

資料科學的第一堂課

心法、案例分析及團隊建立

陳昇瑋

中央研究院資訊科學研究所





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

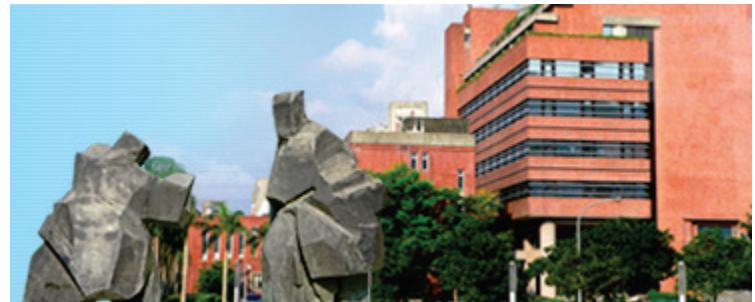
中央研究院資訊科學研究所



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



研究領域

演算法

語音處理

生物資訊

系統技術

中文認知

軟體驗證

智慧型代理人

多媒體

機器學習

多媒體網路與系統實驗室

多媒體網路與系統實驗室

<http://mmnet.iis.sinica.edu.tw>

■ 研究領域

- 使用者經驗
- 多媒體系統
- 計算社會學

中央研究院 | 資訊科學研究所 | 多媒體網路與系統實驗室

Login English

MM Net 多媒體網路與系統實驗室

首頁 研究領域 實驗室成員 研究著作 研究計劃 檔案下載 合作夥伴 學術資源

■ 最新消息 - News

- 徵才啟事：研究助理／博士後研究人員／軟體工程師 [2013-04-04]
- 徵才啟事：中央研究院暨遊戲橘子聯合研究中心 [2013-04-04]
- IEEE Spectrum: Reducing World of Warcraft Power Consumption [2010-08-27]
- 平台發表：Quadrant of Euphoria [2009-10-27]
- 成果發表：Automatic comic storytelling system for online game experience [2009-09-17]
- 服務上線：無名小站：文字繪 [2009-05-20]
- 服務上線：無名小站情報分析事務所 [2009-04-25]
- 服務上線：惡魔快手打字練習 [2008-12-29]

■ 關於我們 - MMNET Laboratory

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？:-)

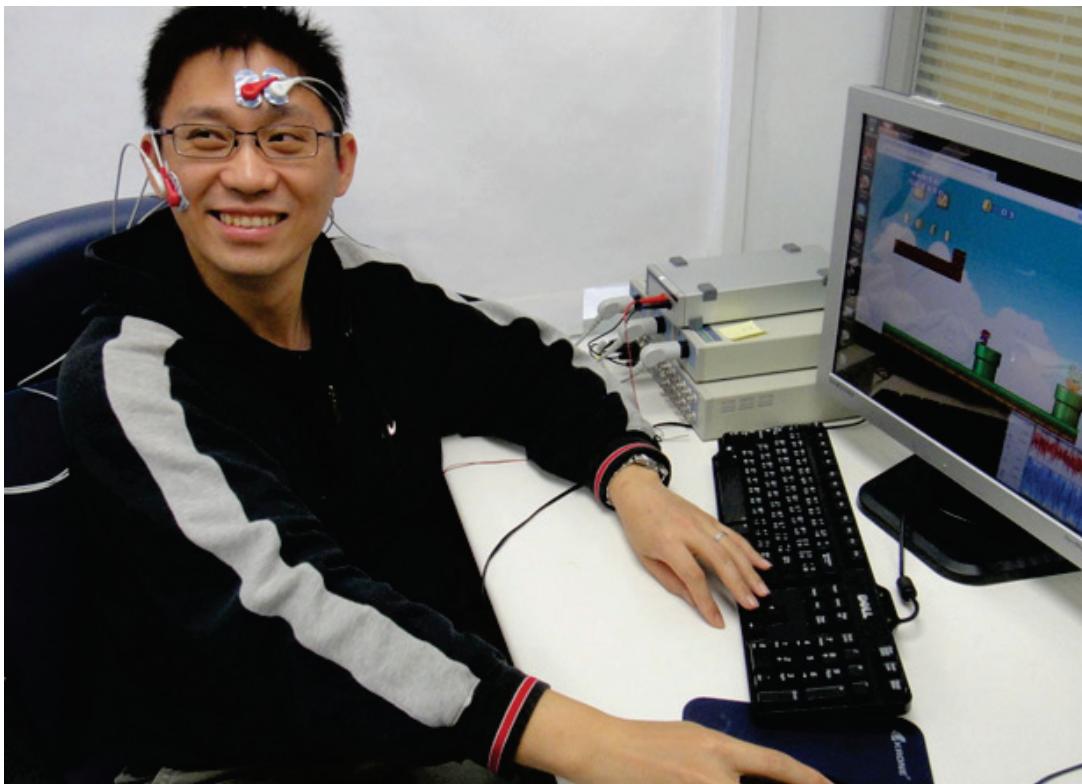


■ 研究領域 - Research Areas

- 使用者經驗
- 多媒體系統
- 社群計算
- 人智運算

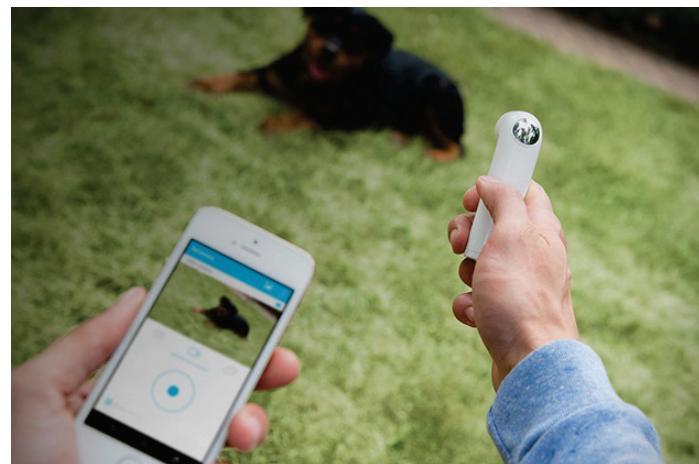
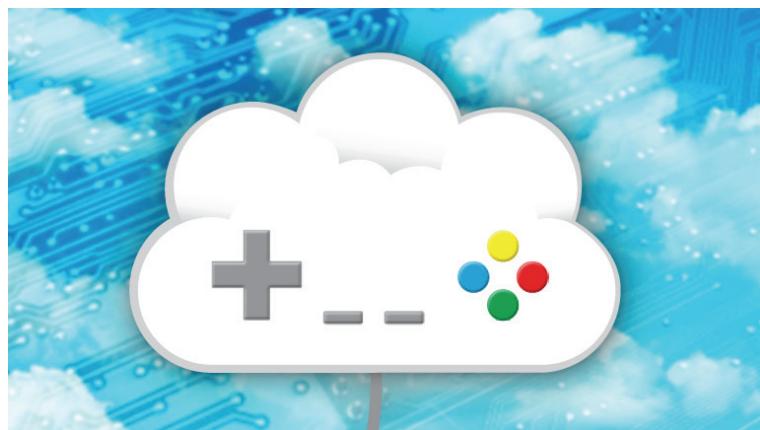
Area 1: Quality of Experience

- 使用情緒量測技術來預言線上遊戲的成與敗

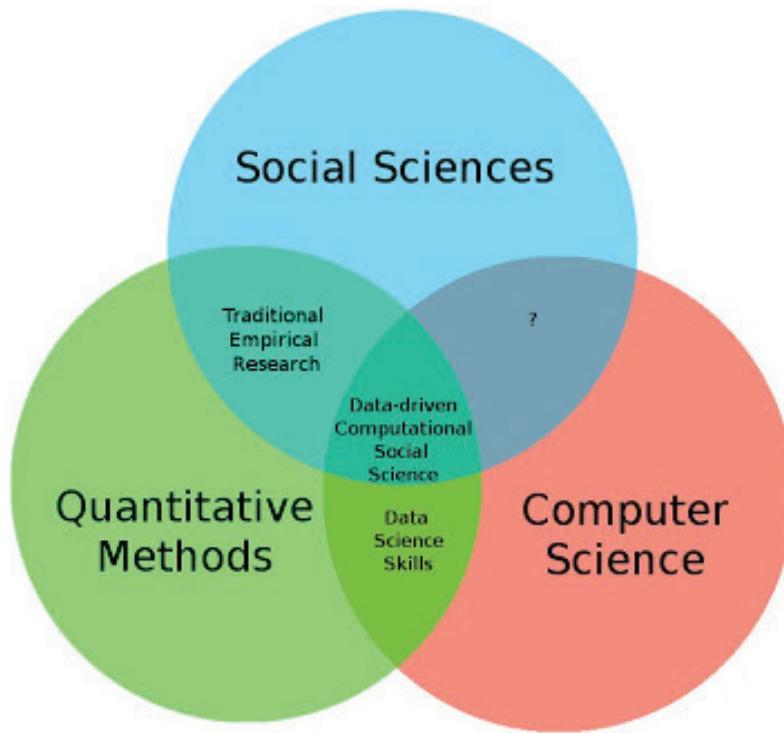


[1] Jing-Kai Lou, Kuan-Ta Chen, Hwai-Jung Hsu, and Chin-Laung Lei, Forecasting Online Game Addictiveness, IEEE/ACM NetGames 2012.

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.

好，進入正題

LET'S SOLVE THIS PROBLEM BY
USING THE BIG DATA NONE
OF US HAVE THE SLIGHTEST
IDEA WHAT TO DO WITH



@markatoonist.com

Definition of “Science”

“

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

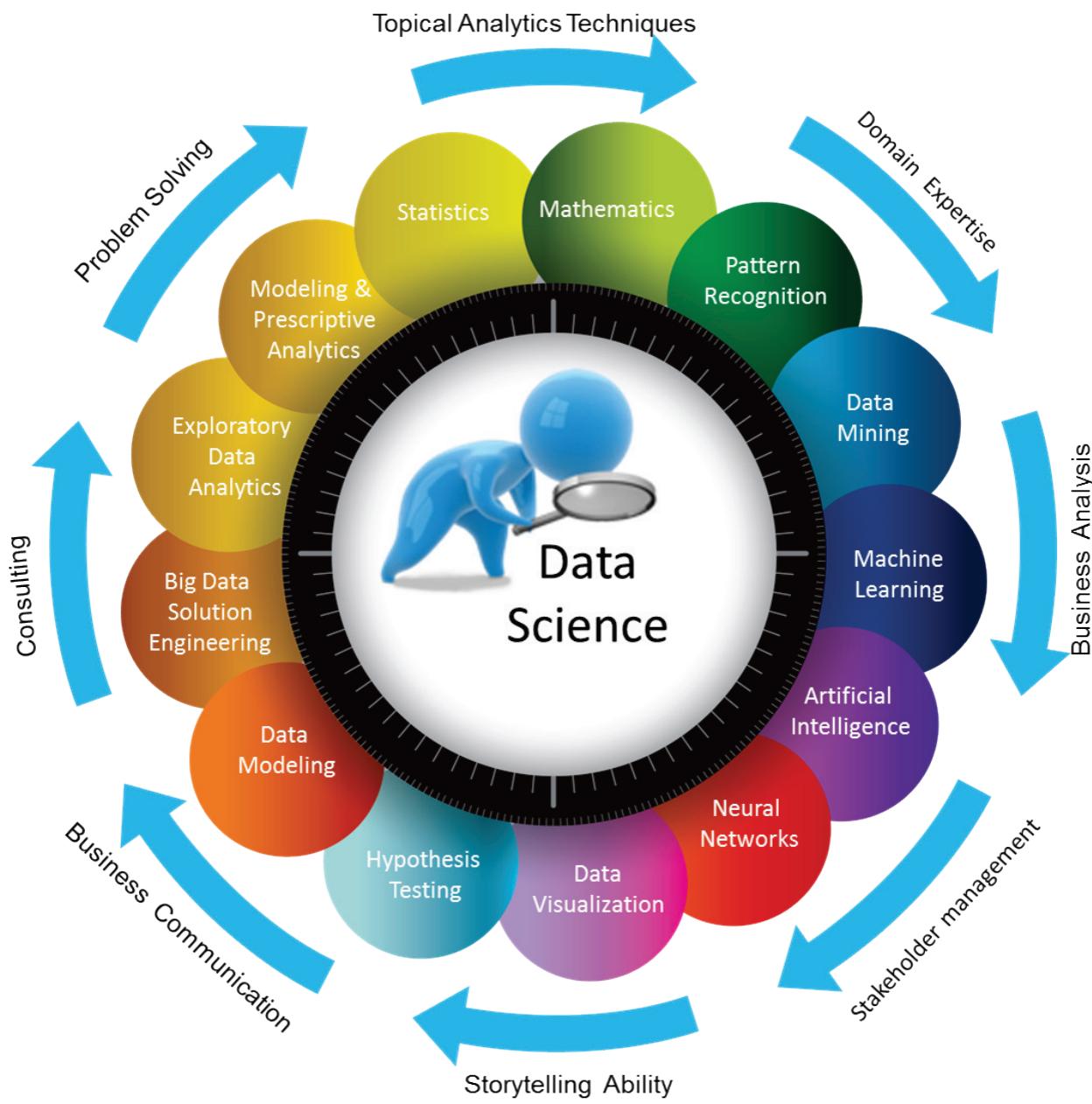
In modern usage "science" most often refers to a **way of pursuing knowledge**, not only to the knowledge itself. Over the course of the 19th century, the word "science" became increasingly associated with the **scientific method** itself.





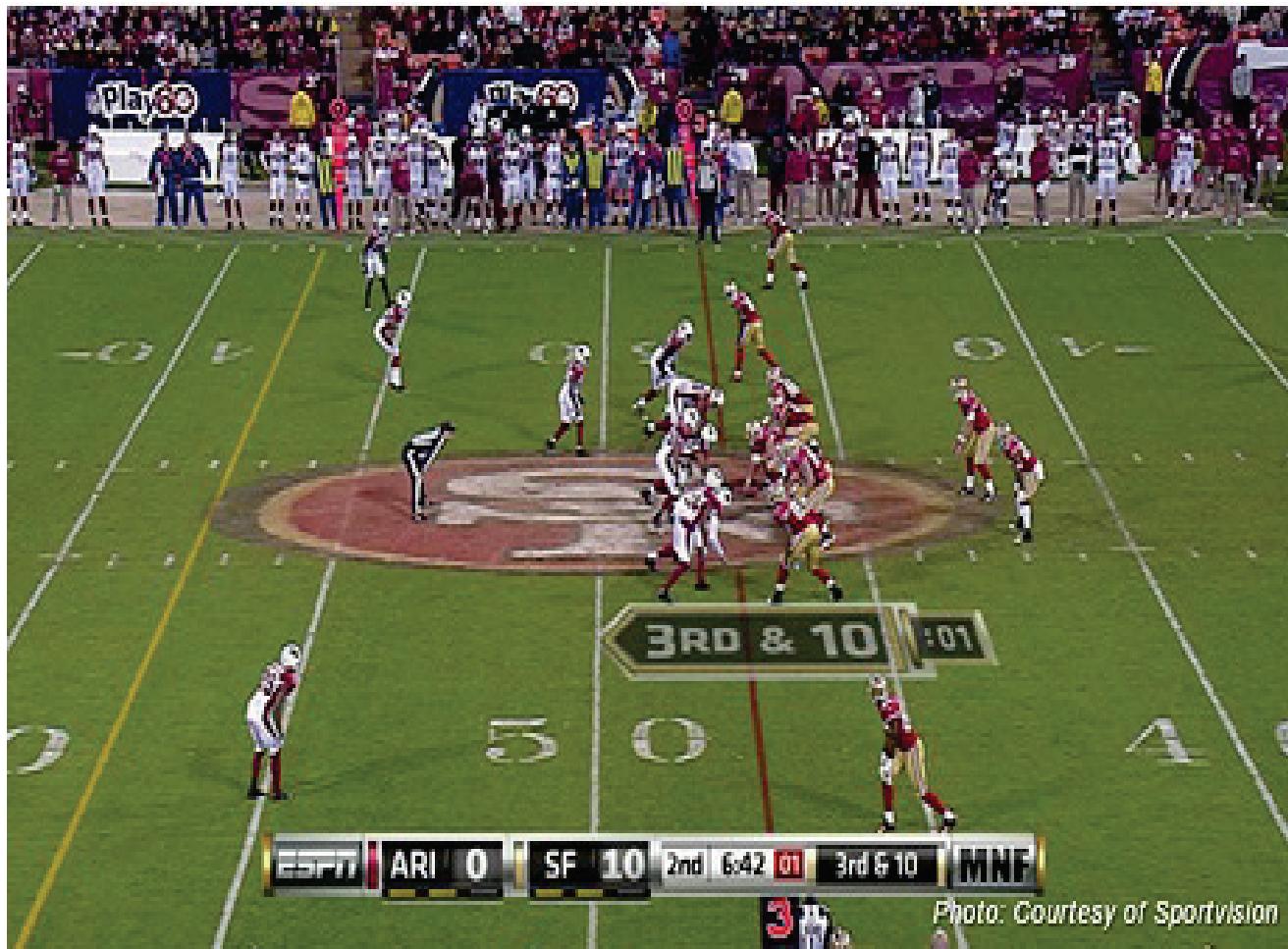


陳昇璋 / 資料科學的第一堂課





Computer vision in sports



SportVision: improving viewer experiences

Computer vision in sports



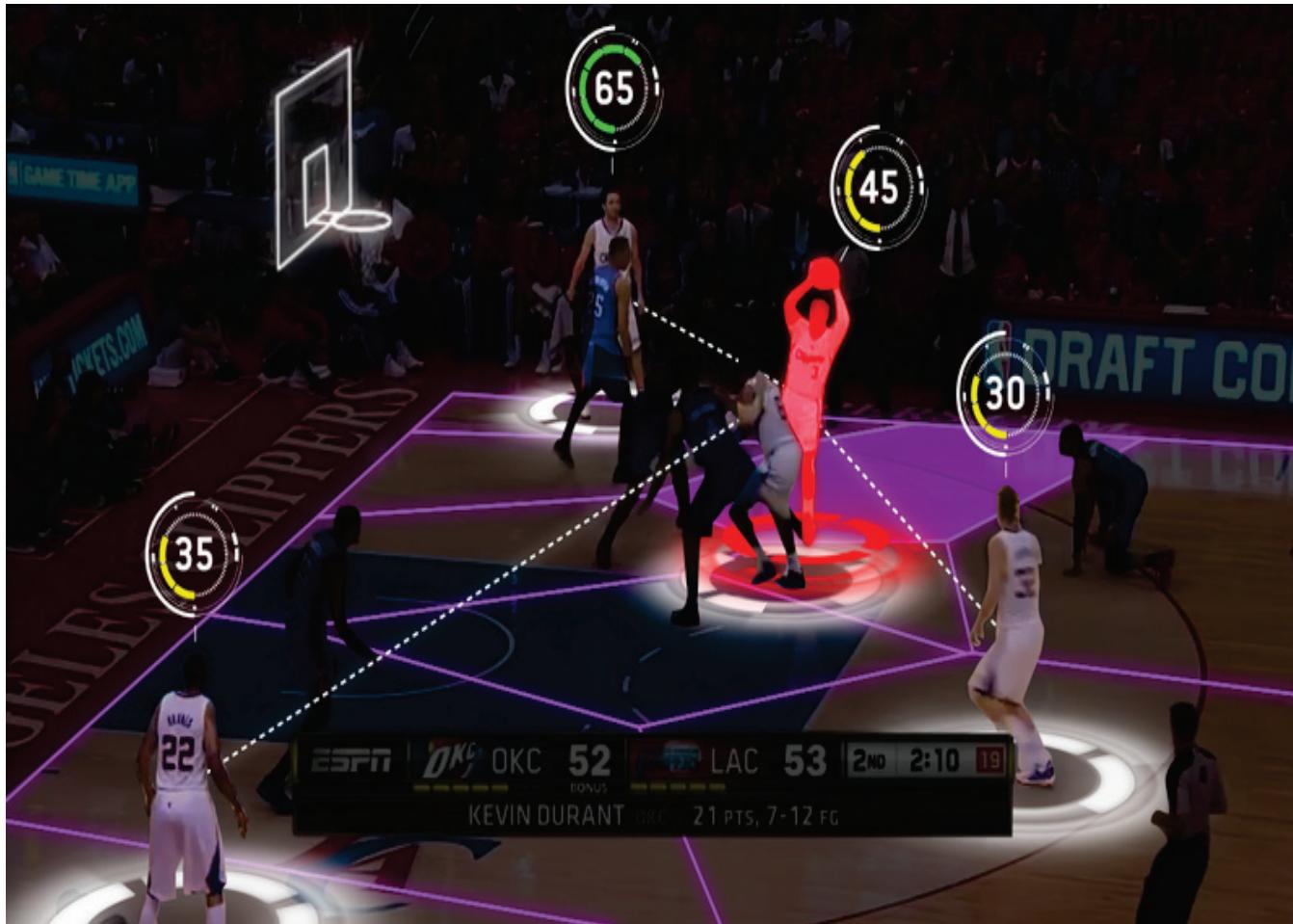
Replay Technologies: improving viewer experiences

Computer vision in sports



Play tracking

Computer vision in sports



Second Spectrum: visual analytics

Matching street clothing photos in online Shops



SEARCH

NEW IN CLOTHING SALE LOOKBOOK

HOME / WALG KNOT TIE FLORAL DRESS

WALG KNOT TIE FLORAL DRESS
CODE - WG 6184
£33.00

CORAL

S SIZE GUIDE

ADD TO BAG

ADD TO WISHLIST

DETAILS

WALG DRESS

- BODYCON FABRIC
- FLORAL PRINT
- SHORT SLEEVES
- COWL NECKLINE
- NO FASTENINGS-SLIP ON/OFF

MATERIAL:
- 95% POLYESTER

<http://walg.co.uk/walg-knot-tie-floral-dress.html>

BUY IT!

HIPPO



HIPPO-driven → Data-driven



Let data drive decisions, not the Highest Paid Person's Opinion.

#HowGoogleWorks

HowGoogleWorks.net

Why Data Science Is Popular?

Big Data

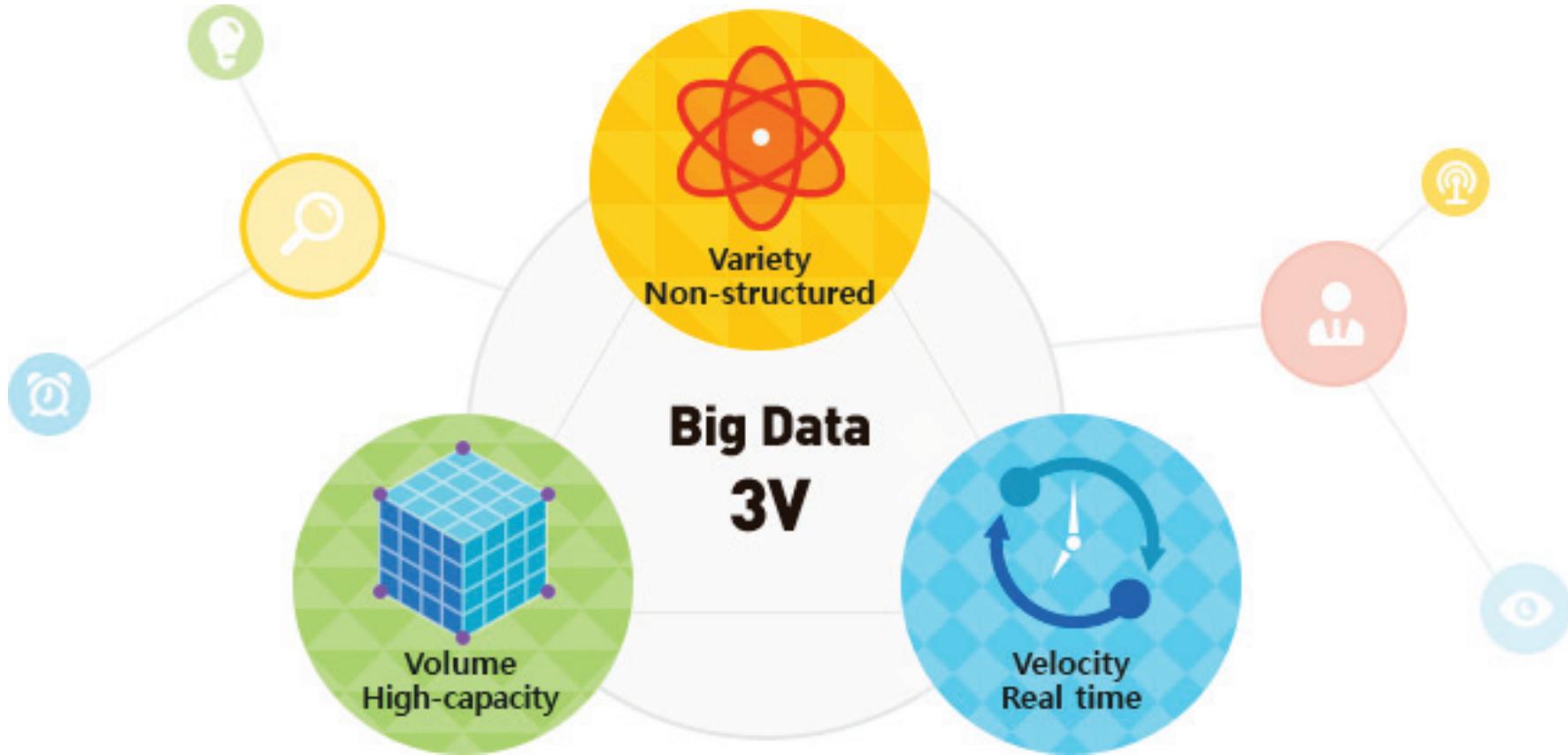
Streaming
Data System

Data
Discovery



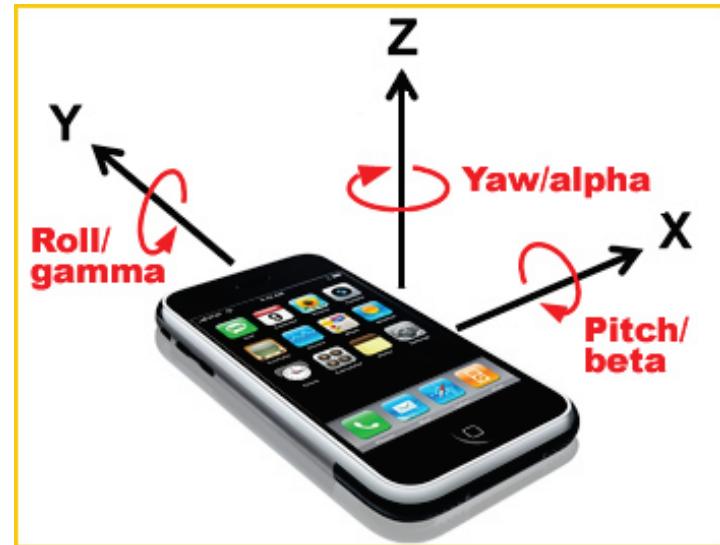
Why Data Science Is Popular? (#1)

#1. Big Data

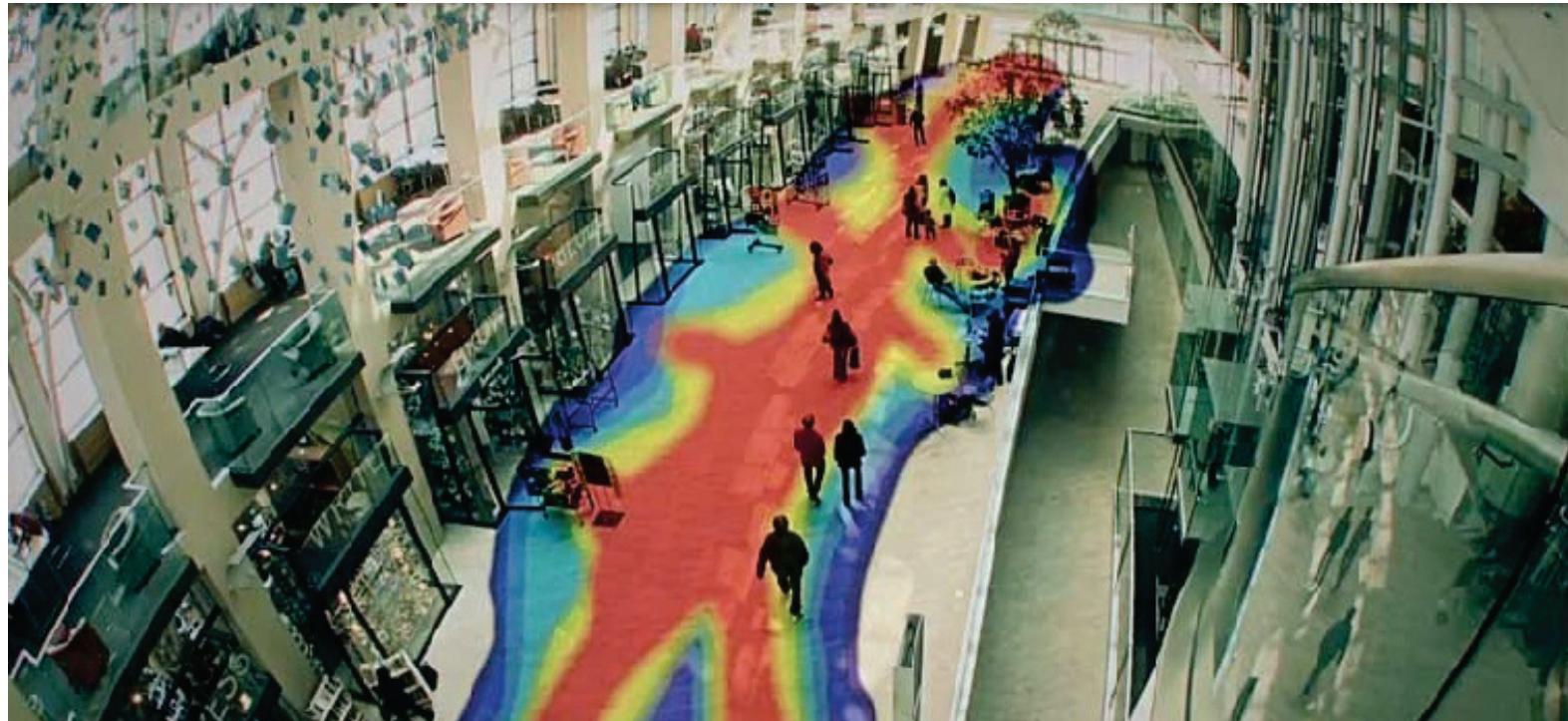
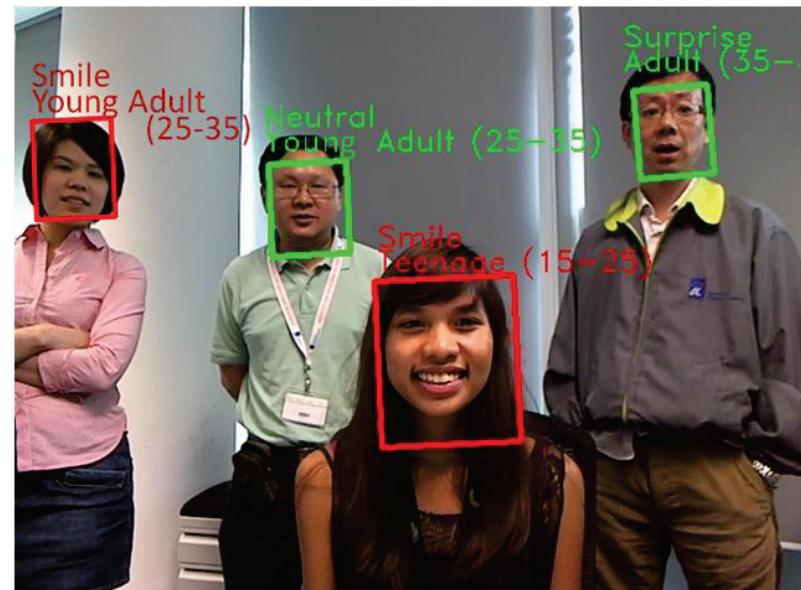


Why Big Data Arises?

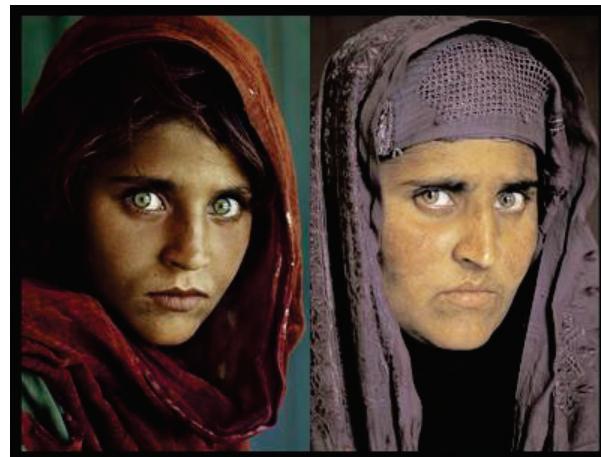
- Collecting & processing data is much cheaper now
- Massive number of online users
- Open data
- New types and wide deployed sensors



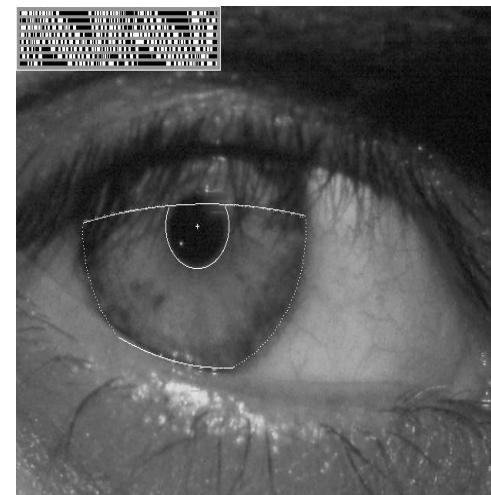
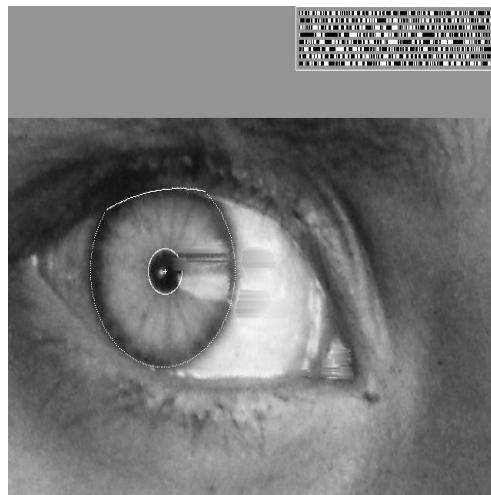
5 - Tills



Vision-based Biometrics



“How the Afghan Girl was Identified by Her Iris Patterns” Read the [story wikipedia](#)



Computer vision for healthcare



Video magnification

Heart Rate = 128 BPM

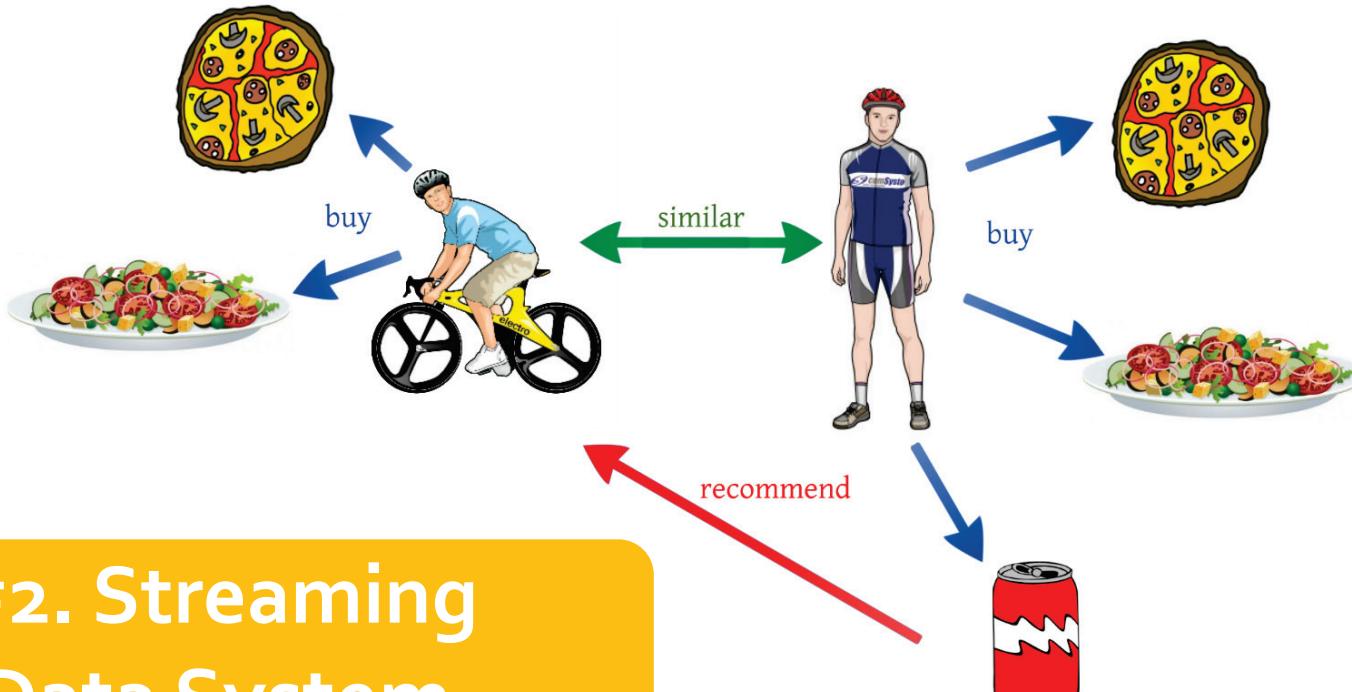


<https://www.youtube.com/watch?v=QbXgEbeceJI>



Why Data Science Is Popular? (#2)

- E.g., recommender systems, anomaly detection systems
- Advances in streaming data infrastructure





Target's Pregnancy Index

“

[Pole] ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole's colleagues noticed that women on the baby registry were **buying larger quantities of unscented lotion around the beginning of their second trimester**. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like **calcium, magnesium and zinc**. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and **extra-big bags of cotton balls, in addition to hand sanitizers and washcloths**, it signals they could be getting close to their delivery date.

”

“

As Pole's computers crawled through the data, he was able to identify about **25 products** that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

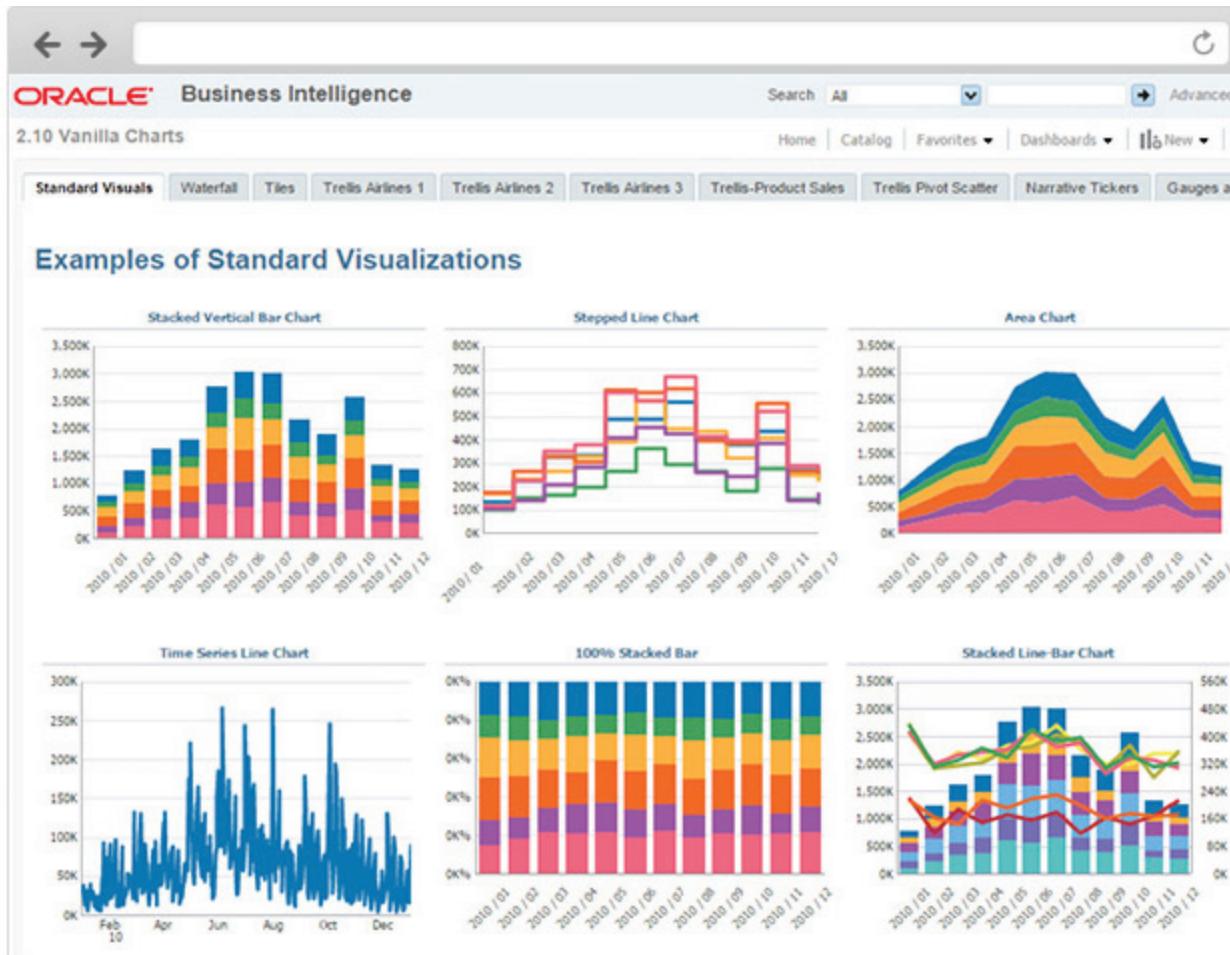
”

One Target employee I spoke to provided a hypothetical example. Take a fictional Target shopper named Jenny Ward, who is 23, lives in Atlanta and **in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug**. There's, say, an 87 percent chance that she's pregnant and that her delivery date is sometime in late August.

Why Data Science Is Popular? (#3)

- New
- Advanced
- Information

Q: 如何打



#3. Data DISCOVERY

提升研發效率

提升回頭率

提升行政效率

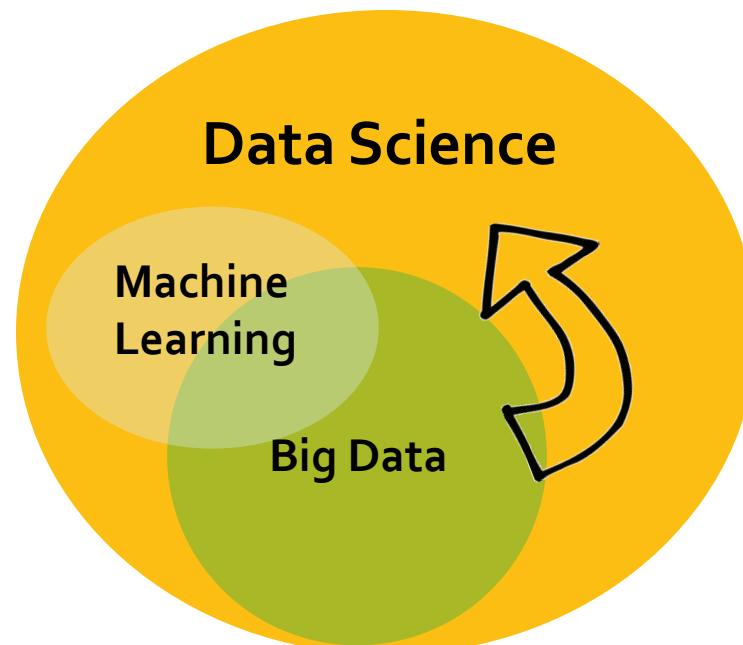
打壓對手 XD

Business is of

成本

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 is More Than ...

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



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 modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

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



PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing packages, e.g., R
- ★ Databases: SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ★ Able to engage with 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

2 days × 800 ppl



8/30 – 8/31, 2014

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

<https://www.facebook.com/twdsconf>



演講議程講師群



2014 台灣資料科學愛好者年會
8/30-31。中央研究院人文科學館

7/7
開始報名

- 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

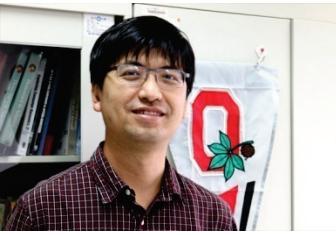
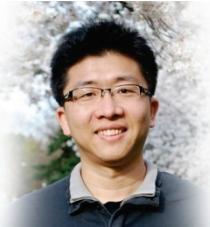
2015

台灣資料科學 愛好者年會

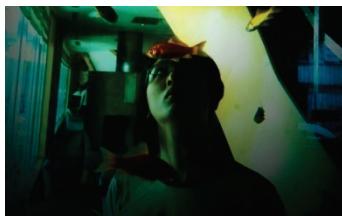
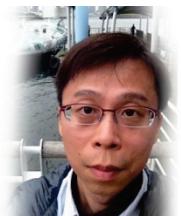
8/20 – 8/23, 2015

<http://datasci.tw/>





資料科學家 x 24





Hadoop/Spark
資料科學快速體驗營



R 語言
資料分析上手課程

Taiwan R User Group

SheetHub



地圖資料視覺化課程



開放文化基金會

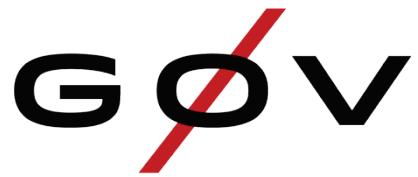


資料新聞實戰營
dBootcamp Taipei

 智庫驅動

The logo consists of the lowercase letters 'cSp' in a stylized, rounded font, with 'c' in green and 'Sp' in blue. To the right of the letters, the words '智庫驅動' are written in a large, blue, sans-serif font.

DSP 資料開竅・企業論壇

 g0v

The logo consists of the lowercase letters 'g0v' in a bold, black, sans-serif font. A thick red diagonal slash is positioned across the 'g' and '0'.

g0v 零時政府黑客松





資料分析可以拿來
解決什麼問題？



1. 計算社會學
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

當電腦科學家遇上社會科學

陳昇瑋

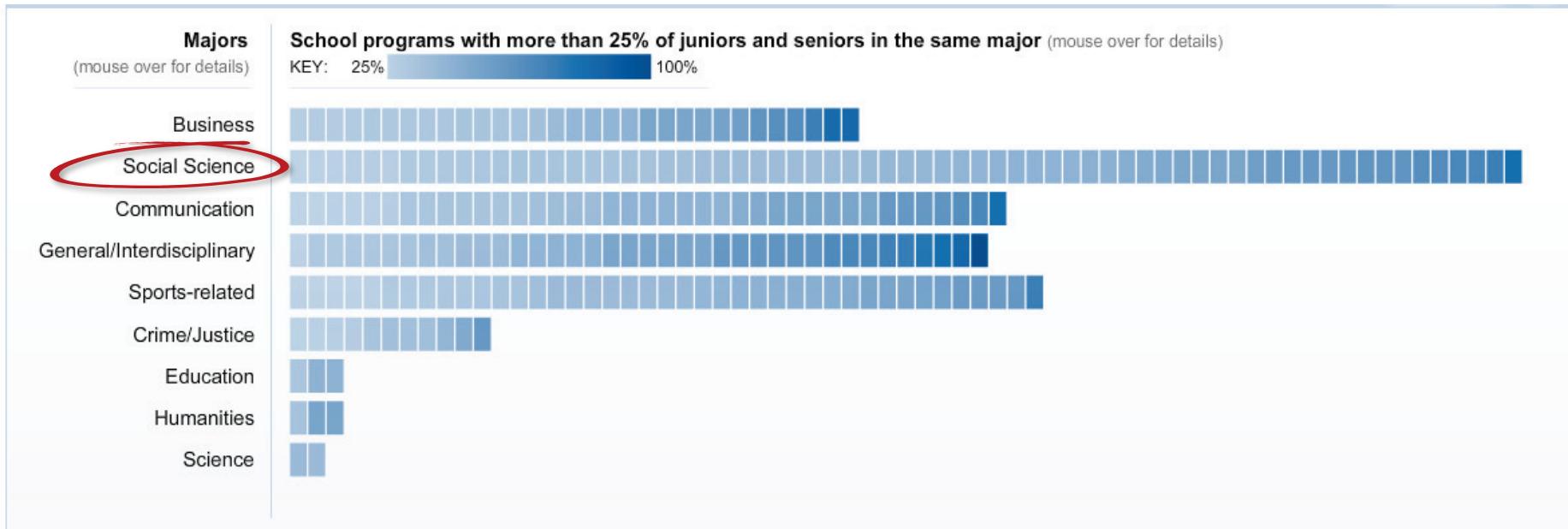
中央研究院資訊科學研究所



PEOPLE

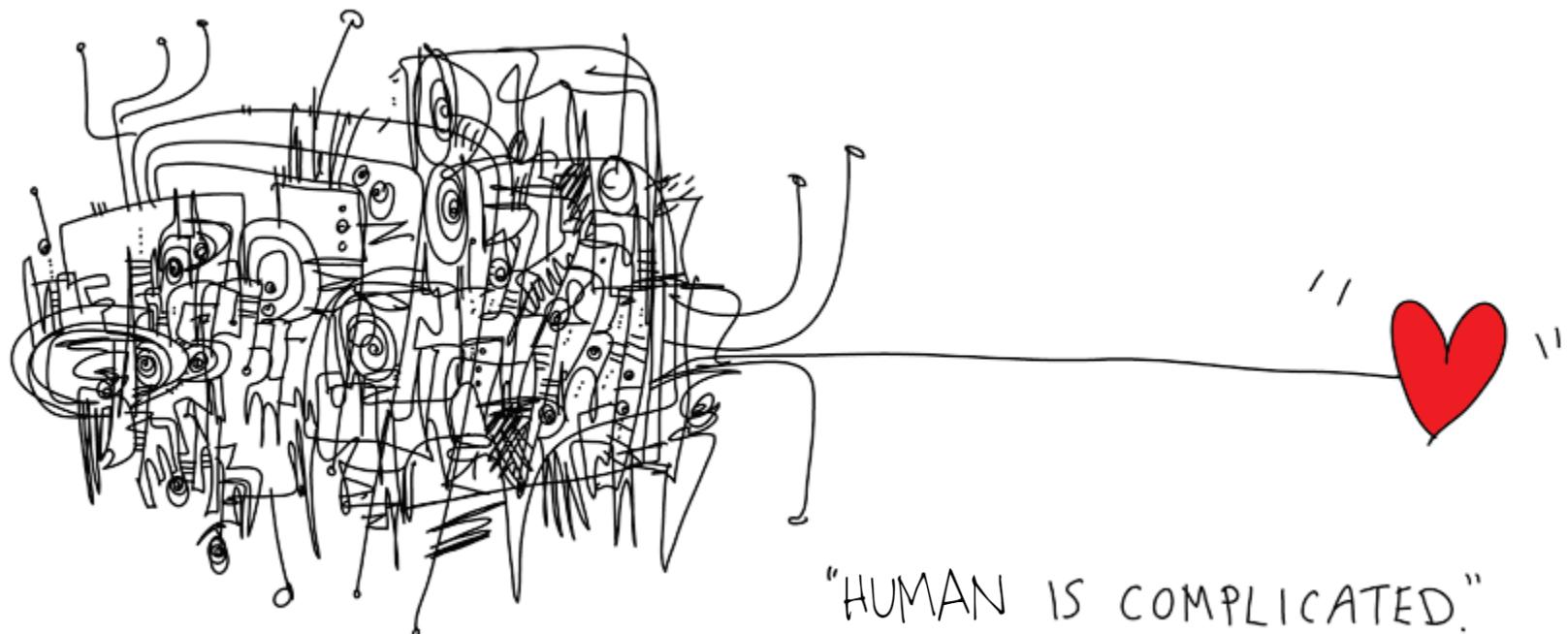


The Favorite Major for US College Athletes



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

Social Science



@gapingvoid

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

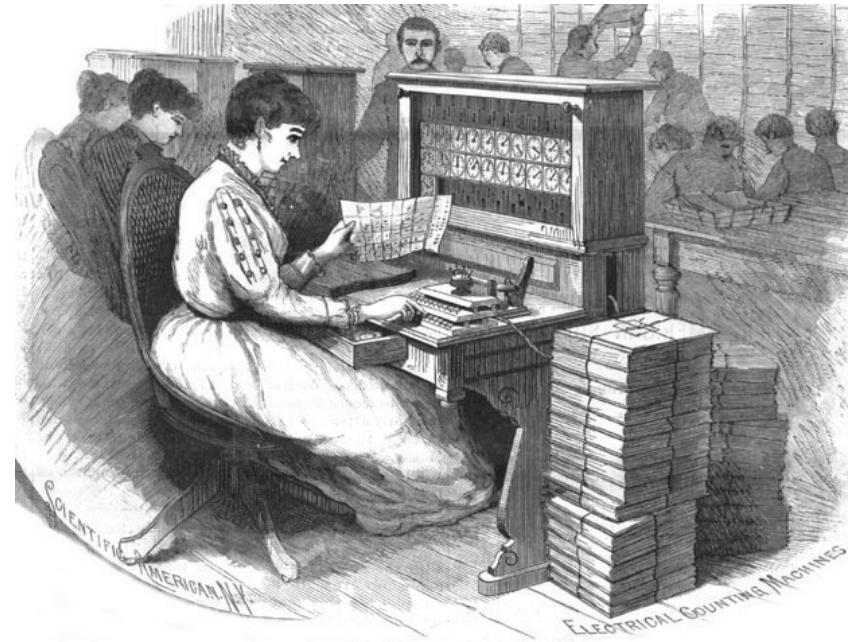
1890 US Census

US 1890 Census Population Schedules

FAMILY SCHEDULE—1 TO 10 PERSONS.

C44

Supervisor's District No. 3		[7-850 e.]		Eleventh Census of the United States.					
Enumeration District No. 78		SCHEDULE No. 1.							
		POPULATION AND SOCIAL STATISTICS.							
Name of city, town, township or post office, and number of civil division: <i>Perryville, Perry Co., Alabama</i>		County: <i>Perry</i>		State: <i>Alabama</i>					
Street and No.: <i>—</i>		Ward: <i>—</i>		Name of Institution: <i>—</i>					
Enumerated by me on the <i>4th</i> day of June, 1890.		<i>Lily J. Davis</i>							
A.—Number of Dwelling-houses in the order of visitation.	<i>144</i>	B.—Number of families in the dwelling-houses.	<i>1</i>	C.—Number of persons in this dwelling-house.	<i>8</i>	D.—Number of Families in the order of visitation.	<i>44</i>	E.—No. of Persons in this family.	<i>8</i>
INQUIRIES.									
	<i>389^{b1}</i>	1	2	3	4	5	6	7	8
1	Christian name in full, and initials or middle names.	<i>William W.</i>	<i>Eliza A.</i>	<i>Charles A.</i>	<i>Young E.</i>	<i>George P.</i>			
2	Married.	<i>Sophia</i>	<i>Smith</i>	<i>Smith</i>	<i>Smith</i>	<i>Smith</i>			
3	Relationship to head of family.	<i>Conf. Sol.</i>	<i>+ +</i>	<i>+ +</i>	<i>+ +</i>	<i>+ +</i>			
4	Whether white, black, mulatto, Indian, Japanese, or Chinese.	<i>White</i>	<i>White</i>	<i>White</i>	<i>White</i>	<i>White</i>			
5	Sex.	<i>Male</i>	<i>Female</i>	<i>Male</i>	<i>Male</i>	<i>Male</i>			
6	Age at nearest birthday, if less than one year, give age in months.	<i>55-</i>	<i>84</i>	<i>23</i>	<i>21</i>	<i>19</i>			
7	Whether single, married, widowed, or divorced.	<i>Married</i>	<i>Married</i>	<i>Single</i>	<i>Single</i>	<i>Single</i>			
8	Whether married during the census year (June 1, 1890, to May 31, 1891).	<i>No</i>	<i>No</i>	<i>+ +</i>	<i>+ +</i>	<i>+ +</i>			
9	Mother of how many children, number of those children living.	<i>x</i>	<i>14-28</i>	<i>+</i>	<i>+</i>	<i>+</i>			
10	Place of birth.	<i>Alabama</i>	<i>Alabama</i>	<i>Alabama</i>	<i>Alabama</i>	<i>Alabama</i>			
11	Place of birth of Father.	<i>South Carolina</i>	<i>Georgia</i>	<i>Alabama</i>	<i>Alabama</i>	<i>Alabama</i>			
12	Place of birth of Mother.	<i>South Carolina</i>	<i>Tennessee</i>	<i>Alabama</i>	<i>Alabama</i>	<i>Alabama</i>			
13	Number of years in the United States.	<i>x</i>	<i>x</i>	<i>x</i>	<i>x</i>	<i>x</i>			
14	Whether naturalized.	<i>x</i>	<i>x</i>	<i>x</i>	<i>x</i>	<i>x</i>			
15	Whether naturalization papers have been taken out.	<i>x</i>	<i>x</i>	<i>x</i>	<i>x</i>	<i>x</i>			
16	Profession, trade, or occupation.	<i>Farm</i>	<i>Homemg</i>	<i>Farm</i>	<i>Farm</i>	<i>Farm</i>			
17	Months unemployed during the census year (June 1, 1890, to May 31, 1891).	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>			
18	Age at which first married during the census year (June 1, 1890, to May 31, 1891).	<i>+ +</i>	<i>+ +</i>	<i>+ +</i>	<i>18 months</i>	<i>8 months</i>			
19	Able to Read.	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>			
20	Able to Write.	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>			
21	Able to speak English. If not, the language or dialect spoken.	<i>English</i>	<i>English</i>	<i>English</i>	<i>English</i>	<i>English</i>			
22	Whether suffering from some disease or infirmity, and if so, name of disease and length of time.	<i>No</i>	<i>No</i>	<i>No</i>	<i>No</i>	<i>No</i>			
23	Whether deafened in mind, sight, hearing, speech, or limb, or crippled, palsied, or disabled, with name of defect.	<i>No</i>	<i>No</i>	<i>No</i>	<i>No</i>	<i>No</i>			
24	Whether blind, or nearly so, congenital, hereditary, or acquired.	<i>No</i>	<i>No</i>	<i>No</i>	<i>No</i>	<i>No</i>			
25	Supplementary 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 house visited.									
CENSUS—1890 (1890) 1-18									



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, ...

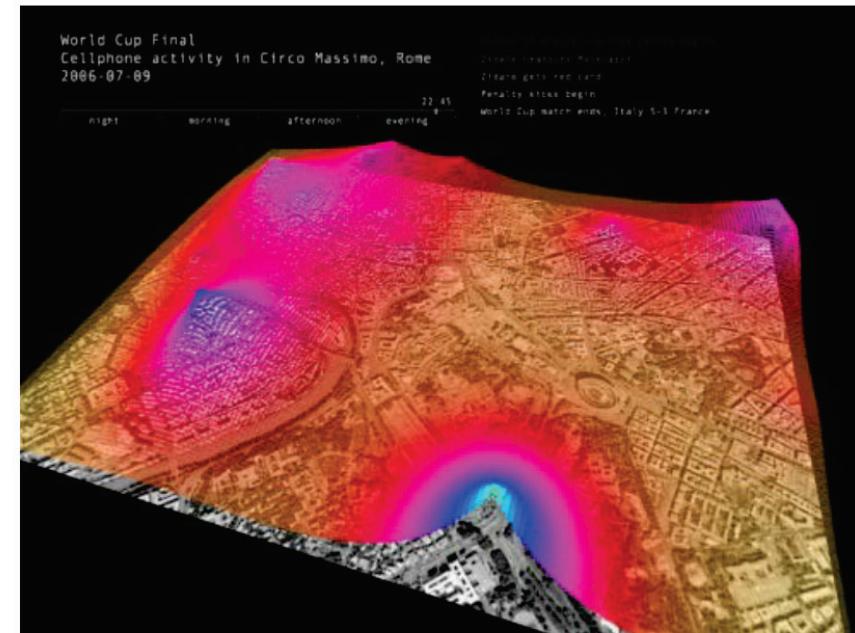


New kinds of data

Urban movement analysis from GPS/phone data

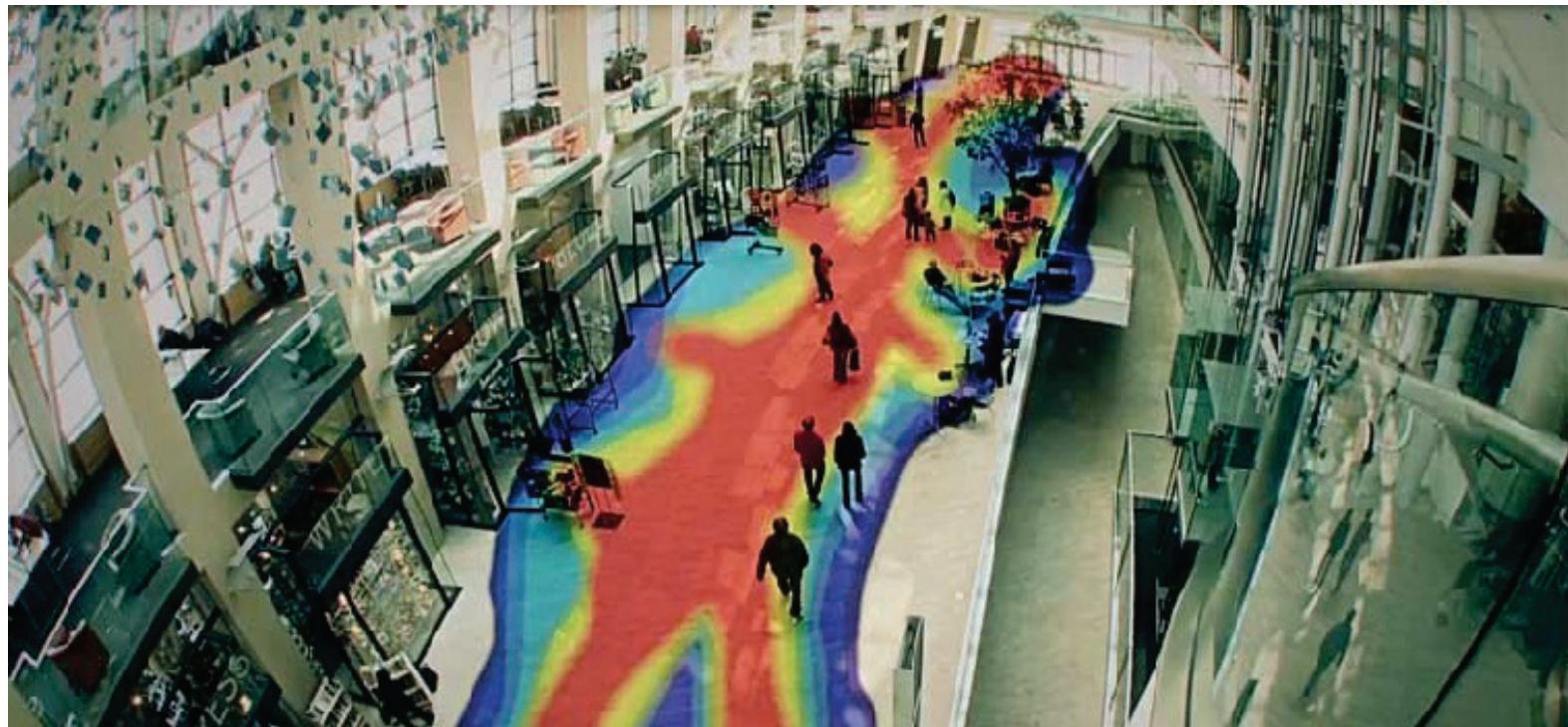
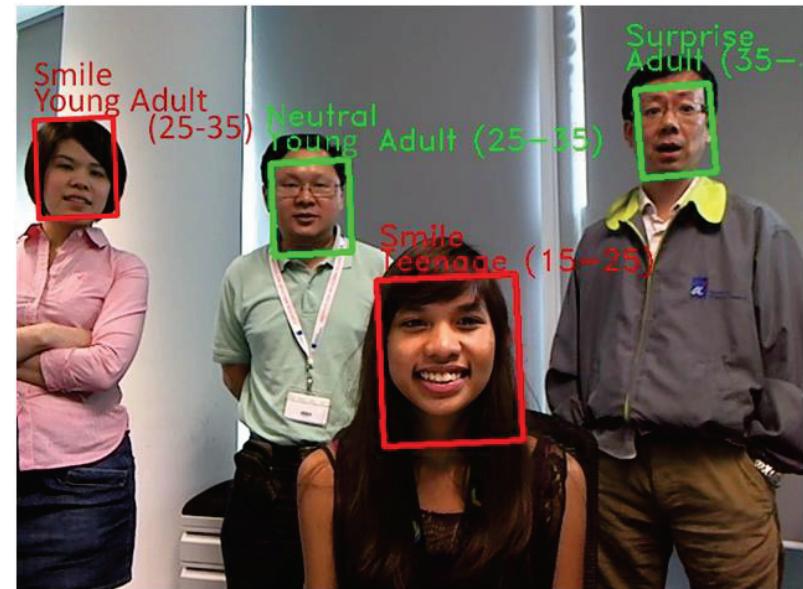


The Amsterdam Real Time Project



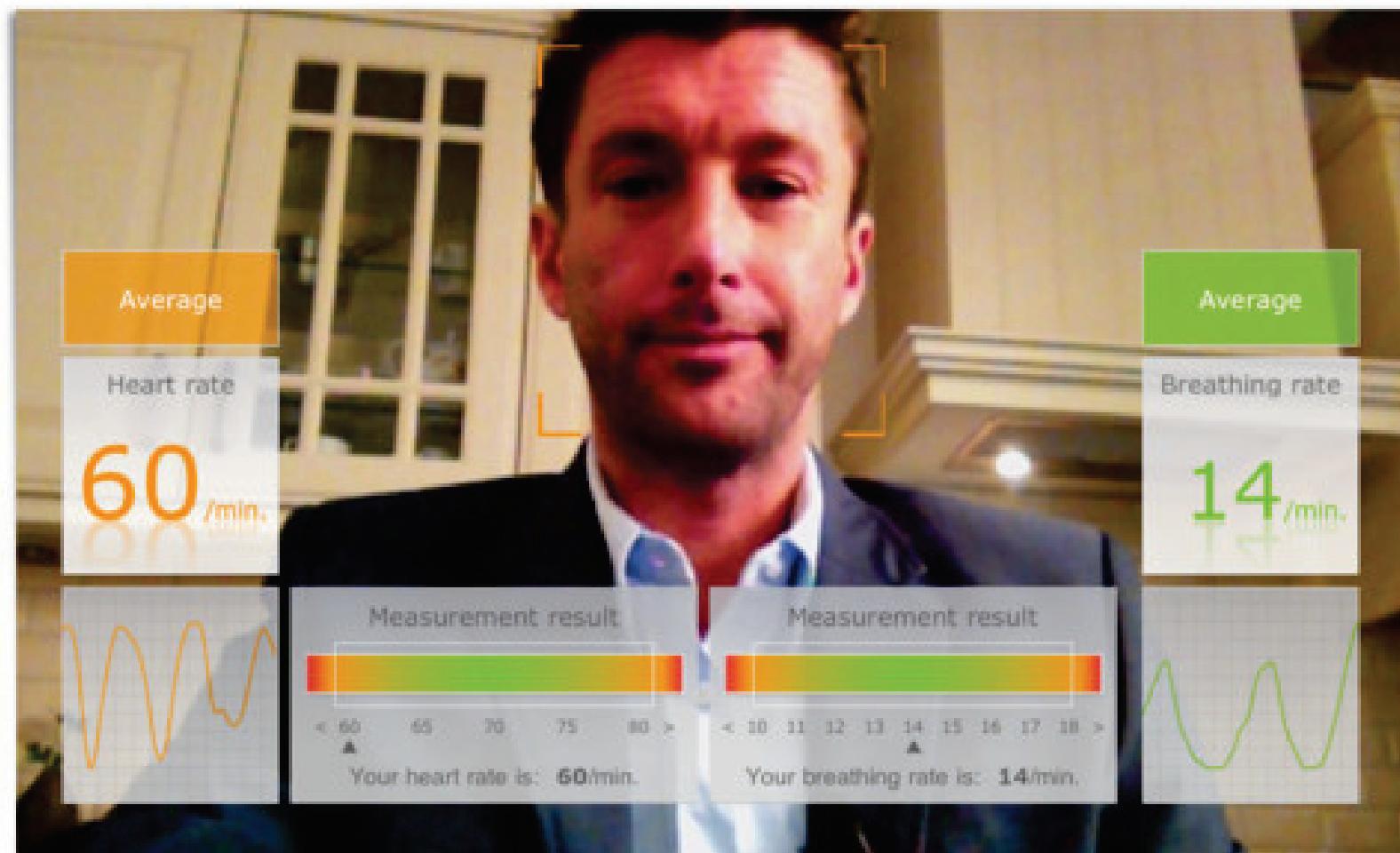
Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-time urban monitoring using cell phones: A case study in Rome, *IEEE Transactions on Intelligent Transportation Systems*, 12(1), 141-151.

5 - Tills



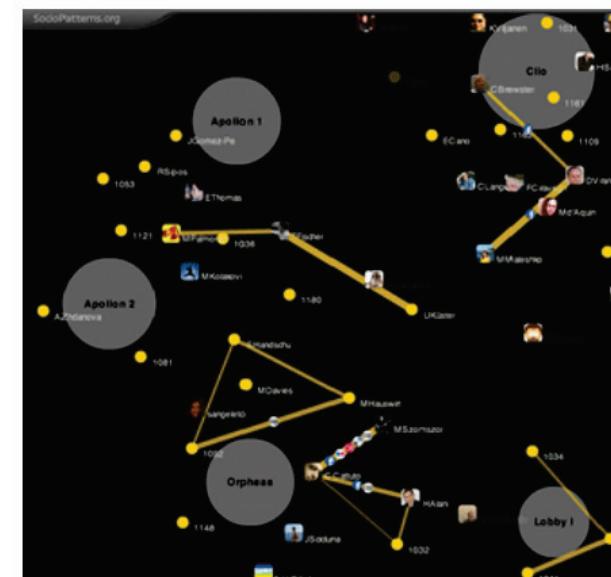
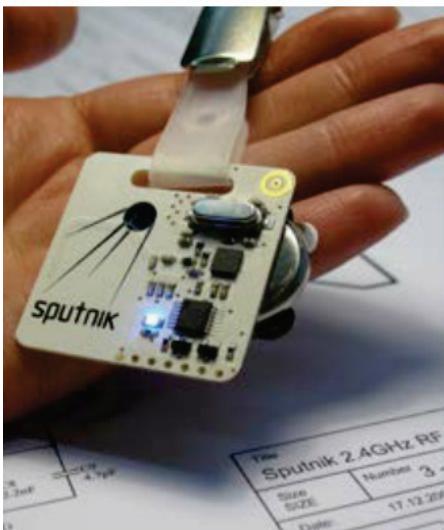


(a) Input



New kinds of data

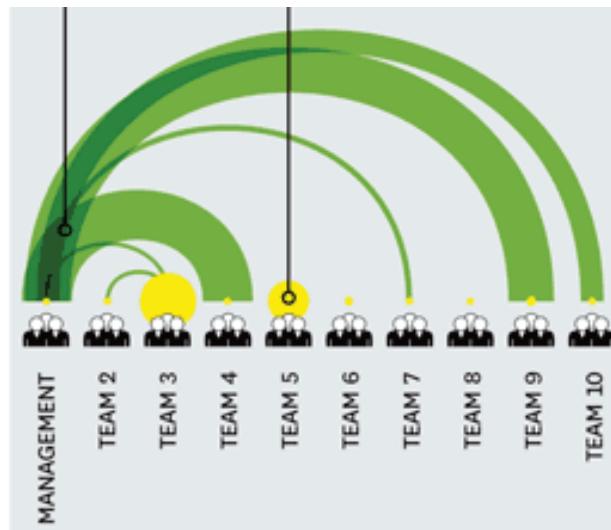
Social Sensing via RFID



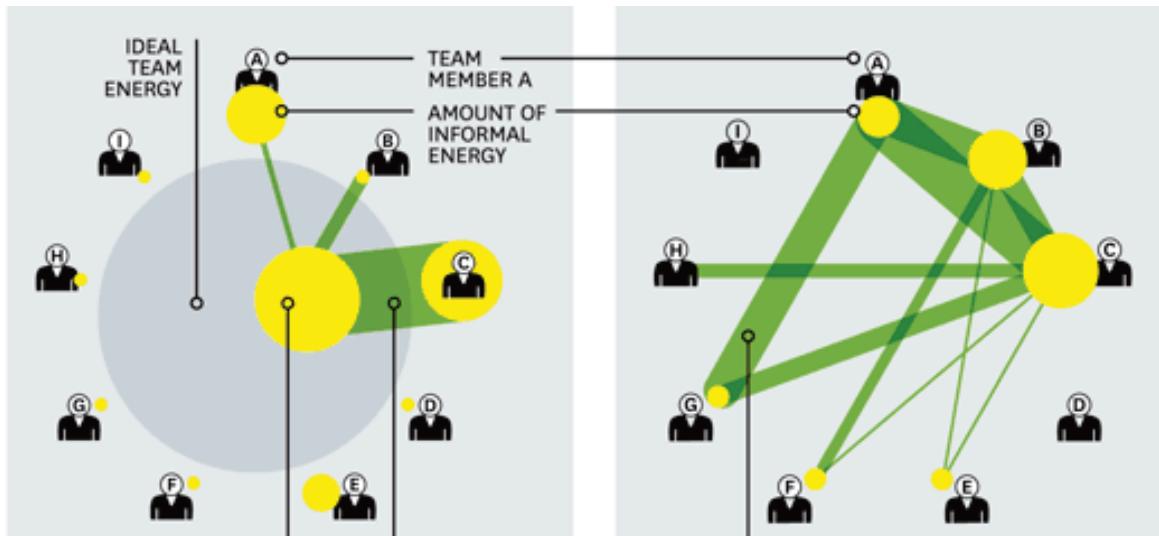
SocioPatterns

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 (Foursquare, Gowalla, Twitter, ...)

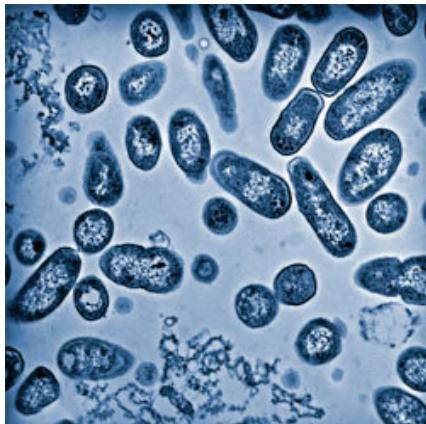
Cheng, Zhiyuan, et al. "Exploring Millions of Footprints in Location Sharing Services." ICWSM 2011 (2011): 81-88.

Computational Social Science

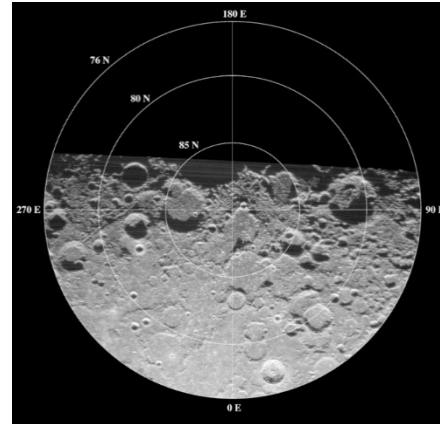
The science that investigates social phenomena through the medium of computing and statistical data management and processing.

Computational Social Science

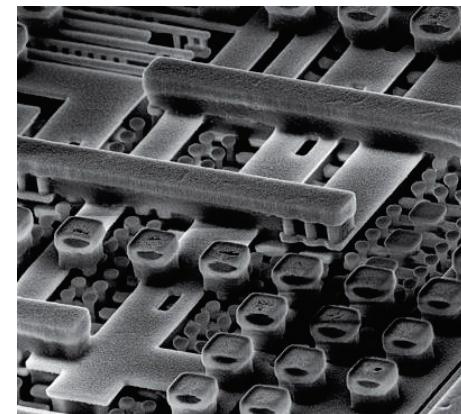
- An instrument-enabled scientific discipline



microbiology
microscope



radio astronomy
radar



nanoscience
electron microscope







**USING IT
IS THE HARDEST PART.**



Technical Challenges

- Computational infrastructures for dealing with
 - **More data:** analyzing large amounts of data
 - **Fuzzy data:** cleaning up imprecise 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



Macroscopic

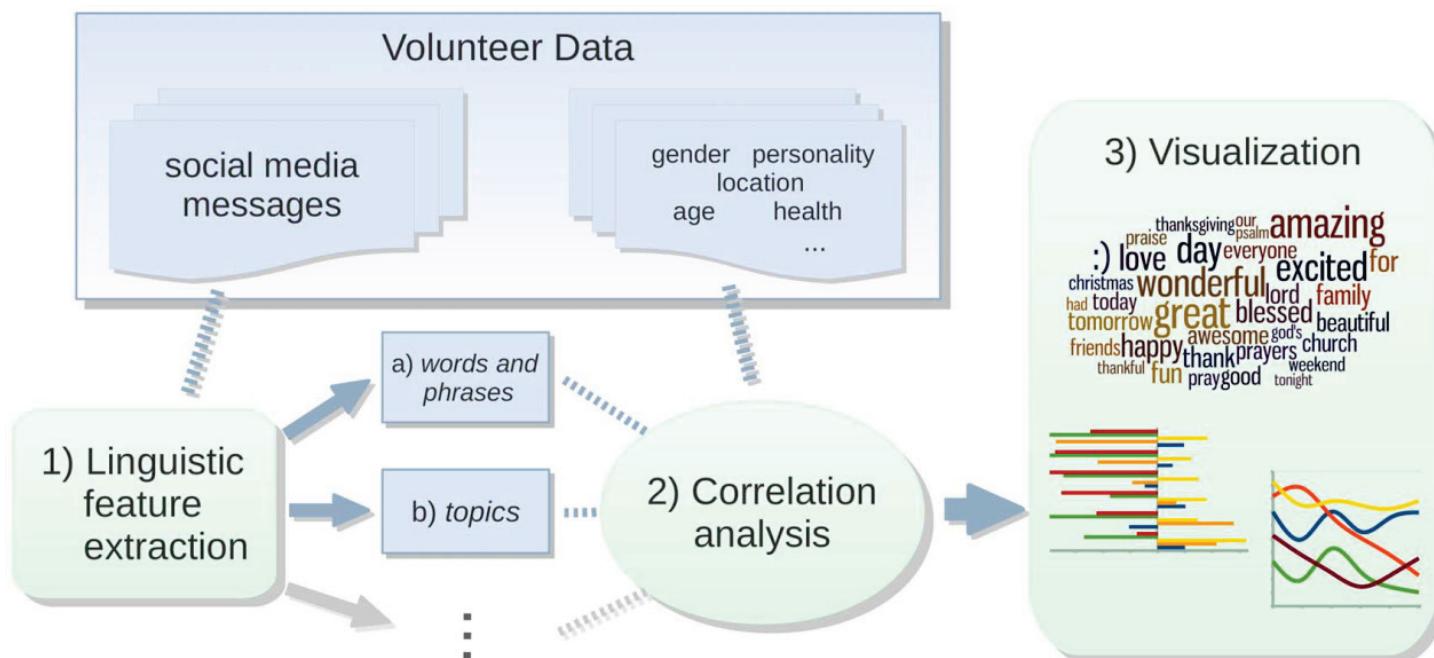
Linguistics

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



relative frequency



correlation strength

male



prevalence in topic

How Old Are You? (#1)

13 - 18



19 - 22

How Old Are You? (#2)

23 - 29



30 - 65



Personality Traits

Extraversion

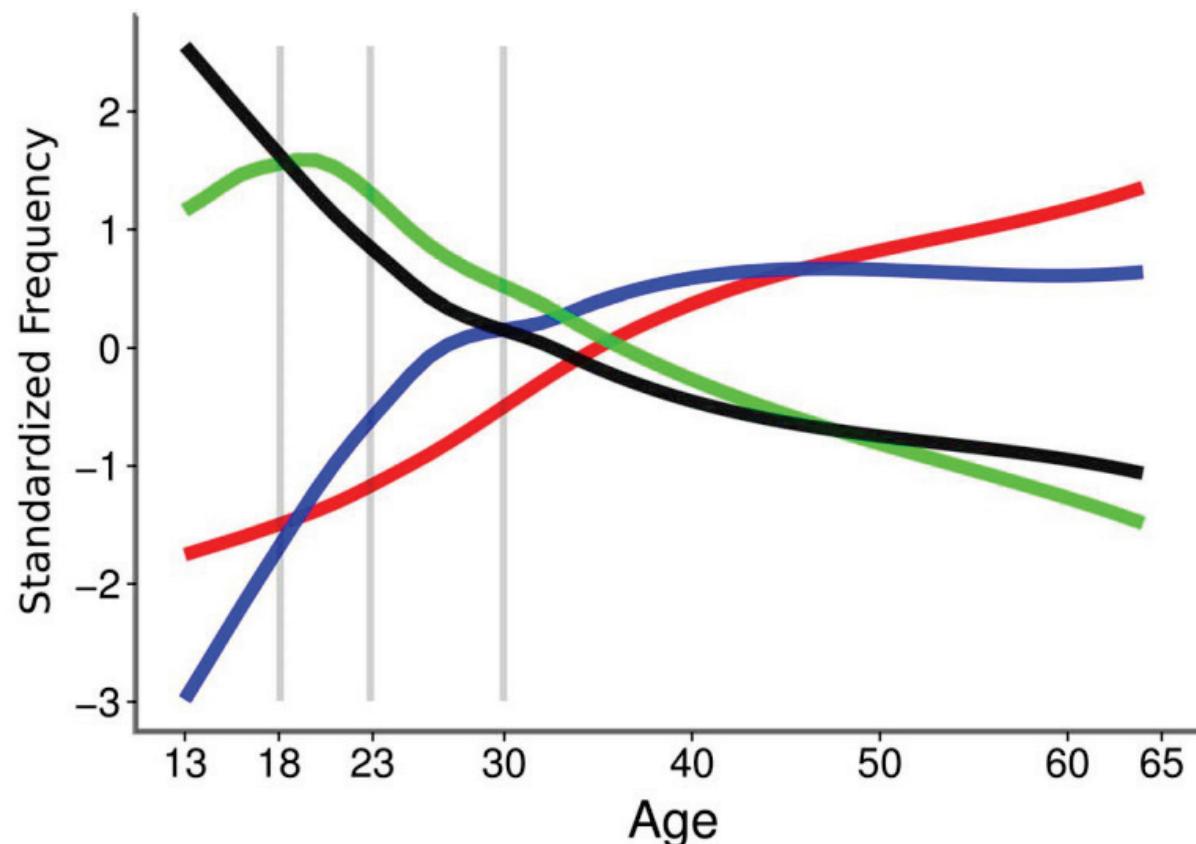


Introversion

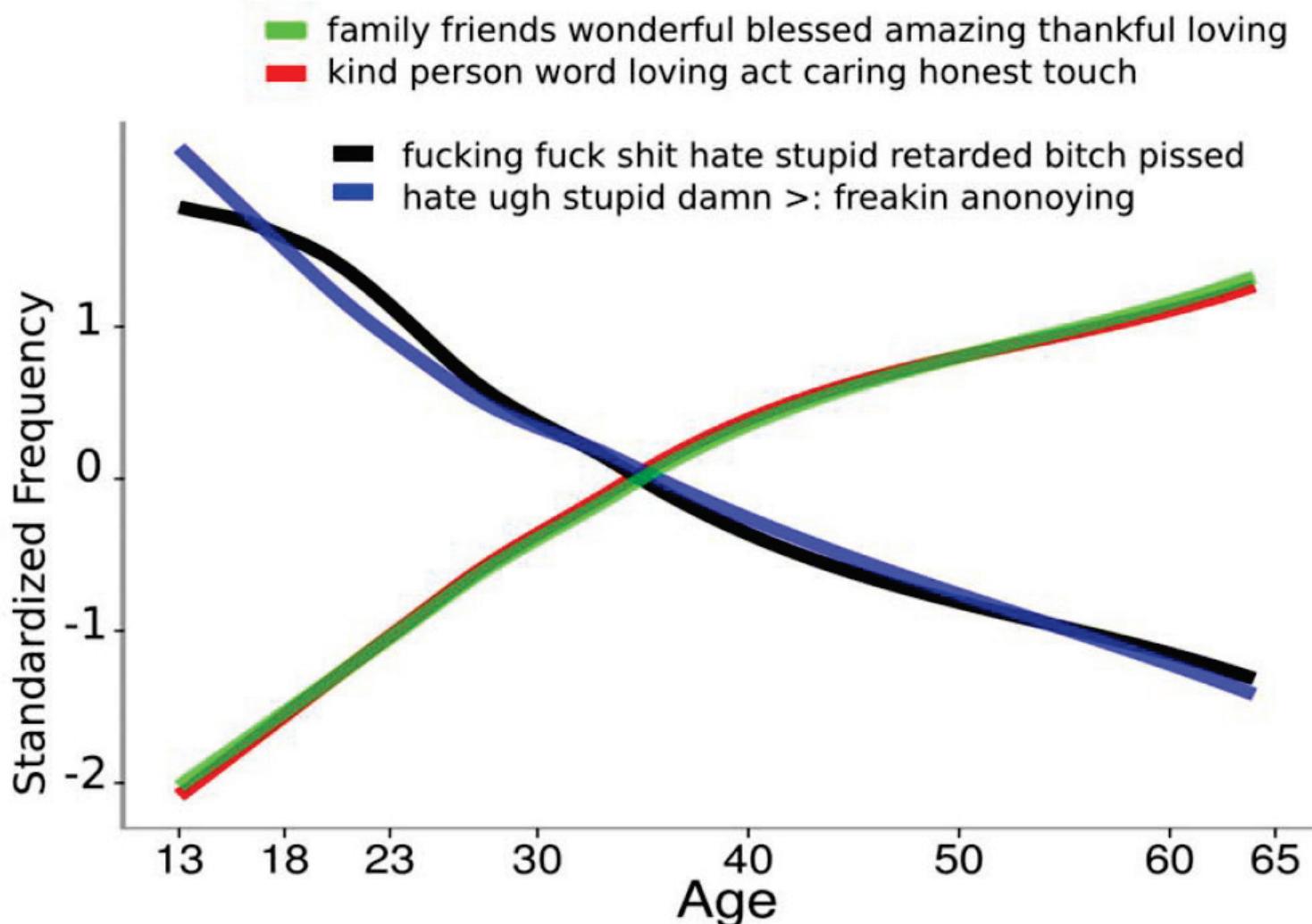


Topics Across 4 Age-groups

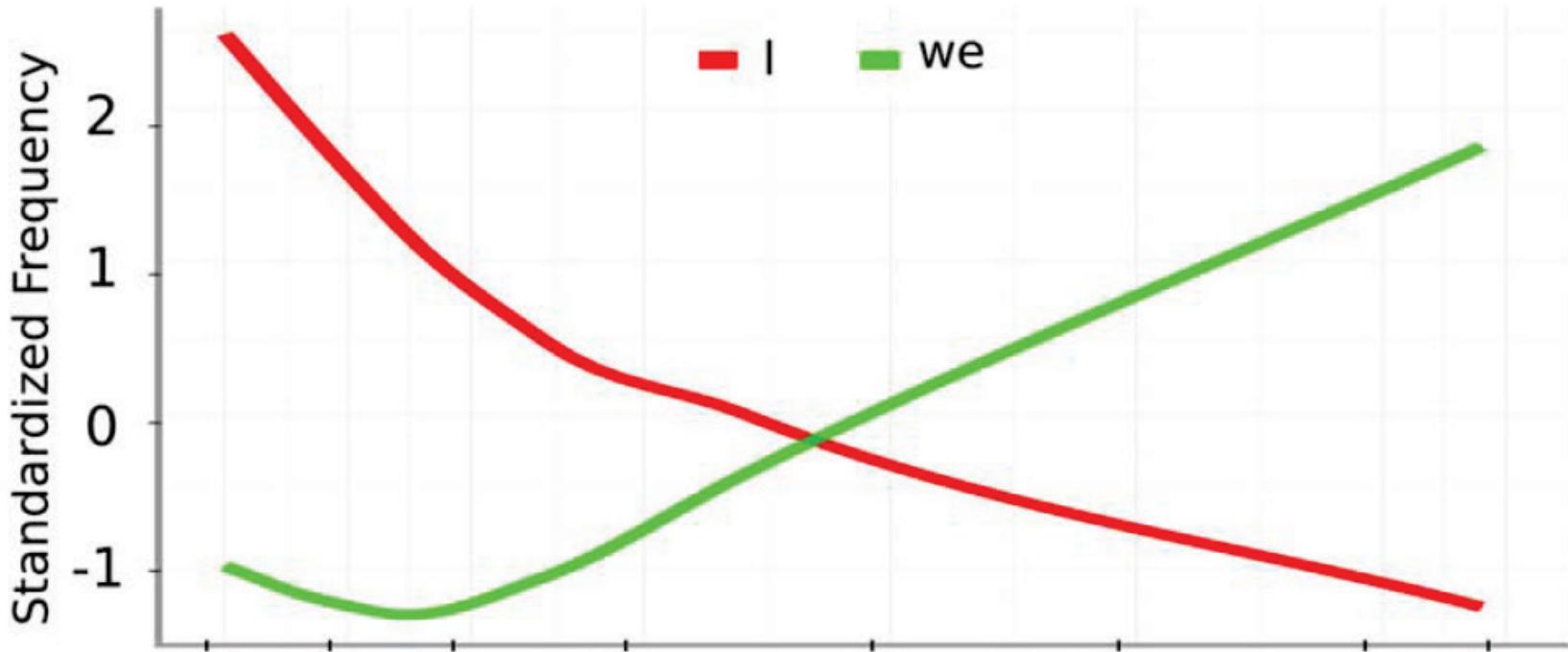
- (30 to 65) son daughter father mother proud oldest data youngest
- (23 to 29) job position company manager interview experience office assistant
- (19 to 22) classes semester class college schedule summer registered taking
- (13 to 18) haha lol :p :D ;) hehe jk ;p



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





Virtual Lab

Social Psychology

MOOD CONTAGION (& MANIPULATION) ON FACEBOOK

Kramer, Adam D.I., 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)



A screenshot of a Facebook news feed illustrating the mood contagion experiment. At the top, a user's post reads "Feeling Happy" with a smiling emoji and the caption "feeling happy". Below it, another user's post reads "Sadona Cassedona" with a sad emoji and the caption "feeling angry". A third user's post below that reads "This is CRAZY-!" followed by a long, irrelevant political rant about House Bill -195. A "Like" button is visible at the bottom left.

Feeling Happy 😊 feeling happy

May 30 ·

Happiness
Laying by
With my b

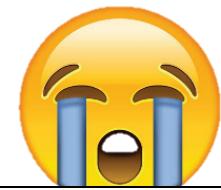
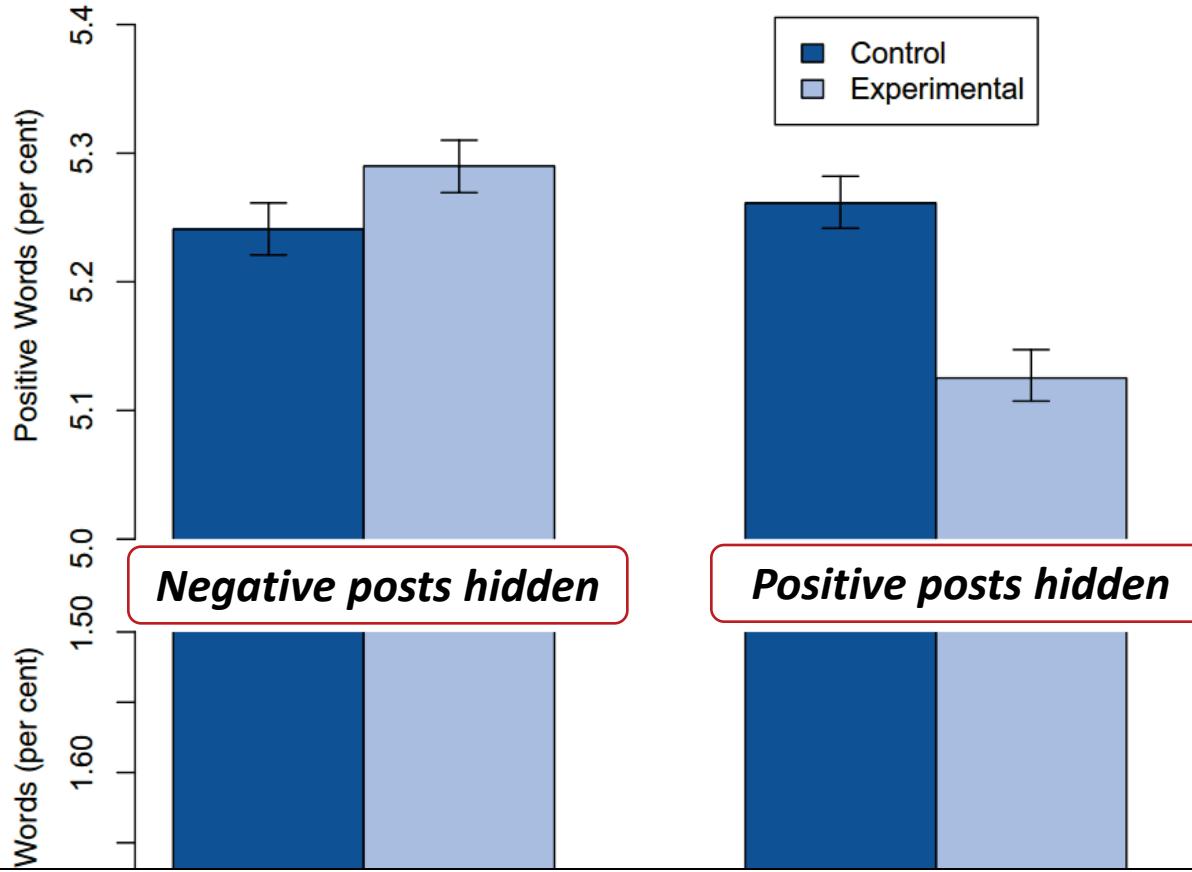
Sadona Cassedona 😢 feeling angry

13 hrs ·

This is CRAZY-!
This 'House Bill -195' makes it illegal to buy second hand items, No Matter
What They Might Be, for CASH-I-I-I!
YES you read this correctly! !
What happened to the adage that: "CASH IS KING"-I ? ? ?

Like

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)

Today is Election Day

What's this? • close

Informational

 Find your polling place on the U.S. Politics Page and click the "I Voted" button to tell your friends you voted.

01155376 People on Facebook Voted

Social

Today is Election Day

What's this? • close

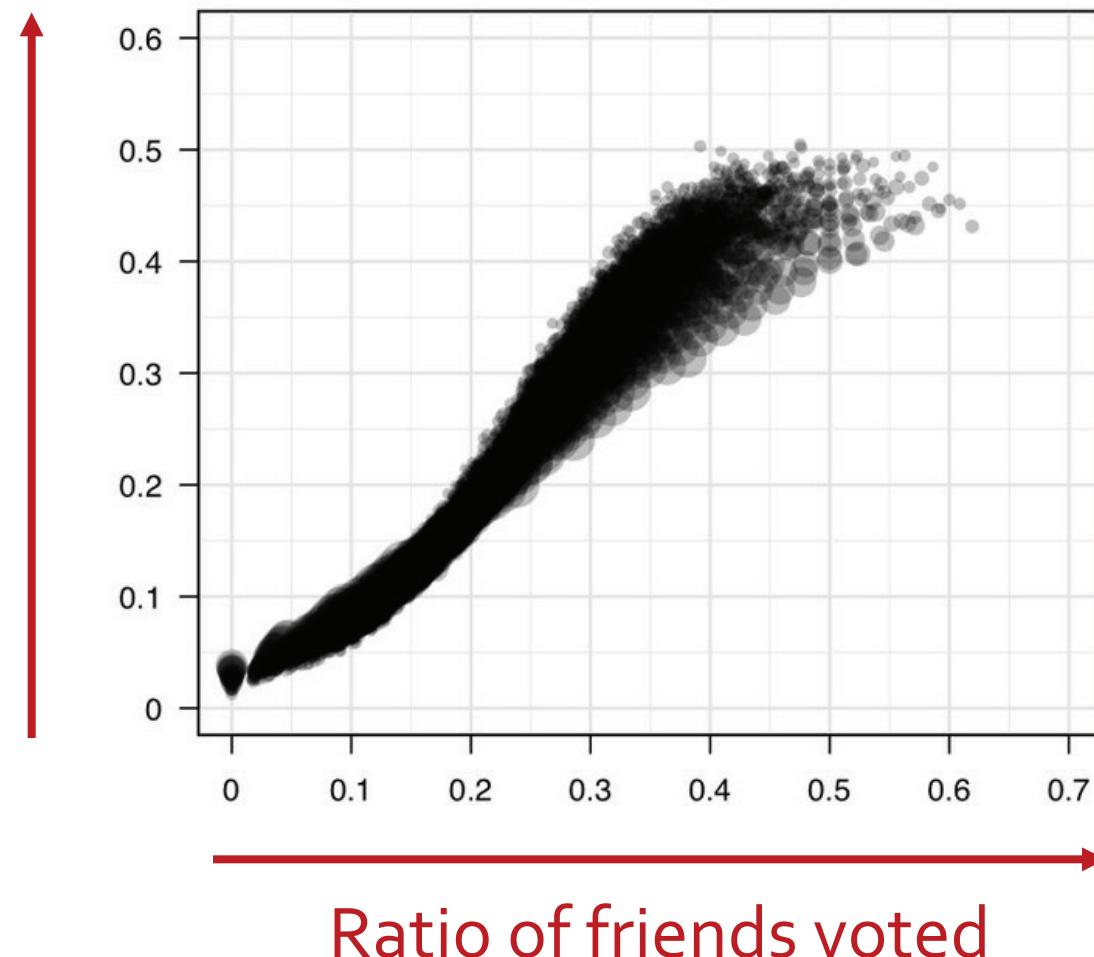
 Find your polling place on the U.S. Politics Page and click the "I Voted" button to tell your friends you voted.

