IBM **Client Center** Montpellier

Private AI Architecture
PowerAI Portfolio and Strategic Directions

Benoit MAROLLEAU - Cloud Architect
Cognitive Systems
IBM Client Center Montpellier, France
benoit.marolleau@fr.ibm.com

linkedin.com/in/benoitmarolleau

@MarolleauBenoit

IBM PowerAI

IBM

Agenda

What is AI ? Machine Learning ? Deep Learning?
AI = Cloud ?

IBM “One AI” & PowerAI Solutions
PowerAI (WML-CE) , PowerAI Enterprise (WML-A) & AC922 HW

Private AI Solutions Overview

- PowerAI Vision & Intelligent Video Analytics
- h2O.ai Driverless AI
- Watson Studio & ICP for Data

Appendices/Bonus to go further:

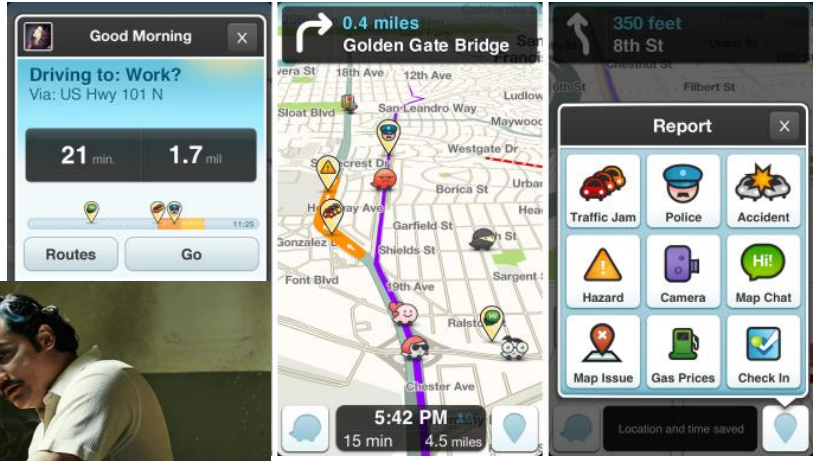
AC922 details & pricing, PowerAI Public References
Demos & Illustrations:

- AI with IBM Cloud Private w/ PowerAI, AI Vision, H2O.ai, Watson Studio (DSX) Local :
- H2O.ai Driverless AI quick demo on ICP/POWER
- DSX & Predictive Maintenance Example

IBM

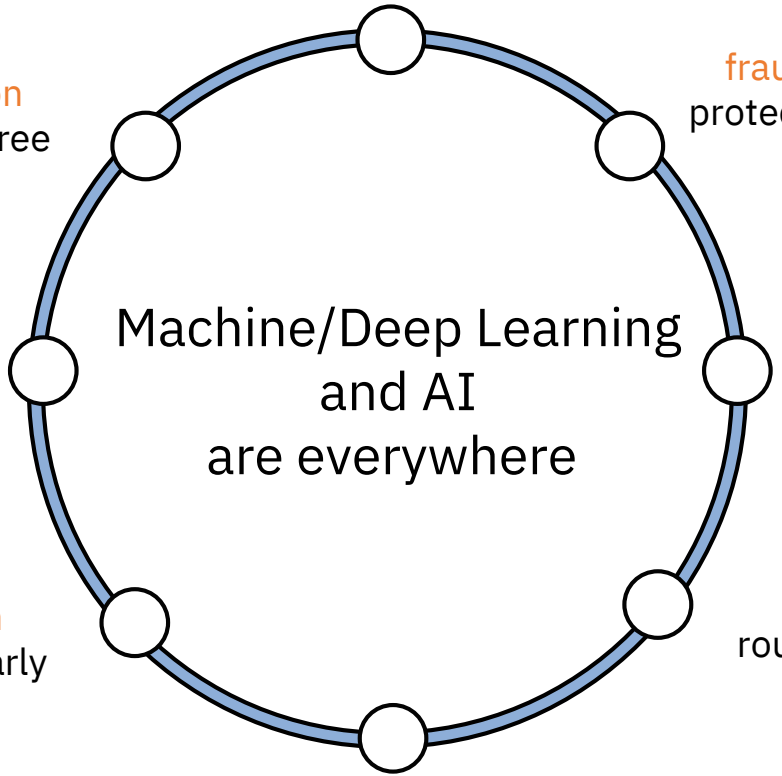
Machine learning is everywhere – influencing nearly everything we do

Netflix provides personalized movie recommendations



Waze provides a personalized driving experience for its users

The future is now



facial recognition
unlocks your phone

fraud detection
protects your credit

recommendations
help you shop faster

chat bots
route calls quicker

spam detection
unclogs your Inbox

machine vision
detects cancer early

autonomous vehicles
detect pedestrians

speech recognition
lets you go hands-free

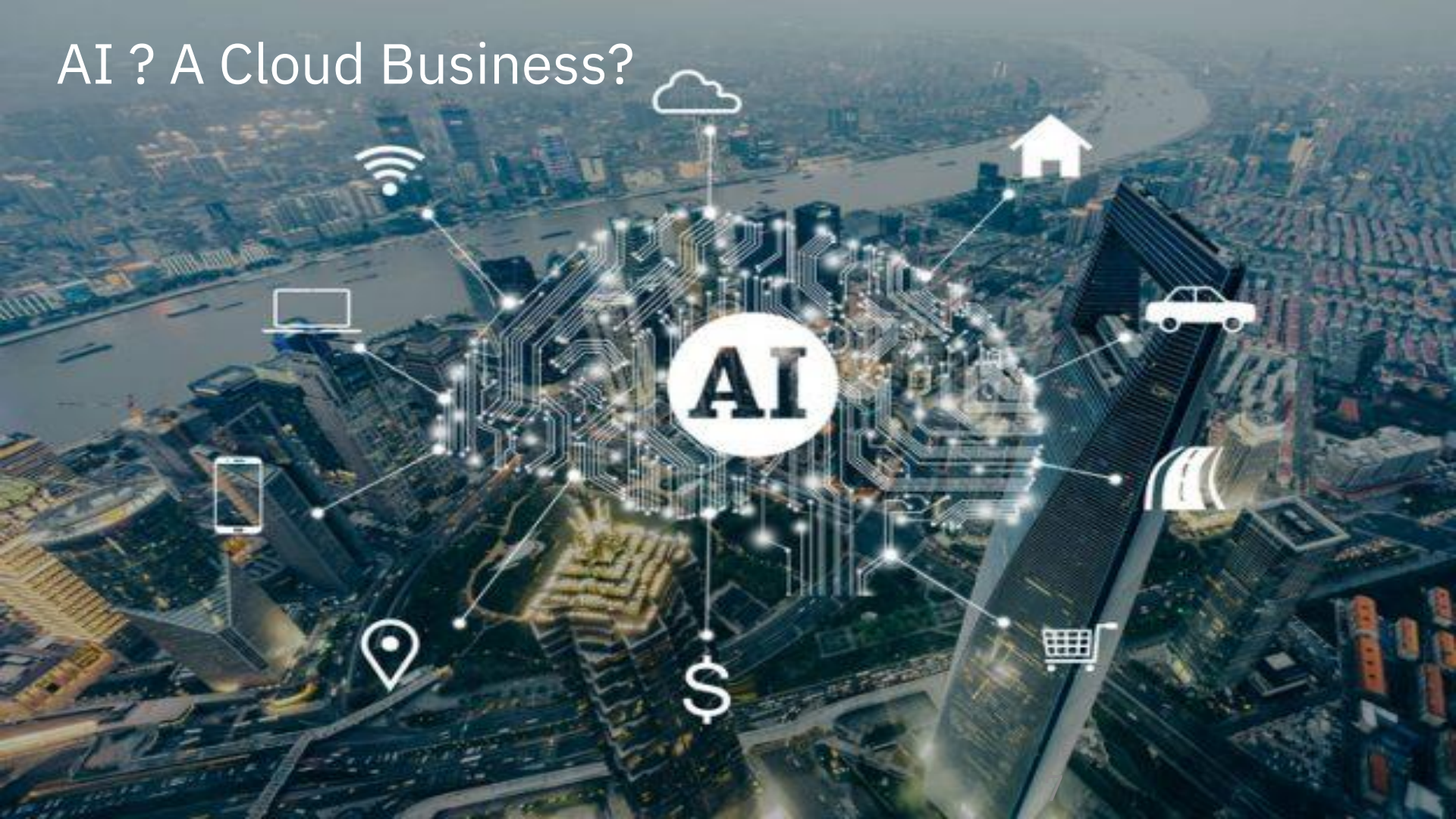
ML Use Cases?

Marketing: AI for Real Time Data
Retail Sales / Automotive: AI for Voice and Image Search coupled with AR/VR
Customer Support: AI for Natural Language (NLP)
Manufacturing: AI Powers Smart Robots (Vision) – Predictive Maintenance – Asset Inspection
Supply Chains: AI for Management – Predictive Maintenance
Information Technology Management: AI Helps Routing (Pattern)
Cybersecurity: AI Protects Assets
Financial Service: AI Enables Intelligent Processing
Life Sciences and Medicine: AI Leverages Algorithms (Vision, NLP)
Smart Cities: AI Optimizes Myriad Functions (Cognitive IoT)

500+ POWERAI CLIENTS WITHIN ONE YEAR

40% NEW TO POWER

AI ? A Cloud Business?



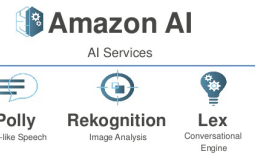
AI, a Cloud Business? Yes but…

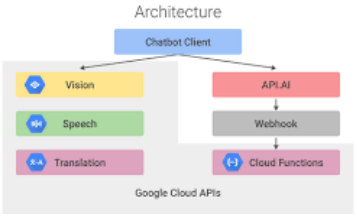
Set of Pre-trained Models
API Invocation
Ready to use
Majority of use cases

Additional Offerings to build new Models
using State of the art Frameworks

Limitations:
Data Gravity
Compliance & Regulations
…

=> AI will be **Public** + **Private** Cloud





Entity Extraction
Sentiment Analysis
Emotion Analysis (Beta)
Keyword Extraction
Concept Tagging
Taxonomy Classification
Author Extraction
Language Detection
Text Extraction
Microformats Parsing
Feed Detection
Linked Data Support
Concept Expansion
Concept Insights
Dialog
Document Conversion
Language Translation
Natural Language Classifier
Personality insights

Big Data Sources

Business Data (Sensitive, Personal Data)

Enterprises Embracing Open-Source AI Software

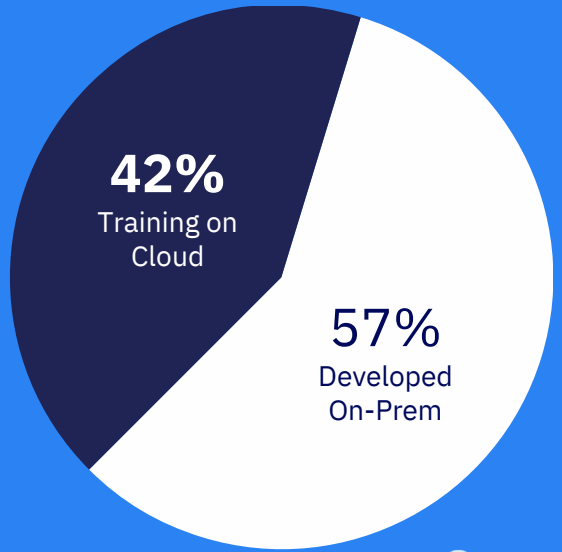
Enterprises building
Data Science teams

Most using Open-Source software:
TensorFlow is most popular

$14B for AI Servers
+ ~$50B for AI Software
(2021, IDC)

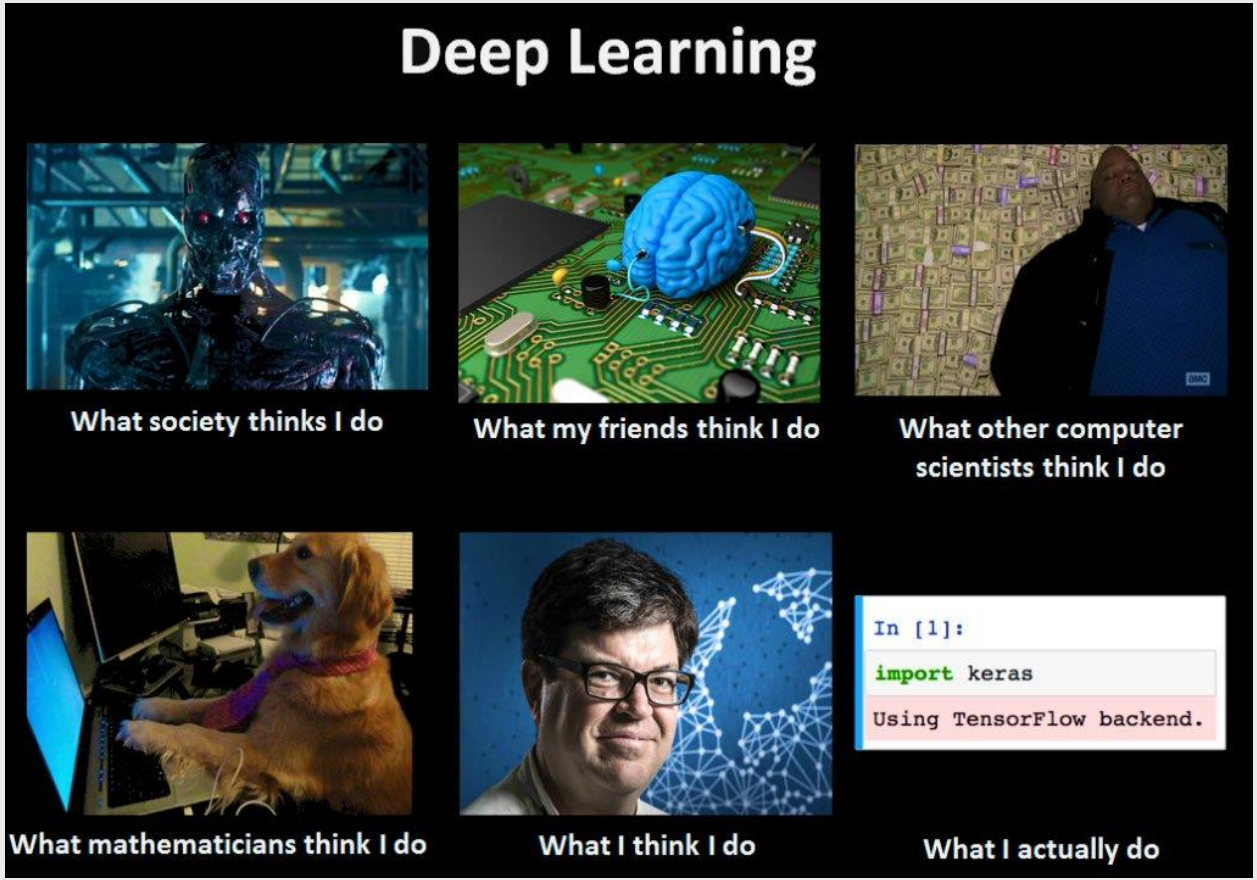
57%

AI Developed
On-Prem



Gartner

What is Machine Learning ? Deep Learning?

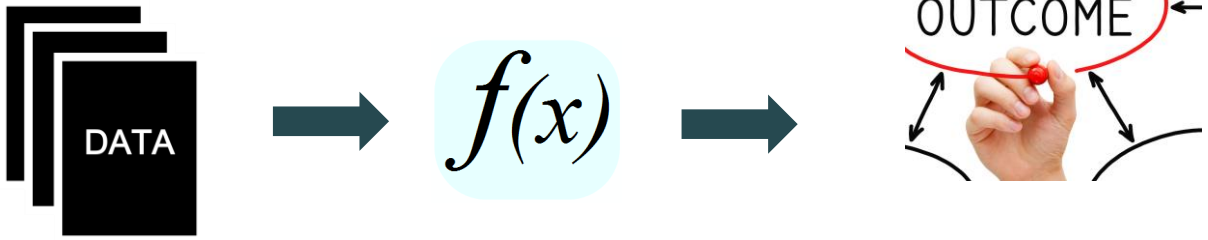


Decision Making

- Credit card transaction
- Loan application
- MRI image
- House data

- Fraudulent vs. legitimate
- Approve vs. reject
- Tumor benign vs. malignant
- House appraisal value

Machine Learning



- Credit card transaction
- Loan application
- MRI image
- House data

- Fraudulent vs. legitimate
- Approve vs. reject
- Tumor benign vs. malignant
- House appraisal value

- **Representing pattern by a mathematical function**
- **Machine learning is just a bunch of math**

Data – Estimate House Price

Sq Ft	Bedroom	Bathroom	Price
2000	3	2	$350,000
1500	2	2	$280,000
2200	3	3	$400,000
…	…	…	…

- Every column except last is a feature
- Last column is a label
- This is a labeled data set

Data – Loan Application

Feature	Feature	Feature	Feature	Label
Loan Requested	**Income**	**Own House**	**Outstanding Debt**	**Decision**
$20,000	$100,000	Y	0	Approve
$50,000	$70,000	N	$20,000	Reject
$5,000	$150,000	Y	$10,000	Approve
...	...	...	...	...

$$\begin{matrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{matrix} \quad \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{bmatrix} \quad \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix}$$

Machine Learning Objective

Use training data to derive f(x) so that

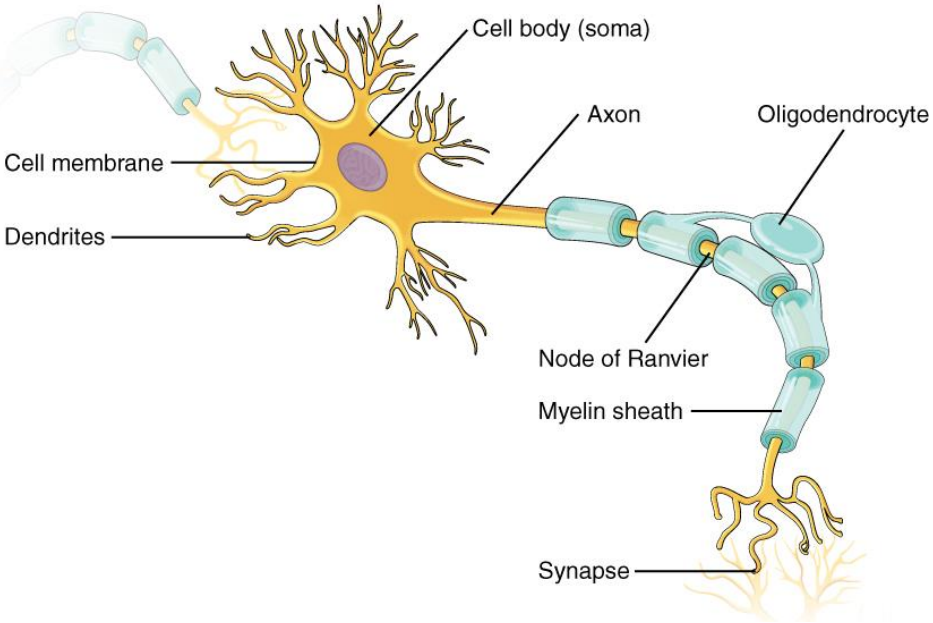
Minimize (Actual - f(x))

or mathematically

$$\min \; 1/n \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2$$

Deep Learning = Training Artificial Neural Networks

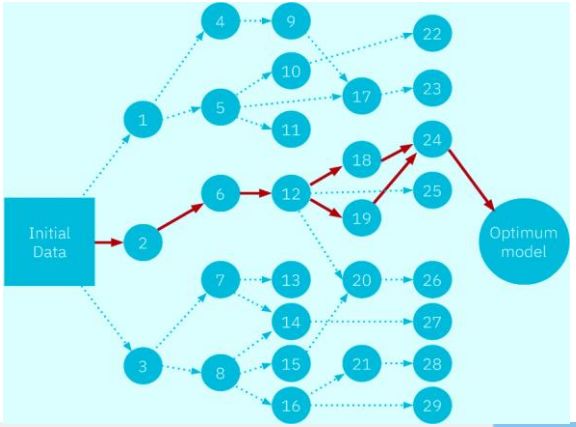
Based on biological neurons. Artificial neurons learn by recognizing patterns in data.



A human brain has:

- 200 billion neurons
- 32 trillion connections between them

➞ Artificial neural networks have far fewer

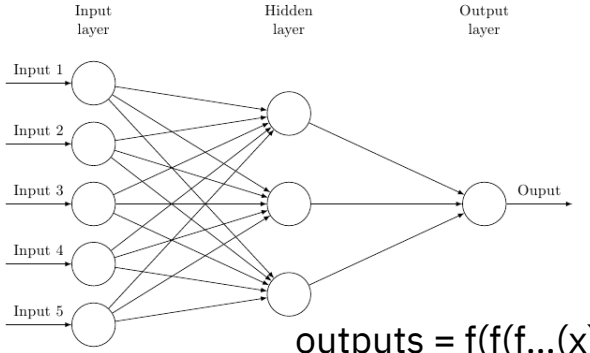


Deep Learning = Training Artificial Neural Networks

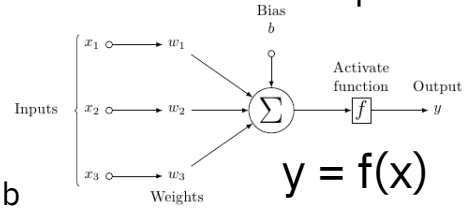
Based on biological neurons. Artificial neurons learn by recognizing patterns in data.

