Quantitative Social Science in the Age of Big Data and AI

Lecture 0: Introduction(updated version)

Zhaopeng Qu
Hopkins-Nanjing Center
February 26 2025

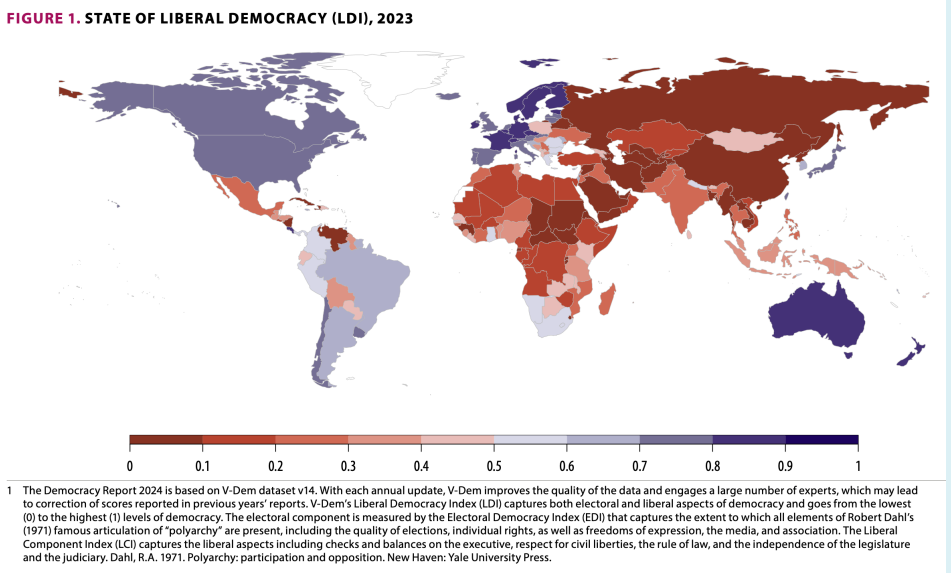
Today's Agenda

1. Interesting Questions in Social Sciences
2. What is Quantitative Social Science?
3. Why Should You Learn(Take) the Course?
4. Course Logistics(How to learn effectively)
5. Understanding Social Science Data

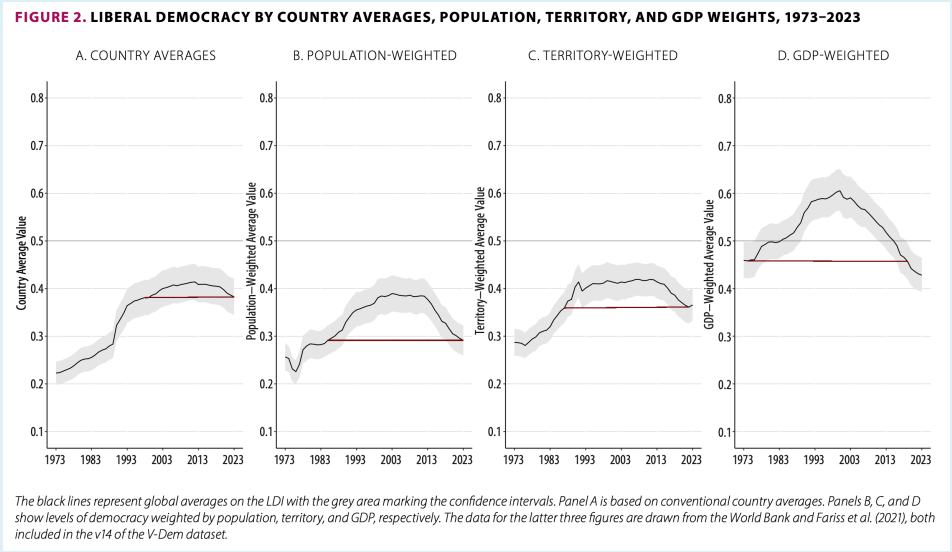
Interesting Questions in Social Sciences

Question #1: Defining and Measuring Democracy

- **Democracy** is a critical concept widely used in political science and other social sciences.
- Measuring the state of democracy across the world helps us understand the extent to which people have political rights and freedoms.



V-Dem Index across the world in 2023



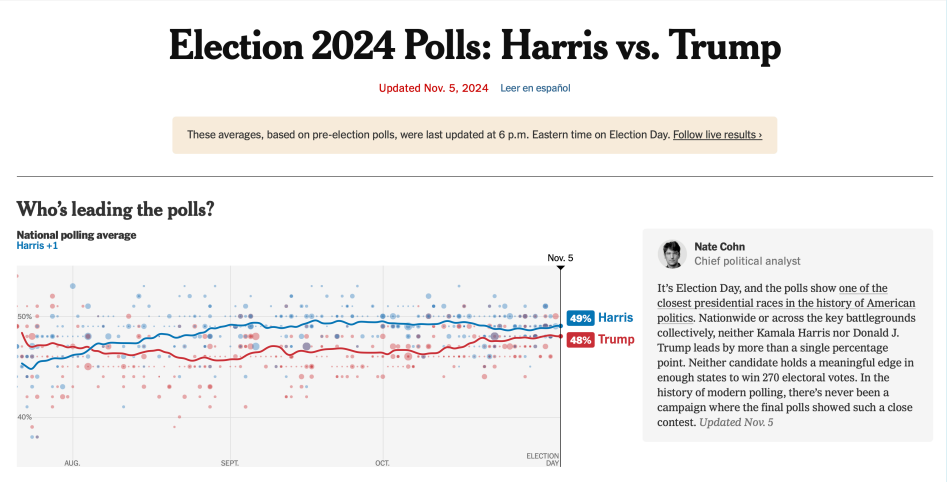
V-Dem Index of the world: 1973-2023

- Nord, Marina et al(2024). Democracy Report 2024: Democracy Winning and Losing at the Ballot.V-Dem Institute.

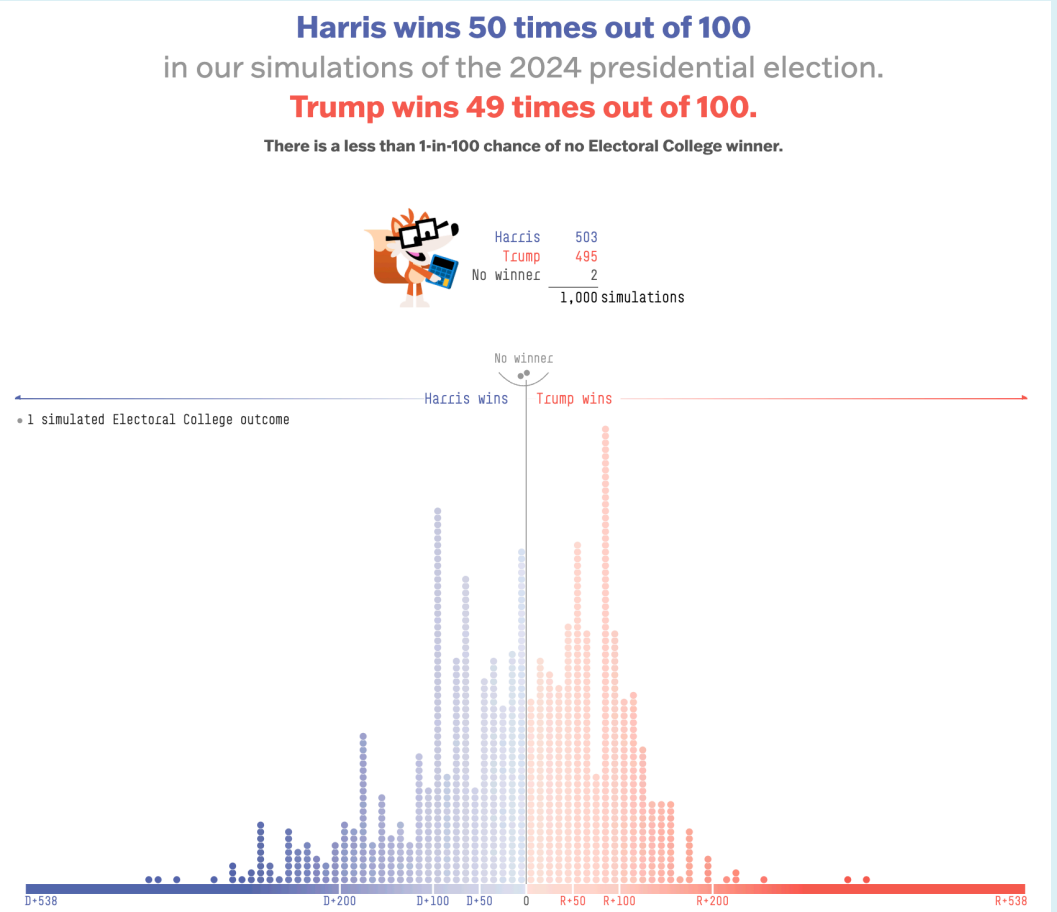
Question #2: Class Size and Student Performance

- A Classical Topic in Economics of Education: **Class Size and Student Performance**
  - *Is there a gap of students' performance between large-size classes and small-size classes?*
- Turn it into an empirical or policy question:
  - *What is the quantitative effect of reducing class size on student achievement?*
  - Further, *how much is the effect by 5 student per class? or 10?*

Question #3: Election Prediction



Source: NYT's Election Forecast



Source: Fivethirtyeight

Question #3: Election Prediction

Source:politico

Source:BBC

- Understanding all these news articles and questions involves making **predictions** using quantitative methods.

Quantitative Answers to Questions

- **Other Similar Questions**:
  - Air pollution and Health?
  - Credit regulation on housing price
  - Coupon on products sales
  - Trade War...
  - Pandemic...
- Quantitative answers to both qualitative and quantitative questions:
  - **a relative or acceptable quantitative measure**
  - **a numerical answer to the question**
  - **a measure of the precision of the answer or the confidence level in the answer**
- This is the role of **Quantitative Social Science**.

What is Quantitative Social Science?

An Overview of Quantitative Social Science

What are Social Sciences?

- Knowledge System of Human Societies

Category	Description	Disciplines
Sciences	Study natural phenomena, laws, and principles, relying on experiments and observations to explain and predict them.	Physics, Astronomy, Chemistry, Biology, Medicine, etc.
Social Sciences	Explore human social behavior, structures, and cultures to understand social phenomena, organizations, and relationships.	*Economics*, *Political Science*, *Sociology*, *Psychology*, *Anthropology* , etc.
Humanities and Arts	Focus on human culture, creativity, and interpretive expression, emphasizing history, criticism, and aesthetic values.	Literature, History, Philosophy, Arts, Religious Studies, etc.

