DATA SCIENCE (資料科學) FALL, 2020

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

Top Apps Worldwide for September 2019 by Downloads (Non-Game)

SensorTower

	Overall Downloads	App Store Downloads	Google Play Downloads
1	WhatsApp	YouTube	WhatsApp
2	TikTok	TikTok	Messenger
3	Messenger	Instagram	TikTok
4	Facebook	WhatsApp	Facebook
5	Instagram	Google Maps	SHAREit
6	YouTube	Facebook	Likee
7	SHAREit	Messenger	Instagram
8	Likee	Pinduoduo	Snapchat
9	Snapchat	Gmail	UC Browser
10	Wish	Jianying Vlog	Club Factory

Note: Does not include downloads from third-party Android stores in China or other regions.

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

Top Social Networking Apps Worldwide for May 2019 by Downloads

SensorTower

	Overall Downloads	App Store Downloads	Google Play Downloads
1	Facebook	Snapchat	Facebook
2	TikTok	TikTok	TikTok
3	Snapchat	Instagram	Snapchat
4	Instagram	Facebook	Instagram
5	Likee	Twitter	Likee
6	Helo	Pinterest	Helo
7	HAGO	Little Red Book	HAGO
8	Twitter	WeChat	Sharechat
9	Pinterest	QQ	Twitter
10	Sharechat	Discord	Pinterest

Note: Does not include downloads from third-party Android stores in China or other regions.

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。

★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。

★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

但，J個4...

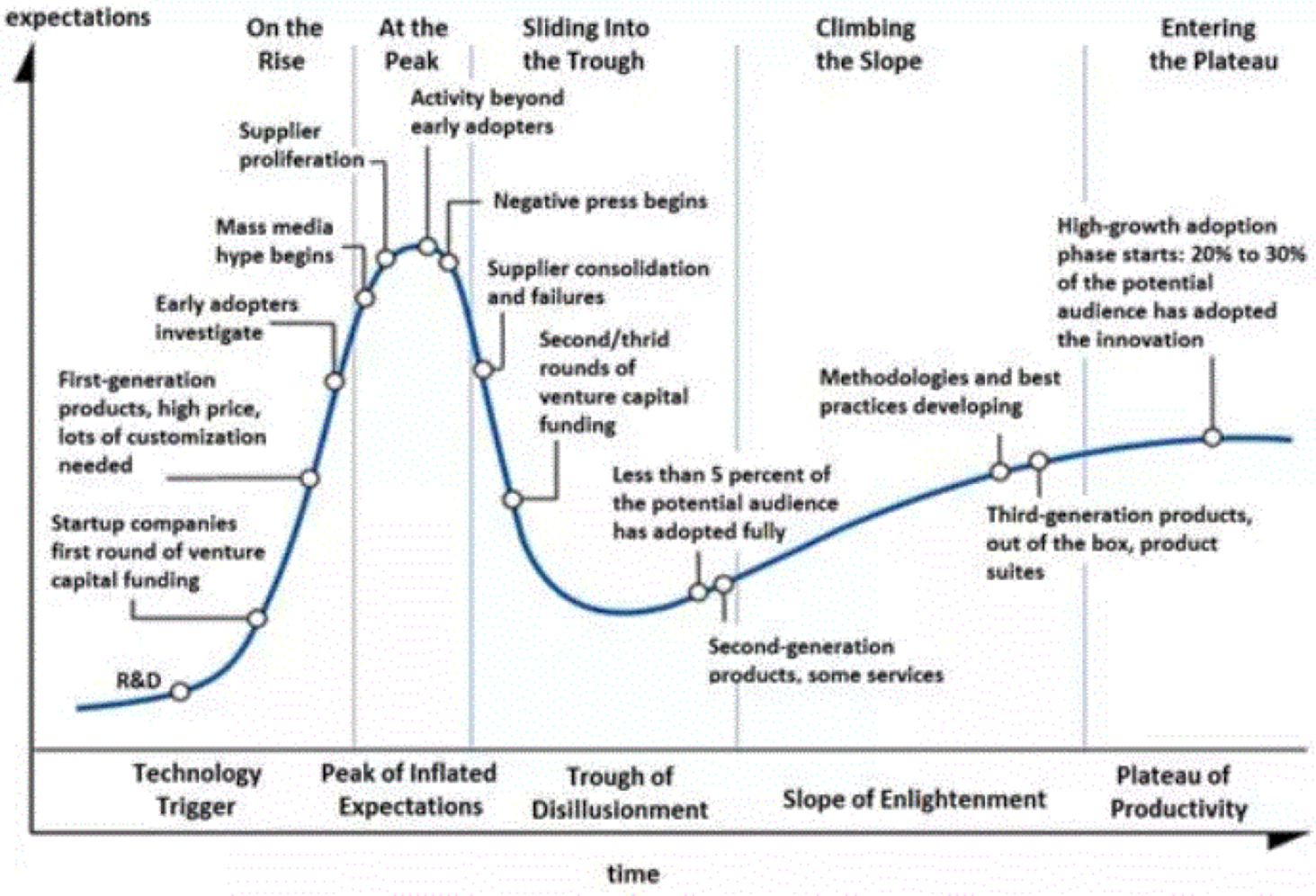
人才擠爆數據科學行業！五年前的「最性感」職業如今邁向泡沫化

2019/03/05 讚 3,629 分享

大數據文摘

https://buzzorange.com/techorange/2019/03/05/how-to-be-a-data-scientists/

SINCE 2012


expectations
On the Rise
At the Peak
Sliding into the Trough
Climbing the Slope
Entering the Plateau
Activity beyond early adopters
Supplier proliferation
Negative press begins
Mass media hype begins
Supplier consolidation and failures
Early adopters investigate
Second/thrid rounds of venture capital funding
First-generation products, high price, lots of customization needed
Less than 5 percent of the potential audience has adopted fully
Methodologies and best practices developing
High-growth adoption phase starts: 20% to 30% of the potential audience has adopted the innovation
Startup companies first round of venture capital funding
Third-generation products, out of the box, product suites
R&D
Second-generation products, some services
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

