SoDA 501: Approaches and Issues in Big Social Data
Spring 2018

Burt L. Monroe
Office: Sparks B002 (The Databasement) or Pond 207
Appointments: (http://burtmonroe.youcanbook.me)
Contact: burtmonroe@psu.edu, 814-867-2726 or 814-865-9215

Description

This seminar is part of the core seminar series for students in the Social Data Analytics dual-title PhD and doctoral minor. The primary objective of the seminar is interdisciplinary exposure to, engagement with, and integration of the tools, practices, language, and standards used in the collection and management of data in the component disciplines of the Social Data Analytics field.

Each of you is well on your way toward a PhD – formal certification as an "expert" – in one of the component disciplines of Social Data Analytics and has in your coursework and research become well versed in one or more of the many computational, informational, statistical, visual analytic, or social scientific approaches to data, and the issues faced by those approaches. Here, we are interested in trying to integrate your multidisciplinary expertise, particularly in the context of data that are *social* (about, or arising from, human interaction) and *big* or *intensive* (of sufficient scale, variety, or complexity to strain the informational, computational, or cognitive limits of conventional approaches to data collection, management, manipulation, or analysis).

The SoDA core seminars are organized around the metaphor of the *social data stack*. The social data stack consists of three fuzzily boundaried layers: the "data layer," the "analytics layer," and the "relevance layer" (Fig. 1).

The *data layer* is comprised of the processes and technologies by which human interactions are translated into data about human interactions. These are the themes emphasized in SoDA 501, "Approaches and Issues in Big Social Data," offered in the spring semester. Some SoDA / IGERT students will take more in depth seminars with focus on computational and informational aspects (primarily in Information Sciences & Technology, Geography, or engineering departments) and research design aspects (primarily in social science departments or Statistics) of the data layer.

The *analytics layer* is comprised of the processes and technologies by which social data are translated into knowledge about society. These are the themes emphasized in SoDA 502, "Approaches and Issues in Social Data Analytics." Some of you will take more in-depth seminars on machine / statistical learning, visual analytics, or other statistical or social scientific approaches to inference. The *relevance layer* is comprised of the processes and technologies by which knowledge about society is translated into value for science or society. Within the SoDA seminars, this is addressed through primarily through exposure to and participation in projects that require an interdisciplinary team science approach.

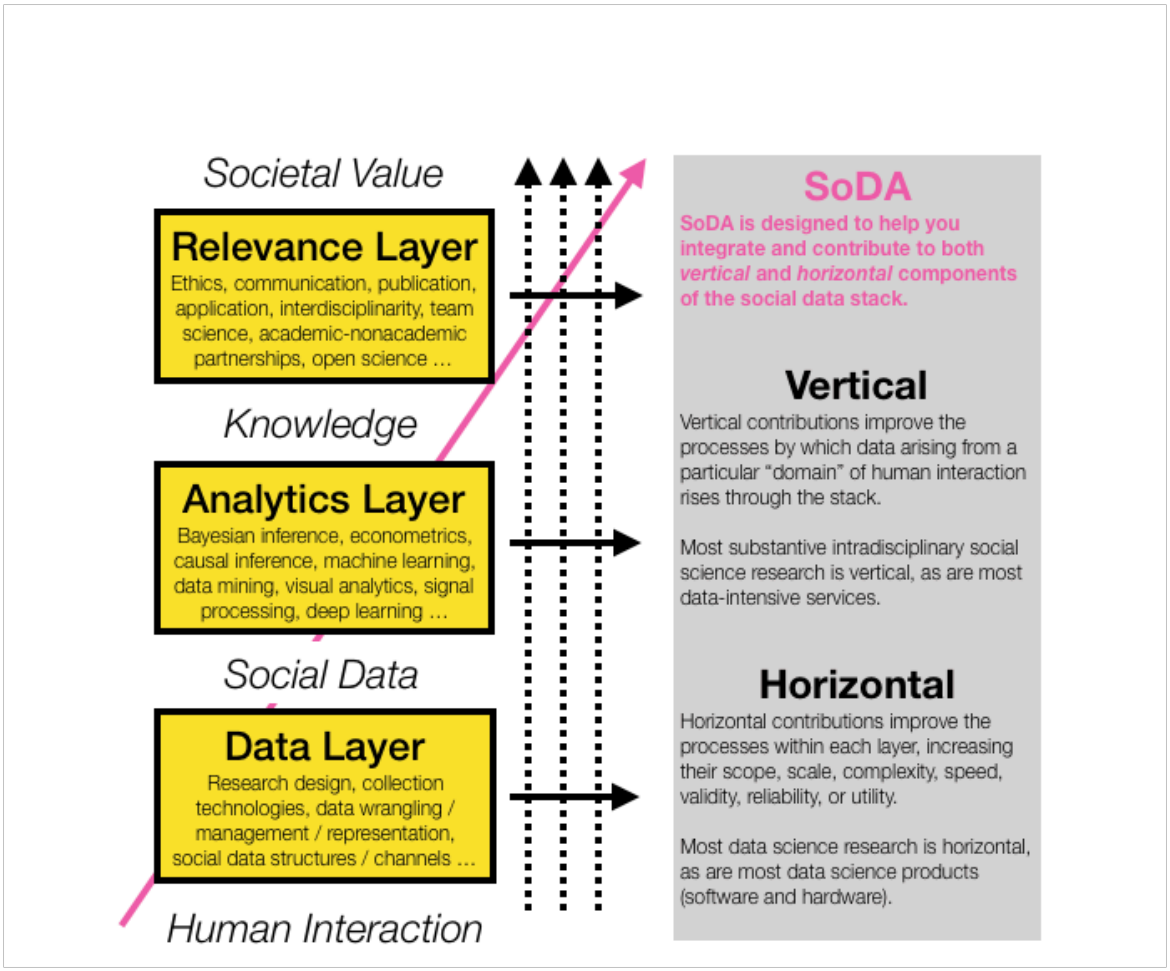

Figure 1: SoDA and the social data stack

Assignments and Grades

The latter leads us to the main pedagogical components of SoDA 501:

- **Engagement in Seminar - 40%**
    - **Guest Speakers** For half of the session most weeks, we will host a guest speaker (typically a member of the Graduate Faculty in Social Data Analytics, drawn from the full range of participating disciplines), discussing an active research project or related topic that touches on one or more areas of concern in the course. *For each speaker, we will have two or more of you acting as "designated respondents," with extra responsibility for having questions for discussion with the speaker.*
    - **Readings and Seminar Discussion** The readings, discussion, and what lecturing I will do, will focus on interdisciplinary integration. In part, this involves identifying those concepts that may be new to some of you in this setting – e.g., how "big" or "machine

learning" or "visual analytics" approaches challenge conventional social science methodology, or how social scientific thinking challenges emerging practices and conventional wisdom in data science – and tools associated with those concepts. In part, this involves interdisciplinary arbitrage and translation – identifying common concepts and structure that may go by slightly different names in different disciplines and settings. *To this end, I want you each to send me – by Wednesday 7:00am each week, by email – lists of terms / concepts that you encountered in that week's reading in three categories: (1) terms/concepts that were new but you think you now understand, (2) terms/concepts that seem to be used differently than in the context of your home discipline, and (3) terms/concepts you still find confusing.*

  - **Grading Criteria** Full 30 points if you are present every week, have made a good faith effort to provide your lists of confusing terms and concepts on time, have thoughtfully read all of the assignments, are prepared to talk about the week's readings and themes and consistently contribute in ways that are productive to the discussion (good questions, thoughtful responses, etc.), with all of that weighted more heavily when you are a designated respondent. If you don't do any of that, 0 points. Sliding scale in between.

- **Exercises - 20%** It is explicitly *not* an objective of this course to "train" you in all of the tools we will mention, much less those that we could mention, a task that would take years. In the interest of collectively "moving the ball forward" for each of you, however, we will have a small number of assigned exercises. Early in the semester, these will be done in (assigned) interdisciplinary teams. Later exercises will be individual.

- **Semester Team Project - 40%** You will, in a team consisting of at least three disciplines, create, gather, and/or organize/manipulate/prepare for analytics a "big" "social" dataset. The data must be at least partly social (arise from human interactions). There must be some nontrivial computational or informational element to the project. There need not be a final analysis of the data, but there must be some basic calculation of descriptive statistics over the data, and some demonstration of the validity of the data for the (or an) intended scientific purpose – e.g., representativeness (and of what), balance / randomization, measurement validity, etc.

  - **March 1 - Deadline for approval of teams and (proposed) projects**
  - **March 15 - 25% Project Review**
  - **March 29 - 50% Project Review**
  - **April 12 - 75% Project Review**
  - **April 26 - Team Project Presentations**
  - **May 3 - White Paper and Data / Replication Archive Due** Submit a 4-5 page paper that documents what was done and why, discusses problems you encountered, provides an assessment of the validity of the data for an analytic purpose (or how it might be validated), and discusses what further work might be done to make the data into a useful resource for others and/or to publish an analysis based on the data. Share your code and data with me in maximally documented and reproducible form (ideally a notebook stored on github or similar).

