Research Design and Applications for Data Analysis

Daniel Kahneman, Thinking, Fast and Slow. Publisher: Farrar, Straus and Giroux; Reprint edition (April 2, 2013), ISBN 978-0374533557

Darrell Huff and Irving Geis, How to Lie with Statistics, Publisher: W. W. Norton & Company; Reissue edition (October 17, 1993), ISBN 978-0393310726

Brian McDonald, Invisible Ink: A Practical Guide to Building Stories that Resonate, Publisher: Libertary Company (January 11, 2010), ISBN 978-0984178629

Edward Tufte, Visual Display of Quantitative Information, Publisher: Graphics Press; 2nd edition (May 2001), ISBN 978-0961392147

Allison, Graham, and Philip Zelikow. Essence of Decision, 2nd edition. Longman, 1999.

EMC Data Science Community, Data Science Revealed: A Data-Driven Glimpse into a Burgeoning New Field. 2012.

LaValle, Steve, Eric Lesser, Rebecca Shockley, Michael S. Hopkins, and Nina Kruschwitz. “Big Data Analytics and the Path from Insights to Value.” MIT Sloan Management Review 52, no. 2 (Winter 2011). http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/

Le Grand, Julian, and Zack Cooper. “The Geeks Must Quash the Believers in Gut Instinct.” Financial Times (February 21, 2012). http://www.ft.com/intl/cms/s/0/5a996db2-5c93-11e1-8f1f-00144feabdc0.html.

Davenport, Thomas H. “Competing on Analytics.” Harvard Business Review (January 2006).

Martin, Roger. “Beyond the Numbers: Building Your Qualitative Intelligence.” Rotman School of Management, 2010.

Salsburg, David. The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. Holt Paperbacks, 2002, chapter 2.

Dutcher, Jenna. “What is Big Data?” Berkeley Data Science Program (September 2014). http://datascience.berkeley.edu/what-is-big-data/

Anderson, Chris. “The End of Theory, Data Deluge Makes Scientific Method Obsolete.” Wired (July 2008). http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Voytek, Bradley. “Automated Science, Deep Data, and the Paradox of Information.” O’Reilly Radar (March 30, 2012). http://radar.oreilly.com/2012/03/data-science-deep-data-information-paradox.html

Shah, Shvetank, Andrew Horne, and Jaime Capellá. “Good Data Won't Guarantee Good Decisions.” Harvard Business Review (April 2012).

Alamar, Benjamin, and Vijay Mehrotra. “Beyond ‘Moneyball’: The Rapidly Evolving World of Sports Analytics, Part I.” Analytics Magazine (September–October 2011).

Heuer, Richard J., Jr. “Psychology of Intelligence Analysis.” Center for the Study of Intelligence, 1999, chapter 6.

Lewis, Michael. “Beane Counter.” Sports Illustrated (May 12, 2003).

Loveman, Gary. “Diamonds in the Data Mine.” Harvard Business Review (May 2003).

Hunter, Kathryn Montgomery. Doctors Stories. Princeton University Press, 1993, 21–26 and 51–57.

Allison, Graham, and Philip Zelikow. Essence of Decision, 2nd edition. Longman, 1999, chapters 3 and 4.

Neustad, Richard, and Ernest May. Thinking in Time: The Uses of History for Decision Makers, 2nd edition. Free Press, 1988, chapter 1.

Stauffer, David. “How Good Data Leads to Bad Decisions.” Harvard Business Publishing Newsletters (2002).

Davenport, Thomas H. “Make Better Decisions.” Harvard Business Review (November 2009).

Hammond, John S., Ralph L. Keeney, and Howard Raiffa. “The Hidden Traps in Decision Making.” Harvard Business Review (September–October 1998).

Kahneman, Daniel. Thinking, Fast and Slow. Farrar, Strauss and Giroux, 2011, chapters 10–18.

Allison, Graham, and Philip Zelikow. Essence of Decision, 2nd edition. Longman, 1999, chapters 5 and 6.

Beal, Dave. “For Numbers Crunchers, Minnesota Twins' Old-School Methods Don't Add Up.” Twin Cities Pioneer Press (June 27, 2012). http://www.twincities.com/twins/ci_20952060/numbers-crunchers-twins-old-school-methods-dont-add.

Davenport, Thomas H., and Brook Manville. Judgment Calls: Twelve Stories of Big Decisions and the Teams That Got Them Right. Harvard Business Review Press, 2012, chapter 2, “WGB Homes: How Can We Sell This House?” and chapter 8, “Mabel Yu and the Vanguard Group: Should We Recommend This Bond to Investors?”

Davenport, Thomas H., and Brook Manville. Judgment Calls: Twelve Stories of Big Decisions and the Teams That Got Them Right. Harvard Business Review Press, 2012, chapter 10, “Should We Restructure for a New Strategy?”

Heuer, Richard J., Jr. “Psychology of Intelligence Analysis.” Center for the Study of Intelligence, 1999, chapters 2, 9–13.

Kahneman, Daniel, and Gary Klein. “Conditions for Intuitive Expertise: A Failure to Disagree.” American Psychologist 64 (2009).

Kahneman, Daniel, and Amos Tversky. “Prospect Theory: An Analysis of Decision under Risk.” Econometrica 47 (1979).

Neyer, Rob. “Phillies Keep Winning without Your Fancy Numbers.” Baseball Nation (March 2, 2012). http://mlb.sbnation.com/2012/3/2/2839053/phillies-keep-winning-without-your-fancy-numbers.

Tverksy, Amos, and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases," Science 185 (1974).

Verducci, Tom. “The Art of Winning an (Even More) Unfair Game.” Sports Illustrated (September 26, 2011). http://sportsillustrated.cnn.com/vault/article/magazine/MAG1190632/index.htm.

Engineering and Public Policy Committee on Science. On Being a Scientist: A Guide to Responsible Conduct in Research. National Academies Press, 2009.

Burton, Robert. On Being Certain. St. Martin’s Griffin, 2009, chapters 1 and 2.

Kuhn, Thomas. The Structure of Scientific Revolutions. University of Chicago Press, 2012, chapter 12.

Creswell, John W. Research Design: Qualitative, Quantitative, and Mixed Methods. Sage Publications, 2008, chapters 1 (approach), 6 (research questions), 7 (theory/RQs), 8 (quantitative methods).

Rao, Venkatesh. “The Dangerous Art of the Right Question.” Trailblazers (July 20, 2010).
http://blog.trailmeme.com/2010/07/the-dangerous-art-of-the-right-question/

Huff, Darrell Huff, and Irving Geis. How to Lie with Statistics. W. W. Norton, 1993.

Best, Joel. Stat-Spotting: A Field Guide to Identifying Dubious Data. University of California Press, 2008, part 1 (p.3-13).

Panger, Galen. “Why the Facebook Experiment is Lousy Social Science Research.”
https://medium.com/@gpanger/why-the-facebook-experiment-is-lousy-social-science-8083cbef3aee

de Vaus, David. Research Design in Social Research. Sage Publications, 2001, chapters 1–3.

