
IBM
– 2018 IBM 科技論壇 –
迎戰未來
機器學習與架構平台的年度盛宴

– 2018 IBM 科技論壇 –

迎戰未來

機器學習與架構平台的年度盛宴

IBM

釋放AI伺服器與機器學習潛力

李永輝

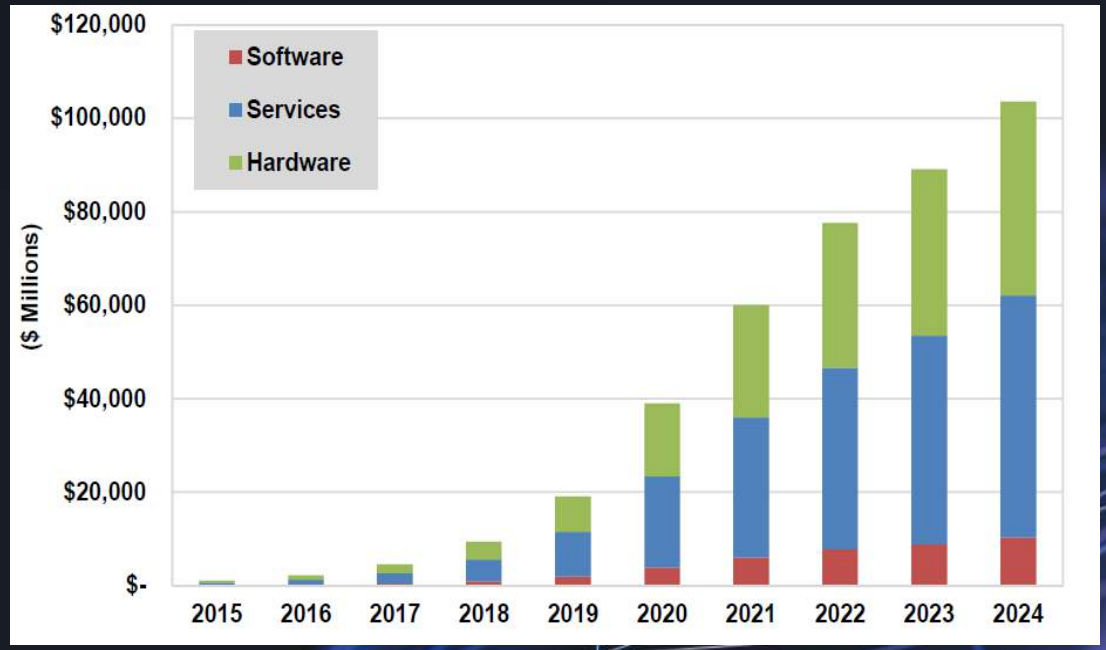
IBM大中華區硬體系統部首席技術官暨傑出工程師

勢不可擋的AI市場

IBM

*"By 2020, 80% of Big Data and Analytics deployments will need distributed micro analytics and 40% of all business analytics software will incorporate prescriptive analytics built on **cognitive computing** functionality. Both of these trends require a dramatic increase in processing power that could be enabled by GPUs."*

