# Georgia Tech

## Course syllabus

## Special Topics: Big Data & Public Policy

### PUBP8813 | ECON8803

**Dr. Omar Isaac Asensio**
**School of Public Policy**

**Room: Skiles 246, T Th 1:30pm-2:45pm**
**Term: Spring 2018**

**This version: 03.12.18**

## Contact Information

Office hours: T Th 3:00pm-4:15pm or by appt
Office location: DM Smith 317, E-mail: asensio@pubpolicy.gatech.edu
TA: Austin McTier, e-mail: amctier@gatech.edu
Course website: canvas.gatech.edu

## Course Information

This course will provide an introduction to "big data" for social science applications. My goal is for you to walk out of this class with hands-on tools to analyze public policies and economic behavior using the wealth of newly available data sources being generated by business and government. Students will learn to conduct experiments and to identify causal mechanisms in large-scale social and administrative data. You will have opportunities to critically evaluate both experimental and observational research through readings and cases as well as develop an original data science proposal 3.000 credit hours, 3.000 lecture hours

## Learning Objectives

Upon successful completion of this course, you should be able to:
- Identify compelling uses of big data to solve social and economic problems;
- Distinguish between research designs for prediction versus inference;
- Experiment with cloud technologies for storage, analysis and computation;

- Develop case studies related to big data and sustainable communities;
- Gain experience in presenting and defending research;
- Consider the protection of human subjects in commissioned studies.

## Methods Requirement

*Big Data and Public Policy* is targeted for Ph.D. or advanced M.S. students in Public Policy; M.S. students in Economics, and M.S. students in Cybersecurity. Suggested prerequisites include at least 1 course in elementary statistics and probability. Some prior programming experience is helpful, but is not required.

## SLS Course

This course is part of Georgia Tech's Serve-Learn-Sustain (SLS) initiative. SLS works with all six colleges to offer courses and programs connecting sustainability and community engagement with real-world partners and projects on the theme of "sustainable communities." More information about SLS can be found at www.serve-learn-sustain.gatech.edu. Visit the website to sign up for the SLS Email List, view the full list of affiliated courses, and find links to Facebook, Instagram and Twitter.

## Grading

Your final grade will be determined as follows:

| Assignment | Points | Date |
|------------|--------|------|
| Problem Set 1 | 10 | Wk 4 |
| Problem Set 2 | 10 | Wk 7 |
| Problem Set 3 | 10 | Wk 12 |
| Midterm | 20 | Wk 8 |
| Final Research Proposal (presentation) | 10 | Wks 13-14 |
| Final Research Proposal (written) | 30 | Wk 16 |
| IRB Training/SLS Events | 5 | Wk 2-3 |
| Participation/CIOS +1 | 5 | Ongoing |

Grade points will be assigned to a letter grade according to the following scale:

| Grade | Score |
|-------|-------|
| A | 80+ |
| B | 70-79 |
| C | 60-69 |
| D | 50-59 |
| F | less than 50 |

You will get individual performance feedback on the problem sets, the midterm and your final proposal. The problem sets will be based on concepts from the

readings and lectures. They will include short exercises and may include hands-on programming tasks. The midterm will be closed book and closed notes. It will include conceptual and quantitative questions and you will be asked to analyze aspects of a manuscript or recent peer-reviewed article. The midterm will test your ability to find appropriate research designs to answer specific questions. Your final project will consist of a written proposal and presentation in your policy area of interest. This year's project themes will align with SLS and the UN Sustainable Development Goals. You will be allowed to partner with 1 or 2 students on your final project should you choose to do so. The final written proposal and presentation will be worth 40% of your grade.

You will also have a bonus opportunity during the semester (and no later than Week 13) to submit a draft version of your research proposal for suggestions. This will be ungraded, but will be your chance to improve the quality of your final project.

**Please come talk to me during office hours if you run into any issues and please do not be afraid to ask questions. Above all, work hard, be creative and have fun!**

## Course policies

### Academic Integrity

We abide by Georgia Tech's academic honor code. I encourage you to collaborate and exchange ideas with your peers. As plagiarism is easily discoverable with software, I expect you to turn in original work and cite your sources regularly in all assignments. Your best scholarship is in your own words.

