DATA SCIENCE (資料科學)

Shuai, Hong-Han (帥宏翰)

Assistant Professor

Department of Electrical and Computer Engineering

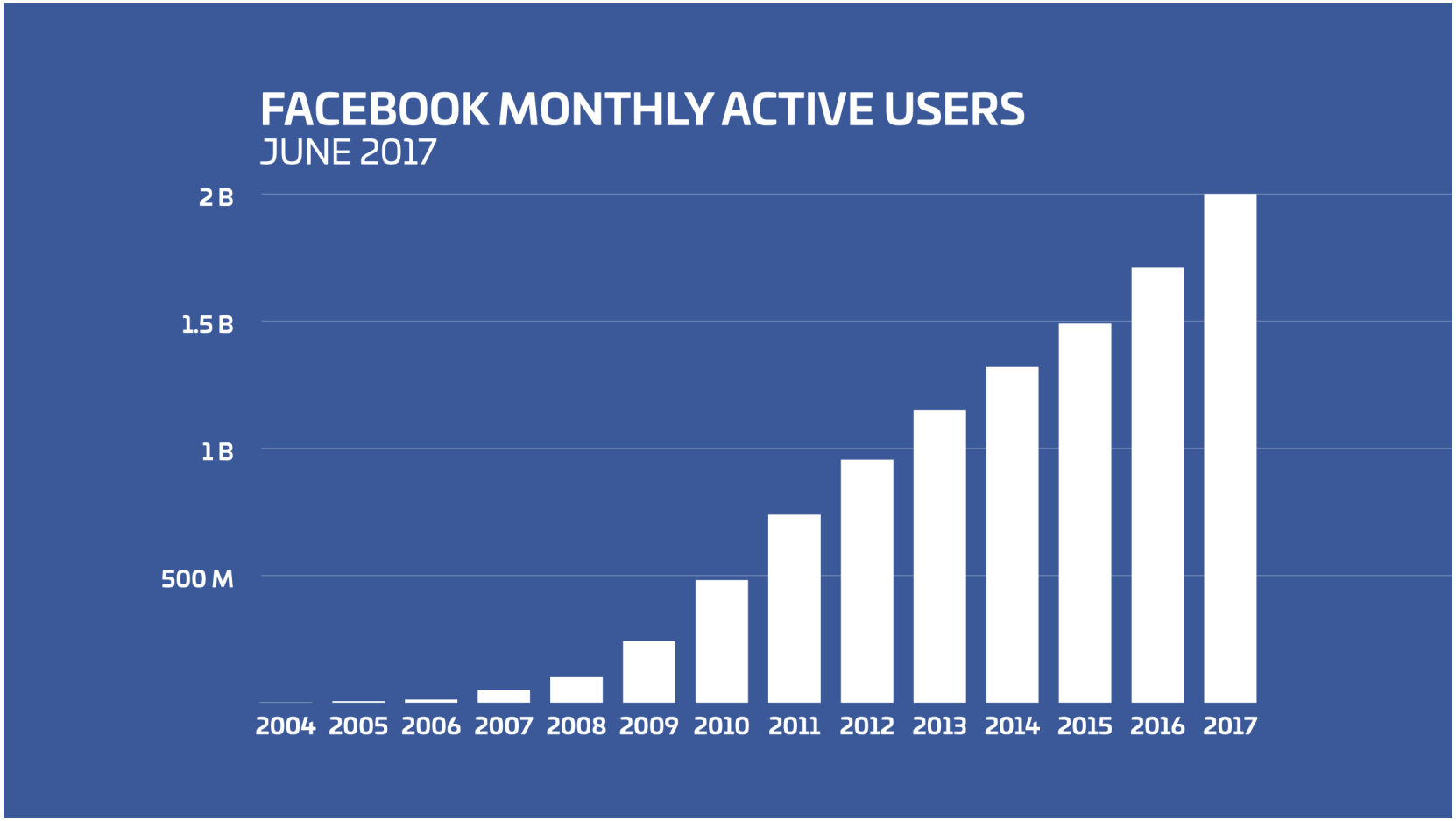
National Chiao Tung University

ABOUT ME

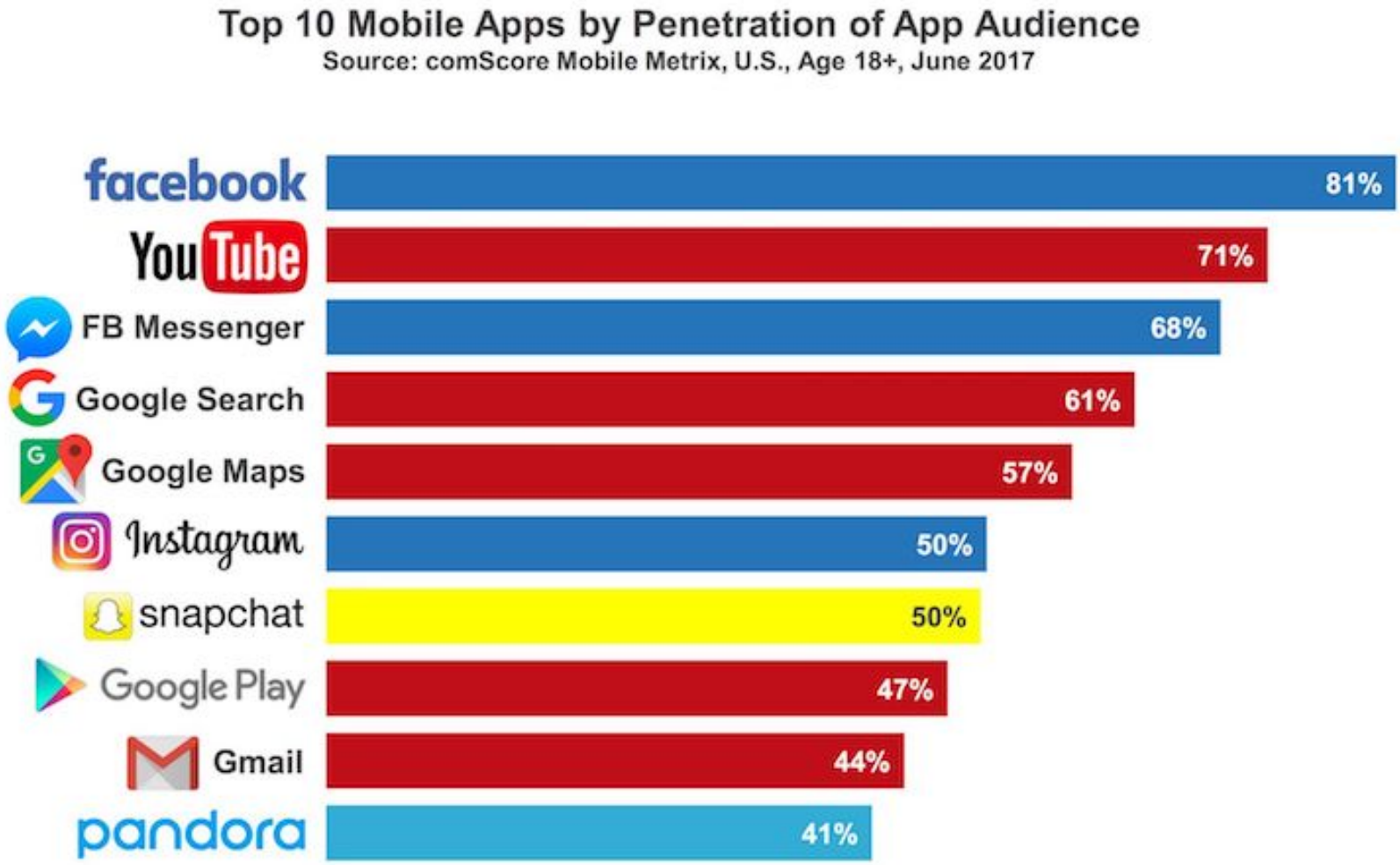
- **Hong-Han Shuai (帥宏翰)**
  - Joined NCTU ECE on 2016/8/1
- **Research Interests:**
  Data Mining, Big Data Analytics, Machine Learning, and Social Network Analytics
- **Office:** ED-807
- **Tel:** 54530
- **E-Mail:** hhshuai@nctu.edu.tw

EMERGENCE FOR THE ROLE OF DATA

- **Web (search) -> Cloud Computing -> Big Data -> Data Science**
  - Google (24PB/day), Facebook (7.9 Billion Comments/day)
  - We are buried in big data, but looking for knowledge
- **The science and technology of data, encompassing techniques of**
  - Database
  - Machine learning
  - Statistics



