Applied Micro-Econometrics

Lecture 0: Introduction

Zhaopeng Qu

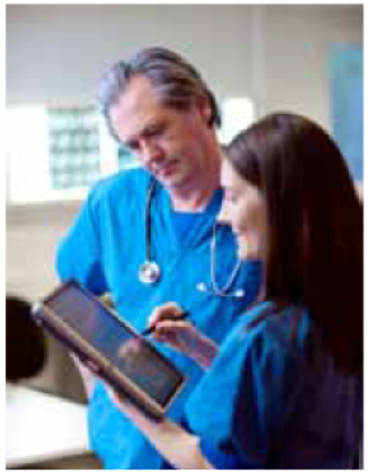
Business School,Nanjing University

Sep.22,2021

Outlines

Introduction: A Scientific Framework of Rational Knowledge

An Complex and Dynamic World

- Living in an unprecedentedly complex and dynamic world, we need rational knowledge(理性认知) and scientific predicition(科学预测).
- Economics you have learned suggests important relationships, often with policy implications, but *virtually never suggests quantitative magnitudes of causal effects*.

How can we have rational knowledge(judgment)?

- Anecdotes(轶事） or Intuition(直觉）
- Theory(理论/逻辑推理）
  - Systematical methodology: Hypothesis, Logical deduction...
- Empirics （数据实证）
  - Statistical inference from data.

An Example: Smoke and Mortality

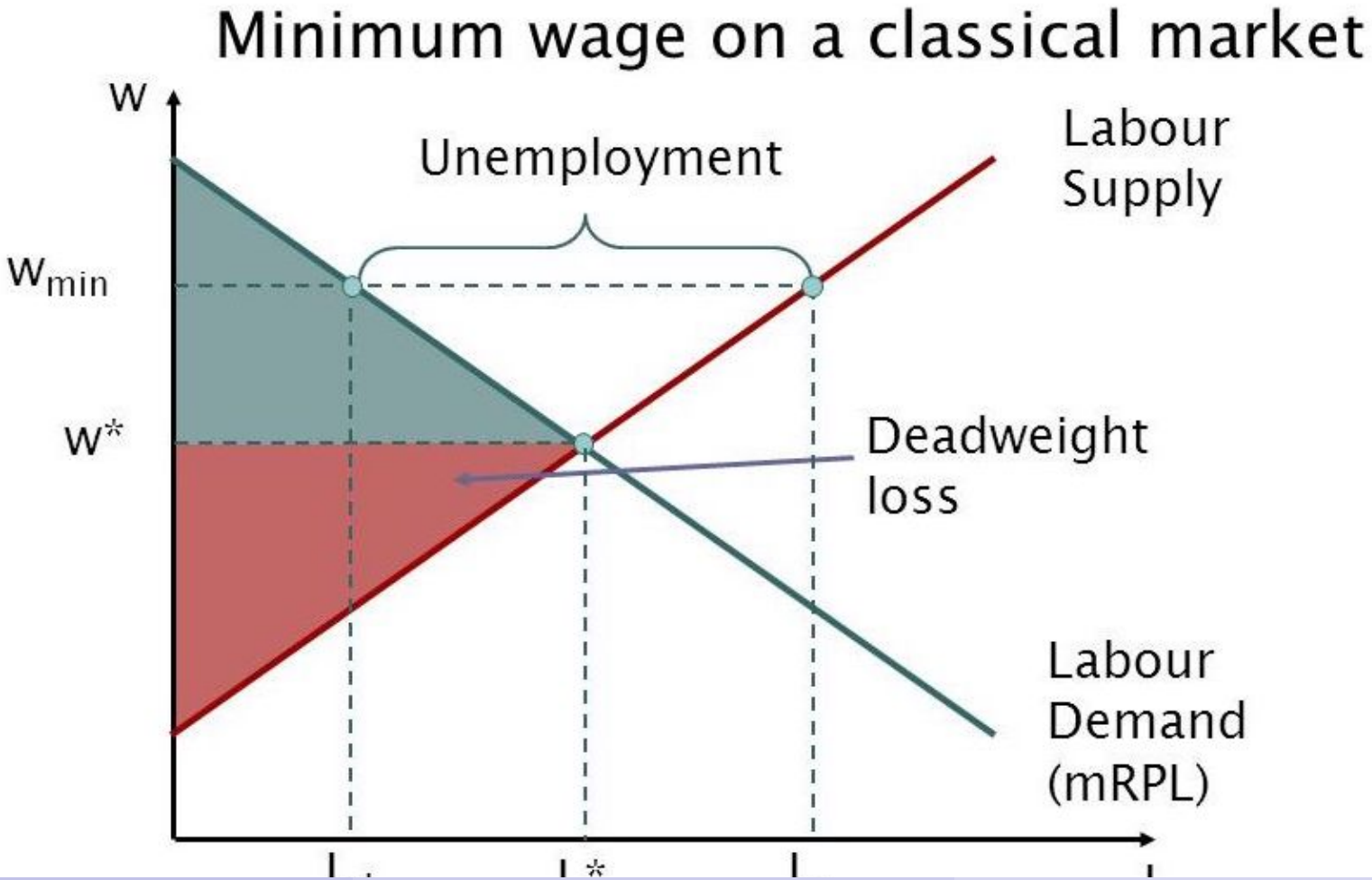
- Anecdotes(轶事) or Intuition(直觉)
  - eg. "My grandmother smoked two packs a day and lived until she was 95 years old."
- Theory
  - 因为香烟中含有尼古丁、焦油、甲醛等致癌物，所以...
- Empirics
  - 做实验或者通过抽样调查等方式收集数据，再用统计或计量方法来验证。

