NOTES FOR THE BOOK, *PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE* — REVISED AND UPDATED EDITION

This document provides citations and comments pertaining to the book's chapters.

Rather than clicking, many URLs below that span multiple lines must be copied to your browser, and extraneous spaces manually removed.

For more information about the book, see the book's website: www.thepredictionbook.com.

For this book's primary list of additional reading and other learning resources, see its "Hands-On Guide: Resources for Further Learning," located at the end of the book itself (also published online as an article: http://bit.ly/1OHGggC).

INTRODUCTION: The Prediction Effect

Note: All case examples in the Introduction chapter appear in this book's Central Tables of examples—see the Central Table Notes (http://www.PredictiveNotes.com) for their citations.

An overview article by the author about the field of predictive analytics:

- Eric Siegel, "4 Things Digital Data Collection Predicts for 2016," *BigThink,* January 27, 2016. http://bigthink.com/big-think-edge/4-predictions-about-2016-provided-by-digital-data-collection.

The author on the organizational process required to successfully make use of predictive analytics:

- Eric Siegel, "The Trick to Predictive Analytics: How to Bridge the Quant/Business Culture Gap," *Analytics Magazine*. http://www.predictiveanalyticsworld.com/patimes/trick-predictive-analytics-bridge-quantbusiness-culture-gap/8177/.

CHAPTER 1: Liftoff! Prediction Takes Action *(deployment)*

Most of John Elder's story comes from a direct interview. For an autobiographical essay by John Elder, see:

- Mohamed Medhat Gaber, *Journeys to Data Mining: Experiences from 15 Renowned Researchers* (Springer, 2012), 61–76.

Regarding terrified spouses of astronauts:

- Mike Mullane, *Riding Rockets: The Outrageous Tales of a Space Shuttle Astronaut* (Scribner, 2006).

$10 million Grand Challenge:

- The Defense Advanced Research Projects Agency (DARPA); est. 1958. See: www.darpa.mil/.

Driverless cars:

- Daniel Nasaw, "Driverless Cars and How They Would Change Motoring," *BBC News Magazine*, May 10, 2012. www.bbc.co.uk/news/magazine-18012812.

Quote from Stephen Dubner:

- Stephen J. Dubner, "Why Can't We Predict Earthquakes?" Freakonomics Radio, March 31, 2011. www.freakonomics.com/2011/03/31/why-cant-we-predict-earthquakes-full- transcript/.

Quote from Mehmet Oz:

- Dr. Mehmet Oz, "10 Questions for Dr. Oz," *Time Magazine Online*, Health. www.time.com/time/specials/packages/article/0,28804,1912201_1912244_1913478,00.html.

Advertisement optimization example project:

- "Case Study: How Predictive Analytics Generates $1 Million Increased Revenue," case study provided by Prediction Impact, Inc. (Note that in this project a positive response actually entailed an opt-in and click, rather than just a click as with most online advertisements.) www.predictiveanalyticsworld.com/casestudy.php.

To view Malcolm Gladwell's entire speech, "Choice, Happiness and Spaghetti Sauce," visit:

- Malcolm Gladwell, "Choice, Happiness and Spaghetti Sauce," TEDTalks Online. www.ted.com/talks/malcolm_gladwell_on_spaghetti_sauce.html. Video file: www.ted.com/talks, February 2006.

Davenport and Harris quote:

- Thomas Davenport and Jeanne Harris, *Competing on Analytics: The New Science of Winning* (Harvard Business School Press, 2007).

"Survey results have in fact shown that a tougher competitive environment is by far the strongest reason organizations adopt this technology":

- David White, "Predictive Analytics: The Right Tool for Tough Times," Aberdeen Group White Paper, February 2010. www.aberdeen.com/aberdeen-library/6287/RA-predictive-analytics-customer-retention.aspx and www.targusinfo.com/files/PDF/outside_research/PredictiveAnalyticsReport.pdf.

John Elder includes "temporal leak" in his list of top 10 data miner mistakes—see:

- Robert Nisbit, John Elder, and Gary Miner, *Handbook of Statistical Analysis and Data Mining Applications* (Academic Press, 2009), chap. 20.

John Elder's PhD thesis:

- John Elder, PhD, "Efficient Optimization through Response Surface Modeling: GROPE Algorithm." Dissertation, School of Engineering and Applied Sciences, University of Virginia (1993).

The business case for PA:

- Patrick Turgeon, "Predictive Analytics—A Case for Early Adoption," Figurs*, February 6, 2012.

www.figurs.ca/pdf/Predictive_Analytics-A_case_for_early_adoption.pdf.

CHAPTER 2: With Power Comes Responsibility: Hewlett-Packard, Target, the Cops, and the NSA Deduce Your Secrets (*ethics*)

Some content of this chapter originated from valuable interviews with Gitali Halder, Anindya Dey, and Alex Beaux of Hewlett-Packard.

Ethical issues of marketing analytics; quote from Gartner's Andrew Frank:

- James Henderson, Insight: Why Marketing Analytics is Big Data's "under-hyped" topic," *Computerworld*, February 16, 2015. www.computerworld.co.nz/article/566365/insight-why-marketing-analysts-big-data-under-hyped-topic/.

Hewlett-Packard "Flight Risk" score:

- Gitali Halder, Hewlett-Packard, and Anindya Dey, Hewlett Packard, "Attrition Driver Analysis," Predictive Analytics World London Conference, November 30, 2011, London, UK. www.predictiveanalyticsworld.com/london/2011/agenda.php#day1–5a.

Overview of HR analytics industry practices:

- Don Peck, "They're Watching You at Work," *The Atlantic*, December, 2013 Issue. www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.

Best practices on managing privacy sensitivity when applying human resources analytics:

- Chantrelle Nielsen, "Collect Your Employees' Data Without Invading Their Privacy," *Harvard Business Review*,

September 24, 2014. https://hbr.org/2014/09/collect-your-employees-data-without-invading-their-privacy/.

Regarding ethical issues in data-driven human resource decisions:

- Don Peck, "They're Watching You at Work," *The Atlantic*, December, 2013 Issue. www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.

Refuting the claim Target's pregnancy prediction revealed a teen pregnancy:

- Gregory Piatetsky, KDNuggets, "Did Target Really Predict a Teen's Pregnancy? The Inside Story," *Predictive Analytic Times*, May 9, 2014 (originally published *KDNuggets*). www.predictiveanalyticsworld.com/patimes/target-really-predict-teens-pregnancy-inside-story/.

Regarding Target's use of PA to predict customer pregnancy:

- "Target Knew Teen Was Pregnant Before Her Dad." www.predictiveanalyticsworld.com/target-on-fox. Original broadcast, Fox News, February 24, 2012.
- Andrew Pole, Target, "How Target Gets the Most out of Its Guest Data to Improve Marketing ROI," Predictive Analytics World Washington, DC, Conference, October 19, 2010, Washington, DC. www.predictiveanalyticsworld.com/Target.
- Charles Duhigg, "How Companies Learn Your Secrets," *New York Times Magazine*, February 16, 2012. www.nytimes.com/2012/02/19/magazine/shopping-habits.html.
- Rachel Nolan, "Behind the Cover Story: How Much Does Target Know?" *New York Times*, February 21, 2012. http://6thfloor.blogs.nytimes.com/2012/02/21/behind-the- cover-story-how-much-does-target-know/.
- "Will Big Data and Big Money Mean Big Trouble?" *To the Point*, KCRW Los Angeles Public Radio 89.9, Los Angeles, CA. Host Warren Olney, April 2, 2012. www.kcrw.com/news/programs/tp/tp120402will_big_data_and_bi.
- Steven Cherry, "Your Favorite Stores Know You All Too Well." IEEE Spectrum's Techwise Conversations Online, March 30,

2012. http://spectrum.ieee.org/podcast/ computing/software/your-favorite-stores-know-you-all-too-well.

- "Target's Deep Customer Data Mining Raises Eyebrows," Martin Moylan, Minnesota Public Radio, MPR News, St. Paul, Minnesota, March 7, 2012. http://minnesota.publicradio.org/display/web/2012/03/07/target-data-mining-privacy/.
- "Targeted Marketing in Your Womb, Digital Product Placement, Pervasive Gaming," *Download This Show*, Australian Broadcasting Corporation's RadioNational, Sydney, Australia, February 26, 2012. www.abc.net.au/radionational/programs/downloadthisshow/ep5/3851666.
- "The Word—Surrender to a Buyer Power," *Colbert Report*, February 22, 2012.www.colbertnation.com/the-colbert-report-videos/408981/february-22–2012/the-word---surrender-to-a-buyer-power.

The value of a user's data:

- Alexis Madrigal, "How Much Is Your Data Worth? Mmm, Somewhere Between Half a Cent and $1,200," *The Atlantic*, March 19, 2012. www.theatlantic.com/technology/archive/2012/03/how-much-is-your-data-worth-mmm-somewhere-between-half-a-cent- and-1-200/254730/.

Overview of data aggregation and selling; data privacy issues:

- Adam Tanner, *What Stays in Vegas: The World of Personal Data—Lifeblood of Big Business—and the End of Privacy as We Know It"* (PublicAffairs, 2014).

Quote from David Sobel, privacy advocate:

- David Sobel, Senior Counsel at Electronic Frontier Foundation, https://www.eff.org/about/staff/david-sobel.

One in four Facebook users enters false data (i.e., lie):

- Michael Crider, "1 in 4 Facebook Users Lies over Privacy Concerns," *Slashgear Magazine Online*, May 7, 2012. www.slashgear.com/1-in-4-facebook-users-lies-over-privacy-concerns-07226506/.

Quote from Alexander Furnas ("... content is wholly subsidized by targeted..."):

- Alexander Furnas, "It's Not All about You: What Privacy Advocates Don't Get about Data Tracking on the Web." *The*

Atlantic, March 15, 2012. www.theatlantic.com/technology/archive/2012/03/its-not-all-about-you-what-privacy-advocates-dont-get-about-data-tracking-on-the-web/254533/.

Banks and lenders collect your social media activity as information to score your level of creditworthiness:

- Ken Lin, "What Banks and Lenders Know about You from Social Media," Mashable Social Media, October 7, 2011. http://mashable.com/2011/10/07/social-media-privacy-banks/.

Quote from Professor Tom M. Mitchell, Carnegie Mellon University:

- Tom M. Mitchell, "Mining Our Reality." *Science* 326, no. 5960 (December 18, 2009), 1644–1645. www.sciencemag.org/content/326/5960/1644.summary and www.cs.cmu.edu/Btom/pubs/Science2009_perspective.pdf.

Data mining does not drill down:

- Letter from the Executive Committee on ACM Special Interest Group on Knowledge, Discover and Data Mining (ACM SIGKDD), "Data Mining Is NOT Against Civil Liberties," June 30, 2003, revised July 28, 2003. www.sigkdd.org/civil-liberties.pdf.

Civil rights and big data:

- Alistair Croll, "Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It," O'Reilly Radar Online, August 2, 2012. http://radar.oreilly.com/2012/08/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it.html.

HP as one of the largest employers:

- "Top Companies: Biggest Employers," *CNNMoney*, July 25, 2011. http://money.cnn.com/magazines/fortune/global500/2011/performers/companies/biggest/.

"I'm surprised" job interview cartoon:

- The cartoon in Chapter 2 originally appeared in *Computerworld* magazine. Copyright 2003 John Klossner, www.jklossner.com.

Crime prediction by police:

- Colleen McCue, PhD, *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis* (Butterworth-Heinemann, 2007).
- Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith and John S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (Rand Corporation, 2013).
- Melissae Fellet, "Cops on the Trail of Crimes That Haven't Happened," *New Scientist*, October 12, 2011. www.newscientist.com/article/mg21128333.400-cops-on-the-trail-of-crimes-that-havent-happened.html.
- Lev Grossman, Cleo Brock-Abraham, Nick Carbone, Eric Dodds, Jeffrey Kluger, Alice Park, Nate Rawlings, Claire Suddath, Feifei Sun, Mark Thompson, Bryan Walsh, and Kayla Webley, "The 50 Best Inventions," *Time*, Technology, November 28, 2011. www.time.com/time/magazine/article/0,9171,2099708–1,00.html.
- Jessica M. Pasko, "*Time* Magazine Names Santa Cruz Predictive Policing Program One of the Year's Top Inventions," *Santa Cruz Sentinel*, November 23, 2011. www.santacruzsentinel.com/ci_19400300?source5pkg.
- Erica Goode, "Sending the Police Before There's a Crime." *New York Times on the Web*, August 15, 2011. www.nytimes.com/2011/08/16/us/16police.html.
- Stephen Hollifield, Information Services Manager, Richmond Police Department, "Curbing Crime with Predictive Analytics," Predictive Analytics World for Government, Washington, DC, September 12, 2011. www.predictiveanalyticsworld.com/gov/2011/agenda.php#day1–1115–1145.
- Carrie Kahn, "At LAPD, Predicting Crimes Before They Happen," National Public Radio Online, August 14, 2011. www.npr.org/2011/11/26/142758000/at-lapd-predicting-crimes-before-they-happen.
- David Smiley, "Not science fiction: Miami wants to predict when and where crime will occur," *Miami Herald*, April 22, 2015. www.miamiherald.com/news/local/community/miami-dade/article19256145.html.
- Sergeant Christopher Fulcher, Vineland, NJ, Police Department, "To Predict and Serve: Predictive Intelligence Analysis," Part I, July 5, 2011, http://policeledintelligence.com/2011/07/05/to-predict-and-serve-predictive-intelligence-analysis-part-i/. Part II,

July 6, 2011, http://policeledintelligence.com/2011/07/06/to-predict-and-serve-predictive-intelligence-analysis-part-ii/.

- "Burlington Police Dept. employs cutting-edge policing technique," *MyFOXBoston,* FOX 25, February 18, 2013. www.myfoxboston.com/story/21237083/2013/02/18/burlington-police-dept-employs-breaking-technique-in-policing.
- G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self- Exciting Point Process Modeling of Crime," *Journal of the American Statistical Association* 106, issue 493 (2011). www.math.ucla.edu/Bmbshort/papers/crime3.pdf.

Predictive policing that is closer to "pre-crime" – targets individuals more likely to commit future crimes (beyond the predictive patrolling of neighborhoods) - front page New York Times article:

- John Eligon and Timothy Williams, "Police Program Aims to Pinpoint Those Most Likely to Commit Crimes," *The New York Times*, September 24, 2015. www.nytimes.com/2015/09/25/us/police-program-aims-to-pinpoint-those-most-likely-to-commit-crimes.html?_r=0.

The pros and cons of predictive policing - Front page New York Times article:

- John Eligon and Timothy Williams, "Police Program Aims to Pinpoint Those Most Likely to Commit Crimes," *The New York Times*, September 24, 2015. www.nytimes.com/2015/09/25/us/police-program-aims-to-pinpoint-those-most-likely-to-commit-crimes.html?_r=0.

LA crime prediction:

- Nate Berg, "Predicting Crime, LAPD-style, *The Guardian*, June 25, 2014. www.predictiveanalyticsworld.com/patimes/predicting-crime-lapd-style/.

NYC crime prediction; says "... already being used in Philadelphia, New Castle, Del. and Lincoln, Neb.":

- Laura Nahmias and Miranda Neubauer, "NYPD testing crime-forecast software," *Capital New York*, July 8, 2015.

www.capitalnewyork.com/article/city-hall/2015/07/8571608/nypd-testing-crime-forecast-software.

- NYC Bulletin, "Developing the NYPD's Information Technology," modified June 25, 2015. www.nyc.gov/html/nypd/html/home/POA/pdf/Technology.pdf.

On the ethics of crime prediction:

- Andrew Guthrie Ferguson, "Predicting Predictive Policing in NYC," *The Huffington Post*, July 8, 2015. www.huffingtonpost.com/andrew-guthrie-ferguson/predicting-predictive-pol_b_7757200.html.

Overview of predictive policing:

- Walter L. Perry, Brian McInnis, Carter C. Price, Susan Smith and John S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*, RAND Corporation, RR-233-NIJ, 2013. As of April 17, 2015: www.rand.org/pubs/research_reports/RR233.html.

Philadelphia and Baltimore predict which parolee will commit murder:

- Jennifer Brown and Karen E. Crummy, "Technology, quick-reaction programs guiding parole reform in other states," *The Denver Post*, September 23, 2013. www.denverpost.com/parole/ci_24152942/technology-quick-reaction-programs-guiding-reform-other-states.

Cheating sumo wrestlers:

- Stephen J. Dubner, "Sumo: More of the Same," Freakonomics Radio, February 2, 2011. www.freakonomics.com/2011/02/02/sumo-more-of-the-same/.

Students cheating:

- Stephen J. Dubner, "Fish Gotta Swim, Teachers Gotta Cheat?" Freakonomics Radio, January 26, 2010. www.freakonomics.com/2010/01/26/fish-gotta-swim-teachers-gotta- cheat/.

Twitter bombs:

- Eni Mustafaraj and Panagiotis Metaxas, "From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search," in *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 26–27, 2010, Raleigh, NC*. http://journal.webscience.org/317/.

On the cost of fraud:

- Progressive Insurance Special Investigative Unit Report from the National Insurance Crime Bureau (NICB). www.progressiveagent.com/claims/special-investigations-unit.aspx.

Federal Trade Commission on fraud report incidents:

- Michael B. Sauter, "America's Worst States for Fraud," 24/7 Wall St., April 11, 2012. http://247wallst.com/2012/04/11/americas-worst-states-for-fraud/.

Estimated that nation's banks experience over $10 billion per year in attempted check fraud:

- Jay Zhou, Data Miners, LLC, "Building In-Database Predictive Scoring Model: Check Fraud Detection Case Study," Predictive Analytics World Washington, DC, Conference, October 20, 2009, Washington, DC. www.predictiveanalyticsworld.com/dc/2009/agenda.php#day1–19.

Aggregate fraud loss in the United States sees estimates from $100 billion to $1 trillion:

- Robert Nisbit, John Elder, and Gary Miner, *Handbook of Statistical Analysis and Data Mining Applications* (Academic Press, 2009), 347.

PayPal fraud:

- "The New World of Massive Data Mining," *Diane Rehm Show*, WAMU 88.5, Washington, DC, April 2, 2012. http://thedianerehmshow.org/shows/2012–04–02/ new-world-massive-data-mining.

1–800-FLOWERS fraud detection:

- Auren Hoffman, CEO, Rapleaf, "Leveraging Social Media Data to Manage Fraud Risk," Predictive Analytics World San Francisco Conference, February 16, 2010, San Francisco, CA.

www.predictiveanalyticsworld.com/sanfrancisco/2010/agenda.php#day1–10.

Mexican Tax Administration fraud detection:

- Luis BeltrÆn Farías, Director, Risk Models Office, Mexican Tax Administration, Ministry of Finance, Mexico, "A Model for Fraud Prevention in VAT Refunds in Mexico," Predictive Analytics World for Government Washington, DC, Conference, September 12, 2011, Washington, DC. www.predictiveanalyticsworld.com/gov/2011/agenda.php#day2–315–335.

U.S. Defense Finance and Accounting Service fraud detection:

- Dean Abbott, Abbott Analytics, "Case Study: Defense Finance and Accounting Service— Invoice Fraud Detection." Thanks to Dean Abbott, Abbott Analytics, www.abbottanalytics.com (2004–2012), for this case study. www.abbottanalytics.com/data-mining-case-study-1.php.
- Dean Abbott, Haleh Vafaie, PhD, Mark Hutchins, and David Riney, "Improper Payment Detection in Department of Defense Financial Transactions," Federal Data Mining Symposium, Washington, DC, March 29, 2000. www.abbottanalytics.com/assets/pdf/ Abbott-Analytics-Improper-Payment-Detection-2000.pdf.

Elder Research fraud modeling project for the IRS (thanks to John Elder for pointing out the fraud scheme as an example detectable only by way of analyzing social data):

- Mohamed Medhat Gaber, *Journeys to Data Mining: Experiences from 15 Renowned Researchers* (Springer, 2012), 61–76.

Regarding cyber-security analytics:

- Seth Robertson, Eric V. Siegel, Matt Miller, and Salvatore J. Stolfo, "Surveillance Detection in High Bandwidth Environments," Third DARPA Information Survivability Conference and Exposition (DISCESSIII), Washington, DC, April 2003. http://secgate-b.hacktory.cs.columbia.edu/sites/default/files/SD-DiscexIII.pdf.

Source for leading research in PA for detection of online attacks and cybercrime:

- *Security Informatics Journal*'s most viewed articles: http://www.security-informatics.com/mostviewed

Spam filter subversion:

- Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles Sutton, J. D. Tygar, and Kai Xia, "Exploiting Machine Learning to Subvert Your Spam Filter," University of California, Berkeley, April 4, 2008. http://static.usenix.org/event/leet08/tech/full_papers/nelson/nelson_html/.