Neural net Framework
& Python Programming

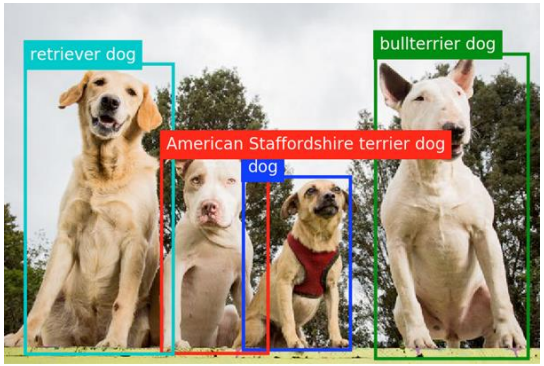
Each input value is a pixel
RGB Value (3 Channels)



outputs = f(f(f...(x))



y = f(x)

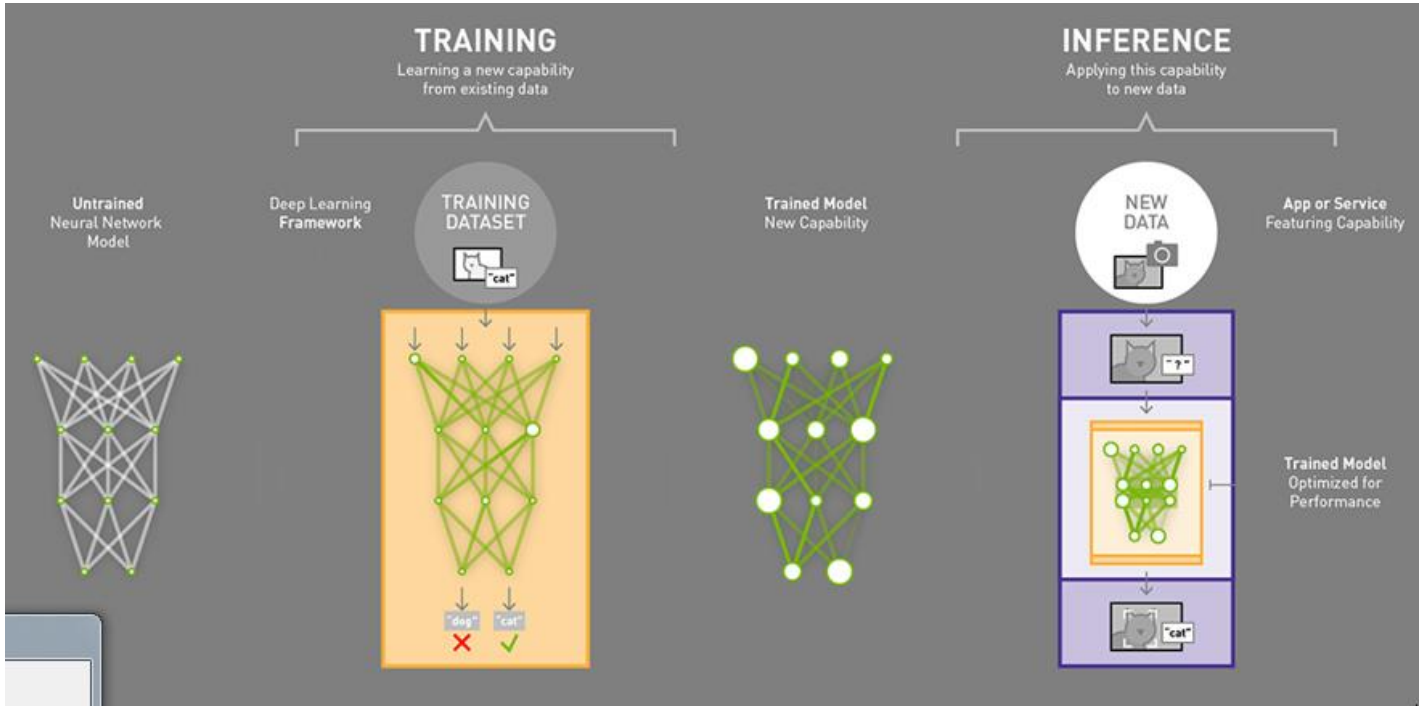


Each output contains:
detection_boxes
detection_scores
detection_classes
num_detections

Training = Adjust the parameters (w,b)
To get the best validation results
Vs. ground truth (Training set).
Inference = apply the model to the input (w,b fixed) to get the output (classification, detection, ...)

Deep Learning = Training Artificial Neural Networks

Based on biological neurons. Artificial neurons learn by recognizing patterns in data.

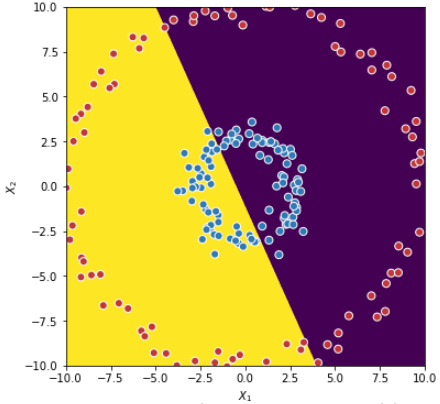


18

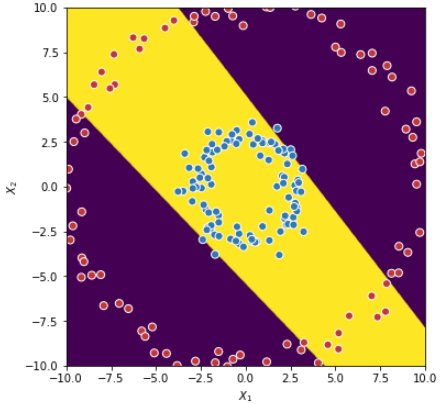
Deep Learning = Training Artificial Neural Networks

Example solving a non linear classification problem

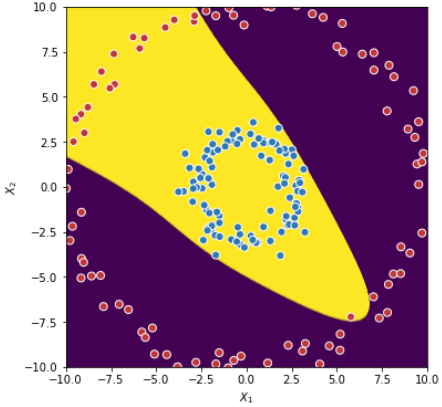
boundary formation with 1 layer and non linear “relu” activation function



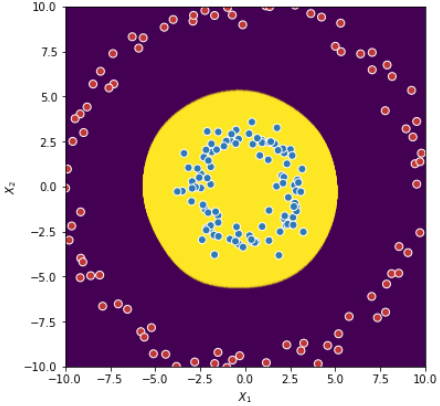
mlp = MLPClassifier(hidden_layer_sizes=(1),
max_iter=500000,activation=**’identity’**,learning_rate_init=0.01)



1 layer and 2 neurons

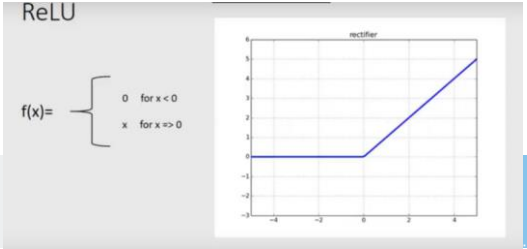


1 layer and 3 neurons



Decision plane with 30 neurons

Rectified Linear Unit (ReLu)

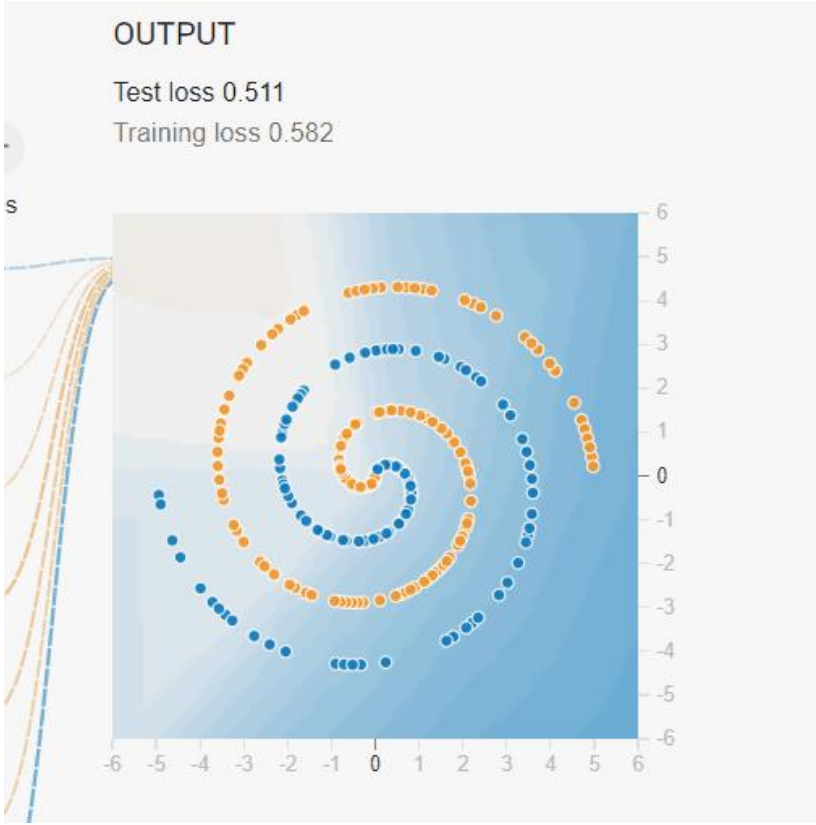


[Neural Nets Simulator:
http://playground.tensorflow.org](http://playground.tensorflow.org)

Deep Learning = Training Artificial Neural Networks

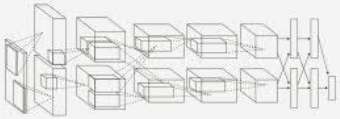
Example solving a non linear classification problem

boundary formation with 3 layers and ‘relu’ activation



Neural Nets Simulator:
http://playground.tensorflow.org

Convolution Networks



Encoder/Decoder

ReLu

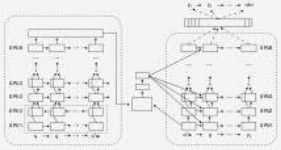
BatchNorm

Concat

Dropout

Pooling

Recurrent Networks



LSTM

GRU

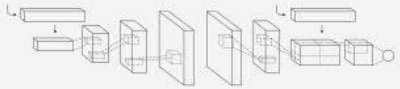
Beam Search

WaveNet

CTC

Attention

Generative Adversarial Networks



3D-GAN

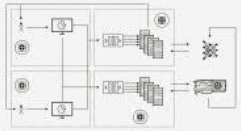
MedGAN

ConditionalGAN

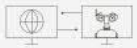
Attention

Speech Enhancement

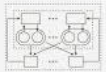
Reinforcement Learning



DQN

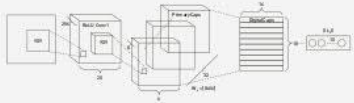


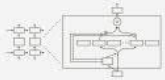
Simulation



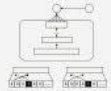
DDPG

New Species





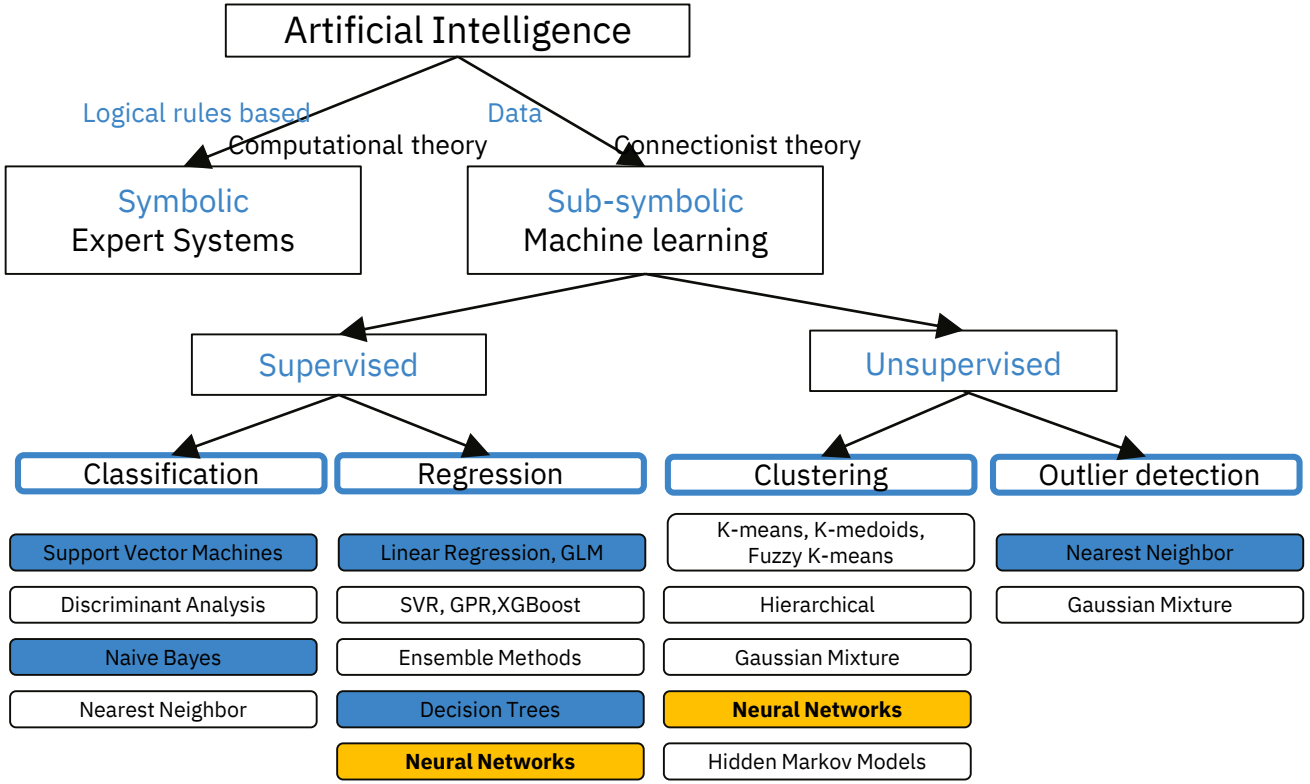
Mixture of Experts



Neural Collaborative Filtering

Block Sparse LSTM

The Machine Learning Tasks & algorithms



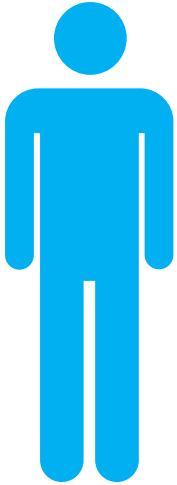
22

Humans

2011

26% Errors

Machine Learning Based



5% Error

2016

3% Errors

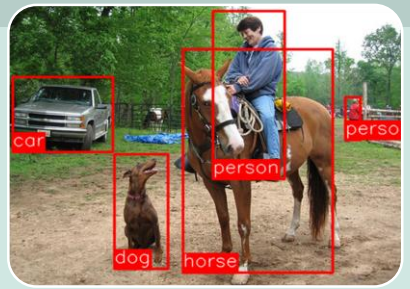
Deep Learning Based

CNN Models detect more and more details in images/videos

Classifier

- Dog: 98%
- Cat: 20%

- Ex : VGG, ResNet, GoogleNet, AlexNet …



Object Detection

- Car 1 : 90% {x1,y1,x2,y2}
- Dog 2 : 80% {x3,y3,x4,y4}
- Person 1: 80% {x1,y1,x2,y2}
- Person 2 : 70% {x1,y1,x2,y2}
- Horse 1 : 70%

Semantic segmentation

- Category « car » : pixel fields 1 90%
- Category « Road » : pixel field 2 97%
- Category « person » : …

- Ex : U-Net, PSPNet, DeepLab

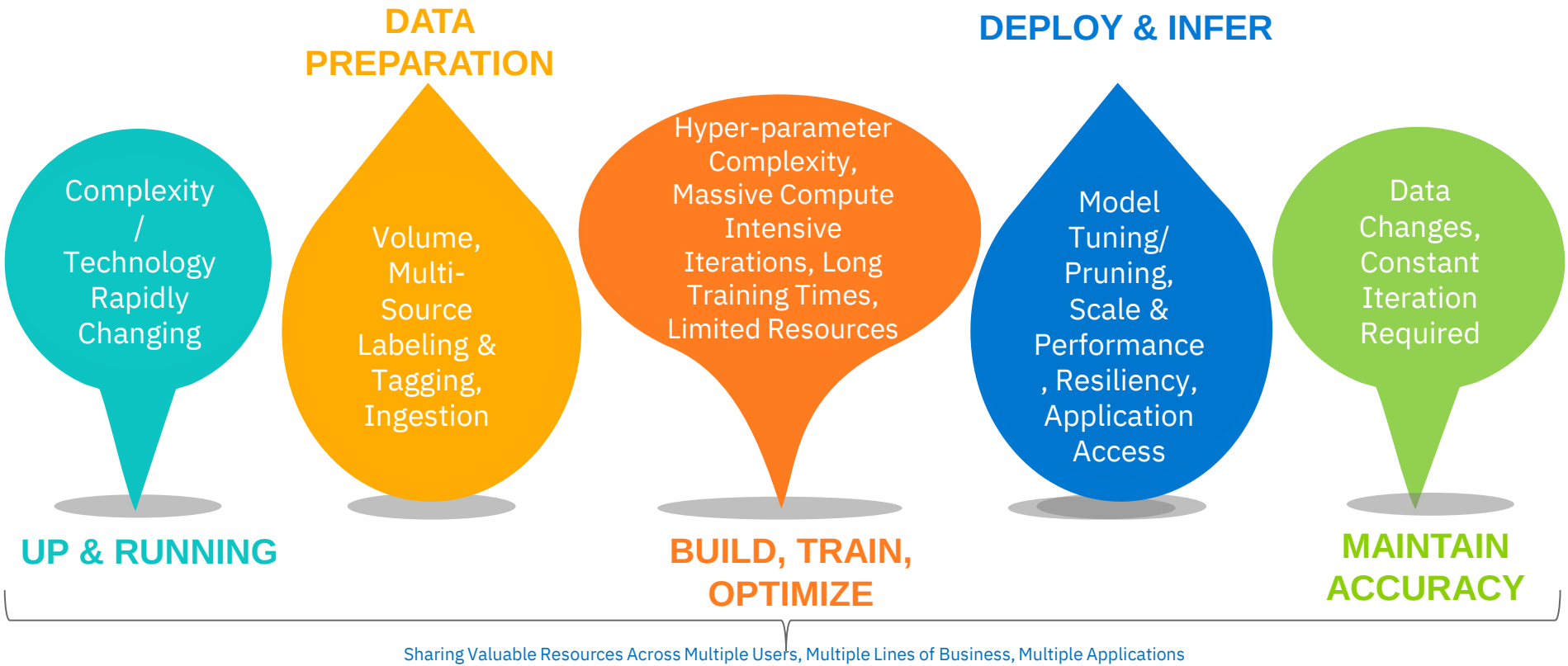
Instance segmentation

- Instance Category « car » 1 : pixel field 1 90%
- Instance Category « car » 1 : pixel field 1 90%
- …

- Ex : Mask-RCNN

Pain Points - Deep Learning Pipeline

Made with PowerAI Vision – Sport / Advertising



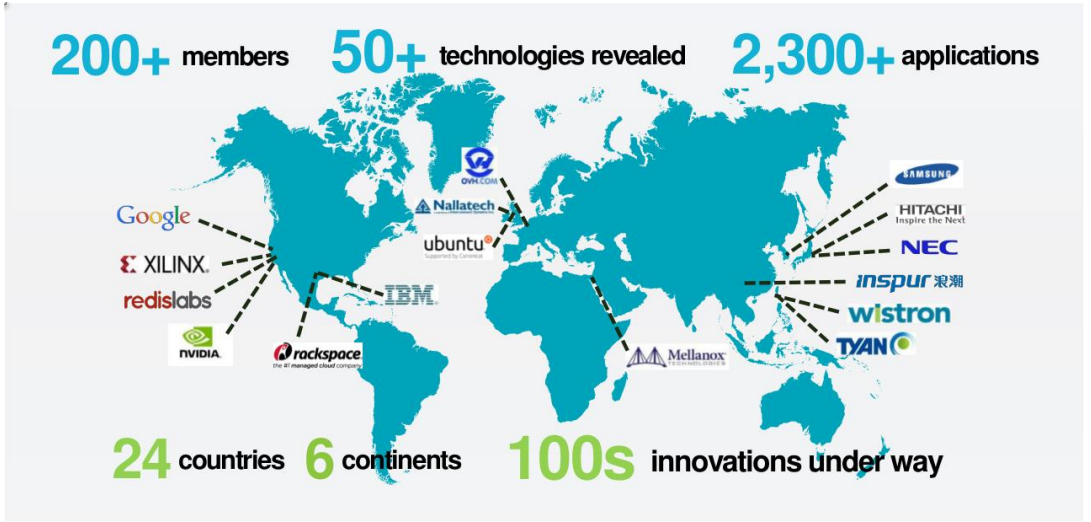
IBM & AI
Introducing AI On Premises & Private AI
IBM PowerAI

IBM PowerAI

IBM

OpenPower Foundation – After 6 years of existence…

- Drivers: Innovation vs. Moore law
- Collaborative Approach
- Domains:
  - Cloud & Scale Out Architecture
  - Research/ High Perf Computing
  - Analytics, Big Data
  - Machine Learning / Deep Learning
- IBM Sells OpenPower servers
  - Power Systems POWER9…



https://www.ibm.com/cloud/bare-metal-servers/power
https://console.bluemix.net/docs/services/PowerAI-IBM/

