Data Science & AI for Economists

Lecture 0: Introduction

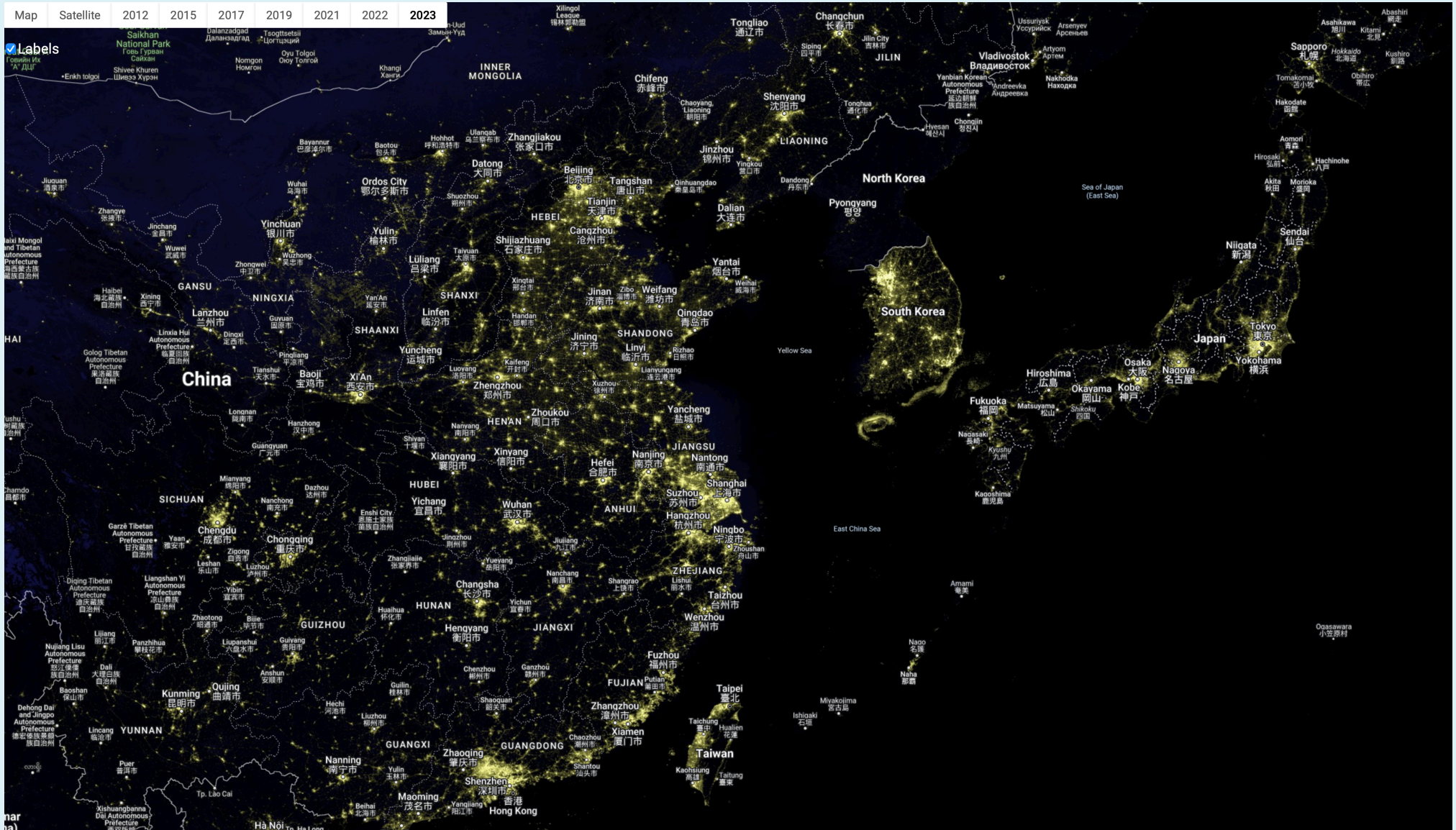
Zhaopeng Qu
Business School,Nanjing University
September 19 2024

Introduction to Data Science and AI

Case #1: Alternative economic indicators

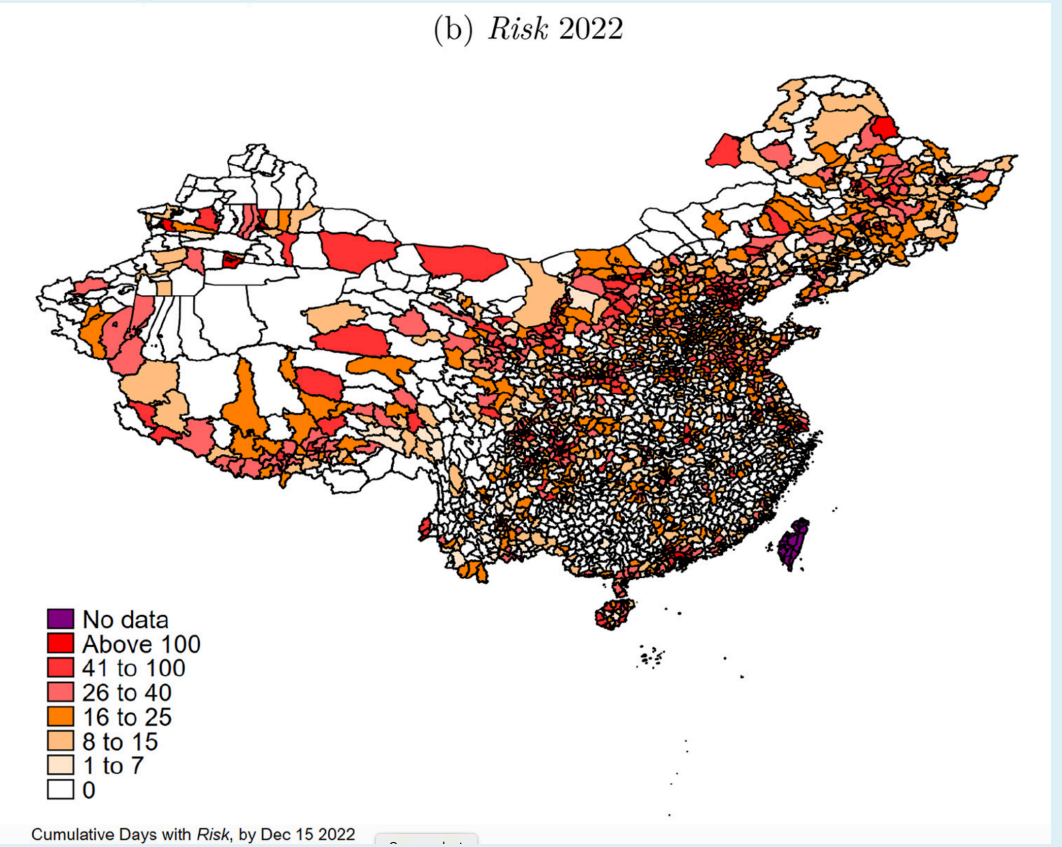
- **Question**: How do we measure a country's economic performance without published or convincing statistic, like North Korea?
- **Answer**: Use alternative indicators, such as satellite images of nighttime lights, to estimate economic activity.

