

生成式AI於邊緣運算的應用發展趨勢

施柏榮

產業顧問兼副主任

產業情報研究所

財團法人資訊工業策進會

2024.04.18



簡報大綱

- 生成式AI於邊緣運算的發展背景
- 生成式AI於邊緣運算的應用案例
- 結論：未來方向觀測與布局建議
- 附件

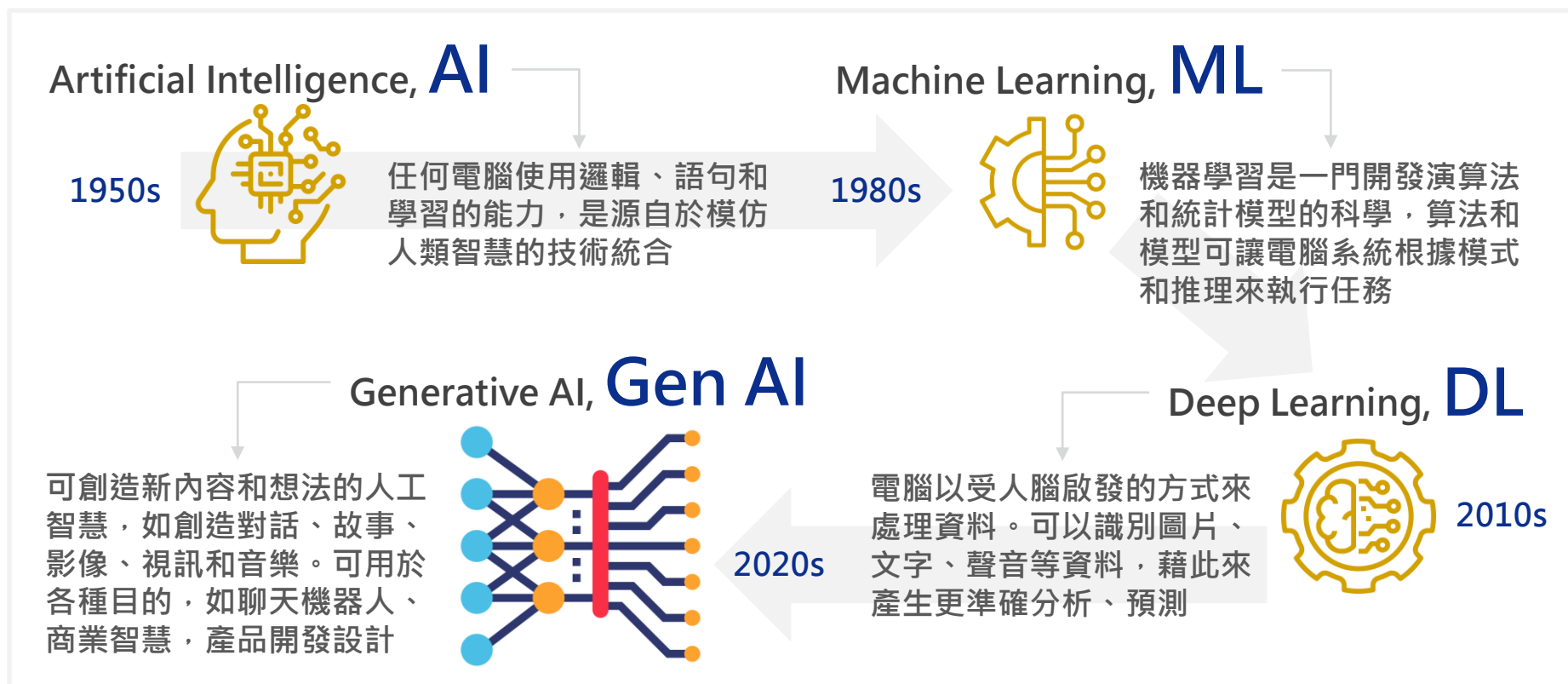


生成式AI於邊緣運算的發展背景





AI科技演進快速並滲透人類生活



備註：按AWS對於生成式AI的觀察，生成式AI具有四項基本優點：加快研究速度、提升客戶體驗、找出最佳化業務程序、提升員工生產力

資料來源：Shanaka Baduge et al., (2022)、Wenwen Li & Hsu (2022)、AWS (2024)、MIC整理，2024年4月

- AI發展可追溯於1950年代，近年AI技術大幅成長與深度學習（DL）的落地應用有關
- 2020年代初期發展出的「生成式AI」受大量關注，原因在於生成式AI功能並非「辨識」（recognition）物件，而是可藉多種數據、語料「生成」（Generative）新的內容



AI改變人類感官與知識傳遞模式



馬素·麥克魯漢

人類知識傳播的歷史階段



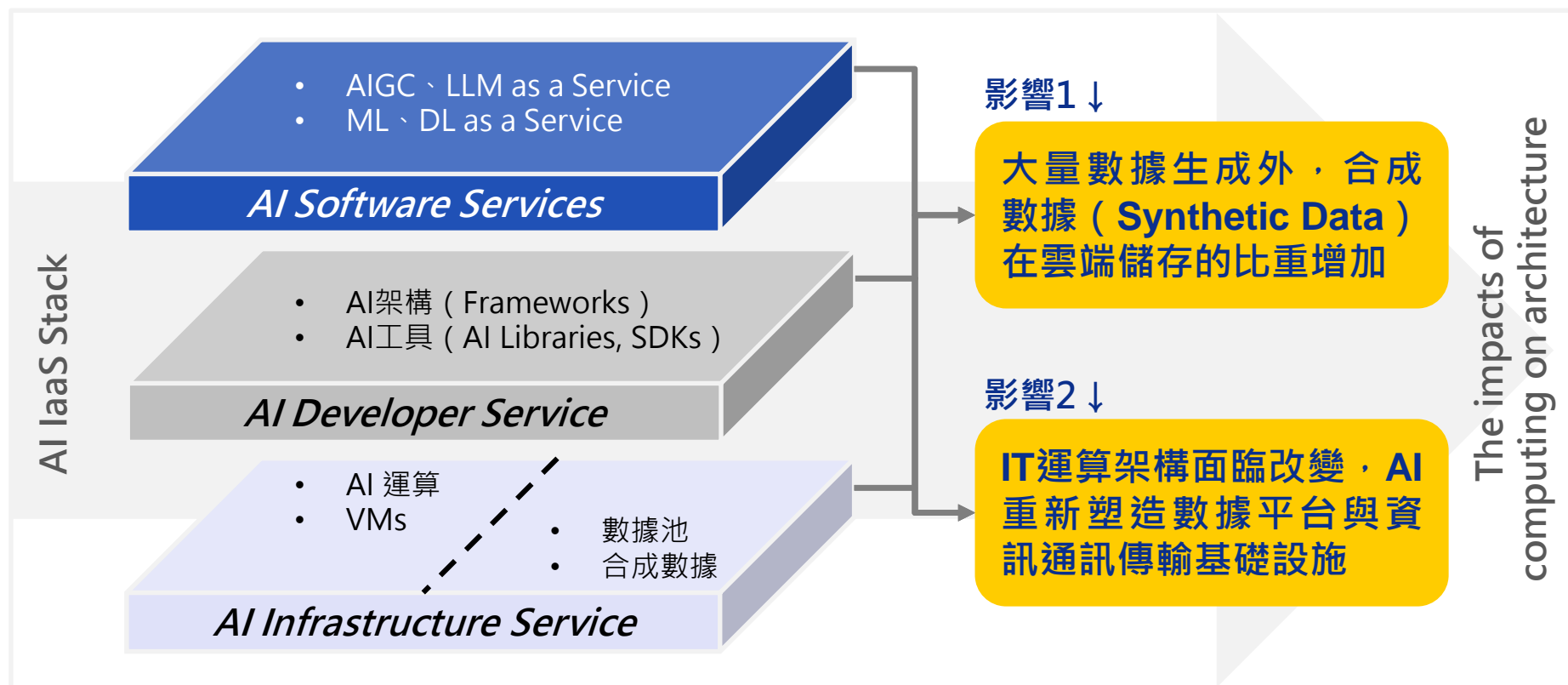
備註：麥克魯漢《古騰堡星系：印刷文明的誕生》指出『媒介及資訊 (知識)』，媒介、資訊、知識、感官四者相互連結

資料來源：Herbert Marshall McLuhan (1962)、Marek Sokolowski & Regina Ershova (2024)、MIC整理，2024年4月

- 麥克魯漢《古騰堡星系：印刷文明的誕生》揭示科技對於人類資訊、知識的影響
- 人類已經過四個歷史階段，而AI的出現則預示新的歷史階段，此一階段，不僅是傳遞的出現變化，資訊與知識生成、生產方式也產生巨變，大量資訊被壓縮在數位「地球村」



AI已重新塑造數據與IT基礎設施



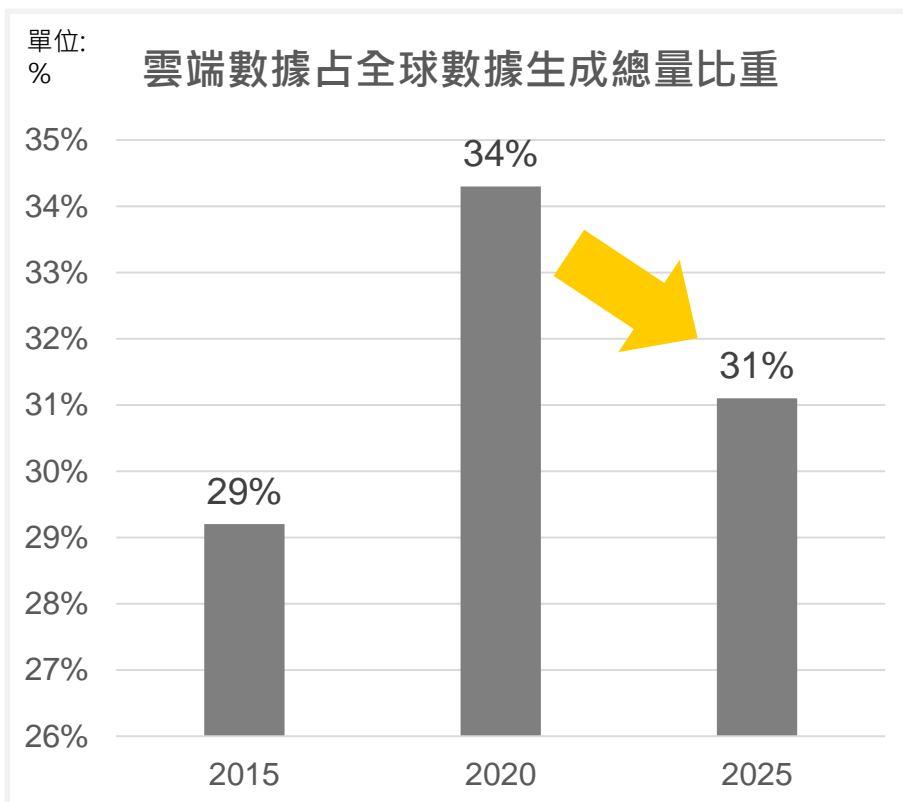
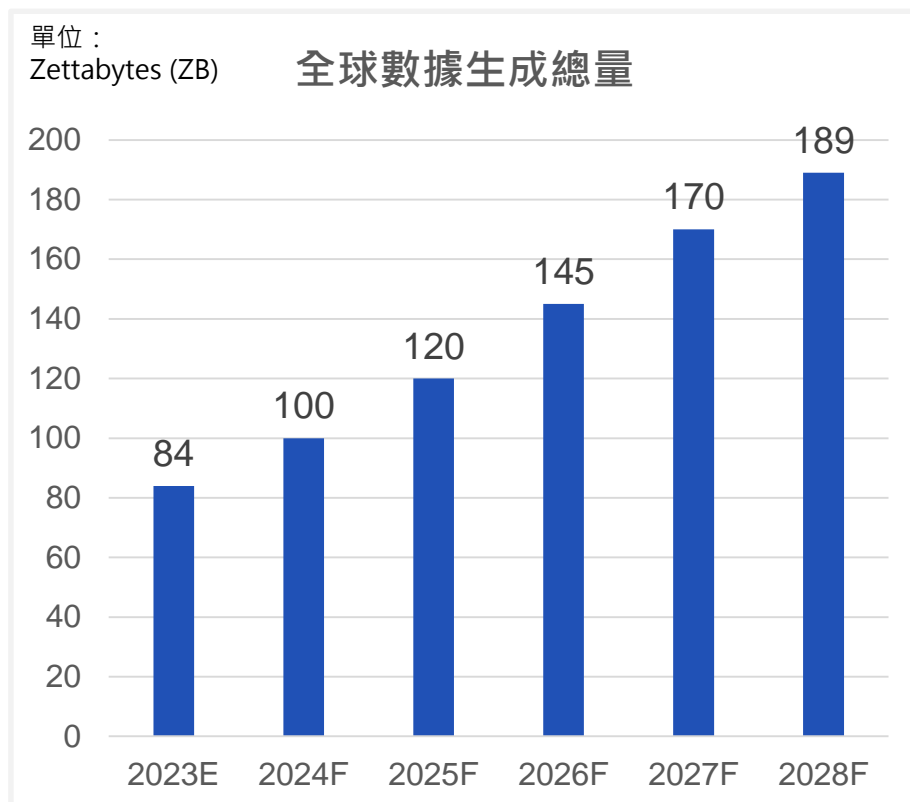
備註：2015年之後，IEEE等平台或相關研究年會，並以AI+IaaS為主題，探討AI對於IT、雲端運算架構可能造成的影響

