The History of Big Data
Professor Benjamin Schmidt
http://benschmidt.org/bigdata20
Email: bs145@nyu.edu
Office Hours: Tuesday, 2-4pm, 20 Cooper Square 520.

Overview

This course places contemporary excitement and fears about "Big Data" in a long historical context. Much is new about the way corporations, governments, and individuals use massive computational resources to search for patterns. But those who use big data draw on legacies from well before the computer age for data management, visualization, and analysis.

We will trace the long history of big data through four parallel strands:

1. The rise of massive systems of data collection by **states** in the 19th century through institutions like the census and the military.
2. The attempts of **businesses** to collect and use data to control their markets and their workers.
3. The relationship of data to the **sciences.**
4. The different eras of **computing** in the last 80 years, and the ways that social forces shaped the development of computing.

This class is listed as a lecture, but will be run in a hybrid lecture-discussion format.

A note on readings and schedules

The schedule printed in this syllabus is likely to change. The course website listed on the front page of the paper documents will reflect the most recent available information.

Course Goals

Like all history courses, this course aims to impart both knowledge about a specific subject and 'transferable' skills.

1. Give you a stronger vocabulary for interacting with data as a primary source and thinking about its origins, setting, and biases.
2. Conduct and communicate archival research.
3. Communicate clearly and respectfully in an oral setting.
4. Write clearly and informatively about non-textual artifacts like datasets, data visualizations, and account books with a focus on clear, succinct, and precise *description*.
5. Debate and describe the ways that contemporary practices of "Big Data" are shaped by and differ from a long historical context;
6. Apply sophisticated historical models of technology shapes social change, and vice versa.
7. Interpret historical sources of data, and recast them into contemporary terms you and your peers can understand; and
8. Understand some of the major turning points in the history of computing, data collection, and social control.

Requirements

Online responses

There are 14 remaining classes in the Coronavirus era of this class.

For five of them, I want you to post a short (~250 words) **reading responses** that zeros in on a point or two in the readings. Don't try to be comprehensive; try to be focused and interesting.

These responses must be posted by 6pm the evening before class.

For a *different* five of them, write a *peer response* to your peers in the time period before class meets. (Leave me a half hour before class to read everything).

Readings and classroom participation

You must complete all the readings for the course and attend class prepared to discuss them. Your peers are counting on you to do so. If for any reason you can't do the reading done by class, you should let me know in advance and still attend class.

This course relies on active, engaged participation in class activities and discussions. We will not be building toward an exam, but we will be calling back through the semester to the base of knowledge we have gained. You should come to every class having read all of the required reading (or watched the required videos, etc.) and prepared to discuss them with your colleagues. We will assess your reading and course engagement through in-class writing exercises (some collected for a grade and others not), reading quizzes, in-class group work, and related assignments.

Maintaining an active class conversation also requires that the class be present, both physicallyvirtually and mentally. To that end: you may miss two classes without penalty over the course of the semester. *Please note:* We make absolutely no distinction between excused and unexcused absences, so use your allotted absences wisely. You may not miss two classes early in the semester and then petition for additional excused absences afterward. When you must miss class, **it is your responsibility to find out what you missed and to make up any pertinent assignments**. You may not make up quizzes or in-class work. If you take one of your excused absences, we simply will not grade any in-class work you missed. If you miss an applied computing activity due to an excused absence you should attempt to make up the work. Once beyond your allotted absences you will receive a zero for any in-class work or computing activities missed.

Participation Self Assessments The second self assessment is in flux as we see how online discussions can be carried out.

It's hard to talk in class. But it's as important a form of intellectual engagement as any other.

You will complete two self assessments of your participation over the course of the semester. Instructions will be passed out in the second week of class.

Note that these are "self assessments," not "self-assessments." That is, I am not asking you to assess *yourself* personally, but to give an honest assessment of the quality and quantity of your engagement and *also* of your peers and myself. We are all working together to build a constructive discussion environment.

In-class responses **These are frozen. We will not be doing any more**.

