Power Week

Université IBM i 2019

22 et 23 mai

IBM Client Center Paris

S04 - PowerAI, Watson et IBM i – Partie 1 Introduction, Cas d'usage et solutions

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

linkedin.com/in/benoitmarolleau
@MarolleauBenoit

Online version – EN & FR
https://ibm.biz/bma-wiki

Agenda

- PARTIE 1
  - Introduction au Machine Learning et cas d'usage
  - Les solutions
    - IBM i, Open Source et Machine Learning
    - IBM "One AI" & solutions dédiées : PowerAI & WML

- PARTIE 2 (S12 – Prochaine Session)
  - Exemples et demonstrations – Technologies en action sur IBM i

    Machine Learning sur IBM i: Scikit-Learn demo

    Machine Learning accéléré sur H2O.ai Driverless AI

    Deep Learning (Visual Recognition) avec PowerAI Vision, IBM i et Node.js

Power Week

Université IBM i

22 et 23 mai 2019

Machine Learning ?
Pourquoi ?

IBM

Introduction – Avant propos

Avant de commencer, quelques précisions sur le Machine Learning / Deep Learning :

- Le Machine Learning apprend des données, évite de programmer explicitement les règles métiers
- Le ML établit des relations de causalité dans les données, alors qu'observe un phénomène (complexe?) qui se répète.
- L'expertise (métier) est nécessaire en phase d'apprentissage pour façonner l'algorithme, le modèle.
- La puissance de calcul, des librairies et Framework sont nécessaires.
- Domaine à la croisée de la robotique, la philosophie, la sociologie, les mathématiques…

L'Intelligence Artificielle en se résume pas à cela, c'est l'orchestration de modèles ML, DL (divers domaines: images, langage, …) et de techniques traditionnelles (moteur de règles, systèmes expert, applications métiers).

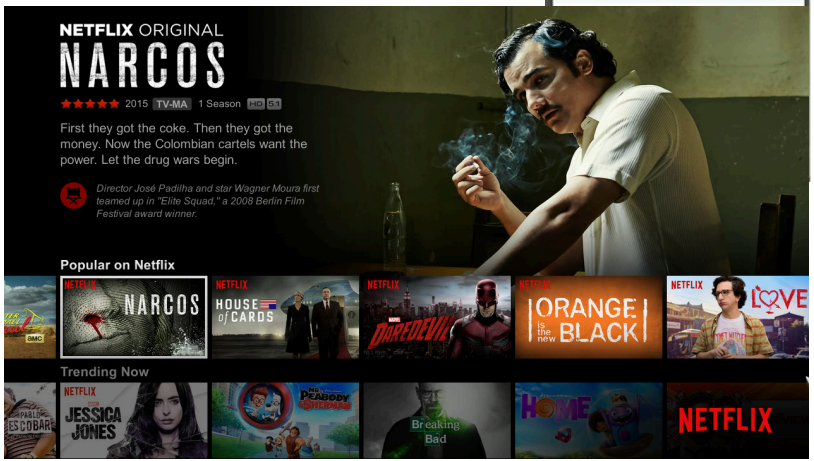
Introduction – Avant propos

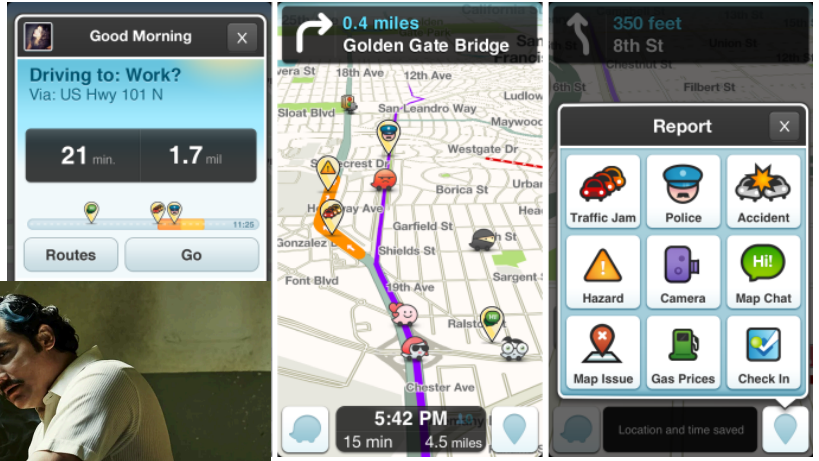
Pourquoi s'y intéresser? Quelques exemples…

- Aide à la décision
  - Détection d'un évènement (images, données structurées), Classification, segmentation des clients
- User Experience (UX)
  - User Interface par l'image, la voix, le texte, helpdesk, chabot (NLP)
- Augmentation (aide) et automatisation
  - Taches répétitives, ex: sur les images, Inspection, Accès difficile, complexe (lien avec la robotique)
  - Un médecin passerait des centaines d'heures à lire les publications utiles pour sa prise de décision: l'IA peut être utile pour l'aider.

Le Machine learning est partout ...
et influence notre vie de tous les jours!

Netflix provides personalized movie recommendations





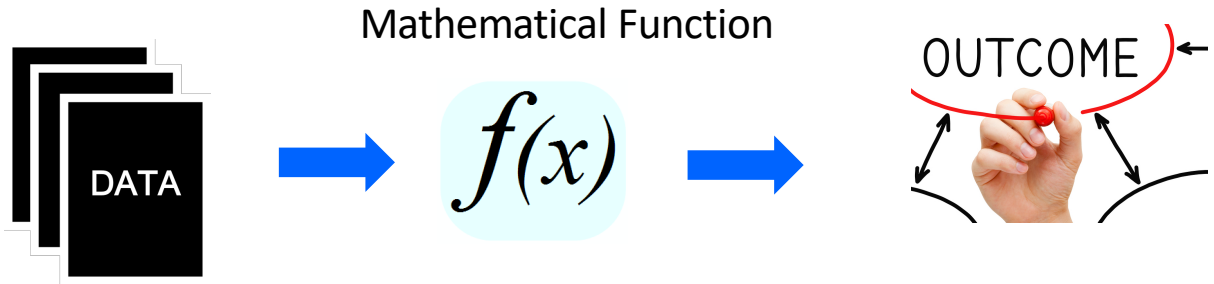
Waze provides a personalized driving experience for its users

Aide à la décision

- 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

- **Fonction décrivant des pattern par des fonctions mathématiques**
- **Le Machine learning est fondé sur les mathématiques**

L’objectif d’un algorithme Machine Learning

Use training data to derive f(x) so that

Minimize (Actual - f(x))

or mathematically

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

- Pendant la phase d’apprentissage, on cherche à minimiser la distance entre le résultat obtenu par l’algorithme f(x) et la réalité (apprentissage supervisé).
- Le volume et la qualité des données est crucial (Training Dataset). Connaissance du phénomène observé.
- Chaque algorithme à ses **paramètres**, qui vont être ajustés durant l’apprentissage en fonction de l’erreur.
- `Le choix de l’algorithme est clé, l’initialisation des **hyperparamètres** également.
- Des outils existent afin d’assister dans ce travail.

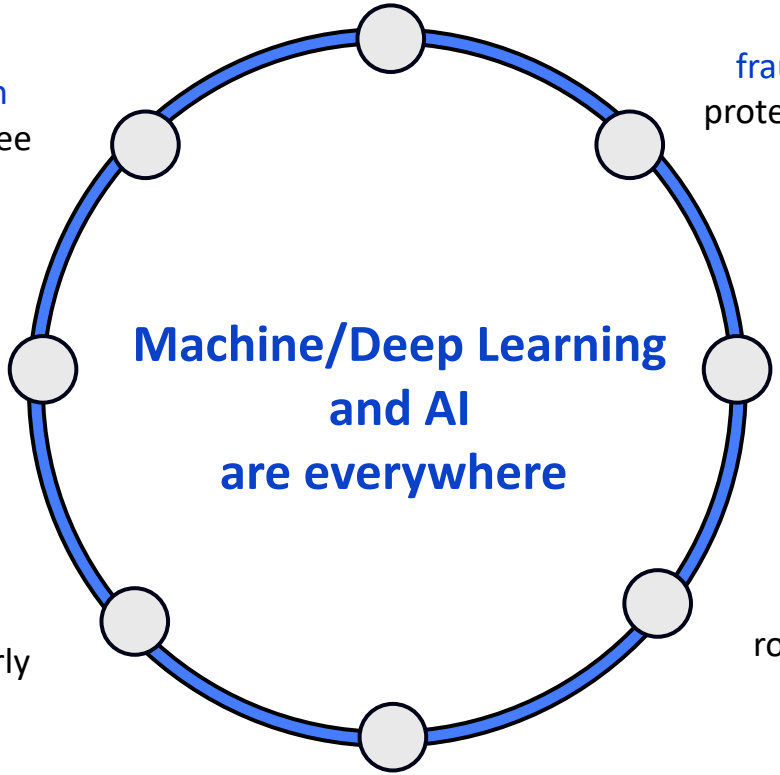
Exemple: Acceptation de prêt

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
…	…	…	…	…

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

Le Machine learning peut prendre en compte des centaines de “Feature”

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

Impact du Machine Learning: Exemple simple

***Direct* marketing — 1% response rate**		
Send marketing mail to 1,000,000 customers at cost of $2 per mailing to sell a $220 service.	$2 x 1,000,000	$2,000,000
One percent response rate means 10,000 customer will buy service.	$220 x 10,000	$2,200,000
	Profit*	**$200,000**
***Predictive* direct marketing — 3% response rate**		
Send marketing mail to 250,000 customers *predicted most likely to buy* at cost of $2 per mailing to sell a $220 service.	$2 x 250,000	$500,000
Three percent response rate means 7,500 customer will buy service.	$220 x 7,500	$1,650,000
	Profit *when using a ML model**	**$1,150,000**

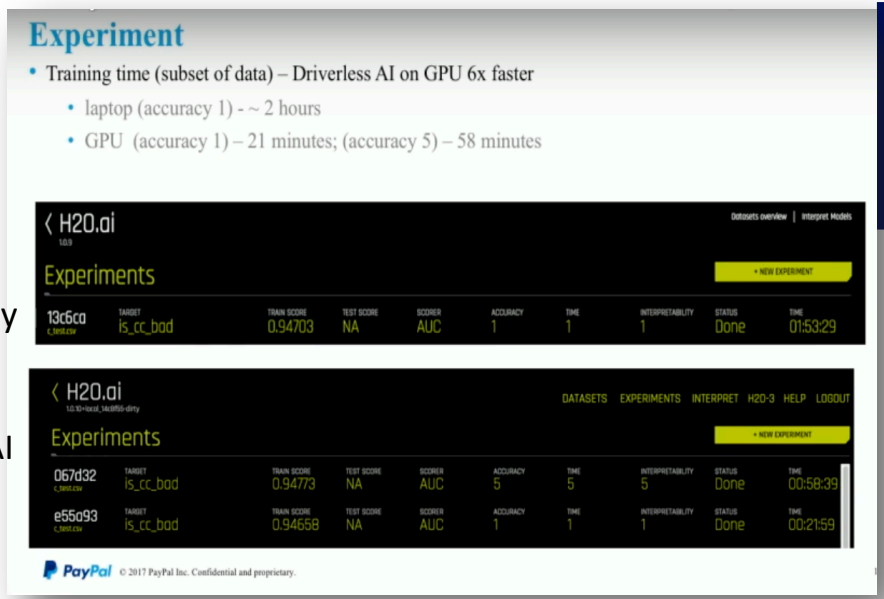
Traditional

Machine learning

*Profit calculation does not include other expenses.

Exemple: Detection de fraude

- 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 "

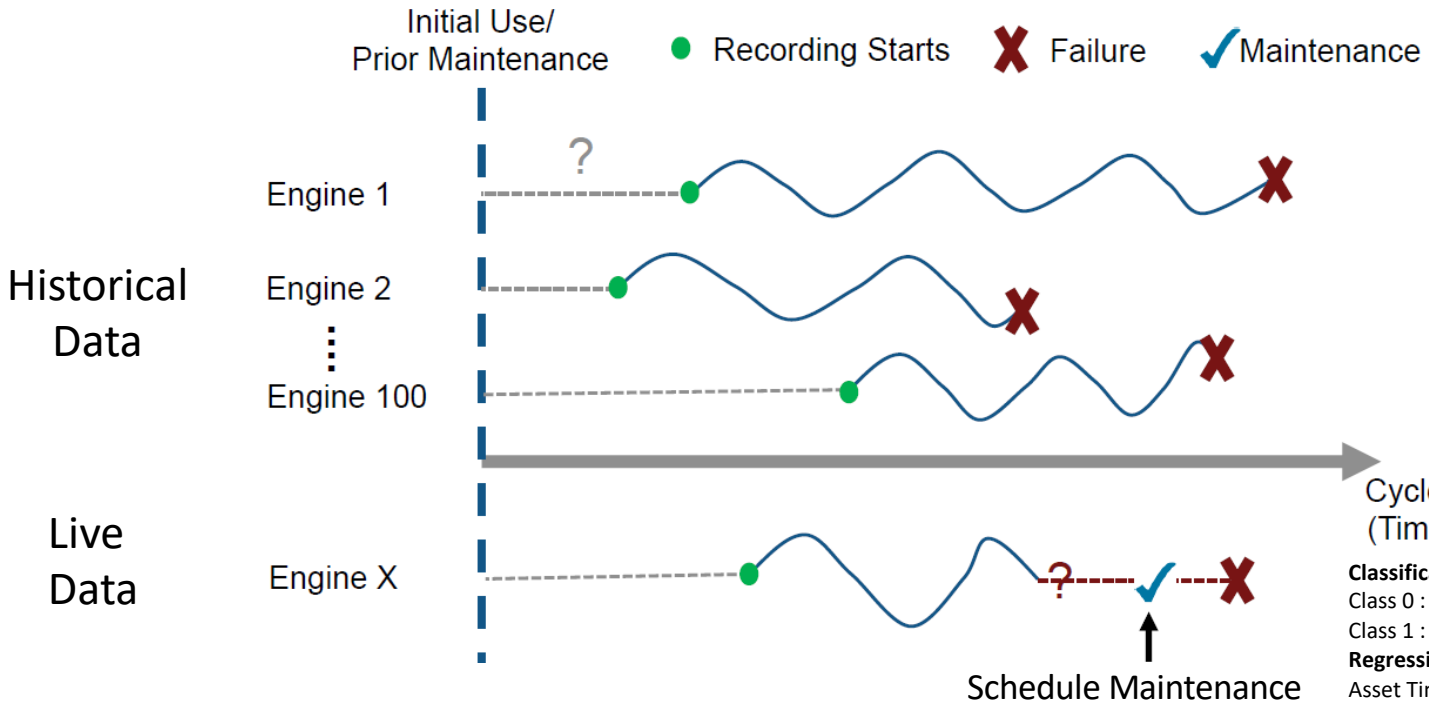
Venkatesh Ramanathan
Senior Data Scientist, PayPal

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.

