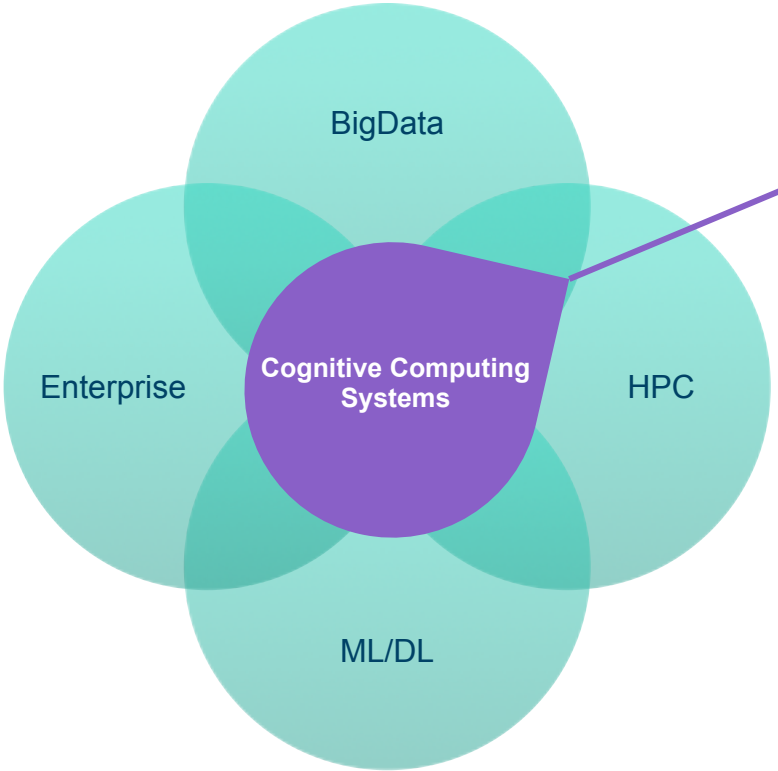
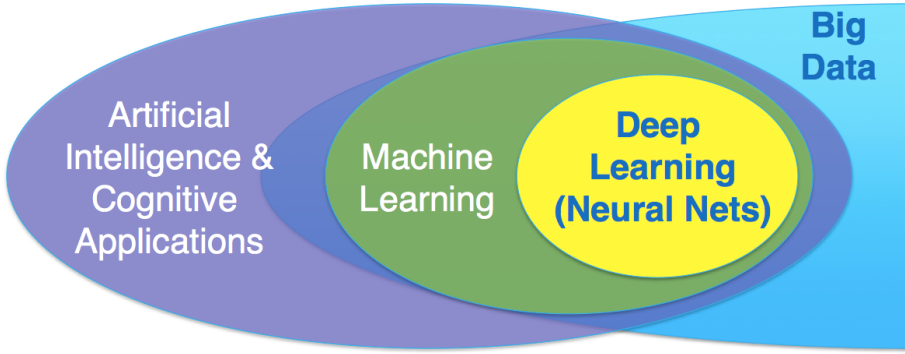
flattening the Deep Learning time to value curve

Franz Bourlet
POWER Systems Technical Sales
IBM Belgium & Luxembourg
Franz_Bourlet@be.ibm.com

Agenda

- Introduction to machine learning / deep learning
- Watson vs PowerAI
- POWER for cognitive solutions
- Cognitive solutions on POWER
- PowerAI demo
- Use cases / references
- AI in Belux





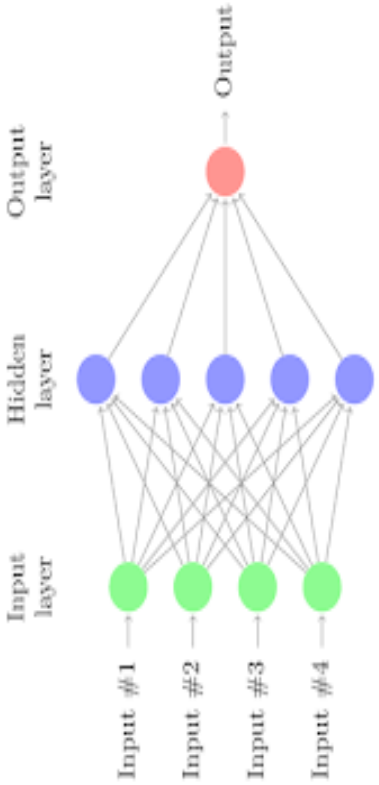
Artificial Intelligence (AI)
- Intelligence exhibited by machines or software

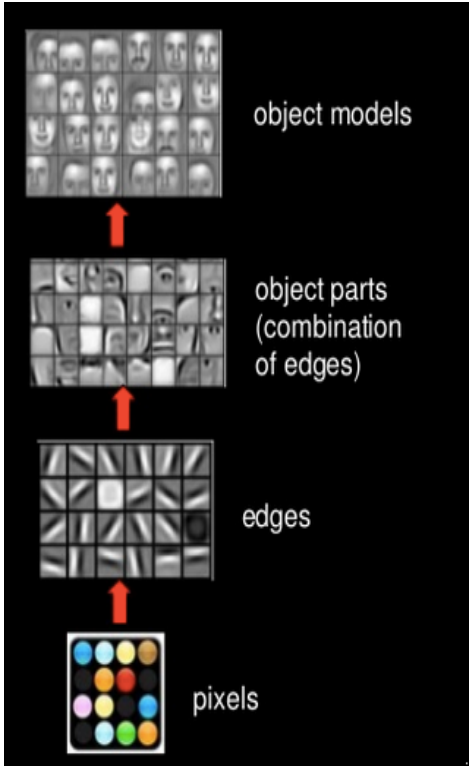
Cognitive Computing
- Computing systems exhibiting acts of perception, memory, judgment or reasoning

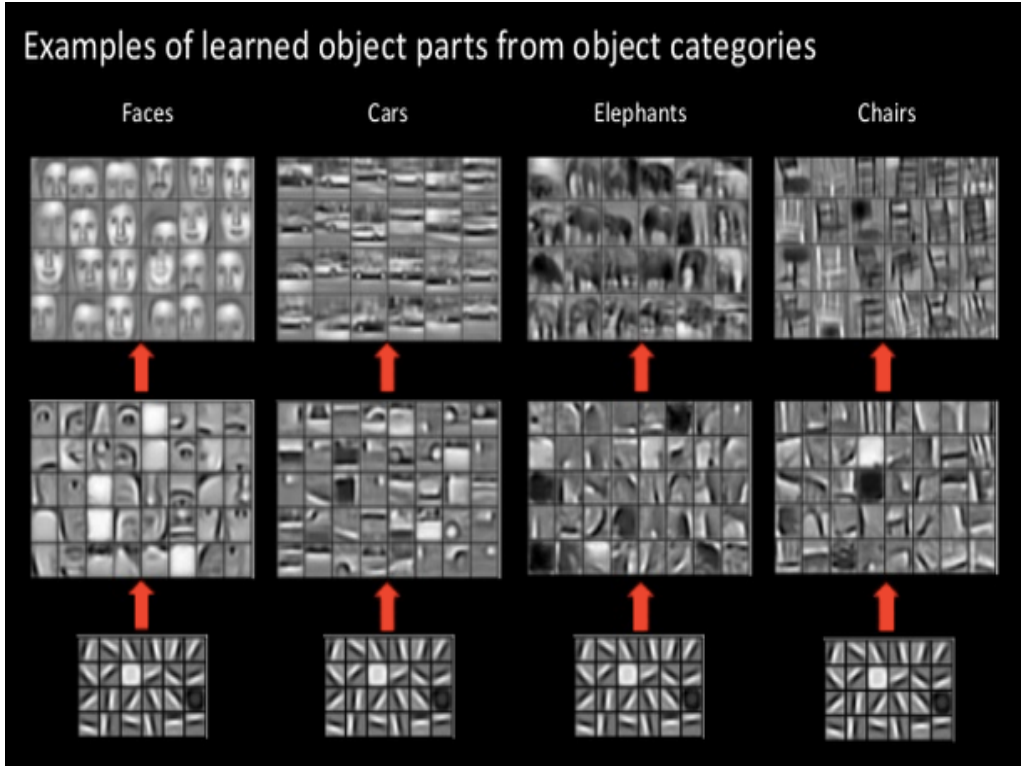
Machine Learning (ML)
- Type of AI that enables computers to learn without being explicitly programmed

Deep Learning (DL)
- Type of ML, based on neural networks loosely modeled after the brain learns features and representations of data

Deep Learning….Under the Hood





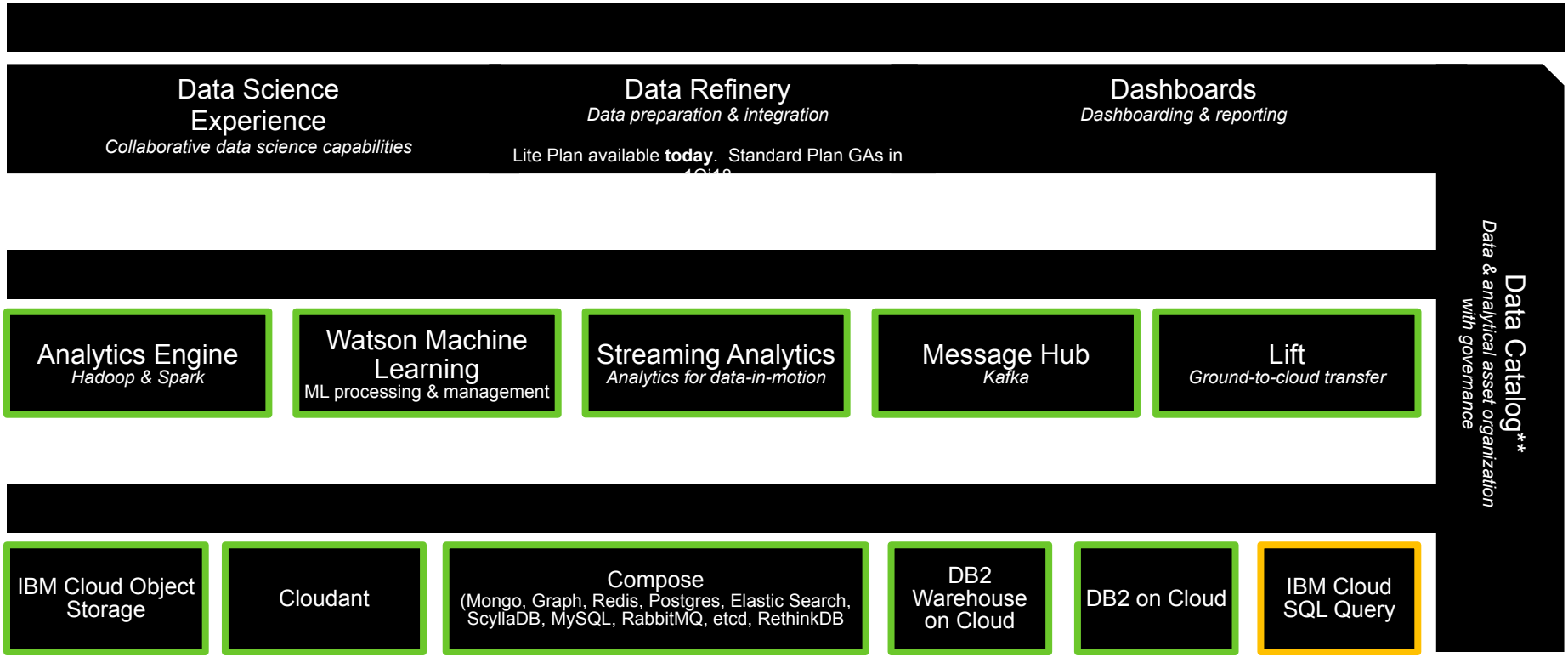


What is Watson?

- **It is a brand**
- **Multiple products in Watson family**
  - A diverse set of cognitive technologies covering AI, machine learning, and deep learning
  - Evolve from the Jeopardy Game and cancer research applications
  - Organic and acquired technologies
- **Cloud only**

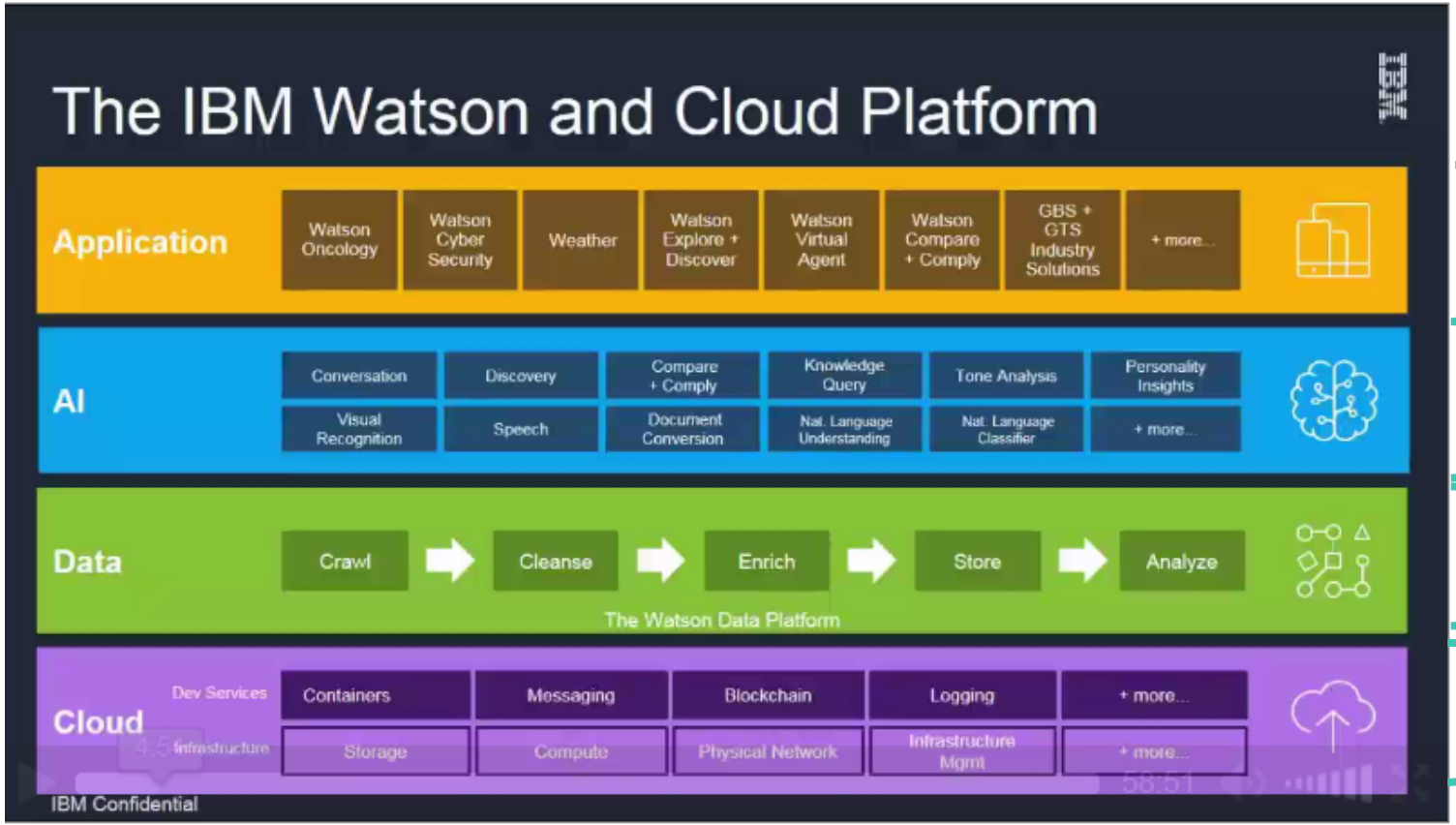
The Watson Data Platform Portfolio

Legend:
Available Now
Generally Available in December
Generally Available in 1H 2018