- Others: Formal sciences like Mathematics and Logic, Applied sciences like Engineering and Applied social sciences like Management.

An Overview of Quantitative Social Science

A Framework for Scientific Method

- Anecdotes(轶事）or Intuition(直觉) about a phenomenon or issue(Topics).
  - Propose meaningful or interesting questions(It does matter or we care about)
- Theory/Model/Hypothesis(理论/模型/假说)
  - Systematical methodology: Theory, Model, Logical deduction
  - To obtain a preliminary conclusion or propose an hypothesis.
- Empirical evidence（经验证据）
  - Collect data: Survey, Experiment, Interview, Observation, etc.
  - Use data to test the hypothesis by case study or quantitative analysis methods.

An Overview of Quantitative Social Science

An Example: Smoke and Mortality

- Anecdotes(轶事）or Intuition(直觉）
  - eg. "My grandmother smoked two packs a day and lived until she was 95 years old."
- Theory and Hypothesis
  - Because Cigarettes contain carcinogens(致癌物) such as nicotine, tar, and formaldehyde(尼古丁、焦油、甲醛等), then...
- Empirical Evidence
  - Collecting data through experiments or surveys, and then use statistical or econometrical methods to verify whether and how cigarettes can harm our health.
- Conclusion:
  - Cigarettes can harm our health and quantify the extent of the harm.

An Overview of Quantitative Social Science

Qualitative vs. Quantitative Analysis

- Qualitative Analysis(定性分析)
  - focus on subjective experiences and insights.
  - often involves in-depth interviews, observations, and case studies, which is not easy to quantify.
  - data is often non-numerical and difficult to quantify,like interview transcripts,diary entries, and field notes,etc.
- Quantitative Analysis(定量分析)
  - focus on objective and measurable data.
  - often uses statistical and mathematical models to analyze data.
  - data is often numerical and can be quantified,like survey data, census data, and experimental results.

An Overview of Quantitative Social Science

What is Quantitative Social Science? (量化社会科学)

- Within the framework of the scientific method, all social sciences that use quantitative methods can be unified under the term *Quantitative Social Science*.
- **Quantitative Social Science** (QSS) involves using quantitative methods and analyzing data to understand and solve problems related to society and human behavior.
- Quantitative Social Science is prevalent
  - Many programs are named "Quantitative Social Science" at both undergraduate and graduate levels.
  - Many courses titled "Quantitative Social Science" or "Quantitative Methods for Social Sciences" are offered in various disciplines.
- In practice, scholars who identify as QSS are more likely to work in political science, sociology, and education rather than economics.

Axioms of Quantitative Social Science

- Axioms of Quantitative Social Science:
  1. Social realities or phenomena can be measured and quantified.
  2. Social realities or phenomena occur in a systematic and regular manner.
  3. Past behavior can be used to predict future behavior (at least to some extent).
- Grounded in these axioms, QSS has three main missions:
  - **Measurement and Description**(测度与描述)
  - **Causal Inference**(因果推断)
  - **Prediction and Forecasting**(预测)

An Overview of Quantitative Social Science

Three Missions of QSS

- **Measurement and Description**: *What happened or is happening in the world?*
  - Collect data in a systematic way.
  - Select and construct indices and indicators to measure complex social concepts.
  - Summarize and visualize data by descriptive statistics.
- Tools: **Statistics**
- eg: Using Democracy Index to measure the level of democracy in the world.

An Overview of Quantitative Social Science

Three Missions of QSS

- **Causal Inference**: *Why did it happen?*
    - Identify the causal relationship between variables, such as the effect of education on income, the impact of minimum wage on employment, and the effect of smoking on health.
- Tools: **Econometrics/Causal Inference**
- eg: Class Size on Test Scores

An Overview of Quantitative Social Science

Three Missions of QSS

- **Prediction and Forecasting**: *What will happen?*
  - Predict future outcomes based on models and past data
  - predict the stock market, forecast the weather, predict the outcome of an election, etc.
- Tools: **Time Series/Machine Learning**
- eg: Predicting the outcome of the election

Quantitative Social Science in the Age of Big Data and AI

QSS in the Age of Big Data and AI

Revolutions in Social Sciences

- Social sciences have experienced two methodological **revolutions** over the past few decades.

- **No.1: Credibility Revolution**
  - A movement that emphasizes the goal of obtaining secure causal inferences in social sciences.
  - The revolution started from around the 1990s, pioneering in economics, then spread over to other empirical social sciences such as sociology, political science,public policy, etc., which has entirely changed empirical social science and business research.

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

© Nobel Prize Outreach. Photo: Paul Kennedy
David Card
Prize share: 1/2

© Nobel Prize Outreach. Photo: Risdon Photography
Joshua D. Angrist
Prize share: 1/4

© Nobel Prize Outreach. Photo: Paul Kennedy
Guido W. Imbens
Prize share: 1/4

Revolutions in Social Science

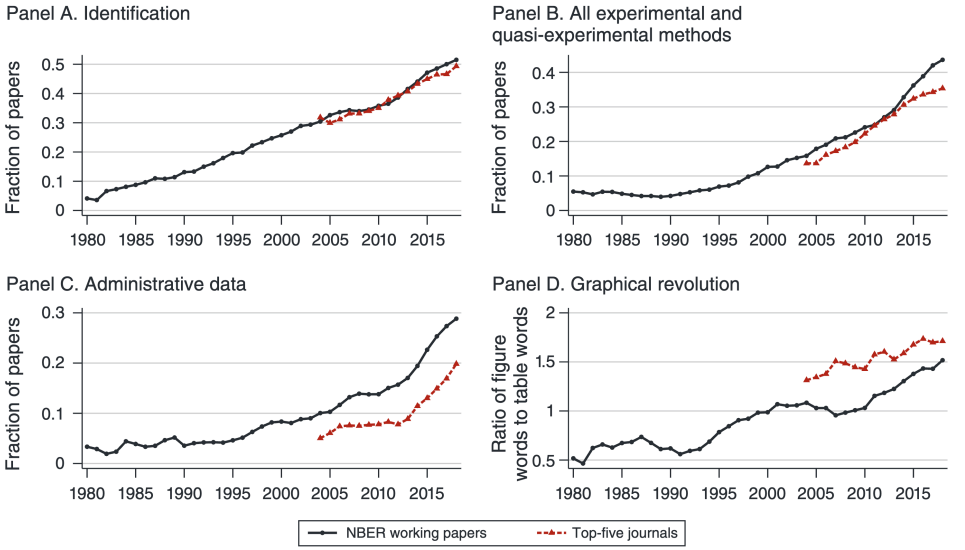


FIGURE 2. THE CREDIBILITY REVOLUTION

Notes: This figure shows different dimensions of the "credibility revolution" in economics: identification (panel A), all experimental and quasi-experimental methods (panel B), administrative data (panel C), and the graphical revolution (panel D). Panel D shows the ratio of the number of "figure" terms to the number of "table" terms mentioned. See Table A.I for a list of terms. The series show five-year moving averages.

Key words for CR

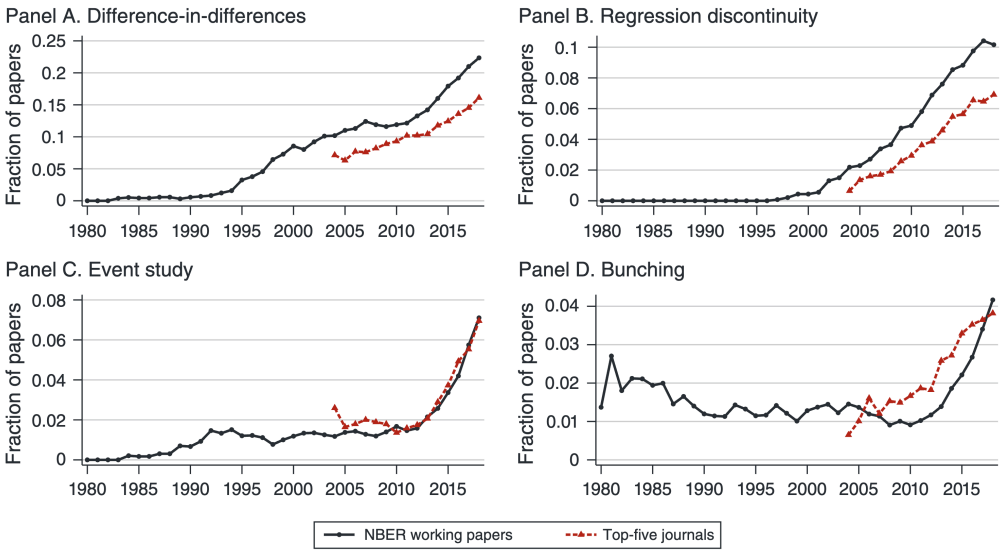


FIGURE 4. QUASI-EXPERIMENTAL METHODS

Notes: This figure shows the fraction of papers referring to each type of quasi-experimental approach. See Table A.I for a list of terms. The series show five-year moving averages.

Quasi-experimental methods

- Currie, J., Kleven, H., & Zwiers, E. (2020). Technology and Big Data Are Changing Economics: Mining Text to Track Methods. AEA Papers and Proceedings, 110, 42–48

QSS in the Age of Big Data and AI

Revolutions in Social Science

- **No.2: Big Data Revolution**
  - How our increasing ability to produce, collect, store and analyze vast amounts of data is going to transform our understanding of the human affairs. (Schonberger and Cukier, 2014)

- Viktor Mayer-Schnberger and Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work and Think

- Data sources and types are changing, which makes new methods to obtain, process, analyze and visualize data necessary.

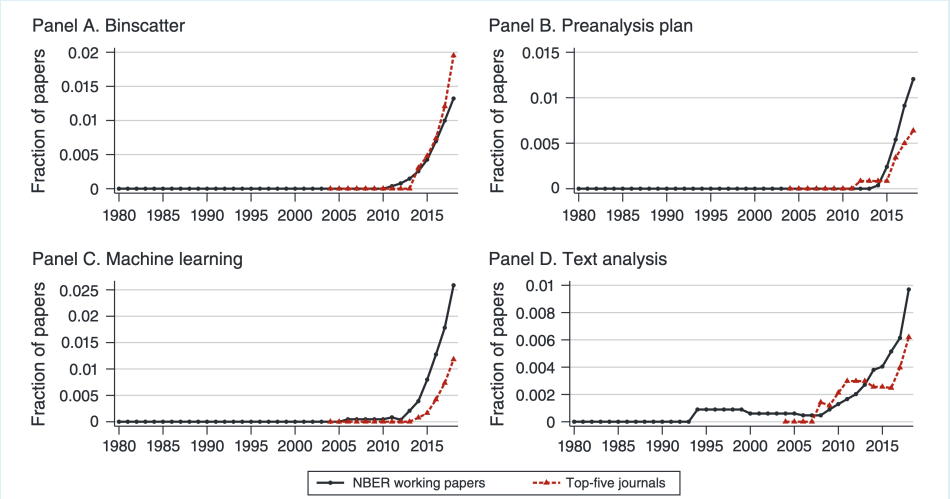


FIGURE 6. WHAT'S NEW?