01155376 People on Facebook Voted

 Jaime Settle, Jason Jones, and 18 other friends have voted.

Effect of Manipulation

Prob. of oneself claimed voted

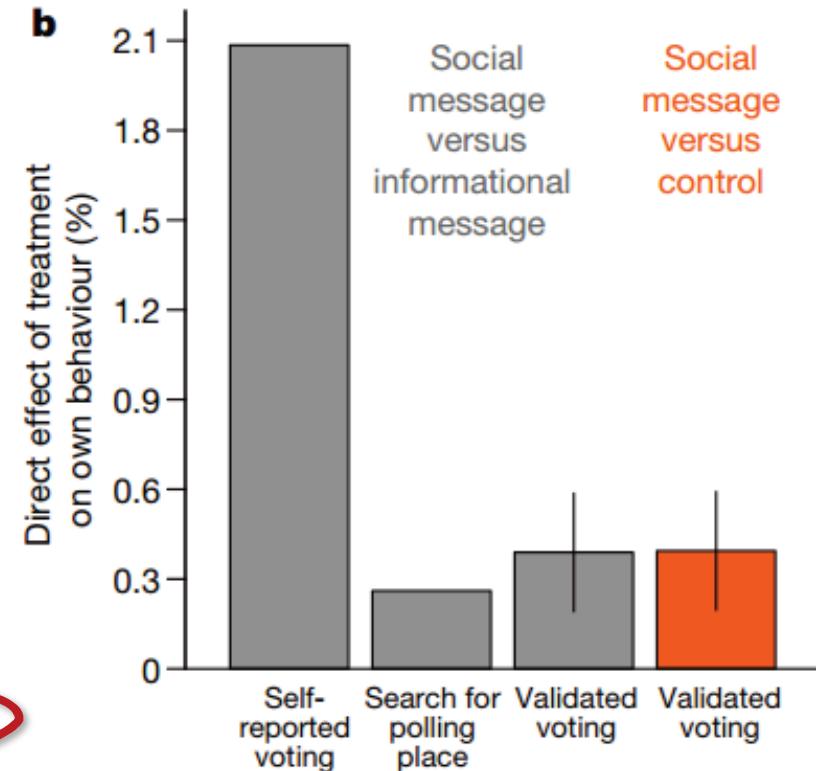


a

Informational message



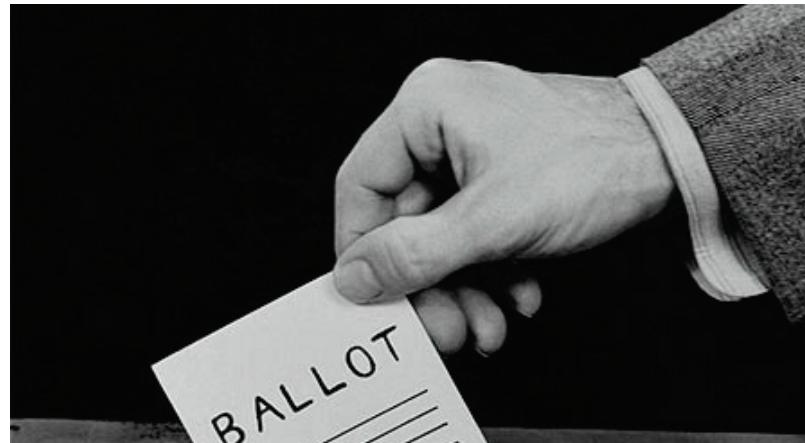
Social message

**b**

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

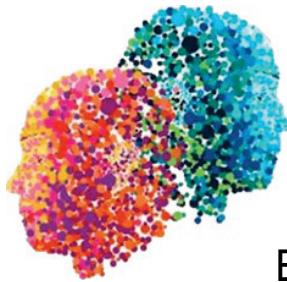


Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer²,
Mark S. Smolinski¹ & Larry Brilliant¹

¹Google Inc. ²Centers for Disease Control and Prevention

Nature 457, 1012-1014 (2009)



Empirical Modeling

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.

Datasets

■ Heart disease

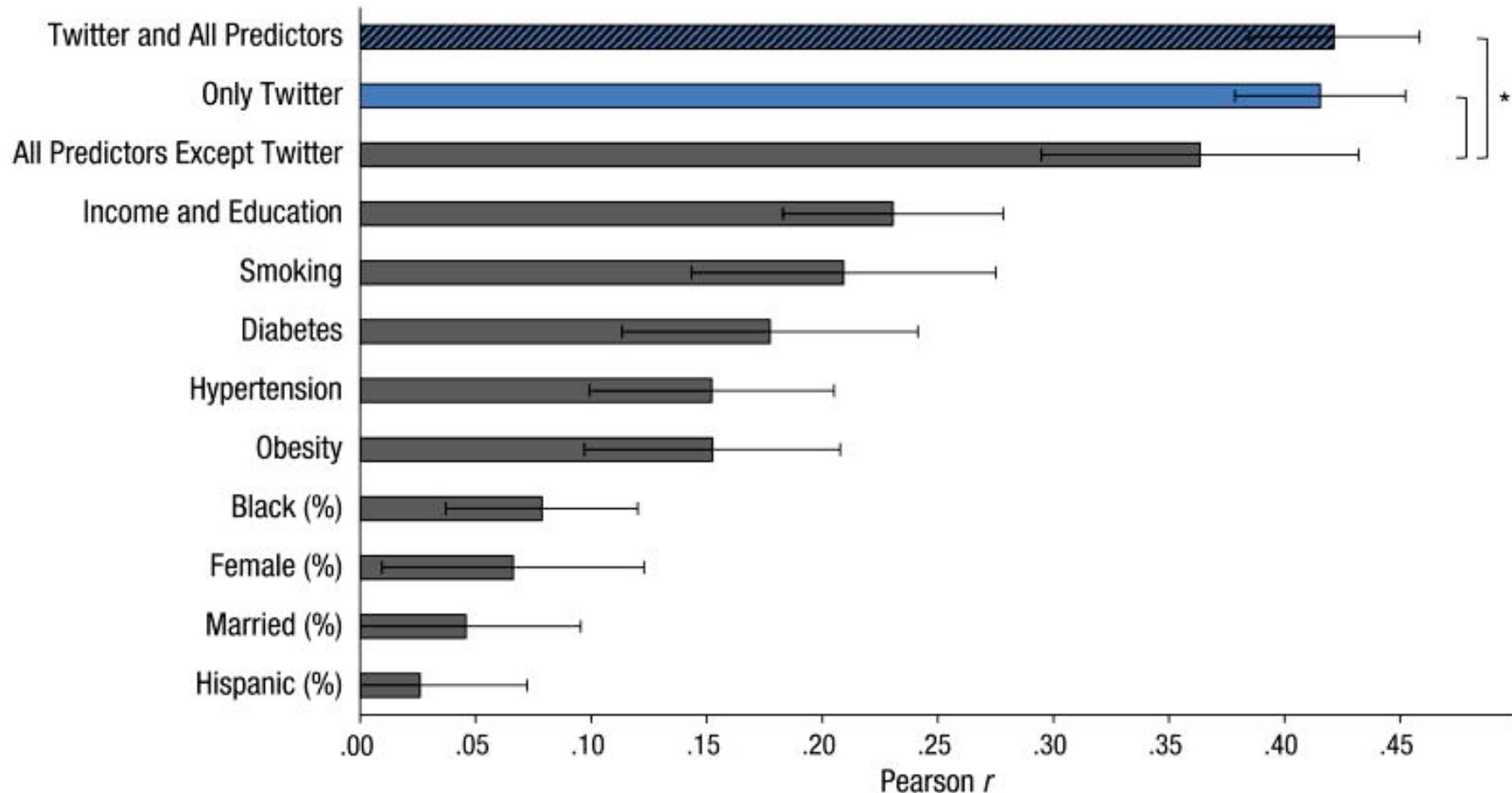
- Arteriosclerotic heart disease mortality rates during 2009 -- 2010



■ 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

bullshit
shits fuckfuckin
bitches damn fucked
fucks fucking bitch
shit shitty ass dude
pissed

$r = .18$

Hate,
Interpersonal
Tension

jealousy mad
bitches
envy hate jealous
hating haters
lovers famous hatin
hater phase ya'll
hated

$r = .16$

Boredom,
Fatigue

sooooooo
boring text hmu
entertain insanely
yawn entertainment
extremely bored stiff
boredom entertained
incredibly bore

$r = .18$

Hostility,
Aggression

bullshit
shits
bitches
fuck
fuckin
damn
fucked
fucks
 fucking
ass
shit
shitty
dude
pissed

$r = .18$

Hate,
Interpersonal
Tension

jealousy
mad
bitches
envy
hate
jealous
hating
haters
lovers
famous
hater
phase
hated
ya'll

$r = .16$

Boredom,
Fatigue

sooooooo
boring
text
entertain
insanely
yawn
entertainment
extremely
bored
boredom
incredibly
entertained
bore

$r = .18$

Skilled
Occupations

skills
development
information
design
management
process
marketing
communication
business
learning
technology
engineering
education
analysis

$r = -.14$

Positive
Experiences

changing
wonderful
experienced
judgment
enjoyable
journey
experiences
exciting
learning
painful
experience
pleasant
share
bound

$r = -.14$

Optimism

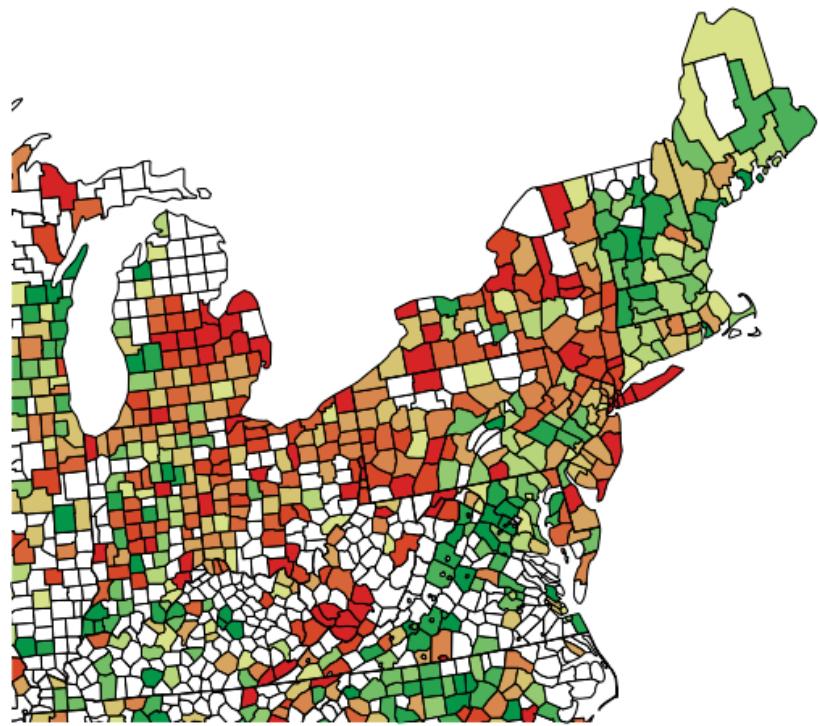
opportunity
possibilities
talents
opportunities
discover
possibility
);
challenge
improve
create
endless
experience
potential
ability
explore

$r = -.12$

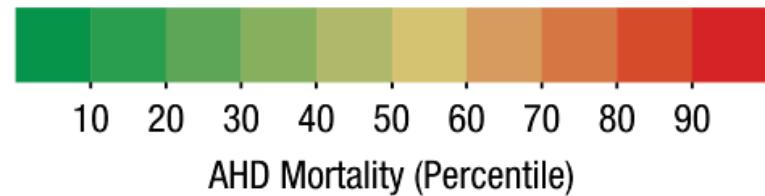
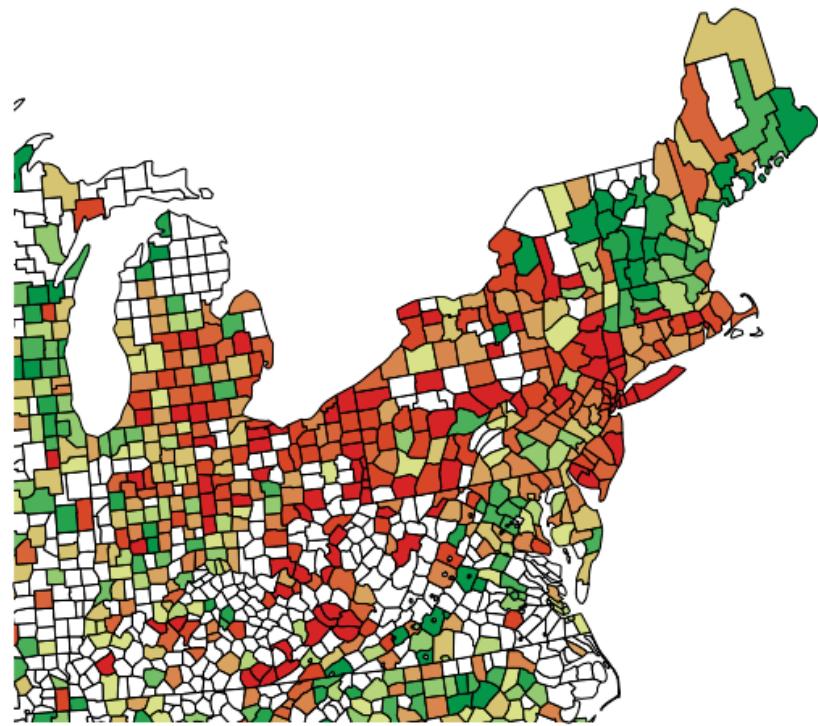
Language Use in Tweets

Language variable	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]**

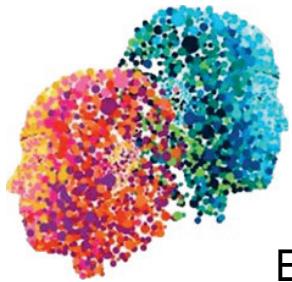
CDC-Reported AHD Mortality



Twitter-Predicted AHD Mortality



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



Empirical Modeling

Social Psychology

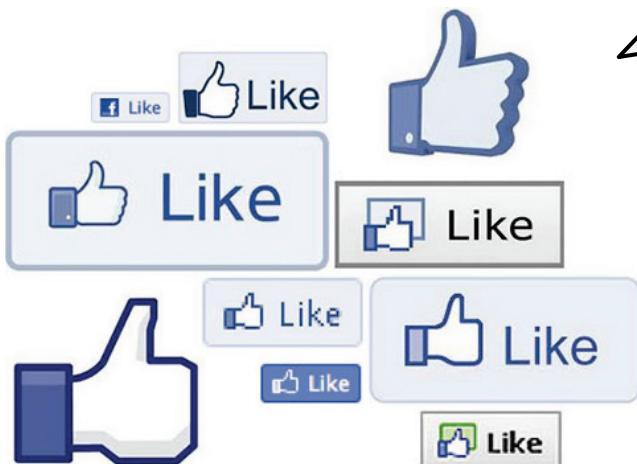
YOU ARE WHAT YOU LIKE

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.	<input type="radio"/>				
I am always prepared.	<input type="radio"/>				
I get stressed out easily.	<input type="radio"/>				
I have a rich vocabulary.	<input type="radio"/>				
I don't talk a lot.	<input type="radio"/>				
I am interested in people.	<input type="radio"/>				
I leave my belongings around.	<input type="radio"/>				
I am relaxed most of the time.	<input type="radio"/>				
I have difficulty understanding abstract ideas.	<input type="radio"/>				
I feel comfortable around people.	<input type="radio"/>				
I pay attention to details.	<input type="radio"/>				
I worry about things.	<input type="radio"/>				

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.

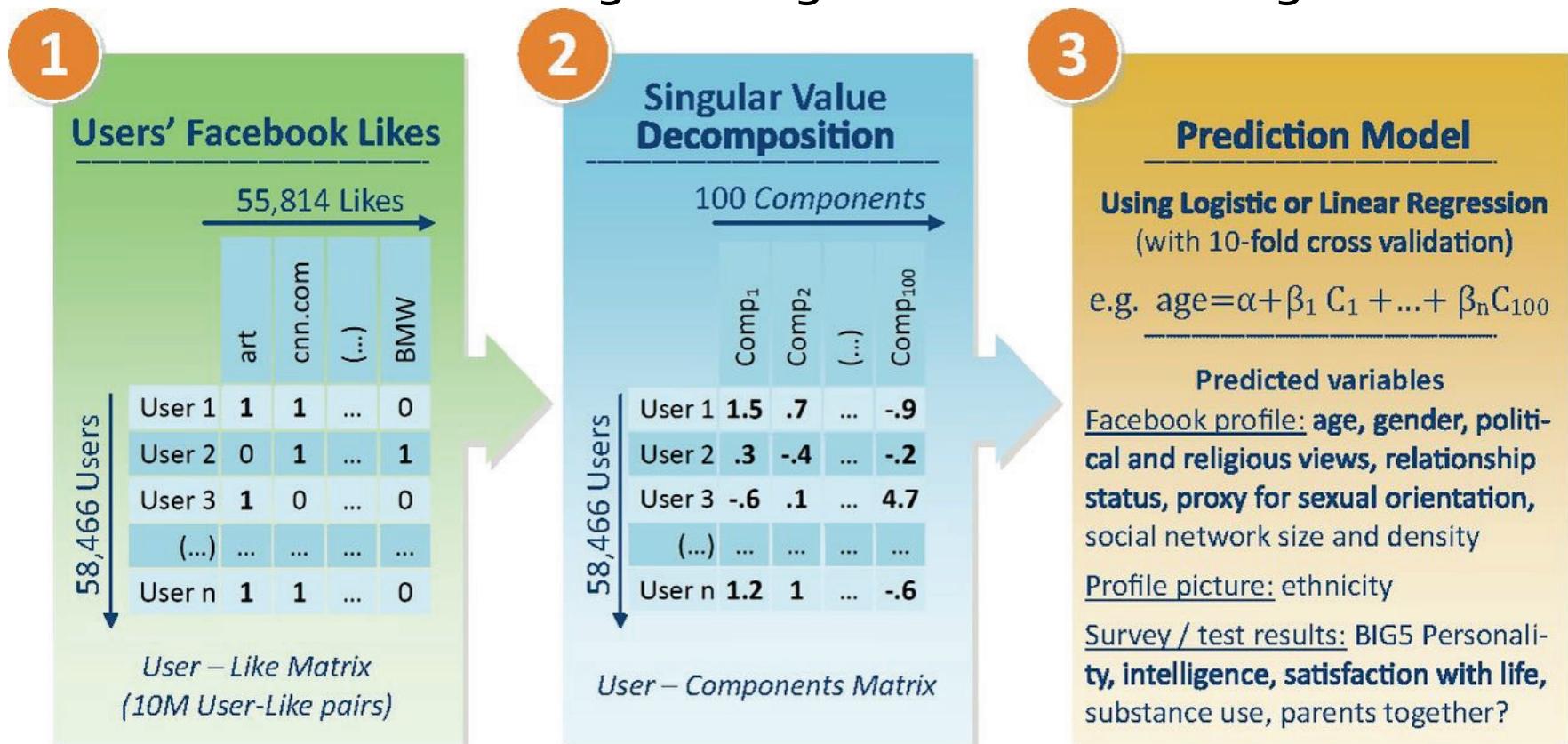
So far I have gotten the important things I want in life.

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

- 31 - 35 Extremely satisfied
- 26 - 30 Satisfied

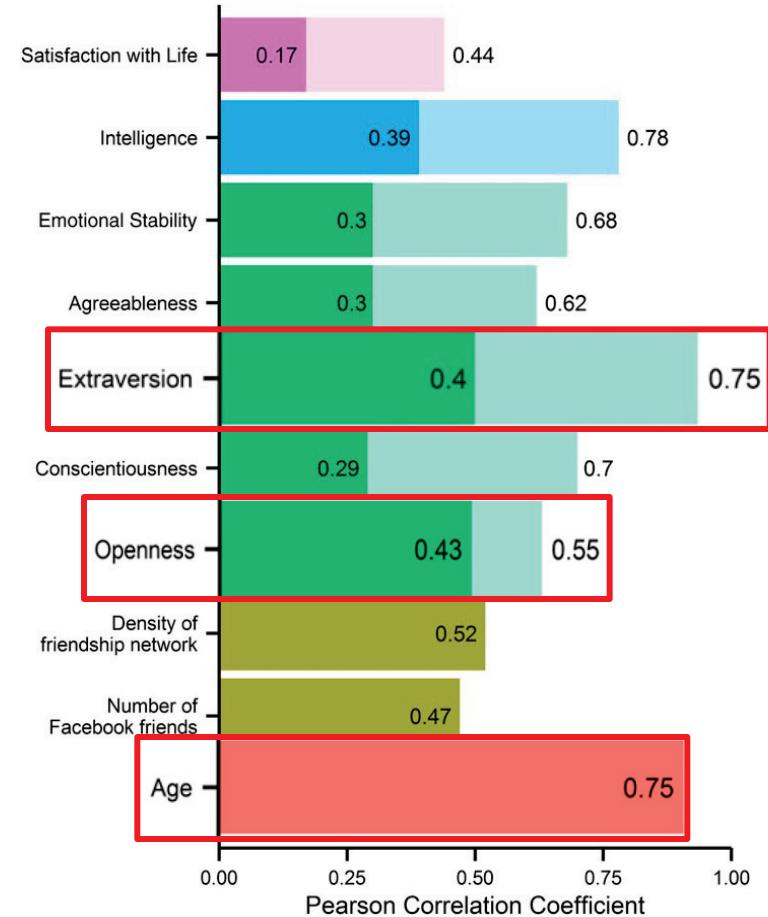
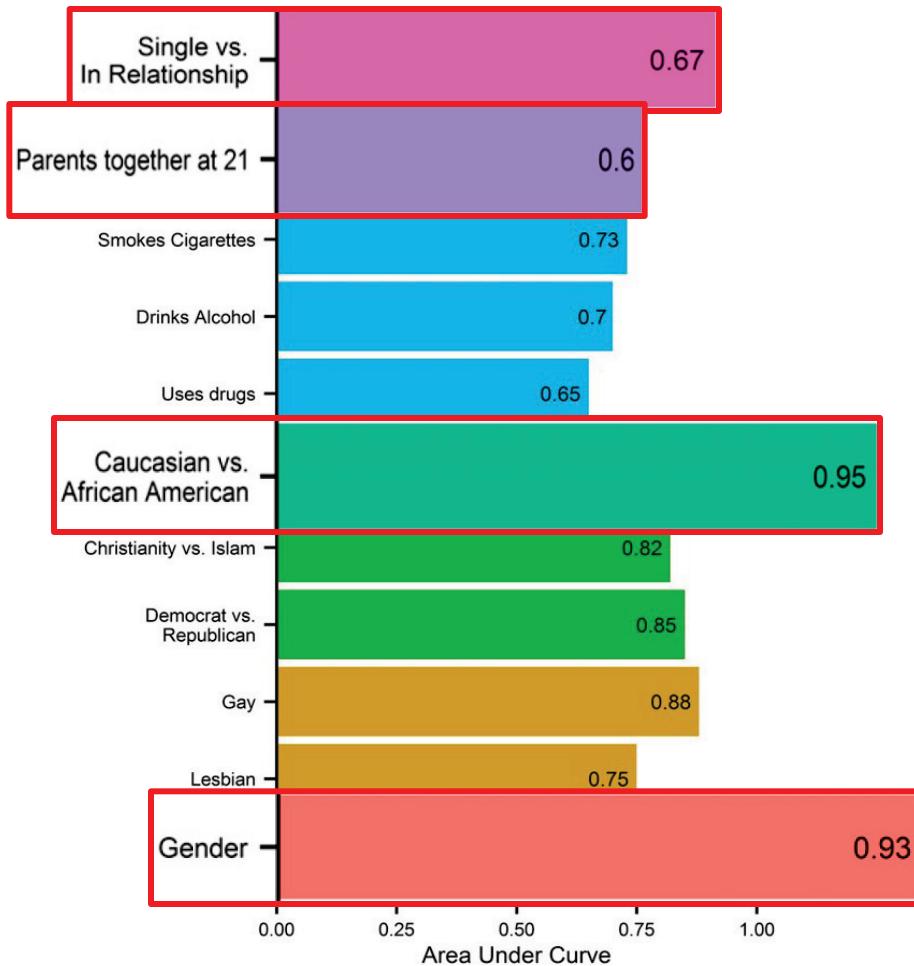
Methodology

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

Discriminative Likes (#2)

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

Discriminative Likes (#3)

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

Li

打鈎

✓ 接收通知
✓ 暫示在螢幕消息中
設定
新增剝與權主題清單.....



卡提諾正妹抱報
Entertainment Website

Sign Up Liked Message ...

Timeline About Photos Likes Videos

#1)

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 >



抬起頭你說色狼

Post Photo / Video

Write something...

Post

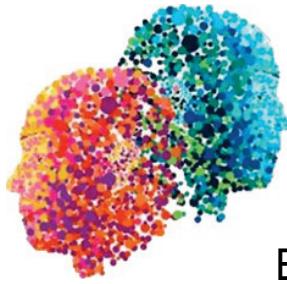


女

Likes are Culture-Dependent (#2)

已婚	嬰兒與母親 懷孕生產情報站	學生愛打工	未婚
	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



Empirical Modeling

資料科學如何幫我們更瞭解 捐款人？



x 3,518

**in 10.5 years
(since May 2003)**

獨養3殘障兒 病母：要撐下去

基金會編號：A3744

45歲的劉玉娟陪她極度多障大女兒阿珊穿支撐背架，23歲的阿珊不停發出「嗚咽」聲，劉玉娟安撫：「等等，快好了。」客廳裡，還有她兩個將升國一的大兒子小祥和小五小兒子小小，

劉工指點說，今年初51歲牛生阿廣沒留下半句話，突然心肌梗塞發作猝逝，留下女兒謹此叩頭。

「長大會照顧你」

雖然子女各有我況，但劉玉茹說，孩子慢傳教都有進步，人文系阿嬤先天聽及舌障，並動手，歷多次手術，心智如四歲女童，生理恢復發育有感童真，「以前我告她到唱歌學校讀到特教高中畢業，她會比手語上用嘴、吃飯，我跟她都頗汗顏。」

劉玉娟又說，大兒子原是中度智障，從3歲上課到小學資深班，前年鑑定已改善度；有先天代謝極度異常的小兒子小歲到2



A3653 黃清貴 結案報告

夫宿逝

三女顯靈供訖萬口

【李佳玲／高雄報導】5月7日《联合报》报导指出，台湾地区今年1至4月的外汇存底增加13.8%，达1031亿美元，其中美元占821亿元，占80%以上。此数据被指虚报。

捐 詢 明 紙

見 ■ **女頭**
32-050-00-00



的表女兒（右二）「最讓我擔心。」



■劉玉娟（右二）帶2子住院照顧大女兒
阿珊（右一）做脊椎側彎手術。

劉玉炳的7旬老媽過，前兩年搬進她，
並奉點養料阿媽，如今及她，「是引外孫女
阿昭住院，我奶奶到來吃陳地（台語：沒
法收養法理），我老人家已不知怎麼辦才
好？」

捐 款 人 資 料	蘋果日報慈善基金會	
● 指定信用卡捐款：請填妥表格，寄回或傳真至蘋果日報慈善基金會，欲進行國稅局連線報稅請填具身分證字號。		
姓名：	電話：	
地址：		
信用卡持卡人姓名：		
信用卡號碼：		
有效日期至：		
捐款數額NT\$：	專金會編號：	
主效簽名：		
● 電子支票：請按「預未日報慈善基金會」		
地址：(114)台北市內湖區興安路141巷32號		
● 郵政劃撥：請洽詢蘋果日報慈善基金會		
表單： <input checked="" type="checkbox"/> 郵局： <input type="checkbox"/> 二每次捐款立即扣款 三年合併一張存摺 二不要寄收據給我 三採用網上匯款，不寄收據 身分證字號		
● 電話：(02)6601-6407、0609-000865 ● 電郵：http://applefbg.km03333.com.tw ● 吉祥號：(02)8801-6999、0609-000836		

AppleDaily Charity Case Dataset

3000+ cases along with detailed description and donation records

2幼子畫圖為癌父打氣	2012/09/19	已結案	445992
單親媽罹癌 豁9歲女成孤	2012/09/18	已結案	761857
幼女幫灌食 癌父「不能倒」	2012/09/17	已結案	474141
38公斤病婦 憔悴等換心	2012/09/16	已結案	990050
單親母突癱「怎討生活」	2012/09/14	已結案	518607
夫罹癌 孕妻挺肚拼家計	2012/09/13	已結案	497717
雙親病倒 獨生女苦打家	2012/09/12	已結案	535150
單親母罹癌 豁4兒無依	2012/09/11	已結案	970125
憨女顧癱母「要乖乖的」	2012/09/10	已結案	516438
夫脊傷妻肝病 像天塌下	2012/09/07	已結案	634566
單親母罹癌 豁稚女失依	2012/09/06	已結案	644243
夫驟逝 病妻打工養幼女	2012/09/05	已結案	612903
夫癱兒瞎獨 病婦頓失依	2012/09/04	已結案	487075
慈父癱2孝子日夜照護	2012/09/03	已結案	504805
孝子罹癌 挂心憨弱母兄	2012/09/02	已結案	1224056



賴曉芝（左）不知何時可換心，「女兒怕我忘了吃藥，會一直提醒我。」

開心，對單親媽賴曉芝如今來說不容易，「讀小二的大女兒跟我說，媽媽，好久沒看到妳笑了，但我想自己病一天天加重，兩女兒都小，心裡難過。」開心活命，對31歲的她是必要的，醫師評估若她不能換心，可能僅剩1年時間。

報導、攝影／向高彬

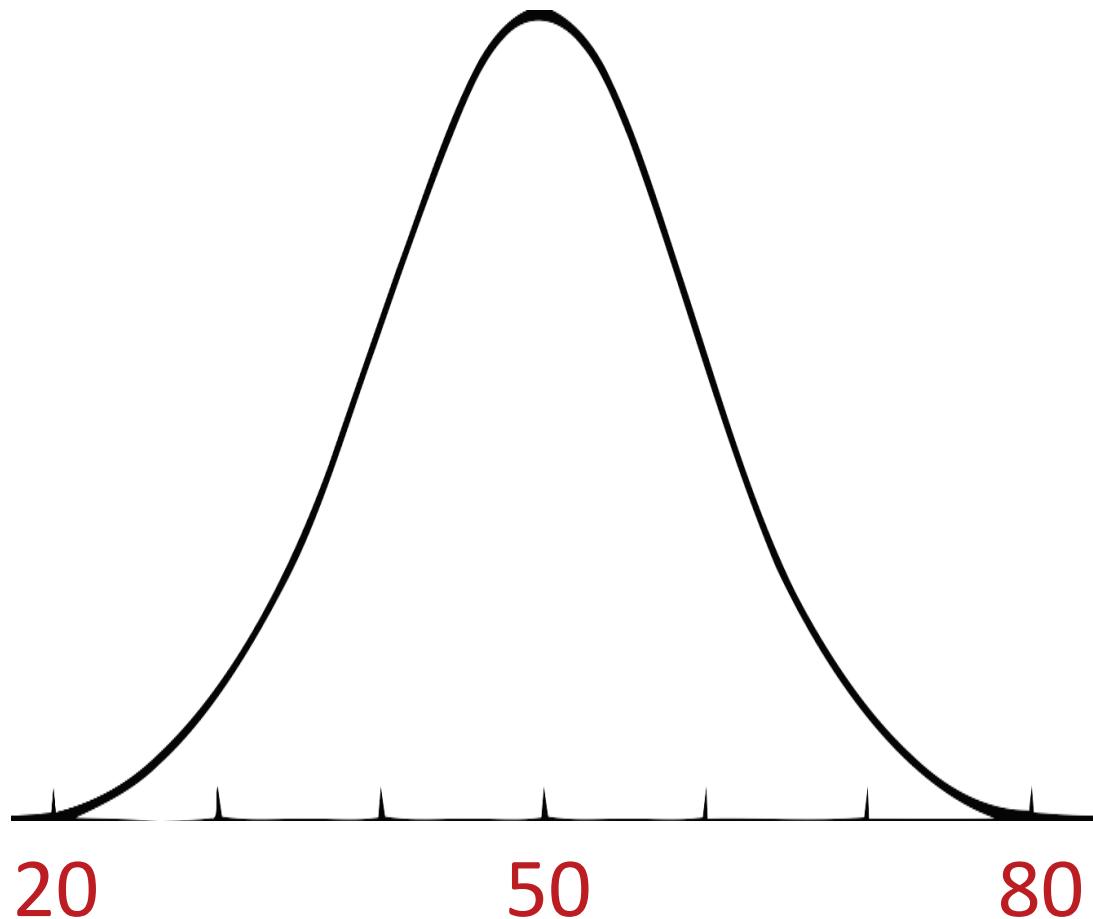
近期住院等待換心的賴曉芝模樣憔悴，體重剩38公斤。她說：「現在只能一天等一天，等待醫院移植通知的奇蹟出現。」

20萬手術費難籌

賴曉芝說，前年不時胸悶疼痛，起先她靠吃成藥止痛，去年4月突然休克，轉診大醫院查出有心臟瓣膜問題，手術後返家休養，按時

捐款人姓名	累計(元)	捐款明細
劉鈺笙	20000	2012/9/14
王白祿	10000	2012/9/17
方育哲	10000	2012/9/27
宣松建材企業(股)公司	6000	2012/9/21
周珮詩	5400	2012/9/13
陳勝和	5000	2012/9/13
吳增榮	5000	2012/9/13

捐款金額分布 (每戶個案家庭)





A3770	<u>男半癱腦損 看影片認兒</u>	2015/10/09	已結案	?
A3768	<u>夫癱兒浙 嫩送報養2孫</u>	2015/10/07	已結案	?
A3767	<u>稚兒畫卡片 為癌父加油</u>	2015/10/06	已結案	?
A3766	<u>養殘妻稚女 脊傷男拼復健</u>	2015/10/05	已結案	?
A3765	<u>壯漢癱如嬰兒 癌母把屎尿</u>	2015/10/04	已結案	?
A3764	<u>單親癌爸皮包骨 就醫難</u>	2015/10/02	已結案	?

DATA COLLECTION

Crawling

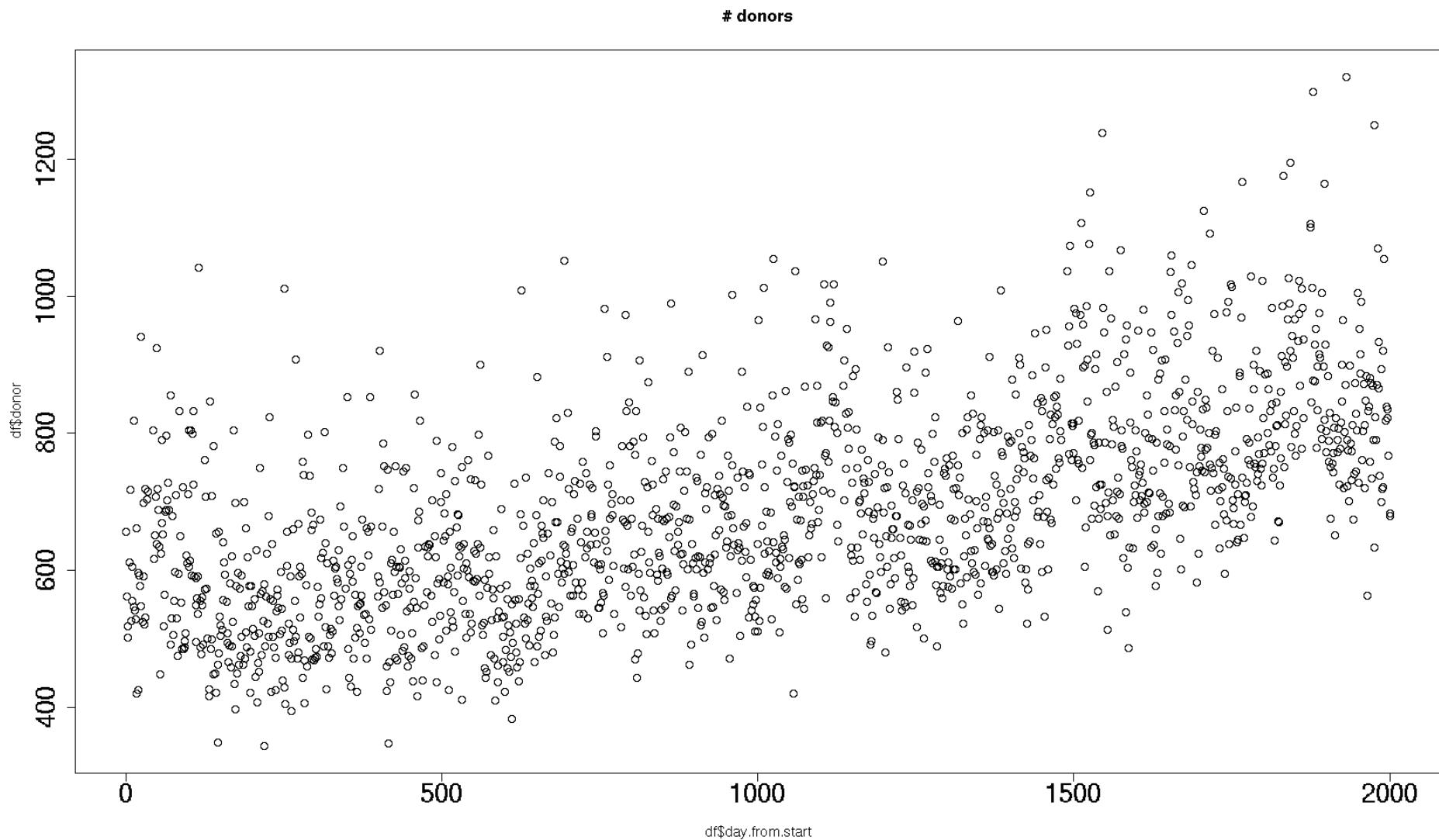
捐款進度報告

累計 3321 筆，共 167 頁 目前在 第15頁 ▾					
編號	報導標題	刊登日期	狀態	累計(元)	捐款明細
A3204	癌父苦育子 靠麵線填肚	2013/10/21	已結案	829461	明細
A3203	攤父難養家 哀嘆好痛苦	2013/10/20	已結案	511776	明細
A3202	癌復發2女病婦無助	2013/10/18	已結案	653351	明細
A3200	父腰傷 國三女陪母打工	2013/10/17	已結案	504568	明細
A3201	兄煎髮撐家 9歲女顧媽	2013/10/16	已結案	609932	明細
A3198	苦顧老父慇兒 癌男抱病上工	2013/10/15	已結案	734220	明細
A3195	沒力氣抱兒 重癱去自責	2013/10/14	已結案	563365	明細
A3199	母癌父殘 高三女撐6口	2013/10/13	已結案	930004	明細
A3197	去B肝變癌 心疼妻吃苦	2013/10/11	已結案	539112	明細

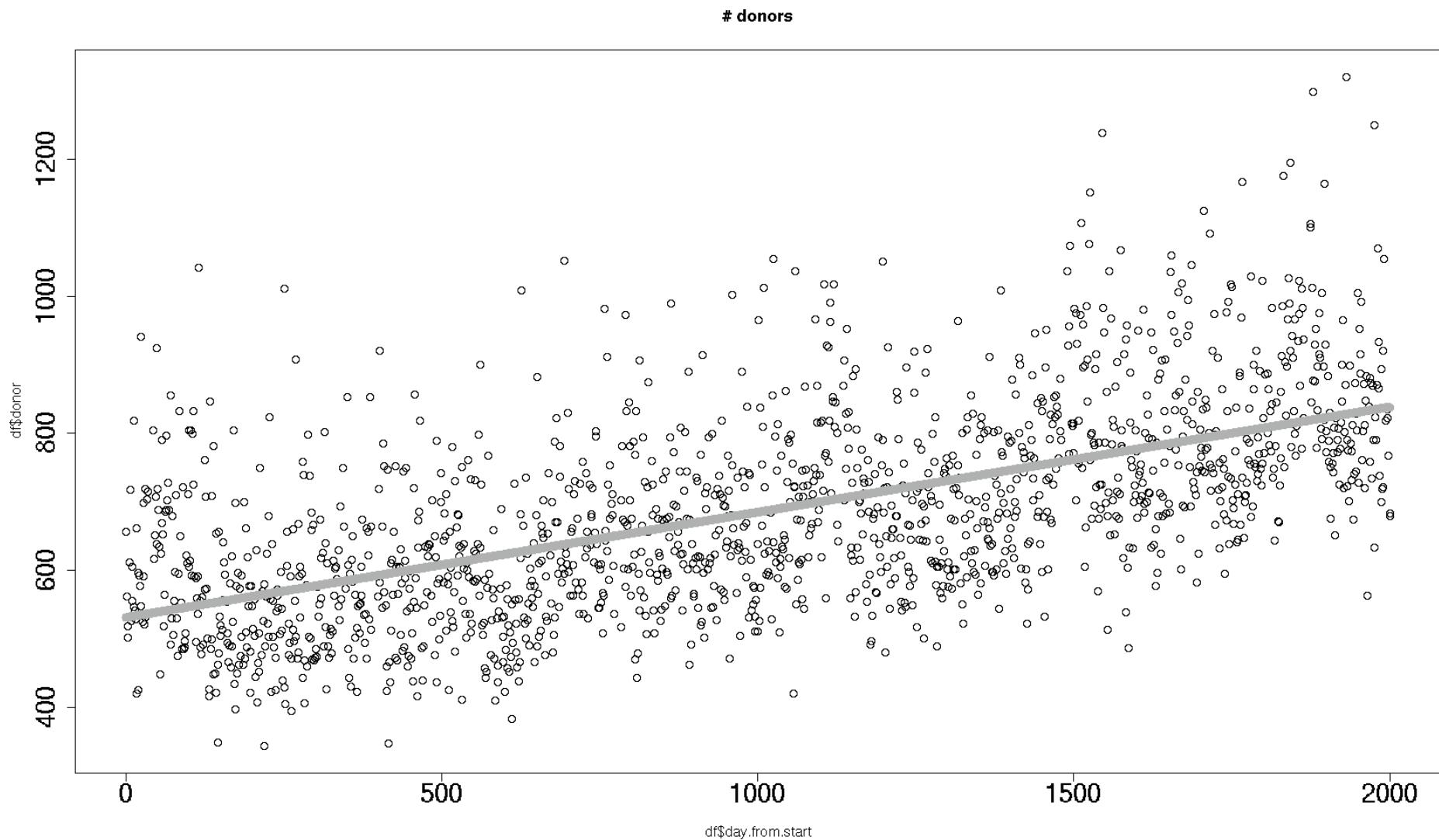
Web page parsing

lid	dt.published	dt.funded	donation	donor	title	journalist	donation.mean	donation.max	n.words	1
A1472	1/3/2008	2/11/2008	253336	502	夫猝逝 婦扛20萬債茫然	韓旭爾	504.6533865	10000	902	
A1473	1/4/2008	2/15/2008	297159	519	雙親病 兒待哺 貧漢難撐	孫敏聿	572.5606936	50000	922	
A1476	1/9/2008	2/19/2008	240616	526	半百父擺攤 顧甦醒癱兒	韓旭爾	457.4448669	10000	933	
A1478	1/8/2008	2/19/2008	360417	718	癌末父抱病養家癱躺入院	張嘉恬	501.9735376	10000	799	
A1479	1/7/2008	2/15/2008	293580	612	夫兒皆精障 妻憂愁成疾	孫敏聿	479.7058824	10000	867	
A1480	1/2/2008	2/11/2008	267223	562	跛男病纏身 無力顧癱母	孫敏聿	475.4857651	10000	833	
A1481	1/17/2008	2/26/2008	335056	661	癌男忍痛 苦撐顧癱母	向高彬	506.892587	10000	941	
A1482	1/10/2008	2/19/2008	309411	605	單親媽愁沒錢治臉傷娃	張嘉恬	511.4231405	10100	1046	
A1483	1/1/2008	2/20/2008	354794	657	翁顧慇兒女 5口擠陋屋	韓旭爾	540.021309	15000	895	
A1484	1/15/2008	2/21/2008	306252	542	父打零工 遲緩女難就醫	孫敏聿	565.0405904	20000	900	
A1485	1/11/2008	2/20/2008	281992	556	中風癱婦愁醫費兒學費	張嘉恬	507.1798561	10000	974	
A1486	1/18/2008	2/27/2008	202164	421	失智翁棲塌屋 隨地便溺	張嘉恬	480.1995249	8000	1007	
A1487	1/14/2008	2/21/2008	458203	818	癌嬪養孫「拼到人生最後」	向高彬	560.1503667	20015	930	
A1488	1/13/2008	2/22/2008	285906	547	夫成漸凍人 妻不捨苦撐	韓旭爾	522.6800731	10015	1033	
A1489	1/16/2008	2/21/2008	244100	529	殘嬪拄杖 咬牙照料癱兒	韓旭爾	461.436673	8000	956	
A1490	1/22/2008	2/29/2008	320824	593	寡婦打零工 苦撐5口家	孫敏聿	541.0185497	10000	926	
A1492	2/3/2008	3/7/2008	449731	704	聽障生顧5老殘 愁生活費	韓旭爾	638.8224432	20015	1099	
A1493	1/23/2008	2/28/2008	305694	577	孫幼 兩兒病 嫩苦撐4口家	韓旭爾	529.7989601	10000	1002	
A1494	1/20/2008	2/27/2008	309658	597	婦打工顧6口 無錢醫癌夫	張嘉恬	518.6901173	10100	878	
A1495	1/24/2008	2/29/2008	296873	548	1家3子殘「剩老的在撐」	張嘉恬	541.7390511	10000	1078	

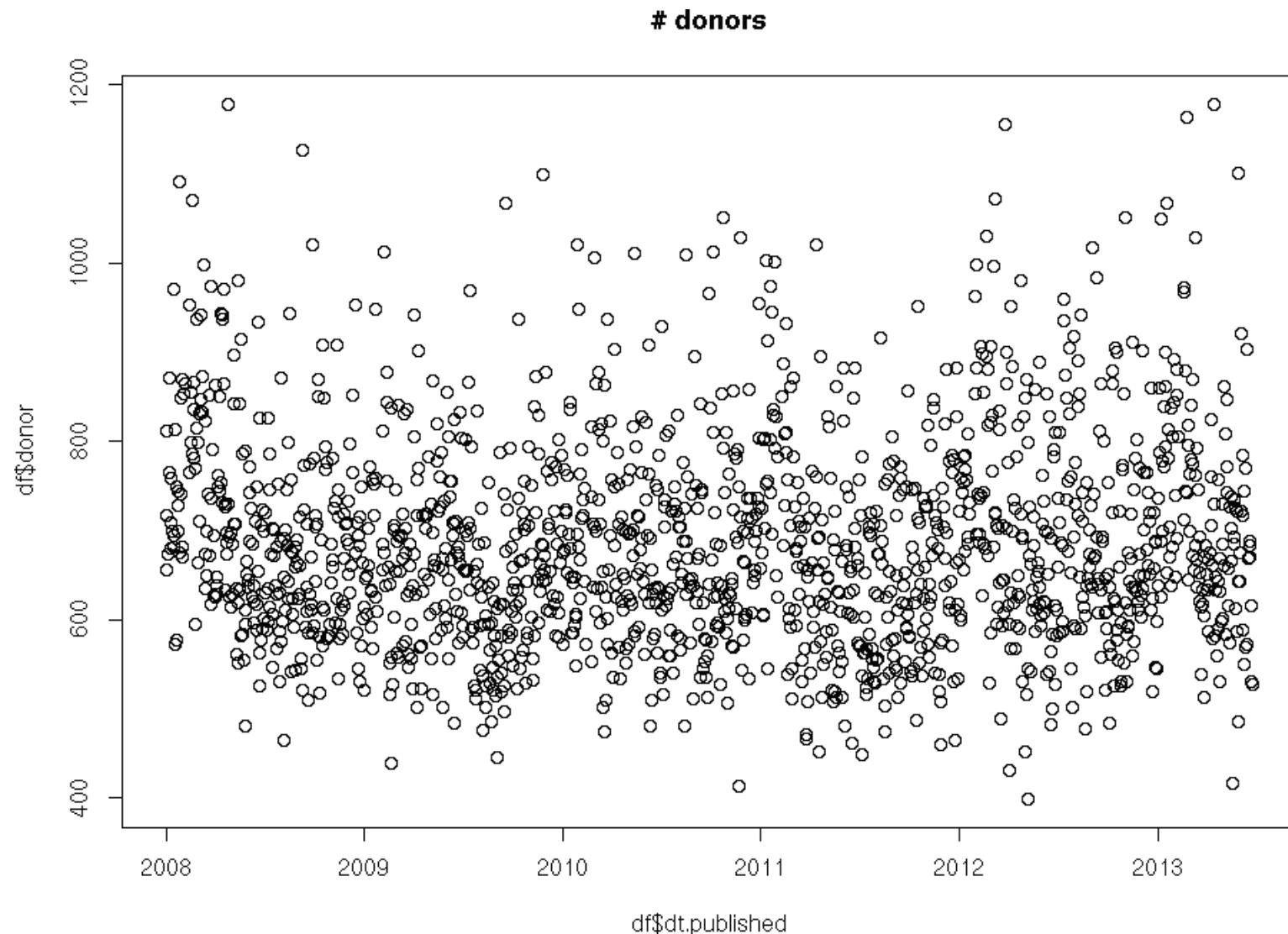
donors



donors w/ linear fitting



Adjusted Time Series



ANNOTATION

人工編碼平台

Online

水腦兒3次進出加護病房 爸媽失業「只想讓她順利長大」

2009年05月31日



黃秋香（右）細心地幫小芬換藥，「醫生交代要注意術後傷口清潔，萬一感染就糟糕了。」

看著女兒小芬終於能開口有說有笑，媽媽黃秋香說：「小芬因水腦症，3次進出加護病房，我的心好像吊在半空中，現在雖然出院，但裝在頭部的引流管恐怕要跟著她一輩子了。」報導・攝影／韓旭爾

「我不求孩子功課好，只想讓她順利長大。」40歲的黃秋香說，今年初，就讀國小六年級的小芬抱怨時常頭痛、頭暈，她並不以為意，直到2月開學時，學校老師發現小芬寫字時，手會不停地發抖，她才察覺女兒健康出了問題，在地區醫院看了兩個多月，也找不出病因，「後來小芬嚴重頭痛，整天都想睡覺，送到大醫院檢查，結果發現水腦症。」

家中尚有三名老弱

黃秋香說，醫生表示小芬的腦脊髓液因不明原因堆積在腦部，若不及時處理，會引起腦部傷害甚至死亡，現已開刀為小芬裝上引流系統，把腦脊髓液從腦室引流到腹腔，「未來小芬每天都要量體溫，只要一有頭痛、嘔吐或發燒就要趕快送醫。」黃秋香說，小芳到大醫院檢查，每次就要耗時一整天，「掛號費、來回交通開銷將近1000元，負擔真的好大。」

黃秋香說，她與40歲的丈夫阿義育有3個子女，最小女兒就讀小四，家中還有69歲的年邁公公，一家生計原靠阿義在養豬場工作維持，每月收入約2萬元，而黃秋香也在資源回收場打零工，每



受訪者：是 | 否

姓名：葉育廷

年齡：41

婚姻狀態：未婚 | 已婚 | 失婚

現與受訪者同住：是 | 否

目前收入來源：殘障補助

經常性月支出：

*若未提及年齡，請以學制推算。[?]

是否為學生：是 | 否

經常性月收入：0.3

萬元

萬元

目前育兒人數：

人

身心狀況

肢體障礙

智能障礙

視覺障礙

精神障礙

言語障礙

聽覺障礙

味覺障礙

嗅覺障礙

皮膚受損

日常狀態

固定回診（例：洗腎、化療、復健）

現正住院（例：醫院、安養院）

臥病在家（例：休養、行動不便）

身體癱瘓（有意識）

植物人（無意識）

失聯

死亡

在外成家或工作

看護家人

無穩定工作（兼職）

有穩定工作（正職）

受家人看護

疾病

腦部受損（例：中風、癲癇）

呼吸道疾病（例：肺結核、氣喘）

肝臟疾病

關節疾病（例：痛風）

精神疾病（例：憂鬱、躁鬱）

腎臟疾病

脊髓／脊椎受損（例：小兒麻痺）

糖尿病

癌症（例：惡性腫瘤）

炎症（例：組織炎）

心血管疾病（例：高血壓、心臟病）

其他疾病

生活習慣

吃檳榔

抽菸

喝酒

賭博

吸毒

過勞

事件

休學／輟學

失業

入獄

外遇／婚變

家暴

離婚

喪偶

天然災害（例：水災、地震、傳染病）

人為事故（例：酒駕、火災）

先天／遺傳病症

急性病症

慢性病症

自殺（例：自殺死亡、自殺未遂、自殘）

債務問題（例：房貸、學貸）

意外事故（例：車禍、工傷）

\$ 10,00,00



Bounty Workers

線上微型案件媒合平台

\$ 9,00,00



➥ 截至目前 Bounty Workers 已成功執行 3250 次任務，並且發出了 99,275 元的獎勵

<http://bountyworkers.net/>



可執行任務

可以藉由完成這些任務，從中獲取豐富獎勵！



會議問卷建檔

by Salmon

\$ 50 元 / 約 20 分鐘

➥ 可執行 3 次



公益文章捐款意願調查

by Jason

\$ 30 元 / 約 10 分鐘

➥ 可執行 2 次

⌚ 剩餘 26 天



跨平台影片播放使用者滿意度調查

by MMNET

\$ 40 元 / 約 25 分鐘

➥ 可執行 1 次

⌚ 剩餘 3 天



關懷弱勢 傳愛慈善公益粉絲團

by 賴

\$ 30 元 / 約 60 分鐘

➥ 可執行 1 次

人工編碼成果



Sample Annotations

tag	id	gender	generation	layer	is_respondent	name	age	marriage	live_with_respondent	lowIncome	isStudent	income_resource
person	1	F	elder	1	1	阿蘭	73	NA	NA	1	NA	低收等補助
person	2	M	adult	2	0	阿炳	48	married		1	1	NA
person	3	M	elder	1	0	老伴	77	married		1	1	NA
person	4	F	child	3	0	長女	14	unmarried		1	1	1
person	5	M	child	3	0	弟弟	NA	unmarried		1	1	1
person	7	M	adult	2	0	次子	NA	NA	NA	1	NA	送貨
person	8	F	adult	2	0	妻子	NA	married	NA	1	NA	工廠零工

income	outcome	num_children	isVerif	attr
16400	0	NA	0	condition: 看護家人
0	10000	NA	0	disability: 精神障礙; condition: 植物人（無意識）; disease: 腦部受損; event: 先天／遺傳病症; event
0	0	NA	0	disability: 精神障礙; condition: 受家人看護; disease: 腦部受損; event: 先天／遺傳病症; condition: 臥
0	0	NA	0	
0	0	NA	0	
0	0	NA	0	condition: 在外成家或工作; condition: 看護家人; condition: 無穩定工作
0	0	NA	0	condition: 看護家人; condition: 無穩定工作

tag	label	id
relation	結婚	1_3
relation	生育	1_2
relation	生育	2_4
relation	生育	2_5
relation	生育	1_7
relation	隔代撫養	1_4
relation	隔代撫養	1_5
relation	結婚	7_8

Variables we got (290+)

[1] "f.tle.有捐款幫助的想法"	"f.tle.不舒服"	"f.tle.沮喪,無力"
[5] "f.tle.難過"	"f.tle.可憐"	"f.tle.生活艱苦,極需幫助"
[9] "f.tleximg.有捐款幫助的想法"	"f.tleximg.誠實"	"f.tleximg.不舒服"
[13] "f.tleximg.沮喪,無力"	"f.tleximg.同情"	"f.tleximg.可憐"
[17] "f.tleximg.可信任"	"f.tleximg.吸引力"	"f.tleximg.體型"
[21] "f.meta.writer.記者1"	"f.meta.writer.記者3"	"f.meta.writer.記者4"
[25] "f.meta.writer.記者5"	"f.meta.writer.記者7"	"f.meta.writer.記者8"
[29] "f.meta.day.of.week.Fri"	"f.meta.day.of.week.Sun"	"f.meta.day.of.week.Thu"
[33] "f.meta.day.of.week.Tue"	"f.meta.word.count"	"f.meta.word.count.scale"
[37] "f.meta.image.count"	"f.meta.month.Aug"	"f.meta.month.Dec"
[41] "f.meta.month.Feb"	"f.meta.month.Jul"	"f.meta.month.Jun"
[45] "f.meta.month.Mar"	"f.meta.month.Nov"	"f.meta.month.Oct"
[49] "f.meta.month.Sep"	"f.any.adult.condition"	"f.any.adult.disease"
[53] "f.any.adult.event"	"f.any.adult.habit"	"f.any.child.disability"
[57] "f.any.child.disease"	"f.any.child.event"	"f.any.child.infant.condition"
[61] "f.any.child.infant.disability"	"f.any.child_infant.disease"	"f.any.child_infant.habit"
[65] "f.any.condition.失聯"	"f.any.condition.在外成家或工作"	"f.any.condition.死亡"
[69] "f.any.condition.身體癱瘓"	"f.any.condition.受家人看護"	"f.any.condition.臥病在家"
[73] "f.any.condition.看護家人"	"f.any.condition.現正住院"	"f.any.condition.無穩定工作"
[77] "f.any.dead.adult"	"f.any.dead.child"	"f.any.dead.elder"
[81] "f.any.dead.elder_adult"	"f.any.dead.infant"	"f.any.disability.言語"
[85] "f.any.disability.味覺"	"f.any.disability.肢體"	"f.any.disability.視覺"
[89] "f.any.disability.嗅覺"	"f.any.disability.精神"	"f.any.disease.心血管"
[93] "f.any.disease.失智"	"f.any.disease.肝"	"f.any.disease.炎"
[97] "f.any.disease.脊椎"	"f.any.disease.腎"	"f.any.disease.精神"
[101] "f.any.disease.糖尿病"	"f.any.disease.癌"	"f.any.disease.鬱"
[105] "f.any.elder.condition"	"f.any.elder.disability"	"f.any.elder.event"
[109] "f.any.elder.habit"	"f.any.elder_adult.condition"	"f.any.elder_adult.disease"
[113] "f.any.elder_adult.event"	"f.any.elder_adult.habit"	"f.any.event.入獄"
[117] "f.any.event.天然災害"	"f.any.event.失業"	"f.any.event.急性病症"
[121] "f.any.event.家暴"	"f.any.event.婚變"	"f.any.event.債務"
[125] "f.any.event.意外事故"	"f.any.event.慢性病症"	"f.any.event.遺傳病症"
[129] "f.any.event.離婚"	"f.any.habit.吃檳榔"	"f.any.habit.抽菸"
[133] "f.any.habit.喝酒"	"f.any.habit.過勞"	"f.any.infant.condition"
[137] "f.any.infant.disability"	"f.any.infant.disease"	"f.any.infant.habit"
[141] "f.any.student"	"f.any.subsidized"	"f.subject.condition.失聯"
[145] "f.subject.condition.在外成家或工作"	"f.subject.condition.有穩定工作"	"f.subject.condition.身體癱瘓"
[149] "f.subject.condition.受家人看護"	"f.subject.condition.固定回診"	"f.subject.condition.看護家人"
[153] "f.subject.condition.現正住院"	"f.subject.condition.植物人"	"f.subject.disability.皮膚"
[157] "f.subject.disability.言語"	"f.subject.disability.味覺"	"f.subject.disability.智能"
[161] "f.subject.disability.視覺"	"f.subject.disability.嗅覺"	"f.subject.disability.聽覺"
[165] "f.subject.disease.心血管"	"f.subject.disease.失智"	"f.subject.disease.呼吸道"
[169] "f.subject.disease.炎"	"f.subject.disease.脊椎"	"f.subject.disease.腦"
[173] "f.subject.disease.精神"	"f.subject.disease.糖尿病"	"f.subject.disease.關節"
[177] "f.subject.disease.鬱"	"f.subject.event.人為事故"	"f.subject.event.天然災害"
[181] "f.subject.event.失業"	"f.subject.event.自殺"	"f.subject.event.家暴"
[185] "f.subject.event.婚變"	"f.subject.event.喪偶"	"f.subject.event.意外事故"
[189] "f.subject.event.慢性病症"	"f.subject.event.轉學"	"f.subject.event.離婚"
[193] "f.subject.gender"	"f.subject.habit.吃檳榔"	"f.subject.habit.抽菸"
[197] "f.subject.habit.喝酒"	"f.subject.habit.過勞"	"f.subject.student"
[201] "f.subject.subsidized"	"f.subject.event.零工"	"f.age.max"

Methodology

- Predict # donors and donation amount
- Feature selection based on mutation information
- Using libsvm to do 2-class classification
 - Classifying top 25% and bottom 25% cases by removing the middle 50% cases
 - 10-fold cross validation
- Find out significant factors that determine the dependent variable(s)

Factor Categories

- Subject
- Structure
- Finance
- Member
- Presentation
- Meta

Factor – Members Category

■ Subject & Member

- Age, gender, marital status
- Disability, disease, accident, habit, status

```
[1] "f.subject.disability.皮膚"
[4] "f.subject.disability.肢體"
[7] "f.subject.disability.嗅覺"
[10] "f.subject.disease.心血管"
[13] "f.subject.disease.呼吸道"
[16] "f.subject.disease.腎"
[19] "f.subject.disease.糖尿病"
[22] "f.subject.disease.鬱"
[25] "f.subject.event.天然災害"
[28] "f.subject.event.急性病症"
[31] "f.subject.event.喪偶"
[34] "f.subject.event.慢性病症"
[37] "f.subject.event.離婚"
[40] "f.subject.habit.吸菸"
[43] "f.subject.habit.過勞"
[46] "f.subject.status.在外成家或工作"
[49] "f.subject.status.身體癱瘓"
[52] "f.subject.status.臥病在家"
[55] "f.subject.status.植物人"
[58] "f.subject.零工"
[61] "f.subject.age.L3_49_92"
[64] "f.subject.gender.age.F_L3_49_92"
[67] "f.subject.gender.age.M_L3_49_92"
[70] "f.subject.marital.status.NA."
[ ] "f.subject.disability.言語"
[ ] "f.subject.disability.智能"
[ ] "f.subject.disability.精神"
[ ] "f.subject.disease.失智"
[ ] "f.subject.disease.炎"
[ ] "f.subject.disease.腦"
[ ] "f.subject.disease.癌"
[ ] "f.subject.event.人為事故"
[ ] "f.subject.event.失業"
[ ] "f.subject.event.家暴"
[ ] "f.subject.event.債務"
[ ] "f.subject.event.輟學"
[ ] "f.subject.gender"
[ ] "f.subject.habit.抽菸"
[ ] "f.subject.habit.賭博"
[ ] "f.subject.status.有穩定工作"
[ ] "f.subject.status.受家人看護"
[ ] "f.subject.status.看護家人"
[ ] "f.subject.status.無穩定工作"
[ ] "f.subject.age.L1_3_40"
[ ] "f.subject.gender.age.F_L1_3_40"
[ ] "f.subject.gender.age.M_L1_3_40"
[ ] "f.subject.marital.status.lose.married"
[ ] "f.subject.marital.status.unmarried"
[ ] "f.subject.disability.味覺"
[ ] "f.subject.disability.視覺"
[ ] "f.subject.disability.聽覺"
[ ] "f.subject.disease.肝"
[ ] "f.subject.disease.脊椎"
[ ] "f.subject.disease.精神"
[ ] "f.subject.disease.關節"
[ ] "f.subject.event.入獄"
[ ] "f.subject.event.自殺"
[ ] "f.subject.event.婚變"
[ ] "f.subject.event.意外事故"
[ ] "f.subject.event.遺傳病症"
[ ] "f.subject.habit.吃檳榔"
[ ] "f.subject.habit.喝酒"
[ ] "f.subject.status.失聯"
[ ] "f.subject.status.死亡"
[ ] "f.subject.status.固定回診"
[ ] "f.subject.status.現正住院"
[ ] "f.subject.subsidized"
[ ] "f.subject.age.L2_41_48"
[ ] "f.subject.gender.age.F_L2_41_48"
[ ] "f.subject.gender.age.M_L2_41_48"
[ ] "f.subject.marital.status.married"
[ ] "f.subject.marital.status.unmarried"
```

Factor – Structure Category

■ Structure

- Count and ratio of particular types of family members
- Relationships between members

```
[1] "f.count.adult"
[4] "f.count.child"
[7] "f.count.child_infant"
[10] "f.count.elder"
[13] "f.count.elder_adult"
[16] "f.count.female"
[19] "f.count.infant.male"
[22] "f.count.minority"
[25] "f.count.student"
[28] "f.ratio.adult.male"
[31] "f.ratio.child.male"
[34] "f.ratio.child_infant.male"
[37] "f.ratio.elder.male"
[40] "f.ratio.elder_adult.male"
[43] "f.ratio.infant.female"
[46] "f.ratio.minority"
[49] "f.ratio.student"
[52] "f.relation.any.kids.dead"
[55] "f.relation.any.siblings"
[58] "f.relation.has.divorce"

[1] "f.count.adult.female"
[4] "f.count.child.female"
[7] "f.count.child_infant.female"
[10] "f.count.elder.female"
[13] "f.count.elder_adult.female"
[16] "f.count.infant"
[19] "f.count.male"
[22] "f.count.minority.female"
[25] "f.ratio.adult"
[28] "f.ratio.child"
[31] "f.ratio.child_infant"
[34] "f.ratio.elder"
[37] "f.ratio.elder_adult"
[40] "f.ratio.female"
[43] "f.ratio.infant.male"
[46] "f.ratio.minority.female"
[49] "f.relation.any.divorce"
[52] "f.relation.any.marriage"
[55] "f.relation.both.parents.alive"
[58] "f.relation.has.grandparenting"

[1] "f.count.adult.male"
[4] "f.count.child.male"
[7] "f.count.child_infant.male"
[10] "f.count.elder.male"
[13] "f.count.elder_adult.male"
[16] "f.count.infant.female"
[19] "f.count.members"
[22] "f.count.minority.male"
[25] "f.ratio.adult.female"
[28] "f.ratio.child.female"
[31] "f.ratio.child_infant.female"
[34] "f.ratio.elder.female"
[37] "f.ratio.elder_adult.female"
[40] "f.ratio.infant"
[43] "f.ratio.male"
[46] "f.ratio.minority.male"
[49] "f.relation.any.kids"
[52] "f.relation.any.parents.alive"
[55] "f.relation.has.adoption"
[58] "f.relation.is.cohabitation"
```

Factor – Finance Category

■ Finance

- Is the family below the poverty line?
- Regular income & expense

```
[1] "f.finance.below.poverty.line"  
[2] "f.finance.income.source"  
[3] "f.finance.income.level.L1_300_33500"  
[4] "f.finance.income.level.L2_34000_65400"  
[5] "f.finance.income.level.L3_67230_100000"  
[6] "f.finance.income.level.NA"  
[7] "f.finance.expense.level.L1_40_64000"  
[8] "f.finance.expense.level.L2_80000_127000"  
[9] "f.finance.expense.level.L3_200000_200000"  
[10] "f.finance.expense.level.NA"
```

Factor – Presentation

■ Presentation

- Currently, only title and images are evaluated
- Subjective ratings from human subjects

[1]	"f.tle.有捐款幫助的想法"	"f.tle.不忍觀看.想要迴避"
[3]	"f.tle.不舒服"	"f.tle.沮喪.無力"
[5]	"f.tle.難過"	"f.tle.同情"
[7]	"f.tle.可憐"	"f.tle.生活艱苦.極需幫助"
[9]	"f.tleximg.有捐款幫助的想法"	"f.tleximg.不忍觀看.想要迴避"
[11]	"f.tleximg.誠實"	"f.tleximg.不舒服"
[13]	"f.tleximg.沮喪.無力"	"f.tleximg.難過"
[15]	"f.tleximg.同情"	"f.tleximg.可憐"
[17]	"f.tleximg.可信任"	"f.tleximg.生活艱苦.極需幫助"
[19]	"f.tleximg.吸引力"	"f.tleximg.體型"

Title & picture rating

義消車禍成殘 4女待養



您會如何描述此照片中的主角？

體型 (未選)

(過瘦) 1

7 (過胖)

吸引力 (未選)

(完全不具吸引力) 1

7 (非常具吸引力)

誠實 (未選)

(完全不誠實) 1

7 (非常誠實)

生活艱苦，極需幫助 (未選)

(完全不需幫助) 1

7 (非常需要幫助)

可信任 (未選)

(完全不可信任) 1

7 (非常可信任)

可憐 (未選)

(完全不可憐) 1

7 (非常可憐)

此照片及文字描述給您的感覺為？

同情 (未選)

(完全不同情) 1

7 (非常同情)

難過 (未選)

(完全不難過) 1

7 (非常難過)

沮喪/無力 (未選)

(完全不沮喪) 1

7 (非常沮喪)

不舒服(未選)

(完全舒服) 1

7 (非常不舒服)

不忍觀看/想要迴避 (未選)

(不會想要迴避) 1

7 (非常想要迴避)

有捐款幫助的想法(未選)

(完全沒有) 1

7 (非常強烈)

Factor – Meta Information

■ Meta information

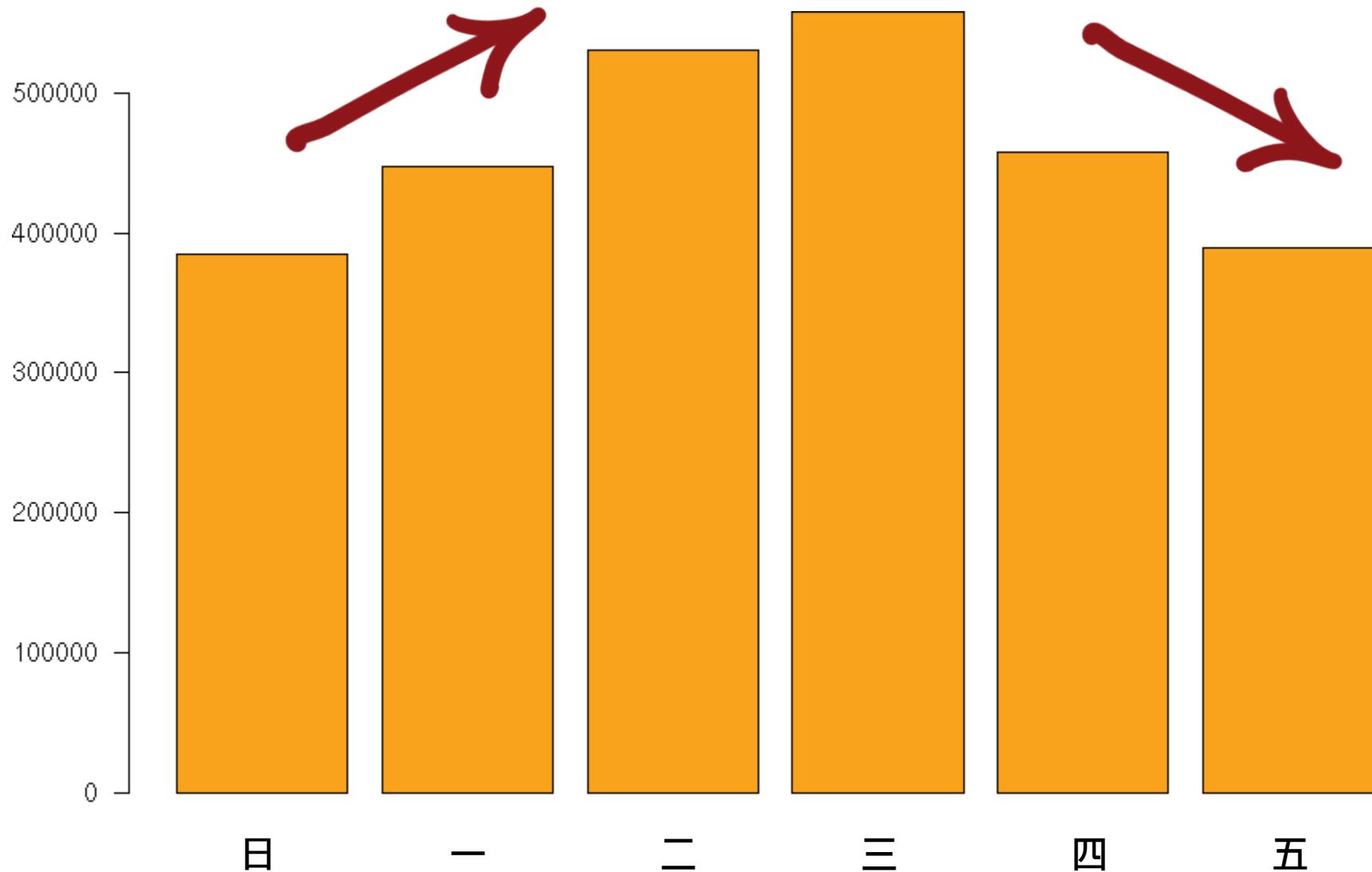
- Information unrelated to the family & its situation
- E.g., article writer and when was the article published

KEY FINDINGS

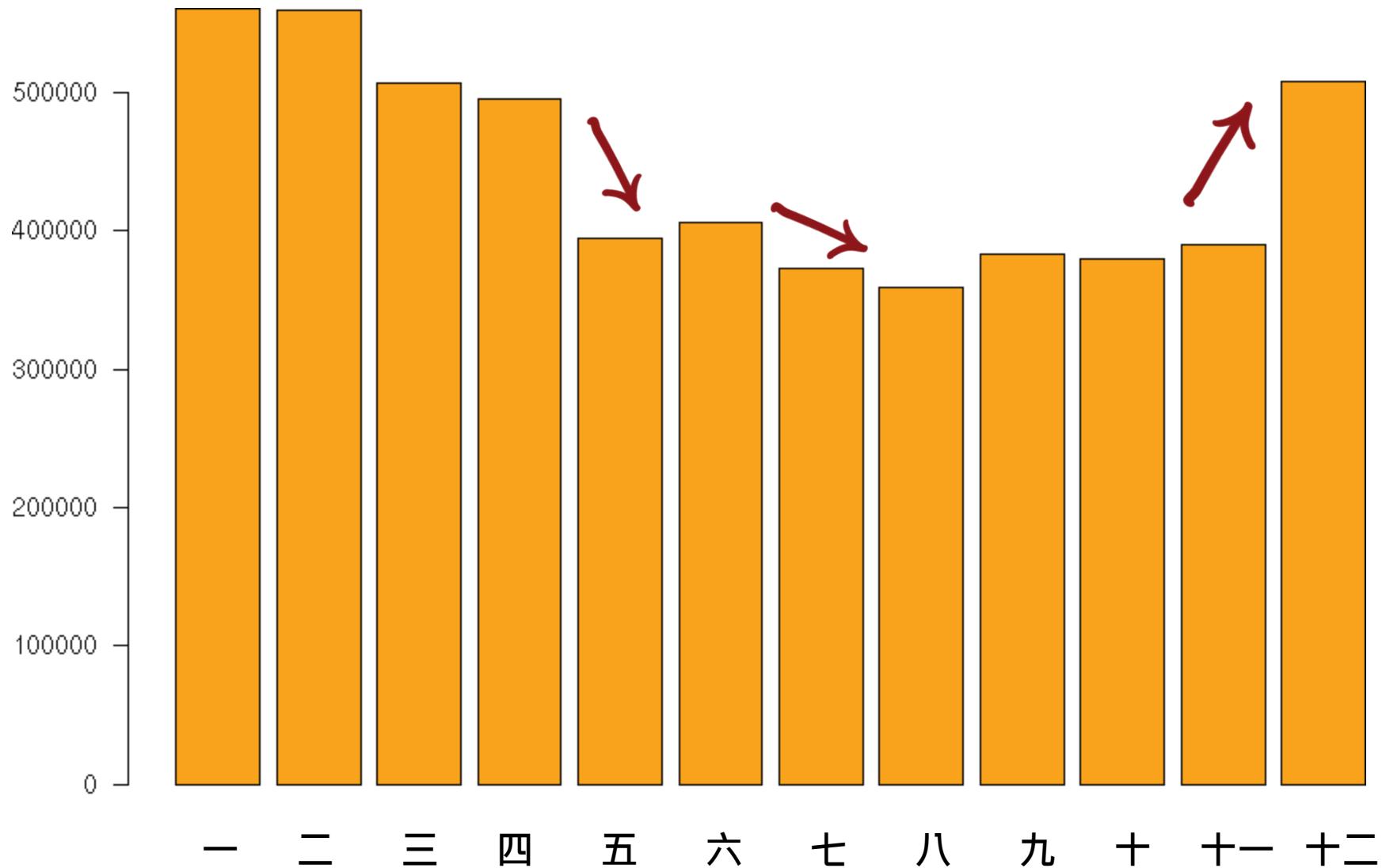
捐款意願與時間點 高度相關



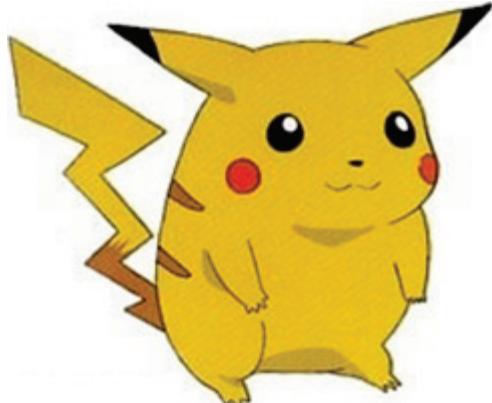
星期幾很重要

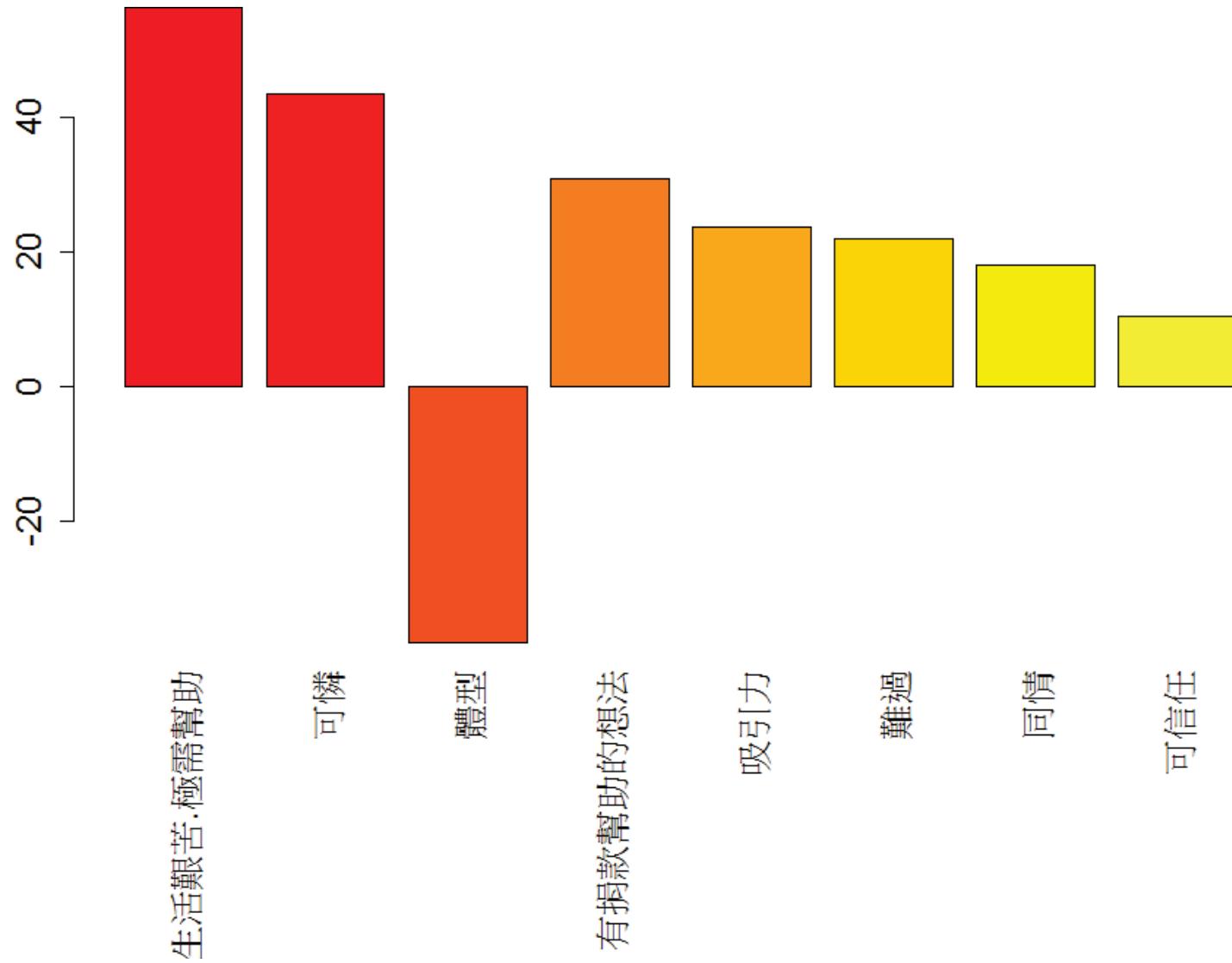


哪個月份也重要

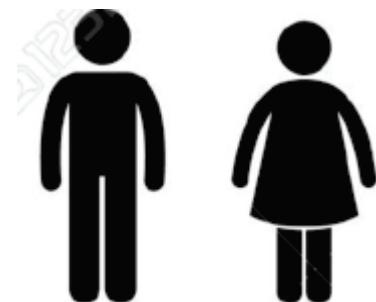


受訪者的胖瘦會影響 捐款決策





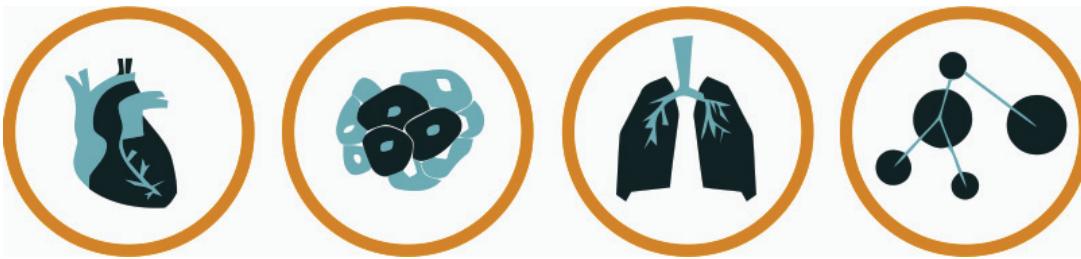
誰收到較多捐款？



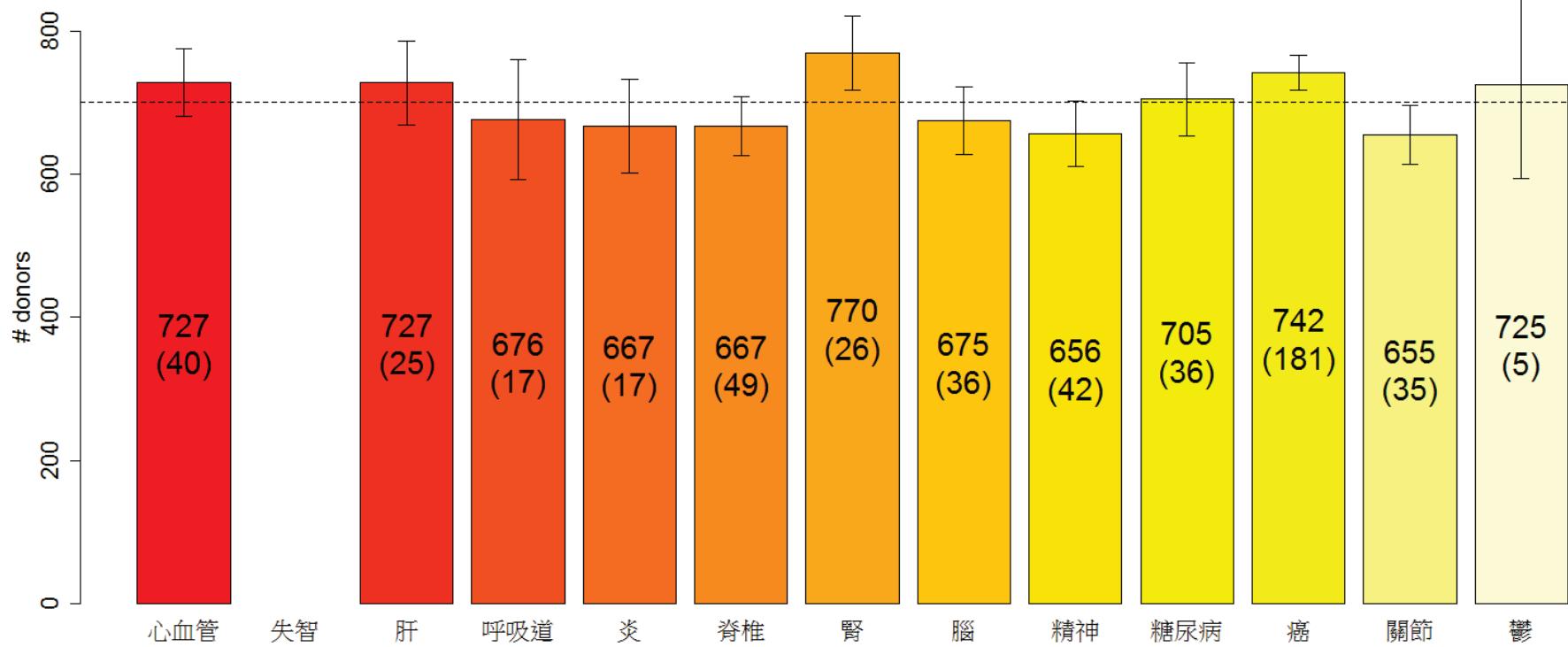
老弱婦孺及單身者能 收到較多捐款



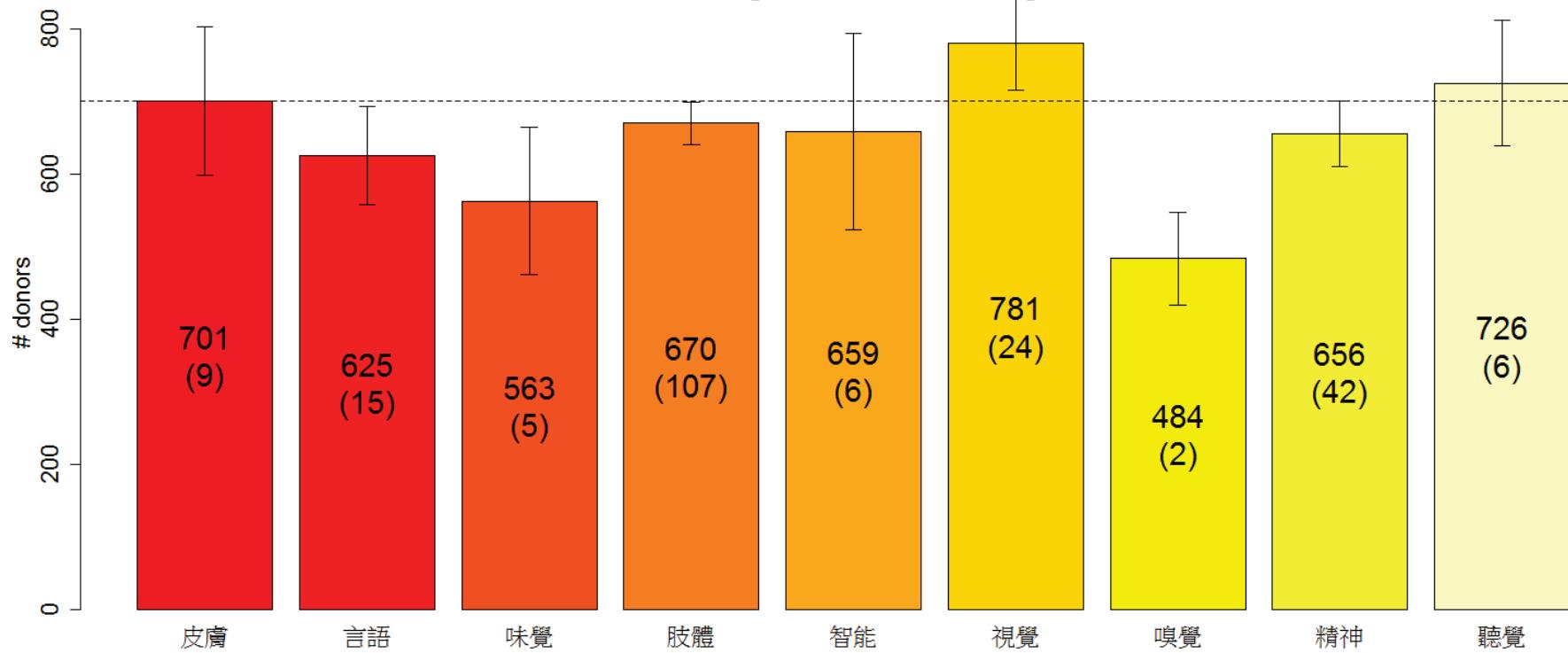
捐款人對各式疾病及 身心障礙有差別待遇



f.subject.disease.



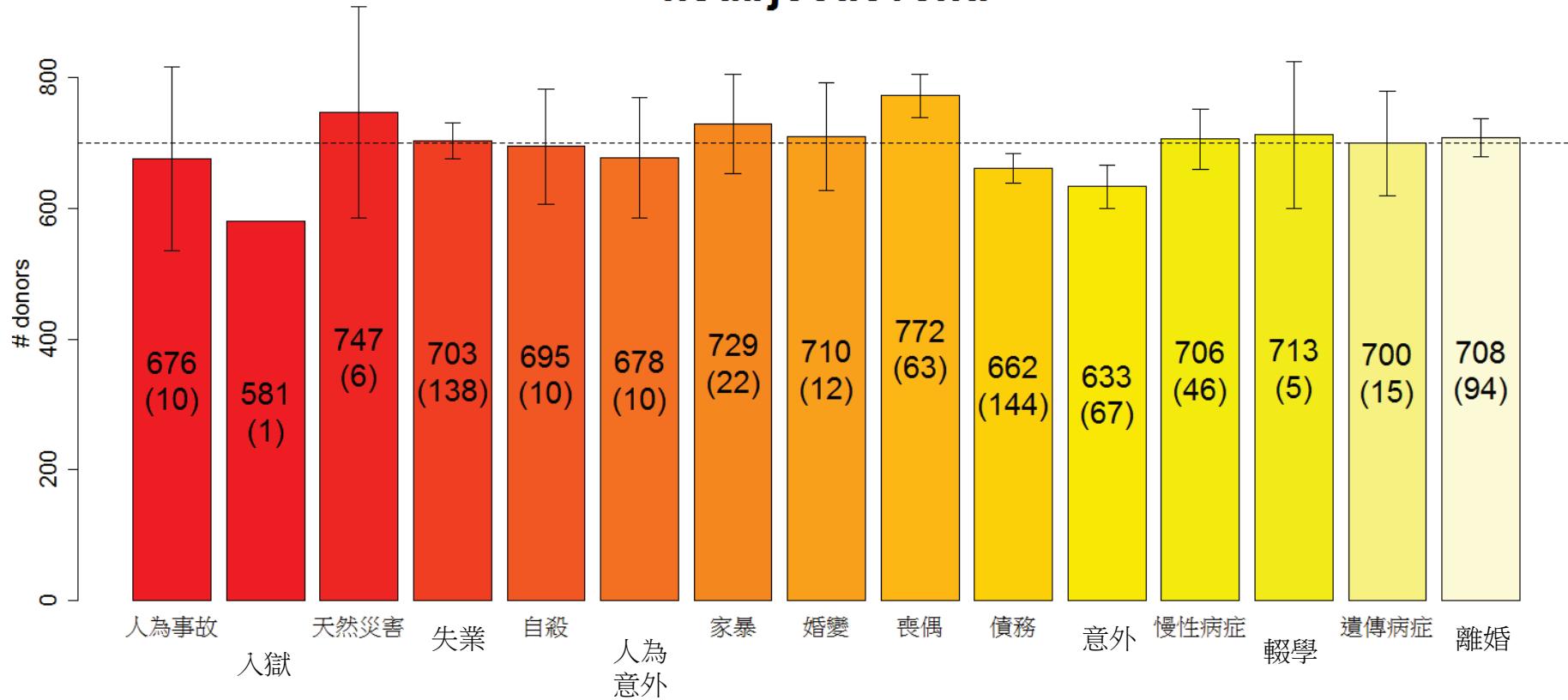
f.subject.disability.



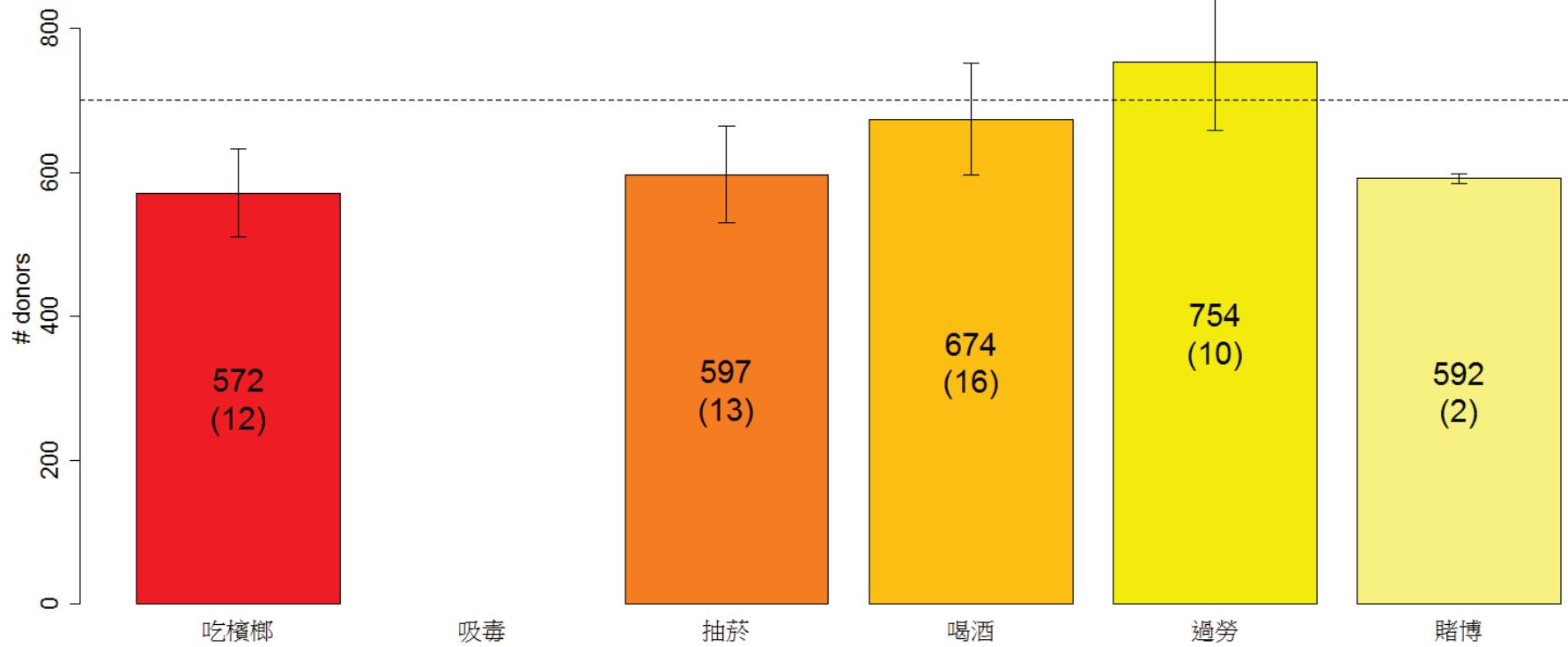
不可抗力因素 較讓人同情



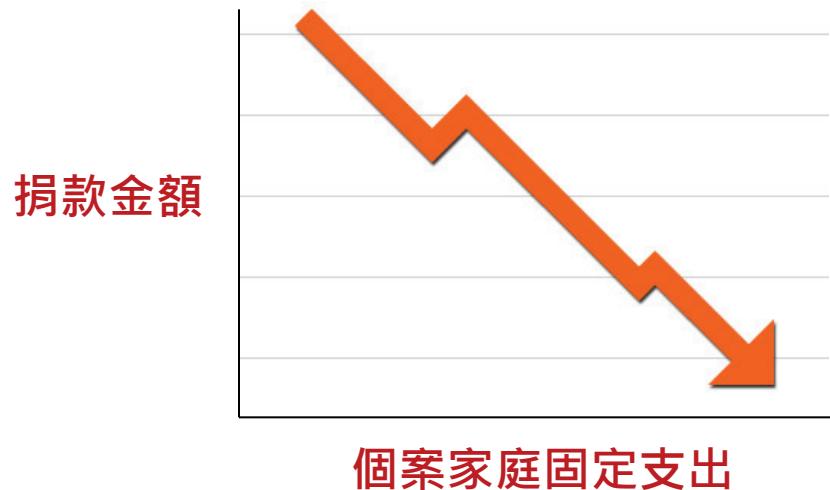
f.subject.event.



f.subject.habit.



