Big Data: Revolution and Reality
Byrne Seminar 01:090:101, section 73, index 18175
Wednesdays, 10:20-11:40 am, Library of Science and Medicine Conference Room

Javier Cabrera  Ryan Womack
Prof. of Statistics  Data Librarian
cabrera@stat.rutgers.edu  rwomack@rutgers.edu
Hill Center  Alexander Library reference offices
Office Hours: Wed 4:30-6:30 pm  Office Hours: before or after class, or by appointment

The topics and readings on this syllabus are provisional, and could change during the course. The date of the class excursion is also provisional, and could change. This syllabus is the FINAL (March 23) version.

Course Description: Talk of Big Data is ubiquitous, but what does this mean in practice? This seminar explores the impact that large scale data collection and analytics are having in academia and business. Web data collection is transforming marketing and economic production. Massive genomic databases are transforming medical research. And text mining is transforming the study of humanities. Meanwhile, merged databases of administrative records increase the potential for both greater social understanding and reduced privacy. New technological tools and approaches are required to handle massive data arrays in physics and astronomy. In business, large databases collect information in real-time and are mined for instant decision-making, such as credit card fraud detection, requiring speed and accuracy. What unique challenges in statistical methodology and computing does Big Data bring? What are the tools of this new trade and what are the traps and tricks of Big Data analytics? What kinds of jobs and careers are being created in Big Data fields, and what skills and degrees do they require?

Course Objectives: Readings and class discussions will explore the implications of Big Data in various areas, and student presentations will allow each student to explore a topic of interest in more detail. Students will emerge with an appreciation for the realities and potential of Big Data to transform our collective future.

Prerequisite(s): None.
Credit Hours: 1
Text: None required. Links to readings will be provided in the course Sakai site.

Grading: This is a Pass/No Credit course. Participation in class and completion of the final presentation are required to pass the course. Students must attend and participate in a minimum of 8 of the 10 class sessions. One or two articles will be assigned as readings for each session. These readings will form the basis of the discussion and are required. Students who do not participate orally during class will be asked to contribute written comments on the Sakai site. The final presentation is described on the next page.
**Final Presentation:** Teams of up to 4 students each will be formed to investigate topics relevant to big data of the team’s own choosing. Teams should submit their proposed topic by the 5th class on February 22. During the final class on March 30, each team will give a brief 10 minute presentation. Presentations should show some independent research and investigation of the topic, and discuss the current issues, developments and controversies relating to the topic. Each team member should have a clearly defined contribution to the presentation. The presentation is an opportunity for the students to develop their own thinking, but there is no need to create elaborate powerpoint presentations or conduct exhaustive surveys of findings.

**Academic Integrity:** Students are expected to adhere to the Rutgers Academic Integrity Policy (see [http://academicintegrity.rutgers.edu](http://academicintegrity.rutgers.edu)). In particular, this requires proper citation and acknowledgment of others’ work, recognition of others’ contributions, ethical treatment of other students, and ensuring that all work submitted as one’s own is “produced without the aid of impermissible materials or impermissible collaboration.”

**Attendance:** Attendance is required as described under Grading, and will be taken each class. It is the absentee’s responsibility to gather any information or materials missed.

**Useful Links:**

- Hands-on Big Data presentation and scripts: [https://github.com/ryandata/bigdata](https://github.com/ryandata/bigdata)
- History of Technology at US Census: [https://www.census.gov/history/www/innovations/technology/](https://www.census.gov/history/www/innovations/technology/)
- Netflix Prize: from Netflix, [Wikipedia](https://en.wikipedia.org/wiki/Netflix_Prank), and Forbes
- How much data does Google store: [article xkcd](https://xkcd.com/1445/)
- Video Tours of Data Centers: Facebook, Google
- NYC Taxis and Uber: [R analysis](https://rpubs.com/jeffleek/nyc_taxis)
- Google ngrams
- [Raj Chetty (economist)](https://rajchetty.org) and Census Research Data Centers
- Big Data in Humanities: [Literary History](https://wwwсходное.com/literaryhistory), [Historical Analysis](https://wwwсходное.com/historicalanalysis), and [Art](https://wwwсходное.com/art)
- Genomics: [NCBI](https://www.ncbi.nlm.nih.gov), [Bioconductor](https://bioconductor.org), and [Genecards](https://www.genecards.org)
- [Brain Initiative (NIH)](https://www.braininitiative.nih.gov)
- Astronomy: [Sloan Digital Sky Survey](https://www.sdss.org), [GalaxyZoo](https://www.galaxyzoo.org), and software mentioned in Feb. 17 articles
**Tentative Course Outline:**

<table>
<thead>
<tr>
<th>Week</th>
<th>Content</th>
</tr>
</thead>
</table>
| Jan 20 | **Introductions** (to Big Data, the course, and each other)  
No Reading |
| Jan 27 | **Big Data in Business**  
“To Facebook, you’re worth $80.95” [http://blogs.wsj.com/cio/2012/05/03/to-facebook-youre-worth-80-95/](http://blogs.wsj.com/cio/2012/05/03/to-facebook-youre-worth-80-95/)  
| Feb 3 | **Big Data in the Social Sciences and Humanities**  
Einav L and Levin J. “Economics in the age of big data” Science 7 November 2014: 1243089 |
| Feb 10 | **Big Data in the Life Sciences**  
Marx V. “Biology: The big challenges of big data” Nature 498, 255-260 (13 June 2013) and additional readings on Sakai |
| Feb 17 | **Big Data in the Physical Sciences + Computing for Big Data**  
Zhang Y and Zhao Y. “Astronomy in the Big Data Era” Data Science Journal, vol 14, no. 11, 2015  
| Feb 24 | **Using R and other software tools for big data**  
David Donoho, “50 Years of Data Science” and John Tukey “The Future of Data Analysis”  
Learning R and Intro to Big Data |
| March 2 | **Site Visit - RDI2 (Rutgers Discovery Informatics Institute)**  
**Meet in CoRE 626.** No Reading |
| March 9 | **Methods and Challenges in working with Big Data**  
| March 23 | **Big Data Skills and Careers/Future of Big Data**  
‘Data Science Job Skills’  
‘Big Data Skills’ |
| March 30 | **Final Presentations and Wrapup**  
No Reading |