

INFS 289-5 — Data Informatics

Draft January 1st, 2016.

Course information:

Number: INFS 289-5
Title: Data Informatics
ID: 628535205
Quarter: Winter 2016
Location: 245, GSE&IS Bldg.
Time: Tuesdays, 9h-12h30.
Web site: <http://ccl.e.ucla.edu/course/view/16W-INFSTD289-5/>

Instructor information:

Instructor: Jean-François Blanchette
Office: room 218, GSE&IS Bldg.
Phone: 310 267 5137
Fax: 310 206 4460
Email: blanchette@ucla.edu
Web: <http://polaris.gseis.ucla.edu/blanchette>
Office hours: Tuesdays 1-3

1. Overview

Under the name of ‘Big Data’ and ‘self quantification’, we are currently under the sway of an avalanche of numbers of epic proportions, a process that aims to reconstruct no less than the very nature of personal and institutional decision making. This deluge of data results from the convergence of several trends and data sources, among others the proliferation of mobile devices and of cheap sensors wirelessly connected; the ease of data collection in online environments; mass digitization projects, the structuration of previously unstructured data sources, and massive data creation projects (e.g., Google Street View); crowdsourcing of reviews, rankings and ratings; the Open Data movement and the availability of new tools for data processing.

Each of these sources and trends offer opportunities for those who seek to leverage data’s capacity in the service of personal and institutional change. To do so however, it is necessary to go beyond the prevalent portrayal of data as immaterial, untainted, and naturally fluid information that objectively records characteristics of phenomena—something we might call ‘data positivism.’¹ Such a characterization actively obscures the careful and attentive labor that is necessary to turn phenomena into decisions. Elaborate social and technical processes, including sensing and measurement, standardization, normalization, aggregation, classification, description, algorithmic processing, and curation, must intervene before an event can be written into a database and form the basis of a decision by an individual or an institution.

¹ See for example James Gleick, *The Information: A History, A Theory, A Flood* (Pantheon, 2011) for an eloquent presentation of this view.

This course will analyze the current deployment of data-driven decision making systems from an informatics perspective, a perspective grounded in three premises: (a) information systems are shaped by both technological evolution, individual practices, and institutional norms; (b) this shaping is highly contextual and (c), it proceeds from the actions of multiple stakeholders with highly diverse (and divergent) interests. As well, instead of a positivist view of data, this course will assume instead that:

1. Data is not a thing or rather, everything is potentially data. Instead of asking “What is data?” it is therefore more productive to ask “When is data?” that is, to follow the processes by which an inscription becomes actionable;²
2. Data has mass and its production, circulation, and consumption therefore consumes (scarce) resources;
3. Data must always be structured at some level and its production, circulation, and consumption therefore requires substantial technical, institutional, and cultural work;
4. Data is never purely descriptive, but also actively shapes the worlds it seeks to describe.

From these premises, this course will provide practical and conceptual tools for participants to explore questions operating on four different registers:

1. *Infrastructure*: What kind of data are individuals and institutions currently collecting, using what means, and for what purposes? What infrastructure (technical, institutional) make such collection possible?
2. *Power*: How are these data leveraged by stakeholders to alter the balance of power in decision making processes?
3. *Warrant*: How does these data practices conform to, challenge, and/or shape current standards and norms with respect to personal/institutional recordkeeping, accountability, privacy, and transparency?
4. *Prospective*: What new data might allow individuals and institutions to better fulfill their goals and missions? What new infrastructure (e.g., computational, institutional, cultural) would it require?

In exploring these questions, course participants will read about data and the design of information design from a broad range of perspectives. In particular, we will leverage previous historical examples when new modes of quantification led to massive institutional, social and cultural change (e.g., mortality tables and the insurance industry), work from scholars in Science and Technology Studies who have paid careful attention to intellectual and institutional labor involved in the circulation of data, and of course, insights from Library and Information Science with respect to, for example, data description and curation.

² With thanks for Jérôme Denis and Samuel Goëta for this formulation.