Exemple: Maintenance Predictive



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/

Quelques cas d’usage…

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)

Power Week

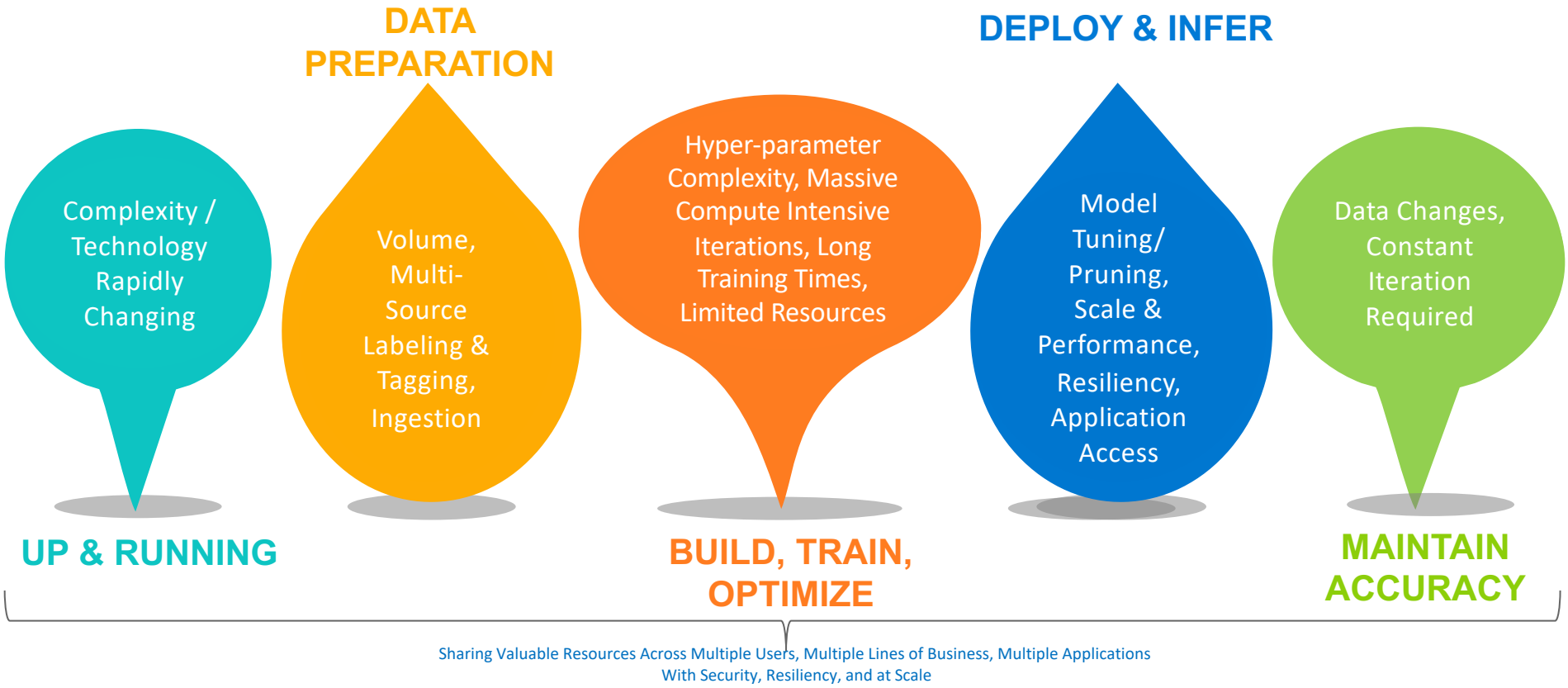
Université IBM i

22 et 23 mai 2019

Machine Learning ? Challenges

IBM

Pain Points – Deep Learning Pipeline



Enterprises generate TONS OF DATA

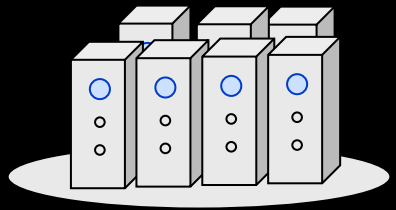
Data that requires governance

Which must be cleaned and shaped for training

…then models must be designed

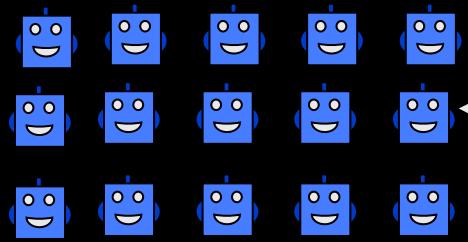
TensorFlow ™ scikit learn PYTORCH SPSS K Keras Caffe

…and trained on high performance compute



To select an optimal model...

…that must be hosted and monitored



Why are enterprises struggling to capture the value of AI?

- Data
- Data resides in silos & difficult to access
- Unstructured and external data wasn't considered

- Governance
- If the data isn't secure, self-service isn't a reality
- Challenge understanding data lineage and getting to a system of truth

- Skills
- Data Science skills are in low supply and high demand
- Nurturing new data professionals is challenging

- Tools & Infrastructure
- Need an environment that enables a "fail fast" approach
- Discrete tools present barriers to productivity

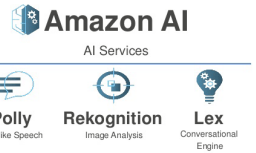
IA dans le Cloud?
Oui mais pas seulement...

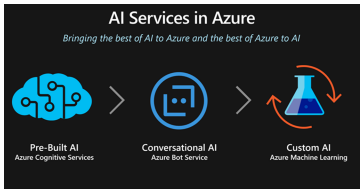
Set of Pre-trained Models
API Invocation
Ready to use
Majority of use cases

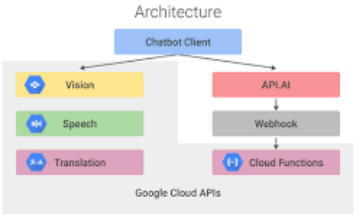
Additional Offerings to build new Models using
State of the art Frameworks

Limitations:
Data Gravity
Compliance & Regulations
...

=> AI will be :
Public Cloud
+ **Private** Cloud (Datacenter)
+ **Edge** (embedded device)







Big Data Sources

Business Data (Sensitive, Personal Data)

Power Week

Université IBM i

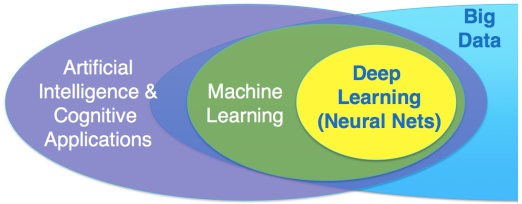
22 et 23 mai 2019

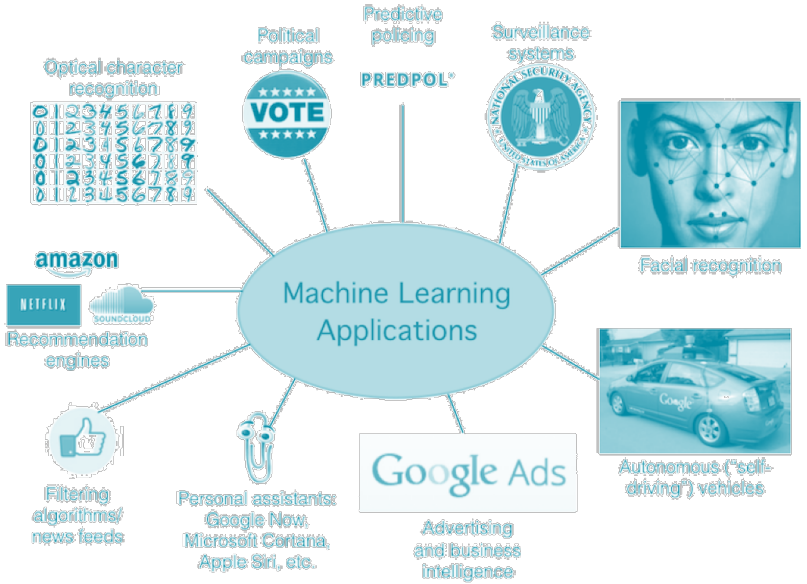
Machine Learning ?
Comment?

IBM

Machine Learning techniques

- Classification: predict class from observations
  - - E.g. Spam Email Detection
- Clustering: group observations into “meaningful” groups
  - - E.g. Amazon Recommendations
- Regression (prediction): predict value from observations
  - - E.g. Energy consumption prediction





and many different technologies and libraries are available:

scikit learn

Spark MLlib

WEKA The University of Waikato

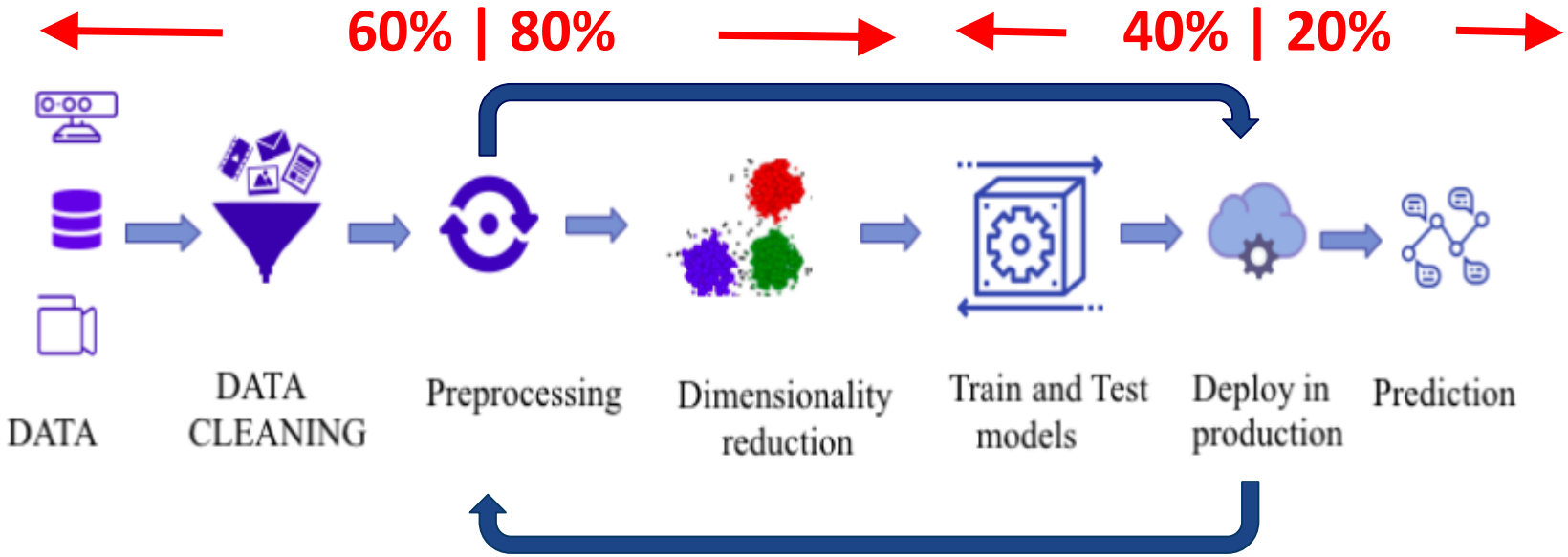
Caffe

SYSTEM ML

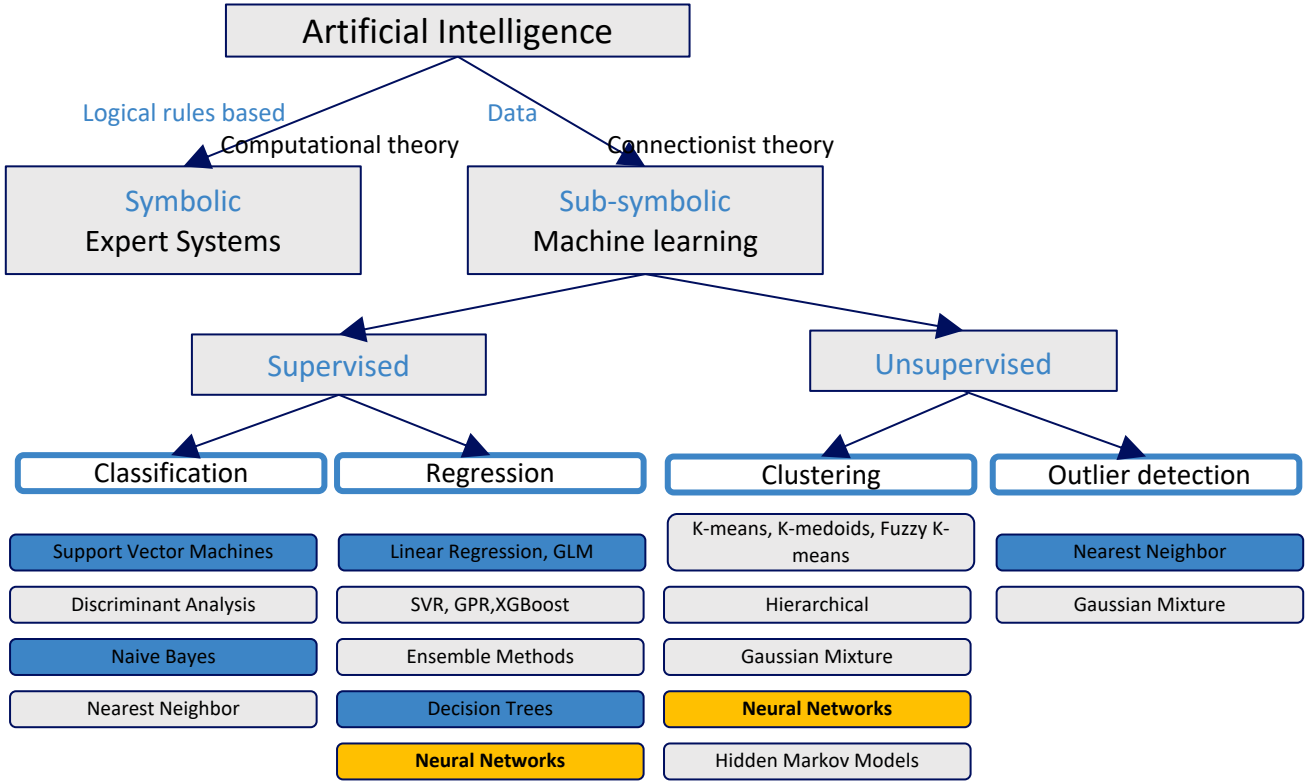
TensorFlow

SPSS®

Machine Learning: A few things

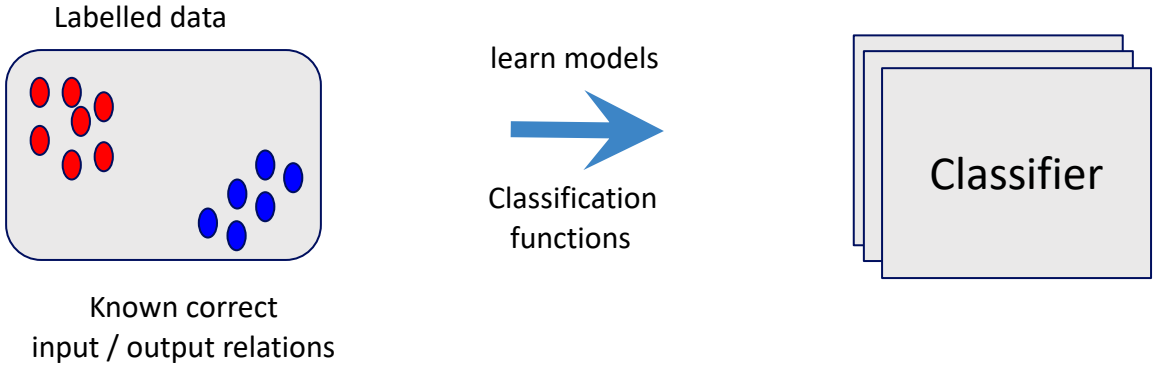


The Machine Learning Tasks & algorithms



Supervised learning

Infer input/output functions/models from labelled data



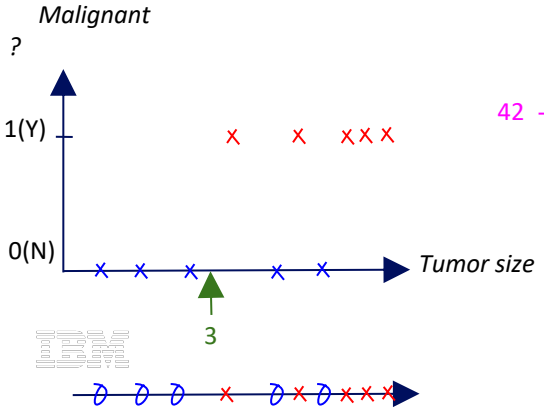
Supervised learning : "right answers" or "ground truth" given

Classification

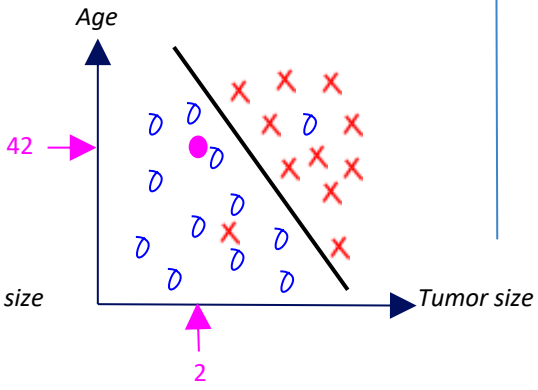
- The output variable takes class labels (discrete valued output)

Breast Cancer (malignant, benign)

1 variable



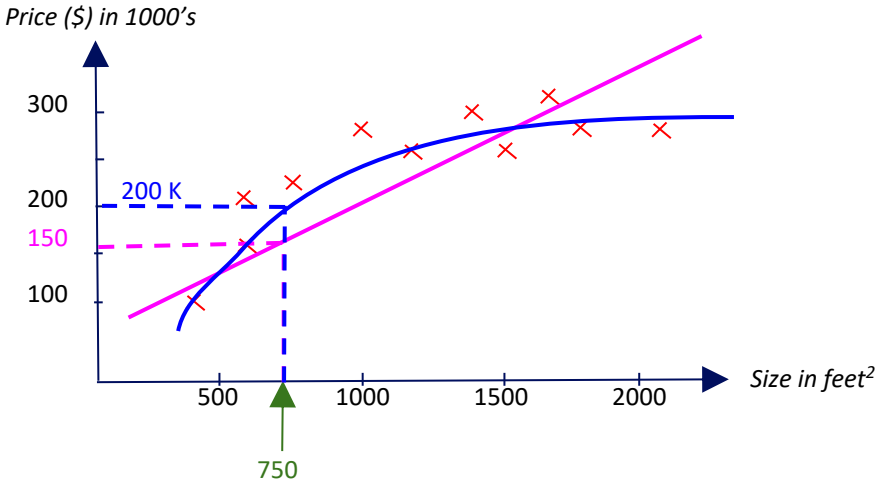
2 variables



Regression

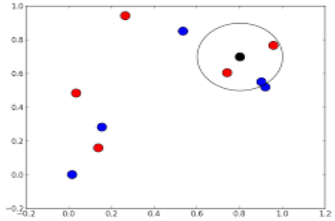
- Predict continuous valued output.

