Research Design, Data, and Intro to R

Elena Barham

Columbia University

October 5, 2020

Roadmap

1. Data Gathering
   - Measurement
   - Sensitive items
   - Data Structure
2. Data Analysis
   - Coding Basics
   - Visualization and Data Exploration
   - Regression
3. Reporting
   - Graphs
   - Tables
4. Questions and Recap

Starting from a hypothesis

To start research design:

- Dependent variable(s)
- Independent variable(s) - variables of theoretical interest
- Units of analysis

Some other questions to ask yourself:

- What is my mechanism?
- What are the competing explanations?

Making decisions as a researcher

What we are looking for:

- Clear mapping between theory and research design.
- Strong identification strategy (if we aim to make causal claims).
- Justified and well conceptualized measurements.
- Attuned to the strengths and weaknesses of our designs.
- Doable!

Operationalizing Variables

Often our hypotheses involve relatively abstract concepts. For example, we might ask: "What is the effect of *natural resource wealth* on *corruption*?

- Conceptualization: defining the core concepts, i.e. What do we mean by corruption?
- Operationalization: developing a research procedure to select and classify *empirical representations* of a concept.
- Measurement: implementing our operationalization strategy.

Good Measures

- **Validity:** Does this measure what it claims to measure?
- **Reliability:** Is this measure consistent?
  - If I measure the same thing multiple times, does the measure yield the same results?
  - Concerns about *objectivity* of measures can often fall under this category of measurement issues.

Good Measures

- **Usability and interpretability:** Is this measure intelligible to my targeted reader? Are results from this measure clear to interpretations?
  - Units?
  - Subject to multiple interpretations (ratios?)
- **Economy:** Is this measurement strategy achievable given my other constraints?

Good Measures - Exercise

Claim: "Oil and other valuable natural resources make states more *corrupt*". How do we measure corruption?

- **Measurement option 1:** Survey data of individuals living under the governments in question on their perceptions of institutional quality and experiences of corruption (i.e. AfroBarometer, AmericasBarometer).
- **Measurement option 2:** External expert coding on their perceptions of institutional quality (i.e. TI's Corruption Perceptions Index).
- **Measurement option 3:** Measures of anti-corruption initiatives joined and legislation passed (i.e. membership in the Extractive Industries Transparency Initiative).

Measuring Sensitive Concepts

- List experiments
- Less sensitive downstream effects
- Expert opinions
- Interviews & trust of respondents

No great options

Discussion: What are some strategies when we are not satisfied with existing measures but we are not able to collect original data?

No great options

Discussion: What are some strategies when we are not satisfied with existing measures?

- Validate across multiple measures
- Create indices of prior measures
- Contextualize our findings given weaknesses of measures

Last Word on Measures

Measures don't need to be perfect, but we need to understand the **trade-offs** and **limitations** involved in our measurement strategies.

- Sometimes data driven.
- Sometimes an improvement in measurement is possible without sacrificing other aspects of the project.
- Sometimes, the measure is the goal.

Common Data Sources

Exercise: In-class brainstorm!

Collecting Data

Some approaches to collecting data:

- Novel survey
- Web scraping
- Quantifying historical or qualitative data
- Aggregating data from disparate sources
- Requesting official data that hasn't been used for research but you suspect exists
- Other?

Example - Data Collection

Research Question: Do distributive political concerns help explain patterns of territorial conservation in Latin America?

- *Measure*: We chose **protected areas** as our unit of analysis, which were the strictest level of conservation and relatively comparable across countries.

Example - Data Collection

Research Question: Do distributive political concerns help explain patterns of territorial conservation in Latin America?

- *Pre-existing data*: We found that the World Database on Protected Areas (WDPA) has a comprehensive list of global protected areas. Great! However, they did not have very much information about these territories...Alas!

Example - Data Collection

Research Question: Do distributive political concerns help explain patterns of territorial conservation in Latin America?

- *New data*: We compiled comprehensive data on who owns and administers each protected area.
  - We wrote to individual environmental ministries to get their data.
  - We reached out to sub-national associations where environmental ministries failed.
  - For the residual protected areas, we tracked down their creation laws and individually coded based on these documents.