初級數據科學家的
供過於求

FACT CHECK

- 對 2018 年 4 月份招聘廣告的研究發現，美國有超過 10000 個職位空缺，傾向有人工智能或機器學習技能的人。
  - 超過 10 萬人開始學習 Fast.ai 提供的深度學習課程。
- LinkedIn 表示市場上缺少 151717 個具有數據科學技能的人才。

Open jobs asking for analytics skills in 2015

2.3M

Forecast of population with analytics skills by 2018

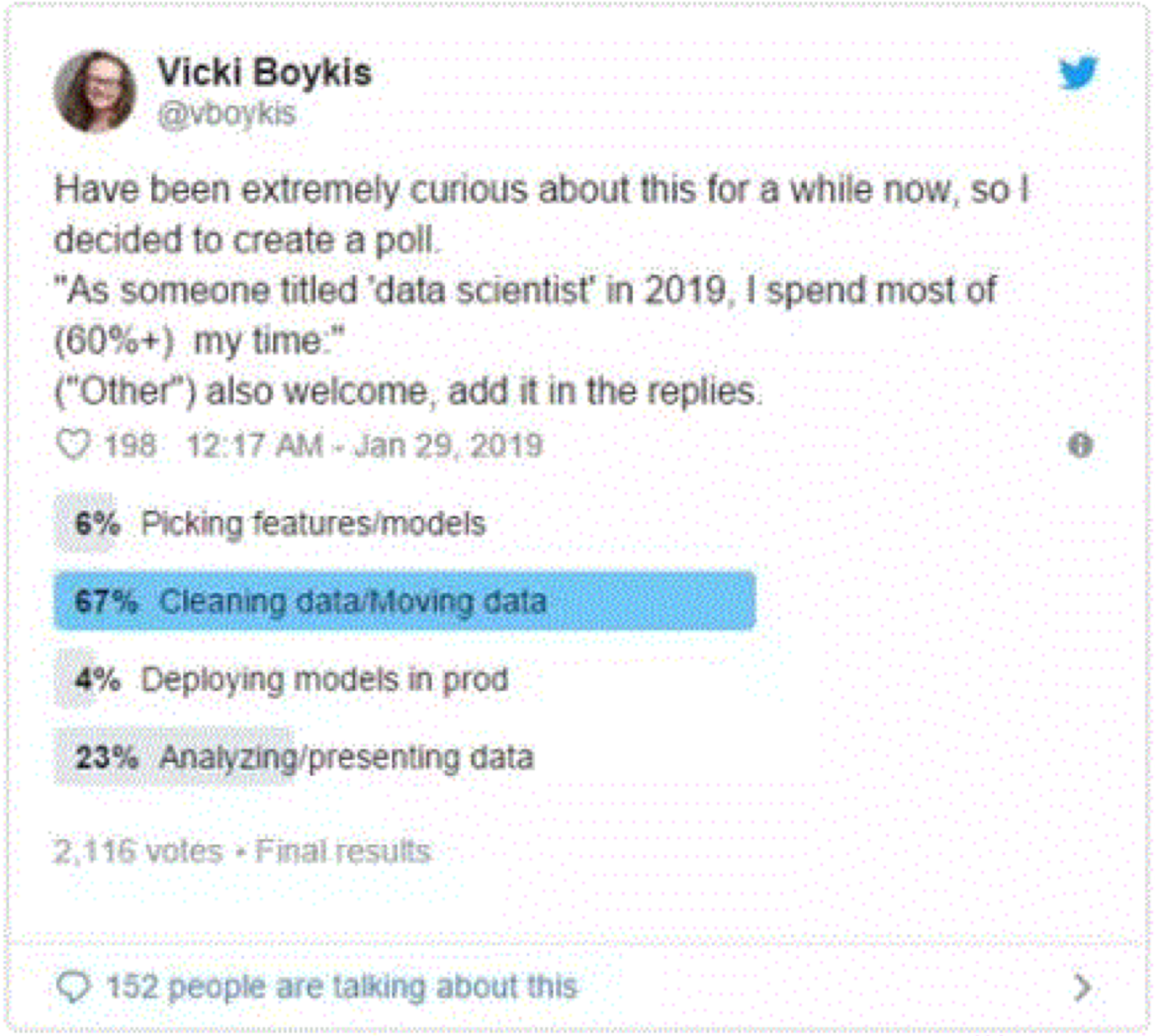
2.9M

Notes: US data only.
Source: Burning Glass Technologies analysis of 26.9 million US job postings from 2015. McKinsey Global Institute, Big Data: The next frontier for innovation, competition, and productivity (June 2011).

