Winter Institute in Data Science and Big Data
Center for Data Science
Government 496/696

Ryan T. Moore*

7 January 2020 at 16:11

Course Information

Government GOVT 496/696
Winter Institute in Data Science and Big Data
January 2 - 11
Kerwin Hall 3

Instructor Information

Ryan T. Moore, Ph.D.
Associate Professor of Government
Office: Kerwin Hall 226
Telephone: 202.885.6470
Homepage: http://www.ryantmoore.org
Email: rtm (at) american (dot) edu

Simon Heuberger, Ph.D. Candidate
Homepage: http://www.simonheuberger.com
Email: sh6943a (at) american (dot) edu

Course Description

This Institute covers the essential basics for doing data science as practiced in the 21st century. Data scientists are expected to know how to obtain relevant data for a substantive problem, clean and explore data, create and evaluate models using data, state inferences, make reliable predictions, and communicate findings to multiple, possibly non-technical, audiences. We will cover each of these steps in 10 intense working days. The course consists of a dynamic mixture of theoretical

*Department of Government, American University, Kerwin Hall 226, 4400 Massachusetts Avenue NW, Washington DC 20016-8130. tel: 202.885.6470; rtm (at) american (dot) edu; http://www.ryantmoore.org.

lectures, guest speakers, and group assignments. The guest lectures include data science leaders from Washington, DC's unparalleled mixture of government, academia, and business. Statistical topics include exploratory methods, graphics, regression, machine learning, ensembles, network analysis, cluster analysis, text analysis, and Bayesian approaches. Specific technical skills include `R`, `Python`, `RMarkdown`, GitHub, and more.

Learning Objectives

By the end of the course, you should be able to

- Use common computing tools for political data science – applied and scholarly
- Visualize, transform, read, wrangle, tidy, analyze data
- Refresh mathematical foundations for modeling
- Learn modern scientific communication tools
- Learn modern version control
- Describe applications of machine learning and other modern statistical data science methods and computing tools
- Do original research using data science methods

Learning Strategies

Computers and Notes in Class

For most class meetings, we will focus our attention on computational implementations of social scientific techniques. There will often be time in class to pose your specific questions about code. As such, you should bring a laptop to class to try out new code, to update your code files, etc.

Requirements and Evaluation

The course is worth 4 university course credits. For students taking the course for credits, the final grade will be based on attendance, participation, and performance on the group project.

On Saturday, January 4 participants will be assigned to groups to begin work on a real data project using a large dataset. The size of the groups will be 3-6 people depending on the the total number participating in the Institute. The purpose is to get experience working in teams solving a real problem in data science and big data. Each group will present their results on Saturday, January 11.

A summary of the course assessments is in Table 1.

Academic integrity is a core value of institutions of higher learning. It is your responsibility to avoid and report plagiarism, cheating, and dishonesty. Please (re-)read the University policy on academic integrity at http://www.american.edu/academics/integrity/code.cfm, particularly Sections I and II.

Assignment	Weight	Due date
Final presentation	30%	January 11
Final project	50%	January 12
Participation	10%	daily
Attendance	10%	daily

Table 1: Course Assessment Summary

Final Project

For the final project, you will conduct original political science research, present your work, and submit the core components of your work (data, code, and description of the analysis). You may select your own topic.

One possibility is that you may use real data that policymakers want to learn about. In conjunction with The Lab @ DC, a research arm of the Executive Office of the Mayor, we will provide you with a handful of data sets pertaining to policies and programs of Washington, DC. Topics will include campaign finance and expenditures, ANC budgeting, public goods and the 311 request system, transit, and affordable housing. These data are available at http://opendata.dc.gov.

With the data you select, you will pose an appropriate political research question that the data can answer with quantitative methods and analyze the data. As appropriate, you will write a data analysis report, and bundle your analysis, data, and original functions for submission. You will present your research to the class in the last meeting.

Your project should represent original data analysis and code development. It should represent quantitative social science at the highest level you can muster. You will work in a team on the final project. Working collaboratively is typical in political data science research.

Software, Statistics, Data, and Literature Support

The primary software for the course is R, but we will spend time with a variety of other software. We will use the RStudio IDE to help us manage our work in R. See http://j.mp/2swvN0p for help getting started with R and RStudio. A brief overview is also available at http://j.mp/2ELPqFO. We will introduce LaTeX and RMarkdown for scientific communication. See http://j.mp/2LWQfQF for an introduction to using LaTeX through R (via `tinytex`). For an introduction to the fuller version of LaTeX, see http://j.mp/2EO0TEM. We will utilize GitHub for version control. See http://j.mp/2ELRKfV for a brief overview.

Support for statistical software is available through CTRL. See http://j.mp/ZrBr2Z for CTRL's workshop schedule.

The Department of Mathematics and Statistics offers statistical consulting services, with extensive hours. For the schedule and contact information, see http://j.mp/1EmVqkY.

The library itself offers support for various software. Our librarian is Olivia Ivey, whom I recommend reaching out to as you formulate a question, search for data, and try to put your question in a larger intellectual or policy context. You can schedule time with her at oliviaivey.youcanbook.me.

