智慧系統與晶片產業發展策略會議

《智慧系統與晶片技術》

報告人

經濟部技術處　羅代處長達生

【智慧系統與晶片產業發展策略會議】

部會座談引言
智慧系統與晶片技術

主辦：經濟部

協辦：科技部、教育部

106/07/11

簡報大綱

一. 背景分析

- 國際發展趨勢
- 國內發展現況與挑戰

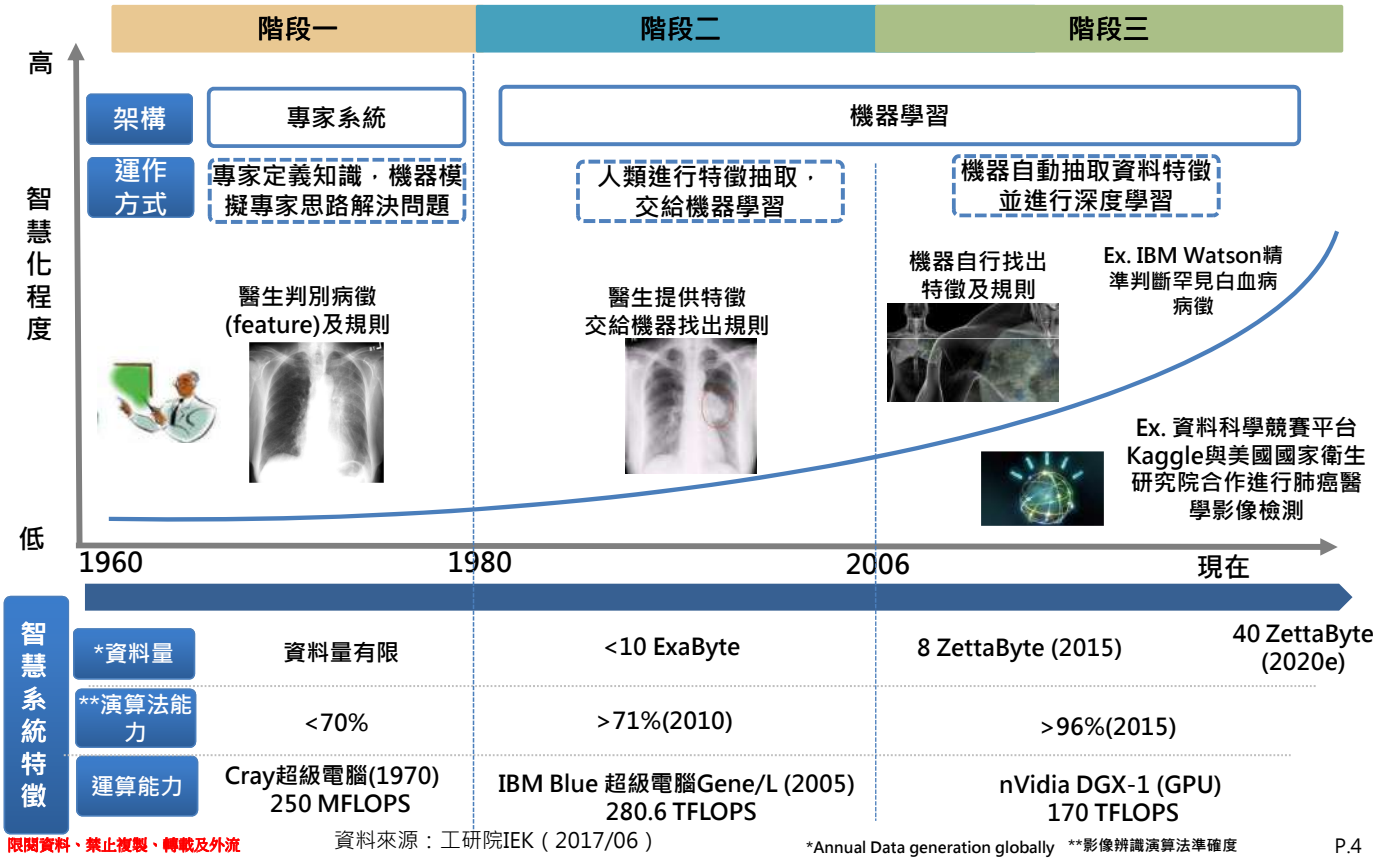
二. 主軸目標與推動作法

- 願景與目標
- 策略作法
- 最終效益(End-Point)

三. 討論題綱

國際發展趨勢(1)

資料量、演算法、運算能力驅動智慧系統發展



資料來源：工研院IEK（2017/06）

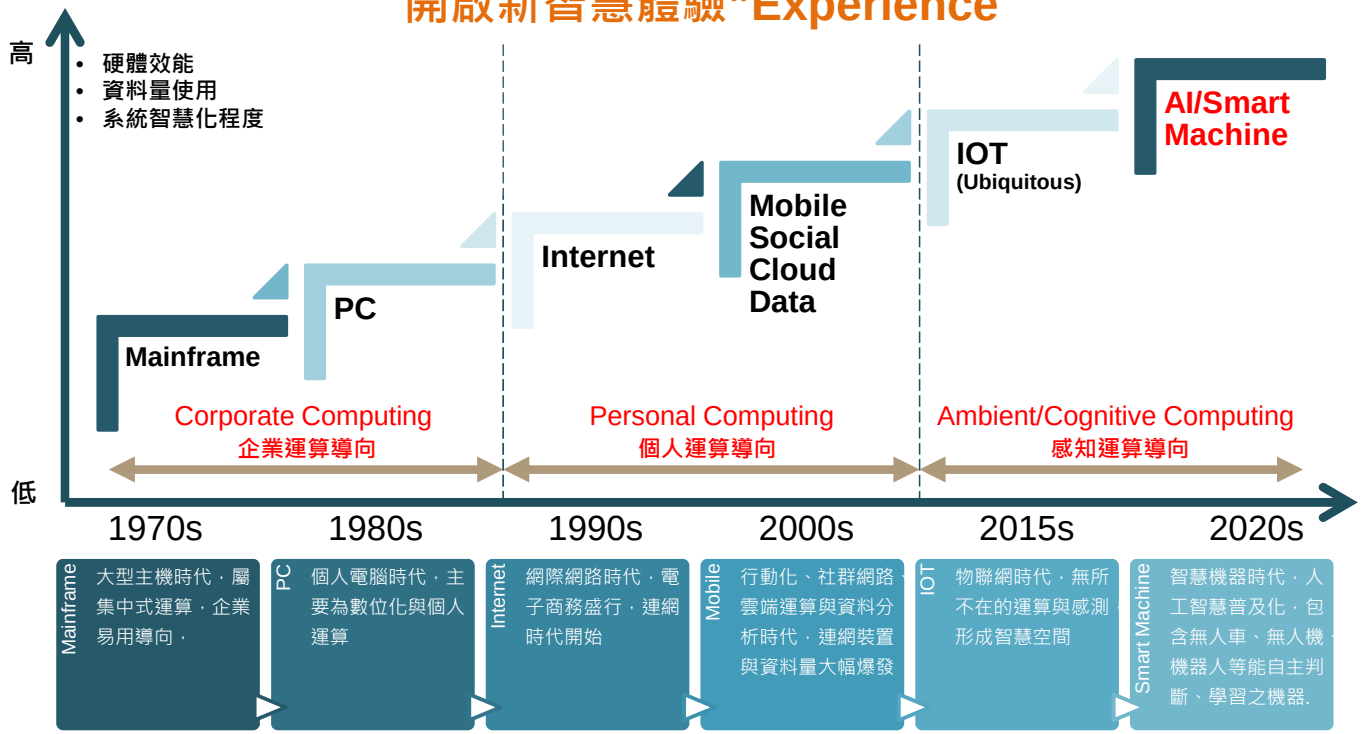
*Annual Data generation globally **影像辨識演算法準確度

國際發展趨勢(2)

新物聯網時代，人工智慧接手

結合Data, Computing, Algorithm

開啟新智慧體驗"Experience"

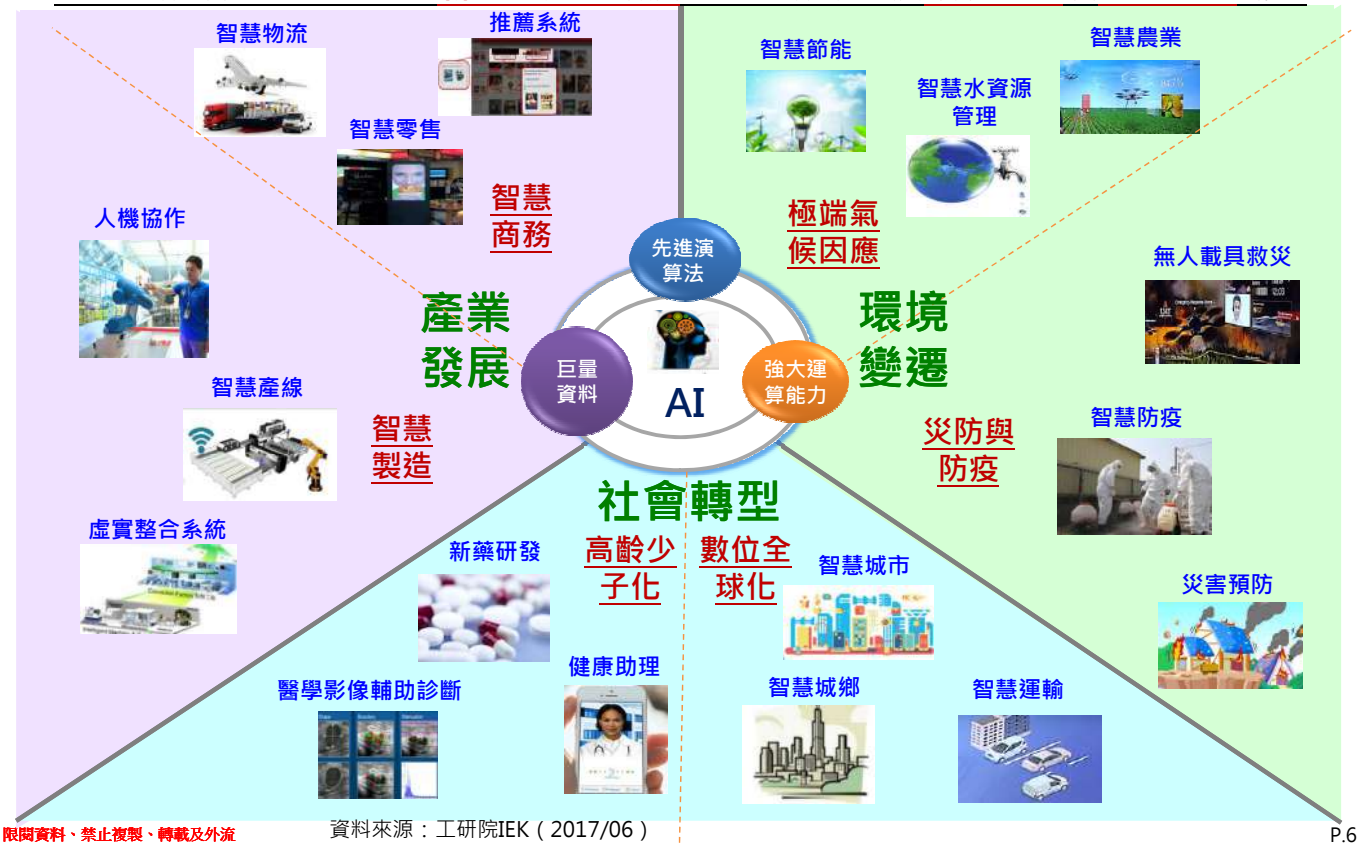


資料來源：工研院IEK（2017/06）

國際發展趨勢(3)