** Data Catalog PNs in SQO **now**

Watson & Cloud Overall Architecture



Industry Solutions – BP/ISV/Client play for PowerAI

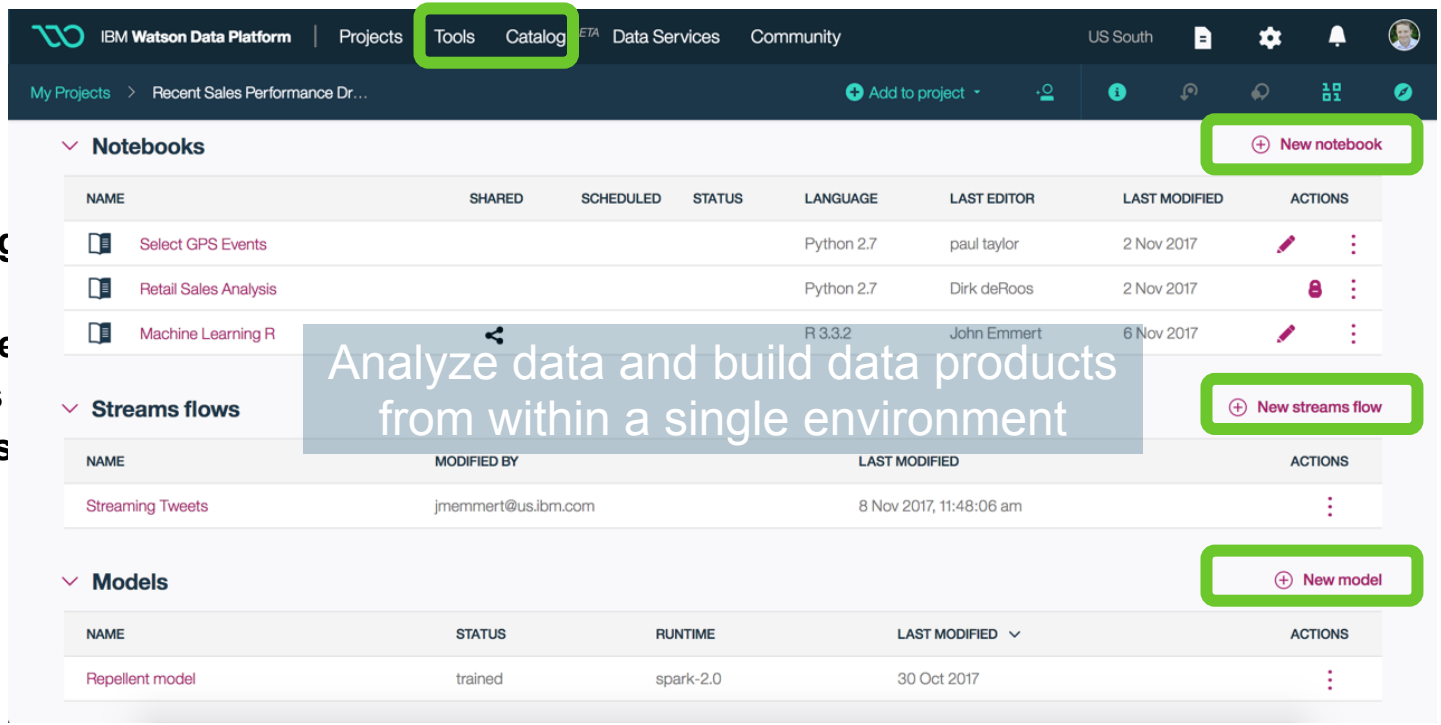
Cognitive Services (comparable to what can be built from open source frameworks that are part of Power AI) – API delivery model

Data Enrichment & Storage services – complimentary to PowerAI

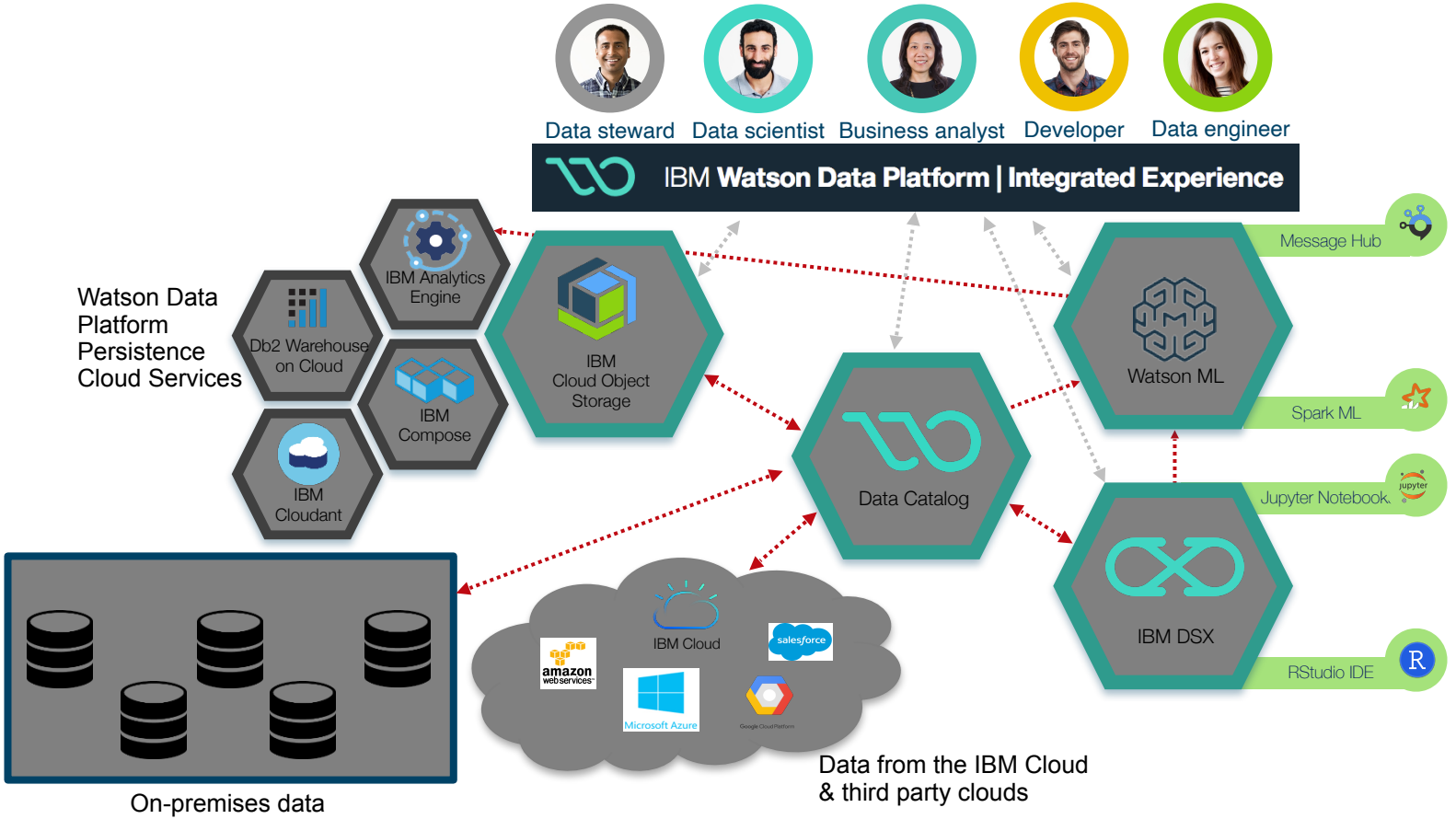
Infrastructure – Only cloud but 2 flavors -
- Public - Multi-tenant, share all
- Dedicated – VPC-driven isolation

Watson Data Platform : an integrated, unified self-service experience

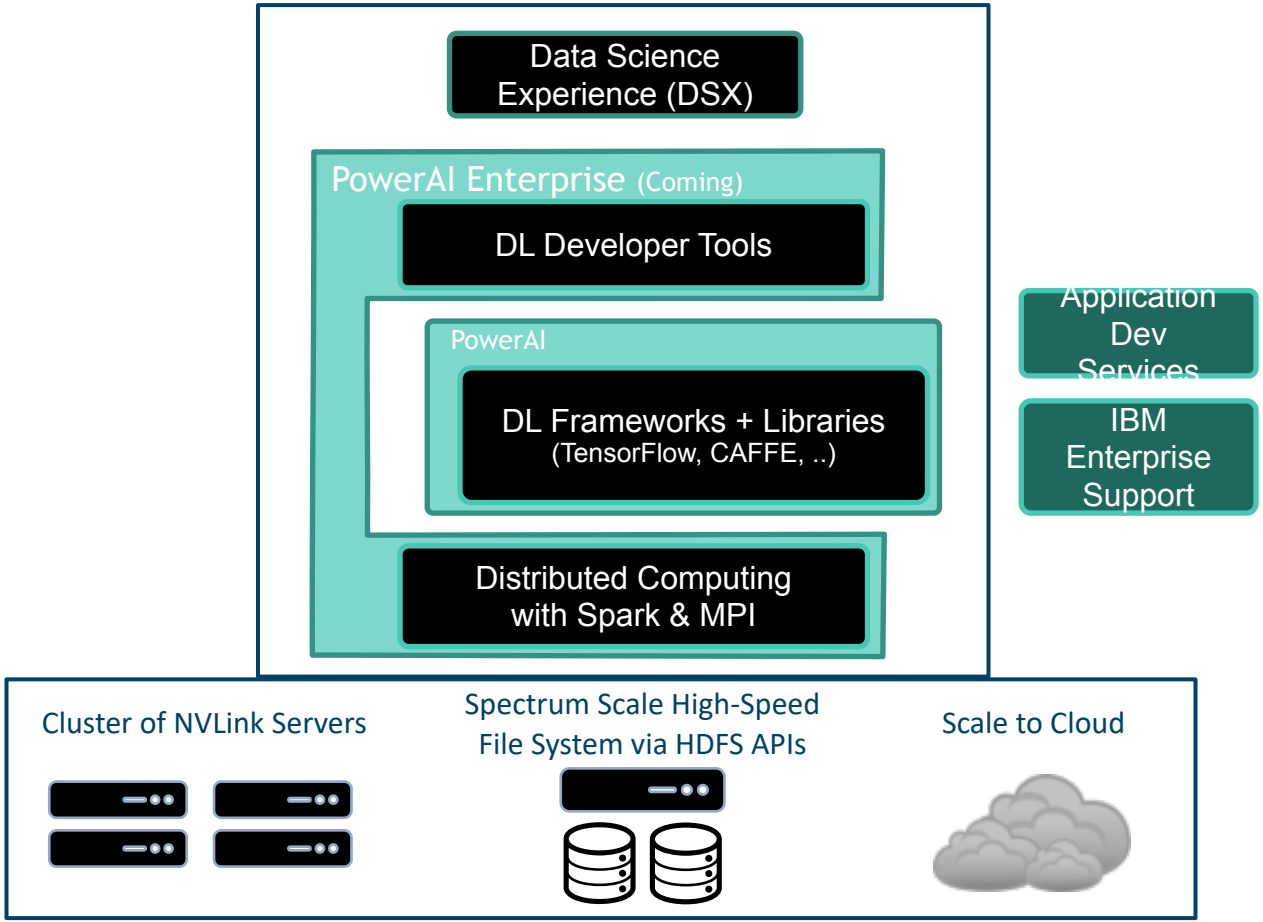
- Shop for data
- Manage policies
- Shape data
- Build dashboards
- Auto model building
- Build ML flows
- Auto-optimize mode
- Develop notebooks
- Streaming pipelines
- Build data apps



Example Deployment Architecture: Data Science



PowerAI Enterprise: Enhancing Developer Experience (& DSX integration)



Positioning --- Infrastructure Location

Watson AI

Cloud
Mainly looking at cloud deployment, minimal data privacy issues, i.e., data can be moved off-premise
Training, validation data can be hosted in a cloud environment
Does not envisage an issue with scaling on the cloud
Use case requires minimal data movement, i.e, for model re-training or the model used is well established
Use cases : chatbots, standard image classification, product recommendation, etc.

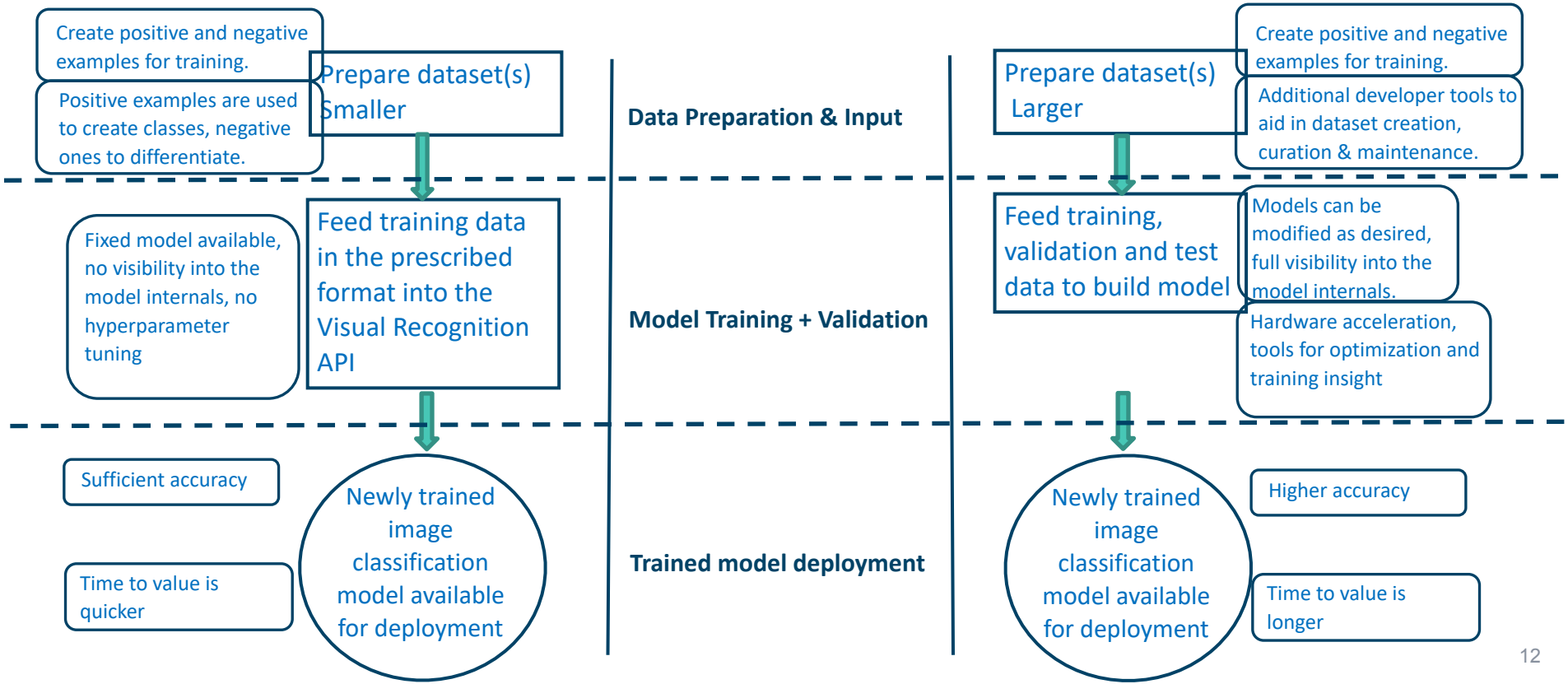
PowerAI

On-prem
Mainly looking at on-prem deployment, strong requirement that data does not move offsite
Training, validation data needs to be kept in-house due to regulatory, competitive or other reasons
Multiple training runs need to be carried out on the model as focus on training is higher
Large amount of data for training and training data is updated frequently
Use cases : fraud detection, credit analysis, image analysis for security, etc.

Positioning --- Model Training

Watson-based flow to create a custom image classification model – Focused on solution building

PowerAI-based flow to create a custom image classification model – build the classifier from ground up



Positioning --- Data

Watson AI	PowerAI
Unstructured Data	**Unstructured and Structured Data**
Unstructured data - systems of engagement	Unstructured data - systems of engagement Structured data - systems of records
Text, audio, images, video	Text, audio, images, video Transactional data, warehouse data
Use cases generally center around add on to business functions: chatbots, translation, visual recognition	Besides Watson AI use cases, it is also used to improve core business functions: churn reduction, fraud detection, product promotion
Use data about what people say	Use data about what people do

Positioning --- Targeted Users

Watson AI

Application Developers
Easier to use. Hide data science complexity.
Easier to retrain
Build applications quickly

PowerAI

Data Scientists
Full data science capabilities
More optimized results with retraining
Slice and dice data to get optimal results Longer training effort

Positioning --- Infrastructure Management

Watson AI

Cloud
Is not keen on visibility into the infrastructure & its management (compute, storage, etc.)
Client is not invested in building a data center and does not plan to
Does not envisage an issue with scaling on the cloud
View the infrastructure management piece as a significant barrier to adoption
Is content with the performance provided by standard configurations (including accelerators) available in the cloud
No push from externalities like regulations, privacy or data as competitive advantage to make the move to an on-prem solution

PowerAI

On-prem
Is very keen to have complete visibility into infrastructure & its management
Already has a data center or is in the process of building one
Considering the impact of scaling on the cloud
Infrastructure management is not a significant barrier to adoption
Is looking to understand the performance benefit provided by non-standard configurations with on-prem infrastructure
Does view regulations, data privacy and/or data as competitive advantage as these issues drive a move to an on-prem solution

THE AI ERA IS HERE.

We continue to experience exponential growth of data and data sources.

Computing has moved beyond ‘post CPU only’ era, giving us vast computational power that was not accessible before.

CIOs are evolving from ‘chief information officer’ to ‘chief intelligence officer’ and the data science organization has continued to gain power and influence.