### Participation and Attendance

Be proactive in doing the assigned readings before class and please come prepared for discussion. Each student will have the opportunity to lead a discussion of one or more readings. In-class participation is essential. Give yourself the best chance to do well with high-quality contributions in class and in your teams if you work with others on a common proposal. At the end of the term, you will be able to rate your team members on their participation and you will assess each other's performance. Your TA will also be looking for evidence of active engagement and will help me assign the participation scores. We document all student engagements and attendance in Canvas, office hours and in class.

### Late Assignments, Missed Exams, Extensions

All assignments will be due on or before 5:00pm on the scheduled delivery date. In general, we will not take late assignments. Also, given the long-term nature of the research proposal, we will not provide extensions save for extenuating circumstances. I would suggest getting started on the planning and writing of your proposal as soon as possible. Make-up exams are discouraged and given

only for <u>documented</u> reasons of illness, family emergency, or participation in Institute approved activities. If you should have any conflicts, please come talk to me outside of class as soon as possible.

### Extra Credit

Be on the lookout for extra credit such as bonus questions appearing on problem sets or the midterm. These questions could be drawn from lectures or the readings and provide additional opportunities to demonstrate that you've absorbed the material covered in class.

### Use of Laptops and Mobile Devices in the Classroom

So as not to disrupt student learning, please turn off the sound on all devices and no texting or messaging.

### Accommodations for Students with Disabilities

If you are a student that should require any special accommodations to help you succeed in this course, please get in touch with the Office of Disability Services at (404) 894-2563 or disabilityservices.gatech.edu and e-mail me as soon as possible to set up a time to discuss your needs.

### Campus Safety

To read the law and guidance on HB280, please visit usg.edu/hb280. On other matters related to personal safety, feel free to download the LiveSafe app at livesafe.gatech.edu or call the GaTechPD at (404) 894-2500.

### Ways to Contact

I encourage you to come chat with me during my office hours and to communicate frequently on the course site. These are great ways to increase your engagement and also gives me a chance to get to know you. You can reach me via email at asensio@pubpolicy.gatech.edu. **Please use your Georgia Tech email address and include "PUBP 8813" or "ECON8803" in the subject line.** I will generally respond promptly to your messages, but please remember I may not be immediately available on weekends. Please also check in with your TA periodically with any questions or concerns.

## Course Materials

### Canvas

We are using Canvas—Georgia Tech's new learning management software. All relevant course materials, discussions and announcements will be uploaded online at canvas.gatech.edu. If you wish, you can download the Canvas app

4

from the App Store or Google Play. Be sure to set your Canvas `notification preferences` to receive electronic updates. I will post the latest information including additional reading materials and resources online, so please check often. **We expect you to comment and engage frequently on Canvas**.

If you run into any technical issues with `Canvas`, students can get help from the 24/7 help desk at (877) 259-8498 or via email `support@instructure.com`. You can also simply click the "Help" button within campus to reach support. Please bear with us as we pilot new features for the benefit of the Institute.

### Library Resources

To assist with your projects, we will hold a class session during week 3 on library resources with our Public Policy subject librarian. Participation in this class session is required. You will also schedule a follow-up consultation directly with our librarian for more personalized assistance. To get you started, please link to the library research guide for our class here: `http://libguides.gatech.edu/PUBP_BigData`. My personal desk copies of course textbooks will also be available on reserve at the library.

### IRB Certification for Human Subjects Research

The Institutional Review Board (IRB) is a federally mandated panel tasked with protecting the rights and welfare of human subjects in research. At Georgia Tech, this process is administered by the Office of Research Integrity Assurance (ORIA). Each student in this course is required to complete the IRB certification process for conducting research using human subjects. To complete the training, follow the instructions online at `http://www.researchintegrity.gatech.edu/irb`. Please note that it may take up to 24 hrs or more after you register to be able to access the course. To activate the training modules, click the `CITI Program` link under "Required Training" and login with your Georgia Tech credentials. Be sure to choose the course titled: "Basic Course in the Protection of Human Research Subjects." The entire program should take about 2-3 hours to complete. To receive credit for participation, you must upload a PDF of your completion certificate or your email confirmation to `Canvas` under the module titled Research Ethics. For more information about IRB, we will hold a course session on IRB training with representatives from ORIA during week 3.

### Required Texts

[**BDSS**] Ian Foster, Rayid Ghani, Ron S Jarmin, Frauke Kreuter and Julia Lane. *Big Data and Social Science: A Practical Guide to Methods and Tools.* CRC Press, Taylor and Francis Group, Boca Raton, FL, 2017.

[**CSS**] R. Michael Alvarez. *Computational Social Science: Discovery and Prediction.* Cambridge University Press, New York, NY 2016.