智慧系統帶動新型態產業及環境發展

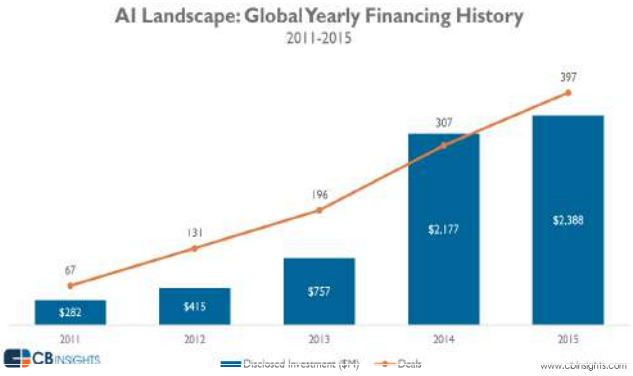
智慧系統技術不僅能帶動新型態產業發展，也能協助解決環境變遷及社會轉型議題



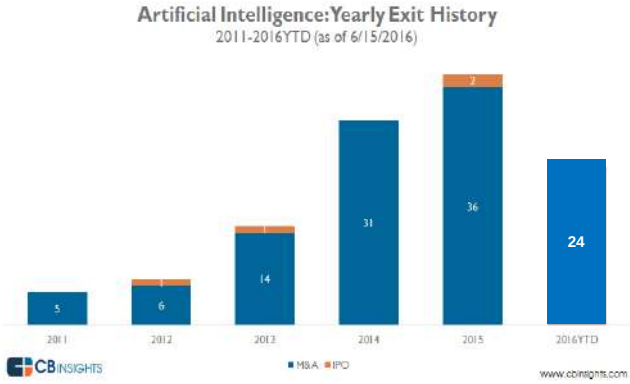
資料來源：工研院IEK（2017/06）

國際發展趨勢(4)

智慧系統帶動新創事業發展、創造新經濟

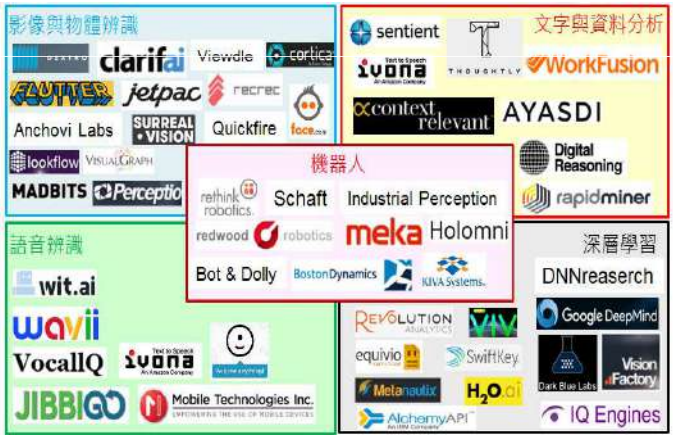


資料來源： CB Insights (募資金額, 新創家數) 2016/06/20



資料來源： CB Insights (併購家數, IPO家數) 2016/06/26

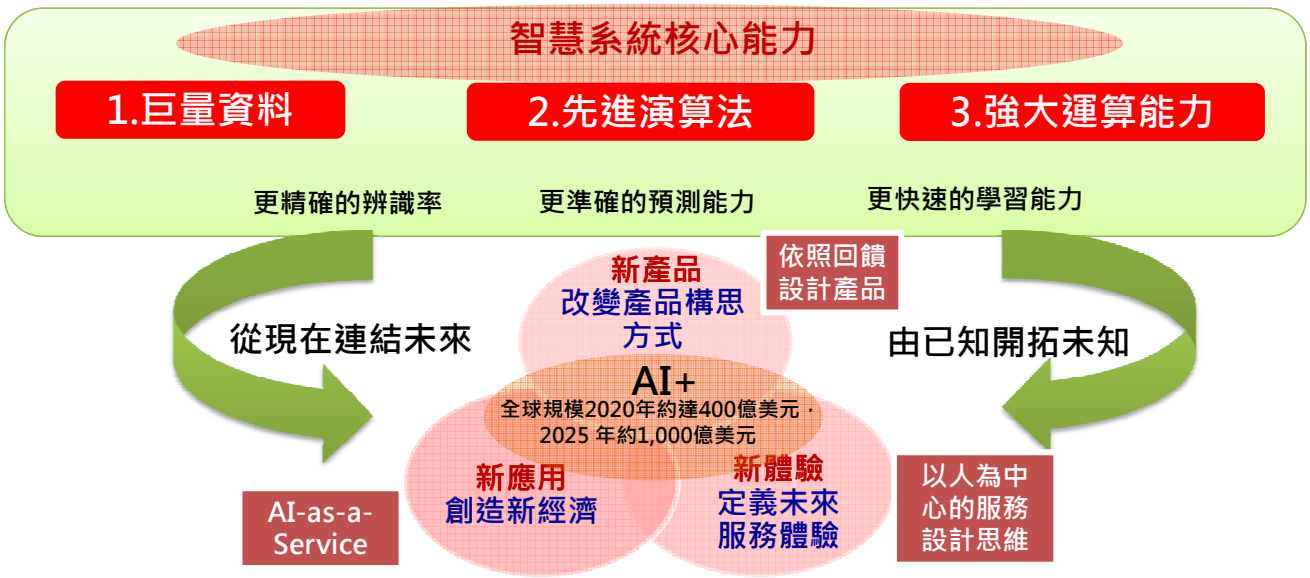
AI的新創公司增長快速，全球IT大廠積極以併購新創公司取得人工智慧技術與人才，意味大廠亦根基未深，正是切入時機點



資料來源：工研院IEK（2017/06）

國際發展趨勢(5)

智慧系統創造出新應用、新產品、新體驗



應用服務

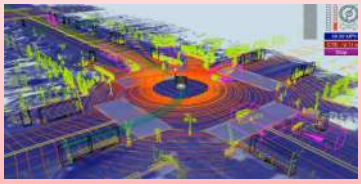
資料來源：Tractica；工研院IEK（2017/06）

Google 布局自駕車深化智慧系統技術

巨量資料

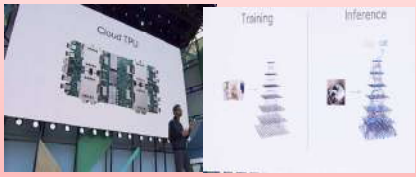
- 無人車近60輛，累計行駛超過 200萬英哩
- 每秒產生數據量高達750MB

先進演算法



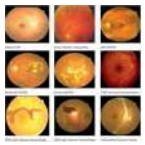
- 每0.1秒做判斷，每1英哩，約做出1,000次決定
- 決定該變換駕駛路線或調整行駛速度

強大運算能力



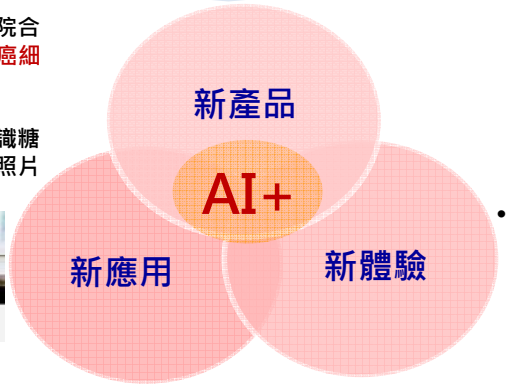
- Tensor Flow Lite：自研獨特計算架構，一塊板子具有 4 個 TPU 計算核心，理論計算力達到了 180 TFlops（萬億次浮點計算）

智慧系統核心能力



- Google Home 家用語音助理硬體
- 醫療領域：與倫敦醫院合作以AI開發自動識別癌細胞的放射治療儀器
- Google.ai
  - 以深度學習辨識糖尿病患者眼底照片並預警

- Google Assistant
  - 語音虛擬助理
- Google for Jobs
  - AI最佳化的求職應用
- Google Email
  - 自動標記和處理垃圾郵件



辨別花種　WiFi自動連線

顯示店家評價

- Google Lens
  - 以影像辨識取代關鍵字搜尋，使用手機相機就能獲得圖像相關資訊
  - 如掃描演唱會海報，能辨識出購票資訊，直接排進行事曆

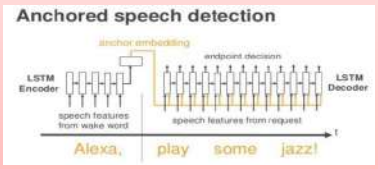
資料來源：Google；工研院IEK整理(2017/06)

Amazon 布局智慧音箱強化智慧系統技術

巨量資料

- 以出貨量及每台每天使用量預估，每天約產生100TB的資料量

先進演算法



- 錨定語音檢測（Anchored Speech Detection）能在多人對話環境中準確辨認出下指令者
- 以Large-scale distributed training加速訓練速度

強大運算能力

- 以80個 GPU 執行約 55 萬幀/秒的速度，每一秒的語音大約 100 幀，一個人要花 16 年的時間來學習 1.4 萬小時語音，Amazon約 3 個小時就可學習完成

智慧系統核心能力

- 智慧物流
  - 導入倉儲機器人Kiva，Amazon倉儲與物流機器人化，減少倉儲成本48%
  - 無人機送貨

- 智慧零售
  - 以機器學習分析用戶偏好，提出智慧推薦與預測服務
  - 實體商店中以車牌辨識技術自動提貨

- 免結帳商店Amazon GO
  - 以相機、麥克風、感測器融合、圖像分析，收集分析用戶購買行為
  - 以大數據分析用戶偏好據以調整庫存
  - Just Walk Out技術自動追蹤商品位置

資料來源：Amazon；工研院IEK整理(2017/06)

Uber以智慧系統技術為後盾，持續衍生新應用市場

- 以技術擴張帶動服務擴張：Uber自行定位為高科技公司，以自研技術--供需媒合、浮動價格計算模型出發，持續投資新技術領域，藉由在地實驗與測試，不斷衍生各種服務，被全球新創仿效
- 藉技術合作與併購布局跨業新服務：除了自研技術，亦與夥伴合作，如無人駕駛技術與Volvo、零組件業者DELPHI、圖資業者TomTom合作，續研發智慧推薦技術，整合不同領域推出跨業新服務

