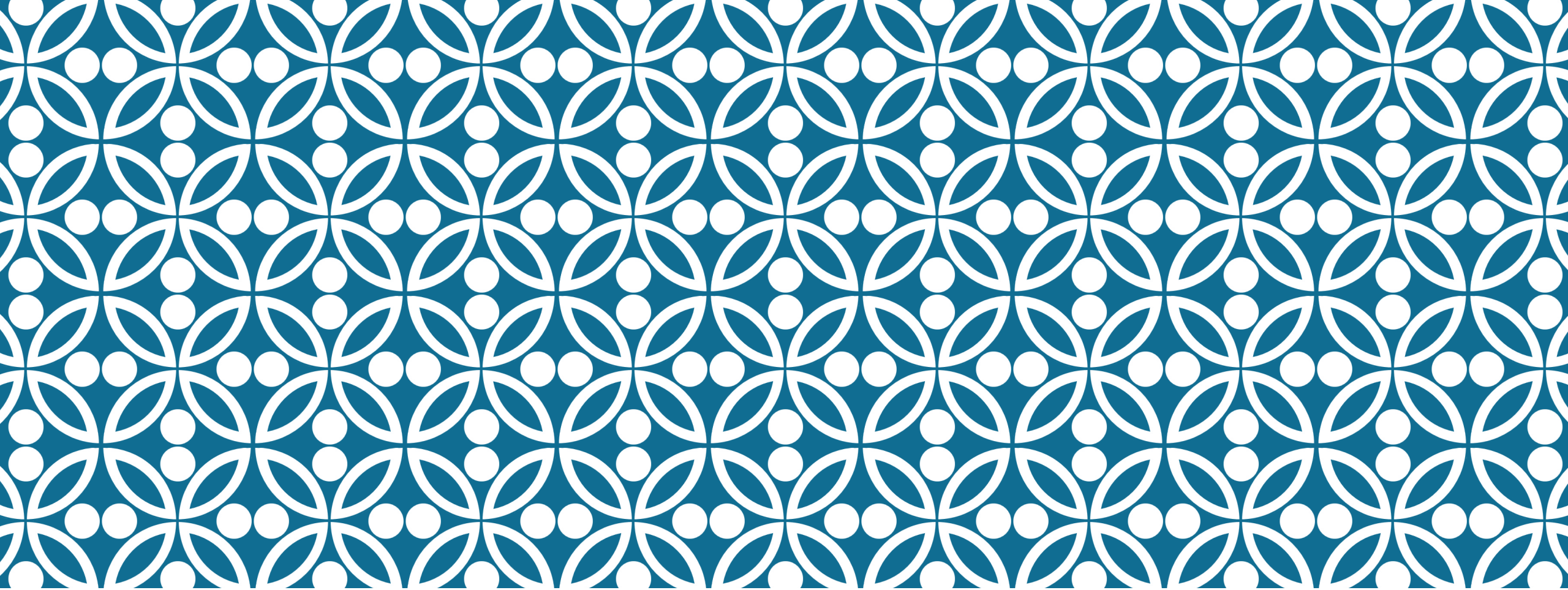


應用機器學習

Brian Chan 陳醒凡

關於我:

我係Brian, 陳醒凡. 土生土長的香港人

背景: BBA in Finance; MSc in Fin. Math.; PhD in Systems Engineering

職業: 曾於對沖基金及投行從事計量研究工作

研究興趣: 金融數據分析、資產定價和交易策略開發

為什麼要任教這門課程?

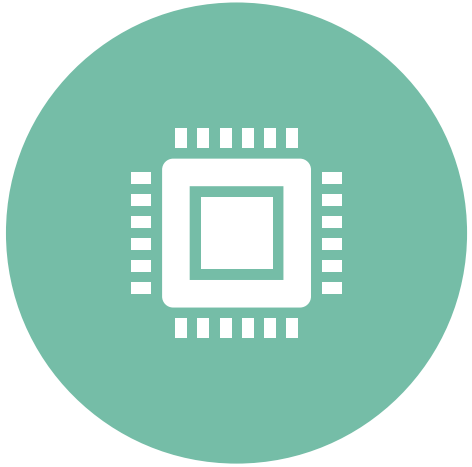
1. 有趣
2. 分享我學到的知識並推動自己梳理一下對機器學習的所知所學
3. 傳授有用的數據處理和機器學習工具

課程目標

1. 了解基本的數據分析
2. 了解基本的機器學習(Machine Learning)方法
3. 掌握Python的基本操作和一些有用的package
4. 處理及從網上下載數據
5. 在Python上應用機器學習

基礎要求

1. MATH: BASIC CALCULUS, PROBABILITY & STATISTICS



2. PROGRAMMING: PYTHON (NUMPY, PANDAS)

MY PLAN (暫定) 2019秋季

第一課	Introduction	講解機器學習的分類及經典應用例子
第二課	Python I	講解 Python 的基本應用
第三課	Python II	講解如何以 Python 作數據處理及常見問題
第四課	Regression	教授迴歸法(Regression)及其應用例子
第五課	Classification	教授分類法(Classification-Logistic regression & SVM)及其應用例子
第六課	Random Forest	教授隨機森林(Random Forest)及其應用例子
第七課	Principal Component Analysis	教授Principal Component Analysis (PCA)及其應用例子
第八課	Model Evaluation	機器學習模型的評估方法(Model Evaluation)
第九課	Deep Learning	簡介深度學習(Deep Learning)及其應用例子

MY PLAN (暫定) 2019夏季

課堂內容

第一課 講解機器學習的分類及經典應用例子

第二課 講解 Python 的基本應用

第三課 講解如何以 Python 作數據處理及常見問題 + 網路爬蟲 (Web scrapping) + Mysql 資料庫

第四課 教授迴歸法(Regression)及其應用例子 (e.g.紅酒評分和股票對沖)

第五課 教授分類法(Classification – Bayesian Classifier)及其應用例子 (e.g."誰在說謊")

第六課 教授分群法(Clustering – K mean classifier)及其應用例子

第七課 機器學習模型的評估方法(Model Evaluation)

第八課 總結各種機器學習模型及其在基金公司內部的應用例子

第九課 介紹人工神經網路(Artificial Neural Network)及其應用例子

今天課堂概要

1. 什麼是機器學習？機器學習的「前世今生」
2. 人工智能技術的前沿
3. 機器學習方法概述
4. 機器學習的編程語言比較
5. 關於機器學習的一些關鍵概念
6. 互動示範（K-means集群）

什麼是機器學習?

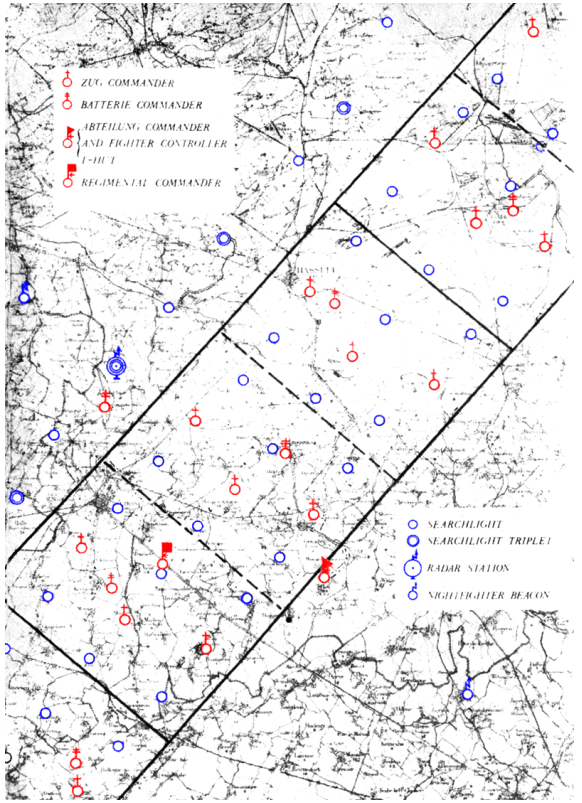
- 主要是設計和分析一些讓電腦可以自動「學習」的演算法。
- 從數據中自動分析獲得規律，並利用規律對未知數據進行預測的演算法。
- 多領域交叉學科，涉及概率論、統計學、逼近論、凸分析、計算複雜性理論等多門學科。
- 因為學習演算法中涉及了大量的統計學理論，機器學習與推斷統計學聯絡尤為密切，也被稱為統計學習理論。
- 機器學習已廣泛應用於數據探勘、電腦視覺、自然語言處理、生物特徵辨識、搜尋引擎、醫學診斷、檢測信用卡欺詐、證券市場分析、DNA序列測序、語音和手寫辨識、戰略遊戲和機械人等領域。

Source: https://zh.wikipedia.org/zh-hk/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92

機器學習的「前世今生」

世界第二次大戰(1939-45)

統計學與運籌學



人工智能的前沿

1. Twitter訊息（文本分析）	Natural Language Processing 自然語言處理
2. 圍棋比賽（AlphaGo / AlphaGo Zero）	Reinforcement Learning 強化學習
3. AI照片生成器	Generative adversarial networks 生成性對抗性網絡
4. 圖像處理（計算抗議人數）	Pattern recognition (Deep learning) 深度學習

