

Kaggle入門心得

2017/11/15(三)

關於Kaggle

關於Kaggle

Kaggle是一個數據建模和數據分析競賽平台。企業和研究者可在其上發布數據，統計學者和數據挖掘專家可在其上進行競賽以產生最好的模型。這一眾包模式依賴於這一事實，即有眾多策略可以用於解決幾乎所有預測建模的問題，而研究者不可能在一開始就了解什麼方法對於特定問題是最為有效的。Kaggle的目標則是試圖通過眾包的形式來解決這一難題，進而使數據科學成為一場運動。(wiki)

The Kaggle logo, featuring the word "kaggle" in a lowercase, blue, sans-serif font.

成立	2010年4月
創辦人	安東尼·戈德布盧姆
代表人物	安東尼·戈德布盧姆 (CEO) 馬克斯·列夫琴 (董事局主席) 傑夫·莫澤 (CTO)
總部	 美國舊金山
標語口號	Making Data Science a Sport 使數據科學成為一項運動
產業	預測建模
網站	www.kaggle.com 

關於Kaggle

Google 收購 Kaggle (2017/3月)



加入 Google 後，我們能夠提供社群 Google 雲技術。這將使大家能利用更強大的基礎設施和部署服務（deployment services），進行可延伸的訓練，並幫助 Kaggle 擁有儲存、抓取大型資料集的能力。

Your Home for Data Science

Kaggle helps you learn, work, and play

Create an account

or

Host a competition



<https://technews.tw/2017/03/10/kaggle-joins-google-cloud/>

Kaggle平台上的資源

豐富的資料集可供學習研究

討論區有很多範例教學說明

獎金 & 工作 & 社群

Kaggle 首頁

<https://www.kaggle.com/>

kaggle

Search kaggle



Competitions

Datasets

Kernels

Discussion

Jobs



Sign In

Sign in or sign up with one click:

We won't share anything without your permission.



or

Use your Kaggle username or email:

[Manually create a new account »](#)

Username or Email

Password

Sign in

☐ Remember me

[Forgot Username / Password](#)

Kaggle Competitions

Active

All


Entered

Sort by Prize


19 active competitions

All Categories


Search competitions



Passenger Screening Algorithm Challenge
Improve the accuracy of the Department of Homeland Security's threat recognition algorithms
Featured · a month to go · terrorism, image, object detection



Zillow Prize: Zillow's Home Value Prediction (Zestimate)
Can you improve the algorithm that changed the world of real estate?
Featured · 2 months to go · housing, real estate



StatOil/C-CORE Iceberg Classifier Challenge
Ship or iceberg, can you decide from space?
Featured · 2 months to go · weather, shipping, binary classification

\$1,500,000

374 teams

\$1,200,000

3,780 teams

\$50,000

879 teams

Kaggle Datasets


1,133 featured datasets

Sort by Hotness

Featured All

Q Search datasets


339



Kaggle ML and Data Science Survey, 2017
A big picture view of the state of data science and machine learning.
Kaggle · updated 14 days ago · data analysis, employment, sociology, artificial intelligence

4,393 downloads
5 comments

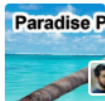
34



The Holy Quran
Understanding God - Sacred Meanings
Zeeshan-ul-hassan Usmani · updated 16 hours ago · languages, data analysis, religious fait...

129 downloads
4 comments

16



Paradise-Panama-Papers
Data Scientists United Against Corruption
Zeeshan-ul-hassan Usmani · updated 2 days ago

104 downloads
2 comments

Kaggle Kernels

Sort by













Hotness

All

Mine

All Languages

All Output Types

725			Comprehensive data exploration with Python run 3 hours ago by Pedro Marcelino in House Prices: Advanced Regression Techniques beginner, eda, data cleaning	Py	250
168			Data ScienceTutorial for Beginners run 5 hours ago by Kaan Can in Pokemon Challenge	Py	77
270			Predicting Employee Kernelover run 9 hours ago by randy lao in Human Resources Analytics statistics, business, finance, eda, data visualization	Py	131
11			Novice to Grandmaster run 2 hours ago by I,Coder in Kaggle ML and Data Science Survey, 2017 data analysis, eda, data visualization, survey analysis	Py	
47			Porto Seguro: The Essential Kickstarter run 5 hours ago by Andrea in Porto Seguro's Safe Driver Prediction eda, data visualization	Py	16
21			PyTorch GPU CNN & BCELoss with predictions run 2 hours ago by QuantScientist in Statoil/C-CORE Iceberg Classifier Challenge	Py	15

