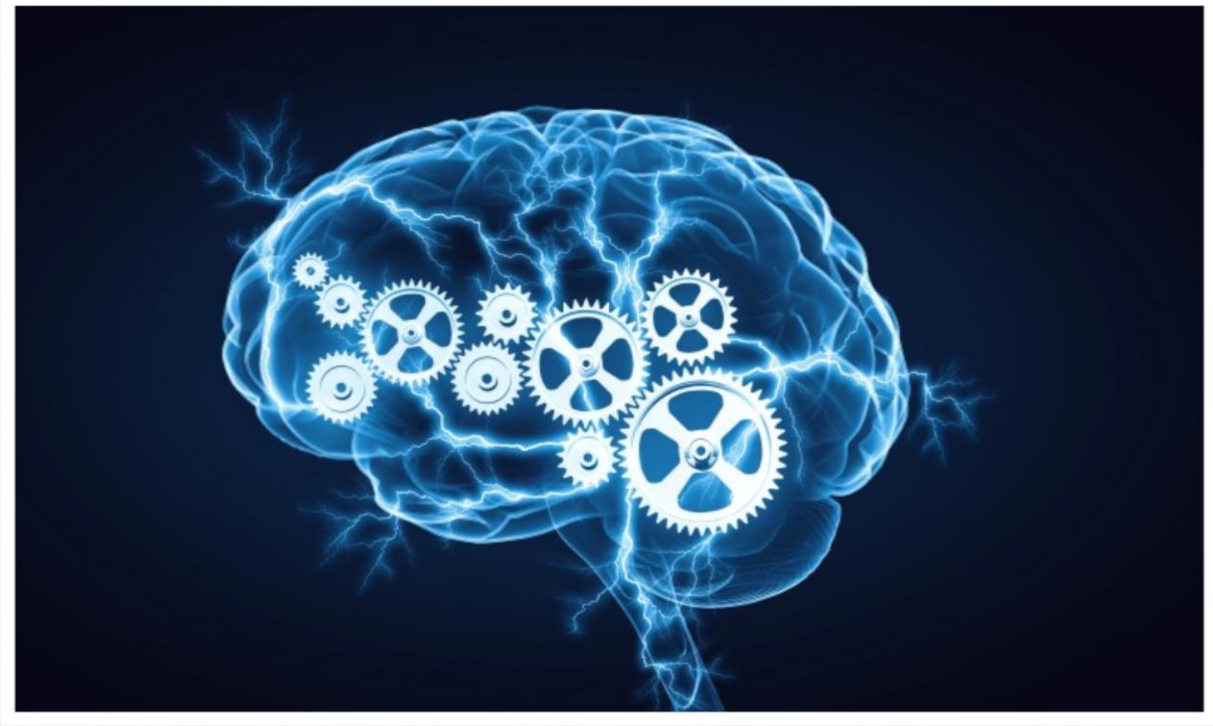
HPC, Big Data, AI: What are the new needs ? Are all infrastructure solutions equal ?



Alain Cyr, PhD

Infrastructure Solutions Architect
Montpellier IBM Client Center

cyralain@fr.ibm.com

@trollnyrd

IBM®

POWER9 OpenPOWER™

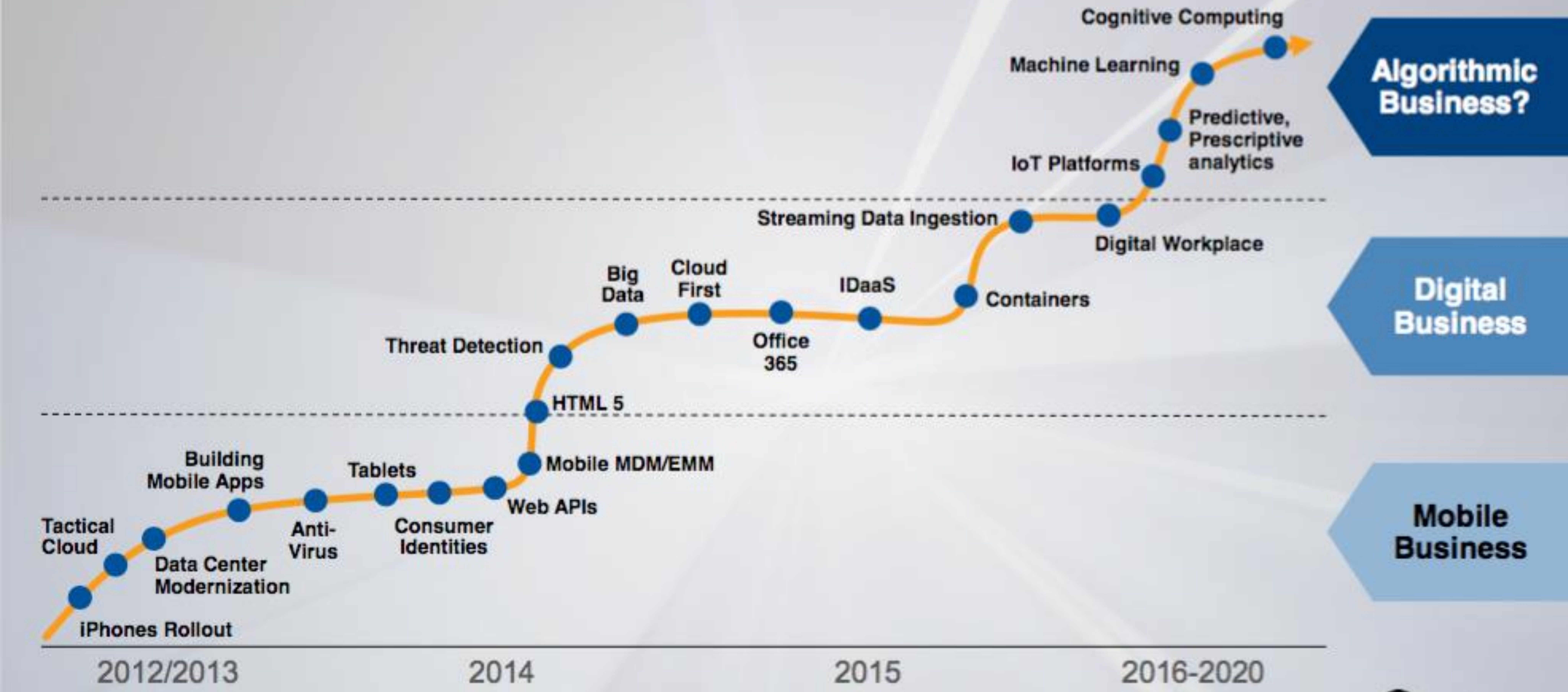
Agenda

- AI and Cognitive Use Cases

- Data driven needs: what are the right platforms ?

- Entreprise class offering for AI: PowerAI

The Path to the Digital Business and Beyond



Source: Gartner, "Architecting the On-Demand Digital Business"; Drue Reeves, Kyle Hilgendorf, Kirk Knoernschild, August 16, 2016

Gartner.

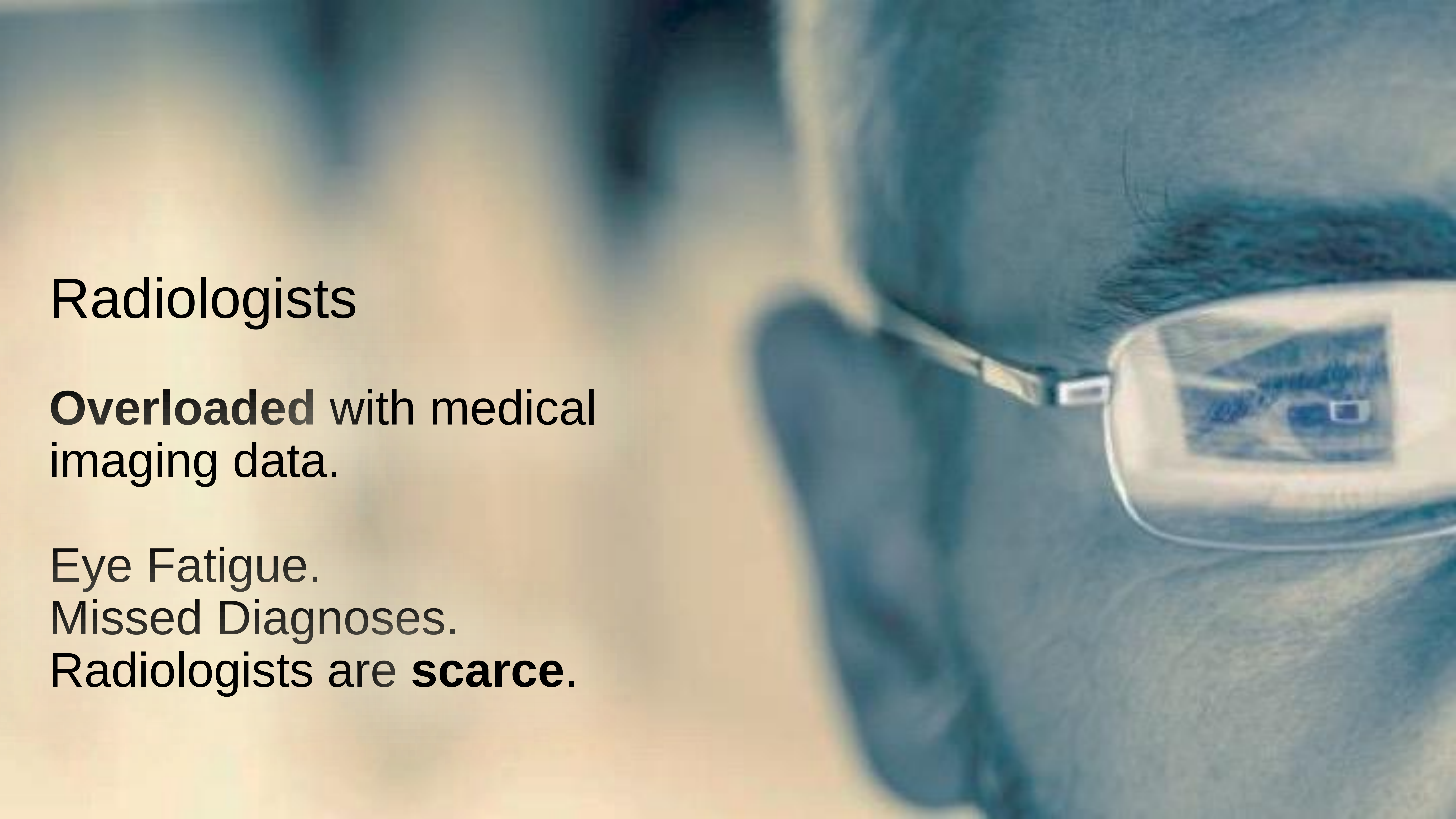
↓ 90% inspection times

↑ 10X number of inspections

↓ accident risk rate

IBMPower

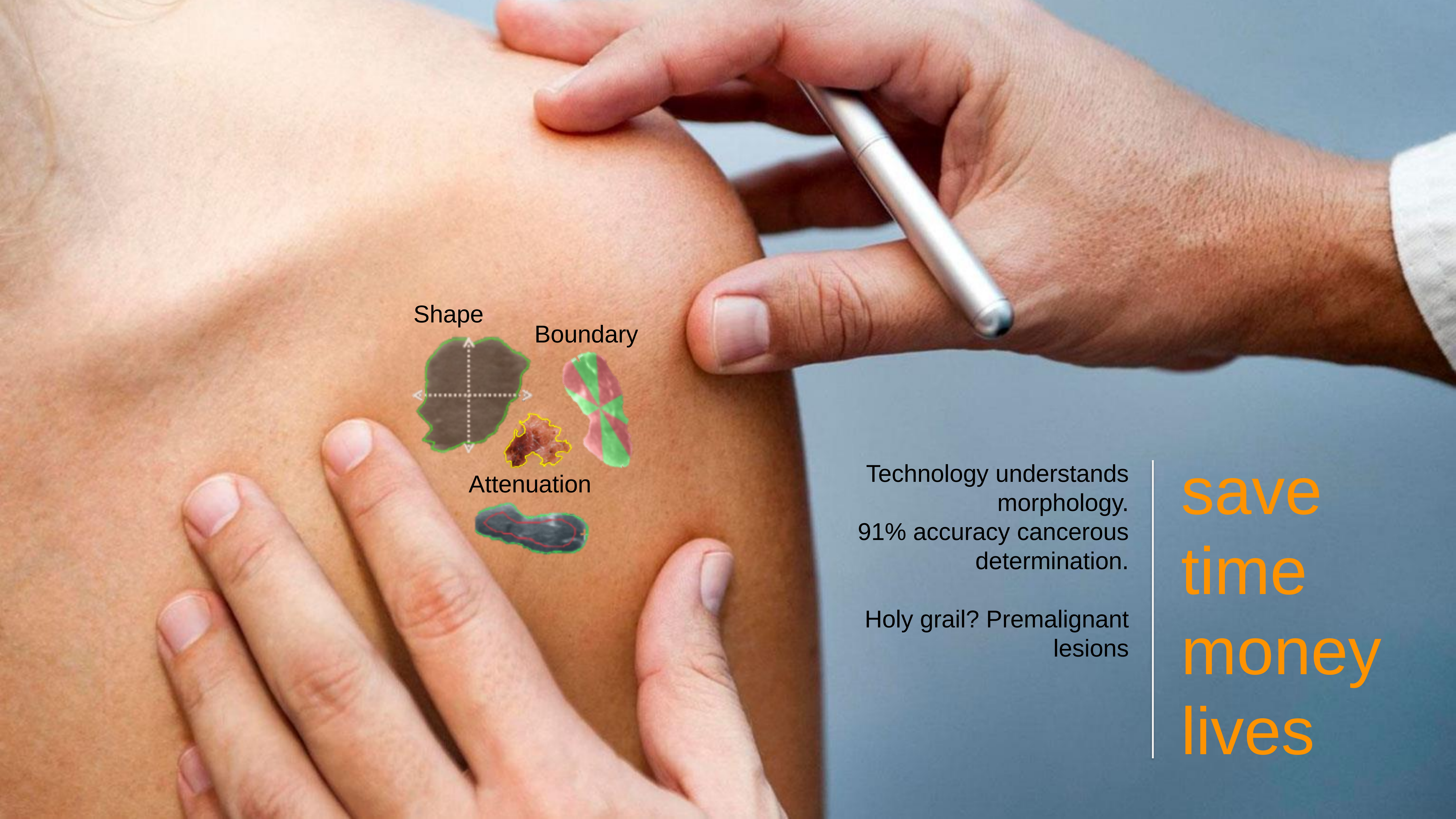
KEPCO
KEPRI



Radiologists

Overloaded with medical imaging data.

Eye Fatigue.
Missed Diagnoses.
Radiologists are **scarce**.



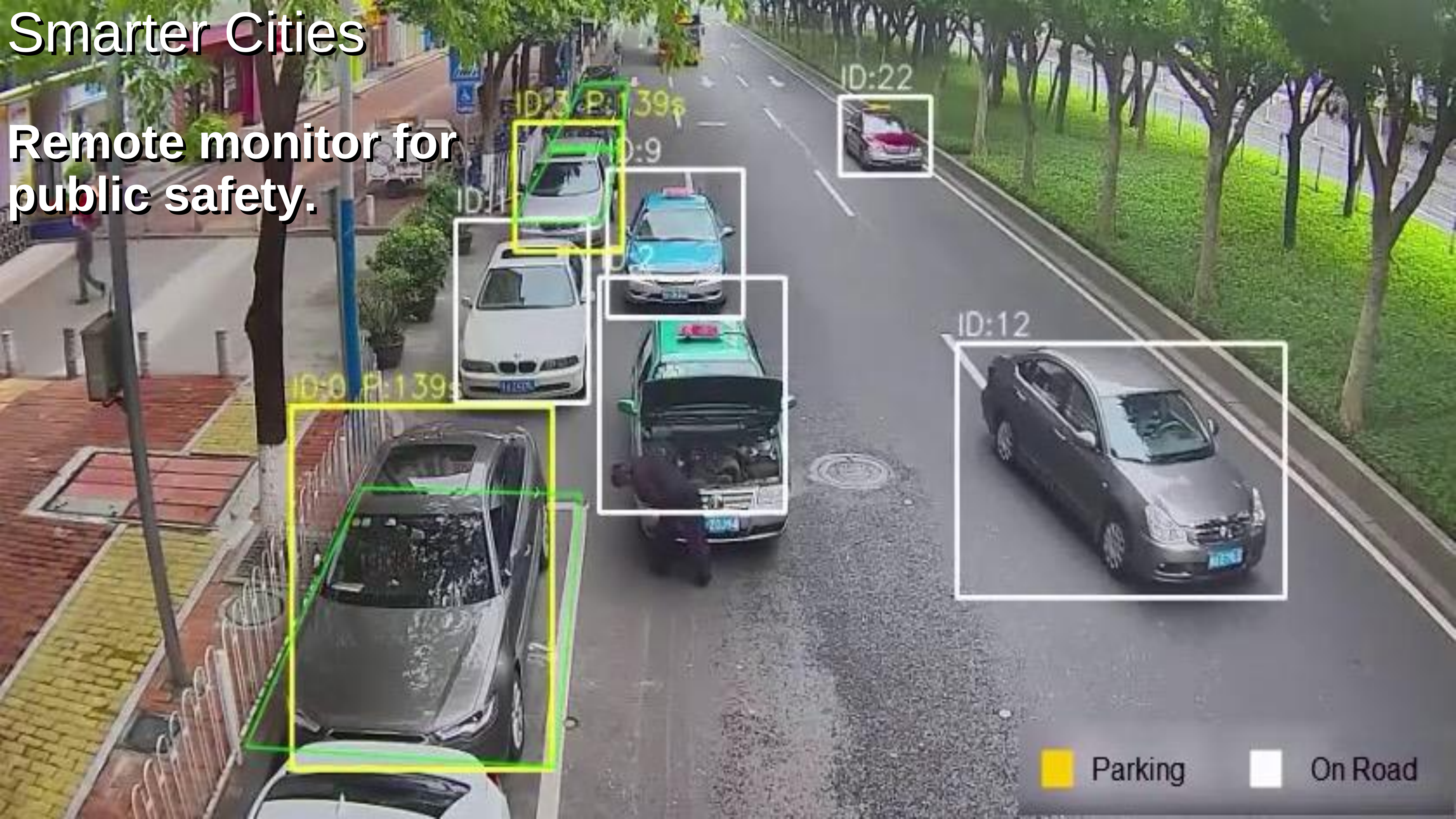
Technology understands morphology.
91% accuracy cancerous determination.