— IDC



組成AI的三個元件

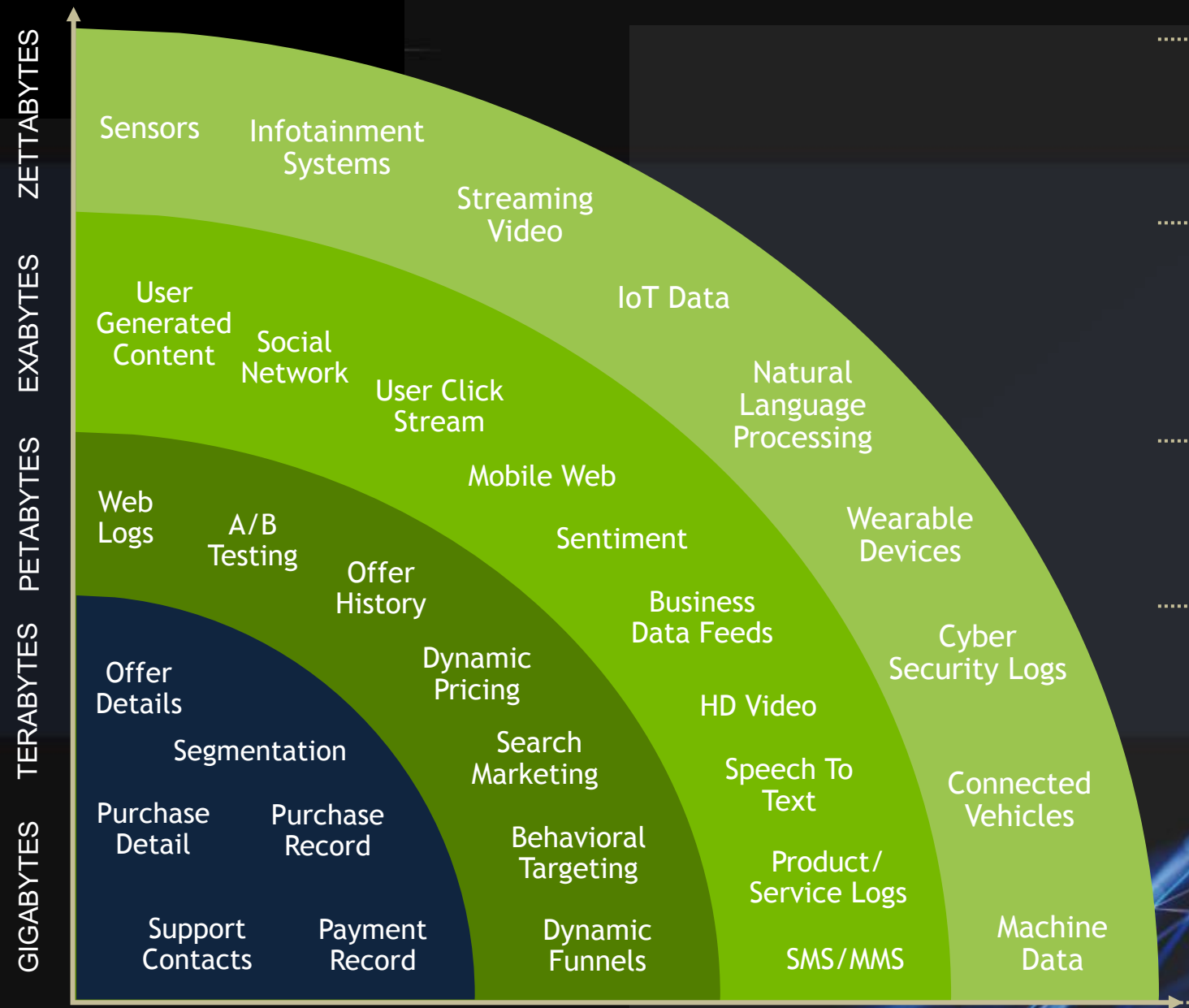
IBM

Data

Compute

Algorithms

各行各業的數據變換



IBM

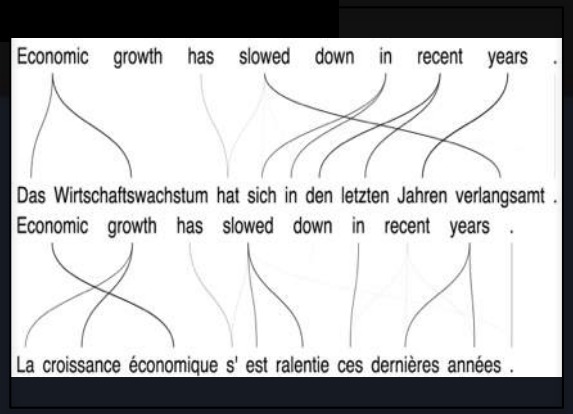
AI

DIGITAL

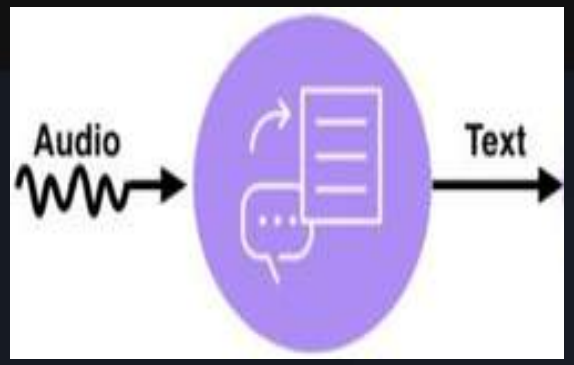
WEB

BUSINESS PROCESS

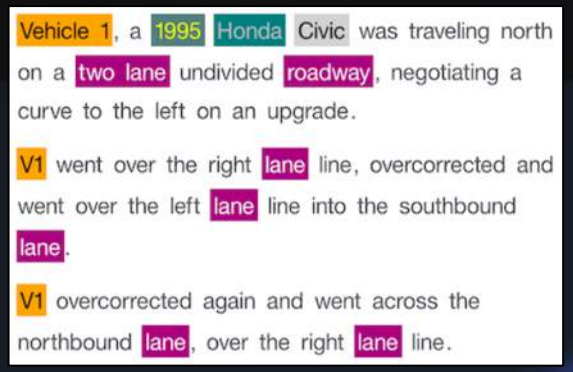
釋放Data的價值



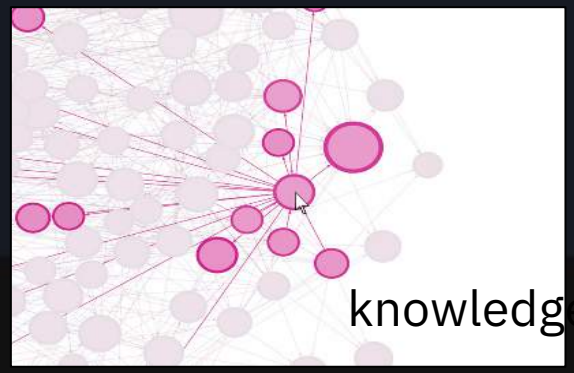
Language Translation



Speech Transcription



Language Understanding

