PSC 8185 · SPRING 2026

Topics in Empirical and Formal Political Analysis

Monday 3:30–6:00pm · Bell Hall 106

Instructor: Derek Holliday
derek.holliday@gwu.edu
derekholliday.com
Office: Monroe Hall 471
Office Hours: Tuesdays 1-3pm, and by appointment

COURSE DESCRIPTION

Political science (and social science more generally) increasingly requires substantial technical skills to ingest and analyze large quantities of novel data. In this course, we will hone those technical skills through three unique but mutually reinforcing modules: (1) data management, (2) statistical learning, and (3) text-as-data. Our unifying theme will be that of measurement: how can we take noisy, messy inputs and transform them into meaningful indicators of theoretical concepts of interest.

You should expect this course to take 2–2.5 hours of class time a week and an average of 8–12 hours of work per week outside of class.

Learning Outcomes

As a result of completing this course, students will:

- Be able to acquire, combine, and manage large and novel sources of data.
- Be proficient in applying modern statistical learning techniques for supervised or unsupervised modeling in the social sciences, especially with textual data.
- Be prudent methodological consumers when it comes to interpreting these types of models.
- Be able to write a paper or dissertation chapter – using the methods discussed in class – capable of being published in a professional journal.

Course Prerequisites

A course in linear regression (OLS) is required. You should have good working familiarity with the R programming language. While not required, knowledge of maximum likelihood estimation would be beneficial.

Materials

There are 3 required books for this course:

- ISL: James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2023). *An Introduction to Statistical Learning with Applications in R*. Second Edition.
  - MORE ADVANCED: Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Second Edition. This book is optional, and only for those who want to go deeper into the models we talk about. A free version of this book is available online via UPenn/author websites.
- PSM: Lauderdale, Ben (2022). *Pragmatic Social Measurement*.
- TAD: Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*.

Other materials (including several online books) will be linked in the syllabus or posted to Blackboard.

Software

We will be using R and a variety of R packages for analysis (you are welcome to use your preferred IDE, but I recommend RStudio for most folks, or Positron for those looking for a more data-science, general purpose IDE). Relatedly, we may at times utilize Python, time permitting.

General Expectations

- Class attendance and participation are mandatory (see below for specific evaluation criteria).

- Collaboration is encouraged; please talk with each other when completing problem sets and projects. Ultimately, however, each student must write their own answer sets and complete their own paper.
- Laptops are acceptable for use in class. Please only use them to consult with class readings or to take notes. Phone use is prohibited unless for medical or emergency purposes.
- This syllabus is subject to change with prior notification.

Evaluation

Traditional grading approaches, in my experience, are poor fits for graduate-level coursework. Grad school is an exercise in failure; we want to push you to the limits of our collective understanding, expose you to the reality that this is hard, and encourage you to take risks. All of that should involve failure, but traditional grading creates an antagonistic relationship between teacher and student that punishes failure and discourages the sort of mentorship relationship we are trying to cultivate.

Therefore, I will be implementing a form of contract grading. Very simply, I will set expectations on the level of effort I expect from you for each grade level (A, A−, B+, B, etc). Grades below B are at the discretion of the instructor.

What does it take to get a passing grade? The contract below specifies a need to receive passing grades. Generally speaking, to receive a passing grade, you must:

- Submit the assignment on time
- Make a good faith attempt at >90% of the problems/components
- Make <1–2 major/breaking errors, and/or a handful of minor errors

Problem Sets and Replications will be graded as Pass, Bare Pass, and Fail. Bare passes count as passes for grading purposes but should signal a need to refocus your efforts.

