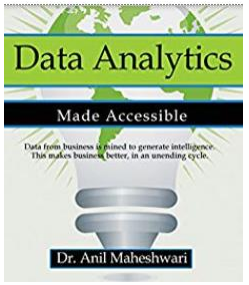
資料科學概論
(Introduction to Data Science)

白敦文 教授(Prof. Tun-Wen Pai)

1) Administrative Details
2) Most in-demand skills around the globe
3) How to explore yourself in the digital world?

Feb 21, 2023

Administrative Details



- Course： **資料科學概論**
- Course ID: 5902312
- Instructor： Dr. Tun-Wen Pai（白敦文 博士)
- Class hours : 202 : 宏裕科技大樓1223(Tue.) ；
  308/309 : 宏裕科技大樓234室(Wed.)
- Reference Textbooks :

1. Python Data Science Handbook by Jake VanderPlas
2. Data Analytics made accessible by Anil Maheshwari
3. An Introduction to Statistical Learning by Gareth James/Daniela Witten/Trevor Hastie/Robert Tibshirani (https://www.statlearning.com/)

- Pre-Requirements: none

- Syllabus:

本課程將初步介紹資料科學之基礎，帶領學程學生學習基本程式設計及工具應用，如何由手上之現有資料或網路資料進行發想及定義問題、學習使用適當演算法技術，由數據中探索資料之特殊樣式，提供最佳決策支援之依據。

- 第 1 週: 課程介紹/資料分析概觀/商業智慧 (0221/0222)
- 第 2 週: 老師出國訪問 (0228放假；0301兩節課於期中考週補上兩小時)
- 第 3 週: 統計基礎/資料倉儲 (0307/0308)
- 第 4 週: 資料探勘 (0314/0315)
- 第 5 週: 資料視覺化/決策樹分析(0321/0322)
- 第 6 週: 迴歸分析(0328/0329)
- 第 7 週: 清明節春假假期(0404/0405)
- 第 8 週: 第一次小考/類神經網路介紹(0411/0412)
- 第 9 週: 補上0301課程及期中考試 (0418/0419)
- 第 10 週: 群集分析(0425/0426)
- 第 11 週: 推薦系統(0502/0503)
- 第 12 週: 關聯規則探勘 (0509/0510)
- 第 13 週: 單純貝式分析(0516/0517)
- 第 14 週: 文字探勘技術(0523/0524)
- 第 15 週: 支持向量機 (0530/0531)
- 第 16 週: 第二次小考/網路探勘(0606/0607)
- 第 17 週: 社交網路數據分析(0613/0614)
- 第 18 週: 期末考試 (2022/06/20)

Grade: 小考(20%)、作業及出席(20%)、期中考(30%)、期末考(30%)

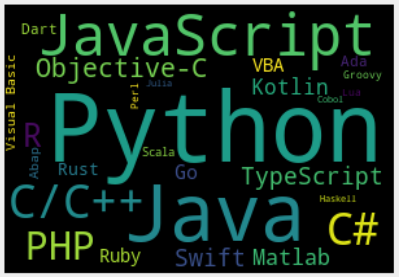
- TA : 吳陽生 wu0306109@gmail.com
- Office hours: 每週一、週二17:00-18:30 @ 1524 (Dr. Pai) / 週四16:00~18:00 @ 1621(吳陽生)
  (宏裕科技大樓Hung-Yu Research Tech. Building)
- Phone: 02-27712171 Ext. 4222(O) Ext. 4269(Lab.) Email: twp@ntut.edu.tw

PopularitY of Programming Language Index (PYPL)

The PYPL PopularitY of Programming Language Index is created by analyzing how often language tutorials are searched on Google.

The more a language tutorial is searched, the more popular the language is assumed to be. It is a leading indicator. The raw data comes from Google Trends.

If you believe in collective wisdom, the PYPL Popularity of Programming Language index can help you decide which language to study, or which one to use in a new software project.