Holy grail? Premalignant lesions

save
time
money
lives

Smarter Cities

Remote monitor for public safety.



25 gigabytes of data per hour is generated by a connected car.

90% of cars will be connected by 2020.

80 million wearable health devices will be available by 2017.

153 exabytes of healthcare data generated by devices in 2013.

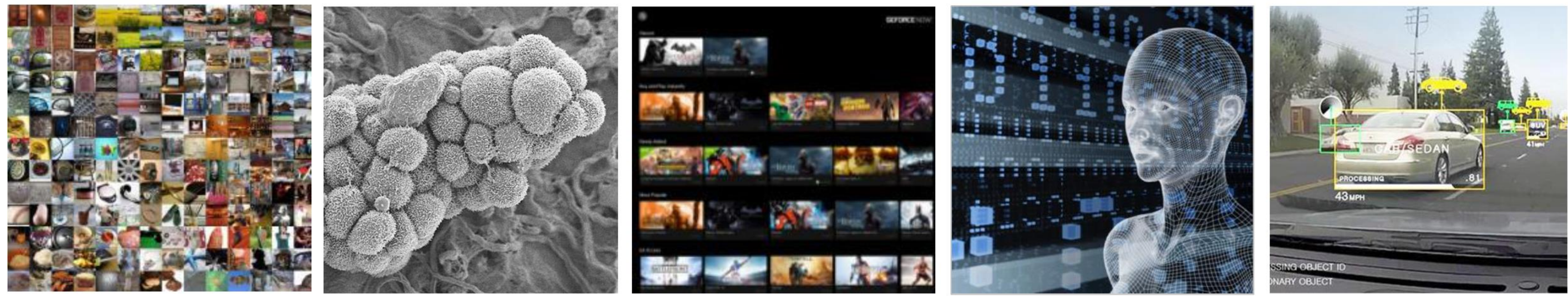
Increasing to **2,314 exabytes** in 2020.

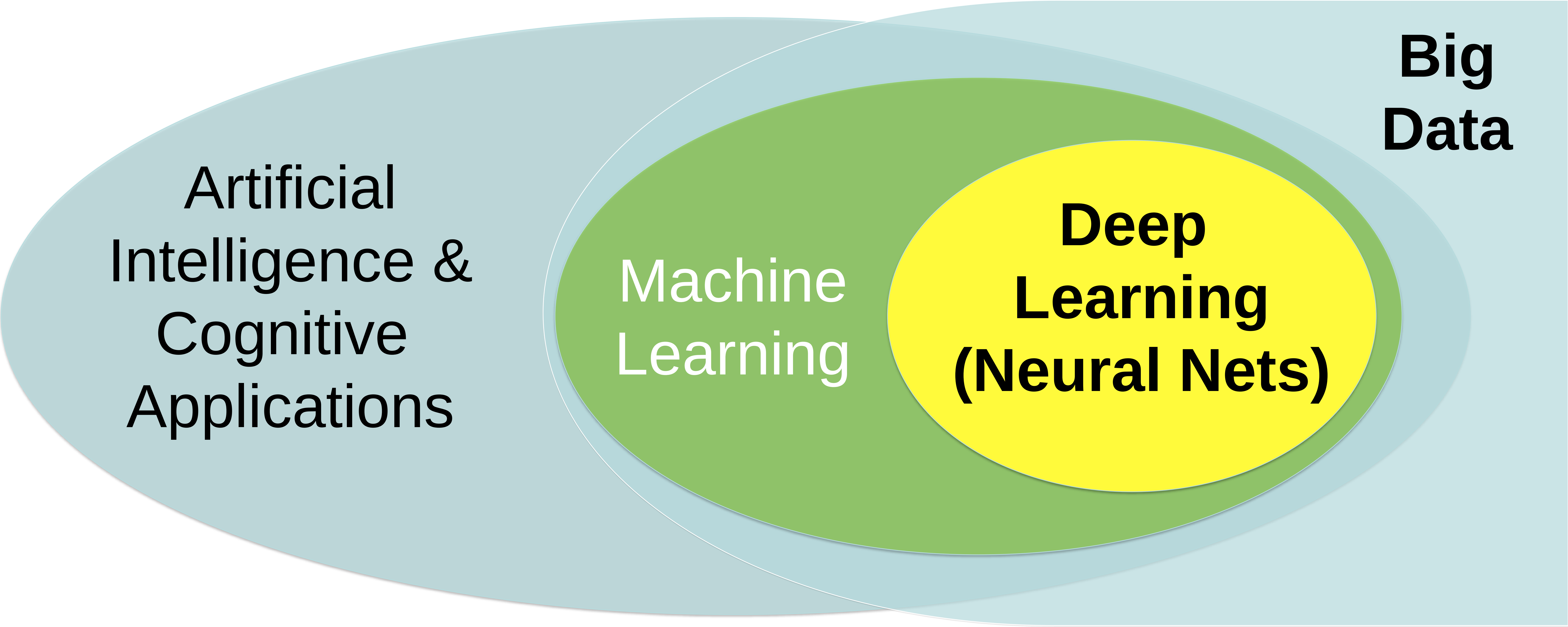
2.5 quintillion bytes of data generated daily by connected machines.

There will be **28 times more sensor-enabled devices than people** by the year 2020.

1.7 megabytes of data per second generated by every human being on the planet by 2020.

The Landscape is Evolving





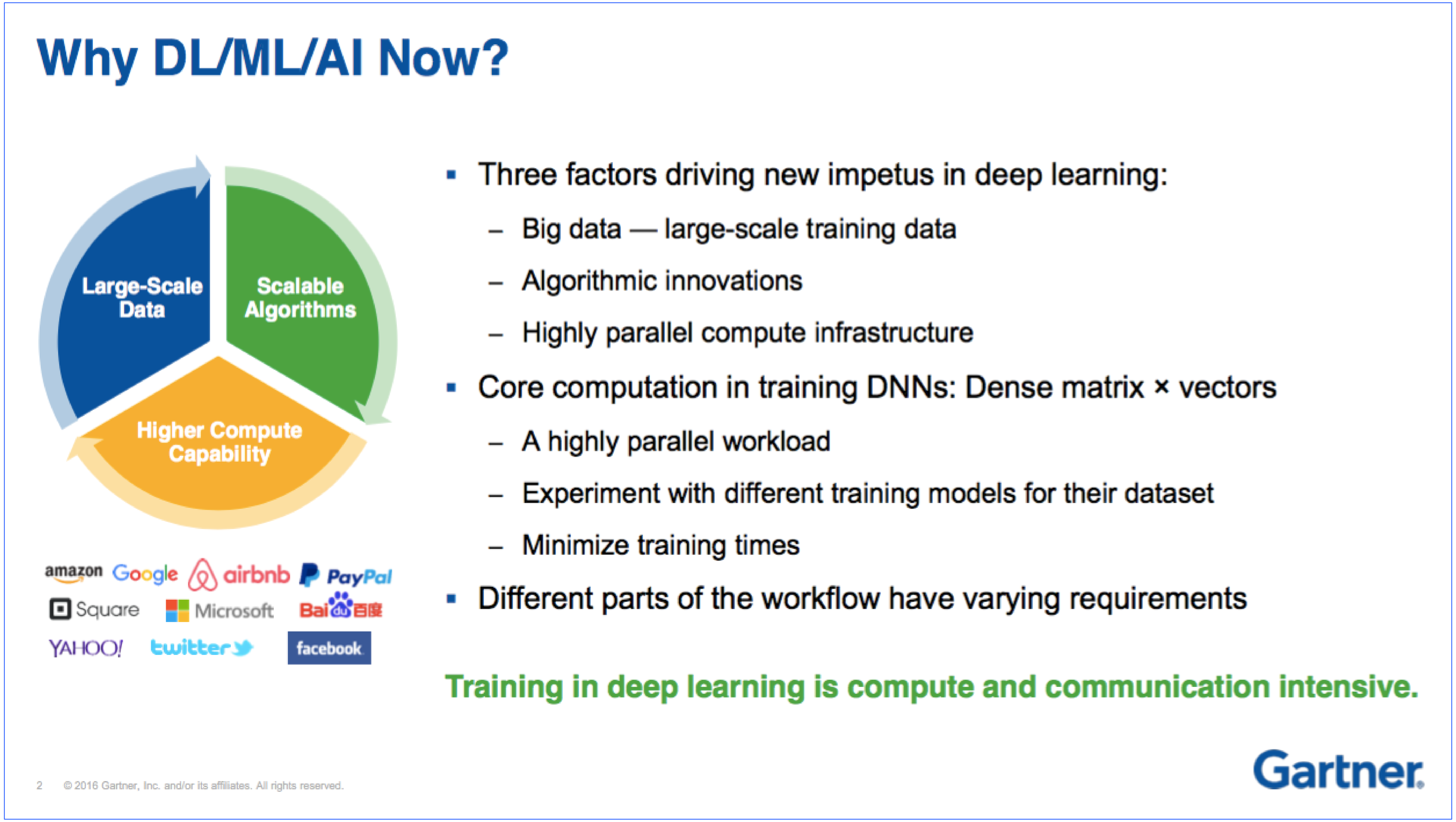
Big Data

Artificial Intelligence & Cognitive Applications

Machine Learning

Deep Learning (Neural Nets)

Gartner Views…



Why DL/ML/AI Now?

Large-Scale Data
Scalable Algorithms
Higher Compute Capability
amazon
Google
airbnb
PayPal
Square
Microsoft
Baidu 百度
YAHOO!
twitter
facebook

- Three factors driving new impetus in deep learning:
  - Big data — large-scale training data
  - Algorithmic innovations
  - Highly parallel compute infrastructure
- Core computation in training DNNs: Dense matrix × vectors
  - A highly parallel workload
  - Experiment with different training models for their dataset
  - Minimize training times
- Different parts of the workflow have varying requirements

Training in deep learning is compute and communication intensive.

Cognitive/AI applies across all industries

AUTOMOTIVE

Self-driving cars,
Driver safety,
Insurance

COMMUNICATIONS

Location-based advertising,
Speech processing

CONSUMER PACKAGED GOODS

Sentiment analysis of
what's hot, product
positioning

FINANCIAL SERVICES

Risk, fraud, surveillance,
product opportunities

EDUCATION & RESEARCH

Interactive learning

HIGH TECHNOLOGY / INDUSTRIAL MFG.