Course Schedule 2018

January 11

- Introductions.
- Syllabus (What this course is and isn't; How this course (hopefully) works).
- Further reference:
  - 10RulesforData; SoftwareCarpentry Lessons.
  - Section: "General Resources for Python and R."
  - Recommended: Python via Anaconda (`https://www.anaconda.com`); R (`https://www.r-project.org`) & RStudio (`https://www.rstudio.com`); Account on ICS-ACI (`https://ics.psu.edu`); Git (`https://git-scm.com`)
- **Exercise #1, Team Updates 1/18, Due 1/25**: BitByBit, Exercise 2.6 (a-g). Teams:
  - **TeamFrancisco**: Sara, Claire, Rosemary, Fangcao
  - **TeamFreelin**: Brittany, Arif, So Young, Xiaoran
  - **TeamKankane**: Shipi, Steve, Lulu, Omer

January 18

- Readings (send list of confusing terms / concepts by 7:00am Jan 17, Wednesday.)
  - BitByBit, Ch. 1 & 2.
  - CompSocSci; Monroe-No; Monroe-5Vs
  - One article, from a different discipline, listed in the "Multidisciplinary Perspectives" section, excluding Business-BigData.
- Further reference: Section "Big Data & Social Data Analytics."
- **Exercise #1 Team Updates**

January 25

- Readings (send list of confusing terms / concepts by 7:00 am Jan 24, Wednesday)
  - GoogleFlu; GoogleBooks; EmbeddingsBias; MachineBias; RacistBot; BDSS-Census
  - ResearchMethodsKB, "Measurement."; Quinn-Topics
- Further reference:
  - Section: "¡Cuidado!"
  - Section: "Measurement Reliability and Validity."
- **Exercise #1 Due; Teams Report**

- Determine "discussion lead" dates.
- **Exercise #2, Due 2/8**: Wikipedia / Google Trends exercise. Teams:
  - **TeamKelling**: Claire, Brittany, Arif
  - **TeamYalcin**: Omer, Lulu, Fangcao
  - **TeamPang**: Rosemary, Shipi, Sara
  - **TeamSun**: Xiaoran, Steve, So Young

February 1

- **Bing Pan (RPTM)**, "Big Data and Forecasting in Tourism"
- Readings (send list of confusing terms / concepts by 7:00 am Jan 31, Wednesday)
  - Bing Pan: "Identifying the Next Non-Stop Flying Market with a Big Data Approach," https://doi.org/10.1016/j.tourman.2017.12.008; "Google Trends and Tourist Arrivals: Emerging Biases and Proposed Corrections," https://doi.org/10.1016/j.tourman.2017.10.014 (See also: "Forecasting Destination Weekly Hotel Occupancy with Big Data," http://journals.sagepub.com/doi/abs/10.1177/0047287516669050; "Forecasting tourism demand with composite search index," https://doi.org/10.1016/j.tourman.2016.07.005.)
  - UnobtrusiveMeasures
  - Multivariate-R, Chapter 1 (Don't get bogged down when the math starts); LatentVariables, Chapter 1 (Skim - don't get bogged down in the math!); NetflixPrize
  - ResearchMethodsKB, "Sampling"; NRCReport, Chapter 8, "Sampling and Massive Data"; BitByBit, Chapter 3, "Asking Questions."
- Further reference:
  - Section: "Indirect / Unobtrusive / Nonreactive Measures, Data Exhaust."
  - Section: "Multiple Measures, Latent Variable Measurement."
  - Section: "Sampling and Survey Design."
  - Section: "Open data, APIs, linked data, ..."
  - Section: "Space and Time" (readings on Time).

February 8

- **Clio Andris (GEOG)**, "What AirBNB and Yelp can teach us about human behavior in cities."
- Readings (send list of confusing terms / concepts by 7:00 am Feb 7, Wednesday)
  - Clio Andris: "Using Yelp to Find Romance in the City: A Case of Restaurants in Four Cities," https://www.dropbox.com/s/uink0zcklwcpo5g/Yelp_Restaurants.pdf?dl=0; "Hidden Style in the City: An Analysis of Geolocated Airbnb Rental Images in Ten Major Cities," https://www.dropbox.com/s/t7y7f6880m4ty7v/AirBNB_Analysis.pdf?dl=0

- InfoRetrieval, Chs. 1, 2, 6 (you may prefer the slides from their classes). My primary hope here is that you understand the "vector space model" and "cosine similarity" (from Chapter 6). My secondary hope is that you understand the basics of Boolean information retrieval (and notions like "index", "inverted index", and "postings"). My tertiary hope is that you are exposed to some basic concepts of text analytics / NLP (including "tokenization", "normalization", "stemming","lemmatization", "stop words", "tf-idf").
- FightinWords (FW was used by Jurafsky in http://firstmonday.org/ojs/index.php/fm/article/view/4944/3863, and a best-selling book *The Language of Food*, for similar applications to Andris based on Yelp reviews, and now appears in his textbook NLP.)

- Further reference:
  - Section: "Space and Time."
  - Section: "Web Scraping."
  - Section: "Data Representations, Data Mappings"
  - Section: "Feature selection, feature extraction, feature engineering, ..."
  - Section: "Databases and data management." (esp SQL)
  - Section: "Language, Text, Speech, Audio."

- **Exercise #2 Due; Teams Report**

February 15 (Postponed)

- **Conrad Tucker (IE)**, "Cybersecurity Policies and Their Impact on Dynamic Data Driven Application Systems."

- Readings (send list of confusing terms / concepts by 7:00 am Feb 14, Wednesday)
  - Conrad Tucker: "Cybersecurity Policies and Their Impact on Dynamic Data Driven Application Systems," http://ieeexplore.ieee.org/document/8064151/; See also "Generative Adversarial Networks for Increasing the Veracity of Big Data," http://ieeexplore.ieee.org/document/8258219/

February 22

- **David Reitter (IST)**: "Computational Psycholinguistics."

- Readings (two weeks worth):
  - David Reitter: "Alignment in Web-Based Dialogue: Studies in Big Data Computational Psycholinguistics." (Based on data from the *Cancer Survivors Network* and *Reddit*) http://www.david-reitter.com/pub/reitter2017alignment.pdf.
  - Similarity
  - PatternRecognition, Ch.2, "Representation."

- MMDS, Chapter 3, "Finding Similar Items" (to understand "minhashing" and "locality sensitive hashing", you'll need to understand "hashing" – Section 1.3.2; also discussed in InfoRetrieval Chapter 3).

- Further reference:
  - Section: "Similarity, Distance, Association, ..."
  - Section: "Derived Data Representations ..."
  - Section: "Record Linkage, Entity Resolution ..."
  - Section: "Clustering, Hashing, Compression, ..."

March 1

- Reading:
  - BitByBit, Ch. 5. ("Creating Mass Collaboration").
  - Crowdsourcing, "Introduction" and Ch.1, "Concepts, Theories, and Cases of Crowdsourcing."
  - HumanComputation, Chs. 1,2,5.
- Further reference:
  - Section: "Crowdsourcing, Human Computation, Citizen Science, Web Experiments."
  - Section: "Making Up Data (smoothing, convolution, kernels, ...)"
  - Section: "Vision, image, video."
- **Semester Projects Must be Approved Before Spring Break**

March 8 - Spring Break

March 15

- **Daniel DellaPosta (SOC)**, "Networks and the Mid-20th Century American Mafia."
- Reading:
  - Dan DellaPosta: "Network Closure in the Mid-20th Century American Mafia." (*Social Networks*, 2017), https://www-sciencedirect-com.ezaccess.libraries.psu.edu/science/article/pii/S037887331630199X; "Between Clique and Corporation: Boundary-Spanning in Solidary Groups." (R&R in *American Journal of Sociology*, in the Box folder).
  - Networks, Chs. 1, 2.
  - MMDS, Ch 5. "Link Analysis."
  - MarkovVisually.
- Further references:
  - Section: "Networks and Graphs."

- Section: "Simulation, resampling, Markov Chains, ..."
- DeepLearning (different kind of network)

- **25% Project Review**

March 22

- Reading:
  - DeepLearning Ch.2, "Linear Algebra."
  - Shalizi-ADA, Chs. 16-18 ("Principal Components Analysis", "Factor Models", "Nonlinear Dimension Reduction.")
- Further reference:
  - Section: "Dimensionality reduction, decomposition, ..."
  - Section: "Multiple measures, latent variable measurement."
  - Section: "Linear algebra / matrix computations."

March 29

- **Naomi Altman (STAT)**, "Generalizing PCA."
- Reading:
  - Altman: "Generalizing PCA." 2015 slides. http://personal.psu.edu/nsa1/AltmanWebpage/PCAToronto.pdf
  - MapReduceIntuition (7 minute video providing "divide and conquer" intuition to MapReduce).
  - TidySAC-Video Hadley Wickham on the tidyverse, "split-apply-combine" with dplyr, and tidy data (1 hour video) (More detail, but perhaps less intuition, in book form discussion of "tidying" data: Chapter 5 of TidyData-R.)
  - MMDS, Ch. 2 (Read the Chapter 2 in the *new* "BETA" version of the book, which also touches on Spark and Tensorflow in section 2.4 "Extensions to MapReduce": http://i.stanford.edu/~ullman/mmdsn.html).
- Further reference:
  - Section: "Nonlinear dimension reduction, manifold learning."
  - Section: "Databases and data management." (esp NoSQL)
  - Section: "Theoretically-structured approaches to data wrangling."
  - Section: "Parallelism, MapReduce, Split-Apply-Combine."
  - Section: "Functional programming."
  - Section: "Cutting and bleeding edge ..."
- **50% Project Review**
- **Exercise #3 (Individual), Due 4/5**

April 5

- **Prasenjit Mitra (IST)**, "Classification of Tweets from Disaster Scenarios."
- Reading:
  - Prasenjit Mitra: Rudra, et al. "Summarizing Situational and Topical Information During Crises" https://arxiv.org/pdf/1610.01561.pdf; Imran, Mitra, & Castillo. "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages." http://mimran.me/papers/imran_prasenjit_carlos_lrec2016.pdf
  - NLP (Jurafsky and Martin), Chapters 15 and 16. "Vector Semantics" and "Semantics with Dense Vectors."
- Further reference:
  - Section: "Language, text, speech, audio."
  - GloVe; word2vec; word2vecExplained; t-SNE.
  - Section: "Mobile devices, distributed sensors, ... "
  - Section: "Crowdsourcing, human computation, citizen science, ..."
  - Section: "Scaling iteration, streaming data, online algorithms."
- **Exercise #3 Due**

April 12

- Reading:
  - BitByBit Ch. 4. "Running Experiments"; review Ch. 2 sections: "Natural experiments in observable data", examples.
  - CausalInference, Chs. 1-2.
- Further reference:
  - Section: "Experimental and observational designs for causal inference."
  - Section: "Crowdsourcing, ..., web experiments."
  - Section: "Human subjects ..."
- **75% Project Review**

April 19

- Reading:
  - BitByBit, Ch. 6, "Ethics."
  - PSU-ORP, "Common Rule and Other Changes." https://www.research.psu.edu/irb/commonrulechanges
  - DataPrivacy

- Reproducibility
- Refresh on MachineBias

- Further reference:
  - Section: "Ethics and Scientific Responsibility in Big Social Data."
  - Section: "¡Cuidado!"