工作內容大多不是寫 code

而是「清洗數據」

FACT CHECK


Vicki Boykis
@vboykis
Have been extremely curious about this for a while now, so I decided to create a poll.
"As someone titled 'data scientist' in 2019, I spend most of (60%+) my time:"
("Other") also welcome, add it in the replies.
198 12:17 AM - Jan 29, 2019
6% Picking features/models
67% Cleaning data/Moving data
4% Deploying models in prod
23% Analyzing/presenting data
2,116 votes • Final results
152 people are talking about this

mat kelcey
@mat_kelcey

for my last few ML projects the complexity hasn't been in the modelling or training, it's been in input preprocessing. find myself running out of CPU more than GPU & in one project i'm actually unsure how to optimise the python further (& am considering c++ for one piece)

♡ 136 6:01 AM - Feb 12, 2019

25 people are talking about this

Katherine Scott
@kscottz

One of the biggest failures I see in junior ML/CV engineers is a complete lack of interest in building data sets. While it is boring grunt work I think there is so much to be learned in putting together a dataset. It is like half the problem.

♡ 580 3:50 AM - Feb 2, 2019

158 people are talking about this

「機器學習工程師」崛起

變相導致數據科學家聲望薪水銳減

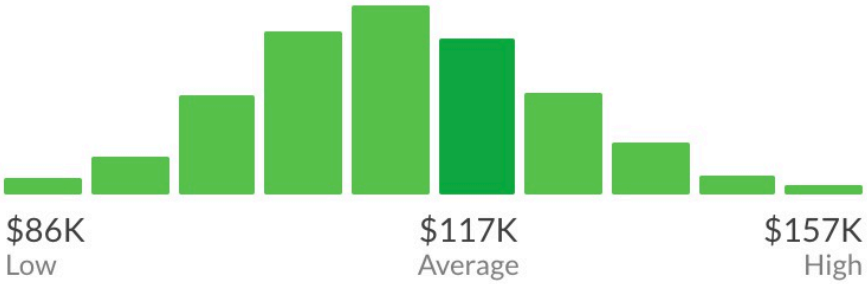
Data Scientist Salaries

4,354 Salaries Updated Sep 8, 2019

Industries Company Sizes Years of Experience

Average Base Pay

$117,345 /yr



Additional Cash Compensation

Average	$11,530
Range	$3,933 - $26,784

How much does a Data Scientist make?
The national average salary for a Data Scientist is $117,345 in United States. Filter by location to see... More

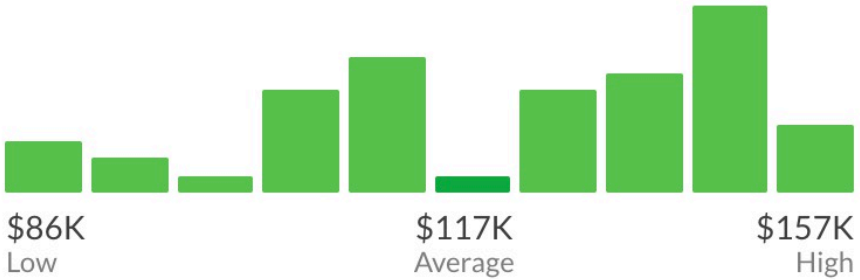
Machine Learning Scientist Salaries

4,354 Salaries Updated Sep 5, 2019

Industries Company Sizes Years of Experience

Average Base Pay

$117,345 /yr



Additional Cash Compensation

Average	$11,530
Range	$3,933 - $26,784

How much does a Machine Learning Scientist make?
The national average salary for a Machine Learning Scientist is $117,345 in United States. Filter by... More

TAIWAN

Average Data Scientist Salary in Taiwan

TWD 771,657

Avg. Salary

TWD 54,000
BONUS

TWD 337,000
PROFIT SHARING

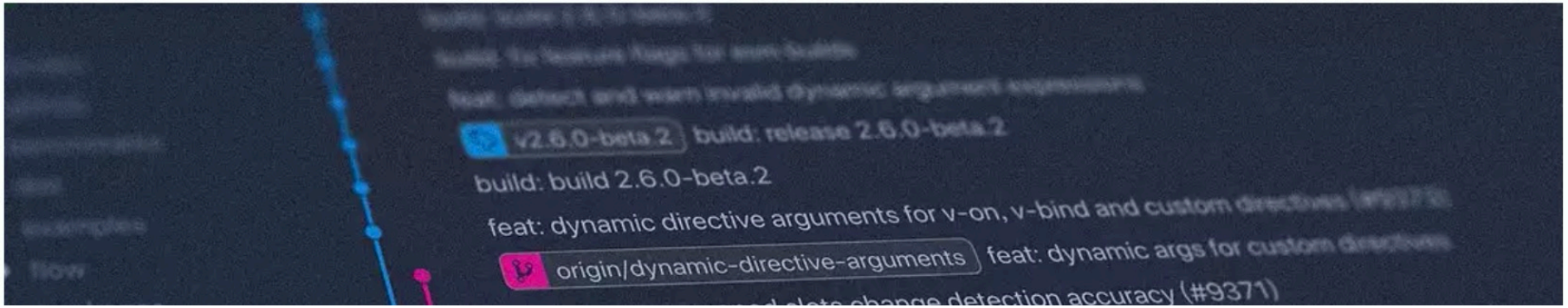
The average salary for a Data Scientist in Taiwan is TWD 771,657.

