"BIG DATA" AND
SOCIAL SCIENCE RESEARCH IN THE DIGITAL AGE

www.johnmclevey.com/475

INTEG 475, Winter 2018
Tuesdays and Thursdays, 1:00 PM - 2:20 PM in EV2 2069
Class Slack Group: integ475-winter2018.slack.com (Join here)

Professor: Dr. John McLevey
Office hours: Thursdays, 3 pm - 4:30 pm or by appointment

Course Description

We are living in an age where digital information is being produced at an unprecedented rate. This explosion of digital data has the potential to revolutionize the way we learn about the world, and how we conduct research related to urgent social and political problems. This course focuses primarily on the knowledge and skills necessary for doing high-quality research with digital data. We will begin by considering the promise and pitfalls of "big data" for social science, and on the intersection of research in social science and data science. Then we will start a research-oriented introduction to the programming language Python, followed by a series of classes on collecting, cleaning, and combining digital datasets. Finally, we will move into a series of classes on analyzing digital datasets using tools from machine learning, text analysis, and social network analysis. There will be an emphasis on good research design and ethics throughout the course.

I assume no previous knowledge or experience of computer programming. Previous courses in research methods and / or statistics are an asset, but are not required. The same goes for previous courses in the social sciences, especially sociology and political science.

In designing this course, I was influenced by (broadly) similar courses offered by the sociologists Matthew Salganik (Course: "Computational Social Science" at Princeton University), Laura Nelson (Course: "Digital Methods for Social Sciences and Humanities" at Northeastern University), and Chris Bail (Course: "Computational Sociology" at Duke University). I was also influenced by the general design of the "Data 8: Foundations of Data Science" course at UC Berkeley.
Learning Objectives

The Schedule section of the syllabus identifies the core learning objectives for every scheduled topic in the course.

Books

There are 4 books that we will use extensively. They are available on reserve at the university library (Porter), in the university bookstore, and online. One of them (Python for Everybody: Exploring Data In Python 3) is freely available online as a PDF file, ebook, or HTML book. Currently, you can read Salganik’s book Bit by Bit online as an HTML book, but I'm not sure if it will stay online once the book becomes officially available in print.


In addition to these books, you may find it occasionally helpful to consult other resources for learning Python, especially if you are interested in some specific area, such as natural language processing. I have plenty of suggestions of further things to read, so don’t hesitate to ask.

If you have no previous knowledge of HTML or CSS, you might consider getting yourself a copy of Jon Duckett's (2011) HTML and CSS: Design and Build Websites. It’s a nice (and beautiful) introduction that assumes no prior knowledge. Of course there are plenty of great resources online. You don’t need to buy the book. Your web scraping skills will advance rapidly as you become more familiar with HTML and CSS.

Any other readings will be available on the course website, on reserve, or through the library database.

Many of our classes will be hands-on labs / workshops. I’ll regularly walk the class through example code, but I will do so assuming that you have done the readings before coming to class.
Computing

You need to have access to a laptop, and you need to bring it to every class meeting from January 16th onwards. If that is not possible, talk to me as soon as you can and I will try to help you work out another option.

Before coming to class on January 16th, you should install Anaconda for Python 3.6 on your laptop. Do not get Python 2.7. Get 3.6. Anaconda will install Python on your system, plus Jupyter Notebooks and many of the packages we will need for this course. We will install additional packages throughout the course as necessary.

Almost all of the computing we do in this course will be in Jupyter Notebooks.