捐款與固定支出 成反比



捐款者期待能看見
「希望」



CASE STUDY

Successful Case

A1568: 氣爆毀容 妙齡女復健路長 (donor = 1179)

rank #1 (out of 1581) in terms of donor

subject gender: Female

subject age: 25

id.subject: # 8

		var	val	decision	change
2	f.tle.生活艱苦.極需幫助	2	0.998	-0.3031	
77	f.subject.disability.視覺	1	0.426	0.2688	
3	f.any.dead.adult	1	0.573	0.1212	
58	f.subject.disability.肢體	1	0.803	-0.1086	
87	f.any.disability.視覺	1	0.587	0.1072	
50	f.meta.image.count	3	0.593	0.1018	
93	f.any.elder.disability	1	0.604	0.0905	
11	f.any.disease.癌	0	0.775	-0.0805	
10	f.count.members	6	0.621	0.0740	
72	f.any.event.意外事故	1	0.768	-0.0738	
30	f.subject.gender.age.F_L1_3_40	1	0.624	0.0706	
9	f.tle.可憐	6	0.624	0.0705	
64	f.ratio.minority	0	0.752	-0.0572	
16	f.count.female	6	0.640	0.0544	
47	f.count.elder_adult.female	6	0.641	0.0540	
14	f.any.dead.elder_adult	1	0.641	0.0538	
12	f.any.condition.死亡	1	0.642	0.0527	
8	f.tle.同情	6	0.644	0.0508	
59	f.any.elder.condition	1	0.647	0.0472	
69	f.ratio.child_infant	0	0.739	-0.0443	

Less Successful Case

A2760: 少年肺膿瘍 不放棄舉重 (donor = 398)

rank #1581 (out of 1581) in terms of donor

subject gender: Female

subject age: 45

id.subject: # 1

		var	val	decision	change
66		f.all.dead.elder	1.000	0.23048	-0.2800
35		f.any.child.disease	1.000	-0.18486	0.1354
37		f.meta.day.of.week.Sun	1.000	0.07156	-0.1211
38		f.any.child_infant.disease	1.000	-0.16402	0.1145
11		f.any.disease.癌	1.000	-0.15182	0.1023
32		f.subject.age.L2_41_48	1.000	0.04939	-0.0989
72		f.any.event.意外事故	1.000	0.02521	-0.0747
88		f.subject.condition.看護家人	1.000	0.02347	-0.0730
14		f.any.dead.elder_adult	1.000	-0.10784	0.0583
8		f.tle.同情	6.000	-0.10659	0.0571
12		f.any.condition.死亡	1.000	-0.10555	0.0561
59		f.any.elder.condition	1.000	-0.10415	0.0547
63	f.subject.marital.status.married		1.000	0.00420	-0.0537
65		f.ratio.adult.male	0.143	-0.09611	0.0466
49		f.tle.難過	5.500	-0.09479	0.0453
57		f.any.condition.看護家人	1.000	-0.00734	-0.0422
85		f.finance.income.level.NA	1.000	-0.00770	-0.0418
45		f.subject.disease.癌	0.000	-0.00968	-0.0398
9		f.tle.可憐	5.500	-0.08801	0.0385
67		f.any.adult.disability	1.000	-0.01281	-0.0367

TEXT MINING APPROACH

C-LIWC簡介

- 從James Pennebaker的LIWC (Linguistic Inquiry and Word Count) 發展而來
- 由台科大與台大心理團隊，依照中文特性增刪類別與語詞，編製而成
- 總計88個類別，6862個詞與詞幹
- 語言特性與寫作風格多少能反應個人特質、影響讀者的感受
- 此文本分析方法，逐漸被廣泛使用在心理學相關研究主題。如：道歉與原諒、測謊、治療過程的語言變化、心理位移等
- C-LIWC官網：<http://cliwc.weebly.com/>

中文版語文探索與字詞計算字典 (C-LIWC)

家庭詞	工作詞				家庭詞		金錢詞		宗教詞		死亡詞		健康詞	
入贊	入學	併吞	專題	精神	大門	擦	一毛不拔	貪財	一貫道	鬼魂	入殮	人命	核磁共振	噎
丈夫	上司	典當	探索	精簡	大樓	翻修	一角	貪婪	人子	鬼影	上吊	下藥	氣喘	墮胎
大伯	上流	初級	推動	綱領	女主人	翻新	九牛一毛	通貨	十字架	鬼魅	上香	大夫	消化	徵候
女主人	上校	協作	教	聚集	女傭	雜務	乞丐	報酬	十字軍	偈誦	亡故	小兒科	消炎	憂鬱
女兒	上課	協商	教育	製作	女僕	寵物	乞討	富有	上帝	基督	大屠殺	不良	疤痕	憂鬱症
女婿	大三	協商會	教育性	製造	工作室	樹	小氣	富裕	上師	基督徒	大體	中毒	疲倦	暴食
小犬	大夫	協會	教室	製造業	公寓	簾	工資	提款機	大乘	基督教	不朽	中風	疲勞	暴飲暴食
小叔	大師	協議	教科書	認識	天井	爐	元	款項	大悲咒	基督教徒	公祭	中醫	倦怠	潰瘍
小姑	大學	命令	教師	遣散	火爐	露台	公司	無價	大殿	基督教徒	公墓	元氣	痺	瘡
已婚	大學生	命題	教務長	需要	功課	露台	分紅	發票	小乘	崇拜	分屍	分裂症	疼痛	瞎
內人	小考	委外	教授	領袖	布幕	鰥夫廳	支出	盜用	不朽	救贖	太平間	切除	疾病	磕藥
公公	小學	委派	教導	劇團	平台		支票	硬幣	中元	教士	火化	反胃	瘻	膠布
公婆	工人	委員	清單	增進	打掃		月入	稅	什葉派	教宗	火葬	幻視	病	醉酒
太太	工作	委員長	率領	審查	休假		月薪	稅收	天主教	教派	古墓	幻覺	病房	劑
太祖	間	委員會	產	履歷	冰箱		欠	稅金	天主教堂	教徒	白包	幻聽	病原	劑量
父	工會	定案	畢業	廣告	地主		欠債	買	天使	教堂	安葬	心血管	病假	整形
父母	工業	店員	畢業生	影印	地址		付	買主	天國	教條	死	心痛	病理	激素
父親	工資	延誤	移交	徵募	地契		付款	買賣	天堂	教會	安樂死	心跳	學	糖尿病
兄	工廠	延遲	組	播音室	地毯		付費	貸	天賜	教義	收殮	心電圖	病態	錠
弟														

家庭詞、死亡詞、健康詞

- 相關：家庭詞、死亡詞、健康詞大致和捐款皆成正相關
- 推論：當事件主題符合傳統價值時較易引起捐款

(r, p-value)	家庭詞	死亡詞	健康詞
log(捐款總額)	(r=0.148, p=0.000)	(r=0.101, p=0.000)	(r=0.056, p=0.026)
捐款人數	(r=0.131, p=0.000)	(r=0.113, p=0.000)	(r=0.058, p=0.021)
每人平均捐款額	(r=0.129, p=0.000)	(r=0.084, p=0.001)	(r=0.007, p=0.771)
範例	母親、婆婆、阿 公、家屬、堂妹、 繼父、雙親	火化、死者、自 殺、告別式、往 生、致死	中風、糖尿病、 結石、住院、安 眠藥

文章總詞數

- 相關：文章總詞數和捐款成正相關
- 推論：將事件敘述越詳盡，越容易募到款

(r , p-value)

log(捐款總額)

捐款人數

每人平均捐款額

總詞數 (word count)

($r=0.101$, $p=0.000$)

($r=0.056$, $p=0.027$)

($r=0.143$, $p=0.000$)

工作詞、成就詞、金錢詞

- 相關：工作詞、成就詞、金錢詞大致和捐款皆成負相關
- 推論：和工作相關的主題，相較不易募得款項

(r, p-value)	工作詞	成就詞	金錢詞
log(捐款總額)	(r=-0.079, p=0.002)	(r=-0.064, p=0.011)	(r=-0.072, p=0.004)
捐款人數	(r=-0.099, p=0.000)	(r=-0.085, p=0.000)	(r=-0.025, p=0.319)
每人平均捐款額	(r=-0.022, p=0.380)	(r=-0.020, p=0.001)	(r=-0.101, p=0.000)
範例	勞工、契約、付費、升遷、職權、權威、裁員、生意、員工、嘉獎、能幹、高層、職業	帳戶、租金、商店、現金、消費、捐贈榮耀	

其它

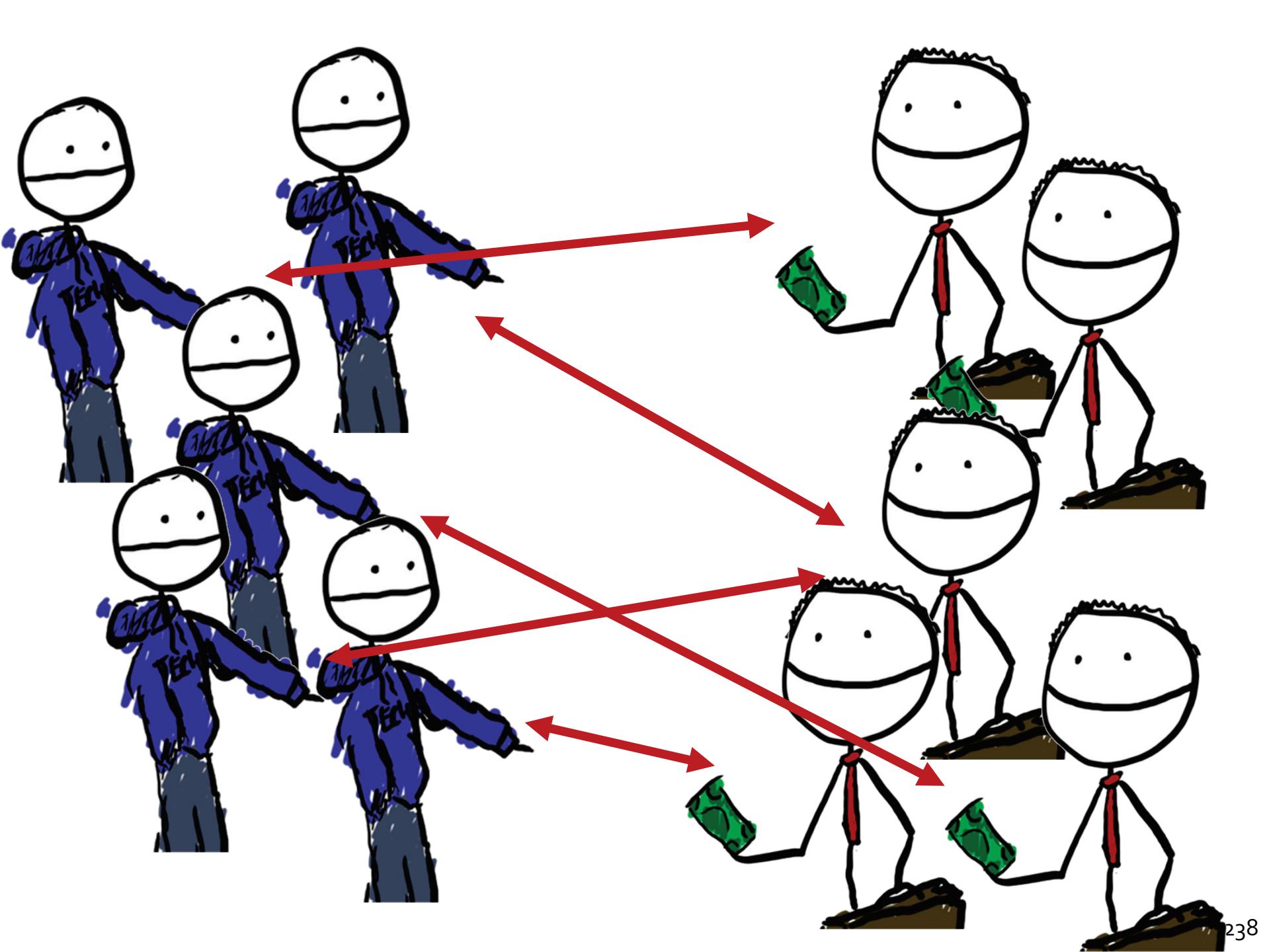
■ 否定詞

- 範例：不滿、不幸、不能、無關、不料、不須
- 相關：和平均每人捐款額呈負相關($r=-0.063, p=0.013$)
- 推論：正面描述較佳

■ 副詞

- 範例：真的、終於、確實、一定、一向、不管、全然
- 相關：和平均每人捐款額呈負相關($r=-0.084, p=0.001$)
- 推論：平實地描述即可，過度誇大或多加贅述易有反效果

ONGOING WORK



Opportunities to explore

- Incentive provisioning
 - Let doners keep track their own donation record
 - Doner profile, like Kiva
 - Re-visit the families being helped
- Viral marketing
- Cognitive biases
 - Anchoring effect
 - Endowed progress effect

(see https://en.wikipedia.org/wiki/List_of_cognitive_biases)

CONCLUSION & OUTLOOK





WE ARE STILL AT
THE VERY START

ROADS

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 (most of) their toolkits
 - Not clear that we will/should make the investment
- Computer scientists have powerful methods but...
 - Trained to propose new algorithms...
 - It seems there are less “methodological” contributions

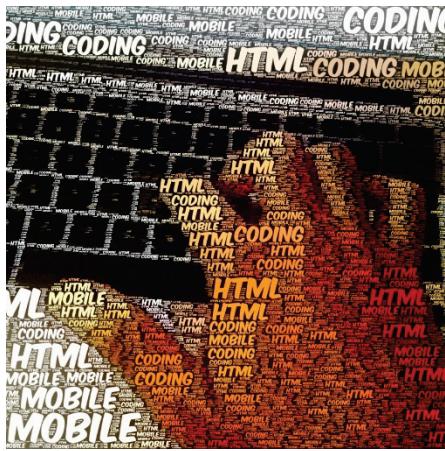
The Challenges

- 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



Ability to process large datasets, algorithms, data mining



Knowledge about social theories, methods, and issues

交流時間





當學術研究者遇見 線上遊戲

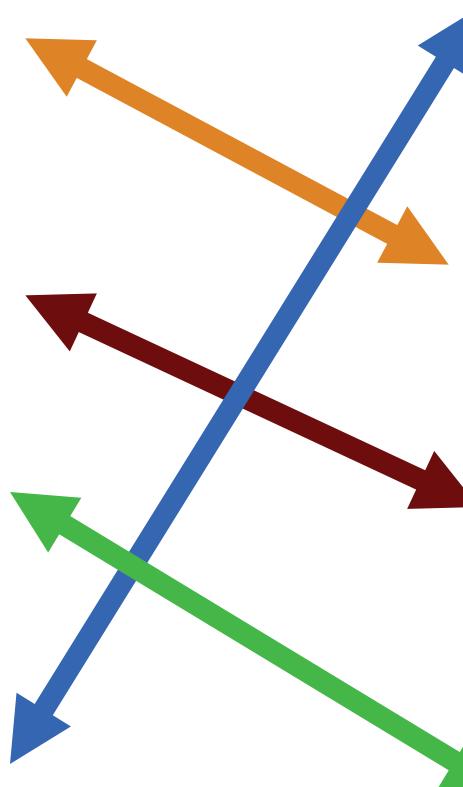
陳昇瑋

中央研究院 資訊科學研究所



Entertainment Market Size (worldwide)

No. 2 US\$ 42 billion



Movie



No. 3 US\$ 35 billion

Video games



No. 1 US\$ 63 billion

Music



No. 4 US\$ 27 billion

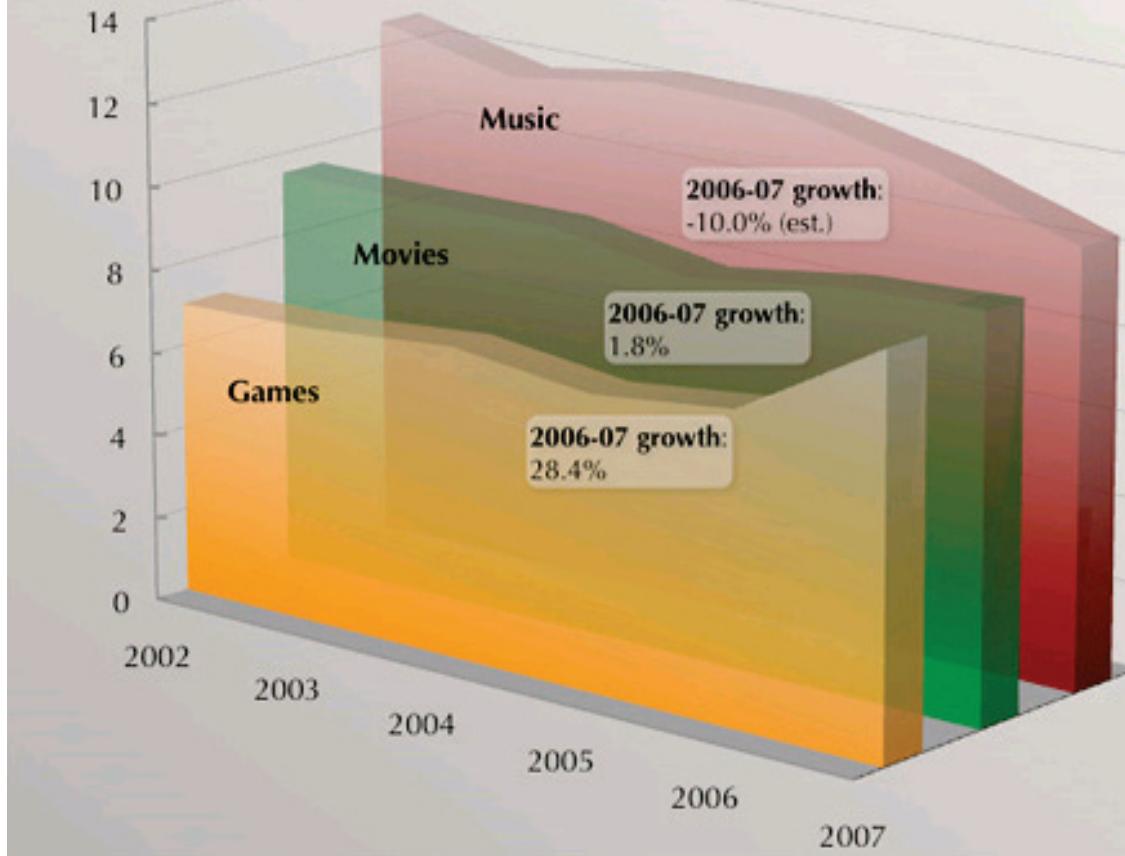
Book



http://vgsales.wikia.com/wiki/Video_game_industry

US music, movie, and gaming revenues — 2002-07

\$ Billions



Game Research: My Own Reasons

As A PC Gamer ...

As A Programmer ...

As A Researcher ...

As A PC Gamer (1)

198



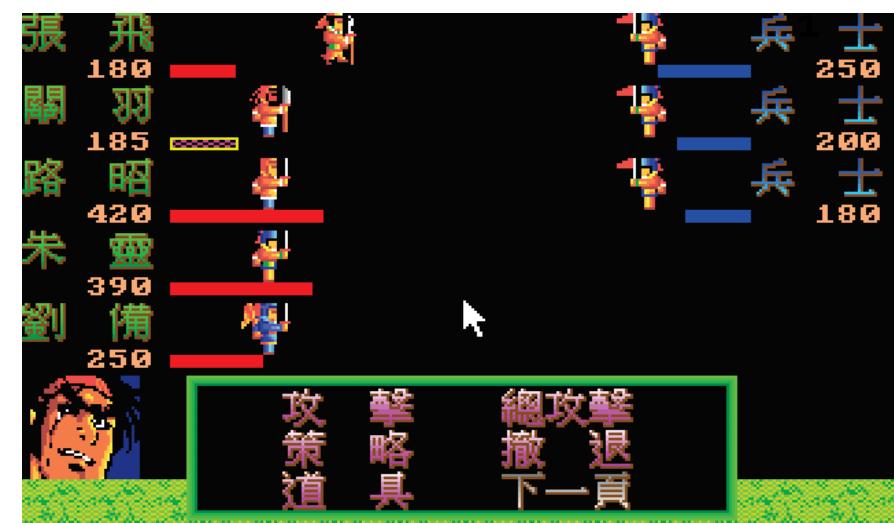
1989



199



199



As A PC Gamer (2)

199



199



199



199



榮獲

八十三學年度全國大學資訊盃

DOOM 項目

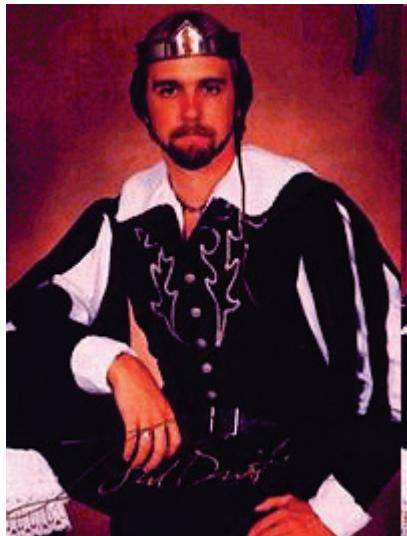
第二名

台大資訊系主辦

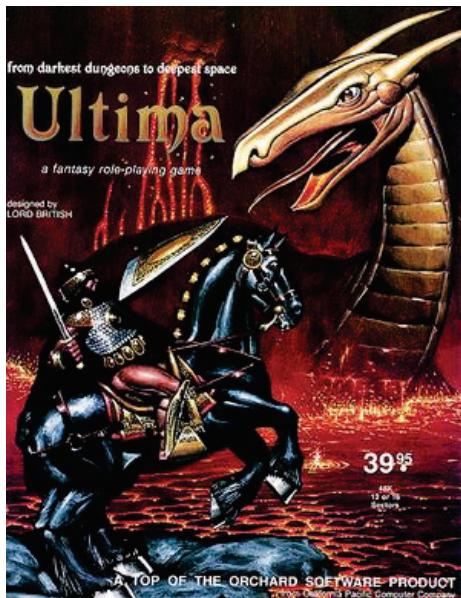
中華民國八十四年四月四日

As A Programmer (1)

- 10 歲寫 football game with ROM BASIC
- 國中寫對打遊戲 with dBASE & Pascal
- 高中寫 RPG with C & Assembly



Richard Garriott



My Role Model in 1990



As A Programmer (2)

- 1999 – 2002 資策會教育訓練課程 (C/C++, Winsock Programming, Delphi, C++Builder) 夾帶遊戲設計課程
- 1999 – 2001 《遊戲設計大師》專欄作家
- 2000 出版《Delphi 深度歷險》
- 2002 出版《C++Builder 深度歷險》



As A Researcher

- **A killer application**
 - 35% Internet users & larger business than movie & music
- **An emerging field**
 - E.g., IEEE Transactions on AI and CI in Games since Sep 2008
- **Asia-based researchers have some niches**
 - Large user base (50%)
 - Lots of local game companies
- **It's fun!**



Security Topics

Game Bot Detection



Game Bots

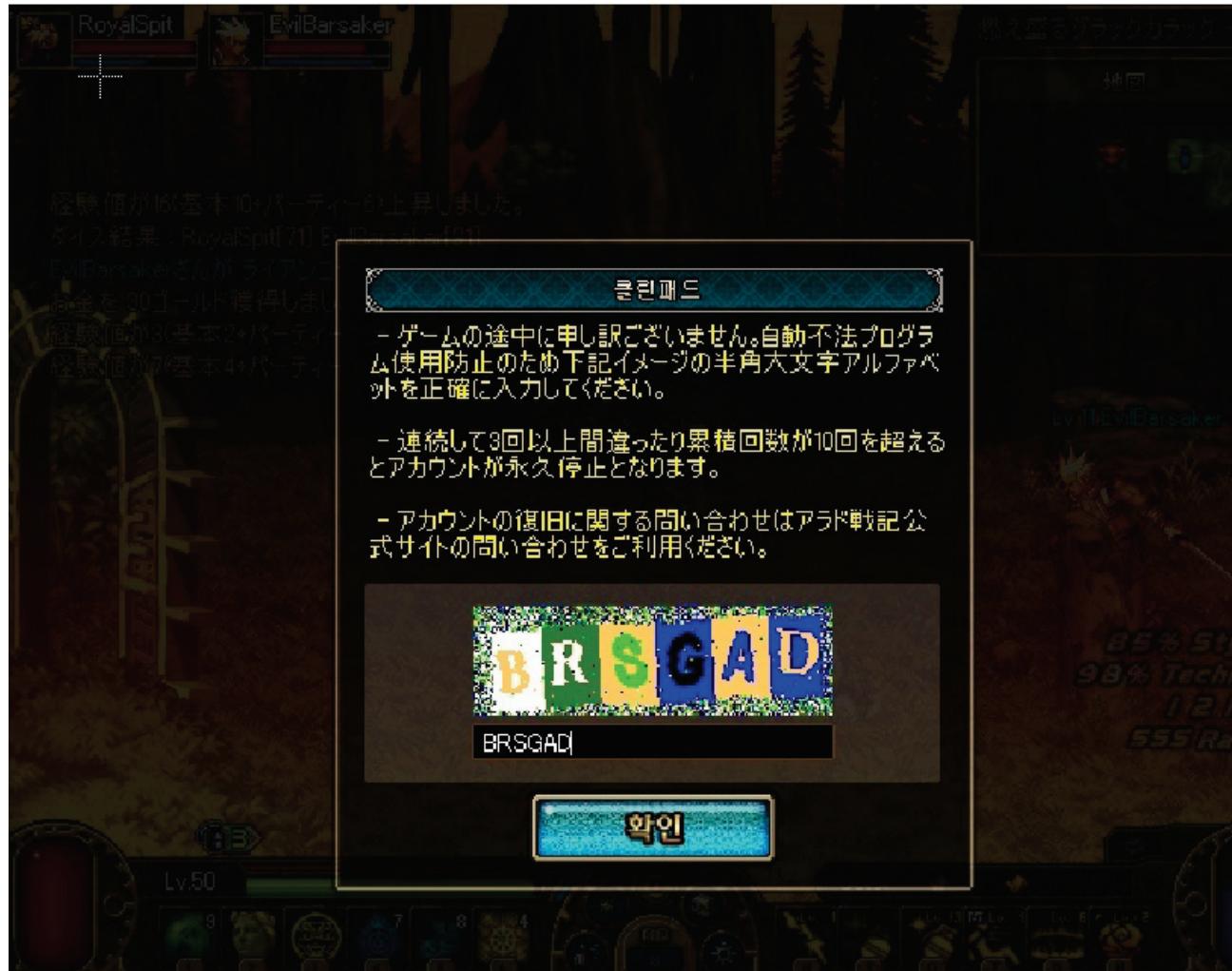
- Game bots: automated AI programs that can perform certain tasks in place of gamers
- Popular in MMORPG and FPS games
 - MMORPGs (Role Playing Games)
accumulate rewards in 24 hours a day
→ break the balance of power and economies in game
 - FPS games (First-Person Shooting Games)
 - a) improve aiming accuracy only
 - b) fully automated
→ achieve high ranking without proficient skills and efforts

Bot Detection

- Detecting whether a character is controlled by a bot is difficult since **a bot obeys the game rules perfectly**
- No general detection methods are available today
- State of practice is identifying via **human intelligence**
 - **Detect by** “bots may show regular patterns or peculiar behavior”
 - **Confirm by** “bots cannot talk like humans”
 - Labor-intensive and may annoy innocent players

CAPTCHA in a Japanese Online Game

(Completely Automated Public Turing test to tell Computers and Humans Apart)



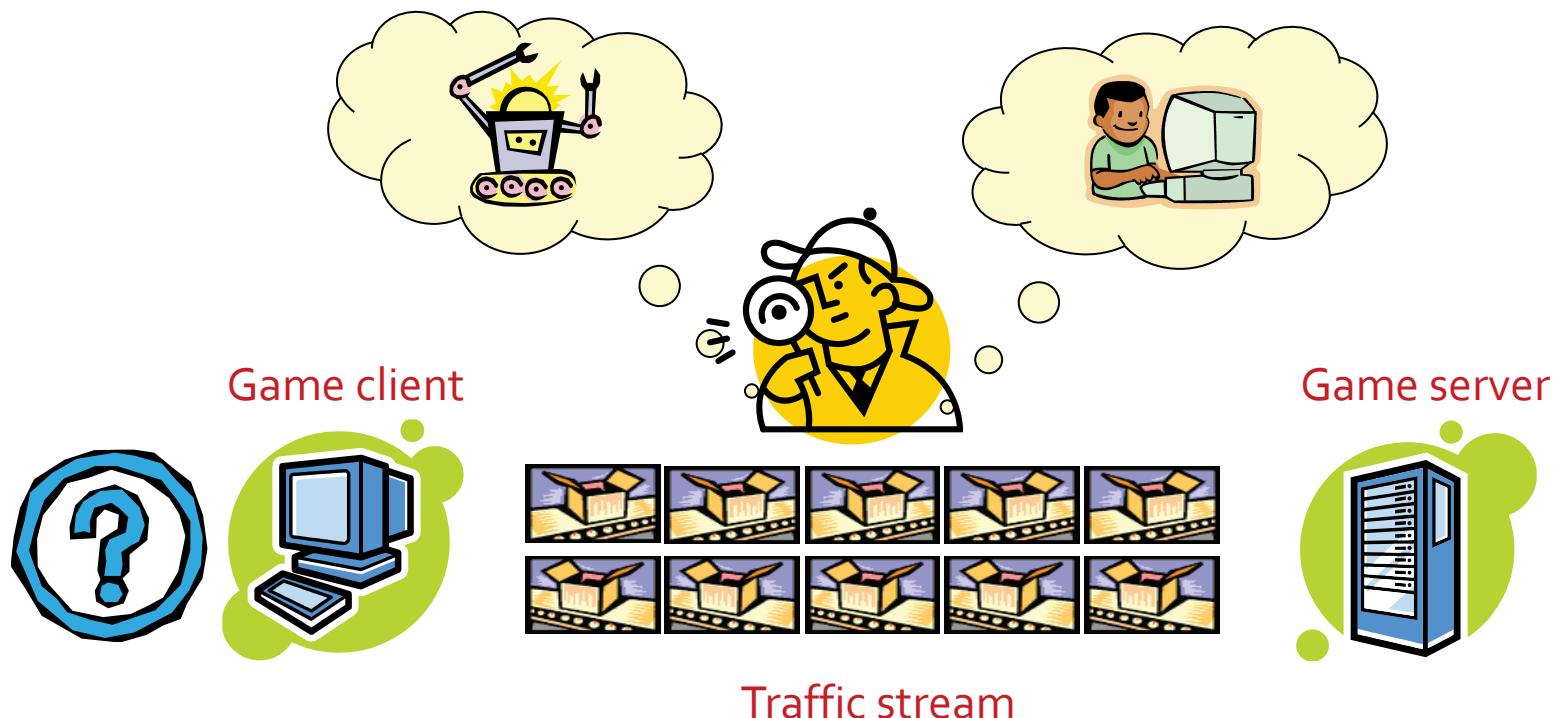
Our Goal of Bot Detection Solutions

- **Passive** detection
 - No intrusion in players' gaming experience
- **No client software support** is required
- **Generalizable** schemes (for other games and other game genres)

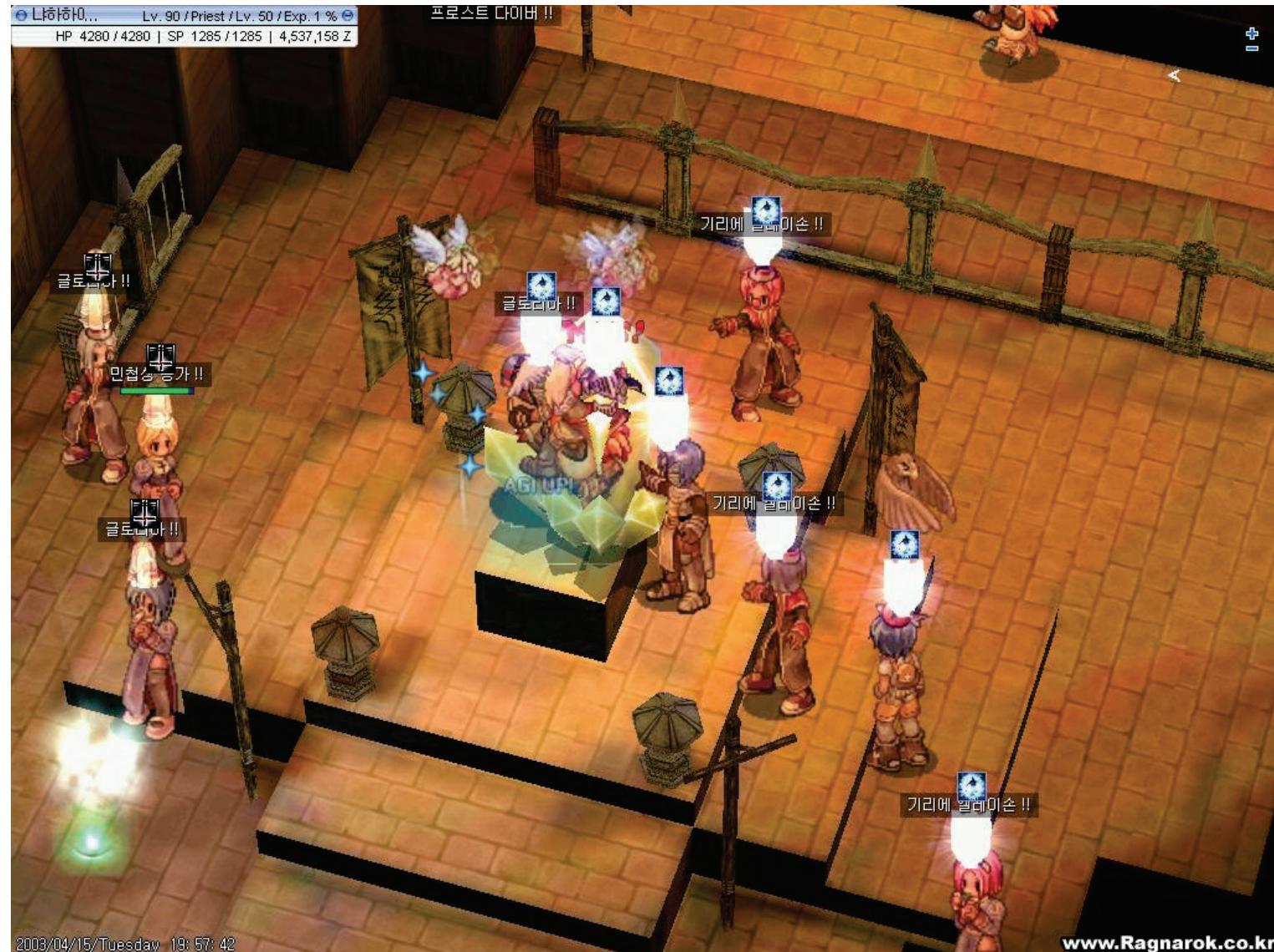
Our Solution I: Traffic Analysis

Q: Whether a bot is controlling a game client given the traffic stream it generates?

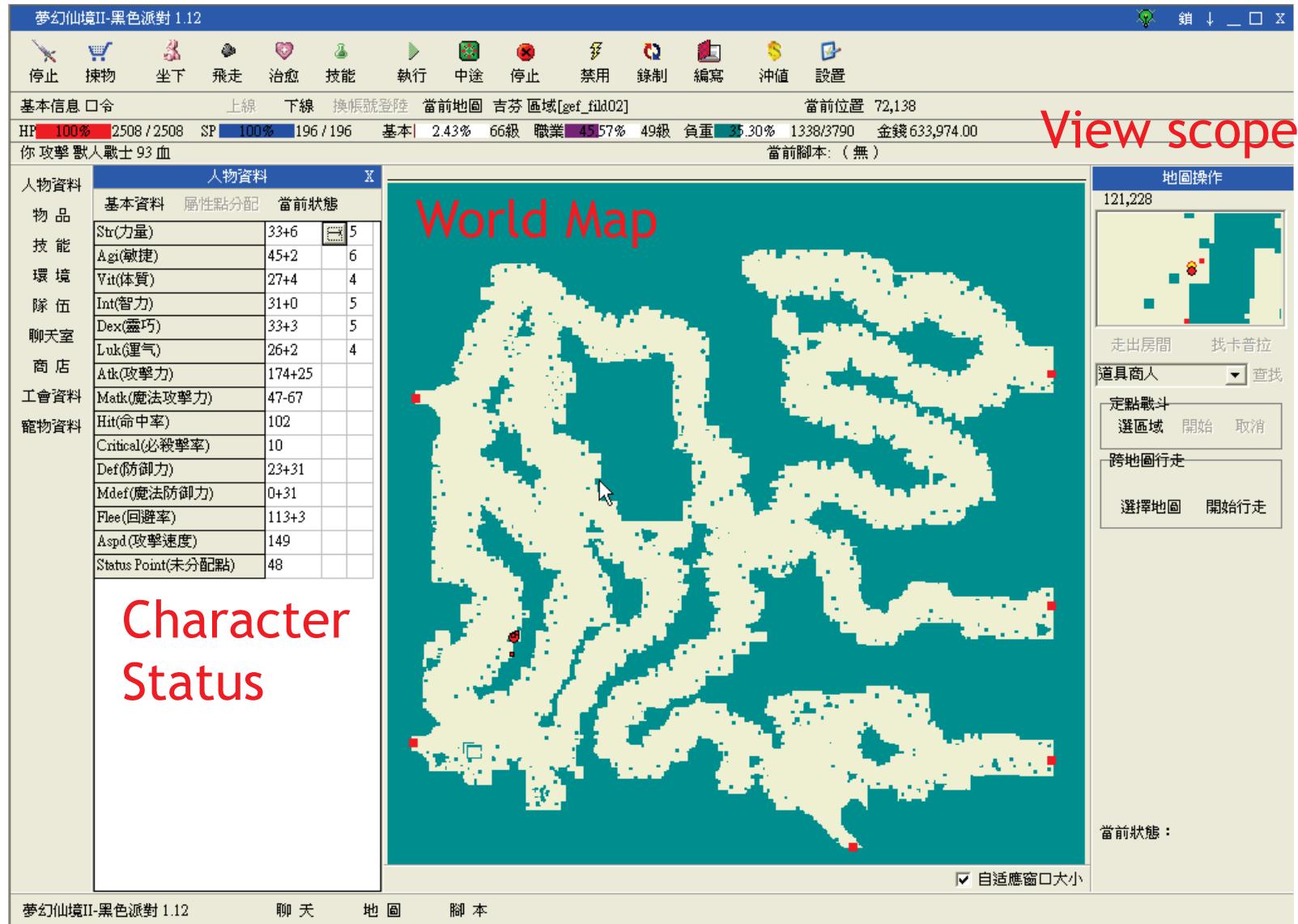
A: Yes or No



Case Study: Ragnarok Online



DreamRO -- A screen shot



Trace Collection

Category	Tr#	ID	Avg. Period	Avg. Pkt rate	Network
Human players	8	A, B, C, D	2.6 hr	1.0 / 3.2 pkt/s	ADSL, Cable Modem, Campus Network
Bots	11	K (Kore) R (DreamRO)	17 hr	1.0 / 2.2 pkt/s	

Heterogeneity in player skills and network conditions

Category	participants	Client pkt rate	Avg. RTT	Avg. Loss rate
Human players	2 rookies 2 experts	0.8 ~ 1.2 pkt/s	45 ~ 192 ms	0.01% ~ 1.73%
Bots	2 bots	0.5 ~ 1.7 pkt/s	33 ~ 97 ms	0.004% ~ 0.2%

207 hours, 3.8 million packets were traced in total

Command Timing

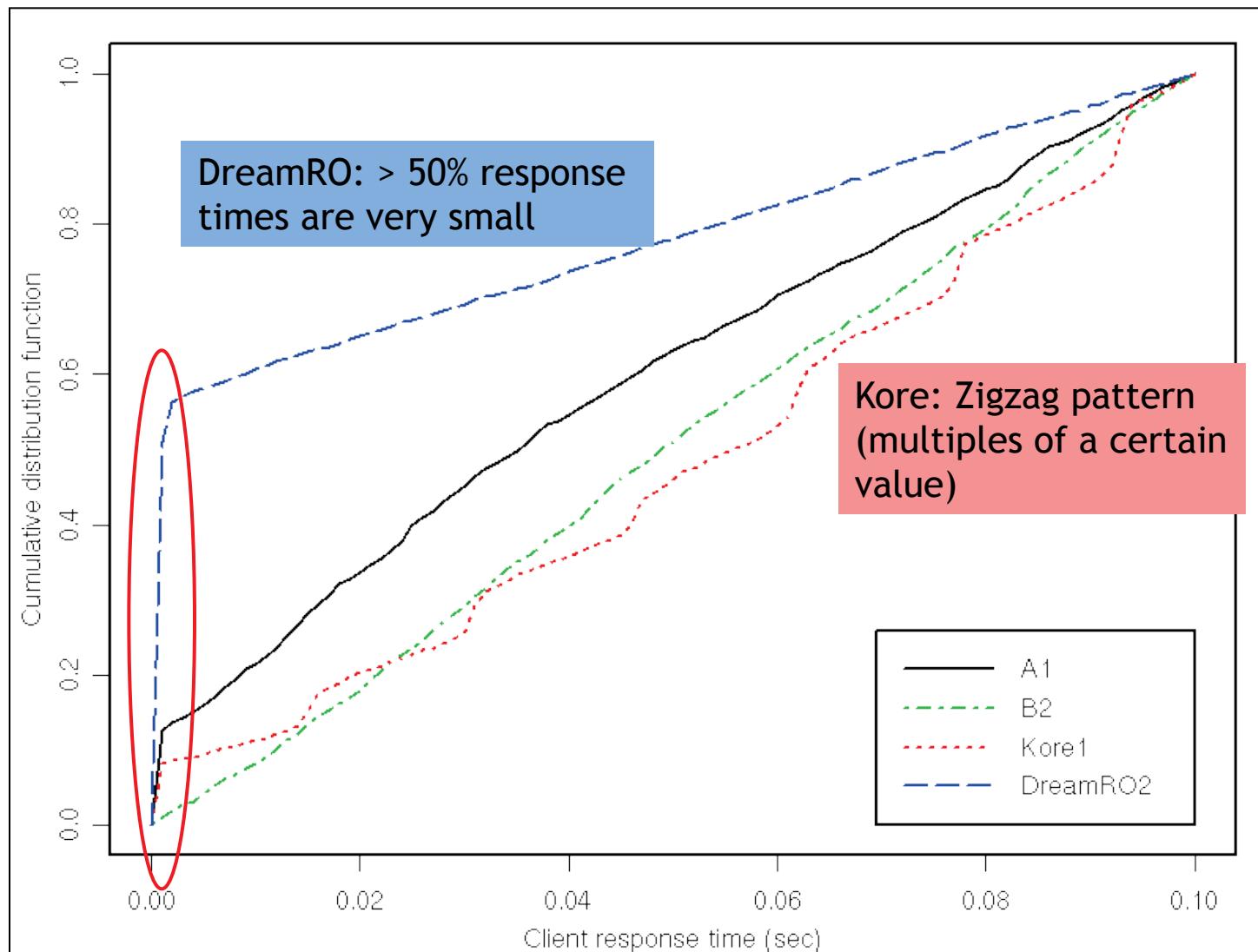


Observation

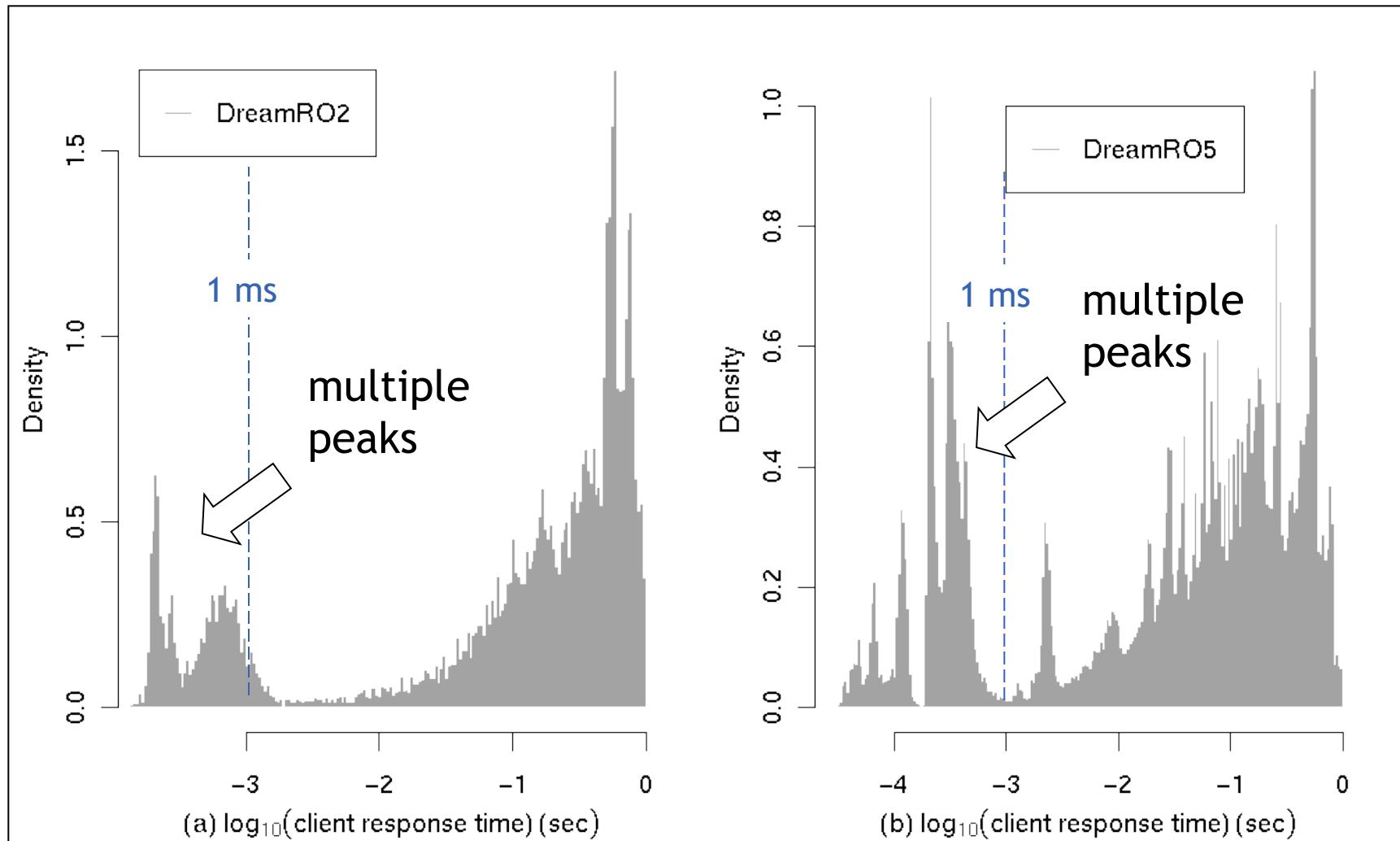
bots often issue their commands based on **arrivals of server packets**, which carry the latest status of the character and environment

- Client response time (response time):
time difference between the client packet departure time and the most recent server packet arrival time
- We expect the following patterns:
 - A large number of small response times (bots respond server packets immediately)
 - Regularity in response times

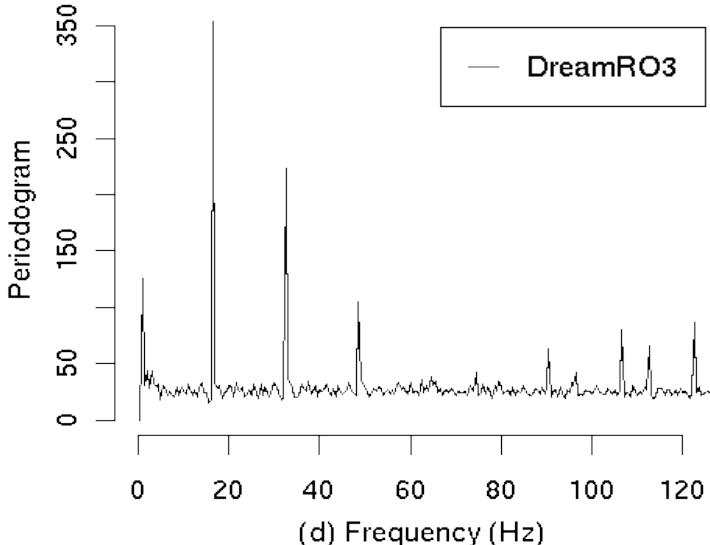
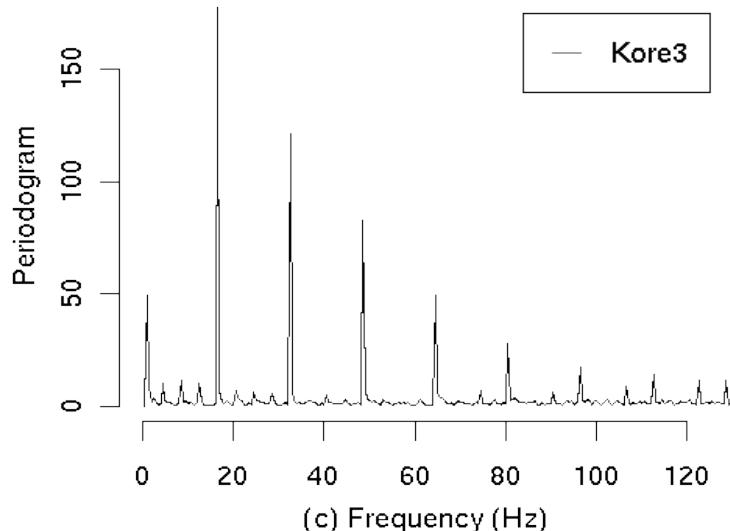
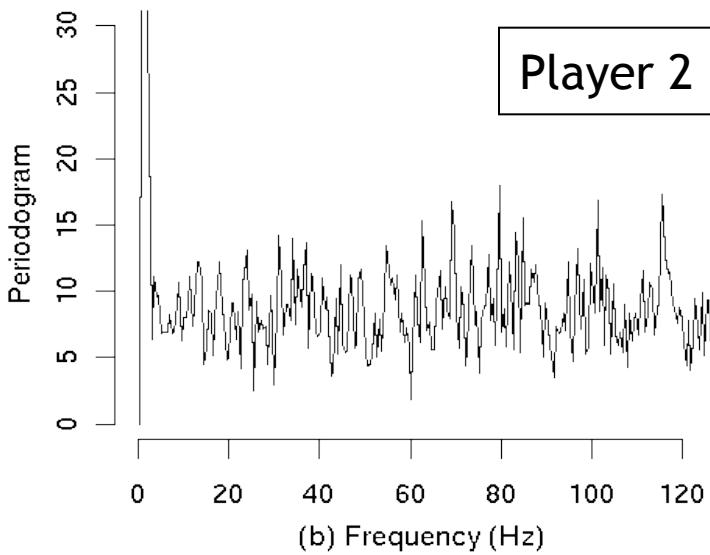
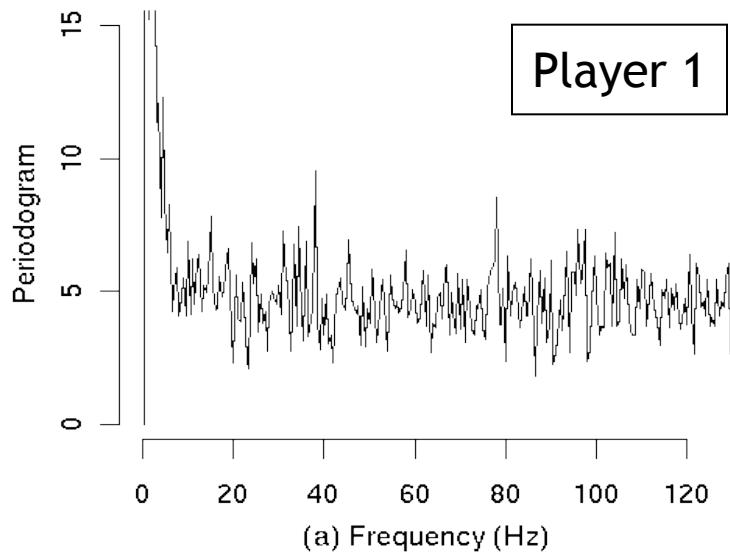
CDF of Client Response Times



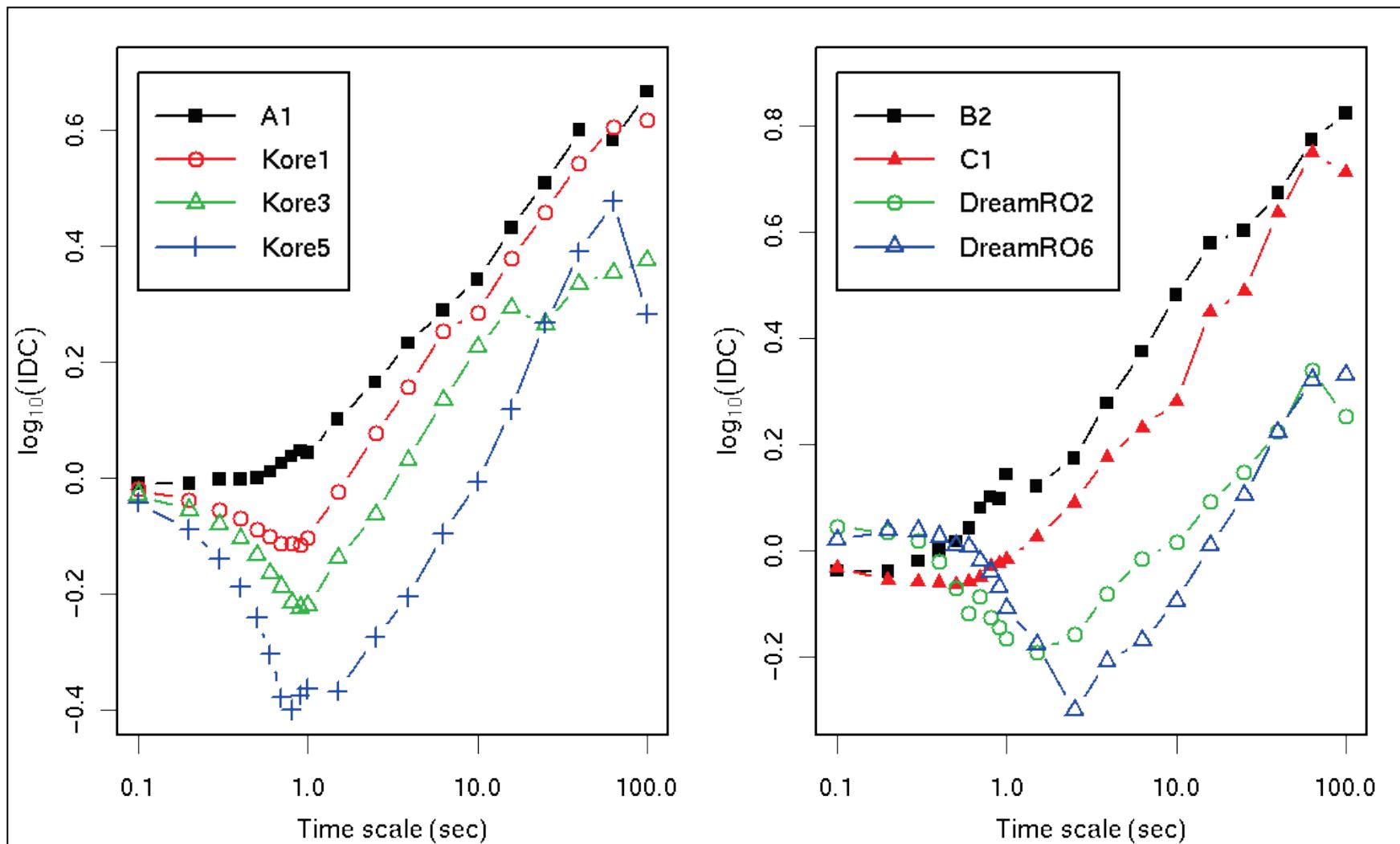
Histograms of Response Times



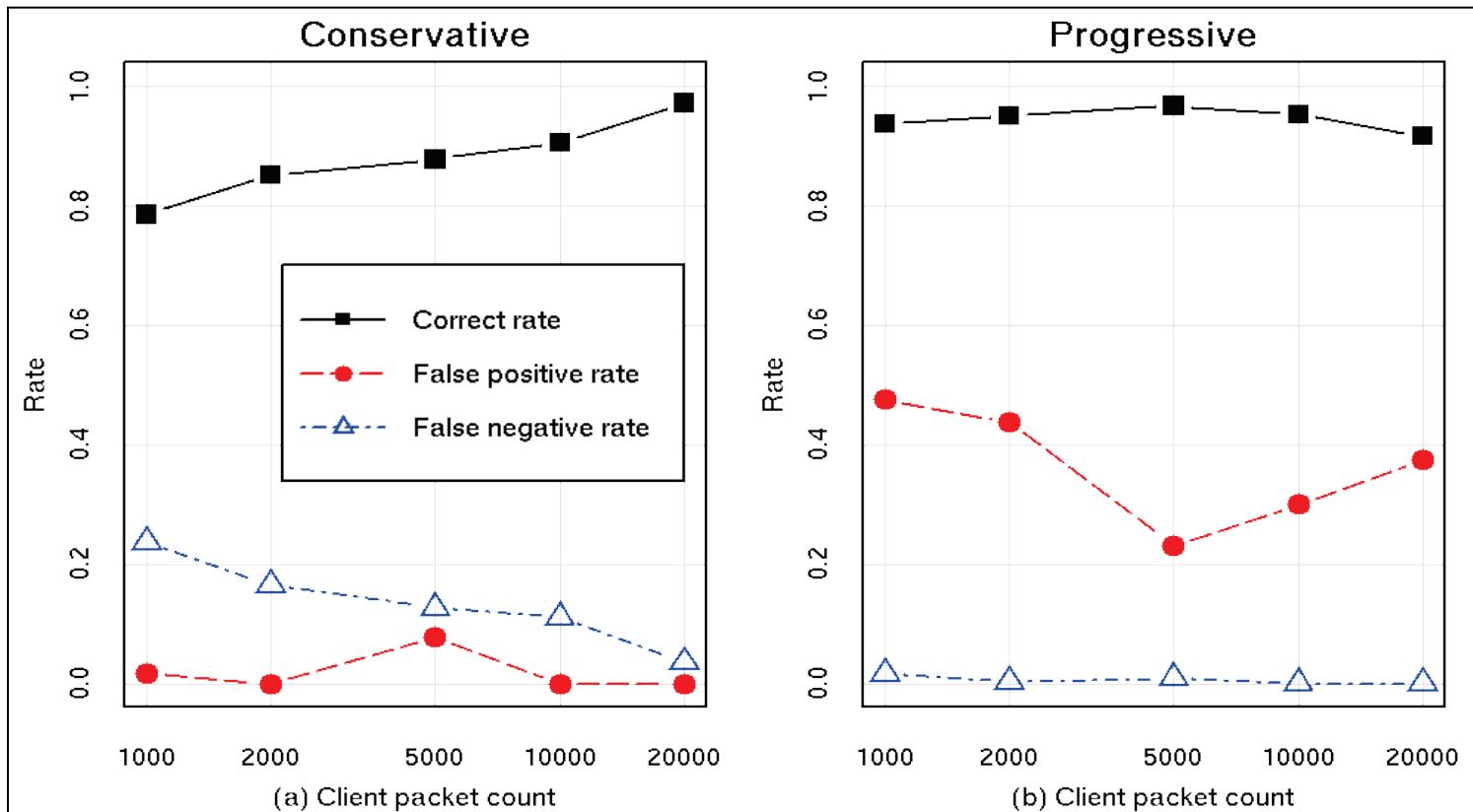
Periodograms of Histograms of Response times



Examining the Trend of Traffic Burstiness



An Integrated Classifier

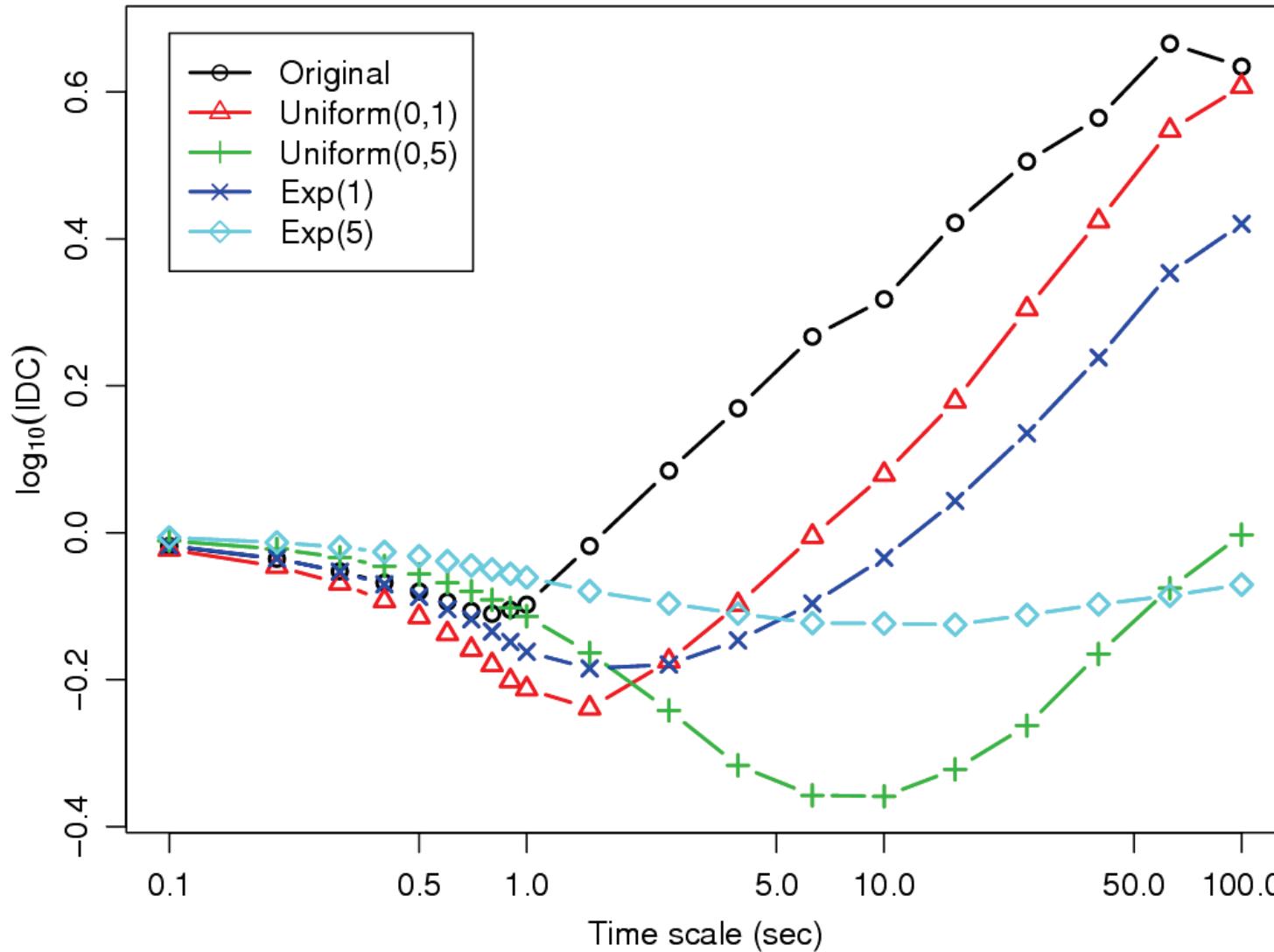


Progressive approach (2000 packets):
false negative rate < 1% and 95% correct rate

Robustness against Counter Attacks

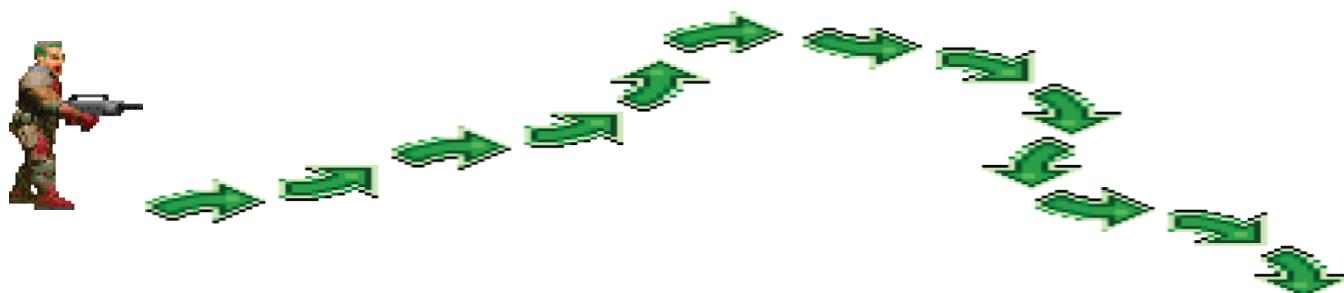
- Adding random delays to the release time of client commands
 - Command timing scheme will be ineffective
 - Schemes based on traffic burstiness and human reaction to network conditions are robust
 - Adding random delay to command timing will not eliminate the regularity unless the added delay is longer than the updating interval by orders of magnitude or heavy-tailed
 - However, adding such long delays will make the bots incompetent as this will slowdown the character's speed by orders of magnitude

The IDC of the original packet arrival process and that of intentionally-delayed versions



Our Solution II: Movement Trajectory

- Based on the **avatar's movement trajectory** in game
- Applicable for all genres of games where players control the avatar's movement directly
- Avatar's trajectory is **high-dimensional** (both in time and spatial domain)



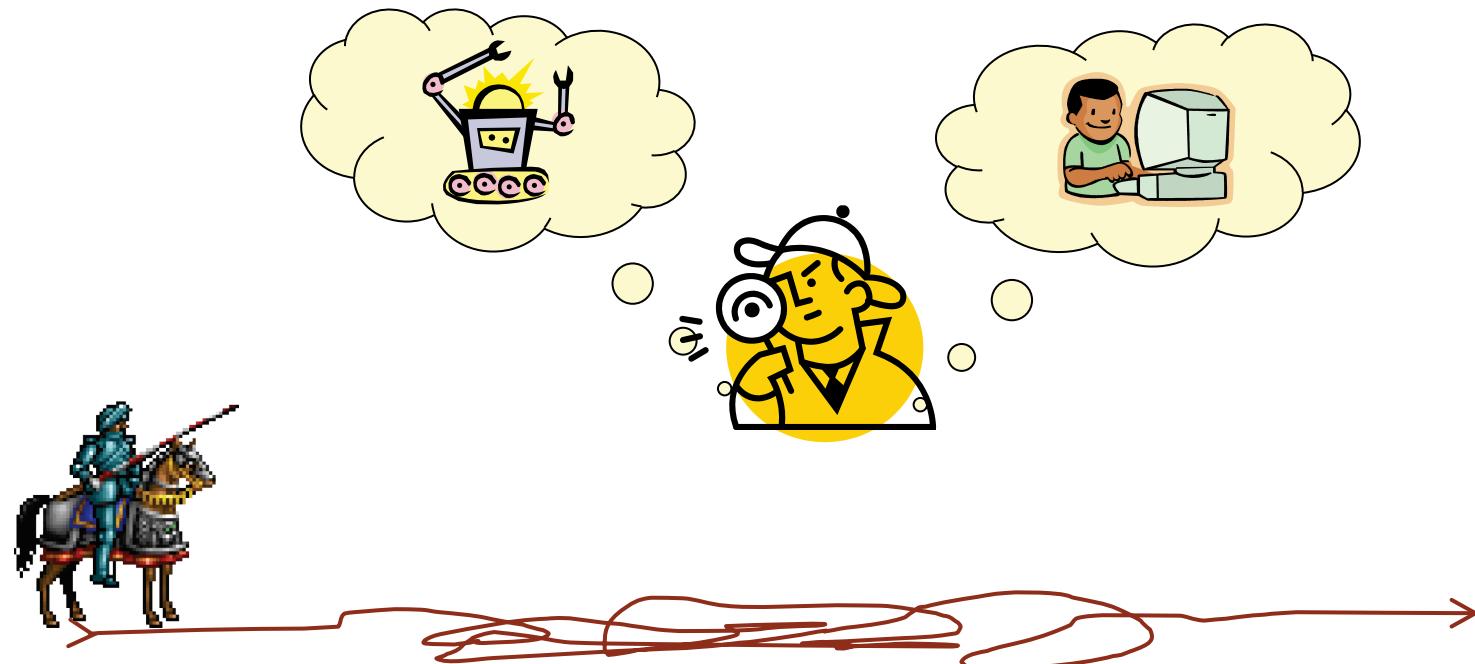
The Rationale behind Our Scheme

- The trajectory of the avatar controlled by a human player is hard to simulate for two reasons:
 - **Complex context information:**
Players control the movement of avatars based on their knowledge, experience, intuition, and a great deal of environmental information in game.
 - **Human behavior is not always logical and optimal**
- How to model and simulate realistic movements (for game agents) is still an open question in the AI field.

Bot Detection: A Decision Problem

Q: Whether a bot is controlling a game client given the movement trajectory of the avatar?

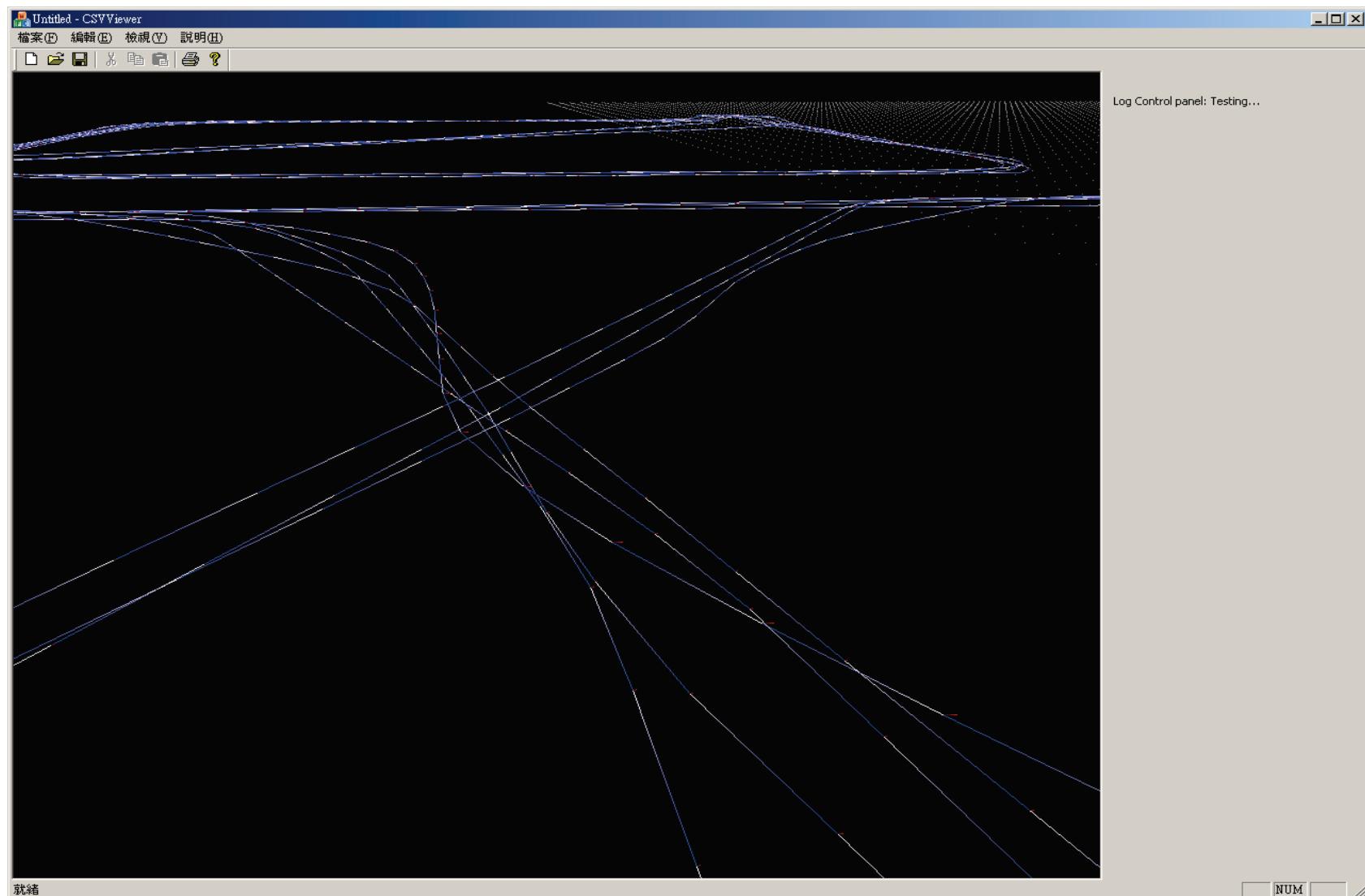
A: Yes / No?



User Movement Trails



3D Path Visualization Tool



Case Study: Quake 2



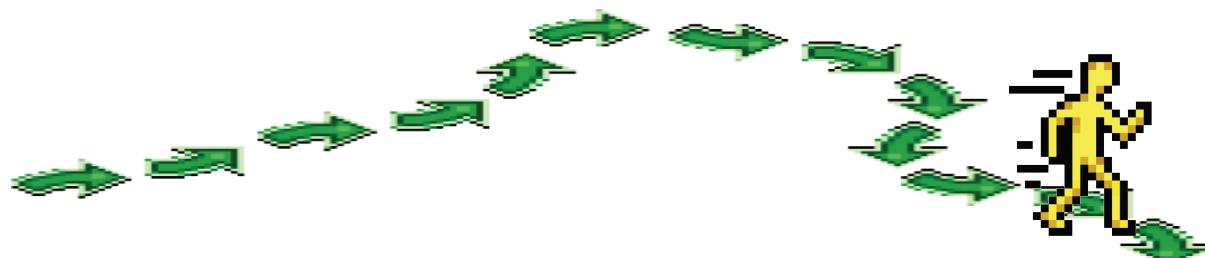
Data Collection

- Human traces downloaded from fan sites including GotFrag Quake, Planet Quake, Demo Squad, and Revilla Quake Site
- Bot traces collected on our own Quake server
 - CR BOT 1.14
 - Eraser Bot 1.01
 - ICE Bot 1.0
- Totally 143.8 hours of traces were collected

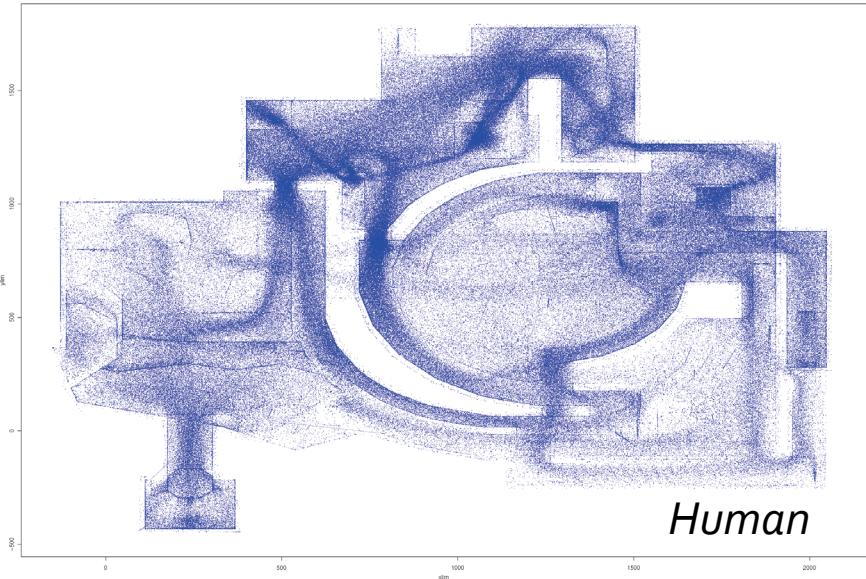
Name	Number	Trace Length	Total	Active
Human	282	1000 seconds	78.0 hours	89%
CR	75	1000 seconds	20.8 hours	89%
Eraser	102	1000 seconds	28.3 hours	92%
ICE	60	1000 seconds	16.7 hours	67%

Data Representation

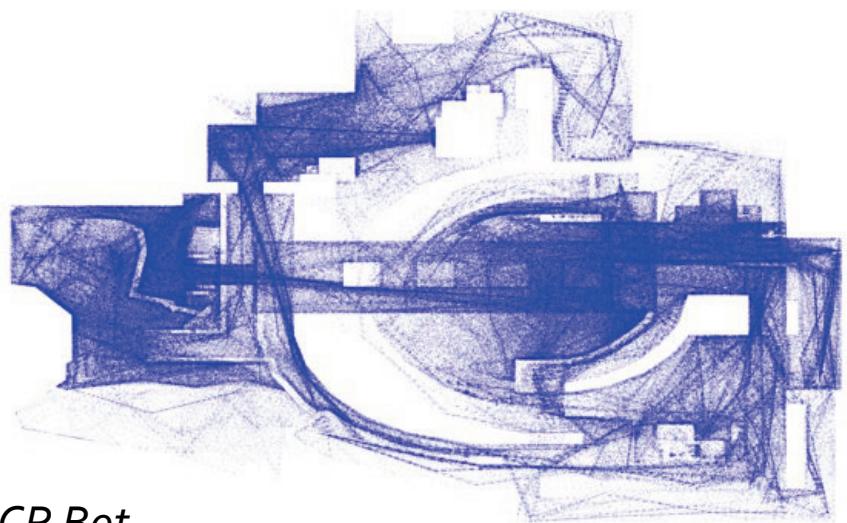
Time(sec.)	Trace			
	S_1	S_2	...	S_N
1	164.87, -258.87	1395.75, -172.25	...	569.37, -45.62
2	159.87, -259.87	1363.50, -171.25	...	586.37, -36.12
3	157.66, -264.42	1340.22, -168.20	...	585.74, -33.67
:	:	:	...	:
i	527.87, 788.00	2045.75, 401.37	...	-5.75, 108.25
:	:	:	...	:
t	984.00, 192.00	497.75, 1289.62	...	511.12, 1433.25
t	(X, Y)	(X, Y)	(X, Y)	



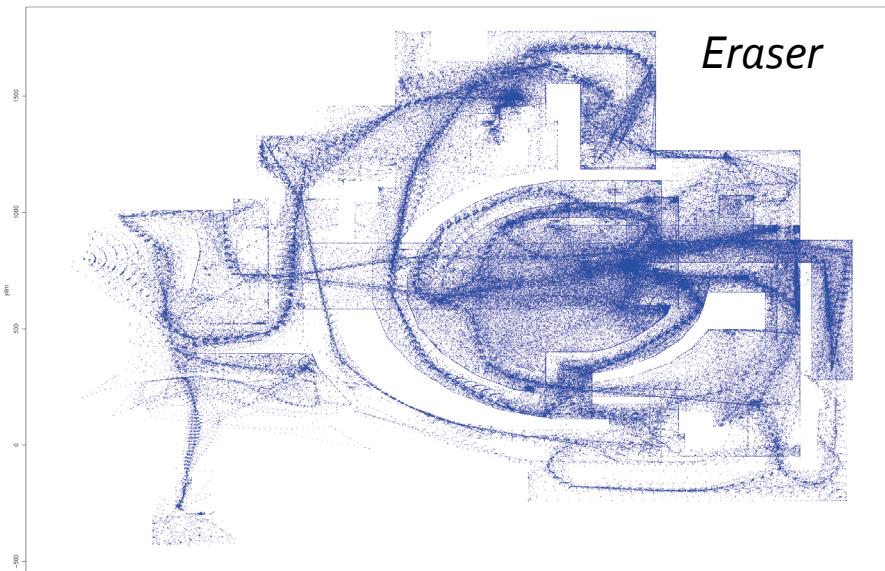
Aggregate View of Trails (Human & 3 Bots)



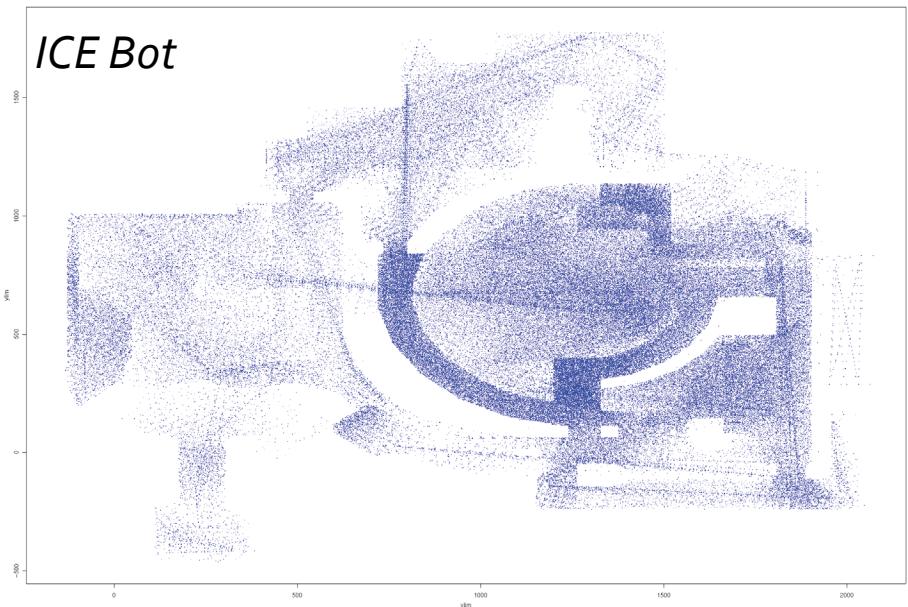
Human



CR Bot

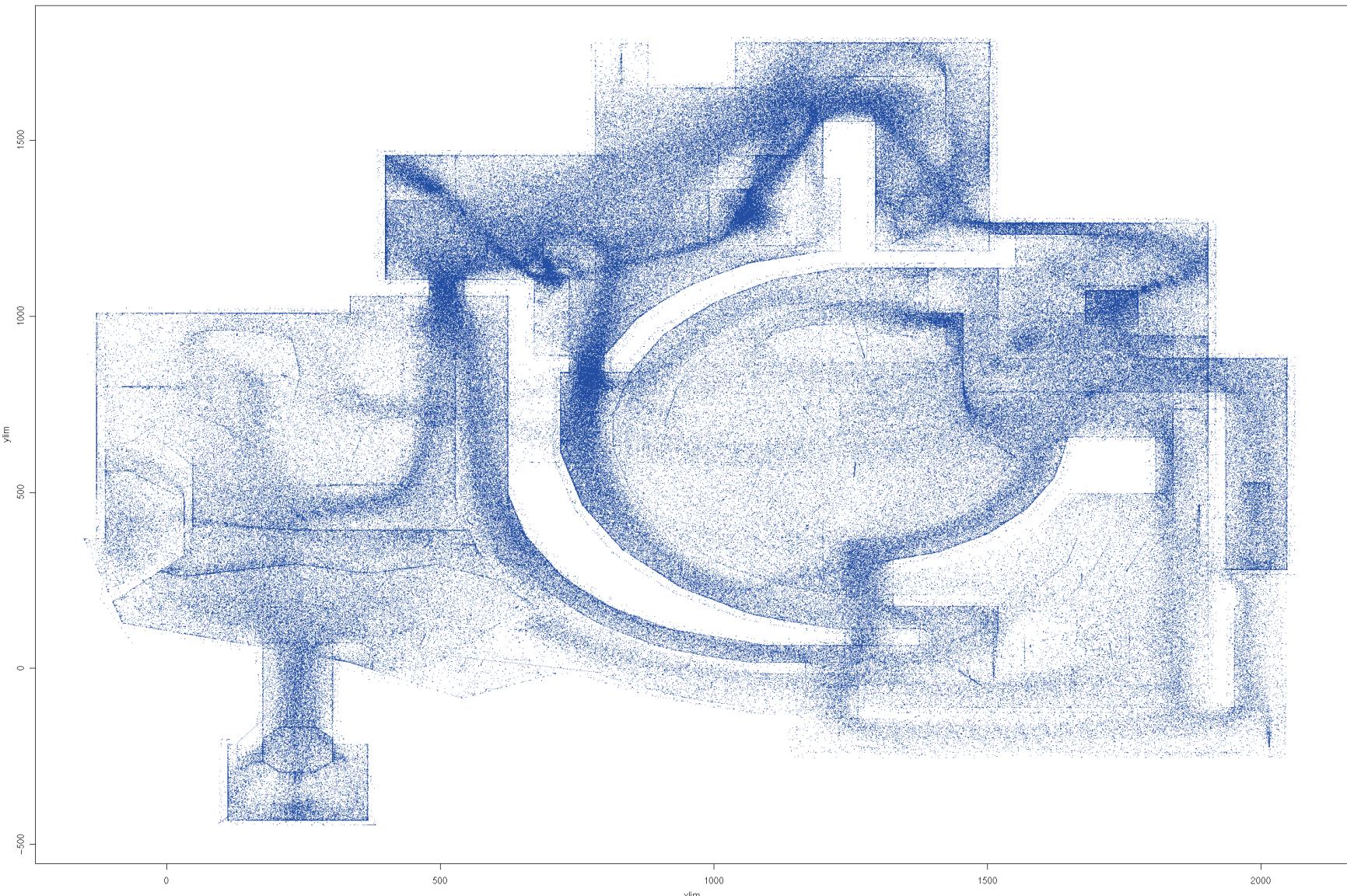


Eraser

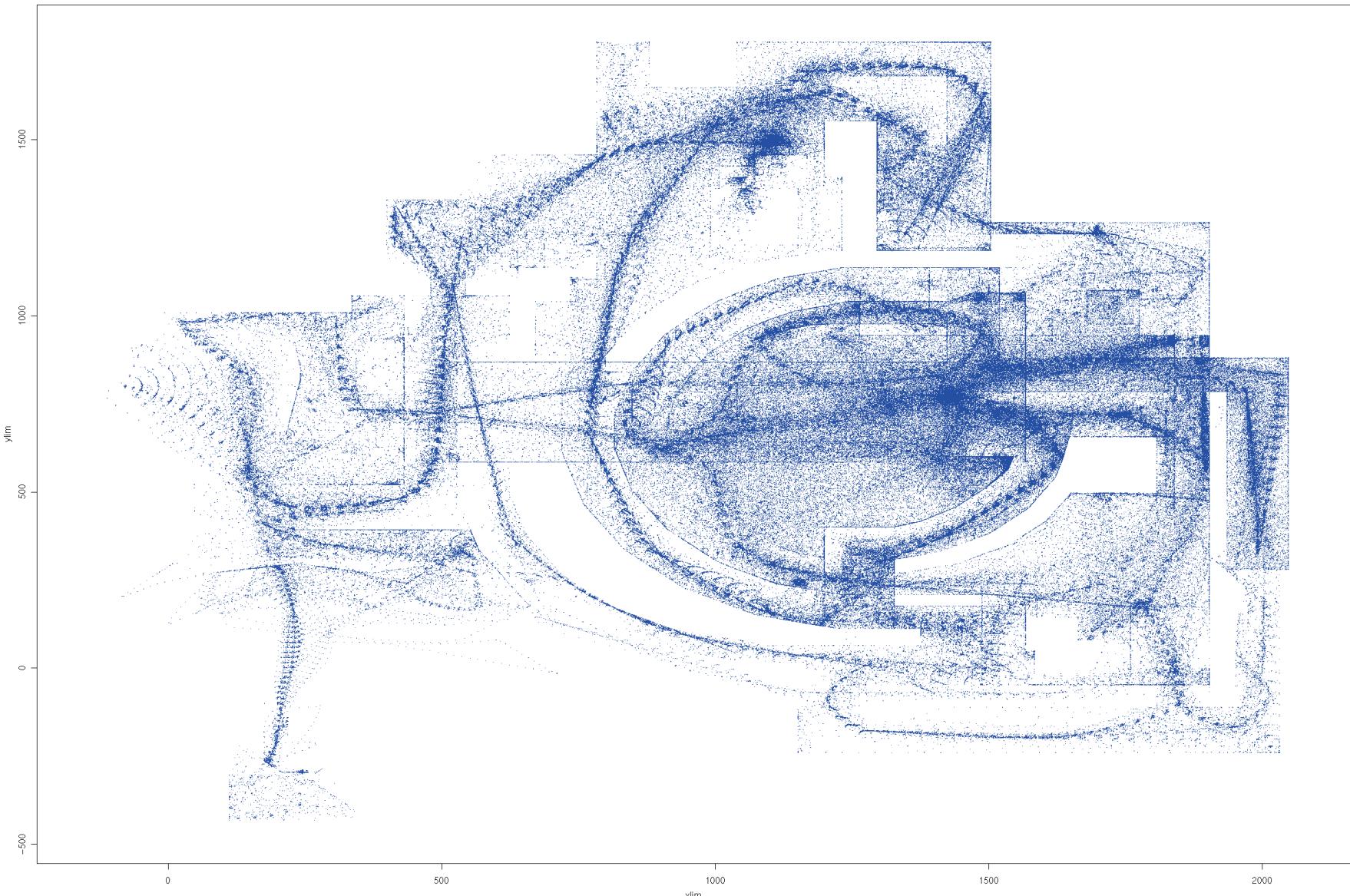


ICE Bot

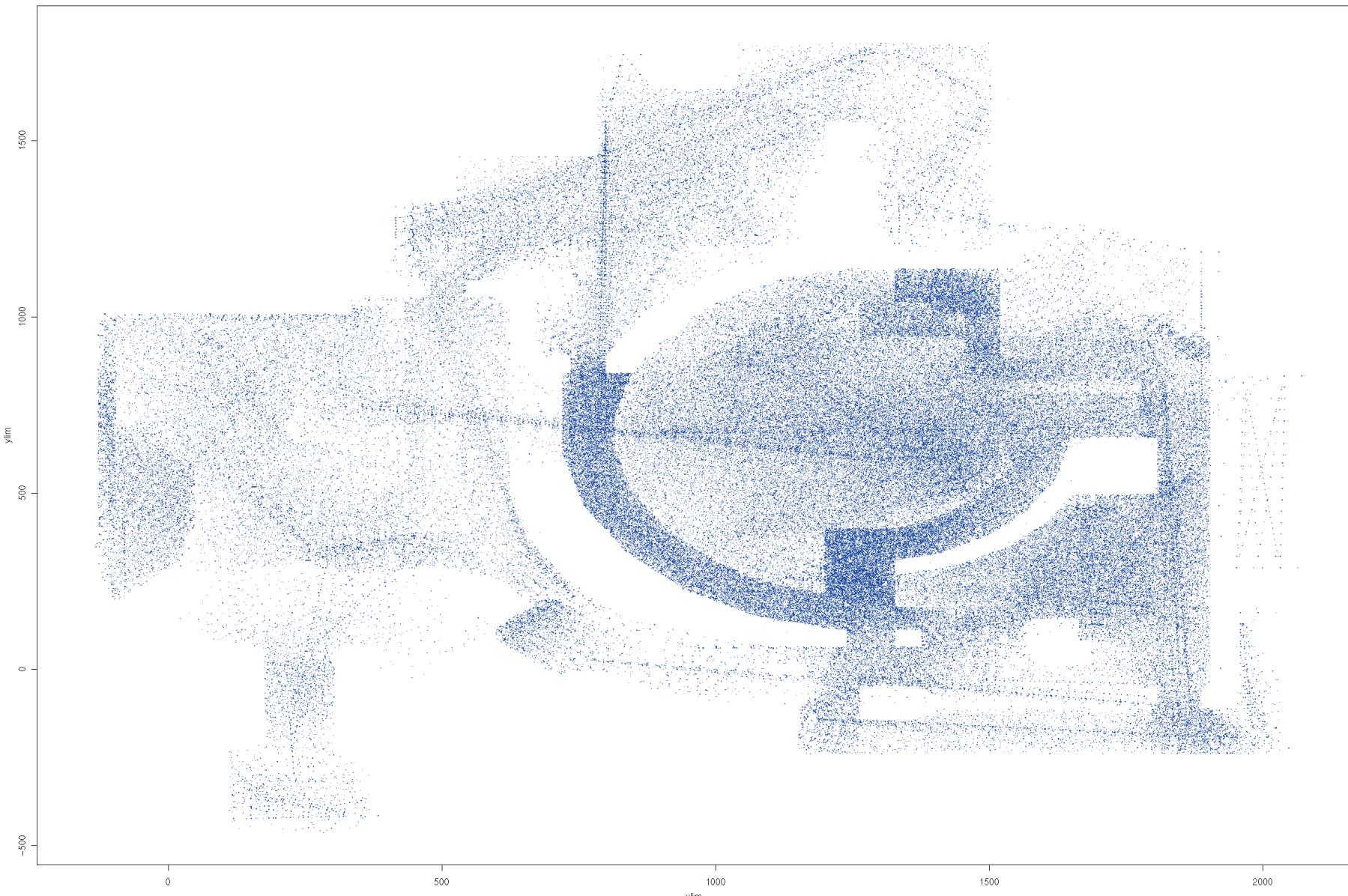
Trails of Human Players



Trails of Eraser Bot



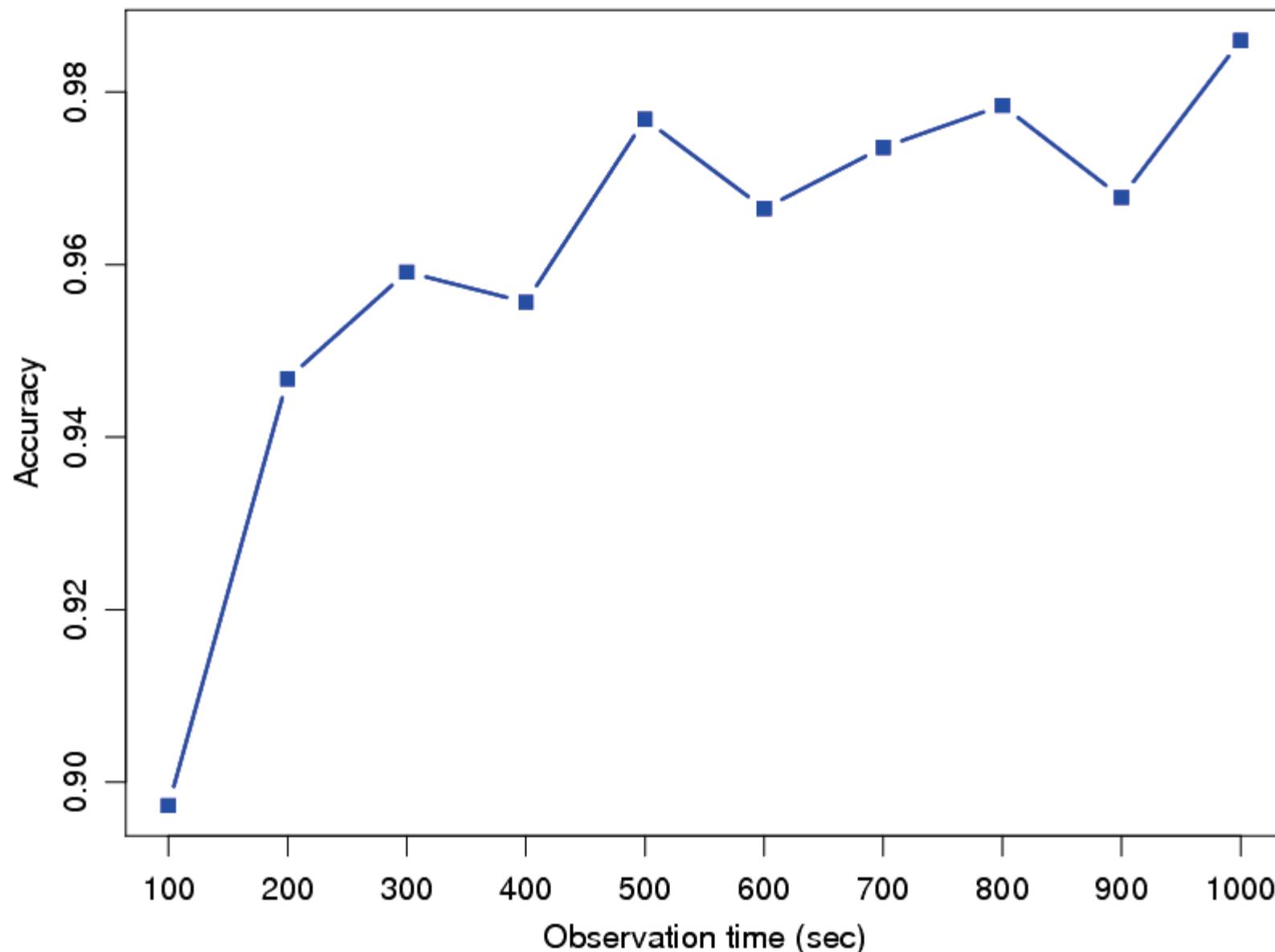
Trails of ICE Bot



Movement Trail Analysis

- Activity
 - mean/sd of ON/OFF periods
- Pace
 - speed/offset in each time period
 - teleportation frequency
- Path
 - linger frequency/length
 - smoothness
 - detourness
- Turn
 - frequency of mild turn, U-turn, ...