Notes: This figure shows the fraction of papers referring to each method. See Table A.I for a list of terms. The series show five-year moving averages.

Currie, J et al(2020)

QSS in the Age of Big Data and AI

Revolutions in Social Science

- Now we are facing the third revolution in social science: **No.3: AI Revolution**.

What is Artificial Intelligence (AI)?

- It is a field of **computer science** that aims to create machines that can perform tasks that typically require human intelligence.It includes several subfields, such as machine learning, natural language processing, computer vision, robotics, and expert systems.
- The most influential breakthrough in AI recently is in the space of **Generative AI** models or the **Large Language Models(LLM)**
  - which are designed to create **text or other forms of media** based on patterns and examples they have been trained on such as **ChatGPT** and many others.
- The ability of the AI model is still quickly evolving and upgrading.

QSS in the Age of Big Data and AI

Revolutions in Social Science

- It is dramatically changing the way of obtaining, processing, analyzing and visualizing infomation and knowledge. So it is also changing the way of doing research.

Category	Task	Usefulness
Ideation and Feedback	Brainstorming	●
	Feedback	◐
	Providing counterarguments	◐
Writing	Synthesizing text	●
	Editing text	●
	Evaluating text	●
	Generating catchy titles & headlines	●
	Generating tweets to promote a paper	●
Background Research	Summarizing Text	●
	Literature Research	○
	Formatting References	●
	Translating Text	●
	Explaining Concepts	◐

The third column reports my subjective rating of LLM capabilities as of September 2023:
○: experimental; results are inconsistent and require significant human oversight
◐: useful; requires oversight but will likely save you time
●: highly useful; incorporating this into your workflow will save you time

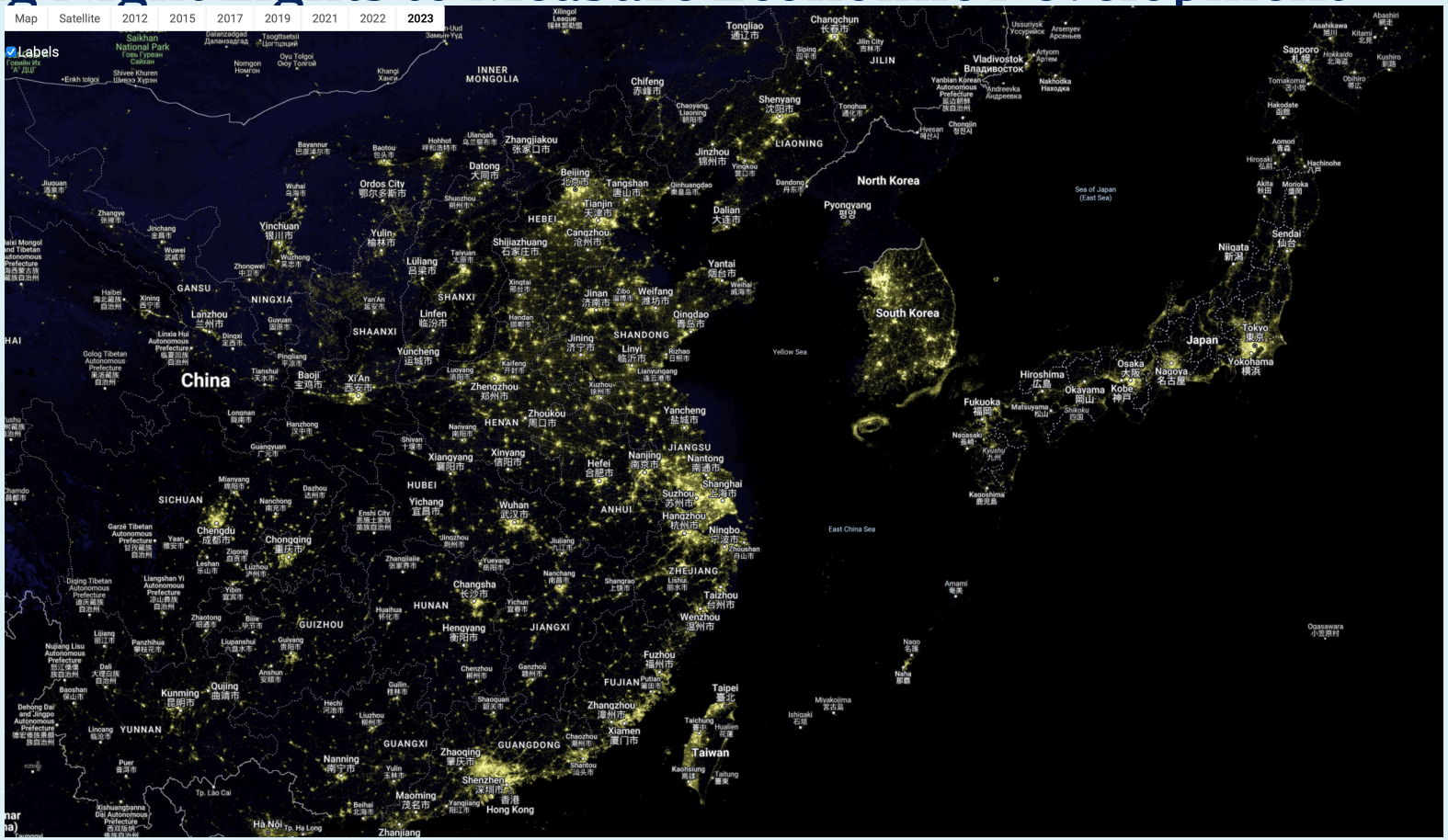
Category	Task	Usefulness
Coding	Writing code	◐
	Explaining code	◐
	Translating code	●
	Debugging code	◐
Data Analysis	Creating figures	◐
	Extracting data from text	●
	Reformatting data	●
	Classifying and scoring text	◐
	Extracting sentiment	◐
	Simulating human subjects	◐
Math	Setting up models	◐
	Deriving equations	○
	Explaining models	◐

The third column reports my subjective rating of LLM capabilities as of September 2023:
○: experimental; results are inconsistent and require significant human oversight
◐: useful; requires oversight but will likely save you time
●: highly useful; incorporating this into your workflow will save you time

- Korinek, A. (2023). Generative AI for Economic Research: Use Cases and Implications for Economists. Journal of Economic Literature, 61(4), 1281–1317

QSS in the Age of Big Data and AI

Case 1: Using Night Lights to Measure Economic Development

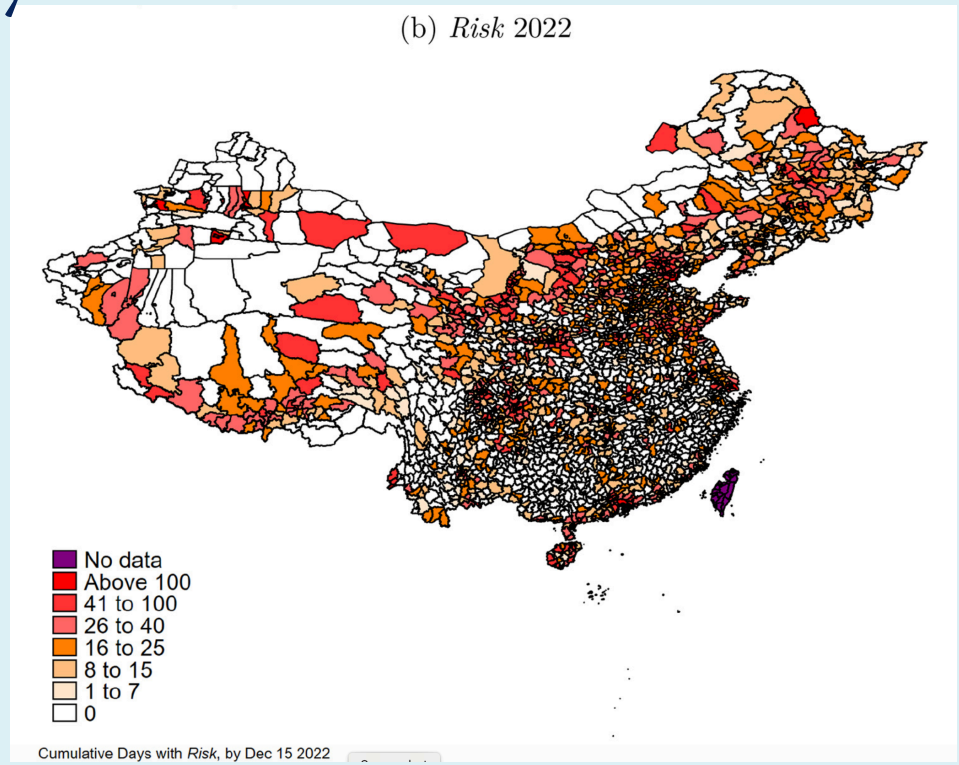


Night Light in NK(NASA,2023)

QSS in the Age of Big Data and AI

Case #2: Measuring the intensity of a policy

- How to measure the intensity of a policy, such as the COVID-19 lockdown policy in China?
- As is well known, there is huge difference for the intensity of the COVID-19 lockdown policy in China across regions and time.
- How to measure the intensity of the policy?



Gong et al(2024)

- Gong, D., Shang, Z., Su, Y., Yan, A. & Zhang, Q. Economic impacts of China's zero-COVID policies. China Econ. Rev. 83, 102101 (2024).

