Big Data & Population Processes
SOC 590 - Autumn 2015

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

Lecture: Mon. 3:30-6:20pm
Savery Hall 157
Office Hours: by appointment

Rapid increases in computational power and the explosion of Internet and mobile phone use have radically transformed our lives, the way we communicate and our behavior. This digital revolution has also fundamentally changed social sciences. "Big Data", simulation models and Web experiments 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 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 and analyze them.

Diversity of Student Backgrounds: Students in this class have different backgrounds. Some students are pursuing a PhD, some others are enrolled in an MA program. 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.

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 big data analysis, 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

Course Requirements and Grading

Participation & student-led discussion	10%
Homework assignments	50%
Term paper	30%
Final presentation	10%
Total	100%

Class Participation & student-led discussions: Class participation 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: active listening, thoughtful preparation, sharing a well-formulated idea after a long pause, stimulating discussion through questions, helping a classmate understand a concept, 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 rather have you speak up than say nothing at all. 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. During the quarter, you will be asked to choose a paper of your choice that is relevant for the general topic of the course, and share it with your classmates in advance of our meeting. In class, you would briefly present the main results, explain why you have chosen the paper and what is relevant about it, and lead a short discussion about the paper.

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. For other assignments, you will be expected to write a short commentary about assigned readings or topics. You may work in small groups (2-3 people) on the assignments.

Term paper: The term project is an empirical research brief on a relevant topic of your choice. You could replicate existing studies or test new ideas. 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, 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.

Final presentation: The last class meeting will be devoted to students' presentations of their term projects. This is your chance to practice your communication skills and to receive constructive feedback from your peers.

Class Conduct

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

- Arriving a few minutes 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 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 we will be able to accommodate you.

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 examples will be provided in R (and in some cases, in Python). Familiarity with R is useful, but not a pre-requisite. The

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 5th - Introduction: Challenges and opportunities for "Big Data" research**
Lab: Data manipulation with R: Baby names data

[1] Hey, T., Tansley, S., & Tolle, K. M. (Eds.). (2009). The Fourth Paradigm: Data-intensive Scientific Discovery.

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

Week 2 **Mon, Oct 12th - Twitter data and difference-in-differences estimation**
Lab: Collect and analyze Twitter data

[1] Flores, R. (2015). Do Anti-immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 using Twitter Data. *Working paper*.

[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 19th – Addressing selection bias in `digital breadcrumbs'**
Lab: Using APIs and adding dimensions to the 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.

Week 4 **Mon, Oct 26th: Web search engine queries**
Lab: Managing large data sets and scalability

[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] Billari, F., D'Amuri, F. and Marcucci, J. (2013) Forecasting Births Using Google. *Meeting of the Population Association of America, 2013*.

[4] Reis, B.Y. and Brownstein, J.S. (2010) Measuring the Impact of Health Policies Using Internet Search Patterns: The Case of Abortion. *BMC Public Health*, 10(1):514, 2010.

Week 5 **Mon, Nov 2nd: Ethical issues and privacy**
Lab: Managing large data sets and scalability (continued)

[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] Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of Large Sparse Datasets. IEEE symposium on Security and Privacy.

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

Week 6 **Mon, Nov 9th - Mobile phones, demography and development**
Lab: Dimensionality reduction in demographic research

[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. and Eagle, N. (2012) Divided we Call: Disparities in Access and Use of Mobile Phones in Rwanda. *Information Technologies & International Development*, 8(2):1-16.

[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 16th: Web experiments**
Lab: Data Visualization

[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.

Week 8 **Mon, Nov 23th: Demographic microsimulation and agent-based models**
Lab: Statistical algorithms

[1] Bruch, E. E., & Mare, R. D. (2006). Neighborhood Choice and Neighborhood Change. *American Journal of Sociology*, *112*(3), 667-709.

[2] Todd, P. M., Billari, F.C. and Simao, J. (2005). Aggregate age-at-marriage patterns from individual mate-search heuristics. *Demography*, *42*(3), 559-574.

[3] Ševčíková, H., Raftery, A. E., & Waddell, P. A. (2007). Assessing uncertainty in urban simulations using Bayesian melding. *Transportation Research Part B: Methodological*, *41*(6), 652-669.

Week 9 **Mon, Nov 30th: Internet and Demographic Behavior**
Lab: TBD

[1] Nie, N.H. and Hilligus, D.S. (2002). The Impact of Internet Use on Sociability. *IT&Society* 1(1):1-20.

[2] Sautter, J.M., Tippett, R.M. and Morgan S.P. (2010). The Social Demography of Internet Dating in the United States. *Social Science Quarterly* 91(2):554-575.

[3] Potarca, G., & Mills, M. (2015). Racial Preferences in Online Dating Across European Countries. *European Sociological Review* 1-16

Week 10 **Mon, Dec 7th: Conclusions and students' presentations**