Bot Detection Performance



Step 1. Pace Vector Construction

- For each trace s_n , we compute the pace (distance) in successive two seconds by

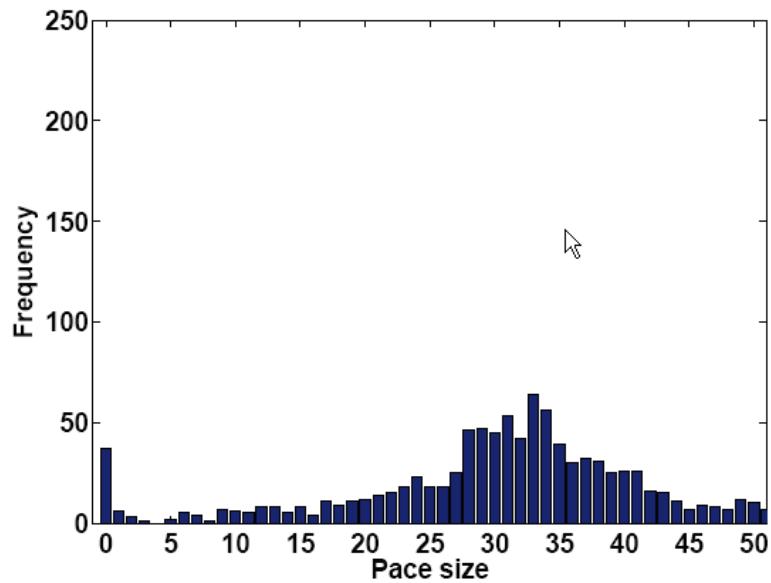
$$\|\mathbf{s}_{n,i+1} - \mathbf{s}_{n,i}\| = \sqrt{(\mathbf{s}_{n,i+1} - \mathbf{s}_{n,i})^T (\mathbf{s}_{n,i+1} - \mathbf{s}_{n,i})}$$

- We then compute the distribution (histogram) of paces with a fixed bin size by

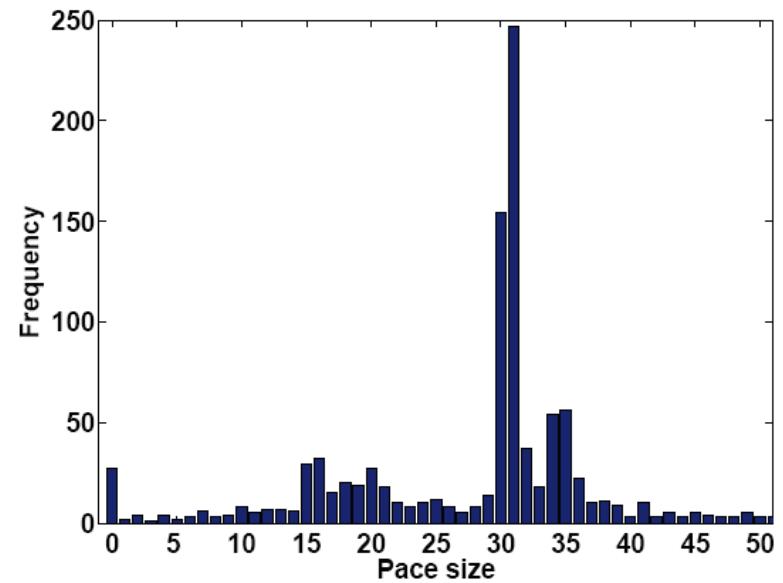
$$\mathbf{F}_n = (f_{n,1}, f_{n,2}, \dots, f_{n,B})$$

where B is the number of bins in the distribution.

Pace Vector: An Example



(A) Human



(B) Bot

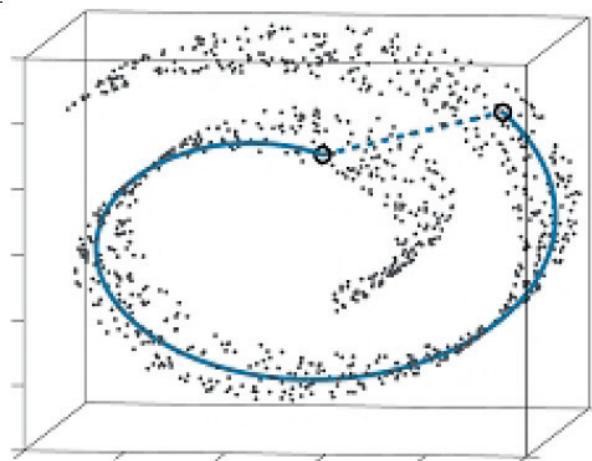
B is set to 200 (dimensions) in this work

Step 2. Dimension Reduction with Isomap

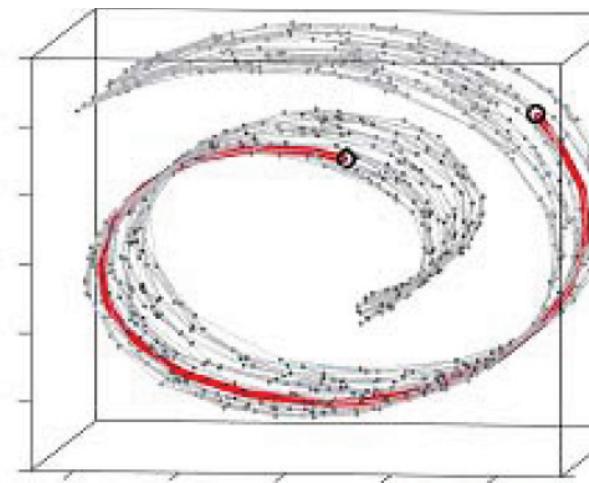
- We adopt **Isomap** for **nonlinear dimension reduction** for
 - Better classification accuracy
 - Lower computation overhead in classification
- **Isomap**
 - Assume data points lie on a manifold

A mathematical space in which every point has a **neighborhood** which resembles Euclidean space, but in which the global structure may be more complicated. (Wikipedia)
 - 1. Construct the neighborhood graph by kNN (k-nearest neighbor)
 - 2. Compute the **shortest geodesic path** for each pair of points
 - 3. Reconstruct data by MDS (multidimensional scaling)

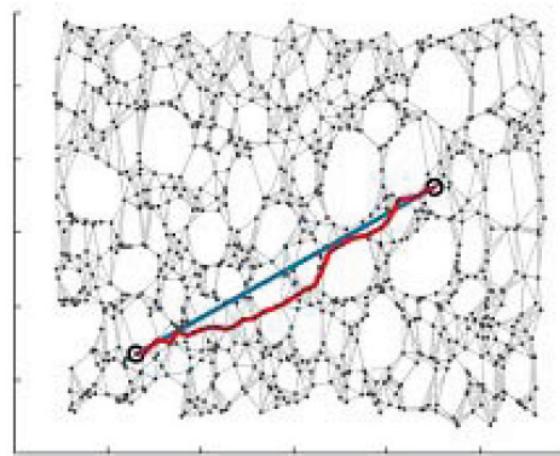
A Graphic Representation of Isomap



(A) A Swiss Roll Data

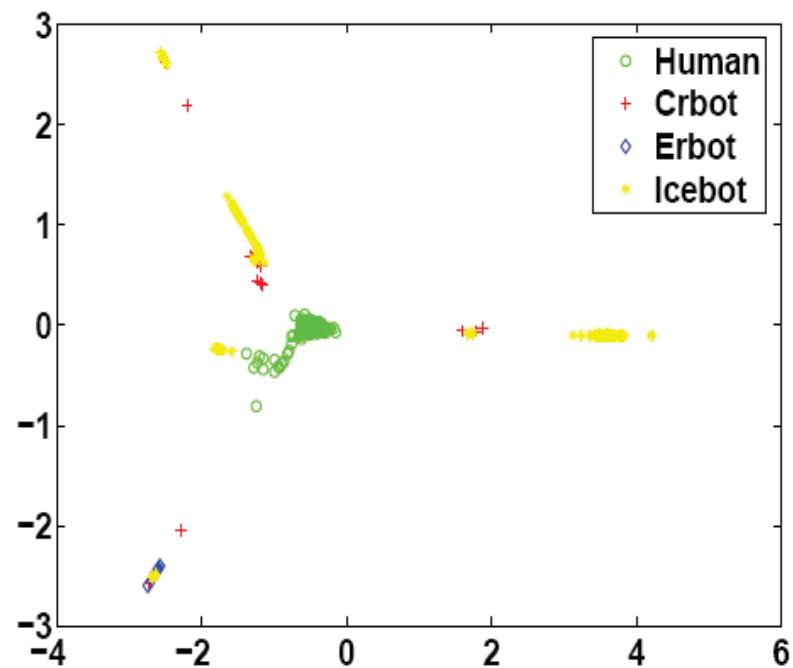
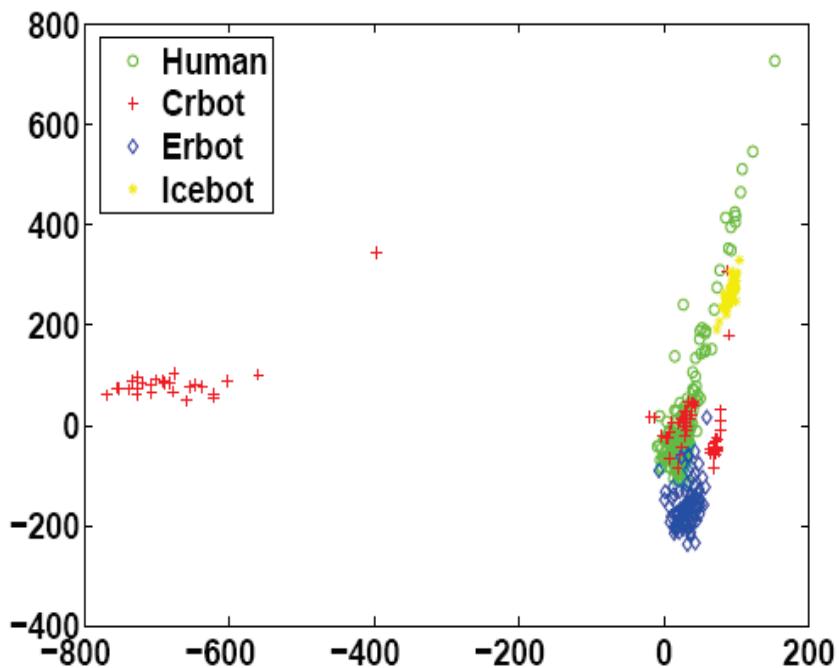


(B) Neighborhood Graph



(C) After Mapping by Isomap

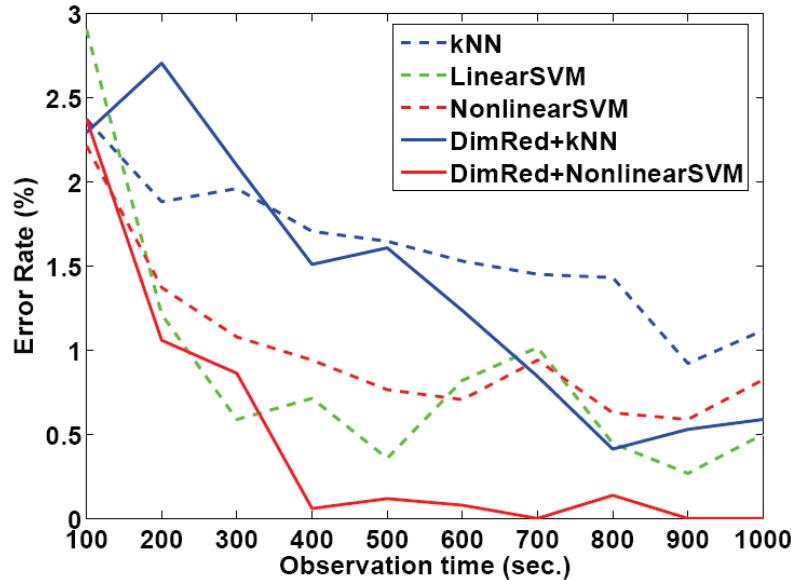
PCA (Linear) vs. Isomap (Nonlinear)



Five Methods for Comparison

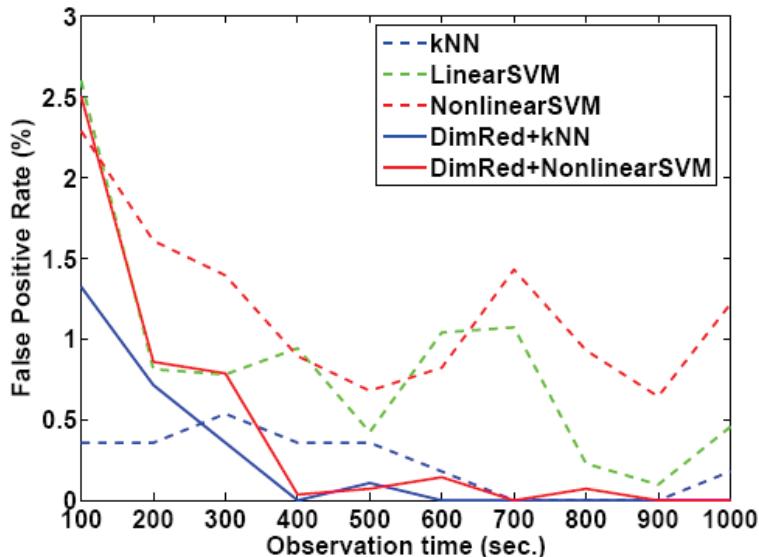
Method	Data Input
kNN	Original 200-dimension Pace Vectors
Linear SVM	Original 200-dimension Pace Vectors
Nonlinear SVM	Original 200-dimension Pace Vectors
Isomap + kNN	Isomap-reduced Pace Vectors
Isomap + Nonlinear SVM	Isomap-reduced Pace Vectors

Evaluation Results

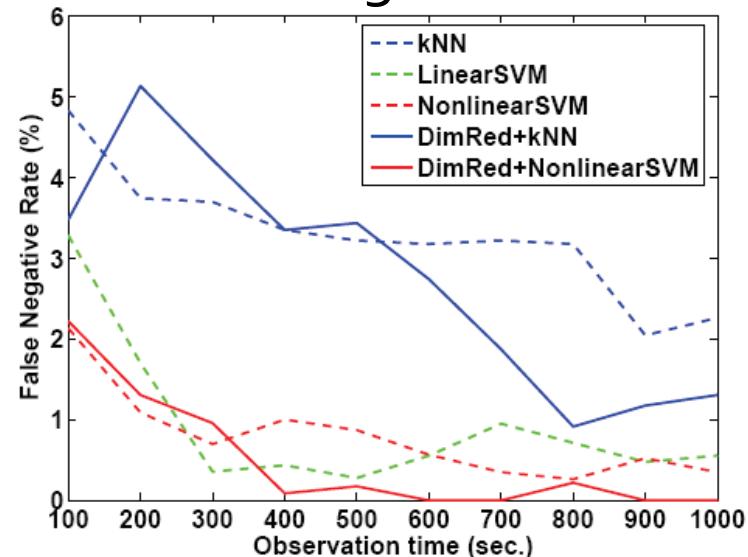


Error Rate

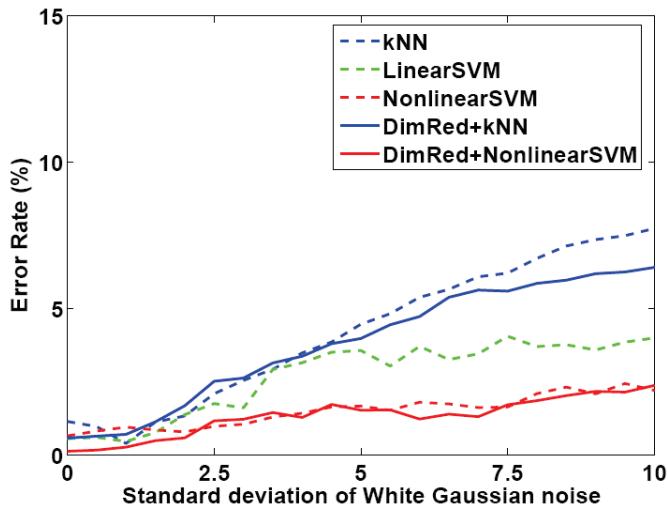
False Positive Rate



False Negative Rate

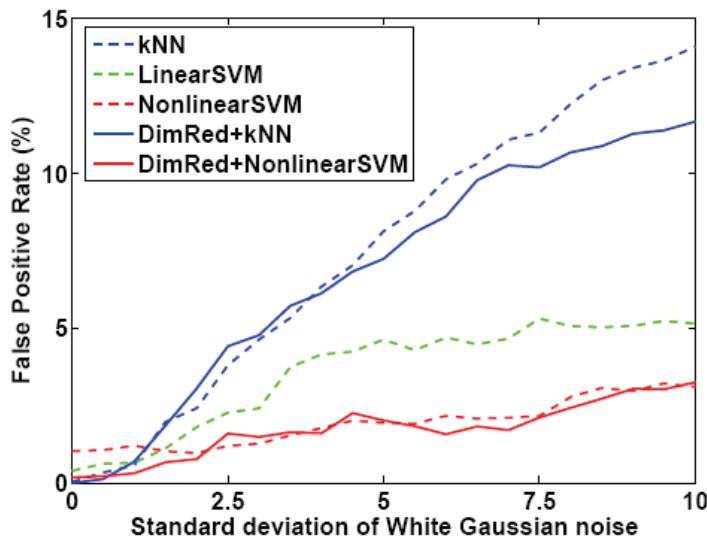


Evaluation Results

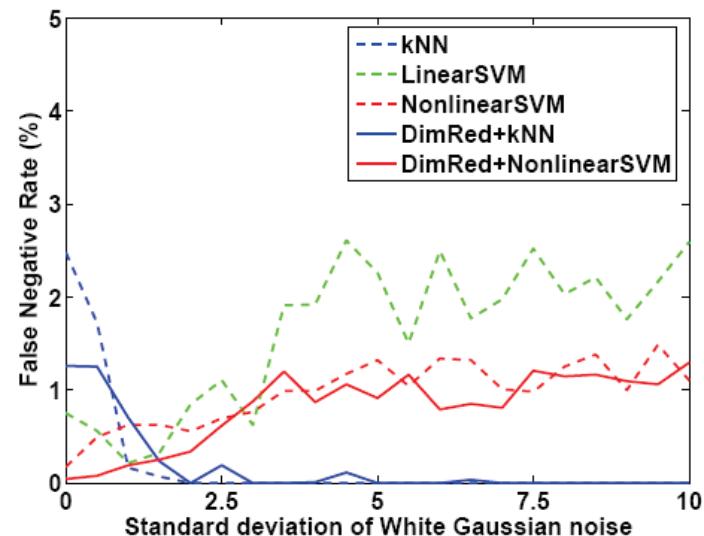


Error Rate

False Positive Rate



False Negative Rate



User Behavior Topics

Game-Play Time Prediction

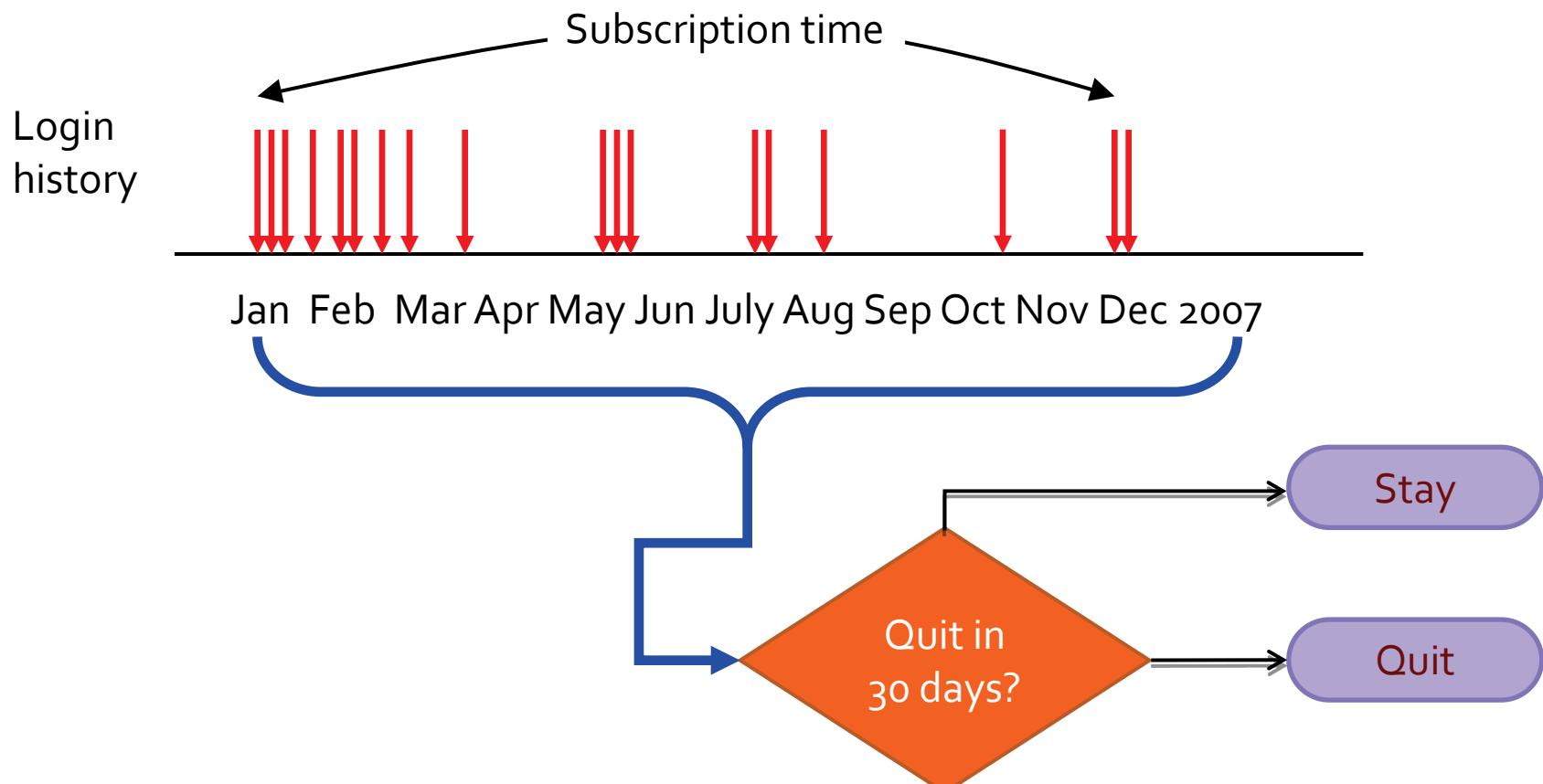


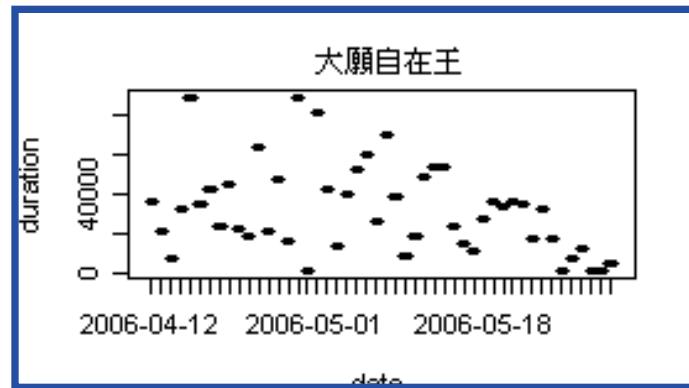
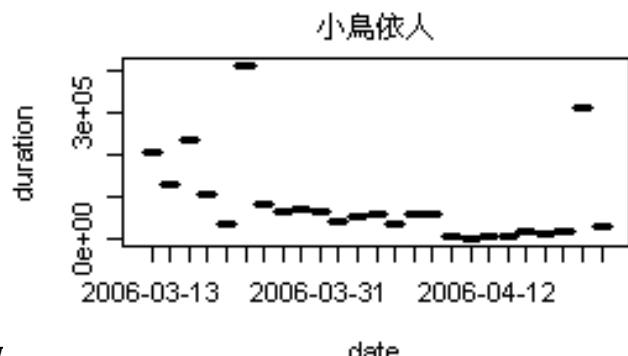
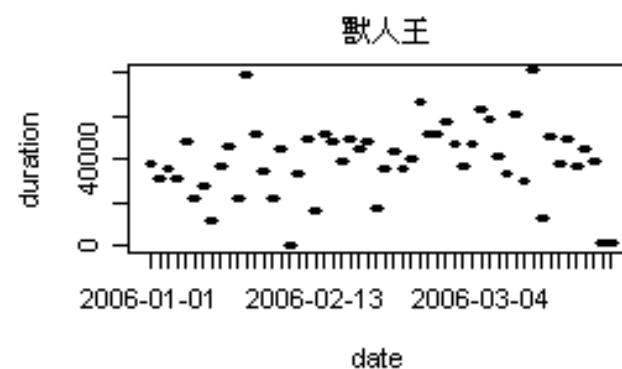
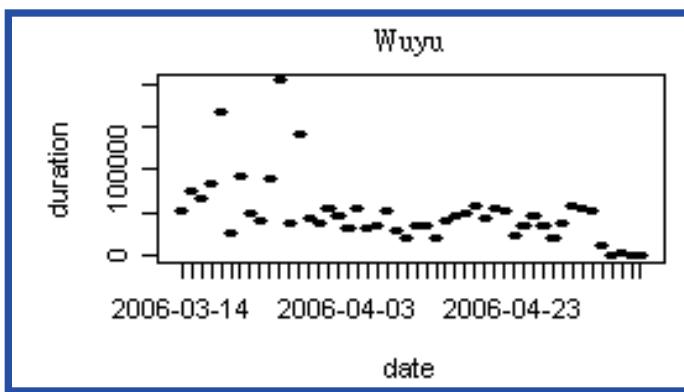
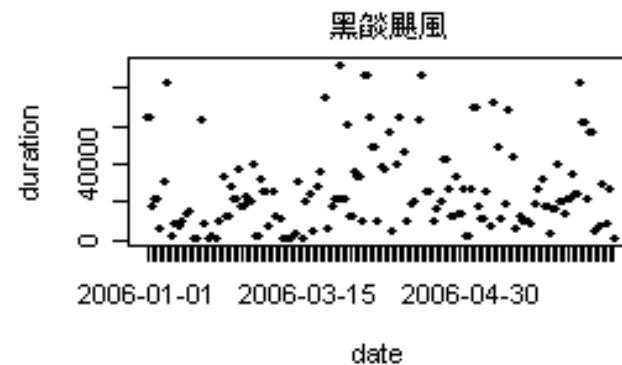
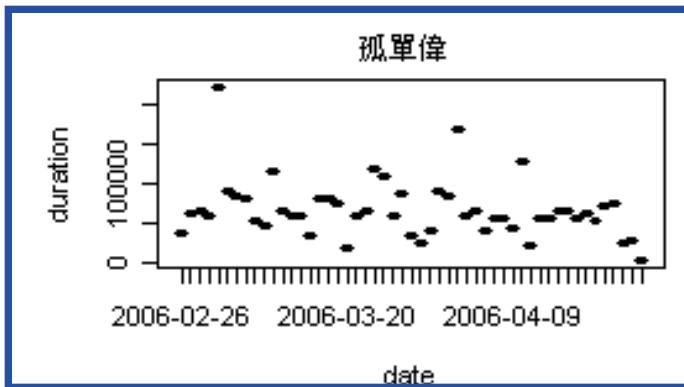
Unsubscription Prediction

- Game improvement
 - Players' unsubscription → low satisfaction
 - Surveys can be conducted to determine the causes of player dissatisfaction and improve the game accordingly
 - More likely to receive useful comments before players quit
- Prevent VIP players' quitting (maintain revenue)
 - For "item mall" model, users' contribution (of revenue) is heavy-tailed
 - Losing VIP players may significantly harm the revenue
- Network/system planning and diagnosis
 - By predicting "which" players tend to leave the game → investigating is there any problem regarding network resource planning, network congestion, or server arrangement

Unsubscription Prediction: Our Proposal

- Rationale: players' satisfaction / enthusiasm / addiction to a game is embedded in her **game play history**



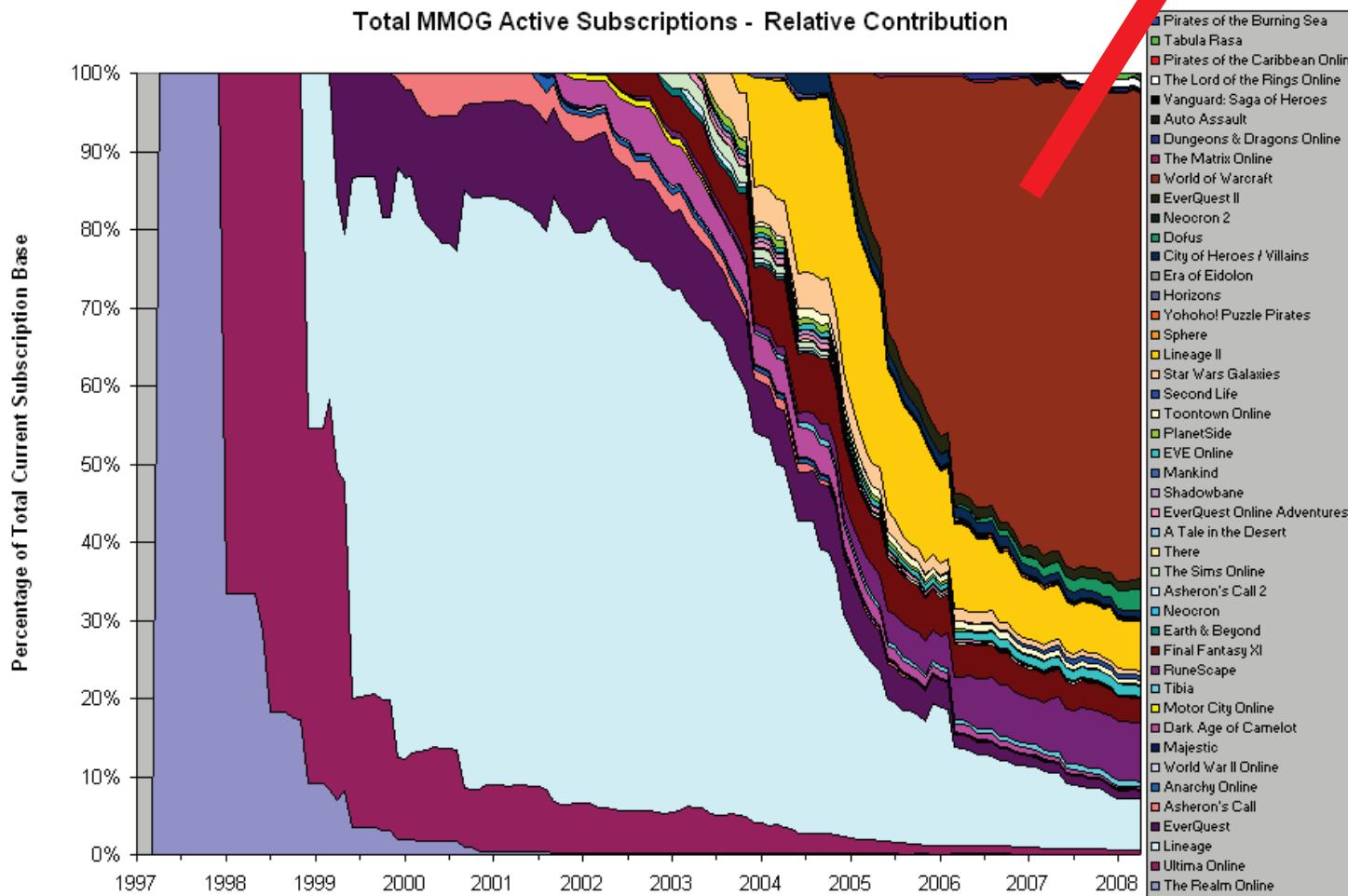




World of Warcraft



- The most popular MMOG for now



Data Collection Methodology

- Create a game character
- Use the command '`\who`'
- The command asks the game server to reply with a list of players who are currently online
- Write a specialized data-collection program (using C#, VBScript, and Lua)



Trace Summary

WoW trace	
Start date	2005-12-22
End date	2007-10-17
Length	664 days
Total sessions	1,672,820
Accounts observed	34,521

福克斯大神之謎？？(1)

網友A：不知道在聖光之願部落的玩家有沒有發現到，在新手村薩滿訓練師的後面，永遠都會站著一個叫「福克斯大神」的獵人玩家！在半年前我到聖光定居時我在新手村見到他，到現在他仍然還是留守在那個地方.....不會暫離，而且可以觀察他= ="

這種事該回報給GM嗎？創新手看到他的時候都覺得好恐怖啊囧

網友B：me too

看到的一瞬間 突然起雞皮疙瘩.....

網友C："已離去"玩家的怨念(怨魂@@)嗎?
還是在悲傷愛情故事裡,癡等所愛的另一人?

^^^^^QQ

網友D：哈 線在好多人在看噢
旁邊為了一大群人@@
觀光景點呀XD

ref. <http://forum.gamebase.com.tw/content.jsp?no=4715&cno=47150002&sno=75201947>
ref. <http://www.wings-of-narnia.com/viewtopic.php?t=3012>

福克斯大神之謎？？(2)

網友E：我剛剛也有去看了一下 開了一個ID叫做“聽說有鬼”的獸人戰士 坐在他面前的桶子一直望著他～忽然！

<暫離>福克斯大神
他蹲下了...隔一分鐘..消失 = ^ ="

..

..

現在我心裡也是毛毛的..

網友F：好猛鬼啊!!!!!!大神的力量好可怕啊,一堆信眾死在他之前！！！！！！

網友G：我上次有開過去看，還遇到了兩位同好，看的時候真的蠻不可思議的...

可以列入魔獸10大世界奇觀吧！

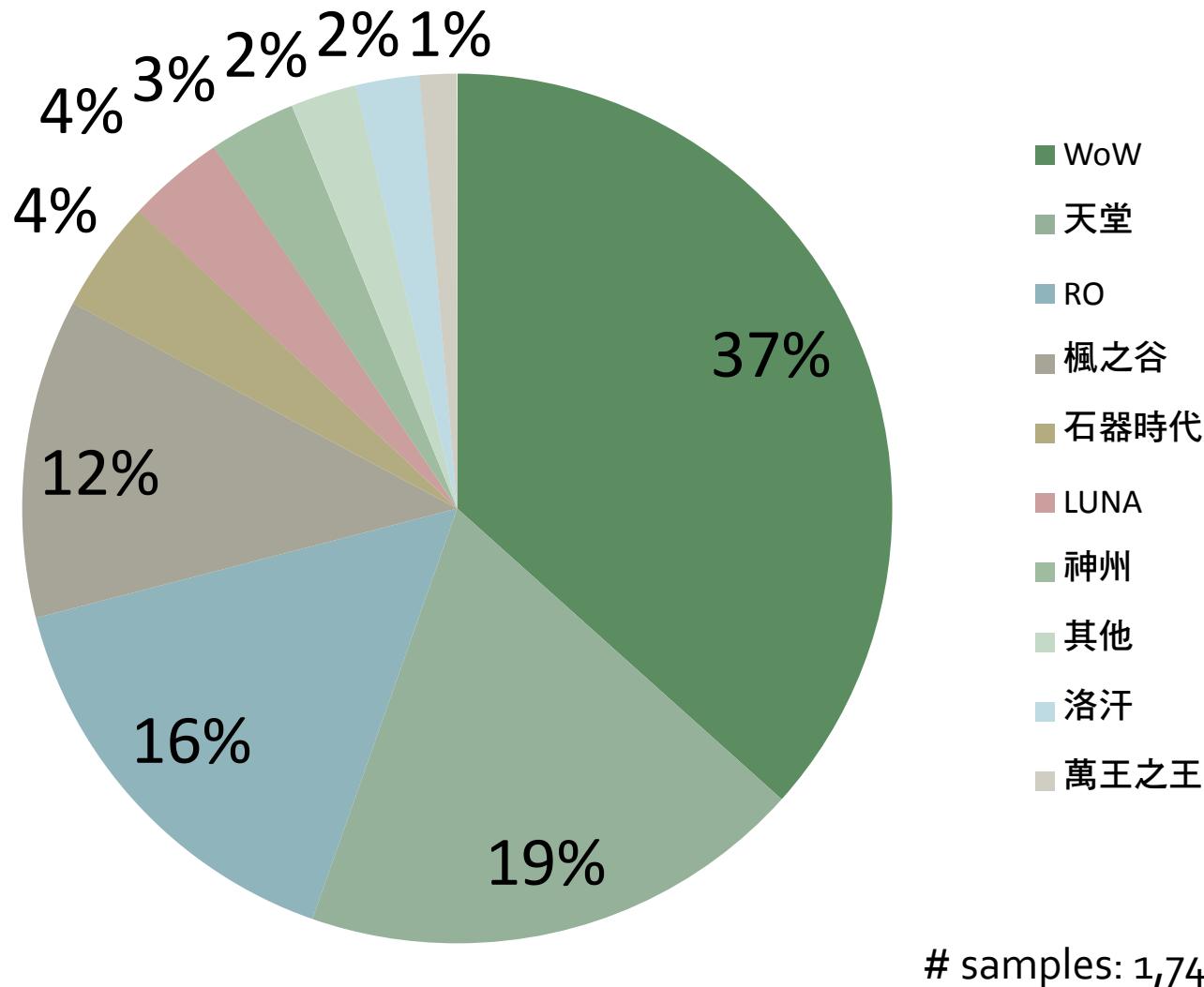
福克斯大神與祂的信眾們 -_-



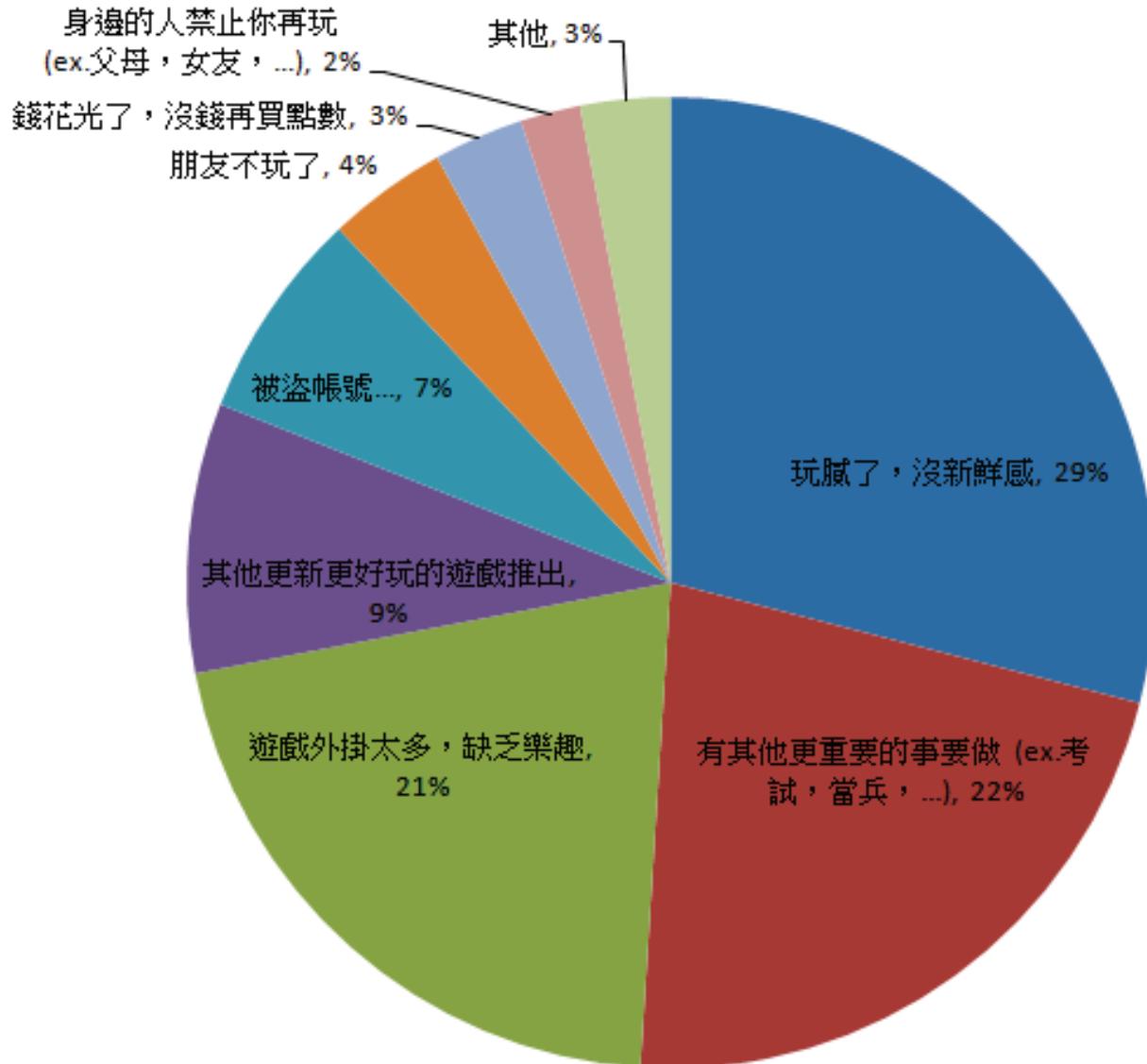




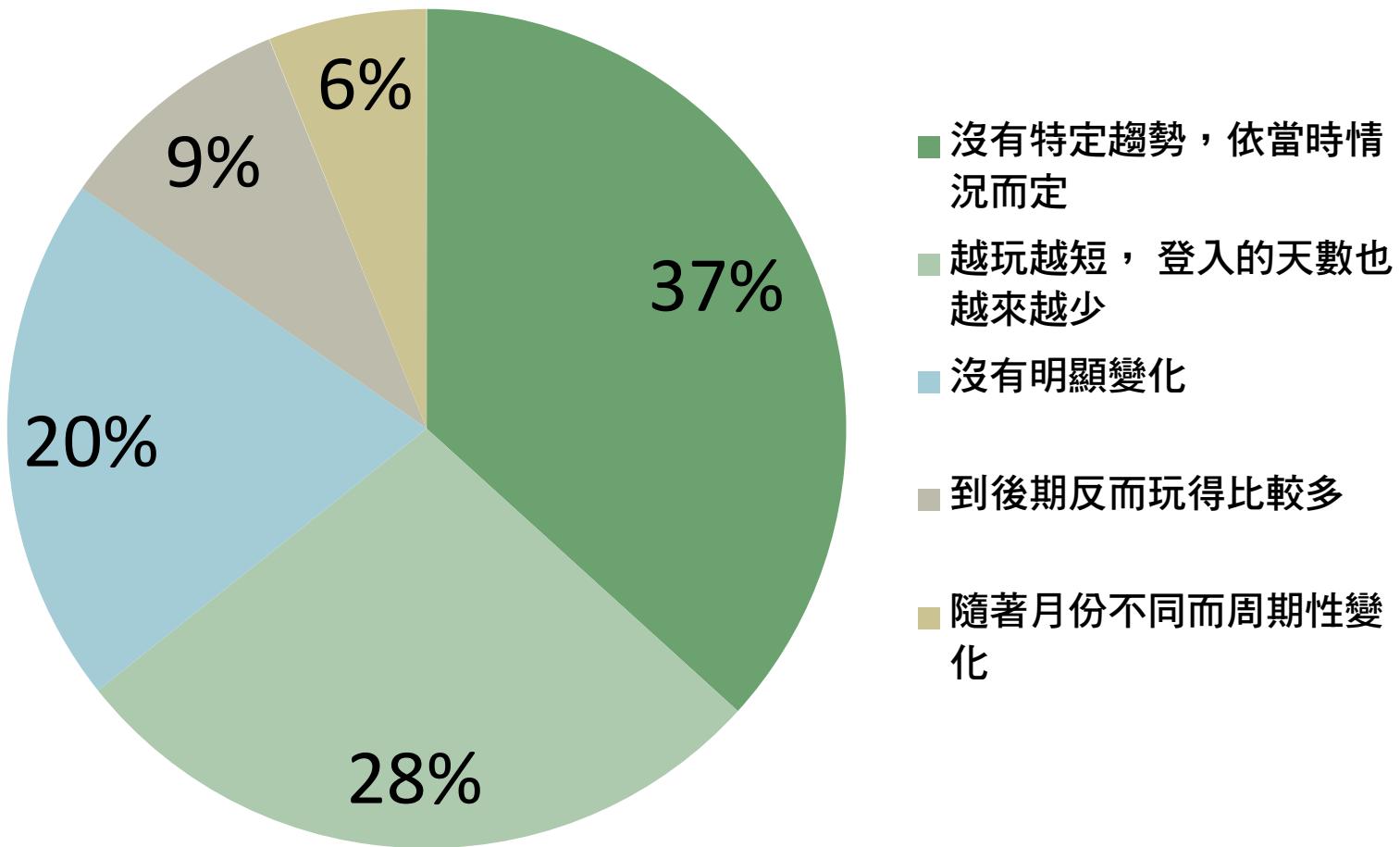
Questionnaire



Reasons for User Unsubscription



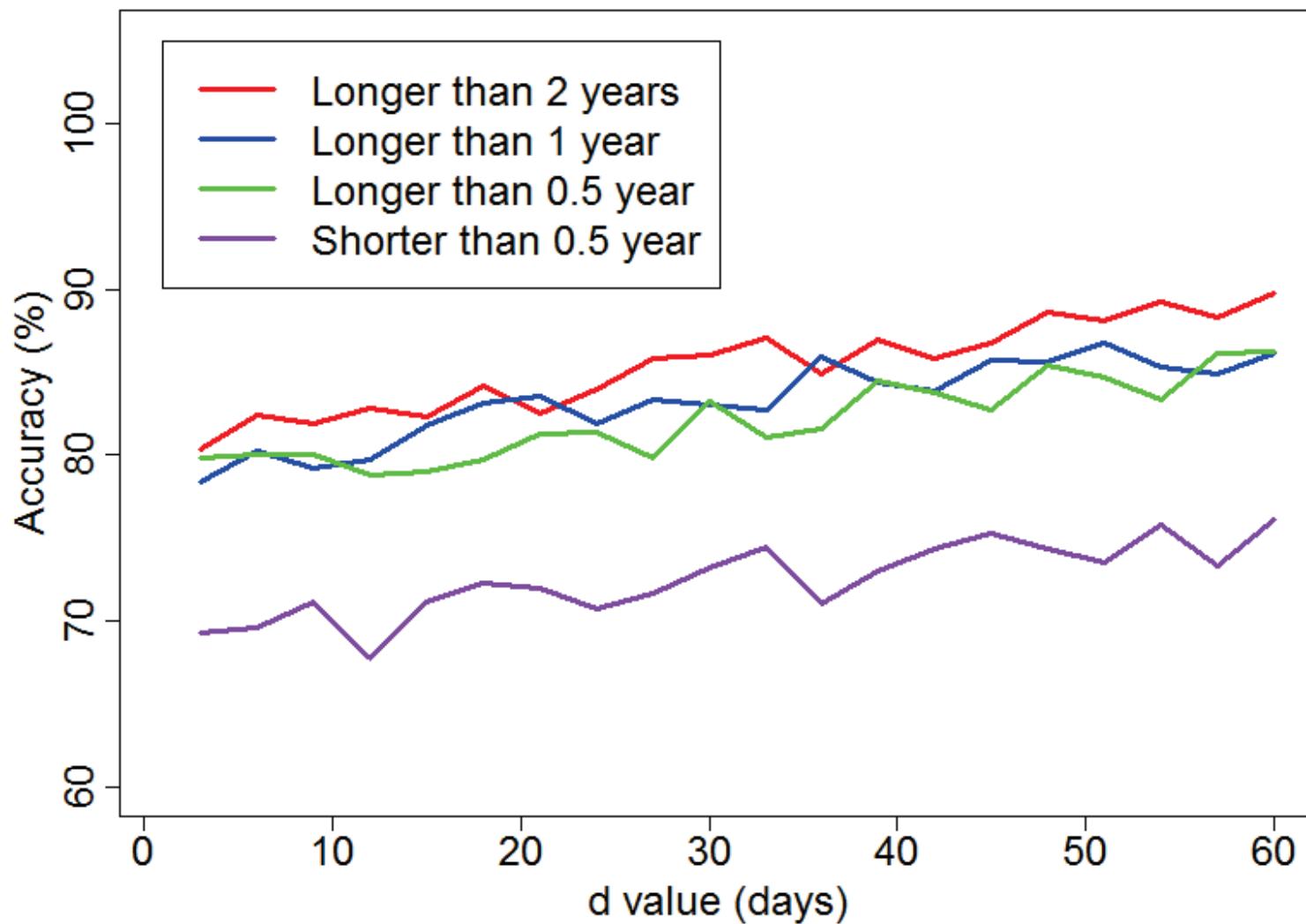
Trend of Game Playing Time



Logistic Regression Model for Unsubscription Prediction

- Significant features (out of > 20 features)
 - Avg. session time
 - Daily session count
 - Variation of the login hour (when the player starts playing a game each day)
 - Variation of daily play time (number of hours)
- A naive logistic regression model achieves approximately 75% prediction accuracy

Unsubscription Prediction Result





交流時間





Forecasting Online Game Addictiveness

Jing-Kae Lou National Taiwan University
Kuan-Ta Chen Academia Sinica
Hwai-Jung Hsu Academia Sinica
Chin-Laung Lei National Taiwan University

World of Warcraft by Blizzard



World of Warcraft by Blizzard

4.5 years and \$63M USD for development before release on 2004*

> \$37M USD for upkeep and expansions during 2004 to 2010**

Thrall yells: The courtyard is ours! Onward to the inner sanctum!

[2. Trade] [Eligoh]: If high tailor Phlay] has invited you to join a group.

Looting changed to Group Loot.

Loot threshold set to Uncommon.

[2. Trade] [Stonedfire]: Stoned goods [Enchanting], [Alchemy] elixir spec, psst me for more

[2. *<http://digitalbattle.com/2006/06/15/world-of-warcraft-cost-63-million/>

**http://online.wsj.com/article/SB10001424052748703467304575383443343071562.html?mod=googlenews_wsj

Grand Theft Auto V (by Rockstar Games)



Grand Theft Auto V (by Rockstar Games)



\$137M USD for development and
100M for marketing

Hit \$1 billion in 3 Days



Witcher 3 (by CDPR)



Witcher 3 (by CDPR)



\$81M USD for development and marketing
3.5 years with 240 staff

Net Profit \$62.5M in 6 weeks

Online Game Industry is Competitive

\$1M to \$200M USD

dev cost per game*

> 200 game titles

each year**

*http://www.gamesetwatch.com/2007/04/mmo_production_costs_how_low_c.php

*<http://www.gamespot.com/news/star-wars-the-old-republic-cost-200-million-to-develop-6348959>

**<http://www.gamespot.com/>

The Terrifying Truth

**Most of them survived
only 4--9 months.**

Usually long before a game's investment could ever be paid off...

The Question

Is a game's lifetime
predictable?

In other words ...

Is a game's addictiveness predictable?

*addictiveness [noun]:
the ability to retain players active in
the game for a long time.*



The Significance

- **STOP** developing hopeless games
- **SUGGEST** better design decisions during development
- **CHOOSE** better games to publish (for game publishers)

State-of-the-Practice

- Intuition of game designers
- Feedbacks from focus groups
- Psychologically inspired methods
 - E.g., the think aloud method

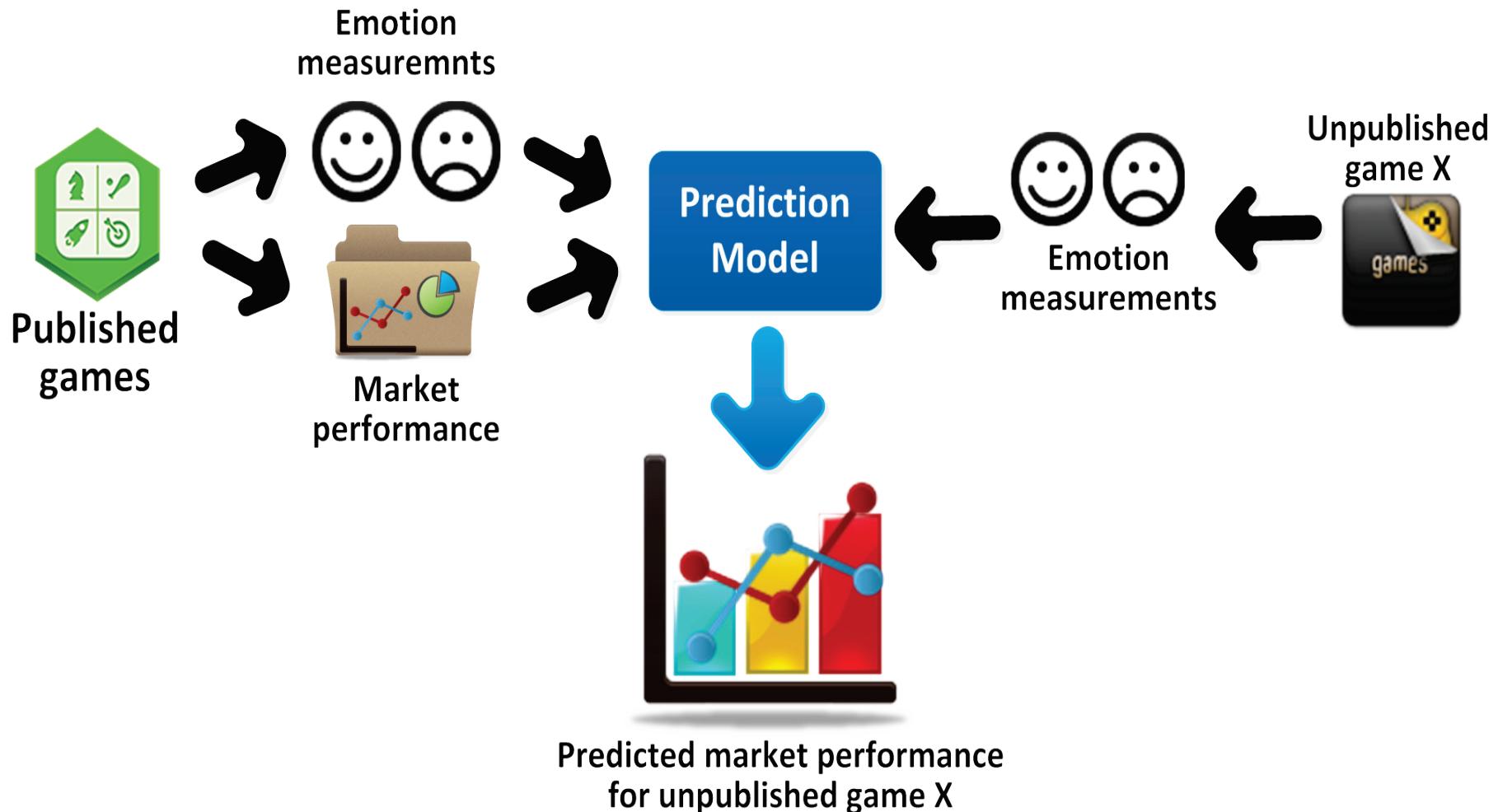
Subjective and thus
tends to be biased

Our rationale

Why a player addicts to an online game?

- *Being entertained*
- *Having various emotions arisen, e.g., joy, excitement, tension*

Our Approach



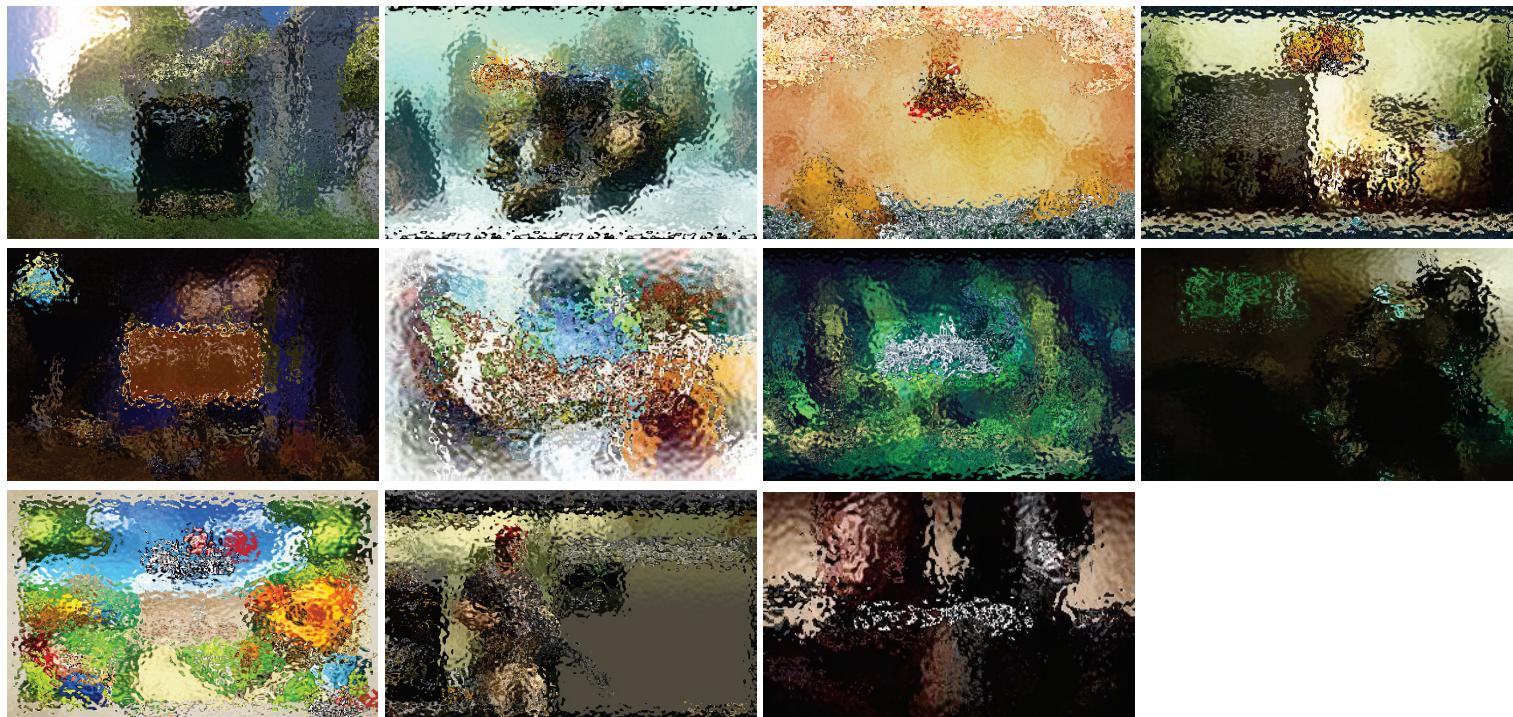
GROUNDTRUTH DATASET DESCRIPTION



Our Collaborator

- Gamania, a top game company in Taiwan
- Gamania released player session information (every player's login and logout events) of 11 games to us

Game + Mania =
gamania



Overview of Games

	Game	Publish Year	Trace Period	# Accounts	User Rating
4 ACT	ACT1	2009	240 days	500K+	8.6
	ACT2	2009	730 days	100K+	8.9
	ACT3	2009	773 days	500K+	8.9
	ACT4	2010	609 days	1,000K+	8.0
2 FPS	FPS1	2009	732 days	1,000K+	8.2
	FPS2	2010	556 days	100K+	7.4
5 RPG	RPG1	2009	385 days	100K+	7.5
	RPG2	2009	323 days	100K+	8.0
	RPG3	2010	486 days	100K+	7.5
	RPG4	2010	732 days	50K+	8.3
	RPG5	2010	820 days	50K+	8.3

Account Activity Records (AAR)

AAR Format

Account	Login Timestamp		Logout Timestamp	
Alex	2009-01-01	11:12:23	2009-01-01	12:40:33
John	2009-01-01	12:31:18	2009-01-01	12:38:47
Serena	2009-01-01	14:05:32	2009-01-02	01:05:25
...
Alex	2009-02-13	11:08:47	2009-02-13	12:23:12
Mary	2009-02-13	12:30:22	2009-02-13	12:45:35

Dataset Overview

# games	11
Total observation days	6,206
Total accounts	8,506,647
Total game sessions	1,311,618,907
Avg. sessions per account	154.2
Avg. sessions per day	211,347

QUANTIFYING GAME ADDICTIVENESS



Attempt #1: Subscription period

INTUITION

A game is more addictive if its gamers tend to play it as much as they can.

- Subscription period
 - The time span (in days) of a player's first and last game sessions.
- Issues
 - The actual time players spent in game is not considered.

Attempt #2: Ratio of Presence

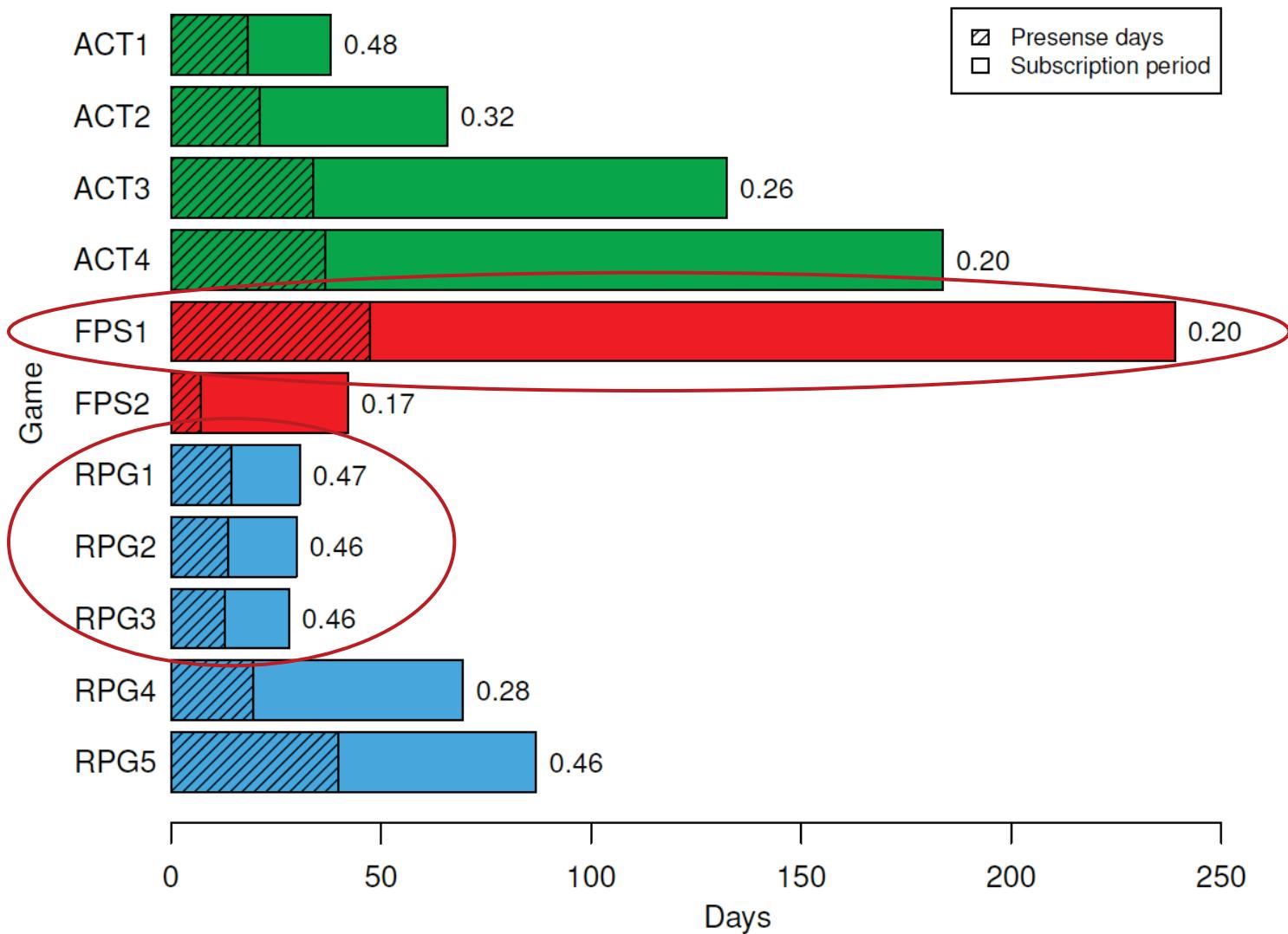
■ Ratio of presence (RoP)

- The total number of days that the gamer entering the game at least once during the subscription period.
- E.g., Entering the game on 20 days with 100 subscription period → $\text{RoP} = 20/100 = 0.2$

■ Issues

- Bias toward games with short subscription periods
- E.g., average 4 online days over 5 subscribed days = RoP 0.8

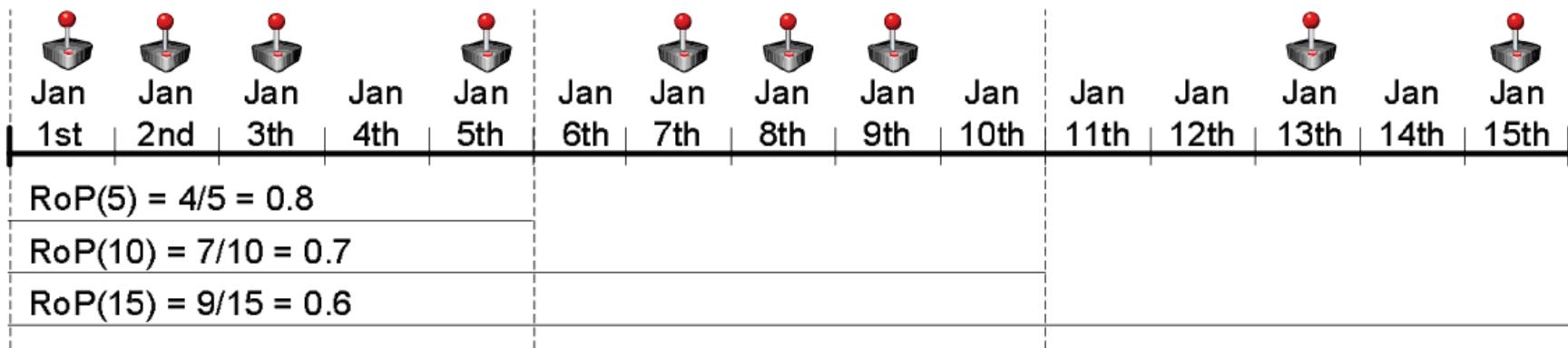
Subscription period and RoP



RoP Generalization

■ RoP(OP)

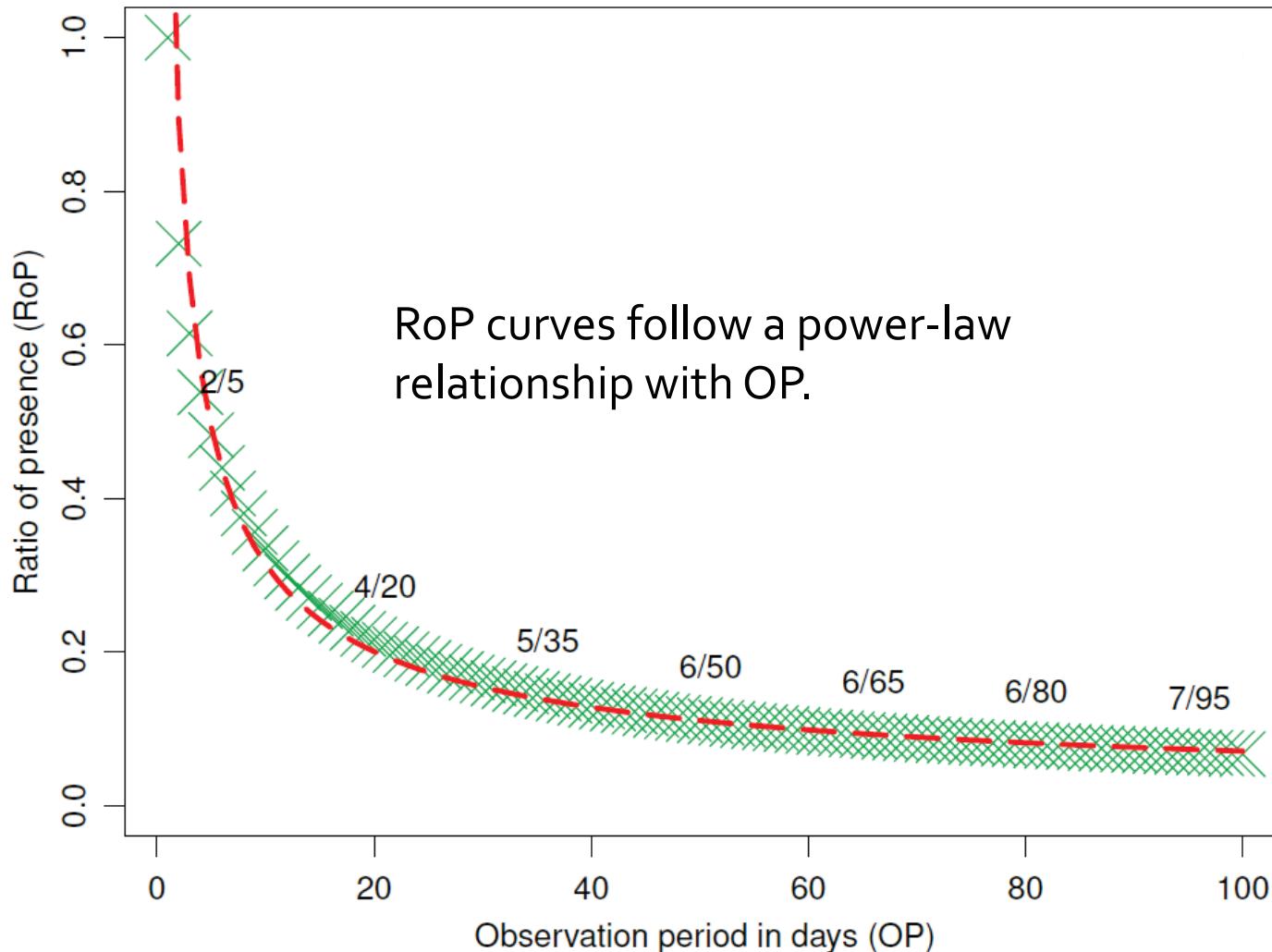
- RoP with a certain observation period

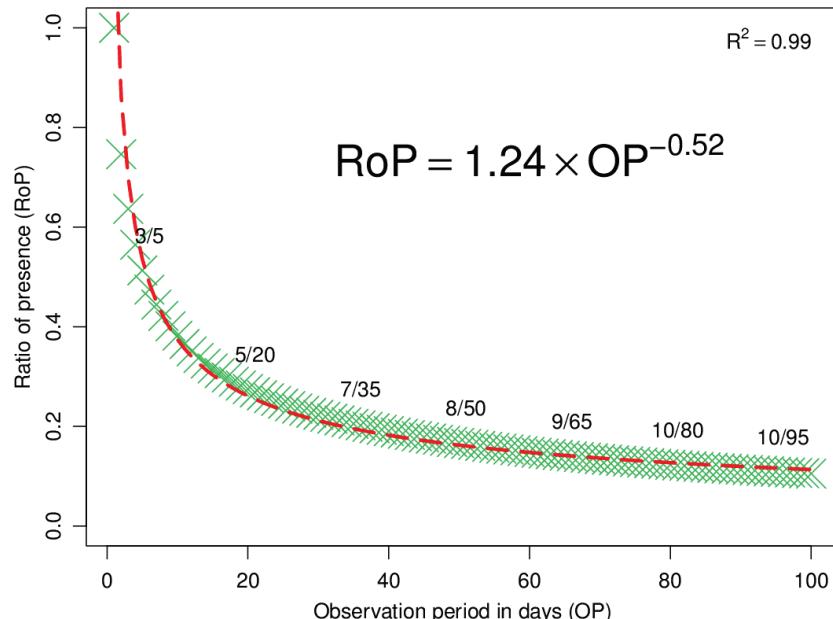
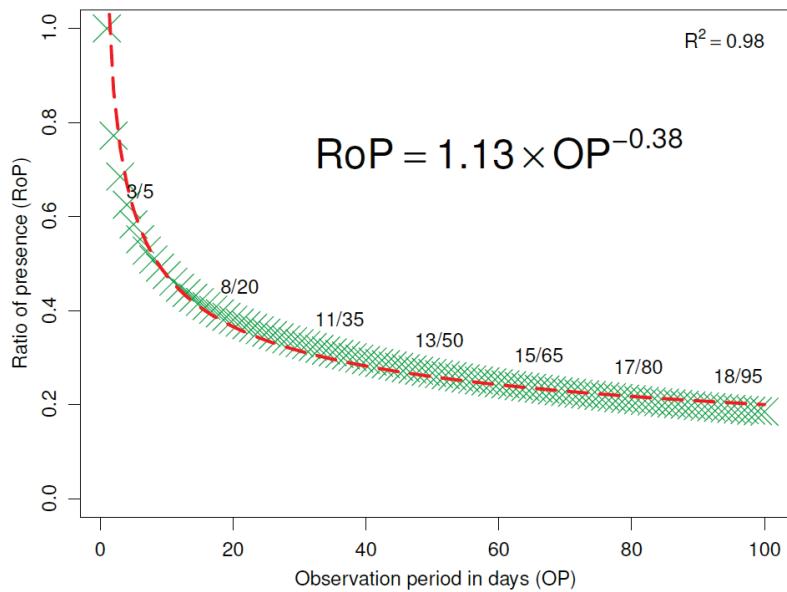
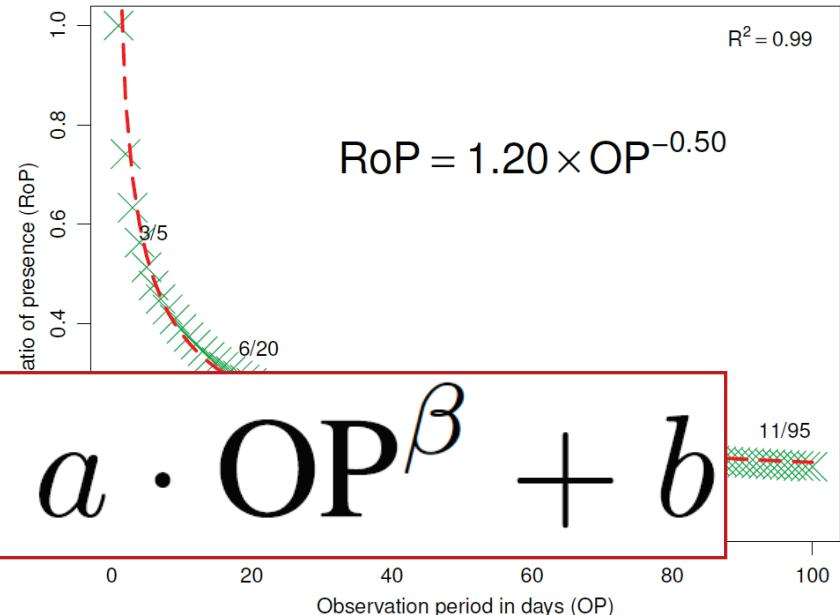
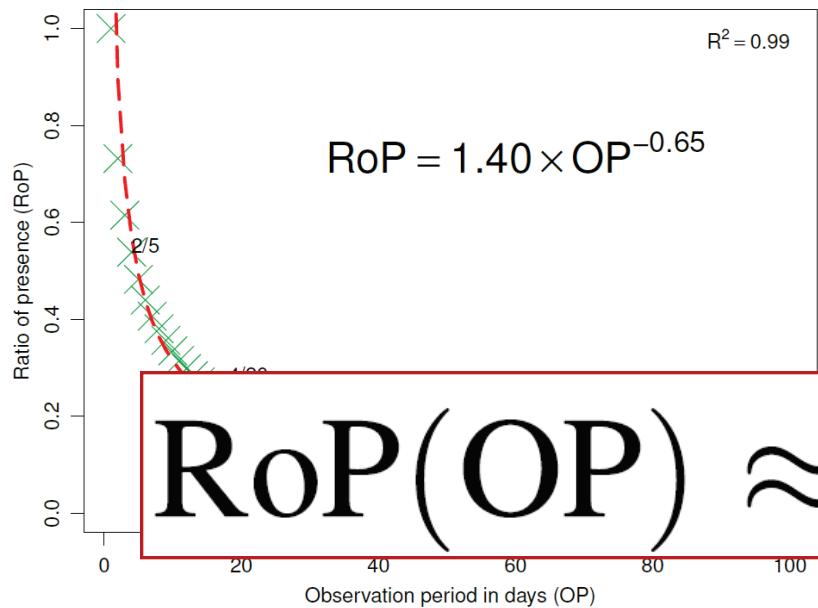


■ RoP curve

- The curve formed by RoPs over a range of OP

The RoP curve of FPS2

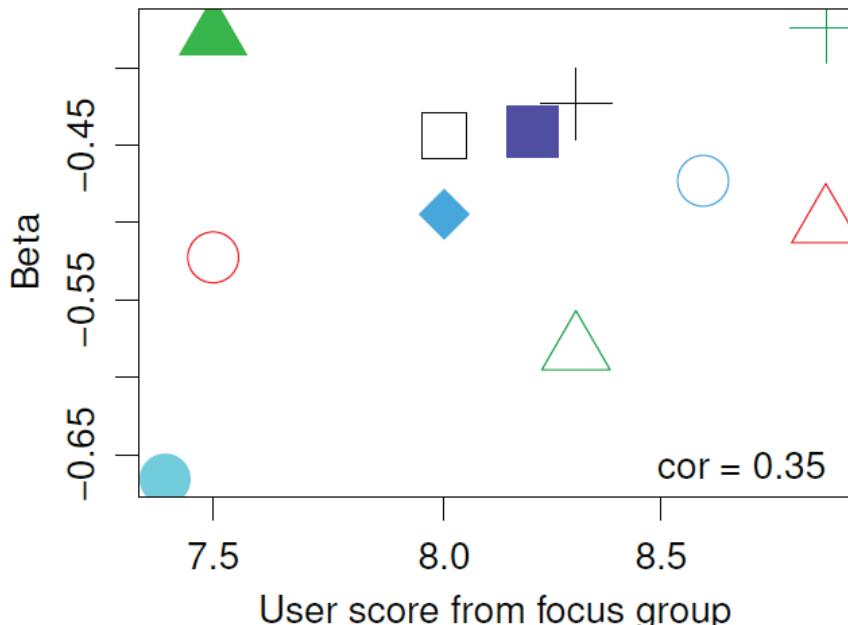
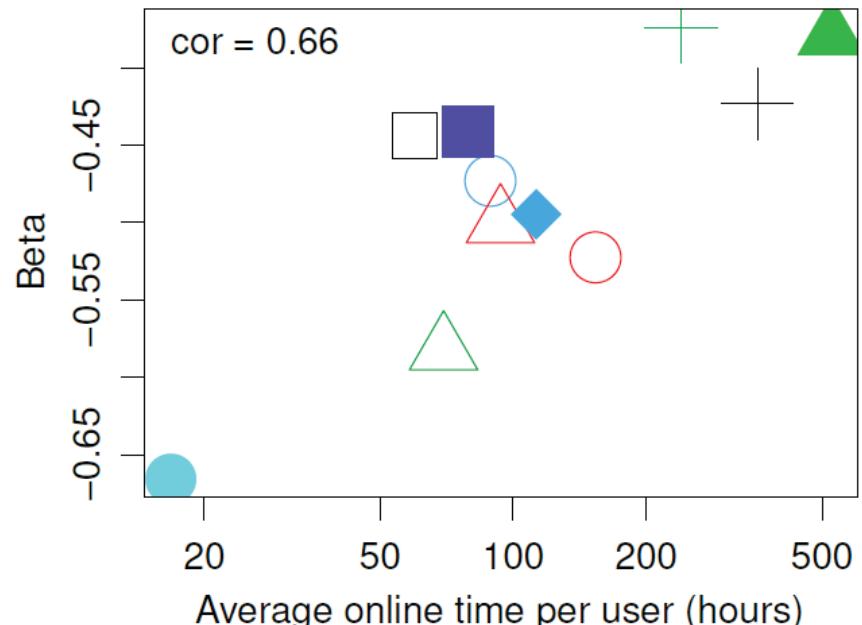
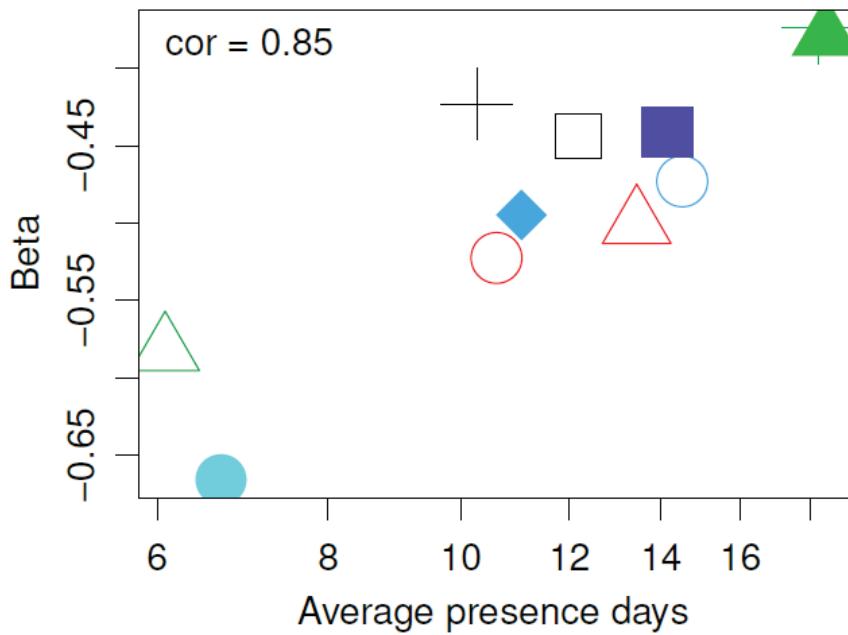
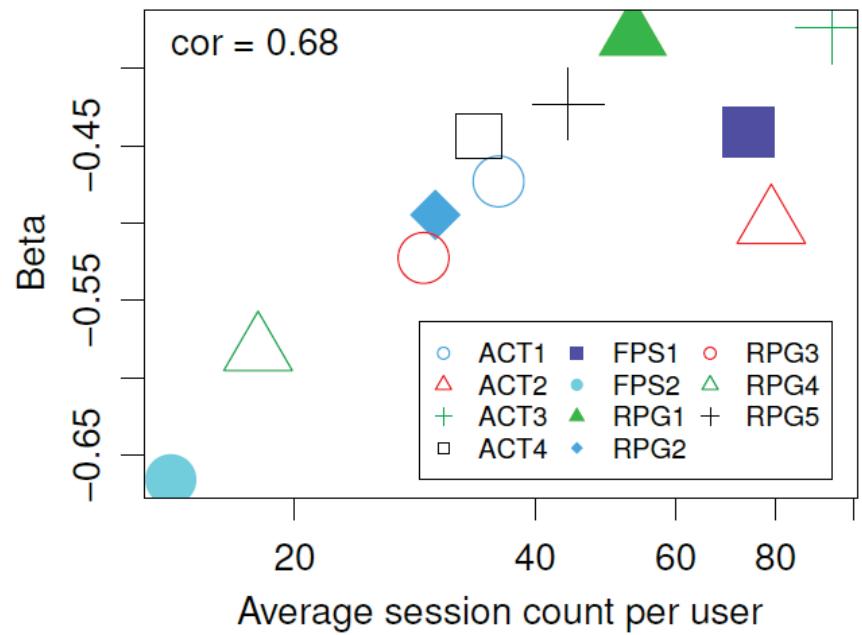


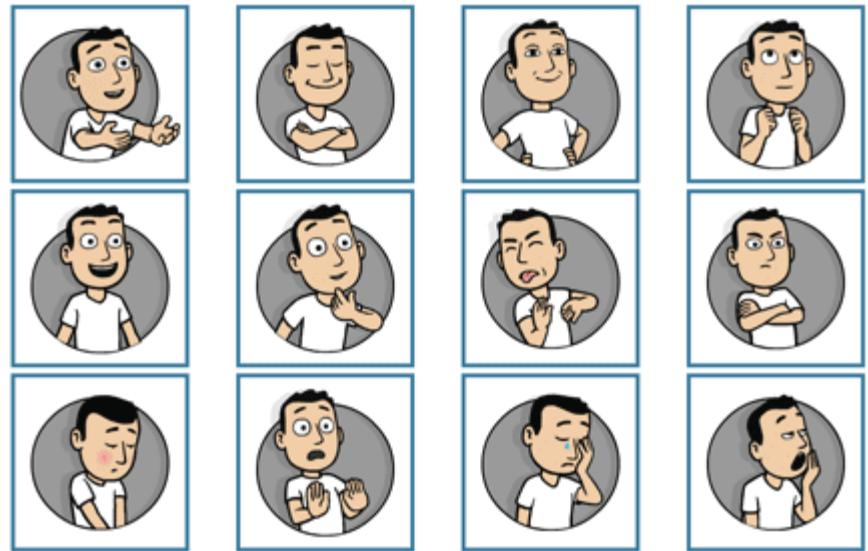


Defining Addictiveness Index

- $\beta \leftarrow$ $\text{RoP}(\text{OP}) \approx a \cdot \text{OP}^\beta + b$
- The **decline rate** of RoP over time
- genre-independent

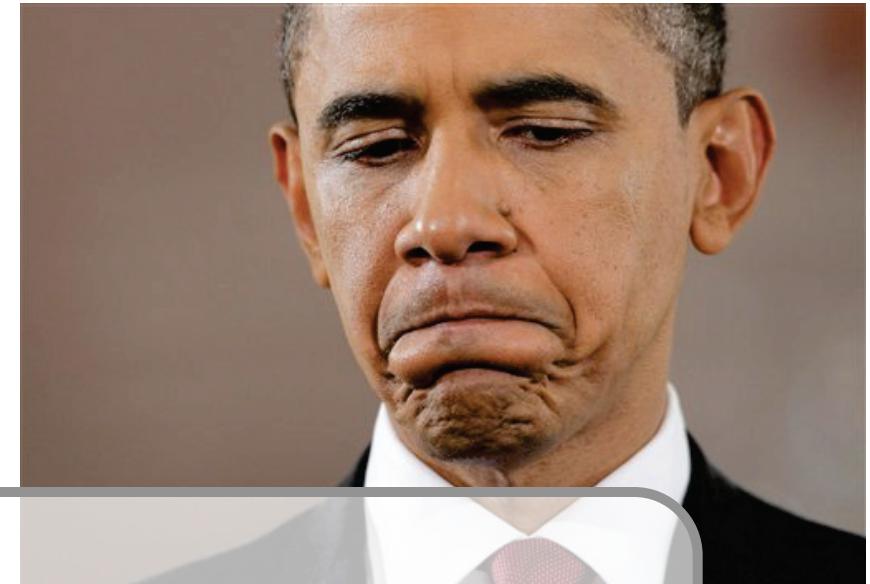
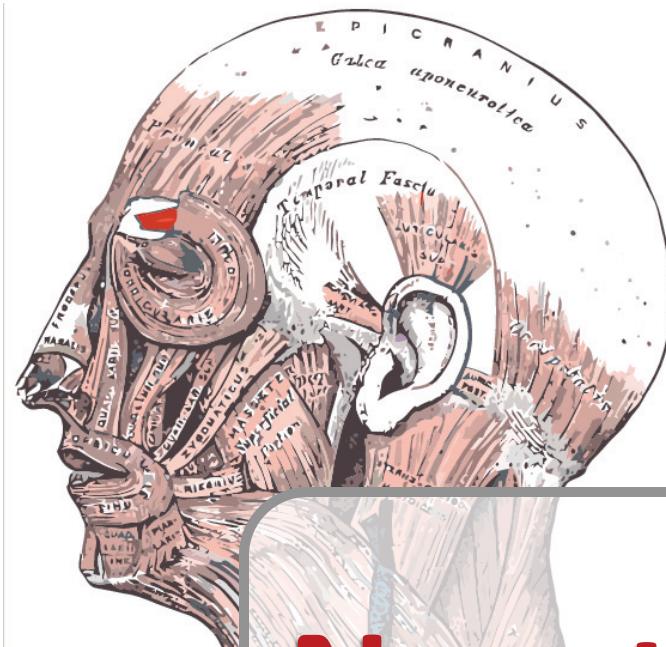
Game	β	R^2	Game	β	R^2
ACT1	-0.44	1.00	RPG1	-0.38	0.98
ACT2	-0.49	1.00	RPG2	-0.50	0.98
ACT3	-0.37	1.00	RPG3	-0.50	0.99
ACT4	-0.43	1.00	RPG4	-0.58	1.00
FPS1	-0.45	0.99	RPG5	-0.43	0.99
FPS2	-0.65	0.99			





MEASURING PLAYER EMOTION

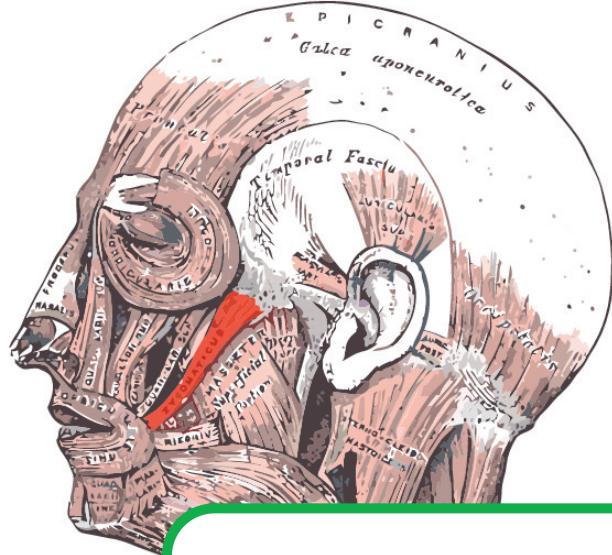
Corugattor supercilli muscle groups



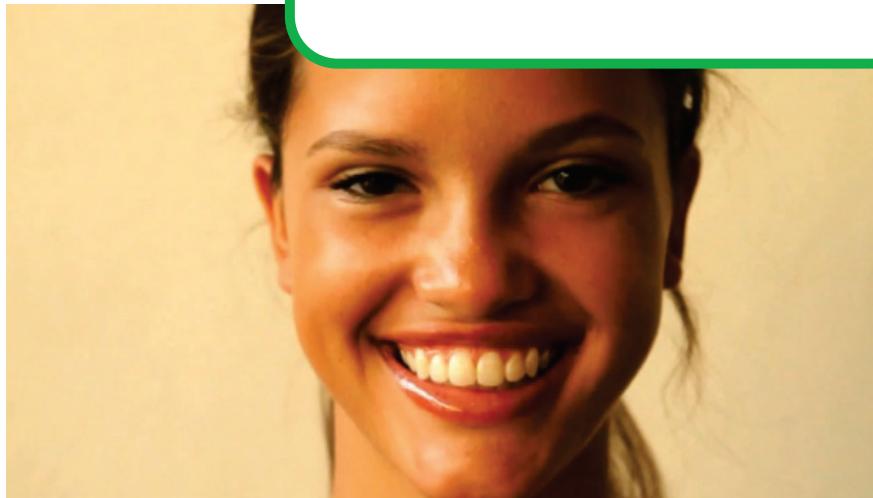
Negative Emotion



Zygomaticus major muscle groups



Positive Emotion



Facial EMG approach

(EMG: Electromyography)

1. Continuous emotion measures (can be at a rate of 1000 Hz or even higher)
2. Does not disturb game play
3. Objective since the emotional indicators are directly measured rather than told by subjects

Facial EMG Measurement Setup



Corrugator Supercilii
muscle
Negative emotions

Zygomaticus Major
muscle
Positive emotions

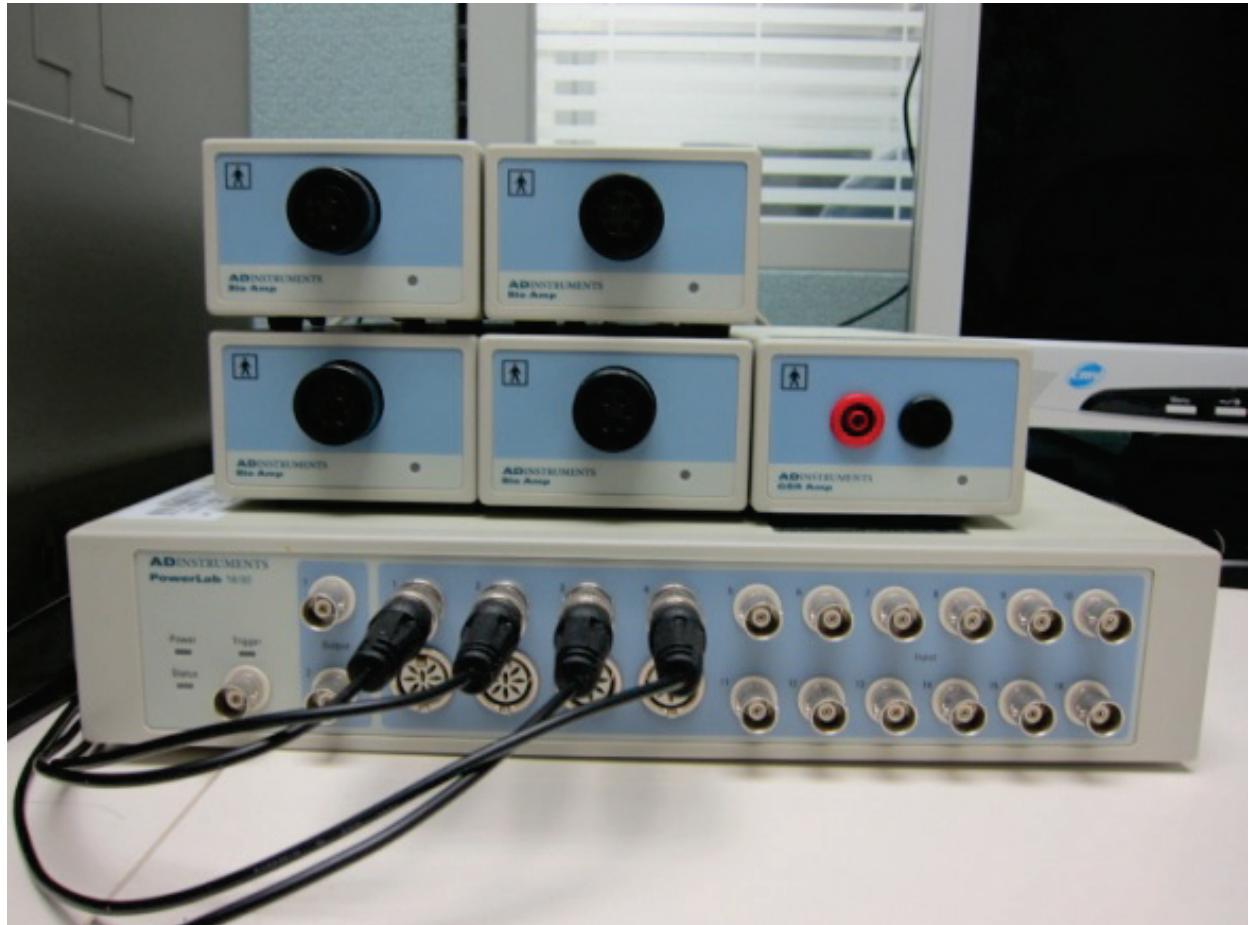
Measurement Devices



Electrodes



Wires



PowerLab 16/30

Measuring Facial EMG during game play



Experiment Design

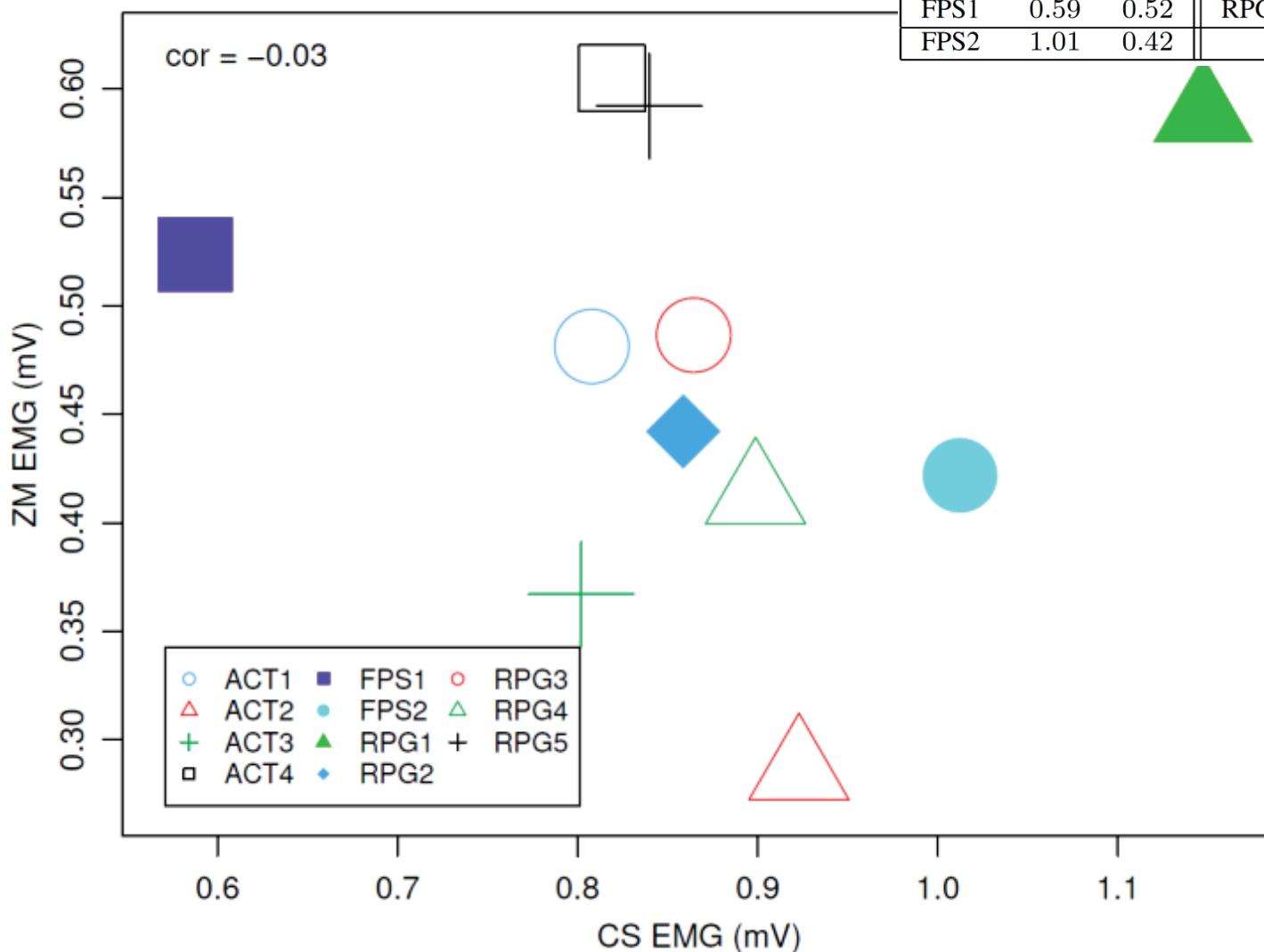
- 84 subjects are asked to play the 11 games
- A subject must be new to the games he played
- Each game session lasts ≥ 45 minutes continuously

# subjects	84
Males	74
Females	10
Ages of the subjects	19–34
# total traces	192
# traces per subject	1–3
# traces per game	15–19
Total hours of traces	155

Quantifying the Measurement

- EMG samples are taken at 1,000 Hz, so a 45-minute trace comprises
 $45 \times 60 \times 1,000 = 2,700,000$ samples
- The average absolute differences between adjacent samples is taken as the representative index
 - Given a time series of electrical potential samples $P = \{p_1, p_2, \dots, p_n\}$
 $f(P) = \text{mean}(\text{abs}(p_2 - p_1), \text{abs}(p_3 - p_2), \dots, \text{abs}(p_n - p_{n-1}))$
- CS: corugattor supercilii muscles → negative emotion
ZM: zygomaticus major muscles → positive emotion

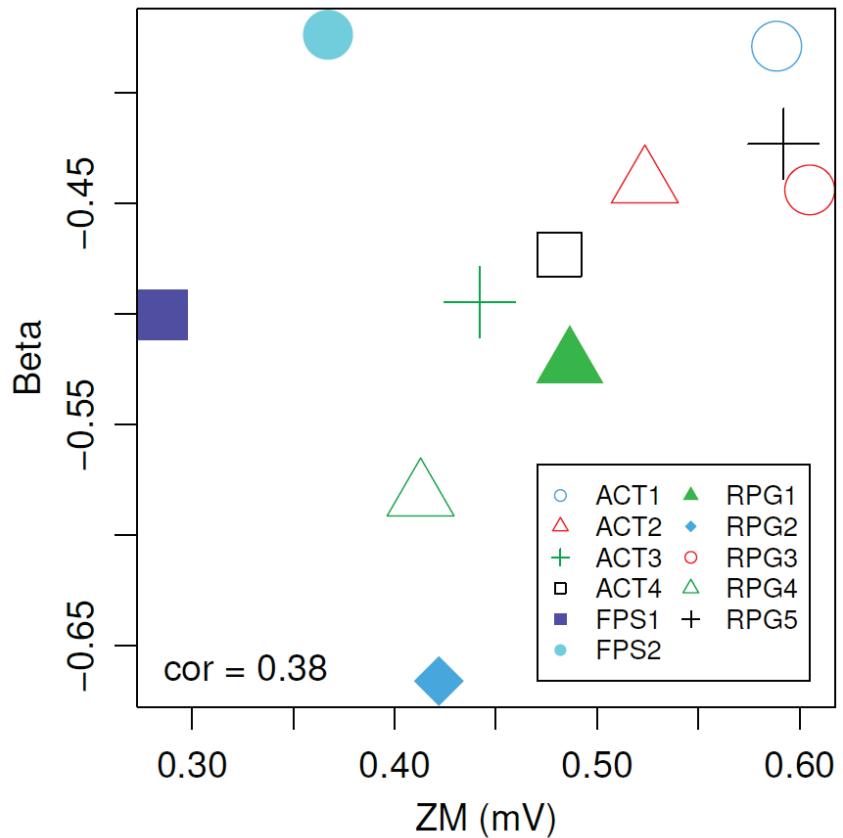
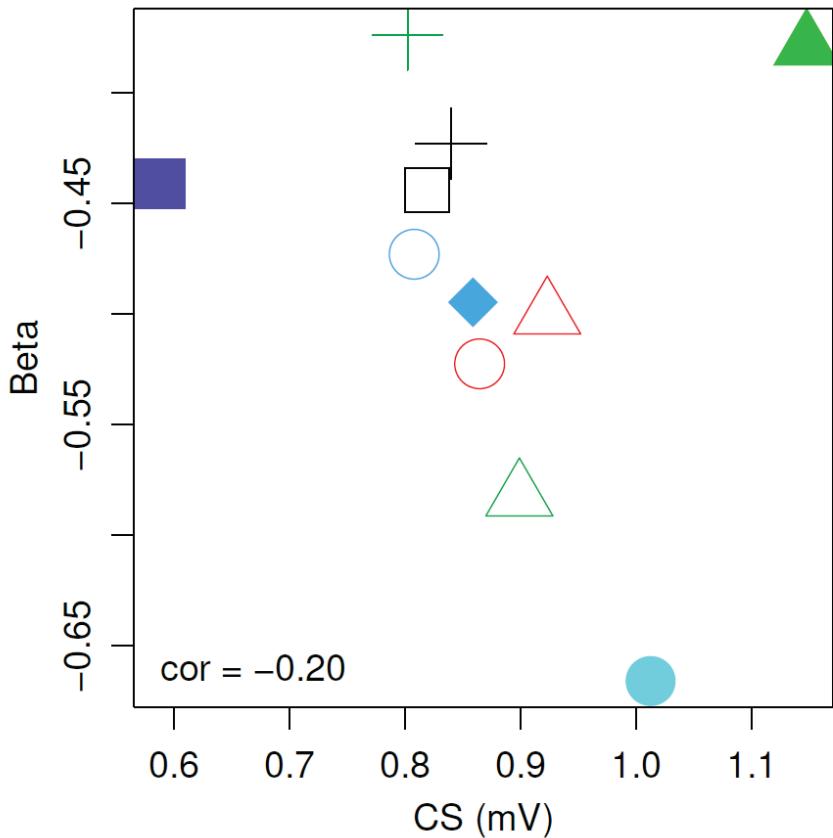
Game	CS	ZM	Game	CS	ZM
ACT1	0.80	0.48	RPG1	1.15	0.59
ACT2	0.92	0.28	RPG2	0.86	0.44
ACT3	0.80	0.36	RPG3	0.86	0.49
ACT4	0.82	0.60	RPG4	0.90	0.41
FPS1	0.59	0.52	RPG5	0.83	0.59
FPS2	1.01	0.42			





FORECASTING GAME ADDICTIVENESS

Emotion vs. Addictiveness



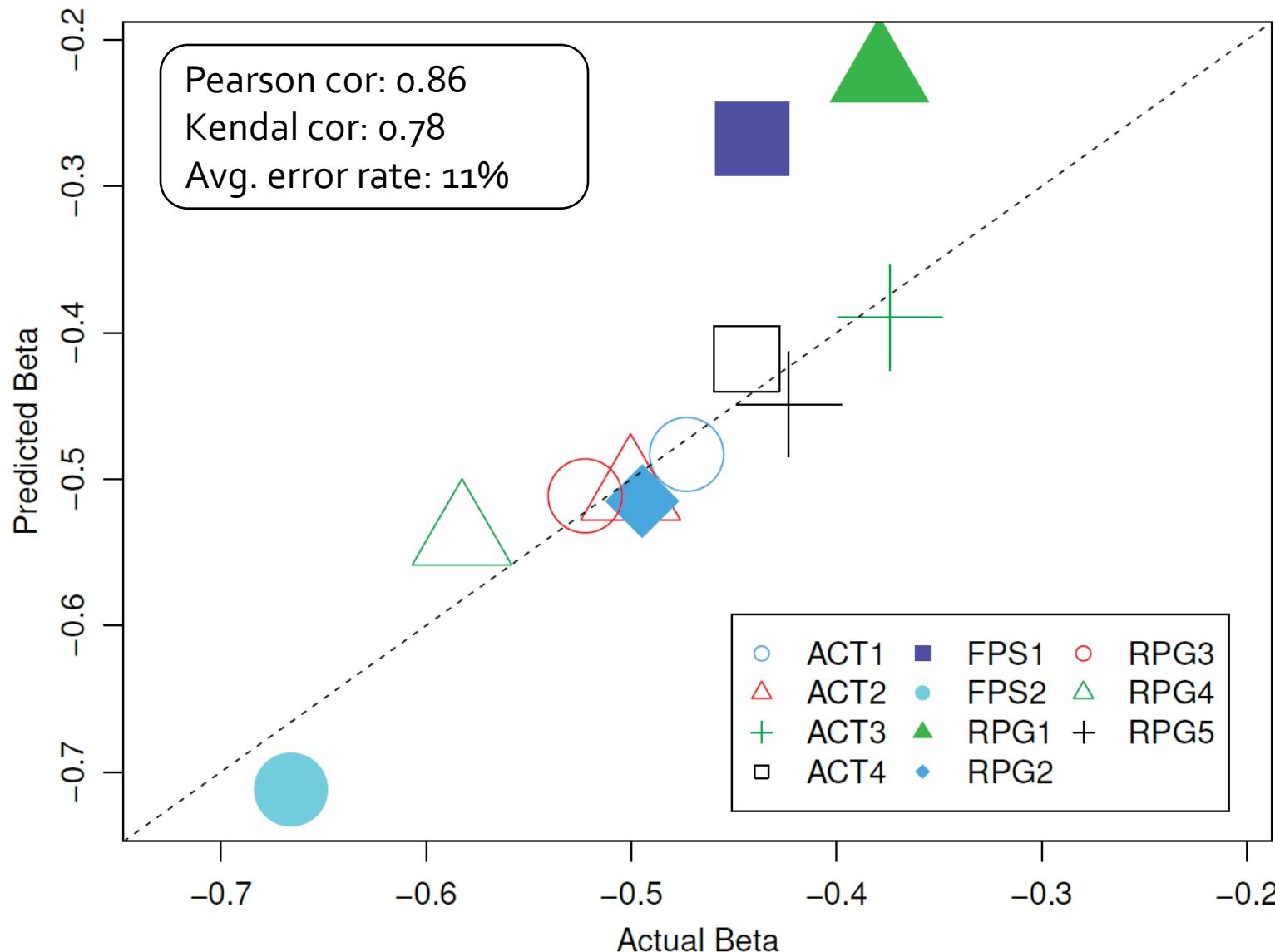
Modeling Game Addictiveness

- ES: the emotional strength
 - $ES = CS + ZM$
 - The combined emotional strength arisen
- $\beta = \omega_0 + \omega_1 \cdot CS + \omega_2 \cdot ZM + \omega_3 \cdot CS:ZM + \omega_4 \cdot CS:ES + \omega_5 \cdot ZM:ES$

Adj. $R^2 = 0.94$

Variable	Coef.	Std. Err	t	Pr > t
(constant)	4.30	0.43	9.87	0.00018
CS	-4.19	0.51	-8.18	0.00044
ZM	-11.59	1.04	-11.07	0.00010
CS:ZM	3.53	1.06	3.31	0.02119
CS:ES	-0.21	0.24	-0.87	0.42263
ZM:ES	4.89	0.66	7.37	0.00072

Leave-One-Out Validation



Applications of the model

- Early evaluation of game design
- Market value assessment before publishing

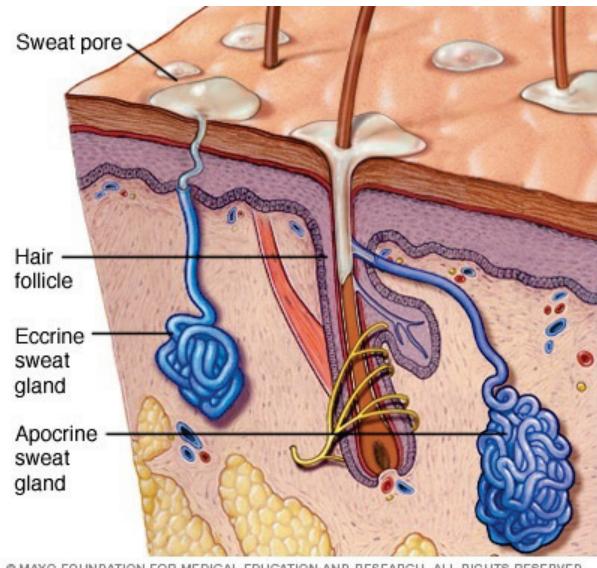
- 1. Optimize the odds of successful investments**
- 2. Target more accurately the provision of better entertaining experience.**

Ongoing Work & Future Plan

- More sophisticated modelings and more validations
 - Game addictiveness may change over a game's lifetime
- Develop models that can explain WHY a game's lifetime is longer than another?
 - Due to particular game designs?
 - Due to commercial promotions or others?