	A	A−	B+	B
Participation	You attend and are engaged in all lectures (1 unexcused absence permitted)	You attend and are engaged in all lectures (2 unexcused absences permitted)	You attend and are engaged in all lectures (2 unexcused absences permitted)	You attend all lectures (3 unexcused absences permitted)
Problem Sets	You receive a passing grade on PS1 and any 5 of PS2–7	You receive a passing grade on PS1 and any 4 of PS2–7	You receive a passing grade on PS1 and any 3 of PS2–7	You receive a passing grade on any 3 of PS1–7
Replications	You receive a passing grade on both replication assignments	You receive a passing grade on both replication assignments	You receive a passing grade on either replication assignment	You receive a passing grade on either replication assignment

Cold Calling: I will occasionally be cold-calling (calling upon specific students to answer questions regardless of whether you volunteer). I understand and empathize with the fact that this can be frightening if you have never experienced this sort of teaching. I want to emphasize that the goal of this is not to "catch you out," but rather to foster engaged, active, and equitable participation. Methods can be hard, and it can be easy to be left behind if you are hesitant to ask a question or participate. Therefore, answers to cold-called questions will never be used to formally evaluate you. If you don't know the answer, simply say "I don't know" or "Pass" and I will move on, no judgement.

Problem Sets: Problem sets will be assigned frequently and meant to be completed over the course of roughly 2 weeks. You should complete problem sets in a .qmd file, then knit to a PDF which you submit alongside your .qmd file. Any mathematical notation should be rendered via LaTeX.

Replications: During the semester, I will assign two replication tasks, which are meant to be completed over a longer time horizon (about the length of a single module). Specific instructions will be given in the assignment sheets, but you should generally expect to conduct an independent replication (with some additional analyses) of papers published in political science related to the module topic.

- *Alternative:* I will allow you to replace ONE of the replication exercises with an application of the module's methods to your own work. The total effort should be at least equivalent to what would have been required in the replication exercise. Please notify me several weeks in advance of the replication deadline if you want to opt-in.

Artificial Intelligence Usage

Generative Artificial Intelligence (GAI) technologies continue to be developed, advanced, and adopted in an increasing number of contexts and industries. While the technology undoubtedly has several industry benefits, both potential and realized, its role in the classroom appears conditional. This class is situated uniquely within this context; the primary goal is to prepare you to be active practitioners and interpreters of these methodologies, which involves generating code and understanding the background processes called by such functions. To put it bluntly, this is an area where you can and will be "found out" quickly with only surface-level understandings of the models facilitated primarily by AI usage. Therefore, I have the following expectations regarding AI usage in this class:

- You MAY use AI to:
  - Check spelling, grammar, or identify synonyms
  - Generate ideas and identify sources
  - Rephrase sentences/reorganize writing that you have already drafted
  - Recommend tweaks to outlines you have drafted yourself
  - Generate snippets of example code that you integrate into your problem set answers
- You may NOT use AI to:
  - Write entire sentences/paragraphs without citation
  - USUALLY: generate complete answers to problem set questions, but especially without checking.

University Policies

Academic Integrity Code

Academic integrity is an essential part of the educational process, and all members of the GW community take these matters very seriously. As the instructor of record for this course, my role is to provide clear expectations and uphold them in all assessments. Violations of academic integrity occur when students fail to cite research sources properly, engage in unauthorized collaboration, falsify data, and otherwise violate the Code of Academic Integrity. If you have any questions about whether particular academic practices or resources are permitted, you should ask me for clarification. If you are reported for an academic integrity violation, you should contact Conflict Education and Student Accountability (CESA), formerly known as Student Rights and Responsibilities (SRR), to learn more about your rights and options in the process. Consequences can range from failure of assignment to expulsion from the University and may include a transcript notation. For more information, refer to the CESA website at students.gwu.edu/code-academic-integrity or contact CESA by email (cesa@gwu.edu) or phone (202-994-6757).

Observance of Religious Holidays

Students must notify faculty during the first week of the semester in which they are enrolled in the course, or as early as possible, but no later than three weeks prior to the absence, of their intention to be absent from class on their day(s) of religious observance. If the holiday falls within the first three weeks of class, the student must inform faculty in the first week of the semester. For details and policy, see provost.gwu.edu/policies-procedures-and-guidelines.

