Gov 51: Project Guide

Finding Your Question, Finding Your Data

Scott Cunningham · Harvard University · Spring 2026

> **Milestone 1 is due Thursday, March 26, 2026.** You need three things: a research question, a dataset, and a 1–2 page proposal. This guide walks you through how to get there.

1 What Is Milestone 1?

For Milestone 1, you need to:

1. **Pick a research question** and classify it as descriptive, predictive, or causal.
2. **Identify a dataset** that can help you answer it (or explain how you plan to collect one).
3. **Write a 1–2 page proposal** laying this out clearly.

That's it. You're not analyzing data yet. You're setting up the project so that when you start the analysis, you know what you're doing and why.

1.1 Submission details

- **Format:** PDF, uploaded to Canvas by 11:59 pm on March 26.
- **Length:** 1–2 pages (no more). Single-spaced is fine.
- **Individual work.** This is not a group assignment.
- **Do not attach the dataset**—just describe it.
- **Grading:** Milestone 1 is graded on clarity, specificity, and feasibility. We are not evaluating your results (you don't have any yet). We're checking that you have a well-defined question and a realistic plan.

2 The Three Types of Studies

Every empirical project falls into one of three categories. Knowing which one you're in determines what methods you use, what counts as success, and how you interpret your results.

2.1 Descriptive

A descriptive study answers: **What does the world look like?**

You're measuring, summarizing, or documenting a pattern. You're not predicting what will happen next, and you're not claiming that one thing causes another. You're just saying: here's what the data show.

- How has the racial wealth gap in the U.S. changed since 1990?
- What topics dominate congressional floor speeches about immigration?
- How do commute times vary across metropolitan areas?

- What share of criminal defendants in Broward County are rearrested within two years?

Descriptive work is underrated. Good description is hard, and it's often the foundation for everything else. Card et al.'s analysis of immigration speeches that we read early in the semester is fundamentally a descriptive project—they measured the tone and content of political rhetoric. That measurement is valuable on its own.

Methods you'd use: Summary statistics, cross-tabulations, visualizations, text analysis, geographic mapping.

2.2 Predictive

A predictive study answers: **Can we forecast an outcome?**

You're building a model that takes inputs and produces a guess about something you haven't observed yet. You don't need to understand *why* the prediction works—you just need it to work well out of sample.

- Can we predict which defendants will be rearrested? (This is what COMPAS tries to do.)
- Can we forecast which countries will experience civil war in the next five years?
- Can we predict election outcomes from polling data?
- Can we identify fraudulent insurance claims from claim characteristics?

The key metric for predictive work is **out-of-sample performance**—RMSE, classification accuracy, or whatever measure fits your outcome. An overfit model that looks great on training data but fails on new data is worthless.

Methods you'd use: Train/test splits, cross-validation, LASSO, Ridge, Elastic Net.

2.3 Causal

A causal study answers: **Does X cause Y?**

> **Important:** We haven't covered causal inference methods yet—they're coming in Weeks 10–12. If you choose a causal question for Milestone 1, you are **not** expected to have an identification strategy. Just state the question and the data. We will work on the identification strategy together once you've learned the tools.

This is the hardest type of study to do well, because you need to isolate the effect of one variable while holding everything else constant. In a perfect world, you'd run a randomized experiment. In the real world, you usually can't, so you need a credible research design. We'll build those skills together.

- Does increasing the minimum wage reduce employment?
- Does body camera adoption reduce police use of force?
- Does access to early childhood education improve long-term earnings?
- Did a specific policy change affect voter turnout?

Methods you'll learn: Randomized experiments, difference-in-differences, instrumental variables, regression discontinuity design.

3 How to Come Up With a Question

This is the part where most students get stuck. Here's what I've found works.

3.1 Start with what bugs you

What's something you've read about, argued about, or wondered about? The best projects come from genuine curiosity. If you're going to spend weeks on this, pick something you actually care about.

- You're frustrated about housing costs → "How have rents changed relative to wages in major cities?"
- You follow criminal justice debates → "Can we predict recidivism better than COMPAS?"
- You're interested in media → "Has newspaper coverage of climate change increased over the past decade?"

3.2 Read the news with a social scientist's eye

Every news article has an implicit empirical claim. When you read that "crime is surging," ask: surging compared to what? Compared to last year? Compared to the 1990s? Compared to what we'd expect given economic conditions? That "compared to what?" is where research questions live.