In addition to Anaconda, you will need a good text editor. I use and love Sublime Text 3 and the Atom text editor. For this class, I recommend that you install Atom, which is completely open source.
Deliverables

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Dates</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Challenges</td>
<td>Will receive on Feb 6, Feb 13, Mar 1, Mar 15, and Mar 22</td>
<td>30</td>
</tr>
<tr>
<td>Self Assessments</td>
<td>Quantitative due every class, qualitative due with data challenges</td>
<td>10</td>
</tr>
<tr>
<td>Tutorial Notebook</td>
<td>Due anytime after reading week and before March 29th</td>
<td>10</td>
</tr>
<tr>
<td>Final Paper / Report</td>
<td>Due on Friday April 13th</td>
<td>40</td>
</tr>
<tr>
<td>Engagement / Participation</td>
<td>Ongoing</td>
<td>10</td>
</tr>
</tbody>
</table>

Data Challenges

From January 16th onwards, there will be small programming exercises that you will work on during and outside of class. In addition to those small assignments, you will collaborate with one or two other people on “data challenges.” These will be slightly more challenging exercises where you will have to draw on the knowledge you have acquired to do something related to content we are covering in the course at the time (e.g. scrape a website of your choosing and write the results to a new dataset, generate and analyze a semantic network from unstructured text). Your submission must be fully reproducible. In other words, I must be able to open your notebook and run the code myself without errors.

The dates for data challenges listed on this syllabus are actually the dates when I will give you the details of the data challenge, not when they are due. I will give you exactly one week to do the challenges. They will not actually take you that long to do, but I realize that you may wish to collaborate face-to-face and that requires dealing with scheduling constraints.

Finally, when you submit your data challenges, each person must include, as an appendix at the end of the notebook, a paragraph long personal self-assessment. More details on this in the section below.

You will submit your data challenge notebooks (and data) as a zipped folder on LEARN and on slack.
Self Assessments

There is a lot of diversity in this class. Some of you have a lot of experience programming, but most of you have none. Some of you have a lot of experience doing research in the social sciences, but most of you do not. It’s always challenging to run a class like this when people have radically different starting points and backgrounds. But... it’s also really fun and intellectually stimulating!

One of the key ways I will try to support everyone’s individual progress is by relying on frequent honest self-assessments. Being honest is in your self-interest, because if you are not honest you will quickly get left behind, and if I have good reason to suspect you being intentionally dishonest, you risk losing most or all of this part of your course grade. Be honest!

The self-assessments will happen in two ways. Every class you will complete a very short survey (<1 min long) consisting only of fixed-response questions and the occasional short answer question requiring no more than a couple of sentences. The second type of assessment you will do is qualitative. Each of the five collaborative data challenges you submit must include a paragraph-long qualitative self-assessment for every person on the team. I want to know what you did, what you learned, how you pushed yourself, and what you think you need to keep working on. You may also include other details as well, if you think they are relevant.

What am I looking for in these self-assessments? I will be using the quantitative assessments to inform how I deliver lectures, facilitate discussions, and run hands-on workshops during class meeting times. I will use the data to keep track of when you are pushing yourself to the limits of your comfort zone (and therefore putting yourself in a position to learn), and when you are not. Similarly, I will use the qualitative data, combined with my own observations, to better understand what you are doing to push yourself to learn. Obviously, the way to get top marks in this part of the course is to work hard and put yourself at the edge of your comfort zones regardless of where that might be, and regardless of where you started from.

Tutorial Notebook

With 1-3 collaborators, you will produce a Jupyter Notebook that introduces your fellow students to material from one of the chapters in Scruping the Web with Python or Big Data and Social Science that we are not already covering in class (e.g. one of the chapters on working with images, or forms and logins). Your notebook should include one or two complete examples with functional Python code and real data. You must cover all the key content from whatever chapter you have selected, but do not simply reproduce the same code and explanations that you find in the text. Draw on knowledge from other parts of the course. You should also include a paragraph in the notebook that describes exactly what each member of your group contributed, and whether or not your group lived up to the original agreements for collaborating. You may also contact me privately if you want to talk about your collaboration.

I will provide a sign-up form at the start of class on February 15th (right before reading week). You will submit your tutorial notebook (and data) as a zipped folder on LEARN and on slack.
Final Paper / Report

With 1-3 collaborators, you will produce a Jupyter notebook that answers an important or interesting research question by analyzing real data. All text, Python code, and output (e.g. graphs, tables) must be included in the Jupyter Notebook that you submit to me, along with your raw data. I should be able to reproduce your entire analysis as I work through your notebook. You can write about anything you want, but consider running the idea by me first.

