Supplementary: Three-Course Reading List

**Part 0: Introduction**

**Data science articulated, data science examples, history and context, technology landscape**

* (example) Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, Albert-László Barabási, [Flavor network and the principles of food pairing](http://www.nature.com/srep/2011/111215/srep00196/full/srep00196.html), Scientific Reports 1, Article number: 196 doi:10.1038/srep00196
* (example) Acerbi A, Lampos V, Garnett P, Bentley RA (2013) [The Expression of Emotions in 20th Century Books](http://www.plosone.org/article/info:doi/10.1371/journal.pone.0059030). PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030
* (example) [Google Flu Trends](http://www.google.org/flutrends/us/#US)
* (example) [Eigenfactor](http://www.eigenfactor.org/" \t "_blank), and [publications](http://www.eigenfactor.org/papers.php)
* (example) [L'Aquila quake: Italy scientists guilty of manslaughter](http://www.bbc.co.uk/news/world-europe-20025626), BBC
* Discussion of data science and data scientists
* [eScience: The Fourth Paradigm](http://research.microsoft.com/en-us/collaboration/fourthparadigm/) (Foreward and Introduction, pages xi - xxxi; Gray's Laws, pages 5-12)
* Chris Anderson, [“The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”](http://www.wired.com/science/discoveries/magazine/16-07/pb_theory) , Wired magazine, 2008
* [Responses to Chris Anderson](http://www.edge.org/discourse/the_end_of_theory.html), 2008

**Part 1: Data Manipulation, at Scale**

**Databases and the relational algebra**

*Readings*

* [How Vertica Was the Star of the Obama Campaign, and Other Revelations](http://citoresearch.com/data-science/how-vertica-was-star-obama-campaign-and-other-revelations)
* E. F. Codd, [1981 Turing Award Lecture, "](http://amturing.acm.org/award_winners/codd_1000892.cfm) [Relational Database: A Practical Foundation for Productivity"](http://amturing.acm.org/award_winners/codd_1000892.cfm), 1981 (Think about which arguments from this short piece are still relevant today.)
* [Advanced] Cohen et al.[“MAD Skills: New Analysis Practices for Big Data”](http://db.cs.berkeley.edu/papers/vldb09-madskills.pdf), 2009
* [Advanced] Erik Meijer, Gavin Bierma [co-Relational Model of Large Shared Data Banks](http://queue.acm.org/detail.cfm?id=1961297), Communications of the ACM, 2011

**MapReduce, Hadoop, relationship to databases, algorithms, extensions, language; key-value stores and NoSQL; tradeoffs of SQL and NoSQL***Readings*

* Ullman, Rajaraman, [Mining of Massive Datasets](http://infolab.stanford.edu/~ullman/mmds.html), Chapter 2
* Stonebraker et al., “[MapReduce and Parallel DBMS’s: Friends or Foes?](http://database.cs.brown.edu/papers/stonebraker-cacm2010.pdf)”, Communications of the ACM, January 2010.
* Dean and Ghemawat, “[MapReduce: A Flexible Data Processing Too](http://cacm.acm.org/magazines/2010/1/55744-mapreduce-a-flexible-data-processing-tool/fulltext)l”, *Communications of the ACM,*January 2010.
* Rick Cattell, “[Scalable SQL and NoSQL Data Stores](http://www.sigmod.org/publications/sigmod-record/1012/pdfs/04.surveys.cattell.pdf)”, *SIGMOD Record*, December 2010 (39:4)
* Optional Technical Background: [The Hadoop Distributed File System](http://developer.yahoo.com/hadoop/tutorial/module2.html)

**Data cleaning, entity resolution, data integration, information extraction***(NOT COVERED IN LECTURES)Readings* */ Talks*