OpenPower Summit 2018

330+ Members

33 Countries

70+ ISVs

Active Membership **From All Layers of the Stack**

100k+ Linux Applications Running on Power

2300 ISVs Written Code on Linux

Partners Bring Systems to Market

169 OpenPOWER Ready Certified Products (Nearly **tripled** since inauguration at 2016 Summit)

20+ Systems Manufacturers

50+ POWER-based systems shipping or in development

100+ Collaborative innovations under way

Source Forbes

Google announced it has deployed POWER-based systems into its data center

IBM POWER SYSTEMS

AC922

An Acceleration Superhighway
Unleash state of the art IO and 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
Flatten the time to AI value curve by accelerating the journey to build, train, and infer deep neural networks

Seamless CPU and Accelerator Interaction

coherent memory sharing

enhanced virtual address translation

POWER9

Broader Application of Heterogeneous Compute

designed for efficient programming models

accelerate complex AI & analytic apps

“vanilla”

Others

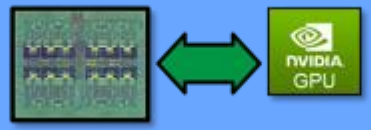
PCIe Gen3

extreme CPU and Accelerator bandwidth

2x

PCIe Gen4

5x

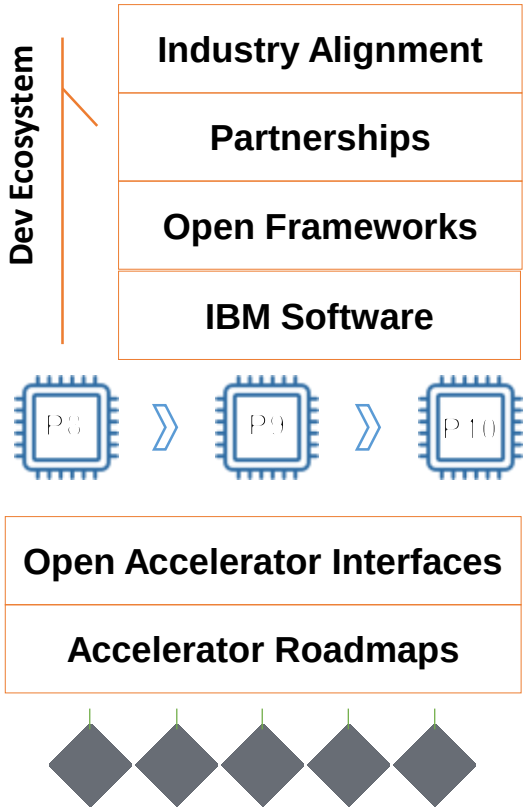


POWER8
with NVLink 1.0

7-10x

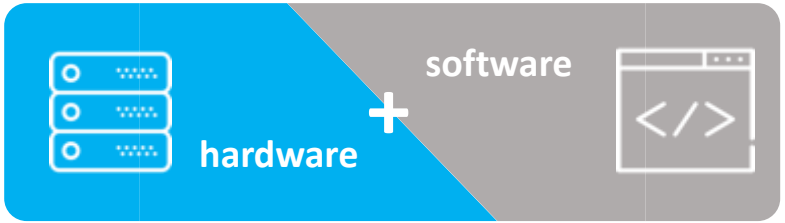
POWER9 with 25G Link
+ NVLink 2.0

Cognitive Systems are built with optimized HW & SW



Not Just About Hardware Design

It's about co-optimized



which ***just work*** for Machine Learning, Deep Learning and AI

POWER9 is the **only** processor with NVLink 2.0 from CPU to GPU

Delivering **5.6X Host-Device bandwidth vs Xeon E5-2640 v4** based systems with CUDA H2D Bandwidth Test

No code changes are required to leverage NVLink capability

"We're excited to see accelerating progress as the **Oak Ridge National Laboratory Summit supercomputer** machine, which we expect will be among the world's fastest supercomputers. The advanced capabilities of the IBM POWER9 CPUs coupled with the NVIDIA Volta GPUs will significantly advance DOE's mission critical applications," says Buddy Bland, Oak Ridge Leadership Computing Facility Director

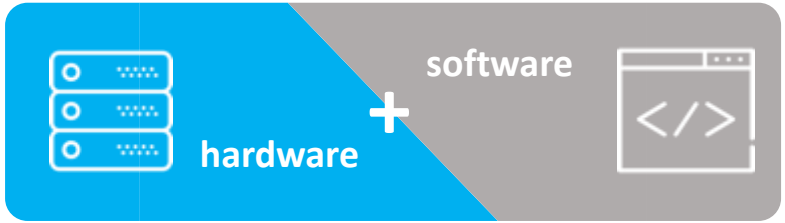
(Free) PowerAI 1.6 Overview (WML-CE)

Linux Native install or docker image on RHEL 7.6 / Ubuntu 18.04
https://hub.docker.com/r/ibmcom/powerai/
Works with CUDA 10 + Nvidia drivers

- Distributed Deep Learning (DDL) 1.3.0
- TensorFlow 1.13.1
- IBM Caffe 1.0.0
- Caffe2 1.0.1
- PyTorch 1.0.1
- Snap ML 1.2.0
- Rapids cuDF /cuML 0.2.0

Not Just About Hardware Design

It's about co-optimized



Available on x86-64 architecture, Optimized for

- IBM AC922 POWER9 system with NVIDIA Tesla V100 GPUs
- IBM S822LC POWER8 system with NVIDIA Tesla P100 GPUs

PowerAI Vision

Auto-ML for Images & Video

Label | Train | Deploy

Evolving with IBM One AI Strategy

PowerAI (WML-CE)

PowerAI: Open Source Frameworks

TensorFlow™ PYTORCH Chainer SnapML

Large Model Support (LMS)

PowerAI Enterprise (WML-A)

Distributed Deep Learning (DDL)

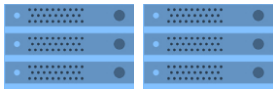
Auto ML (future)

IBM Spectrum Conductor
Cluster Virtualization, Elastic Training
Auto Hyper-Parameter Optimization

Deep Learning Impact (DLI) Module

Data & Model Management, ETL, Visualize, Advise

Accelerated Infrastructure



Accelerated Servers AC922

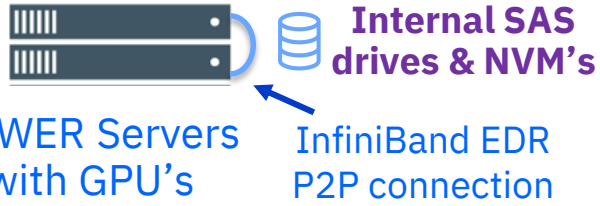
Storage (Spectrum Scale ESS)

PowerAI Top 4 Features

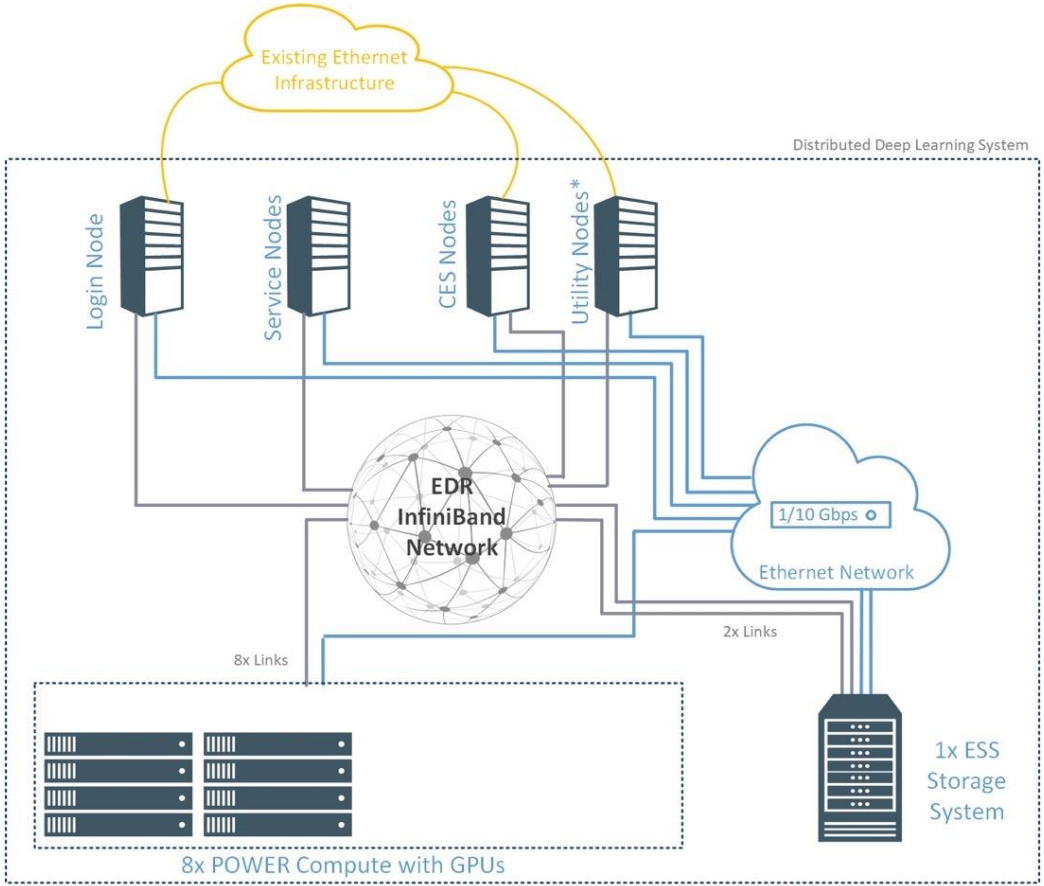
Snap ML & H2O	2x to 40x Faster Machine Learning with Snap ML. H2O Driverless AI automates ML. Most enterprise clients use ML today
Large Model Support (LMS)	Our TensorFlow can handle larger models & datasets; leads to higher accuracy
PowerAI Vision	Enables clients without data scientists to start with AI. Bundle with lab services or service provider or workshop
WML Accelerator	Higher Server & GPU Utilization. Target x86 server install base. We can schedule & manage GPU resources better than other software

Hardware Overview

Starter Deep Learning System

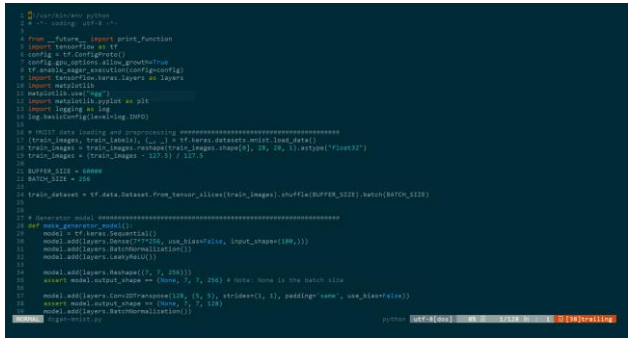


Mid-Size Deep Learning System



Reference Architecture for AI Infrastructure

Experimentation Single Tenant

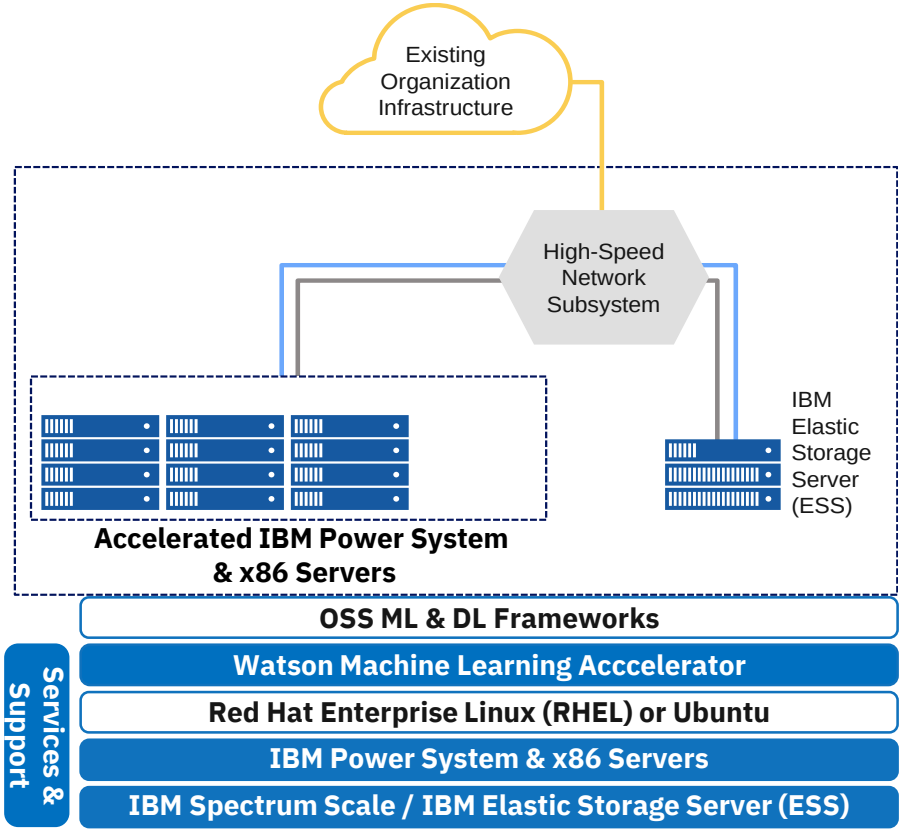


PowerAI

Cuda Drivers – OSS Frameworks

Red Hat Enterprise Linux (RHEL) or Ubuntu

Production – Multi Tenant & Scalability

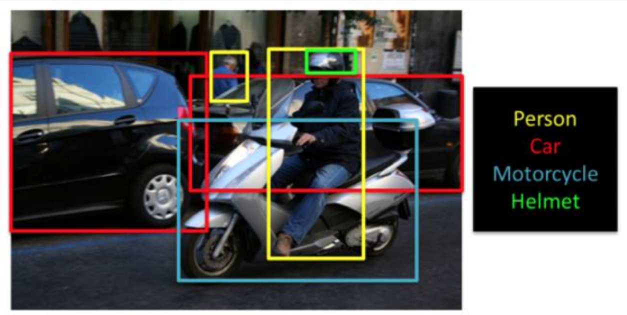


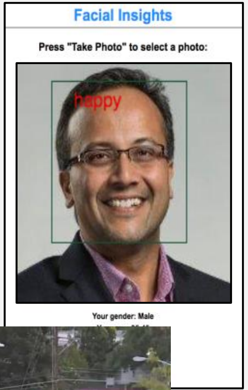
Comparing AI Offerings on Power

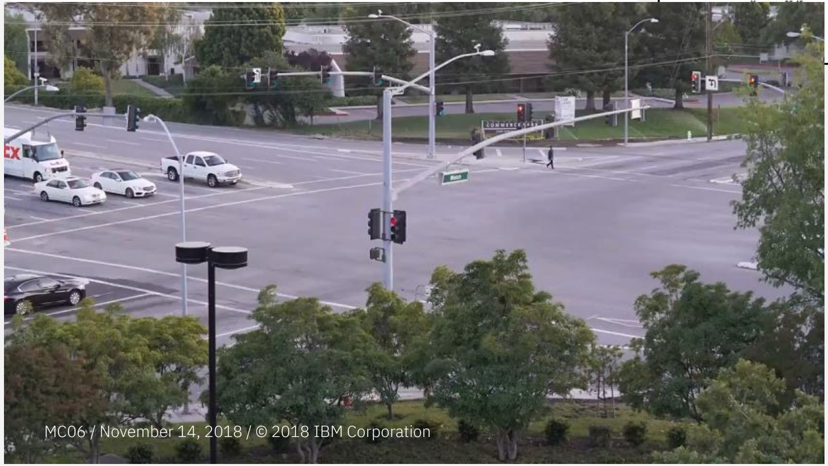
		Deep Learning			ML and DL	Machine Learning
		Power AI Base (WML-CE)	Power AI Enterprise (WML-A)	AI Vision	Watson Studio Local	H2O Driverless AI
Offering	Description	Deep Learning	Deep Learning for the Enterprise	Deep Learning with Video tools	Notebook oriented development environment for ML and DL	Automated Machine learning
	Pricing Model	Free download	Commercial	Commercial	Commercial	Commercial
	Support	Available from IBM	IBM L 1-3 Included	IBM L1-3 Included	Available from IBM	H2O L 1-3
Applications	Text & Numeric	Yes	Yes	No	Yes	Yes
	Images	Yes	Yes	Yes	Yes	No
	Video	-	Optional add-on	Yes		No
	Primary Persona	Data Scientist	Data Scientist	Line of Business	Data Scientist	Data Scientist
	Second persona	IT	IT	IT	IT	Line of Business
	User Skill Level	High	Medium to high	Low	Medium to high	Low to Medium
	Strengths	Rapid deployment, high performance, scale	enterprise grade, High performance, rapid Deployment	Rapid deployment, simple GUI high performance	Notebook based development environment, strong collaboration, model management	Simplified deployment, intuitive user interface, automatic pipelines, "explainability" for models, end to end automation
Platform	Distributed DL (DDL)	1-4 nodes	1-thousands of nodes	Coming	Coming	-
	Large Model Support	Yes	Yes	Coming	Coming	-
	Server(s)	S822LC or AC922	S822LC or AC922	S822LC or AC922	S822LC or AC922, LC922	S822LC, AC922, LC921/922
IBM Products	Spectrum MPI (DDL)	Limited to 4 nodes	Included			Optional add-on
	Spectrum Conductor DLI	Optional add-on	Included	Coming	Optional Add On	Optional add-on
	IBM Watson Studio Local	Optional add-on	Optional add-on	No		Optional add-on
Cloud	IBM Cloud Public	Yes	No	Trial only	Watson Studio	?
	IBM Cloud Private	Yes	Yes	Yes	Yes	Yes

H2O Driverless AI Complements IBM PowerAI Vision

IBM Power AI delivers Deep Learning for Images







H2O Driverless AI is an Automatic Machine Learning

Transactional Data: Store Level

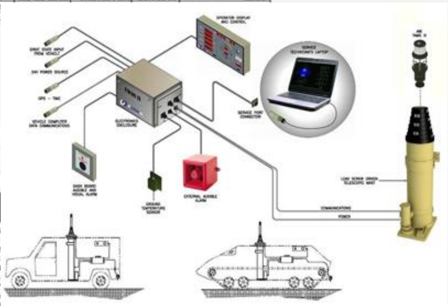
Transactional

Product_Code	Amount
[illegible]	2.32
[illegible]	1.43
[illegible]	1.89
[illegible]	2.2
[illegible]	1.14
[illegible]	3.89
[illegible]	1.6
TOTAL	14.47

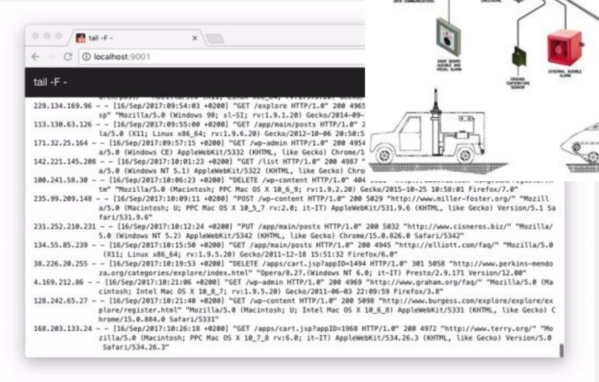
Transaction ID	Date	Time	AMOUNT	Card_number
p3913045367	11/24/2010	3:20:32 AM	14.47525	************2323
q6781991970	12/4/2010	10:30:10 AM	39.46618	************6451
r9178278202	11/2/2010	1:23:21 AM	24.99964	************2179
s7329082178	10/8/2010	12:29:40 AM	19.21324	************5826
t6835491952	10/17/2010	10:36:12 PM	39.56086	************9498
u8333871154	12/5/2010	5:22:59 PM	33.35401	************9379
v1240343519	10/27/2010	4:12:38 AM	49.64481	************5466
w5440123613	11/26/2010	10:36:36 PM	24.23247	************1816
x8906115216	11/3/2010	1:57:33 PM	32.45101	************2662
y1400944560	11/25/2010	7:26:45 PM	32.12293	************4330
z3665080889	10/9/2010	11:09:58 AM	21.93351	************6533
p6738256686	11/23/2010	10:14:36 AM	21.71996	************4615
q8964299443	11/24/2010	1:19:24 AM	15.46741	************7694
r9012945206	11/22/2010	2:02:15 PM	31.14201	************7140
s2116305133	11/23/2010	7:15:13 PM	43.16047	************9208
t7478724264	12/2/2010	12:08:46 PM	40.14018	************2695
u9844400319	10/11/2010	3:10:19 AM	23.70907	************2980
v2388298974				
w567236458				