Your notebook must begin with at least a few paragraphs describing the general problem your team is concerned with, the specific research questions you are addressing, and why the problem and questions are important or interesting. Your notebook must also include a discussion of the data you are analyzing, where it came from, what it's strengths and limitations are, and what methods you are going to use to analyze that data. (It is not necessary to use methods not covered in the course, although you may do so if you wish.) The findings section of your report must include inline visualizations and tables. Finally, you must interpret the meaning of what you found given the questions you asked, describe what you learned from this assignment, and describe any additional data and / or analyses that you think would help your further the work you started with this notebook.

As with other collaborative work in this course, you should also include a paragraph in the notebook that describes exactly what each member of your group contributed, and whether or not your group lived up to the original agreements for collaborating. You may also contact me privately if you want to talk about your collaboration.

Engagement / Participation

The quality of this course -- like any other -- depends on you being engaged. Your participation grade will be based on (a) contributions to class discussion, (b) small group discussion, (c) your involvement in any online discussions, and (d) attendance. If you really don't like speaking up in class, you can participate more online, but you must speak with me about this. Although I will not be assigning a participation grade until the end of the semester, I am happy to provide qualitative feedback on your participation throughout the semester.

If you arrive more than 10 minutes late, you will lose 50% of the credit for attending class. In other words, arriving late twice is equivalent to missing a class. There is no penalty for excused absences, which always require advance notice and generally require a note from a doctor.
COURSE SCHEDULE

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thursday Jan 04</td>
<td>Introduction</td>
</tr>
<tr>
<td>Tuesday Jan 09</td>
<td>Getting started with Jupyter</td>
</tr>
<tr>
<td>Thursday Jan 11</td>
<td>Social science with digital data</td>
</tr>
<tr>
<td>Tuesday Jan 16</td>
<td>Getting started with Python</td>
</tr>
<tr>
<td>Thursday Jan 18</td>
<td>Getting started with Python</td>
</tr>
<tr>
<td>Tuesday Jan 23</td>
<td>Getting started with Python</td>
</tr>
<tr>
<td>Thursday Jan 25</td>
<td>To scrape the web, understand the web</td>
</tr>
<tr>
<td>Tuesday Jan 30</td>
<td>To scrape the web, understand the web</td>
</tr>
<tr>
<td>Thursday Feb 01</td>
<td>Scrape the web</td>
</tr>
<tr>
<td>Tuesday Feb 06</td>
<td>Crawl the web</td>
</tr>
<tr>
<td>Thursday Feb 08</td>
<td>Collecting digital data from APIs</td>
</tr>
<tr>
<td>Tuesday Feb 13</td>
<td>Parsing documents and cleaning dirty data</td>
</tr>
<tr>
<td>Thursday Feb 15</td>
<td>Digital data + asking questions</td>
</tr>
<tr>
<td>Tuesday Feb 20</td>
<td>Reading week, no class</td>
</tr>
<tr>
<td>Thursday Feb 22</td>
<td>Reading week, no class</td>
</tr>
<tr>
<td>Tuesday Feb 27</td>
<td>Experiments and mass collaborations</td>
</tr>
<tr>
<td>Thursday Mar 01</td>
<td>Machine learning: the basics</td>
</tr>
<tr>
<td>Tuesday Mar 06</td>
<td>Text analysis, supervised and unsupervised</td>
</tr>
<tr>
<td>Thursday Mar 08</td>
<td>Text analysis, supervised and unsupervised</td>
</tr>
<tr>
<td>Tuesday Mar 13</td>
<td>Analyzing networks: the basics</td>
</tr>
<tr>
<td>Thursday Mar 15</td>
<td>Analyzing networks: the basics</td>
</tr>
<tr>
<td>Tuesday Mar 20</td>
<td>Combining data from multiple sources</td>
</tr>
<tr>
<td>Thursday Mar 22</td>
<td>Combining data from multiple sources</td>
</tr>
<tr>
<td>Tuesday Mar 27</td>
<td>Scaling up: tools for research computing with big data</td>
</tr>
<tr>
<td>Thursday Mar 29</td>
<td>Student panel: Some NetLab Students, Past and Present</td>
</tr>
</tbody>
</table>

