

機器學習於材料資訊的應用

Machine Learning on Material Informatics

陳南佑(NAN-YOW CHEN)

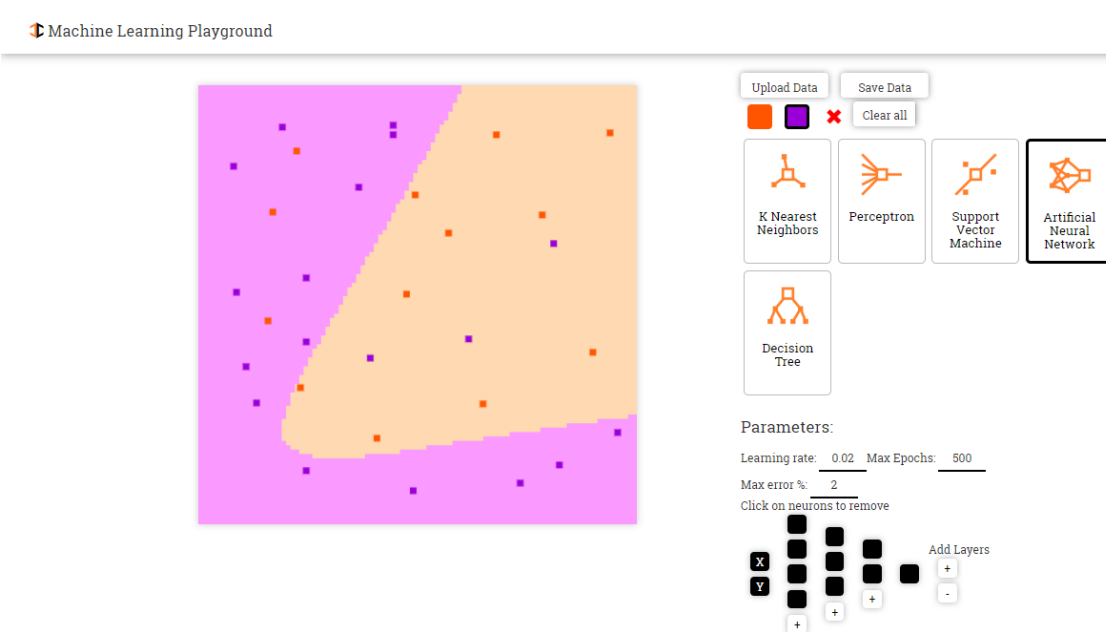
nanyow@narlabs.org.tw

楊安正(AN-CHENG YANG)

acyang@narlabs.org.tw

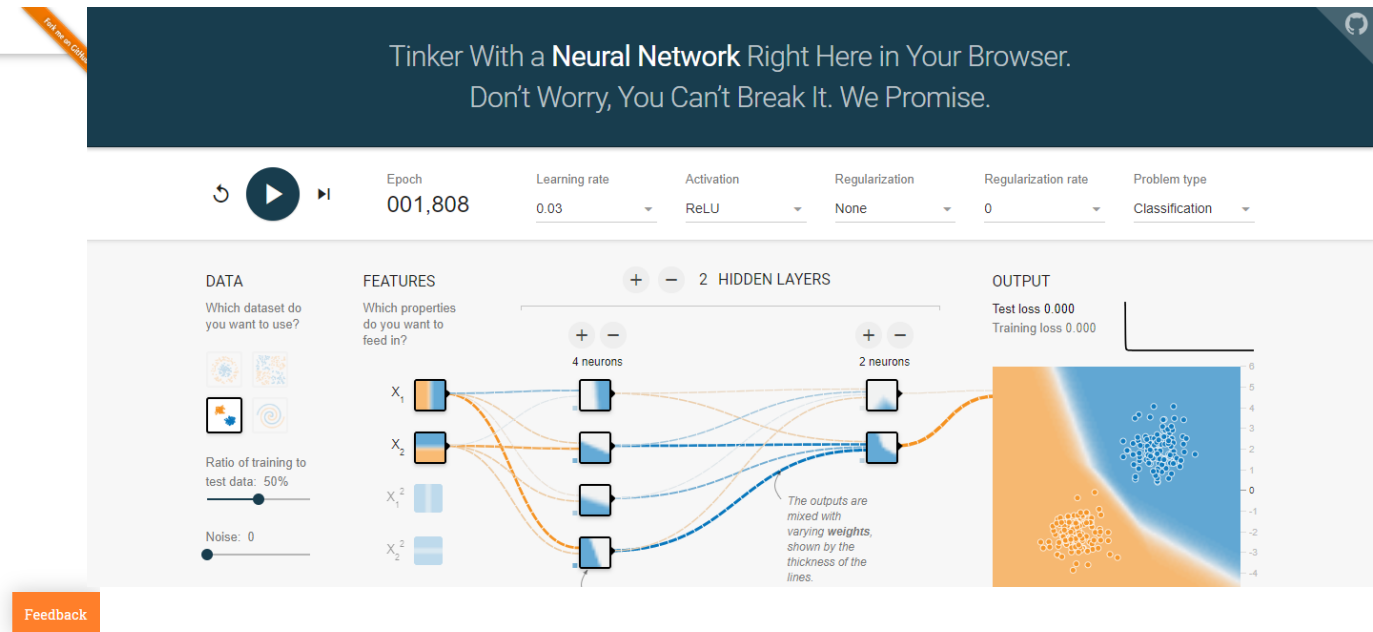
Two online Machine Learning playground

Machine Learning Playground



<http://ml-playground.com>

Play with neural networks!



<https://playground.tensorflow.org>

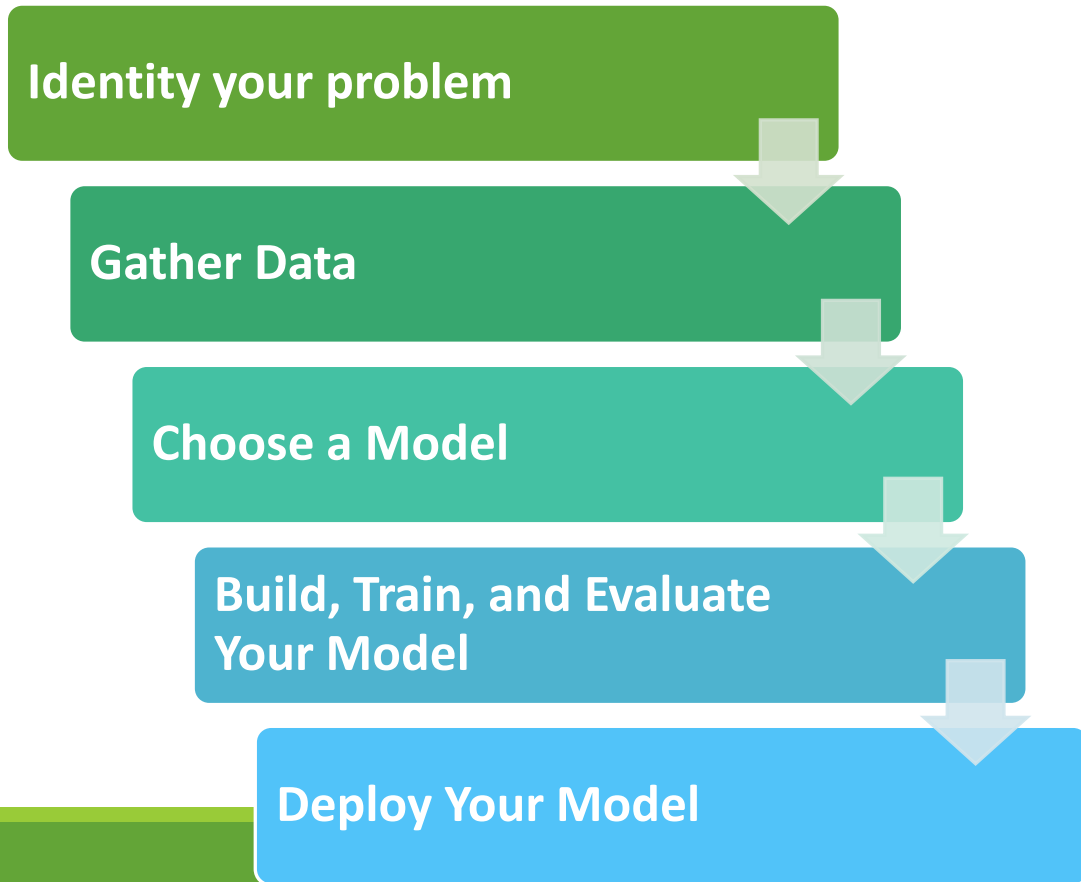
Development Environment For Machine Learning