Several times over the semester, I will ask you to write a short paragraph or two of response at the beginning of class. These will take no more than fifteen minutes. They serve two purposes. One is to get you to think about the issues on your own in a focused way. The other, frankly, is to build a form of incentive to do readings.

I estimate there will be about 10 of these over the course of the semester. The two lowest scores will be dropped.

Archival Project Your first project will grow out of this, and involve a brief in-class presentation followed by a paper on an archival source of data from one of the many outstanding research libraries in the Boston area.

Papers You will write one 6 to 8 page paper for this class, based on the readings; no outside research is expected.

Final Project Final project assignments will be distributed in late March, but you should start thinking early about which one you will want. It will consist of either 1) an 8-10 page paper in which you extend one of the weeks of the course with additional readings; or 2) a digital project in which you analyze a data set created before the year 1994 using modern tools. In either case, you must discuss the project in advance with me.

Behavior

You are required to be respectful to your fellow classmates and professors: listening attentively, not interrupting, and maintaining a civil discourse. Personal attacks, hostility, and mockery will not be tolerated. If you have any issues, please talk to me directly so that I can address them. You are also welcome to consume drinks or minimal food in class, provided they are not distracting. (Use your common sense, please. A quick roll of sushi is fine; a heaping bowl of ramen is not.)

Technical Snafus This course relies heavily on access to computers, specific software, and the Internet. **At some point during the semester you WILL have a problem with technology**: your laptop will crash, a file will become corrupted, a server will go down, a piece of software will not act as you expect it to, or something else will occur. These are facts of twenty-first-century life, not emergencies. To succeed in college and in your career you should develop work habits that take such snafus into account. Start assignments early and save often. Always keep a backup copy of your work saved somewhere secure (preferably off site). None of these unfortunate events should be considered emergencies: inkless printers, computer virus infections, lost flash drives, lost passwords, corrupted files, incompatible file formats. It is *entirely your responsibility* to take the proper steps to ensure your work will not be lost irretrievably; if one device or service isn't working, find another that does. **We will not grant you an extension based on problems you may be having with technological devices or the internet services you happen to use**. When problems arise in the software we are all using for the course, we will work through them together and learn thereby.

Northeastern's Title IX Policy prohibits discrimination based on gender, which includes sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, and faculty and staff. If you or someone you know has been harassed or assaulted,confidential support and guidance can be found through counseling services and religious clergy. By law, those employees are not required to report allegations of sex or gender-based discrimination to the University. Alleged violations can be reported non-confidentially to the Title IX Coordinators. Reporting Prohibited Offenses does NOT commit the victim/affected party to future legal action.

Grading

Grades are going to be weird this semester. These percentages reflect my current thinking, and may change with events before the first of April.

- Presence, participation, and preparation: 10% (Reduced post-COVID.)
- Self assessments: 15%
- In class responses: 4% (Frozen as of COVID time.)
- Reading Responses: 14%
- Peer responses: 7%
- Data Visualization Assignment: 10%
- Archival Presentation: 2%
- Archival Paper: 13%
- Final Paper: 20%

The end of this syllabus includes a longer description of what sort of work will receive an "A," a "B," and so forth.

Schedule

Introductions

Mon, Jan 27 Introductions

Readings

- (In class): Plato, *Phaedrus*, on the invention of writing.

Early Modern Information Overload

Wed, Jan 29 Learning how to Read

Readings

- Ann Blair "Reading Strategies for Coping with Information Overload Ca.1550-1700," *Journal of the History of Ideas* 64, no. 1 (2003): 11–28, https://doi.org/10.1353/jhi.2003.0014.
- Patrick Rael, How to Read a Secondary Source

Mon, Feb 03 Shuffling Paper

Readings

- Staffan Müller-Wille and Isabelle Charmantier "Natural History and Information Overload: The Case of Linnaeus," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43, no. 1 (March 2012): 4–15, https://doi.org/10.1016/j.shpsc.2011.10.021.