Case #1: Alternative economic indicators



Case #2: Measuring the intensity of a policy

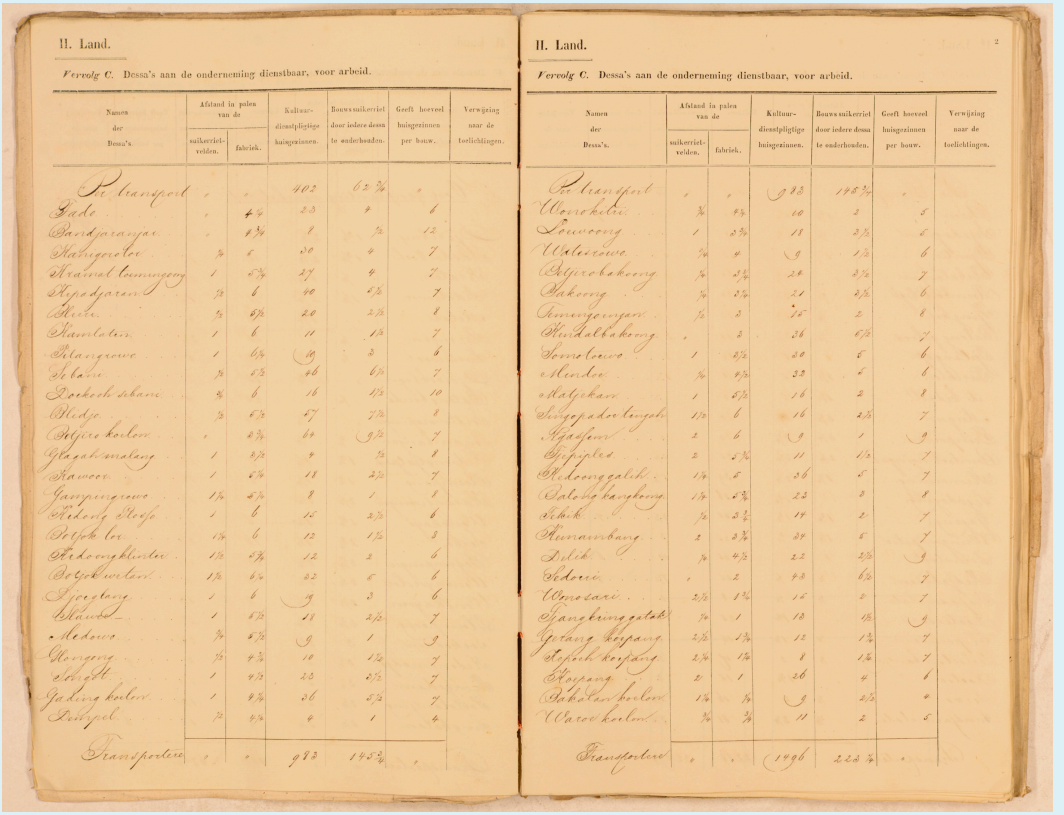
- How to measure the intensity of a policy, such as the COVID-19 lockdown policy in China?
- As is well known, there is huge difference for the intensity of the COVID-19 lockdown policy in China across regions and time.



Gong et al(2024)

Cases #3: Collecting information from documents

- Information (data) recorded for a long time in terms of documents, papers, books and other forms.
- Traditional way to collect data from old documents is by-hands.


II. Land.

Vervolg C. Dessa's aan de onderneming dienstbaar, voor arbeid.

Namen der Dessa's.	Afstand in palen van de suikerriet-velden.	Afstand in palen van de fabriek.	Kultuur-dienstpligtige huisgezinnen.	Bouwsuikerriet door iedere dessa te onderhouden.	Geeft heeren huisgezinnen per bouw.	Verwijzing naar de toelichtingen.
De transport	"	"	402	62 ¾	"	
Dado	"	4 ¼	23	4	6	
Bandjaranjar	"	4 ¾	8	½	12	
Kahrigerolor	¾	6	30	4	7	
Kraton Lamongong	1	6 ¾	27	4	7	
Kepodjaran	¾	6	40	5 ½	7	
Sleko	½	6 ½	20	2 ½	8	
Tamboton	1	6	11	1 ½	7	
Pelangrowo	1	6 ¼	19	3	6	
Sobaru	¾	6 ½	46	6 ½	7	
Dorkoh sbaru	¾	6	16	1 ½	10	
Bledjo	½	6 ½	57	7 ½	8	
Kepoh koelon	"	6 ¾	64	9 ½	7	
Glagahmalang	1	6 ½	4	½	8	
Kawoor	1	6 ¾	18	2 ½	7	
Gampingrowo	1 ¼	6 ¾	8	1	8	
Kedondj Klosjo	1	6	15	2 ½	6	
Pelpoh loe	1 ¼	6	12	1 ½	8	
Kedoengblimbing	1 ½	6 ¾	12	2	6	
Pelpoh wetan	1 ½	6 ¼	22	6	6	
Djogdjang	1	6	9	3	6	
Klaras	1	6 ½	13	2 ½	7	
Medawo	¾	6 ½	9	1	9	
Klangong	½	6 ¾	10	1 ½	7	
Sraglit	1	6 ¾	23	3 ½	7	
Jabang koelon	1	6 ¾	26	5 ½	7	
Dempel	½	6 ¾	4	1	4	
Transportere	"	"	973	145 ¾	"	

II. Land.

Vervolg C. Dessa's aan de onderneming dienstbaar, voor arbeid.