April 26 - LAST CLASS MEETING

- **Team Project Presentations**

May 3 - Team Projects Due

Past Visiting Speakers in SoDA 501 and 502

SoDA 502 (Fall 2017)

- Chris Zorn (PLSC)
- Diane Felmlee (SOC)
- Alan MacEachren (GEOG)
- Eric Plutzer (PLSC)
- James LeBreton (PSYCH)
- Sesa Slavkovic (STAT)
- Guido Cervone (GEOG)
- Ashton Verdery (SOC)
- Maggie Niu (STAT)
- Reka Albert (PHYS)

SoDA 501 (Spring 2017)

- Tim Brick (HDFS). "Towards real-time monitoring and intervention using wearable technology."
- Aylin Caliskan (Princeton). "A Story of Discrimination and Unfairness: Bias in Word Embeddings."
- Jay Yonamine (Google - IGERT alum). "Data Science in Industry."
- Johnathan Rush (Illinois). "Geospatial Data Science Workshop."
- Rick Gilmore (PSYCH). "Toward a more reproducible and robust science of human behavior."
- Glenn Firebaugh (SOC). "Measuring Inequality and Segregation with US Census Data."
- Charles Twardy (Sotera). "Data Science for Search and Rescue."
- Anna Smith (Ohio State). "A Hierarchical Model for Network Data in a Latent Hyperbolic Space."
- Rebecca Passonneau (CSE). "Omnigraph: Rich Feature Representation for Graph Kernel Learning."
- Alex Klippel (GEOG). "Virtual Reality for Immersive Analytics."
- Murali Haran (STAT). "A Computationally Efficient Projection-based Approach for Spatial Generalized Mixed Models."

SoDA 502 (Fall 2016)

- Clio Andris (GEOG). "Integrating Social Network Data into GISystems."
- Jia Li (STAT). "Clustering under the Wasserstein Metric."
- Rachel Smith (CAS). "Stigma Networks / Perceptions of Sociograms."
- Zita Oravecz (HDFS).
- Bethany Bray (Methodology Center). "Latent Class and Latent Transition Analysis."
- Dave Hunter (STAT). "Model Based Clustering of Large Networks."
- Scott Bennett (PLSC). "ABM Model of Insurgency."
- David Reitter (IST).
- Suzanna Linn (PLSC) "Methodological Issues in Automated Text Analysis: Application to News Coverage of the US Economy."

SoDA 501 (Spring 2016)

- Bruce Desmarais (PLSC). "Learning in the Sunshine: Analysis of Local Government Email Corpora."
- Timothy Brick (HDFS). "Mapping and Manipulating Facial Expression."
- Qunying Huang (USC) "Social Media: An Emerging Data Source for Human Mobility Studies."
- Lingzhou Zue (STAT). "An Introduction to High-Dimensional Graphical Models."
- Ashton Verdery (SOC). "Sampling from Network Data."
- Sarah Battersby (Tableau). "Helping People See and Understand Spatial Data."
- Alexandra Slavkovic (STAT). "Statistical Privacy with Network Data."
- Lee Giles (IST). "Machine Learning for Scholarly Big Data."

Readings and References (updated Spring 2018)

We will discuss a relatively small subset of the readings listed here, and this will vary based on topics and readings discussed by visiting speakers and you yourselves. The remainder are provided here as curated references for more in-depth investigation in both the theory and practice related to the topic (with the latter heavily weighted toward resources in Python and R).

† Material that is, at last check, made available through the Penn State library. Most journal links should work if you are logged in to a Penn State machine, or through the Penn State VPN. Some article archives, and most books, require additional authentication through webaccess. If links are broken, start directly from a search via https://www.libraries.psu.edu. Some books require installation of e-readers, like Adobe Digital Editions. Links to lynda.com can be accessed through http:\lynda.psu.edu.

‡ Material that is, at last check, legally provided for free. In some cases, these are the preprint versions of published material.

§ Material for which a legal selection is or will be provided through the class Box folder.

Big Data & Social Data Analytics

Overviews

- ‡**[CompSocSci]**. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. "Computational Social Science." *Science*. 323(5915):721-3 + Supp., Feb 6. http://science.sciencemag.org/content/323/5915/721.full; https://gking.harvard.edu/files/gking/files/LazPenAda09.pdf.
- ‡**[NRCreport]**. National Research Council. 2013. *Frontiers in Massive Data Analysis*. National Academies Press. (Free w/ registration: http://www.nap.edu/catalog.php?record_id=18374). Ch.1 "Introduction"; Ch.2 "Massive Data in Science, Technology, Commerce, National Defense, Telecommunications, and other Endeavors."
- ‡**[MMDS]**. Jure Leskovec, Anand Rajaraman, and Jeff Ullman. 2014. *Mining of Massive Datasets* Cambridge University Press. http://www.mmds.org/. (BETA version of Third Edition: http://i.stanford.edu/~ullman/mmdsn.html)
- §**[BitByBit]**. Matthew J. Salganik. 2018 (Forthcoming). *Bit by Bit: Social Research in the Digital Age* Princeton University Press. Ch. 1 "Introduction"; Ch. 2 "Observing Behavior."
- DSHandbook-Py, Ch. 1 "Introduction: Becoming a Unicorn"; Ch.2 "The Data Science Road Map."

Burt-schtick

- ‡**[Monroe-5Vs]** Burt L. Monroe. 2013. "The Five Vs of Big Data Political Science: Introduction to the Special Issue on Big Data in Political Science." *Political Analysis*. 21(V5): 1–9. https://doi.org/10.1017/S1047198700014315. (Volume, Velocity, Variety, Vinculation, Validity)
- †**[Monroe-No]** Burt L. Monroe, Jennifer Pan, Margaret E. Roberts, Maya Sen and Betsy Sinclair. 2015. "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science." *PS: Political Science & Politics*. 48(1): 71–4. http://dx.doi.org/10.1017/S1049096514001760.
- †**[Quinn-Topics]** Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science*. 54(1): 209–28. http://onlinelibrary.wiley.com/

doi/10.1111/j.1540-5907.2009.00427.x/full (esp. topic modeling as measurement, approach to validation)

- †[**FightinWords**] Burt L. Monroe, Michael Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis*. 16(4): 372-403. https://doi.org/10.1093/pan/mpn018. (esp. the impact of sampling variance and regularization through priors.)
- ‡[**BDSS-Census**] Big Data Social Science @ PSU Team. 2012. "A Closer Look at the Kaggle Census Data." https://burtmonroe.github.io/BDSSKaggleCensus2012/. (esp. the relevance of the social processes by which data come to exist as data.)

Multidisciplinary Perspectives

- †[**Business-BigData**] Andrew McAfee and Erik Brynjolfsson. 2012. "Big Data: The Management Revolution." *Harvard Business Review*. 90(10): 61–8, October https://hbr.org/2012/10/big-data-the-management-revolution; Thomas H. Davenport and D.J. Patel. 2012. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review*. 90(10):70-6, October. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.
- †[**InfoSci-BigData**] C.L. Philip Chen and Chun-Yang Zhang. 2014. "Data-intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data." *Information Sciences*. 275: 314-47. https://doi.org/10.1016/j.ins.2014.01.015.
- †[**Informatics-BigData**] Vasant G. Honavar. 2014. "The Promise and Potential of Big Data: A Case for Discovery Informatics." *Review of Policy Research*. 31(4): 326-330. https://doi.org/10.1111/ropr.12080.
- ‡[**Stats-BigData**] Beate Franke, Jean-François, Ribana Roscher, Annie Lee, Cathal Smyth, Armin Hatefi, Fuqi Chen, Einat Gil, Alexander Schwing, Alessandro Selvitella, Michael M. Hoffman, Roger Grosse, Dietrich Hendricks, and Nancy Reid. 2016. "Statistical Inference, Learning and Models in Big Data." *International Statistical Review*. 84(3): 371-89. http://onlinelibrary.wiley.com/doi/10.1111/insr.12176/full.
- †[**Econ-BigData**] Hal R. Varian. 2013. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives*. 28(2): 3-28. https://doi.org/10.1257/jep.28.2.3.
- †[**GeoViz-BigData**] Alan MacEachren. 2017. "Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier." In Chenghu Zhou, Fenzhen Su, Francis Harvey, and Jun Xu, eds, *Spatial Data Handling in the Big Data Era*, pp 139-155. Springer. https://link-springer-com.ezaccess.libraries.psu.edu/chapter/10.1007/978-981-10-4424-3_10
- †[**Soc-BigData**] David Lazer and Jason Radford. 2017. "Data ex Machina: Introduction to Big Data." *Annual Review of Sociology* 43: 19-39. https://doi.org/10.1146/annurev-soc-060116-053457.
- §[**Politics-BigData**] Keith T. Poole, L. Jason Anasastopolous, and James E. Monagan III. Forthcoming. "The 'Big Data' Revolution in Political Campaigning and Governance." *Oxford Bibliographies in Political Science*.

¡Cuidado! Traps, Biases, Problems, Pains, Perils

- †[**GoogleFlu**] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science*. (343). 14 March. http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf.
- ‡[**MachineBias**] ProPublica. "Machine Bias: Investigating Algorithmic Injustice." Series: https://www.propublica.org/series/machine-bias. See, especially:

    - Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. "Machine Bias." *ProPublica* May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
    - Julia Angwin, Madeleine Varner and Ariana Tobin. 2017. "Facebook Enabled Adverstisers to Reach Jew Haters." https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters

- ‡[**Polling2016**] Doug Rivers. (Nov. 11, 2016). "First Thoughts on Polling Problems in the 2016 US Elections." https://today.yougov.com/news/2016/11/11/first-thoughts-polling-problems-2016-us-elections/.
- †[**EventData**] Wei Wang, Ryan Kennedy, David Lazer, Naren Ramakrishnan. 2016. "Growing Pains for Global Monitoring of Societal Events." *Science*. 353:6307, pp. 1502–1503. https://doi.org/10.1126/science.aaf6758.
- ‡[**GoogleBooks**] Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution." *PLoS One*. https://doi.org/10.1371/journal.pone.0137041.
- ‡[**OkCupid**] Michael Zimmer. 2016. "OkCupid Study Reveals the Perils of Big-Data Science." *Wired*. https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/, May 14.
- †[**CriticalQuestions**] danah boyd and Kate Crawford. 2011. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication, & Society*. 15(5): 662–79. http://dx.doi.org/10.1080/1369118X.2012.678878.
- ‡[**RacistBot**] Daniel Victor. 2016. "Microsoft created a Twitter bot to learn from users. It quickly became a racist jerk." *New York Times*. https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.
- MMDS 1.2.2-1.2.3 on Bonferroni.
- Also, BDSS-Census.