Computers detect computer-generated text:

- Allen Lavoie and Mukkai Krishnamoorthy, "Algorithmic Detection of Computer Generated Text," Cornell University Library, August 4, 2010. http://arxiv.org/abs/1008.0706.

Cheating in online chess tournaments:

- Garry Kasparov, "The Chess Master and the Computer," *New York Review of Books on the Web*, February 11, 2010. www.nybooks.com/articles/archives/2010/feb/11/the-chess-master-and-the-computer/.

This chapter's primary source regarding recidivism prediction (including the quotes from Ellen Kurtz):

- Nadya Labi, "Misfortune Teller," *The Atlantic*, January/February 2012. www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/8846/.

Oregon's online crime-prediction tool:

- "The Public Safety Checklist for Oregon," Criminal Justice Commission, last updated August 11, 2012. https://risktool.ocjc.state.or.us/psc/.

Hungry judges rule negatively:

- Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso, "Extraneous Factors in Judicial Decisions," edited by Daniel Kahneman, Princeton University, Princeton, NJ, February 25, 2011. http://lsolum.typepad.com/files/danziger-levav-avnaim-pnas-2011.pdf and www.pnas.org/content/108/17/6889.

Guilty of a crime you didn't commit (yet):

- Ben Goldacre, "It's Not So Easy to Predict Murder—The Maths," *The Guardian Online*, December 8, 2006. www.guardian.co.uk/science/2006/dec/09/badscience.uknews.

A law professor's op-ed protesting states' adoption of recidivism prediction:

- Sonja B. Starr, "Sentencing, by the Numbers," *The New York Times*, August 10, 2014. www.nytimes.com/2014/08/11/opinion/sentencing-by-the-numbers.html?_r=2.

Former Attorney General Eric Holder opposes data driven sentencing:

- Massimo Calabresi, "Exclusive: Attorney General Eric Holder to Oppose Data-Driven Sentencing," *TIME*, Vol. 184, No. 4, July 31, 2014. www.time.com/3061893/holder-to-oppose-data-driven-sentencing/.

More on recidivism prediction:

- Letter from the Office of the Assistant Attorney General, U.S. Department of Justice, Criminal Division, Washington, DC, to The Honorable Patti B. Saris, Chair, United States Sentencing Commission, dated July 20, 2014. www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annual-letter-final-072814.pdf.

Crime data can be flawed - this article examines issues with FBI data:

- Ben Poston, "FBI crime-reporting audits are shallow, infrequent," *Milwaukee-Wisconsin Journal Sentinel*, August 18, 2012. www.jsonline.com/watchdog/watchdogreports/fbi-crimereporting-audits-are-shallow-infrequent-cg5uvel-166665516.html.

Predicting homicide recidivism:

- Rolf Loeber, Dustin Pardini, D. Lynn Homish, Evelyn H. Wei, Anne M. Crawford, David P. Farrington, Magda Stouthamer-Loeber, Judith Creemers, Steven A. Koehler, and Richard Rosenfeld, "The Prediction of Violence and Homicide in Young Men," *Journal of Consulting and Clinical Psychology* 73, no. 6 (2005): 1074–1088.

www.wpic.pitt.edu/research/famhist/PDF_Articles/APA/H%201.pdf.

- Melanie-Angela Neuilly, Kristen M. Zgoba, George E. Tita, and Stephen S. Lee, "Predicting Recidivism in Homicide Offenders Using Classification Tree Analysis," *Sage Journals, Homicide Studies* 15, no. 2 (May 2011): 154–176. http://hsx.sagepub.com/content/15/2/154.abstract.

Predicting and preventing homicide:

- Rosanna Louise Buttrick, "Homicide: Prediction and Prevention," University of Nottingham, School of Psychology, June 2009. www.psychology.nottingham.ac.uk/staff/ddc/c8cxpa/further/Dissertation_examples/Buttrick_09.pdf.

Regarding UK bank search for terrorists:

- Steven D. Levitt and Stephen J. Dubner, *Superfreakonomics: Global Cooling, Patriotic Prostitutes, and Why Suicide Bombers Should Buy Life Insurance* (William Morrow, 2011).
- Stephen J. Dubner, "SuperFreakonomics Book Club: Ian Horsley Answers Your Questions about the Terrorist Algorithm," September 1, 2010. www.freakonomics.com/2010/09/01/superfreakonomics-book-club-ian-horsley-answers-your-questions-about-the-terrorist- algorithm/.

For more information on crime-detecting models (e.g., biometric models) to identify terror suspects:

- Christopher Westphal, *Data Mining for Intelligence, Fraud & Criminal Detection* (CRC Press, 2009).

Government data collection to fight discrimination:

The Department of Housing and Urban Development's Affirmatively Furthering Fair Housing database:

- Emily Badger, "Obama administration to unveil major new rules targeting segregation across U.S.," *The Washington Post*, [Wonkblog]. www.washingtonpost.com/blogs/wonkblog/wp/2015/07/08/obama-administration-to-unveil-major-new-rules-targeting-segregation-across-u-s/.
- U.S. Department of Housing and Urban Development, Bulletin HUD No. 15-084**, "**HUD Announces Final Rule on

Affirmatively Furthering Fair Housing," July 8, 2015. http://portal.hud.gov/hudportal/HUD?src=/press/press_releases_media_advisories/2015/HUDNo_15-084.

The Federal Housing Financing Agency:

- The Federal Housing Finance Agency's National Mortgage Database Project. www.fhfa.gov/PolicyProgramsResearch/Programs/Pages/National-Mortgage-Database.aspx.

The Consumer Financial Protection Bureau:

- Richard Cordray, "Prepared Remarks of CFPB Director Richard Cordray at the Auto Finance Field Hearing," The Consumer Financial Protection Bureau, September 18, 2014. www.consumerfinance.gov/newsroom/prepared-remarks-of-cfpb-director-richard-cordray-at-the-auto-finance-field-hearing/.

Police misconduct:

- NPR Staff, NPR Digital Media, "Can Big Data Help Head Off Police Misconduct?" *All Things Considered*, July 19, 2016. https://www.wbez.org/shows/all-things-considered/can-big-data-help-head-off-police-misconduct/eadc6dea-50bb-4295-8d03-b55b75e443a2.

The U.S. Department of Education:

- Department of Education Civil Rights Data Collection (CRDC), [Updated] 2013-14 and 2015-16 CRDC - Updated 4/9/2015. www2.ed.gov/about/offices/list/ocr/data.html?src=rt/.
- Department of Education Press Release, "Expansive Survey of America's Public Schools Reveals Troubling Racial Disparities," March 21, 2014. www.ed.gov/news/press-releases/expansive-survey-americas-public-schools-reveals-troubling-racial-disparities.

Concerns in the United Kingdom regarding the discriminatory effect of predictive models:

- Christine Evans-Pughe, "Forecasting Human Behavior Carries Big Risks," *The Guardian*, July 18, 2007. www.guardian.co.uk/technology/2007/jul/19/guardianweeklytechnologysection.it.

Regarding the ethics of criminal prediction:

- Michael Lotti, "Ethics and the Information Age," *Effect Magazine Online*, Winter 2009/2010.

www.larsonallen.com/EFFECT/Ethics_and_the_Information_Age.aspx.

Regarding inadvertent discrimination that can result from analytics (referred to as "big data" in this article):

- Eileen Sullivan, "White House says big data could be used to discriminate against Americans," *PBS NewsHour*, April 26, 2014. www.pbs.org/newshour/rundown/white-house-says-big-data-used-discriminate-americans/.
- Alysa Zeltzer Hutnik, "FTC Highlights Alternative Scoring Products in Privacy Seminar," *AD Law Access,* March 26, 2014. www.adlawaccess.com/2014/03/articles/ftc-highlights-alternative-scoring-products-in-privacy-seminar/.

Chicago police say its computers can tell who will be a violent criminal, but critics say it's nothing more than racial profiling:

- Matt Stroud, "The minority report: Chicago's new police computer predicts crimes, but is it racist?" *The Verge*, February 19, 2104. www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist.

Racial profiling in the United States:

- ACLU Report, "The Persistence of Racial and Ethnic Profiling in the United States: A Follow-up Report To The U.N. Committee on the Elimination of Racial Discrimination," (August, 2009). www.aclu.org/report/persistence-racial-and-ethnic-profiling-united-states.

Quote from Alexander Furnas ("...we grant private entities...greater powers of persuasion than anyone has ever had..."):

- Alexander Furnas, "It's Not All about You: What Privacy Advocates Don't Get about Data Tracking on the Web," *The Atlantic*, March 15, 2012. www.theatlantic.com/technology/archive/2012/03/its-not-all-about-you-what-privacy-eadvocates-dont-get- about-data-tracking-on-the-web/254533/.

Regarding the spotlight on Target's pregnancy PA application:

- KDnuggets poll: “Was Target Wrong in Using Analytics to Find Pregnant Women?” KDnuggets, February 2012 online poll. www.kdnuggets.com/2012/02/new-poll-target- analytics-wrong-to-find-pregnant-women.html.

Eric Schmidt on facial recognition and privacy:

- Bianca Bosker, “Facial Recognition: The One Technology Google Is Holding Back,” *Huffington Post*, June 1, 2011. www.huffingtonpost.com/2011/06/01/facial-recognition-google_n_869583.html.

On the related topic of predicting race, age, and sexual orientation from Facebook likes, gender by way of Twitter followees, and "the chances of an employee suffering a nervous breakdown, selling or leaking of company secrets or predict any type of 'inappropriate behaviour' - including office affairs":

- Michal Kosinski, David Stillwell and Thore Graepel, “Private traits and attributes are predictable from digital records of human behavior,” *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, Vol. 110, No. 15, April 9, 2013. www.pnas.org/content/110/15/5802.abstract.
- University of Cambridge Research, "Online records could expose intimate details and personality traits of millions," University of Cambridge, March 11, 2013. www.cam.ac.uk/research/news/digital-records-could-expose-intimate-details-and-personality-traits-of-millions.
- Puneet Singh Ludu, “Inferring gender of a Twitter user using celebrities it follows,” Cornell University Library, arXiv:1405.6667 [cs.IR] or arXiv:1405.6667v1 [cs.IR], May 26, 2014. http://arxiv.org/abs/1405.6667.
- "Minority office report: Warning over software now being used by bosses that predicts if you're going to steal from the firm, have a nervous breakdown and even have an office affair." http://www.dailymail.co.uk/news/article-2956766/Minority-office-report-Warning-software-used-bosses-predicts-going-steal-firm-nervous-breakdown-office-affair.html.
- Justin Tenuto, “Using Machine Learning to Predict Gender,” KDNuggets Post, November, 2015. www.kdnuggets.com/2015/11/machine-learning-predict-gender.html.

Even Chuck Norris protests government data collection, as the IRS collects taxpayer data from social media:

- Chuck Norris, "IRS Now Robo-audits Your Spending," *Ammoland*, April 9, 2013. www.ammoland.com/2013/04/irs-now-robo-audits-your-spending/.

Should Data Science Become a Profession and Self-regulate?:

- KDNuggets Webcast/Debate: "Should Data Science Become a Profession and Self-regulate?" KDNuggets, April 10, 2013. www.kdnuggets.com/2013/04/webcast-debate-should-data-science-become-a-profession-self-regulate.html.

Special Sidebar on Automatic Suspect Discovery:

For context, here is the complete albeit limited extent of my own prior direct involvement in government intelligence work. Besides many miscellaneous interactions with government and industry workers as a consultant and the founder of the Predictive Analytics World conference series, which includes the annual PAW Government event, I worked directly on only one project in government intelligence. While at a startup company around 2002/2003, I served as the project supervisor for a CIA-funded project to develop a prototype system that automatically groups email addresses into "cliques," according to how often they email one another, and then characterizes the topic areas discussed in each group's communications by way of frequent keywords in email subject lines. The pilot system was tested only on the startup company's own internal email system, and all employees were notified of such. The groups and associated email conversation topics automatically inferred and output by the system were intuitively compelling, e.g., the holiday party organizers were successfully designated as such. Beyond that project, most of the startup company's work applied analytics to network intrusion detection systems (i.e., to detect hackers and viruses). The email-grouping system was part of the Email Mining Toolkit described in the following research papers (I was erroneously left off the publications' author lists despite my involvement, writing contributions, and, specifically, that I designed and drew Figure 1 in them both, and Figure 2 in the latter):

- Salvatore J. Stolfo, Shlomo Hershkop, Chia-Wei Hu, Wei-Jen Li, Olivier Nimeskern, and Ke Wang, *Behavior-based Modeling and its Application to Email Analysis*. Columbia University modified September 4, 2011. http://ids.cs.columbia.edu/sites/default/files/TOIT-EMT.pdf.
- Salvatore J. Stolfo, Shlomo Hershkop, Ke Wang, Wei-Jen Li, Olivier Nimeskern, and Chia-Wei Hu, *Behavior Profiling of Email*. Columbia University modified September 4, 2011. http://ids.cs.columbia.edu/sites/default/files/nsf-nij-emt.pdf.

The author's Newsweek op-ed on NSA bulk data collection:

- Eric Siegel, "A Rogue Liberal: Halting NSA Bulk Data Collection Compromises Intelligence," *Newsweek*, November 28, 2015. www.predictiveanalyticsworld.com/patimes/a-rogue-liberal-halting-nsa-bulk-data-collection-compromises-intelligence/6882/.

The NSA's Utah Data Center is the U.S.'s largest data center for surveillance:

- James Bamford, "The NSA is Building the Country's Biggest Spy Center (Watch What You Say)," Wired Online, March 15, 2012. www.wired.com/2012/03/ff_nsadatacenter/.

The NSA is the world's largest employer of Ph.D. mathematicians:

- Matt Bedan, *"Echelon's Effect: The Obsolescence of the U.S. Foreign Intelligence Legal Regime,"* Federal Communications Law Journal, Vol. 59, Iss.2, Article 7, 425 (2007). www.wanttoknow.info/articles/echelon_ukusa_unconstitutional.pdf.

Mission and Executive Order for the NSA:

- National Security Agency Central Security Service Mission Statement. https://www.nsa.gov/about/mission/index.shtml.
- National Archives, Federal Register, Executive Order 12333—United States intelligence activities. The provisions of Executive Order 12333 of Dec. 4, 1981, appear at 46 FR 59941, 3 CFR, 1981 Comp., p. 200, unless otherwise noted. www.archives.gov/federal-register/codification/executive-order/12333.html.

Two primary resources as backstory on the Edward Snowden disclosures:

- Glenn Greenwald, *No Place To Hide: Edward Snowdon, the NSA and the U.S. Surveillance State* (Metropolitan Books, 2014).
- Luke Harding, *The Snowden Files: The Inside Story of the World's Most Wanted Man* (Vintage Books, 2014).

General treatises on privacy in the age of the Internet and data collection by both corporations and government agencies:

- Robert Scheer, *They Know Everything About You: How Data-Colling Corporations and Snooping Government Agencies are Destroying Democracy* (Nations Books, 2015).
- Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (President and Fellows of Harvard College, 2015).
- Adam Tanner, *What Stays in Vegas: The World of Personal Data—Lifeblood of Big Business—and the End of Privacy as We Know It* (PublicAffairs, 2014).

The move towards increased government data collection and the debate around it pre-dated Snowden:

- Alex Howard, "Predictive analytics and data sharing raise civil liberties concerns," O'Reilly Radar Online, April 11, 2013. http://radar.oreilly.com/2013/04/predictive-big-data-analytics-privacy.html.

Pattern that predicts an attack by Lashkar-e-Taiba:

- Neal Ungerleider, "A Computer Program That Predicts Terrorist Attacks," Co.EXIST. www.fastcoexist.com/1680540/a-computer-program-that-predicts-terrorist-attacks.
- A. Mannes, J. Shakarian, A. Sliva, and V. S. Subrahmanian, "A Computationally-Enabled Analysis of Lashkar-e-Taiba Attacks in Jammu & Kashmir," Laboratory for Computa- tional Cultural Dynamics, July 1, 2011. https://lccd-content.umiacs.umd.edu/main/ papers/let_eisic_camera.pdf.

NSA bulk data collection was, in fact, publicly revealed years before Snowden, in 2006:

- Leslie Cauley, "NSA has massive database of American phone calls," *USA TODAY,* May 11, 2006. http://usatoday30.usatoday.com/news/washington/2006-05-10-nsa_x.htm.

Support for the presumption that the NSA has worked with PA and will continue to do so:

NSA Data Scientist job posting on LinkedIn:

- "The relevant experience must involve data mining, informatics, data science, programming, computational algorithms, information retrieval (i.e., organizing and structuring data), statistical analysis, machine learning, artificial intelligence, software engineering, and/or systems design and analysis."
  www.linkedin.com/jobs2/view/18725605
- NSA Job Description, Data Scientist, Fort George G. Meade, Maryland, Occupational Group: 1550, Computer Science.
  https://www.nsa.gov/psp/applyonline/EMPLOYEE/HRMS/c/HRS_HRAM.HRS_CE.GBL?Page=HRS_CE_JOB_DTL&Action=A&JobOpeningId=1049497&SiteId=1&PostingSeq=1.

Technologies invented by the NSA include capabilities for "data mining" and to "discover patterns":

- National Security Agency 2014 Technology Catalog, NSA Technology Transfer Program (TTP), V2.0, July 2014.
  www.nsa.gov/research/_files/tech_transfers/nsa_technology_transfer_program.pdf

Overview of NSA big data capabilities includes machine learning:

- Michael Hickins, "How the NSA Could Get So Smart So Fast. Modern Computing Is Helping Companies and Governments Accurately Parse Vast Amounts of Data in a Matter of Minutes," *The Wall Street Journal*, June 12, 2013.
  www.wsj.com/articles/SB10001424127887324049504578541271020665666
- Kurt Marko, "The NSA and Big Data: What IT Can Learn," *InformationWeek,* July 1, 2013.
  http://reports.informationweek.com/abstract/81/11055/Business-Intelligence-and-Information-Management/the-nsa-and-big-data-what-it-can-learn-.html.

Total Information Awareness, created in part to "find patterns" and apply data mining, is now nominally defunct, but its legacy continues at the NSA (predictive modeling is also specifically mentioned in the third citation):

- John Markoff, "THREATS AND RESPONSES: INTELLIGENCE; Pentagon Plans a Computer System That Would Peek at Personal Data of Americans," *New York Times,* November 9, 2002. www.nytimes.com/2002/11/09/us/threats-responses-intelligence-pentagon-plans-computer-system-that-would-peek.html.
- Shane Harris, "Giving In to the Surveillance State," *The New York Times,* August 22, 2012. www.nytimes.com/2012/08/23/opinion/whos-watching-the-nsa-watchers.html?_r=0.
- Defense Advanced Research Projects Agency's Information Awareness Office Report to Congress regarding the Terrorism Information Awareness Program. In response to Consolidated Appropriations Resolution, 2003, Pub. L. No. 108-7, Division M, § 111(b), May 20, 2003. https://epic.org/privacy/profiling/tia/may03_report.pdf.

Palantir's intelligence software, purchased by the NSA, provides machine learning capabilities. Peter Thiel, cofounder of PayPal, later cofounded Palantir with the intent to implement PayPal's fraud detection methods within counterterrorism software:

- Shane Harris, "Palantir Technologies spots patterns to solve crimes and track terrorists, *Wired.co.uk,* July 31, 2012. www.wired.co.uk/magazine/archive/2012/09/features/joining-the-dots/viewall.
- Matt Burns, "Leaked Palantir Doc Reveals Uses, Specific Functions and Key Clients," *Tech Crunch Online*, January 11, 2105. http://techcrunch.com/2015/01/11/leaked-palantir-doc-reveals-uses-specific-functions-and-key-clients/.
- Shane Harris, "Killer App. Have a bunch of Silicon Valley geeks at Palantir Technologies figured out how to stop terrorists?" *Washingtonian,* January 31, 2012. www.washingtonian.com/articles/people/killer-app/.
- Kevin Simler, "Palantir: So What Is It You Guys Do?" *The Palantir Blog,* December 4, 2007 [blog post]. https://www.palantir.com/2007/12/what-do-we-do/.

- Quora Question: “How important is machine learning to what Palantir is doing?” Matt Gordon and Anirvan Mukherjee, *Quora*, May 24, 2014. www.quora.com/How-important-is-machine-learning-to-what-Palantir-is-doing.

Regarding efforts to use data mining to "hunt for terrorists..., Cognito sells software to the NSA that the company says can find patterns in massive amounts of data, such as lists of telephone calling records.”:

- Arshad Mohammed and Sara Kehaulani Goo, “Government Increasingly Turning to Data Mining,” *The Washington Post*, June 15, 2006. www.washingtonpost.com/wp-dyn/content/article/2006/06/14/AR2006061402063.html.