QSS in the Age of Big Data and AI

Case 3: Using Mobile Phone Data to assist in humanitarian aid

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Open access | Published: 16 March 2022

Machine learning and phone data can improve targeting of humanitarian aid

Emily Aiken, Suzanne Bellue, Dean Karlan, Chris Udry & Joshua E. Blumenstock ✉

Nature **603**, 864–870 (2022) | Cite this article

54k Accesses | **502** Altmetric | Metrics

Abstract

The COVID-19 pandemic has devastated many low- and middle-income countries, causing widespread food insecurity and a sharp decline in living standards[1]. In response to this crisis, governments and humanitarian organizations worldwide have distributed social assistance to more than 1.5 billion people[2]. Targeting is a central challenge in administering these programmes: it remains a difficult task to rapidly identify those with the greatest need given available data[3,4]. Here we show that data from mobile phone networks can improve the targeting of humanitarian assistance. Our approach uses traditional survey data to train machine-learning algorithms to recognize patterns of poverty in mobile phone data; the trained algorithms can then prioritize aid to the poorest mobile subscribers. We evaluate this approach by studying a flagship emergency cash transfer program in Togo, which used these algorithms to disburse millions of US dollars worth of COVID-19 relief aid. Our analysis compares outcomes—including exclusion errors, total social welfare and measures of fairness—under different targeting regimes. Relative to the geographic targeting options considered by the Government of Togo, the machine-learning approach reduces errors of exclusion by 4–21%. Relative to methods requiring a comprehensive social registry (a hypothetical exercise; no such registry exists in Togo), the machine-learning approach increases exclusion errors by 9–35%. These results highlight the potential for new data sources to complement traditional methods for targeting humanitarian assistance, particularly in crisis settings in which traditional data are missing or out of date.

- Using mobile phone data and AI to identify(predict) who is the greatest need of help during the Pandemic.
- Relative to the traditional trageting methods, it reduced errors by 21% at most.
- Aiken, E., Bellue, S., Karlan, D., Udry, C. & Blumenstock, J. E. Machine learning and phone data can improve targeting of humanitarian aid. Nature 603, 864–870 (2022).

QSS in the Age of Big Data and AI

- There can be many labels for our work...
  - Econometrics(Causal Inference)
  - Statistics
  - Data Mining/Big Data/Data Science
  - Machine Learning(ML) and Artificial Intelligence(AI)
- Along this spectrum, the focus shifts from heavily emphasizing the phenomena being measured to a more practical approach of discovering patterns that are useful and true.
  - The more we move to the right, the more we are interested in prediction and less in causality.
  - The more we move to the left, the more we are interested in causality and less in prediction.
- The **similarities** are much bigger than any distinctions.

QSS in the Age of Big Data and AI

SYMPOSIUM

We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together

Justin Grimmer, *Stanford University*

Information is being produced and stored at an unprecedented rate. It might come from recording the public's daily life: people express their emotions on Facebook accounts, tweet opinions, call friends on cell phones, make statements on Weibo, post photographs on Instagram, and log locations with GPS on phones. Other information comes from aggregating media. News outlets disseminate news stories through online sources, and blogs and websites post content and receive comments from their readers. Politicians and political elites contribute their own messages to the public with advertising during campaigns. The federal government disseminates information about where it spends money, and local governments aggregate information about how they serve their citizens.

The promise of the "big data" revolution is that in these data are the answers to fundamental questions of businesses, governments, and social sciences. Many of the most boisterous claims come from computational fields, which have little experience with the difficulty of social scientific inquiry. As social scientists, we may reassure ourselves that we know better. Our extensive experience with observational data means that we know that large datasets alone are insufficient for solving the most pressing of society's problems. We even may have taught courses on how selection, measurement error, and other sources of bias should make us skeptical of a wide range of problems.

This statement is true; "big data" alone is insufficient for solving society's most pressing problems—but it certainly can help. This paper argues that big data provides the opportunity to learn about quantities that were infeasible only a few years ago. The opportunity for descriptive inference creates the chance for political scientists to ask causal questions and create new theories that previously would have been impossible (Monroe et al. 2015). Furthermore, when paired with experiments or robust research designs, "big data" can provide data-driven answers to vexing questions. Moreover, combining the social scientific research designs makes the utility of large datasets even more potent.

The analysis of big data, then, is not only a matter of solving computational problems—even if those working on big

of statistical techniques. For the analysis of big data to truly yield answers to society's biggest problems, we must recognize that it is as much about social science as it is about computer science.

THE VITAL ROLE OF DESCRIPTION

Political scientists prioritize causal inference and theory building, often pejoratively dismissing measurement—inferences characterizing and measuring conditions as they are in the world—as "mere description" or "induction." Gerring (2012) showed, for example, that 80% of articles published in *American Political Science Review* focus on causal inference. The dismissal of description is ironic because much of the empirical work of political scientists and theories that they construct are a direct product of description. Indeed, political scientists have developed a wide range of strategies for carefully measuring quantities of interest from data, validating those measures, and distributing them for subsequent articles. Therefore, although descriptive inference often is denigrated in political science, our field's expertise in measurement can make better and more useful causal inferences from big data.

The VoteView project is perhaps the best example of political science's expertise with measurement and why purely descriptive projects affect the theories we construct and the causal-inference questions we ask (McCarty, Poole, and Rosenthal 2006; Poole and Rosenthal 1997).[1] VoteView is best known for providing NOMINATE scores—that is, measures of where every representative to serve in the US House and Senate falls on an ideological spectrum. The authors are emphatic that NOMINATE measures only low-dimensional summaries of roll-call voting behavior. Like other measurement techniques, these summaries are a consequence of both the observed data and the assumptions used to make the summary (Clinton and Jackman 2009; Patty and Penn 2015). Extensive validations suggest, however, that the measures are capturing variation in legislators' expressed ideology (Clinton, Jackman, and Rivers 2004; Poole 1984; Poole and Rosenthal 1985; 1997).

The impact of the VoteView project is broad and substan-

- Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. PS: Political Science & Politics, 48, 80-83.

QSS in the Age of Big Data and AI

- **Big Data and AI** can be seen as a **complement** to quantitative social science, providing new tools and methods to analyze data and extract insights that may not be possible with traditional QSS methods alone.
- **Big Data and AI** cannot speak for themselves; they need to be guided by social scientists in the analysis and interpretation of data.
- The combination of QSS and Big Data and AI can be seen as a new field of **Quantitative Social Science 2.0**.
    - Some Scholars especially in socialogy and communication call it **Computational Social Science**.
- I will stick to the term **Quantitative Social Science** in this course for the sake of tradition and consistency.

Course Logistics

About Me, our TA, and the Course

- My name is Zhaopeng Qu(曲兆鹏)
  - Associate Professor, Institute of Population Studies, Business School.
  - Research Fields: Labor Economics and Applied Econometrics
  - **Email**: qu@nju.edu.cn
- Our TA: **Zhengwu Raoxue(饶雪政武)**(very nice and helpful)
  - Majored in Managment of Info System(MIS) in Nanjing University.
  - Proficient in R and Python.
  - Will asistant you in the lab sessions.
- Online Resources
  - **Our Course Website**: https://byelenin.github.io/QSS_2025/
  - **Wechat group**: discuss anything related to the course.

Prerequisite and Procedures

- Although there are no formal prerequisites for the course, it is recommended that you have taken at least one course in **Statistics**.
- It is assumed that you are **comfortable with or at least willing to learn** **dealing with data** and **coding** using **R/Python**.

- The First Part: **Lectures by the instructor**
  - Briefly introduce the underlying theoretical problems and focus heavily on the empirical strategy.
  - Provide specific examples from classical papers with interesting topics.
- The Second Part: **Computer Labs in and after class**
  - Teach basic coding skills in R and Python.
  - Help you to practice the skills using AI tools to coding and data analysis.
  - Provide a platform for you to apply the skills to your own research projects.

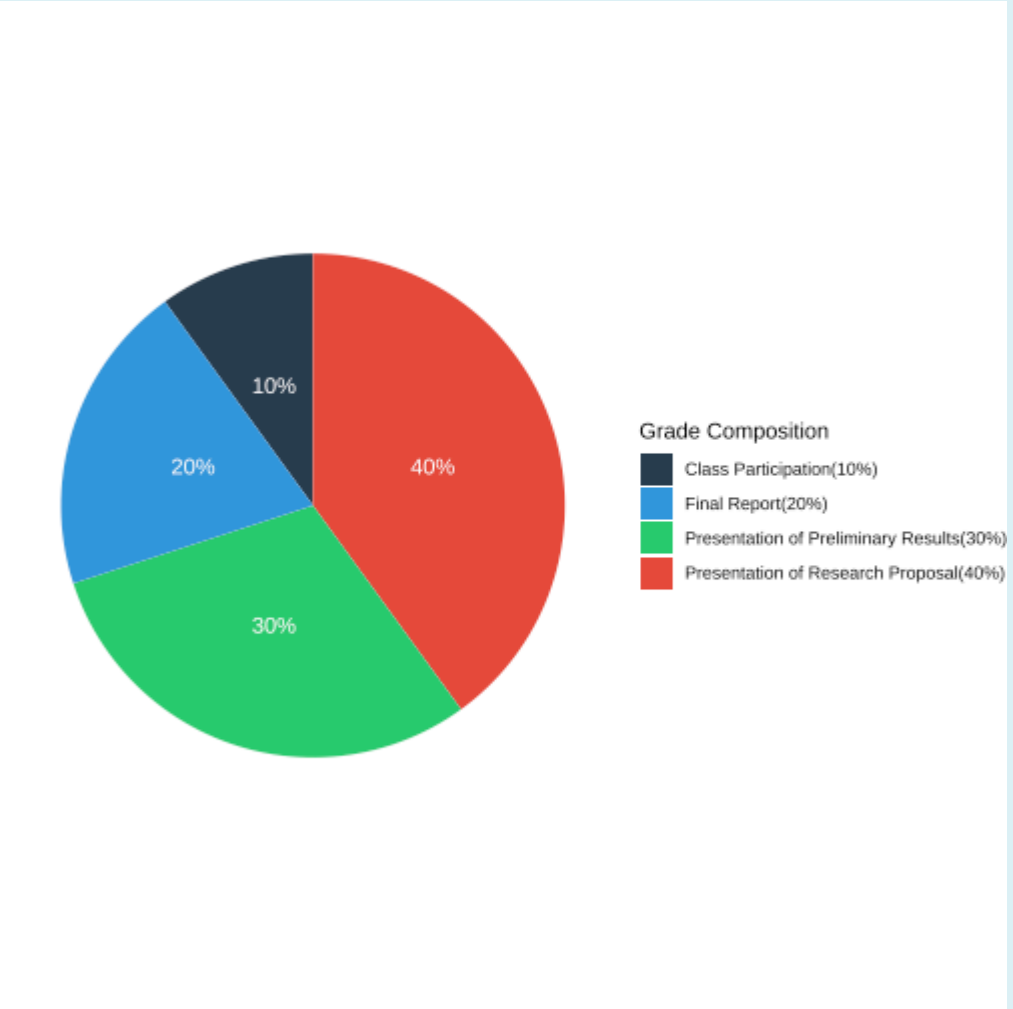
The Procedures

- The Third Part: **Your Own Research Project**
  - Team up with 2-3 classmates to work on a research project.
  - A Presetation for Research Proposal : in mid term
  - A Presentation for Preliminary Results: in the end of the semester
  - Language: *English or Chinese are both acceptable*.

