心理與神經資訊學
(Psychoinformatics & Neuroinformatics)

課號: Psy5261

識別碼: 227U9340

教室:彷彿在雲端

時間: 一789

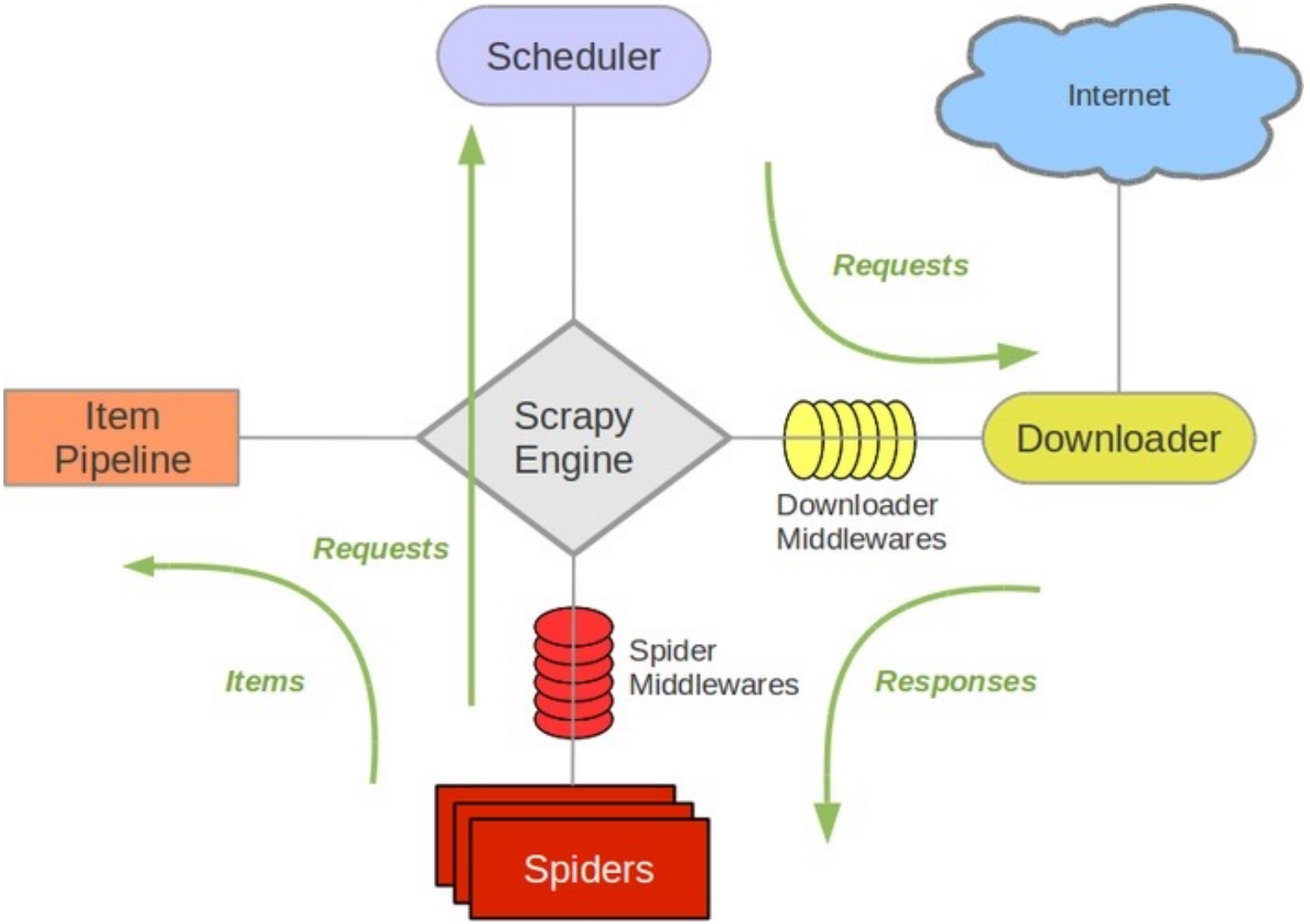

今天來把整個學期串起來
！

總論

(Overview)