Right now, your infrastructure is putting up

ROADBLOCKS

Not equipped for enterprise-level data volumes

Blocks to acceleration

Servers not specifically designed for AI workloads

Not able to easily scale

AI DEMANDS A DIFFERENT TYPE OF SYSTEM

IBM Power Systems provides the cutting-edge advances in AI that data scientists demand, and the critical reliability that IT needs.

IBM Power Systems AC922

Architecture designed for the AI era with advanced IO interfaces

AI at unrivaled scale with what will likely become the worlds most powerful supercomputer

Extends on heritage of performance leadership across AI, HPC and accelerated DBs

Best Server for Enterprise AI

AC922

An Acceleration Superhighway
Unleash accelerated computing potential in the post CPU-only era

Designed for the AI Era
Architected for the modern analytics and AI workloads that fuel insights

Delivering Enterprise-Class AI
Cutting-edge AI innovation data scientists desire, with dependability IT requires

An Acceleration Superhighway

Unleash accelerated computing potential in the post CPU-only era

Fastest Accelerator Performance

NVIDIA 2nd generation NVLink for Power9 and PCIe Gen4 deliver performance advantage vs. PCIE gen3

Most Efficient use of Accelerators

Share resources across CPUs and GPUs, while reducing bottlenecks to more fully-utilize acceleration investments

Simplest path to Acceleration

Coherence simplifies path to acceleration by abstracting data movement and locality

Developed via Collaborative Effort

Benefits of Industry collaboration are etched in silicon via NVLink and OpenCAPI

**Coherence simplifies coding by abstracting data movement and locality for developers*

Designed for the AI Era

Exceptional data-intensive architecture for the types of AI and Analytics workloads that are fueling the AI Era

Exceptional Next-gen Design

Architecture ahead of x86 severs and better-equipped to handle today's data-intensive HPC and analytics workloads

More Data. Faster Data.

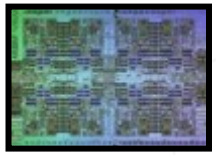
Up to 5.6x more I/O throughput and memory bandwidth, 4x threads*. 4th generation PCIe and 2nd generation NVLink, 2nd generation of CAPI

AI at Unrivaled Scale

AC922 is the backbone of CORAL Summit, meeting milestones to deliver 200+PetaFlops of HPC and 3 ExaFlops of AI as a service performance.

** 5.6x I/O bandwidth claim based on NVIDIA measurement test conducted on a Xeon E5-2640 V4 +P100 vs Power9 + V100 (12 GB/s vs 68 GB/s rated)*

An Acceleration Superhighway: POWER 9 is IBM's Latest Processor

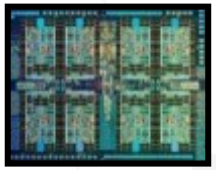


POWER7
45 nm

Enterprise

- 8 Cores
- SMT4
- eDRAM L3 Cache

1H10

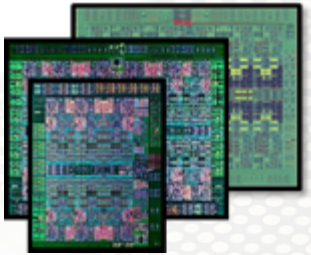


POWER7+
32 nm

Enterprise

- 2.5x Larger L3 cache
- On-die acceleration
- Zero-power core idle state

2H12

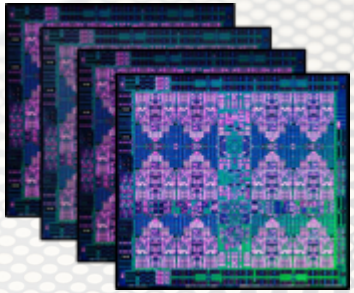


POWER8
Family 22nm

Enterprise &
Big Data Optimized

- Up to 12 Cores
- SMT8
- CAPI Acceleration
- High Bandwidth GPU Attach

1H14 – 2H16



POWER9 Family
14nm

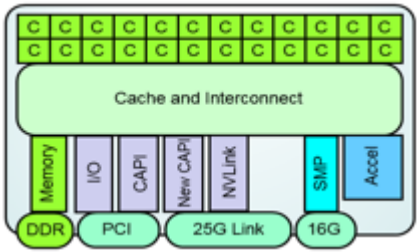
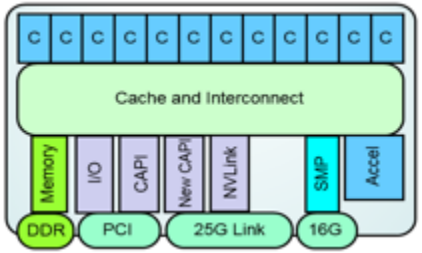
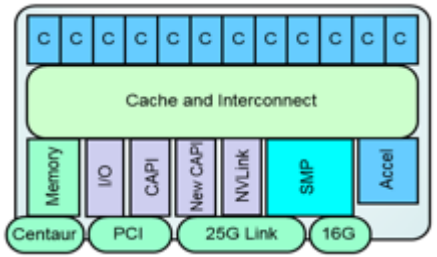
Built for the Cognitive Era

- Only processor with NVLink, PCIe Gen 4 advanced IO interfaces and coherence
- Premier Platform for Accelerated Computing
- Processor Family with Scale-Up and Scale-Out Optimized Silicon

2H17 – 2H18+

POWER9 Processor Family

Core Count / Size

SMP scalability / Memory subsystem	SMT4 Core **24 SMT4 Cores / Chip** Linux Ecosystem Optimized	SMT8 Core **12 SMT8 Cores / Chip** PowerVM Ecosystem Continuity
Scale-Out – 2 Socket Optimized **Robust 2 socket SMP system** **Direct Memory Attach** • Up to 8 DDR4 ports • Up to 170 GB/s memory BW • Commodity packaging form factor		
Scale-Up – 4+-Socket Optimized **Scalable System Topology / Capacity** • Large multi-socket **Buffered Memory Attach** • 8 Buffered channels • Up to 230 GB/s memory BW		

POWER9

An acceleration superhighway.
The only processor specifically designed for the AI era.

4x
Threads per core vs x86

9.5x
Up to 9.5x more I/O bandwidth than x86

2.6x
More RAM possible vs. x86

1st
CPU to deliver PCIe gen 4

An Acceleration Superhighway:
POWER9 offers a variety of Acceleration Options

- **Extreme Processor / Accelerator Bandwidth and Reduced Latency**
- **Coherent Memory and Virtual Addressing Capability for all Accelerators**
- **OpenPOWER Community Enablement – Robust Accelerated Compute Options**

State of the Art I/O and Acceleration Attachment Signaling

– **PCIe Gen 4** x 48 lanes – 192 GB/s duplex bandwidth

– **25G Link** x 48 lanes – 300 GB/s duplex bandwidth

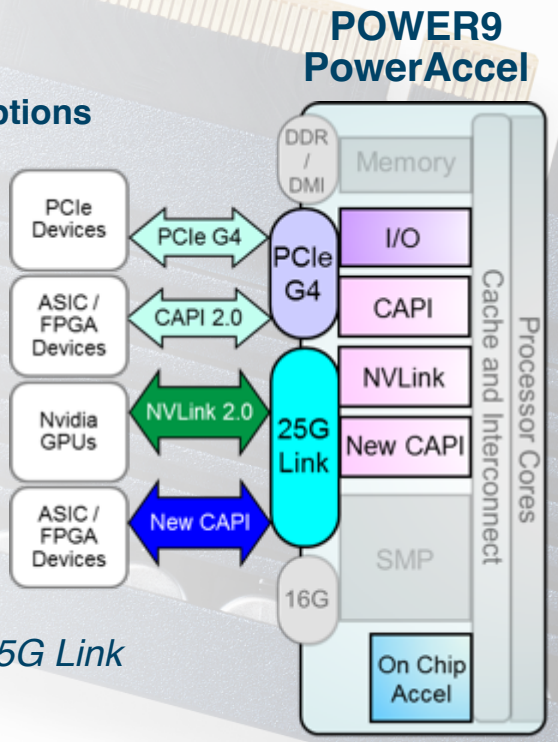
Robust Accelerated Compute Options with OPEN standards

– **On-Chip Acceleration** – Gzip x1, 842 Compression x2, AES/SHA x2

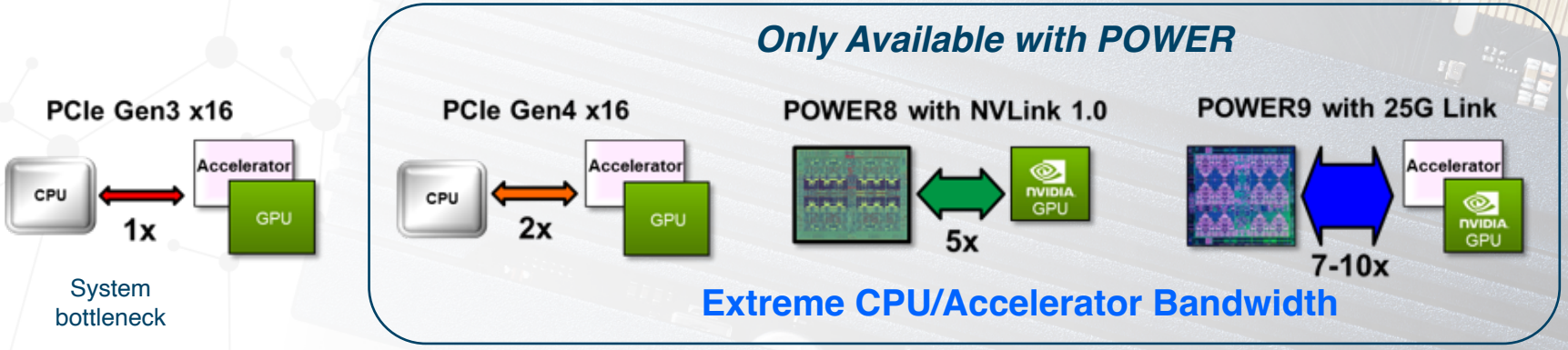
– **CAPI 2.0** – 4x bandwidth of POWER8 using *PCIe Gen 4*

– **OpenCAPI 3.0** – High bandwidth, low latency and open interface using *25G Link*

– **NVLink 2.0** – Next generation of GPU/CPU bandwidth and integration



An Acceleration Superhighway: POWER9 Introduces Acceleration Innovations



Seamless CPU/Accelerator Interaction

- Coherent memory sharing
- Enhanced virtual address translation

Broader Use of Heterogeneous Compute

- Designed for efficient programming models
- Accelerate complex analytic / cognitive applications

Designed for Great Supercomputing & AI Leaders

*Unprecedented performance and application gains with **advanced IO interfaces** integrated into the **NEW P9 processor** delivering **capabilities** not available on x86*

Advanced IO interfaces:

- **2nd Generation CPU - GPU NVLink:** ~5.6X the CPU-GPU bandwidth compared to x86
- **PCIe Gen4/CAPI 2.0:** First to market with PCIe Gen 4 with 2x improvement over PCIe Gen 3 and next gen CAPI for coherent device attachment

Introducing Coherence: Treat system memory just like GPU memory enabling game changing simplification of programing and larger model sizes

Burst Buffer: Start/finish job data staging with high performance storage adapter resulting in significant improvement of computing efficiency

GPU Direct: Peer to peer communication within a cluster for remote data access and transfer (RDMA)

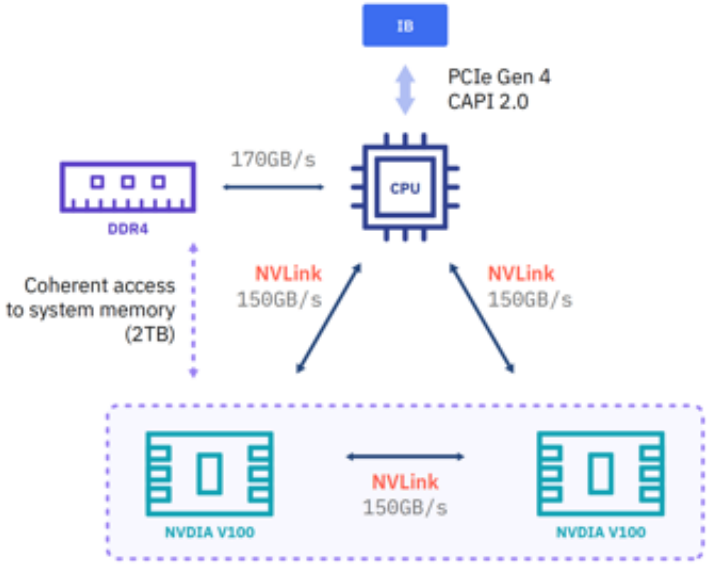
Configuration System Details for 4Q GA

- MTM: 8335-GTG - Specific 4Q Feature Availability:
  - 1.6 TB NVMe high performance storage adapter
  - 2 SFF SATA: HDD (Max 4TB); SSD (Max 3.84TB)
  - IO: EDR InfiniBand, Quad ENET (2x1/2x10 GB), Quad ENET (4x1 GB), 100 GB ENET
  - RHEL 7.4 for P9
  - Air cooled only version available (max 4 GPU's)
  - OpenBMC

- Results are based on IBM Internal Measurements running the CUDA H2D Bandwidth Test
- Hardware: Power AC922; 32 cores (2 x 16c chips), POWER9 with NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU; Ubuntu 16.04. S822LC for HPC; 20 cores (2 x 10c chips), POWER8 with NVLink; 2.86 GHz, 512 GB memory, Tesla P100 GPU
- Competitive HW: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4xTesla V100 GPU, Ubuntu 16.04

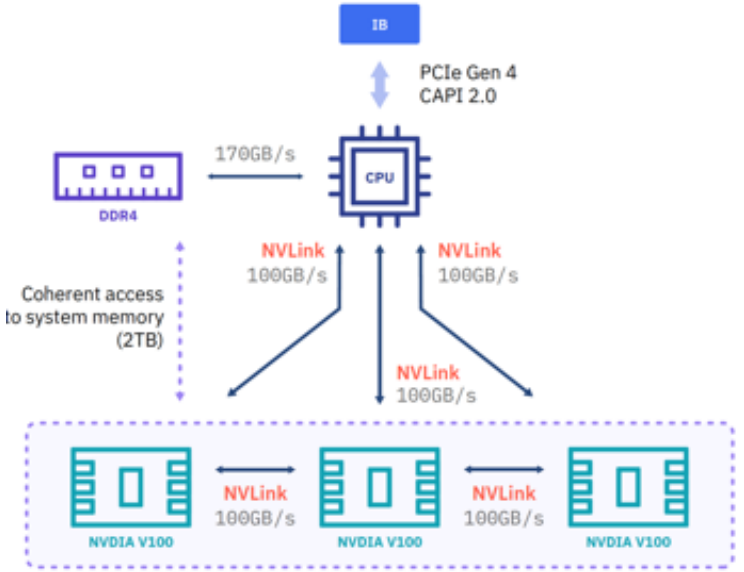
Two AC922 GPU configurations available

4 GPUs - Air (4Q'17)/Water Cooled (Coming)



- Up to 4 GPUs, air/water cooled options
- 150GB/s of bandwidth from CPU-GPU

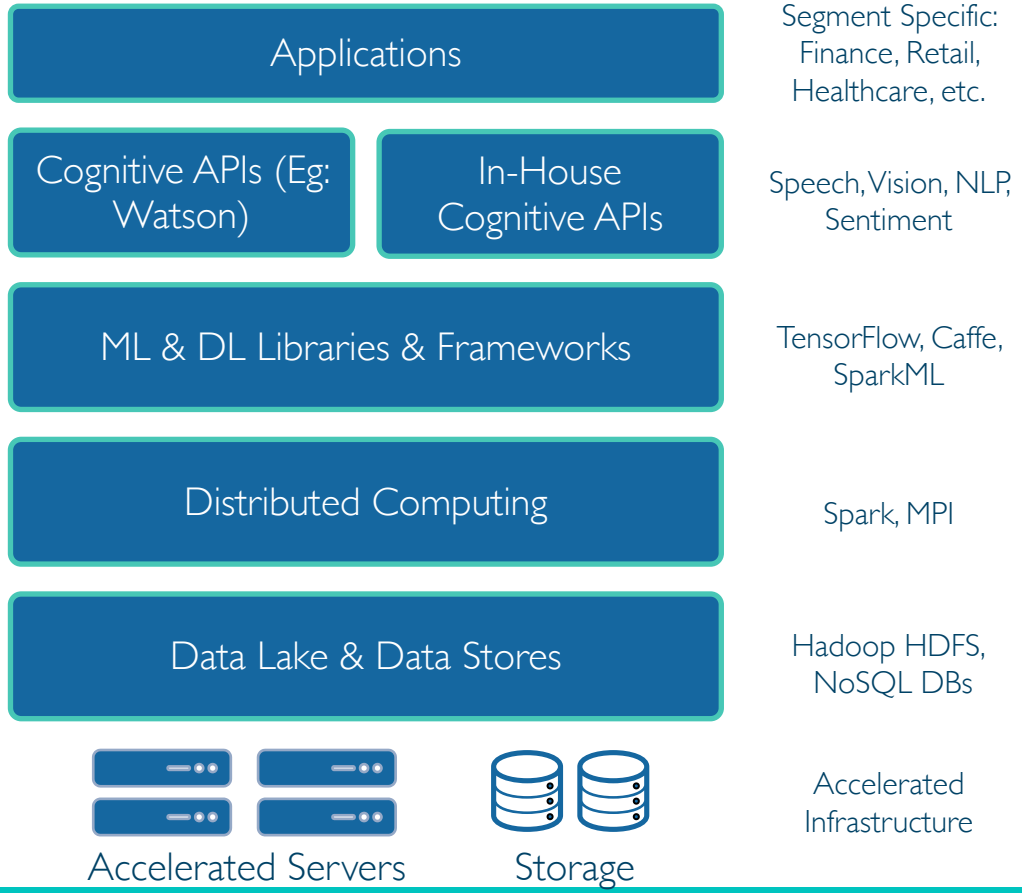
6 GPUs - Water Cooled (coming)