Worldwide, Feb 2023 compared to a year ago:

Rank	Change	Language	Share	Trend
1		Python	27.7 %	-0.7 %
2		Java	16.79 %	-1.3 %
3		JavaScript	9.65 %	+0.6 %
4	↑	C#	6.97 %	-0.5 %
5	↓	C/C++	6.87 %	-0.6 %
6		PHP	5.23 %	-0.8 %
7		R	4.11 %	-0.1 %
8	↑↑	TypeScript	2.83 %	+0.8 %
9		Swift	2.27 %	+0.3 %
10	↓↓	Objective-C	2.25 %	-0.1 %

https://pypl.github.io/PYPL.html

In-demand Hard Skills

- ## Business Analysis

LinkedIn research reveals that every job professional must have some level of business analysis skills. This is regardless of the specific job role. The skill is important as it is all about finding solutions and answers to an organization's problems. It also facilitates those solutions.

- ## Analytical Reasoning

The demand for skills centered on data-driven decision-making is definitely on the rise. As data collection and data analysis are continuing to increase rapidly, most companies are looking for employees who can interpret the data and then take action on that data to facilitate company growth and expansion. This is why if you can prove that you are capable of analytic reasoning, you will be able to show that you are management material and also a cut above the rest.

- ## Affiliate Marketing

It is no secret that the rise of social media and digital marketing has managed to pave the way for affiliate marketing to become a new standard. Did you know that affiliate marketing helps leverage the great power of influencers and organizational partnerships to reach hyper-targeted audiences? Many individuals with affiliate marketing skills command top-dollar for these skills, so it is certainly in your interest to know more about them.

- ## Sales

You may know that selling is one of the most valuable hard skills. Note that the ability to sell is something that comes quite easily to some, while others find it challenging. However, it is worth noting that to stand out from the crowd, you have to show prospective employers that you can understand the sales funnel and the way it works. You can achieve additional credit if you can also demonstrate effective and efficient sales team management.

- ## Cloud Computing

Cloud computing is an important computer system resource that has on-demand availability. Companies use it for purposes of data storage and computing power. Note that if you can show your employer that you have this skill, you will have an updated skill set that employers will appreciate.

https://info.nyif.com/the-most-in-demand-hard-and-soft-skills-of-2021/

In-demand Soft Skills

- **Creativity**

Offering a unique and different perspective and thinking "outside the box" to examine a problem from various perspectives is important to generate new and improved ideas that drive the business forward. It is important to come up with excellent ideas, have brainstorming sessions with your team, and do something unique and original. Keep in mind that creativity is one of the most important and desired soft skills in the modern workplace; so, shore up your imagination and creativity to get yourself out there.

- **Communication**

You probably know that the ability to articulate (verbally and in writing) with colleagues, peers, and clients is considered the most desirable and highly valued skill by most employers. There is no doubt that communication is the key to success. So, become a better communicator by sharing your insights and collaborating with your team.

- **Adaptability**

Many people find it difficult to adapt to change, but in the modern workplace, it is a necessity. When change comes, you will have to adapt to it and also keep moving ahead. If you can't adapt to change and stay rigidly set on a single path or idea, it can detrimentally affect you. Note that adapting to change is a crucial skill that is in very high demand, particularly with the changes we all are experiencing with the COVID-19 health situation.

接受→ 改變 → 離開 are always the three-step keys for every job!!

