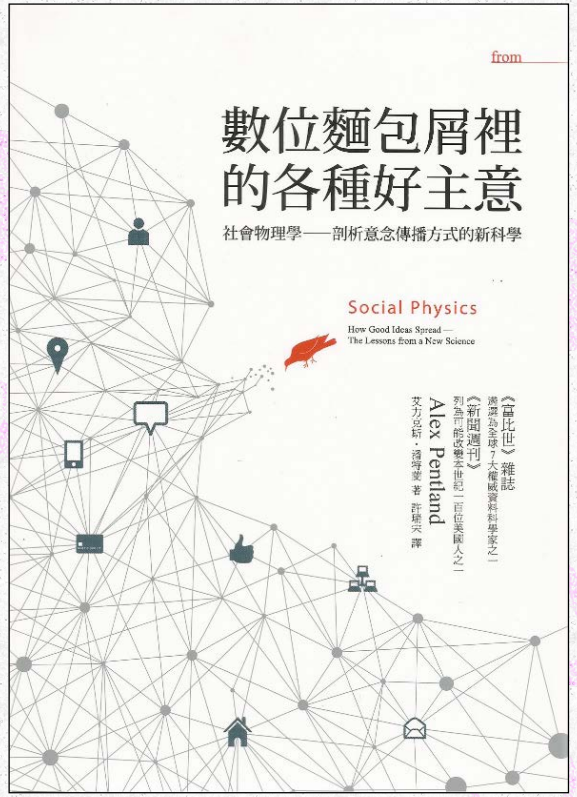
# 好書導讀




**陳昇瑋**

**台灣資料科學協會理事長**

**中央研究院資訊科學研究所研究員**

# 中央研究院

## Academia Sinica

= Chinese in Latin

## = Chinese Academy

- 32 research institutes in 3 major divisions

  1) mathematics, physics, and applied sciences;

  2) life sciences;

  3) humanities and social sciences.
- 1,000 tenure-tracked research fellows
- 4,000 assistants and technicians

# 中央研究院資訊科學研究所

# 中央研究院資訊科學研究所

## 組成

- 40 位研究員
- 30 位博士後研究員
- 300 位研究助理

資料洞察實驗室

## 研究領域

| | | |
|---|---|---|
| 演算法 | 資料科學 | 智慧型代理人 |
| 語音處理 | 中文認知 | 多媒體 |
| 生物資訊 | 系統技術 | 機器學習 |

Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience. Gulping the wine doesn't work. -Daniel B. Wright (2003)

## 最新消息

- 歡迎參加 2015 台灣資料科學愛好者年會！
- 徵才啟事：研究助理 / 博士後研究人員 / 軟體工程師
- 徵才啟事：博士後研究人員
- IEEE Spectrum: Reducing World of Warcraft Power Consumption
- 雜誌專訪：iT人甘苦談—從程式人跨到學術人的深度 歷險

## 研究領域

大數據分析及應用

多媒體系統

使用者經驗

社群計算及計算社會學

群眾外包 / 人智運算

## 關於我們

有人訪問 Prof. H.J. Eysenck (一位知名的心理學家)：

「當某個研究帶來令您滿意的成果時，您的感覺如何 ？」

他的回答是：「那種感覺就好像貓得到奶油一樣，很 難加以形容。我想那是一種充盈的美好感覺，你覺得 一切都是那麼地令人愉快。我個人對自己原先預測的 某些東西終於成為事實，多少會感到驚訝，因為我總 覺得那不太可能，但現在卻成為事實，實在值得慶祝 。所以你會對自己感到高興，對自然、對整個世界感 到高興，你覺得生命真美好，值得繼續活下去。」

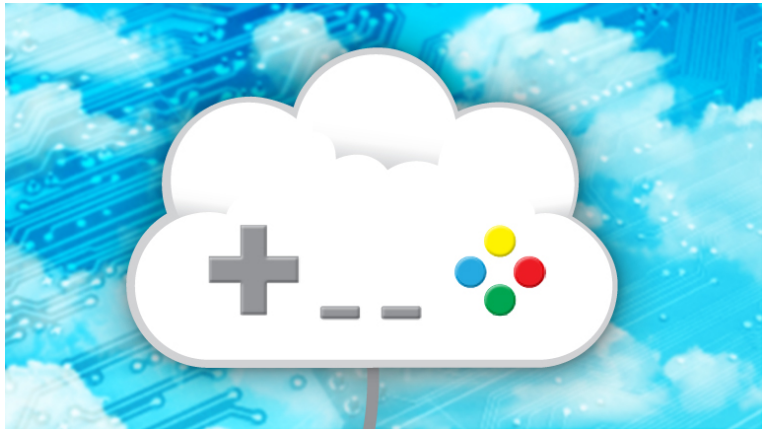
你也想一起來體會「貓 得到奶油」的感覺嗎？:-)

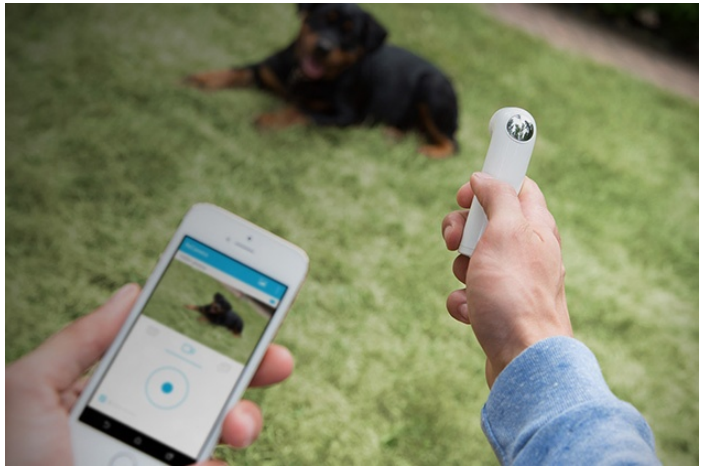
# Area 1: Quality of Experience

- 使用情緒量測技術來預言線上遊戲的成與敗

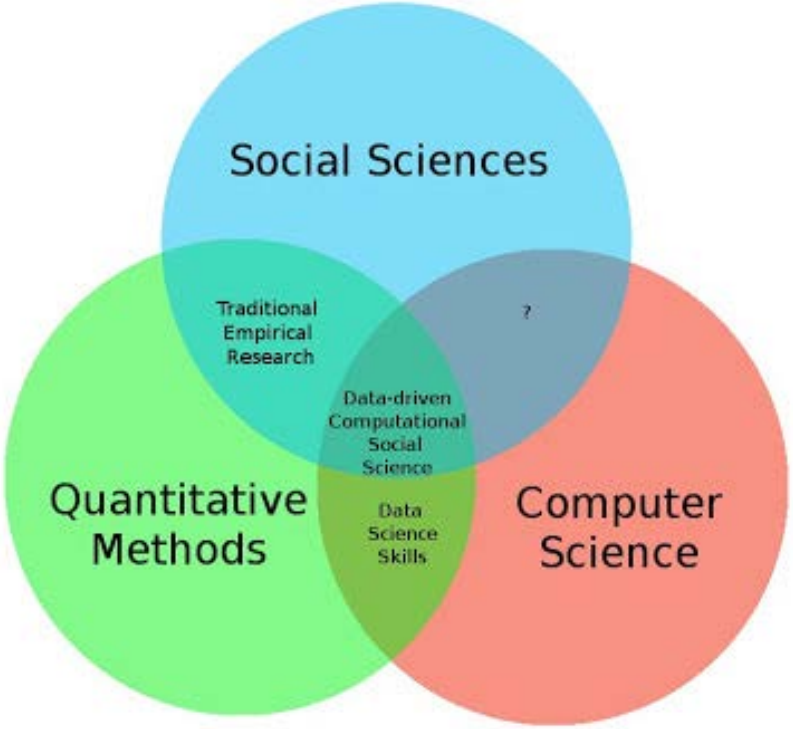
[1] Jing-Kai Lou, Kuan-Ta Chen, Hwai-Jung Hsu, and Chin-Laung Lei, Forecasting Online Game Addictiveness, IEEE/ACM NetGames 2012.

# Area 2: Multimedia Systems





# Area 3: Computation Social Science




"The emerging intersection of the social and computational sciences, an intersection that includes analysis of web-scale observational data, virtual lab–style experiments, and computational modeling" [1].

[1] Duncan J. Watts, Computational Social Science Exciting Progress and Future Directions, Frontiers of Engineering, Winter 2013.

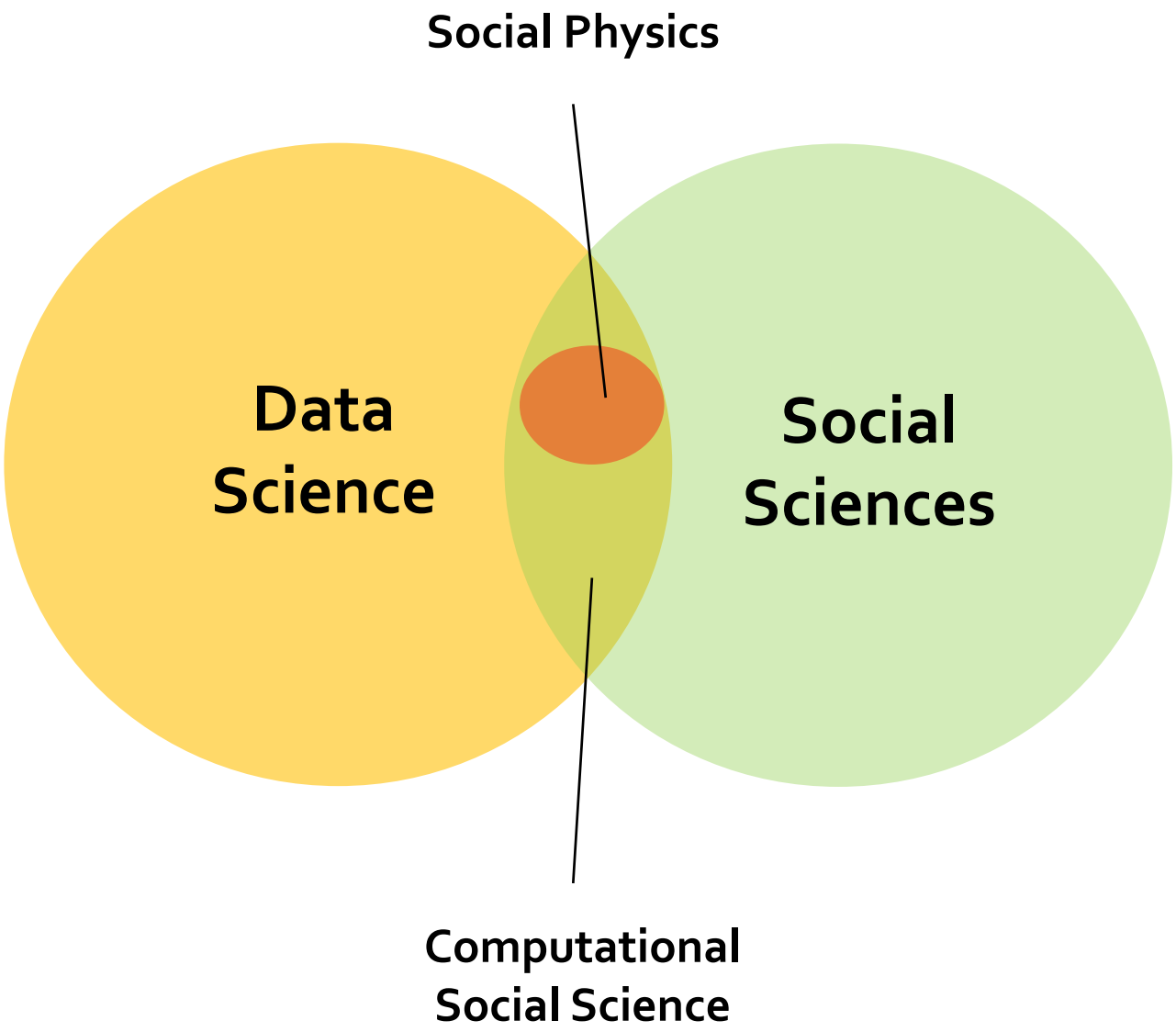
# 資料分析這條路

- Since 2002 (my first PhD year) …
- PhD dissertation: based on a 20-hour game packet trace
- Collaboration & Consulting
  - IT 製造
  - 網路 / 手機遊戲
  - 社群網路服務
  - 多媒體消費性產品
  - 中央 / 地方政府
  - 銀行 / 壽險 / 電子票証

# 完整投影片線上閱讀 / 下載

http://www.iis.sinica.edu.tw/~swc/talk/data_science_overview.html

# 好，進入正題


Social Physics
Data Science
Social Sciences
Computational Social Science

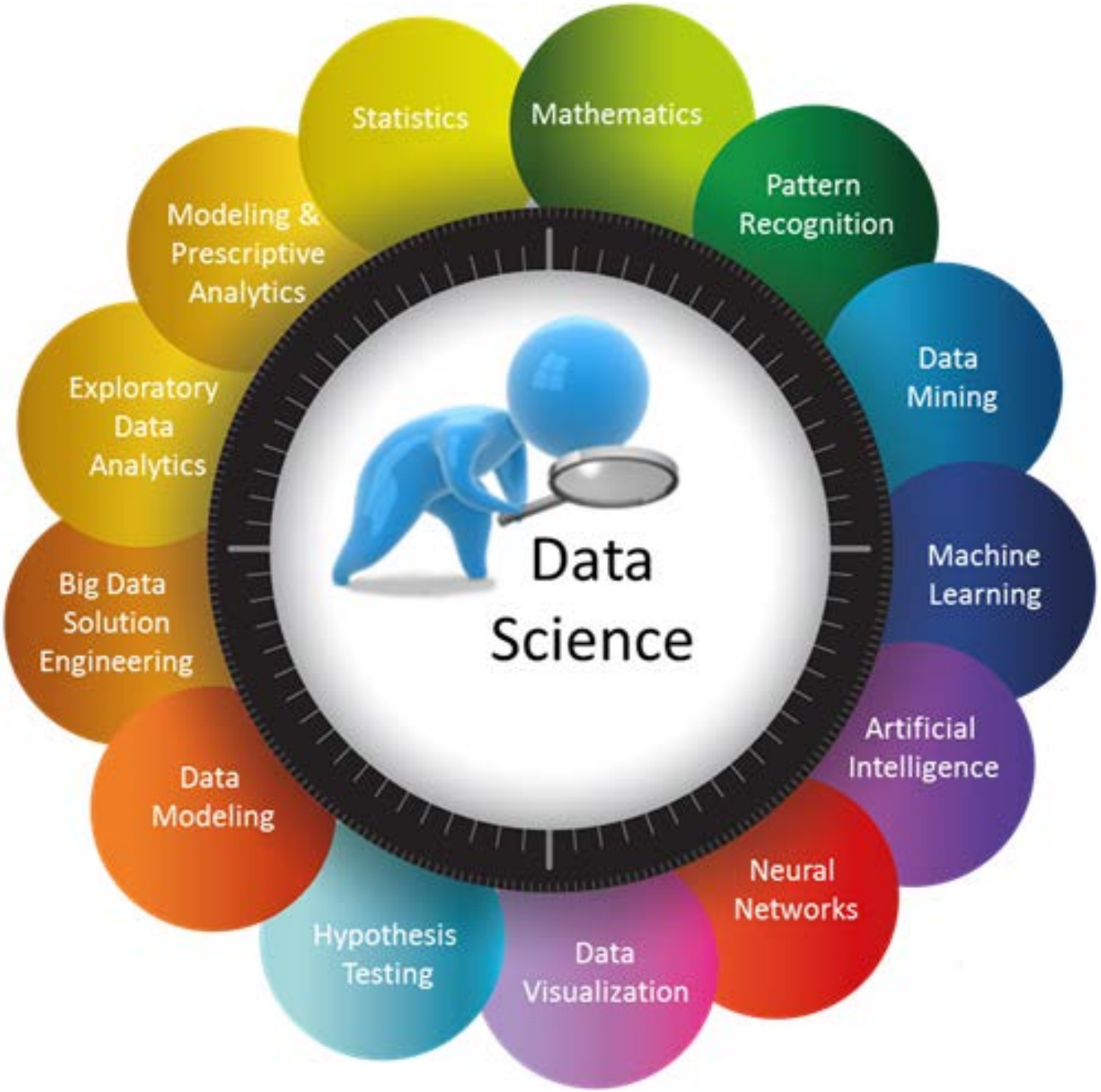
# What's Data Science?

# Definition of "Science"

Science is a systematic enterprise that builds and organizes knowledge in the form of general, measureable and verifiable explanations and predictions about the universe.

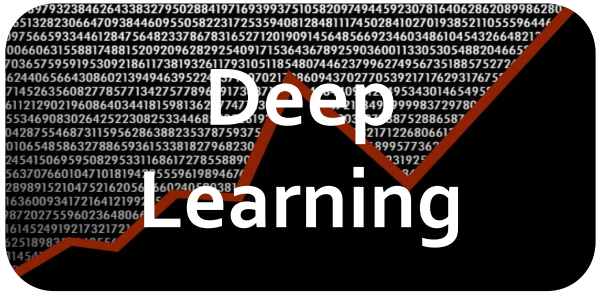
In modern usage "science" most often refers to a way of pursuing knowledge, not only to the knowledge itself. Over the course of the 19th century, the word "science" became increasingly associated with the scientific method itself.

P&H
MinePro

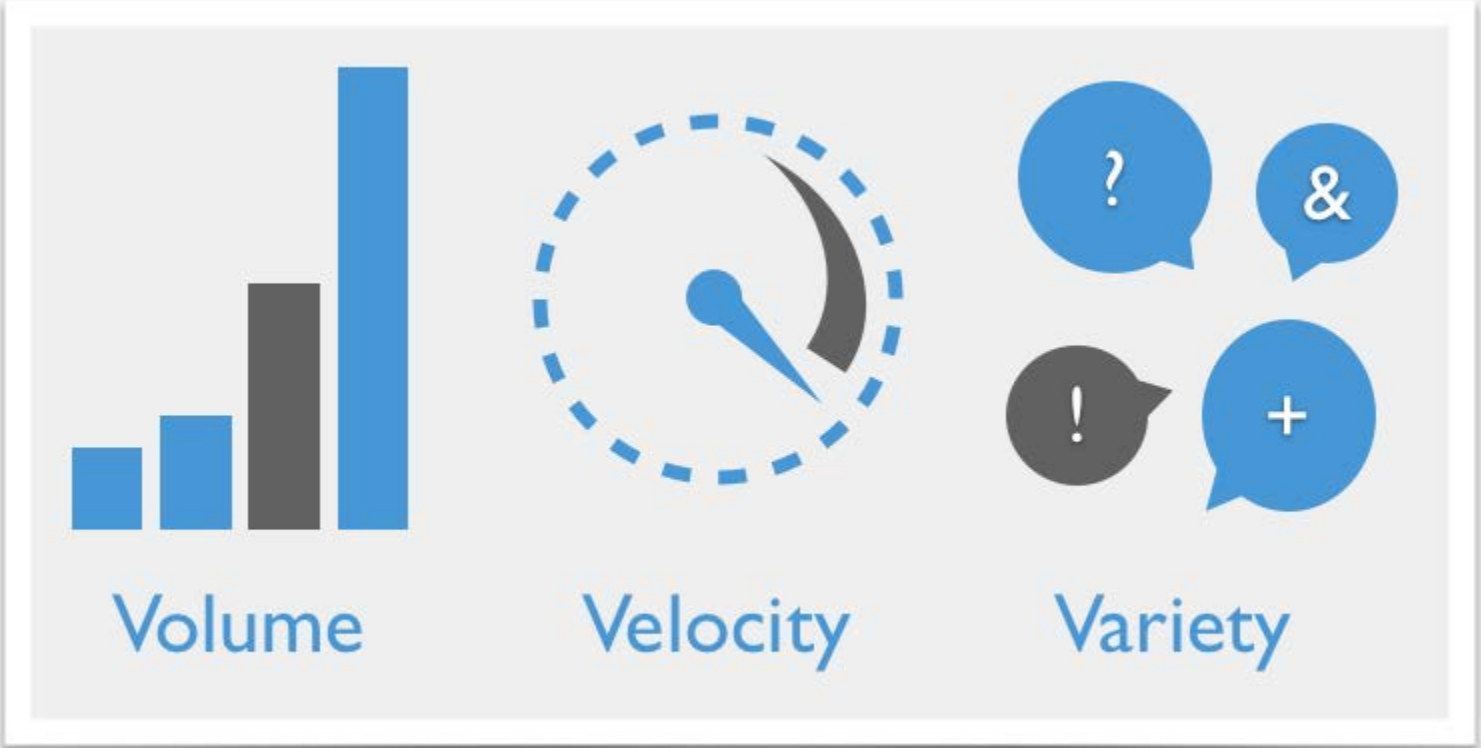

Statistics
Mathematics
Pattern Recognition
Data Mining
Machine Learning
Artificial Intelligence
Neural Networks
Data Visualization
Hypothesis Testing
Data Modeling
Big Data Solution Engineering
Exploratory Data Analytics
Modeling & Prescriptive Analytics
Data Science

Electric
re-charging
lane

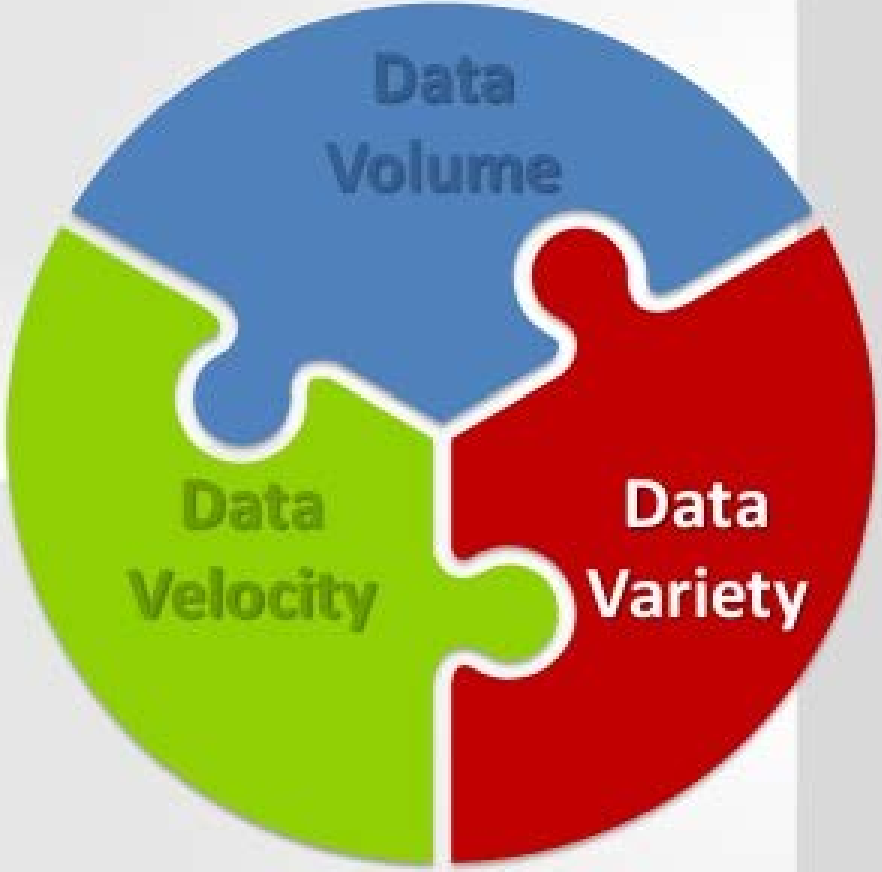
# 3 Major Trends of Data Science



# 3 Major Trends of Data Science (#1)



# 3V of Big Data explained




## Volume

Huge data size, terabytes – petabytes

## Velocity

High speed of data flow, data change and data processing

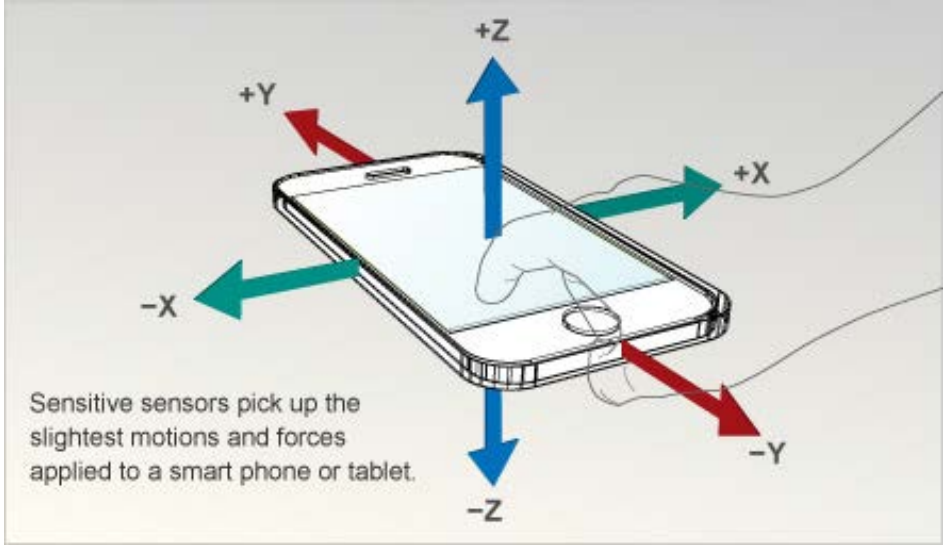
## Variety

Various data sources (social, mobile, M2M, structured and unstructured data)

Copyright: infoDiagram.com

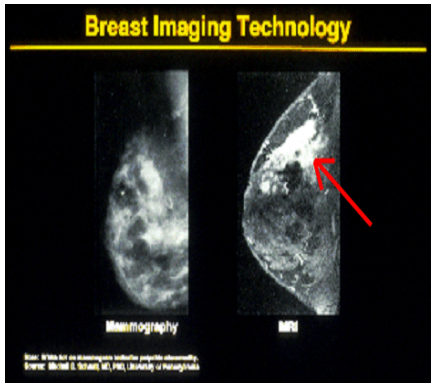
# Why Big Data Arises?

- Collecting & storing data is much cheaper now
- Massive number of online users
- Open data from governments
- New types and wide deployed sensors



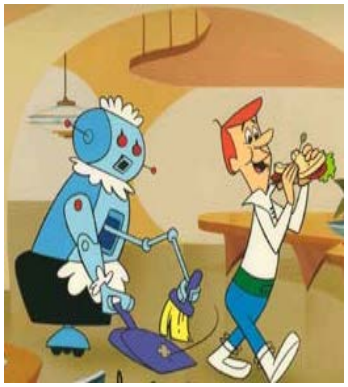
# Computer Vision Matters

Safety




Health

Security



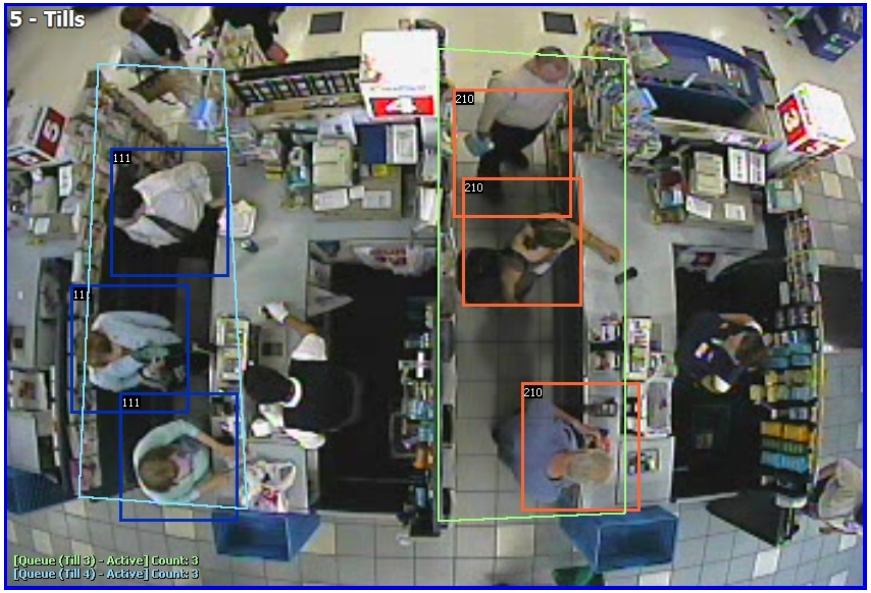
Comfort

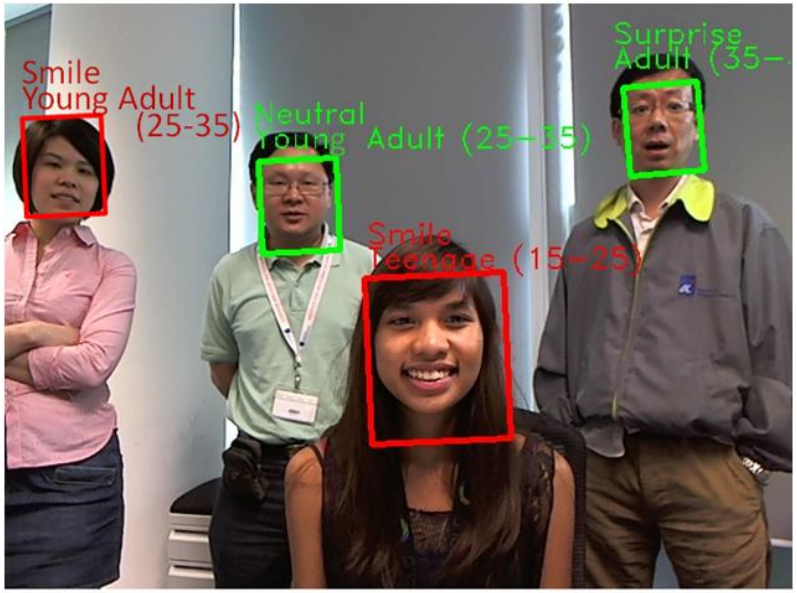
Fun

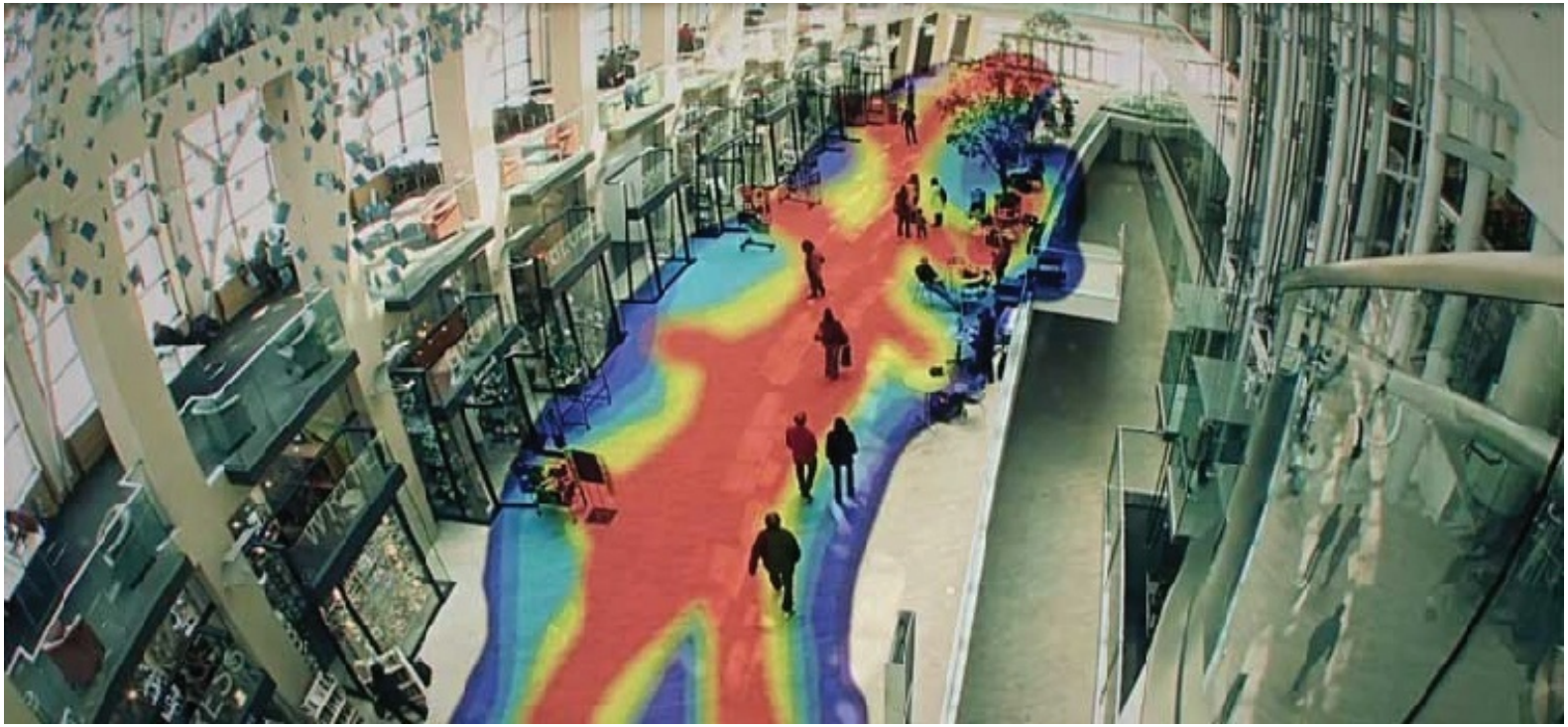
Access



(Slide Credit: Jia-Bin Huang)


5 - Tills
111
111
111
210
210
210
[Queue (Till 3) - Active] Count: 3
[Queue (Till 4) - Active] Count: 3


Smile
Young Adult
(25-35)
Neutral
Young Adult (25-35)
Smile
Teenage (15-25)
Surprise
Adult (35-



# Computer vision in sports

SportVision: improving viewer experiences



(Slide Credit: Jia-Bin Huang)

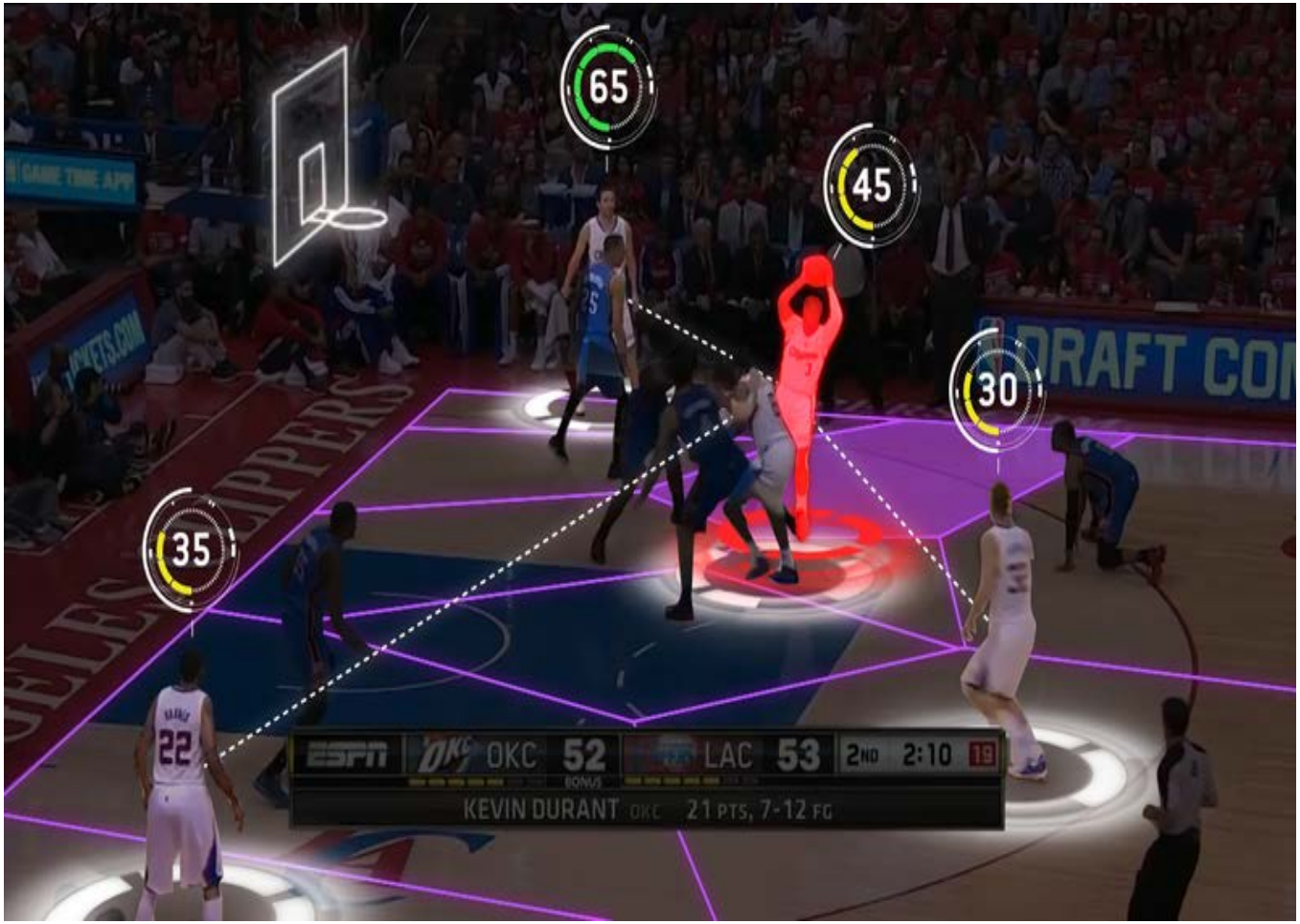
# Computer vision in sports

Play tracking



(Slide Credit: Jia-Bin Huang)

# Computer vision in sports




Second Spectrum: visual analytics



(Slide Credit: Jia-Bin Huang)

# Computer vision in sports

Replay Technologies: improving viewer experiences



(Slide Credit: Jia-Bin Huang)

# Computer vision for healthcare

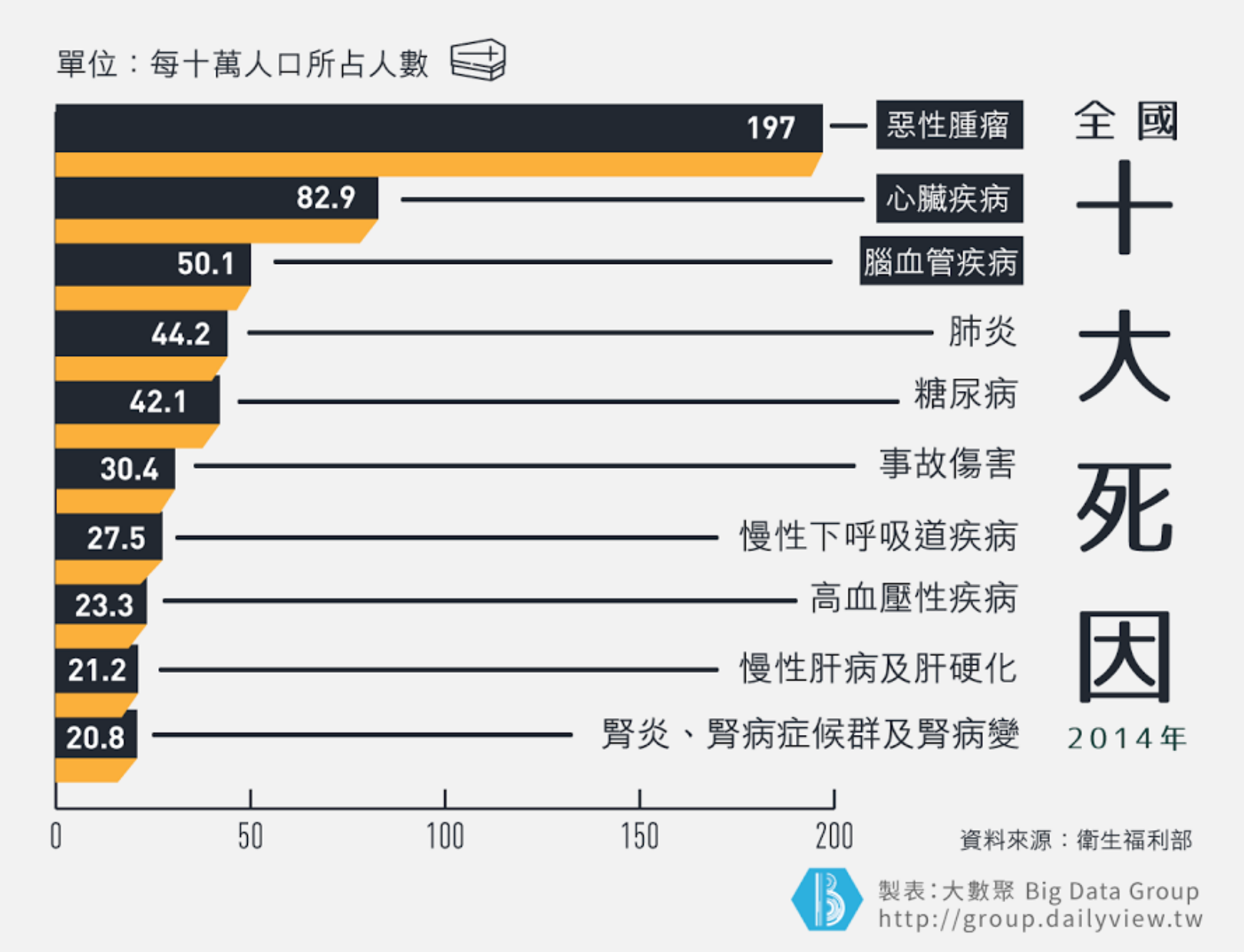
Video magnification

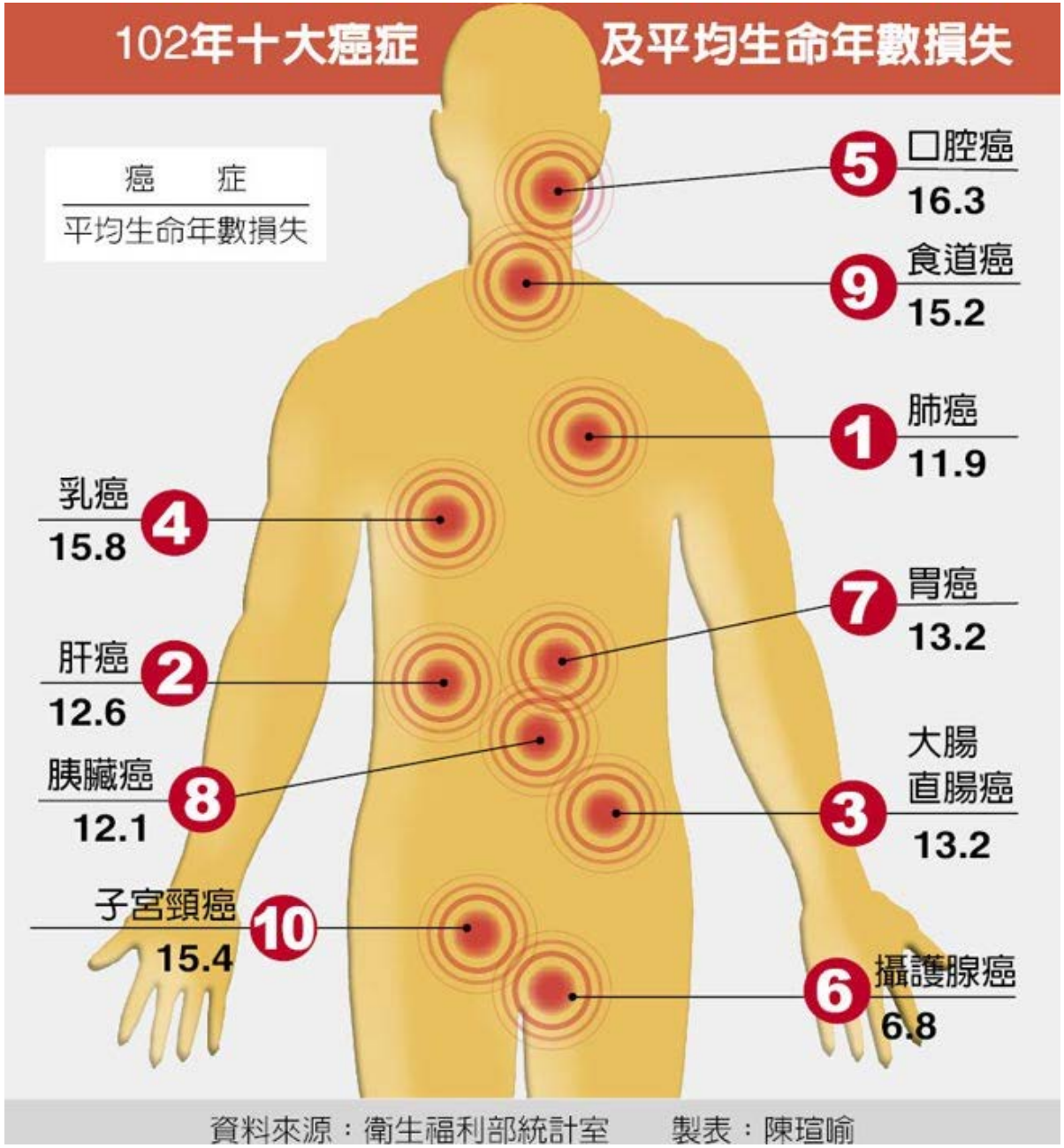


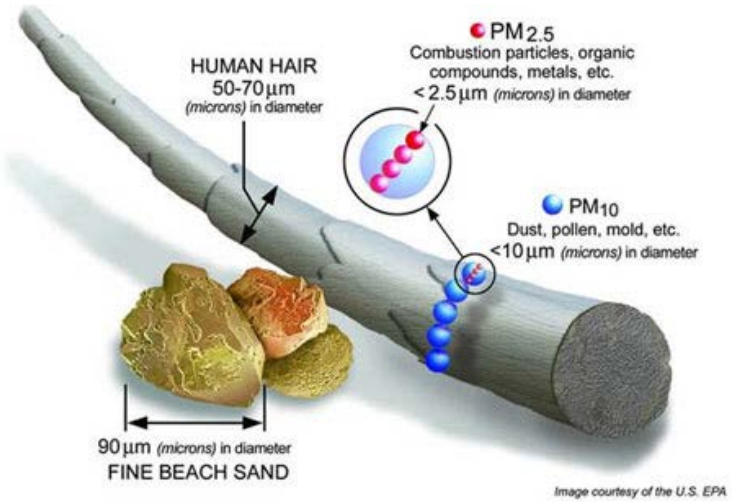
(Slide Credit: Jia-Bin Huang)