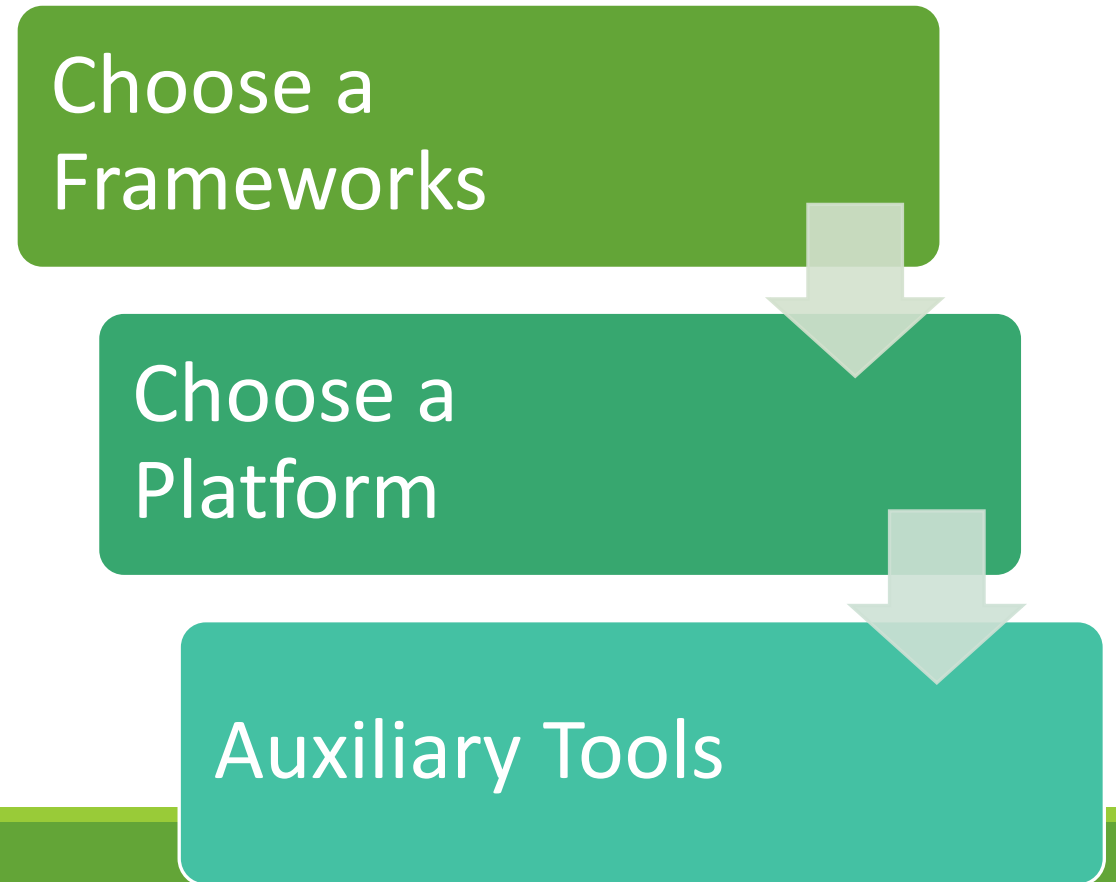
Getting started with machine learning

Two Perspectives

Analysis problem

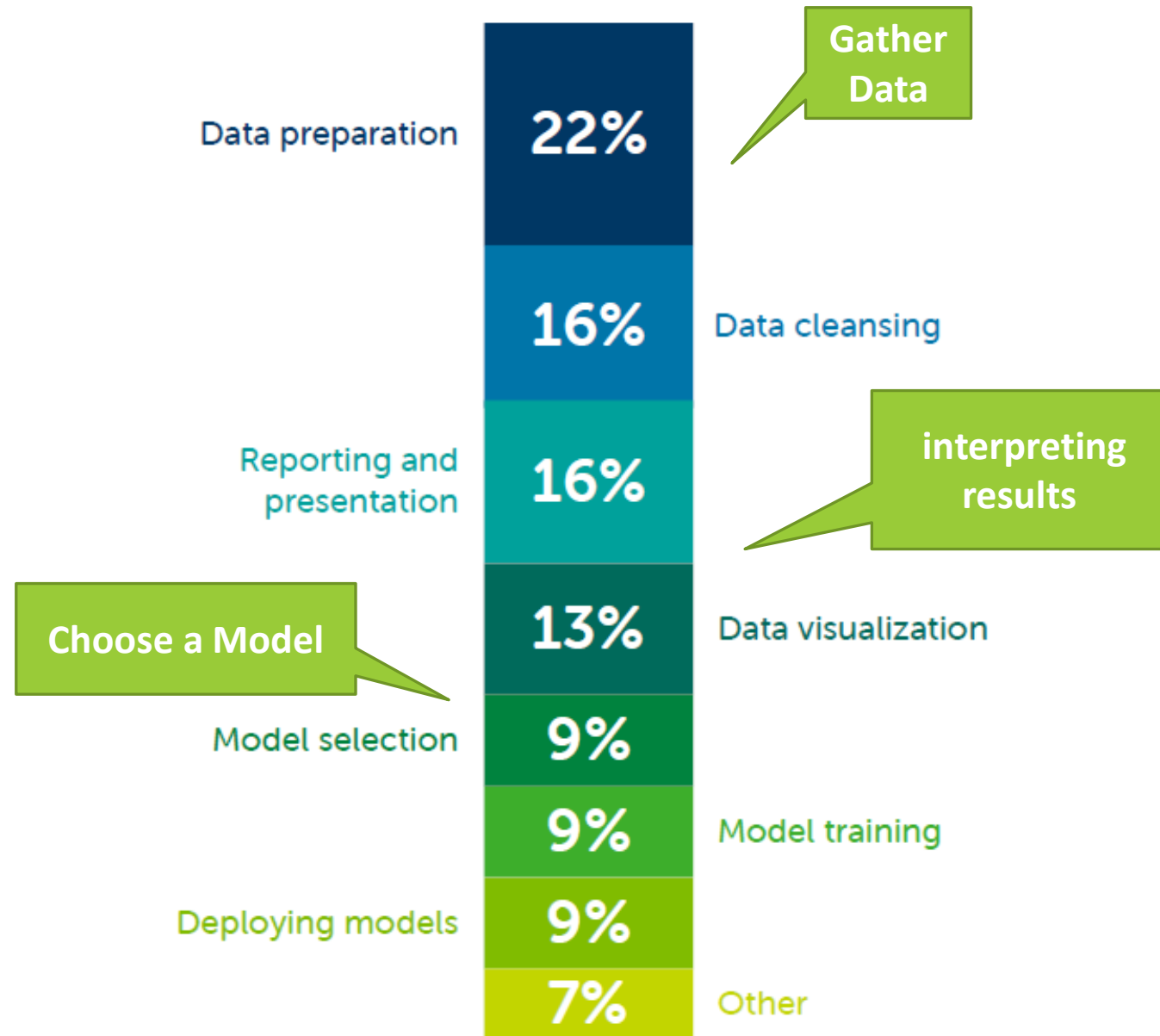


Implementation



How do data scientists spend their time?

- 37.75% of their time on data preparation and cleansing.
- interpreting results 29.19%
- models through selection, training, and deployment takes about 26.44%
- Other 6.62

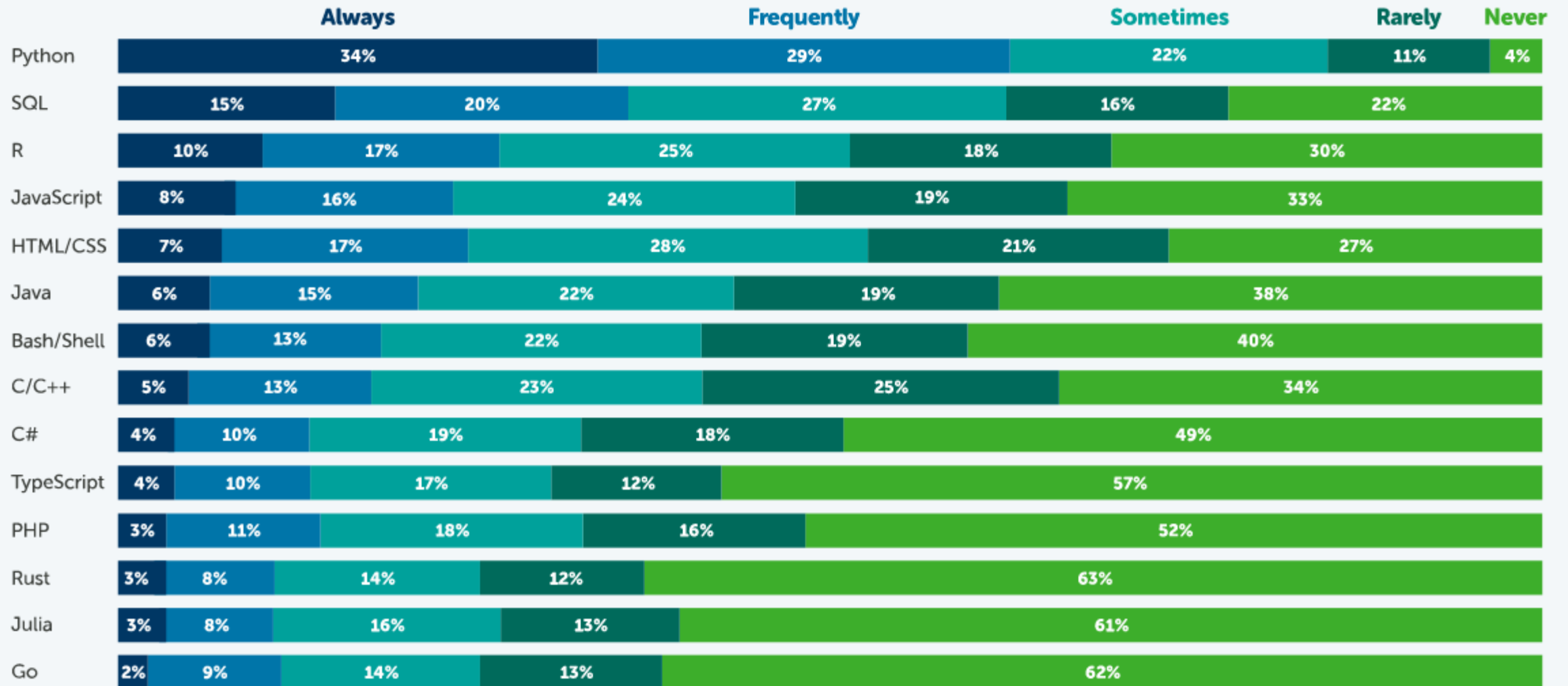


n = 1,966

We asked our respondents how much time they spend on the above tasks, and for each item they entered a number reflecting the percentage of time spent relative to the other options. This is the average of the reported percentages.

Choose a Frameworks

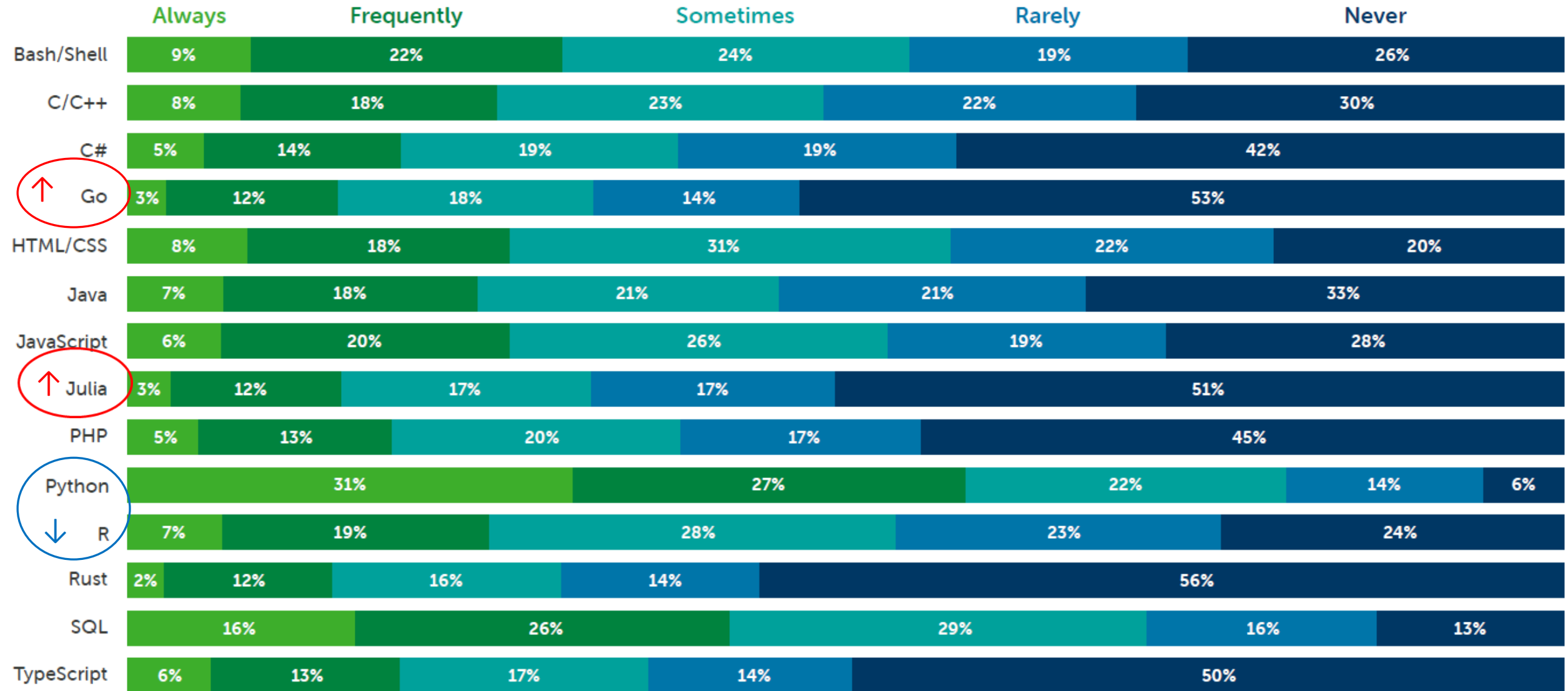
Popular languages in Data Science @ 2021



