

Course syllabus

Data Science for Public Policy

# PUBP 3042

Dr. Omar Isaac Asensio School of Public Policy

Room: Skiles 311, M W 9:30-10:45am Term: Fall 2018

This version: 12.19.18

# Contact Information

Oﬃce hours: M W 11:00am - 12:15pm or by appt

Oﬃce location: DM Smith 317, E-mail: [asensio@pubpolicy.gatech.edu](mailto:asensio@pubpolicy.gatech.edu)

TA: Susie Ha, e-mail: [s.ha@gatech.edu](mailto:s.ha@gatech.edu)

Course website: canvas.gatech.edu

# Course Information

This course provides an introduction to data science tools and quantitative methods for social science applications. My goal is for you to walk out of this course with hands-on learning experiences to be able to identify and empir- ically test theories and hypotheses with large-scale social and administrative data. Students will have an opportunity to engage with both experimental and observational research as well as design original data science experiments. To do this, you will learn fundamentals of prediction, causal inference, statistical programming and applied machine learning for policy analysis. The course cul- minates in a ‘Policy Datathon’ competition at the end of the semester. 3.000 credit hours, 3.000 lecture hours.

# Learning Objectives

Upon successful completion of this course, you should be able to:

* Identify social and administrative data sources to address policy issues;
* Distinguish between models of prediction versus causal inference;
* Collect, pre-process, and analyze data programatically;
* Gain experience in presenting and defending research;
* Consider ethical issues of fairness, privacy, and the protection of human subjects in commissioned studies;
* Evaluate how decisions impact the sustainability of communities.

# Methods Requirement

*Data Science for Public Policy* is an introductory course in computational social science. The course is targeted for advanced undergraduate students in public policy, economics and other social science fields. The course is also suitable for computer science or engineering students looking to apply data science tools in social and policy domains. Students may find it useful to have previously taken at least 1 probability and statistics course (PUBP 3120 or equivalent). Some prior programming experience in Python or R 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 sustainabil- ity 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.](http://www.serve-learn-sustain.gatech.edu/) Visit the SLS website to sign up for the SLS mailing list and view the full list of affiliated courses.

# 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 |
| Problem Set 4 | 10 | Wk 14 |
| Midterm | 20 | Wk 9 |
| Final Project Proposal | 30 | Wk 15-16 |
| IRB/Library/SLS training | 5 | Ongoing |
| Participation score/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 project proposal. The problem sets will be based on concepts from the readings and lectures. They will include multi-part exercises and hands-on tasks with sample datasets. The midterm will be closed book and closed notes. It will include both conceptual and quantitative questions and you will be asked to analyze aspects of a manuscript or recent peer-reviewed article in creative ways. The midterm will test your ability to find appropriate research designs to answer specific questions. Your final project consists of a written project proposal and its presentation in an interactive competition. You will apply data science solutions to tackle a specific policy challenge. The topic varies by semester.

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

# Policy Datathon

Your final project this semester will involve a Policy Datathon competition that will run in conjunction with the Jump into STEM initiative sponsored by the U.S. Department of Energy Building Technologies Oﬃce and Oak Ridge National Lab. The theme of the competition is data science and sustainable communities. The title of the challenge is “Smart/Connected Solutions for Building Performance.” Your task is to develop data-driven solutions to motivate residential and commercial building owners to manage energy use. 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 30% of your grade. Frequently asked questions about qualifying Jump projects are available here: https://web.ornl.gov/sci/buildings/jump/faq/

Please come talk to me during oﬃce hours if you run into any issues and please do not be afraid to ask questions often. 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 col- laborate and exchange ideas and code implementations with your peers. As plagiarism is easily discoverable with software, I expect you to turn in original work and to 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 pre- pared for discussion. Each student will have the opportunity to lead discussions of one or more readings. In-class and online participation in Canvas is essen- tial. Give yourself the best chance to do well with high-quality contributions in class and in your teams. 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 perfor- mance. Your TA will also be looking for evidence of active engagement and will help me assign the participation scores. We document all student engagement, attendance and analytics in Canvas and during oﬃce hours.

Late Assignments, Missed Exams, Extensions

All assignments will generally 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 your final projects, we will not provide extensions save for extenuating circumstances. I would suggest getting started on the planning and writing of your project proposal as soon as possible. Make-up exams are discouraged and given only for documented 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.

Absences

A student may miss class on occasion due to medical issues. Georgia Tech has a web page that describes the expectations, rights, and responsibilities of students, instructors, the Oﬃce of Student Life, and health care providers. The information is intended to give students better direction as to how they should proceed to notify instructors when they are ill and need to miss class and what kind of documentation they should provide and to whom. Students should refer to Georgia Tech policies at the “Student Absence from Class Due to Illness or Personal Emergencies” web page at [www.catalog.gatech.edu/policies/](http://www.catalog.gatech.edu/policies/) student-absence-regulations

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 read- ings 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.

