Multiplying data science with IBM deep learning platforms - faster value creation from deep learning AI

”Data scientists don’t train a model - they train thousands!”

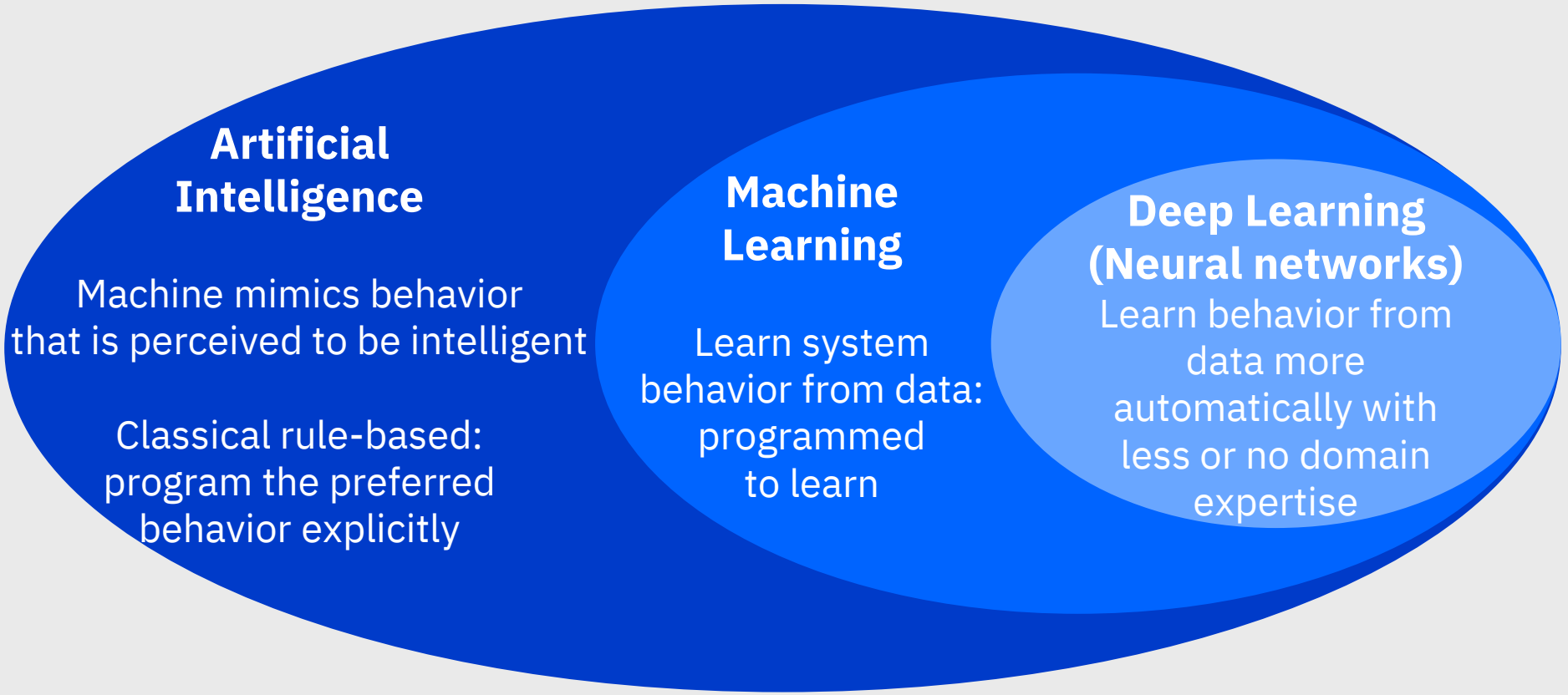
Deep learning development
and computation platforms:

From limited deep learning HW and SW setups
to *productized versatile solutions and scalability*

Jukka Remes, BDE, Dr (Tech.)

IBM

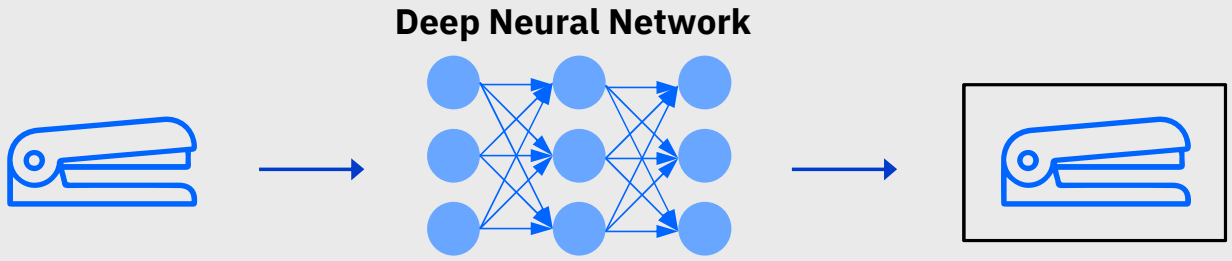
AI Landscape – tools for digital automation



Machine Learning

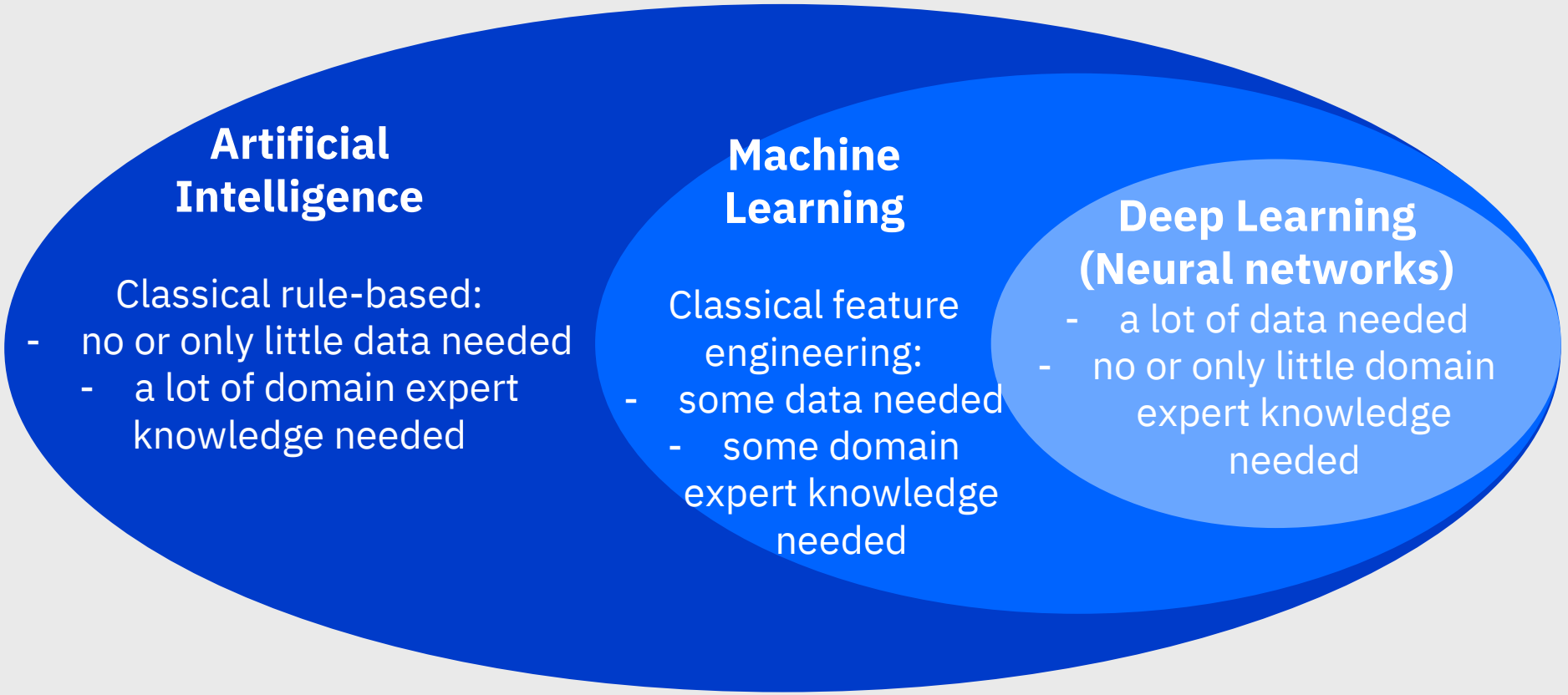
Input → Feature Extraction → Features → Classification → Output

Deep Learning

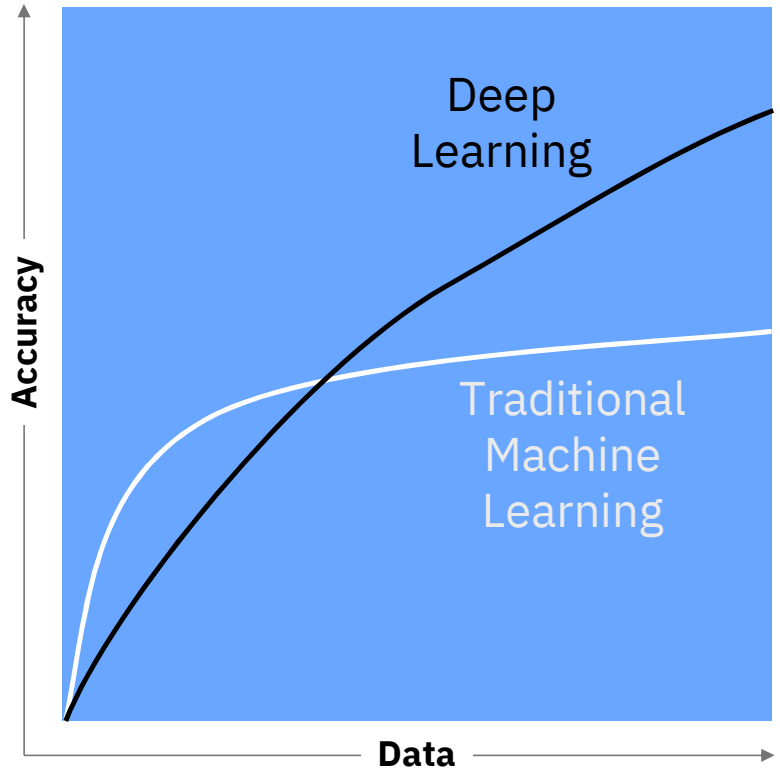


Input → Feature Extraction & Classification → Output

AI pre-requisites: *application domain expert* knowledge and data



Deep Learning Has Revolutionized Machine Learning



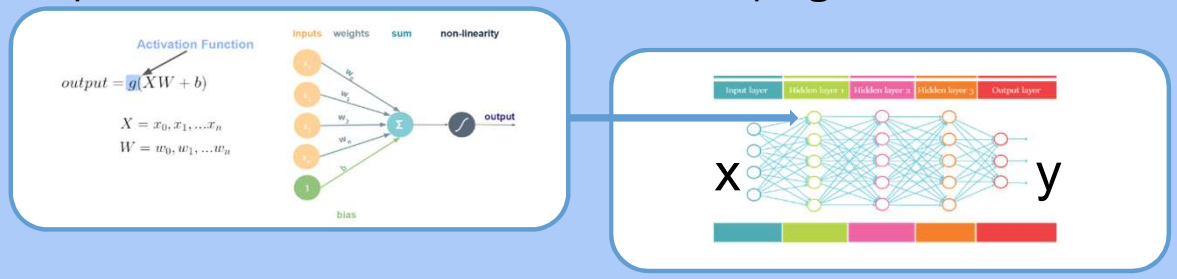
Deep Learning Popularity Growing Exponentially

Source: Google Trends. Search term “Deep Learning”

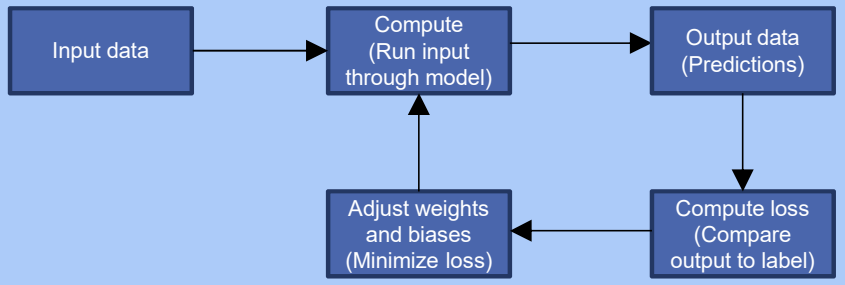
Deep learning starting point..