Namen der Dessa's.	Afstand in palen van de suikerriet-velden.	Afstand in palen van de fabriek.	Kultuur-dienstpligtige huisgezinnen.	Bouwsuikerriet door iedere dessa te onderhouden.	Geeft heeren huisgezinnen per bouw.	Verwijzing naar de toelichtingen.
De transport	"	"	973	145 ¾	"	
Oloerkoelon	¾	7 ¼	60	2	5	
Louwoong	1	5 ¾	18	3 ½	5	
Oldeirowo	¾	4	9	1 ½	6	
Pelgrokakoong	¾	5 ¼	20	3 ½	7	
Pakoong	¾	5 ¼	21	2 ½	6	
Kemingerigan	½	5	15	2	6	
Kedalkakoong	"	5	26	6 ½	7	
Sonsolowo	1	5 ½	20	5	6	
Mendo	¾	4 ½	33	5	6	
Kalpekan	1	5 ½	16	2	8	
Sidjopadoelegak	1 ½	6	16	2 ½	7	
Ngadjoer	2	6	9	1	9	
Sepiples	2	5 ¾	11	1 ½	7	
Kedoenggalih	1 ¼	5	26	5	7	
Salokokangkong	1 ¼	5 ¾	23	3	8	
Kebak	½	5 ¾	14	2	7	
Kemambang	2	5 ¾	24	5	7	
Delik	¾	4 ½	22	2 ½	9	
Sedeni	"	2	43	6 ½	7	
Oloesaré	2 ½	1 ¾	15	2	7	
Sangkringgatak	¼	1	13	1 ½	9	
Gedang koepang	2 ½	1 ¾	12	1 ¾	7	
Sepoh koepang	2 ¾	1 ¾	8	1 ¾	7	
Koepang	2	1	26	4	4	
Pakelan koelon	1 ¼	¼	9	2 ½	9	
Oloerkoelon	¾	5 ¾	11	2	5	
Transportere	"	"	1496	223 ¾	"	

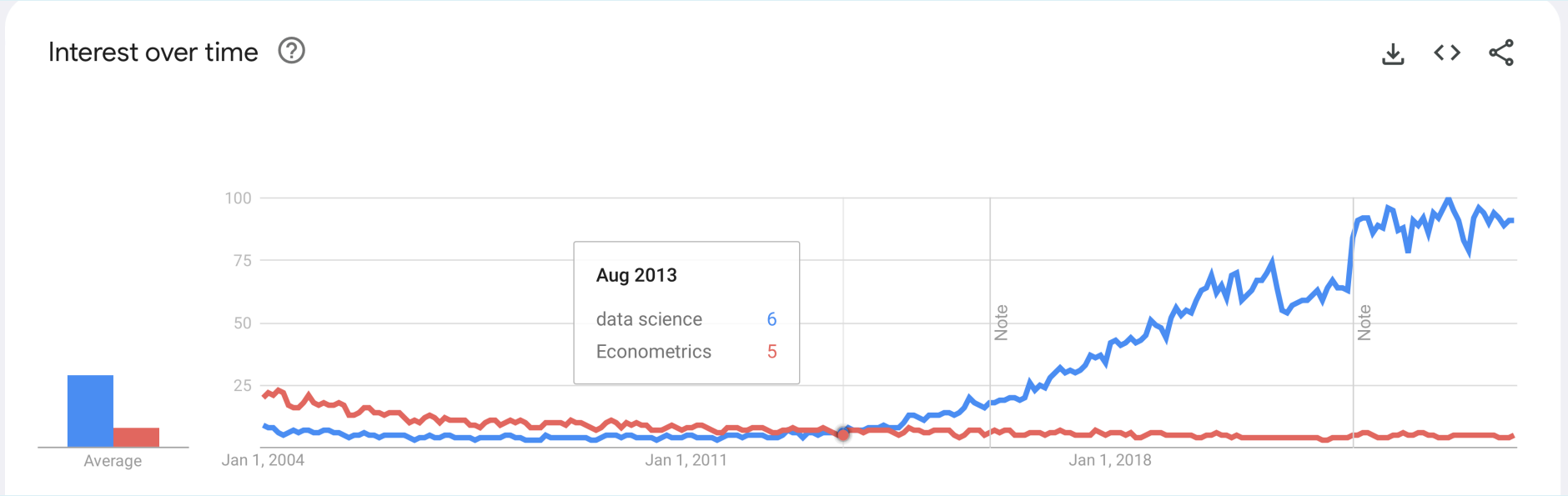
Dell(2020)

New Questions need new tools and skills

- New data sources and neew data types are emerging, which make social science research more challenging and exciting.
- It requires new tools and skills to obtain, process, analyze and visualize various data.
- It is the job of **Data Science**.

What is the Data Science?

- **Data Science** is an **interdisciplinary** academic field that compromises mainly concepts from statistics, computer science, and information science, etc.
- Although scientist have always relied on data to test their theories and make predictions,the term "Data Science" has only recently become popular during the last decade.



Google Trends

What is the Data Science?

Peter Naur(1928-2016)

- Danish computer scientist,Turing Award winner(2005)
- Firstly coined the term "Data Science" in 1960s.

William S. Cleveland(1943-)

- American Computer Scientist and Statistician,
- Formally defining and naming the field of Data Science in 2001.

What is the Data Science?

Joshua Angrist(1960-)

- American Economist and Professor at MIT
- 2021 Nobel Prize co-winner

> "One thing I always say is(that) **econometrics is the original data science**. Before there was data science, there was econometrics." in an public interview at 2021 on Youtube.

What is the Data Science?