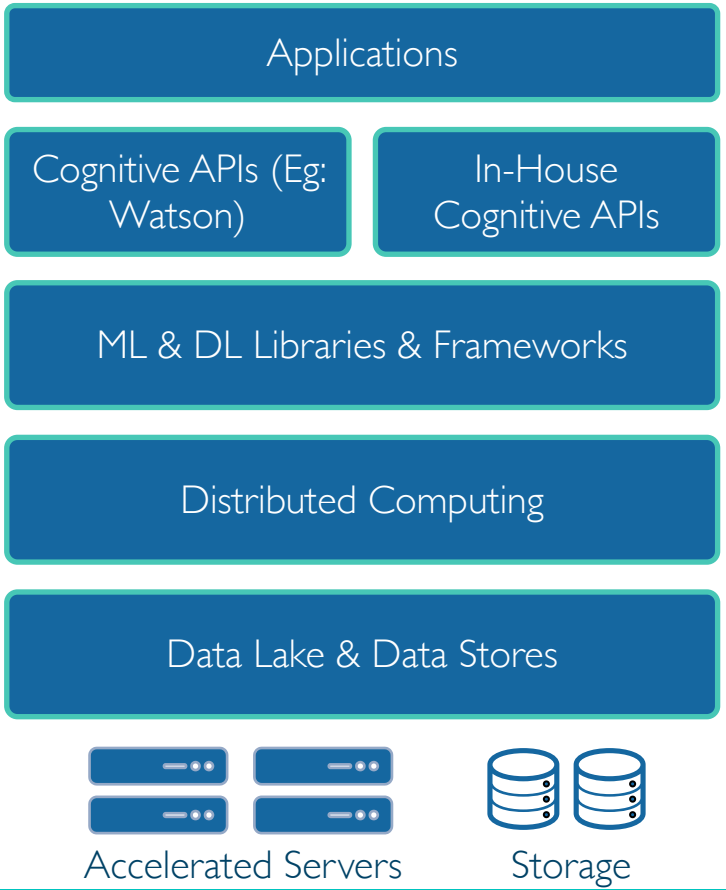
- Up to 6 GPUs, water cooled only
- 100 GB/s of bandwidth from CPU-GPU

- Coherent access to system memory
- PCIe Gen 4 and CAPI 2.0 to InfiniBand
- Water cooled options available in 2Q'18

Layers in AI Infrastructure Stack



Challenges in Building Cognitive Solutions



Use Case Identification, Access to Enough Data

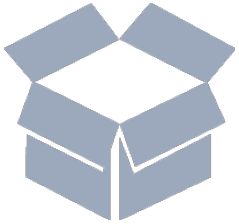
Finding Right "Tagged" Data, Model Integrity

Feature extraction, Selecting Right Model, Hyper-parameter tuning

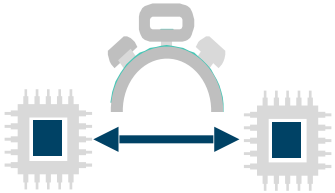
Multi-tenant, Cluster Virtualization, DL Framework Scaling

Data Prep, ETL, Curation, Data Labeling

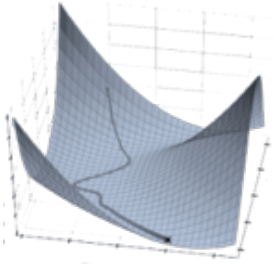
Performance to Reduce Training Time



enterprise-ready
software distribution
built on open source



performance
faster training times
for data scientists



tools for ease
of development

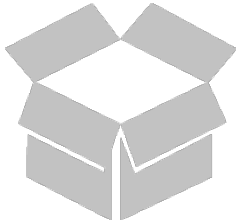
IBM PowerAI

Enterprise-Ready
Software Distribution
Built on Open Source

fast, easy deployment

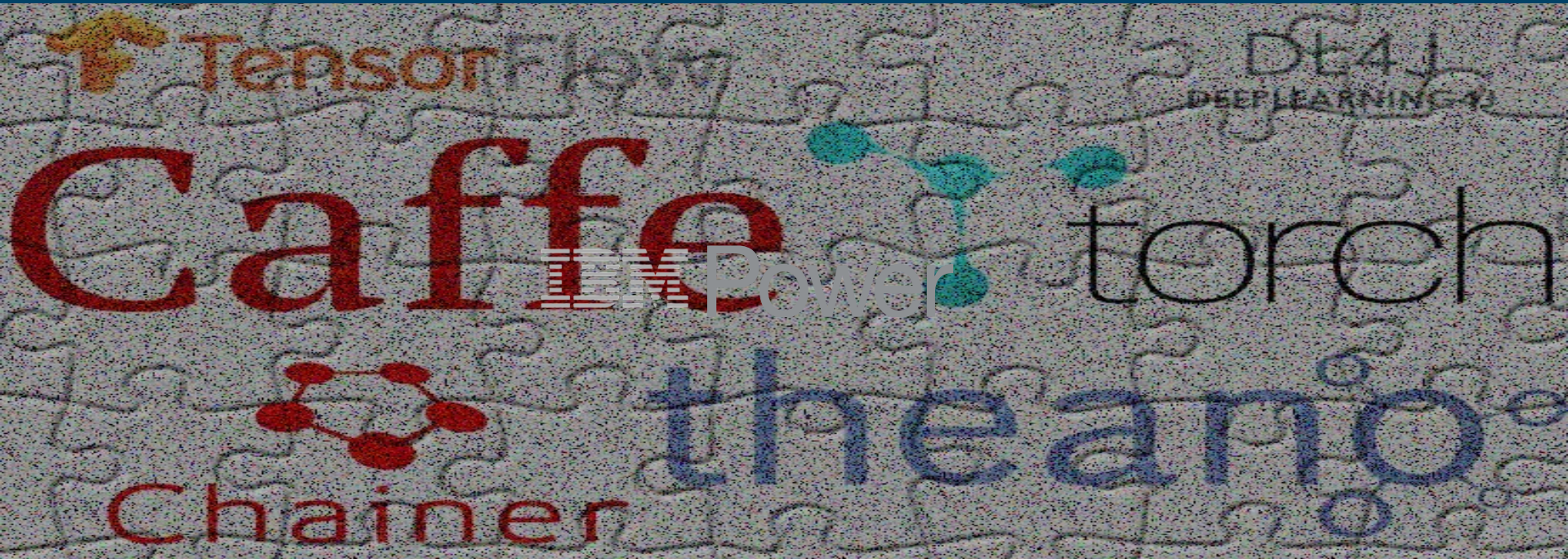
NOT PowerAI!

PowerAI!



Enterprise-Ready Software Distribution Built on Open Source

precompiled and current open source frameworks



PowerAI v4.0 - Deep Learning Software Distribution

Deep Learning Frameworks

Caffe | NVCaffe | IBMCaffe | Torch

TensorFlow | Distributed TensorFlow | Theano | Chainer

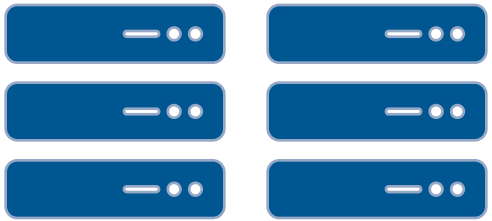
Supporting Libraries

OpenBLAS | Bazel | Distributed Communications | NCCL | DIGITS

Power

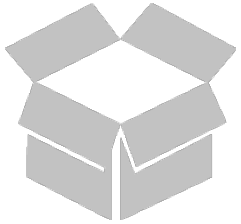
Accelerated Servers and Infrastructure for Scaling

IBM Cluster of NVLink Servers



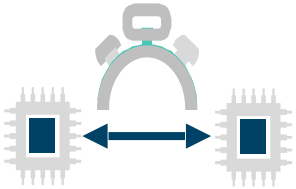
Spectrum Scale: High-Speed Parallel File System

Scale to Cloud



Enterprise-Ready Software Distribution Built on Open Source

available enterprise support for the entire stack



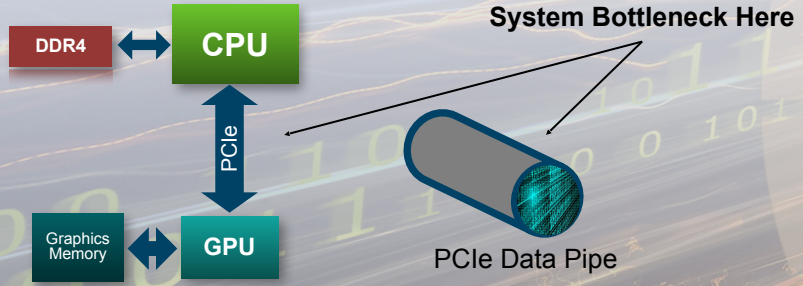
Performance… Faster Training and Inferencing

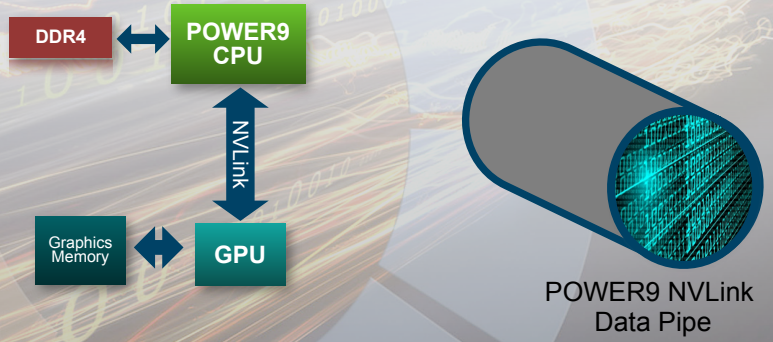
unique innovation through OpenPower collaboration

THE NEXT PLATFORM

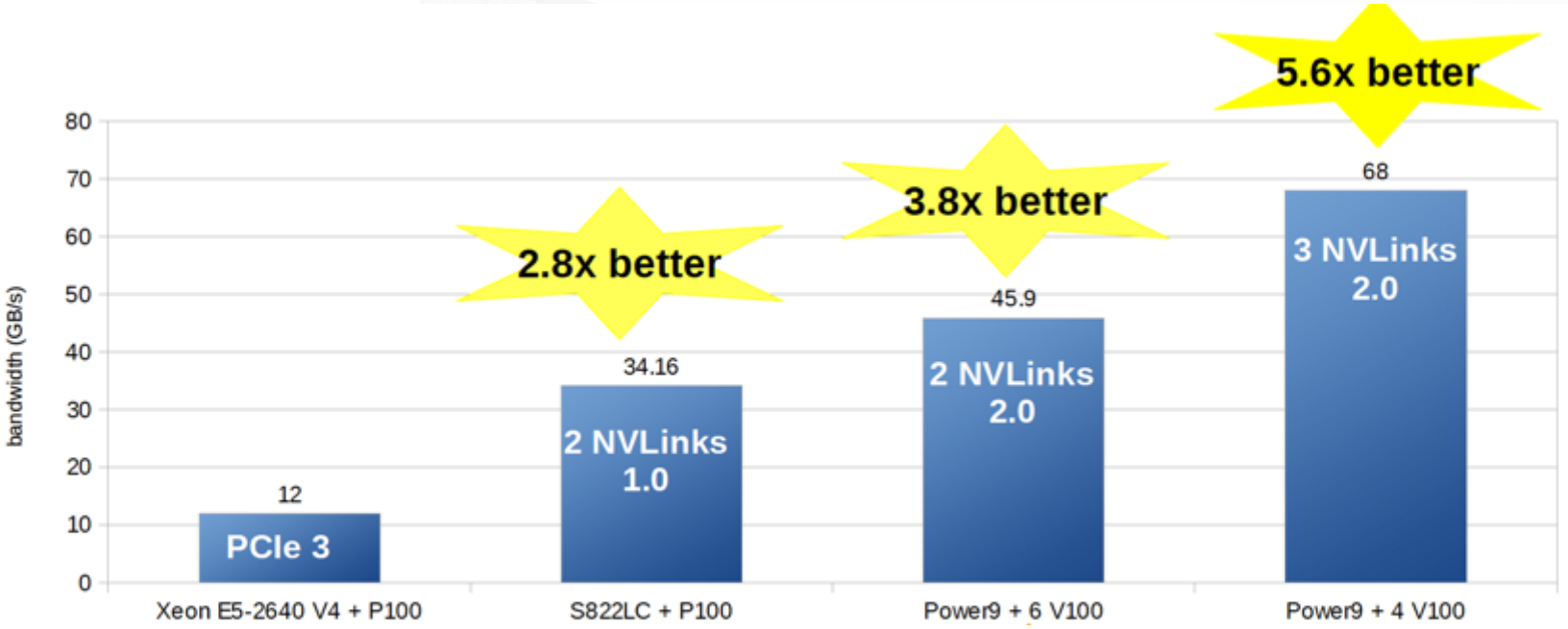
THE SYSTEM BOTTLENECK SHIFTS TO PCI-EXPRESS

POWER9 with NVLink delivers 5.6X the bandwidth



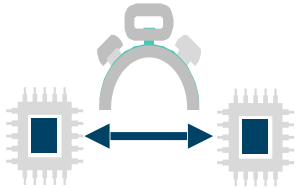


The NVLink difference CPU-GPU



- P9 with 2nd Gen NVLink enables 5.6x faster data movement from CPU-GPU in 4 GPU system
- In 6 GPU system bandwidth is minimally reduced but balanced by higher compute capability

- Results are based on IBM Internal Measurements running the CUDA H2D Bandwidth Test
- Hardware: Power AC922; 32 cores (2 x 16c chips), POWER9 with NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU; Ubuntu 16.04. S822LC for HPC; 20 cores (2 x 10c chips), POWER8 with NVLink; 2.86 GHz, 512 GB memory, Tesla P100 GPU
- Competitive HW: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4xTesla V100 GPU, Ubuntu 16.04



Performance… Faster Training and Inferencing

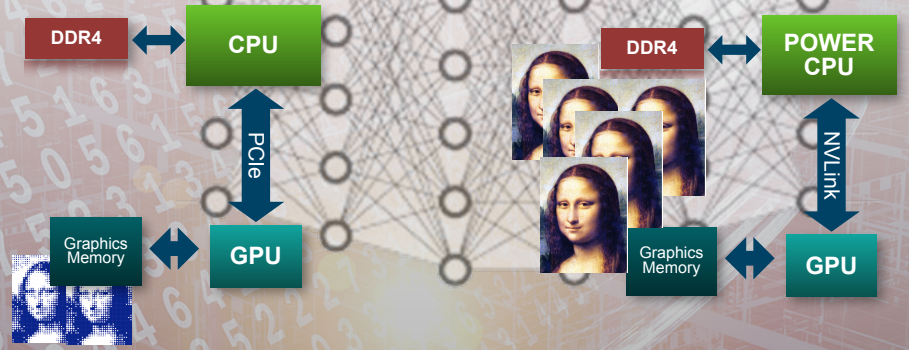
faster training times for data scientists

Distributed Deep Learning

Traditional Model Support → *Large Model Support*

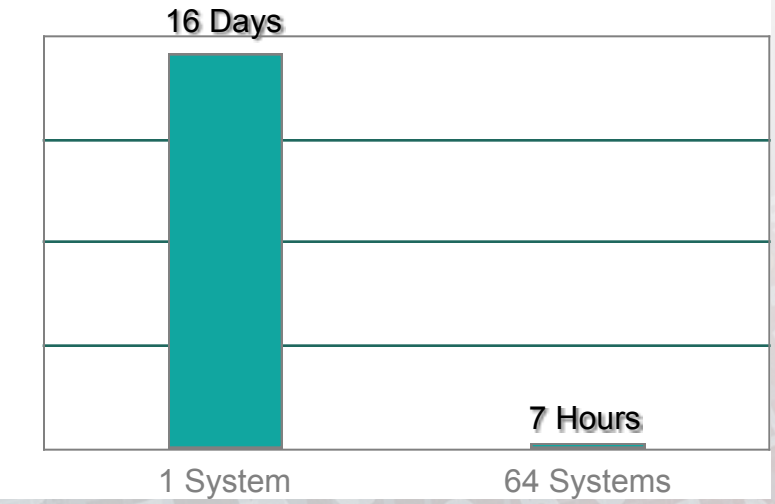
(Competitors)
Limited memory on GPU forces trade-off in model size / data resolution

