SYLLABUS - APPLIED DATA ANALYTICS FOR PUBLIC POLICY

Julia Lane and Daniela Hochfellner

COURSE DESCRIPTION AND OBJECTIVES
The goal of the Applied Data Analytics class is to develop the key data analytics skill sets necessary to harness the wealth of newly-available data. Its design offers hands-on training in the context of real microdata. The main learning objectives are to apply new techniques to analyze social problems using and combining large quantities of heterogeneous data from a variety of different sources. It is designed for graduate students who are seeking a stronger foundation in data analytics.

Objectives:
- Evaluate which data are appropriate to a given research question and statistical need.
- Identify the different data quality frameworks and apply them to public policy problems.
- Learn a broad array of basic computational skills required for data analytics, typically not taught in social science, economics, statistics or survey courses.

The curriculum is structured around four key components:
- **Foundations**: The social science of measurement, Formulating research questions, Basics of program evaluation, Differentiating data sources, "Big Data" - definitions, technical issues, Quality frameworks and varying needs, Introduction to the data that will be used in this class, Case studies, Introduction to Python, Working with Jupyter Notebooks, Web scraping exercises, Exploring data visually.
- **Data Curation**: Introduction to APIs, Database concepts, Database taxonomies, Introduction to characteristics of large databases, Building a data schema, ETL in different databases, Building datasets to be linked, Linkage in the context of big data, Create a big data work flow, Data hygiene: curation and documentation.
- **Data Analysis**: What is machine learning, Examples, process and methods, Fundamentals of network analysis, Directed and undirected graphs, Relational analysis on graphs, Value of text data, Different text analytics paradigms, Discovering topics and themes in large quantities of text data, The importance of geographic information, Basics in spatial data analysis, Mapping your data.
- **Presentation, Inference, and Ethics**: Using graphics packages for data visualization, Error sources specific to found (big) data, Examples of big data analysis and erroneous inferences, Inference in the big data context, Methods to correct for data errors, Big data and privacy, Legal framework, Statistical framework, Disclosure control techniques, Ethical issues, Practical approaches

TEXTBOOK:
Big Data and Social Science: A practical guide to models and tools, Taylor Francis 2016, Ian Foster, Rayid Ghani, Ron Jarmin, Frauke Kreuter and Julia Lane

REQUIREMENTS & PREPARATION

Programming skills
Python: basic knowledge (Intro to Python for Data Science by Data Camp (https://www.datacamp.com/courses/intro-to-python-for-data-science)

COURSE STRUCTURE
The course will be structured in bi-weekly sessions, whereas each session is combined with voluntary lab time. The sessions will consist of lectures and computing exercises, the voluntary lab will give you time to work on your assignments, ask questions, or discuss specific interests or problem sets in more detail with the instructors.
<table>
<thead>
<tr>
<th>Session</th>
<th>Date</th>
<th>Mandatory lecture and exercises</th>
<th>Voluntary lab time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>January 26th, 2017</td>
<td>9am - 12:30pm</td>
<td>1:30-3:30pm</td>
</tr>
<tr>
<td>Session 2</td>
<td>February 9th, 2017</td>
<td>9am - 12:30pm</td>
<td>1:30-3:30pm</td>
</tr>
<tr>
<td>Session 3</td>
<td>February 23rd, 2017</td>
<td>9am - 12:30pm</td>
<td>1:30-3:30pm</td>
</tr>
<tr>
<td>Session 4</td>
<td>March 9th, 2017</td>
<td>9am - 12:30pm</td>
<td>1:30-3:30pm</td>
</tr>
<tr>
<td>Session 5</td>
<td>March, 23rd, 2017</td>
<td>9am - 12:30pm</td>
<td>1:30-3:30pm</td>
</tr>
<tr>
<td>Session 6</td>
<td>April 6th, 2017</td>
<td>9am - 12:30pm</td>
<td>1:30-3:30pm</td>
</tr>
<tr>
<td>Session 7</td>
<td>April 20th, 2017</td>
<td>9am - 12:30pm</td>
<td>1:30-3:30pm</td>
</tr>
</tbody>
</table>

The time in between classes should be used to work on your group research project.