10%	MEDIAN	90%
TWD 36k	TWD 772k	TWD 3m

【AI 工程師淪為數據打雜工？】過來人親曝血淚史，破解業內五大謊言！

2020/09/04

TO 精選觀點



https://buzzorange.com/techorange/2020/09/04/machine-learning-tips/

MrAcurite 323 points · 12 hours ago

Learn as much fancy theory as you want, but at the end of the day, your job is going to be 99% data cleaning and infrastructure work.

PM_ME_STORM_CROW 79 points · 12 hours ago

This is very true. So much so that I thought I was alone. I work mostly in NLP and 99% of my job is labelling data and making some infrastructure in Java / python with the ML being cosine similarity between various language model vectors

Continue this thread →

怎麼樣避免淪為打雜人員？

- Reddit 熱帖下，網友們吐槽最多的，就是學習和工作的巨大差距產生的幻滅感。學了一大堆高上大的理論知識，以為自己能上手酷炫的神經網路，調參、訓練、部署一條龍，但實際上，這些內容大概只佔全部工作的 10%-20%。
- 大家的建議是： 面對這種大量重複勞動，先別急著上手，去查一查資料，很多類似工作已經有成熟快速的方法 。

機器學習路漫漫，要怎麼不吃虧？

- 哪些在工作崗位上很重要的東西，是機器學習課程不會教你的？
- 網友們總結了最重要的 4 條：
  - 正確認識業務
  - 處理凌亂的現實世界數據集
  - 工程導向，而不是在 Jupyter Notebooks 中編寫研究程式碼
  - 資料可視化

機器學習工作變數多，心臟要很強！

- 規劃傳統工程項目時，設定里程碑、期望值是（相對）簡單的，但對於機器學習專案來說，在最初目標和前期階段之後，很難給出具體的計劃，因為變數實在太多了。
- 這就需要在專案初期階段就保持頭腦清醒，不要給自己挖坑，同時也要和 boss 以及同事說明溝通。

打破機器學習界的謊言

- 任何人都可以輕鬆成為數據科學家/機器學習工程師

打破機器學習界的謊言

- 任何人都可以輕鬆成為數據科學家/機器學習工程師
- 軟體工程師可以輕鬆成為數據科學家
- 學習應用「現成庫」就可以輕鬆上手搞 AI
- 搞 AI 無需學習高等數學/統計學
- 一種特定演算法可以應用於任何領域並獲得成功

真下決心走機器學習這條路，應該相信什麼呢？

- 首先是要對行業現狀有大致了解。對於普通開發來說，機器學習崗位數量要比其他開發崗少很多。
- ML 專案大多比較專，比較小眾。你之前的專案經驗，全國可能只有幾家公司感興趣。
- 除基本數學能力外，持續精進專業知識和跨領域的能力很重要。

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/

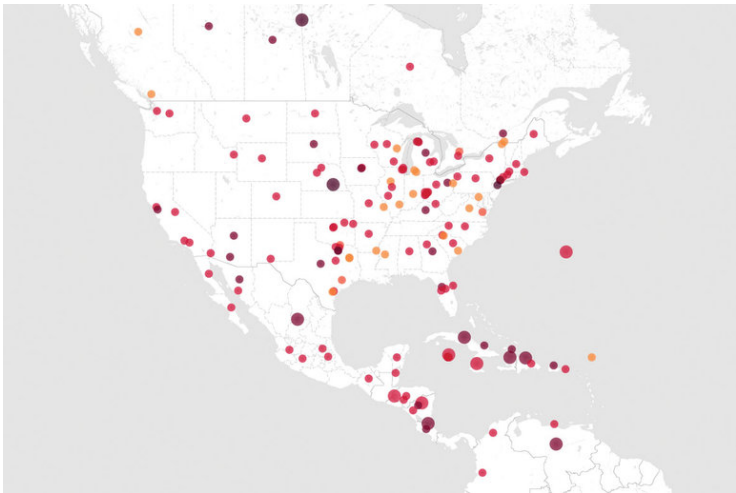
MOGIA FROM GENIC.AI

- 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

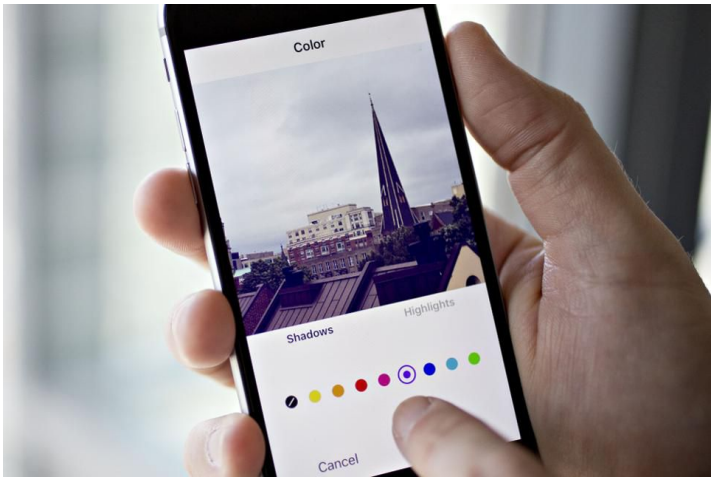
- A paper in **WWW'2016** proposes framework for detecting Social Network Mental Disorder
  - Using only online social network data
  - Cyber-Relationship addiction
  - Information Overload
  - Net Compulsion