Panel Design

list of packages.R | conservation

Filter

NAME	GOV_TYPE	MANG_AUTH	STATUS	STATUS_YR	REP_M_AREA	GIS_M_AREA
Abayuva	Local communities	Not Reported	Designated	1995	0	*NA*
Abra del Acay	Sub-national ministry or agency	Not Reported	Designated	1995	0	0.000000e+00
Acambuco	Sub-national ministry or agency	Not Reported	Designated	2004	0	0.000000e+00
Aconquija	Federal or national ministry or agency	Administración de Parques ...	Designated	2018	0	0.000000e+00
Aconquija	Sub-national ministry or agency	Administración de Parques ...	Designated	2018	0	0.000000e+00
Aconquija	Sub-national ministry or agency	Not Reported	Designated	1936	0	*NA*
Afloramiento Limoso	Sub-national ministry or agency	Not Reported	Designated	2002	0	0.000000e+00
Agua Dulce	Sub-national ministry or agency	Not Reported	Designated	1970	0	0.000000e+00
Aguaray-mí	Private Landowners	Not Reported	Designated	1988	0	0.000000e+00
Aguas Chiquitas	Sub-national ministry or agency	Not Reported	Designated	1982	0	*NA*
Alejandro Orloff Salti...	Sub-national ministry or agency	Not Reported	Designated	1997	0	0.000000e+00
Alto Andina de la Chi...	Sub-national ministry or agency	Not Reported	Designated	1992	0	0.000000e+00
Andino Norpatagónica	Collaborative governance	Not Reported	Designated	2007	0	0.000000e+00
Andres Giai	Private Landowners	Not Reported	Designated	1997	0	0.000000e+00
Angastaco	Sub-national ministry or agency	Not Reported	Designated	1995	0	0.000000e+00
Apipé Grande	Sub-national ministry or agency	Not Reported	Designated	1994	0	0.000000e+00
Araucaria	Sub-national ministry or agency	Not Reported	Designated	1991	0	0.000000e+00
Arboretum L.N. Alem	Collaborative governance	Not Reported	Designated	1994	0	*NA*
Arroyo Ayui Grande	Private Landowners	Not Reported	Designated	2000	0	*NA*
Arroyo El Durazno	Private Landowners	Not Reported	Designated	2011	0	0.000000e+00

Showing 1 to 21 of 7,792 entries, 35 total columns

Missing Data

- Triage across sources.
- Interpolation - strategies vary.
- Work with complete data and theorize effect for results, external validity. What do we lose by dropping missing observations? Does this induce bias, or alter our confidence in our estimates?
- Zoom out to a higher level of aggregation.

Exercise

Try to write out the variable headings of a data set that could help you answer your question or test your hypothesis.

- What are your units of analysis?
- What information do you need about each unit?
- What is the scope of the data set (time periods, which units are in it?)

Intro to R - Packages

What are packages?

- Loadable software within R or RStudio
- Increase and broaden functionality of R
- Make it easier to do things that you could do in base R – package mans someone else has gone to the trouble of writing the command for you!

How to know what package you need?

- Usually we just look it up :)
- Generally a process of trial and error

Tidyverse

Tidyverse is a suite of R packages that can make your life a lot easier. Functionality includes:

- ggplot2 - functionality to make neat graphics
- dplyr - data manipulation for common challenges
- stringr - intuitive tools to work with strings (i.e. text variables)

And more!