Detailed Schedule and Learning Objectives

Most of the class descriptions below include suggested readings and/or videos in addition to required reading and videos. I have included these to help you go a bit further than what we
cover in class. You will not lose any points if you decide not dig into the suggested material. You will not be disadvantaged in class either, since I will not likely bring this material up in class.

thursday jan 04 INTRODUCTION

This class meeting will introduce the core themes and learning objectives for the term. By the end of the class, you should (1) know what the key themes of the course are, (2) be able to clearly explain what I expect from you in the course, and (3) understand what you need to do to succeed.

- This syllabus!
- **Supplementary / Optional:** Chapter 1 "Introduction" from *Big Data and Social Science*

tuesday jan 09 GETTING STARTED WITH JUPYTER

By the end of this class, you should (1) have all the required software installed, and (2) be able to start a Jupyter Notebook. Please try to have the software installed on your laptop before class starts. We will likely lose a lot of class time to a few uncooperative machines.

BTW, the reading for Thursday is a long one, so you should aim to be at least 1/2 through it by this point.

- Chapter 1 "Introduction" from *Bit by Bit*

thursday jan 11 SOCIAL SCIENCE WITH DIGITAL DATA

By the end of this class, you should be able to (1) explain how social scientists and data scientists use digital data in observational studies, and (2) describe the 10 common characteristics of "big data" discussed in Salganik: big, always on, nonreactive, incomplete, inaccessible, non-representative, drifting, algorithmically confounded, dirty, and sensitive.

- Chapter 2 "Observing Behavior" from *Bit by Bit*
**tuesday jan 16 GETTING STARTED WITH PYTHON**

By the end of this class, you should be able to (1) assign things to variables, (2) execute code conditionally, and (3) explain what a function is, use built-in functions, and understand when, why, and how you might want to write your functions.

- Chapter 1 "Why should you learn to write programs?" from *Python for Everyone*
- Chapter 2 "Variables, expressions, and statements" from *Python for Everyone*
- Chapter 3 "Conditional execution" from *Python for Everyone*
- Chapter 4 "Functions" from *Python for Everyone*
- **Supplementary / Optional:** Watch and follow along with Jessica McKellar's "A Hands-On Introduction to Python for Beginning Programmers." If you have no previous Python background, I recommend watching and following along with McKellar's tutorial before coming to class.
- **Supplementary / Optional:** Watch Brian Granger, Chris Colbert, and Ian Rose's 2017 talk "JupyterLab: The Evolution of the Jupyter Notebook" to get a sense of what you can do with Jupyter Lab.

---

**thursday jan 18 GETTING STARTED WITH PYTHON**

By the end of this class, you should be able to (1) understand iteration in Python, (2) manipulate strings, and (3) understand basic data structures (lists, dictionaries, and tuples).

- Chapter 5 "Iteration" from *Python for Everyone*
- Chapter 6 "Strings" from *Python for Everyone*
- Chapter 8 "Lists" from *Python for Everyone*
- Chapter 9 "Dictionaries" from *Python for Everyone*
- Chapter 10 "Tuples" from *Python for Everyone*

---

**tuesday jan 23 GETTING STARTED WITH PYTHON**

By the end of this class, you should be able to (1) read, create, open, modify, and save files; and (2) work with data in the form of Pandas dataframes.