資料來源：Sebastian Lins et al. (2021)、Alvaro Figueira & Bruno Vaz (2022)、DataCamp, Inc. (2024)、MIC整理，2024年4月

- AI技術的落地與應用，會對於既有的數據與IT基礎設施造成影響，或重塑IaaS的架構
- 第一，AI的應用將產生更大量的數據處理需求，合成數據的比重也將持續增加；第二，AI除了對於晶片運算效率的需求之外，IaaS設計架構也將依循AI運行的模式產生變化



數據處理位置從雲端漸移至地端



備註：Covid-19驅動更多智慧化應用，數據生成總量快速成長，未曾放緩
資料來源：Redgate (2024)，MIC整理，2024年4月

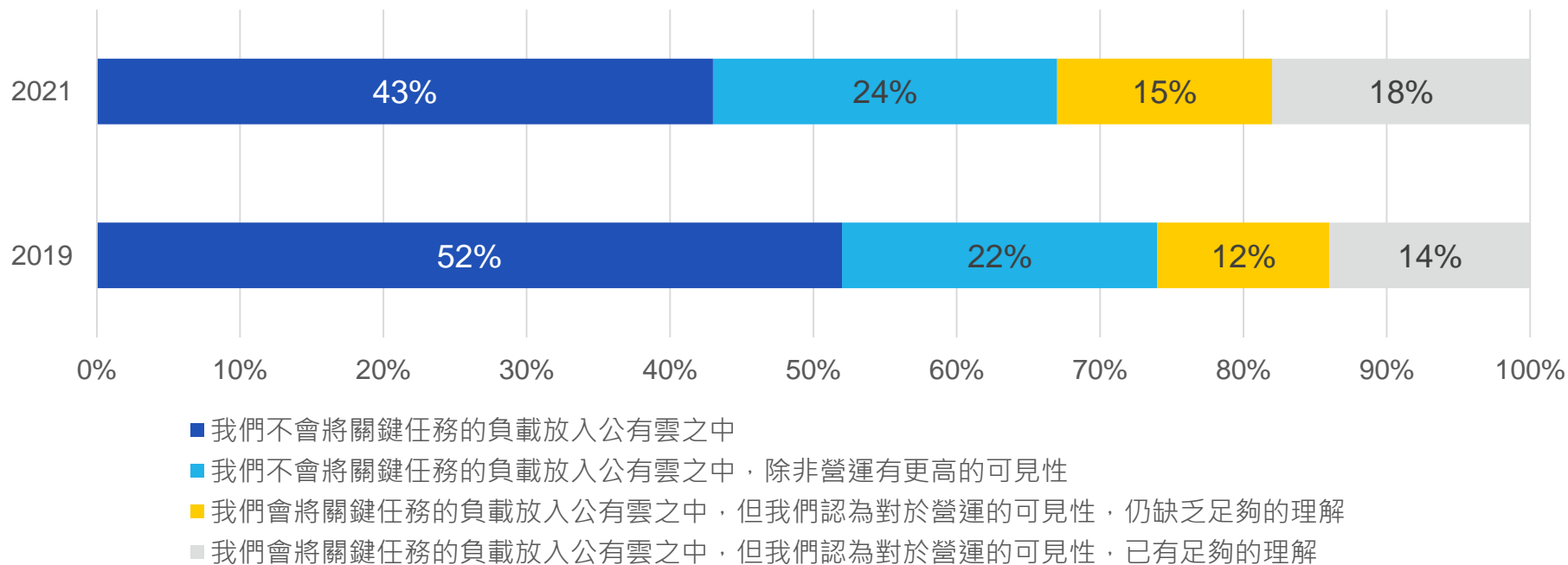
資料來源：Seagate & IDC (2018)、Cisco (2023)，MIC整理，2024年4月

- 全球數據生成總量預期將持續成長，2028年將全球數據生成總量將達到189ZB
- 進一步觀察雲端數據於全球數據生成總量的比重，則可以發現到2020年至2025年出現下降的態勢，這意味著大量的數據生成，邊緣、地端在數據處理的重要性相對提升



信任與可見性仍是關鍵影響因子

關於將組織內部的關鍵任務，放入公有雲環境的思考



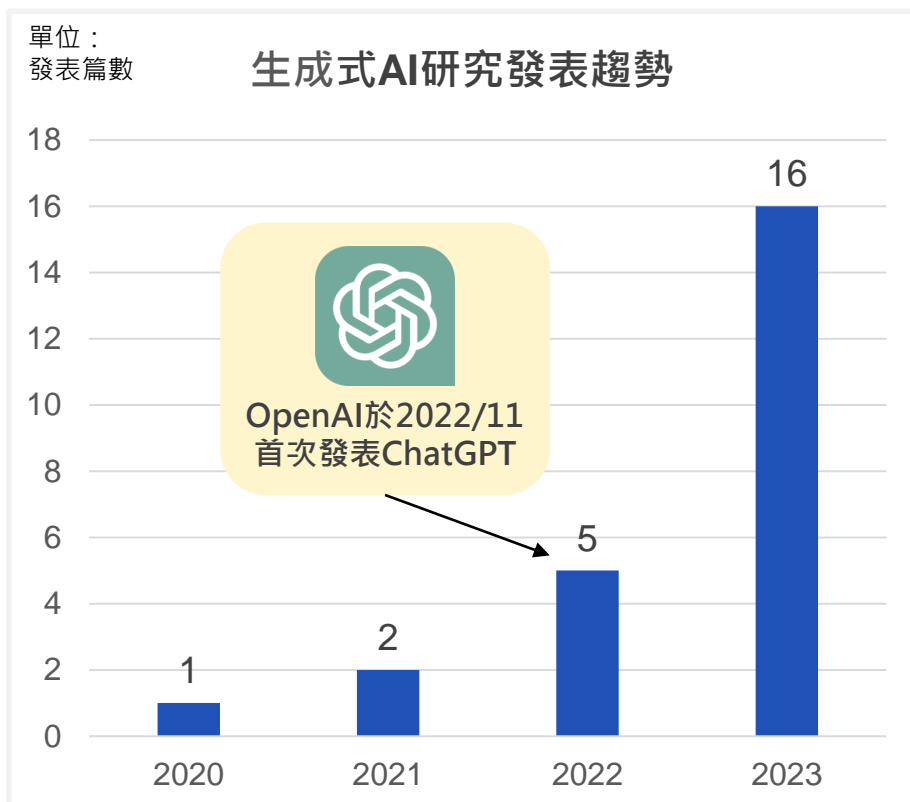
備註：關鍵任務 (Mission-critical Workloads) 意指對於組織、企業來說屬於關鍵生產力要素的IP、方法、技術、製程、流程等知識資產

資料來源：Rhonda Ascierio (2022) · MIC整理，2024年4月

- 雲端運算 (Cloud Computing) 已為當代國家、產業倚賴的重要關鍵基礎設施
- 進一步觀察組織將「關鍵任務」 (Mission-critical Workloads) 放置公有雲的意願，經Covid 19已有提升，但「信任」與可見性 (控制權) 仍是採用雲端與否的考量