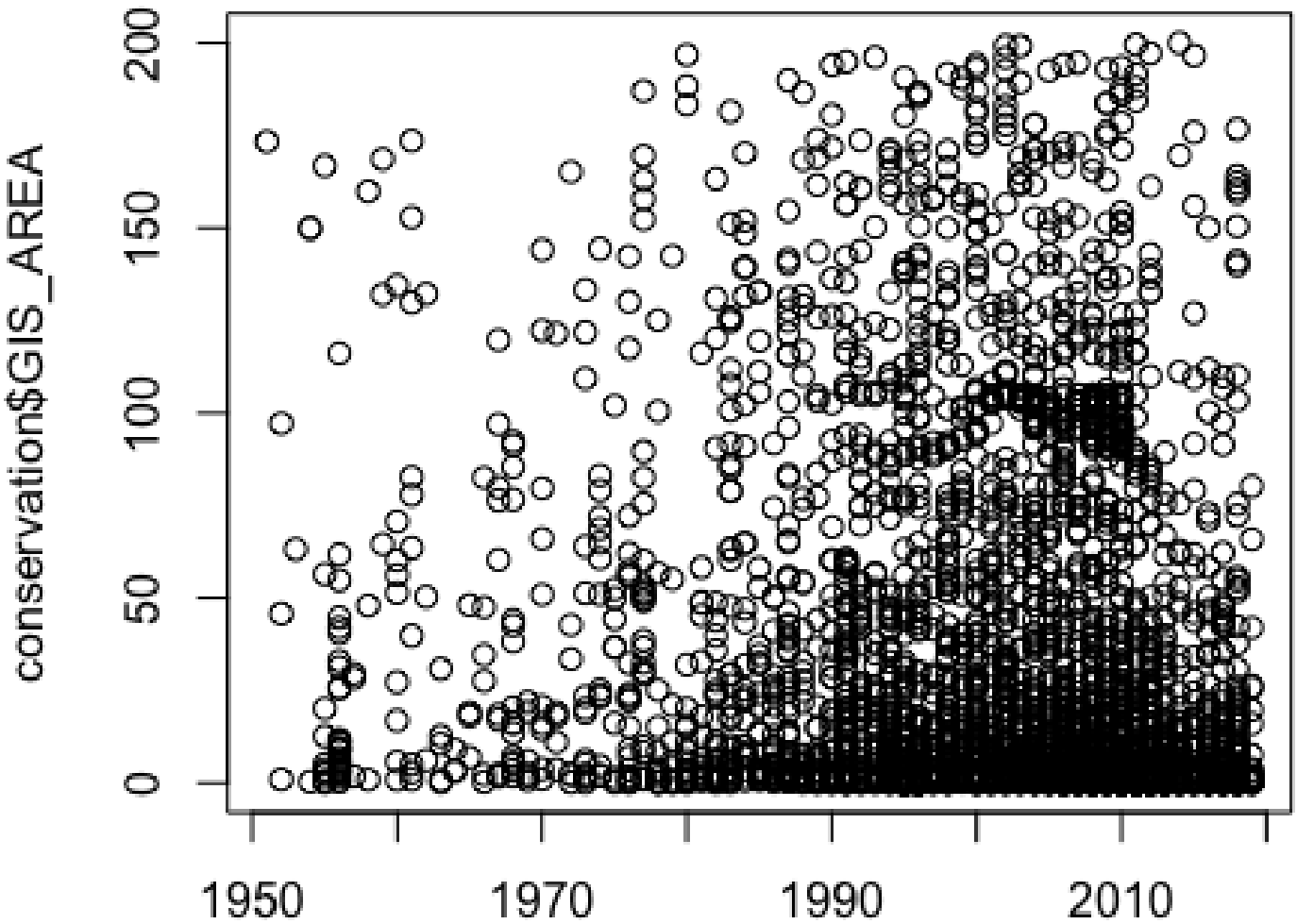
Loading Data

Some possible formats:

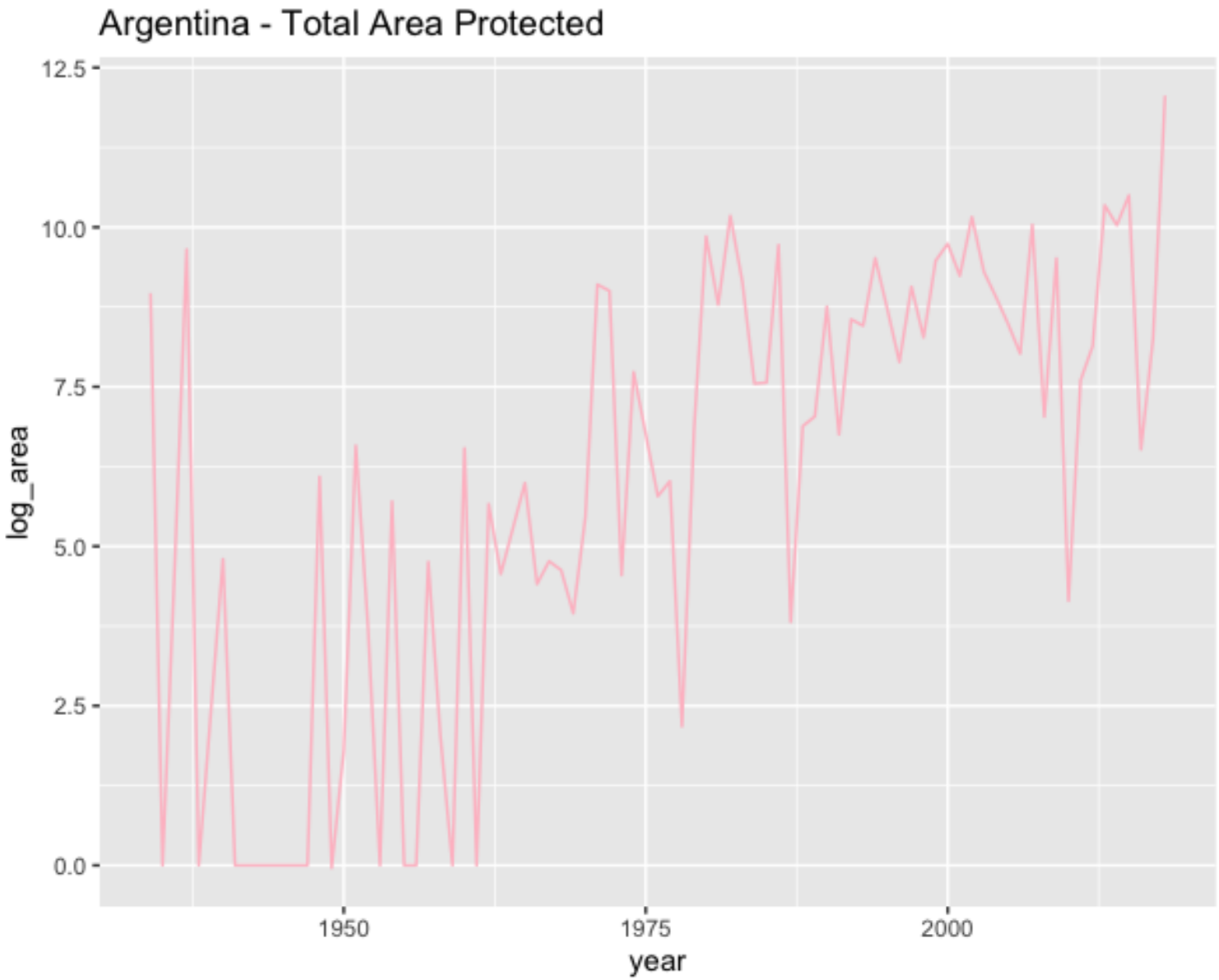
- .csv
- .xlsx
- .Rda
- .dta
- shapefiles

R can import many file types although some require packages to load.

