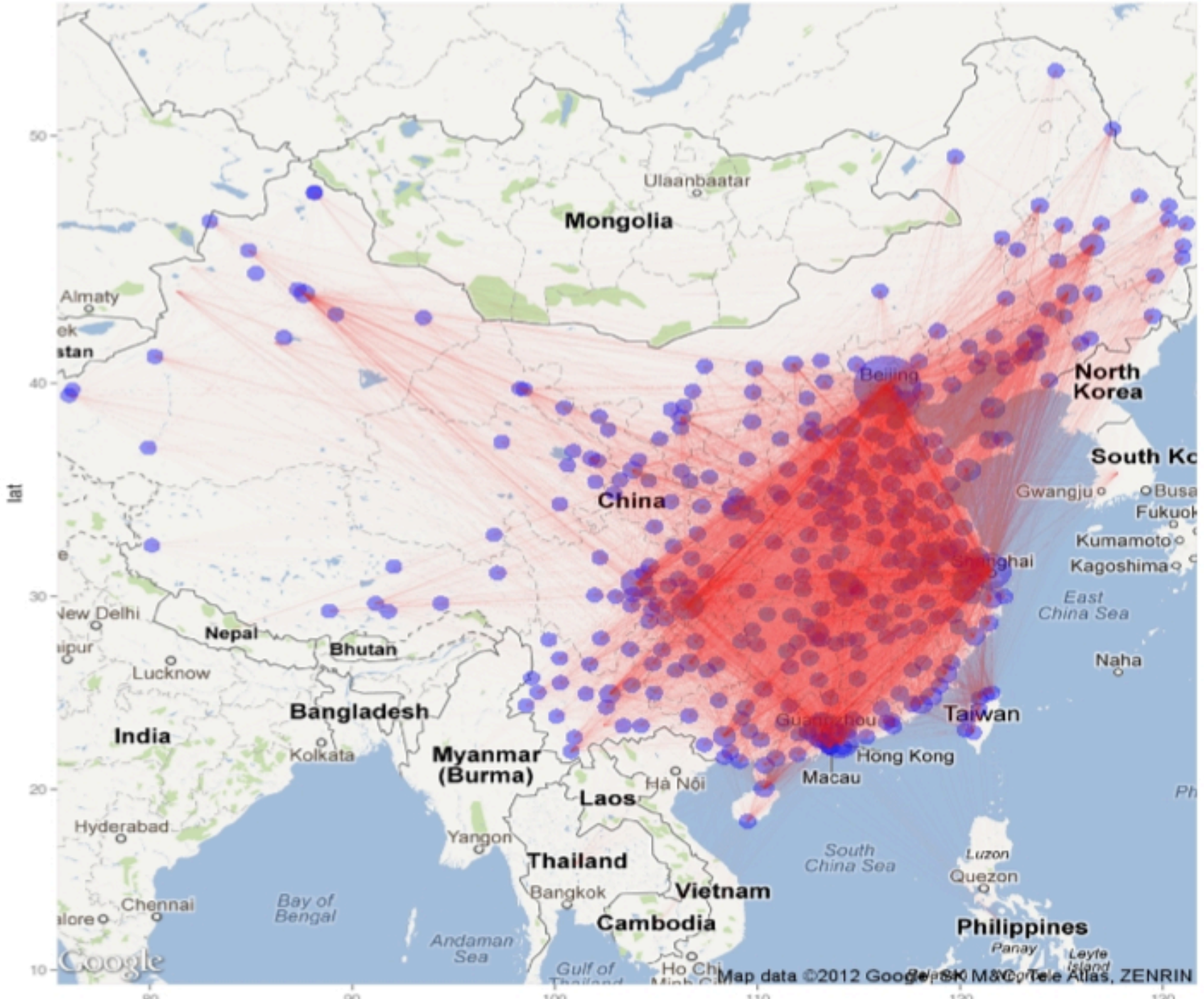
香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Big Data: An Introduction

Qiang Yang
Chair Professor, HKUST
http://www.cse.ust.hk/~qyang


Ulaanbaatar
Mongolia
Almaty
North Korea
Beijing
China
South Ko
Gwangju
Busa
Fukuok
Kumamoto
Kagoshima
East China Sea
Naha
Taiwan
New Delhi
Nepal
Bhutan
Lucknow
Bangladesh
India
Kolkata
Myanmar (Burma)
Hong Kong
Macau
Hà Nội
Laos
Hyderabad
Yangon
Thailand
South China Sea
Luzon
Quezon
Bangkok
Vietnam
Chennai
Bay of Bengal
Cambodia
Philippines
Andaman Sea
Panay
Leyte Island
Gulf of Thailand
Ho Chi Minh City
lat
50
40
30
20
10
Google
Map data ©2012 Google, SK M&C, Tele Atlas, ZENRIN

Baihe.com

- 每年上亿用户
- 5-10短信／天
- 日短信量：千万
- 2012年成功：50万对

What Big Data Means

Big Data: Volume, Velocity, Variety

Social

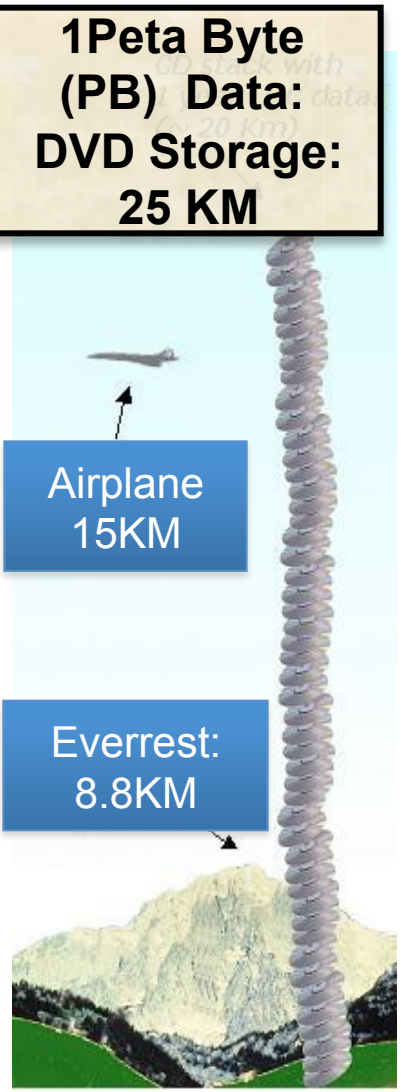
Facebook：1 Billion Users
daily increment: 10 TB

Search

Google：Daily 20 PB

Turing Award Winner: Jim Gray: Fourth Paradigm

- **Doubles every 18 months**
- **IDC: in 2020, will reach 35 X 10^6 PB**

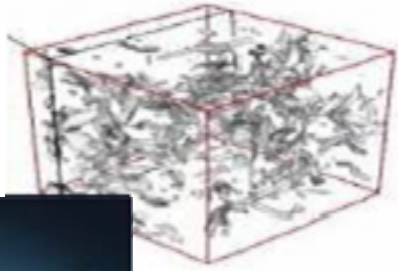


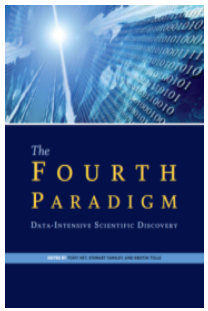
The Fourth Paradigm

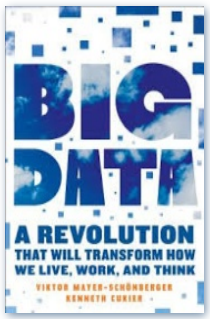
- Thousand Years Ago
  - Empirical observation
- Hundreds Years Ago - Theory
  - Newton Law, Maxwell …
- Decades Ago - Simulation
  - Complex Phenomena
- Today — Data Science

Causal Relation →Statistical Correlation

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$







From: <The Fourth Paradigm>

Data Analytics Tasks

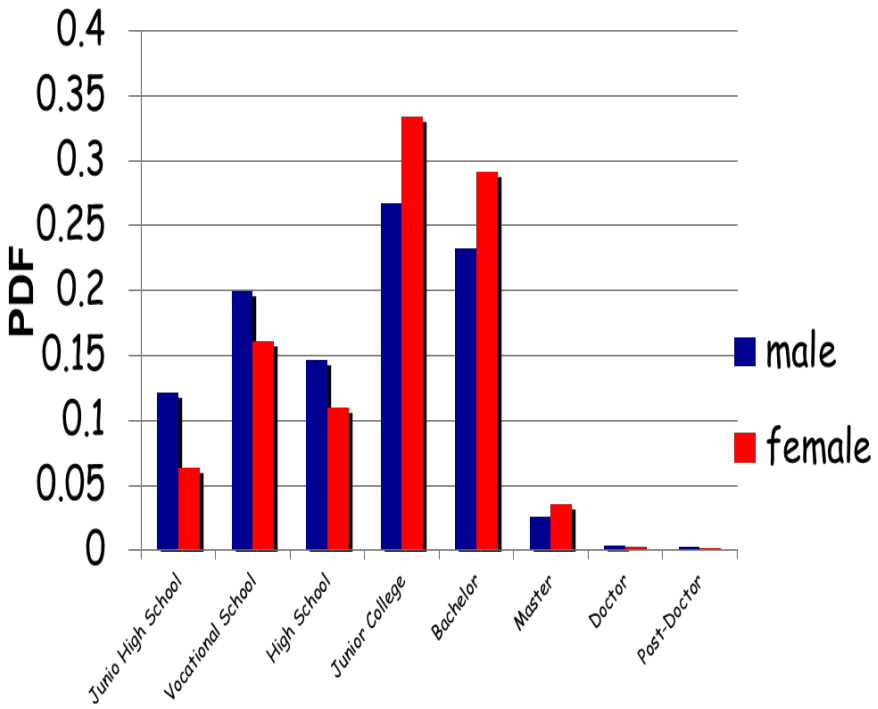
Query Data	Prediction
• Clustering （聚类） • Association （关联） • Segmentation （分群） • Topic Analysis （主题）	• Regression （回归预测） • Classification （分类） • Recommendation (推荐）

Data Analytics Tasks

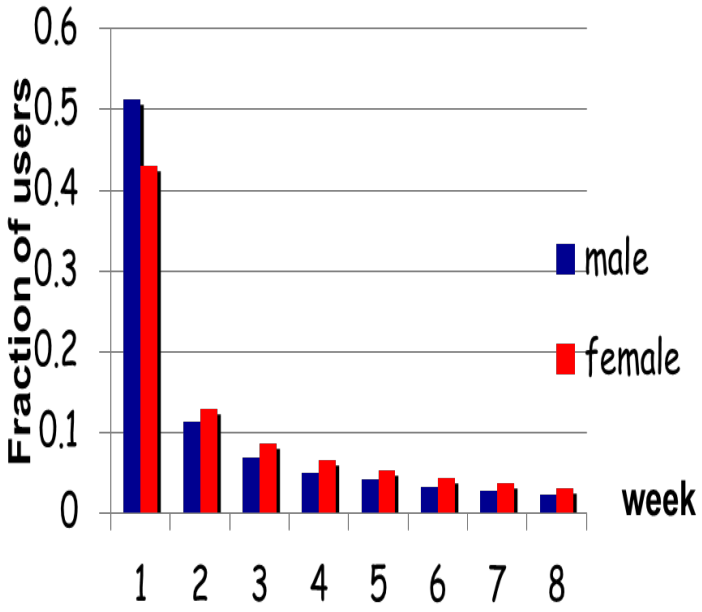
Query Data	
• Summary (汇总） •Clustering （聚类） • Association (关联） • Segmentation (分群） • Topic Analysis （主题）	

Who Date Online?