**SESSION 1: INTRODUCTION TO PROGRAM, DATA AND PROJECTS**

- Tutorial on how to define and scope a research project
  - Example study: Worker Advancement in the Low-Wage Labor Market: The Importance of “Good Jobs” by Fredrik Andersson, Harry J. Holzer and Julia I. Lane: [link](#)
- Introduction to data being used in class
- Overview of the computing environment and project space
  - Basics of using the command line in Linux
- How to work collaboratively in computing environments: Introduction to Git
  - Overview of Git: What is it and how does it work?
  - Getting to know the required git commands to successfully manage a collaborative project

Readings:
- Chapter 1 of textbook
- Worker Advancement in the Low-Wage Labor Market: The Importance of “Good Jobs” by Fredrik Andersson, Harry J. Holzer and Julia I. Lane: [link](#)
- Linux/Unix common terminal commands: [link](#)
- Git: [link to 1-pager](#)

**SESSION 2: DATABASES, SQL, AND PYTHON FOR DATA ANALYTICS**

- Database management and database clients
  - Why using databases?
  - What are databases: types, pro/cons, usage characteristics
- Introduction in SQL
  - Become familiar with the basic syntax, structure, and uses of SQL
  - Writing and running SQL queries, learn descriptive SQL queries
- Python/Pandas basics: Python basics needed for all data analyses done in this class
  - What is Python and Jupyter?
  - Learn to code: variables, data structures – lists and maps, logic – if then else and loops, functions – calling and writing

Readings:
• Chapter 4 of textbook
• Wes McKinney, Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, 2012, pp. 466
• SQL: [link]
• Python for Economists: [http://scholar.harvard.edu/ambell/python-economists](http://scholar.harvard.edu/ambell/python-economists)

More Resources for Python/Pandas (not required as readings):
• Introduction to Python for Econometrics, Statistics and Data Analysis by Kevin Sheppard (free): [link]
• Python: 1-pager from DataCamp & longer version of general Python notes
• Pandas: [link]
• Software Carpentry: [https://swcarpentry.github.io/python-novice-inflammation/](https://swcarpentry.github.io/python-novice-inflammation/)
• Python Tutorial: [https://www.python.org/about/gettingstarted/](https://www.python.org/about/gettingstarted/)

---

SESSION 3: WEB-SCRAPING, APIs AND RECORD LINKAGE

- Overview of two general ways one can retrieve data from data sources on the Internet: API and web scraping.
  - The goal is to become familiar with different types of APIs (GET- and POST- based HTTP APIs), different formats of requests, and how to learn a given API
- Learn the tools used to interact with network based APIs: Understand and use the tools for talking directly with APIs over HTTP connection, introduce libraries that abstract the details of the API and present a simplified programmatic interface
  - Making raw HTTP API requests, Using pre-packaged API client libraries, practical considerations
- Theory and Principles of record linkage
- Pre-processing needed before linking records: How to parse string fields, Introduction into regex

Readings:
Chapter 2 and 3 of textbook
Ryan Mitchell, Web Scraping with Python, O'Reilly Media, 2015
Hernández MA, Stolfo SS 1998, Real-world data is dirty: data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery 2(1), 9-73

More Resources  (not required as readings):
• Ivan P. Fellegi and Alan B. Sunter, A Theory for RecordLinkage, Journal Of The American Statistical Association Vol. 64, Iss. 328,1969
• Record linkage by Herzog, Scheuren and Winkler: [link]
• Winkler WE 2009. Record linkage. D Pfeffermann and CR Rao (Hg.) Handbook of Statistics 29A, Sample Surveys: Design, Methods and Applications Amsterdam: Elsevier
• Python's requests & Beautiful Soup libraries (for web scraping & APIs): [link]
• Regex: [link to PDF](#)
• Python regular expressions: [https://docs.python.org/2/library/re.html](https://docs.python.org/2/library/re.html)
• Online regular expression tester: [https://regex101.com/](https://regex101.com/)

---

SESSION 4: MACHINE LEARNING
Formulation research questions in a machine learning framework: from transformation of raw data to feeding them into a model

How to build, evaluate, compare, and select models

How to reasonably and accurately interpret models

Address biases in machine learning techniques and their consequences for public policy, for example how race biases can lead to unfair treatment of ethnic minorities in public policy.