Example: Flat File

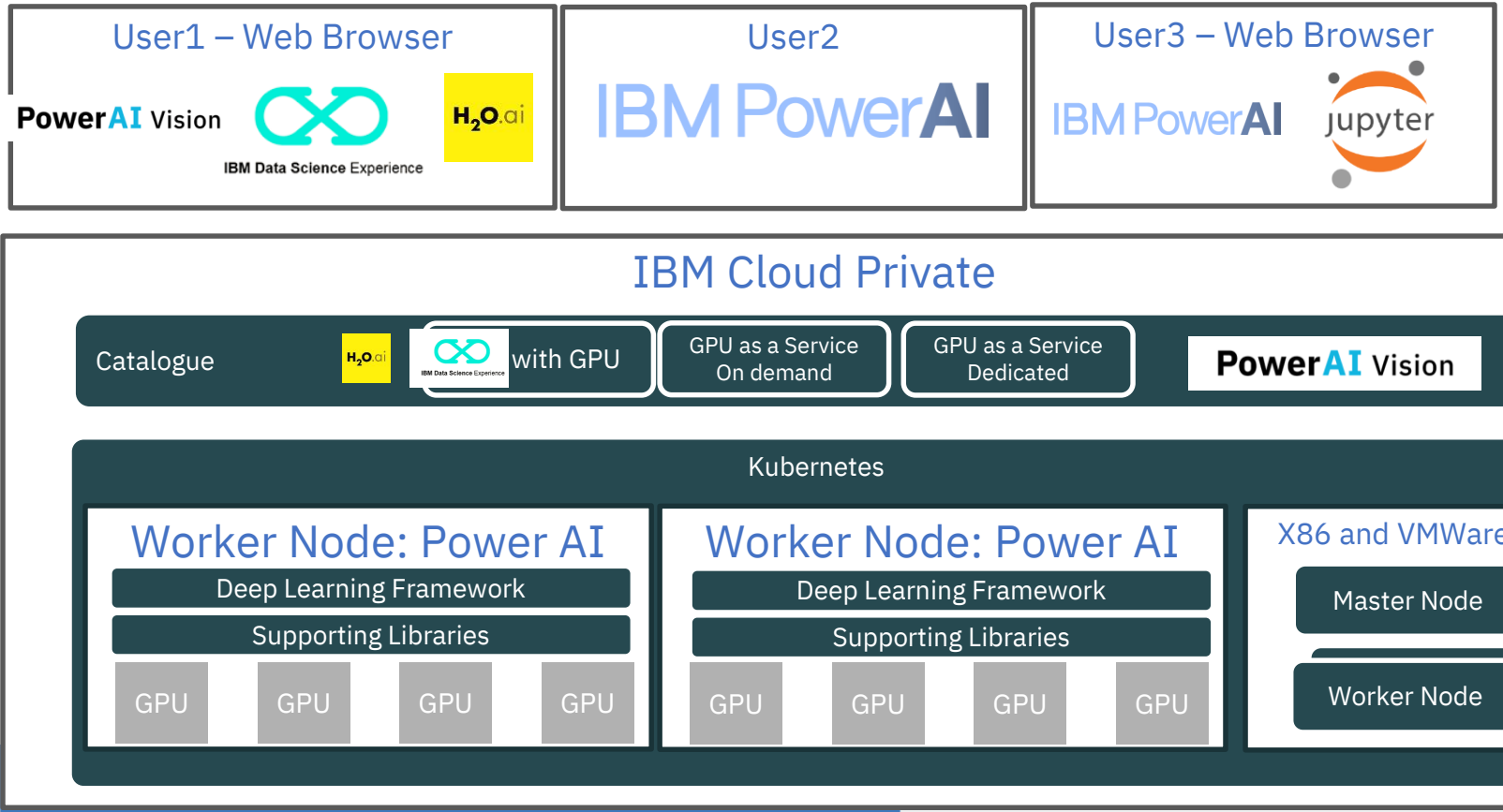
Sensors



Log

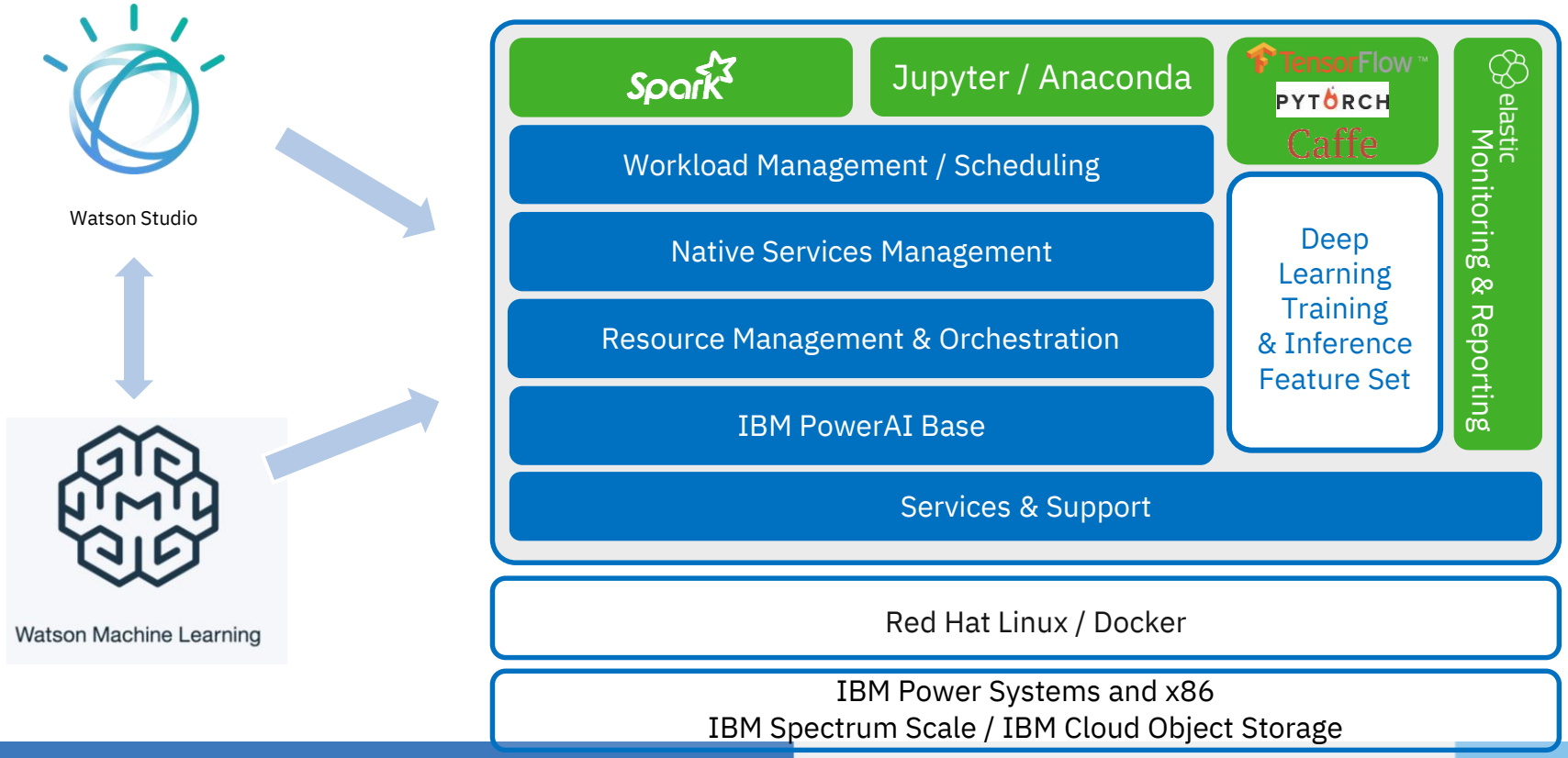


Kubernetes/IBM Private Cloud w/ PowerAI : Build your own AI Private Cloud

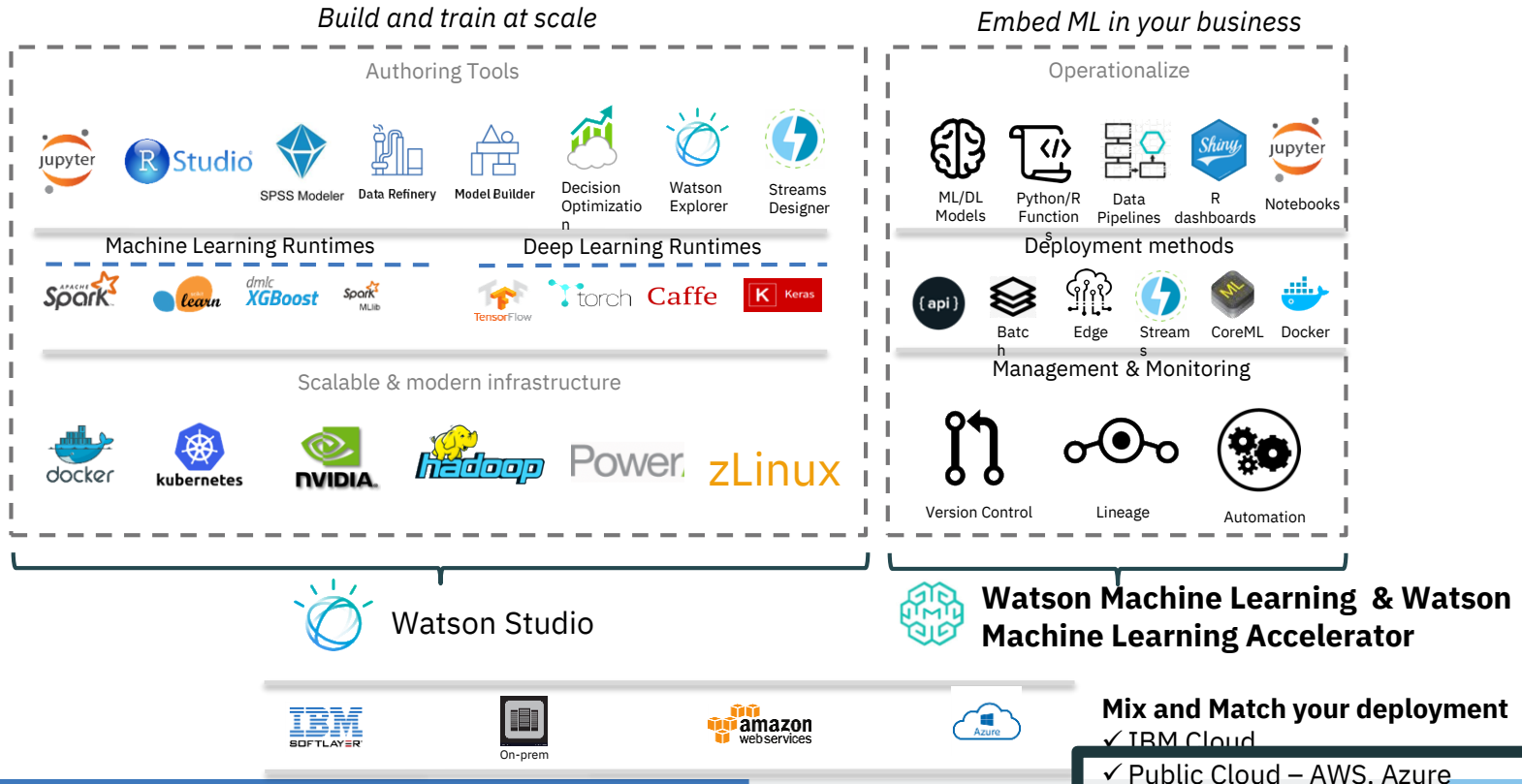


IBM Watson ML Accelerator:

Power AI Base + Spectrum Conductor + Deep Learning Impact



Injecting AI Firepower with IBM Watson Studio and IBM Watson Machine Learning

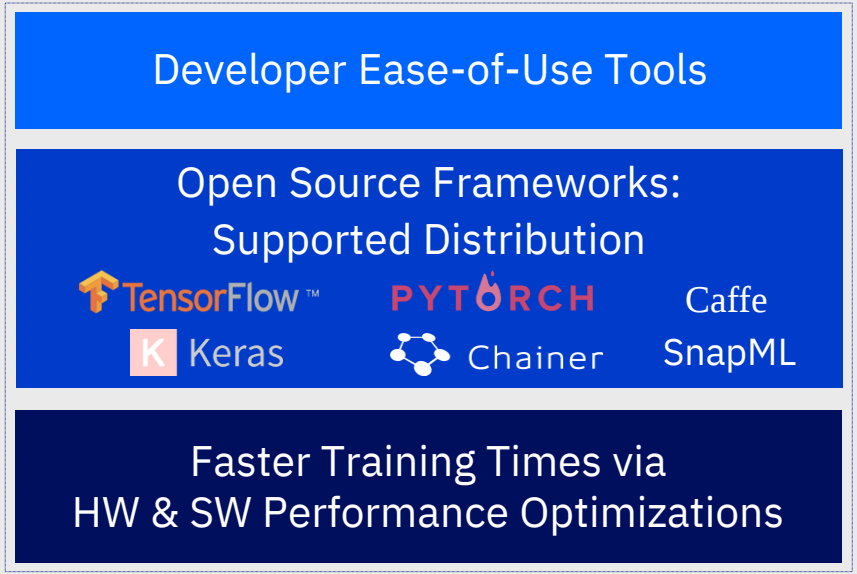


- Focus on PowerAI differentiators
  PowerAI Base WML-CE: LMS, Accelerated ML, Solutions
  PowerAI Enterprise WML-A : Scalability

PowerAI

Open-Source Based Enterprise AI Platform

Integrated & Supported AI Platform
3-4x Speedup for AI Training
Ease of Use Tools for Data Scientists



PowerAI: Enterprise AI Platform

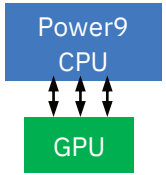
Simplicity: Integrated Platform that Just Works

Curate, Test, and Support Fast Moving Open Source

Provide Enterprise Distribution on RedHat

Easy to deploy Enterprise AI Platform

Ease of Use, Unique Capabilities



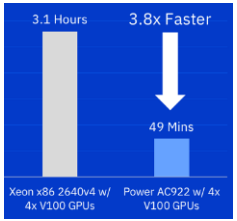
Large data & model support due to NVLink

Acceleration of Analytics & ML

AutoML: PowerAI Vision

Elastic Training: Scale GPUs as Required

Faster Model Training Time

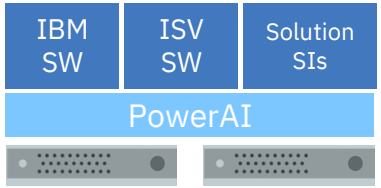


Faster Training Times in Single Server

Scalability to 100s of Servers (Cluster level Integration)

Leads to Faster Insights and Better Economics

Open AI Platform w/ Ecosystem Partners



Platform that Partners can build on

Software Partners: H2O, IBM, Anaconda

SIs, Solution Vendors & Accelerator Partners

Deep learning is memory constrained

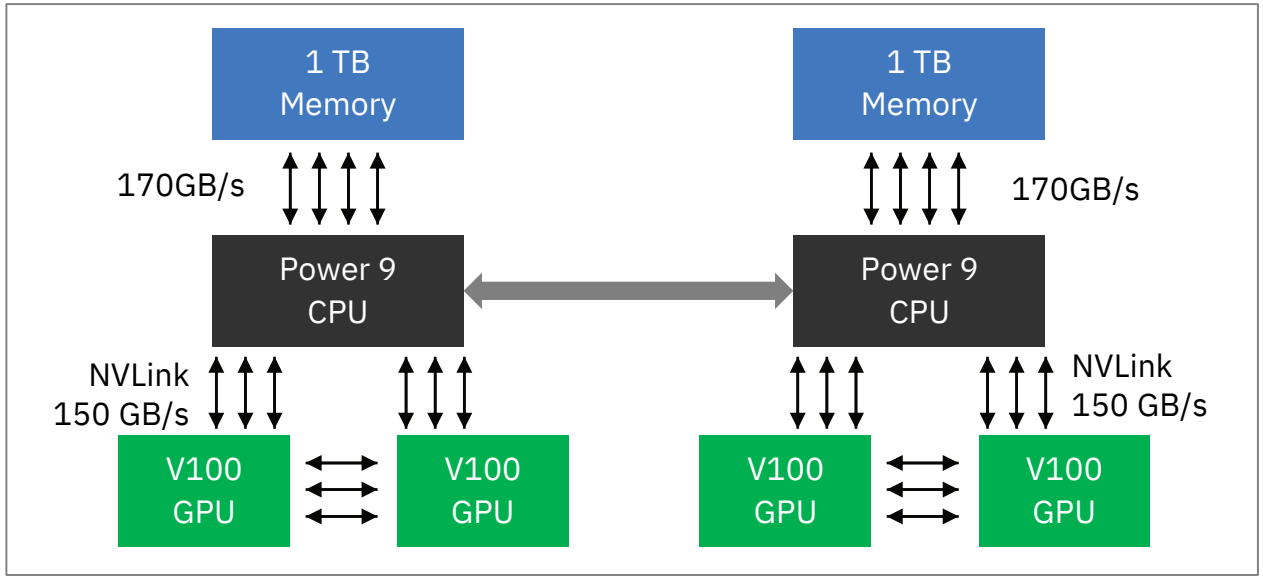
- GPUs have limited memory
- Neural networks are growing deeper and wider
- Amount and size of data to process is always growing

5x Faster Data Communication with Unique CPU-GPU NVLink High-Speed Connection

Store Large Models in System Memory

Fast Transfer via NVLink

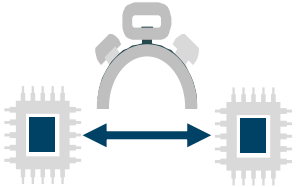
Operate on One Layer at a Time



IBM AC922 Power System

Deep Learning Server (4-GPU Config)

Acceleration in training …. days become hours



Performance…
Faster Training
and Inferencing

Large AI Models Train
~4 Times Faster
POWER9 Servers with NVLink to GPUs
vs
x86 Servers with PCIe to GPUs

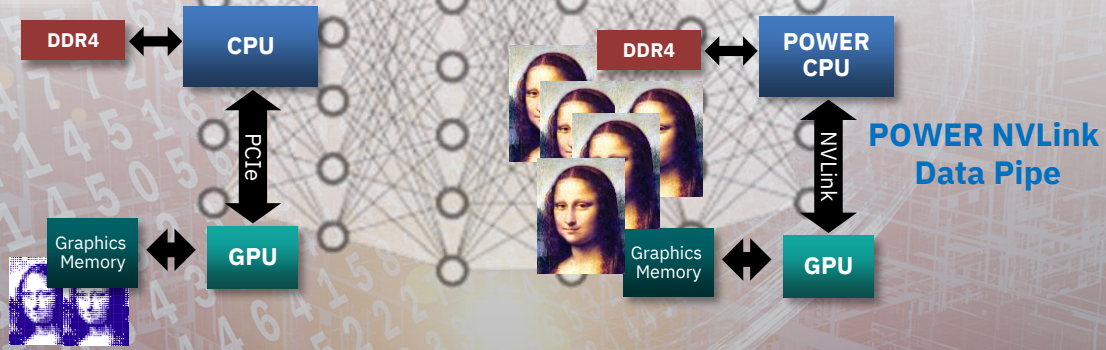
faster training times for data scientists

Distributed Deep Learning

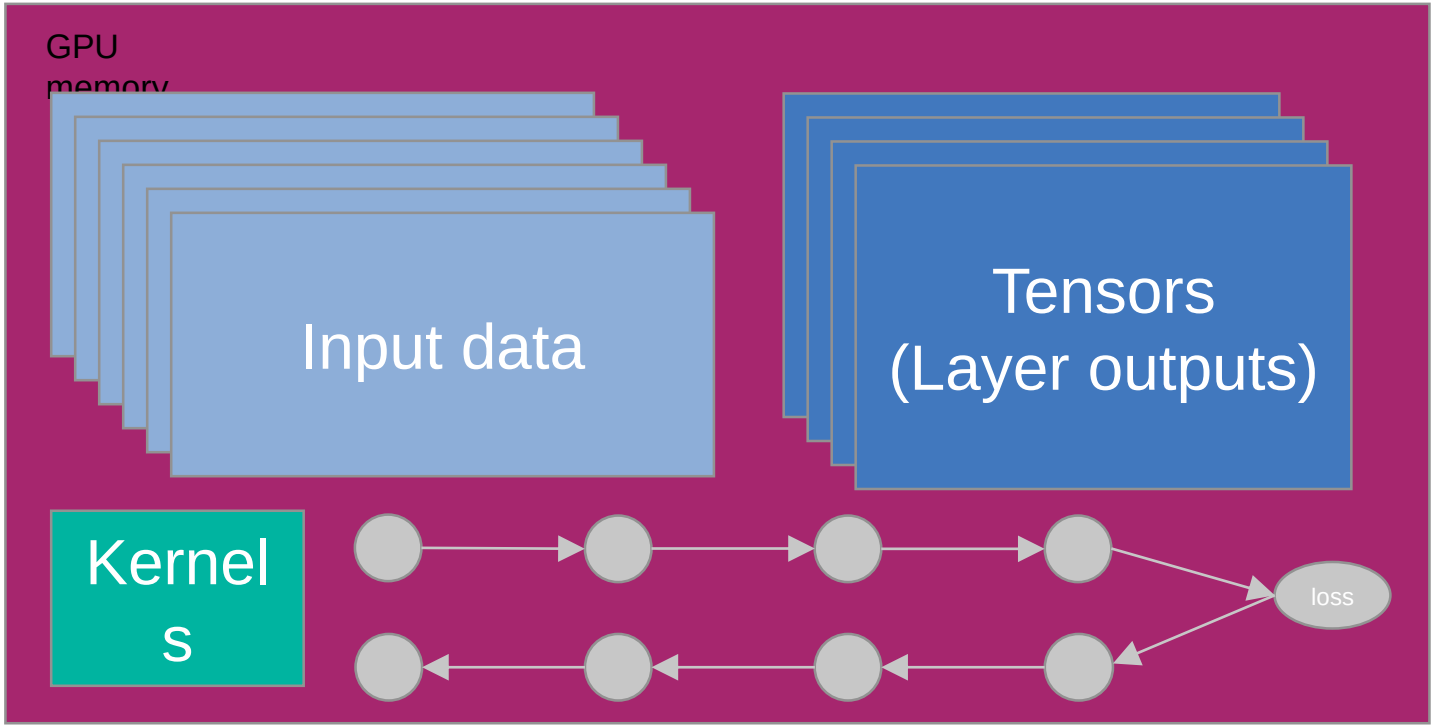
Traditional Model Support → ***Large Model Support***