Heuer, Richard J., Jr. “Psychology of Intelligence Analysis.” Center for the Study of Intelligence, 1999, chapters 4–5, 8.

Juliano, William. “Was Branch Rickey the Father of Sabermetrics?” The Yankee Analysts (March 28, 2011).
http://www.yankeeanalysts.com/2011/03/was-branch-rickey-the-father-of-sabermetrics-27771.

McDonald, B. Invisible Ink: A Practical Guide to Building Stories That Resonate (Libertary, 2013).

Offenhuber, Dietmar. “Visual Anecdote.” Leonardo 43, no. 4 (August 2010): 367–74.

Tufte, Edward. Visual Display of Quantitative Information. Graphics Press, 2001.

Williams, Harold S. “Informing vs. Persuading.” Innovating 1, no. 2. The Rensselaer Institute.

Kleiner, Art, and George Roth. "How to Make Experience Your Company's Best Teacher.” Harvard Business Review (September 1997).

Gray, Jonathan, Liliana Bounegru, and Lucy Chambers. “Data Journalism in Perspective.” The Data Journalism Handbook. 2012.
http://datajournalismhandbook.org/1.0/en/introduction_4.html.

Laurila, David. “Jon “Boog” Sciambi: Broadcasting the Stats.” FanGraphs Baseball (March 12, 2012). http://www.fangraphs.com/blogs/index.php/jon-boog-sciambi-broadcasting-the-stats/.

Steele, Julie, and Noah Iliinsky. Beautiful Visualization: Looking at Data through the Eyes of Experts (Theory in Practice). O’Reilly Media, 2010.

Duhigg, Charles. “How Companies Learn Your Secrets.” New York Times (February 16, 2012). http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Nonaka, Ikujiro. “The Knowledge-Creating Company.” Harvard Business Review (November 1991).

Enriquez, Juan, Gary P. Pisano, and Gaye L. Bok. “In Vivo to in Vitro to in Silico: Coping with Tidal Waves of Data at Biogen.” Harvard Business School, 2002.

Matz, Eddie. “Saviormetrics.” ESPN, The Magazine (August 13, 2012). http://espn.go.com/mlb/story/_/id/7602264/oakland-brandon-mccarthy-writing-moneyball-next-chapter-reinventing-analytics-espn-magazine.

Boudway, Ira. “Baseball: Running the New Numbers.” Bloomberg Businessweek (March 31, 2011). http://www.businessweek.com/magazine/content/11_15/b4223072802462.htm.

Lewis, Peter H. “For the Love of the Technology, the Bay Area Is Reinventing Baseball (Again).” The New York Times (April 26, 2012). http://www.nytimes.com/2012/04/27/us/for-the-love-of-the-technology-san-francisco-is-reinventing-baseball-again.html.

Neyer, Rob. “FIELDf/x Is Going to Change Everything.” ESPN (August 30, 2010). http://espn.go.com/blog/sweetspot/post/_/id/5041/fieldfx-is-going-to-change-everything

Storing and Retrieving Data

The Discipline of Organizing - Glushko Bad

Data Handbook - O'Reilly - McCallum Optional:

Doing Data Science - O'Reilly - O'Neil & Schutt

Data Science for Business - O'Reilly - Provost & Fawcett
Machine Learning for Hackers - O'Reilly - Conway & White

Ferrucci, D., Brown, E., Chu-Carooll, J., Fan, J., Gondek, D., Kalyanpur, A. A., . . . Welty, C. (2010, Fall). Building Watson: An overview of the DeepQA project. AI Magazine, 59–79.

Leskovec, J., Rajaraman, A., & Ullman, J. (2011). Mining of massive datasets. New York, NY: Cambridge University Press. Chapter 1

Manning, C. D., Raghavan, P., & Schuetze, H. (2008). Introduction to information retrieval. New York, NY: Cambridge University Press. Chapter 1

Saracevic, T. (1975, November–December). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 321–343.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1975, reprinted 1990). A re-examination of relevance: Toward a dynamic, situational definition. Information Processing & Management, 26(6), 766–776.

Stonebraker, M. (2009, June 30). The end of a DBMS era (might be upon us). ACM Blogs.

Vu, L. (2012, July 11). Getting value from a trillion electron haystack. iSGTW. o
Journal of the American Society for Information Science, 45(3) (1994, April).
This journal is available via the UCB Library Proxy.

Hey, T., Tansley, S., & Tolle, K. (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: Microsoft Research.

Download this week's chapter: Jim Gray on eScience: A Transformed Scientific Method Jagadish, H. V., et al. "Big Data and Its Technical Challenges".

Communications of the ACM v. 57, n. 7 (July 2014).

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003, October). The Google File System. SOSP'03.

Leskovec, J., Rajaraman, A., & Ullman, J. (2011). Mining of massive datasets. New York, NY: Cambridge University Press. Chapter 4

Lin, J., & Dyer, C. (2010, April). Data-intensive text processing with MapReduce. Manuscript to appear in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Daniel Jurafsky & James H Martin. (2nd edition). "Speech and Language Processing". Chapter 2.1

Cattell, R. (2010, December). Scalable SQL and NoSQL data stores. SIGMOD Record, 39(4).

Manning, C. D., Raghavan, P., & Schuetze, H. (2008). Introduction to information retrieval. New York: NY: Cambridge University Press. (Chap. 1)

Seltzer, M. (2005, April). Beyond relational databases. ACM Queue.

Hoffer, J. A., Ramesh, V., & Topi, H. (2012). Modern database management (11th ed.). Upper Saddle River, NJ: Prentice Hall (Pearson Educational).

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. New York, NY: Cambridge University Press. Revisiting Chapter 1

Kryder, M. H., & Kim, C. S. (2009, October). After hard drives—What comes next? IEEE.

Stonebraker, M. (2010, April). SQL databases v. NoSQL databases. Communications of the ACM, 53(4), 10–11.

Manning, C. D., Raghavan, P., & Schuetze, H. (2008). Introduction to information retrieval. New York: NY: Cambridge University Press.

Jimmy Lin and Chris Dyer, Data-Intensive Text Processing with MapReduce, Morgan

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., . . . Gruber, R. E. (2006). Bigtable: A distributed storage system for structured data. OSDI 2006.

Dean, J., & Ghemawat, S. (2010, January). MapReduce: A flexible data processing tool. Communications of the ACM, 53(1).

Ghemawat, S. (2003, October). The Google File System. In SOSP'03.

Hiemstra, D., & Hauff, C. (2010). MapReduce for information retrieval evaluation: "Let's quickly test this on 12 TB of data." In M. Agosti et al. (Eds.), CLEF 2010, LNCS 6360, 64–69.

Hiemstra, D., & Hauff, C. (2010). University of Twente at TREC 2010: MapReduce for experimental search. In TREC19 Proceedings. Gaithersburg, MD: NIST.

Lin, J., & Dyer, C. (2010). Data-intensive text processing with MapReduce. Morgan & Claypool.