30 High Paying Skills

- Coding And Software Enhancement
- Networking Development
- Soft Skills
- Algorithms Designer
- Cloud Computing
- UI Designer
- Online Frameworks
- Software Computing
- Analyst
- Data Science
- Public Relations
- Economical View
- Video Production
- Audio Production
- Sales (Ex. Affiliate, Offline)
- Digital Marketing & SEO
- Copywriting
- SEO
- Content Creation
- Blockchain
- AR/VR
- Cybersecurity
- Ethical Hacking
- Machine Learning
- Financial Management
- Trading
- Foreign Languages
- Management Consulting
- Art, Design, and Photography
- Sports/Fitness Coach/Nutritionist

https://startuptalky.com/future-top-skills/

IBM 大名鼎鼎的 Watson 也要被賣了，人類的 AI 夢該醒了？

熱門

人類豐滿的 AI 夢，正撞上冰冷的現實。1 月 19 日，據《華爾街日報》引用知情人士報導，IBM 考慮出售 Watson Health 業務，可能的方案包括賣給私募股權公司、醫療企業或與特殊目的收購公司（SPAC）合併。

https://technews.tw/2021/02/22/ibm-watson-ai/

Google、軟銀都陣亡過！盤點 AI 專案失敗的 4 大原因

- 企業都想做 AI，但實際上沒那麼簡單

- 根據《台灣人工智慧學校 AI Academy Taiwan》2019 年針對台灣各大產業 1,095 位業界校友的調查統計，成功導入 AI 人工智慧的台灣企業僅占 20%。放眼國際，許多全球知名企業的 AI 專案也慘遭滑鐵盧：

- Google 在泰國落地測試智慧醫療失敗，拖慢醫療流程；美國杜克大學發布的 PULSE 演算法誤將歐巴馬的頭像還原為白人，引發種族歧視爭議。

https://buzzorange.com/techorange/2021/02/04/ai-project-difficulties/

AI 專案難實際執行，問題出在哪？

1. AI 模型訓練過程中沒有加入實際場域的數據

無論是剛導入 AI 而產生數據處理需求的新手企業，還是已有 AI 專案經驗、為了 retrain 模型再度找有經驗的老手企業，都曾經在同一個地方卡關：AI 數據標註品質有做到位，但 AI 模型卻無法應用落地 。

→原因在於，客戶並未以「實際場景」的數據來進行 AI 模型訓練。

2. AI 數據標註原則定義不夠客觀

與企業工程師對接 AI 數據處理需求時，當詢問這批人臉辨識（Face Recognition）的 AI 數據標註的原則是什麼，常常會接到諸如此類的回答：「頭太小的話，就不要標註數據」。

一般人的邏輯覺得很合理的事情，對於機器學習（Machine Learning）來說卻是一大挑戰。 機器學習需要知道的是趨近「絕對客觀」的原則 ，一旦 AI 數據標註原則不夠客觀，AI 模型很容易隨著人的「主觀認定」來學習，當專案換了一位工程師，機器學習出來的效果可能也會跟著變 。

https://buzzorange.com/techorange/2021/02/04/ai-project-difficulties/

AI 專案難實際執行，問題出在哪？

3. AI 模型訓練（Model Training）沒有循序漸進

以肢體行為辨識（Posture Estimation）為例，Coco Dataset 從一開始只辨識人體 7 大主要關鍵點（Key Point），後來逐步發展成 25 點，甚至快 40 點，有些客戶會希望可以一次就標註 40 個關鍵點，直接拿去機器學習（Machine Learning）。

說起來，機器學習和教小孩很像，一下子給太多的特徵點（Feature Points）反而會「揠苗助長」，導致 AI 模型學到最後分不清楚自己到底在學習什麼。有些客戶，一開始想用難度較高的 Segmentation 方式讓模型學習人的行為，但是人的行為百百種、語意切割（Segmentation）的變異度也高，就比較難學得好。