(PowerAI)
Use system memory and GPU to support more complex models and higher resolution data

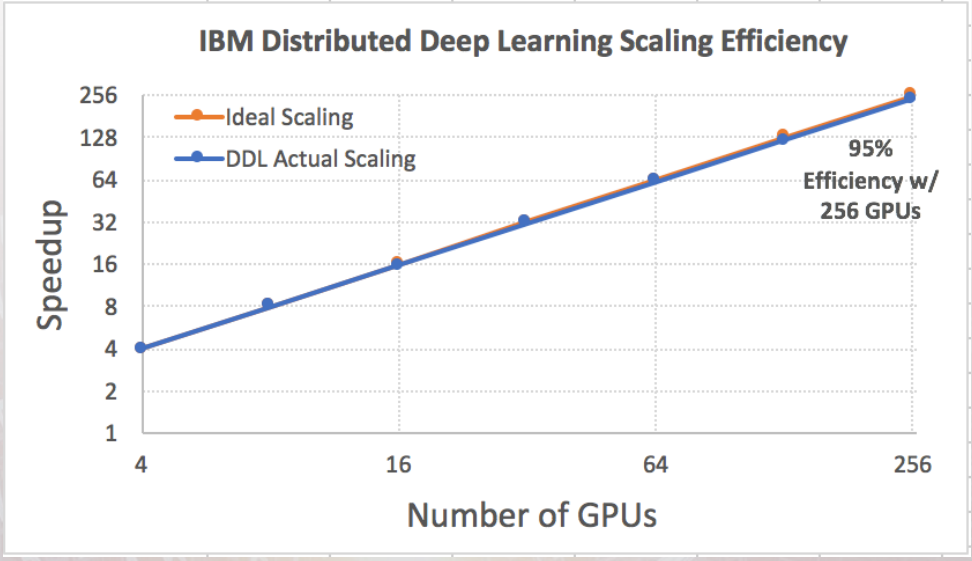


PowerAI Rel. 4 with Distributed Deep Learning

16 Days Down to 7 Hours:
58x Faster

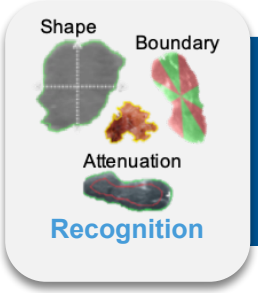


Near Ideal Scaling to 256 GPUs and Beyond

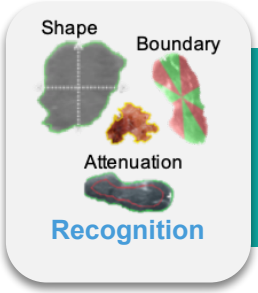


ResNet-101, ImageNet-22K, Caffe with PowerAI DDL, Running on Minsky (S822Lc) Power System

Acceleration training …. days become hours



9 Days



4 Hours | 4 Hours | 4 Hours | 4 Hours | 4 Hours

What will you do?
Iterate more and create more accurate models?
Create more models?
Both?

4 Hours

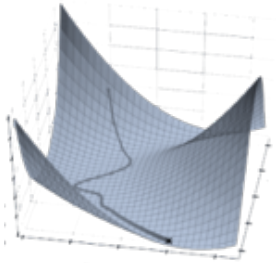
IBM Power

100x

Learning runs with Power 9*

54x

Learning runs with Power 8



Tools for Ease of Development

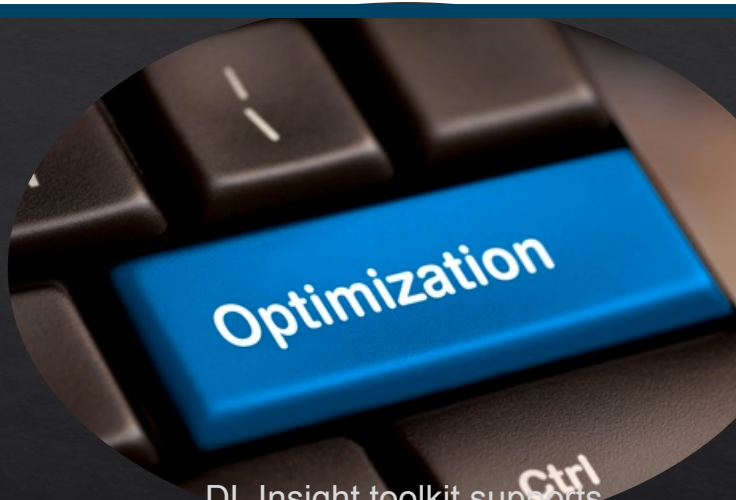
rich advisory and building toolsets to flatten time to value

AI Vision
rich toolset image recognition neural networks

Caffe

Spark

automated deep learning toolkit data preparation



DL Insight toolkit supports auto-training runs for hyper parameter tuning
+++

“IBM Power is a great cognitive platform if not the best out there. The IBM Power team identified the need for and implemented acceleration before anyone else in the industry and are already on their third generation with the highest speed accelerator interconnects and coherent architecture that can share main memory with accelerators.”

Forbes

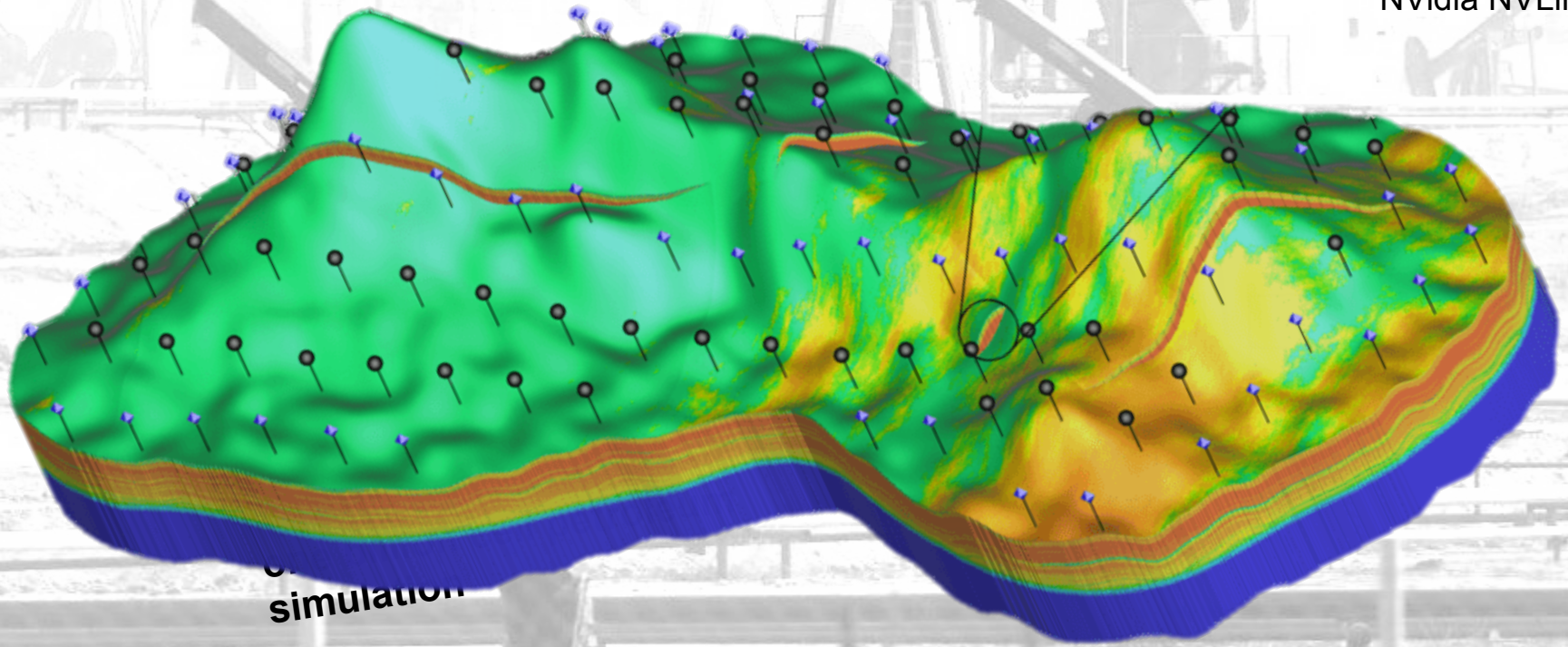
IBM's New PowerAI Features Demonstrate Enterprise AI Leadership …Again

Oil & Gas Billion Cell Reservoir calculation

IBM Power

22,000 computer node cluster
716,000 Intel CPUs

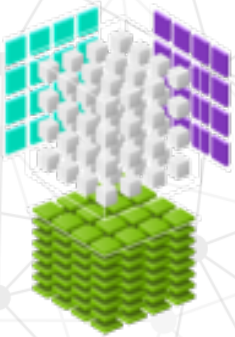
30 computer node cluster
60 Power CPUs
120 NVIDIA Tesla P100 GPUs
NVidia NVLink



simulation

CORAL – Summit and Sierra on day 1

AI Exascale Today



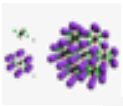
Order of Magnitude Leap in Computational Power

200 PF

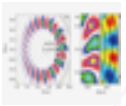
20 PF

Real, Accelerated Science

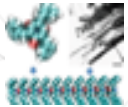

ACME

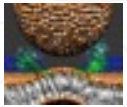

DIRAC


FLASH


GTC

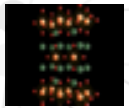

HACC


LSDALTON

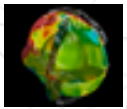

NAMD

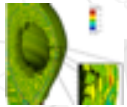

NUCCOR


NWCHEM


QMCPACK


RAPTOR


SPECFEM


XGC

Deployment at Summit

- Significant application performance over Titan (AMD/NVIDIA) Achieved with ¼ the number of servers
- Similar wins at :

3+EFLOPS
Tensor Ops

10X
Perf Over Titan

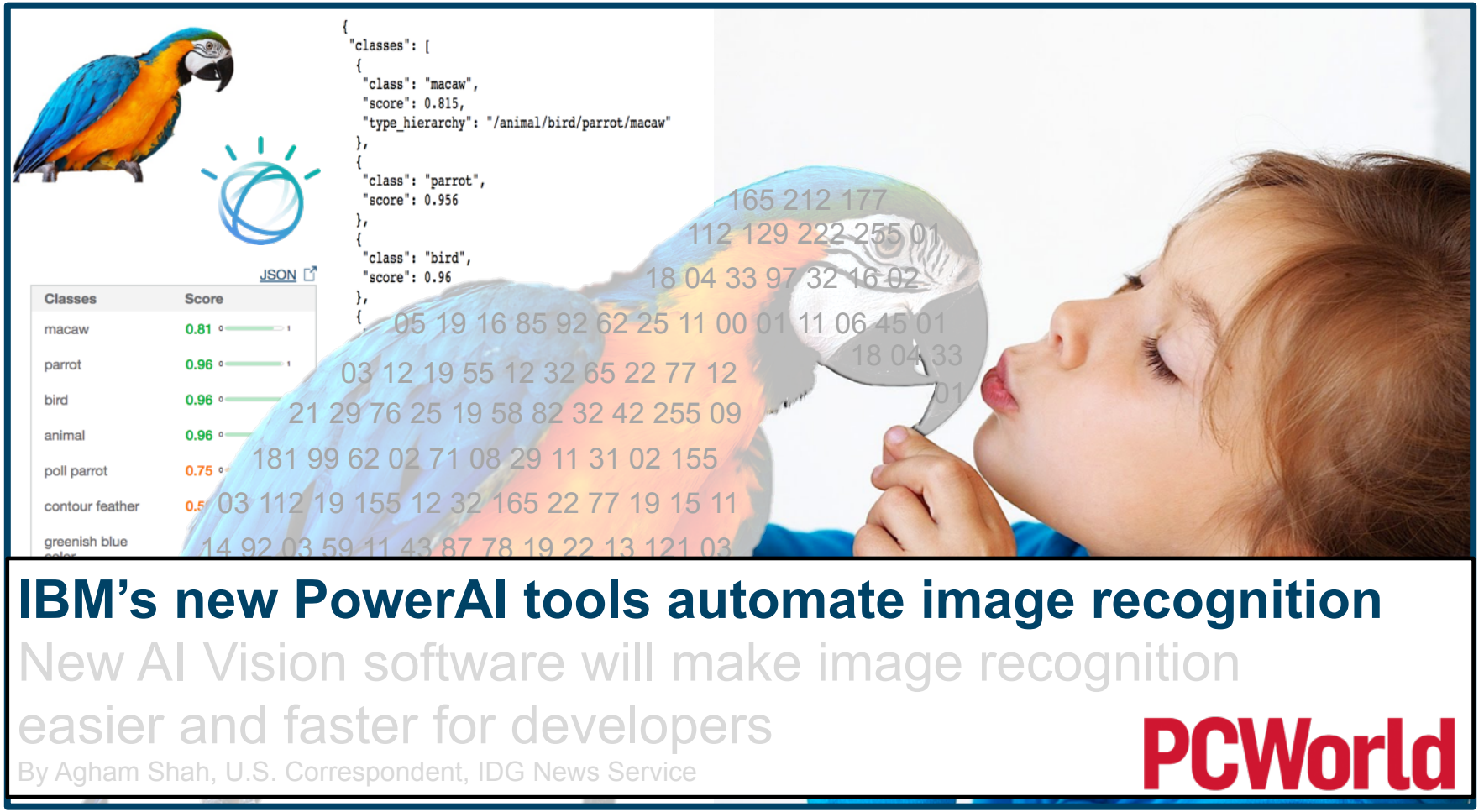
5-10X
Application Perf Over Titan

Office of Science
U.S. Department of Energy

U.S. DEPARTMENT OF ENERGY

Source: Data published by Oak Ridge National Laboratories at https://www.olcf.ornl.gov/summit/
Courtesy of Oak Ridge National Laboratory, U.S. Dept. of Energy

NDA until product announce



IBM's new PowerAI tools automate image recognition

New AI Vision software will make image recognition easier and faster for developers

By Agham Shah, U.S. Correspondent, IDG News Service

PCWorld

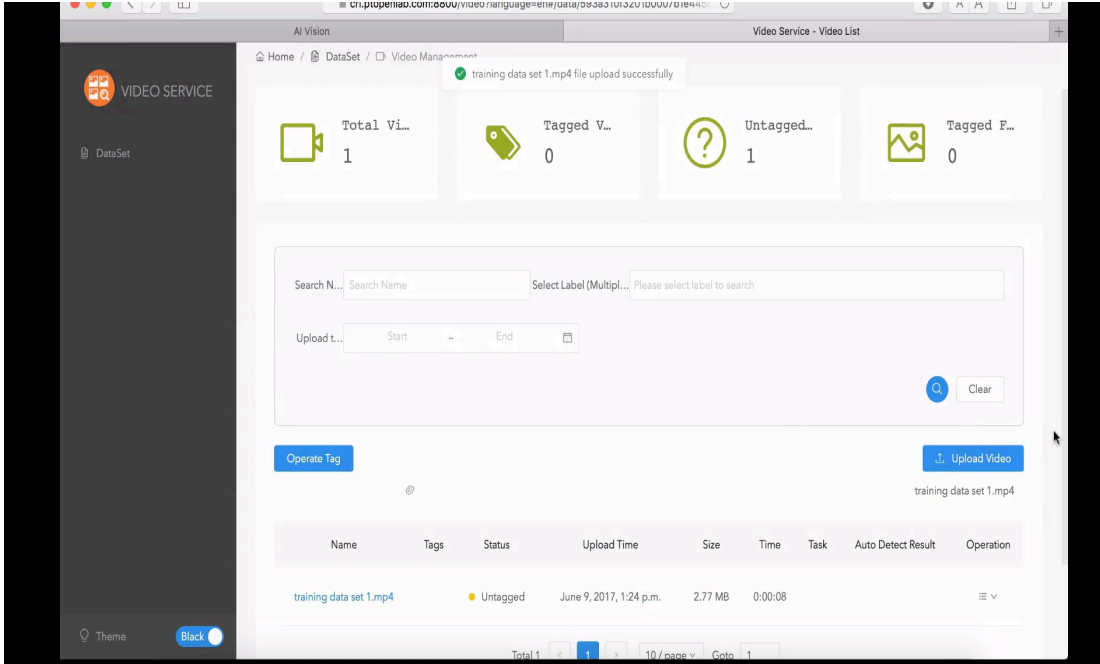
AI Vision toolset

Experts only becomes beginner knowledge requirement to build image-based neural nets