Wed, Feb 05 Ordering the Worlds

Readings

- Jorge Luis Borges "The Analytical Language of John Wilkins," trans. Lilia Graciela Vázquez (Alamut, 1999).
- Michel Foucault *The Order of Things: An Archaeology of the Human Sciences* (New York: Vintage Books, 1994)., Introduction and Chapter 3

Preparation

- What do you *understand* about the Foucault?

Mon, Feb 10 Visualization and Images

Readings

- Go back to Foucault and try to come back with one more observation.
- Joseph Priestley *A Description of a Chart of Biography: By Joseph Priestley. ...* (Printed at Warrington, 1764), http://archive.org/details/adescriptionach00priegoog., 1-18. (I'm putting the whole book online: skim the rest if you like. Read this primarily with the Foucault in your head; what's different about Priestley's episteme that he needs to explain timelines this particular way?)
- After reading Priestley, look at the specimen chart from his book https://upload.wikimedia.org/wikipedia/commons/9/98/PriestleyChart.gif and a scan of the full charthttps://pages.uoregon.edu/infographics/timeli of biography.jpg

In class

- William Playfair 1759-1823. *The Commercial and Political Atlas and Statistical Breviary*, ed. Howard Wainer and Ian Spence 1944- (New York: Cambridge University Press, 2005).

Assignment Distributed: First self assessment

Wed, Feb 12 Sharing Knowledge in Early Modern China

Readings

- Schäfer, Dagmar. "Silken Strands: Making Technology Work in China." In Culture of Knowledge: Technology in Chinese History (Leiden: Brill, 2011), pp. 45–73.
- Carla Nappi, "INTERLUDE. A READER'S GUIDE TO THE BENCAO GANGMU." from *The Monkey and the Inkpot*.

Mon, Feb 17 No class: President's Day

Information Managers

Wed, Feb 19 Accounting for Slavery

Readings

- Ellen Gruber Garvey and Lisa Gitelman "'Facts and FACTS' : Abolitionists' Database Innovations," in *"Raw Data" Is an Oxymoron* (Cambridge: MIT Press, 2013), 89–102.
- Sean Wilentz *Major Problems in the Early Republic, 1787-1848: Documents and Essays* (Lexington, Mass.: D.C. Heath, 1992)., "GW Hammond, Instructions to his Overseer"
- Jessica Marie Johnson "Markup BodiesBlack [Life] Studies and Slavery [Death] Studies at the Digital Crossroads," *Social Text* 36, no. 4 (137) (December 1, 2018): 57–79, https://doi.org/10.1215/01642472-7145658.
- (In class: In class: *American Slavery as it is*, runaway advertisements.)

Mon, Feb 24 Industrial Revolutions

Readings

- James R Beniger *The Control Revolution: Technological and Economic Origins of the Information Society* (Cambridge, Mass.: Harvard University Press, 1986)., Chapter 6, "Industrial Revolution and the Crisis of Control"
- D. O. J. "Mercantile Agencies." *New York Daily Times*, November 7, 1851, http://search.proquest.com/hnpnewyorktimesindex/docview/95765241/abstract/142445A46F336CD6D70/11?accountid=12826.
- T. "Mercantile Agencies." *New York Daily Times*, October 29, 1851, http://search.proquest.com/hnpnewyorktimesindex/docview/95772455/abstract/142445A46F336CD6D70/12?accountid=12826.

Wed, Feb 26 Archives

activity: New York Public Library, archival document session.

Mon, Mar 02 State Capacity

Readings

- James C. Scott *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (New Haven: Yale University Press, 1999)., pp. 11-52, 64-83. (i.e.; Chapter 1, and the last half of chapter 2). On Google Drive, and available online from campus at https://www.jstor.org/stable/10.2307/j.ctt1nq3vk.

note: The Scott is full of some really Big Ideas that we need for the rest of this class, told through several amazingly divergent stories about particular areas (Germany forestry, French land taxes, Filipino surnames, Parisian Streets, and so forth.) Some of these–especially the idea of "legibility"–do not show up until the very end of these selections. The details are fascinating and help you understand the issues; but the specifics here are less important than in, say, Beniger. Do not lose sight of the forestry for the trees.