4. 缺乏管理層的理解與支持

AI 熱潮讓許多企業趨之若鶩，然而 AI 要能夠順利落地，除了上述三項實務建議，企業管理層對於 AI 的認知和支持更是一大關鍵。許多台灣企業的 AI 數位轉型主導者，可能是傳統公司裡面有豐富資歷的 CTO 技術長或管理階層，對於 AI 人工智慧這個全新領域的概念，比較缺乏深度的理解，也沒有類似 AI 模型訓練和測試的相關經驗。

https://buzzorange.com/techorange/2021/02/04/ai-project-difficulties/

ChatGPT機器人，為什麼Google最害怕？

- OpenAI發佈聊天機器人ChatGPT後，兩個月吸引上億人使用。有望成為足以取代Google搜尋的智慧版搜尋系統
- OpenAI推出的聊天機器人─ChatGPT，席捲了社群媒體。許多人發現ChatGPT可以寫詩、編劇，還可以幫忙寫論文。雖然背後的科技已經存在多時，但這是OpenAI第一次將如此強大的語言生成系統開放給大眾使用。
- ChatGPT使用者開始相互較勁，試圖找出最有創意的指令。也有些人找到了ChatGPT的超務實使用方式。例如：工程師請它幫忙寫程式或揪錯。不過，ChatGPT最強勁的功能應該是，為我們提供比Google搜尋更理想的解答。
- Google可能會因此陷入財務災難。Google搜尋功能瀏覽數十億網頁、為內容加上索引，再依據相關程度列出搜尋結果。使用者看到的是一連串可供點選的網頁清單。ChatGPT則為焦躁的網路用戶提供更誘人的搜尋結果：依據自身的研究彙整網頁內容，**提供單一解答**。

GPT就厲害在它能夠基於無監督（unsupervised）的數據，建立起通用的語言模型，接著再針對有監督（supervised）的特定任務逐步微調，如此一來，便成功通過當時科學家遇上的瓶頸。2018年，OpenAI發表論文，主要在討論如利用所謂「通用預訓練（generative pre-training，簡稱為GPT）」改善模型對於語言的理解，第一代GPT模型預訓練的數據量達到約5GB，使用到的參數接近1.2億。隔年（2019）OpenAI發表GPT-2，預訓練的數據量暴漲，直接衝高到40GB，使用到的參數更是來到15億。OpenAI並沒有停下腳步，在2020年又釋出了GPT-3，這次的數據量翻了千倍，達到45TB，而參數量也升級到1,750億。

https://www.cw.com.tw/article/5123990

Gartner headquarters in Stamford

Type	Public company
Traded as	NYSE: IT S&P 500 component
ISIN	US3666511072
Industry	Research & Advisory Services
Founded	1979; 41 years ago by Gideon Gartner in Stamford, Connecticut

Headquarters	Stamford, Connecticut[*citation needed*]
Key people	Gene Hall (CEO) Craig Safian (CFO) Mike Diliberto (CIO)
Products	Research Consulting Conferences
Revenue	▲US$4.25 billion (2019)[1] ▲US$4.1 billion (2018)[2]
Operating income	▲US$(4.3) billion (2017)[2]
Net income	▼US$3.3 billion (2017)[2]
Total assets	▲US$7.3 billion (2017)[2]
Total equity	▼US$983 million (2017)[2]
Number of employees	15,173 (2018)

Nearly 17,000

associates

We're a people business, powered by independent analysis and unmatched depth and breadth of expertise.

Meet Our Leaders

Check Out Career Opportunities

100

countries

We're a global company, serving more than 14,000 client enterprises around the world.

See Our Locations

$4.2B

revenue

We're a member of the S&P 500, with clients in 77% of the Global 500 and organizations of every size.