- Chapter 7 "files" from *Python for Everyone*
- Watch Daniel Chen’s SciPy 2016 tutorial Introduction to Pandas

---

**thursday jan 25 TO SCRAPE THE WEB, UNDERSTAND THE WEB**

By the end of this class, you should be able to (1) explain the basics of what happens when you visit websites, (2) understand the basics of HTML markup.
• Skim the introduction and first five chapters ("Introduction," "Structure," "Text," "Lists," "Links," and "Images") from Jon Duckett's *HTML and CSS: Design and Build Websites.* (There are a lot of images in this book, and the chapters are very short and easy to understand.)
• Watch "What is the Internet?" from the code.org playlist "How the Internet Works."
• Watch "IP Addresses and DNS" from the code.org playlist "How the Internet Works."
• Watch "Wires, Cables, and Wifi" from the code.org playlist "How the Internet Works."
• Watch "Packets, Routing, and Reliability" from the code.org playlist "How the Internet Works."
• Watch "HTTP and HTML" from the code.org playlist "How the Internet Works."
• Supplementary / Optional: "Encryption and Public Keys" from the code.org playlist "How the Internet Works."
• Supplementary / Optional: "Cybersecurity and Crime" from the code.org playlist "How the Internet Works."
• Supplementary / Optional: "How Search Works" from the code.org playlist "How the Internet Works."

**tuesday jan 30 TO SCRAPE THE WEB, UNDERSTAND THE WEB**

By the end of this class, you should be able to (1) understand more complex HTML markup, (2) explain how CSS works, and (3) understand how to collect data from websites by scraping or using Application Programming Interfaces (API).

• Chapter 2 "Working with Web Data and APIs" from *Big Data and Social Sciences*
• Skim Chapters 6 - 10 ("Tables," "Forms," "Extra Markup," "Flash, Audio, and Video," and "Introducing CSS") from Jon Duckett's *HTML and CSS: Design and Build Websites.* (There are a lot of images in this book, and the chapters are very short and easy to understand.)

**thursday feb 01 SCRAPE THE WEB**

By the end of this class, you should be able to (1) put your new knowledge of HTML and CSS to work in developing your first web scraper.

• Chapter 1 "Your First Web Scraper" from *Web Scraping with Python*
• Chapter 2 "Advanced HTML Parsing" from *Web Scraping with Python*
• Supplementary / Optional: Watch Ryan Mitchell's DEF CON 23 talk "Separating Bots from the Humans." I suggest watching this video if you really want to get into scraping complex web pages. If the material we have covered so far is all new to you, this video may be a bit too advanced, but it’s still worth watching.
- **Supplementary / Optional**: Watch Corey Schafer’s Beautiful Soup tutorial

---

tuesday feb 06 CRAWL THE WEB

By the end of this class, you should be able to (1) develop a web scraper that can crawl!

- Chapter 3 "Starting to Crawl" from *Web Scraping with Python*

---

thursday feb 08 COLLECTING DIGITAL DATA FROM APIS

By the end of this class, you should be able to (1) explain what APIs are, (2) find APIs for data you are interested in, and (3) collect data from APIs.

- Chapter 4 "Using APIs" from *Web Scraping with Python*

---

tuesday feb 13 PARSING DOCUMENTS AND CLEANING DIRTY DATA

By the end of this class, you should be able to (1) parse data from documents rather than web pages, and (2) clean up messy unstructured data from web pages and documents.

- Chapter 6 "Reading Documents" from *Web Scraping with Python*
- Chapter 7 "Cleaning Your Dirty Data" from *Web Scraping with Python*

---

thursday feb 15 DIGITAL DATA + ASKING QUESTIONS

By the end of this class, you should be able to (1) describe the role that survey methods and the 'total survey error framework' play in the digital age; (2) accurately describe historical changes in sampling strategies, and explain why Salganik thinks new approaches to non-probability sampling differ from earlier approaches; (3) describe innovations in how we ask questions; and (4) differentiate between "enriched asking" and "amplified asking."