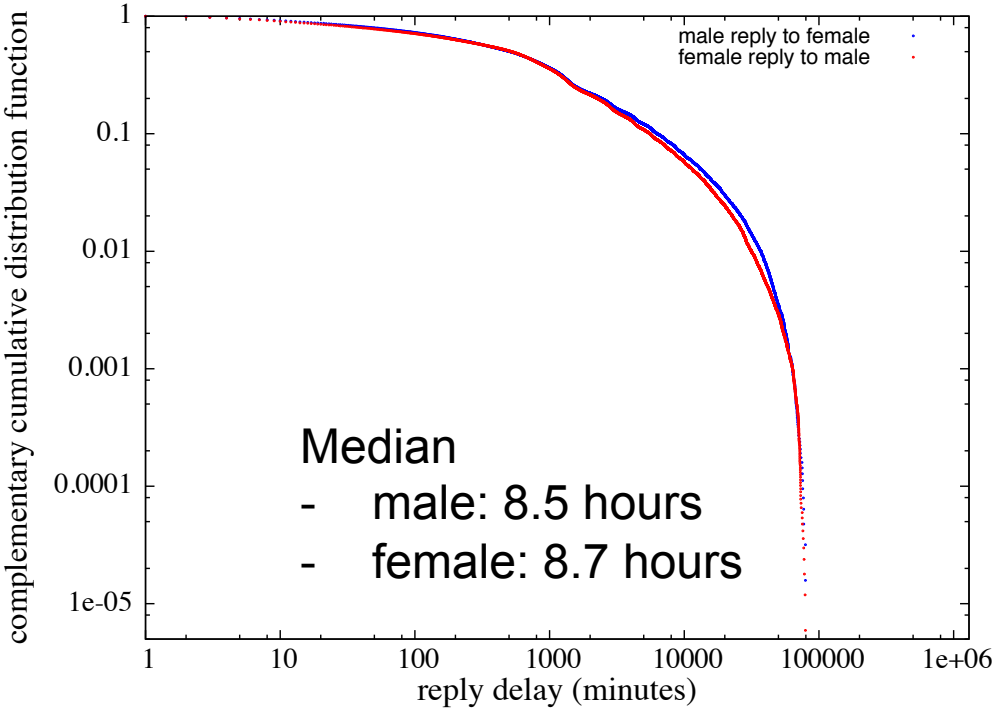
- Relevant to Education Level?



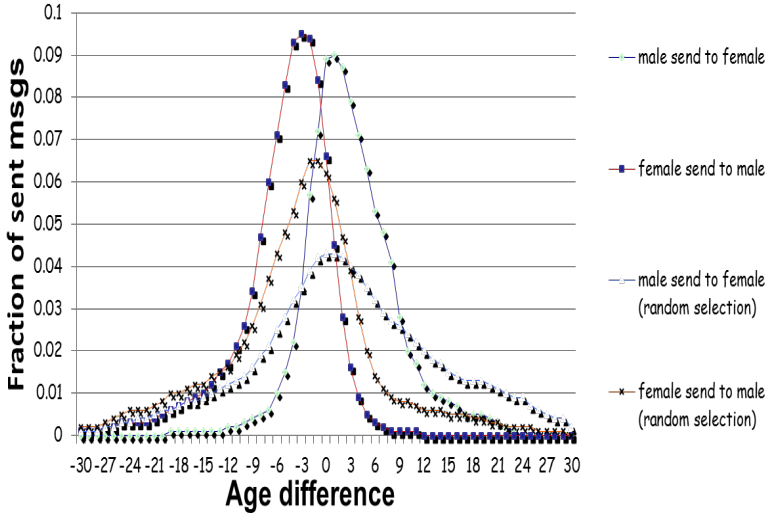
- How patient are they?



How Long Love Waits?

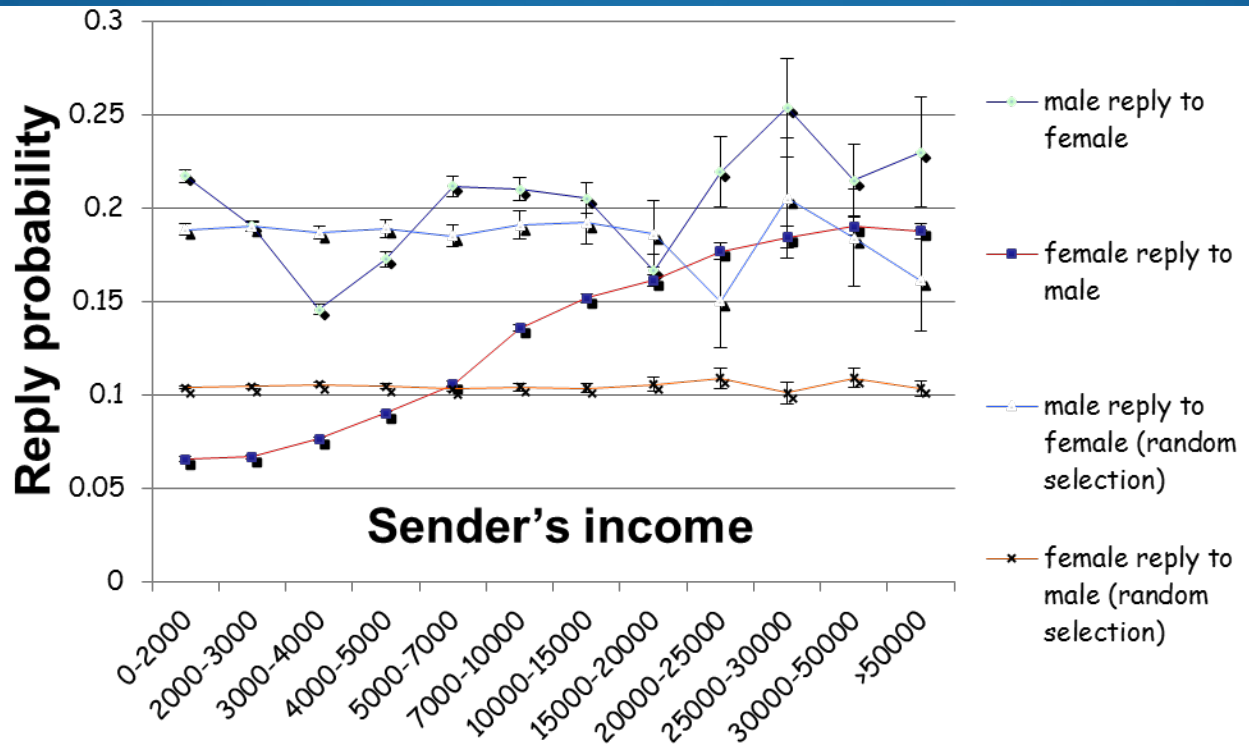


Median
- male: 8.5 hours
- female: 8.7 hours



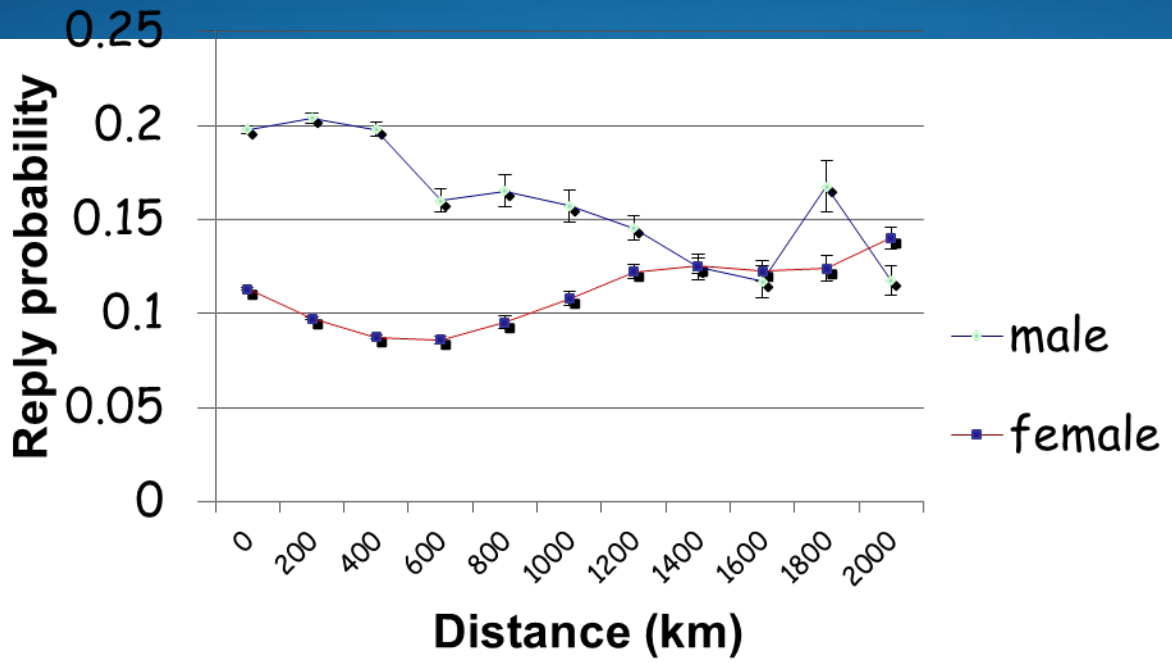
Males tend to look for younger females and females tend to look for older males

Does $$ Matter in Love?