Housing price prediction



Machine Learning Methods (examples)

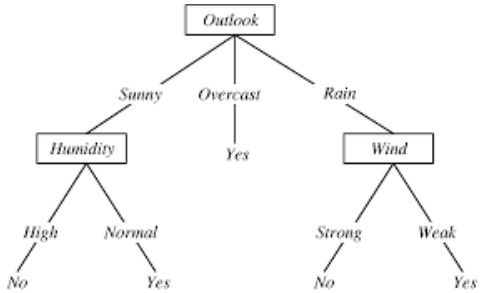
K Nearest Neighbours



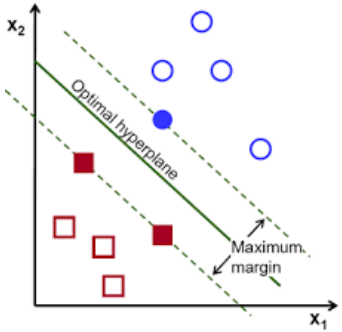
Bayesian Classifiers

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

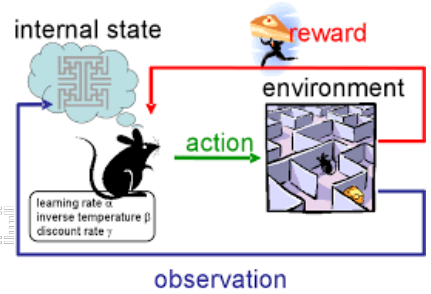
Decision trees



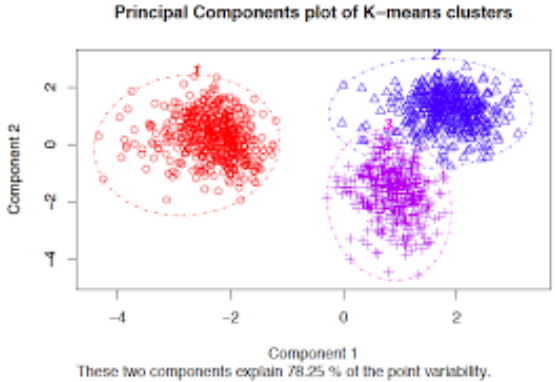
Support Vector Machines



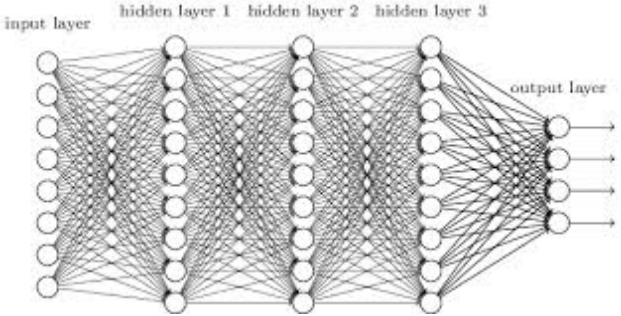
Reinforcement learning



Cluster Analysis

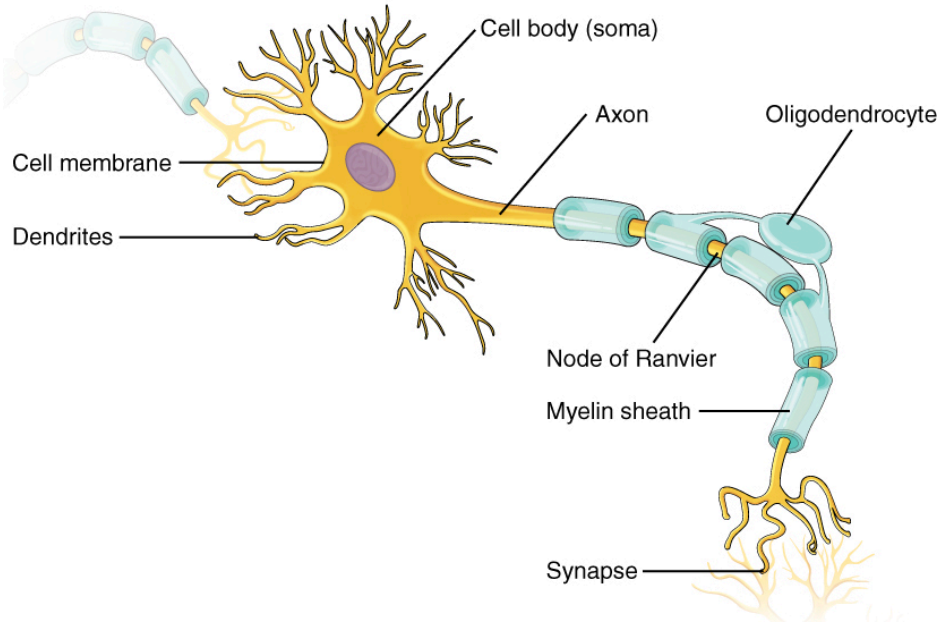


Neural Networks / Deep learning



Deep Learning = Training Artificial Neural Networks

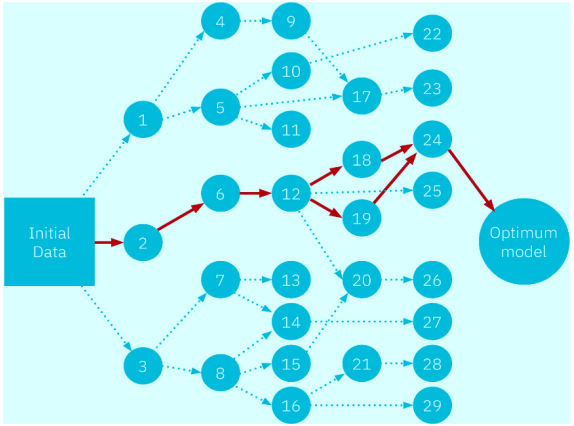
Based on biological neurons. Artificial neurons learn by recognizing patterns in data.



A human brain has:

- 200 billion neurons
- 32 trillion connections between them

Artificial neural networks have far fewer

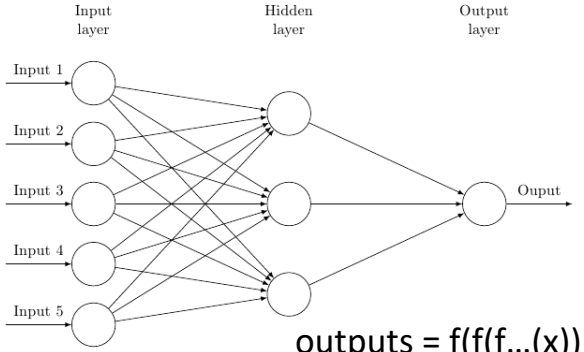


Deep Learning = Training Artificial Neural Networks

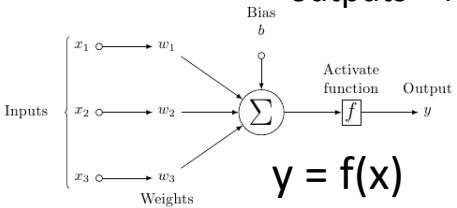
Based on biological neurons. Artificial neurons learn by recognizing patterns in data.

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

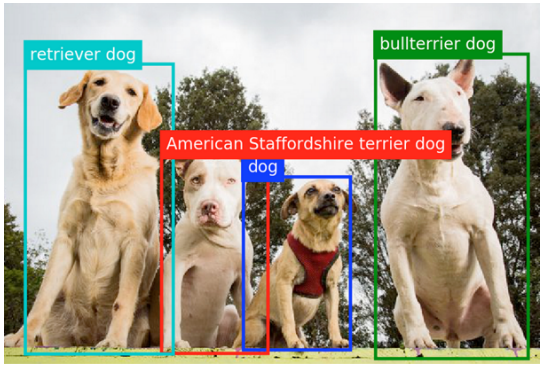
Neural net Framework
& Python Programming



outputs = f(f(f…(x))



y = f(x)

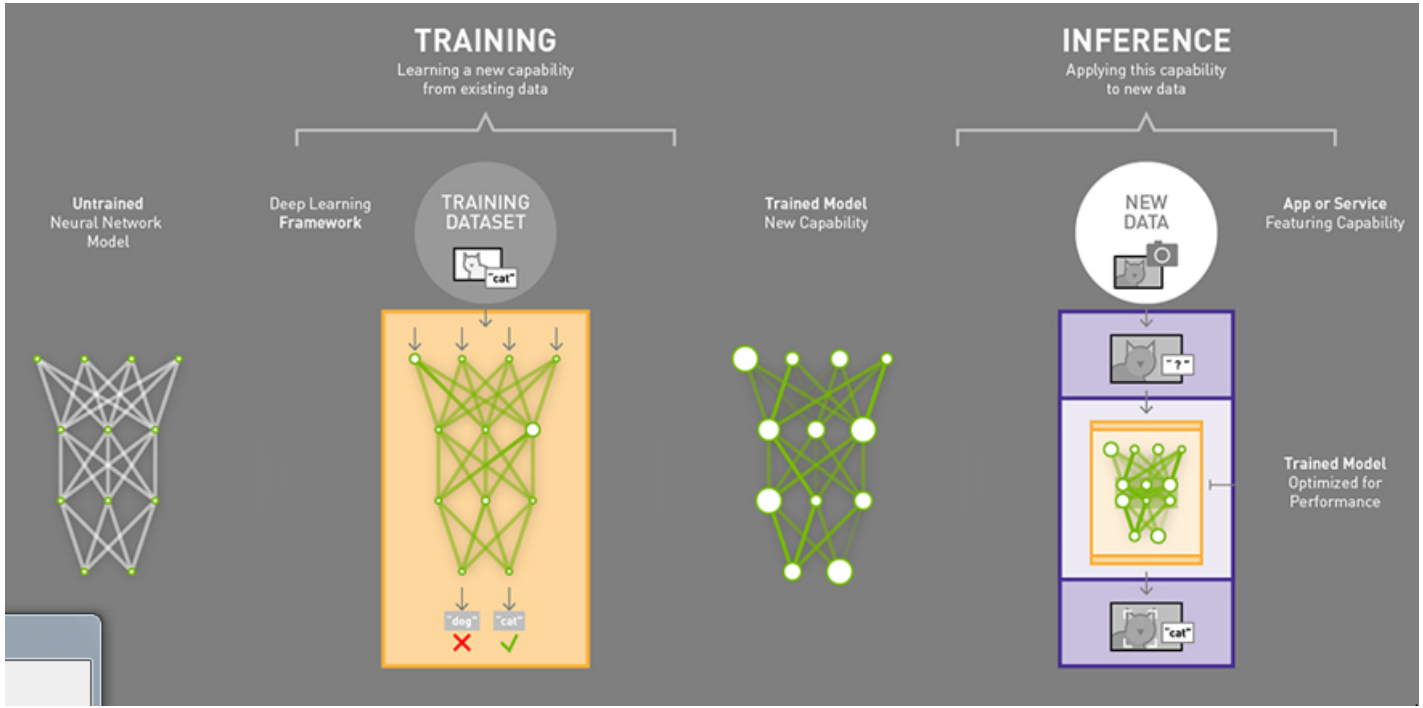


Each output contains:
detection_boxes
detection_scores
detection_classes
num_detections

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

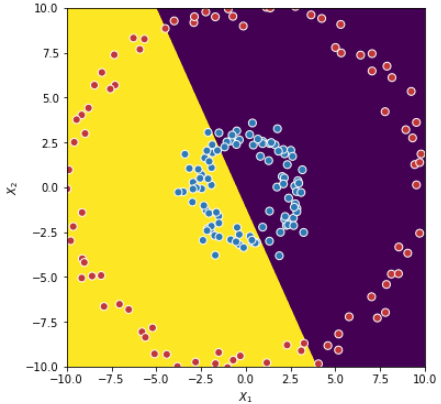
Based on biological neurons. Artificial neurons learn by recognizing patterns in data.