It Is Just The Beginning

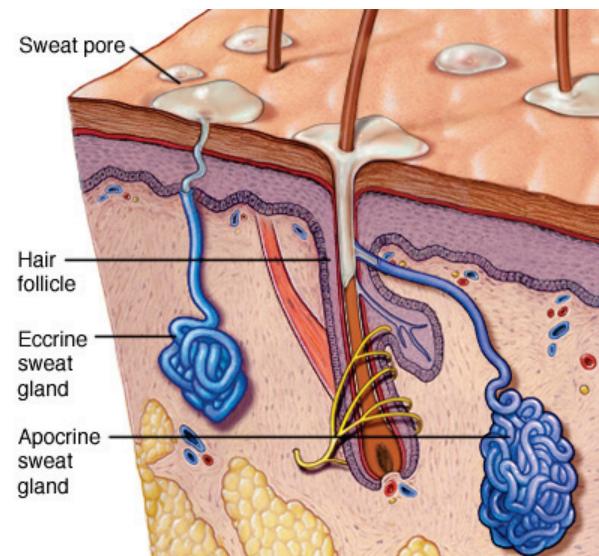
- We are now digging into psychophysiology
- Exploring various possibility to read one's emotion
 - Brain activity
 - Eye movement
 - Heart activity
 - Respiration
 - Sweat secretion
 - And so on
- Also the mechanisms related to fun and addiction
 - Reward process
 - ...



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Sweat Secretion

- Apocrine
 - Hormonal change
 - Active for stress and sexual excitement
- Eccrine
 - Themoregulation
 - Excretion
 - Protection
 - Reflection of emotion change
 - Palms and soles



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

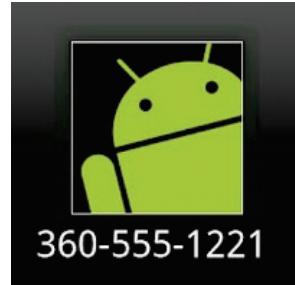


交流時間





Tring Tring
Who's calling?



未知號碼來電怎麼辦？

陳昇瑋

中央研究院資訊科學研究所

An everyday annoyance...



Who's calling?

Answer

Ignore

Available Solutions



WhoCallsMe?

Phone Number Search

Get Details



Technologies Adopted

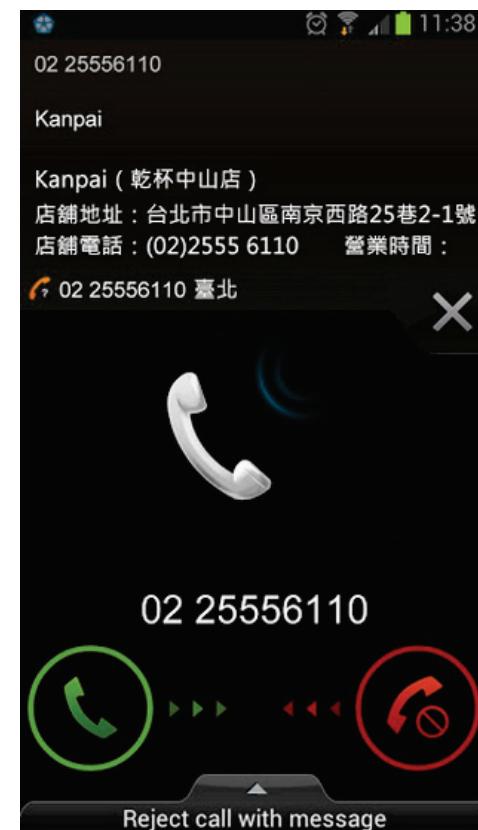
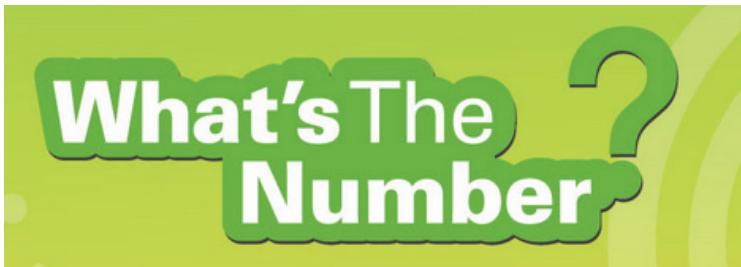
- Yellow pages
 - HiPage, YP.com
 - Yelp, Google Places
 - 104.com.tw, 好評網
- Users' address books
- Google search (!)



Search Phone Numbers on Google

- 02-2311-3731
- 02-27883799
- 0933-555770
- 0277064034
- 0987772305
- ...

Available Solutions in Taiwan





■ Real-time caller ID identification

- based on Google search and **user reports / tags**



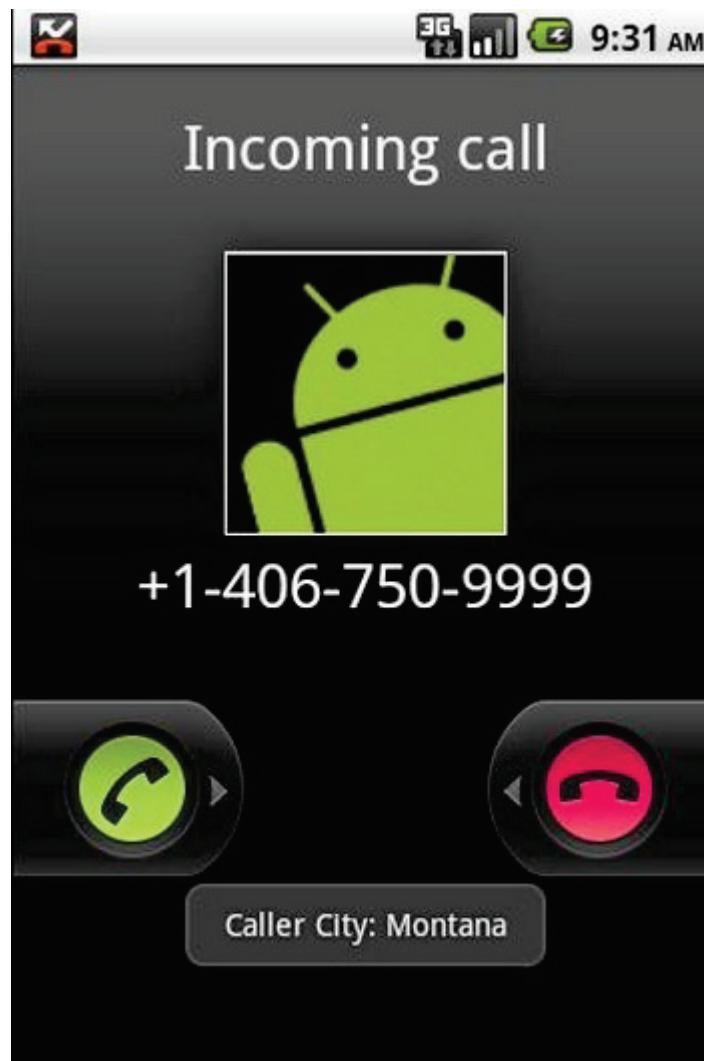
THUS I BECAME A USER OF



TWO YEARS AGO...

**BUT, ... GOOGLE SEARCH AND USER
TAGGING ARE NOT ENOUGH**

Frequently, I still see the following screen:



e.g., 0910889139

**THUS, I WROTE AN EMAIL TO
WHOSCALL CUSTOMER SERVICE**



Whoscall 很實用，但是若 Google search 找不到
也沒有人回報的未知號碼你們就沒戲唱了對吧 :P

沒錯 XD



那我來幫忙做這個功能好了... :D

好啊 :)





A JOINT RESEARCH PROJECT

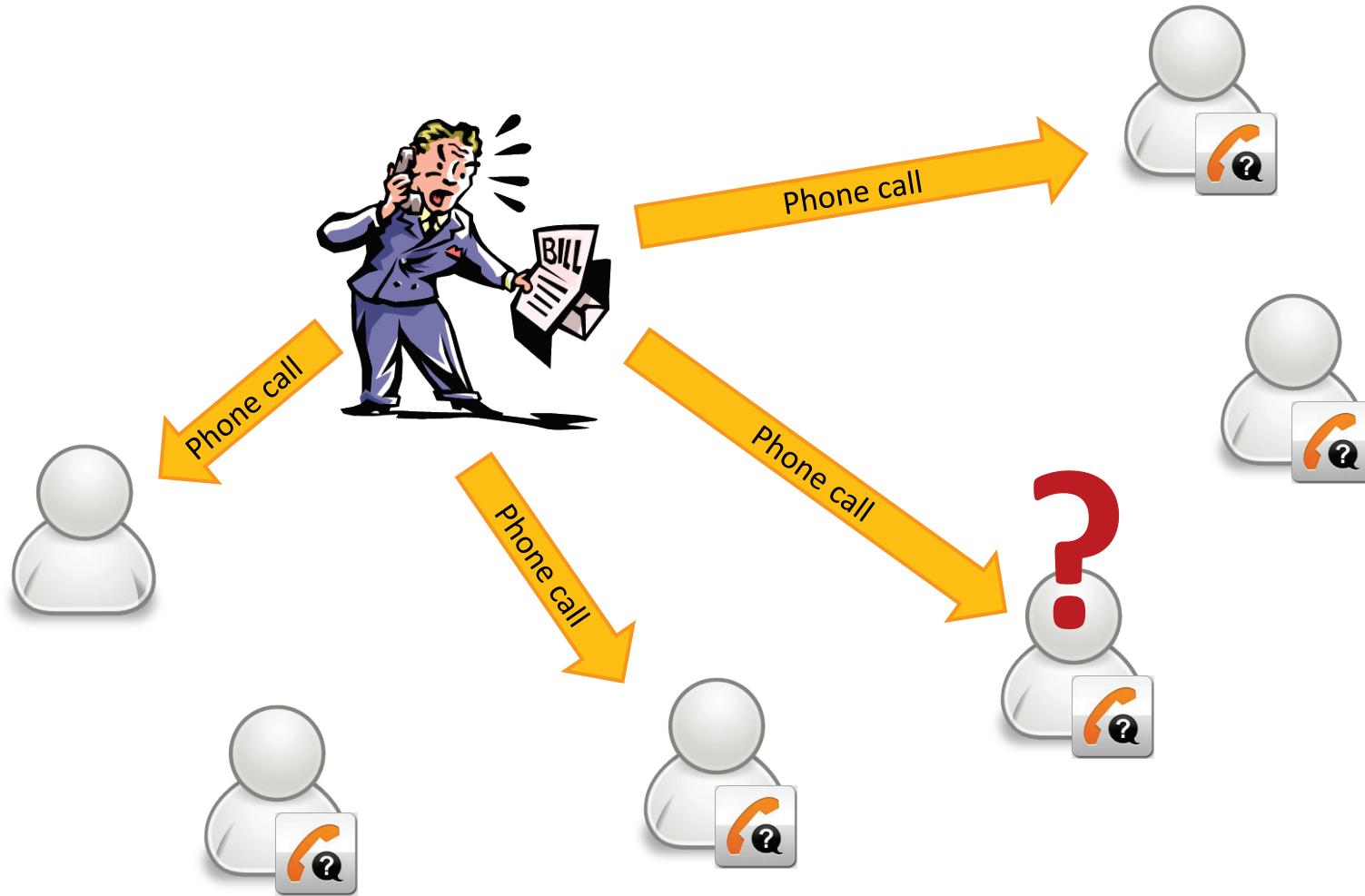
The research problem

- For a unknown phone number
 - No google results (or no useful information)
 - No user tags / reports
 - Not a Whoscall user
- Can we determine if it's a malicious number?
 - 推銷電話?
 - 詐騙電話?
 - 色情電話?
 - ~~打錯電話?~~

Rationale

- We believe it's possible to identify a malicious number because of ...
- Whoscall userbase (= **potential sensors**)
 - 4 million installations
 - 1 million active users (daily)
 - 10 million phone calls (daily)
- So, when a phone number reaches a Whoscall user, we could possibly determine whether the number is malicious or not based on its **previous call behavior**.

The Scenario



Our Steps

- Recruit a group of voluntary Whoscall users as our sensors
- Collect phone call logs from these sensors for a month
- Compare these phone call logs with user reports
(封鎖記錄)
- Use machine learning techniques to build a predictor for unknown phone numbers

Privacy Concerns

- **User privacy** is kept the highest priority
- Phone numbers are stored as MD5 hash codes
(therefore unable to be reversed)

User reports

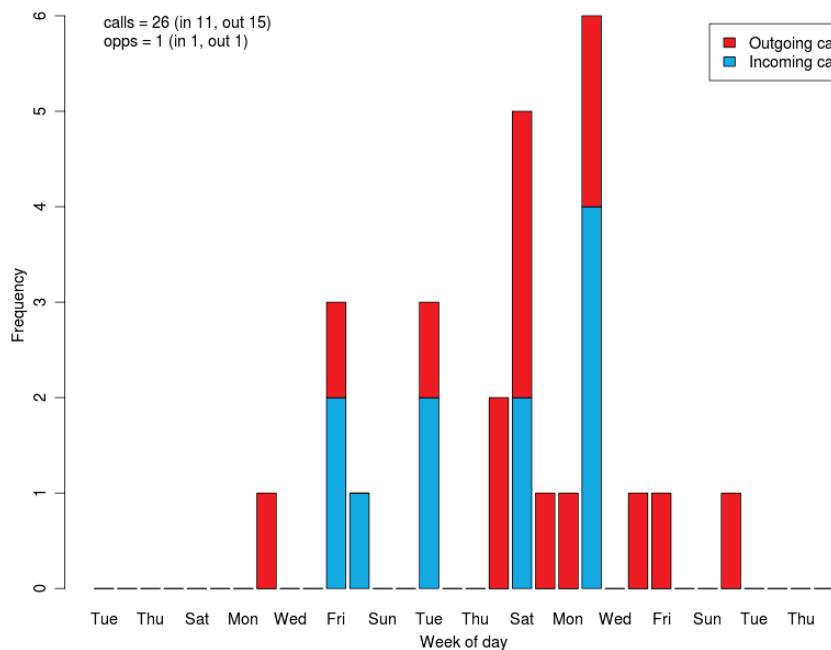
一接就掛斷	嚴重騷擾	色情交友	摩門
一打來就掛掉	國外莫名來電	色情交友電話	撥了馬上掛掉
一接對方馬上掛斷	國際電話偽裝台北區碼???	色情人肉市場	擾亂電話
一接就掛	地下期貨公司	色情仲介	擾人電話
一接就掛掉	地下錢莊	色情傳播	收數
一接就掛斷	地下錢莊推銷	色情垃圾簡訊	收視率調查
一接就掛斷的吵人電話	地下非法期公司	色情外送	放款簡訊
一接就掛電話	地下非法期貨公司	色情妹妹電話	放款電話
一接聽就掛掉	地產	色情媒介	政府宣導
一接起來就掛斷電話	垃圾	色情宣傳	政府立案單身
一接起來，就說打錯	垃圾件	色情干擾	敲一聲而已
一直傳廣告簡訊	垃圾廣告	色情廣告	整人電話
一直打錯	垃圾簡訊	色情廣告擾人	新光保全
一直打錯電話	垃圾訊息	色情廣告簡訊	日制
一直收到沒顯示的APP	垃圾電話	色情拉客妹	日產フィナンシャル
一直狂打錯電話	城市理財	色情按摩	日豐車行Sales
一聲	基隆美髮	色情推銷	星展
一聲不響，就掛掉，有問題	填問卷	色情推銷廣告	星展借貸
一聲就掛	壽險	色情推銷簡訊	星展推消
一聲掛斷	外勞	色情推銷電話	星展銀行
一聽收線	外崎砂斗美	色情援交外送	星展銀行推廣
一響即掛	多次接聽冇人回應，數秒後	色情敗類	星展銀行貸款
一響就掛	夜半打給不認識的在那亂	色情服務	淫媒仲介
		色情業廣告	

Data Summary

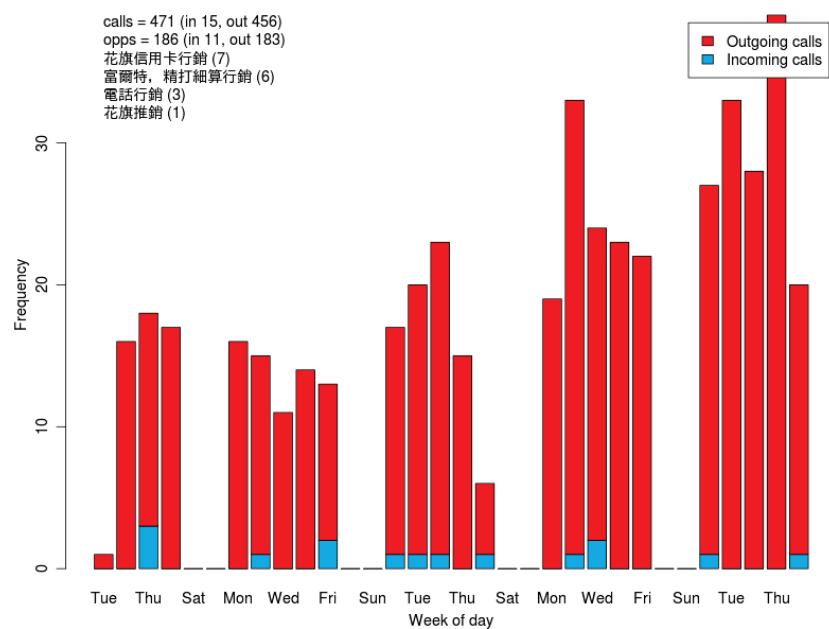
# total numbers	9,017
# normal numbers	5,000
# marketing numbers	3,207 推銷電話
# fraud numbers	810 詐騙電話
# reports for marketing numbers	19,833 (mean 9.8 per number)
# reports for fraud numbers	4,100 (mean 8.4 per number)
# calls	83,174 (mean 92.2 med 8.0)
# missed calls	35,151 (0.42)
Avg calls per opposite party	1.6
Call duration	mean 3.5 med 0.9 minutes

Call Pattern Observation

Normal

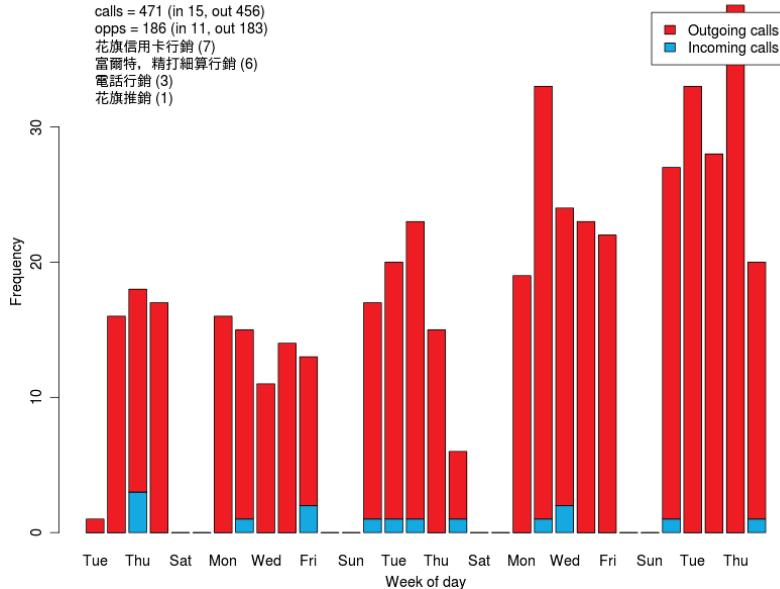


Spam

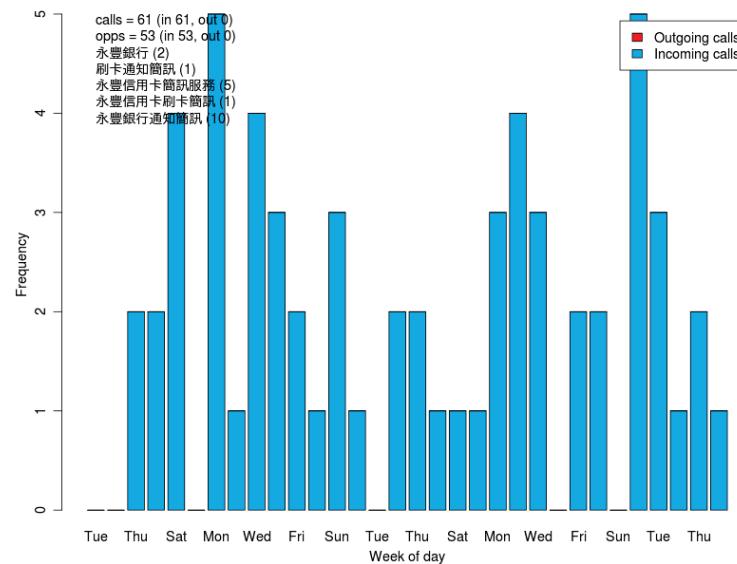


Two Modes of Spam Calls

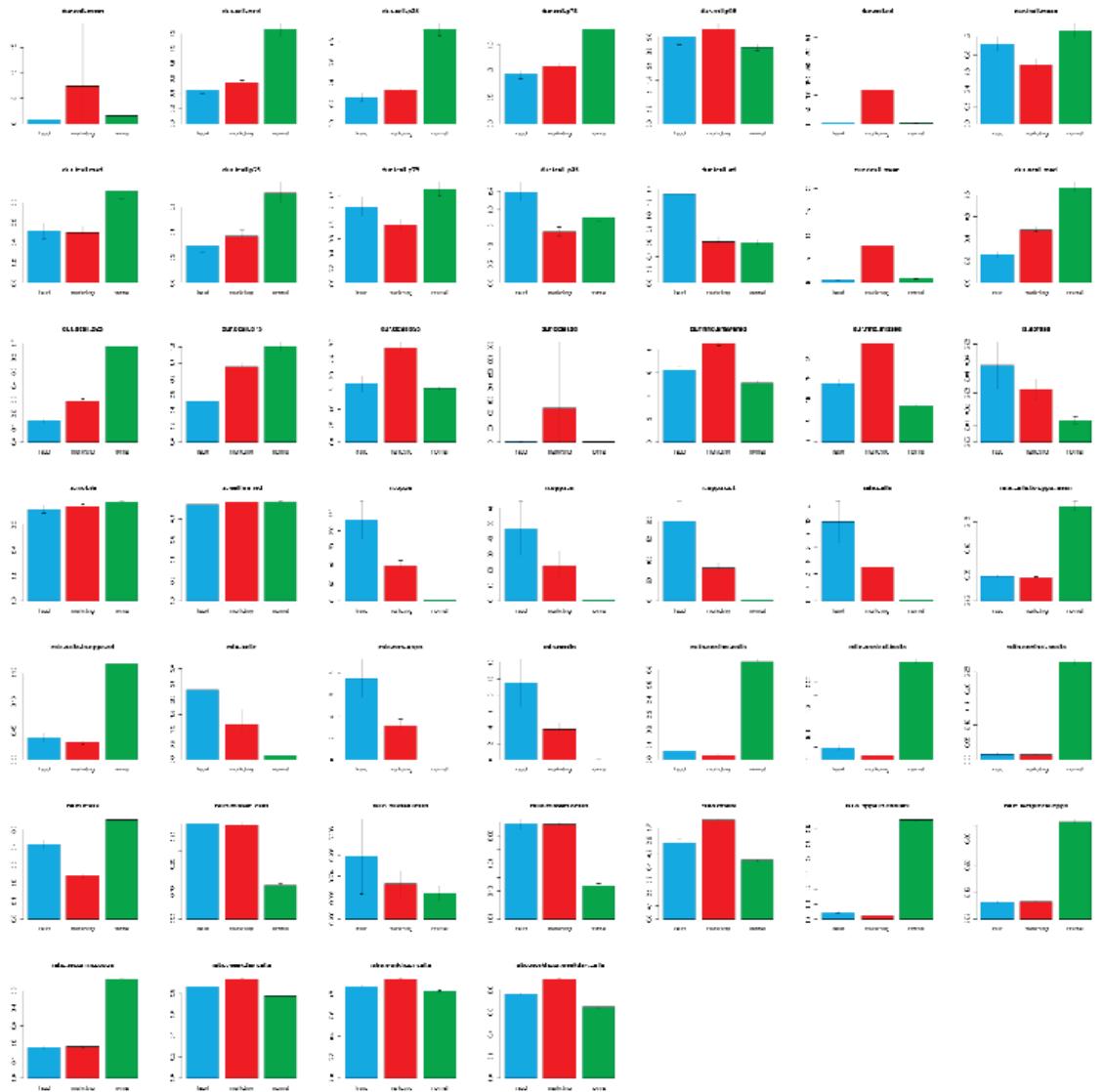
Type 1



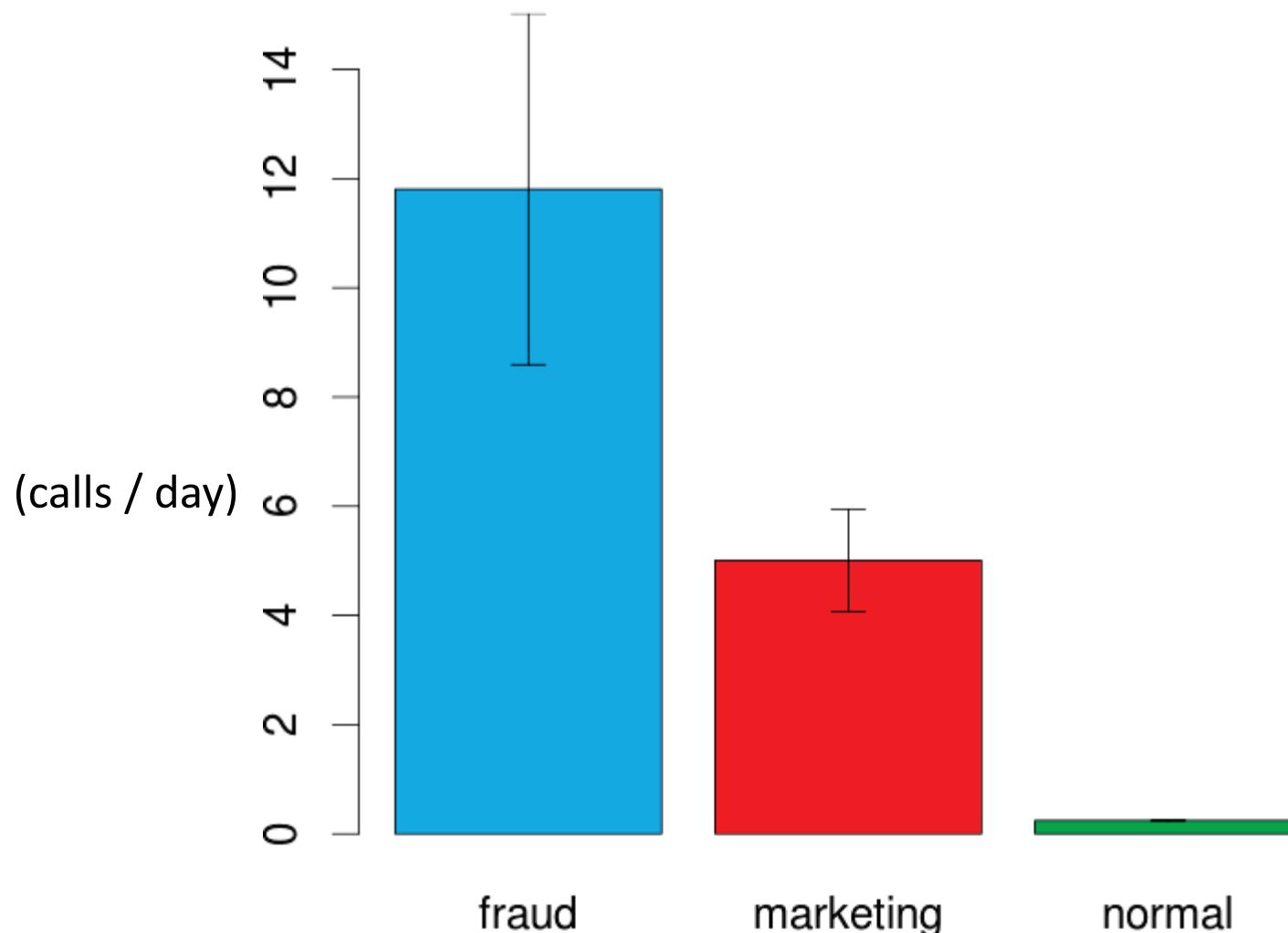
Type 2



A Side-by-Side Comparison

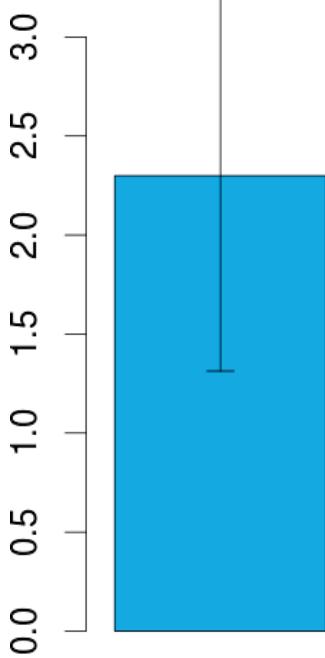


rate.calls



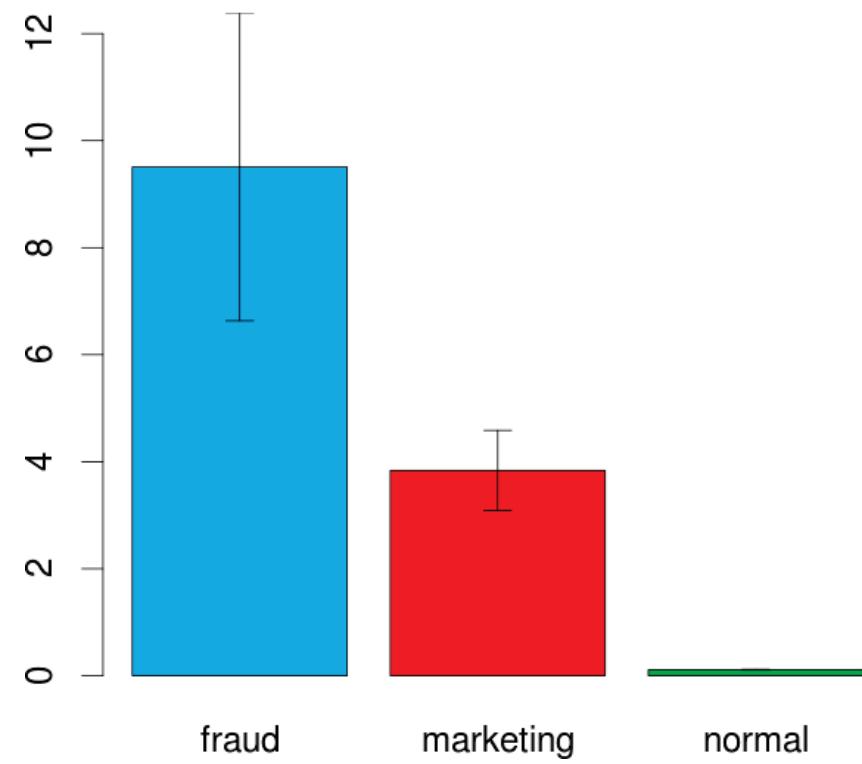
(calls / day)

rate.icalls

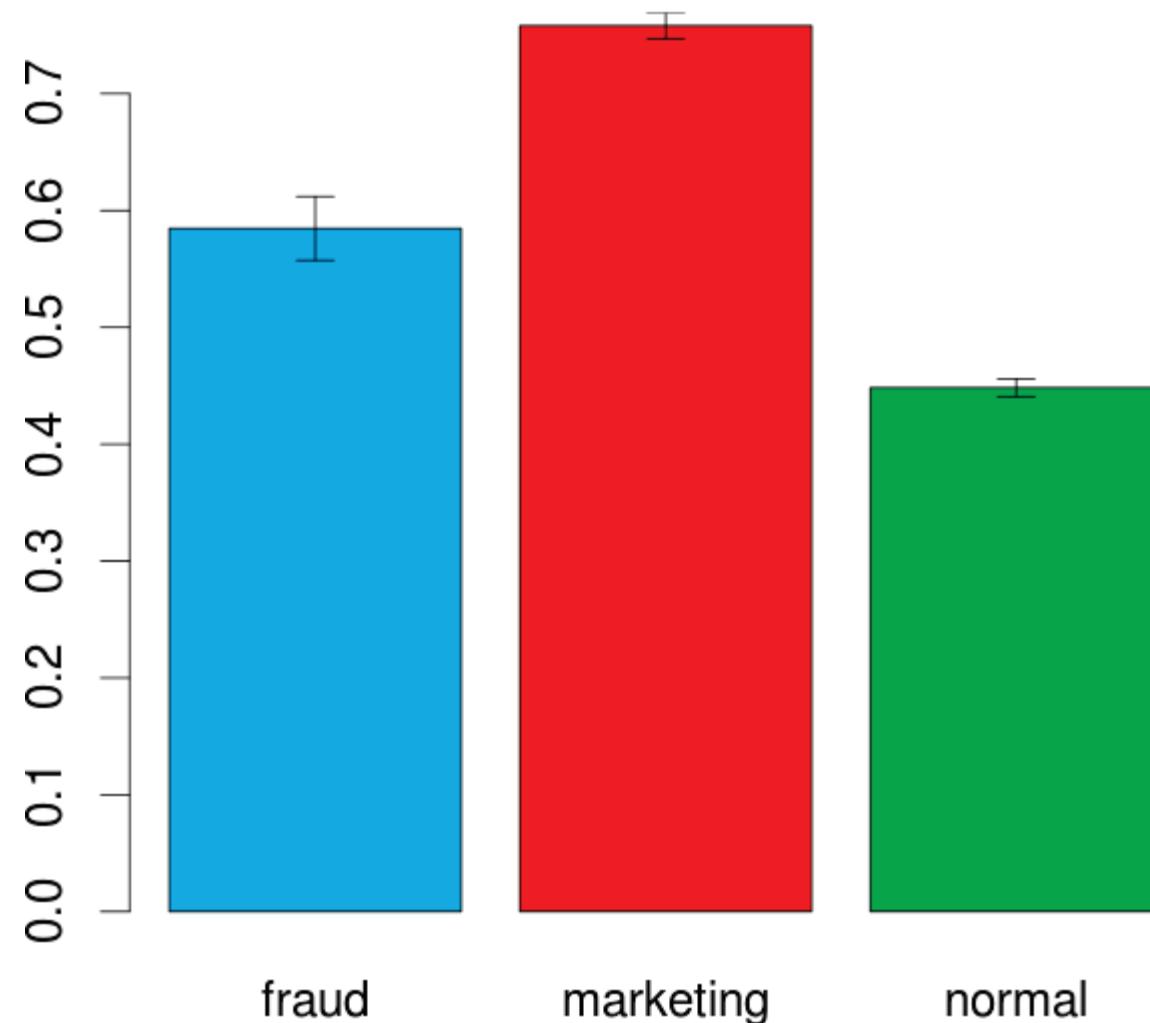


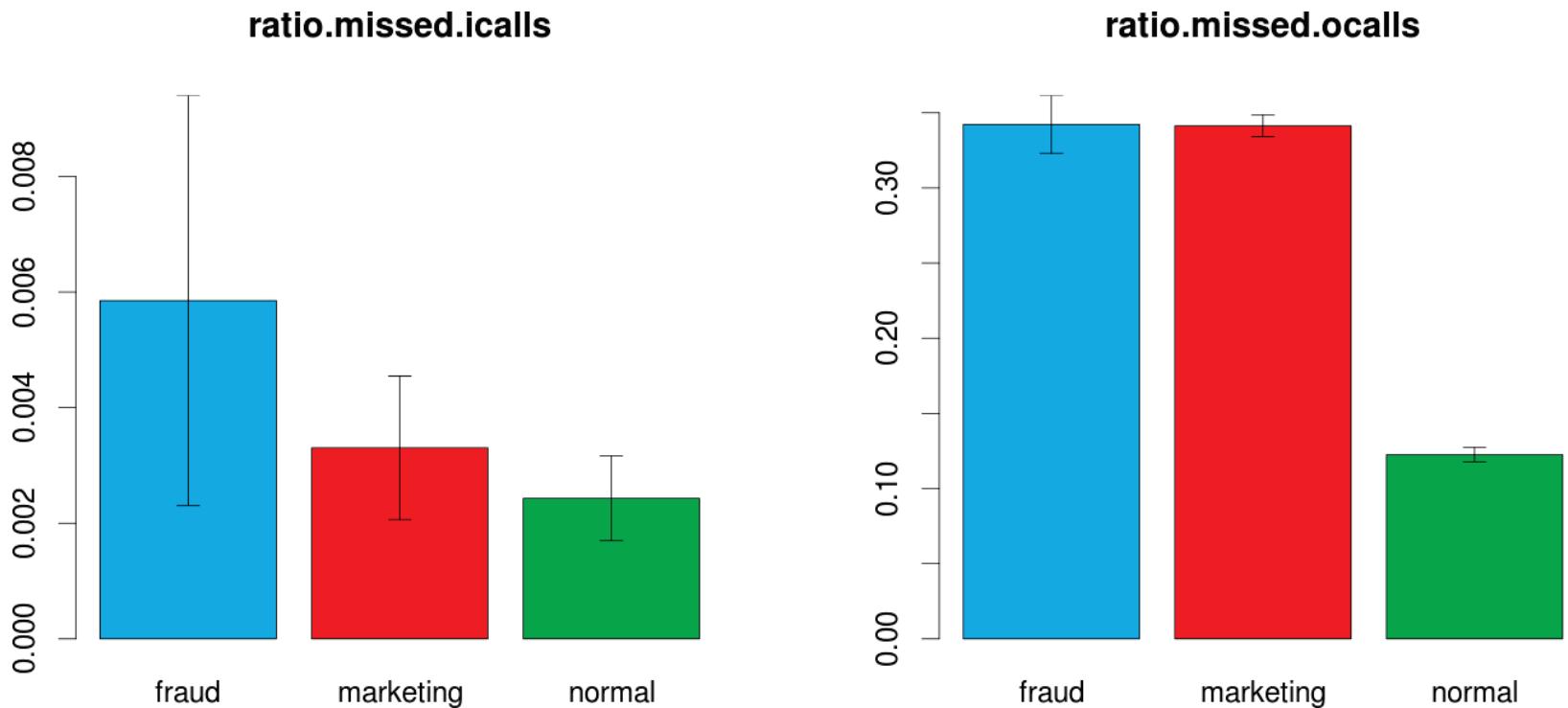
(calls / day)

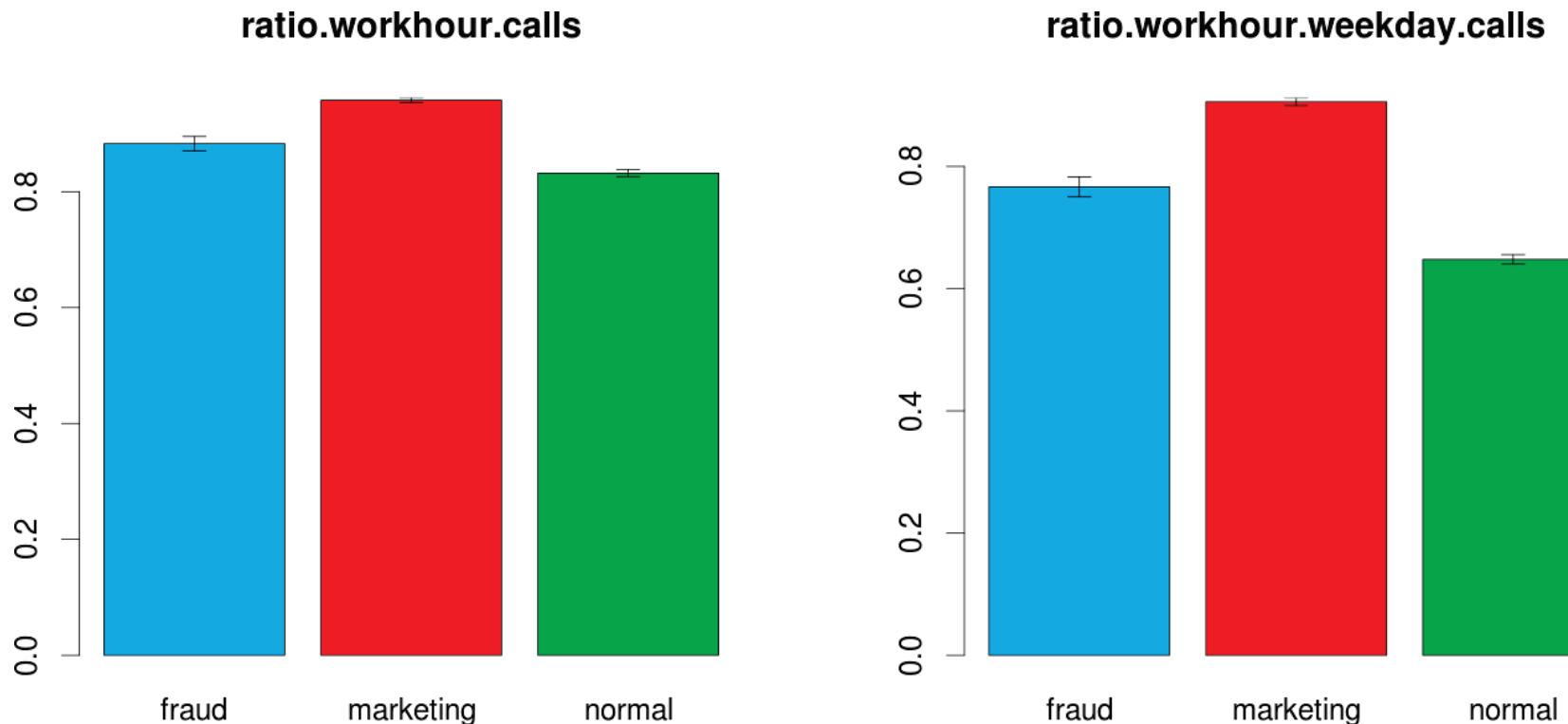
rate.ocalls



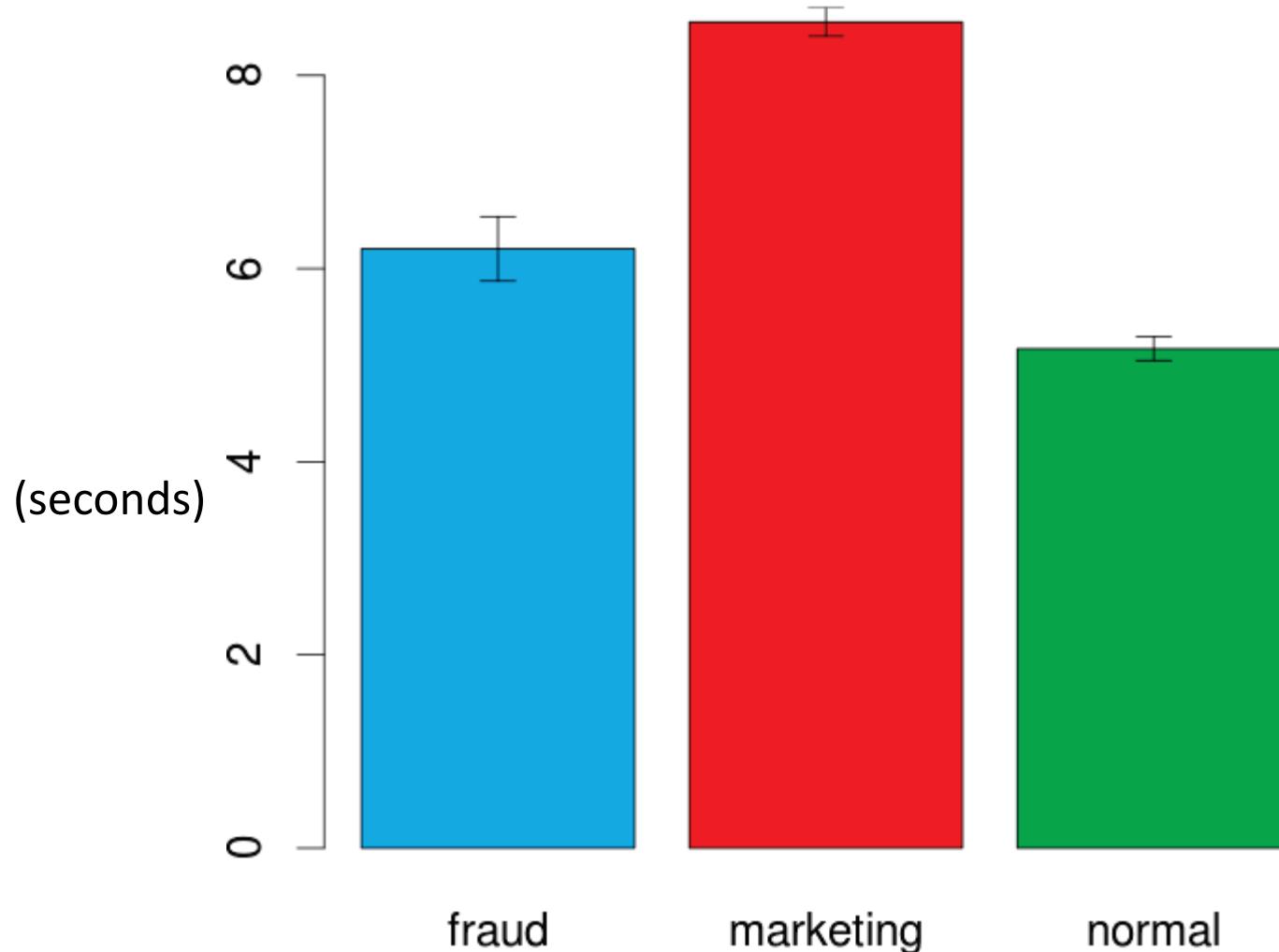
ratio.ocalls



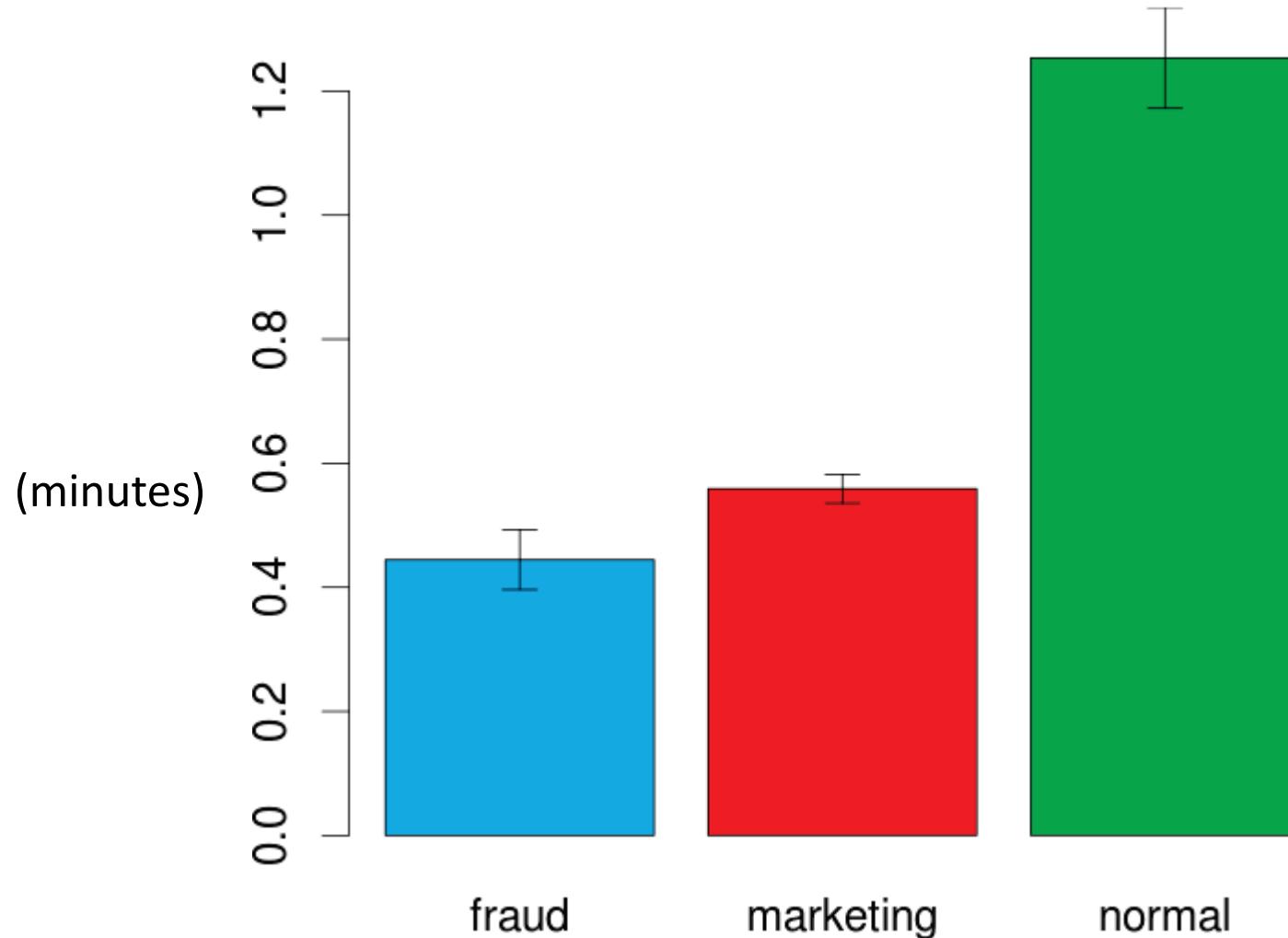


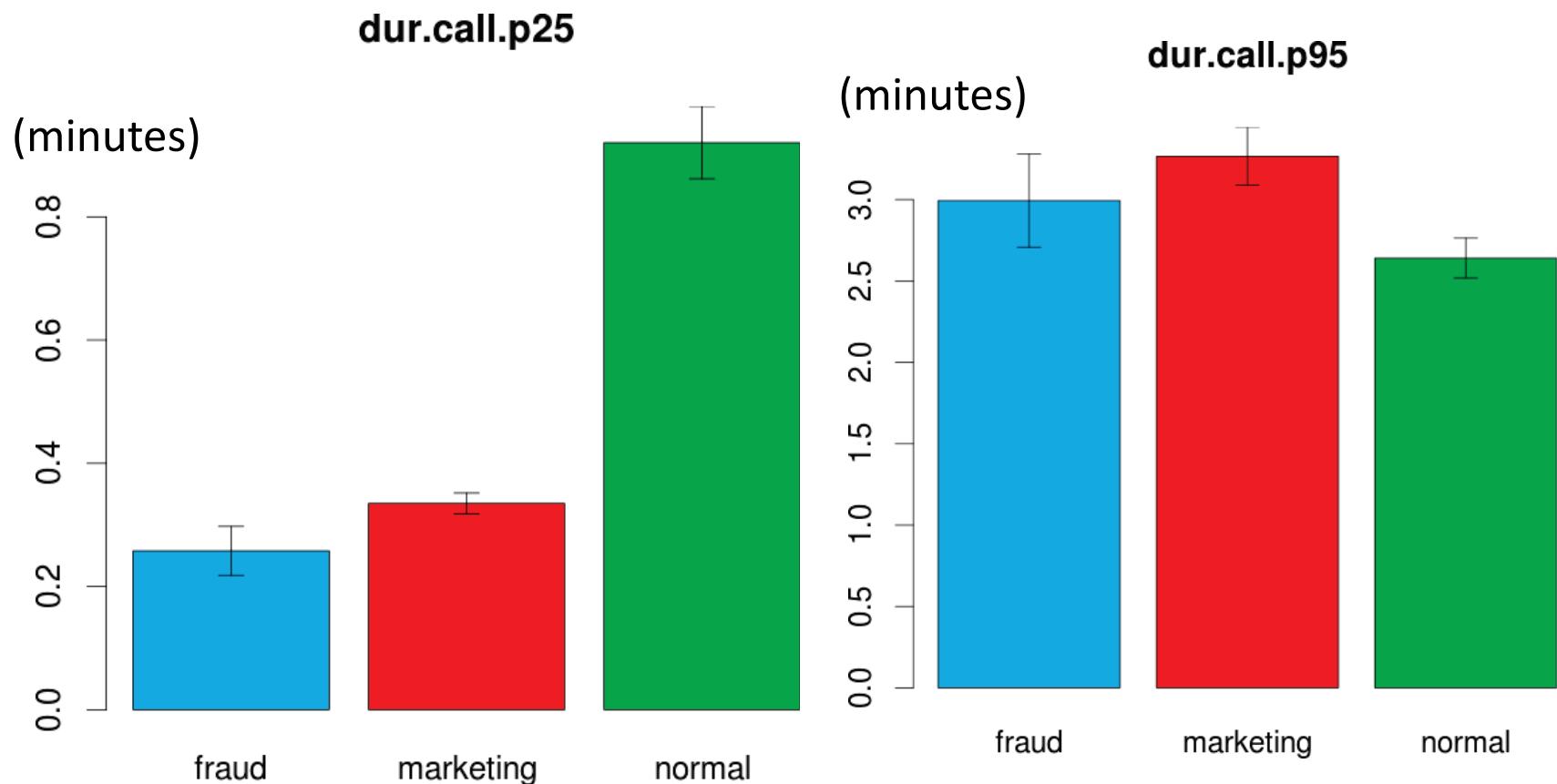


dur.ing.answered

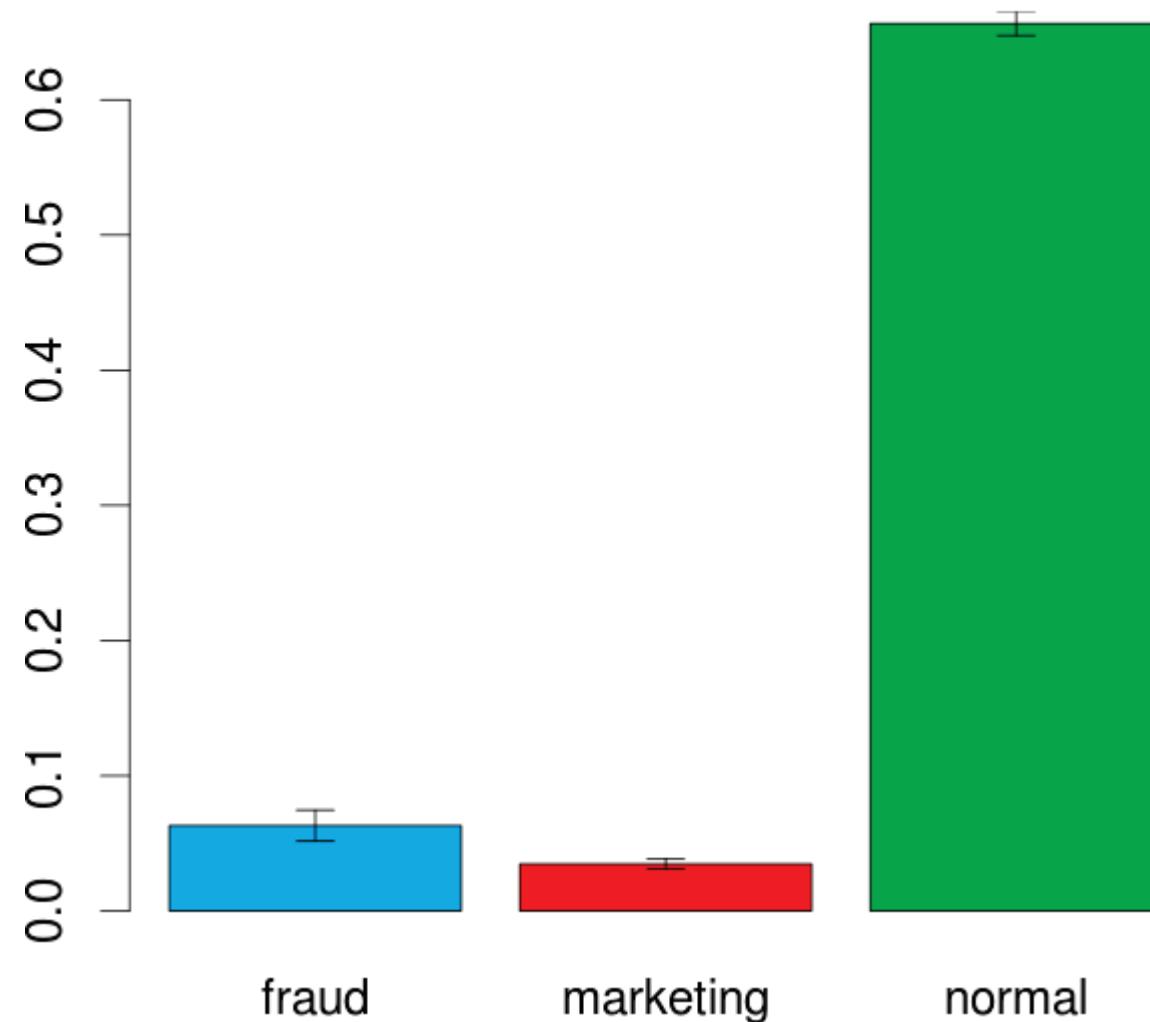


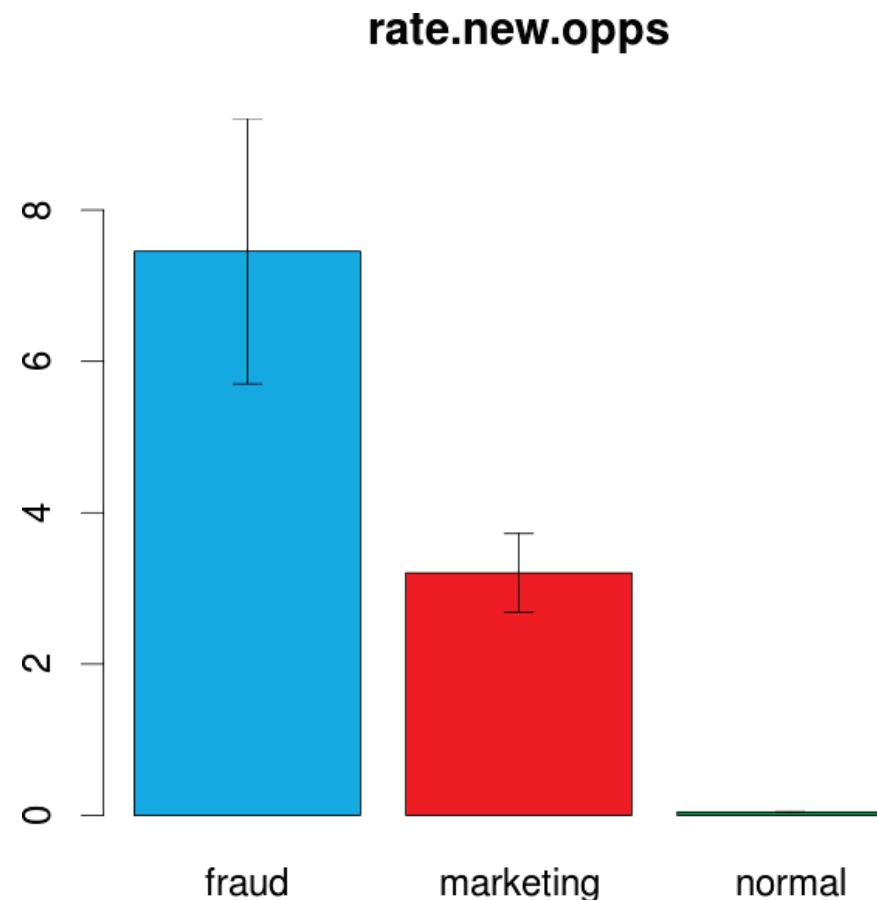
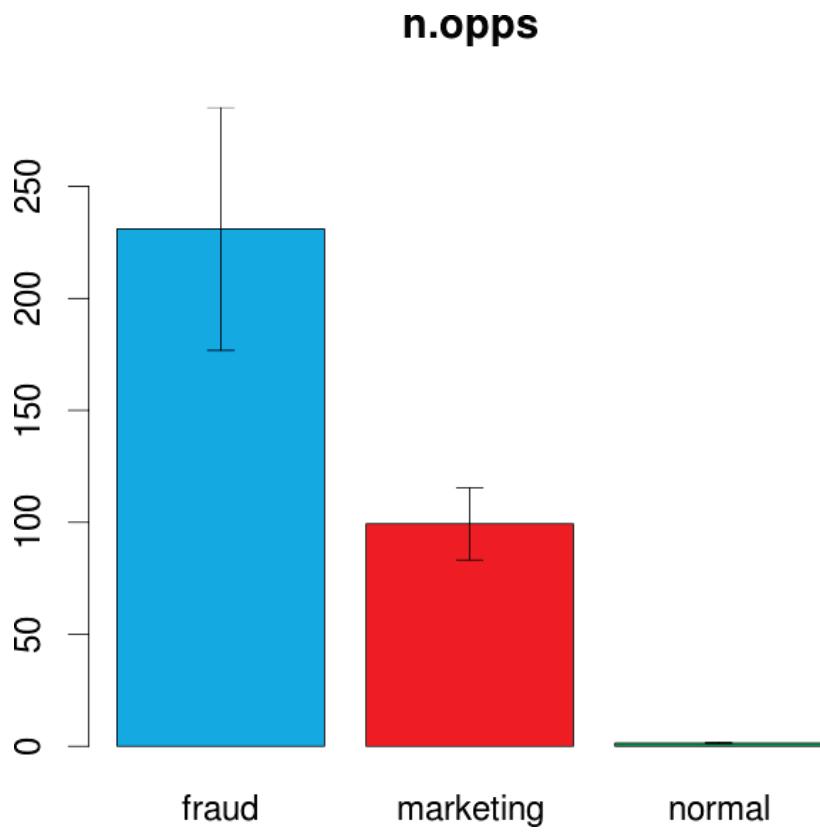
dur.call.med



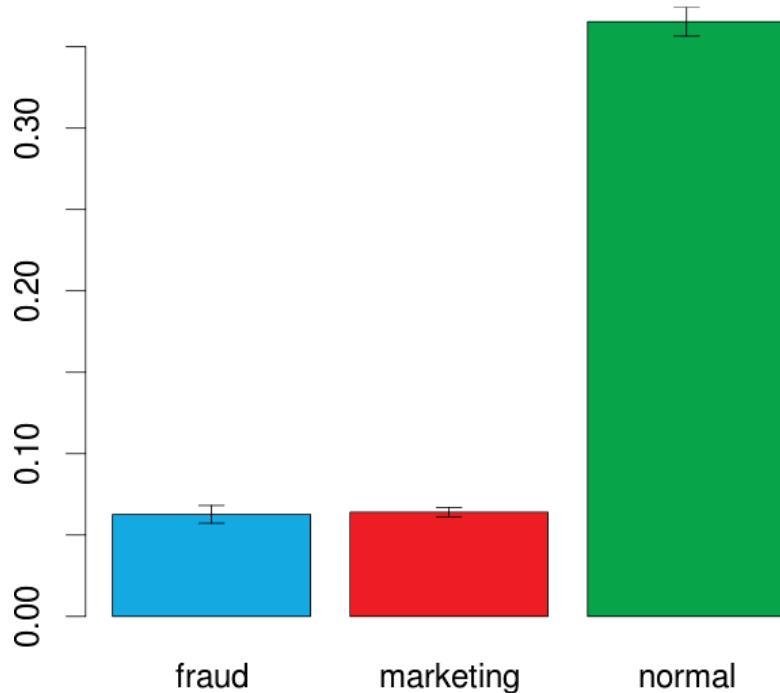


ratio.contact.calls

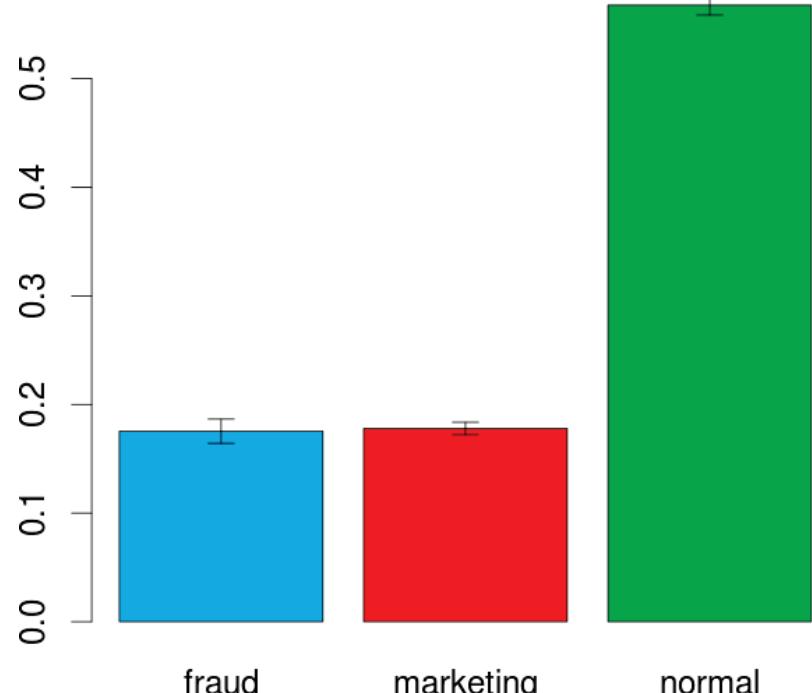




ratio.reciprocal.opps

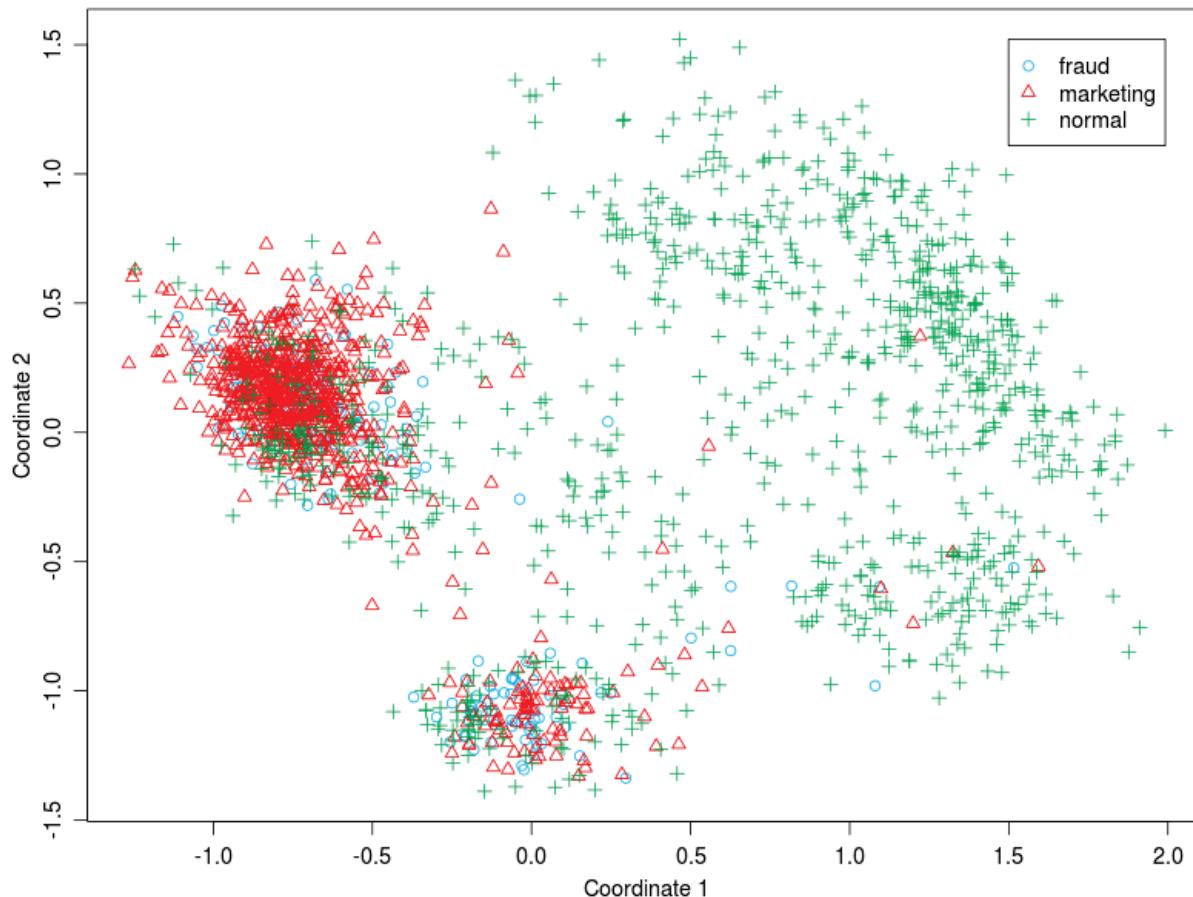


ratio.recurring.opps



Dimension Reduction

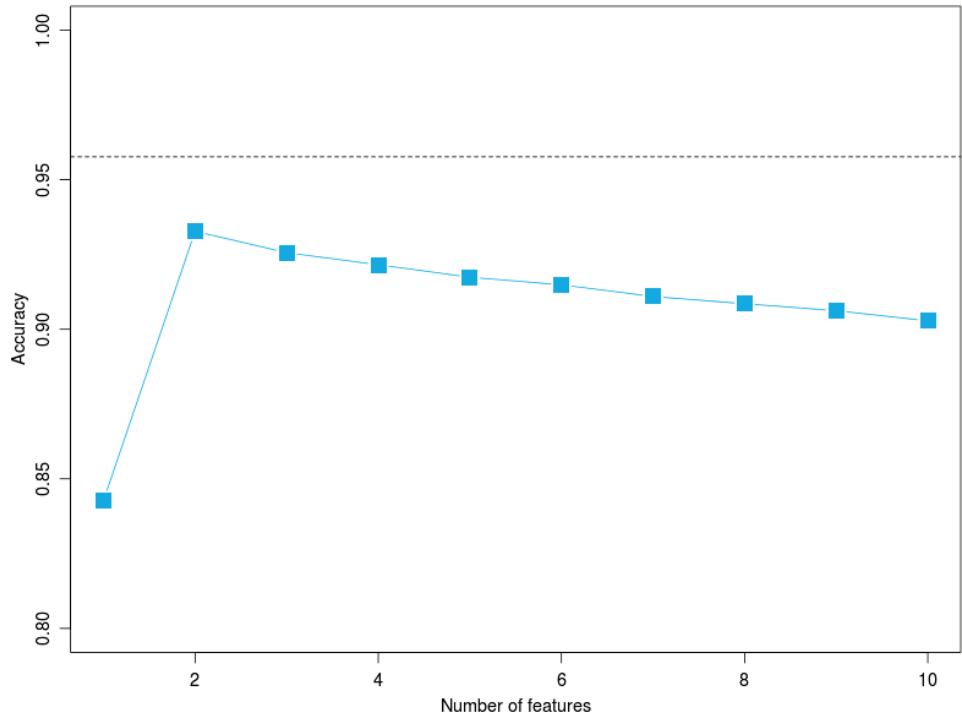
- 46 dimensions => 2 dimensions
- using classical MDS (multi-dimensional scaling)



Feature selection

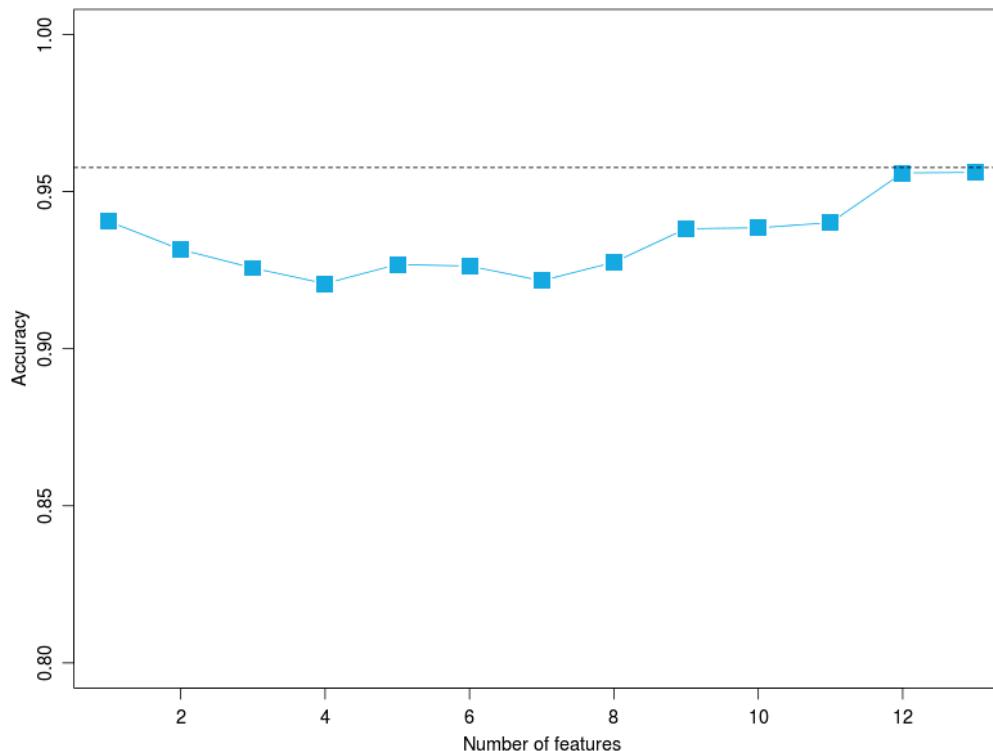
■ Using 2-norm SVM (support vector machine)

rate.calls	202.8896577	rate.new.oppss	201.6991125
rate.icalls	166.9199731	rate.ocalls	145.6960172
dur.call.sd	132.8181019	dur.ocall.sd	124.7282845
dur.icall.sd	73.2379431	dur.call.med	-62.7574665
dur.call.p25	-57.1773352	dur.call.p75	-53.3864202
dur.ocall.p25	-41.8495747	ratio.opp.in.contact	-40.4253833
rate.calls.to.oppss.mean	-38.3358544	ratio.recurring.oppss	-37.4713811
ratio.reciprocal.oppss	-36.7153847	dur.ocall.med	-35.1171411
ratio.missed.calls	33.5161904	ratio.missed.ocalls	31.2717160
dur.ring.missed	30.7990978	dur.ring.answered	30.0789343
dur.icall.p95	26.6858521	ratio.workhour.weekday.calls	26.1263872
dur.call.mean	-24.8529271	ratio.contact.ocalls	-23.8576452
ratio.contact.calls	-23.7470703	ratio.icalls	-19.0362513
ratio.ocalls	19.0362513	dur.icall.med	-18.1230396
dur.ocall.p75	-18.0652520	ratio.missed.icalls	17.7723104
ratio.weekday.calls	16.6667708	dur.ocall.p95	15.4224822



Feature Selection (cont)

```
vars.selected = c("rate.new opps", "rate.icalls", "rate.ocalls", "dur.ocall.mean",
"ratio.opps.in.contact", "rate.calls.to.opps.mean", "dur.icall.med",
"dur.ring.answered", "ratio.recurring.opps", "ratio.workhour.weekday.calls",
"ratio.reciprocal.opps", "ratio.missed.ocalls", "ratio.contact.ocalls")
```



你以為這樣就可以
收工了嗎？
太天真了...