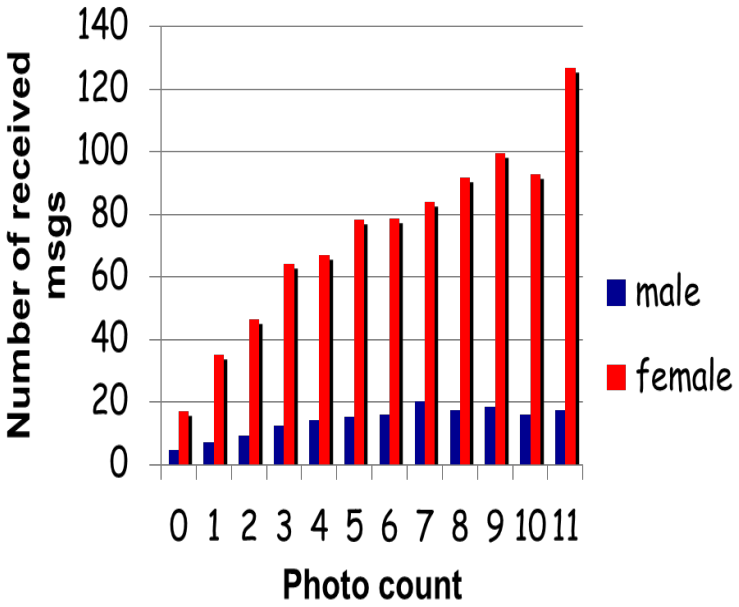
- Female reply probability increases with male senders' income,
- Males don't care much about the income of a female

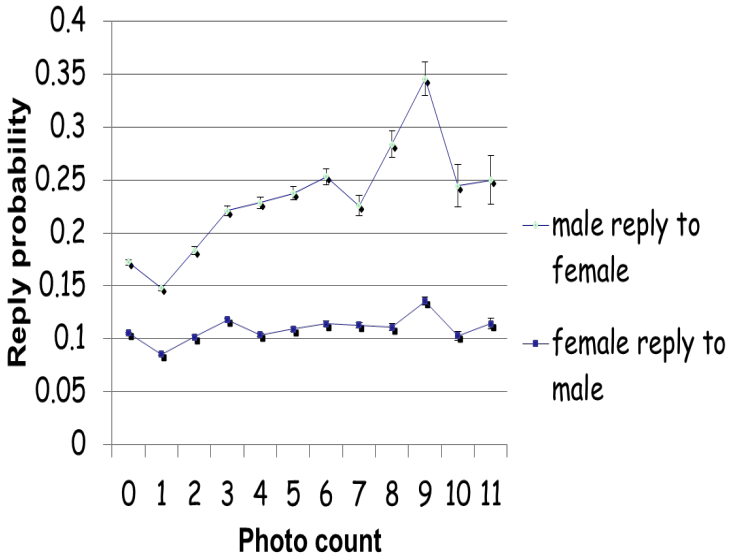
Does Distance Matter?



- Male reply probability decreases with distance
- Female reply probability first decreases with distance but increases in the range from 800 to 1,400km

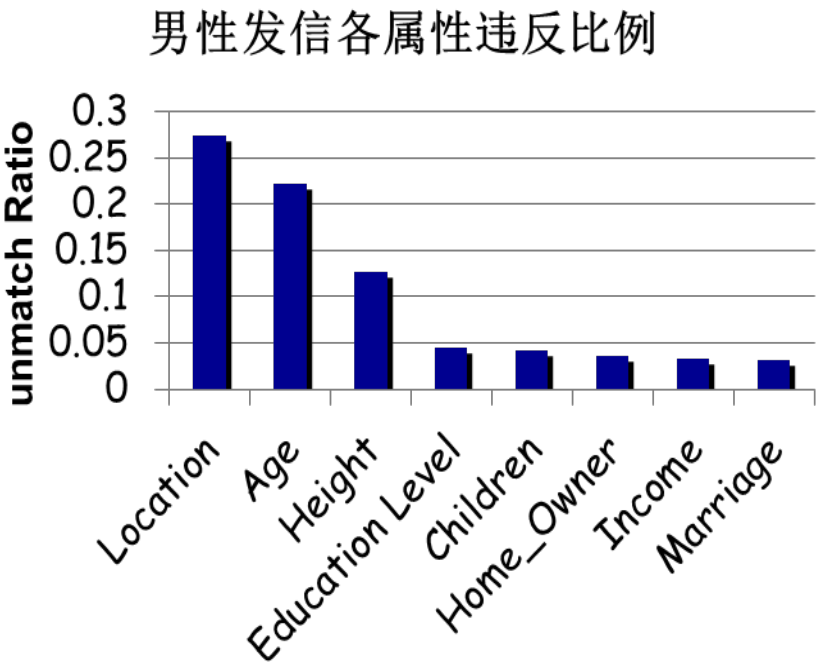
Are Looks important?

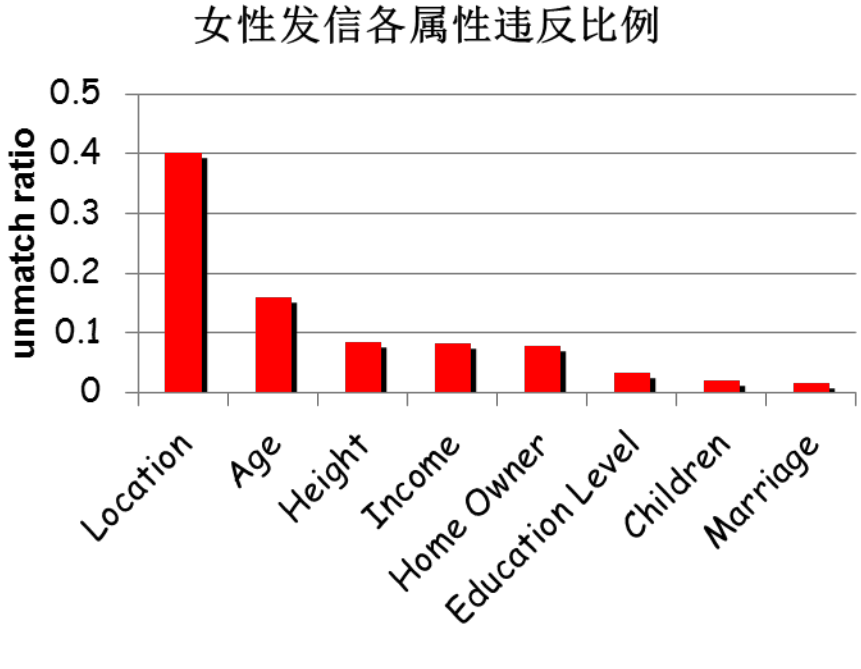




- One receives more msgs if posted more photos
- More so for females

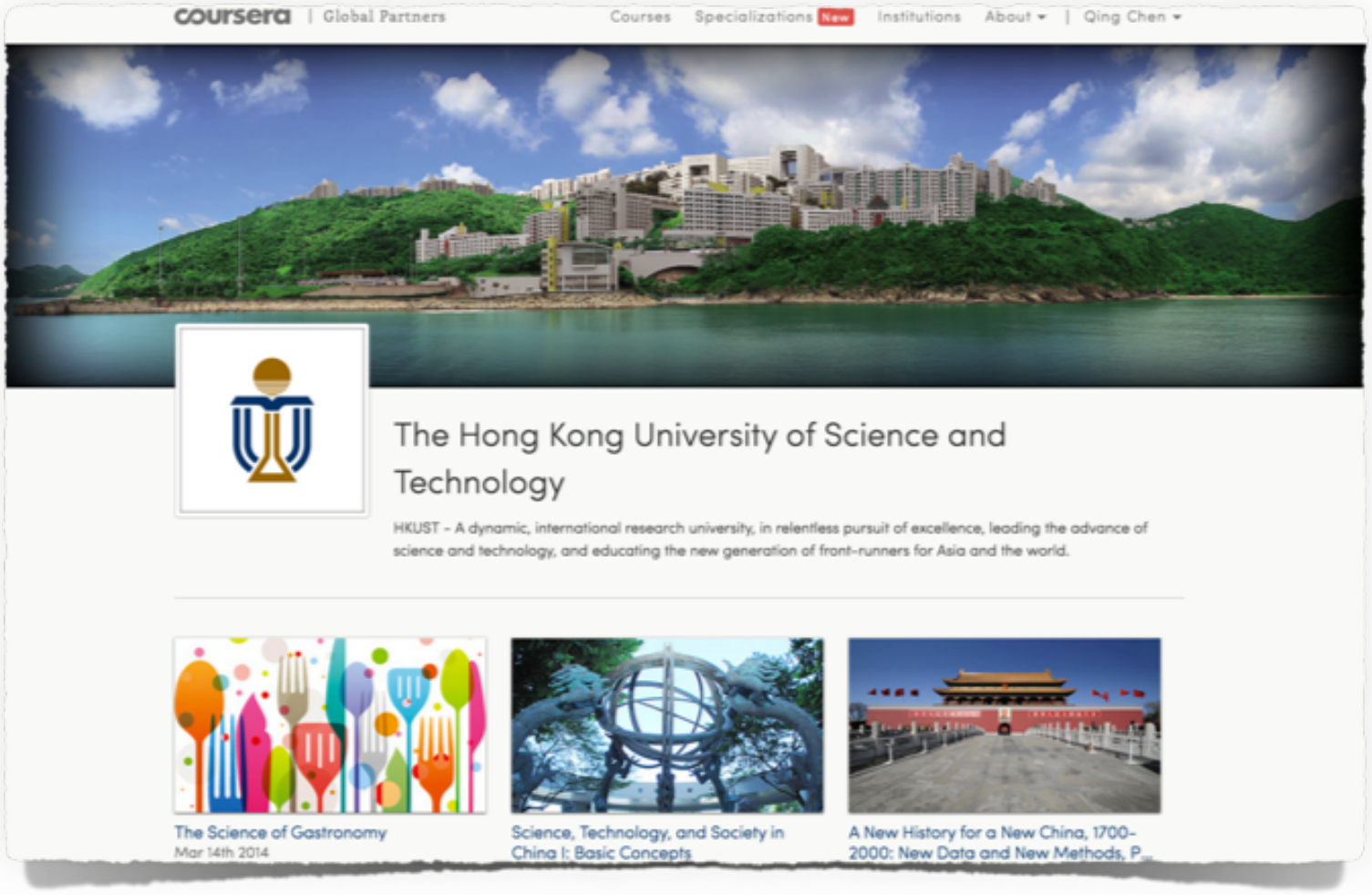
Does your heart follow your mind?





- Both male and females do not care about: age, location and height, contrary to their stated preferences
- Females are the most strict with marriage and children status, as well as education level of male senders.

HKUST on Coursera




A New History for a New China, 1700-2000
Revolutionary Victims in Shuangcheng and Elsewhere
0:00 / 9:30
Stacked
Stream
Expanded
play
pause
seeked
ratechange
stalled
error
0'50"
1'40"
2'30"
3'20"
4'10"
5'00"
5'50"
6'40"
7'30"
8'20"
9'10"

Which course has more difficult homework assignments?