Typical target:
Task automation development through supervised learning by input-output mapping approximation available using neural nets

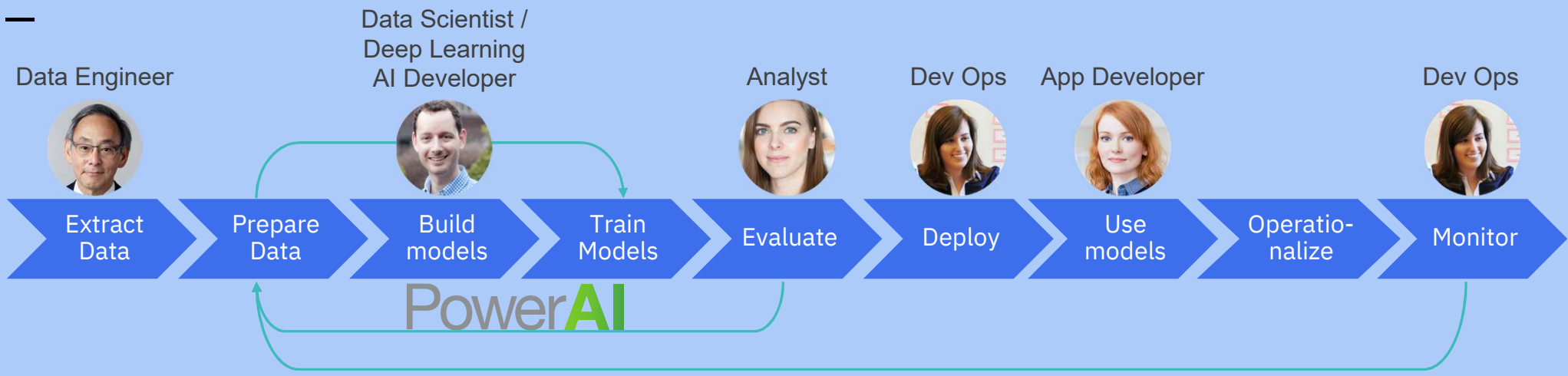
- X=medical image/signal/variables, Y=disease class/severity/prognosis
- X=customer feedback/preferences/.., Y=service action/campaign.., etc.



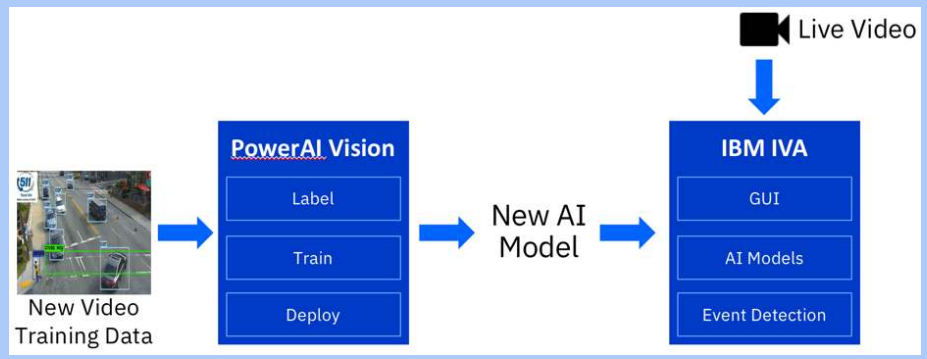
Calculation and data(/IO) intensive computing => GPU-training (thousands of cores) on high bandwidth systems



Systematic data science: deployment, integration and continuous AI development/model improvement



Interested parties need to be able to consume the trained model and the model needs to be supported



https://medium.com/@sumitg_16893/deep-insights-with-ai-for-video-analytics-5464fd30ebe1

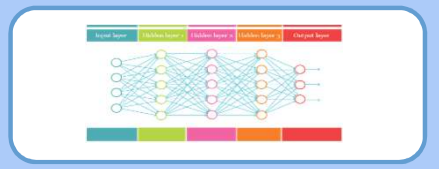
Machine learning and deep learning in AI development - dependencies

Subject matter expertise

Input data (e.g. images, audio and/or text data)

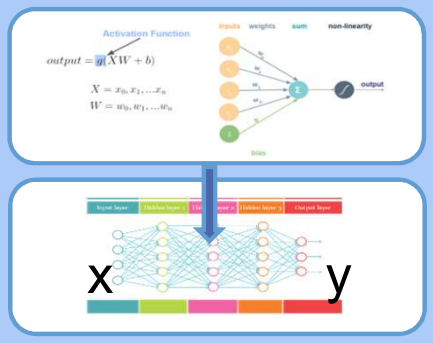
1. AI need / development problem

10. Trained model to be used as AI/part of AI

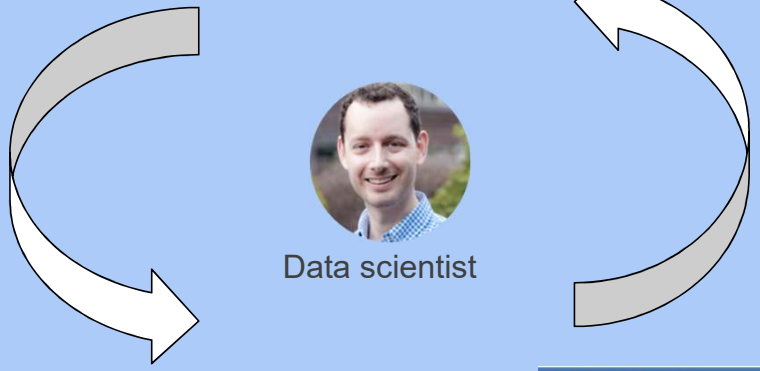


arXiv.org

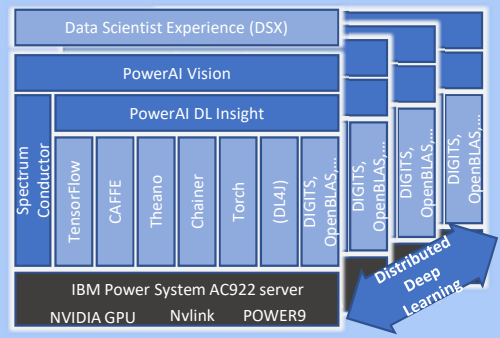
Research publications



Mathematical models (e.g. deep neural nets)



Data scientist



Computing environment

python™

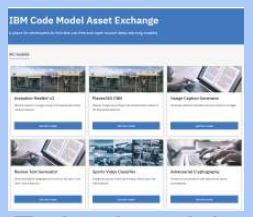
Programming language

TensorFlow

GitHub

Example software source code

Software libraries / machine learning frameworks



Trained models

jupyter

SPYDER

Development environments

IBM

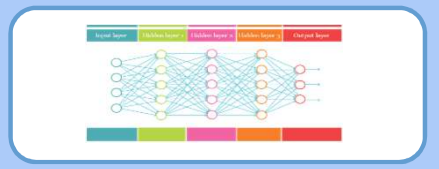
Machine learning and deep learning in AI development - dependencies

Subject matter expertise

Input data (e.g. images, audio and/or text data)

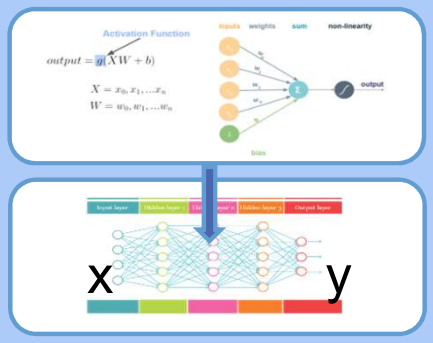
1. AI need / development problem

10. Trained model to be used as AI/part of AI



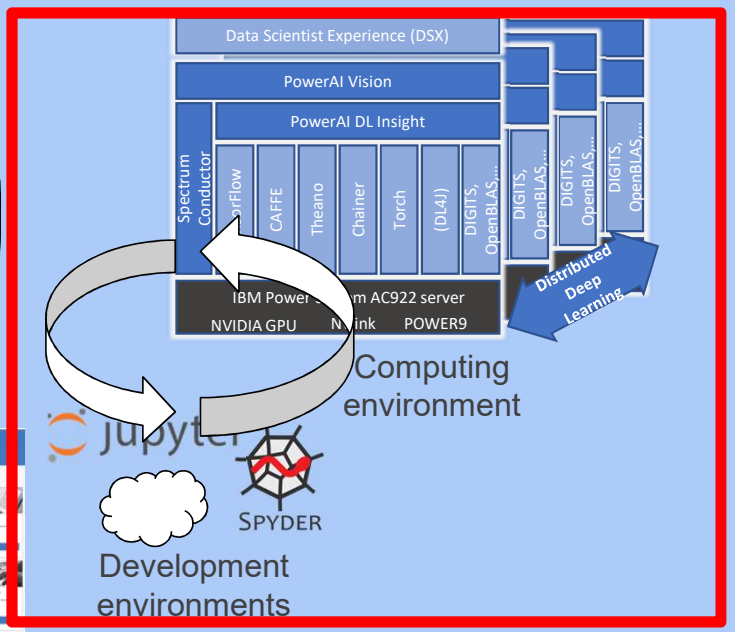
arXiv.org

Research publications



Mathematical models (e.g. deep neural nets)

