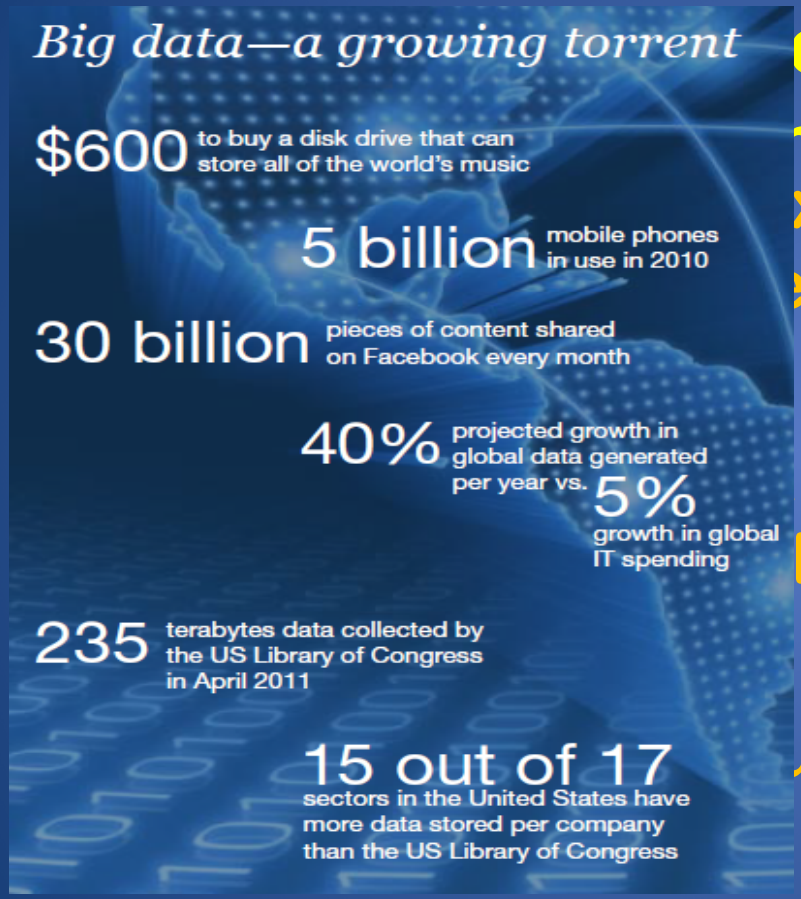
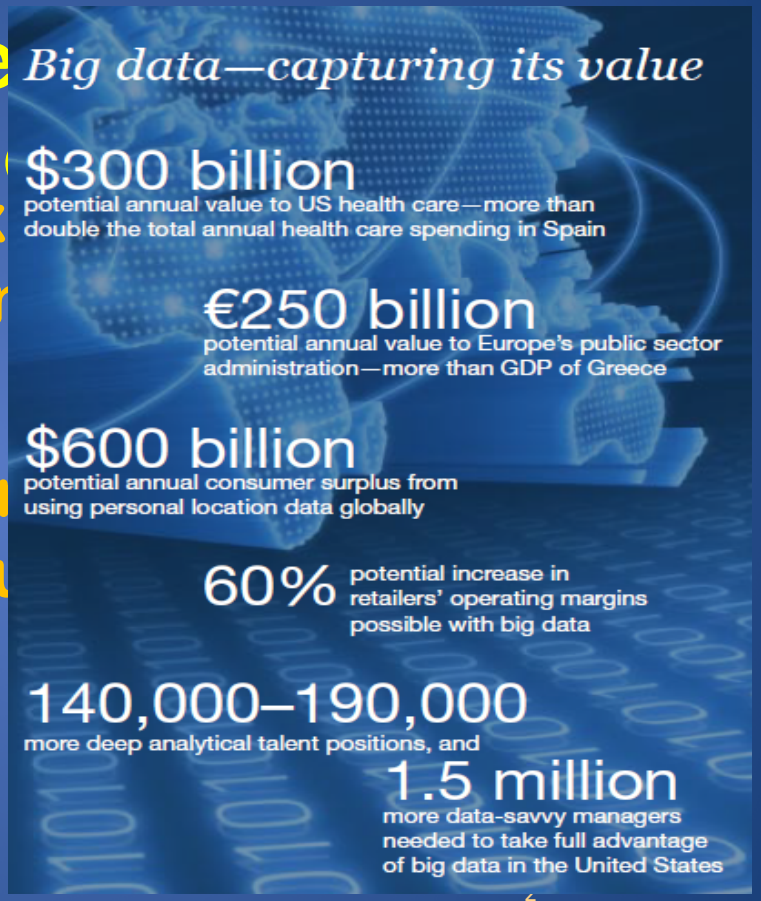
大数据时代的管理决策

Lecture 0: Introduction to Big Data Analytics

Zhaopeng Qu
Nanjing University Business School
March 08 2024

Big Data Create Great Business


Big data—a growing torrent
$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated per year vs. 5% growth in global IT spending
235 terabytes data collected by the US Library of Congress in April 2011
15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

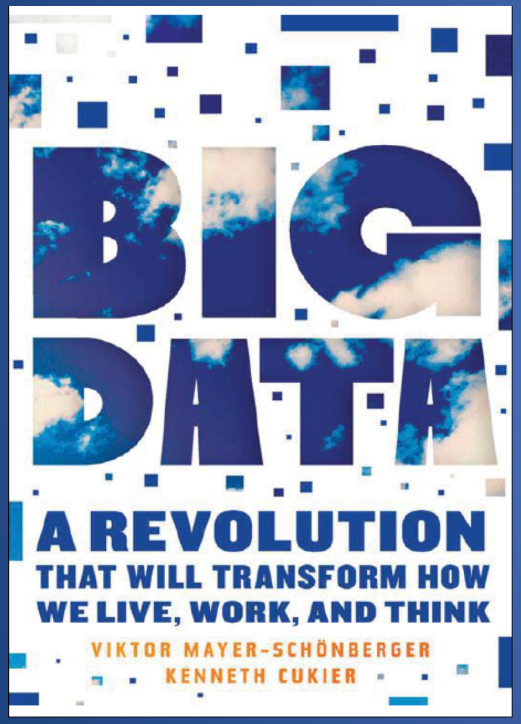

Big data—capturing its value
$300 billion potential annual value to US health care—more than double the total annual health care spending in Spain
€250 billion potential annual value to Europe's public sector administration—more than GDP of Greece
$600 billion potential annual consumer surplus from using personal location data globally
60% potential increase in retailers' operating margins possible with big data
140,000–190,000 more deep analytical talent positions, and
1.5 million more data-savvy managers needed to take full advantage of big data in the United States

What do we mean by “big data”

McKinsey&Company

- “Big data” refers do datasets whose size beyond the ability of typical database software tools to capture, store, manage and analyse.