Download this week's chapter 2: MapReduce Basics

Download this week's chapter 7: Limitations on MapReduce

Stonebraker, M., Abadi, D., DeWitt, D. J., Madden, S., Paulson, E., Pavlo, A., & Rasin, A. (2010, January). MapReduce and parallel DBMSs: Friends or foes? Communications of the ACM, 53(1).

Salton, G. (1981, Fall). A blueprint for automatic indexing. ACM SIGIR Forum in Cornell University's newsletter, 16(2), 22–38.

Olston, C., Reed, B., Srivastava, U., Kuma, R., & Tomkins, A. (2008, June). Pig Latin: A not-so-foreign language for data processing. SIGMOD'08, Vancouver, BC, Canada.

Pavlo, A., Paulson, E., Rasin, A., Abadi, D. J., DeWitt, D. J., Madden, S., & Stonebraker, M. (2009, June–July).

A comparison of approaches to large-scale data analysis. SIGMOD'09, Providence, Rhode Island.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, Fall). From data mining to knowledge discovery in databases. AI Magazine, 17(3).

Pazzani, M. J. (2000, March/April). Knowledge discovery from data?. IEEE Intelligent Systems.

Inmon, W. H. (2000). Building the data warehouse: Getting started.

Meijer, E., & Bierman, G. (2011, March). A co-relational model of data for large shared data banks. Programming Languages, 9(3).

Abadi, D. J. et al. (2003). Aurora: A new model and architecture for data stream management. The VLDB Journal, 12(2), 120–139.

Carney, D., Cetintemel, U., Cherniack, M., Convey, C., Lee, S., Seidman, G., . . . Zdonik,

S. (2002). Monitoring streams—A new class of data management applications. Proceedings of the 28th VLDB Conference Hong Kong.

Lin, J. (2013, March). MapReduce is good enough? Big Data, 1(1). Mary Ann Liebert, Inc.

Exploring and Analyzing Data

Field, Andy, Jeremy Miles, and Zoe Field. Discovering Statistics Using R. SAGE Publications, 2012.

Bernard, Russell H. Social Research Methods: Qualitative and Quantitative Approaches, 2nd ed. SAGE Publications, 2000, chapter 5 “Sampling.”

Freedman, David, Robert Pisani, Roger Purves. Statistics, 4th ed. W. W. Norton & Company, Inc., 2007, chapter 13 “What Are the Chances?” and chapter 14 “More about Chance.”

Freedman, David, Robert Pisani, Roger Purves. Statistics, 4th ed. W. W. Norton & Company, Inc. 2007, chapter 1 “Controlled Experiments.”

Dahlia K. Remler, Gregg G. Van Ryzin. Research Methods in Practice: Strategies for Description and Causation, chapter 13 “Natural and Quasi Experiments.”

Applied Machine Learning

Read Halevy, A., Norvig, P., & Pereira, F. (2009).The unreasonable effectiveness of data.

Intelligent Systems (IEEE).

Optional: Provost and Fawcett. Data science for business.

Read Feynman, R. (1974, June). Cargo cult science. Engineering and Science 37(7).

Read Domingos, P. (2012). A few useful things to know about machine learning.

Communications of the ACM.

Skim Hawkins, D. (2004).The problem of overfitting. Journal of Chemical Information and Computer Sciences.

Read Paul Graham on Naive Bayes in 2002.

Skim Michael Collins tutorial on Naive Bayes (with math), see pages 1–4.

Skim Kanich, C. et al. (2008). Spamalytics: An empirical analysis of spam marketing conversion. ACM conference on Computer and Communications Security.

Read as much as you can Carter, T. (2001, June). An introduction to information theory and entropy.

Read blog post from yhat about predicting churn. Read short introduction to Adaboost

Read Chapter 5 of Schutt & O'Neill. (2013). Doing data science. Read Section 4.6 of Whitten, Frank, & Hall. Data mining.

Optional: Chapter 3 (Sections 3.1 and 3.2), Chapter 4 (especially section 4.4), and Chapter 6 (sections 6.16.3) of Friedman, Hastie, & Tibshirani. The elements of statistical learning.

Read Chapter 6 of Daum, H. A course in machine learning.

Read Chapter 8 of Daume. A course in machine learning.

Read Chapter 7 (section 7.4) of Whitten, Frank, & Hall. Data mining.

Optional Cosma Shalizi SVM lecture notes.

Read An empirical comparison of supervised learning algorithms.

Skim On comparing classifiers: Pitfalls to avoid and a recommended approach. Skim SKLearn classifier comparisons for toy problems.

Read Chapter 7 (sections 7.1–7.3) of Rajarman et al. Mining of massive datasets.

Read Whitten, Frank, & Hall. Chapter 4.8.

Optional: Zhao, Y., Karypis, G., & Fayyad, U. (2005).Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery.

Optional: Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genomewide expression patterns. Proceedings of the National Academy of Sciences, 95: 14863–14868.

Read Tibshirani lecture notes on EM.

Read Doug Reynolds original paper on GMMs for speaker identification.

Read Turk & Pentland. (1991). Eigenfaces for recognition.

Read Chapter 11 (sections 11.1–11.3) of Rajarman et al. Mining of massive datasets.

Read Chapter 7 (section 7.4) of Whitten, Frank, & Hall. Data mining.

Optional: Chapter 14 (sections 14.2, 14.5–14.10) of Friedman, Hastie, & Tibshirani. The elements of statistical learning.

Read Godbole, N. et. al. (2007). Largescale sentiment analysis for news and blogs. International Conference on Weblogs and Social Media.

Read Page, L. et al. (1999). The PageRank citation ranking: Bringing order to the web. Stanford.

Read Chapter 8 of Schutt & O'Neill. (2013). Doing data science.

Read Chapter 9 of Rajarman et al. Mining of massive datasets.

Optional: Koren, Y. (2009). The BellKor solution to the Netflix grand prize.

Optional: Resnick et al. (1994). GroupLens: An open architecture for collaborative filtering of netnews. CSCW: 175–186.

Optional: Bell, R. M., & Koren Y. (2007). Lessons from the Netflix prize challenge. ACM SIGKDD Explorations Newsletter.

Read Chapter 5 of Whitten, Frank, & Hall.

Visualizing and Communicating Data

Miller,J.E.(2012).TheChicagoGuidetoWritingaboutMultivariate Analysis, second edition, Chicago: University of Chicago Press. ("CGWMA" below)

Few, Stephen. Show Me the Numbers, Analytics Press, 2012. ("SMTN" below)

Tufte,Edward.TheVisualDisplayofQuantitativeInformation,2nd Edition. Cheshire, CT: Graphics Press.

Strunk, W. and E.B. White, The Elements of Style, 3rd Edition, Macmillan, 1979 (or 4th edition, or Illustrated edition, or 50th Anniv. edition—or what you already own).

TheChicagoManualofStyle,availablefreetostudents online,
http://www.chicagomanualofstyle.org/16/contents.html

Tufte, E.R. (2001). "Graphical Excellence" (pp. 13-51) and "Graphical Integrity" (pp. 52-77) from The Visual Display of Quantitative Information, 2nd Edition. Cheshire, CT: Graphics Press.

Miller,CGWMA,"Chapter2,SevenBasicPrinciples."