[**MW**] Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge University Press, Cambridge, UK, 2007.

**Other Helpful Resources**

Alex Bell, *Python for Economists.* Available online: [http://scholar.harvard.edu/files/ambell/files/python_for_economists.pdf](http://scholar.harvard.edu/files/ambell/files/python_for_economists.pdf)

Kaggle Datasets: [https://www.kaggle.com/datasets](https://www.kaggle.com/datasets)

## Schedule

### Wk 1: The Case for Big Data in Social Science

| DATE | DESCRIPTION | DUE: |
|------|-------------|------|
| Jan 9 | **First class meeting** <br> Course introduction and overview; assign teams. | In-class survey |
| Jan 11 | **Big Data Prediction Versus Inference** <br><br> • Read CSS: Intro 1-24; BDSS: Ch.1, 1-19 <br> • S. Athey, Beyond prediction: Using big data for policy problems. *Science* (2017) 355(6324):483-485. <br> • McKinsey Global Institute. *The Age of Analytics: Competing in a data-driven world.* Full report (2016) <br> • McKinsey Global Institute. *Big data: The next frontier for innovation, competition, and productivity.* Executive summary (2011). | Bring 1 peer-reviewed article to class you are willing to defend. |

### Wk 2: Data Responsibility

| Jan 16 | **Consumer Targeting and Measurement.** <br> [GUEST SPEAKER] Becca Marco, comScore <br><br> • C. Warshaw, The application of big data in surveys to the study of elections, public opinion, and representation. in CSS: 27-50. <br> • A. Pentland, Reality mining of mobile communications: Toward a new deal on data. *Global Information Technology Report* (2009) World Economic Forum. 75-80. <br> • D. Greenwood et al., The new deal on data: A framework for institutional controls. https://doi.org/10.1017/CBO9781107590205.012 available online at http://civics.com/NewDealOnData/ <br> • Kramer et al., Experimental evidence of massive-scale emotional contagion through social networks. *PNAS*, (2014) 111(24):8788-8790. | IRB training online registration |
| Jan 18 | **Big Data in the Government Sector** <br><br> • R.S. Jarmin and A.B. O'Hara, Big data and the transformation of public policy analysis. *Journal of Policy Analysis and Management* (2016) 35(3): 715-721. <br> • G.H. Kim et al, Big-data applications in the Government sector. *Communications of the ACM* (2014) 57(3): 78-85. <br> • [CASE STUDY]: Predicting Fuel Economy | Complete IRB training course |

**Wk 3: Privacy, Confidentiality and IRB Training**

| Jan 23 | **Institutional Review Board** [GUEST SPEAKER] Carolyn Sims, GT Office of Research Integrity Assurance (ORIA). <br><br> • Best Predictors and Conditional Probability <br> • Read BDSS: 299-311. <br> • Read The Belmont Report. Ethical principles and guidelines <br> • 45 CFR part 46, Federal policy for the protection of human subjects, subpart A: "The Common Rule" | Submit topic for final project |
|---|---|---|
| Jan 25 | **Library resources, R Tutorial** [GUEST SPEAKER] Ameet Doshi, GT Public Policy subject librarian. <br><br> • Bounding Identification regions <br> Meet in the Homer Rice classroom, ground floor of GT Library. | |

**Wk 4: Causal Inference and Counterfactuals**

| Jan 30 | **Causality and Counterfactuals** <br><br> • Read MW: Ch. 1.1 to 1.4 <br> More advanced students might want to read: <br> • Imai et al., Misunderstandings between experimentalists and observationalists about causal inference. *Statistics in Society Series A* (2008) 171(2): 481-502. | |
|---|---|---|
| Feb 1 | **Causal Graphs I: Identifying models** <br><br> • Read MW: Ch 3 <br> • CF Manski, Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* (1993) 60(3): 531-542. | Problem Set 1 due |

**Wk 5: Identification in Large Datasets**

| Feb 6 | **Causal Graphs II: Non-identifying models** <br><br> • Read MW: Ch. 4 <br> • E. Bareinboim and J. Pearl, Causal inference and the data-fusion problem. *PNAS* 113(27): 7345-7352. | |
|---|---|---|
| Feb 8 | **Randomization and the Selection Problem** <br><br> [PYTHON DEMO] Data Capture and web scraping <br> • Read BDSS: 23-31 <br> • Additional reading: Nitsch et al. *American Journal of Epidemiology* 163(5): 397-403 | |