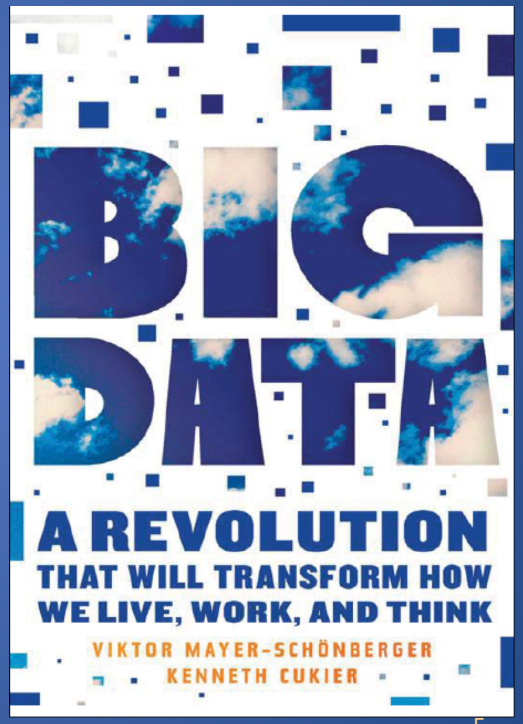
What do we mean by “big data”



- Big data refers to things one can do at large scale that cannot be done at a smaller one. Mayer-Schonberger(2013)

How Big Data Changed us?

- 大数据更多意味着
  - 创新商业模式和竞争规则
  - 促进政府管理的变革
  - 创造新知识和新思维



Why big data?

- The Planets Are Ready for Analytics
  - Powerful IT
  - Data critical mass
  - Skills sufficiency
  - Business need

Why big data?

- Very Powerful IT Infrastructure
- Moore' rule（摩尔定律）
- Kryder' law (克莱德定律）
- More intelligent softwares

Why big data?

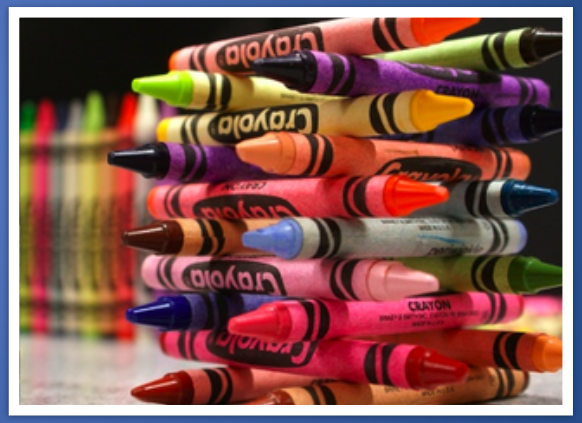
- Mass Data Storage
  - ERP, CRM…
  - Pos Machine
  - Credit Card
  - Club
  - Online Store
  - 共享单车
  - Many many Apps
- Most of them actually are a by-product.

Why Big data?

- Methods
  - Statistics
  - Econometrics
  - Machine Learning

Why big data?

- Updated Business Mode
  - Customers in the centrality or asymmetric to symmetric information
  - The Long Tail Theor（长尾理论）



Analytical Competitors：
Old Hands Polishing Their Edge

- Wal-Mart （沃尔玛）
  - 1990s-“啤酒与尿布”
  - 2000s-“飓风与Pie”
- Target (塔吉特)
  - 预测怀孕指数

Analytics in Las Vegas

Analytics in Sports

- Baseball
  - Red Socks
- Basketball
  - Huston Rocket
- Soccer
  - AC Millian
- American Football…

Analytical Competitors：Number-Crunchers from Birth

- Netflix — Movie preference algorithms
- Farecast.com—Flight price forecasting（预测机票）

Analytical Users across Industries

Consumer Products
- Kraft（卡夫食品）
- Mars（玛氏）
- P&G （宝洁）

Financial Services
- Barclay（巴克莱）
- Capital One

Telecom
- O2

Industrial Products
- Deere(约翰迪尔)
- Cemex（西麦斯）

Retail
- Tesco
- Wal-Mart

Transport
- FedEx(联邦快递)
- UPS

In China：京东的实践

- 市场推广活动：定向营销

用户 A： 男性、28 岁、北京、累计购买金额 13428 元、没有投诉记录、最近 2 个月购买过 ipad4 MD513CH，购买过图书三体，搜索过莫言、剃须刀、HDMI 转接线、手机等关键词，关注 Sony KDL-46HX750 3D LED 液晶电视，促销偏好度高.....

用户 B： 女性、33 岁、上海、累计购买金额 3420 元、曾有过投诉记录，记录关键词为安装慢、退货等，近 2 个月购买过 ONLY 圆领立体剪裁无袖修身连衣裙 E(黑)，蓝月亮亮白增艳自然清香洗衣液 3000g，关注飞利浦 PT720 三刀头电动剃须刀，搜索过雅培、多美滋，促销偏好度低.....