1. TWITTER訊息

https://news.cnyes.com/news/id/3054033

"*Twitter mood predicts the stock market*",

Journal of computational science 2 (2011) 1-8

〈分析〉Twitter情緒指數 果真是預測市場走勢新法寶？

鉅亨網編譯李業德 綜合外電　2012/02/20 21:57

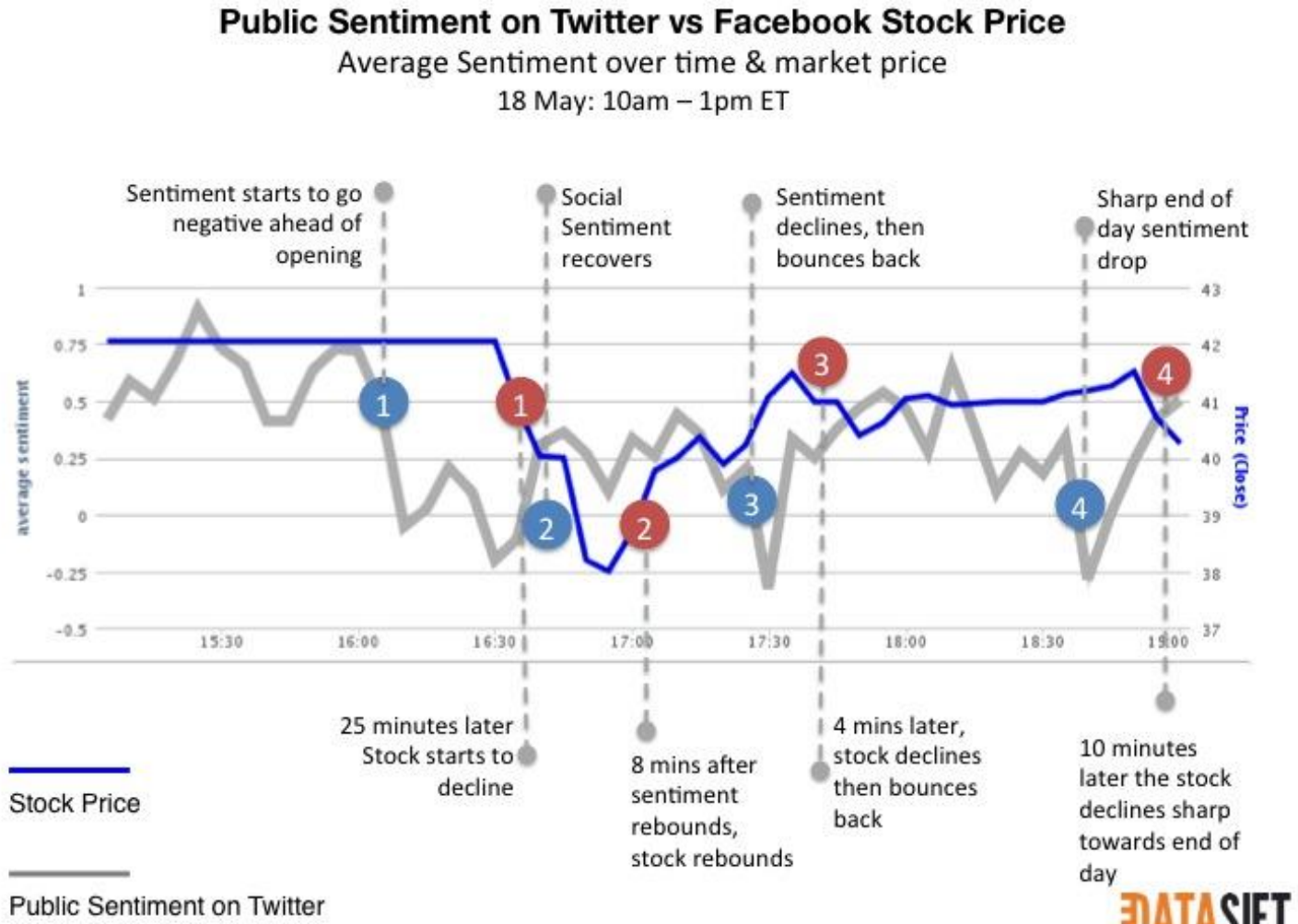
以Twitter分析大眾情緒，可預測股市走勢。(圖為Twitter首頁)

市場研究機構 MarketPsych 主管 Richard Peterson 約在 8 年前，便對避險基金經理人們宣稱，社交媒體可用以預測股市走向，但當時觀念過於新穎難為人接受。

Peterson 主張，社交媒體的動向可用以網羅民眾想法以及觀感，而這些資訊可轉化為有力的投資理念。

經理人們對他說到：「你瘋了，基金怎麼可能分析社交媒體？」沒有人認真考量他的觀點。

	date	user_loc	message	full_name	country	country_code	geo_code	tokens	no_pauls
0	2018-04-09 23:52:02	London, England	' look back today the first time paul ryan said " " without hint irony	Wandsworth, London	United Kingdom	GB	[-0.19372, 51.451656]	['', look, back, today, first, time, paul, ryan, said, ", ", without, hint, irony]	[look, back, today, first, time, said, without, hint, irony]
1	2018-04-09 23:56:39	NaN	paul ryan and mitch mcconnell are big sissies	Pennsylvania, USA	United States	US	[-77.604684, 41.117936]	[paul, ryan, mitch, mcconnell, big, sissies]	[mitch, mcconnell, big, sissies]
2	2018-04-09 23:57:41	Florida, USA	were are mcconell and paul ryan your silence sickening	Boynton Beach, FL	United States				
3	2018-04-10 00:00:04	New York City	over under paul ryan potus cinco impeacho	Manhattan, NY	United States				
4	2018-04-10 00:03:42	Miami, FL	where are paul ryan mitch mcconnell	Miami, FL	United States				

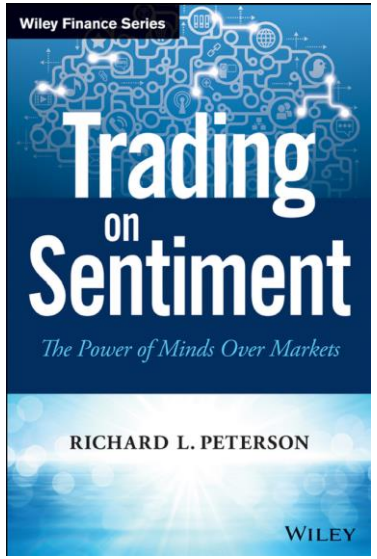


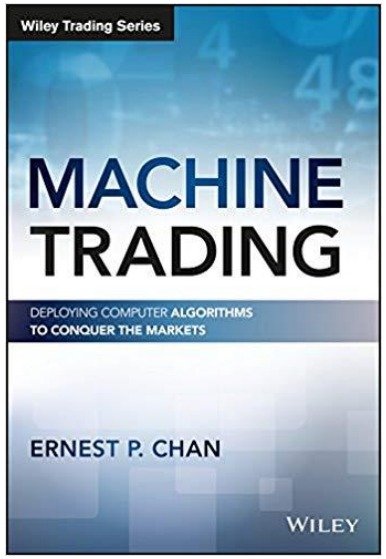
Trading on Sentiment: The Power of Minds Over Markets Book by Richard L. Peterson

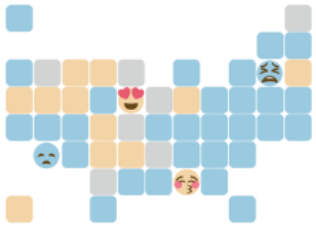
https://onlinelibrary.wiley.com/doi/book/10.1002/9781119219149

Machine Trading: Deploying Computer Algorithms to Conquer the Markets

https://www.bookdepository.com/Machine-Trading-Ernest-P-Chan/9781119219606?ref=grid-view&qid=1568814342827&sr=1-1







WALL STREET JOURNAL/IHS MARKIT

U.S. Social Sentiment Index

Measuring the content of millions of Twitter messages each hour is one way to gauge the nation's changing mood in near-real time. Here, the Wall Street Journal and IHS Markit have plotted that sentiment, comparing it to recent norms. How this is determined

By ***WSJ Graphics***

Last updated June 21, 2019 at 8:08 a.m. ET

Related article: A New Index Tracks Our National Mood One Tweet at a Time

U.S. SENTIMENT, FOR 11 A.M. ET

92 | -2 pts.

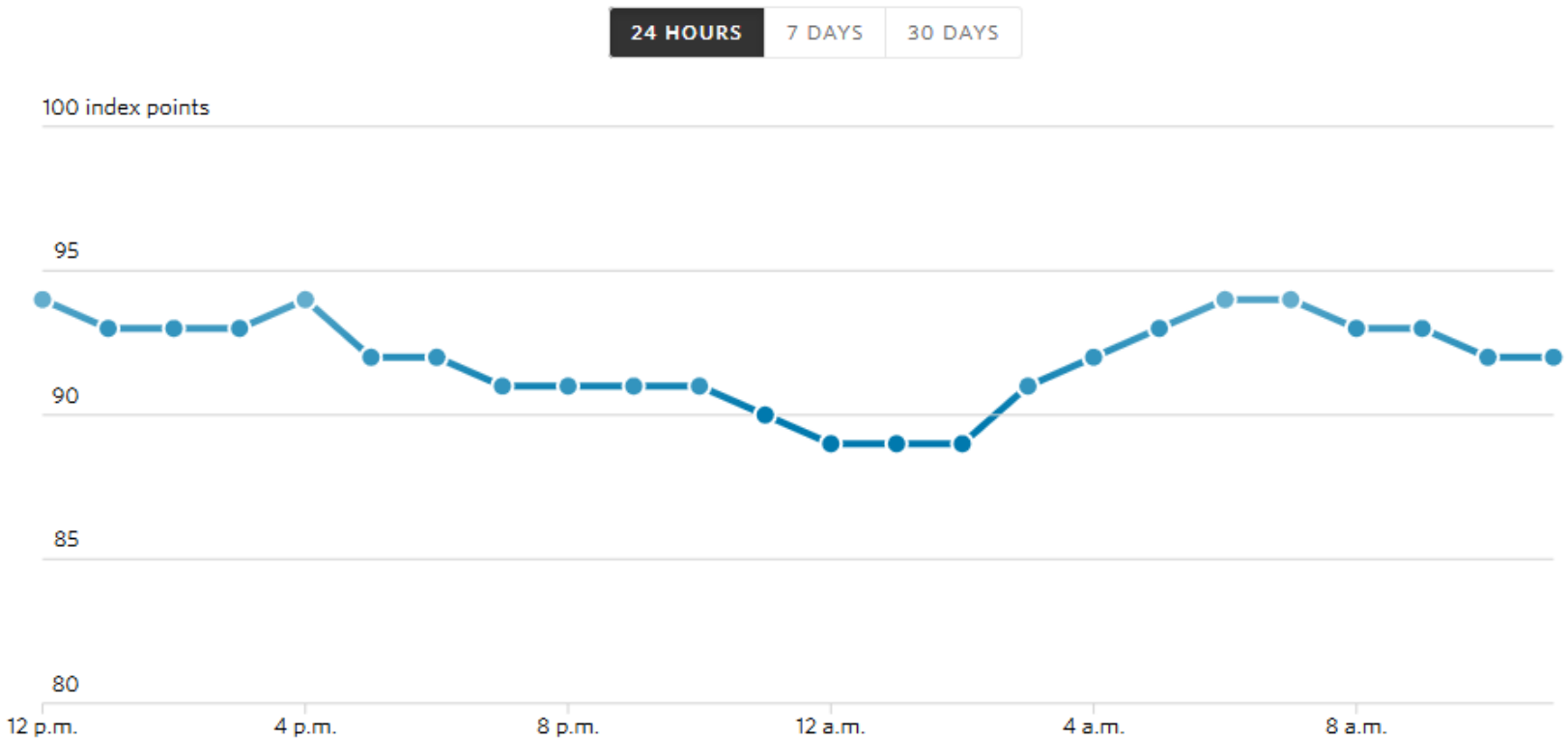
Lower than the average Saturday at this hour

This section, updated hourly, compares current sentiment to the historical average.