Data scientist



Computing environment

Development environments

python™

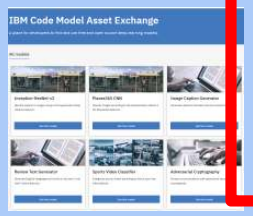
Programming language

TensorFlow

GitHub

Example software source code

Software libraries / machine learning frameworks



Trained models

IBM

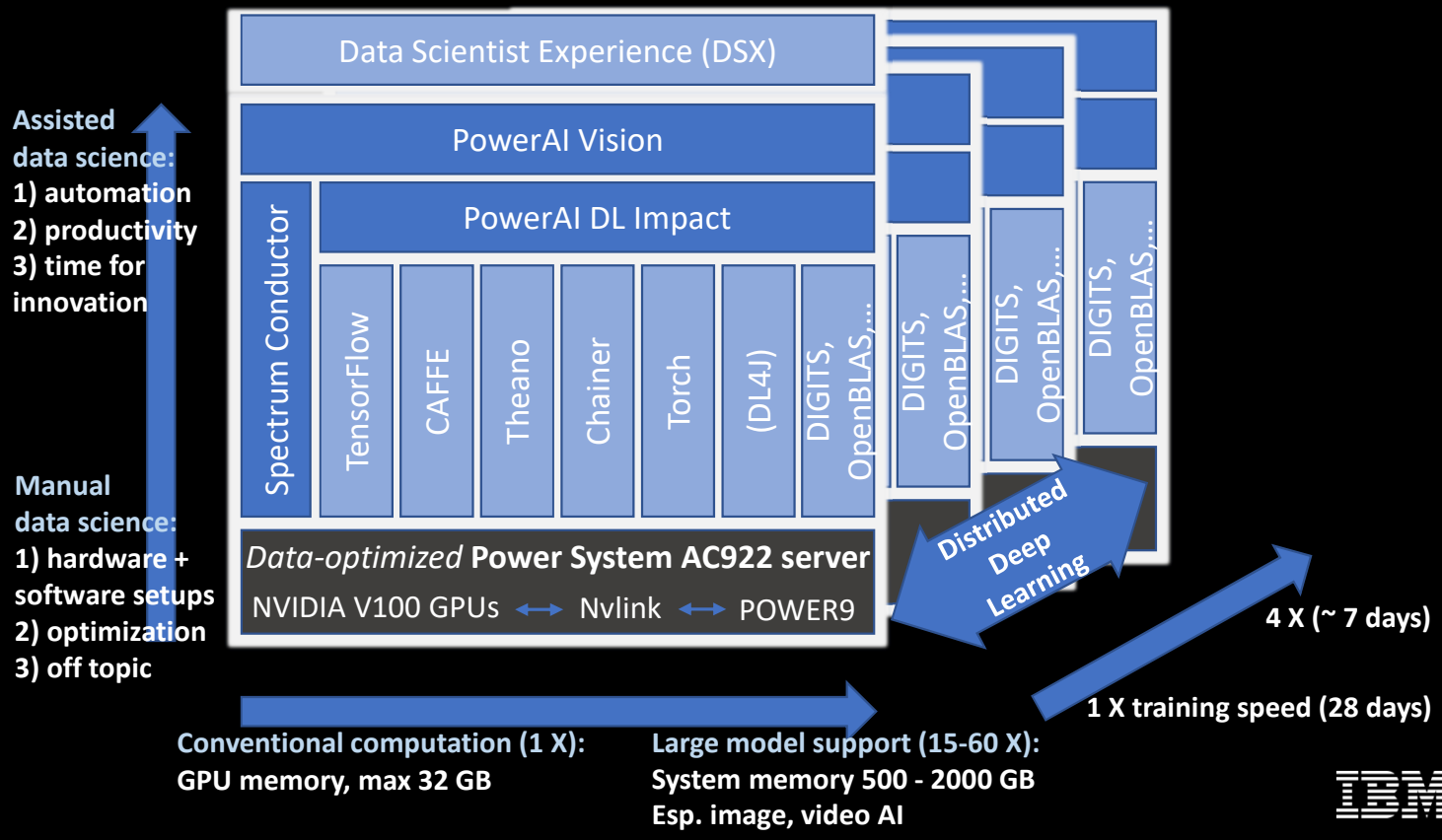
Multiplying data science with IBM PowerAI - faster value creation from deep learning AI

”Data scientists don’t train a model - they train thousands!”

Options for deployment

a) Self-operated system or cluster suitable for IT + user departments
b) Services from IBM business partner ecosystem and networks

PowerAI package & tools – 3 axes of differentiation and productivity



Jukka Remes, BDE, Dr (Tech.)
Teppo Seesto, Solution Architect

IBM

Multiplying data science with IBM PowerAI - faster value creation from deep learning AI

”Data scientists don’t train a model - they train thousands!”

SW stack - use the level you need:

1) **Full data science environment with GUI (DSX) (and/or cluster management IBM Cloud Private local installation)**
2) **Dedicated computer vision solution for low threshold DL adoption (PowerAI Vision)**
3) **DL experiment automation, e.g. hyperparameter optimization (PowerAI DL Impact)**

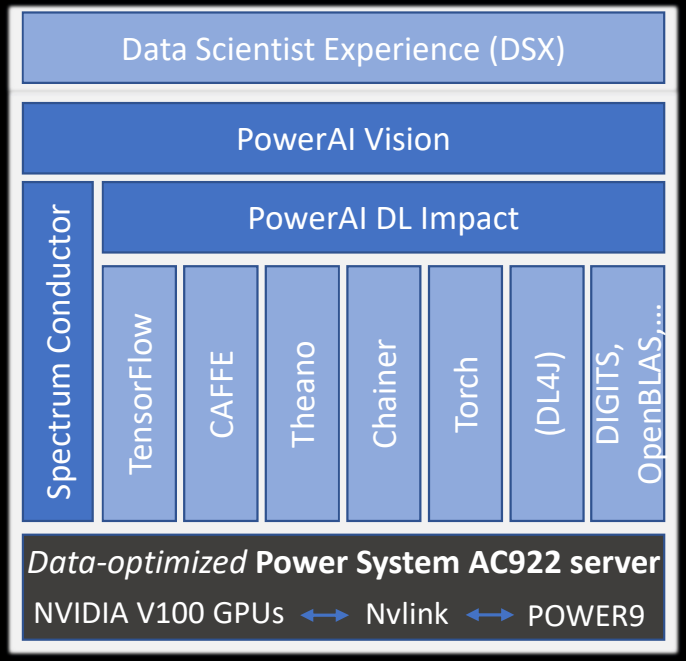
and/or

4) **Conventional use of deep learning frameworks (TF, Caffe, Torch, Keras etc.) out-of-box (e.g. Linux command line, Python, tools of your choosing)**

Assisted data science:
1) automation
2) productivity
3) time for innovation

Manual data science:
1) hardware + software setups
2) optimization
3) off topic

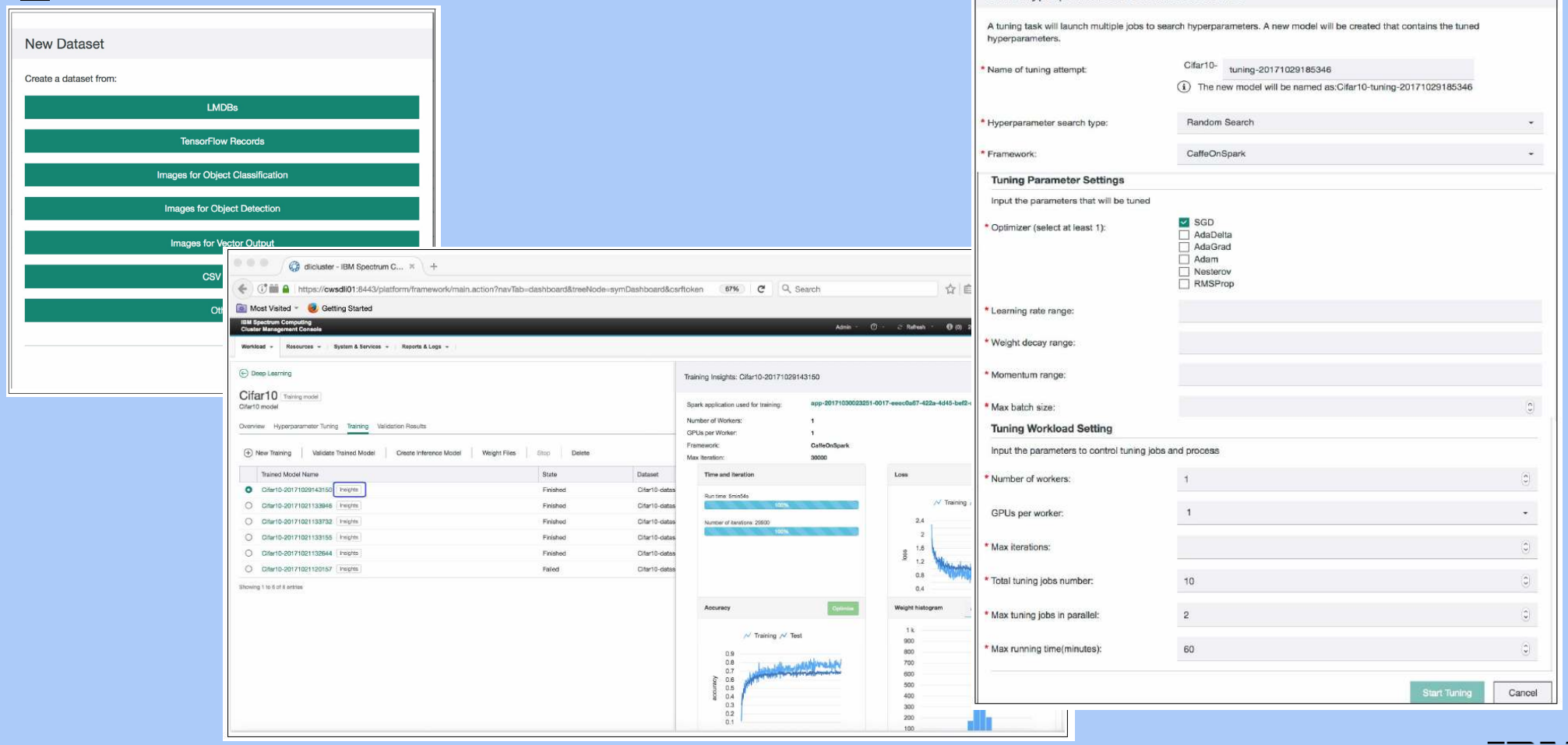
PowerAI package & tools – 1st axis: different options for usage



Jukka Remes, BDE, Dr (Tech.)
Teppo Seesto, Solution Architect

IBM

Deep Learning Impact (DLI), WML Accelerator (including model deployment)