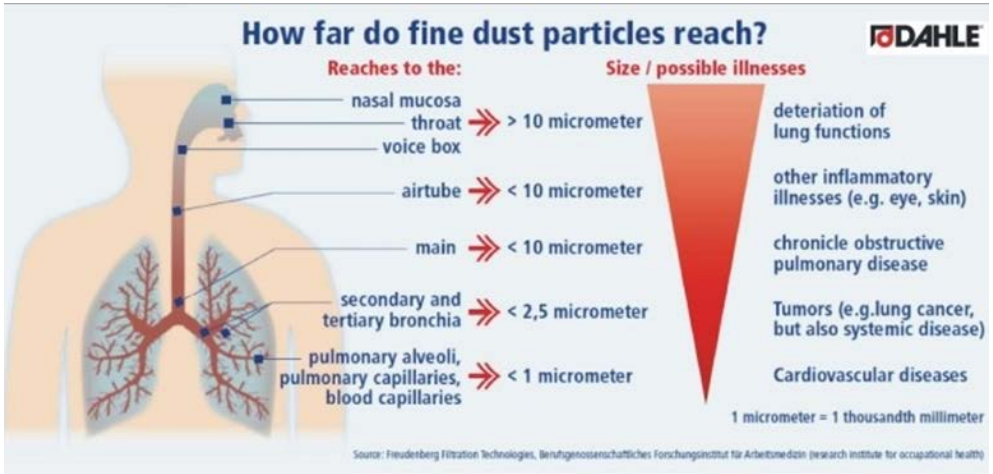
https://www.youtube.com/watch?v=QbXgEbeceJI

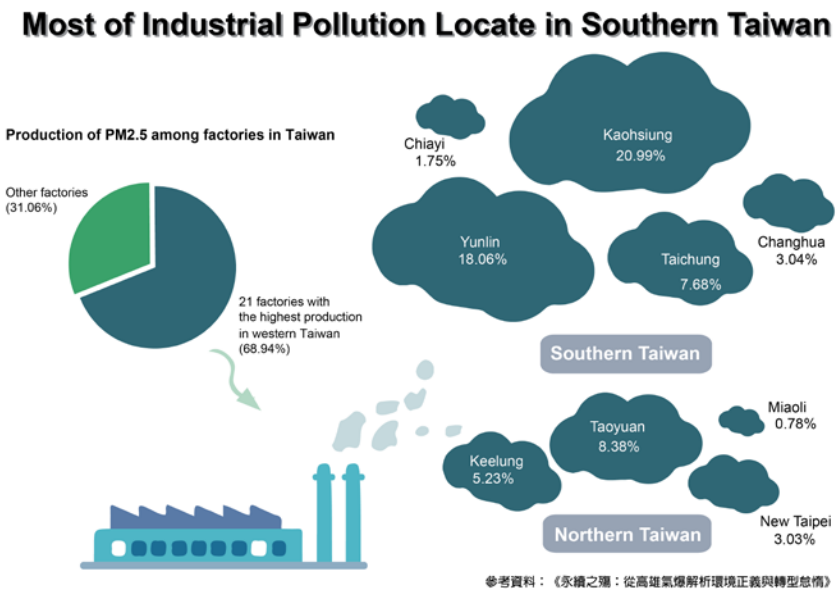
(Credit: Jia-Bin Huang)

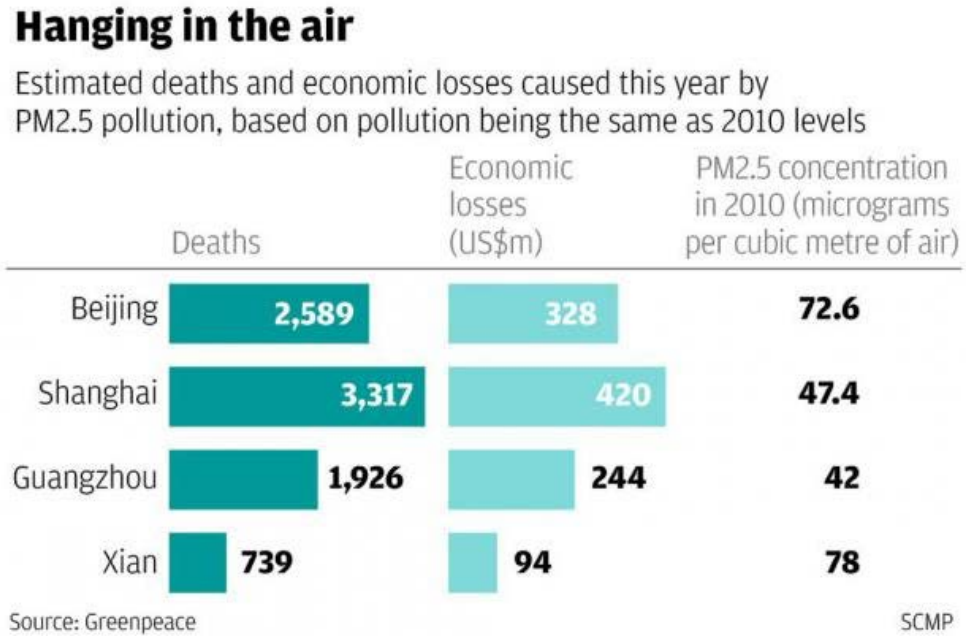

單位：每十萬人口所占人數
全國
十大死因
2014年
197
惡性腫瘤
82.9
心臟疾病
50.1
腦血管疾病
44.2
肺炎
42.1
糖尿病
30.4
事故傷害
27.5
慢性下呼吸道疾病
23.3
高血壓性疾病
21.2
慢性肝病及肝硬化
20.8
腎炎、腎病症候群及腎病變
0
50
100
150
200
資料來源：衛生福利部
製表：大數聚 Big Data Group
http://group.dailyview.tw


102年十大癌症 及平均生命年數損失
癌 症
平均生命年數損失
5 口腔癌 16.3
9 食道癌 15.2
1 肺癌 11.9
4 乳癌 15.8
7 胃癌 13.2
2 肝癌 12.6
8 胰臟癌 12.1
3 大腸直腸癌 13.2
10 子宮頸癌 15.4
6 攝護腺癌 6.8
資料來源：衛生福利部統計室 製表：陳瑄喻

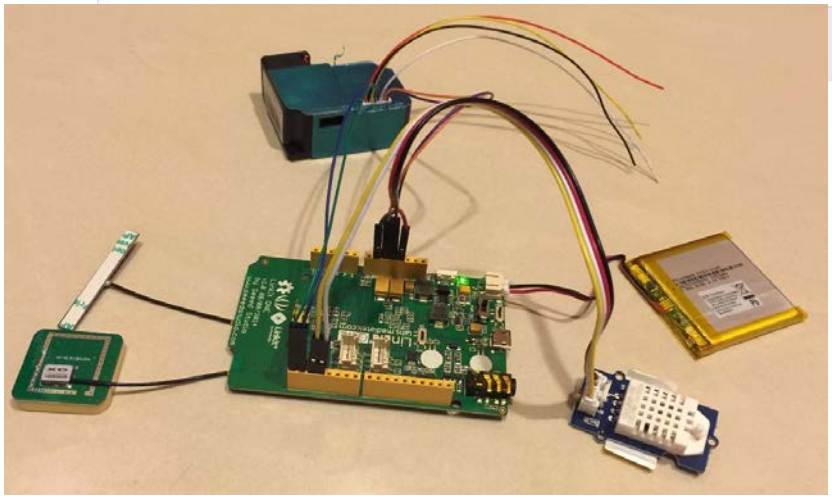

HUMAN HAIR
50-70µm
(microns) in diameter
PM2.5
Combustion particles, organic compounds, metals, etc.
<2.5µm (microns) in diameter
PM10
Dust, pollen, mold, etc.
<10µm (microns) in diameter
90µm (microns) in diameter
FINE BEACH SAND
Image courtesy of the U.S. EPA


How far do fine dust particles reach?
DAHLE
Reaches to the:
Size / possible illnesses
nasal mucosa
throat
voice box
> 10 micrometer
airtube
< 10 micrometer
main
< 10 micrometer
secondary and tertiary bronchia
< 2,5 micrometer
pulmonary alveoli, pulmonary capillaries, blood capillaries
< 1 micrometer
deteriation of lung functions
other inflammatory illnesses (e.g. eye, skin)
chronicle obstructive pulmonary disease
Tumors (e.g.lung cancer, but also systemic disease)
Cardiovascular diseases
1 micrometer = 1 thousandth millimeter
Source: Freudenberg Filtration Technologies, Berufsgenossenschaftliches Forschungsinstitut für Arbeitsmedizin (research institute for occupational health)

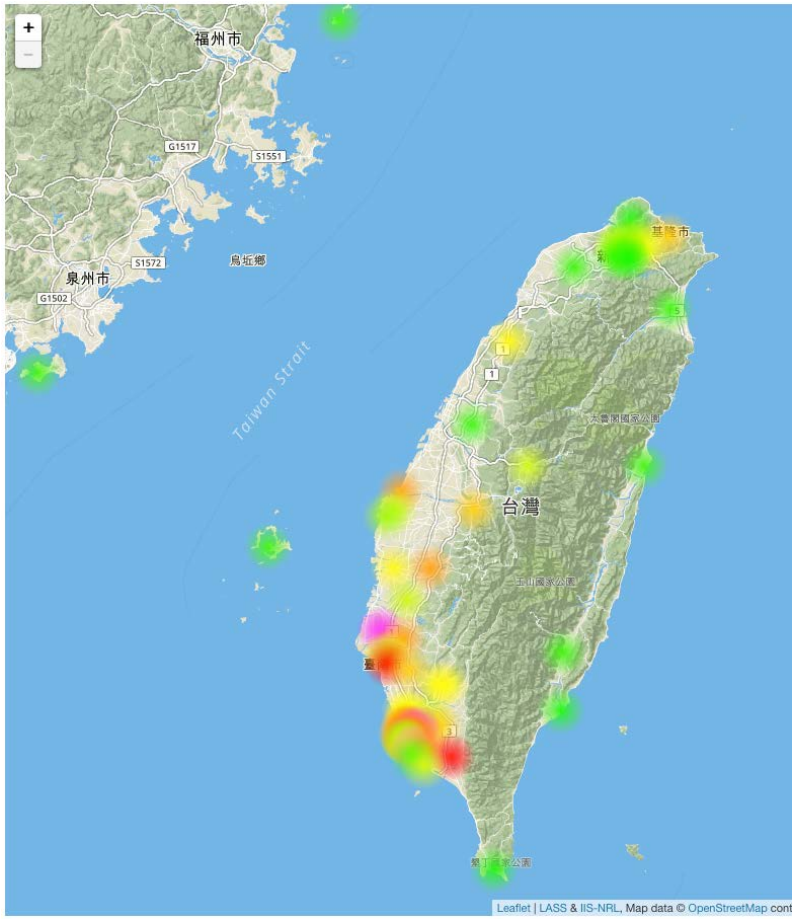

Most of Industrial Pollution Locate in Southern Taiwan
Production of PM2.5 among factories in Taiwan
Other factories (31.06%)
21 factories with the highest production in western Taiwan (68.94%)
Chiayi 1.75%
Kaohsiung 20.99%
Yunlin 18.06%
Taichung 7.68%
Changhua 3.04%
Southern Taiwan
Miaoli 0.78%
Taoyuan 8.38%
Keelung 5.23%
New Taipei 3.03%
Northern Taiwan
參考資料：《永續之殤：從高雄氣爆解析環境正義與轉型怠惰》


Hanging in the air
Estimated deaths and economic losses caused this year by PM2.5 pollution, based on pollution being the same as 2010 levels
Deaths
Economic losses (US$m)
PM2.5 concentration in 2010 (micrograms per cubic metre of air)
Beijing 2,589 328 72.6
Shanghai 3,317 420 47.4
Guangzhou 1,926 244 42
Xian 739 94 78
Source: Greenpeace
SCMP

LASS
Location Aware
Sensor System
LASS-開源公益的環境感測器網路
公開社團
已加入
分享
通知
討論區
成員
活動
相片
檔案
搜尋這個社團
撰寫貼文
新增相片／影片
建立票選活動
新增檔案
留個言吧.....
新增成員
+ 輸入姓名或電子郵件地址.....
成員
1,680 位成員 （183 位新成員）
傳送訊息 · 以電子郵件邀請

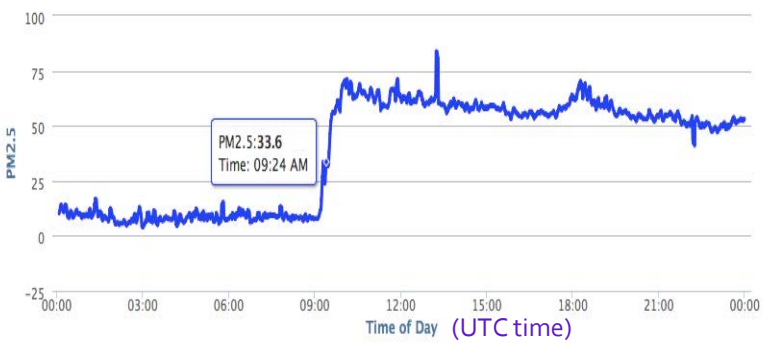


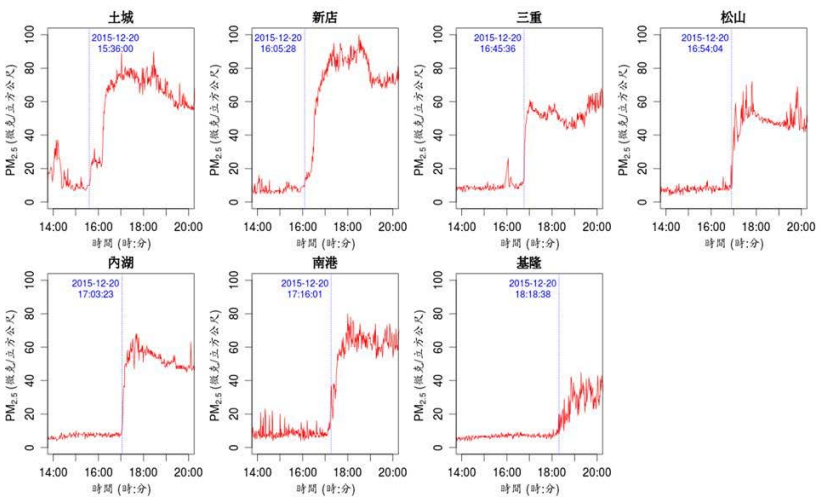
# LASS: Current Status

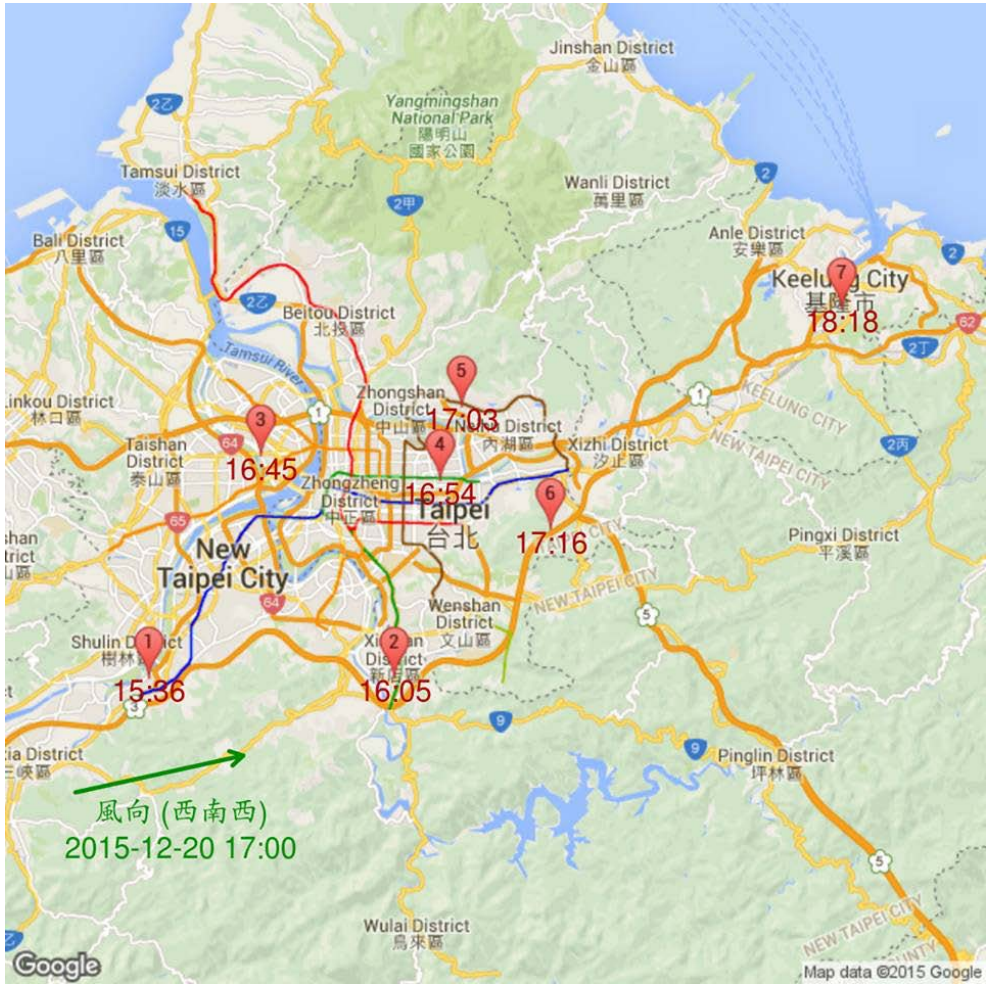



- Total 136 nodes
- Retrieving open data from TW-EPA (36) and KH-Webduino (17)
- Providing JSON API and real-time heatmap
- Expecting to reach 1,000 nodes in Jan 2017

# LASS: Case Study



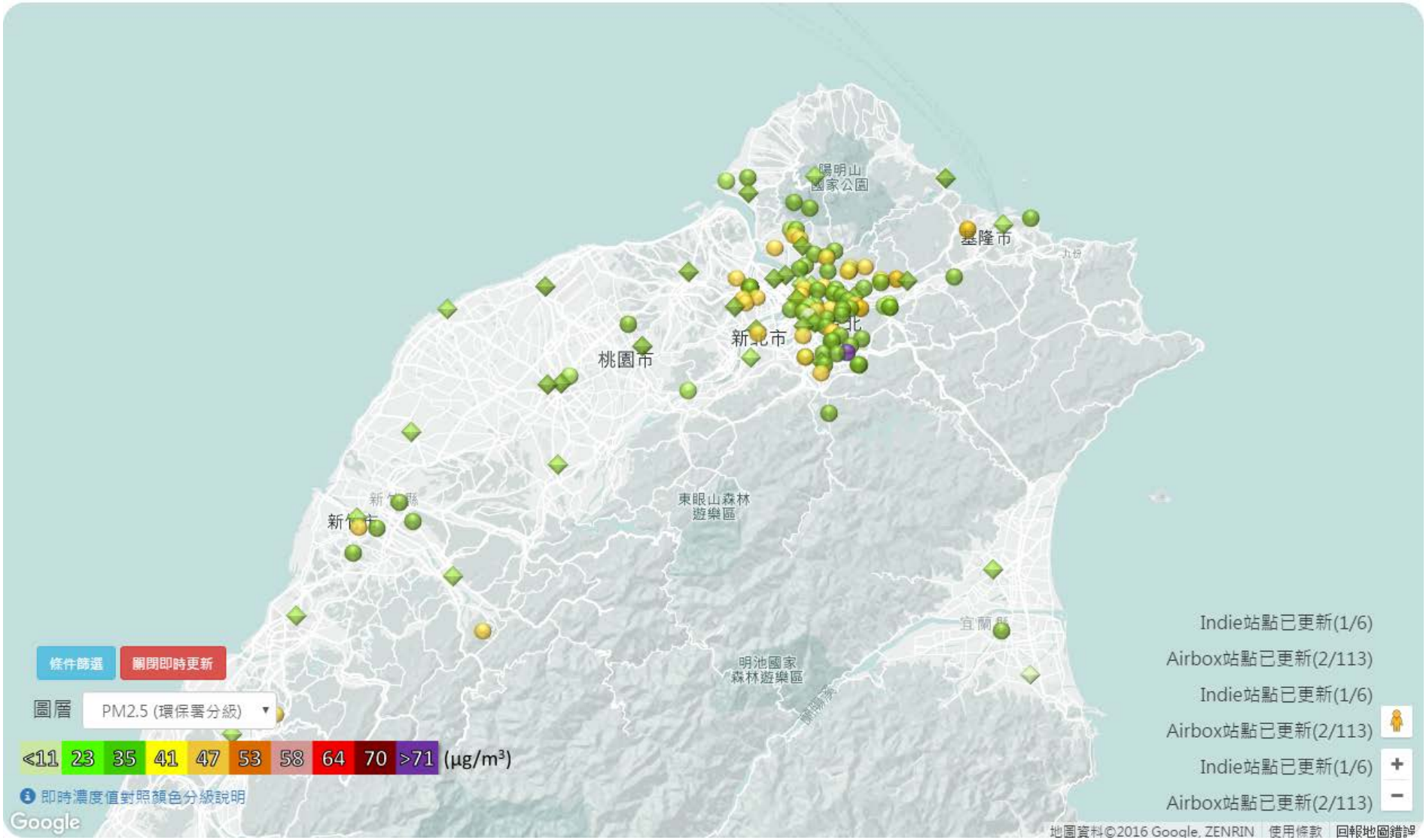



空氣盒子
PM 2.5
啟動記者會
共同主辦單位
TAIPEI
ASUS
REALTEK
台北智慧城市
空氣盒子 PM 2.5
發表會
REALTEK
雖江街、內江街
COMMUNI
中
古亭
DIST
萬大路

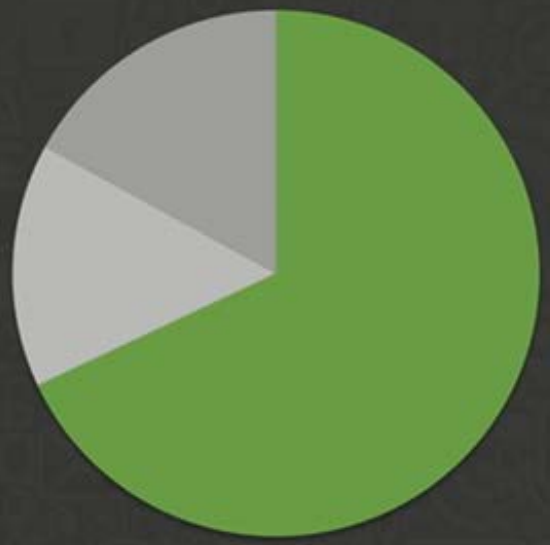
GØV 零時空汙觀測網 自造站點持續招募中!

http://govairmap.3203.info/map.html



# Data variety: Our biggest challenge

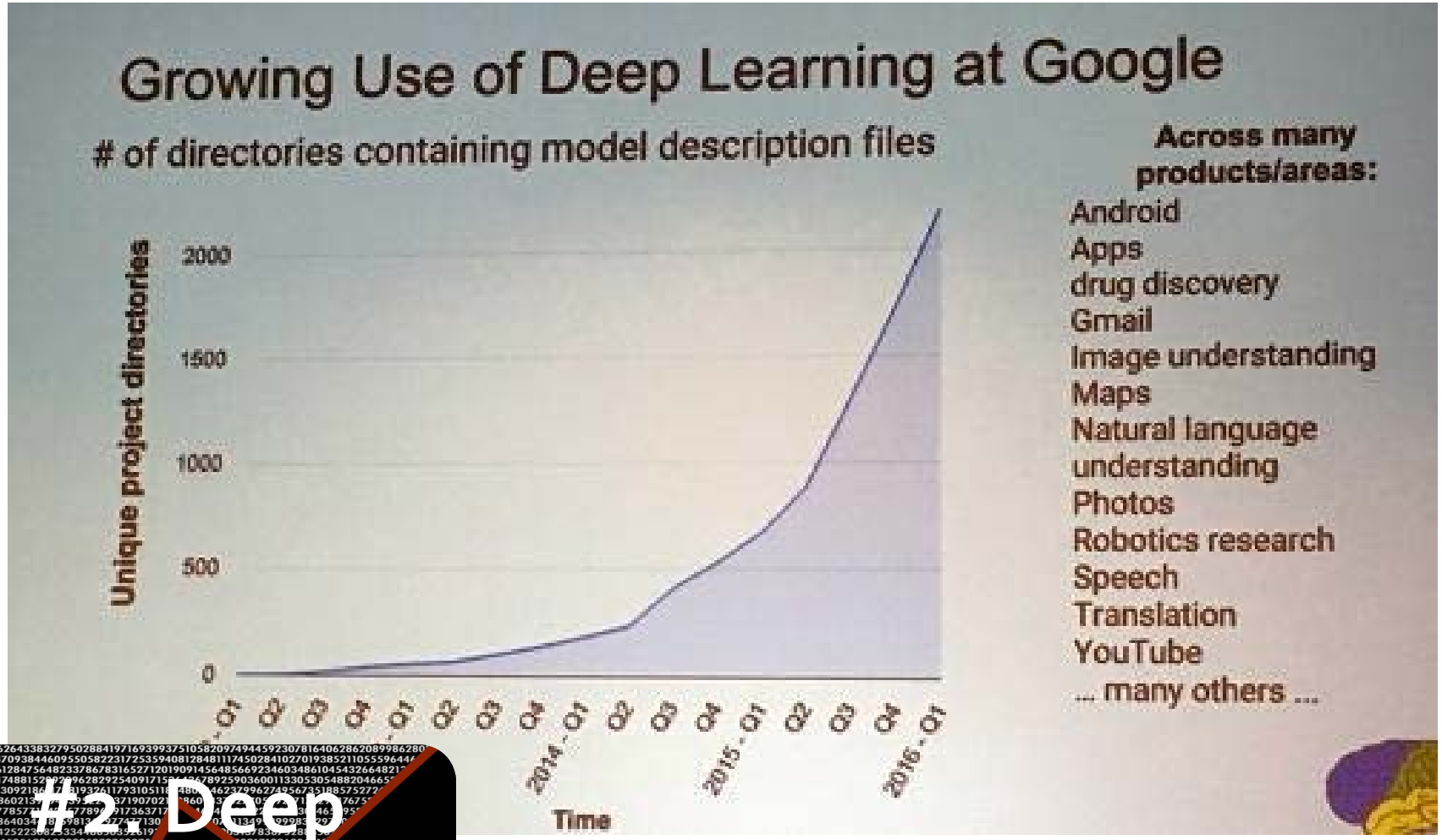
Top Big Data Issues



- Data Variety (68%)
  *Diverse, streaming or new data types*
- Data Volume (15%)
  *Greater than 100TB*
- Other Data (17%)
  *Less than 100TB*

*"Of Gartner's "3Vs" of big data (volume, velocity, variety), the* ***variety of data sources*** *is seen by our clients as both the greatest challenge and* ***the greatest opportunity****."*

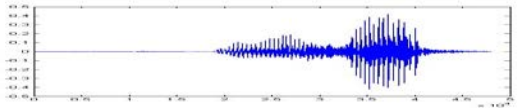
# 3 Major Trends of Data Science (#2)




**#2: Deep Learning**

# Three Steps for Deep Learning

- Speech Recognition

$f^*($  $) =$ "你好"

- Handwritten Recognition

$f^*($ $) =$ "2"

- Playing Go

$f^*($ $) =$ "5-5"
(step)

- Dialogue System

$f^*($ "Hi" $) =$ "Hello"
(what the user said) (system response)

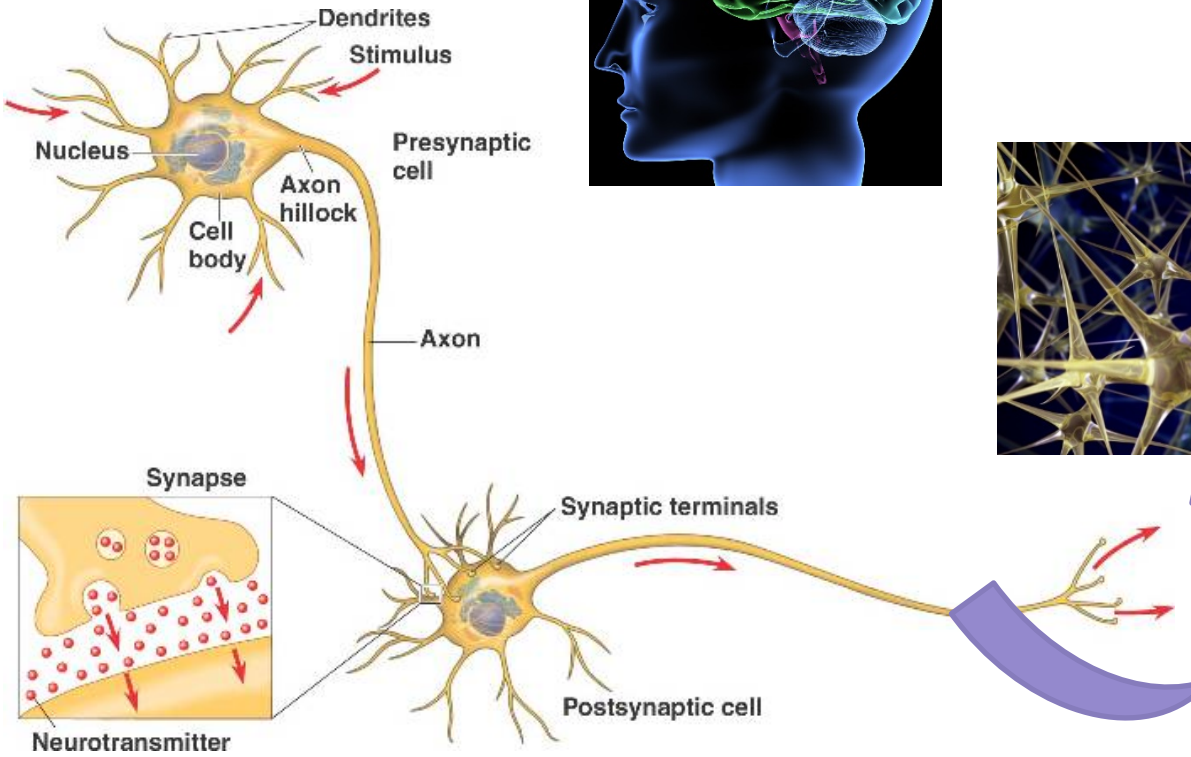
Step 3: Learn!

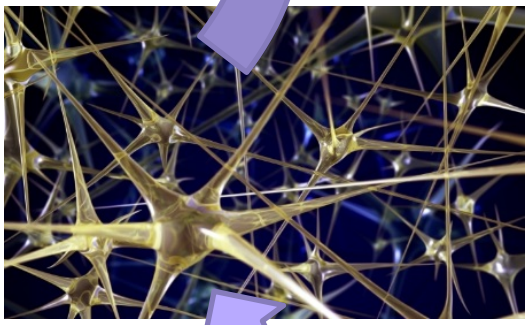
Pick the best function f*



(Slide Credit: Hung-Yi Lee)

# Human Brains



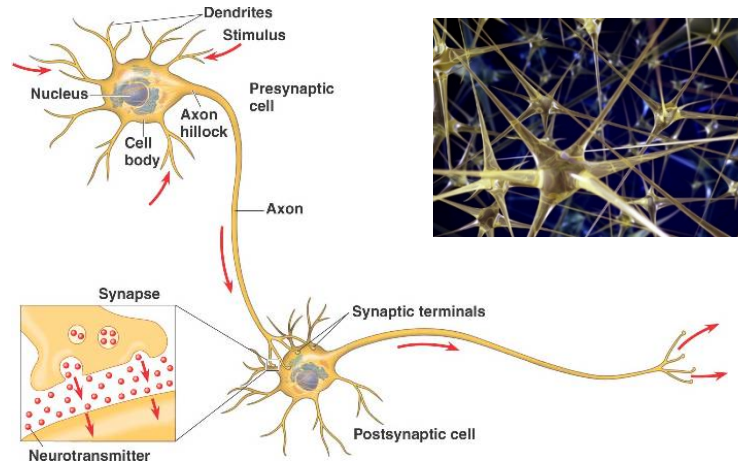



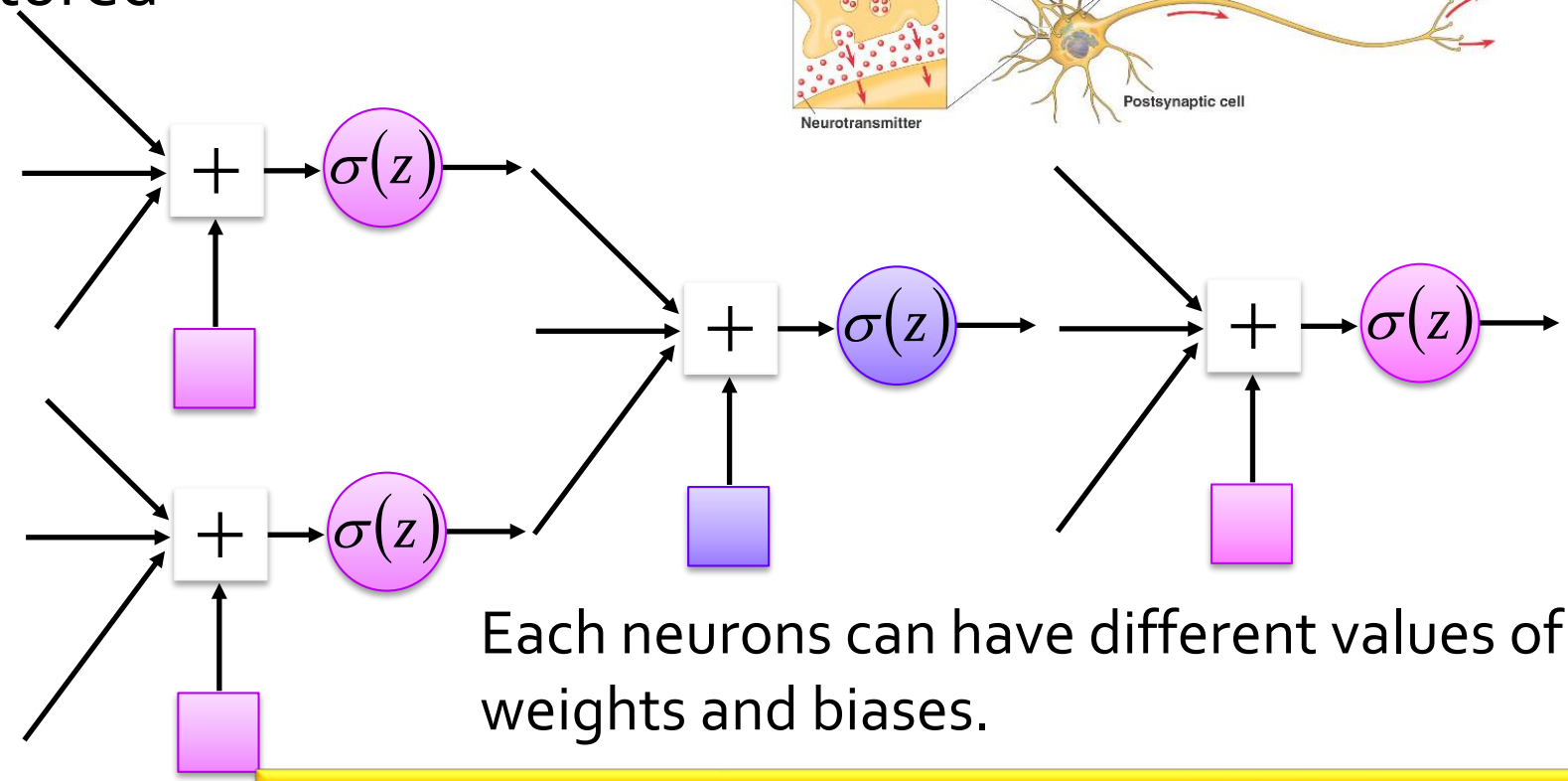


(Slide Credit: Hung-Yi Lee)

# Artificial Neural Network

Different connections leads to different network structured





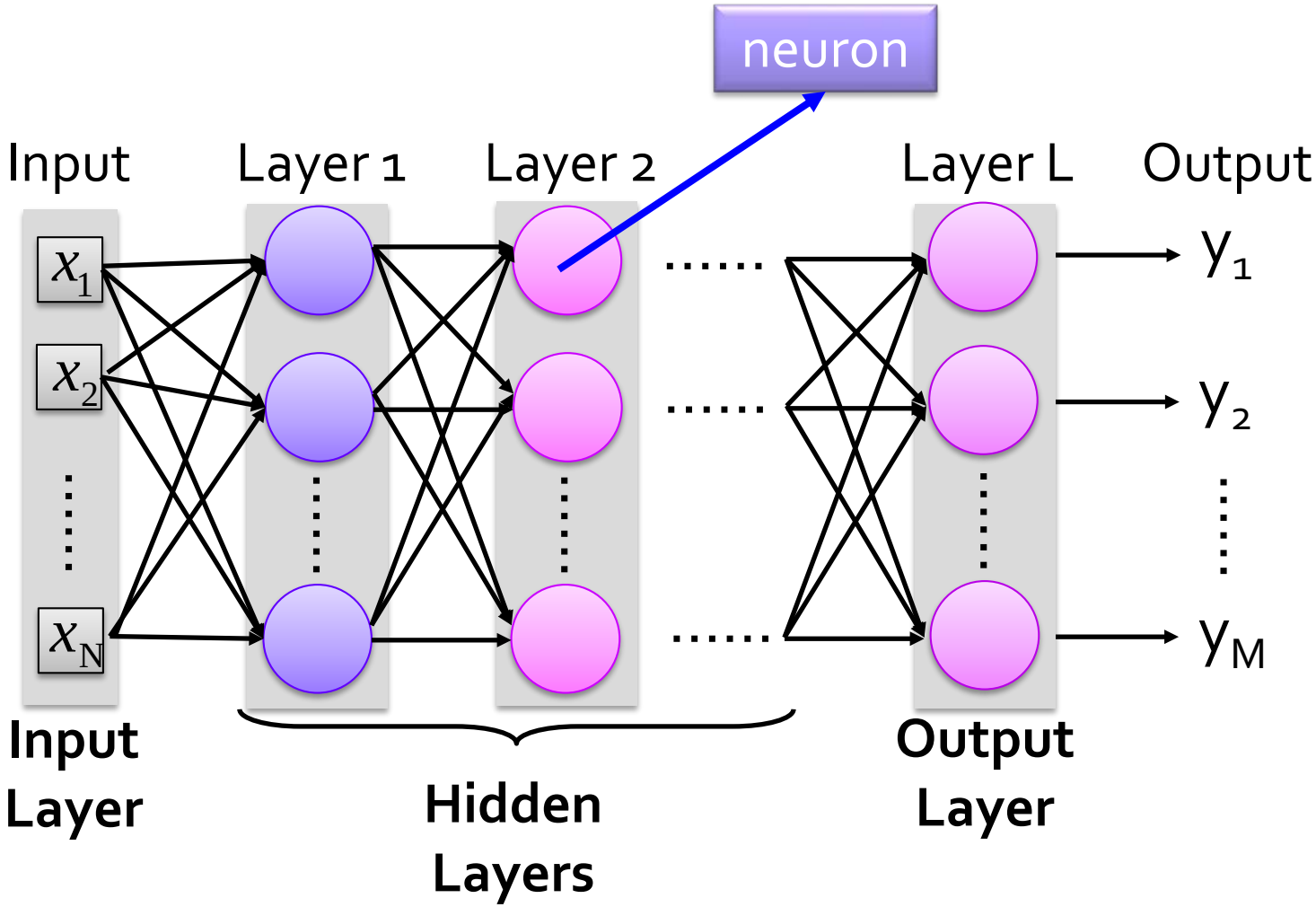

Each neurons can have different values of weights and biases.