Research Design and Measurement

Overviews

- BitByBit.
- ‡[**ResearchMethodsKB**] William M. Trochim. 2006. *The Research Methods Knowledge Base* http://www.socialresearchmethods.net/kb.
- [**7Rules**] Glenn Firebaugh. 2008. *Seven Rules for Social Research* Princeton University Press.

Measurement Reliability and Validity

- ResearchMethodsKB "Measurement."
- 7Rules Ch. 3 "Build Reality Checks into Your Research."
- Reliability, see also FightinWords.
- Validity, see also Quinn10-Topics; Monroe-5Vs.

Indirect / Unobtrusive / Nonreactive Measures, Data Exhaust

- †[**UnobtrusiveMeasures**] Raymond M. Lee. 2015. "Unobtrusive Measures." *Oxford Bibliographies*. https://doi.org/10.1093/OBO/9780199846740-0048. (Canonical cite is Eugene J. Webb. 1966. *Unobtrusive Measures*, or Webb, Donald T. Campbell, Richard D. Schwartz, Lee Sechrest. 1999. *Unobtrusive Measures*, rev. ed., Sage.)
- BitByBit Examples in Chapter 2.

Multiple Measures, Latent Variable Measurement

- 7Rules Ch. 4 "Replicate Where Possible."
- †[**LatentVariables**] David J. Bartholomew, Martin Knott, and Irini Moustaki. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley. http://site.ebrary.com.ezaccess.libraries.psu.edu/lib/pennstate/detail.action?docID=10483308) (esp Ch.1, "Basic ideas and examples.")
- †[**Multivariate-R**] Brian Everitt and Torsten Hothorn. 2011. *An Introduction to Applied Multivariate Analysis with R*. Springer. http://link.springer.com.ezaccess.libraries.psu.edu/book/10.1007%2F978-1-4419-9650-3
- Shalizi-ADA, Ch.17 "Factor Models."
- ‡[**NetflixPrize**] Edwin Chen. 2011. "Winning the Netflix Prize: A Summary."
- ‡CRAN: https://cran.r-project.org/web/views/Multivariate.html; https://cran.r-project.org/web/views/Psychometrics.html; https://cran.r-project.org/web/views/Cluster.html

Sampling and Survey Design

- †[**Sampling**] Steven K. Thompson. 2012. *Sampling*, 3rd ed. http://sk8es4mc2l.search.serialssolutions.com/?sid=sersol&SS_jc=TC_024492330&title=Wiley%20Desktop%20Editions%20%3A%20Sampling
- ResearchMethodsKB "Sampling."
- NRCreport. Ch. 8, "Sampling and Massive Data."
- BitByBit Ch. 3, "Asking Questions."
- ‡[**MSE**] Daniel Manrique-Vallier, Megan E. Price, and Anita Gohdes. 2013. In Seybolt, et al. (eds). *Counting Civilians*. "Multiple Systems Estimation Techniques for Estimating Casualties in Armed Conflicts." Preprint: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.469.939&rep=rep1&type=pdf)
- †[**NetworkSampling**] Ted Mouw and Ashton M. Verdery. 2012. "Network Sampling with Memory: A Proposal for More Efficient Sampling from Social Networks." *Sociological Methodology*. 42(1):206–56. https://dx.doi.org/10.1177%2F0081175012461248
- See also, FightinWords (re hidden heteroskedasticity in sample variance).
- ‡[**HashDontSample**] Mudit Uppal. 2016. "Probabilistic data structures in the Big data world (+ code)." (re "Hash, don't sample.") https://medium.com/@muppal/probabilistic-data-structures-in-the-big-data-world-code-b9387cff0c55.
- MMDS: Hash Functions (1.3.2); Sampling in streams (4.2).
- ‡CRAN. https://cran.r-project.org/web/views/OfficialStatistics.html ("Complex Survey Design"; "Small Area Estimation").

Experimental and Observational Designs for Causal Inference

- ‡**[CausalInference]** Miguel A. Hernán, James M. Robins. Forthcoming (2017). *Causal Inference.* Chapman & Hall/CRC. Preprint: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. Chs. 1. "A definition of causal effect."; Ch. 2, "Randomized experiments."; Ch. 3, "Observational studies." (See Ch. 6, "Graphical representation of causal effects" for integration with Judea Pearl approach.)
- BitByBit Chapter 4. "Running Experiments"; Ch. 2, Natural experiments in observable data, examples.
- ResearchMethodsKB "Design."
- Firebaugh-7Rules Ch. 2, "Look for Differences that Make a Difference, and Report Them."; §Ch. 5 "Compare Like with Like."; Ch. 6 "Use Panel Data to Study Individual Change and Repeated Cross-Section Data to Study Social Change."
- Shalizi-ADA Part IV "Causal Inference."
- Monroe-No.
- CRAN. https://cran.r-project.org/web/views/ExperimentalDesign.html

Technologies for Primary and Secondary Data Collection

Mobile Devices, Distributed Sensors, Wearable Sensors, Remote Sensing

- †**[RealityMining]** Nathan Eagle and Alex (Sandy) Pentland. 2006. "Reality Mining: Sensing Complex Social Systems." *Personal and Ubiquitous Computing* 10(4): 255–68. https://doi.org:10.1007/s00779-005-0046-3.
- †**[QuantifiedSelf]** Melanie Swan. 2013. "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery." *Big Data* 1(2): 85–99. https://doi.org:10.1089/big.2012.0002.
- †**[SensorData]** Charu C. Aggarwal (Ed.). 2013. *Managing and Mining Sensor Data* Springer. http://link.springer.com.ezaccess.libraries.psu.edu/book/10.1007%2F978-1-4614-6309-2. Ch.1 "An Introduction to Sensor Data Analytics."
- †**[NightLights]** Thushyanthan Baskaran, Brian Min, Yogesh Uppal. 2015. "Election cycles and electricity provision: Evidence from a quasi-experiment with Indian special elections." *Journal of Public Economics* 126:64-73. https://doi.org/10.1016/j.jpubeco.2015.03.011.

Crowdsourcing, Human Computation, Citizen Science, Web Experiments

- †**[Crowdsourcing]** Daren C. Brabham. 2013. *Crowdsourcing.* MIT Press. http://site.ebrary.com.ezaccess.libraries.psu.edu/lib/pennstate/detail.action?docID=10692208. "Introduction"; Ch.1 "Concepts, Theories, and Cases of Crowdsourcing."
- BitByBit. Ch. 5. "Creating Mass Collaboration."
- NRCreport. Ch. 9. "Human Interaction with Data."
- †**[MTurk]** Krista Casler, Lydia Bickel, and Elizabeth Hackett. 2013. "Separate but Equal? A Comparison of Participants and Data Gathered via Amazon's MTurk, Social Media, and Face-to-Face Behavioral Testing." *Computers in Human Behavior* 29(6): 2156–60. http://doi.org/10.1016/j.chb.2013.05.009.
- †**[LabintheWild]** Katharina Reinecke and Krzysztof Z. Gajos. 2015. "LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples." In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15).* 1364–1378. http://dx.doi.org/10.1145/2675133.2675246.

- †[**TweetmentEffects**] Kevin Munger. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39(3): 629–49. http://doi.org/10.1016/j.chb.2013.05.009.
- †[**HumanComputation**] Edith Law and Luis van Ahn. 2011. *Human Computation* Morgan & Claypool. http://www.morganclaypool.com.ezaccess.libraries.psu.edu/doi/pdf/10.2200/S00371ED1V01Y201107A

Open Data, File Formats, APIs, Semantic Web / Linked Data

- DSHandbook-Py Ch. 12. "Data Encodings and File Formats."
- ‡[**OpenData**] Open Data Institute. "What Is Open Data?" https://theodi.org/what-is-open-data.
- ‡[**OpenDataHandbook**] Open Knowledge International. *The Open Data Handbook* http://opendatahandbook.org. Includes appendix "File Formats": http://opendatahandbook.org/guide/en/appendices/file-formats/.
- ‡[**APIs**] Brian Cooksey. 2016. *An Introduction to APIs* https://zapier.com/learn/apis/.
- ‡[**APIMarkets**] RapidAPI / mashape API marketplaces. https://docs.rapidapi.com; https://market.mashape.com. ProgrammableWeb. https://www.programmableweb.com.
- †[**LinkedData**] Tom Heath and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool. http://www.morganclaypool.com.ezaccess.libraries.psu.edu/doi/abs/10.2200/S00334ED1V01Y201102WBE001 Ch. 1 "Introduction"; Ch. 2 "Principles of Linked Data."
- †[**SemanticWeb**] Nikolaos Konstantinos and Dimitrios-Emmanuel Spanos. 2015. *Materializing the Web of Linked Data* Springer. http://link.springer.com.ezaccess.libraries.psu.edu/book/10.1007%2F978-3-319-16074-0. Ch. 1, "Introduction: Linked Data and the Semantic Web."
- †Morgan & Claypool. *Synthesis Lectures on the Semantic Web: Theory and Technology* http://www.morganclaypool.com/toc/wbe.1/1/1.

Web Scraping

- †[**TheoryDrivenScraping**] Richard N. Landers, Robert C. Brusso, Katelyn J. Cavanaugh, and Andrew B. Colmus. 2016. "A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research." *Psychological Methods* 4: 475–492. http://dx.doi.org.ezaccess.libraries.psu.edu/10.1037/met0000081.
- ‡[**Scraping-Py**] Al Sweigart. 2015. *Automate the Boring Stuff with Python: Practical Programming for Total Beginners*. Ch. 11: Web-Scraping https://automatetheboringstuff.com/chapter11/.
- †[**Scraping-R**] Simon Munzert, Christian Rubba, Peter Meißner, and Dominic Nyhuis. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Wiley. http://onlinelibrary.wiley.com/book/10.1002/9781118834732.
- ‡CRAN: https://cran.r-project.org/web/views/WebTechnologies.html.