觀察法撈大數據

Scrapy使用Twisted的async I/O

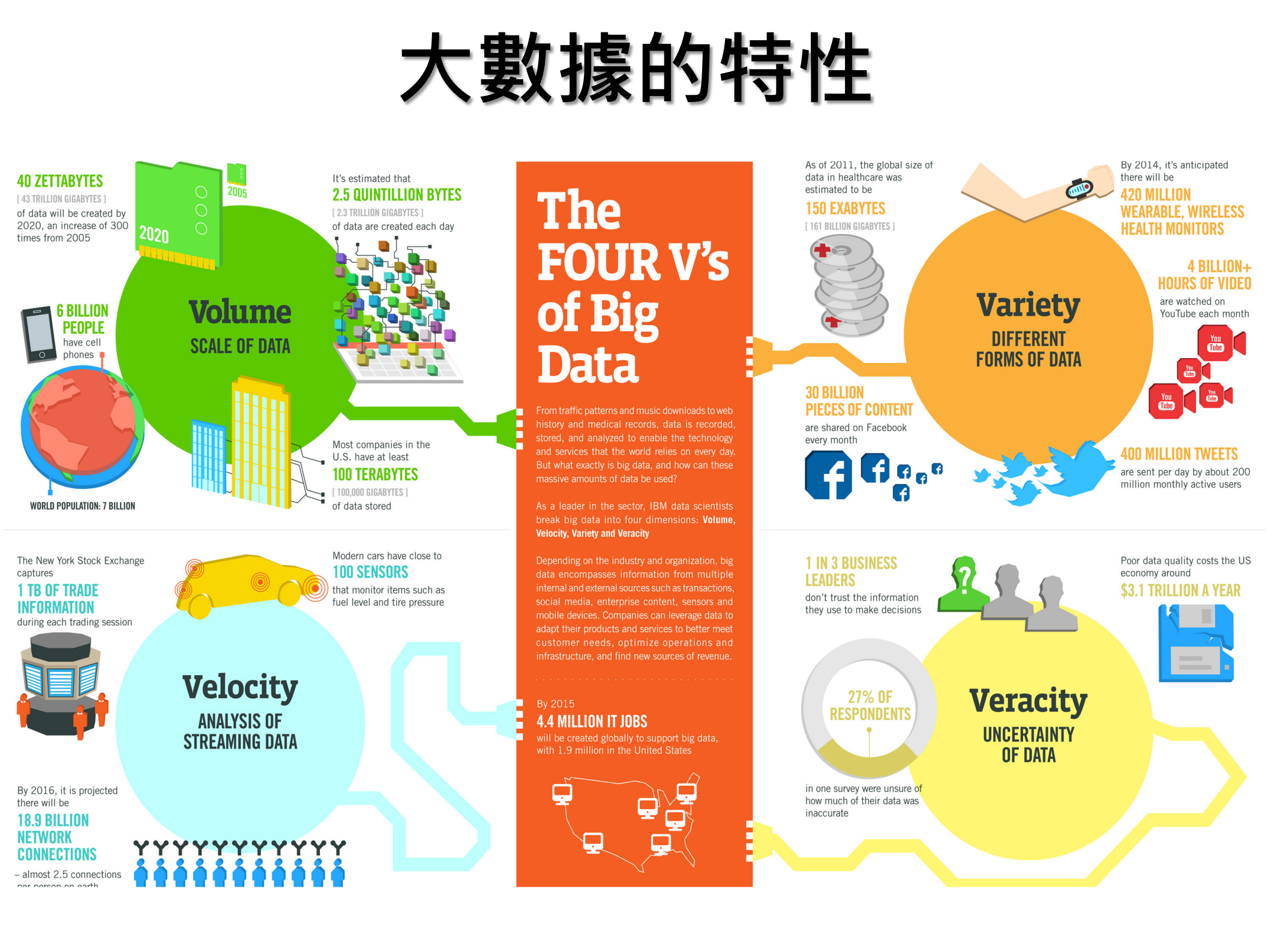


實驗法收大數據

Javascript & Node.js使用async I/O

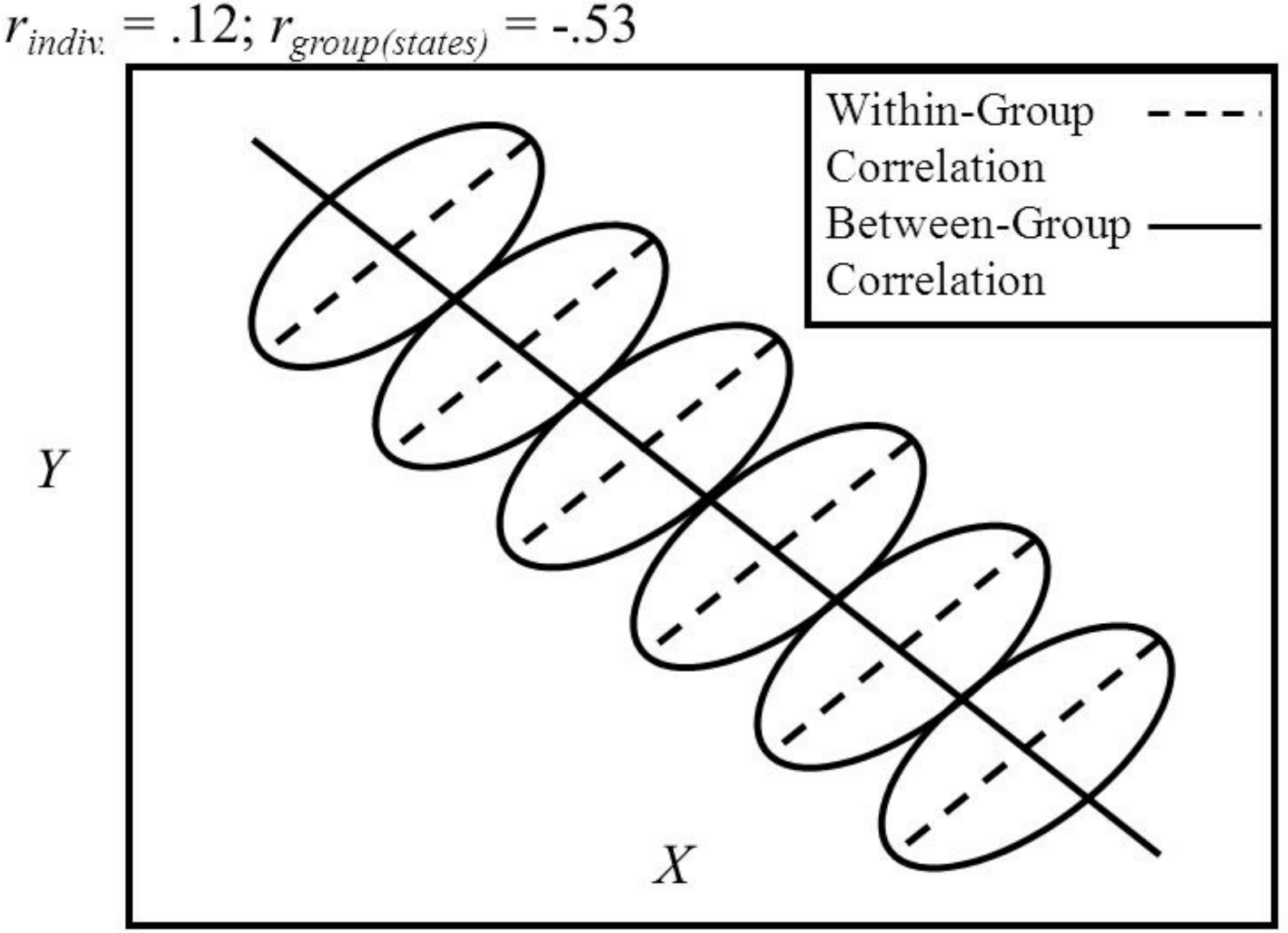
```
function hell(win) {
 // for listener purpose
 return function() {
   loadLink(win, REMOTE_SRC+'/assets/css/style.css', function() {
     loadLink(win, REMOTE_SRC+'/lib/async.js', function() {
       loadLink(win, REMOTE_SRC+'/lib/easyXDM.js', function() {
         loadLink(win, REMOTE_SRC+'/lib/json2.js', function() {
           loadLink(win, REMOTE_SRC+'/lib/underscode.min.js', function() {
             loadLink(win, REMOTE_SRC+'/lib/backbone.min.js', function() {
               loadLink(win, REMOTE_SRC+'/dev/base_dev.js', function() {
                 loadLink(win, REMOTE_SRC+'/assets/js/deps.js', function() {
                   loadLink(win, REMOTE_SRC+'/src/' + win.loader_path + '/loader.js', function() {
                     async.eachSeries(SCRIPTS, function(src, callback) {
                       loadScript(win, BASE_URL+src, callback);
                     });
                   });
                 });
               });
             });
           });
         });
       });
     });
   });
 };
}
```

大數據的特性


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The Four V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
Variety
DIFFERENT FORMS OF DATA
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

大數據幫助看到全貌

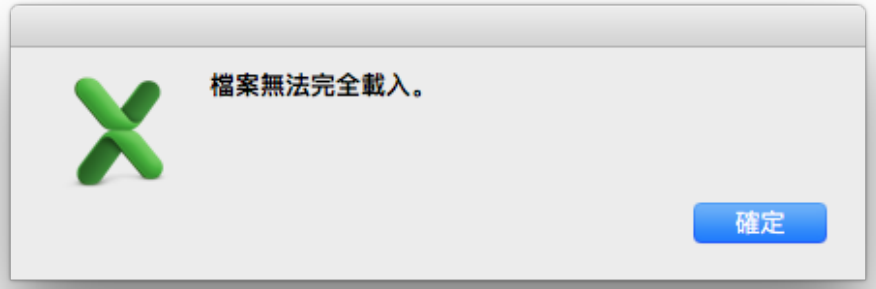
例如局部正相關但整體負相關



多大叫做大?

試看看16_Materials.zip

network.txt (5.0 GB): 284,884,514列
network2.txt (167 MB): 10,000,000列
network3.txt (32 MB): 2,000,000列
network4.txt (129 KB): 10,000列



Excel最多只能讀2^{20}=1048,576行

序列計算vs.平行計算

調 調 調 調 調 調 調 調 調 調

0 10 20 30 40 50 60 70 80 90 100

(只有1條生產線)

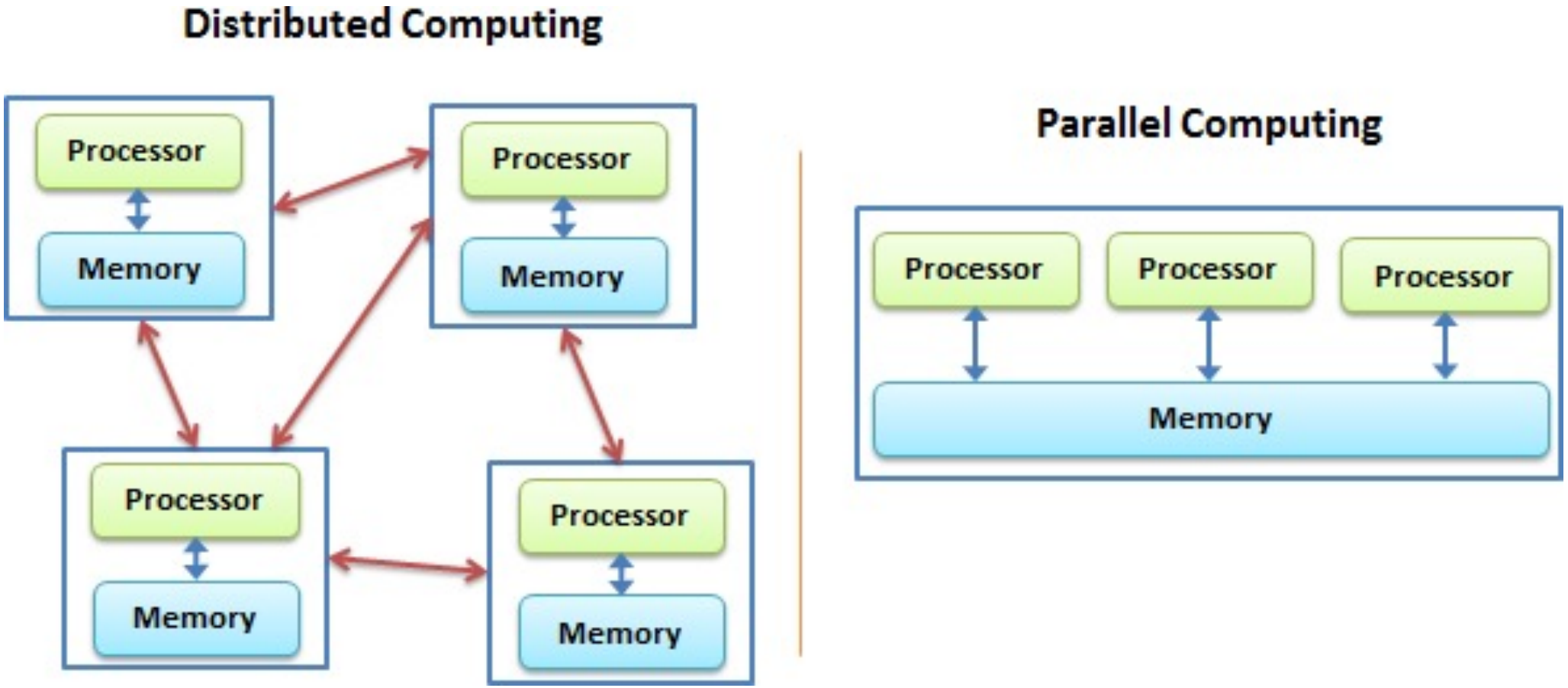
調 調 調 調 調 調 調 調 調 調 調

0 10 20 30 40 50 60 70 80 90 100

(共有11條生產線)