Intellectual Property

Course content is the intellectual property of the instructor or student who created it, and may not be recorded or distributed without consent.

Course Evaluation

The course evaluation will take place in class towards the end of the semester. Please take a few minutes to provide this valuable feedback.

Further Information for American University Students

For further detailed information on the important issues of academic integrity, emergency preparedness, academic support, discrimination, and use of social media, please see here.

Calendar

- Day 1: Thursday, January 2

9:00 - 9:30 Introduction to Data Science, Statement of Objectives, Map of Skills

9:30 - 10:00 Installing `R`, `RStudio`, `Anaconda`, `Python`

10:00 - 12:00 Introducing `R` and the `tidyverse`

12:15 - 1:45 Lunch

1:45 - 2:30 Introducing literate programming with TEX and `RMarkdown`

2:45 - 4:15 Exploratory Data Analysis. Data Wrangling. Cleaning and coarsening data.

- Day 2: Friday, January 3

9:00 - 9:30 Introduction to Today's Goals and Discussion

9:30-11:30 Math Refresher

11:30 - 12 Good Programming Practices

12:15 - 1:45 Lunch

1:45 - 2:15 Defining and Producing Reproducible Research

2:30 - 4:30 Version Control with `git` and GitHub

- Day 3: Saturday, January 4

9:00 - 9:30 Introduction to Today's Goals and Discussion

9:30 - noon Introduction to Statistical Inference: null hypothesis significance testing, samples versus populations, distributions

12:15 - 1:45 Lunch

1:45 - 4:00 Docker: computing in containers (*Hersh Gupta*, DC Department of Human Services and The Lab @ DC)

– Comparative computing: R, Python, Stata, and the shell (*Simon Heuberger*, American University)

- Group project teams

- Day 4: Sunday, January 5
  - Group Meetings to Work on Project

- Day 5: Monday, January 6

9:15 - 9:30 Introduction to Today's Goals and Discussion

9:30 - 12:00 The Linear Regression Model (*Dr. Jeff Gill*, American University)

12:15 - 1:45 Lunch

1:45 - 4:00 Generalized Linear, Additive Models (*Dr. Jeff Gill*, American University)

- Day 6: Tuesday, January 7

9:15 - 9:30 Introduction to Today's Goals and Discussion

9:30 - 10:00 Inference Versus Prediction, Statistical Philosophies and Objectives (*Dr. Jeff Gill*, American University)

10:15 - 12:00 Bayesian Inference and computing, priors, posteriors, MC, MCMC, rejection (*Dr. Jeff Gill*, American University)

12:15 - 1:45 Lunch

2:00 - 3:30 Social Network Analysis (*Dr. Hans Noel*, Georgetown University)

3:45 - 5:00 Data visualization and summarization in `R` and `Python`, Including: `plotly`, `matplotlib`, `seaborn` in `Python` and `ggplot2` in `R` (*Dr. Donna Dietz*, American University)

- Day 7: Wednesday, January 8

9:30 - 10:00 Introduction to Today's Goals and Discussion

10:00 - 10:30 Dimension Reduction, Principal Components Analysis (*David Gerrard*)

10:45 - 12:15 Introduction to Machine Learning: Supervised Learning, Unsupervised Learning

12:15 - 1:30 Lunch

1:30 - 2:45 Images as Data, Convolutional Neural Networks, Autotaggers, Transfer Learning, Computer Vision (*Emily Bello-Pardo*, American University)

3:00 - 3:45 Model Selection and Validation (*Peter Casey*, Catalist)

3:50 - 4:30 Data Science for Political Campaigning (*Peter Casey*, Catalist)

- Day 8: Thursday, January 9

9:00 - 9:30 Introduction to Today's Goals and Discussion

9:30 - 10:45 Using Cloud Computing Resources for Big Data: Code Ocean (*Simon Heuberger*, American University)

10:45 - noon Cluster Analysis: Hierarchical Clustering, Divisive and Agglomerative Clustering, DBSCAN, Kmeans Clustering, K-Nearest Neighbours

12:15 - 1:45 Lunch

1:45 - 2:15 Basics of Text Wrangling in `R`

2:15 - 4:30 Text As Data, Natural Language Processing (*Rebecca Johnson*, Dartmouth University and US Office of Evaluation Sciences)

- Day 9: Friday, January 10

9:00 - 9:15 Introduction to Today's Goals and Discussion

9:15 - 10:15 Important Issues in Data Science: privacy, security, ethics, (*Dr. Jane Wall*, American University)

10:30 - noon AWS, Big Data in `R` (*Jim Harner*)

12:00 - 1:15 Lunch

1:15 - 4:30 Hadoop, Hive, Spark, Machine Learning in `R` (*Jim Harner*)

- Day 10: Saturday, January 11

9:00 - 9:30 Introduction and Discussion

9:30 - noon Presentation of Group Projects

12:15 - 1:45 Lunch

2:00 - 3:00 Conclusions