Miller,CGWMA,"AppendixA,Implementing'Generalization,Example, Exceptions"

ACMCodeofEthics,
http://www.acm.org/about/code-of-ethics

PrinciplesandRulesofCopyright,AdobeSystemsInc.,2013.

Ph.D.comic,"TheScienceNewsCycle,"May18,2009,
http://www.phdcomics.com/comics/archive.php?comicid=1174

Miller,CGWMA,"Chapter20,WritingforAppliedAudiences."

Gould,StephenJay(1980)"ThePanda'sThumb,"inThePanda's Thumb: More Reflections in Natural History, pp. 19-26.

Freedman,David,andRobertPisaniandRogerPurves(2007),"TheLaw of Averages", Chapter 16 in Statistics, 4th Edition, pp. 273-287. 3.

GraphDesignI.Q.Test,StephenFew,2009.
http://www.perceptualedge.com/files/GraphDesignIQ.html

Few,SMTN,“Chapter3,DifferingRolesofTablesandGraphs.”

Few,SMTN,“Chapter4,FundamentalVariationsofTables.”

Few,SMTN,“Chapter6,FundamentalVariationsofGraphs.”

Tufte,Edward(2000).Exerptfrom“Words,Numbers,Images—Together,” in Beautiful Evidence, Graphics Press, pp. 97-101, 106-109, 114-121.

Heer,J.,Bostock,M.,Ogievetsky,V.“ATourThroughtheVisualization Zoo,” ACM Queue, 2010.

http://xkcd.com/833/

“Aclassificationofvisualrepresentations",Lohse,Biolsi, Walker, Reuter, CACM 1994.

Nielsen,Jakob,“Chapter5,UsabilityHeuristics”and“Chapter7, Usability Testing” in Usability Engineering, 1993.

Strunk, W. and E.B. White, The Elements of Style, 3rd Edition, Macmillan, 1979.

Murray,Scott,InteractiveDataVisualizationfortheWeb,O’Reilly,2013, Chapters 5-8.
http://chimera.labs.oreilly.com/books/1230000000345/index.html

Few,SMTN,“Chapter5,VisualPerceptionandGraphical Communication”

Tufte,E.R.(1990).Exerptfrom“ColorandInformation,”inEnvisioning Information. Cheshire, CT: Graphics Press, pp. 80-85, 90-95

Borland,D.andR.M.Taylor,“RainbowColorMap(Still)Considered Harmful,” in IEEE Computer Graphics and Applications, March/April 2007, pp. 14-17.

Ware,C.“Chapter1,FoundationsforanAppliedScienceofData Visualization,” in Information Visualization: Perception for Design, 3rd Edition, Morgan Kaufman, Waltham, MA, 2013.

Albers,Josef,InteractionofColor,YaleUniversityPress,NewHaven, 1975, pp.1-11.

Tinkel,Kathleen,"TakingItIn:WhatMakesTypeEasytoReadandWhy," Adobe Magazine, March/April 1996, pp. 41-45.

Krause,Jim,DesignBasicsIndex,HowDesignBooks,Cincinnati,2004, pp. 15-19, 21-23, 34-35, 41-45, 63-69, 74-77, 79-85.

Shneiderman,Ben.“TheEyesHaveIt:ATaskbyDataTypeTaxonomy for Information Visualizations,” Proc. IEEE Conference on Visual Languages, Boulder 1996.

Wickham,H.,Cook,D.,Hofmann,H.,andBuja,A.“Graphicalinference for infovis.” In IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '10), vol. 16, no. 6, 2010, pp. 973–979. http://vita.had.co.nz/papers/inference-infovis.pdf

Few,SMTN,“Chapter7,GeneralDesignforCommunication.”

Miller,CGWMA,“Chapter7,ChoosingEffectiveExamplesand Analogies.”

Miller,CGWMA,“Chapter8,BasicTypesofQuantitativeComparisons.”

Miller,CGWMA,“Chapter12,WritingIntroductions,Conclusions,and Abstracts.”

Miller,CGWMA,“Chapter13,WritingaboutDataandMethods.”

Miller,CGWMA,“Chapter3,Causality,StatisticalSignificance,and Substantive Significance.”

Miller,CGWMA,“Chapter4,FivemoreTechnicalPrinciples.”

Miller,CGWMA,“Chapter11,ChoosingHowtoPresentStatisticalTest Results.”

Miller,CGWMA,“Chapter14,WritingaboutDistributionsand Associations.”

Cham,Jorge,“YourConferencePresentation,”Ph.D.Comics, http://www.phdcomics.com/comics/archive.php?comicid=1553

Few,SMTN,Chapter12,“MultivariateAnalysis.”

Fisher,D.,“Chapter19,AnimationforVisualization:Opportunitiesand Drawbacks,” in Beautiful Visualization, pp. 329-352.

Inselberg,A.“MultidimensionalDetective,”IEEE,June1997,pp.100-107.

Marx,V.,“DataVisualization:AmbiguityasaFellowTraveler,”Nature Methods, Vol. 10, No. 7, July 2013, pp. 613-615.

(optional)Gratzl,S.,Lex,A.,Gehlenborg,N.,Pfister,H.andStreit,M.

“LineUp: Visual Analysis of Multiattribute Rankings,” IEEE Transactions on Visualization and Computer Graphics, Vol. 19, No. 12, December 2013. http://data.icg.tugraz.at/caleydo/publication/2013_InfoVis_Gratzl_Line Up.pdf, also http://www.youtube.com/watch?v=iFqCBI4T8ks

Wattenberg,M.“VisualExplorationofMultivariateGraphs,” CHI 2006 Proceedings, Montreal, pp. 811-819.

Few,Stephen,“Time-SeriesAnalysis.”inNowYouSeeIt,AnalyticsPress, 2009, 143-188.

Aigner,W.,S.Miksch,W.Müller,etal.,“Visualmethodsforanalyzing Time-oriented data,” IEEE Transactions on Visualizations and Computer Graphics, Jan/Feb 2008, pp. 47-60.

Schmandt,M.“Chapter1:IntroductiontoGISandMapping,”GIS Commons: An Introductory Textbook on Geographic Information Systems, http://giscommons.org.

Monmonier,M.(1996).Chapter2:ElementsoftheMap.InHowtoLie with Maps. Chicago, University Of Chicago Press.

Few,Stephen.“Chapter3,ThirteenCommonMistakesinDashboard Design” and “Chapter 7, Designing Dashboards for Usability”, in Information Dashboard Design: The Effective Visual Communication of Data, O’Reilly Media, 2006.

WAI-ARIA Introduction http://www.w3.org/TR/wai-aria/introduction

WCAG2.0ataGlance http://www.w3.org/WAI/WCAG20/glance/

Forreference: http://www.w3.org/TR/wai-aria/

http://www.w3.org/TR/WCAG20/ http://www.section508.gov

Viégas,F.B.andM.Wattenberg,“ArtisticDataVisualization:Beyond Visual Analytics,” HCII 2007.