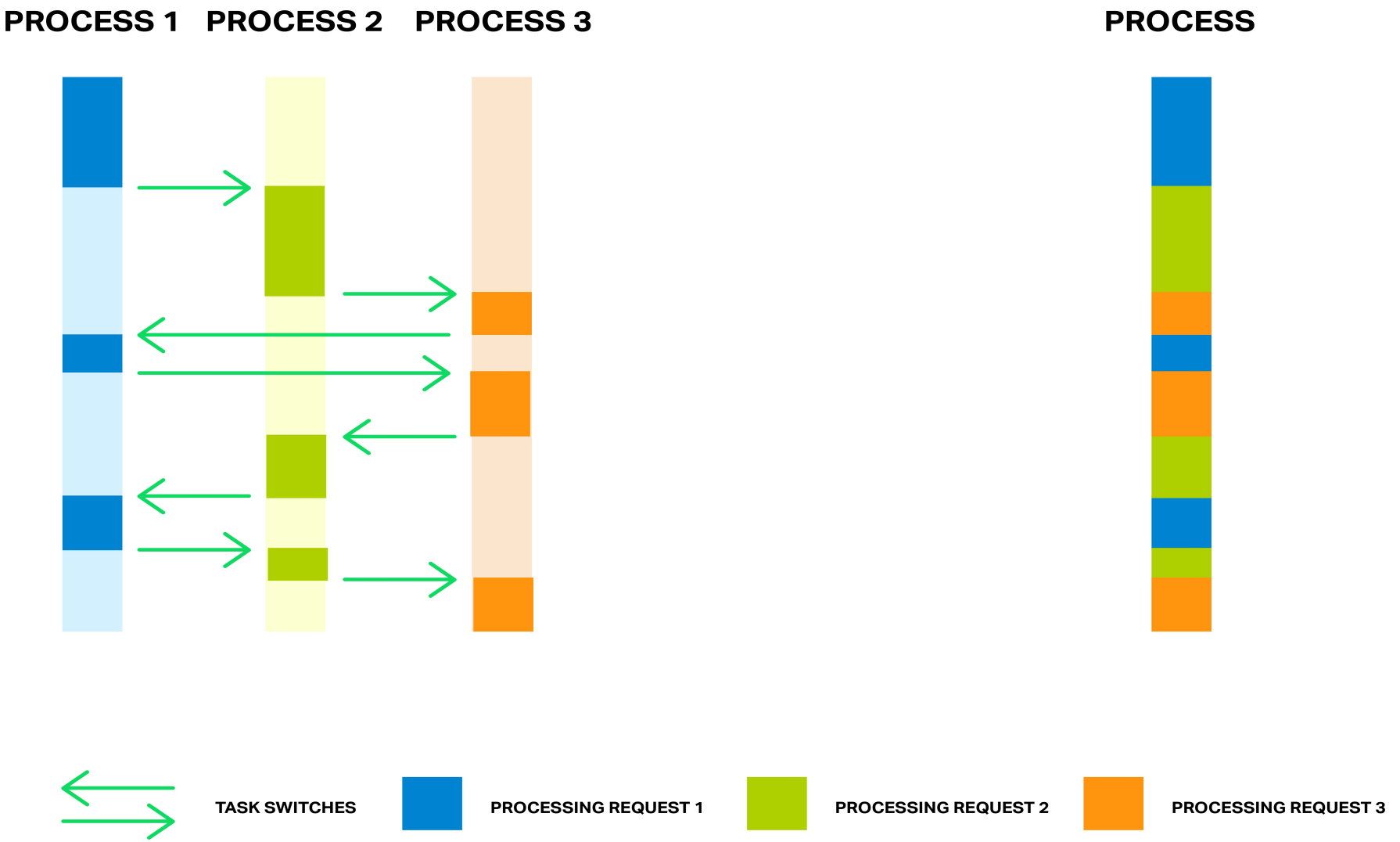
平行計算vs.分散式計算

前者用一台後者用多台電腦(如：邊緣運算&聯邦學習)



平行/分散計算vs.異步執行

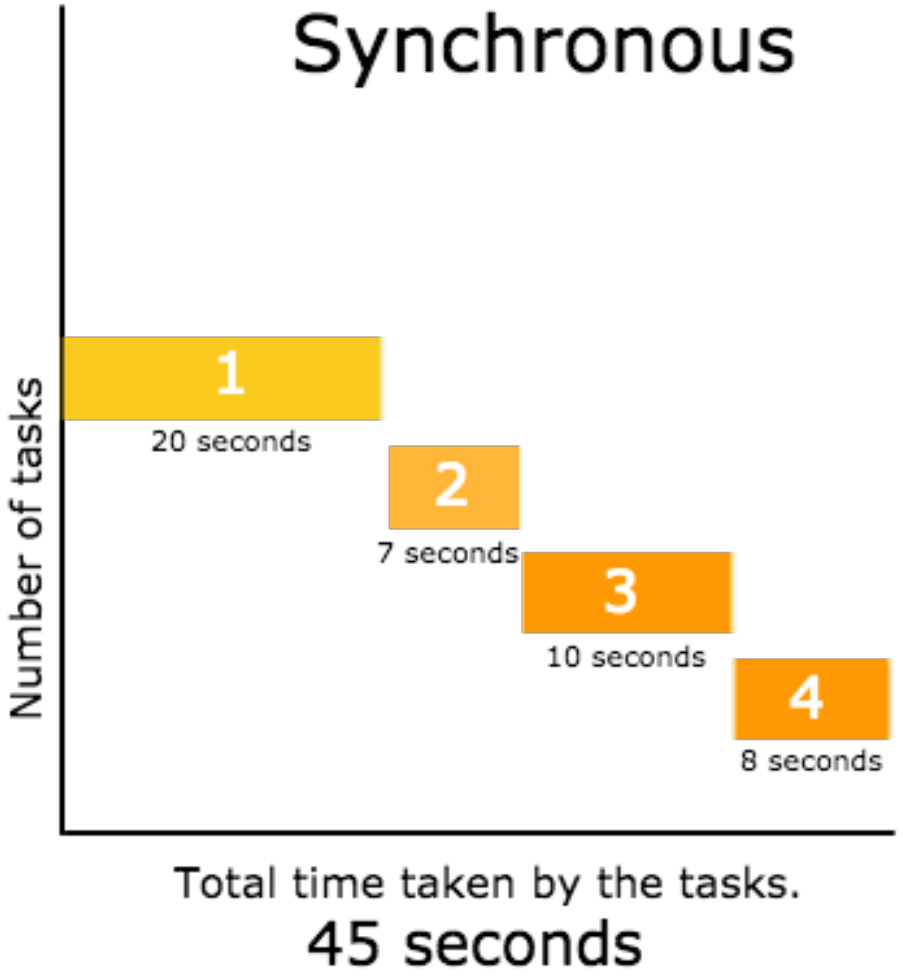
平行/分散計算執行效率有可能輸單程序異步執行

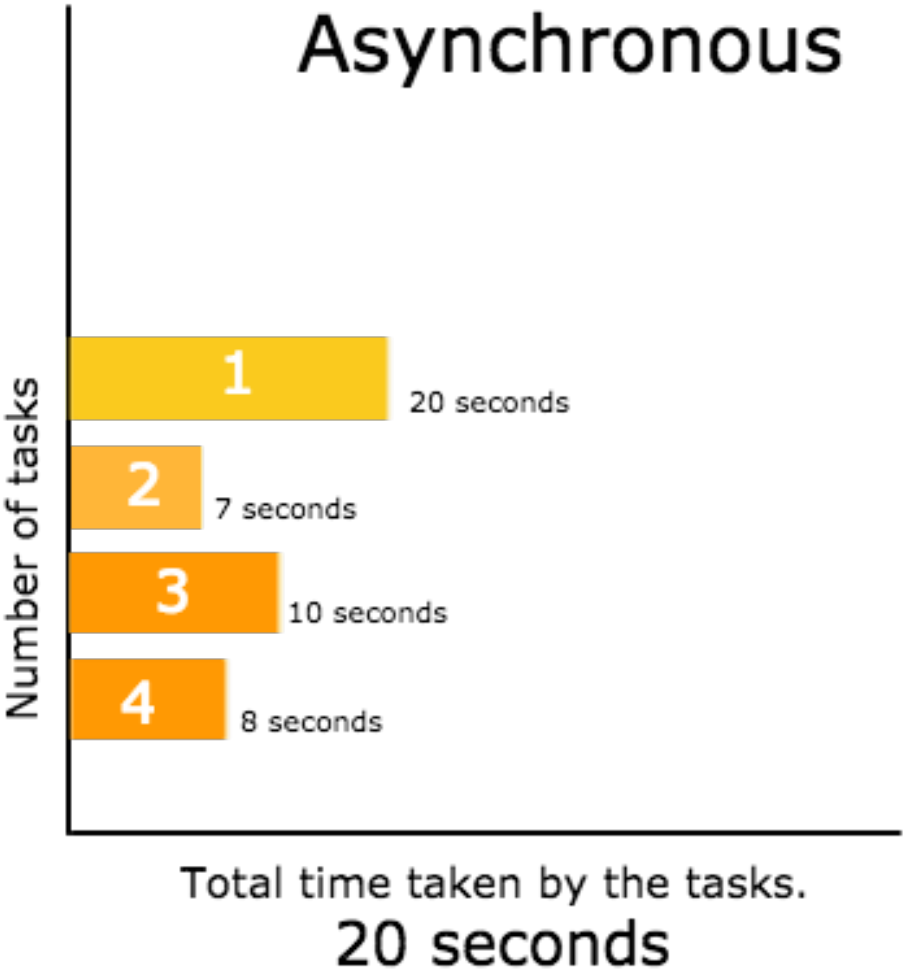


異步執行
(Asynchronous Execution)

同步執行vs.異步執行(1/2)

多,慢, 且不相依I/O(如網頁請求)的情況適合異步執行





同步執行vs.異步執行(2/2)