Commitment to Diversity

The Ivan Allen College of Liberal Arts supports the Georgia Tech’s commit- ment to creating a campus free of discrimination on the basis of race, color, religion, sex, national origin, age, disability, sexual orientation, gender identity, or veteran status. We further aﬃrm the importance of cultivating an intellec- tual climate that allows us to better understand the similarities and diﬀerences of those who constitute the Georgia Tech community, as well as the necessity of working against inequalities that may also manifest here as they do in the broader society.

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 GaTech PD at (404) 894-2500.

Ways to Contact

I encourage you to come chat with me during my oﬃce hours and to commu- nicate 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.](mailto:asensio@pubpolicy.gatech.edu) Please use your Georgia Tech email address and include “PUBP 4803” 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 on- line at canvas.gatech.edu. You can download the Canvas app from the App Store or Google Play. Be sure to set your Canvas notification preferences to receive electronic updates on your phone or by email. I will post the lat- est 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.](mailto:support@instructure.com) You can also simply click the “Help” button within canvas 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 4 on library resources including an R bootcamp with our Public Policy subject librarian. Participation in this class session is required. To get you started with your projects, please link to the library research guide for our class here: http:// libguides.gatech.edu/PUBP4803. My personal desk copy 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 certifi- cation 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 Pro- tection 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 2.

Required Text

[QSS] Kosuke Imai. *Quantitative Social Science: An Introduction*. Princeton University Press, Princeton, NJ 08540, 2017.

Other Helpful Texts

[CSS] R. Michael Alvarez. *Computational Social Science: Discovery and Pre- diction*. 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 Uni- versity Press, Cambridge, UK, 2007.

Other Helpful Resources

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

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

# Schedule

Wk 1: Introduction

|  |  |  |
| --- | --- | --- |
| Date | Description | Due: |
| Aug 20 | First class meeting  Course introduction and overview; assign teams. | In-class  survey |
| Aug 22 | Data Science in Public Policy.   * Review Probability Fundamentals, QSS: 6.1 to 6.2 * Video short: What is computational social science? * R.S. Jarmin & A.B. O’Hara, Big data and the transformation of public policy analysis. *Journal of Policy Analysis and Management* (2016) 35(3): 715- 721. * G.H. Kim et al, Big-data applications in the Gov- ernment sector. *Communications of the ACM* (2014) 57(3): 78-85. * G. King, Big Data is not about the data! * S. Athey, Beyond prediction: Using big data for policy problems. *Science* (2017) 355(6324):483-485. * J. Kleinberg et al, Prediction policy problems.   *American Economic Review* (2015) 105(5): 491-495. | Write a 1/2  page motivation statement; complete IRB online registration. |

Wk 2: Ethical Issues and Data Responsibility

|  |  |  |
| --- | --- | --- |
| Aug 27 | Institutional Review Board  [guest speaker] Carolyn Sims, GT Oﬃce of Research Integrity Assurance (ORIA).   * Read The Belmont Report. Ethical principles and guidelines * 45 CFR part 46, Federal policy for the protection of human subjects, subpart A: "The Common Rule" * Kramer et al., Experimental evidence of massive- scale emotional contagion through social networks. *PNAS*, (2014) 111(24):8788-8790. | Complete  IRB training course |
| Aug 29 | Conditional Prediction   * Best Predictors and Conditional Probability * Read QSS: 6.1 to 6.5 * [Programming Bootcamp]: Intro to R. | Problem set  1 announced |

Wk 3: Programming Fundamentals

|  |  |  |
| --- | --- | --- |
| Sep 3 | No class meeting today (Labor Day) |  |
| Sep 5 | Introduction to Python   * Bounding identification regions * [Programming Bootcamp]: Intro to Python. |  |

Wk 4: Causal Inference and Counterfactuals

|  |  |  |
| --- | --- | --- |
| Sep 10 | Causality and Counterfactuals   * Read QSS: 2.1 to 2.4   More advanced students might want to read:   * Imai et al., Misunderstandings between experimen- talists and observationalists about causal inference. *Statistics in Society Series A* (2008) 171(2): 481-502. [Python demo] Data capture and web scraping |  |
| Sep 12 | Introduction to R and Library Resources  [guest speaker] Ameet Doshi, GT Public Policy subject librarian.  Meet in the Homer Rice classroom, ground floor of GT Library. | Problem  Set 1 due |

Wk 5: Randomized Controlled Trials

|  |  |  |
| --- | --- | --- |
| Sep 17 | Identification in large data sets   * Harrison and List, Field Experiments. *Journal of Economic Literature*, (2004) 42(4): 1009-1055. * List, Why economists should conduct field exper- iments and 14 tips for pulling one oﬀ. *Journal of Economic Perspectives*, (2011) 25(3): 3-16. * Asensio and Delmas, Nonprice incentives and energy conservation. *PNAS*, (2015) 112(6): E510-E515 * [Data Demo] Racial Discrimination in the Labor Market |  |
| Sep 19 | Regression and Causation   * Read QSS: 4.1 to 4.3 * Swirl Exercises R, Causality2 * Threats to Validity I-IV |  |