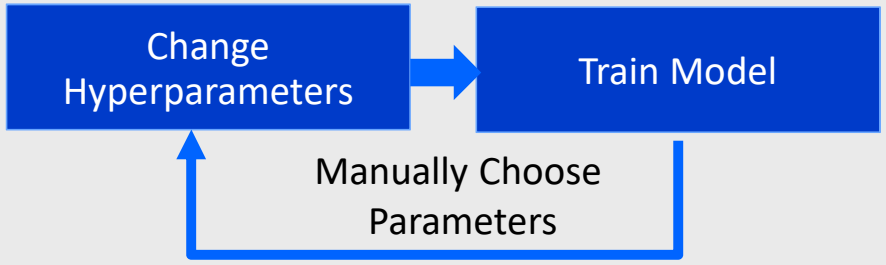


http://www.redbooks.ibm.com/abstracts/sg248409.html?Open

Auto Hyper-Parameter Optimization (HPO) in WML Accelerator / PowerAI Enterprise

Manual Process
Can take weeks

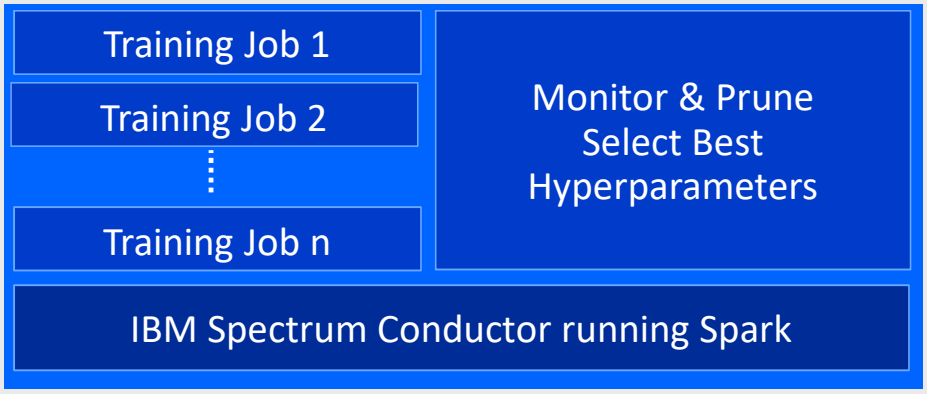
Run Model Training 100s of Times



Lots of Hyperparameters:

Learning rate, Decay rate, Batch size, Optimizers (Gradient Descent, Momentum, ..)

Auto-Hyperparameter Optimizer (Auto-HPO)
Done in Hours

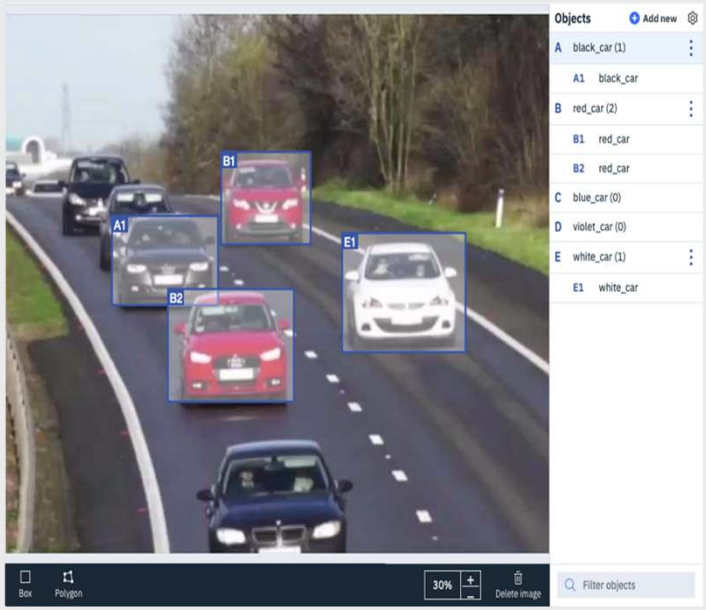


Auto-HPO has 3 search approaches

Random, Tree-based Parzen Estimator (TPE), Bayesian

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

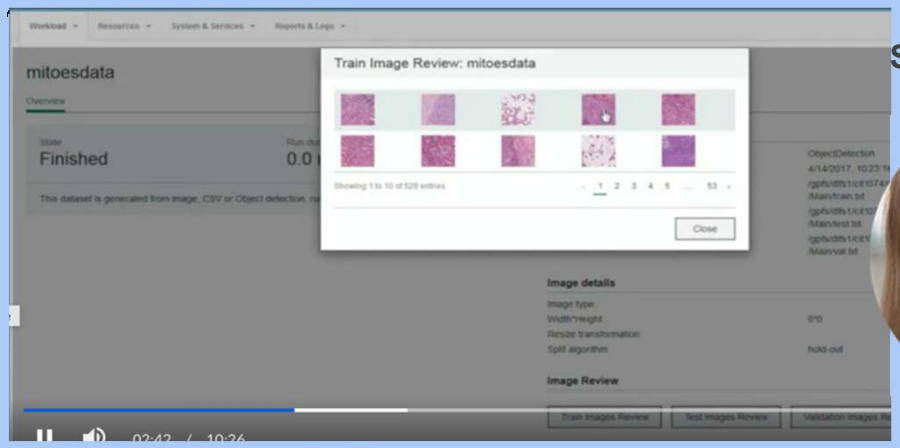
Label Image or Video Data



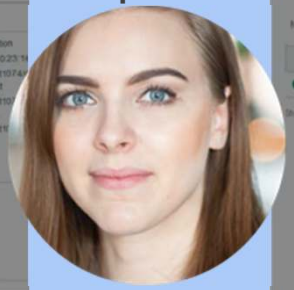
Auto-Train AI Model

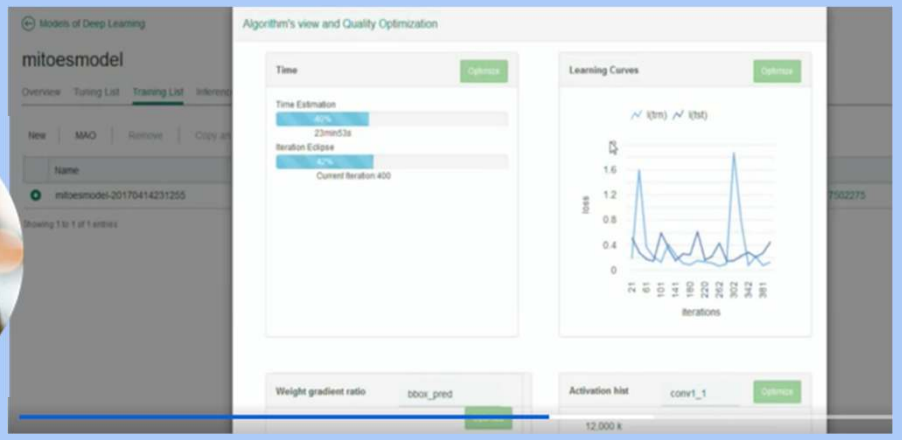
Package & Deploy AI Model

… hiding all together the machine learning details (getting greater group involved in AI development)

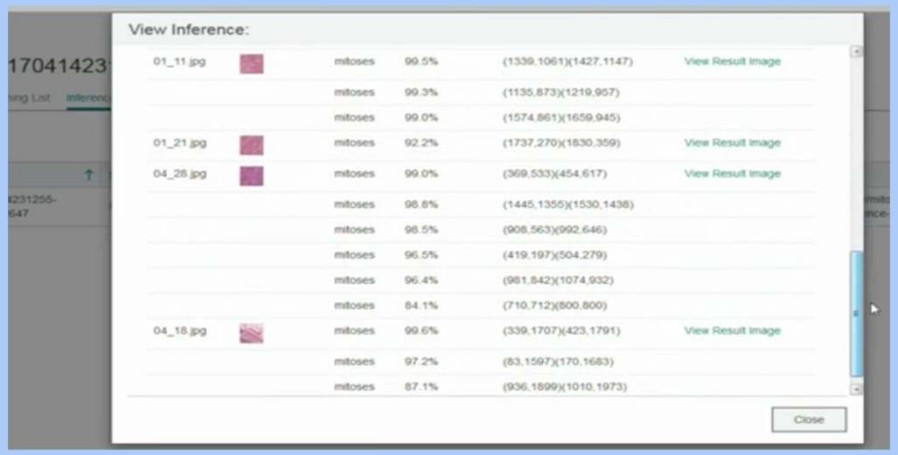


Subject matter experts

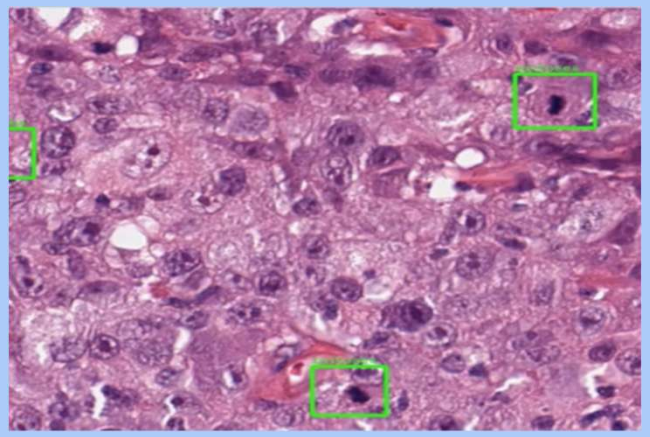




Application developers



Data scientists



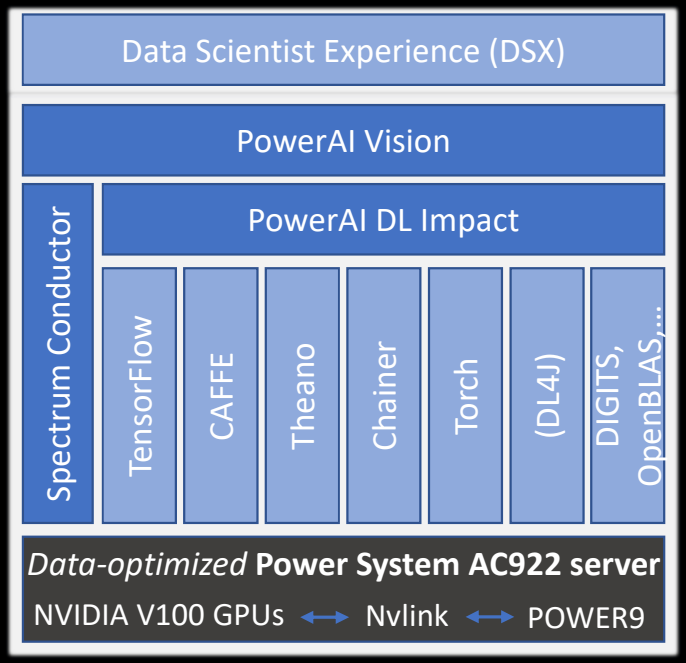
Tumor Proliferation Assessment – mitosis detection
Images from electron-microscope
Size of image - 70K * 60K

https://developer.ibm.com/linuxonpower/2018/05/11/breast-cancer-classification-ibm-powerai-vision/

Multiplying data science with IBM PowerAI - faster value creation from deep learning AI

”Data scientists don’t train a model - they train thousands!”