(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



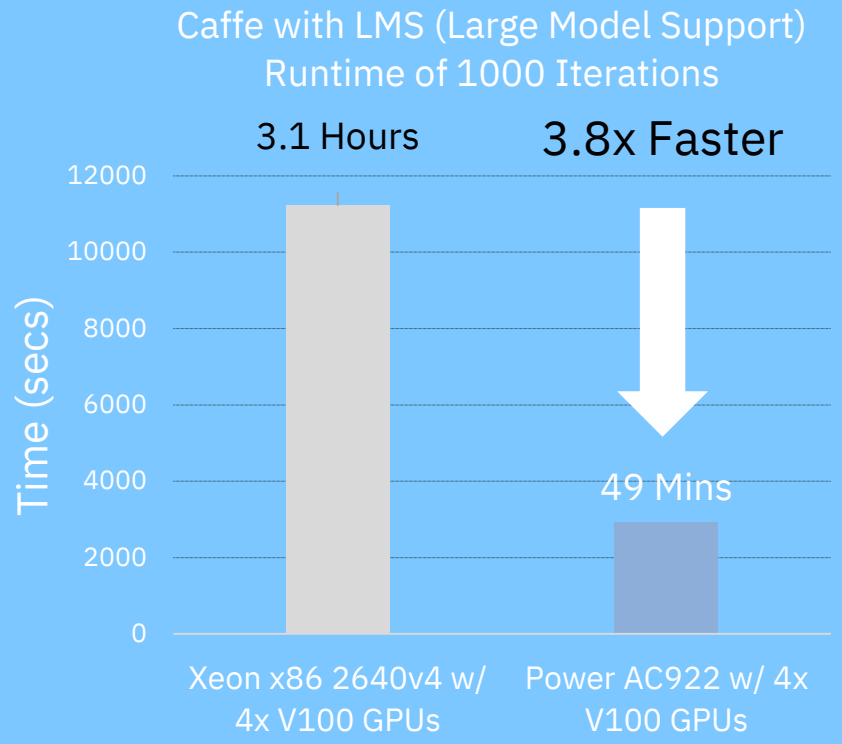
Gpu memory usage



Large AI Models Train ~4 Times Faster

POWER9 Servers with NVLink to GPUs
vs
x86 Servers with PCIe to GPUs

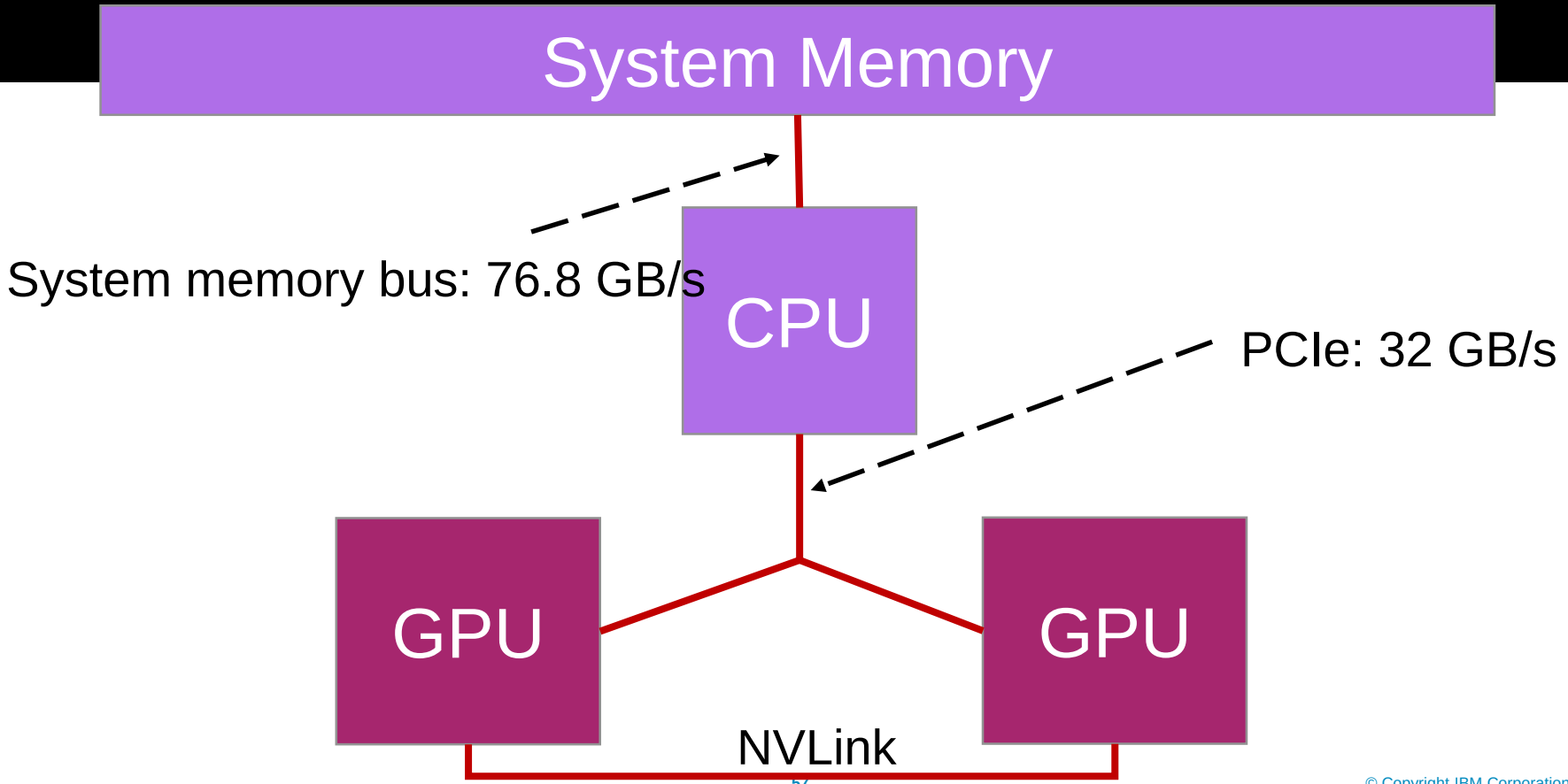
Detailed Benchmark Information in Back



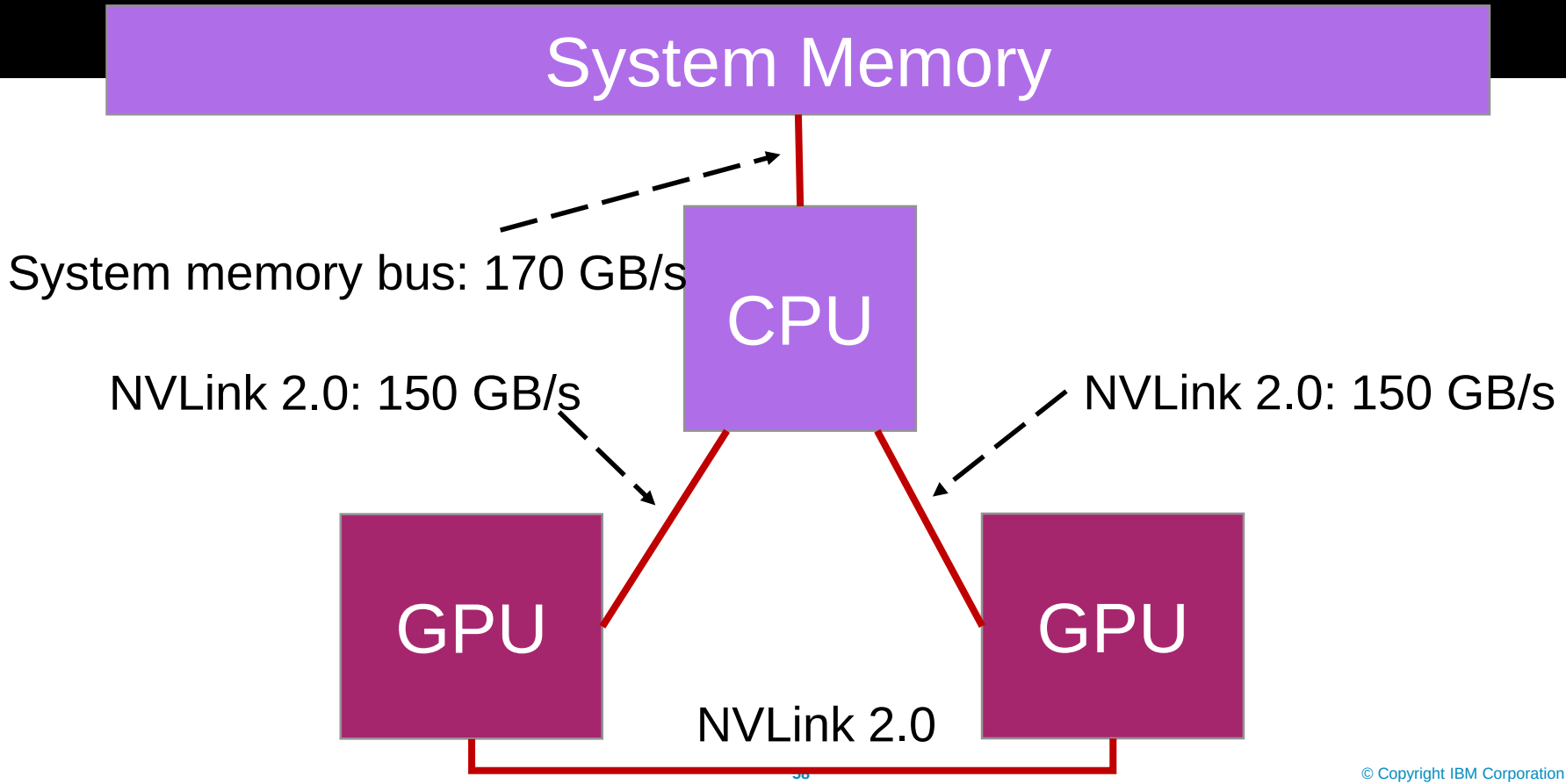
GoogleNet model on Enlarged ImageNet Dataset (2240x2240)

55

Typical GPU connectivity



Nvlink cpu to GPU connectivity



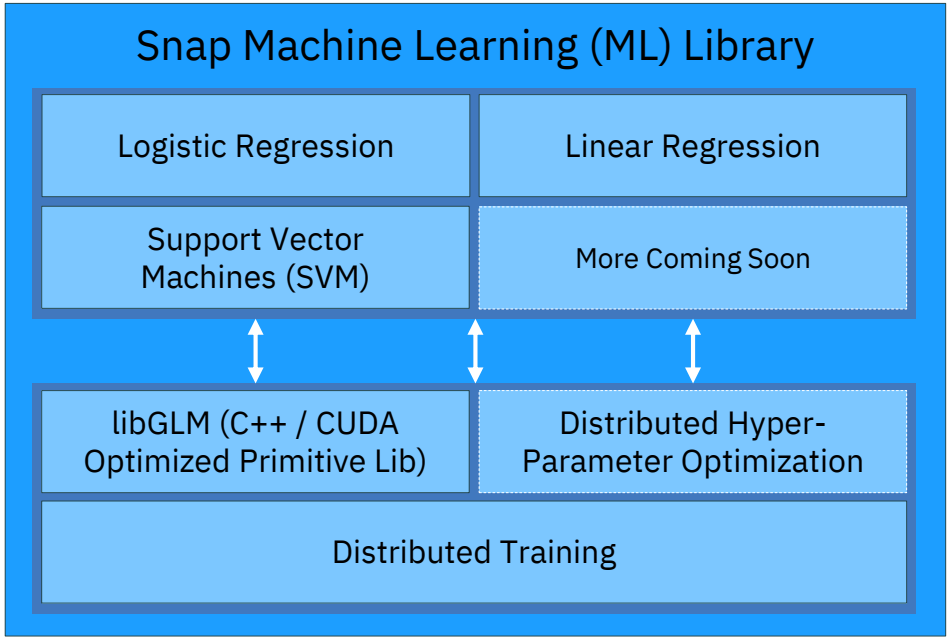
- Accelerated Machine Learning

IBM PowerAI

IBM

Snap ML

Distributed GPU-Accelerated Machine Learning Library



APIs for Popular ML Frameworks

python

scikit learn

(coming soon)

APACHE Spark

4 APIs

pai4sk API (Included in PAI 1.5.4)
snap-ml-local API
snap-ml-mpi API
snap-ml-spark API

- Snap ML: Training Time Goes From An Hour to Minutes

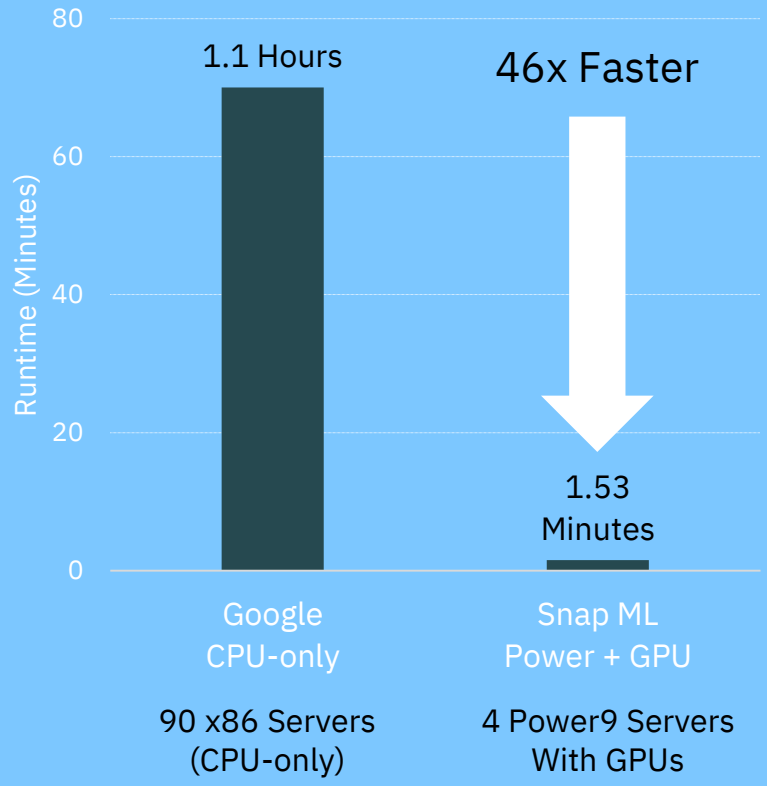
46x faster than previous record set by Google

Workload: Click-through rate prediction for advertising

Logistic Regression Classifier in Snap ML using GPUs vs TensorFlow using CPU-only

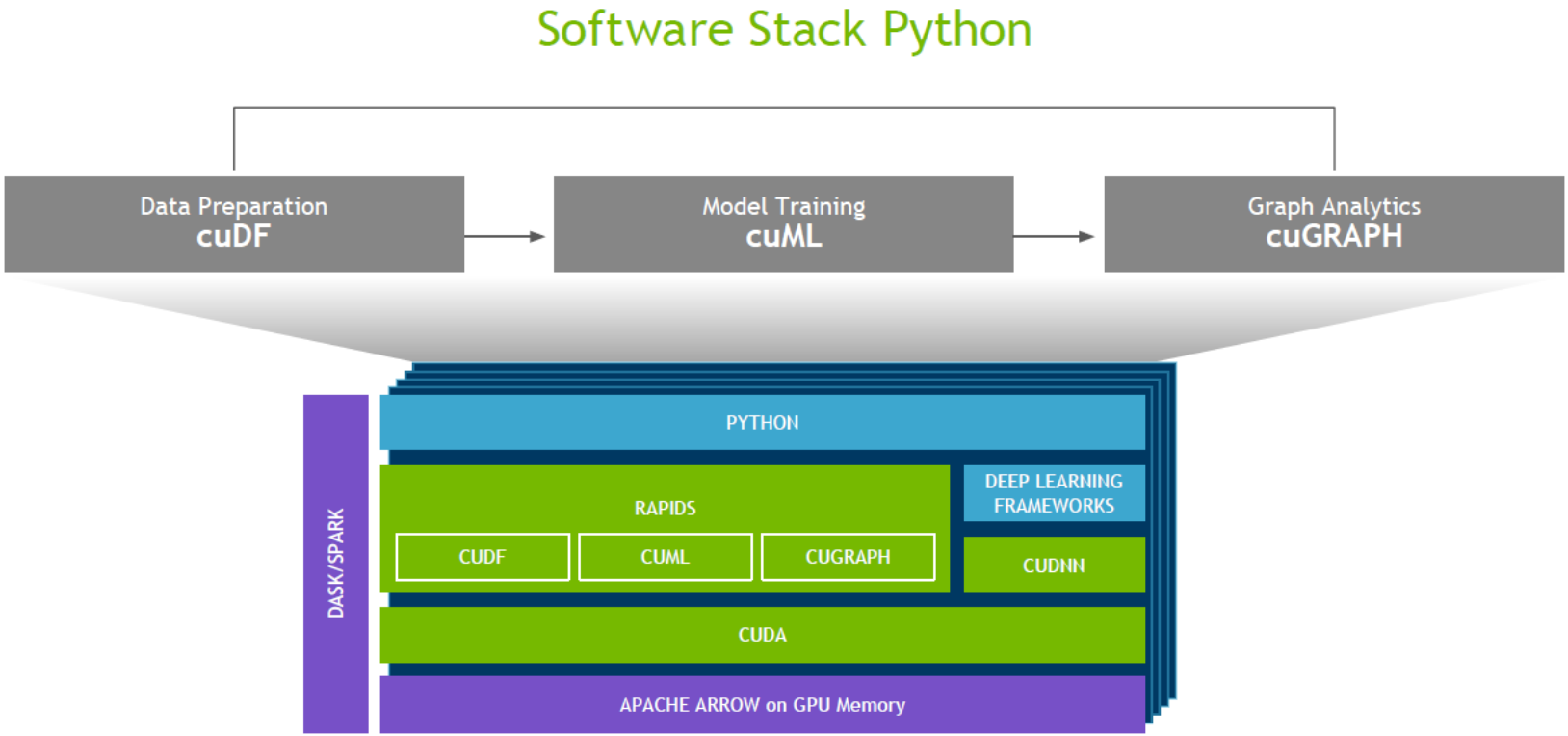
Dataset: Criteo Terabyte Click Logs (http://labs.criteo.com/2013/12/download-terabyte-click-logs/)
4 billion training examples, 1 million features
Model: Logistic Regression: TensorFlow vs Snap ML
Test LogLoss: 0.1293 (Google using Tensorflow), 0.1292 (Snap ML)
Platform: 89 CPU-only machines in Google using Tensorflow versus 4 AC922 servers (each 2 Power9 CPUs + 4 V100 GPUs) for Snap ML
Google data from this Google blog

Logistic Regression in Snap ML (with GPUs) vs TensorFlow (CPU-only)



Nvidia RAPIDS – Open GPU Data Science

PowerAI 1.6 support



- Solutions based on PowerAI

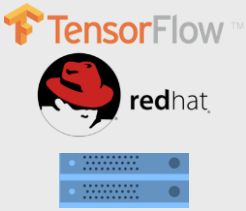
  Accelerate ++ & increase productivity

IBM PowerAI

IBM

PowerAI: Enterprise AI Platform

Simplicity: Integrated Platform that Just Works

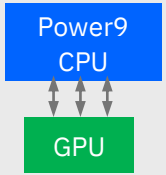


Curate, Test, and Support Fast Moving Open Source

Provide Enterprise Distribution on RedHat

Easy to deploy Enterprise AI Platform

Ease of Use, Unique Capabilities



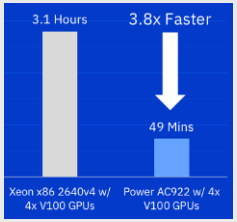
Large data & model support due to NVLink

Acceleration of Analytics & ML

AutoML: PowerAI Vision

Elastic Training: Scale GPUs as Required

Faster Model Training Time

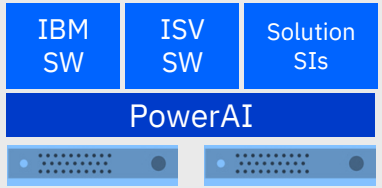


Faster Training Times in Single Server

Scalability to 100s of Servers (Cluster level Integration)

Leads to Faster Insights and Better Economics

Open AI Platform w/ Ecosystem Partners



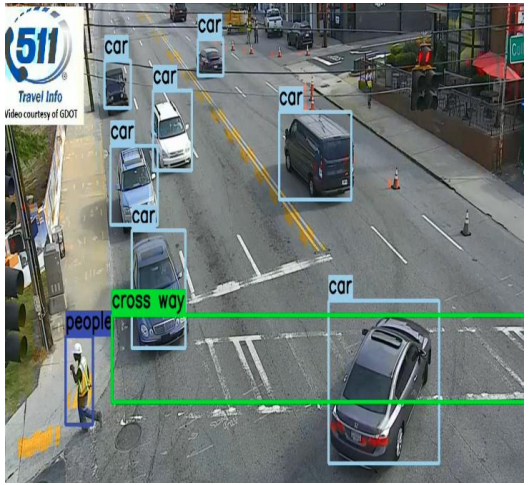
Platform that Partners can build on

Software Partners: H2O, IBM, Anaconda

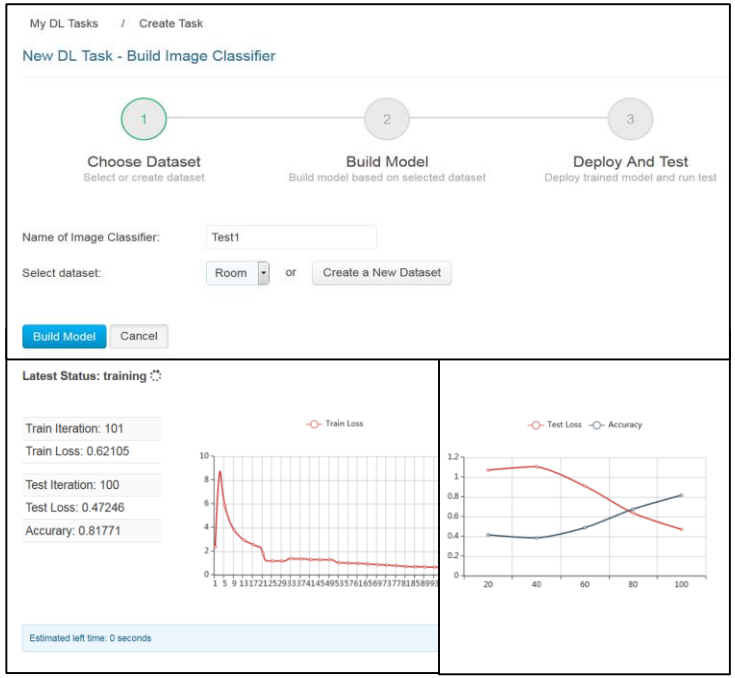
SIs, Solution Vendors & Accelerator Partners