Classical Theory Question: Human Capital v.s Signal

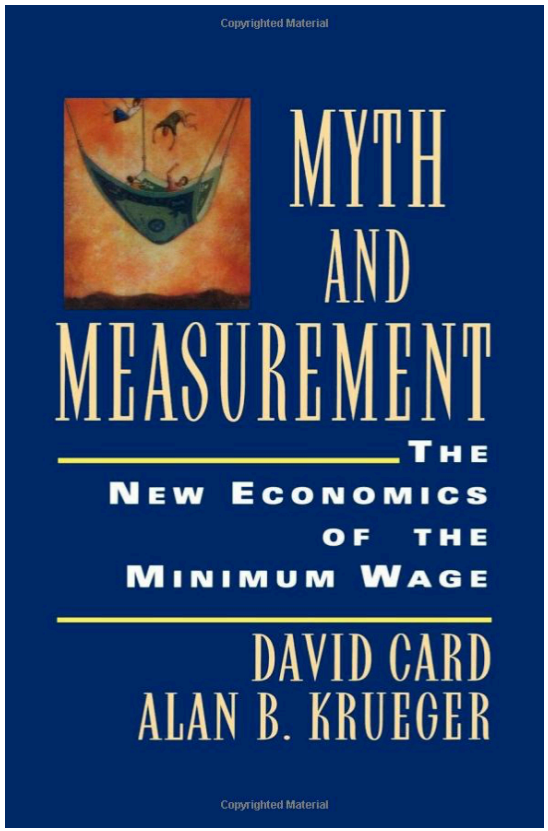
- A common phenomenon in labor markets can be observed all over the world.
  - **Higher education, Better pay!**
- Two classical theory to explain
  - **Human Capital**: Education improves work productivity.
  - **Signal**: Education does not increase the productivity. It simply serves as a signal of the individuals' innate ability.
- Question: which one is right?

Hot Public Policy Debate: Minimum Wage and Unemployment

- Classical Supply-Demand Model tell us



Hot Public Policy Debate: Minimum Wage and Unemployment



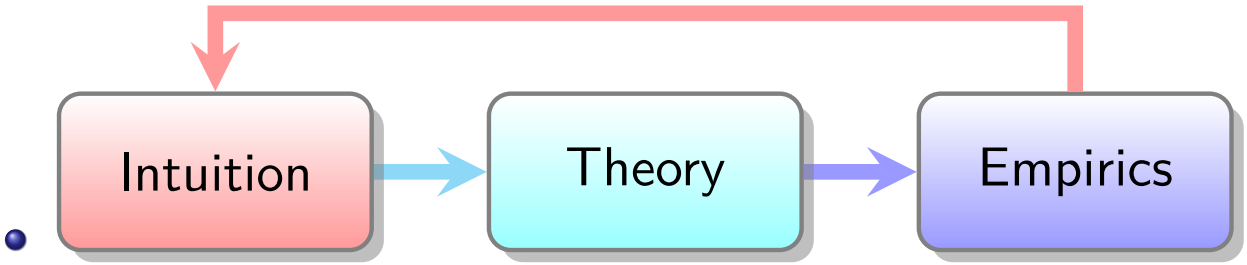
- One famous empirical evidence challenged the theory by Card and Krueger(1994)
- They found that increases in the minimum wage do NOT lead to job losses.

Other Similar Questions

- Air pollution and Health?
- Credit regulation on housing price
- Coupon on products sales
- Trade War...
- COVID19...

An 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 quantitive methods to test your theory or conclusion.



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

Theory, Empirics and Math

- Economic theories sole are not enough to explain social phenomenon.
  - It should be tested by empirical evidence.
  - When having competing theories, we need more test to justify which one is more likely right.
- When mechanism is clear, math is unnecessary, but empirical evidence is unexpendabley.
  - Having a mathematical model is better, but sometimes not necessary$\cdots$

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 theory may provide clues about the direction of the answer.
  - But decisions require quantitative answers to quantitative questions.
- Therefore, we have develop a framework and find a practical method that provide
  - a numerical answer to the question
  - a measure of how precise the answer is.
- *It is the job of* ***Econometrics***

What is Econometrics?

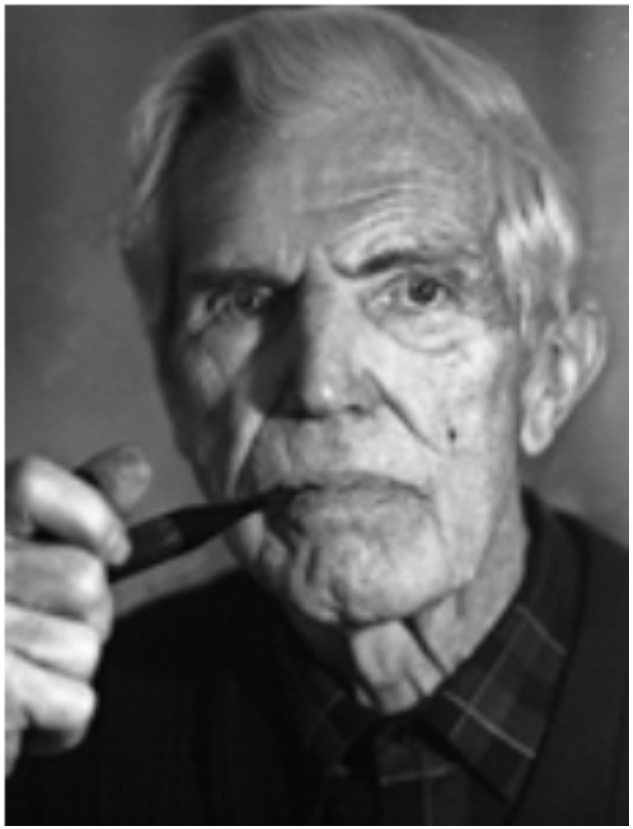
Introduction: Econometrics

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

Introduction: Econometrics

- "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."
  - (Ragnar Frisch, ***Econometrica***, 1933, volume 1, pages. 1-2)