Robotics, Mfg. quality,
Warranty analysis

LIFE SCIENCES

Drug reactions, drug
discovery

MEDIA/ENTERTAINMENT

Viewers / advertising
effectiveness

ON-LINE SERVICES / SOCIAL MEDIA

Dialogue, image
processing, sentiment

HEALTH CARE

Patient monitoring,
diagnostics

OIL & GAS

Exploration,
simulation efficiency

RETAIL

Consumer sentiment
Demand forecasting

TRAVEL & TRANSPORTATION

Traffic and safety
management

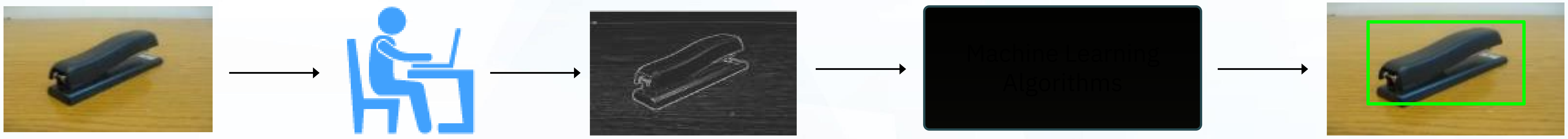
UTILITIES

Smart Meter analysis
for network capacity,

LAW ENFORCEMENT & DEFENSE

Threat analysis - social
media monitoring, photo
analysis

Machine Learning



Input

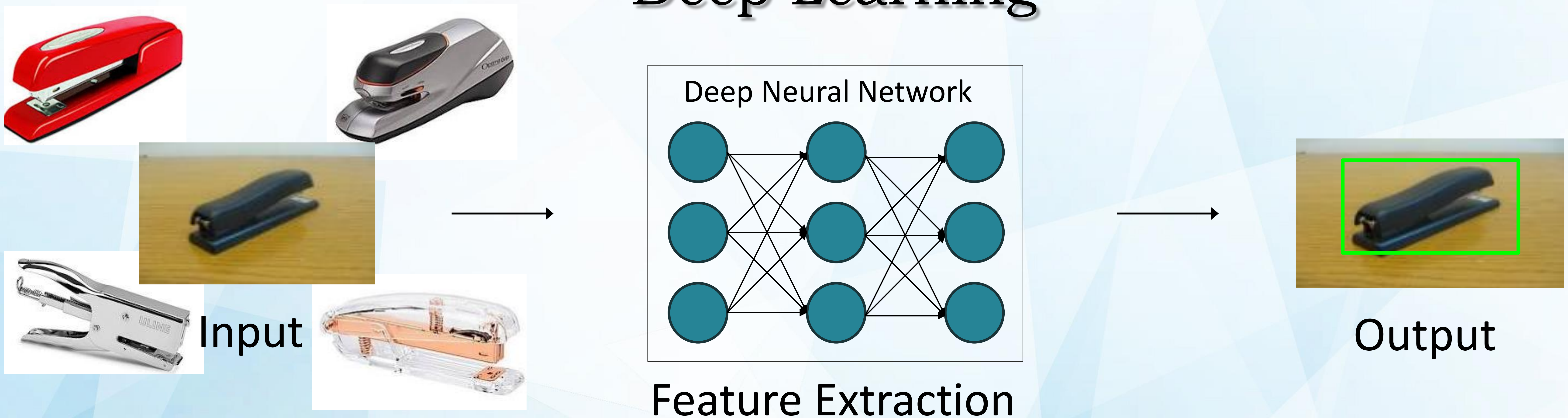
Feature Extraction

Features

Classification

Output

Deep Learning

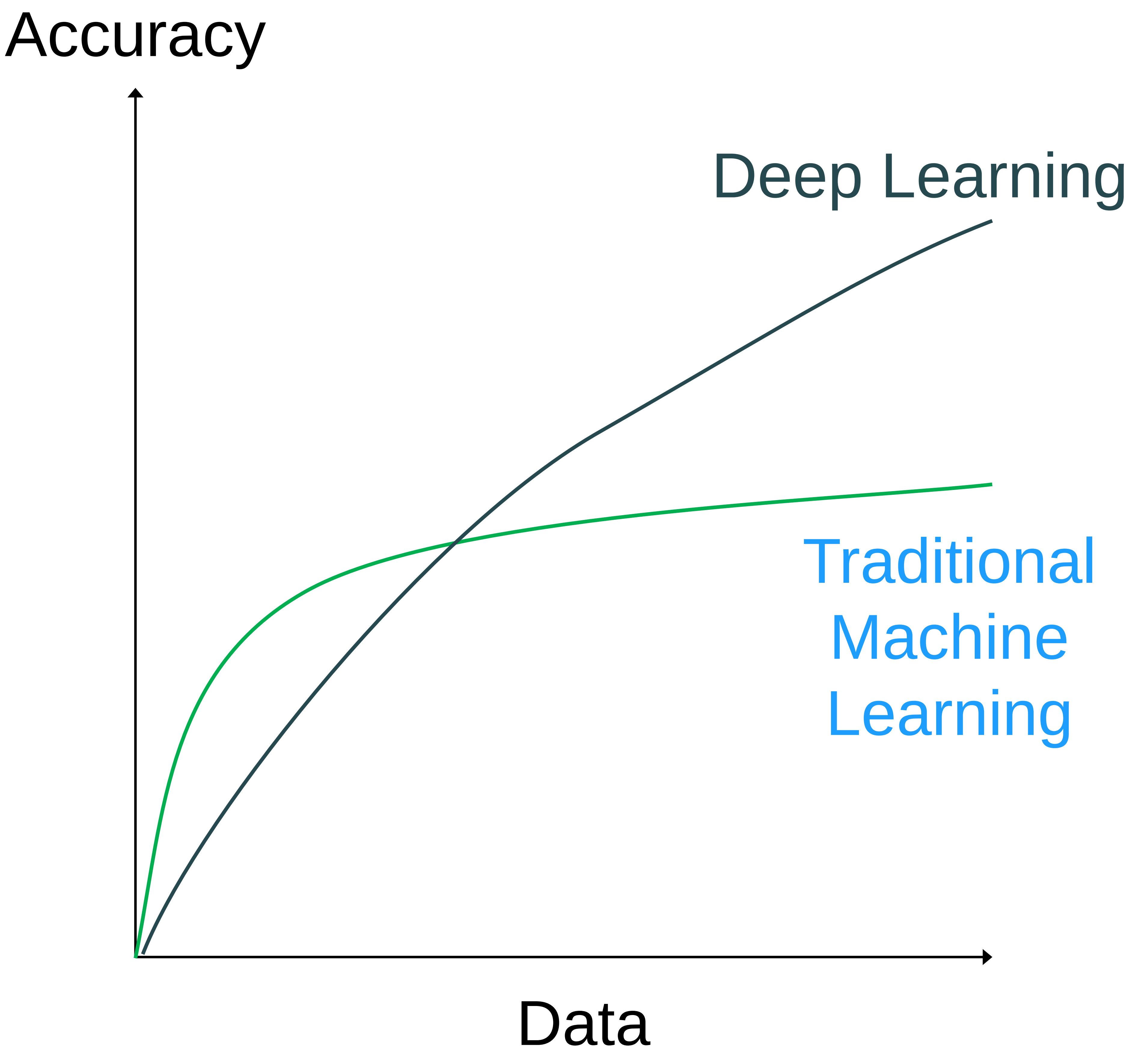


Input

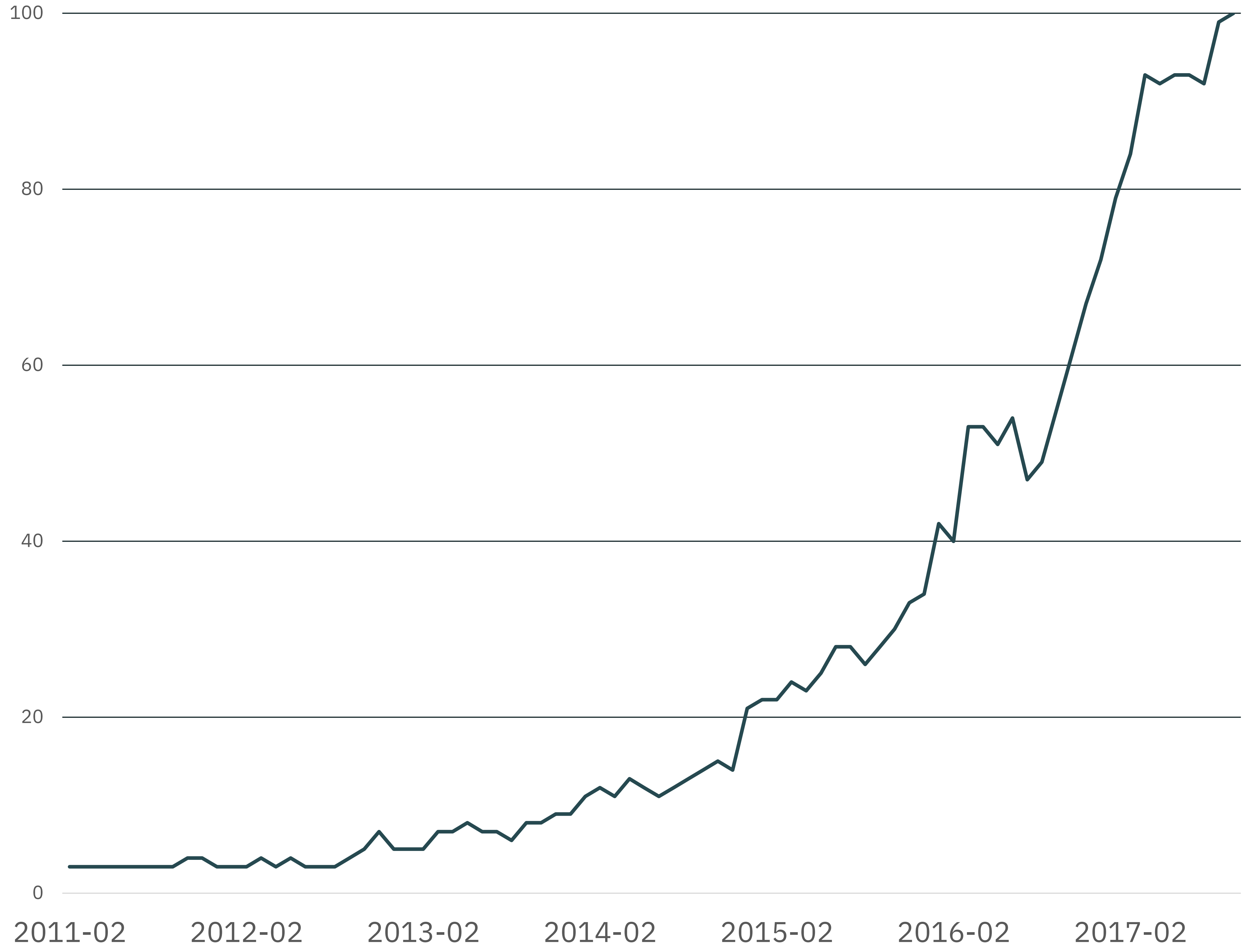
Feature Extraction & Classification

Output

Deep Learning Has Revolutionized Machine Learning

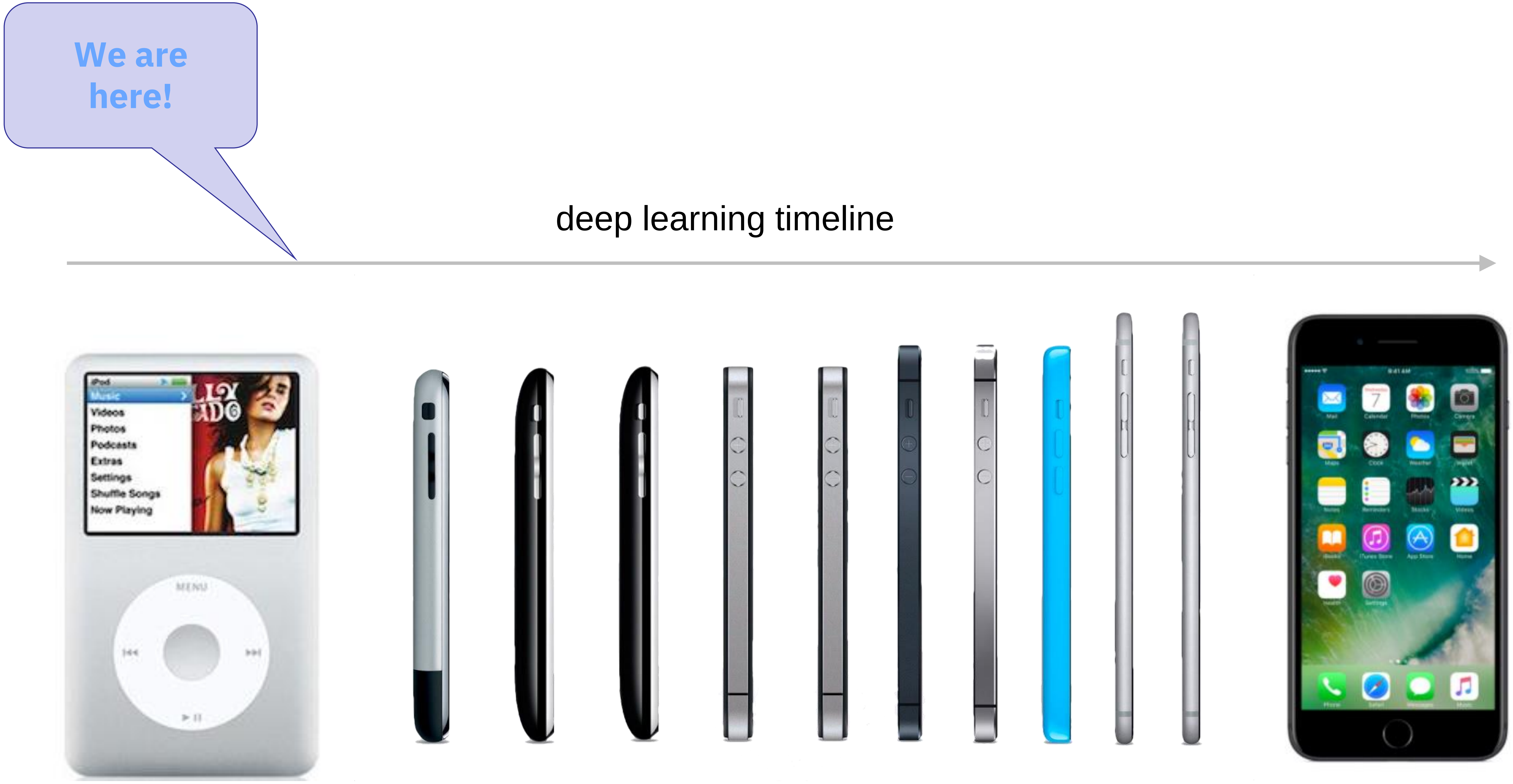


# of Searches for Deep Learning from 2011 to 2017



Source: Google Trends. Search term “Deep Learning”

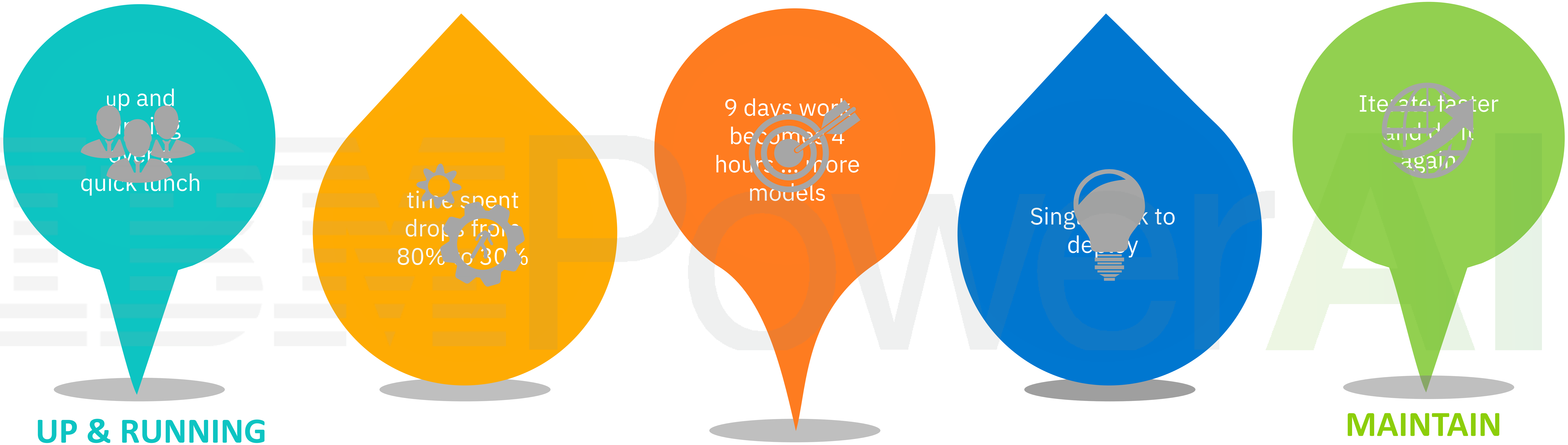
More iPod than iPhone X



We are here!

deep learning timeline

Realizing business value comes with its challenges



up and running over a quick lunch

time spent drops from 80% to 30%

9 days work becomes 4 hours … more models

Single click to deploy

Iterate faster and do it again

UP & RUNNING
weeks to months

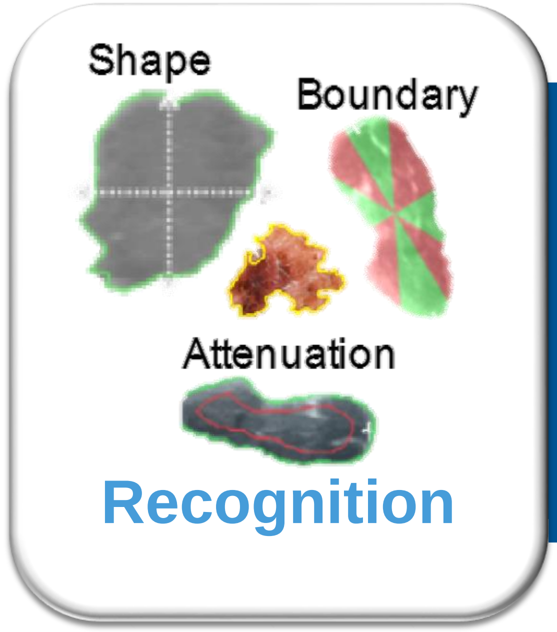
DATA PREPARATION
most time spent here

BUILD, TRAIN, OPTIMIZE
very iterative

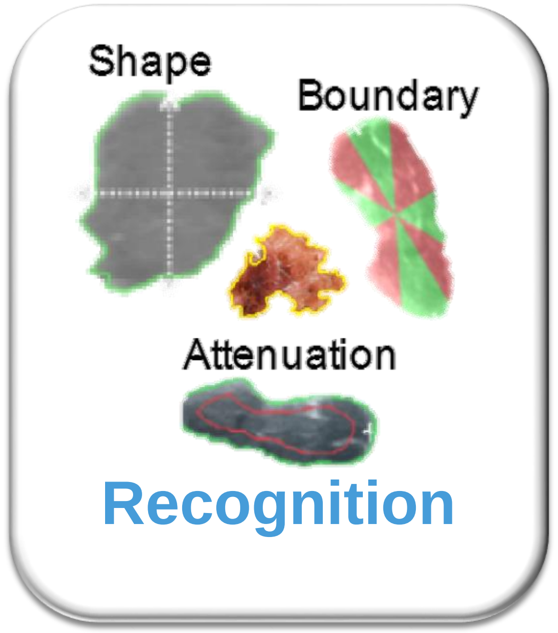
DEPLOY & INFER
requires different skills

MAINTAIN ACCURACY
experience all that pain again

Faster Training Time with Distributed Deep Learning



9 Days



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

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

4 Hours | 4 Hours | 4 Hours | 4 Hours

100x

Learning runs with Power 9*

54x

Learning runs with Power 8

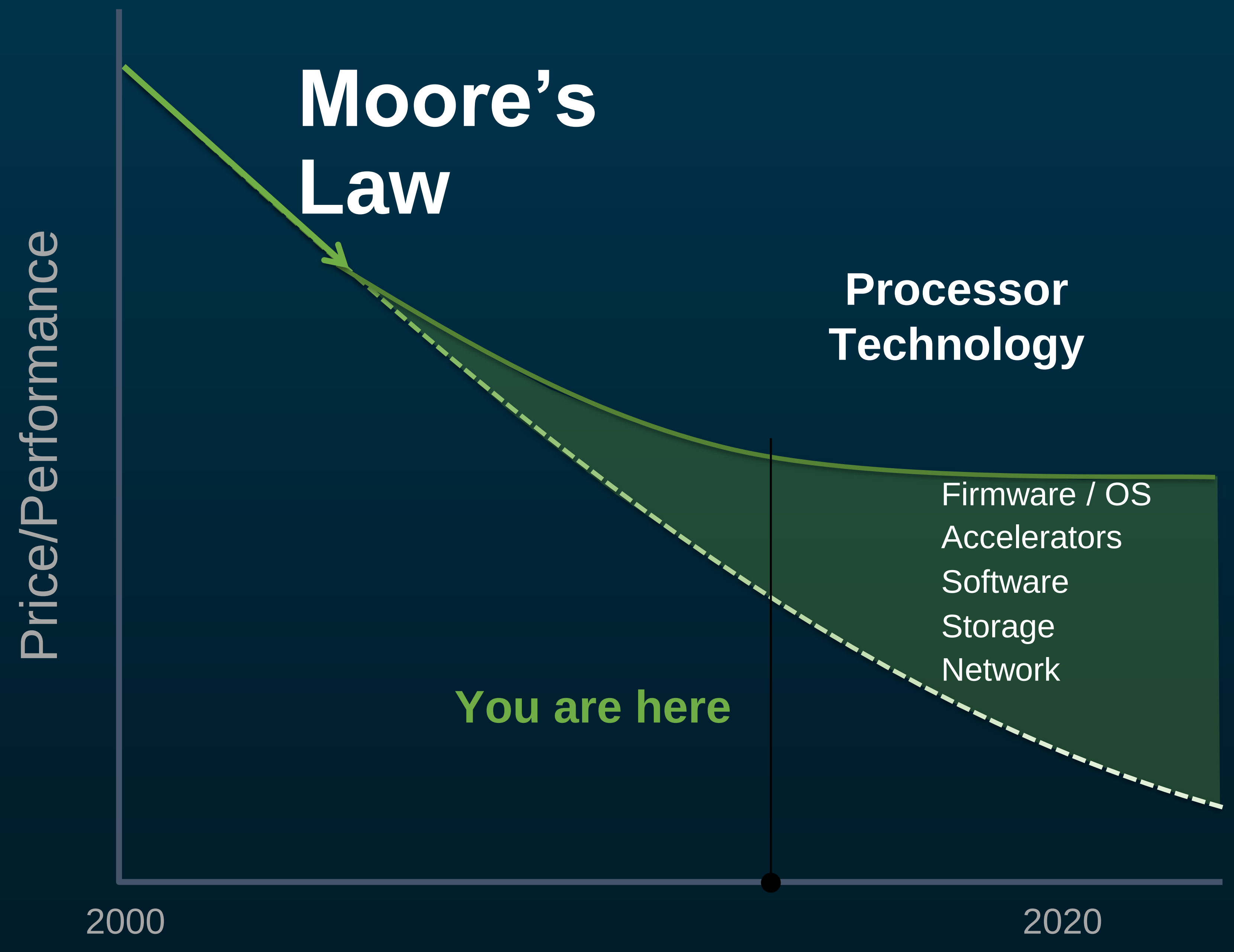
IBM PowerAI

Houston, we have a problem...