Readings:
- Chapter 6, textbook

SESSION 5: NETWORK ANALYSIS AND TEXT ANALYSIS

Introduction into network analysis: What is network analysis? Representation of networks, network measures, centrality metrics, cliques, community detection
  ○ Strategies for detecting potential network data in relational data sets

Introduction in text analysis: Information retrieval, clustering and text categorization, text summarization, machine translation
  ○ How to transform a corpus of text into a matrix on which NLP can be applied
  ○ Learn how to implement topic modeling
  ○ Document tagging and evaluation of document tagging

Readings:
- Chapter 7 and Chapter 8, textbook
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. O'Reilly, 2009

SESSION 6: INFORMATION VISUALIZATION AND INFERENCE/ERRORS

Theory of information visualization
  ○ Communication tool
  ○ Choosing a chart type
  ○ Labeling and information overload
  ○ Color consideration

Visualizing analytical results

How to deal with inference and the errors associated with big data

Problems of Big data and the errors resulting from it

The total error paradigm: Traditional models and their implication for big data research

Readings
- Chapter 9 and 10, textbook

SESSION 7: PRIVACY, CONFIDENTIALITY, AND ETHICS

• Recognize where and understand why ethical issues can arise when applying analytics to policy problems starting with collection and moving through the management, sharing, and analysis of data
• Plan, execute, and evaluate a research project along privacy concerns and ethical obligations
• Key technical, ethical, policy, and legal terms and concepts that are relevant to a normative assessment of novel analytic techniques and tools for mitigating or managing the ethical concerns.

Readings:
• Chapter 11, textbook

EVALUATION

The evaluation is based on individual assignments (30% of the grade), group projects with progress reports produced every three weeks (30%), final project (20%) and contribution to code and metadata documentation and participation (20%).

The goal of group projects is to demonstrate the ability to develop a joint research project over time according to academic principles. This includes formulating a research question, writing (extended) abstracts, documenting research progress, and presenting results. At the beginning of the course, students will be assigned/volunteer to join a research group. The group work will constitute 50% of the final grade. For group projects, all members will receive the same grade. However, if it is apparent that a given member of a group has contributed much more or much less, that student’s grade will go up or down accordingly.

• Group progress report 1 (10%): Each group will be required to write a two-page research memo. Research memos should outline the research question and agenda of the group project, short description of data and methods used, and expected outcomes and applications for public policy. This report is due before the beginning of Session 3 and should be submitted through NYU classes.
• Group progress report 2 (10%): Each group will be required to write a two-page research progress report to document the group work. This report is due before the beginning of Session 5 and should be submitted through NYU classes.
• Group presentation (10%): Students will work in groups through the entire class on a research project addressing a public policy topic. At the end of the semester groups are required to present their results.
• Group final paper (20%): Each group will be required to submit a final research paper at the end of the semester. This paper should outline the analysis, results, implications, potential impact as well as
recommendations for further research according to following structure: abstract, introduction/problem definition, literature review and related previous work, data description and methodology, results and Implications, limitations, conclusion. This report is due act the end of the semester and should be submitted through NYU classes.

In addition to the group assignments sessions will have individual assignments which consist of a problem set based on the topic addressed in class. The statistical package used to work on the assignments is Python and SQL. Code and Output has to be turned in and will be evaluated. The individual assignments will constitute 50% of the final grade.

- Assignment Session 2 (5%)
- Assignment Session 3 (5%)
- Assignment Session 4 (5%)
- Assignment Session 5 (5%)
- Assignment Session 6 (5%)
- Assignment Session 7 (5%)
- Code and Metadata documentation (20%)

It is imperative that you come to class on time, have read the reading assignment, and are prepared to discuss concepts and questions in class. Attendance will only be taken once: at the very beginning of every class. If you miss class, you must notify the instructors in advance and it is up to you to get notes and materials from another student. Regular attendance & contributive participation in class will constitute 10% of the final grade.

All group projects and individual assignments should be posted on NYU Classes at least 24-hours prior to the beginning of the following session.

Plagiarism

All students must produce original work. Outside sources are to be properly referenced and/or quoted. Lifting copy from websites or other sources and trying to pass it off as your original words constitutes plagiarism. Such cases can lead to academic dismissal from the university.