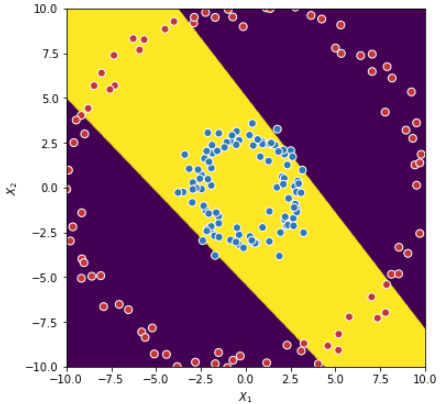
Deep Learning = Training Artificial Neural Networks

Example solving a non linear classification problem

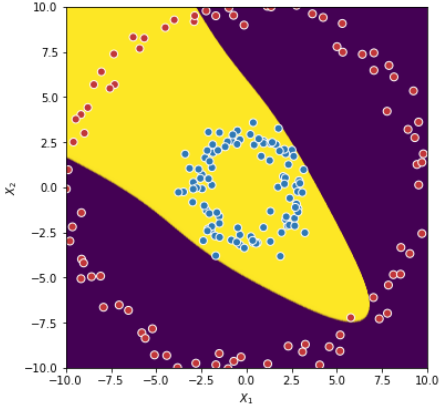
boundary formation with 1 layer and non linear “relu” activation function



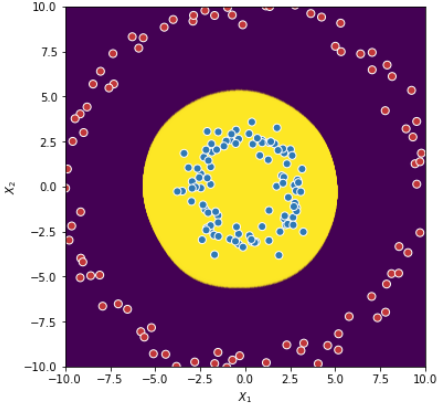
mlp = MLPClassifier(hidden_layer_sizes=(1),
max_iter=500000,activation=’**identity**’,learning_rate_init=0.01)



1 layer and 2 neurons

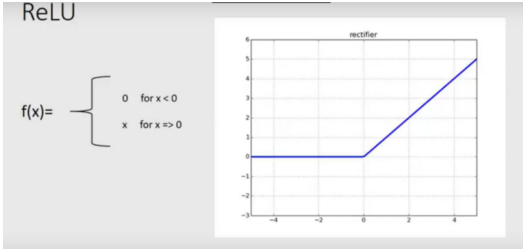


1 layer and 3 neurons



Decision plane with 30 neurons

Rectified Linear Unit (ReLu)

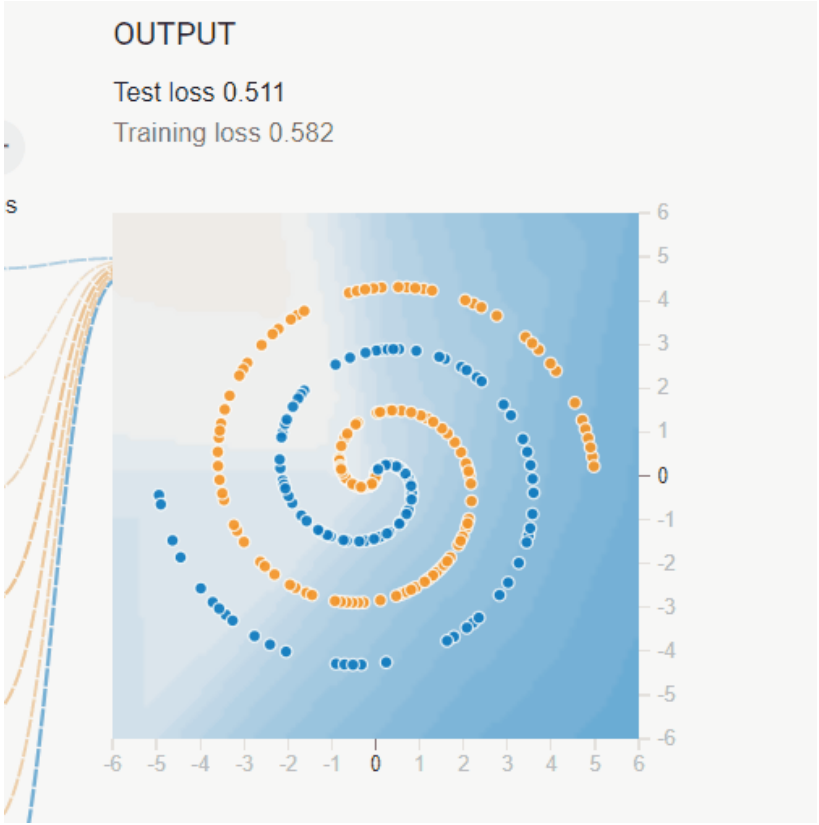


[Neural Nets Simulator:
http://playground.tensorflow.org](http://playground.tensorflow.org)

Deep Learning = Training Artificial Neural Networks

Example solving a non linear classification problem

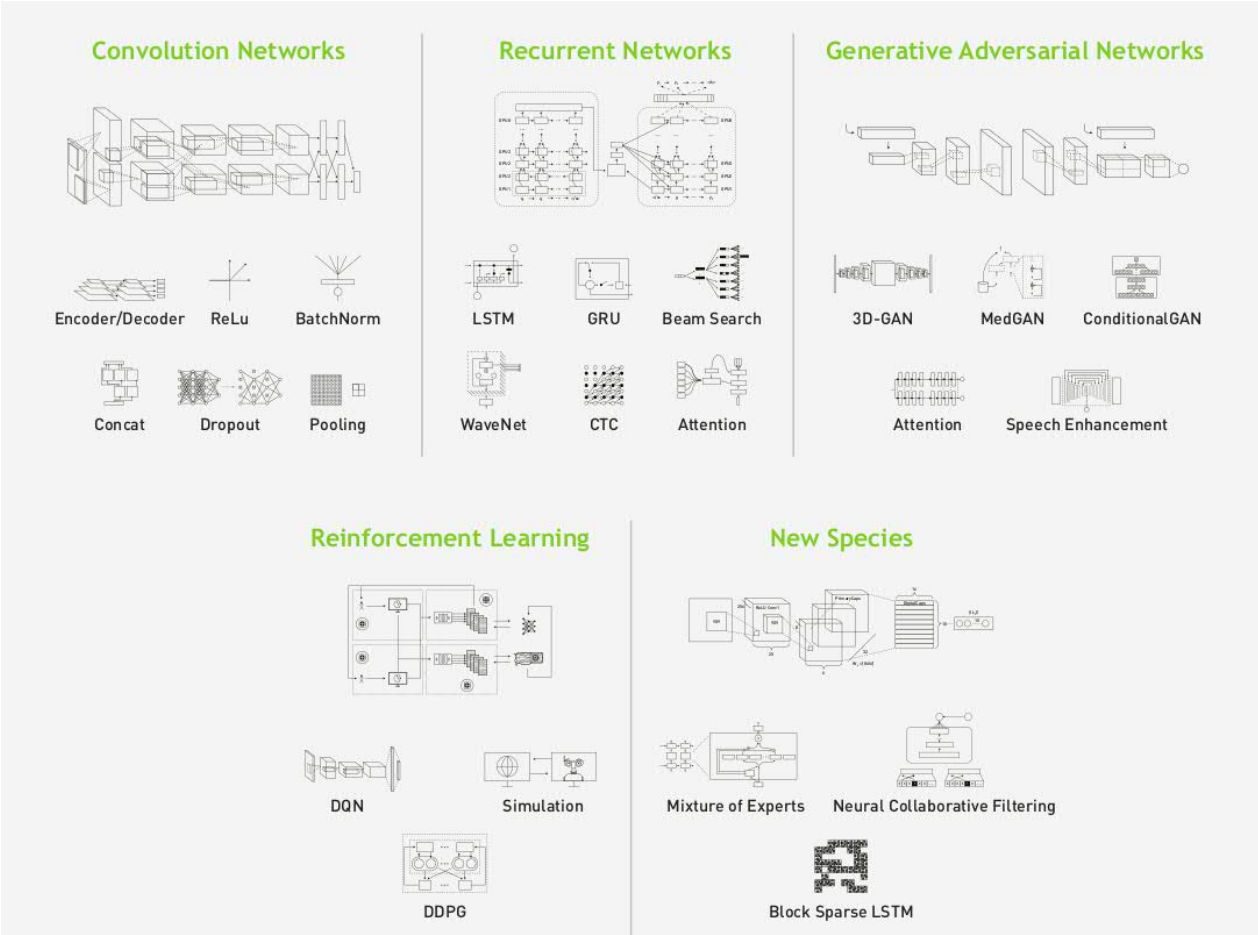
boundary formation with 3 layers and ‘relu’ activation



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

Deep Learning = Training Artificial Neural Networks

A few NN architectures



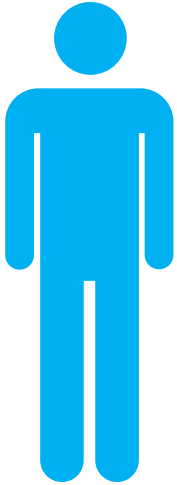
Le Machine learning répond aux besoins métier

Humans

2011

26% Errors

Machine Learning Based



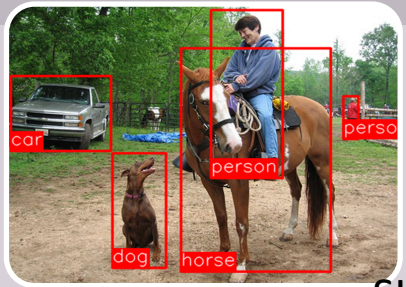
5% Error

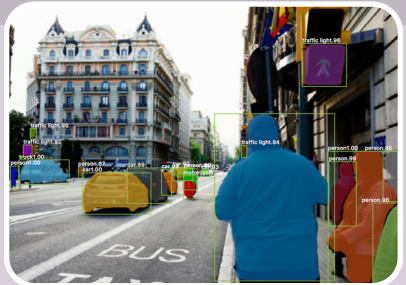
2016

3% Errors

Deep Learning Based

CNN Models detect more and more details in images/videos





SI68399

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% {x1,y1,x2,y2}

- Ex :Faster-RCNN, YOLO v2

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

Power Week

Université IBM i

22 et 23 mai 2019

Solutions : IBM «One AI»

Scénario 1: ML sur IBM i w/ Technologies Open Source
Scénario 2: AI dans le Cloud avec Watson / Data Platform
Scénario 3: AI On Premises accéléré & Private AI Cloud avec IBM PowerAI / WML-CE / WML-A

IBM vous supporte dans vos choix, dans votre Datacenter ou dans le Cloud public avec UNE solution cohérente.

IBM PowerAI

IBM

IBM i, Machine Learning & Solutions IBM

Quelques questions à se poser

Machine Learning use cases & relevance for business ?

- Business Analyst & Data Expertise
- Model Precision / Accuracy
- Model evaluation, monitoring & re-training (full lifecycle)

Deployment Options ? (vs. regulation, cost, performance, skills, other technical aspects)

- AI in the Public Cloud
- AI in the Datacenter (on premise)
- AI on the edge

Data Science Phases ?

- Data Preparation & Model building (which data sources, language & EDI, framework, libraries, algorithm?)
- Model training & validation
- Model evaluation, deployment for Inference (REST API? Batch? Streaming?)

Performance / Cost / Time to market ?

- Accelerated ML (GPU, FPGA,...) or not for model training? Model Inference?

Regulation & Auditability vs. model understanding ?

- Model Fairness
- Model Interpretability
- Model monitoring

IBM i, Machine Learning & Solutions IBM

Solutions Cloud ou dans le datacenter

2

PaaS/SaaS model, Pay per use

Great functionalities, quickly available, CPU and GPU

IBM Watson

Watson Studio + WML + Data Platform

IBM Cloud

1

ML Capabilities on IBM i

Simple, Efficient

No GPU, runs on CPUs

for Business

Machine Learning Libraries

Advanced users (i.e. Cluster with 4+ PowerAI)

- **WML-A for Hyperparameter optimization**
- **Unique Linear Scalability : Distributed DL (DDL)**
- **State of the art EGO job Scheduler from HPC**

3

Plugin – Training & Inference

WML-A

AC922

python

PowerAI Vision

H_2O.ai

Watson Studio +WML

IBM Watson

PowerAI Base (WML-CE) – optional ICP K8s Deployment

docker

kubernetes

AC922

AC922

Accelerated ++

Divide Training time by x vs. x86

DL: Large Model Support & NVLink

ML: Accelerated ML with SnapML and Nvidia Rapids

Free PowerAI Supported & Optimized Frameworks –

Simple Docker or conda based deployment. Optional K8s/ICP

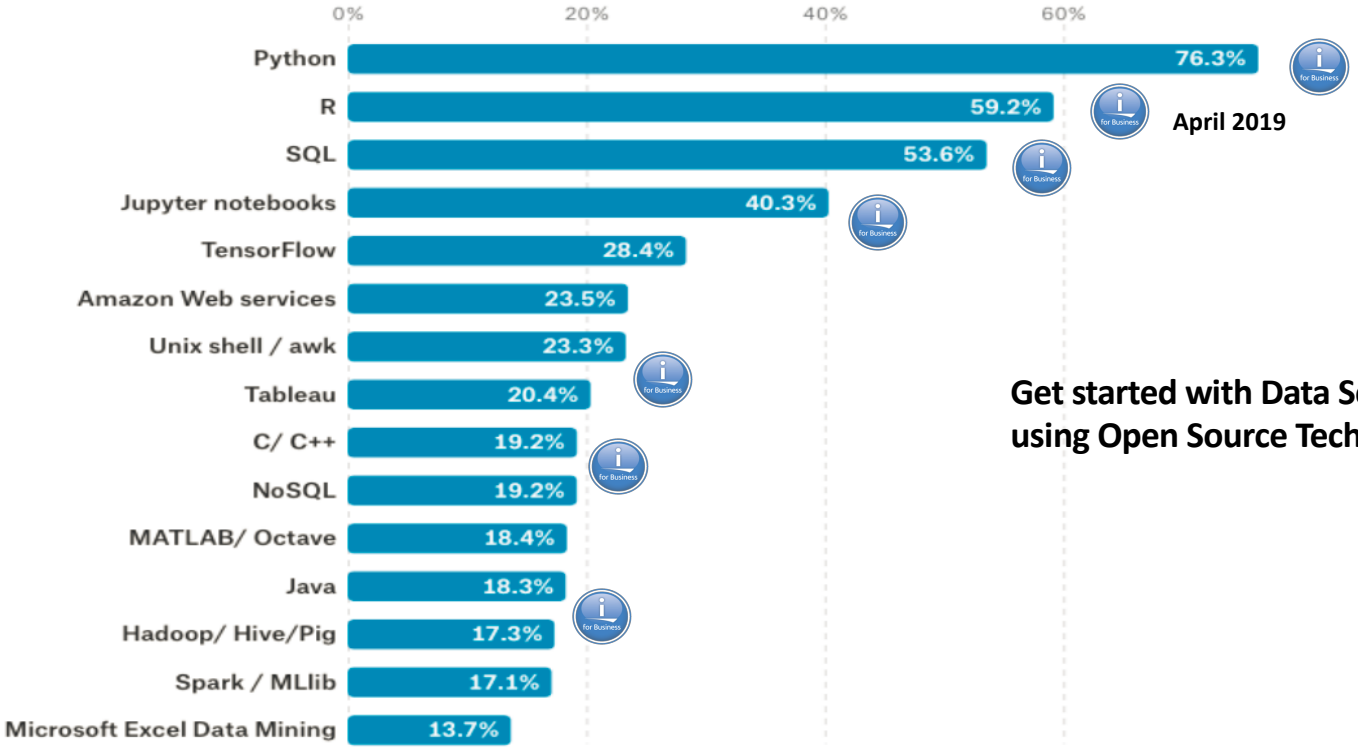
On top of PowerAI Base (WML-CE):

Watson Studio, H2O Driverless AI, PowerAI Vision…

Get Results in minutes

Data Science tools & technologies

Kaggle 2017 Data Science Tools Survey



April 2019

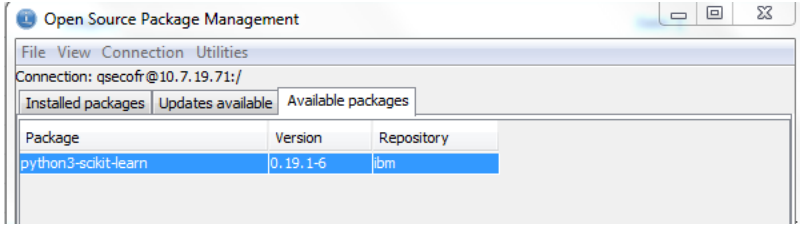
Get started with Data Science on IBM i using Open Source Technologies

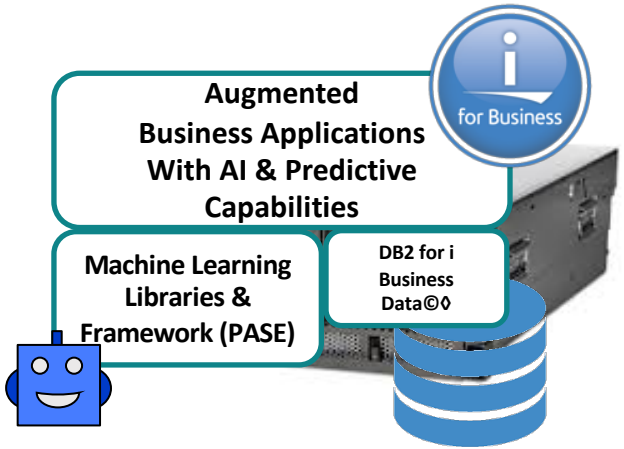
IBM ML Technologies & IBM i

Scenario 1: Utiliser les frameworks et langages disponible sur IBM i 7.2+

Data & Scientific Packages Available

- Numpy, Pandas : Data Processing
- **Scipy, Scikit-Learn**
- IPython : interactive Python
- NLTK : Natural Language Processing
- Matplotlib, jupyter : Data Visualization
- R Language (Interpreter, Runtime)
- More to come? ☺