Use of Electronic Course Materials and Class Recordings

Students are encouraged to use electronic course materials, including recorded class sessions, for private personal use in connection with their academic program of study. Electronic course materials and recorded class sessions should not be shared or used for non-course related purposes unless express permission has been granted by the instructor. Students who impermissibly share any electronic course materials are subject to discipline under the Student Code of Conduct. Contact the instructor if you have questions regarding what constitutes permissible or impermissible use of electronic course materials and/or recorded class sessions. Contact Disability Support Services at disabilitysupport.gwu.edu if you have questions or need assistance in accessing electronic course materials.

Academic Support

Academic Commons. Academic Commons is the central location for academic support resources for GW students. To schedule a peer tutoring session for a variety of courses visit go.gwu.edu/tutoring. Visit

academiccommons.gwu.edu for study skills tips, finding help with research, and connecting with other campus resources. For questions email academiccommons@gwu.edu.

GW Writing Center. GW Writing Center cultivates confident writers in the University community by facilitating collaborative, critical, and inclusive conversations at all stages of the writing process. Working alongside peer mentors, writers develop strategies to write independently in academic and public settings. Appointments can be booked online at gwu.mywconline.

Support for Students In and Outside the Classroom

Disability Support Services (DSS), 202-994-8250. Any student who may need an accommodation based on the potential impact of a disability should contact Disability Support Services at disabilitysupport.gwu.edu to establish eligibility and to coordinate reasonable accommodations.

Student Health Center, 202-994-5300, 24/7. The Student Health Center (SHC) offers medical, counseling/psychological, and psychiatric services to GW students. More information about the SHC is available at healthcenter.gwu.edu. Students experiencing a medical or mental health emergency on campus should contact GW Emergency Services at 202-994-6111, or off campus at 911.

GW Campus Emergency Information

GW Emergency Services: 202-994-6111. For situation-specific instructions, refer to GW's Emergency Procedures guide.

GW Alert. GW Alert is an emergency notification system that sends alerts to the GW community. GW requests students, faculty, and staff maintain current contact information by logging on to alert.gwu.edu. Alerts are sent via email, text, social media, and other means, including the Guardian app. The Guardian app is a safety app that allows you to communicate quickly with GW Emergency Services, 911, and other resources. Learn more at safety.gwu.edu.

Protective Actions. GW prescribes four protective actions that can be issued by university officials depending on the type of emergency. All GW community members are expected to follow directions according to the specified protective action. The protective actions are Shelter, Evacuate, Secure, and Lockdown (details below). Learn more at safety.gwu.edu/gw-standard-emergency-statuses.

- **Shelter:** Protection from a specific hazard. The hazard could be a tornado, earthquake, hazardous material spill, or other environmental emergency. Specific safety guidance will be shared on a case-by-case basis.
  - *Action:* Follow safety guidance for the hazard.
- **Evacuate:** Need to move people from one location to another. Students and staff should be prepared to follow specific instructions given by first responders and University officials.
  - *Action:* Evacuate to a designated location. Leave belongings behind. Follow additional instructions from first responders.
- **Secure:** Threat or hazard outside of buildings or around campus. Increased security, secured building perimeter, increased situational awareness, and restricted access to entry doors.
  - *Action:* Go inside and stay inside. Activities inside may continue.
- **Lockdown:** Threat or hazard with the potential to impact individuals inside buildings. Room-based protocol that requires locking interior doors, turning off lights, and staying out of sight of corridor windows.
  - *Action:* Locks, lights, out of sight. Consider Run, Hide, Fight.

Classroom emergency lockdown buttons. Some classrooms have been equipped with classroom emergency lockdown buttons. If the button is pushed, GWorld Card access to the room will be disabled, and GW Dispatch will be alerted. The door must be manually closed if it is not closed when the button is pushed. Anyone in the classroom will be able to exit, but no one will be able to get in.