3.3 Think about "compared to what?"

Every good question has an implicit comparison:

- "Is the racial wealth gap large?" → compared to what baseline?
- "Does education reduce crime?" → compared to not being educated?
- "Are commute times getting longer?" → compared to when?

Making the comparison explicit helps you figure out what data you need.

3.4 Narrow it down

The most common mistake is picking a question that's too big. "Inequality" is not a research question. "Has the Gini coefficient in U.S. counties changed differently in states that expanded Medicaid versus states that didn't?" is a research question. You should be able to state your question in one sentence.

3.5 Don't be afraid to start with the data

Sometimes the best approach is backwards: find an interesting dataset first, then ask what questions it can answer. There's no shame in this. Some of the best research in social science started with someone stumbling across an interesting dataset and asking what it could tell us.

4 Where to Find Data

You don't need to use all of these. Find one that fits your question.

4.1 Large Survey and Census Data

IPUMS (ipums.org)—The gold standard for U.S. microdata. Census, American Community Survey, Current Population Survey, and more. We used ACS data from IPUMS for PS1. If your question involves demographics, income, employment, housing, or migration, start here.

ANES (electionstudies.org)—The American National Election Studies. Surveys of American voters going back to 1948. Political attitudes, voting behavior, partisan identification, public opinion.

GSS (gss.norc.org)—The General Social Survey. Social attitudes, demographics, and behavior since 1972. Great for questions about social change over time.

Pew Research Center (pewresearch.org)—Extensive surveys on politics, media, technology, and social trends. Many datasets are publicly available.

Cooperative Election Study (cces.gov.harvard.edu)—Massive election survey (60,000+ respondents per wave). Voting behavior, policy attitudes, demographics. Widely used in political science. Free.

4.2 Political Science and Conflict Data

V-Dem (v-dem.net)—Varieties of Democracy. The standard cross-national dataset on democracy, covering 200+ countries with hundreds of indicators. If your question involves democratic institutions, regime change, or governance, start here.

Voteview (voteview.com)—Congressional roll-call voting data from the 1st Congress to the present. Ideal for questions about polarization, party loyalty, or legislative behavior.

MIT Election Data + Science Lab (electionlab.mit.edu)—Cleaned U.S. election returns at the state, county, and precinct level. If your question involves election outcomes, this is the cleanest source.

Correlates of War (correlatesofwar.org)—Interstate and intrastate conflict data going back to 1816. Standard dataset for international relations research.

4.3 Social Science Archives

ICPSR (icpsr.umich.edu)—The Inter-university Consortium for Political and Social Research. Over 16,000 studies. Harvard gives you access.

NACJD (icpsr.umich.edu/NACJD)—The National Archive of Criminal Justice Data, housed within ICPSR. Crime, policing, courts, corrections, victimization.

Harvard Dataverse (dataverse.harvard.edu)—Replication data for published papers. Especially useful if you want to extend existing research.

4.4 Government and International Data

Data.gov (data.gov)—Federal open data. EPA air quality, USDA crop data, HUD housing statistics.

World Bank (data.worldbank.org)—International development indicators for virtually every country, going back decades.

BLS (bls.gov)—Employment, wages, prices, productivity. The labor market and inflation.

FBI Crime Data Explorer (crime-data-explorer.fr.cloud.gov)—Crime statistics from the Uniform Crime Reporting program.

4.5 Journalism and Community Datasets

ProPublica Data Store (propublica.org/datastore)—Investigative journalism datasets. The COMPAS data we use in PS3 came from here.

FiveThirtyEight (github.com/fivethirtyeight/data)—Clean, well-documented datasets on politics, sports, and culture.

Kaggle (kaggle.com/datasets)—Community-curated datasets on almost every topic. **Use with caution:** many Kaggle datasets have undocumented provenance, missing codebooks, and unclear sampling frames. A Kaggle dataset is acceptable only if you can identify and document the original source of the data.

4.6 Replication Data from Published Studies

Some of the best datasets come from published research. You can use these for your project—**but you must ask a new question, not just replicate the original paper.** The data is a starting point; the question has to be yours.

