Galaxy Software Services

以Machine Learning技術提升現有資訊系統智能服務

Wayne Tien

2018.10.12

GSS Galaxy Software Services 叡揚資訊

國家產業創新獎
卓越中堅企業

2018

李開復：5 秒準則

如果一項工作需要的思考決策，能在5秒鐘內做出決定，就有很大可能被人工智慧技術全部或部分取代

AI是協助我們更高效率地完成手上的某些任務

省下的精力時間去作高價值的事情

大綱

Table Of Content

一、機器學習

二、案例分享

三、續章

國家產業創新獎
卓越中堅企業

Chapter

1

機器學習 (Machine Learning)

機器學習二三事

- 機器學習程序
- 監督式機器學習 (Supervised Learning)
- 非監督式機器學習 (Unsupervised Learning)

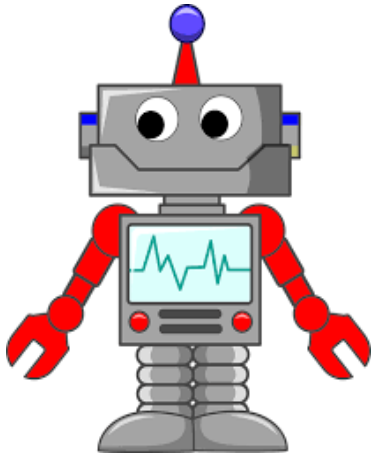
機器學習 (Machine Learning)

透過從過往的資料
和經驗中學習
並找到其運行規則
最後達到人工智慧
的方法

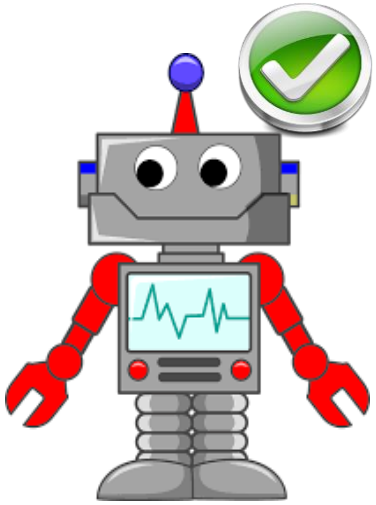
機器學習包含透過樣本資料來訓練機器辨識出**運作模式(演算法)**，而不是採用特定的**規則(Rule)**。

