Data!

**Find some data**

Individual-level (survey) data

* GSS and ANES
* ISSP
* \*barometers
* World Values Survey
* CU Library
* World Bank/EBRD enterprise surveys (BEEPS)
* Pew, Annenberg, Gallup

Observational data

* ICPSR
* World Bank, LSMS
* QoG
* NYC OpenData
* Database of Political Institutions
* CU Library
* US Census Factfinder
* Voteview

Replication data

* Harvard Dataverse
* Acemoglu, Johnson, and Robinson 2001
* Easterly and Levine; Fearon
* Ask for it!

**Build your own dataset**

* Scrape websites
* Use APIs: Twitter, New York Times
* Historical data coding
* Google Ngram, Correlate, Search data

R packages: XML, rvest

Python: scrapy, beautifulsoup

**Create your own data** (IRB!)

* Survey individuals with Mechanical Turk
* Lab experiments (Salvo Nunnari)
* Survey experiments
* Qualitative data (+BIQQ)

**Cleaning and Manipulating Your Data**

Excel (or OpenOffice/LibreOffice)

Stata: reshape, collapse, carryforward, egen, merge, sort

R packages: foreign, xlsx, plyr, memisc, reshape2

**Look at Your Data**

Plot it, look at cross-tabs, histograms

Check for coding errors

Look at raw correlations

Identify and double-check outliers

**Presenting Your Data** (but don’t worry about this until later!)

Stata: outreg2

R packages: stargazer, xtable, coefplot, maptools

Do:

* *Think creatively about what sorts of data can be used to make your arguments*. As pointed out in KKV, political science presumes a certain amount of generalizability and abstraction to arguments and theories. One consequence of that is that there could be interesting ways of showing the underlying elements of your arguments that go beyond the specific cases or manifestations that you are primarily interested in. Think: “Okay, so if my argument works in countries A and B like I think it does, maybe it would also work in US State X and Y or in India or in the 1800s.”
* *Use quantitative data to support premises and facets of your arguments*. Remember that your reader doesn’t have the same background knowledge in this topic that you do. To the extent that you can show rather than tell readers that Important Fact #1 or Crucial Assumption #2 hold in the real world, the better they will be able to follow you as you move beyond that.
* *Consider demonstrating the logic of (or mechanisms behind) your arguments in more than one way or with more than one source/type of data*. Showing that your arguments hold at multiple levels or with disparate kinds of data can be very powerful.
* *Google a lot.* There are lots of nice and simple tutorials out there. There’s probably also someone who had the same question as you or, if you’re lucky, who has already solved your problem with some software that is now available.
* *Ask us for help or suggestions.* Even with mundane stuff like cleaning or organizing data, it can be very helpful for you to ask for suggestions – especially if you’re facing a task that you think will be long and tedious.

Don’t:

* *Lose track of the research question you are really interested in*. Of course you should make sure your question is one that you can answer satisfactorily using evidence you will be able to get. But that doesn’t mean you should prematurely abandon or mangle your research because you’re concerned about finding data. There are many creative solutions to these sorts of problems.
* *Lose sight of the forest for the trees*. Having lots of nice data is great, but focus on honing your arguments first. The data can always come later. You can always get additional or better variables once you’ve got a good foundation that supports your arguments.
* *Let yourself get bogged down in searching for and working with data*.