Weights and biases are network parameters $\theta$



(Slide Credit: Hung-Yi Lee)

# Fully Connect Feedforward Network




Deep means many hidden layers



(Slide Credit: Hung-Yi Lee)

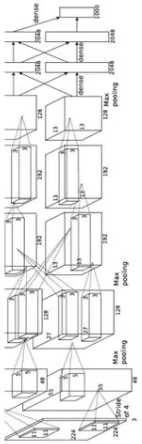
(Slide Credit: Hung-Yi Lee)

# Ultra Deep Networ

http://cs231n.stanford.edu/slides/winter1516_lecture8.pdf

8 layers

16.4%



AlexNet (2012)

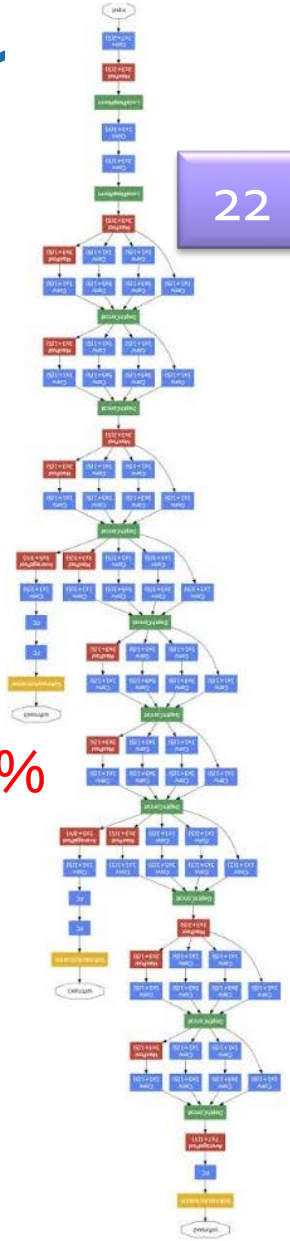
19 layers

7.3%

VGG (2014)

22 layers

6.7%



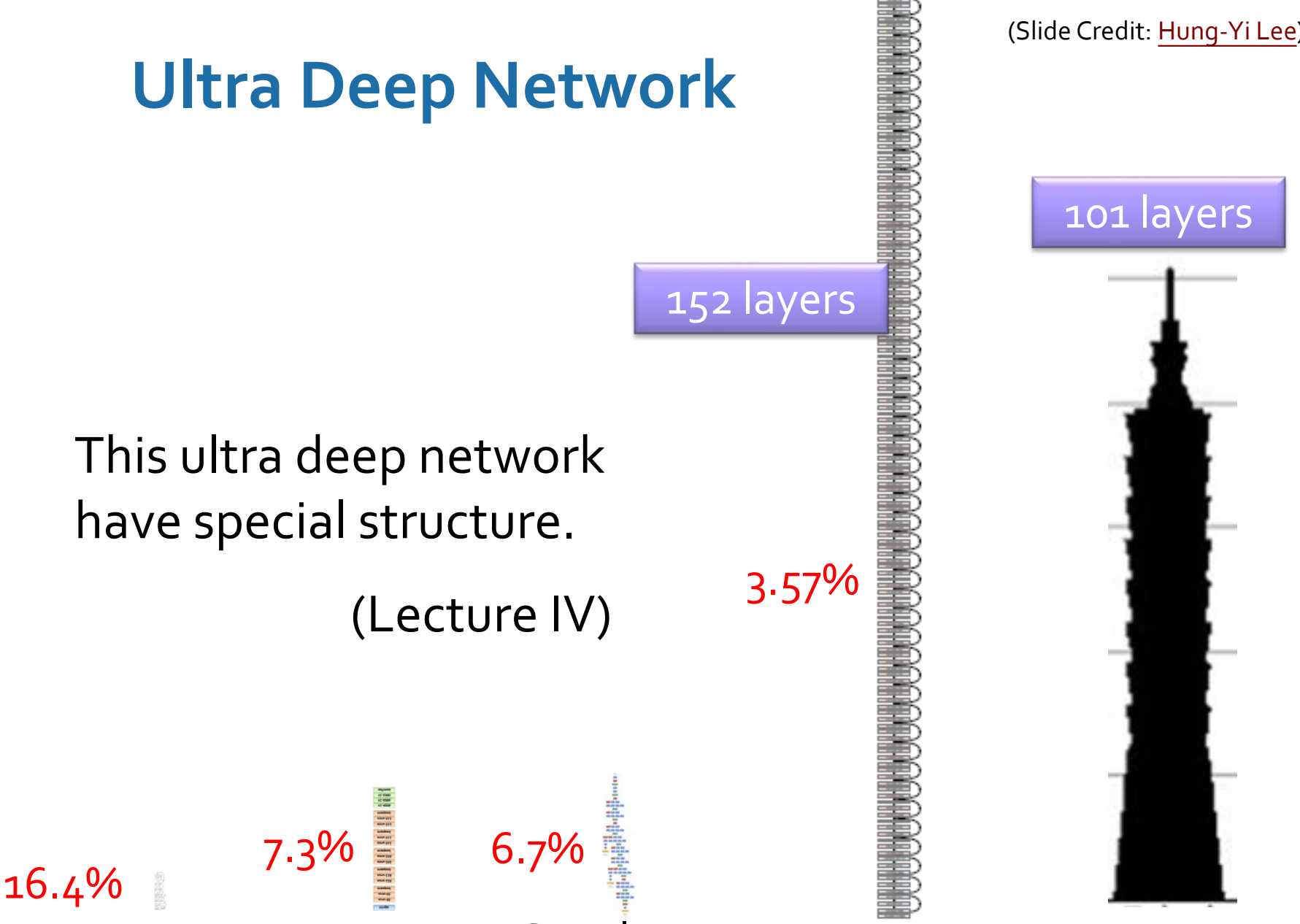
GoogleNet (2014)

(Slide Credit: Hung-Yi Lee)

# Ultra Deep Network

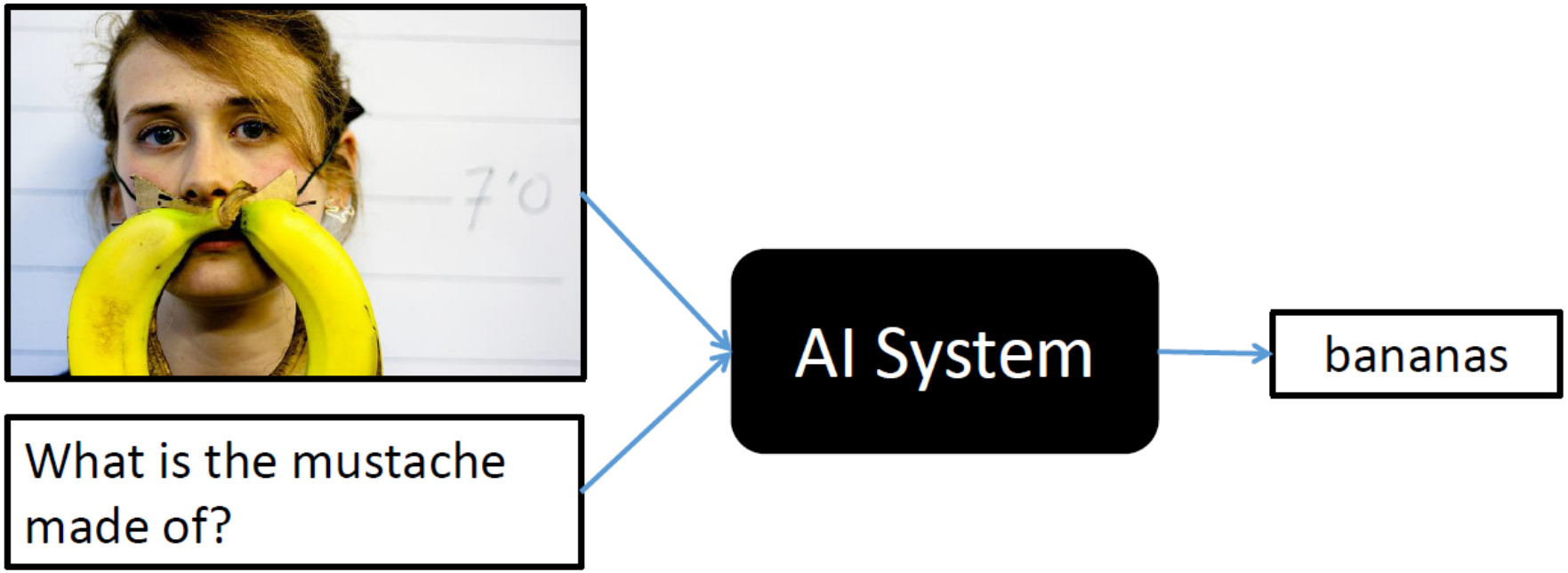
This ultra deep network have special structure.

(Lecture IV)




AlexNet (2012) | VGG (2014) | GoogleNet (2014) | Residual Net (2015) | Taipei 101

# Visual Question Answering

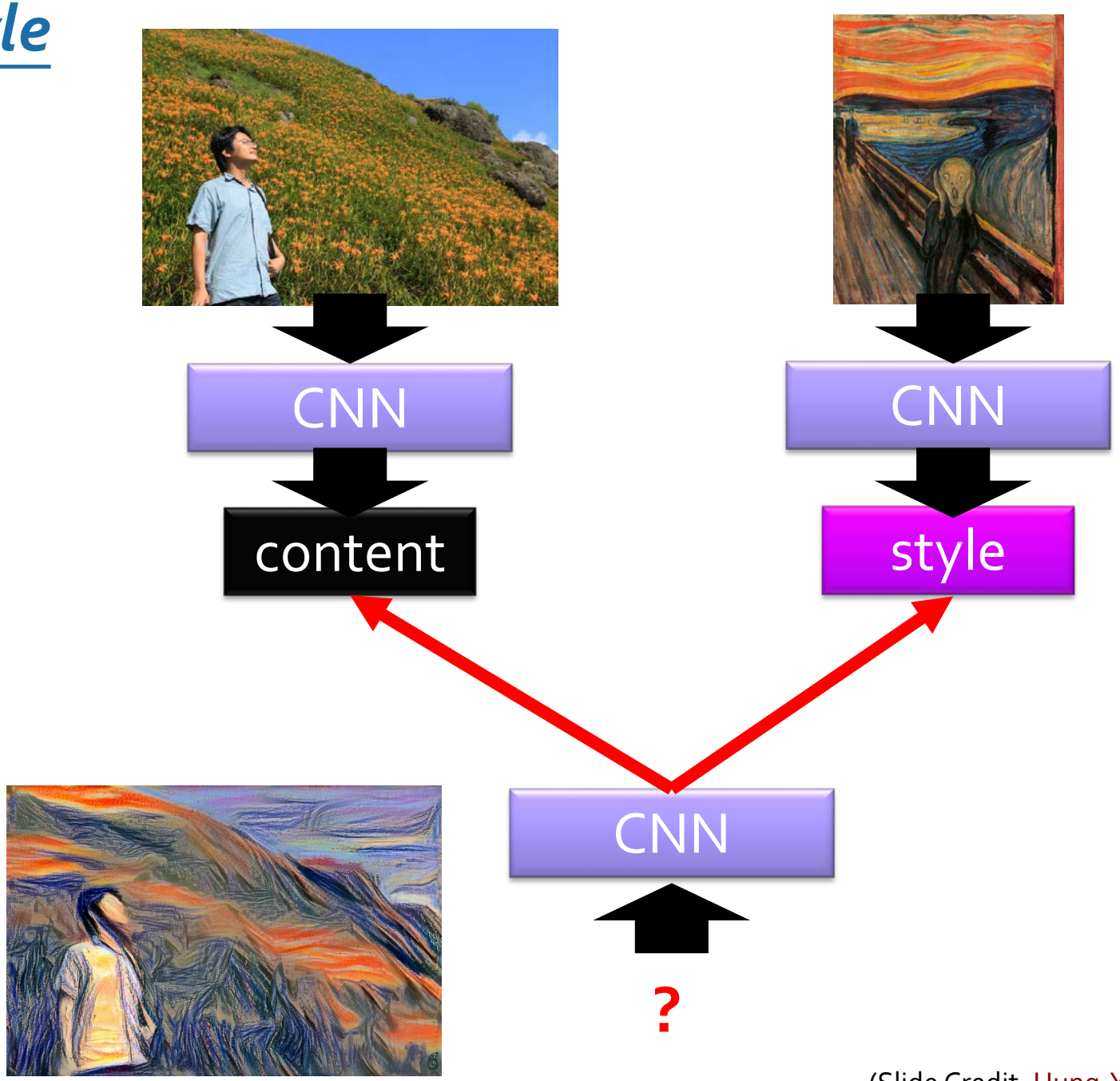



source: http://visualqa.org/



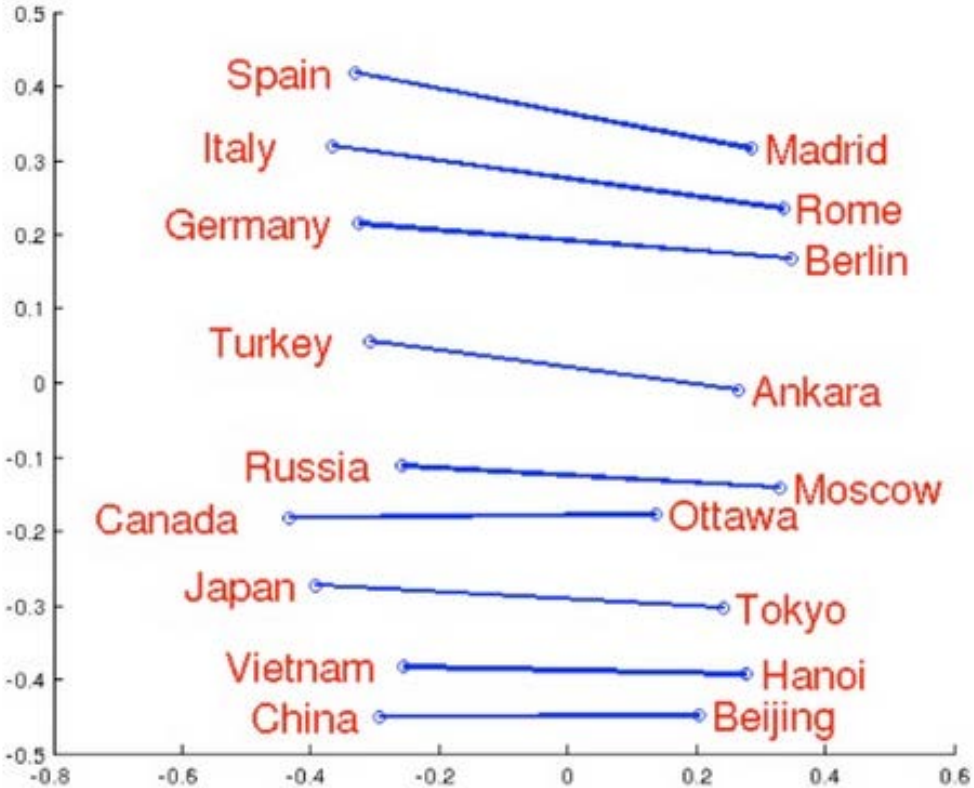
(Slide Credit: Hung-Yi Lee)

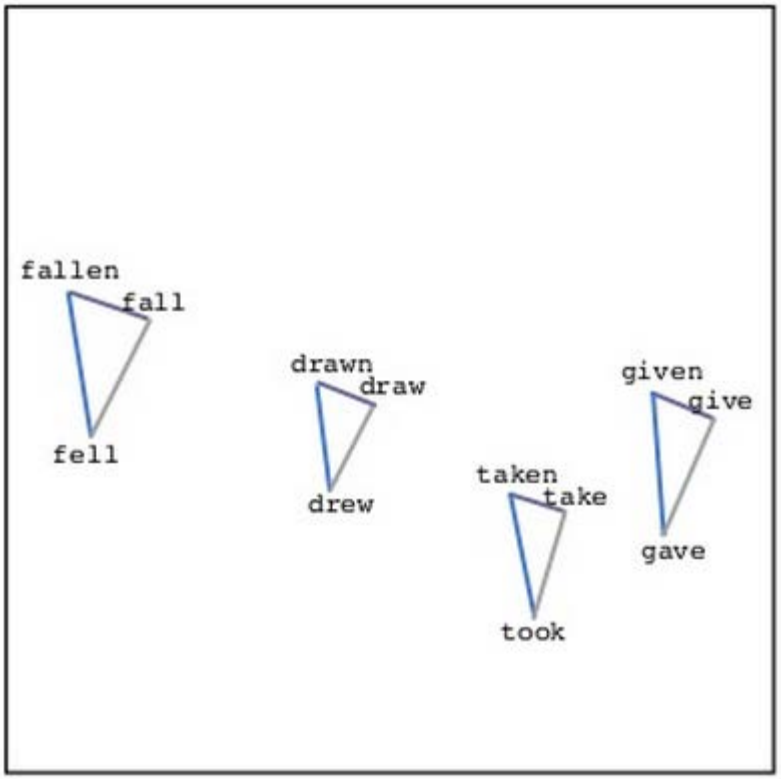
# Deep Style




(Slide Credit: Hung-Yi Lee)

# Word Vector




Source: http://www.slideshare.net/hustwj/cikm-keynotenov2014

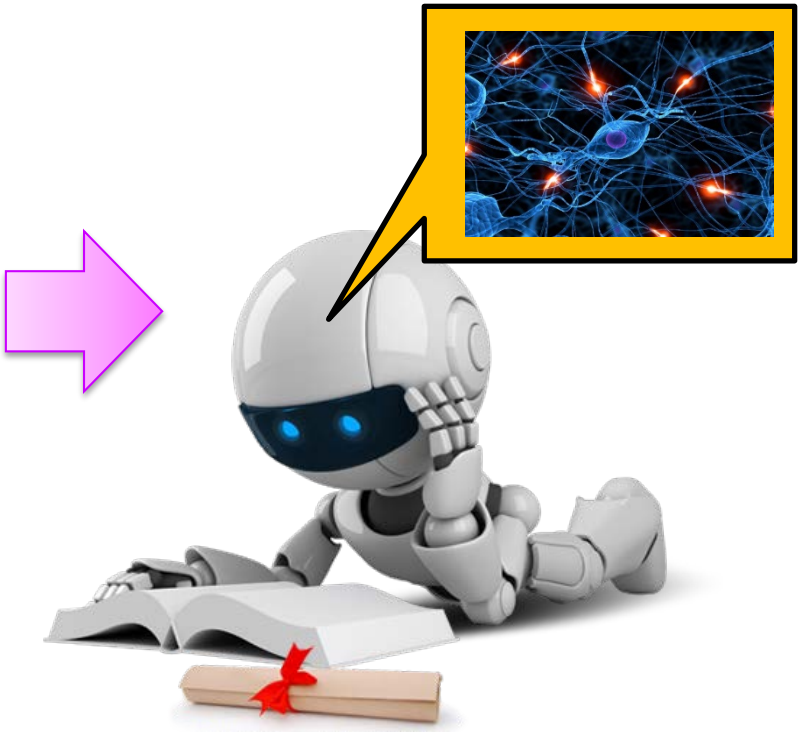


(Slide Credit: Hung-Yi Lee)

# Machine Reading

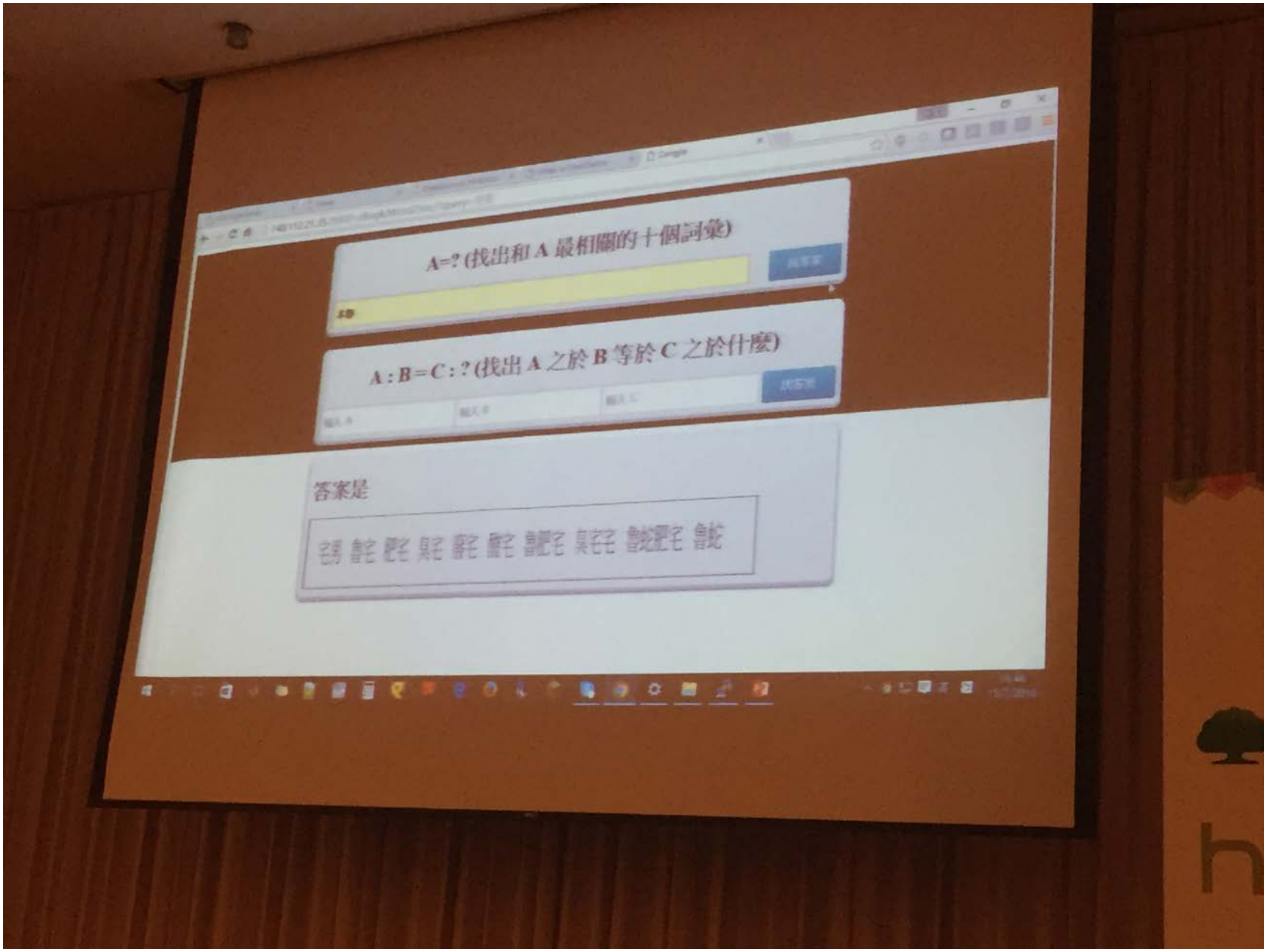
- Machine learn the meaning of words from reading a lot of documents without supervision

Machine learns to understand 鄉民用語 via reading the posts on PTT





(Slide Credit: Hung-Yi Lee)


A=? (找出和 A 最相關的十個詞彙)
A : B=C : ? (找出 A 之於 B 等於 C 之於什麼)
答案是

# 如果你想"深度學習 深度學習"

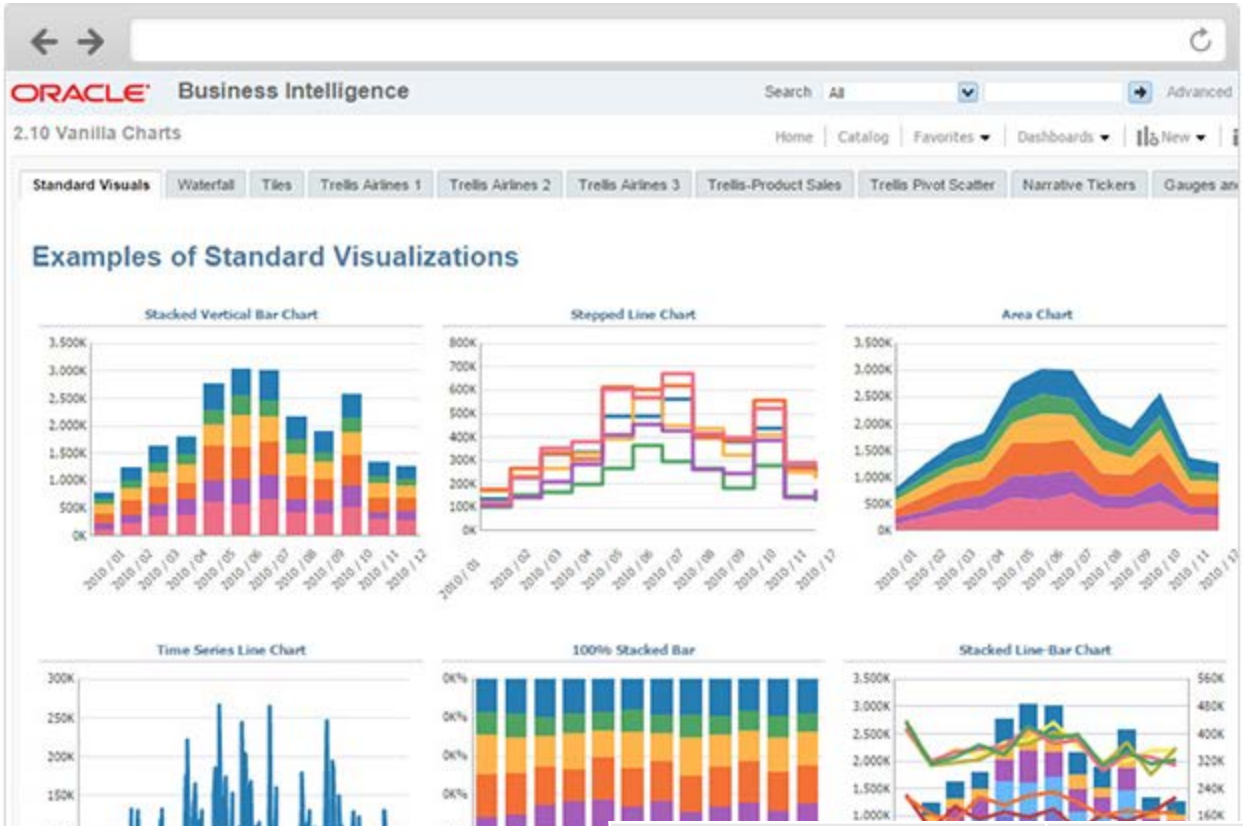
- "Neural Networks and Deep Learning"
  - written by Michael Nielsen
  - http://neuralnetworksanddeeplearning.com/
- "Deep Learning"
  - Written by Yoshua Bengio, Ian J. Goodfellow and Aaron Courville
  - http://www.iro.umontreal.ca/~bengioy/dlbook/
- Course: Machine learning and having it deep and structured
  - http://speech.ee.ntu.edu.tw/~tlkagk/courses_MLSD15_2.html



(Slide Credit: Hung-Yi Lee)

# 3 Major Trends of Data Science (#3)

- Nev
- Adv ng
  tech




iness is
of

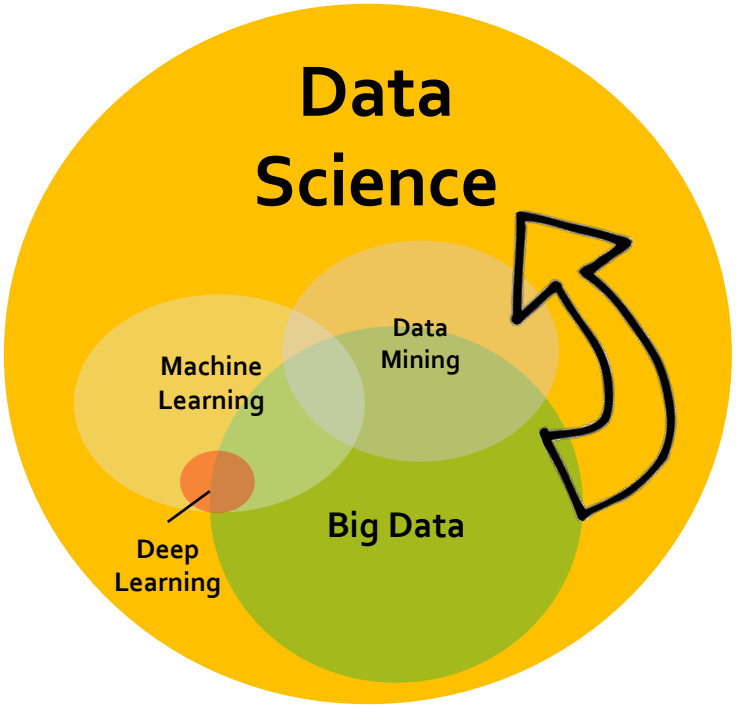
Q: 如何

| | |
|---|---|
| 提升產品品質？ | 加強包裝？ |
| 加強行銷？ | 降低生產成本？ |
| 提升研發效率？ | 提升行政效率？ |
| 提升回頭率？ | 技術水準？ |

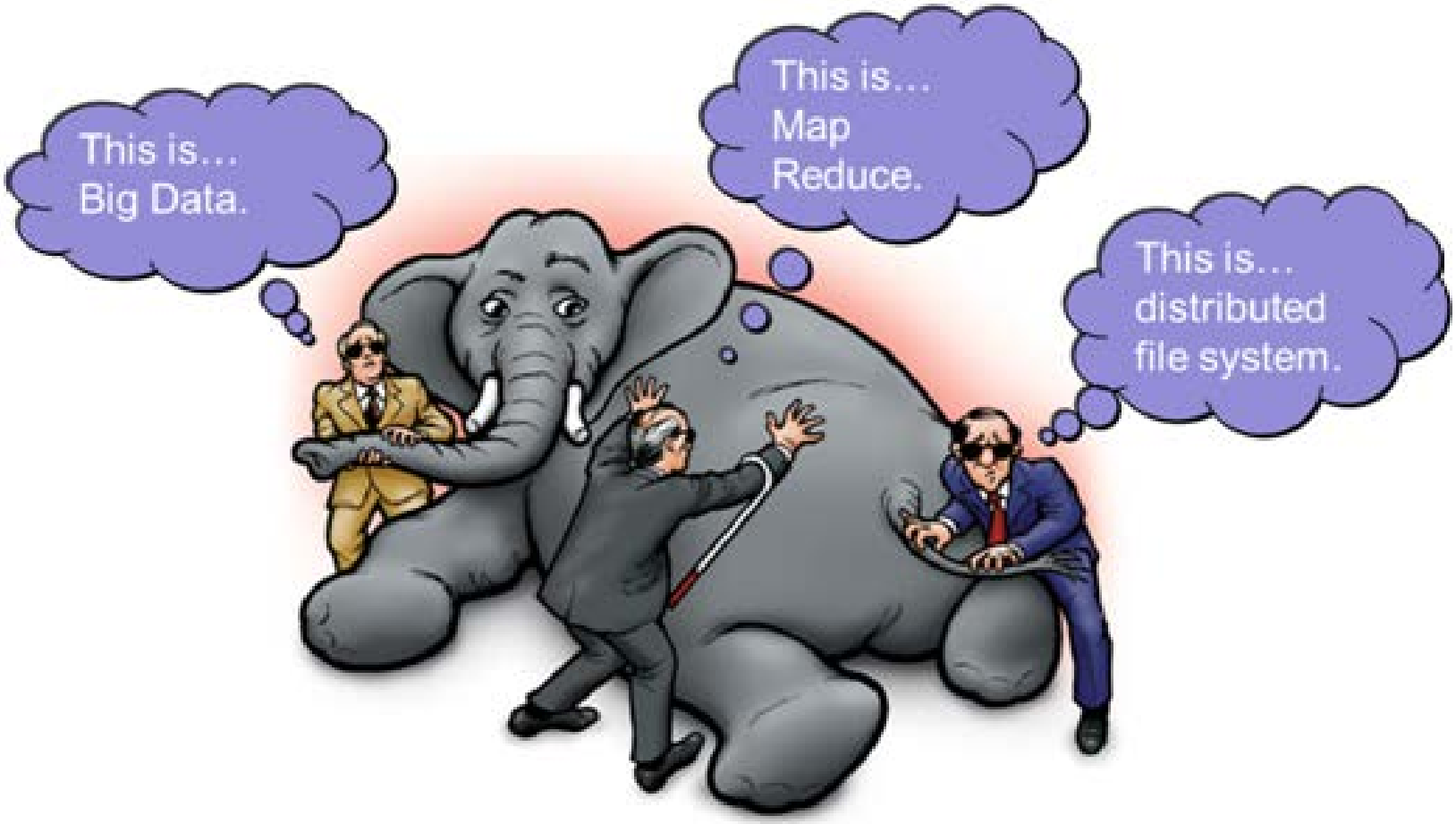
## #3. Data Discovery

# Data Science vs. Big Data

- Data Science is a superset of Big Data.
- However, the rise of Big Data draws people's attention to Data Science.



# Data Science from Individuals' Views



## Goal Definition

## Design & Collection

## Data Preparation

## EDA

## Variables? Methods?

## Evaluation, Validation & Model Selection

## Model Use & Reporting

(Slide Credit: Galit Shumueli)

# Data Science is More Than …

- Big data (small data also do)
- Statistics / machine learning
- Data analysis languages (e.g., R, Python)
- Data infrastructure (e.g., NoSQL, Hadoop, Spark)
- Data visualization

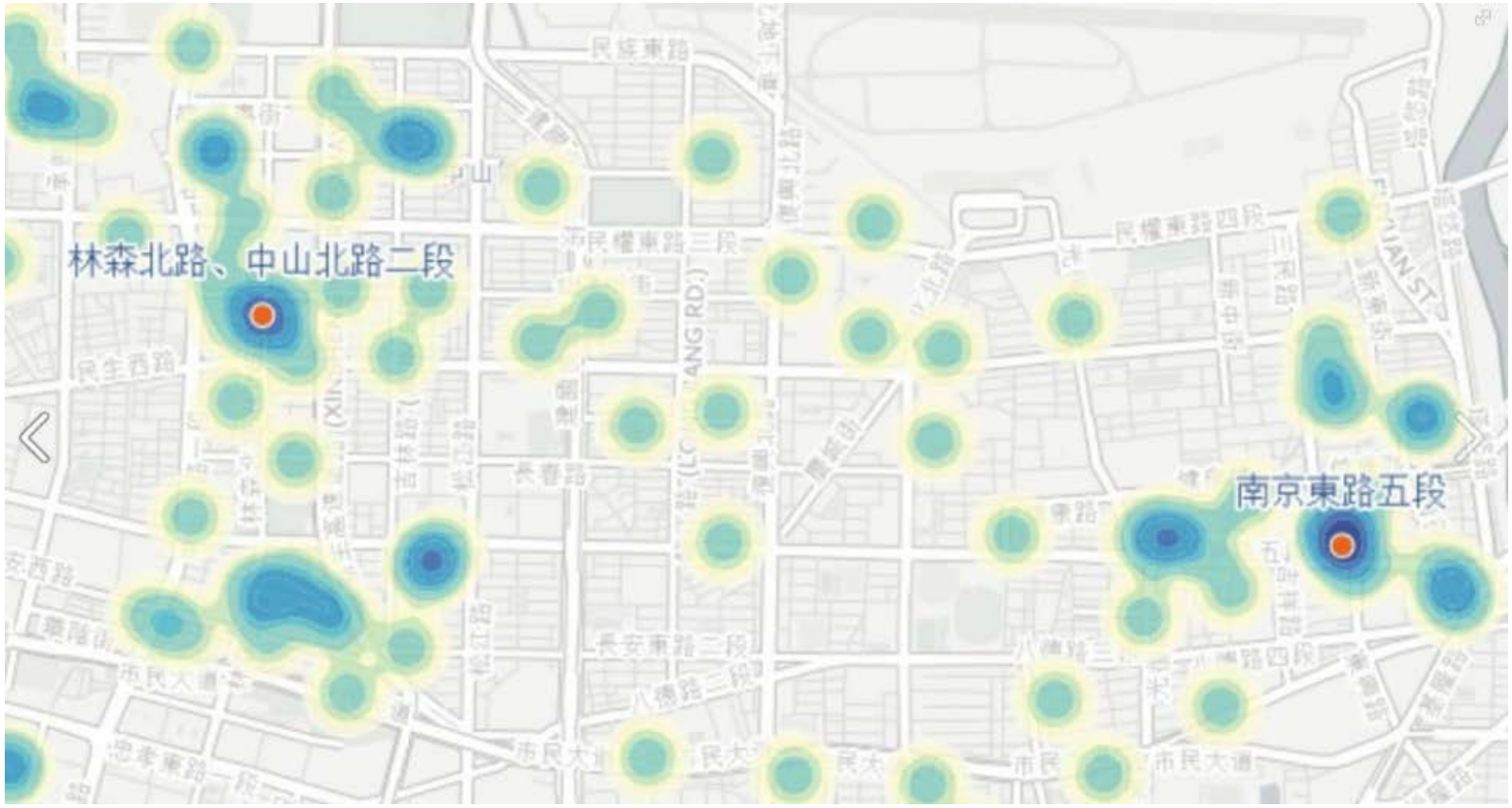
# Data Visualization vs. Data Analysis

Visualization is the act or process of interpreting in visual terms or of putting into visible form.

Analysis indicates a careful study of something to learn about its parts, what they do, and how they are related to each other.

