Data Science & Population Processes
SOC 401/538 - Autumn 2017

Instructor: Emilio Zagheni
E-mail: emilioz@uw.edu
Phone: (206) 616-1173
Office: Savery Hall 235

Lecture: Mon. 3:30-5:50pm
LOW 114
Office Hours: by appointment

Teaching Assistant: Connor Gilroy
E-mail: cgilroy@uw.edu
Office hours: TBA

Rapid increases in computational power and the explosion of Internet, social media and mobile phone use have radically changed our lives, the way we interact and our behavior, including demographic choices. The digitalization of our lives has led to the so-called “data revolution” that is transforming social sciences. “Big Data” and data science tools offer social scientists the opportunity to address core social questions in new ways.

In this course, we will study how traditional methods used in social sciences can help us make sense of new data sources, and how these new data sources may require new approaches and research design. There will be a mix of lectures, student-led discussions and demos, and hands-on computational activities with various tools. The course covers substantive topics relevant to demographic research as well as a selection of data science tools to extract Internet data, manage large data sets, analyze them, and visualize them.

Goals: In this course, we will discuss a number of substantive topics related to the emergence of (big) data-driven discovery in social sciences, with emphasis on population processes. By the end of the course, students will be familiar with relevant literature at the intersection of demographic research and computational social science. The main goals of the course are i) to develop critical thinking about the emergent field of big data analysis, ii) to learn some of the methods, approaches and tools of data science in the context of population research, iii) to identify research questions in your own area of interest that could be addressed with innovative data sources and to devise an appropriate research plan.

Diversity of Student Backgrounds: Students in this class have different backgrounds. Some students are pursuing a PhD or an MA, some others are undergraduates. Some students may have strong computational and statistical skills, some others may not. Some students may be familiar with population studies, some others not. To accommodate the range of backgrounds, I emphasize substance, and key statistical and computational

concepts. There will also be different types of homework assignments. Some of them will involve computing and coding. Some others will be critical reflections about the readings. In short, I facilitate and encourage the participation of students who do not have extensive background in statistics, or computational methods, but are eager to learn.

Course Requirements and Grading

Participation & Contribution	10%
Homework assignments	50%
Term paper	30%
Project presentations	10%
Total	100%

Class Participation & Contribution: Class participation & Contribution will count towards your final grade. Please help create a constructive learning environment. Different people have different ways in which they participate best, all of which are valid: thoughtful preparation, sharing a well-formulated idea after a long pause, stimulating discussion through questions, helping a classmate understand a concept, asking a classmate for help, discussing ideas and challenges during office hours, sharing news articles with the class, etc. I strongly encourage you to interact with me and the other students. Even if you feel uncertain about how to express something, I would prefer that you speak up. Listen to your peers, wait for your turn to speak, and refrain from using discriminatory language. If you are a talker, make sure that your quieter peers get a chance to speak. If you are shy, remember that if you have a question, most likely there is at least one other person with the same question who would be happy to listen to the answer.

One particularly meaningful way to contribute to the class is by sharing what you know and what interests you. For example, if you are interested in a journal article that is relevant for the general topic of the course, but is not in the syllabus, please share it with your classmates via the Discussion board on Canvas and include a short paragraph where you present the main results and explain why you have chosen to share that the paper and what is relevant about it. Similarly, if you have developed a data science tool or have been using one that you think is relevant for other people in the class, please share it via Canvas with a brief explanation so that other students can follow the code as if it were a mini-demo or tutorial.

Homework assignments: There will be homework assignments almost every week. For some assignments, you will be expected to work on some technical problems related to statistical concepts or computational tasks discussed in class. These assignments would require coding, mostly in R. For other assignments, you may be expected to write a short commentary about assigned readings or topics. You will be asked to work in small groups (2-3 people) on the assignments, but each person of the group must submit a copy of the assignment and report the names of all the group participants.

Term paper: The term project is an empirical research brief on a relevant topic of your choice. You could replicate existing studies, test new ideas, develop a new visualization that is relevant for your own research, etc.. You could use existing data or collect your own. I encourage you to be adventurous. The style and sophistication of analysis depend on the student's background. In terms of format and length, you should follow the guidelines for submission to the *Descriptive Findings* series of *Demographic Research*: http://www.demographic-research.org/info/general_information.htm. Early in the term, you should discuss your ideas with me, so that we can define a feasible plan. The term project can be done individually or as a small group. The term paper is due on Monday, December 11, at 12noon.