Python中async I/O可使用asyncio套件

check if done

switch

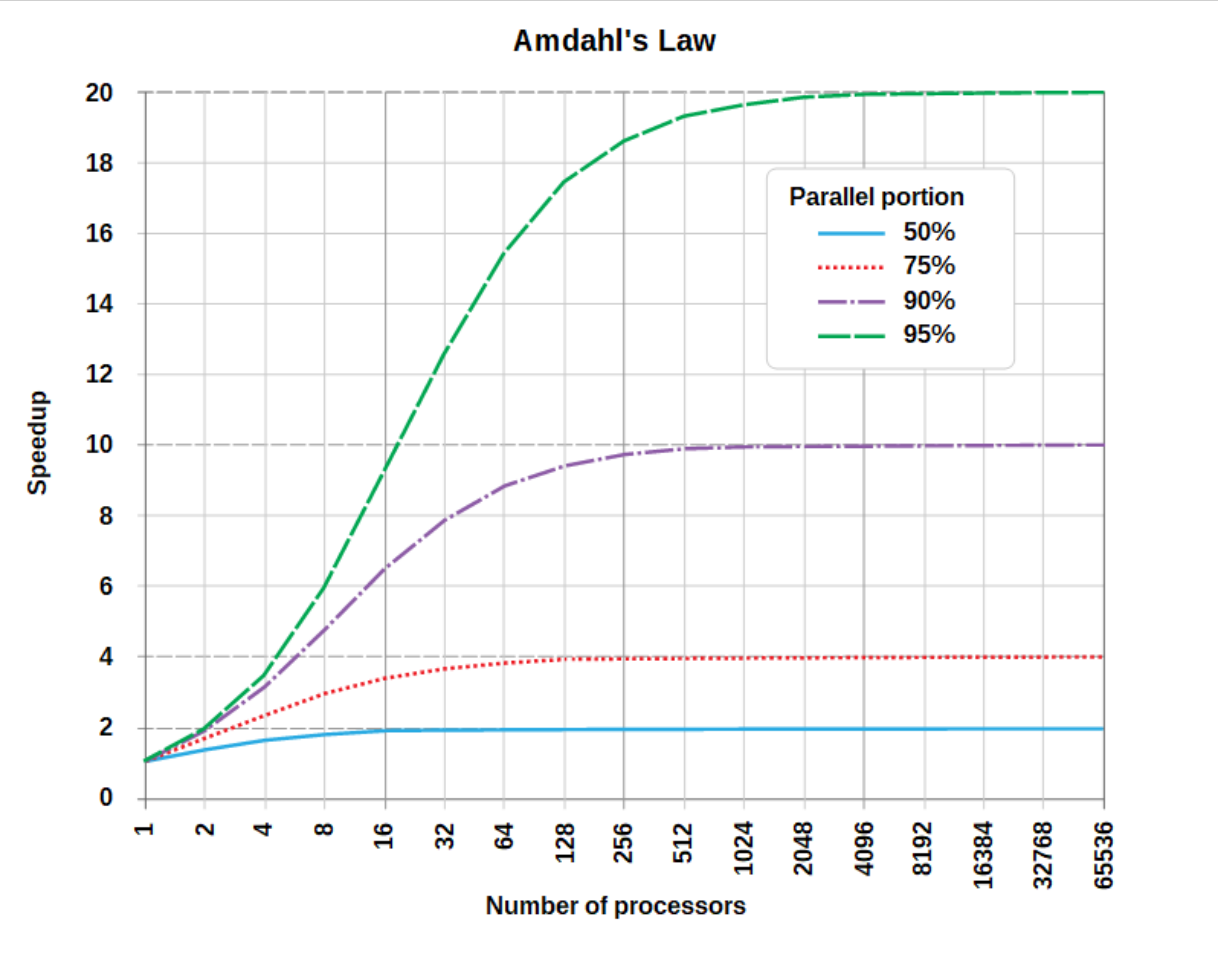
await

cf. JS’s call-back func.

平行計算
(Parallel Computing)

阿姆達爾定律(Amdahl's Law)

無法平行化的部分使得N個workers不會N倍快

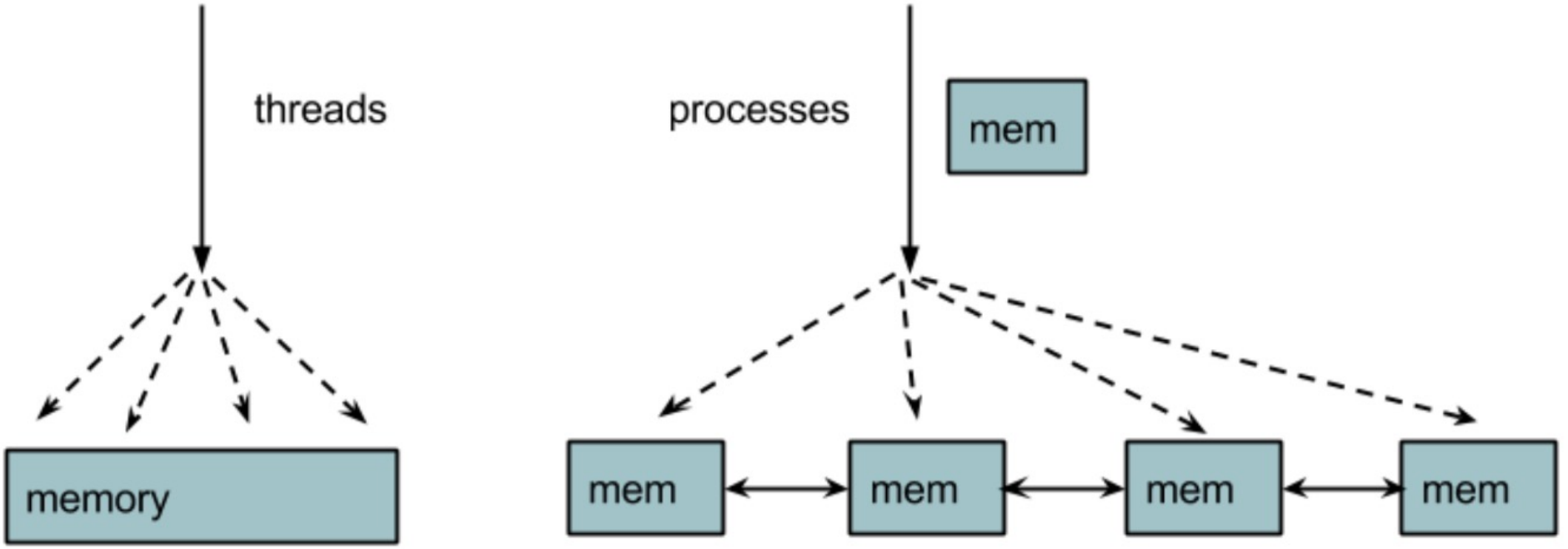


$$T(n) \geq S * T(1) + \frac{P * T(1)}{n}$$

$$T(\infty) \approx S * T(1)$$

Processes vs. Threads

一行程下可有共享記憶體的多執行緒



跑processes有更多的創建/通訊overheads
但能更有效率使用multi-core CPUs for 高計算量工作

第二週介紹的內建函數map

```
import math
def adjust_score(old):
 new=math.sqrt(old)*10
 return new

print(list(map(adjust_score,range(0,101,10))))
```

套上去

multiprocessing的map才是真的平行計算
以後講Big Data的時候會再看到類似觀念

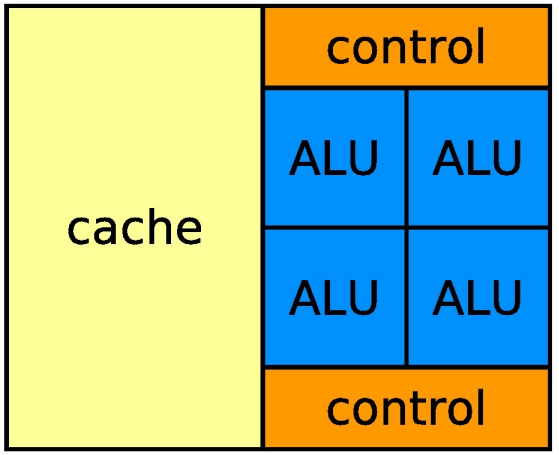
Python: concurrent.futures

在Python 3.2版前是使用threading與multiprocessing

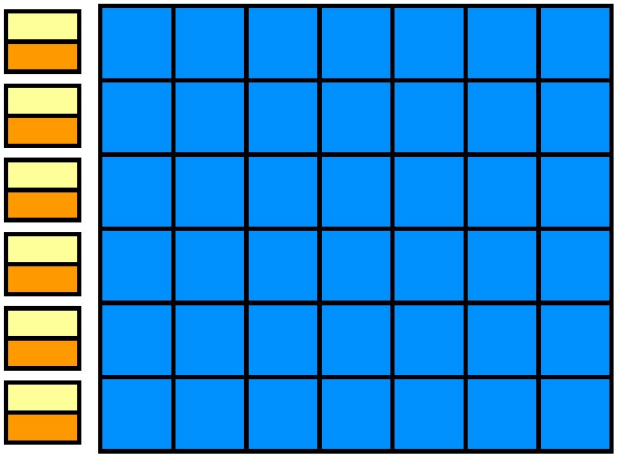
```
import math, concurrent.futures as cf

def adjust_score(old):
 new=math.sqrt(old)*10
 return new

with cf.ThreadPoolExecutor(max_workers=2) as
pool:
 new=pool.map(adjust_score,range(100))

list(new)
```