PowerAI Vision: ”Point-and-Click” AI for Images & Video

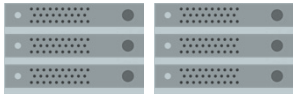
Label Image or Video Data



Auto-Train AI Model

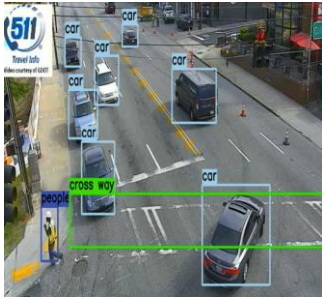


Package & Deploy AI Model



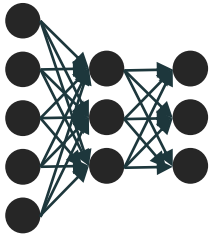
Semi-Automatic Labeling using PowerAI Vision

Manually Label



Define Labels Manually Label Some Images / Video Frames

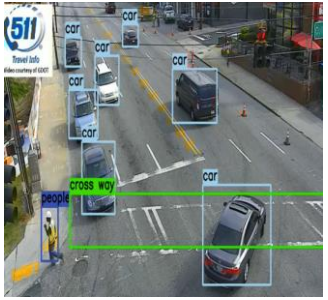
Train DL Model



Use Trained DL Model

Run Trained DL Model on Entire Input Data to Generate Labels

Correct Labels on Some Data



Manually Correct Labels on Some Data

Repeat Till Labels Achieve Desired Accuracy

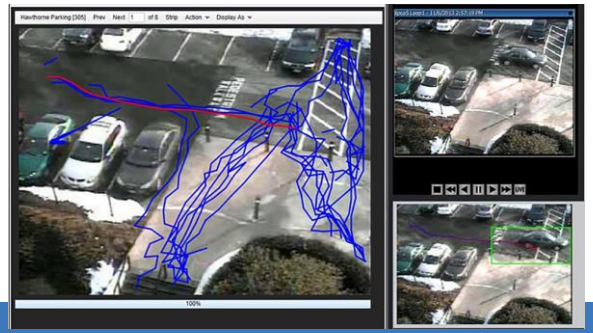
IBM Intelligent Video Analytics (IVA)

Video Analytics Software with Pre-Trained AI Models

Complex Event Monitoring with GUI-based Configuration

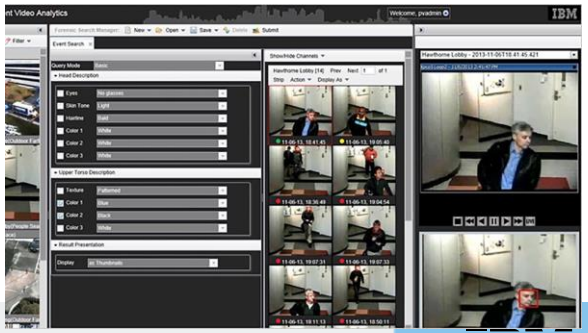
Targeted at Public Safety, Remote Monitoring, etc

Detect Changes to Patterns

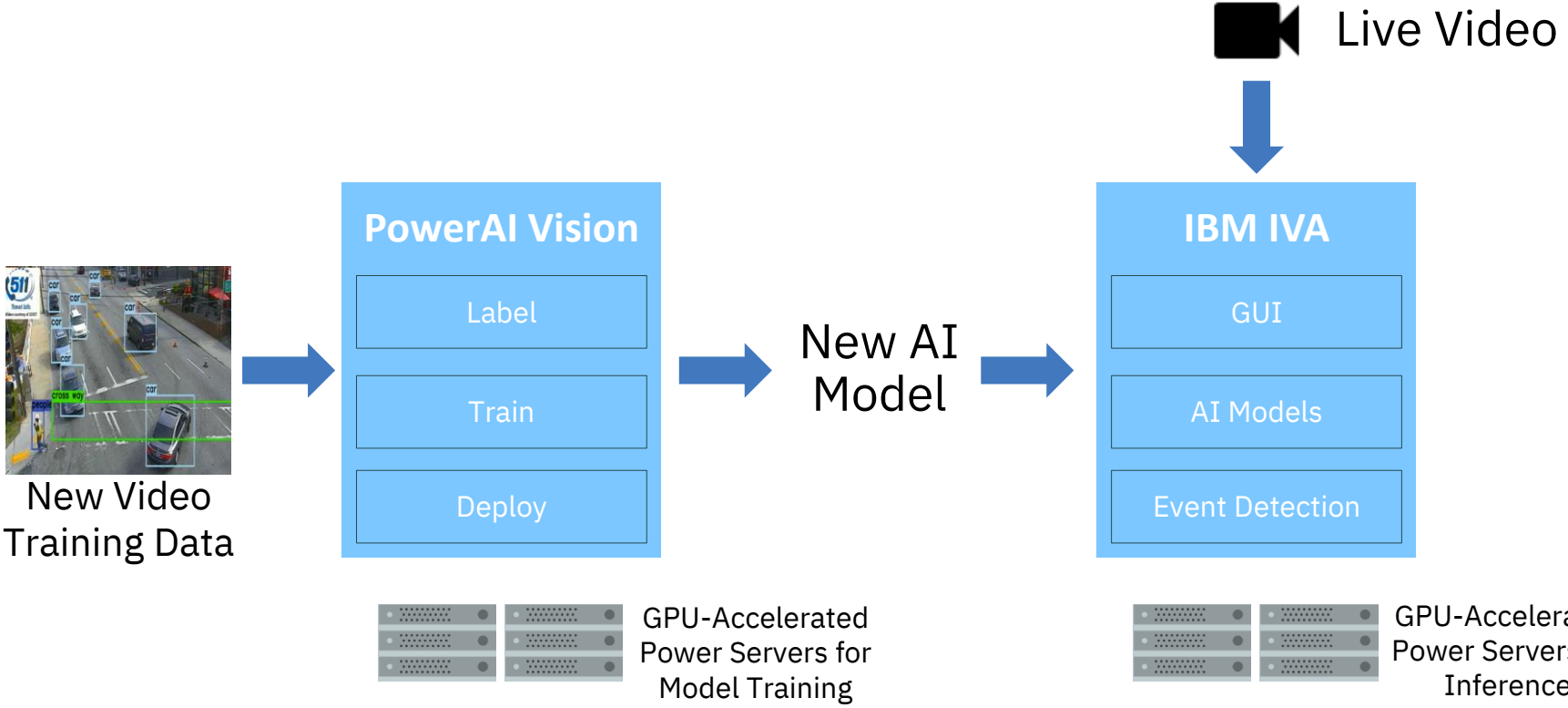


Redaction of Faces

Facial Recognition & People Search



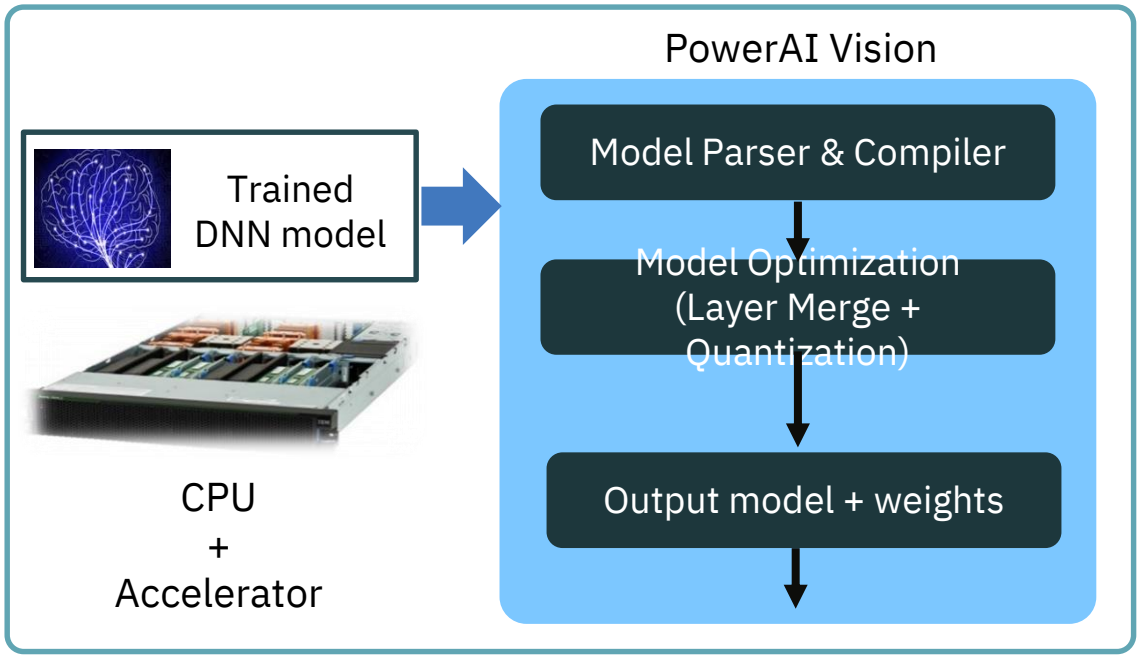
Announcing PowerAI Vision Integration with IBM IVA



68

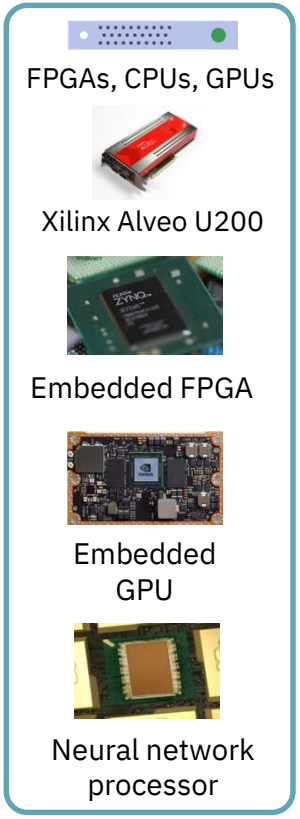
Deploying Trained Models

Data Center: Train model & Compile to Edge

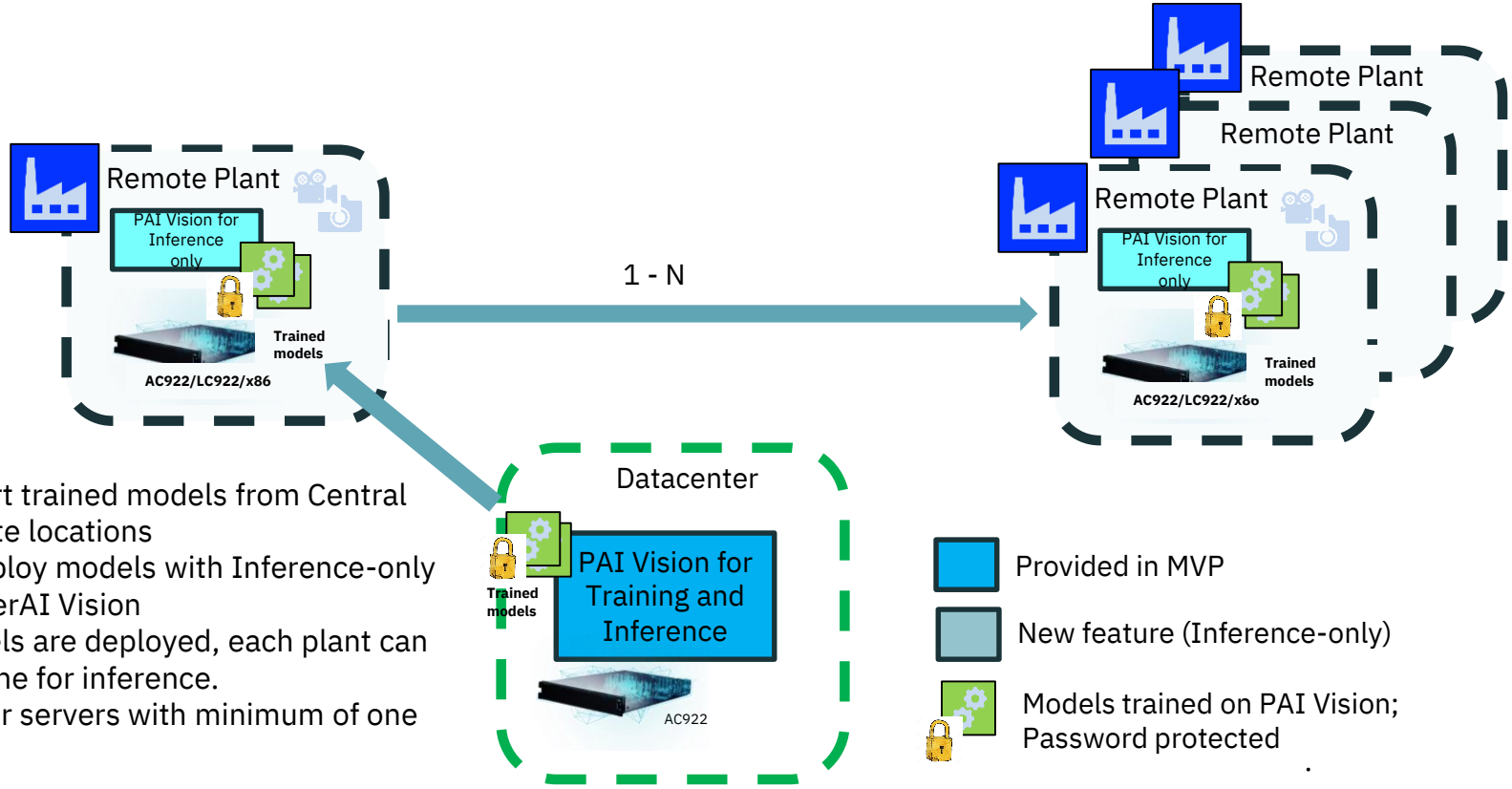


Map to Different Platforms

Cloud or Edge



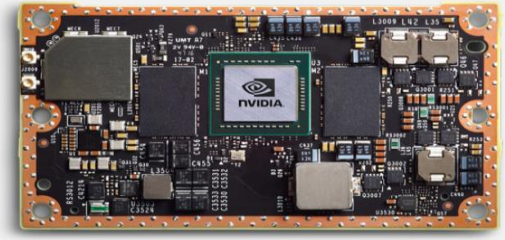
Train on central server but deploy on several remote servers



1. Manually export trained models from Central server to remote locations
2. Import and deploy models with Inference-only license of PowerAI Vision
3. Once the models are deployed, each plant can work stand alone for inference.
4. Supports Power servers with minimum of one GPU

Accelerated embedded edge devices

NVIDIA Jetson TX2 Module



Jetson TX2 is the fastest, most power-efficient embedded AI computing device. This 7.5-watt supercomputer on a module brings true AI computing at the edge. It's built around an NVIDIA Pascal™-family GPU and loaded with 8GB of memory and 59.7GB/s of memory bandwidth. It features a variety of standard hardware interfaces that make it easy to integrate it into a wide range of products and form factors.

Train on PowerAI Vision but infer on embedded devices

Retail Analytics

Track how customers navigate store, identify fraudulent actions, detect low inventory for items

IVA generates track summary, flags missing objects, alerts on suspicious behavior

Worker Safety Compliance

Utilize surveillance cameras to ensure worker safety compliance

IVA enables zone monitoring, heat maps, detection of loitering

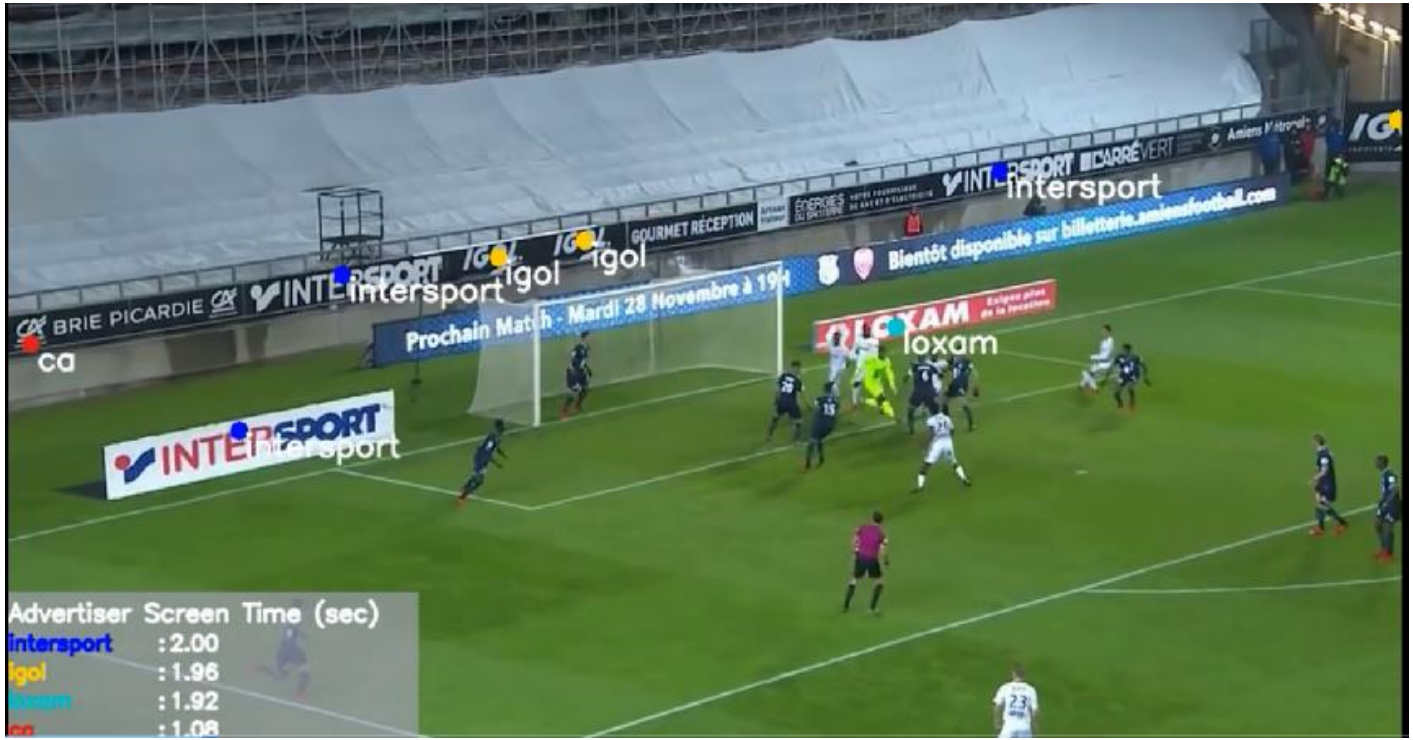
Remote Inspection & Asset Management

Identify faulty or worn out equipment in remote & hard to reach locations

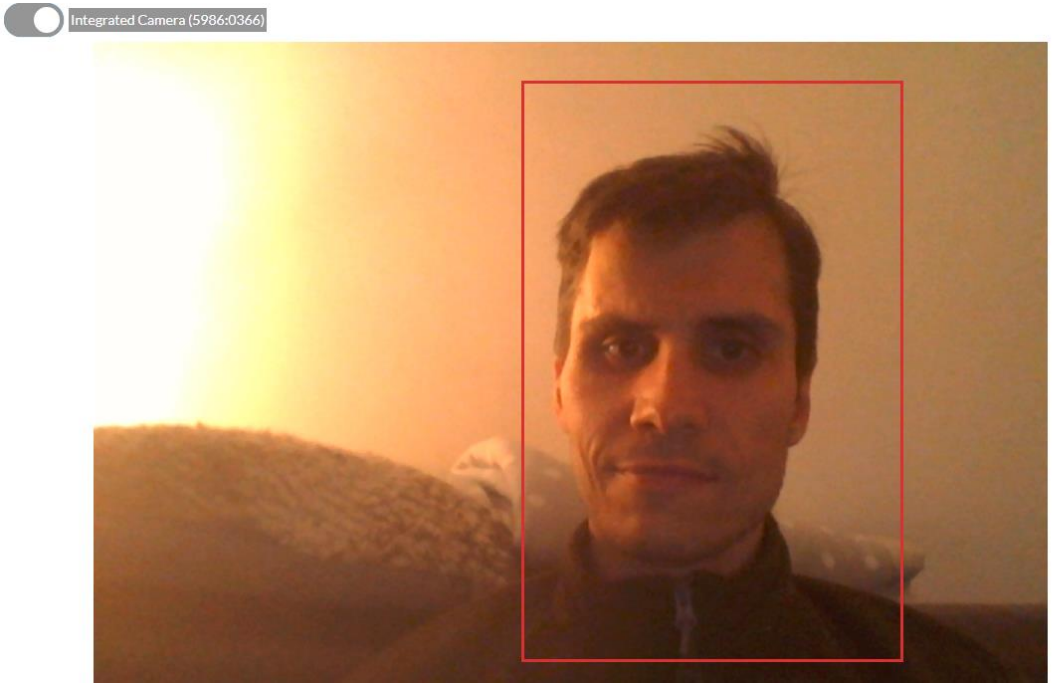
IVA can alert to schedule maintenance job, along with critical infrastructure security