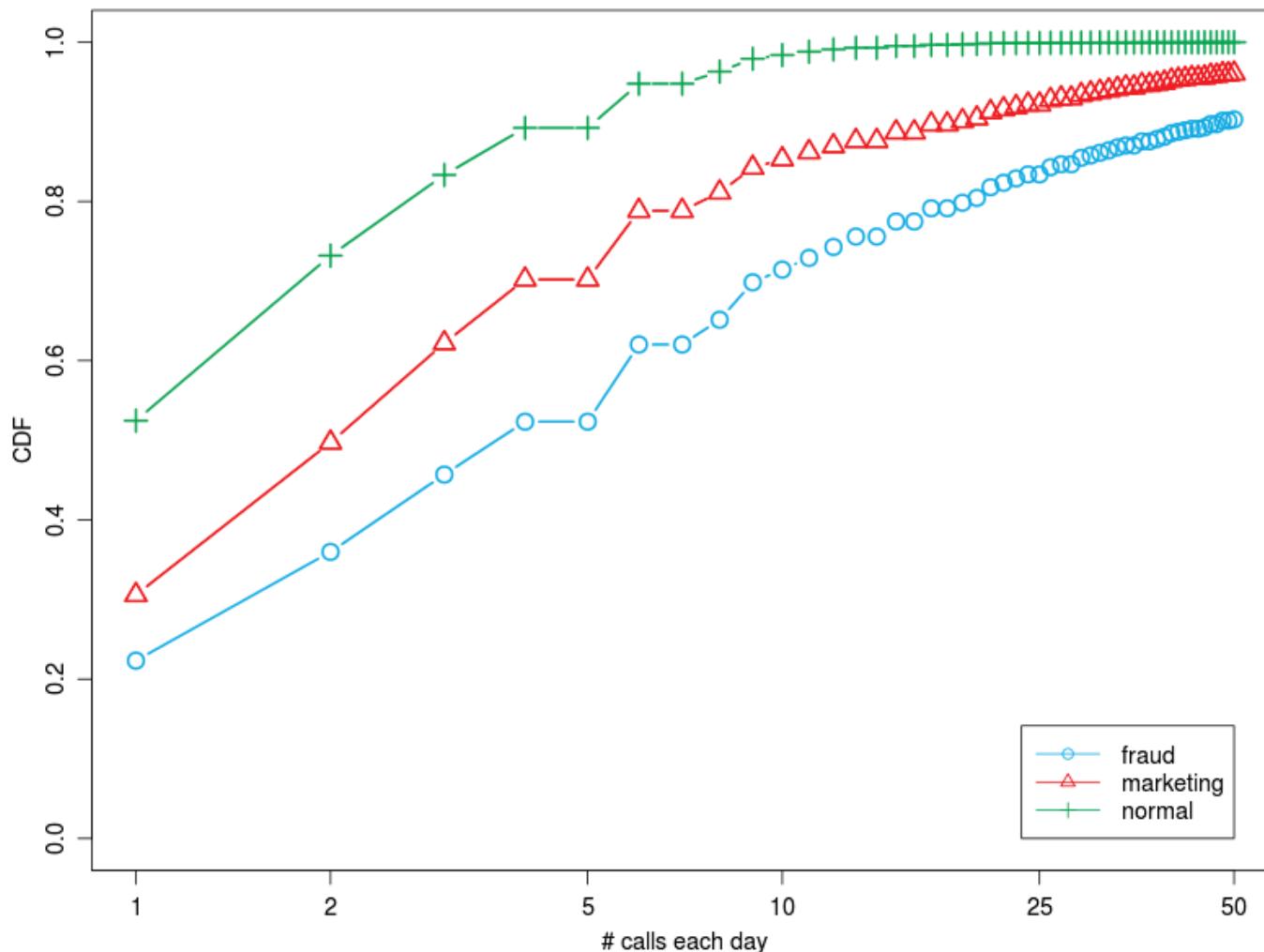


收工？

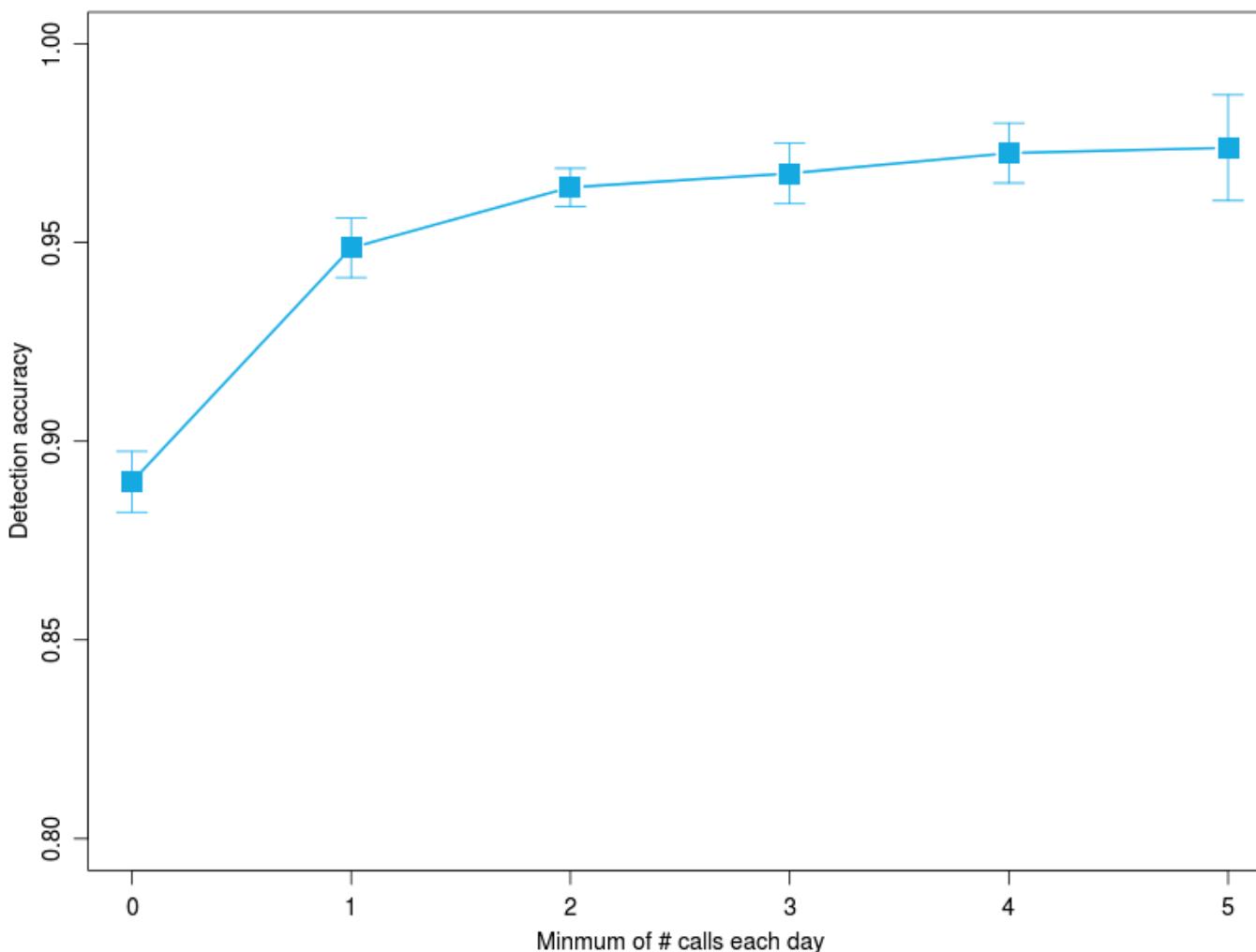
Our Goal

- Predict whether a number is malicious **as EARLY as possible**
 - In order to prevent further victims...
- Our goal: **accurate and FAST detection**

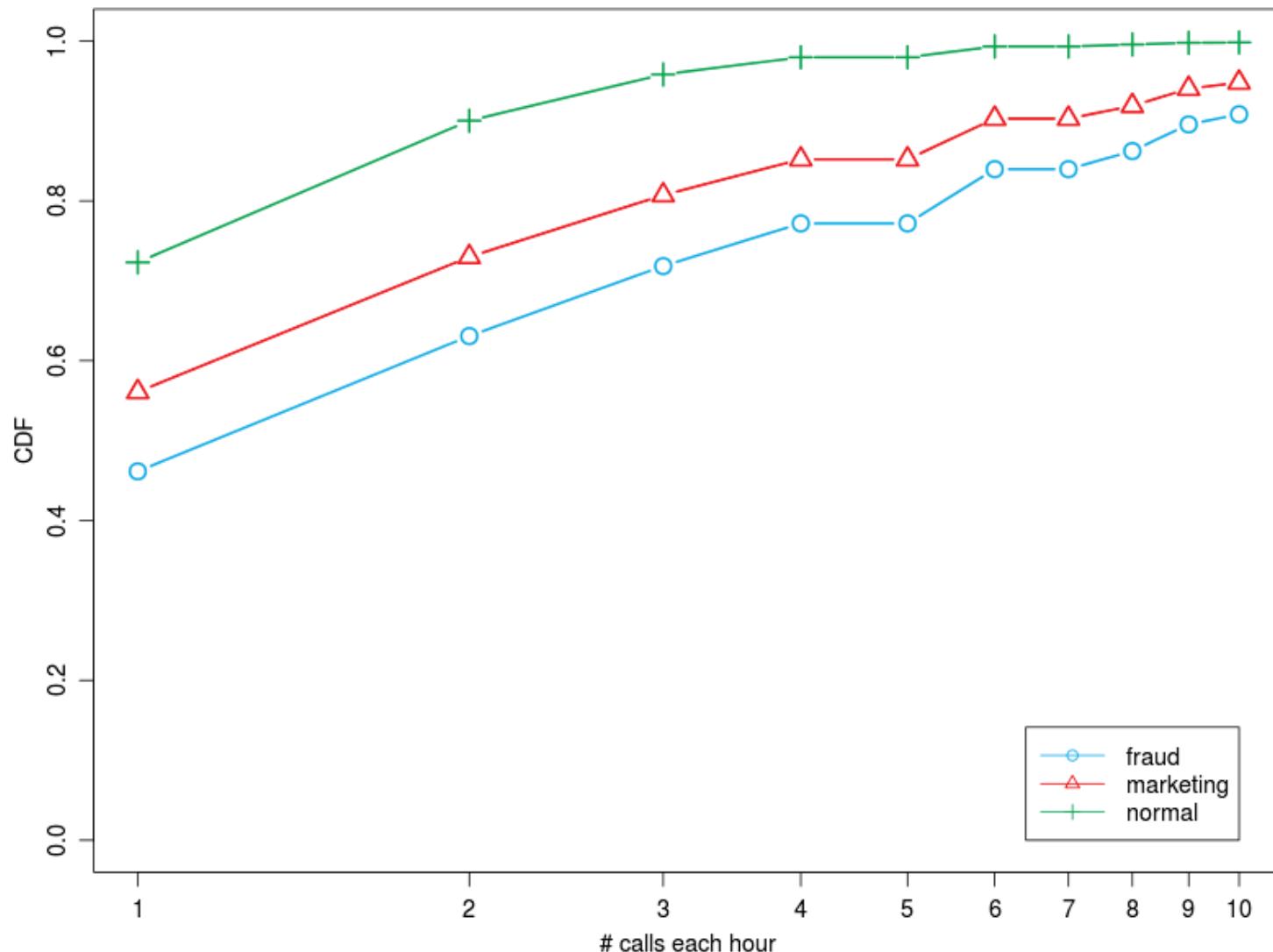
calls observed each day



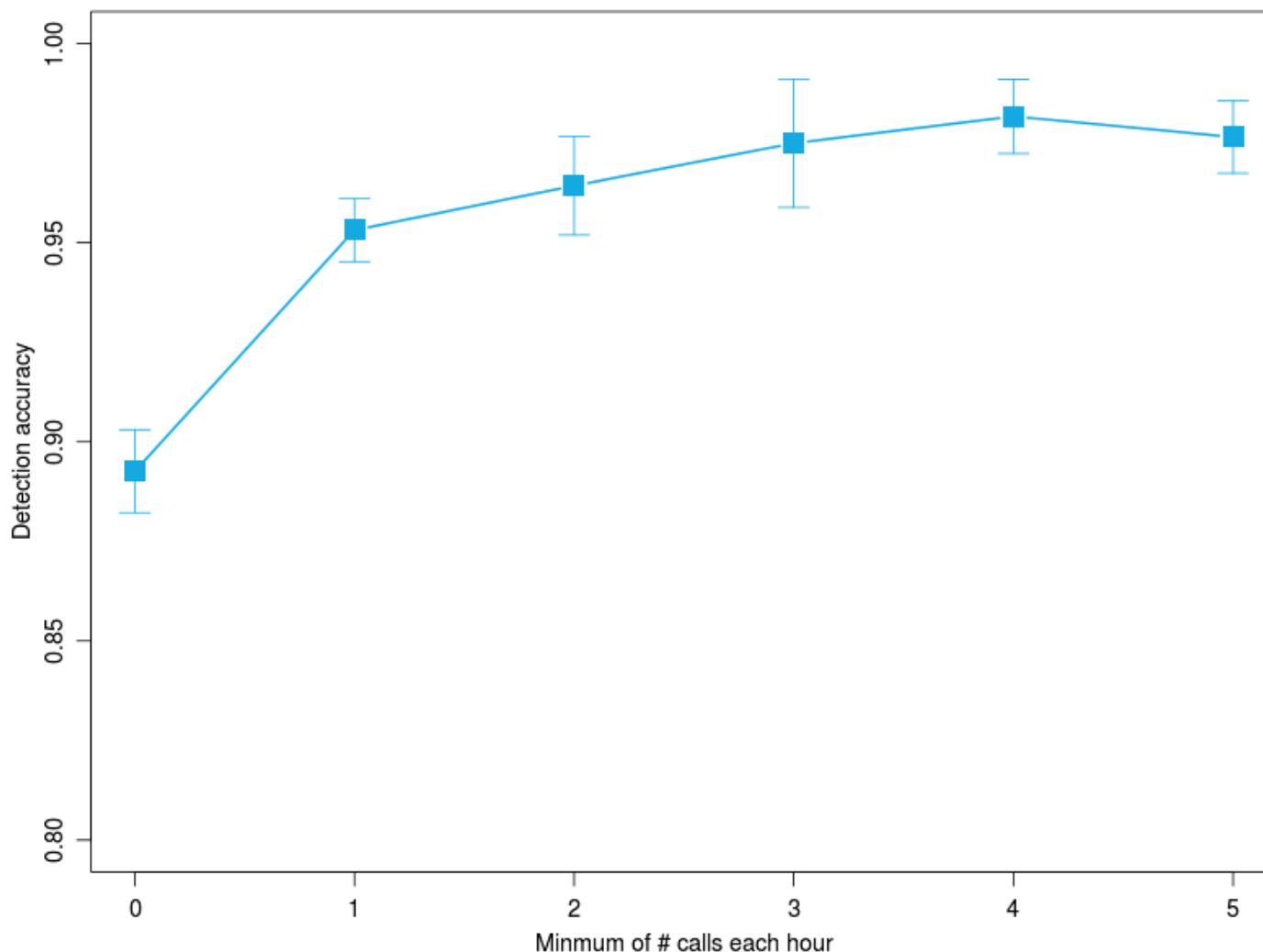
Observation time: Month → Day



calls observed each hour

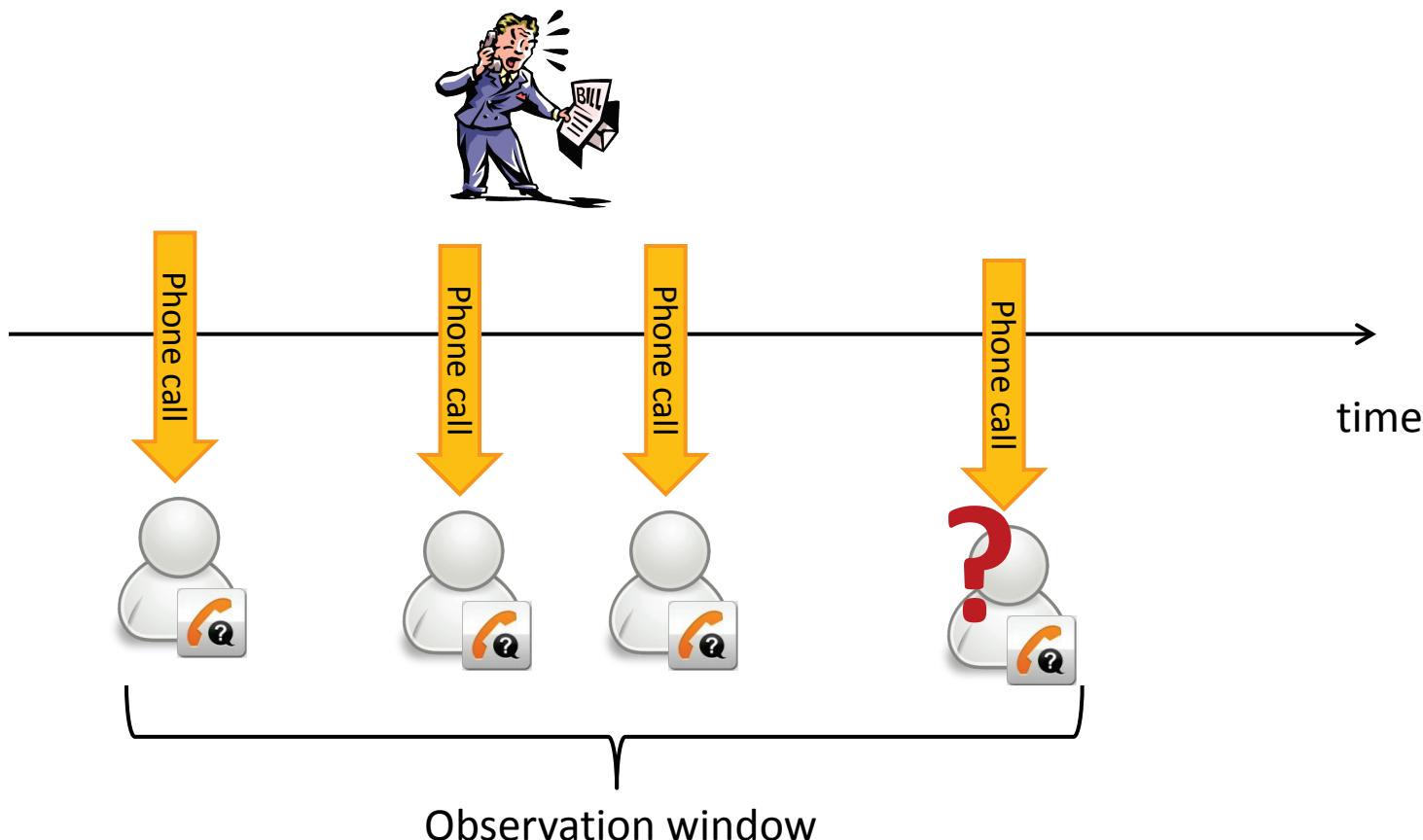


Observation time: Day → Hour

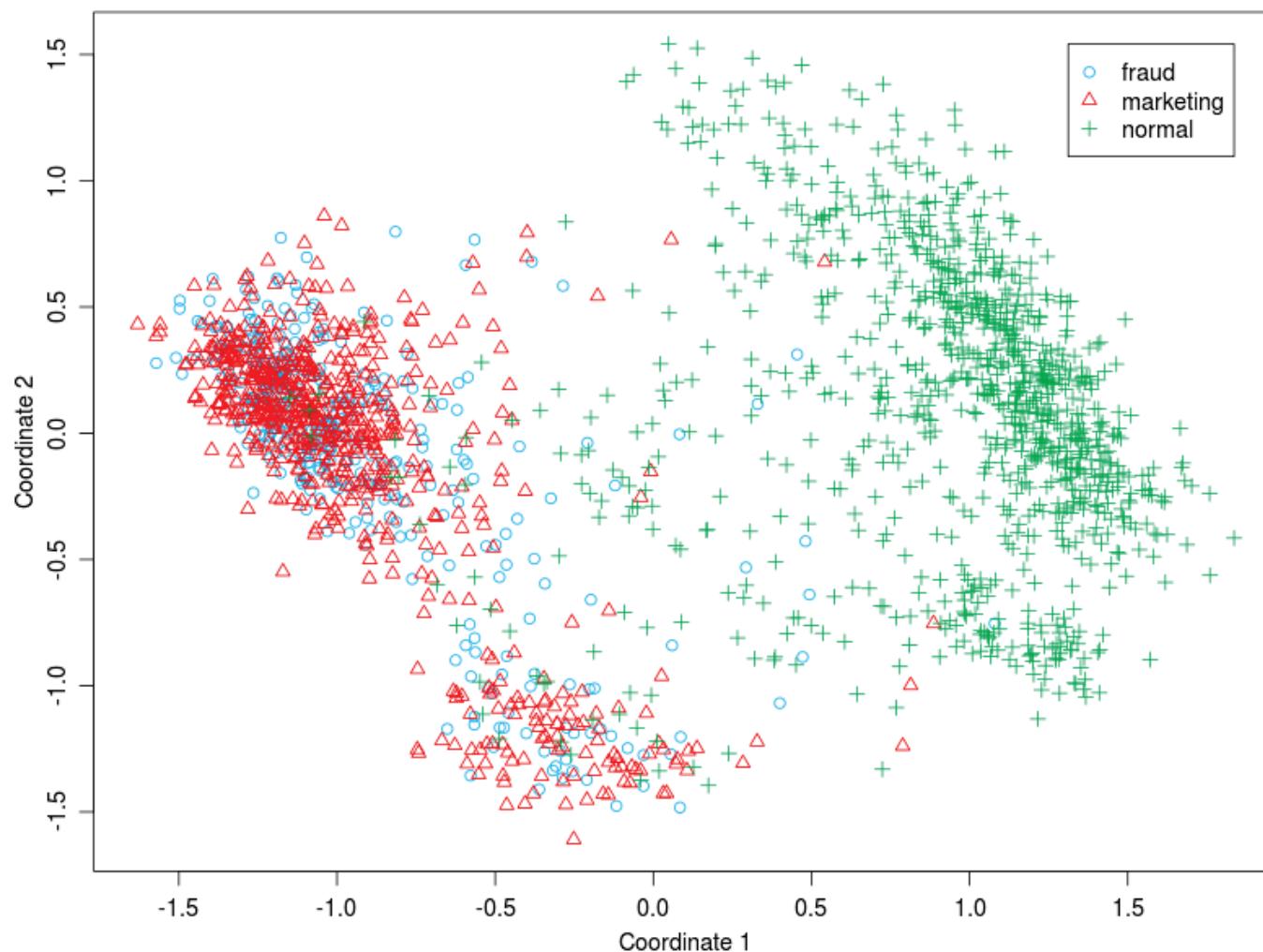


Dynamic observation period

- When we require malicious number prediction?
Ans: The time a phone call reaches a Whoscall user

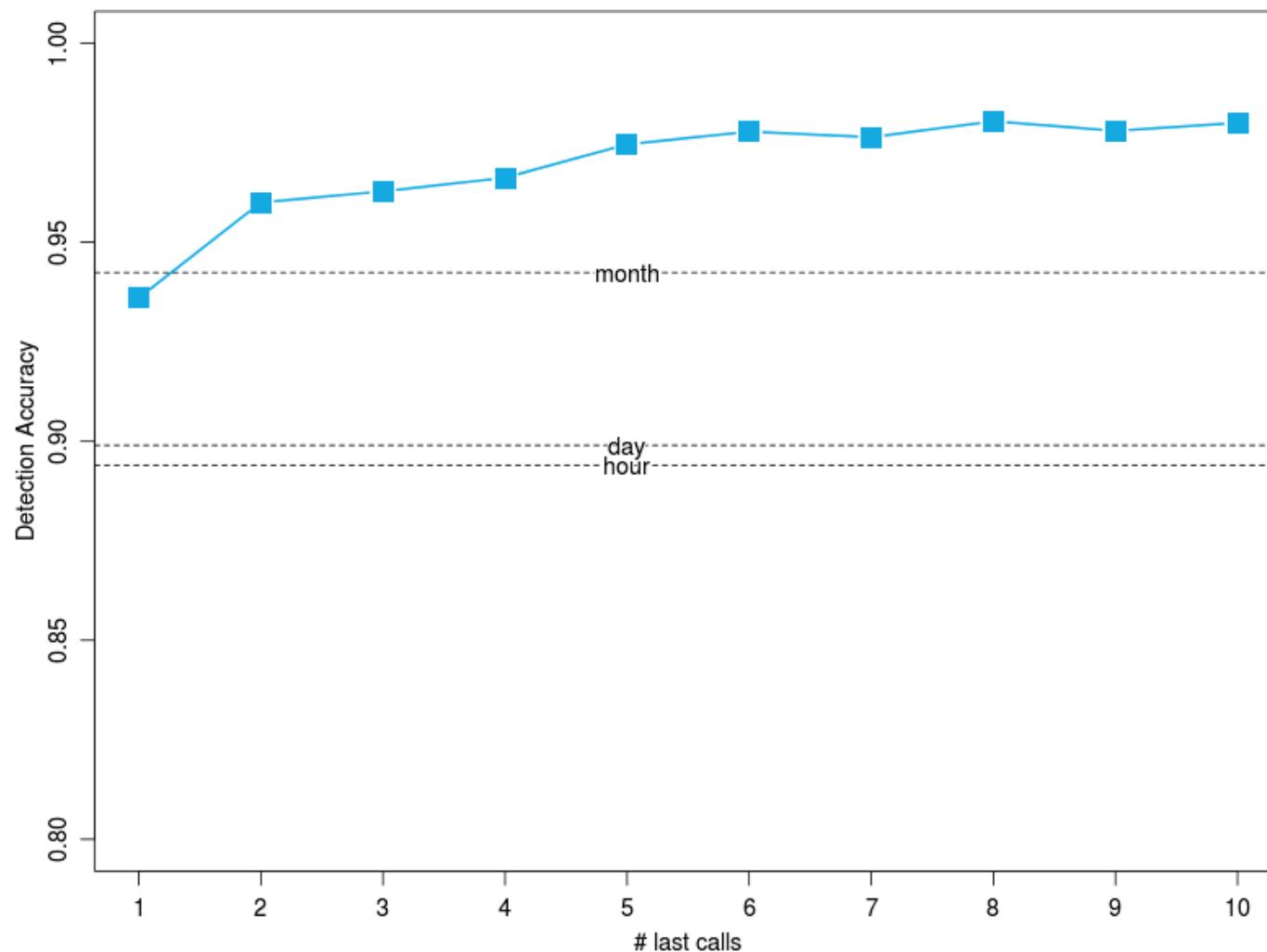


Observation time: The last 5 calls



dur.icall.p75	-150.813717	dur.icall.med	-147.846702
dur.icall.mean	-135.542911	ratio.recurring.opps	-119.340540
dur.icall.p95	-109.285196	dur.call.med	-97.421974
dur.icall.sd	-97.191078	dur.ocall.med	-96.747295
ratio.opps.in.contact	-94.542820	dur.ocall.p25	-91.305330
dur.icall.p25	-88.392960	dur.call.p75	-86.944857
ratio.contact.calls	-84.769716	rate.new.opps	83.520355
ratio.reciprocal.opps	-75.914189	rate.ocalls	75.236755
ratio.contact.icalls	-71.759195	dur.call.mean	-70.161483
dur.ocall.mean	-62.424721	dur.call.p25	-61.584172
ratio.contact.ocalls	-56.983550	ratio.missed.icalls	55.815130
dur.ocall.p75	-55.245537	rate.calls	54.187597
dur.call.p95	-47.966343	rate.calls.to.opps.mean	-40.440625
ratio.workhour.calls	38.906191	dur.call.sd	-35.599057
rate.icalls	-33.949360	ratio.workhour.weekday.calls	30.221138
ratio.missed.calls	27.973029	ratio.missed.ocalls	23.188426

Prediction based on the last N calls



真的可以收工了嗎？

Work in Progress

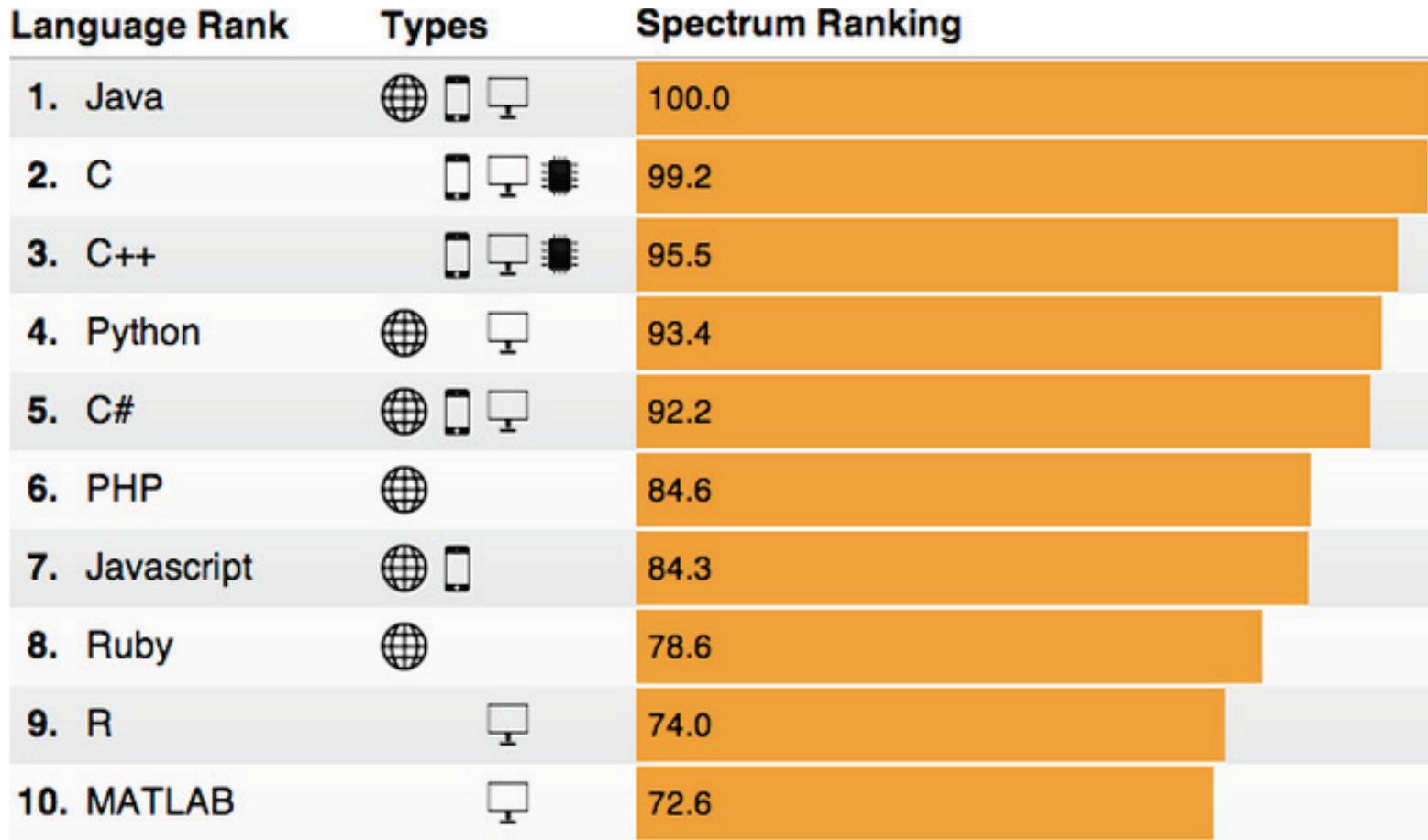
- Feature selection
- Anti-countermeasures
- Online learning
- Personalized penalty setting
- Crowdsourced tag correction mechanisms
- And much more...



(Tools I used in this project: awk + PHP + R)

A SHORT PROMO ON R

Why R ?



[1] IEEE Spectrum: The Top Programming Languages in 2014

<http://spectrum.ieee.org/static/interactive-the-top-programming-languages#index>

Starting R

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for generating a diamond pricing plot. The code includes library imports, data summaries, and a ggplot2 call.
- Console:** Displays the output of the R code, including summary statistics for variables x, y, and z, and the generated ggplot2 object p.
- Workspace:** Shows the diamonds dataset with 53,940 observations and 10 variables, along with other objects like aveprice, clarity, and p.
- Plots:** A scatter plot titled "Diamond Pricing" showing Price vs. Carat. The plot is color-coded by Clarity, with points ranging from I1 (red) to IF (pink).
- Bottom Status Bar:** Shows the date "1/20/12" and the time "9:56 AM".

Learning R....



Once you become an R expert...

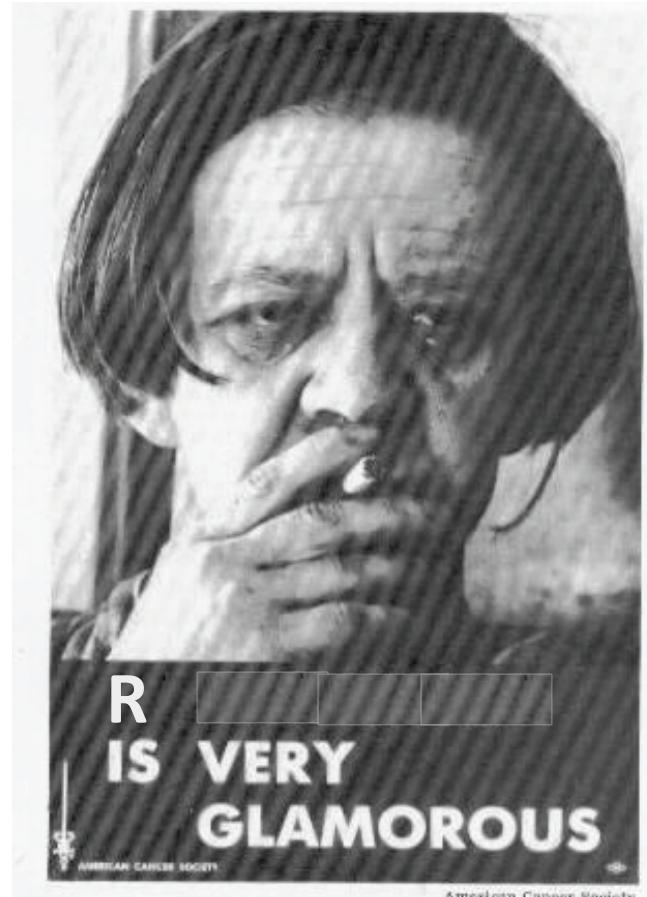


Demo of R Basics

- examine data.frame
- table, hist, ecdf, color
- plot, cor
- barplot, boxplot on dur.call.med

Final Words of Warning

- “Using R is a bit akin to smoking. The beginning is difficult, one may get headaches and even gag the first few times. But in the long run, it becomes pleasurable and even addictive. Yet, deep down, for those willing to be honest, there is something not fully healthy in it.” --Francois Pinard



R [REDACTED]
IS VERY
GLAMOROUS

American Cancer Society

American Cancer Society

TW.R 社群 & MLDM Monday

- 聚會資訊

- 時間：每週一晚上七點辦
- 地點：政大創立方
- 報名網址：<http://www.meetup.com/Taiwan-R/>

- 社群資訊：



Taiwan R User Group
507 likes · 126 talking about this

Liked Message ⚙

Community
近期活動查詢與報名請至 <http://www.meetup.com/Taiwan-R>
過期活動影片觀賞請至 <https://www.youtube.com/user/TWuseRGroup>

About – Suggest an Edit

Photos Likes



507



交流時間





有沒有人在偷用你的臉書？

陳昇瑋

中央研究院資訊科學研究所

The Prevalence of Social Network Services

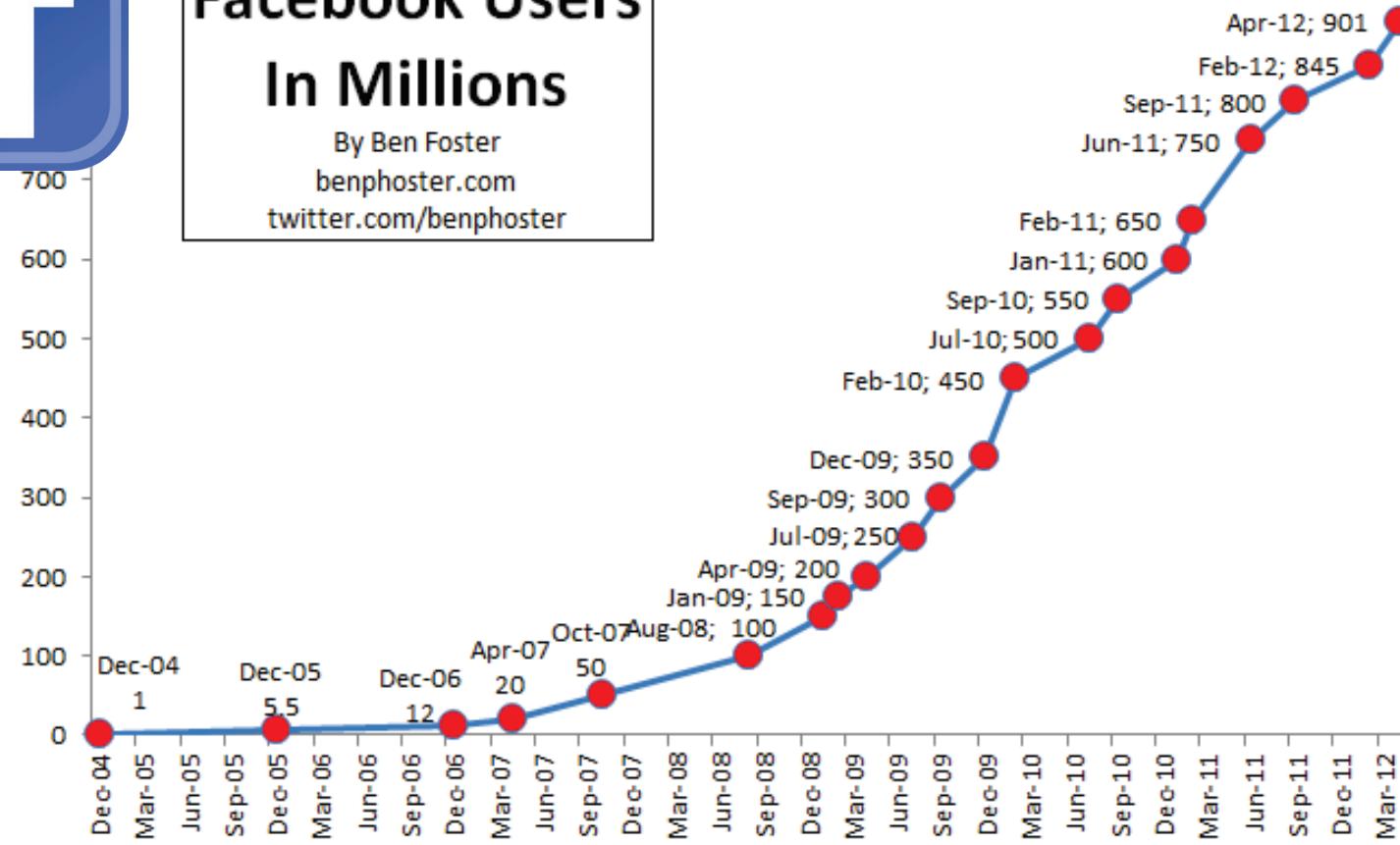


Facebook Users In Millions

By Ben Foster

benphoster.com

twitter.com/benphoster



Sensitive Info on SNS: A LOT!

- Personal info
 - Photos, Diary, Schedule
 - Groups, Pages, Likes
- Connections with friends
 - Friends' information
 - Friends' photos, demographics, and so on
- Interactions with friends
 - Conversations



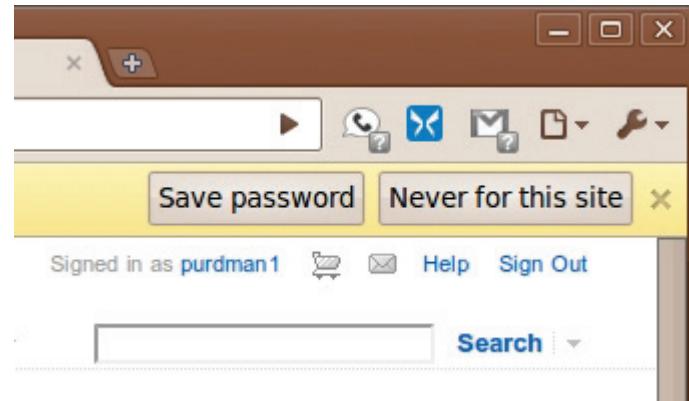
Are those information safe?

stealthy use of SNS accounts
is commonly seen.



Stealthy Use: Tips 1, 2, 3!!

- People let browsers to manager their passwords
- Entering password on mobile devices is cumbersome
- People left SNS logged on when they're temporarily away

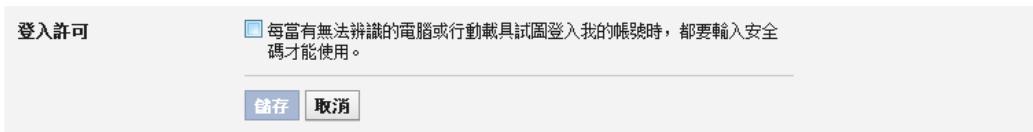


Existing Measures of Facebook

- 紀錄 IP address 、作業系統及瀏覽器種類



- 註冊裝置：經過簡訊回傳認證碼驗證裝置



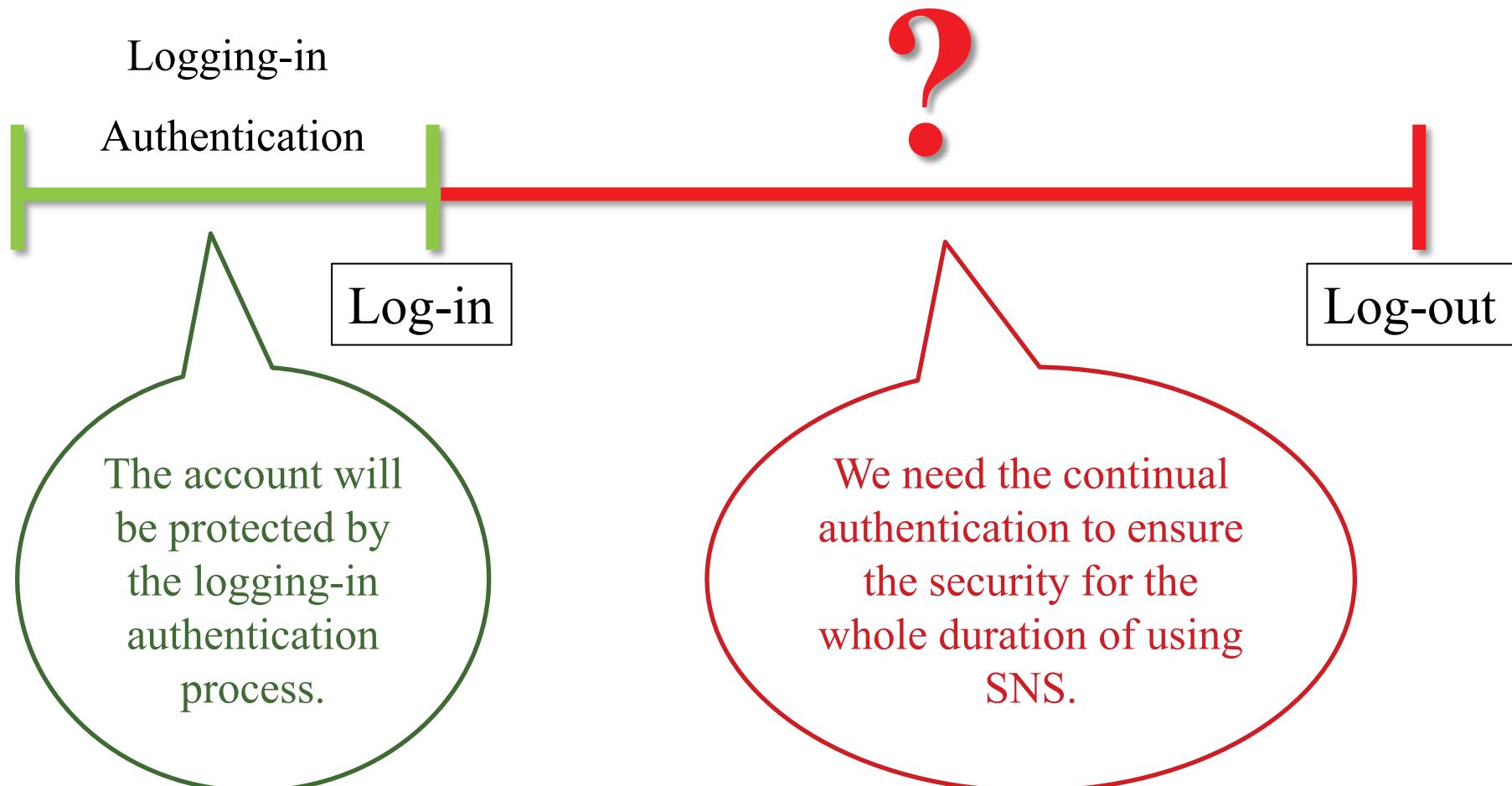
- 然而，這些方法都無法辨別一台已註冊的電腦，是否目前為註冊者本人操作，被盜用時無法即時得知。

Our Approach

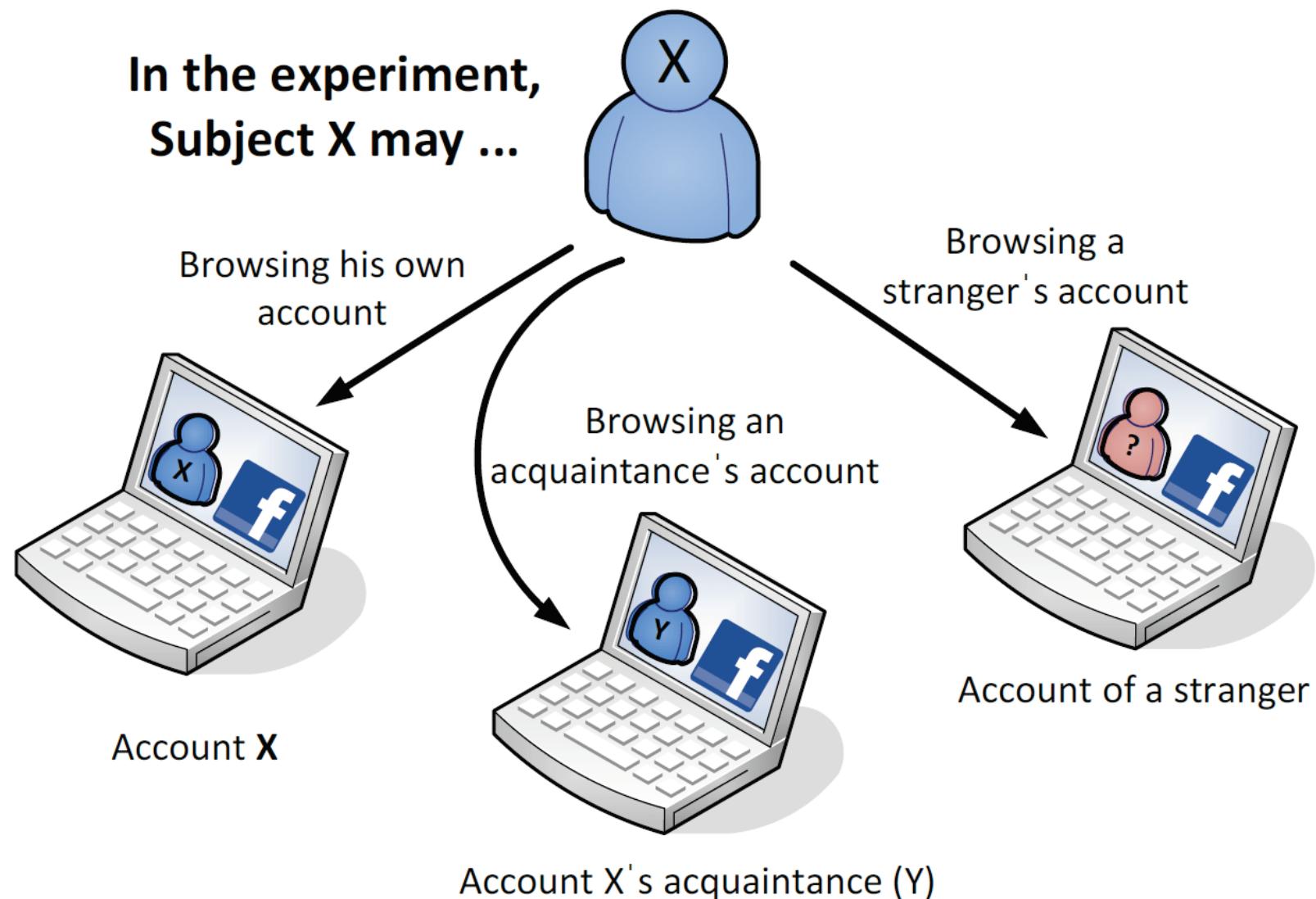
- Rationale: 不同人使用同一個帳號時，瀏覽行為也會不同。
 - 會特別注意某位朋友的資訊嗎？
 - 會多少時間瀏覽新資訊？
 - 會如何瀏覽過時資訊？
- 透過機器學習，判斷瀏覽行為是否為帳號擁有者所進行。
- 當偵測到異常的行為時，透過行動電話或是電子郵件通知帳號擁有者，以確保帳號安全。

The Loophole Of User Identity Process

The whole duration of using SNS



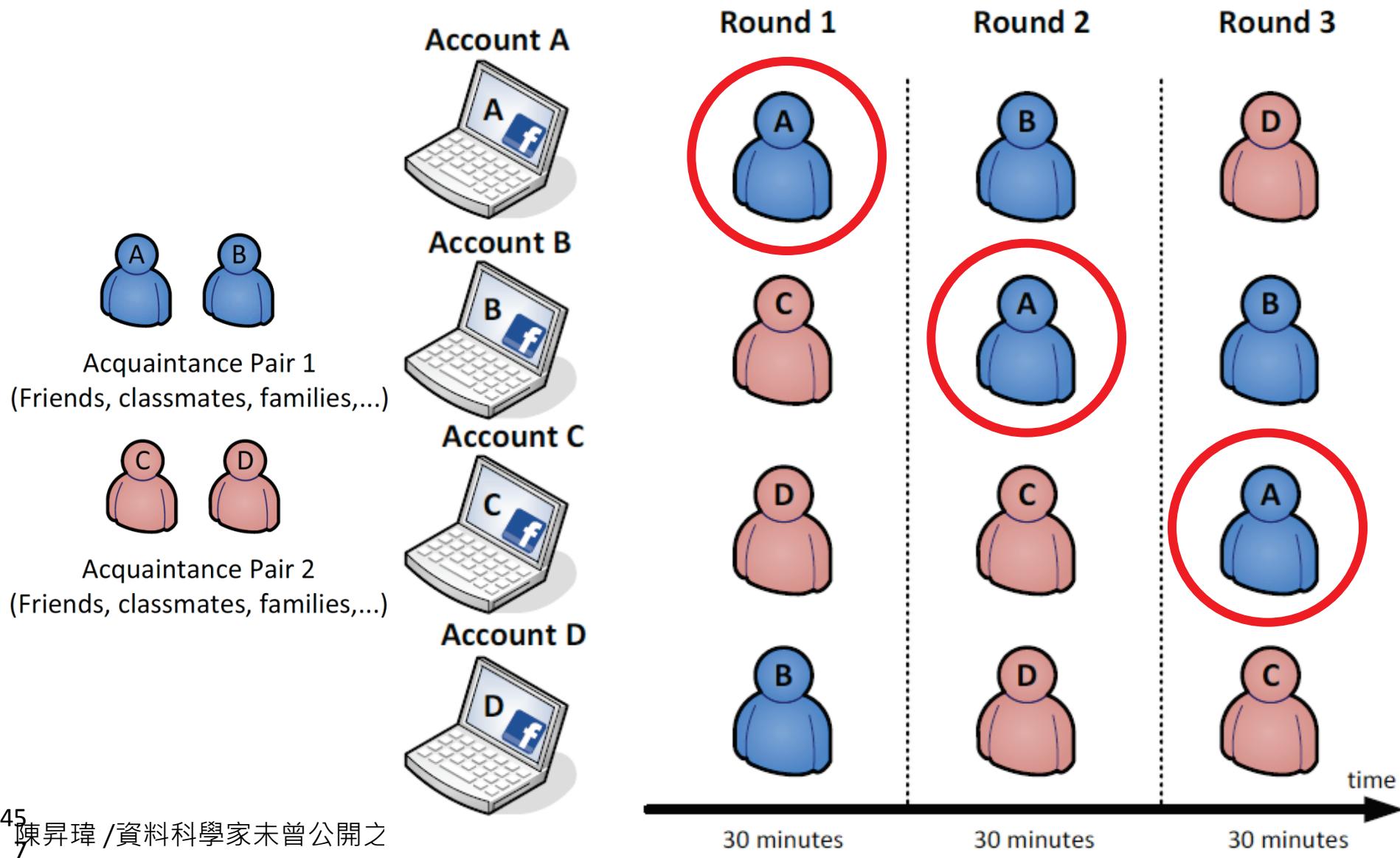
The 3 Different Roles Of A Subject



User Studies

1. 受測者必須兩兩認識的人為一對，關係可以是家人、朋友、情侶、同事或同學。
2. 每位使用者登入自己的 Facebook 帳號，並瀏覽個人朋友清單。
3. 接下來實驗分為三階段，每一階段約30分鐘，並隨機安排位置，每個人有可能使用非自己的帳號。
4. 記錄下每一筆與 Facebook 主機間的 http request 及 response。
5. 實驗完成後，請使用者填寫個人基本資訊，包含年齡，性別，與同組夥伴的關係。

Experiment Procedures



Data Collection: HTTP Spying

- Intercept all HTTP communications (including AJAX req. and resp.) between the subject's PC and Facebook servers

The screenshot shows the Fiddler application interface. The main window displays a list of captured web sessions. The 'Web Sessions' table has columns for #, Result, Protocol, Host, URL, and Body. Most entries are for Facebook, with one for fiddler2.com. Session 6 is highlighted in green. The right side of the interface shows detailed views for the selected session, including 'Request Headers', 'Cache', 'Client', and 'Cookies / Login' tabs.

#	Result	Protocol	Host	URL	Body
1	200	HTTP	www.fiddler2.com	/fiddler2/updatedcheck.asp?isBeta=False	450
5	200	HTTP	www.facebook.com	/tonyminghung.wang	298,906
4	200	HTTP	www.facebook.com	/ajax/typeahead/search/bootstrap.php?filter[0]=app&fil...	95
4	200	HTTP	www.facebook.com	/ajax/typeahead/search/bootstrap.php?filter[0]=user&...	95
5	200	HTTP	www.facebook.com	/ai.php?aed=AQKfxgjy-S37fZzfiFEgCRbmf-sQNBy22CTo...	110
6	200	HTTP	www.facebook.com	/images/spacer.gif	43
6	200	HTTP	www.facebook.com	/ajax/chat/user_info.php?ids[0]=552172983&ids[1]=56...	5,417
8	200	HTTP	0-ect.channel.face...	/pull?channel=p_100001894876481&seq=1&partition=0...	35
8	200	HTTP	www.facebook.com	/ajax/pagelet/generic.php/PersonalProfileWallTabPagele...	65,905
8	200	HTTP	www.facebook.com	/ajax/hovercard/user.php?id=100001894876481&endp...	3,467
8	502	HTTP	pixel.facebook.com	/ajax/hovercard/shown.php?asyncSignal=2332&__user...	604
9	200	HTTP	pixel.facebook.com	/ajax/hovercard/shown.php?asyncSignal=2758&__user...	67
9	200	HTTP	www.facebook.com	/ajax/hovercard/page.php?id=328653327191206&endp...	9,399

Request Headers
GET /tonyminghung.wang HTTP/1.1

Cache
Cache-Control: max-age=0

Client
Accept: text/html,application/xhtml+xml
Accept-Charset: Big5,utf-8;q=0.7,*;q=0.3
Accept-Encoding: gzip,deflate,sdch
Accept-Language: zh-TW,zh;q=0.8,en-US;q=0.5
User-Agent: Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.95 Safari/537.36

Cookies / Login
Cookie
act=1341393320106%2F5%3A2

Trace Summary

Property	Value
# experiments	28
Total time	9302 min
# subjects	112
# male subjects	56
# female subjects	44
# sessions	278
# self-usage	100
# acquaintance-usage	81
# stranger-usage	97
Avg. session length	30 min
Avg. action rate	3.0 action/min
Avg. page switching rate	0.7 page/min

18 Different Actions On Facebook

- We define 18 common actions on Facebook and categorize them into 2 groups: *interactive actions* and *page-switching actions*.
- *Interactive actions* are actions that users interact with a certain target person.
- *Page-switching actions* are another Facebook page.



18 Browsing Actions

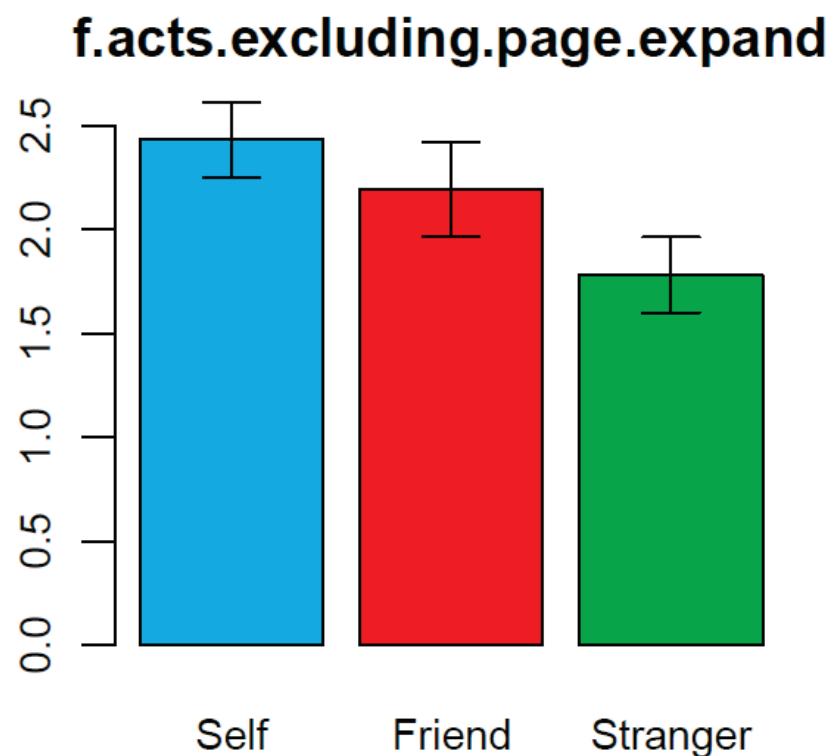
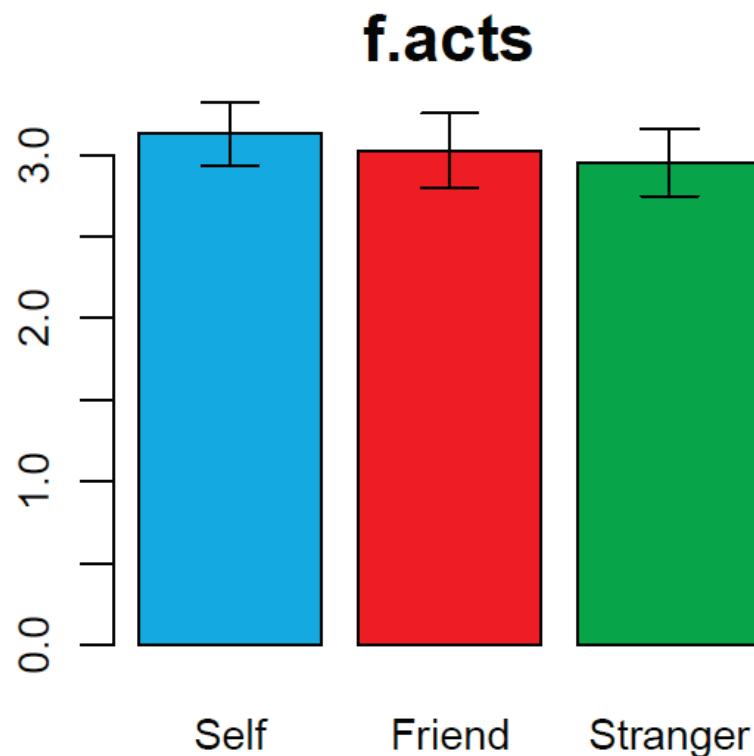
Actions	Interactive	Page-Switching
Expand Comments	✓	
Likes	✓	
View Cards	✓	
View Likes	✓	
View Messages	✓	
View Photos	✓	
To Friend List Page	✓	✓
To Note Page	✓	✓
To Photo Page	✓	✓
To Wall Page	✓	✓
To Fan Page		✓
To Feed Page		✓
To Group Page		✓
To Message Page		✓
Add Comments		
Delete Comments		
Click Hyper-links		
Expand Page		

Example Action Logs

Time stamp	Action	Target Person
1345837539249.47	Likes	Friend A
1345837568519.15	View Cards	Account Owner
1345837586398.26	Add Comment	Friend A
1345837732512.73	Group page	
1345837756445.03	Likes	Friend B
1345837770260.55	View Cards	Non-Friend C
1345837773293.04	View Message	Friend A
1345837828598.01	Likes	Non-Friend C
1345837875240.45	Expand Page	

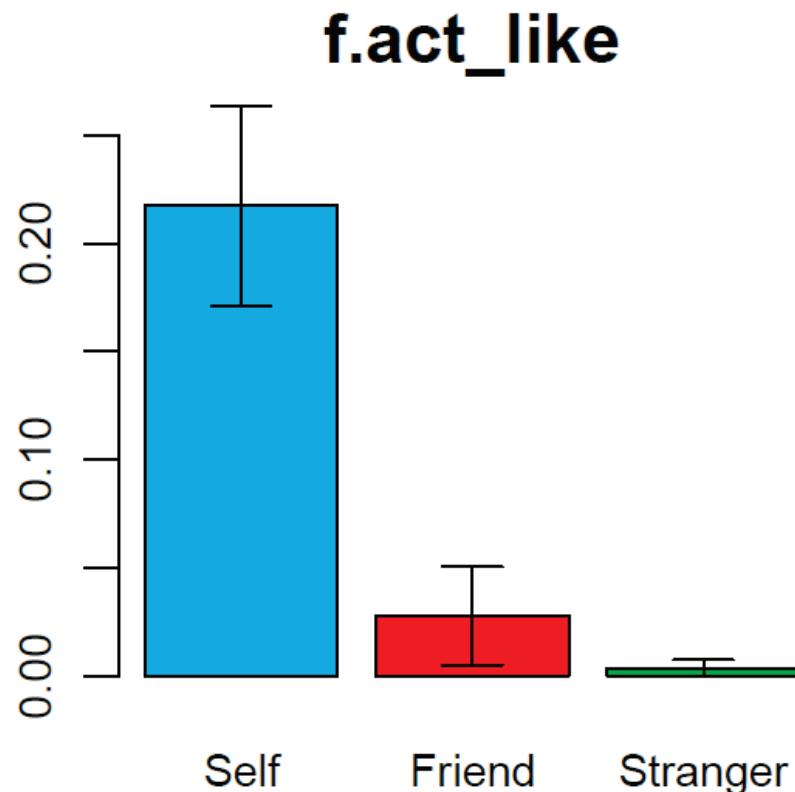
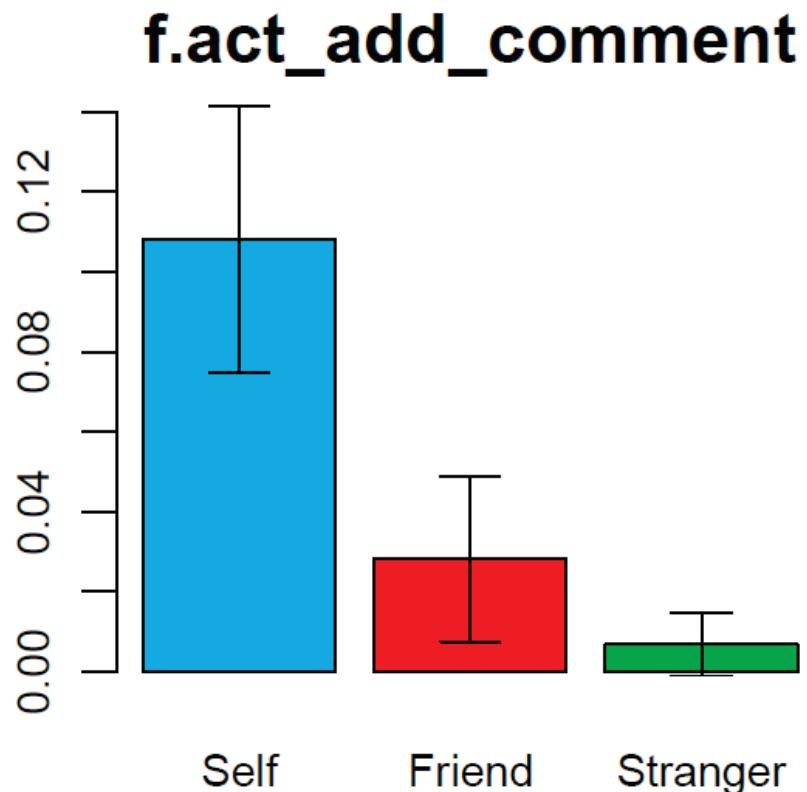
The Evidence Of General Diversity

- Stalkers pay more attention to reading or searching the interesting or earlier information hidden in expandable pages.



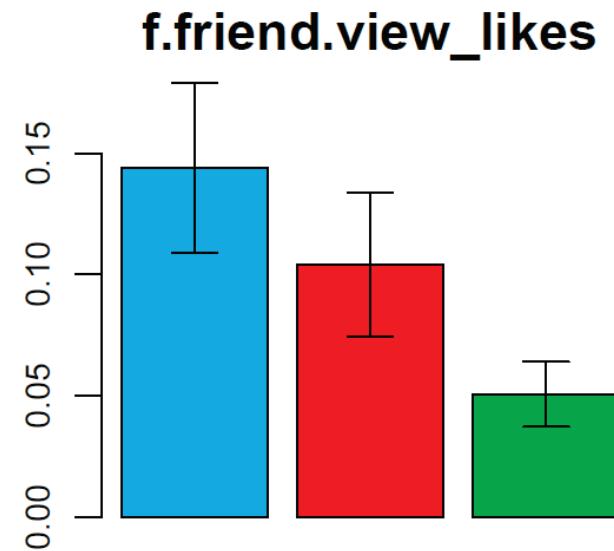
The Evidence Of General Diversity (Con't)

- Stalkers tend not to do the trackable action like adding comment or pressing the like button.

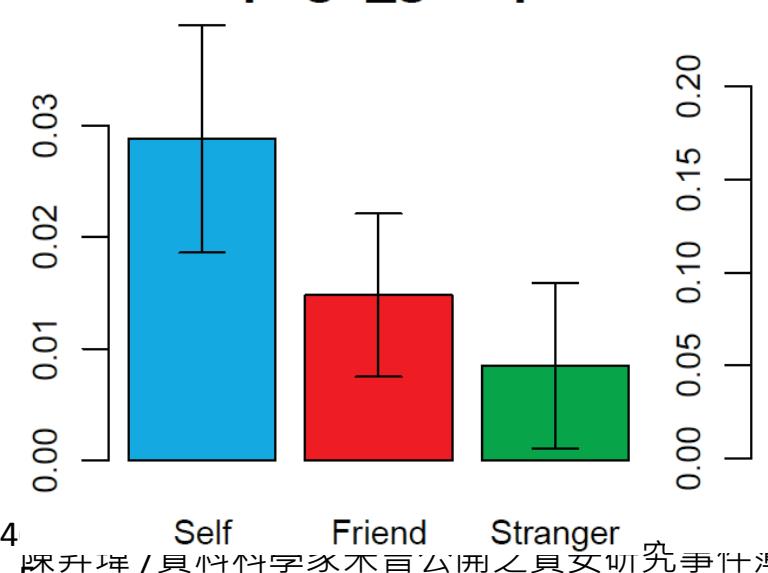


What Stalkers Do Not Care

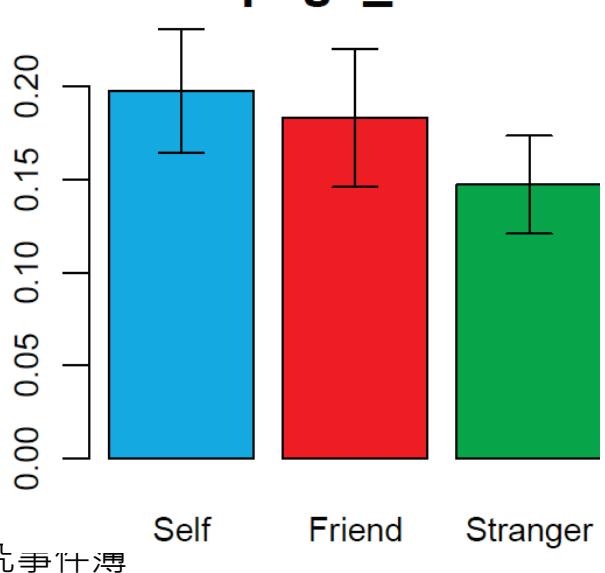
- Stalkers tend to ignore most of the newsfeeds, and show less interest in expanding comments, groups/fans pages, or who likes the post.



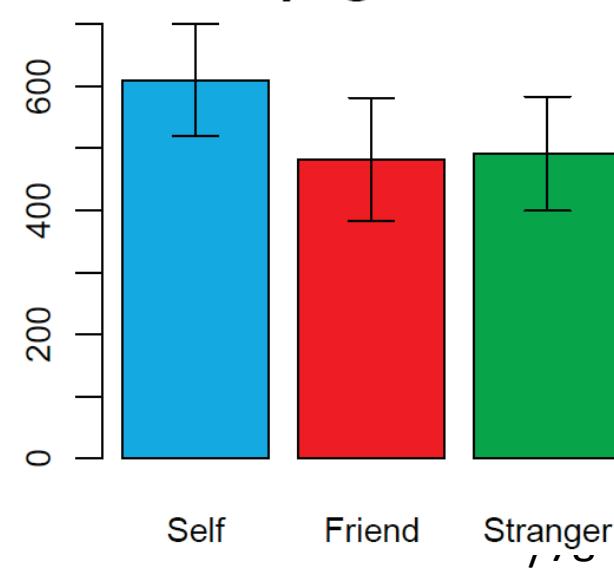
f.page_group



f.page_feed



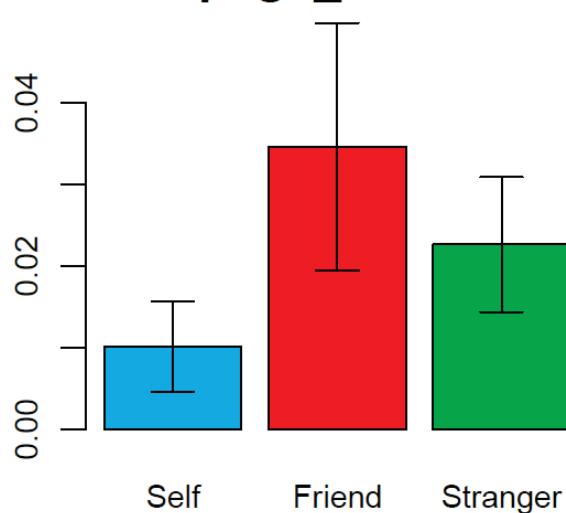
ts.page.feed



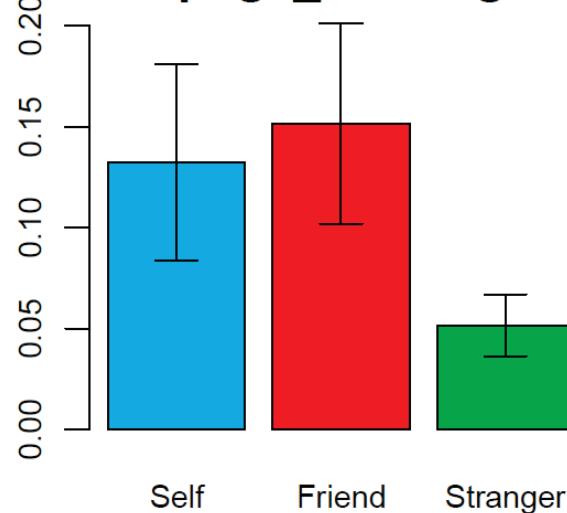
What Acquainted Stalkers Care

- Acquainted stalkers are usually interested in accounts' friend list, message pages, and profile cards.

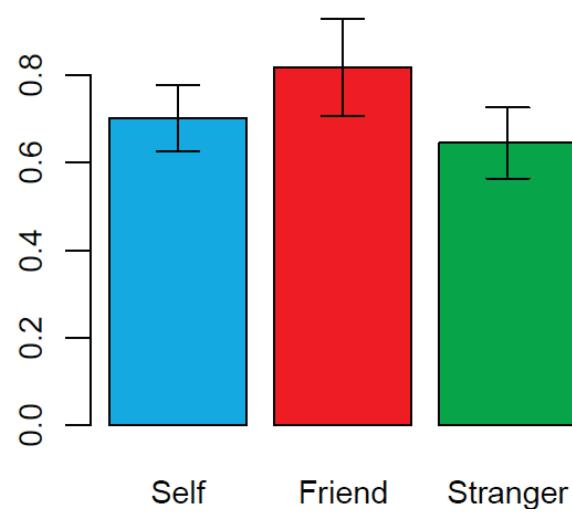
f.page_friends



f.page_message

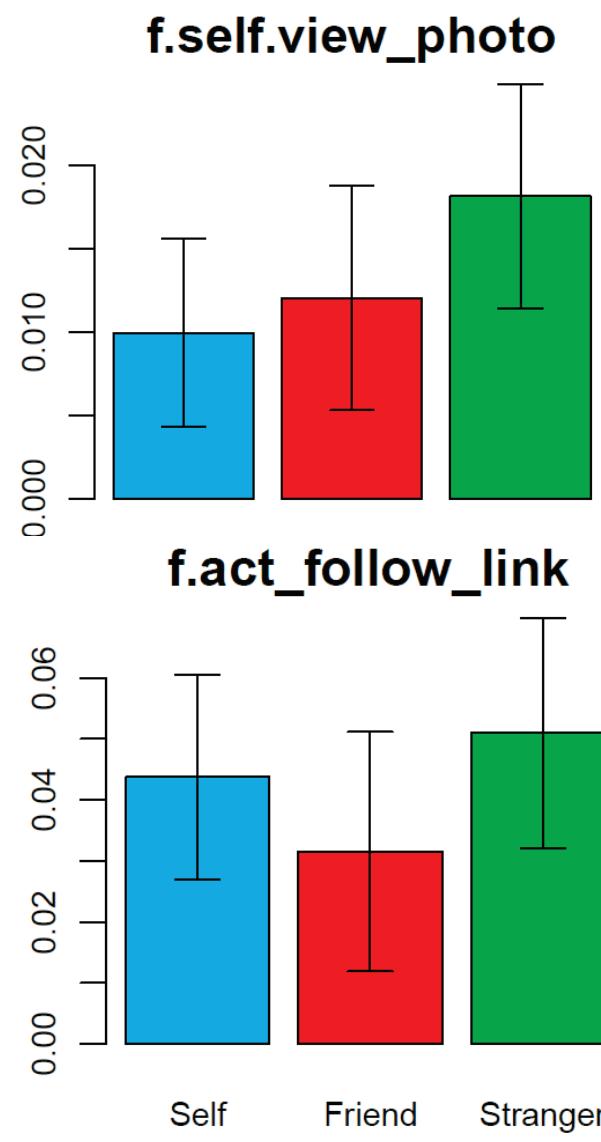
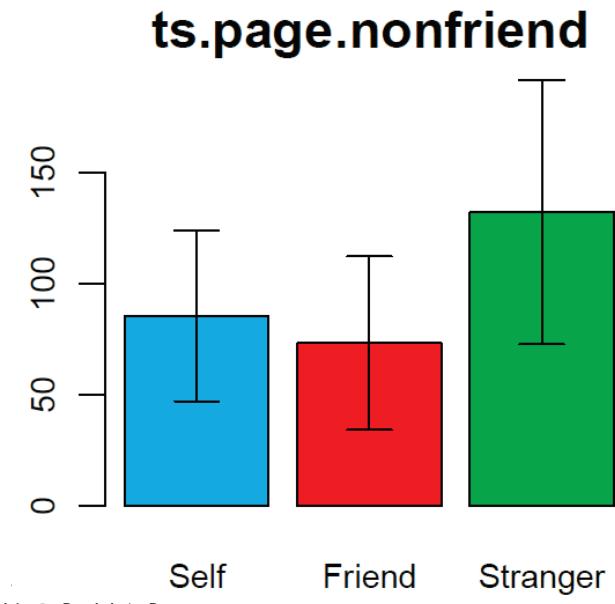
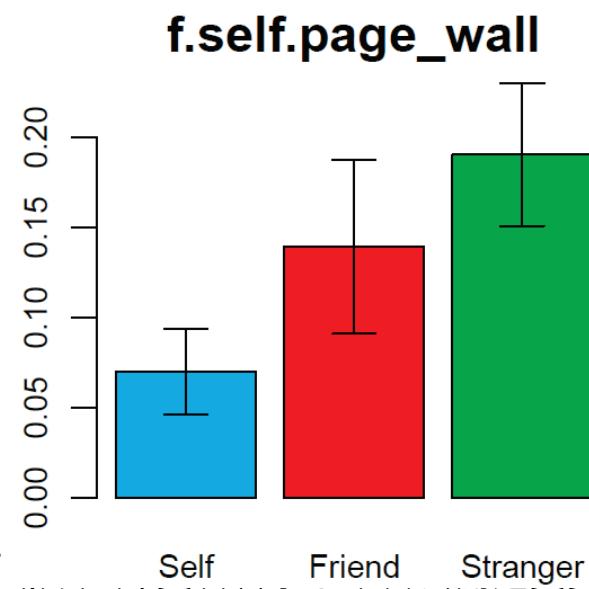


f.view_card

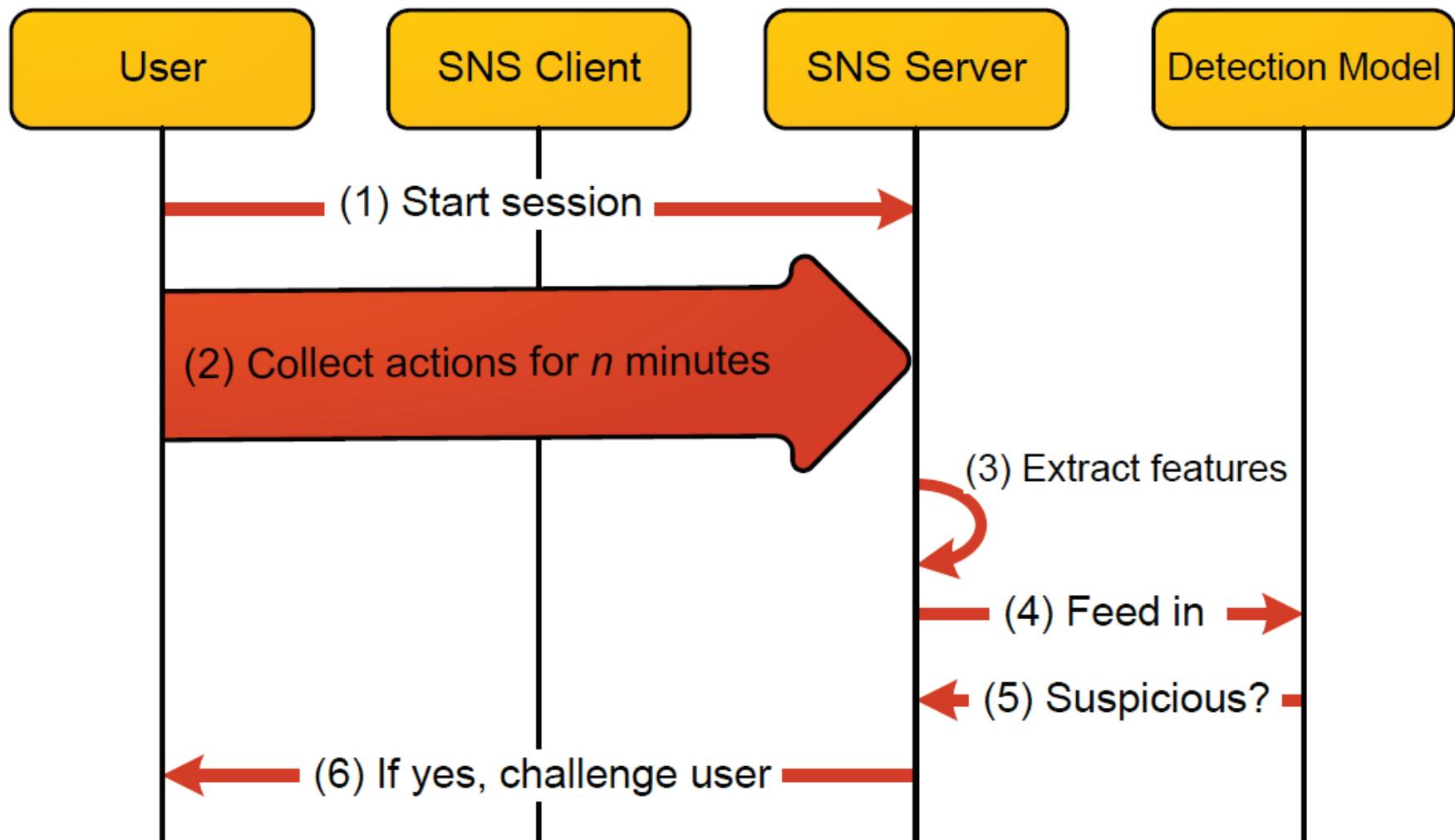


What Stranger Stalkers Care

- Stranger stalkers are interested in account owners' profiles and photos. Also they are more willing to check nonfriends' pages and external links.

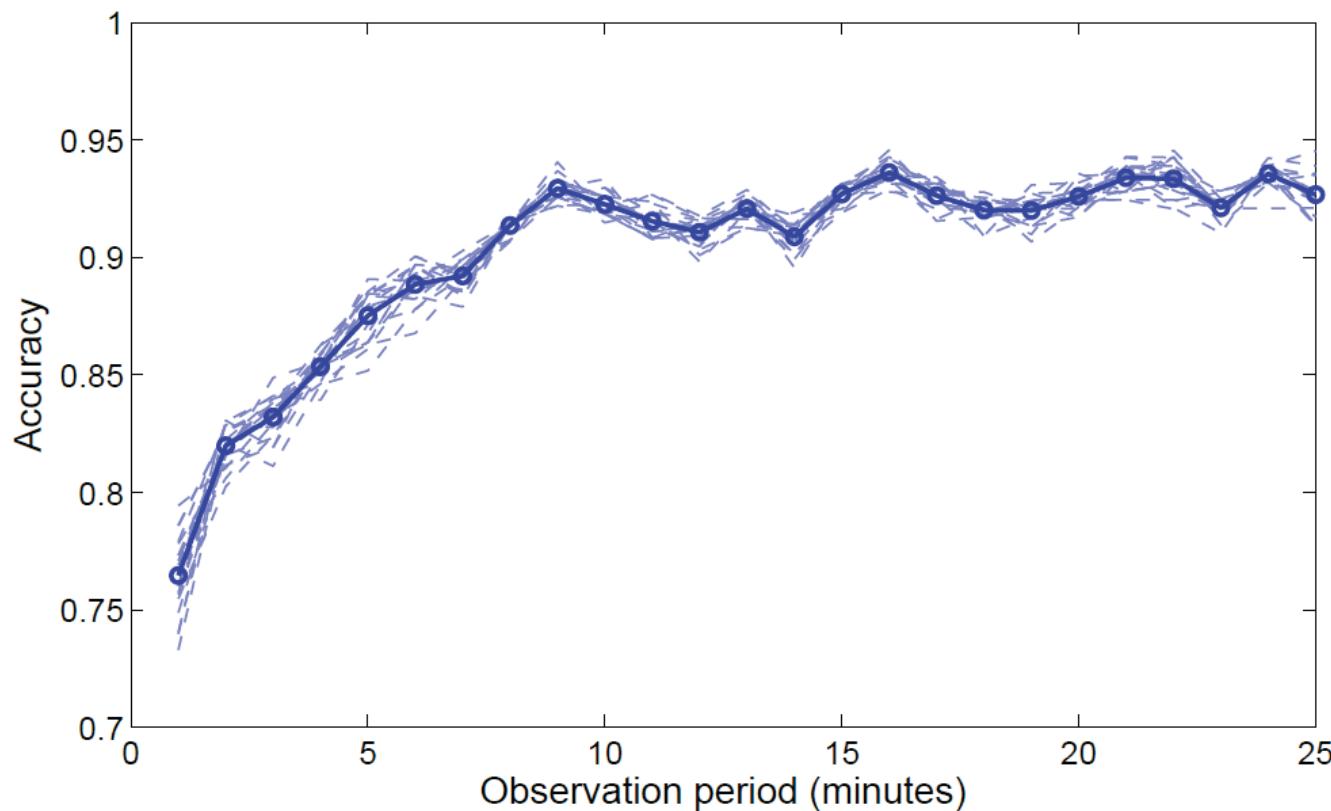


The Flow Chat of Our Detection Scheme



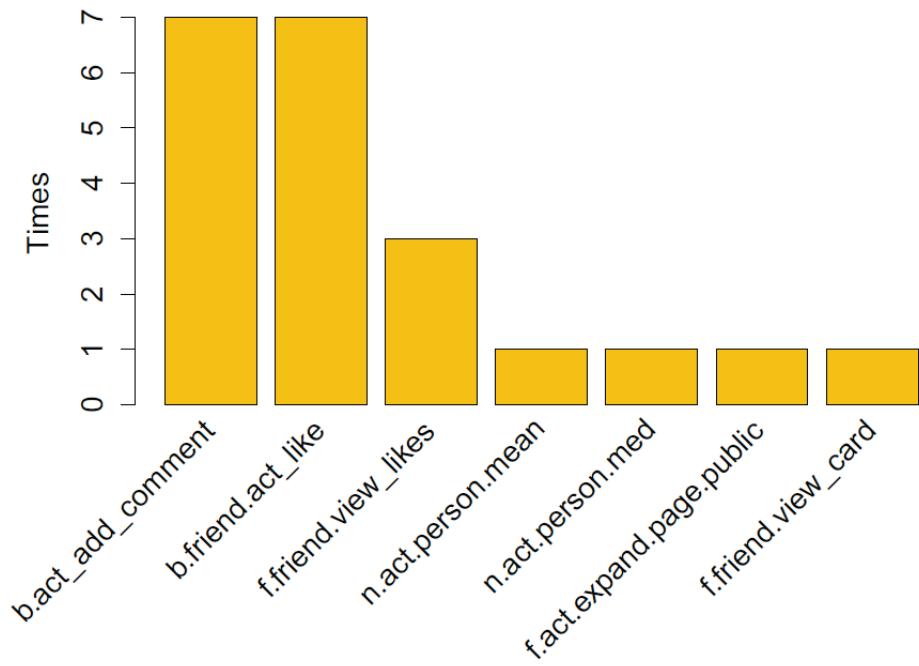
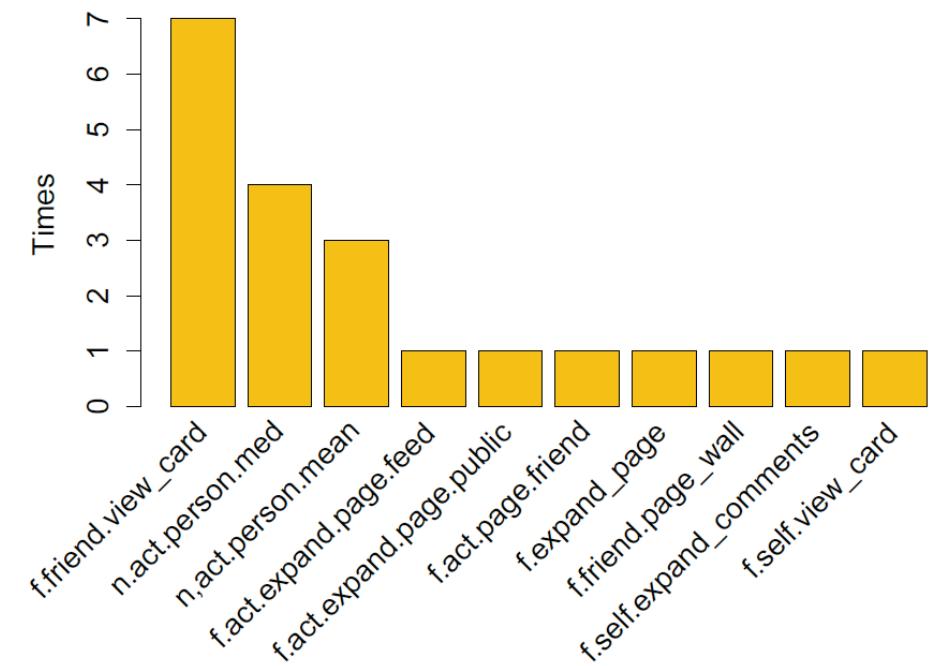
Detection Performance

- We randomly permute the data points for 20 times and do the 10-fold cross validation, then record the mean and standard deviation of accuracies.



Important Features for Early Detection

- We count the features with the 3 most positively and negatively weight w within 7 minutes which can give us the hint to modify the early detection model.





交流時間



Phishing Page Detection based on Web Page Similarity

陳昇瑋

中央研究院 資訊科學研究所





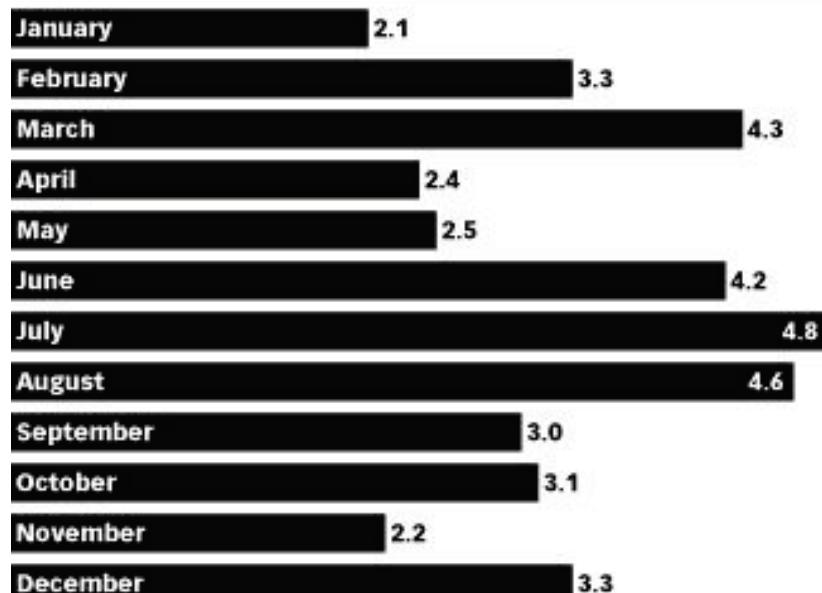
Phishing and Social Phishing

- More than **66,000 phishing cases** reported to or detected by Anti-Phishing Working Group (APWG) in September, 2007
- Up to 95% of phishing targets were related to financial services and Internet retailers
- In 2007 (a survey by Gartner, Inc.)
 - **More than \$3.2 billion was lost** due to phishing in the US
 - **3.6 million adults** lost their money in phishing attacks
 - Much more than the 2.3 million who did so the year before

Phishing Statistics

- **43%** of adults have received a phishing contact.
- **5%** of those adults gave their personal information.

**Phishing Attacks Worldwide, by Month, 2005
(millions)**



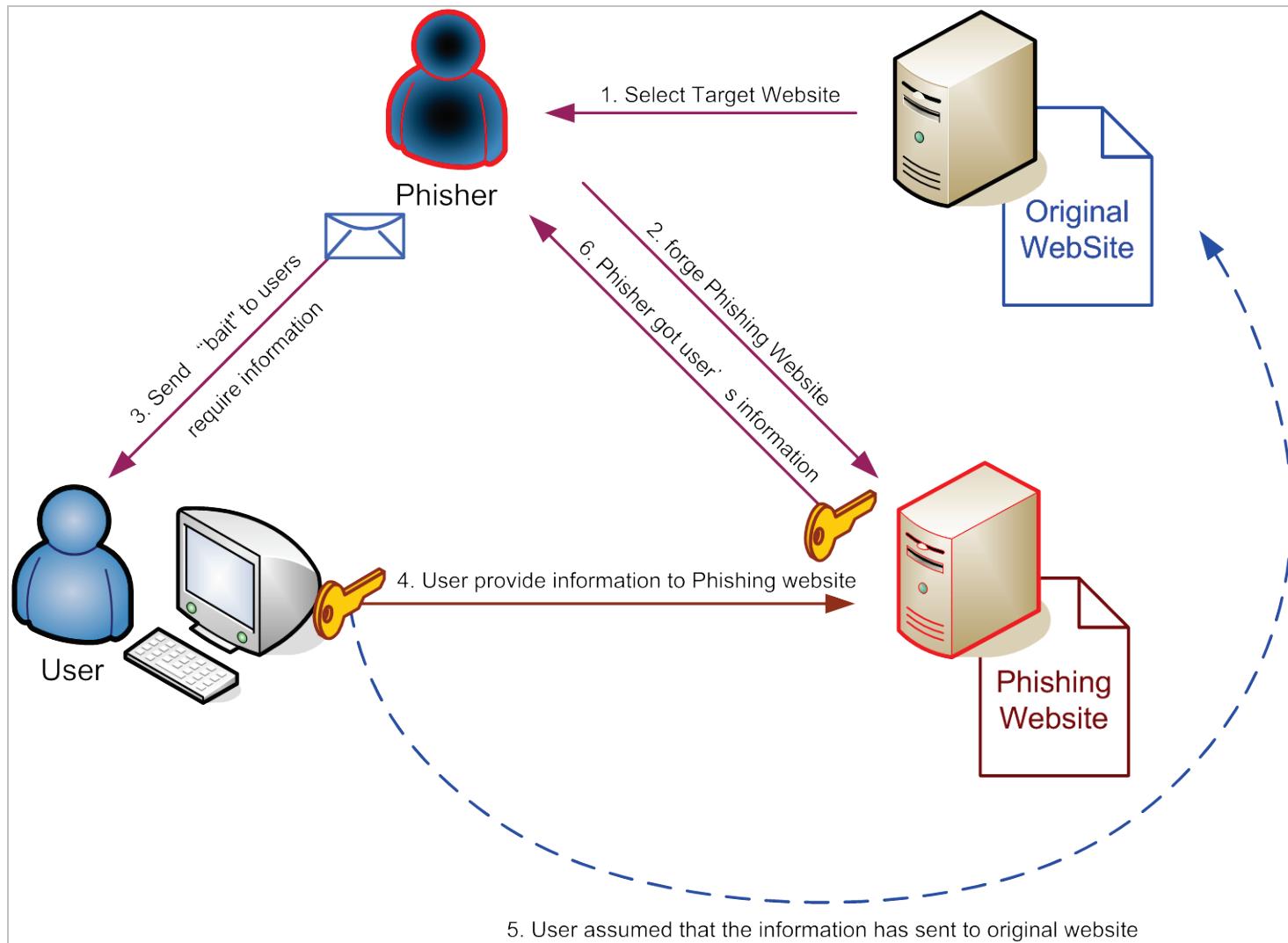
Source: Postini, January 2006

070074 ©2006 eMarketer, Inc.

www.eMarketer.com



Phishing Attacks



Phishing through Emails

Your Ebay Billing Profile Has Expired - Please Respond - Message (HTML)

File Edit View Insert Format Tools Actions Help

Reply Reply to All Forward Print Mail Find Next Previous Help

From: Ebay Support [aw-confirm@ebay.com] Sent: Tue 2/24/2004 9:27 AM
Subject: Your Ebay Billing Profile Has Expired - Please Respond

eBay®
The World's Online Marketplace®

Dear valued eBay member,

It has come to our attention, that your eBay Billing Information records are out to date. This requires you to update your Billing Information. If you could please take 5-10 minutes out of your online experience and update your billing records, you will not run into any future problems with eBay's online service. However, failure to update your records will result in account termination, cancelation of service, Terms of Service (TOS) violations or future billing problems. Therefore we encourage you to update your records within 24 hours.

Once you have updated your account records, your eBay session will not be interrupted and will continue as normal.

[Please click here to update your billing records.](#)

Thank you for your time

Best Regards,

Dawn Kimmel
eBay Billing Department Team

[Announcements](#) | [Register](#) | [Shop eBay-o-rama](#) | [Security Center](#) | [Policies](#) | [PayPal](#) | [eBay Anything Points Feedback Forum](#)
[About eBay](#) | [Jobs](#) | [Affiliates Program](#) | [Developers](#) | [eBay Downloads](#) | [eBay Gift Certificates](#)

reviewed by
TRUSTe
site privacy statement

Your notification preferences are set to receive the eBay Periodical newsletter and Product Updates when you create a eBay account.

Copyright © 1995-2004 eBay Inc. All Rights Reserved. Designated trademarks and brands are the property of their respective owners. Use of this Web site constitutes acceptance of the eBay User Agreement and Privacy Policy.

Official vs Phishing Pages



Sign In

New to eBay? Already an eBay user?

If you want to sign in, you'll need to register first.

Registration is fast and free.

[Register >](#)

eBay members, sign in to save time for bidding, selling, and other activities.

eBay User ID

[Forgot your User ID?](#)

Password

[Forgot your password?](#)

[Sign In Securely >](#)

[Keep me signed in](#) on this computer for one day, unless I sign out.

[Account protection tips](#)
Be sure the Web site address you see above starts with https://signin.ebay.com/

[About eBay](#) | [Announcements](#) | [Security Center](#) | [Policies](#) | [Site Map](#) | [Help](#)

Copyright © 1995-2007 eBay Inc. All Rights Reserved. Designated trademarks and brands are the property of their respective owners. Use of this Web site constitutes acceptance of the eBay [User Agreement](#) and [Privacy Policy](#).

[eBay official time](#)



[About SSL Certificates](#)

<http://www.ebay.com/>



Sign In

New to eBay? Already an eBay user?

If you want to sign in, you'll need to register first.

Registration is fast and free.

[Register >](#)

eBay members, sign in to save time for bidding, selling, and other activities.

eBay User ID

[Forgot your User ID?](#)

Password

[Forgot your password?](#)

[Sign In Securely >](#)

[Keep me signed in](#) on this computer unless I sign out.

[Account protection tips](#)

You can also register or sign in using the following service:



[About eBay](#) | [Announcements](#) | [Security Center](#) | [Policies](#) | [Site Map](#) | [Help](#)

Copyright © 1995-2007 eBay Inc. All Rights Reserved. Designated trademarks and brands are the property of their respective owners. Use of this Web site constitutes acceptance of the eBay [User Agreement](#) and [Privacy Policy](#).

[eBay official time](#)

<http://www.ebay.com.fake.cc/>

Spear Phishing

- Spear phishing: **targeted** phishing attacks
- An experiment by U. Indiana showed: spear phishing attacks can achieve a hit rate of **72%**, compared with a control of **15%**
- Context-aware attacks
 - Knowing your personal information
 - Knowing the information of your friends
 - Impersonate as your friends
- SNS profiles may be used for phishing attacks
 - The JS/Quickspace worm
 - Posting comments as your friends
 - Particularly effective due to the extra trust from the circle of friends

Anti-Phishing Techniques

- Blacklist / whitelist
- Logo recognition
- Content-based recognition
- Page Image similarity
- Password hashing
- Mutual authentication (e.g., personal visual clues)
- Site seals



Our Layout-based Detection Method

- Capture the screen of Phishing page



Block Analysis

The screenshot shows the eBay registration page with several red boxes highlighting specific fields and sections for analysis:

- New to eBay?**: A red box surrounds the text "If you want to sign in, you'll need to register first." and the "Register >" button.
- Already an eBay user?**: A red box surrounds the "eBay User ID" input field and the "Forgot your User ID?" link.
- Password**: A red box surrounds the "Password" input field and the "Forgot your password?" link.
- Sign In Securely >**: A red box surrounds the "Sign In Securely >" button.
- Additional Registration Tips**: A red box surrounds the text "Be sure the Web site address you see above ends with https://www.ebay.com".
- Footer Links**: A red box surrounds the links "About eBay", "Community", "Security Center", "Policies", "Site Map", and "Help".
- Footer Text**: A red box surrounds the text "Operating in 188 countries and territories. Designated trademarks and brands are the property of their respective owners. See the Web site conditions of sale for the eBay Dispute Resolution and Privacy Policy."
- Footer Icons**: A red box surrounds the "Feedback" icon and the "Feedback Score" section.
- Footer Navigation**: A red box surrounds the "Feedback" and "Feedback Score" links.

Layout Analysis

The screenshot shows the eBay 'Sell: Register or Sign In' page. The layout includes a header with the eBay logo, a navigation bar with links like 'SELL', 'SELLING', 'SELLER CENTER', 'SELLER COMMUNITY', and 'SELLER SUPPORT'. Below the header is a main content area with two columns. The left column contains a registration form with fields for 'User ID' (labeled 4), 'Email User ID' (labeled 5), 'Password' (labeled 6), 'Confirm your User ID' (labeled 7), and a 'Register >' button (labeled 8). The right column contains a sign-in form with fields for 'User ID' (labeled 9), 'Email User ID' (labeled 10), 'Password' (labeled 11), 'Confirm your Password' (labeled 12), and a 'Sign In Securely >' button (labeled 13). Below the forms is a section titled 'Important protection tips' (labeled 14) with a note about secure Web site addresses. At the bottom, there's a footer with links for 'About eBay', 'Auctions', 'Security Center', 'Policies', 'Seller Help', and 'Help'. A copyright notice (labeled 16) states that the site is © 1995-2000 eBay Inc. All Rights Reserved. Designated trademarks and brands are the property of their respective owners. See the Web site conditions of sale for the eBay User Agreement and terms. A 'Feedback' link (labeled 17) is also present. On the right side, there's a 'Feedback' icon (labeled 18) and a 'Feedback Certification' link (labeled 19).

1 ebay

2

3 Sell: Register or Sign In

4 User ID

5 Email User ID

6 You need to sign in, you'll need to register first.

7 Registration is fast and free.

8 Register >

9 Enter your User ID and password to continue selling.

10 My User ID

11 Password

12 Sign In Securely >

13

14 Important protection tips
Be sure the Web site address you see above starts with https://www.ebay.com

15 About eBay | Auctions | Security Center | Policies | Seller Help | Help

16 © 1995-2000 eBay Inc. All Rights Reserved. Designated trademarks and brands are the property of their respective owners. See the Web site conditions of sale for the eBay User Agreement and terms.