< 91	91-93	94-96	97-99	100	101-103	104-106	107-109	110 +

◄ Negative mood Historical average sentiment Positive mood ►

http://graphics.wsj.com/twitter-sentiment/


24 HOURS
7 DAYS
30 DAYS
100 index points
95
90
85
80
12 p.m.
4 p.m.
8 p.m.
12 a.m.
4 a.m.
8 a.m.

MOST POSITIVE STATES

State	Sentiment Index	vs. U.S. average
Ark.	100	+8
Neb.	98	+6
S.D.	98	+6
Ala.	97	+5
N.D.	97	+5
Iowa	96	+4
Idaho	96	+4
Utah	95	+3
N.H.	95	+3
Vt.	95	+3

Five highest scores shown

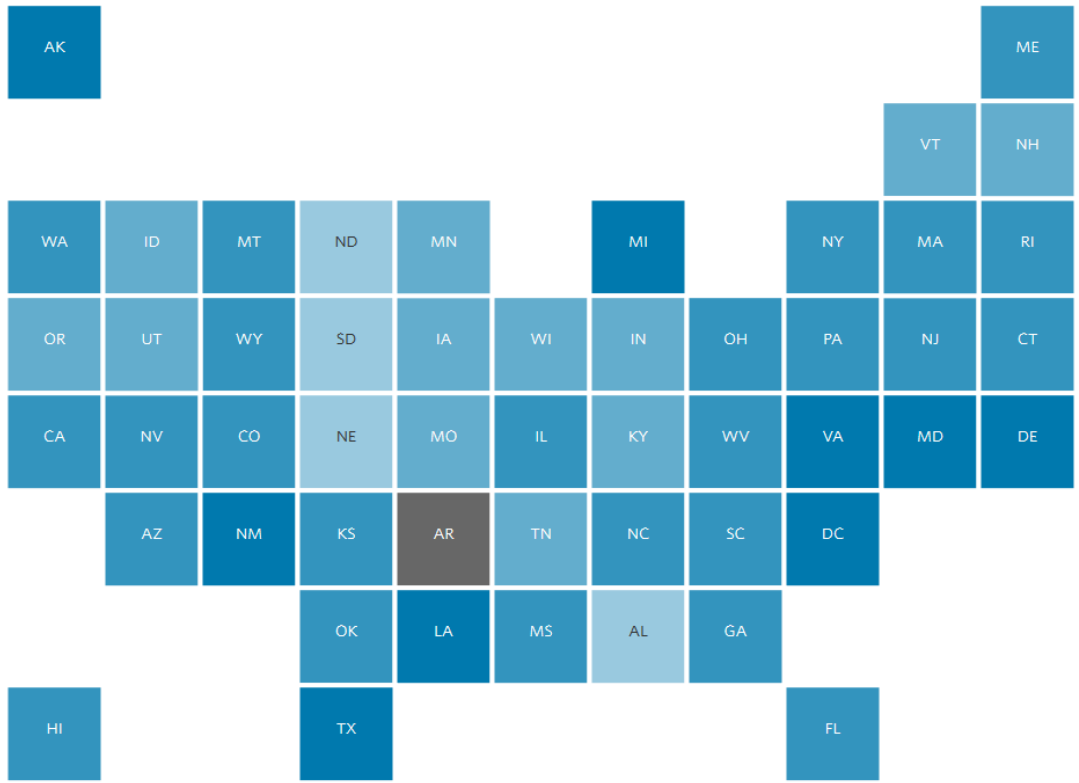
MOST NEGATIVE STATES

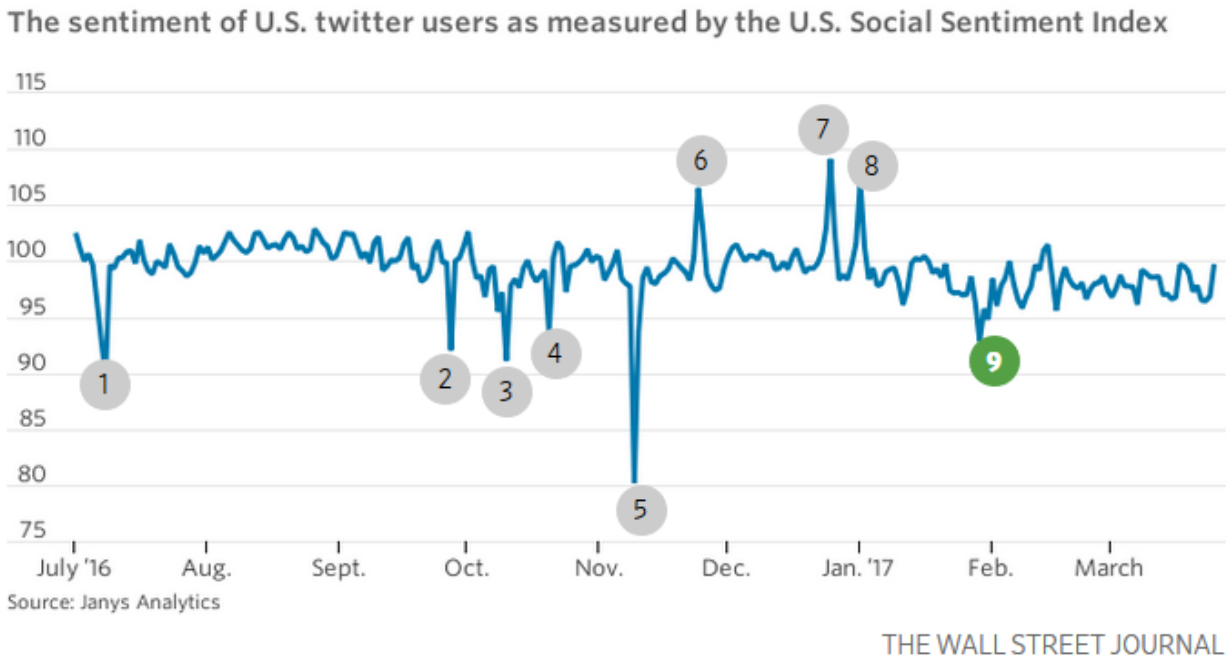
State	Sentiment Index	vs. U.S. average
Del.	87	-5
N.M.	88	-4
La.	88	-4
Va.	88	-4
Md.	88	-4
D.C.	89	-3
Alaska	90	-2
Texas	90	-2
Mich.	90	-2
Conn.	91	-1
S.C.	91	-1
Ariz.	91	-1
Fla.	91	-1
Miss.	91	-1
N.J.	91	-1
Wash.	91	-1
Ga.	91	-1

Five lowest scores shown

SENTIMENT BY STATE

U.S. average: 92 (24 hours through 11 a.m. ET on Feb. 2, 2019)





Wall street journal : "*a-new-index-tracks-our-national-mood-one-tweet-at-a-time*"

1. July 8, 2016	The day after five police officers were killed in Dallas.
2. Sept. 27, 2016	The day after the first presidential debate between Hillary Clinton and Donald Trump.
3. Oct. 10, 2016	The day after the second presidential debate.
4. Oct. 20, 2016	The day after the third presidential debate.
5. Nov. 9, 2016	The day after Donald Trump's presidential-election win.
6. Nov. 24, 2016	Thanksgiving Day.
7. Dec. 25, 2016	Christmas Day.
8. Jan. 1, 2017	New Year's Day.
9. Jan. 29, 2017	The weekend of President Donald Trump's executive order on immigration.

DOES IT LOOK LIKE SENTIMENT INDEX?

2. ALPHAGO ZERO – REINFORCEMENT LEARNING

Google DeepMind研發的AlphaGo:

在2016年3月，AlphaGo以4:1戰胜韓國頂尖棋手李世乭九段，讓AI成為了目前最熱門的話題之一。

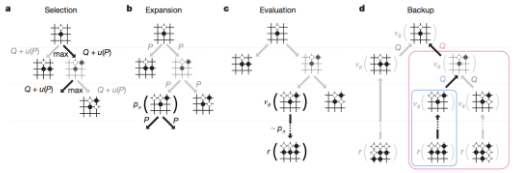
在2017年5月，新版AlphaGo以3:0戰勝當今世界圍棋第一人，中國的柯潔九段。所有棋手都同意它已全面勝過人類，但它仍需要人類棋譜作為訓練的前期輸入。

在2017年10月，名為AlphaGo Zero的最新版已能***完全脫離人類棋譜***，從零開始，純粹依靠自我探索，自我對弈，就能實現超越此前所有版本的棋力。

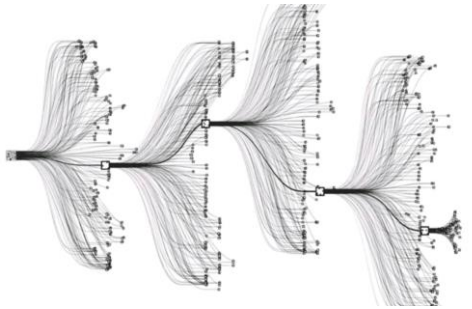
在2017年12月，DeepMind還將AlphaGo Zero的方法用於國際象棋、日本將棋，稱為AlphaZero。它僅需幾個小時的訓練，就打敗了此前世界最強的程序，證明AlphaGo方法的通用性極強。

1. AlphaGo 參考了10萬套棋譜去訓練模型。

2. 據DeepMind透露， AlphaGo Zero只用了四十天去訓練模型。



近來較新研發的AI 可在24小時內就能趕上學習人類棋譜的AlphaGo，並在40小時內超越(九成勝率)與李世乭對戰的AlphaGo。



AlphaGo Zero 成功的其中一些關鍵:

1. TPU (~40 GTX1018 Ti GPU)
2. Monte Carlo Search
3. 生成式對抗網絡 (Generative Adversarial Networks - GAN)

Source: https://zhuanlan.zhihu.com/p/32378765

Google DeepMind

3. AI照片生成器

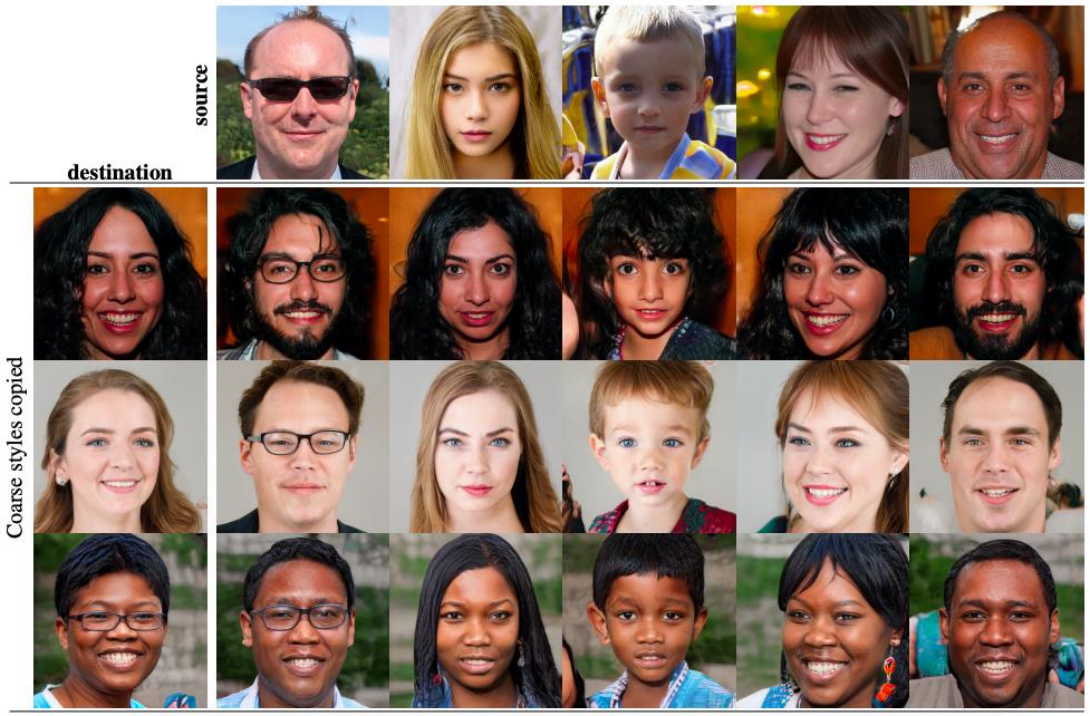
1. This Person Does Not Exist

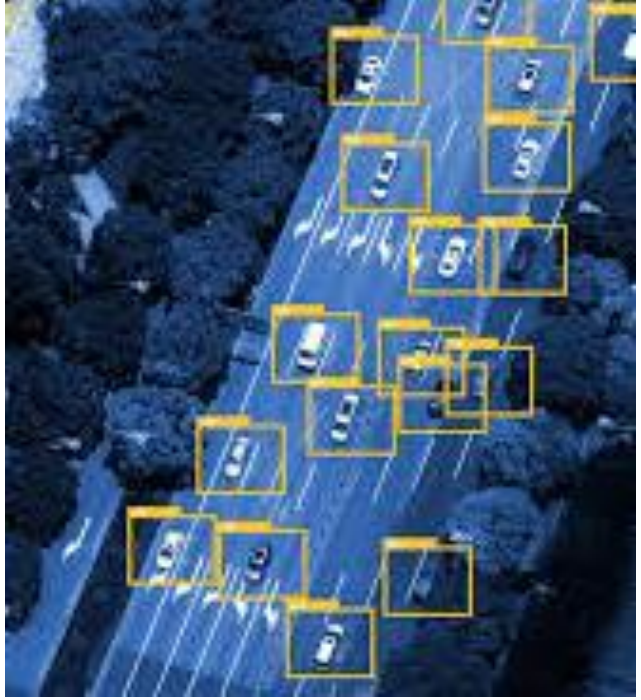
https://www.thispersondoesnotexist.com/

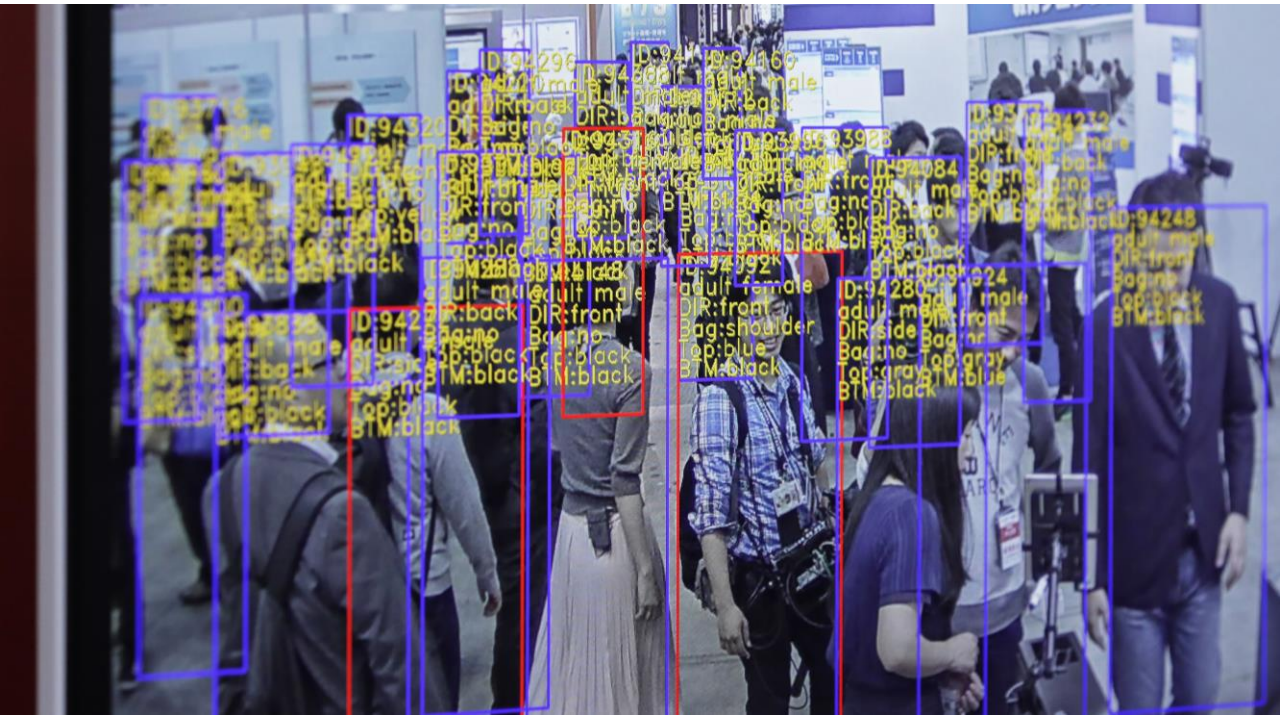
2. Face transform

Source:
https://www.theverge.com/2018/12/17/18144356/ai-image-generation-fake-faces-people-nvidia-generative-adversarial-networks-gans

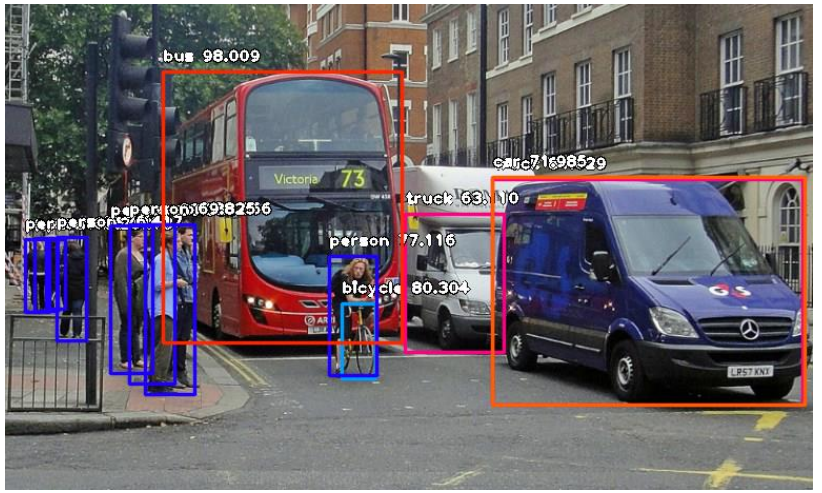
Source:
Youtuhttps://www.youtube.com/watch?v=kSLJriaOumA