用户 C：

京东商城要做红酒专场活动，请问上述哪个用户更可能是目标客户群。

In China：京东的实践

模型的线下测试效果

- 涉及用户数：9832608
- 购买概率大于 0.34 用户数：303641
- 未来 5 天实际购买用户数：14290
- 预测命中用户数：10337

对用户： 最小程度地打扰客户，提高客户体验

对企业： 减低营销成本，提高客户忠诚度

In China：京东的实践

筛选的客户我们还需要做以下工作

渠道： 网页直接推荐、邮件推送（提醒）、移动客户端推荐、短信告知、站内提醒

时间： 工作日、周末、节日、日间、晚间等

方式： 直减、满减、活动、优惠券、捆绑销售等

Government and Public Service

- Government
  - Mexico-(Progresa: Program for Education Health and Nutrition)
- School:
  - STAR in US
- Hospital:
  - The debate of evidence based medicine(循证医学)

In China

- “The world's biggest camera surveillance network” and “Social

Conclusion

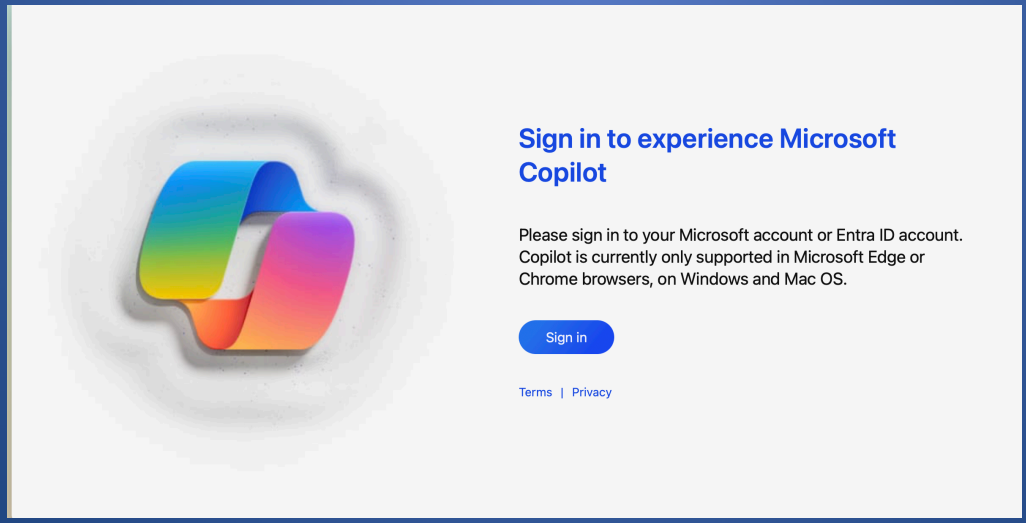
- 今天，各种组织都在搜集目标客户的信息和数据，并以科学客观的技术来分析，以便更加准确、严谨、深刻地做出决策并掌握先机。

- 无论企业还是政府、医院、学校……，仅仅靠直觉和经验决策的时代已经一去不复返了，在新一代的商业经营中，今天的管理者比以往任何时候都应该更倚重数据分析。

- 越来越多的企业在广告、定价、产品性能及雇佣政策方面大量而广泛地采用数据分析，新一代的管理决策者必须具有现代数据分析的思维，而这种利用科学方法来分析数据的管理决策思维正在掌握和改变这个世界。

The Revolution of AI

ChatGPT



What is Competing on Analytics?

- “数据竞争分析法”： 指广泛地使用数据、统计及计量等定量分析方法，利用解释与预测的模型，来管理商业的经营和制定决策方案的竞争方法。

- **In General, Searching for Business Logic from Data and Make Decisions based on it.**

Business Intelligence & Data Analysis

Why Competing on Analytics

- 数据分析法可以用在几乎所有的业务流程中
  - 内部流程
    - 财务
    - 生产
    - 研发
    - 人力资源
  - 外部业务
    - 客户
    - 供应商
    - Marketing: 广告、促销

Two Major Missions

1. Analyzing the causals and evaluating the effects.
2. Forecasting the trends and making the predictions.

- Final Goal: optimizing our decisions.

Summary

- 面对大数据时代的挑战，数据竞争分析法是基于业务和战略层面的基本分析方法，是大数据时代的基本商业思维和逻辑。

- 作为新一代的职业经理人 ，要想全面理解数据分析方法
  - 理解数据分析的基本原理：统计、计量和机器学习。
  - 掌握数据分析的基本技能：使用统计软件处理、分析和展示数据的基本动手能力。

Our Goal

- 本课程的目的也许不是直接培养大家成为数据科学家(Data Scientist or Data Analyst)，而是要让大家从因果推断(Causal Inference)为的角度更深刻的理解数据分析原理及在商业中的应用。

- 开阔视野，锻炼动手能力，激发各位的创新性思维。

- 最终希望能够帮助各位在实际工作中有效地利用数据分析思维，提高管理决策的创造性和精确性。

Introduction: A Scientific Framework of Rational Knowledge

Question #1: Student's Performance and Class Size

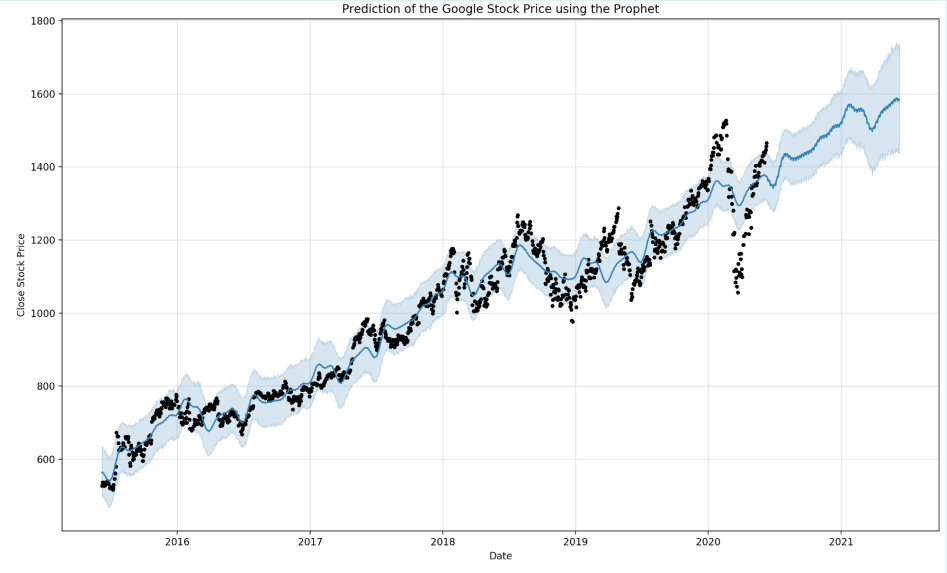
- A Classical Issue in Economics of Education: Is there a gap of students' performance between large-size classes and small-size classes?
- Turn it into an empirical or policy question:
  - What is the quantitative effect of reducing class size on student achievement?
  - Like by 5 student per class? or 10?