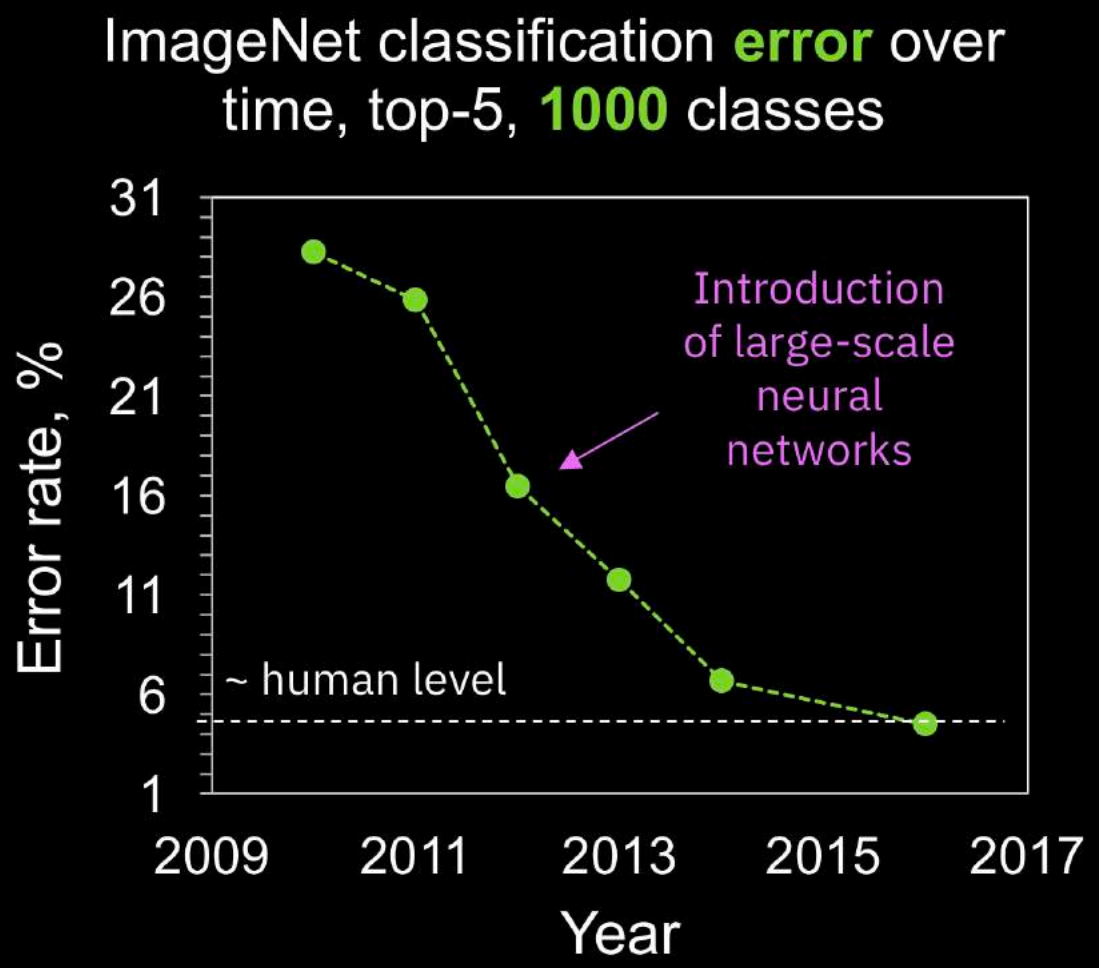


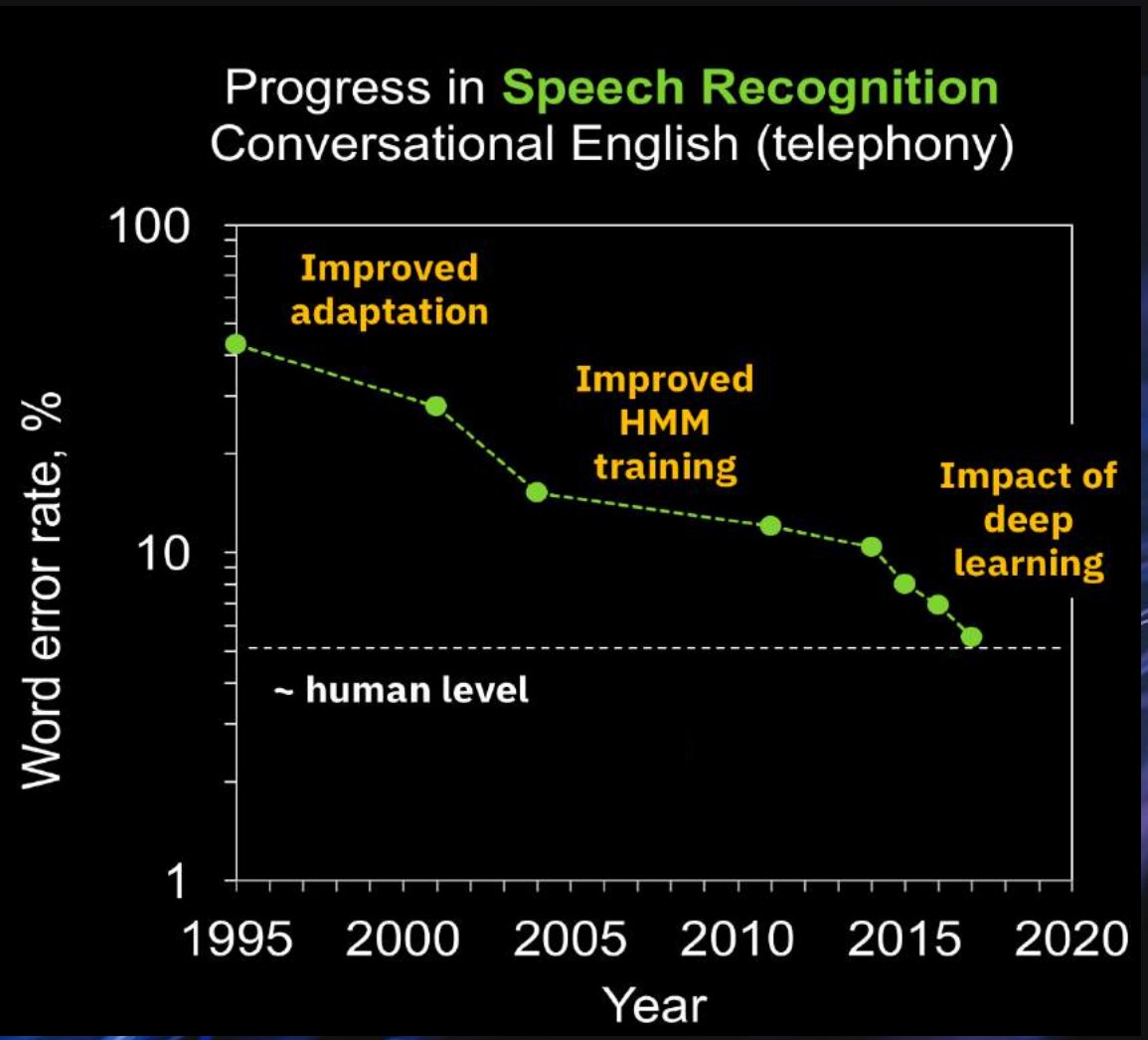
Machine Reasoning

Object Detection

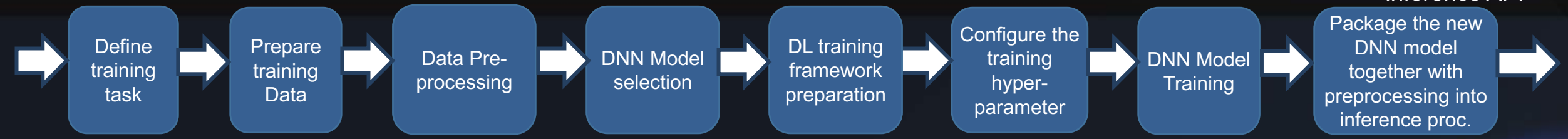
Face Recognition

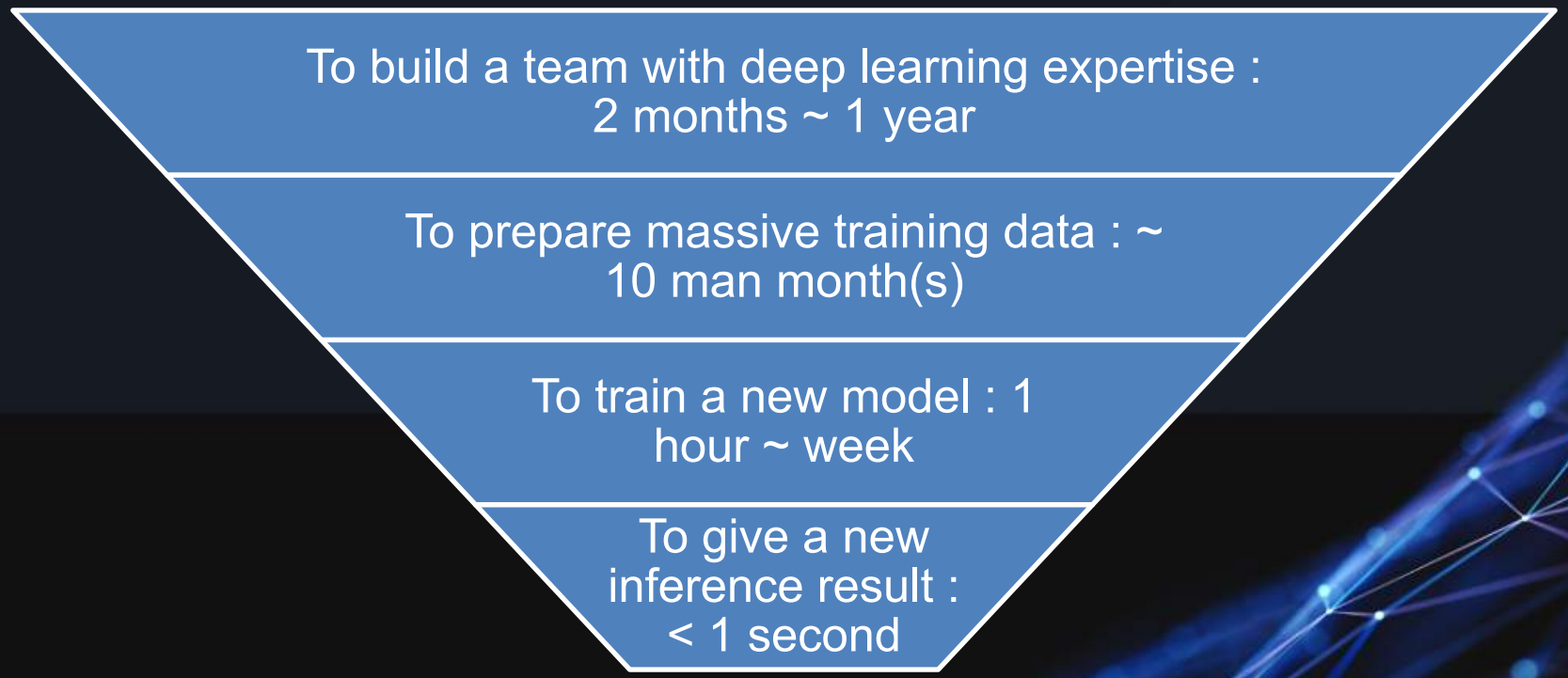
科技案例進展


ImageNet classification error over time, top-5, 1000 classes
Error rate, %
31
26
21
16
11
6
1
Introduction of large-scale neural networks
~ human level
2009
2011
2013
2015
2017
Year