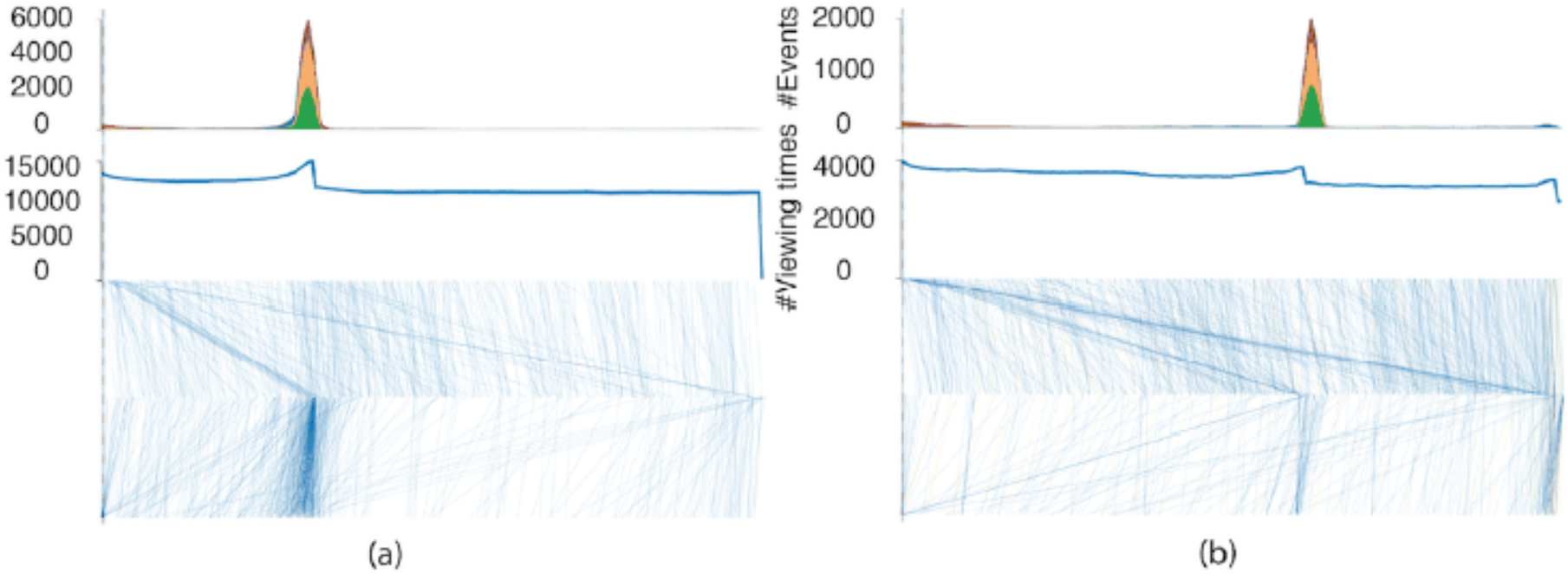


Figure 3: Comparison between the Content-based views of two videos with a in-video question. We can see that: there are a consid-

US and Chinese Students Behaves Differently?

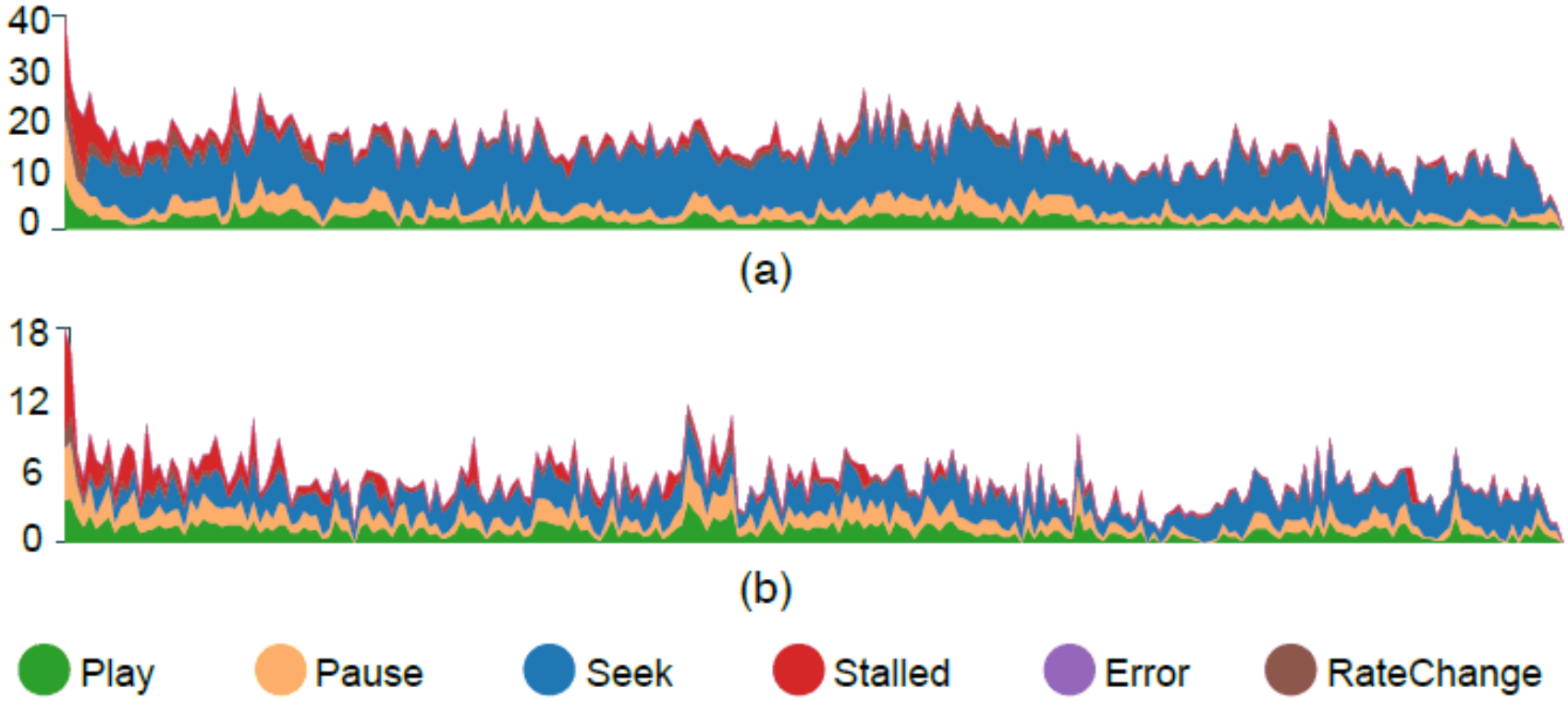
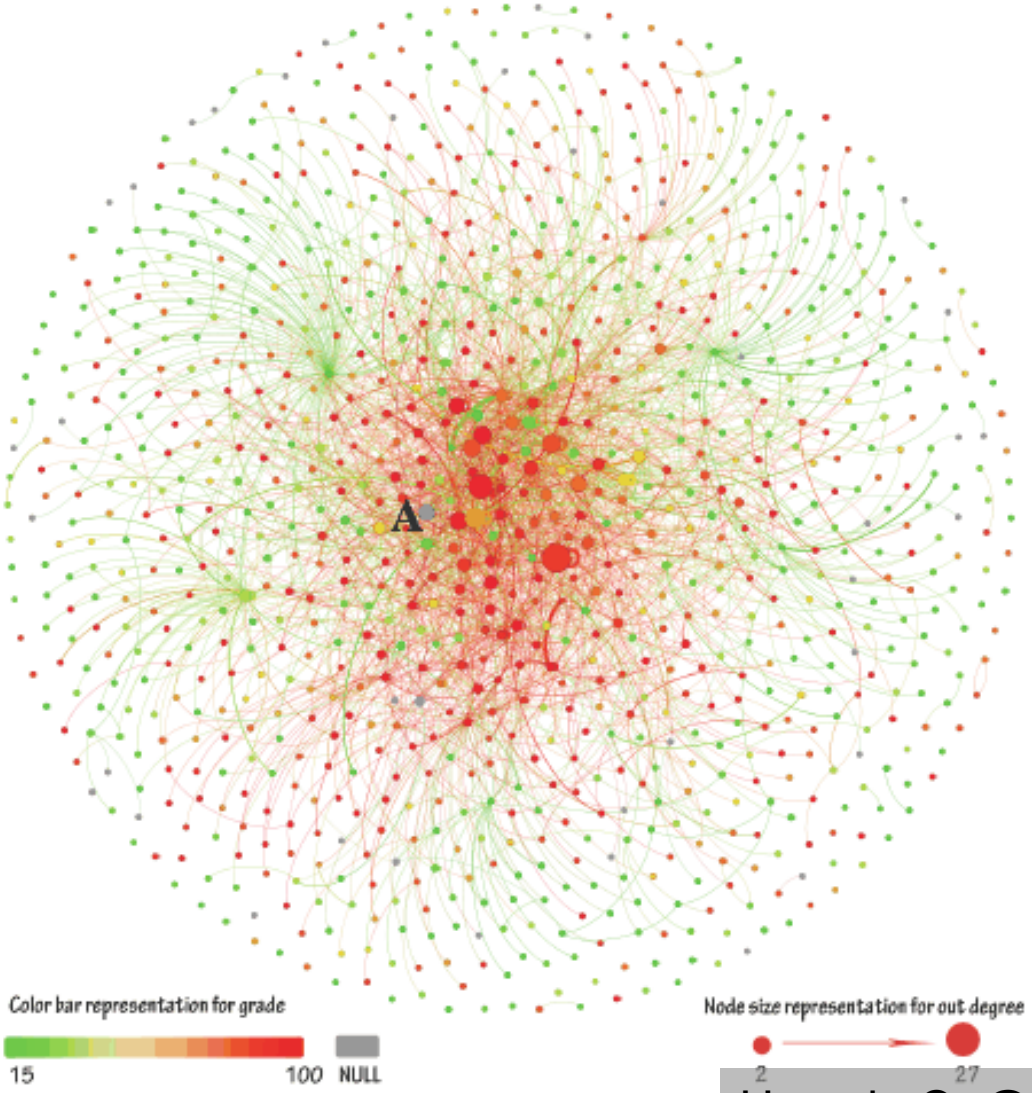
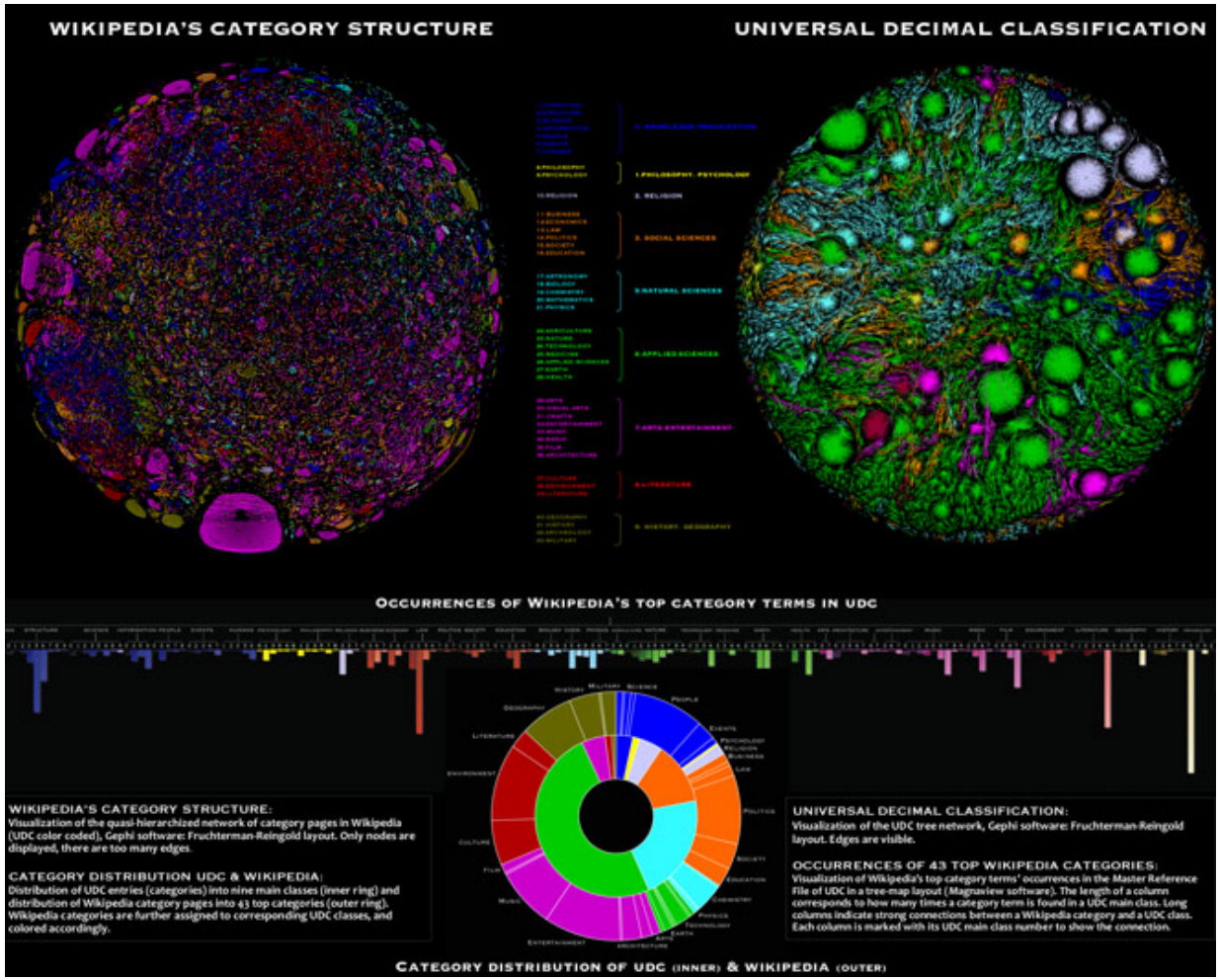