Manovich,Lev,TheAnti-SublimeIdealinDataArt,self-published,2002. http://www.manovich.net/DOCS/data_art.doc. Originally published as The Anti-Sublime Ideal in New media”, in the online journal Chair et métal 7 (2002).

Williams,JosephM.,Chapter9,“Elegance,”inStyle:TowardClarity and Grace, University of Chicago Press, Chicago, 1990, pp. 153-166.
Reinard,J.C. “The empirical study of the persuasive effects of evidence: The status after fifty years of research,” Human Communication Research, Vol. 15, No.1, Fall 1988, pp. 3-30 (31-59 optional).

Field Experiments

FE: Field Experiments: Design, Analysis, and Interpretation, by Alan S. Gerber and Donald P. Green

MHE: Mostly Harmless Econometrics: An Empiricist's Companion, by Joshua D. Angrist and Jörn-Steffen Pischke (MHE).

MTGI: More Than Good Intentions, by Dean Karlan and Jacob Appel. This is a popular-press book rather than a textbook; it introduces us to many examples of valuable experiments in development economics.

NYTimes HRT article

Feynman

three news articles

Lewis and Reiley [through section III.B]

Karlan and Appel book: focus on chapters 1, 5, 8, 9.

Lewis and Rao [sections 1, 3.1, 3.2, 4.1, 4.2]

Ayres *et al.* (Opower)

Skim List and Lucking-Reiley

Johnson, Lewis, and Reiley (Sections 1, 2, 3.1, 4.3)

Goodson

Gerber and Green 2005

Johnson, Lewis, and Reiley (Sections 3.2-4.1, 5)

Miguel and Kremer (Sections 1-3,8-9)

Blake and Coey (Sections 2 and 3)

DiNardo and Pischke (skim)

Simonsohn *et al.* (skim)

incinerator synopsis (DID)

Washington 2008 (natural experiment) (skim)

Lalive (RD) (skim)

Allcott and Rogers

Sherman et al.

Freedman: “Shoe Leather”

Legal, Policy, and Ethical Considerations for Data Scientists

Skloot, Rebecca. 2013. “The Immortal Life of Henrietta Lacks, the Sequel.” The New York Times. http://www.nytimes.com/2013/03/24/opinion/sunday/the-immortal-life-of-henrietta-lacks-the-sequel.html.

“The Henrietta Lacks Foundation.” 2014. http://henriettalacksfoundation.org/.
boyd, danah, and Kate Crawford. 2012. “Critical Questions for Big Data.”

Information, Communication & Society 15 (5): 662–79. doi:10.1080/1369118X.2012.678878.

Harding, S. (1991). *Whose Science? Whose Knoweldge? Thinking from Women’s Lives*. (Ch. 1)

Jurgenson, N. (2014). “The View from Nowhere.” *The New Inquiry*. http://thenewinquiry.com/essays/view-from-nowhere/

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979).

Ethical principles and guidelines for the protection of human subjects of research [The Belmont Report]. Department of Health, Education, and Welfare.

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979.

The Belmont Report - Office of the Secretary, Ethical Principles and Guidelines for the Protection of Human Subjects Research. Washington, DC.

http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

Crawford, Kate. 2013. “The Hidden Biases in Big Data.” Harvard Business Review. http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/.

Davila, Florangela. 2002. “USDA Disqualifies Three Somalian Markets from Accepting Federal Food Stamps.” The Seattle Times. http://community.seattletimes.nwsource.com/archive/?date=20020410&slug=somalis10m.

Boston pothole reporter app: http://www.streetbump.org/.

Secretary’s Advisory Committee on Automated Personal Data Systems. 1973.

Records Computers and the Rights of Citizens. Washington, DC. http://www.justice.gov/sites/default/files/opcl/docs/rec-com-rights.pdf.

OECD. 2013. “Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data (2013)”, 11–37.

Bond, Carol S, Osman Hassan Ahmed, Martin Hind, Bronwen Thomas, and Jaqui Hewitt-Taylor. 2013. “The Conceptual and Practical Ethical Dilemmas of Using Health Discussion Board Posts as Research Data.” Journal of Medical Internet Research 15 (6): e112. doi:10.2196/jmir.2435.

Hayden, Erika Check. 2013. “THE GENOME.” Nature 497: 172–74. http://www.nature.com/polopoly_fs/1.12940!/menu/main/topColumns/topLeftColumn/pdf/497172a.pdf.

Rothstein, Mark A., and Abigail B. Shoben. 2013. “Does Consent Bias Research.” American Journal of Bioethics 13. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2244990.

Facebook Contagion Study

Kramer, Adam D I, Jamie E Guillory, and Jeffrey T Hancock. 2014. “Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks.”

Proceedings of the National Academy of Sciences of the United States of America 111 (24): 8788–90. doi:10.1073/pnas.1320040111.

Tufekci, Z.: Facebook & Engineering the public
https://medium.com/message/engineering-the-public-289c91390225

Tufekci, Z.: Engineering the Public: Big Data, Surveillance and Computational Politics (forthcoming in First Monday)
http://technosociology.org/wp-content/uploads/2014/06/Zeynep-Computational-Politics-and-Engineering-the-Public.pdf *(OPTIONAL ADDITIONAL READING)*

Grimmelmann, J.: "Illegal, Immoral, & Mood-Altering" - https://medium.com/@JamesGrimmelmann/illegal-unethical-and-mood-altering-8b93af772688 *(OPTIONAL ADDITIONAL READING)*

Terrorist Watch List:
https://firstlook.org/theintercept/article/2014/08/05/watch-commander/

OkCupid Experiments:
http://blog.okcupid.com/index.php/we-experiment-on-human-beings/

John Oliver and Native Advertising:
https://www.youtube.com/watch?v=E_F5GxCwizchttps://www.youtube.com/watch?v=E_F5GxCwizc

16 CFR Part 312 - Children's Online Privacy Protection Rule. USA. http://www.law.cornell.edu/cfr/text/16/part-312.

John Kropf, Public Information and Privacy in a Global Society, BNA Privacy & Security Law Report, March, 24, 2014
http://privacylaw.bna.com/pvrc/7057/split_display.adp?fedfid=43383440&vname=pvlrnotallissues&wsn=499496500&searchid=22477163&doctypeid=6&type=oadate4news&mode=doc&split=0&scm=7057&pg=0

Health and Human Services. 2003. OCR HIPAA Privacy-Research. Vol. 512. Washington, DC. http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/research/research.pdf.

45 CFR 164.514 - OTHER REQUIREMENTS RELATING TO USES AND DISCLOSURES OF PROTECTED HEALTH INFORMATION. USA. http://www.law.cornell.edu/cfr/text/45/164.514 .

In re GOOGLE INC. STREET VIEW ELECTRONIC COMMUNICATIONS LITIGATION. 2011 794 F. Supp. 2d 1067 (N.D. CA). http://www.leagle.com/decision/In%20FDCO%2020110630A55.xml/IN%20RE%20GOOGLE%20INC.%20STREET%20VIEW%20ELECTRONIC%20COMM

Kravets, David. 2013. "Google 's Wi-Fi Sniffing Might Break Wiretap Law, Appeals Court Rules." Wired. http://www.wired.com/2013/09/googles-wifi-wiretapping/.