GPU vs. CPU

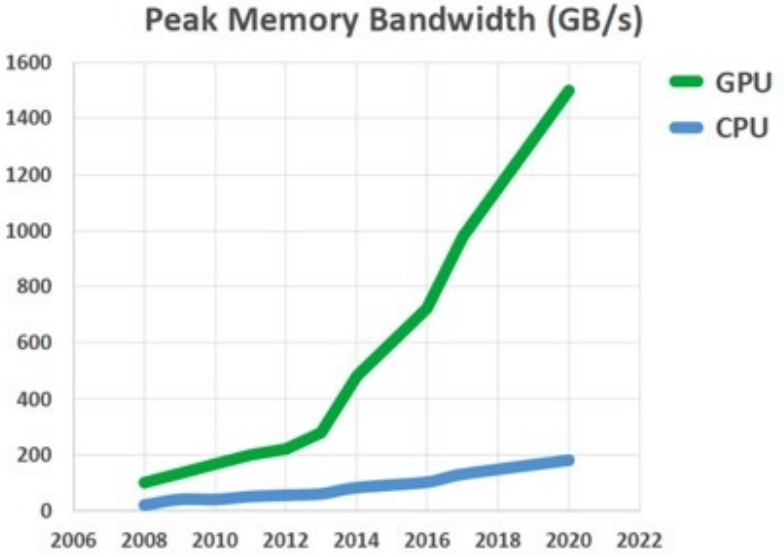
GPU有更多簡化的核心做平行計算達到10倍以上的效能

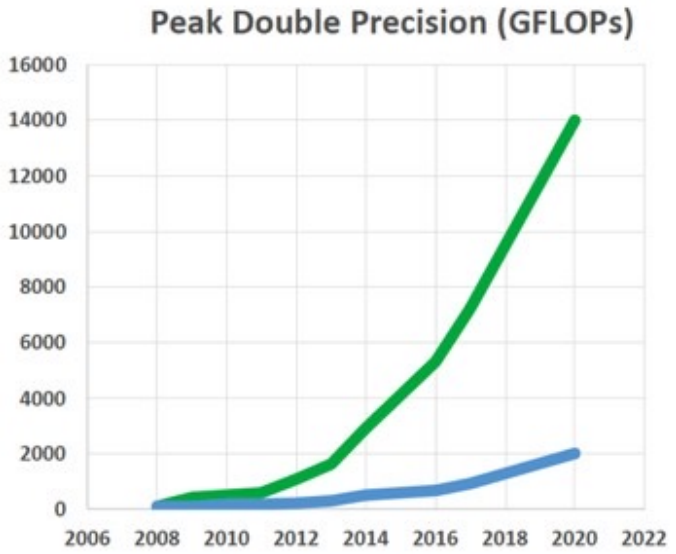


CPU



GPU



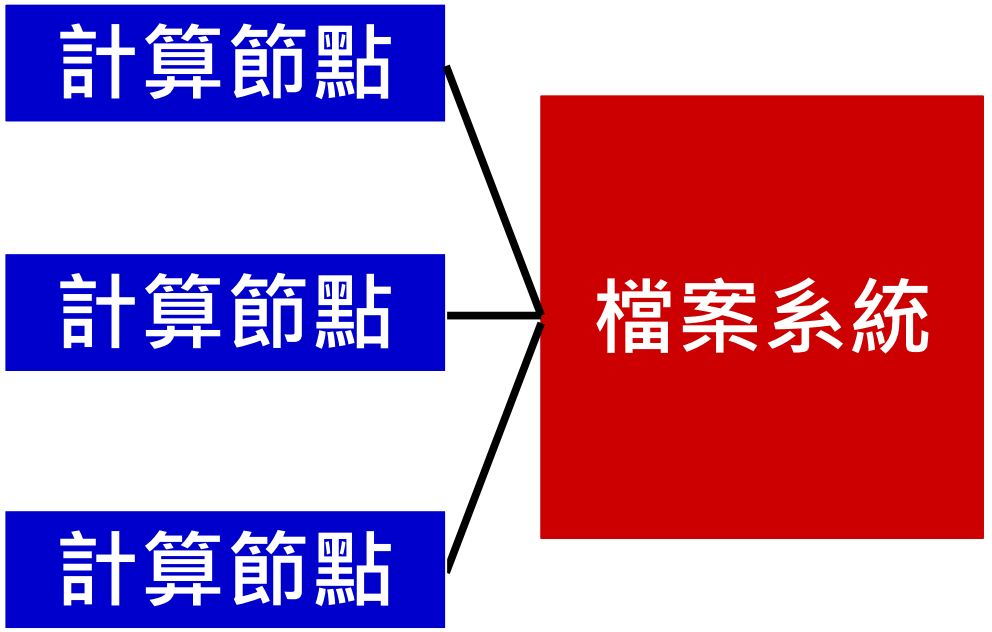


分散式計算架構/環境
(Distributed Computing)

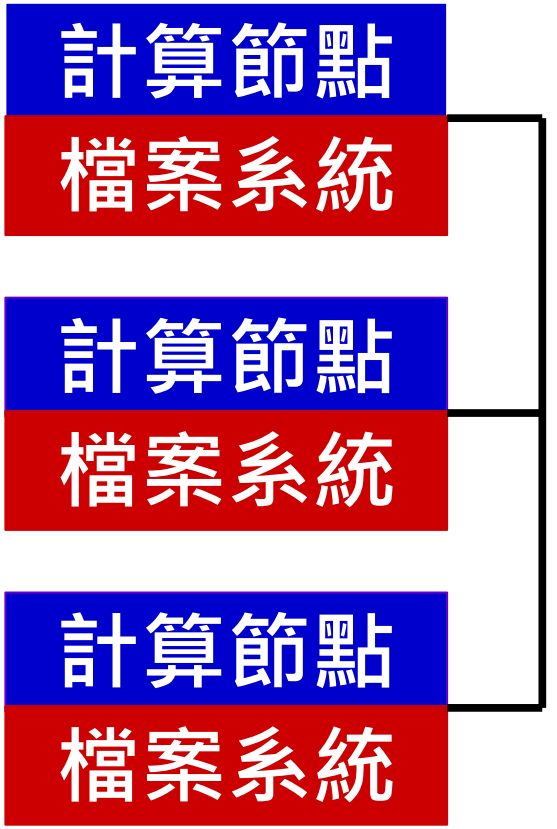
電腦叢級架構

大數據分析: 資料在哪裡, 就在哪裡分析

傳統的科學計算叢集
(資料集中式儲存)

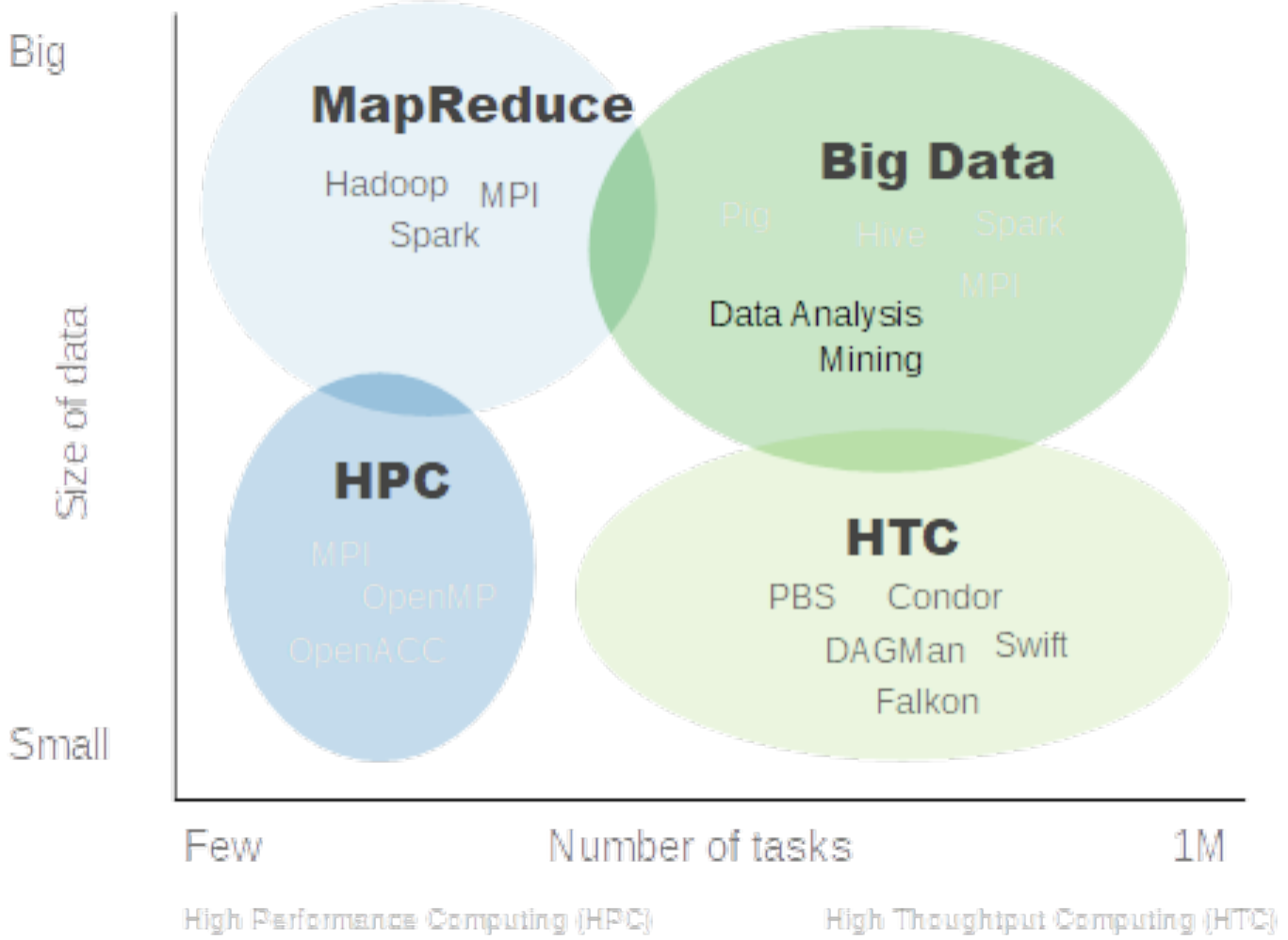


新式的大資料叢集
(資料分散式儲存)