Kaggle Discussion

Site Forums



Kaggle Forum

Events and topics specific to our community

last post 2 hours ago



Getting Started

The first stop for new Kagglers

last post an hour ago



Product Feedback

A feature wish list for our engineers

last post 16 hours ago



Questions & Answers

Technical advice from other data scientists

last post an hour ago



Datasets

Requests for and discussion of open data

last post 4 hours ago



InClass

Kaggle InClass questions, answers, and general information

last post 2 days ago

30,203 topics

Sort by Hotness

All Mine | Upvoted

All Categories Topics

Search topics



48



5 things I learned from this competition

[Bert Carremans](#) 5d ago in Porto Seguro's Safe Driver Prediction

last comment by

[Marc](#) 2h ago

15

14



Collaborative RAMP team

[Balazs Kegl](#) 4d ago in Porto Seguro's Safe Driver Prediction

last comment by

[Balazs Kegl](#) 7m ago

22

Kaggle jobs

Featured Posts



★ Machine Learning / AI Architect – Research & Develop...

Citrix Systems Inc. · Patras, Greece

posted 3 days ago

496

views



★ Senior Data Scientist

Telenor Norway · Oslo

posted 3 days ago

405

views



★ Research Data Scientist

University of Sydney · Sydney

posted 4 days ago

535

views



★ Data Scientist

Haraj · Riyadh, Saudi Arabia

posted 5 days ago

595

views

Kaggle Community

› The Official Blog of
Kaggle.com

🔍 Search

📖 Categories

[DATA SCIENCE NEWS](#) (61)

[KAGGLE NEWS](#) (133)

[KERNELS](#) (40)

[OPEN DATASETS](#) (4)

[TUTORIALS](#) (45)

[WINNERS' INTERVIEWS](#) (214)

No Free Hunch

KAGGLE.COM

Introducing Kaggle's State of Data Science & Machine Learning Report, 2017

[Mark McDonald](#) | 10.30.2017



In 2017 we conducted our first ever extra-large, industry-wide survey to capture the state of data science and machine learning. As the data science field booms, so has our community. In 2017 we hit a new milestone of reaching over 1M registered data

September Kaggle Dataset Publishing Awards Winners' Interview

[Mark McDonald](#) | 10.25.2017

This interview features the stories and backgrounds of our \$10,000 Datasets Publishing Award's September winners—Khuram Zaman, Mitchell J. and Dave Fisher-Hickey. If you're inspired to publish your own datasets on Kaggle and vie for next month's prize, check out this page for more details. First Place, Religious Texts Used By ISIS by Fifth Tribe (Khuram Zaman) Can you tell us a little about your background? I'm the CEO of a digital agency called Fifth Tribe based out of 1776 in Crystal ...

Kaggle Competition Rules

One account per participant

You cannot sign up to Kaggle from multiple accounts and therefore you cannot submit from multiple accounts.

No private sharing outside teams

Privately sharing code or data outside of teams is not permitted. It's okay to share code if made available to all participants on the forums.

Team Mergers

Team mergers are allowed and can be performed by the team leader. In order to merge, the combined team must have a total submission count less than or equal to the maximum allowed as of the merge date. The maximum allowed is the number of submissions per day multiplied by the number of days the competition has been running.

Team Limits

There is no maximum team size.