Project presentations: In Week 6, you (and your group if applicable) will be asked to give a brief presentation of your term project idea. As we will all be aware of what everyone else is working on, we can offer feedback to each other. During the last class meeting, in Week 10, you will be asked present your term project and your results. This is your chance to practice your communication skills and to receive further constructive feedback from your peers. The write-up for the final paper is due about a week after your presentation. That way, you can incorporate the feedback that you received after your presentation.

Class Conduct

The class atmosphere will be quite relaxed. Just a few guidelines to make sure:

- Arriving a bit late is tolerated as long as you make an effort to minimize the disturbance for other students.

- Eating and drinking in class is allowed, but please make sure that you are not disturbing others.

- Please turn off your cellphone or put it on silent mode.

- If you cannot make it to class for whatever reason, make sure that you know what happened during the lecture and lab that you missed.

- If you are having trouble with the course material or personal problems that are hindering your performance in the class, please come and talk to me so that we can solve the problem before it is too late. It is better to bring up any concerns as early as they arise.

- Please always show respect to your fellow classmates.

Students with Disabilities

Please inform me as soon as possible of special needs that you may have.
The sooner you notify me, the better I will be able to make appropriate arrangements.

Academic Integrity

A fundamental tenet of all educational institutions is academic honesty. Students must do all their work within the boundaries of acceptable academic norms. See the UW statement about student academic responsibility prepared by Committee on Academic Conduct in the College of Arts and Sciences (https://depts.washington.edu/grading/pdf/AcademicResponsibility.pdf). Students found guilty of plagiarism or academic dishonesty will be subject to appropriate disciplinary actions.

Course schedule, format and reading list

Each session will be a mix of lecture, discussion, and lab (hands-on computational activities). I will provide source code and material for the lab on a weekly basis. Most of the coding examples will be in R. Some familiarity with R is useful for this course.

Below is the course schedule and list of readings. The reading list may change. Additional readings, including news reports, demos and tutorials may be added during the course of the quarter, depending on students' interests and time availability.

Week 1 **Mon, Oct 2nd - Introduction: Challenges and opportunities for "Big Data" research**
Lab: Data manipulation with R: Baby names data

[1] Ginsberg, J., Mohebbi, M.H., Patel R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2008) Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, 457(7232):1012-1014.

[2] Lazer, D. M., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science* 343(6176):1203-1205

[3] Chae, D.H., Clouston, S. et al. (2015) Association between an Internet-based Measure of Area Racism and Black Mortality. *PloS ONE* 10(4)

[4] Ruths, D., Pfeffer, J. (2014). Social Media for Large Studies of Behavior. *Science*, *346*(6213), 1063-1064.

[5] Billari, F, Zagheni, E. (2017) Big Data and Population Processes: a Revolution? *Proceedings of the Conference of the Italian Statistical Society 2017.*

[6] Lazer, D. and Radford, J. (2017) Data ex Machina: Introduction to Big Data. *Annual Review of Sociology* (43):19-39

[6] Torfs, P. and Brauer, C. (2014). A (very) Short Introduction to R. https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

Week 2 **Mon, Oct 9th - Twitter data and difference-in-differences estimation**
Lab: Collect and analyze Twitter data

[1] Flores, R. (2017, forthcoming). Do Anti-immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 using Twitter Data. *American Journal of Sociology*.

[2] Zagheni, E., Garimella, K., Weber, I. and State, B. (2014). Inferring International and Internal Migration Patterns from Twitter Data. *Proceedings of ACM WWW (Companion):* 439-444.

Week 3 **Mon, Oct 16th – Guest lecture by Connor Gilroy**
General principles for using APIs; Accessing Facebook data

Week 4 **Mon, Oct 23th – Addressing selection bias in `digital breadcrumbs'**
Lab: Shell scripting for processing data

[1] Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting Elections with Non-Representative Polls. *International Journal of Forecasting.* 31:980-991.