Figure 2: The Event Graphs showing the clickstream data of the same course during the same time period but for learners different countries. a) Learners from the U.S; b) Learners from China.

Are Good Students More Vocal?



Are Concepts Different in Wikipedia and Library?



Almila Akdag Salah/Cheng Gao/Krzystof Suchecki/Andrea Scharnhorst/ Den Haag/Knowledge Space Lab [high-resolution version]

Where Do Scientists Collaborate Most Often?

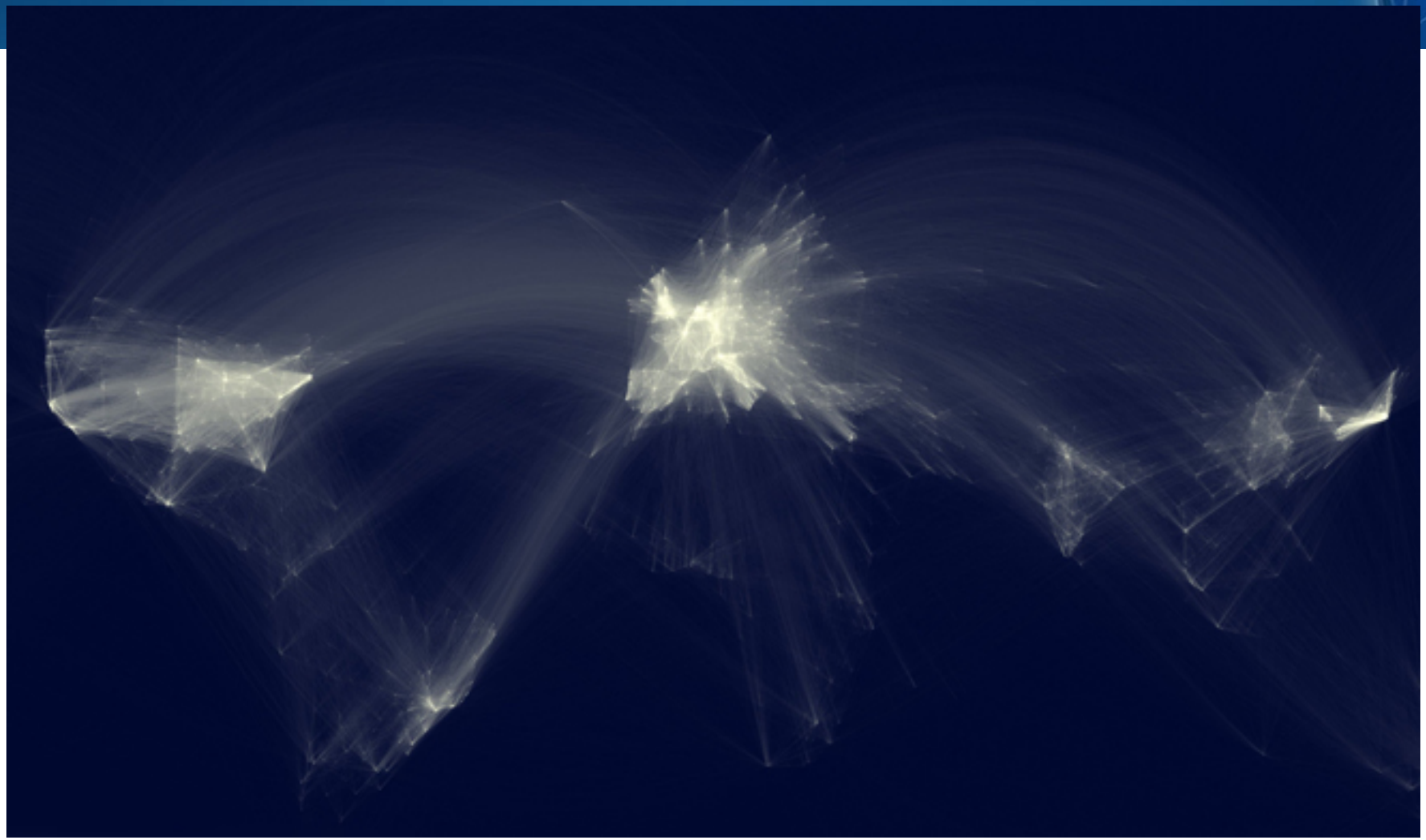
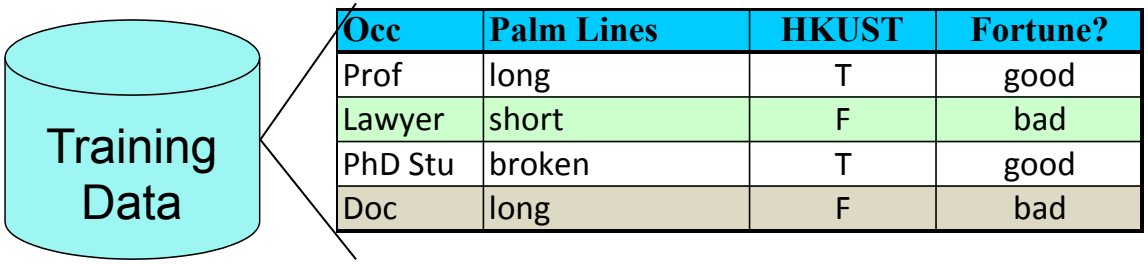


Image: Olivier H. Beauchesne/Science-Metrix [high-resolution version]