Progress in Speech Recognition
Conversational English (telephony)
Word error rate, %
100
10
1
Improved adaptation
Improved HMM training
Impact of deep learning
~ human level
1995
2000
2005
2010
2015
2020
Year

AI訓練的數據處理





Cognitive System:

Provide optimized SW + HW design, and tool chains to significantly enhance

- Productivity
- Performance
- Time to market

具備URLI的智能系統

IBM

Understand

Natural Language Processing (NLP)

Unstructured Information Management

Reasoning

Knowledge Representation and Reasoning

High Performance Data Analytics

Learning

Deep Learning / Machine Learning

Image / Video / Voice Recognition

Interactive

Text To Speech (TTS) / Speech to Text (STT)

Question Answering Technology

組成AI的三個元件

IBM

Data

Compute

Algorithms

邁向 AI 之路

IBM Systems 的解決方案專為顛覆當今最高級的資料應用而設計，其中不僅包括您目前所運行的任務關鍵應用，還包括下一代 AI 工作負載。

任務關鍵工作負載 | **大資料工作負載** | **企業 AI 工作負載**

Z14 伺服器

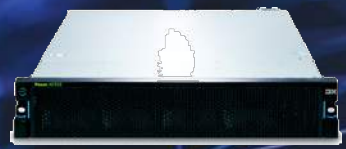
LinuxONE 伺服器

Power Systems UNIX 伺服器

彈性存儲伺服器 (ESS)

向外擴展伺服器（面向 NoSQL/Hadoop）

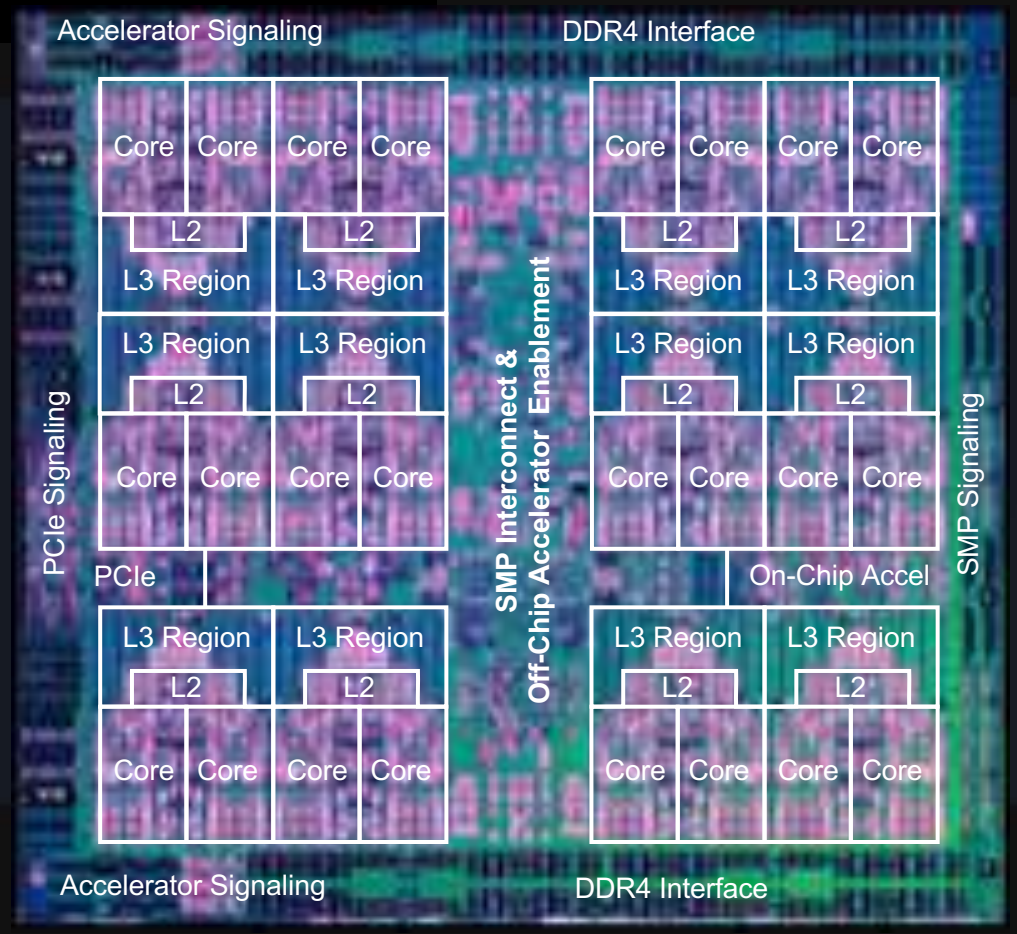
CPU + GPU 伺服器（面向 AI/HPC）



核心基礎架構 → **下一代 AI 工作負載**

全新 POWER9 處理器

專為認知業務而設計



- 14nm finFET 半導體處理器
- 80 億個電晶體、17 層金屬堆疊
- 120 MB eDRAM、12 x 20 路關聯區域
- 芯片上頻寬達 7 TB/s

Improved Thread Performance with SMT8/4, Shorter Pipeline

I/O Throughput Increases 61% @ 366GB/s

Accelerator : PCIe Gen4, CAPI 2.0, OpenCAPI, NVLINK 2.0 & on-Chip

Energy Efficiency Improves 45%* & Workload Optimized Frequency

POWER9 的創新加速技術

NVLINK 2.0 CPU-GPU 連接

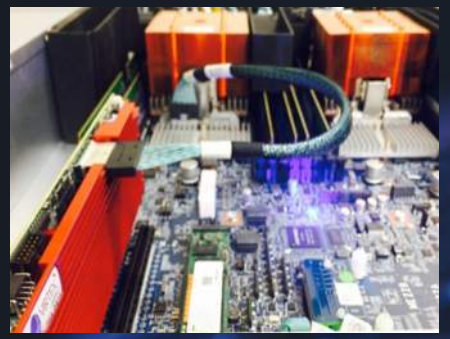
- 最大輸送量高達 300 GB/秒
- CPU-GPU / GPU-GPU
- 統一記憶體訪問
- DL、HPC、GPU DB 的理想之選