Ethics & Scientific Responsibility in Big Social Data

Human Subjects, Consent, Privacy

- BitByBit Ch. 6 ("Ethics").
- ‡[**PSU-ORP**] Penn State Office for Research Protections (under the VP for Research)
  - Human Subjects Research / IRB: https://www.research.psu.edu/irb

- Revised Common Rule: https://www.research.psu.edu/irb/commonrulechanges
- Responsible Conduct of Research: https://www.research.psu.edu/education/rcr
- Research Misconduct: https://www.research.psu.edu/researchmisconduct
- SARI @ PSU (Scientific and Research Integrity Training): https://www.research.psu.edu/training/sari

- ‡[**MenloReport**] David Dittrich and Erin Kenneally (Center for Applied Interneet Data Analysis). 2012. "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research." and companion report "Applying Ethical Principles ..." https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted.
- ‡[**BigDataEthics-CBDES**] Jacob Metcalf, Emily F. Keller, and danah boyd. 2016. "Perspectives on Big Data, Ethics, and Society." Council for Big Data, Ethics, and Society. http://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/.
- ‡[**AoIRReport**] Annette Markham and Elizabeth Buchanan (Association of Internet Researchers). 2012. "Ethical Decision-Making and Internet Research, Recommendations of the AoIR Ethics Working Committee (Version 2.0)." https://aoir.org/reports/ethics2.pdf and guidelines chart https://aoir.org/wp-content/uploads/2017/01/aoir_ethics_graphic_2016.pdf.
- ‡[**BigDataEthics-Wired**] Sarah Zhang. 2016. "Scientists are Just as Confused about the Ethics of Big-Data Research as You." *Wired*. https://www.wired.com/2016/05/scientists-just-confused-ethics-big-data-research/.
- †[**BigDataEthics-HerschelMori**] Richard Herschel and Virginia M. Mori. 2017. "Ethics & Big Data." *Technology in Society* 49: 31-36. http://doi.org/10.1016/j.techsoc.2017.03.003.

The Science of Data Privacy

- †[**BigDataPrivacy**] Terence Craig and Mary E. Ludloff. 2011. *Privacy and Big Data* O'Reilly Media. http://pensu.eblib.com/patron/FullRecord.aspx?p=781814.
- ‡[**DataAnalysisPrivacy**] John Abowd, Lorenzo Alvisi, Cynthia Dwork, Sampath Kannan, Ashwin Machanavajjhala, Jerome Reiter. 2017. "Privacy-Preserving Data Analysis for the Federal Statistical Agencies." A Computing Community Consortium white paper. https://arxiv.org/abs/1701.00752.
- †[**DataPrivacy**] Stephen E. Fienberg and Aleksandra B. Slavković. 2011. "Data Privacy and Confidentiality." *International Encyclopedia of Statistical Science.* 342–5. http://doi.org/978-3-642-04898-2_202.
- ‡[**NetworksPrivacy**] Vishesh Karwa and Aleksandra Slavković. 2016. "Inference using noisy degrees: Differentially private β-model and synthetic graphs." *Annals of Statistics.* 44(1): 87-112. http://projecteuclid.org/euclid.aos/1449755958.
- †[**DataPublishingPrivacy**] Raymond Chi-Wing Wong and Ada Wai-Chee Fu. 2010. *Privacy-Preserving Data Publishing: An Overview.* Morgan & Claypool. http://www.morganclaypool.com.ezaccess.libraries.psu.edu/doi/pdfplus/10.2200/S00237ED1V01Y201003DTM002.
- DataMatching, Ch 8. "Privacy Aspects of Data Matching."

Transparency, Reproducibility, and Team Science

- ‡[**10RulesforData**] Alyssa Goodman, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, David W. Hogg, Vinay Kashyap, Ashish Mahabal, Aneta Siemiginowska, and Aleksandra Slavkovic (2014) "Ten Simple Rules for the Care and Feeding of Scientific Data." *PLoS Computational Biology* 10(4): e1003542. https://doi.org/10.1371/journal.pcbi.1003542 (Note esp, curated resources for reproducible research.)

- †[**Transparency**] E. Miguel, C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, M. Van der Laan. 2014. "Promoting Transparency in Social Science Research." *Science* 343(6166): 30–1. https://doi.org/10.1126/science.1245317.
- †[**Reproducibility**] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behavior* 0021(2017). https://doi.org/10.1038/s41562-016-0021.
- ‡[**SoftwareCarpentry**]. esp. "Lessons." http:\software-carpentry.org/lessons.
- DSHandbook-Py Ch. 9, "Technical Communication and Documentation"; Ch. 15, "Software Engineering Best Practices."
- †[**TeamScienceToolkit**] Vogel AL, Hall KL, Fiore SM, Klein JT, Bennett LM, Gadlin H, Stokols D, Nebeling LC, Wuchty S, Patrick K, Spotts EL, Pohl C, Riley WT, Falk-Krzesinski HJ. 2013. "The Team Science Toolkit: enhancing research collaboration through online knowledge sharing." *American Journal of Preventive Medicine* 45: 787-9. http://10.1016/j.amepre.2013.09.001.
- §[**Databrary**] Kara Hall, Robert Croyle, and Amanda Vogel. Forthcoming (2017). *Advancing Social and Behavioral Health Research through Cross-disciplinary Team Science.* Springer. Includes: Rick O. Gilmore and Karen E. Adolph. "Open Sharing of Research Video: Breaking the Boundaries of the Research Team." (See http://databrary.org)
- CRAN https://cran.r-project.org/web/views/ReproducibleResearch.html.

Social Bias / Fair Algorithms

- MachineBias
- ‡[**EmbeddingsBias**] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human Biases." *Science*. https://arxiv.org/abs/1608.07187.
- ‡[**Debiasing**] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." https://arxiv.org/abs/1607.06520.
- ‡[**AvoidingBias**] Moritz Hardt, Eric Price, Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." https://arxiv.org/abs/1610.02413.
- ‡[**InevitableBias**] Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." https://arxiv.org/abs/1609.05807.

Databases and Data Management

- NRCreport. Ch. 3. "Scaling the Infrastructure for Data Management."
- †[**SQL**] Jan L. Harrington. 2010. *SQL Clearly Explained* Elsevier. http://www.sciencedirect.com.ezaccess.libraries.psu.edu/science/book/9780123756978.
- DSHandbook-Py Ch. 14, "Databases."
- †[**noSQL**] Guy Harrison. 2015. *Next Generation Databases: NoSQL, NewSQL, and Big Data* Apress. https://link-springer-com.ezaccess.libraries.psu.edu/book/10.1007/978-1-4842-1329-2.
- †[**Cloud**] Divyakant Agrawal, Sudipto Das, and Amr El Abbadi. 2012. *Data Management in the Cloud: Challenges and Opportunities* Morgan & Claypool. http://www.morganclaypool.com.ezaccess.libraries.psu.edu/doi/pdfplus/10.2200/S00456ED1V01Y201211DTM032.

- †Morgan & Claypool *Synthesis Lectures on Data Management* http://www.morganclaypool.com/toc/dtm/1/1

Data Wrangling

Theoretically-structured approaches to data wrangling

- ‡**[TidyData-R]** Garrett Grolemund and Hadley Wickham. 2017. *R for Data Science* (http://r4ds.had.co.nz/), esp Ch. 12 , "Tidy Data."; also Wickham. 2014. "Tidy Data." *Journal of Statistical Software* 59(10). http://www.jstatsoft.org/v59/i10/paper. Tools: The tidyverse, http://tidyverse.org.
- ‡**[DataScience-Py]** Jake VanderPlas. 2016. *Python Data Science Handbook* (https://github.com/jakevdp/PythonDSHandbook-Py/), esp Ch 3 on pandas: http://pandas.pydata.org
- ‡**[DataCarpentry]** Colin Gillespie and Robin Lovelace. 2017. *Efficient R Programming.* https://csgillespie.github.io/efficientR/, esp Ch. 6 on "efficient data carpentry."
- §**[Wrangler]** Joseph M. Hellerstein, Jeffrey Heer, Tye Rattenbury, and Sean Kandel. 2017. *Data Wrangling: Practical Techniques for Data Preparation.* Tool: Trifacta Wrangler, http://www.trifacta.com/products/wrangler.

Data wrangling practice

- DSHandbook-Py, Ch. 4, "Data Munging: String Manipulation, Regular Expressions, and Data Cleaning."
- †**[DataSimplification]** Jules J. Berman. 2016. *Data Simplification: Taming Information with Open Source Tools.* Elsevier. http://www.sciencedirect.com.ezaccess.libraries.psu.edu/science/book/9780128037812.
- †**[Wrangling-Py]** Jacqueline Kazil; Katharine Jarmul. 2016. *Data Wrangling with Python.* O'Reilly. http://proquestcombo.safaribooksonline.com.ezaccess.libraries.psu.edu/9781491948804.
- †**[Wrangling-R]** Bradley C. Boehmke. 2016. *Data Wrangling with R.* Springer. http://link.springer.com.ezaccess.libraries.psu.edu/book/10.1007%2F978-3-319-45599-0.

Record Linkage / Entity Resolution / Deduplication

- †**[DataMatching]** Peter Christen. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* http://link.springer.com.ezaccess.libraries.psu.edu/book/10.1007%2F978-3-642-31164-2 (esp. Ch. 2, "The Data Matching Process")
- DataSimplification, Chapter 5 "Identifying and Deidentifying Data."
- ‡**[EntityResolution]** Lise Getoor and Ashwin Machanavajjhala. 2013. "Entity Resolution for Big Data." Tutorial, KDD. http://www.umiacs.umd.edu/~getoor/Tutorials/ER_KDD2013.pdf.
- ‡**[SyrianCasualties]** Peter Sadosky, Anshumali Shrivastava, Megan Price, and Rebecca C. Steorts. 2015. "Blocking Methods Applied to Casualty Records from the Syrian Conflict." https://arxiv.org/abs/1510.07714 (For more on blocking, see DataMatching, Ch. 4 "Indexing.")
- CRAN Task Views: https://cran.r-project.org/web/views/OfficialStatistics.html "Statistical Matching and Record Linkage."