Submission Limits

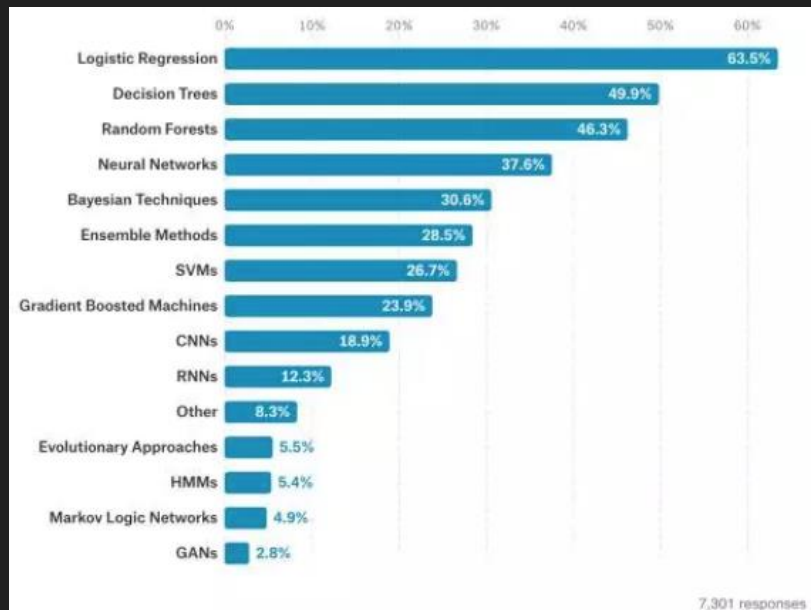
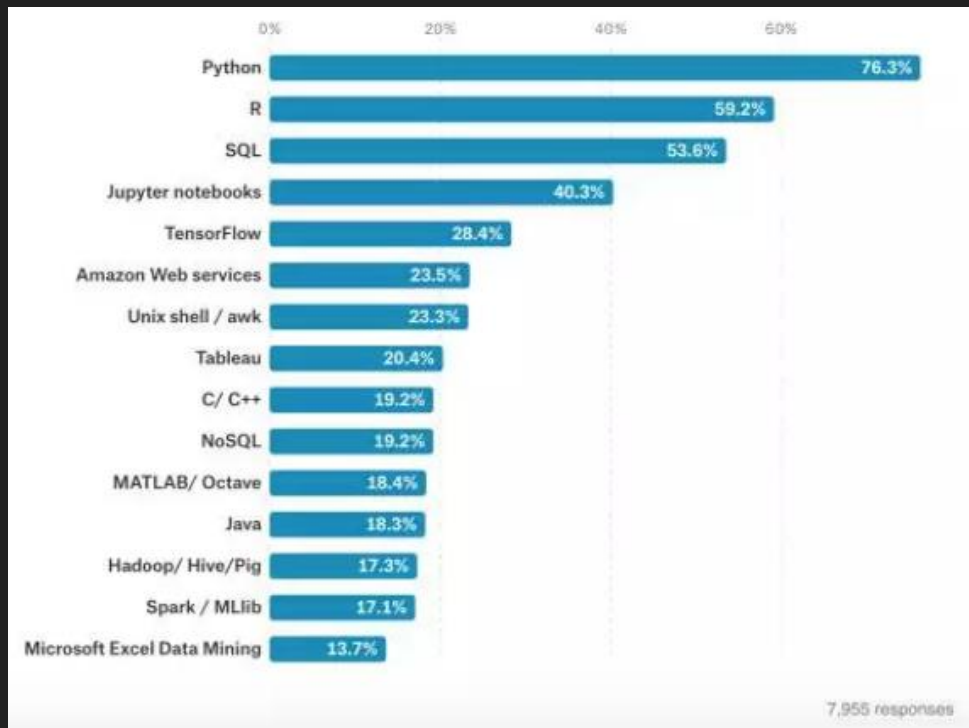
You may submit a maximum of 10 entries per day.

You may select up to 5 final submissions for judging.

Competition Timeline

如何開始

Kaggle問卷調查(常用的語言及演算法)



事前準備

電腦配備:

ML: 至少 8G RAM / DL: 需有支援Nvidia CUDA 的顯卡

主流語言與工具:

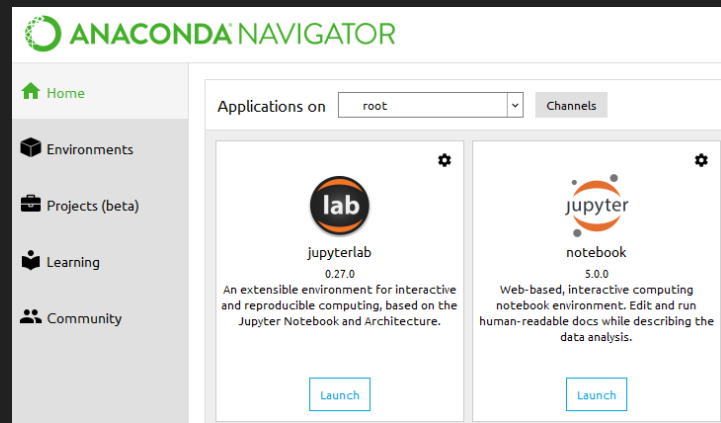
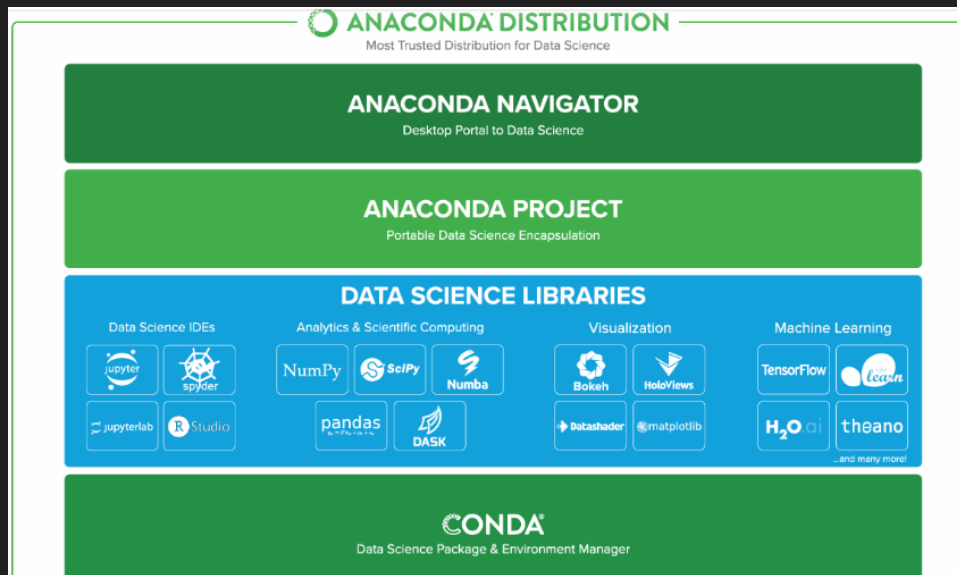
python:

scikit-learn, pandas, matplotlib, seaborn, xgboost,
tensorflow, keras...etc.

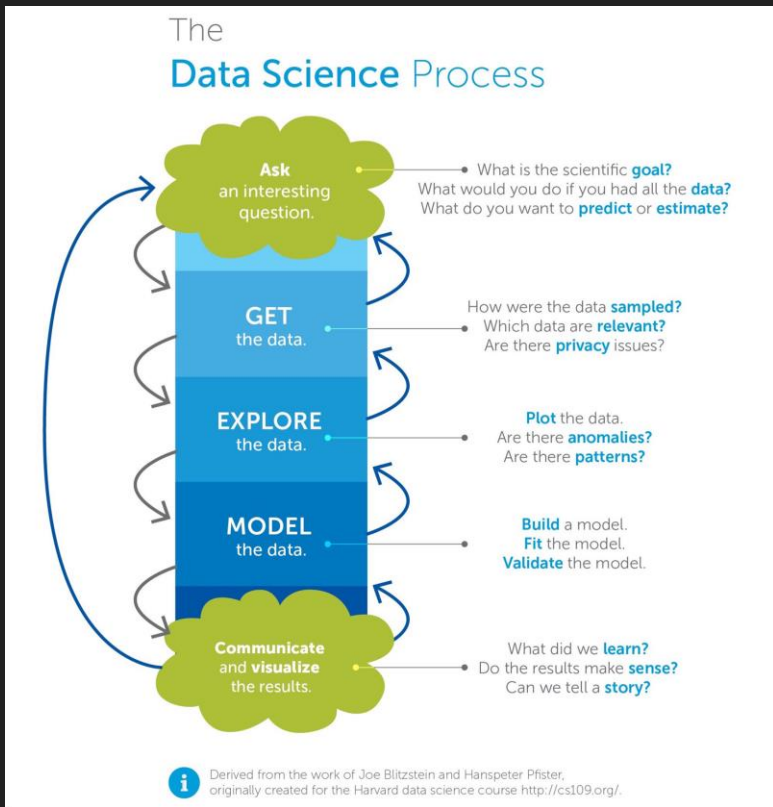
python 環境安裝

IDE: Anaconda <https://www.anaconda.com/>

安裝步驟請參考: http://tensorflowkeras.blogspot.tw/2017/08/tensorflowkeraswindows_29.html