資料來源：各公司；工研院IEK(2017/06)

重點國家之國家級人工智慧政策比較

➢ 美、日、韓等國已將AI列入國家重要科技發展項目，並陸續發表推動計畫揭櫫未來發展願景與布局重點

	政策綱領或推動方案(發布時間)	技術布局項目	關鍵推動措施	目標

美國	• 白宮發布國家人工智慧研發策略計畫 (2016/10)	• 數據為中心的演算法
		• 增強人工智慧感知能力
		• 高性能服務機器人
			• 長期投資AI技術
			• 聯邦政府提高使用AI能力
			• 健全AI制度和法案
			• 發展AI共享資料環境
			• 評估AI對就業市場影響
			• 評估AI人力需求
				• 以人工智慧提升美國總體競爭力
				  • 製造業
				  • 物流業
				  • 金融業
				  • 運輸業
				  • 農業
				  • 行銷媒體業

日本	• 文部科學省發布先進整合智慧平台計畫(AIP項目) (2016/5)	• 自然語言處理
		• 影像分析
		• 機器學習演算法
		• 預測型安全技術
			• 成立產官學研合作計畫
			• 在法人「理化學研究所」之下成立革新智慧統合研究中心，解決人才不足問題，並研究製造與醫療合作兩大AI方向
				• 解決國內高齡化、防災等社會問題
				• 四大AI產業出口為目標
				  • 製造業
				  • 行動生活
				  • 醫療/照護/健康
				  • 零售商務

韓國	• 科學技術戰略委員會宣布未來五年投入1兆韓幣（約美金八億八千萬）建立人工智慧研發中心(2016/3)	• 視覺分析工具
		• 自動翻譯
		• 自然語言辨識
		• 大數據
		• 雲端運算
			• 產官學合作
			  • 政策鼓勵企業使用AI
			  • 成立公私合營公司AIRI，打造Big Data、AI和Cloud的共通平台以開發行業應用
			  • ETRI負責AI技術
				• 開發可有效解決社會問題的解決方案
				• 培育世界級AI人才、提供實作場域

資料來源：工研院IEK (2017/06整理)

國際各國近期AI推動策略構面

- 五大策略構面推動人工智慧：全球政府近期推動之人工智慧政策內容，不外乎奠定基礎技術、研發智慧機器、推動行業應用及鼓勵新創企業，另也關注AI對產業就業結構、社會與倫理之影響
- 技術紮根與環境整備並進：各國也著手整備產業環境，如實作場域提供、人才培育、推動公開資料、產官學協同推動及國際接軌等面向也是各國政策探討重點

1 奠定基礎技術

將演算法、視覺技術、語言分析等技術納入國家研究計畫，加強發展機器學習、深層學習、自然語言處理、圖影像分析等基礎技術

2 研發智慧機器

聚焦於機器人和自駕車的自動控制、視覺辨識、主被動安全等技術研發，並以智慧機器的功能和表現水準，作為國家人工智慧技術強弱比較基礎

3 推動行業應用

界定產業鏈，促進人工智慧技術產業化，推動製造業、金融業、醫療業、網路、機器人等領域，以及連結物聯網技術發展相關應用

4 鼓勵新創企業

公立投資銀行提撥基金宣導創業，各大學開展創業計畫，提供學生育成場所，政府也廣設新創園區，大學積極培育巨量資料、人工智慧、物聯網等領域人才

5 關注產業與就業結構變化、社會與文化的影響、倫理與道德觀念的變遷

資料來源：工研院IEK整理(2017/06)

國內發展現況與挑戰—智慧系統

挑戰三:產業對AI效益採取觀望
1.成功實例說服力仍有限
2.需藉由整合開發平台降低業界投入門檻與風險，提高投入誘因

	智慧系統範疇	投入業者	投入概況
應用服務	垂直領域應用服務 Vertical Application and Services	製造業、金融業、零售業、服務業、…	導入應用領域仍有限，以影像及客服系統導入較快
系統整合	資料分析與運算 Intelligent Analytics	沛星互動科技、精誠知意圖、御言堂、賽維科技、意藍資訊、…	業者解決方案智慧化程度有待強化
雲	雲端管理及系統營運平台 Cloud Computing and Platform	廣達、緯創、英業達、華碩雲端、宏碁、研華、…	伺服器業者已切入AI伺服器產品，然多採國際晶片方案
網	通訊異質網路 Heterogeneous Network	合勤、智邦、研華、中磊、正文、富鴻網、....	產業已布局邊緣運算技術但產品與AI整合度仍低
端	感知裝置、系統晶片 Sensor Device and SoC	聯發科、威盛、原相、奇景、… / 鴻海、宏達電、華碩、宏碁、和碩、…	廠商已投入但尚在先期研發階段

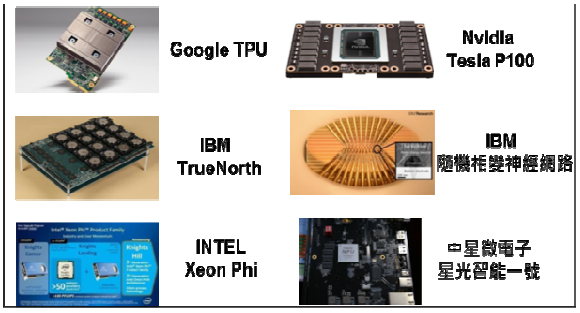
挑戰二:須強化資料掌握度
1. 需擴大數據量能以養成數據訓練模型
2. 如何掌握資料標記技

挑戰一:欠缺AI系統自主能力
1.系統晶片感測技術 2. 先進伺服器雲平台技術 3. 智慧運算分析技術 4. 智慧系統整合技術 5. AI安全隱私保護技術

國內發展現況與挑戰—智慧晶片

- **AI晶片產業範疇**：智慧晶片產業包含AI Chip, 感測器與智慧感知模組，異質整合與晶片安全技術。其中AI Chip起步較晚但業界已有實績；感測器與智慧感知模組發展已久，已掌握一定基礎；晶片製造及封裝異質整合強；晶片安全技術等亦有業者勤於耕耘利基市場
- **結合系統廠商進行場域驗證**：由於AI晶片架構與系統應用需求特性息息相關，需有系統廠商配合提出規格並合作進行場域驗證

國際大廠搶先投入AI Chip



我國智慧晶片產業與現況與挑戰

	AI Chip	Sensor	異質整合	晶片安全
布局現況	智慧語音助理AI晶片、車用電子	如3D感測，預計將用於人臉辨識	2.5D或3D晶片製造及封裝	如「晶片指紋」創新技術PUF及TPM安全晶片等
產業能量				

挑戰
- 晶片與系統應用介接仍不多，需結合系統廠商進行驗證
- 新技術增加額外成本與設計複雜度，產品接受度易受限

資料來源：Tractica；Synopsys；工研院IEK(2017/06)

簡報大綱

一. 背景分析
- 國際發展趨勢
- 國內發展現況與挑戰

二. 主軸目標與推動作法
- 願景與目標
- 策略作法
- 最終效益(End-Point)

三. 討論題綱

主軸目標與推動作法

願景　促進台灣成為全球AI創新研發樞紐

建立智慧系統與晶片技術，支持跨業整合，促進台灣進入全球AI系統+晶片領先群

三大目標

目標一:建立AI應用關鍵標記資料技術，促進新創發展，落實產業AI化

目標二 :布局下世代AI自主系統產品，落實AI系統產業化

目標三 :開展AI晶片產業，落實AI晶片自主化

推動策略	策略一：健全資料環境	策略二：提升系統智慧化程度	策略三：紮根自主晶片
推動作法	• 深分析 -強化深度學習系統技術 • 加強深度學習發展關鍵，包含數據收集、深度類神經網路、高能效運算 • 固安全 -建構軟硬體安全防護技術 • 強化智慧安全隱私與自動化軟體安全技術	• 健雲端 -建立符合未來所需雲端平台 • 物聯網、巨量資料應用，皆仰賴雲端平台，走向高度軟硬整合，並需提升擴充性、可用性 • 厚應用 -布局利基次系統與服務 • 垂直整合環境感知、智慧終端、深度分析、服務提供	• 興晶片 -發展自主AI晶片/感測晶片 • 促進AI chip、工業感測器、晶片型LiDAR自主化，並強化晶片安全相關研發

策略一：健全資料環境

深分析-強化深度學習系統技術

產業需求

- 需提高DNN model運算速度並降低運算功耗
  - ➢ 提高深度學習模型的訓練速度、準確率，縮短訓練時程與提高硬體間資料傳輸速度
- 需降低DNN訓練時間與所需人力
  - ➢ 較少數據建立的DNN系統或是更快的資料收集與標註，可增加我國資通硬體產業在特定領域的應用與技術門檻
- 大量的標記資料
  - ➢ 以台廠而言，普遍缺乏發展DNN所需之大量資料，且大量資料取樣不易

技術發展方向

- ✓ 高效能DNN 訓練系統/裝置所需技術：
  - ➢ 發展創新訓練系統產品技術(如cluster/appliance)與整合開發環境
- ✓ 高效率DNN資料標記技術
  - ➢ 非監督式(Unsupervised)與增強型(Reinforcement)架構
  - ➢ DNN系統間之協作技術(如Generative Adversarial Network)
- ✓ 開放式資料標記平台
  - ➢ 建立協作平台，提升資料收集等能力

產學研協作(經濟部)

學：前瞻演算法 → 研：DNN系統技術(如整合開發環境、加速學習、學習經驗轉移、深度學習影像辨識技術) → 產：
- DNN設備業者：發展新產品
- 系統廠商：應用嵌入式系統發展新產品
- 資料系統服務：在地資料分析運用

前瞻技術布局(科技部)

人機介面技術、知識探索技術、處理技術、裝置運作技術、辨識技術

策略一：健全資料環境

固安全-建構軟硬體安全防護技術

產業需求

- **惡意軟體數量增長成指數成長**
  - 被動式的防毒軟體已無法滿足主動式惡意入侵防禦需求
- **企業機密資料常漫遊企業可管控範圍外**
  - 機密資料易受不肖雲端提供者(或其員工)及大環境政策改變等因素而外洩，進而威脅到企業的生存
- **以開源碼為基礎發展資安為關鍵**
  - 將使台灣智慧系統與資安系統於國際市場更具公信力

技術發展方向

- ✓ **惡意程式阻絕(白名單方式)**
  - **特定聯網服務所需「零」誤判率攔截之創新白名單作業系統架構技術**
  - **預防性駭客攻擊測試技術**
- ✓ **資料與軟體安全：**
  - **在加密資料上直接做資料處理**
  - **「自動化」程式弱點偵測、修補與攻擊**
- ✓ **加強開源資安軟體的掌握與投入**
  - **成為國際開源社群之貢獻者**