* Elmagarmid, et. al. [Duplicate Record Detection: A Survey](https://www.cs.purdue.edu/homes/ake/pub/survey2.pdf),
* Koudas, et. al.[Record Linkage: Similarity Measures and Algorithms](http://disi.unitn.it/~p2p/RelatedWork/Matching/aj_recordLinkage_06.pdf)

**Part 2: Analytics**

**Topics in statistical modeling and experiment design** *Readings*

* Statistics is Easy! Dennis Sasha, Manda Wilson, Morgan and Claypool
* Chapter 3 of A Handbook of Statistical Analyses Using R
* [Gregory Park on overfitting to the leaderboard in a Kaggle Competition](http://blog.kaggle.com/2012/07/06/the-dangers-of-overfitting-psychopathy-post-mortem/)

**Introduction to Machine Learning, supervised learning, decision trees/forests, simple nearest neighbor***Readings*

* Xindong Wu et al., [Top 10 Algorithms in Data Mining, Knowledge and Information Systems](http://www.cs.uvm.edu/~icdm/algorithms/index.shtml), 14(2008), 1: 1-37. (read section on C4.5)
* Ullman, Rajaraman, [Mining of Massive Datasets](http://infolab.stanford.edu/~ullman/mmds/book.pdf) , Chapter 1
* Pedro Domingos, [A Few Useful Things to Know about Machine Learning](http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf), CACM 55(10), 2012

**Unsupervised learning: k-means, multi-dimensional scaling**

*Readings*

* Xindong Wu et al., [Top 10 Algorithms in Data Mining, Knowledge and Information Systems](http://www.cs.uvm.edu/~icdm/algorithms/index.shtml), 14(2008), 1: 1-37. (read section on k-means)

**Part 3: Interpreting and Communicating Results**

**Visualization, visual data analytics***Readings (well, watchings)*

* Hans Rosling, [The Joy of Stats](http://www.gapminder.org/videos/the-joy-of-stats/)
* Pat Hanaran, [Tools for Data Enthusiasts](http://vimeo.com/50723101)
* Jeffrey Heer, Michael Bostock, Vadim Ogievetsky, [A Tour through the Visualization Zoo](http://queue.acm.org/detail.cfm?id=1805128), Communications of the ACM, Volume 53 Issue 6, June 2010

**Backlash: Ethics, privacy, unreliable methods, irreproducible results**

* Howard Wen, "[Big Ethics for Big Data](http://strata.oreilly.com/2012/06/ethics-big-data-business-decisions.html)", O'Reilly Media
* John Markoff, New York Times, [Unreported Side Effects of Drugs Are Found Using Internet Search Data](http://www.nytimes.com/2013/03/07/science/unreported-side-effects-of-drugs-found-using-internet-data-study-finds.html?_r=0), March 13, 2013
* Mike Loukides, [Data Skepticism](http://strata.oreilly.com/2013/04/data-skepticism.html), O'Reilly Media, April 2013
* Gary Marcus and Ernest Davis, [Eight (No, Nine!) Problems With Big Data](http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html), New York Times, April 6, 2014
* Tim Harford, [Big data: are we making a big mistake?](http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html), March 28, 2014
* K.N.C., [The backlash against big data](http://www.economist.com/blogs/economist-explains/2014/04/economist-explains-10), The Economist, Apr 20th 2014 (very short)
* See also: [Gartner Hype cycle](http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp)
* George Johnson, [New Truths That Only One Can See](http://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html), New York Times, January 20, 2014
* John P. A. Ioannidis, [Why Most Published Research Findings Are False](http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124), PLOS One, August 30, 2005
* Dan Mckinley, [Whom the Gods Would Destroy, they First Give Real-Time Analytics](http://mcfunley.com/whom-the-gods-would-destroy-they-first-give-real-time-analytics)

**Part 4: Graph Analytics**

*Readings*

* Sherif Sakr, [Processing large-scale graph data: A guide to current technology](http://www.ibm.com/developerworks/library/os-giraph/os-giraph-pdf.pdf), June 2013