生成式AI將帶動新的數據增長



備註：IEEE資料庫關鍵字檢索，包括期刊、展會等發表研究與文件
資料來源：IEEE (2024)，MIC整理，2024年4月

全球生成式AI主要技術提供商

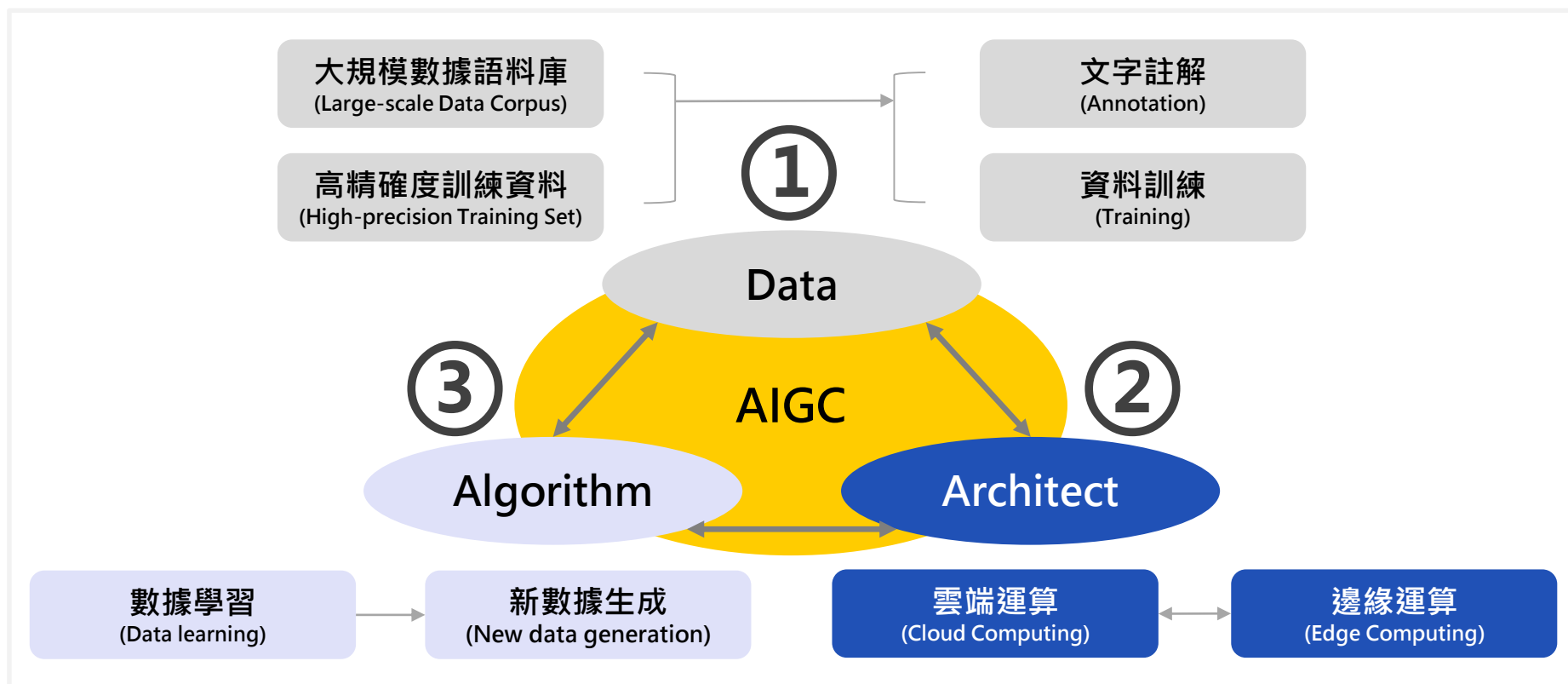
推出公司	產品名稱	應用與服務
OpenAI	ChatGPT	文本生成、聊天機器人
Google	LaMDA	問答和聊天機器人
NVIDIA	StyleGAN	圖像生成與設計
Microsoft	Turing-NLG	摘要、翻譯和問答
DeepMind	DVD-GAN	影片生成
Stability AI	Stable D.	文字到圖像
Eleuther AI	GPT-Neo	文本生成
Baidu	ERNIE	問答和聊天機器人

資料來源：Jiayang Wu et al. (2023)，MIC整理，2024年4月

- 生成式AI隨著OpenAI之ChatGPT發布之後，在2022、2023年受到市場大量的關注
- 不僅OpenAI，雲端服務提供商如Microsoft、Google等也投入相關產品，新興的應用服務如文本生成、圖片與影片生成，勢必帶動新一波數據處理需求的擴張



生成式AI同樣將重塑IT基礎設施



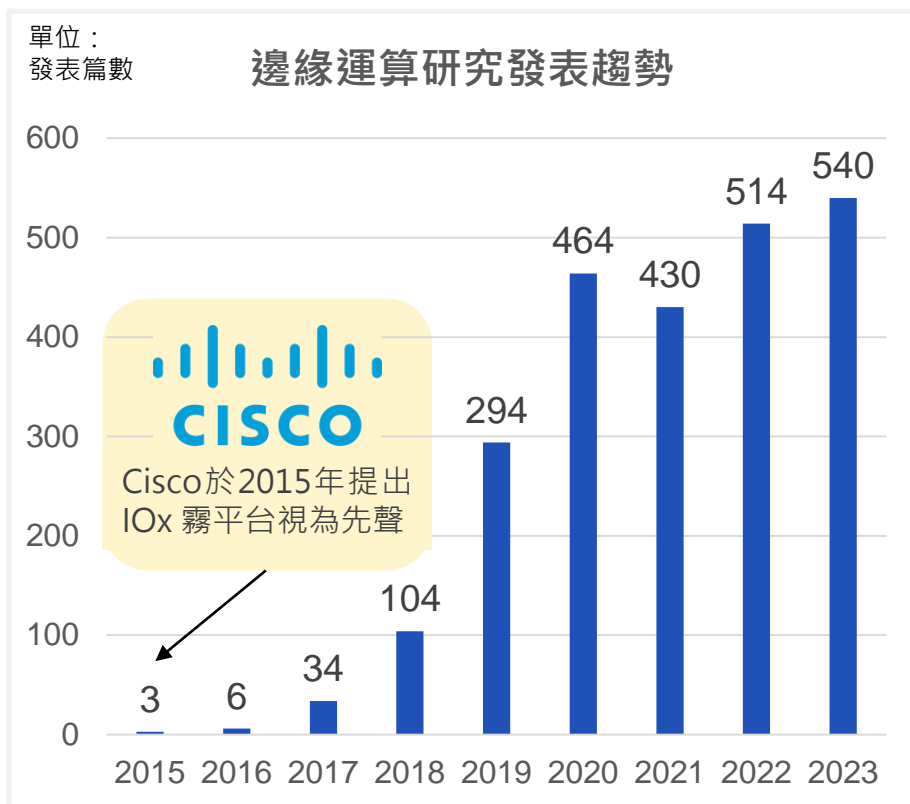
備註：AIGC (Artificial Intelligence Generated Content) 主要利用ML及DL演算法與模型，採用適宜的運算架構，分析大量的數據資料庫

資料來源：Jiayang Wu et al., (2023) · MIC整理 · 2024年4月

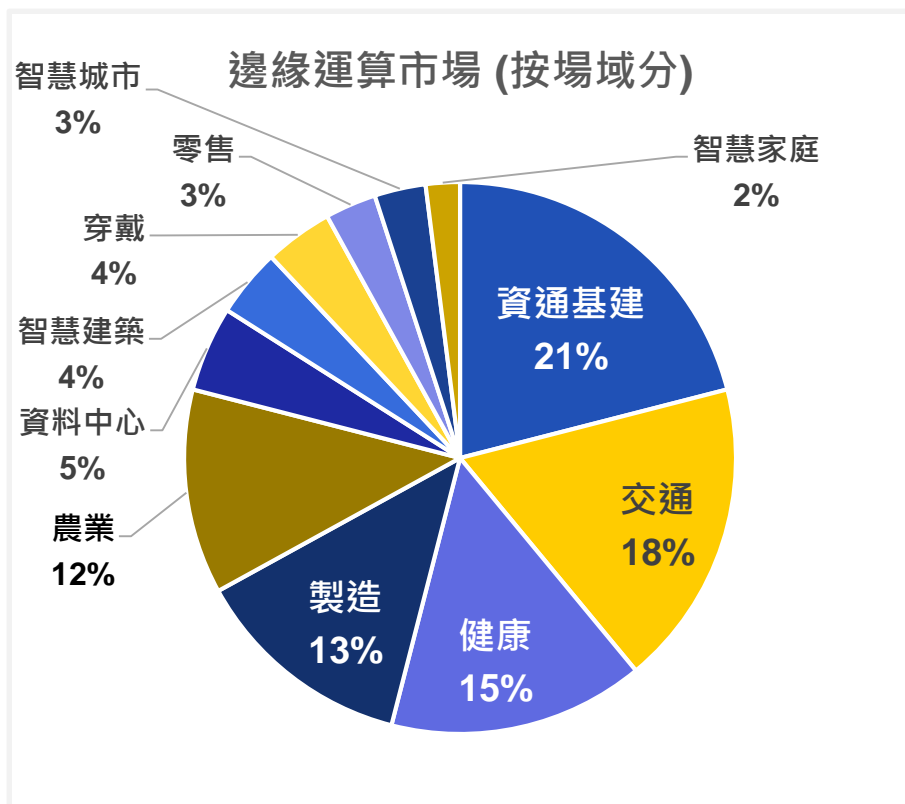
- AIGC除晶片、硬體基礎設備之外，亦包含「數據」、「演算法」、「架構」三個技術實現元素的組成；「架構」除了雲端運算之外，也包含「邊緣運算」運算架構
- 三個技術實現元素也顯示出 - **AIGC也將重新塑造IT基礎設施、運算架構的型態**



邊緣為生成式AI落地要件 (1/2)



備註：IEEE 資料庫關鍵字檢索，包括期刊、展會等發表研究與文件
資料來源：IEEE (2024)，MIC 整理，2024 年 4 月



備註：資通基建包含 5G 小型基地台等通訊基礎設施
資料來源：OpenFog (2018)、IIC (2023)，MIC 整理，2024 年 4 月

- 邊緣運算 (Edge Computing) 架構，最早可追溯至 2010 年代中心，近年仍持續在不同應用領域中驗證，包括資通基建 (含通訊)、交通、健康、製造皆為主要的落地應用
- AIGC 與邊緣運算 (Edge AIGC) 的討論在 2023 年底，已在 IEEE 等技術討論逐漸增加



邊緣為生成式AI落地要件 (2/2)



組織、企業採用邊緣運算誘因

數據處理層

降低數據壅塞

採用分層數據處理原則，可減少傳輸至雲端處理的數據量，同時也降低雲端產生數據壅塞情況

資訊數位信任

①偏屬於封閉系統，可應用資訊安全較為敏感應用；②使用者對數據的掌握、可見性程度較高

運算資源層

網路延遲率

採用分散式的架構，可於近終端進行資料的擷取、分析、儲存、過濾，適合延遲率敏感的應用

服務不間斷性

邊緣形成的系統可成為一個區域運算架構；在缺乏雲端或是間歇連線情況，提供不間斷服務

資源設備限制

可以在未與雲端網路連結情況下，執行某些特定的雲端功能；同時也可以更彈性化進行調整

備註：雲端運算、邊緣運算哪一種運算架構的資訊安全表現較高，向來IT、OT服務提供者略有爭議，因此，改以資訊「數位信任」來進行表述
資料來源：IIC (2023)、Inside Industry Association (2024)、Jiayang Wu et al., (2023)、MIC整理，2024年4月

- 觀察邊緣運算採用因素，延遲率、服務不間斷性（缺乏雲端支援，仍可提供不間斷運算服務）、解決資源設備限制、解決數據壅塞、資訊數位信任皆為主要採用的主要因素
- AIGC採用邊緣運算主要的誘因，則與降低數據壅塞、「資訊數位信任」有關

生成式AI於邊緣運算的應用案例





企業使用生成式AI延伸眾多議論

Bloomberg

Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak

- Employees accidentally leaked sensitive data via ChatGPT
- Company preparing own internal artificial intelligence tools



By [Mark Gurman](#)



2023年5月2日 at 上午8:48 [GMT+8]



Updated on 2023年5月2日 at 下午1:54 [GMT+8]

Save

THE WALL STREET JOURNAL.

Apple Restricts Employee Use of ChatGPT, Joining Other Companies Wary of Leaks

The iPhone maker is concerned workers could release confidential data as it develops its own similar technology

By [Aaron Tilley](#) [Follow](#) and [Miles Kruppa](#) [Follow](#)

Updated May 18, 2023 7:35 pm ET

← Bloomberg、WSJ媒體報導部分企業禁止員工，將敏感資料與程式碼上傳到人工智慧聊天機器人程式

reddit 平台，大量出現工作組織禁止員工用人工智慧聊天機器人程式工作的討論帖



r/consulting • 9 mo. ago
r_hrubby

My company banned ChatGPT 🤖



reddit ...