Introduction: Econometrics



- ***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." (***Econometrica***, 1944, volume 12, pages. 1-2)

Introduction: Econometrics

- ***James Stock and Mark Watson(2014)***
  - "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: Econometrics

- My View: **In general, a series of scientific methods to searching for economic logics from data.**
- It could include two jobs
  - *Making a causal inference*,such as
    - Testing economic theories.
    - Estimating causal effects.
    - Using data to give policy recommendations.
  - *Forecasting or predicting* future values
- 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**.

Methodological Revolutions in Social Science

- Social science(firstly started by Economics) is experiencing **two methodological "revolutions"** over the past few decades.
- On the one hand, there is the **"Credibility Revolution"**
  - A movement that emphasizes the goal of obtaining secure causal inference(Angrist and Pischke, 2010)
- On the other hand, there is the **"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)
- Obviously, Econometrics are playing key roles in both two revolutions.

Credibility Revolution in Social Science

- Before the revolution, researchers frequently relied on attempts to statistically model the world to make inferences from **observational data**.
  - In essence, they would rely heavily on **ex post** statistical analysis to make causal inferences. Now we have acknowledged that this is not a "real" causal infernce.
- The revolution centered around the idea that the only way to truly account for possible sources of bias is to remove the influence of all confounders **ex ante through better research design.**
  - Thus, since the revolution, researchers have attempted to design studies around sources of random or as-if random variation, either with **experiments** or what have become known as **"quasi-experimental"** designs.
- The revolution started from around **1990s pioneering in economics** then spread over other empirical social sciences such as sociology, policital science, education, legal studies, etc, which *has entirely changed empricial social science*.

Big Data Revolution in Social Science

- The name of Big Data comes from computer scientists working to do aggregation on data that is too big to fit on a single machine.
- Now **Data Science** is a umbrella term including many aspects of the challenges from the massive-scale datasets.
- Big Data Analytics is just *Applied Data Science*,especially focused on business and industrial applications.
- The big data present exciting opportunities for the study of social science, but at the same time its size and heterogeneity present significant challenges.
- Big Data analytic tools are highly correlated with econometrics
  - partially straight out of previous statistics and econometrics classes(Cluster, Regression and Matching)
  - some are totally new(such as decision trees and neural networks).

Big Data,Data Science and Econometrics

- So many labels for what we do...
  - Econometrics
  - Data Mining/Big Data/Data Science
  - Machine Learning(ML)
  - Artificial Intelligence(AI)
- Along this spectrum, you move from heavy focus on what things you are measuring (what real phenomena they correspond to) to a more practical 'useful is true' pattern discovery approach.
- *The similarities are much bigger than any distinctions.*

Econometrics: sub-fields or sub

- **Theoretical Econometrics**
  - It is concerned with methods, both their properties and developing new ones.
  - It is closely related to mathematical statistics, and it states assumptions of a particular method, its properties etc.
  - We could call *theoretical econometricians* as the ***producer*** of econometrics.
- **Applied Econometrics**
  - More orientated to applied work, such as choice of technique and interpretation of research finding.
  - But it should be also based on **a solid conceptual foundation** and some **practical experiences** plus a little bit **skills of computer**.
  - Most of us are the ***consumers*** of econometrics.

Wrap Up

- Econometrics is a collection of a series of scientific methods to searching for economic logics from data.
- It could include two jobs: *causal inference and prediction.*
- It are playing key roles in revolutions of both social science and business practice.

Methodological revolutions in Social Science

- Social science(firstly started by Economics) is experiencing **two methodological "revolutions"** over the past few decades.
- On the one hand, there is the **"credibility revolution"**
  - A movement that emphasizes the goal of obtaining secure causal inferences(Angrist and Pischke, 2010)
- On the other hand, there is the **"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)
- Obviously, Econometrics are playing a key role in these two revolutions.

Credibility Revolution in Social Science

- Before the revolution, researchers frequently relied on attempts to statistically model the world to make inferences from **observational data**.
  - In essence, they would rely heavily on **ex post** statistical analysis to make causal inferences. Now we have acknowledged that this is not a "real" causal infernce.
- The revolution centered around the idea that the only way to truly account for possible sources of bias is to remove the influence of all confounders **ex ante through better research design.**
  - Thus, since the revolution, researchers have attempted to design studies around sources of random or as-if random variation, either with **experiments** or what have become known as **"quasi-experimental"** designs.
- The revolution started from around **1990s pioneering in economics** then spread over other empirical social sciences such as sociology, policital science, education, legal studies, etc, which *has entirely changed empricial social science*.

Big Data Revolution in Social Science

- The name of Big Data comes from computer scientists working to do aggregation on data that is too big to fit on a single machine.
- Now **Data Science** is a umbrella term including many aspects of the challenges from the massive-scale datasets.
- Big Data Analytics is just *Applied Data Science*,especially focused on business and industrial applications.
- The big data present exciting opportunities for the study of social science, but at the same time its size and heterogeneity present significant challenges.
- Big Data analytic tools are highly correlated with econometrics
  - partially straight out of previous statistics and econometrics classes(Cluster, Regression and Matching)
  - some are totally new(such as decision trees and neural networks).