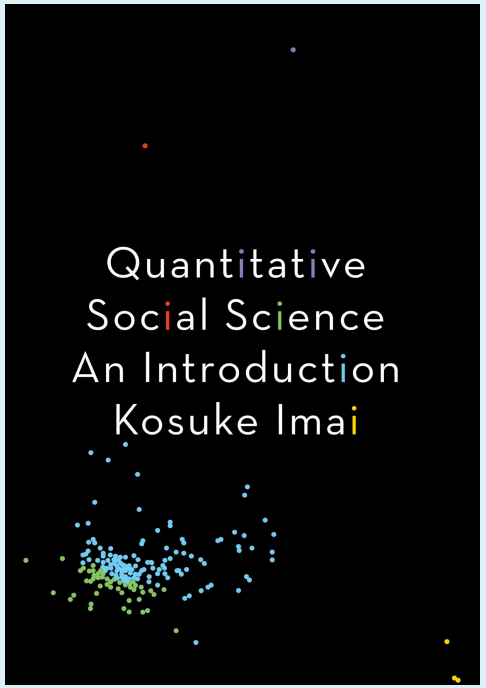
Evaluation

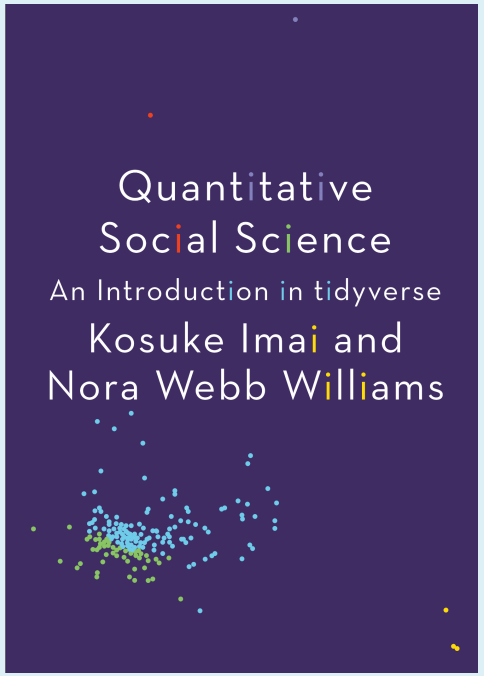
- The final grade will be based on the following components:

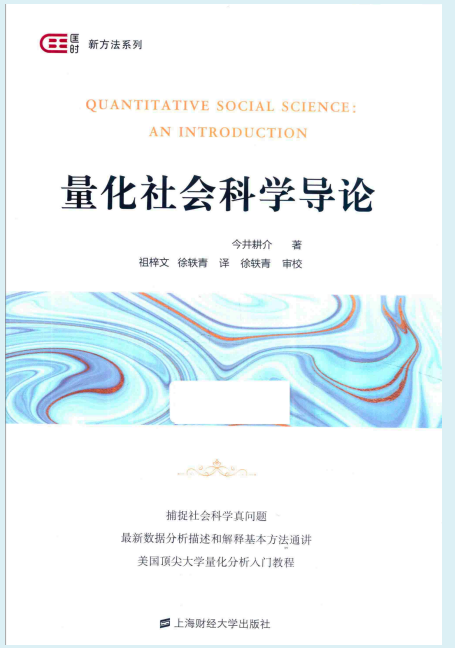
- **Class Participation**
- **Midterm Presentation of Research Proposal**
- **Final Presentation of Preliminary Results**
- **Final Report**



Required Textbooks

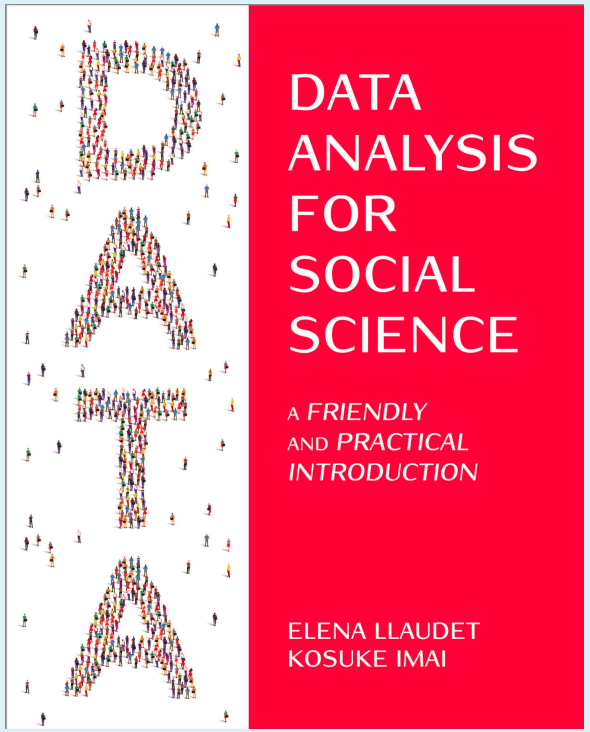


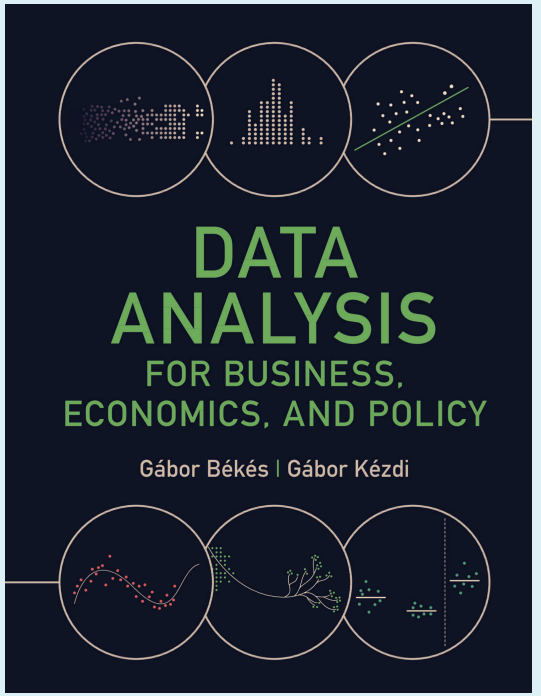




- **Quantitative Social Science: An Introduction**, Kosuke Imai, 2017. Princeton University Press.
- **Quantitative Social Science: An Introduction with Tidyverse** , Kosuke Imai and Nora Webb Willimas, 2022. Princeton University Press.
- **《量化社会科学导论》**，今井耕介著，祖梓文和徐轶青译， 中信出版社，2022年，上海财经大学出版社。

Supplementary Textbooks





- Llaudet, Elena and Kosuke Imai(2023),Data Analysis for Social Science: A Friendly and Practical Introduction, Princeton University Press.
- Gábor Békés and Gábor Kézdi(2021), Data Analysis for Business, Economics, and Policy,Cambridge University Press.

Interesting Books for Reading

- Steven D. Levitt and Stephen J. Dubner, SuperFreakonomics: Global Cooling, Patriotic Prostitutes, and Why Suicide Bombers Should Buy Life Insurance, 2009.（中译本《超爆魔鬼经济学》,中信出版社。）
- Raymond Fisman & Edward Miguel, Economic Gangsters: Corruption, Violence, and the Poverty of Nations, 2010.(中译本:《经济黑帮：腐败、暴力的经济学》，中信出版社。）
- Abhijit V. Banerjee & Esther Duflo, Poor Economics A Radical Rethinking of the Way to Fight Global Poverty, 2011.(中译本:《贫穷的本质：我们为什么摆脱不了贫穷》，中信出版社。）
- Abhijit V. Banerjee & Esther Duflo, Good Economics for Hard Times,2019.(中译本:《好的经济学》，中信出版社。)
- *Help you to understand how empirical economists work and how to do research in empirical economics.*

Computing Tools

- The main computing tools used in the course are **R**. And **Python** is also used occasionally.

R

- Pros:
  - More popular in academia while also used in industry.
  - More powerful and flexible in data analysis and visualization.
- Cons:
  - Less multi-functional and flexible than Python.
  - Less supported other than economics, statistics, and data analysis.

Python

- Pro
  - More multi-functional and flexible.
  - More general-purpose and supported by many fields.
- Con
  - More steep learning curve.
  - Less popular in academia while more used in industry.
  - Less powerful and flexible in data analysis and visualization.

Promise and Expectation

What I promise to offer you

- Prepare lectures as well as possible.
- One to one interaction on topics covered in the course, especially for your own topics.
- Help you start to using R to analyze data sets in China.
- **A good score?**
  - **It depends on you**.

What I expect to you

- Class participation with a little bit aggressive attitude.
  - More questions, more scores!
- Self-motivated learning by doing.

One More Thing

- **Don't ever cheat on your projects!**
- And how to use AI tools properly in your study and research? please follow my instruction.

Welcome contact me

Why and Who should take the course?

Why take the course?

Course Goals

- Understand and master the **basic concepts and fundamental methods** of QSS.
- Learn to use softwares such as **R/Python** for data collection, cleaning, analysis, and visualization.
- Gain a preliminary understanding of **version control concepts** and learn to use **AI** tools like **Github Copilot/ChatGPT/Claude/Deepseek** to assist in programming and data analysis tasks.
- Learn about how to use **non-traditional data sources** such as satellite images, web scraping, and OCR text, etc. in QSS research.

Why take the course?

The Purpose of the course

- Introduce you to quantitative social science in a way that is **accessible** to students who have not taken a prior course in related fields.
- The course focuses more on intuition and practical examples, while **keeping the mathematical content to a minimum**. Notably,**matrix notation is not used** and probability theory and statistics are not required but recommended.
  - Laying down a solid theoretical background in **introductory level** QSS,
  - Inquiring about some basic instruments and some latest development in the **empiricist's toolbox**.
  - Developing an ability to implement modern QSS methods with AI tools to real-world data.

Why take the course?

Hopefully, taking this course can help you:

- Develop your quantitative analysis skills to analyze real-world data.
- Cultivate a **critical thinking** about empirical studies and applications in social sciences.
- Help yourself enjoy to learn some new ideas in an empiricist's mindset.

