Computational Social Science Syllabus

Instructor

- Jacob Eisenstein, College of Computing

Goal

This graduate seminar focuses on text and network analysis of data with applications to domains such as political science, sociolinguistics, sociology, and public health.

Readings

Readings will be drawn from survey papers, course notes, and the textbook "Quantitative Social Science: An Introduction" by Kosuke Imai.

Grading

- **Three problem sets, worth 15% each** = 45%. These are intended to build and assess your understanding of the core technical concepts in the class.
- **A final project, worth a total of 25%**. This is intended to assess your ability to apply the technical concepts in the class to a challenging problem.
- **Weekly reading responses**, of which you must do ten, **worth 2% each = 20%**. These are intended to facilitate and assess understanding of the reading material. Reading responses must be posted to the private Wordpress site by 1pm on the wednesday of the class.
- **Class participation, worth 10%**. This is intended to motivate an engaging classroom discussion.

Prerequisites

Students should have strong programming skills, and background in probability and statistics.

Learning objectives and educational outcomes

As part of this course, students will learn to:

- understand and discuss basic social science questions and quantitative research methodology;
- design and implement computational techniques for analyzing socially-relevant data and metadata, especially text and social networks;
- select and apply these techniques to novel datasets to answer socially interesting questions.

Academic integrity

Students are encouraged to discuss the problem sets and readings outside of class. However, everyone must submit their own work, and you may not share code or answers. If your discussion with another student helped you make a breakthrough on a difficult problem, that is fine, but give credit!

The final project may be performed in teams of at most three students.

Suspected cases of honor code violations will be handled through the Office of Student Integrity. If you have a question about collaboration policy, please ask.

Learning accommodations

If needed, we will make classroom accommodations for students with documented disabilities. These accommodations must be arranged in advance and in accordance with the Office of Disability Services (http://disabilityservices.gatech.edu).

Excused absences policy

http://www.catalog.gatech.edu/rules/4/

Outline of topics

The schedule that follows is from the 2015 edition of the course CS8803-CSS. There are a few different categories of reading:

- **Reading**: This is required.

- **Supplemental reading**: Optional. These are usually research papers that relate to the topic of the class, and may be discussed in class. Look here for final project ideas.
- **Catch-up reading**: Optional. This is tutorial/survey material to help students catch up with the technical concepts in the class.
- **Also see**: Optional, non-peer-reviewed material. This includes blogposts and popular press articles.

Part 1: Data science

In this section, we will use contemporary research papers to get up to speed on statistical methods. If you have trouble keeping up with these readings, you may consult Think Stats, a free online textbook.

1/5: Computational social science

- **Reading**: Computational social science by Lazer et al; Six provocations for Big Data by boyd and Crawford.
- **Supplemental reading**: No silver bullet: De-identification still doesn't work by Narayanan and Felten; Computational Social Science: Toward a Collaborative Future by Wallach, 2015.
- **Also see**: The view from the other side: perspectives on computational sociology from Fabio Rojas. Good comments, too; Sticky data: Why even 'anonymized' information can still identify you.

1/7: Provocations; counting random events

We will cover basic probability and statistics, including random events, probability mass functions, cumulative distribution functions, and hypothesis testing.

- **Reading**: Six provocations for Big Data by boyd and Crawford; Think Stats chapters 1-3.

Here is a stats refresher iPython notebook, focusing on tests of statistical significance. You may find this useful to look at before doing problem set 1.

1/12: Multiple comparisons; randomized tests; correlation

Further discussion of hypothesis testing, with emphasis on corrections for multiple comparisons, randomized tests, and measures of correlation.

- **Reading**: Censorship and deletion practices in Chinese social media, by Bamman, O'Connor, and Smith, sections 1-4;
- **Catch-up reading**: Think Stats, chapters 5 and 7.

- **Also see**: Ages and names; XKCD on multiple comparisons; Spurious correlations

1/14: Regression

- **Reading**: *Wasserman chapter on linear regression* (available on T-square, ask me if you can't access it); More tweets, more votes by Joseph DiGrazia et al, 2013; How (not) to predict elections by Metaxas et al, 2011 (skim).
- **Supplemental reading**: The cost of racial animus on a Black presidential candidate: Using Google search data to find what surveys miss by Stephens-Davidowitz, 2013; Understanding the political representativeness of Twitter users; Online and social media data as a flawed continuous panel survey by Diaz et al 2014.
- **Catch-up reading**: Think stats, chapters 4 and 8; Notes on regression by Andrew Ng, sections 0, 3, and 5.
- **Also see**: this Vox piece by Ezra Klein on the use of controls in studies of racial and gender discrimination.
- **Problem set 1** out. Due January 23 at 5pm.

1/19: No class

Celebrate Martin Luther King Day

1/21: Classification and clustering

- **Reading**: A few useful things to know about machine learning by Domingos, 2012. A computational approach to politeness by Cristian Danescu-Nicolescu-Mizil et al, 2013. Data carpentry (very short!)
- **Problem set 1** due on January 23 at 5pm.

Part 2: Network Analysis

In this section of the course, monday readings will be drawn from the textbook Networks, Crowds, and Markets by Easley and Kleinberg (abbreviated E&K). Free PDFs of each chapter are available by following the link.