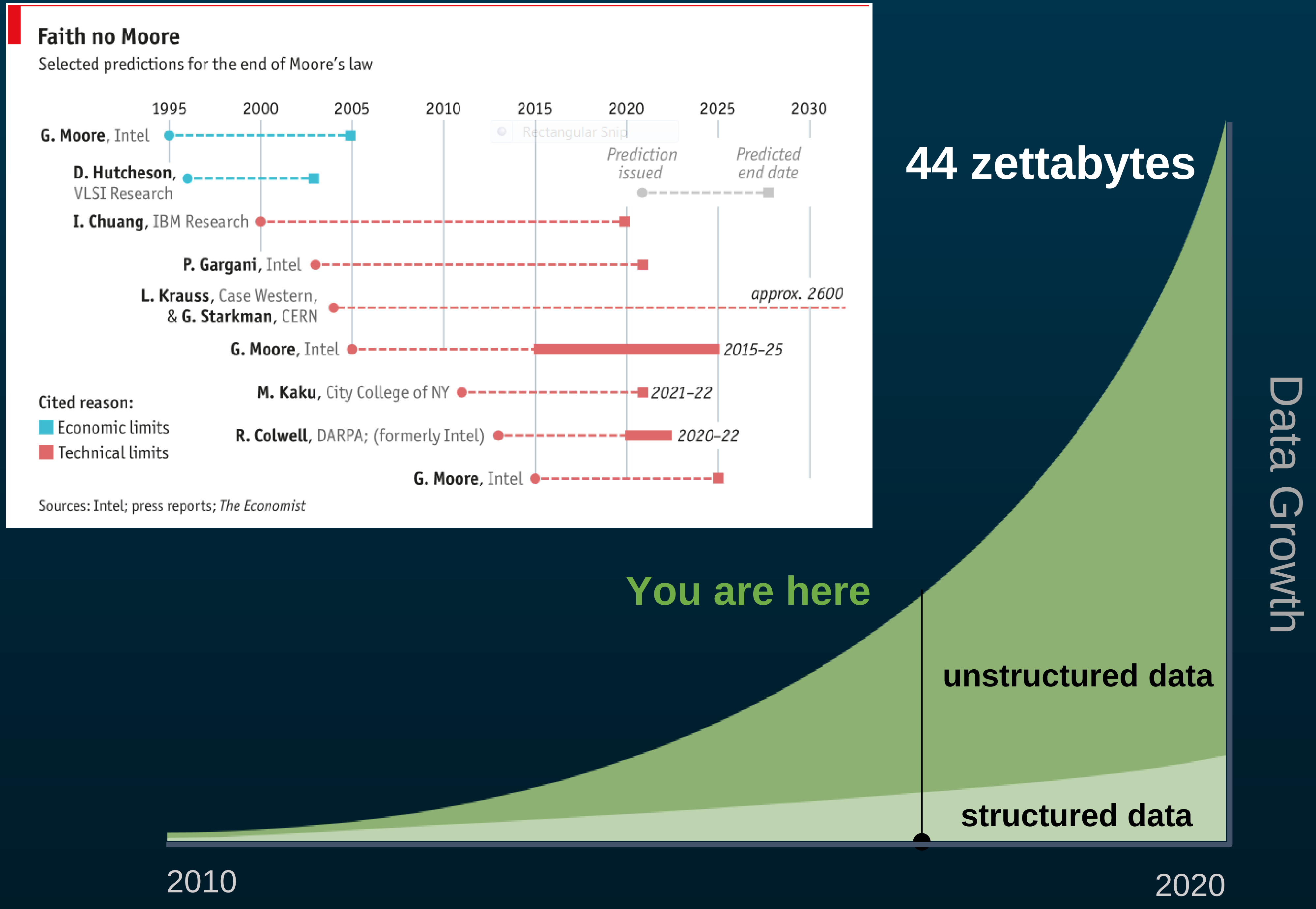
OpenPOWER™

Today’s challenges demand innovation

Full system and stack open innovation required



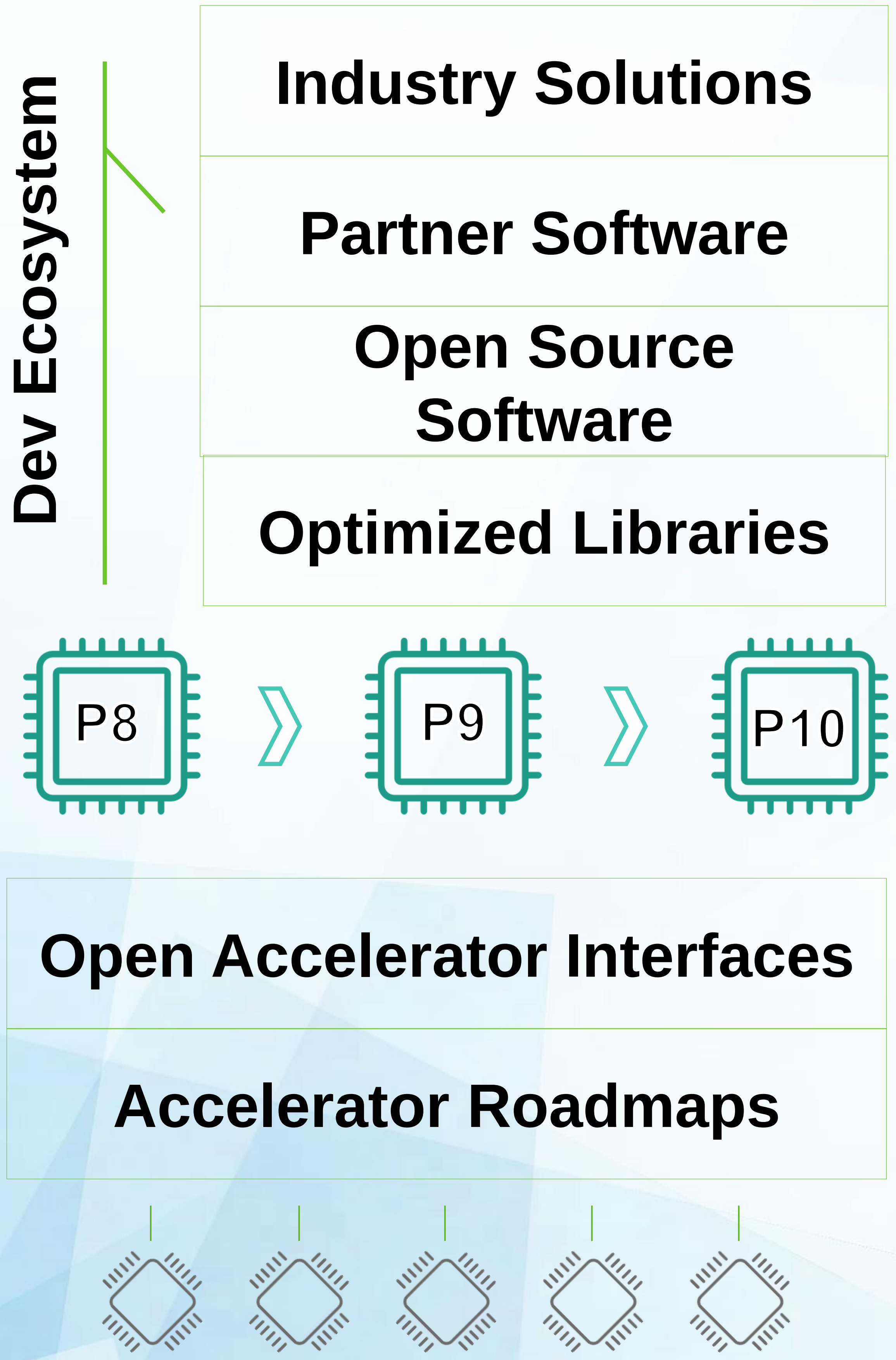
Data holds competitive value


Faith no Moore

Selected predictions for the end of Moore’s law

Sources: Intel; press reports; *The Economist*

Cognitive Systems are built with optimized hardware and software



mongoDB

APACHE Spark

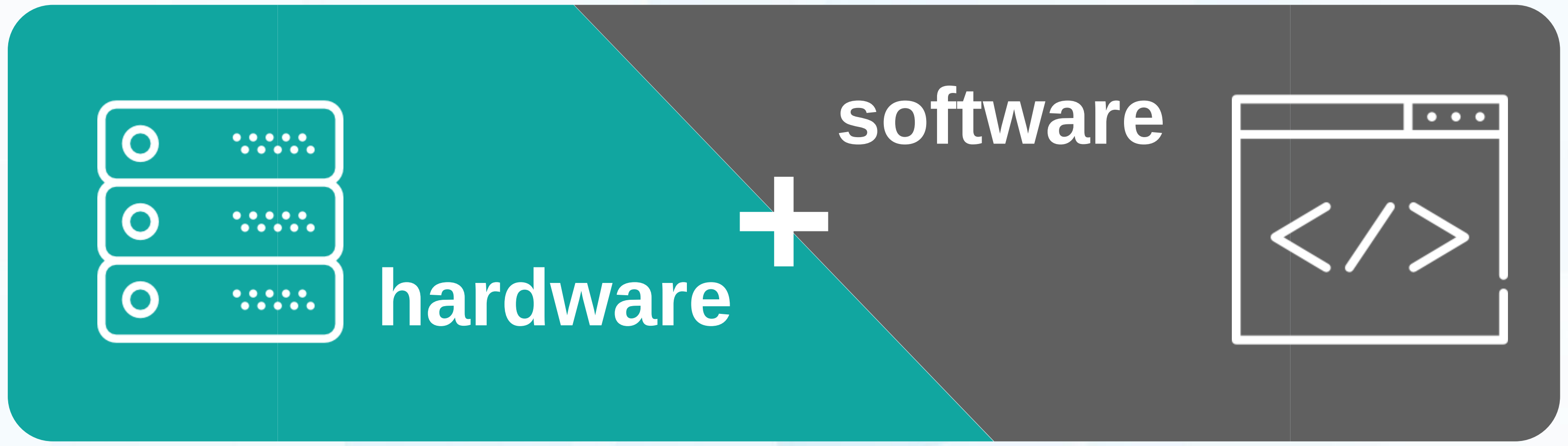
hadoop

EDB POSTGRES

cassandra

Not Just About Hardware Design

It's about co-optimized



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

TensorFlow™ Caffe theano

torch Chainer

IBM Power Systems: open to the core

OpenPOWER

>340 members

Mellanox TECHNOLOGIES

XILINX

NVIDIA.

Open Source Workloads

mongoDB.

APACHE Spark

hadoop

EDB POSTGRES

cassandra

OpenCAPI

AMD

Hewlett Packard Enterprise

Google

DELL

Open Frameworks

torch

theano

Chainer

TensorFlow™

Caffe

Making machine learning and AI more affordable

OpenPOWER™

This is what a r-evolution looks like

Implementation / HPC / Research

ASU Arizona State University
ASTRI
Agency for Science, Technology and Research SINGAPORE
CAMBRIDGE CARES
Carnegie Mellon University
cfms
CINECA
GSIC
Hartree Centre
ICCS-NTUA
iiit-b
ickm
JÜLICH FORSCHUNGSZENTRUM
Lawrence Livermore National Laboratory
LSU Louisiana State University
M University of Michigan
NUS National University of Singapore
Nanyang Technological University
Oak Ridge National Laboratory
OSU Oregon State
OVH.COM
rackspace the #1 managed cloud company
Centre de Calcul ROMEO
Maison de la Simulation de Champagne-Ardenne
RICE Unconventional Wisdom
rzg
SASTRA UNIVERSITY
Sandia National Laboratories
SDSC San Diego Supercomputer Center
Symbiosis Institute of Computer Studies & Research
TEES Texas A&M Engineering Experiment Station
UF University of Florida
TACC
USC University of Southern California
清華大學 Tsinghua University
University of Arkansas
UNICAMP
University of Oregon
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

Software

American Megatrends
BYOSOFT
FIXSTARS
redhat
SUSE
FreeBSD
Google
SANMARCO INFORMATICA
gpudb
红旗 Linux
RedHadoop
redislabs
synerscope
Groupe T2i innovation informatique
ubuntu Supported by Canonical

System / Integration

ARROW ELECTRONICS, INC.
ASETEK
AVNET
Bull an atos company
CIARA Speed | Performance | Passion for Innovation
Cirrascale
CSPi
DRC A Security First Corp Company
E4 COMPUTER ENGINEERING
IBM
inspur 浪潮
Mark III Systems
Microway Technology you can count on since 1982.
NEC
Neu Cloud Oriental
OCF
ONE STOP SYSTEMS
PENGUIN COMPUTING
RAPTOR ENGINEERING
rikor
RTDS Technologies
STACK VELOCITY
TEXHOПPOM
UNISOURCE
YADRO
中太数据 ZOOM NETWORKS

I/O / Storage / Acceleration

ALGO-LOGIC
ALPHA DATA
AVAGO TECHNOLOGIES
BittWare FPGA COMPUTING SYSTEMS
blazegraph
BLUEBEE
BROCADE
Chelsio Communications
CONVEY COMPUTER
DataDirect NETWORKS
edico genome
EVERSPIN TECHNOLOGIES The MRAM Company
FUSION-IO
HGST a Western Digital company
HITACHI Inspire the Next
Inphi Think fast.
Interface Masters TECHNOLOGIES Innovative Network Solutions
MAXELER Technologies
Mellanox TECHNOLOGIES
MEMBLAZE
Micron
Microsemi
MYRICOM Network Products
Nallatech
NVIDIA
PMC
QLOGIC
SAMSUNG
SanDisk
Semptian 恒扬科技
SK hynix
SOLARFLARE
XILINX

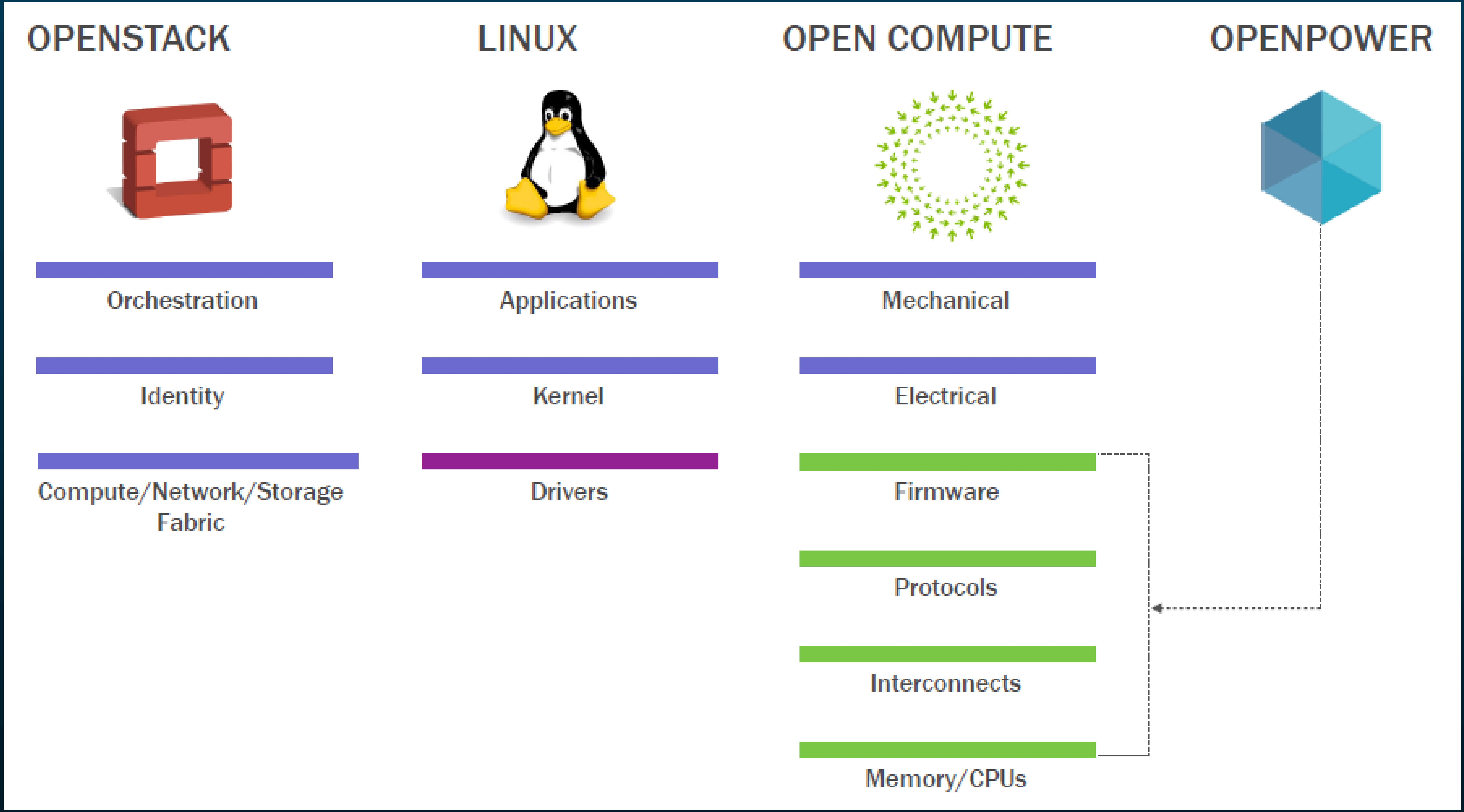
Boards / Systems

Celestica
创和通讯
IBM
Inventec
msi
柏飞电子
TYAN
wistron

Chip / SOC

IBM
IDT
Infineon
POWERCORE
SYNAPSE design
VeriSilicon

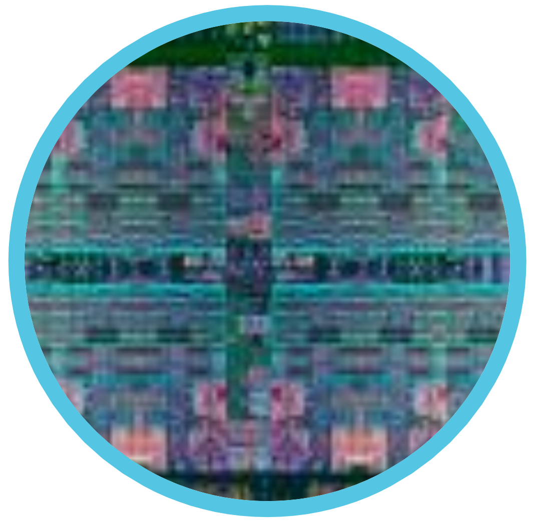
Cross community collaboration is essential