Made with PowerAI Vision – Drone Inspection

Made with PowerAI Vision – Sport / Advertising



Made with PowerAI Vision – Real time detection on IBM i



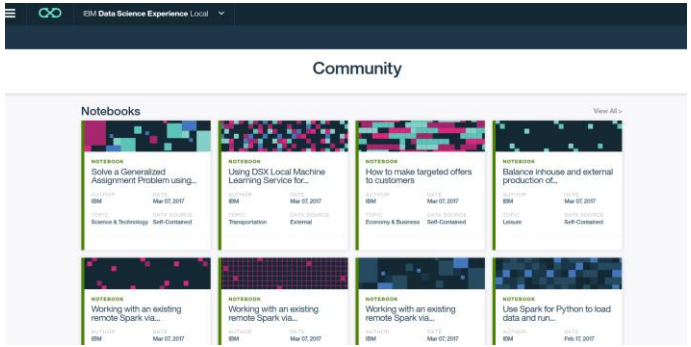
Visual Recognition Demo on IBM i + PowerAI Vision

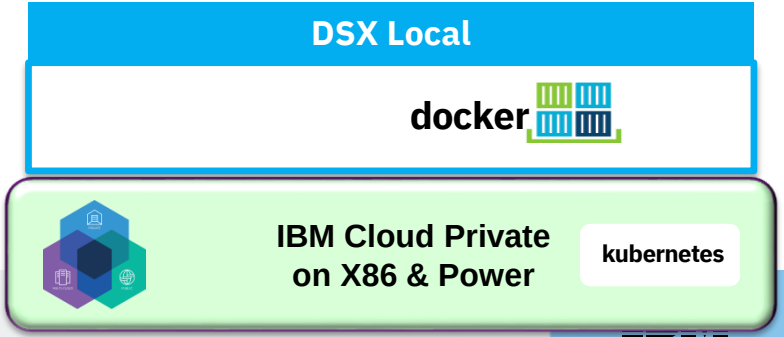
- IBM Watson Studio Local & ICP for Data

  Collaborative toolset for increasing productivity
  End to end AI project lifecycle
  Takes benefit of the PowerAI innovations &optimized frameworks

Core Attributes of Watson Studio

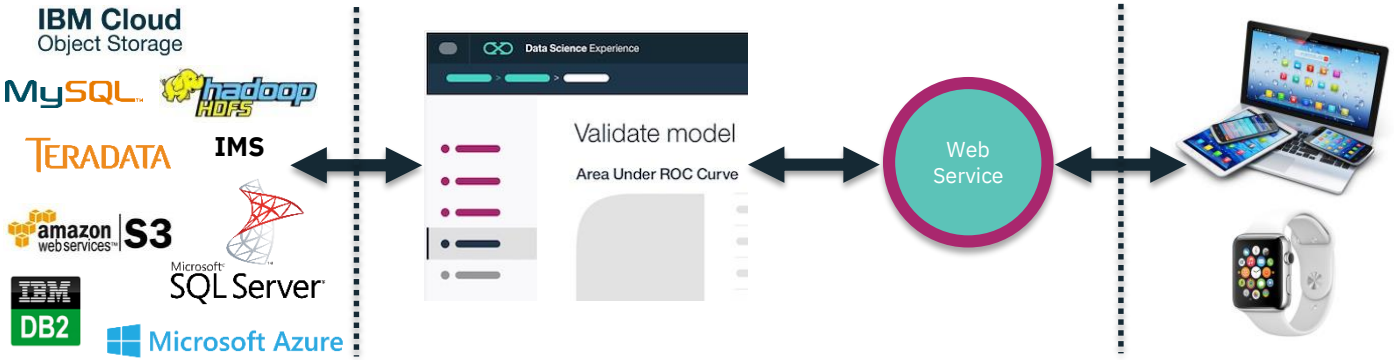
- IBM Watson Studio (aka DSX) is available
  - ❑ As a cloud offering aka **Watson Studio**
  - ❑ As a desktop application
    - Free, disconnected mode
  - ❑ As an on-premises solution
    - ❑ **DSX Local now Watson Studio Local** on x86/Power
    - ❑ Power: Scale-out LC Systems with PowerAI + GPU / Nvlink acceleration
    - ❑ Possible private cloud deployment with IBM Cloud Private





Build Predictive models with Watson Studio

IBM Machine Learning



Data Access:

- Easily connect to Behind-the-Firewall and Public Cloud Data
- Catalogued and Governed Controls through Watson Data Platform

Creating Models:

- Single UI and API for creating ML Models on various Runtimes
- Auto-Modelling and Hyperparameter Optimization

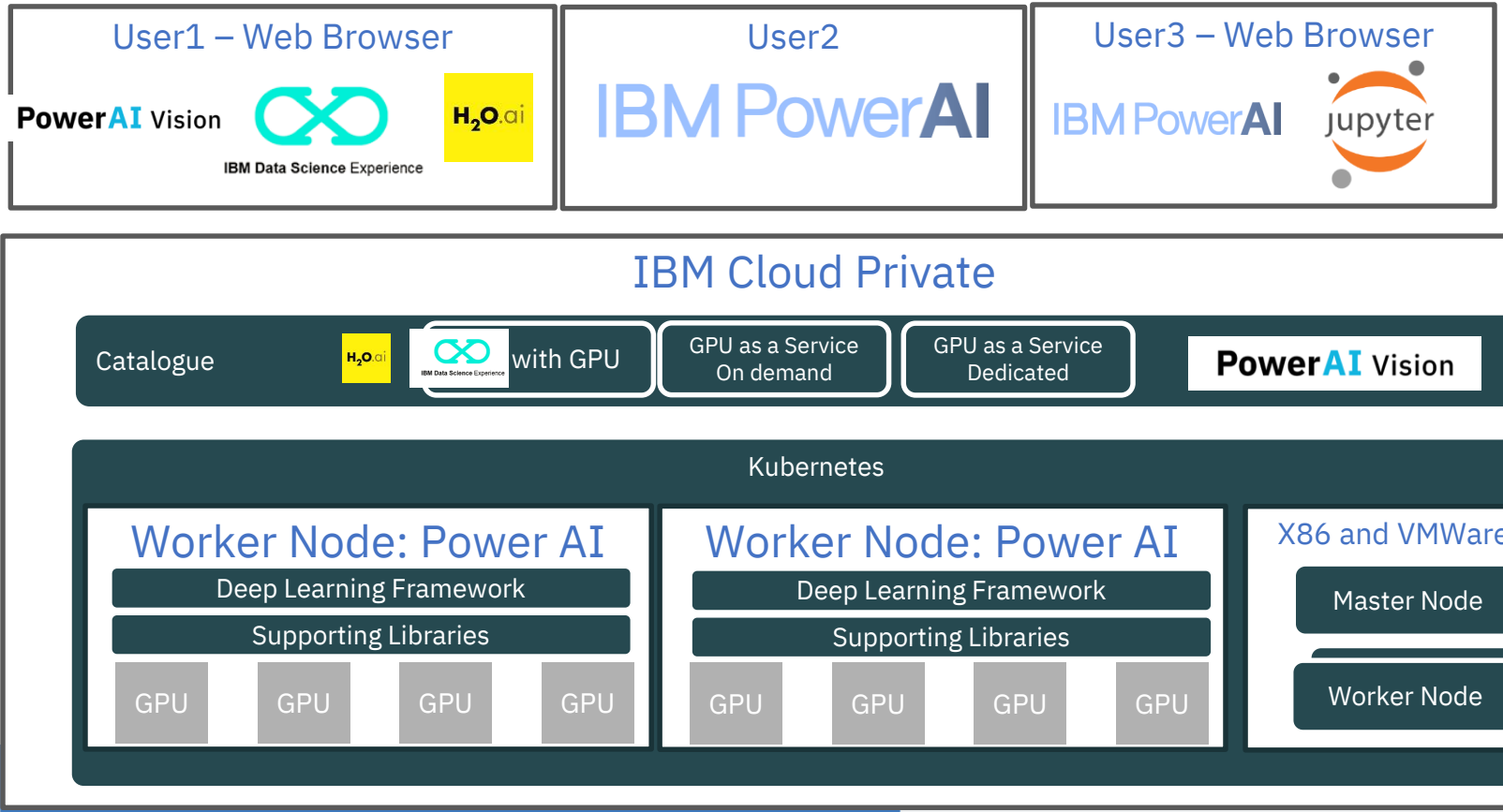
Web Service:

- Real-time, Streaming, and Batch Deployment
- Continuous Monitoring and Feedback Loop

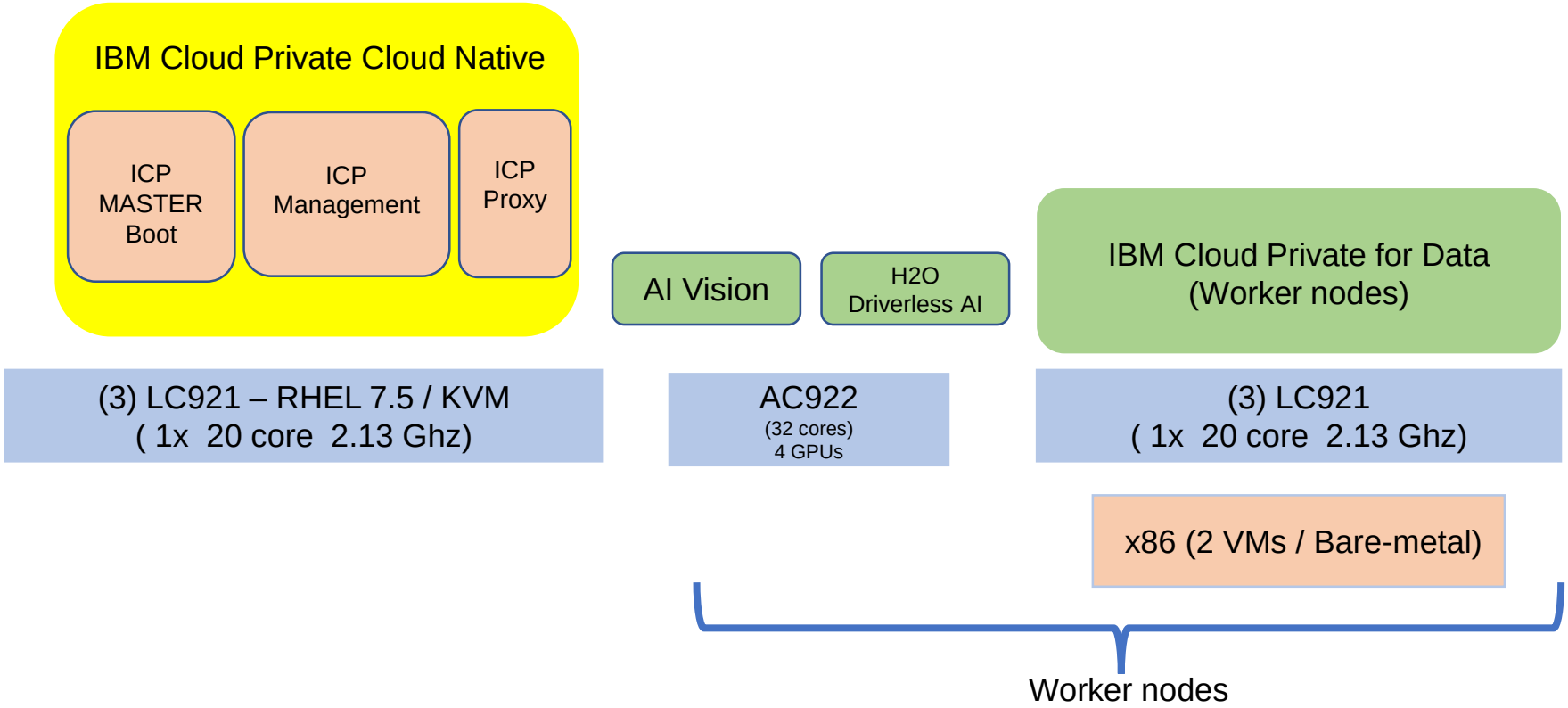
Intelligent Apps:

- Integrate ML models with apps, websites, etc.
- Continuously Improve and Adapt with Self-Learning

Kubernetes/IBM Private Cloud w/ PowerAI : Build your own AI Private Cloud



ICP + ICP4Data + PowerAI Vision + H2O Driverless AI

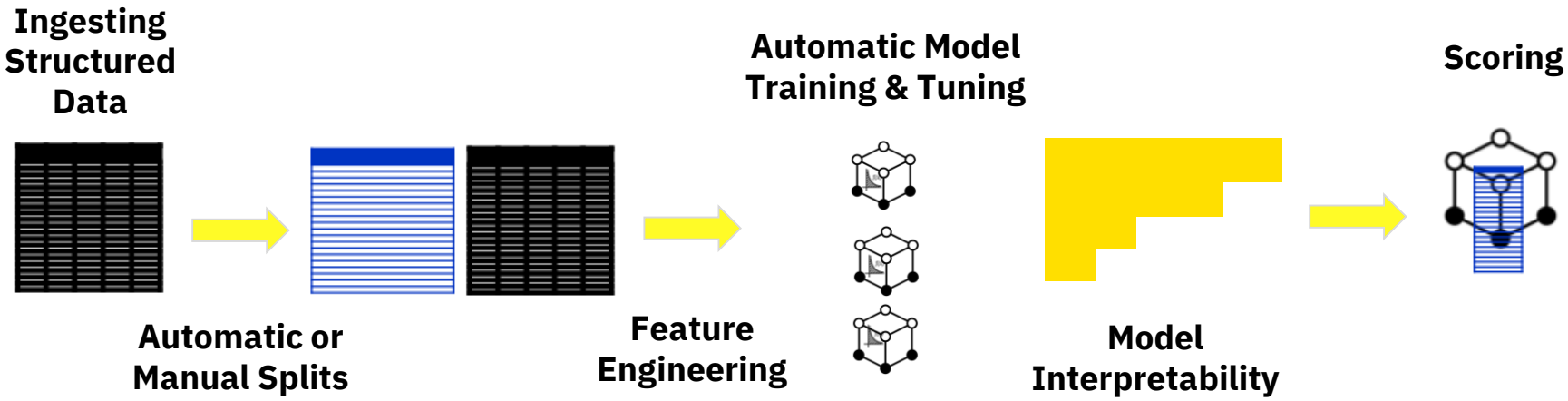


- H2o.ai Driverless AI on Power
  Toolset for increasing productivity
  Accelerated ML
  End to end AI project lifecycle

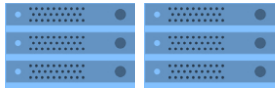
IBM PowerAI

IBM

H2O Driverless AI: Automated Machine Learning



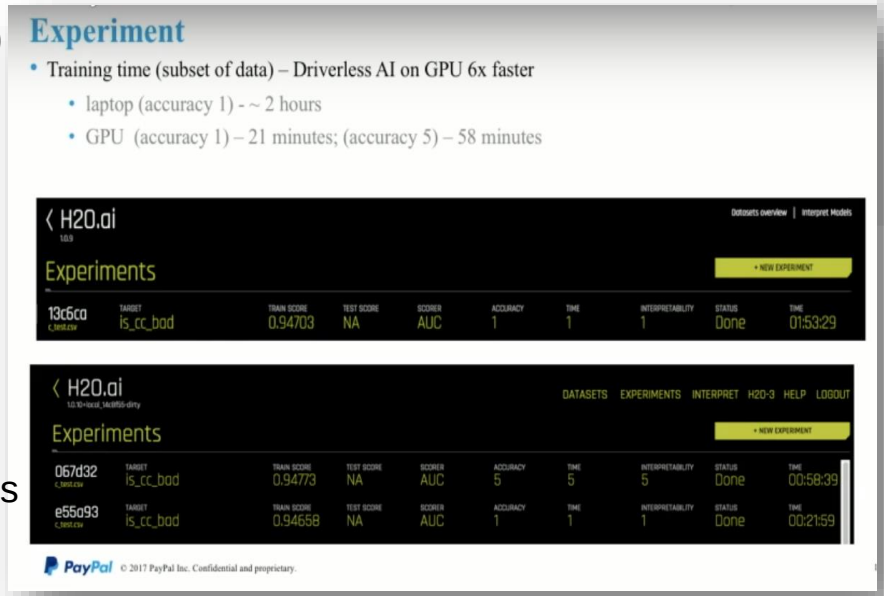
Optimized for GPU-Accelerated Power9 Servers



H2O.ai

Financial Fraud Detection

- Driverless AI matched **10 years** of expert feature engineering
- Increased accuracy from **0.89 to 0.947 (6%)** in detecting fraudulent activity
- **6X** speed up when using H2O4GPU with Driverless AI



"Driverless AI is giving amazing results in terms of feature and model performance "

Venkatesh Ramanathan
Senior Data Scientist, PayPal

Leader in Gartner's 2018 Data Science Quadrant

H2O Driverless AI and IBM POWER9 GPU Systems are bringing together the best of breed AI innovation. To handle the increasingly complex workloads of AI you need an integrated system of software and hardware:

- IBM POWER9 supports nearly 2.6x more RAM, 9.5x more I/O bandwidth than comparable systems.
- Nearly 2X the data ingest speed and over 50% faster feature engineering.
- With GPU accelerated machine learning delivering nearly 30X speedup on model building.
- Support for up to 6 V100 GPUs on a single system.

H2O Driverless AI: How it Works

1 **Drag and drop data**

Ingest data from cloud, big data and desktop systems

HDFS

SQL

Local

Amazon S3

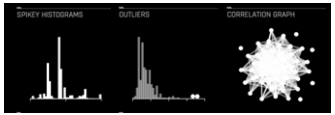
Snowflake

Google BigQuery

Azure Blog Storage

2 **Automatic Visualization**

Understand the data shape, outliers, missing values, etc.



Modelling

X Y

Dataset

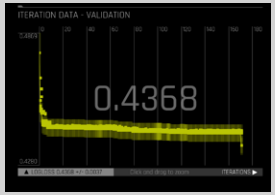
3 **Automatic Machine Learning**

Use best practice model recipes and the power of high performance computing to Iterate across **thousands of possible models** including advanced feature engineering and parameter tuning

Automatic Machine Learning

Model Recipes:
- i.i.d. data
- Time-series
- More on the way

Survival of the Fittest



Advanced Feature Engineering + **Algorithm** + **Model Tuning**

4 **Automatic Scoring Pipelines**

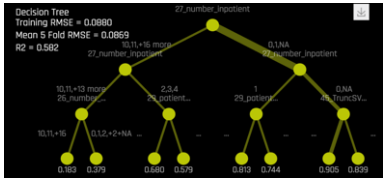
Deploy ultra-low latency Python or Java Automatic Scoring Pipelines that include feature transformations and models.