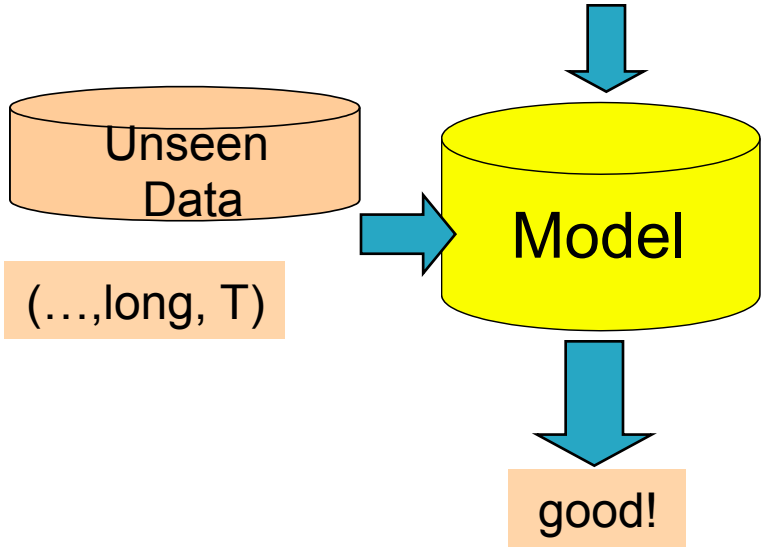
Data Analytics

	Prediction
	• Regression • Classification • Recommendation

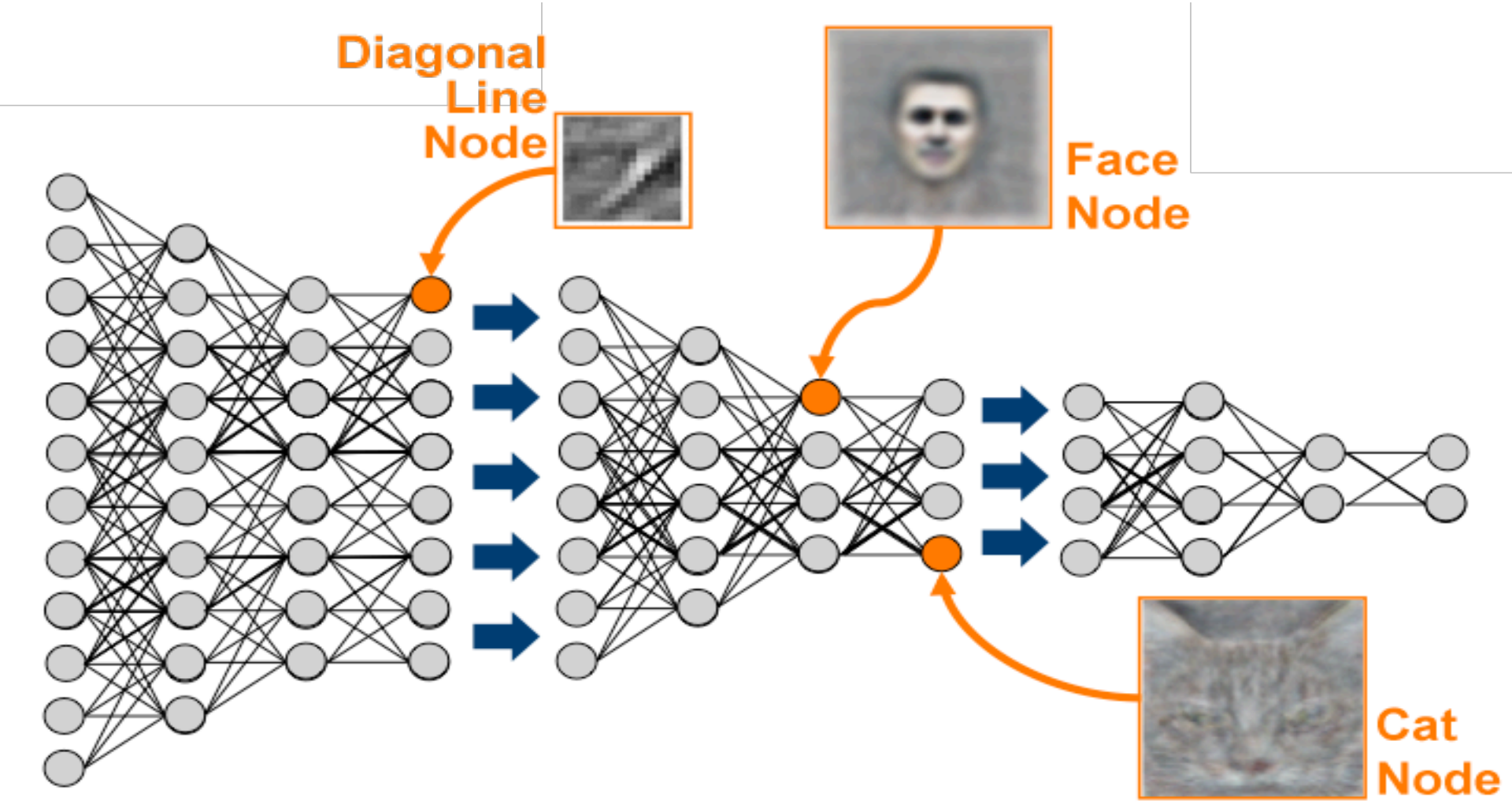
What's Common between a Data Scientist and a Fortune Teller?

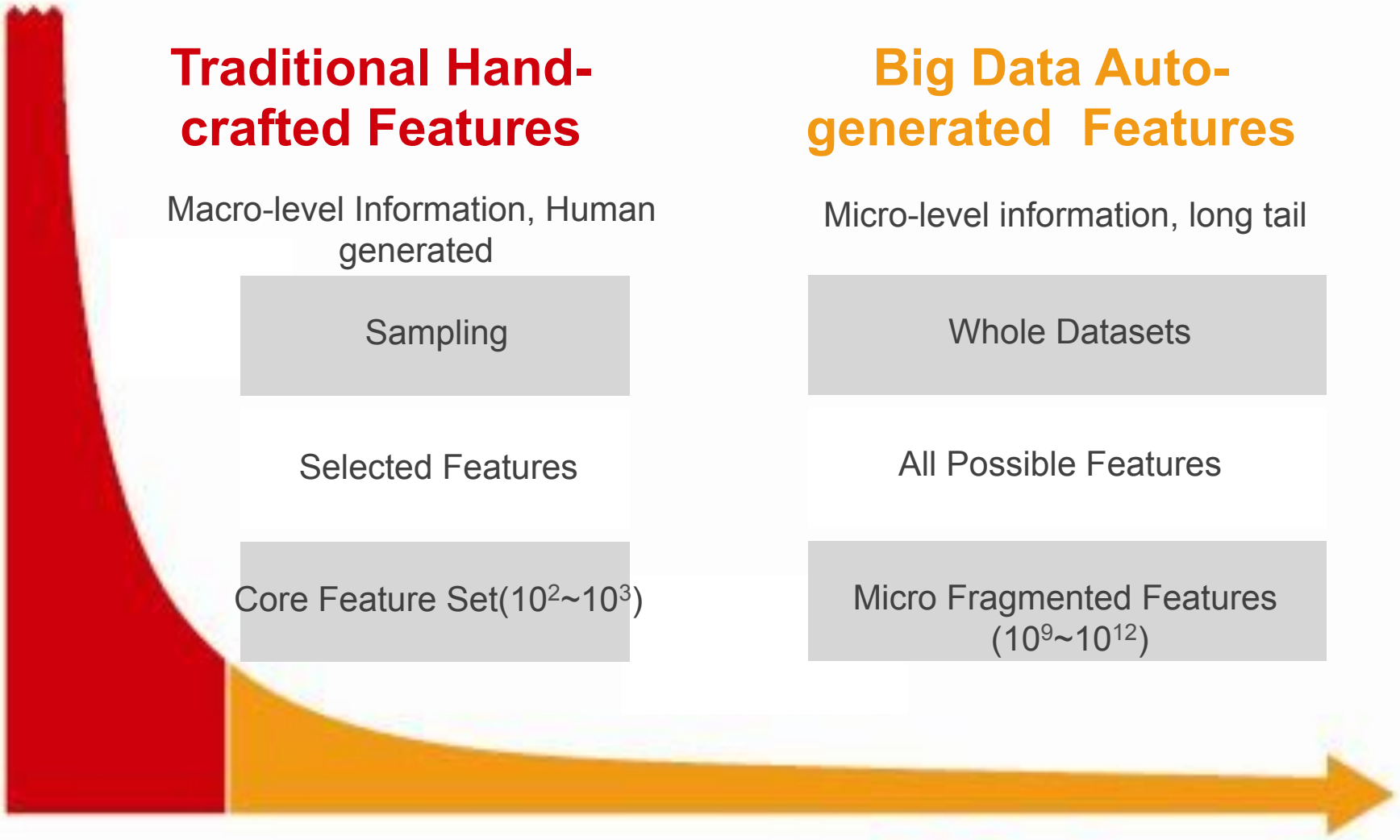


Occ	Palm Lines	HKUST	Fortune?
Prof	long	T	good
Lawyer	short	F	bad
PhD Stu	broken	T	good
Doc	long	F	bad



Deep Learning




Traditional Hand-crafted Features
Macro-level Information, Human generated
Sampling
Selected Features
Core Feature Set(10^2~10^3)
Big Data Auto-generated Features
Micro-level information, long tail
Whole Datasets
All Possible Features
Micro Fragmented Features (10^9~10^{12})

Feature Engineering

Over 100 Billion Features

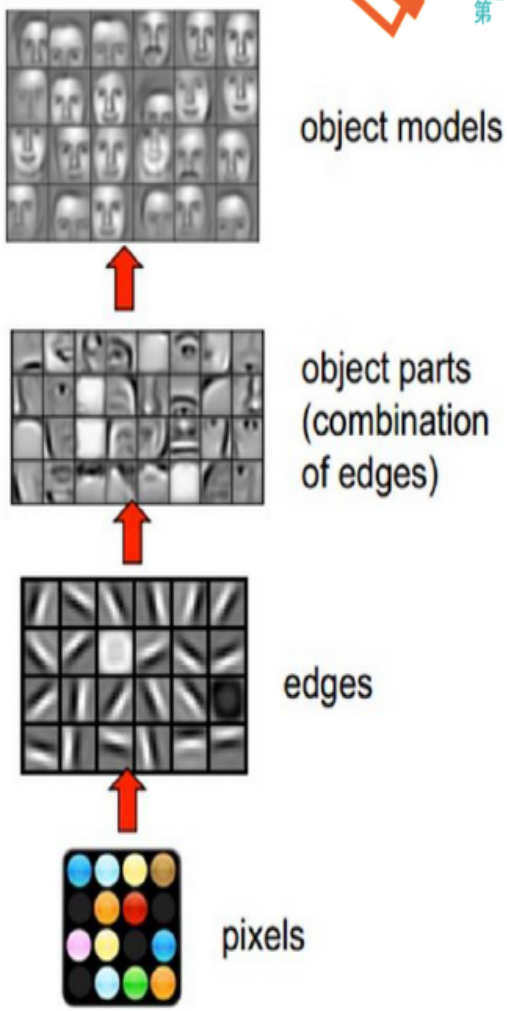
$$x = (1,0,0,0,1,\ldots,1,0,\ldots,0,1,0,0)$$