- 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

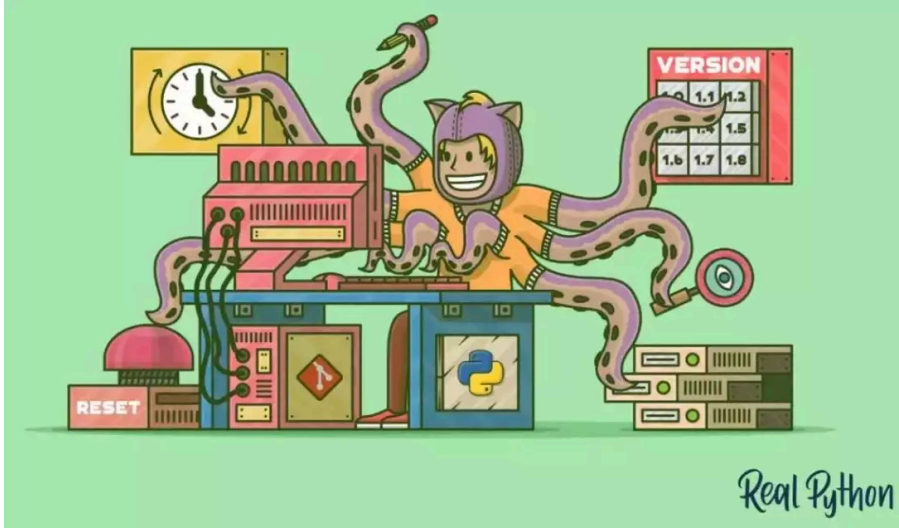
- The research conducted by Andrew G. Reece from Harvard University and Christopher M. Danforth from the University of Vermont said that Instagram may offer clues about depression.
- 166 individuals, who agreed to share their Instagram data and whether they already had a clinical diagnosis of depression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

PREREQUISITE



- Python

https://github.com/jackfrued/Python-100-Days

第一階段，Python 語言基礎（學習週期 15 天）

第二階段，Python 語言進階（週期 15 天）

第四階段，玩轉 Linux 操作系統（週期 5 天）

第九階段，爬蟲開發（週期 10 天）

第十階段， 數據處理和機器學習（週期 15 天）

- Machine Learning/Deep Learning

SYLLABUS

Week	Contents
1	Introduction to Data Science
2	Data Crawling [HW1: Crawling Releases]
3	Natural Language Processing (I): Word Representation
4	Natural Language Processing (II): Recurrent Neural Network
5	Natural Language Processing (III): Sentiment Analysis and Recommendation System (HW2: Attractiveness Prediction)
6	Natural Language Processing (IV): Text Generation
7	Natural Language Processing (V): Question Answering
8	Midterm

2018第一篇文章

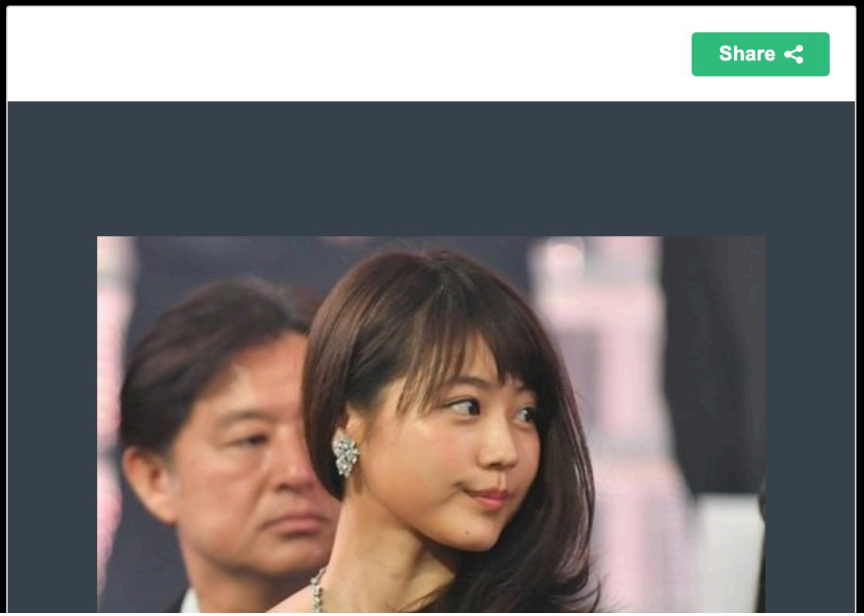
[正妹] 有村架純

https://www.ptt.cc/bbs/Beauty/M.1514740613.A.FF1.html

批踢踢實業坊 › 看板 Beauty　　聯絡資訊 關於我們

作者 as314 (幽默販賣機)　　看板 Beauty
標題 [正妹] 有村架純
時間 Mon Jan 1 01:16:50 2018

https://i.imgur.com/RIo8fVu.jpg



5 AUG 2020

Unsupervised Translation of Programming Languages

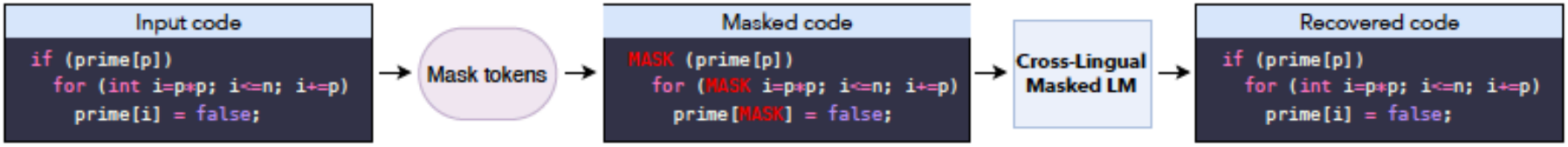
Baptiste Roziere*
Facebook AI Research
Paris-Dauphine University
broz@fb.com