Data Visualization

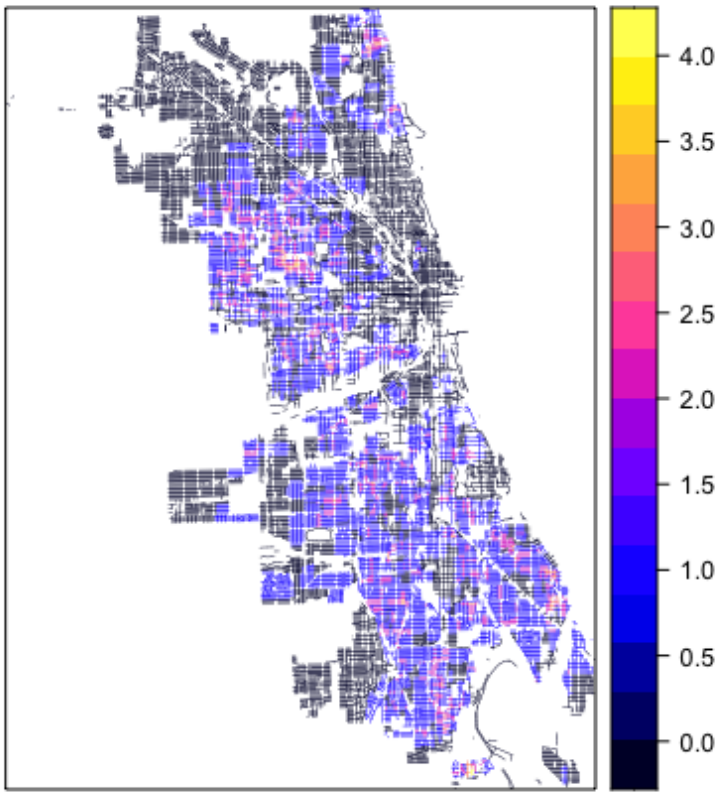


Data Visualization



Spatial Visualization

Figure 1: Gang Density Per Block in Chicago (2015)



Why visualize spatially?

- Are there spatial patterns to the data?
- Are units independent (r.e. potential outcomes framework)?
- Are there potential spillovers?

Regression Basics

$y_i = \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$

In R:

- lm(y $\sim x_1 + x_2$, data = data)

Other packages for regression:

- glm - generalized linear models
- experiment - experimental design and analysis
- ivmodel, ivtools - for instrumental variable analysis

Fixed Effects Models

Fixed effects: Accounts for *time-invariant* traits of units not captured by controls

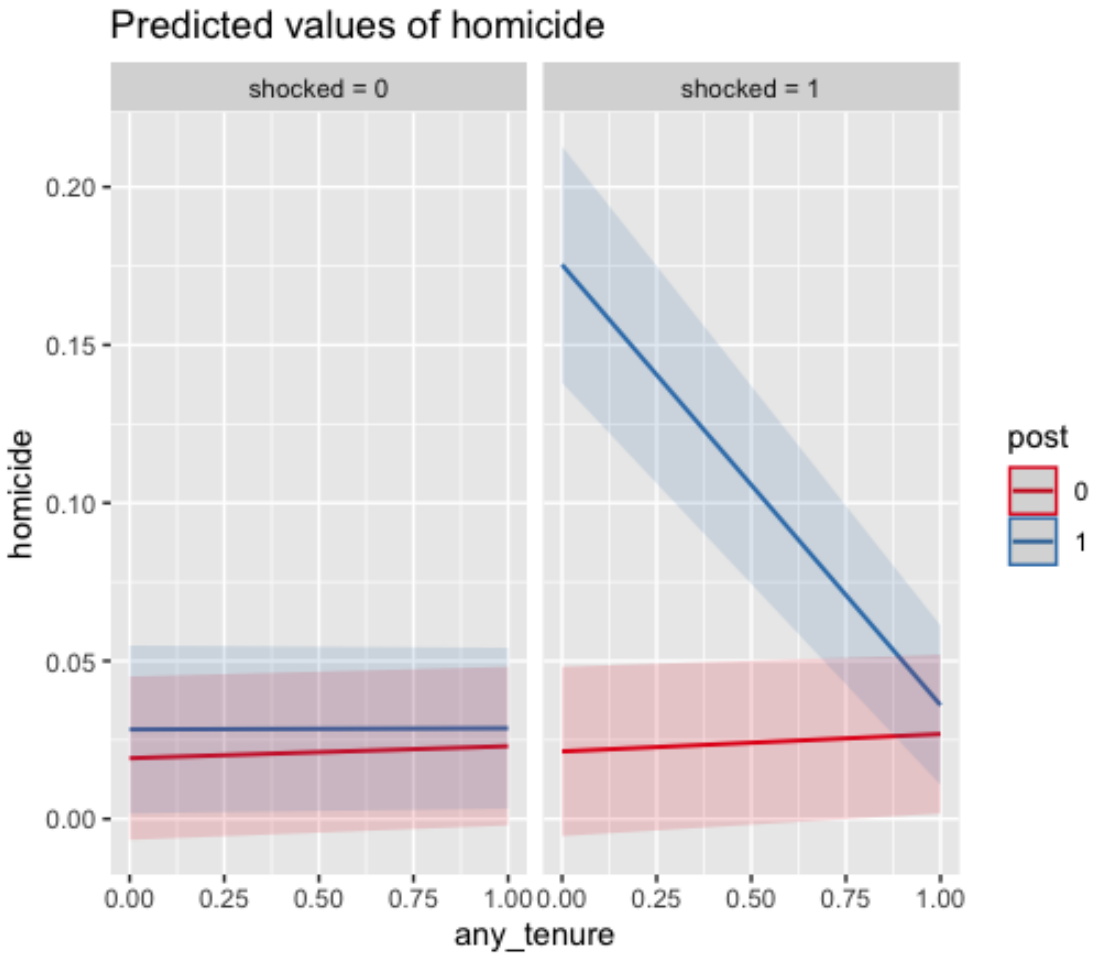
- Compares variation within unit of FEs
- Can help with omitted variable bias
- Level of FEs is a *theoretical decision*
- Multiple FEs may be justified