https://techcrunch.com/2017/06/27/facebook-2-billion-users/



http://www.businessofapps.com/data/app-statistics/

WHAT IS DATA SCIENCE

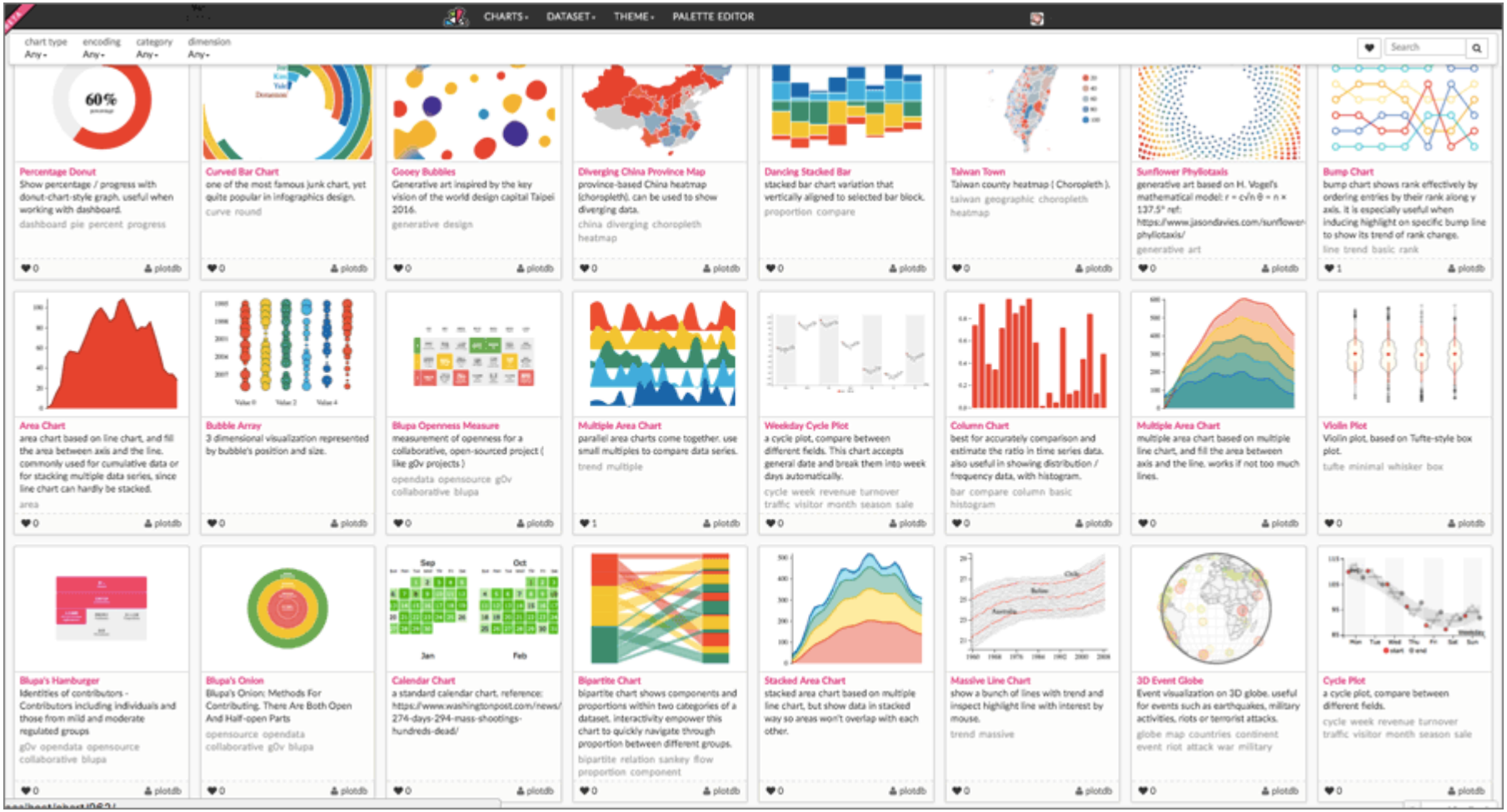
- Data science is:
  - An interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms
  - Either structured or unstructured [1][2]
  - A continuation of some data analysis fields
    - such as statistics, machine learning, data mining, and predictive analytics
  - Similar to Knowledge Discovery in Databases (KDD)

[1] Dhar, V. (2013). "Data science and prediction". Communications of the ACM. 56 (12): 64.
[2] Jeff Leek (2013-12-12). "The key word in "Data Science" is not Data, it is Science". Simply Statistics.