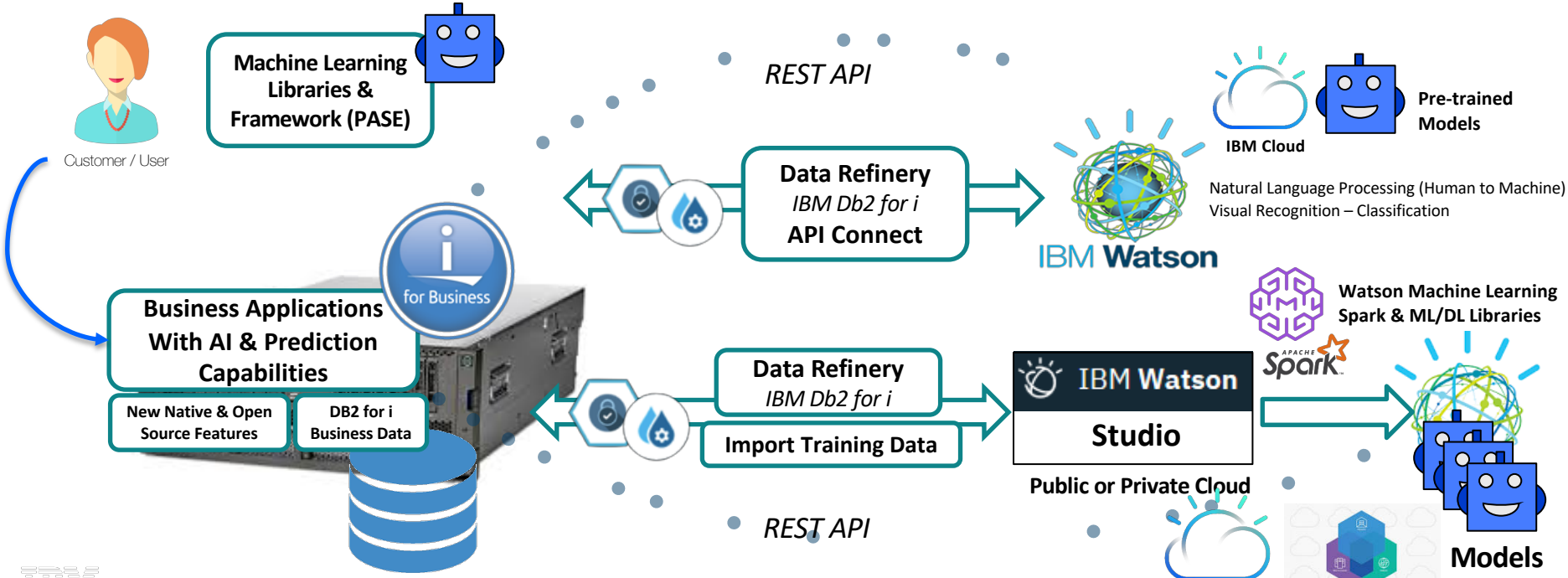


Alternative vs. dataset size, investment on ML etc.

- Training & Inference on IBM i
- Training on PowerAI / WML , Inference on IBM i (<->) (Public or Private Cloud)
- Data preparation on IBM i, Training & Inference on Accelerated Servers (PowerAI, WML-A)

IBM i & Artificial Intelligence

Scenario 2: Utiliser Watson Developer Cloud (API) , IBM Watson Studio (Cloud ou Local)



Demo: https://ibm.box.com/v/ibmi-opensource-watsonstudio

- Data is the key in all AI projects: your business data resides on IBM i
- Use pre-trained & customizable models with IBM Watson (Developer Cloud) services in IBM Cloud
- Build your own use case & business specifics models with IBM Watson Studio - IBM Cloud / on premises (Studio Local w/ Cloud Private)

Power Week

Université IBM i

22 et 23 mai 2019

Solutions : IBM «One AI»

Scénario 3: AI On Premises accéléré & Private AI Cloud avec IBM PowerAI / WML-CE / WML-A

IBM vous supporte dans vos choix, dans votre Datacenter ou dans le Cloud public avec UNE solution cohérente.

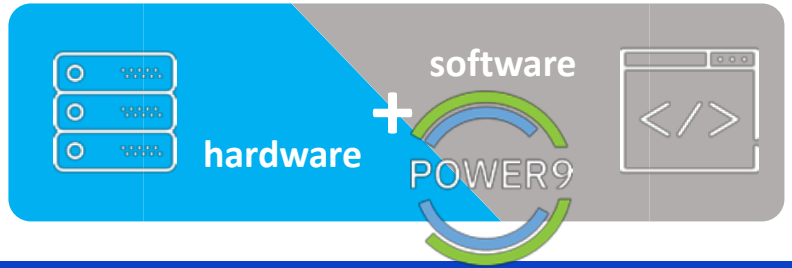
IBM PowerAI

IBM

IBM i & Artificial Intelligence

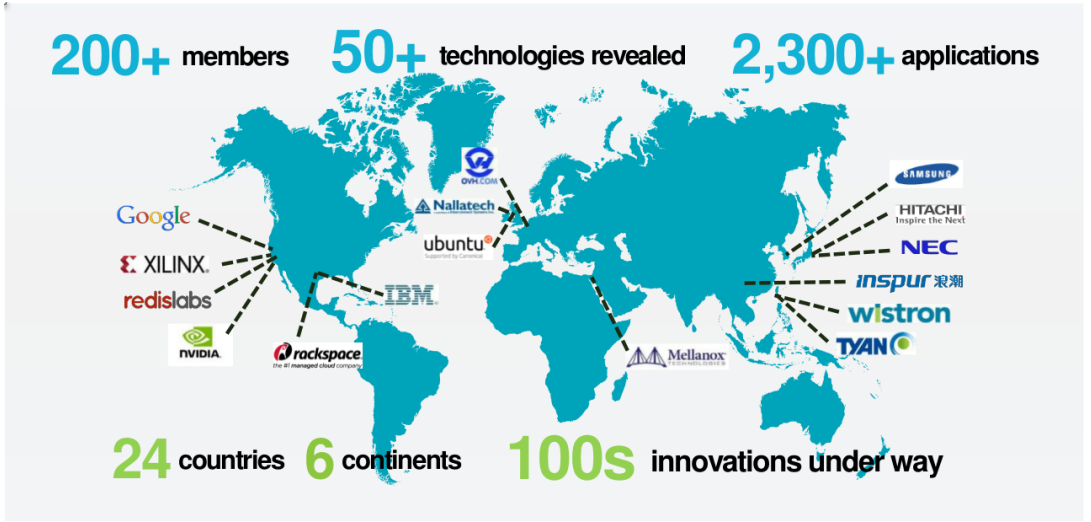
Scenario 3: Utiliser l'état de l'art du Private Accelerated ML/DL avec PowerAI / WML-A, s'intégrant automatiquement avec des solutions comme Watson Studio Local, PowerAI Vision, Driverless AI.

- L'objectif est de ne plus payer à l'usage, d'être autonome, et d'avoir une technologie pérenne et performante
  - Prix compétitifs (vs. TCA/TCO, vs. offres Cloud ou « on premise » concurrentes)
  - Obtenez la précision nécessaire dans vos modèles
- Solution Matérielle et logicielle Open source, supportée par IBM.
  - Solution designée pour le Machine Learning et Deep Learning accéléré. Base du supercalculateur CORAL Summit
  - Disponible « on premise » , dans le Cloud public.
- Facilement Cloudifiable en Cloud Privé via Kubernetes / IBM Cloud Private (Hébergeurs, Consolidation des environnements AI)



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

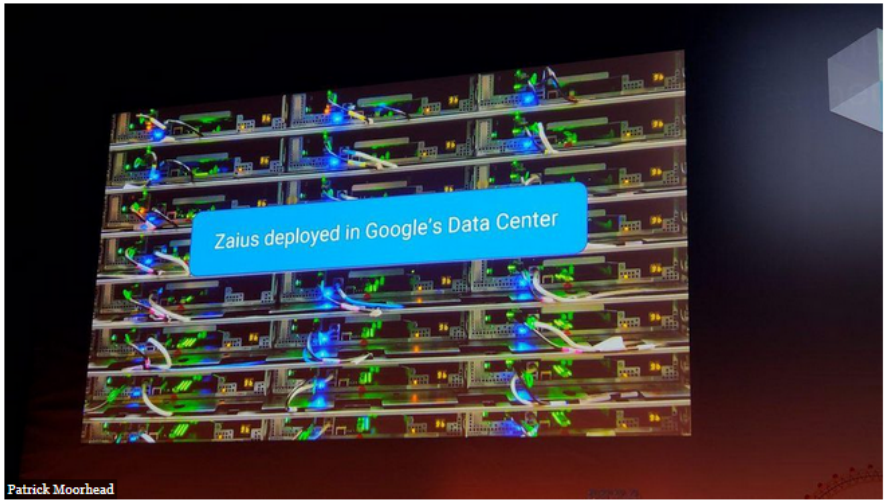
169 OpenPOWER Ready Certified Products (Nearly **tripled** since inauguration at 2016 Summit)

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

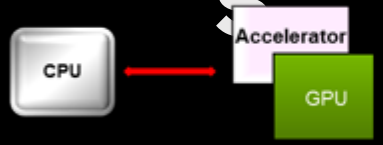
POWER9

Broader Application of Heterogeneous Compute

designed for efficient programming models
accelerate complex AI & analytic apps

"vanilla"

Other



- PCIe Gen3

extreme CPU and Accelerator bandwidth

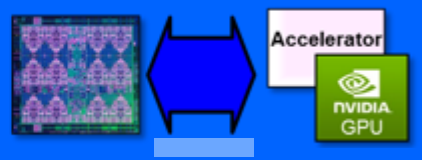
2x

- PCIe Gen4

5x

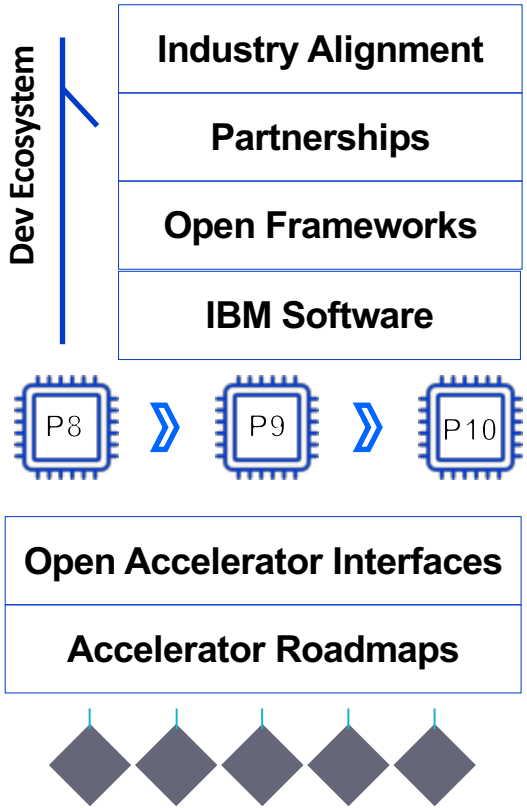
POWER8
with NVLink 1.0

7-10x



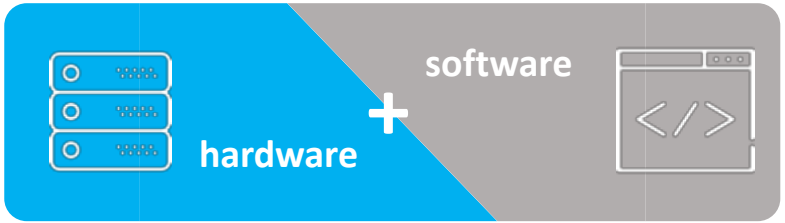
POWER9 with 25G
Link + NVLink 2.0

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

500+ POWERAI CLIENTS WITHIN ONE YEAR

40% NEW TO POWER

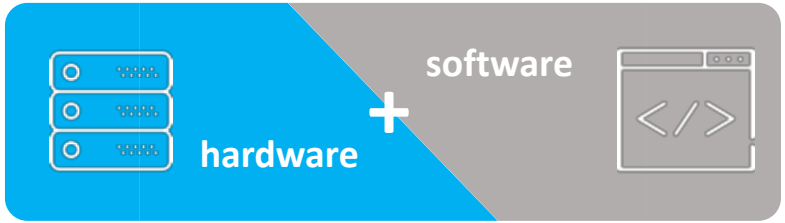
(Free) PowerAI 1.6.1 (WML-CE) Overview

5765-PAI - Linux Native install (conda) 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

Announcement Letter

PowerAI Vision

Auto-ML for Images & Video

Label | Train | Deploy

H2O.ai

Watson Studio

Evolving with IBM One AI Strategy

PowerAI (WML-CE)

PowerAI: Open Source Frameworks

TensorFlow™ PYTORCH Chainer SnapML

Large Model Support (LMS)

Distributed Deep Learning (DDL)

Auto ML (future)

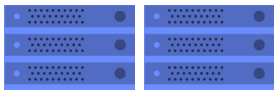
IBM Spectrum Conductor
Cluster Virtualization, Elastic Training
Auto Hyper-Parameter Optimization

Deep Learning Impact (DLI) Module

Data & Model Management, ETL, Visualize, Advise

PowerAI Enterprise (WML-A)

Accelerated Infrastructure

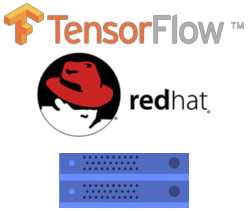


Accelerated Servers AC922

Storage (Spectrum Scale ESS)

PowerAI: Enterprise AI Platform

Simplicity: Integrated Platform that Just Works

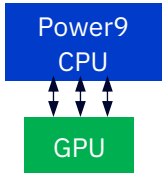


Curate, Test, and Support Fast Moving Open Source

Provide Enterprise Distribution on RedHat

Easy to deploy Enterprise AI Platform

Ease of Use, Unique Capabilities



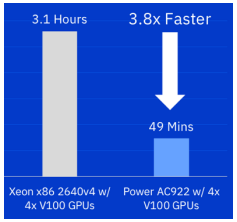
Large data & model support due to NVLink

Acceleration of Analytics & ML

AutoML: PowerAI Vision

Elastic Training: Scale GPUs as Required

Faster Model Training Time

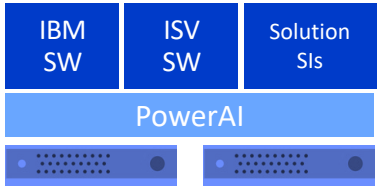


Faster Training Times in Single Server

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

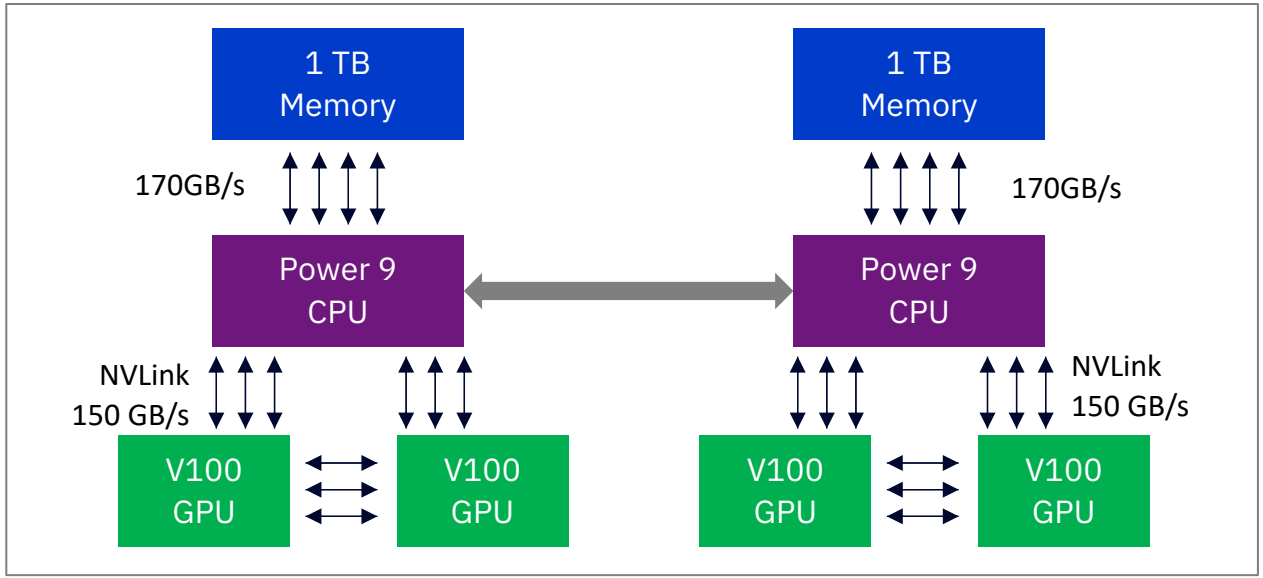
SIs, Solution Vendors & Accelerator Partners

