UCLA Departments of Design & Media Arts, Statistics, & Information Studies present:

Data design & aesthetics, Spring 2007

(Information Studies IS 274; Design|Media Arts 259-M; Statistics: M237)

Instructors
Mark Hansen (cocteau@stat.ucla.edu)
Jean-François Blanchette (blanchette@ucla.edu)
Office hours: (Mark), Tuesdays 1-3 (Jean-François)

Class hours: Tuesday, 9h00-12h30
Class location: 5261 Broad Building.
Lab hours: Tuesday, 6-7, Instructional Computing Lab, Boelter Hall 9413
Map: http://www.computerlabs.ucla.edu/Map.asp

Course homepage: http://courses.gseis.ucla.edu/course/view.php?id=152

Approach and Objectives

Almost every aspect of our lives is, in some form or another, captured, described, and rendered in data. New technologies for collection (e.g., embedded sensors), echange (the Internet), and display (e.g., GIS) have generated an explosion of data. Today, professional, research, and creative practices increasingly depend on data and data processing, on the ability to understand and manipulate of large datasets, on drawing conclusions from or in some way adapting to complex quantitative observations of the physical world, on organizing, describing, exchanging, preserving, and searching vast amount of digital resources. For example:

- New data collection technologies have made it easy to record continuous, high-resolution measurements of our physical environment (weather patterns, seismic events, the human genome);
- Embedded sensors (e.g., GPS) enable the constant monitoring our movements through and interactions with our physical surroundings (automobile and air traffic, large-scale land use, advanced manufacturing facilities);
- Interactions in computer-mediated settings depend crucially on or consist entirely of complex digital data (networked games, peer-to-peer technologies, Web site and Internet usage);
- Access, use, and preservation of digital resources is entirely predicated on their description, organized through metadata schemas, that promise to greatly expand the interoperability of information resources across systems and time boundaries
- Media art relies on complex data sets generated through

This course will address some significant stops along the "data pipeline", from collection technologies, to transmission, storage, visual analysis, modeling and decision-making. Some of the questions we will address along the way:

- How do physical objects, phenomena, and people get translated into objective measurements/descriptions? How can a measurement be understood as a social object?

- What are the competing models (deterministic mathematical, probabilistic, or data-based representations) for objects and phenomena and how do they organize or "expose" the information they carry?

- Who has access to the data or views of the data, and at what resolution; and what is the role of legislation in setting these limits? What technologies might promote the sharing of data in "safe" ways? Is there, or should there be, a centralized authority that is charged with data collection and dissemination or are data produced and organized in a more organic fashion?

- How are data or derived analyses presented to the general public for decision making? How are decisions made or, abstractly, how are optimization problems solved, when the underlying data are noisy or uncertain?

- If the flow consists of large quantities of complex, dynamic data, how do "subscribers" understand or express patterns or regularities, and how do these forms of expression affect their view of the phenomena being described?

The insight that guides this course is that decisions along the pipeline should not be made in an isolated way: choosing a database schema, data formats and protocols, ultimately decide the kind of analysis that can be performed; and, run in reverse, modeling often drives choices about what data to collect and how it is represented. While this interconnectedness may seem an obvious finding, it is rarely made explicit when training researchers whose practices tap into and contribute to the flow of data. Because no single discipline can claim ownership of the entire "pipeline", translating this observation into a deep understanding of the nature of the relationships that operate the data pipeline requires an intersciplinary approach and commitment to both critical and practical exploration of technologies and theories.

Recommended textbooks:
David M. Kroenke, *Database Concepts*, 3rd Edition, Prentice Hall, 2007
Excellent and succint exposition of relational database design concepts.

Recommended software/skills:
http://www.processing.org

Evaluation
The main evaluation component of the course will be a group project (4-5 students) that will explore some of the issues touched in the class. The project will comprise a design|media art component, a statistics component, and a IS component (with relative weights depending on the specific group). It should innovate in at least one of those dimensions. Your innovation could be a new measurement, a new query, or a new visualization of the data, or any combination of those. Examples of possible projects will be presented during lectures/discussions. **The theme for this iteration of the class will be Forgetting/Remembering.**

Lab sessions

During lectures, we will provide some basic understanding of technical topics like data modeling (ER diagrams), SQL, statistical programming and visualization, and GIS. We will hold practical lab sessions Tuesdays 6-7 in the Statistics Department's Teaching facility.

Week 1 (Tue. April 3): Overview

- 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), pp. 227-288.