A FISA court document requesting authorization for additional data collection describes two "major tools" for applying analytics to metadata: The first is the chaining of linked phone numbers, and the second is redacted, with PA being a possibility. On page 10, after discussing chaining linked phone numbers. The second major tool analysts can use with an archive of collected metadata is [redacted phrases of perhaps one or two substantive sentences].":

- Foreign Intelligence Surveillance Court document, "Memorandum of Law in Support of Application for Certain Tangible Things for Investigations to Protect against International Terrorism." Exhibit C, Docket Number: BR: 06-05, 1871(c) (2), July 2009. www.aclu.org/files/section215/12122013/2006.5.23%20Govt.%20Memo.%20of%20Law.pdf.

The “ABC Terrorism Prediction Model” by Psynapse Technologies predicts terrorist attacks that fall within an ongoing campaign with about 85 percent accuracy (first citation); the NSA uses this model (second citation):

- Dr. William Forrest Crain (Technical Chair), “The Global War on Terrorism: Analytical Support, Tools, and Metrics of Assessment (Unclassified),” Military Operations Research Society Workshop, US Naval War College, August 11, 2005. http://www.dtic.mil/dtic/tr/fulltext/u2/a442451.pdf.
- William M. Arkin, “Telephone Records are just the Tip of NSA’s Iceberg,” Global Research, May 14, 2006.

www.globalresearch.ca/telephone-records-are-just-the-tip-of-nsa-s-iceberg/2444?print=1.

The DARPA Information Awareness Office (IAO) pursues "real time learning, pattern matching."

- The DAPRA Information Awareness Office (IAO) Announcement**,** The DARPA Information Awareness Office (IAO), August 2, 2002. http://web.archive.org/web/20020802012150/http://www.darpa.mil/iao.

NSA technologies and patents for data mining and data analysis, respectively:

- John Markoff, "Taking Spying to Higher Level, Agencies Look for More Way to Mine Data," *The New York Times,* February 25, 2006. www.nytimes.com/2006/02/25/technology/25data.html?pagewanted=all&_r=0.
- Alex Salkever, "NSA Patents Analysis and Visualization," KDnuggets.com Post, September 26, 2015. www.kdnuggets.com/2015/09/nsa-patents-analysis-visualization.html.

NSA-patented Renoir tool enables data mining:

- National Security Agency Technology Transfer Program Technology Catalog, March 2015. https://www.nsa.gov/research/_files/tech_transfers/nsa_technology_transfer_program.pdf.

The FBI applies PA to assign terrorism "risk scores" to suspects:

- Martin H. Bosworth, "FBI Uses Data Brokers, "Risk Scores" To Hunt Terrorists," *Consumer Affairs*, July 11, 2007. www.consumeraffairs.com/news04/2007/07/fbi_risk_scores.html.
- Ellen Nakashima, "FBI Plans Initiative To Profile Terrorists," *The Washington Post*, July 11, 2007. www.washingtonpost.com/wp-dyn/content/article/2007/07/10/AR2007071001871.html.

U.S. Armed Forces conduct research to analytically predict terrorist attacks:

- *See the listings under "U.S. Armed Forces" in the Notes for Central Table 5.*

Arguments that intelligence agencies should and will turn to predictive analytics to fight terrorism and other crime:

- Manilio Allegra, "Where High-accuracy Wireless Location is (and isn't) Headed in 2014," *Directions Magazine*, January 13, 2014. www.directionsmag.com/entry/where-high-accuracy-wireless-location-is-and-isnt-headed-in-2014/377368.

The NSA is positioned to apply face recognition, which can provide further individual-level data for both suspects and non-suspects, and also itself requires capabilities in machine learning to execute:

- James Risen and Laura Poitras, "N.S.A. Collecting Millions of Faces From Web Images," The New York Times, May 31, 2014. http://www.nytimes.com/2014/06/01/us/nsa-collecting-millions-of-faces-from-web-images.html?_r=0.

Further speculation on the use of PA by the NSA and other government agencies:

- Andrew Couts, "What's the NSA Picking out of Your Phone Calls? Just 'Unvolunteered Truths,'" Digital Trends Online, August 31, 2013. www.digitaltrends.com/mobile/whats-the-nsa-picking-out-of-your-phone-calls-just-unvolunteered-truths/.
- Adam Mazmanian, "What the NSA Can't do with Your Data (probably)," FCW Online, June 12, 2013. www.predictiveanalyticsworld.com/patimes/what-the-nsa-cant-do-with-your-data-probably/.
- Dean Abbott, "The NSA, Link Analysis and Fraud Detection," Predictive Analytics Times Online, July 1, 2013. www.predictiveanalyticsworld.com/patimes/the-nsa-link-analysis-and-fraud-detection/.
- Alex Howard, "Predictive analytics and data sharing raise civil liberties concerns," O'Reilly Radar Online, April 11, 2013.

www.predictiveanalyticsworld.com/patimes/predictive-analytics-and-data-sharing-raise-civil-liberties-concerns/.

A former NSA cyber expert on the need for predictive analytics for fighting terrorism:

- Jason Murdock, "Former NSA cyber expert: We are all enlisted in the cyber war," *V3.co.uk*, November 20, 2015. www.v3.co.uk/v3-uk/feature/2435732/former-nsa-cyber-expert-we-are-all-enlisted-in-the-cyber-war.

The U.S. Special Operations Forces, which is dedicated to the war on terrorism, uses predictive analytics to dive hiring decisions:

- Jeanne Sahadi, "The cost of fighting terrorism," *CNN Money*, November 16, 2015. http://money.cnn.com/2015/11/16/news/economy/cost-of-fighting-terrorism/index.html.
- Dean Abbott, President, Abbott Analytics, "Hiring and Selecting Special Forces Personnel Using Predictive Analytics," Predictive Analytics World for Workforce San Francisco Conference, March 31, 2015, San Francisco, CA. www.predictiveanalyticsworld.com/workforce/2015/agenda.php#day1-1030b. Full presentation slide deck: www.predictiveanalyticsworld.com/workforce/2015/presentations/pdf/1030_ Track2_Abbott.pdf.

Predictive policing that is closer to "pre-crime" – targets individuals more likely to commit future crimes (beyond the predictive patrolling of neighborhoods) – the first was a front page New York Times article:

- John Eligon and Timothy Williams, "Police Program Aims to Pinpoint Those Most Likely to Commit Crimes," *The New York Times*, September 24, 2015. www.nytimes.com/2015/09/25/us/police-program-aims-to-pinpoint-those-most-likely-to-commit-crimes.html?_r=0.
- Justin Jouvenal, "The New Way Police are Surveilling You: Calculating Your Threat 'Score'," *The Washington Post*, January 10, 2016. https://www.washingtonpost.com/local/public-safety/the-new-way-police-are-surveilling-you-calculating-your-threat-score/2016/01/10/e42bccac-8e15-11e5-baf4-bdf37355da0c_story.html.

USA Freedom Act, shut down of bulk data collection, and the associated debate:

NSA bulk collection of phone records authorized (only) until November 29, 2015 per the USA Freedom Act:

- The United States Department of Justice, National Security Division 15-1061, Office of Public Affairs Press Release, *Joint Statement by the Department of Justice and the Office of the Director of National Intelligence on the Declassification of the Renewal of Collection Under Section 215 of the USA Patriot Act as Amended by the USA Freedom Act,* August 28, 2015. www.justice.gov/opa/pr/joint-statement-department-justice-and-office-director-national-intelligence-declassificati-1.

Legislators debate the issues related to passing the USA Freedom Act:

- Jennifer Steinhauer and Jonathan Weisman, "U.S. Surveillance in Place Since 9/11 Is Sharply Limited," *The New York* Times, June 2, 2015. www.nytimes.com/2015/06/03/us/politics/senate-surveillance-bill-passes-hurdle-but-showdown-looms.html?_r=1.

The National Research Council of the Academy of Sciences argues that no technology can achieve the same effect in lieu of bulk data and recommends the NSA to predictively target surveillance, "to monitor trends or patterns in communications that might lead to intelligence insights":

- National Research Council of the National Academies, *Bulk Collection of Signals Intelligence,* (National Academies Press 2015). www.nap.edu/download.php?record_id=19414#.

NSA says ending bulk collection will "significantly reduce their operational capabilities":

- National Security Agency Director Michael Rogers Senate Select Intelligence Committee testimony on Intelligence and Cybersecurity Issues on September 24, 2015, *C-SPAN*, September 24, 2015. www.c-span.org/video/?328309-1/hearing-cybersecurity.

Bulk data's "primary potential investigative value: identifying new intelligence targets in the future from a massive pool of previously collected information":

- Spencer Ackerman, “Technology offers no magic solutions to bulk data collection issues, says panel,” *The Guardian*, January 15, 2015. www.theguardian.com/world/2015/jan/15/technology-no-magic-solution-bulk-data-collection-panel.

European Union concern over its citizens' data being accessed by the NSA:

- Patrick Condren, “European Court of Justice’s AG Rules ‘Safe Harbour’ Invalid Impacting Global Use Data,” *icrunchdata News*, October 6, 2015. https://icrunchdatanews.com/european-court-justice-rules-safe-harbour-invalid-impacting-global-data-use/.

Arguments the USA Freedom Act still allows "too much" continuation of the data collection possible under the Patriot Act:

- The Editorial Board, “A Surveillance Bill That Falls Short,” *The New York Times*, May 22, 2014. www.nytimes.com/2014/05/23/opinion/a-surveillance-bill-that-falls-short.html.
- Mattathias Schwartz, “Three Big Questions About the N.S.A’s Patriot Act Powers,” *The New Yorker,* June 2, 2015. www.newyorker.com/news/news-desk/three-big-questions-about-the-n-s-a-s-patriot-act-powers.
- Ali Watkins and Jessica Schulberg, “Senate Votes To Turn NSA Spying Back On, But With Reforms,” *Huffington Post*, June 2, 2015. www.huffingtonpost.com/2015/06/02/nsa-surveillance-congress_n_7496100.html.
- Edward Felten presents the first talk in the speaker series "Surveillance in the Age of Big Data," co-sponsored by ISTS and the Computer Science Colloquium, April 7, 2014. https://www.youtube.com/watch?v=tyPL_sN776k (see content around 3:20).
- RT Op-Edge Interview with Ted Rall, “‘Freedom Act will make no difference’ to NSA surveillance,” *RT.com*, June 3, 2015. http://rt.com/op-edge/264605-us-freedom-act-snooping/.

- USA Freedom Act, Wikipedia.com. http://en.wikipedia.org/wiki/USA_Freedom_Act.
- Michael Brenner, "The NSA's Second Coming," *The Huffington Post*, June 8, 2015. www.huffingtonpost.com/michael-brenner/the-nsas-second-coming_b_7535058.html.
- Paul Waldman, "A reality check on the future of government spying," *The Washington Post*, June 3, 2015. www.washingtonpost.com/blogs/plum-line/wp/2015/06/03/a-reality-check-on-the-future-of-government-spying/.
- Lauren Walker, "NSA to Destroy Data Collected From Mass Phone Surveillance," *Newsweek*, July 27, 2015. www.newsweek.com/nsa-destroy-data-collected-mass-phone-surveillance-357500.
- Peter Baker and David E. Sanger, "Why the N.S.A. Isn't Howling Over Restrictions," *The New York Times*, May 1, 2015. www.nytimes.com/2015/05/02/us/politics/giving-in-a-little-on-national-security-agency-data-collection.html.
- Norman Solomon, "The USA Freedom Act Is a Virtual Scam," *The Huffington Post,* June 5, 2015. www.huffingtonpost.com/norman-solomon/the-usa-freedom-act-is-a_b_7519046.html.

"The bigger haystack makes it harder to find the needle":

- Mattathias Schwartz, "The Whole Haystack: The N.S.A. claims it needs access to all our phone records. But is that the best way to catch a terrorist?" *The New Yorker*, January 26, 2015. www.newyorker.com/magazine/2015/01/26/whole-haystack.

The case for further opening healthcare research data:

- John Wilbanks, "Is Too Much Privacy Bad For Your Health?" *Part 4 of the TED* Radio Hour *episode* The End of Privacy, NPR Radio, January 31, 2014. www.npr.org/2014/01/31/265700003/is-too-much-privacy-bad-for-your-health.
- John Wilbanks, "Let's pool our medical data," TEDGlobal 2012. (*This talk was presented at an official TED conference),* filmed June, 2012. www.ted.com/talks/john_wilbanks_let_s_pool_our_medical_data.
- Matt Reaney (BigCloud.io), "Big Data Could Revolutionize Healthcare. Will We Let It?" KDnuggets.com [Blog Post], January 31, 2015. www.kdnuggets.com/2015/01/big-data-revolutionize-healthcare.html.

- I. Glenn Cohen, Ruben Amarasingham, Anand Shah, Bin Xie and Bernard Lo, "The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care," HealthAffairs Online, doi: 10.1377/hlthaff.2014.0048 *Health Aff July 2014 vol. 33 no. 7 1139-1147*. http://content.healthaffairs.org/content/33/7/1139.abstract?=right.

Ethics in predictive healthcare:

- Glen Cohen, Ruben Amarasingham, Anand Shah, Bin Xie and Bernard Lo, "The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care," *HealthAffairs*, Vol. 33, No. 7, July, 2014. http://content.healthaffairs.org/content/33/7/1139.abstract.

Treatises arguing for broad-scoped NSA data collection:

- Dianne Feinstein, "The NSA's Watchfulness Protects America," *The Wall Street Journal*, October 13, 2013. www.wsj.com/articles/SB10001424052702304520704579125950862794052.
- Meta S. Brown, "NSA Mass Surveillance: Biggest Big Data Story," *Forbes,* August 27, 2015. www.forbes.com/sites/metabrown/2015/08/27/nsa-mass-surveillance-biggest-big-data-story/.

Brazil's president, Dilma Rousseff, on privacy:

- Raymond Colitt and Arnaldo Galvao, "Rousseff Calls Off U.S. Visit Over NSA Surveillance," Bloomberg Business Online, September 17, 2013. www.bloomberg.com/news/articles/2013-09-17/brazil-said-to-call-off-state-visit-in-retaliation-for-spying.
- Julian Borger, "Brazilian president: US surveillance a 'breach of international law,'" *The Guardian,* September 24, 2013. www.theguardian.com/world/2013/sep/24/brazil-president-un-speech-nsa-surveillance.

Regarding the question of when/whether data can itself form sufficient particularized suspicion for an individual to warrant seizure or detention of the individual:

- Guthrie Andrew Ferguson, "Big Data and Predictive Reasonable Suspicion," University of Pennsylvania Law Review, Vol. 163 U. Pa. L. Rev. 327 (2015). http://ssrn.com/abstract=2394683 and http://dx.doi.org/10.2139/ssrn.2394683.

- Kate Vinton, "Court Rules Police Don't Need A Warrant To Get Your Cellphone Records," *Forbes*, May 5, 2015. www.forbes.com/sites/katevinton/2015/05/05/court-rules-cops-dont-need-a-warrant-to-get-your-cellphone-records/.

CHAPTER 3: The Data Effect: A Glut at the End of the Rainbow *(data)*

Media that reported an orange used car is least likely to be a lemon:

- Marcus Wohlsen, "San Francisco startup makes data science a sport." *The Seattle Times*, April 14, 2012. http://old.seattletimes.com/html/businesstechnology/2017983961_apusscienceassport.html.
- Viktor Mayer-Schönberger, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Eamon Dolan/Mariner Books; Reprint Edition, March 4, 2014).
- Marcus Wohlsen, "Kaggle Competitions Make Data Science A Sport," *The Huffington Post*, April 15, 2012. www.huffingtonpost.com/2012/04/15/kaggle-competitions-data-science_n_1427204.html.
- Deborah Gage, "Big Data Uncovers Some Weird Correlations," *The Wall Street Journal*, March 23, 2014. www.wsj.com/articles/SB10001424052702303369904579423132072969654.
- Quentin Hardy, "Bizarre Insights From Big Data," *The New York Times*, March 28, 2012. http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/?_r=0
- NPR Morning Edition, "The 'Big Data' Revolution: How Number Crunchers Can Predict Our Lives," National Public Radio Online, March 7, 2013. www.npr.org/2013/03/07/173176488/the-big-data-revolution-how-number-crunchers-can-predict-our-lives.

The competition and public data set for predicting bad buys among used cars, sponsored by Carvana, a data driven company that modernizes onine used car buying:

- Kaggle, "Predict if a car purchased at auction is a lemon," Competition, September 30, 2011. www.kaggle.com/c/DontGetKicked.

Telecom companies using social data to predict customer defection:

- Auren Hoffman, "Birds of a feather use the same phone service," Summation [Blog Post], November 4, 2009. http://blog.summation.net/2009/11/birds-of-a-feather-use-the-same-phone-service-provider.html.
- Michael Driscol, PhD, Principal, Founder, Dataspora LLC, "*The Social Effect: Predicting Telecom Customer Churn with Call Data,*" Predictive Analytics World Conference San Francisco, February 16, 2010, San Francisco, CA. www.predictiveanalyticsworld.com/sanfrancisco/2010/agenda.php#day1-12.

Banks assess creditworthiness via social media data:

- Ken Lin, "What Banks and Lenders Know About You From Social Media," *Mashable*, October 7, 2011. http://mashable.com/2011/10/07/social-media-privacy-banks/.
- Richard Satran, "IRS High-Tech Tools Track Your Digital Footprints," *U.S. News & World Report*, April 4, 2013. http://money.usnews.com/money/personal-finance/mutual-funds/articles/2013/04/04/irs-high-tech-tools-track-your-digital-footprints.
- Knowledge@Wharton, "The Surprising Ways that Social Media Can Be Used for Credit Scoring," The Wharton School of the University of Pennsylvania, November 5, 2014. http://knowledge.wharton.upenn.edu/article/using-social-media-for-credit-scoring.
- "Your smartphone's secrets could help you bag a bank loan," *New Scientist,* April 22, 2015, Issue 3018. www.newscientist.com/article/mg22630182.400-your-smartphones-secrets-could-help-you-bag-a-bank-loan.html#.VVFOsdpVikp.
- Daniel Björkegren and Darrell Grissen, "Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment," Brown University, Department of Economics, July 13, 2015. http://dan.bjorkegren.com/files/danbjork_creditscoring.pdf.

The IRS considers social media data to target tax audits:

- Richard Satran, "IRS High-Tech Tools Track Your Digital Footprints," *U.S. News & World Report*, April 4, 2013. http://money.usnews.com/money/personal-finance/mutual-funds/articles/2013/04/04/irs-high-tech-tools-track-your-digital-footprints.
- Dean Silverman, Senior Advisor to the Commissioner and Director, Office of Compliance Analytics, IRS, "Making Analytics Pay," Predictive Analytics World for Government Conference, September 17, 2012, Washington, DC. http://www.predictiveanalyticsworld.com/gov/2012/agenda.php#day1-115.
- Chuck Norris, "IRS Now Robo-audits Your Spending," *Ammoland,* April 9, 2013. www.ammoland.com/2013/04/irs-now-robo-audits-your-spending/.

On the value of data: "Harrah's parent company, Caesar's Entertainment, declared bankruptcy as a consequence of overextension and growing competition. During proceedings, creditors appraised Caesar's vast store of customer data as the company's most valuable asset, worth about $1 billion."

- Andrew Thompson, "Engineers of Addiction," *The Verge*, May 6, 2015. www.theverge.com/2015/5/6/8544303/casino-slot-machine-gambling-addiction-psychology-mobile-games.

Twitter for measuring unemployment:

- Thomas Claburn, "Tweets Tell Whether You Have A Job," *InformationWeek,* November 20, 2014. www.predictiveanalyticsworld.com/patimes/tweets-tell-whether-job/.

A survey of stock market predictions using the web:

- Michaela Nardo, Marco Petracco-Giudici, and Minás Naltsidis, "Walking Down Wall Street with a Table: A Survey of Stock Market Predictions Using the Web," *Journal of Economic Surveys*, January 16, 2015, ISI Journal Citation Reports © Ranking: 2014: 78/333 (Economics), Online ISSN: 1467-6419. http://onlinelibrary.wiley.com/doi/10.1111/joes.12102/full.

In this book's first edition, this chapter (Chapter 3) covered the following research project in detail—note its statistical soundness is criticized—yet its general area of exploration held as interesting—by the next citation immediately below:

- Eric Gilbert and Karrie Karahalios, Department of Computer Science, University of Illinois at Urbana–Champaign, "Widespread Worry and the Stock Market," Association for the Advancement of Artificial Intelligence (www.aaai.org), March 12, 2010. http://comp.social.gatech.edu/papers/icwsm10.worry.gilbert.pdf.