HARVARD BUSINESS REVIEW-DATA SCIENTIST: THE SEXIEST JOB OF THE 21ST CENTURY

★資料視覺化分析師： 將大量資料經過演算、建立預測模型， 再透過如Tableau、QlikView、Spotfire、PlotDB等工具，進行視覺化轉換，強化資料的易讀性。

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century


CHARTS
DATASET
THEME
PALETTE EDITOR
chart type
encoding
category
dimension
Any
Search
60%
Percentage Donut
Show percentage / progress with donut-chart-style graph, useful when working with dashboard.
dashboard pie percent progress
Curved Bar Chart
one of the most famous junk chart, yet quite popular in infographics design.
curve round
Gooey Bubbles
Generative art inspired by the key vision of the world design capital Taipei 2016.
generative design
Diverging China Province Map
province-based China heatmap (choropleth). can be used to show diverging data.
china diverging choropleth heatmap
Dancing Stacked Bar
stacked bar chart variation that vertically aligned to selected bar block.
proportion compare
Taiwan Town
Taiwan county heatmap (Choropleth).
taiwan geographic choropleth heatmap
Sunflower Phyllotaxis
generative art based on H. Vogel's mathematical model: r = c√n θ = n × 137.5° ref: https://www.jasondavies.com/sunflower-phyllotaxis/
generative art
Bump Chart
bump chart shows rank effectively by ordering entries by their rank along y axis. it is especially useful when inducing highlight on specific bump line to show its trend of rank change.
line trend basic rank
Area Chart
area chart based on line chart, and fill the area between axis and the line. commonly used for cumulative data or for stacking multiple data series, since line chart can hardly be stacked.
area
Bubble Array
3 dimensional visualization represented by bubble's position and size.
Blupa Openness Measure
measurement of openness for a collaborative, open-sourced project (like gOv projects)
opendata opensource gOv collaborative blupa
Multiple Area Chart
parallel area charts come together. use small multiples to compare data series.
trend multiple
Weekday Cycle Plot
a cycle plot, compare between different fields. This chart accepts general date and break them into week days automatically.
cycle week revenue turnover traffic visitor month season sale
Column Chart
best for accurately comparison and estimate the ratio in time series data. also useful in showing distribution / frequency data, with histogram.
bar compare column basic histogram
Multiple Area Chart
multiple area chart based on multiple line chart, and fill the area between axis and the line. works if not too much lines.
Violin Plot
Violin plot, based on Tufte-style box plot.
tufte minimal whisker box
Blupa's Hamburger
Identities of contributors - Contributors including individuals and those from mild and moderate regulated groups
gOv opendata opensource collaborative blupa
Blupa's Onion
Blupa's Onion: Methods For Contributing. There Are Both Open And Half-open Parts
opensource opendata collaborative gOv blupa
Sep
Oct
Jan
Feb
Calendar Chart
a standard calendar chart. reference: https://www.washingtonpost.com/news/274-days-294-mass-shootings-hundreds-dead/
Bipartite Chart
bipartite chart shows components and proportions within two categories of a dataset. interactivity empower this chart to quickly navigate through proportion between different groups.
bipartite relation sankey flow proportion component
Stacked Area Chart
stacked area chart based on multiple line chart, but show data in stacked way so areas won't overlap with each other.
Massive Line Chart
show a bunch of lines with trend and inspect highlight line with interest by mouse.
trend massive
3D Event Globe
Event visualization on 3D globe. useful for events such as earthquakes, military activities, riots or terrorist attacks.
globe map countries continent event riot attack war military
Cycle Plot
a cycle plot, compare between different fields.
cycle week revenue turnover traffic visitor month season sale
plotdb