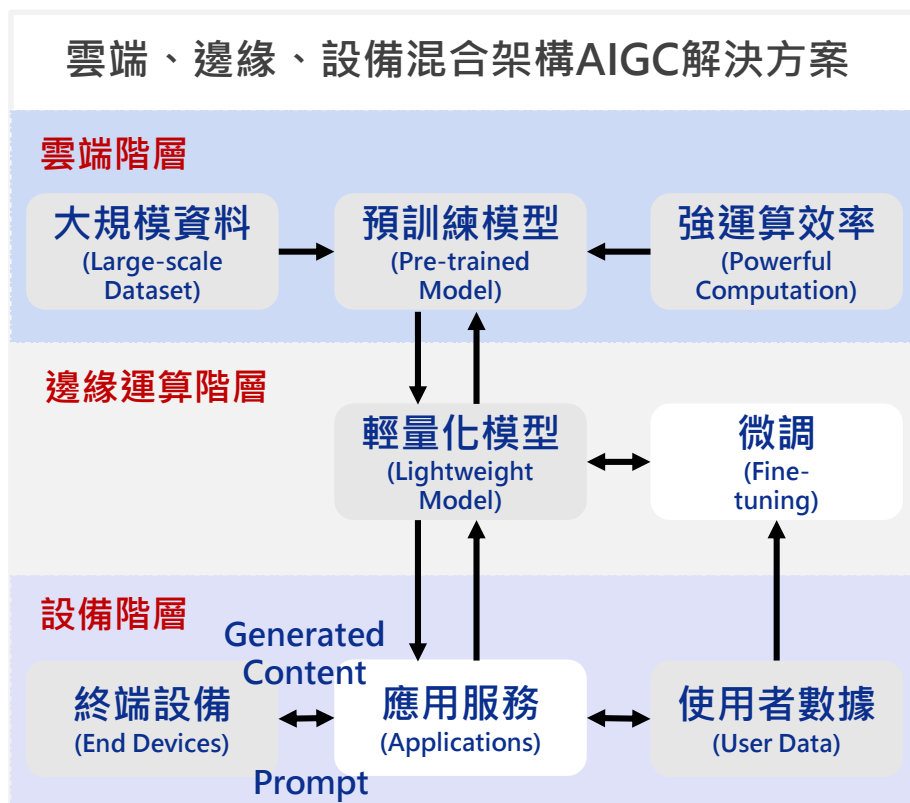
Hi all, I am new here, literally signed up to write this post. I work at a Tier 2 strategy consultancy located on the East Coast. I used ChatGPT a lot but now following announcements from Accenture and PwC my firm decided to issue a company-wide ban because of data security concerns... I can't access OpenAI's website anymore. I wonder if any of you are in similar shoes... Do you see use any secure alternatives?

備註：上述新聞資訊來源，多數來自於媒體，多數被指涉的企業並未直接證實

資料來源：Bloomberg (2023)、The Wall Street Journal (2023)、reddit, r_hrubby, joetaylorland (2023, 2024)、MIC整理，2024年4月



Edge AIGC運算架構情境 (1/2)



備註：上述為一般性的AIGC分層情境，並不完全適用於各種AIGC情境
資料來源：Y. C. Wang et al., (2024) · MIC整理 · 2024年4月

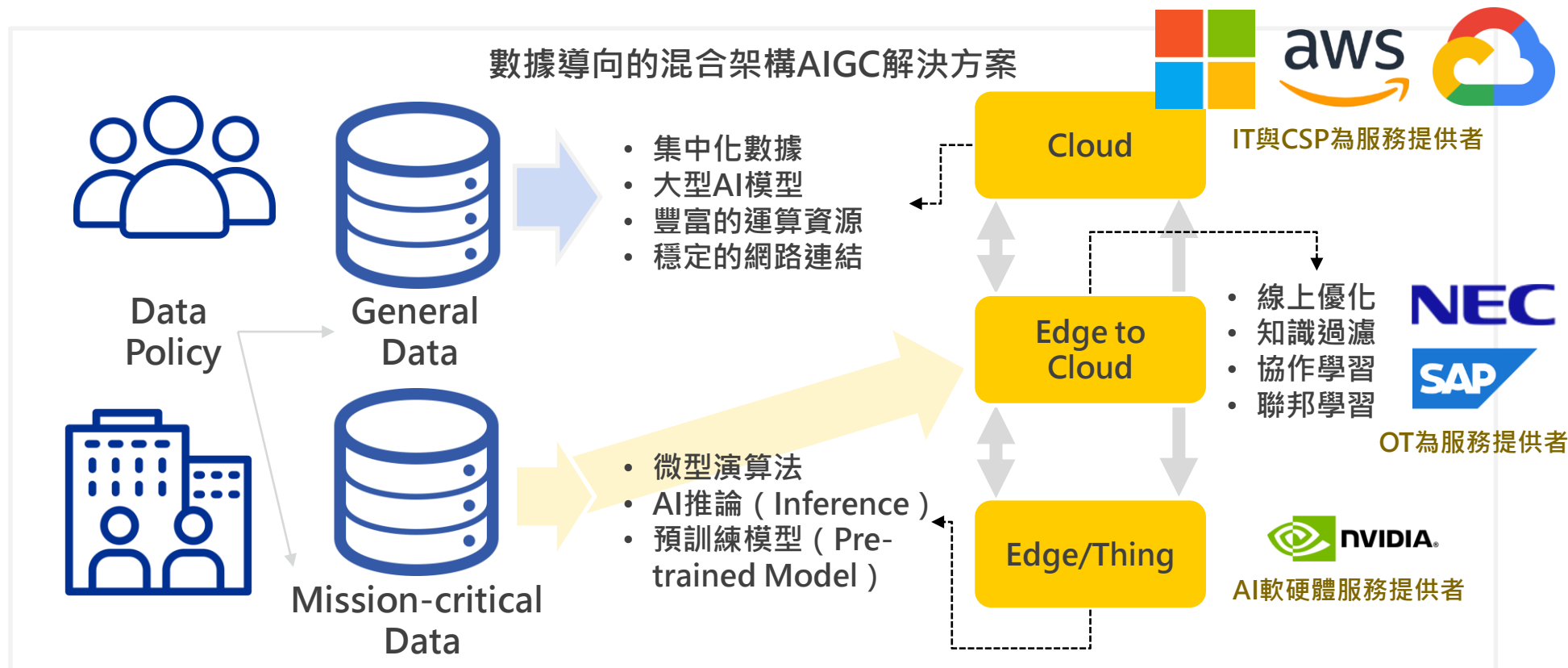
邊緣運算架構的資源處理優勢與限制		
	優勢	限制
雲端	<ul style="list-style-type: none">• 運算資源較高• 服務品質較低• 資訊安全維護較佳• 操作較為方便	<ul style="list-style-type: none">• 高延遲 (Latency)• 資源集中化• 個人隱私風險較高
邊緣	<ul style="list-style-type: none">• 低延遲• 資源分散化• 客戶導向• 客戶掌握程度較高	<ul style="list-style-type: none">• 運算資源受限• 服務品質受限• 資訊安全維護受限• 前期須有設備建置
設備 D2D	<ul style="list-style-type: none">• 個人化導向• 個人隱私風險較低• 高度彈性化	<ul style="list-style-type: none">• 運算資源較低• 服務品質較低• 有電力維持問題• 有設備維修問題

備註：運算資源包括數據運算能力、數據儲存能力等
資料來源：Minrui Xu et al. (2023) · MIC整理 · 2024年4月

- 邊緣運算具「客戶導向」特徵，但在一般性情境，相對雲端亦有運算資源受限的缺點
- 觀察IEEE等技術報告，目前Edge AIGC的運算架構多半採用邊緣雲 (Edge to Cloud) 架構，將雲端、邊緣、設備三端運算資源進一步融合，並依情境產生架構的差異設計



Edge AIGC運算架構情境 (2/2)



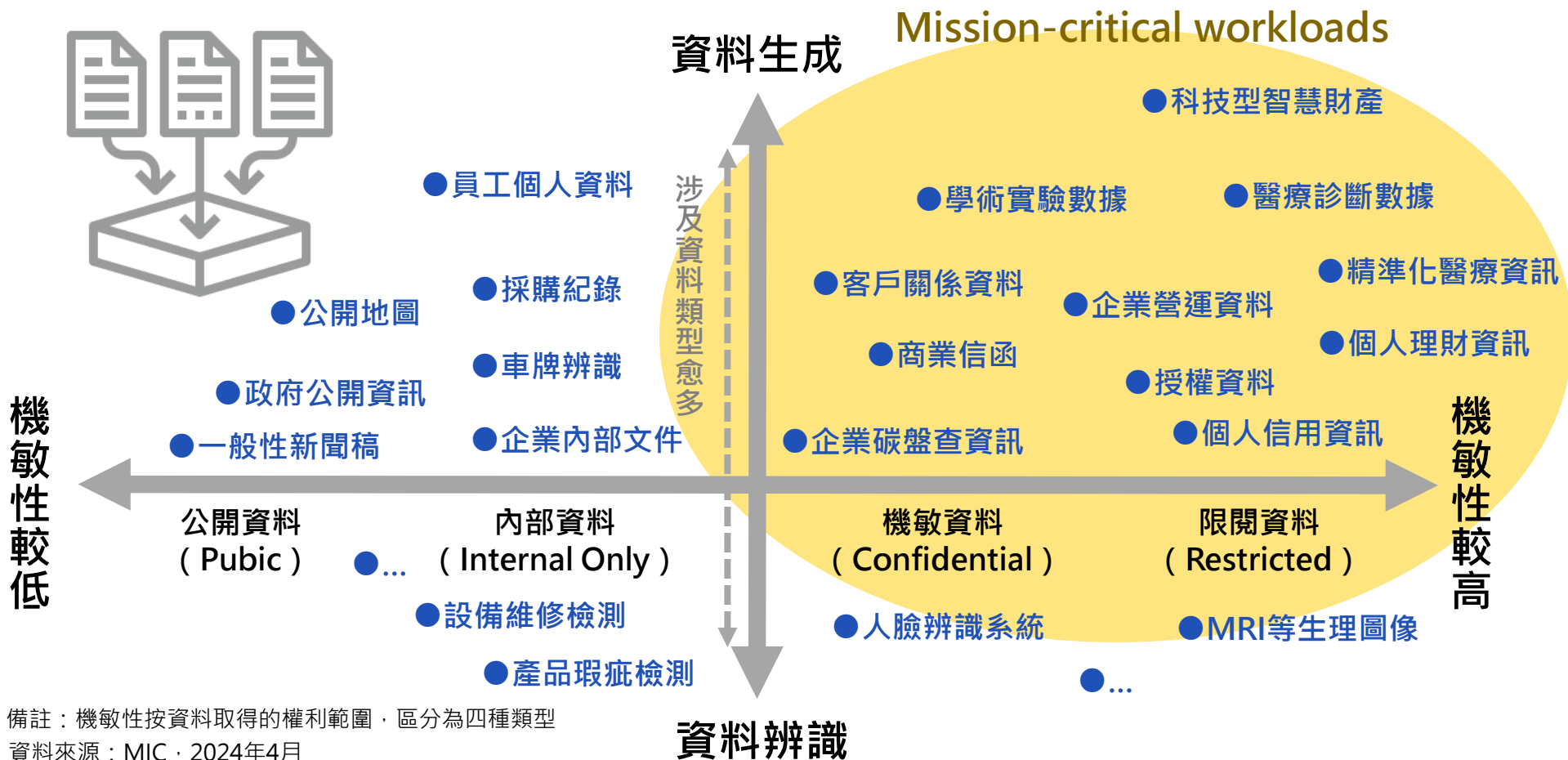
備註：一般性、關鍵任務數據的區分，主要在於資料的機敏性，多半會由該組織所建構的數據政策來進行明確定義