PCIe Gen 4 Industry First

- 2 倍輸送量
- 後向相容
- I/O 週邊設備連接
- 實際 I/O 標準

CAPI 2.0 Enhancement

OpenCAPI Collaboration



- 一致性加速處理器介面
- 計算加速（機器學習、視頻、生物資訊）
- 存儲加速（記憶體中資料庫、存儲演算法）
- 網絡加速（壓縮、加密）

IBM與客戶、合作夥伴的POWER9創新案例

Google

“谷歌對 IBM 能夠在最新 POWER 技術研發上取得進展感到非常高興。POWER9 OpenCAPI 匯流排及大記憶體功能為穀歌資料中心的創新帶來了新的機會。

BART SANO
谷歌平臺副總裁

NVIDIA

“IBM 致力於為客戶交付高性能的解決方案，例如適於 NVIDIA® Tesla V100 GPU 和 NVLink 的 POWER9 解決方案，旨在説明客戶加速深度學習工作負載。

IAN BUCK
NVIDIA 總經理、副總裁
加速計算

Limelight NETWORKS

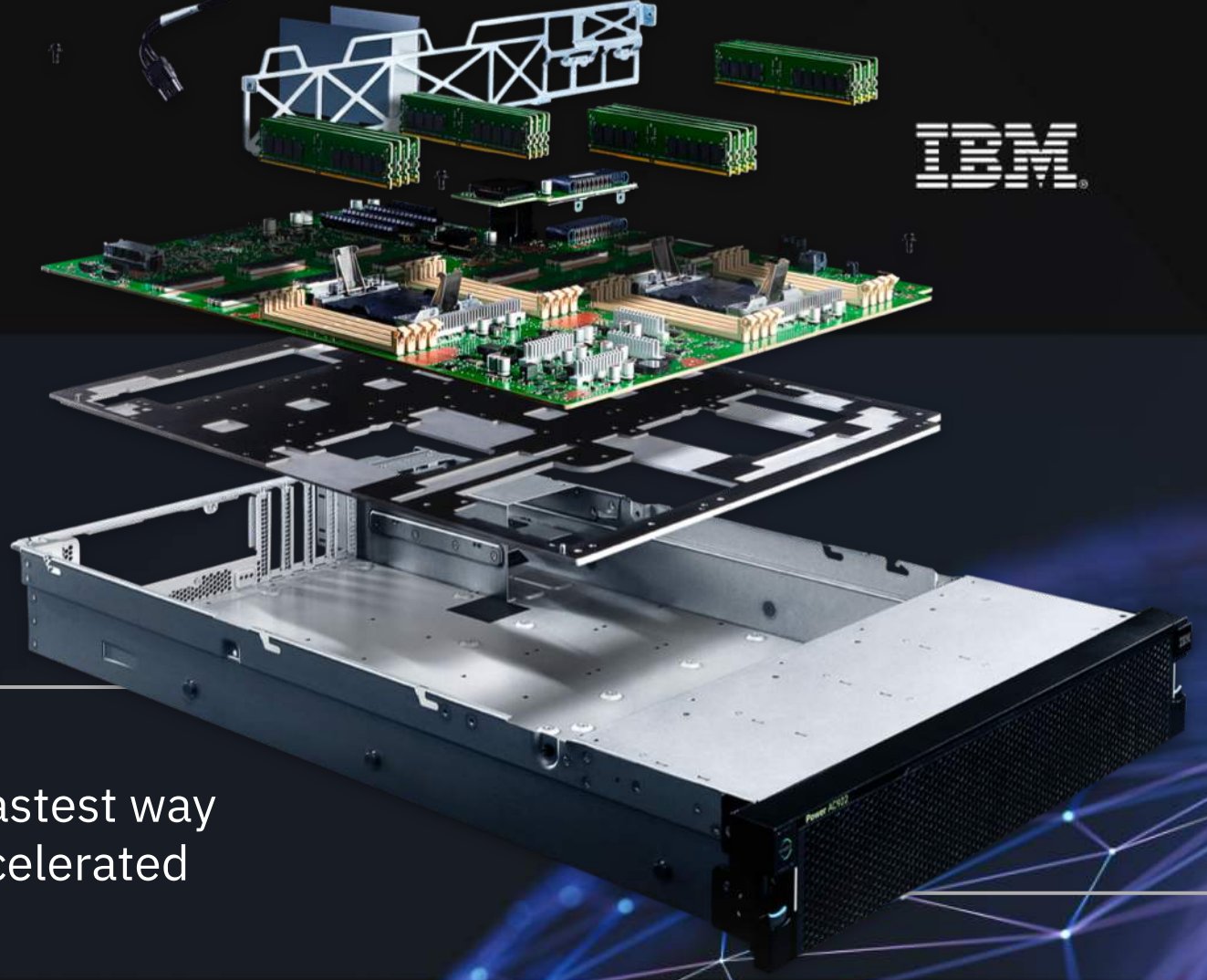
“LimeLight 是一家致力於為客戶（如 BBC 和 Marvel Comics）提供各種各樣的高效工具，説明他們提升數位內容流處理水準的公司；該公司表示，OpenPOWER（以及基於 POWER9 的 PCIe Gen4）幫助他們克服了 PCIe Gen3 與其他伺服器一同使用時常常會出現的瓶頸，不僅加快了流處理速度，還縮短了緩衝時間。

OpenPOWER 2018 高峰會文章

https://www.forbes.com/sites/patrickmoorhead/2018/03/19/headed-into-its-fifth-year-openpower-has-momentum-into-the-power9-generation/2/#74af40ef795a