PlotDB

HARVARD BUSINESS REVIEW-DATA SCIENTIST: THE SEXIEST JOB OF THE 21ST CENTURY

★商業智慧分析師： 具備Hadoop、Hive及HBase等軟體使用經驗，能分析企業資料倉儲的各種不同類型資料，從中洞察客戶行為、市場趨勢，進而擬定策略。

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

HARVARD BUSINESS REVIEW-DATA SCIENTIST: THE SEXIEST JOB OF THE 21ST CENTURY

★資料管理師：企業內所有資料的「進」與「出」，都需要經過他認證與管理。也必須確保資料的安全性，甚至具備資料備援的專業技能。

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

HARVARD BUSINESS REVIEW-DATA SCIENTIST: THE SEXIEST JOB OF THE 21ST CENTURY

★資料工程師： 需懂資料庫、資料結構、自然語言處理、數據採礦、數據模型等技術，協助建構大數據的資料平台架構。

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

HARVARD BUSINESS REVIEW-DATA SCIENTIST: THE SEXIEST JOB OF THE 21ST CENTURY

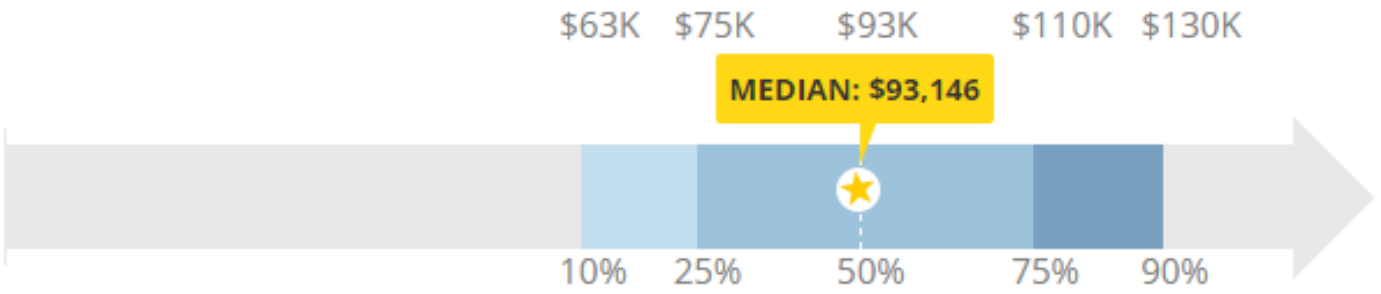
★資料科學家：具備統計學、數學等專業，能將大量資訊運用電腦演算，轉換成具有商業價值的資料，並具備優秀的溝通力，能分析、解釋資料，影響企業決策。

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Data Scientist, IT Salary (United States)

The average pay for a Data Scientist, IT is $93,147 per year. Experience has a moderate effect on income for this job. Most people move on to other jobs if they have more than 10

Read More

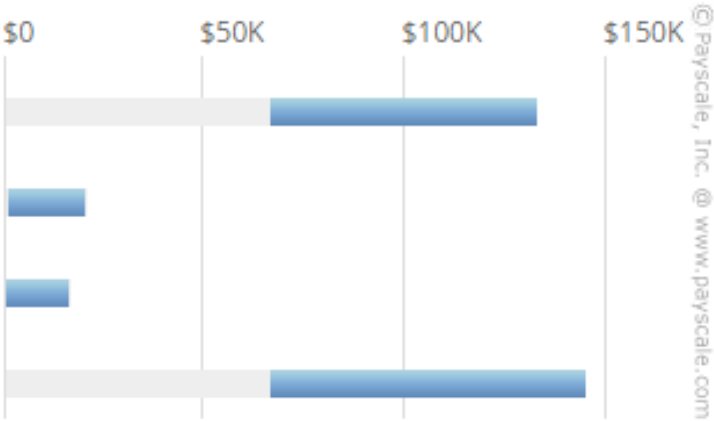


Add this chart to your site: 640px 300px

+ city + experience + skill + employer job

Show Hourly Rate

National Salary Data (?)	
Salary	$62,953 - $126,182
Bonus	$1,030 - $19,432
Profit Sharing	$481.77 - $15,586
Total Pay (?)	**$62,833 - $137,870**



Country: United States | Currency: USD | Updated: 12 Jan 2016 | Individuals Reporting: 1,569