Big Data,Data Science and Econometrics

- So many labels for what we do...
  - Econometrics
  - Data Mining/Big Data/Data Science
  - Machine Learning(ML)
  - Artificial Intelligence(AI)
- Along this spectrum, you move from heavy focus on what things you are measuring (what real phenomena they correspond to) to a more practical 'useful is true' pattern discovery approach.
- *The similarities are much bigger than any distinctions.*

Econometrics: sub-fields or sub

- **Theoretical Econometrics**
  - It is concerned with methods, both their properties and developing new ones.
  - It is closely related to mathematical statistics, and it states assumptions of a particular method, its properties etc.
  - We could call *theoretical econometricians* as the ***producer*** of econometrics.
- **Applied Econometrics**
  - More orientated to applied work, such as choice of technique and interpretation of research finding.
  - But it should be also based on **a solid conceptual foundation** and some **practical experiences** plus a little bit **skills of computer**.
  - Most of us are the ***consumers*** of econometrics.

Wrap Up

- Econometrics is a collection of a series of scientific methods to searching for economic logics from data.
- It could include two jobs: *causal inference and prediction.*
- It are playing key roles in revolutions of both social science and business practice.

Why and who should take the course?

The Purpose

- To build a bridge from learning courses to doing research
  - Inquire some basic instruments in empiricist' s toolbox.
  - Read papers to develop a good taste and identify some good research topics.
  - Use knowledge you learned to propose your own research topics.

Who pursue an academic career

- Congratulations! You are in one of the most promising and internationalizing areas in economic research of China.
  - Master econometrics knowledge will improve your research greatly.
  - Your research is judged on how convincing it is.
  - Econometrics helps ensure and formalize credibility.
  - Overwhelming majority of top journal articles are quantitative.

Who enter industry job market

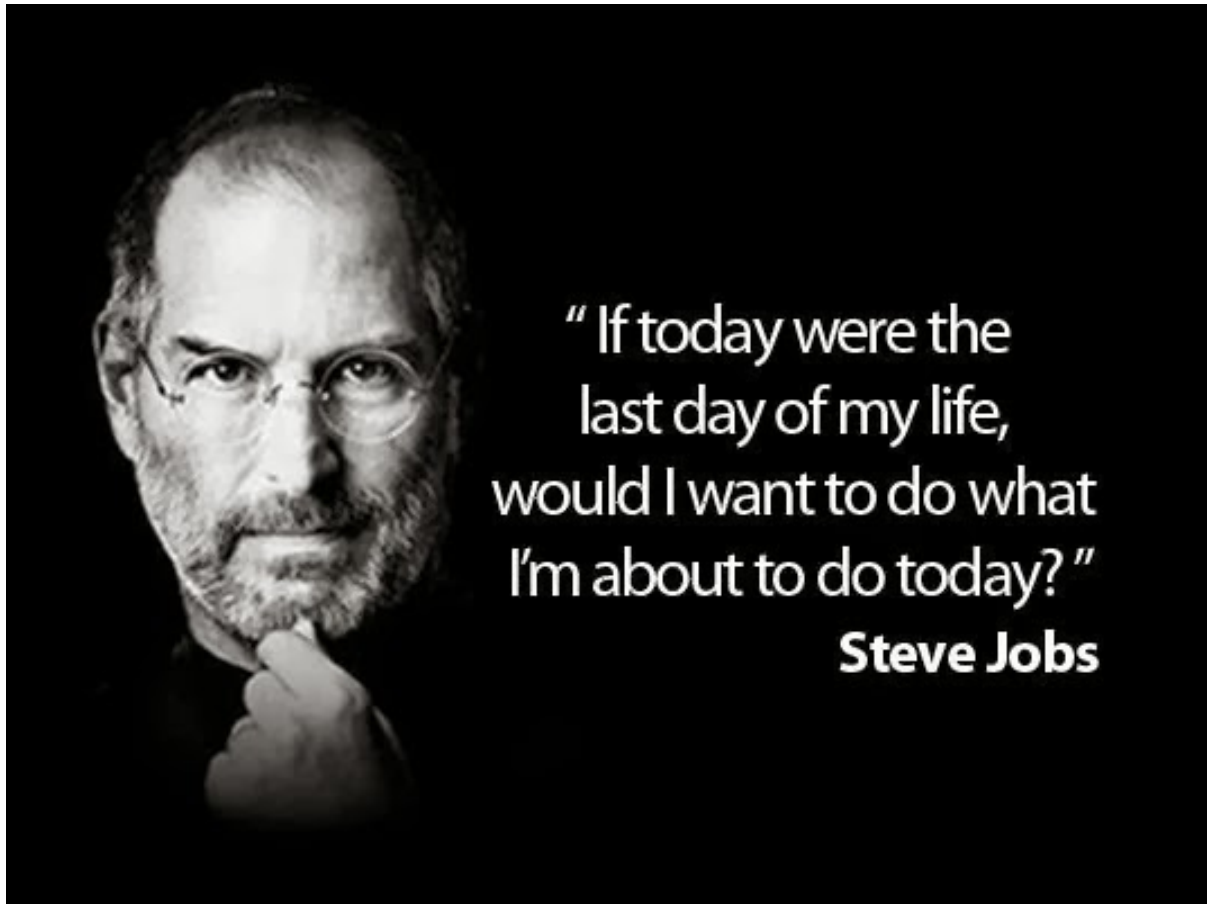
- Who want to work in industry: mastering econometrics can help you get a good job!
  - Credit cards, POS terminals and ERP systems widely used in supermarkets, banks and factories also revolutionize the business (or management) practices in marketing, accounting, management operation etc.
  - A lot of internet giants even hire economists to lead their special R&D department. Such as Google/Microsoft/eBay/Amazon/阿里巴巴/腾讯/百度/....
  - Besides in finance, **Big data** are also grow vigorously in Consulting and Business areas.