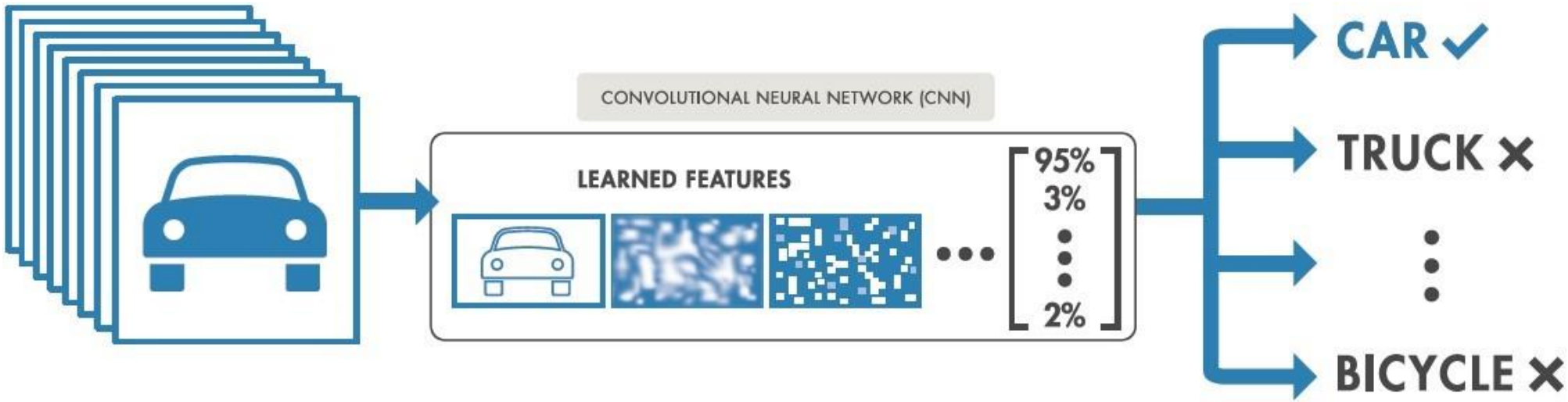


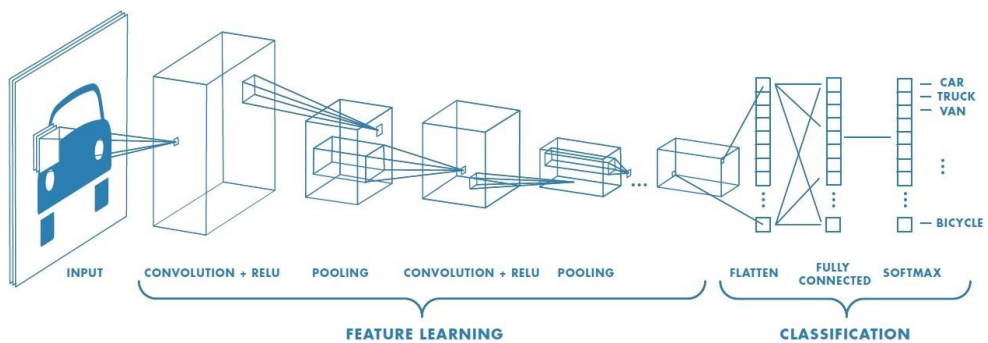
4.圖像處理

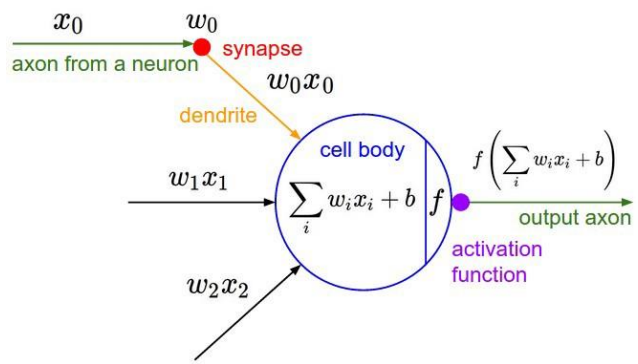


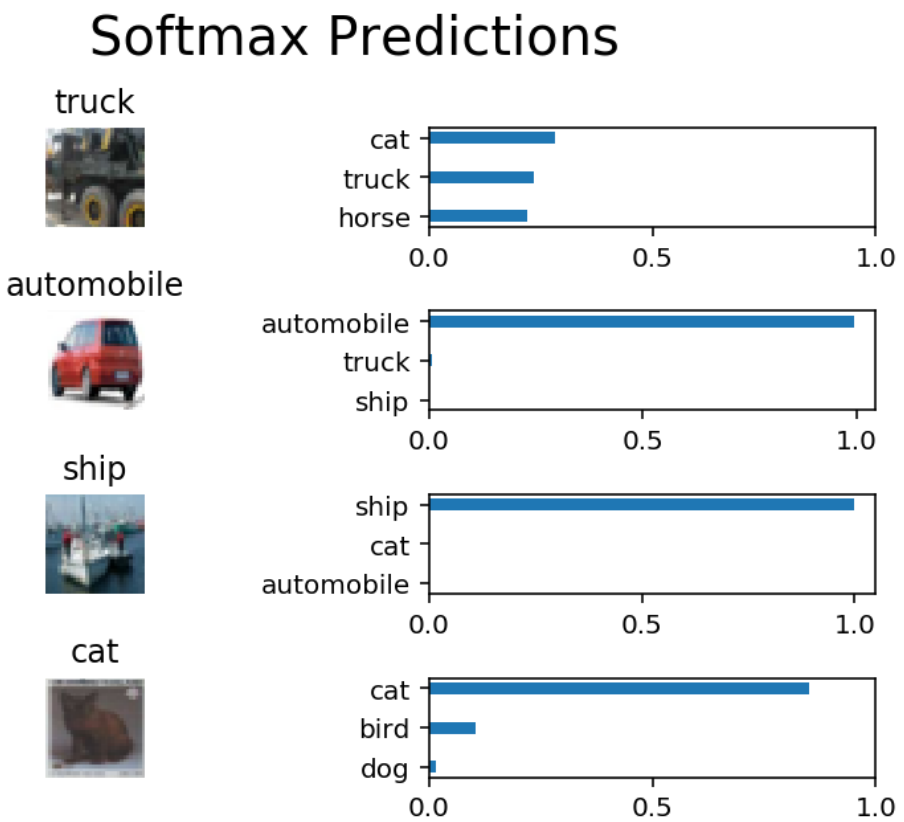
計算遊行人數

https://hk.news.appledaily.com/local/realtime/article/20190611/59697822


CONVOLUTIONAL NEURAL NETWORK (CNN)
LEARNED FEATURES
95%
3%
2%
CAR ✓
TRUCK ✗
BICYCLE ✗

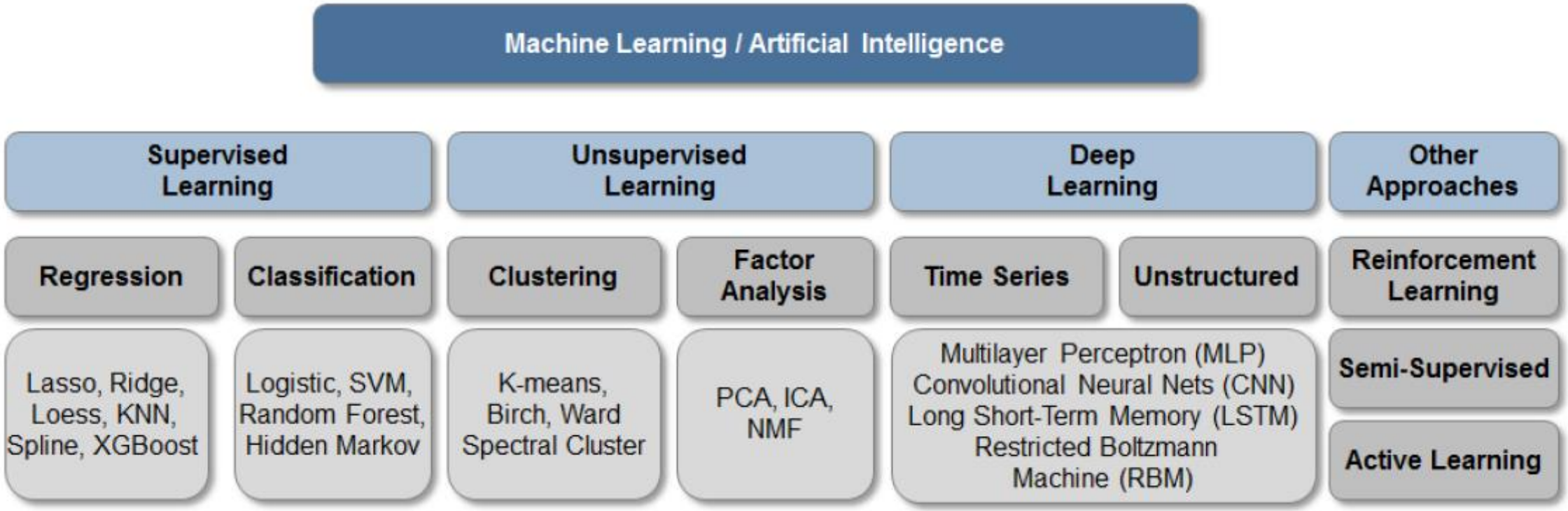

CAR
TRUCK
VAN
BICYCLE
INPUT
CONVOLUTION + RELU
POOLING
CONVOLUTION + RELU
POOLING
FLATTEN
FULLY CONNECTED
SOFTMAX
FEATURE LEARNING
CLASSIFICATION


x_0
w_0
axon from a neuron
synapse
$w_0 x_0$
dendrite
cell body
$w_1 x_1$
$\sum_i w_i x_i + b$
f
$f\left(\sum_i w_i x_i + b\right)$
output axon
activation function
$w_2 x_2$