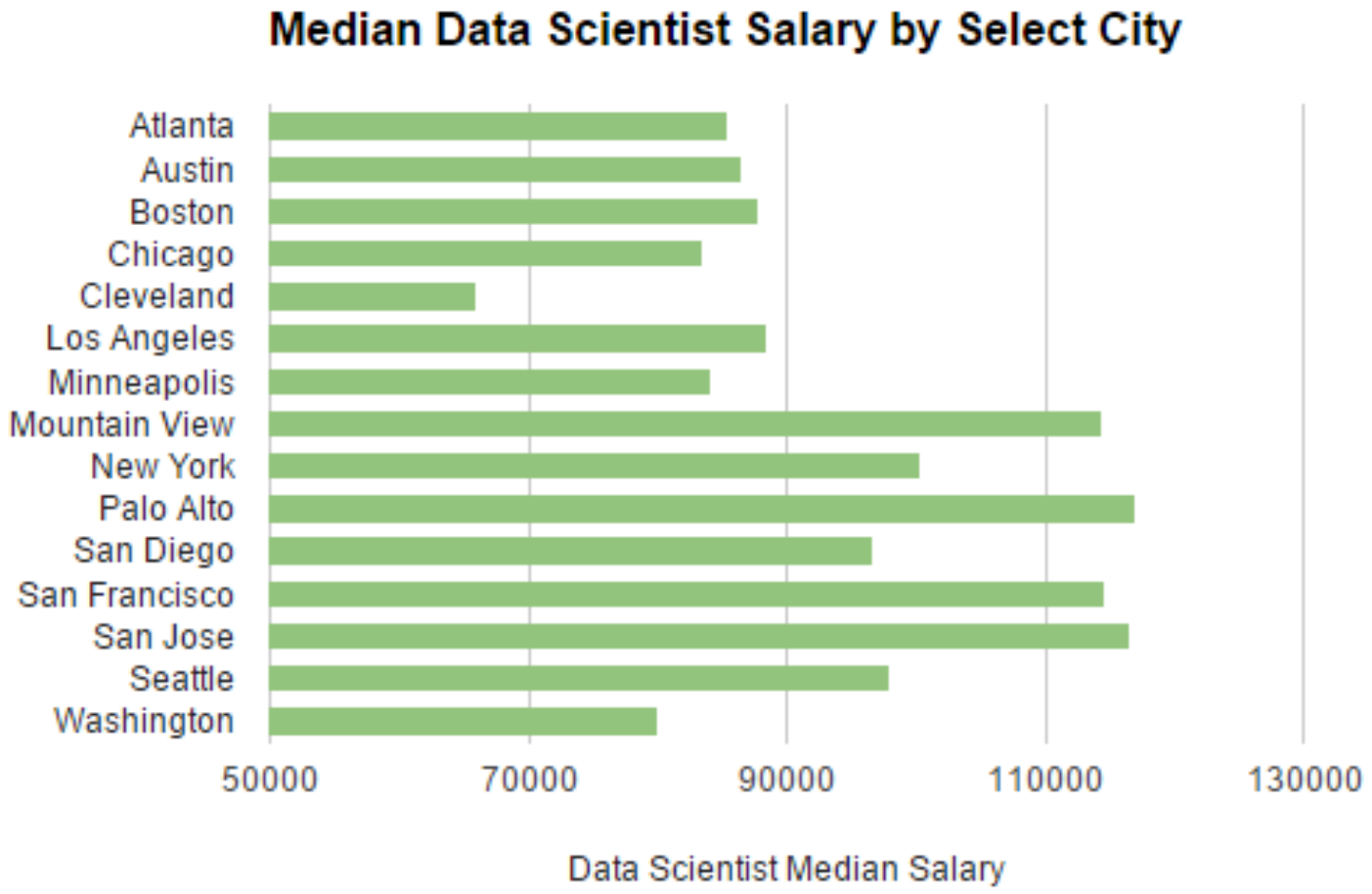


Figure 3. Median Data Scientist Salary by Select City (Interactive)

科技報橘(TECHORANGE)

- Gartner 於指出，未來兩年內，全球超過半數大型組織企業，皆需靠資料分析一決勝負；然而，根據 2016 年全台企業 IT 趨勢大調查，超過七成的企業仍面臨大數據與資料分析的應用困境。即便台灣的開放資料量為世界第一，全台的資料科學家卻僅有 1,092 人，在未來兩年內更將面臨需求人數大增 477%。
- 根據 104 人力銀行統計，台灣目前資料科學人才短缺，2018 年 國內資料人才需求將達 10萬。.
- 新加坡商鈦坦科技開出「資料科學家：年薪 250 萬」的徵才廣告。

https://buzzorange.com/techorange/2016/07/17/taiwan-data-science-conf-2016/

WHAT CAN DATA SCIENCE HELP?

- **Business:** to make more money
  - Social Influence, Social Media, Opinion, Sentiment, 抓带風向的, 或自己下去带風向, etc
  - Trustworthiness, Interests, Trajectory Patterns
  - High-Frequency Trading
- **Politics:** early identification of events
  - Revolution in Egypt, Tunisia protests
- **Health and Well-beings:**
  - Mental Disorder Detection, Virus Outbreaks
- **Events and Habits:**
  - Earthquakes, Smoking Habit, Fat
- **Security**：
  - Face Recognition, Acoustic Keystroke Identification, Smartphone input, etc

IMPACTS ON BUSINESS

- Twitter speaks, markets listen and fears rise (**New York Times, BBC**)
  - After a Twitter hoax that claimed President Obama was injured in an explosion at the White House. That report caused the Dow Jones industrial average to drop temporarily by 150 points, erasing $136 billion in market value

- Facebook friends could change your credit score (**CNN Money**)
  - A handful of tech startups are using social data to determine the risk of lending to people who have a difficult time accessing credit.
- In August 2012, an Italian journalist set up a fake Twitter account for a member of Russia's government and tweeted that the president of Syria had been killed, causing brief fluctuations in the oil markets (**CNN**)

IMPACT ON POLITICS

- Egyptian Revolution Began on Facebook (**New York Times**)
  - "We Are All Khaled Said" (a page created on Facebook) helped ignite an uprising that led to the resignation of President Hosni Mubarak and the dissolution of the ruling National Democratic Party.