Question #2: Discrimination in Labor Market

- What is "discrimination"?
  - **Equal work, unequal pay?**
  - **Equal work ability or productivity, unequal pay?**

- Many types of discrimination in labor market:
  - Racial Discrimination for housing loans.
  - Gender Discrimination in labor market
  - "Hukou" Discrimination in China
- **How to prove the existence of discrimination?**
- **How to quantitatively measure the degree of discrimination?**

Question #3: How Will Stock Market Head?



- It seems that people always want a sneak preview of the future.
  - What will sales be next year at a firm that is considering investing in new equipment?
  - Will the stock market go up next month, and, if it does, by how much?
  - How confident can we be in our predictions?

Questions need both quantity and quality answers

- **Other Similar Questions**:
  - Air pollution and Health?
  - Credit regulation on housing price
  - Coupon on products sales
  - Trade War...
  - Pandemic...
- Living in an unprecedentedly complex and dynamic world, we need to make decisions by
  - **Rational cognition**(理性认知)
  - **Scientific predicition**(科学预测)

A Scientific Framework for Rational Cognition

How to obtain rational knowledge(judgment)?

- Anecdotes(轶事）or Intuition(直觉)
- Theory(理论/逻辑推理）
  - Systematical methodology: Hypothesis, Logical deduction...
- Empirical evidence（经验证据）
  - statistical inference from data.

An Example: Smoke and Mortality

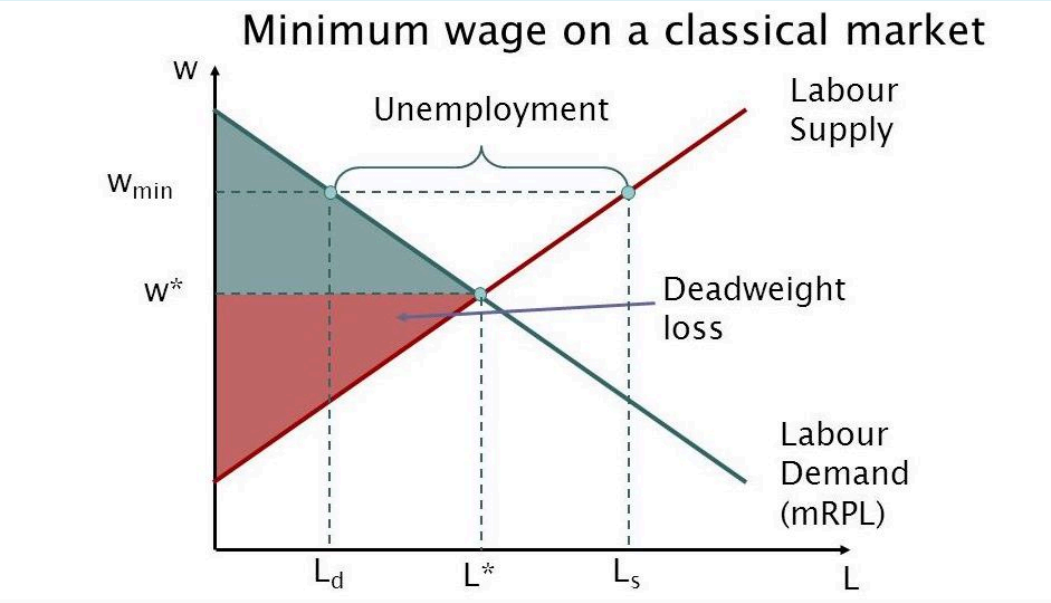
- Anecdotes(轶事）or Intuition(直觉）
  - eg. “My grandmother smoked two packs a day and lived until she was 95 years old.”
- Theory
  - Because Cigarettes contain carcinogens(致癌物) such as nicotine, tar, and formaldehyde(尼古丁、焦油、甲醛等), then...
- Empirics
  - Collecting data through experiments or surveys, and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

Classical Question: Human Capital v.s Signal

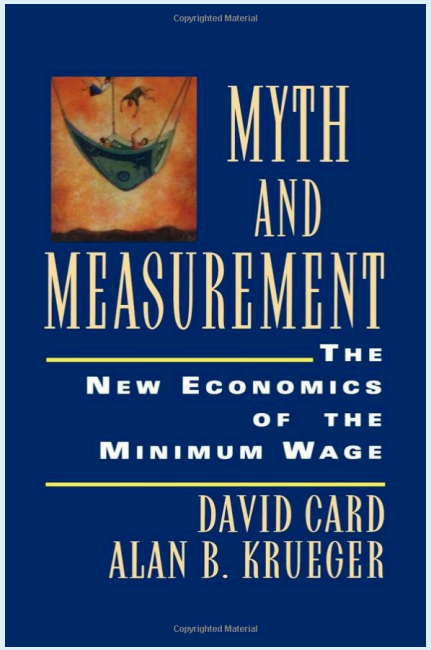
- A common phenomenon in labor markets can be observed across countries.
  - Higher education, Better pay!

- Two classical theories to explain
  - **Human Capital**: Education improves work productivity.
  - **Signal**: Education does not increase the productivity. It simply serves as a signal of the individuals' innate ability.
- **Question**: which one is right?

