BIG DATA FOR FEDERAL AGENCIES

SURV 699Y

CLASS STRUCTURE AND COURSE CONCEPT

The amount of digital data generated as a by-product in society is growing fast, e.g., data from satellites, sensors, transactions, administrative processes, social media and smartphones. This type of data is characterized by high volume, high velocity, high variety and is often called big data. The hope is to gain insights from this data for different areas such as e.g., health and crime prevention, planning of infrastructures, and business decisions. Big Data is of interest for agencies that produce statistics to find alternative data sources either to reduce cost, to improve estimates or to produce estimates in a more timely fashion. In particular on the economic statistics side, this interest in growing rapidly. The change in the nature of the new types of data, their availability, the way in which they are collected, and disseminated are fundamental. The change constitutes a paradigm shift for agencies that in the past relied primarily on survey research. However, data quality frameworks well established in statistics production still hold. Thus the goal of this lecture is to equip the next generation of social and economic scientists as well as survey researchers with the right tools to face these data. The lecture uses specially curated data sets and a working example that runs through the entire course. The lecture is paired with a hands-on lab session (SURV699U), in which students apply all learned techniques through a worked example relevant to core work of the Federal Statistical Agencies.

GRADING METHOD (regular, letter). Course will fulfill the requirement of an elective in the JPSM Master and Certificates Programs.

PREREQUISITES (familiarity with survey and administrative data; mastery of a statistical software package; completion of an online Python programming class or pre-class bootcamp)

LEARNING OUTCOME

Learn how to think about data analysis to solve social problems using and combining large quantities of heterogeneous data from a variety of different sources. Learn how to evaluate which data are appropriate to a given research question and statistical need. Learn the
different data quality frameworks and learn how to apply them. Learn the basic computational skills required for data analytics (for text-mining, large-scale data integration and visualization), typically not taught in social science, economics, statistics or survey courses. Learn how to apply statistical and data quality frameworks to big data problems. Identify new approaches to creating and displaying information for Federal Agencies.

**WEEKLY STRUCTURE**

- **Friday 9am – noon**: new material will be introduced (see detailed schedule below)
- **Friday 1pm – 3pm**: Lab session on new material
- **Wednesday 9am - noon**: Lab session on homework problems and projects
- **Ongoing**: video lectures on python, machine learning

**Location (Friday Session)**

Summit LL, 601 New Jersey Ave NW
- with Ethernet connection
- Video connection capabilities
- Recording capabilities
- Data on secure server at NYU/CUSP

**Wednesday Sessions on Data Projects via Adobe Connect**

**GRADING**

- Weekly online quizzes (worth 10% total)
- Class participation (10% of grade)
- Course project and homework (60%) in one of two options
  - Option a: Heavier emphasis on projects. Students self-select into groups or choose to do individual projects. Baseline homework expected to be done every week
  - Option b: Heavier emphasis on homework. Baseline homework expected to be done every week; additional, more in depth, problems assigned tied to addressing agency problems

- Class presentations (20% of grade)

Dates of when assignment will be due are indicated in the syllabus. Late assignments will not be accepted without prior arrangement with the instructors.

**READINGS**

TBD

**ACADEMIC CONDUCT**

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at the University of Maryland Graduate School web site
Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

**CLASS SCHEDULE**

**9/4 1. INTRODUCTION AND MOTIVATION (JULIA LANE AND FRAUKE KREUTER)**  
The goal of this session is to introduce students to the overall approach, to the data, what will be learned

1. The social science of measurement  
2. “Big Data”: Definitions, technical issues  
3. Quality Frameworks and varying needs (relevance, timeliness, accuracy, accessibility, clarity, comparability, coherence)  
4. Introduction to big data and overview of existing analysis  
5. Introduction to the data that will be used in the class

**9/11 2. DATABASE BASICS (JULIANA FRIERE AND JULIA LANE)**  
The goal of this session is for students to be able to read, write, and build different types of databases using heterogeneous sources of data.