Course Schedule

January 12 — Course Introduction + Basic Skills

INTRODUCTION TO STATISTICAL LEARNING/MEASUREMENT

- ISL Chapters 1 and 2
- PSM Chapters 1 and 2

TIDYVERSE IN R, WRITING CLEAN CODE
- *R for Data Science*, Chapters 1 and 5

VISUALIZING DATA
- *R for Data Science*, Chapter 3

TIDYMODELS
- Get started with tidymodels

January 19 — NO CLASS (MLK Jr. Day)

Due Problem Set 1 due Friday, Jan. 23 by 11:59pm Eastern

Module 1: Data Acquisition, Transformation, and Management

January 26 — Web-Scraping and APIs

WEB-SCRAPING
- *R for Data Science*: Web scraping
- rvest package

APIS
- httr2 package

PRACTICAL GUIDE
- This book is slightly dated, but if you have never worked with HTML, XML, etc, the early chapters are a good introduction for non-programming types (available online through the GWU library):
  - Munzert, Simon, Christian Rubba, Peter Meissner, and Dominic Nyhuis (2014). *Automated Data Collection with R: A Practical Guide to Web-Scraping and Text Mining*. John Wiley & Sons.

February 2 — Database Management, Cloud Computing, and Record Linkage

Due Problem Set 2 due Friday, Feb. 6 by 11:59pm Eastern

SQL
- What is SQL?
  - OPTIONAL: very brief SQL course

APACHE ARROW
- Scaling Up With R and Arrow

DUCKDB
- *DuckDB in Action* Chapters 1–2 (see Blackboard)
- duckdb package
- duckplyr package

OPTIONAL: GWU HIGH-PERFORMANCE COMPUTING
- HPC Onboarding (note: Pegasus access is limited to PIs. If you think you might need HPC access, you will need to be added to a PI's lab)

CLOUD COMPUTING
- Posit Cloud
- AWS S3

RECORD LINKAGE
- Enamorado, Ted, Benjamin Fifield, and Kosuke Imai (2019). "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." *American Political Science Review*.
  - OPTIONAL: more recent developments in the field:
    - Kaufman, Aaron R. and Aja Klevs (2022). "Adaptive Fuzzy String Matching: How to Merge Datasets with Only One (Messy) Identifying Field." *Political Analysis*.
    - Ornstein, Joseph T. (Forthcoming). "Probabilistic Record Linkage Using Pretrained Text Embeddings." *Political Analysis*.

February 9 — OCR and Basic LLM Usage

OCR+

- Colner, Jonathan (forthcoming). "Detecting Formatted Text: Data Collection Using Computer Vision." *Political Analysis.*
- Hegghammer, Thomas (2022). "OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment." *Journal of Computational Social Science.*
  - tesseract package
  - daiR package (for Google Document AI)
  - paws package (for Amazon Textract)

LOCAL LLMS WITH OLLAMA

- Documentation

LLM APIS THROUGH R

- ellmer package
- mall package
- rollama package

USING LLMS PRACTICALLY

- Students get free access to GitHub Pro, which includes Copilot. If you are using Positron as an IDE, you can hook this up to Positron Assistant for AI-generated code snippets.
  - Positron Assistant is also one of many LLM coding "agents," such as Claude Code or Factory's Droid (which can also be installed as an extension to Positron, or called via terminal). This allows you to set more complicated tasks and approve steps along the way.

February 16 — NO CLASS (Presidents' Day)

Due Problem Set 3 due Friday, Feb. 20 by 11:59pm Eastern

Module 2: Statistical Learning and Measurement

February 23 — Supervised Scaling

SUPERVISED SCALING MEASUREMENT

- PSM Chapters 7 and 8
- OPTIONAL: PSM Chapter 9 (if you want to learn about indices)

APPLICATION

- Hopkins, Dan and Hans Noel (2022). "Trump and the shifting meaning of 'conservative': Using activists' pairwise comparisons to measure politicians' perceived ideologies." *American Political Science Review.*