產學研協作(經濟部)

學	研	產
軟體框架、偵測機制	• 高強度軟體資安防護與偵測系統 • 資安OS晶片核心	• 電信營運商：提供企業BYOD整合解決方案 • 服務提供者：高資安環境提供流暢雲端服務 • 系統廠：發展高安全性終端

前瞻技術布局(科技部)

智慧化攻防技術、智慧化分析與生成、智慧化數位鑑識、資安大數據、基礎建設安全

策略二：提升系統智慧化程度

健雲端-建立符合未來所需雲端平台

產業需求

- **深度學習之訓練需先進雲端平台支持**
  - AI+各行業應用亟需大規模雲端服務，其中雲端平台效能需能支撐深度學習訓練資料等運算所需
- **更具彈性的系統架構以因應蓬勃創新服務所需**
  - 傳統的雲端平台，需從使用者需求、軟體系統開發與整合費力部署調試；未來蓬勃創新應用，有賴具更彈性的雲端平台，以降低成本並創造更佳的商業價值

技術發展方向

- ✓ **人工智慧訓練所需可擴充/高效能系統：**
  - **超高速資料中心網路架構**
  - **全快閃記憶體儲存管理**
- ✓ **更具彈性調度與動態擴展之系統技術：**
  - **容器型與虛擬機型混和虛擬化**
  - **公共私有混和雲**
  - **FPGA-based 運算平台與軟體工具**
  - **大量冷資料儲存需求的冷儲存系統**

產學研協作(經濟部)

學	研	產
虛擬化系統軟體技術	• 雲端服務維運技術 • 應用效能管理技術	• 伺服器/儲存/網通業者：建立基礎設施服務及平台服務 • 軟體業者：提供雲端平台維運服務、DevOps服務、雲端服務效能監控服務

前瞻技術布局(科技部)

雲端服務與網路技術、虛擬化技術、記憶體內運算、資料串流處理、雲端效能管理、霧運算

策略二：提升系統智慧化程度

厚應用-布局利基次系統與服務

產業需求

- 建構補足所需次系統、核心軟體與服務平台，結合既有優勢，轉型邁向下世代AI產業

2025年自動駕駛車市場產值估達420億美元

無人機市場2020年達112億美元

工業物聯網市場從2015年至2025年成長7倍

虛擬專家
虛擬客服助理
虛擬理財專員

技術發展方向

- ✓ 發展具利基之次系統與核心軟體系統，如：
  - ➢ 核心軟體：如商用無人機隊執行警政勤務，達到自動化輪值，與高精準避障
  - ➢ 自動駕駛感知次系統：因應未來國際各車廠之共通需求，並適合資通電子產業
  - ➢ 服務型虛擬助理：虛擬理財專員、虛擬客服助理
- ✓ 共通性關鍵模組：如新型辨識與控制模組等

產學研協作(經濟部)

學	研	產
基礎辨識演算法與知識工程	• 軟硬整合智慧辨識與服務系統技術 • 大量資料即時運算技術	• 終端業者：建立裝置聯網與資料管理、資料分析與可視化、平台經營維運等服務，提升為服務/系統提供者 • 服務業者：發展具預測能力之服務(如虛擬理財專員、產線自動排程、虛擬客服助理)

前瞻技術布局(科技部)

網宇實體系統技術、環境感測器、多重感測融合技術、決策及控制演算法

策略三：紮根晶片前瞻技術

興晶片-建立自主AI晶片/感測晶片

產業需求

- 人工智慧運算帶來新的晶片發展契機
  - ➢ 台灣IC設計以數位晶片為核心強項，需降低嵌入式系統晶片進行DNN模型推論演算的功耗及時間延遲，期帶來嵌入式系統晶片的產業加值
- 智慧系統實現，有賴感測訊息精度與多元來源：
  - ➢ 目前所需工業級感測器幾乎皆受制於國外，而車用、無人機等智慧終端亦將帶動新興感測器需求，追求體積小、精度高和成本效益

技術發展方向

- ✓ 低功耗高效能之人工智慧處理器晶片/公版
  - ➢ 聚焦系統運作時之終端推理
  - ➢ 深度學習模型Compiler and Optimizer
- ✓ 晶片安全相關技術：
  - ➢ 例如車規等級高安全性車載晶片平台
- ✓ 感測晶片
  - ➢ 新型光感測器(如晶片型LiDAR)
  - ➢ 工業力、扭力、光、電磁感測器等

產學研協作(經濟部)

學	研	產
DNN深度學習模型最佳化運算	高能效DNN推論處理器晶片技術	• 晶片設計業者：應用處理器、影像處理器、微控制器引擎、高效能Security OS核心 • 系統廠：新型人工智慧終端 • 感測器廠商：消費級感測器切入工業級

前瞻技術布局(科技部)

智慧終端晶片技術(含低功耗深度學習AI晶片架構、低耗能記憶體、環境感知系統、RF收發機晶片等)

最終效益(End-Point)

一. 厚植AI系統產業化技術，推動我國成為AI系統輸出國

- ✓ DNN system 產品: appliance +integrated development environment
- ✓ 先進雲端平台系統 (CPU/GPU/NPU/FPGA智慧運算設施、容器管理系統、應用效能分析系統、混合雲協同運作、超高速資料中心網路系統)
- ✓ 高強度軟體資安防禦系統
- ✓ 控制模組與機器人關鍵模組

二. 完備產業AI化開發平台，建構我國成為全球AI應用新創群聚典範

- ✓ 利基次系統與核心軟體
- ✓ 關鍵標記技術與開放式協作平台
- ✓ AI應用示範案例

三. 建立AI晶片產業鏈，促進我國擠身AI晶片自主國：

- ✓ AI processor/compiler for inferencing
- ✓ 異質整合型光通訊晶片、異質模組整合型晶片、晶片安全核心(Kernel)
- ✓ 新興感測晶片，如：LiDAR、工業感測晶片、辨識晶片等

簡報大綱

一. 背景分析
- 國際發展趨勢
- 國內發展現況與挑戰

二. 主軸目標與推動作法
- 願景與目標
- 策略作法
- 最終效益(End-Point)

三. 討論題綱

討論題綱

- 議題一：發展智慧(AI)系統及晶片技術與產品，應聚焦那些優勢/利基項目?
- 議題二：推動產業AI應用系統，如何強化在地與國際應用之連結?
- 議題三：建立AI晶片產業鏈，如何協助晶片業者參與應用系統與場域之整合驗證?