Hadley Wickham(Posit)

- **Committee of Presidents of Statistical Societies(COPSS) Presidents' Award** winner(2019).
- Chief Scientist at **Posit**(Former named as RStudio).
- One of Leading figure in the field of Data Science with some most popular packages in R.
  - **ggplot2** and **tidyverse**.

> "Data science is an exciting discipline that allows you to transform raw data into understanding, insight, and knowledge." in *R for Data Science(2e)*.

What is the Data Science?

- **My Own View**

> "All knowledge and skills of gaining and communicating insights from complex data through digital techniques. It is a blend of principles, algorithms, and systems to extract knowledge and insights from structured and unstructured data."

- Main contents of Data Science:
  - **Data Collection**: surveys, sensors, web scraping and OCR etc.
  - **Data Wrangling** : cleaning, transforming, merging,filtering, aggregating, and summarizing.
  - **Data Analysis** : descriptive statistics, causal inference, and predictive analytics.
  - **Data Visualization** : graphs, charts, and maps.
  - **Data Communication** : reports, dashboards, and presentations
- As a student that analyzes data, most of it you do already. However...

What is Artificial Intelligence (AI)?

- **AI**, thus the *Artificial Intelligence*, is a field of computer science that aims to create machines that can perform tasks that typically require human intelligence.
- **Many Areas**:
  - Machine Learning
  - Natural Language Processing (NLP)
  - Computer Vision
  - Robotics
  - 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.
- It is dramatically changing the lives of human beings, and so is how we do research.

Why Economics need Data Science and AI

Revolutions in Social Science

- Social sciences(*firstly started by Economics*) 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.(Angrist and Pischke, 2010)
  - The revolution started from around the 1990s, pioneering in economics, then spread over to other empirical social sciences such as sociology, political science, education, 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

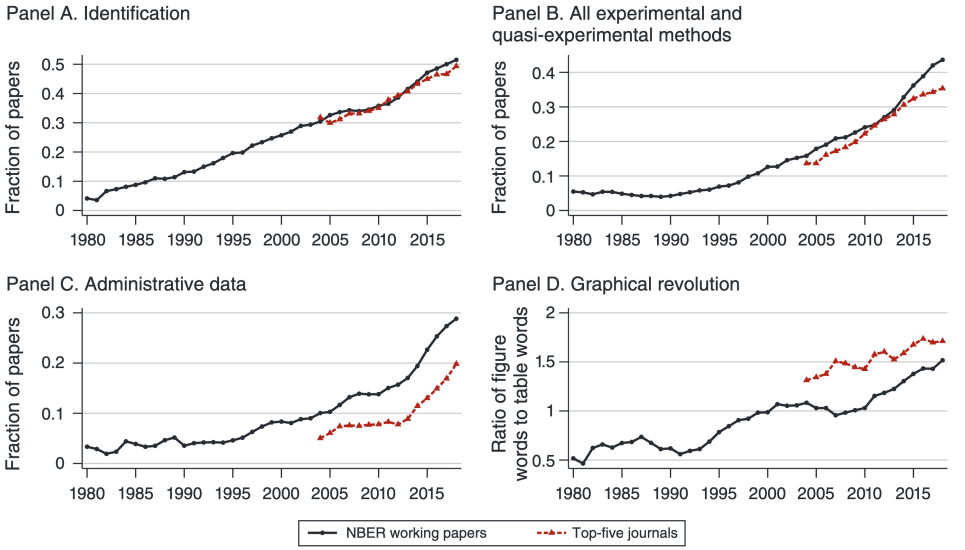


FIGURE 2. THE CREDIBILITY REVOLUTION

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

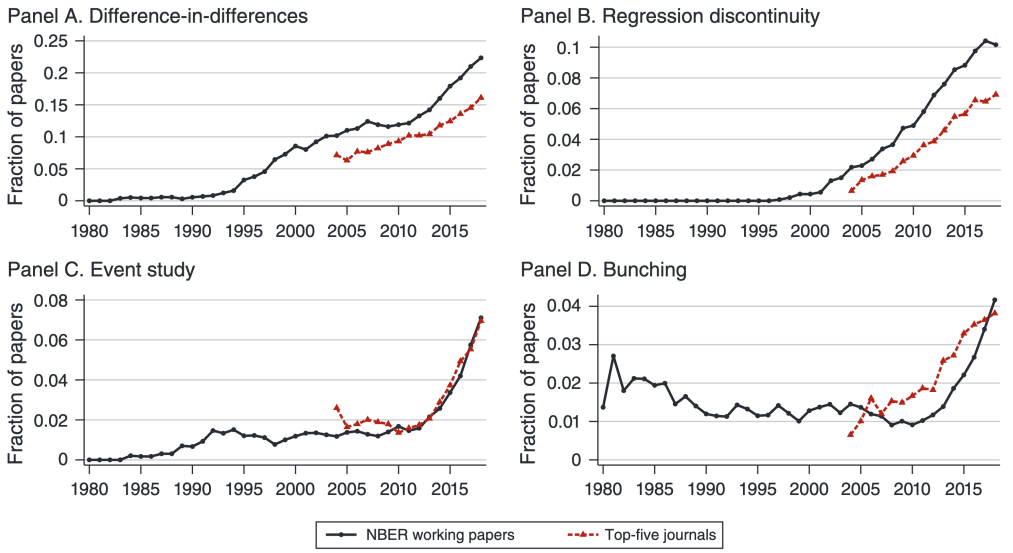


FIGURE 4. QUASI-EXPERIMENTAL METHODS

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

Revolutions in Social Science

- **No.2: Big Data Revolution**
  - A movement that emphasizes that 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, 2013)