—Merriam-Webster's Dictionary

# 維護市民安全！ 北市開放犯罪地圖

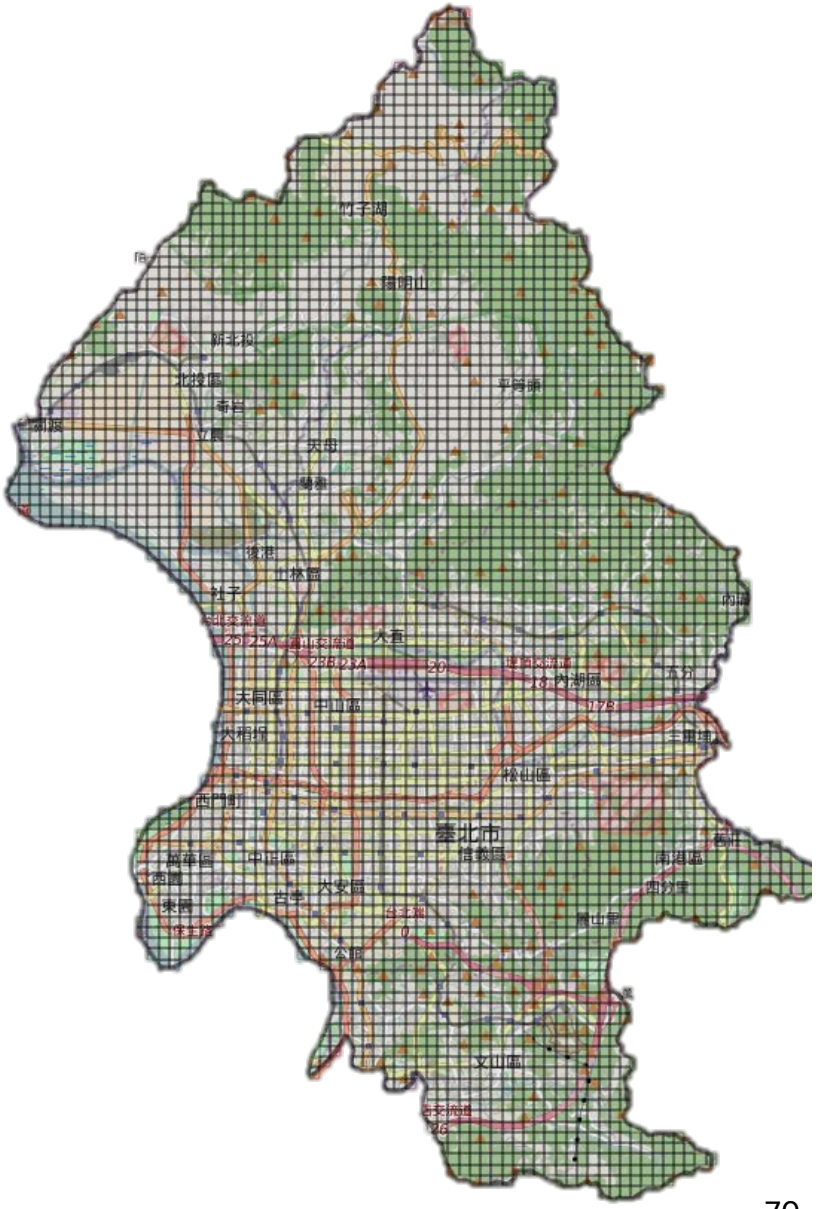


# 定義空間尺度與問題

- 以長寬各 500 公尺的方格為單位
  - 共 1081 個方格
  - 剔除無建築物之方格後 ➔ 823 個

- 犯罪機率
  = 累計犯罪紀錄/建築物數量
  - 共 396 個方格內有犯罪紀錄

- 藉由每方格內的基礎建設和鄉里資料，定義方格特性



# Dataset

- **每個里的**
  - 面積
  - 人口
  - 性別比例
  - 婚姻狀態
  - 教育程度
  - 平均收入
  - 平均年齡和老化指數
  - 平均戶數和每戶平均人數
- **地點資料 (地址)**
  - 公園
  - 公廁
  - 銀行
  - 學校
  - 公車站
  - 捷運站
  - 警察局
  - 便利商店
  - 戶外停車場
  - 路邊停車格
  - 路邊監視器
  - 建築物高度

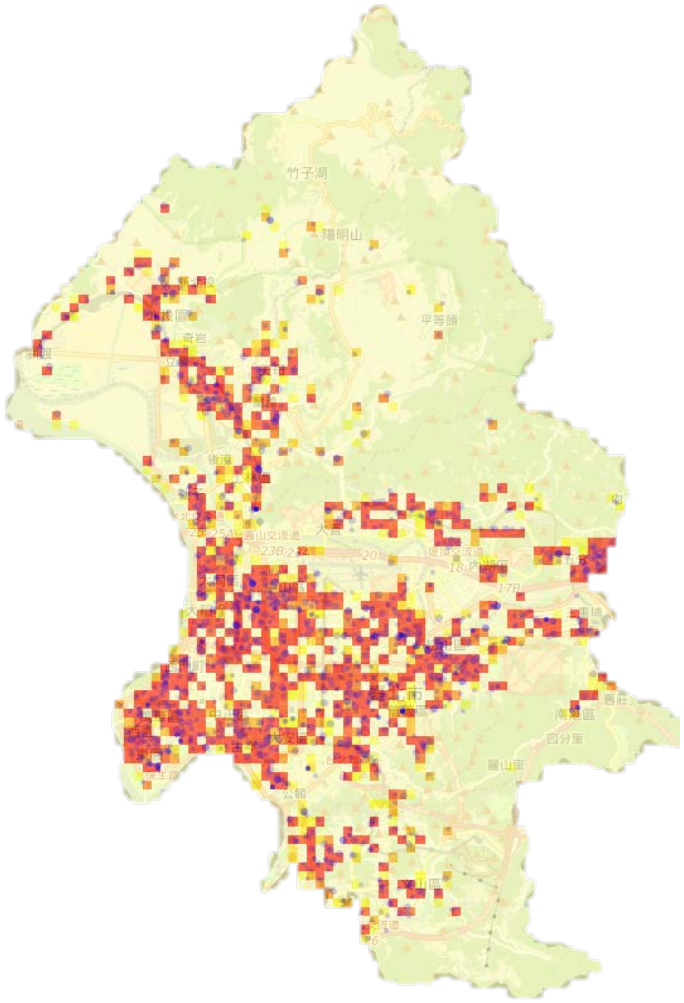
# 與犯罪嚴重程度相關的變數

- **正相關**
  - 性別比 (男:女) ***
  - 離婚人口比例 ***
  - 離最近的學校距離 ***
  - 離最近的銀行距離 ***
  - 離最近的警察局距離 ***
  - 離最近的捷運站距離 ***
- **負相關**
  - 路燈數量 ***
  - 監視器數量 ***
  - 公車站牌數量 ***
  - 每家戶平均人數 ***
  - 教育程度 (就讀大學比例) ***

Significance code: * = 0.05; ** = 0.01; *** = 0.001

# 犯罪嚴重程度分類

- 將有犯罪紀錄的方格區分為
  - 高危險
  - 低危險
- 以 SVM Recursive Feature Elimination 挑選出 17 個 features
- 以 SVM classification 作為預測模型
  - Accuracy: 0.722
  - AUC: 0.730



# Data Science in Taiwan

# Back to 2012...

# Data Scientist:

## *The Sexiest Job of the 21st Century*

***Meet the people who can coax treasure out of messy, unstructured data.***

*by Thomas H. Davenport and D.J. Patil*

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

# MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

## MATH & STATISTICS

- Machine learning
- Statistical modeli
- Experiment desig
- Bayesian inferenc
- Supervised learni random forests, l
- Unsupervised lea dimensionality re
- Optimization: gra variants

## PROGRAMMING & DATABASE

ndamentals
.g. Python
packages, e.g., R
NoSQL
nd parallel query
s
g
like AWS

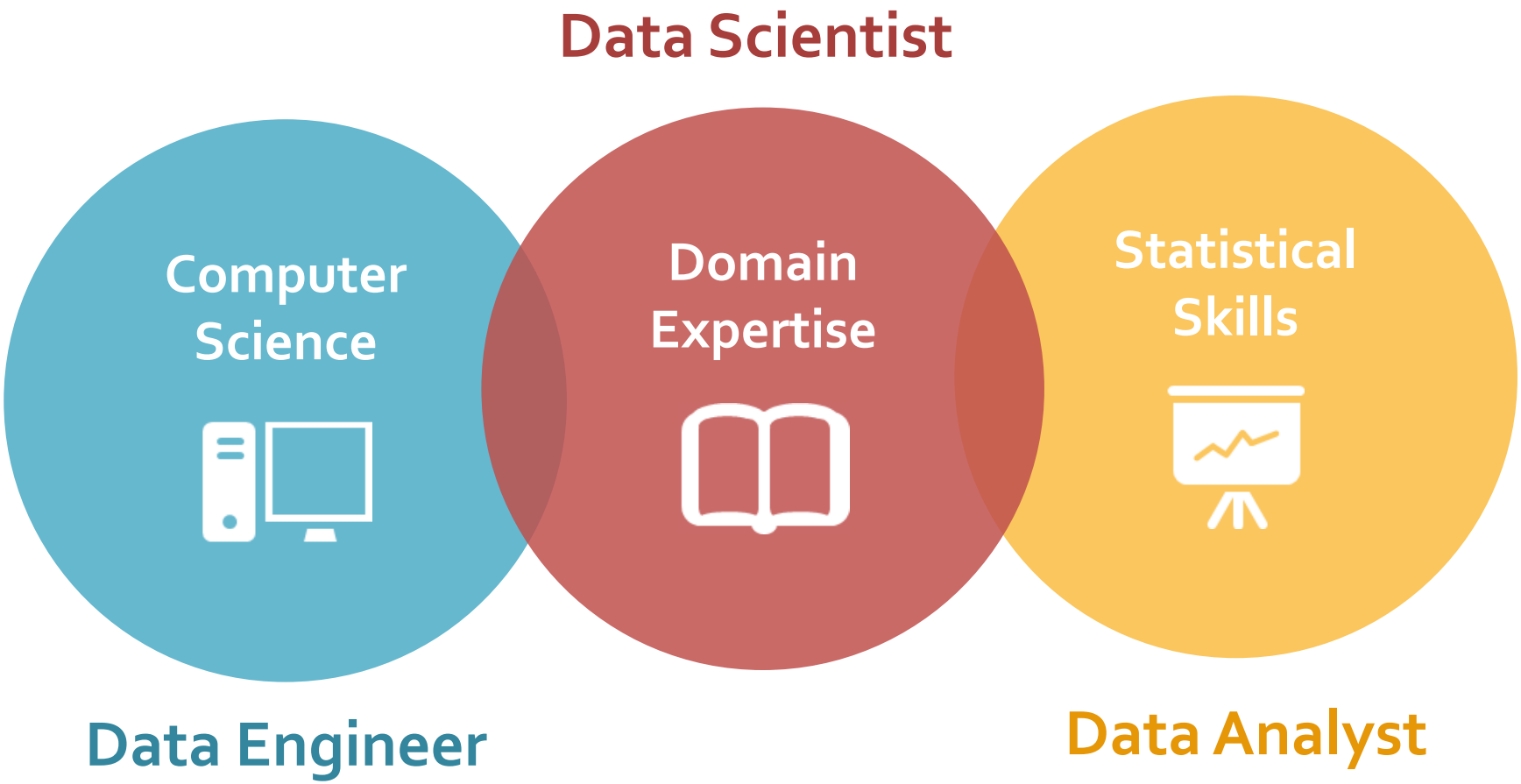
## DOMAIN KN & SOFT SKIL

- Passionate about
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

## TION ION

senior management

- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau


Data Scientist
Computer Science
Domain Expertise
Statistical Skills
Data Engineer
Data Analyst

Data Librarian
Data Journalist
Data Analyst
Data Engineer
Data Steward
Data Archivist
Data Librarian
Data Journalist
Data Analyst
Data Engineer
Data Steward
Data Archivist

# 2 days x 800 ppl

# 2014 台灣資料科學愛好者年會

8/30 – 8/31, 2014

http://twconf.data-sci.org/

https://www.facebook.com/twdsconf

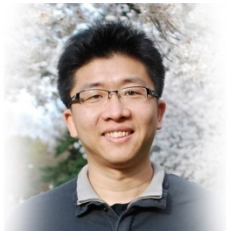
# 演講議程講師群

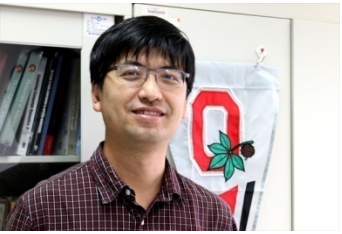
- 林智仁 (Chih-Jen Lin), 國立臺灣大學資訊工程學系特聘教授
- 高嘉良 (Chia-liang Kao), g0v.tw 台灣零時政府共同創辦人
- 劉嘉凱 (Chia-Kai Liu), 御言堂總經理
- 陳君厚 (Chun-Houh Chen), 中央研究院統計科學研究所研究員兼副所長
- 趙國仁 (Craig Chao), Vpon 行動數據科技數據科學家
- 潘美連 (Mei-Lien Pan), 台灣醫學資訊學會祕書長
- 劉家宏 (Chia-Hung Liu), 華聯生物科技股份有限公司 研發部副理
- 林大利 (Da-Li Lin), 特有生物研究保育中心助理研究員
- 郭建甫 (Jeff Kuo), Gogolook 走著瞧公司創辦人兼執行長
- 高義銘 (Yimin Kao), Gogolook 走著瞧公司資料科學家
- 彭啟明 (Chi Ming Peng), 天氣風險管理開發公司總經理
- 吳牧恩 (Mu-En Wu), 東吳大學數學系助理教授
- 呂俊宏 (Enrico Lu), 資訊工業策進會創新應用服務研究所研究顧問
- 洪進吉 (Gene Hong), 台灣數位文化協會顧問
- 黃孝文 (Norman), Yahoo! Taiwan Senior Data Engineer
- 林于聖 (Jason Lin), Yahoo! Taiwan Senior Data Engineer
- 余孟勳 (MengHsun Simon Yu), 台灣公益責信協會發起人兼理事長

# 4 days x 1300 ppl

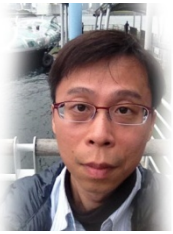
# 2015 台灣資料科學愛好者年會

8/20 – 8/23, 2015





# 資料科學專家 x 24



SOGI 逗機王

Hadoop/Spark
資料科學快速體驗營

R 語言
資料分析上手課程

SheetHub

地圖資料視覺化課程

開放文化基金會
Open Culture Foundation

資料新聞實戰營
dlBootcamp Taipei

dsp 智庫驅動

DSP 資料開竅・企業論壇

GØV

g0v 零時政府黑客松

系列活動

# 台灣資料科學愛好者年會

# 玩轉交易 (Be a Trade"R")！

日期: 2015/ 11/14 (六) 地點: 中央研究院人文社會科學館

| | |
|---|---|
| 9:00 - 10:30 | 馬丁格爾的聖盃！ |
| 10:30 - 10:50 | 茶點時間 |
| 10:50 - 12:20 | 凱利賭徒: 模擬與分析 |
| 12:20 - 13:10 | 午餐 |
| 13:10 - 14:40 | 輕鬆把玩金融資料：Quantmod Basic |
| 14:40 - 15:00 | 茶點時間 |
| 15:00 - 16:30 | 金融資料探勘 |
| 16:30 - 17:00 | 互動論壇 |

# 機 器 學 習 初 探

日期: 2015/ 12/12 (六) 地點: 中央研究院人文社會科學館

| | |
|---|---|
| 9:00 - 10:30 | 從資料中學習<br>(Learning from Data) |
| 10:30 - 10:50 | 茶點時間 |
| 10:50 - 12:20 | 基礎機器學習模型<br>(Fundamental Machine Learning Models) |
| 12:20 - 13:10 | 午餐 |
| 13:10 - 14:40 | 惱人的過度調適<br>(Hazard of Overfitting) |
| 14:40 - 15:00 | 茶點時間 |
| 15:00 - 16:30 | 現代機器學習模型<br>(Modern Machine Learning Models) |

# 親手打造 Google 級深度學習模型

日期: 2015/12/29 (二)　地點: 中央研究院人文社會科學館

## 邱中鎮 博士

**Google Brain 軟體工程師**

邱中鎮博士在 2014 年於美國南加大取得博士學位，畢業後即加入 Google Brain 小組，目前專注於打造深度學習系統與設計深度學習模型來解決 Google 等級的難題。

| | |
|---|---|
| 9:00 - 10:30 | **深度學習入門**<br>(Introduction to Deep Learning) |
| 10:30 - 10:50 | 茶點時間 |
| 10:50 - 12:20 | **親手打造 Google 級深度學習模型**<br>(Build Your Own Google-Class Deep Learning Machine) |

# 電腦視覺一二三

日期: 2016/1/12 (二)　地點: 中央研究院人文社會科學館

## 黃嘉斌

伊利諾大學香檳分校電機與電腦工程博士候選人

黃嘉斌為伊利諾大學香檳分校電機與電腦工程博士候選人, 研究主軸為電腦視覺以及計算攝影學。他的個人網頁為 http://www.jiabinhuang.com/

| | |
|---|---|
| 9:00 - 10:30 | 淺談電腦視覺<br>(Introduction to Computer Vision) |
| 10:30 - 10:50 | 茶點時間 |
| 10:50 - 12:20 | 基礎與應用<br>(Fundamentals and Applications) |

# 資料科學的第一堂課：心法、案例分析與團隊建立

日期: 2016/1/30 (六)　地點: 中央研究院人文社會科學館

| | |
|---|---|
| 9:00 - 10:30 | 資料科學簡介 |
| 10:30 - 10:50 | 茶點時間 |
| 10:50 - 12:20 | 資料分析實戰案例分享 |
| 12:20 - 13:10 | 午餐 |
| 13:10 - 14:40 | 資料科學家的養成之路 |
| 14:40 - 15:00 | 茶點時間 |
| 15:00 - 16:00 | 企業文化以及資料科學團隊的建立 |
| 16:00 - 17:00 | 互動論壇<br>與談人<br>陳昇瑋 / 年會總召, 中央研究院資訊科學研究所研究員<br>彭啟明 / 天氣風險管理開發公司總經理<br>邱銘彰 / 台灣威瑞特公司技術長 |

# http://datasci.tw/event/

## 系列活動

### 一天搞懂深度學習 (2016/5/21)

這門課程將由台大電機系李宏毅教授利用短短的六個小時簡介深度學習技術。在本課程中，李教授會先帶大家熟悉深度學習的基本概念與工具，接著剖析深度學習和其它機器學習方法的潛在優勢，以及分享實作技巧及實戰經驗，希望幫助大家不只能了解深度學習，也可以有效率地上手深度學習，用在手邊的問題上。無論是從未嘗試過深度學習的新手，還是已經有一點經驗想更深入學習，都可以在這門課中有所收穫。 活動介紹

### 實戰 D3.js — 最熱門的互動資料視覺化技術 (2016/4/30)

D3.js 立基於網頁技術、並從設計的核心納入互動、即時等概念，可說是當代最先進的一組視覺化工具。本課程很榮幸邀請 infographics.tw 與 g0v.tw 共同發起人吳泰輝來進行 D3.js 的教學。在這堂課中，你可以學到：D3.js 資料模型的完整概念、製作動畫與變形圖表、使用 D3.js 重新解譯資料、製作熱度地圖、Circular Treemap 等進階圖表。 活動介紹

### R 語言翻轉教室 (2016/4/23, 2016/4/30)

本課程很榮幸邀請到 Taiwan R User Group 的共同創辦人吳齊軒擔任講者，利用一套用 R 學 R 的教案 -- R 語言翻轉教室，來引領無程式開發經驗的學員，從實作中瞭解 R 語言的基本操作以及初階資料分析。這樣的訓練將幫助學員學會如何掌握資料的脈絡，進而輔助決策，目標是改善同學們學R語言基礎知識的體驗。 活動介紹

### 視訊訊號處理與深度學習應用 (2016/4/23)

本課程很榮幸邀請到陳維超博士擔任講者，課程內容集合了電腦視覺的概念簡介、多媒體資料的分類與檢索方式、深度學習於視訊資料的應用與挑戰、與加速這些技術的 GPU 平行算運算平台等，並將介紹實際的 IoT 應用例子，與實作所需之程式工具等，為理論與實務並重的安排。 活動介紹

### 資料視覺化之理論、賞析與實作 (2016/3/26)

這個課程濃縮了康教授在台大所開設的『資料視覺化』課程的精華，總共包含賞析、理論、設計與實作四節課程。 活動介紹

# 4 days x 1718 ppl

2016

台灣資料科學
愛好者年會

資料科學的第一堂課：
心法、案例分析與團隊建立

從資料到知識：
從零開始的資料探勘

一天搞懂深度學習

手把手教你
R 語言資料分析實務

R 語言資料工程及探勘實務

2016
台灣資料科學
愛好者年會

# 資料科學專家 x 50

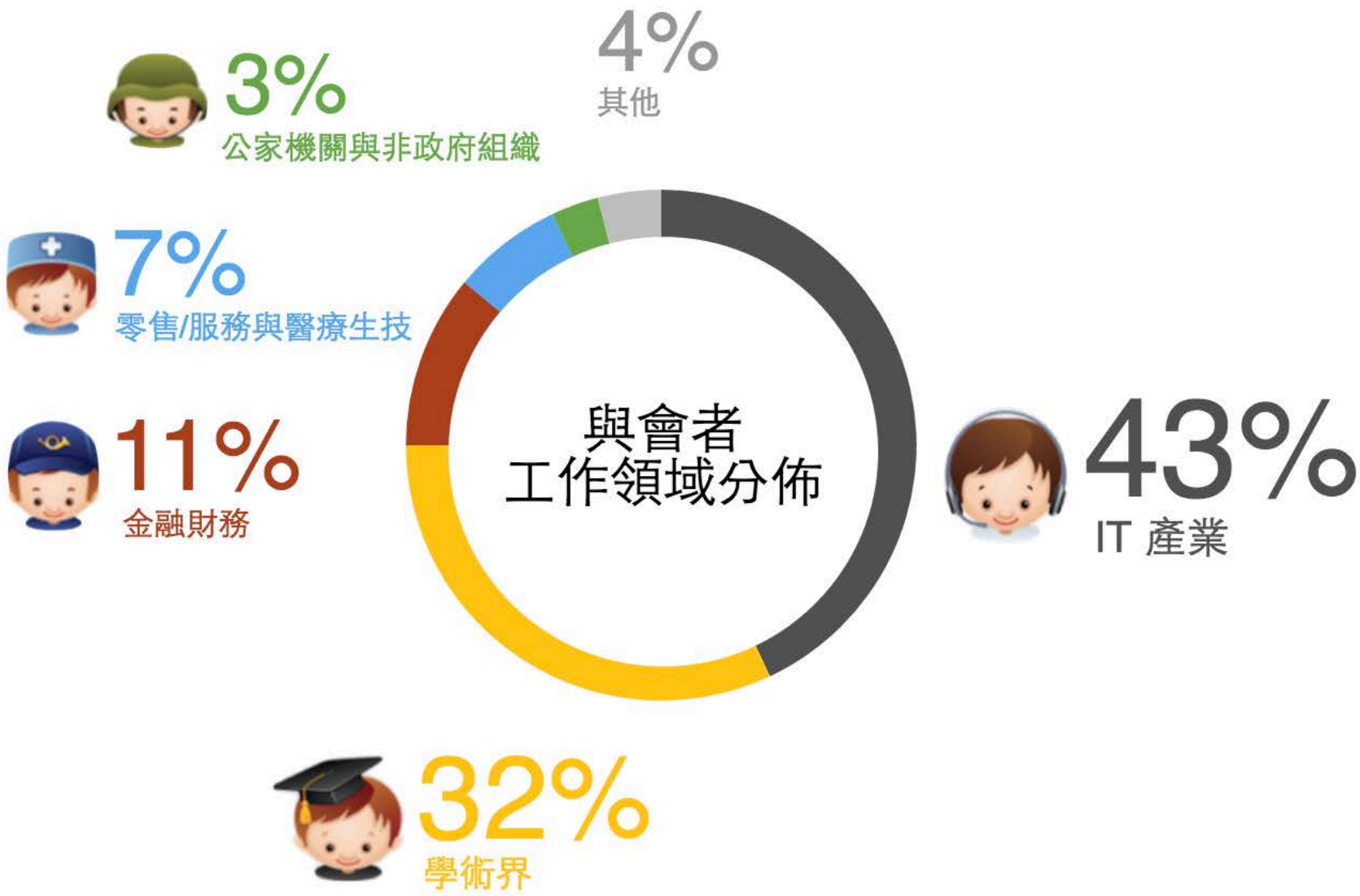
- 產官學合作媒合交流
- 人才媒合場次
- 有志一同交流場次 x 10
  - 社群網路分析
  - 電商、零售及網路行銷
  - 資料視覺化
  - 資訊安全
  - 健康醫療
  - 教育大數據
  - 財務金融
  - 未來城市的交通運輸
  - 人工智慧/機器學習/深度學習
  - 開放資料及個資保護

# 1,718 與會者

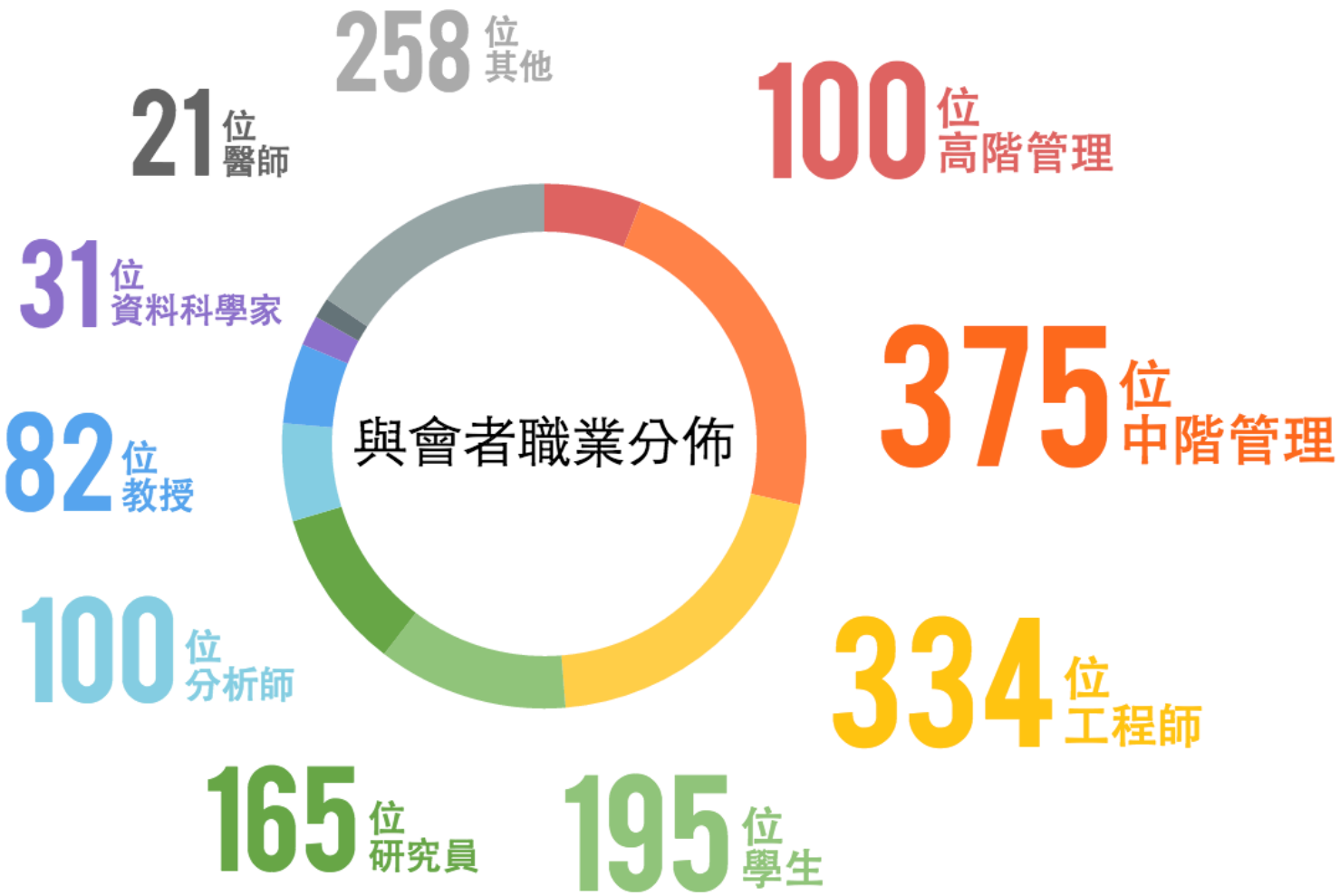
## 活動票券

| 票種 | 販售時間 | 價格 |
| --- | --- | --- |
| 年會主議程 (含交流晚宴) (六)、(日) | 2016/06/01 12:00 ~ 2016/06/17 23:59 結束販售 | TWD$5,480 |
| 年會主議程 (六)、(日) | 2016/06/01 12:00 ~ 2016/06/17 23:59 結束販售 | TWD$4,680 |
| 資料科學的第一堂課：心法、案例分析與團隊建立 (四) | 2016/06/01 12:00 ~ 2016/06/17 23:59 | TWD$1,980 |
| 從資料到知識：從零開始的資料探勘 (四) | 2016/06/01 12:00 ~ 2016/06/17 23:59 結束販售 | TWD$1,980 |
| R 語言資料工程及探勘實務 (四)、(五) | 2016/06/01 12:00 ~ 2016/06/17 23:59 結束販售 | TWD$3,980 |
| 一天搞懂深度學習 (五) | 2016/06/01 12:00 ~ 2016/06/17 23:59 | TWD$1,980 |
| 手把手教你 R 語言資料分析實務 (五) | 2016/06/01 12:00 ~ 2016/06/17 23:59 結束販售 | TWD$2,480 |

一小時完售


4%
其他
3%
公家機關與非政府組織
7%
零售/服務與醫療生技
11%
金融財務
與會者
工作領域分佈
43%
IT 產業
32%
學術界

# 共有 1661 位與會者提供職位銜



# 協辦單位

中央研究院
資訊科技創新研究中心

中央研究院
資訊服務處

中央研究院
統計科學研究所

國際科學與技術資料委員會
中華民國委員會

財團法人資訊工業策進會
數據科技與應用研究所

中華民國計算語言學學會

FINDIT

Sudo Recruit

中華機率統計學會

台大智慧聯網
創新研究中心

工業技術研究院
巨量資訊科技中心

國家高速網路與計算中心

# 贊助企業

## 鑽石級

國泰人壽 Cathay Life Insurance

funP Innovation Group 雲沛創新集團

新加坡商 鈦坦科技 TITANSOFT

Eit 東森信息科技股份有限公司 Eastern Information Technology Co., Ltd

htc

WAVENET 潮網科技

玉山金控 E.SUN FHC

IBM

## 白金級

Vpon 威朋

## 黃金級

Deloitte. 勤業眾信

Google

BRIDGEWELL 宇匯知識科技

gamania

Migo Marketing with Intelligence

MERIDIGEN CREATION / POSSIBILITY / VITALITY 宣捷生物科技股份有限公司 MERIDIGEN BIOTECH CO., LTD

mozilla

Weather Risk 天氣風險管理開發 WeatherRisk Explore Inc.

GMobi GENERAL MOBILE CORPORATION

icAn entertainment

if
else
王昱舜

Vpon
Come Join Us !
Vpon 威朋大數據集團
Vpon
Asia's Leading
Big Data Mobile Advertising
Vpon 威朋
大數據
由我掌
Vpon 威朋

2016
台灣資料科學
愛好者年會
玉山金控
funP
htc
Google
mozilla

Analytics,
Big Data, and
Data Science
Courses

討論區 成員 活動 相片 檔案

搜尋這個社團

撰寫貼文 新增相片／影片 建立票選活動 更多

留個言吧.....

最新動態

陳昇瑋更新了社團相片。
10 小時

新增成員

+ 輸入姓名或電子郵件地址.....

成員 152 位成員（18 位新成員）

說明 編輯

2016 台灣資料科學年會已圓滿結束，四天活動共將有 1800 人次參與。今年的年會做了一些新的嘗試，我們舉辦三種形式的交流場次…… 更多

# 資料分析可以拿來解決什麼問題？

1. 計算社會科學
2. 遊戲外掛偵測
3. 遊戲玩家忠誠度分析
4. 線上遊戲市場表現預測
5. 未知號碼電話該不該接？
6. 有沒有人在偷用你的臉書？
7. 釣魚網頁偵測
8. 如何輔助線上遊戲虛寶銷售

# 計算社會科學初探 - 當電腦科學家遇上社會科學

陳昇瑋
台灣資料科學協會理事長
中央研究院資訊科學研究所研究員

# Computational Social Science

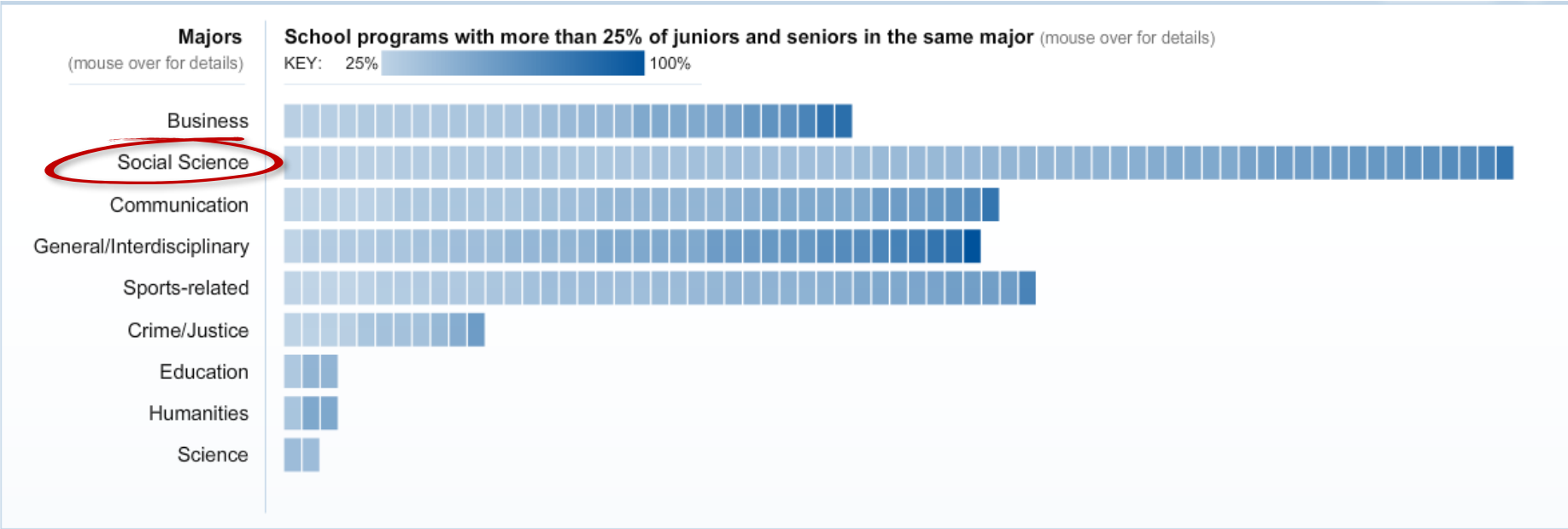
***The Collaborative Futures of Big Data, Computer Science, and Social Sciences***

Sheng-Wei (Kuan-Ta) Chen

Institute of Information Science
Academia Sinica

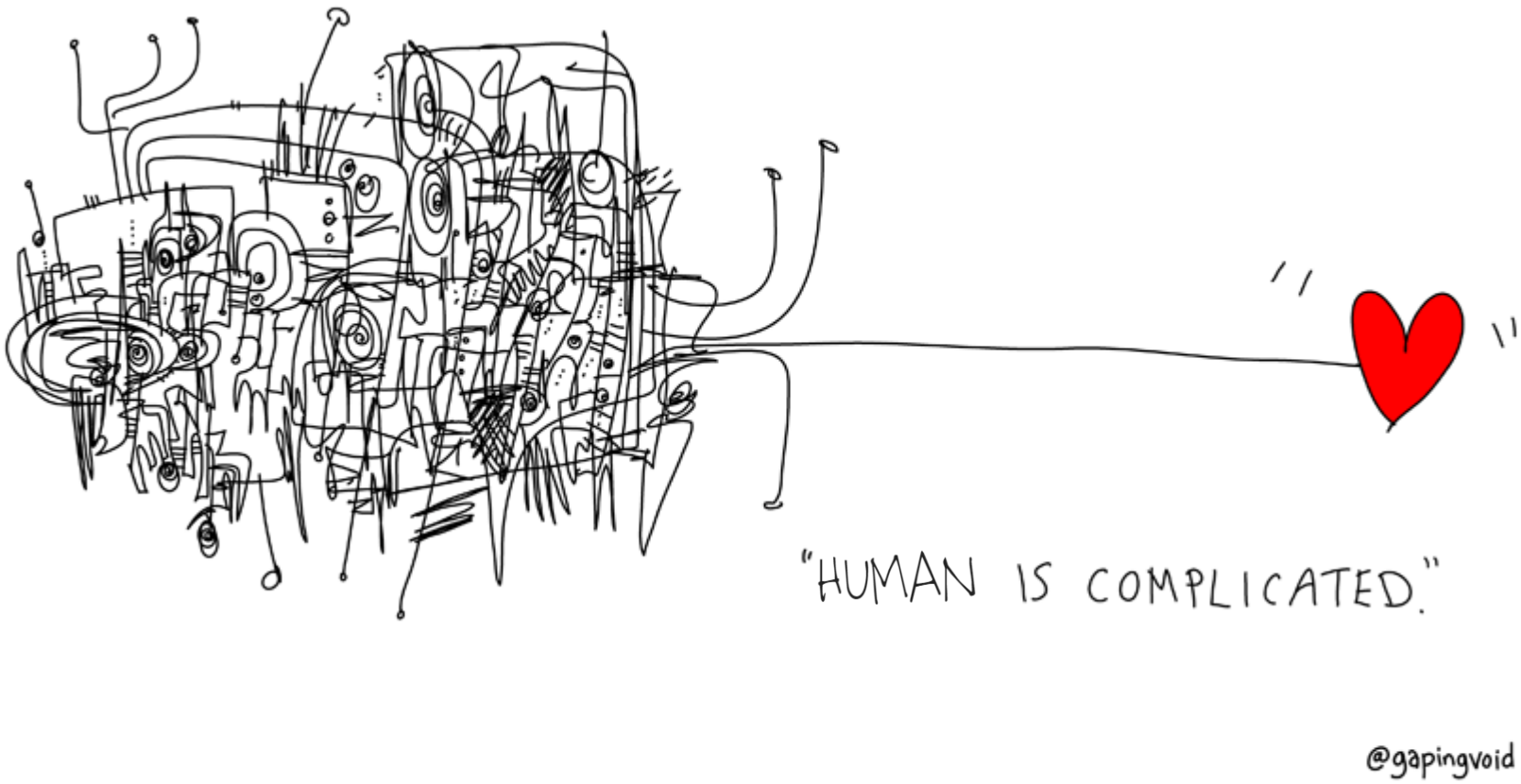
# PEOPLE

# The Favorite Major for US College Athletes




(Source: USA Today, http://usatoday30.usatoday.com/sports/college/2008-11-18-majors-graphic_N.htm)

# Social Science



# Social Life is Hard to See

- We can interview *friends*, but we cannot interview a *friendship*
  - Fleeting interaction
  - In private
  - Tedious to record over time, especially in large groups

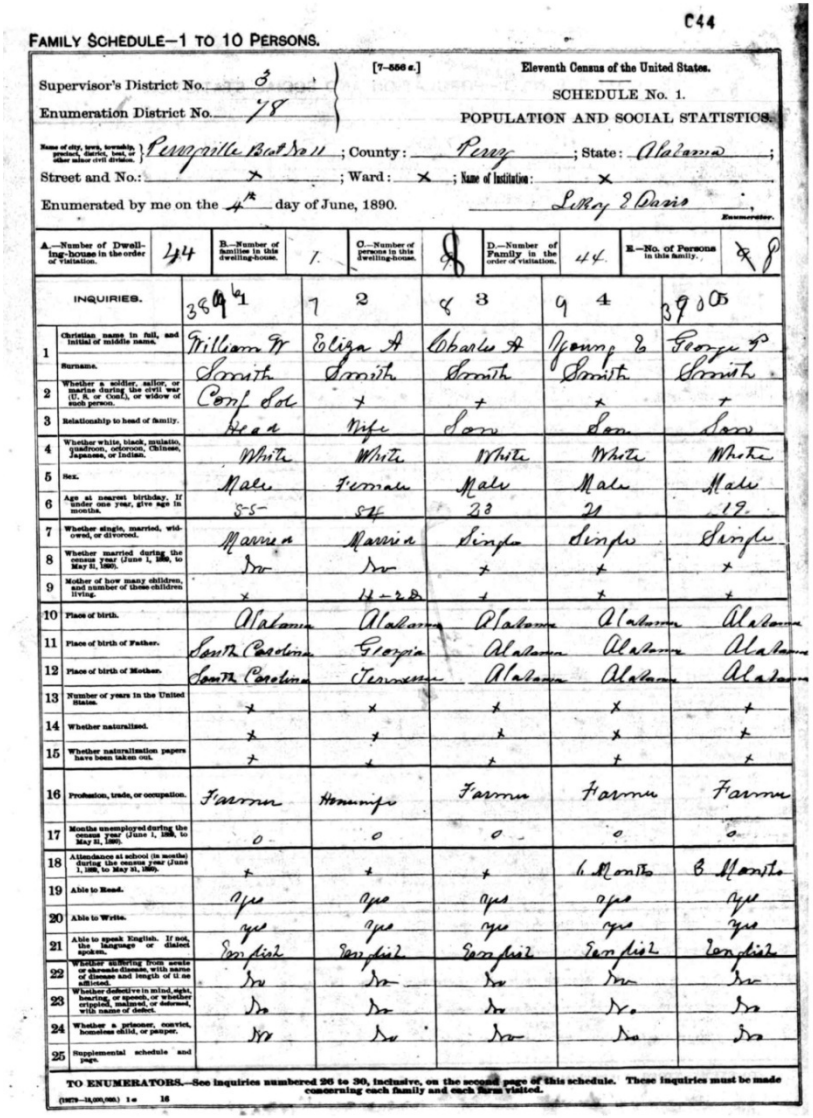
# Bigger Problems

- Social phenomena involve many individuals interacting to produce *collective entities*
  - firms, markets, cultures, political parties, social movements, audiences
  - "**Micro-Macro**" problem (aka "Emergence")

- Micro-macro problems are hard to study empirically
  - Difficult to collect observational data about individuals, networks, and populations **at same time**
  - Even more difficult to do "macro" scale experiments

# 1890 US Census

US 1890 Census Population Schedules



C44

FAMILY SCHEDULE—1 TO 10 PERSONS.

[7-556 a.]

Eleventh Census of the United States.

SCHEDULE No. 1.

POPULATION AND SOCIAL STATISTICS.

Supervisor's District No. 3

Enumeration District No. 78

Name of city, town, township, precinct, district, ward, or other minor civil division: Perryville Beat No 11; County: Perry; State: Alabama;

Street and No.: x; Ward: x; Name of Institution: x

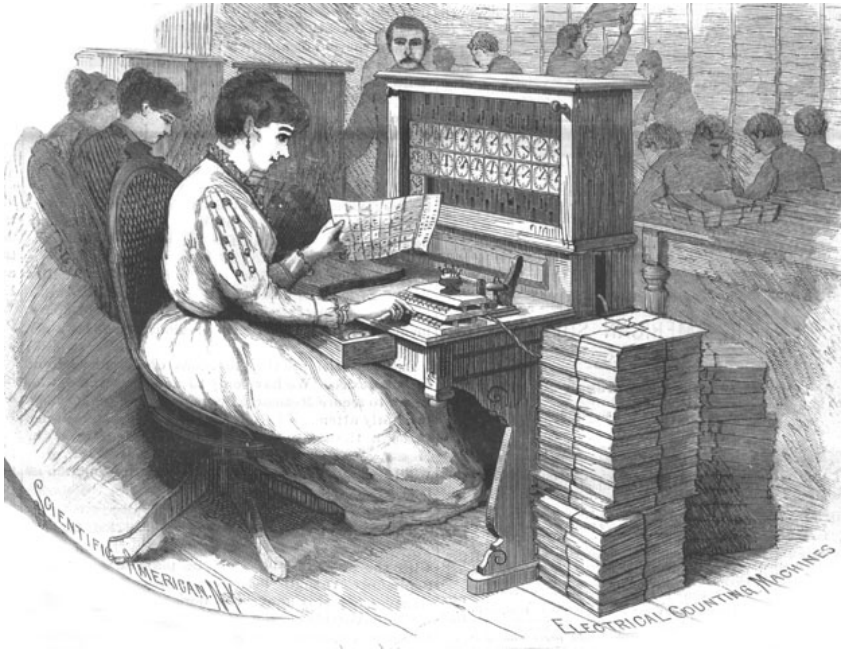
Enumerated by me on the 4th day of June, 1890. Lilley L Davis, Enumerator.

A.—Number of Dwelling-house in the order of visitation: 44. B.—Number of families in this dwelling-house: 1. C.—Number of persons in this dwelling-house: 8. D.—Number of Family in the order of visitation: 44. E.—No. of Persons in this family: 8

| INQUIRIES. | 38 1 | 7 2 | 8 3 | 9 4 | 39 5 |
|---|---|---|---|---|---|
| 1 Christian name in full, and initial of middle name. | William M | Eliza A | Charles A | Young B | George P |
| Surname. | Smith | Smith | Smith | Smith | Smith |
| 2 Whether a soldier, sailor, or marine during the civil war (U. S. or Confd.), or widow of such person. | Confd Sol | x | x | x | x |
| 3 Relationship to head of family. | Head | Wife | Son | Son | Son |
| 4 Whether white, black, mulatto, quadroon, octoroon, Chinese, Japanese, or Indian. | White | White | White | White | White |
| 5 Sex. | Male | Female | Male | Male | Male |
| 6 Age at nearest birthday. If under one year, give age in months. | 55 | 54 | 23 | 21 | 19 |
| 7 Whether single, married, widowed, or divorced. | Married | Married | Single | Single | Single |
| 8 Whether married during the census year (June 1, 1889, to May 31, 1890). | No | No | x | x | x |
| 9 Mother of how many children, and number of those children living. | x | 4-2 D | x | x | x |
| 10 Place of birth. | Alabama | Alabama | Alabama | Alabama | Alabama |
| 11 Place of birth of Father. | South Carolina | Georgia | Alabama | Alabama | Alabama |
| 12 Place of birth of Mother. | South Carolina | Tennessee | Alabama | Alabama | Alabama |
| 13 Number of years in the United States. | x | x | x | x | x |
| 14 Whether naturalized. | x | x | x | x | x |
| 15 Whether naturalization papers have been taken out. | x | x | x | x | x |
| 16 Profession, trade, or occupation. | Farmer | Housekeeper | Farmer | Farmer | Farmer |
| 17 Months unemployed during the census year (June 1, 1889, to May 31, 1890). | 0 | 0 | 0 | 0 | 0 |
| 18 Attendance at school (in months) during the census year (June 1, 1889, to May 31, 1890). | x | x | x | 6 Months | 6 Months |
| 19 Able to Read. | yes | yes | yes | yes | yes |
| 20 Able to Write. | yes | yes | yes | yes | yes |
| 21 Able to speak English. If not, the language or dialect spoken. | English | English | English | English | English |
| 22 Whether suffering from acute or chronic disease, with name of disease and length of it as afflicted. | No | No | No | No | No |
| 23 Whether defective in mind, sight, hearing, or speech, or whether crippled, maimed, or deformed, with name of defect. | No | No | No | No | No |
| 24 Whether a prisoner, convict, homeless child, or pauper. | No | No | No | No | No |
| 25 Supplemental schedule and page. | | | | | |

TO ENUMERATORS.—See inquiries numbered 26 to 30, inclusive, on the second page of this schedule. These inquiries must be made concerning each family and each farm visited.