簡報完畢

BOST

智慧系統與晶片產業發展策略會議

《智慧系統與晶片技術》

引言人

力旺電子　徐清祥董事長

ememory

智慧產業之晶片安全技術

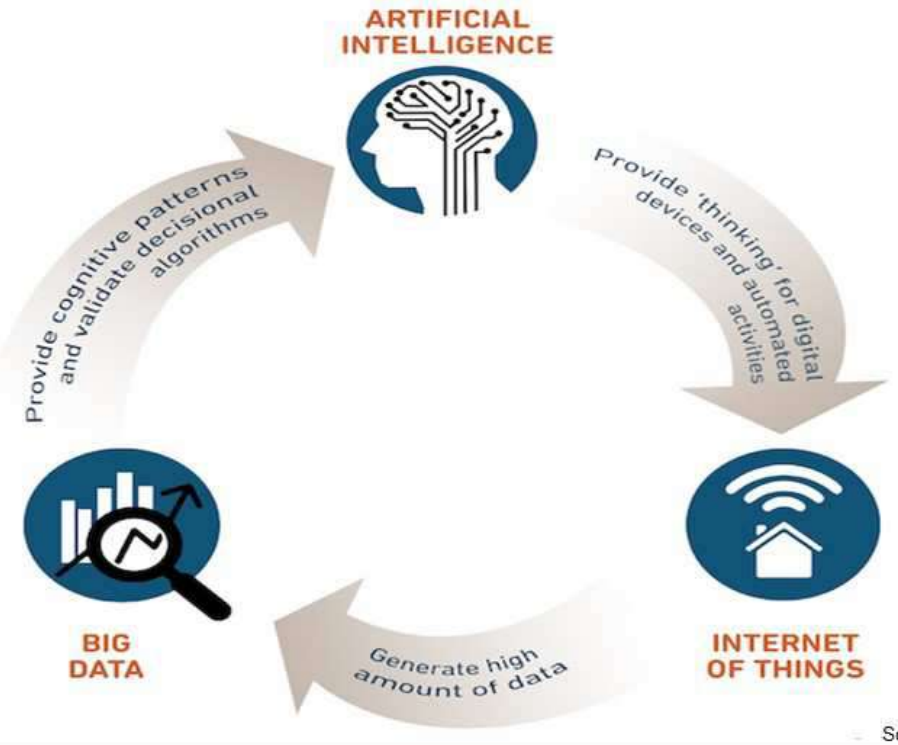
力旺電子 徐清祥

July 11th , 2017

內容綱要

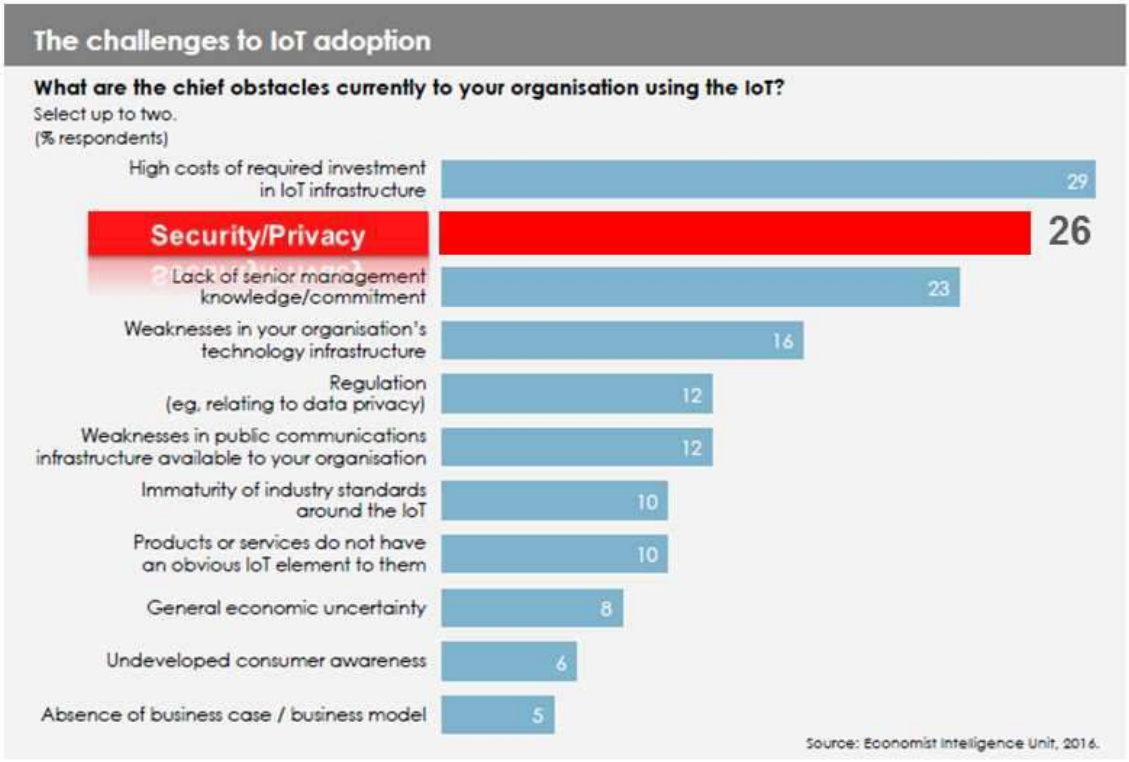
- IoT/AI智慧產業發展趨勢
- 正視物聯網安全問題
- 政府的角色

影響未來世界的三大科技趨勢

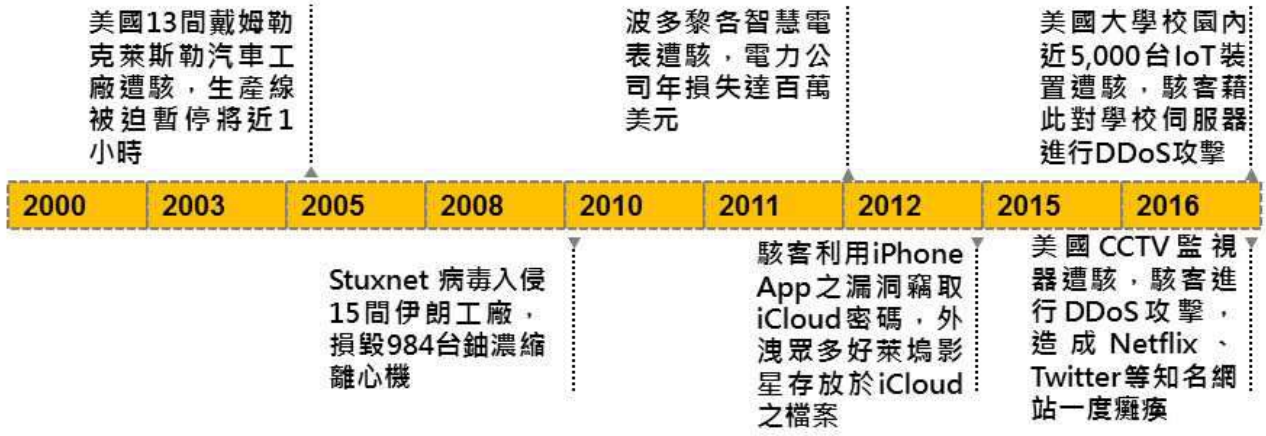


Source: DiploFoundation

物聯網市場發展之挑戰

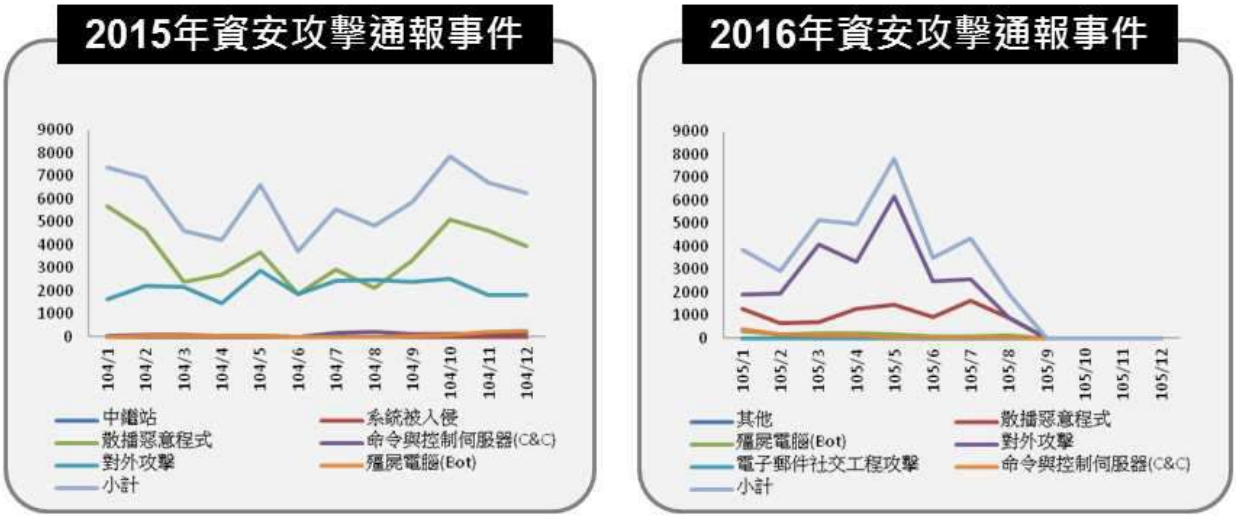


全球重大資安攻擊事件



Source: rcrwireless, campustechnology, tacnetsol, Technewsworld, IBM Security Intelligence, CNN

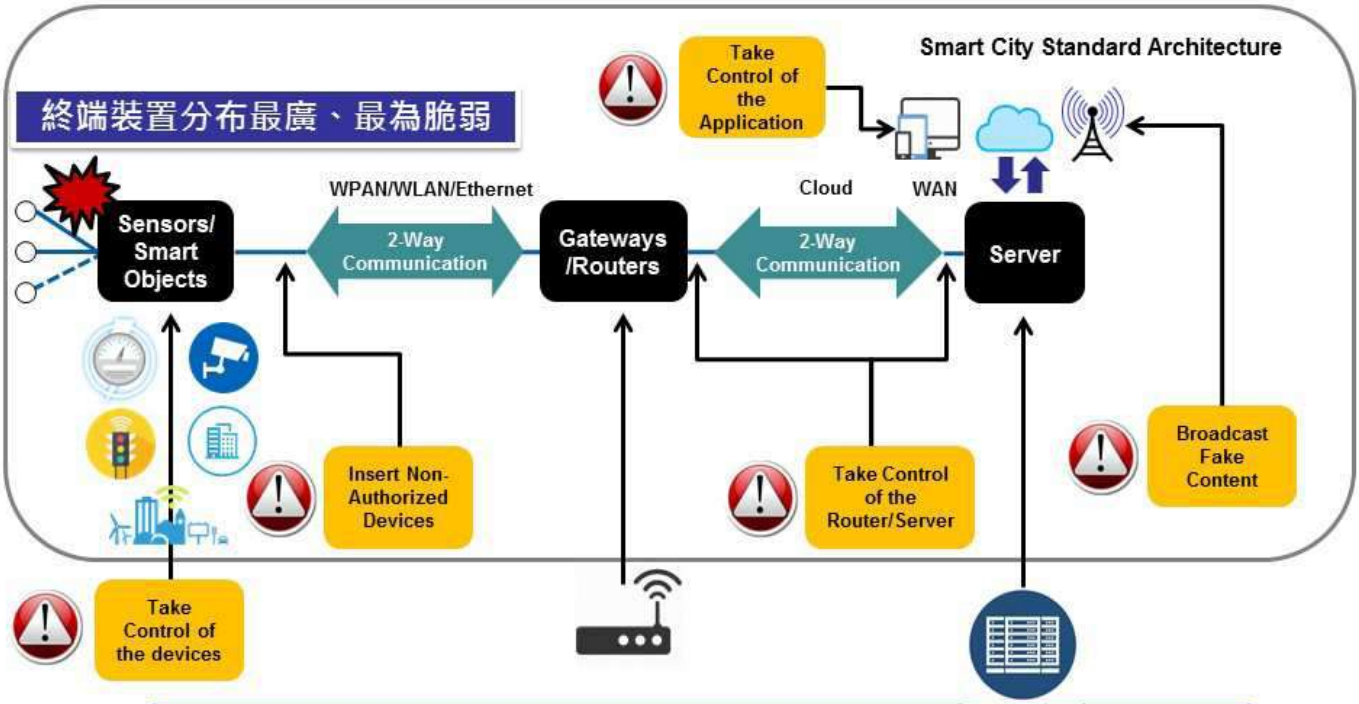
台灣資安攻擊事件



過去兩年, 台灣發生超過 40,000 件資安攻擊事件

資料來源:國家通訊傳播委員會,台灣電腦網路危機處理暨協調中心, TrendMicro, Tacert

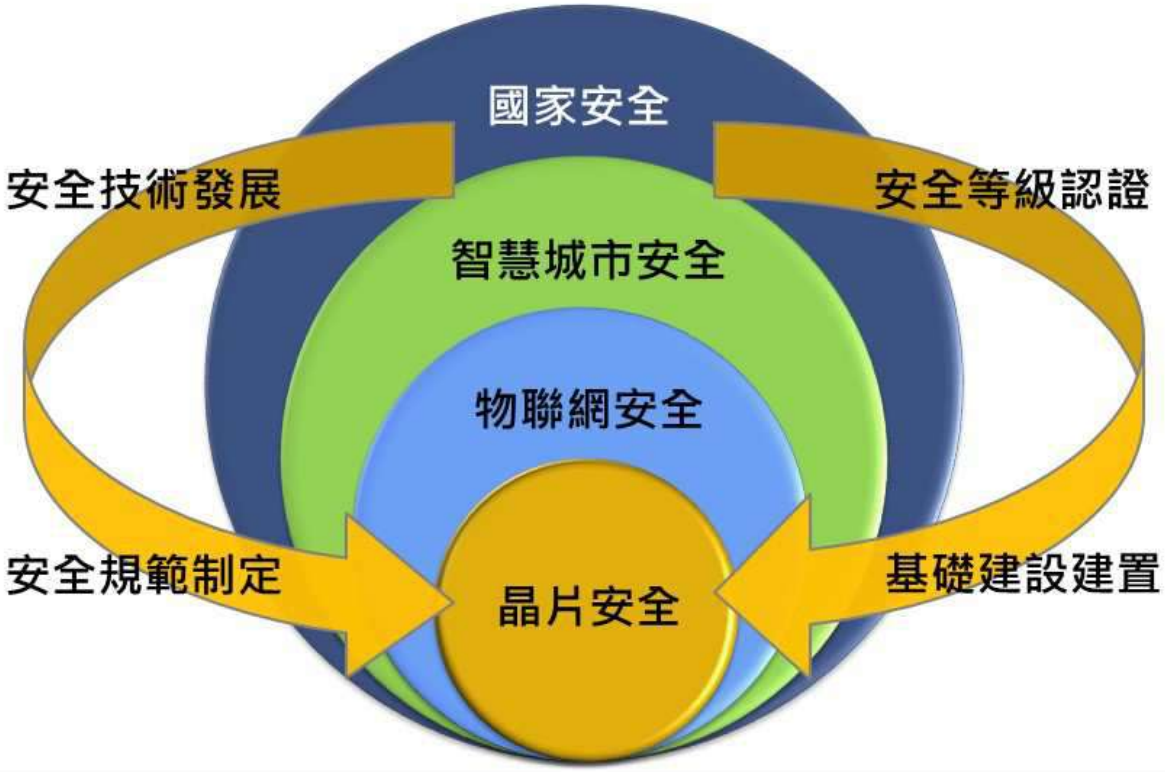
物聯網安全威脅無所不在



物聯網安全為智慧城市發展之基礎

Source: Inside Secure, Incipio, Bolt, Seeicons,iconfinder, SAP Hana, Libe Riot, Comodo, Keywordsuggest, Pcmode, Github, Newdesignfile