User Demographic

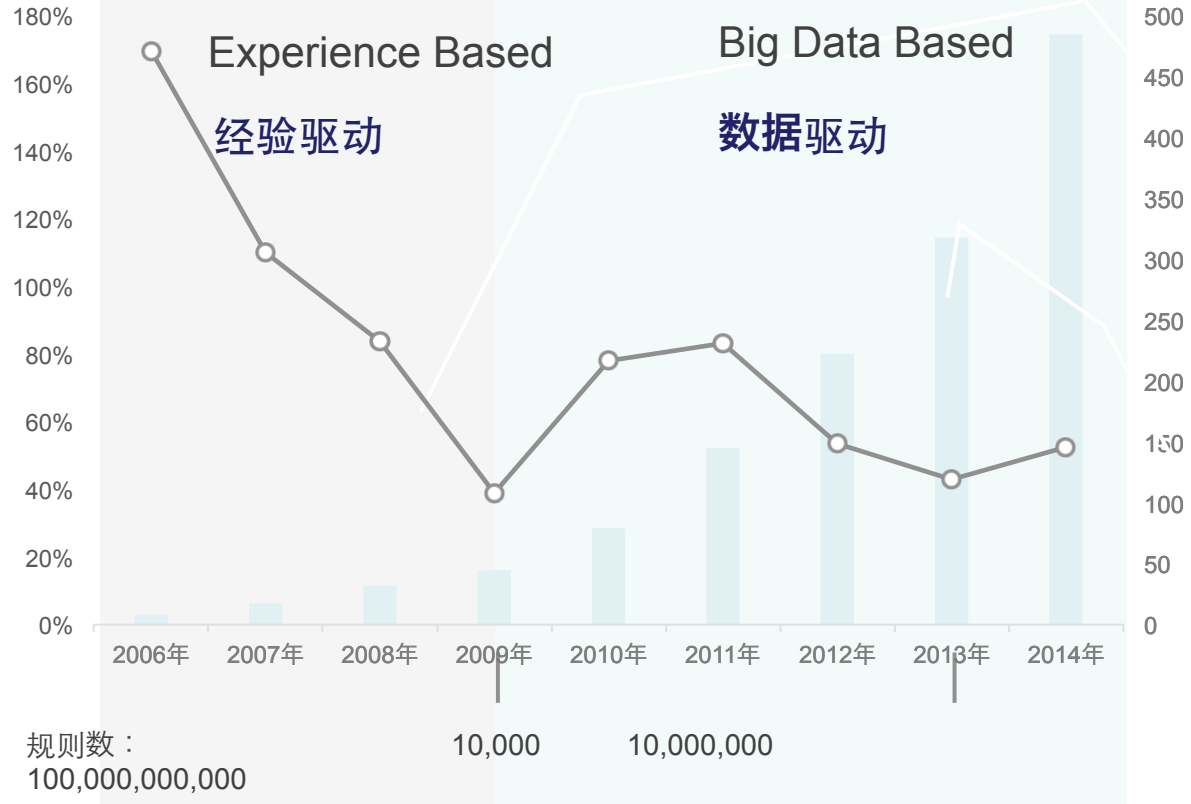
User Relevance to Target

Recent User Behavior

Human Managed
Vs.
Automatically Generated and Maintained



Baidu：Big Data ➔ New Growth



Driving Force：

Internet Users an Ads **Increase**

Effectiveness Gain from Big Data

2009：1 server online, #features: 10,000, income increase = **40%**

2013：20,000 server cluster, # features: 100,000,000,000, income increase = **800%**

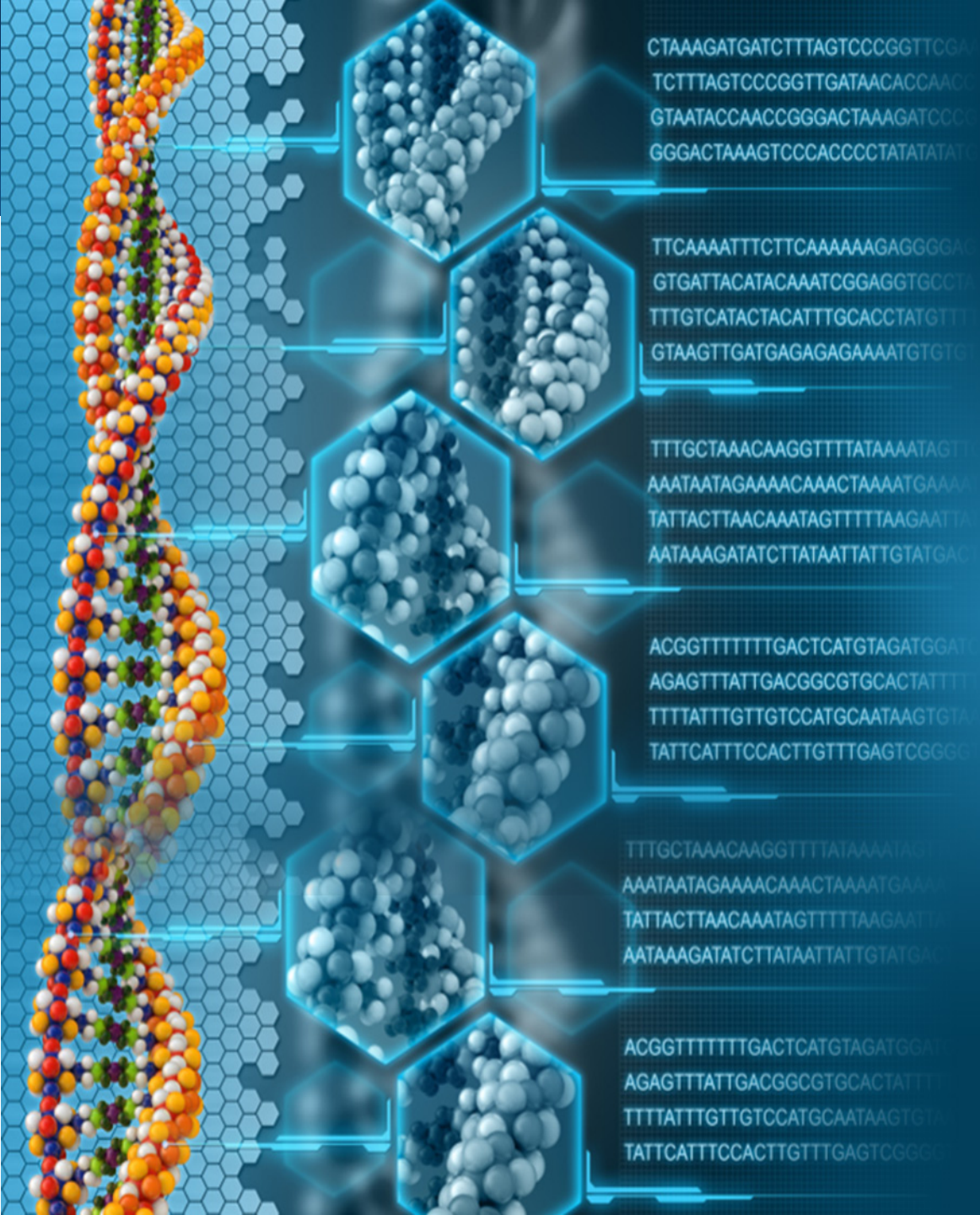
Transaction Installment Precision Marketing

Target: Millions of credit card transactions per day,

Action: Installment recommendation via SMS

Income:**+61.7%**

Note：same ratio of SMS



Genetic Big Data ATCG vs 0101

DNA: Data Source

DNA: Most Effective Data Storage

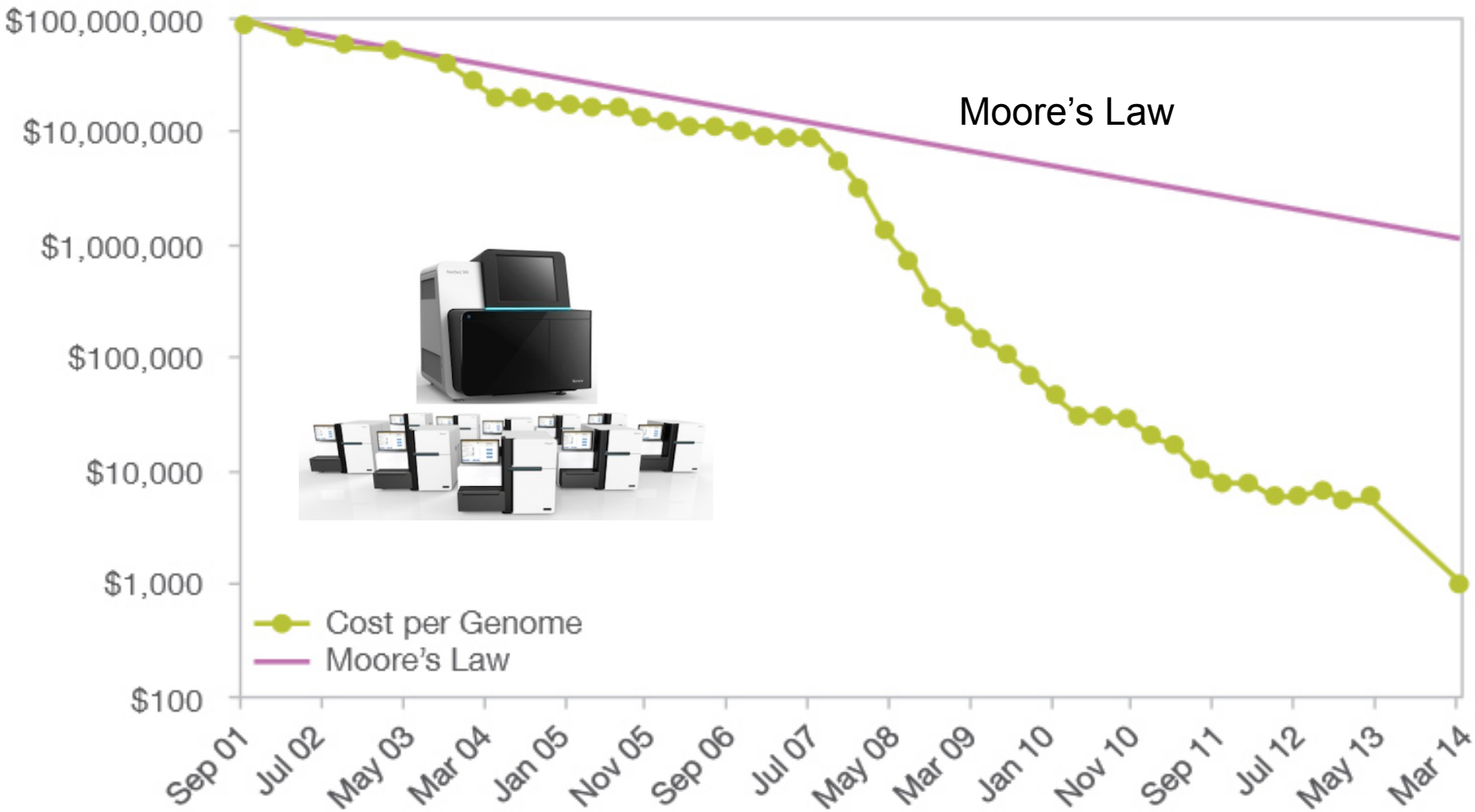
0.000,000,000,001,5g DNA
3,000,000,000 Bytes

Distance between: 3.4 Å (0.34nm)

Under certain conditions,
DNA data can be stored for
several million years!

Wang Jun@BGI Shenzhen

We can now read our genome at low cost, but understanding it is enormously difficult



Professor Mingjie Zhang

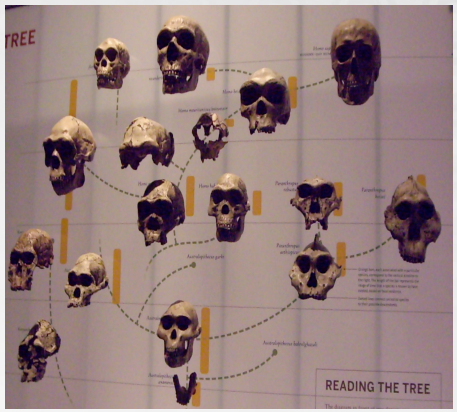
But Understanding it is the most difficult!