1. Database concepts  
2. Database taxonomies  
3. ETL in different databases  
4. Data hygiene: curation and documentation

**9/18 3. VISUALIZATION (CATHERINE PLAISANT)**  
The goal of this session is to provide an overview of different approaches to conveying information in an intuitive and statistically valid manner. The chapter will focus on the uses of visualization for analytical purposes as well as for communication purposes.

1. Overview of visualization approaches  
2. Using graphics packages for data visualization, including network geolocation and GIS software to display shape files

**9/25 4. GROUP PRESENTATIONS**

**10/2 5. UNDERSTANDING THE USES OF SOCIAL MEDIA AND USING APIs (CAMERON NEYLON)**  
The goal of this session is to examine how to make use of social media APIs. The particular application is capturing information about the transmission of knowledge and ideas
1. Introduction to APIs
2. Use the PLOS Search API to obtain a list of DOIs for relevant papers.
3. Use clustering algorithms to identify subsets of articles that appear to have differing “impact signals”
4. Identify followers
5. Application of visualization techniques
6. Discussion of coverage and qualitative enrichment of quantitative scores

10/9 6. Programming with Big Data (Claudio Silva)
This session shows how to combine large datasets and optimize programming. It will provide an introduction to NOSQL, MapReduce and Hadoop

1. Introduction to characteristics of large databases
2. Building datasets to be linked
3. Create a big data sample work flow
4. Example of record linkage with MapReduce

10/16 7. Networks (Jason Owen Smith)
The goal of this session is to develop new units of analysis (networks) and show how to measure different types of networks

1. Directed and undirected graphs
2. Relational analysis on graphs
3. Changes in structure, size and orientation of networks as a result of federal grants

10/23 8. Data Linkage (Stefan Bender and Josh Tokle)
The goal of the session is to go through the basics of data linkage across heterogeneous datasets

1. Linkage in the context of big data
2. Social science approaches
3. Computer science approaches
4. Strengths and weaknesses

10/30 9. Machine Learning (Rayid Ghani and Malte Schierholm)
The goal of this session is to show students how to apply machine learning ideas, processes and methods and how to build a conceptual framework within the context of the goals and data available for the class.

1. What is machine learning
2. Examples, process and methods
3. Advanced topics

11/6 10. Text Analysis (Jordan Boyd Graber and Evgeny Klochikhin)
The goal of this session is to introduce students to the potential for using text analysis to identify topics and themes.

1. Value of text data
2. Different text analytics paradigms
3. Discovering topics and themes in large quantities of text data

11/13 GROUP PRESENTATIONS

11/20 11. NON-RANDOM SAMPLES AND STATISTICAL INFERENCE (PAUL BIEMER)
This session will provide an overview of using data from a non-randomly selected sample for inference.

1. Review of total survey error (the TSE paradigm); general discussion on effects of inference
2. Discussion of error sources specific to found (big) data
3. Examples of big data analysis and erroneous inferences caused by ignoring data errors (Google trends, etc)
4. Big Data and Errors
   a. Issues for inference assuming perfect data (incidental correlation, incidental endogeneity, noise accumulation)
   b. Issues for inference assuming fallible data (classification error effects on subgroup identification and cluster analysis, effects of uncorrelated and correlated errors on correlations, regressions, data mining etc., effects of missing data and sample selectivity)
5. Overview of methods to correct for data errors
   a. Super-scaled data scrubbing and editing
   b. Adjust the data via a cleaned reference sample
   c. Modeling approaches
      (e.g. hierarchical modeling for correlated errors)
   d. Weighting to a reference set to address selectivity
   e. Other approaches
      (mass imputation, likelihood and Bayesian methods, etc.)
6. Open Issues in the analysis of found data

11/27 12. THANKSGIVING

12/4 13. PRIVACY AND CONFIDENTIALITY (FRAUKE KREUTER, JULIA LANE AND STEFAN BENDER)
This session will provide an overview of the privacy and confidentiality issues associated with big data

1. Legal Framework
2. Statistical Framework
3. Practical approaches

12/11 14. GROUP PRESENTATIONS