Public Policy: Minimum Wage and Unemployment



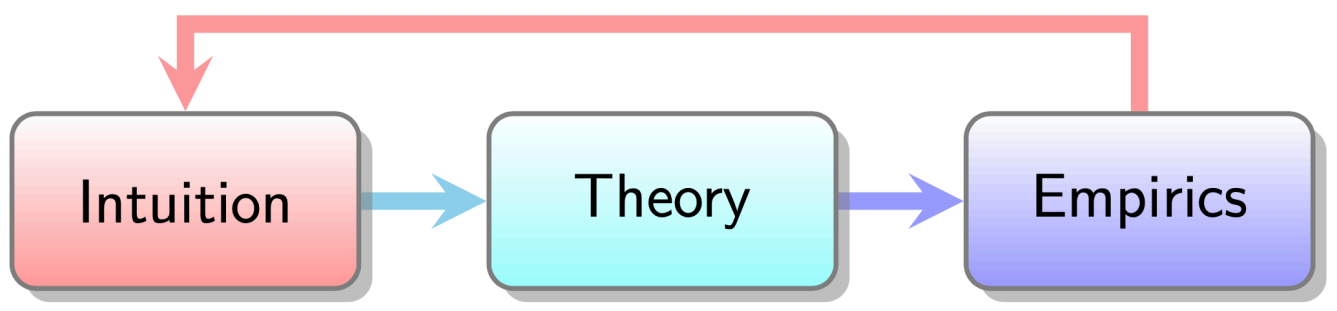
- The classical supply-demand model tell us that
  - Minimum wage will definately increase unemployment.



- One famous empirical evidence challenged the theory by David Card and Alan Krueger(1994)
- They found that increases in the minimum wage do **NOT** lead to job losses.

A Scientific Workflow to Analyzing

- By **Intuition**: Propose meaningful or interesting questions(It does matter or we care about)
- By **Theory**: Obtain a preliminary conclusion or proposal an hypothesis
- By **Empirics**: use data and quantitative methods to test your theory or conclusion.



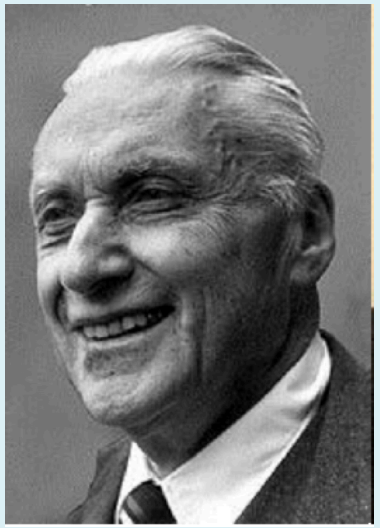
- Once we have a theory (or cause) which has been testified by empirical works, then we can manipulate the cause to obtain the effect.

Quantitative Answers to Quantitative Questions

- Many decisions in economics, business and government hinge on understanding the relationship among variables in the world around us.
  - Economic and manage theories may provide .hi-slate[clues about the direction] of the answer.But making decisions require .hi-pink[quantitative answers to quantitative questions].
- Therefore, we have develop a framework and practical methods that provide:
- **a reasonable answer to the question**
  - whether the effect is ZERO or not?
- **a numerical answer to the question**
  - whether the effect is positive or negative and how large the effect is?
- **a measure of how precise the answer is**
  - how confident we are in the answer

Econometrics and Big Data Analytics

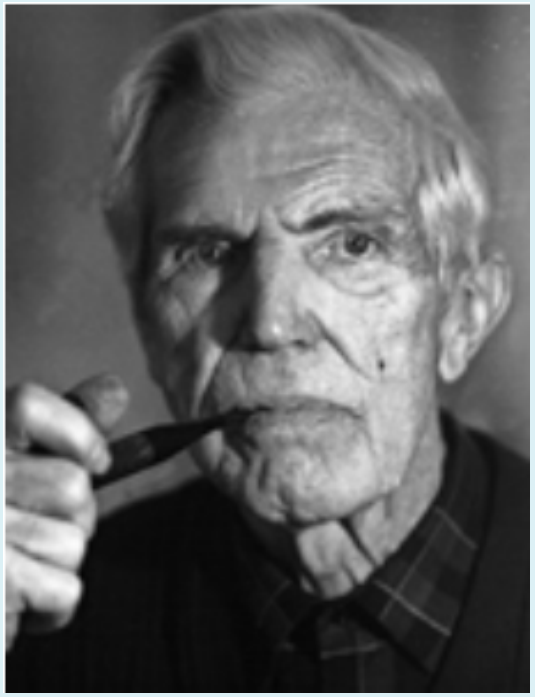
Econometrics: A Brief History



- The term is attributed to **Ragnar Frisch(1895-1973)** who is the 1969 **Nobel Prize** co-winner(the first year for Economics).
- Although the term coins by a combination of economics("Econ-") and metrology("-Metrics"), it is special enough in social science and science at that time.

> "Econometrics is by no means the same as **economic statistics**. Nor is it identical with what we call general **economic theory**, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonymous with **the application of mathematics** to economics. Experience has shown that each of these three view-points, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And **it is this unification that constitutes econometrics**". in *Econometrica*, 1933, volume 1, pages1-2.

Econometrics: A Brief History



- **Trygve Haavelmo(1911-1999)**
- 1989 Nobel Prize winner

> "The method of econometric research aims, essentially, at **a conjunction** of **economic theory** and **actual measurements**, using the theory and technique of **statistical inference** as a bridge pier.". in *Econometrica*, 1944, volume 12, pages1-2.

Econometrics: A Brief History

James Stock(Havard)

Mark Watson(Princeton)

> "Ask a half dozen econometricians what econometrics is—you could get a half dozen different answers.At a broad level, it is a **science and art** of using **economic theory** and **statistical techniques** to analyze **economic data**.", *Introduction to Econometrics*, 4th edition.

Econometrics: A Brief History

- **My Own View**

> "In general, a series of **scientific methods** to searching for economic logic from data.It could include two broad jobs

- Making a **causal inference**
  - Testing economic theories.
  - Estimating causal effects.
  - Policy evaluation.
- **Predicting and Forecasting**
  - 'Causal' prediction
  - Forecasting for future outcomes
- More and more prevalence in
  - other social science such as political science, sociology, law and education studies etc
  - and business practice, like the hottest one: Data Science.

Econometrics,Big Data and Social Science

- Social science(firstly started by Economics) has experienced two methodological **revolutions** over the past few decades. Econometrics has been playing a critical role for revolutions.

- **Credibility revolution**
  - A movement that emphasizes the goal of obtaining secure causal inferences(Angrist and Pischke, 2010)

- **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)

Econometrics,Big Data and Social Science

- 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.

Why and Who should take the course?

Why Econometrics is so important?