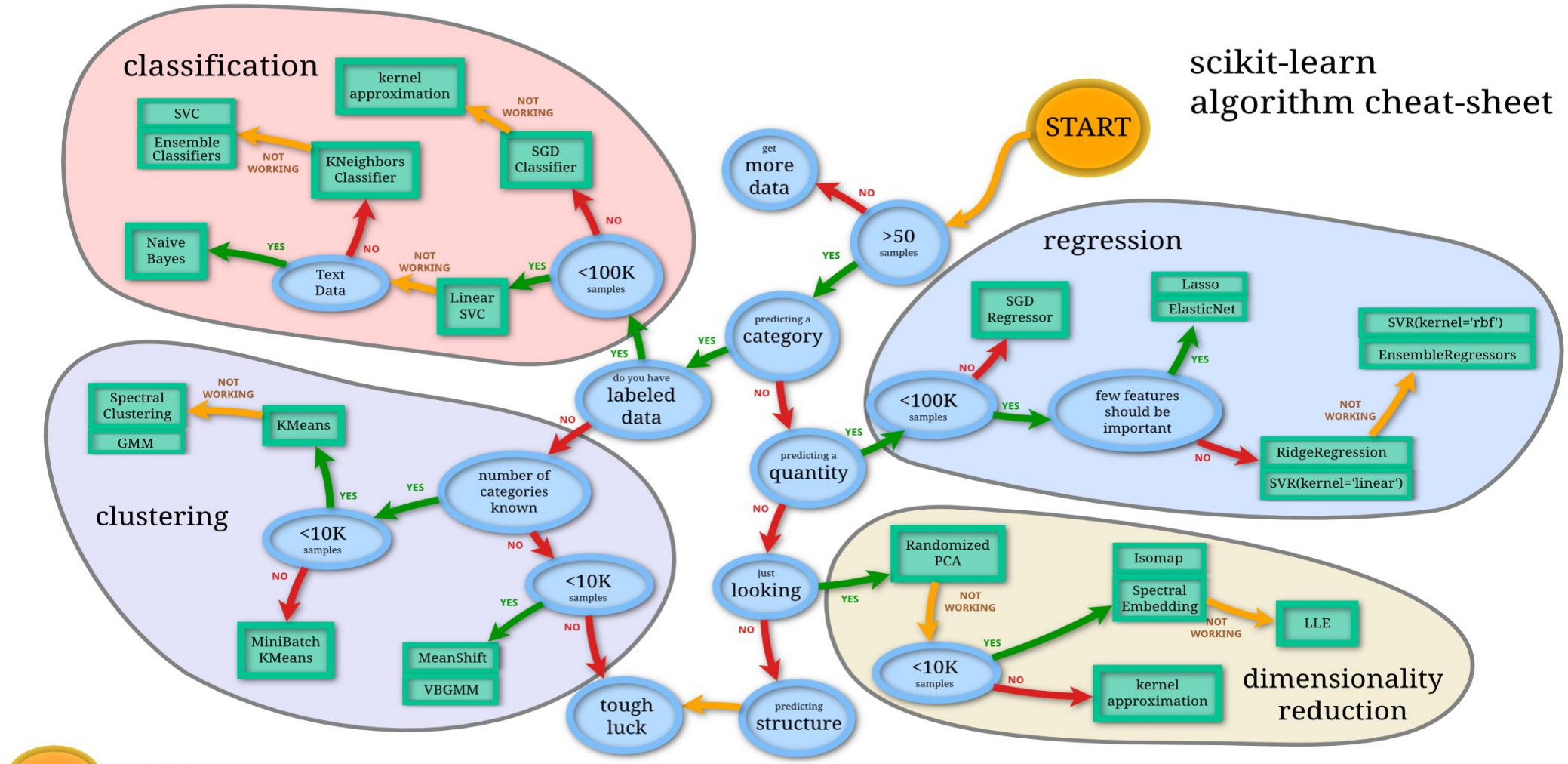


Medium: Dhanoop Karunakaran

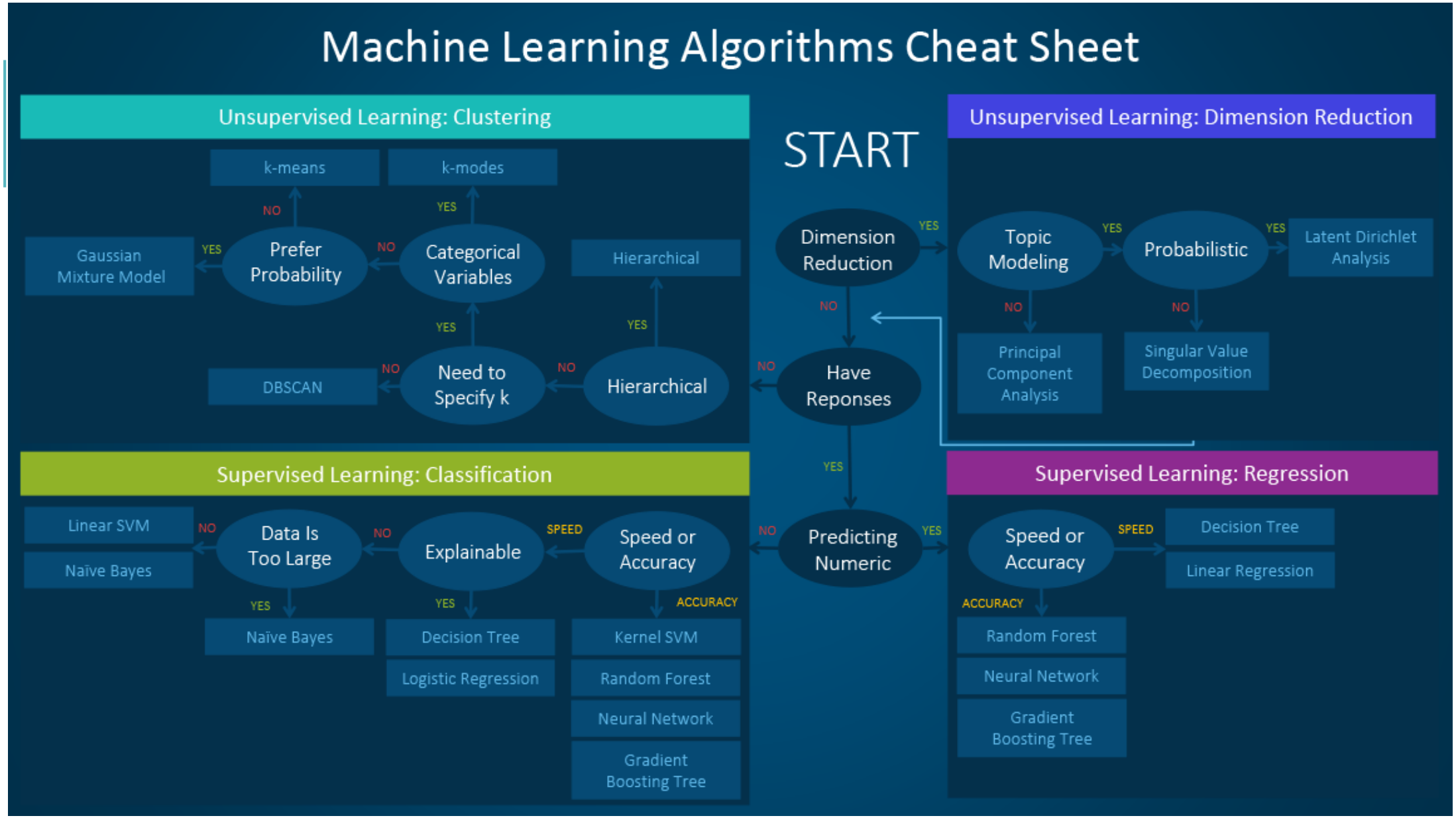
機器學習方法概覽


Machine Learning / Artificial Intelligence
Supervised Learning
Unsupervised Learning
Deep Learning
Other Approaches
Regression
Classification
Clustering
Factor Analysis
Time Series
Unstructured
Reinforcement Learning
Lasso, Ridge, Loess, KNN, Spline, XGBoost
Logistic, SVM, Random Forest, Hidden Markov
K-means, Birch, Ward Spectral Cluster
PCA, ICA, NMF
Multilayer Perceptron (MLP)
Convolutional Neural Nets (CNN)
Long Short-Term Memory (LSTM)
Restricted Boltzmann Machine (RBM)
Semi-Supervised
Active Learning

J.P. Morgan



Source: https://scikit-learn.org/stable/tutorial/machine_learning_map/



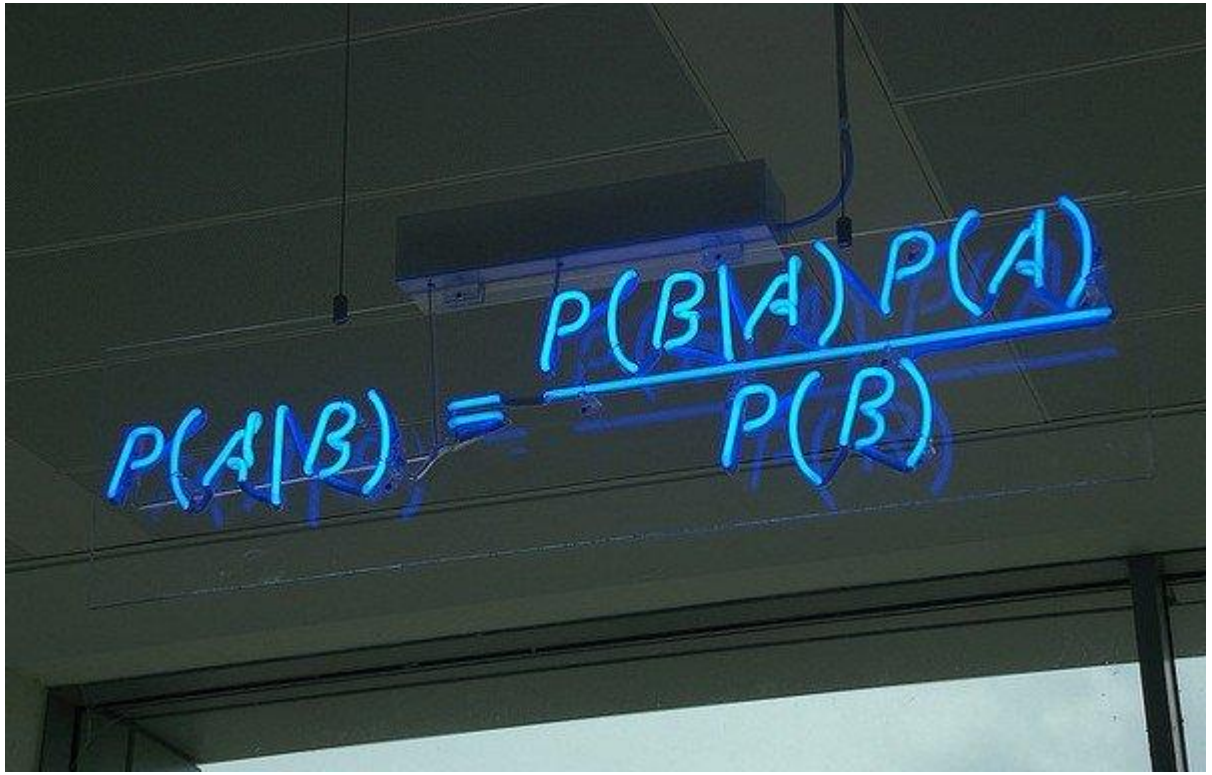
Source: https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