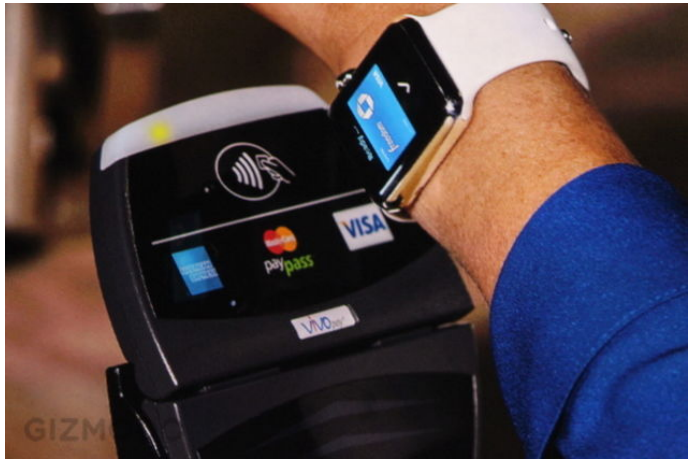
(10279—15,000,000.) 1-c 16




1st time *Hollerith* machines were used to tabulate US Census data (population: 62,947,714)

# The Era of Big Data

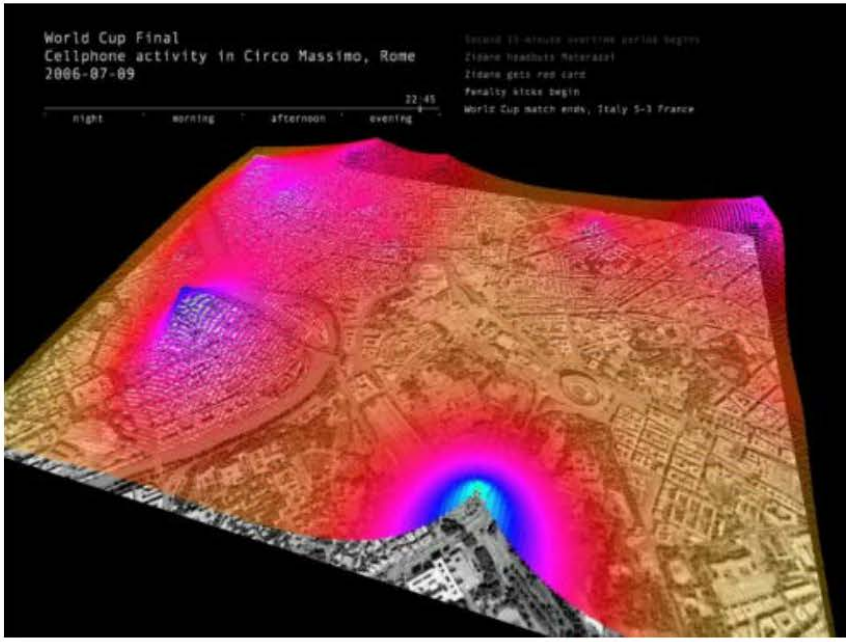
- Past: Government data, national survey data
- Today: A variety of new data sources
  - Economic data: trade, finance, e-cash / e-wallet, …
  - GIS data: satellite, GPS loggers, laser scanning cars, …
  - Sensor data: video surveillance, smart phones, wearable devices, mobile apps, beacons, …



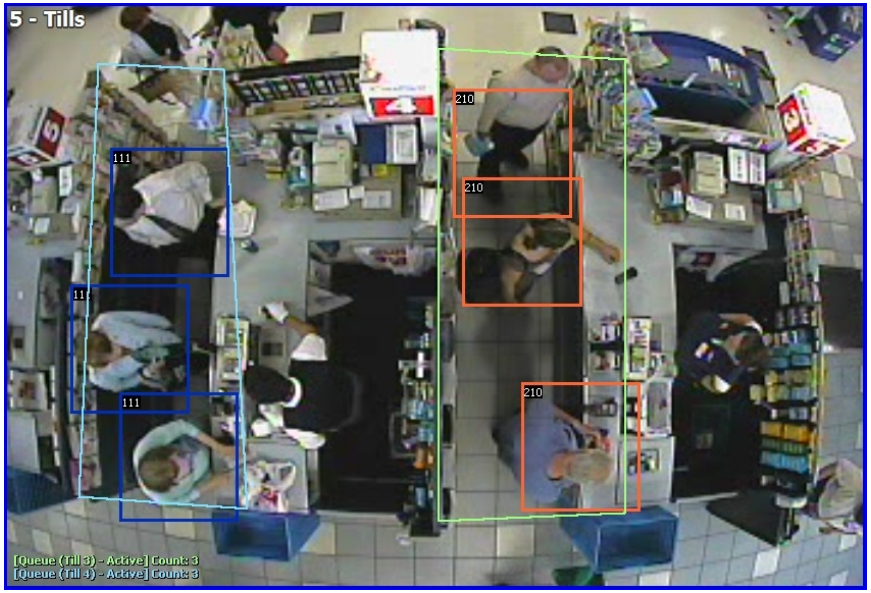
# New kinds of data

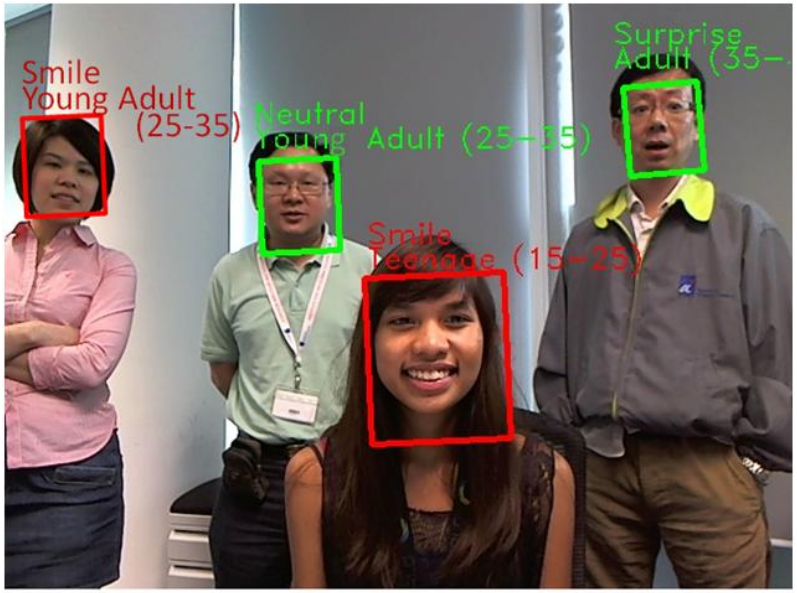
*Urban movement analysis from GPS/phone data*

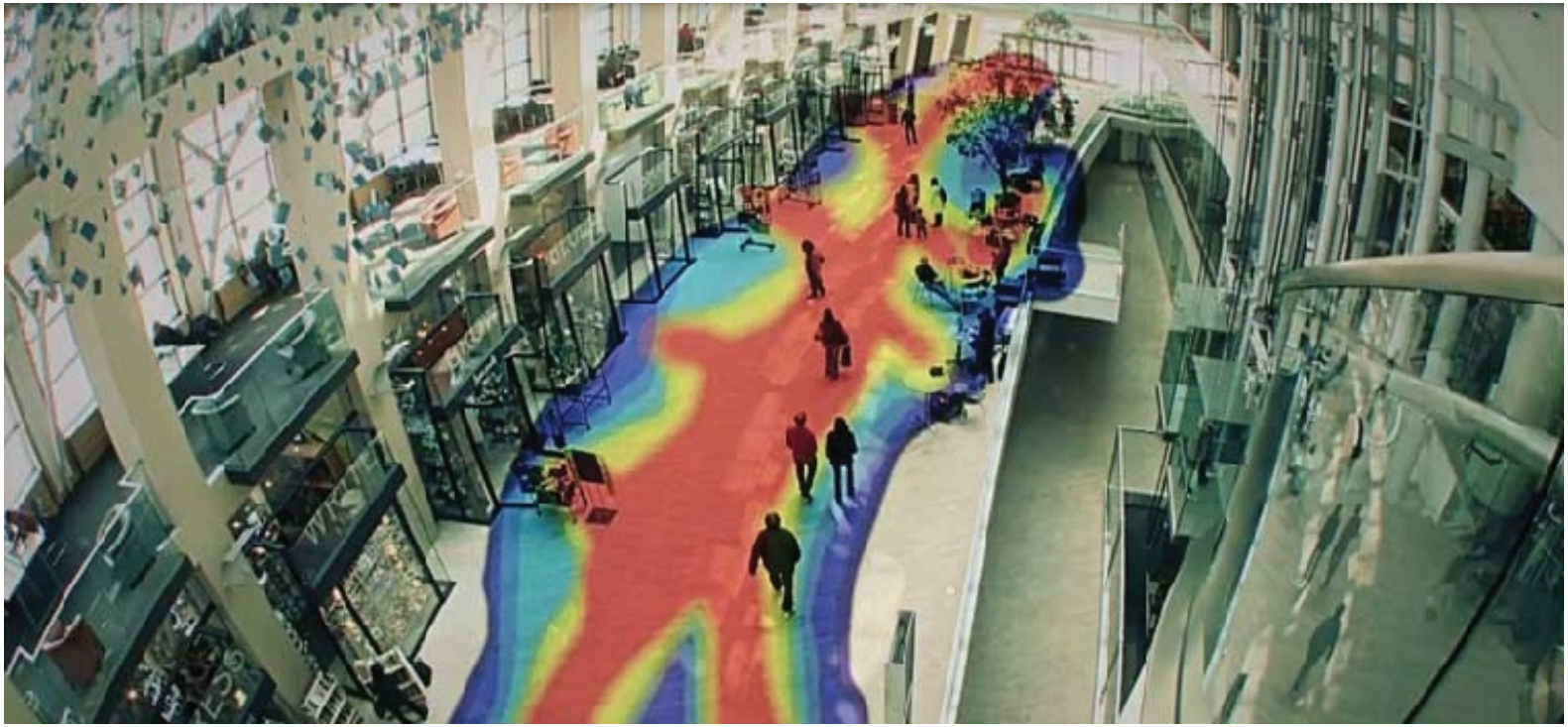
The Amsterdam Real Time Project




Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-time urban monitoring using cell phones: A case study in Rome, *IEEE Transactions on Intelligent Transportation Systems, 12*(1), 141-151.


5 - Tills
111
111
111
210
210
210
[Queue (Till 3) - Active] Count: 3
[Queue (Till 4) - Active] Count: 3


Smile
Young Adult
(25-35)
Neutral
Young Adult (25-35)
Smile
Teenage (15-25)
Surprise
Adult (35-



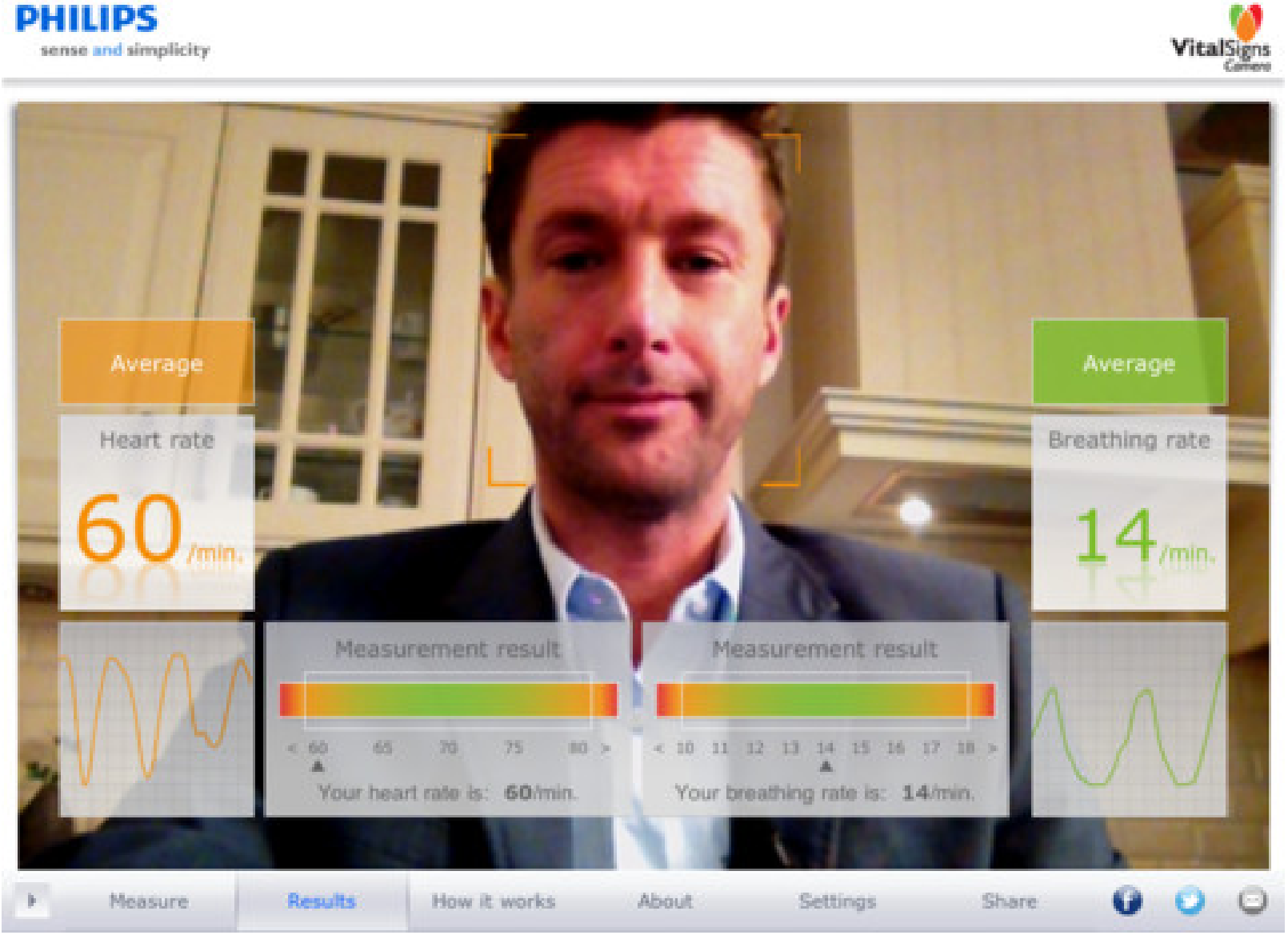
(a) Input

# Computer vision for healthcare

Video magnification

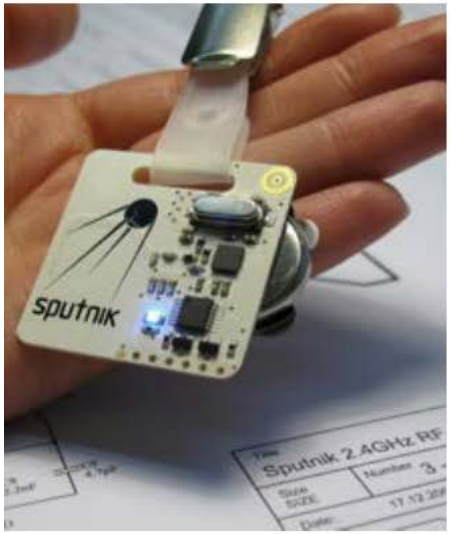


(Slide Credit: Jia-Bin Huang)

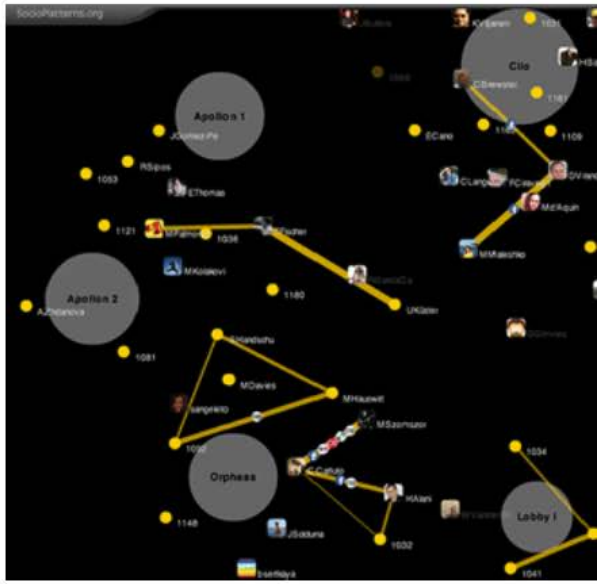

PHILIPS
sense and simplicity
VitalSigns Camera
Average
Heart rate
60 /min.
Average
Breathing rate
14 /min.
Measurement result
< 60 65 70 75 80 >
Your heart rate is: 60/min.
Measurement result
< 10 11 12 13 14 15 16 17 18 >
Your breathing rate is: 14/min.
Measure
Results
How it works
About
Settings
Share

# New kinds of data

*Social Sensing via RFID*

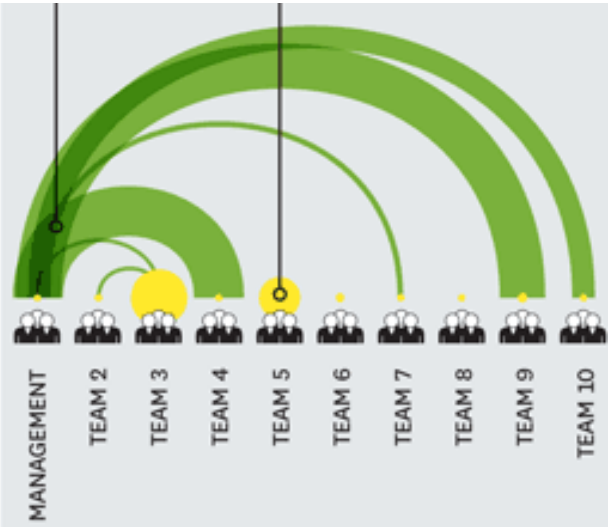



SocioPatterns

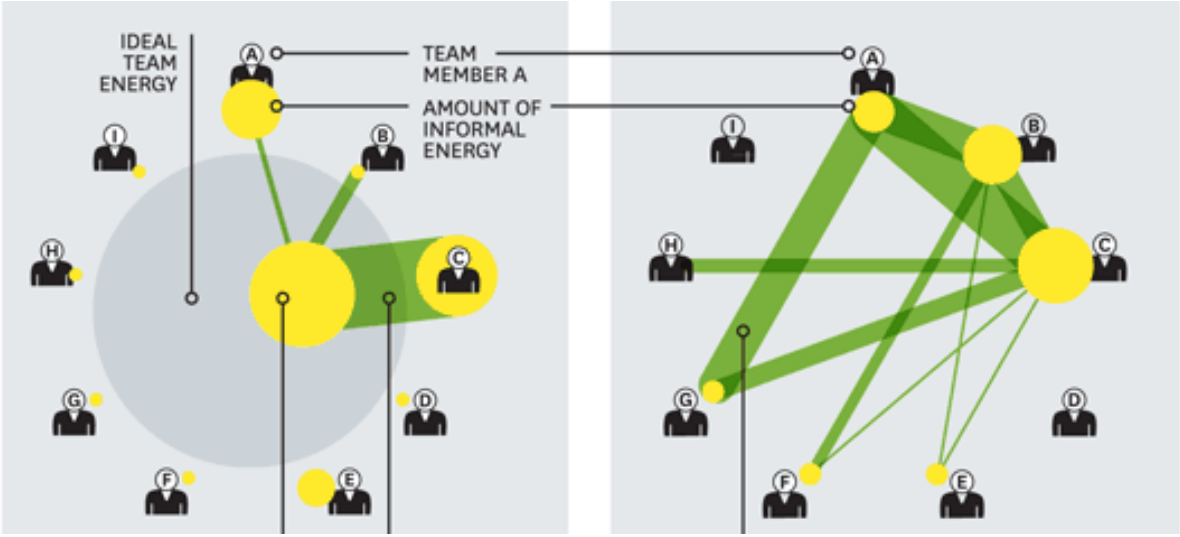



Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J. F., & Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PloS one*, *5*(7), e11596.

# Engagement and Exploration




- Standing face-to-face?
- Physical distance
- Hand gesture, posture
- Conversation patterns
- Frequency of interruptions



# Web as a Record of Social Interaction

- Public web pages / discussions
- Twitter, Facebook, blogs, news groups, wikis, MMOGs, Instagram, LastFM, Flickr, Spotify
- Private email, Whatsapp, LINE, Slack
- Text, images, sounds: speeches, commercials

# New kinds of data

*Human mobility in societies*

Check-ins (Foursqaure, Gowalla, Twitter, ...)

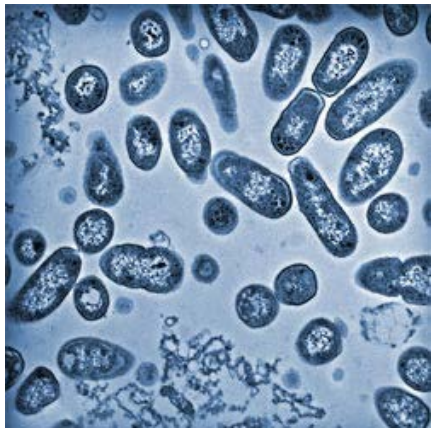
Cheng, Zhiyuan, et al. "Exploring Millions of Footprints in Location Sharing Services." *ICWSM* 2011 (2011): 81-88.

# Computational Social Science

*The science that investigates social phenomena through the medium of computing and statistical data processing.*

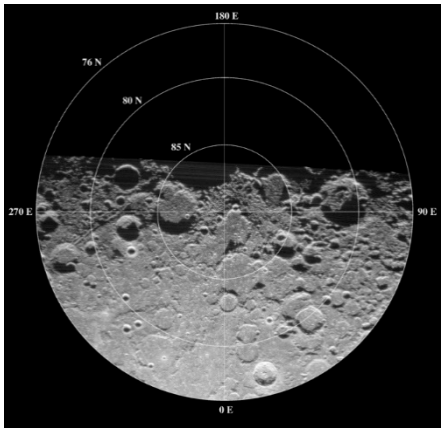
# Computational Social Science

- An **instrument-enabled** scientific discipline



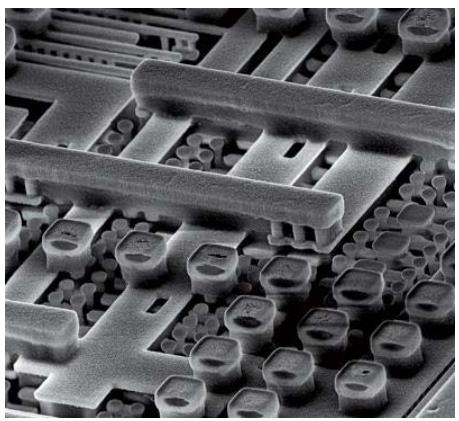
microbiology

microscope




radio astronomy

radar



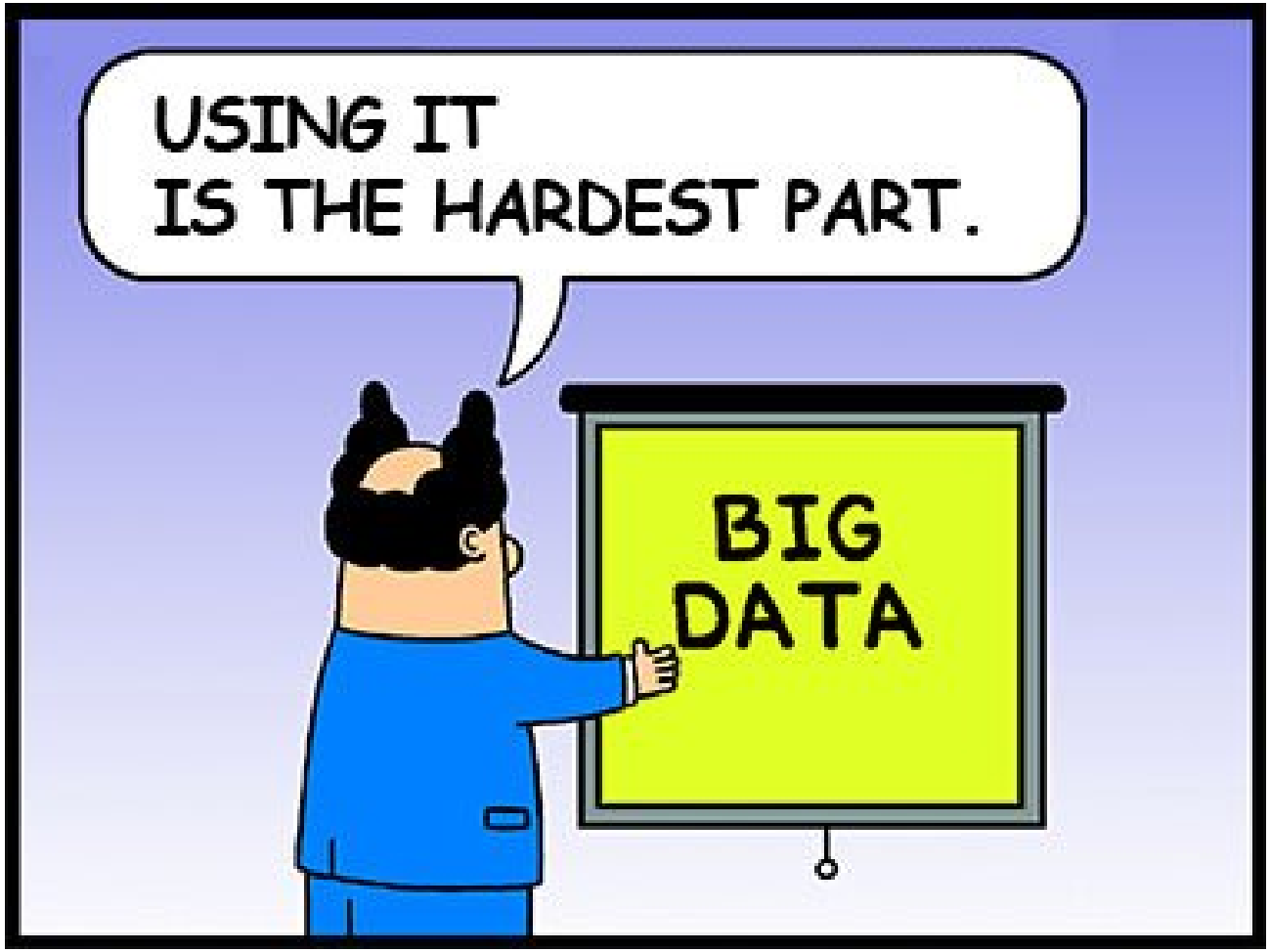
nanoscience

electron microscope

understanding
Complex
global
Social
complex
network
new
emergence
like
Press
level
Special
research
Social
models
levels
see
work
society
process
computational
information
Topics
systems
behaviour
complexity
Computational
ICT
data
study
science
experiments
people
analysis
model
entities
different
University
cultural
phenomena
modelling
need
understand
Science
networks
Journal
one
also
dynamics
properties

BIG
DATA
INFORMATION
PETABYTES
SEARCH
INTERNET
ANALYSIS
RESEARCH
LEARNING
WORLD
EVERY
BASED
MANAGEMENT
TECHNOLOGIES
GOVERNMENT
PROCESS
VOLUME
SIZE
EXABYTES
LARGE
STORAGE
SOFTWARE
SAN
NAS
MILLION
SETS
DEVELOPMENT
TECHNOLOGY
SCIENCE
FUTURE
BILLION
NEW
INITIATIVE
ECONOMIC
PARALLEL
TERABYTES
FRAMEWORK
BETWEEN
BUSINESS
DATABASES
PROCESSING
ANALYTICS
VISUALIZATION
STATISTICS
DEFINITION
USING
MASSIVE
FACTOR
ONLINE
DAY
SYSTEMS
MAKING
INTEGRATION
WORLDWIDE
RESULTS
SOCIAL
YEARS
DRIVE
REPLICATION
HANDLE
SOURCES
CHALLENGES
LIMITS
COMPLEX
ANNOUNCED
PAST
CAPACITY
HAND
ONE
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
SECOND
OUT
INCLUDE
NETWORKS
HYPOTHESIS
END
MOST
NEARLY
MAPREDUCE
CRITIQUES
STATE
CRITIQUE
PEOPLE
TRAFFIC
USE
CURRENTLY
APPROACH
MANY
NATIONAL
FLOW
INTELLIGENCE
APPROACHES
PROCESSES
PER
PRIVATE
QUERIES
DETERMINE
WELL
SET
DEPARTMENT
HIGHER
ALGORITHMS
COMPANIES
SIMULATIONS
HIGH
USERS
AMOUNT
REAL
RATE
STRUCTURE
WORK
APPLICATIONS
INFRASTRUCTURE
TERMS
YEAR
CAPABILITIES
SUCCESSFUL
NEEDED
USES
LESS
HUGE
COMPARED
VARIOUS
TIMES
SHARED
ARCHITECTURE
PARADIGM
FUNDING
RECORDS
COST
INSIGHT
TOOLS
CENTER
RELATED
TIME
MARKET
SENSOR
VARIETY
WOULD
PROJECT
ANNUAL
NOW
DISTRIBUTED
RELEVANT
USED

HUMANITIES
HEALTH
BIOLOGY
Economics
SOCIOLOGY
CALCULUS
LITERATURE
Grammar
ANATOMY
PHYSICS
World History


USING IT
IS THE HARDEST PART.
BIG
DATA

# Technical Challenges

- Computational infrastructures for dealing with
  - **More data**: analyzing large amounts of data
  - **Fuzzy data**: cleaning up inprecise and noisy data
  - **New kinds of data**: processing real-time sensor streams and web data

- Need for new substantive ideas
- Need for new statistical methods (WHY in addition to WHAT and HOW)

# 3 Common Approaches

Macroscope

Virtual Lab

Empirical Modeling

# APPROACHES

## #1 MACROSCOPE

## #2 VIRTUAL LAB

## #3 EMPIRICAL MODELING

Macroscope

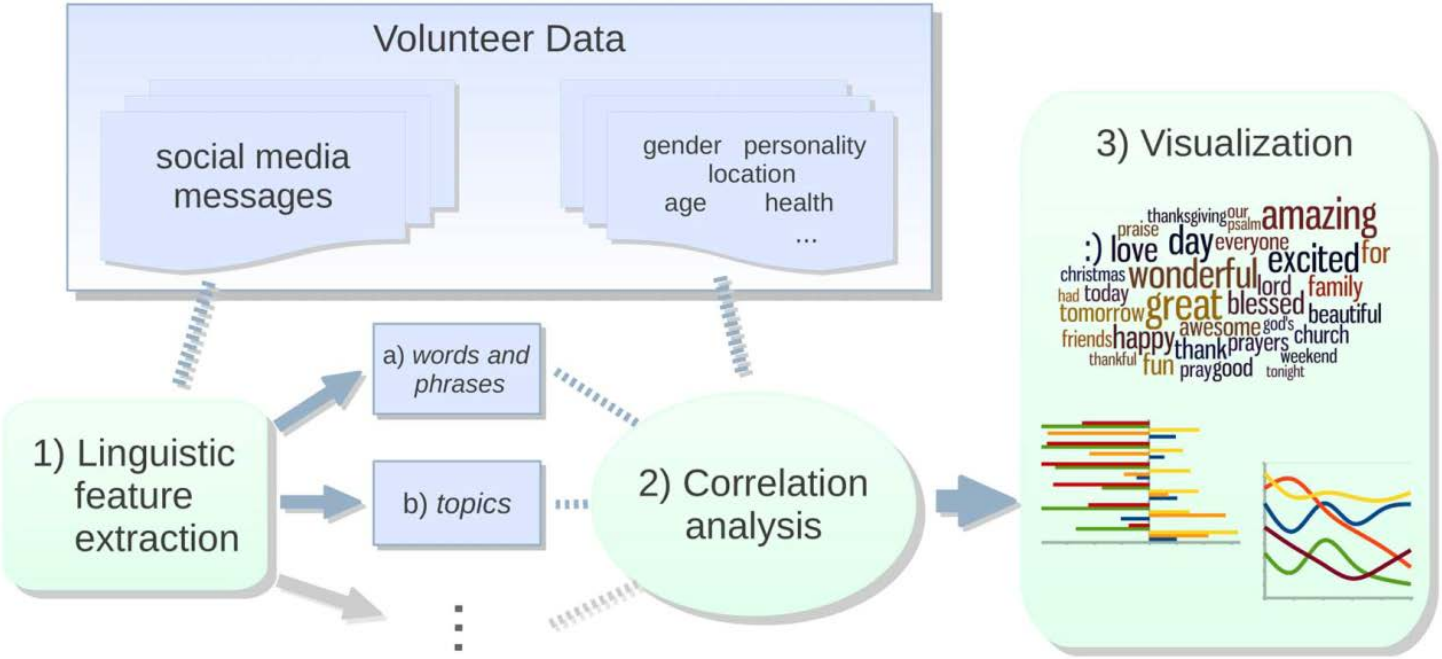
Linguistics

# WE ARE WHAT WE SAY

Schwartz, H. Andrew, et al. "Personality, gender, and age in the language of social media: The open-vocabulary approach." *PloS one* 8.9 (2013): e73791.

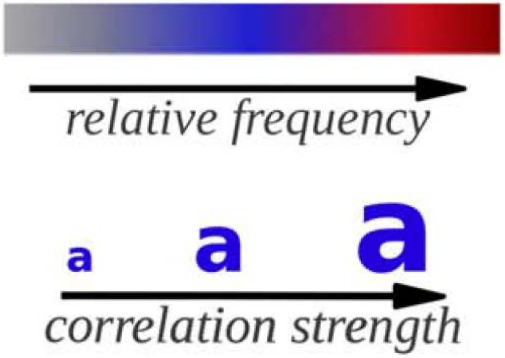
# Dataset

- 700 million words, phrases, and topic instances collected from 75,000 volunteers' FB posts
- Record users' personality (5-factor), gender and age



# What Words Do You Use?

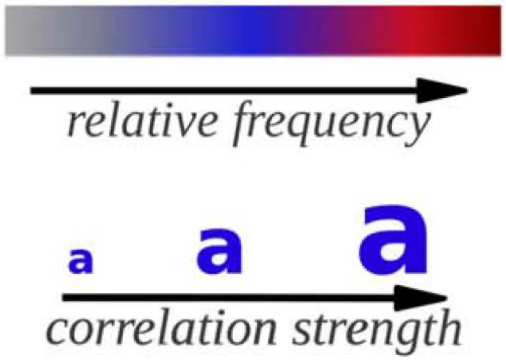
female



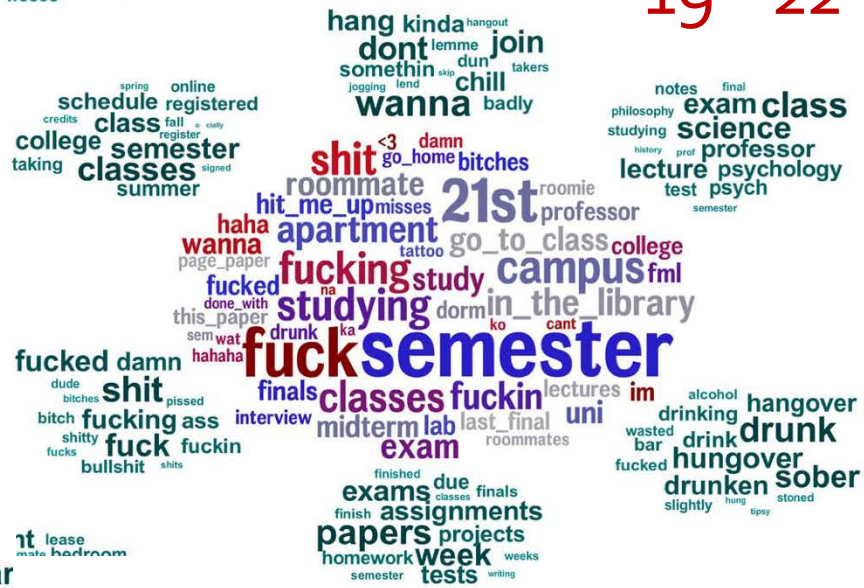

male

# How Old Are You? (#1)

13 - 18

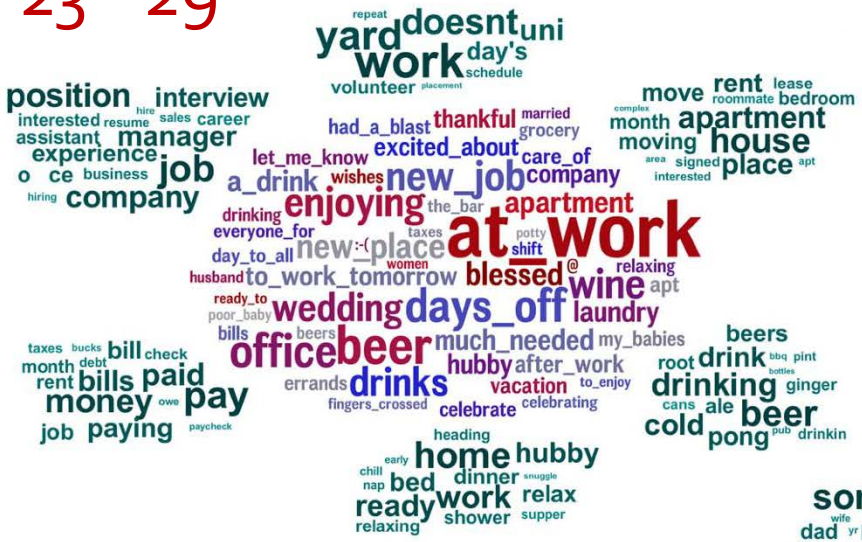


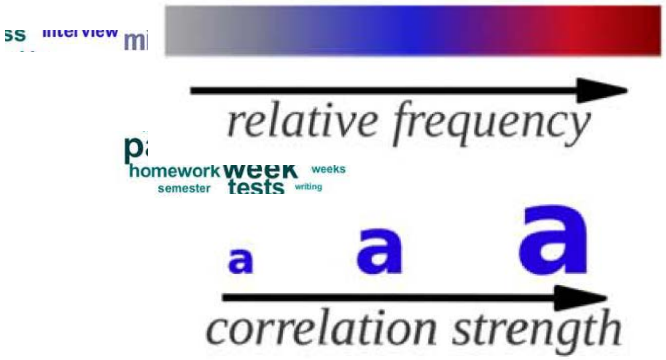

19 - 22



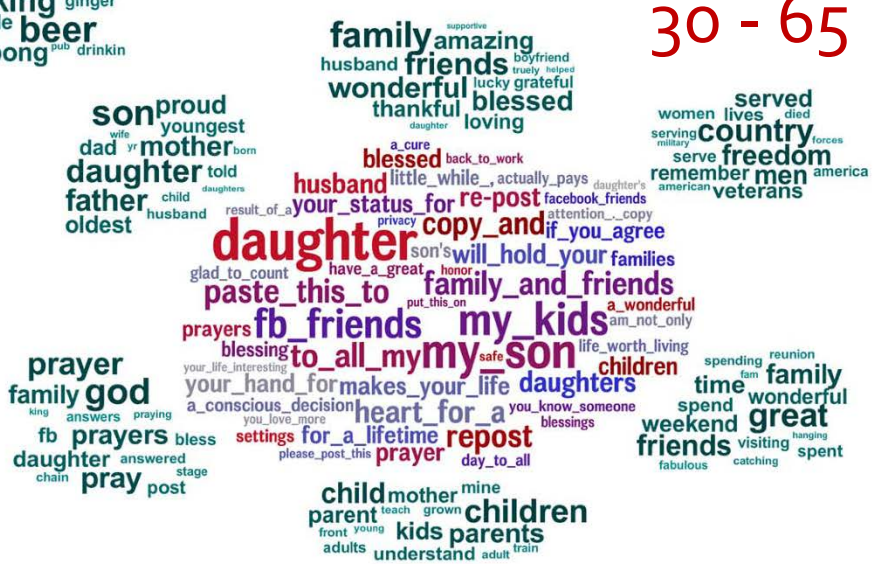
# How Old Are You? (#2)