換句話說，機器學習是一種弱人工智慧(narrow AI)，它從資料中得到複雜的函數(或樣本)來學習以創造演算法(或一組規則)，並利用它來做預測。

華倫·巴菲特
Warren Buffett



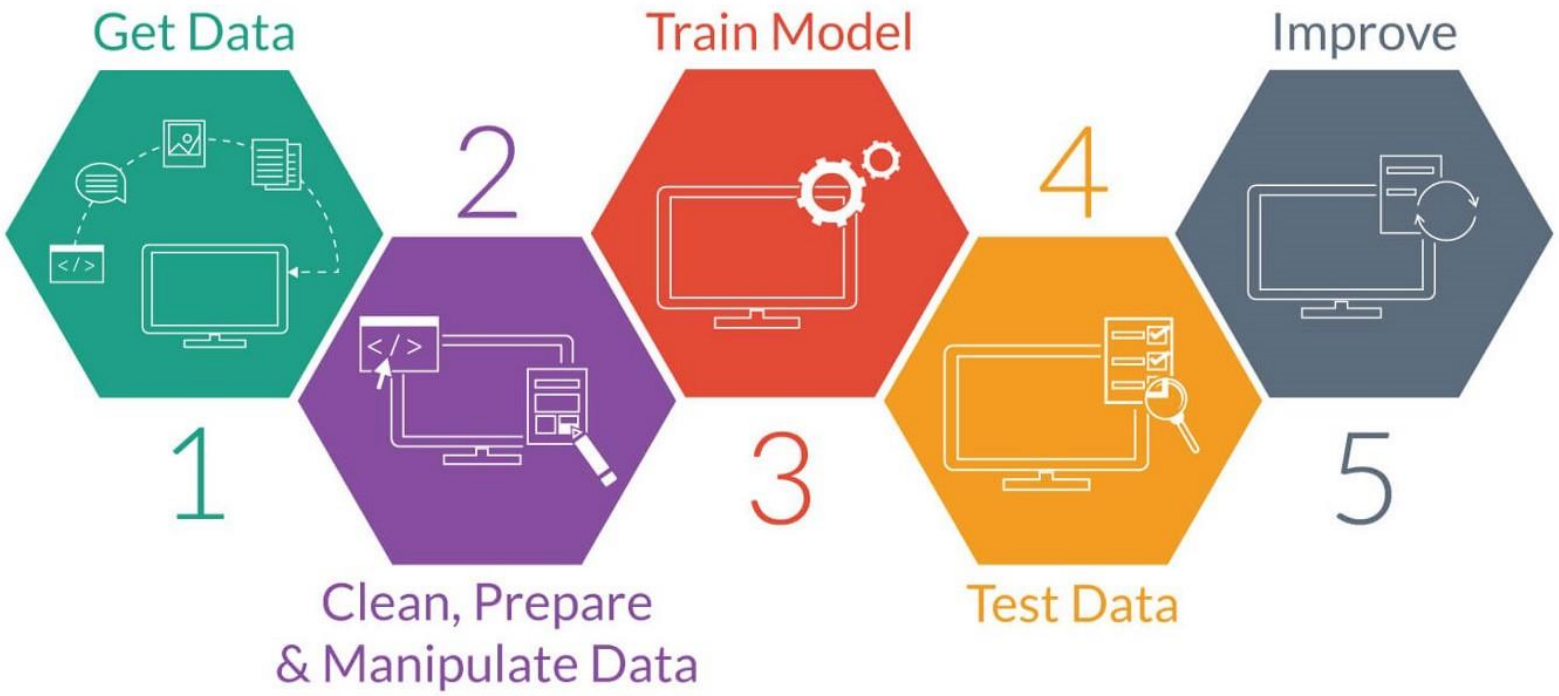
Buffett
選股機器人



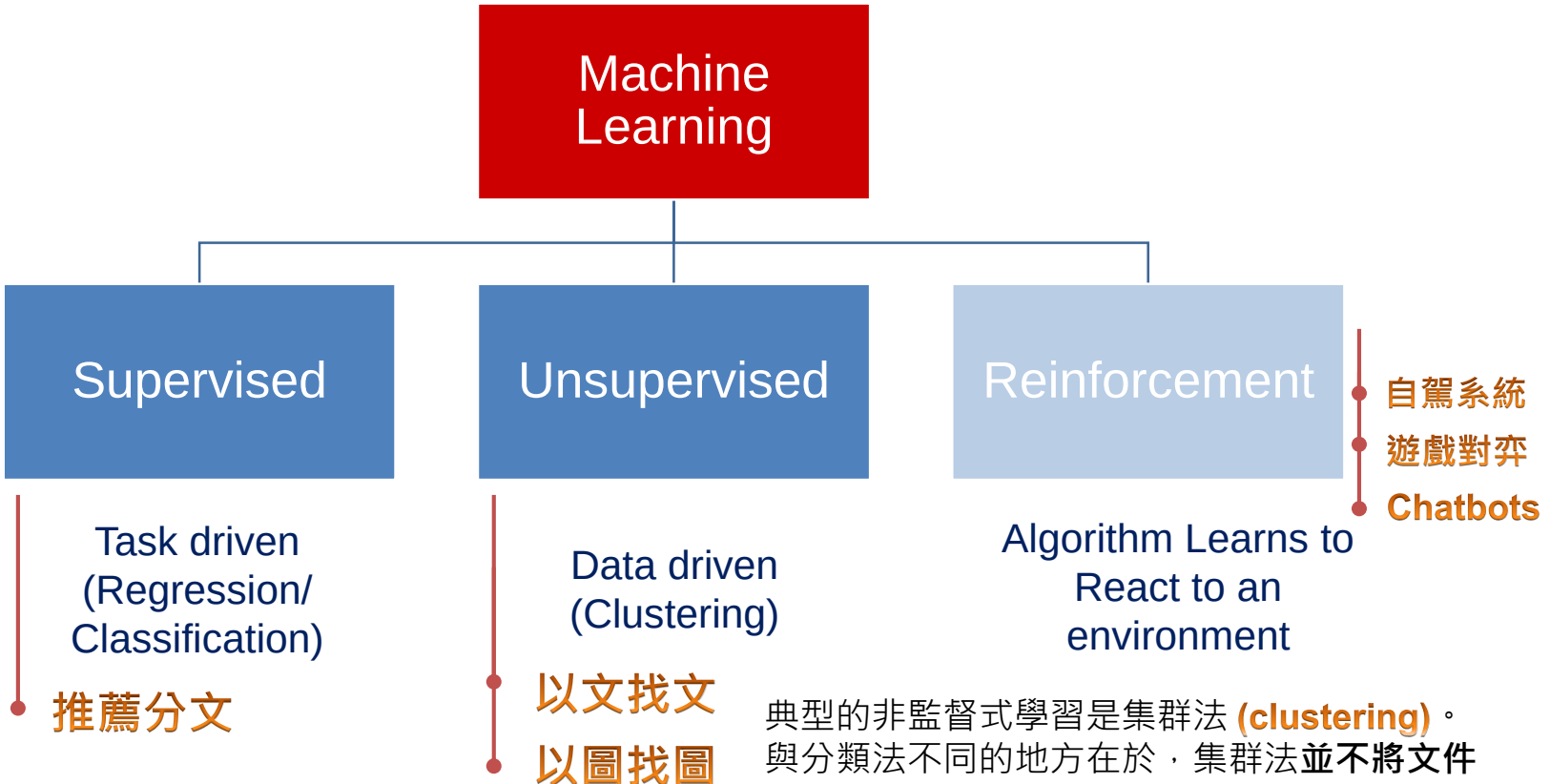
優質
選股機器人

機器學習程序

Typical Machine Learning Workflow



Types of Machine Learning



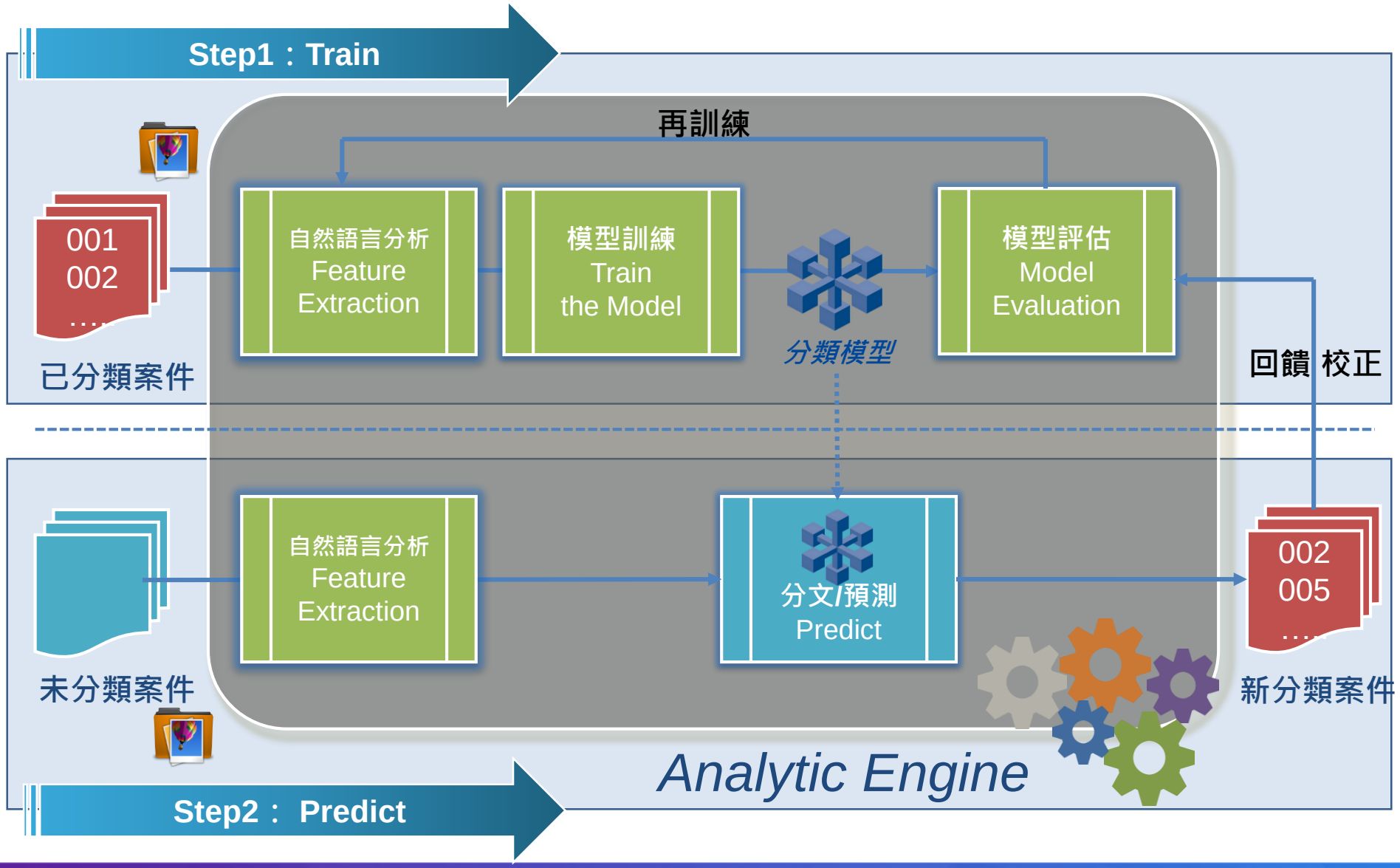
監督式學習使用經標記之訓練集 **(labeled training dataset)** 來建構系統模型，具有預先定義好之輸入與已知之輸出。

在監督式學習中，典型的任務是**文件分類**和**迴歸分析**。

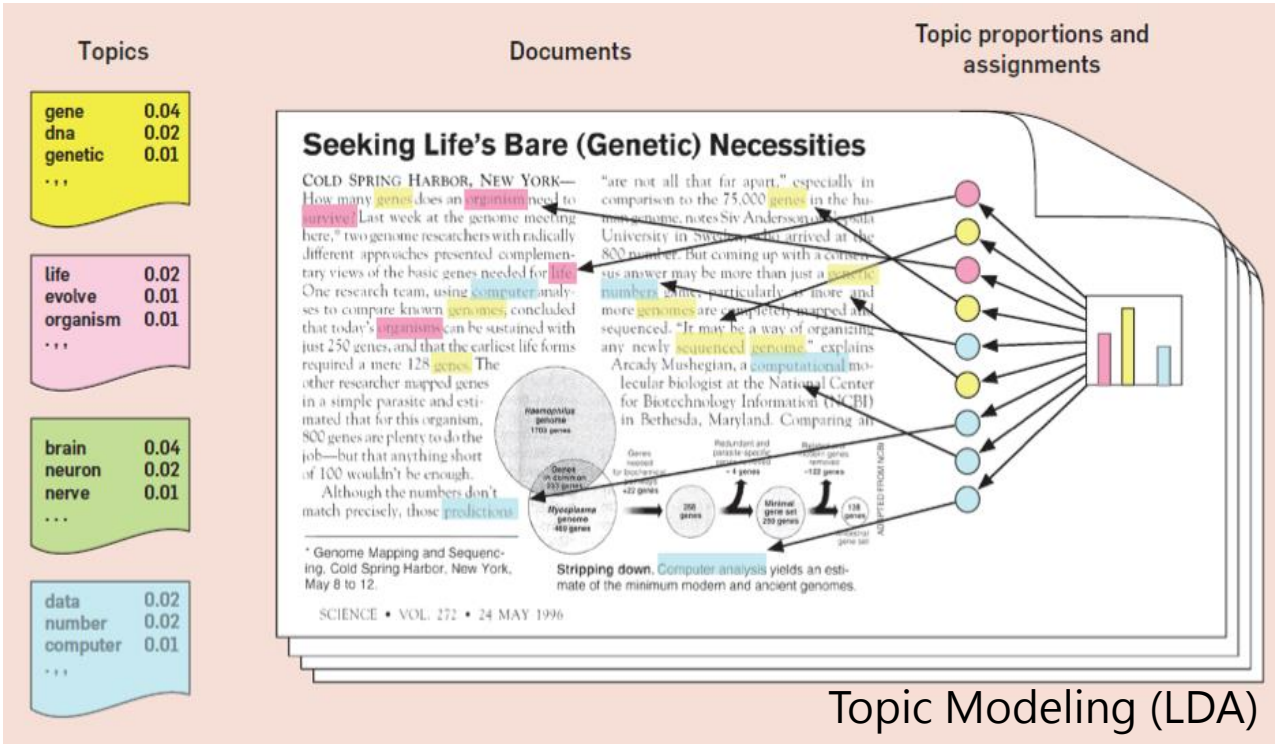
典型的非監督式學習是集群法 **(clustering)**。與分類法不同的地方在於，集群法**並不將文件預先分類**，而是讓演算法依據每個文件相似度與相異程度，將文件區分成不同之群集。