Who look for fun

- Applied MicroEconometrics could not be a boring and demanding variant of a mathematics course, but **interesting and having fun**.
- You can just enjoy it by thinking in an empiricist' s way in your daily life.
  - have novel ideas or new perspectives about our world.
  - Econometrics is kind of a bible or philosophy of economists.
- We will cover many very interesting stories
  - Interesting Examples
    - Eg. Crime and Abortion in *Freakonomics* written by Steven Levitt.
    - Eg. What is the value to be the president's son(or daughter)? in *Economic Gangster* written by Raymond Fisman and Edward Miguel.

Whoever and Whatever

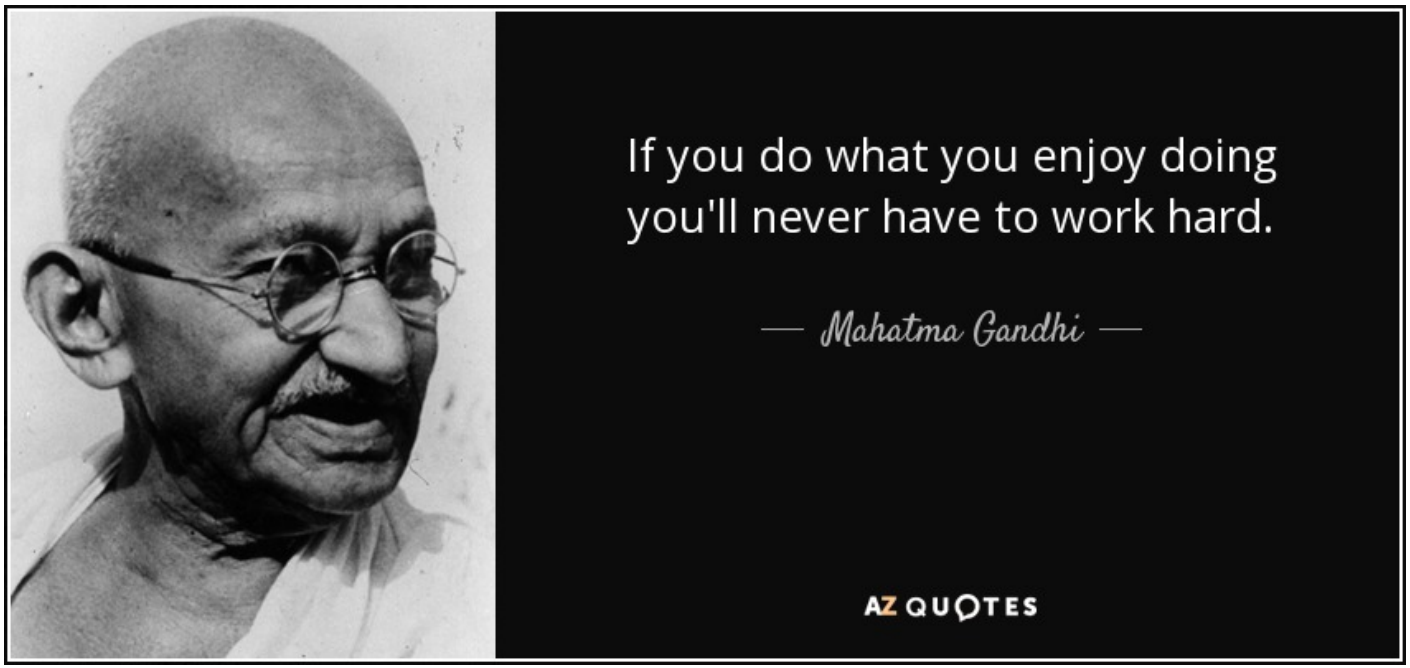
- **Whoever you would like to be or whatever you want**



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

Whoever and Whatever

- **Whoever you would like to be or whatever you want**



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

Hard and Soft Skills

- You 'd better learn or improve several important skills during gradute studies.
  - **Hard Skills**
    - Language
    - Computer
  - **Soft Skills**
    - Critical Thinking
    - Presentation and Writing
    - Teamwork
- **Fortunately, you will learn/practice almost all above skills in our class.**

Conclusion

- In a word, *Applied Econometrics* is a **veryvery vary** important and interesting but also a "hardcore" course.
  - Please think over before your take it!
  - Once you take it, please work hard on it!
  - And please enjoy working hard on it!

Logistics to the Course

About Me

- My name is *Zhaopeng Qu*
  - Position and Affiliation: *Associate Professor, Institute of Population Studies, Business School.*
  - Research Fields: *Labor Economics and Applied Econometrics*
  - Office: Room 2017, Anzhong Building
  - Office Hour: Make an appointment in advance
  - Tel: 83621349
  - Email: qu@nju.edu.cn
- Online resources
  - Our Course Website: https://byelenin.github.io/MicroEconometrics/
  - Wechat group: discuss anything about the course.(optional)

How to learn

- Read required and related materials before lecture.
- Get well-prepared for the presentation before the class.
- Discuss with your classmates often.
- Make good friends with data and Stata or R
- Think about the content related with your personal knowledge and experience.
- Whenever you have questions, ask them.