23 - 29



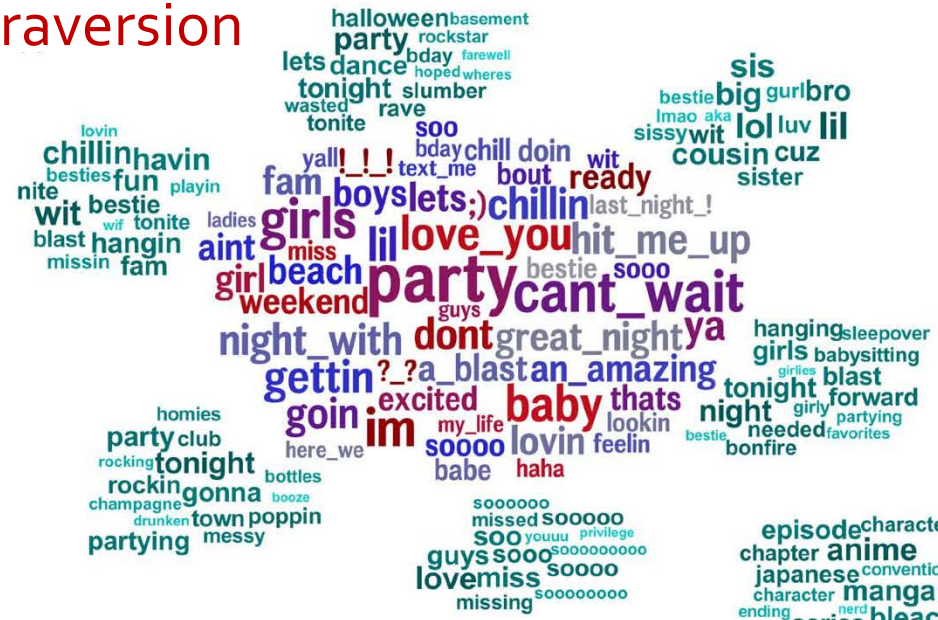



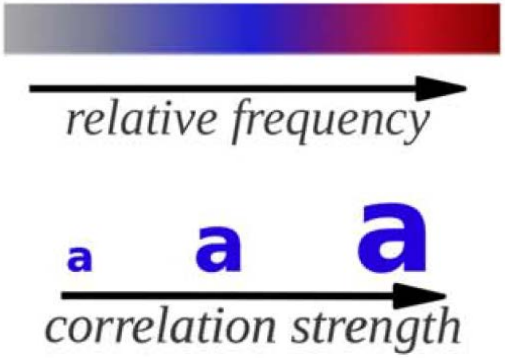

30 - 65



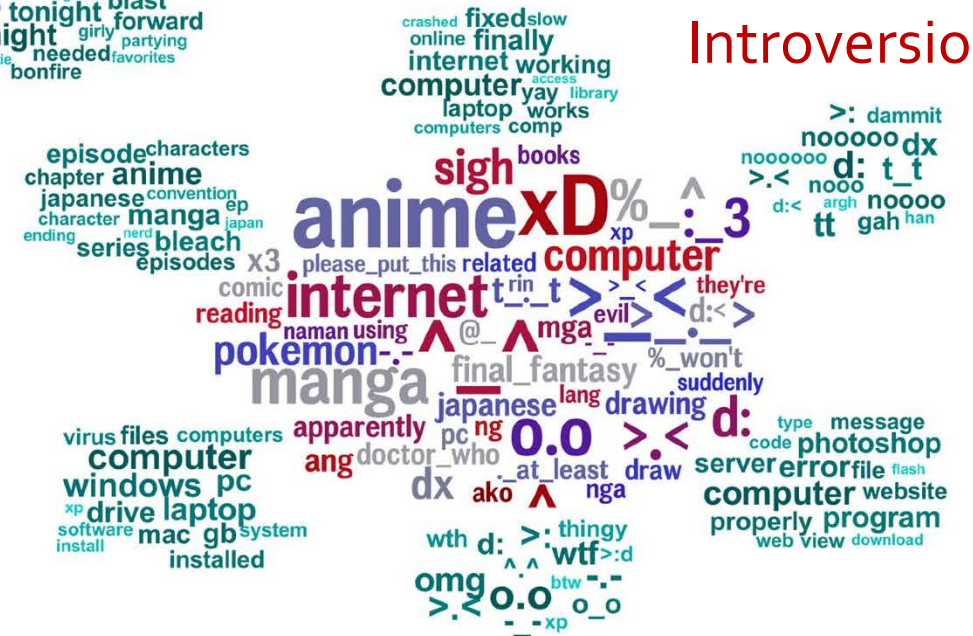
# Personality Traits

Extraversion




Introversion



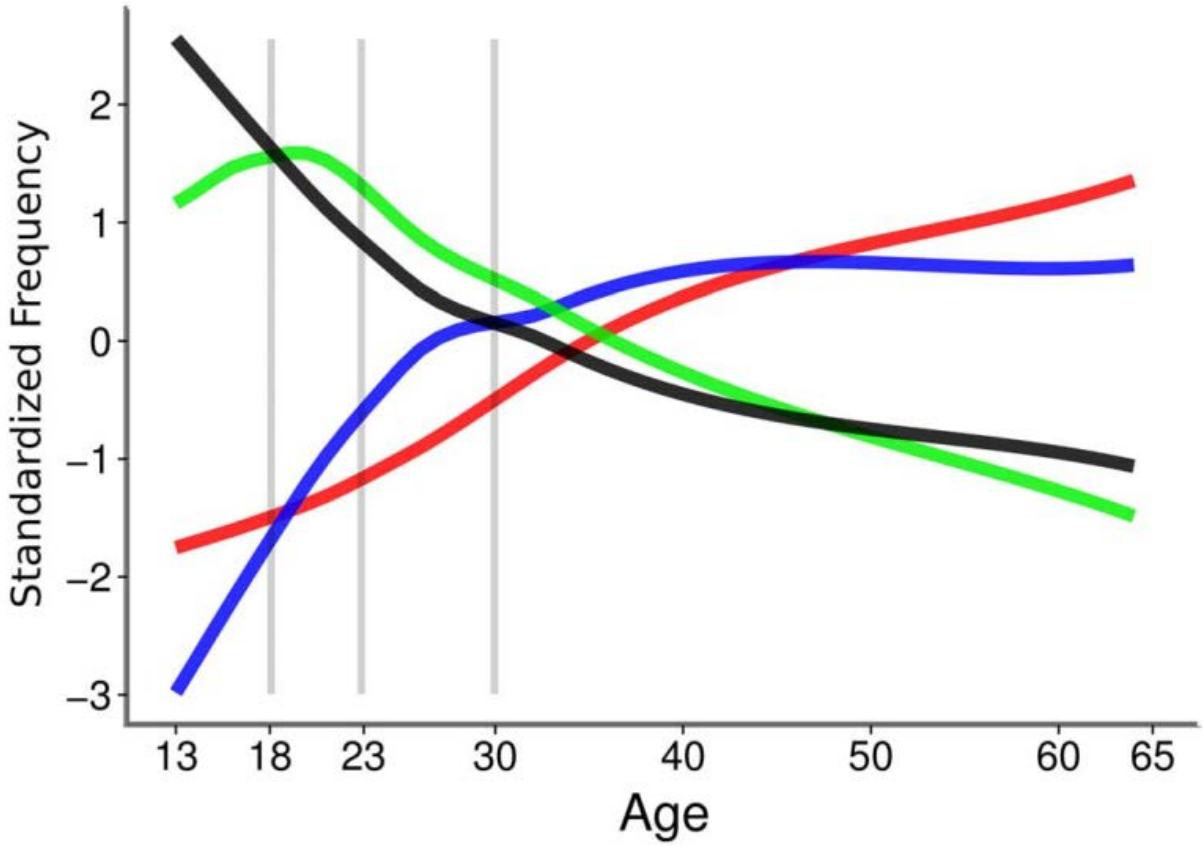
# Topics Across 4 Age-groups

*(30 to 65)* son daughter father mother proud oldest data youngest

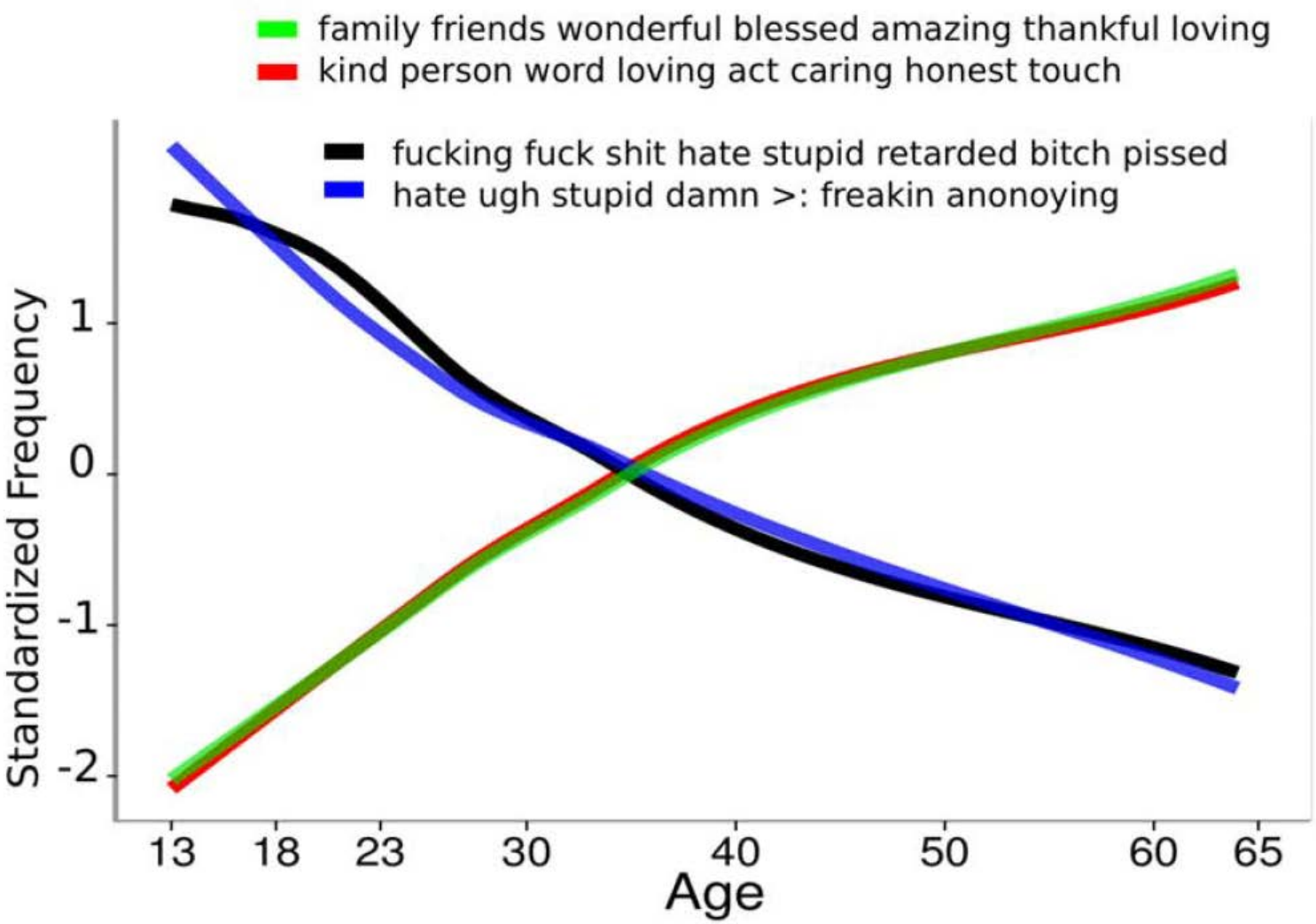
*(23 to 29)* job position company manager interview experience office assistant

*(19 to 22)* classes semester class college schedule summer registered taking

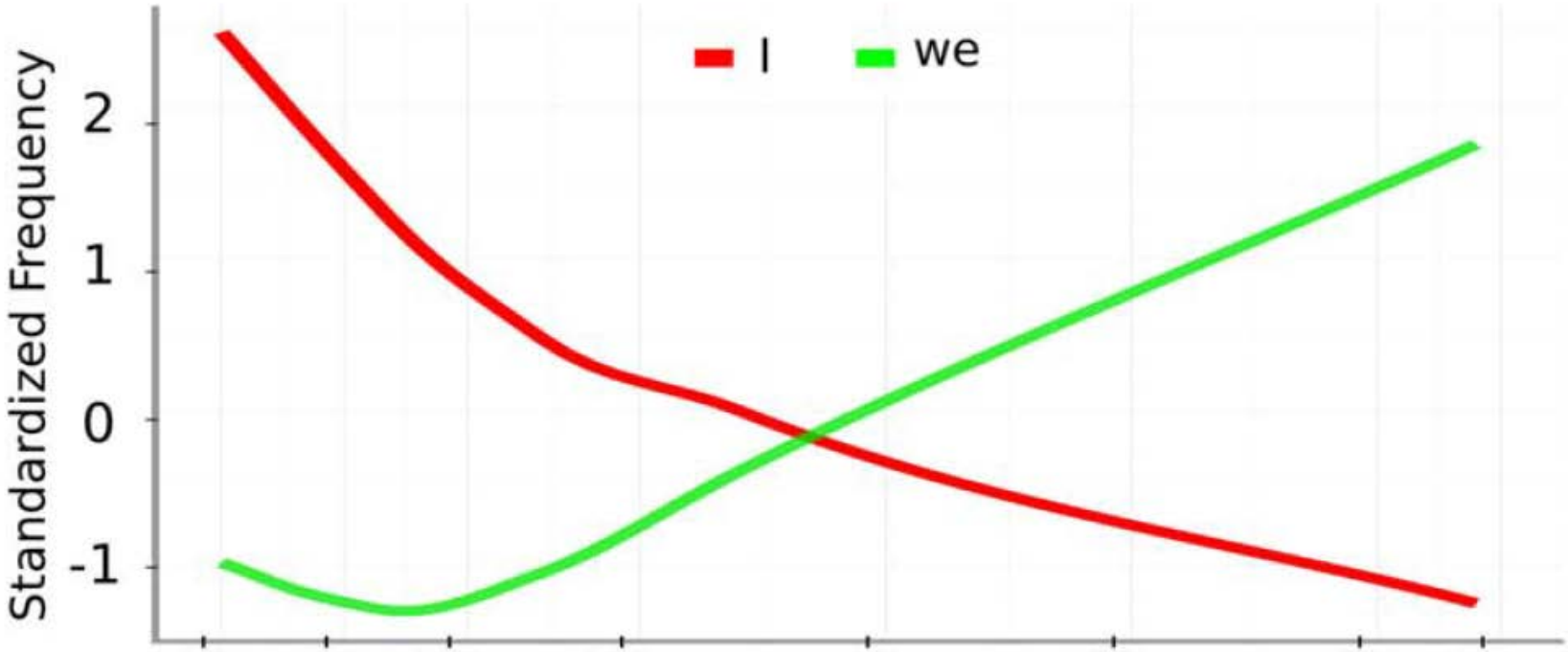
*(13 to 18)* haha lol :p :D ;) hehe jk ;p



# Warm and Negative Words



# Usage of "I" & "We"




**Huge-volume data + simple analysis → crystal clear language use patterns**

# APPROACHES

#1 MACROSCOPE

**#2 VIRTUAL LAB**

#3 EMPIRICAL MODELING

# Scaling up the Lab

- Social science experimental heavily constrained by scale and speed
  - Unit of analysis was individuals or small groups
  - Experiments took months to design and run

- Potentially "virtual labs" lift both constraints
  - State of the art ~ 5000 workers, but in principle could construct subject panel ~ 100K – 1M
  - Could shrink hypothesis-testing cycle to days or hours

amazon mechanical turk™
Artificial Artificial Intelligence

Virtual Lab

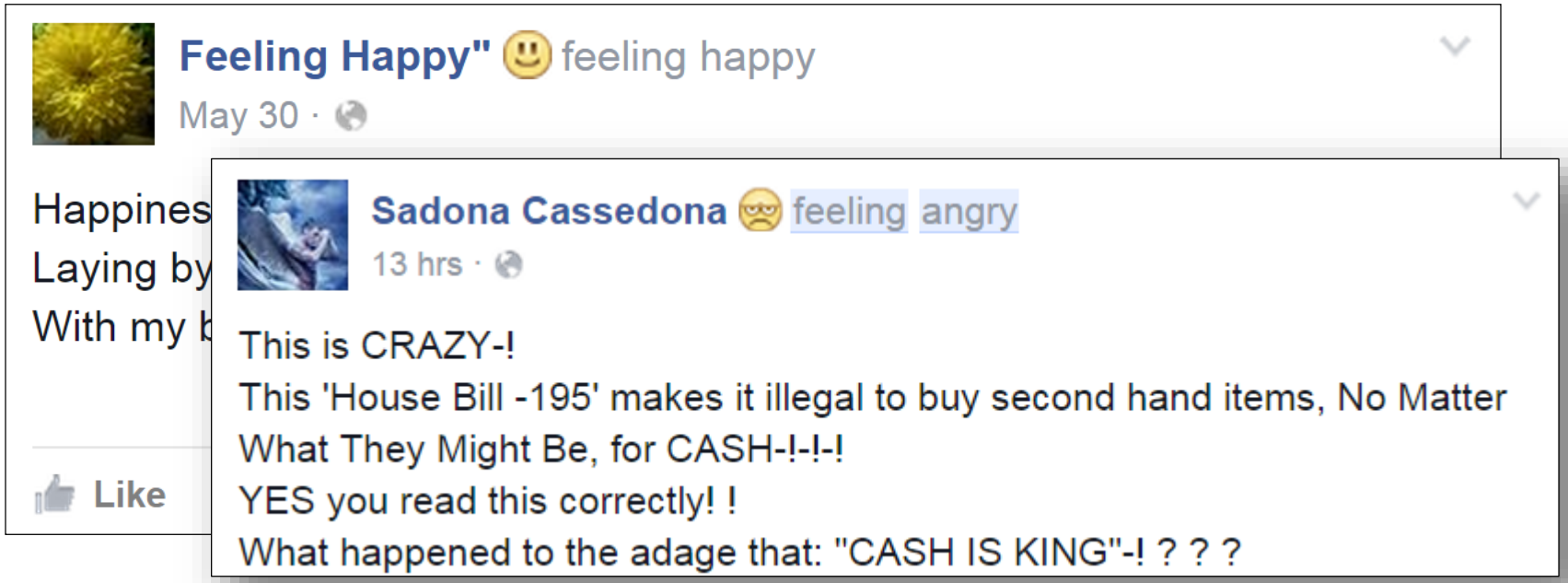
Social Psychology

# MOOD CONTAGION (& MANIPULATION) ON FACEBOOK

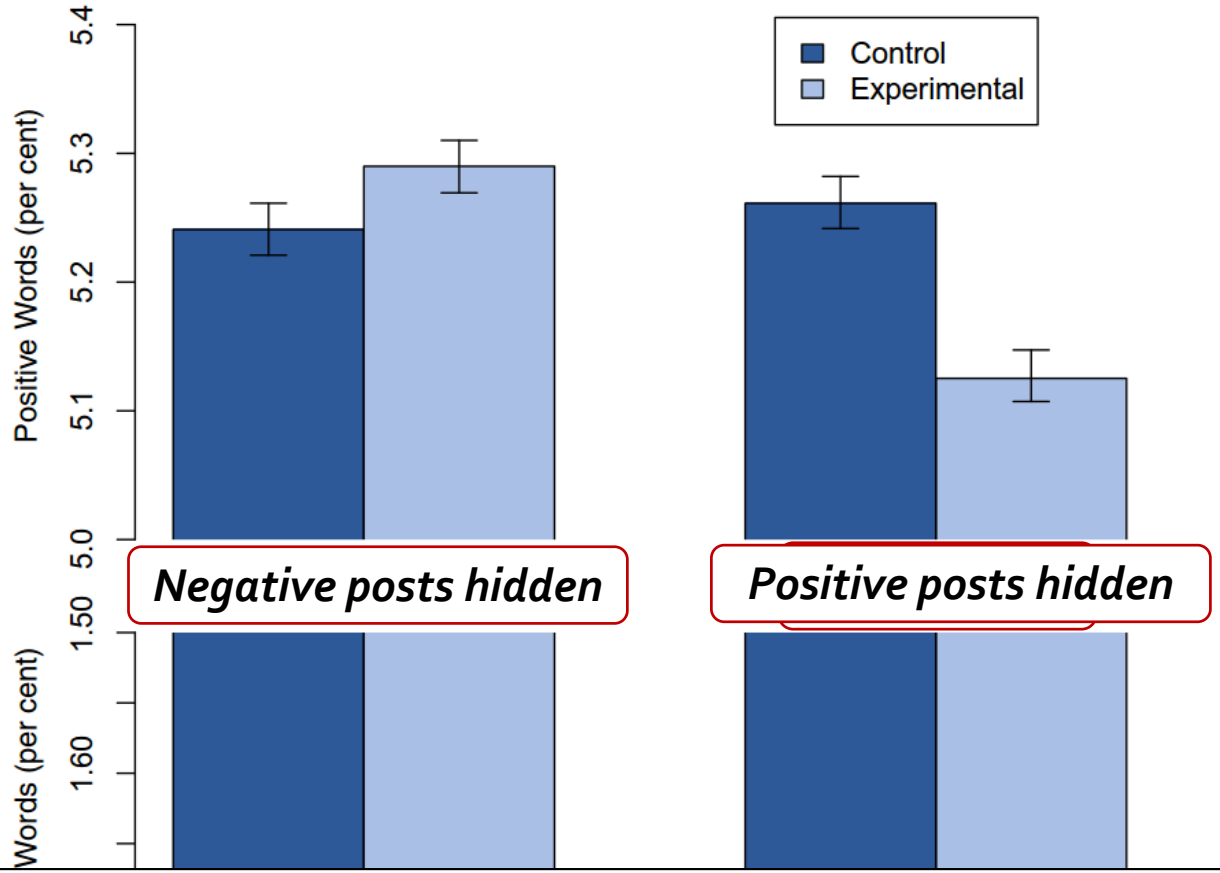
Kramer, Adam DI, Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences*111.24 (2014): 8788-8790.

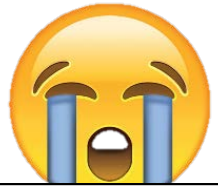
# Facebook Mood Contagion

- 0.7 million (~ 0.04%) users on Facebook
- 3 million posts manipulated in one week
- **Hide** some "positive" or "negative" emotional posts from users (in the experimental group)



# Observations



**Facebook users' emotion can be easily manipulated by changing *ALGORITHMS***

# Ethical Issues (!)

- Unethical experiment because it's conducted without users' consent

> *Well, Facebook's data use policy states that users' information will be used "**for internal operations, including troubleshooting, data analysis, testing, research and service improvement**," meaning that any user can become a lab rat.*

- Serious invasion of users' perceptions about their friend circles (and the society)

Virtual Lab

Social Psychology & Politics

# FACEBOOK "I VOTED" BUTTON

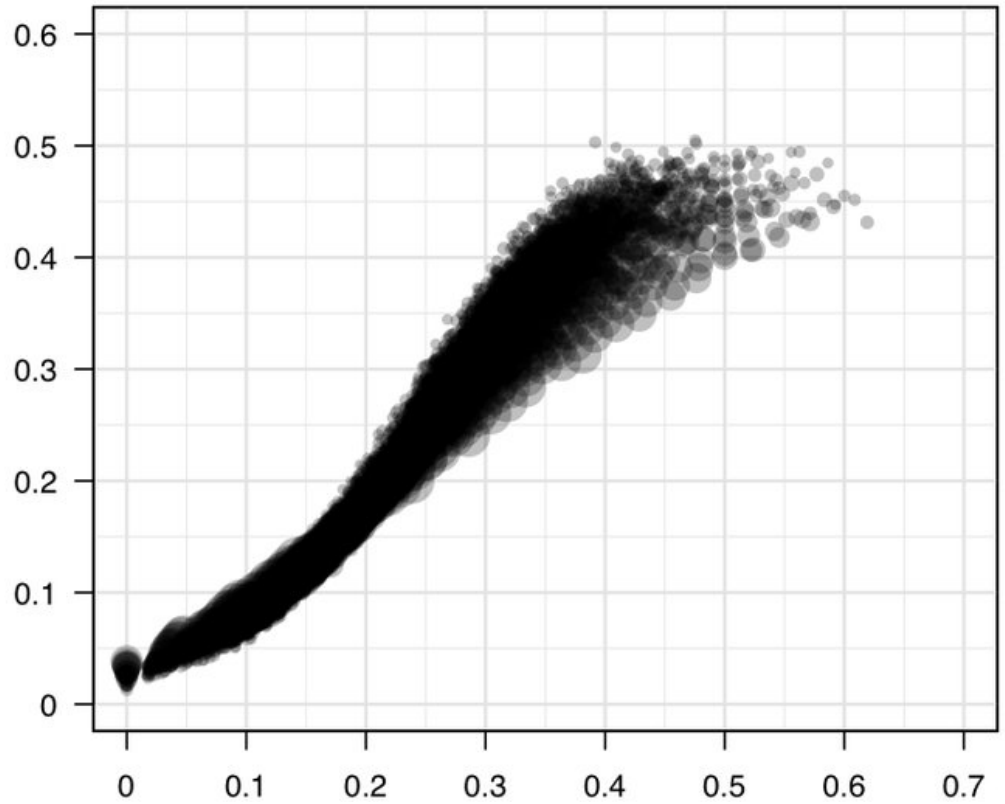
Bond, Robert M., et al. "A 61-million-person experiment in social influence and political mobilization." Nature 489.7415 (2012): 295-298.

# "I Voted" Button

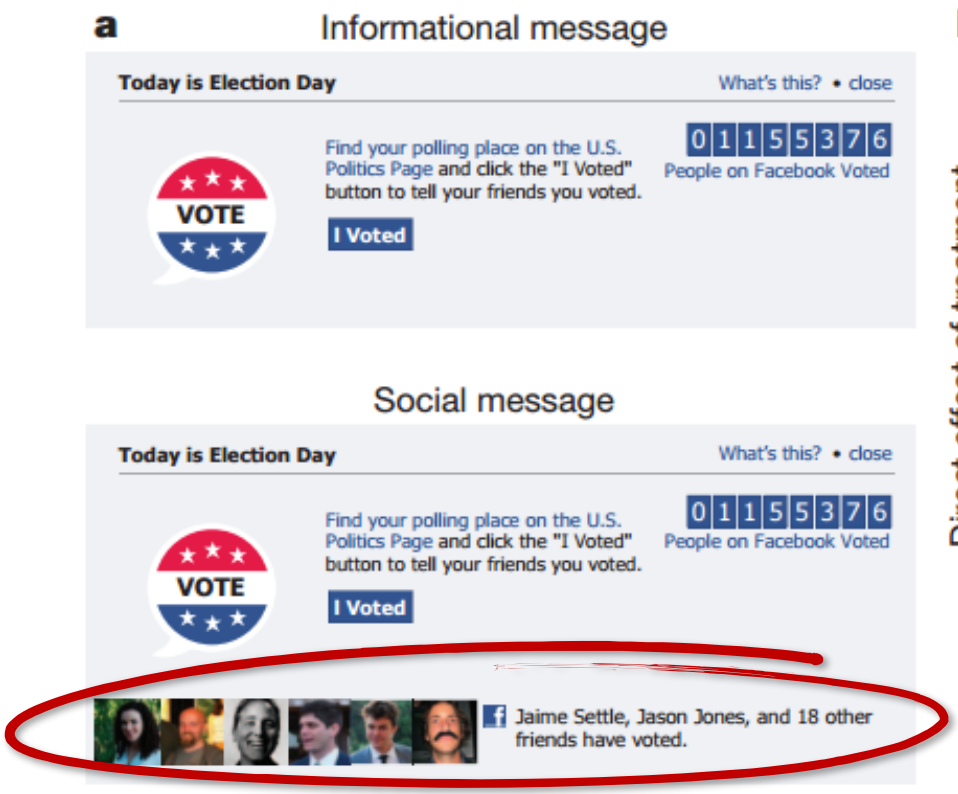
- Direct messages to 61 million users on FB
  - **Informational**: 1% users received
  - **Social**: 98% users received
  - **Control group**: 1% (no message received)

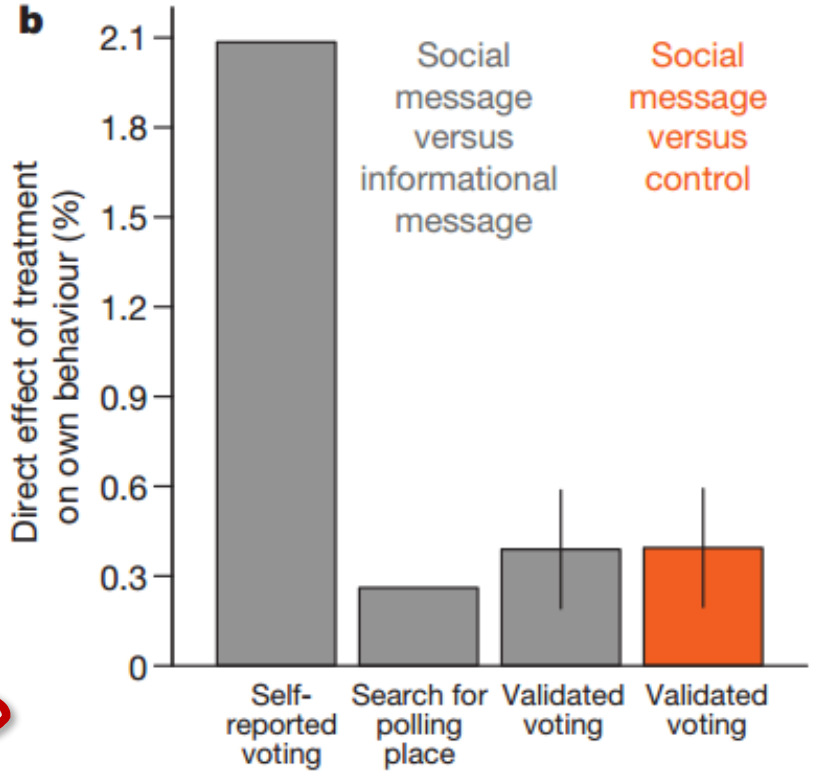
# Effect of Manipulation

Prob. of oneself claimed voted




Ratio of friends voted

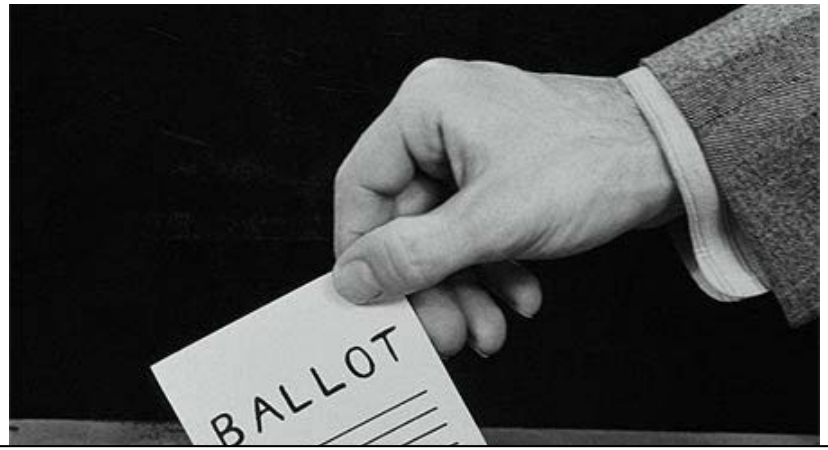



2% more likely to click "I voted" button and 0.3% more likely to seek information about a polling place, and **0.4%** more likely to head to the polls.

# Real-world Consequence (!)

- In total there were about 60,000 votes of turnout, and estimated 280,000 indirect turnout (out of 61 million users)




**What if Facebook did not randomize the control/experimental groups?**

# APPROACHES

#1 MACROSCOPE

#2 VIRTUAL LAB

**#3 EMPIRICAL MODELING**

# Empirical Modeling

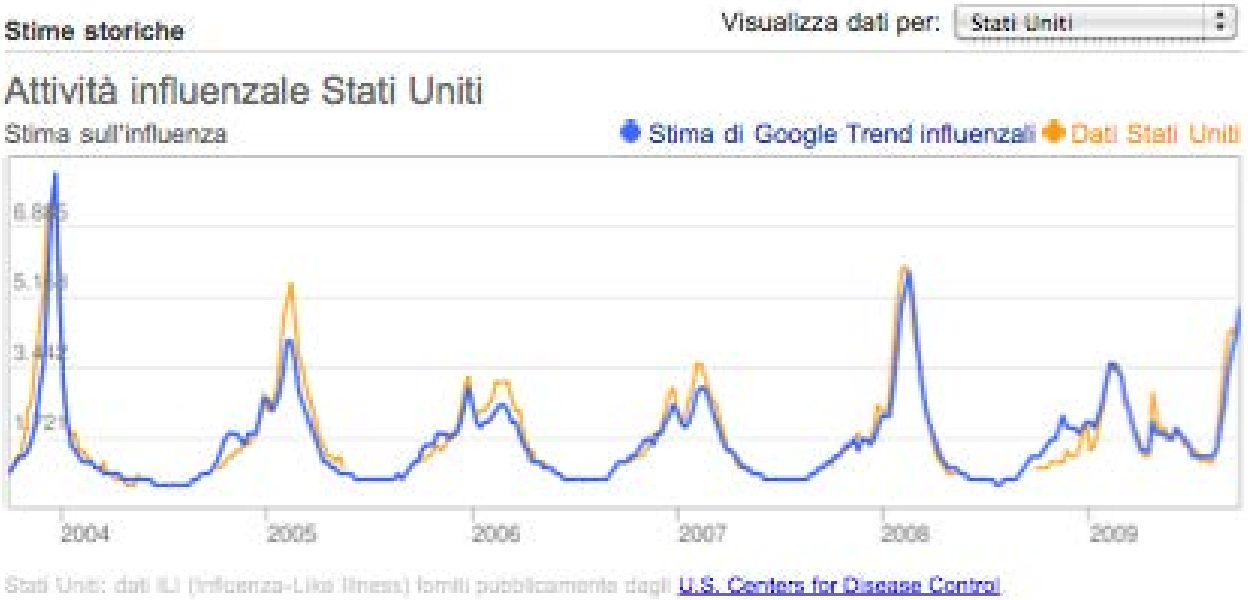
- Traditional mathematical or computational modeling
  - Tends to rely on many, often unrealistic, assumptions
  - Not generally tested in detail against data
  - Result is proliferation of models that exist in parallel and are often incompatible with each other

- New sources/scales of data allow both to learn/test models and also calibrate them
  - Observations → Models → Lab → Field → Observations

**PREDICTION IS VERY DIFFICULT, ESPECIALLY ABOUT THE FUTURE.**

*Niels Bohr*

# Google Flu Trends




Google

## Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]Google Inc. [2]Centers for Disease Control and Prevention

Nature 457, 1012-1014 (2009)

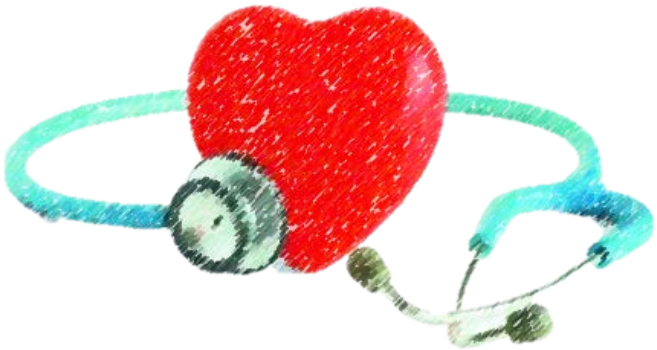
Medicine and Linguistics

# PREDICTION OF COUNTY-LEVEL HEART DISEASE MORTALITY

Eichstaedt, Johannes C., et al. "Psychological language on twitter predicts county-level heart disease mortality." Psychological science 26.2 (2015): 159-169.

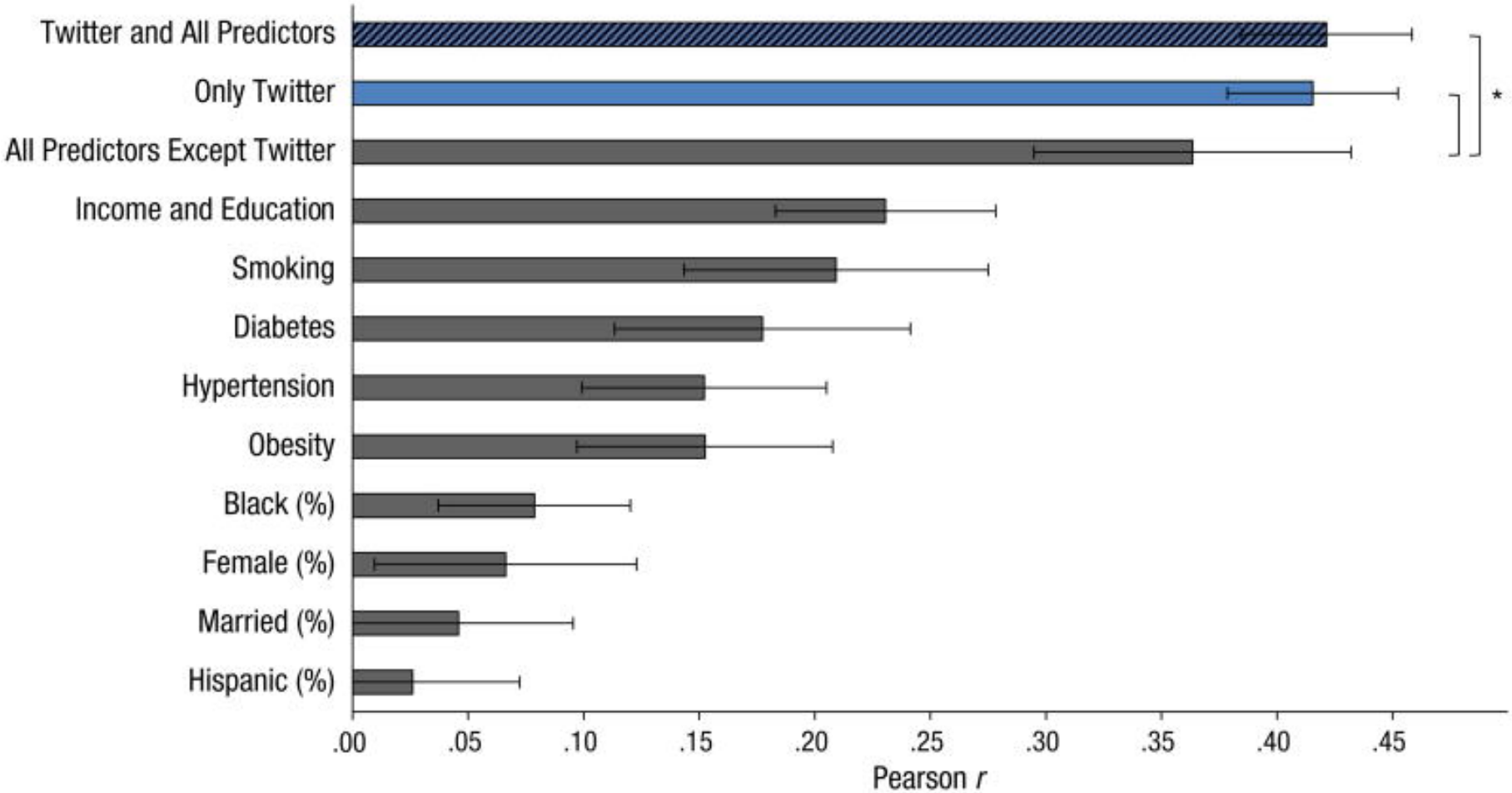
# Datsets

- Heart disease
  - Arteriosclerotic heart disease mortality rates during 2009 -- 2010

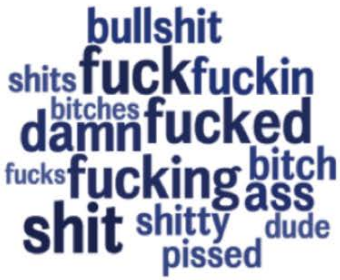


- Predictors
  - 826 million tweets collected between June 2009 and March 2010
  - Socioeconomic (income and education)
  - Demographic (percentages of Black, Hispanic, married, and female residents)
  - Health status (diabetes, obesity, smoking, and hypertension)

# Prediction Accuracy



Hostility, Aggression




$r = .18$

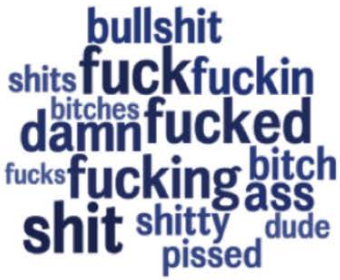
Hate, Interpersonal Tension

$r = .16$

Boredom, Fatigue

$r = .18$

Hostility, Aggression




$r = .18$

Skilled Occupations

$r = -.14$

Hate, Interpersonal Tension

$r = .16$

Positive Experiences

$r = -.14$

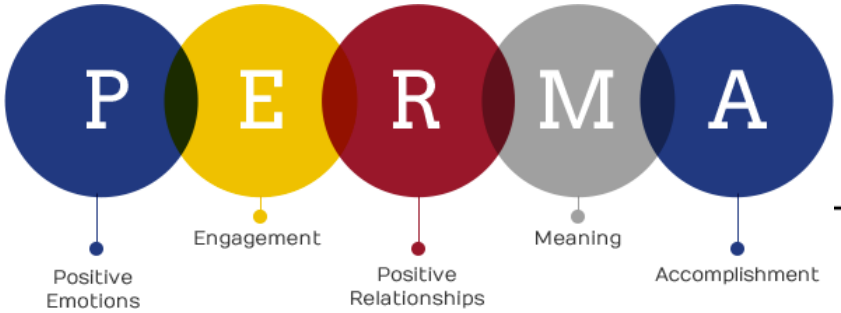
Boredom, Fatigue

$r = .18$

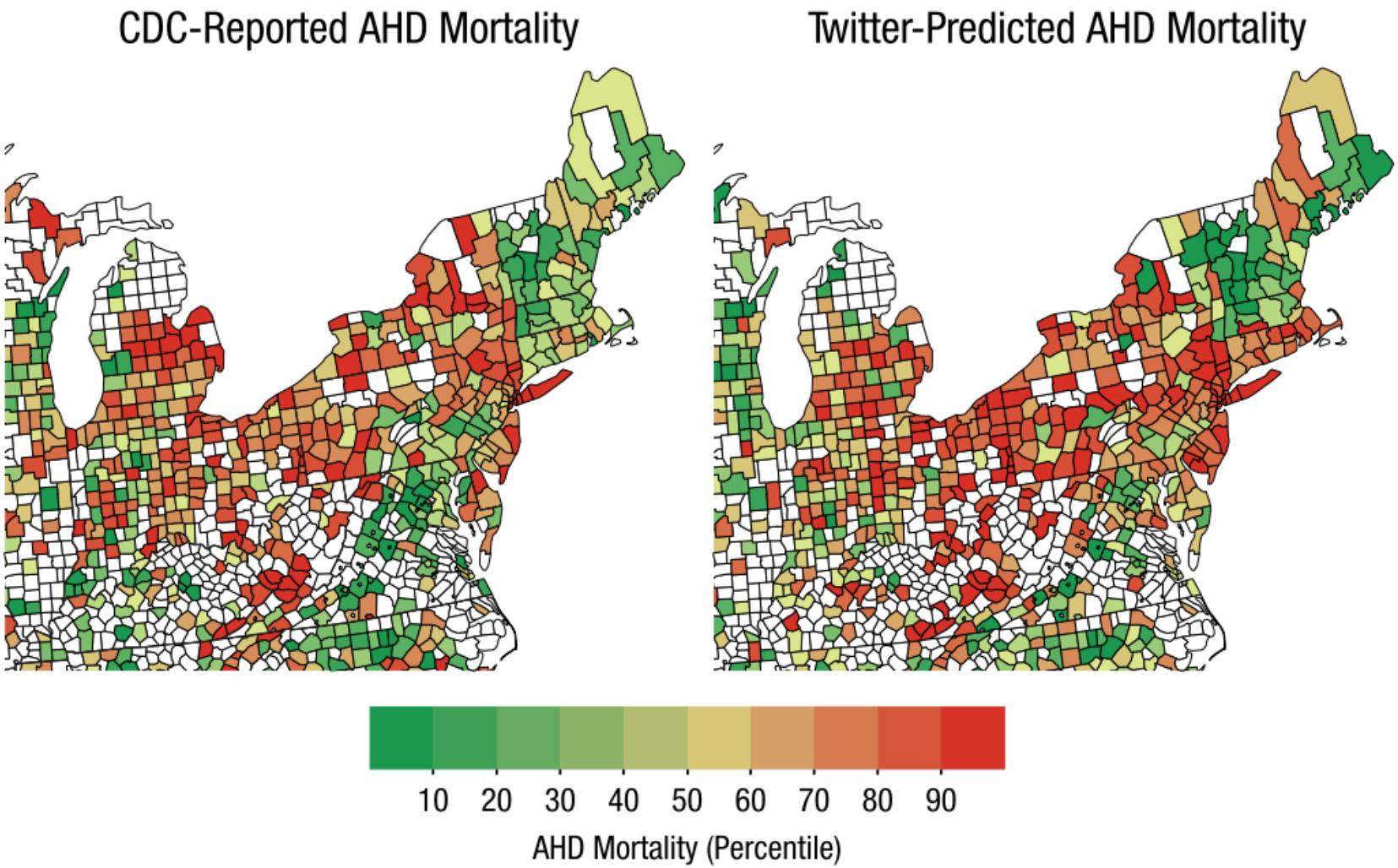
Optimism

$r = -.12$

# Language Use in Tweets




| | Correlation with AHD mortality |
|---|---|
| Risk factors | |
| Anger | .17 [.11, .22]** |
| Negative relationships | .16 [.11, .21]** |
| Negative emotions | .10 [.05, .16]** |
| Disengagement | .14 [.08, .19]** |
| Anxiety | .05 [.00, .11]† |
| Protective factors | |
| Positive relationships[a] | .02 [−.04, .07] |
| Positive emotions | −.11 [−.17, −.06]** |
| Engagement | −.16 [−.21, −.10]** |




**Social media opens up a new window of what humans actually feel and think**

Empirical Modeling

Social Psychology

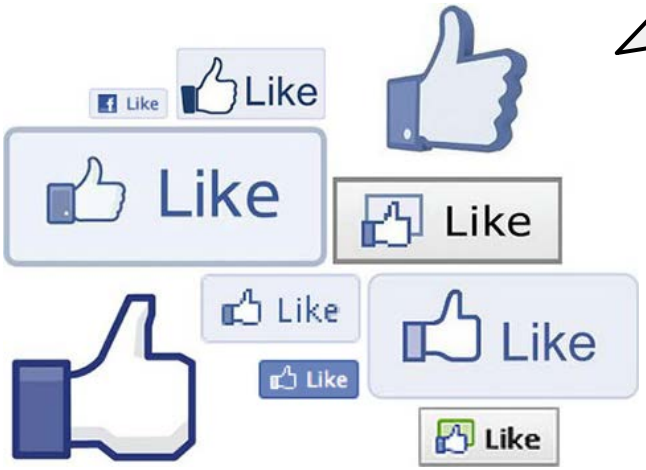
# YOU ARE WHAT YOU LIKE

Kosinski, Michal, David Stillwell, and Thore Graepel. "Private traits and attributes are predictable from digital records of human behavior." Proceedings of the National Academy of Sciences 110.15 (2013): 5802-5805.