關鍵在於讓同一群集內文件間的差異盡量小，彼此相似程度高；至於不同群集間的文件則有大差異，從而可將每一文件歸屬至一個特定群集，對於文件檢索之效率與準確度之提升有極大助益。

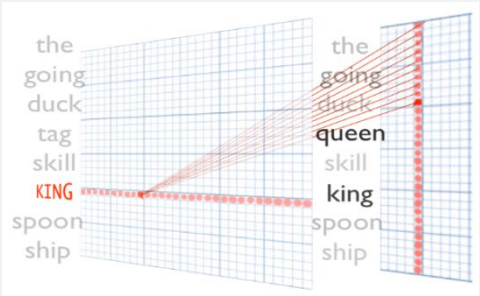
監督式機器學習 (Supervised Learning)



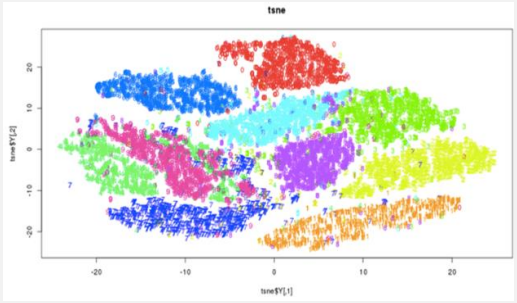
非監督式機器學習(Unsupervised Learning)



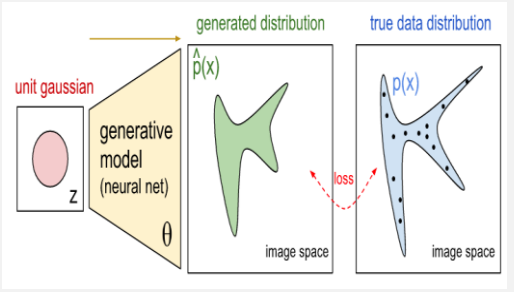
Topic Modeling (LDA)



Similar Word Extraction



Clustering



Generative Models

Chapter

2

案例分享

智慧局產業專利知識平台

- 推薦分類
- 以文找文
- 專同辭庫

智慧局產業專利知識平台實際應用

智慧局

局處長官
產業分析員
審查人員

專利推薦分類

產業專利知識網

企業

高階主管
市場分析員
研發人員

IPKM

產業專利知識平台

專利趨勢分析

六局專利
共通格式

技術辭庫

專利檢索

社群網路

專利推薦

以文找文

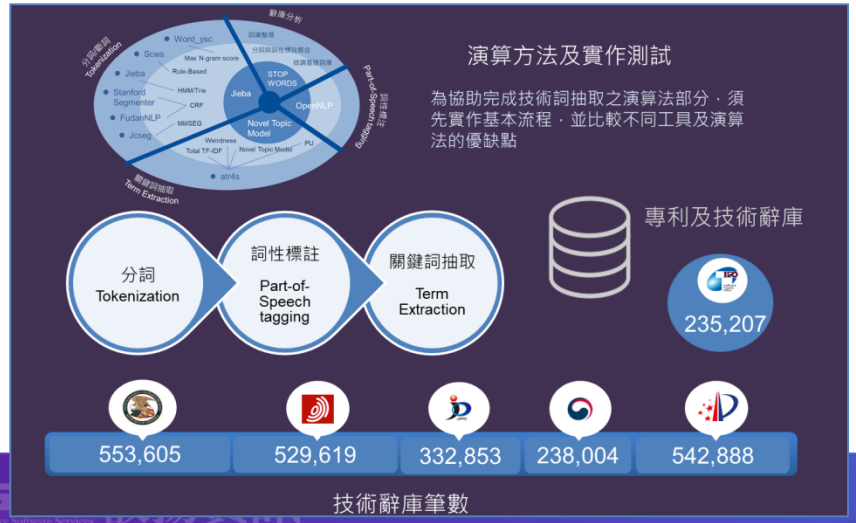
- 建置專利自動分類機制，以減輕專利分類人力負荷並提高分類正確性
- 分析，蒐集、彙總六局專利技術名詞，建立專利技術辭庫，做為後續高階檢索應用發展基礎

- 協助企業快速關注全球專利趨勢走向、強化智財管理與提升競爭力
- 聚焦自身產業專利搜尋，避免研發資源重複投入、監控競爭對手專利及針對產品實施前侵權風險評估

機器學習實作

專利以文找文

專利技術辭庫



專利自動分類

知識平台

智慧局產業專利知識平台

專利自動推薦
對於有興趣的專利
可進行收藏

個人檢索/瀏覽歷程
以文找文功能

瀏覽及收藏排行

追蹤訂閱快速
瀏覽區

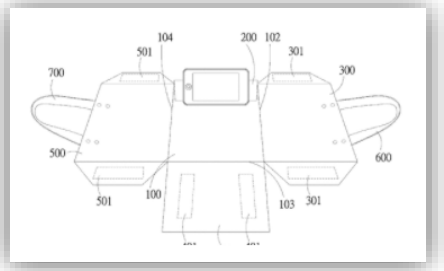
智慧局產業專利知識平台

產業別分類主題檢索及瀏覽

關鍵字檢索

專利推薦分類 - Why ?

專利發明



多功能摺疊餐袋

專利細分69,000種分類

專 利 分 類

階層		類號	編排	總數	範例	說明
一	Section	A~H	部	8個	A	人類生活需要
二	Class	二位數	主類	128個	45	手攜物品或旅行品
三	Subclass	一個大寫的英文字母	次類	628個	C	小包；行李箱；手提袋
四	Main group	1至3位數加/00	主目	69,000個	011	午餐或野餐盒或類似物
五	Sub group	將” / ”後的00改為2~4個數字	次目		/20	

專利推薦分類文件特性

中華民國第 I453740 發明專利

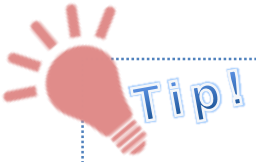
專利名稱：

「光學儲存媒體和從光學儲存媒體閱讀 資料之方法，以及包括閱讀光學儲存媒體上 資料所用的光學拾波器之裝置」，

摘要：

「本案光學儲存媒體包括基片層(2)，和 設置在基片層(2)上之資料層(3)，資料層(3)包 括資料，配置在磁軌內，成為標記和空格。 保護碼(PC)寫碼於一磁軌或磁軌之一部份，該碼包括第一尺寸(w1)之標記(P)和較小的第二 尺寸(w2)之標記(P)。第二尺寸(w2)之標記尤 其是寬度，較第一尺寸(w1)的標記之寬度 小。欲得保護碼(PC)，使用此方法包括步驟 為，以第一雷射功率閱讀磁軌或磁軌(T)之一 部份，得第一資料訊號(D1)，在另一步驟以 不同於第一功率之第二功率，閱讀同樣磁軌 或磁軌(T)之同樣部份，得第二資料訊號 (D2)，並考量第一和第二資料訊號，計算保 護碼。保護碼之計算，可例如利用各資料閱 讀裝置之微處理器進行。」。

本件專利係關 於DVD之專利，然無論「發明名稱」或「摘 要」皆未出現「DVD」、「CD」、或 「CD-R」等用語。



不同知識領域資料內容特性多有不同，有其特殊詞彙或是專有名詞

針對不同知識領域會採取不同的對策，並經多過次模擬去決定策略及演算手法

沒有一種模型 (model) 可直接套用於各個知識領域

專利推薦分類流程


專利說明書XML解析
分詞/斷詞
Jieba
Stop Words 移除
特徵化
特徵值演算
TFT-IDF
詞轉換成向量
Word2Vec
三階分類訓練建立模型
五階分類訓練建立模型
80.3%
分類預測
70.1%

推薦分文機器學習(Supervised)流程一

推薦分文機器學習(Supervised)流程二

以文找文

Tip!