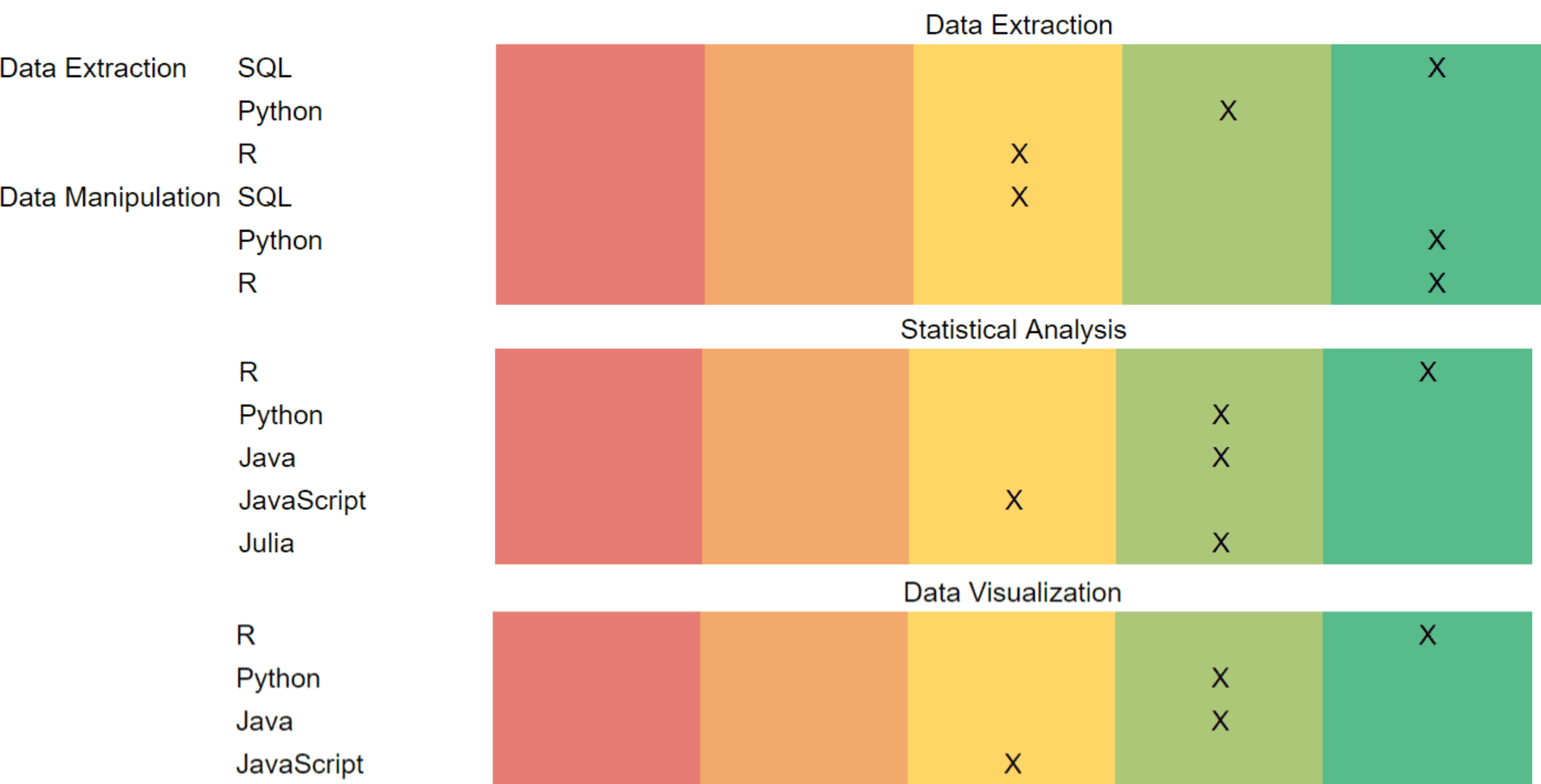
*n=3104

<https://www.anaconda.com/state-of-data-science-2021>

Choose a Frameworks



n = 2,274



Modeling/ML

Python				X
R			X	
Java/JavaScript				X
C/C++			X	
Julia				X
TypeScript				X

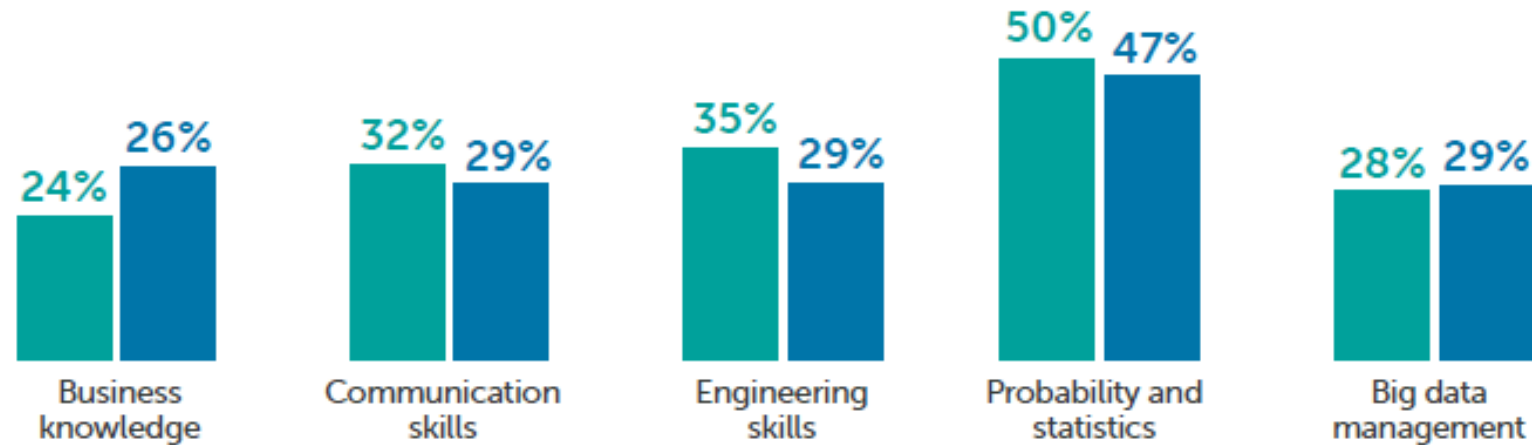
Model Deployment

Python				X
Java				X
JavaScript				X
C#				X
HTML/CSS		X		
PHP			X	
Rust			X	
GoLang			X	

Automation

Python				X
Bash/Shell		X		
Java				X
C#			X	
R		X		

Top five most important skills or areas of expertise?



■ **Educator respondents:**
What topics, tools, or skills is your institution teaching students of data science and machine learning?

■ **Student respondents:**
What topics, tools, or skills are covered in your courses in preparation for entering the data science/ML field?

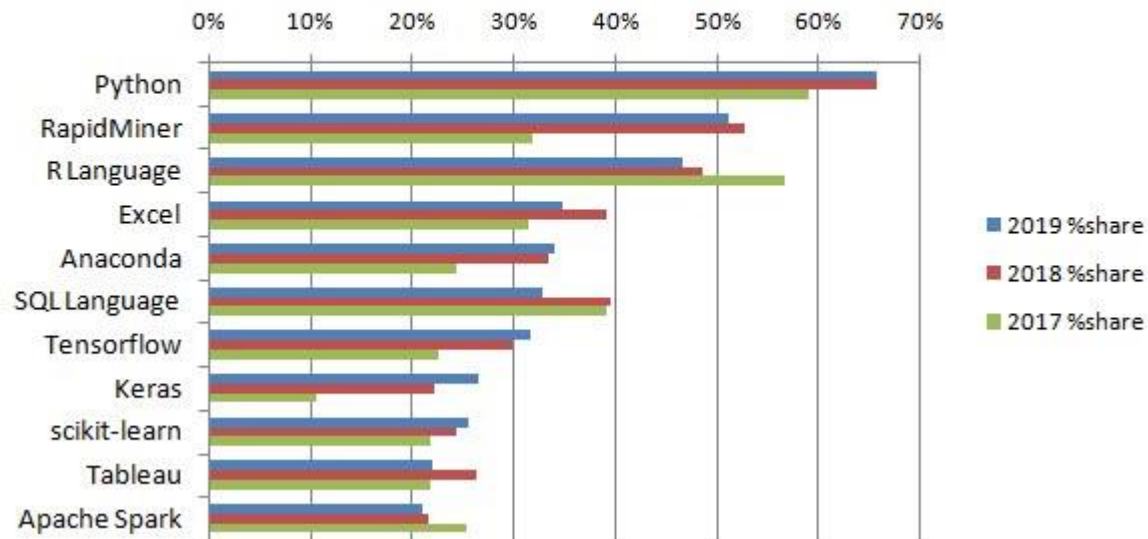
n = 517 and n = 407

Choose a Frameworks

Popular languages in Data Science

Top Analytics/Data Science/ML Software in 2019 KDnuggets Poll

Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll



Software	2019 % share	2018 % share	2017 % share
Python	65.8%	65.6%	59.0%
RapidMiner	51.2%	52.7%	31.9%
R Language	46.6%	48.5%	56.6%
Excel	34.8%	39.1%	31.5%
Anaconda	33.9%	33.4%	24.3%
SQL Language	32.8%	39.6%	39.2%
Tensorflow	31.7%	29.9%	22.7%
Keras	26.6%	22.2%	10.7%
scikit-learn	25.5%	24.4%	21.9%
Tableau	22.1%	26.4%	21.8%
Apache Spark	21.0%	21.5%	25.5%