資料來源：Yun-Cheng Wang et al., (2023)、Tao Wang et al., (2023)、MIC整理，2024年4月

- 以企業、組織實務的情境為例，在一般性的情境之中，**企業將生產出「一般性數據」、
「關鍵任務數據」兩種類型的數據**，後者多半具有高度的機敏性，或定義為數位資產
- 「關鍵任務數據」因數位信任、客戶控制權等因素，對於邊緣運算的需求相對較高



藉資料屬性識別Edge AIGC應用

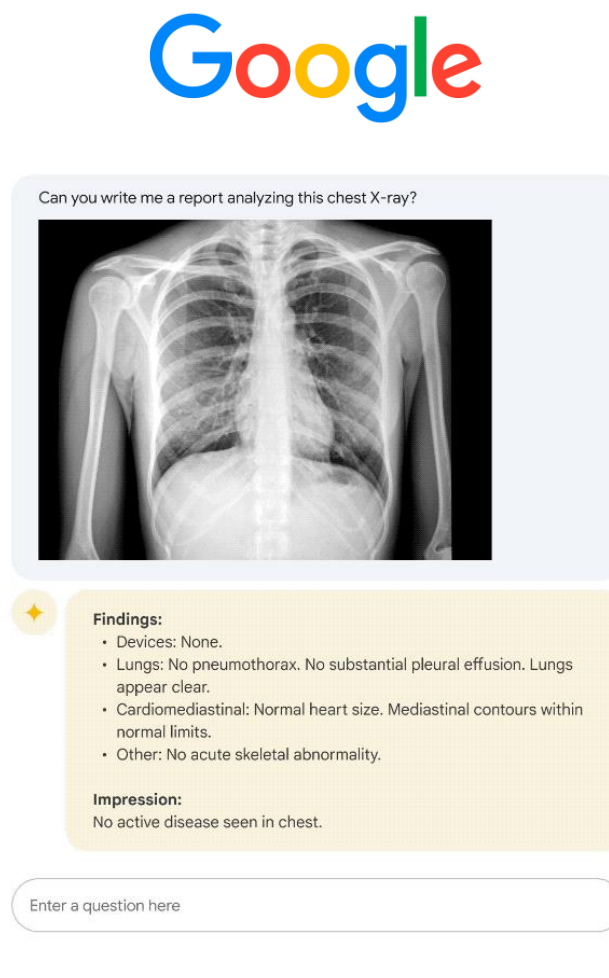


- 藉由以「**資料機敏性**」、「**資料處理型態**」（應用資料生成、資料辨識）兩軸來區分出資料情境分類，如應用情境：**具高機敏性、應用於資料生成者，為Edge AIGC適用情境**
- 包括：科技型智慧財產、醫療診斷、企業營運等，皆為可能的應用情境與場景



Edge AIGC應用案例_醫療資料

案例說明	基礎內容
產品名稱	Google Med-PaLM / Med-PaLM 2
應用情境	理解並分析各種複雜醫學內容，再從醫療資料歸納出診療重點方向、應對問題
基礎資訊	<ul style="list-style-type: none">推出時間：2022特殊紀錄：第一個在美國醫療執照考試，取得及格成績的AI模型
初期採用	TRL：6 CRI：2
核心技術	大型語言模型（LLM） 自然語言處理（NLP）
案例簡介	<ul style="list-style-type: none">簡介：Med-PaLM是醫療用大型語言模型，內容涵蓋：專業醫學檢查、醫學研究等，共包含七個「醫學數據語料庫」。對象為醫事專業人員，期望生成有效的醫療答案效益：<ul style="list-style-type: none">① 降低醫事人員進行知識檢索的時間② 提供醫事人員臨床支持③ 從大量研究中彙整發現



備註：Karan Singhal等人於2023年7月於Nature期刊上，嘗試驗證Med-PaLM回答醫學問題取得科學共識的比例

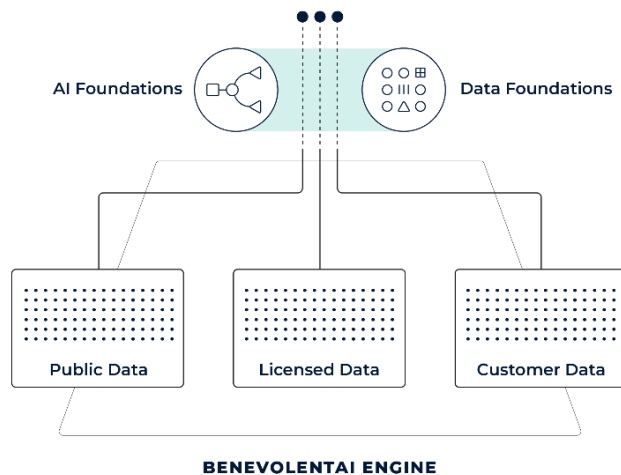
資料來源：Google (2023)、Google Research (2024)、MIC整理，2024年4月



Edge AIGC應用案例_藥物開發

案例說明	基礎內容
產品名稱	BenevolentAI Platform
應用情境	透過端到端人工智慧平台功能來提供新型候選藥物
基礎資訊	<ul style="list-style-type: none">推出時間：2022特殊紀錄：企業創立三年，BenevolentAI 估值快速突破10億美元
發展階段	TRL：6 CRI：2
核心技術	大型語言模型（LLM） 濕實驗室（Wet Lab）
案例簡介	<ul style="list-style-type: none">簡介：BenevolentAI將文獻和數據整合到其生物醫學圖譜，以藥物開發人員為對象藉由該平台探討疾病，如特發性肺纖維化（IPF）潛在機制，以快速識別新標的效益：<ul style="list-style-type: none">① 加速藥物發現的時間② 相對於傳統方法，提供成功機率更高的新型候選藥物評估方法

Benevolent^{AI}



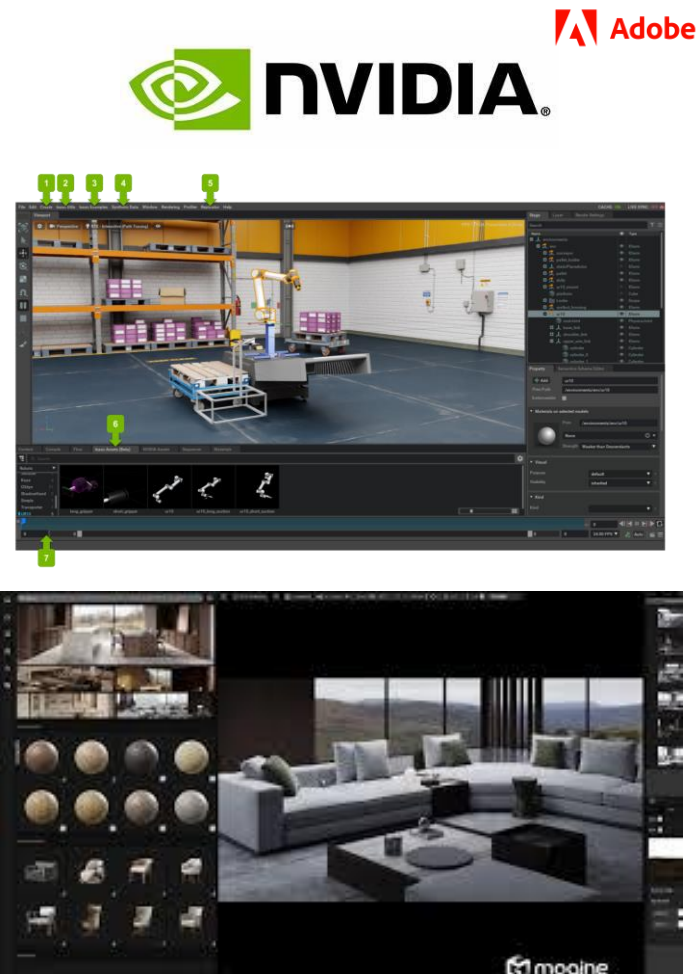
備註：BenevolentAI成立於2013年，為全球基於AI進行藥物開發的領導企業之一，加速生物醫學創新的速度

資料來源：BenevolentAI (2024)，MIC整理，2024年4月



Edge AIGC應用案例_工程建模

案例說明	基礎內容
產品名稱	NVIDIA Omniverse
應用情境	促進3D建模開發，用於創建虛擬場景、建築、產品原型等
基礎資訊	<ul style="list-style-type: none">推出時間：2023（整合AIGC技術升級）特殊紀錄：COMPUTEX Metaverse& XR Application（2023年）類別獎
發展階段	TRL：5 CRI：2
核心技術	AIGC應用於通用場景描述（OpenUSD） 深度學習超高取樣（DLSS）
案例簡介	<ul style="list-style-type: none">簡介：Omniverse於2023年升級，進一步整合OpenUSD、AIGC強化3D建模流程，主要客戶製造、工廠的設計與開發商，可用於流程自動化的設計、3D產品設計等效益：<ul style="list-style-type: none">① 增加廠房、產品的開發人員開發速度② 可應用於更為複雜的廠房、產品開發的項目，降低測試、驗證的難度

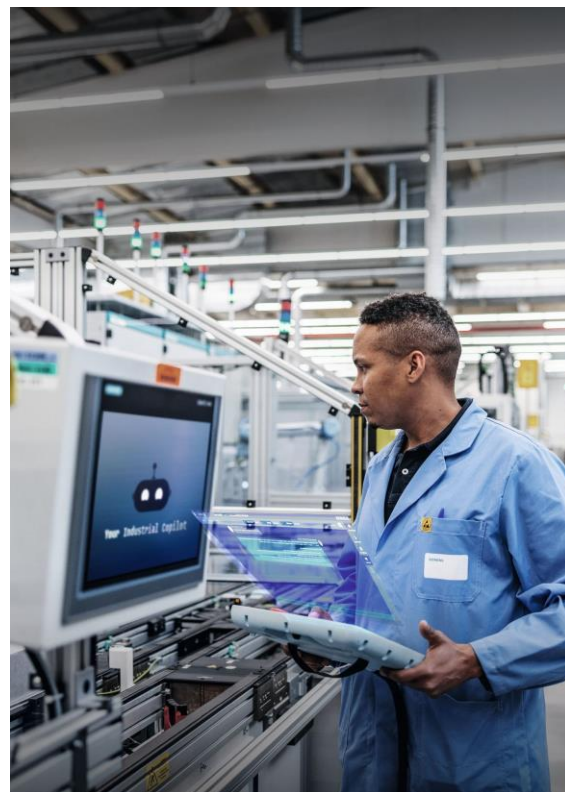


備註：NVIDIA 2023年宣布將Adobe生成式人工智慧模型系列Adobe Firefly，納入在 Omniverse 平台中的 API，提供給開發人員和創作者客戶使用
資料來源：NVIDIA (2023)、Adobe (2024)、MIC整理，2024年4月