使用者的想法

「尋找資料花太多時間，想快點找到相關的資訊」

「不太清楚要下什麼關鍵字 ，才能找到我要的東西」

「也許給我一些相關(可能沒看過或一時沒想到)的內容也不錯」

以文找文實作（輸入）

輸入一段文字或一篇文件 → 文字前處理 → 演算分析 → 主題關聯延伸 → 相似度權重排列 → 檢索結果呈現

根據本發明一實施例，提供一種半導體晶片封裝構件，包含有一基板，具有一晶片安裝面；複數個焊接墊，設於該晶片安裝面上；一第一虛設接墊以及一與該第一需設接墊間隔開的第二虛設接墊，設於該晶片安裝面上；一防焊遮罩，設於該晶片安裝面上，並部分覆蓋該焊接墊、該第一虛設接墊與該第二虛設接墊；一晶片封裝，安裝在該晶片安裝面上，並透過設於該複數個焊接墊上的複數個錫球電連接該基板；一分立元件，設於該晶片封裝與該基板之間，該分立元件具有一第一連接端與一第二連接端；一第一焊錫，將該第一連接端、該第一虛設接墊與該晶片封裝連接起來；以及一第二焊錫，將該第二連接端、該第二虛設接墊與該晶片封裝連接起來。其中該分立元件包含一基板側電容、一去耦電容、一電阻或一電感。

檢索 取消

半導體 晶片 封裝 構件 基板 晶片 ~~安裝 面 複數~~ **焊接** ~~墊 設於~~ **晶片** ~~安裝 面上 虛設 接墊 需 設接 墊 間隔 開 虛設 接墊 設於 該 晶片 安裝 面上~~ **防焊 遮罩** 設於 **晶片** ~~安裝 面上 部分 覆蓋~~ **焊接** ~~墊 虛設~~ **接墊** ~~虛設 接墊~~ **晶片** ~~封裝 安裝 在~~ **晶片** ~~安裝 面上 透過 設於 複數~~ **焊接** ~~墊 上 複數~~ **錫球電** 連接 **基板** ~~分立~~ **元件** ~~設於~~ 該 **晶片 封裝 基板** ~~之間 分立~~ **元件** ~~連接 連接 端~~ **焊錫** ~~將~~ 連接 ~~端 虛設~~ 接墊 **晶片 封裝** 連接 **焊錫** 連接 ~~端 虛設 接墊~~ **晶片 封裝** 連接 ~~分立 元件~~ **基板** 側 **電容** ~~去~~ 耦 **電容電阻 電感**