Marie-Anne Lachaux*
Facebook AI Research
malachaux@fb.com

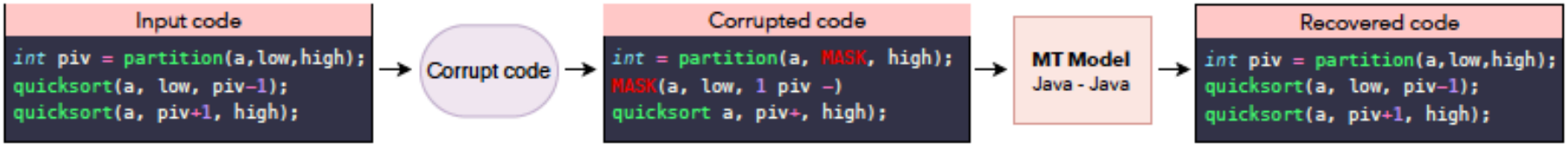
Lowik Chanussot
Facebook AI Research
lowik@fb.com

Guillaume Lample
Facebook AI Research
glample@fb.com

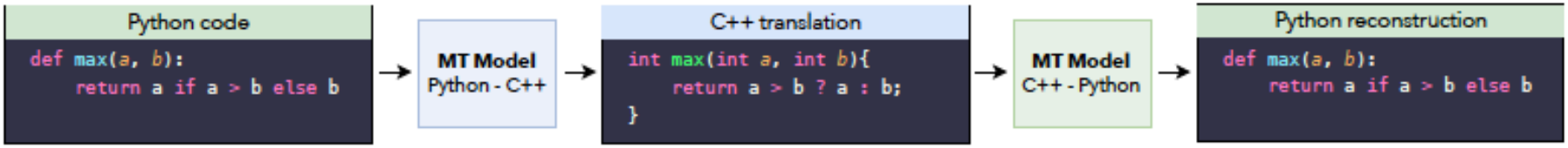
Cross-lingual Masked Language Model pretraining