- Chapter 3 "Asking Questions" from *Bit by Bit*
- Watch Matt Salganik’s 2017 talk "Survey research in the digital age" from the Summer Institute in Computational Social Science
• **Supplementary / Optional:** Watch Chris Bail's 2017 talk "Apps for Social Science Research" from the Summer Institute in Computational Social Science

• **Supplementary / Optional:** Christopher Bail. 2015. "Taming Big Data Using App Technology to Study Organizational Behavior on Social Media." *Sociological Methods & Research* 46(2):1-29.

• **Supplementary / Optional:** Read Chapter 10 "Errors and Inference" from *Big Data and Social Science*

---

tuesday feb 20 READING WEEK, NO CLASS

• None!

---

thursday feb 22 READING WEEK, NO CLASS

• None!

---

tuesday feb 27 EXPERIMENTS AND MASS COLLABORATIONS

By the end of this class, you should be able to (1) explain the role of experiments in social science and data science and explain how they are changing in the digital age, and (3) explain the role of mass collaborations in social science and data science and explain how they are changing in the digital age.

• Chapter 4 "Running Experiments" from *Bit by Bit*

• Chapter 5 "Creating Mass Collaboration" from *Bit by Bit*

• Watch Devah Pager's overview of her experimental audit study "The Mark of a Criminal Record"


---

thursday mar 01 MACHINE LEARNING: THE BASICS

By the end of this class, you should be able to (1) define machine learning and explain how it fits into the larger analysis landscapes of social science and data science; and (2) define some key machine learning vocabulary, concepts, and processes.

• Chapter 6 "Machine Learning" from *Big Data and Social Science*


tuesday mar 06 TEXT ANALYSIS, SUPERVISED AND UNSUPERVISED

By the end of this class, you should be able to (1) explain the potential contributions that computational text analysis can make to research in social science and data science, (2) differentiate between supervised and unsupervised text analysis, and (3) prepare unstructured text data for analysis in Python.

- Chapter 7 "Text Analysis" from *Big Data and Social Science*
- David Zentgraf "What every programmer absolutely, positively needs to know about encodings and character sets to work with text"
- Watch Jennifer Pan's Databite talk "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument"


thursday mar 08 TEXT ANALYSIS, SUPERVISED AND UNSUPERVISED

By the end of this class, you should be able to (1) identify important words in documents, (2) measure the similarity of different texts / documents, (3) develop a topic model in Python, and (4) use natural language processing and network analysis to construct and explore semantic networks in Python.

- Christopher Bail. 2015. "Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media." *PNAS*
- Alix Rule, Jean-Philippe Cointet, and Peter Bearman. 2015. "Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014." *PNAS* 112(35)
tuesday mar 13 ANALYZING NETWORKS: THE BASICS

By the end of this class, you should be able to (1) describe some of the key concepts, theories, and methods from social network analysis; and (2) accurately characterize similarities and differences between network analysis in the social sciences and the natural and computational sciences.

- Watch Filiz Garip on network effects and social inequality
- Watch Mark Granovetter on social networks and getting a job
- Watch Sandra Gonzalez-Bailon describe her new book Decoding the Social World: Data Science and the Unintended Consequences of Communication

Supplementary / Optional: Watch Sandra Gonzalez-Bailon's 2017 talk "Decoding the Social World" from the Summer Institute in Computational Social Science. It's pretty advanced relative to what you have learned so far, so if you watch this video, focus on trying to understand the problem that Dr. Gonzalez-Bailon's is addressing, the logic of her research design, and her key arguments.


thursday mar 15 ANALYZING NETWORKS: THE BASICS

By the end of this class, you should be able to (1) explain how social networks shape the diffusion of information, ideas, and behaviors; (2) create and modify networks in Python; (3) compute basic network statistics, and identify groups, positions, and roles.
• Chapter 8 "Networks: The basics" from *Big Data and Social Science*
• Watch James Evans talk about political echo chambers and consumption of science.