Improve your skills in graduate studies

- **Hard Skills**
  - **Language**
  - **Computer**
  - **Presentation and Writing**
- **Soft Skills**
  - **Critical Thinking**
  - **Teamwork**

- Fortunately, you could practice all above skills more or less in our class.

Why take the course?

For those pursuing an academic career:

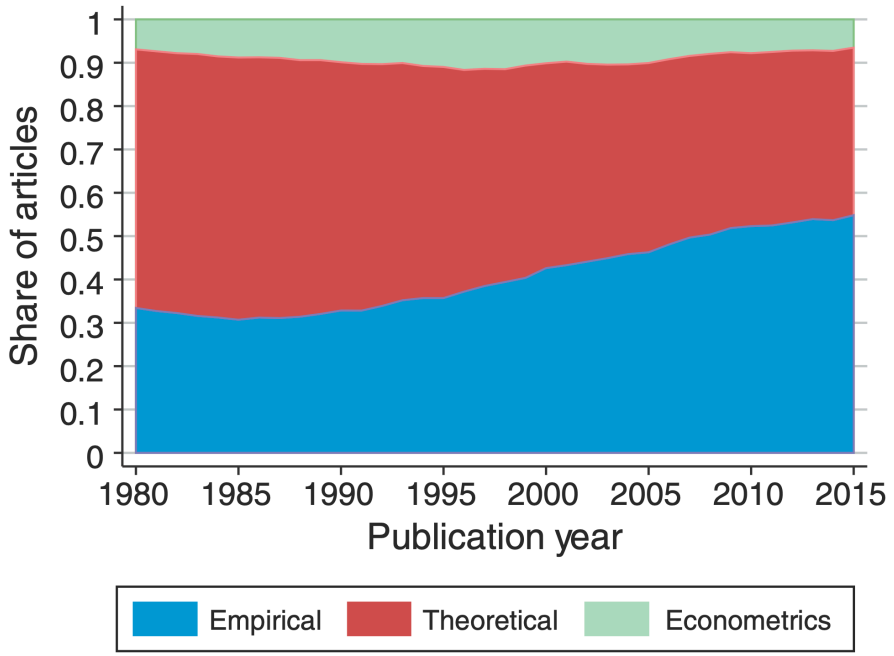


FIGURE 6. WEIGHTED PUBLICATIONS BY STYLE

Angrist et al(2017)

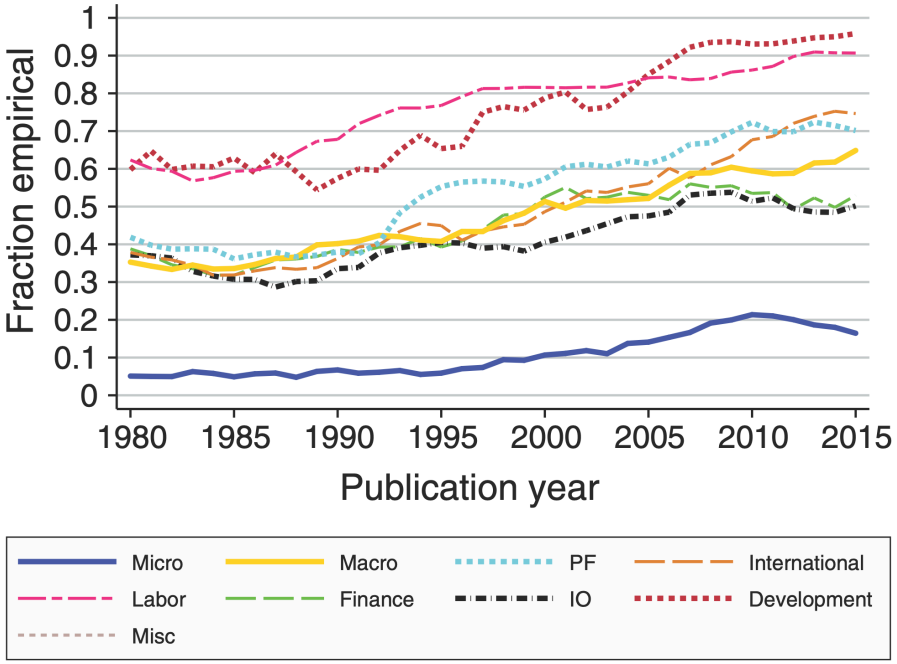


FIGURE 4. WEIGHTED FRACTION EMPIRICAL BY FIELD

Angrist et al(2017)

- The proportion of empirical studies in economics is **increasing more and more**.

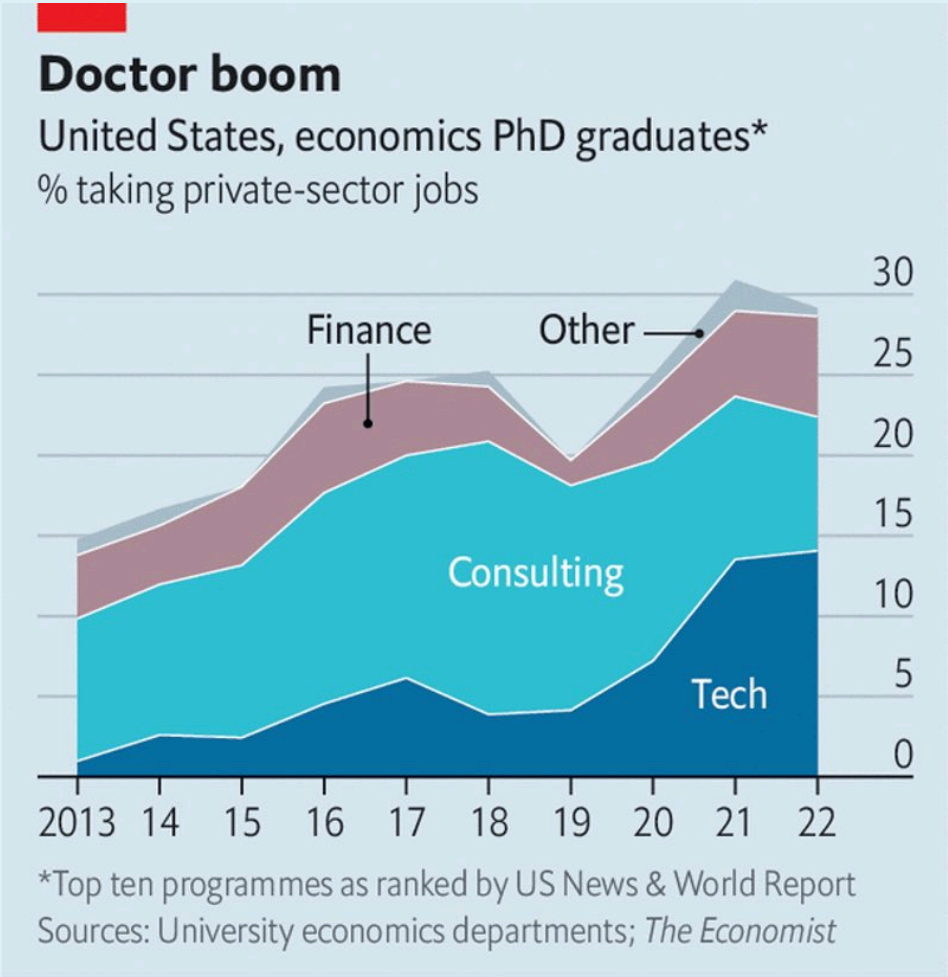
Why take the course?

Who want enter industry job market

- Who want to work in industry: mastering econometrics may help you **get a good job!**
- A lot of internet giants even hire economists to lead their special R&D department. Such as
  - Google,Microsoft,eBay,Baidu,Alibaba,Tencent,Tiktok
- **Data Analyst/Data Scientist** is **the hottest job** in consulting, business areas as well as financial industry right now.

Why take the course?

The Change of EcoPhD Job Markets in the US by The Economist



The Economist(2022): Why economists are flocking to Silicon Valley

Why take the course?

Who want enter industry job market: Apple Job Wanted

Economist/Core Data Scientist
Apple · Beijing, Beijing, China

Apply Save ···

Key Qualifications

Strong background in statistics or econometrics, regression analysis, causal inference, time series analysis, GLM, logistic regression, probability theory, regularization, interest in machine learning algorithms

Develop internal visualization and modeling tools to facilitate data-driven decisions

Present results and other analytical findings to business partners

Strong statistical background and experience with causal inference, time series analysis (e.g. ARIMA, exponential smoothing, time series regression methods etc.), forecasting, and data analysis

Experienced R/Python programmer also proficient in other languages important to the ETL data pipeline (e.g. SQL)

Experience with data visualization packages (e.g. ggplot2, plotly) and advancing multiple projects at once on a tight schedule

Ability to share results with a non-technical audience

Experience in bayesian statistics and modeling (e.g. bayesian structural time series, dynamic linear models)

Advocate and practitioner of version control and reproducible code

Excellent verbal and written communication skills, in both Mandarin Chinese and English

Description

- Work with various teams to understand business problems and provide business solutions
- Build models to causal impact of new programs release across different scenarios
- Develop internal visualization and modeling to facilitate data-driven decisions
- Present results and other analytical findings to business partners

Education & Experience

- PhD in Economics or related fields
- M.S. in related field with 5+ years experience applying econometric models to business problems.

Why take the course?