Top 10 Programming Languages For Data Scientists

2018	2021
Python	C/C++
Java	Julia
R	Java
Julia	JavaScript
SAS	Lisp
SQL	MATLAB
MATLAB	Python
Scala	R
C	SQL
F#	Scala

<https://analyticsindiamag.com/top-10-programming-languages-data-scientists-learn-2018/>

<https://analyticsindiamag.com/top-programming-languages-for-data-scientists-in-2021/>

Advantages and Disadvantages of Python

Advantages

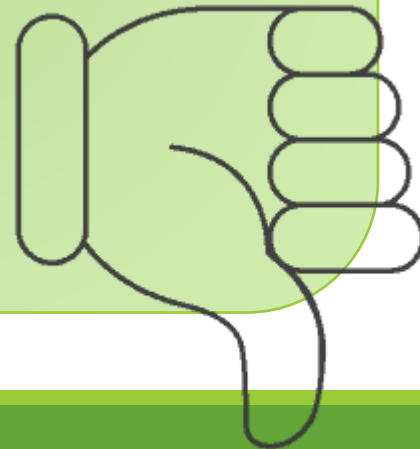
Extensive Support Libraries

Interactive
Object-Oriented, Procedure –
Oriented, Functional
programming
Extensible in C++ & C



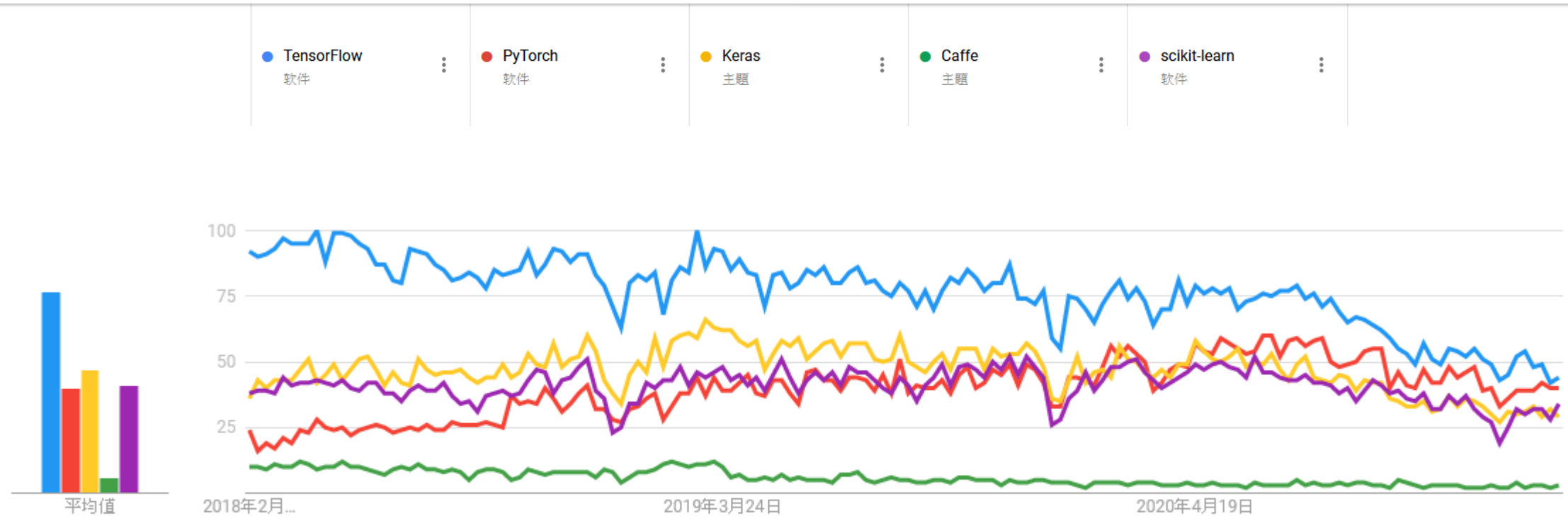
Disadvantages

~~Difficulty in Using Other
Languages~~
~~Gets Slow in Speed~~
Run-time Errors



Choose a Frameworks

Popular Machine Learning Frameworks (python)



[https://trends.google.com/trends/explore?date=2018-02-23%202021-02-](https://trends.google.com/trends/explore?date=2018-02-23%202021-02-23&q=%2Fg%2F11bwp1s2k3,%2Fg%2F11gd3905v1,%2Fg%2F11c1r2rvnp,%2Fg%2F11g6ym8nbt,%2Fm%2F0h97pvq)

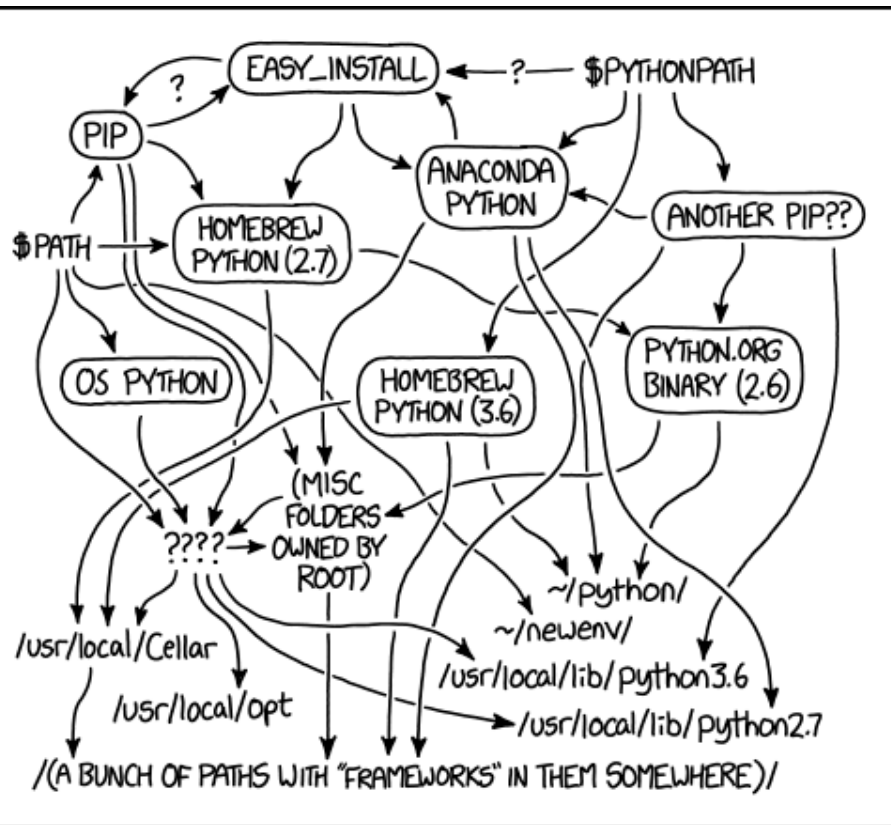
[23&q=%2Fg%2F11bwp1s2k3,%2Fg%2F11gd3905v1,%2Fg%2F11c1r2rvnp,%2Fg%2F11g6ym8nbt,%2Fm%2F0h97pvq](https://trends.google.com/trends/explore?date=2018-02-23%202021-02-23&q=%2Fg%2F11bwp1s2k3,%2Fg%2F11gd3905v1,%2Fg%2F11c1r2rvnp,%2Fg%2F11g6ym8nbt,%2Fm%2F0h97pvq)

Snakes in my computer?



Build your own Development Environment

Use python + pip + virtualenv



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

Windows

1. Download and install python.
2. Install virtualenv via pip
3. Activate virtualenv
4. Install package in virtualenv



Mac & Linux

Don't use system python
Don't use system python
Don't use system python

1. Download and compiler python.
2. Install virtualenv via pip
3. Activate virtualenv
4. Install package in virtualenv

Build your own Development Environment

Use Miniconda/Anaconda

Conda is a package management system.

Miniconda/Anaconda is a distribution for python.

Anaconda is owned by Continuum Analytics™.

Packages in Conda are released by binary not source code. This means that if conda decide not to release certain package you will ...



ANACONDA®

Anaconda Navigator
File Help

ANACONDA NAVIGATOR

Sign in to Anaconda Cloud

Home

Environments

Learning

Community

Documentation

Developer Blog



Applications on scikit-learn

Channels

Refresh



Jupyter

Notebook

5.7.4
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Launch



anypoint

1.1.1

Install



dioplas

0.4.1

Install



Glueviz

0.14.0

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Install



hyperspyui

1.1.0



JupyterLab

0.35.3

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.



Orange 3

3.20.1

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows



pysplot-gui

1.2.3

Small Practice

Create a new environment

1. Install numpy scipy matplotlib
2. Install scikit-learn
3. Install jupyter or spyder

Launch jupyter or spyder

Try to run some code

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
# Compute the x and y coordinates for points on a sine curve
```