Tooling lets non-techies label data – brings expertise to algorithm from LOB and mitigates errors

Choose best model and framework to apply based on data set

IBM Power
Tools for Ease of Development



Easily upload multiple videos which will be used to create labels for objects

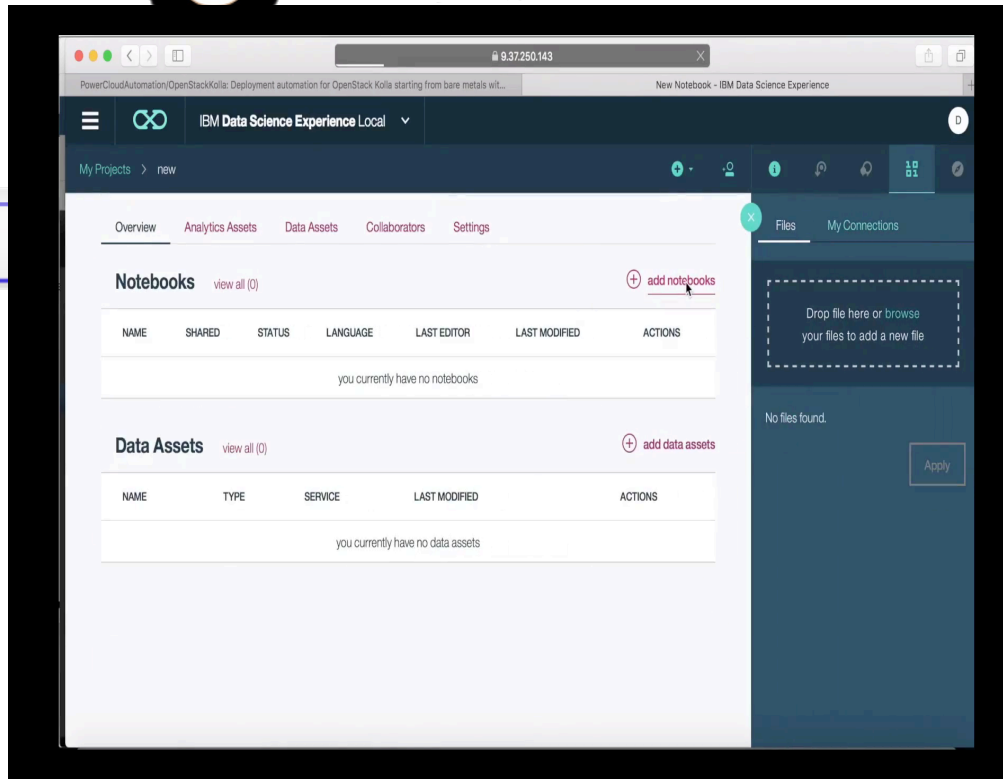
IBM **Data Science Experience**

~~**In my dreams**~~

I'm coding in
an open data science
framework,
running on Spark and
Power

…in minutes

IBM Power
Tools for Ease
of Development



Spark

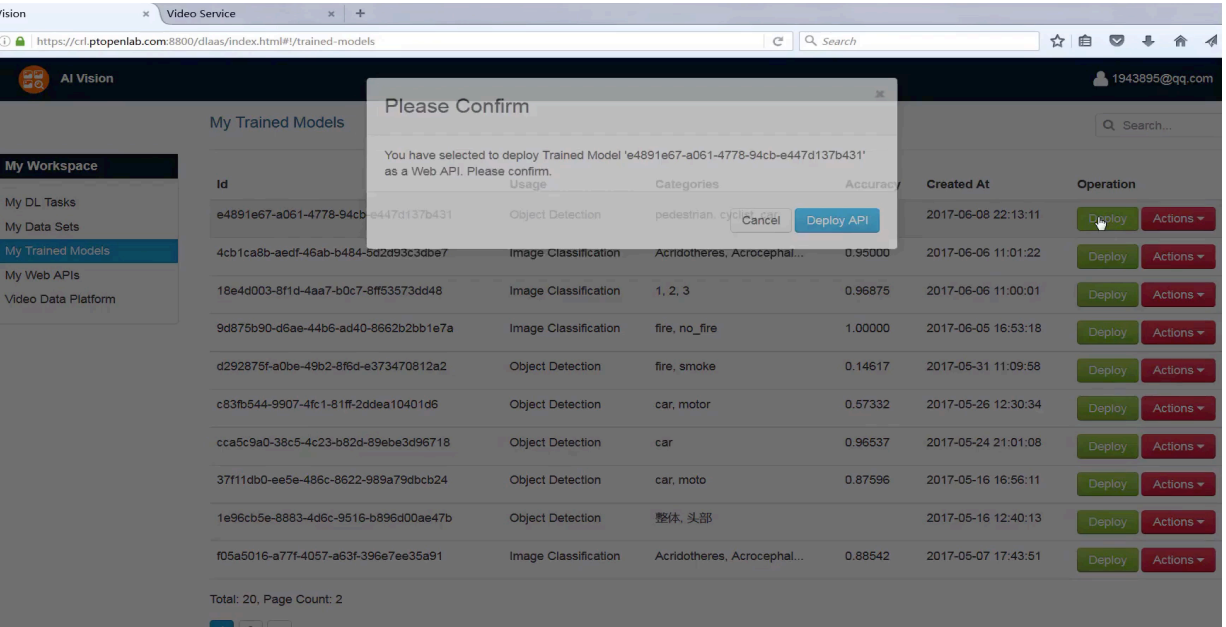
Learn
Create
Collaborate

POWER

IBM Power

Tools for Ease of Development

Build and deploy with mouse clicks



Built training set with a mere 3 MB of video

Deployed as a REST-API with a mouse clicks

Data is inferenced with a single mouse click

Deep Learning

What you and I (our brains) do without even thinking about it…..we recognize a bicycle





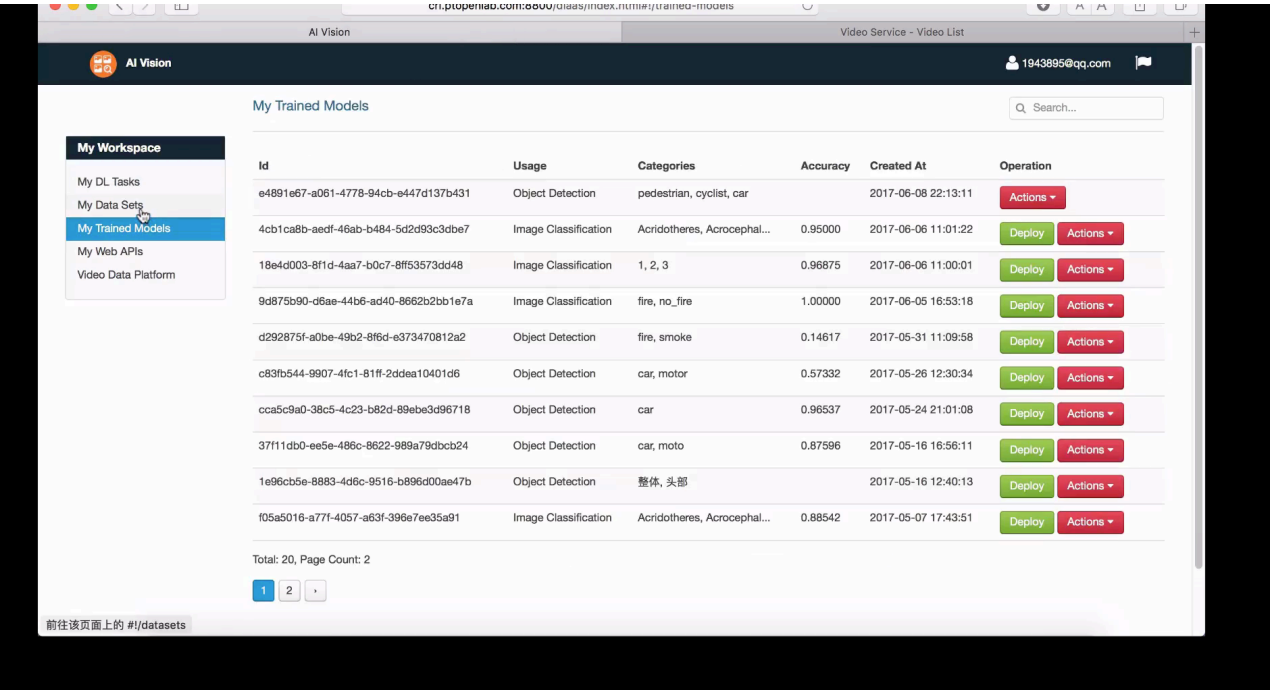



IBM Power

Tools for Ease of Development

Point and click label tuning

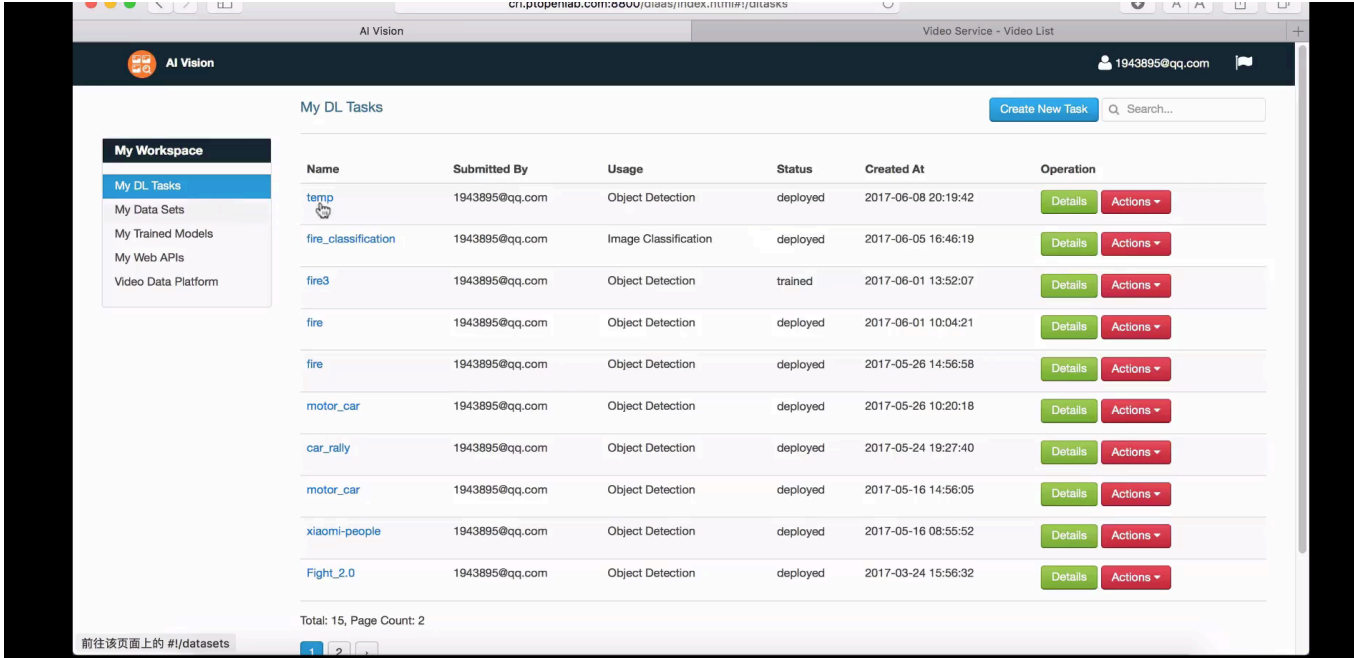
10x faster to create a large labeled data set compared to traditional methods



AI Vision toolset

IBM Power
Tools for Ease of Development

AI Vision assistants help to rapidly iterate over models and makes suggestions along the way to improve a model's accuracy



Create a new task to train the model

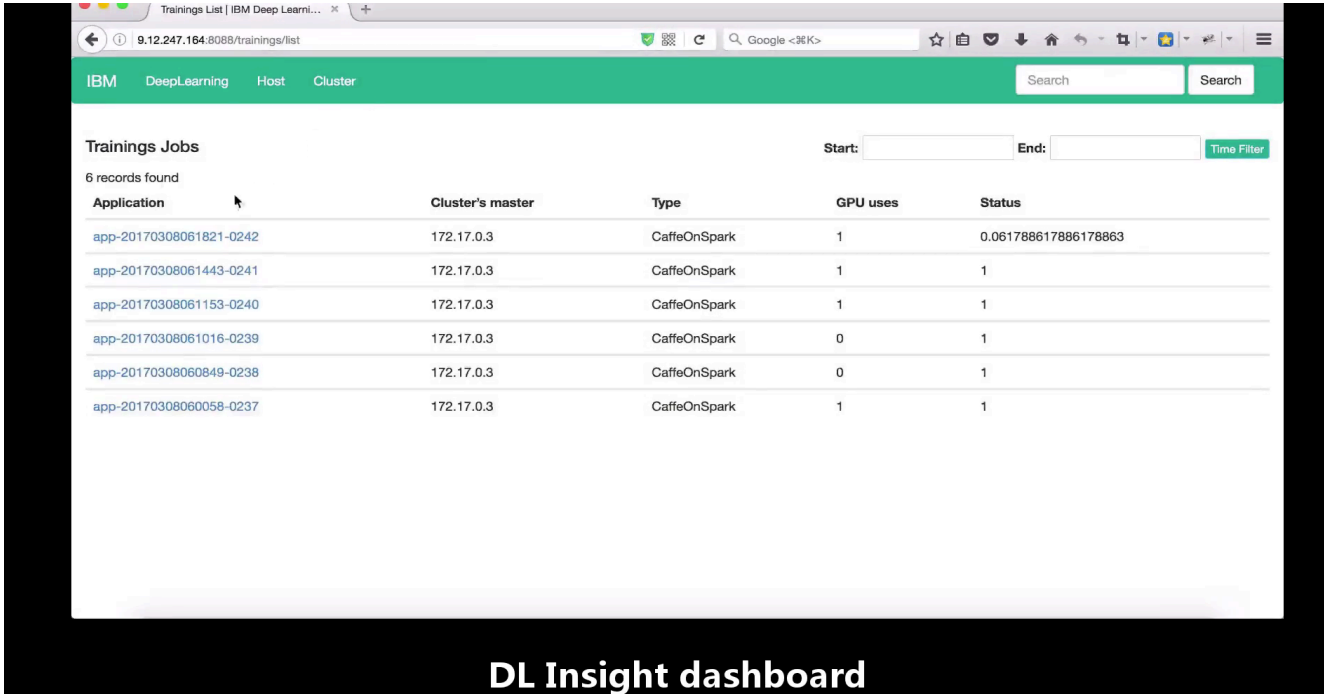
Real time monitoring of hyper parameters

IBM Power
Tools for Ease of Development

Expert optimization advice for hyper parameter selection and tuning

Traffic light alerting for required parameter optimization with early stop advice and more