晶片安全攸關國家安全



政府應強力主導建立安全的物聯網環境

- 建置安全的物聯網基礎建設
- 制定物聯網產業鏈安全規範
- 獎勵投入晶片安全技術之產業
- 設立安全認證中心

Embedded Wisely, Embedded Widely

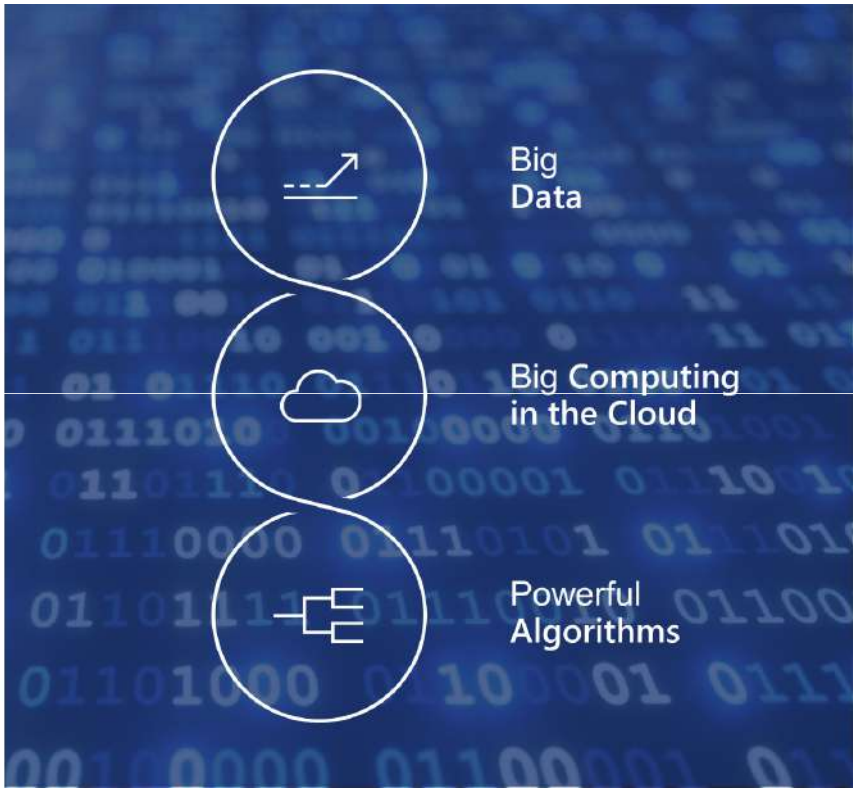
智慧系統與晶片產業發展策略會議

《智慧系統與晶片技術》

引言人

微軟研究院　潘天佑副院長

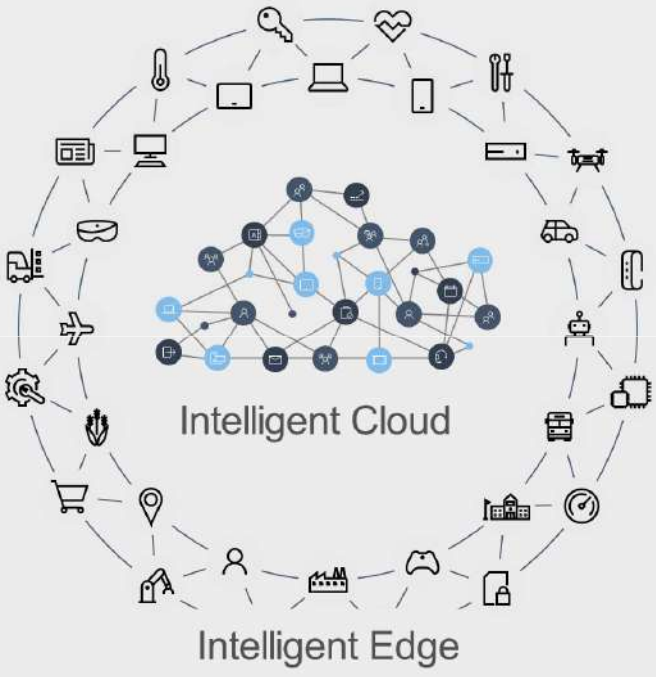
Why AI?
Why now?

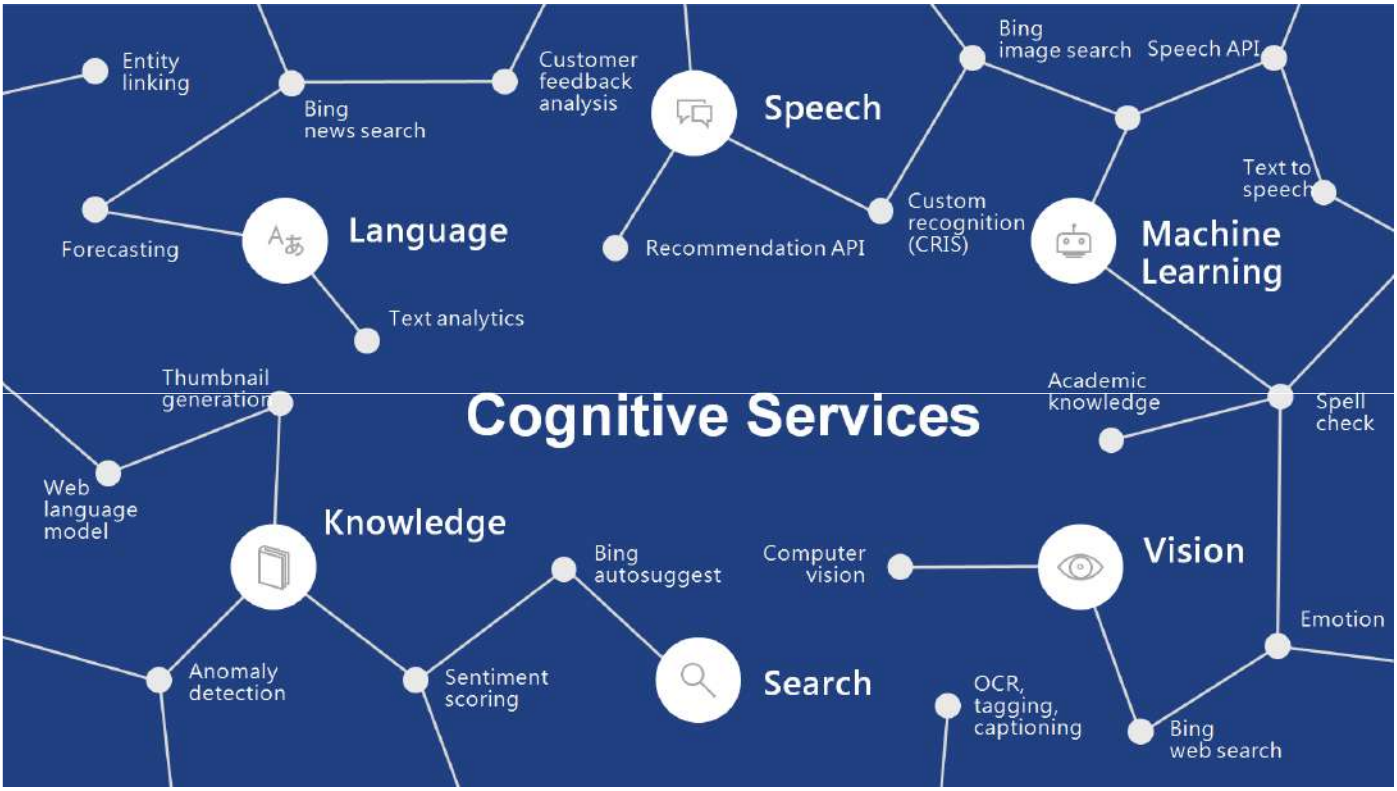


Multi-device

Artificial Intelligence

Serverless





Thoughts about AI

- Security – managed or controlled?
- Innovation – connected or segregated?
- Data – shared or reserved?

Opportunities of AI

- Advance public services with AI.
- Democratize AI for all businesses.
- Build intelligent edge industry.