```
x = np.arange(0, 3 * np.pi, 0.1)
```

```
y = np.sin(x)
```

```
# Plot the points using matplotlib
```

```
plt.plot(x, y)
```

```
plt.show() # You must call plt.show() to make graphics appear.
```

Choose a Platform



Google Colaboratory

Popular Machine Learning Platform



Amazon
Machine
Learning



IBM Cloud



Google Cloud

Amazon ML

Microsoft Machine
Learning Studio

IBM Cloud

CloudLinux

<https://cloud.google.com/>

<https://aws.amazon.com/tw/machine-learning/>

<https://azure.microsoft.com/en-us/services/machine-learning-studio/>

<https://www.ibm.com/cloud>

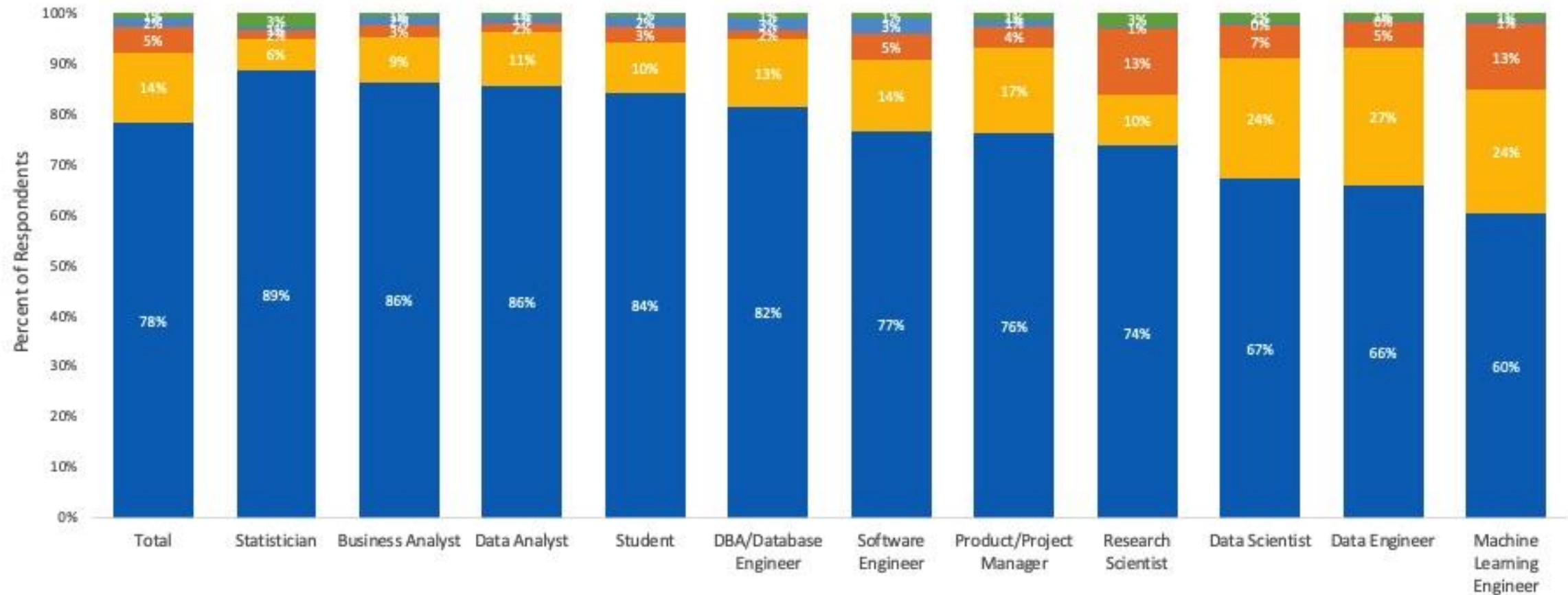
[CloudLinux | For Hosting Providers & Data Centers](#)

<https://analyticsindiamag.com/5-alternatives-to-google-colab-for-data-scientists/>

<https://towardsdatascience.com/5-top-cloud-computing-platforms-with-certification-programs-956c48991738>

What type of computing platform do you use most often for your data science projects?

■ A personal computer or laptop ■ A cloud computing platform (AWS, Azure, GCP, hosted notebooks, etc) ■ A deep learning workstation (NVIDIA GTX, LambdaLabs, etc) ■ None ■ Other



Note: Data are from the 2020 Kaggle ML and Data Science Survey. You can learn more about the study here: <https://www.kaggle.com/c/kaggle-survey-2020/overview>; A total of 17029 respondents answered this question.

Auxiliary Tools-VCS

What is Version Control System & Why do you need it?

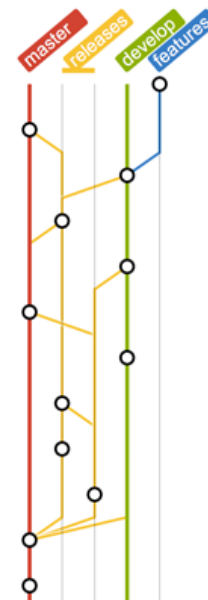
Folder / Filename style

VCS : git, svn

EVERY DESIGNER IN THIS WORLD



Git Flow Chart



FEATURE/JRA-1234	Working on JRA-1234	Remie Bolte	23:02:08 today	49a339b
MASTER	1.2.0 Merge branch 'release/1.2.0'	Remie Bolte	22:38:11 today	a38b6ec
DEVELOP	RELEASE/1.3.0 Merge branch 'release/1.2.0' into develop	Remie Bolte	22:38:11 today	07bf32f
	Merge branch 'master' into release/1.2.0	Remie Bolte	22:36:54 today	4b464c4
	Merge branch 'hotfix/1.1.3' into develop	Remie Bolte	22:36:41 today	8958747
1.1.3	Merge branch 'hotfix/1.1.3'	Remie Bolte	22:36:40 today	aee21f0
	Adding a feature specific file	Remie Bolte	22:34:46 today	3e8f9d1
	Merge branch 'hotfix/1.1.3' into release/1.2.0	Remie Bolte	22:33:55 today	55fb1b2
	Removing obsolete file	Remie Bolte	22:33:39 today	a93a486
	Adding release notes	Remie Bolte	22:33:08 today	b1b1d85
SIGNED_TAG	Copy C.zip as D.zip	Anna Buttfield	2011-02-03	0a943a2
	Add binary file C.zip	Anna Buttfield	2011-02-03	e0a3f6d

Auxiliary Tools-VCS

Git client

	msysGit+ TortoiseGit	SmartGit	SourceTree	Eclipse +egit
windows	⦿	⦿	⦿	⦿
mac		⦿	⦿	⦿
linux		⦿		⦿
GUI	⦿	⦿	⦿	⦿
portable	⦿	⦿		⦿
license	GNU GPL v2	Free for non-commercial purposes	Free for non-commercial purposes	EPL 1.0

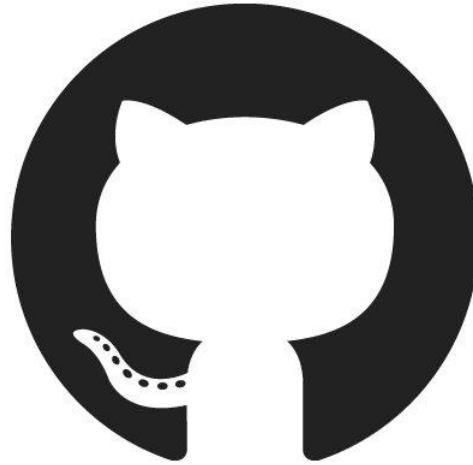
Auxiliary Tools-VCS

Popular Software repository



gitlab

<https://about.gitlab.com/>



github

<https://github.com/>



bitbucket

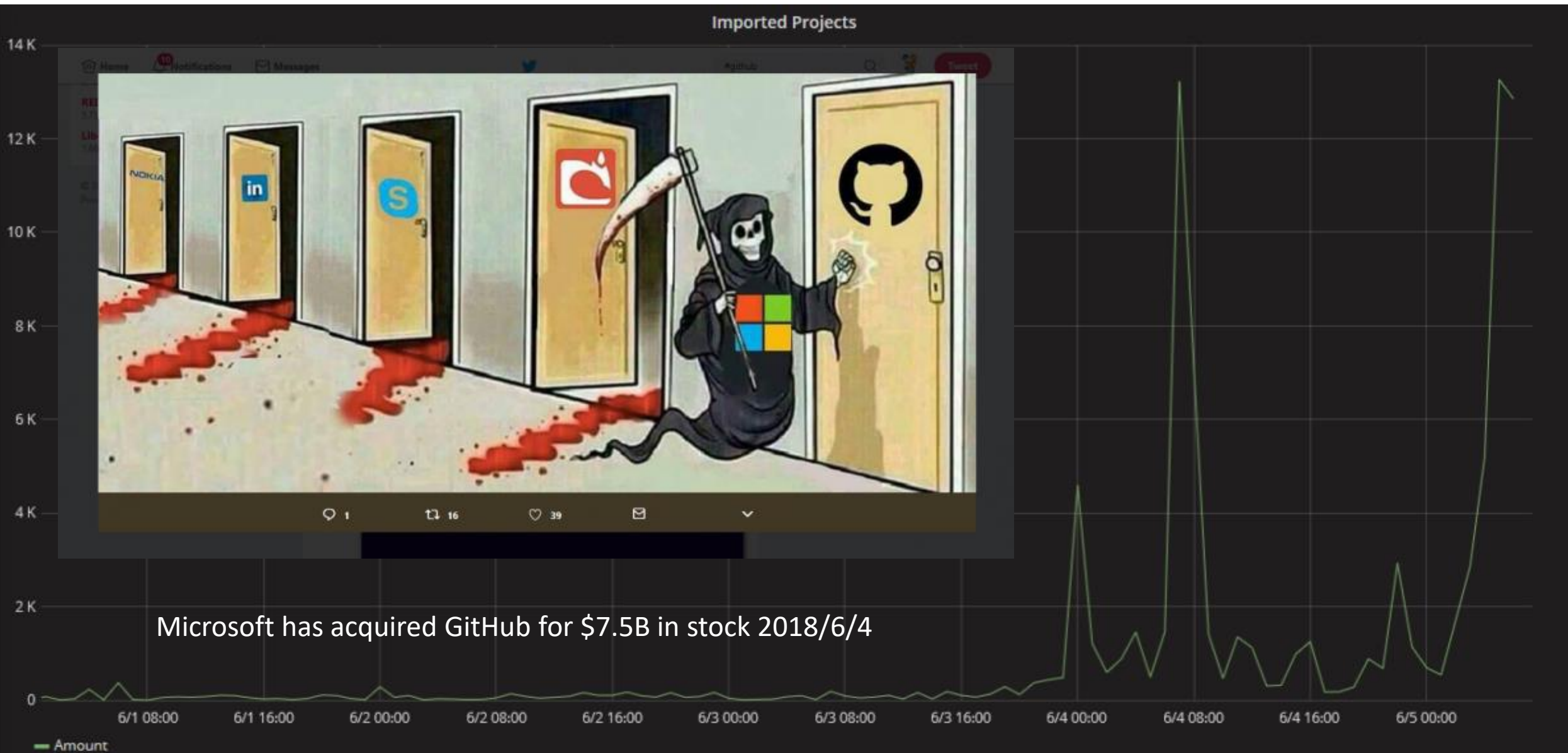
<https://bitbucket.org/>

Auxiliary Tools-VCS

Popular Software repository

gitlab	github	bitbucket
GitLab 是一個利用 Ruby on Rails 開發的開源應用程式，實現一個自託管的 Git 項目倉庫，可通過 Web 介面進行訪問公開的或者私人項目。	GitHub 是第一個供「用Git進行版本控制系統的軟體開發項目」使用的基於Web的代碼託管服務，是目前全球最大的開源社交編程及代碼託管網站。	BitBucket 是 2008 年創建的原始碼託管網站，採用 Mercurial 和 Git 作為分布式版本控制系統，同時提供免費帳戶和商業計劃。2010 年被 Atlassian 收購。
Git	Git，SVN，HG，TFS	Git，HG，SVN
cloud-hosted plan 允許無限數量的用戶在無限數量的公共和私有項目上進行協作，並且每個存儲庫有 10GB 的空間限制	對磁碟使用沒有限制。但是，項目不能超過 1 GB和單個文件不能超過 100 MB。	Small teams plan 允許 5 個成員加入，公有/私有倉庫均免費。

Imported Projects of gitlab



ChatGPT



Yah, it's me again!!

由OpenAI開發的一個人工智慧聊天機器人程式。

基於GPT-3.5架構

在英語中效果最好，也能使用部分其他語言，但效果不一。但還沒有相關paper。

目前官方訓練資料集僅有2021前的資料。

2019年7月22日微軟投資OpenAI 10億美元，雙方攜手合作替Azure雲端平台服務開發人工智慧技術。

2020年6月11日宣布了GPT-3語言模型，微軟於2020年9月22日取得獨家授權。

讓火爆全網的ChatGPT寫學測題目他能拿幾分呢？

<https://www.youtube.com/watch?v=2N4KouAIJDc>

請用python產生繪製test.csv當中第二欄的圓餅圖程式碼

這是一個繪製圓餅圖的範例程式碼：