Who enter industry job market: ByteDance Job Wanted

国际电商–经济学家/数据科学家

上海 | 正式 | 产品 – 数据分析 | 职位 ID：A117677

职位描述

我们欢迎有创造力、探索精神、且具备基本经济学、统计学素养的人才加入，和业务方共创并推动项目的上线落地。我们的合作业务方包括推荐算法、产品、运营、资源管理等。

主要职责：

把商业问题转化为可解的模型问题。通过经济学视角的思考和科学的方法（因果推断、AB实验、求解理论模型、预测等）来推动搜推策略、产品功能、资源分配等相关决策:

1、因果性的衡量各类策略、政策的效果，衡量长期影响，并形成系统性的方法论；

2、对数据现象现象进行归因，对用户、商家的决策链路做深入探索，总结洞察和建议，帮助各决策方建立认知；

3、优化各类资源分配（流量、营销补贴）；

4、优化国际电商的生态环境，包括但不限于经营环境，用户体验，内容生态，供给生态，持续助力商达成长和用户增长。

职位要求

1、经济学、统计学、运筹学、金融学、或者其他的相关的量化学科背景；

2、掌握 R 或Python等至少一项数据分析必备的编程语言，以及基础SQL能力；

3、有一定的解决商业问题、构建可落地的系统性解决方案、复杂项目管理、协调多方决策的经验；

4、良好的写作沟通能力；

5、以下领域的相关的科研、或者业界项目经历: reduced-form 因果推断、预测、causal ML、劳动经济学、健康经济学、教育经济学、行为经济学、金融经济学、产业组织学。

投递

商业化数据科学家–因果推断

上海 | 正式 | 产品 – 数据分析 | 职位 ID：A56405

职位描述

1、通过积累日常使用经验、阅读相关学术论文和公开资料等，沉淀并向数科团队输出对因果推断方法论的深入理解和使用经验，澄清常见的使用误区，提供标准应用流程指南，以保障方法在团队内应用的科学性，提高使用效率；

2、关注具有方法论共性或场景共性的相似业务问题，主导专项探索，与其他业务方向数科同学紧密配合，从宏观视角优化资源分配效率或策略，优化产品策略，或针对相似问题抽象可复用的、普适的分析框架或解决方案，提升团队分析、决策效率；

3、对宏观战略问题进行拆解、定义，通过数据描述、可视化、挖掘、统计建模等方法，提炼有效的数据洞察和产品战略建议，指导科学的决策与迭代。

职位要求

1、本科以上学历，统计学、数学、计量经济学、数据科学、计算机等量化分析相关专业优先，硕士、博士优先；

2、具备扎实的统计学/计量经济学/机器学习/因果推断等数据科学理论基础及应用经验；精通SQL，熟练掌握Python/R中的一种，可进行数据清洗、可视化和分析；

3、具备快速学习能力，能够快速理解产品逻辑，并具备较强的逻辑思维能力，在较大不确定性的问题中可以构建分析框架，将数据转化为有效的商业洞察；

4、能够主动、独立思考的同时，具备良好的团队协作能力与责任心，善于与其他协作团队沟通，有主人翁意识；

5、具备强烈的好奇心与自我驱动力，乐于接受挑战，追求极致和创新，富有使命感。

投递

Why take the course?

Who find a job in public sectors in China(shang'an movement)

Source:SCMP

- To be honest with you, the course may **NOT** help you succeed in the examination directly.
- However, in the long run, it provides valuable knowledge and skills that offers a broader understanding of the subject matter and enhances your critical thinking abilities.
- It ultimately benefits your career, as well as our people, our country, and the world.

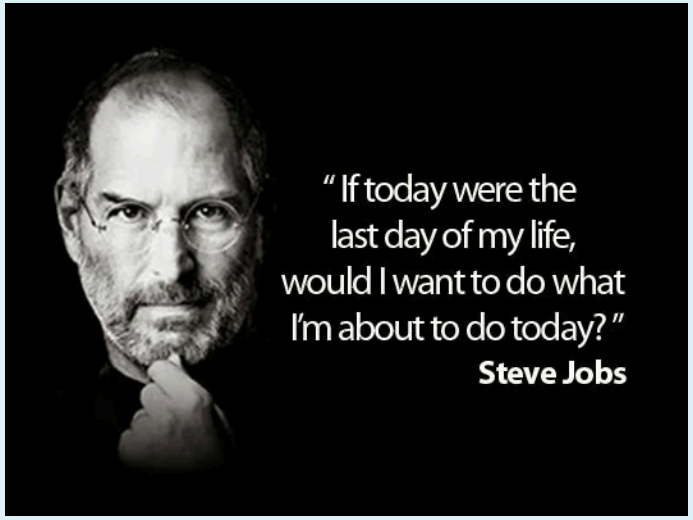
Why take the course?

Who look for fun

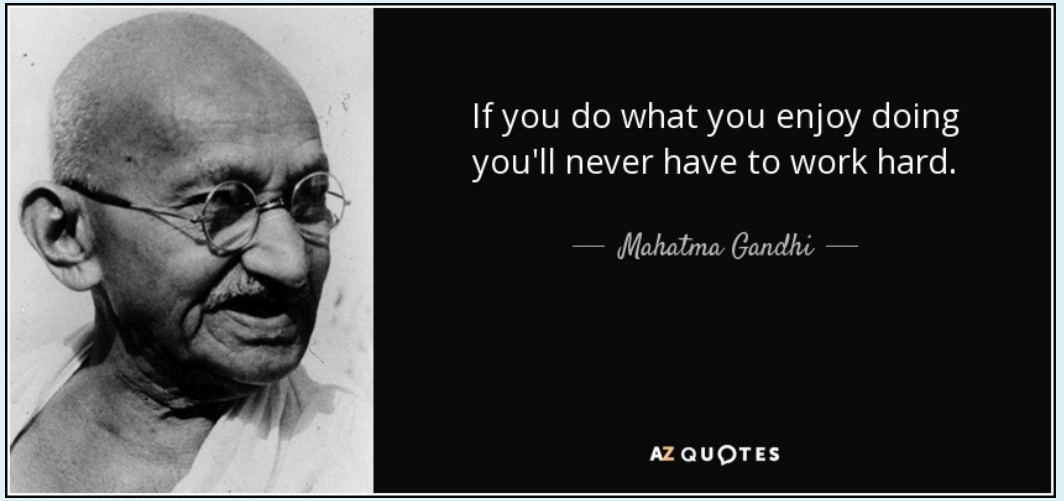
- I will try my best to make the course an interesting and enjoyable class rather than a boring and demanding variant of a statistics or computer science course.
- Also covering several interesting and insightful stories like
  - Eg. **Crime and Abortion** in *Freakonomics* written by Steven Levitt.
  - Eg. What is **the economic value** to be the president's son(or daughter)? in *Economic Gangster* written by Raymond Fisman and Edward Miguel.

Whoever and Whatever

Whoever you would like to be or whatever you want



- **Every choice you make has an opportunity cost,try your best to make a wise one**.



- **Enjoy doing something seriously and cultivate a special quality for yourself!**

Wrap up

- QSS is an **essential and intriguing yet challenging** course.
  - **Please consider carefully before enrolling**
  - **Once committed,please work hard on it!**
  - **And remember, enjoy the process of working hard!**

An Introduction to Economic Data

What is Data

- Data is a collection of facts or information, which can be presented in various forms such as *numbers*, *tables*, *words*, *graphs*, *pictures*, or even *sounds* and *videos*.
- And it can be processed and analyzed to produce knowledge and insights either by itself or after structuring, cleaning, and analysis.
- Data is most straightforward to analyze if it forms a single **data table**(a matrix).
  - It consists of **observations**(观测值) and **variables**(变量).
  - Observations are also known as cases, or row.
  - Variables are sometimes called **features** or covariates.
- Normally, in a data table the *rows are the observations*, *columns are variables*.

A Simple Example: CA School Data

TABLE 1.1 Selected Observations on Test Scores and Other Variables for California School Districts in 1999

Variables

Observation (District) Number	District Average Test Score (fifth grade)	Student–Teacher Ratio	Expenditure per Pupil ($)	Percentage of Students Learning English
1	690.8	17.89	$6385	0.0%
2	661.2	21.52	5099	4.6
3	643.6	18.70	5502	30.0
4	647.7	17.36	7102	0.0
5	640.8	18.67	5236	13.9
. . .	. . .	. . .	. . .	. . .
418	645.0	21.89	4403	24.3
419	672.2	20.20	4776	3.0
420	655.8	19.04	5993	5.0

ID of Observations

One case

Note: The California test score data set is described in Appendix 4.1.

Data: Basic Characteristics

- First, we need to distinguish between the **population** and the **sample**.
  - **Population**: the entire group of individuals, objects, or events of interest.
  - **Sample**: a subset of the population that is supposed to be representative of the population.
- Second, we need to know **the unit of observation** and **the unit of analysis**.
  - **Unit of Observation**: the entity about which data are collected.
  - **Unit of Analysis**: the entity about which analysis/inferences are made.

A Simple Example: CA School Data

Topic: the effect of **class size** on **student achievement** in California public schools.

- Ideally,
  - **Population**: all the students in the CA public school system.
  - **Sample**: a subset of the students in the CA public school system by a **random sampling**.
- **Unit of Observation** and **Unit of Analysis**: each student.
- In reality, based on the availability of data and the way data is collected
  - **Population**: all the schools(districts), which can be equivalent to all the students in the CA public school system.
  - **Sample**: a subset, 420, of the all schools(districts) by a random sample.
- **Unit of Observation** and **Unit of Analysis**: each school(district)
- Actually, the unit of observation and unit of analysis can be **different**.
  - eg.the unit of observation is the student, but the unit of analysis are both the student and class.

Data: Sources and Types

Data Sources

- Traditional Collecting Methods:
  - Statistical Reports or Documents
  - Survey or Census
  - Administrative Data
  - Lab or Field Experimental Data
- Collecting Data in Digital Times:
  - Online Transactions or Activities
  - Social Media
  - Geolocations or Geographic Data
  - Online Documents or Texts
- Always need to figure out
  - what is the **population and sample** of the data.
  - what is the **unit of observation and unit of analysis** of the data.

Data: Sources and Types

Data Quality

- Including
  - Content
  - Accuracy
  - Completeness
  - Consistency
- **Garbage in, Garbage out**
  - **Prioritize data, then methods.**

Ethical and Legal Issues

- Including
  - Privacy and Confidentiality
  - Data Security
  - Data Ownership
  - Data Sharing and Open Data

Data Types

Experimental V.S. Observational

- **Experimental** data come from experiments designed to evaluate a treatment or policy or to investigate a causal effect.
- **Observational** data come from non-experimental settings, such as surveys,administrative records and other sources.

Data Structure

- Cross-sectional data
- Time series data
- Panel/longitudinal data
- Pool-cross sectional data

1.Cross-Sectional Data: (Major Focus)

- Units: individuals, households, firms, cities, states, countries, etc.
- Data on multiple agents at a single point in time

$$\{x_i, y_i \dots\}_{i=1}^{N}; N = Sample\ Size$$

- Usually obtained by random sampling from the underlying population.It means

$$\{x_i, y_i \perp x_j, y_j\}, i \neq j \in N$$

- Cross-sectional data are widely used in economics and other social sciences:
  - labor economics, public finance, industrial economics, urban economics, health economics...