## Wk 6: Randomized Experiments

| Feb 13 | **Causal Graph Examples and RCTs**<br><br>[Class Discussion] Causal graph examples | Causal graph for proposal due |
|---|---|---|
| Feb 15 | **Randomized Controlled Trials**<br><br>• Harrison and List, Field Experiments. *Journal of Economic Literature*, (2004) 42(4): 1009-1055.<br>• List, Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives*, (2011) 25(3): 3-16.<br>• Asensio and Delmas, Nonprice incentives and energy conservation. *PNAS*, (2015) 112(6): E510-E515<br>• [Data Demo] Real-time smart meter data | Proposal outline |

## Wk 7: Quasi-Experiments

| Feb 20 | **Matching Algorithms for Causal Effects**<br>• Read MW: Ch. 5<br>• Asensio and Delmas, The effectiveness of US energy efficiency building labels. *Nature Energy* 2: 17033. | |
|---|---|---|
| Feb 22 | **Causal Graphs III and Midterm Review**<br>• Additional materials will be posted on Canvas | Problem set 2 due |

## Wk 8: Midterm and Microsoft Azure Cloud Technologies

| Feb 27 | **Implementation Demo Using Cloud Technologies**<br>[GUEST SPEAKER] Giota Asensio, Keysight Technologies (formerly HP labs)<br><br>• Additional materials will be posted on Canvas | |
|---|---|---|
| Mar 1 | **In-class Midterm**<br>Closed book and closed notes. | Midterm |

**Wk 9: Mid-Course Assessments and Quasi-Experiments II**

| Mar 6 | **Midterm solutions and student feedback**<br><br>• Midterm returned and course survey results | |
|---|---|---|
| Mar 8 | **Regression Discontinuity Designs**<br><br>• Lee, Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, (2008) 142: 675-697.<br>• D. Caughey and J. Sekhon, Elections and the regression discontinuity design: Lessons from close U.S. house races. *Political Analysis* (2011) 19: 385-408. | In-class survey 2 |

**Wk 10: Machine Learning and Google Cloud Technologies**

| Mar 13 | **ML and Google Cloud Platform**<br>[GUEST SPEAKER] Sanjay Agravat, Google<br><br>• Supervised ML and Big Data on Google cloud<br>• See cloud.google.com | Proposal outline due |
|---|---|---|
| Mar 15 | **ML and Policy Evaluation**<br>[SUPERVISED MACHINE LEARNING] Hands on session<br><br>• S. Athey and G. Imbens, The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2): 3-32.<br>• Prof. Asensio at UC Berkeley. Keith will give interactive demo of transportation policy case study | |

**Wk 11: Spring Break**

| Mar 20 | **No classes for spring break** | |
|---|---|---|
| Mar 22 | **No classes for spring break**<br>Use time to submit or work on draft proposals. | |

**Wk 12: Big Data Applications**

| Mar 27 | **Big Data Applications I:**<br>**Prediction using IBM Watson Cloud**<br>[SUPERVISED MACHINE LEARNING] Hands on session<br>Materials to be posted online. | Submit draft proposal for feedback. |
|---|---|---|
| Mar 29 | **Big Data Applications II: Data Visualization**<br>[GUEST SPEAKER] Ximin Mi, Data Viz Lab<br>See libguides.gatech.edu/dataviz/intro | |

**Wk 13: Big Data Applications Cont'd and Begin Final Presentations**

| Apr 3 | **Big Data Applications III:** **Proposal Workshop** Interactive session. | Problem Set 3 due |
|---|---|---|
| Apr 5 | **Student Presentations Day 1** Participation is required. | Student peer-review |

**Wk 14: Final Projects**

| Apr 10 | **Student Presentations Day 2** Participation is required. Final in-class assessment. | Student peer-review |
|---|---|---|
| Apr 12 | **Student Presentations Day 3** Participation is required. Turn in SLS Showcase submissions (data visualization, caption, summary) | Student peer-review. End of policy datathon. |

**Wk 15: Final Projects**

| Apr 17 | **No class today.** Use time to incorporate peer evaluation feedback. | |
|---|---|---|
| Apr 19 | **SLS In-Class Focus Group.** Led by Carol Thurman, Center for SLS. | |

**Wk 16: End of term**

| Apr 24 | **Georgia Tech Final Instructional Day** | |
|---|---|---|
| Apr 30 | **Final Written Proposal Due 5:00PM** | Proposal due. Final survey. |