- Luebke, D. M. and S. Milton (1994). "Locating the victim: An overview of census-taking, tabulation technology, and persecution in Nazi Germany." *IEEE Annals of the History of Computing* **16**(3): 25-39.

- Lev Manovich, "Database as Symbolic Form", *Convergence*, **5**(2):80-99.

Additonal readings

- "Math will rock your world", *Business Week*, January 23 2006. (be sure to read the comments as well).

Week 2 (Tue. April 10): The origin of data

How do physical objects, phenomena, and people get translated into "objective" measurements and descriptions? How can a measurement be understood as a social object?

- Hacking, I. (1990). *The taming of chance*. Cambridge University Press.
  Chapter 3: "Public amateurs, secret bureaucrats;"
  Chapter 13: "Regimental chests."
- Porter, T. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, N.J.: Princeton University Press.
  Chapter 2: "How Social Numbers Are Made Valid."
- Cole, S. A. (2002). *Suspect identities: A history of fingerprinting and criminal identification*. Cambridge, Mass.: Harvard University Press.
  Chapter 2: "Measuring the Criminal Body"

Additonal readings

- Alfred D. Chandler, Jr. *The Visible Hand — The Managerial Revolution in American Business*. Harvard University Press, 1977.
  Chapter: "The Railroads: The First Modern Business Enterprises, 1850s-1860s".

Lab: relational databases
http://dev.mysql.com/downloads/mysql/5.0.html#downloads

Week 3 (Tue. April 17): Statistical Concepts

The first four readings consider averages in some way, from the theory of means by Quetelet to composite photography by Galton. The next three readings consider variation (or variability).

Average:

- Alphonse Quetelet, "Letters addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha on the Theory of probabilities as applied to the moral and political sciences," London, p. 38-98.
- Galton, F. "Generic images", *Proceedings of the Royal Institution*, 1879, **9**(161-70).
- Galton, F. "Composite portraits", *Journal of the Anthropological Institute,* 1879, **8**(132-144) — Samples at http://galton.org/composite.htm

Variation:

- M. Turk, "A Random Walk through Eigenspace," *IEICE Transactions on Information and Systems*, Vol. E84-D, No. 12, December 2001, pp. 1586-1595.
- Cesar F. Caiafa, Araceli N. Proto, Daniel Vergani, Zulma Stanganelli (2005) " Development of individual recognition of female southern elephant seals, Mirounga leonina, from Punta Norte Península Valdés, applying principal components analysis" *Journal of Biogeography* **32**(7), 1257–1266.

Additonal readings

- Dahlia S. Cambers, "Normman and Norma: Looking for Mr. and Mrs. America," *Cabinet* **15**:69.
- Mary Coffey, "American Adonia: Eugenics, statistics, and the controversial paunch," *Cabinet* **15**:70-71.
- Mâns Wrange/Ombud, "Meet Marianne: The Average Citizen Project," *Cabinet* **15**:72.
- Lefevbre, Henri, *Rythmanalysis: Space, time, and everyday life,* 2004, Continuum.

Lab: Introduction to the R language.
http://www.r-project.org

Week 4 (Tue. April 24): Representational practices

Data acquires its meaning when put in relation with other data, through classification and categorization. A major philosophical assumption underlying much computer science practice is that the world can be faithfully represented on a computer, if enough data, of sufficient precision can be captured (Agre). Race (Campell-Kelly), disease, mental illness exist across a continuous spectrum (Bowker and Star). A major question regards the discontinuities across the spectum are "real" observable discontinuities, or whether they are pure social constructions (Haslam).

- Agre, P. (1997). "Beyond the Mirror World: Privacy and the Representational Practices of Computing" in *Technology and Privacy: The New Landscape.* P. Agre and M. Rotenberg. Cambridge, Mass., The MIT Press: 29–61.
- Bowker, G. C. & Star, S. L. (1999) *Sorting Things Out: Classification and Its Consequences.* MIT Press: Cambridge, MA.
  Chapter 2: "The Kindness of Strangers"
  Chapter 3: "Classification, Coding, and Coordination."
- Haslam, Nick, "Kinds of Kinds: A Conceptual Taxonomy of Psychiatric Categories", *Philosophy, Psychiatry, & Psychology* - Volume 9, Number 3, September 2002, pp. 203-217

Additonal readings

- Campbell-Kelly, M. (1990). "Punched-Card Machinery" in *Computing Before Computers.* (W. Aspray, ed.) Ames, Iowa University Press: 122-155.
- O'Sullivan, G. (2002). "The South African Truth and Reconciliation Commission: Database Representation", chapter 4 (including appendixes) *in Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*, Patrick Ball, Herbert F. Spirer, and Louise Spirer (eds.) American Association for the Advancement of Science, 2000.