> **Opportunity Insights** (opportunityinsights.org/data)—Raj Chetty's team publishes tract-level data on income mobility, college attendance, incarceration rates, patent rates, and more—broken down by parental income, race, and gender. Free CSV downloads, no registration. This is probably the richest and most accessible dataset on this list.
> *Example questions:* "Do neighborhoods that produce high-earning adults also produce high patent rates?" or "How does income mobility differ between college towns and non-college towns?"

Abramitzky, Boustan, Jácome & Pérez—Replication data for "Intergenerational Mobility of Immigrants in the US over Two Centuries" (*AER*, 2021). Millions of father-son pairs from historical censuses. Free on openICPSR (project 120490). This connects directly to the Card et al. immigration work we studied.

AEA Replication Packages (openicpsr.org)—The American Economic Association requires authors to post data and code for published papers. Thousands of datasets across every topic in economics—all free.

Important rule: If you use data from a replication package, your project must ask a question the original authors did not answer. You're not replicating—you're using their data to investigate something new.

5 Collecting Your Own Data

You're not limited to existing datasets.

Fair warning: All three approaches below require significantly more time and technical skill than using an existing dataset. If you're considering any of them, talk to me *before* Milestone 1 so we can assess feasibility.

Web scraping. If the data you want exists online but isn't downloadable, you can scrape it. Python's `BeautifulSoup` or R's `rvest` make this feasible. Be respectful: check the site's `robots.txt`, don't overwhelm their servers, and be aware of terms of service. Budget more time than you think—JavaScript-rendered pages and rate limits can slow you down.

Text classification with LLMs. If you have text data and want to classify it (sentiment, topic, stance), large language models can serve as zero-shot classifiers. This is cutting-edge methodology—published papers in top journals use this approach. Come talk to me if you're interested; I'll help you get started, but expect a learning curve.

Surveys. You can design and administer your own survey using Qualtrics (Harvard has a license) or Google Forms. **Important:** surveying human subjects requires IRB approval, which can take weeks at Harvard. If you're considering a survey, talk to me immediately—the timeline is tight.

6 Practical Tips

6.1 Check that your data actually works

Before you commit to a dataset, load it into R and make sure:

- You can read the file without errors
- The key variables you need are present and populated
- There are enough observations for meaningful analysis (at least a few hundred; more is better)
- There's enough variation in your outcome and key predictors

There's nothing worse than writing a proposal around a dataset you can't actually use.

6.2 Documentation matters

When you download data, save the codebook or documentation alongside the data file. Future-you will thank present-you when you're trying to remember what `EMPSTAT = 3` means.

6.3 Think about the final product

Your final report will include descriptive statistics, visualizations, and some form of statistical analysis. As you pick your question and data, think about whether you can do something interesting with the tools we've learned.

6.4 Think about ethics

Some of the most interesting research questions involve sensitive topics—criminal justice, health, race, immigration. If your project touches on these areas, your final report should acknowledge the ethical dimensions of the work. Who could be helped or harmed by these findings? What are the limitations of using data to make predictions about people? You saw this tension in the COMPAS data: a prediction algorithm can be accurate on average while being systematically unfair to particular groups. Good empirical work isn't just technically correct—it's honest about what the numbers can and cannot tell us.

6.5 It doesn't need to be groundbreaking

You're not writing a dissertation. A well-executed descriptive analysis of an interesting question is a great project. A clean prediction exercise that compares models on a real dataset is a great project. Don't feel pressure to solve a major social problem—just do careful, honest empirical work on something that interests you.

7 What to Submit for Milestone 1

Your proposal should be 1–2 pages and include:

1. **Research question:** One clear sentence stating what you want to know.
2. **Study type:** Is this descriptive, predictive, or causal? Why?
3. **Data:** What dataset will you use? Where does it come from? How many observations does it have? What are the key variables?
4. **Brief plan:** In 2–3 sentences, describe what your analysis will look like. What will you show? What methods will you use?

If you're still deciding between two questions, it's fine to submit both and ask for feedback. That's what Milestone 1 is for.

> The project is your chance to do something with the tools we've been building all semester. Every problem set has been practice. The project is the game. Pick something you'll enjoy working on. The best projects aren't the most ambitious—they're the ones where the student was genuinely curious and did careful work.
>
> If you're stuck, come to office hours. We'll figure it out together.