March 2 — Machine Learning I

Due Problem Set 4 due Friday, March 6 by 11:59pm Eastern

RESAMPLING METHODS

- ISL Chapter 5

SHRINKAGE

- ISL Chapter 6 (only through 6.2)

NON-LINEAR MODELS

- ISL Chapter 7

CURRENT POLITICAL SCIENCE LITERATURE

- Cranmer, Skyler J. and Bruce A. Desmarais (2017). "What Can We Learn from Predictive Modeling?" *Political Analysis.*
- Mellon, Jonathan and Christopher Prosser (2025). "Regularized Regression Can Reintroduce Backdoor Confounding: The Case of Mass Polarization." *American Political Science Review.*

March 9 — NO CLASS (Spring Break)

March 16 — Machine Learning II

TREE-BASED MODELS
- ISL Chapter 8

DEEP LEARNING
- ISL Chapter 10

CURRENT POLITICAL SCIENCE LITERATURE
- Montgomery, Jacob M. and Santiago Olivella (2018). "Tree-Based Models for Political Science Data." *American Journal of Political Science.*
- Montgomery, Jacob M. et al. (2015). "An Informed Forensics Approach to Detecting Vote Irregularities." *Political Analysis.*

March 23 — Latent Traits and Scales I

Due Problem Set 5 due Friday, March 27 by 11:59pm Eastern

PCA/FACTOR ANALYSIS
- PSM Chapter 11
- ISL Chapter 12 (12.1-12.2)
- Alternative: Jackman, Simon (2008). Chapter 6: Measurement (Section 6). In *The Oxford Handbook of Political Methodology.*

CLUSTERING
- PSM Chapter 13
- ISL Chapter 12 (12.4)

APPLICATIONS
- Magyar, Zsuzsanna B. (2022). "What Makes Party Systems Different? A Principal Component Analysis of 17 Advanced Democracies 1970–2013." *Political Analysis.*
- Ahlquist, John S. and Christian Breunig (2011). "Model-based Clustering and Typologies in the Social Sciences." *Political Analysis.*

March 30 — Latent Traits and Scales II

Due Replication #1 due Friday, April 3 by 11:59pm Eastern

BASIC IRT MODELS
- PSM Chapter 12
- Jackman, Simon (2008). Chapter 6: Measurement (Section 7). In *The Oxford Handbook of Political Methodology.*

ADVANCED MEASUREMENT MODELS
- Jackman, Simon (2008). Chapter 6: Measurement (Sections 8–10). In *The Oxford Handbook of Political Methodology.*

APPLICATIONS
Read what interests you.
- Clinton, Joshua, Simon Jackman, and Douglas Rivers (2004). "The statistical analysis of roll call data." *American Political Science Review.*
- Tsai, Tsung-han and Chang-chi Lin (2017). "Modeling Guessing Components in the Measurement of Political Knowledge." *Political Analysis.*
- Martin, Andrew D. and Kevin M. Quinn (2002). "Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999." *Political Analysis.*
- Caughey, Devin and Christopher Warshaw (2016). "The dynamics of state policy liberalism, 1936–2014." *American Journal of Political Science.*
- Grumbach, Jacob M. (2023). "Laboratories of Democratic Backsliding." *American Political Science Review.*
- Treier, Shawn and Simon Jackman (2008). "Democracy as a Latent Variable." *American Journal of Political Science.*

Module 3: Text-As-Data

April 6 — Basics of Textual Data

For simplicity, we'll follow TAD pretty linearly. This looks like a lot, but the chapters are quite short. For example, in total chapters 1–9 are about 97 pages. For some light companion code, see Joe Ornstein's website.

WHY TEXTUAL DATA?
- TAD Chapters 1–2

SELECTION AND REPRESENTATION
- TAD Chapters 3–9

COMPUTATION
- Silge, Julia and David Robinson (2017). *Text Mining with R*. O'Reilly Media, Inc. Chapters 1–3, 5
  - MORE ADVANCED: Hvitfeldt, Emil and Julia Silge (2022). *Supervised Machine Learning for Text Analysis in R*. Chapters 2–4

APPLICATIONS

Skim most of these, focusing on your interests.