Wed, Mar 04 CLASS CANCELLED, instructor illness.

Mon, Mar 09 The Census

Readings

- Thomas P. Kinnahan "Charting Progress: Francis Amasa Walker's Statistical Atlas of the United States and Narratives of Western Expansion," *American Quarterly* 60, no. 2 (2008): 399–423, https://doi.org/10.1353/aq.0.0012.
- Margo J Anderson *The American Census: A Social History* (New Haven: Yale University Press, 1988)., chapter on industrial census.
- Bring the Scott readings as well.

Cultures of Data/BEGINNING OF CORONAVIRUS REMOTE INSTRUCTION.

Wed, Mar 11 Fordism

Readings

- Stephen Meyer *The Five Dollar Day: Labor Management and Social Control in the Ford Motor Company, 1908-1921*, Suny Series in American Social History (Albany, N.Y: State University of New York Press, 1981).

in_class: chaplin_modern_1936, first fifteen minutes

Mon, Mar 16 No class: Spring Break

Wed, Mar 18 No class: Spring Break

Mon, Mar 23 Ordinary Americans

activity: Please post online onto NYU courses an image of something interesting/unexpected from your archival dataset. We will talk about these in small groups on Zoom for the first 45 minutes. Then we will talk about the Igo readings about Middletown for about 45 minutes.

Readings

- Sarah Elizabeth Igo *The Averaged American: Surveys, Citizens, and the Making of a Mass Public* (Cambridge, Mass: Harvard University Press, 2007)., Introduction, Chapter 1 and 2. Note that all six chapters of the Igo are online in the Google docs to accomodate anyone who cannot access the online versions through the NYU library. But I would prefer you access directly through the NYU library, where you can read/download the full book with NYU login.. Please be in touch if you have trouble with network issues, etc.

Wed, Mar 25 Quantifying Publics

activity: In class presentations

Readings

- Igo, Chapters 3-4 and epilogue.

Mon, Mar 30 State statistics

Reading

- Ghosh, Arunabh. 2018. "Lies, Damned Lies, and (Bourgeois) Statistics: Ascertaining Social Fact in Mid-century China and the Soviet Union." Osiris 33 (1): 149-168

Computing Culture

Wed, Apr 01 Imagining Computers

Readings

- Vannevar Bush "As We May Think," *The Atlantic*, July 1945, http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.
- Vannevar Bush "Memex Revisited," in *From Memex to Hypertext*, ed. James M. Nyce and Paul Kahn (San Diego, CA, USA: Academic Press Professional, Inc., 1991), 197–216, http://dl.acm.org/citation.cfm?id=132180.132193.

Mon, Apr 06 Making Programmers

Readings

- "The Computer Girls", *Cosmopolitan*, 1967
- Jennifer Light "When Computers Were Women," *Technology and Culture* 40, no. 3 (1999): 455.
- *Desk Set* (20th Century Fox, 1957)., in class

Wed, Apr 08 Data-Mania

Readings

- Arthur Raphael Miller *The Assault on Privacy: Computers, Data Banks, and Dossiers* (Ann Arbor, University of Michigan Press, 1971), http://archive.org/details/assaultonprivacy00mill., pp. 34-69, 254-274

Mon, Apr 13 Punching in

Readings

- Stephen Lubar, "Do not fold, spindle, or mutilate" 1992. link
- More time on Miller and privacy.

Personal Computing

Wed, Apr 15 Database Populism

Readings

- Database Populism
- Ted M. Lau, "Total Kitchen Information System", *Byte Magazine*, 1977

Mon, Apr 20 The Spreadsheet

Readings

- Stephen Levy, "A Spreadsheet way of knowing", *Harper's Magazine*, 1984. (You can find a newer copy of this republished online, but read the PDF of the original.