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

Store Large Models in System Memory

Fast Transfer via NVLink

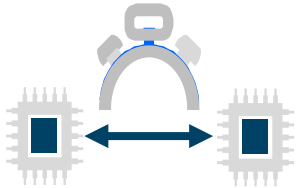
Operate on One Layer at a Time



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

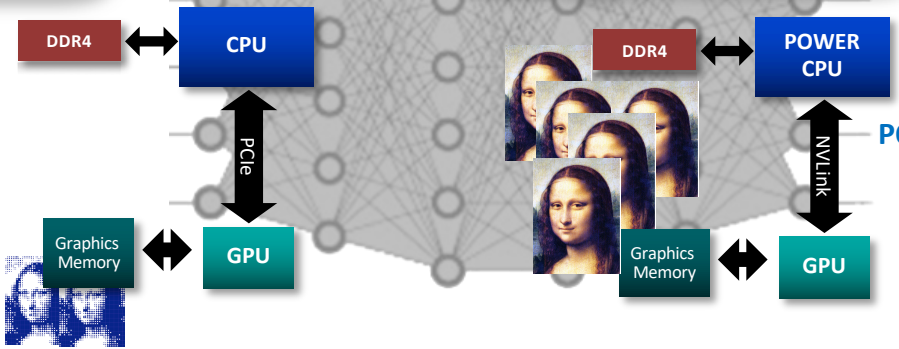
faster training times for data scientists

Distributed Deep Learning

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

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

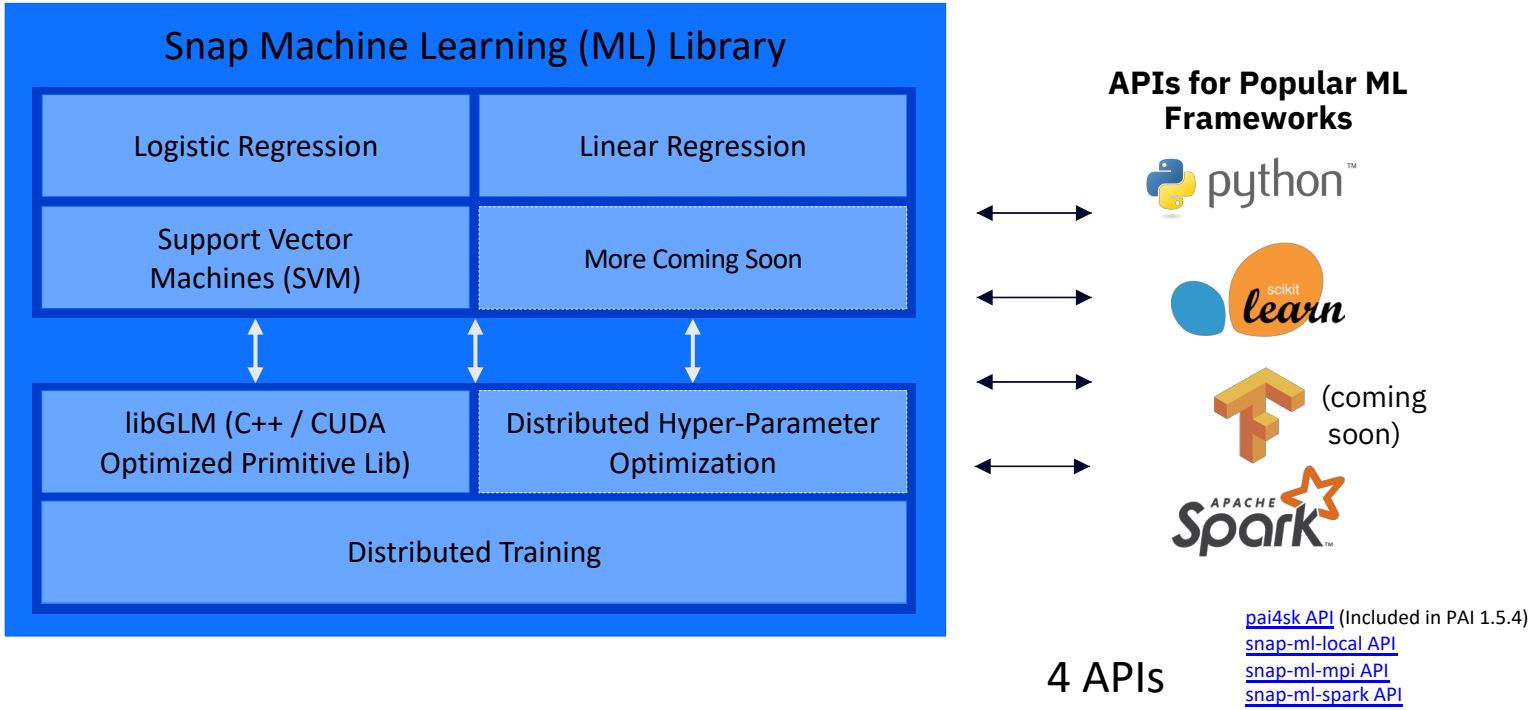
(PowerAI)
Use system memory and GPU to support more complex models and higher resolution data



POWER NVLink Data Pipe

Snap ML

Distributed GPU-Accelerated Machine Learning Library



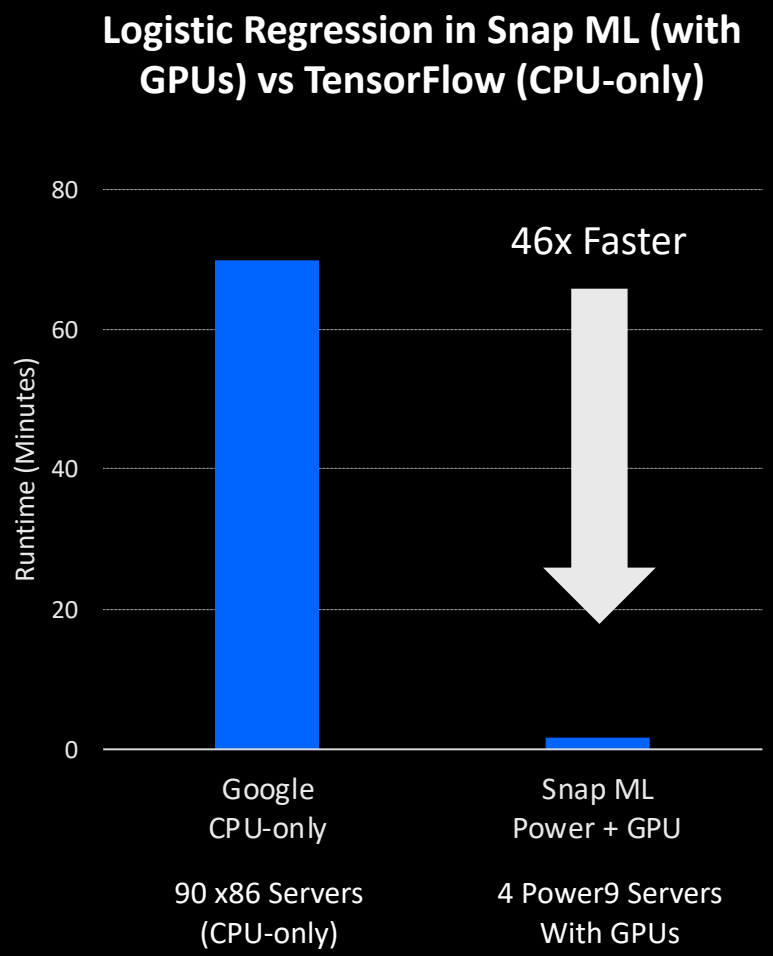
- Snap ML: Training Time Goes From An Hour to Minutes

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

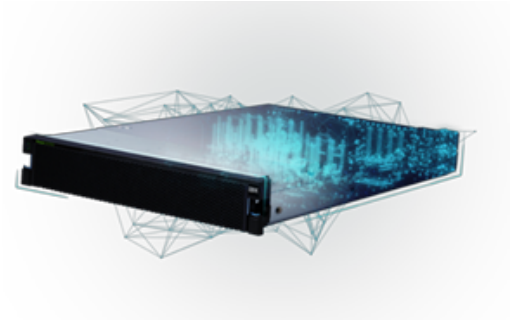


En résumé: WML-CE/ PowerAI Top 4 Features

Accelerated ML (H2O, Snap ML)	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

Architecture de Référence – Private AI

Expérimentation « Single Tenant »
AC922 – Disques internes



PowerAI

Cuda Drivers – OSS Frameworks

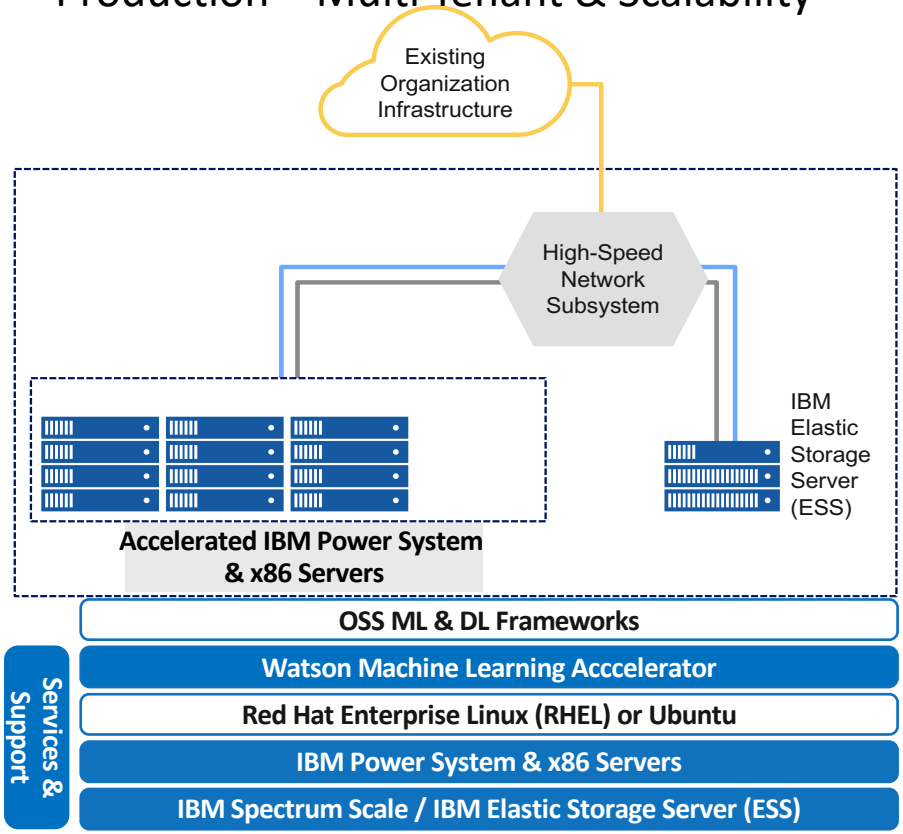
Red Hat Enterprise Linux (RHEL) or Ubuntu

Internal SAS drives & NVM's

POWER Servers with GPU's

InfiniBand EDR P2P connection

Production – Multi Tenant & Scalability



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

User1 – Web Browser

PowerAI Vision

IBM Data Science Experience

H2O.ai

User2

IBM PowerAI

User3 – Web Browser

IBM PowerAI

jupyter

IBM Cloud Private

Catalogue

H2O.ai

IBM Data Science Experience with GPU

GPU as a Service On demand

GPU as a Service Dedicated

PowerAI Vision

Kubernetes

Worker Node: Power AI

Deep Learning Framework

Supporting Libraries

GPU GPU GPU GPU

Worker Node: Power AI

Deep Learning Framework

Supporting Libraries

GPU GPU GPU GPU

X86 and VMWare

Master Node

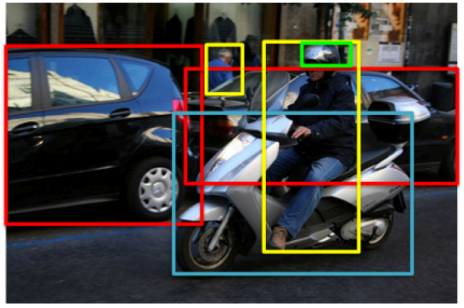
Worker Node

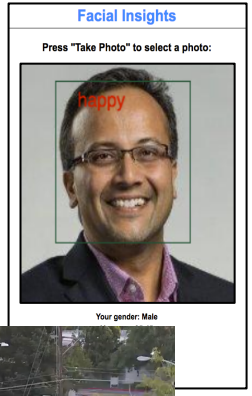
Comparing AI Offerings on Power

		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

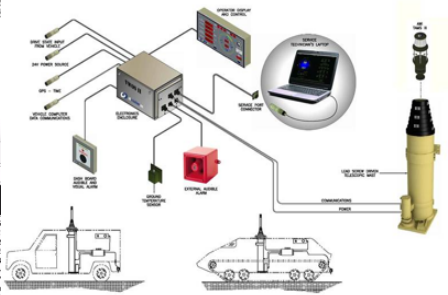
Transactional

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
s2136305133	11/23/2010	7:15:13 PM	43.16047	**********9208
t7478724264	12/2/2010	12:08:46 PM	40.14018	**********2695
u9844403319	10/11/2010	3:10:19 AM	23.70907	**********2980

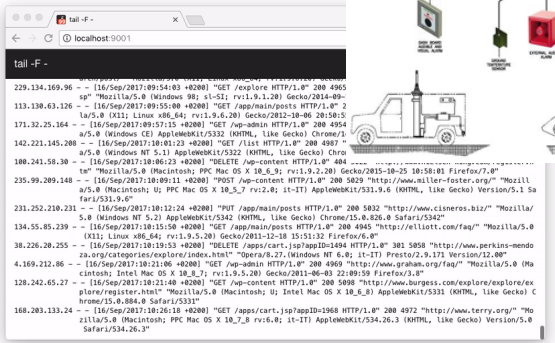
Product_Code	Amount
G324516281	2.32
S820571874	1.45
R922779511	1.89
T835381161	2.21
X361194534	1.14
A229041651	3.89
B217229145	1.45
TOTAL	14.47