資料科學分析流程



新手常見問題:

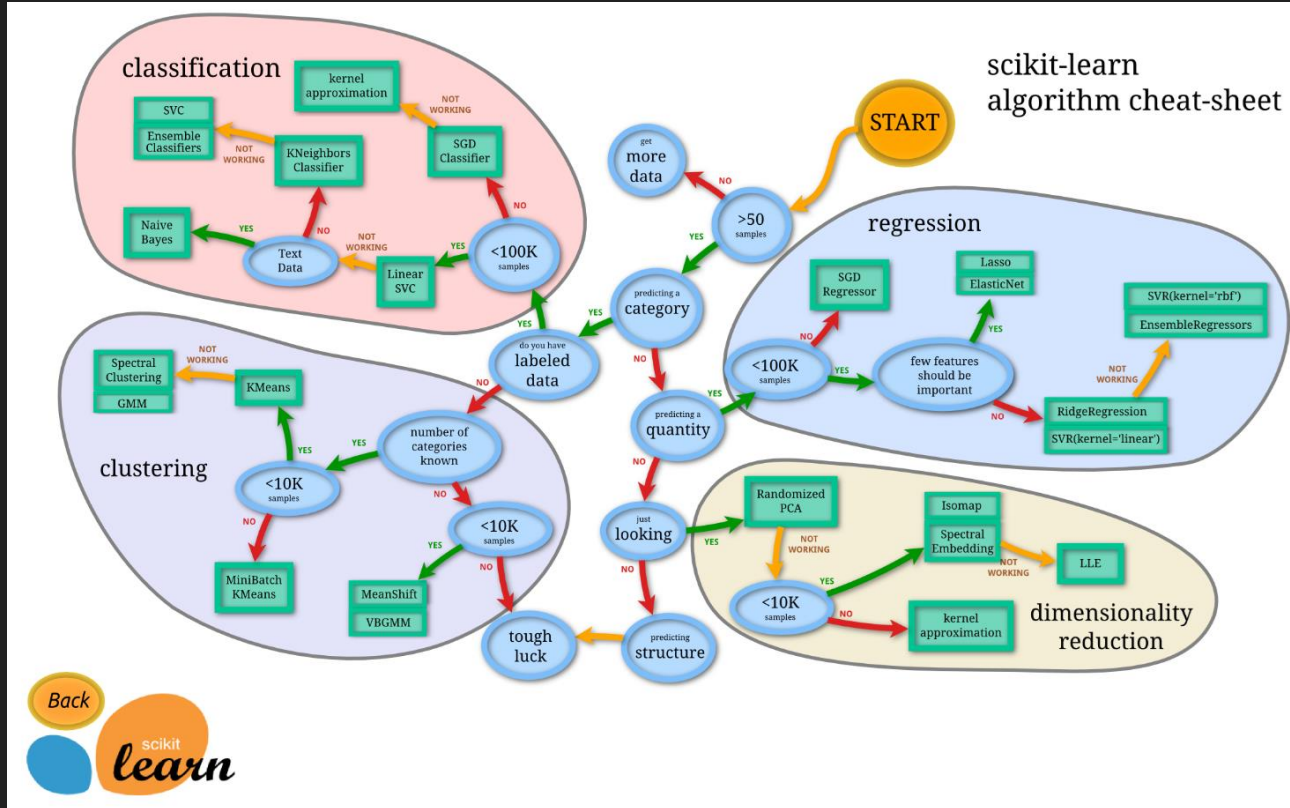
Data Clean: 缺值(missing value)該怎麼填補?

Feature engineering:
怎麼把lable(文字)資料轉成數值?
Feature/變數,如何取捨?

Modeling:
模型選擇? 參數調校?

Ensemble:
整體學習建模的方法?