- Several Common Questions about Econometrics?
  - Why we should study econometrics?
  - How is studying econometrics helpful in understanding social science?
  - Especially, *can one excel in the economics without learning econometrics?*
- The answer to the last question is simple.
  - NO! because one hardly to learn modern economics without the knowledge of econometrics.
- Econometrics is **one of three core courses** required in almost every economics department worldwide.
  - The other two are Microeconomics and Macroeconomics.

Why Econometrics is so important?

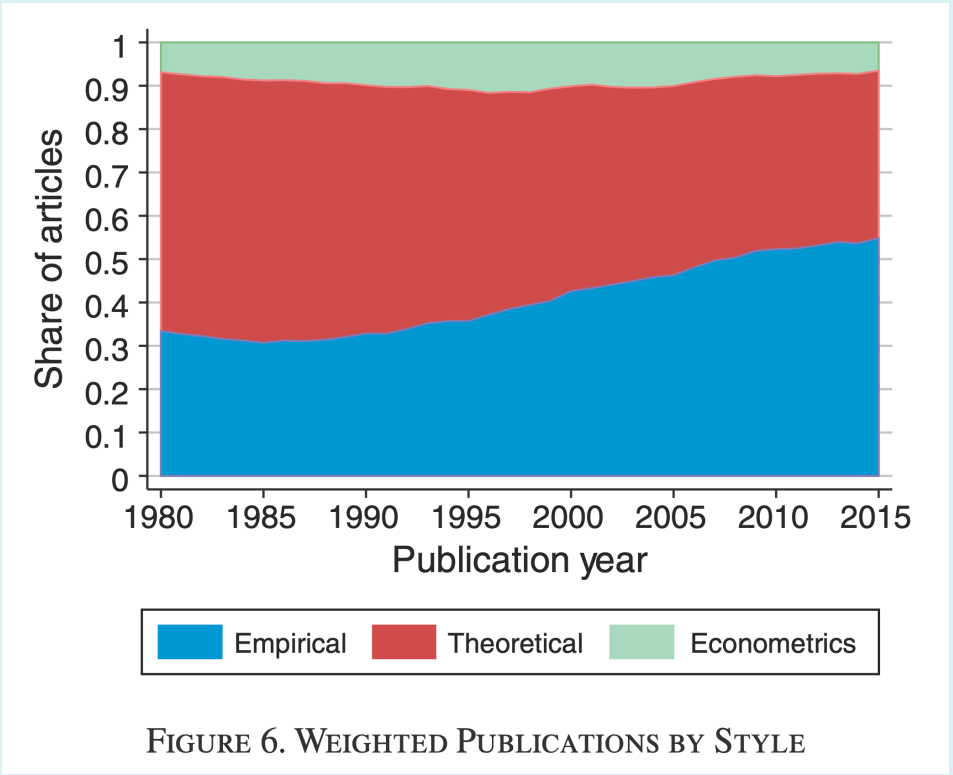


FIGURE 6. WEIGHTED PUBLICATIONS BY STYLE

Angrist et al(2017)

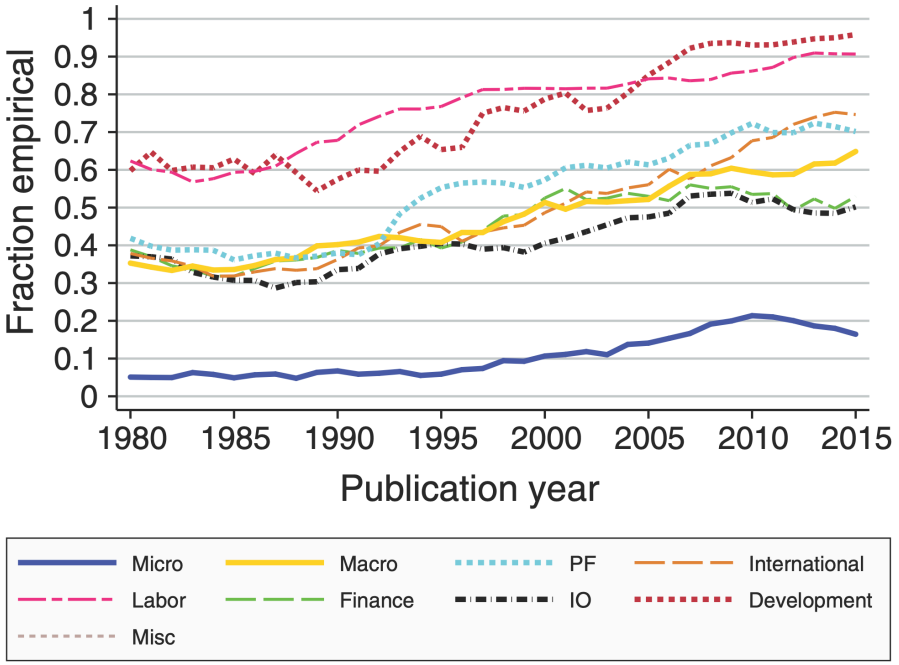


FIGURE 4. WEIGHTED FRACTION EMPIRICAL BY FIELD

Angrist et al(2017)

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

Why Econometrics is so important?

- 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 Econometrics is so important?

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 Econometrics is so important?

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 Econometrics is so important?

Who working in public sectors

- 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.

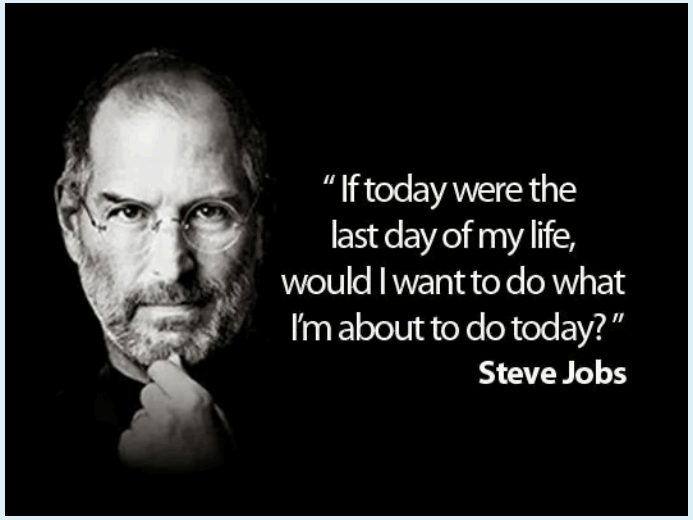
Why Econometrics is so important?