Kravets, David. 2012. "An Intentional Mistake : The Anatomy of Google's Wi-Fi Sniffing Debacle." Wired. http://www.wired.com/2012/05/google-wifi-fcc-investigation/.

Ohm, Paul. 2014. "Should Sniffing Wi-Fi Be Illegal ?" IEEE On the Horizon 12 (1): 73–76. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=06756905.
boyd, danah, and Alice Marwick. 2011. "Social Privacy in Networked Publics : Teens ' Attitudes , Practices , and Strategies", 1–29. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1925128.

Schwartz, Paul M, and Daniel J Solove. 2013. "Reconciling Personal Information in the United States and European Union", 877–916.

http://dx.doi.org/10.2139/ssrn.2271442 . (excerpt pp. 6-12)

Narayanan, Arvind, and Vitaly Shmatikov. 2009. "De-Anonymizing Social Networks." IEEE S&P. http://www.cs.utexas.edu/~shmat/shmat_oak09.pdf.

Jernigan, Carter, and Behram Mistree. 2009. "Gaydar: Facebook Friendships Expose Sexual Orientation." First Monday 14 (10). http://firstmonday.org/article/view/2611/2302.

Berkeley, UC. 2011. "Committee for Protection of Human Subjects -- Informed Consent." http://cphs.berkeley.edu/consent.pdf. [Skim]

"UC Berkeley Research Administration and Compliance -- Human Research Protection Program." http://cphs.berkeley.edu/review.html.

Zimmer, Michael. 2010. "'But the data is already public'': On the Ethics of Research in Facebook." *Ethics and Information Technology* 12: 313–325.

Underwood, Marion K., et al. "The BlackBerry project: capturing the content of adolescents' text messaging." Developmental psychology 48.2 (2012): 295.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289770/

http://www.forbes.com/sites/kashmirhill/2012/04/18/a-texas-universitys-mind-boggling-database-of-teens-daily-text-messages-emails-and-ims-over-four-years/

http://www.michaelzimmer.org/2012/04/25/research-ethics-and-the-blackberry-project/

Certificates of Confidentiality http://grants.nih.gov/grants/policy/coc/index.htm
Molloy, Jennifer C. 2011. “The Open Knowledge Foundation: Open Data Means Better Science.” PLoS Biology 9 (12): e1001195.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3232214&tool=pmcentrez&rendertype=abstract.

Janssen, Marijn, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. “Benefits, Adoption Barriers and Myths of Open Data and Open Government.” Information Systems Management 29 (4): 258–68.

http://www.tandfonline.com/doi/abs/10.1080/10580530.2012.716740.
Big Data Panel II: Deep dive on new opportunities and challenges in health and education

Planet lab http://www.planet-lab.org/
http://www.cs.princeton.edu/~llp/policy.pdf
CHAPTER 22.1. Privacy Rights for California Minors in the Digital World [22580 - 22582]. 2014. State of California.
http://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=BPC&sectionNum=22581.

California Office of Privacy Protection. 2012. “Recommended Practices on Notice of Security Breach Involving Personal Information January 2012.”
http://oag.ca.gov/sites/all/files/agweb/pdfs/privacy/recom_breach_prac.pdf.
(excerpt pp. 5-15)

Perlroth, Nicole. 2014. “Hackers Lurking in Vents and Soda Machines.” New York Times, 18–21. http://www.nytimes.com/2014/04/08/technology/the-spy-in-the-soda-machine.html.

US Department of Health and Human Services Office of Civil Rights. 2010. “Annual Report to Congress on Breaches of Unsecured Protected Health Information For Calendar Years 2009 and 2010 As Required by the Health Information Technology for Economic and Clinical.” http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/breachrept.pdf.

In re Facebook, Complaint, FTC File No. 092 3184 [PDF]

In the Matter of Eli Lilly [PDF]

Faden, Ruth R, Nancy E Kass, Steven N Goodman, Peter Pronovost, Sean Tunis, and Tom L Beauchamp. 2013. “An Ethics Framework for a Learning Health Care System: A Departure from Traditional Research Ethics and Clinical Ethics.” The Hastings Center Report Spec No (February): S16–27. doi:10.1002/hast.134.

http://onlinelibrary.wiley.com/doi/10.1002/hast.134/abstract

Willis, James E., John P. Campbell, and Matthew D. Pistilli. 2013. “Ethics, Big Data, and Analytics: A Model for Application.” EDUCAUSE Review 48 (3). http://www.educause.edu/ero/article/ethics-big-data-and-analytics-model-application.

Stephens-Davidowitz, Seth. 2013. “UNREPORTED VICTIMS OF AN ECONOMIC.” http://static.squarespace.com/static/51d894bee4b01caf88ccb4f3/t/51e22f38e4b0502fe211fab7/1373777720363/childabusepaper13.pdf.

Jane Robbins, The Ethics of MOOCs, Inside Higher Ed March 25, 2013 http://www.insidehighered.com/blogs/sounding-board/ethics-moocs

The Asilomar Convention for Learning Research in Higher Education http://asilomar-highered.info/

Wen, Miaomiao, Diyi Yang, and Carolyn Penstein Rosé. “Sentiment Analysis in

MOOC Discussion Forums : What Does It Tell Us ?" Proceedings of Educational Data Mining. http://www.cs.cmu.edu/~mwen/papers/edm2014-camera-ready.pdf.

Gillespie, Tarleton. "The Relevance of Algorithms." In Media Technologies, edited by Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot. Cambridge, MA: MIT Press. http://www.tarletongillespie.org/essays/Gillespie - The Relevance of Algorithms.pdf.

Ziewitz, Malte. 2011. How to think about an algorithm: Notes from a not quite random walk. http://zwtz.org/files/ziewitz_algorithm.pdf

Barocas, Hood, and Ziewitz. 2013. Governing Algorithms: A Provocation Piece. http://governingalgorithms.org/resources/provocation-piece/

Kraemer, Felicitas, Kees Overveld, and Martin Peterson. 2010. "Is There an Ethics of Algorithms?" Ethics and Information Technology 13 (3): 251–60. doi:10.1007/s10676-010-9233-7. http://link.springer.com/10.1007/s10676-010-9233-7

The Constitution Project. 2010. "Principles for Government Data Mining Preserving Civil Liberties in the Information Age." http://www.constitutionproject.org/wp-content/uploads/2012/09/DataMiningPublication.pdf.

Stuart, G., "Databases, Felons, and Voting: Errors and Bias in the Florida Felons Exclusion List in the 2000 Presidential Elections" (September 2002). KSG

Working Paper Series RWP 02-041. Read pp. 22-40

United States of America (for the Federal Trade Comm'n) v. Spokeo Inc., Civ. No. CV12-05001 (C.D. Cal. June 12, 2012).