CPU, GPU, memory utilization info, comms overhead, +++

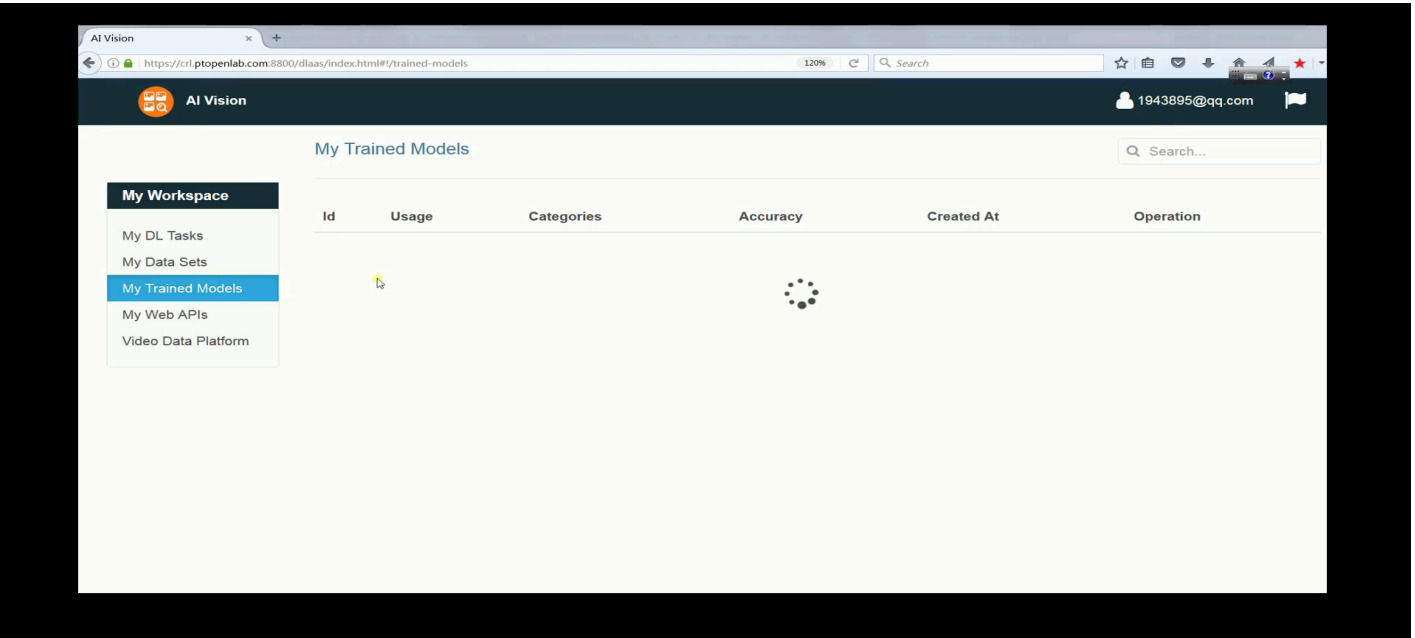


DL Insight dashboard

IBM Power

Tools for Ease of Development

Point and click model deployment



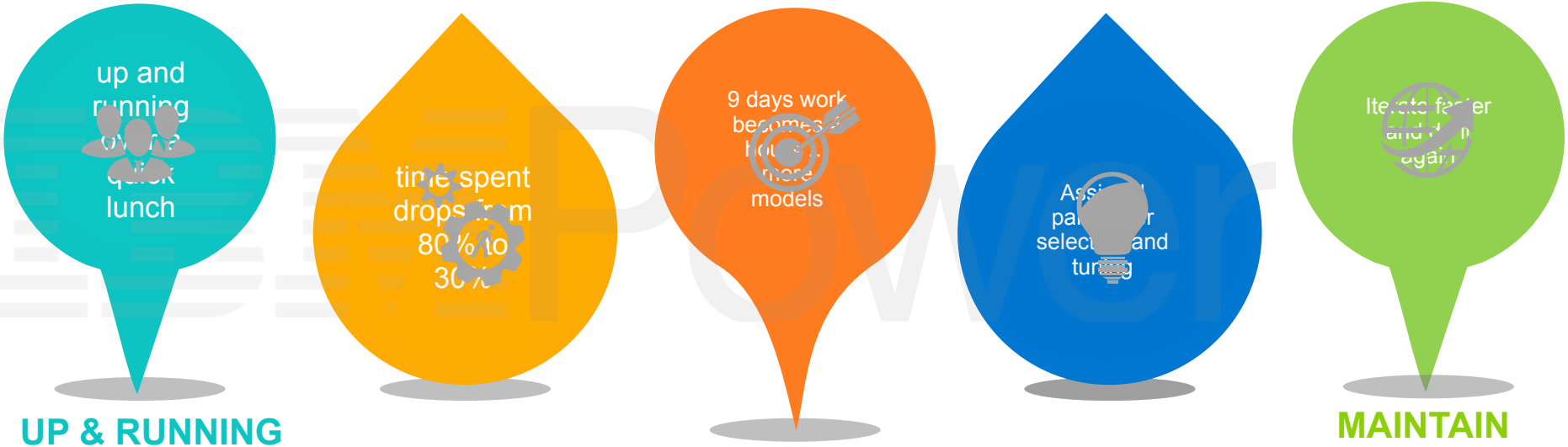
Hardcore coding days become clicks to expose the model as a REST-API endpoint and scored from anywhere

IBM Power


AI Vision in Data Center
Real-time detection of objects for which the Deep Learning model was trained (Car, Motor & Pedestrian)
2013/01/31 13:54:26

DATA PREPARATION
most time spent here

DEPLOY & INFER
requires different skills



UP & RUNNING
weeks to months

BUILD, TRAIN, OPTIMIZE
very iterative

MAINTAIN ACCURACY
experience all that pain again

IBM POWER AC922 pricing

65K street price, including :

- 2 X POWER9 (40 cores)
- 4 X Tesla V100
- 1TB RAM
- NVLink everywhere

Comparable to DGX Station
Superior performance

Superb value for the money!

Deep Learning / AI Enterprise Use Cases

AUTOMOTIVE

Auto sensors reporting location, problems

COMMUNICATIONS

Location-based advertising

CONSUMER PACKAGED GOODS

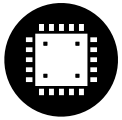
Sentiment analysis of what's hot, problems

FINANCIAL SERVICES

Risk & portfolio analysis
New products

EDUCATION & RESEARCH

Experiment sensor analysis



HIGH TECHNOLOGY / INDUSTRIAL MFG.

Mfg. quality
Warranty analysis

LIFE SCIENCES

Clinical trials

MEDIA/ENTERTAINMENT

Viewers / advertising effectiveness

ON-LINE SERVICES / SOCIAL MEDIA

People & career matching

HEALTH CARE

Patient sensors, monitoring, EHRs

OIL & GAS

Drilling exploration sensor analysis

RETAIL

Consumer sentiment

TRAVEL & TRANSPORTATION

Sensor analysis for optimal traffic flows

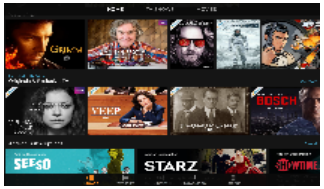
UTILITIES

Smart Meter analysis for network capacity,

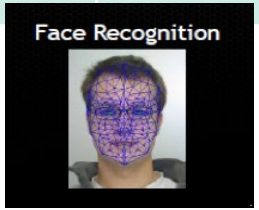
LAW ENFORCEMENT & DEFENSE

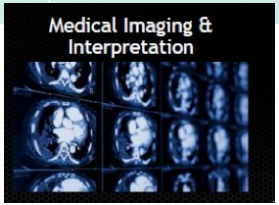
Threat analysis - social media monitoring, photo analysis

Deep Learning in Industries



Automotive and Transportation	Broadcast, Media and Entertainment	Consumer Web, Mobile, Retail	Security and Public Safety	Medicine and Biology
• Autonomous driving: • Pedestrian detection • Accident avoidance	• Captioning • Search • Recommendations • Real time translation •	• Image tagging • Speech recognition • Natural language • Sentiment analysis	• Video Surveillance • Image analysis • Facial recognition and detection	• Drug discovery • Diagnostic assistance • Cancer cell detection
Auto, trucking, heavy equipment, Tier 1 suppliers (Hyundai, Toyota, Komatsu, General Motors, Volvo)	Consumer facing companies with large streaming of existing media, or real time content	Hyperscale web companies, large retail (Google photos, Twitter, Woolworths, Aeon)	Local and national police, public and private safety/ security (ADT, IViz, Pinkerton, Sentry)	Pharmaceutical, Medical equipment, Diagnostic labs (Takeda, Asian Pharma, Pfizer)





Deep Learning in Banking Industry

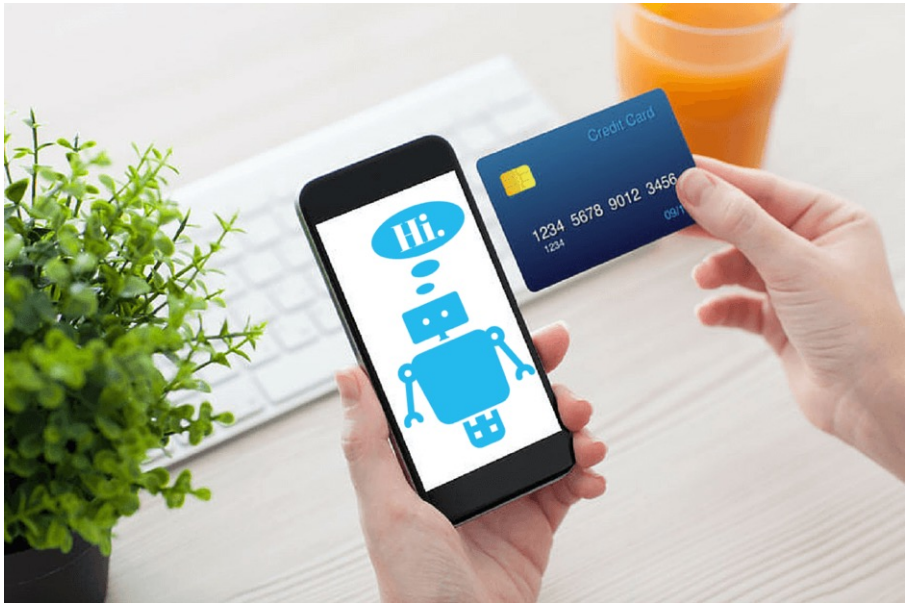
1. Predictive Chat Boots for Customer Support
2. Customer recommendations
3. Fraud Detection
4. Algorithmic trading
5. Credit Risk

http://ieeexplore.ieee.org/document/7359417/

Contextual Chat Boot

What is provided:

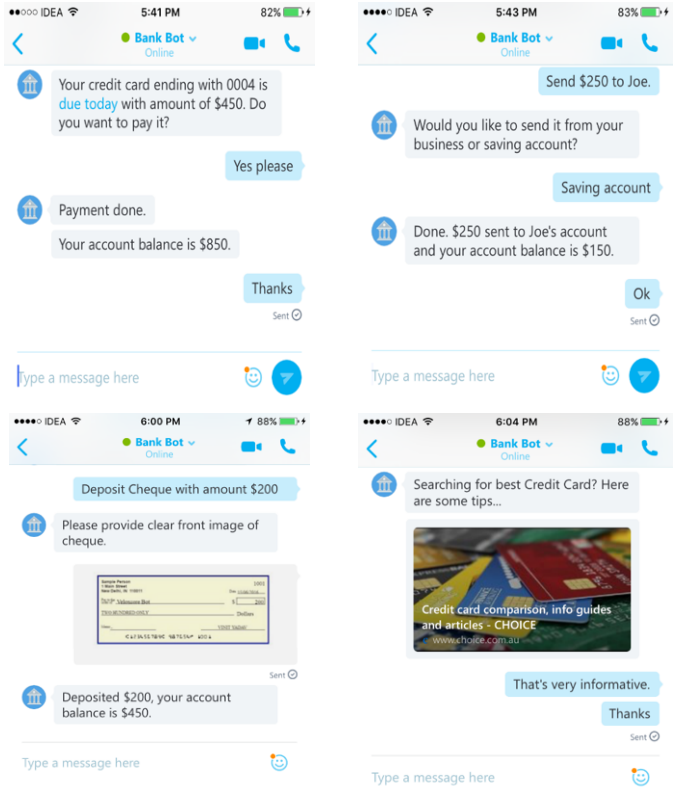
- Ability to search for customer financial data
- Provide excellent answers to the top 5 customer support questions
- Handle other questions reasonably well
- Small talk on a basic level

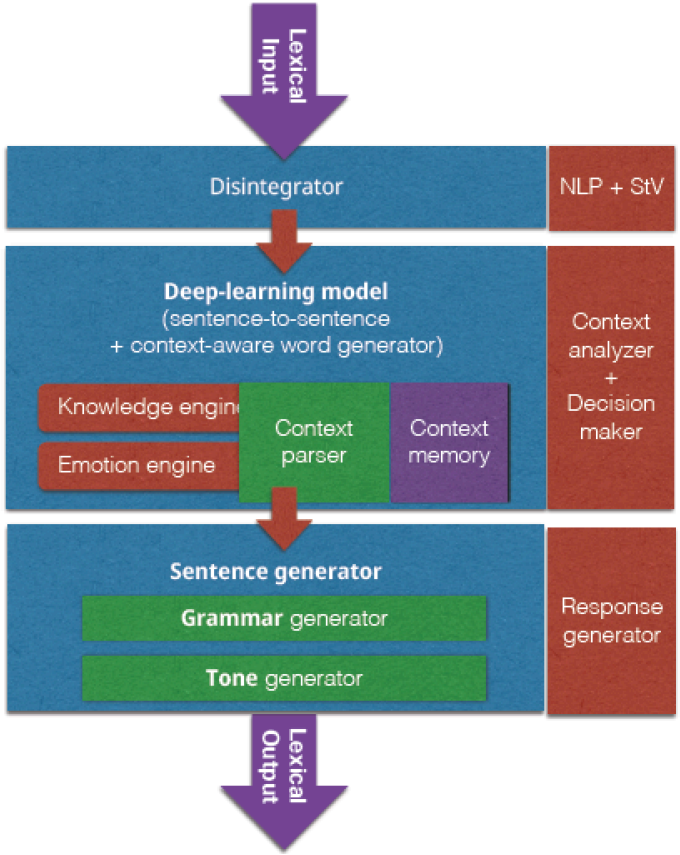


Attributes:

- Enhance
- Understand
- Knowledge
- Decide

Contextual Chat Boot

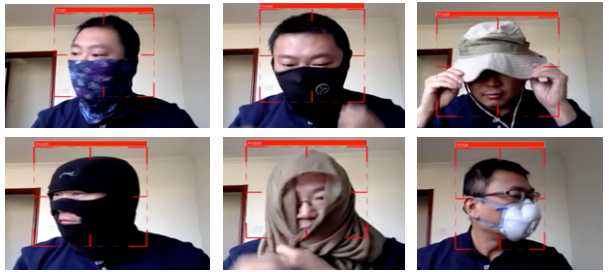




Fraud Detection – Masked Face Detection at ATM's

Challenge

- Unattended ATMs become target of crimes
- Masked face detection could help recognize potential criminal actions and then trigger alarm or limit function
- Traditional pattern identification algorithms are not so effective to resolve such many diversified and changing possibilities



Benefits

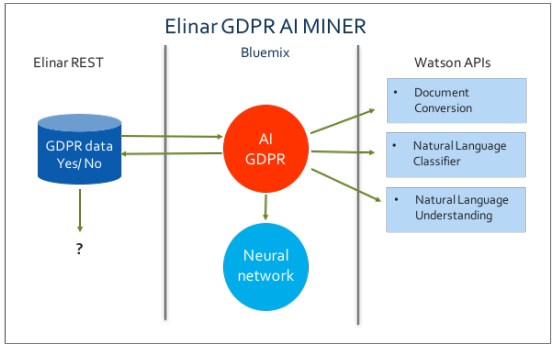
- Masked Face detection for ATM proved production ready and significantly improve the security of ATMs and banks

Solution

- A cluster of IBM Minsky servers (S822LC) with Nvidia P100 GPU and NVLink
- PowerAI stack for the deep learning framework, running over Spectrum Conductor with Spark and Spectrum Scale
- Modeling services of face occlusion detection thru Caffe
- Sampled 150+ face occlusion videos to generate 1500+ images as the training and testing dataset
- Real time video recording and auto-detect face occlusion, support multiple ways of recognition simultaneously

EU GDPR discovery using Deep Learning

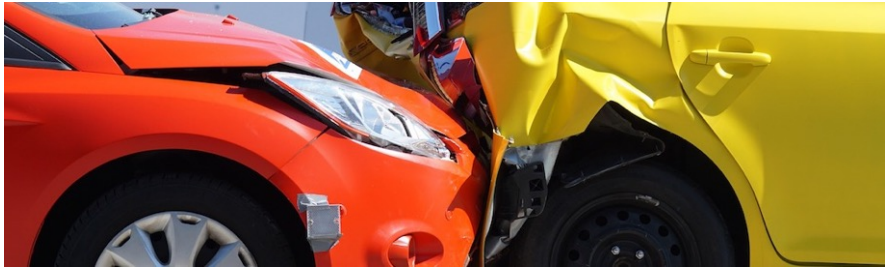
- **Use Case:** *How to find and identify GDPR data of single individuals from multiple data sources?*
- **Addressable Market:** In general (a) public authorities, (b) organizations that engage in large scale systematic monitoring, or (c) organizations that engage in large scale processing of sensitive personal data. The GDPR not only applies to organisations located within the EU but it will also apply to organisations located outside of the EU if they offer goods or services to, or monitor the behaviour of, EU data subjects. It applies to all companies processing and holding the personal data of data subjects residing in the European Union, regardless of the company's location.
- **Offerings:** ELINAR / PowerAI based solution for GDPR discovery using text mining and deep Learning



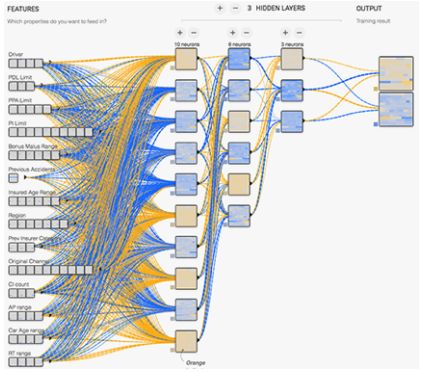
http://on-demand.gputechconf.com/gtc-eu/2017/presentation/23120-ari-juntunen-gdpr-discovery-using-text-mining-and-deep-learning.pdf

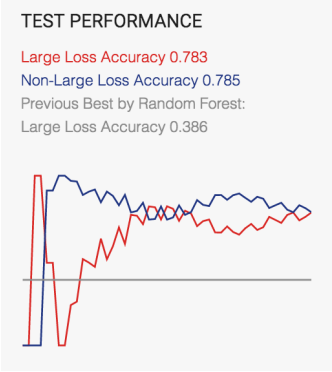
Insurance Price Optimization

Approximately 7-10% of AXA's customers cause a car accident every year. Most of them are small accidents involving insurance payments in the hundreds or thousands of dollars, but about 1% are so-called large-loss cases that require payouts over $10,000. As you might expect, it's important for AXA adjusters to understand which clients are at higher risk for such cases in order to optimize the pricing of its policies.