Who look for fun

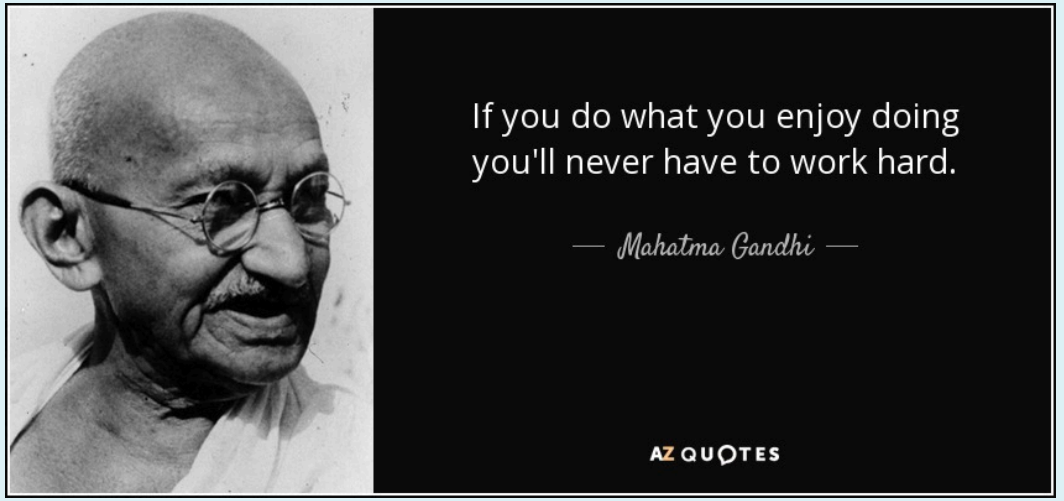
- The course could not be a boring and demanding variant of a mathematics course, but an interesting and enjoyable class.
- Help yourself enjoy life by adopting an empiricist's mindset in your daily activities.
  - novel ideas or new perspectives on our world.
- Also covering several interesting and 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!**

Hard and Soft Skills

You COULD learn or improve several important skills during your graduate studies.

- 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

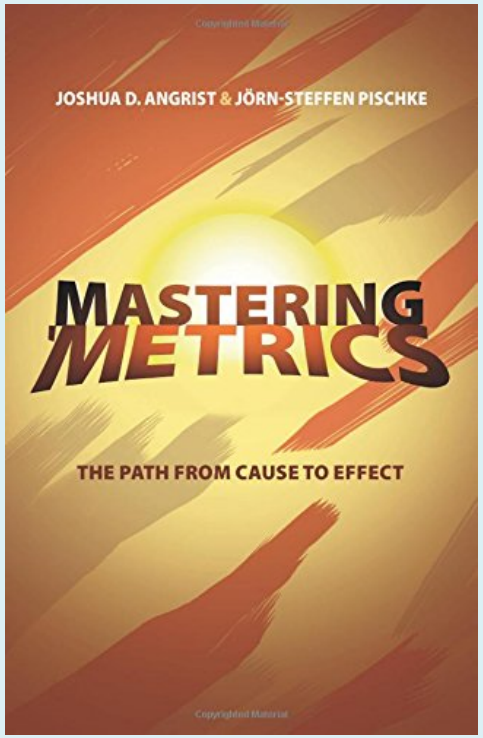
- In essence, **Big Data Analytics** is an **essential and intriguing yet challenging** course.
  - **Please consider carefully before enrolling**
  - **Once committed,please work hard on it!**
  - **And remember, enjoy the process of working hard!**

Course Logistics

About Me, our TA and the Course website

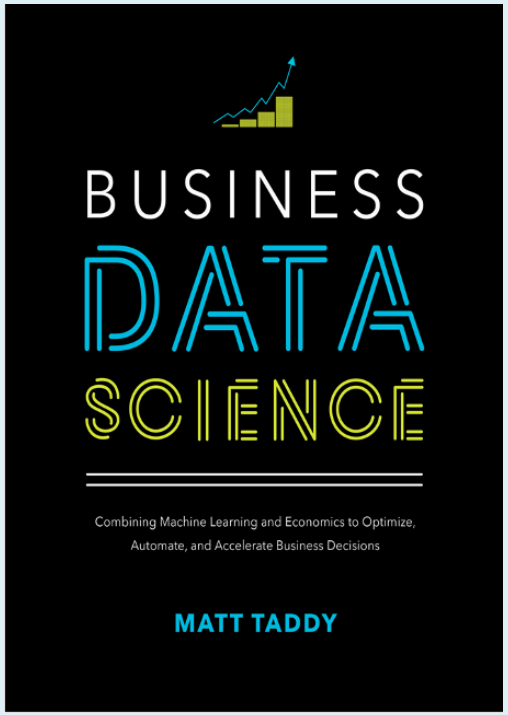
- My name is Zhaopeng Qu(曲兆鹏)
  - Associate Professor, Institute of Population Studies, Business School.
  - Research Fields: Labor Economics and Applied Econometrics
  - Email: qu@nju.edu.cn
- Our TA: Zishu Wang(王紫澍)
  - Graduate students in the first grade, proficient in Stata and R
  - Help you with your team projects
- **Our Course Website**: https://byelenin.github.io/MBA_Big_Data/
  - You could find the syllabus, slides, and other materials on the website.

