INTEG 475

Open Science & Technology

Tuesday and Thursday, 1:00 - 2:20 in the KI Studio
#slack: f2015-integ475.slack.com

John McLevey, PhD
Knowledge Integration
Sociology & Legal Studies
University of Waterloo

DESCRIPTION

This is a collaborative and project-driven course. Students will learn how to collect, analyze, and visualize data to answer policy relevant questions, or questions of broad social scientific / sociological interest. We will work with open data provided by governments or organizations, and data mined from web sites and social media (e.g. Twitter). While having some prior experience with research methods and / or programming is advantageous, it is **not** necessary to do well in this course. For students without any previous experience programming, this course will provide an applied – and friendly – introduction to programming for data collection and analysis using Python. For students with some programming experience, this course will provide an introduction to applied data analysis for policy and social science research.

LEARNING OUTCOMES

There are two general, but very important, learning objectives in this course. **First**, students who do not have any experience programming will learn how to think about collecting and analyzing data like a programmer. This requires learning some basic programming concepts. On the other hand, students who do have some experience programming will lean how to think about collecting and analyzing data like a social scientist or policy researcher. Again, this will require learning basic theory and methods from the social sciences. I will emphasize sociology because I am a sociologist, but you are free to look to other social sciences for ideas.

Second, students will learn how to write simple programs in Python and R for collecting and analyzing data. In particular, students will learn how to collect and manage data, how to do basic text and network analyses, and how to visualize data effectively. Almost everything will be done in Python written in Jupyter Notebooks or in R Markdown files. **I am assuming no prior knowledge of Python or R.**

DELIVERABLES AND EVALUATION

Assignment	Deadline	Value
6 collaborative mini-projects	Ongoing	30%
10 self-assessment reports	Every week for weeks 2-11	10%
1 article review (500 words)	Due the day before class	10%
1 collaborative data analysis project	Dec. 1	50%

There will be 6 in class data challenges. Students will collaborate in groups of 3-5 to answer a specific empirical questions using computational tools learned in the course. Each of these challenges will be worth 5% of the total course grade, and should be discussed (whenever appropriate) in self-assessment reports. Team members will all get the same grade if the collaboration seems equitable and inclusive.

Students will complete self-assessment reports every week from week 2 through 11. Each self-assessment should be thorough and honest, and should make *explicit* reference to experiences in class and in collaborative groups.

Each student will pick 1 empirical article from the list of social science articles listed in the first part of the "Download Links" section to review. Reviews should emphasize issues related to research design, with a particular focus on how the author(s) used data and methods to answer their research questions. Reviews are due by noon the day before the relevant class.

Finally, students will design their own collaborative projects, due at the end of the course. We will discuss these projects over the course of the semester. The most important requirements are (1) that they are informed by a social science literature, and (2) that they they use data collection and analysis techniques learned in the course. The final projects may be submitted in the form of a Jupyter Notebook or a final paper. If the project is submitted as a final paper, it must include all code and data used in the analysis.

READINGS

Almost all of the readings for this course are articles, and are therefore available online via the University of Waterloo library website. In addition, I recommend downloading a free copy of:

Downey, Allen (2008) *Think Python: How to Think Like a Computer Scientist*. Needham, Massachusetts: Green Tea Press. Green Tea Press makes a PDF of the book available for free on their website.

SOFTWARE

We will have an "installation party" on the second day of class, but please try and install as much of this as you can before class. There is quite a lot to install, but you do not have to pay for anything.

Software	Download Link
Python	active link, see electronic version
VirtualBox	active link, see electronic version
Vagrant	active link, see electronic version
Java 8	active link, see electronic version
Atom text editor	active link, see electronic version
LaTeX	active link, see electronic version
Haskell	active link, see electronic version
Pandoc	active link, see electronic version
R	active link, see electronic version
Rstudio	active link, see electronic version

SUBMITTING WORK & LATE POLICY

You will submit all work electronically in the general channel on #slack **and** on LEARN. **You may not submit hard copies or Microsoft Word documents under any circumstances**. Please submit PDF files instead. 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.

COMMUNICATION

We will be using the collaboration tool slack for all class communication. Of course you are free to email me, but I tend to respond to slack messages from students faster than I respond to emails. Sign up with and sign into slack by going to slack.com. There are slack apps for Mac OS X, iOS, and Android. If you are a Linux or Windows user, there is a very good web app.

Feedback

I will solicit brief, informal, and confidential course evaluations three or four times 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

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.

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.

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.