"Making up data": Imputation, Smoothers, Kernels, Priors, Filters, Teleportation, Negative Sampling, Convolution, Augmentation, Adversarial Training

- †[**Imputation**] Yi Deng, Changgee Chang, Moges Seyoum Ido, and Qi Long. 2016. "Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data." *Scientific Reports*. 6(21689). https://www.nature.com/articles/srep21689.
- ‡[**Overimputation**] Matthew Blackwell, James Honaker, and Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Overview and Applications." *Sociological Methods & Research*. 46(3) 303-341. http://gking.harvard.edu/files/gking/files/measure.pdf.
- Shalizi-ADA, Sect. 1.5 (Linear Smoothers), Ch. 8 (Splines), Sect. 14.4 (Kernel Density Estimates); DataScience-Python NB 05.13, "Kernel Density Estimation."
- FightinWords (re Bayesian priors as additional data, impact of priors on regularization).
- DeepLearning, Section 7.5 (Data Augmentation), 7.12 (Dropout), 7.13 (Adversarial Examples), Chapter 9 (Convolutional Networks)
- †[**Adversarial**] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. 2015. "Explaining and Harvesting Adversarial Examples." https://arxiv.org/abs/1412.6572.
- See also resampling and simulation methods.
- See also feature engineering / preprocessing.
- CRAN Task Views: https://cran.r-project.org/web/views/OfficialStatistics.html "Imputation."

(Direct) Data Representations, Data Mappings

- NRCreport. Ch. 5. "Large-Scale Data Representations."
- †[**PatternRecognition**] M. Narasimha Murty and V. Susheela Devi. 2011. *Pattern Recognition: An Algorithmic Approach* Springer. https://link-springer-com.ezaccess.libraries.psu.edu/book/10.1007%2F978-0-85729-495-1 Section 2.1: "Data Structures for Pattern Representation."
- ‡[**InfoRetrieval**] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *Introduction to Information Retrieval* Cambridge University Press. http://nlp.stanford.edu/IR-book/. Ch 1, 2, 6. (also note slides used in their class).
- †[**Algorithms**] Brian Steele, John Chandler, Swarna Reddy. 2016. *Algorithms for Data Science* Wiley. http://link.springer.com.ezaccess.libraries.psu.edu/book/10.1007%2F978-3-319-45797-0, Ch. 2 "Data Mapping and Data Dictionaries."
- See also Social Data Structures.

Similarity, Distance, Association, Covariance, The Kernel Trick

- PatternRecognition Section 2.3 "Proximity Measures."
- ‡[**Similarity**] Brendan O'Connor. 2012. "Cosine similarity, Pearson correlation, and OLS coefficients." https://brenocon.com/blog/2012/03/cosine-similarity-pearson-correlation-and-ols-coefficients/
- For relatively comprehensive lists, see also:
  - M-J Lesot, M Rifqi, and H Benhadda. 2009. "Similarity measures for binary and numerical data: a survey." *Int. J. Knowledge Engineering and Soft Data Paradigms* 1(1): 63-. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.212.6533&rep=rep1&type=pdf
  - Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert. 2010. "A Survey of Binary Similarity and Distance Measures." *Journal of Systemics, Cybernetics, & Informatics* 8(1):43-8. http://www.iiisci.org/Journal/CV$/sci/pdfs/GS315JG.pdf

    - Sung-Hyuk Cha. 2007. "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions." *International Journal of Mathematical Models and Methods in Applied Sciences* 4(1): 300-7. http://csis.pace.edu/ctappert/dps/d861-12/session4-p2.pdf.
    - Anna Huang. 2008. "Similarity Measures for Text Document Clustering." *Proceedings of the 6th New Zealand Computer Science Research Student Conference* 49-56. http://www.nzcsrsc08.canterbury.ac.nz/site/proceedings/Individual_Papers/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf

- Michael B. Jordan. "The Kernel Trick." (Lecture Notes) https://people.eecs.berkeley.edu/~jordan/courses/281B-spring04/lectures/lec3.pdf
- Eric Kim. 2017. "Everything You Ever Wanted to Know about the Kernel Trick (But Were Afraid to Ask)." http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick_blog_ekim_12_20_2017.pdf

Derived Data Representations - Dimensionality Reduction / Compression / Decomposition / Embeddings

The groupings here, and under the measurement / multivariate statistics section, are particularly arbitrary. For example, "k-Means clustering" can be viewed as a technique for "dimensionality reduction," "compression," "feature extraction," "latent variable measurement," "unsupervised learning," "collaborative filtering" ...

Clustering, hashing, quantization, blocking, compression

- ‡**[Compression]** Khalid Sayood. 2012. *Introduction to Data Compression*, 4th ed. Springer. http://www.sciencedirect.com.ezaccess.libraries.psu.edu/science/book/9780124157965, (e.g., coding, blocking via vector quantization).
- MMDS, Ch.3 "Finding Similar Items." (minhashing, locality sensitive hashing)
- Multivariate-R Ch. 6, "Clustering."
- DataScience-Python NB 05.11, "k-Means Clustering"; NG 05.12, "Gaussian Mixture Models."
- ‡**[KMeansHashing]** Kaiming He, Fang Wen, Jian Sun. 2013. "K-means Hashing: An Affinity-Preserving Quantization Method for Learning Binary Compact Codes." *CVPR*, https://www.cv-foundation.org/openaccess/content_cvpr_2013/papers/He_K-Means_Hashing_An_2013_CVPR_paper.pdf.
- †**[Squashing]** Madigan, D., Raghavan, N., Dumouchel, W., Nason, M., Posse, C., and Ridgeway, G. (2002). "Likelihood-based data squashing: A modeling approach to instance construction." *Data Mining and Knowledge Discovery*, 6(2), 173-190. http://dx.doi.org.ezaccess.libraries.psu.edu/10.1023/A:1014095614948.
- †**[Core-sets]** Piotr Indyk, Sepideh Mahabadi, Mohammad Mahdian, Vahab S. Mirrokni. 2014 "Composable core-sets for diversity and coverage maximization." *PODS '14*. 100-8. https://doi.org/10.1145/2594538.2594560.

Feature selection, feature extraction, feature engineering, weighting, preprocessing

- PatternRecognition Sections 2.6-7. "Feature Selection; Feature Extraction."
- DSHandbook-Py Ch. 7, "Interlude: Feature Extraction Ideas."
- DataScience-Python NB 05.04, "Feature Engineering."
- *Features in text*: NLP and InfoRetrieval re tf.idf and similar; FightinWords.
- *Features in images*: DataScience-Python NB 05.14, "Image Features."

Dimensionality reduction, decomposition / factorization, change of basis / reparameterization, matrix completion, latent variables, source separation

- MMDS: Ch. 9, "Recommendation Systems."; Ch.11 "Dimensionality Reduction."
- DSHandbook-Py Ch. 10, "Unsupervised Learning: Clustering and Dimensionality Reduction."
- ‡**[Shalizi-ADA]** Cosma Rohilla Shalizi. 2017. *Advanced Data Analysis from an Elementary Point of View*: Ch. 16 ("Principal Components Analysis"); Ch. 17 ("Factor Models"). http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/.

  See also Multivariate-R Ch. 3 "Principal Components Analysis"; Ch. 4 "Multidimensional Scaling"; Ch. 5 "Exploratory Factor Analysis"; Latent.
- DeepLearning Ch. 2 "Linear Algebra"; Ch 13 "Linear Factor Models."
- ‡**[GloVe]** Jeffrey Pennington, Richard Socher, Christopher Manning. 2014. "GloVe: Global vectors for word representation." *EMNLP* https://nlp.stanford.edu/projects/glove/.
- †**[NMF]** Daniel D. Lee and H. Sebastian Seung. 1999. "Learning the parts of objects by non-negative matrix factorization." *Nature*. 401:788-791. http://dx.doi.org.ezaccess.libraries.psu.edu/10.1038/44565.
- †**[CUR]** Michael W. Mahoney and Petros Drineas. 2009. "CUR matrix decompositions for improved data analysis." *PNAS*. http://www.pnas.org/content/106/3/697.full.
- ‡**[ICA]** Aapo Hyvärinen and Erkki Oja. 2000. "Independent Component Analysis: Algorithms and Applications." *Neural Networks* 13(4-5): 411-30. https://www.cs.helsinki.fi/u/ahyvarin/papers/NN00new.pdf.
- **[RandomProjection]** Ella Bingham and Heikki Mannilla. 2001. "Random projection in dimensionality reduction: applications to image and text data." *KDD* https://doi.org/10.1145%2F502512.502546.
- Compression, e.g., "Transform coding", "Wavelets."

Nonlinear dimensionality reduction / Manifold learning

- DeepLearning Section 5.11.3 , "Manifold Learning."
- DataScience-Python NB 05.10, "Manifold Learning." (Locally linear embedding [LLE]; Isomap)
- ‡**[KernelPCA]** Sebastian Raschka. 2014. "Kernel tricks and nonlinear dimensionality reduction via RBF kernel PCA." http://sebastianraschka.com/Articles/2014_kernel_pca.html.
- Shalizi-ADA Ch. 18 "Nonlinear Dimensionality Reduction." (LLE)
- ‡**[LaplacianEigenmaps]** Mikhail Belkin and Partha Niyogi. 2003. "Laplacian eigenmaps for dimensionality reduction and data representation." *Neural Computation*. 15(6): 13731396. h.ttp://web.cse.ohio-state.edu/~belkin.8/papers/LEM_NC_03.pdf
- DeepLearning Ch.14 ("Autoencoders").
- ‡**[word2vec]** Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013 "Efficient Estimation of Word Representations in Vector Space." https://arxiv.org/abs/1301.3781.
- ‡**[word2vecExplained]** Yoav Goldberg and Omer Levy. 2014. "word2vec Explained: Deriving Mikolov et al.s Negative-Sampling Word-Embedding Method." https://arxiv.org/abs/1402.3722
- ‡**[t-SNE]** Laurens van der Maaten and Geoffrey Hinton. 2008. "Visualising Data using t-SNE." *Journal of Machine Learning Research*. 9: 2579-2605. http://jmlr.csail.mit.edu/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

Computation and Scaling Up

Scientific computing, computation at scale

- NRCreport. Ch. 10. "The Seven Computational Giants of Massive Data Analysis."
- DSHandbook-Py Ch. 21, "Performance and Computer Memory"; Ch. 22, "Computer Memory and Data Structures"
- ‡**[ComputationalStatistics-Python]** Cliburn Chan. Computational Statistics in Python: https://people.duke.edu/~ccc14/sta-663/index.html.
- CRAN Task View: High Performance Computing https://cran.r-project.org/web/views/HighPerformanceComputing.html

Numerical computing

- **[NumericalComputing]** Ward Cheney and David Kincaid. 2013. *Numerical Mathematics and Computing*, 7th ed. Brooks/Cole Cengage Learning.
- CRAN Task View: Numerical Mathematics https://cran.r-project.org/web/views/NumericalMathematics.html

Optimization (e.g., MLE, gradient descent, stochastic gradient descent, EM algorithm, neural nets)

- DSHandbook-Py Ch. 23, "Maximum Likelihood Estimation and Optimization" (gradient descent)
- DeepLearning Ch. 4, "Numerical Computation"; Ch. 6, "Deep Feedforward Networks"; Ch. 8, "Optimization for Training Deep Models."
- CRAN Task View: Optimization and Mathematical Programming https://cran.r-project.org/web/views/Optimization.html.