```
python
import matplotlib.pyplot as plt
import pandas as pd

# 讀取 csv 檔案
df = pd.read_csv('test.csv')

# 繪製圓餅圖
plt.pie(df.iloc[:, 1], labels=df.iloc[:, 0], autopct='%1.1f%%')
plt.axis('equal')

# 顯示圖片
plt.show()
```

註：這個程式碼假設 `test.csv` 的第一欄是每個項目的標籤，第二欄是對應的數值。

做 AI 的喊捉 AI

GPT-2 Output Detector

<https://openai-openai-detector.hf.space/>

ai-text-classifier

<https://platform.openai.com/ai-text-classifier>

Text

深夜裡，月亮照耀著我心，
我獨自想起你溫柔的微笑。
我想念你，如此之深，
你是我生命中最美的一部份。

我愛你，如同大海愛太陽，
你是我生命中唯一的光明。
我會永遠守護你，直到地老天荒，
你是我心中永恆的情人。

夜色漸漸深入，星星閃爍著，
我在這裡祈禱，願我們永遠相隨。

By submitting content, you agree to our [Terms of Use](#) and [Privacy Policy](#). Be sure you have appropriate rights to the content before using the AI Text Classifier.

Submit

Clear

The classifier considers the text to be **likely** AI-generated.

LaMDA (Language Model for Dialogue Applications) vs. Apprentice Bard

2M1207 b



2023/2/8在巴黎的展示翻車，導致google股價大跌超7%，市值蒸發1000多億美元（約3兆新台幣）。

問題：

「James Webb 太空望遠鏡有什麼新發現？」

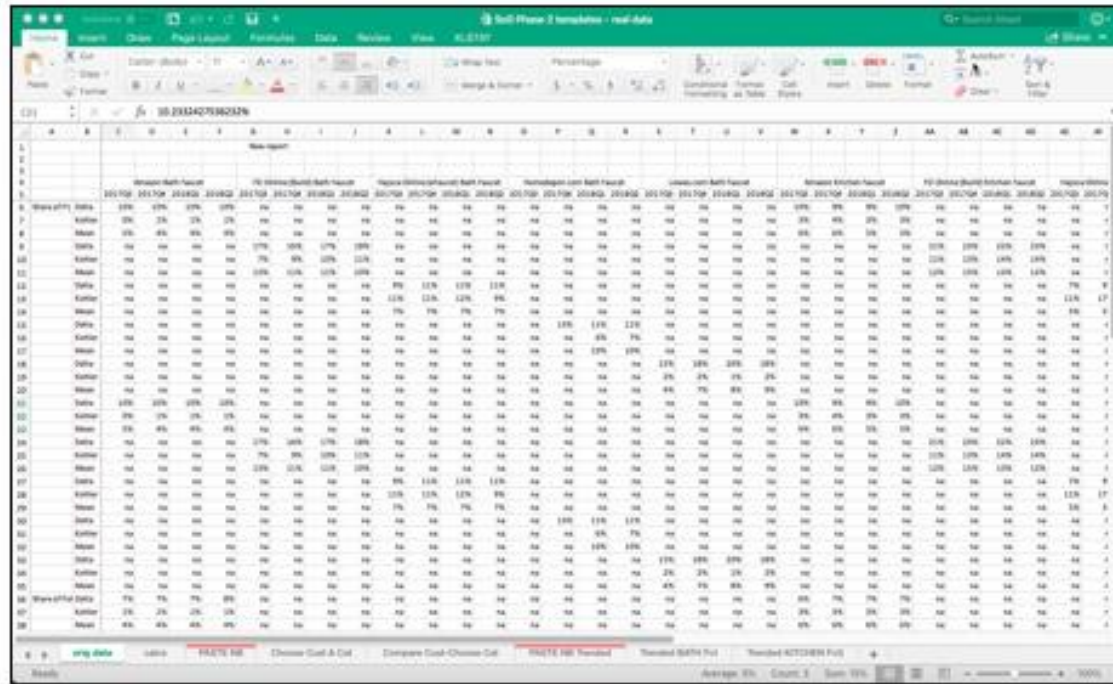
Bard給出其中一個答案是：「韋伯太空望遠鏡拍攝了第一張太陽系外行星的照片。」

但NASA證實，第一張太陽系以外行星的照片是由歐洲南方天文臺的Very Large Telescope (VLT)甚大望遠鏡在2004年拍攝的。

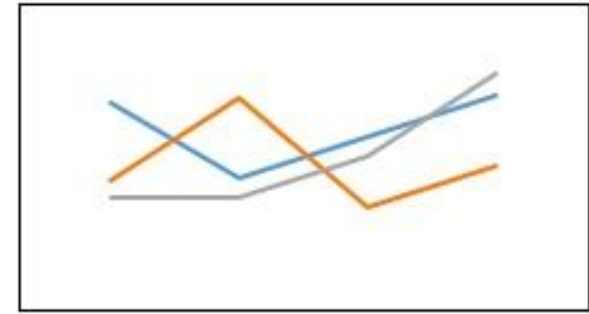
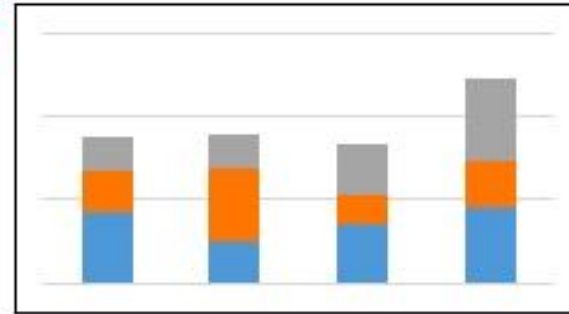
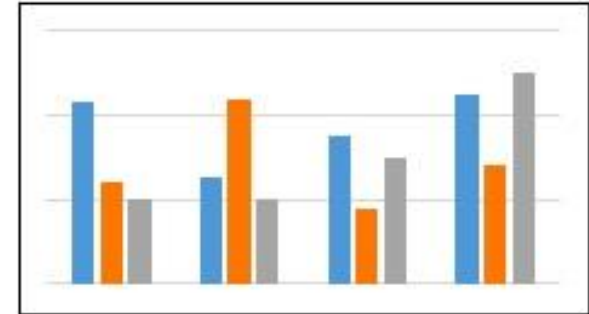
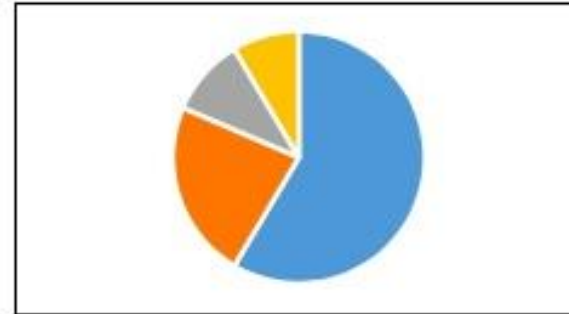
Bard錯誤的回答直接導致Google當日股價大跌超7%，市值蒸發1000多億美元（約3兆新台幣），與此同時也加劇外界對其相關工具尚未準備好整合進搜尋引擎裡的擔憂，就目前來看，Bard似乎難與強大的ChatGPT匹敵。

Auxiliary Tools-visualization

One picture worth ten thousand words --- Data visualization



The screenshot shows an Excel spreadsheet with a green title bar and a ribbon menu. The main area contains a large table with columns labeled 'Sales Report' and rows of numerical data. A red arrow points from the table towards the right, indicating the flow of data into the visualizations.



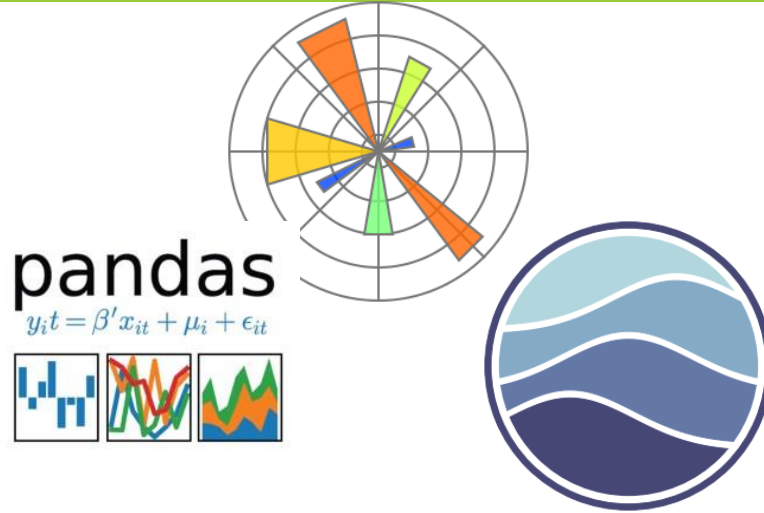
Auxiliary Tools-visualization

Data Analytic and visualization



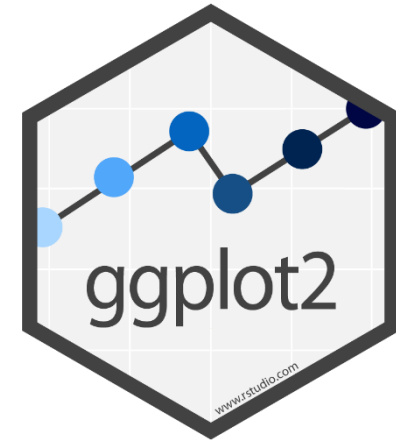
plotly

<https://plot.ly/>



Matplotlib, Pandas, Seaborn

<https://matplotlib.org/>
<https://pandas.pydata.org/>
<https://seaborn.pydata.org/>



ggplot2

<https://ggplot2.tidyverse.org/>

20 Best Data Visualization Software Solutions of 2019