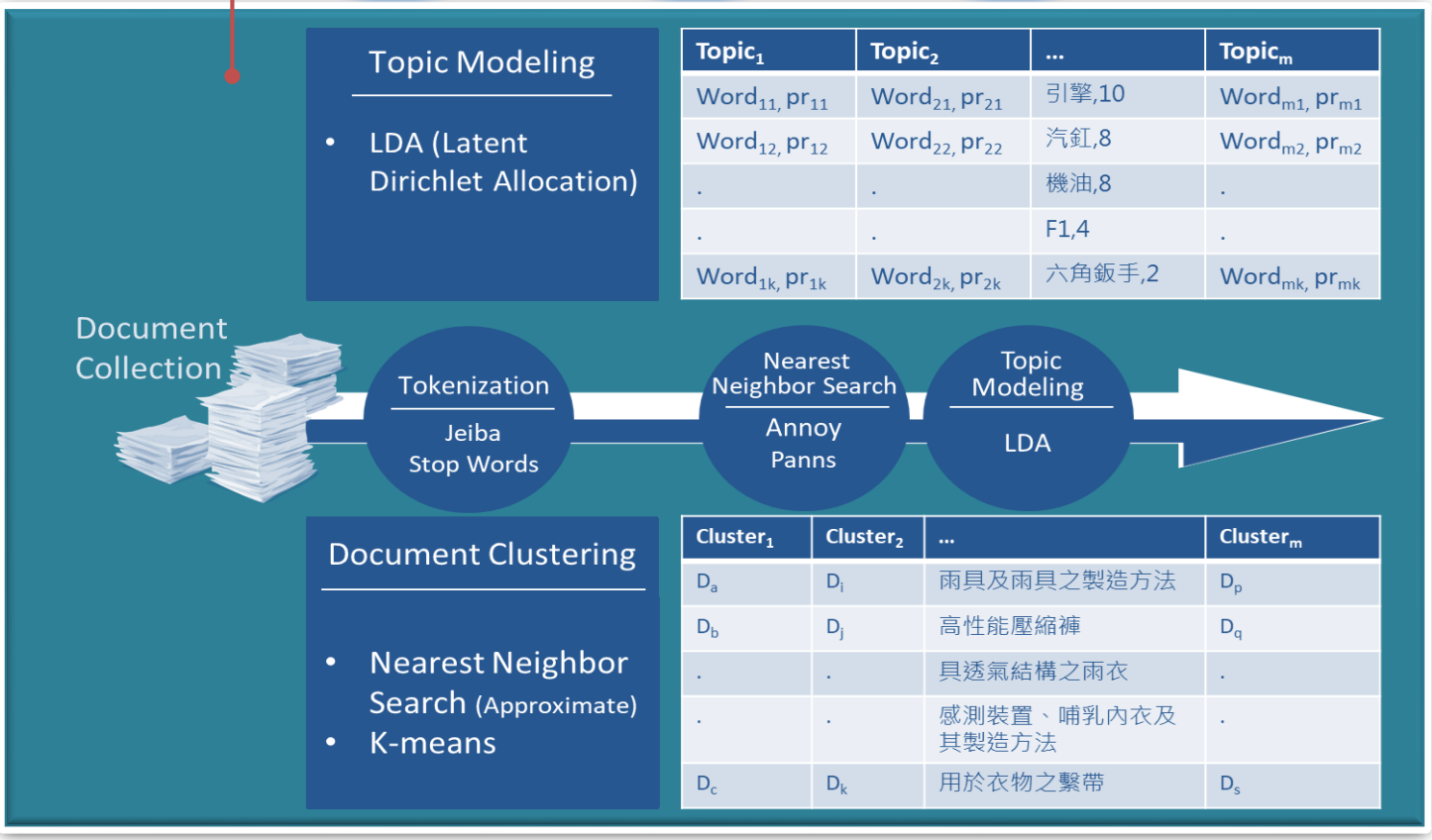
以文找文實作（分析學習）

輸入一段文字或一篇文件 → 文字前處理 → 演算分析 → 主題關聯延伸 → 相似度權重排列 → 檢索結果呈現

A 文件特徵向量化

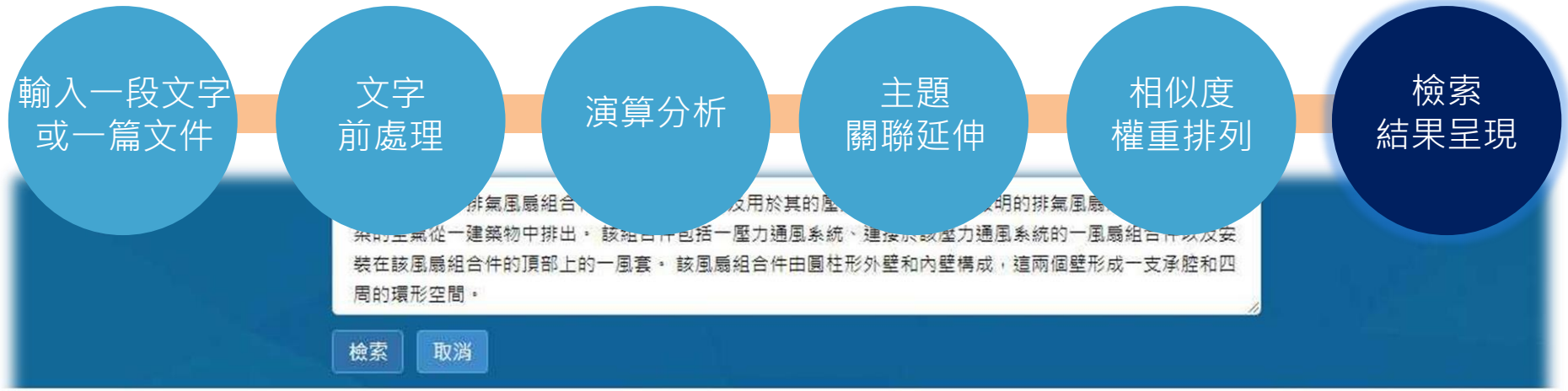
B 文件特徵索引

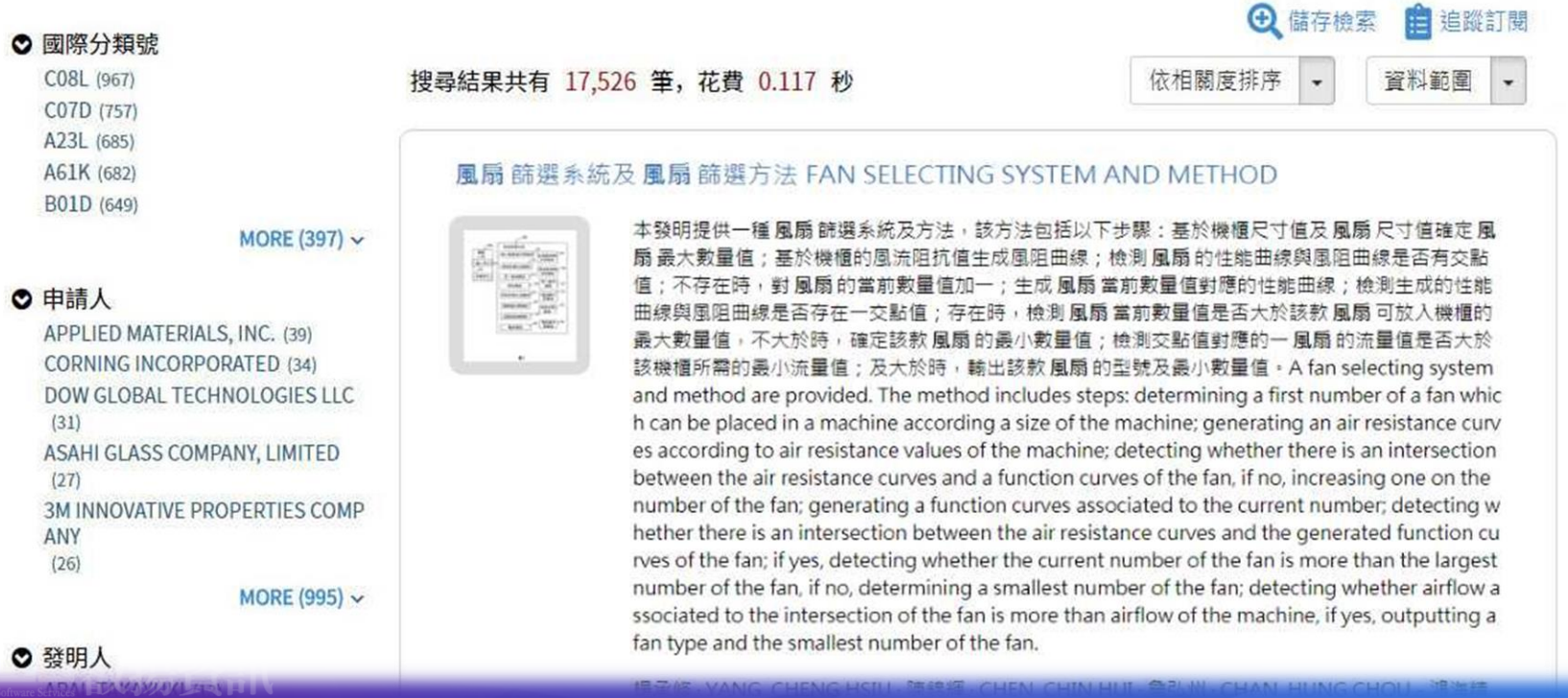
C 最鄰近搜尋



以文找文實作（輸出結果）

輸入一段文字或一篇文件 → 文字前處理 → 演算分析 → 主題關聯延伸 → 相似度權重排列 → 檢索結果呈現

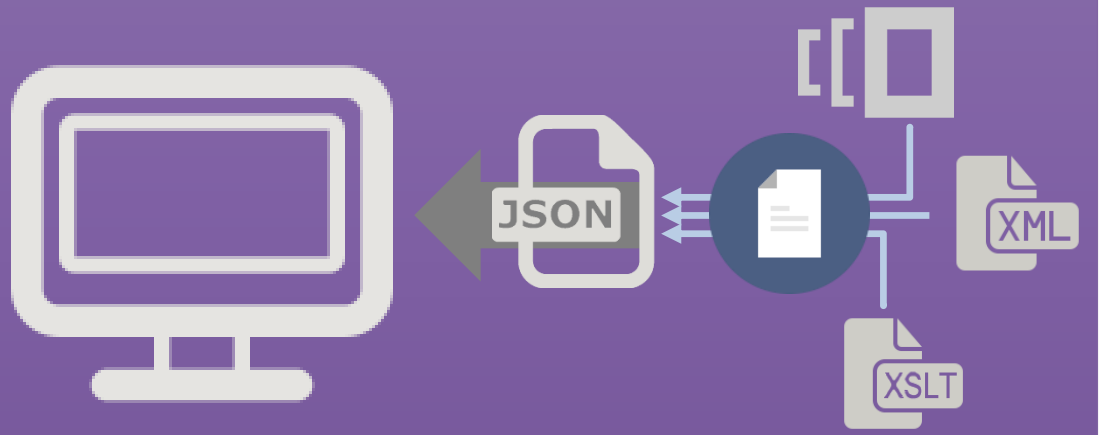




多國專利文件轉置成果

106 年 1 月 ~ 106 年 7 月
發明專利公開公報轉置資料數

- 193,640
- 88,907
- 104,195
- 61,484
- 25,608



575,080

技術辭庫學習流程

辭庫分析

詞庫整理

- 採用專利技術名詞 **中英對照詞庫**
- 刪除不必要之詞彙及片段
- 另加入 **Jieba** 工具簡體中文詞量

分詞與詞性標註的媒合

- **詞性標註** 是指對於句子中的每個詞指派適合的詞性；名詞、動詞、形容詞等等
- 詞庫與標註詞性在訓練模組時需要將兩者調整為相同標準

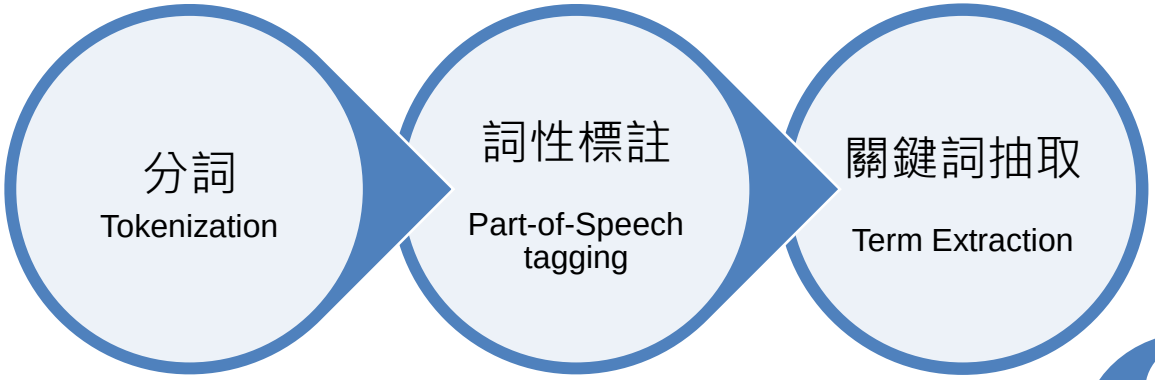
微調基礎詞庫

- 詞庫內容混雜，需要由關鍵詞抽取結果來找尋詞彙是否錯誤

分詞 (Tokenization)
- Jieba 演算法工具

標註詞性 (POS tagging)
- OpenNLP 演算法工具

關鍵詞抽取 (Term-Extraction)
- Atr4s 演算法工具

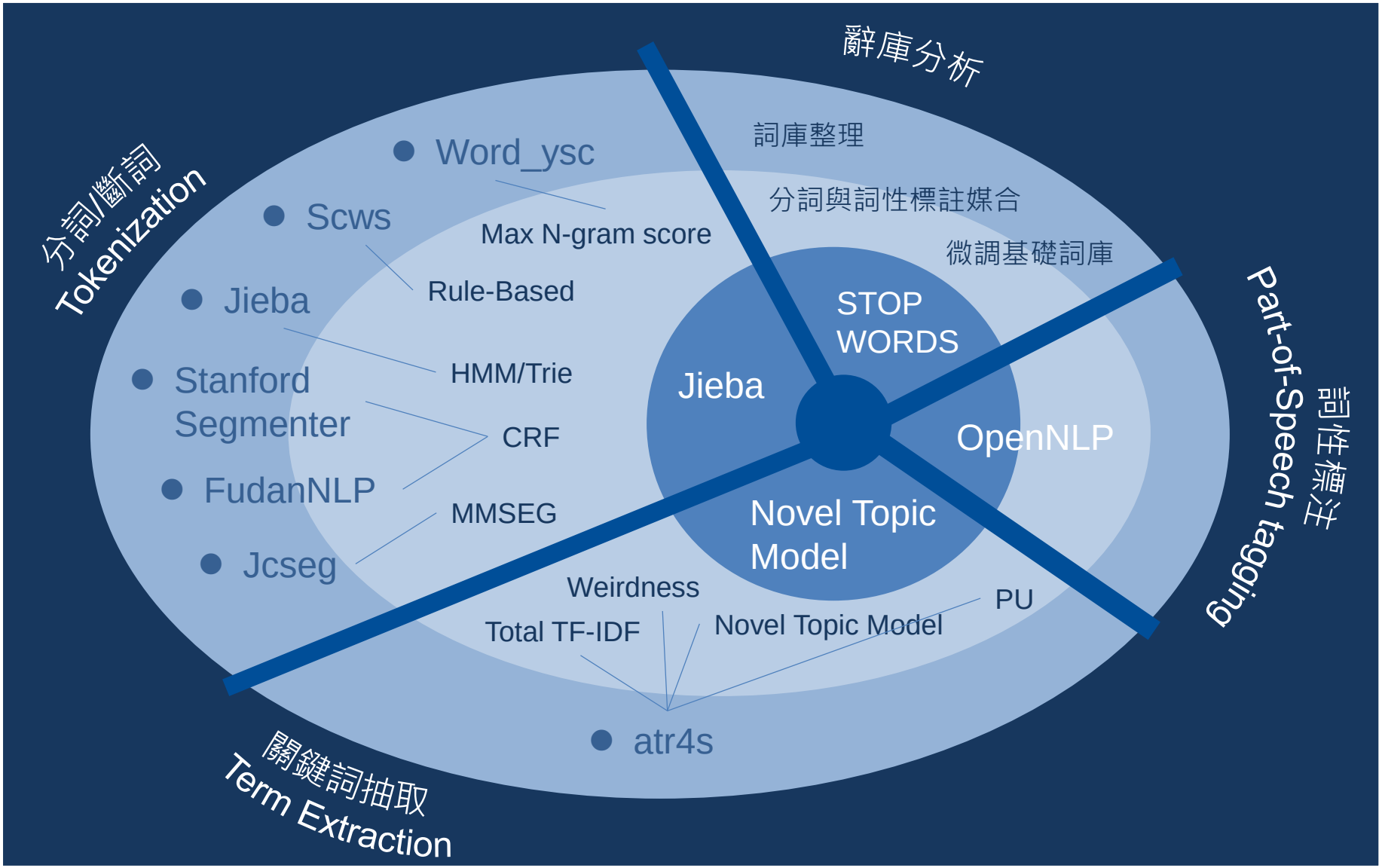


技術辭庫筆數

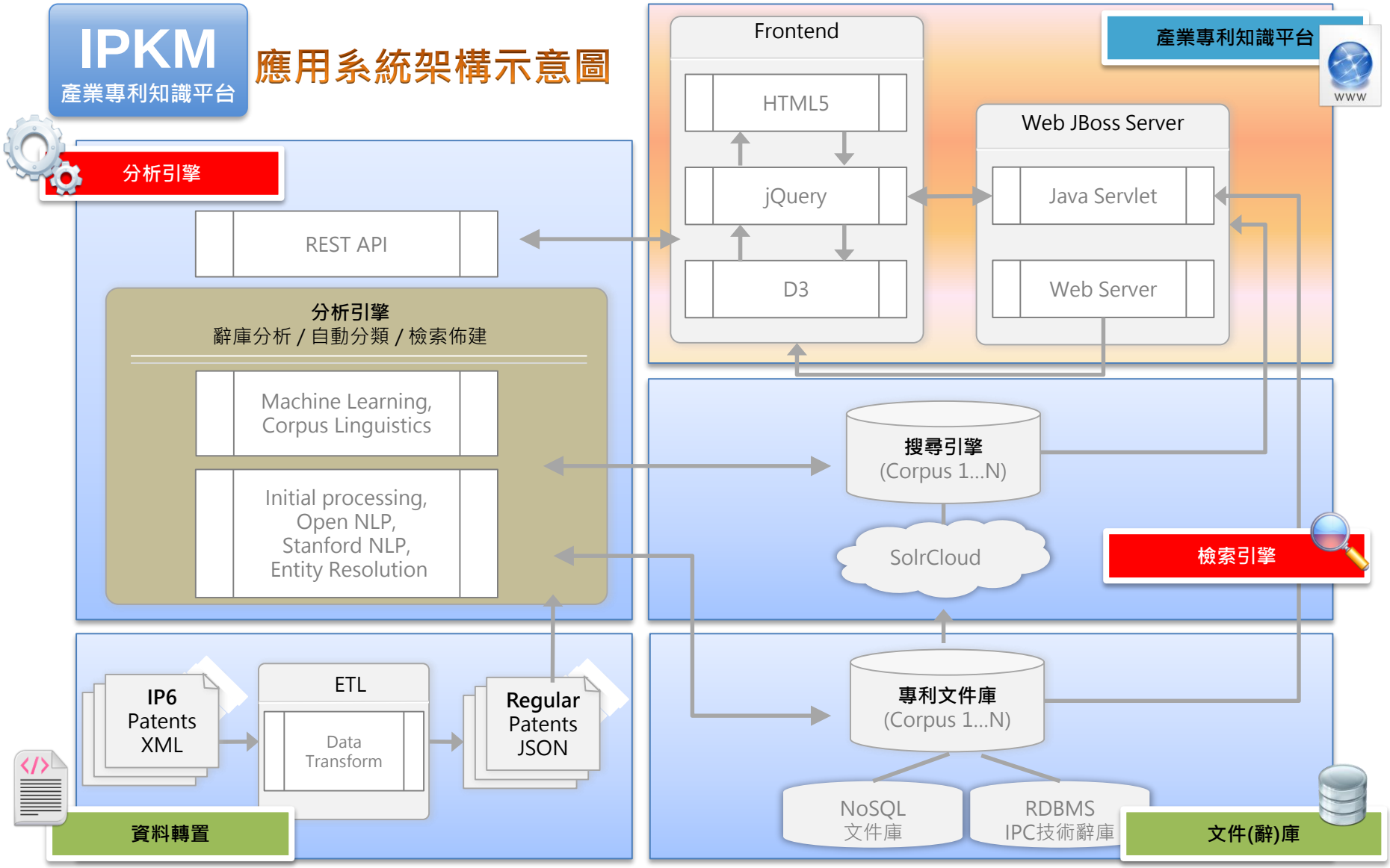