OPENSTACK	LINUX	OPEN COMPUTE	OPENPOWER
Orchestration	Applications	Mechanical	
Identity	Kernel	Electrical	
Compute/Network/Storage Fabric	Drivers	Firmware	
		Protocols	
		Interconnects	
		Memory/CPUs	

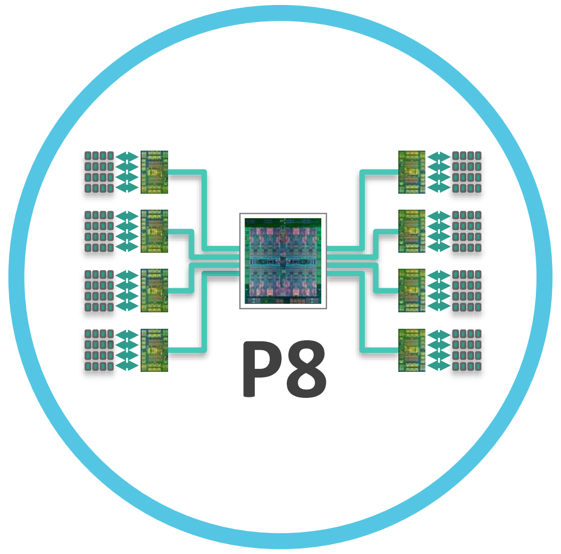
Balanced Systems Designed for Cognitive

Power high-performance core, bandwidth, accelerator differentiation



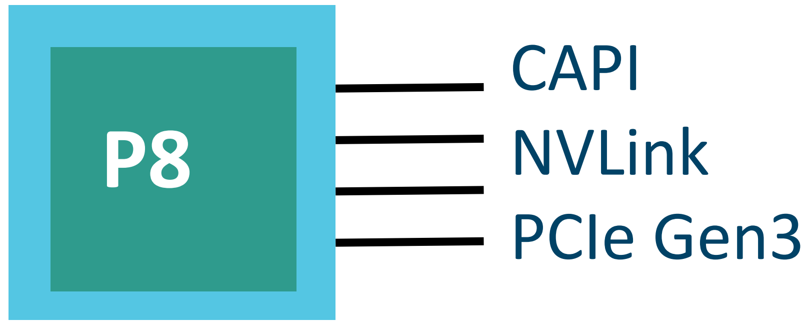
High Performance Cores

>2X vs. Intel



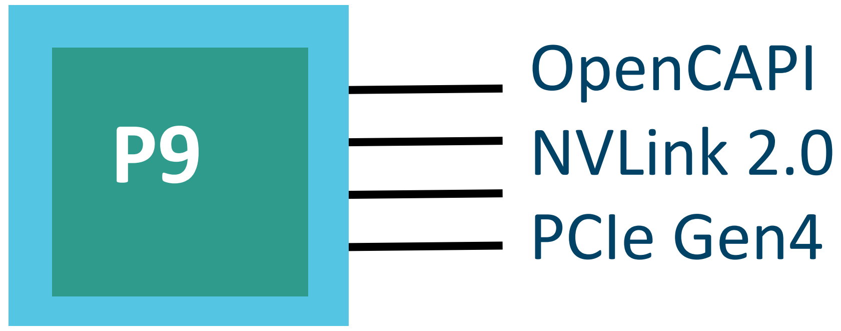
Fast & Large Memory System

4-5X Memory Bandwidth
3X Cache vs. Intel



Fast PowerAccel Interconnects for Accelerators

NVLink 1.0
5x vs. Intel



Faster PowerAccel Interconnect for Accelerators

OpenCAPI / NVLink 2.0
9.5x vs. Intel

Power partnerships with accelerator industry players:
Nvidia GPU, Xilinx, AMD GPU, open to other options

FROM HERE TO AI

Our 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

Big Data Workloads

Enterprise AI Workloads

S922/S914/S924,
H922/H924,
L922

E950/H950

E980/H980

LC922/LC921

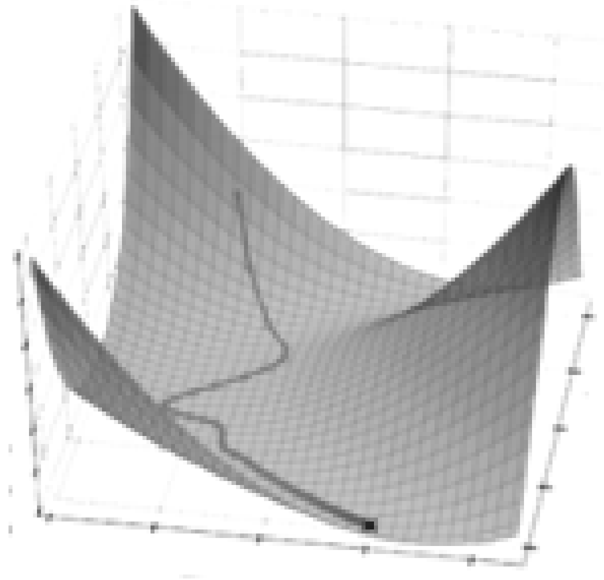
AC922

Core Infrastructure

Next Gen AI Workloads

enterprise-ready
software distribution
built on open source

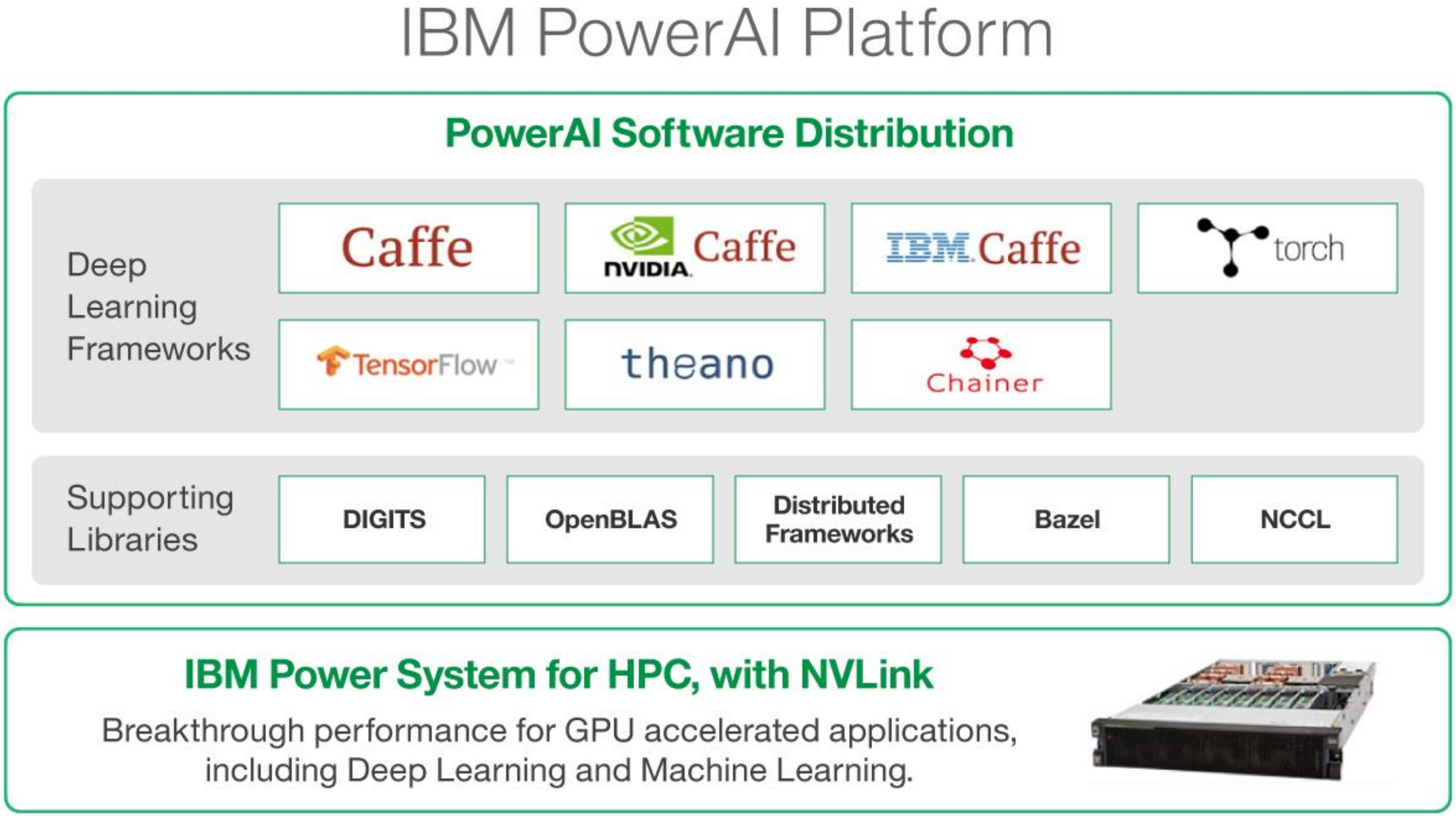
performance
faster training times
for data scientists



tools for ease
of development

IBM PowerAI

Enterprise Ready Build on Open Source



IBM PowerAI Platform

PowerAI Software Distribution

Deep Learning Frameworks

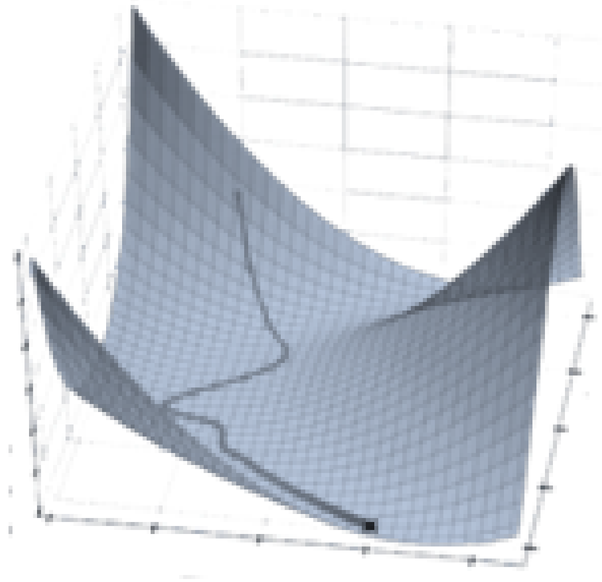
- Caffe
- NVIDIA Caffe
- IBM Caffe
- torch
- TensorFlow™
- theano
- Chainer

Supporting Libraries

- DIGITS
- OpenBLAS
- Distributed Frameworks
- Bazel
- NCCL

IBM Power System for HPC, with NVLink

Breakthrough performance for GPU accelerated applications, including Deep Learning and Machine Learning.



Tools for Ease of Development

rich advisory and building toolsets to flatten time to value

AI Vision
rich toolset image recognition neural networks

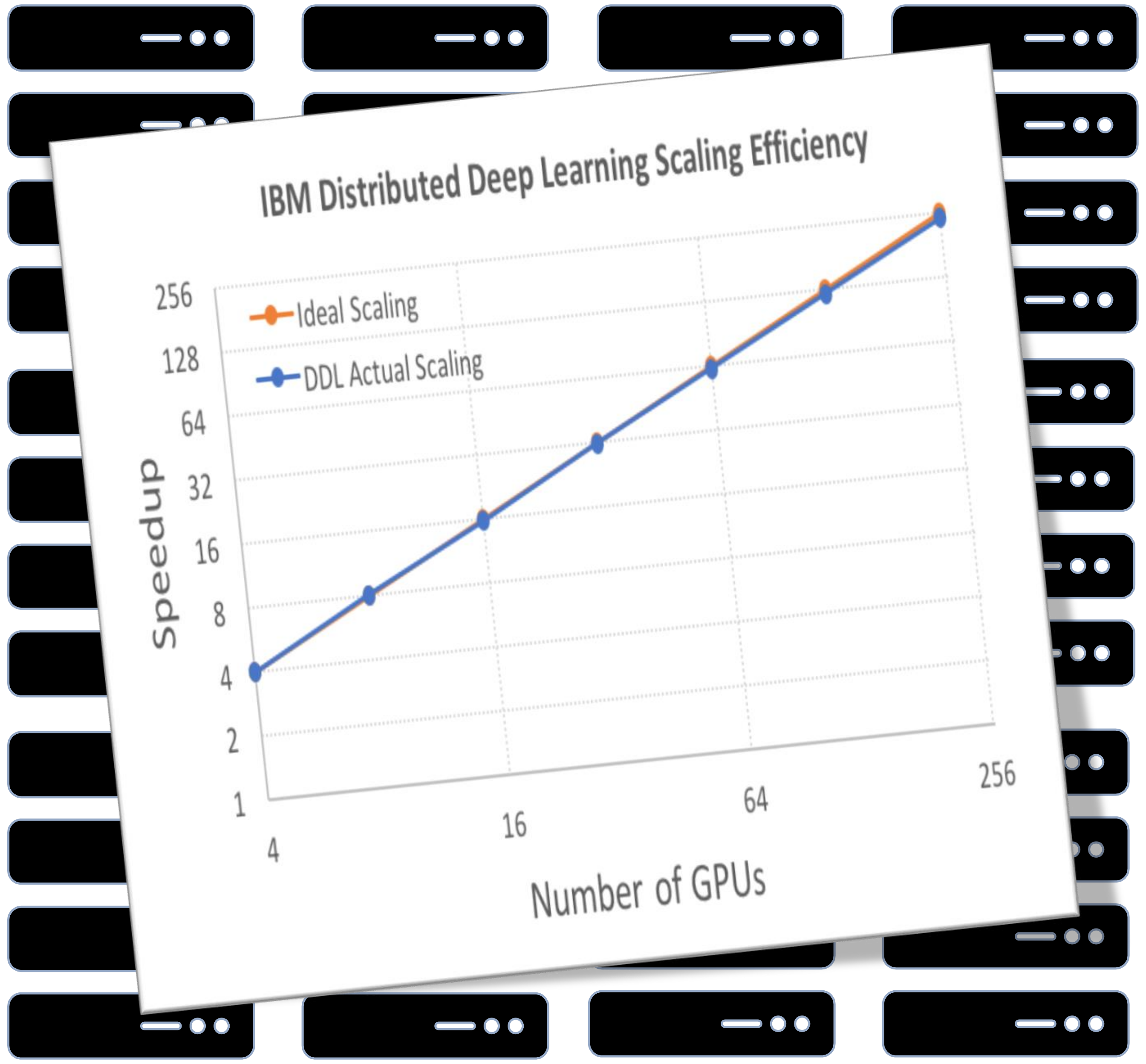
automated deep learning toolkit data preparation

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

Performance… Faster Training and Inferencing

faster training times for data scientists

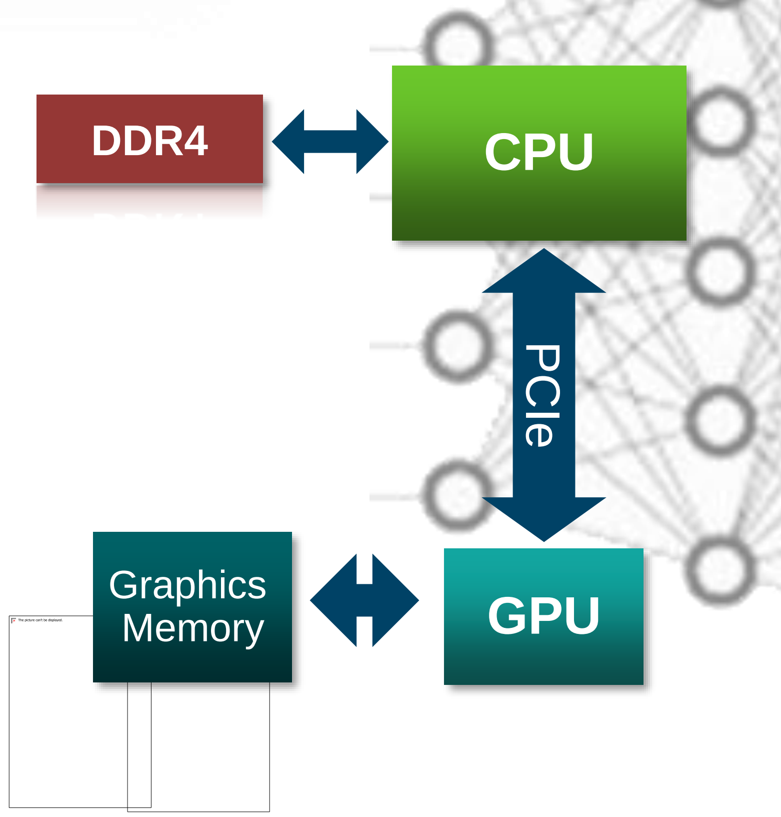
Distributed Deep Learning



Traditional Model Support → *Large Model Support (LMS)*

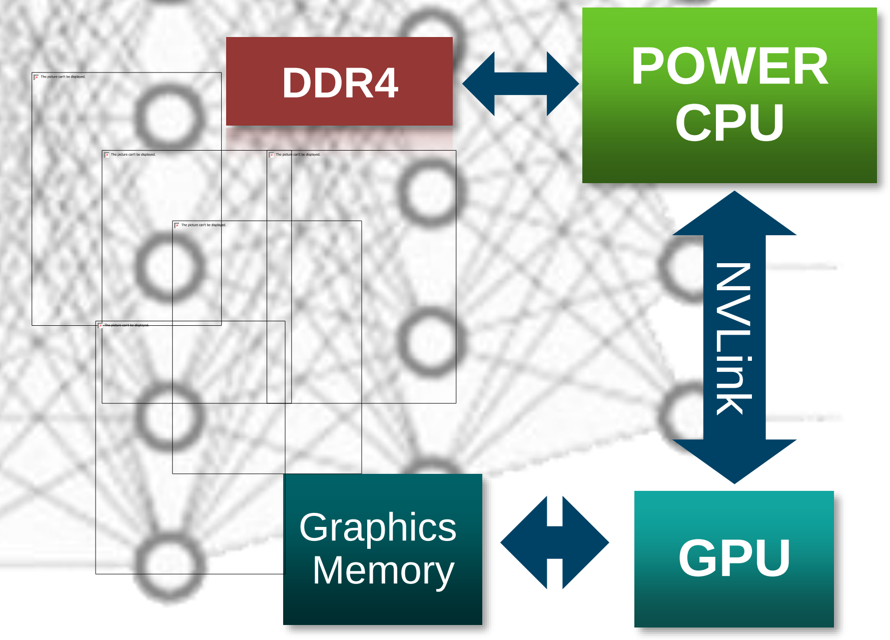
(Competitors)