<https://financesonline.com/data-visualization/>

<https://pbpython.com/visualization-tools-1.html>

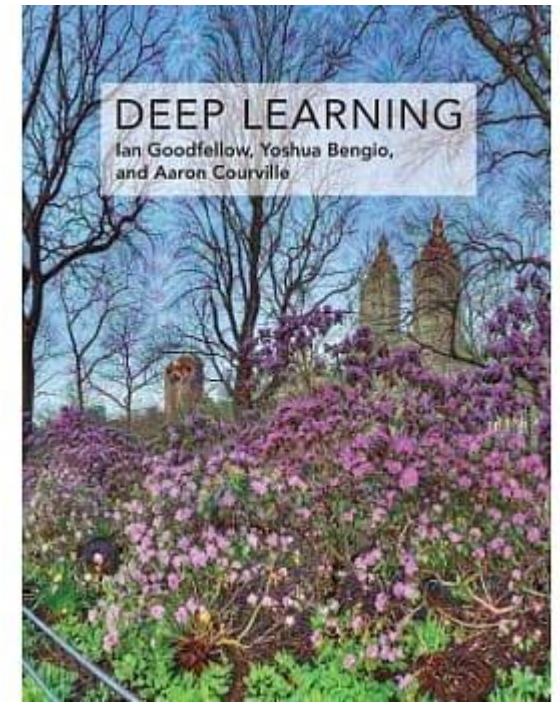
Learning Resources

Book

1. Deep Learning (ISBN : 9780262035613)
2. <https://d2l.ai/>, <https://github.com/d2l-ai>
3. 深度學習 (ISBN : 9789865021924)
4. Python程式設計：從入門到進階應用 (ISBN : 9789865030025)
5. 實戰TensorFlow：Google深度學習系統 (ISBN : 9789864764730)

Online Learning Resources

- [10 Best and Free Machine Learning Courses, Online](#)
- Coursera : <https://zh-tw.coursera.org/learn/machine-learning>
- data-flair : <https://data-flair.training/blogs>
- Tensorflow : <https://www.tensorflow.org/tutorials>
- scikit-learn : <https://scikit-learn.org/stable/tutorial/index.html>
- Keras : <https://keras.io/>



Yann LeCun

YOSHUA BENGIO

GEOFFREY E HINTON

AI不是萬能的



IBM 大名鼎鼎的 Watson 也要
被賣了，人類的 AI 夢該醒了？

1975擊敗西洋棋王的Deep Blue輸入了人類的200多萬局棋譜，使用超強的運算硬體，進行窮舉搜索，並不是「真AI」。

Watson 是 IBM AI 業務的招牌，也是人類最初充滿野心的 AI 夢代表。

2011 年，IBM 的 Watson就在智力競答節目《危險邊緣》，擊敗節目史上最成功的兩位人類冠軍。第二天，IBM 宣布 Watson 的新職業目標：AI 醫生。

Watson先理解自然語言（患者的電子病歷），然後檢索資料庫（治療方案和最新醫學文獻），最終得出答案。

AI不是萬能的

Watson 挑戰的是診斷，且還是醫學難度最大的腫瘤治療領域，Watson Health 面臨資料和 AI 智慧的雙重挑戰。



資料層面，大部分醫療資料是非結構化資訊，要將這些非結構化資料整理為機器的訓練資料就花費難以想像的人力物力。(資料準備清洗很花成本)

圖靈獎得主Yoshua Bengio曾悲觀表示，AI 的自然語言理解能力進步飛快，但比人類依然差很多，AI 無法理解醫學文本歧義，也無法找到人類醫生會注意到的細微線索。(從資料海萃取特徵不是容易的事)

本來應該餵給 Watson 大量真實數據找到新治療方法，但罕見病例本就缺乏，Watson 被灌入一堆沒什麼用的假設數據，並不是真正的病人數據。訓練用真實病例數量很少，最多的肺癌也僅 635 例，最少的卵巢癌更只 106 例。(資料擴增要有節制地被使用)

AI 的本質是統計學，得出的推論局限於提供的數據，無法像專業醫生，獨立生成新的見解。

Watson 只能比人類專家更快得出相同結果，無法治療人類醫生治不了的病。

人工智慧為應用科學帶來的改變

越來越多領域的研究人員，把研究上所遇到的難題改用機器學習技術解決。

- 但機器學習和資料驅動的模型要為人類在科學探索上使用還有很大一段路要走。

機器學習就像個黑盒子，知其然不知其所以然。

- 這點出了機器學習的一項缺點，若不知其所以然，又要如何從學理出發進行模型的改良或精進？

機器學習就像微積分等數學工具一樣，是人類發展用來探索萬物道理的一種工具，現在還在還在這個方法發展的初期，但即使如此機器學習也已經在許多研究難題上展現一條新的道路。

現在正是需要有更多人力投入發展之時，你準備好了嗎？

<https://playground.tensorflow.org>

Play with neural networks!

Tinker With a **Neural Network** Right Here in Your Browser.
Don't Worry, You Can't Break It. We Promise.



Epoch
001,808

Learning rate

0.03

Activation

ReLU

Regularization

None

Regularization rate

0

Problem type

Classification

DATA

Which dataset do you want to use?

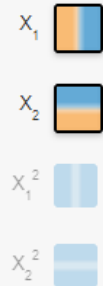


Ratio of training to test data: 50%