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

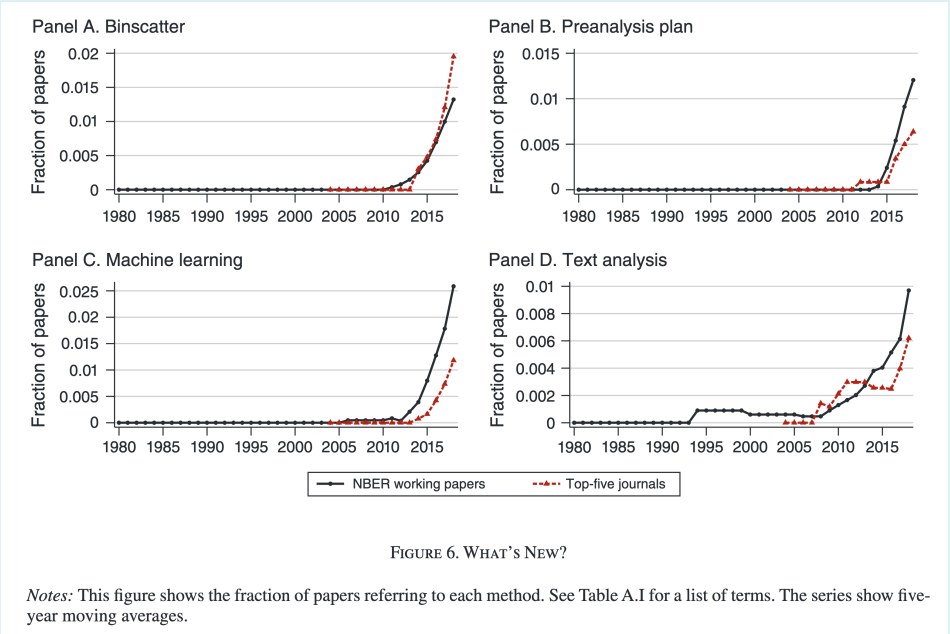


FIGURE 6. WHAT'S NEW?

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)

Revolutions in Social Science

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

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

- The ability of the AI model is still quickly evolving and upgrading.

Why Economics needs Data Science and AI

- Economics and Data Science have many in common
  - require strong analytical skills
  - rely heavily on data and statistical analysis
  - need domain knowledge to interpret results

- The main difference between Economics and Data Science

1. Data collection and processing
   - Econ: more **structured** and **cleaned**
   - DS: more **unstructured** and **messy**

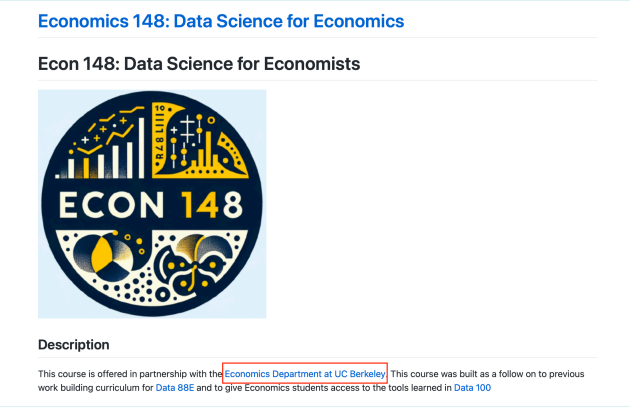
2. methods and tools
   - Econ: more **theoretical** and **causal inference**
   - DS: more **practical** and **prediction**

Why Economics needs Data Science and AI

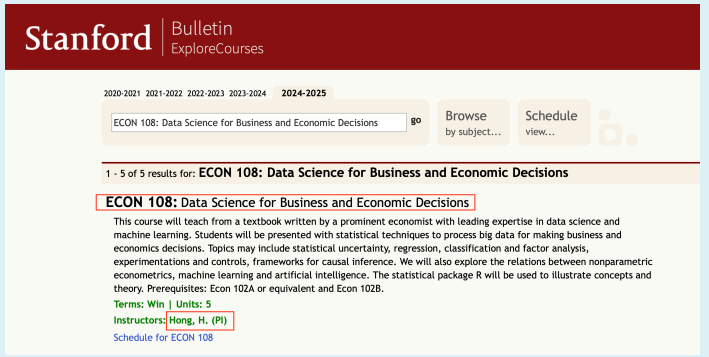
- Data Science can be seen as a **complement** to Economics, providing new tools and methods to analyze data and extract insights that may not be possible with traditional econometric methods alone.
- Economics theory and Econometrics methods can also provide a **foundation** for Data Science, helping to guide the analysis and interpretation of data.
- The combination of Economics and Data Science can lead to more **robust** and **comprehensive** analyses that can provide valuable insights into complex economic and social issues.
- **AI revolution** especially the GAI or LLM models are the result of the developing combination of Social Science, Data Science and Computer Science.
- It is dramatically changing the way we do research in Economics and Data Science
  - With the help with AI tools, we can finish the work in Data Science and Economics more **efficiently and accurately**.

Economics and Data Science Everywhere

- Many double-majored programs in **Economics and Data Science** on both undergraduate and graduate levels have been established in many universities worldwide during the past 5 years
  - "The **most popular** double major in UC Berkeley"
- Many programs in Economics at top universities also provide the courses in Data Science and AI for their students.
  - MIT, Harvard,Stanford, Chicago,and UC Berkeley etc.



@UCBerkeley



@Standford

The Most Popular undergraduate course

Top 3 in Harvard University

- **Economics 10b**: "Principles of Economics"
- **Life Sciences 1b**: "An Integrated Introduction to the Life Sciences"
- **Economics 1152**: "Using Big Data to Solve Economic and Social Problems"

Raj Chetty

Who'd better take the course?

Who' d better take the course?

The Purpose of the course

- Introduce you the foundational concepts of data science and artificial intelligence, emphasizing their **practical applications** in economics and social sciences.
- Unlike traditional econometrics courses, the focus is to learn some new tools and methods that can develop your ability to work with **non-traditional** economic data.
- The course is designed to be accessible to students with a wide range of backgrounds, including those with **little or no prior experience** in data science or economics.
- Help yourself enjoy to learn some new ideas in an empirical social scientist's mindset.

Why take the course?