1.Cross-Sectional Data: (Major Focus)

TABLE 1.1 Selected Observations on Test Scores and Other Variables for California School Districts in 1999

Observation (District) Number	District Average Test Score (fifth grade)	Student–Teacher Ratio	Expenditure per Pupil ($)	Percentage of Students Learning English
1	690.8	17.89	$6385	0.0%
2	661.2	21.52	5099	4.6
3	643.6	18.70	5502	30.0
4	647.7	17.36	7102	0.0
5	640.8	18.67	5236	13.9
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
418	645.0	21.89	4403	24.3
419	672.2	20.20	4776	3.0
420	655.8	19.04	5993	5.0

Note: The California test score data set is described in Appendix 4.1.

- **Questions?**: observations,variables,the sample size?

$$x_i = STRatio_i;\ y_i = TestScore_i;\ N = 420$$

2.Time Series Data:(Minor Cover)

- Observations on a variable (or several variables) over time, thus data on a single agent at multiple points in time

$$\{x_t, y_t \dots\}_{t=1}^{T}; T = Sample\ Size$$

- Examples:
  - stock prices, money supply
  - consumer price index(CPI)
  - gross domestic product(GDP)
  - automobile sales
- Data frequency: minutes, hourly, daily, weekly, monthly, quarterly, annually.
- Economic observations can rarely be assumed to be independent across time. So we have to account for the dependent nature of economic time series.

2.Time Series Data:(Minor Cover)

TABLE 1.2 Selected Observations on the Growth Rate of GDP and the Term Spread in the United States: Quarterly Data, 1960:Q1–2013:Q1

Observation Number	Date (year:quarter)	GDP Growth Rate (% at an annual rate)	Term Spread (% per year)
1	1960:Q1	8.8%	0.6%
2	1960:Q2	−1.5	1.3
3	1960:Q3	1.0	1.5
4	1960:Q4	−4.9	1.6
5	1961:Q1	2.7	1.4
.	.	.	.
.	.	.	.
.	.	.	.
211	2012:Q3	2.7	1.5
212	2012:Q4	0.1	1.6
213	2013:Q1	1.1	1.9

Note: The United States GDP and term spread data set is described in Appendix 14.1.

- **Questions?** observations,variables,the sample size?

$$x_t = Date(quarter);\ y_t = GDP\ Growth\ Rate;\ N(T) = 213$$

3.Panel or Longitudinal Data(Minor Cover)

- Time series for each cross-sectional member in the data set, thus data on multiple agents at multiple points in time.
- The same cross-sectional units (individuals, firms, countries, etc.) are followed over a given time period.

$$\{x_{it}, y_{it} \dots\}_{i=1,t=1,}^{NT}$$

- Advantages of panel data:
  - Controlling for (time-invariant) unobserved characteristics
  - Consideration of the effects of lag variables

3.Panel(or Longitudinal) Data(Minor Cover)

TABLE 1.3 Selected Observations on Cigarette Sales, Prices, and Taxes, by State and Year for U.S. States, 1985–1995

Observation Number	State	Year	Cigarette Sales (packs per capita)	Average Price per Pack (including taxes)	Total Taxes (cigarette excise tax + sales tax)
1	Alabama	1985	116.5	$1.022	$0.333
2	Arkansas	1985	128.5	1.015	0.370
3	Arizona	1985	104.5	1.086	0.362
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
47	West Virginia	1985	112.8	1.089	0.382
48	Wyoming	1985	129.4	0.935	0.240
49	Alabama	1986	117.2	1.080	0.334
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
96	Wyoming	1986	127.8	1.007	0.240
97	Alabama	1987	115.8	1.135	0.335
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
528	Wyoming	1995	112.2	1.585	0.360

Note: The cigarette consumption data set is described in Appendix 12.1.

- **Questions?**: observations,variables,the sample size?

$$x_{it} = Total\ Taxes_{it};\ y_{it} = Cigarette\ Sales_{it};\ N \times T = 48 \times 11 = 528$$

4.Pool Cross-Sectional Data(Not Cover)

- Pooled cross sections can be generated by combining two or more years cross-sectional data.
  - Cross-sectional data in each year is independent with other years.
  - While the data come from a same population in different time,the data does not necessarily track the respondent multiple times.
- For it has both cross-sectional and time series features, so allows consideration of changes in key variables over time.
- Simple pooling may also be used when the number of observations of a single cross section is small.
- It is widely used in:
  - Cohort studies
  - Difference-in-differences analyses
  - Cross-sectional analyses

4.Pool Cross-Sectional Data(Not Cover)

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

- **Questions?**: observations,variables,the sample size?

$$x_{ijt} = hprice_{i,1993}, hprice_{j,1995}; y_{ijt} = proptex_{i,1993}, proptex_{j,1995};$$
$$N = N_i + N_j = 250 + 270 = 520$$

Variables or Features in Data

- Variables are the characteristics or properties of the observations in the data set, which can be broadly classified into two types:
- **Quantitative**: continuous or discrete numerical values. These numbers represent a quantity or amount.
  - **Scale**: the difference between two values is **meaningful and consistent** in most cases.
- **Qualitative**: specific numbers used to represent different groups or categories.
  - Here, the scale is mostly **meaningless or arbitrary**.
- The number of variables is the number of columns in the data set,represents by p,
  - which is normally **less than** the number of observations in the **traditional data set**, represents by N

$$p \ll N$$

  - However, in the **big data** era, the number of variables can be **larger than** the number of observations, represents by $p \gg N$

Quantitative Variables v.s. Qualitative Variables

TABLE 7.1 A Partial Listing of the Data in WAGE1

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

- **Quantitative Variables**: wage, education, experience...
- **Qualitative Variables**: female and married...

Features in Big Data

TABLE 14.1 Variables in the 817-Predictor School Test Score Data Set

Main variables (38)	
Fraction of students eligible for free or reduced-price lunch	Ethnicity variables (8): fraction of students who are American Indian, Asian, Black, Filipino, Hispanic, Hawaiian, two or more, none reported
Fraction of students eligible for free lunch	Number of teachers
Fraction of English learners	Fraction of first-year teachers
Teachers' average years of experience	Fraction of second-year teachers
Instructional expenditures per student	Part-time ratio (number of teachers divided by teacher full-time equivalents)
Median income of the local population	Per-student expenditure by category, district level (7)
Student–teacher ratio	Per-student expenditure by type, district level (5)
Number of enrolled students	Per-student revenues by revenue source, district level (4)
Fraction of English-language proficient students	
Ethnic diversity index	
+ Squares of main variables (38)	
+ Cubes of main variables (38)	
+ All interactions of main variables ($38 \times 37/2 = 703$)	
Total number of predictors $= k = 38 + 38 + 38 + 703 = 817$	

Data Types and QSS Methods

Causal Inference(因果推断)

- Cross-Sectional
- Pool Cross-Sectional
- Short Panel(large N, small T)

Times Series Analysis(时间序列分析)

- Times series
- Long Panel(small N, large T)

Big Data Econometrics(Machine Learning Econometrics)

- High-dimensional data(**larger p**, **large N**, **large T**)
  - eg. variables selection to avoid the curse of dimensionality
- Unstructured data(**unstructured N**, **unstructured p**, **unstructured T**, **many missing values**)
  - eg. extract information from text data, image data, audio data, video data, etc.

Data Sets in International Political Economy

- Topic: Measuring democracy
  1. V-Dem(Varieties of Democracy)
  2. Freedom House(Freedom in the World)
  3. Polity IV
  4. EIU(Economist Intelligence Unit)
- Topic: Conflict, War and Terrorism
  1. Correlates of War Project
  2. Uppsala Conflict Data Program
  3. PRIO(The Peace Research Institute Oslo)
  4. Global Terrorism Database
- Individual Data:
  - World Value Survey(WVS)
  - The General Social Survey(GSS)
- Comprehensive Data Platforms:
  - Harvard Dataverse
  - ICPSR(Inter-university Consortium for Political and Social Research)
  - Dataverse

Typical Data sets in China

- Survey Data
  - China Family Panel Survey(CFPS)
  - China Health and Retirement Longitudinal Study(CHARLS)
- Administrative Data:
  - Census: 全国人口普查数据；全国1%人口抽样调查
  - China Industrial Survey Data: 工业企业数据库
  - Chinese Custom Transaction Data 海关交易数据库
  - 全国工商企业登记数据库
- Online Big Data:
  - Transaction data on Tabao,JD,Tmall(淘宝、京东、天猫...）
  - Movie Data on Douban.com(豆瓣\猫眼电影数据)
  - Night-Lights Data(夜间灯光数据) and Air Quality: PM2.5(空气质量数据）
  - Land Transaction Markets(土地交易市场数据）
  - Geolocations Data(地理位置数据）: Baidu Map, Didi, Mobike, Ofo...
  - Social Media Data(微博、微信、知乎、豆瓣、贴吧、论坛、博客、新闻、评论、问答、社交网络...）

Homework(not required)

Homework 1: Household Surveys in China

- Register on one of the following database websites and download the data set.
  - China Household Income Project(CHIP):中国居民收入调查
  - China Health and Nutrition Survey(CHNS):中国健康与营养调查
  - China Family Panel Survey(CFPS):中国家庭追踪调查
  - China Health and Retirement Longitudinal Study(CHARLS):中国健康养老追踪调查
  - Chinese General Social Survey(CGSS):中国综合社会调查
  - China Labor-force Dynamics Survey(CLDS):中国劳动力动态调查
  - China Household Financial Survey(CHFS):中国家庭金融调查
  - China Urban Labor Survey(CULS):中国城市劳动力调查

Homework 1: Household Surveys in China

- Understand the purpose and main content of the sample survey, as well as basic information such as the scope of sampling, the method, and the sample size. Determine the data structure to which this data belongs.(了解抽样调查的目的和主要内容，以及抽样范围、方式、样本量等等基本信息，判断该数据属于哪种数据结构?)
- Download the survey questionnaire and comprehend the specific information it contains.(下载调查的问卷，详细了解调查有哪些具体的信息)
  - Identify the questions you are interested in and locate them in the questionnaire.(首先确定自己感兴趣的问题，然后到问卷中去寻找)
  - Alternatively, review the questionnaire first and find the specific information of interest.(或者先看问卷，找到自己感兴趣的具体信息)
- Download the corresponding data and perform preliminary data cleaning and statistical analysis (to be completed after the computer class)(下载相应数据，进行初步的数据清理和统计分析）
- Prepare for the research proposal project at the end of semester(为期末的研究项目做准备）

Homework 2: IT Preparations

- Ensure you have a **stable and reliable** INTERNERT connection.
- Install Git on your computer.
- Sign up for GitHub as a student and send your GitHub ID to our TA.
- Install R and RStudio on your computer.
- If you have any questions, please feel free to ask our TA or me.