

Winter Institute in Data Science and Big Data
Center for Data Science
SPA 020/420/620

Ryan T. Moore*

5 January 2026 at 08:30

Course Information

School of Public Affairs SPA 020/420/620
Winter Institute in Data Science and Big Data
15-16 December 2025, 5-10 January 2026
Remote on Zoom

Instructor Information

Ryan T. Moore, Ph.D.
Associate Professor of Government
Office: Kerwin Hall 234
Telephone: +1 202 885 6470
Homepage: <http://www.ryantmoore.org>
Email: `rtm (at) american (dot) edu`
Office Hours: Team work sessions (first half); by appointment before class

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 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, Quarto, SQL, GitHub, and more.

*Department of Government, American University, Kerwin Hall 234, 4400 Massachusetts Avenue NW, Washington DC 20016-8130. tel: +1 202 885 6470; `rtm (at) american (dot) edu`; <http://www.ryantmoore.org>.

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.

In the opening days, 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 the last day of the course.

A summary of the course assessments is in Table 1.

Assignment	Weight	Due date
Final presentation	20%	10 January
Final project	50%	10 January
Final peer evaluation	10%	10 January
Participation	10%	daily
Attendance	10%	daily

Table 1: Course Assessment Summary

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.

Final Project

For the final project, you will conduct original 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.¹

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 L^AT_EX and Quarto for scientific communication. See <http://j.mp/2LWQfQF> for an introduction to using L^AT_EX through R (via `tinytex`). For an introduction to the fuller version of L^AT_EX, 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 course. 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](#).

¹One possibility is that you may use real data that policymakers want to learn about. Topics may 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>.

Calendar

- Day 1: Monday, 15 December

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**
12:00 - 13:00 Lunch break
13:00 - 14:00 Introducing literate programming with **T_EX** and **Quarto**
14:15 - 16:15 Exploratory Data Analysis. Data Wrangling. Cleaning and coarsening data.
16:30 - 17:00 Final project teams. Orientation, posing questions, finding data. Office hours.

- Day 2: Tuesday, 16 December

9:00-11:30 Math/Statistics Refresher
11:30 - 12:00 Introduction to (Linear) Modeling

- Day 3: Monday, 5 January

9:00 - 9:20 Questions, Discussion, Troubleshooting
9:30 - 11:00 Introducing Python: Basics, **NumPy**, and **pandas** (*Ehsan Habibpour*, American University)
11:15 - 12:15 Good Programming Practices
12:15 - 13:25 Lunch break
13:30 - 15:45 Linear Regression Model Theory and Applications (*Dr. Jeff Gill*, American University)
16:00 - 17:00 Basic Text Wrangling in R

- Day 4: Tuesday, 6 January

9:00 - 9:20 Questions, Discussion, Troubleshooting
9:30 - 12:00 Generalized Linear Models, Bayesian Models, Artificial Intelligence (*Dr. Jeff Gill*, American University)
12:00 - 13:00 Lunch break
13:00 - 15:00 Version Control with **git** and GitHub
15:15 - 16:45 Social Network Analysis (*Dr. Hans Noel*, Georgetown University)
16:45 - 17:00 Final project team work

- Day 5: Wednesday, 7 January

9:00 - 9:10 Questions, Discussion, Troubleshooting
9:15 - 10:45 Big Data and Cloud Computing in the Social Sciences (*Dr. Joshua Ferrer*, American University)
11:00 - 12:00 Intermediate Bash: **make** files (*Dr. David Gerard*, American University)
12:00 - 13:00 Lunch break

13:05 - 14:05 Large Language Models for Social Science: Foundations and Applications (*Ali Amini*, American University)
 14:15 - 15:00 Cluster Analysis: Hierarchical Clustering, Divisive and Agglomerative Clustering, DB-SCAN, K-means Clustering
 15:15 - 16:45 Data Visualization Using `ggplot2` in R (*Dr. Donna Dietz*, American University)
 16:45 - 17:00 Final project team work

Reading:

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215, 2001

- Day 6: Thursday, 8 January

9:00 - 9:20 Questions, Discussion, Troubleshooting
 9:30 - 10:45 Introduction to Machine Learning: Supervised Learning, Unsupervised Learning
 11:00 - 12:30 Large Language Models in Social Science Research (*Dr. Jonathan Colner*, American University)
 12:30 - 13:55 Lunch break, final project team work
 14:00 - 16:30 *E pluribus, veritum*: An Introduction to Ensemble Models via Random Forests (*Dr. Ren Massari*, The Lab @ DC)
 16:30 - 17:00 Final project team work

- Day 7: Friday, 9 January

9:00 - 9:10 Questions, Discussion, Troubleshooting
 9:15 - 10:45 Images as Data, Convolutional Neural Networks, Autotaggers, Transfer Learning, Computer Vision (*Dr. Em Bello-Pardo*, Senior Director of Research and Data Science, YouGov Blue)
 11:00 - 12:30 Deploying R-Based Data Solutions on AWS (*Tyler Sanders*, Red Oak Strategic)
 12:30 - 14:10 Lunch break, final project team work
 14:15 - 15:45 Webscraping with `rvest`: Case Studies from Journalism, (*Aarushi Sahejpal*, American University)
 16:00 - 17:00 Final project team work

- Day 8: Saturday, 10 January

9:00 - 9:30 Discussion
 9:30 - noon Presentation of Group Projects
 noon - 15:00 Final project team work
 15:00 Final Report Submission Deadline