Edge AIGC應用案例_製造助理

案例說明	基礎內容
產品名稱	Siemens Industrial Copilot
應用情境	將AI與AIGC導入產品生命週期 (PLM) 管理，強化工廠自動化
基礎資訊	<ul style="list-style-type: none">推出時間：2023特殊紀錄：德國航太、汽車之軸承製造商舍弗勒集團為初期採用者
發展階段	TRL：6 CRI：2
核心技術	自然語言處理 (NLP) AI嵌入自動化軟體 (Automation Software)
案例簡介	<ul style="list-style-type: none">簡介：將AI、AIGC導入到產品生命週期、產品設計、製造生命週期的管理，並藉由Teams進行連結，以Teams為載體讓現場工作人員可以更容易存取、回饋資料效益：<ul style="list-style-type: none">① 降低耗時、昂貴的產品調整② 讓眾多的技術、操作人員，能在日常工作為設計和製造流程做出貢獻



備註：Siemens藉由Siemens Xcelerator獲取自動化和流程模擬資訊，但指出客戶對其資料保持完全控制，並且不會應用於訓練底層AI、AIGC模型
資料來源：Siemens (2023)、Microsoft Community Hub (2024)、MIC整理，2024年4月



Edge AIGC應用案例_業務優化

案例說明	基礎內容
產品名稱	NEC Digital Platform (日文LLM)
應用情境	NEC以客戶內部業務為目標，協助客戶將LLM導入到內部業務，建立客戶企業的內部知識
基礎資訊	<ul style="list-style-type: none">推出時間：2023特殊紀錄：2023年NEC與10家企業和大學合作推出「NEC生成式AI進階客戶計畫」
發展階段	TRL：6 CRI：2
核心技術	日文、術語型的大型語言模型 (LLM) Edge Cloud混合運算架構
案例簡介	<ul style="list-style-type: none">簡介：藉生成式AI服務選單，提供一站式服務項目給客戶，為客戶提供業務諮詢、假設驗證等服務，快速協助客戶能將LLM融合在客戶業務、協助客戶建立內部知識效益：<ul style="list-style-type: none">① 降低客戶內部導入LLM的時間② 協助客戶建立知識模型，同時讓客戶對於NEC的服務產生黏著度



NEC Generative AI Service Menu (NEC Digital Platform)			提供予定
サービス	コンサルティングサービス	LLM活用シナリオの企画に関するサービスを提供開始	7月
	ナレッジエンジニアリングサービス	Microsoft Azure OpenAI Serviceを活用した仮設検証サービスを中心に提供開始	7月
	教育サービス	LLM活用に関するリテラシー教育サービスを提供開始	8月
	環境構築サービス	Microsoft Azure OpenAI Serviceを用いた環境構築サービスを提供開始	7月
ソフトウェア	NEC Generative AI Framework	プロンプト生成・質問管理を支援するソフトウェアを提供開始	10月
LLM	NEC開発のLLM	NEC Advanced Customer Programを対象に	8月 (先行評価開始)
ハードウェア	NEC Generative AI Appliance Server	オンプレミス利用が可能となるハードウェア基盤を提供開始	10月 (評価開始)
データセンター	NEC 印西データセンター	Microsoft Azure ExpressRouteへの接続拠点により低遅延/セキュアな接続を実現するサービスを提供中(NEC LLMと連携予定)	提供中

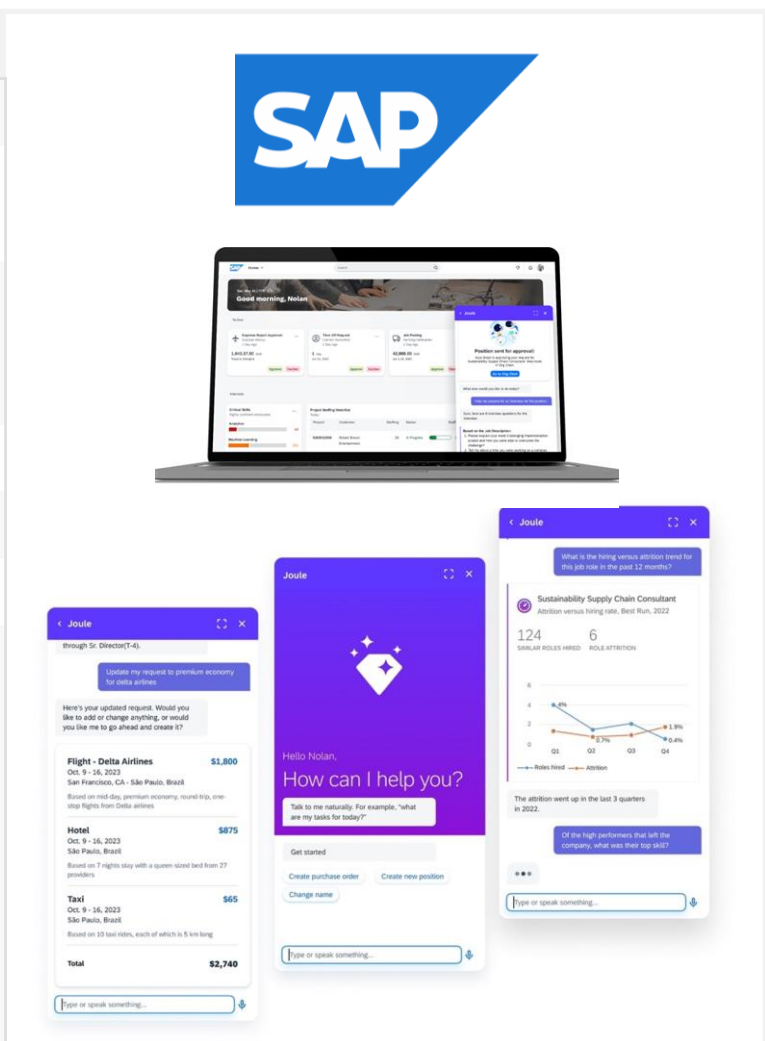
備註：NEC將生成式AI鑲入在Digital Platform之中，並且在2024年1月份宣布建立cotomi新的企業品牌，以日本本地市場的企業為主要訴求客群

資料來源：NEC (2024)、The Japan Times (2023)、MIC整理，2024年4月



Edge AIGC應用案例_商業智慧

案例說明	基礎內容
產品名稱	SAP Joule
應用情境	整合企業私有雲的資料，並加上第三方資料，提供企業專屬的商業洞察
基礎資訊	<ul style="list-style-type: none">推出時間：2023特殊紀錄：2024年3月SAP與NVIDIA 宣布擴展合作夥伴關係
發展階段	TRL：6 CRI：2
核心技術	自然語言生成式AI助理 微服務 (Microservices)
案例簡介	<ul style="list-style-type: none">簡介：將Joule嵌入於SAP企業雲端之中，提供企業智能化的商業洞察。為企業呈現營運智慧方案，包含供應鏈管理、採購、客戶體驗管理等商業智慧洞察效益：<ul style="list-style-type: none">① 強化雲端服務的競爭優勢② 協助客戶針對關鍵業務，提出最適化流程與最佳化解決方案



備註：SAP、NVIDIA預計在2024年底，整合NVIDIA AI foundry與SAP雲端解決方案，並且採用NVIDIA NIM微服務，在企業端部署應用程式
資料來源：SAP (2024)、NVIDIA (2024)、MIC整理，2024年4月



Edge AIGC潛在與創造中的應用

其他Edge AIGC可能、驗證中的應用情境

①

元宇宙服務

將AIGC藉由MEC布署在邊緣，提供個人化元宇宙應用服務



②

個人沉浸式遊戲

整合AR、MR、AIGC等打造個人沉浸式遊戲，強化個人遊戲互動



③

次世代智慧座艙 車用數位儀表

將車用娛樂、導航等系統，藉由AIGC技術整合至車用儀表平台



④

虛擬人即服務 (DHaaS)

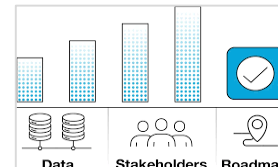
藉由MEC、AIGC等，在現場端打造虛擬人，與人產生臨場互動



⑤

企業碳盤查追蹤

整合採購、供應鏈、ERP等，於企業建立內部碳盤查分析報告



備註：上述應用情境，為IEEE等研究期刊，曾針對Edge AIGC的未來應用進行情境與應用描繪，但尚未有明確企業宣布、公布產品推出期程者
資料來源：BinaryX (2023)、Minrui Xu et al., (2023)、Hongyang Du et al., (2024)、Jiani Fan et al., (2024)、Zi Qin Liew et al., (2024)、KDDI (2021)、Panasonic Automotive Systems (2022)，MIC整理，2024年4月

結論：未來方向觀測與布局建議





無縫化運算架構實現更多元情境

Edge AIGC現有與未來潛在情境

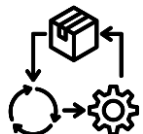
現在 / 已有的Edge AIGC案例



- 契約、企劃、客戶關係文件
- 標準作業程序 (SOP)



- 研究資料重點摘要與總結
- 研究流程設計



- 產品生命週期管理
- 生產線、廠房生產流程規劃



- 程式碼生成
- 合成式數據與產品測試
- 結合CAD產品建模與輔助設計

未來 / 潛在的Edge AIGC案例



- 個人化理財助理
- 個人化醫療、健康顧問



- 智慧車載、航空儀表板
- 車載娛樂系統



- 個人化 (無劇本) 遊戲
- AR、MR、XR高沉浸式遊戲



- 元宇宙、虛擬世界建構
- 個人化虛實交融生活、娛樂體驗

涉及數據語料庫愈多元化，同時，對於IaaS與Edge to Cloud的無縫化技術需求愈高

備註：潛在的Edge AIGC也同時具有高度個人化、客製化的服務特徵

資料來源：Minrui Xu et al., (2023)、Hongyang Du et al., (2024)、Jiani Fan et al., (2024)、MIC整理、2024年4月