課程範圍

機器學習方法

1. 迴歸模型 Regression model
2. 分類模型 Classification
3. 分群模型 Clustering

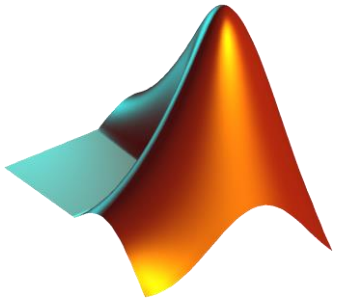
機器學習更多介紹

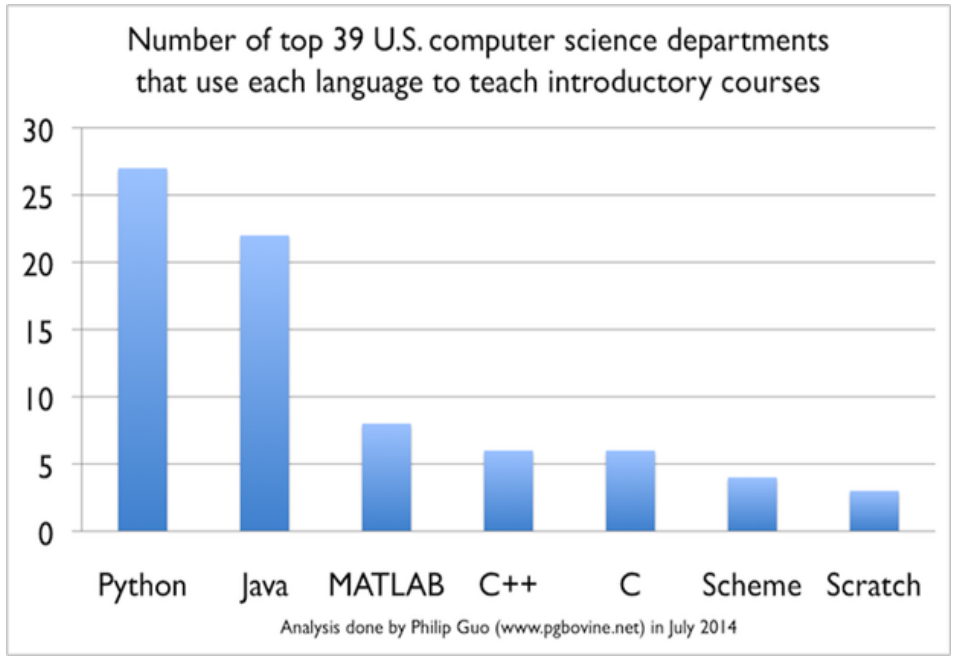


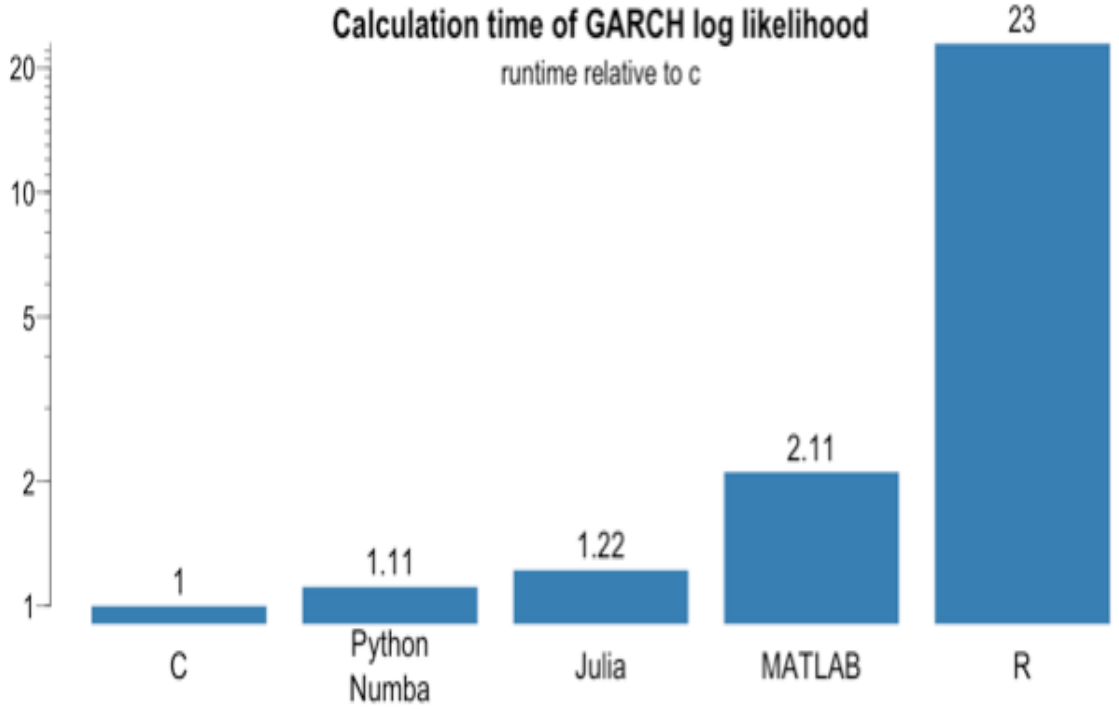
https://bigdatafinance.tw/index.php/tech/564-2018-03-28-09-55-07

機器學習的編程語言比較

	Speed	Support	Easy to learn (to me only!)	Popularity	Cost
Python	Fast	Best	Fair	Highest	Free
R	Slow	Good	Fair	Fair	Free
Matlab	Fast	Good	Easiest	Fair	$800/yr

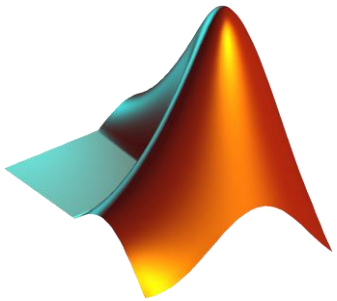



Number of top 39 U.S. computer science departments
that use each language to teach introductory courses
30
25
20
15
10
5
0
Python
Java
MATLAB
C++
C
Scheme
Scratch
Analysis done by Philip Guo (www.pgbovine.net) in July 2014


Calculation time of GARCH log likelihood
runtime relative to c
20
10
5
2
1
1
1.11
1.22
2.11
23
C
Python
Numba
Julia
MATLAB
R


jupyter
spyder
ANACONDA


R



關於機器學習的一些關鍵概念

Development flow: *Data – Model – Evaluation - Apply*

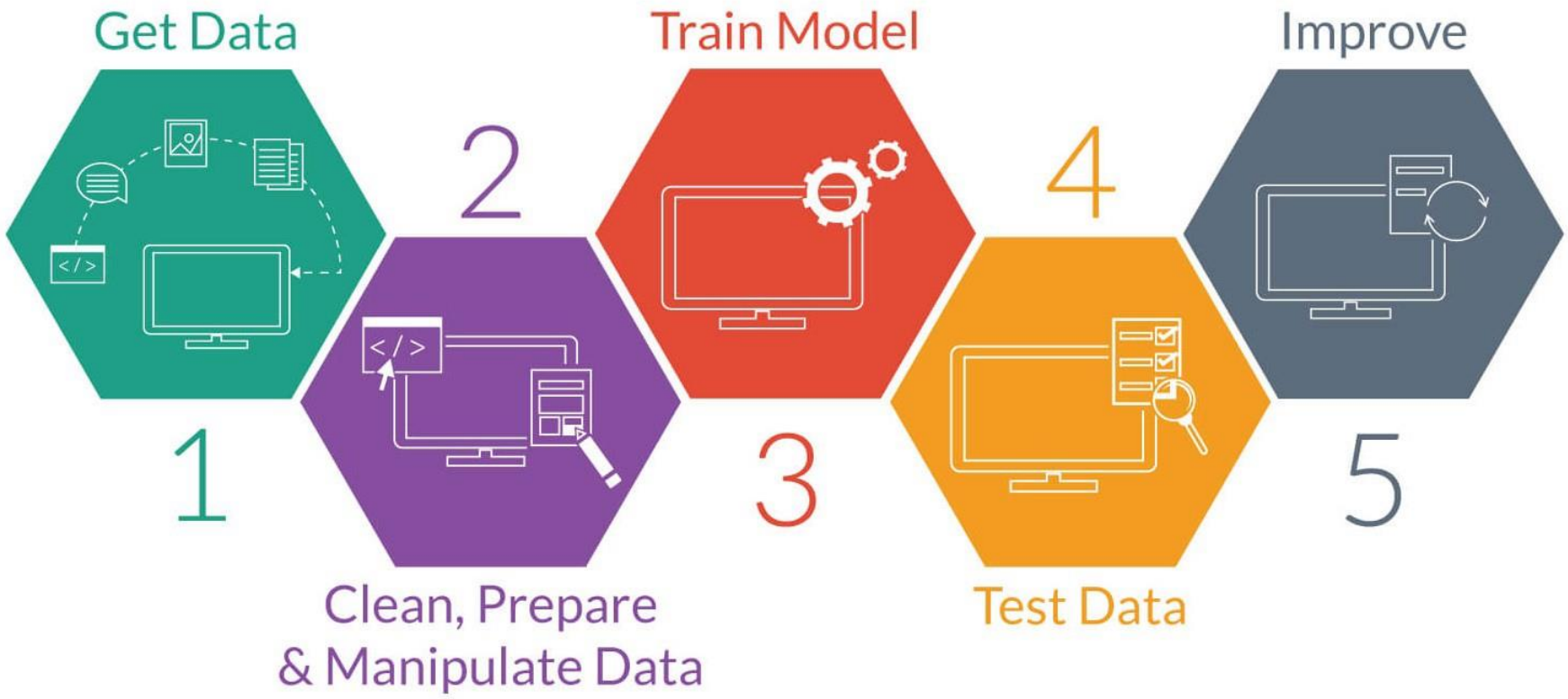
Over-fit vs under-fit

Over-parametrization: *Penalization*

Performance evaluation: *Train set, Test set, Validation set*

Enhancing methods: *Cross Validation, Bagging, Boosting, Ensembling*

開發流程


Get Data
1
Clean, Prepare
& Manipulate Data
2
Train Model
3
Test Data
4
Improve
5

矩陣修覆

	Inside Out	Good Will Hunting	Mean Girls	Terminator	Titanic	Warrior
						
Tina Fey 	3	1	5	1	?	1
Helen Mirren 	2	?	?	2	5	1
Sylvester Stallone 	1	3	1	4	2	5
Tom Hanks 	?	3	1	?	4	3
George Clooney 	2	2	1	3	1	4