Statistical faults in prior work claiming to predict the stock market from social media:

- Rapheal Olaniyan, Daniel Stamate and Doina Logofatu, "Social Web-Based Anxiety Index's Predictive Information on S&P 500 Revisited," *Statistical Learning and Data Sciences*, Volume 9047, pp 203-213, April 3, 2015. http://link.springer.com/chapter/10.1007/978-3-319-17091-6_15.
- Στενĕη Πăν, "The junk science behind the 'Twitter Hedge Fund," Buy the Hype blog, April 14, 2012. http://sellthenews.tumblr.com/post/21067996377/noitdoesnot.
- Στενĕη Πăν, "An open letter to Johan Bollen," Buy the Hype blog, July 1, 2012. http://sellthenews.tumblr.com/post/26259141917/open-letter-to-bollen.
- Ben Gimbert, "Sour Grapes: Seven Reasons Why "That" Twitter Prediction Model is Cooked," Some Ben? Blog, May 13, 2011. http://blog.someben.com/2011/05/sour-grapes-seven-reasons-why-that-twitter-prediction-model-is-cooked/.

Questions about the legitimacy of social media based prediction:

- Snarky skeptical statistics blog post, "The Junk Science behind the 'Twitter Hedge Fund,'" April 14, 2012. http://sellthenews.tumblr.com/post/21067996377/noitdoesnot.
- Ben Gimpert, "Sour Grapes: Seven Reasons Why 'That' Twitter Prediction Model Is Cooked," Some Ben? blog post, May 13, 2011. http://blog.someben.com/2011/05/ sour-grapes-seven-reasons-why-that-twitter-prediction-model-is-cooked/.

Governments such as Bhutan's measure mass mood, e.g., via their *gross national happiness* index, as a means to track prosperity:

- Jyoti Thottam/Thimphu, "The Pursuit of Happiness," *TIME*, Vol. 180, No. 17, October 22, 2012. www.time.com/time/magazine/article/0,9171,2126639,00.html.

Source of figure showing Calm and Happy across Election Day and Thanksgiving:

- J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2, no. 1 (March 2011), 1–8, doi:10.1016/j.jocs.2010.12.007, arxiv: abs/1010.3003.

Predicting by social media:

- Sitaram Asu and Bernardo A. Huberman, "Predicting the Future with Social Media," Cornell University Library, March 29, 2010, arXiv.org, arXiv:1003.5699. http://arxiv.org/abs/ 1003.5699/.
- Anshul Mittal and Arpit Goel, "Stock Prediction Using Twitter Sentiment Analysis," Stanford University Libraries, December 16, 2011. http://cs229.stanford.edu/proj2011/ GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf.
- Allison Aubrey, "Happiness: It Really Is Contagious," NPR News, *All Things Considered*, December 5, 2008. www.npr.org/templates/story/story.php?storyId597831171.
- Shea Bennett, "Can Twitter Beat the Stock Market? Tweet Sentiment Trading API Bets That It Can," Mediabistro, July 5, 2012. www.mediabistro.com/alltwitter/twitter-trading-api_b24992.
- Jim Giles, "Blogs and Tweets Could Predict the Future," *New Scientist Online*, June 21, 2010. www.newscientist.com/article/mg20627655.800.
- Ciara Byrne, "Social Media Popularity Can Predict Stock Prices," Venture Beat Online, March 17, 2011. http://venturebeat.com/2011/03/17/study-social-media-popularity- can-predict-stock-prices/.
- Nicholas A. Christakis and James H. Fowler, "Happiness Is Contagious," *Edge: The Third Culture*, December 4, 2008. www.edge.org/3rd_culture/christakis_fowler08/ christakis_fowler08_index.html.
- Zhi Da, Joseph Engelberg, and Pengjie Gao, "In Search of Attention," University of Notre Dame, June 3, 2010. nd.edu/Bpgao/papers/Google_04june2010.pdf.

- Global Pulse, "Unemployment through the Lens of Social Media," Global Pulse Research, in partnership with S.A.S. www.unglobalpulse.org/projects/can-social-media- mining-add-depth-unemployment-statistics.
- Yigitcan Karabulut, "Can Facebook Predict Stock Market Activity?" Miami University, first draft: August 29, 2011, second draft: October 17, 2011. http://bus.miami.edu/ umbfc/_common/files/papers/Karabulut.pdf.
- Brian Womack, "Twitter Feeds on Netflix Prompt Stock Service," Bloomberg Online, January 23, 2012. www.bloomberg.com/news/2012–01–23/social-media-predicting- stock-moves-spawn-sentiment-firms-tech.html.
- Charlie White, "Who Will Win a Grammy? Twitter Predicts the Future," Mashable Entertainment, February 11, 2012. http://mashable.com/2012/02/11/grammy-twitter- predictions/.
- Yuan Zhang, Jie Tang, Jimeng Sun, Yiran Chen, and Jinghai Rao, "MoodCast: Emotion Prediction via Dynamic Continuous Factor Graph Model," Tsinghua University, China, October 11, 2010. http://keg.cs.tsinghua.edu.cn/jietang/publications/ICDM10-Zhang- et-al-MoodCast.pdf.
- Yifu Huang, Shuigeng Zhou, Kai Huang, Jihong Guan, "Boosting Financial Trend Prediction with Twitter Mood Based on Selective Hidden Markov Models, *Database Systems for Advanced Applications, Lecture Notes in Computer Science,* April 9, 2015, Volume 9050, 2015, pp 435-451. http://link.springer.com/chapter/10.1007/978-3-319-18123-3_26.
- Xue Zhang, Hauke Fuehres, and Peter A. Gloor, "Predicting Stock Market Indicators through Twitter: 'I Hope It Is Not as Bad as I Fear,'" *Procedia—Social and Behavioral Sciences* 26 (2011): 55–62. www.sciencedirect.com/science/article/pii/S1877042811023895.

Regarding stock prediction:

- Lisa Grossman, "Twitter Can Predict the Stock Market?" *Wired Magazine Online*, October 19, 2010. www.wired.com/wiredscience/2010/10/twitter-crystal-ball/.
- Bradley Hope, "How Computers Trawl a Sea of Data for Stock Picks," *The Wall Street Journal*, April 1, 2015. www.wsj.com/articles/how-computers-trawl-a-sea-of-data-for-stock-picks-1427941801.

- NPR Staff, "Need Stock Tips? Read Your Tweets," NPR News, *All Things Considered*, October 24, 2010. www.npr.org/templates/story/story.php?storyId5130793579.
- "Can Twitter Predict the Future?" *The Economist*, June 2, 2011. www.economist.com/node/18750604.
- Randy Saaf, AlphaGenius, "Using Twitter & the Social Internet to Obtain Above Market Returns," Text Analytics World San Francisco Conference, March 7, 2012, San Francisco, CA. www.textanalyticsworld.com/sanfrancisco/2012/agenda/full-agenda#day1430–450–2.

A venture capital firm has appointed a computer algorithm to its board of directors:

- BBC News Staff, "Algorithm appointed board director," BBC News, May 16, 2014. www.bbc.com/news/technology-27426942.

On the author's mom, Lisa Schamberg, and making art of compost:

- Lisa Schamberg, "Self-Composed," *Seven Days: Vermont's Independent Voice*. www.7dvt.com/lisa-schamberg.
- Lisa Schamberg, "Self-Composed: Photography by Lisa Schamberg." www.lisaschamberg.com.

On the age of big data:

- Steve Lohr, "The Age of Big Data," *New York Times*, February 11, 2012. www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html.

A nice timeline summary of some of the key junctures in the cultural development of today's data movement(s):

- Gil Press, "A Very Short History of Data Science," *Forbes*, May 28, 2013. www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/.

Online dating analytics:

- Christian Rudder, *Dataclysm: Love, Sex, Race and Identity—What Our Online Lives Tell Us about Our Offline Selves* (Broadway Books, 2015).
- Ravi Kalakota, "Love, Sex and Predictive Analytics: Tinder, Match.com, and OkCupid," PracticalAnalytics.com [Blog Post],

May 29, 2015. https://practicalanalytics.wordpress.com/2015/05/29/love-sex-and-predictive-analytics-tinder-match-com-and-okcupid/.

Mandate for U.S. federal agencies to release data sets:

- Wendy R. Ginsberg, Analyst in Government Organization and Management, "The Obama Administration's Open Government Initiative: Issues for Congress," Congressional Research Service, 7–5700, www.crs.gov, R41361, January 28, 2011. www.scribd.com/ doc/48966713/CRS-The-Obama-Administration%E2%80%99s-Open-Government- Initiative-Issues-for-Congress.

Regarding the number of blogs:

- Computing Community Consortium blog, "Big Data," facts and figures posted by IBM, June 28, 2012. www.cccblog.org/wp-content/uploads/2012/06/Big-Data-Final.jpeg.
- Pingdom Team, "Internet 2011 in Numbers," Pingdom blog, January 17, 2012. http:// royal.pingdom.com/2012/01/17/internet-2011-in-numbers/.

Number of daily tweets:

- Twitter blog, "200 Million Tweets per Day," Twitter, June 20, 2011. http://blog.twitter.com/2011/06/200-million-tweets-per-day.html.

Estimation of number of web pages:

- WorldWideWebSize.com, "The Size of the World Wide Web (the Internet)," updated daily. www.worldwidewebsize.com/.

Millions of retail transactions per hour:

- John Naughton, "Why Big Data Is Now Such a Big Deal," *The Guardian*'s *The Observer Online*, March 17, 2012. www.guardian.co.uk/technology/2012/mar/18/big-data-storage-analysis-internet.

Number of photographs on Facebook:

- Quora, "How Many Photos Are Uploaded to Facebook Each Day?" January 2011. www.quora.com/How-many-photos-are-uploaded-to-Facebook-each-day.
- Jonathan Good, "How Many Photos Have Ever Been Taken?" 1000memories blog, September 15, 2011.

http://blog.1000memories.com/94-number-of-photos-ever-taken-digital-and-analog-in-shoebox.
- Instagram, "Photoset," Instagram blog, September 2, 2012. http://blog.instagram.com/.

Regarding femto-photography:

- Ramesh Raskar, Moungi G. Bawendi, Andreas Velten, Everett Lawson, Amy Fritz, Di Wu, Matt O'Toole, Diego Gutierrez, Belen Masia, and Elisa Amoros, "Femto-Photography: Visualizing Photons in Motion at a Trillion Frames per Second," MIT Media Lab, June 14, 2012. http://web.media.mit.edu/Braskar/trillionfps/.
- "Ramesh Raskar: Imaging at a Trillion Frames per Second," YouTube, Web, July 26, 2012. https://www.youtube.com/watch?v5Y_9vd4HWlVA.

Number of mobile devices:

- Matt Warman, "50 Billion Devices Online by 2020," *The Telegraph*, January 31, 2012. www.telegraph.co.uk/technology/internet/9051590/50-billion-devices-online-by-2020.html.
- Zachary Davies Boren, "There are officially more mobile devices than people in the world," *The Independent*, July 26, 2015. www.independent.co.uk/life-style/gadgets-and-tech/news/there-are-officially-more-mobile-devices-than-people-in-the-world-9780518.html.
- Karen Tillman, "How Many Internet Connections are in the World? Right. Now." Cisco blog, July 29, 2013. http://blogs.cisco.com/news/cisco-connections-counter.

Daily data growth:

- "The New World of Massive Data Mining," *The Diane Rehm Show*, WAMU 88.5, Washington, DC, April 2, 2012. http://thedianerehmshow.org/shows/2012–04–02/ new-world-massive-data-mining.

Regarding the cost of information storage:

- Matthew Komorowski, "The History of Storage Cost," MKomo.com. www.mkomo.com/cost-per-gigabyte.

Covering land masses with data:

- Marin Hilbert, "That Giant Sifting Sound: A Short History of Big Data," *The Economist*, The Ideas Economy, June 7–8, 2011, Santa Clara, CA. http://fora.tv/2011/06/07/

That_Giant_Sifting_Sound_A_Short_History_of_Big_Data#What_Drives_the_Global_Growth_of_Information.

- Martin Hilbert and Priscila Lopez, "The World's Technological Capacity to Store, Communicate, and Information," *Science* 332, no. 6025 (April 1, 2011), 60–65. doi:10.1126/ science.1200970. www.sciencemag.org/content/332/6025/60.

NSA accumulating Internet traffic:

- James Bamford, "The NSA Is Building the Country's Biggest Spy Center (Watch What You Say)," *Wired Magazine Online*, March 15, 2012. www.wired.com/threatlevel/2012/ 03/ff_nsadatacenter/all/1.

More on the massive accumulation of data:

- Jeffrey Marlow, "What to Do with 1,000,000,000,000,000,000 Bytes of Astronomical Data per Day," *Wired Magazine Online*, April 2, 2012. www.wired.com/wiredscience/2012/04/ what-to-do-with-1000000000000000000-bytes-of-astronomical-data-per-day/.
- Thomas H. Maugh II, "For the First Time, Researchers Track Manta Rays with Satellites," *Los Angeles Times*, May 12, 2012. http://articles.latimes.com/2012/may/12/science/la-sci-sn-manta-satellites-20120512.
- Steven James Snyder, "Inside the Kurzweil SXSW Keynote: On Infinite Mind Power, Robotic Overlords and Immortality," *Time*, Techland, March 14, 2012. http://techland.time.com/2012/03/14/kurzweil-south-by-southwest-keynote-speech-grossman/.
- Bernard Marr, Best-Selling Author, Keynote Speaker and Leading Business and Data Expert at Advanced Performance Institute, "Big Data – 25 Amazing Facts Everyone Should Know, Slideshare.net post, September 24, 2014. www.slideshare.net/BernardMarr/big-data-25-facts.

Eight zettabytes by 2012:

- Dan Vesset, Benjamin Woo, Henry D. Morris, Richard L. Villars, Gard Little, Jean S. Bozman, Lucinda Borovick, Carl W. Olofson, Susan Feldman, Steve Conway, Matthew Eastwood, and Natalya Yezhkova, "Worldwide Big Data Technology and Services 2012–2012 Forecast," *ICD Analyze the Future*, March 2012, Doc #233485. www.idc.com/getdoc.jsp?containerId5233485.

Prior to our naming this The Data Effect, Tom Khabaza pointed out that there are always patterns:

- Tom Khabaza, "Nine Laws of Data Mining—Part 2," *Data Mining & Predictive Analytics*, edited by Tom Khabaza, January 14, 2012. http://khabaza.codimension.net/index_files/Page346.htm.

This Economist article frames The Data Effect in another way, giving examples where there are "little things that mean a lot":

- Schumpeter, "Little things that mean a lot," *The Economist*, July 19, 2014. www.economist.com/news/business/21607816-businesses-should-aim-lots-small-wins-big-data-add-up-something-big-little.

"Personal data is the new oil of the Internet and the new currency of the digital world":

- Meglena Kuneva, European Consumer Commissioner, March 2009, "Personal Data: The Emergence of a New Asset Class," An Initiative of the World Economic Forum, January 2011. http://gerdleonhard.typepad.com/files/wef_ittc_personaldatanewasset_report_2011.pdf.

Table: Bizarre and Surprising Insights—Consumer Behavior (Chapter 3)

Guys literally drool over sports cars:

- James Warren, "Just the Thought of New Revenue Makes Mouths Water," *New York Times*, September 29, 2011. www.nytimes.com/2011/09/30/us/just-the-thought-of- new-revenue-makes-mouths-water.html.
- Christopher Shea, "Mouth-Watering Consumer Goods," *Wall Street Journal*, August 24, 2011. http://blogs.wsj.com/ideas-market/2011/08/24/mouth-watering-consumer-goods/.
- David Gal, "A Mouth-Watering Prospect: Salivation to Material Reward," *Journal of Consumer Research* 38, no. 6 (April 2012): 1022–1029. www.jstor.org/discover/10.1086/ 661766.

If you buy diapers, you are more likely to also buy beer:

- Daniel J. Power, "What Is the 'True Story' about Data Mining, Beer and Diapers?" *DSS News* 3, no. 23 (November 10, 2002). www.dssresources.com/newsletters/66.php.

- Ronny Kohavi, "Crossing the Chasm: From Academic Machine Learning to Commercial Data Mining," International Conference on Machine Learning 1998, Slide 21. http://robotics.stanford.edu/Bronnyk/chasm.pdf.

Dolls and candy bars:

- Christopher Palmeri, "Believe in Yourself, Believe in the Merchandise," *Forbes Magazine Online*, September 8, 1997. www.forbes.com/forbes/1997/0908/6005118a.html.

Pop-Tarts before a hurricane:

- Constance L. Hays, "What Wal-Mart Knows About Customers' Habits," *The New York Times*, November 14, 2004. http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html?_r=0.
- Olivia Katrandjian, "Hurricane Irene: Pop-Tarts Top List of Hurricane Purchases," *ABC World News,* August 27, 2011. http://abcnews.go.com/US/hurricanes/hurricane-irene-pop-tarts-top-list-hurricane-purchases/story?id=14393602.

Staplers reveal hires:

- Based on an example presented by Wayne Eckerson, formerly with TDWI, now founder and principal consultant at BI Leader Consulting. http://tdwi.org/blogs/wayne-eckerson/ list/waynes-blog.aspx and www.bileader.com/.

Higher crime, more Uber rides:

- Voytek, "How Crime Location Knowledge is a Proxy for Uber Demand," Transportation Network Companies News for Drivers, [Blog Post], September 13, 2011. http://www.tnc-drivers.org/lost-uber-web-pages.

Mac users book more expensive hotels:

- Dana Mattioli, "On Orbitz, Mac Users Steered to Pricier Hotels," *Wall Street Journal*, June 26, 2012, and updated August 23, 2012. http://online.wsj.com/article/SB10001424052702304458604577488822667325882.html.

Your inclination to buy varies by time of day:

- Esco Strong, "Search Conversion Rates by Daypart," *Atlas Digital Marketing Insight*, March 2, 2007. http://atlassolutions.com/insights.

Your e-mail address reveals your level of commitment:

- From the author's own work.

Banner ads affect you more than you think:

- Usama Fayyad, PhD, "Search Marketing and Predictive Analytics: SEM, SEO and On- Line Marketing Case Studies," Predictive Analytics World Washington, DC, Conference, October 20, 2010, Washington, DC. www.predictiveanalyticsworld.com/dc/2010/agenda.php#day2–10.

Companies win by not promoting customers to think *(Chapter 7 focuses on this topic)*:

- Jessica Tsai, "Mail Model of the Year: U.S. Bank Accrues Savings by Only Pitching Customers Who Aren't Already Likely to Show Interest," *CRM Magazine*, destinationCRM. com, April 2010. www.destinationcrm.com/Articles/Columns-Departments/REAL- ROI/Mail-Model-of-the-Year-66123.aspx.

Your web browsing reveals your intentions:

- Cell phone contract expiration example courtesy of Akin Arikan, "Multichannel Marketing Metrics with Akin." www.multichannelmetrics.com/.

Friends stick to the same cell phone company (a social effect):

- Thanks to Tim Manns. For details, see the Predictive Analytics World presentation: Tim Manns, "Know Your Customers by Knowing Who They Know, and Who They Don't: Leveraging the Power of Social Interaction." www.predictiveanalyticsworld.com/TimManns_CaseStudyOptus.pdf.
- Auren Hoffman, "Birds of a Feather Use the Same Phone Service Provider," Summation blog post, November 4, 2009. http://blog.summation.net/2009/11/birds-of-a-feather- use-the-same-phone-service-provider.html.
- Michael Driscoll, PhD, "The Social Effect: Predicting Telecom Customer Churn with Call Data," Predictive Analytics World San Francisco Conference, February 16, 2010, San Francisco, CA. www.predictiveanalyticsworld.com/sanfrancisco/2010/agenda.php#day1–12.

Transient web shoppers are less likely to return:

- Niche retail website customers who make their first purchase after typing in the website URL by hand are twice as likely to return for a second purchase in the future than those who arrived from a Google search (author's internal case study).

Table: Bizarre and Surprising Insights—Finance and Insurance (Chapter 3)

Low credit rating, more car accidents:

- Herb Weisbaum, "Insurance Firms Blasted for Credit Score Rules," ConsumerMan on NBC News, January 27, 2010. www.msnbc.msn.com/id/35103647/ns/business-consumer_news/t/insurance-firms-blasted-credit-score-rules/#.UET1ppbuiSp.
- Dan Collins, "Credit Ratings Drive Car Insurance Costs," CBS News, February 11, 2009. www.cbsnews.com/2100–18563_162–522755.html.
- Federal Trade Commission, "FTC Releases Report on Effects of Credit-based Insurance Scores," FTC News, July 24, 2007. www2.ftc.gov/opa/2007/07/facta.shtm.

Your shopping habits foretell your reliability as a debtor:

- Charles Duhigg, "What Does Your Credit-Card Company Know about You?" *New York Times*, May 12, 2009. www.nytimes.com/2009/05/17/magazine/17credit-t.html.

Typing with proper capitalization indicates creditworthiness:

- Steve Lohr, "Banking Start-Ups Adopt New Tools for Lending," *The New York Times*, January 18, 2015. http://www.nytimes.com/2015/01/19/technology/banking-start-ups-adopt-new-tools-for-lending.html.

A small business's credit risk depends on the owner's behavior as a *consumer*:

- Vernon Gerety, "Credit Scoring in the Leasing Industry," ELA Credit & Collection Management Conference, June 9–11, 2002. www.elfaonline.org/cvweb_elfa/Product_Downloads/Credit_Scoring.pdf.