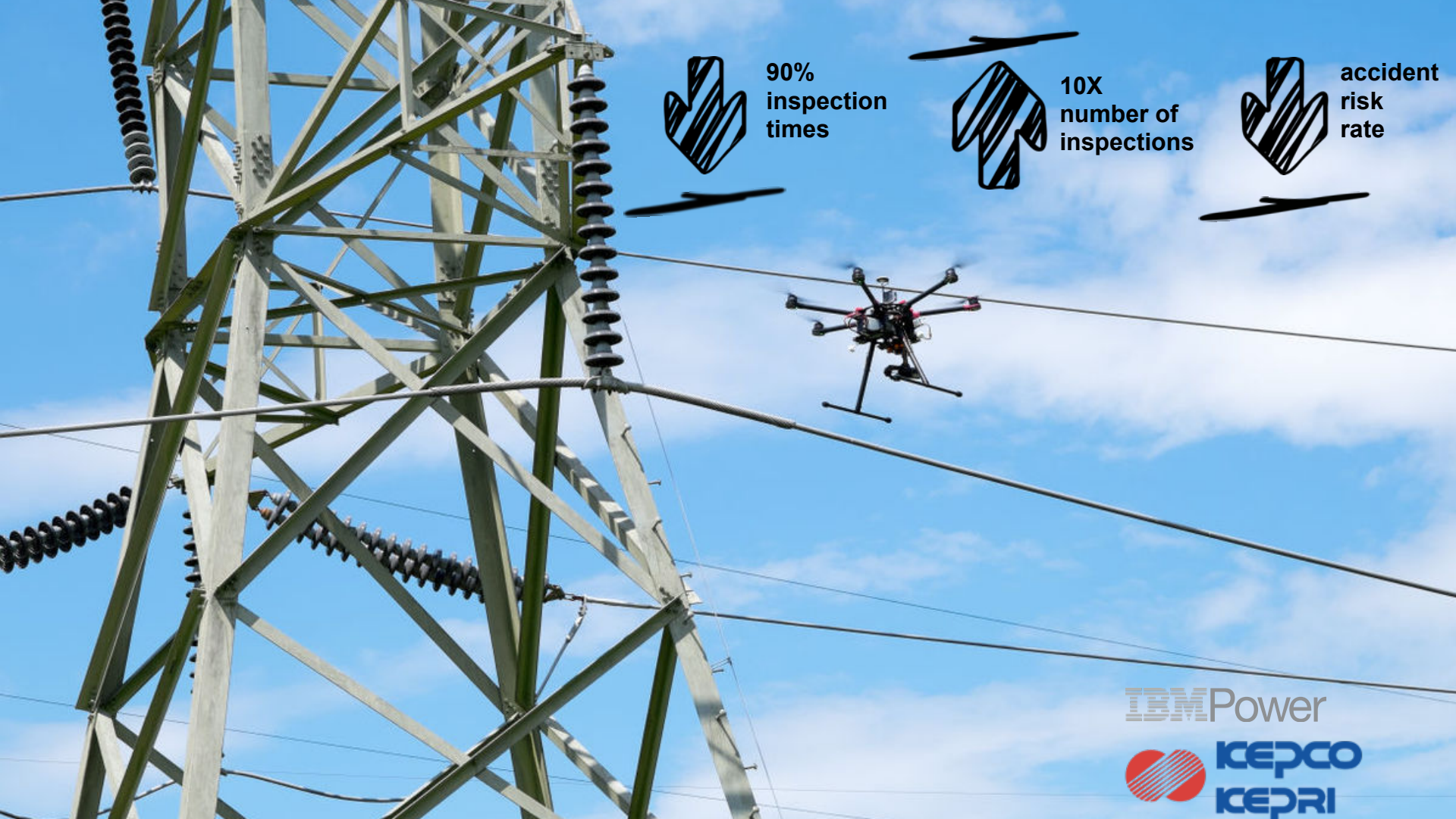
At the right, you can see there are about 70 values as input features including the following.
•Age range of the driver
•Region of the driver's address
•Annual insurance premium range
•Age range of the car





AXA entered these features into a single vector with 70 dimensions and put it into a deep learning model in the middle. The model is designed as a fully connected neural network with three hidden layers, with a ReLU as the activation function. This use case has end with a accuracy of 78%

https://cloud.google.com/blog/big-data/2017/03/using-machine-learning-for-insurance-pricing-optimization


90%
inspection
times
10X
number of
inspections
accident
risk
rate
IBM Power
KEPCO
KEPRI

CREDIT SUISSE

AI as a Service – Fraud Surveillance

Financial services organisation evaluates high-performance PowerAI solution to combat fraud.

Winning Solution

- IBM Power System S822LC
- IBM Power AI Development Platform
- RHEL

Total Contract Value: $92k

Competition: Nvidia DGX-1

Solution Benefits

- The shared, standardised & agile on prem Cloud AI PaaS Service offers CS LoBs cost benefits & time to market competitive advantage for new innovative AI services
- Significantly reduce the cost of Fraud & Fraud Surveillance & Detection
- Time to value with one click installation
- Level 3 support based on Enterprise grade AI development platform
- Significantly improved model accuracy enabled by large ML/DL model support leveraging IBM unique hardware acceleration architecture
- Significantly reduced model training times leveraging unique Distributed Deep Learning capability & graphical monitoring of training accuracy enabled by Deep Learning Impact
- Most cost effective Distributed Computing Framework – IBM Spectrum Computing

Business Partner: Recarta

About the customer

Credit Suisse is one of the world's leading financial services providers and offers clients its combined expertise in the areas of private banking, investment banking and asset management. Credit Suisse provides advisory services, comprehensive solutions and innovative products to companies, institutional clients and high-net-worth private clients globally, as well as to retail clients in Switzerland. Credit Suisse is headquartered in Zurich and operates in about 50 countries worldwide.

Business challenge and solution

Credit Suisse has been providing low latency HPC Monte Carlo Simulation PaaS services based on IBM's market leading workload scheduling & shared resource management middleware platform (Symphony); & Data Analytics PaaS services based on Hadoop/ MapReduce to the Capital Markets Trading LoBs for many years; to support Risk Analytics, Pricing & FRTB driven XVA calcs . These mission critical services are deployed on a global 60,000 + core commodity computing grid cluster, with a few nodes using GPU hardware acceleration for specific Trading Risk analytics use cases. Within the last 6 months, the CS SAMI organisation has extended this PaaS capability to include a ML/DL AI PaaS service for Fraud Surveillance & Detection of all Trading email communication, using NLP algorithms based on the Theano framework. The challenges CS are facing is to scale the ML/DL AI PaaS service efficiently, to meet the performance demands of an increasing number of ML/DL uses across the firm, from Trading Risk Analytics & Market Predictions, extending the Fraud use case to include voice to text; to many operational efficiency use cases. The solution undergoing evaluation is IBM's enterprise grade AI application development framework, PowerAI, Vision AI, Distributed Deep Learning & Deep Learning Impact, leveraging IBM's unique hardware acceleration platform Power8 Minsky, based on Nvidia GPUs & NvLink technology.

Why did the Client choose IBM Systems

CS have selected IBM's AI platform for evaluation as the AI platform of choice, based on Vision & Strategy alignment with IBM's highly differentiated SDI Vision & Strategy. This leverages a common middleware workload scheduling & shared resource management platform, IBM's Spectrum Computing Framework. This uniquely manages not only traditional compute intensive low latency HPC & data intensive analytics workloads, but also large scale ML/DL AI model workloads, leveraging IBM's unique Power8 Minsky hardware acceleration capability, allowing CS to meet the wide range of AI use case performance demands cost effectively.

Winning Solution
IBM Power AI AC922

Total Contract Value:
50k$

Competition:
x86

Solution Benefits:
- 2x performance over x86
- NVLink technolgy with communication between CPUs and GPUs

Business Partner:
UMB

Use Case Contact: Rene Bersier

About the customer
LGT Group is the private banking and asset management group of the princely House of Liechtenstein. Originally known as The Liechtenstein Global Trust, LGT is the largest family-owned private wealth and asset manager in the world, wholly owned by the Prince of Liechtenstein Foundation. LGT is headquartered in Vaduz, Liechtenstein

Business challenge and solution
LGT Bank want to handle their data faster and want to get more out of their data to save time and use also new ways of AI. They have seen our new announcement from the PowerAI AC922 and liked the idea to start with deep learning to program neural networks and to analyse their data faster. Their idea is to analyse their data and also to learn more about deep learning. Specific use cases will be defined in March/April.

Why did the client choose IBM Systems
LGT Bank have bought one E880C and get the PowerAI included. With the new Announcement they understand how easy it is to set up Open Source Frameworks, which are pre-packaged and easy to implement. They like the fact that large Deep Learning jobs can be clustered across serval servers with PowerAI. The key for them was the benefit in the NVLink technology, which creates a direct, ultra-fast connection between CPU and GPU, with up to 5.6 x faster communication. With that they get a faster output of their data.

OTP Bank Hungary

GPU accelerated server win

Financial services institution evaluates GPU accelerated computing in several use cases

Winning Solution
2 pcs of AC922 servers, each with:
- 32 core Power9 processor
- 1 TB memory,
- 4x NVIDIA Volta GPU,
- 100Gb EDR IB connection

Total Contract Value: $ 110,000

Competition
- x86 vendors

Solution Benefits
- NVLink integrated GPU cards offer superior performance in the same price range as x86 competitors

Business Partner: InterComputer

Use-case contact: Jozsef Suranyi (IBM HU)

About the customer
Owing to economic and legal considerations, OTP Group provides its universal financial services through several subsidiaries. In Hungary, traditional banking operations are performed by the Bank while specialized services, including car leasing and investment funds are developed and offered by the Bank's subsidiaries. OTP Bank has completed several successful acquisitions in the past years, becoming a key player in the region. OTP Group currently operates in Bulgaria (DSK Bank), Croatia (OTP banka Hrvatska), Romania (OTP Bank Romania), Serbia (OTP banka Srbija), Slovakia (OTP Banka Slovensko), Ukraine (CJSC OTP Bank), Russia (OAO OTP Bank, Donskoy Narodny Bank) and Montenegro (Crnogorska komercijalna banka AD) via its subsidiaries.

Business challenge and solution
OTP Bank has a small, independent special “Research and Innovation” IT team to test and evaluate leading IT technologies. They purchased a Power8 Minsky server in 2016 and tested under several workloads, ranging from in-house developed GPU assisted algorithms to GPU database for relational or non-relational data. Convinced by the superior performance of the server, they ordered the Power9 follow-on within two weeks of the announcement. They will further test MapD Core GPU db on Power and also porting Ethereum blockchain technology to Linux on Power.

Why did the Client choose IBM Systems
Designed to deliver extremely high performance for resource-intensive workloads, the IBM Power System AC922 is the only architecture with the latest NVIDIA NVLink technology, creating a direct, ultra-fast connection between CPU and GPU. AC922 is available on the same price range as brand x86 servers, but delivers superior performance and technical design.

AI in Belux

- AI Brussels meetup : 1175 members
- AI Mons : 179 members
- AI Ghent : 50 members
- Machine learning Luxembourg : 339 members
- Data science Luxembourg : 1605 members
- Applied deep learning (Antwerp) : 143 members
- TensorFlow Belgium : 892 members
- **AI 4 business conference** in Brussels on February 27, IBM presence

PowerAI in Europe

- Nordics : nearly 700 members in total (4 cities)
- London : 587 members
- Brussels : created 21/2

RoboVision

ROBOVISION
DEEP LEARNING APPLIED

Located in Ghent

12 employees, all data scientists / AI experts

Use cases in manufacturing, finance, agriculture, retail

References : Audi, BNP Paribas Fortis

Partnering with IBM since Q2 2017

POC in Montpellier/Poughkeepsie in Q3

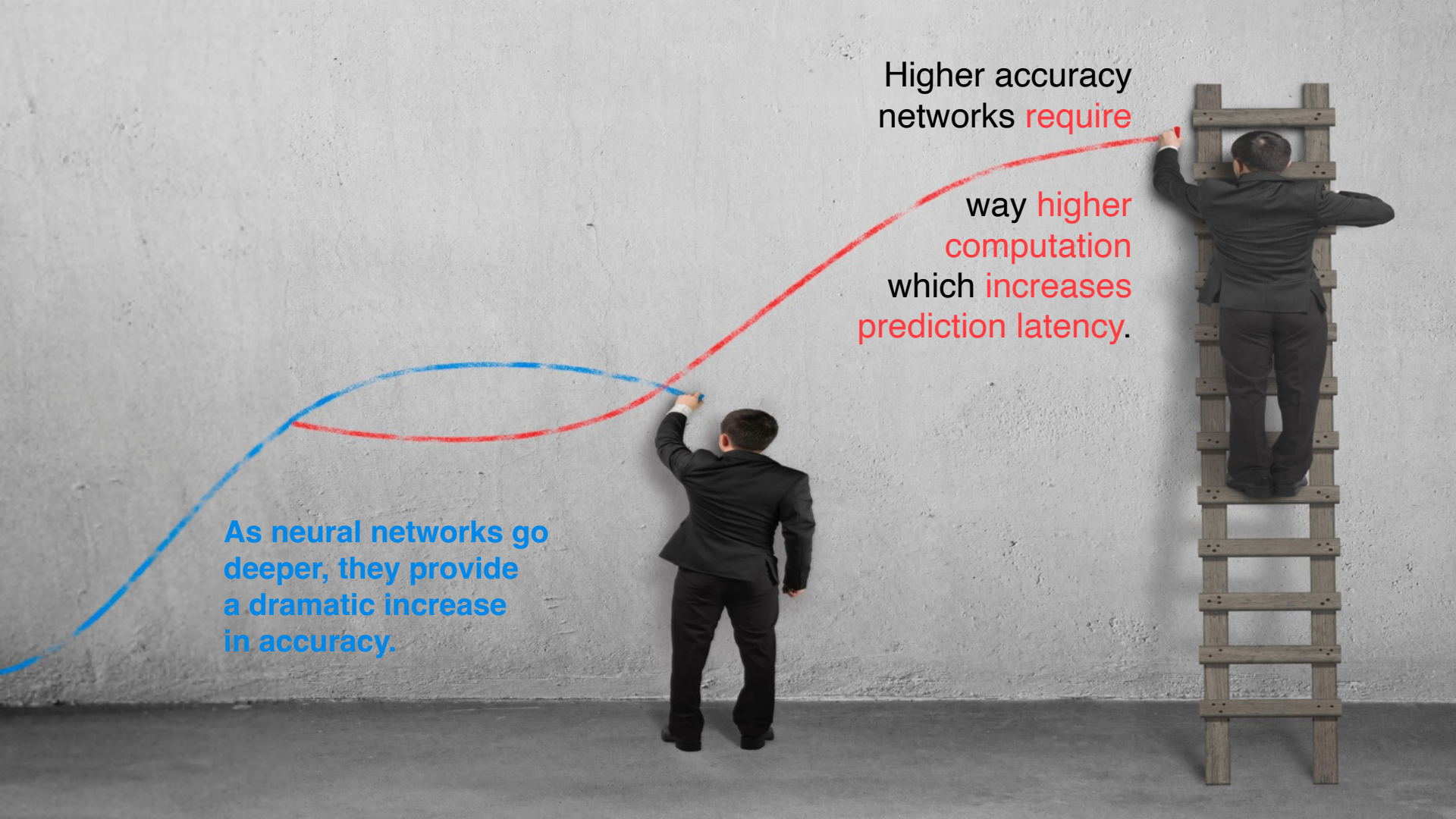
On premise POC in Q4/Q1 2018

In an independent benchmark by industry expert, Brytlyt's GPU Database outperformed all other vendors by a factor of four or more with its PostgreSQL fork tapping into the super computing power of IBM Minsky Hardware.

Based on PostGreSQL
Easy to integrate into existing landscape
Rich functionality
Results based on (sub-optimal) POWER8 hardware

https://www.brytlyt.com/blog/brytlyt-gpu-database-smashes-benchmark-record/
http://tech.marksblogg.com/billion-nyc-taxi-rides-brytlytdb-ibm-minsky.html?utm_source=brytlyt%20website&utm_medium=MinskyBenchmark

brytlyt


Higher accuracy networks require way higher computation which increases prediction latency.
As neural networks go deeper, they provide a dramatic increase in accuracy.