553,605	529,619	332,853	238,004	542,888

235,207

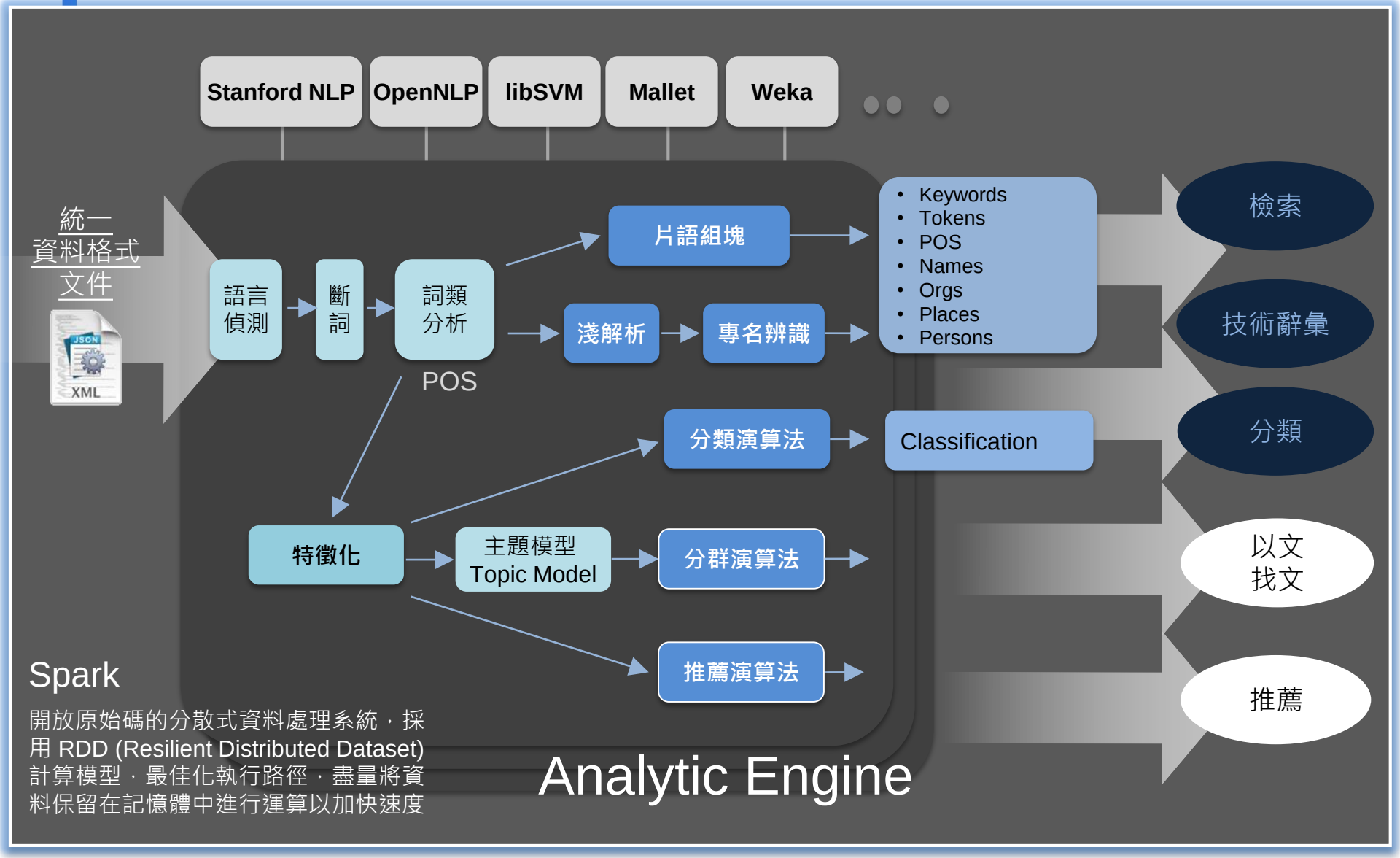
技術辭庫學習


辭庫分析
詞庫整理
分詞與詞性標註媒合
微調基礎詞庫
分詞/斷詞
Tokenization
Word_ysc
Scws
Max N-gram score
Rule-Based
Jieba
HMM/Trie
Stanford
Segmenter
CRF
FudanNLP
MMSEG
Jcseg
STOP
WORDS
Jieba
OpenNLP
Novel Topic
Model
詞性標注
Part-of-Speech tagging
Weirdness
Total TF-IDF
Novel Topic Model
PU
atr4s
關鍵詞抽取
Term Extraction

應用系統架構


IPKM
產業專利知識平台
應用系統架構示意圖
Frontend
HTML5
jQuery
D3
產業專利知識平台
WWW
Web JBoss Server
Java Servlet
Web Server
分析引擎
REST API
分析引擎
辭庫分析 / 自動分類 / 檢索佈建
Machine Learning,
Corpus Linguistics
Initial processing,
Open NLP,
Stanford NLP,
Entity Resolution
搜尋引擎
(Corpus 1...N)
SolrCloud
檢索引擎
IP6
Patents
XML
ETL
Data
Transform
Regular
Patents
JSON
專利文件庫
(Corpus 1...N)
NoSQL
文件庫
RDBMS
IPC技術辭庫
資料轉置
文件(辭)庫

分析引擎模型


Stanford NLP
OpenNLP
libSVM
Mallet
Weka
統一
資料格式
文件
JSON
XML
語言
偵測
斷
詞
詞類
分析
POS
片語組塊