生成式AI的採用原則與風控措施

企業使用生成式AI所須依循的原則與建議措施

原則	潛在風險	建議措施
隱私性	共享機密資訊可能會導致侵犯隱私	<ul style="list-style-type: none">在工作場所使用人工智慧工具和聊天機器人制定明確的政策和協定，並確保它們符合GDPR等相關數據隱私法規限制可以上傳到 AI 聊天機器人的數據量，並確保員工知道什麼樣的數據被認為是敏感和機密的
使用者問題	工作場所使用 AIGC，人員需要監督和培訓	<ul style="list-style-type: none">實施組織控制，根據所處理數據的敏感性限制對 AI 系統的訪問確保在工作場所使用人工智慧工具和聊天機器人時，有人工監督和控制指定具有適當專業知識的員工來監控人工智慧工具的使用方式，並確保他們有能力在誤用或事故發生時進行干預
問責性	違反公司數據政策，和潛在的法律法規	<ul style="list-style-type: none">為在工作場所使用人工智慧工具和聊天機器人制定明確的政策和協定，並確保員工接受負責任使用的培訓追究員工對涉及人工智慧工具的任何濫用或事故的責任。組織內建立報告和處理 AI 相關事件的流程，確保及時回應並從事件中吸取教訓
可靠性	導致意想不到後果，如偏見、不準確或其他可靠性問題	<ul style="list-style-type: none">改進 AI 系統文件和培訓材料，強調不與 AI 服務共用敏感資訊確保定期審核和評估 AI 模型的可靠性和準確性；監控和分析數據輸入，以檢測任何偏差或不準確之處，並及時解決；在具有代表性和多樣化的數據上訓練 AI 模型，以避免偏差並確保準確性

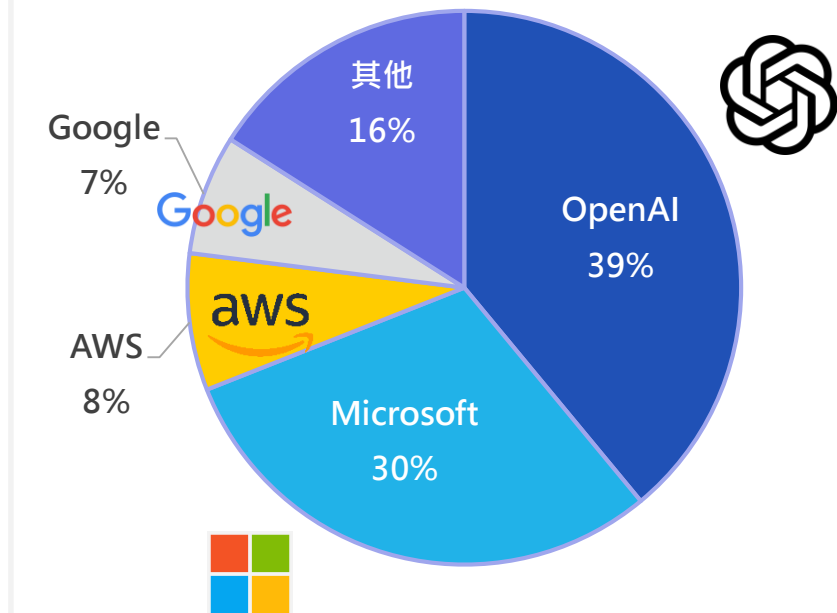
備註：上述DPEX所列出的潛在風險，主要參考經驗案例為三星使用ChapGPT所延伸出來的議題，MIT也同樣指出相似的案例

資料來源：DPEX Network (2024)、MIT Sloan School of Management (2024)、MIC整理，2024年4月



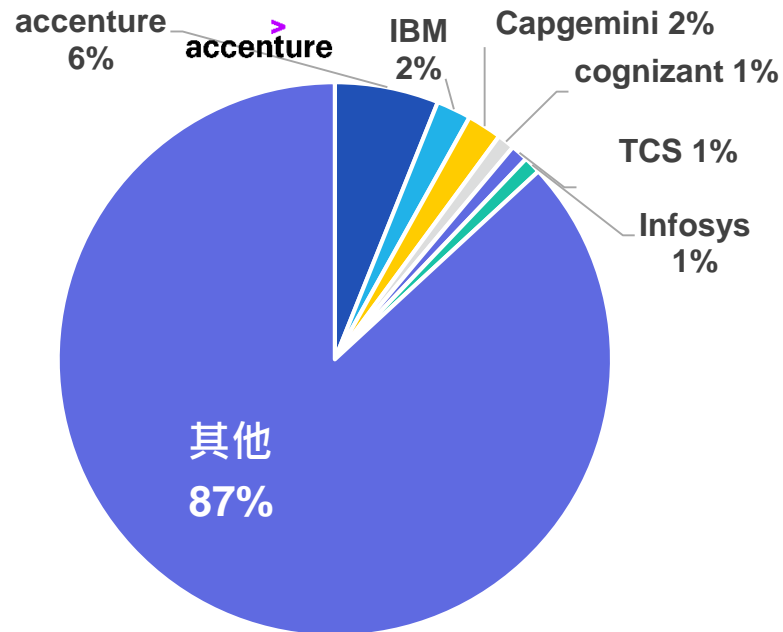
生成式AI服務為近地、利基市場

全球生成式AI模型、平台領導企業



備註：其他依占比包括Anthropic、AI21 Labs、cohere等
資料來源：IoT Analytics (2023)，MIC整理，2024年4月

全球生成式AI服務領導企業



備註：TCS為Tata顧問服務 (Tata Consultancy Services)
資料來源：IoT Analytics (2023)，MIC整理，2024年4月

- 藉由解決全球AIGC的市場占有率，可以發現到「模型、平台」前兩大領導企業，大約佔據近70%以上的市場規模，但「服務」則呈現出高度分散化的市場發展態勢
- 顯示AIGC的服務具有高度的近地、屬地、客製化的市場發展特徵



Edge AIGC服務提供商發展建議

①

以「數據顧問」(Data Advisor) 為新職能定位

AIGC選擇採用Edge架構，其主要因素在於「機敏數據」，以及背後「數位資產意識」。服務提供商須知覺此變化，並給予自身新定位

②

由「專業駐點」團隊、與客戶建立協作模式

Edge AIGC意味服務提供商，必須提供客戶「客製化」的服務，這也意味須掌握數據端「領域知識」，駐點與協作服務模式將成關鍵



③

藉「邊緣雲」(Edge-Cloud) 融合IT、OT兩端

Edge端運算能力不如Cloud，因此服務提供商必須掌握「微服務」、「超融合基礎架構」技術，以整合IT、OT兩端的運算等資源池

④

投入「商業智慧」(BI) 的數據語料庫的建置

單一企業、組織的商業智慧語料庫可能不足，因此，服務提供商是否提供有效的「第三方語料庫」，將成為市場區隔與市場競爭力來源

備註：邊緣雲 (Edge-Cloud) 按GCP定義為：「將雲端基礎架構 (VM) 延伸至邊緣和資料中心，符合資料主權，能在專屬環境使用雲端原生服務
資料來源：IEEE Communications Society (2024)、MIT HAN Lab (2024)、IEEE Computer Society (2023)、MIC整理，2024年4月



Edge AIGC潛在採用者發展建議

①

建構企業「數據政策」

一般情境之下，Edge AIGC成本與技術最佳化模式為混合架構，因此，企業是否建立自身的「數據政策」將是發展自身知識庫的關鍵所在

以「數據長」或Level C人員為企業推動引擎

②

企業、組織內部所用的Edge AIGC預期會涉及到企業內部不同的數據孤島 (Silo)，如無Level C以上的人員投入，推動速度將可能放緩

③

Edge AIGC前三年將有「邊緣設備投資支出」

Edge為導向的架構，在臨場端預期出現邊緣伺服器、微型資料中心的建置需求，在投入前期預期會出現新的設備投資支出，須通盤規劃



將「AI+BI」作為中長期的企業發展願景、目標

④

Edge AIGC可被視為企業將AI導入BI的一種數位轉型方案，這也意味企業須將「AI+BI」為中長期目標，朝向「數位資產」的智慧化發展

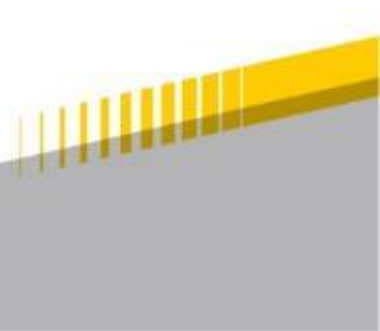
備註：數據政策 (Data Policy) 是針對企業內部的所有數據、數位資產，舊資料的機敏性、重要性、QoS等進行分類、分級規範的企業發展行為
資料來源：MIT Sloan Management Review (2024)、IMD (2024)、IEEE Computer Society (2023)、MIC整理，2024年4月



結論

- AI、AIGC在數據生成、運算需求，**AI已經改變IT運算基礎設施**，AIGC尤其會進一步驅動數據增量（文本生成）、語料資料遷移的需求，加上客戶之於「**數據資產**」的「**數據控制權**」意識興起，具有「**客戶導向**」特徵的**邊緣運算**，成為關鍵的技術解決方案
- Edge AIGC多採邊緣雲（Edge to Cloud）架構，將雲端、邊緣、設備資源進行分層與融合，**適用內部「關鍵任務」（Mission-critical Data）數據情境**，如企業內部業務優化、商業智慧等
- 未來Edge AIGC預期朝元宇宙、高沉浸遊戲等個人化應用發展，但隨著**數據語料庫更加多元化**，**邊緣雲最佳化設計將成為關鍵**
- **Edge AIGC將為工業控制、資訊服務業者，創造出新市場商機**，不過對於服務提供商來說，服務提供模式、自身定位也將改變
- 企業導入Edge AIGC，可視為企業將數位資產「智慧化」的實踐，但無可跨越的第一步是 - **企業首先必須建構自身「數據政策」**





附件





邊緣運算相關之技術定義與內涵

邊緣運算、多重接取邊緣運算定義

組織	定義
	多重接取邊緣運算 (Multi-access Edge Computing) 為開發人員和內容供應商，直接提供、建構如同雲端運算的能力，並且具有網路邊緣的IT服務特性，擁有超低延遲、高頻寬，以及即時存取、即時回應等功能
	邊緣運算 (Edge Computing) 是一種去「中心化」的運算基礎設施，其中運算資源和應用服務，可以從數據資料的來源一直連結到雲端，可以進行資料的蒐集，或者可讓使用者執行某些低延遲的操作，在「邊緣」建構一定的資源池，以滿足運算需求
	邊緣運算 (Edge Computing) 的架構，是將運算資源從中心化的資料中心或雲端位置上，轉移到更靠近設備的地方，以支援具有較低延遲需求的應用，同時，更有效地處理資料，以節省網路成本，當資料在邊緣，而不是雲端處理時，回程 (backhaul) 成本就會降低
	邊緣運算 (Edge Computing) 是在更加靠近數據生成位置的地方，使用數據儲存以及運算能力，這是因為數據資料的總量，已超過現有運算架構的能力，透過邊緣運算系統，能顯著提高了應用程式效能、降低頻寬需求，並提供了更快的即時洞察

備註：ETSI除了多重接取邊緣運算之外，亦提出行動邊緣運算 (Mobile Edge Computing)，皆以MEC作為技術架構的簡稱

資料來源：ETSI (2015, 2023)、Industry IoT Consortium (2024)、Cisco (2022)、AWS (2022, 2024)、MIC整理，2024年4月

- 邊緣運算的發展可追溯到2010年代中期，主要推動者為ETSI、OpenFog、IIC
- 邊緣運算的基礎定義與內涵有：①將原先放置於雲端、或部分雲端運算的服務，遷移至鄰近數據生成點的位置；②在鄰近數據生成點的位置上，建置運算與儲存的資源池