Explore Investor Information

View Our Corporate Fact Sheet

Gartner：2023年十大戰略技術趨勢

- Gartner 揭示 2023 年十大戰略技術趨勢，聚焦在三個主題上：**最佳化（optimize）、規模化（scale）與開拓新領域（pioneer）**

- 優化 - 優化企業的韌性、營運、可信度。
- 擴展 - 擴展垂直應用、交付方式、價值實現。
- 開拓 - 開拓生態圈應用、高韌性智慧、全新商業領域。

2023 年，僅僅是科技的應用還不足夠。 Gartner 認為這些主題受到**環境、社會和治理（ESG）的影響**，需要達成可持續性的共同責任。為了我們的子孫後代，企業每進行一項技術投資，都需要抵消它所產生的環境影響，來實現『默認可持續性』這一目標。

https://www.cio.com.tw/gartner-research-report-2023-strategic-technology-trends/ /

1. **數位免疫系統（Digital Immune System）** - 76% 負責數位產品的團隊現在還需要對營收負責，因此企業正在尋找新的實踐和方法，使其團隊能夠在實現高商業價值的同時，降低風險和提高客戶滿意度。因此，**數位應用的可靠度**成為企業的關鍵訴求，而數位免疫系統能夠滿足企業的這一個需求。

數位免疫系統是一連串的技術與手法，透過可觀測性（Observability）、AI 增強型測試（AI-Augmented Testing）、混沌工程（Chaos Engineering）、自我修復（Autoremediation）、站點可靠性工程（Site Reliability Engineering）和應用供應鏈安全（Apps Supply Chain Security）等技術，來大幅提高雲原生系統的韌性和穩定性。

2. **可觀測性應用（Applied Observability）** 企業採取任何行動時，都會產生具備數位化特徵的可觀測數據，例如日誌、使用軌跡、API 調用、停留時間、下載和文件傳輸等。可觀測性應用以一種高度統籌和整合的方式將這些可觀測的特徵數據進行處理，藉由“讓數據說話”，來創造出一個決策迴圈，讓企業快速地做決策來提升營運效率。 Gartner 認為可觀測性應用可以讓企業利用他們的數據特徵來獲得競爭優勢。 它能夠讓企業在正確的時間藉由觀測數據來做決策，是一種強大的工具。 如果能夠在戰略中予以規劃並成功執行，可觀測性應用將成為數據驅動型決策的最強大來源。

https://www.cio.com.tw/gartner-research-report-2023-strategic-technology-trends/

3. **AI 信任、風險和安全管理（Trust, Risk and Security Management）** 許多企業未做好管理 AI 風險的充分準備。 Gartner 在美國、英國和德國開展的一項調查顯示，41% 的企業曾經歷過 AI 隱私洩露或安全事件。但該調查也發現積極管理 AI 風險、隱私和安全的企業在 AI 專案中取得了更好的成果。 與未積極管理這些功能的企業的 AI 專案相比，在這些企業中有更多的 AI 項目能夠從概念驗證階段進入到生產階段並實現更大的業務價值。

4. **產業雲平台（Industry Cloud Platforms）** 產業雲平台是支援特定產業（金融、製造、醫療等等）的公有雲平台。這些平台提供**軟體即服務（SaaS）、平台即服務（PaaS）和基礎設施即服務（IaaS）**等服務，提供產業所需的應用場景的模組化能力。 企業可以利用產業雲平台的服務，更快速的搭建基礎模組和實現數位業務，提升敏捷性和推動創新。Gartner 預測，到 2027 年，超過 50% 的企業將使用產業雲平台來加速他們的業務專案。

5. **平台工程（Platform Engineering）** 平台工程是一套機制和架構，將 Infrastructure 服務化，讓軟體開發團隊在軟體交付作業時能夠自助式的使用 Infrastructure 服務。過去開發團隊和 Infrastructure 壁壘分明，開發團隊求快但是 Infrastructure 團隊求穩，因此時常有摩擦。 許多企業的 IT 部門均已在開發部門和 Infrastructure 部門之外，成立平台工程的部門，負責平台工程。**平台工程能夠優化開發者體驗並加快產品團隊為客戶創造價值的速度。** Gartner 預測，到 2026 年，80% 的企業將建立平台工程團隊，其中 75% 將包含開發者自助服務的 Portal。