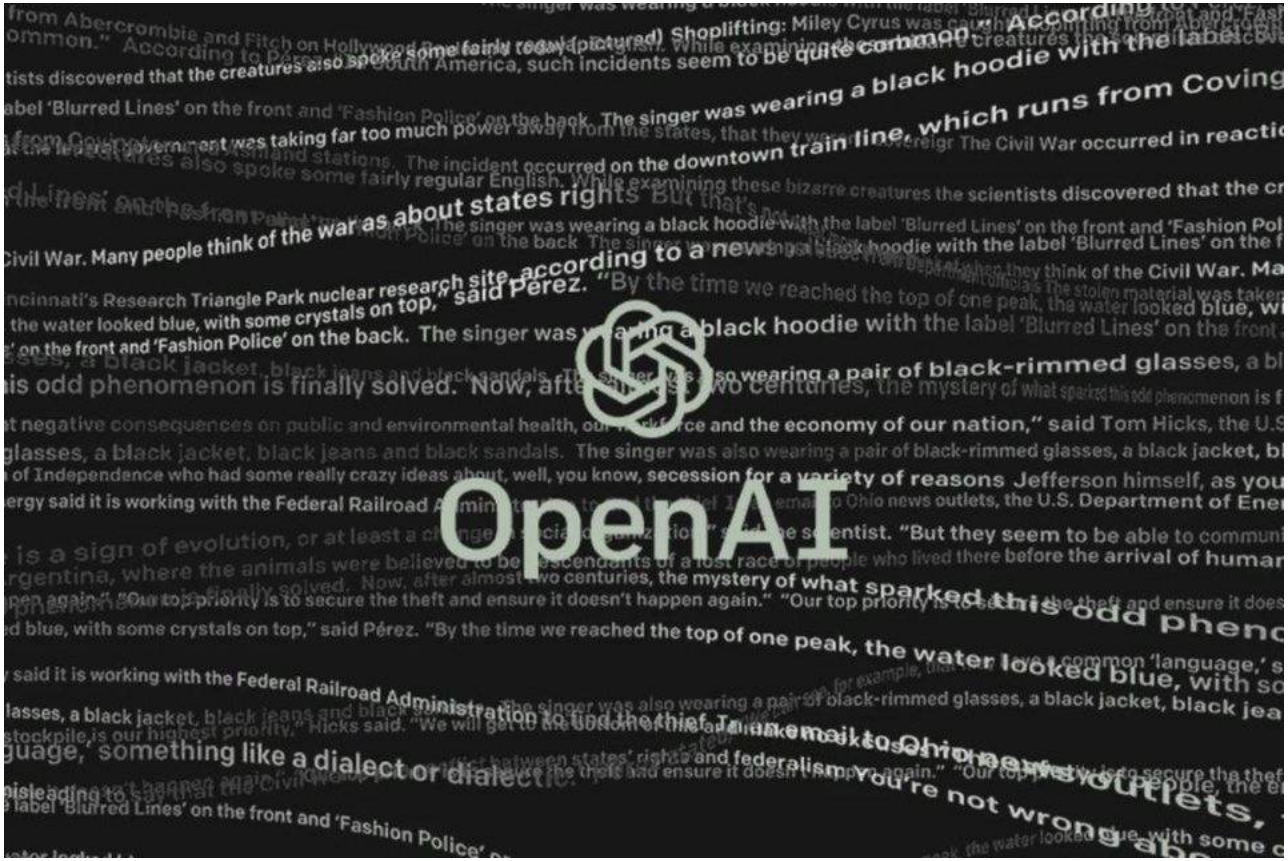
Denoising auto-encoding



Back-translation



1750 億參數，史上最大 AI 模型 GPT-3 上線



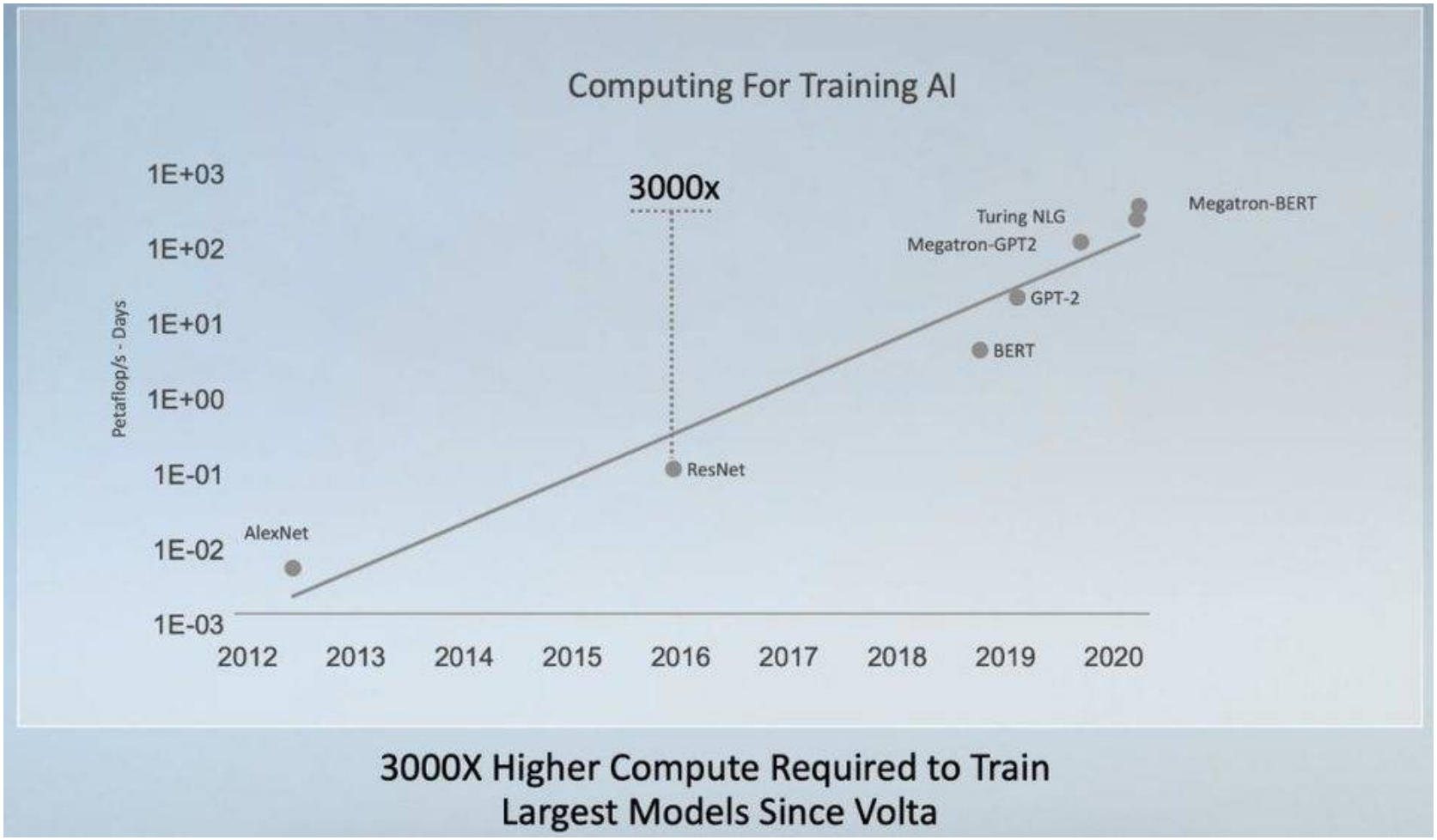
GPT-3

- 比全球最大深度學習模型 Turing NLP大上十倍
- 使用的最大數據集在處理前容量達到了 45TB
- GPT-3 需要比AlphaGoZero大兩倍以上的算力
- Paper (72 pages):
- https://arxiv.org/abs/2005.14165

Code:

- https://github.com/openai/gpt-3
- About 700GB for a pretrained model
- 翻譯、問答和文本填空任務

Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model. In Section 4 we characterize the impact of the remaining overlaps, and in future work we will more aggressively remove data contamination.



NEWS GENERATION

Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

USING A WORD TO MAKE SENTENCES

- 給出新單詞「Gigamuru」（表示一種日本樂器）。
  GPT-3 給出的句子是：叔叔送了我一把 Gigamuru，我喜歡在家彈奏它。

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

- 給出新單詞「screeg」（揮劍，擊劍）。
  GPT-3 造出的句子是：我們玩了幾分鐘擊劍，然後出門吃冰淇淋。

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

GRAMMATIC CORRECTION

Poor English input: The patient was died.
Good English output: The patient died.