Wed, Apr 22 Surveillance Statism

Readings

- Following up from Monday: on your time, go to https://archive.org/details/mac_Paint_2. Make a drawing and also save a file, exploring the operating system.
- This American Life, Photo Op. (Episode 493). https://www.thisamericanlife.org/493/picture-show/act-one-0

Mon, Apr 27 The Information Superhighway

Readings

- Tim Berners-Lee and Mark Fischetti *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor* (San Francisco: HarperSanFrancisco, 1999)., Introduction; Chapters 1, 2, and 3

Social Computing

Wed, Apr 29 Information Overload Revisited

Readings

- James Gleick *The Information: A History, a Theory, a Flood* (New York: Pantheon Books, 2011).
- Siva Vaidhyanathan *The Googlization of Everything: (And Why We Should Worry)* (Berkeley: University of California Press, 2011)., Chapter 2

Mon, May 04 Big Data and the Sciences

Readings

- The End of Theory, Wired Magazine, 2008 link
- The Norvig-Chomsky Debate, 2012: Norvig and Chomsky

Wed, May 06 Coronadata

Readings

- Feb 15: To Tame Coronavirus, Mao-Style Social Control Blankets China https://www.nytimes.com/2020/02/15/business/china-coronavirus-lockdown.html

Mon, May 11 Surveillance Capitalism

Readings

- Shoshana Zuboff, *Surveillance Capitalism*, excerpts.

2020-05-15 Assignment: Final Papers due over NYU classes by 5pm. I will respond to any requests for comments on draft first paragraphs sent by 4pm, Tuesday.

Policies

Academic Integrity

Plagiarism is the serious intellectual sin in the humanities. All work you submit must be your own in accordance with CAS guidelines (https://cas.nyu.edu/content/nyu-as/cas/academic-integrity.html).

Accessibility

Please inform me privately as soon as possible if you needs that need accomodating.

New York University provides reasonable accommodations to qualified students who disclose their disability to the Moses Center. Reasonable accommodations are adjustments to policy, practice, and programs that provide equal access to NYU's programs and activities. Accommodations and other related services are determined on a case-by-case basis, taking into consideration each student's disability-related needs and NYU program requirements.

Religious observances

Should a due date or class meeting fall on a religious observance that is not an NYU holiday, please let me know and we can make accomodations. NYU's policy on religious observances is online: https://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/university-calendar-policy-on-religious-holidays.html.

Counseling and Wellness

If you experience any health or mental health issues during this course, I encourage you to utilize the support services of the 24/7 NYU Wellness Exchange 212-443-9999.

If you are having mental health problems that are preventing you from attending class or completing assignments, please let me know as soon as possible.

Use of Electronic Devices

Laptops and tablets are allowed in class, and it is permissible to use them to refer to notes and readings. Nonetheless, I strongly encourage you to print out readings if you are able to do so; if you find the expense prohibitive, I am happy to print up to three week's worth of readings in advance for any student who comes to office hours. At some points in discussions, I will ask everyone to put all screens away.

Web browsing, e-mail, etc., are not allowed. **Not even when the activity is directly related to class discussion.** If you think it's critically important that you get a reference from Wikipedia or wherever to contribute to class, you must ask first.

Phones must remain in bags, pockets, etc. If I see you using a cell phone, I will mentally note a zero for the day in class participation. I may ask you to put it away, but often I will not say anything because to do so would be insulting to the peers you are ignoring.

You are not as sneaky texting under the table as you think you are.

Acknowledgements

Elements of this class draw on courses by Emily Thompson, Shannon Mattern, Lauren Klein, and others.

Language in this syllabus comes from a variety of other sources, especially Ryan Cordell at Northeastern University.

The paper syllabus uses a template by Andrew Goldstone.

Anderson, Margo J. *The American Census: A Social History*. New Haven: Yale University Press, 1988.

Beniger, James R. *The Control Revolution: Technological and Economic Origins of the Information Society*. Cambridge, Mass.: Harvard University Press, 1986.