For those pursuing an academic career:

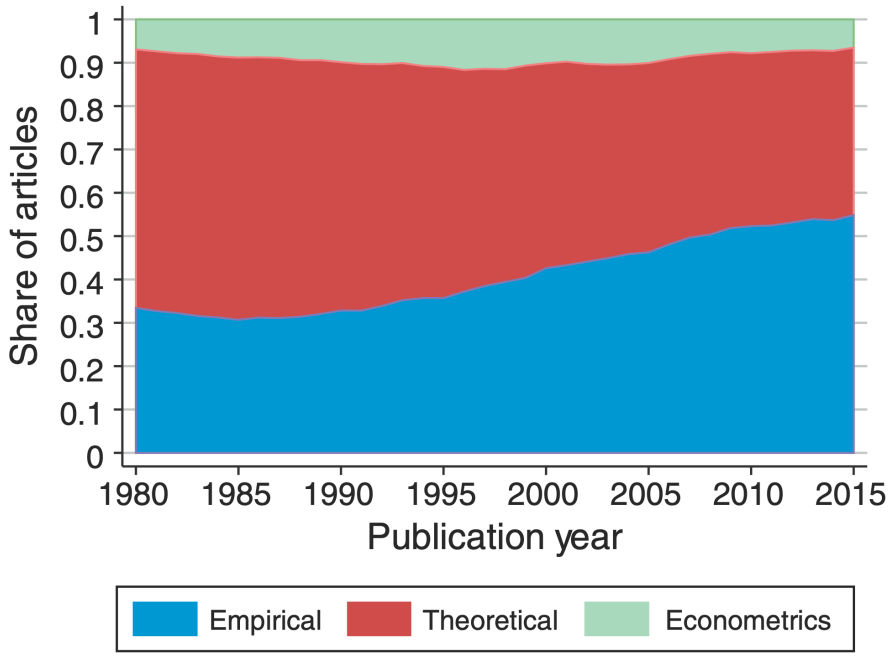


FIGURE 6. WEIGHTED PUBLICATIONS BY STYLE

Angrist et al(2017)

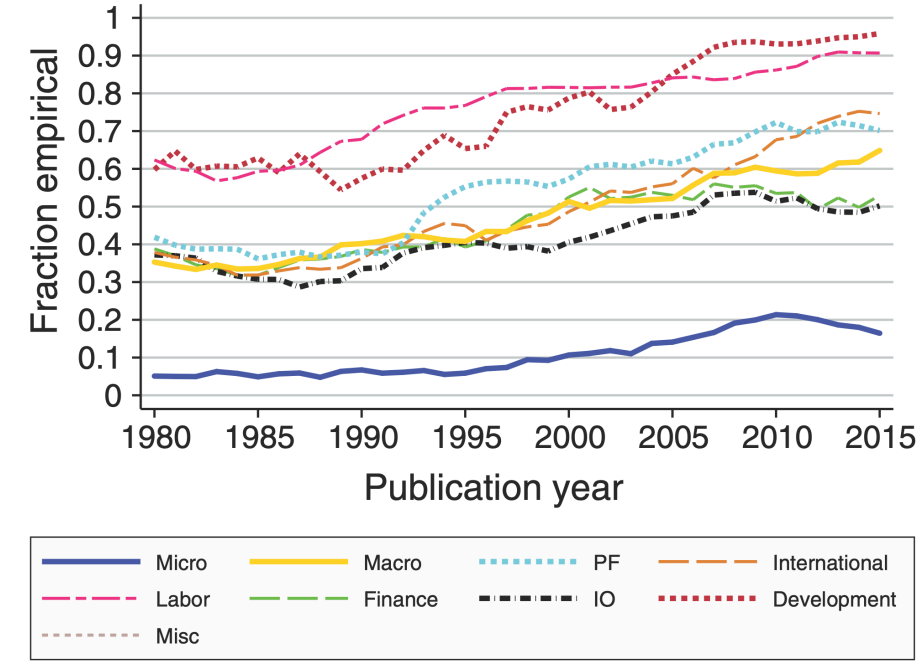


FIGURE 4. WEIGHTED FRACTION EMPIRICAL BY FIELD

Angrist et al(2017)

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

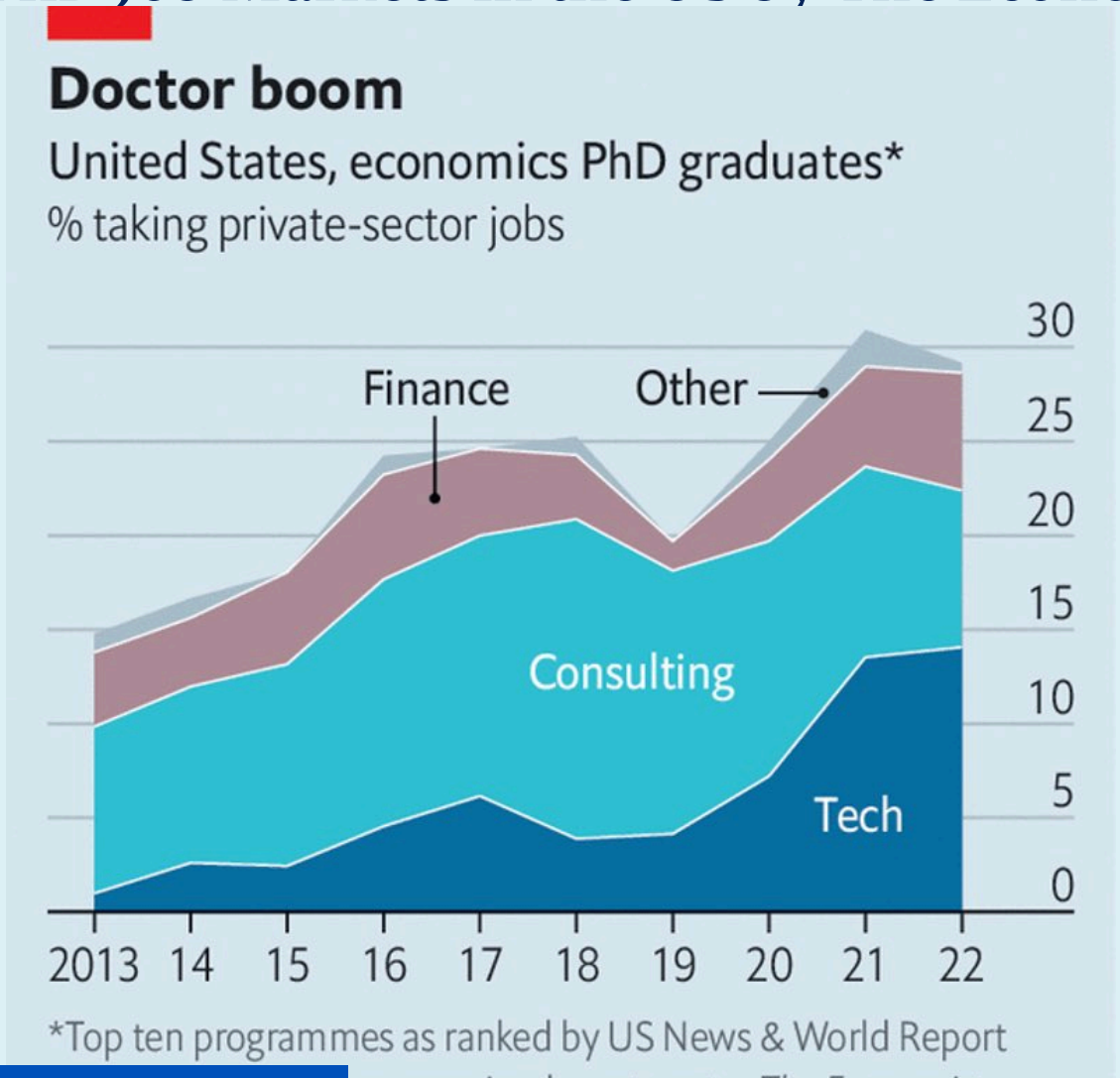
Why take the course?