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

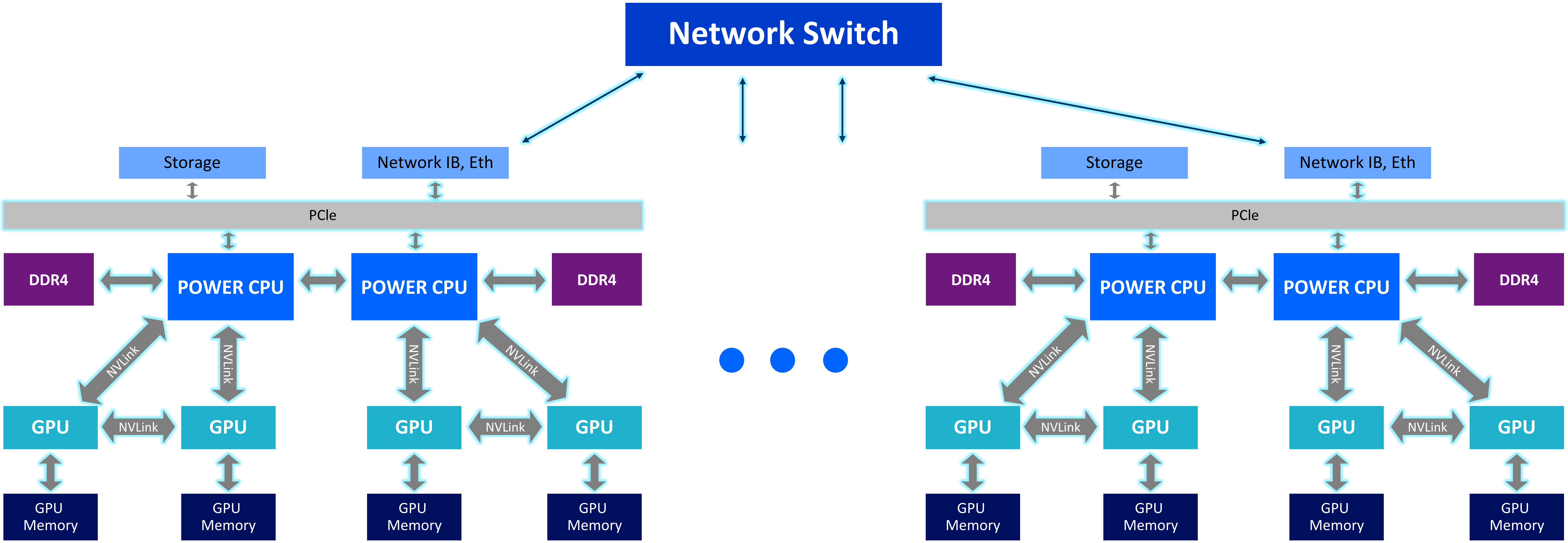


(PowerAI)

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

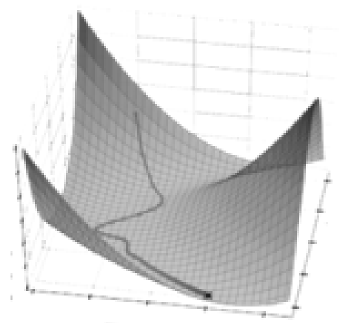


COMMUNICATION PATHS



Power AI DDL: Fully utilize bandwidth for links within each node and across all nodes

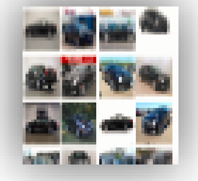
→ Learners communicate as efficiently as possible



Deep Learning Impact: Monitor, Adviser and Optimizer

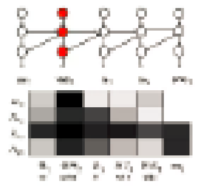
Deep Learning Applications

Image recognition

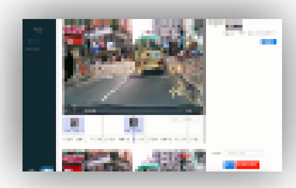


Object detection

Translation



Others

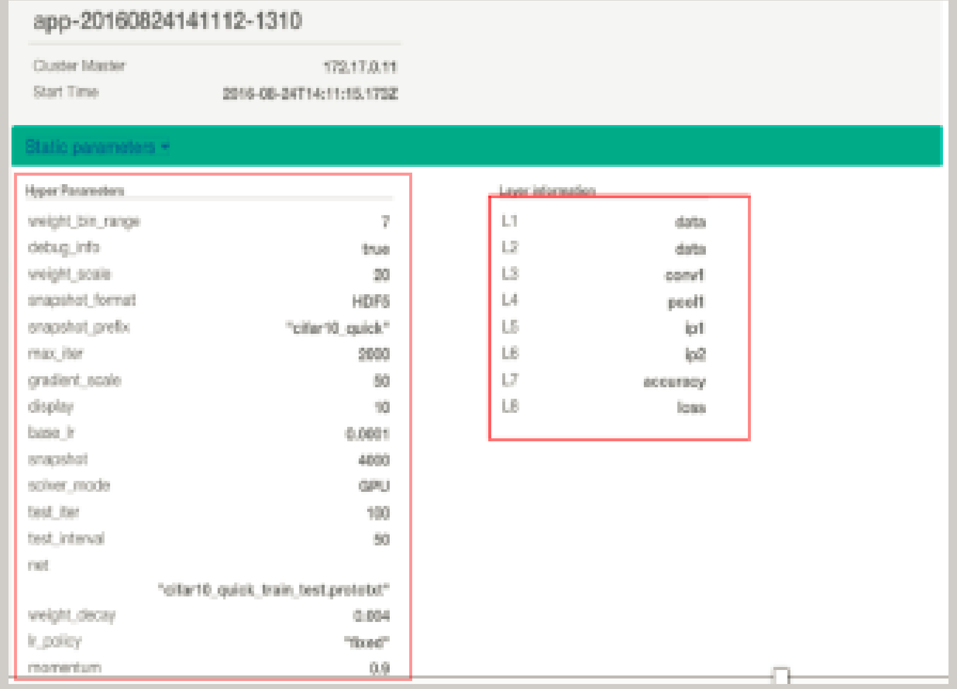


Optimize

Hyper-parameters Optimizer

optimizing procedure parameters

criterion and model parameters



app-20160824141112-1310

Cluster Master	172.17.0.11
Start Time	2016-08-24T14:11:15.173Z

Hyper Parameters

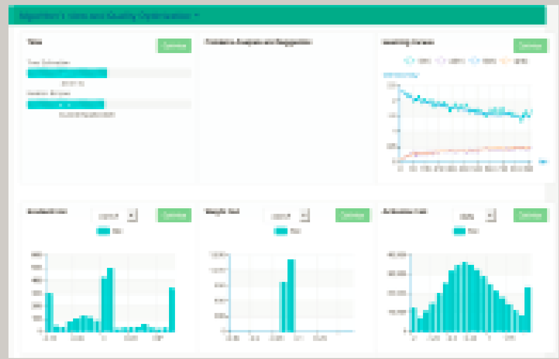
weight_bin_range	7
debug_info	true
weight_scale	20
snapshot_format	HDF5
snapshot_prefix	"cifar10_quick"
max_iter	2000
gradient_scale	50
display	10
base_lr	0.0001
snapshot	4000
solver_mode	GPU
test_iter	100
test_interval	50
net	"cifar10_quick_train_test.prototxt"
weight_decay	0.004
lr_policy	"fixed"
momentum	0.9

Layer information

L1	data
L2	data
L3	conv1
L4	pool1
L5	ip1
L6	ip2
L7	accuracy
L8	loss

Monitor

Real-time monitor for running application



learning curve

weight/gradient/activation histogram and norm

worst cases of training samples

Advice

Real-time adviser for running application

check the training process

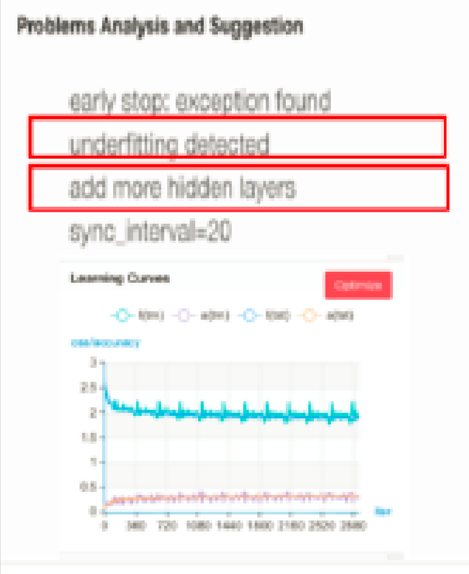
overflow

underfitting

overfitting

divergence

convergence



Problems Analysis and Suggestion

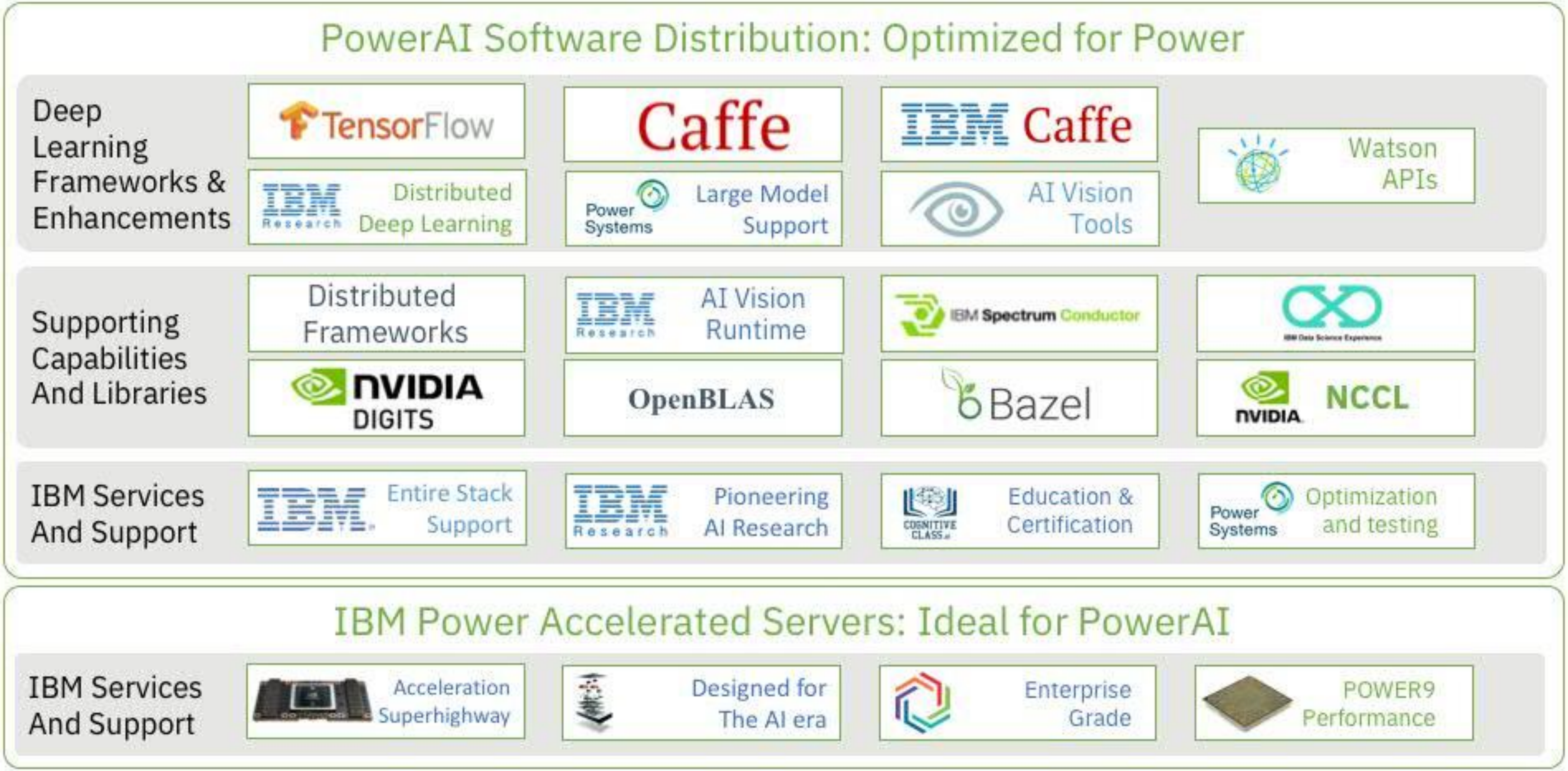
early stop: exception found

underfitting detected

add more hidden layers

sync_interval=20

IBM PowerAI Platform



PowerAI Software Distribution: Optimized for Power

Deep Learning Frameworks & Enhancements

- TensorFlow
- IBM Research Distributed Deep Learning
- Caffe
- Power Systems Large Model Support
- IBM Caffe
- AI Vision Tools
- Watson APIs

Supporting Capabilities And Libraries

- Distributed Frameworks
- NVIDIA DIGITS
- IBM Research AI Vision Runtime
- OpenBLAS
- IBM Spectrum Conductor
- Bazel
- IBM Data Science Experience
- NVIDIA NCCL

IBM Services And Support

- IBM Entire Stack Support
- IBM Research Pioneering AI Research
- COGNITIVE CLASS.ai Education & Certification
- Power Systems Optimization and testing

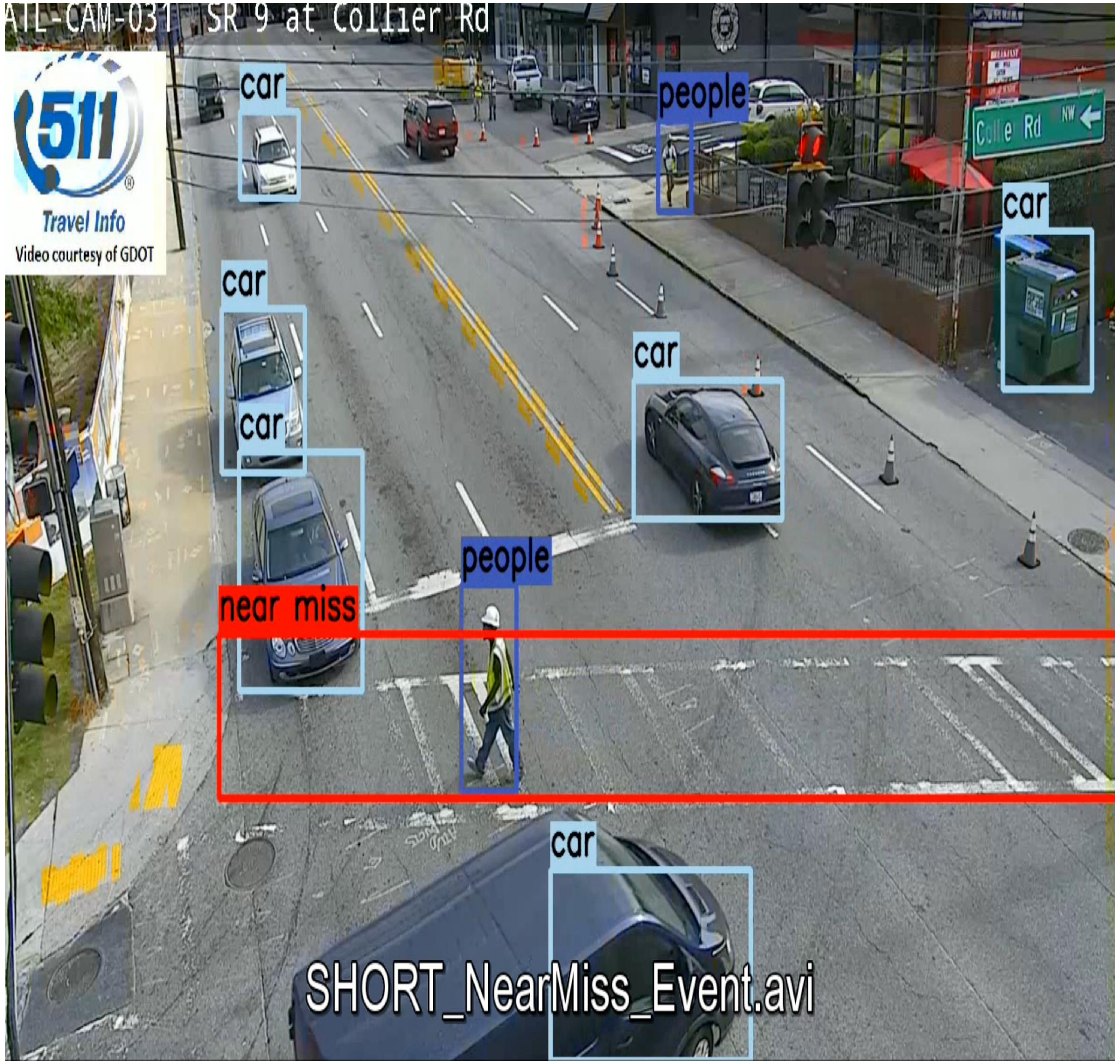
IBM Power Accelerated Servers: Ideal for PowerAI