IBM POWER SYSTEMS
AC922

The best server of enterprise AI

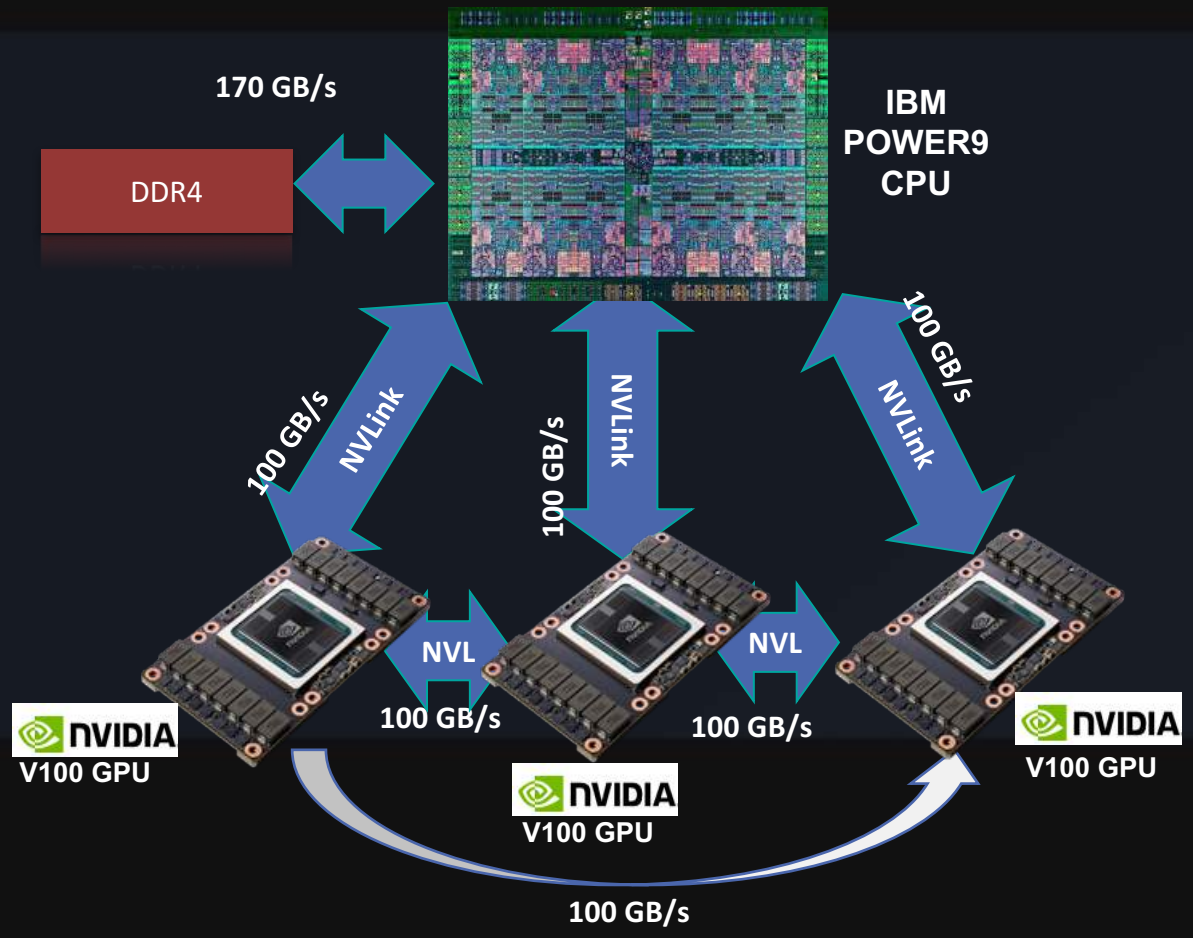


The IBM Power Systems AC922 offers the fastest way to deploy deep learning frameworks and accelerated workloads – with enterprise class support.

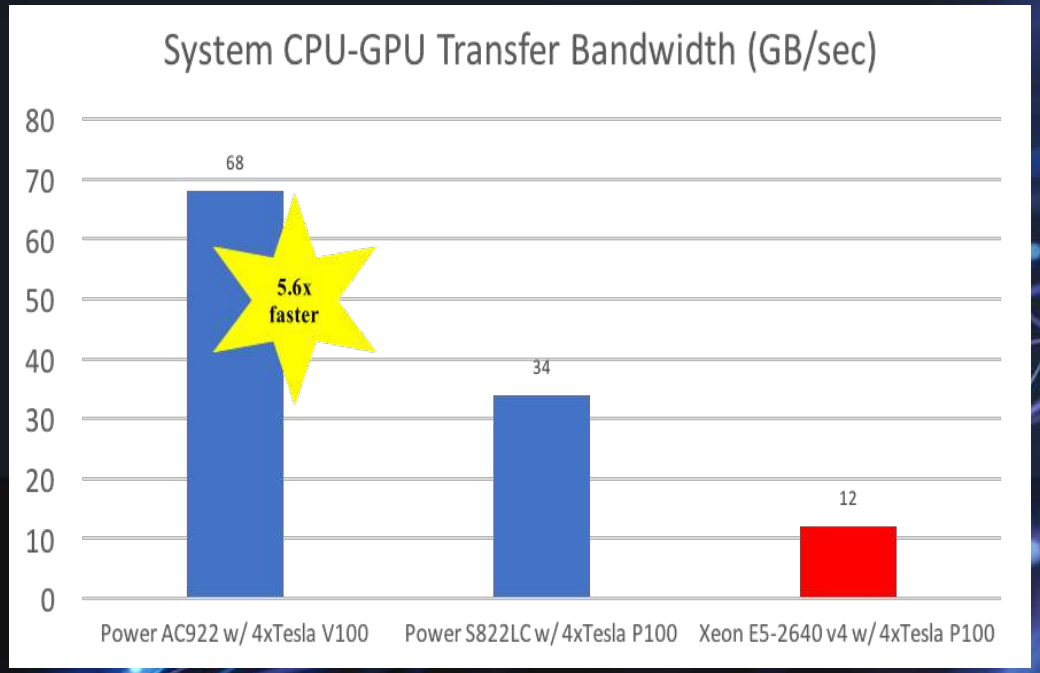
Up to 3.8X
reduction in AI model training for deep learning frameworks

1.8X
better performance of accelerated databases

新一代的 NVLINK GPU 加速



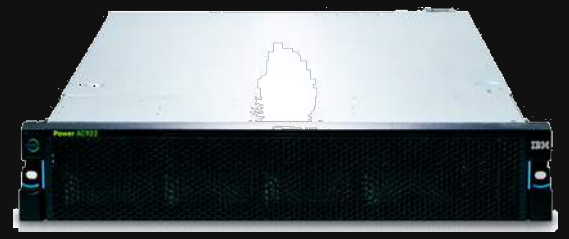
IBM POWER9™ 和 NVLink 2.0 有助於解決您在代碼方面的 PCI-E 瓶頸；相比參與測試的 x86 平臺的 CUDA 主機設備頻寬，可將資料傳輸速度提升 5.6 倍。POWER9 是市場上唯一一款面向 NVLink 2.0 而推出的處理器，從 CPU 到 GPU 均是如此。



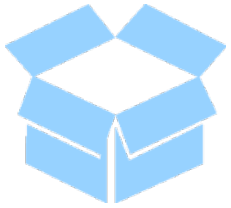
Source - https://developer.ibm.com/linuxonpower/perfcol/perfcol-technical/

世界上最快的超級電腦 IBM Power AC922

- POWER9 CORAL 系統 – Summit：橡樹嶺國家實驗室 (ORNL)
- 計算速度超過 150 Peta FLOPS，將會成為全球速度最快的超級電腦之一
- 相比 Titan，僅需四分之一的節點，便可實現 5-10 倍的應用性能提升
- ~3,500 x IBM Power AC922 (POWER9 + NVIDIA V100 GPU + Mellanox InfiniBand) + IBM ESS 存儲 + IBM Spectrum Computing HPC Software Stack

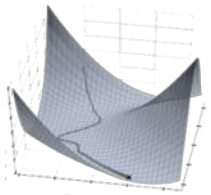


IBM PowerAI 最新業企版 1.5



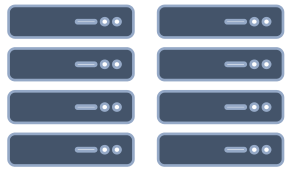
PowerAI : Enterprise Software Distribution

Binary Package of Major Deep Learning Frameworks with Enterprise Support



AI Vision* / DL Impact : Tools for Ease of Dev.