- Tunisian protests fueled by social media networks (**CNN**)

- A tweet doesn't just trigger financial panic, it can also strain diplomatic relations, as the U.S. Embassy in Cairo found out in April when the official Twitter account posted a link to a Daily Show segment critical of Egyptian President Mohammed Morsi (**CNN**)

- In March, someone posing as the U.S. ambassador to Moscow tweeted a criticism of the Russian presidential election process, which was picked up by the news media in Russia before it was revealed as a hoax. The U.S. government responded with official statements in both incidents (**CNN**)

WAEL GHONIM

- 「**首先，我們不知道如何應對謠言。**那些謠言表現了人們的偏見，並被相信和散播。」
- 「**其次，我們創造了自己的同溫層。**我們往往只和觀點相同的人溝通，在社群媒體的協助下，我們取消關注或屏蔽意見不同的人們。」
- 「**第三，線上討論會很快激起人們的憤怒。**這讓我們忘了，螢幕後面的，是活生生的人，而不是阿凡達。」
- 「**第四，由於社交媒體快速、簡短的特性，我們很快就跳到了結論。**在此情況下，很難表達出複雜、犀利的觀點。」
- 「**最後，也是我認為最重要的一點，在於社群媒體的特性。**」戈寧說道，「我們的社群媒體被設計為利於傳播而非參與，利於張貼而不是溝通，利於淺薄的觀點而非深度的討論。就好像我們認為，自己是來這裏說教而非對話。」

https://www.ted.com/talks/wael_ghonim_let_s_design_social_media_that_drives_real_change?language=en

媒體小農

https://mediafarmers.org/

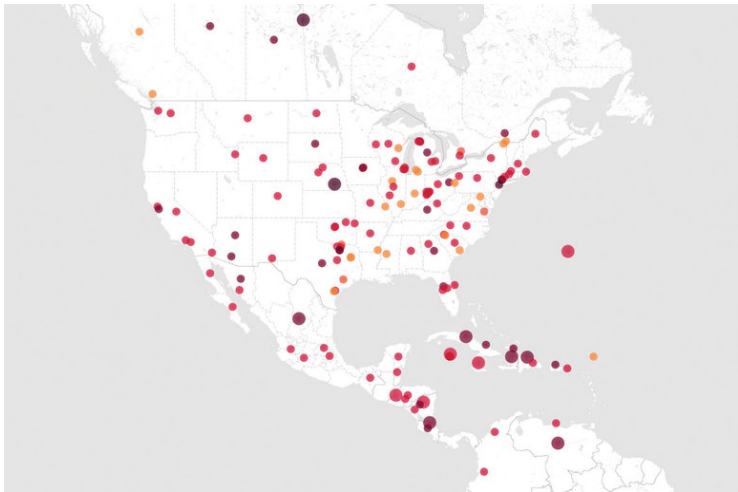
MOGIA FROM GENIC.AI

- Data from Google, Facebook, Twitter, and YouTube
- MoglA predicted Trump's victory in October, even before the FBI announced it was examining new Clinton emails following WikiLeaks revelations about impropriety.
- The CTO mentions that it is hard to detect the messages of sarcasm.

HEALTH AND WELLBEING

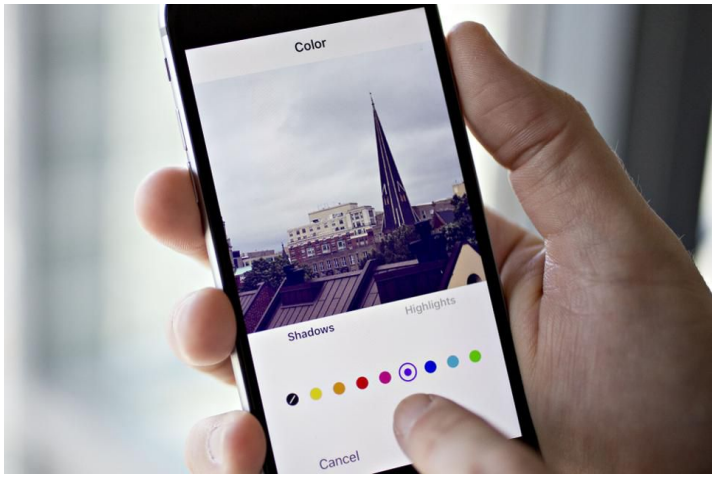
- A paper in **WWW'2016** proposes framework for detecting Social Network Mental Disorder
  - Using only online social network data
  - Cyber-Relationship addiction
  - Information Overload
  - Net Compulsion

- An algorithm spotted the EBOLA outbreak 9 days before WHO announced it (**NEWSWEEK**)
  - Monitoring social media sites, local news reports, medical workers' social networks and government websites to track instances of disease



HEALTH AND WELLBEING

- The research conducted by Andrew G. Reece from Harvard University and Christopher M. Danforth from the University of Vermont said that Instagram may offer clues about depression.
- 166 individuals, who agreed to share their Instagram data and whether they already had a clinical diagnosis of depression (71 had a history of depression).