The Procedure:

- **The First Part-Lecture**
  - Introduce the underlying theoretical problems briefly and focus on the empirical strategy heavily.
  - Focus on some specific examples in classical papers with interesting topics in our field.

The Procedure:

- **The Second Part: Presentation for Papers**
  - Read 10-15 papers which publised in top journals in economics.
  - Every time we will discuss **three** papers. And **two-person-team** for **one paper** to present.

The Procedure:

- **The Third Part: Your Own Research**
  - A Research Proposal: in mid term
  - And Preliminary Results: in the end
  - Goal: A decent term paper

The Procedure:

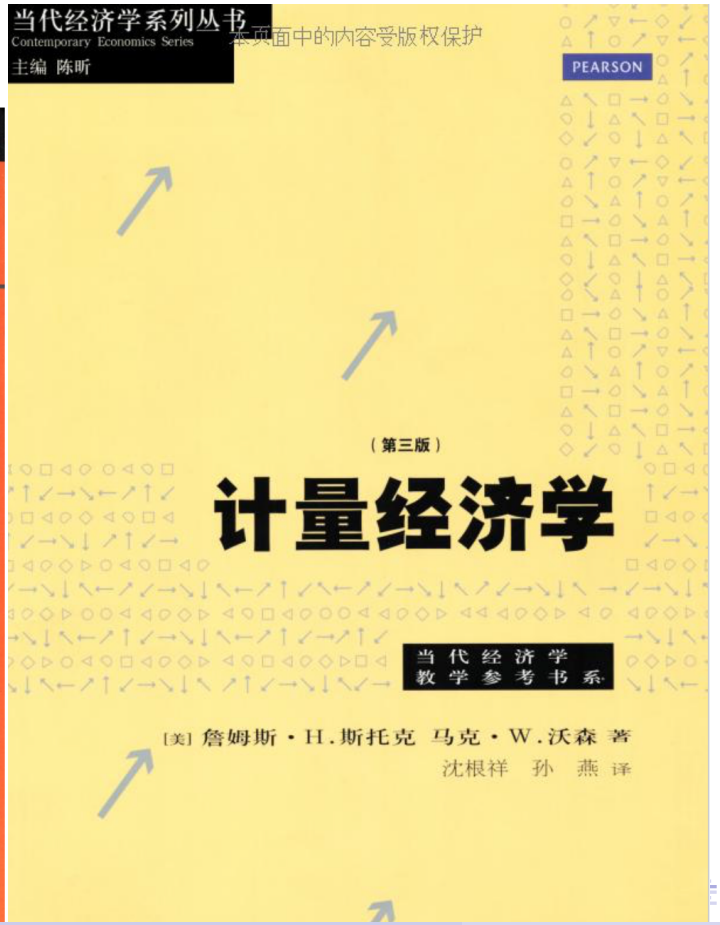
- **Last but not least: Presentation Structure**
  - Motivation
  - Literature
  - Empirical Strategy
  - Data
  - Results

Prerequisite

- I assume that you **should** be comfortable some basic concepts of **probability theory** and **statistics**, such as
  - random variable
  - expectation
  - variance
  - probability density function, p.d.f.
  - cumulative distribution function, c.d.f
  - covariance
  - L.L.N and C.L.T
  - Estimator
  - Unbiased and Consistent
  - Asymptotic Normality
  - Hypothesis test
- We will not review these basic concepts and formulas in the lectures. You should review it by yourself anyway.

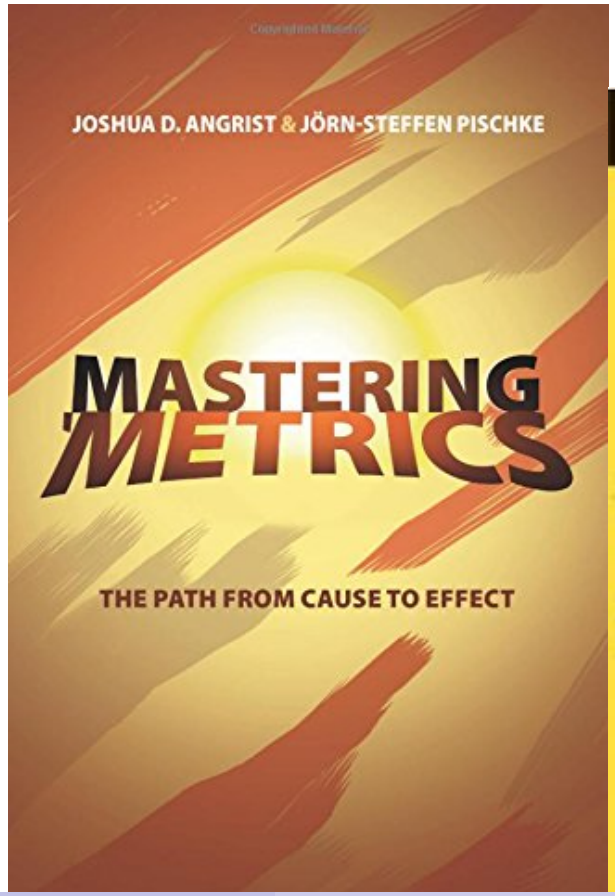
Reference Textbooks

- Required Textbook: James H. Stock & Mark W. Watson, (2012). *Introduction to Econometrics*, 3rd Edition, Pearson Education.
  - 影印版：格致出版社 / 上海人民出版社。
  - 中文版：世纪出版集团 / 上海人民出版社。