Table: Bizarre and Surprising Insights—

Healthcare (Chapter 3)

Genetics foretell cheating wives:

- Christine E. Garver-Apgar, Steven W. Gangestad, Randy Thornhill, Robert D. Miller, and Jon J. Olp, "Major Histocompatibility Complex Alleles, Sexual Responsivity, and Unfaithfulness in Romantic Couples," *Psychological Science* 2006 17: 830, October 27, 2006. http://psy2.ucsd.edu/Bmgorman/garver.pdf.
- Special to *World Science*, "Gene May Help Predict Infidelity, Study Reports," *World Science*, November 30, 2006. www.world-science.net/exclusives/061128_infidelity-genes.htm.

Early retirement means earlier death:

- Andreas Kuhn, Jean-Philippe Wuellrich, and Josef Zweimüller, "Fatal Attraction? Access to Early Retirement and Mortality," IZA Discussion Paper, October 2010. www.econ.uzh.ch/faculty/kuhn/wp_499_10_2010.pdf.

Skipping breakfast correlates with coronary heart disease:

- Leah E. Cahill, PhD; Stephanie E. Chiuve, ScD; Rania A. Mekary, PhD; Majken K. Jensen, PhD; Alan J. Flint, MD, DrPh; Frank B. Hu, MD, PhD; and Eric B. Rimm, Sc, *"Prospective Study of Breakfast Eating and Incident Coronary Heart Disease in a Cohort of Male US Health Professionals,"* Circulation, American Heart Association Journals, July 23, 2013, Volume 128, Issue 4. *Circulation. 2013;128:337-343, doi:10.1161/CIRCULATIONAHA.113.001474*. http://circ.ahajournals.org/content/128/4/337.full.pdf.

Google search trends predict disease outbreaks:

- Google Trends, "Flu Trends," www.google.org/flutrends/about/how.html. See also: 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 (February 19, 2009), 1012–1014. doi:10.1038/nature07634. www.nature.com/nature/journal/v457/n7232/full/nature07634.html.

Smokers suffer less from repetitive motion disorder:

- State Compensation Insurance Fund, "Microbreaks," *Ergomatters* 3, no. 4. www.statefundca.com/safety/ErgoMatters/Microbreaks.asp.

Positive health habits are contagious (a social effect):

- Alice Park, "Quitting Smoking Is Contagious," *Time*, May 21, 2008. www.time.com/time/health/article/0%2c8599%2c1808446%2b05.html.
- Nicholas A. Christakis, MD, PhD, MPH, and James H. Fowler, PhD, "The Spread of Obesity in a Large Social Network over 32 Years," *New England Journal of Medicine* 357 (November 1, 2007): 1866–1868. www.nejm.org/doi/full/10.1056/NEJMsa066082#t5letters.

Happiness is contagious (a social effect):

- Nicholas A. Christakis and James H. Fowler, "Social Networks and Happiness," *Edge*, December 4, 2008. www.edge.org/3rd_culture/christakis_fowler08/christakis_fowler08_index.html.
- Allison Aubrey, "Happiness: It Really Is Contagious," NPR News, *All Things Considered*, December 5, 2008. www.npr.org/templates/story/story.php?storyId597831171.

Knee surgery choices make a big difference:

- Lars Ejerhed, MD, Jüri Kartus, MD, PhD, Ninni Sernert, Kristina Köhler, and Jon Karlsson, MD, PhD, "Patellar Tendon or Semitendinosus Tendon Autografts for Anterior Cruciate Ligament Reconstruction," *American Journal of Sports Medicine* 31, no. 1 (January 2003): 19–25. http://ajs.sagepub.com/content/31/1/19.short.

Music expedites poststroke recovery and improves mood:

- T. Särkämö, M. Tervaniemi, S. Laitinen, A. Forsblom, S. Soinila, M. Mikkonen, T. Autti, H. M. Silvennoinen, J. Erkkilä, M. Laine, I. Peretz, and M. Hietanen, "Music Listening Enhances Cognitive Recovery and Mood after Middle Cerebral Artery Stroke," *NCBI*, March 2008, 131 (Part 3), 866–876. www.ncbi.nlm.nih.gov/pubmed/18287122.

Yoga improves your mood:

- Kazufumi Yoshihara, Tetsuya Hiramoto, Nobuyuki Sudo, and Chiharu Kubo, "Profile of Mood States and Stress-Related Biochemical Indices in Long-Term Yoga Practitioners," *NCBI*, *BioPsychoSocial Medicine* 5, no. 6 (2011). Published online June

3, 2011, doi:10.1186/1751-0759-5-6. www.ncbi.nlm.nih.gov/pmc/articles/PMC3125330/.

- Similar results apply also with other athletic activities: Oxford Food & Fitness Dictionary, www.answers.com/topic/profile-of-mood-states.

Table: Bizarre and Surprising Insights—Crime and Law Enforcement (Chapter 3)

Suicide bombers do not buy life insurance:

- Stephen J. Dubner, "SuperFreakonomics Book Club: Ian Horsley Answers Your Questions about the Terrorist Algorithm," September 1, 2010. www.freakonomics.com/2010/09/01/superfreakonomics-book-club-ian-horsley-answers-your-questions- about-the-terrorist-algorithm/.

Unlike lightning, crime strikes twice:

- G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self- Exciting Point Process Modeling of Crime," *Journal of the American Statistical Association* 106, no. 493 (March 2011), Applications and Case Studies, doi:10.1198/jasa.2011.ap09546. www.math.ucla.edu/Bmbshort/papers/crime3.pdf.

Crime rises with public sporting events:

- Daniel I. Rees and Kevin T. Schnepel, "College Football Games and Crime," Working Papers Series, Department of Economics, University of Colorado, Denver, January 2008. www.ilr.cornell.edu/cheri/workingPapers/upload/cheri_wp109.pdf.

Crime rises after elections:

- Katherine Wells, "Wildfires, Cops and Keggers: A New Marketplace Podcast on Election- Time Mischief," Freakonomics Radio, November 2, 2011. www.freakonomics.com/ 2011/11/02/wildfires-cops-and-keggers-a-new-marketplace-podcast-on-election-time- mischief/.

Phone card sales predict Congo massacres:

- Quentin Hardy, "Bizarre Insights from Big Data," *New York Times*, Bits blog, March 28, 2012.

http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/.

Hungry judges rule negatively:

- Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso, "Extraneous Factors in Judicial Decisions," edited by Daniel Kahneman, Princeton University, Princeton, NJ, February 25, 2011. http://lsolum.typepad.com/files/danziger-levav-avnaim-pnas-2011.pdf and www.pnas.org/content/108/17/6889.

Table: Bizarre and Surprising Insights—Miscellaneous (Chapter 3)

Music taste predicts political affiliation:

- Brian Whitman, "How Well Does Music Predict Your Politics?" Variogr.am blog post, July 12, 2012. http://notes.variogr.am/post/26869688460/how-well-does-music-predict- your-politics.

Online dating: Be cool and unreligious to succeed:

- Christian Rudder, "Exactly What to Say in a First Message," OkTrends, OkCupid, September 14, 2009. http://blog.okcupid.com/index.php/online-dating-advice-exactly-what-to-say-in-a-first-message/.

Hot or not? People consistently considered attractive get *less* attention:

- Christian Rudder, OkTrends, "Surprising Statistics about Hot People versus Ugly People," Business Insider, January 13, 2011. www.businessinsider.com/surprising-statistics- about-hot-people-versus-ugly-people-2011–1.

Users of the Chrome and Firefox browsers make better employees *(Thanks to Michael Housman, Chief Analytics Officer at Cornerstone OnDemand, Inc., for additional information regarding this example.)***:**

- Stephen J. Dubner, "The Maddest Men of All: A New Freakonomics Radio Podcast," Freakonomics Online, February 26, 2015. http://freakonomics.com/2015/02/26/the-maddest-men-of-all-a-new-freakonomics-radio-podcast/.
- E.H., "How might your choice of browser affect your job prospects?" *The Economist*, April 10, 2013. http://www.economist.com/blogs/economist-

explains/2013/04/economist-explains-how-browser-affects-job-prospects.

A job promotion can lead to quitting (see Chapter 2 for details on this case study):

- Gitali Halder, Hewlett-Packard, and Anindya Dey, Hewlett Packard, "Attrition Driver Analysis," Predictive Analytics World London Conference, November 30, 2011, London, UK. www.predictiveanalyticsworld.com/london/2011/agenda.php#day1–5a.

More engaged employees have fewer accidents (Shell):

- Thanks to Esther Bongenaar, Lead HR Analytics at Shell International, for this example.

Higher status, less polite:

- Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts, *"A Computational Approach to Politeness with Application to Social Factors,"* Cornell University Library, June 25, 2013, arXiv:1306.6078 [cs.CL]. http://arxiv.org/abs/1306.6078 *and* http://www.mpi-sws.org/~cristian/Politeness.html.

See also, the determination of controversial Wikipedia articles:

- Taha Yasseri, Anselm Spoerri, Mark Graham, and János Kertész, *"The most controversial topics in Wikipedia: A multilingual and geographical analysis"* Cornell University Library, May 23, 2013, arXiv:1305.5566 [physics.soc-ph]. http://arxiv.org/abs/1305.5566.

Vegetarians miss fewer flights:

- *The Economist* Staff, "A Different Game: Information Is Transforming Traditional Businesses," *The Economist*, February 25, 2010. www.economist.com/node/15557465.
- Quentin Hardy, "Bizarre Insights from Big Data," *New York Times*, Bits blog, March 28, 2012. http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/.

Smart people like curly fries:

- Michal Kosinski, David Stillwell, and Thore Graepel, "*Private traits and attributes are predictable records of human behavior*," PNAS Online (Proceedings of the National Academy of Sciences

of the United States), April 9, 2013, Vol. 110, No. 15. http://www.pnas.org/content/110/15/5802.abstract.

- Jennifer Golbeck, "The Curly Fry Conundrum: Why social media "likes" say more than you might think," TEDtalks Online, October, 2013. http://www.ted.com/talks/jennifer_golbeck_the_curly_fry_conundrum_why_social_media_likes_say_more_than_you_might_think.

A photo's quality is predictable from its caption:

- Quentin Hardy, "Bizarre Insights from Big Data," *New York Times*, Bits blog, March 28, 2012. http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/.

Female-named hurricanes kill more people than male hurricanes:

- Kiju Jung, Sharon Shavitt, Madhu Viswanathan, and Joseph M. Hilbe, *"Female hurricanes are deadlier than male hurricanes,"* PNAS Online (Proceedings of the National Academy of Sciences of the United States), June 17, 2014, Vol. 111, No. 24. http://www.pnas.org/content/111/24/8782.abstract.
- Jason Samenow, "Female-named hurricanes kill more than male hurricanes because people don't respect them, study finds," *The Washington Post*, June 2, 2014. http://www.washingtonpost.com/blogs/capital-weather-gang/wp/2014/06/02/female-named-hurricanes-kill-more-than-male-because-people-dont-respect-them-study-finds/.

Men on the *Titanic* faced much greater risk than women:

- Kaggle. "Titanic: Machine Learning from Disaster" Competition, September 28, 2012. www.kaggle.com/c/titanic-gettingStarted.

Rock Star Mortality:

- Kim Painter, "Solo Rock Stars at Greater Risk to Die Early," *USA Today, Health and Wellness*, December 19, 2012. www.usatoday.com/story/news/nation/2012/12/19/rock-star-early-death-solo-artists/1780027.
- Mark A. Bellis, Karen Hughes, Olivia Sharples, Tom Hennell, and Katherine A. Hardcastle, "Dying to Be Famous: Retrospective Cohort Study of Rock and Pop Star Mortality and its Association with Adverse Childhood Experiences," *BMJ Open*, *BMJ Group*, December 19, 2012, Volume 2, Issue 6,

2:e002089 doi:10.1136/bmjopen-2012-002089. http://bmjopen.bmj.com/content/2/6/e002089.full?sid58f58c788-57df-4d54-a256-e9a6785f831c.

Chapter 3, Continued (after its Tables):

Shark attacks increase when ice cream sales increase:

- Story from BBC News, "Just Because," *BBC News Magazine*, September 3, 2008. http://news.bbc.co.uk/2/hi/uk_news/magazine/7592579.stm.

Regarding hormone replacement therapy:

- Debbie A. Lawlor, George Davey Smith, and Shah Ebrahim, "Commentary: The Hormone Replacement–Coronary Heart Disease Conundrum: Is This the Death of Observational Epidemiology?" *International Journal of Epidemiology* 33, no. 3 (2004): 464–467, doi:10.1093/ije/dyh124. First published online: May 27, 2004. http://ije.oxfordjournals.org/content/33/3/464.

Regarding the growing number of retracted journal publications:

- G. Naik, "Mistakes in Scientific Studies Surge," *Wall Street Journal*, August 10, 2011. http://www.wsj.com/articles/SB10001424052702303627104576411850666582080.

Does smoking cause cancer?:

- David Salsburg, *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century* (W.H. Freeman & Company, 2001), 183.

The argument that, although we "cannot rely on correlation alone... insisting on absolute proof of causation is too exacting a standard":

- Tim Harford, "Cigarettes, Damn Cigarettes and Statistics," *The Financial Times*, April 10, 2015. http://timharford.com/2015/04/cigarettes-damn-cigarettes-and-statistics/.

More insight regarding inferring causality:

- Adam Kelleher, "If Correlation Doesn't Imply Causation, Then What Does?" FreeCodeCamp, (blog post) June 27, 2016.

https://medium.freecodecamp.com/if-correlation-doesnt-imply-causation-then-what-does-c74f20d26438#.xwbn7t3fe.

Further Reading on Vast Search (final portion of Chapter 3):

NOTE: ***Vast search and statistical significance are fundamental yet elusive and far-reaching topics—if you are considering diving into the following citations for more depth on these issues and the statistical methods that address them, I suggest reading all of this book's Chapters 3 and 4 first, since they provide the broader perspective and framework within which to understand the more technical aspects.***

Media that reported an orange used car is least likely to be a lemon:

- Marcus Wohlsen, "San Francisco startup makes data science a sport." *The Seattle Times*, April 14, 2012. http://old.seattletimes.com/html/businesstechnology/2017983961_apusscienceassport.html.
- Viktor Mayer-Schönberger, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Eamon Dolan/Mariner Books; Reprint Edition, March 4, 2014).
- Marcus Wohlsen, "Kaggle Competitions Make Data Science A Sport," *The Huffington Post*, April 15, 2012. www.huffingtonpost.com/2012/04/15/kaggle-competitions-data-science_n_1427204.html.
- Deborah Gage, "Big Data Uncovers Some Weird Correlations," *The Wall Street Journal*, March 23, 2014. www.wsj.com/articles/SB10001424052702303369904579423132072969654.
- Quentin Hardy, "Bizarre Insights From Big Data," *The New York Times*, March 28, 2012. http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/?_r=0
- NPR Morning Edition, "The 'Big Data' Revolution: How Number Crunchers Can Predict Our Lives," National Public Radio Online, March 7, 2013. www.npr.org/2013/03/07/173176488/the-big-data-revolution-how-number-crunchers-can-predict-our-lives.

The competition and public data set for predicting bad buys among used cars, sponsored by Carvana, a data driven company that modernizes onine used car buying:

- Kaggle, "Predict if a car purchased at auction is a lemon," Competition, September 30, 2011. www.kaggle.com/c/DontGetKicked.

The Nassim Taleb book quoted and mentioned:

- Nassim Nicholas Taleb, *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets* (Random House, 2004).

Nate Silver on the peril of big data:

- Stephen J. Dubner, "Nate Silver Says: "Everyone Is Kind of Weird: A New Freakonomics Radio Podcast," Freakonomics.com Podcast, April 23, 2015. http://freakonomics.com/2015/04/23/nate-silver-says-everyone-is-kind-of-weird-a-new-freakonomics-radio-podcast/.

Here is where to read more on *target shuffling*, the simulation-based method employed to reassess the significance of the "orange lemons" claim. Target shuffling was also the method John Elder employed to gauge the system integrity of his stock market prediction system described in Chapter 1. It validated the system's robustness, informing the decisions to deploy it, keep it running for years, and eventually to shut it down:

- John Elder and Ben Bullard, "Are Orange Cars Really not Lemons?" Elder Research, Inc., April 2013. http://datamininglab.com/orange-car.

Video overviews of target shuffling:

- Dr. John Elder, "Target Shuffling Video," Elder Research, 2016. http://www.elderresearch.com/target-shuffling-video.
- Dr. John Elder, CEO and Founder, Elder Research, Inc., "Video: The Power (and Peril) of Predictive Analytics," Predictive Analytics World for Business Boston Conference, October 5-9, 2014, Boston, MA.

http://www.predictiveanalyticsworld.com/patimes/video-the-power-and-peril-of-predictive-analytics/.

- "HBO Teaches You How to Avoid Bad Science," *Predictive Analytics Times*, July 7, 2016 (includes *HBO's* "Last Week Tonight with John Oliver" video). http://www.predictiveanalyticsworld.com/patimes/hbo-teaches-you-how-to-avoid-bad-science/7753/.

More on the tribulations of p-values - "If you observe a P value close to 0.05, your false discovery rate will not be 5%. It will be at least 30% and it could easily be 80% for small studies.":

- David Colquhoun, "On the hazards of significance testing. Part 2: the false discovery rate, or how not to make a fool of yourself with ***P*** values," DC's Improbably Science [Blog Post], March 24, 2014. www.dcscience.net/2014/03/24/on-the-hazards-of-significance-testing-part-2-the-false-discovery-rate-or-how-not-to-make-a-fool-of-yourself-with-p-values/.

How common (and, perhaps, easy) it is to confuse the specific meaning of p-value:

- Andrew Gelman, "A bad definition of statistical significance from the U.S. Department of Health and Human Services, Effective Health Care Program," Statistical Modeling, Causal Inference, and Social Science blog, July 21, 2015. http://andrewgelman.com/2015/07/21/a-bad-definition-of-statistical-significance-from-the-u-s-department-of-health-and-human-services-effective-health-care-program/

One and a half minute video visually explaining the problem of vast search:

- P.T.S, K.N.C, and O.M., "Unlikely results," *The Economist*, October 21, 2013. www.economist.com/blogs/graphicdetail/2013/10/daily-chart-2.

Refuted finding: Childhood leukemia was four times higher among those that lived closest to high-voltage power lines:

- Catharine Paddock, Ph.D., “No Childhood Leukemia Risk from Power Lines,” *Medical News Today*, February 7, 2014. http://www.medicalnewstoday.com/articles/272330.php.

More entertaining spurious correlations (albeit within an article discussing correlation versus causation):

- Statwolf, “Does Eating Cheese Make You More Likely to Die in Your Sleep? The Risks in Inferring From Data,” Statwolf.com (blog post), September 23, 2016. http://www.statwolf.com/blog/does-eating-cheese-make-you-more-likely-to-die-in-your-sleep-the-risks-in-inferring-from-data.

"Even if something is published in the flagship journal of the leading association of research psychologists, there's no reason to believe it. The system of scientific publication is set up to encourage publication of spurious findings." (Andrew Gelman, professor of statistics and political science at Columbia University):

- Andrew Gelman, “Too Good to Be True,” *Slate*, July 24, 2013. http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_multiple_comparisons_give_spurious_results.html.

More examples of spurious claims:

- Andrew Gelman and Kaiser Fung, “The Power of the “Power Pose,”” *Slate*, Jan 19, 2016. http://www.slate.com/articles/health_and_science/science/2016/01/amy_cuddy_s_power_pose_research_is_the_latest_example_of_scientific_overreach.html.
- Victoria Woollaston, “Does Sour Cream Cause Bike Accidents? No, but it LOOKS like it does: Hilarious graphs reveal how statistics can create false connections,” *DailyMail*, May 28, 2014. http://www.dailymail.co.uk/sciencetech/article-2640550/Does-sour-cream-cause-bike-accidents-No-looks-like-does-Graphs-reveal-statistics-produce-false-connections.html

- Ky Harlin, "The 10 Most Bizarre Correlations," *BuzzFeed*, April 11, 2013. http://www.buzzfeed.com/kjh2110/the-10-most-bizarre-correlations

After 150 Years, the American Statistical Association Says No to p-values:

- Dr. Norman Matloff, "After 150 Years, the American Statistical Association Says No to p-values," KDNuggets post, March 2016. http://www.kdnuggets.com/2016/03/asa-says-no-p-values.html.

FiveThirtyEight on vast search (P-hacking) - interactive demo:

- Christie Aschwanden, "Science Isn't Broken," *FiveThirtyEightScience*, August 19, 2015. http://fivethirtyeight.com/features/science-isnt-broken/#part1.