IBM Services And Support

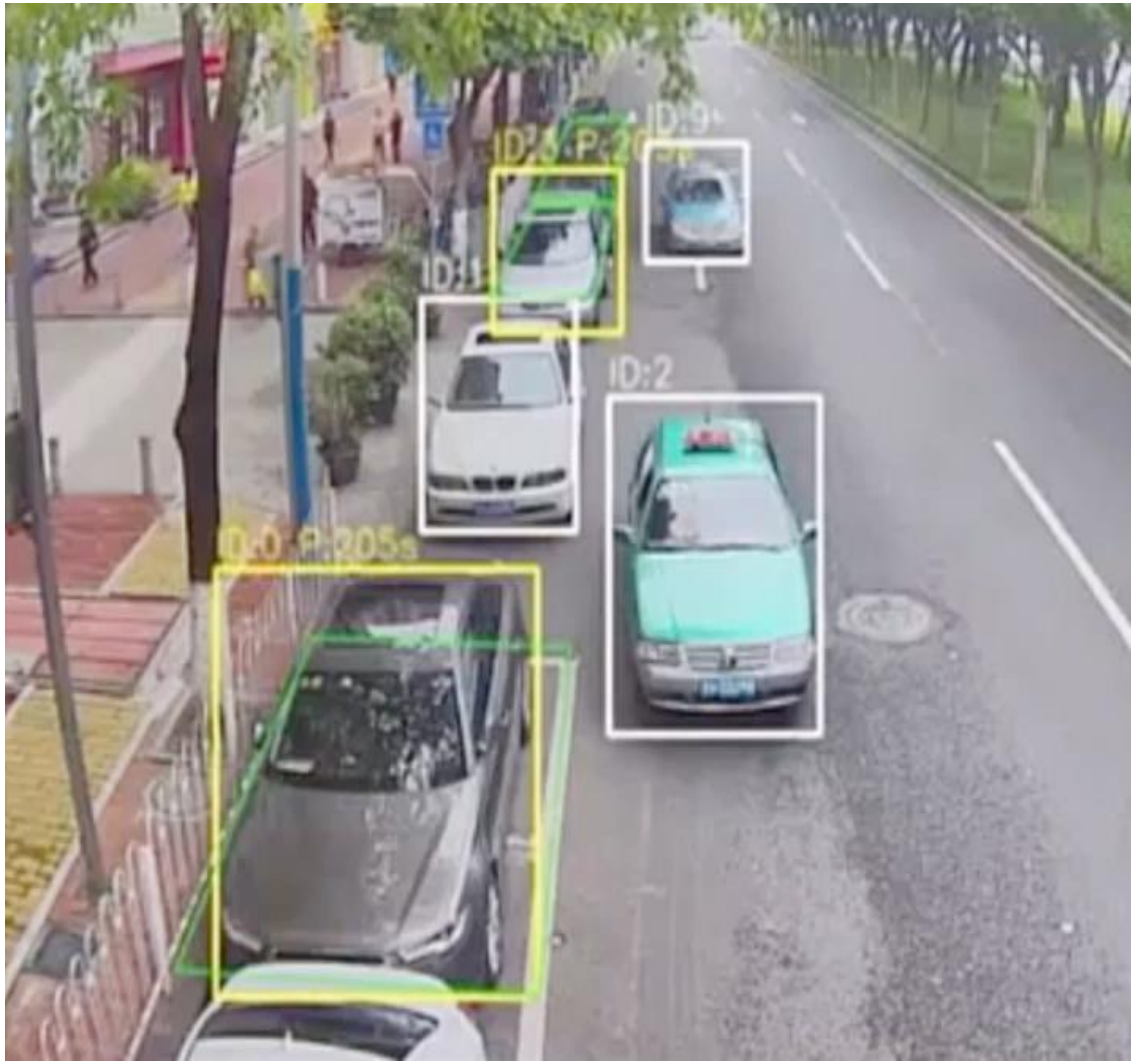
- Acceleration Superhighway
- Designed for The AI era
- Enterprise Grade
- POWER9 Performance

Smarter & Safer Cities

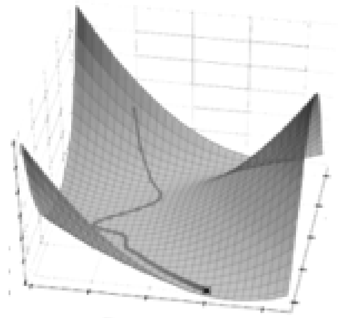
Near miss at intersections

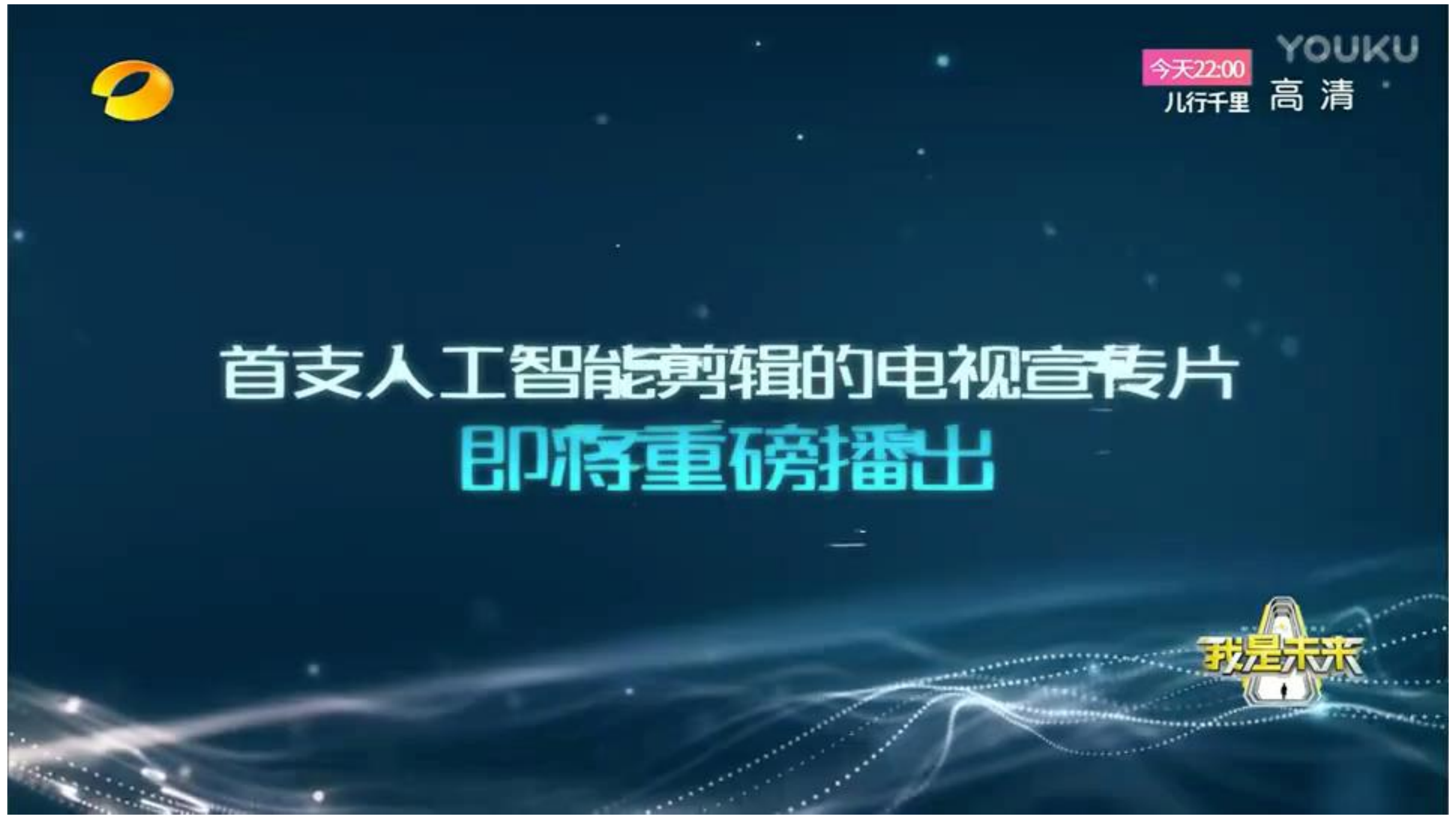


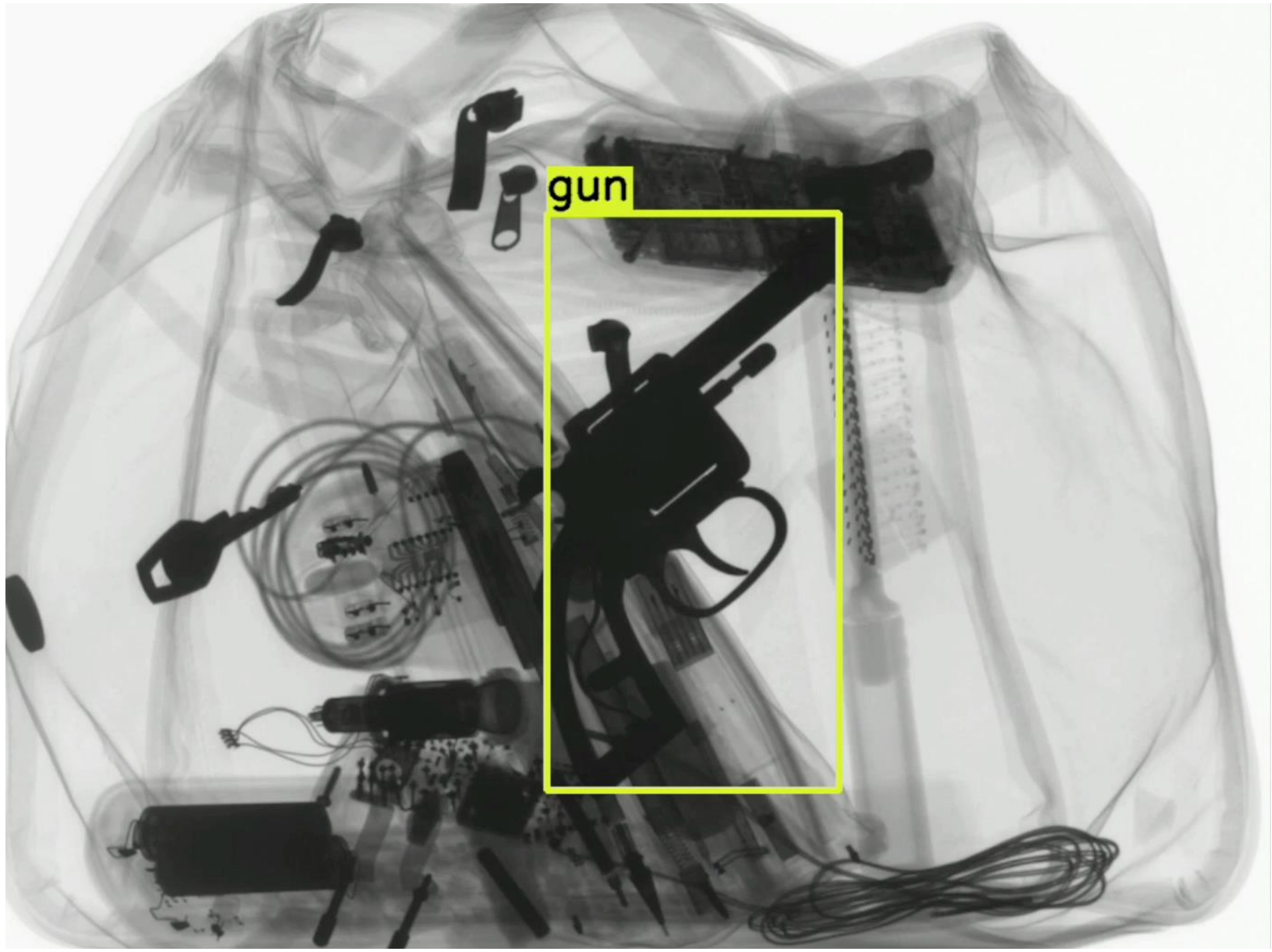
Monitor and Impose regulations



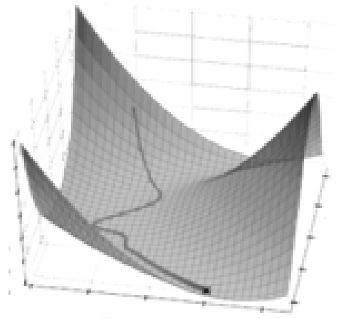
More Use Cases

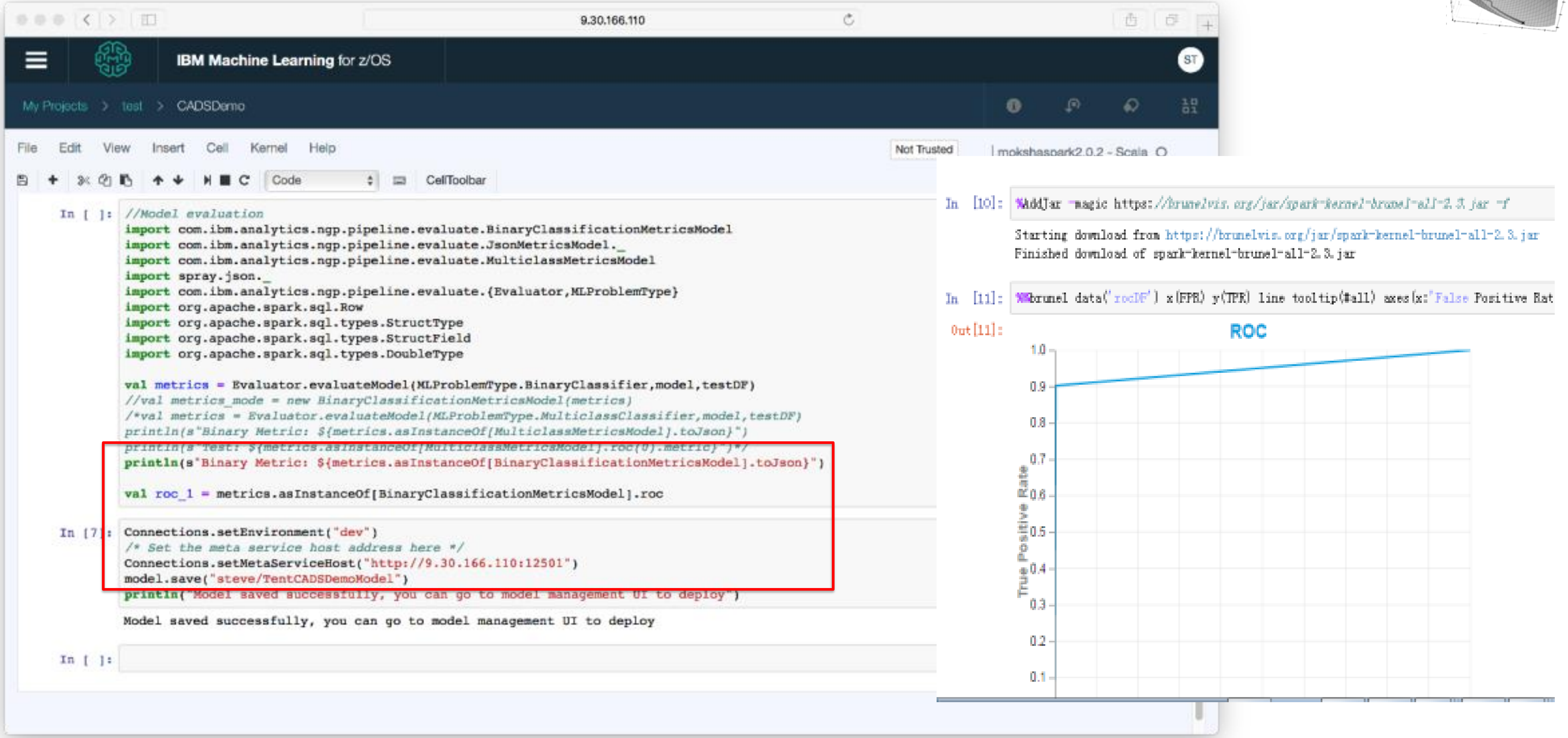






Programmatic Approach: Using Jupyter Notebooks





In [ ]:

```
//Model evaluation
import com.ibm.analytics.ngp.pipeline.evaluate.BinaryClassificationMetricsModel
import com.ibm.analytics.ngp.pipeline.evaluate.JsonMetricsModel._
import com.ibm.analytics.ngp.pipeline.evaluate.MulticlassMetricsModel
import spray.json._
import com.ibm.analytics.ngp.pipeline.evaluate.{Evaluator,MLProblemType}
import org.apache.spark.sql.Row
import org.apache.spark.sql.types.StructType
import org.apache.spark.sql.types.StructField
import org.apache.spark.sql.types.DoubleType

val metrics = Evaluator.evaluateModel(MLProblemType.BinaryClassifier,model,testDF)
//val metrics_mode = new BinaryClassificationMetricsModel(metrics)
/*val metrics = Evaluator.evaluateModel(MLProblemType.MulticlassClassifier,model,testDF)
println(s"Binary Metric: ${metrics.asInstanceOf[MulticlassMetricsModel].toJson}")
println(s"Test: ${metrics.asInstanceOf[MulticlassMetricsModel].roc(0).metric}")*/
println(s"Binary Metric: ${metrics.asInstanceOf[BinaryClassificationMetricsModel].toJson}")

val roc_1 = metrics.asInstanceOf[BinaryClassificationMetricsModel].roc
```