PowerAI package & tools – 2nd axis: go beyond GPU memory limits



Conventional computation (1 X):
GPU memory, max 32 GB

Large model support (15-60 X):
System memory 500 - 2000 GB
Esp. image, video AI

Jukka Remes, BDE, Dr (Tech.)
Teppo Seesto, Solution Architect

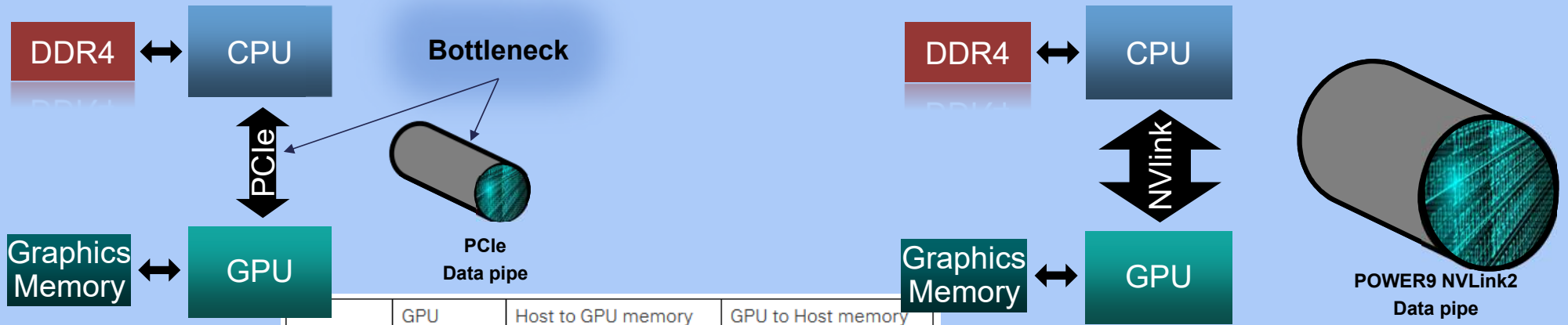
IBM

Large Model Support (LMS) on PowerAI (only) 1/2

GPU memory (max 8-16/32 GB) has to host active experiment (including neural net parameters) as well training data

System main memory much larger in server systems (128 GB – 2000 GB)

NVLINK provides both fast GPU interconnection and in Power servers also fast connection between GPUs and system main memory - IO-advantage in itself for feeding training data to GPUs



	GPU utilization	Host to GPU memory copy throughput	GPU to Host memory copy throughput
x86	29.1%	10.9 GB/s	11.7 GB/s
AC922 (PowerAI)	73.5%	64.4 GB/s	60.4 GB/s

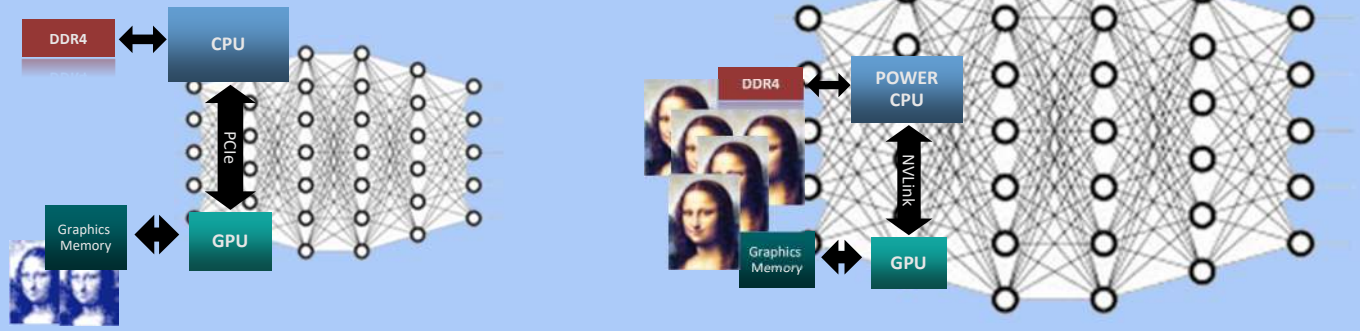
https://developer.ibm.com/linuxonpower/2018/07/27/tensorflow-large-model-support-case-study-3d-image-segmentation/

Large Model Support (LMS) on PowerAI (only) 2/2

“Large Model Support” (LMS) – larger experiments than GPU memory:

- Caffe: swap chunks of data between GPU and system main memory

- Tensorflow: computation graph rewriting in order to swap results between GPU and system main memory



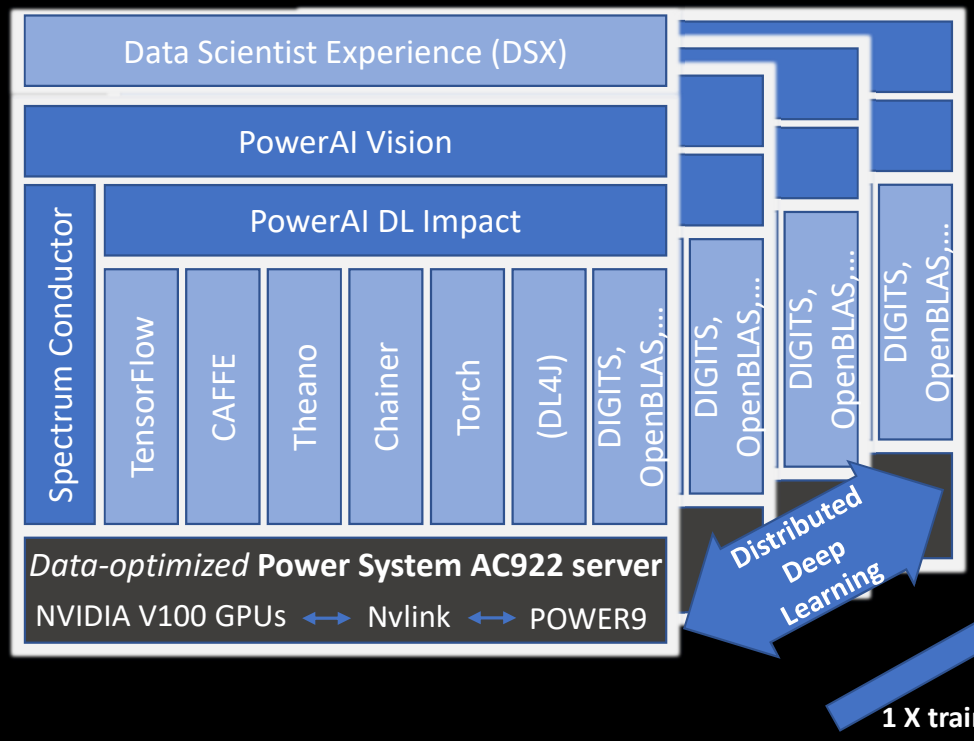
https://github.com/ibmsoe/caffe/tree/master-lms
https://developer.ibm.com/linuxonpower/2018/07/27/tensorflow-large-model-support-case-study-3d-image-segmentation/
https://arxiv.org/abs/1807.02037
Technology preview of LMS Tensorflow support on PowerAI: /opt/DL/tensorflow/doc/README-LMS.md

IBM

Multiplying data science with IBM PowerAI - faster value creation from deep learning AI

”Data scientists don’t train a model - they train thousands!”

PowerAI package & tools – 3rd axis: easily divide workloads across cluster



Distributed DL: HW and SW support readily included

Jukka Remes, BDE, Dr (Tech.)
Teppo Seesto, Solution Architect

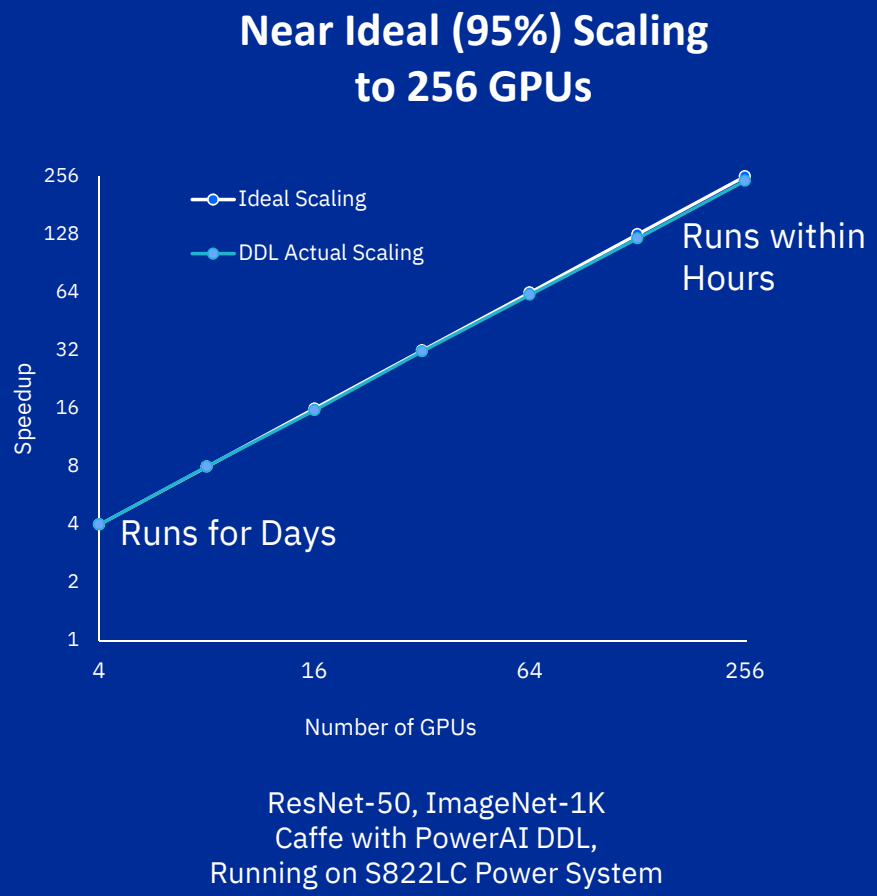
IBM

Distributed Deep Learning (DDL)