1/26 and 1/28: Networks

- **Monday reading**: E&K chapters 1 and 2
- **Wednesday reading**: Structural diversity in social contagion
- **Supplemental reading**: Chapter 6 of Newman is more rigorous and more detailed.

2/2 and 2/4: Strong and weak ties

- **Monday reading**: E&K chapter 3
- **Wednesday reading**: The Role of Social Networks in Information Diffusion by Bakshy et al, 2012;

2/9 and 2/11: Homophily

- **Monday reading**: E&K chapter 4
- **Wednesday reading**: Inferring social ties from geographic coincidences by Crandall et al, 2010; Find me if you can by Backstrom et al
- **Supplemental reading**: Birds of a Feather: Homophily in Social Networks by McPherson et al, 2001; Homophily and Contagion Are Generically Confounded in Observational Social Network Studies by Shalizi and Thomas, 2011;

2/16: No class

- **Problem set 2** out. Due March 1 at 5pm.

2/18: Signed social networks

- **Reading**: E&K chapter 5
- **Supplemental reading**: The Slashdot Zoo: Mining a social network with negative edges by Kunegis et al, 2009.

2/23: Signed social networks in social media and literature

- **Reading**: Signed networks in social media by Leskovec et al, 2010.
- **Supplemental reading**: "You're Mr. Lebowski, I'm the Dude": Inducing Address Term Formality in Signed Social Networks by Krishnan and Eisenstein, 2014; Extracting Signed Social Networks From Text by Hassan, Abu-Jbara, and Radev, 2012. Exploiting social network structure for person-to-person sentiment analysis by West et al, 2014.

3/2: Statistical models of networks: ERGMs

- **Reading**: A survey of statistical network models, sections 2-3.8.
- **Problem set 2** due on March 1 at 5pm.
- **Supplemental reading**: Inferential Network Analysis with Exponential Random Graph Models by Cranmer and Desmarais, 2011.

3/4: Statistical models of networks: Stochastic Blockmodels

- **Reading**: A survey of statistical network models, sections 2-3.8.
- **Supplemental reading**: A Multiscale Community Blockmodel for Network Exploration by Ho, Parikh, and Xing, 2012; Document Hierarchies from Text and Links by Ho, Eisenstein, and Xing, 2012.

Part 3: Text

3/9: Word counting

- **Reading** The psychological meaning of words: LIWC and computerized text analysis methods by Tausczik and Pennebaker (2010); Text as data sections 2-4
- **Supplemental reading**: Linguistic Models for Analyzing and Detecting Biased Language by Recasens et al; Shedding (a thousand points of) light on biased language by Yano et al; Detecting and modeling local text reuse.
- **Also see**: Fairness versus freedom

3/11: Text classification and regression

Reading: Narrative framing of consumer sentiment in online restaurant reviews - **Supplemental reading**: Phrases that signal workplace hierarchy by Gilbert; Political ideology detection using recursive neural networks by Iyyer et al; More than Words: Syntactic Packaging and Implicit Sentiment by Greene and Resnick. - **Problem set 3** out on Thursday 3/12, due 3/29.

3/16 and 3/18: No class, spring break

3/23 and 3/25: Statistical models of text

- **Monday reading**: Fighting words by Monroe, Colaresi, and Quinn. iPython notebook.
- **Wednesday reading**: Data Analysis with Latent Variable Models by Blei, 2014. Read sections 1-3.3, then read Probabilistic topic models by Blei, 2012.
- **Supplemental reading**: The rest of Data Analysis with Latent Variable Models, especially if you want to know how these things really work; Identifying regional dialects in online social media by Eisenstein (2015); Sparse additive generative models of text by Eisenstein et al (2011); Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach by Schwartz et al (2014); Care and feeding of topic models by Boyd-Graber, Mimno, and Newman (2014).

- **Problem set 3** due 3/29 at 11:59pm.

3/30 and 4/1: Topic models and metadata

- **Monday reading**: Probabilistic topic models by Steyvers and Griffiths
- **Wednesday reading**: Topic and Role Discovery in Social Networks...
- **Supplemental reading**: Learning to extract international relations from political context by O'Connor et al, 2013; Hierarchical relational models for document networks by Chang and Blei, 2009; A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases by Grimmer, 2009; Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression by Taddy, 2013.
- Final project proposal due Friday 4/3 at 11:55pm.

4/6 and 4/8: Language in a social context

- **Monday reading**: Predicting crime using Twitter and Kernel Density Estimation by Gerber, 2013.
- **Wednesday reading**: Echoes of power: Language effects and power differences in social interaction by Danescu-Niculescu-Mizil, Lee, Pang, Kleinberg, 2012
- **Supplemental reading**: The Bayesian Echo Chamber: Modeling Social Influence via Linguistic Accommodation by Guo et al, 2015.

Part 4: Final projects

4/13 and 4/15: Final project check-ins

I will use a Google doc to schedule meetings with each group.

4/20: Final project presentations

4/22: Contemporary topics in computational social science

4/24: Final project writeups due at 11:55pm