Wk 6: Quasi-Experiments and Observational Studies

|  |  |  |
| --- | --- | --- |
| Sep 24 | Matching Algorithms for Causal Eﬀects   * Asensio and Delmas, The eﬀectiveness of US energy eﬃciency building labels. *Nature Energy* 2: 17033. * [In-Class Case Study] Behavioral science and policy | Swirl  Probability1 Probability2 bonus due |
| Sep 26 | Regression Discontinuity Designs   * Read QSS: 4.3 to 4.4 * Lee, Randomized experiments from non-random selection in U.S. House elections. *Journal of Econo- metrics*, (2008) 142: 675-697. * D. Caughey and J. Sekhon, Elections and the regression discontinuity design: Lessons from close   U.S. house races. *Political Analysis* (2011) 19: 385-408.   * Additional readings TBA |  |

Wk 7: Observational Studies and Final Project Planning

|  |  |  |
| --- | --- | --- |
| Oct 1 | Jump into STEM Kickoﬀ   * Guest lecture: Caroline Hazard & Dr. Im Piljae * Meet in Clough 262 Classroom * Read QSS: 2.5 to 2.7   https://jump.ideascale.com/a/campaign-home/22230 |  |
| Oct 3 | R Interactive class   * R Programming Demo |  |

Wk 8: Big Data Applications: Augmented Reality

|  |  |  |
| --- | --- | --- |
| Oct 8 | No Class. Fall Recess | Problem  set 2 due  Oct 8, 5pm |
| Oct 10 | Spatial Data: Augmented Reality Data   * Guest lecture: Jamie Clarke, COO Skignz * Meet in Clough 323 Classroom * Read QSS 5.3 * Additional materials will be posted on Canvas |  |

Wk 9: Supervised Machine Learning: Intro to Cloud Computing

|  |  |  |
| --- | --- | --- |
| Oct 15 | ML Implementation Demo Using Cloud Technologies (Microsoft Azure)  [guest speaker] Giota Asensio, Keysight Labs   * Additional materials to be posted on Canvas | Final project  outline + 15 references due |
| Oct 17 | Midterm review   * Additional materials TBA |  |

Wk 10: Midterm

|  |  |  |
| --- | --- | --- |
| Oct 22 | Midterm  Closed book and closed notes. |  |
| Oct 24 | Midterm exam review |  |

Wk 11: Classification: Natural Language Processing

|  |  |  |
| --- | --- | --- |
| Oct 29 | ML Process overview   * Best predictors and loss functions * Variance-Bias Tradeoﬀ * Metrics for ML Evaluation criteria * Khun and Johnson, *Applid Predictive Modeling*, Springer, NY, 2013 |  |
| Oct 31 | Text as Data I  [supervised machine learning] Hands on session   * Read QSS: 5.1 * [Python demo]: IBM Watson Cloud |  |

Wk 12: Sentiment Analysis and Prediction & Visualization

|  |  |  |
| --- | --- | --- |
| Nov 5 | Text as Data II   * Spatial Data: Read QSS 5.3 * [Python demo]: Emotion Analysis using Watson |  |
| Nov 7 | Big Data Applications II: Data Visualization using Tableau Server  [guest speaker] Ximin Mi, Data Viz Lab See libguides.gatech.edu/dataviz/intro | Problem  set 3 due |

Wk 13: Applied Machine Learning I: Cloud Technologies

|  |  |  |
| --- | --- | --- |
| Nov 12 | Data Viz Workshop  Dicussion session on data visualization for final projects | Submit draft  proposal for feedback. |
| Nov 14 | Microsoft Azure ML Studio  [supervised machine learning] Hands on session with Microsoft Azure ML cloud   * Predict Plugscore using random forest models |  |

Wk 14: Applied Machine Learning II

|  |  |  |
| --- | --- | --- |
| Nov 19 | Convolutional Neural Networks for Text Clas- sification   * Asensio, O.I. et al, (2018) Public sentiment of elec-   tric vehicle owners in the United States.  •[Azure ML cloud demo] Sentiment prediction |  |
| Nov 21 | No class | Problem  set 4 due. |

Wk 15: Final Projects: Policy Datathon

|  |  |  |
| --- | --- | --- |
| Nov 26 | Student Presentations Day 1.  Attendance and participation is required. |  |
| Nov 28 | Student Presentations Day 2.  Attendance and participation is required. |  |

Wk 16: End of term

|  |  |  |
| --- | --- | --- |
| Dec 3 | Project Proposal Workshop (optional)  Attendance and participation is required. |  |
| Dec 5 | Reading Period  Use time to incorporate peer evaluation feedback. |  |

Final written proposal due no later than time of scheduled final exam:

Friday December 7, 2018, 10:50AM. No exceptions.