The complaint, http://www.ftc.gov/sites/default/files/documents/cases/2012/06/120612spokeocmpt.pdf

Stipulation http://www.ftc.gov/sites/default/files/documents/cases/2012/06/120612spokeostip.pdf

The Consent Decree
http://www.ftc.gov/sites/default/files/documents/cases/2012/06/120612spokeoorder.pdfhttp://www.ftc.gov/sites/default/files/documents/cases/2012/06/120612spokeoorder.pdf

The Leadership Conference. 2014. “Civil Rights Principles for the Era of Big Data.” http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html.
Barocas & Selbst - “Big Data’s Disparate Impact”
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899http://link.springer.com/10.1007/s10676-010-9233-7

Gandy, Oscar H. 2009. “Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems.” Ethics and Information Technology 12 (1): 29–42. doi:10.1007/s10676-009-9198-6. http://link.springer.com/10.1007/s10676-009-9198-6

Center for Media Justice. 2013. Consumers , Big Data , and Online Tracking in the Retail Industry A CASE STUDY OF WALMART. http://centerformediajustice.org/wp-content/files/WALMART_PRIVACY_.pdf.

Dwork, Cynthia, and Deirdre K Mulligan. 2013. “It’s Not Privacy, and It's Not Fair”, 35–40. http://www.stanfordlawreview.org/sites/default/files/online/topics/DworkMullliganSLR.pdf.

Calo, Ryan. 2013. “CONSUMER SUBJECT REVIEW BOARDS: A THOUGHT EXPERIMENT.” Stanford Law Review Online, 97–102. http://www.stanfordlawreview.org/sites/default/files/online/topics/Calo.pdf.

Lerman, Jonas. 2013. “Big Data and Its Exclusions.” Stanford Law Review Online, 55–63. http://www.stanfordlawreview.org/sites/default/files/online/topics/66_stanlrevonline_55_lerman.pdf.

Chalabi “Why We Don’t Know The Size of the Transgender Population”
http://fivethirtyeight.com/features/why-we-dont-know-the-size-of-the-transgender-population/

Crawford, Kate, and Jason Schultz. 2013. “Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms.” Boston College Law Review 55 (1). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2325784&download=yes.

Citron, Danielle Keats. 2007. “Technical Due Process.” Washington University Law Review 86: 1249–1313.
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1012360.

Big Data Panel III Algorithms: Transparency, Accountability, Values and Discretion

Global Network Initiative Principles
http://globalnetworkinitiative.org/principles/index.php and Implementation Guidelines http://globalnetworkinitiative.org/implementationguidelines/index.php

Big Data Panel IV: Governance Roundtable

C. Raab and D. Wright, Surveillance: Extending the Limits of Privacy Impact Assessments, in Privacy Impact Assessment, (D. Wright and P. DeHert eds) Springer 2012. §§17.2.6 through end

DHS Privacy Threshold Analysis pp. 3-7 Department of Homeland Security Privacy Office. “Privacy Threshold Analysis (PTA).”
http://www.dhs.gov/sites/default/files/publications/privacy-dhs-pta-template-20140123.pdf. (Pages 5-19)

Series of Readings on “Do Not Track”
http://www.techpolicy.com/Blog/June-2012/If-you-choose-not-to-decide,-your-web-browser-will.aspx
http://blog.mozilla.org/privacy/2012/05/31/do-not-track-its-the-users-voice-that-matters/
http://www.w3.org/TR/tracking-dnt/
http://www.law360.com/articles/531445/do-not-track-group-finally-nails-down-tech-standard
http://www.law360.com/articles/531445/do-not-track-group-finally-nails-down-tech-standard:

Series of Readings on Tesla
http://www.teslamotors.com/blog/most-peculiar-test-drive

http://www.forbes.com/sites/kashmirhill/2013/02/19/the-big-privacy-takeaway-from-tesla-vs-the-new-york-times/https://www.privacyassociation.org/privacy_perspectives/post/the_strange_and_unmarked_road_ahead_for_privacyhttp://www.teslamotors.com/sites/default/files/pdfs/tmi_privacy_statement_external_6-14-2013_v2.pdf
http://www.teslamotors.com/sites/default/files/pdfs/tmi_privacy_statement_external_6-14-2013_v2.pdf

Feist v. Rural, 499 U.S. 340 (1991) [HTML]

Pro-CD v. Zeidenberg, 86 F.3d 1447 (1996) [HTML]

Executive Office of the President. 2014. "BIG DATA : SEIZING OPPORTUNITIES, PRESERVING VALUES."
http://www.whitehouse.gov/issues/technology/big-data-review.

Scaling Up! Really Big Data

Sicular, S. (2013, March 27). Gartner's big data definition consists of three parts, not to be confused with three "V"s. Forbes.

The four Vs of big data. IBM: The Big Data & Analytics Hub, Infographics & Animations. http://www.ibmbigdatahub.com/infographic/four--vs--big--data

Cloud computing. Wikipedia. http://en.wikipedia.org/wiki/Cloud_computing

fasp benchmarks. Aspera. http://asperasoft.com/resources/benchmarks/

Sharpe, J. (2014, April 22). Use InfoSphere streams as a sensory interface to Watson. IBM: developerWorks.

Silver, N., & McCann, A. (2014, May 29). How to tell someone's age when you know her name. FiveThirtyEightLife.

What is cloud? IBM Cloud.
http://www.ibm.com/cloud--computing/us/en/what--is--cloud-- computing.html

Arcangelim, A. (2008). Using Linux as a hypervisor with KVM. Qumranet Inc.

Schulz, G. (2011, December 5). Cloud, virtual and storage networking conversations part IV. Toolbox.com.

Hemmings, K. (2012, December 27). 3 types of cloud service models. appcore.

SoftLayer API: http://sldn.softlayer.com/article/SoftLayer--API--Overview

Eadline, D. (n.d.). Moving HPC to the cloud. Admin New Day HPC.

Getting started with SoftLayer, an IBM Company.
http://knowledgelayer.softlayer.com/gettingstarted/meet--softlayer
http://knowledgelayer.softlayer.com/gettingstarted/how--to
http://knowledgelayer.softlayer.com/gettingstarted/how--to/set--up--your--account

Bryce, J., Wilson, D., Fischer, C., & Aubuchon, G. (2013, November 8). OpenStack keynote featuring Concur, DigitalFilm Tree, Shutterstock (Video). OpenStack Summit. Summit conducted in Hong Kong.

Sabbah, D. (2013, November 8.) IBM keynote: Managing the next era of computing with an open cloud architecture (Video). OpenStack Summit. Summit conducted in Hong Kong.

DevStack—an OpenStack community production: http://devstack.org/

OpenStack documentation: https://wiki.openstack.org/wiki/Documentation/HowTo

OpenStack architecture design guide: http://docs.openstack.org/arch--design/content/

OpenStack command--line interface (CLI) reference: http://docs.openstack.org/cli--reference/content/

HDFS architecture: http://hadoop.apache.org/docs/current/hadoop--project--dist/hadoop--hdfs/HdfsDesign.html

Apache Hadoop Main 2.5.2 API: http://hadoop.apache.org/core/docs/current/api/

Schmuck, F., & Haskin, R. (2002, January). GPFS: A shared--disk file system for large computing clusters. In Proceedings of the FAST 2002 Conference on File and Storage Technologies.