Here is a somewhat random sampling of just a few of the many assorted sources explaining and exploring the vast search issue as it arises across assorted industries/fields:

- Graphpad Statistics Guide, "Multiple Comparisons Traps," Graphpad Online, February 24, 2015. www.graphpad.com/guides/prism/6/statistics/index.htm?stat_ multiple_ comparisons_ traps.htm.
- David Lane, "The Problem of Too Many Statistical Tests: Subgroup Analyses in a Study Comparing the Effectiveness of Online and Live Lectures," *Numeracy,* Vol. 6: Issue 1, Article 7 (2013). DOI: http://dx.doi.org/10.5038/1936-4660.6.1.7. http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1118&context=numeracy.
- John H. McDonald, "Multiple Comparisons," Handbook of Biological Statistics [Blog Post], July 20, 2015. www.biostathandbook.com/multiplecomparisons.html.
- Lemuel A. Moyé, *Multiple Analyses in Clinical Trials: Fundamentals for Investigators (Statistics for Biology and Health),* (Springer-Verlag, 2003).
- Pedro R. Peres-Neto, "How many statistical tests are too many? The problem of conducting multiple ecological inferences revisited," *Marine Ecology Progress Series,* Vol.

176, pages 303-306, 1999. www.int-res.com/articles/meps/176/m176p303.pdf.

CHAPTER 4: The Machine That Learns: A Look inside Chase's Prediction of Mortgage Risk *(modeling)*

Some content of this chapter originated from valuable interviews with Dan Steinberg, PhD, Chief Executive Officer and President of Salford Systems.

For further technical details behind this chapter's Chase case study:

- Ali Moazami and Shaolin Li, "Mortgage Business Transformation Program Using CART-based Joint Risk Modeling," Salford Systems Data Mining Conference, 2005. www.salford-systems.com/doc/MoazamiLi.pdf.

For Further Technical Reading on Predictive Modeling:

For a technical introduction to predictive modeling methods:

- Robert Nisbit, John Elder, and Gary Miner, *Handbook of Statistical Analysis and Data Mining Applications* (Academic Press, 2009), 347.

For an academic textbook with the complete probability and math:

- Tom M. Mitchell, *Machine Learning*, Science/Engineering/Math (McGraw-Hill, 1997).

For a visual, conceptual breakdown of modeling methods that's clickable:

- Saed Sayad, PhD, "An Introduction to Data Mining." http://chem-eng.utoronto.ca/B datamining/dmc/data_mining_map.htm.

Merger that made Chase the largest U.S. bank holding company in 1996:

- JPMorgan Chase & Co. website, "History of Our Firm." www.jpmorganchase.com/corporate/About-JPMC/jpmorgan-history.htm.

The history of risk scoring:

- Vitalie BUMACOV and Arvind ASHTA, "The Conceptual Framework of Credit Scoring from Its Origins to Microfinance (Draft)," Burgundy School of Business, June 2, 2011. http://tinyurl.com/c7fs9cz.
- David Durand, *Risk Elements in Consumer Instalment Financing* (National Bureau of Economic Research, 1941), ISBN: 0-870-14124-4. www.nber.org/books/dura41–1.

A perspective on micro-risk management:

- James Taylor, JT on EDM. "Risk by Risk—A Decision-Centric Approach to Risk Management," DecisionManagementSolutions.com blog post, February 15, 2010. http:// jtonedm.com/2010/02/15/risk-by-risk-a-decision-centric-approach-to-risk-management/.
- James Taylor with Neil Raden, *Smart (Enough) Systems* (Prentice Hall, 2007).
- James Taylor. *Decision Management Systems* (IBM Press, 2011).

Insurance companies using predictive modeling:

- Dr. Thomas Ashley, Dr. Guizhou Hu, and Chris Stehno, "How Does Mortality Modeling Affect Life Insurance Preferred Risk Selection?" ReFocus Conference, Las Vegas, NV, February 27–March 2, 2011. www.soa.org/files/pd/2011-lv-refocus-b3.pdf.

Predictive modeling and insurance companies:

- *Insurance Journal* Staff, "How Predictive Modeling Has Revolutionized Insurance," *Insurance Journal*, June 18, 2012. www.insurancejournal.com/news/national/2012/06/18/251957.htm.

Insurer's use of predictive analytics:

- Charles Nyce, PhD, CPCU, API, "Predictive Analytics White Paper," AICPACU/IIA, September 28, 2007. http://www.hedgechatter.com/wp-content/uploads/2014/09/predictivemodelingwhitepaper.pdf.

Eric Webster quote:

- Gregory Piatetsky-Shapiro, "Discussion with State Farm's Eric Webster: Insurance and Data Mining," KDnuggets.com, April 15, 2009. www.kdnuggets.com/news/2009/n08/ 3i.html.

Stanford's free machine learning course; its Professor Andrew Ng founded Coursera:

- Terence Chea, "MOOCs from Elite Colleges Transform Online Higher Education," *Huffington Post Online*, August 5, 2012. www.huffingtonpost.com/2012/08/05/mooc- massive-open-online-courses_n_1744430.html.
- Andrew Ng, "The Online Revolution: High-Quality Education for Everyone," ACM San Francisco Bay Area Professional Chapter, Cupertino, CA, May 16, 2012. www.sfbayacm.org/event/online-revolution-high-quality-education-everyone.

Polls showing decision tree popularity among predictive modeling methods:

- KDnuggets poll: Algorithms for data analysis/data mining, November 2011. www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html.
- Rexer Analytics: Data Miner Survey, August 2012. www.predictiveanalyticsworld.com/ survey-signup.php.

The 30-second rule:

- Audrey Fukuman and Andy Wright, "You Dropped Food on the Floor—Do You Eat It? The 30-Second Rule: A Decision Tree," SFFoodie, blogs.SFWeekly.com, January 29, 2010. http://blogs.sfweekly.com/foodie/2010/01/you_dropped_food_on_the_floor.php.

Decision tree trained to predict votes on U.S. Supreme Court rulings by former Justice Sandra Day O'Connor:

- Andres D. Martin, Kevin M. Quinn, Theodore W. Ruger, and Pauline T. Kim, "Competing Approaches to Predicting Supreme Court Decision Making," *Symposium: Forecasting U.S. Supreme Court Decisions* 2, no. 4 (December 2004), 761–767. http:// wusct.wustl.edu/media/man1.pdf.
- Dylan Matthews, "This computer program can predict 7 out of 10 Supreme Court decisions," *Vox Media*, August 4, 2014. www.vox.com/2014/8/4/5967147/how-a-computer-model-got-to-predict-70-of-supreme-court-decisions.
- Daniel Martin Katz, Michael James Bommarito, II and Josh Blackman, "Predicting the Behavior of the Supreme Court of the

United States: A General Approach," Social Science Research Network, July 21, 2014. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2463244.

Oliver G. Selfridge quote:

- Marti A. Hearst Haym Hirsh, "AI's Greatest Trends and Controversies," *IEEE Intelligent Systems* 15, issue 1 (January 2000), 8–17, doi:10.1109/5254.820322. http://dl.acm.org/citation.cfm?id5630511.

For an example of a visual, rich decision tree to play the guessing game:

- Shana Mlawski and Carlos A. Hann Commander, "Female Character Flowchart," Overthinkingit.com, October 12, 2010. http://big.assets.huffingtonpost.com/flowchart.jpeg.

This New York Times article addresses what most predictive analytics practitioners refer to as data preparation. Although it uses other terms to refer to that, it draws a good birds-eye-view vantage on the topic, and covers an arena of emerging start-up companies aiming to focus on it:

- Steve Lohr, "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights," *The New York Times,* August 17, 2014. www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

Butter in Bangladesh:

- Jason Zweig, "Data Mining Isn't a Good Bet for Stock-Market Predictions," *Wall Street Journal*, August 8, 2009. http://online.wsj.com/article/SB124967937642715417.html.
- Jason Zweig interviews David Leinweber, author of *Nerds on Wall Street*. Live: The *Wall Street Journal*, August 7, 2009. http://live.wsj.com/video/an-interview-with-nerds-on- wall-street-author/5062DA68-FCF6–42AC-AC62-AE6046BA40AC.html#!5062DA68-FCF6–42AC-AC62-AE6046BA40AC.
- Laura Washington, "What's the Stock Market Got to Do with the Production of Butter in Bangladesh?" *CNNMoney Magazine*, March 1, 1998. http://money.cnn.com/magazines/moneymag/moneymag_archive/1998/03/01/238606/index.htm.

Buried treasure discoveries predicted the Dow Jones Industrial Average:

- David Leinweber, "Back into the Data Mine," Nerds on Wall Street blog post, August 2009. http://nerdsonwallstreet.typepad.com/my_weblog/2009/08/back-into- the-data-mine.html.

More humorous examples of spurious and misunderstood correlations:

- Peter Bruce, "Swimming Pools and Nicholas Cage: A Look at How We Misunderstand Correlation," Data Science Central Blog, February 9, 2015. www.datasciencecentral.com/m/blogpost?id=6448529%3ABlogPost%3A255381.
- Tyler Vigen, *Spurious Correlations*, (Hachette Books, 2015). www.tylervigen.com/spurious-correlations.
- Tyler Vigen, "Number of people who drowned by falling into a swimming-pool correlates with Number of films Nicolas Cage appeared in," Spurious Correlations blog. www.tylervigen.com/view_correlation?id=359.

More examples of spurious discoveries being detected after publication:

- Michael Walker, "Bad Data Science and Woody Allen," Data Science Central blog, May 12, 2014. www.datasciencecentral.com/profiles/blogs/bad-data-science-and-woody-allen.

Another related example of overlearning:

- Mike Moffatt, "Does the Superbowl Predict Economic Growth?" About.com Economics, April 10, 2012. http://economics.about.com/od/economicindicatorintro/a/superbowl.htm.

Quote by Nassim Nicholas Taleb:

- Nassim Nicholas Taleb, *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*, 2nd ed. (Random House Trade Paperbacks, 2005).

Regarding evolving Tetris players:

- Eric V. Siegel and Alexander D. Chaffee, "Genetically Optimizing the Speed of Programs Evolved to Play Tetris," in *Advances in Genetic Programming*, ed. Peter J. Angeline and Kenneth E. Kinnear Jr. (MIT Press, 1996), 279–298, ISBN: 0-262-01158-1. http://dl.acm.org/citation.cfm?id5270220.

Inductive bias:

- Tom M. Mitchell, "The Need for Biases in Learning Generalizations," CBM-TR 5–110, Rutgers University, New Brunswick, NJ. Converted to electronic version by Roby Joehanes, Kansas State University, November 8, 2007. http://dml.cs.byu.edu/Bcgc/docs/mldm_tools/Reading/Need%20for%20Bias.pdf.

No free lunch theorem:

- David H. Wolpert, "The Lack of a Priori Distinctions between Learning Algorithms," *Neural Computations* 8, issue 7 (October 1, 1996): 1341–1390. http://dl.acm.org/citation.cfm?id51362128.

Early work in decision trees:

- James N. Morgan and John A Sonquist, "Problems in the Analysis of Survey Data, and a Proposal," *Journal of the American Statistical Association* 58, issue 202 (June 1963): 415–434. www.cs.nyu.edu/Broweis/csc2515–2006/readings/morgan_sonquist63.pdf.

Decision tree overlearning figures derived from Figure 3.6 of:

- Tom M. Mitchell, *Machine Learning*, Science/Engineering/Math (McGraw-Hill, 1997).

NOTE: Here is a gotcha/pitfall regarding the use of a test set. If you repeat the train/test cycle enough times, the test set effectively serves as training data, since the "best" model is the one that did best on the test data. Therefore, that test set is no longer unseen and unbiased. To put it another way, with enough iterations, you could be conducting* vast search *on the test set instead of only the training set. In this case, either a new, truly unseen dataset may be used to validate, or we must turn to an evaluation method that does take vast search explicitly into consideration, such as* target shuffling*, referenced in the prior chapter.

Classification and regression trees:

- Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen, *Classification and Regression Trees* (Chapman & Hall/CRC, 1984).

The history of CART:

- Ronny Kohavi, "Crossing the Chasm: From Academic Machine Learning to Commercial Data Mining," International Conference on Machine Learning 1998, Slide 21. http:// robotics.stanford.edu/Bronnyk/chasm.pdf.

Chase's acquisition of JPMorgan:

- Andrew Garfield, "Chase Acquires JPMorgan for $34 bn, but 10,000 Jobs May Go," *The Independent*, September 14, 2000. www.independent.co.uk/news/business/news/chase- acquires-jp-morgan-for-34bn-but-10000-jobs-may-go-699819.html.

JPMorgan Chase as the largest bank in the United States:

- Hugh Son, "BofA Loses No. 1 Ranking by Assets to JPMorgan as Chief Moynihan Retreats," Bloomberg, October 18, 2011. www.bloomberg.com/news/2011–10–18/ bank-of-america-loses-no-1-ranking-by-assets-to-jpmorgan-chase.html.
- Halah Touryalai, "America's Biggest Banks: Bank of America Fails to Win Back Top Spot," *Forbes*, March 26, 2012. www.forbes.com/sites/halahtouryalai/2012/03/26/ americas-biggest-banks-bank-of-america-fails-to-win-back-top-spot/.

Regarding the macroscopic level and the failures leading to the 2008 recession:

- Saul Hansell, "How Wall Street Lied to Its Computers," *New York Times*, Bits blog, September 18, 2008. http://bits.blogs.nytimes.com/2008/09/18/how-wall-streets-quants-lied-to-their-computers/.

Quote from Bart Baesens:

- Bart Baesens, PhD, "Building Bulletproof Models," *sascom* Magazine, 3rd quarter, 2010. www.sas.com/news/feature/FSmodels.html.

CHAPTER 5: The Ensemble Effect: Netflix, Crowdsourcing, and

Supercharging Prediction *(ensembles)*

Layperson competitors for the Netflix Prize:

- Eric Siegel, PhD, "Casual Rocket Scientists: An Interview with a Layman Leading the Netflix Prize, Martin Chabbert," September 2009. www.predictiveanalyticsworld.com/ layman-netflix-leader.php.

$1 million Netflix Prize:

- Netflix Prize, September 21, 2009. www.netflixprize.com/.

Seventy percent of Netflix movie choices based on recommendations:

- Jeffrey M. O'Brien, "The Netflix Effect," *Wired Magazine Online*, December 12, 2002. http://www.wired.com/wired/archive/10.12/netflix.html.
- Michael Liedtke, "Netflix Recommendations Are About to Get Better, Say Execs," *Huffington Post Online*, April 9, 2012. www.huffingtonpost.com/2012/04/09/netflix-recommendations_n_1413179.html.

Predicting movie ratings and recommender systems:

- Arkadiusz Paterek, "Predicting Movie Ratings and Recommender Systems," PDF file, June 19, 2012. http://arek-paterek.com/book/.

Netflix Prize team BellKor's Pragmatic Chaos:

- "BellKor's Pragmatic Chaos Is the Winner of the $1 Million Netflix Prize!!!!" September 17, 2009. www2.research.att.com/Bvolinsky/netflix/bpc.html.

Regarding SpaceShipOne and the XPrize:

- XPrize Foundation, "Ansari X Prize," XPrize Foundation, updated April 25, 2012. http://space.xprize.org/ansari-x-prize.

Netflix Prize team PragmaticTheory:

- PragmaticTheory website. https://sites.google.com/site/pragmatictheory/.

Netflix Prize team BigChaos:

- Istvan Pilaszy, "Lessons That We Learned from the Netflix Prize," Predictive Analytics World Washington, DC, Conference, October 21, 2009, Washington, DC. www.predictiveanalyticsworld.com/dc/2009/agenda.php#day2–13.
- Clive Thompson, "If You Liked This, You're Sure to Love That," *New York Times*, November 21, 2008. www.nytimes.com/2008/11/23/magazine/23Netflix-t.html.

Netflix Prize team The Ensemble:

- Blog post by Aron, "Netflix Prize Conclusion," The Ensemble, September 22, 2009. www.the-ensemble.com/content/netflix-prize-conclusion#comments.

Recount of the dead heat completion of the Netflix Prize:

- Gregory Piatetsky-Shapiro, "Netflix Prize Contest Dramatic Finish—and the Winner Is...," *KDnuggets News*, July 28, 2009. www.kdnuggets.com/news/2009/n14/1i.html.
- PragmaticTheory blog post, "By a Nose...or Is It a Hair..." PragmaticTheory blog, July 28, 2009. http://pragmatictheory.blogspot.com/2009/07/by-nose-or-is-it-hair_28.html.
- Blog post by Aron, "Final Submission Countdown," The Ensemble, July 29, 2009. www.the-ensemble.com/content/final-submission-countdown.

Netflix Prize Team Just a Guy in a Garage:

- Gavin, "Supercomputing Made Super Easy," Just a Guy in a Garage blog, July 14, 2011. http://justaguyinagarage.blogspot.com.

Contest to automatically grade student-written essays:

- Kaggle, "Develop an Automated Scoring Algorithm for Student-Written Essays," Competition, February 10, 2012. www.kaggle.com/c/asap-aes.
- Mark D. Shermis and Ben Hamner, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis," Contrasting Essay Scoring, April 12, 2012. http://dl.dropbox.com/u/44416236/NCME%202012%20Paper3_29_12.pdf.

Contest to predict the distribution of dark matter in the universe:

- Kaggle, "Mapping Dark Matter," Competition, May 23, 2011. www.kaggle.com/c/mdm.

Heritage Health Prize:

- Kaggle, "Improve Healthcare, Win $3,000,000," Competition, April 4, 2011. www.kaggle.com/host/casestudies/hpn.

More Kaggle case studies:

- Anthony Goldbloom, "Machines Learn, but Can They Teach?" Predictive Analytics World Boston Conference, October 2, 2012, Boston, MA. www.predictiveanalyticsworld.com/boston/2012/agenda.php#day2–445a.
- Anthony Goldbloom, "Prediction Competitions: Far More Than Just a Bit of Fun," Predictive Analytics World London Conference, November 16, 2010, London, UK. www.predictiveanalyticsworld.com/london/2010/agenda.php#day2–11.
- Karthik Sethuraman, Kaggle, and Diederik Van Liere, "Solving the Last Mile: Focusing Global Intelligence on Your Data," Predictive Analytics World San Francisco Conference, March 5, 2012, San Francisco, CA. www.predictiveanalyticsworld.com/sanfrancisco/2012/agenda.php#day1–7.
- Mehul Patel, Kaggle, and Giles Pavey, "Predicting Retail Customer Behaviour," Predictive Analytics World London Conference, December 1, 2011, London, UK. www.predictiveanalyticsworld.com/london/2011/agenda.php#day1–16a.

Crowdsourcing in general, beyond analytics projects:

- Jeff Howe, *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business* (Three Rivers Press, 2008).

Quote from Anthony Goldbloom about Kaggle's crowdsourcing:

- Tanya Ha, "Lucrative Algorithms," Catalyst Online, August 18, 2011. www.abc.net.au/ catalyst/stories/3296837.htm.

Regarding the shortage of analytics experts:

- James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, May 2011. www.mckinsey.com/ Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

More on Kaggle and crowdsourcing predictive analytics:

- Kaggle, "About Us: Our Team," www.kaggle.com/about.
- Karthik Sethuraman, Kaggle, "Crowdsourcing Predictive Analytics: Why 25,000 Heads Are Better Than One," Predictive Analytics World Chicago Conference, June 25, 2012, Chicago, IL. www.predictiveanalyticsworld.com/chicago/2012/agenda.php#day1– 250a.
- Beth Schultz, "Kaggle Makes Sport of Predictive Analytics," All Analytics Online, October 7, 2011. www.allanalytics.com/author.asp?section_id51411&doc_id5234238.
- Beth Schultz, "Challengers & Competitors Get Their Predictive Analytics Games On," All Analytics Online, October 10, 2011. www.allanalytics.com/author.asp?section_id51411&doc_id5234275.
- Anna Brown, "Competition Breeds the Best in Analytics," SAS Voices blog, October 21, 2011. http://blogs.sas.com/content/sascom/2011/10/21/competition-breeds-the-best- in-analytics/.
- Quentin Hardy, "Bizarre Insights from Big Data," *New York Times*, Bits blog, March 28, 2012. http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/.

For other data mining competitions, see:

- KDnuggets: Analytics, Data Mining Competitions. www.kdnuggets.com/competitions/ index.html.
- DrivenData: Data science competitions to save the world—competitions that address the world's biggest social challenges. http://www.drivendata.org/
- Crowdsourcing for Search and Data Mining (CSDM 2011): A workshop of the Fourth ACM International Conference on Web Search and Data Mining (WSDM 2011), Hong Kong, China, February 9, 2011. http://ir.ischool.utexas.edu/csdm2011/.
- CSDM 2011 Crowdsourcing for Search and Data Mining, Hong Kong, China. http://ir.ischool.utexas.edu/csdm2011/.
- CrowdANALYTIX. www.crowdanalytix.com/welcome.
- Netflix Prize, September 21, 2009. www.netflixprize.com/.
- Salford Systems. www.salford-systems.com/en/resources/case-studies/119-teradata- center-for-crm-at-duke-competition.
- Duke University Department of Statistical Science. http://stat.duke.edu/datafest/.