Graphical tools to Enhance Data Scientist Developer Experience



DL Impact / DDL / LMS : Faster Training Times

Performance Optimized for Single Node & Distributed Computing Scaling

PowerAI Software Distribution

Deep Learning Frameworks: Caffe | NVIDIA Caffe | IBM Caffe | torch | TensorFlow | DL4J | theano | Chainer

Supporting Libraries: DIGITS | OpenBLAS | Distributed Frameworks | Bazel | NCCL

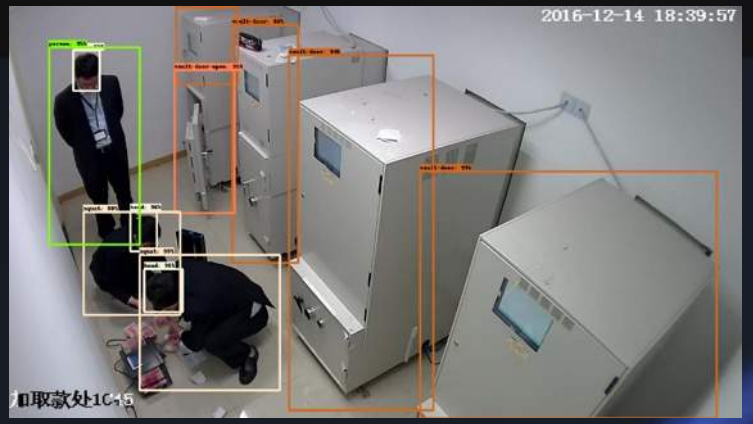
IBM Power System for HPC, with NVLink

Breakthrough performance for GPU accelerated applications, Including Deep Learning and Machine Learning

IBM PowerAI 客戶案例

電力公司採用無人機檢測線路

智能監控系統



醫療保健領域檢查

股票指數預測

組成AI的三個元件

IBM

Data

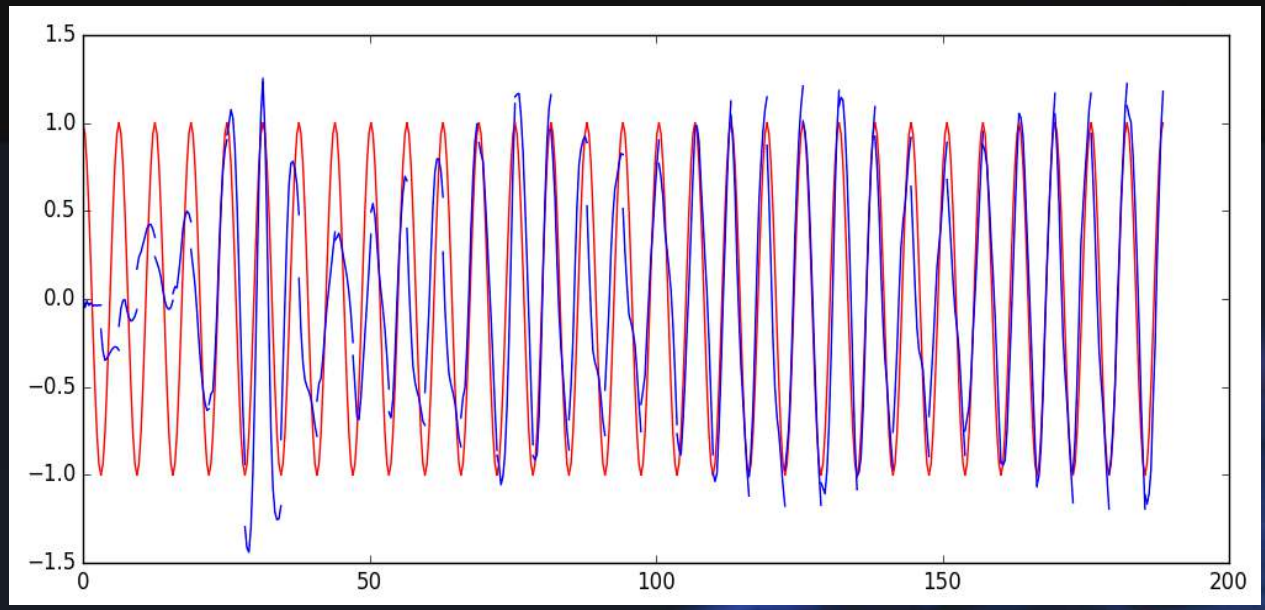
Compute

Algorithms

股票指數趨勢預測

Predict next 10mins，Accuracy 80%+
Predict next 50mins，Accuracy 70%+

RNN (LSTM) & linear regression Model



Model Adoption to Future Exchange & Foreign Exchange

Future Exchange Trading Code	Future Exchange Description	class=2, (up, down) 10 Min Forecast accuracy (%)	class=3, 幅度0.1% (up 0.1%, flat, down 0.1%) 10 Min Forecast accuracy (%)
if888	CSI 300 Index (滬深300股指連續)	81.66%	78.85%
rb888	Steel Bar (螺紋鋼連續)	81.94%	76.68%
cf888	Cotton (棉花連續)	81.24%	74.71%