Deep learning training takes days to weeks

DDL in WML CE extends TensorFlow & enables scaling to 100s of servers

Automatically distribute and train on large datasets to 100s of GPUs



ResNet-50, ImageNet-1K
Caffe with PowerAI DDL,
Running on S822LC Power System

Elastic Distributed Training (EDT) in WML Accelerator / PowerAI Enterprise

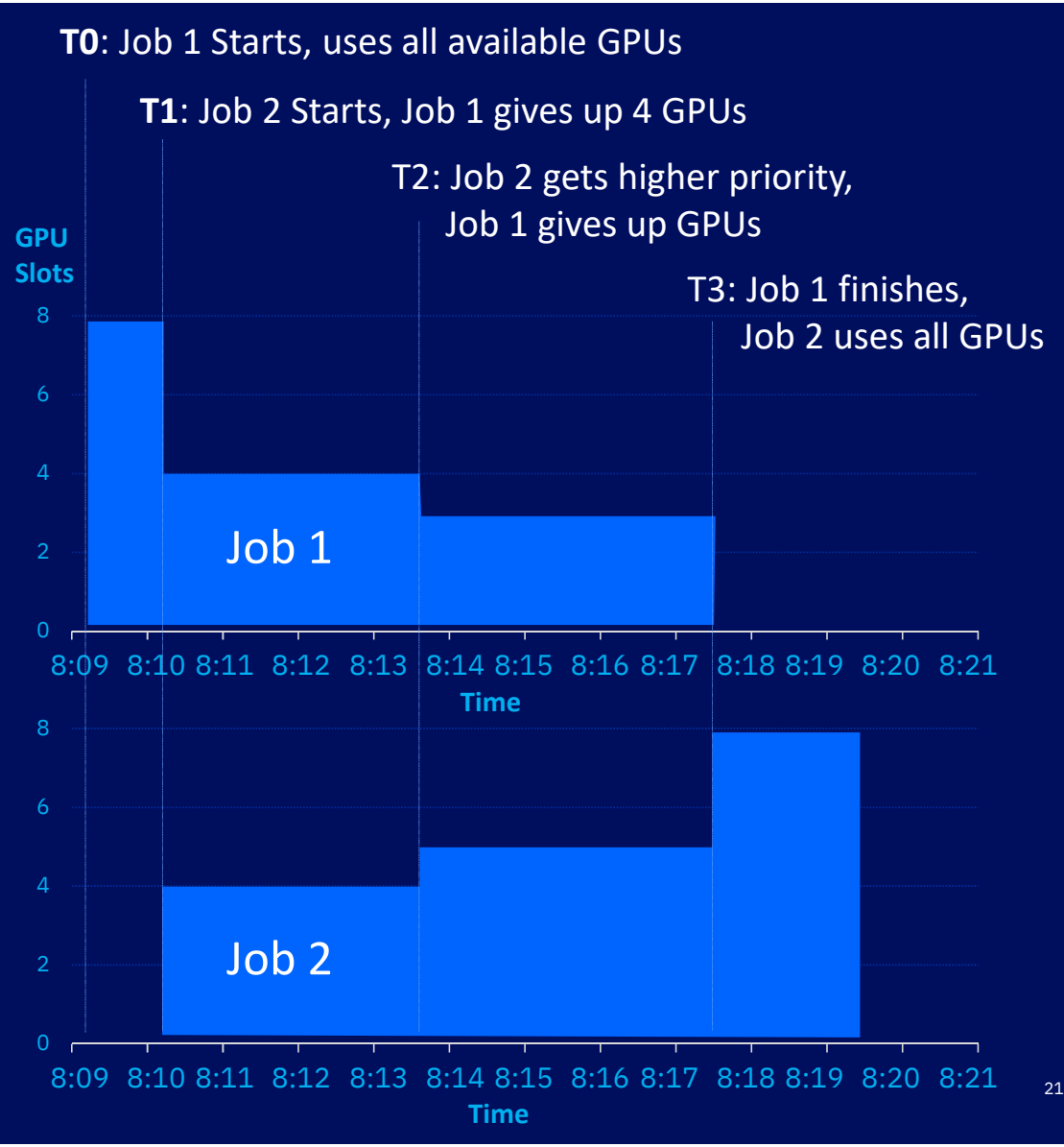
Dynamically Reallocates GPUs within milliseconds

Increases Job Throughput and Server / GPU Utilization

Works with Spark & AI Jobs

Works with Hybrid x86 & Power Cluster

2 Servers with 4 GPUs each: total 8 GPUs
Available Policies: Fair share, Preemption, Priority



Multiplying data science with IBM PowerAI - faster value creation from deep learning AI

"Data scientists don't train a model - they train thousands!"

Options for deployment

a) Self-operated system or cluster suitable for IT + user departments
b) Services from IBM business partner ecosystem and networks

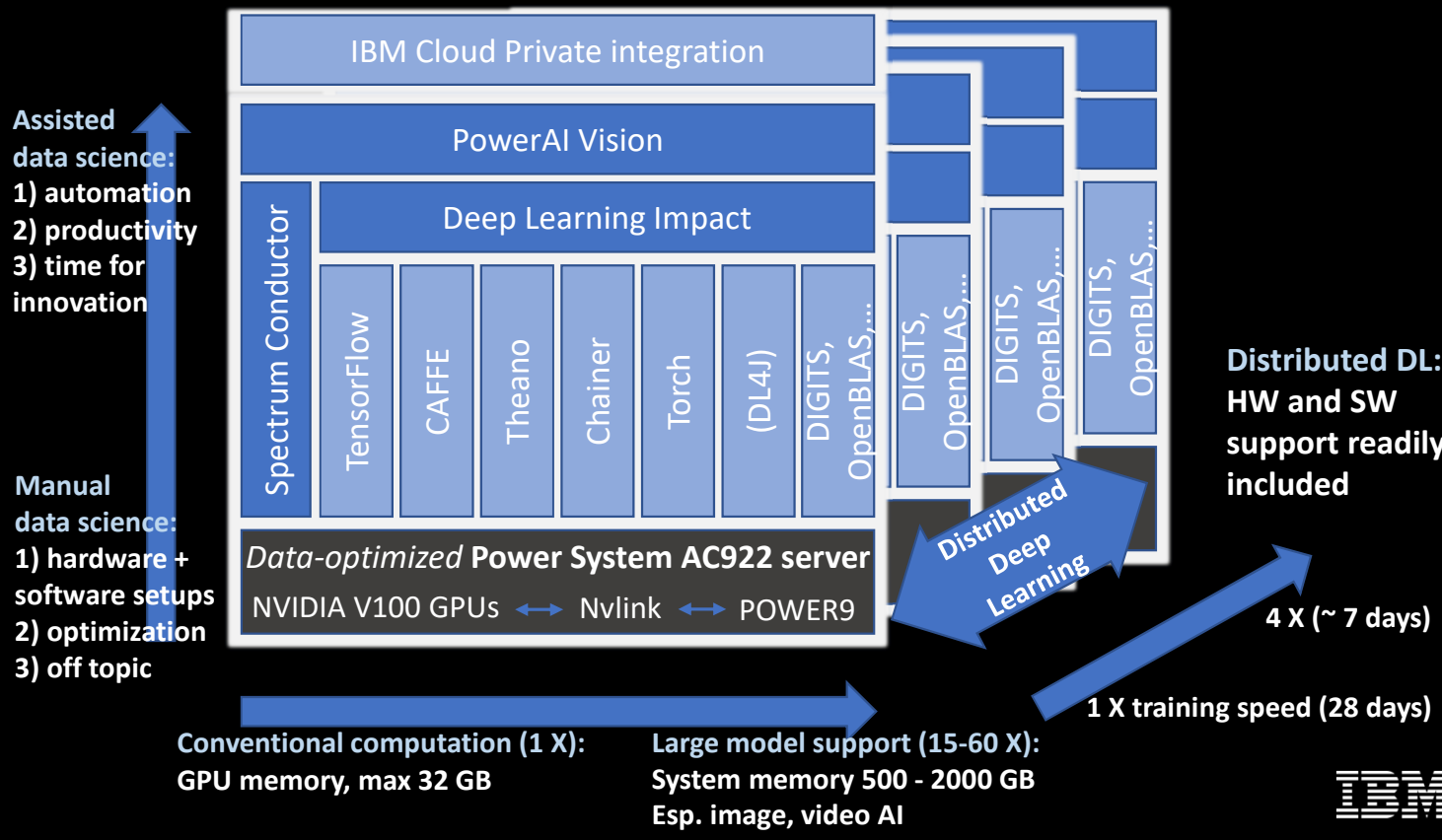
SW stack - use the level you need:

1) Full data science environment with GUI (DSX) (and/or cluster management IBM Cloud Private local installation)
2) Dedicated computer vision solution for low threshold DL adoption (PowerAI Vision)
3) DL experiment automation, e.g. hyperparameter optimization (PowerAI DL Impact)

and/or

4) Conventional use of deep learning frameworks (TF, Caffe, Torch, Keras etc.) out-of-box (e.g. Linux command line, Python, tools of your choosing)

PowerAI package & tools – 3 axes of differentiation and productivity



Jukka Remes, BDE, Dr (Tech.)
Teppo Seesto, Solution Architect

IBM

IBM Open Source Based AI Stack

Auto-AI software: PowerAI Vision, IBM Auto-AI

Watson Studio

WML CE

Data Preparation
Model Development
Environment

Watson Machine Learning

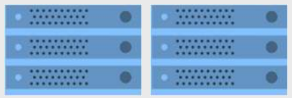
Watson ML Accelerator

Watson ML CE

Runtime Environment
Train, Deploy, Manage Models

Watson OpenScale

Model Metrics, Bias,
and Fairness
Monitoring



Accelerated AC922
Power9 Servers

Storage
(Spectrum Scale ESS)

Runs on x86 & other storage too
Available on Private Cloud or Public Cloud

Previous Names:
WML Accelerator = PowerAI Enterprise
WML Community Ed. = PowerAI-base

Our Focus: Ease of Use & Faster Model Training Times

Watson ML Accelerator

Distributed Deep Learning (DDL)

Auto Hyper-Parameter Optimization (HPO)

Elastic Distributed Training (EDT) & Elastic Distributed Inference (EDI)

IBM Spectrum Conductor
Apache Spark, Cluster Virtualization, Job Orchestration

Watson ML

Model Management & Execution

Model Life Cycle Management

Watson ML Community Edition WML CE

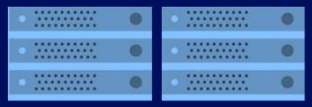
WML CE: Open Source ML Frameworks

TensorFlow™ PYTORCH Chainer Snap ML

Large Model Support (LMS)

DDL-16

Infrastructure Designed for AI



Power9 or x86 Servers with GPU Accelerators

Storage (ESS)