不同領域的資料/計算特性

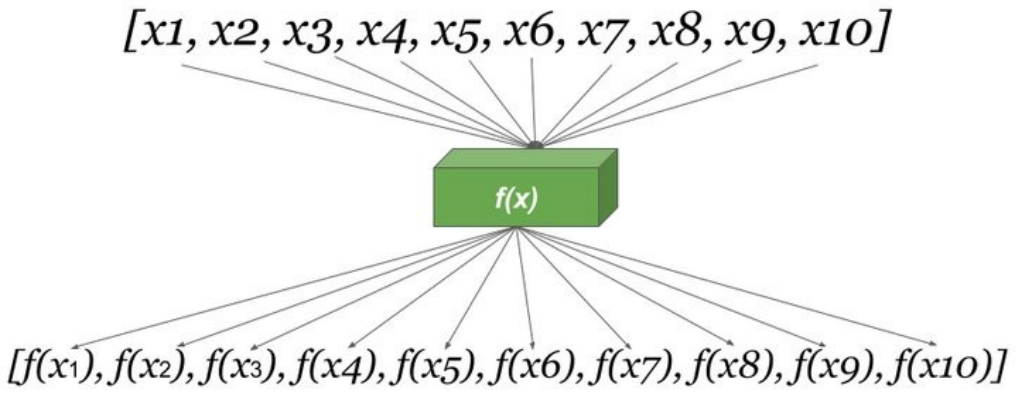
傳統科學計算資料少計算多; 大數據分析資料多計算多



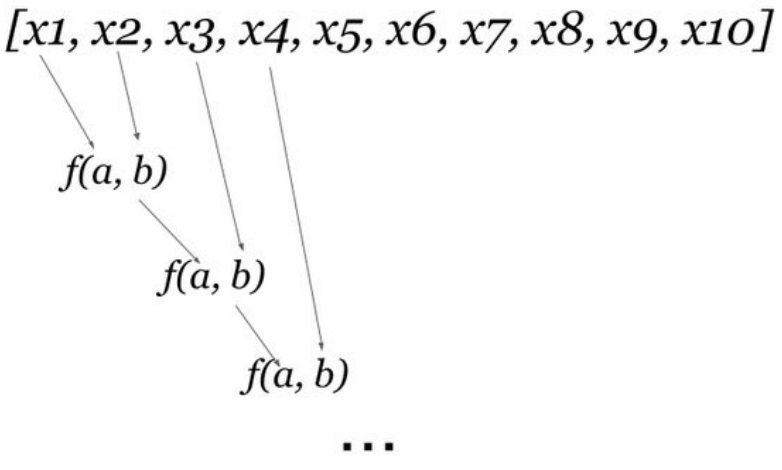
Map+Reduce

就是分而治之之後再彙整所有資訊

Map

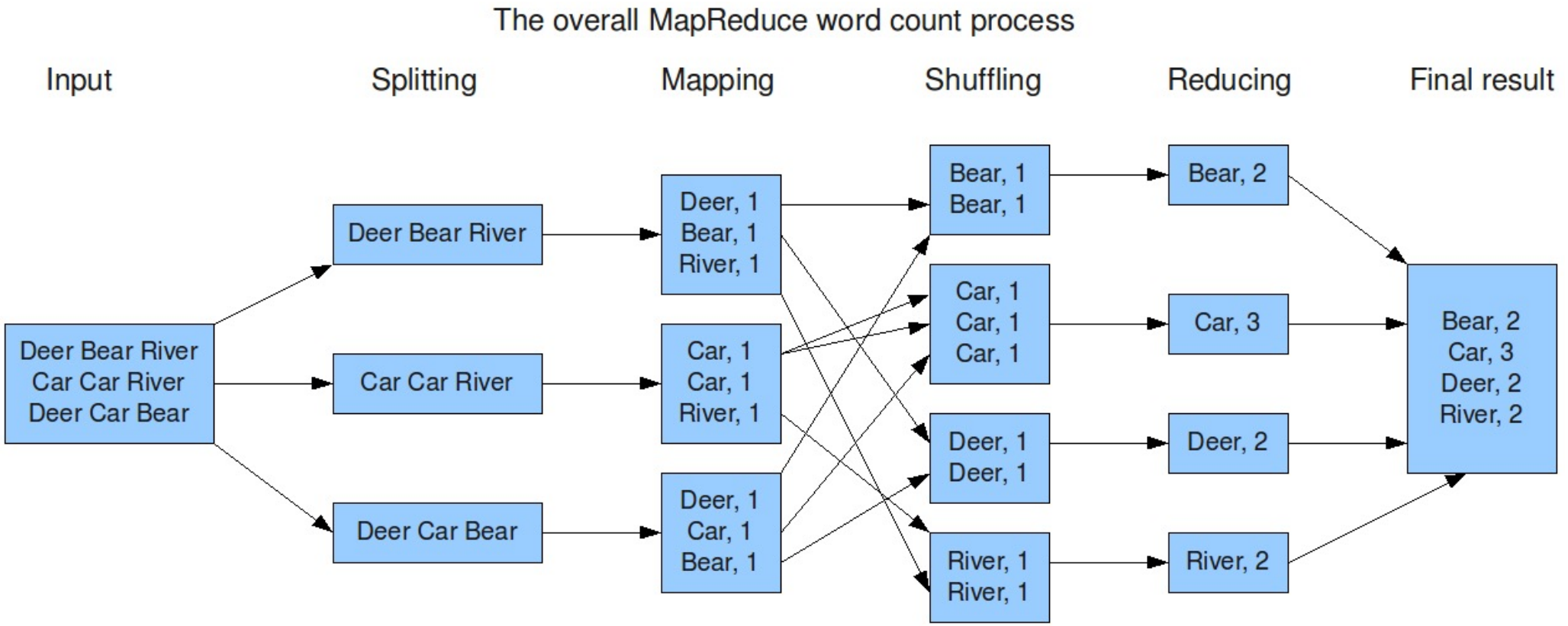


Reduce



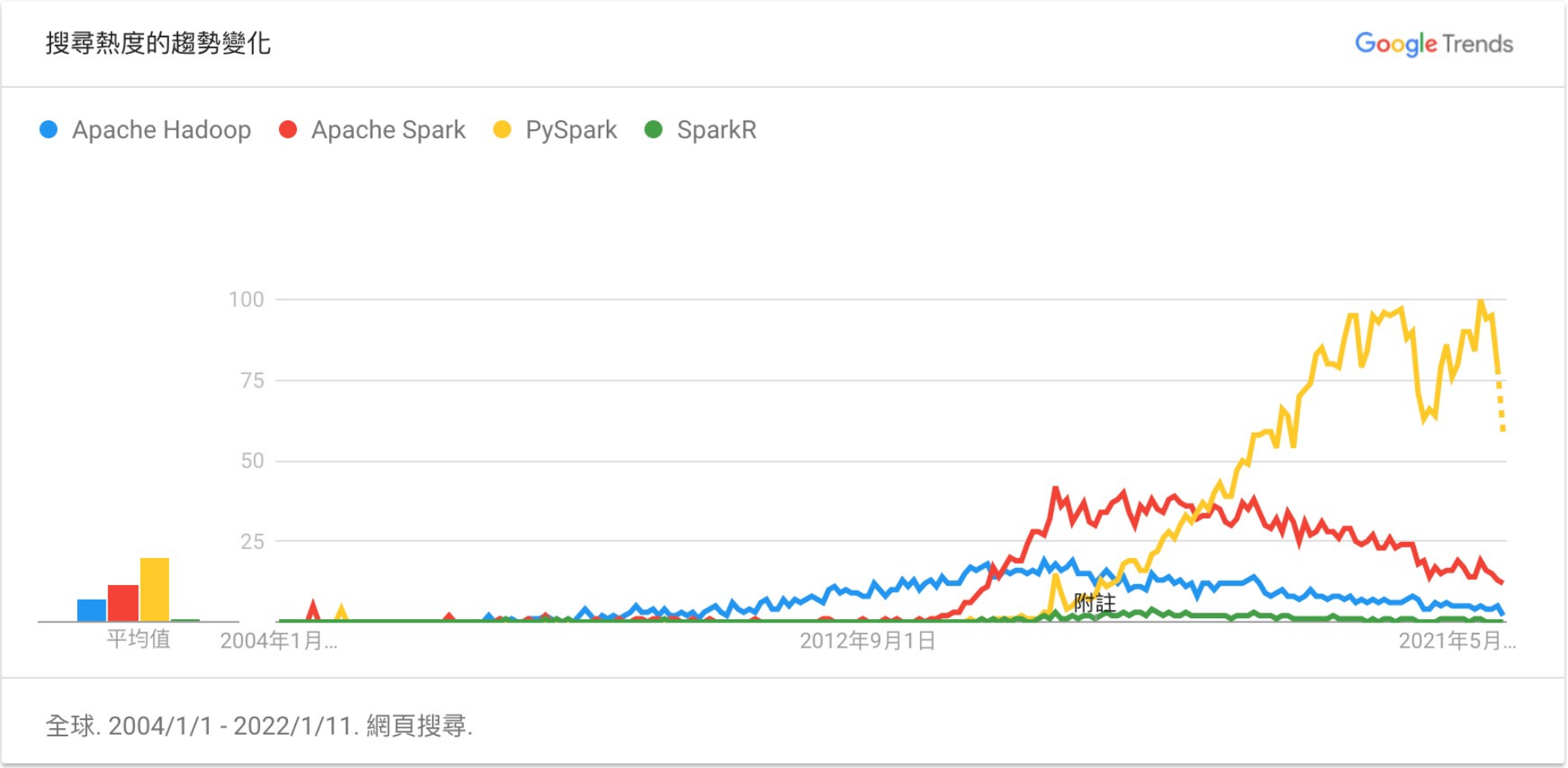
MapReduce範例

光是計算字頻這麼簡單的事都很麻煩



MapReduce的計算環境: 私有雲

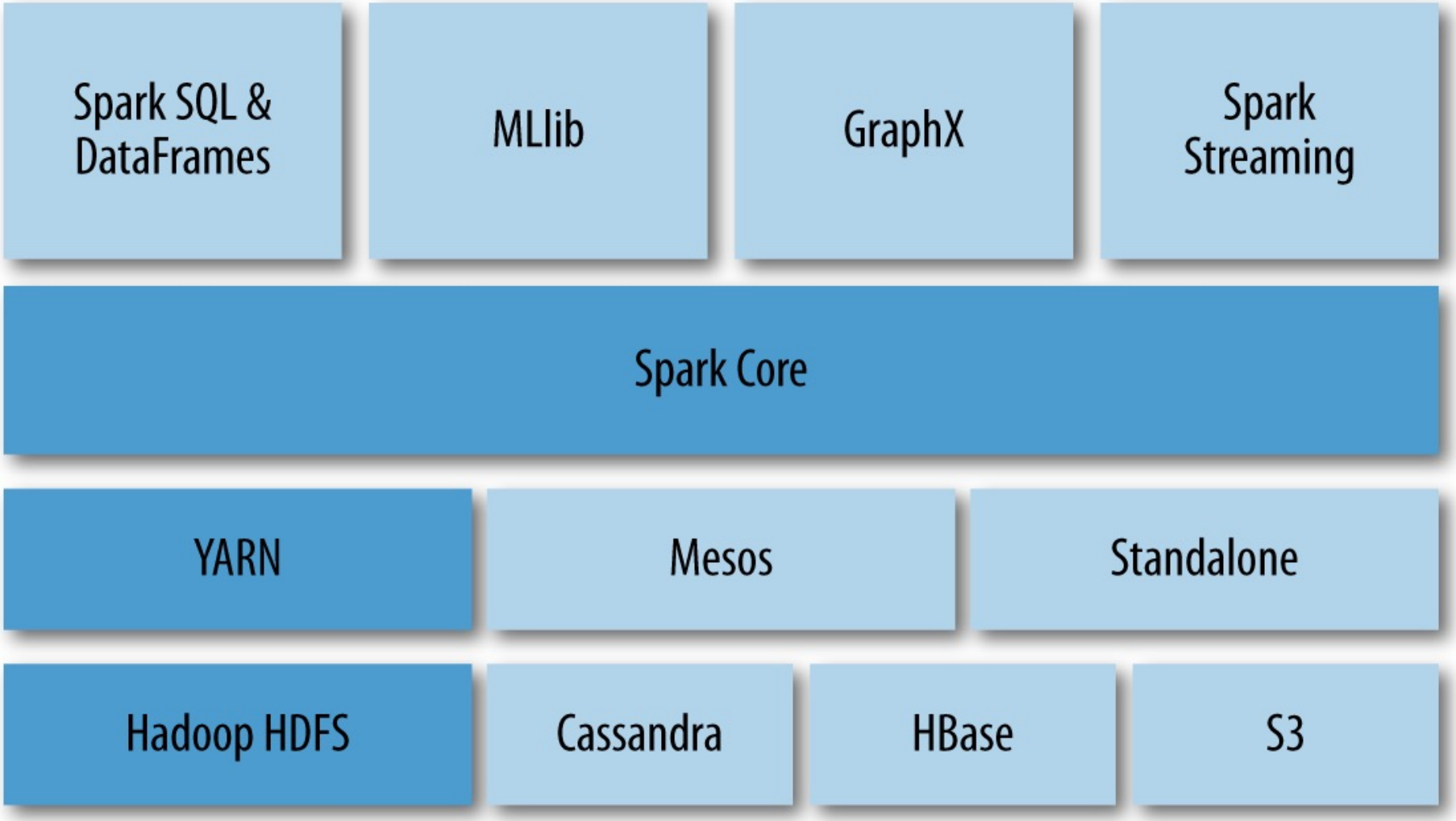
Hadoop → Spark → SparkR、PySpark



PySpark其實很容易在官網學

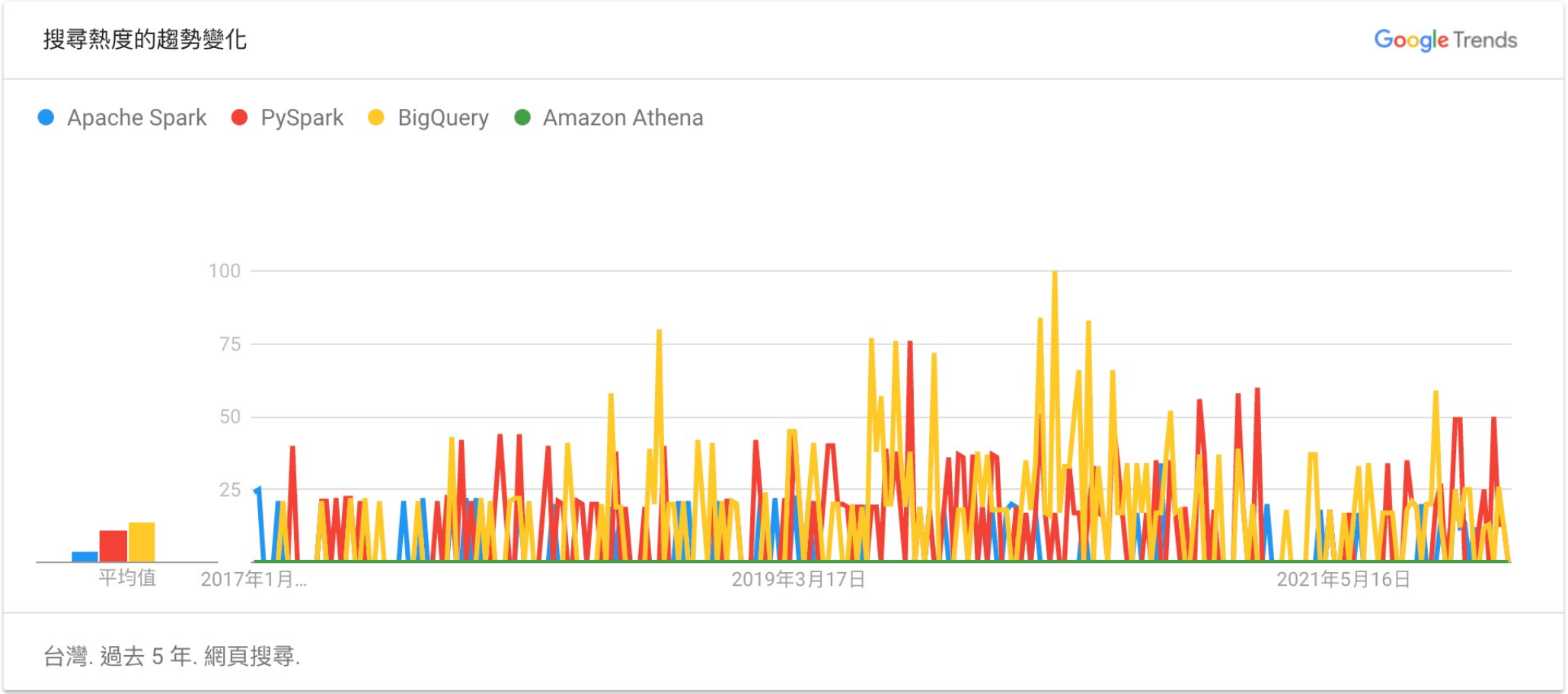
Spark Ecosystem