Noise: 0

FEATURES

Which properties do you want to feed in?



+ - 2 HIDDEN LAYERS

+ -

4 neurons

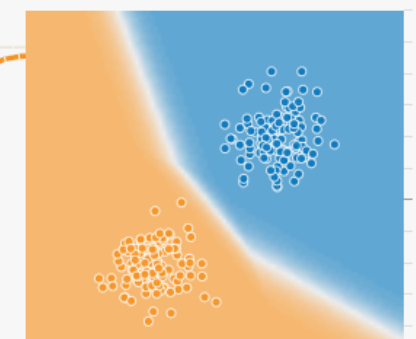
+ -

2 neurons

The outputs are mixed with varying weights, shown by the thickness of the lines.

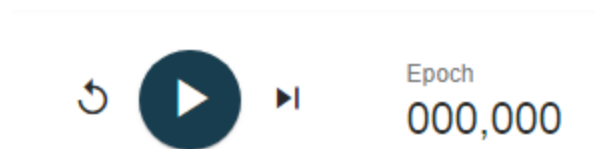
OUTPUT

Test loss 0.000
Training loss 0.000



訓練控制區塊

上面工具列的左邊，是執行步數與顯示，你可以重置訓練，持續訓練，單次訓練，查看已執行的周期數。



訓練參數設定

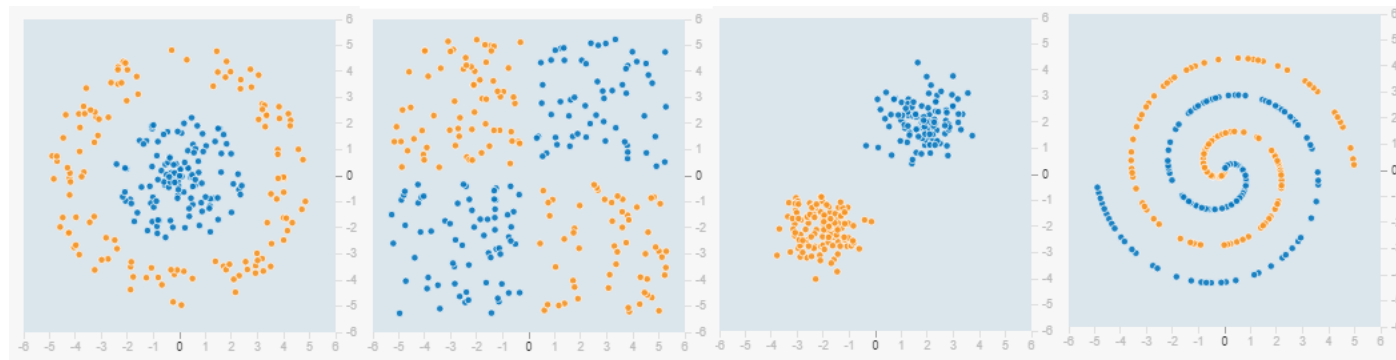
上面工具列的右邊，是訓練的設定，一開始比較常設的就是學習率，最後一個問題型態會改變下方左邊資料集的選項。有 **Classification** (分類問題資料集) 和 **Regression** (回歸問題資料集) 二種。

Learning rate	Activation	Regularization	Regularization rate	Problem type
0.03	Tanh	None	0	Classification

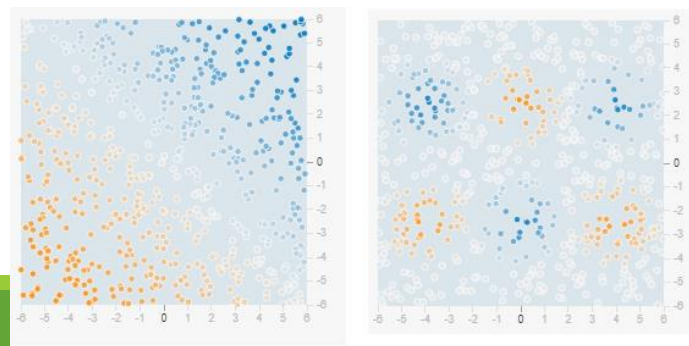
訓練資料集

下方左邊是資料集的設定，

Classification (分類問題資料集)，有4個資料集(Circle / XOR / Gaussian / Spiral)。



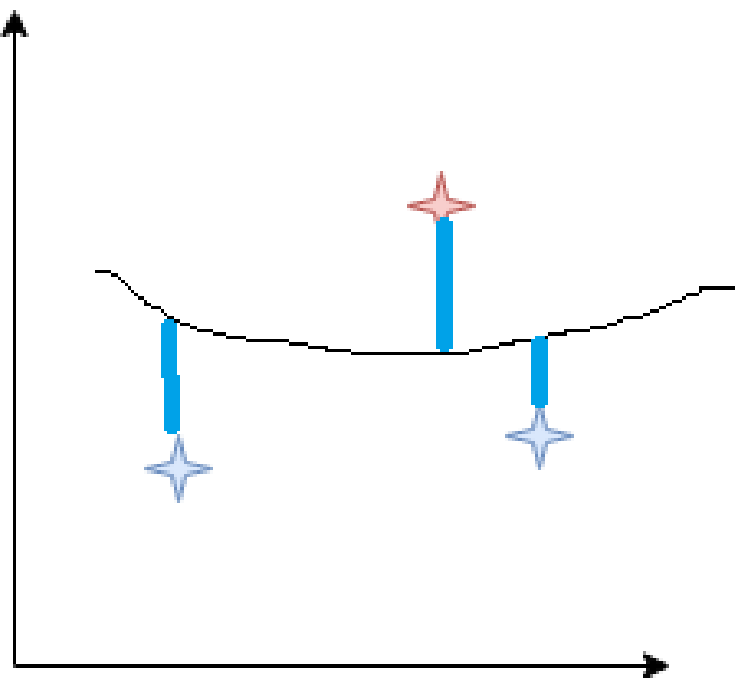
Regression (回歸問題資料集)，有2個資料集(Plane / Multi gaussian)。



Classification v.s. Regression

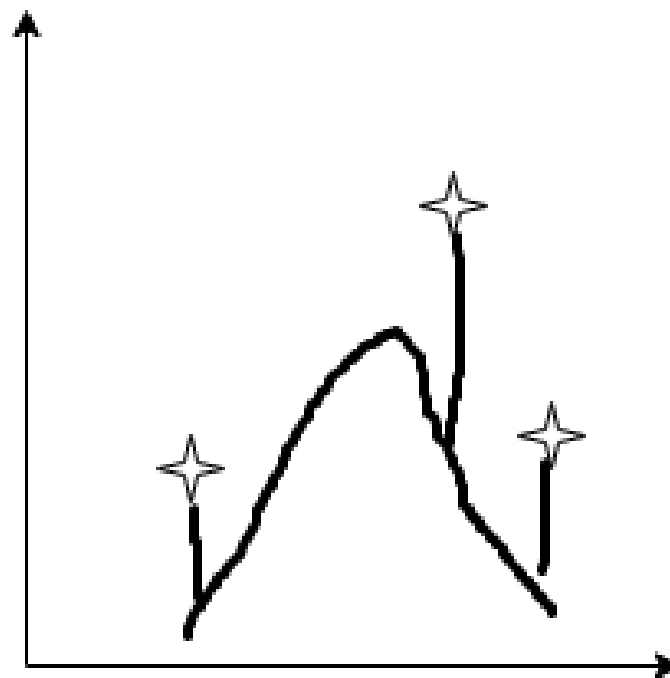
Classification

找出一條線(面)，讓資料點到線(面)的距離總和最大



Regression

找出一條線(面)，讓資料點到線(面)的距離總和最小



取用資料集設定


設定資料集的訓練設定，要拿多少資料出來訓練，S/N ratio，batch的大小





特徵選取


FEATURES


Which properties do you want to feed in?


X_1 


X_2 

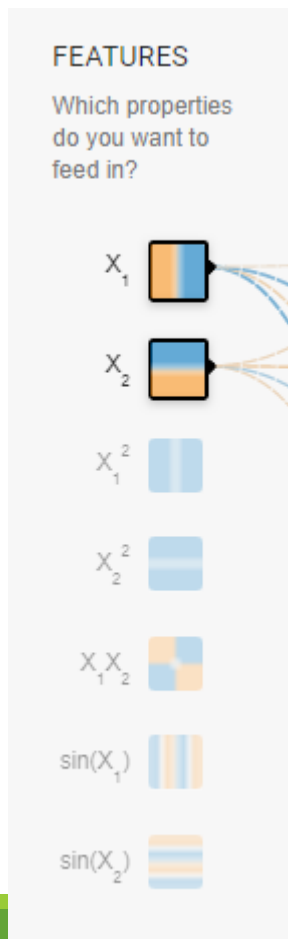
X_1^2 

X_2^2 

$X_1 X_2$ 

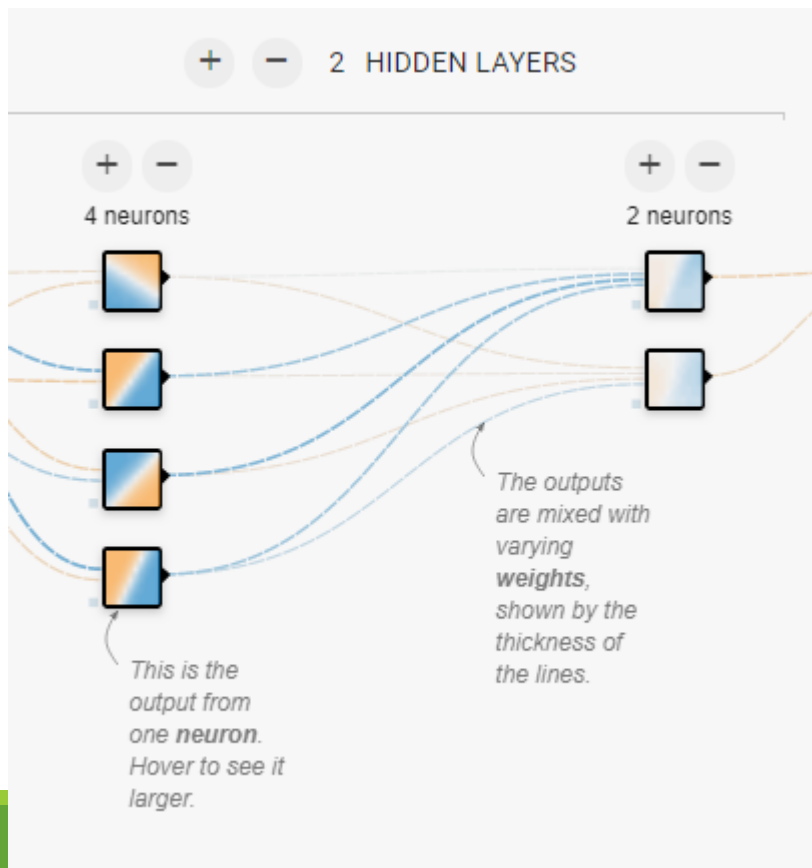
$\sin(X_1)$ 

$\sin(X_2)$ 

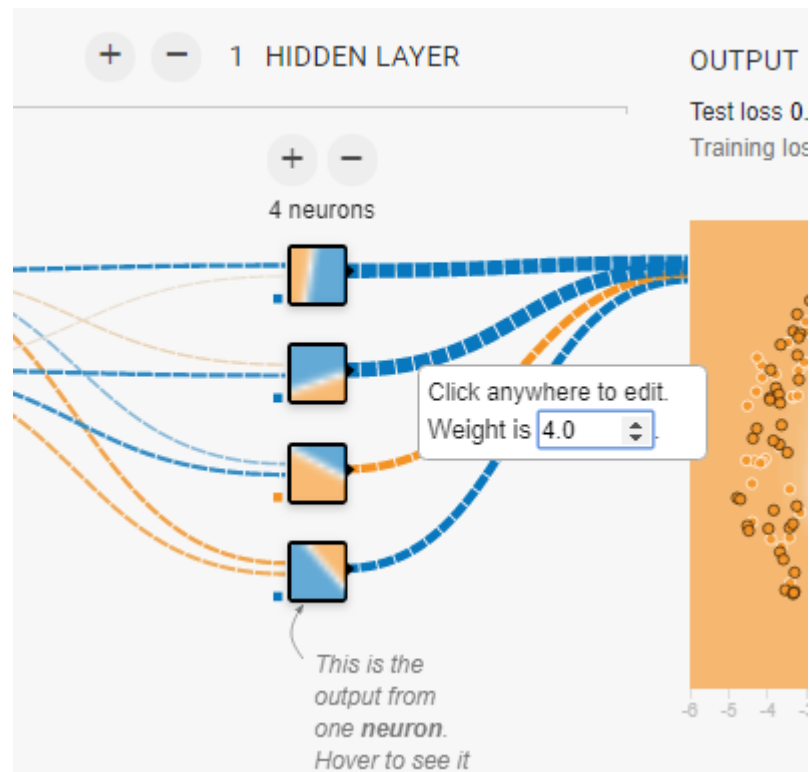


網路設定

可以調整網路的深度和廣度

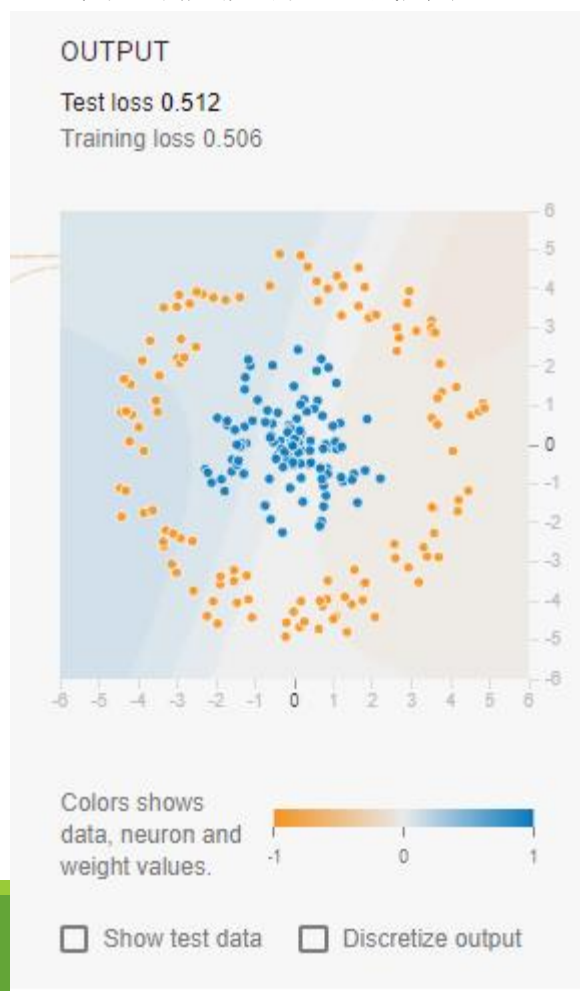


調整個別連結權重



訓練結果

下面右方是輸出結果的顯示，可以看到損失值的情況，也可以用測試集看看模型效果。



Playground Hands on(使用XOR dataset)

Task 1: 使用單一隱藏層單一神經元作為model，這model能表現這個非線性的XOR dataset嗎?

Task 2: 增加隱藏層的神經元數量(2~3)，同時也變更activation function，再看看這model能表現這個非線性的XOR dataset嗎?

Task 3: 任意調整隱藏層數量、神經元數量，及各項訓練設定(learning rates, regularization...)。你能用盡可能少的層數和神經元數量得到小於0.177的test loss嗎?

Task 4: 使用單一隱藏層、3個神經元作為模型， learning rates=0.01 , activation= ReLU, regularization=L2，執行四到五次，每次至少500 steps。(每次開始前記得按"Reset the network button")，觀察不同run之間，model的輸出變化。這帶給你什麼啟示?

Task 5: 增加隱藏層數量、神經元數量，重複Task 4，這對模型穩定性有無幫助?

Playground Exercises(使用noisy spiral dataset)

Task 1: 在只使用 X_1 和 X_2 作為輸入的features的前提下，調整網路層數和神經元數量，你可以訓練出最好的model什麼？

Task 2: 增加輸入的features 像是 $X_1 \cdot X_2$, $\sin(X_1)$, $\sin(X_2)$ 。你覺得增加features和調整網路層數和神經元數量誰對模型的好壞影響較為顯著？