Lab: Relationships in relational databases.

Week 5 (Tue. May 1): Text processing

While data is often implicitely understood as numbers, text itself can be usefully analysed through numerical methods.

- William Mitchell, *Placing Words: Symbols, Space and the City*. MIT Press, 2005, pp. 3-19.
- Italo Calvino, *If on a Winter's Night a Traveller*, Harcourt Brace & Company, pp. 168-198.
- Frederick Mosteller, David Wallace, *Inference and Disputed Authorship: The Federalist*, 1964, Addison-Wesley, Chapter 1.
- Bell, Cleary, Witten, *Text Compression*, Prentice Hall, 1990, chapters 1.

Additonal readings

- Frederick Mosteller, David Wallace, *Inference and Disputed Authorship: The Federalist*, 1964, Addison-Wesley, Chapter 2, 3.
- Bell, Cleary, Witten, *Text Compression*, Prentice Hall, 1990, chapters 2, 4.

Lab: regular expressions, natural language toolkit (http://nltk.sourceforge.net/).

Web sites:

- Trigger Happy: http://www.dot-store.com/pages/thap.html
- Listening Post: http://www.earstudio.com/projects/listeningPost.html
- Word Count: http://www.wordcount.org/main.php
- Word count: http://www.wordcount.org/querycount.php
- Axis: http://artport.whitney.org/commissions/codedoc/Levin/axis.html
- Secret lives of numbers: http://www.turbulence.org/Works/nums/
- Baby name wizard: http://babynamewizard.com/namevoyager/lnv0105.html
- Google trends: http://www.google.com/trends
- Google AdWords happening: http://www.iterature.com/adwords/
- Amazon text stats: http://www.amazon.com/gp/search-inside/text-readability-help.html/ref=sib_ab_help/102-4582744-1528156
- Text Arc: http://textarc.org/
- Nielsen Buzzmetrics Blogpulse: http://www.blogpulse.com/

- http://www.marumushi.com/apps/newsmap/newsmap.cfm
- http://www.cs.umd.edu/hcil/treemap/
- 2007 State of the Union Address: http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html
- http://www.research.ibm.com/visual/projects/history_flow/
- http://twittervision.com/
- http://whocalled.us/
- http://r-s-g.org/carnivore/
- http://www.wefeelfine.org/
- http://artport.whitney.org/commissions/thedumpster/

Week 6 (Tue. May 8): Machine learning and data mining

The first pair give a history of where data mining came from as a discipline; they are followed by a critique of some kdd practices by Gandy. We then read about the underlying reasoning behind machine learning and use a couple chapters from game design to talk about Bayesian networks and neaural networks

- Lisa Jevbratt , "The Prospect of the Sublime in Data Visualizations" *YLEM Journal*, number 8 volume 24 (July/August 2004), pp. 4-8.
- Padhraic Smyth, "Data mining: Data analysis on a grand scale", Technical Report, Dept. of Computer Science, UC Irvine.
- David Bourg and Glenn Seemann, *Artificial Intelligence for Game Developers,* O'Reilly. Chapters 12 and 14.

Additonal readings

- Harman, Kulkarni, *Reliable Reasoning: Induction and Statistical Learning Theory*, MIT Press, Chapter 1 and 2.

Lab: Python (www.python.org).

Week 7 (Tue. May 15): Data networks

Distributed networks as technology, social form, and identity.

- Forster, P. and King, J. L., (1995), "Information Infrastructure Standards in Heterogeneous Sectors: Lessons from the Worldwide Air Cargo Community", in Kahin, B., & Abbate, J. (eds). *Standards policy for information infrastructure*. Cambridge, Ma: MIT Press.
- Yates, J. (1989). "Communication technology and the growth of internal communication", in *Control through communication: The rise of system american management.* Baltimore: Johns Hopkins University Press.

- Galloway, Alexander, Chapter 1 and 2: "Physical Media" and "Form", in *Protocol,* MIT Prress, 2004.

Additonal readings

- Bamford, J. (2001). "Muscle." In *Body of secrets: Anatomy of the ultra-secret national security agency: From the cold war through the dawn of a new century* (pp. 406-451). New York: Doubleday.
- "Content Standard for Digital Geospatial Metadata", Metadata Ad Hoc Working Group, Federal Geographic Data Committee, 1998.
- Liu, Alan, Chapter 1: "The idea of knowledge work", in *The Laws of Cool*, University of Chicago Press,

Lab: Ggobi (www.ggobi.org).
Activity: group consultations with instructors.