Who want to enter the 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 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



Why take the course?

Who want to 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 want to 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?

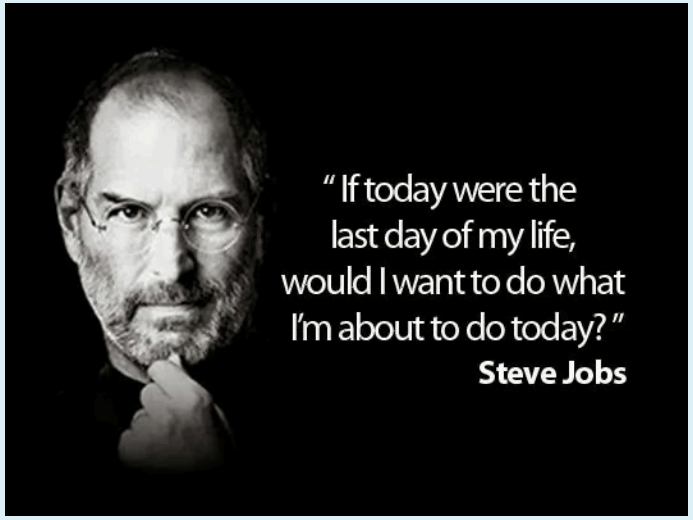
Who find a job in public sectors(shang'an movement)

Source:SCMP

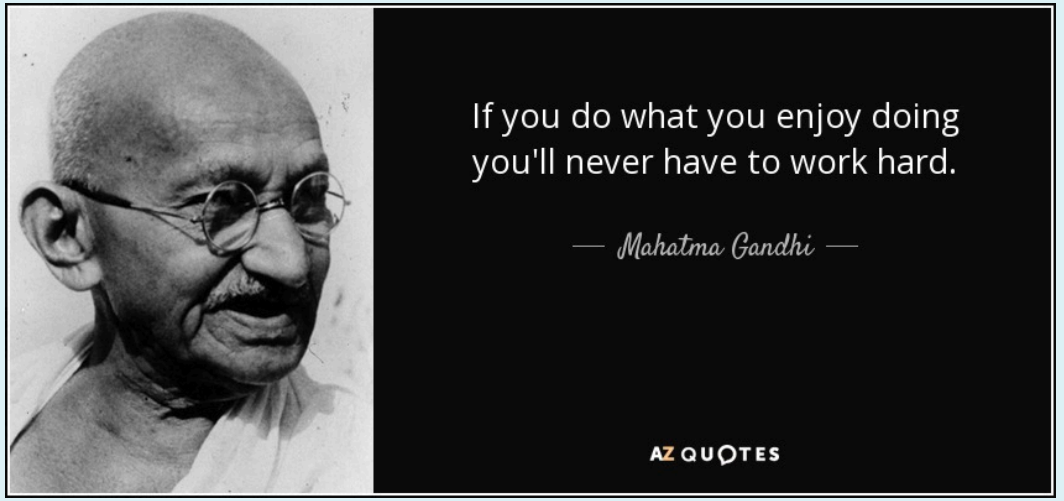
- To be honest with you, the course does not help you succeed in the examination.
- However, in the long run, it provides valuable knowledge and skills that offers a broader vision and abilities.
- It ultimately benefits your career, as well as our people, our country, and the world.

Whoever and Whatever you want to be

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!**

Hard and Soft Skills

You COULD learn or improve several important skills during your college.

- **Hard Skills**
  - **Language**
  - **Computer**
  - **Presentation and Writing**
- **Soft Skills**
  - **Critical Thinking**
  - **Teamwork**
- Fortunately, you could learn/practice almost all above skills in our class.

Wrap up

- In essence, DSAI is an **cutting-edge 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!**

Course Logistics

About Me and the Course

- My name is Zhaopeng Qu(曲兆鹏)
  - Position and Affiliation:Associate Professor, Institute of Population Studies, Business School.
  - Research Fields: Labor Economics and Applied Econometrics
  - Office: Room 2017, Anzhong Building
  - Tel: 83621349;Email: qu@nju.edu.cn
- 南京大学"人工智能"建设系列课程:
  - *专业选修课* for 2nd/3rd year undergraduate students in Economics or other social sciences.
  - Welcome other students who are interested in the course.
  - 2024年秋季第一次开课!, so everything is new and fresh.
- **Course Website**: (https://byelenin.github.io/DSAI_2024/)

Prerequisite and Procedures

- Although there is no formal prerequisite for the course, I **recommend** that you' d better take at least one course in **Statistics**, **Data Analysis** or **Econometrics**.
- And I assume that you should be **comfortable** to **dealing with data** and **coding experiences** by using **Python/R/Stata**.