可用大家熟悉的資料結構(DataFrames)與資料庫語法(SQL)並進行(簡單的)機器學習(Mllib)



公有雲(Public Cloud)的崛起

不用建置電腦叢集就可以使用分散式SQL



GCP BigQuery

100B Benchmark with 3 wildcards

Query Editor | UDF Editor

```
1 SELECT language, SUM(views) as views
2 FROM [bigquery-samples:wikipedia_benchmark.Wiki100B]
3 WHERE REGEXP_MATCH(title, "G.*o.*o.*g")
4 GROUP BY language
5 ORDER BY views desc;
```

No Cached Results

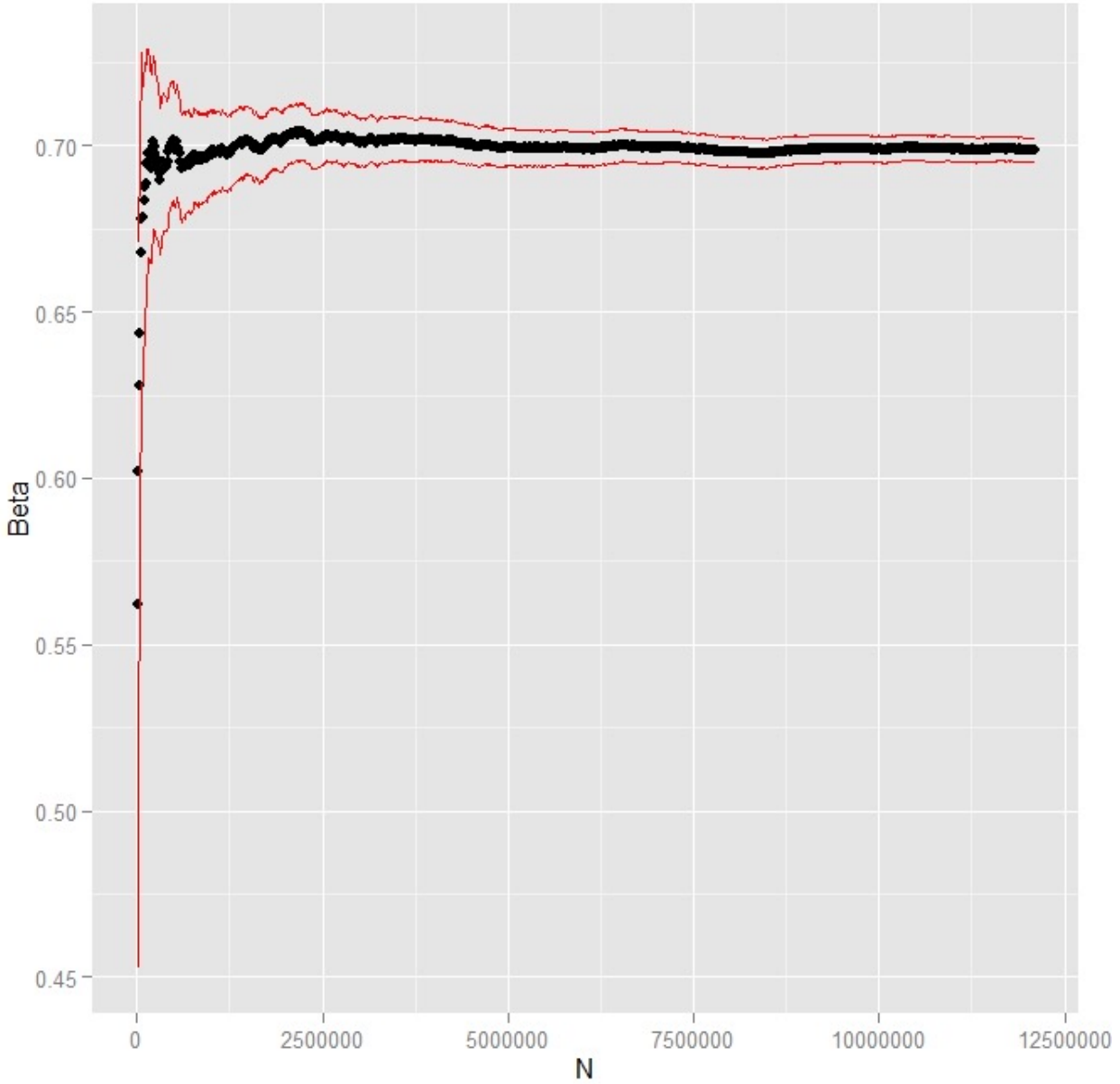
RUN QUERY | Save Query | Save View | Format Query | Show Options