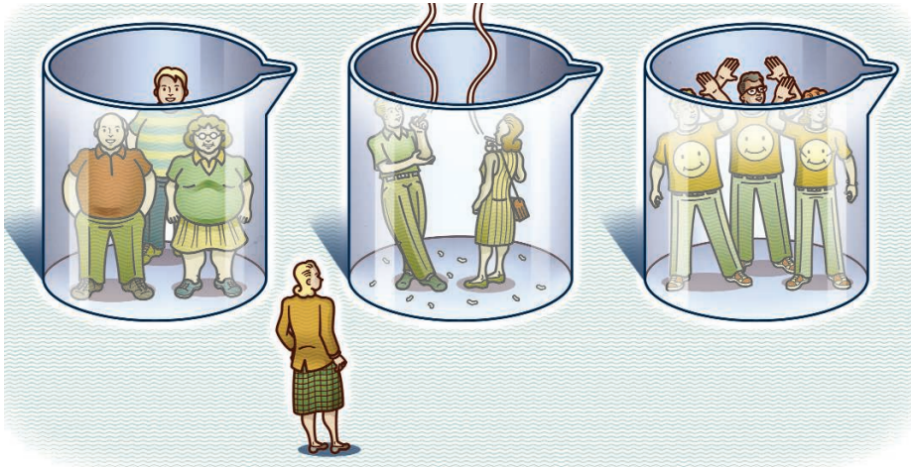


9 CRITERIA

- (1) Post bluer, darker, and grayer photos
- (2) Post more frequently
- (3) Have more comments on their Instagram posts
- (4)Have fewer likes on their Instagram posts
- (5) Post photos with human faces
- (6) Show less of their face, when including a photo with their face.
- (7) Not use Instagram filters to adjust the photo's brightness and coloring.
- (8) Use the Inkwell filter (which would make the photo black and white) when they did use filters.
- (9)Not use Valencia, filter that lightens the tint of the photo

EVENT, HABITS, SECURITY

- **Earthquakes:**
  - Using social media (Twitter) to detect earthquakes (**WWW'2010**)
- **Habits:**
  - Friendship as a Health Factor (**Science'2009**)
  - Smoking, overweight, emotion spread through online social networks
- **Security:**
  - Keystroke identification with acoustic data
  - Smart phone input



SYLLABUS

Week	Contents
1	Introduction to Data Science
2	Intro Python / Data Crawling [H1: Crawling Release]
3	Data Mining (Frequent Pattern –Apriori + FP-Tree)
4	Data Mining (Classification) [HW2: Data Mining]
5	Data Mining (Classification)
6	Data Mining (Clustering) + Evaluation
7	Statistical Measurements/Feature Selection/Dimension Reduction [HW3: dimension reduction/feature selection]
8	Machine Learning/Deep Learning
9	Midterm

SYLLABUS

Week	Contents
10	Deep Learning [HW4]
11	Advanced Technologies for Deep Learning (1)
12	Advanced Technologies for Deep Learning (2)
13	NLP [HW5]
14	Multimedia Processing
15	Graph Theory/SN analysis [HW6]
16	Graph Theory/SN analysis
17	Invited Talk
18	Final presentation

GRADING

- **Homework:** 48% (6 @ 8% each)
  - 6 assignments (e.g., coding, paper reading)
- **Midterm:** 25%
  - In-Class exam
- **Final Project (in groups):** 27%
  - Analyzing real datasets (can be the one you crawled)
  - Interesting ideas are preferred
  - Grading **NOT** based on accuracy
    - But… please make the accuracy be above a minimum threshold
  - Project presentation (about 20 min/team) is required
- Up to 6 points for class participation

EXAM, HOMEWORK, PROJECT, PLEASE…

- Do not copy the others' homework
- Write your own code, please
- In project, please:
  - Work as a team
  - Contribute your ideas
  - Implement your part
  - Do join the discussions
- You can always come knock my door
  - I would be glad to help you

FEEDBACKS

- I'm glad to know any kinds of feedbacks
  - 教太快(太慢)、教太難(太簡單)、太冷、講話陷入奇妙的迴圈、太多笑話...等。
- Please let me know!
  - E-mail, phone, etc

TA

- 4EF @ ED-716
- 許家銘 ming2242@gmail.com
- 陳泓仁 0226.hjc@gmail.com

- https://www.facebook.com/groups/300586450730445/

Questions