# Personality Prediction

## Personality traits

- Gender, age, relationship status, # friends
- Sexual orientation, ethnicity, religion, political inclination
- Addictive substances (alcohol, drugs, cigarette), parental separation
- IQ, 5-Factor model, satisfaction with Life



# Data Collection

- 9,939,220 Likes (55,814 unique ones) from 58,466 Facebook volunteers
  - Sports
  - Music
  - Books
  - Restaurants
  - Popular websites

# Ground truth

- Political Inclination

| Democrat | Republican |
|---|---|
| Democratic | GOP (Grand Old Party) |
| Democratic Party | Republican Party |

- Sexual Orientation

| Homosexual | Heterosexual |
|---|---|
| 1 / 0 | 1 / 0 |

# Ground truth

- 5-Factor Model
  - Openness
  - Conscientiousness
  - Extraversion
  - Agreeableness
  - Stability

This is an interactive version of the IPIP Big-Five Factor Markers.

| | Disagree | | Neutral | | Agree |
|---|---|---|---|---|---|
| I feel little concern for others. | ○ | ○ | ○ | ○ | ○ |
| I am always prepared. | ○ | ○ | ○ | ○ | ○ |
| I get stressed out easily. | ○ | ○ | ○ | ○ | ○ |
| I have a rich vocabulary. | ○ | ○ | ○ | ○ | ○ |
| I don't talk a lot. | ○ | ○ | ○ | ○ | ○ |
| I am interested in people. | ○ | ○ | ○ | ○ | ○ |
| I leave my belongings around. | ○ | ○ | ○ | ○ | ○ |
| I am relaxed most of the time. | ○ | ○ | ○ | ○ | ○ |
| I have difficulty understanding abstract ideas. | ○ | ○ | ○ | ○ | ○ |
| I feel comfortable around people. | ○ | ○ | ○ | ○ | ○ |
| I pay attention to details. | ○ | ○ | ○ | ○ | ○ |
| I worry about things. | ○ | ○ | ○ | ○ | ○ |

# Ground truth

- Satisfaction with Life (SWL)

Below are five statements that you may agree or disagree with. Using the 1 - 7 scale below, indicate your agreement with each item by placing the appropriate number on the line preceding that item. Please be open and honest in your responding.

- 7 - Strongly agree
- 6 - Agree
- 5 - Slightly agree
- 4 - Neither agree nor disagree
- 3 - Slightly disagree
- 2 - Disagree
- 1 - Strongly disagree

____ In most ways my life is close to my ideal.

____ The conditions of my life are excellent.

____ I am satisfied with my life.

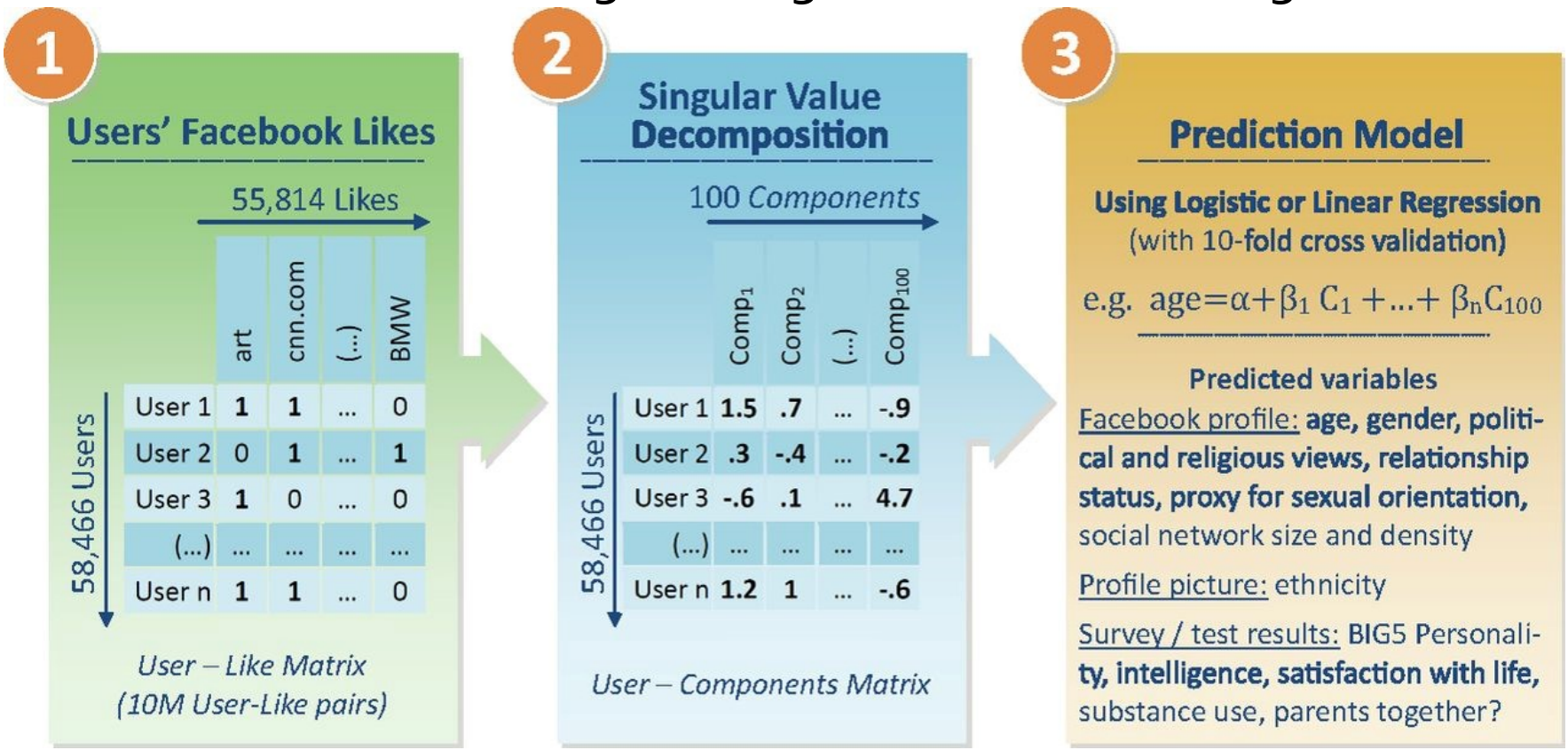
____ So far I have gotten the important things I want in life.

____ If I could live my life over, I would change almost nothing.

- 31 - 35 Extremely satisfied
- 26 - 30 Satisfied

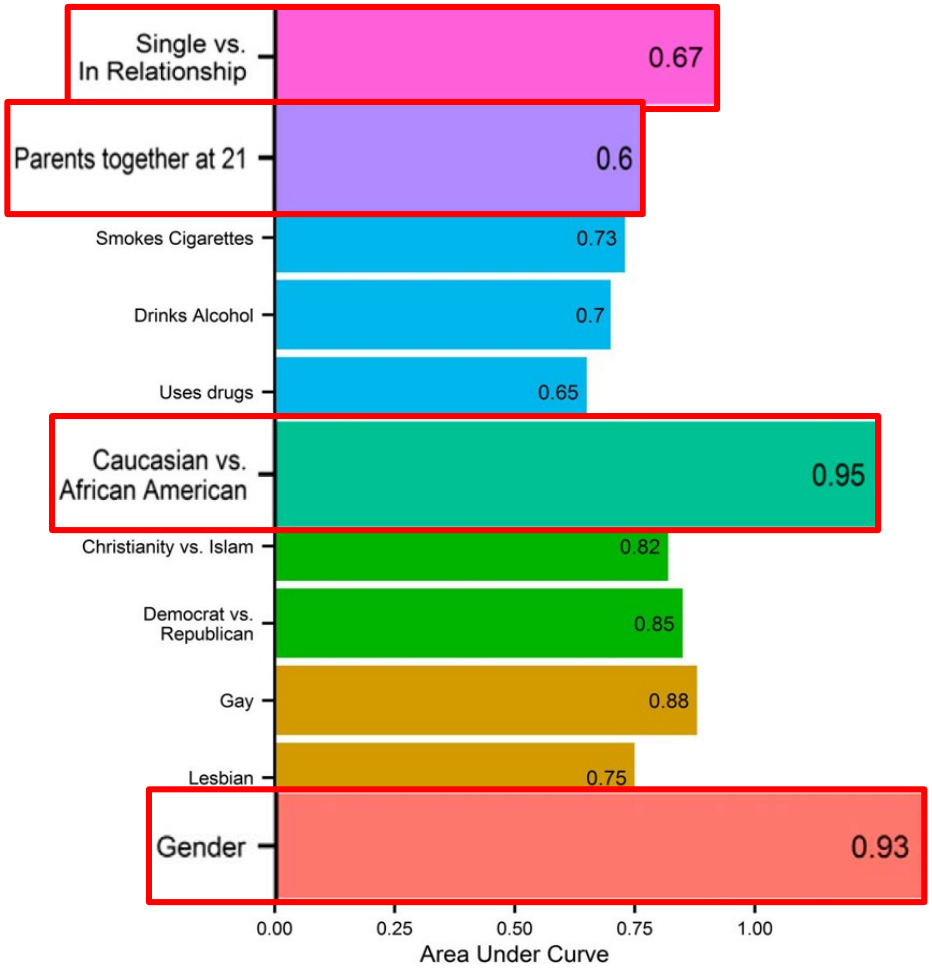
# Methodology

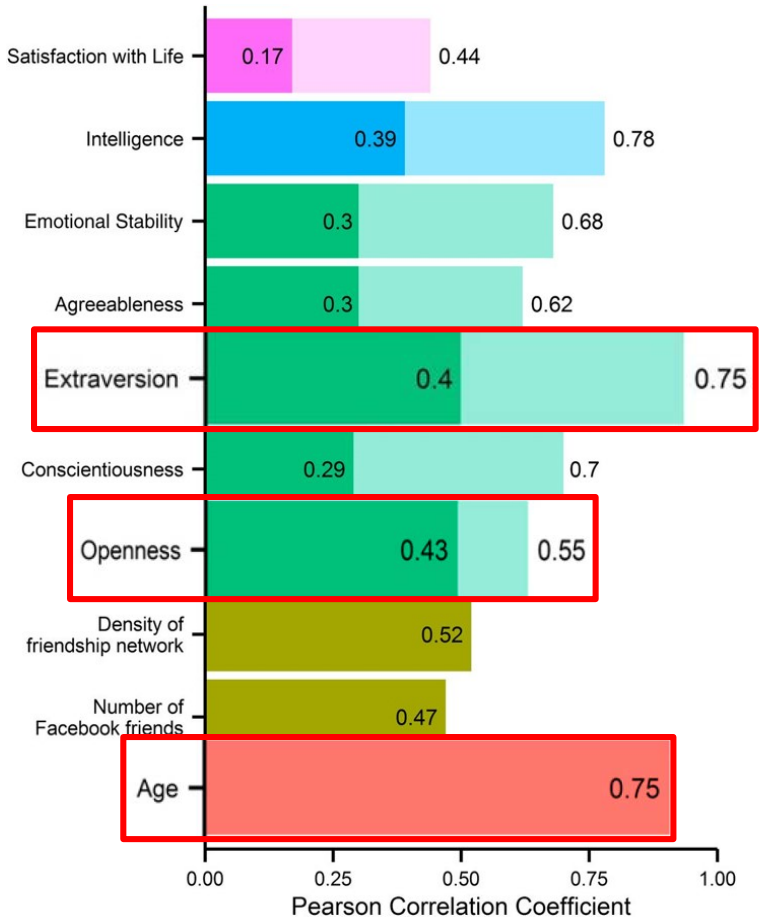
- User-Like matrix dimension reduction: Singular Value Decomposition (SVD)
- Prediction models: Logistic Regression & Linear Regression



# Prediction Results

Solid: Pearson corr. coef. between pred. & actual values
Transparent: baseline acc. of the questionnaire, in terms of test-retest reliability



# Discriminative Likes (#1)

| | | | | |
|---|---|---|---|---|
| **Gender** | ***Female*** | Tv Fanatic | Modern Warfare 2 | ***Male*** |
| | | Chiq | ESPN | |
| | | Gillette Venus | Sportscenter | |
| | | Shoedazzle | Band Of Brothers | |
| | | Bebe | Starcraft | |
| | | Proud To Be A Mom | Deadliest Warrior | |
| | | Covergirl | Dos Equis | |
| | | Wet Seal | Red Vs Blue | |
| | | Aerie By American Eagle | X Games | |
| | | Mall World | Bruce Lee | |

# Discriminative Likes (#2)

| | | | | |
|---|---|---|---|---|
| **Friends** | ***Many*** | Mojo-Jojo<br>Biology<br>Dollar General<br>Hillary<br>106 & Park<br>Jennifer Lopez<br>Paid In Full<br>Yo Gotti<br>The Dollar You Are Holding Could've Been In A Stripper's Butt Crack | The Dark Knight<br>In'n'out Burger<br>Hard Rock<br>Honey, Where Is My Supersuit<br>Hating ICP<br>Minecraft<br>Iron Maiden<br>Walking With Your Friend & Randomly Pushing Them Into Someone/Something | ***Few*** |

# Discriminative Likes (#3)

| | | | | |
|---|---|---|---|---|
| IQ | *High* | The Godfather | Jason Aldean | *Low* |
| | | Mozart | Tyler Perry | |
| | | Thunderstorms | Sephora | |
| | | The Colbert Report | Chiq | |
| | | Morgan Freemans Voice | Bret Michaels | |
| | | The Daily Show | Clark Griswold | |
| | | Lord Of The Rings | Bebe | |
| | | To Kill A Mockingbird | I Love Being A Mom | |
| | | Science | Harley Davidson | |
| | | Curly Fries | Lady Antebellum | |

# Li #1)

男 女

讚
打鉤
按收通知
顯示在動態消息中
設定
新增到興趣主題清單

卡提諾正妹抱報
Entertainment Website

Sign Up | Liked | Message | •••

Timeline | About | Photos | Likes | Videos

397,729 people like this
吳長鋼 and 55 other friends

Invite friends to like this Page

ABOUT

這裡是卡提諾論壇正妹區！ 這裡是正妹的海洋！爆紅人氣正妹、女明星寫真、氣質網路正妹都在這裡！有正妹~絕不私藏~!!http://goo.gl/mUZa9m

Ask for 卡提諾正妹抱報's website

PHOTOS

Post | Photo / Video

Write something...

Post

卡提諾正妹抱報
11 hrs ·

阿斯！約哥好猛喔

# Likes are Culture-Dependent (#2)

| 已婚 | | 未婚 |
|---|---|---|
| | 嬰兒與母親 懷孕生產情報站 | 學生愛打工 |
| | 味全 MyWei | Duncan |
| | Estee Lauder Taiwan 雅詩蘭黛 | 林俊傑 JJ Lin |
| | 光泉"HOT"鮮奶 | Cherng |
| | 舒潔溫柔心感動 | Byebyechuchu |
| | 綠巨人 | Dcard |
| | AVON Taiwan 雅芳粉絲團 | 田馥甄 Hebe |
| | 人人玩遊戲 | 彭于晏 Eddie Peng |
| | Creative Baby - 台灣 | Dorothy |

**Unprecedented opportunity to observe individuals in a society**

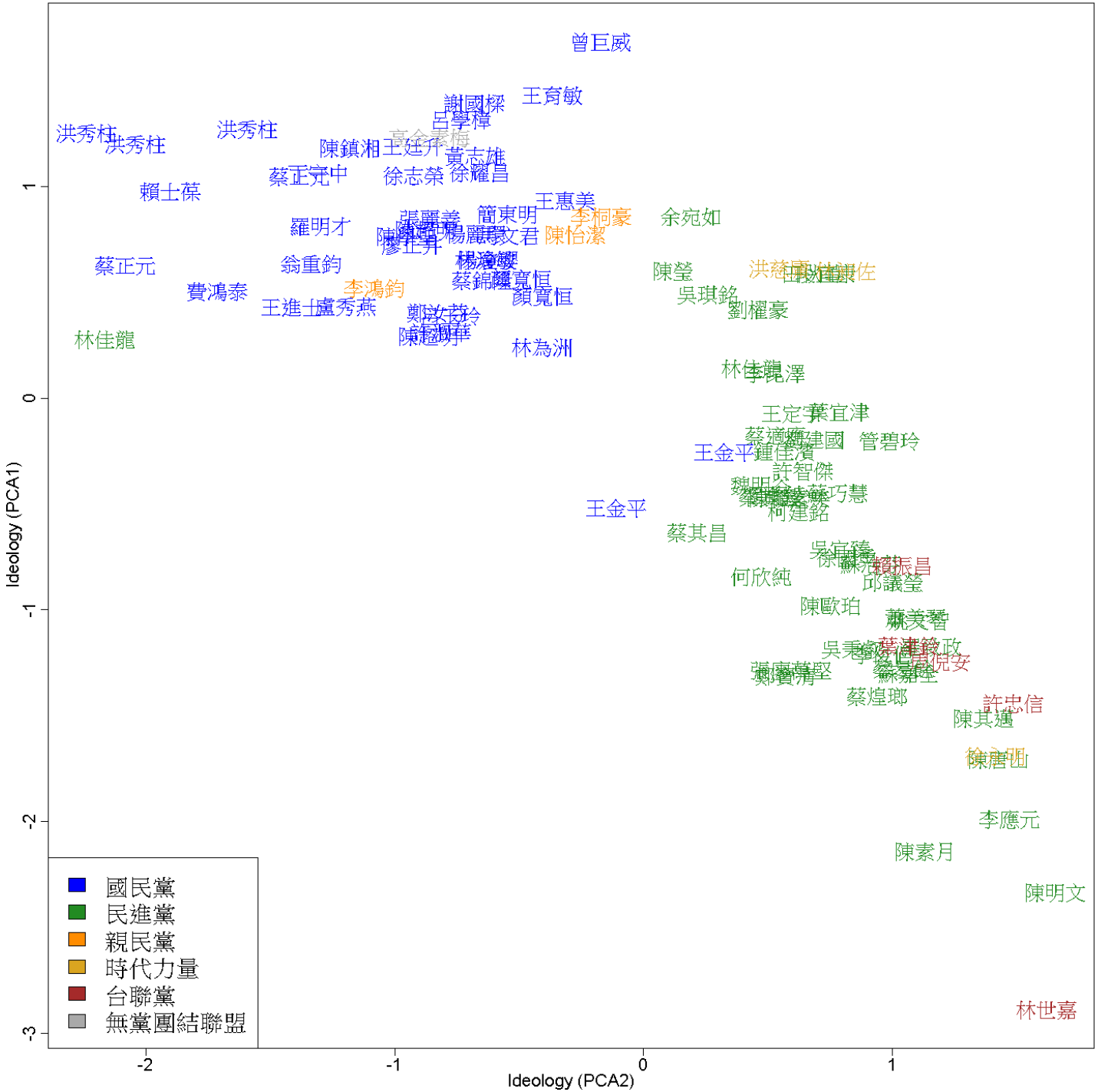

曾巨威
王育敏
謝國樑
呂學樟
洪秀柱
洪秀柱
洪秀柱
陳鎮湘
王廷升
黃志雄
蔡正元
徐志榮
徐耀昌
賴士葆
王惠美
簡東明
李桐豪
余宛如
羅明才
陳怡潔
蔡正元
翁重鈞
陳瑩
費鴻泰
李鴻鈞
蔡錦隆
顏寬恒
吳琪銘
劉櫂豪
王進士
盧秀燕
林佳龍
林為洲
林佳龍
李俊俋
王定宇
葉宜津
蔡適應
管碧玲
王金平
許智傑
王金平
蔡其昌
何欣純
邱議瑩
陳歐珀
蔡煌瑯
許忠信
陳其邁
陳唐山
李應元
陳素月
陳明文
林世嘉
國民黨
民進黨
親民黨
時代力量
台聯黨
無黨團結聯盟
Ideology (PCA1)
Ideology (PCA2)
1
0
-1
-2
-3
-2
-1
0
1

**Table 3.** Top Polarized Pages in Community Category

| Label | Page name | Ideology | Politicalness |
|---|---|---|---|
| CB1 | 熊熊想到不對勁 | -1.46 | 0.87 |
| CB2 | 挺洪秀柱參選中國國民黨黨主席 | -1.43 | 0.80 |
| CB3 | 正義無敵・力挺馬英九 | -1.37 | 0.86 |
| CB4 | 馬英九總統後援會 | -1.33 | 0.89 |
| CB5 | 中華民國-馬英九網路後援團 | -1.33 | 0.85 |
| CG5 | 無限期支持戰神國昌老師 | 0.81 | 0.89 |
| CG4 | 臺澎國際法法理建國連線 | 0.81 | 0.80 |
| CG3 | 台大法律學生挺 318 | 0.82 | 0.83 |
| CG2 | 捍衛台灣文史青年組合 | 0.82 | 0.85 |
| CG1 | 譴責方仰寧及其長官違憲剝奪公投盟集會權利 | 0.83 | 0.90 |

**Table 4.** Top Polarized Pages in Musician/Band Category

| Label | Page name | Ideology | Politicalness |
|---|---|---|---|
| MB1 | Taipei Wind Orchestra & Symphonic Band | -0.10 | 0.53 |
| MB2 | 臺北市立國樂團 | -0.08 | 0.65 |
| MB3 | 周華健 | -0.08 | 0.42 |
| MB4 | 金佩姍 | -0.07 | 0.60 |
| MB5 | 台北香頌室內樂團 | -0.06 | 0.55 |
| MG5 | 2HRs | 0.74 | 0.52 |
| MG4 | 農村武裝青年 | 0.74 | 0.73 |
| MG3 | 拍謝少年 | 0.75 | 0.60 |
| MG2 | 淡水男孩-Tamsui BoyZ(TBZ) | 0.78 | 0.85 |
| MG1 | 濁水溪公社 LTK Commune | 0.79 | 0.73 |

**Table 5.** Top Polarized Publishers

| Label | Page name | Ideology | Politicalness |
|---|---|---|---|
| PB1 | 道聲出版社 | 0.05 | 0.37 |
| PB2 | 大田出版 | 0.08 | 0.49 |
| PB3 | 四也童書出版公司 | 0.08 | 0.70 |
| PB4 | 小熊出版社 LittleBearBooks | 0.13 | 0.53 |
| PG4 | 唐山書店／唐山出版社 | 0.66 | 0.58 |
| PG3 | 衛城出版 | 0.74 | 0.60 |
| PG2 | 玉山社出版公司 | 0.78 | 0.75 |
| PG1 | 前衛出版社 | 0.81 | 0.80 |

**Table 6.** The Ideology and Politicalness of Mass Media in Taiwan

| Label | Page name | Ideology | Politicalness |
|---|---|---|---|
| N1 | 中時電子報【即時新聞】頻道 | -0.46 | 0.68 |
| N2 | 中視新聞 | -0.34 | 0.53 |
| N3 | 大愛全球新聞 | -0.11 | 0.38 |
| N4 | 聯合晚報 | -0.07 | 0.57 |
| N5 | GOOD TV 好消息電視台 | -0.01 | 0.38 |
| N6 | 自由時報 | 0.21 | 0.37 |
| N7 | 大紀元時報 - 台灣 (The Epoch Times - Taiwan) | 0.23 | 0.58 |
| N8 | 新唐人亞太電視台 | 0.30 | 0.78 |
| N9 | 年代新聞 | 0.31 | 0.58 |
| N10 | 三立新聞 | 0.37 | 0.60 |

**Table 7.** Largest Ideology Difference between Pages and The Mentioned Politicians

| Page name | Mentioned | Ideo. diff. | Type |
|---|---|---|---|
| 賴清德有種不進議會，沒種不進市府！ | 賴清德 | -1.84 | antagonistic |
| 驅逐柯建銘離開國會聯盟 | 柯建銘 | -1.82 | antagonistic |
| 大家一起去踹邱議瑩的門，表達司法不公的憤怒 | 邱議瑩 | -1.79 | antagonistic |
| 無限期等待賴清德市長遵守政治承諾！政治要誠信不可耍無賴！ | 賴清德 | -1.66 | antagonistic |
| 蔡英文炒地權貴、民進黨貪汙收賄媚日爛黨 | 蔡英文 | -1.65 | antagonistic |
| 柯文哲有病 | 柯文哲 | -1.64 | antagonistic |
| 無限期支持連勝文參選總統 (Support Sean Lien for President Forever) | 連勝文 | 1.63 | sarcastic |
| 胡志強的芭樂園 (The Guava Farm[6] of Hu) | 胡志強 | 1.62 | sarcastic |
| 終結國民黨！馬英九下台！還給台灣公平正義！ | 馬英九 | 1.61 | antagonistic |
| 我愛連勝文 (I Love Sean Lien) | 連勝文 | 1.60 | sarcastic |

# CONCLUSION & OUTLOOK

WE ARE STILL AT
THE VERY START
START

# LOTS of Big Questions

- The polarization of global economic inequality
- What explains the success of social movements?
- The emergence of pro-sociality behavior
- The causality of video gaming and propensity of violence?
- The politics of censorship
- The causality of social selection and social influence?
- …

# The Data Divide

- Social scientists have **good questions** but…
  - IT tools are not part of their toolkits
  - Not clear that we will/should make the investment

- Computer scientists have **powerful methods** but…
  - Trained to resolve technical problems
  - It seems there are less "methodological" contributions

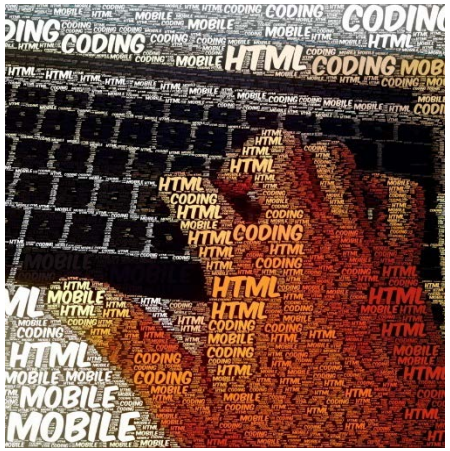
# The Challenges

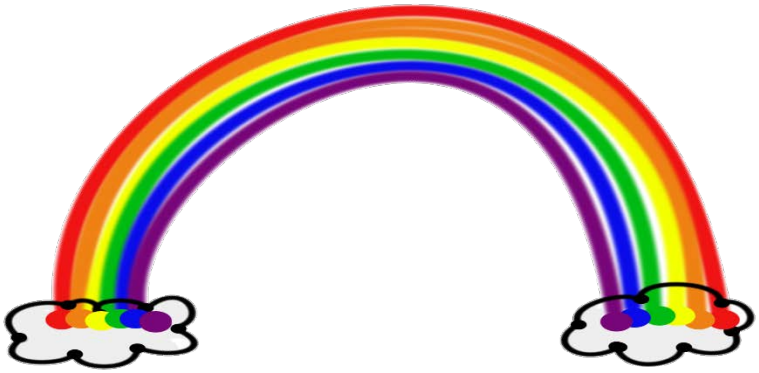
- Education and habits of social and computer scientists
  - Different ways of thinking
  - Different methodologies
  - Differences in framing questions and defining contributions
- Data access and **fragmentation** issue
- Data privacy issue
- Ethics issue
- Organizational issue

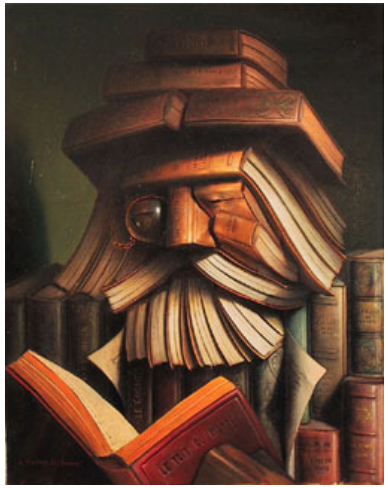
# Institutional Innovations

- New platforms and protocols for data management
  - Better coordination of data collection, storage, sharing
  - Recruitment and management of subject pools, field panels
- Integrated research designs
  - Coordination across theoretical, experimental and observational studies
- Collaborative interdisciplinary teams
  - For a given data set, often unclear what the most interesting question is
  - For a given question, often unclear how to collect the right data

# Computational Social Science



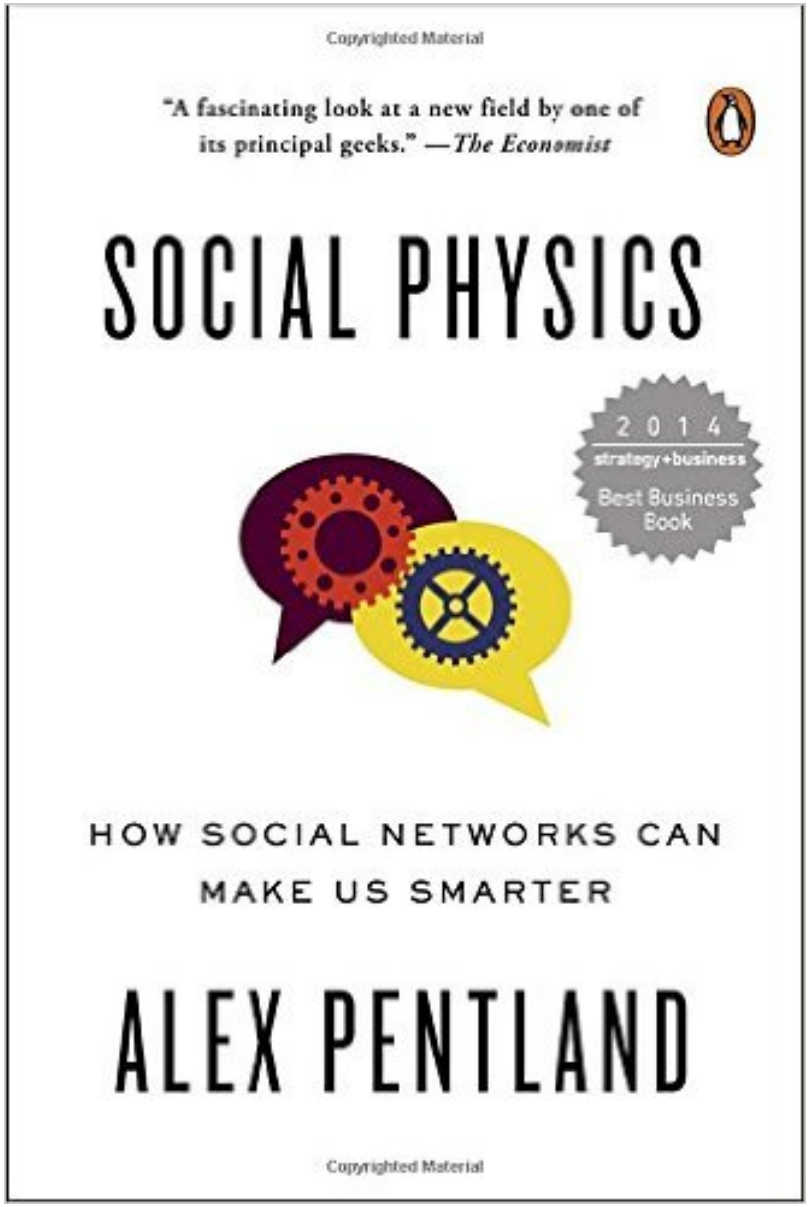


Techniques to collect, manage, and process large datasets

Knowledge about social theories, methods, and issues

# 交流時間

# 社會物理學


Copyrighted Material
"A fascinating look at a new field by one of its principal geeks." —The Economist
SOCIAL PHYSICS
2014
strategy+business
Best Business Book
HOW SOCIAL NETWORKS CAN MAKE US SMARTER
ALEX PENTLAND
Copyrighted Material

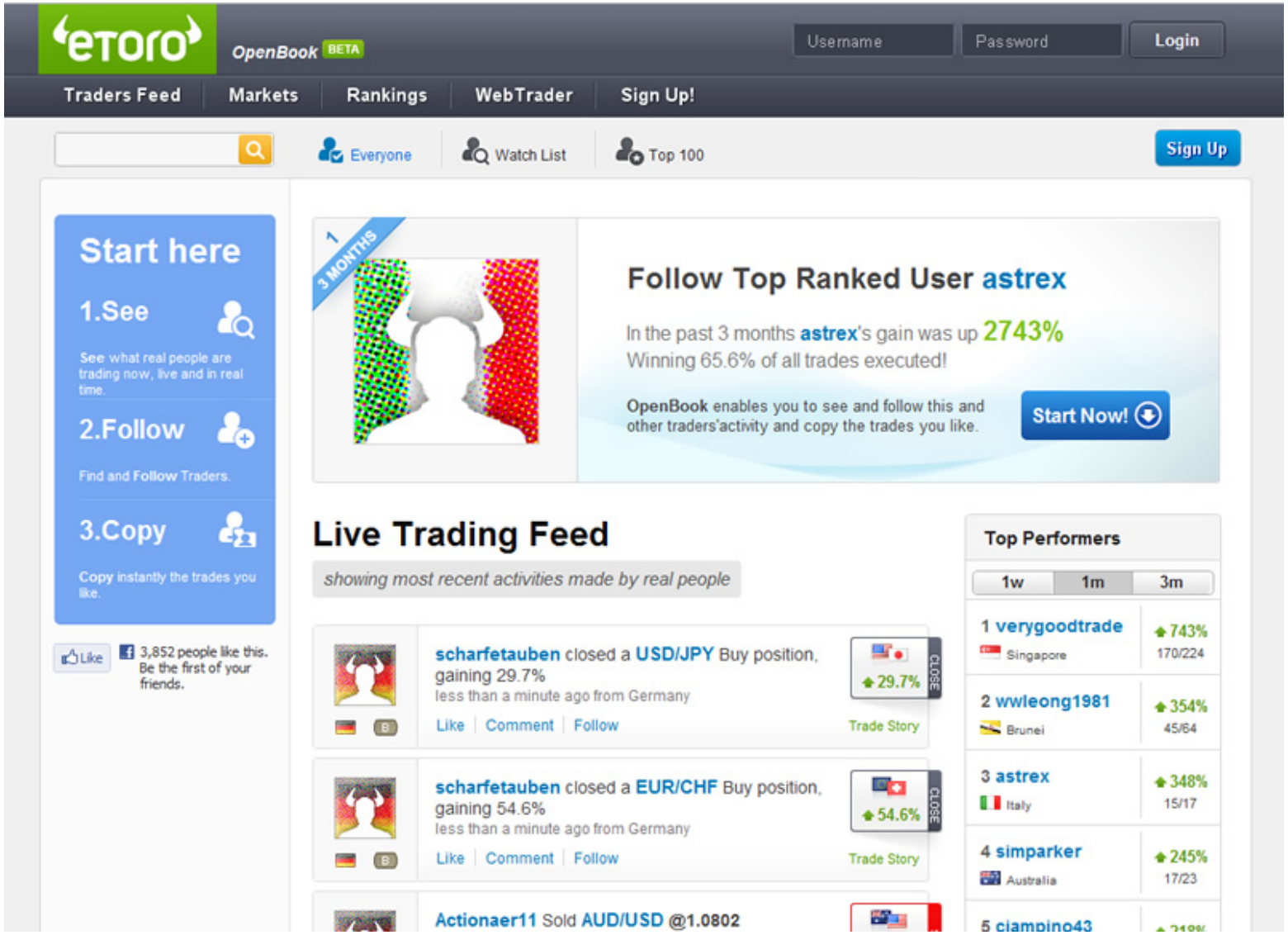
# 社會物理學

- 現實探勘 (reality mining) - 以巨量行為資料來解釋社會行為的新科學

- 不僅是複雜數學與量化預測，更是現實情境下可應用的實踐科學
  - 社會學習 (social learning)
  - 社群網絡中的意念流 (idea flow)

Alex “Sandy” Pentland

# eToro + OpenBook



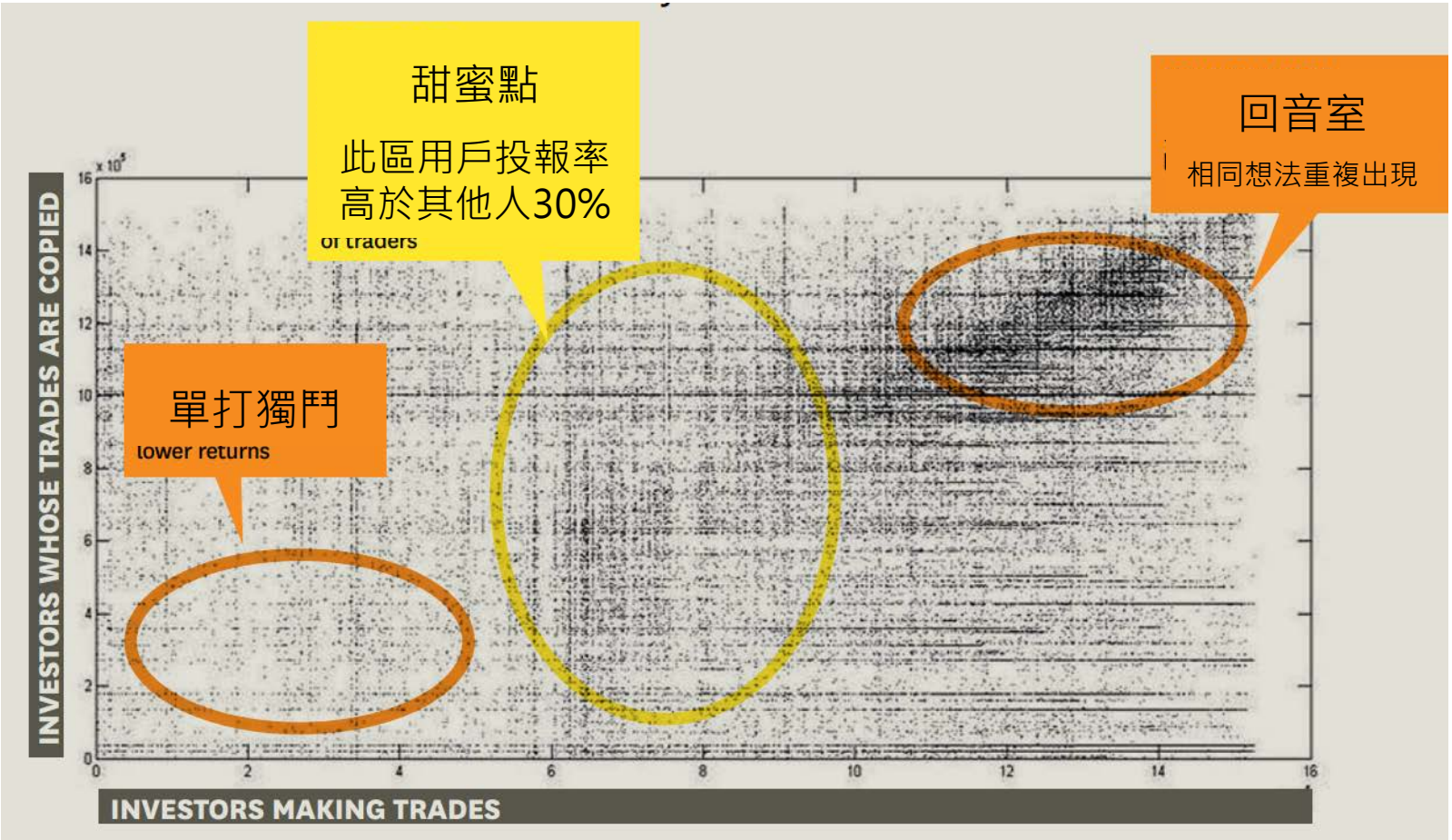

www.tradermaker.com/wp-content/themes/tradermarket/images/reviewscreens/etoro-openbook.jpg

# eToro + OpenBook

- 用戶可以查看 / 模仿其他用戶的交易、投資組合和績效紀錄，但不能看到其他用戶模仿誰的交易

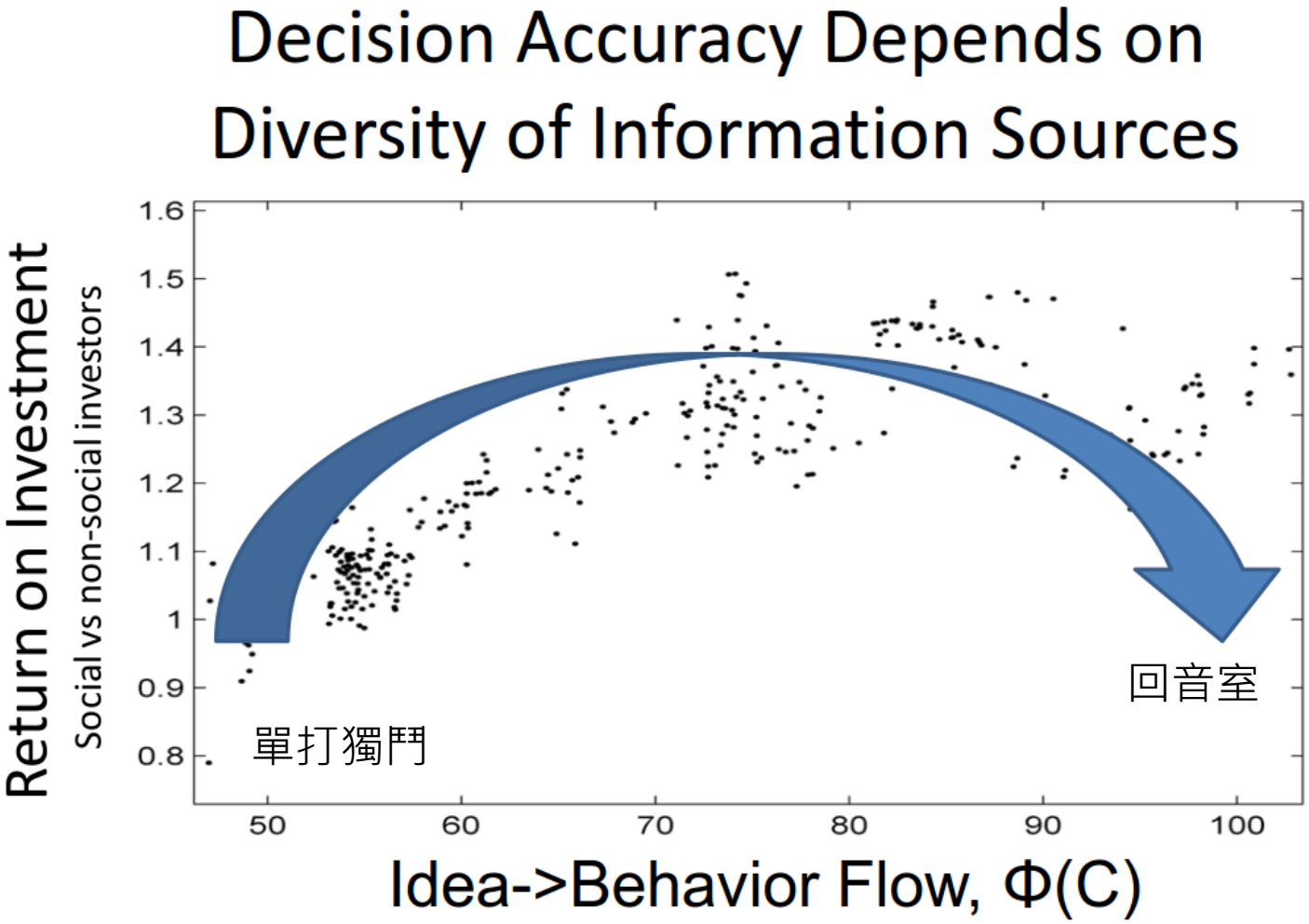
- 投資效益分析
  - 收集 2011 年裡 160 萬名用戶、近 1000 萬筆的美元 / 歐元交易行為資料