Query complete (24.7s elapsed, 4.06 TB processed)

Big Data的迷思

更多數據≠更好的結果

加入更多數據不見得會增加signal to noise ratio



大數據時代可重相關不重因果!?

「知道what(相關)就夠了,沒必要知道why(因果)」

Google提示相關字

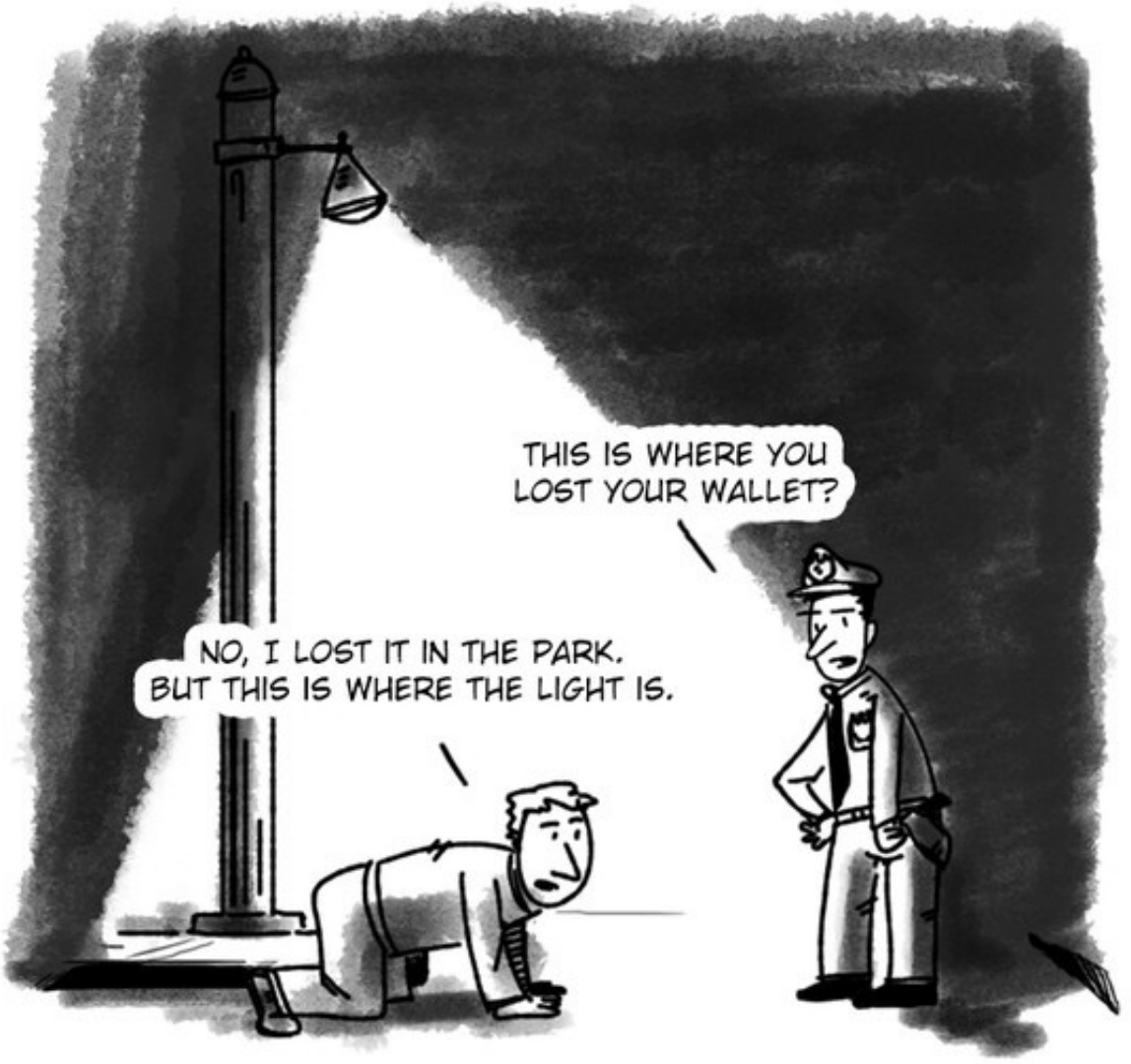
信用與行為的相關

懷孕與購物型態相關

啤酒與尿布常被合購

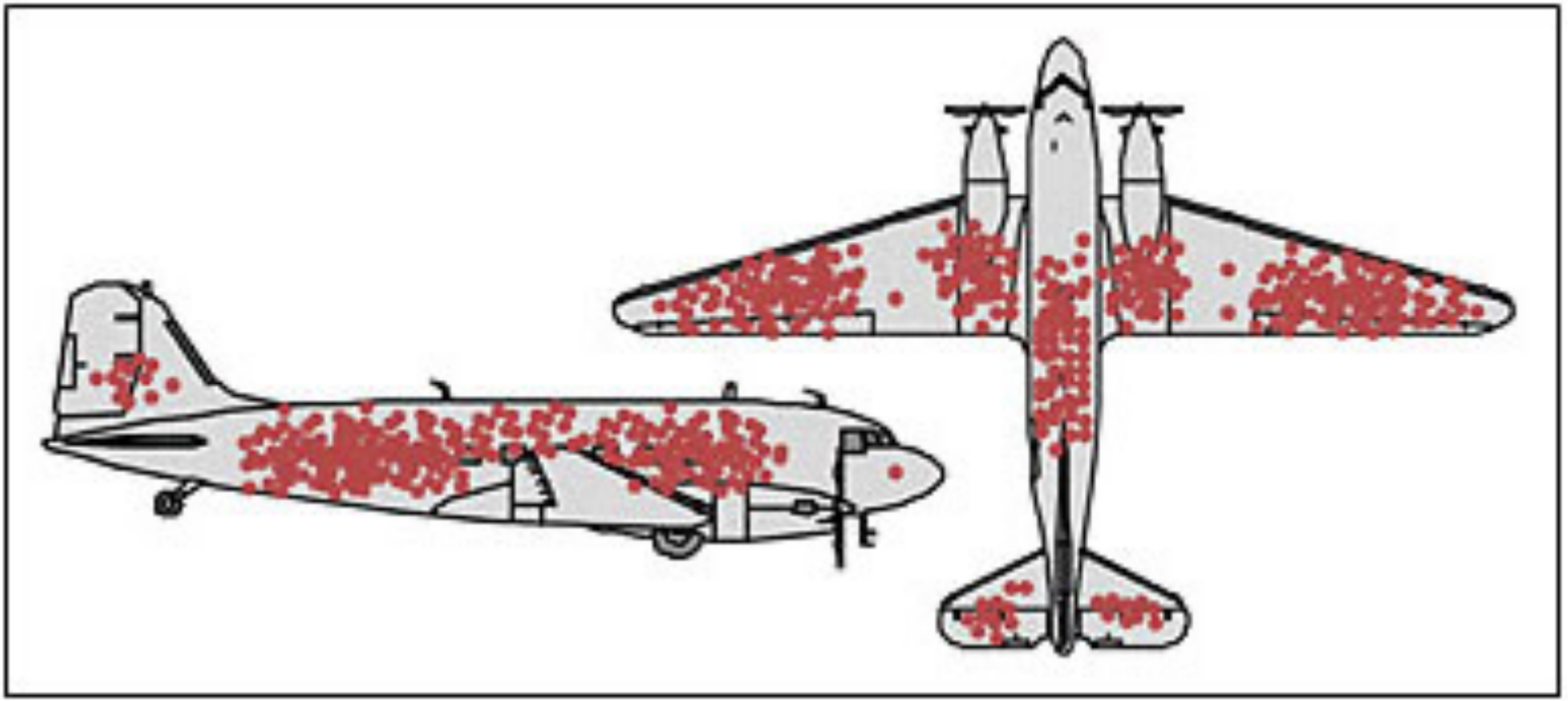
有搜集到關鍵資料嗎? (1/2)

因果關係中的「因」API有提供嗎?



有搜集到關鍵資料嗎? (2/2)

應強化機頭, 機尾, 機翼, 或機身來避免失事?



即便資料搜集完整,答案不見得在資料中

That's all for the semester!

Game Over