- Denny, Matthew J. and Arthur Spirling (2018). "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis.*
- Rodriguez, Pedro L. and Arthur Spirling (2022). "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *Journal of Politics.*

April 13 — What Text Can Show Us

Due Problem Set 6 due Friday, April 17 by 11:59pm Eastern

DISCOVERY WITH TEXT
- TAD Chapters 10–14

COMPUTATION
- Silge, Julia and David Robinson (2017). *Text Mining with R*. O'Reilly Media, Inc. Chapter 6
- Hvitfeldt, Emil and Julia Silge (2022). *Supervised Machine Learning for Text Analysis in R*. Chapter 5

APPLICATIONS
- Mosteller, Frederick and David Wallace (1963). "Inference in an Authorship Problem." *Journal of the American Statistical Association.*
- Monroe, Burt, Michael Colaresi, and Kevin Quinn (2008). "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis.*
- Ban, Pamela et al. (2019). "How Newspapers Reveal Political Power." *Political Science Research and Methods.*
- Das, Sanmay et al. (2022). "All (Mayoral) Politics is Local?" *Journal of Politics.*

April 20 — Measurement with Text

MEASUREMENT
- TAD Chapters 15–21
- Hvitfeldt, Emil and Julia Silge (2022). *Supervised Machine Learning for Text Analysis in R*. Section II (Chapters 6–7)

APPLICATIONS
- King, Gary, Jennifer Pan, and Margaret E. Roberts (2013). "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review.*
- Grimmer, Justin (2010). "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis.*
- Parthasarathy, Ramya, Vijayendra Rao, and Nethra Palaniswamy (2019). "Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies." *American Political Science Review.*
- Ying, Luwei, Jacob M. Montgomery, and Brandon M. Stewart (2022). "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." *Political Analysis.*

April 27 — Inference with Text + LLMs

INFERENCE
- TAD Chapters 22–27

DEEP LEARNING/LLMS/AI
- Hvitfeldt, Emil and Julia Silge (2022). *Supervised Machine Learning for Text Analysis in R*. Section III (Chapters 8–10)

APPLICATIONS
- Egami, Naoki et al. (2022). "How to make causal inferences using texts." *Science Advances*.
- Fong, Christian and Justin Grimmer (2023). "Causal Inference with Latent Treatments." *American Journal of Political Science*.
- Gilardi, Fabrizio, Meysam Alizadeh, and Mael Kubli (2023). "ChatGPT outperforms crowd workers for text-annotation tasks." *Proceedings of the National Academy of Sciences*.
- Bisbee, James et al. (2024). "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis*.
- You should really just skim the First View section of *Political Analysis* for the most up-to-date LLM publications… they are coming out so quickly that this section will be outdated by the time we get to it.

April 29 (Designated Monday) — More Non-Tabular Data (Images and Audio)

Due Problem Set 7 due Friday, May 1 by 11:59pm Eastern

Due Replication #2 due Friday, May 8 by 11:59pm Eastern

This is all quite advanced and requires computational power (and technical prowess) that I can't fairly expect you to have. My goal is more to expose you to frontier models in the field and give you a place to start should you want to pursue this further.

IMAGES
- Basic Image Classification (with Tensorflow/Keras in R)
- Torres, Michelle and Francisco Cantu (2022). "Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data."

AUDIO
- voice package (under development)
- For most speech-to-text applications, I would suggest using an AI API or installing a model locally and running something like the audio.whisper package.
- Knox, Dean and Christopher Lucas (2021). "A Dynamic Model of Speech for the Social Sciences." *American Political Science Review*.
- Dietrich, Bryce J., Matthew Hayes, and Diana Z. O'Brien (2019). "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." *American Political Science Review*.