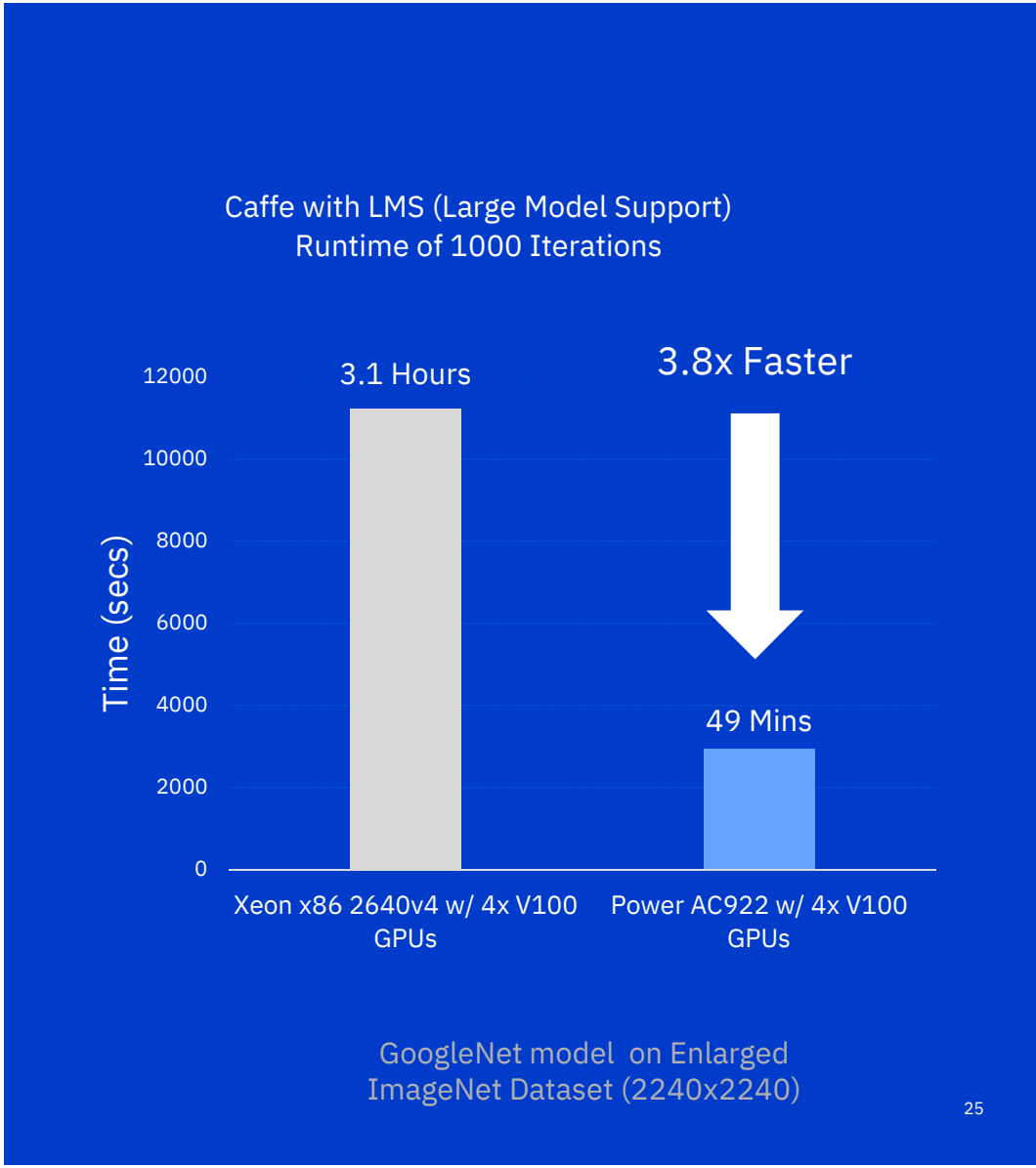
Large AI Models Train ~4 Times Faster

-

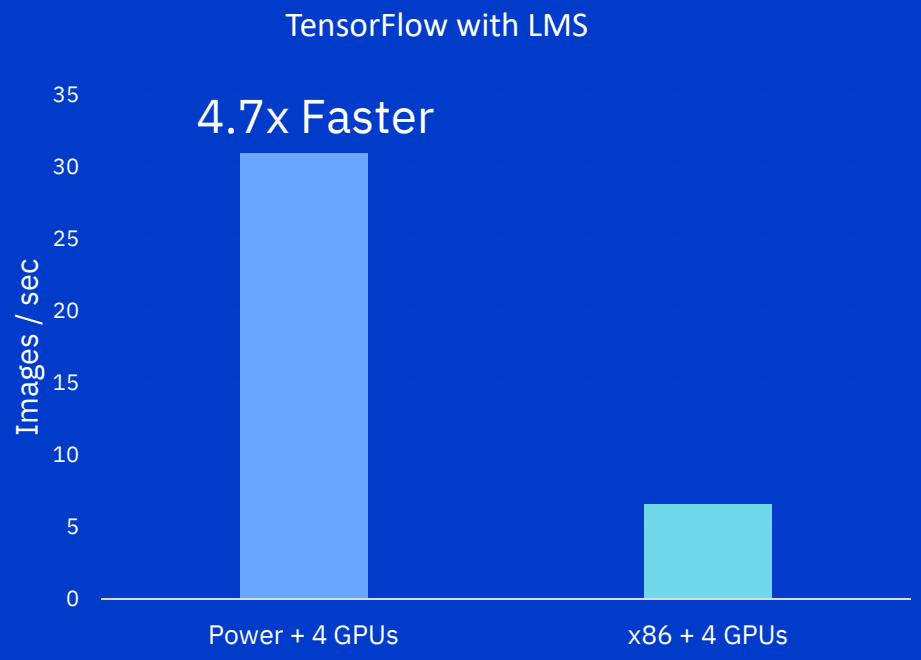
Time-to-Results Drops from Months to Weeks

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

Detailed Benchmark Information is available



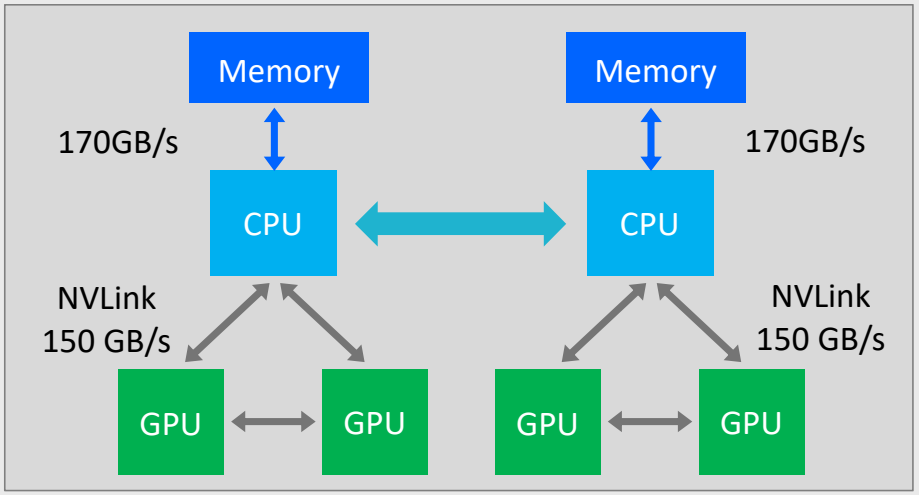
Large Model Support (LMS) Enables Higher Accuracy via Larger Models



500 Iterations of Enlarged GoogleNet model on Enlarged ImageNet Dataset (2240x2240), mini-batch size = 15
Both servers with 4 NVIDIA V100 GPUs

Store Large Models & Dataset in System Memory

Transfer One Layer at a Time to GPU



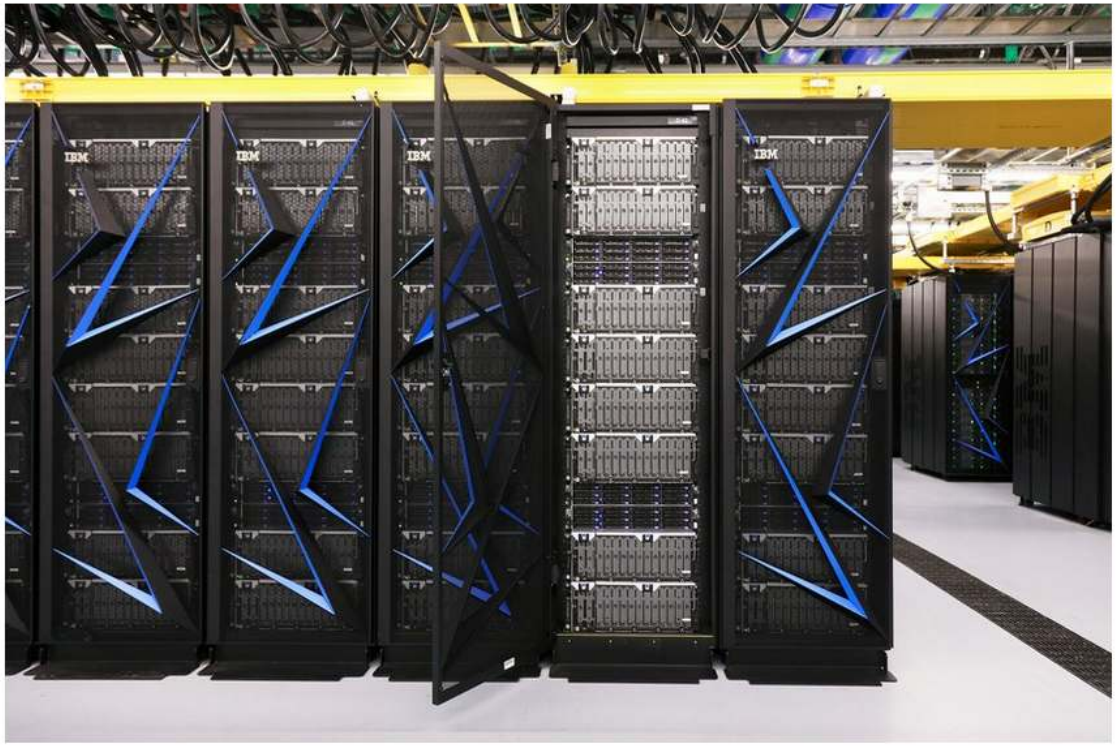
IBM AC922 Power9 Server

CPU-GPU NVLink 5x Faster than Intel x86 PCI-Gen3

WIRED

TOM SIMONITE BUSINESS 06.08.18 12:00 PM

THE US AGAIN HAS THE WORLD'S MOST POWERFUL SUPERCOMPUTER



The IBM-built Summit supercomputer is the world's smartest and most powerful AI machine. Its racks are connected by over 185 miles of fiber-optic cables.

GENEVIEVE MARTIN/OAK RIDGE NATIONAL LABORATORY

https://www.wired.com/story/the-us-again-has-worlds-most-powerful-supercomputer/

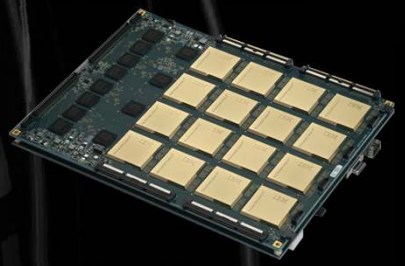
The Future of Accelerated Computing

IBM PowerAI

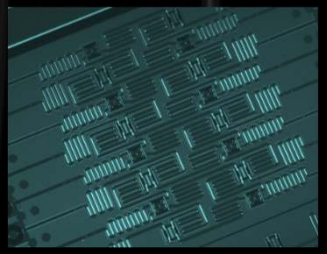
Today:
Graphic Processors
NVIDIA Volta

Power System

New:
FPGA
Xilinx Alveo



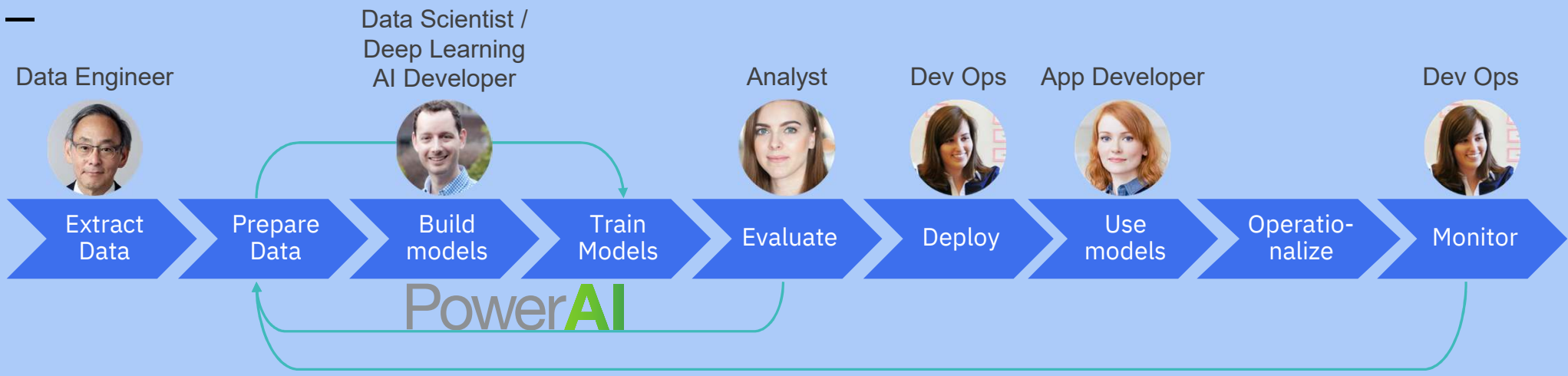
Tomorrow:
Neuromorpic Processors
IBM TrueNorth