- The First Part: **Lectures by the instructor**
  - Introduce the underlying theoretical concepts briefly and focus on applications heavily.
  - Provide some specific examples in classical papers in the topics.
  - Normally, everyone who take the course is **required attending the class**.
- The Second Part: **Coding practices in and after the class**
  - Hands-on coding sessions to apply theoretical concepts.
  - Use of real-world datasets for practical experience.
  - Encourage collaboration and discussion among peers.

Course Overview

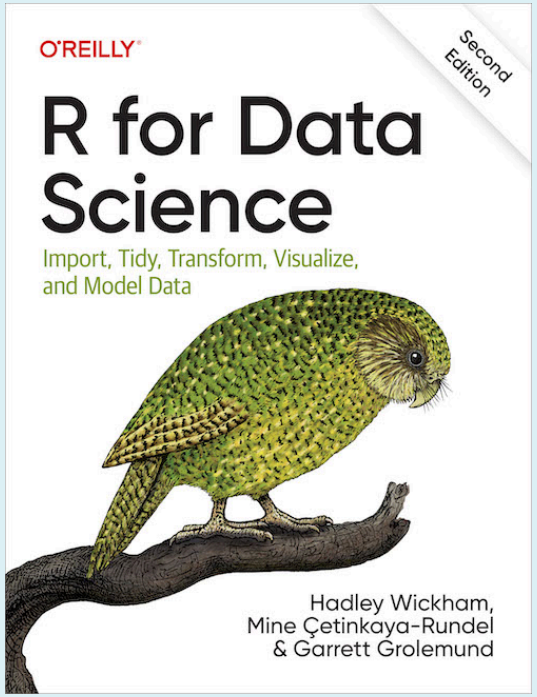
- Basic Tools
  - Github,IDE and AI tools
- Introduction to R and Python for Data Analysis
  - Data wrangling, visualization, and analysis
- Topic1: Spatial Data Analysis
  - Mapping, geocoding, and basic spatial analysis
  - NASA: Nightlight Data
- Topic 2: Web Scraping and Text Analysis
  - Scraping websites, analyzing text data
  - Text recognition and sentiment analysis
- Topic 3: OCR and Image Analysis
  - Extracting text from images, analyzing image data
  - Image recognition and classification

Evaluation

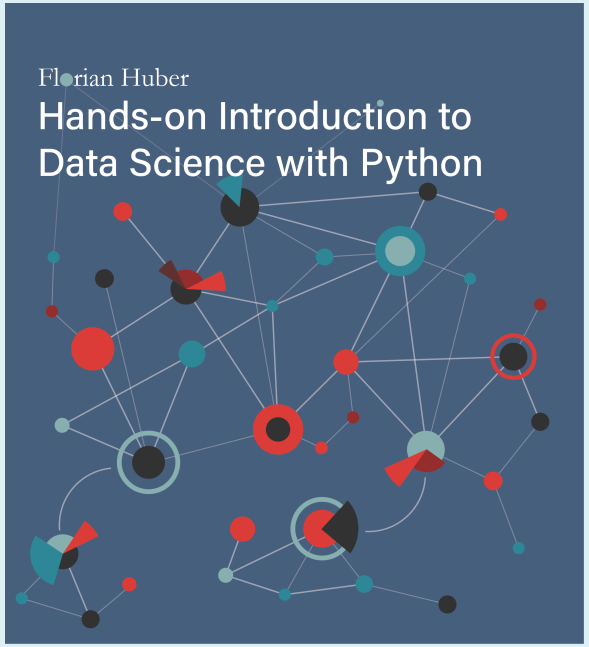
- The final grade will be based on the following components:
  - **Class Participation(10%)**
  - **Proposals Presentation(40%)**
  - **Team Project Presentation(30%)**
  - **Final Report(20%)**
- About **Team Projects and Proposals**
  - Students are required to form a team of **2-3** members to work on a research proposal.
  - **Midterm Presentations**: a presentation of the proposal.
  - **Final Presentations**: a presentation of the preliminary results.
  - **Final Report**: a final report of the project.

Textbooks

- There are **NO** required textbooks for the course.However, the following textbooks are useful for reference:



- R for Data Science by Hadley Wickham,Mine Cetinkaya-Rundel and Garrett Grolemund(2e)



- Hands-on Introduction to Data Science with Python by Florian Huber

Computing Tools

- The main computing tools used in the course are **Python/R**, optionally.

- **R**
  - Pro
  - Con
- **Python**
  - Pro
  - Con

- If your coding experience is limited to **Stata**, there's no need to worry.
  - I am currently learning **Python** and **R** as a beginner, and we can learn together.
  - Additionally, with the assistance of **AI tools**, anyone can become a proficient coder.

Promise and Expectation

What I promise to offer you

- Prepare lectures as well as possible.
- Provide timely feedback on your projects.
- Provide additional resources and references.

- **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.
  - More practices, more scores!

Two Iron Rules

- Don't ever cheat on your assignments!

- Don't ever snitch your teachers to help political repression!

Welcome contact me

An Introduction to Economic Data

Two Axioms of Data Analysis

- **Axiom 1**: Any economy can be seen as a **stochastic process** governed by a certain probability law.
  - The economy's future state is not deterministic but can be described in terms of probabilities.
- **Axiom 2**: Economic phenomena, often summarized in form of data, can be interpreted as a **realization** of this stochastic data generating process.
  - By studying historical data, we can infer patterns, trends, and the probability distributions that describe the stochastic process, thereby gaining insights into the economy's underlying dynamics.
- It highlights the importance of probabilistic models in economics and provides a theoretical basis to use statistical tools and models to analyze 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: 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

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 Data

- **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

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

- 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?

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

Data Types and Sub-Econometrics

Micro-Econometrics(微观计量经济学)

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

Big Data

- Large N and T
- High Frequency Data
- Large P thus the number of variables

Macro-Econometrics(宏观计量经济学)

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

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

- Find a stable and reliable internet connection.
- Sign up to GitHub as a student and install Git on your computer.
- Then send your Github ID to me.