智慧系統與晶片產業發展策略會議

《智慧系統與晶片技術》

引言人

國立清華大學　吳誠文教授

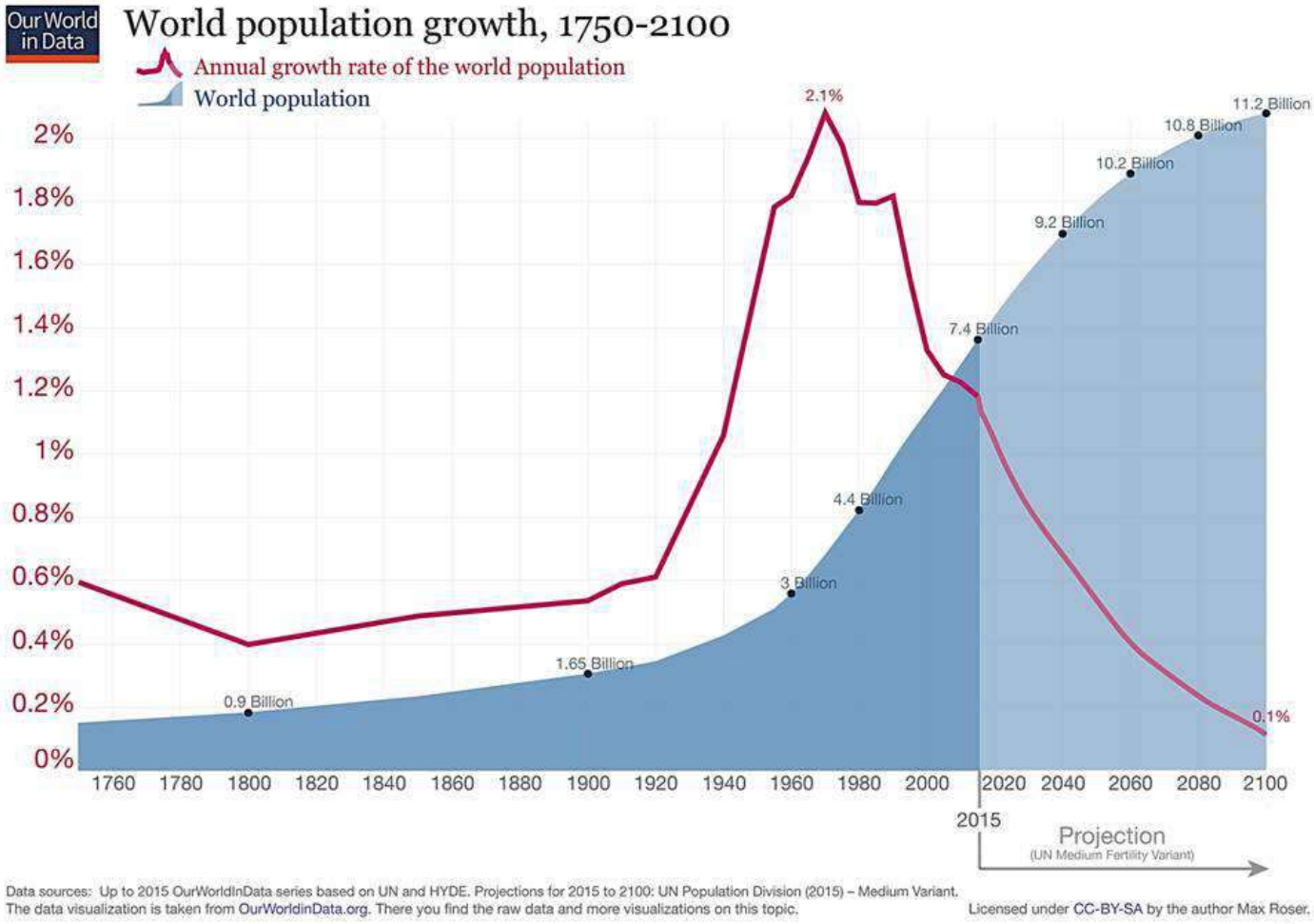
智慧系統與晶片產業發展策略會議

Leveraging AI/IOT By Semiconductor

Cheng-Wen Wu (吳誠文)

07/11/2017

國立清華大學
NATIONAL TSING HUA UNIVERSITY



IOT Growth Prediction Adjusted

- In 2011, Cisco predicted 50B things to be connected by 2020
- In 06/16, Cisco adjusted the 2020 estimation
  - ✓ 50B → 26.3B
- Imagine a world with connected devices 2-3 orders of magnitude more than today
  - ✓ And the huge amount of data they generate
- How can it be possible without dramatic increase in **spending** or **energy** supply?

◆ Grand challenges:

- **Fast system design and manufacturing**
- **Cost/energy consumption containment**

David Patterson Interviewed by CNBC

- Four years ago, Google worried that if every Android user had 3 minutes of conversation translated a day using machine learning, they'd have to double their data centers
- Alphabet is spending $10B a year on capital expenses, largely tied to Google data centers
- TPU is running 15-30x faster and 30-80x more (energy) efficient than conventional processors

Source: www.cnbc.com, 5/6/2017

New Territory for Chip Makers

- A GPU company is currently the market leader in deep learning neural network platforms
  - ✓GPU-based DNN accelerators are installed in major cloud datacenters
  - ✓And, most autonomous vehicle AI platforms
- A leading Semiconductor company has been making FPGAs as AI accelerators
  - ✓It has also invested in ASIC technology (by M&A)
  - ✓And opened Advanced Vehicle Labs
- Dozens of competing AI chip startups have emerged in the past two years
- More solutions will evolve in the future

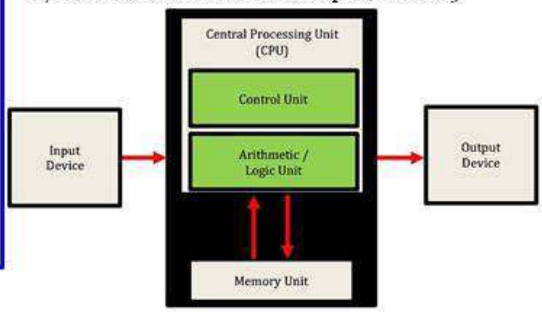
Deep Learning and Deep Neural Network

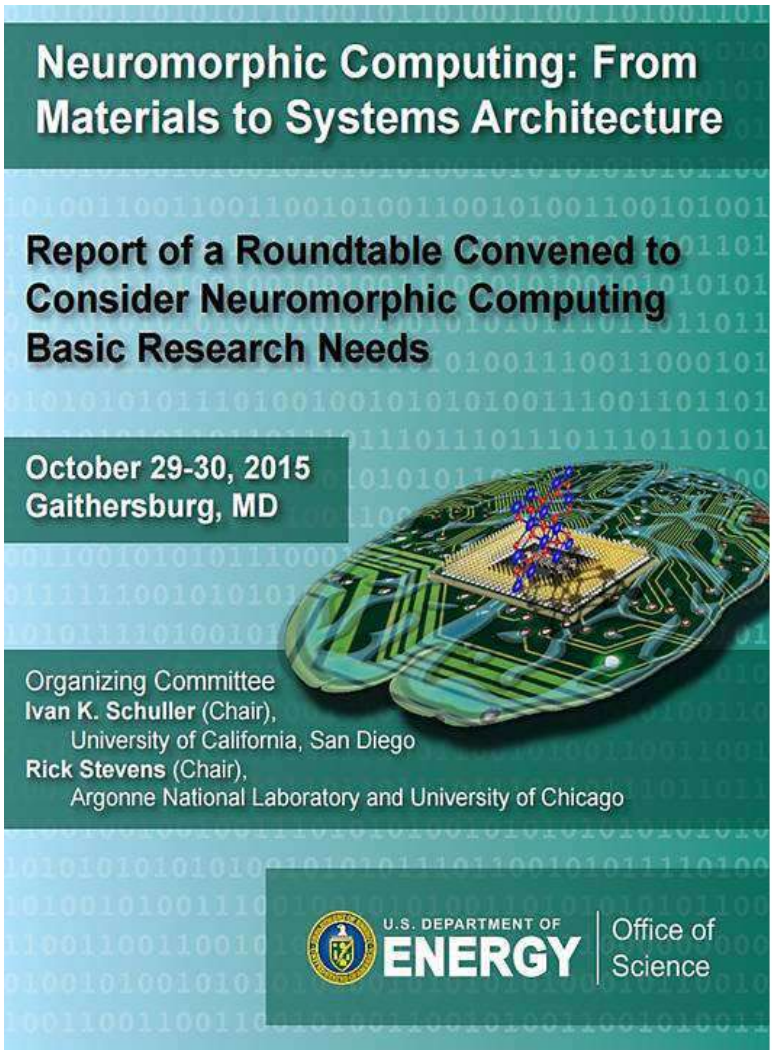
- Neural network inspired machine learning
  - ✓Training and inference
  - ✓Energy efficiency increased by orders of magnitude
    - ➢ Neuromorphic architectures: 6 orders or more expected
- ➢Key success factors:
  - ✓Big (effective) training data
  - ✓Scalable DNN models
  - ✓Efficient hardware

1. ***Simulating*** neural function in software on conventional von Neumann hardware
2. ***Emulating*** neural function in dedicated ANN hardware

von Neumann Architecture

Source: Neuromorphic Computing: From Materials to Systems Architecture. US DOE Report. Oct. 2015





- ➢ Development of novel materials and devices incorporated into unique architectures will allow a revolutionary technological leap toward a fully "**neuromorphic**" computer
- ➢ New neuromorphic architectures and novel devices needed to produce **lower energy consumption** and **enhanced computation**

Federal Vision for Future Computing

[White paper released by DOE, NSF, DOD, NIST, IC (July 29, 2016)]

- Emerging computing architecture platforms, neuromorphic, quantum, etc.
  - Significantly enhancing performance while reducing energy consumption by over 6 orders of magnitude (from MWatts to Watts)
- Intelligent big data sensor: autonomous and reprogrammable
- Machine intelligence for scientific discovery
- Cybersecurity

Final Reminder

- This is a finite world, with limited capacity and resources
  - ✓ Industrial growth limited by economy and energy scales
- It's all about replacement
  - ✓ You replace others (live) or be replaced (die)
- AI will be everywhere, with IOT
  - ✓ Applications and services taking advantage of DL
- Business opportunities from
  - ✓ Training data
  - ✓ DNN models
  - ✓ Efficient Semiconductor
    - ◆ Neuromorphic computing eventually
- Taiwan: Leverage AI/IOT by semiconductor

智慧系統與晶片產業發展策略會議

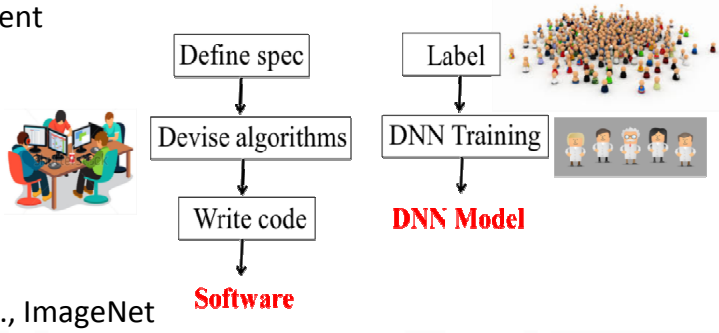
《智慧系統與晶片技術》

引言人

工業技術研究院　關志克所長

Introduction to AI System

- AI ~ Deep Neural Network (DNN)-based Machine Learning
- AI system: a system with analysis and synthesis capabilities powered by DNN-based models running at the edge or in the cloud
- Machine learning: a universal algorithm for building a functional mapping between sample inputs and associated outputs
  - A new paradigm of software development
  - Learn from many normal people vs. Design with few gifted experts
- From AI Winter to AI Everywhere
  - Algorithmic breakthrough that enables training of deep neural network
  - Large high-quality training data set, e.g., ImageNet
  - Availability of high-performance GPU

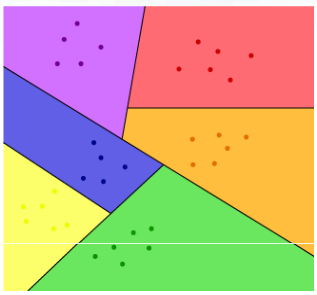


Overarching Strategies

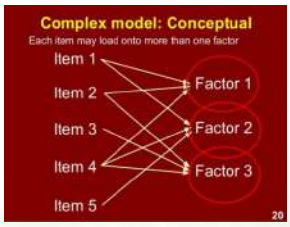
- 產業AI化: Apply modern AI techniques to improving the value and efficiency of existing industry segments
  - Medicine: diagnostics, drug synthesis, smart hospital
  - Manufacturing: defect detection, equipment maintenance, robotic manipulation
  - Finance: credit assessment, personal investment, trading algorithm
  - Commerce: advertisement, retail analytics, logistics planning
- AI產業化: Convert modern AI techniques into new systems and products that enable applications of AI
  - High-performance DNN training
  - Real-time low-power DNN inferencing
  - DNN-based systems: autonomous driving vehicle, autonomous drone, robot, personal virtual assistant

