, 108
  techniques, 104
ratio, 48
ray
  unit, 241
Reagan, Ronald, 326
rearrangement operations, 238
recall, 216, 221
receiver-operator characteristic curve, 218
rectified linear units, 381
rectifier, 381
redundancy, 393
regression, 16
  application, 359
  for classification, 290
  LASSO, 287
  logistic, 289, 292
  ridge, 286
regression models, 272
  removing outliers, 272
regularization, 286, 376
reinforcement learning, 372
Richter scale, 131
right-sizing training data, 404
road network, 322
robustness, 3, 96, 359
Roosevelt, Franklin D., 326
Rota, Gian-Carlo, 237

sabermetrics, 22
sample space, 27
sampling, 132, 403, 404
  beyond one dimension, 133
  by truncation, 404
scalar
  multiplication, 242
scale invariant, 132
scales
  Likert, 298
  ordinal, 297
scaling constant, 293
scatter plots, 98, 177
  best practices, 177
  three-dimensional, 179
Schaumann, Jan, 351
scientist, 2
scores, 95
  threshold, 218
scores vs. rankings, 100
scoring functions, 95
security, 418
self-loop, 322
semi-supervised learning, 374
Shakespeare, William, 326
sharp, 215
Sheep Market, 86
shortest paths, 324
signal to noise ratio, 37
significance level, 139
Silver, Nate, 203
similarity graphs, 341
similarity matrix, 342
simple graph, 322
single-command program, 224
single-pass algorithm, 402
singular value decomposition, 258
sketching, 403
skew, 413
Skiena, Len, ix
small evaluation set, 226
social media, 392
  analysis, 21
  data, 394
Social Network–movie, 105
spam, 393
spam filtering, 393
Spearman rank correlation coefficient, 42, 108
Spears, Britney [566], 262
spectral clustering, 343

spidering, 67
splitters, 329
sports performance, 38
SQL databases, 63
standard deviation, 36, 125, 132, 138, 227, 403
statistical analysis, 121, 230
statistical distributions, 122
statistical proxy, 7
statistical significance, 135
statistics, 29
stochastic gradient descent, 285, 296, 382
stock market, 37, 79, 126
Stony Brook University, 20
stop words, 414
storage hierarchy, 401
streaming, 402
summary statistics, 157, 159
supervised learning, 372
supervision, 372
    degrees of, 372
support vector machines, 352, 366
support vectors, 368
Surowiecki, James, 82
sweat equity, 66

T-test, 137
tangent line, 282
target scaling, 274
    sublinear, 276
taxi
    records from New York, 11
    tipping model, 286
    tipping rate, 13, 277
taxi driver, 277
terms of service, 68
test statistic, 138
text analysis, 253
theory of relativity, 144
Tikhonov regularization, 287
time unification, 75
Titanic, 183, 358
top-k success rate, 219
topic modeling, 373
topological graph, 323
topological sorting, 110, 324
transparency, 417
tree, 298, 336
trees
    agglomerative, 337
true negatives, 214
true positives, 214
truncation, 405
Tufte, Edward, 155, 162
Twitter, 392, 404, 405

U.S. presidential elections, 187, 203
uncertainty, 175
undirected graph, 321
uniform distribution, 406
uniform sampling, 405
uninvertible, 400
unit conversions, 72
UNIX time, 75
unlabeled graphs, 323
unrepresentative participation, 393
unsupervised learning, 372
unweighted graph, 322
urban transportation network, 11

validation data, 95
value prediction, 210
value prediction models
    evaluating, 221
variability measures, 36
variance, 36, 202, 403
    interpretation, 37
variation coefficient, 137
variety, 394
vectors, 240
    unit, 240
velocity, 394
veracity, 395
vertex, 319
vertices, 319
visualization, 404
    chart types, 170
    critiquing, 189
    interactive, 195
    tools, 160
visualization aesthetic, 162

chartjunk, 162, 165
colors, 163, 168
data-ink ratio, 162, 163
lie factor, 162, 164
repetition, 163, 169
scaling and labeling, 162, 167
volume, 394
Voronoi diagrams, 315
voting, 363
with classifiers, 363

web crawling, 68
weighted average, 84
weighted graph, 322
Welch's t-statistic, 138
Wikipedia, 20, 79, 116, 326
wisdom, 19
wisdom of crowds, 81
Wolfram Alpha, 59, 144
word embeddings, 254, 383
word2vec, 384

XML, 62

Z-score, 103, 308, 376
Zipf's law, 131, 357