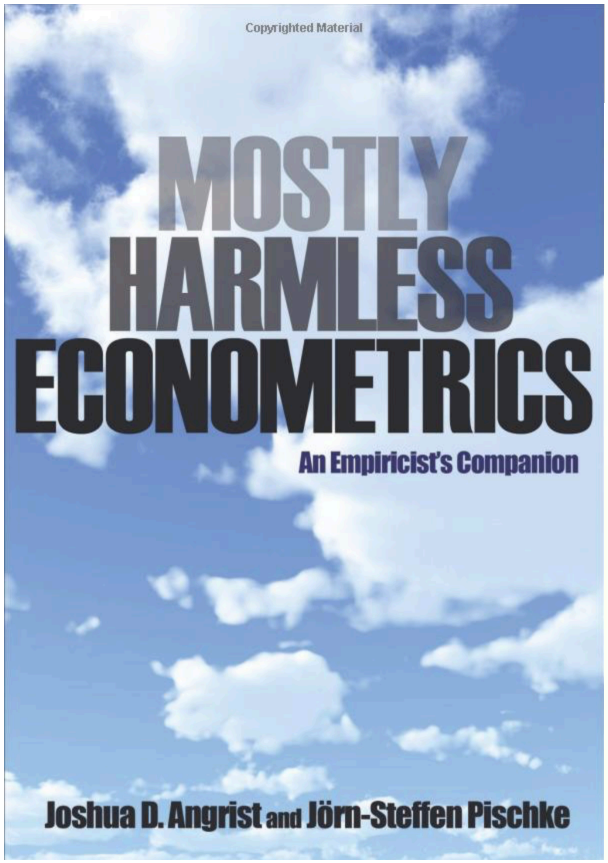
Reference Textbooks

- Supplementary textbook
  - 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）



Reference Textbooks

- Supplementary textbook
  - Joshua D. Angrist & Jorn-Steffen Pischke, (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press. (中译本：《基本无害的计量经济学》，格致出版社，2018)



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。）
- 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。)
- Angus Deaton, *The Great Escape:Health,Wealth,and the Origins of Inequality*,2015.(中译本:《逃离不平等: 健康、财富及不平等的起源》，中信出版社。)

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 Stata or R to analyze some popular 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!
  - Interrupt me as often as necessary! (but I know most of you are not comfortable to this)
  - Got a dumb question? Please assume that you are the smartest person in the class, and you eventually will be!
- Read required materials and finish homeworks.

Evaluation

- Class Participation (10%)
- Presentation(20%)
- Midterm: A Proposal and Presentation (30%)
- Final Exam: Preliminary Results (30%)
- Final Draft(10%)

Schedual

Schedual

Q & A

The Structure of Economic Data

Introdution

- Data Structure
  - Cross-sectional data
  - Time series data
  - Pooled cross-sectional data
  - Panel 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.

- $x = STRatio;\ y = TestScore;\ N = 420$

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

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

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

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.

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

3.Pool(Repeat) 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.
  - It means that data does not track the respondent multiple times.
  - But the data come from a same population in different time.
- 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:
  - Before-after comparisons of a government policy
  - Cohort studies
  - Cross-sectional analyses

3.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

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

4.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.

$$\{x_{it}, yi_t \cdots\}_{i=1,t=1,}^{NT}$$

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

4.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.

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

Corresponding sub-fields or sub-course

- 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)

Source

- Traditional Collecting Way:
  - Survey(调查)
  - Administrative data(官方业务数据)
- Collecting Data in Digital Times:
  - Mass or Big data:
    - Online documents
    - Social Media
    - Geolocations or Geographic data
  - High Frequency Data
    - Stock,future or other financial tractional data

Big Data: Examples

Big Data Application examples in different Industries:

Retail/Consumer

- Merchandizing and market basket analysis
- Campaign management and customer loyalty programs
- Supply–chain management and analytics
- Event– and behavior–based targeting
- Market and consumer segmentations

Finances & Frauds Services

- Compliance and regulatory reporting
- Risk analysis and management
- Fraud detection and security analytics
- Credit risk, scoring and analysis
- High speed arbitrage trading
- Trade surveillance
- Abnormal trading pattern analysis

Web and Digital media

- Large–scale clickstream analytics
- Ad targeting, analysis, forecasting and optimization
- Abuse and click–fraud prevention
- Social graph analysis and profile segmentation
- Campaign management and loyalty programs

Health & Life Sciences

- Clinical trials data analysis
- Disease pattern analysis
- Campaign and sales program optimization
- Patient care quality and program analysis
- Medical device and pharmacy supply–chain management
- Drug discovery and development analysis

Telecommunications

- Revenue assurance and price optimization
- Customer churn prevention
- Campaign management and customer loyalty
- Call detail record (CDR) analysis
- Network performance and optimization
- Mobile user location analysis

Ecommerce & customer service

- Cross–channel analytics
- Event analytics
- Recommendation engines using predictive analytics
- Right offer at the right time
- Next best offer or next best action

Big Data: Volumn

- Big in volumn in terms of
  - the number of observations(size n)
  - the number of variables(dimension p)
- Eg. Wechat(微信) in 2018
  - 1.082 billion active users every month
  - 45 billion messages every day
  - 0.41 billion video calls every day