Model Documentation

Automatic Scoring Pipeline

Deploy Low-latency Scoring to Production

Machine learning Interpretability



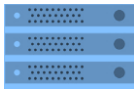
Powered by GPU Acceleration

H2O.ai NVIDIA IBM

Made with Driverless AI to make smarter IBM i apps

Sales + CRM data

Optimized for GPU-Accelerated Power9 Servers



Recommendation Engine
Scoring Pipeline - REST API

node.js Java python


```
[14:57:33][sp4i-appserver.power.com][/home/BENOIT/DAI]# uname -a
OS400 sp4i-appserver 2 7 00100006C37T
[14:57:37][sp4i-appserver.power.com][/home/BENOIT/DAI]# ./run_h2o_prediction_from_ibmi.sh M 21.0 Married Retail
This example script demonstrates how to communicate with the Driverless AI Scoring Service via HTTP from IBM i - PASE / Open Source / Any REST Client
The protocol used is JSON-RPC 2.0.
----------------------------------------
Name Type Range Value
----------------------------------------
GENDER object - M
AGE float32 [17.0, 69.0] 21.0
MARITAL_STATUS object - Married
PROFESSION object - Retail
----------------------------------------
Scoring individual rows from IBM i App...
{"jsonrpc": "2.0", "id": 1, "result": [0.7058841586112976, 0.03077976033091545, 0.04269004985690117, 0.022570349276065826, 0.19807571172714233]}
Get the target labels
{"jsonrpc": "2.0", "id": 1, "result": ["Camping Equipment", "Golf Equipment", "Mountaineering Equipment", "Outdoor Protection", "Personal Accessories"]}
```

Comparing AI Offerings on Power

		Deep Learning			ML and DL	Machine Learning
		Power AI Base (WML-CE)	Power AI Enterprise (WML-A)	AI Vision	Watson Studio Local	H2O Driverless AI
Offering	Description	Deep Learning	Deep Learning for the Enterprise	Deep Learning with Video tools	Notebook oriented development environment for ML and DL	Automated Machine learning
	Pricing Model	Free download	Commercial	Commercial	Commercial	Commercial
	Support	Available from IBM	IBM L 1-3 Included	IBM L1-3 Included	Available from IBM	H2O L 1-3
Applications	Text & Numeric	Yes	Yes	No	Yes	Yes
	Images	Yes	Yes	Yes	Yes	No
	Video	-	Optional add-on	Yes		No
	Primary Persona	Data Scientist	Data Scientist	Line of Business	Data Scientist	Data Scientist
	Second persona	IT	IT	IT	IT	Line of Business
	User Skill Level	High	Medium to high	Low	Medium to high	Low to Medium
	Strengths	Rapid deployment, high performance, scale	enterprise grade, High performance, rapid Deployment	Rapid deployment, simple GUI high performance	Notebook based development environment, strong collaboration, model management	Simplified deployment, intuitive user interface, automatic pipelines, "explainability" for models, end to end automation
Platform	Distributed DL (DDL)	1-4 nodes	1-thousands of nodes	Coming	Coming	-
	Large Model Support	Yes	Yes	Coming	Coming	-
	Server(s)	S822LC or AC922	S822LC or AC922	S822LC or AC922	S822LC or AC922, LC922	S822LC, AC922, LC921/922
IBM Products	Spectrum MPI (DDL)	Limited to 4 nodes	Included			Optional add-on
	Spectrum Conductor DLI	Optional add-on	Included	Coming	Optional Add On	Optional add-on
	IBM DSX Local	Optional add-on	Optional add-on	No		Optional add-on
Cloud	IBM Cloud Public	Yes	No	Trial only	Watson Studio	?
	IBM Cloud Private	Yes	Coming	Yes	Yes	Coming

500+ POWERAI CLIENTS WITHIN ONE YEAR

40% NEW TO POWER

- How to get Started? Pricing

AI Developer Box & AI Starter Kit

Power AI Developer Box	AI Starter Kit for Data Scientists
Free 30-Day Licenses for PowerAI Vision & WML Accelerator (free to Academia)  Power9 + GPU Desktop PC: $3,449 Order from: https://raptorcs.com/POWERAI/	WML Accelerator Pre-installed (formerly called PowerAI Enterprise)  2 AC922 Power9 + GPU Servers + 1 Linux Storage Server

Target street price $230K

IBM S822LC
POWER8
20 cores
4 - P100 GPUs
256GB RAM

Hardware List Price
$67,054

TCA Street Price
$53,654

TensorFlow
70

TCA Price/Performance
$766

IBM AC922
POWER9
32 cores
4 – V100 GPUs
256GB RAM

Hardware List Price
$68,854

TCA Street Price
$55,083

TensorFlow
140

TCA Price/Performance
$393

DELL C4130
x86 Broadwell
28 cores
4 – V100 GPUs
256GB RAM

Hardware List Price
$87,569

TCA Street Price
$70,055

TensorFlow
70

TCA Price/Performance
$1001

POWER / x86 Price Index

	POWER8 vs. Broadwell	POWER9 vs. Broadwell
Hardware List Price Ratio	0.77	0.77
TCA Street Price Ratio	0.77	0.77
TCA Price Performance Ratio	0.77	0.39

POWER9
HIGHLIGHTS

- Getting questions on GPU pricing compared to on-line pricing? IBM's net price is competitive!
  - Dell List price for a Tesla V100 GPU is $16,919 vs. $11,499 in AC922
- Drive home AC922 GPU efficiency versus commodity x86 servers … CPU←→NVLink and LMS is a differentiator
- P8 to P9 accelerated computing servers are a price point replace at 256GB memory capacities

Start your Journey

PowerAI Vision

Offering details

- Introduction to IBM PowerAI Vision
- Developer Works

Access to Software

- Trial editon, FREE for 30 days
- Low cost accelerated HW from Raptor
- As a service on NIMBIX & MeridianIT
- Demo instance on Client Demo Center
- Partners access Software catalog
- Forum for support, ask us for help

Developer Journeys

- Counting cars
- Train models to detect flavors of Coke
- Classify Photo resist wafers

Videos on YouTube

- Train models for Classification and Object detection
- Train models for Advanced Driving Assistant Systems
- Continuous learning for data labeling
- Follow our channel on YouTube

Agenda

Introduction – AI Trends & Challenges

PowerAI & PowerAI Enterprise (WML-A) & AC922 HW Introduction
Technology Differentiators, Accelerated Machine Learning

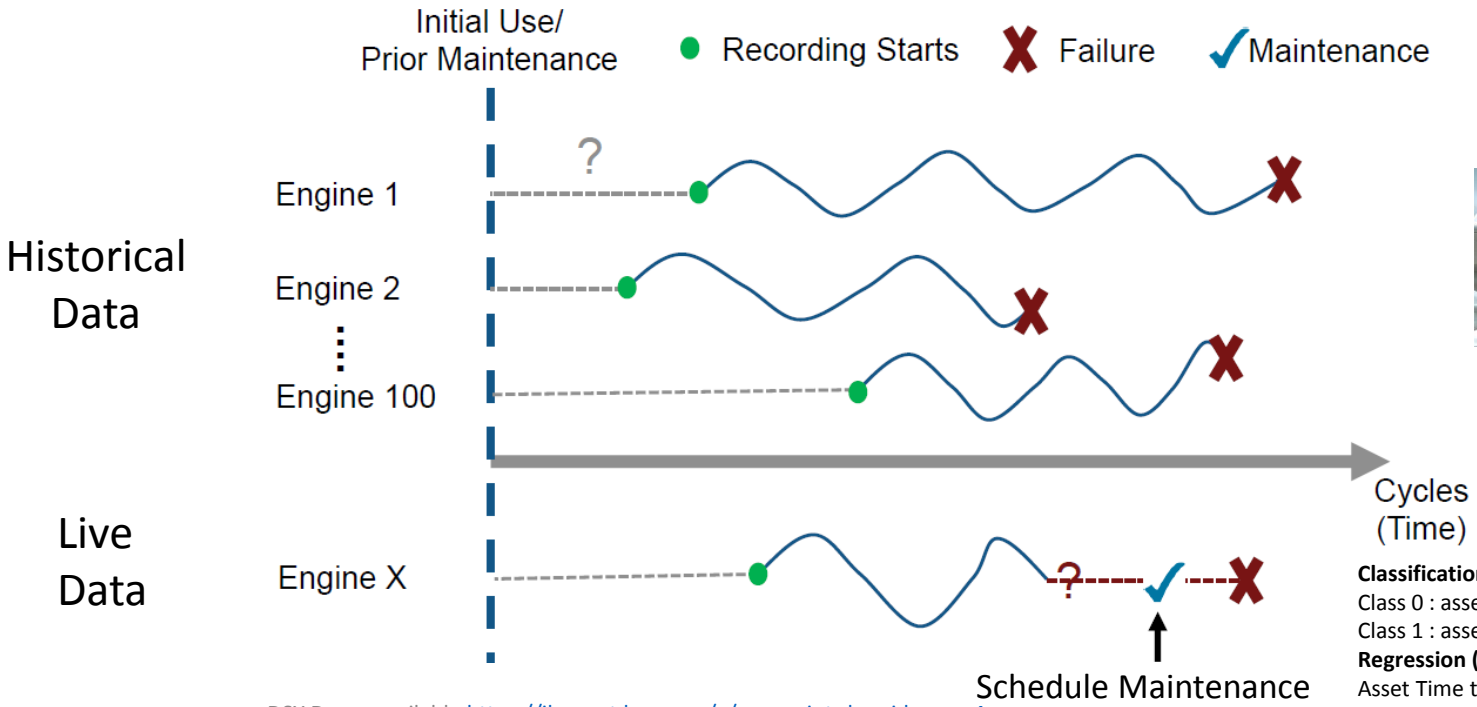
Private AI Solutions Overview based on PowerAI :

PowerAI Vision & Intelligent Video Analytics
Watson Studio, ICP4Data
h2O.ai Driverless AI

Appendices/Bonus:

AC922 details & pricing, PowerAI Public References
Demos & Illustrations:
AI with IBM Cloud Private w/ PowerAI, AI Vision, H2O.ai, Watson Studio (DSX) Local :
H2O.ai Driverless AI quick demo on ICP/POWER
DSX & Predictive Maintenance Example

Example: Predictive Maintenance



Classification:
Class 0 : asset will fail within the next 15 cycles
Class 1 : asset won't fail within the next 15 cycles
Regression (prediction):
Asset Time to failure = 18 cycles

DSX Demo available https://ibm.ent.box.com/v/power-iot-dsx-video-mp4
Historical Data from NASA https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/

Get Started Today

PowerAI Developer Portal
https://developer.ibm.com/linuxonpower/deep-learning-powerai/technology-previews/powerai-vision/

AI Vision Object Detection **Demo**
https://www.youtube.com/watch?v=19vaot75JCY
https://github.com/IBM/powerai-counting-cars/blob/master/notebooks/counting_cars.ipynb

AI Vision / Public Cloud – Get Started **demo**
https://github.com/IBM/powerai-vision-object-detection

PowerAI **FAQ**
https://developer.ibm.com/linuxonpower/deep-learning-powerai/faq/

PowerAI 1.5.3 on IBM Cloud **Free trial**
https://console.bluemix.net/catalog/services/powerai

PowerAI Vision 1.1.1 **Free trial**
Register for a free 3-day trial of PowerAI Vision

Power-Up your journey to **AI** with IBM **Cognitive Systems Lab** Montpellier

Get Started Today with Machine & Deep Learning

PoC/Support /Enablement/Workshops…

Engage us

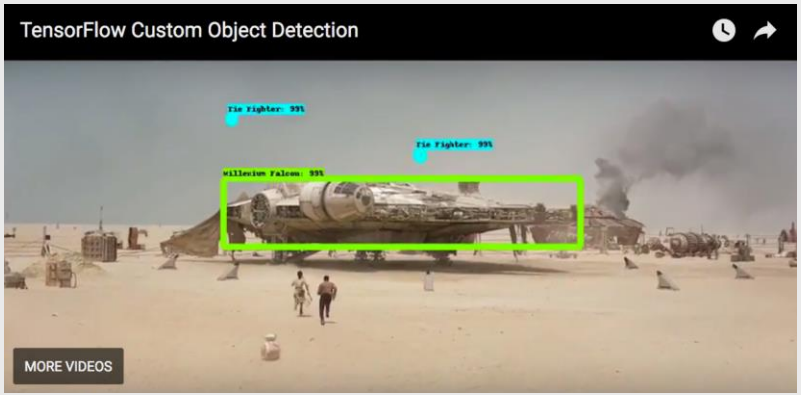
Alain Roy
Cognitive Systems Lab Engagement Leader
IBM Client Center Montpellier
a2roy@fr.ibm.com

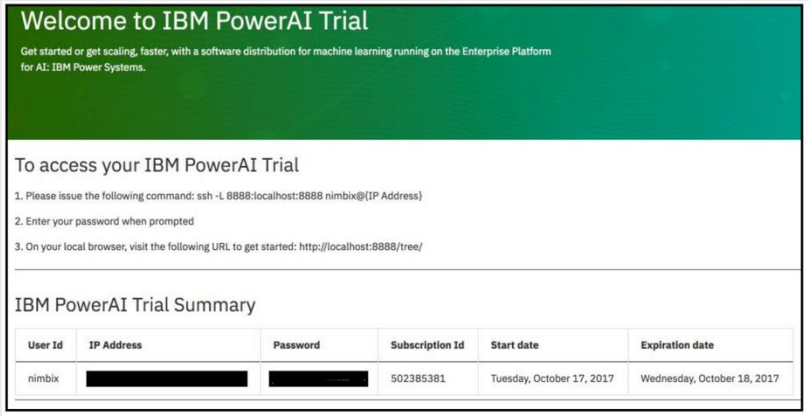
Philippe Chonavel
Cognitive Systems Lab Manager
IBM Client Center Montpellier
p_chonavel@fr.ibm.com

1. Build a Data Science Team
   Your Developers Can Learn
   http://cognitiveclass.ai
2. Identify a Low Hanging Use Case
3. Figure Out Data Strategy
4. Consider Pre-Built AI APIs
5. **Contact Us**
6. Get Started Today at
   www.ibm.biz/poweraideveloper

IBM PowerAI

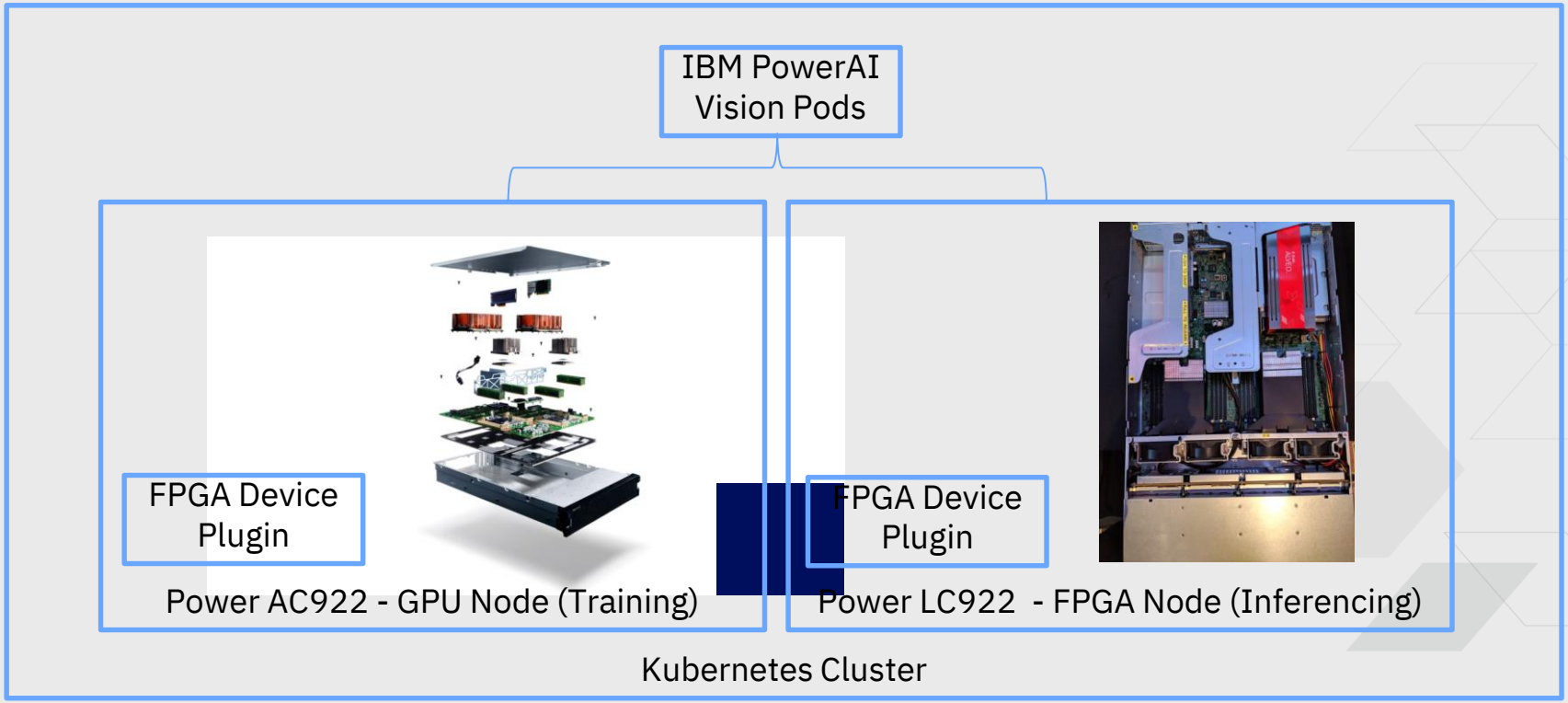
ibm.biz/poweraidevelopper





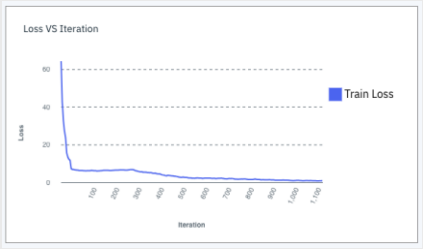
Backup

Kubernetes Cluster



Integration with Xilinx MLSuite

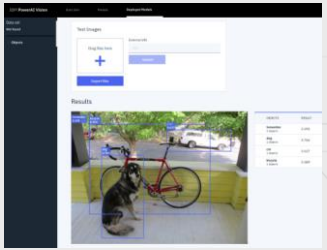
Train YOLO



Deploy

Schedule Available FPGA

Run Inference



Inference Container

{ REST API }

Deploy.py

GPU or FPGA?

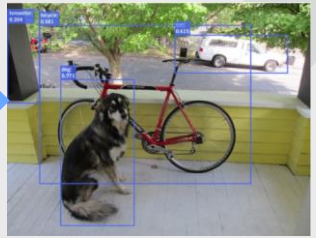
xfDNN Compile & Quantize (if needed)

xfDNN Runtime

Caffe Prototxt, Weights, & Anchors

Sample Training Images


```
{webAPIId: "bf1c53fa-b931-4697-9ef7-985137b2e080",…}
  classified: [,…]
    0: {confidence: 0.304, ymax: 246, label: "tvmonitor", image_id: "image.
    1: {confidence: 0.771, ymax: 549, label: "dog", image_id: "image.png",
      confidence: 0.771
      image_id: "image.png"
      label: "dog"
      xmax: 314
      xmin: 132
      ymax: 549
      ymin: 188
    2: {confidence: 0.615, ymax: 172, label: "car", image_id: "image.png",
    3: {confidence: 0.581, ymax: 445, label: "bicycle", image_id: "image.pn
  imageUrl: "https://github.com/pjreddie/darknet/raw/master/data/dog.jpg"
  result: "success"
  webAPIId: "bf1c53fa-b931-4697-9ef7-985137b2e080"
```

POWER9 Systems Roadmap

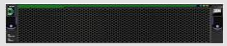
2017

POWER9 (SO)

7965-S42 Rack

P9 SO SMT4

AC922



Linux

2018

POWER9 (SO & SU)

E980

P9 SU SMT8

AIX

i for Business

Linux

S924, H924
S914

S922 / H922

L922

AIX

i for Business

Linux

PowerVM

P9 SO SMT8

PowerVM

Linux

P9 SO SMT8

E950

P9 SU SMT8

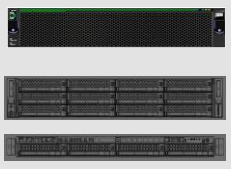
AIX

Linux

AC922 (GA2)

LC922

LC921



Linux

P9 SO SMT4

4Q

1Q

2Q

3Q

4Q

Power Systems Product Portfolio

POWER9 servers and solutions are built to crush today's most advanced data applications … from the mission critical applications you run today to the next generation of AI workloads

Mission Critical Workloads

Entry	Mid	Enterprise	Big Data Workloads	Enterprise AI Workloads
S922/S914/S924 H922/H924/L922 **Q1 2018**	E950/H950 **Q3 2018**	E980/H980 **Q3 2018**	LC922/LC921 **Q2 2018**	AC922 **Q4 2017**

Core Infrastructure

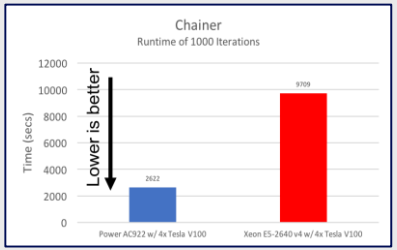
Next Gen AI Workloads

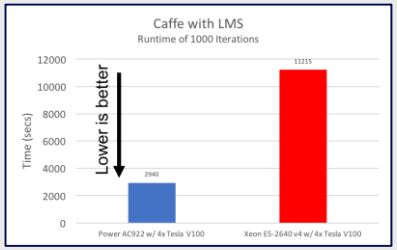
IBM **Power AI Systems** Performance

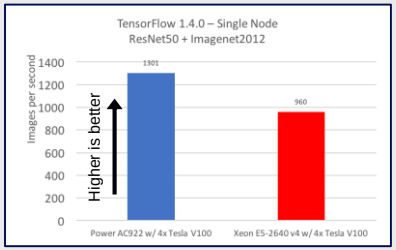
IBM **PowerAI Enterprise**

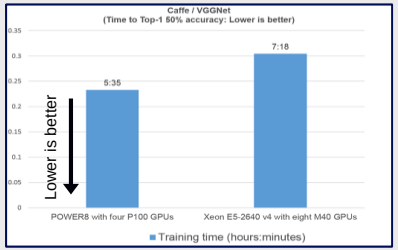
Faster training times than x86[1]

- **3.7x** faster running Chainer[2]
- **3.8x** faster running Caffe[3]
- **2.3x** faster running TensorFlow[4] and **35%** more images per second[5]
- Four (4) GPUs on Power is faster than eight (8) GPUs on Intel[6]









Extended references in notes section [1]
https://developer.ibm.com/linuxonpower/perfcol/perfcol-mldl/

Public References

PowerAI client examples: using deep learning for better outcomes

- Fraud and crime prevention for Automated Tellers, facial recognition
- Detects attempts to disguise, or hide facial features

- Manufacturing process inspection for hot steel rolling
- Reducing defects through early detection

- Technology incubator for AI in healthcare
- Increase diagnostic accuracy, reduce examination time

- Reducing credit approval turn around time, while at the same time decreasing potential risk of default
- Improved personalization for customer offers: next offer, next best offer

PowerAI client examples: using deep learning for better outcomes

- Personalized quotes based on vehicle inspection and insurance certificates
- Minimize lead time for insurance quotes and limit lost sales opportunities

- New video analytics model assembles “trailer” for TV program
- From 30 days down to a day

- Natural language and data pattern analysis for e-discovery
- Quickly respond and adapt to privacy/right to be forgotten requests

- Deep learning model improves accuracy of real-time bio simulations
- View accurate rendering of a patient’s heart before the procedure

IBM S822LC
POWER8
20 cores
4 - P100 GPUs
256GB RAM

Hardware List Price
$67,054

TCA Street Price
$53,654

TensorFlow
70

TCA Price/Performance
$766

IBM AC922
POWER9
32 cores
4 – V100 GPUs
256GB RAM

Hardware List Price
$68,854

TCA Street Price
$55,083

TensorFlow
140

TCA Price/Performance
$393

DELL C4130
x86 Broadwell
28 cores
4 – V100 GPUs
256GB RAM

Hardware List Price
$87,569

TCA Street Price
$70,055

TensorFlow
70

TCA Price/Performance
$1001

POWER / x86 Price Index

	POWER8 vs. Broadwell	POWER9 vs. Broadwell
Hardware List Price Ratio	0.77	0.77
TCA Street Price Ratio	0.77	0.77
TCA Price Performance Ratio	0.77	0.39

POWER9
HIGHLIGHTS

- Getting questions on GPU pricing compared to on-line pricing? IBM's net price is competitive!
  - Dell List price for a Tesla V100 GPU is $16,919 vs. $11,499 in AC922
- Drive home AC922 GPU efficiency versus commodity x86 servers … CPU←→NVLink and LMS is a differentiator
- P8 to P9 accelerated computing servers are a price point replace at 256GB memory capacities