SCHEDULE & READINGS

	Date	Substantive Topic	Hands On!
1	T, Sept. 5	Introduction	None
1	Th, Sept. 17	Installation Party!	None
2	T, Sept. 22	The Command Line for Social Scientists	Navigating the file system
2	Th, Sept. 24	Plain Text for Social Scientists	Bash, Markdown, Pandoc
3	T, Sept. 29	Reproducible Social Science, Python	Bash, Python, Git
3	Th, Oct. 1	Data Mining: Twitter	Python: *Notebooks*, *Twitter API*
4	T, Oct. 6	Working with Data: Twitter	Python: *pandas*
4	Th, Oct. 8	Working with Data: Twitter	Python: *pandas*
5	T, Oct. 13	Soc. of Knowledge, (Big) Data Mining	Python: *metaknowledge*, *pandas*
5	Th, Oct. 15	Soc. of Knowledge, Data Visualization	R: *knitr*, *dplyr*, *ggplot2*
6	T, Oct. 20	Soc. of Knowledge, Clustering, Networks	R: *iGraph*, *ggplot2*
6	Th, Oct. 22	Network Visualization	R: *iGraph*
7	T, Oct. 27	**Collaborative Data Analysis**	Python, R
7	Th, Oct. 29	**Collaborative Data Analysis**	Python, R
8	T, Nov. 3	Reproducible Workflows, Dynamic Documents	GNU Make, R: *knitr*, *pander*
8	Th, Nov. 5	Intro to Analyzing Textual Data	R: *tm*, *topicmodels*, *mallet*
9	T, Nov. 10	**Collaborative Data Analysis**	R
9	Th, Nov. 12	Mining Textual Data	Python: *BeautifulSoup*
10	T, Nov. 17	**Collaborative Data Analysis**	Python, R
10	Th, Nov. 19	Analyzing Textual Data	R: *tm*, *topicmodels*, *mallet*
11	T, Nov. 24	**Collaborative Data Analysis**	Python, R
11	Th, Nov. 26	Interactive Data Visualizations	R: *ggiz*, *shiny*, *d3Network*
12	T, Dec 1	**Collaborative Data Analysis**	Python, R
12	Th, Dec 3	Debrief, Collaborative Redesign	None

Readings by Week Number

All **bolded** articles below are eligible for article reviews.

1. **Lazer et al. (2009)** "Life in the network: the coming age of computational social science"
2. Milligan and Baker (2014) "Introduction to the Bash Command Line," Baker and Milligan (2014) "Counting and mining research data with Unix," Kieran Healy (2014) "Plain Text, Papers, Pandoc," and Tenen and Wythoff (2014) "Sustainable Authorship in Plain Text using Pandoc and Markdown."
3. Gentzkow and Shapiro (2014) "Code and Data for the Social Sciences: A Practitioner's Guide" and **DiGrazia et al. (2013)** "More tweets, more votes." Required Video: Jessica McKeller's talk "A hands on introduction to Python for beginners". Optional reading: Russell (2013) *Mining the Social Web*.
4. **Bail (2014)** "The cultural environment: measuring culture with big data" and Julia Evans (2014) "Pandas Cookbook." Optional reading: McKinney (2012) *Python for Data Analysis*.
5. **James Evans and Foster (2011)** "Metaknowledge" and **Healy and Moody (2014)** "Data Visualization in Sociology." Optional Videos: Hadley Wickham's (1) "expressing yourself in R", and (2) tutorial on dplyr.
6. **Carrington (2015)** "Social Networks Research" (posted on slack) and Ognyanova (2014) "Network Visualization with R." Suggested video: James Evan's talk "How Science Thinks, and

How to Think Better." Optional readings: Bellotti (2012) "Getting funded: Multi-level network of physicists in Italy," Verd and Lozares (2015) "Reconstructing Social Networks through Text Analysis."

7. **Roger Peng (2015)** "Data Analysis as Art," "The Epicycles of Analysis," and "Stating and Refining the Question."
8. **Miller (2013)** "Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach," and Chen (2011) "Introduction to Latent Dirichlet Allocation."
9. **Blei (2012)** "Probabalistic Topic Models," **Farrell (2015)** "Corporate funding influences the production and thematic content of ideological polarization," and **King, Pan, and Roberts (2013)** "How censorship in China allows government criticism but silences collective expression"
10. **Nelson (2015)** "Political Logics as Cultural Memory: Cognitive Structures, Local Continuities, and Women's Organizations in Chicago and New York City," **Mohr and Bogdanov (2013)** "Topic models: What they are and why they matter," and Wieringa (2012) "Intro to Beautiful Soup"
11. **Hanna (2013)** "Computer-aided content analysis of digitally enabled movements" and **Golder and Macy (2014)** "Digital footprints: opportunities and challenges for online social research." Watch Hadley Wickham's talk "The State of Interactive Graphics."
12. **Watts (2013)** "Computational social science"

USEFUL "CHEAT SHEETS" FOR R

Rstudio has useful cheatsheets for R Markdown, data visualization using ggplot2, interactive visualizations with Shiny, and data wrangling with dplyr. You can get them at `https://www.rstudio.com/resources/cheatsheets/`.