Elastic storage (GPFS web link): http://www.ibm.com/systems/platformcomputing/products/gpfs/

Apache Hadoop: http://hadoop.apache.org/

Apache Pig: http://pig.apache.org/

IBM Platform Computing blog: https://www--304.ibm.com/connections/blogs/platformcomputing/?lang=en_us

Dignan, L. (2014, March 31). Cloudera raises $900 million, plots expansion. ZDNet.

White, T. (2012). Hadoop, the definitive guide. Yahoo Press.

Spark overview: http://spark.apache.org/docs/latest/

Hornbeck, R. L. (2013, February 18). Batch versus streaming: Differentiating between tactical and strategic big data analytics.

Spark programming guide (code examples):
http://spark.apache.org/docs/latest/programming--guide.html

Resilient distributed datasets (RDDs):
http://spark.apache.org/docs/latest/programming--guide.html#resilient--distributed--datasets--rdds

MLlib—Basic statistics: http://spark.apache.org/docs/latest/mllib--statistics.html

GraphX programming guide

Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2014). Learning Spark. O'Reilly Media.

Sliwa, C. (n.d.). Troubleshooting and identifying data storage performance bottlenecks. TechTarget.

Shafer, J. (2010). I/O virtualization bottlenecks in cloud computing today. In Proceedings of the 2nd Conference on I/O Virtualization.

Vajgel, P. (2009, April 30). Needle in a haystack: Efficient storage of billions of photos. Facebook.

Boudjnah, C. (2013, February 12). Ceph and Swift: Why we are not fighting. eNovance.

FASP. Apera. http://asperasoft.com/technology/transport/fasp/

Metz, C. (2012, August 8). If Xerox PARC invented the PC, Google invented the Internet. Wired.

DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., . . . Vogels, W. (n.d.). Dynamo: Amazon's highly available key--value store.

NoSQL. Wikipedia. http://en.wikipedia.org/wiki/NoSQL

Cloudant For Developers interactive tutorial
Reading and writing, primary index, secondary indexes, search indexes

Cloudant API Reference (skim)

Stoica, I., Morris, R., Karger, D., Kaashoek, M. F., & Balakrishnan, H. (2001, August). Chord: A scalable peer-to-peer lookup service for Internet applications. SIGCOMM'01, San Diego.

Gheawat, S., Gobioff, H., & Leung, S--T. (2003, October). The Google File System. SOSP'03, New York.

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. OSDI.

Chang, F., Dean, J., Ghemaat, S., Hsieh, W., Wallach, D. A., . . . Gruber, R. E. (2006, November). Bigtable: A distributed storage system for structured data. OSDI'06, Seattle.

Cloudant Online Training Presentations

Zaharia, M., Das, T., Li, H., Shenker, S., & Stoica, I. (n.d.). Discretized streams: An efficient and fault--tolerant model for stream processing on large clusters.

Stream computing in the cloud (IBM InfoSphere Streams):
https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw--infomgt&S_PKG=ov24587&S_TACT=109HF53W&S_CMP=is_bdebook8

Amazon Kinesis (Amazon Web Services Blog post about launch): Launching Kinesis

Amazon Kinesis: http://aws.amazon.com/kinesis/

Amazon Kinesis Service API Reference:
http://awsdocs.s3.amazonaws.com/kinesis/latest/kinesis--api.pdf

Goetz, P. T. (2014, August 11). Apache storm vs. Spark streaming (Video). Retrieved from
http://www.slideshare.net/ptgoetz/apache--storm--vs--spark--streaming?qid=dfbb7c09--6f87--40ca-- a69d--d76837efd236

Goetz, P. T. (2014, April 7). Hadoop Summit Europe 2014: Apache Storm architecture (Video). Retrieved from
http://www.slideshare.net/ptgoetz/storm--hadoop--summit2014

Zaharia, M., Das, T., Li, H., Shenker, S., & Stoica, I. (n.d.). Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters. Retrieved from http://tinyurl.com/dstreams

Spark streaming programming guide:
http://spark.incubator.apache.org/docs/latest/streaming--programming--guide.html

ETE 2012—Nathan Marz on Storm (Video). Retrieved from
https://www.youtube.com/watch?v=bdps8tE0gYo

Apache Spark tutorial: https://storm.incubator.apache.org/documentation/Tutorial.html

AMPLab Camp on Spark streaming: http://ampcamp.berkeley.edu/wp--content/uploads/2013/07/Spark--Streaming--AMPCamp--3.pptx

AWS CloudFormation: http://aws.amazon.com/cloudformation/

AWS OpsWorks: http://aws.amazon.com/opsworks/

AWS Elastic Beanstalk: http://aws.amazon.com/elasticbeanstalk/

IBM Bluemix: http://ibm.biz/HackBluemix

OpenStack Heat: https://wiki.openstack.org/wiki/Heat

CloudSoft Brooklyn walkthrough: http://brooklyncentral.github.io/start/walkthrough/index.html

GigaSpaces: Cloudify 3.0 getting started: http://getcloudify.org/guide/3.0/quickstart.html

Van de Geijn, R. A., & Watts, J. (n.d.). SUMMA: Scalable universal matrix multiplication algorithm.

Message Passing Interface (MPI and 2D Cartesian communicators): https://computing.llnl.gov/tutorials/mpi/

Introduction to parallel computing (HPC tutorial): https://computing.llnl.gov/tutorials/parallel_comp/

MPI tutorial: A comprehensive MPI tutorial resource: http://mpitutorial.com

Brin, S, & Page, L. (n.d.). The anatomy of a large--scale hypertextual web search engine.

Floyer, D. (2014, July 5). The growth and management of unstructured data. Wikibon.org.

Google: Crawling & indexing: http://www.google.com/intl/en/insidesearch/howsearchworks/crawling--indexing.html

Solr tutorial: http://lucene.apache.org/solr/4_10_0/tutorial.html

Nutch and Lucene framework (Presentation). Retrieved from http://tinyurl.com/k79hofu

Baeza-Yates, R., & Cambazoglu, B. B. (2014). Scalability and efficiency challenges in large-scale web search engines. Yahoo Labs: Tutorial at SIGIR 2014, Gold Coast, Australia.

Computational genomics. Wikipedia.

Human population genomics. IBM Research.

Langmead, B., Schatz, M. C., Lin, J., Pop, M., & Salzberg, S. L. (2009). Searching for SNPs with cloud computing. Genome Biology, 10, R134.

Wall, D. P., Kudtarkar, P., Fusaro, V. A., Pivovarov, R., Patil, P., & Tonellato, P. J. (2010, May). Cloud computing for comparative genomics. BMC Bioinformatics, 11, 259.

Ferrucci, D., Brown, E., Chu Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., . . . Welty, C. (2010, Fall). Building Watson: An overview of the DeepQA project. AI Magazine.

Shenoi, M. (2014, September 18). IBM Watson analytics—Powerful analytics for everyone (Blog).

WatsonPaths. IBM Research.

IBM Watson. (2014, August 27). Introducing IBM Watson Discovery Advisor. (Video).