Interaction Terms

Interaction terms: These are useful when you think the effect of one variable depends on the value of another variable

- Differences-in-differences frameworks
- Heterogeneous treatment effects
- Different to mediation analysis

Displaying Results



Regression Tables

Stargazer can help make readable and nice tables in LaTeX!

Table 1: Shocks increase violence in peripheral turf

	Dependent variable:
	homicide
shocked*post	0.006**
	(0.002)
tot_gangs	0.002*
	(0.001)
shocked	0.003**
	(0.002)
post	0.004**
	(0.002)
Constant	0.010
	(0.014)
Observations	69,492
Regional & Seasonal FEs	Yes
Block controls	Yes
R^2	0.004
Adjusted R^2	0.003
Note:	*p<0.1; **p<0.05; ***p<0.01

Questions?

Questions?

Research and Epistemology

- Positivism
  - Core belief that concepts (like democracy) exist independently of inquiry and that social scientists' job is to measure and study these concepts to understand patterns within and across cases. Objective causes and effects exist.
  - Experimental and quasi-experimental methods, focused on identifying causal effects
  - Descriptive, focused on identifying trends, associations, and relationships among theoretical variables of interest
  - Goal of absolute knowledge that accumulates to demonstrate objective truths

Research and Epistemology

- Interpretivism
  - Core belief that concepts (like democracy) are socially embedded in the context of social scientists who study it, and whose goal is to clarify the way these concepts are used, meant, and produced in their historical and political contexts.
  - Descriptive, focused on conceptualization of specific, unique, and relative meanings
  - Goal of generating knowledge (relative to context, time, culture, etc) to describe subjective truths

Research Approaches for Positivists

What do we mean by causal identification?

- Potentially "causally identified":
  - Randomized controlled trials (RCTs)
  - Quasi-experimental designs (differences-in-differences, regression discontinuity, matched comparisons/synthetic controls, instrumental variables analysis)
- Non-identified but still positivist:
  - Correlation and regression
  - Qualitative methods with causal claims (ethnography, interview-based methods, case studies)

Research Approaches for Interpretivists

- Quantitative approaches
  - Digital/automated text analysis
- Qualitative approaches
  - Ethnography/participant observation (in-person and digital)
  - Discourse analysis
  - Case-study analysis
  - Natural language interviews
  - Analysis of subjectivity and positionality (including of the researcher)

Internal and External Validity

- **Internal validity:** Does our the evidence produced by our design clearly support a claim of cause and effect?
- **External validity:** To what extent do we expect these findings to generalize to other groups or cases?