6. **無線技術價值實現（Wireless Value Realization）** 無線的技術（WIFI, 5G, Bluetooth, RFID, ...）已經發展多年並已經非常成熟。但是各個技術的適用場景不同，不可能有任何無線技術能夠佔據主導地位。企業應該訂定策略，善用各種無線解決方案來滿足辦公室、行動裝置服務、低功耗服務以及無線電連接等各種場景的需求，為企業帶來價值。 Gartner 預測，到 2025 年，60% 的企業將同時使用五種以上的無線技術。 網路的功能將不再僅限於純粹的連接，它們將使用內置的分析功能提供洞察，而某些新世代的低功耗系統（例如 Wiliot Pixel）甚至能夠從網路電波中獲取能量即可運作，不需任何外接電源。 這意味著無線的技術將能夠實現創新的業務模式，創造商業價值。

7. 超級應用（Superapps） 超級應用是一個集應用、平台和生態系統功能於一身的行動應用。 它不僅有自己的功能，而且還為第三方提供了一個開發和發佈微應用的平台。例如大陸的微信就是一個最顯著的成功案例，微信除了是一個傳遞訊息的行動應用，也是一個生態系統和平台，有數十萬個第三方開發的微應用能夠安裝在微信內，為使用者提供食衣住行的各項服務。Gartner 預測，到 2027 年，全球 50% 以上的人口將成為多個超級應用的日活躍使用者。 Gartner 建議服務業開始評估超級應用的可行性，利用超級應用來搭建生態系統，取得先機並達成數位創新。

8. **自適應 AI（Adaptive AI）** 自適應 AI 系統透過不斷反覆訓練模型，自動使用新的數據進行學習，來迅速適應在最初開發過程中無法預見的現實世界變化。 這些系統根據即時反饋，來動態調整它們的學習和目標，因此能夠適應外部環境快速變化，及企業目標的不斷變化。自適應 AI 能夠避免 AI 模型的偏移，而造成業務的負面影響。 業界最著名的案例，是美國房地產公司 Zillow 的慘痛教訓。Zillow 推出了 Zillow Offers 的線上服務，用 AI 來對房地產做線上即時報價。這項服務在初期很成功，對公司帶來了許多營收。但是由於疫情和市場變化，當初的 AI 模型開始偏移，但是 Zillow 渾然不知，繼續用遠高於市場行情的價格報價和購入房地產。最終的結果是 Zillow 虧損數十億美金，並終止 Zillow Offers 的業務。自適應 AI 能夠確保 AI 的持續精準性，避免類似問題。

9. **元宇宙（Metaverse）** Gartner 將元宇宙定義為一個由透過虛擬技術，將實體和數位現實融合而成的虛擬共享空間。 這個空間具有持久性，能夠提供增強沉浸式體驗。 Gartner 預計完整的元宇宙將獨立於設備並且不屬於任何一家廠商。 它將產生一個由數位貨幣和非同質化通證（NFT）推動的虛擬經濟體系。 Gartner 預測，到 2027 年，全球超過 40% 的大型企業將在基於元宇宙的專案中使用 Web3、增強現實（AR）雲和數位孿生的組合來增加收入。 元宇宙仍需許多年才會趨於成熟（Gartner 的 Hype Cycle 認為需要十年以上）。即便沒有立即商機，企業仍需開始研究元宇宙的發展，並思考未來在元宇宙的商業模式。

10. **可持續性（Sustainability）** 可持續性貫穿 2023 年的所有戰略科技趨勢。 在 Gartner 最近的一項調查中，執行長們表示環境和社會變化已成為投資者的三大優先事項之一，僅次於利潤和收入。 這意味著為了實現可持續性目標，高階主管必須加大 ESG 相關技術和服務的投資力道。 為此，企業需要新的可持續技術框架來提高 IT 服務的能源和材料效率，透過可追溯性、分析、可再生能源和人工智慧（AI）等技術實現企業的可持續發展，同時還要部署幫助客戶實現其可持續性目標的 IT 解決方案。