2. Course outcomes and professional preparation

The course will prepare participants for a wide range of opportunities in areas of professional practice that relate to open data, digital curation and preservation, data science, human-computer interaction, records management, information governance, management of archives and libraries, etc. Upon satisfactory completion of this course, participants will have demonstrated their understanding of, and familiarity with:

- The process of quantification and production of data from a cultural, institutional, and technical perspective;
- The data lifecycle—creation, description, classification, standardization, processing, visualization/action, annotation, duration, storage—and the opportunities and challenges for institutional/personal change afforded at each step;
- How to analyze, question, and advocate for processes of quantification and statistical decision making;
- The notion of the ‘residual’, i.e., the phenomena which fails to be captured by sensors, data models, classification systems, algorithms, data visualizations, etc.;
- The role of ‘data friction’³ in the systems development process;
- How numbers/data become (or fail to become) trustworthy at each step of the lifecycle;
- The relationship between news forms of data collection, processing, and distribution, and new forms of statistical processing (e.g., machine learning), software packages, and hardware platforms;
- The development of new modes of institutional governance appropriate to data-driven decision making and their effects on power relations among stakeholders.

3. Method

The course will rely on the three distinct pedagogical methods.

- (1) In class discussion of assigned readings that will help us gradually develop a theoretical and practical framework for an informatics approach to data-driven decision making systems. To keep our assumptions in check, we will also discuss data-related current events drawing from blogs, the popular press, film, novels, etc.;
- (2) Throughout the term, invited speakers (from either academia, industry, or other relevant institutions) will share about their data-centered projects. We will use these presentations to get an on-the-ground understanding of how data is currently used by businesses and institutions, the obstacles, the

³ “The costs in time, energy and attention required simply to collect, check, store, move, receive, and access data.” Paul Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. MIT Press, 2010, p. 84.

- relation between real world phenomena and their representation as data, the inferences made from the data;
- (3) Assignment — There will be a single class project (see “Final Assignment” handout), counting for 60%, and 40% awarded for class attendance and participation. Each week, we will collectively discuss the application of the readings to the final project, report on progress and obstacles, and draft the deliverable. All written materials will be shared with all course participants.

4: Course Requirements

- Come to class prepared to discuss the readings. See “How to Read a Book,” (<http://pne.people.si.umich.edu/PDF/howtoread.pdf>).
- Forfeit the use of your laptop and other electronic devices during class lectures. In group exercises, one laptop per team is allowed.
- Participate in discussions. You are particularly encouraged to question the assumptions of the readings, the instructor, and your fellow students, as long as you do so respectfully. In doing so, you will sharpen your ability for critical thinking, innovation, debate, and public speaking, skills fundamental to your future professional life.
- Written work should be of high quality. If you have concerns about writing, address them early. A useful resource is UCLA’s Graduate Writing Center (<http://gsrc.ucla.edu/gwc/>).
- Assignments must be turned in according to the scheduled due dates. In particular, no incompletes will be given.
- If you feel that you may need an accommodation for a disability or have any other special needs, make an appointment to discuss this with the instructor. I will best be able to address special circumstances if I know about them early in the term. The website for the UCLA Office for Students with Disabilities (www.osd.ucla.edu) contains a wealth of useful information as well as official policies about this issue.

5. Schedule of readings and speakers

Week 1 (January 5): Beyond Data Positivism

This lecture will survey the concepts that undergird the current resurgence of data positivism and outline what an informatics-based approach offers.

Scott K. Johnson, “[The little box that controls half of your home’s energy use.](#)” *Ars Technica*, Dec. 14, 2015.

Kenneth Cukier and Viktor Mayer-Schönberger, “The Rise of Big Data,” *Foreign Affairs*, May/June 2013, pp. 28-40.

Paul Edwards, "Friction," in *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. MIT Press, 2010

See also:

Lisa Gitelman and Virginia Jackson, "Introduction," in *"Raw Data" is an Oxymoron*, (Lisa Gitelman, ed.), pp. 1-14, The MIT Press, 2012.

Danah Boyd & Kate Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," (June 2102) *Information, Communication, and Society* 15(5):662-679.