Linear algebra / matrix computations

- **[MatrixComputations]** Gene H. Golub and Charles F. Van Loan. 2013. *Matrix Computations*. 4th ed. Johns Hopkins University Press.
- ‡**[NetflixMatrix]** Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. "Matrix Factorization Techniques for Recommender Systems." *Computer* August, 42-9. https://datajobs.com/data-science-repo/Recommender-Systems-[Netflix].pdf
- ‡**[BigDataPCA]** Jianqing Fan, Qiang Sun, Wen-Xin Zhou, Ziwei Zhu. "Principal Component Analysis for Big Data." http://www.princeton.edu/~ziweiz/pca.pdf.
- ‡**[RandomSVD]** Andrew Tulloch. 2009. "Fast Randomized SVD." https://research.fb.com/fast-randomized-svd/
- ‡**[FactorSGD]** Rainer Gemulla, Peter J. Haas, Erik Nijkamp, and Yannis Sismanis. 2011. "Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent." *KDD* http://www.cs.utah.edu/~hari/teaching/bigdata/gemulla11dsgd.pdf
- ‡**[Sparse]** Max Grossman. 2015. "101 Ways to Store a Sparse Matrix." https://medium.com/@jmaxg3/101-ways-to-store-a-sparse-matrix-c7f2bf15a229
- ‡**[DontInvert]** John D. Cook. 2010. "Don't Invert that Matrix!" https://www.johndcook.com/blog/2010/01/19/dont-invert-that-matrix/

Simulation-based inference, resampling, Monte Carlo methods, MCMC, Bayes, approximate inference

- ‡[**MarkovVisually**] Victor Powell. "Markov Chains Explained Visually." http://setosa.io/ev/markov-chains/.
- DSHandbook-Py Ch. 25, "Stochastic Modeling" (Markov chains, MCMC, HMM).
- Bayes; ComputationalStatistics-Py.
- DeepLearning Ch. 17, "Monte Carlo Methods"; Ch. 19, "Approximate Inference."
- ‡[**VariationalInference**] Jason Eisner. 2011. "High-Level Explanation of Variational Inference." https://www.cs.jhu.edu/~jason/tutorials/variational.html
- CRAN Task View: Bayesian Inference https://cran.r-project.org/web/views/Bayesian.html.

Parallelism, MapReduce, Split-Apply-Combine

- ‡[**MapReduceIntuition**] Jigsaw Academy. 2014. "Big Data Specialist: MapReduce." https://www.youtube.com/watch?v=TwcYQzFqg-8&feature=youtu.be.
- MMDS, Ch 2. "Map-Reduce and the New Software Stack." (3rd Edition, discusses Spark & TensorFlow.)
- Algorithms Ch. 3"Scalable Algorithms and Associative Statistics.", Ch. 4. "Hadoop and MapReduce."
- NRCreport. Ch. 6. "Resources, Trade-offs, and Limitations."
- TidyData-R (Split-apply-combine is the motivating principle behind the "tidyverse" approach.) See also Part III "Program" (pipes, functions, vectors, iteration).
- ‡[**TidySAC-Video**] Hadley Wickham. 2017. "Data Science in the Tidyverse." https://www.rstudio.com/resources/videos/data-science-in-the-tidyverse/.
- See also: FactorSGD

Functional Programming

- ComputationalStatistics-Python "Functions are first class objects" through first exercises.
- DSHandbook-Py Ch. 20 "Programming Language Concepts."
- See also Haskell; .

Scaling iteration, streaming data, online algorithms (Spark)

- NRCreport. Ch. 4. "Temporal Data and Real-Time Algorithms."
- MMDS: Ch. 4. "Mining Data Streams."
- ‡[**BDAS**] AMPLab. *BDAS: The Berkeley Data Analytics Stack* https://amplab.cs.berkeley.edu/software
- †[**Spark**] Mohammed Guller. 2015. *Big Data Analytics with Spark: A Practitioners Guide to Using Spark for Large-Scale Data Processing, Machine Learning, and Graph Analytics, and High-Velocity Data Stream Processing* Apress. https://link-springer-com.ezaccess.libraries.psu.edu/book/10.1007/978-1-4842-0964-6.
- DSHandbook-Py Ch. 13, "Big Data."
- DeepLearning Ch. 10, "Sequence Modeling: Recurrent and Recursive Nets."

General Resources for Python and R

- †[**DSHandbook-Py**] Field Cady. 2017. *The Data Science Handbook* (Python-based) Wiley. http://onlinelibrary.wiley.com.ezaccess.libraries.psu.edu/book/10.1002/9781119092919.
- ‡[**Tutorials-R**] Ujjwal Karn. 2017. "A curated list of R tutorials for Data Science, NLP, and Machine Learning." https://github.com/ujjwalkarn/DataScienceR
- ‡[**Tutorials-Python**] Ujjwal Karn. 2017. "A curated list of Python tutorials for Data Science, NLP, and Machine Learning." https://github.com/ujjwalkarn/DataSciencePython

Cutting and Bleeding Edge of Data Science Languages

- [**Scala**]. †Vishal Layka and David Pollak. 2015. *Beginning Scala.* https://link-springer-com.ezaccess.libraries.psu.edu/book/10.1007%2F978-1-4842-0232-6; †Nicolas Patrick. 2014. *Scala for Machine Learning.* https://ebookcentral.proquest.com/lib/pensu/detail.action?docID=1901910; ‡Scala site https://www.scala-lang.org. (See also .)
- [**Julia**]. †Ivo Baobaert. 2015. *Getting Started with Julia.* https://ebookcentral.proquest.com/lib/pensu/detail.action?docID=1973847; ‡Douglas Bates. 2013. "Julia for R Programmers." http://www.stat.wisc.edu/~bates/JuliaForRProgrammers.pdf; ‡Julia site https://julialang.org.
- [**Haskell**]. †Richard Bird. 2014. *Thinking Functionally with Haskell.* https://doi-org.ezaccess.libraries.psu.edu/10.1017/CBO9781316092415; †Hakim Cassimally. 2017. *Learning Haskell Programming.* https://www.lynda.com/Haskell-tutorials/Learning-Haskell-Programming/604926-2.html; †James Church. 2017. *Learning Haskell for Data Analysis* https://www.lynda.com/Developer-tutorials/Learning-Haskell-Data-Analysis/604234-2.html; Haskell site https://www.haskell.org.
- [**Clojure**]. †Mark McDonnell. 2017. *Quick Clojure: Essential Functional Programming.* https://link-springer-com.ezaccess.libraries.psu.edu/book/10.1007%2F978-1-4842-2952-1; †Akhill Wali. 2014. *Clojure for Machine Learning.* https://ebookcentral.proquest.com/lib/pensu/detail.action?docID=1674848; †Arthur Ulfeldt. 2015. https://www.lynda.com/Clojure-tutorials/Up-Running-Clojure/413127-2.html; ‡Clojure site https://clojure.org.
- [**TensorFlow**]. ‡Abadi, et al. (Google). 2016. "TensorFlow: A System for Large-Scale Machine Learning." https://arxiv.org/abs/1605.08695; ‡Udacity "Deep Learning." https://www.udacity.com/course/deep-learning--ud730; ‡TensorFlow site https://www.tensorflow.org.
- [**H2O**] Darren Cool. 2016. *Practical Machine Learning with H2O Powerful, Scalable Techniques for Deep Learning and AI.* O'Reilly; Arno Candel and Viraj Parmar. 2015 *Deep Learning with H2O*; ‡H2O site https://www.h2o.ai

Social Data Structures

Space and Time

- †[**GIA**]. David O'Sullivan, David J. Unwin. 2010. *Geographic Information Analysis, Second Edition* John Wiley & Sons. http://onlinelibrary.wiley.com/book/10.1002/9780470549094.
- [**Space-Time**] Donna J. Peuquet. 2003. *Representations of Space and Time* Guilford.
- Roger S. Bivand, Edzer Pebesma, Virgilio Gmez-Rubio. 2013. *Applied Spatial Data Analysis in R* Free through library: http://link.springer.com.ezaccess.libraries.psu.edu/book/10.1007%2F978-1-4614-7618-4. See also Edzer Pebesma. 2016. "Handling and Analyzing Spatial, Spatiotemporal and Movement Data in R." https://edzer.github.io/UseR2016/.
- ‡[**PySAL**]. Sergio J. Rey and Dani Arribas-Bel. 2016. "Geographic Data Science with PySAL and the pydata Stack." http://darribas.org/gds_scipy16/

- DataScience-Python. NB 04.13 "Geographic Data with Basemap."
- *Time: "Longitudinal," intra-individual data as common in developmental psychology*: †[**Longitudinal**] Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs, editors. 2009. *Longitudinal Data Analysis* Chapman & Hall / CRC. http://pensu.eblib.com/patron/FullRecord.aspx?p=359998.
- *Time: "Time Series / Time Series Cross Section / Panel," as common in economics, political science, and sociology.* DataScience-Python Ch. 3 re pandas; DSHandbook-Py Ch. 17, "Time Series Analysis"; Econometrics; GelmanHill.
- *Time: "Sequential Data / Streams" as in NLP, machine learning* Spark and other "streaming data" readings,
- *Space-Time: Data with continuity, connectivity, neighborhood structure* See DeepLearning Chapter 9 re convolution.
- CRAN Task Views: Spatial https://cran.r-project.org/web/views/Spatial.html; Spatiotemporal https://cran.r-project.org/web/views/SpatioTemporal.html; Time Series https://cran.r-project.org/web/views/TimeSeries.html.