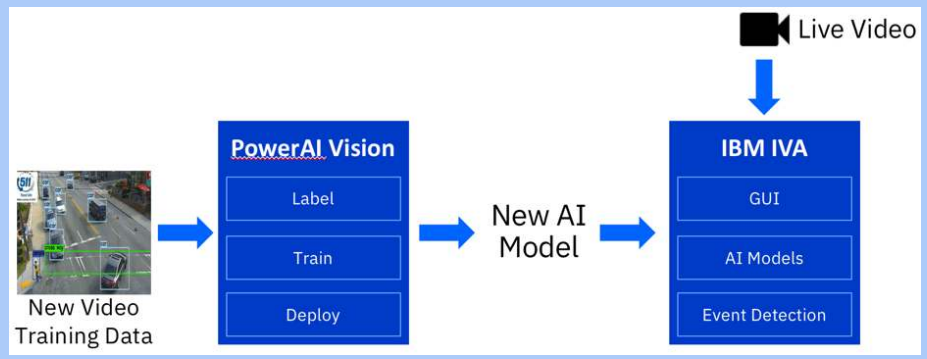
Future:
Quantum Computer
IBM Q


OpenPOWER™
Implementation / HPC / Research
Software
System / Integration
I/O / Storage / Acceleration
Boards / Systems
Chip / SOC

Systematic data science: deployment, integration and continuous AI development/model improvement

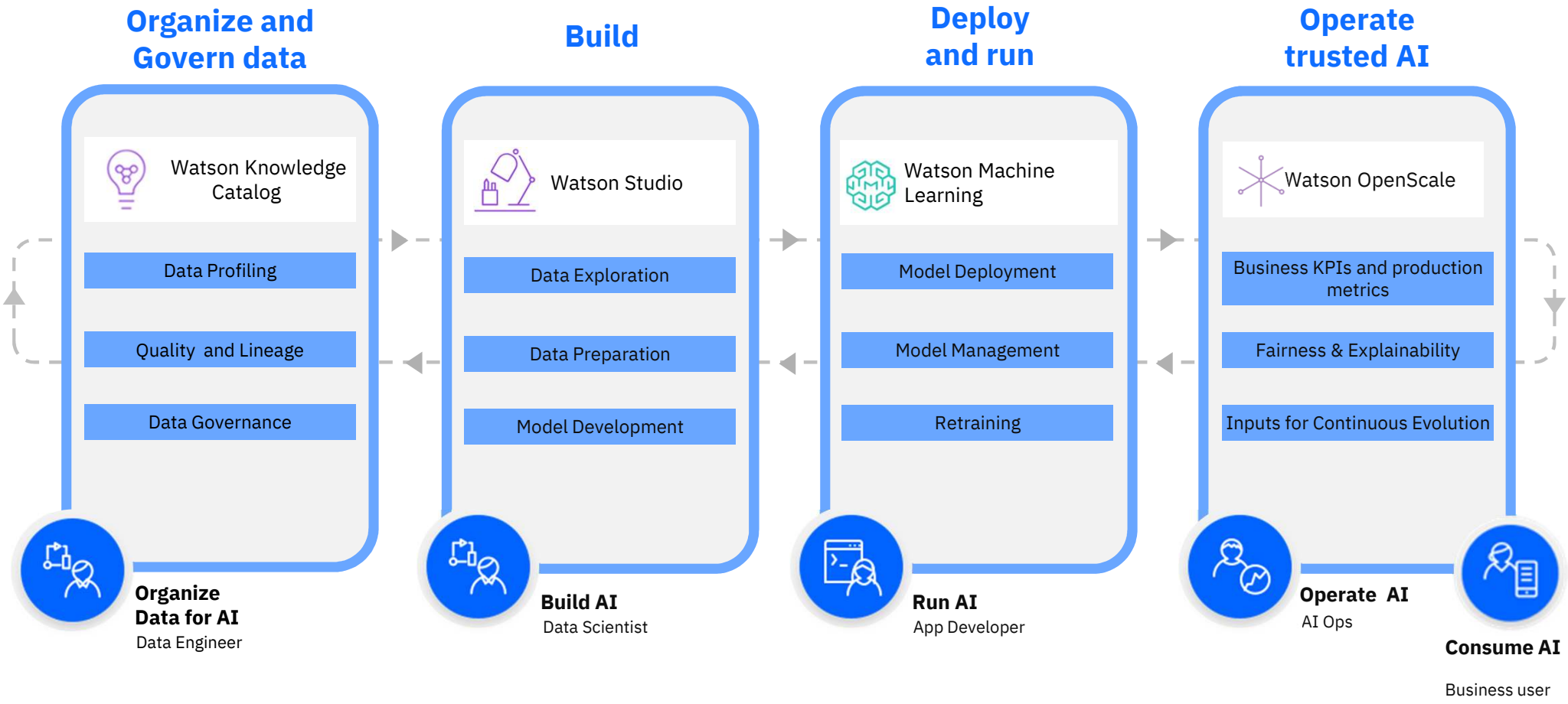


Interested parties need to be able to consume the trained model and the model needs to be supported



https://medium.com/@sumitg_16893/deep-insights-with-ai-for-video-analytics-5464fd30ebe1

Watson tool portfolio addresses all the stages of AI development and utilization



Thank you!

Questions?

Multiplying data science with IBM PowerAI - faster value creation from deep learning AI

"Data scientists don't train a model - they train thousands!"

Options for deployment

a) Self-operated system or cluster suitable for IT + user departments
b) Services from IBM business partner ecosystem and networks

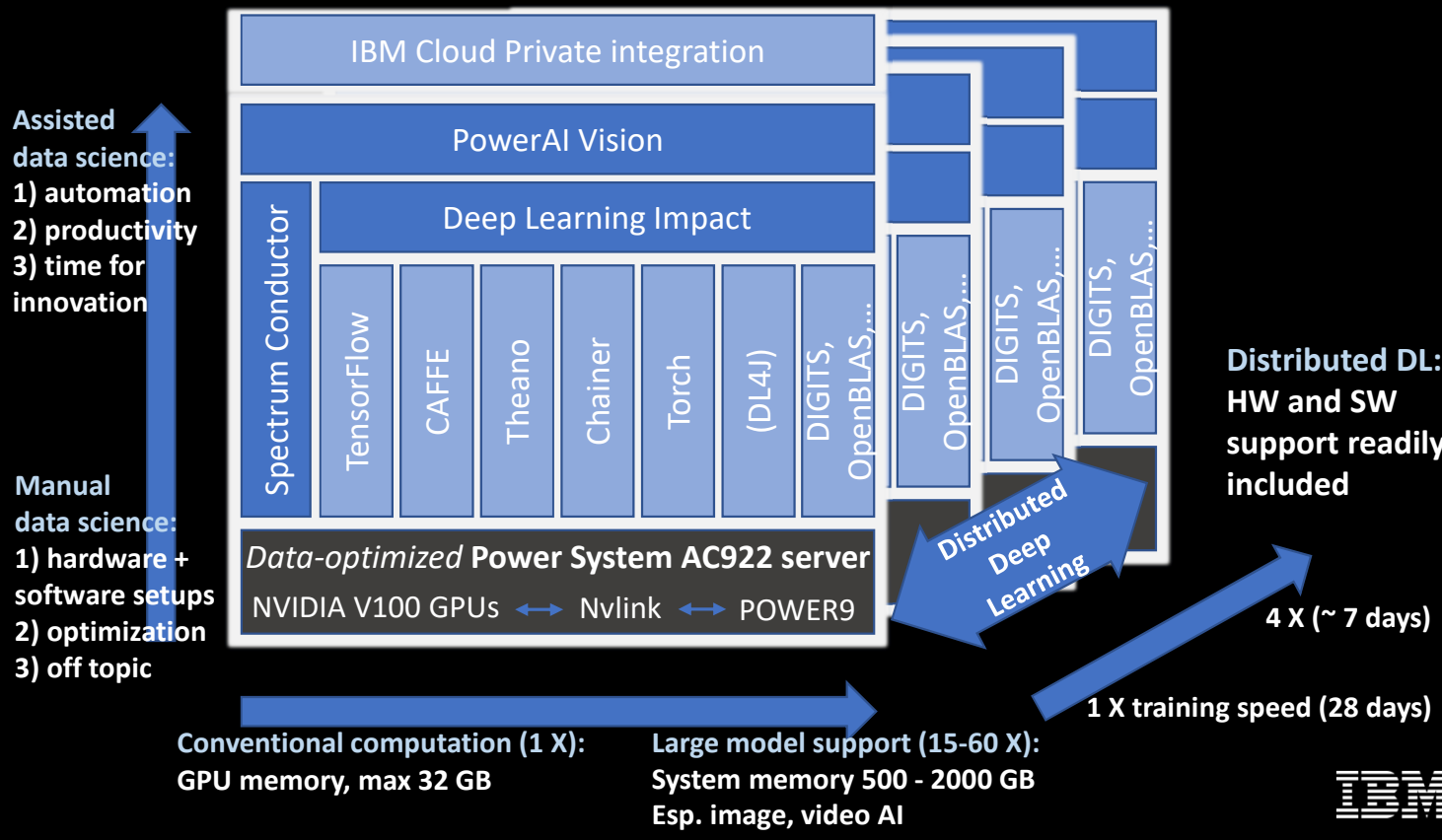
SW stack - use the level you need:

1) Full data science environment with GUI (DSX) (and/or cluster management IBM Cloud Private local installation)
2) Dedicated computer vision solution for low threshold DL adoption (PowerAI Vision)
3) DL experiment automation, e.g. hyperparameter optimization (PowerAI DL Impact)

and/or

4) Conventional use of deep learning frameworks (TF, Caffe, Torch, Keras etc.) out-of-box (e.g. Linux command line, Python, tools of your choosing)

PowerAI package & tools – 3 axes of differentiation and productivity



Jukka Remes, BDE, Dr (Tech.)
Teppo Seesto, Solution Architect

IBM