• **Supplementary / Optional:** Watch Lada Adamic's talk "Social Networks as Information Filters." The talk is advanced for what you know so far. Focus on understanding the problem that Dr. Adamic is addressing, her research design, and her core findings.
• **Supplementary / Optional:** Watch Jure Leskovec’s 2011 talk "Rhythms of Information Flow Through Networks." Again, it’s advanced, so focus on understanding the big picture.
• **Supplementary / Optional:** Andrey Rzhetskya, Jacob Foster, Ian Foster and James Evans (2015) "Choosing experiments to accelerate collective discovery." PNAS. 112(47):14569–14574.
• **Supplementary / Optional:** Lingfei Wu, Dashun Wang, James Evans (2017) "Large Teams Have Developed Science and Technology; Small Teams Have Disrupted It" arXiv:1709.02445
• **Supplementary / Optional:** Watch a tutorial on the python package *networkx*. This tutorial by Rob Chew and Peter Baumgartner covers much of the same material we will cover in class. "Connected: A Social Network Analysis Tutorial with NetworkX"

tuesday mar 20 COMBINING DATA FROM MULTIPLE SOURCES

By the end of this class, you should be able to (1) explain what record linkage and discuss the challenges and opportunities it introduces, and (2) differentiate between approaches to record linkage.

• Chapter 3 "Record Linkage" from *Big Data and Social Science*
• Watch Patrick Ball’s Databite talk "Understanding Patterns of Mass Violence with Data and Statistics"
thursday mar 22 COMBINING DATA FROM MULTIPLE SOURCES

By the end of this class, you should be able to (1) understand and modify the code for a record linkage workflow in Python, including preparing datasets for record linkage, comparing records, classifying records, fusing datasets, and evaluating linkages.

- From the NetLab blog, read: Jillian Anderson's "Introduction to recordlinkage"
- From the NetLab blog, read: Jillian Anderson's "Pre-processing with recordlinkage"
- From the NetLab blog, read: Joel Becker's "Indexing Candidate Links with recordlinkage"
- From the NetLab blog, read: Joel Becker's "Sorted Neighbourhood Indexing with recordlinkage"
- From the NetLab blog, read: Jillian Anderson's "Record Comparison with recordlinkage"
- If available, from the NetLab blog, read posts on classification, fusion, and evaluation.

tuesday mar 27 SCALING UP: TOOLS FOR RESEARCH COMPUTING WITH BIG DATA

By the end of this class, you will be able to (1) compare databases with other ways of storing and accessing data; (2) explain the basic ideas behind Hadoop, MapReduce, and Spark, and time permitting; (3) define and understand how to use the dataframe API for PySpark.

- Skim Chapter 4 "Databases" from Big Data and Social Science
- Skim Chapter 15 "Using Databases and SQL" from Charles Severance’s Python for Everyone
- Read Chapter 5 "Programming with Big Data" from Big Data and Social
- **Supplementary / Optional:** Watch Andrew Ray’s talk "Data Wrangling with PySpark for Data Scientists Who Know Pandas."
STUDENT PANEL: SOME NETLAB STUDENTS, PAST AND PRESENT

If we can make their schedules work, we will have a student panel with some of my former students and NetLab researchers: Jillian Anderson (now in a data science MSc at Simon Fraser), Joel Becker (now a data scientist at Shopify), Tahin Monzoor (current NetLab student, previously worked on a data science team at Interact), Steve McColl (currently a data scientist with the federal government), Tiff Lin (currently a research fellow at the Berkman Klein Center for Internet and Society at Harvard University), and Reid McIlroy-Young (currently in a computational social science Master's program at the University of Chicago).

- Chapter 6 "Ethics" from *Bit by Bit*
- Chapter 7 "The Future" from *Bit by Bit*

Course Policies

Submitting Work & Grading Process

I will only grade work that you submit electronically.

Text matching software will be used to screen assignments in this course. This is being done to verify that use of all materials and sources in assignments is documented. Students will be given an option if they do not want to have their assignment screened by Turnitin. If you do not wish to submit your work to Turnitin, you can schedule a meeting with me to discuss your submissions in person.