Network, Graphs

- ‡[**Networks**] David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* Cambridge University Press. http://www.cs.cornell.edu/home/kleinber/networks-book/.
- MMDS: Ch. 5, "Link Analysis"; Ch. 10, "Mining Social-Network Graphs."
- †[**Networks-R**] Douglas Luke. 2015. *A User's Guide to Network Analysis in R*. Springer. https://link-springer-com.ezaccess.libraries.psu.edu/book/10.1007/978-3-319-23883-8
- †[**Networks-Python**] Mohammed Zuhair Al-Taie and Seifedine Kadry. 2017. *Python for Graph and Network Analysis*. Springer. https://link-springer-com.ezaccess.libraries.psu.edu/book/10.1007%2F978-3-319-53004-8

Hierarchy, Clustered Data, Aggregation, Mixed Models

- †[**GelmanHill**] Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models* Cambridge University Press. http://pensu.eblib.com/patron/FullRecord.aspx?p=288457.
- †[**MixedModels**] Eugene Demidenko. 2013. *Mixed Models: Theory and Applications with R*, 2nd ed. Wiley. http://site.ebrary.com.ezaccess.libraries.psu.edu/lib/pennstate/detail.action?docID=10748641

Social Data Channels

Language, Text, Speech, Audio

- ‡[**NLP**] Daniel Jurafsky and James H. Martin. Forthcoming (2017). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, 3rd ed. Prentice-Hall. Preprint: https://web.stanford.edu/~jurafsky/slp3/
- InfoRetrieval.
- ‡[**CoreNLP**] Stanford CoreNLP. https://stanfordnlp.github.io/CoreNLP/

- †[**TextAsData**] Grimmer and Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*. https://doi.org:10.1093/pan/mps028.
- Quinn-Topics
- FightinWords.
- †[**NLTK-Python**] Jacob Perkins. 2010. *Python Text Processing with NLTK 2.0 Cookbook*. Packt. http://site.ebrary.com.ezaccess.libraries.psu.edu/lib/pennstate/detail.action?docID=10435387; †[**TextAnalytics-Python**] Dipanjan Sarkar. 2016 *Text Analytics with Python* Springer. http://link.springer.com.ezaccess.libraries.psu.edu/book/10.1007%2F978-1-4842-2388-8;
- DSHandbook-Py Ch. 16, "Natural Language Processing."
- Matthew J. Denny. http://www.mjdenny.com/Text_Processing_In_R.html
- CRAN Natural Language Processing: https://cran.r-project.org/web/views/NaturalLanguageProcessing.html
- †[**AudioData**] Dean Knox and Christopher Lucas. 2017. "The Speaker-Affect Model: Measuring Emotion in Political Speech with Audio Data." http://christopherlucas.org/files/PDFs/sam.pdf; https://www.youtube.com/watch?v=Hs8A9dwkMzI.
- †Morgan & Claypool *Synthesis Lectures on Human Language Technologies* http://www.morganclaypool.com/toc/hlt/1/1.
- †Morgan & Claypool *Synthesis Lectures on Speech and Audio Processing* http://www.morganclaypool.com/toc/sap/1/1.

Vision, Image, Video

- ‡[**ComputerVision**] Richard Szelinski. 2010. *Computer Vision: Algorithms and Applications* Springer. http://szeliski.org/Book/
- MixedModels. Ch. 11 "Statistical Analysis of Shape."; Ch. 12 "Statistical Image Analysis."
- *Audio/Video/Volumetric*: See DeepLearning Chapter 9 re convolution.
- †Morgan & Claypool. *Synthesis Lectures on Image, Video, and Multimedia Processing* http://www.morganclaypool.com/toc/ivm/1/1
- †Morgan & Claypool. *Synthesis Lectures on Computer Vision* http://www.morganclaypool.com/toc/cov/1/1

Approaches to Learning from Data (The Analytics Layer)

- NRCreport. Ch. 7. "Building Models from Massive Data."
- MMDS. (data-mining)
- CausalInference.
- ‡[**VisualAnalytics**] Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann, editors. 2010. *Mastering the Information Age: Solving Problems with Visual Analytics* The Eurographics Association: Goslar, Germany. http://www.vismaster.eu/book/. See also †Morgan & Claypool. *Synthesis Lectures on Visualization* http://www.morganclaypool.com/toc/vis/2/1
- ‡[**Econometrics**] Bruce E. Hansen. 2014. *Econometrics* http://www.ssc.wisc.edu/~bhansen/econometrics/

- ‡[**StatisticalLearning**] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R* Springer. http://www-bcf.usc.edu/~gareth/ISL/index.html.
- [**MachineLearning**] Christopher Bishop. 2006. *Pattern Recognition and Machine Learning* Springer.
- †[**Bayes**] Peter D. Hoff. 2009. *A First Course in Bayesian Statistical Methods* Springer. http://link.springer.com/book/10.1007%2F978-0-387-92407-6.
- †[**GraphicalModels**] Søren Højsgaard, David Edwards, Steffen Lauritzen. 2012. *Graphical Models with R* Springer. https://link-springer-com.ezaccess.libraries.psu.edu/book/10.1007%2F978-1-4614-2299-0
- [**InformationTheory**] David Mackay. 2003. *Information Theory, Inference, and Learning Algorithms* Springer.
- ‡[**DeepLearning**] Ian Goodfellow, Yoshua Bengio and Aaron Courville. 2016. *Deep Learning* MIT Press. http://www.deeplearningbook.org/. (see also Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521(7553), 436–444. https://doi.org:10.1038/nature14539.)

  See also ‡[**Keras**] Keras: The Python Deep Learning Library https://keras.io Francoios Chollet. Directory of Keras tutorials: https://github.com/fchollet/keras-resources.
- †[**SignalProcessing**] Jose Maria Giron-Sierra. 2013. *Digital Signal Processing with MATLAB Examples* (Volumes 1-3.) Springer. https://link-springer-com.ezaccess.libraries.psu.edu/book/10.1007/978-981-10-2534-1

Penn State Policy Statements

Academic Integrity

Academic integrity is the pursuit of scholarly activity in an open, honest and responsible manner. Academic integrity is a basic guiding principle for all academic activity at The Pennsylvania State University, and all members of the University community are expected to act in accordance with this principle. Consistent with this expectation, the Universitys Code of Conduct states that all students should act with personal integrity, respect other students dignity, rights and property, and help create and maintain an environment in which all can succeed through the fruits of their efforts.

Academic integrity includes a commitment by all members of the University community not to engage in or tolerate acts of falsification, misrepresentation or deception. Such acts of dishonesty violate the fundamental ethical principles of the University community and compromise the worth of work completed by others.

Disability Accomodation

Penn State welcomes students with disabilities into the Universitys educational programs. Every Penn State campus has an office for students with disabilities. Student Disability Resources (SDR) website provides contact information for every Penn State campus (http://equity.psu.edu/sdr/disability-coordinator). For further information, please visit the Student Disability Resources website (http://equity.psu.edu/sdr/).

In order to receive consideration for reasonable accommodations, you must contact the appropriate disability services office at the campus where you are officially enrolled, participate in an intake interview, and provide documentation: See documentation guidelines at (http://equity.psu.edu/sdr/guidelines). If the documentation supports your request for reasonable accommodations, your campus disability services office will provide you with an accommodation letter. Please share this letter with your instructors and discuss the accommodations with them as early as possible. You must follow this process for every semester that you request accommodations.

Psychological Services

Many students at Penn State face personal challenges or have psychological needs that may interfere with their academic progress, social development, or emotional wellbeing. The university offers a variety of confidential services to help you through difficult times, including individual and group counseling, crisis intervention, consultations, online chats, and mental health screenings. These services are provided by staff who welcome all students and embrace a philosophy respectful of clients cultural and religious backgrounds, and sensitive to differences in race, ability, gender identity and sexual orientation.

- Counseling and Psychological Services at University Park (CAPS) (http://studentaffairs.psu.edu/counseling/): 814-863-0395
- Penn State Crisis Line (24 hours/7 days/week): 877-229-6400
- Crisis Text Line (24 hours/7 days/week): Text LIONS to 741741

Educational Equity

Penn State takes great pride to foster a diverse and inclusive environment for students, faculty, and staff. Consistent with University Policy AD29, students who believe they have experienced or observed a hate crime, an act of intolerance, discrimination, or harassment that occurs at Penn State are urged to report these incidents as outlined on the Universitys Report Bias webpage (http://equity.psu.edu/reportbias/).

Background Knowledge

Students will have a variety of backgrounds. Prior to beginning interdisciplinary coursework to fulfill Social Data Analytics degree requirements, including SoDA 501 and 502, students are expected to have advanced (graduate) training in at least one of the component areas of Social Data Analytics, and a familiarity with basic concepts in the others.

With regard to specialization, students are expected to have advanced (graduate) training in ONE of the following:

- quantitative social science methodology and a discipline of social science (as would be the case for a second-year PhD student in Political Science, Sociology, Criminology, Human Development and Family Studies, or Demography); OR
- statistics (as would be the case for a second-year PhD student in Statistics); OR
- information science or informatics (as would be the case for a second-year PhD student in Information Science and Technology, or a second-year PhD student in Geography specializing in GIScience); OR
- computer science (as would be the case for a second-year PhD student in Computer Science and Engineering).

This requirement is met as a matter of meeting home program requirements for students in the dual-title PhD, but may require additional coursework on the part of students in other programs wishing to pursue the graduate minor.

With regard to general preparation, students are expected to have ALL of the following technical knowledge:

- basic programming skills, including basic facility with R &/or Python; AND
- basic knowledge of relational databases &/or geographic information systems; AND
- basic knowledge of probability, applied statistics, &/or social science research design; AND
- basic familiarity with a substantive or theoretical area of social science (e.g., 300-level coursework in political science, sociology, criminology, human development, psychology, economics, communication, anthropology, human geography, social informatics, or similar fields).

It is not unusual for students to have one or more gaps in this preparation at time of application to the SoDA program. Students should work with Social Data Analytics advisers to develop a plan for timely remediation of any deficiencies, which generally will not require formal coursework for students whose training and interests are otherwise appropriate for pursuit of the Social Data Analytics degree. Where possible this will be addressed at time of application to the Social Data Analytics program.

To this end, some free training materials are linked in the reference section, and there are also abundant free high-quality self-paced course-style training materials on these and related subjects available through edX (http://www.edx.org), Udacity (http://www.udacity.com), Codecademy (http://codecademy.com) and Lynda (http://www.lynda.psu).