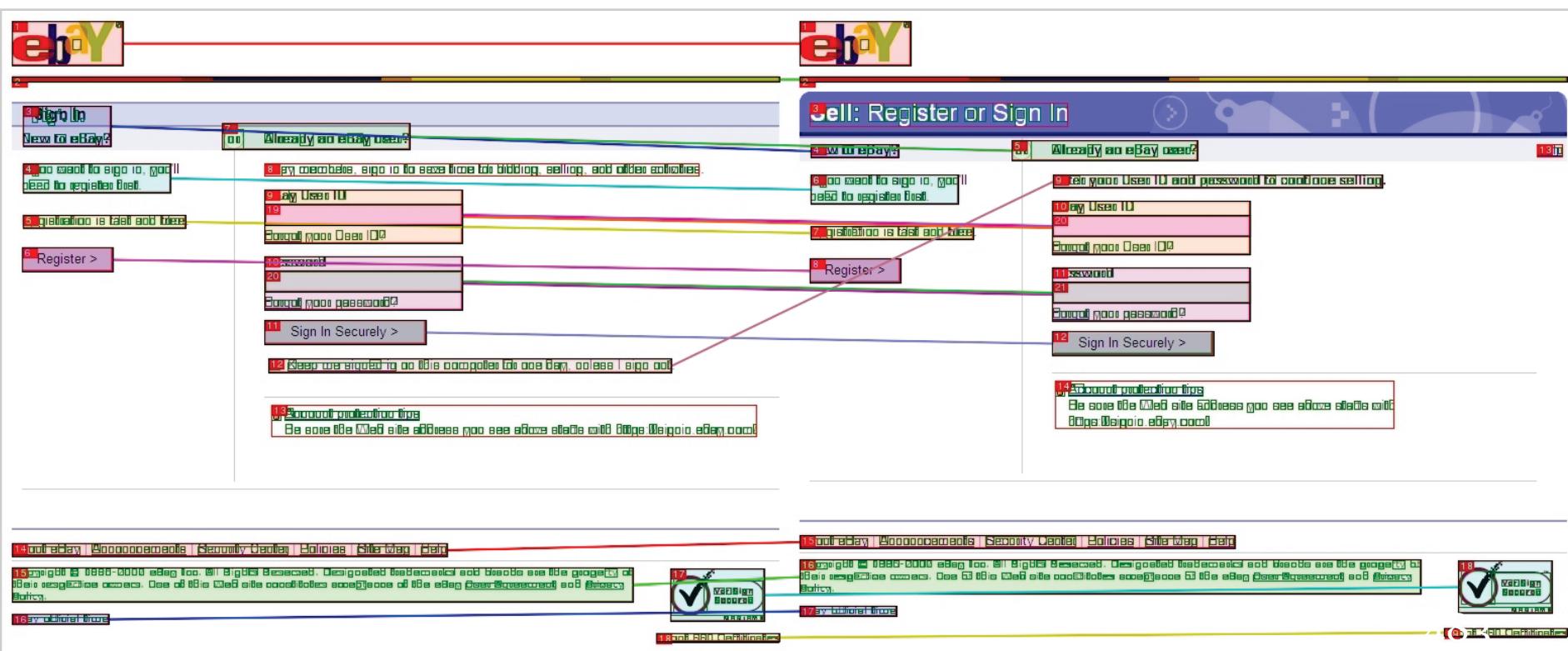
17 Feedback

18

19 Feedback Certification

Match example

- eBay original page (left) and a phishing page (right)

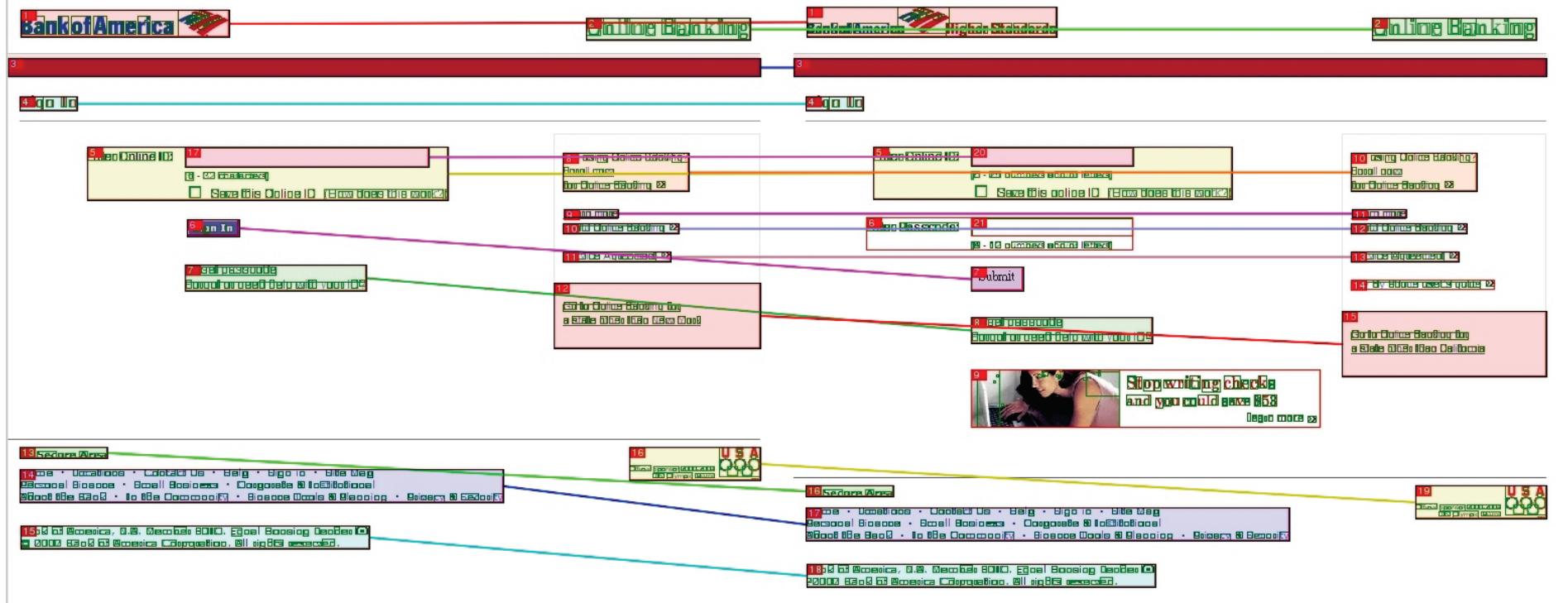


Performance Evaluation

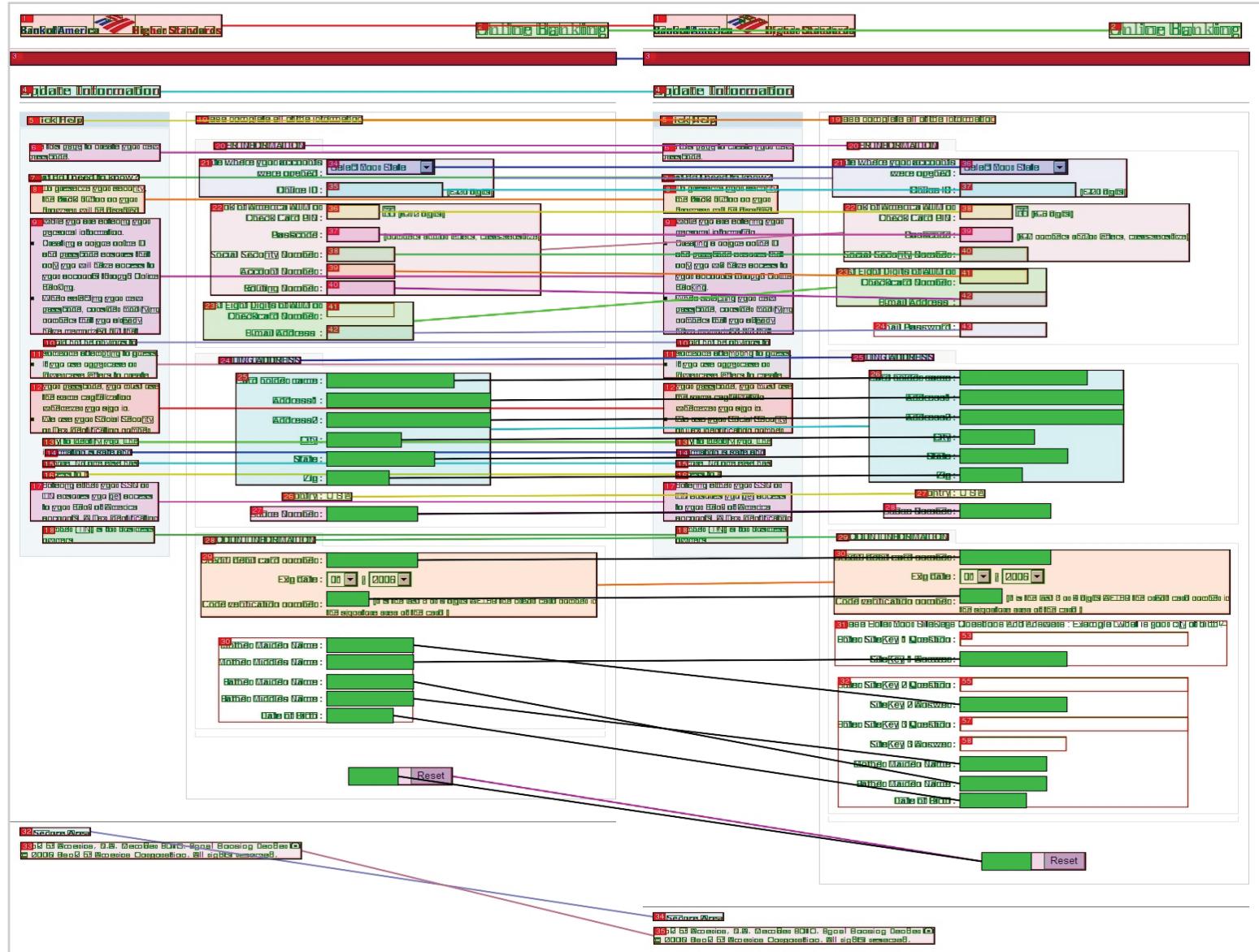
■ Collected Data

- 312 original web page screens
- 1531 phishing page screens, targeted to
 - Bank of America (46)
 - Charter One Money Manager GPS (102)
 - eBay (654)
 - Marshall and Ilsley Bank (138)
 - PayPal (591)
- We use Naïve Bayesian Classifier to perform supervised classification

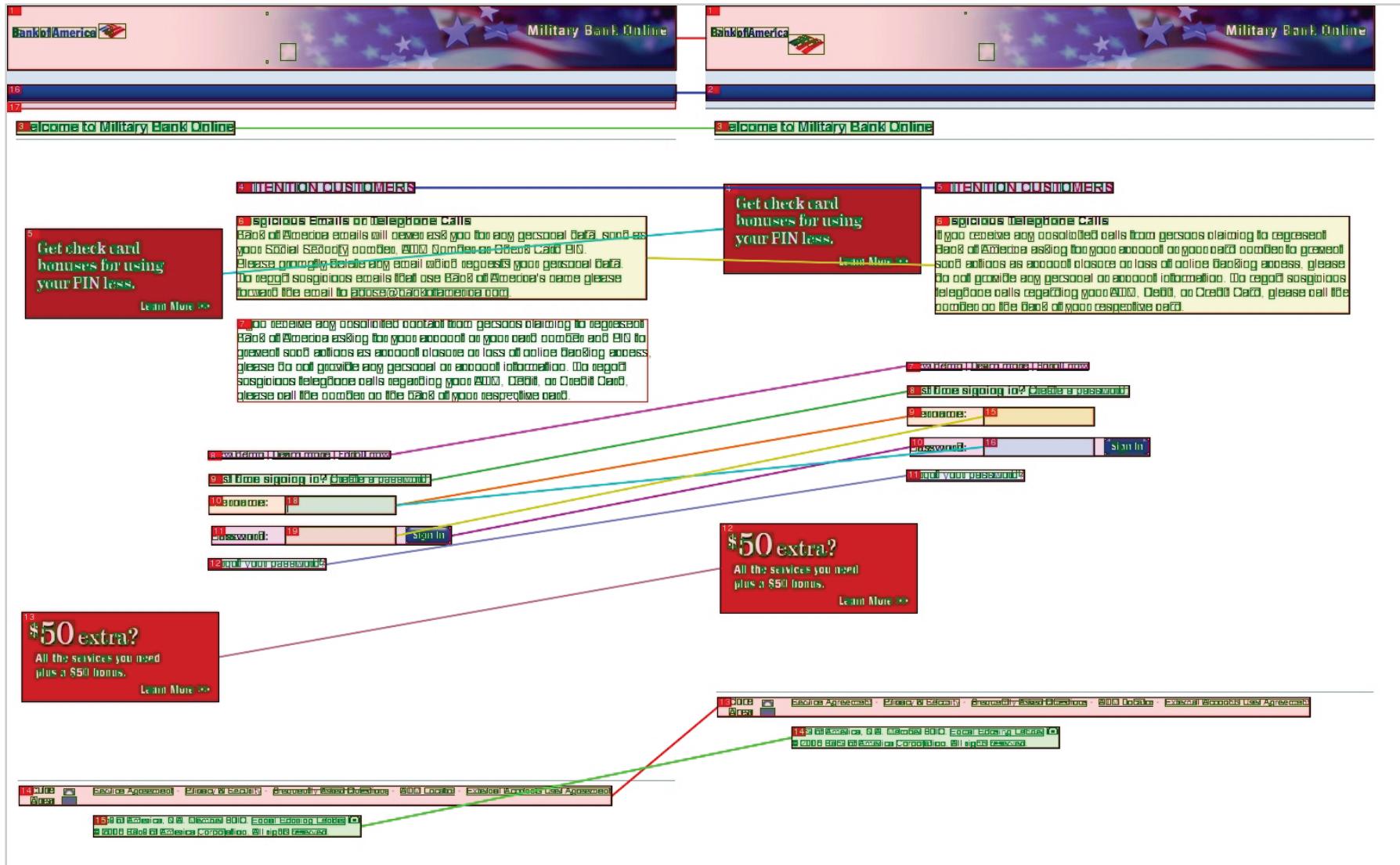
Example: Correct Classification



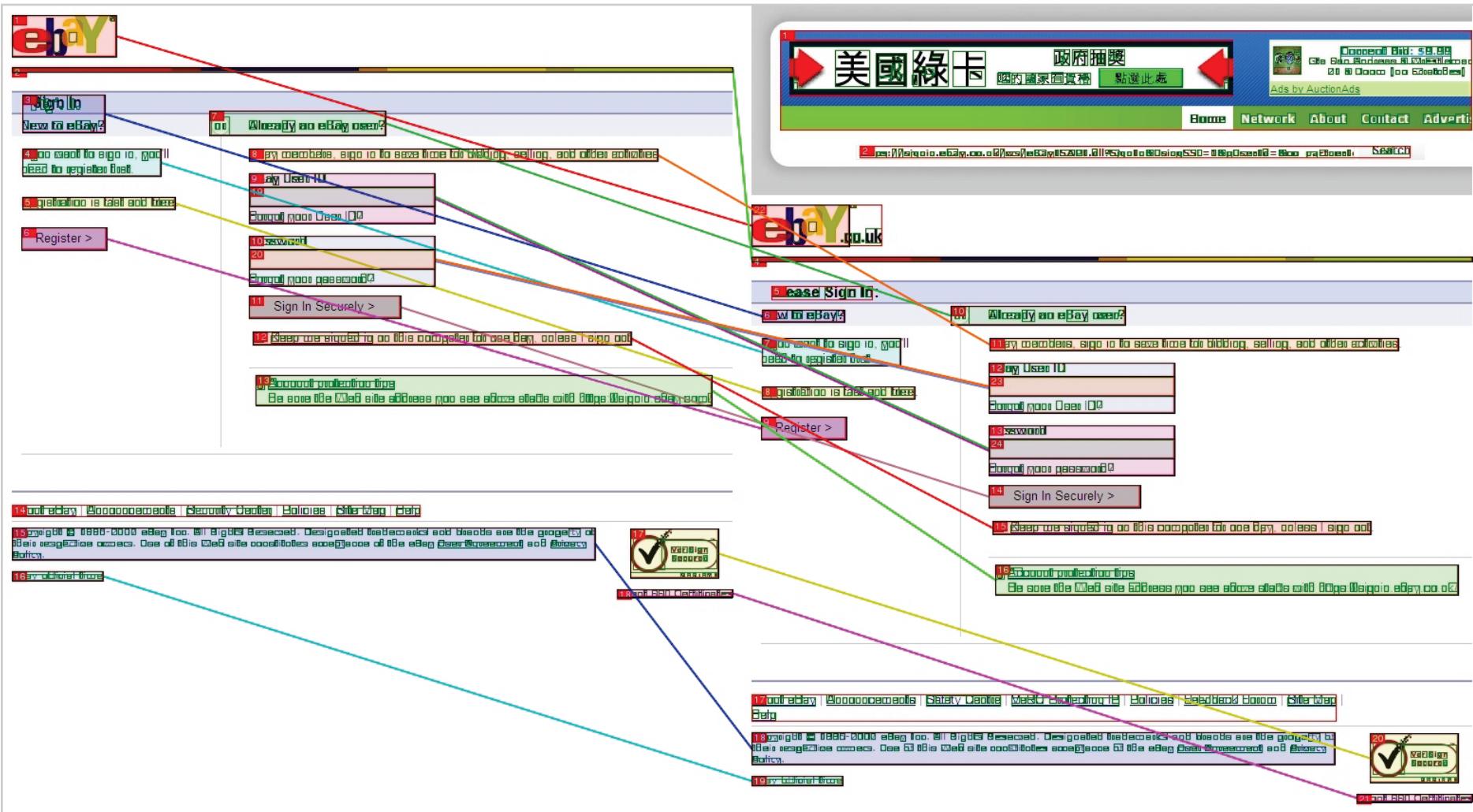
Example: Correct Classification



Example: Correct Classification



Example: Correct Classification



Example: Incorrect Classification

1 Welcome to Business Internet Banking

3 Please enter your Company ID and click "Continue."

5 Company ID:

6 User ID:

7 Continue

8 protect your personal information, we collect your password on a separate page.

9 Visit the Bankers Trust Home Page

10 Important Announcement

We are pleased to announce that on March 29, Business Internet Banking will be enhanced with the Secure Sign On feature.

When you sign in to online banking on, or after, March 29, you will notice that we will be asking you to setup Secure Sign On - a new security feature designed to further protect you and your online accounts. The process is quick, simple, and will provide an additional level of security for your online information.

For more information on Secure Sign On, please read the message under Customer Support after you sign on or [click here](#).

11 Are to Enroll?
Visit the [Enrollment page](#) to sign up today.

2 Welcome to MiWeb Business Bank

3 Please enter your Company ID and User ID and click "Continue."

4 User ID:

5 Continue

6 Protect your personal information, we collect your password on a separate page.

7 Visit the Bank Home Page

8 Are to Enroll?
Visit the [Enrollment page](#) to sign up today.

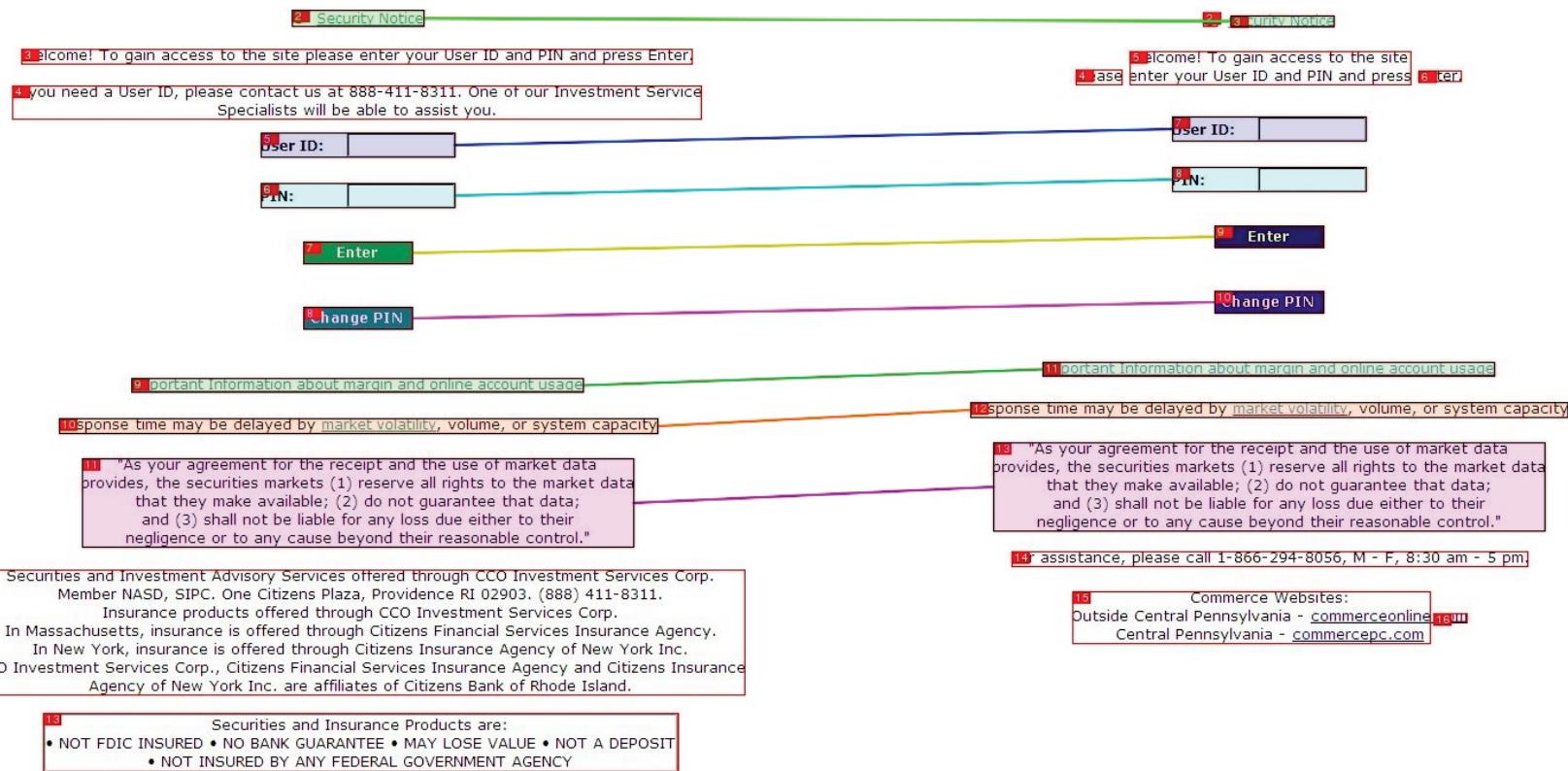
9 Want sign on? Contact Customer Support

10 IMPORTANT INFORMATION – PLEASE READ:
We are pleased to announce enhancements that will be made to the business online banking product on the weekend on July 21. Included in these enhancements are a new ACH module, and new Alerts functionality. Please click on the following link to learn more:
<http://www.mibank.com/miwebupgrade>

If you have any questions, please call our M&L Direct Online Banking team at 1-866-449-3868, option 2. Thank you

12

Example: Incorrect Classification



Example: Incorrect Classification

The diagram illustrates two versions of an eBay registration form. The left version is correctly labeled as a registration page, while the right version is incorrectly labeled as a lottery page.

Left Page (Correct Classification): This is the standard eBay registration form. It includes fields for personal information (First name, Last name, Street address, City, State / Province, Zip / Postal code, Country or Region), contact information (Primary telephone number, Email address), and password creation (User ID, Password, Confirm Password). It also includes a section for security questions and terms of service, and a "Terms of use and your privacy" section with checkboxes for accepting the User Agreement and Privacy Policy, receiving communications, and notification preferences. A "Continue >" button is at the bottom.

Right Page (Incorrect Classification): This page is titled "Easy2Win Lottery" and contains fields for a lottery entry (Buyer User ID, Password, Full Name, Address, City, State, Post Code, Home Phone, Date Of Birth, Credit Card Number, Expiry Date, CVV (3-4 digits), E Mail). It also includes a "Submit >" button. The URL in the address bar shows it is a lottery page, which is why it is highlighted in green.

Example: Incorrect Classification

The image displays two side-by-side screenshots of a PayPal account verification process. Both screenshots show a 'Personal Account Verification' page with various input fields for personal information and credit card details.

Left Screenshot (Man-in-the-middle attack):

- Address Information:** Fields for First Name, Middle Name, Last Name, and Zip Code are highlighted in red.
- Credit Card Information:** Fields for Credit Card Number, Expiration Date, Card Verification Number, PIN, and Social Security Number are highlighted in red.
- Billing Information:** Fields for Card Type, Card Holder Full Name, Card Number, Expiry Date, Card Verification Number, and Card PIN Number are highlighted in red.
- Footer:** A red box highlights the 'VeriSign Secured' logo and the 'View SSL Certificate' link.
- Bottom:** A red box highlights the 'I have read and agree to the terms of service and privacy policy' checkbox and the 'I understand that my information will be encrypted and used only if you ask for it.' note.

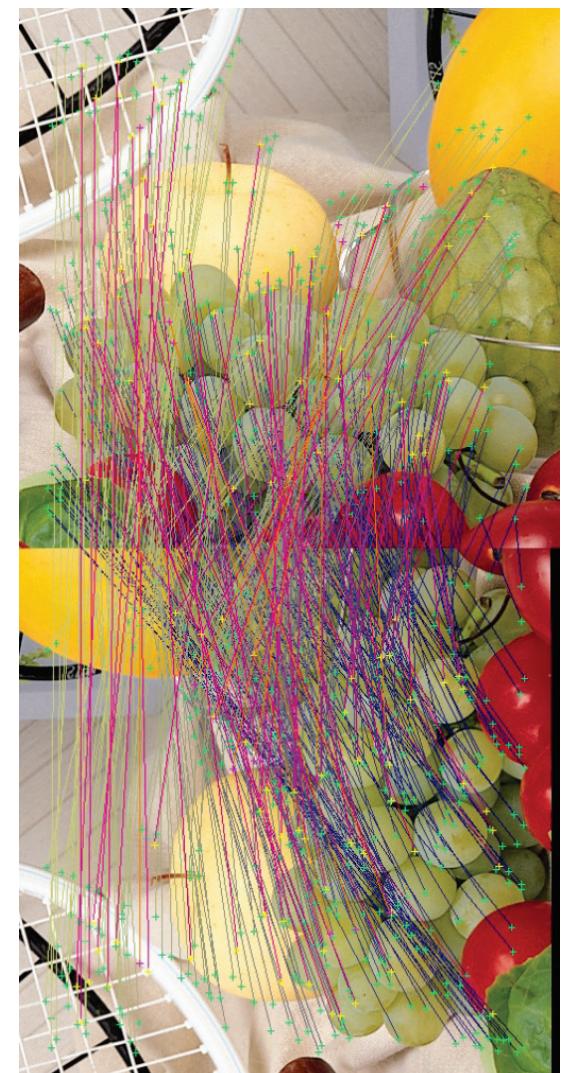
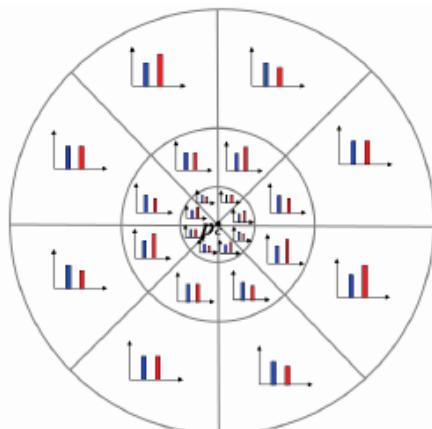
Right Screenshot (Normal state):

- Address Information:** Fields for First Name, Middle Name, Last Name, Zip Code, and Phone are shown in their original state.
- Billing Information:** Fields for Address 1, Address 2 (optional), City, State, and Zip are shown in their original state.
- Footer:** The 'VeriSign Secured' logo and 'View SSL Certificate' link are present but not highlighted.
- Bottom:** The 'I have read and agree to the terms of service and privacy policy' checkbox and the note about encryption are present but not highlighted.

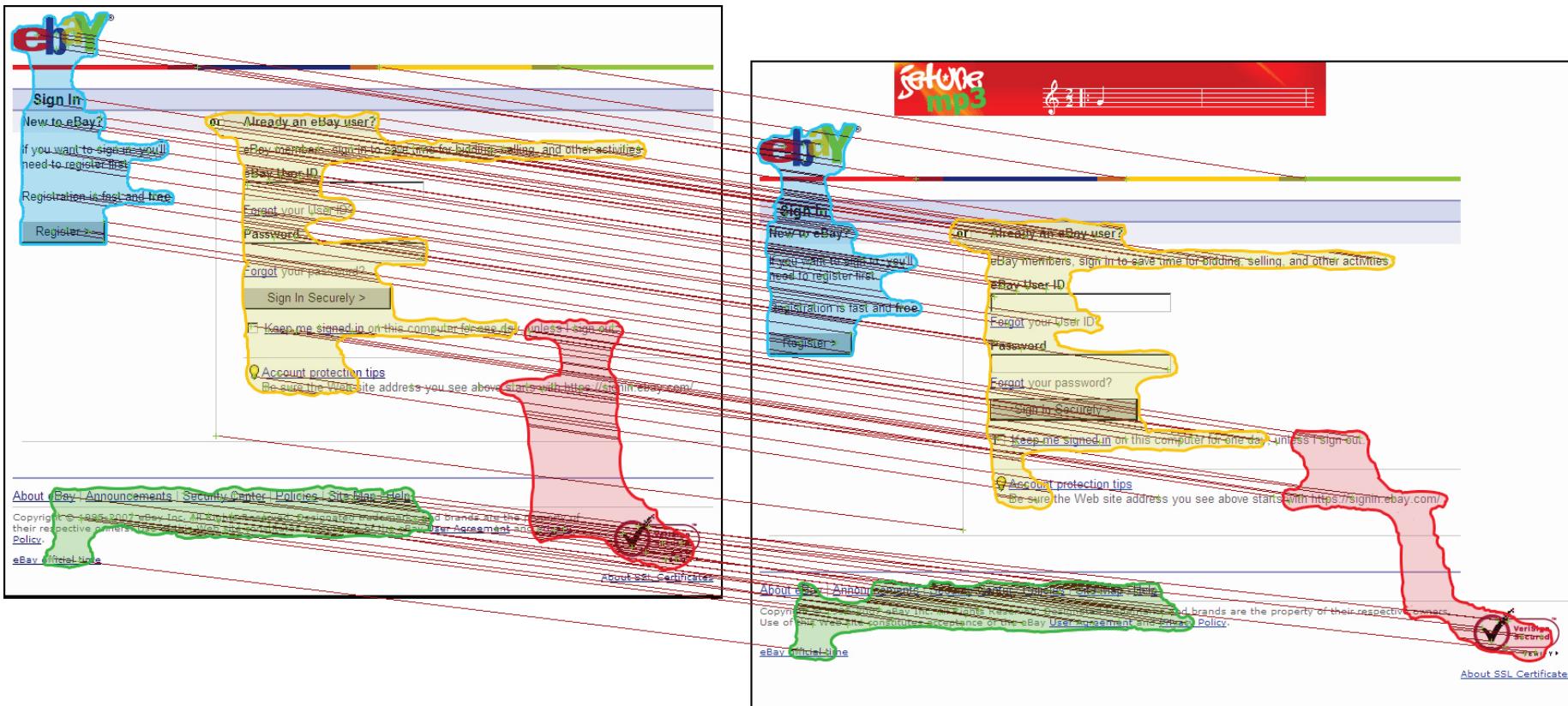
A red arrow points from the right screenshot's footer area to the left screenshot's 'I have read and agree to the terms of service and privacy policy' area, indicating that the user has agreed to the terms while being presented with a modified form.

Our Local-Feature-based Detection Method

- Step 1: Visual assessment with local content descriptors
 - Context Contrast Histogram (CCH)
 - invariant to scale, rotation, etc.
 - even more efficient than SIFT, the most well-known descriptor for its excellent performance
- Step 2: Page scoring & classification
 - Scoring Criteria
 - correct matching rate
 - ratio of matched area
 - Naïve bayesian classification

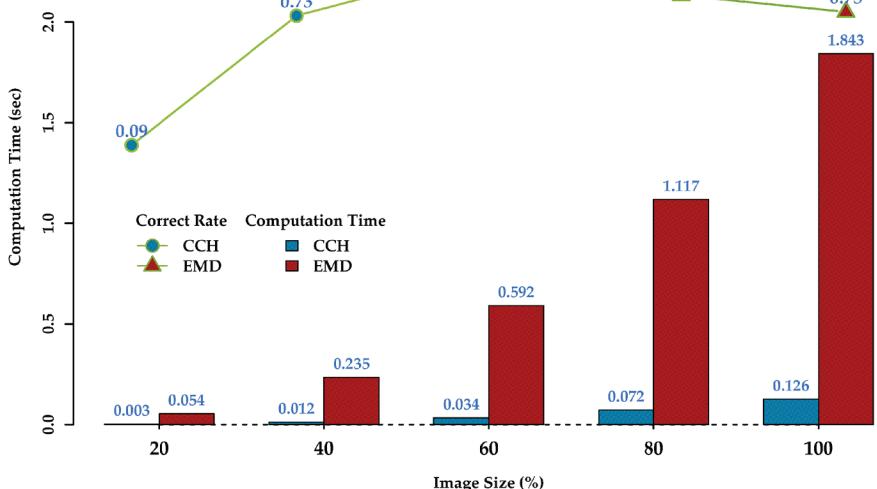
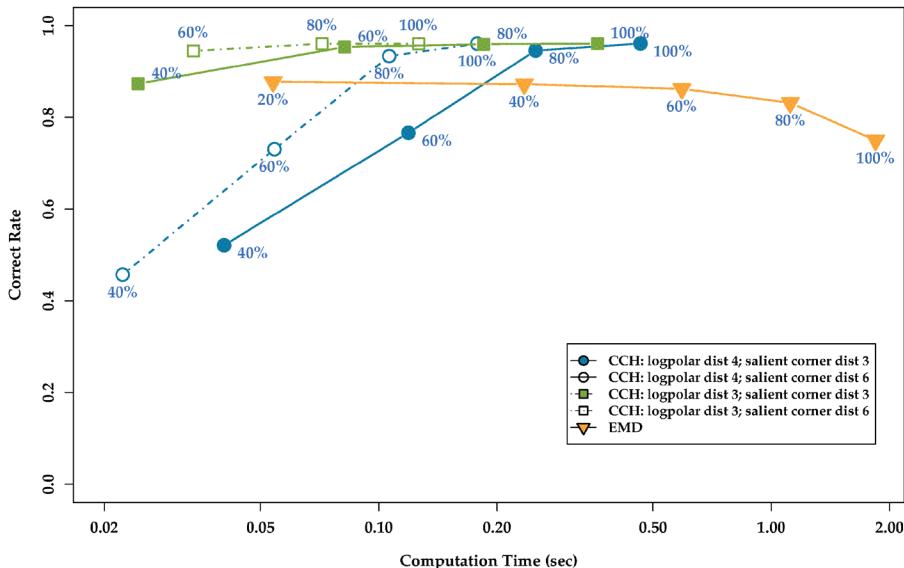
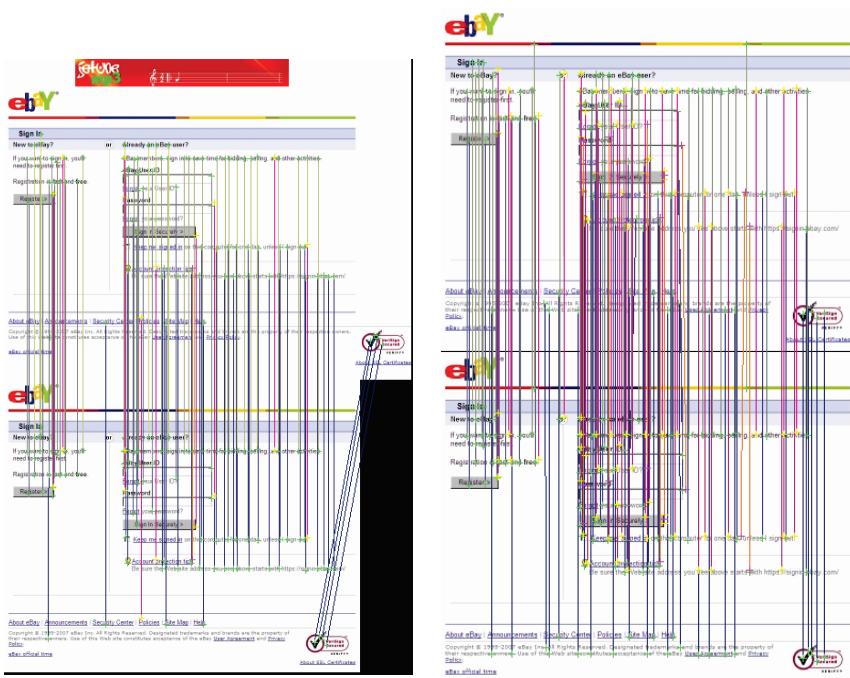


Phishing Page Matching (Classification)



Performance Evaluation

- Superior to EMD (Earth-Mover's Distance) scheme (IEEE TDSC, 2006)



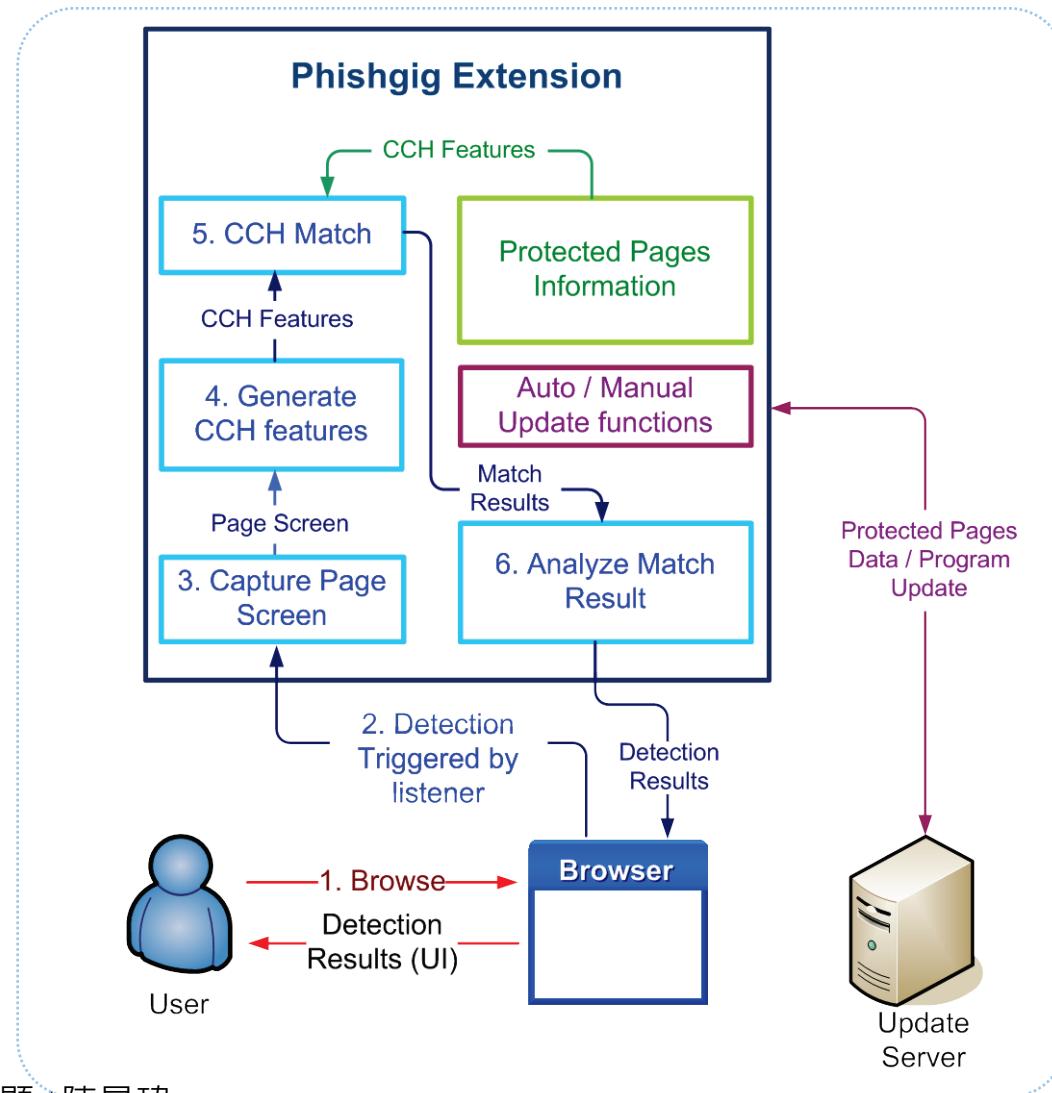


What is Phishgig?

- An Anti-Phishing tool
 - Implemented as browser extension (XUL + XPCOM)
 - for Firefox (3.0 or later)
 - Current version: 0.6.1 (Under development)
- Built in phishing detection?
 - Based on black list
 - A phishing page needs 24 hrs to be reported
- Real-time phishing protection
 - Using robust local feature-based scheme
 - Configurable protection
 - Update (Auto or Manual)



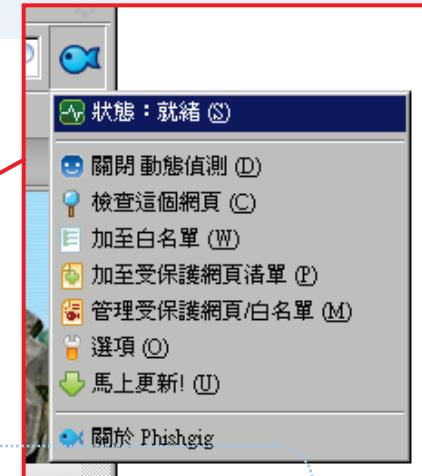
How Phishgig works



Phishgig

<http://mmnet.iis.sinica.edu.tw/proj/phishgig/>

1. Installed in Firefox 3.0.1



2. Live Status



3. Protected Pages Management

A screenshot of the 'Phishgig 受保護網頁 / 白名單管理員' window. It lists various websites under '目標網站' and their corresponding '網域', '類別', and '啓用' (Enabled) status. An 'eBay' login page is embedded within the window, showing the 'Welcome to eBay' and 'Ready to bid and buy? Register here' sections.



Phishgig in Action

1. Legitimate eBay login page

分析資訊

- 狀態: 釣魚網頁!
- 網路釣魚目標: ebay_0
- 相似度: 77%

網頁資訊

- 網址: http://larvikfutsal.com/templates/rhuk_milkyway/eBayISAPI.dll&SignIn
- 基本網域: larvikfutsal.com
- IP位置: 195.159.134.237

2. Fake eBay login page

分析資訊

- 狀態: 釣魚網頁!
- 網路釣魚目標: ebay_0
- 相似度: 77%

網頁資訊

- 網址: http://larvikfutsal.com/templates/rhuk_milkyway/eBayISAPI.dll&SignIn
- 基本網域: larvikfutsal.com
- IP位置: 195.159.134.237



交流時間



PHISHDEF: URL NAMES SAY IT ALL

PHISHDEF: URL NAMES SAY IT ALL

○ Authors:

- Anh Le, Athina Markopoulou
(University of California, Irvine)
- Michalis Faloutsos
(University of California, Riverside)

○ Source:

- to appear in IEEE INFOCOM 2011 Mini Conference,
Shanghai, China, April 10-15, 2011. (poster, tech
report)

OUTLINE

- Introduction
- Dataset and Feature Extraction
- Classification Algorithms
- Evaluation Results
- System Deployment
- Conclusion

INTRODUCTION

- “How well can one detect phishing URLs using only lexical features compared to using full features?”
- PhishDef Properties:
 - High accuracy:
 - 96%-97%
 - Light-weight:
 - Low latency
 - Imposes a modest overhead
 - Proactive approach
 - As opposed to reactively relying on blacklist
 - Resilience to noise
 - 95%-86% accuracy when there is 5%-45% noise

DATASET AND FEATURE EXTRACTION

○ Dataset

- Malicious URLs
 - PhishTank
 - MalwarePatrol
- Legitimate URLs
 - Yahoo Directory
 - Open Directory (DMOZ)

○ External Feature Collection

- WHOIS
- Team Cymru

DATASET AND FEATURE EXTRACTION(CONT.)

○ Feature Extraction

- Automatically selected features
 - Delimiters: '/', '?', '.', '=', '_', '&' and '-'.
 - Four parts:
 - Domain Name
 - Directory
 - File Name
 - Argument
- Obfuscation-resistant lexical features
 - Four different URL obfuscation techniques
 - Five categories of hand-selected lexical features

URL OBFUSCATION TECHNIQUES

- (I) Obfuscating the host with an IP address
- (II) Obfuscating the host with another domain
- (III) Obfuscating with large host names
- (IV) Domain unknown or misspelled

Type	Descriptive Examples
I	http://210.80.154.30/~test3/.signin.ebay.com/ebayisapidllsignin.html http://0xd3.0xe9.0x27.0x91:3030/www.paypal.com/uk/login.html
II	http://21photo.cn/https://cgi3.ca.ebay.com/eBayISAPI.dllSignIn.php http://2-mad.com/hsbc.co.uk/index.html
III	http://www.volksbank.de.custsupportref1007.dllconf.info/r1/vm http://sparkasse.de.redirector.webservices.aktuell.lasord.info
IV	http://www.wamuweb.com/IdentityManagement/ http://mujweb.cz/Cestovani/1om3/SignIn.html?r=7785

Phishing URLs characteristics

www.paypal.creasconsultores.com/www.paypal.com/Resolutioncenter.php

shevkun.org/css/paypal.com/cgi-bin/cmd%3D_login-submit/css/websc.php

us-mg6.mail.yahoo.com.dwarkamaigroup.com/Yahoo.html

emailloans.hostingventure.com.au/bankofamerica.com

nitkowski.pl/components/wellsfargo/questions.php

The registered domain has no relationship with the rest of the URL

http://4ld.3ld.mld.ps /path1/path2?key1=value1&key2=value2

- Most parts of URLs can be freely defined
- Except the **registered domain**: main level domain + public suffix

HAND-SELECTED FEATURES

- Features related to the full URL
 - Length of the URL (Type II)
 - Number of dots in the URL (Type II)
 - Blacklisted words (Type IV)
 - confirm, account, banking, secure, ebayisapi, webscr, login and signin
 - Paypal, free, lucky and bonus

- Features related to the domain name
 - Length of the domain name (Type III)
 - IP or port number is used in the domain name (Type I)
 - Number of tokens of the domain name (Type III)
 - Number of hyphens used in the domain name (Type III)
 - The length of the longest token (Type III)

- Features related to the directory
 - Length of the directory (Type II)
 - Number of sub-directory tokens (Type II)
 - Length of the longest sub-directory token (Type II)
 - Maximum number of dots and other delimiters used in a sub-directory token (Type II)

HAND-SELECTED FEATURES

- Features related to the file name
 - Length of the file name (Type II)
 - Number of dots and other delimiters used in the file name (Type II)
- Features related to the argument part
 - Length of the argument part
 - Number of variables
 - Length of the longest variable value
 - The maximum number of delimiters used in a value

URL	www.naturenilai.com/form2/paypal/webscr.php?cmd=_login
Auto-Selected	name=www, name=naturenilai, tld=com, dir=form2, dir=paypal file=webscr, ext=php, arg=cmd, arg=login
Obfuscation-Resistant	URL len=54, n_dot=3, blacklist=1
	Domain Name len=19, IP=0, port=0, n_token=3, n_hyphen=0, max_len=11
	Directory len=14, n_subdir=2, max_len=6, max_dot=0, max_delim=0
	File Name len=10, n_dot=1, n_delim=0
	Argument len=11, n_var=1, max_len=6, max_delim=1

- Summary of dataset

CLASSIFICATION ALGORITHMS

○ Batch Learning

- Support Vector Machine (SVM)

○ Online Learning

- Online Perception (OP)
- Confidence Weighted (CW)

$$(\boldsymbol{\mu}_{t+1}, \Sigma_{t+1}) = \arg \min_{\boldsymbol{\mu}, \Sigma} D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \parallel \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)),$$

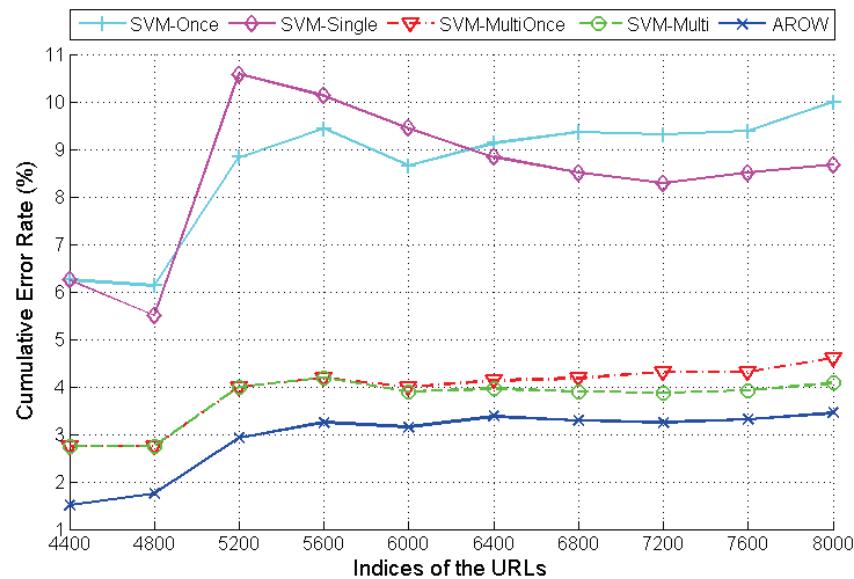
$$\text{s.t. } \Pr_{\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)}[y_t(\mathbf{w} \cdot \mathbf{x}_t)] \geq \eta.$$

- Adaptive Regularization of Weights (AROW)

$$(\boldsymbol{\mu}_{t+1}, \Sigma_{t+1}) = \arg \min_{\boldsymbol{\mu}, \Sigma} D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \parallel \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)) + \lambda_1 l_{h^2}(y_t, \boldsymbol{\mu} \cdot \mathbf{x}_t) + \lambda_2 \mathbf{x}_t^T \Sigma \mathbf{x}_t,$$

EVALUATION RESULTS

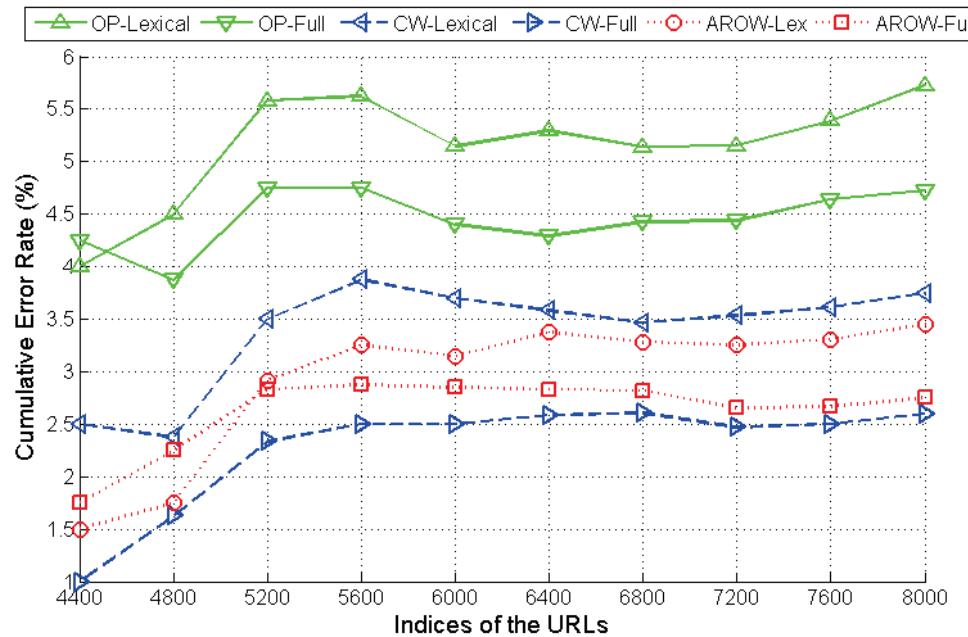
- Batch-based vs. Online algorithms
 - SVM vs. AROW
 - Yahoo-Phish



EVALUATION RESULTS(CONT.)

○ Lexical Features vs. Full Features

- OP, CW and AROW
- Yahoo-Phish



EVALUATION RESULTS(CONT.)

⦿ Obfuscation-Resistant Lexical Features

- Performance of AROW with/without OR features after the last URL

Dataset	Cumulative Error Rate (%)			# Mis-Classified Benign URLs (FP)			# Mis-Classified Malicious URLs (FN)		
	w/o OR Ftrs	with OR Ftrs	Gain (Gain Pctg)	w/o OR Ftrs	with OR Ftrs	Gain	w/o OR Ftrs	with OR Ftrs	Gain
Yahoo-Phish	3.92 ($\lambda = 0.5$)	3.45 ($\lambda = 0.5$)	0.37 (9%)	47	55	-8	110	83	27
Yahoo-Malware	3.70 ($\lambda = 5.0$)	3.05 ($\lambda = 5.0$)	0.65 (18%)	88	77	11	60	45	15
DMOZ-Phish	4.05 ($\lambda = 5.0$)	3.60 ($\lambda = 50$)	0.45 (11%)	50	53	-3	112	91	21
DMOZ-Malware	5.12 ($\lambda = 0.5$)	3.75 ($\lambda = 0.5$)	1.37 (27%)	22	41	-19	183	109	74



交流時間





資料科學如何輔助線上遊戲 虛寶銷售

facebook

Gift Shop

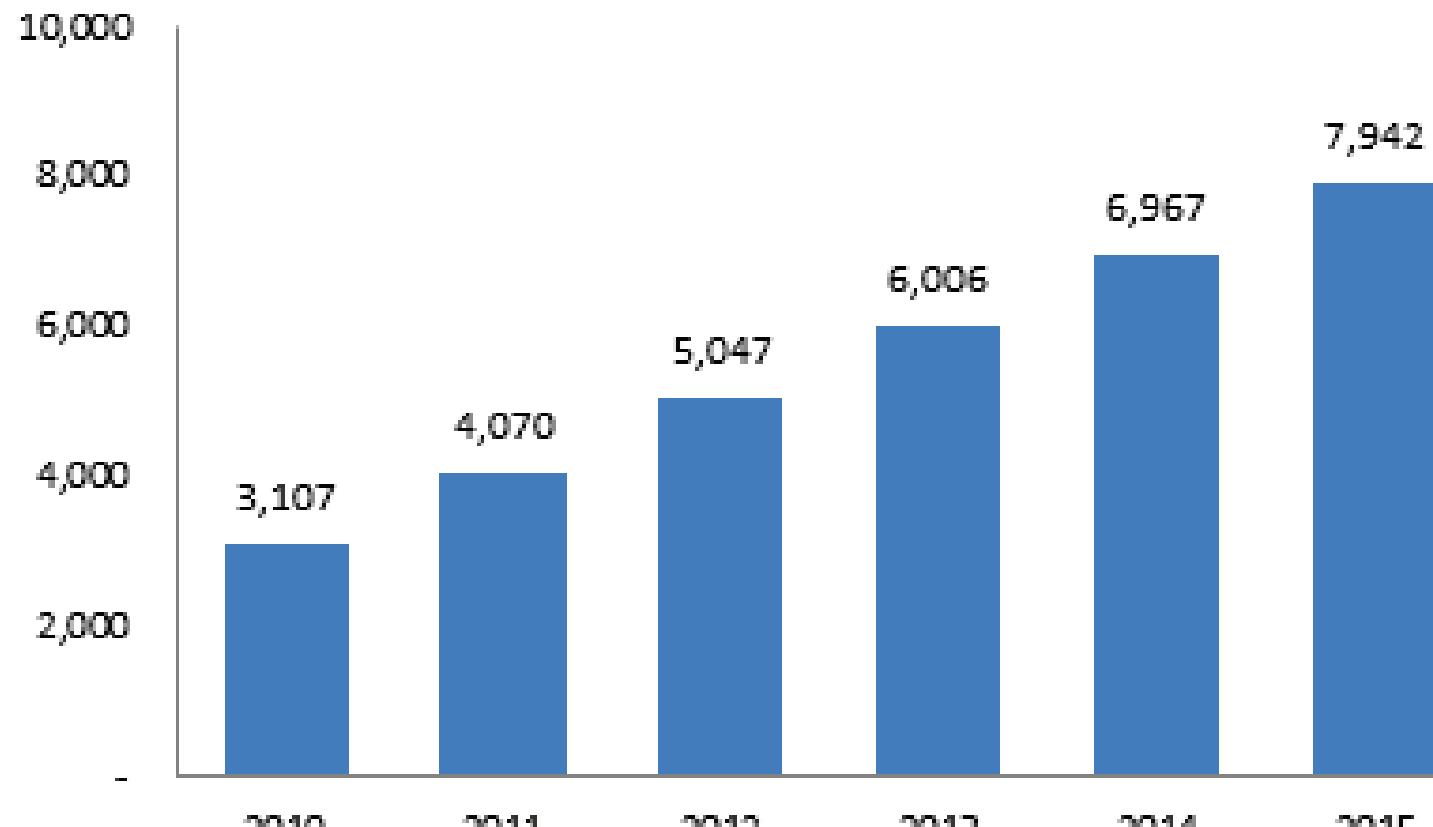


Displaying 1 - 28 of 84 gifts.

[1](#) [2](#) [3](#) [next](#)



Worldwide Virtual Goods Revenue (in millions USD)



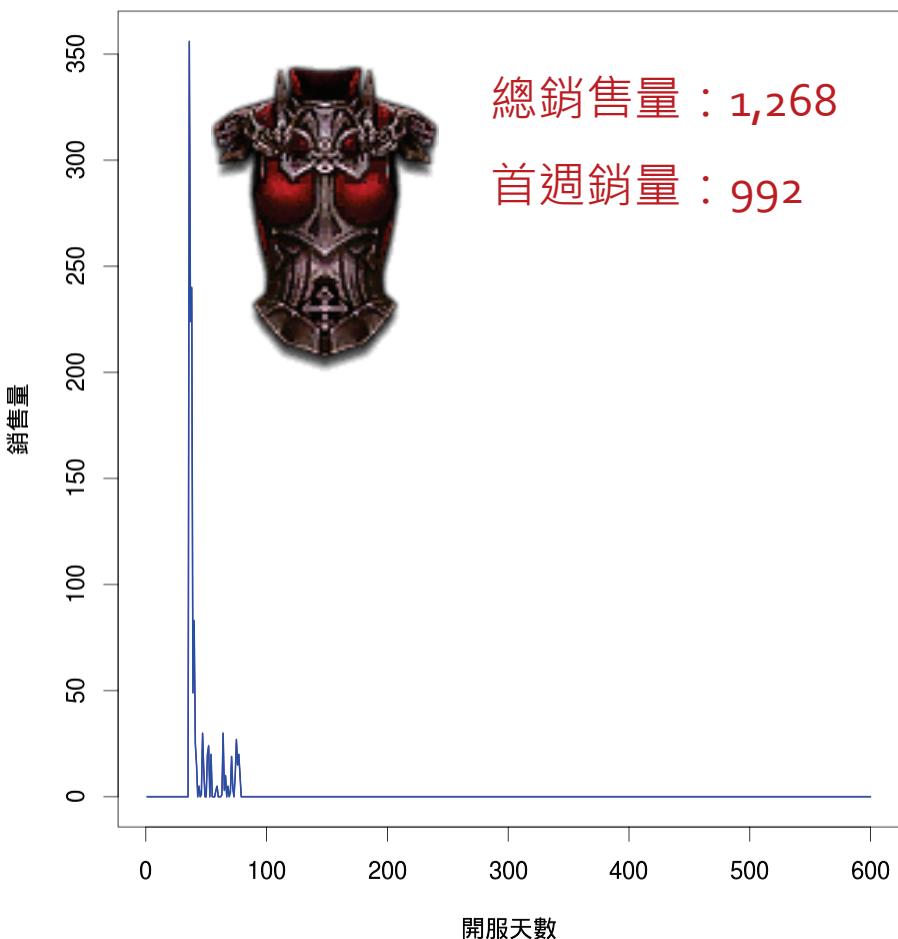
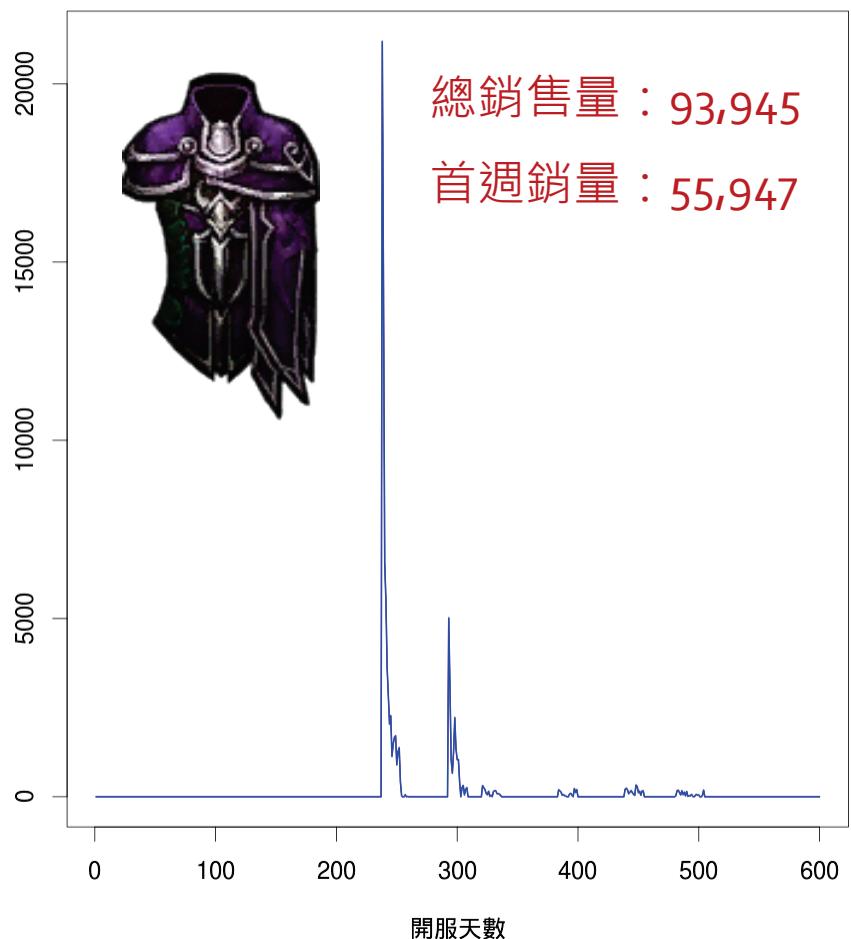
Source: eMarketer, Jiang Zhang

哪一件銷量最好？



<http://diablo.inegamers.com>

商品銷售差異



資料分析團隊該通常做些什麼？

■ 玩家層面

- DAU, WAU, MAU
- 上線時間
- 平均花費

■ 玩家 vs. 商品

- 玩家對於特定商品的偏好
- 玩家屬性(性別、年紀、等級、職業、是否 VIP)、購買期間與商品的關係

■ 商品層面

- 每個商品的交易量
- 每個商品隨著時間交易量演進

■ 行銷作法

- 使用推薦系統來做個人化推薦商品給玩家

其實我們很想知道一個問題...



<http://diamondincgamers.com>

以資料分析幫助設計虛擬商品

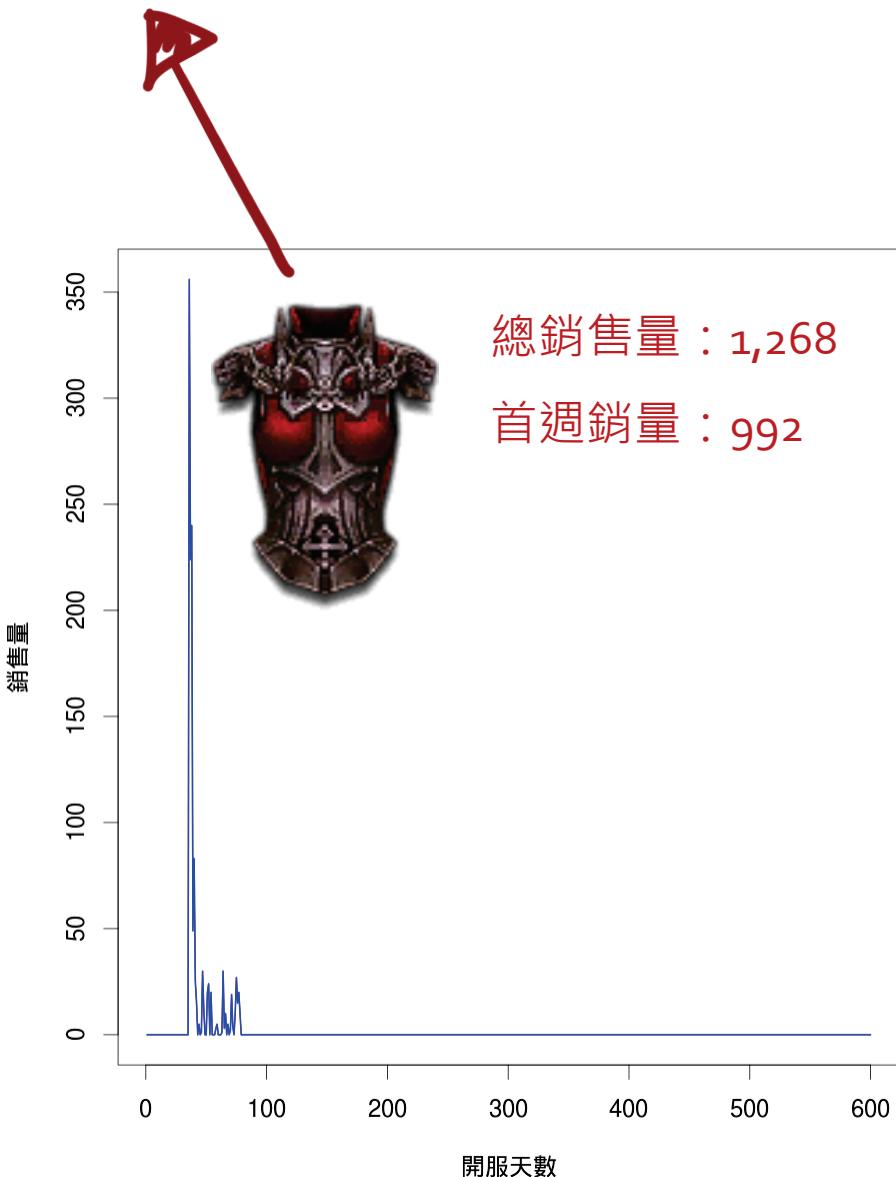
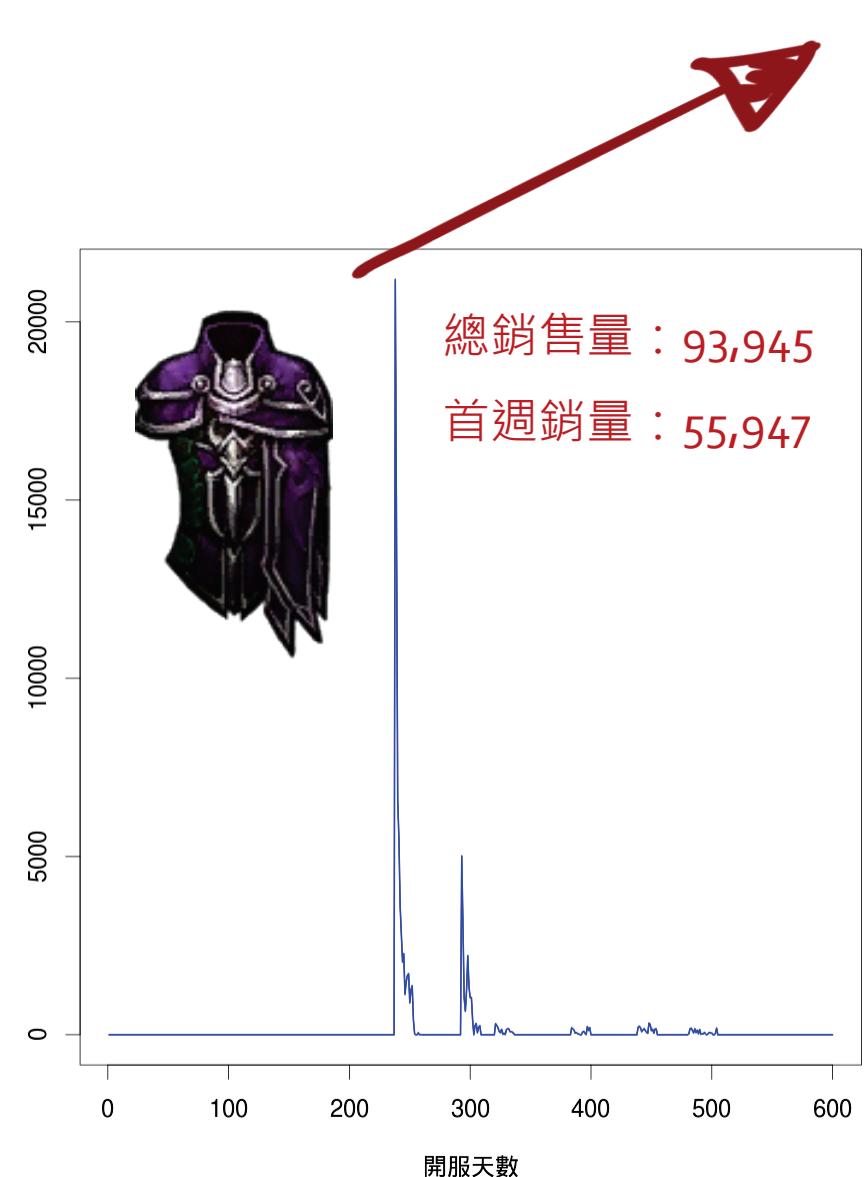
- 量化影響虛擬商品銷售好壞的要素
 - 主觀要素
 - 影像訊號要素
- 提供可以讓設計師參考的**設計指引**
- 建構一套系統化的方法，為運行在不同**區域, 國家**的遊戲，提供調整虛擬商品設計的準則

目標

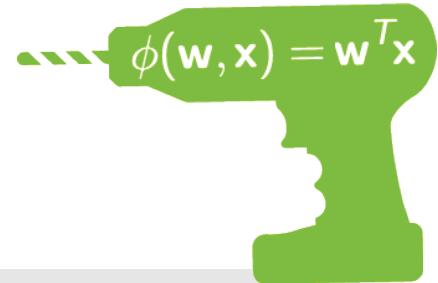


設計熱銷的虛擬商品

Semantic attributes are needed



Feature Engineering



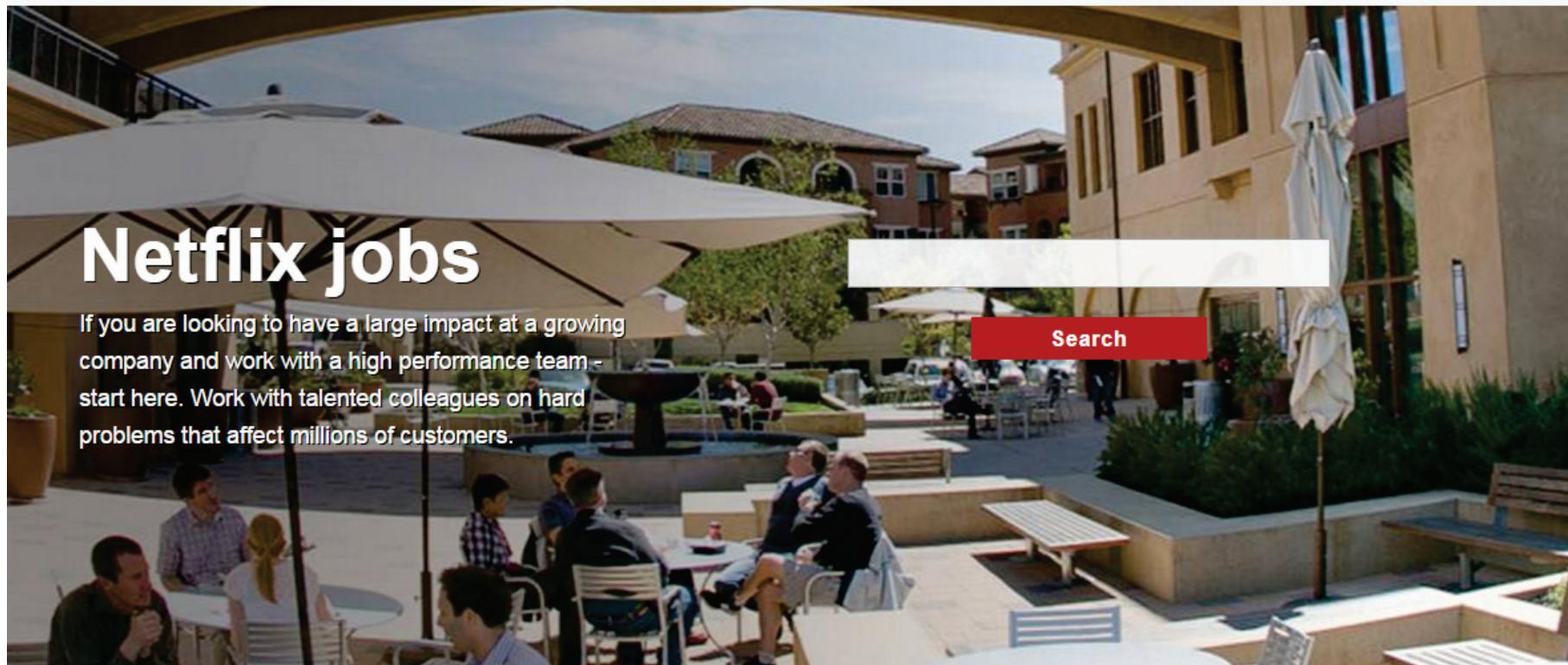
“

A feature is a piece of information that might be useful for prediction. Any attribute could be a feature, as long as it is useful to the model.

"...some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used."

—Pedro Domingos,
"A Few Useful Things to Know about Machine Learning"





Netflix jobs

If you are looking to have a large impact at a growing company and work with a high performance team - start here. Work with talented colleagues on hard problems that affect millions of customers.

Search

<http://jobs.netflix.com/jobs.php?id=NFXo1466>

 Engineering

 Data Science and
Engineering

 IT Operations

 User Experience /
Design

Content

UK/Ireland Tagger

Product Management

London, GBR

Netflix, the world's leading internet channel for movies and TV is launching a hunt for a UK/Ireland based tagger to join its enhanced content team.

Netflix Taggers

- 聘請專人依照 SOP (36 pages) 觀賞並標註影片
- 555 個標籤，76,897 種組合 (2014年一月)
- 以標籤為基礎建立影片推薦系統



Netflix Micro-genres for Videos

177	Critically-acclaimed British Dramas based on real life	196	TV Comedies from the 1990s
178	Romantic Comedies Featuring a Strong Female Lead	197	Steamy Independent Dramas
179	Cerebral Documentaries	198	Suspenseful TV Mysteries
180	Emmy-winning TV Dramas	199	Dark Romantic British Dramas based on Books
181	Witty TV Comedies	200	Because you watched Luther
182	Exciting 20th Century Period Pieces based on real life	201	Suspenseful TV Dramas
183	TV Comedies	202	Quirky TV Shows
184	Violent Movies	203	Witty TV Comedies
185	Scary Suspenseful Movies	204	Suspenseful TV Sci-Fi & Fantasy
186	Exciting Movies	205	Critically-acclaimed Visually-striking Movies
187	Crime TV Shows	206	Visually Striking Ominous Movies
188	Violent Race Against Time Movies	207	Period Dramas with a Strong Female Lead
189	Disney Comedies	208	Critically-acclaimed Movies about Art & Design
190	Quirky Sitcoms	209	Visually-striking Exciting Sci-Fi and Fantasy
191	Visually-Striking Independent Dramas	210	Military 20th Century Period Pieces
192	Irreverent TV Comedies from the 1990s	211	Exciting Action Sci-Fi & Fantasy
193	Exciting Crime TV Shows	212	Witty Romantic Comedies Featuring a Strong Female Lead
194	Visually-striking Dark Movies	213	Dramas Featuring a Strong Female Lead
195	Family-friendly Comedies	214	Gritty Movies

科技三箭？



Crowdsourcing

= Crowd + Outsourcing

*“soliciting solutions via open calls
to large-scale communities”*

A more formal definition

“Crowdsourcing is the act of taking a **job** traditionally performed by a designated agent (usually an employee) and outsourcing it to an **undefined**, generally large **group of people** in the form of an open call.” [1]

[1] Howe, Jeff. Crowdsourcing: A Definition, <http://crowdsourcing.typepad.com/>

Image Semantics

- Reward: 0.04 USD / task



Instructions: Provide information about the following image(s) by accurately answering the following questions.

Guidelines:

- Specific terms are preferred (Disneyland vs amusement park)
- Correct spelling is required (Hint: Use Firefox for spellchecker functionality)
- Don't repeat terms

main theme?
key objects?
unique attributes?

Main Theme
What is the MAIN theme of the image?

Key Objects
What are the key objects in the image?

Unique Attributes
What are some unique attributes about this image? (actions, emotions, colors)



find out photos of revolvers!

Main Unsure? Look up in Google Wikipedia

Click on the photos that contain:

revolver, six-gun, six-shooter: a pistol with a revolving cylinder (usually having six chambers for bullets)

Note: Please pick as many as possible, otherwise your submission may be rejected. You may receive a bonus up to \$0.04 based on the quality of your submission. It is OK to have OTHER objects in the photo. PICK ONLY PHOTOS -- NO DRAWINGS OR COMPUTER GRAPHICS.



Below are the photos you have selected. Click to deselect.



| < < page 1 of 2 > > |

0.02 USD/ task

Human Skeleton



0.01 USD/ task

Photo Orientation



Mechanical Turk Project

If you're using the turk, Be sure to copy the text back into the HIT page so that you can be credited.

- Photo should be rotated 90 degrees left (counter-clockwise)
- Photo should be rotated 90 degrees right (clockwise)
- Photo should be turned upside down
- Photo is oriented properly

Please describe the picture in the box using 10 words or more:

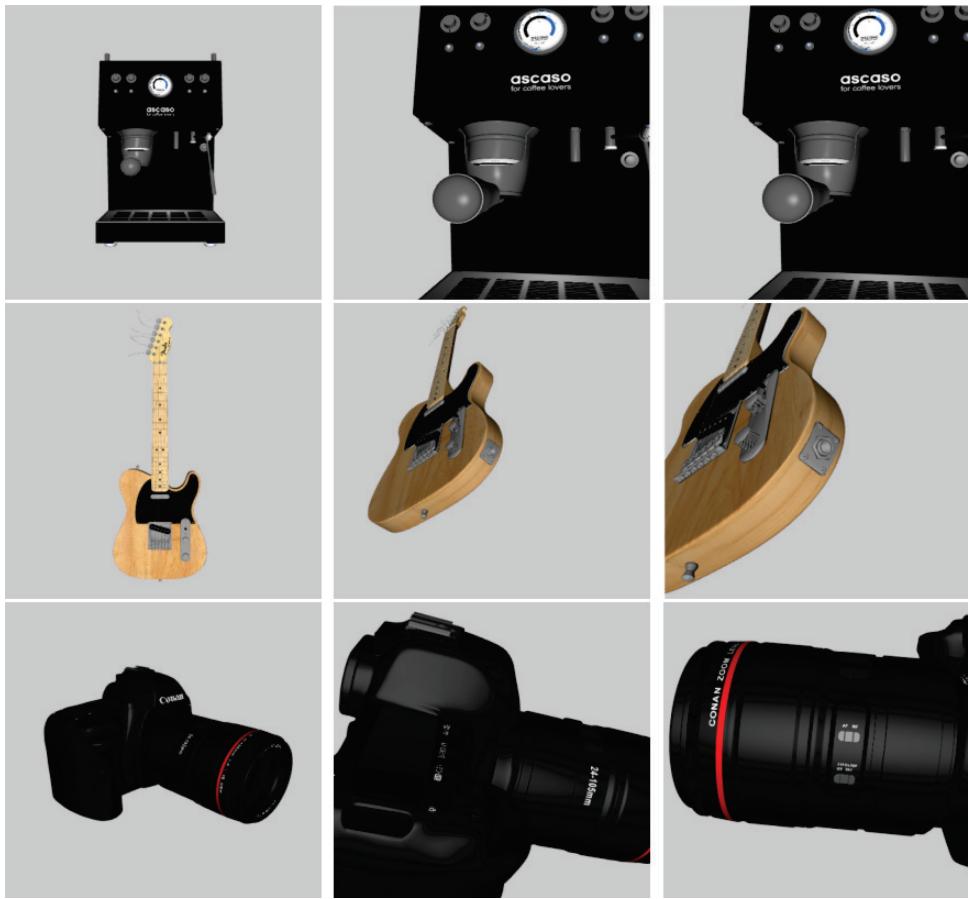
shells

[Submit Turk](#) | [Skip / Load a different photo](#)

The submit button MUST be clicked!

0.01 USD/ task

Perspectives for 3D Objects



Thi Phuong Nghiem, Axel Carlier, Geraldine Morin, and Vincent Charvillat,
"Enhancing online 3D products through crowdsourcing," ACM
CrowdMM'12.

Web Site Classifier



Site Navigation

- » Main Page
- » Asian Movies Only
- » Asian Pictures Only
- » Japanese Sex
- » Free Asian XXX
- » Hot AV Idols
- » Thai girls and porn
- » Hentai Toons
- » Full Text Version

G (general audience) **PG** (parental guidance) **R** (restricted) **X** (porn)

12 USD / hour

Panos Ipeirotis, "Crowdsourcing using Mechanical Turk: Quality Management and Scalability," Invited Talk at CSDM 2011.

Photographers' Intention



- to support a task?
- to capture a bad feeling?
- to preserve a good feeling?
- to recall later on?
- to publish it online?
- to show it to friends and family?

Mathias Lux, Mario Taschwer, and Oge Marques, "A Closer Look at Photographers' Intentions: a Test Dataset," ACM CrowdMM'12.

Linguistic Affective Judgement

- Affective response (Snow et al. 2008)

Headline: **Closings and cancellations top advice on flu outbreak**

How much does this headline evoke the following emotions?

Anger (0-100)

Disgust (0-100)

Fear (0-100)

Joy (0-100)

Sadness (0-100)

Surprise (0-100)

In general, how positive or negative is this headline, on a scale of:

-100 (very negative) <--- 0 (neutral) ---> 100 (very positive)

Comment

“Closing and cancellations
top advice on flu outbreak”

USD 0.4 to label 20 headlines (140 labels)

A Lot More Examples

- Document relevance evaluation
- Document rating collection
- Noun compound paraphrasing
- Person name resolution
- Among others...

\$ 10,00,00



Bounty Workers

線上微型案件媒合平台

\$ 9,00,00



截至目前 Bounty Workers 已成功執行 3250 次任務，並且發出了 99,275 元的獎勵

<http://bountyworkers.net/>



可執行任務

可以藉由完成這些任務，從中獲取豐富獎勵！



會議問卷建檔

by Salmon

\$ 50 元 / 約 20 分鐘

可執行 3 次



公益文章捐款意願調查

by Jason

\$ 30 元 / 約 10 分鐘

可執行 2 次

◎ 剩餘 26 天



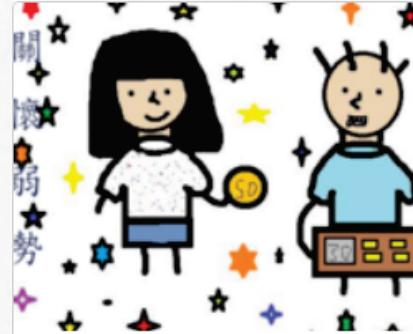
跨平台影片播放使用者滿意度調查

by MMNET

\$ 40 元 / 約 25 分鐘

可執行 1 次

◎ 剩餘 3 天



關懷弱勢 傳愛慈善公益粉絲團

by 賴

\$ 30 元 / 約 60 分鐘

可執行 1 次



侍者、變裝、僕從、
小妹、遐想、貓女、
短裙、萌萌、長腿、
長襪、女僕、俏麗、
甜美、奪目、可愛、
幫傭、女侍、女佣、
服從、服務、迷裙



交流時間



Data Librarian



Data Journalist



Data Analyst



Data Engineer



Data Steward



Data Archivist



資料科學人才的養成

陳昇瑋

中央研究院資訊科學研究所

Major Roles in a Data Team



Data Project Manager



Data Scientist



Data Analyst



Data Engineer

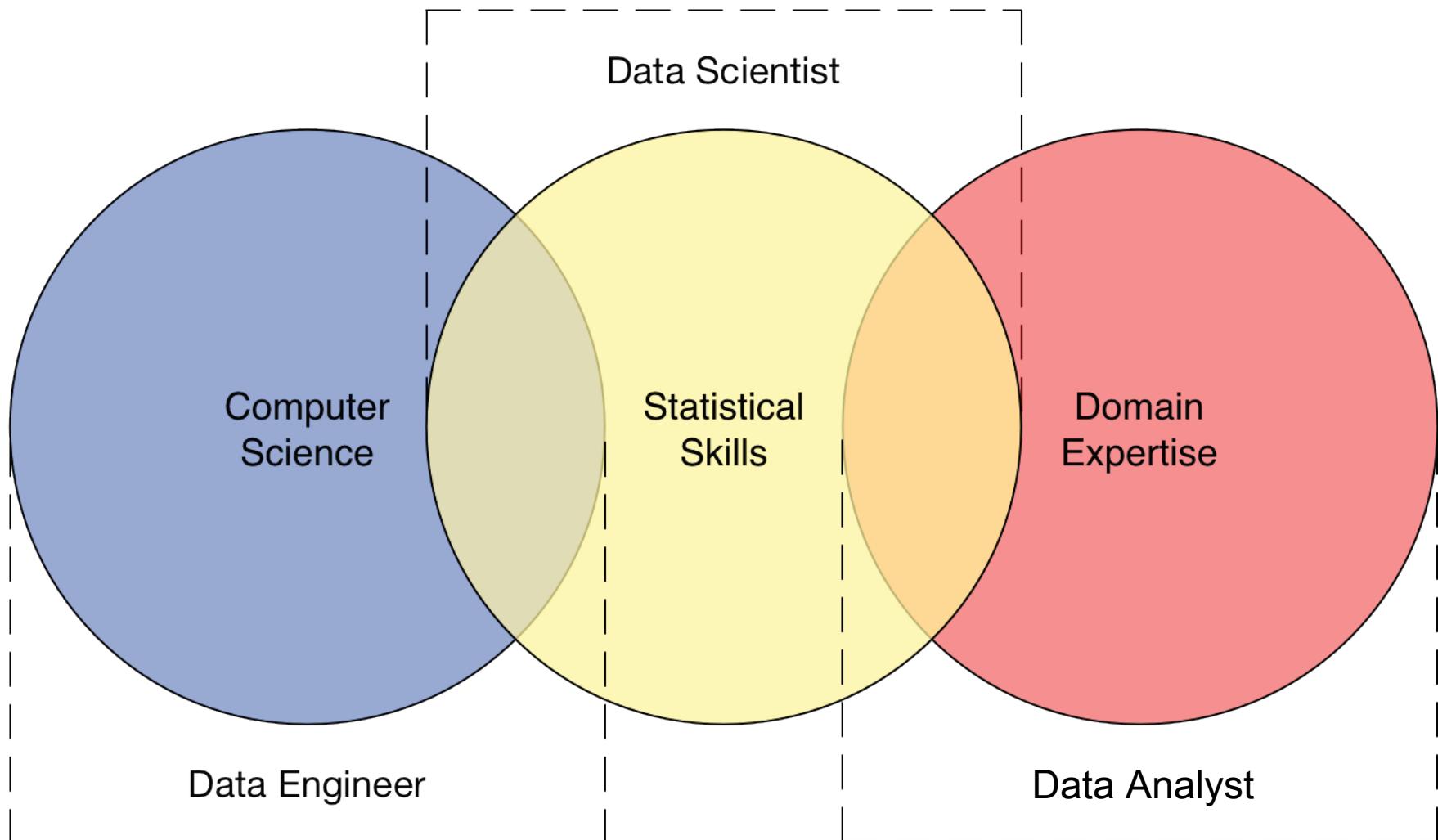
技術背景

資料科學家 / 分析師

- Statistics
- Statistical packages (e.g., R, Python)
- Machine learning
- Domain-specific data mining techniques
- Data visualization

資料工程師

- UN*X / Web programming
- DBMS
- Data crawling / parsing
- Data cleansing
- Data visualization techniques (e.g., d3.js)



~~理~~夢想中的資料科學家



Data Analyst

■ 資料分析師

- 統計分析、建模
 - 報表及視覺資料呈現
 - 機器學習

■ 大數據資料分析家

- 懂得分析文字、影片或圖像等非結構化資料
 - 知道如何引入外部資料來做結合

駭客

- 會寫程式
 - 能掌握大數據技術架構

● 科學家

- 科學性思維
 - 探索未知，定義問題
 - 設計實驗，驗證假設

商業專家

- 企業如何運作、如何賺錢？
 - 對於要把資料分析與大數據運用在哪些層面很有看法

■ 可靠的顧問

- 良好的溝通能力與人際技巧
 - 懂得發問，能快速掌握問題的核心及評估可行性



在奇異公司，我們發現，具備兩到三種專業知識的資料科學家，做起事來最有效率。有幾個原因可證明此事為真。

首先，懂得多領域的知識，似乎能夠在創造力方面帶來可觀的優勢。我聽說這叫做「有利的位置」(*coign of vantage*)，此處應該將之視為「建築物的基石」。

矗立在建築物外部轉角處的基石，雖然無法看到建築物的每一面，卻占有同時看到兩面的優勢，這也導致這樣的人比只能看到其中一面的人擁有莫大的創造優勢。

資料科學家也是如此。

資料素養

- 瞭解資料的(潛在)價值

看似簡單的難題

- 如何提昇印度女性地位？
 - 墮胎 vs. 犯罪率？
 - 槍枝越多，犯罪越少？
 - 鯊魚殺的人多還是大象殺人多？
 - 足球罰球時，踢哪個方位最可能進球？
 - 兒童汽車座椅安全還是安全帶安全？
 - 酒醉只有開車才危險嗎？
- 假裝知道：傳統思維謬誤 / 道德羅盤 / 從眾與偏見



假裝知道

■ 未來其實難以預測

- 股市專家數年間超過 6,000 個預測，整體準確率僅為 47.4%

你真的喝得出貴的葡萄酒？

■ 哈佛學者學會上的盲目品酒測試

- 四個醒酒壺

1	貴葡萄酒 A
2	貴葡萄酒 B
3	便宜葡萄酒 C
4	貴葡萄酒 A



- 結果：四壺平均評分相近，且 1 號壺與 4 號壺評分差距最大！

你真的喝得出貴的葡萄酒？

■ Robin Goldstein 的實驗

- 在幾個月內到全美各地進行 17 項雙盲品酒測試
- 參加人數超過 500 人，包括入門人士、侍酒師與酒商
- 測試 523 種酒，每支酒價格從 1.65 美元至 150 美元不等
- 結果
 - 較貴的酒沒有獲得比較高分
 - 平均而言，昂貴葡萄酒的分數稍低於便宜的酒
 - 樣本中 12% 的參與者受過品酒訓練，但這些人並未偏好便宜的酒，也沒有明顯特別偏好昂貴的酒

人們為什麼自殺？

- 近年來美國兇殺率與交通死亡率均創新低，但自殺率幾乎不變，數十年間 15~24 歲的自殺率甚至增為 3 倍
 - 紐澤西 Richard Stockton 學院的心理學家 David Lester，透過 2 千 5 百多篇學術發表，探索自殺與其他事物的關聯：酒精、憤怒、抗憂鬱劑、星座、生物化學、血型、體型、憂鬱症、藥物濫用、槍枝控管、快樂、假期、網路使用、智商、心理疾病、偏頭痛、月亮、音樂、國歌歌詞、性格類型、抽煙、性靈、看電視、開闊空間
- 研究了這麼多，還是**不知道**到底人們為何自殺
 - David Lester 的結論：「**沒有**特定事物可以怪罪」

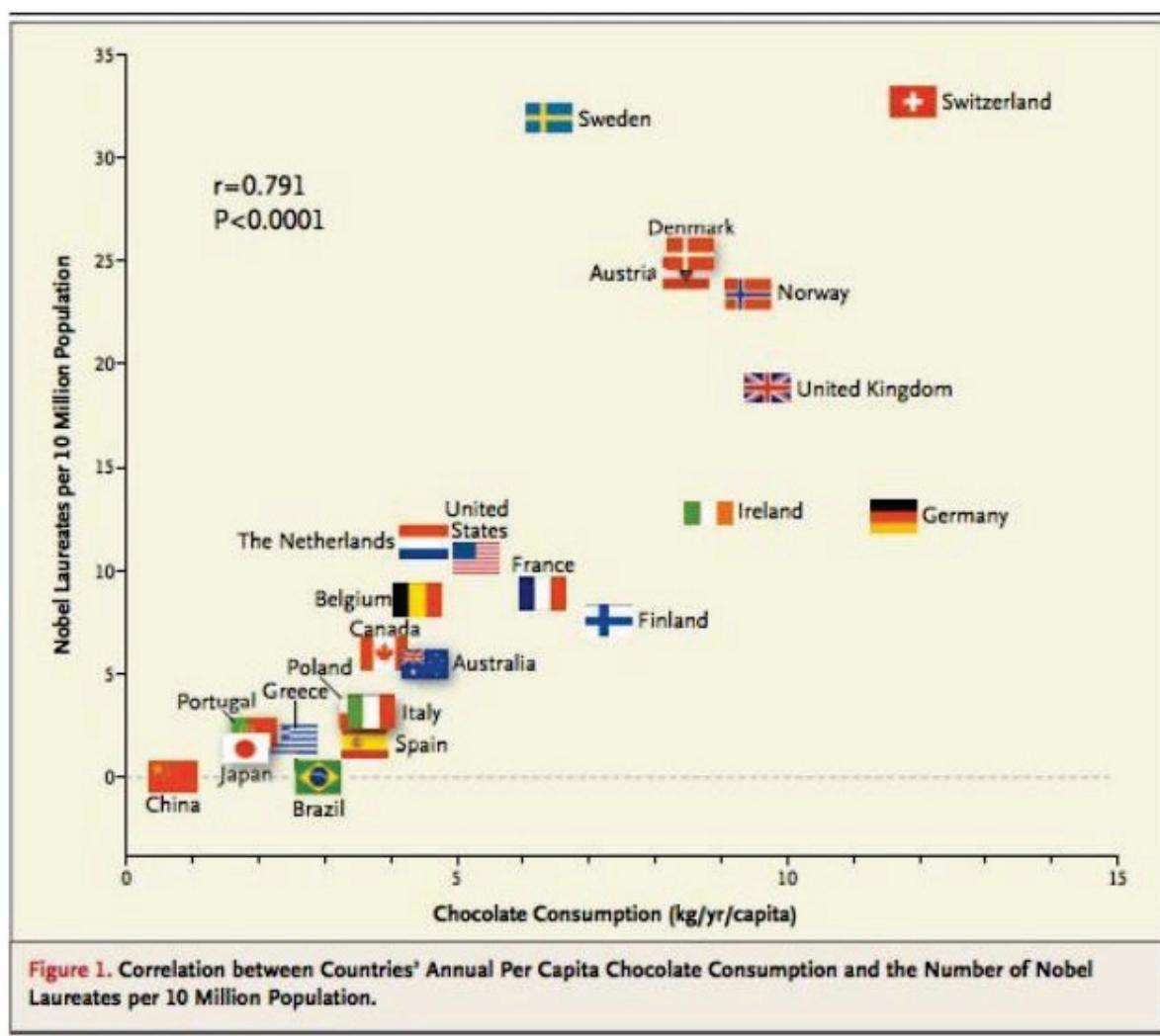
測量才可能確認真相

相關 ≠ 因果

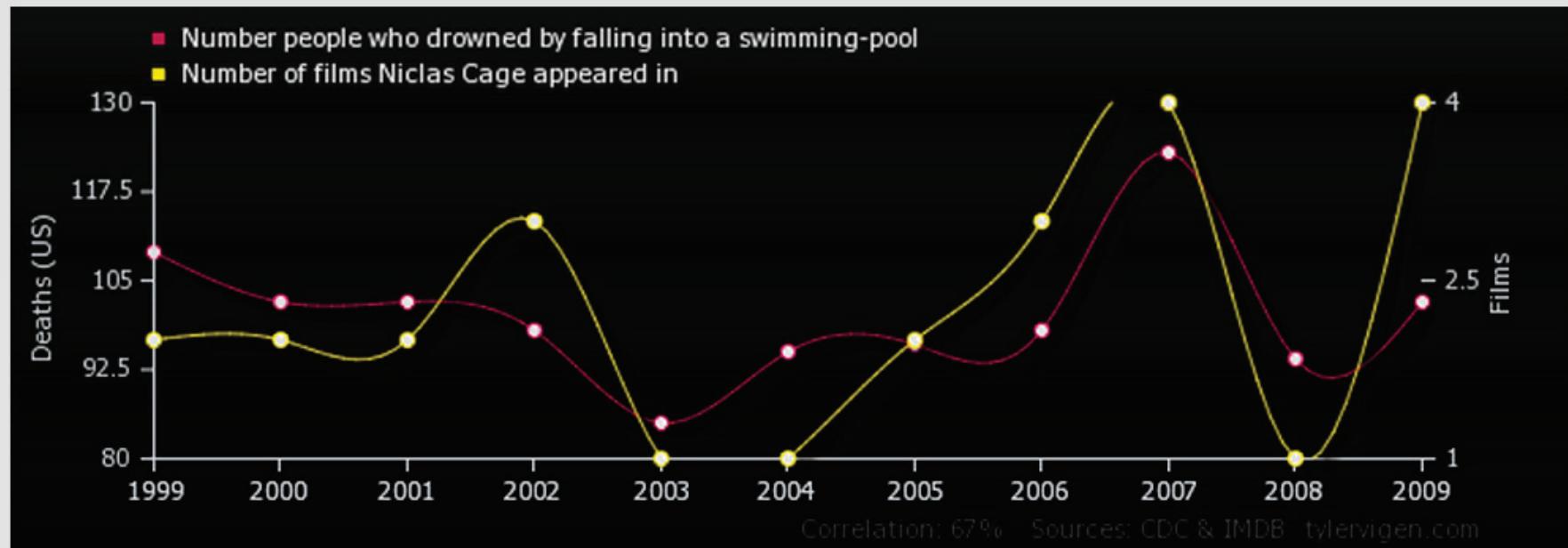
X 與 Y 相關

- X 導致 Y ?
- Y 導致 X ?
- 或另有變數同時導致 X & Y ?

巧克力消耗量 vs. 諾貝爾得獎數



Number people who drowned by falling into a swimming-pool correlates with Number of films Nicolas Cage appeared in



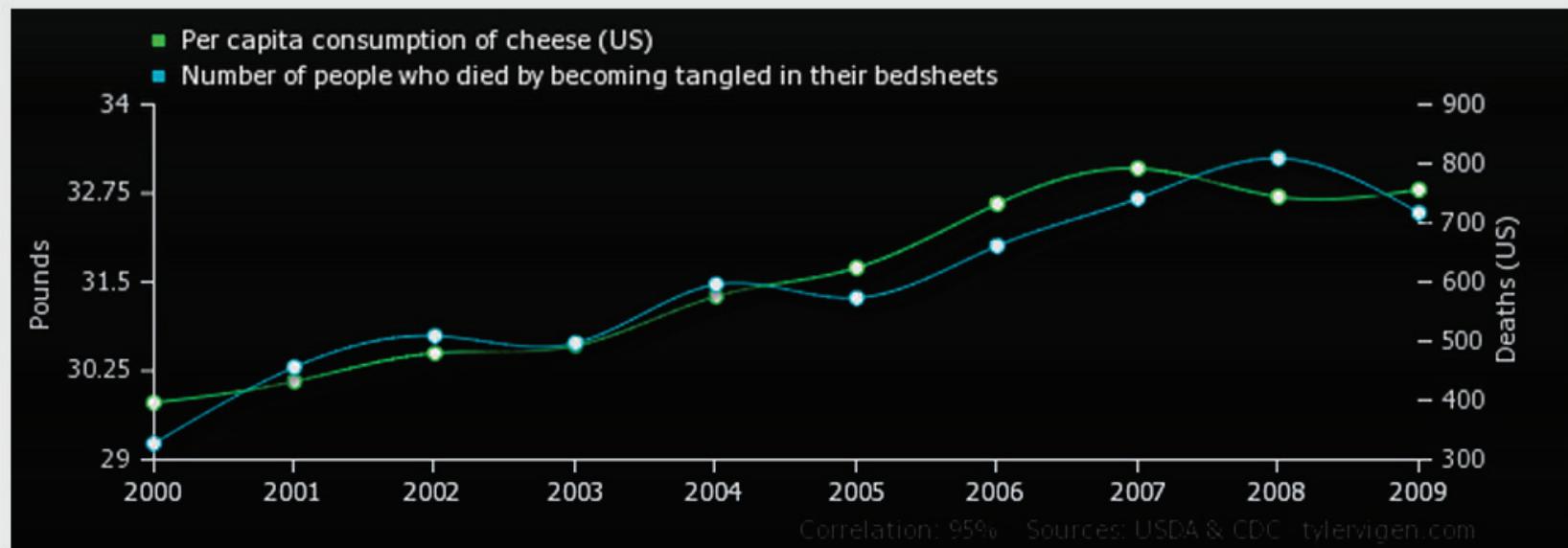
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool Deaths (US) (CDC)	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in Films (IMDB)	2	2	2	3	1	1	2	3	4	1	4

Correlation: 0.666004

Per capita consumption of cheese (US)

correlates with

Number of people who died by becoming tangled in their bedsheets



	<u>2000</u>	<u>2001</u>	<u>2002</u>	<u>2003</u>	<u>2004</u>	<u>2005</u>	<u>2006</u>	<u>2007</u>	<u>2008</u>	<u>2009</u>
<i>Per capita consumption of cheese (US)</i> Pounds (USDA)	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
<i>Number of people who died by becoming tangled in their bedsheets</i> Deaths (US) (CDC)	327	456	509	497	596	573	661	741	809	717

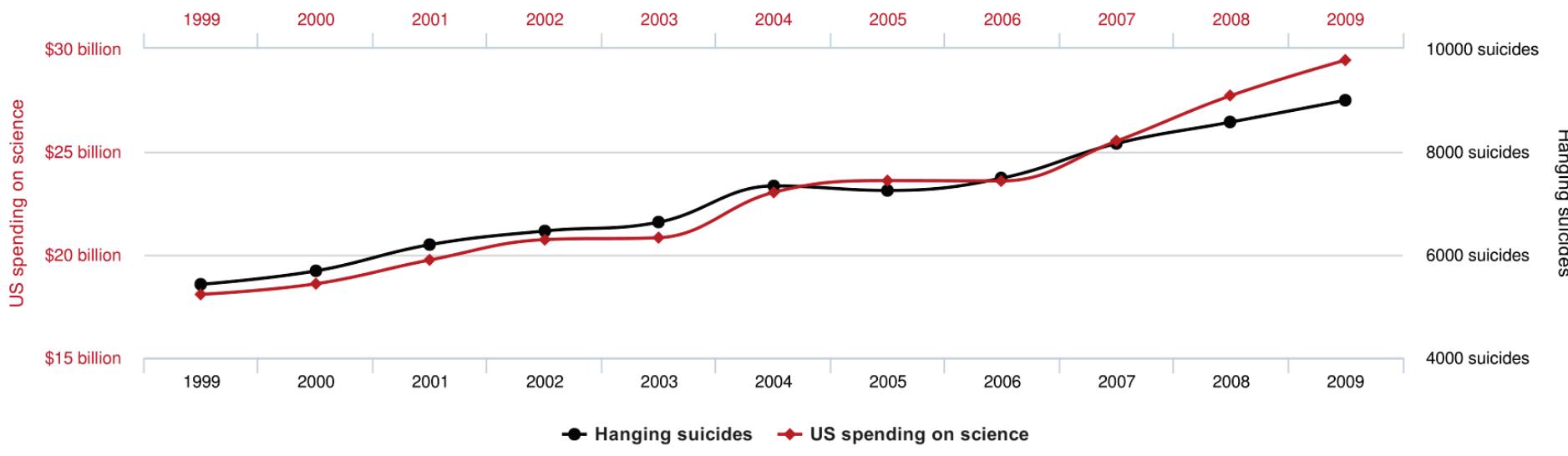
Correlation: 0.947091

上吊自殺 vs. 科學經費

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation



$$r=0.99789126$$

金錢有助勝選？

■ 花費高的候選人的確較常當選

- 是金錢讓人贏得選舉？
- 抑或領袖魅力引來捐款和選票？

■ 候選人吸引力如何量化？

- 檢視 1972 以來美國國會選舉，相同候選人連兩次對決比較
- 連續兩次 A vs. B 的情形約有 1,000 件，在候選人吸引力相對 穩定下，即可測量出金錢的作用
 - 勝者就算經費削減一半得票率僅減少 1%
 - 敗者儘管經費加倍，也不過多爭取到 1% 的得票率



父母對子女成績的影響？

■ 幼兒長期研究計畫

- 美國 1990 年代晚期，全國各地選出共 2 萬名以上學童詳細調查背景資料，並測量由幼稚園到五年級的學業進步情形

■ 迴歸分析結果

- ~~家中藏書豐富，是否讓小孩在學校表現優良？~~
- 家中藏書豐富的小孩，是否比沒有書的小孩表現好？
→ 家中藏書豐富的小孩，成績優於沒書的小孩
- 但家中藏書或許只反應家長所得高低，成績高低可能有其它變數影響

回到父母對子女成績的影響

■ 哪些是與考試成績高度相關的家庭因素？

- 父母教育程度高
- 家庭關係親密
- 父母社經地位高
- 最近搬到較好的社區
- 母親生第一胎時 30 歲以上
- 小孩出生時體重偏低
- 小孩參加過學前輔導
- 母親在小孩出生後到上幼稚園前沒有上班
- 父母在家中說英語
- 父母會定期帶小孩上博物館
- 小孩為領養
- 小孩常挨打
- 父母參與學校家長會
- 小孩常看電視
- 家裡有很多書
- 父母幾乎天天唸書給小孩聽

父母對子女成績的影響

重要的是家長「是」怎樣的人，而非家長「做」了什麼

家長「是誰」高度相關	家長「做什麼」低度相關：
教育程度高	家庭關係親密
社經地位高	最近搬到較好的社區
母親生第一胎時 30 歲以上	母親在小孩出生後到上幼稚園前沒有上班
小孩出生時體重偏低	小孩參加學前輔導
在家中說英語	定期帶小孩上博物館
小孩為領養	小孩常挨打
參與學校家長會	小孩常看電視
家裡有很多書	幾乎天天唸書給小孩聽

養父母的影響？

- 養父母通常較親生父母聰明、教育水準、收入也較高，但這些優點通常對養子女的學業成績**沒有貢獻**
- 然而養子女上大學、從事待遇高的工作、成年後結婚的比率較高，顯示養子女成年後能擺脫純由 IQ 所預測的命運軌跡

美國黑人罹患心血管病機率為何較高？

- 美國黑人得高血壓機率較白人高 50%
- 明顯的刺激因子：飲食、抽煙、貧窮等都無法解釋
- 加勒比海黑人高血壓率亦較高，但現居非洲的黑人，統計上患病機率則和美洲白人無異

美國黑人罹患心血管病機率為何較高？

- 哈佛經濟學者 Roland Fryer 的觀察



美國黑人罹患心血管病機率為何較高？

- 昔日奴隸貿易的篩選，可能是美國黑人心血管疾病罹患率較高的根本原因
 - 奴隸從非洲運送至美洲常中途死亡，脫水是主因
 - 「鹽敏感性」高的人，較不容易脫水，體質能留住鹽分，就能留住更多水分
 - 商人（舔臉）找出鹽敏感性高的奴隸，降低風險
 - 此種鹽敏感性體質是高度遺傳特徵

以資料來輔助誘因設計

誘因

「道德不會改變人的行為，價格才會！」

——歐巴馬總統經濟顧問 Austan Goolsbee

- 解決問題的基本步驟：瞭解特定情境下，所有相關人士的誘因
- 勿聽其言，而要觀其行

聽其言，觀其行

加州居民節約能源的原因？

■ 電話訪問：在您決定節能時，下列因素的重要程度？

3 ■ 省錢

1 ■ 環保

2 ■ 對社會有益

4 ■ 很多人正在做

聽其言，觀其行 (cont.)