股票指數趨勢預測

Data

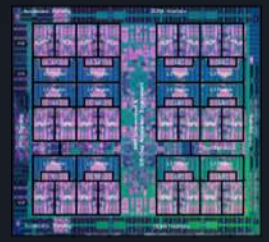
Stock Index Historical Data

- High
- Low
- Open
- Close
- Volume
- K-Line
- MACD
- DIF

Compute

IBM Power AC922 Cognitive Systems

IBM POWER9 CPU



NVIDIA TESLA V100 GPU

Algorithm

IBM PowerAI & IBM Deep Learning Impact

IBM PowerAI

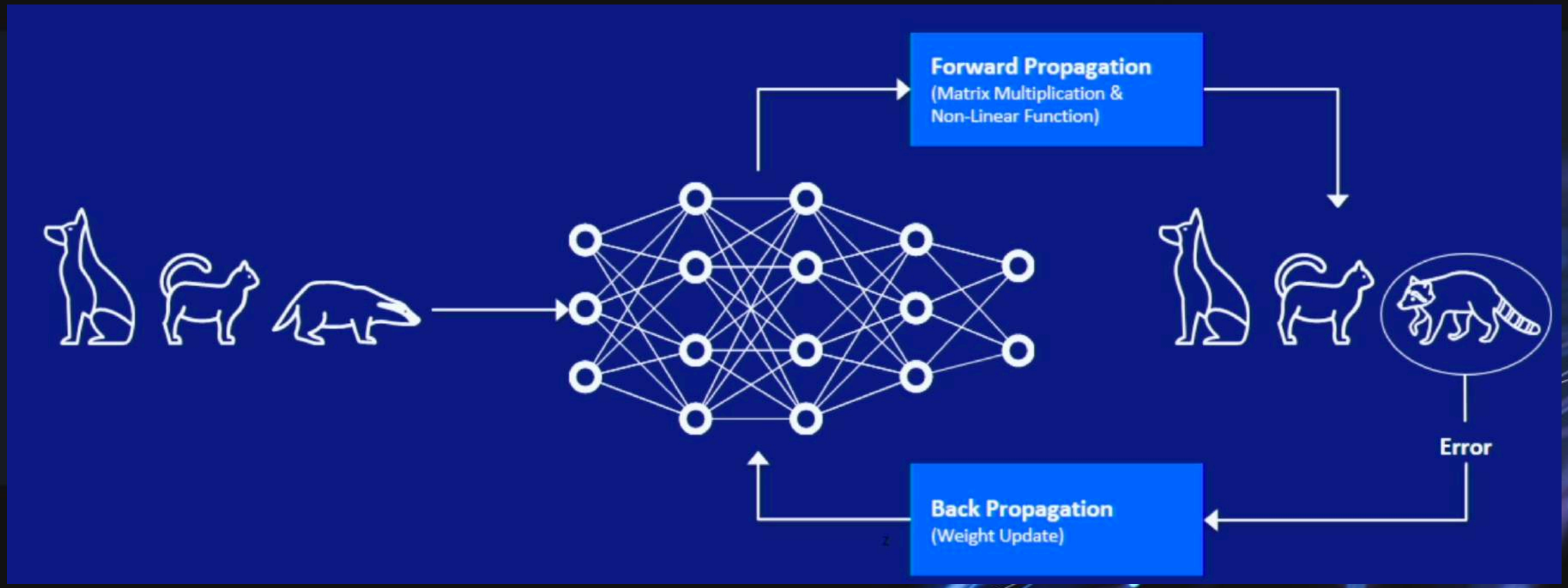
TensorFlow

TensorFlow

Programming : Python

Signal Model : RNN (LSTM)

深度學習 Deep Learning



Advanced Driver Assistance Systems (ADAS) System

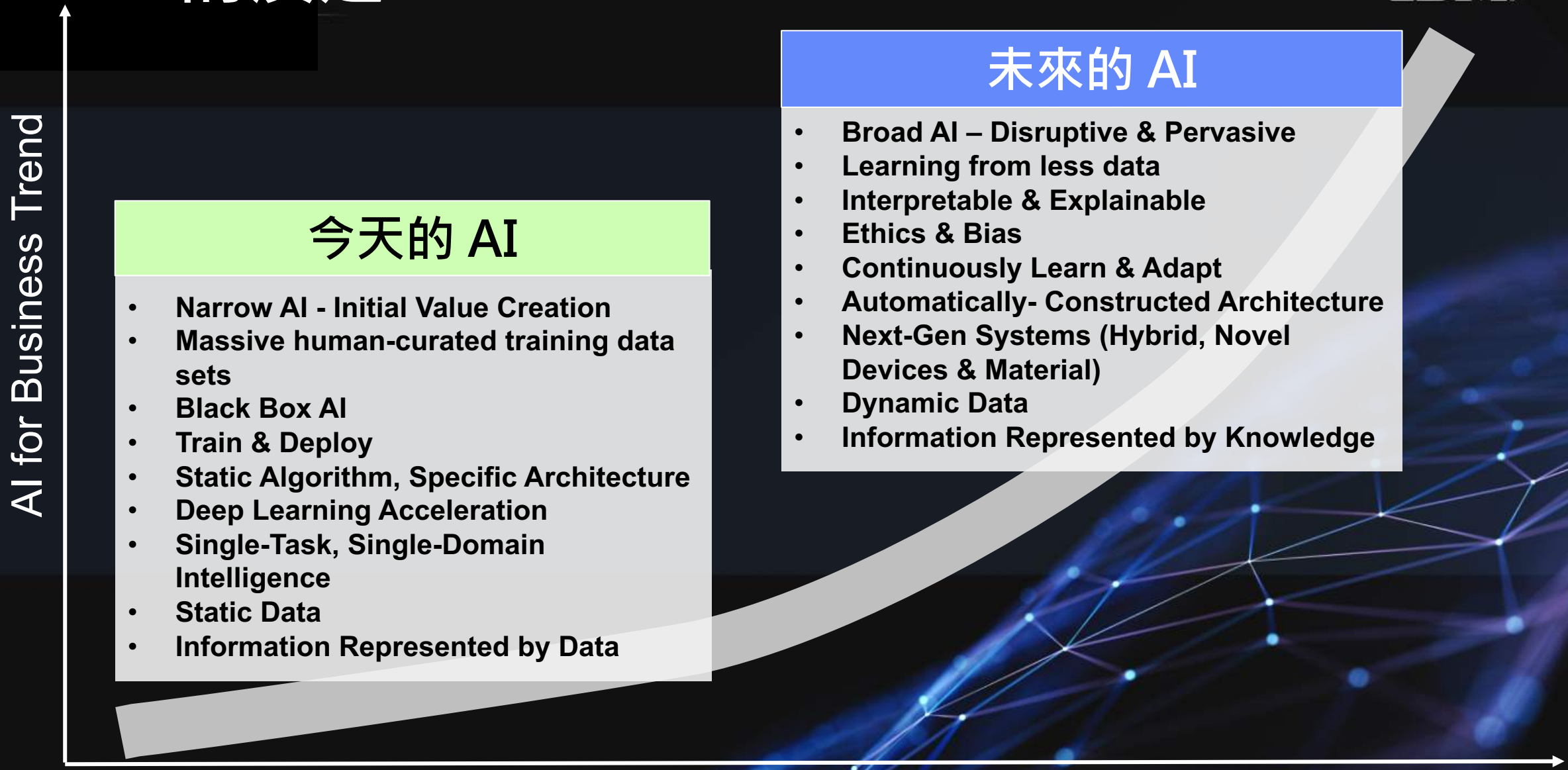
Block out front car

Block out passing cars

Anti-glare on traffic sign

Bad Weather Condition

AI的演進



今天的 AI

- Narrow AI - Initial Value Creation
- Massive human-curated training data sets
- Black Box AI
- Train & Deploy
- Static Algorithm, Specific Architecture
- Deep Learning Acceleration
- Single-Task, Single-Domain Intelligence
- Static Data
- Information Represented by Data

未來的 AI

- Broad AI – Disruptive & Pervasive
- Learning from less data
- Interpretable & Explainable
- Ethics & Bias
- Continuously Learn & Adapt
- Automatically- Constructed Architecture
- Next-Gen Systems (Hybrid, Novel Devices & Material)
- Dynamic Data
- Information Represented by Knowledge

– 2018 IBM 科技論壇 –

迎戰未來

機器學習與架構平台的年度盛宴

IBM

謝謝

讓 IBM 與您攜手共同釋放 AI 潛力
迎戰未來！