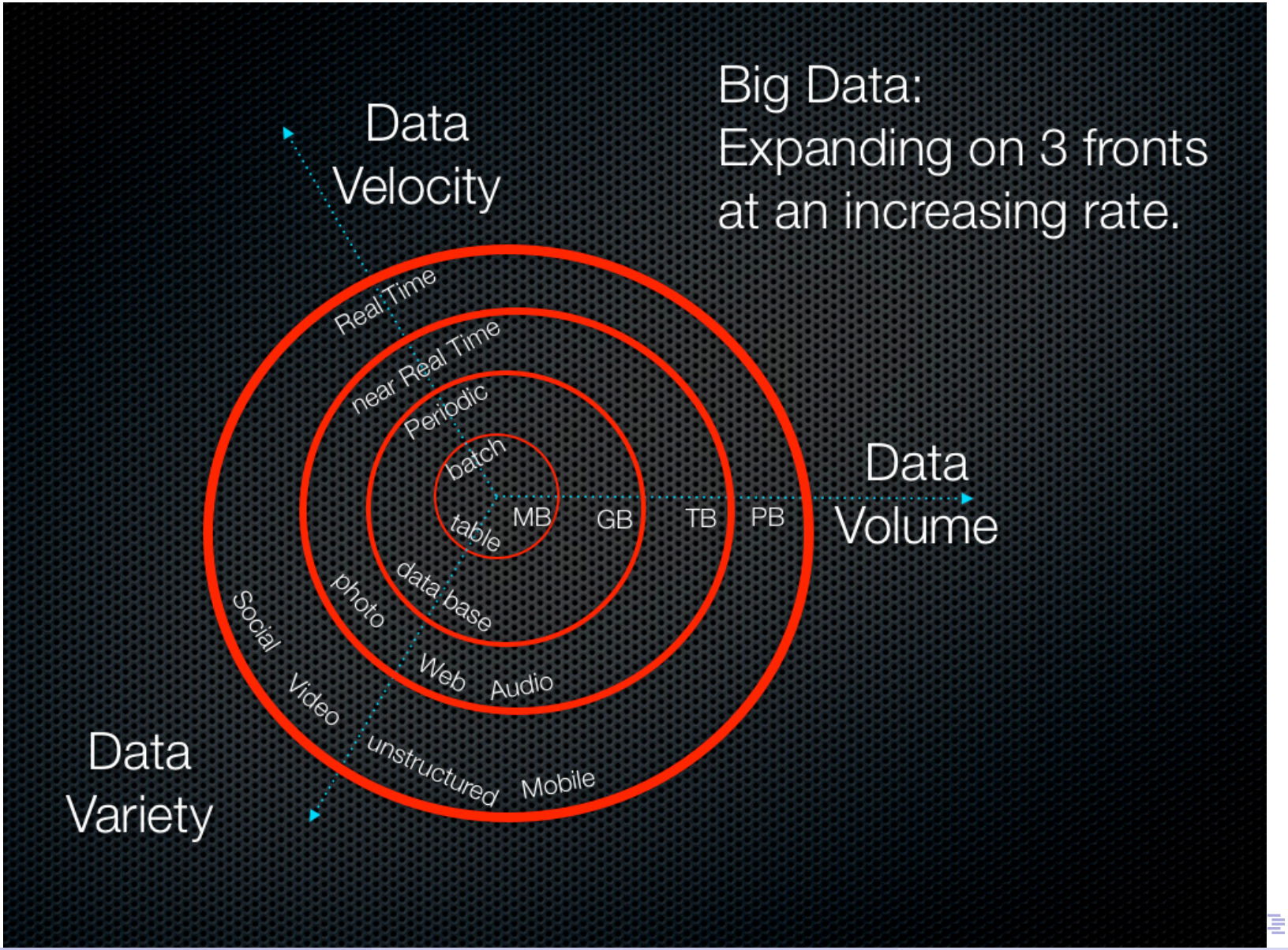
Big Data: Variety

- Big in variety in terms of
  - classical data(numbers in tables)
  - photo/audio/video
  - text
  - map
  - sensors

Big Data: Velocity

- Velocity is the measure of how fast the data is coming in.
  - Eg. Facebook(脸书) users upload more than 900 million photos a day in 2016
  - Eg. Wechat(微信) in 2018
    - 45 billion messages every day
    - 0.41 billion video calls every day

Big Data: 3Vs


Big Data:
Expanding on 3 fronts
at an increasing rate.
Data
Velocity
Real Time
near Real Time
Periodic
batch
table
MB
GB
TB
PB
Data
Volume
data base
photo
Social
Web
Audio
Video
unstructured
Mobile
Data
Variety

Big Data means to analytics

- Big in the number of observations: size n

$$n \to +\infty$$

- Big in the number of variables: dimension p

$$p \to +\infty$$

- Tools in Big Data are
  - partially straight out of previous statistics and econometrics classes(Cluster, Regression and Matching)
  - some are totally new(such as decision trees and neural networks)

Data in China

- **Survey Data**
  - China Household Income Project(CHIP)
  - 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；
  - 工业企业数据库；
  - 海关交易数据库；
- **Online data**:
  - Shopping dataon Tabao,JD,Tmall
  - Movie Data on Douban.com(豆瓣电影数据)
  - Air Quality: PM2.5(空气质量数据)
  - Night-Lights Data(夜间灯光数据)
  - Land Transaction Markets(土地交易市场数据)

Homework(not required)

Homework

- **到如下数据库网站选择其一登记注册**
  - 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

Homework

- 了解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。
- 下载调查的问卷，详细了解调查有哪些具体的信息。
  - 首先确定自己感兴趣的问题，然后到问卷中去寻找。
  - 或者先看问卷，找到自己感兴趣的具体信息。
  - 下载相应数据，进行初步的数据清理和统计分析 (待上机课之后)
  - 为期末的研究项目做准备。