淺解析
專名辨識
• Keywords
• Tokens
• POS
• Names
• Orgs
• Places
• Persons
分類演算法
Classification
特徵化
主題模型
Topic Model
分群演算法
推薦演算法
檢索
技術辭彙
分類
以文
找文
推薦
Spark
開放原始碼的分散式資料處理系統，採用 RDD (Resilient Distributed Dataset) 計算模型，最佳化執行路徑，盡量將資料保留在記憶體中進行運算以加快速度
Analytic Engine

Chapter

3

續章

關於 AI 的迷思
我們的現在與未來

AI 常見的迷思

「透過 google 就可以取得海量資料?」
「得到的這些資料，就可以提供機器學習使用?」
「AI 就是要省人工，如果資料得先整理不就開倒車?」

AI 需要一定量、能有效解析的樣本資料作為學習，隨之而來的前置作業會有一大堆資料「淨化」工作， 並不是只要有大量資料就可以。

在很多領域裏，它的效能並不會隨著訓練樣本數量的增加而逼近完美。能在訓練資料中抓出所有相關的輸入特徵尤為重要。

機器學習的用意在於模擬人類的分類和預測能力，不同知識領域資料內容特性多有不同，有其獨特性與技術專門性。

針對不同知識領域（資料內容），需要採取不同的對策，並經多次模擬後，才能決定策略及演算手法。

對 AI 的正確認知

「AI 改變了什麼？沒改變什麼？」
「分析引擎建立後，只會越來越聰明?」
「AI 這麼棒，為什麼不盡量把問題都丟給它解？」

AI 不會創造新的商業模式，但提高了效率。

機器學習不是魔術，你自己都不知道使用者的問題如何解決的時，資料科學是不會跑出結果的。

沒有「開箱即用」的 AI 應用；資料不同、解析/學習結果不同。沒有一種模型 (model) 可直接套用於所有知識領域。

不是「找到分析模型」就可結案，而不需再投入；相反地，學習成果的維持需要持續使用、評估、調整。

AI模擬人 ，人會出錯，機器也會出錯，人可作即時的修正就馬上的補正，但AI.....