[2] Zagheni, E. and Weber, I. (2012). You are where you E-mail: Using E-mail Data to Estimate International Migration Rates. *Proceedings of ACM Web Science*

[3] State, B., Rodriguez, M., Helbing, D. and Zagheni, E. (2014) Migration of Professionals to the US: Evidence from LinkedIn Data. *Proceedings of Social Informatics.*

[4] Zagheni, E. and Weber, I. (2015) Demographic Research with non-representative Internet Data. *International Journal of Manpower*. 36(1):13-25.

[5] Zagheni, E., Weber, I. and Gummadi, K. (2017). Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review*.

Week 5 **Mon, Oct 30th - Guest Lecture by Lee Fiorio**
Studying Spatial Mobility in the Digital Age

Week 6 **Mon, Nov 6th - Mobile phones, demography and development**
Lab: Scalability

Student lightning talks about their term projects.

[1] Blumenstock, J.E. (2012). Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda. *Information Technology for Development*, 18(2):107-125.

[2] Blumenstock, J.E., Cadamuro, G and On, R. (2015) Predicting Poverty and Wealth from Mobile Phone Metadata. *Science*, 350:1073-1076.

[3] Palmer, J.R.B., Espenshade, T.J., Bartumeus, F., Chung, C.Y., Ozgencil, N.E., and Li K. (2012). New Approaches to Human Mobility: Using Mobile Phones for Demographic Research. *Demography* (50):1105-1128.

[4] Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D. and Tatem, A.J. (2014) Dynamic Population Mapping Using Mobile Phone Data. *Proceedings of the National Academy of Sciences* 111(45):15888-15893.

Week 7 **Mon, Nov 13th: Ethical issues, privacy and reproducible research**
Lab: Scalability (continued) and version control

[1] Zimmer, M. (2010). But the Data is Already Public: On the Ethics of Research in Facebook. *Ethics and information technology*, *12*(4), 313-325.

[2] Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24):8788-8790.

[3] Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Carnegie Mellon University Data Privacy Working Paper 3. Pittsburgh 2000.

[4] Ullman, J. D., Leskovec, J., & Rajaraman, A. (2011). *Mining of Massive Datasets* (pp. 305-338). Cambridge University Press.

Week 8 **Mon, Nov 20th: Web Experiments, Regression Discontinuity and Research Design**
Lab: Interactive Data Visualization with Rshiny

[1] Lewis, R., Rao, J.M. and Reiley, D. (2008). Measuring the Effects of Advertising: The Digital Frontier

[2] Salganik, M.J. and Watts, D.J. (2009) Web-based Experiments for the Study of Collective Social Dynamics in Cultural Markets. *Topics in Cognitive Science*, 1(3):439{468, 2009.

[3] Luca, M. (2011) Reviews, Reputation and Revenue: The Case of Yelp.com. Working Paper 12-016. Harvard Business School.

[4] King, G., Pan, J. and Roberts, M.E. (2014) Reverse-engineering censorship in China: Randomized Experimentation and Participant Observation, *Science* 345(6199).

Week 9 **Mon, Nov 27th: The digitalization of the life course**
Lab: TBA, depending on students' interest

[1] Trudeau, J. (2016). The Role of New Media on Teen Sexual Behaviors and Fertility Outcomes – The case of 16 and Pregnant. *Southern Economic Journal*, *82*(3), 975-1003.

[2] Bönisch, P. and Hyll, W. (2015). Television Role Models and Fertility – Evidence from a Natural Experiment. *SOEPpapers* 752.

[3] Dettling, L.J. (2016). Broadband in the Labor Market: The Impact of Residential High-Speed Internet on Married Women's Labor Force Participation. *ILR Review*.

[4] Bellou, A. (2015) The Impact of Internet Diffusion on Marriage Rates: Evidence from the Broadband Market. *Journal of Population Economics* 28, 265-297.

[5] Cacioppo, J.T., Cacioppo, S., Gonzaga, G.C., Ogburn, E.L. and VanderWeele, T.J. (2013). Marital Satisfaction and Break-ups Differ across On-line and Off-line Meeting Venues. Proceedings of the National Academy of Sciences 110(25):10135-10140.

[6] Freese, J., Rivas, S. and Hargittai, E. (2006) Cognitive Ability and Intenet Use among Older Adults. *Poetics* 34:236-249.

Week 10 **Mon, Dec 4th: Conclusions and students' presentations**

Monday, Dec 11**: The term paper is due via canvas by Monday, Dec 11 at 12noon.**