- Data-Mining-Cup, DMC Competition. www.data-mining-cup.de/en/review/dmc-2012/.
- ACM KDD Cup. www.sigkdd.org/kddcup/.
- Nokia Research Center, Nokia Mobile Data Challenge. http://research.nokia.com/page/12000 and Nokia Mobile Data Challenge 2012 Workshop (18–19.6.2012). http://research.nokia.com/page/12340.
- "Predicting Customer Acquisition & Retention Using Structured and Semi-Structured Data." KDnuggets. www.kdnuggets.com/2012/01/wcai-research-opportunity-siriusxm- predicting-customer-acquisition-retention.html.
- EMVIC 2012: "The First Eye Movement Identification and Verification Competition." www.emvic.org/.
- Aspiring Minds Machine Learning Competition. www.aspiringminds.in/mlCompetition/.
- Information Security Amazon Data Security Competition. https://sites.google.com/site/amazonaccessdatacompetition/.

There is a constant flow of acquisitions, startups, and company activity in the analytics space - for the latest, see this Twitter feed:

- Big data and analytics company roll-ups: tinyurl.com/bigdatarollups

Approaches to the Netflix Prize:

- Clive Thompson, "If You Liked This, You're Sure to Love That," *New York Times*, November 21, 2008. www.nytimes.com/2008/11/23/magazine/23Netflix-t.html.

Regarding collaboration rather than competition on the Netflix Prize:

- Jordan Ellenberg, "This Psychologist Might Outsmart the Math Brains Competing for the Netflix Prize," *Wired*, February 25, 2008. www.wired.com/techbiz/media/magazine/16-03/mf_netflix.

Overview of several uses of ensembles by Netflix Prize teams:

- Todd Holloway, "Ensemble Learning Better Predictions through Diversity," ETech 2008, March 11, 2008. http://abeautifulwww.com/EnsembleLearningETech.pdf.

Andreas Töscher from Netflix Prize team BigChaos:

- "Advanced Approaches for Recommender System and the Netflix Prize," Predictive Analytics World San Francisco

Conference, February 28, 2009, San Francisco, CA. www.predictiveanalyticsworld.com/sanfrancisco/2009/agenda.php#advancedapproaches.

Netflix Prize leaderboard:

- Netflix Prize leaderboard results, September 21, 2009. www.netflixprize.com/leaderboard.

Winning the Netflix Prize by a matter of 20 minutes:

- Steve Lohr, "A $1 Million Research Bargain for Netflix, and Maybe a Model for Others," *New York Times*, September 1, 2009. www.nytimes.com/2009/09/22/technology/internet/22netflix.html.

Quote from James Surowiecki:

- James Surowiecki, *The Wisdom of Crowds* (Anchor, reprint ed., 2005).

A wise crowd guesses the dollars at Predictive Analytics World:

- Dean Abbott, "Another Wisdom of Crowds Prediction Win at eMetrics/Predictive Analytics World," Data Mining and Predictive Analytics blog, April 26, 2012. http://abbottanalytics.blogspot.com/2012/04/another-wisdom-of-crowds-prediction-win.html.

InnoCentive website:

- www.innocentive.com/.

Foldit:

- Zoran Popovic, "Massive Multiplayer Games to Solve Complex Scientific Problems," TED2013 @ Vancouver. http://talentsearch.ted.com/video/Zoran-Popovic-Massive-multiplay.
- "Check Out Exactly What Competitive Protein Folding Is All About!" YouTube, May 8, 2008, uploaded by UWfoldit. www.youtube.com/watch?v5lGYJyur4FUA.

For more crowdsourcing projects, see Wikipedia's list of dozens:

- Wikipedia, "List of Crowdsourcing Projects." http://en.wikipedia.org/wiki/List_of_crowdsourcing_projects.

Ensemble modeling is often considered the most important predictive modeling advancement of this century's first decade:

- Giovanni Seni and John Elder, *Ensemble Methods in Data Mining Improving Accuracy through Combining Predictions* (Morgan & Claypool Publishers, 2010).

"The most useful and powerful result of statistical research over the past decade has come to be known as ensemble modeling":

- Robert Nisbit, John Elder, and Gary Miner, *Handbook of Statistical Analysis and Data Mining Applications* (Academic Press, 2009), 304.

The ensemble modeling method called bagging:

- Leo Breiman, "Bagging Predictors," *Machine Learning* 24, no. 2 (1996), 120–140, doi:10.1007/BF00058655. www.springerlink.com/content/l4780124w2874025/.

The visual of decision boundaries approximating a circle:

- John Elder and Greg Ridgeway [Combining Estimators to Improve Performance], KDD Tutorial Notes of the Fifth CM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999. http://datamininglab.com/media/pdfs/kdd99_elder_ridgeway.pdf.

The improvement gained by ensemble models:

- Giovanni Seni and John Elder, *Ensemble Methods in Data Mining Improving Accuracy through Combining Predictions* (Morgan & Claypool Publishers, 2010).
- John Elder and Greg Ridgeway [Combining Estimators to Improve Performance], KDD Tutorial Notes of the Fifth CM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999. http://datamininglab.com/media/pdfs/kdd99_elder_ridgeway.pdf.
- Robert Nisbit, John Elder, and Gary Miner, *Handbook of Statistical Analysis and Data Mining Applications* (Academic Press, 2009); see chaps. 13 and 18.

More regarding ensemble methods:

- S. S. Lee and John Elder, "Bundling Heterogeneous Classifiers with Advisor Perceptrons," October 14, 1997. http://datamininglab.com/media/pdfs/lee_elder_97.pdf.
- Dean Abbott, Case Study: Air Force Research Laboratory, "Integrated Toxicity Assessment System (ITAS)," AbbottAnalytics.com. www.abbottanalytics.com/data-mining-case-study-0.php.

- Dean Abbott, Data Mining Resources: White Papers, "The Benefits of Creating Ensembles of Classifiers," AbbottAnalytics.com. www.abbottanalytics.com/white-paper-classifiers.php.
- Dean Abbott, "Comparison of Algorithms at PAKDD2007," Data Mining and Predictive Analytics blog, May 1, 2007. http://abbottanalytics.blogspot.com/2007/05/comparison-of-algorithms-at-pakdd2007.html.
- Dean Abbott, "PAKDD-10 Data Mining Competition Winner: Ensembles Again!" Data Mining and Predictive Analytics blog, May 27, 2010. http://abbottanalytics.blogspot.com/2010/05/pakdd-10-data-mining-competition-winner.html.
- Mohamed Medhat Gaber, *Journeys to Data Mining: Experiences from 15 Renowned Researchers* (Springer, 2012), 61–76.
- Dean Abbott, "How to Improve Customer Acquisition Models with Ensembles," Predictive Analytics World San Francisco Conference, February 18, 2009, San Francisco, CA. www.predictiveanalyticsworld.com/sanfrancisco/2009/agenda.php#nra.
- Philip K. Chan and Salvatore J. Stolfo, "On the Accuracy of Meta-Learning for Scalable Data Mining," *Journal of Intelligent Information Systems* 8, issue 1 (January/February 1997): 5–28. http://dl.acm.org/citation.cfm?id5251006.
- Leo Breiman, "Bagging Predictors," *Machine Learning* 24, no. 2 (1996), 123–140, doi:10.1007/BF00058655. www.springerlink.com/content/l4780124w2874025/.
- Michael Murff, PayPal, and Hui Wang, "Ensembles for Online Analytic Scoring Engine," Predictive Analytics World San Francisco Conference, March 14, 2011, San Francisco, CA. www.predictiveanalyticsworld.com/sanfrancisco/2011/agenda.php#day1–11a.
- Mike Kimel, "Revenge of the Clueless: Combining Many Poor Estimates into an Expert Forecast," Predictive Analytics World Toronto Conference, April 26, 2012, Toronto, Canada. www.predictiveanalyticsworld.com/toronto/2012/agenda.php#day2–17a.

TreeNet has won multiple PA competitions, including:

- Salford Systems, "Teradata Center for CRM at Duke Competition: Predicting Customer Churn with TreeNet," Challenge. www.salford-systems.com/en/resources/case-studies/119-teradata-center-for-crm-at-duke-competition.

Both first- and second-place winners of KDD Cup 2010 employed ensemble models:

- KDD Cup 2012 Educational Data Mining Challenge, "Winners of KDD Cup 2010: Educational Data Mining Challenge," hosted by PSLC DataShop. http://pslcdatashop.web.cmu.edu/KDDCup/results.jsp.

The generalization paradox:

- John F. Elder IV, "The Generalization Paradox of Ensembles," *Journal of Computational and Graphical Statistics* 12, no. 4, Statistical Analysis of Massive Data Streams (December 2003): 853–864. http://datamininglab.com/media/pdfs/Paradox_JCGS.pdf.

CHAPTER 6: Watson and the *Jeopardy!* Challenge *(question answering)*

Some content of this chapter originated from a valuable interview with David Gondek, PhD, of IBM Research.

For a historical account of the people behind the development of Watson:

- Stephen Baker, *Final Jeopardy!: Man vs. Machine and the Quest to Know Everything* (Houghton Mifflin Harcourt, 2011).

Way back in 2003, researchers developed a system to answer the trivia questions on another TV quiz show, *Who Wants to Be a Millionaire?* The system achieved 75 percent accuracy, in part by ensembling across multiple search engines. However, this show's questions are multiple-choice, so the whittling down of candidate answers is already mostly solved for a question answering system:

- Shyong (Tony) K. Lam, David M. Pennock, Dan Cosley, and Steve Lawrence, "1 Billion Pages51 Million Dollars? Mining the Web to Play 'Who Wants to be a Millionaire?'" UAI '03, Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence, Acapulco, Mexico, August 7–10, 2003, pages 337–345. http://arxiv.org/ftp/arxiv/ papers/1212/1212.2477.pdf.

Regarding "Time flies like an arrow":

- Gilbert Burck, *The Computer Age and Its Potential for Management* (Harper & Row, 1965).

My (the author's) PhD research pertained to the "have a car/baby" example (temporal meaning in verbs):

- Eric V. Siegel and Kathleen R. McKeown, "Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights," *Computational Linguistics* 26, issue 4 (December 2000). doi:10.1162/089120100750105957, http://dl.acm.org/citation.cfm?id5971886.

Googling only 30 percent of the *Jeopardy!* questions right:

- Stephen Baker, *Final Jeopardy: Man vs. Machine and the Quest to Know Everything* (Houghton Mifflin Harcourt, 2011), 212–224.

Quote about Google's book scanning project:

- George Dyson, *Turing's Cathedral: The Origins of the Digital Universe* (Pantheon Books, 2012).

Natural language processing:

- Dursun Delen, Andrew Fast, Thomas Hill, Robert Nisbit, John Elder, and Gary Miner, *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications* (Academic Press, 2012).
- James Allen, *Natural Language Understanding*, 2nd ed. (Addison-Wesley, 1994).

Regarding the translation of "The spirit is willing, but the flesh is weak":

- John Hutchins, "The Whisky Was Invisible or Persistent Myths of MT," *MT News International* 11 (June 1995), 17–18. www.hutchinsweb.me.uk/MTNI-11–1995.pdf.

Ruminations on Apple's Siri versus Watson from WolframAlpha's creator:

- Stephen Wolfram, "Jeopardy, IBM, and WolframAlpha," Stephen Wolfram blog, January 26, 2011. http://blog.stephenwolfram.com/2011/01/jeopardy-ibm-and-wolframalpha/.

Knowledge Representation Rhapsody:

- Eric Siegel, PhD, "Knowledge Representation Rhapsody," Columbia University, January 9, 2008. www.cs.columbia.edu/Bevs/ai/songKRR.html.

The leading university classroom textbook on artificial intelligence:

- Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. (Prentice Hall, 2002).

Quote from Douglas Adams:

- Douglas Adams, *The Hitchhiker's Guide to the Galaxy* (Harmony Books, 1979).

Larry Tesler's quote, "Intelligence is whatever machines haven't done yet":

- Larry Tesler, "CV: Adages & Coinages," Larry Tesler blog, (ca. 1970) www.nomodes.com/Larry_Tesler_Consulting/Adages_and_Coinages.html.

Theft of the Lion of Nimrud ivory carving:

- Fiachra Gibbons, "Experts Mourn the Lion of Nimrud, Looted as Troops Stood By," *The Guardian*, April 30, 2003. www.guardian.co.uk/world/2003/apr/30/internationaleducationnews.arts.

IBM's revenue and size:

- IBM, "Who We Are." www.ibm.com/ibm/us/en/.

IBM's rank among top U.S. companies:

- "The World's Biggest Public Companies," edited by Scott DeCarlo, *Forbes*, April 18, 2012. www.forbes.com/global2000/#p_1_s_d6_All%20industries_United%20States_All %20states_.

TREC QA system performance:

- David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty, "Building Watson: An Overview of the DeepQA Project," *AI Magazine*, Fall 2012. www.aaai.org/Magazine/Watson/watson.php.
- TREC Proceedings. trec.nist.gov/proceedings/proceedings.html.

Alex Trebek on IBM's investment to build Watson:

- "Trebek Says IBM Computer Could Answer Wrong on *Jeopardy!*" Bloomberg Television, February 14, 2012. Online video clip, YouTube, accessed March 23, 2012. www.youtube.com/watch?v5tYguL82wk78.

IBM's estimated spend to develop Watson:

- Don Tennant, "'Final Jeopardy' Author Stephen Baker on the Impact of IBM's Watson," ITBusinessEdge, February 14, 2011. www.itbusinessedge.com/cm/community/features/interviews/blog/final-jeopardy-author-stephen-baker-on-the-impact-of-ibms-watson/.

How Watson works:

- IBM, "IBM Watson: Ushering in a new era of computing," IBM Innovations, April 11, 2012. www-03.ibm.com/innovation/us/watson/.
- IBM, "The DeepQA Project," IBM *Jeopardy!* Challenge, April 22, 2009. www.research.ibm.com/deepqa/deepqa.shtml.
- IBM, "Competing at *Jeopardy!* Is Just the First Step," IBM DeepQA Project, July 1, 2008. www.research.ibm.com/deepqa/index2.shtml.
- IBM, "The Science Behind Watson," *IBM Innovations*, April 11, 2012. www-03.ibm.com/innovation/us/watson/science-behind_watson.shtml.
- David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty, "Building Watson: An Overview of the DeepQA Project," *AI Magazine* 31, no. 3. www.aaai.org/ojs/index.php/aimagazine/article/view/2303/0.
- D. C. Gondek, A. Lally, A. Kalyanpur, J. W. Murdock, P. A. Duboue, L. Zhang, Y. Pan, Z. M. Qiu, and C. Welty, "A Framework for Merging and Ranking of Answers in DeepQA," *IBM Journal of Research and Development* 56, issue 3.4 (May–June 2012), 14:1– 14:12. http://ieeexplore.ieee.org/xpl/login.jsp?reload5true&tp5&arnumber56177810.
- David C. Gondek, IBM, "Building Watson—An Overview of the DeepQA Project." Predictive Analytics World New York Conference, October 19, 2011, New York, NY. www.predictiveanalyticsworld.com/newyork/2011/agenda.php#keynote-jeopardy.

- Edward Nazarko, IBM, "Putting IBM Watson to Work," Predictive Analytics World Toronto Conference, April 26, 2012, Toronto, Canada. www.predictiveanalyticsworld.com/toronto/2012/agenda.php#day2–1kn.
- Robert Jewell, IBM, "Putting IBM Watson to Work," Predictive Analytics World Boston Conference, October 2, 2012, Boston, MA. www.predictiveanalyticsworld.com/ boston/2012/agenda.php#day2–1kn.
- Clive Thompson, "What Is IBM's Watson?" *New York Times*, June 16, 2010. www.nytimes.com/2010/06/20/magazine/20Computer-t.html.
- "Smartest Machine on Earth," *Nova*, PBS, originally aired May 2, 2012. www.pbs.org/ wgbh/nova/tech/smartest-machine-on-earth.html.

Roger Craig, the PA practitioner and *Jeopardy!* winner:

- Roger Craig, "Data Science Meets the Quiz Show *Jeopardy!*" Predictive Analytics World Chicago Conference, June 26, 2012, Chicago, IL. www.predictiveanalyticsworld.com/ chicago/2012/agenda.php#day2–11.
- Chris Jones, "If You Could Master All the Data in the World...," *Esquire*, January 25, 2012. www.esquire.com/features/roger-craig-jeopardy-champ-0212.
- NPR Staff, "How One Man Played 'Moneyball' with *Jeopardy!*," NPR, *All Things Considered*, November 20, 2011. www.npr.org/2011/11/20/142569472/how-one-man- played-moneyball-with-jeopardy.
- Ryan Tate, "How a Geek Cracked the *Jeopardy!* Code," Gawker.com, November 16, 2011. http://gawker.com/5860275/how-a-geek-cracked-the-jeopardy-code.
- Ned Potter, "'Jeopardy!' Champ Wins Jackpot with Web App," ABCNews, Technology Review, November 17, 2011. http://abcnews.go.com/blogs/technology/2011/11/ jeopardy-champ-wins-jackpot-with-web-app/.
- Alexandra Carmichael, "Roger Craig Wins *Jeopardy* Championship with Knowledge Tracking," *Quantified Self*, November 17, 2011. http://quantifiedself.com/2011/11/ roger-craig-on-knowledge-tracking/.

Ken Jennings used Roger Craig's study system:

- Ken Jennings, “Map Tourism,” Ken-Jennings.com blog post, November 17, 2011. http://ken-jennings.com/blog/archives/3385.

Ken Jennings quotes:

- Ken Jennings, “My Puny Human Brain,” Slate.com, February 16, 2011. www.slate.com/ articles/arts/culturebox/2011/02/my_puny_human_brain.html.

View complete Watson *Jeopardy!* match video:

- Day 1: http://www.dailymotion.com/video/xj487i_jeopardy-february-14-2011-ibm-challenge-day-1_shortfilms
- Day 2: http://www.dailymotion.com/video/xj489j_jeopardy-february-15-2011-ibm-challenge-day-2_shortfilms
- Day 3: http://www.dailymotion.com/video/xj48ai_jeopardy-february-16-2011-ibm-challenge-day-3_shortfilms

List of questions of the Watson *Jeopardy!* matches:

- J! Archive, “Show #6086—Monday, February 14, 2011.” www.j-archive.com/showgame.php?game_id53575.
- J! Archive, “Show #6087—Tuesday, February 15, 2011.” www.j-archive.com/showgame.php?game_id53576.
- J! Archive, “Show #6088—Wednesday, February 16, 2011.” www.j-archive.com/ showgame.php?game_id53577.

Ken Jenning's subsequent TED talk:

- "Watson, Jeopardy and me, the obsolete know-it-all." https://www.ted.com/talks/ken_jennings_watson_jeopardy_and_me_the_obsolete_know_it_all?language=en

Applying Watson for healthcare:

- YouTube, “Progress Report: IBM Watson Utilization Management Pilot at WellPoint, Inc.,” posted by IBMWatsonSolutions, May 15, 2012. www.youtube.com/watch?v57d6c-5kJdGk.

AI palindrome generation system:

- Eric V. Siegel, “Iambic IBM AI: The Palindrome Discovery AI Project,” White Paper, CiteSeer Online, November 7, 2011. http://citeseerx.ist.psu.edu/viewdoc/summary?doi510.1.1.34.2132. *Note—here’s another pertinent palindrome the system derived—the world’s only self-referential, automatically generated, palindromic joke:* Net safety? Byte fasten.

CHAPTER 7: Persuasion by the Numbers: How Telenor, U.S. Bank, and the Obama Campaign Engineered Influence *(uplift)*

This chapter benefited greatly from valuable interviews with Patrick Surry, Nicholas Radcliffe, and Neil Skilling.

Another overview of uplift modeling by the author that somewhat complements this chapter and includes an (often-updated) bullet list of links to pertinent videos and conference sessions:

- Eric Siegel, Ph.D., "Personalization Is Back: How to Drive Influence by Crunching Numbers," *Predictive Analytics Times*, January 8, 2014. www.predictiveanalyticsworld.com/patimes/personalization-is-back-how-to-drive-influence-by-crunching-numbers/.

For a more applied overview of uplift modeling, see the author's white paper:

- Eric Siegel, PhD, "Uplift Modeling: Predictive Analytics Can't Optimize Marketing Decisions Without It," Prediction Impact, Inc. White Paper, June 29, 2011. www.predictiveanalyticsworld.com/signup-uplift-whitepaper.php.

Telenor case study:

- Suresh Vittal, "Optimal Targeting through Uplift Modeling: Generating Higher Demand and Increasing Customer Retention While Reducing Marketing Costs," Forrester Research White Paper, 2008. www.portraitsoftware.com/resources/white_papers/optimal-targeting-through-uplift-modeling.
- Suresh Vittal, "Optimizing Customer Retention Programs," Forrester Research White Paper, 2008. www.portraitsoftware.com/resources/white-papers/optimizing-customer- retention-programs.