Machine Learning Basics

- **Supervised Learning:** from sample input-output pairs
  - Labeling a training data set ➔ knowledge acquisition
  - Training to get a functional model ➔ knowledge transfer & abstraction
  - Applying a learned model ➔ knowledge application
  - Ask the right question: Set up a proper optimization objective function: "Like this?"
  - Training corresponds to multi-variable non-linear optimization
    - Universal Approximation Theorem
    - Gradient descent-based search



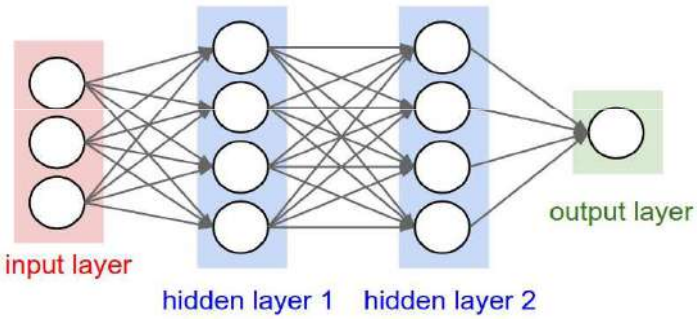
- **Unsupervised Learning**
  - Clustering
  - Factor analysis
  - Auto encoding

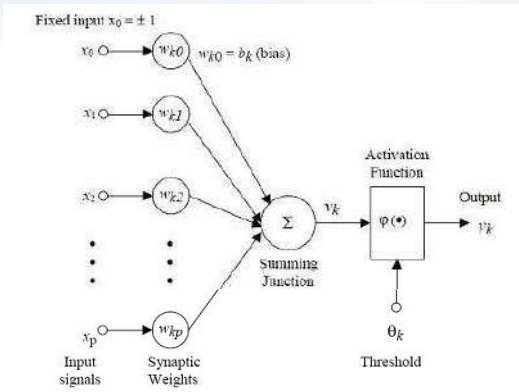


工業技術研究院 Industrial Technology Research Institute

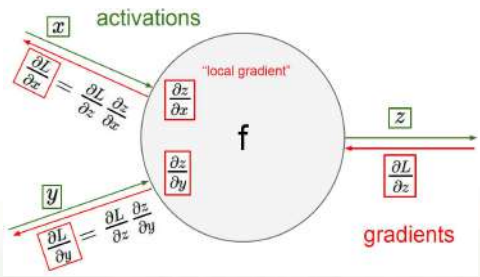
Training and Inference of Neural Network

Training: Backward Propagation ➔





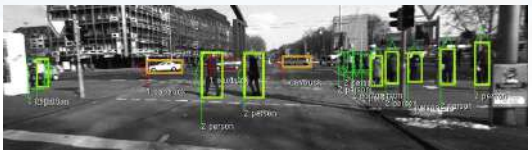
Inference: Forward Propagation ➔

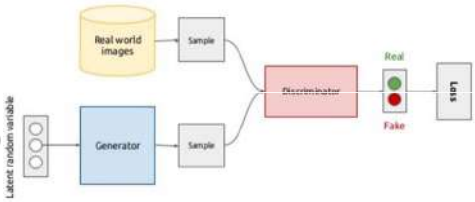


工業技術研究院 Industrial Technology Research Institute

Direction 1: DNN Model Training

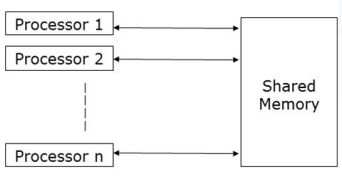
- DNN model is software program of the future
- **Model quality**
  - Innovative DNN architectures and training techniques
    - Residual, Reinforcement and One-shot learning
  - Cost-effective acquisition of high-quality training data set
    - Semi-automatic training data collection and labeling
    - Automatic generation: Generative adversarial network (GAN)
    - Remove the need for training data: Unsupervised learning
    - Examples: (1) Retail spending of individual persons and families, (2) Up-to-date street views for all for-driving roads, (3) Food orders in restaurants by every individual or family, and (4) Commute trajectory of every driving commuter
- **Training speed**
  - Minimize the number of rounds required in the training process
    - Integrated development environment for DNN training
  - Minimize the computation overhead associated with each training round
    - High-performance distributed DNN training appliance: Nvidia's DGX-1



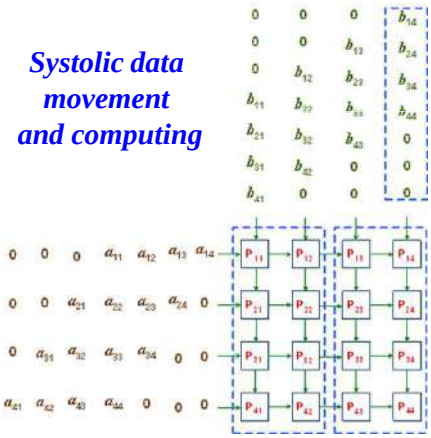


Direction 2: DNN Inference Engine

- Apply DNN model using specialized processor
- Cloud-based DNN inference processor
  - Example: Google's TensorFlow processing unit (TPU)
- **Embedded DNN inference processor**
  - Low-power, low-cost and real-time
  - Example: Nvidia's Xavier, Intel's Movidus, and KAIST's CNNP
  - Systolic array architecture for memory access reduction
  - Make the easy case fast: Big/Little
  - Incremental computing: Make the already seen case fast
- **DNN compiler: DNN model ➔ executable code**
  - DNN model compression to reduce model complexity
  - Memory access scheduling to maximize pre-fetch and reuse
- **Next major battleground: Vehicle computing platform for autonomous driving**
  - Advanced sensor IC: e.g. solid-state radar and LiDAR

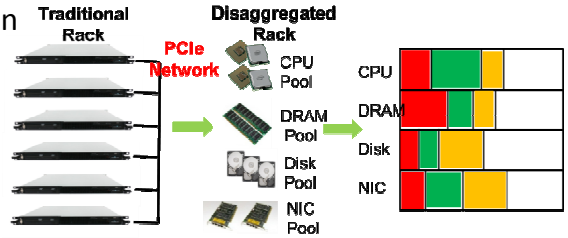


Systolic data movement and computing



Direction 3: Scalable Cloud Data Center

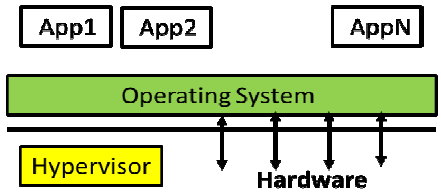
- DNN training will become a major workload of future cloud data centers
- **In-memory computing**
  - Heterogeneous: CPU, GPU, FPGA and smartphone SOC
  - FPGA as a DNN training platform: Compiling program into circuit
  - Container-based and hypervisor-based virtualization
  - Disaggregated architecture
- **DRAM-speed storage**
  - All flash array and 3D Xpoint: SAS ➔ NVMe
  - NVDIMM and beyond



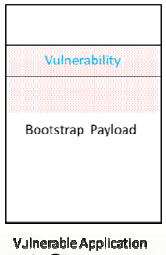
- **Very-high-performance networking**: Network is becoming the bottleneck
  - Switch/router = network OS + merchant switch IC
  - Software-based and user-programmable switch and NIC
- **Cold storage**: How to cost-effectively store and manage ever growing data that will never be deleted

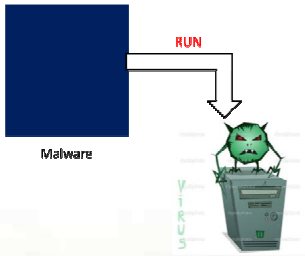
Direction 4: Analysis-based Cyber Security

- AI systems are network connected and thus prone to cyber attacks
- **White-list based cyber defense:** Only allowed programs are permitted to run
  - Why? No. of malwares >> no. of goodwares
  - Programs in the form of executable binaries, scripts, dynamically linked libraries, kernel modules, and macros embedded in documents
  - Checks on program inputs and configuration files
  - Tamper-proof white-list enforcement even in the presence of kernel rootkits



- **Program analysis-based software security**
  - Given a buggy program, automatically find out
    - Where is the bug? Is it a vulnerability?
    - How to exploit it?
    - How to patch it?
    - How to develop an intrusion detection signature for it?

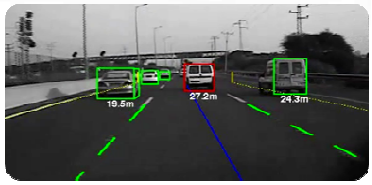




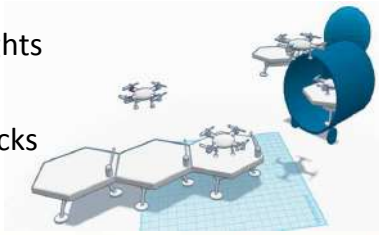
- **Encryption domain data processing**

Direction 5: Human-Competitive AI Systems

- Develop commercially compelling advanced AI systems using technological building blocks created in Directions 1-4



- **Autonomous driving vehicle (ADV)**
  - Surrounding sensing, Driving decision making and Vehicle control
  - Real-time video analysis for object recognition and tracking
  - Multi-sensor fusion and integration
  - Semi-automatic ADV training data collection for Taiwan: 100,000 KMs
- **Self-flying long-service-time drone fleet**
  - Autonomous take-off, landing, hovering and flying during the nights
  - Coordinated turn-taking with self-charging
  - Real-time video analysis for bridge/electric tower inspection checks



- **Personal virtual assistant**
  - Voice-based natural language interactions
  - Senior citizen care based on living quality measurement and analysis

工業技術研究院 Industrial Technology Research Institute

Summary

- This AI renaissance is largely attributed to advances in DNN
- Myriad applications of DNN technology are practical and compelling
  - Autonomous driving car, self-flying drone, industry/service/investment robot, etc.
  - AI applications for business/enterprise problems are promising: legal document analysis, patent application analysis, medical record and literature analysis, etc.
  - From analysis to synthesis
- Systems enabling DNN training and inference are also commercially promising
  - Competition in DNN inference processor design is like the GPU competition circa 2000
  - Compiler able to perform DNN model analysis and transformation plays a crucial role
  - Data center and customized appliance tailored to large-scale DNN training
- On the theory front, our understanding of why DNN is so effective in certain application areas is still unsatisfactory: more tricks than theories
  - Reminiscent of our limited understanding of human brains
  - Continual innovations in DNN architectures: InceptionNet, ResidualNet, etc.

工業技術研究院 Industrial Technology Research Institute

Thank You!

Questions and Comments?

tcc@itri.org.tw