Week 8 (Tue. May 22): Geospatial data and mapping

- Linda L. Hill, *Georeferencing — The Geographical Associtions of information*, MIT Press. Chapter 1, 2, 3.
- Nadine Schuurman, "Trouble in the heartland: GIS and its critics in the 1990s", *Progress in Human Geography* **24**,4 (2000) pp. 569–590.
- Sarah Elwood, "Critical Issues in Participatory GIS: Deconstructions, Reconstructions, and New Research Directions", *Transactions in GIS* , 2006, **10**(5): 693–708

Additonal readings

- Ian McHarg, *Design with Nature*, Wiley (1969). (Landscape architect, largely credited as the first one to use "layers", a critical part of GIS).
- David A. Crowder, *Google Earth for Dummies*, Wiley. Chapter 1, 2, 3.
- Sarah Elwood, "Beyond Cooptation or Resistance: Urban Spatial Politics, Community Organizations, and GIS-Based Spatial Narratives" *Annals of the Association of American Geographers*, **96**(2), 2006, pp. 323–341.

Web sites:

- http://worldprocessor.com/
- http://imapla.lacity.org/Viewer/GIS/Viewer.asp
- http://gmapsflighttracker.com/
- Community mapping: http://nkla.sppsr.ucla.edu/
- KML: http://code.google.com/apis/kml/documentation/
- Laura Kurgan Monochrome landscapes: http://www.l00k.org/monochromes_proj/
- GPS drawing, "One year in London": http://www.gpsdrawing.com/info/catalogue.htm
- GPS drawing, "Vegas Dollar": http://www.gpsdrawing.com/gallery/land/usa/nv/vegas_dollar.htm
- Bio Mapping: http://biomapping.net/index.htm
- City in a Soundwalk: http://www.treetheater.org/nysoundwalk/
- Place Blogger: http://www.placeblogger.com/

Activity:
Form groups consisting of one existing project representative per group. (so someone from project 1, someone from project 2, etc). this should work out just about evenly although there might be a group that is short of a project or a group with two people on one project. anyway, it should work out about even. Each person takes turns describing what their project is about, the overall goals, the overall plan and their contribution. in short, they need to be able to articulate the full project, have connection with the full project... this will add pressure for each discipline to know what the others are doing... so dma can't ignore the backend processing and stat can't ignore a dma user interface.

Lab: consultations with instructors.

Week 9 (Tue. May 29): Data visualization

Guest speaker: Aaron Koblin (http://aaronkoblin.com/)

- Andrea Polli, "Atmospherics/Weather Works: Artistic Sonification of Meteorological Data", *YLEM Journal*, number 8 volume 24 (July/August 2004), pp. 9-13.

Week 10 (Tue. June 5): Data curation

- Depocas, Alain, Jon Ippolito, and Caitiln Jones. *Permanence Through Change: The Variable Media Approach*. New York: Guggenheim Museum, 2003.
- Karasti, Helena & Karen Baker. "Infrastructuring for the Long-Term: Ecological Information Management." Hawaii International Conference on System Sciences 2004 (HICSS'37), Hawaii, January 5-8 2004.
- Lyman, Peter. "Archiving the World Wide Web" in *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*. Washington, DC: Council on Library and Information Resources, 2002, pp. 38-51

Additonal readings

- Rothenberg, Jeff (2000), "Preserving Authentic Digital Information", in *Authenticity in a Digital Environment*, Washington, DC: The Council on Library and Information Resources.
- Reference Model for an Open Archival Information System (OAIS) – CCSDS.

Activity:
We would like to use the last class on data curation as an opportunity to reflect on the problem of preserving the group projects. I will assign some readings, but would like to discuss these issues as they apply concretely to your projects. For this purpose, I would be grateful if you could provide me with a list of all of the technologies (hardware and software), as well as data standards involved in your project. For example:

For example:
HTML for output,
Java applet for output,

Special algorithm to analyze guitar solos,
Python for real-time processing of the data,
HTTP standard (processing of URLs)

Unknown data standard used by Match.com to categorize data profiles.
Unknown data standard used by FutureMe.com to organize emails.
60" plasma screen for presentation.

If there is anything that is not generic, specify how so.

Write this into a draft document that describes all the technical components of your projects, and the steps you took/are taking to process the data. This will be a useful document at all levels.

Week 11 (Tue. June 12): Final Presentations

Presentation in class.