Daniel Rosenberg (2012), "Data Before the Fact," in *"Raw Data" is an Oxymoron*, (Lisa Gitelman, ed.), pp. 15-40, The MIT Press, 2012.

Week 2 (January 12): Where is data?

This lecture will survey the wide range of sources from which data is currently obtained (including sentiment analysis) and the various challenges such data collection faces, including noise, missing or faulty infrastructure, institutional obstacles, etc.

Readings

James Rufus Koren, "[Some lenders are judging you on much more than finances](#)", Los Angeles Times, December 19, 2015.

James Somers, "Why New York Subway Lines are Missing Countdown Clocks," *The Atlantic*, November 13, 2015.

Kathryn Gates, Dawn Wilkins, Sumali Conlon, Susan Mossing and Maurice Eftink, "Maximizing the Value of Student Ratings Through Data Mining," in *Educational Data Mining: Applications and Trends*, (Alejandro Peña-Ayala, ed.) Springer, 2014.

See also:

Seth Roberts, "Self-experimentation as a source of new ideas: Ten examples about sleep, mood, health, and weight" (2004). *Behavioral and Brain Sciences* 27(2):227-288.

Week 3 (January 19): Quantifying

Christopher Mims, "Creating a 'Fire Alarm' for Terrorist Attacks," *The Wall Street Journal*, November 23, 2015.

Rob Walker, "The Song Decoders," *The New York Times Magazine*, October 14, 2009. <http://www.nytimes.com/2009/10/18/magazine/18Pandora-t.html>

Alexis C. Madrigal, "How Netflix Reversed Engineered Hollywood," *The Atlantic*, January 2, 2014.

<http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/>

Theodore M. Porter, "Making Things Quantitative," (1994) *Science in Context*, 7(3):389-407. (Q175.4 .S343)

"Rarely, if ever, are preexisting qualities simply made more precise by being quantified. At issue, rather, is the creation of new entities, made impersonal and (in this sense) objective when widely scattered people are induced to count, measure, and calculate in the same way."

Joseph O'Connell, "Metrology: The Creation of Universality by the Circulation of Particulars" (1993) *Social Studies of Science* 23:129-173.

Christopher Groskopf, "The Quartz Guide to Bad Data," *Quartz*, December 15, 2015. <http://qz.com/572338/the-quartz-guide-to-bad-data/>

Week 4 (January 26): Opening

Guest speaker: Morgan Currie, Doctoral Candidate, UCLA Department of Information Studies.

Statewide Open Data Portal

<http://www.legtrack.com/bill.html?bill=201520160SB573>

Larry Lessig, "Against Transparency," *The New Republic*, October 9, 2009.

<http://www.newrepublic.com/article/books-and-arts/against-transparency>

Samuel Goëta and Tim Davies, "The Daily Shaping of State Transparency: Standards, Machine-Readability, and The Configuration of Open Government Data Policies."

Jérôme Denis and Samuel Goëta, "Exploration, Extraction and 'Rawification'. The Shaping of Transparency in the Back Rooms of Open Data."

Week 5 (February 2): Standardizing

Alain Desrosières, "How to be Real and Conventional: A Discussion of the Quality Criteria of Official Statistics." *Minerva* (2009) 47:307-322.

James C. Scott (1998), "Nature and Space" and "Cities, People, and Language," in *Seeing Like a State*, Yale University Press.

Paul Edwards, "Making Global Data", chapter 7 in *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*, pp. 187-227. Cambridge, Mass.: The MIT Press, 2010.

Paul Edwards, "Making Data Global", chapter 10 in *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*, pp. 250-285. Cambridge, Mass.: The MIT Press, 2010.

Week 6 (February 9): Trusting

Guest speaker: Lisa Federer, Research Data Informationist, National Institutes of Health Library.

Jeff Elder, "Inside a Twitter Robot Factory," *Wall Street Journal*, Nov. 24, 2013.

David Streitfeld, "The Best Book Reviews Money Can Buy," *New York Times*, Aug. 25, 2012.