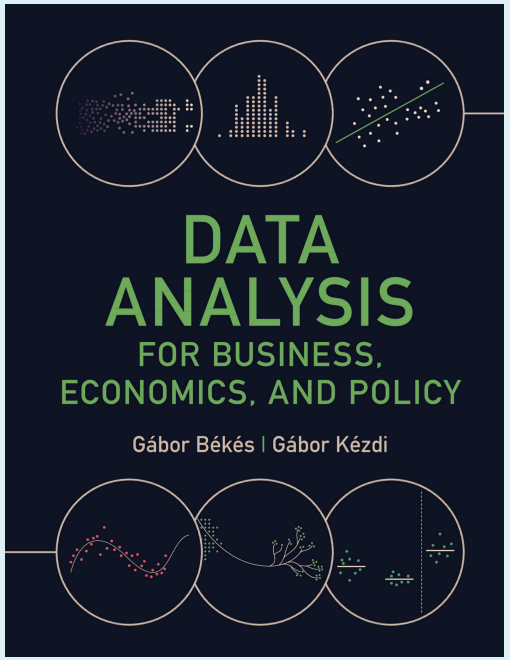
Textbooks



- Joshua D. Angrist & Jorn-Steffen Pischke, (2014). Mastering 'metrics: The Path from Cause to Effect. Princeton University Press. （中译本：《精通计量：从原因到结果的探寻之路》，格致出版社，2018）

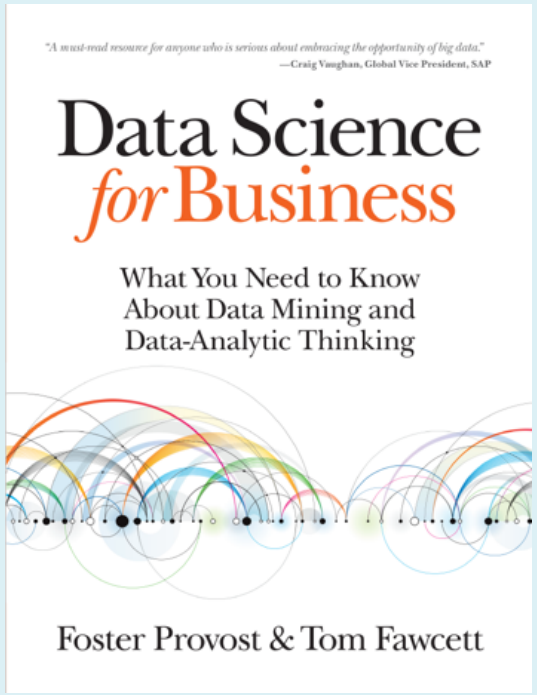
Supplementary Textbooks

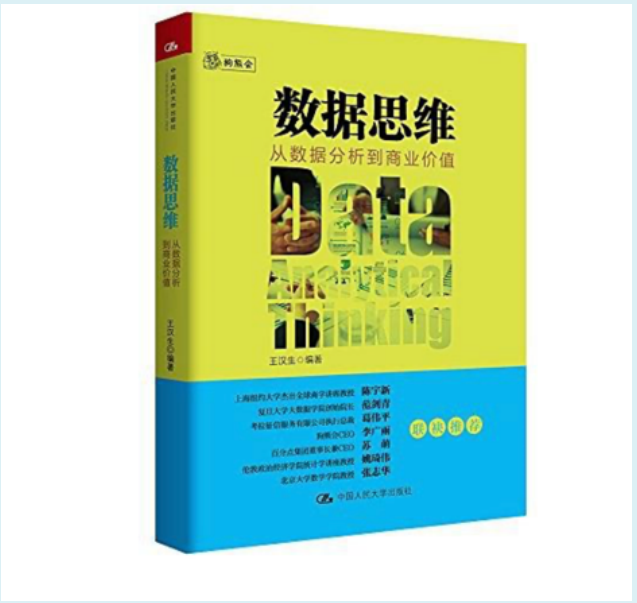




- Matt Taddy(2019),Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions, McGraw-Hill.
- Gábor Békés & Gábor Kézdi(2021). Data Analysis for Business, Economics, and Policy. Cambridge University Press.

Supplementary Textbooks





- Foster Provost and Tom Fawcett,(2013). Data Science for Business, O'Reilly. （中译本：《商战数据挖掘：你需要了解的数据科学与分析思维》,人民邮电出版社，2019）
- 王汉生,《数据思维：从数据分析到商业价值》，中国人民大学出版社，2017年9月。

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》，斯蒂夫•列维特、斯蒂芬•都伯纳著，中信出版社，2010年1月。）
- Ian Ayres, *Super Crunchers: Why Thinking-By-Numbers is the New Way To Be Smart*, 2007. (中译本《超级数字天才》，伊恩•艾瑞斯著，中国青年出版社，2008年1月。)
- Viktor Mayer-Schonberger and Kenneth Cukier, *Big Data: A Revolution that will Transform How We Live, Work and Think*. (中译本《大数据时代：生活、工作与思维的大变革》，浙江人民出版社，2013年1月)
- Nate Silver, *The Signal and the Noise: Why so many predictions fail—but some don't*,2012. (中译本《信号与噪声》，纳特•艾瑞斯著，中信出版社，2013年11月。)
- **涂子沛,“数据系列”：《大数据》、《数据之颠》、《数文明》以及《数商》**
  - 作者长期在硅谷工作，曾回国担任阿里集团副总裁。

Course Description

- This will introduce you to modern econometrics (causal inference) in an accessible way, even if you haven't taken a prior course in econometrics.
  - This is not an applied business statistics course, though it does require prerequisite knowledge in the subject.
  - It is also not a machine learning course, which focuses more on prediction.
- The course focuses more on intuition and practical examples, while **keeping the mathematical content to a minimum**. Notably.
  - Laying down a simple theoretical background in introductory level econometrics.
  - Inquiring about some basic instruments and some latest development ideas in econometrics.
  - Linking the econometric methods to the real-world business problem by cases.
- Learn to think **like an empiricist** when facing business problems in the Big Data era.

Evaluation

- Class Participation(20%)
- Team Project: A research proposal(80%)
  - A team consists of **no more than 5** students.
  - **Oral Presentations(40%)**: 10 minutes for 10 pages slides.
  - **Written Proposals (40%)**: 5 pages in A4, approximately 2500-3000 words.

Two Iron Rules

- Don't ever cheat on your assignments!

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

Welcome contact me and out TA

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

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

- 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?

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

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)

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

- 结合自身公司经营或行业特点，找到一份相关的数据集，描述该数据集的基本特征，包括数据收集方式、数据类型、数据结构、数据频率、样本量等等。
- 从该数据集出发，尝试提出一个研究问题，描述该问题的背景、意义、研究目的、研究方法等等。