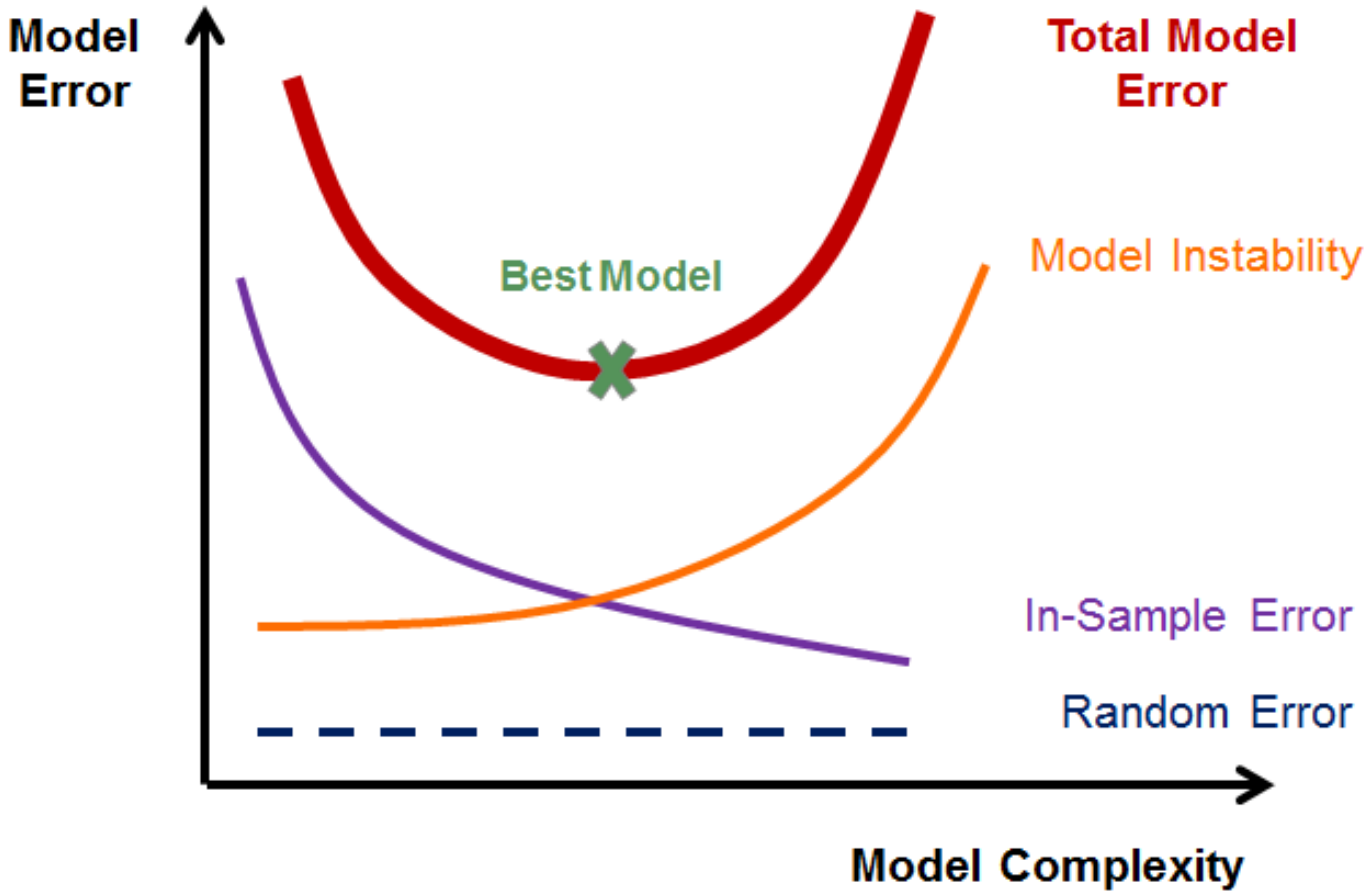
開發流程

0. 收集您的數據。

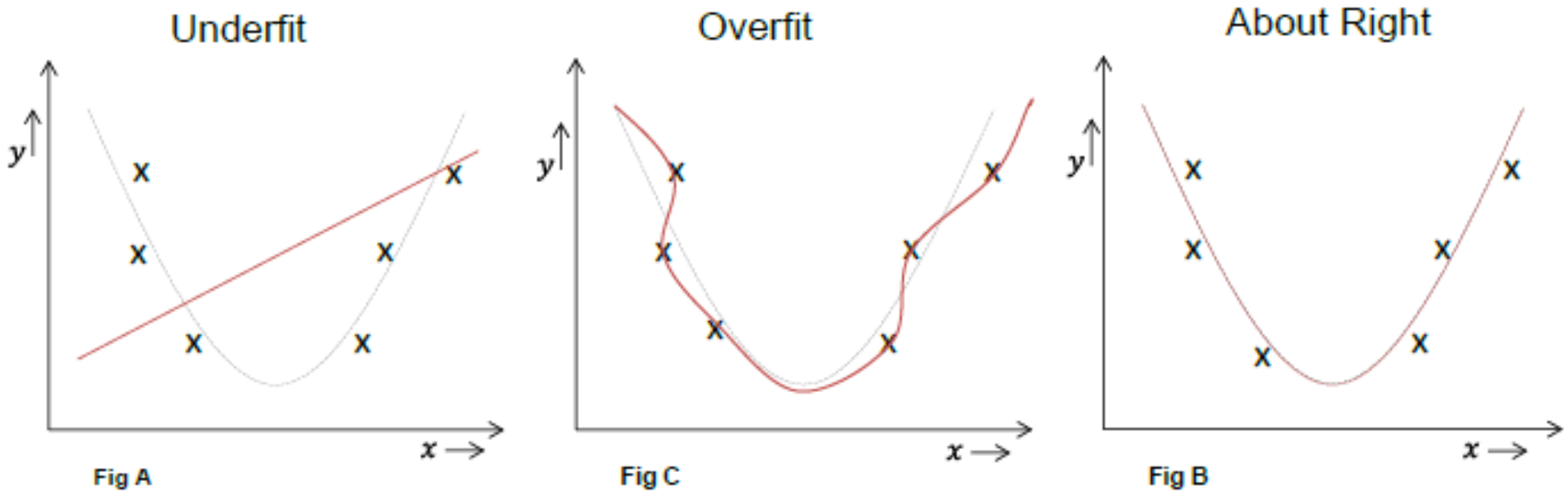
1. 數據預處理 (data preprocessing)
2. 開發數據模型
   - 選擇方法
   - 訓練模型
   - 評估模型準確性
   - 調整超參數 (hyperparameters)
3. 模型評估：訓練集，測試集和驗證集 (Train set, Test set and Validation set)
4. 應用訓練好的模型

These stages are iterative

過度訓練vs訓練不足

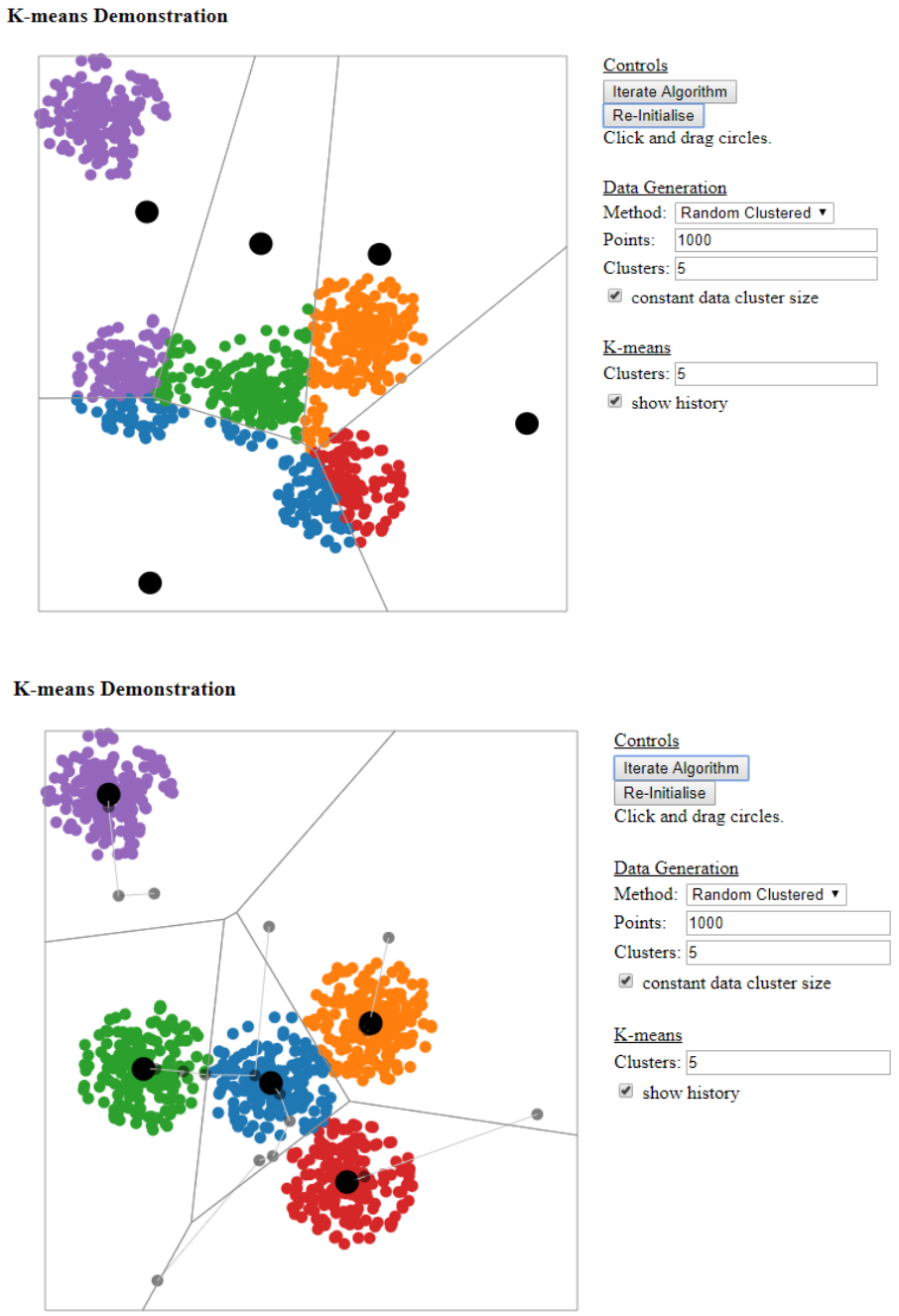

Model
Error
Total Model
Error
Best Model
Model Instability
In-Sample Error
Random Error
Model Complexity

過度訓練vs訓練不足



互動演示（K-MEANS集群）

Source: http://alekseynp.com/viz/k-means.html



今天課堂總結

1. 什麼是機器學習？機器學習的「前世今生」
2. 人工智能技術的前沿
3. 機器學習方法概述
4. 機器學習的編程語言比較
5. 關於機器學習的一些關鍵概念
6. 互動示範（K-means集群）

下一課...

Python基本操作：

1. 基本統計分析
2. 數據輸入和輸出
3. 有用的package（例如numpy和dataframe）
4. 定義函數
5. 製作圖表