Joseph M. Reagle, Jr. "Manipulated: 'Which Ice Cube is the Best?'," in *Reading the Comments: Likers, Haters, and Manipulators at the Bottom of the Web*. The MIT Press, 2015, pp. 43-72.

Week 7 (February 16): Fairness, ethics, and reactivity

Brian K. Roberts, "Will Traffic NIMBYs Ruin Waze?," *Latimes.com*, accessed November 30, 2015,

Joe Flint, "In L.A., One Way to Beat Traffic Runs Into Backlash," *Wall Street Journal*, November 14, 2015, sec. Page One.

Gingras, Yves. "Criteria for Evaluating Indicators." In *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, edited by Blaise Cronin and Cassidy R Sugimoto, 109–26. MIT Press, 2014.

Wendy Nelson Espeland and Michael Sauder, "Rankings and Reactivity: How Public Measures Recreate Social Worlds," *American Journal of Sociology* 113, no. 1 (2007): 1–40.

Week 8 (February 23): Opposing

Guest speaker: Brittany Paris, Jennifer Pierre,, etc. UCLA Department of Information Studies.

Jen Graves, "[What only artists can teach us about technology, data, and surveillance](#)" *The Stranger*, December 10, 2104.

Kitchin, Rob and Lauriault, Tracey P., [Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work](#) (July 30, 2014). The Programmable City Working Paper 2; pre-print version of chapter to be published in Eckert, J., Shears, A. and Thatcher, J. (eds) *Geoweb and Big Data*. University of Nebraska Press. Forthcoming.

Finn Brunton and Helen Nissenbaum (2013), "Political and Ethical Perspectives on Data Obsfucation," in *Privacy, Due Process and the Computational Turn* (Mireille Hildebrandt and Katja de Vries, eds.)

Week 9 (March 1): Prospecting

David Kravets, "How the NFL—Not the NSA—is Impacting Data Gathering Well Beyond the Gridiron," *Ars Technica*, September 13, 2015.
<http://arstechnica.com/tech-policy/2015/09/the-nfl-is-reshaping-the-surveillance-society-xbox-one-experience-and-gambling/>

Melanie Swan, "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery," (2013) *Big Data* 1(2):85-99.

Phil Agre, "Living Data," (November 1994) *Wired* 2(11).
<http://www.wired.com/wired/archive/2.11/agre.if.html>

Week 10 (March 8)

Collective presentation of final report to the Faculty.

INFS 289-5 — Data Informatics: Final Assignment

*Jean-François Blanchette,
Draft Fall 2015*

1. Overview

The class assignment will consist in performing a complete data audit of the department and offer recommendations with respect the implications of current and future data practices for governance of the Department. The audit will seek to answer the following four questions:

1. What data is currently collected by the Department, using what means, and for what purposes?
2. How are these data leveraged in the process of individual and institutional decision making by Departmental stakeholders?
3. How does these data practices conform to or challenge current institutional norms with respect to recordkeeping, personal and institutional accountability, privacy, and transparency?
4. How can the Department leverage this (or other) data to better fulfill its institutional mission?

The empirical and analytical material generated these questions will be integrated in a report to be presented to the Chair and Faculty of the Department.

2. Rationale for site

The Department is an excellent real-life example of the kind of institutional context in which graduates of the program are most likely to perform their professional duties: (a) it is not overly large, with 5-6 staff, 12 core faculty, and 100-200 students; (b) it collects and uses data from a wide range of sources, embedded in institutional contexts at many levels of the UC systems (in-house, school-wide, campus-wide, UC-wide, federally mandated, etc.); (c) it collects, stores, exchanges, processes, displays and provide access to the datasets using a broad range of technological systems, themselves governed by distinct institutional units (e.g., the Dept. itself, graduate division, Office of the Registrar, GSE&IS financial office, US Dept. of education, etc.); (d) datasets and the practices associated with them are governed by the particular interests of various stakeholders, including staff, faculty, university administration on multiple levels, students, parents, etc.