Late Policy

I will deduct **5 points** a day for every day, or part of a day, that your work is late, including weekends. I will not make exceptions without a medical note.

Laptops and the Facebook Penalty

Laptops may be used in the classroom on the honors system, but you must sit in the designated laptop section. I reserve the right to modify this policy if laptops appear to be interfering with student learning. If I see Facebook, email, an IM client other than #slack, a newspaper story, a blog, or any other content not related to the class, I will remove 1 point from your participation grade on the spot.

Communication

We will be using the collaboration tool #slack for regular class communication. I use the do not disturb settings on #slack, so I will not see any messages you send me outside of normal working
hours. You are free to email me, of course, but I tend to respond to slack messages from students faster than I respond to emails. There are free #slack apps for Mac OS X, Linux (beta), Windows (beta), iOS, and Android.

Preferred Chat System, by XKCD

Feedback

I will solicit brief, informal, and confidential course evaluations throughout the semester. These will only take a few minutes of your time. The purpose is to make sure that we are moving at a comfortable pace, that you feel you understand the material, and that my teaching style is meeting your needs. I will use this ongoing feedback to make adjustments as the course progresses. Although you are not obligated to do so, please fill out the evaluations so that I can make this the best learning experience for you, and the best teaching experience for me.

On Campus Resources

Access Ability Services

The AccessAbility Office, located in Needles Hall, Room 1132, collaborates with all academic departments to arrange appropriate accommodations for students with disabilities without compromising the academic integrity of the curriculum. If you require academic accommodations to lessen the impact of your disability, please register with the AccessAbility Office at the beginning of each academic term.

Mental Health

The University of Waterloo, the Faculty of Environment, and our Departments consider students’ well-being to be extremely important. We recognize that throughout the term students may face health challenges -- physical and / or emotional. Please note that help is available. Mental health is a serious issue for everyone and can affect your ability to do your best work. Counselling Services is an inclusive, non-judgmental, and confidential space for anyone to seek
support. They offer confidential counselling for a variety of areas including anxiety, stress management, depression, grief, substance use, sexuality, relationship issues, and much more.

The Writing Centre

Although I will be giving you feedback on your work throughout the term, I encourage you to make appointments with people at the writing centre. Their services are available to all UW students.

University Policies

Academic Integrity

In order to maintain a culture of academic integrity, members of the University of Waterloo community are expected to promote honesty, trust, fairness, respect and responsibility.

We will all uphold academic integrity policies at University of Waterloo, which include but are not limited to promoting academic freedom and a community free from discrimination and harassment. You can educate yourself on these policies -- and the disciplinary processes in place to deal with violations -- on the Office of Academic Integrity website.

A student is expected to know what constitutes academic integrity, to avoid committing academic offense, and to take responsibility for his/her actions. A student who is unsure whether an action constitutes an offense, or who needs help in learning how to avoid offenses (e.g., plagiarism, cheating) or about 'rules' for group work / collaboration should seek guidance from the course professor, academic advisor, or the Undergraduate Associate Dean. For information on categories of offences and types of penalties, students should refer to Policy 71, Student Discipline. For typical penalties, check Guidelines for Assessment of Penalties.

Grievances and Appeals

A student who believes that a decision affecting some aspect of his / her university life has been unfair or unreasonable may have grounds for initiating a grievance. Read Policy 70: Student Petitions and Grievances, Section 4. When in doubt please contact your Undergraduate Advisor for details.

A decision made or penalty imposed under Policy 70 -- Student Petitions and Grievances (other than a petition) or Policy 71 -- (Student Discipline) may be appealed if there is a ground. A student who believes he/she has a ground for an appeal should refer to Policy 72 (Student Appeals).

Religious Observances

Student needs to inform the instructor at the beginning of term if special accommodation needs to be made for religious observances that are not otherwise accounted for in the scheduling of classes and deliverables.