Berners-Lee, Tim, and Mark Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. San Francisco: HarperSanFrancisco, 1999.

Blair, Ann. "Reading Strategies for Coping with Information Overload Ca.1550-1700." *Journal of the History of Ideas* 64, no. 1 (2003): 11–28. https://doi.org/10.1353/jhi.2003.0014.

Borges, Jorge Luis. "The Analytical Language of John Wilkins." Translated by Lilia Graciela Vázquez. Alamut, 1999.

Bush, Vannevar. "As We May Think." *The Atlantic*, July 1945. http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

———. "Memex Revisited." In *From Memex to Hypertext*, edited by James M. Nyce and Paul Kahn, 197–216. San Diego, CA, USA: Academic Press Professional, Inc., 1991. http://dl.acm.org/citation.cfm?id=132180.132193.

Desk Set. 20th Century Fox, 1957.

Foucault, Michel. *The Order of Things: An Archaeology of the Human Sciences*. New York: Vintage Books, 1994.

Garvey, Ellen Gruber, and Lisa Gitelman. "'Facts and FACTS' : Abolitionists' Database Innovations." In *"Raw Data" Is an Oxymoron*, 89–102. Cambridge: MIT Press, 2013.

Gleick, James. *The Information: A History, a Theory, a Flood*. New York: Pantheon Books, 2011.

Igo, Sarah Elizabeth. *The Averaged American: Surveys, Citizens, and the Making of a Mass Public*. Cambridge, Mass: Harvard University Press, 2007.

J., D. O. "Mercantile Agencies." *New York Daily Times*. November 7, 1851. http://search.proquest.com/hnpnewyorktimesindex/docview/95765241/abstract/142445A46F336CD6D70/11?accountid=12826.

Johnson, Jessica Marie. "Markup BodiesBlack [Life] Studies and Slavery [Death] Studies at the Digital Crossroads." *Social Text* 36, no. 4 (137) (December 1, 2018): 57–79. https://doi.org/10.1215/01642472-7145658.

Kinnahan, Thomas P. "Charting Progress: Francis Amasa Walker's Statistical Atlas of the United States and Narratives of Western Expansion." *American Quarterly* 60, no. 2 (2008): 399–423. https://doi.org/10.1353/aq.0.0012.

Light, Jennifer. "When Computers Were Women." *Technology and Culture* 40, no. 3 (1999): 455.

Meyer, Stephen. *The Five Dollar Day: Labor Management and Social Control in the Ford Motor Company, 1908-1921*. Suny Series in American Social History. Albany, N.Y: State University of New York Press, 1981.

Miller, Arthur Raphael. *The Assault on Privacy: Computers, Data Banks, and Dossiers*. Ann Arbor, University of Michigan Press, 1971. http://archive.org/details/assaultonprivacy00mill.

Müller-Wille, Staffan, and Isabelle Charmantier. "Natural History and Information Overload: The Case of Linnaeus." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43, no. 1 (March 2012): 4–15. https://doi.org/10.1016/j.shpsc.2011.10.021.

Playfair, William, 1759-1823. *The Commercial and Political Atlas and Statistical Breviary*. Edited by Howard Wainer and Ian Spence 1944-. New York: Cambridge University Press, 2005.

Priestley, Joseph. *A Description of a Chart of Biography: By Joseph Priestley. ...* Printed at Warrington, 1764. http://archive.org/details/adescriptionach00priegoog.

Scott, James C. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press, 1999.

T. "Mercantile Agencies." *New York Daily Times*. October 29, 1851. http://search.proquest.com/hnpnewyorktimesindex/docview/95772455/abstract/142445A46F336CD6D70/12?accountid=12826.

Vaidhyanathan, Siva. *The Googlization of Everything: (And Why We Should Worry)*. Berkeley: University of California Press, 2011.

Wilentz, Sean. *Major Problems in the Early Republic, 1787-1848: Documents and Essays*. Lexington, Mass.: D.C. Heath, 1992.