# 社會學習的證據




hbr.org/2012/04/the-new-science-of-building-great-teams

# Idea Flow vs. RoI




Copyright alex pentland 2012

sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_082159.pdf

Altshuler,Pentland

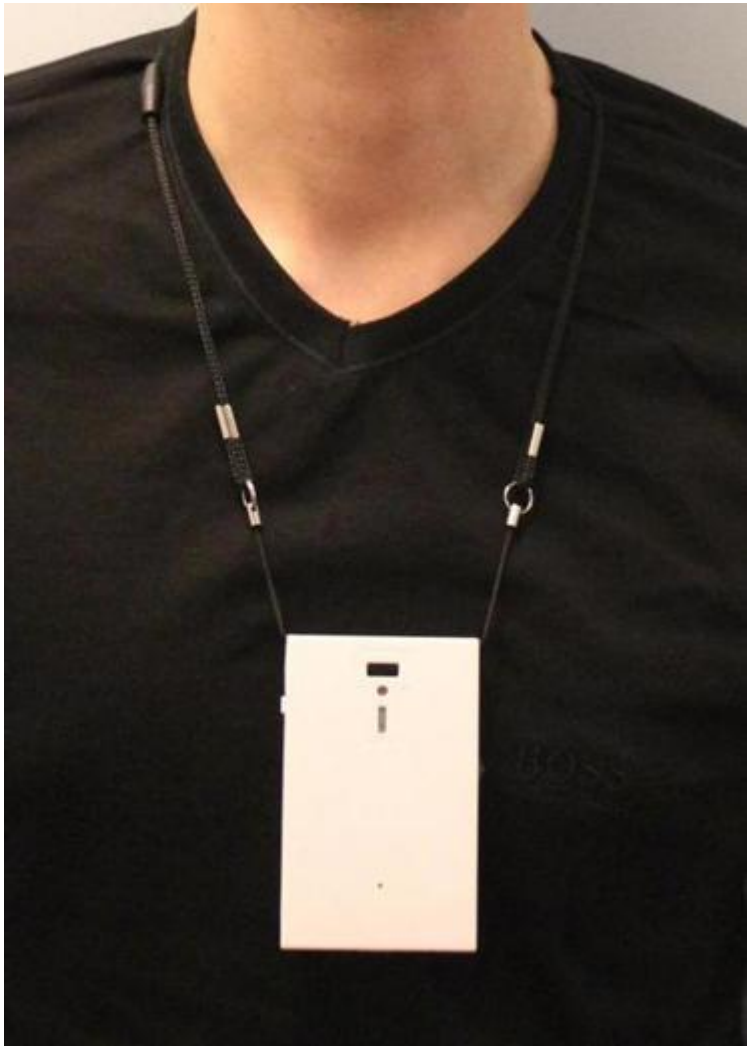
# 量化群體智慧

- 為什麼有些企業比其它企業來得有開創性？

- 決定群體表現的因素
  - 專業能力？
  - 凝聚力？
  - 成就感？
  - 薪水？
  - 領導者風格？
  - 文化？

# 社會計量識別牌 (Sociometer)

- 與誰互動以及互動行為
  - 談話語氣
  - 是否面對面 (距離)
  - 手勢多寡
  - 交談時聆聽和 (被) 打斷頻率
- 「對話輪替」的均等程度



www.bostonglobe.com/business/2013/11/02/breakthrough-management-tool-big-brother-workplace/WKMDFFieBC9M98EWUPbFZL/story.html

# 伺服器銷售公司

- 為期 1 個月，23 人，約 1,900 小時的互動觀察

- 客製化訂單任務派工：紀錄任務開始和結束的確切時間 → 衡量每名業務助理每項任務的確切花費時間

- 參與程度排名前 1/3 的員工

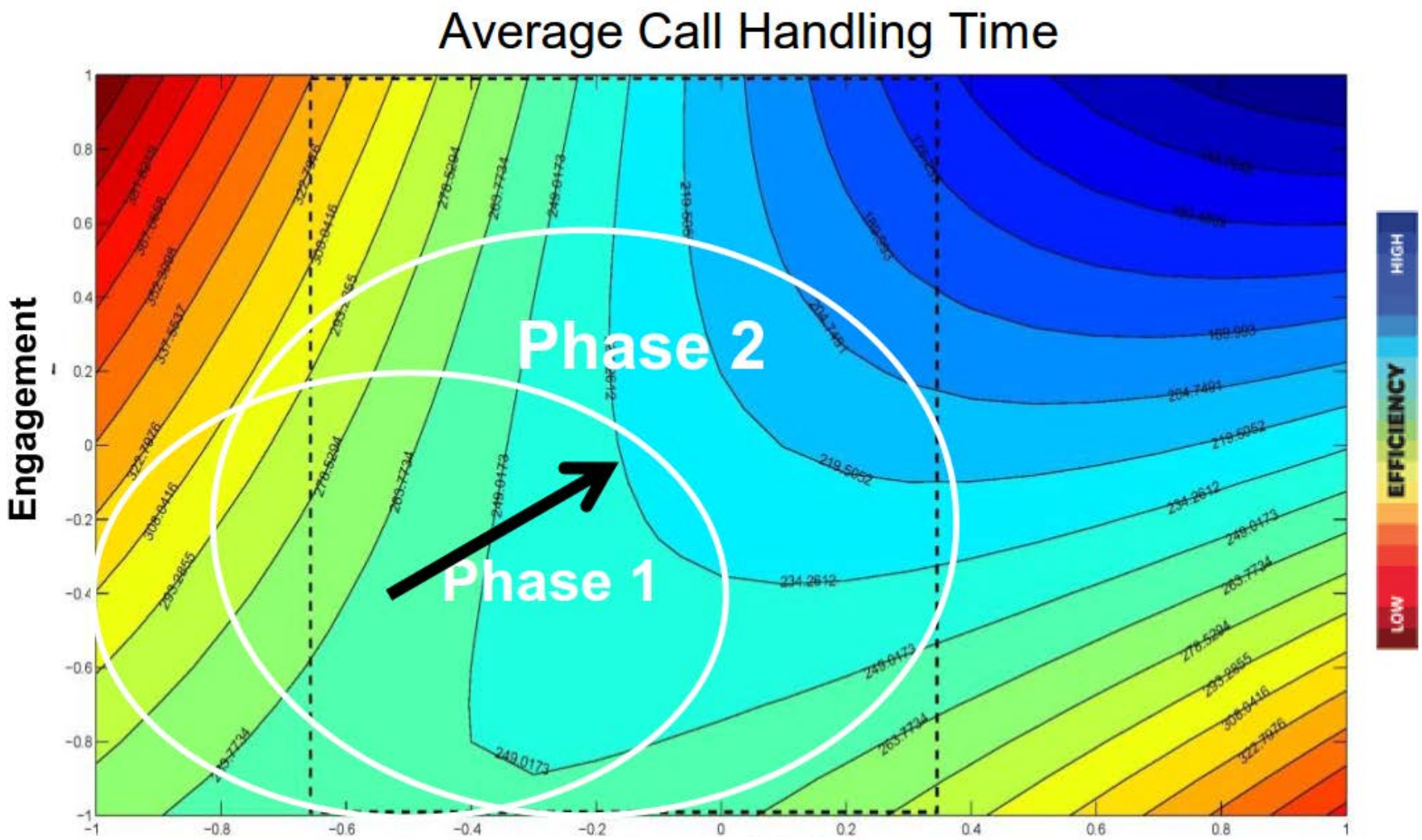
➔ 生產力較一般員工高出 10%

# Bank of America 電話客服中心

- 為期 6 週，每組 20 人，共 4 組的客服人員行為資料
- 效率指標 - 個案的平均處理時間
  - 若降低平均處理時間 5 % → 每年節省 USD $1M

- 從 idea flow 角度來改善
  - 客服輪流休息改為團隊輪值 ➔增加客服之間的互動和參與
  - 提昇 30% 參與程度 ➔ 平均效率提升 8% (20% for the previously worst case)
  - 估計有 USD $15M 效益 (given 3,000 位客服人員)

# 調整輪休時間後工作效率提昇



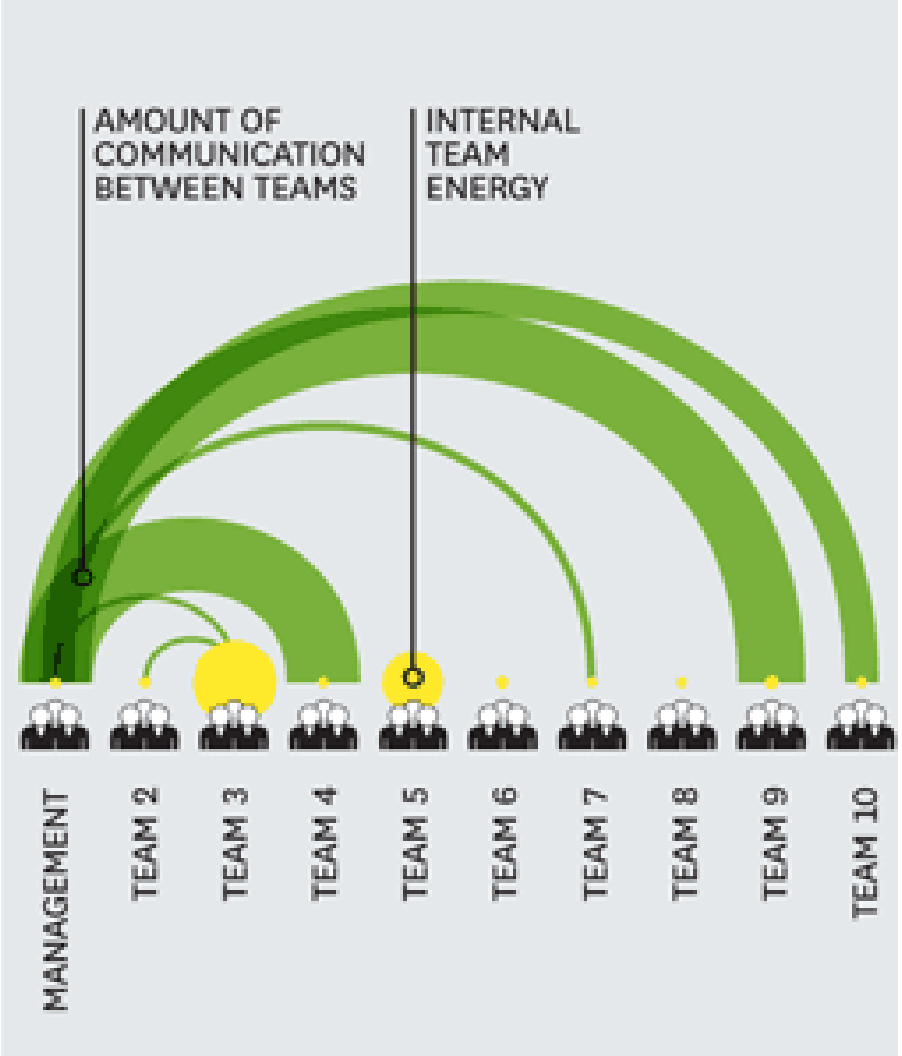

pyright alex pentland 2012 all rights res

Olguin, Waber, Kim, Pentland

sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_082159.pdf

# 量化團隊參與及探索

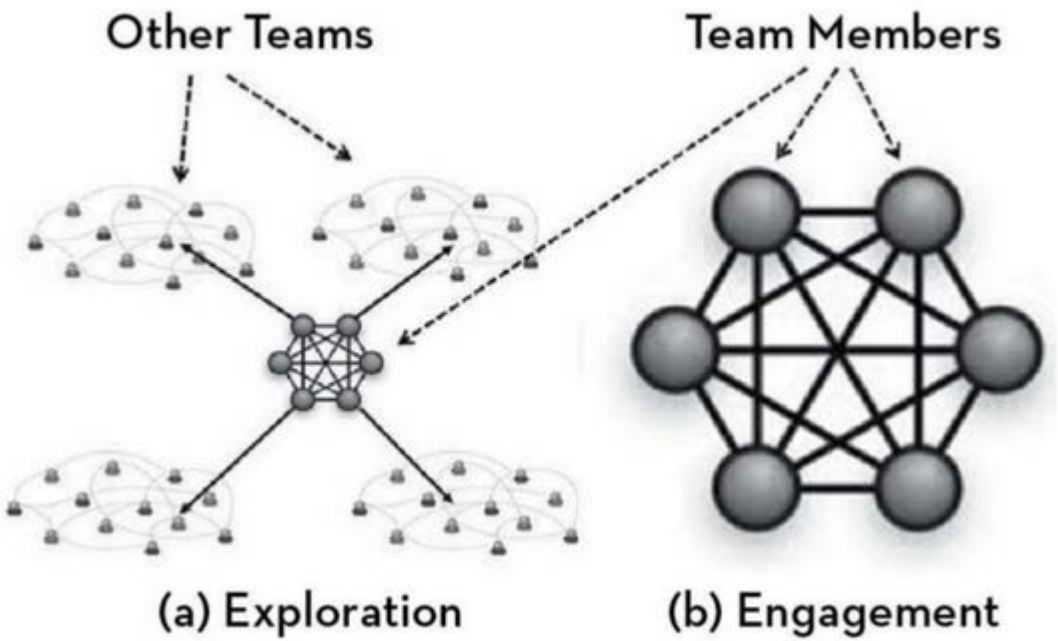
- 參與 (engagement)
  團隊內的互動
- 探索 (exploration)
  跨團隊的交流




hbr.org/2012/04/the-new-science-of-building-great-teams

# 參與和探索行為

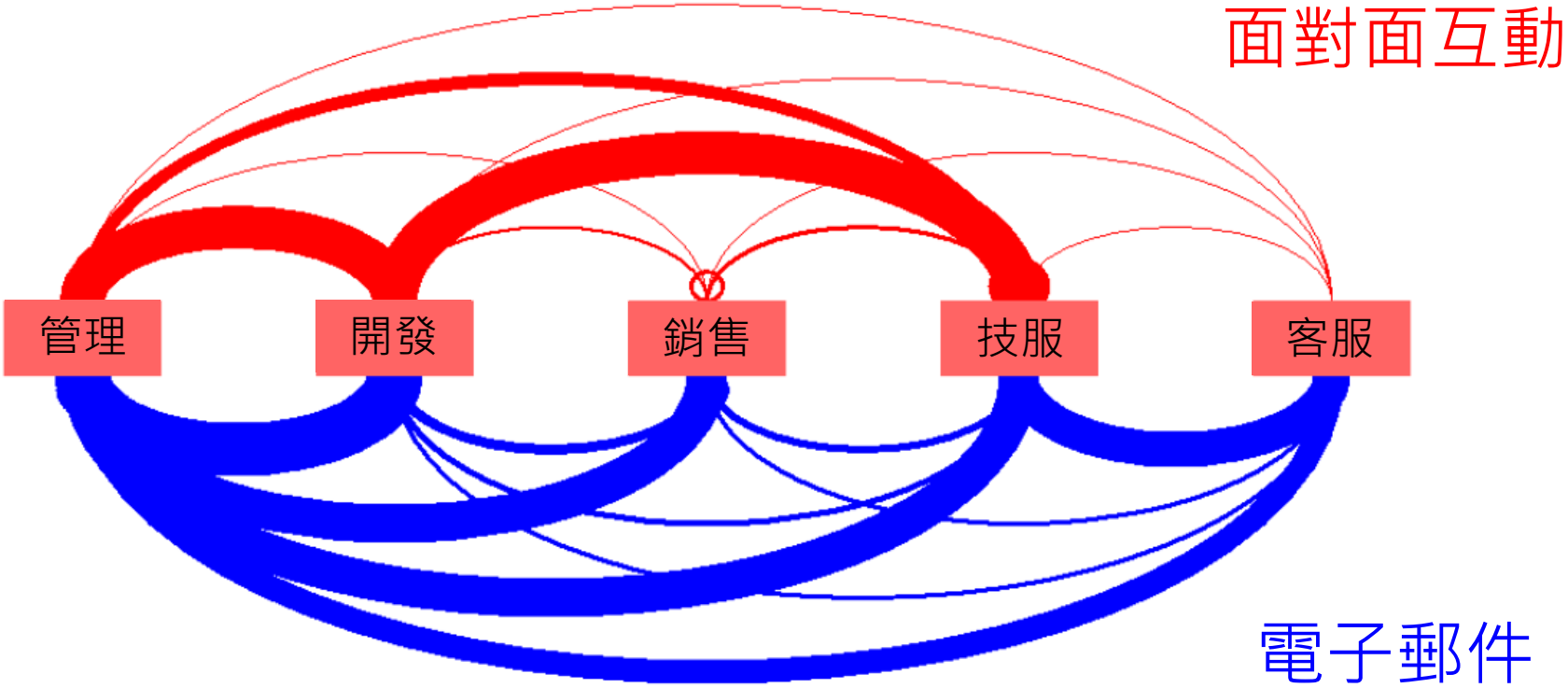
- 星狀網路：產生團隊以外的新意念流
- 密集互連：豐沛互動，有助檢視新意念，並融入團隊的規範和習慣之中




Alex "Sandy" Pentland, Social Physics

# 一個典型的企業架構

- 為期 1 個月，5 個團隊，22 名員工，2,200 小時的資料變化，並監控電子郵件流量，共 880 封郵件。




sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_082159.pdf

# 一個典型的企業架構 (cont)

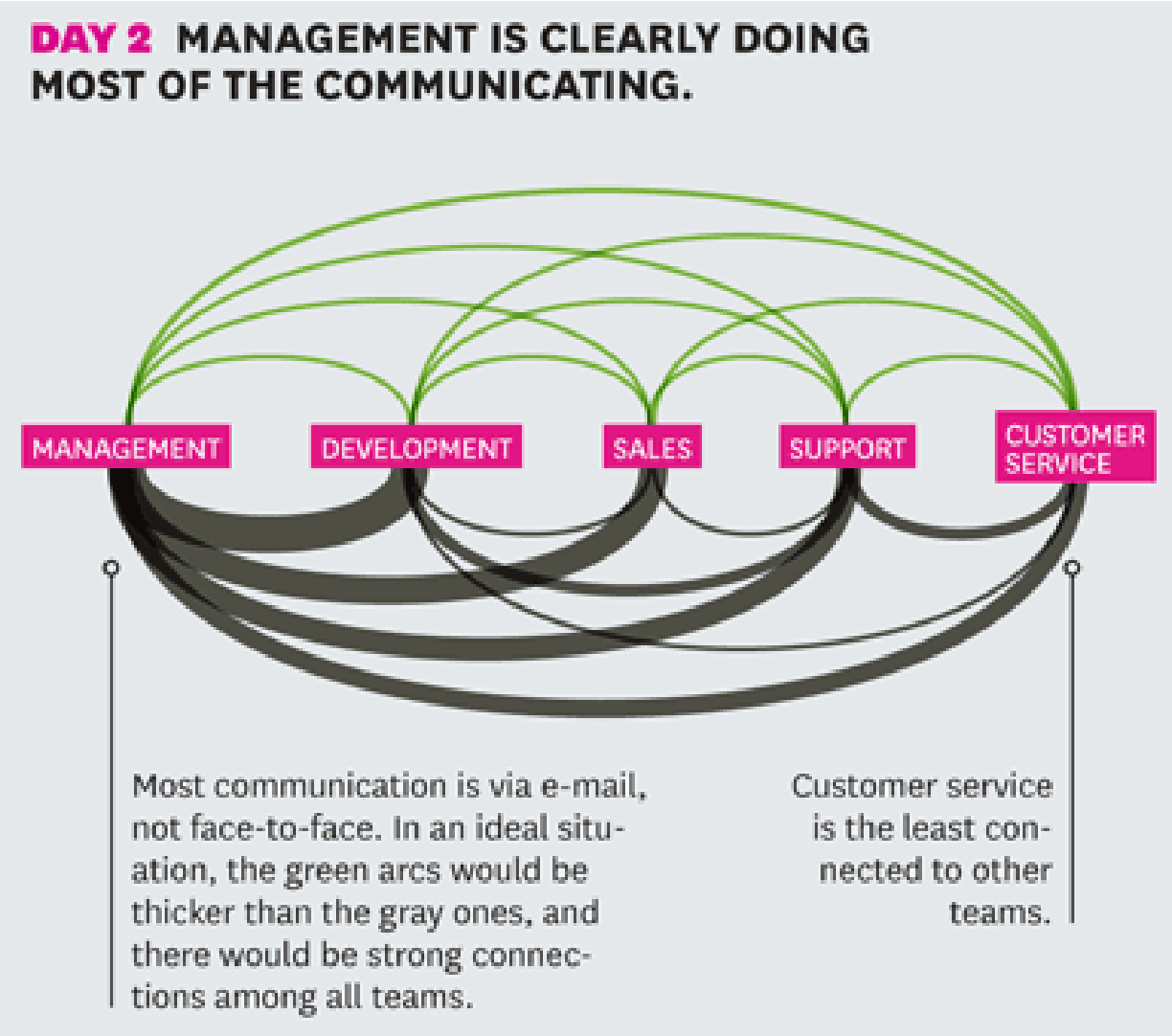
- 設計新行銷專案的團隊在探索和參與兩種模式間擺盪
- 負責製作的團隊則否，主要是團隊內部互動，新想法很少流入

- 意念流黑洞
  - 其他部門很少與客服部門面對面交談
  - 可能解法：改變座位安排，確保所有人都在互動交流圈中，得以改善部門間協調問題

# 一個失敗的專案

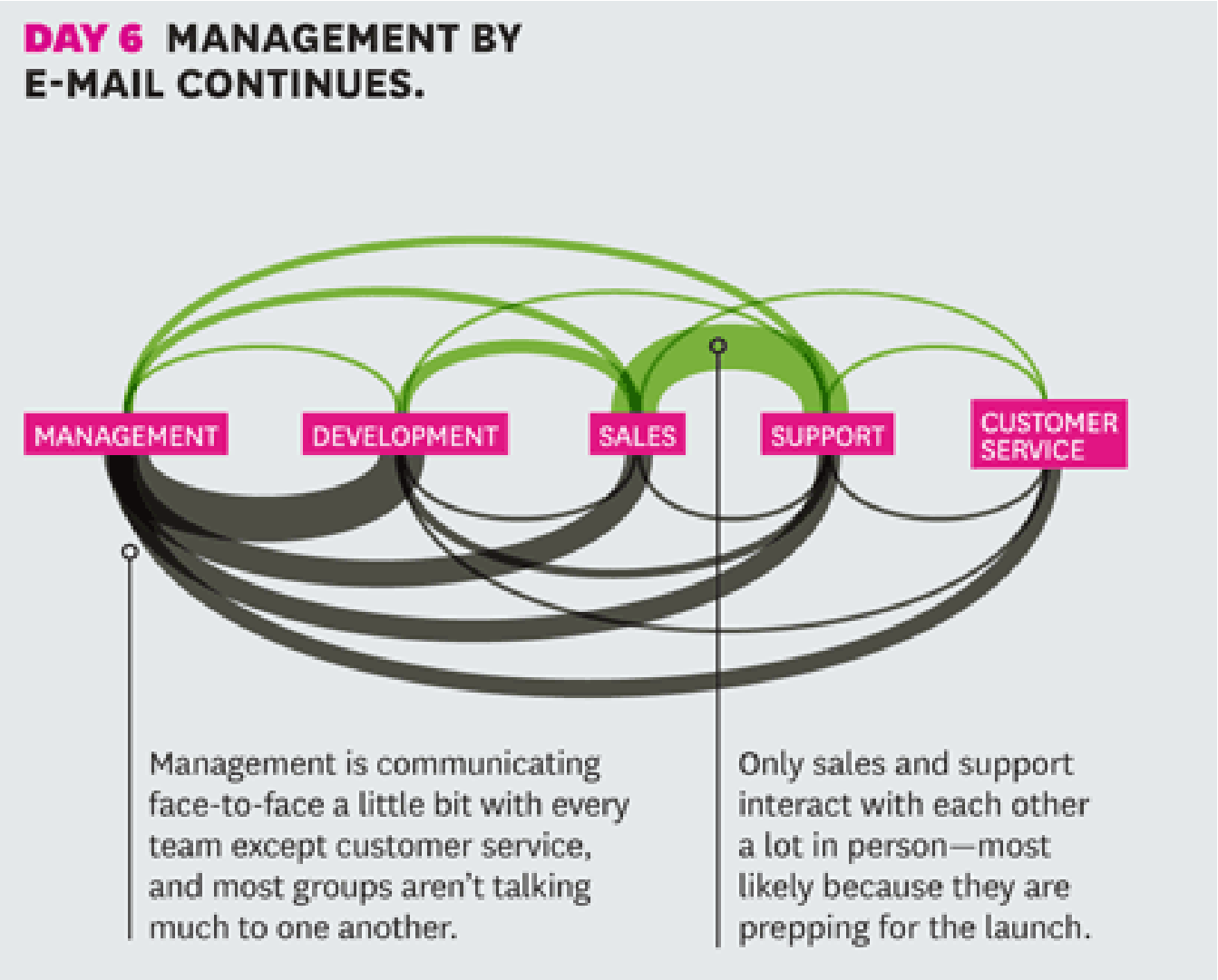
- 20 天的專案監控
- 可從專案起始觀測意念流隨時間的變化，看出不健康的、互動性低的意念流表現

# 專案初始：意念流由管理團隊發出



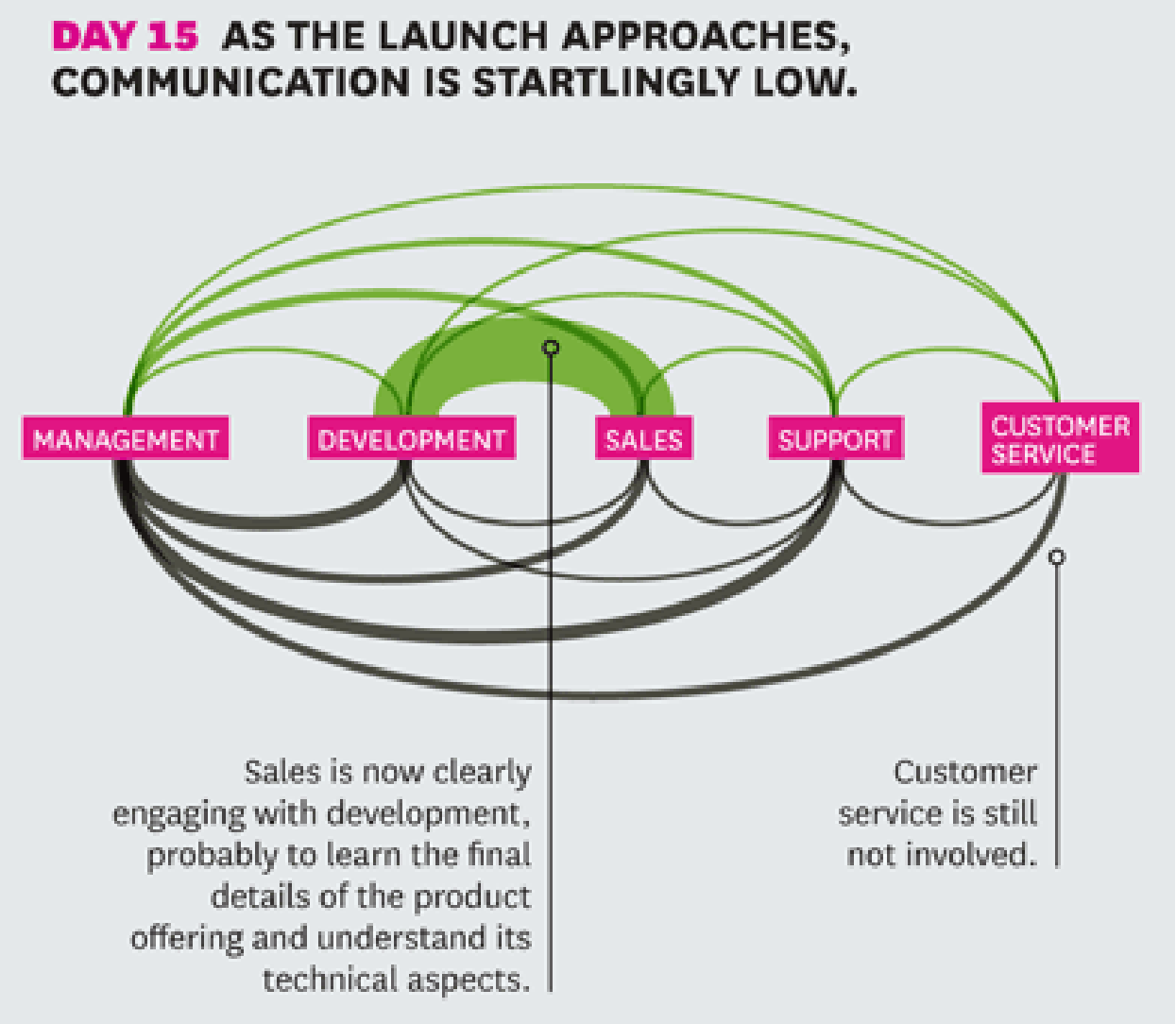

hbr.org/resources/images/article_assets/hbr/1204/R1204C_B_LG.gif

# 僅銷售和支援部門有較多當面溝通



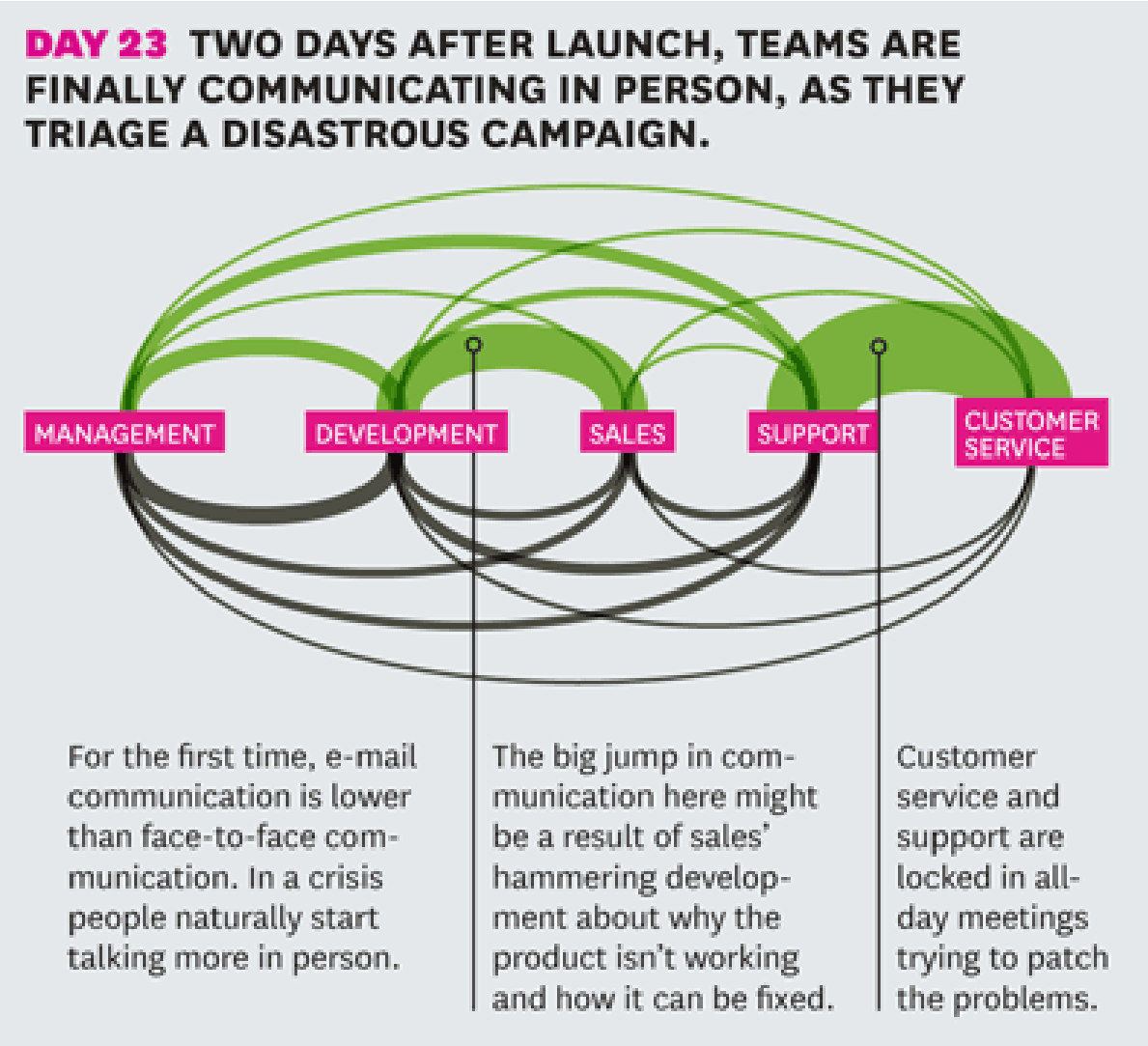

hbr.org/resources/images/article_assets/hbr/1204/R1204C_B_LG.gif

# 接近結案期限，面對面互動量大幅降低




hbr.org/resources/images/article_assets/hbr/1204/R1204C_B_LG.gif

# 交貨發生問題後，部門間開始大量溝通




hbr.org/resources/images/article_assets/hbr/1204/R1204C_B_LG.gif

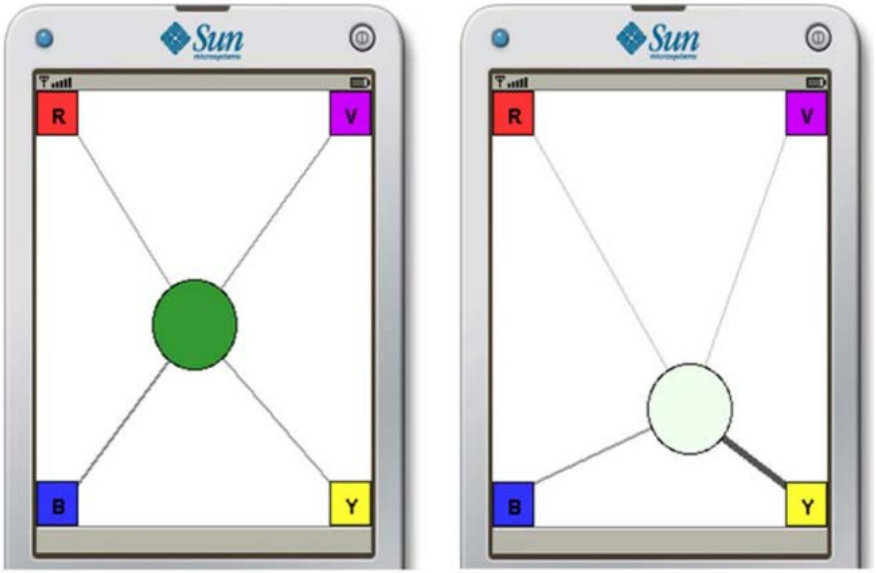
# 改善工作團隊的意念流

- 周五下午 4:30pm 開啤酒趴？
- 把員工餐廳的方桌改成長桌？

# 不僅是觀測，希望進一步改善

- 會議即時反饋系統：社會計量識別牌 + 互動視覺化
- 利用即時視覺反饋鼓勵群體中均衡、高度的參與

alumni.media.mit.edu/~taemie/research.htm
vismod.media.mit.edu//tech-reports/TR-623.pdf

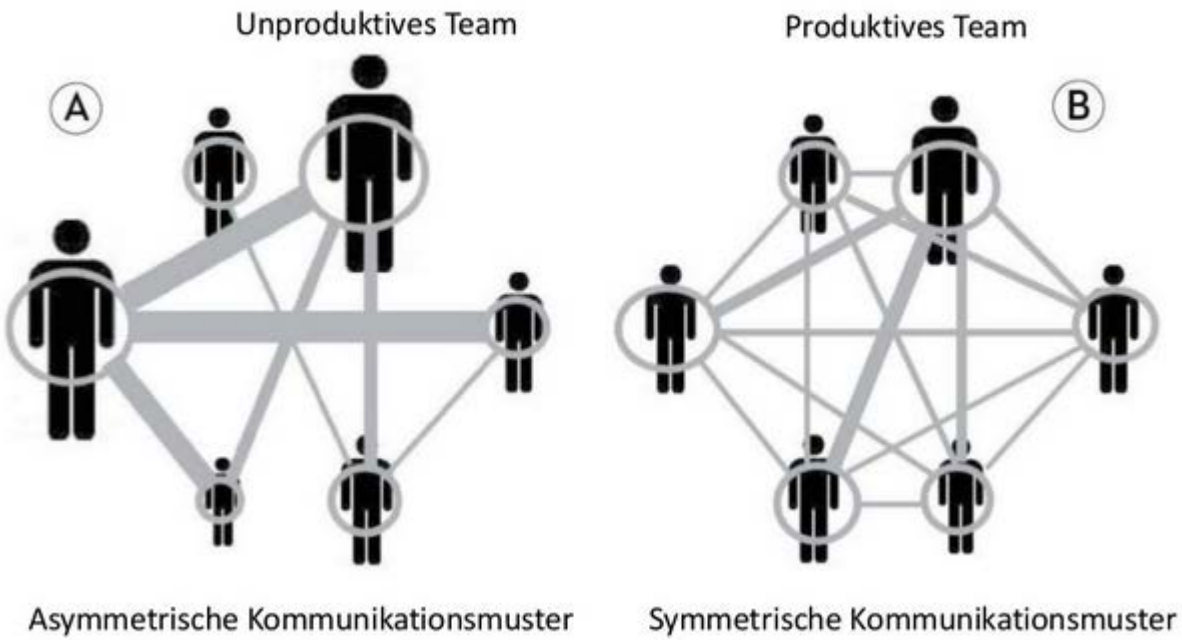



(a) 參與程度高

(b) 特定人士主導

# 高效能表現來自良好的互動型態

- **點子很多**：貢獻簡短意見，而非只有少數長篇大論
- **密集互動**：即時短評（支持或否定），幫助建立共識
- **主意多樣性**：個人參與互動程度相對平均




Alex "Sandy" Pentland, Social Physics

# 「貝爾明星」研究

- 卓越 v.s. 平凡
  - 人脈網絡多樣性 (diversity)
  - 預備式探索 (preparatory exploration)

http://www.thestevensmithblog.com/153/how-can-reaching-out-to-others-build-a-community-and-solve-business-issues/

# 找到魅力型連結者

## 魅力型連結者

- 意念蒐集者，充滿好奇，積極發問
- 精力充沛、推動對話
- 有系統地與他人互動，非支配討論，而是鼓勵良好的意念流型態
- 使意念得以跨越群體的界線流通

≠

## 派對動物

- 口若懸河但總是言不及義
- 注重表象，跟隨流行熱潮
- 好出鋒頭，喜歡成為眾人焦點

# 資料科學用之於企業管理

- 找出未來新星
- 找出可能相處有問題的小團隊
- 找出無法融入族群的新人
- 更準的面試方法
- 預測離職
- 預測人 vs 人 and 人 vs 團隊的速配度
- 預測決策的效果 (e.g., 預測市場)

# 交流時間

Important Message

1. *Hidden Information May Be Revealed with Data Analysis*

2. *Unstructured Data Can Be Valuable*

3. *Small Data May Contain BIG Values*

# 資料科學的五大迷思

- 資料科學是門新學問
- 資料科學總會計算出最好的結果
  - 當公司沒有足夠的資料時，機器學習是個非常糟糕的選擇
- 資料科學會幫我設計產品
  - 資料科學並不能取代專業知識
  - 若團隊本身沒有相關經驗，機器學習無法幫你找到答案
- 使用大數據的解決方案比較好
- 三個月內速成資料科學家

https://blog.alphacamp.co/2015/08/04/data-science-myth/ by 蕭瑟寡人

# 別讓大數據變玄學

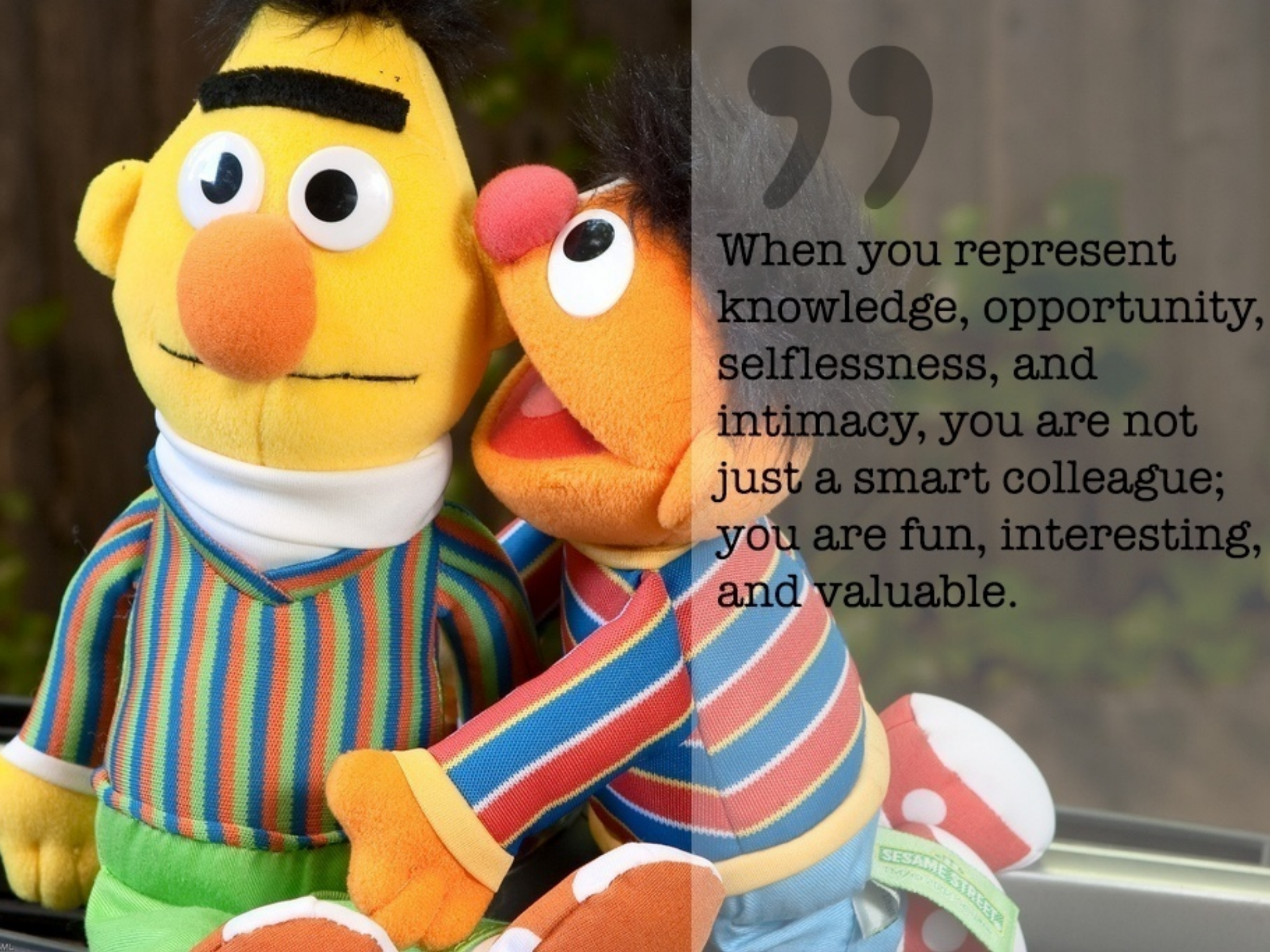
2016-03-08 02:54:10
經濟日報 陳昇瑋

原文網址

Big-data is like teenage sex:

everyone talks about it,
nobody knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it!

**大數據是個經世致用之學，若對你來說仍是個虛無飄緲，好像很厲害，什麼都可以解決，或好像要投資個幾億元幾千萬元才能有回收，建議換個老師／顧問吧。**好老師及壞老師怎麼辨別？很簡單，用「細節測試法」：請對方列出他在數據分析專案內親身進行的成功／失敗步驟，列不出20項具體細節者，所謂的功力大概不是從別人的書上學來，就是從下屬繼承而來的。**（作者為中央研究院資訊科學研究所研究員、台灣資料科學協會理事長）**




When you represent knowledge, opportunity, selflessness, and intimacy, you are not just a smart colleague; you are fun, interesting, and valuable.

2017

~~2016~~

# 台灣資料科學愛好者年會

# 善用資料，創造價值

陳昇瑋

台灣資料科學協會
中央研究院資訊科學研究所