Poor English input: We think that Leslie likes ourselves.
Good English output: We think that Leslie likes us.

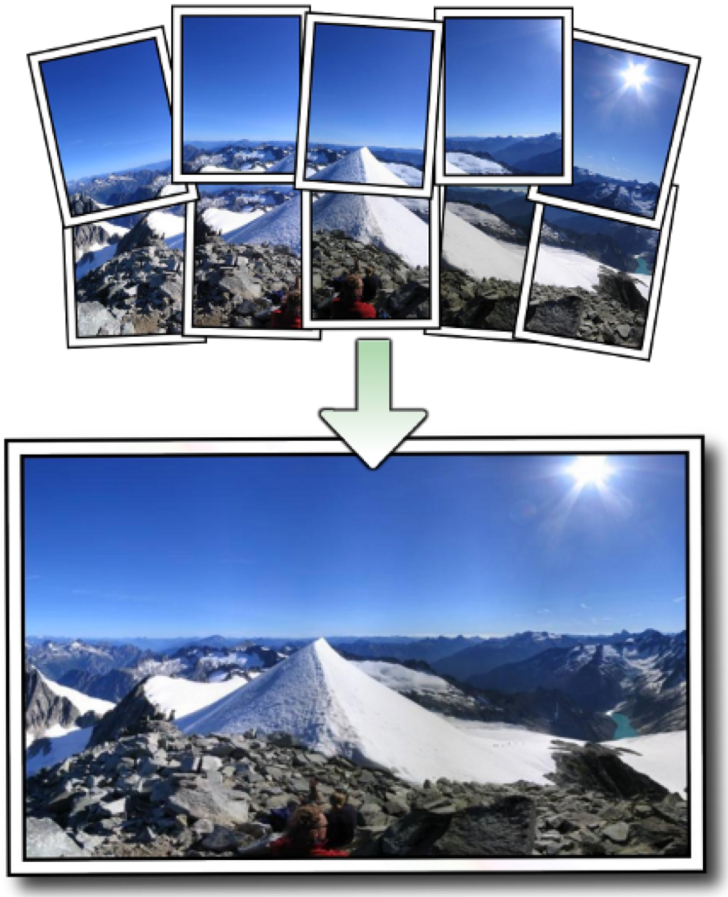
MATH

- **2 digit addition (2D+)** – The model is asked to add two integers sampled uniformly from $[0, 100)$, phrased in the form of a question, e.g. "Q: What is 48 plus 76? A: 124."
- **2 digit subtraction (2D-)** – The model is asked to subtract two integers sampled uniformly from $[0, 100)$; the answer may be negative. Example: "Q: What is 34 minus 53? A: -19".
- **3 digit addition (3D+)** – Same as 2 digit addition, except numbers are uniformly sampled from $[0, 1000)$.
- **3 digit subtraction (3D-)** – Same as 2 digit subtraction, except numbers are uniformly sampled from $[0, 1000)$.
- **4 digit addition (4D+)** – Same as 3 digit addition, except uniformly sampled from $[0, 10000)$.
- **4 digit subtraction (4D-)** – Same as 3 digit subtraction, except uniformly sampled from $[0, 10000)$.
- **5 digit addition (5D+)** – Same as 3 digit addition, except uniformly sampled from $[0, 100000)$.
- **5 digit subtraction (5D-)** – Same as 3 digit subtraction, except uniformly sampled from $[0, 100000)$.
- **2 digit multiplication (2Dx)** – The model is asked to multiply two integers sampled uniformly from $[0, 100)$, e.g. "Q: What is 24 times 42? A: 1008".
- **One-digit composite (1DC)** – The model is asked to perform a composite operation on three 1 digit numbers, with parentheses around the last two. For example, "Q: What is 6+(4*8)? A: 38". The three 1 digit numbers are selected uniformly on $[0, 10)$ and the operations are selected uniformly from {+,-,*}.

SYLLABUS

Week	Contents
9	Traditional Computer Vision: [HW3: Panorama Releases]
10	Advanced Computer Vision (I): Generative Adversarial Networks and Glow
11	Advanced Computer Vision (II): Capsule Networks
12	Advanced Computer Vision (III): Object Detection (HW4: Salient Object Detection)
13	Advanced Computer Vision (IV): Crowd Estimation and Style Transfer
14	Social Network Analysis (I): User modeling and graph embedding
15	Social Network Analysis (II): Fake news on social media
16	Final Project Presentation (2 days)

PANORAMA



GRADING

- **Homework: 48**% (4 @ 12% each)
  - 4 assignments (e.g., coding, paper reading)
- **Midterm: 27**%
  - In-Class exam
- **Final Project (in groups):** 25%
  - 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 15 min/team) is required
  - Open Competition is also a good candidate.
- 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

TA

- Tuesday 13:00-15:00 @ ED-716
- 宋韻筑 yunzhusong.eed07g@nctu.edu.tw
- 吳易倫 w86763777@gmail.com
- 陳義瑄 yisyuan.chen.ece@gmail.com

- FB: https://www.facebook.com/groups/763230681108210
- Streaming: https://meet.google.com/nuj-kvfx-qup

Questions