- 田野實驗：登門拜訪，發放小標語掛在居民門上
 - 能源節約（對照組）
 - 節約能源，保護環境（道德動機）
 - 盡你的責任，替子孫節省能源（社會責任）
 - 節約能源也省錢（財務動機）

#1 和你的鄰居一起節約能源（從眾心理）

觀察才能得知真相

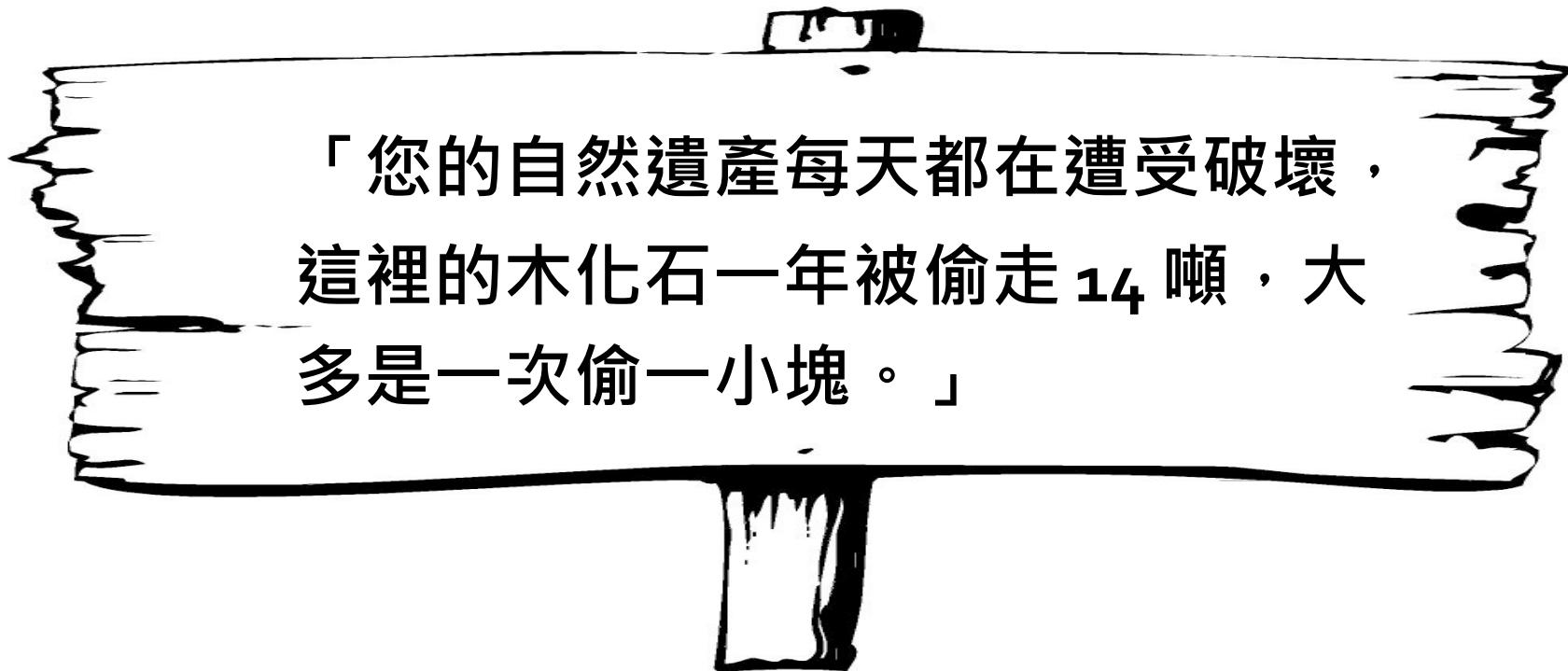
■ 十個常用美國房地產廣告字眼中，哪些與最終售價高度正相關？

- 絶佳 (Fantastic) ✓ 花崗岩 (Granite)
- 寬敞 (Spacious) ✓ 最先進 (State-of-the-Art)
- ✓ 可麗耐建材 (Corian) • “！”
- 迷人 (Charming) ✓ 饗宴 (Gourmet)
- ✓ 楓木 (Maple) • 環境優美 (Great neighborhood)

■ 分析 10 萬筆芝加哥郊區售屋資料

- 3,000 筆房仲銷售自宅，控制地點、屋況等變數後，平均銷售時間多 10 天，相同屋況最終售價高 3%

誘因：錯誤示範



- 美國亞利桑那州，化石森林國家公園的警示標語
 - 立有警告標語小徑的失竊率，是沒有標語小徑的 3 倍！
 - 錯誤的行為，因為很多人都在做，而被合理化了。

眼鏡蛇效應

- 印度被殖民時期，為減少當地眼鏡蛇數量，英國政府懸賞殺眼鏡蛇換獎金
- 越南被法國殖民也有類似的案例：減少鼠害
- 墨西哥波哥大為解決塞車問題，政府規定，每天只有部份車牌號碼可以上路，以降低車流量

慈善募款：我只煩你一次

- 微笑列車成立於 1949 年，到 2007 年止已為 76 個國家 38 萬的唇齶裂兒童提供免費治療，工作的重點地區是中國和印度
- 策略：「只要現在捐一次，我們將永遠不會再請您捐錢」
- 一般募款希望培養重複性捐款人，怎麼能為了短期進帳而犧牲長期捐款？



SmileTrain

為什麼我只該煩你一次？

Repayer

"I give to my alma mater"

"I support organizations that have had an impact on me or a loved one"

23% of donors

Casual Giver

"I give to well known nonprofits because it isn't very complicated"

18% of donors

High Impact

"I support causes that seem overlooked"

"I give to nonprofits I feel are doing the most good"

16% of donors

Faith Based

"We give to our church"

"We only give to organizations that fit with our religious beliefs"

16% of donors

See the Difference

"I think its important to support local charities"

"I give to small organizations where I feel I can make a difference"

13% of donors

Personal Ties

"I give when I am familiar with the people who run an organization"

14% of donors

慈善募款：我只煩你一次

- 微笑列車回覆卡選項：

- 1/3 • 「這是唯一一次捐款，請寄給我報稅收據，別再請我捐款」
- 2/3 [• 「我願意每年收到兩次微笑列車訊息，請尊重我的意願，限制寄給我的郵件數量」
• 「讓我知道微笑列車行動的最新進展，定期寄給我通訊」]

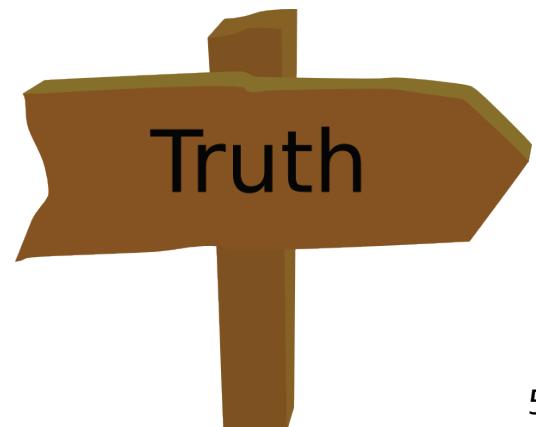
- 結果：

- 首次捐款的機率是一般DM的 2 倍，平均首捐金額也較高
- 整體捐款率竟然提昇 46%!



IN SEARCH OF THE TRUTH

- 保持赤子之心，不要帶入自己的假設或偏見
- 善用資料及統計工具
- 自然實驗可遇不可求，必要時設計實驗來驗證假設
- 會慢慢再接近「真相」一點...😊



創意人的訓練

- 創意的產生是有方法可循的

如何成為創意人

idea

陳昇瑋 / 中央研究院

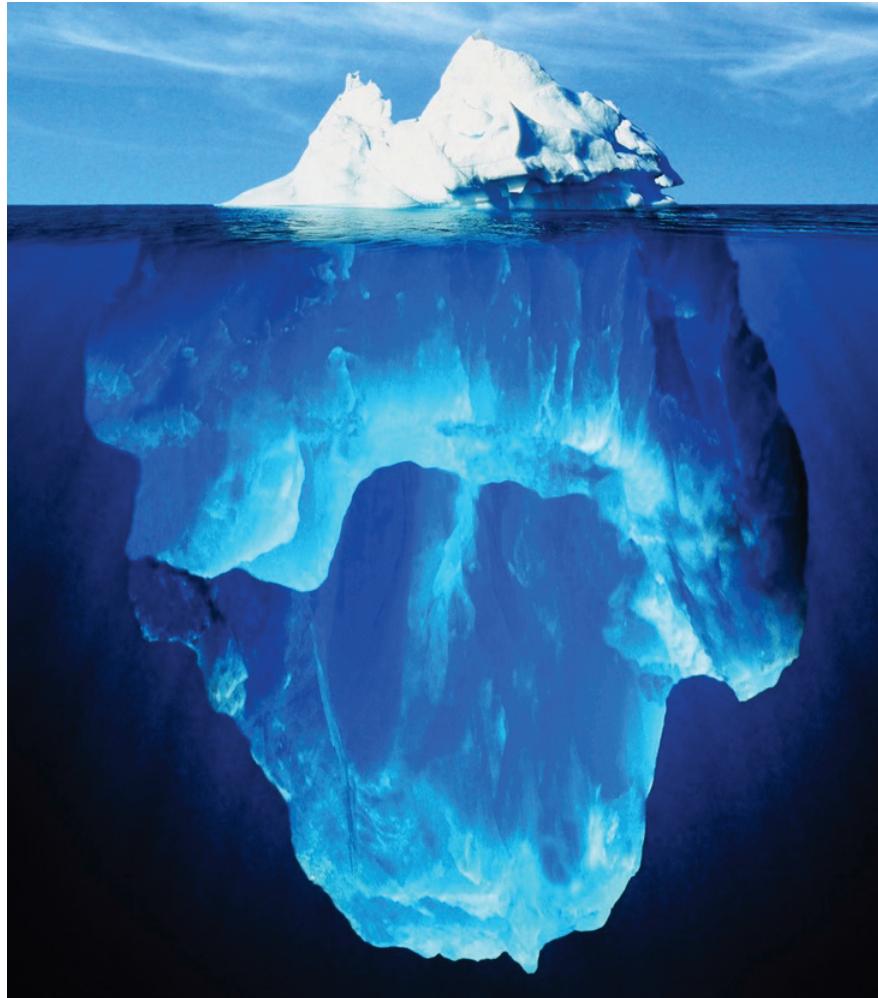
創意的發生



讀太多書會限制 創意嗎？

魔島理論

by James Webb Young (楊傑美)



魔島理論 - 別人怎麼說

“創意似乎不能超乎一個人的**經驗**之外”

“創意人就像乳牛，不吃草就分泌不出乳汁”

“百分之九十九的努力，加上百分之一的靈感”

但，靈感怎麼來的？

一定還有些因素決定它
冒不出海面.....

創意的形式 (1)

■ 拼圖遊戲

- 不相干事物的「相干性」
- 隨身聽：走路 + 音樂
- 果汁汽水：果汁 + 汽水
- 論文主題產生器？



■ 改變用途

- 不龜手之藥：染布工人→軍隊
- 心理學、社會學→廣告業（爭取消費者）→政治

創意的形式 (2)

■ 階段再定義

- 眼光是新的，東西就是新的
- 創意不見得是改變東西，有時候只是改變自己
- 「認知的改變」是重要的創新來源
- 情勢律 (Law of Situation)
 - 年代影視：製作者 → 提供者 → 規劃者
 - 窗帘 → 調節光線
 - 影印機 → 辦公室自動化
 - 大賣場 → 商品訊息 / 遊戲休閒
 - 手機, Google, Facebook, ...

創意的自我訓練



巴黎司機訓練法

- 強迫自己觀察，直到觀察成為生活的一部分
 - 上班休息時間，觀察每一位同事打電話的姿勢
 - 用餐時間，觀察每一位食客吃飯的細節
- 當你看的東西與人不同，你想的東西也就與眾不同



杜拉克式問句

- 簡化問題，並集中精神於真正的問題上
 - 問題要淺
 - 問題要清楚
 - 判斷問題的重要性
- 「好的問題，就等於答對了一半」
- 我來說一個故事 ...



“What if ...” 訓練法

- 給自己大膽的假設，試想各種可能的狀況
- 如果...會怎麼樣...
 - 如果台灣持續乾旱，我們生活用水該怎麼辦？
 - 如果不小心睡過頭了，上班遲到會怎麼樣？



反分析訓練法

- 分析與綜合，要彼此互相支援
 - 分析是「**同中求異**」，把看起來相同的東西說成不相干
 - 綜合是「**異中求同**」，把看起來不同的東西說成相關
 - 運用「分析」的能力，將東西拆成不同成分，再運用「綜合」的能力，將這些成分重新排列組合

- 試著找出兩個(看似)不相干事物的共同之處
 - 戒指 vs. 仙人掌
 - 音響 vs. 茶杯
 - 信用卡 vs. 早餐
 - 皮夾 vs. 螞蟻
 - 鉛筆 vs. 溜滑梯

重新定義訓練法

- 創意的來源，有時只是「認知的改變」
- 如果解釋是新的，舊的東西也能變成新的
- 漸距推遠
 - 賣豆漿的人 → 供應早餐的人 → 供應外出人士方便快速用早餐的人
- 平行重定義
 - 百貨公司擁有者 → 建築物的地主
 - 賣東西給消費者的商店 → 為消費者選擇生活用品的人

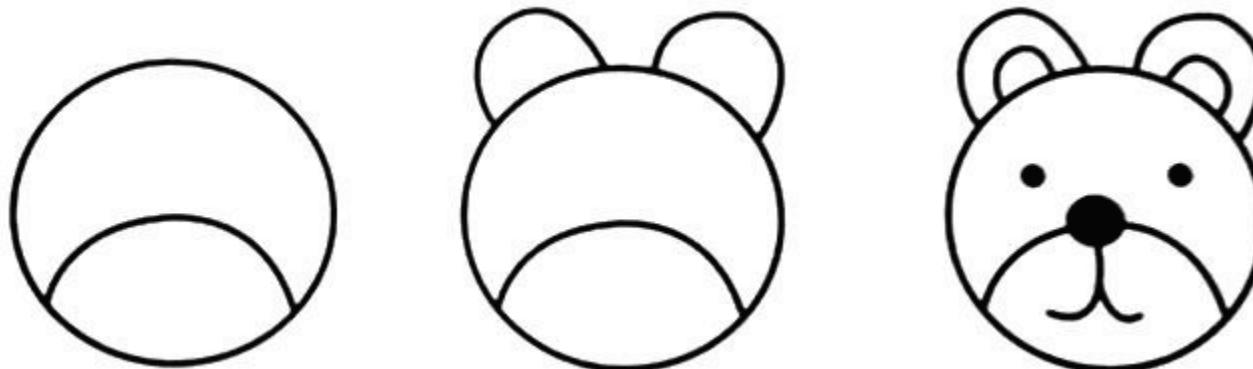


創意的產出

三個階段

“ 預備期 → 潛伏期 → 發光期 ”

- Helmholtz (德國哲學家)



三個階段 (cont.)



籌備 → 培養 → 靈感 → 事實驗證



- Hoshe F. Rubinstein (USC)



1. 收集原始資料
2. 在心裡咀嚼這些資料
3. 儘你所能的將主題拋開，把問題徹底忘掉
4. 不知道從哪裡點子就竄出來了
5. 將你新生的點子付諸實踐，然後看看它是不是會成功



- James Young (廣告人)

三個階段 (cont.)

“

把你自己的心沉浸在你正在進行的計畫中，達到一個飽和的狀態，然後開始等待。

並不是停下來休息或停下來開始看一個星期的電視，我說的是忘了它，去做別的工作。

”

- Lloyd Morgan

“

所有研究室的發現、發明都是經過一段時間的緊密思考和收集資料後，在放鬆的時刻以「靈感」的方式出現。

”

- C.G. Suits (GE)

三個階段 (cont.)

“

當卡在某個案子時，**就去做下一個**，讓非意識的部分來發揮功用。當你再回到這個案子時，你會很驚訝地發現，10 次裡有 9 次問題都解決了，你甚至不知道是怎麼解決的。

”

- Carl Sagan (天文科學家)

Incubation

“

(noun.)

1610s, "brooding," from Latin incubationem (nominative incubatio) "a laying upon eggs," noun of action from past participle stem of incubare "to hatch," literally "**to lie on, rest on**," from in- "on" (see in- (2)) + cubare "to lie" (see cubicle). The literal sense of "**sitting on eggs to hatch them**" first recorded in English 1640s.

<http://dictionary.reference.com/browse/incubation>

如何產出好構想？

量中取質

“

在一切相等的前提下，每單位時間內，若有人能產生很多構想，則得到好構想的機會比別人大。

”

- J.P. Guilford

“

得到一個好構想的最好方法，就是要有很多構想。

”

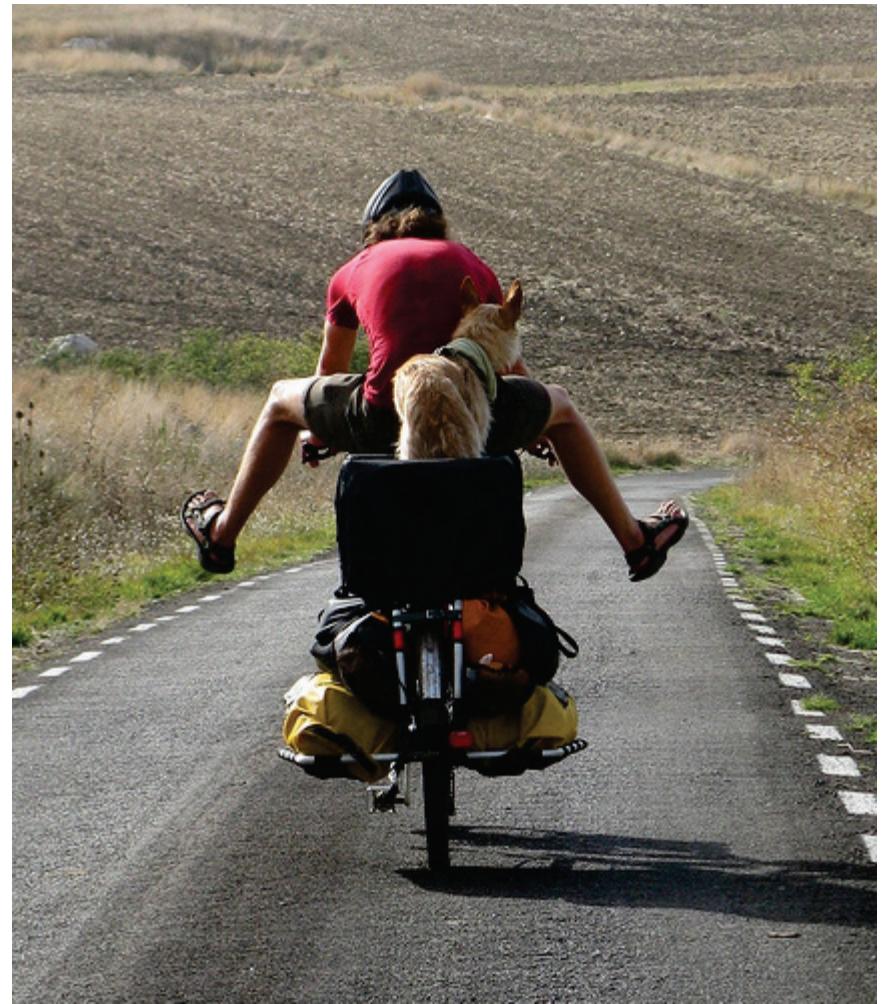
- Linus Pauling (Nobel Laureate)

如採珠者採蚌

- 愛迪生
- 5000 產品名 (70 人)
- 610 個書名
- 100 個社論標題
- 3800 個橋名



自由運轉



“Free wheeling” (coined by Dr. J. R. Killian Jr.)

自由運轉模式

“

如果理智對意念檢核得太緊密的話，創造性的意念就將躲藏起來。

”

- Friedrich Von Schiller

- 不要同時踩煞車和踏油門，不對任何觀念做任何評斷
- 儘量想出一大堆構想，儘可能以最快的速度將其列出。
- 搭便車 → on top of others' ideas
- 反面思考
- 唯一目的：「數量，數量，更多的數量！」

面對未知的時候

“

當你不確定一個問題是否有答案，要找答案就難了；當你知道**有很多答案**，要找到一兩個就容易多了。

”

- Emile Coue (法國心理學家)

“

當一個科學家面對一個問題，他確定有答案時，他的態度就轉變了，那等於已經找到 50% 的答案。

”

- Norbert Wiener (數學家)

再十個點子再去吃飯！

■ 瑞士刀的戶外看板廣告

- 一天的時間夠不夠？
- 午休時間夠不夠？



「以量求質」的其它形式

“

當我年輕時，我發現所做的十件事情中，失敗的總有九件。我不想成為失敗者，所以我總是做十倍的事情。

”

- George Bernard Shaw

我的點子筆記本



創意的絆腳石



血統主義

- 「這是不可能的；大家都不這麼做。」
- 抗拒新元素的加入，每一件事都應遵循既有的規則
- 我們的創造力因而受到阻礙，破壞我們思考的流暢性和彈性



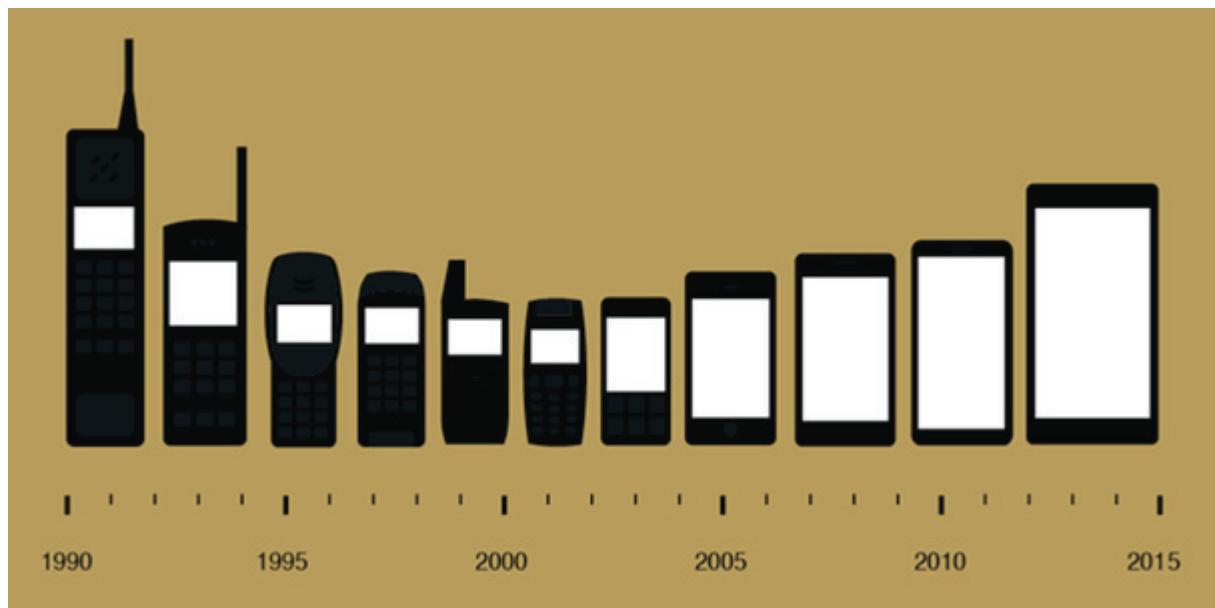
逆變心理

- 習慣領域 「從前，我有一次...」
- Comfort zone
- 我們得學習與「改變」一起生活



直線主義

- 不介意加入新元素，但總以為新元素加入的變化都是循著直線前進
- 歷史往往不是直線發展的
 - 電腦→大→小→便宜→快速→多功能
 - 電視→尺寸→成像品質→高傳真



如何毀掉一場動腦會議

- **讓老闆先說**：只要老闆先說，就註定這場動腦會議失敗了，因為大家會傾向猜測與說出老闆喜歡的方向
- **大家輪流依序發言**：大概輪個一次或兩次就結束了
- **只讓專家或技術人員發言**：動腦會議最好由不同性質的人組成，匯聚各領域人才，理想人數約為 5~8 人，如果成員中有與主題有關的專家，比例為半數以下較為恰當，因為集合各領域人才，對於擴大發想內容更有幫助。
- **遠離辦公室**：在海灘想出來的點子通常會離題太遠
- **不允許笨想法**：如果每個想法都要能實行才能提出，我敢保證這場動腦會議會超級冷
- **一五一十記錄會議內容**：只要記錄重點與建議事項即可，而且不可由主持人擔任。

二十條創意守則 by Charles Thompson

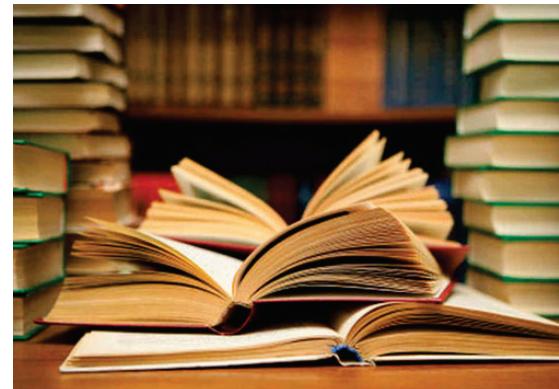
1. 只要想出走在時代前十五分鐘的點子，不必想出比時代早幾個光年的點子。
2. **得到偉大點子的最佳方法，就是先想出許許多點子，然後再把壞點子淘汰。**
3. 不要只尋求唯一的正確答案。
4. 如果一時想不出來……暫時休息一下。
5. **一想到點子，馬上紀錄下來，免得忘記。**
6. 如果每個人都認為你錯了，你就比他們早了一步；如果每個人都取笑你的點子，那麼你就比他們早了兩步。
7. 當你提出一個笨問題時，通常可以得到一個聰明的答案。
8. 每個問題都有答案，只要問對問題，答案自然顯現。
9. 絕對不要以最基本的看法來解決問題。
10. 在問題未解決之前，先想像困難解決之後的景像。
11. 成功的創意家通常用反證法來解決問題或發想創意。
12. 向傳統想法挑戰，可化不利為機會點。
13. 如果套上不同的鞋子不管用的話，試著從直昇機或太空船上看待事情。
14. 用大自然的角度觀看目標或問題，可大大提昇眼界，得到不同的解決方案。
15. 把握擷取別人一流的創意原則，精益求精。
16. **對失敗的懲罰，絕對不可重於對不做任何事的懲罰！**
17. 通常點子的有趣特質導向創新，而非正面或負面評價。
18. **把你的點子寫下來，就像把錢存在銀行裡。**
19. 在六十分鐘會議前，請做一分鐘頭腦熱身運動。
20. 把洗澡當作一件樂事吧！也許就在你刷刷洗洗.哼哼唱唱之間，靈感就來了。



個人建議

建議 #1 - 大量閱讀

- 先找一些名著墊底
- 要把閱讀範圍延伸到專業之外
- 應立足于個人靜讀
- 讀書卡片不宜多做
 - 書中真正深切觸動你的內容，想丟也丟不掉，對此你要有更多的洒脫和自信
 - 「早歲讀書無甚解，晚年省事有奇功」 by 蘇轍
- 有空到書店走走，逛逛圖書館也很好



[1] 余秋雨〈青年人的閱讀〉

建議 #2 - 不放過所有的發想

- 隨時可記錄，從來不刪除的筆記方式
- 定時瀏覽記錄，重新檢視所有的發想
- 一有機會就讓旁人幫忙驗證



建議 #3 - 杜拉克問句的練習

“ 做學問要於**不疑處有疑**，待人要於**有疑處不疑**。

”

- 胡適

- 一答接一問，在答案中起問題
 - 追根究柢
 - 務必追到問題核心
-
- 很棒的附加價值 - 再也不怕參加社交活動！

建議 #4 - 獨處與熱情

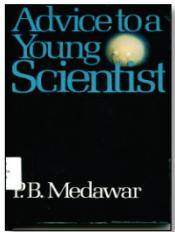
“ 舒適的**獨處**，快樂的孤寂對原創性思想有多麼大的幫忙啊。 ”

“ **熱情**會提高我們的知覺力，讓我們能體會最細微的表現。就像一個戀人，每天在他的愛身上，都可以發現新的事物。 ”

一個創新人才需具備的基本特質！

講了這麼多，都是在分享我們這些年做事情的經驗及理念。常常有人問我們，怎麼樣的人適合做我們這行？首先一定要是總是充滿好奇心（Always Curious），同時具備同理心（Be Empathetic），謙虛（Stay Humble）。一直很喜歡日本人用「初心」這個講法，外面有太多事情我還不懂，有太多事情我還要嘗試，要永遠保持一顆初心面對世界。真正了解以上這一切你才能去創造價值。

延伸閱讀



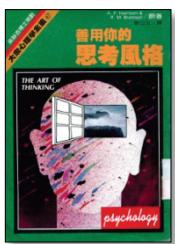
■ Advice to a Young Scientist

P.B. Medawar, BasicBooks, 1979.



■ 科學之路：科學家的心路歷程

貝弗里奇 著/ 楊新北 譯, 長堤出版社, 1984.



■ 善用你的思考風格

■ 哈里森(Harrison, A. F.), 布朗森(Bramson, R. M.) 著/廖立文 譯,
遠流出版公司, 1985.



■ 創造與人生

■ Robert Olson 著/呂勝瑛, 翁淑緣 譯, 遠流出版公司, 1985.

延伸閱讀



應用想像力

Osborn, Alex Fraickney 著/邵一杭譯, 協志工業叢書, 1987.



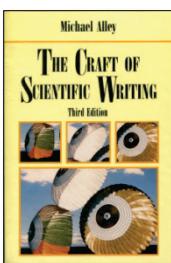
The Grace of Great Things

Robert Grudin, Ticknor, Fields, 1990.



如何撰寫零錯誤程式

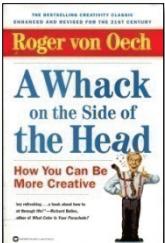
Steve Maguire 著/施威銘研究室譯, 旗標, 1994.



The Craft of Scientific Writing

Michael Alley, Springer, 1996.

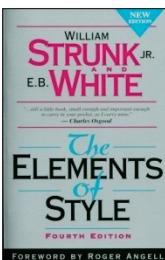
延伸閱讀



- A Whack on the Side of the Head
Roger von Oech, Warner Books, 1998.



- 創意人：創意思考的自我訓練
詹宏志, 臉譜文化, 1998.

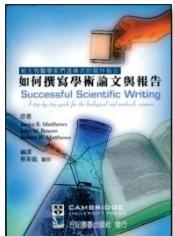


- The Elements of Style
William Strunk Jr., Longman, 1918.



- 文案自動販賣機：第一本本土廣告文案寫作指南
楊梨鶴, 商周出版, 2000.

延伸閱讀



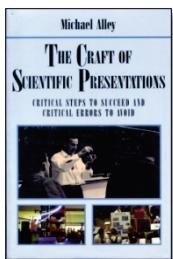
如何撰寫學術論文與報告

- Janice R. Matthews, John M. Bowen, Robert W. Matthews 著 / 蔡東龍 譯, 合記圖書出版社, 2002.



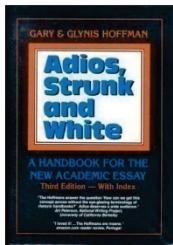
如何閱讀一本書

- Mortimer J. Adler, Charles Van Doren 著 / 郝明義, 朱衣 譯, 台灣商務印書館, 1972.



The Craft of Scientific Presentations

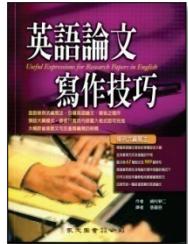
Michael Alley, Springer, 2003.



Adios, Strunk and White

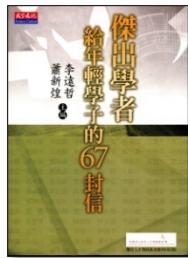
Gray, Glynis Hoffman, Verve press, 2003.

延伸閱讀



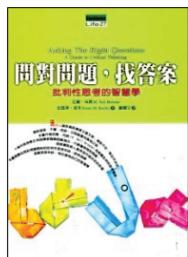
英語論文寫作技巧

崎村耕二 著/張嘉容 譯, 眾文圖書公司, 2003.



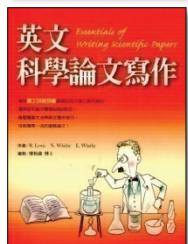
傑出學者給年輕學子的67封信

李遠哲, 蕭新煌, 天下文化, 2003.



問對問題，找答案：批判性思考的智慧學

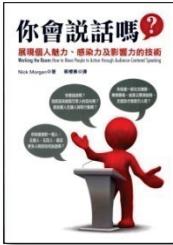
M. Neil Browne, Stuart M. Keeley, 商智文化, 2006.



英文科學論文寫作

R. Lewis, N. Whitby, E. Whitby, 眾文圖書公司, 2007.

延伸閱讀



你會說話嗎

Nick Morgan 著/蔡櫻素 譯, 臉譜文化, 2006.



研究科學的第一步：給年輕探索者的建議

Santiago Ramon y Cajal, 究竟出版社, 2007.



撰寫論文的第一本書

周春塘, 書泉出版社, 2007.



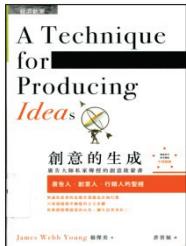
英語論文〔句型、片語〕表現集

小田麻里子, 味園真紀 著/馮慧瑛 譯, 眾文圖書公司, 2007.

延伸閱讀



- 英文研究論文寫作文法指引
廖柏森, 眾文圖書公司, 2007.

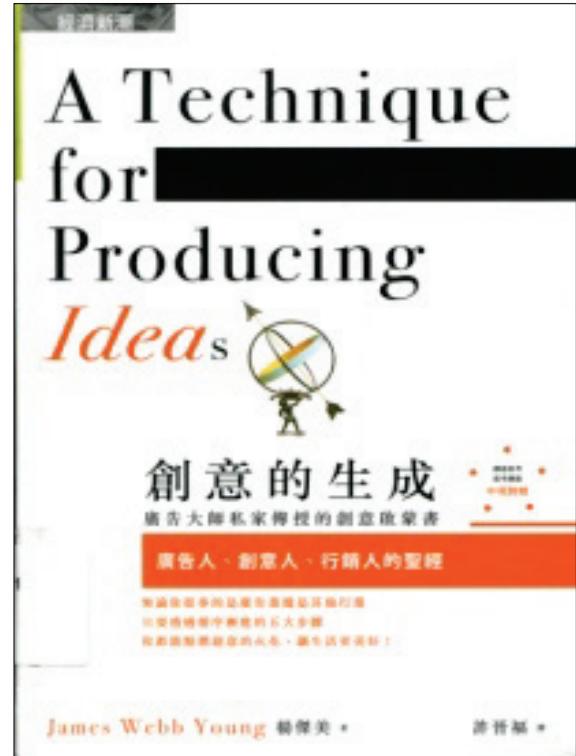
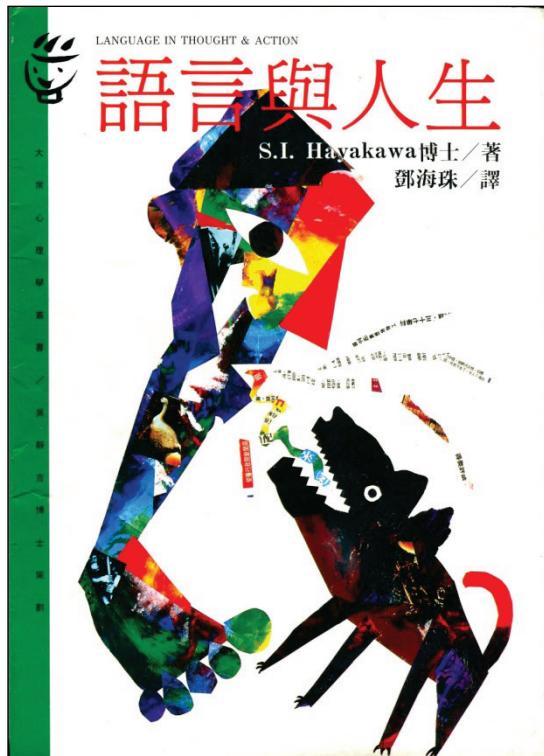
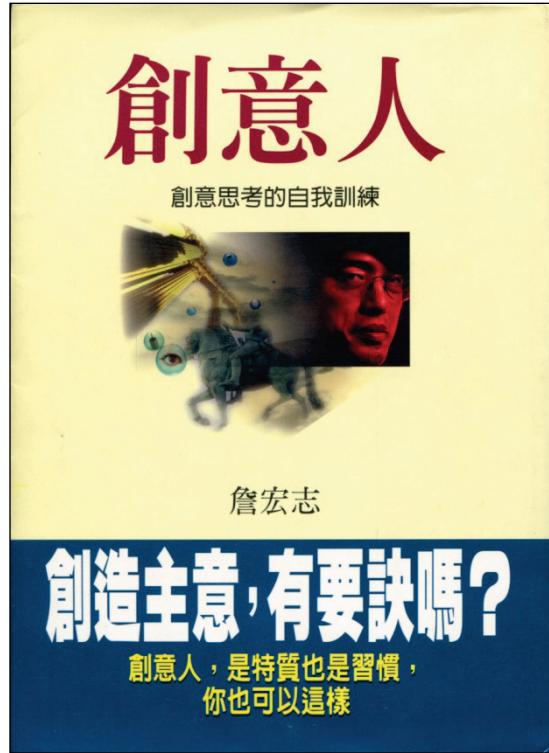


- 創意的生成
楊傑美 著/許晉福譯, 經濟新潮社, 2009.



- 語言與人生
S.I. Hayakawa 著/鄧海珠譯, 遠流出版公司, 1994

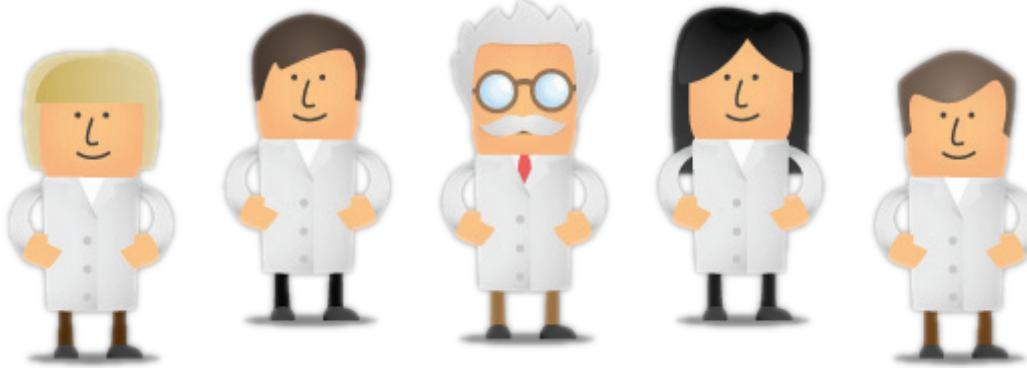
建議閱讀





交流時間





資料科學團隊的建立

陳昇瑋

中央研究院資訊科學研究所

今天談些什麼

- 資料科學團隊的組成
- 先佈置廚房還是先做菜？
- 如何分工合作？
- 企業組織及文化
- 社會物理學與企業管理

找不到有經驗的專家怎麼辦？

- 三個出發點：資訊，數學統計，問題領域
 - 專精一項就很不錯，專精兩項即少見
- 個人特質
 - 細心^① yet 富創意^③
 - 溝通能力^②
- 好的成員可以耳濡目染學習其它的面向；
好的領導者可以把不同面向的成員組合起來。

最小團隊組成

■ 理想的初始團隊規模



PM

Data
Scientist

Data
Engineer

Data
Engineer

Visualization
Designer

- 兩個不嫌少，先求有再求好
- 但也不要忽略 **Data Project Manager** - 對於資料分析技術及流程、目標設定能有掌握度的 PM

先佈置廚房還是先做菜？

「大」數據處理平台？

- 對於許多組織而言，「**大**」並非最重要的特質。
- 根據 2012 年由 New Vantage Partners 針對大型組織的五十名經理人所做的一項調查，在大公司裡，他們所處理的較屬於「**資料缺乏結構**」的問題，而非「**資料過於龐大**」的問題。
 - 30% 的大數據問題主要在於「必須分析來自多個來源的資料」；
 - 22% 的受訪者則主要聚焦於「分析新型態的資料」；
 - 12% 的人主要是「分析動態的資料串流」；
 - 只有 28% 的受訪者是以分析大於 1TB 的資料集為主要工作，且當中有 13% 是處理介於 1TB 與 100TB 間的資料集。

那一年我們一起追的 Hadoop



這頭象
也不好騎



有時候這是真的

Command-line tools can be 235x faster than your Hadoop cluster

Sat 25 January 2014 by [Adam Drake](#)

最小工作平台

- 隨機抽樣是我們的好朋友
- A workstation + R/Python is normally enough
- Use FULL dataset after analytics is finished over small samples
 - Except for deep learning and similar methods that require large-scale dataset to be effective

如何分工合作？

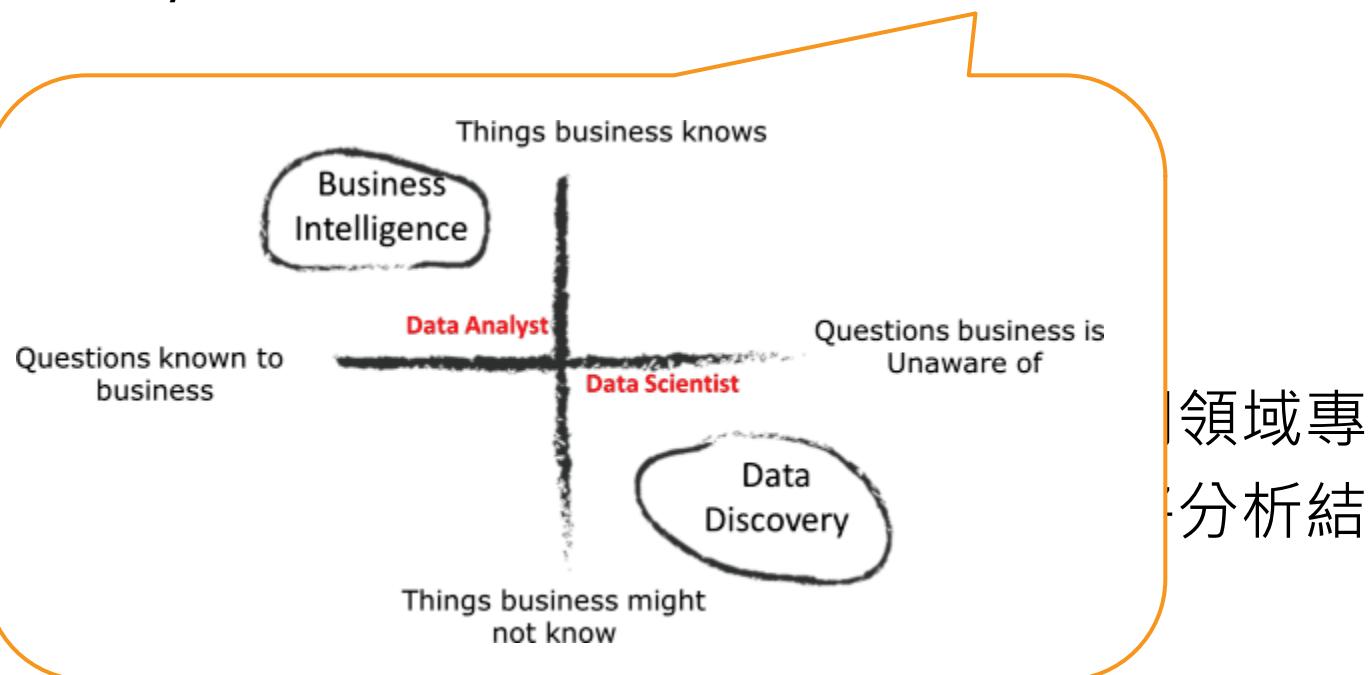
資料科學團隊 ≠ 資料倉儲團隊

■ 資料倉儲團隊

- 管理 / 整合資料
- 處理行銷 / 業務 / 管理團隊的資料 / 報表需求
- 資料庫 / 欄位 / 報表方式會變，但多數問題是事先定義的

■ 資料科學

- 資料倉儲
- 企業領導（家）定義
結果導入良



資料團隊與領域專家

- 領域專家負責**發問（或指出方向）**
 - 要問出對的/重要的問題是最困難的一件事
- 資料團隊負責**重新定義問題及尋找答案**
 - 問題的形式有時決定該問題能否得到解決
 - 拿掉人為的假設，找到最有效益的問題來聚焦
 - e.g., 怎麼提升利潤？

提升產品品質

加強包裝

加強行銷

降低生產成本

提升工作效率

找到對的人

提升回頭率

打壓對手 XD

資料科學團隊 ≠ 報表產生器

■ 授權團隊

- 把**報告撰寫和基本資料處理**從資料科學家的工作中剝離開來，讓他們可以集中於更有效的工作。

■ 培養對資料好奇的文化

- 教導**所有的員工**使用工具（例如儀表板），消除數據的壁壘，激發他們的好奇心，告訴他們每個人如何可以更好地利用數據。
- 類似行為有助於改變他們把統計報告當做是臨時請求的思想，可以解放資料科學家。

護才與養才

- 把資料科學家規範得太緊，他們不會有好表現。
- 與負責產品與服務的高階主管，而非督導業務職能的人建立關係。應該多花些時間參與技術社群/研討會及進行技術分享。
- 為公司增加的最大價值，不在於寫出報告或向資深高階主管做簡報，而是在**面對顧客的產品和流程上創新**。

混出資料科學家

- 「每周要跟管理業務的負責人吃兩次飯，最起碼兩次，這就是你的 KPI。」
- 商業敏感是要靠「混」出來的，它並不會憑空出現。更一般性來說，數據部的人要和業務部的人經常在一起，不只是一同開會，更要一起喝茶、吃飯。

資料團隊可以 scale up 嗎？



- 7人公司 → 3000人公司的過程
- 我們一開始是集中式的，團隊中提供互相學習的機會，保持一致的工作目標。

隨著時間的推移，我們被看成一種資源，被要求提供數據，而沒有能夠主動思考未來的機會。

- 所以我們決定用嵌入式的安排，我們仍然遵循集中的管理，但是我們打破了自己的小組，讓數據團隊的夥伴更直接同工程師、設計師、產品經理、行銷人員等等溝通。
- 我們的團隊成員在決策過程中被視為合作夥伴，而不僅僅是統計採集。

企業組織及文化

If you want to build a data organization,
everybody has to first believe in data.



資料必須是一等公民

- 資料不只是配角，不是程式設計師 debug 使用，也不是要符合主管機關的要求而收集 / 保存而已。
- 資料**收集**、**保存**及**提供**也是系統規格的一部分
- 由資料科學團隊在事前檢視資料收集完整性及品質，但要另有資料倉儲團隊來負責資料的整合維護

讓資料成為企業資產

- 企業資產，而非部門資產，或是沒爹沒娘的孤兒...
- 理想作法：程式 / 資料透明化 / 共有
- 可行作法
 - 所有資料由單一團隊統一管理
 - 資料團隊為戰略編組，高層全力支援
- 真實案例：以上皆非

Start with simple data analysis then moving to more complex ones



complex

made

simple

Draft Zero



Draft Zero



資料科學團隊 KPI

「你沒測量過的東西，是無法管理的。」

-- W. Edwards Deming

■ 績效量化

- 通常不是無成本的，需要額外的**投資**，且需要**時間累積**。
- A/B testing is our good friend
- 唯有如此，效果才能夠真實地呈現, e.g., # users, # session time, # transactions

KPI 的共享

■ 建立績效共享制度

- 資料蒐集團隊 / 資料倉儲團隊
- 發生 / 提出問題的團隊
- 實作資料產品的團隊



為什麼導入資料團隊這麼困難？

典範移轉 (Paradigm Shift)



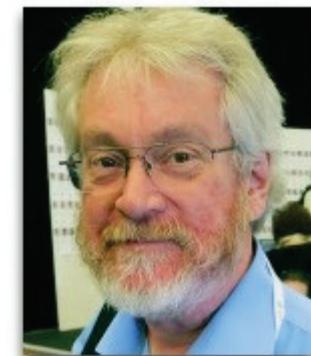
小結

- 資料科學團隊的組成有挑戰性
- 但要提供良好的工作環境讓資料科學團隊得以發揮，需要更大的變革及改造

社會物理學與企業管理

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

The screenshot shows the eToro + OpenBook platform interface. At the top, there's a navigation bar with the eToro logo, a 'OpenBook BETA' button, and login fields for Username and Password. Below the navigation bar, there are tabs for Traders Feed, Markets, Rankings, WebTrader, and Sign Up!.

In the main content area, there's a 'Start here' section with three steps: 1.See (See what real people are trading now), 2.Follow (Find and Follow Traders), and 3.Copy (Copy instantly the trades you like). It also features a 'Follow Top Ranked User astrex' section, which highlights a user who has gained 2743% in the past 3 months, winning 65.6% of all trades. A 'Start Now!' button is available to follow this user.

The central part of the screen displays a 'Live Trading Feed' showing recent activities made by real people. Two examples are listed:

- scharfetauben closed a **USD/JPY** Buy position, gaining 29.7% less than a minute ago from Germany. Buttons for Like, Comment, and Follow are shown, along with a 'Trade Story' button and a 'CLOSE' button.
- scharfetauben closed a **EUR/CHF** Buy position, gaining 54.6% less than a minute ago from Germany. Buttons for Like, Comment, and Follow are shown, along with a 'Trade Story' button and a 'CLOSE' button.

At the bottom of the feed, it says Actionaer11 Sold AUD/USD @1.0802. To the right, there's a 'Top Performers' section with a table showing the top 5 traders based on performance over 1w, 1m, and 3m periods:

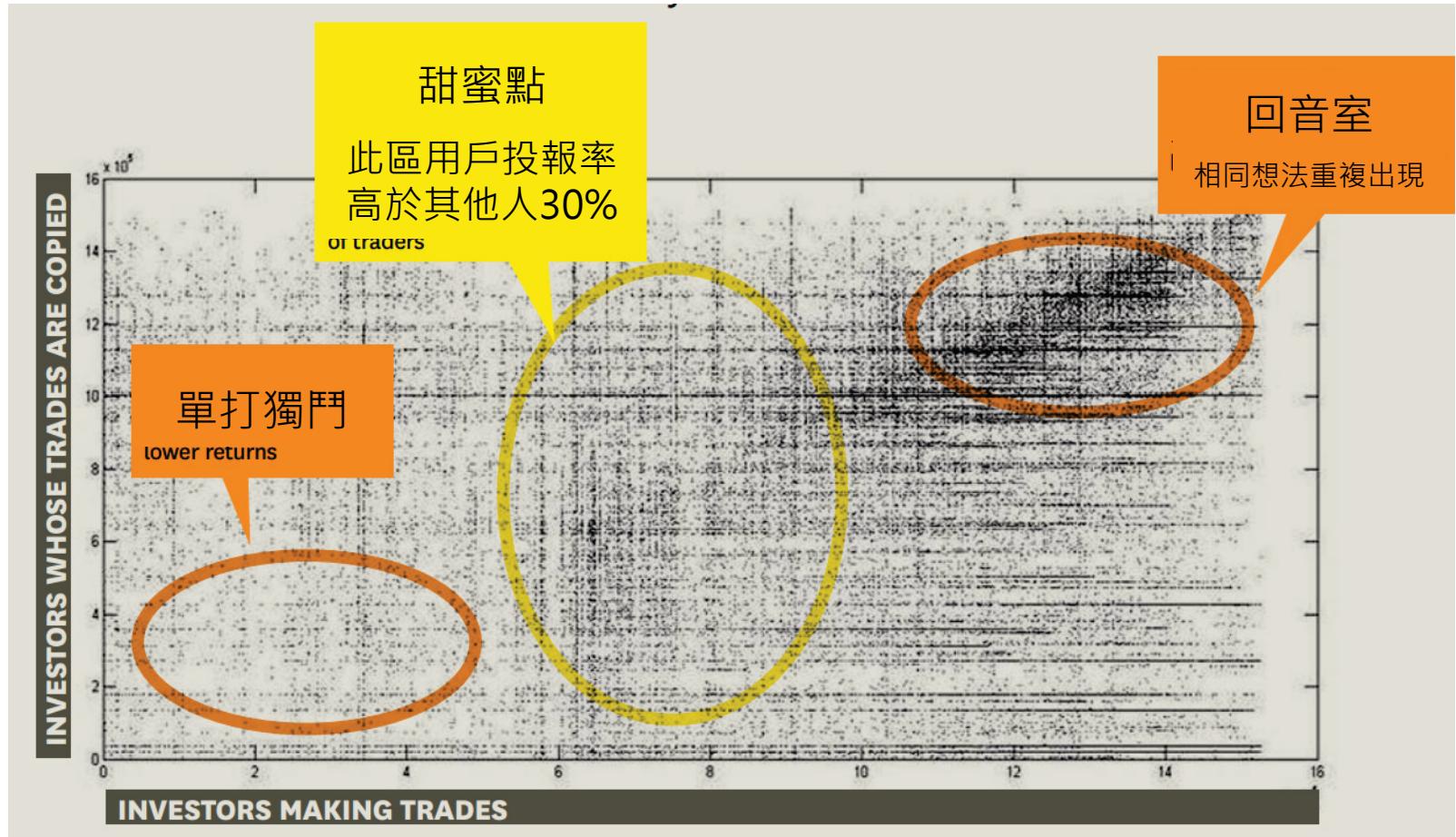
Rank	User	Performance	Location	Trades
1	verygoodtrade	743%	Singapore	170/224
2	wwleong1981	354%	Brunei	45/64
3	astrex	348%	Italy	15/17
4	simparker	245%	Australia	17/23
5	ciampino43	240%		

www.trademaker.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

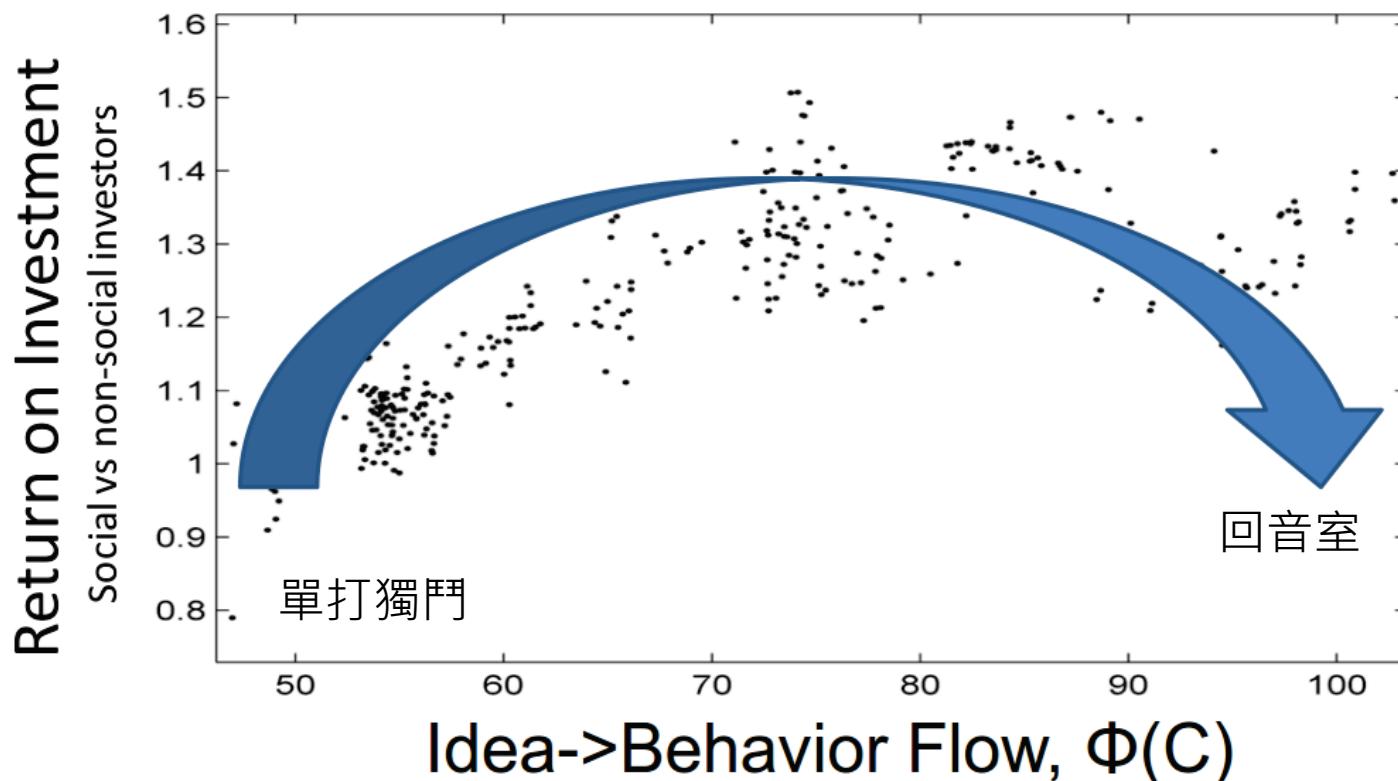
社會學習的證據



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

Idea Flow vs. RoI

Decision Accuracy Depends on
Diversity of Information Sources



量化群體智慧

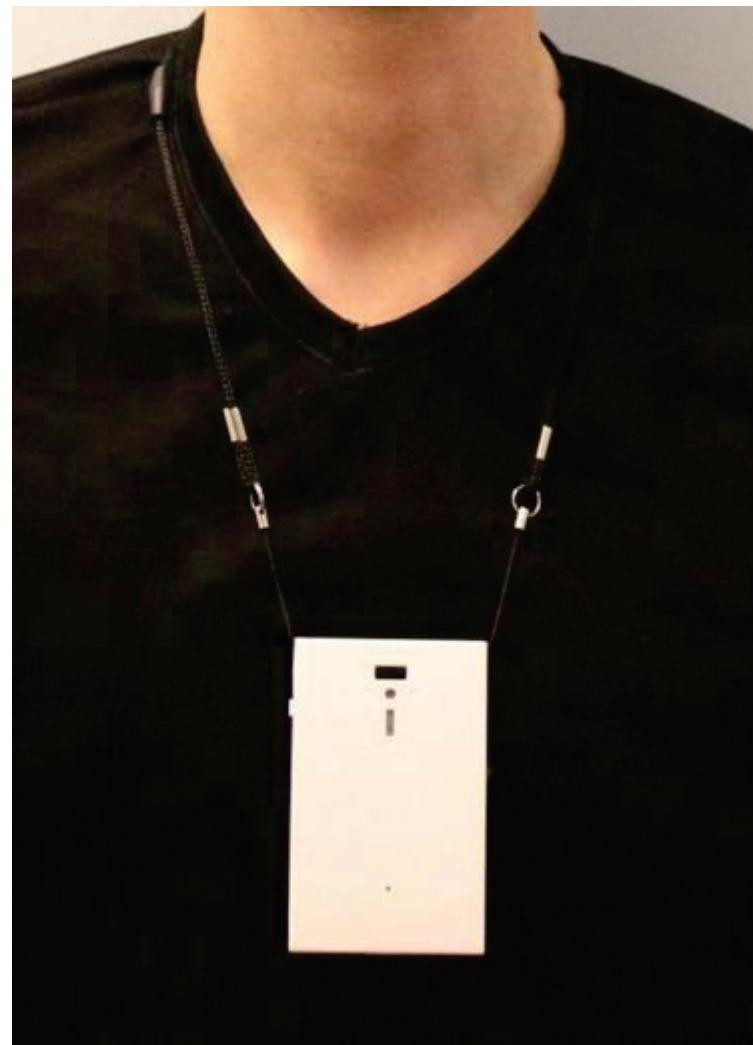
- 為什麼有些企業比其它企業來得有開創性？
- 決定群體表現的因素
 - 專業能力？
 - 凝聚力？
 - 成就感？
 - 薪水？
 - 領導者風格？
 - 文化？

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



伺服器銷售公司

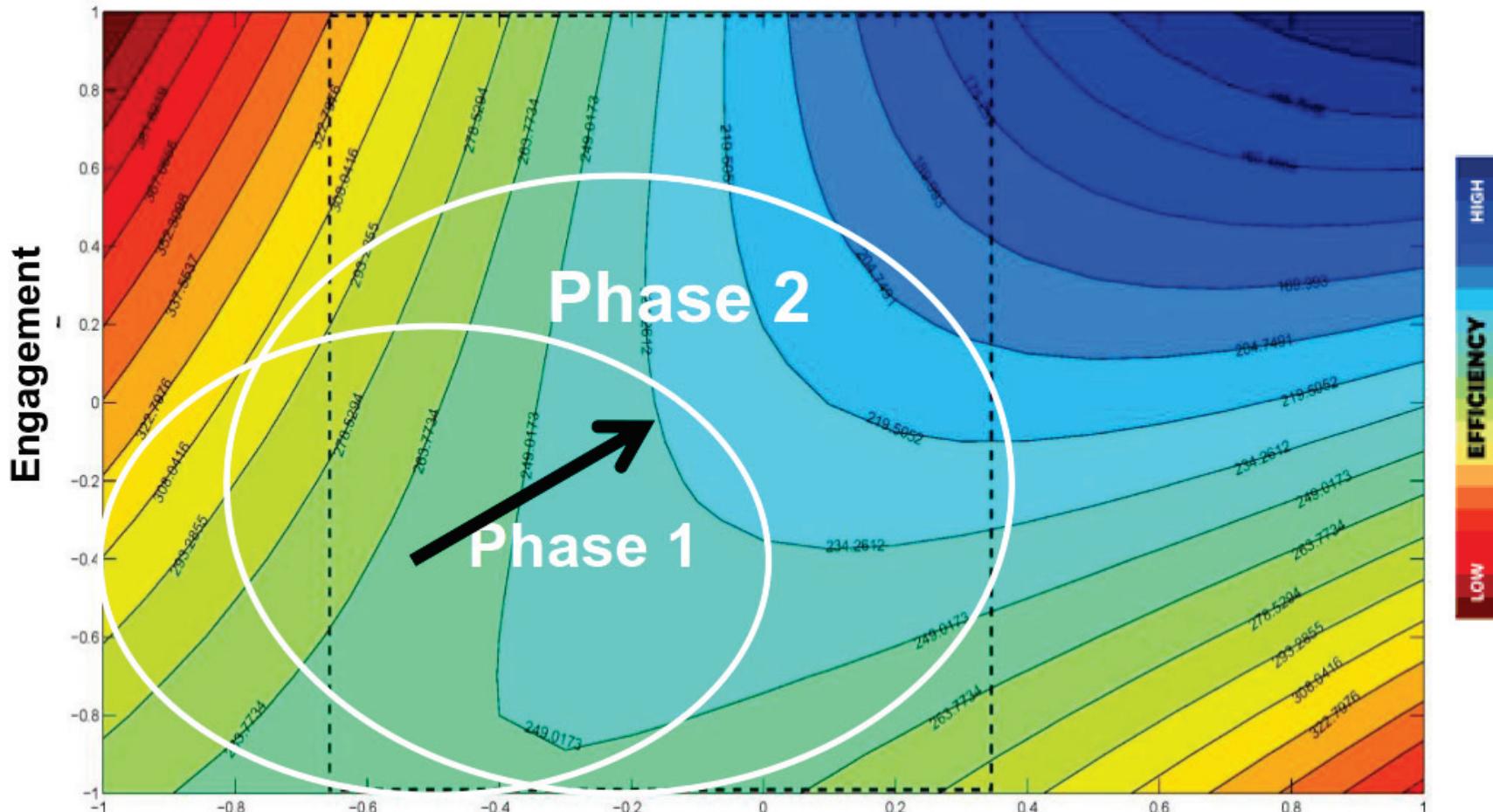
- 為期 1 個月，23 人，約 1,900 小時的互動觀察
- 客製化訂單任務派工：紀錄任務開始和結束的確切時間 → 衡量每名業務助理每項任務的確切花費時間
- 參與程度排名前 1/3 的員工
→ 生產力較一般員工高出 10%

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

Average Call Handling Time



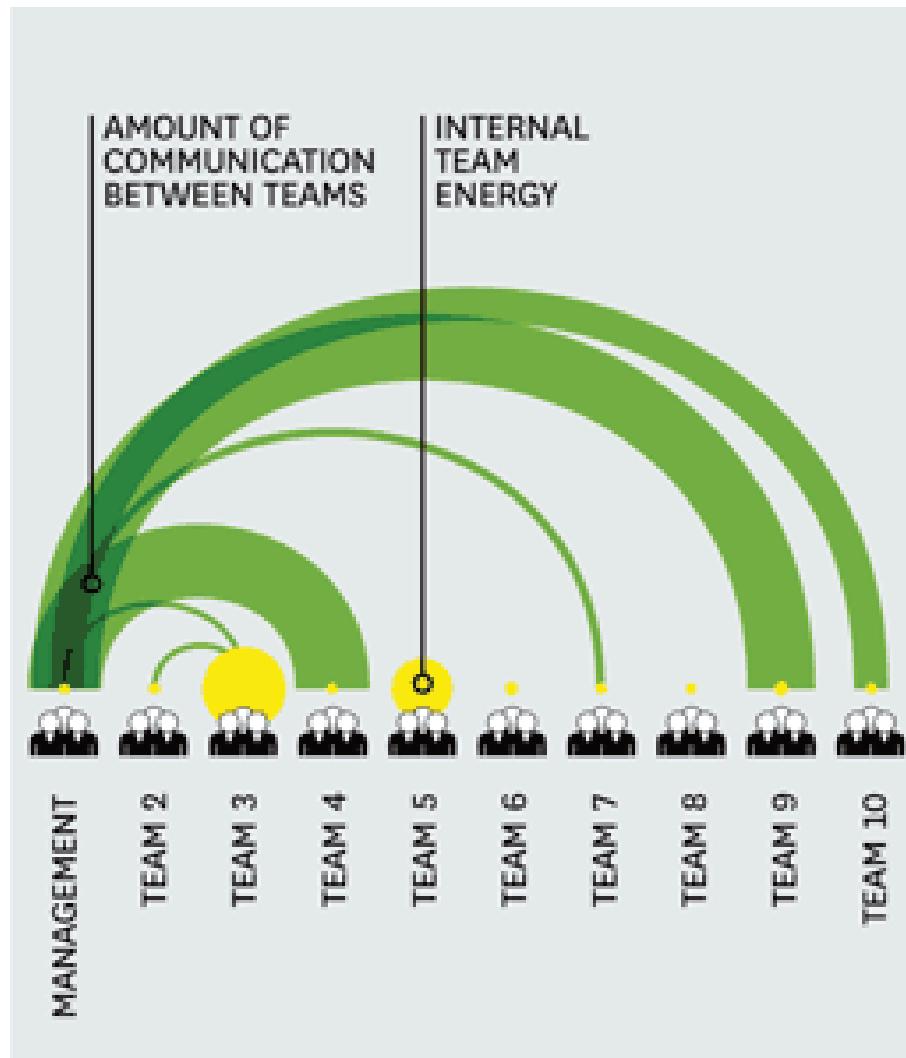
Copyright alex pentland 2012 all rights reserved

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

Olguin, Waber, Kim, Pentland

量化團隊參與及探索

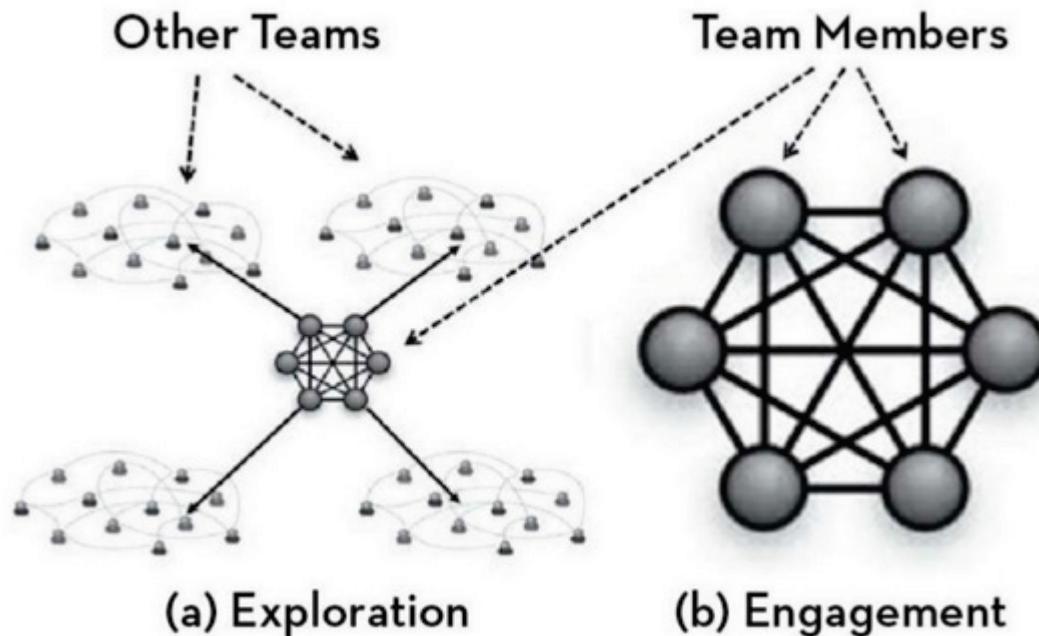
- 參與 (engagement)
團隊內的互動
- 探索 (exploration)
跨團隊的交流



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

參與和探索行為

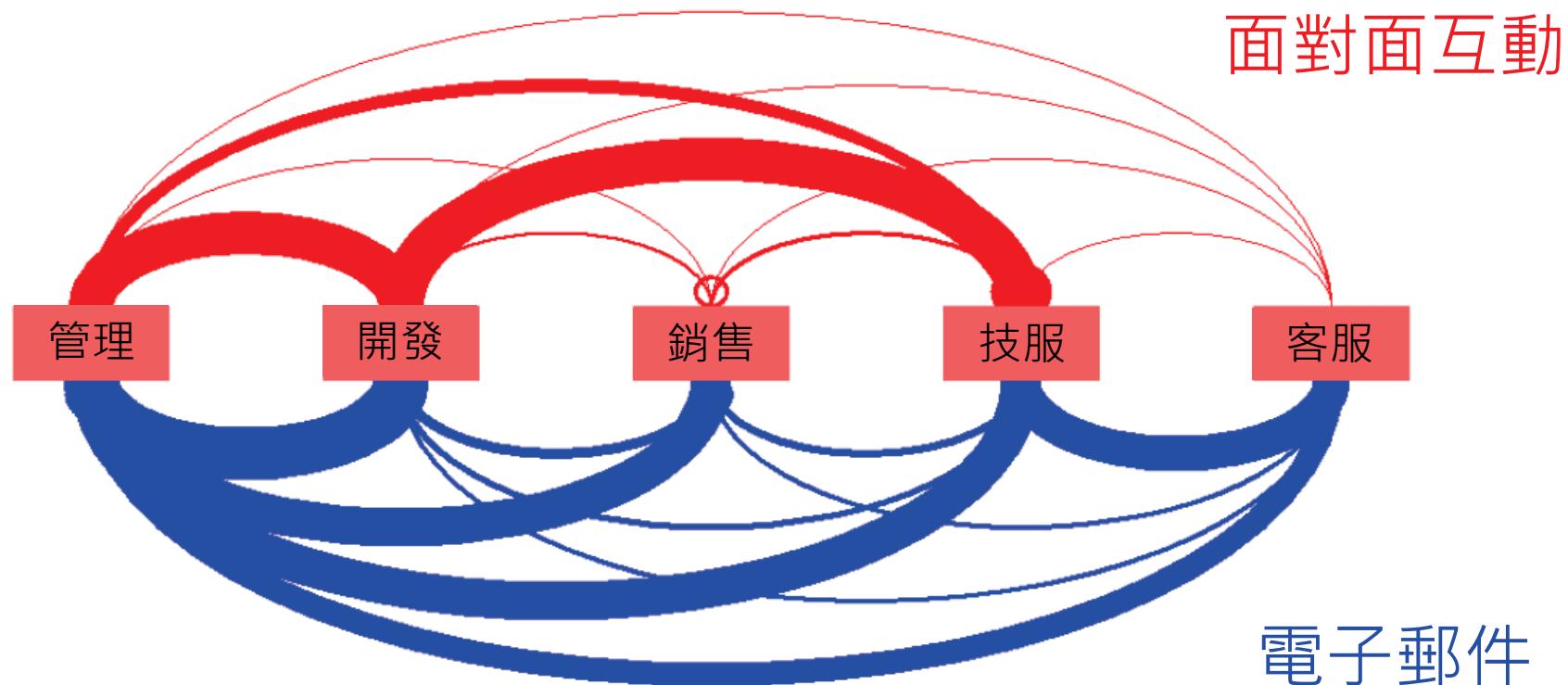
- 星狀網路：產生團隊以外的新意念流
- 密集互連：豐沛互動，有助檢視新意念，並融入團隊的規範和習慣之中



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。



一個典型的企業架構 (cont)

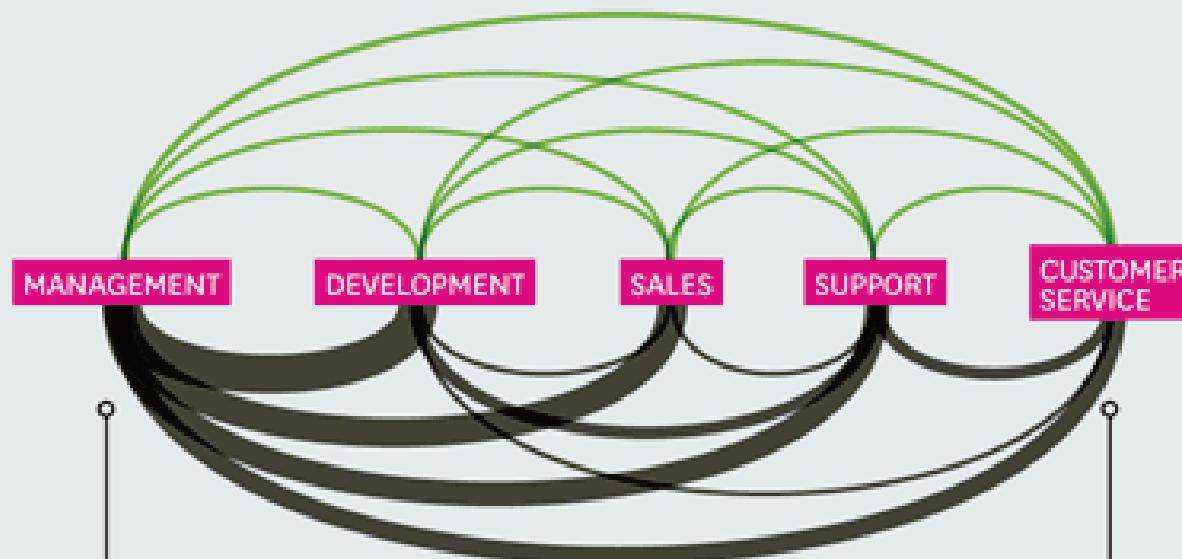
- 設計新行銷專案的團隊在探索和參與兩種模式間擺盪
- 負責製作的團隊則否，主要是團隊內部互動，新想法很少流入
- 意念流黑洞
 - 其他部門很少與客服部門面對面交談
 - 可能解法：改變座位安排，確保所有人都在互動交流圈中，得以改善部門間協調問題

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

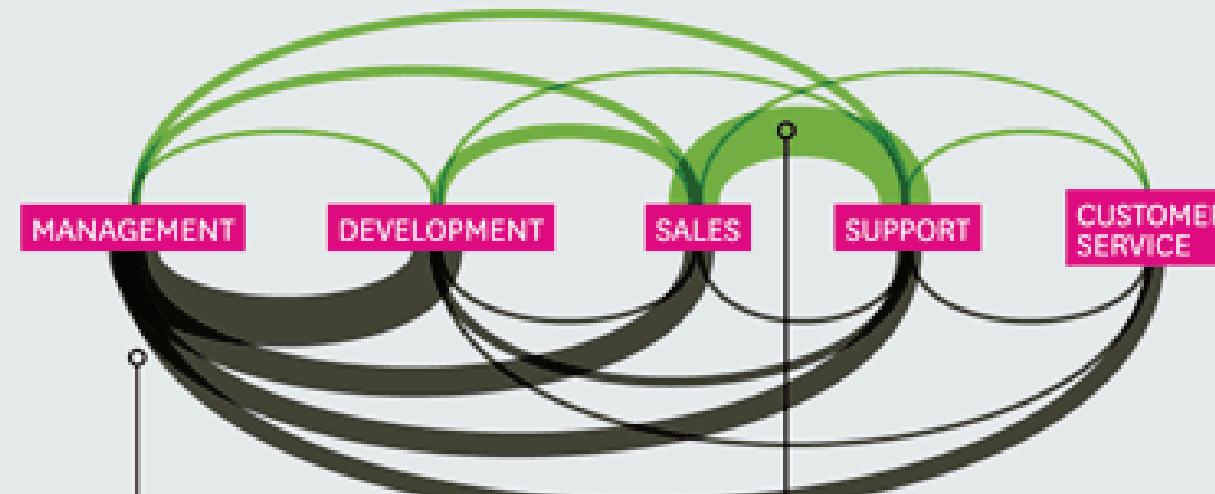
DAY 2 MANAGEMENT IS CLEARLY DOING MOST OF THE COMMUNICATING.



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

僅銷售和支持部門有較多當面溝通

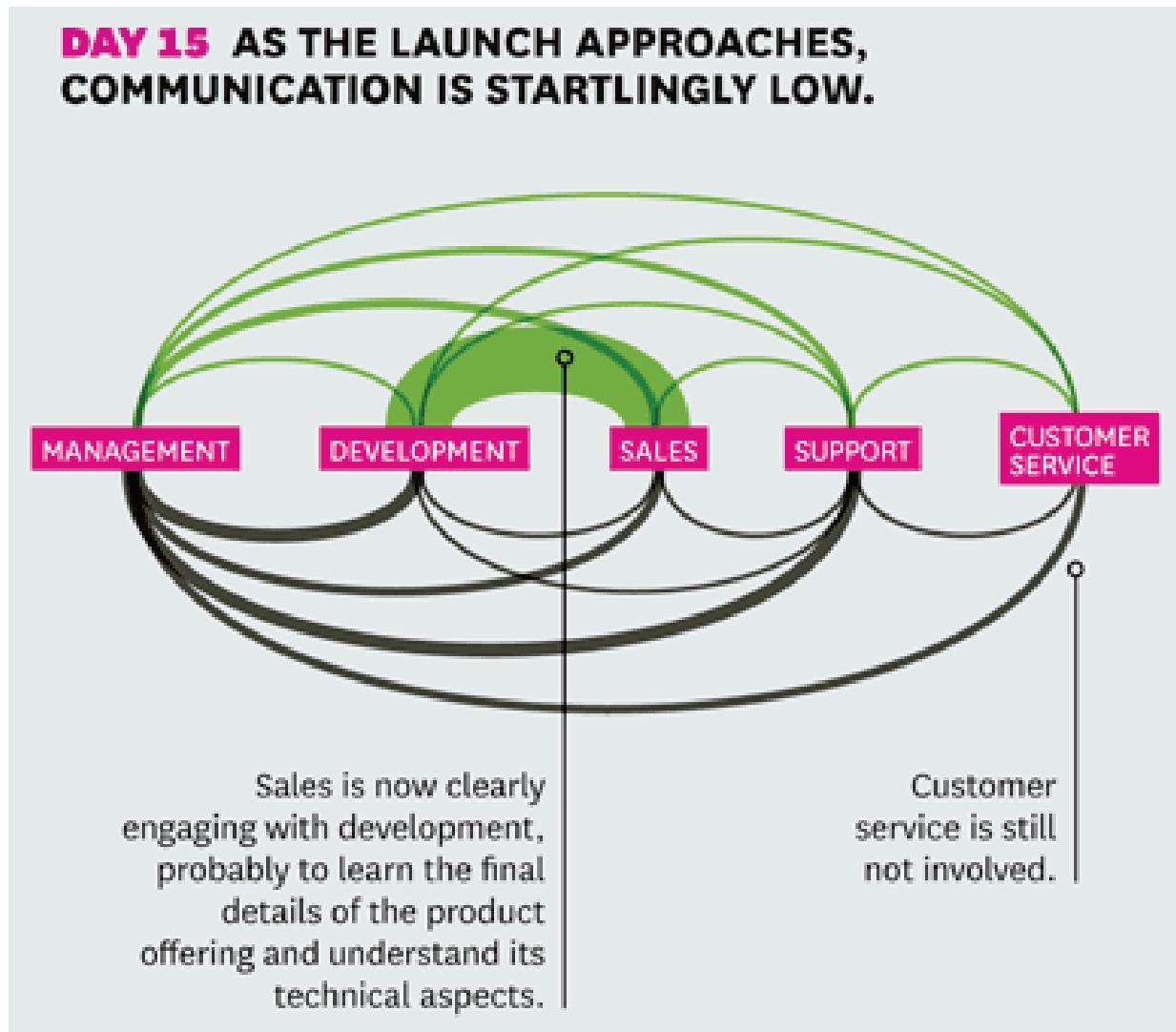
DAY 6 MANAGEMENT BY E-MAIL CONTINUES.



Management is communicating face-to-face a little bit with every team except customer service, and most groups aren't talking much to one another.

Only sales and support interact with each other a lot in person—most likely because they are prepping for the launch.

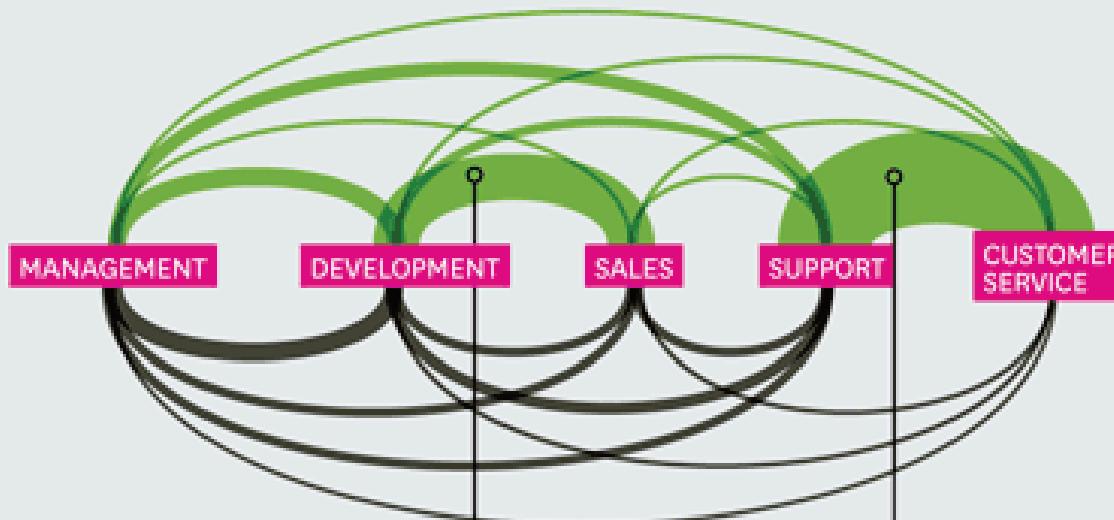
接近結案期限，面對面互動量大幅降低



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

DAY 23 TWO DAYS AFTER LAUNCH, TEAMS ARE FINALLY COMMUNICATING IN PERSON, AS THEY TRIAGE A DISASTROUS CAMPAIGN.



For the first time, e-mail communication is lower than face-to-face communication. In a crisis people naturally start talking more in person.

The big jump in communication here might be a result of sales' hammering development about why the product isn't working and how it can be fixed.

Customer service and support are locked in all-day meetings trying to patch the problems.

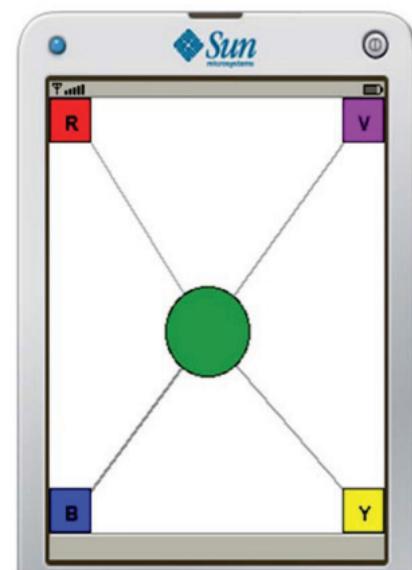
hbr.org/resources/images/article_assets/hbr/1204/R1204C_B_LG.gif

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？

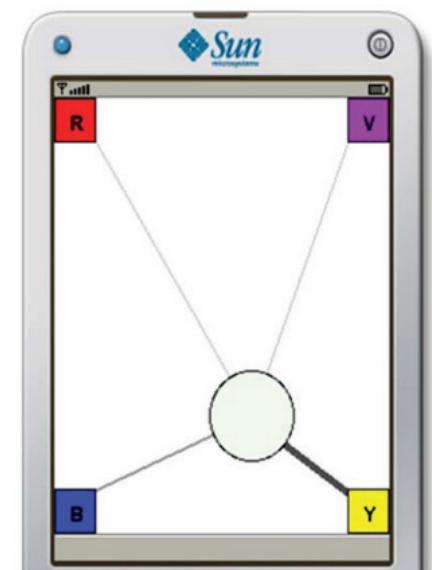
不僅是觀測，希望進一步改善

- 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



(a)

參與程度高



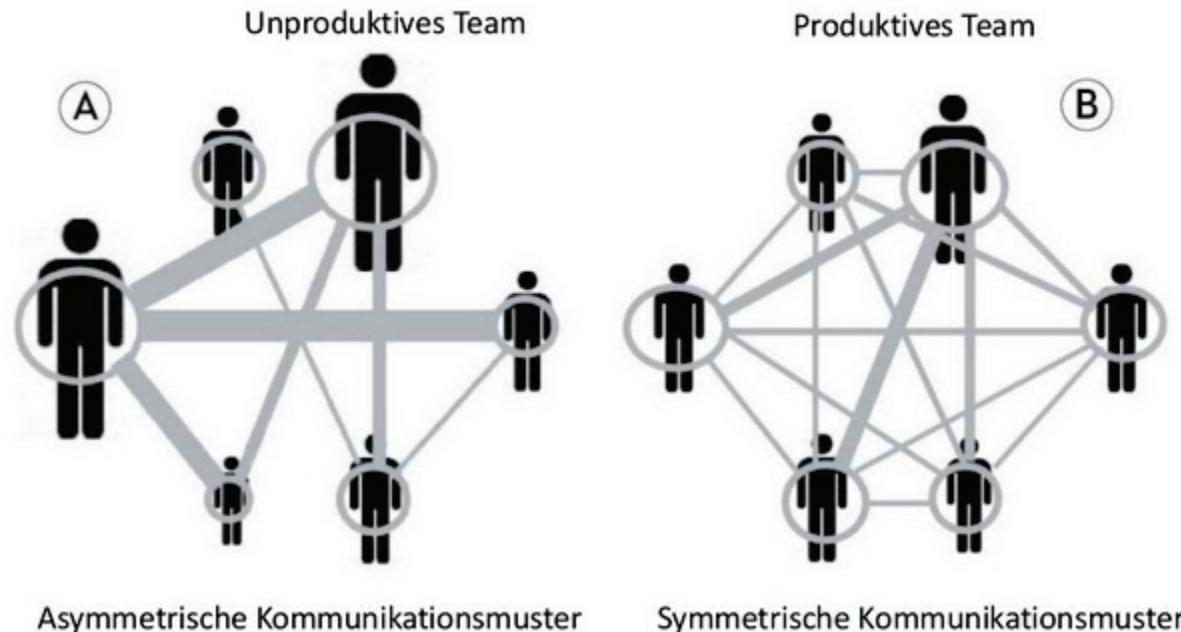
(b)

特定人士主導

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

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

- 預備式探索 (preparatory exploration)
- 人脈網絡多樣性 (diversity)



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入
- 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



Points

■ Dealing with unstructured data (非結構性資料)

- 通話行為
- 蘋果日報慈善捐款
- 虛擬寶物的資料化

■ Data re-targeting and re-use (資料重用)

- SWIFT 全球銀行電匯系統
- 商店監視器

■ Data integration (異質性資料結合)

- 便利商店補貨記錄 + 天氣
- 就診記錄 + Twitter
- 手機訊號 + 病例通報



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



林軒田 / Hsuan-Tien Lin

國立臺灣大學資訊工程學系 / 副教授

林軒田教授於2001年取得台灣大學資訊工程系學士學位，並在2005及2008年相繼在美國加州理工學院取得碩博士學位。2008年返回台灣大學資訊工程系擔任助理教授，2012年升等副教授，2013與2014年間擔任台灣人工智慧協會秘書長，2014年起在沛星科技兼任技術顧問。

研究方向為「機器學習的理論基礎與演算法設計」，他合著了在亞馬遜線上書店暢銷的機器學習入門教科書「Learning From Data」，並依此書在Coursera平台上開設了「機器學習基石」與「機器學習技法」兩門熱門的大型公開線上課程。



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 (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 (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 (Introduction to Computer Vision)
10:30 - 10:50	茶點時間
10:50 - 12:20	基礎與應用 (Fundamentals and Applications)

資料科學面面觀：理論、案例及企業導入方法

日期: 2016/1/23 (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	<p>互動論壇 與談人</p> <p>陳昇瑋 / 年會總召, 中央研究院資訊科學研究所研究員 彭啟明 / 天氣風險管理開發公司總經理 邱銘彰 / 台灣威瑞特公司技術長</p>

- 資料視覺化心法課程 (3/26)
- 從電腦視覺到虛擬實境 (4/23)
- D3.js 互動式資料視覺化實戰 (4/30)
- R 語言的翻轉教室 (4/23, 4/30)
- 深入淺出深度學習 (5/21)
- 紿資料工程師的統計 123
- 社群資料處理上手
- 文字探勘技術上手
- 資料視覺化技術初探
- 敬請期待...

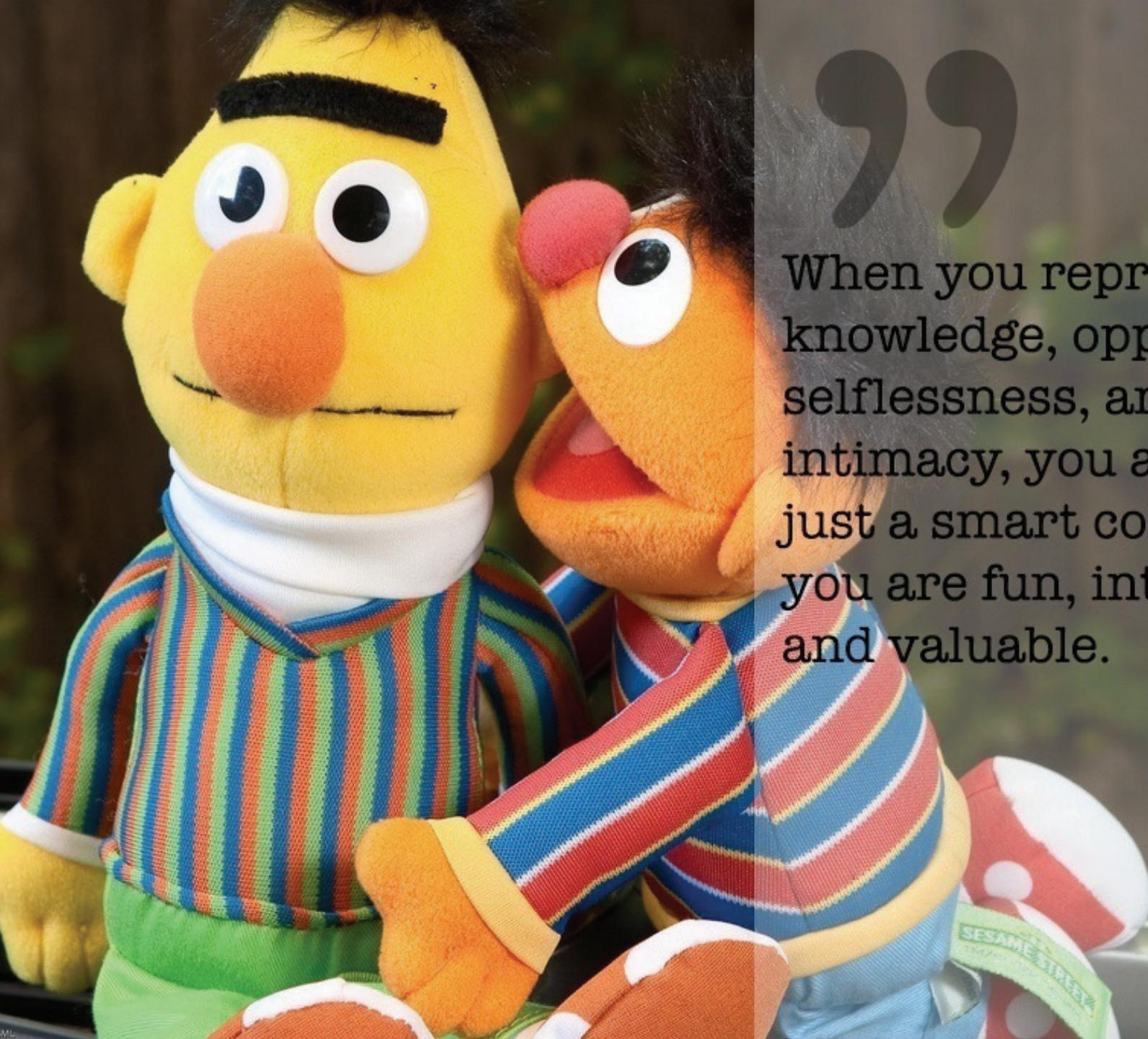
2016
~~2015~~

台灣資料科學 愛好者年會

7/14 - 7/17, 2016

<http://datasci.tw/>

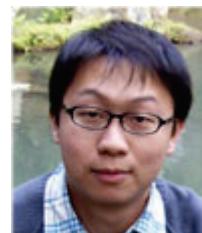
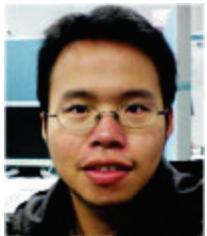
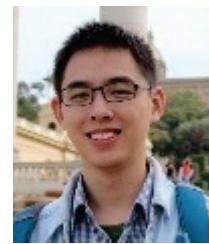




“

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

誌謝



善用資料，創造價值



陳昇瑋

中央研究院
資訊科學研究所