What to say in an online dating message:

- Christian Rudder, "Exactly What to Say in a First Message," OkTrends, OkCupid, September 14, 2009. http://blog.okcupid.com/index.php/online-dating-advice-exactly-what-to-say-in-a-first-message/.

Influencing for ecologically and environmentally sound behaviors:

- Robert B. Cialdini, "Crafting Normative Messages to Protect the Environment," *Association for Psychological Science*, July 9, 2003. www.psychologicalscience.org/pdf/cialdini.pdf.
- Noah J. Goldstein, Steve J. Martin, and Robert B. Cialdini, *Yes! 50 Scientifically Proven Ways to Be Persuasive* (Free Press; reprint ed., 2009).
- Robert B. Cialdini, PhD, *Influence: The Psychology of Persuasion*, rev. ed. (HarperBusiness, 2006).
- M. Hunecke, S. Haustein, S. Bohler, and S. Grischkat, "Attitude-Based Target Groups to Reduce the Ecological Impact of Daily Mobility Behavior," *Community-Based Social Marketing, Environment and Behavior* 42 (1): 3–43. www.cbsm.com/articles/attitudebased1target1groups1to1reduce1the1ecological1impact1of1daily1mobility1behavior_ 7771.

Choice of graft source for ACL-reconstructing knee surgery:

- Lars Ejerhed, MD, Juri Kartus, MD, PhD, Ninni Sernert, Kristina Kohler, and John Karlsson, MD, PhD, "Patellar Tendon or Semitendinosus Tendon Autografts for Anterior Cruciate Ligament Reconstruction? A Prospective Randomized Study with a Two-Year Follow-Up," *American Journal of Sports Medicine* 31, no. 1 (January 2003): 19–25. http://ajs.sagepub.com/content/31/1/19.short.

Google comparing 41 shades of blue:

- Laura M. Holson, "Putting a Bolder Face on Google," *New York Times*, February 28, 2009. www.nytimes.com/2009/03/01/business/01marissa.html.

Facebook optimizes display rules:

- Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic, "Role of Social Networks in Information Diffusion," International World Wide Web Conference Committee (IW3C2), WWW 2012, April 16–20, 2012, Lyon, France. ACM 978-1-4503-1229-5/12/04. arXiv:1201.4145v2 [cs.SI], February 28,

2012. http://cameronmarlow.com/media/bakshy-the_role-2012b.pdf.

"Prescriptive analytics" is a problematic term (regarding footnote 5 on page 267):

- See the following article as well as the reader comment placed there by the author of this book (at the bottom of the article webpage): Bill Vorhies, "Prescriptive versus Predictive Analytics – A Distinction without a Difference?" November 14, 2014, Predictive Analytics Times. http://www.predictiveanalyticsworld.com/patimes/prescriptive-versus-predictive-analytics-distinction-without-difference/4259/.

U.S. Bank case study:

- Thanks to Michael Grundhoefer, Patrick Surry, and Pitney Bowes Software for this example case study.
- Jessica Tsai, "Mail Model of the Year: U.S. Bank Accrues Savings by Only Pitching Customers Who Aren't Already Likely to Show Interest," *CRM Magazine*, destinationCRM.com, April 2010. www.destinationcrm.com/Articles/Columns-Departments/REAL-ROI/Mail-Model-of-the-Year-66123.aspx.

Uplift decision trees:

- Nicholas Radcliffe and Patrick Surry, "Real-World Uplift Modelling with Significance- Based Uplift Trees," *Stochastic Solutions Limited*, March 29, 2011. http://stochasticsolutions.com/sbut.html.

Other approaches to uplift modeling (beyond uplift trees, other analytical approaches to uplift modeling include specialized *k*–nearest neighbor (*k*-NN) classifiers, a specialized version of naive Bayes, and other uplift methods):

- Victor S. Y. Lo, "The True Lift Model—A Novel Data Mining Approach to Response Modeling in Database Marketing," *ACM SIGKDD Explorations Newsletter* 4, issue 3 (2003), 78–86. www.sigkdd.org/explorations/issues/4–2–2002–12/lo.pdf.
- Kim Larsen, "Net Lift Models: Optimizing the Impact of Your Marketing," Workshop at Predictive Analytics World San Francisco Conference, San Francisco, CA, March 9, 2012. www.predictiveanalyticsworld.com/uplift.

Measuring impact in sociology:

- James J. Heckman and Richard Robb Jr., "Alternative Methods for Evaluating the Impact of Interventions: An Overview,"

Journal of Econometrics 30, issues 1–2 (October– November 1985): 239–267. http://faculty.smu.edu/millimet/classes/eco7377/papers/heckman%20robb.pdf.

Personalized medicine:

- Brian Claggett, Lihui Zhao, Lu Tian, Davide Castagno, and L. J. Wei, "Estimating Subject-Specific Treatment Differences for Risk-Benefit Assessment with Competing Risk Event-Time Data," Harvard University Biostatistics Working Paper Series, Working Paper 125, March 2011. biostats.bepress.com/harvardbiostat/paper125.
- Ross E. McKinney Jr., MD, George M. Johnson, MD, Kenneth Stanley, MD, Florence H. Yong, MS, Amy Keller, Karen J. O'Donnell, PhD, Pim Brouwers, PhD, Wendy G. Mitchell, MD, Ram Yogev, MD, Diane W. Wara, MD, Andrew Wiznia, MD, Lynne Mofenson, MD, James McNamara, MD, and Stephen A. Spector, MD, the Pediatric AIDS Clinical Trials Group Protocol 300 Study Team, "A Randomized Study of Combined Zidovudine-Lamivudine versus Didanosine Monotherapy in Children with Symptomatic Therapy-Naive HIV-1 Infection," *Journal of Pediatrics* 133, issue 4 (October 1998): 500–508. www.sciencedirect.com/science/article/pii/S0022347698700575.
- Ron Winslow, "Major Shift in War on Cancer: Drug Studies Focus on Genes of Individual Patients; Testing Obstacles Loom," *Wall Street Journal*, June 5, 2011. http://online.wsj.com/article/SB10001424052702304432304576367802580935000.html.
- Ilya Lipkovich, Alex Dmitrienko, Jonathan Denne, and Gregory Enas, "Subgroup Identification Based on Differential Effect Search—A Recursive Partitioning Method for Establishing Response to Treatment in Patient Subpopulations," *Statistics in Medicine* 30, issue 21 (September 20, 2011): 2601–2621. onlinelibrary.wiley.com/doi/10.1002/sim.4289/abstract.
- J. C. Foster, J. M. Taylor, and S. J. Ruberg, "Subgroup Identification from Randomized Clinical Trial Data," U.S. National Library of Medicine National Institute of Health, *Statistics in Medicine* 30, no. 24 (October 30, 2011): 2867–2880. doi:10.1002/sim.4322. Epub, August 4, 2011. www.ncbi.nlm.nih.gov/pubmed/21815180.
- Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li, "Subgroup Analysis via Recursive Partitioning," *Journal of Machine Learning Research* 10 (2009):

141–158.
http://jmlr.csail.mit.edu/papers/volume10/su09a/su09a.pdf.
- Andrew Pollack, "A Push to Tie New Drugs to Testing," *New York Times*, December 26, 2011. www.nytimes.com/2011/12/27/health/pressure-to-link-drugs-and50-companion-diagnostics.html.

Social influence on mobile phone customer cancellations:

- Auren Hoffman, "Birds of a Feather Use the Same Phone Service Provider," Summation.com blog post, November 4, 2009. http://blog.summation.net/2009/11/birds-of-a- feather-use-the-same-phone-service-provider.html.

Jeffrey Strickland's book has a formalization of uplift in its final chapter:

- Jeffrey Strickland, *Predictive Modeling and Analytics* (Lulu.com, August 2014). http://www.lulu.com/shop/jeffrey-strickland/predictive-modeling-and-analytics/paperback/product-21748775.html.

More on uplift modeling:

- Victor S. Y. Lo, "New Opportunities in Marketing Data Mining," in *Encyclopedia of Data Warehousing and Mining*, 2nd ed. (Idea Group Publishing). doi:10.4018/978-1-60566-010- 3.ch218, ISBN13: 9781605660103, ISBN10: 1605660108, EISBN13: 9781605660110. www.irma-international.org/viewtitle/11006/.
- Nicholas Radcliffe, "Generating Incremental Sales: Maximizing the Incremental Impact of Cross-Selling, Up-Selling and Deep-Selling through Uplift Modelling," Stochastic Solutions Limited, February 16, 2008. http://stochasticsolutions.com/cross-sellPaper.html.
- Nicholas Radcliffe, "Nicholas Radcliffe's Uplift Modeling FAQ," Scientific Marketer blog post, September 27, 2007. http://scientificmarketer.com/2007/09/uplift-modelling- faq.html.
- Nicholas Radcliffe, "Hillstrom's MineThatData Email Analytics Challenge: An Approach Using Uplift Modelling," Stochastic Solutions Limited, May 4, 2008. http://stochasticsolutions.com/etailPaper.html.
- Nicholas Radcliffe, "Uplift Modelling: You Should Not Only Measure but Model Incremental Response," Predictive Analytics World London Conference 2010, November 15, 2010, London,

UK. www.predictiveanalyticsworld.com/london/2010/agenda.php#day1–16.

- Neil Skilling, "Uplift Modelling: You Should Not Only Measure but Model Incremental Response," Predictive Analytics World London Conference 2011, December 1, 2011, London, UK. www.predictiveanalyticsworld.com/london/2011/agenda.php#day2–10.
- Kathleen Kane, Victor S. Y. Lo, Jane Zheng, and Alex Arias-Vargas, "True-Lift Modeling: Mining for the Most Truly Responsive Customers and Prospects," Predictive Analytics World San Francisco Conference March 14, 2011, San Francisco, CA. www.predictiveanalyticsworld.com/sanfrancisco/2011/agenda.php#day1–12b.
- Kim Larsen, "Response Modeling Is the Wrong Modelling: Maximize Impact with Net Lift Modeling," Predictive Analytics World San Francisco Conference 2010, February 17, 2010, San Francisco, CA. www.predictiveanalyticsworld.com/sanfrancisco/2010/agenda.php#day2–2.
- Eric Siegel, PhD, "Persuasion by the Numbers: Optimize Marketing Influence by Predicting It," Predictive Analytics World San Francisco Conference 2012, March 4, 2012, San Francisco, CA. www.predictiveanalyticsworld.com/sanfrancisco/2012/agenda.php#keynote-es.
- Patrick Surry, "Predict Who Is Persuadable Before You Market," Predictive Analytics World Chicago Conference 2012, June 25, 2012, Chicago, IL. www.predictiveanalyticsworld.com/chicago/2012/agenda.php#day1–45.

Additional technical papers on uplift modeling (added to these Notes in November 2016):

- Taiyeong Lee, Rutwen Zhang, Xiangxiang Meng, and Laura Ryan, "Incremental Response Modeling Using SAS® Enterprise Miner™," SAS Institute, Inc., Paper 096-2013 (2013). http://support.sas.com/resources/papers/proceedings13/096-2013.pdf.
- Juthika Khargharia, "How Incremental Response Modeling Can Help You Reach the Right Target Group More Precisely,"

Subconscious Musings, July 12, 2013. http://blogs.sas.com/content/subconsciousmusings/2013/07/12/how-incremental-response-modeling-can-help-you-reach-the-right-target-group-more-precisely/.

- Behram Hansotia and Brad Rukstales, "Incremental Value Modeling," *Journal of Interactive Marketing*, Vol. 16, No. 3, (2002). DOI: 10.1002/dir.10035. http://www.sciencedirect.com/science/article/pii/S1094996802701618/part/first-page-pdf
- James J. Heckman, "Sample Selection Bias as a Specification Error," *Econometrica*, Vol. 47, No. 1, (Jan. 1979), pp. 153-161. http://www.jstor.org/stable/1912352?origin=JSTOR-pdf&seq=1#page_scan_tab_contents.
- Kim Larsen, "Net Lift Models: Optimizing the Impact of Your Marketing Efforts," *SAS Training, Business Knowledge Series,* (2010). https://support.sas.com/edu/schedules.html?id=2591&ctry=US#s1=1.
- Victor S.Y. Lo, "The True Lift Model – A Novel Data Mining Approach to Response Modeling in Database Marketing," (January 2002). *SIGKDD Exploration*, Vol. 4, Issue 2, pp 78-86. https://www.researchgate.net/profile/Victor_Lo3/publication/220520042_The_True_Lift_Model_-_A_Novel_Data_Mining_Approach_to_Response_Modeling_in_Database_Marketing/links/00463530681bd8f8de000000.pdf.
- Nicholas J. Radcliffe and Patrick D. Surry, "Differential Response Analysis: Modeling True Responses by Isolating the Effect of a Single Action," *Credit Scoring and Credit Control VI*, (October 2013), cscc99-1. http://www.maths.ed.ac.uk/~mthdat25/uplift/cscc99-1.
- Naeem Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring,* (Wiley, 2005).
- Nicholas J. Radcliffe, "Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Modeling," *Direct Marketing Analytics Journal*, 2007, pp 14-21. http://www.research.ed.ac.uk/portal/en/publications/using-control-groups-to-target-on-predicted-lift-building-and-assessing-uplift-model(f6976305-b4cf-416b-9686-28fcc8334d7f)/export.html.

- Victor S.Y. Lo, "*New Opportunities in Marketing Data Mining*," Encyclopedia of Data Warehousing and Mining, Second Edition (2009). DOI: 10.4018/978-1-60566-010-3.ch218.
- Tianxi Cai, Lu Tian, Peggy H. Wong, L.J. Wei, "Analysis of Randomized Comparative Clinical Trial Data for Personalized Treatment Selection," *Oxford Journals, Science & Mathematics, Biostatitics*, Vol. 12, Issue 2, pp 270-282. http://biostatistics.oxfordjournals.org/content/12/2/270.abstract.
- Florence Hiu-Ling Yong, (2015). "Quantitative Methods for Stratified Medicine." Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences. https://dash.harvard.edu/handle/1/17463130.
- Houssam Nassif, Vitor Santos Costa, Elizabeth S. Burnside, and David Page, "Relational Differential Prediction," (2012). *Machine Learning and Knowledge Discovery in Databases*, Vol. 7523 of the series *Lecture Notes in Computer Science*, pp 617-532. http://link.springer.com/chapter/10.1007/978-3-642-33460-3_45.
- Houssam Nassif, Y. Wu, David Page, Elizabeth Burnside, "Logical Differential Prediction Bayes Net, Improving Breast Cancer Diagnosis for Older Women," *AMIA Annu Symp Proc*, 2012;2012:1330-9. Epub 2012 November 3. https://www.ncbi.nlm.nih.gov/pubmed/23304412.
- Houssam Nassif, Finn Kuusisto, Elizabeth S. Burnside, David Page, Jude Shavlik and Vitor Santos Costa, "Score As You Lift (SAYL): A Statistical Relational Learning Approach to Uplift Modeling," Proceedings of the European Conference on Machine Learning (ECML PKDD 2013). DOI: 10.1007/978-3-642-40994-3_38. http://ftp.cs.wisc.edu/machine-learning/shavlik-group/nassif.ecml13.pdf.
- Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page and Jude Shavlik, "Support Vector Machines for Differential Prediction," *Machine Learning and Knowledge Discovery in Databases,* Vol. 8725 of the series *Lecture Notes in Computer Science,* pp 50-56 (2014). https://www.ncbi.nlm.nih.gov/pubmed/26158123.
- Lukasz Zaniewicz and Szymon Jaroszewicz, "Support Vector Machines for Uplift Modeling," IEEE 13th International

Conference on Data Mining Workshops (2013), TX, USA, December 7-10, 2013 pp: 131-138, DOI Bookmark: http://doi.ieeecomputersociety.org/10.1109/ICDMW.2013.23.

Certain work in reinforcement learning has begun to approach the same problem uplift modeling sets out to solve:

- John Langford, "Contextual Bandits," Machine Learning (Theory), hunch.net (blog post), October 27, 2007. http://hunch.net/?p=298.
- Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, Ed Snelson, "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, *Journal of Machine Learning Research*, (2013) pages 3207-3260, published November, 2013. http://jmlr.org/papers/volume14/bottou13a/bottou13a.pdf.
- Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, Ed Snelson, "Counterfactual Reasoning and Learnings Systems," Cornell University Library, arXiv:1209:2355v5 [cs.LG], September 11, 2012 (v1), last revised July 27, 2013. http://arxiv.org/abs/1209.2355.
- Alex Strehl, John Langford, Sham Kakade, Lihong Li, "Learning from Logged Implicit Exploration Data," Cornell University Library, arXiv:1003.0120v2 [cs.LG], February 27, 2010, last revised June 14, 2010. http://arxiv.org/abs/1003.0120.
- David A. Freedman, "Statistical Models and Shoe Leather," *Sociological Methodology,* Volume 21 (1991), pages 291-313. www.math.rochester.edu/people/faculty/cmlr/Advice-Files/Freedman-Shoe-Leather.pdf.
- Jens Kober, J. Andrew Bagnell, Jan Peters, "Reinforcement Learning in Robotics: A survey," *The International Journal of Robotics Research,* July, 2013. www.ri.cmu.edu/pub_files/2013/7/Kober_IJRR_2013.pdf.

Sidebar—How Persuasion Modeling Helped Obama Win:

- Thanks to Rayid Ghani for valuable interviews that informed this sidebar.

- Christi Parsons and Kathleen Hennessey, "Obama campaign's investment in data crunching paid off." *Los Angeles Times*, November 13, 2012. http://articles.latimes.com/2012/nov/13/nation/la-na-obama-analytics-20121113.

- Michael Scherer, “Inside the Secret World of the Data Crunchers Who Helped Obama Win.” *TIME Magazine*, November 07, 2012. http://swampland.time.com/2012/11/07/ inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/.
- *Colbert Nation*, www.colbertnation.com. Stephen Colbert interviews Nate Silver, *New York Times* blogger about his book, *The Signal and the Noise: Why So Many Predictions Fail— but Some Don’t*. http://www.colbertnation.com/the-colbert-report-videos/420765/ november-05-2012/nate-silver.
- Peggy Noonan, “They’ve Lost That Lovin’ Feeling.” *Wall Street Journal*, July 30, 2011. http://online.wsj.com/article/SB10001424053111904800304576474620336602248.html.
- Jack Gillum, “Mitt Romney Uses Secretive Data Mining To Identify Wealthy Donors.” *Huffington Post*, August 24, 2012. www.huffingtonpost.com/2012/08/24/mitt-romney- data-mining_n_1827318.html.

Obama for America 2012's Dan Porter full conference presentations on persuasion modeling - the second one is similar, but has less high level overview of the topic and more technical details; to dig deep you could watch both:

- Daniel Porter, “Pinpointing the Persuadable: Convincing the Right Customers and the Right Voters (Video),” *Predictive Analytics Times*, December 12, 2014, (originally presented at Predictive Analytics World Conference in Boston, October 5-9, 2014). www.predictiveanalyticsworld.com/patimes/video-pinpointing-persuadables-convincing-right-customers-right-voters/.
- Daniel Porter, “Pinpointing the Persuadable: Convincing the Right Voters to Support Barack Obama (Video),” *Predictive Analytics Times*, January 8, 2014, (originally presented at Predictive Analytics World Conference in Boston, September 30-October 1, 2013). www.predictiveanalyticsworld.com/patimes/pinpointing-the-persuadables-convincing-the-right-voters-to-support-barack-obama/.

Analytics job postings for Hillary for America beyond its own site:

- Kaggle Jobs Post, “Hillary for America – Data Scientist, Modeling Analyst and Survey Methodologist (Brooklyn, NY),” Kaggle Jobs, May, 2015. www.kaggle.com/jobs/14002/hillary-for-america-data-scientist-modeling-analyst-and-survey-methodologist.
- KDNuggets Jobs Post, “Hillary for America: Data Scientist, Modeling Analyst, and Survey Methodologist,” KDNuggets Jobs, May, 2015. www.kdnuggets.com/jobs/15/05-04-hillaryclinton-data-scientist-modeling-analyst-survey-methodologist.html.

Hillary for America 2016's campaign's use of uplift modeling:

- Eric Siegel, “How Hillary for America Is (Almost Certainly) Using Uplift Modeling,” *Scientific American*. http://www.predictiveanalyticsworld.com/patimes/hillary-america-almost-certainly-using-uplift-modeling/8147/.
- Top of Mind with Julie Rose, “How Data Science Could Determine the 2016 Election,” October 5, 2016. https://www.byuradio.org/episode/0a40e117-7678-422b-a700-5ce1e8b1dd37/top-of-mind-with-julie-rose-big-data-in-2016-election-financial-decisions-roller-coasters?playhead=65&autoplay=true.