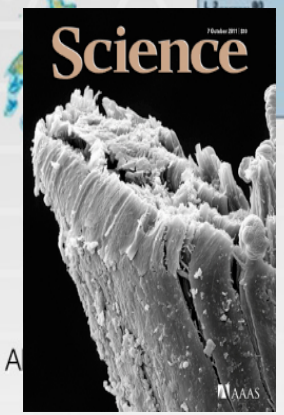
Life's Equation!

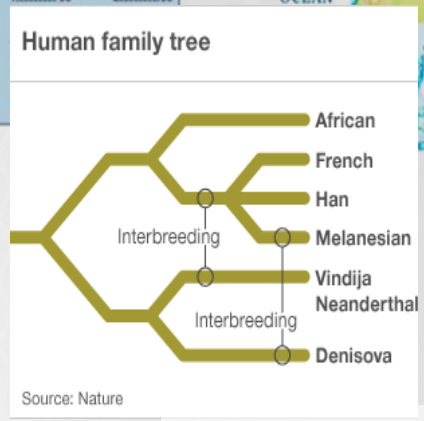
Wang Jun@BGI Shenzhen

Wang Jun@BGI Shenzhen

Understand How Humans Migrate







Wang Jun@BGI Shenzhen

Common Variations by GWAS

Published Genome-Wide Associations through 12/2012
Published GWA at p≤5X10^{-8} for 17 trait categories

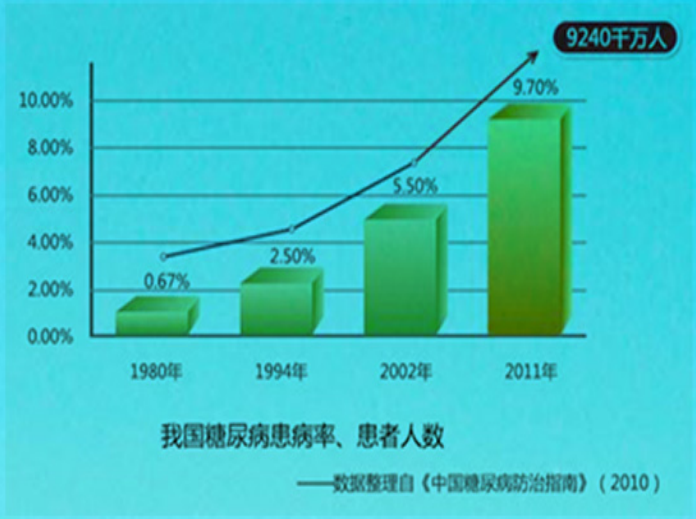
SNPS and GWAS

- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

National Human Genome Research Institute

EMBL-EBI

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/



ARTICLE

doi:10.1038/nature11450

A metagenome-wide association study of gut microbiota in type 2 diabetes

Junjie Qin[1]*, Yingrui Li[2]*, Zhiming Cai[3]*, Shenghui Li[1]*, Jianfeng Zhu[1]*, Fan Zhang[2]*, Suisha Liang[2], Wenwei Zhang[2], Yuanlin Guan[2], Dongqian Shen[1], Yangqing Peng[2], Dongya Zhang[2], Zhuye Jie[1], Wenxian Wu[2], Youwen Qin[1], Wenbin Xue[2], Junhua Li[2], Lingchuan Han[2], Donghui Lu[4], Peixian Wu[4], Yali Dai[4], Xiaojuan Sun[2], Zesong Li[2], Aifa Tang[4], Shilong Zhong[4], Xiaoping Li[5], Weineng Chen[1], Ran Xu[1], Mingbang Wang[1], Qiang Feng[2], Meihua Gong[2], Jing Yu[2], Yanyan Zhang[2], Ming Zhang[1], Torben Hansen[7], Gaston Sanchez[8], Jeroen Raes[7,8], Gwen Falony[7,8], Shujiro Okuda[7,8], Mathieu Almeida[9], Emmanuelle LeChatelier[9], Pierre Renault[9], Nicolas Pons[9], Jean-Michel Batto[9], Zhaoxi Zhang[2], Hua Chen[1], Ruifu Yang[3,10], Weimou Zheng[1], Songgang Li[2], Huanming Yang[2], Jian Wang[2], S. Dusko Ehrlich[9], Rasmus Nielsen[6], Oluf Pedersen[5,11,12], Karsten Kristiansen[1,13] & Jun Wang[2,3,13]

Cure for Diabetes?

Wang Jun@BGI Shenzhen


Thus, we need
big data!
Life Big Data
碳基
Wang Jun@BGI Shenzhen

Power

Endurance

VO2max

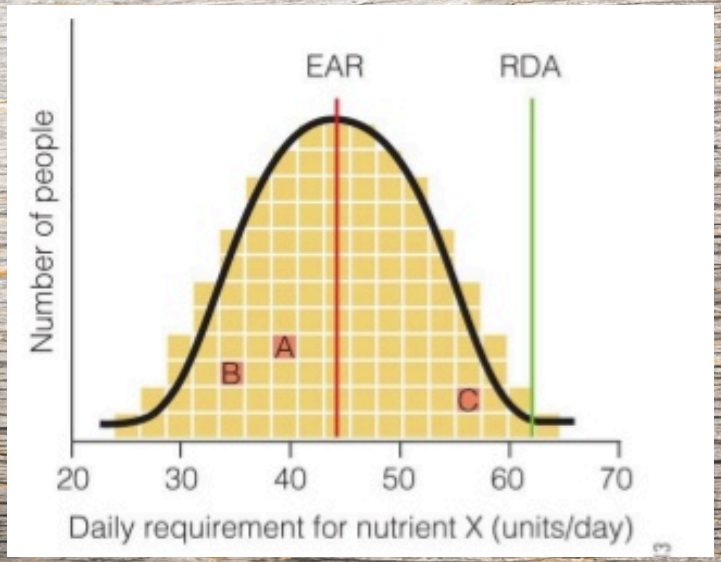
Recovery

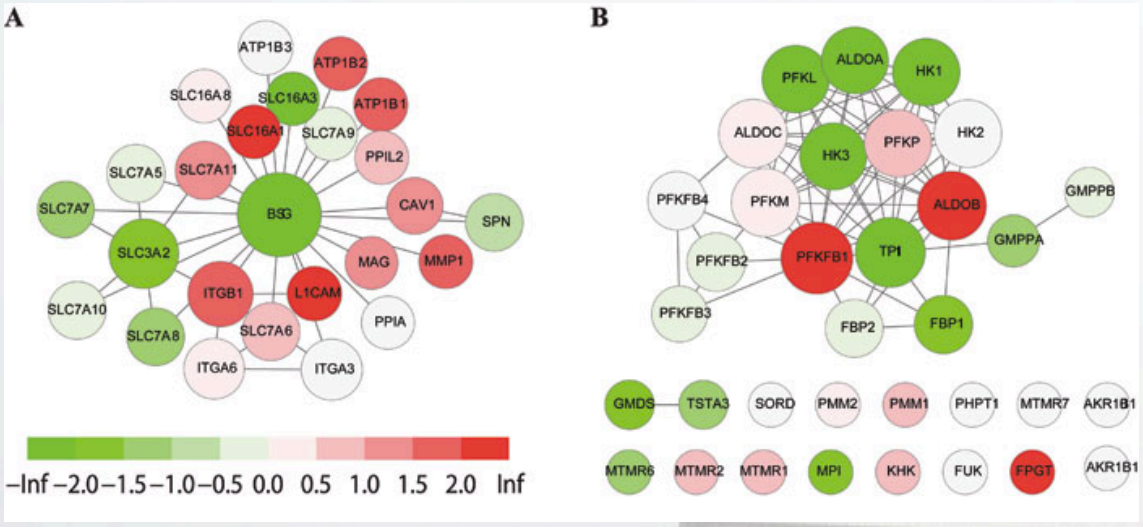
Sickness Risks

Gene+ Life Style

Wang Jun@BGI Shenzhen

Gene + Life Style





Gene + Environment

Wang Jun@BGI Shenzhen

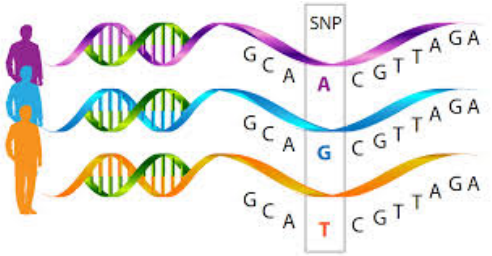
Next-generation sequencing and big data analysis of human diseases (Alzheimer's disease)

Next-generation sequencing

AD Controls

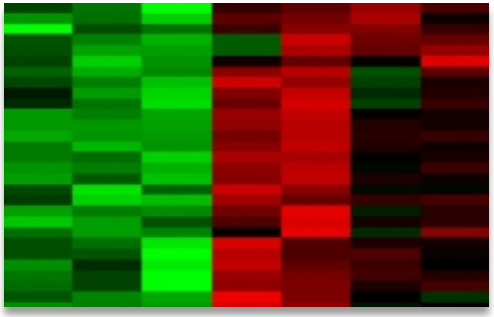
Whole blood samples

Big data analysis



Identify genetic variations

Big data analysis



- Quantitative analysis
- Differential expression of genes

Potential biomarkers

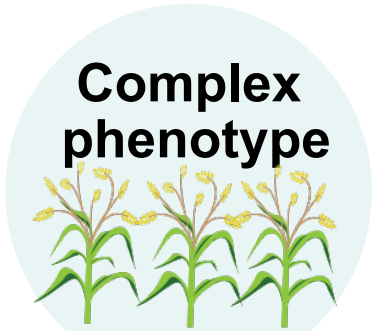
Nancy Ip@HKUST

Project Millet: HKUST and BGI Shenzhen

Input: Very high dimension and low sample size labeled data ($N \approx 2000$, $D \approx 240K$)

Machine Learning

Task: Train an accurate phenotype predictor using genetic data.



Accelerate Hybridization Breeding

E.g. Facilitate understand biology process

markers underlying specific phenotype.

How to Educate a Data Scientist?

- Find structure in data
  - Can use statistical learning tools
  - Can generate new features
    - Combinational features
    - From 10 to 10 Billion
    - Samples and Features
- Build good models
  - Model must fit type of data
    - Text, signal, images, speech, social media, tables, etc.
  - Knows how to trade off between model complexity and data quantity

How to Educate a Data Scientist?

- Focus on **questions**, not methods
  - Given the data, what questions are interesting?
  - For questions, what tools are available?
- Knows how to **interpret** results
  - Explain black box models to non-computer experts

Conclusions: Big Data + Science = Data Science

- Data Scientists
  - Collect Data
  - Understand the data
  - Generate Features
  - Build and test models
  - Explain the results
  - Find new knowledge!