Working with the Department offers multiple advantages: (1) ease of access and coordination of work among students; (2) immediate relevance to students' perspective, as stakeholders in the data practices they are analyzing; (3) leverages and enhances the skills of IS students in the service of the UCLA IS community; (4) immediate experience of the need to manage power relations (including issues of confidentiality, institutional authority, etc.), and how these must be accounted for in both the crafting and presentation of findings and recommendations. While some students may find it challenging to conduct such

an audit where they are involved with many stakeholders in a number of overlapping professional relationships, awareness of these constraints, and reflection on how to work with and around them is itself a professional skill that will be usefully develop by students in this class. Indeed, beyond the theoretical, methodological, and empirical knowledge generated by the assignment, the audit will allow class participants to practice several important professional skills in a real-world context, including: business writing, interviewing, negotiating diverse and conflicting stakeholders interest in consultancy work, working in teams under tight deadlines, public speaking and presentation, and prospective analysis.

3. Process and structure of the audit

The particular structure of the assignment will depend on the number of students enrolled. All class participants will be involved in the assignment, which will be coordinated by the instructor. Each section of the report will be assigned to a team (possibly of one), who will be responsible for researching, writing up, and editing of that section. The audit will proceed in four separate phases:

(1) Identification and data gathering

In the first phase, project teams will identify an initial list of datasets that are central to the operation of the Department. They will gather preliminary information about these datasets from existing sources, including legislation, research papers, online sources, etc. They will identify individuals in the organization responsible for, and stakeholders impacted by, these datasets.

Through interviews with stakeholders, participant-observation, analysis of technical systems, and other methods, project teams will gather empirical material relative to the context of production, standardization, storage, access, circulation, visualization, and long-term preservation of the data. In particular, project teams will:

- Identify all datasets currently managed by the Department;
- Identity the institutional warrant (if any) for these datasets, at the legislative, administrative, and cultural level;
- Describe the workflow of these datasets, from capture to consumption by users, including description, standardization, visualization,
- Describe the technological context for these datasets;
- Identify stakeholders with respect to data practices in the Dept.: students, staff, faculty, general public, etc. What kind of power does each group have with respect to data practices, how are they impacted by the evolution of such practices;
- Describe the conditions (institutional, technological, etc.) for user access, (including visualization) of these datasets;
- Identify current institutional circumstances at the Departmental, School, campus, and UC level that impact data governance practices;
- Identify how affect, cultural values, and cognitive dispositions may impact data practices;

(2) Contextual and workflow analysis

In the second phase, teams will use the empirical material they have gathered to analyze the context and workflow of the datasets. They will determine whether additional information (e.g., follow-up interviews) are required. They will write-up the result of their analysis in a form suitable for the final report. The analysis should:

- Include a description of the current mode of governance of the Department and the role played by each particular dataset;
- Articulate and identify concerns with the impact of current Departmental data practices on recordkeeping norms, privacy, accountability, and transparency;
- Describe at what point are datasets leveraged as evidence in institutional decision making;
- Identify obstacles and enabling factors to effective data governance.

(3) Prospective analysis and recommendations

In the third phase, teams will gather the result of their analysis and collectively propose:

1. Engage in prospective analysis and propose future scenarios for the Department's engagement with data governance, given the possible evolution of technologies, cultural norms, legislation, institutional and personal data collection practices, etc.
2. Recommendations for actions by the various stakeholders, in order to achieve proposed outcomes, including next steps, long term vision, and existing opportunities that can be leveraged.

(4) Delivery and presentation of the report

The final deliverable will be a report presented to the Chair and Faculty of the Department. The report will consist of (a) a cover letter, explaining the context for the report; (b) an executive summary outlining the objectives, process, and main recommendations of the report; (c) introduction, analysis, and conclusion; (d) prospective analysis; and (e) recommendations.

During the final class, participants will present the main finding and recommendations of the report to a panel of MLIS students, PhDs and faculty.

5. Timeline/deadlines

A portion of every class will be devoted to work on the project, beginning with the first class and ending with a final group presentation of the report in the final class.