FULL REFERENCES

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3-4): 465–82.

Baker, James, and Ian Milligan. 2014. *Counting and Mining Research Data with Unix*. Programming Historian. `http://programminghistorian.org/lessons/research-data-with-unix.html`.

Bellotti, Elisa. 2012. "Getting Funded: Multi-Level Network of Physicists in Italy." *Social Networks* 34(2): 215–29.

Blei, David. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4): 77–84.

Carrington, Peter. 2015. "Social Networks Research." In *Mixed Methods Social Networks Research: Design and Applications*, eds. Silvia Domínguez and Betina Hollstein.

Chen, Edwin. 2011. *Introduction to Latent Dirichlet Allocation*. `http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/`.

DiGrazia, Joseph, Karissa McKelvey, Johan Bollen, and Fabio Rojas. 2013. "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior." *PLOS One*.

Downey, Allen. 2008. "Think Python: How to Think Like a Computer Scientist."

Evans, James, and Jacob Foster. 2011. "Metaknowledge." *Science*.

Evans, Julia. 2014. *Pandas Cookbook*. http://jvns.ca/blog/2013/12/22/cooking-with-pandas/.

Farrell, Justin. 2015. "Corporate Funding Influences the Production and Thematic Content of Ideological Polarization." *in progress*.

Gentzkow, Matthew, and Jesse Shapiro. 2014. *Code and Data for the Social Sciences: A Practitioner's Guide*. http://web.stanford.edu/~gentzkow/research/CodeAndData.pdf.

Golder, Scott, and Michael Macy. 2014. "Digital Footprints: Opportunities and Challenges for Online Social Research." *Sociology* 40(1): 129.

Hanna, Alexander. 2013. "Computer-Aided Content Analysis of Digitally Enabled Movements." *Mobilization: An International Quarterly* 18(4): 367–88.

Healy, Kieran. 2014. *Plain Text, Papers, Pandoc*. http://kieranhealy.org/blog/archives/2014/01/23/plain-text/.

Healy, Kieran, and James Moody. 2014. "Data Visualization in Sociology." *Annual review of sociology* 40: 105–28.

King, Gary, Jennifer Pan, and Margaret E Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(02): 326–43.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and others. 2009. "Life in the Network: The Coming Age of Computational Social Science." *Science (New York, NY)* 323(5915): 721.

McKinney, Wes. 2012. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc."

Miller, Ian Matthew. 2013. "Rebellion, Crime and Violence in Qing China, 1722–1911: A Topic Modeling Approach." *Poetics* 41(6): 626–49.

Milligan, Ian, and James Baker. 2014. *Introduction to the Bash Command Line*. Programming Historian. http://programminghistorian.org/lessons/intro-to-bash.html.

Mohr, John W, and Petko Bogdanov. 2013. "Topic Models: What They Are and Why They Matter." *Poetics* 41(6): 545–69.

Nelson, Laura K. 2015. "Political Logics as Cultural Memory: Cognitive Structures, Local Continuities, and Women's Organizations in Chicago and New York City." *Revise and Resubmit at American Journal of Sociology*.

Ognyanova, Katherine. 2014. "Network Visualization with R." *POLNET 2015 Workshop, Portland OR*.

Peng, Roger. 2015. *The Art of Data Science: A Guide for Anyone Who Works with Data*. LeanPub.

Russell, Matthew. 2013. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. " O'Reilly Media, Inc."

Tenen, Dennis, and Grant Wythoff. 2014. *Sustainable Authorship in Plain Text Using Pandoc and Markdown*. Programming Historian. http://programminghistorian.org/lessons/sustainable-authorship-in-plain-text-using-pandoc-and-markdown.html.

Verd, Joan Miquel, and Carlos Lozares. 2015. "Reconstructing Social Networks Through Text Analysis: From Text Networks to Narrative Actor Networks." In *Mixed Methods Social Networks Research: Design and Applications*, eds. Silvia Domínguez and Betina Hollstein.

Watts, Duncan J. 2013. "Computational Social Science: Exciting Progress and Future Directions." *The Bridge on Frontiers of Engineering* 43(4): 5–10.

Wieringa, Jeri. 2012. *Intro to Beautiful Soup*. Programming Historian. http://programminghistorian.org/lessons/intro-to-beautiful-soup.html.