Example: Flat File

Sensors

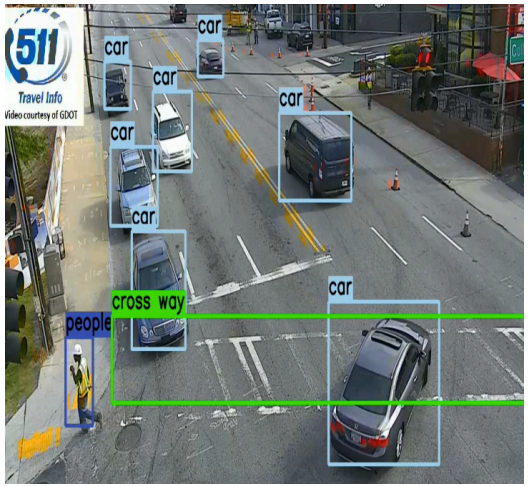


Log

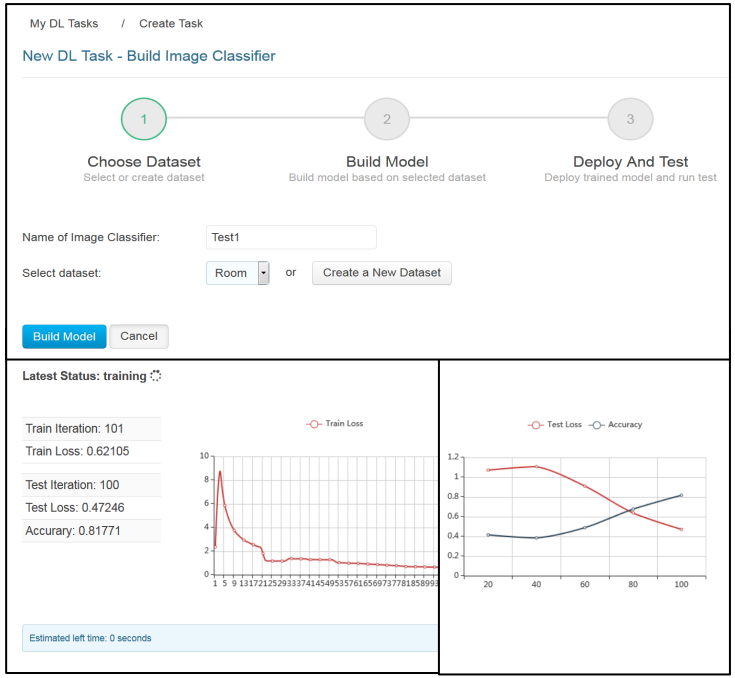


PowerAI Vision: ”Point-and-Click” AI for Images & Video

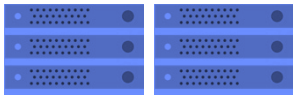
Label Image or Video Data



Auto-Train AI Model

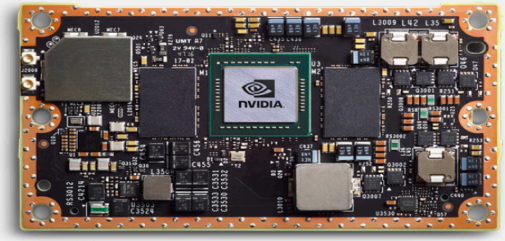


Package & Deploy AI Model



Accelerated embedded edge devices

NVIDIA Jetson TX2 Module

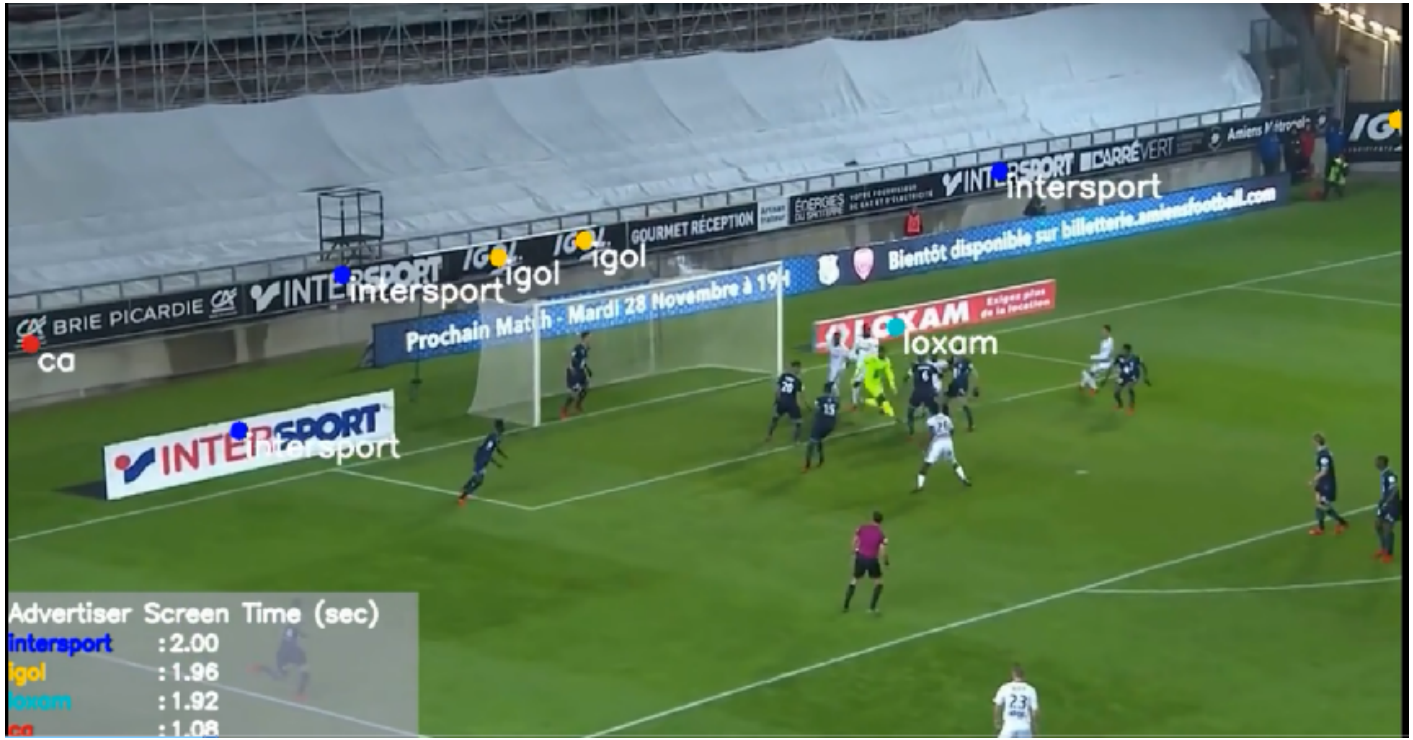


Jetson TX2 is the fastest, most power-efficient embedded AI computing device. This 7.5-watt supercomputer on a module brings true AI computing at the edge. It's built around an NVIDIA Pascal™-family GPU and loaded with 8GB of memory and 59.7GB/s of memory bandwidth. It features a variety of standard 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

Made with PowerAI Vision – Drone Inspection

Made with PowerAI Vision – Sport / Advertising



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

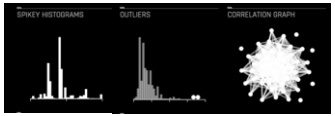
Snowflake

Google BigQuery

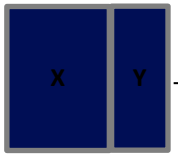
Azure Blog Storage

2 Automatic Visualization

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



Modelling



Dataset

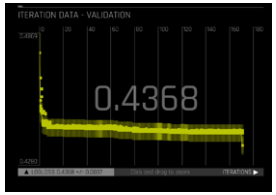
3 Automatic Machine Learning

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**

Powered by GPU Acceleration

4 Automatic Scoring Pipelines

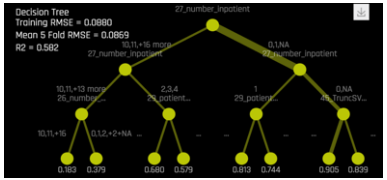
Deploy ultra-low latency Python or Java Automatic Scoring Pipelines that include feature transformations and models.

Model Documentation

Automatic Scoring Pipeline

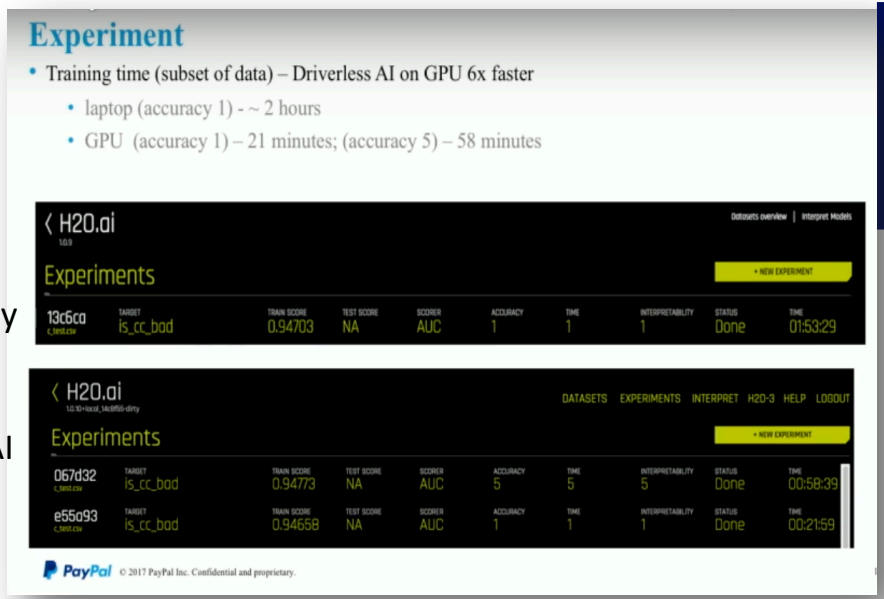
Deploy Low-latency Scoring to Production

Machine learning Interpretability



Exemple: Détection de fraude

- 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.

Comment bien démarrer: Contactez-nous!

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 & Jupyter notebook

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/

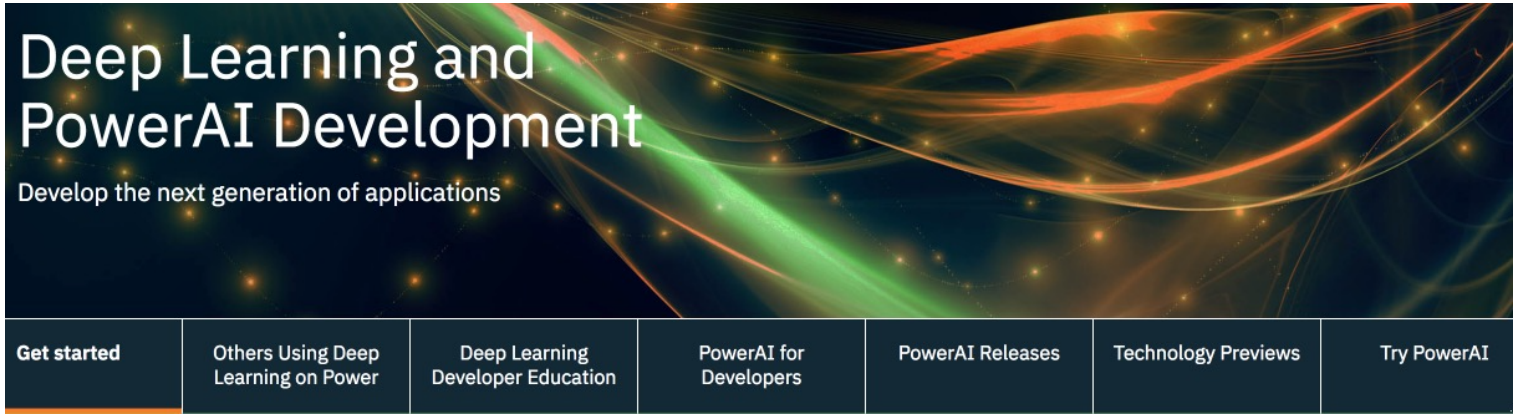
PowerAI Vision 1.1.1 **Free trial**
Register for a free 3-day trial of PowerAI Vision

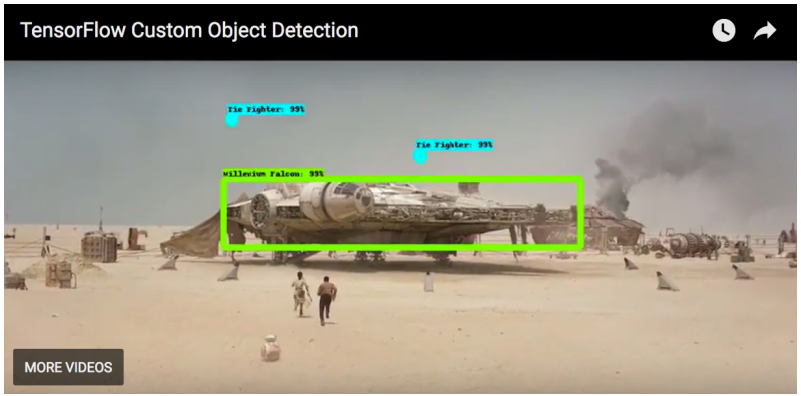
H2o.ai **Driverless AI** (Trial 21 days)
https://www.h2o.ai/products/h2o-driverless-ai/

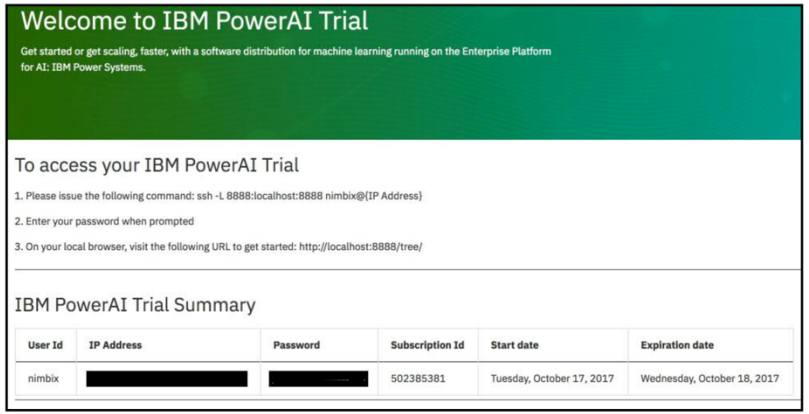
Presentations, Demo Replays : https://ibm.biz/bma-wiki

Want to know more? Need support ?
Montpellier team & AI Environment for your PoC / Tests: Driverless/PowerAI/AI Vision Remote Access
Contact us : a2roy@fr.ibm.com / benoit.marolleau@fr.ibm.com

ibm.biz/poweraidevelopper







Montpellier Cognitive Systems Lab

Manager : Philippe Chonavel

Coordinator : Alain Roy

Team PowerAI ATS Europe

Jean-Armand Broyelle

Regis Cely

Sebastien Chabrolles

Sylvain Delabarre

Maxime Deloche

Benoit Marolleau

Team CAPI/SNAP (FPGA)

Bruno Mesnet

Alexandre Castelane

Fabrice Moyen

Team HPC

Ludovic Enault

Pascal Vezolle

Montpellier Client Center (France)

Our Offerings

Technical Consultancy & Assistance

Co-Creation Lab Workshops

Hands-On Technical Enablement

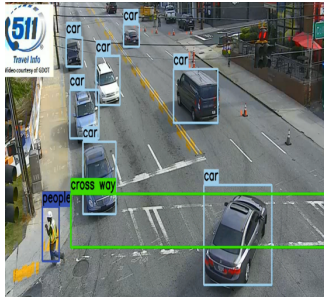
Demonstration/PoT

PoC/Benchmark


MERCI

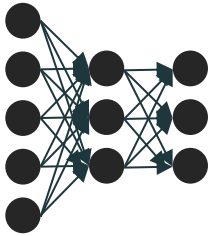
Semi-Automatic Labeling using PowerAI Vision