適合 AI 的應用與問題?

「AI 會讓我的工作消失?」
「先決定明年是 AI 元年，再來看怎麼 AI ?」

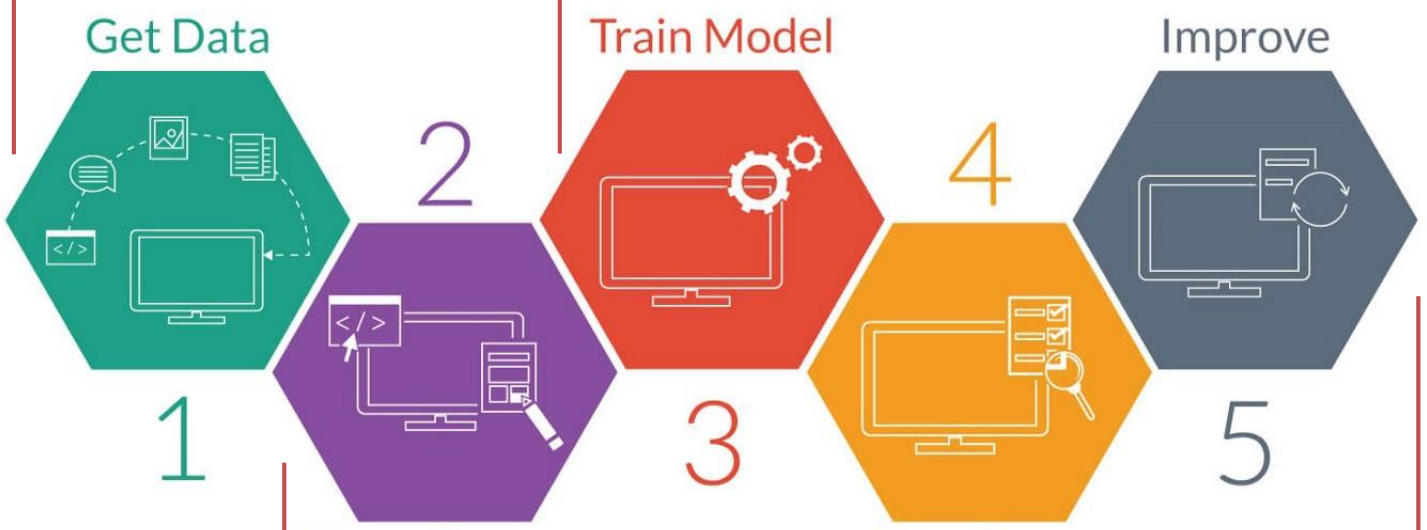
理想應用標的：「如果我們能知道『...』就好了」，
鎖定有價值的問題，而非「找一件事讓 AI 解決」，
價值愈高的問題，吸引愈多投入企圖與資源。

商業中需要複雜數學能力或 AI 解決的問題還很少。大部分事情透過 AI 解決，反而把事情複雜化。構建這個能力的工作量可能是直接解決它的難度的一百倍、一千倍。

設定易達成的短期目標；
不急於投入大量基礎建設；但要從主要目標分解出易於達成及驗證成果的階段性計畫。

適合用 AI 處理的問題特徵

Typical Machine Learning Workflow



複雜問題 人工難以建立完整邏輯規則

專屬策略 針對個別知識領域找出適切的策略

資料適足 分析所需資料夠新、夠乾淨

學習是個犯錯的過程，沒有終點須持續精進 允許犯錯

哪些任務最適合 AI？

標記界定明確的輸入和輸出，能以學習函式將其對應起來的任務

分類、預測；例如「分析一份入學申請來預測未來合格的可能性」

有明確的目標和度量標準、提供清晰反饋的任務

訓練資料明確度有多高或有明確界定的全系統表現尤為重要

例如「優化全內湖範圍內而不是民權東路交叉路口的交通流量」

不需要對於決策過程進行細緻解釋的任務

系統可提供建議，但不代替決策，人類本身就會做出不完美的決策!

例如「醫療診斷」

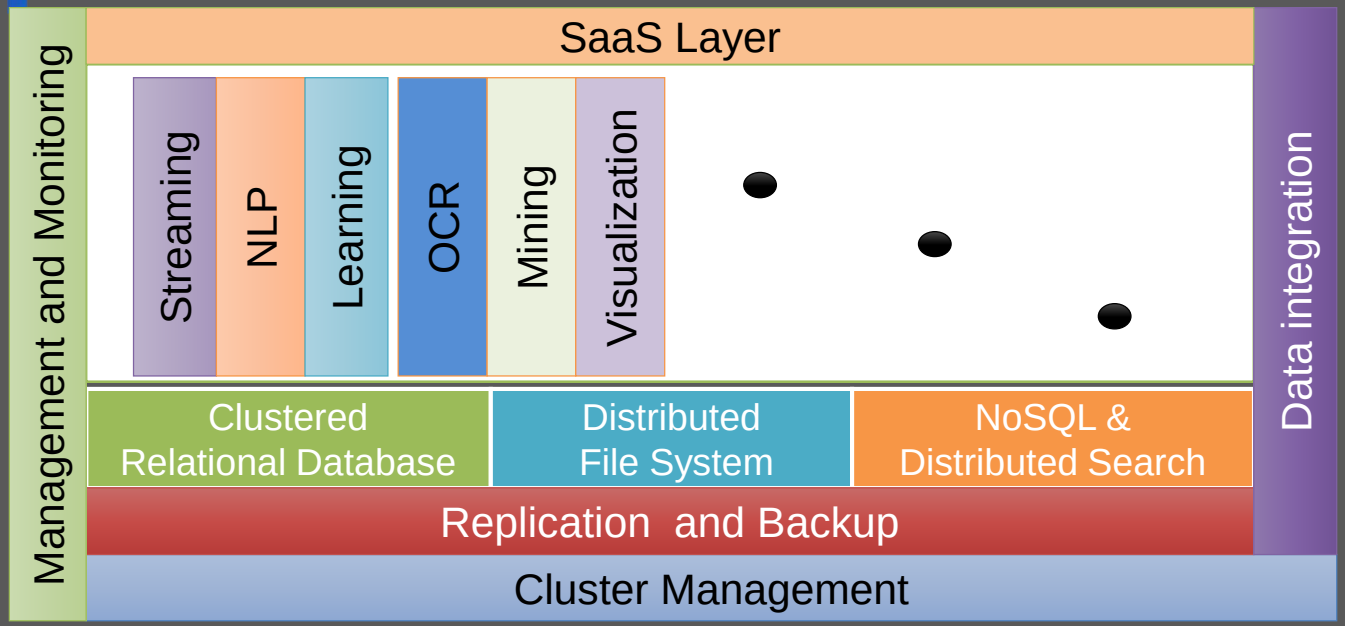
能夠容忍錯誤、不需要可證實的正確度或最優解決方案的任務

幾乎所有的機器學習演算法都是從統計學和概率上得出解決方案

因而，要把它們訓練到百分之百的準確度幾乎不可能

深度研發、創新研究

創新研究所 – NLP Topics



大型分散式系統

NE Recognition (專名辨識)

Parser (解析器)

Tokenization (斷詞)

Recommender Systems (推薦系統)

POS Tagging (詞類標示)

Summarization (總結)

Cluster/Classification (叢集/分類)

Emotion Detection (情緒偵測)

Keyword Extraction (關鍵字擷取)

自然語言機器學習
相關研究主軸

2016

2018

Thank You

感謝您的聆聽，敬請指教

GSS 叡揚資訊
Galaxy Software Services

國家產業創新獎
卓越中堅企業

GSS 叡揚資訊

Vital 雲端服務家族

GSS 技術部落格