生成式AI相關之技術定義與內涵

生成式AI定義	
組織	定義
 Now Part of Digital Twin Consortium	生成式人工智慧 (AIGC 、 AI Generated Content) 是指廣泛建基於人工智慧的方法、模型和應用程式，主要利用大量資料，包括大型語言模型，來產生新內容，這些內容可能包括文字、影片、圖像、音訊、程式碼，近年快速地在全球進行推展
	生成式人工智慧 (AIGC) 是一種深度學習模型，可以根據訓練資料，產生高品質的文字、圖像和其他內容，資料隱私以及信任、透明度的問題，是當前全球組織、企業採用生成式AI的最大障礙，同時也產生新的技術需求
	生成式人工智慧 (AIGC) 模型使用神經網路，來識別現有資料中的模式和結構，藉此產生新的原創內容，使用者能夠根據各種輸入的指令，快速產生新內容，而模型輸入和輸出形式可包括文字、圖像、聲音、動畫、3D模型等其他多元類型的資料
	生成式人工智慧 (AIGC) 是人工智慧的一種形式，可以根據訓練的資料，產生文字、圖像和各種內容，可以分為基於Transformer的模型 (如GPT-3、GPT-4)、生成式對抗網路 (GAN)、變分自編碼器 (Variational Autoencoder, VAE) 等

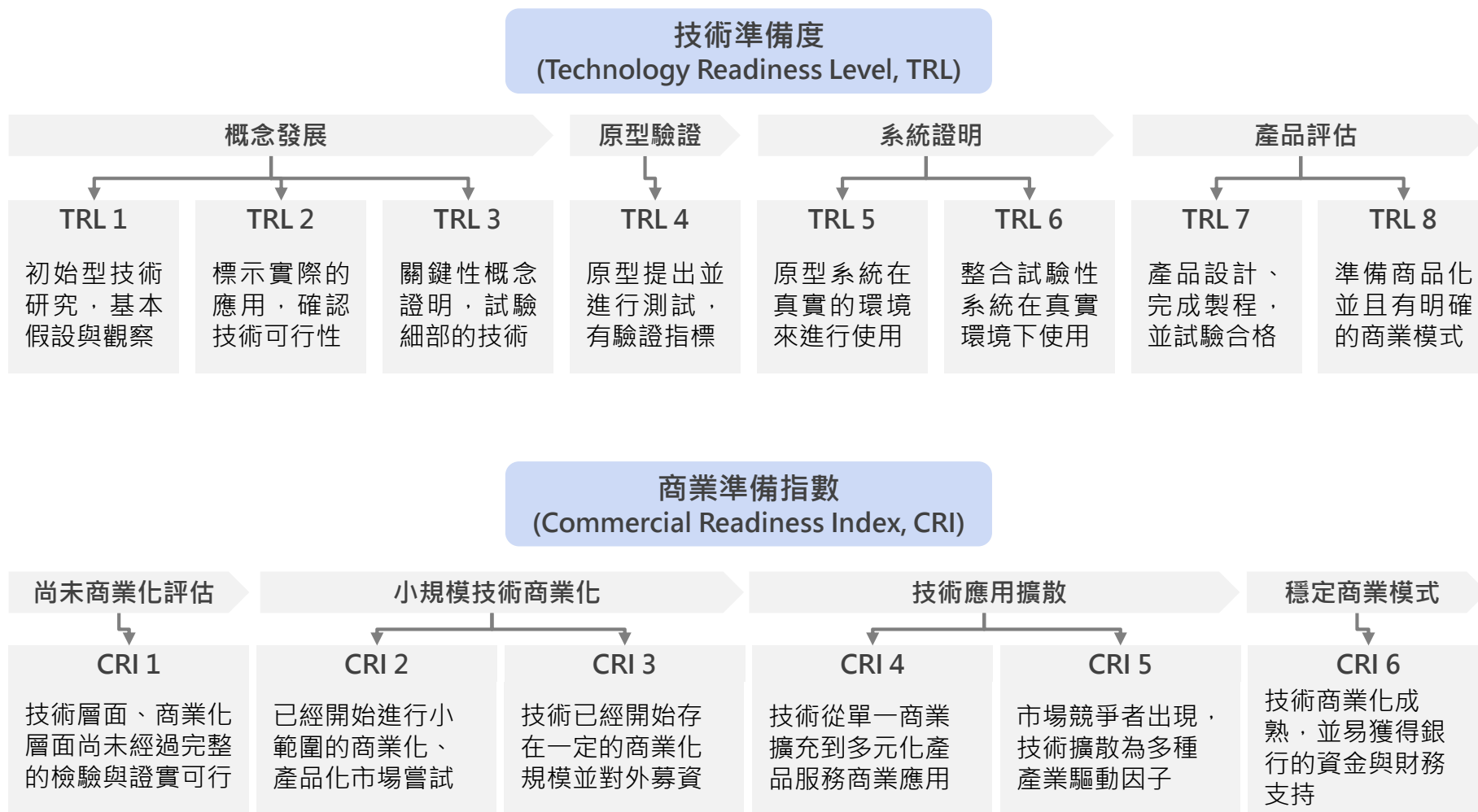
備註：ETSI除了多重接取邊緣運算之外，亦提出行動邊緣運算 (Mobile Edge Computing)，皆以MEC作為技術架構的簡稱

資料來源：Industry IoT Consortium (2024)、IBM (2024)、NVIDIA (2024)、SAP (2024)、MIC整理，2024年4月

- 生成式AI的發展可追溯到2020年代初期，主要推動者為OpenAI、SAP等
- 生成式AI的基礎定義與內涵有：①利用大量數據語料，包括大型語言模型，進一步產生新內容；②使用者可以藉由各種輸入的指令，來影響生成式AI所產生的新內容



產品技術與商業準備度指標體系



備註：包括IEA等機構已有相關產品、技術準備度的評估報告，同時採用TRL、CRI兩種指標來進行技術創新階段的評估，融合技術與商業發展思維

資料來源：NASA (2012)、Bezuidenhout, L et al., (2017)、IEA (2023)，MIC整理，2024年4月

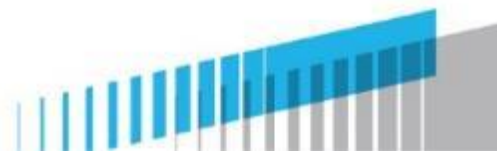


MIC 產業提昇的關鍵力量
Thank You

施柏榮 產業顧問兼副主任

pojungshih@iii.org.tw

產業情報研究所



智慧財產權暨引用聲明

- 本活動所提供之講義內容或其他文件資料，均受著作權法之保護，非經資策會或其他相關權利人之事前書面同意，任何人不得以任何形式為重製、轉載、傳輸或其他任何商業用途之行為
- 本講義內容所引用之各公司名稱、商標與產品示意照片之所有權皆屬各公司所有
- 本講義全部或部分內容為資策會產業情報研究所整理及分析所得，由於產業變動快速，資策會並不保證本活動所使用之研究方法及研究成果於未來或其他狀況下仍具備正確性與完整性，請台端於引用時，務必注意發布日期、立論之假設及當時情境



AISP 情報顧問服務

Advisory & Intelligence Service Program

產業情報顧問服務AISP為資策會MIC最核心的產業情報資料庫服務，運用最先進數位平台服務技術，提供產業在資訊與通訊（ICT）領域最完善的新知識、新技術、新方向的產業情報資訊服務平台。服務內容包括「產業情報資訊、突發事件觀察剖析、關鍵議題焦點評論、產業議題深度研究、國際大展情報蒐集分析、前瞻趨勢」等。隨時觀察產業發展動態與趨勢，觀測掌握全球重要的產業發展動態，並依據產業需求規劃研究範疇與議題，開展符合產業需求的產業情報資料庫。

推薦資料庫



Application IC & Components 應用IC與關鍵零組件

本產品研究範疇包含應用IC與關鍵零組件於新興應用之發展，聚焦新興應用晶片與零組件技術、應用晶片與技術發展、晶片大廠競合與熱門議題等。透過技術分析、應用分析、產品分析與市場動態等不同面向，探討半導體晶片與關鍵零組件導入新興應用之相關發展。

研究範疇

- 應用IC與關鍵零組件於新興應用之發展分析

研究重點

- 新興應用晶片與零組件技術
- 應用晶片與技術發展
- 晶片大廠競合與熱門議題

研究構面

- 技術分析
- 應用分析
- 產品佈局分析
- 市場動態分析

Performance Computing 運算系統

本產品針對電腦主機板、桌上型電腦與伺服器等資訊系統產品，並新增高效能運算、資料中心、邊緣運算與雲端服務大廠之重要議題，除原本產銷訪查與趨勢分析，另針對重要議題之產業發展、產品動態進行研究剖析，以協助上下游業者掌握運算系統產業未來商機。

研究範疇

- 一般資訊運算暨高效能運算系統產品之產業趨勢與市場前景

研究重點

- 桌上型個人電腦與其主機板
- 伺服器與企業資訊運算系統
- 資料中心技術與應用發展
- 邊緣運算與分散式架構
- 雲端運算產業與政策研析

研究構面

- 市場分析
- 產銷分析
- 產品發展分析
- 關鍵晶片分析
- 產業競爭分析

AISP情報顧問服務網
<https://mic.iii.org.tw/aisp>

瞭解更多

Artificial Intelligence 人工智慧

本產品以「AI產業化」及「產業AI化」兩大主軸進行研究，在「AI產業化」上探討各式新興AI及生成式AI演算法之應用、AlaaS服務、人工智慧硬體及晶片、軟體框架等議題；「產業AI化」則探討不同行業AI及生成式AI技術於場域之議題、產業AI化動態及新創應用案例等內容。

研究範疇

- AI產業化之相關軟硬體、治理議題，以及產業AI於不同垂直領域之應用發展

研究重點

- 新興AI軟硬體及平台
- AI新興算法與服務
- AI大廠領域佈局動向
- 重點應用領域趨勢
- 可信任AI與AI評測

研究構面

- 技術趨勢前瞻
- 標竿廠商動向
- 國際重點政策
- 產品發展分析
- 標竿應用案例

MIC到府簡報服務

趨勢洞察力 決定 企業競爭力

MIC 協力為您促進 組織 / 人才 再升級

組織人才前瞻力的提升，儼然已成為現今企業突破轉型的新顯學。為成功協助企業菁英掌握瞬息萬變的市場趨勢，特別針對產業熱門議題以及MIC重點研究，提供研究顧問至貴公司「到府簡報」之服務，期盼能將MIC多年凝聚累積的研究能量及專業精闢的情報服務，深耕企業內部員工，加速提升組織競爭力，共創企業新價值，與企業組織人才攜手找出迎向新經濟的解方



15 大
議題精選

- 產經趨勢
- 資訊產業
- 半導體產業
- 5G/B5G
- 數位經濟
- 金融科技
- 科技應用
- 電動車
- 人工智慧
- 數位轉型
- 資安防護
- 智慧城市
- 智慧製造
- 智慧醫療
- 能源與環境

點擊詳閱
MIC到府簡報議題



欲瞭解詳情，請洽MIC會員服務中心，由專人為您服務

☎ (02)2378-2306 ✉ members@iii.org.tw

MIC 資策會 | 產業情報研究所