Manually Label



Define Labels Manually Label Some Images / Video Frames

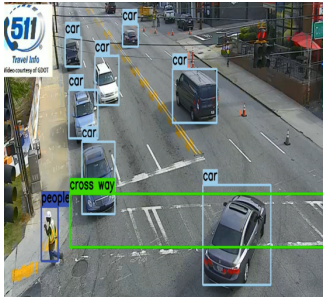
Train DL Model



Use Trained DL Model

Run Trained DL Model on Entire Input Data to Generate Labels

Correct Labels on Some Data

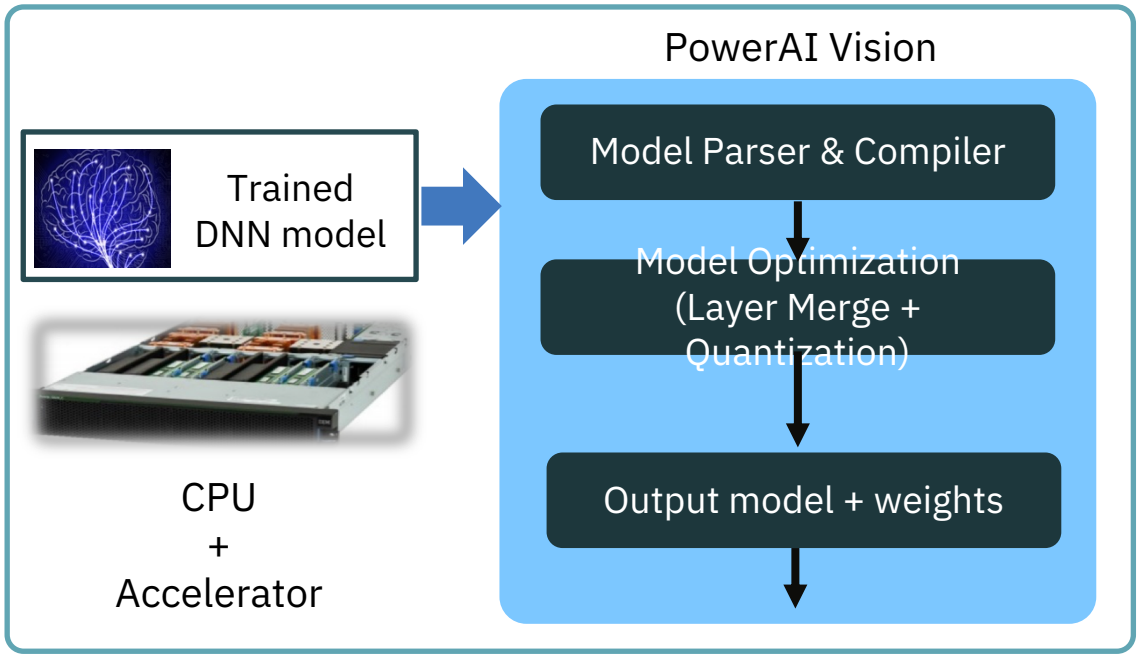


Manually Correct Labels on Some Data

Repeat Till Labels Achieve Desired Accuracy

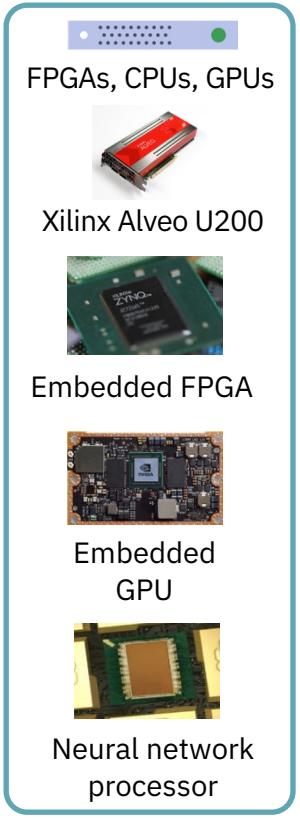
Deploying Trained Models

Data Center: Train model & Compile to Edge

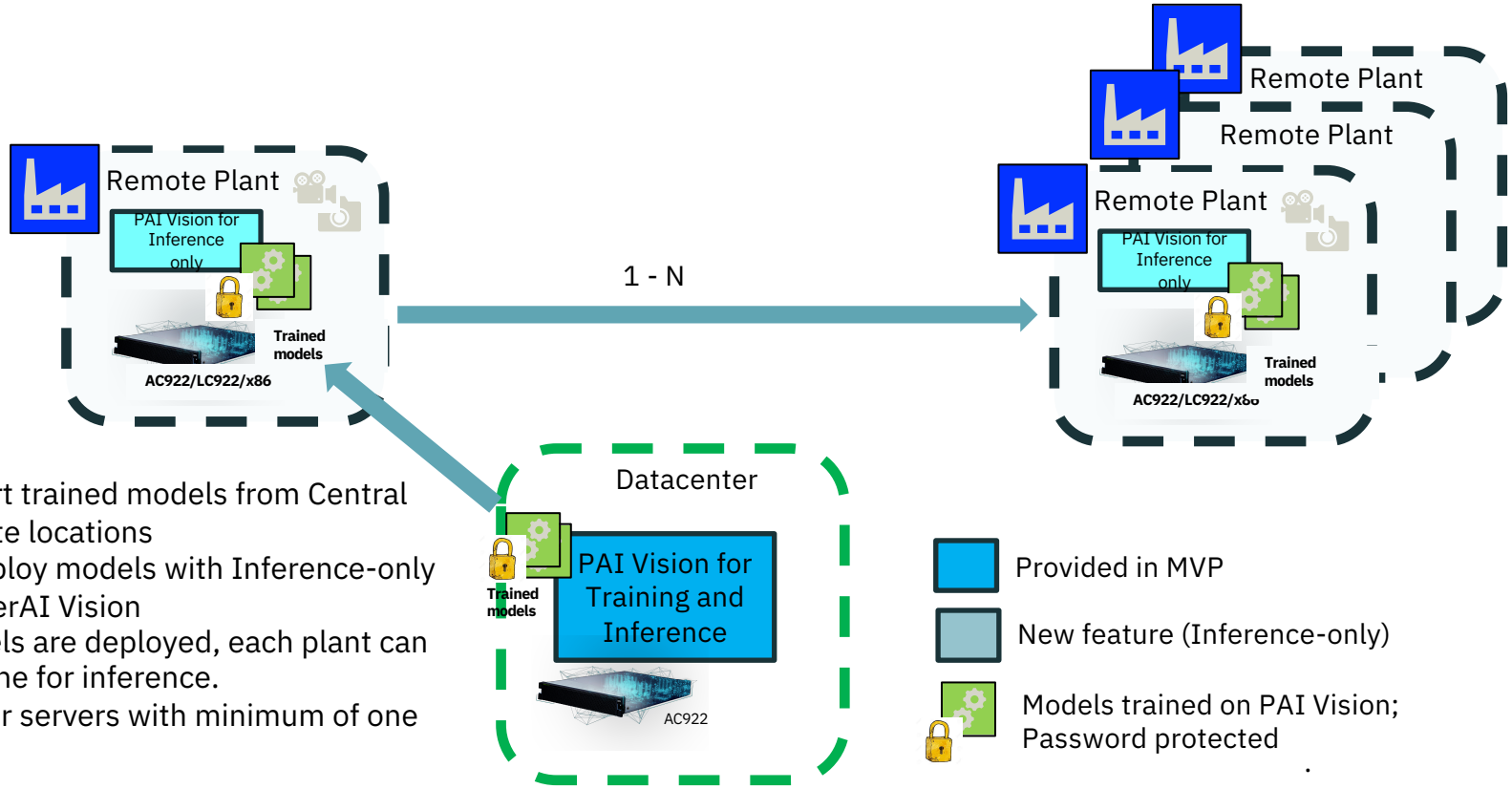


Map to Different Platforms

Cloud or Edge



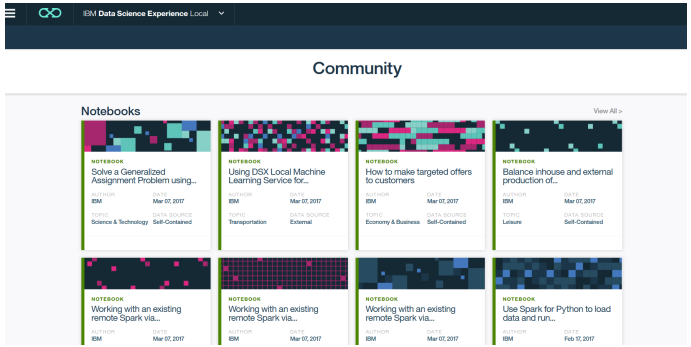
Train on central server but deploy on several remote servers

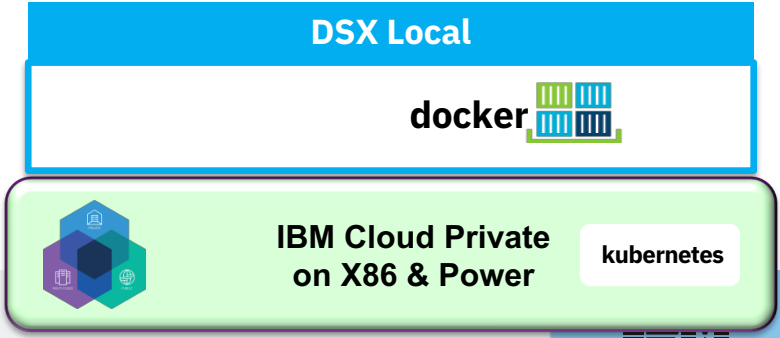


1. Manually export trained models from Central server to remote locations
2. Import and deploy models with Inference-only license of PowerAI Vision
3. Once the models are deployed, each plant can work stand alone for inference.
4. Supports Power servers with minimum of one GPU

Core Attributes of Watson Studio

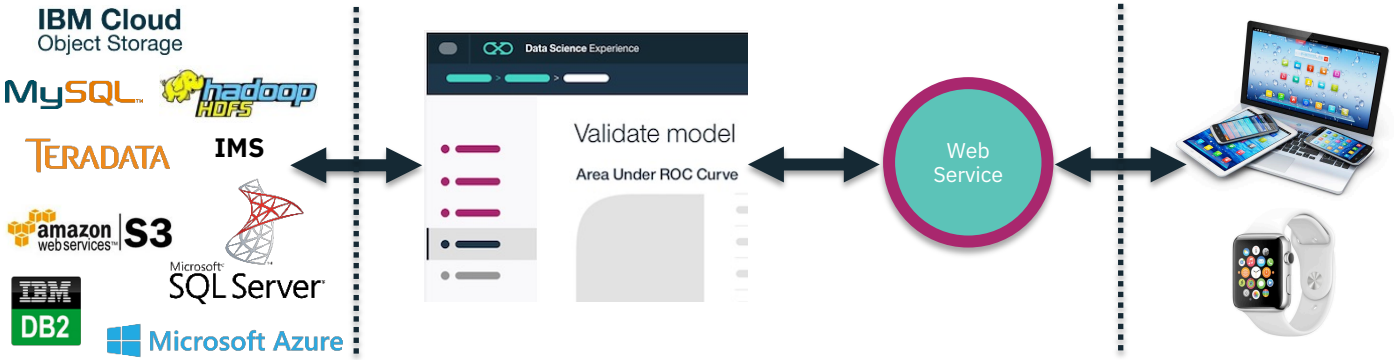
- IBM Watson Studio (aka DSX) is available
  - As a cloud offering aka **Watson Studio**
  - As a desktop application
    - Free, disconnected mode
  - As an on-premises solution
    - **DSX Local now Watson Studio Local** on x86/Power
    - Power: Scale-out LC Systems with PowerAI + GPU / Nvlink acceleration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

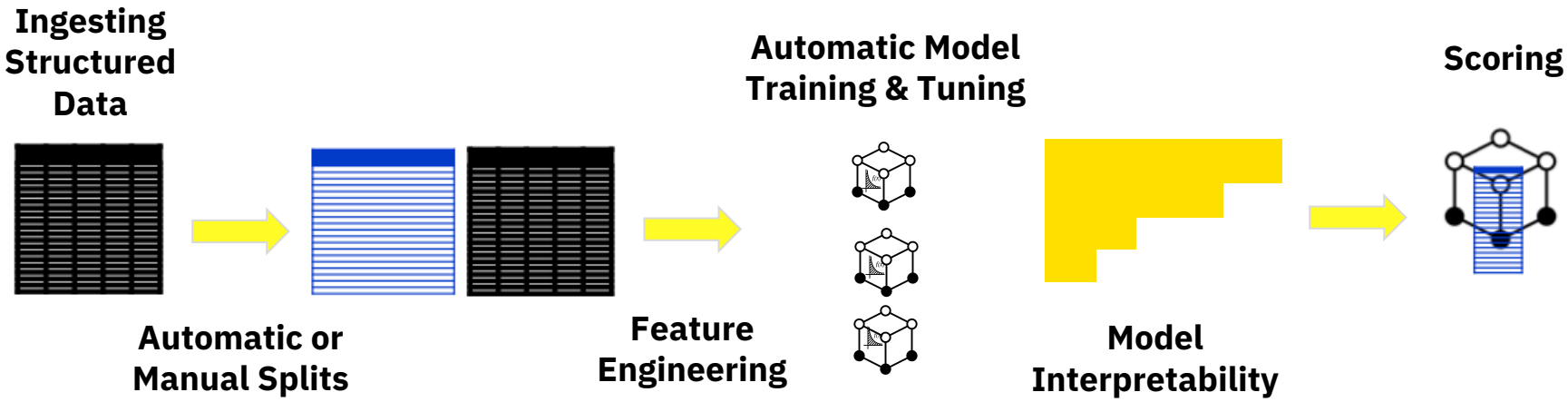
Web Service:

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

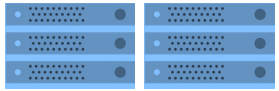
Intelligent Apps:

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

H2O Driverless AI: Automated Machine Learning



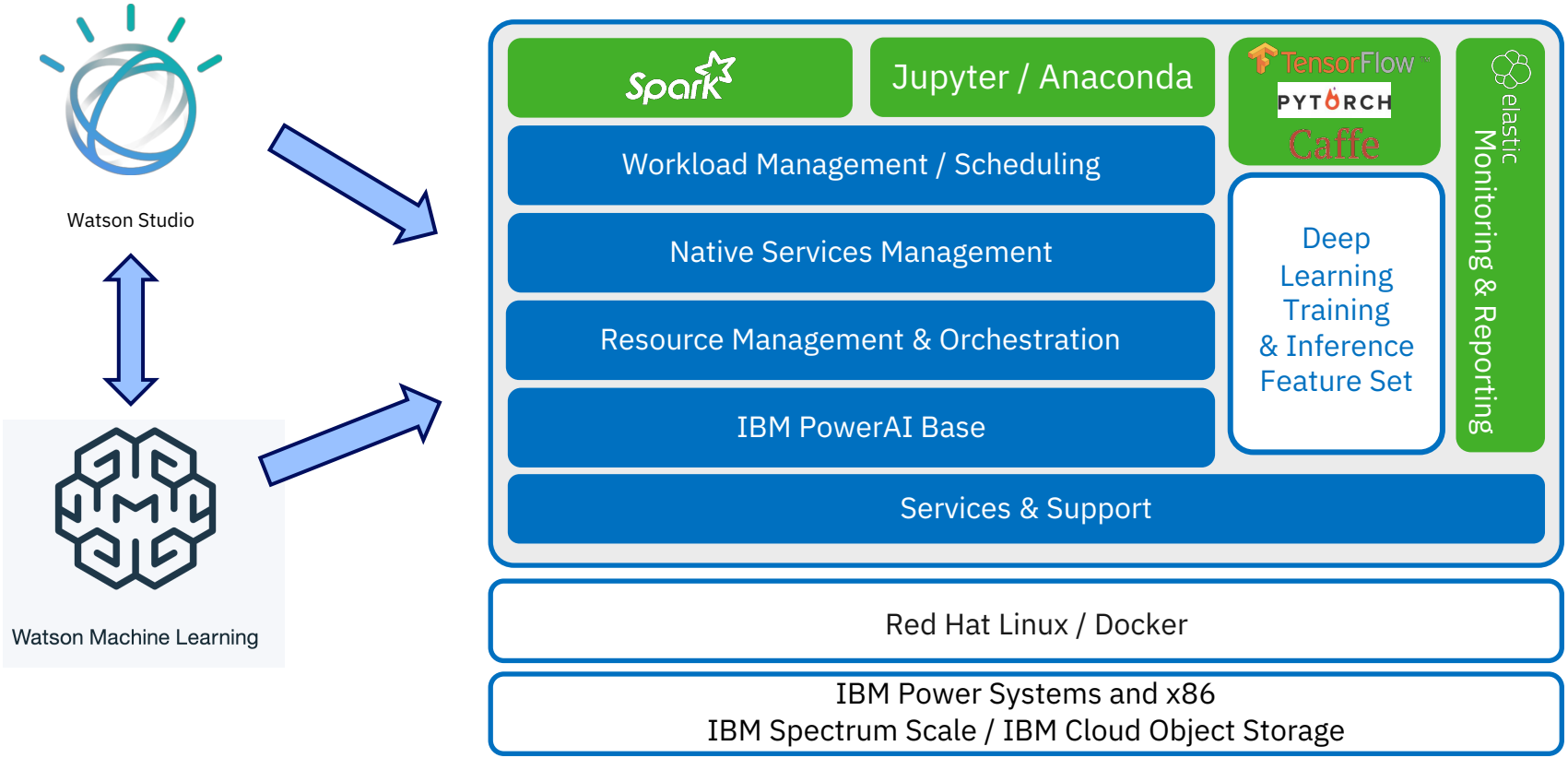
Optimized for GPU-Accelerated Power9 Servers



H2O.ai

IBM Watson ML Accelerator:

Power AI Base + Spectrum Conductor + Deep Learning Impact



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