Model selection



程式範例

https://github.com/stuser/temp/tree/master/kaggle_intro

1213 lines (1212 sloc) | 255 KB

Raw

Blame

History



```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

```
In [2]: import warnings
warnings.filterwarnings('ignore')
```

資料載入與探索

```
In [3]: # Load in the train datasets
train = pd.read_csv('input/train.csv', encoding = "utf-8", dtype = {'type': np.int32})
test = pd.read_csv('input/test.csv', encoding = "utf-8")
submission = pd.read_csv('input/submission.csv', encoding = "utf-8", dtype = {'type': np.int32})
```

```
In [4]: train.head(3)
```

```
Out[4]:
```

	id	花萼長度	花萼寬度	花瓣長度	花瓣寬度	屬種	type
0	1	5.4	3.7	1.5	0.2	Iris-setosa	1
1	2	4.8	3.4	1.6	0.2	Iris-setosa	1
2	3	4.8	3.0	1.4	0.1	Iris-setosa	1

相關學習資源

學習資源

1. 資料科學年會 & 人工智慧年會 (中研院) <http://datasci.tw>

2. Sharecourse > Python機器學習與深度學習實作 <http://www.sharecourse.net>



台灣人工智慧年會 × 台灣資料科學年會

人工智慧 · 智慧台灣

2017 台灣資料科學年會

2017/11/11 (六) ~ 2017/11/12 (日)

中央研究院 · 台北南港

台灣人工智慧學校

啟蒙課程



ShareCourse 學聯網

Python 機器學習與深度學習實作

Python 機器學習與深度學習實作課程介紹影片 1

即刻報名

觀念 x 討論 x 實作

彈性學習的線上課程

Python 機器學習與深度學習實作
Python for Machine Learning & Deep Learning

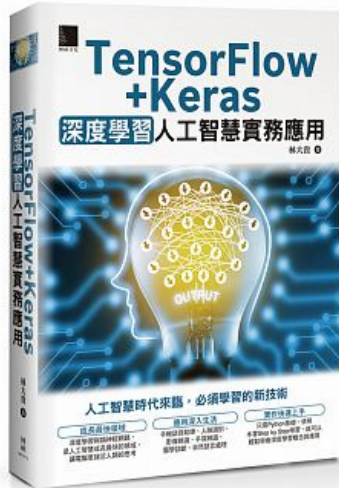
課程開始：	2017-08-25
課程結束：	2017-10-05
課程費用：	NT\$ 6800 元
課程時數：	每週約5小時
課程證書：	證書申請說明與規範
瀏覽人次：	23398

分享到：

Facebook
Google
LINE

推薦參考書

博客來 > 中文書 > 電腦資訊 > 概論科技趨勢 > 人工智慧/機器學習 > 商品介紹



TensorFlow+Keras深度學習人工智慧實務應用

作者：林太壹 [追蹤作者](#) [?](#)

出版社：博碩 [訂閱出版社新書快訊](#) [?](#)

出版日期：2017/06/09

語言：繁體中文

定價：590元

優惠價：**79折 466元**

優惠期限：2017年12月31日止

再折扣 11/8-11/10 中文書暢銷千大，結帳滿699再95折！

【分級買就送】分級VIP會員買就送OPENPOINT(部份除外) 詳情

運送方式：

可配送點：台灣、蘭嶼、綠島、澎湖、金門、馬祖、全球

可取貨點：台灣、蘭嶼、綠島、澎湖、金門、馬祖
香港、澳門、新加坡

中午前訂 可明天 拿

庫存 > 10

放入購物車

直接結帳

加入下次再買清單

我要寫評鑑

分享

讚 73

【facebook粉絲團】

www.facebook.com/TensorflowKeras/

【部落格】

tensorflowkeras.blogspot.tw/

reference

KAGGLE ENSEMBLING GUIDE <https://mlwave.com/kaggle-ensembling-guide/>

Approaching (Almost) Any Machine Learning Problem | Abhishek Thakur

<http://blog.kaggle.com/2016/07/21/approaching-almost-any-machine-learning-problem-abhishek-thakur/>

Facebook - Kaggle粉專: <https://www.facebook.com/kaggle/>

Facebook社團 - 台灣 Kaggle 交流區 <https://www.facebook.com/groups/kaggletw/>

Facebook社團 - Go, Kaggle讀書會交流區 <https://www.facebook.com/groups/384340325250613/>

Facebook社團 - 線上讀書會 <https://www.facebook.com/readbook999/>

以下不負責閒聊

閒聊~~

新機會出現到資訊效率化 -- 成功模式被學習, 團隊分工, 算力產能

Library使用: 早期ML library (CPU), 現在DL library (GPU)

比賽期限內你所能實驗的參數組合 (算力決定你所能解決問題的能力)

時間分配

聽再多，不如動手作