In [7]:

```
Connections.setEnvironment("dev")
/* Set the meta service host address here */
Connections.setMetaServiceHost("http://9.30.166.110:12501")
model.save("steve/TentCADSDemoModel")
println("Model saved successfully, you can go to model management UI to deploy")
```

```
Model saved successfully, you can go to model management UI to deploy
```

In [ ]:

In [10]:

```
%AddJar -magic https://brunelvis.org/jar/spark-kernel-brunel-all-2.3.jar -f
```

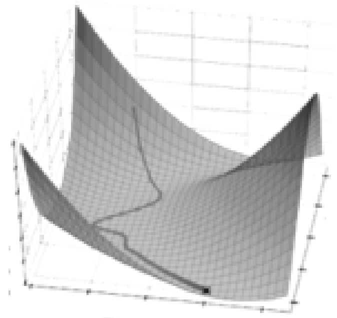
```
Starting download from https://brunelvis.org/jar/spark-kernel-brunel-all-2.3.jar
Finished download of spark-kernel-brunel-all-2.3.jar
```

In [11]:

```
%%brunel data('rocDF') x(FPR) y(TPR) line tooltip(#all) axes(x:'False Positive Rat
```

Out[11]:

IBM AI & Data Science Workbench

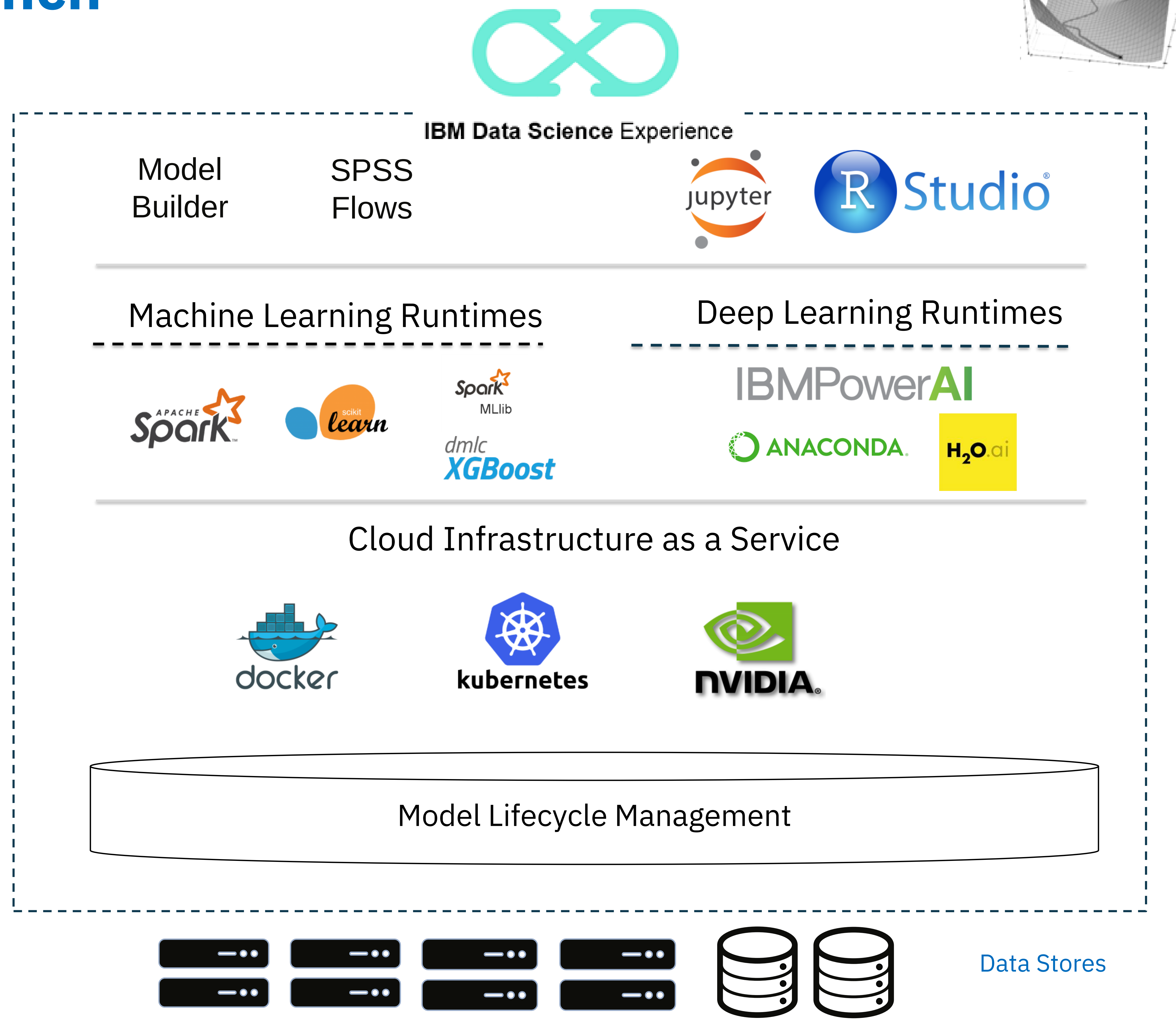


Democratizing ML/DL

Operationalizing ML/DL

Fit-for-purpose

Integration with Watson functions



IBM Data Science Experience

Model Builder

SPSS Flows

Machine Learning Runtimes

Deep Learning Runtimes

Cloud Infrastructure as a Service

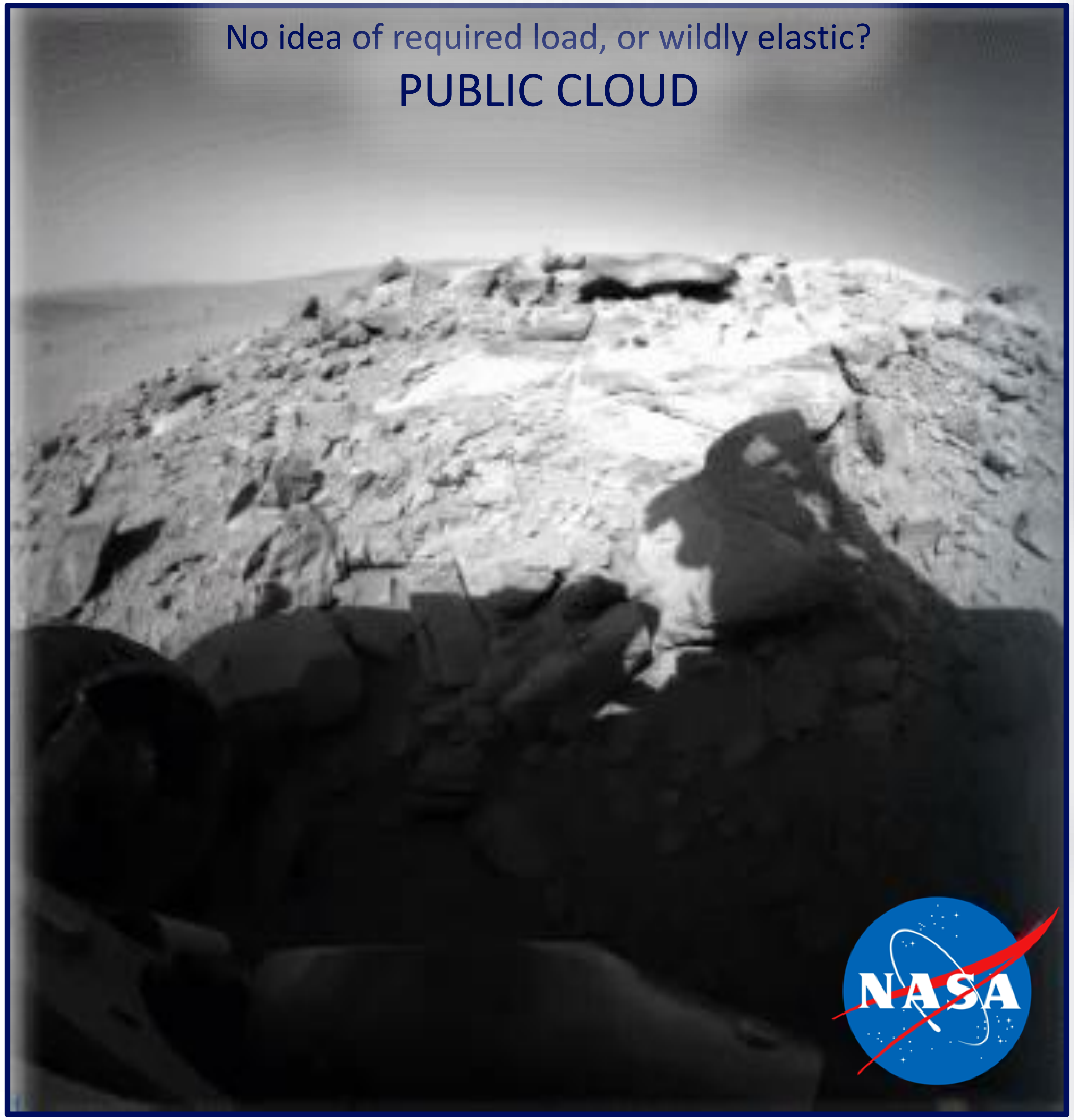
Model Lifecycle Management

Data Stores

Cognitive Systems

No idea of required load, or wildly elastic?

PUBLIC CLOUD



HOTELS AIRBNB

Staying in NYC for a weekend?

PUBLIC CLOUD

Staying in NYC 6 months?

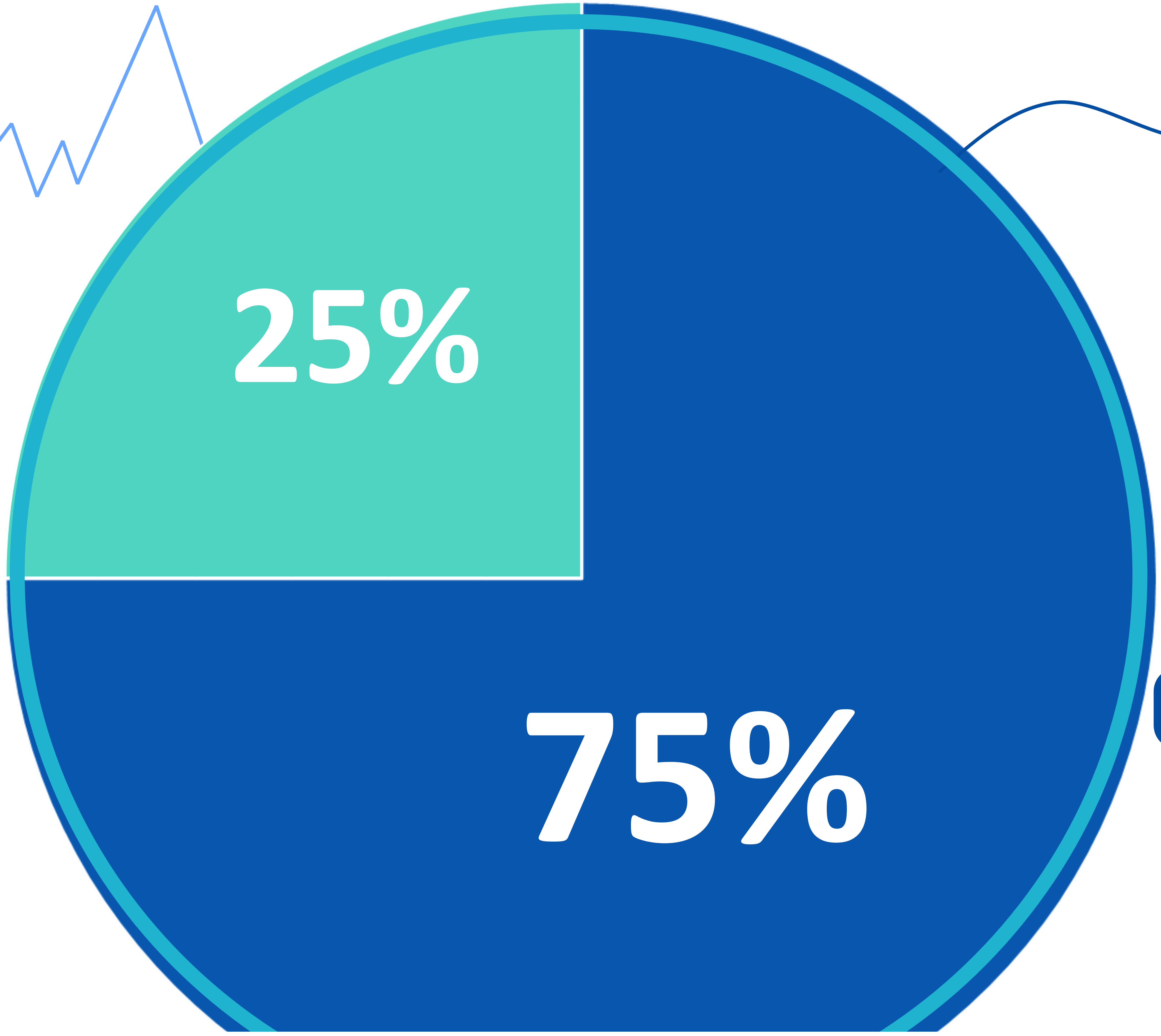
PRIVATE or PUBLIC?

Staying in NYC 3 years?

PRIVATE CLOUD

Elastic Workloads

Spin up and down resources
on the public cloud



Predictable Workloads

Lower costs with private cloud
infrastructure

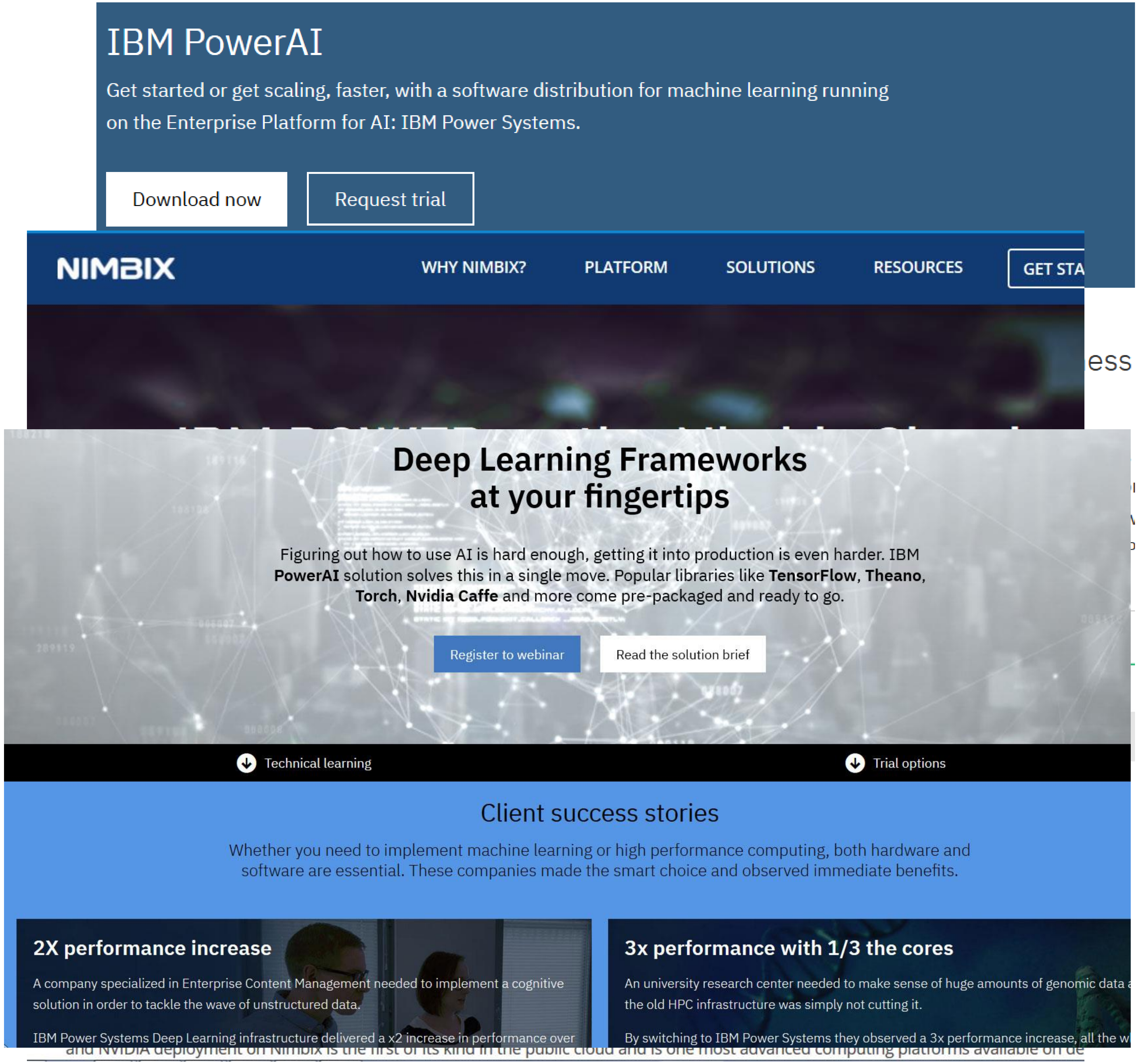
balance **owning** and **renting** for today’s enterprise workloads

What’s next ?

- **Getting PowerAI ?** **http://ibm.biz/powerai**

- **Testing PowerAI in the Cloud ?**
  https://power.jarvice.com/

- **Need for any further reference or contact ?**
  https://www-03.ibm.com/systems/uk/power/hardware/hpc/outthink.html



IBM PowerAI

Get started or get scaling, faster, with a software distribution for machine learning running on the Enterprise Platform for AI: IBM Power Systems.

Deep Learning Frameworks at your fingertips

Figuring out how to use AI is hard enough, getting it into production is even harder. IBM **PowerAI** solution solves this in a single move. Popular libraries like **TensorFlow**, **Theano**, **Torch**, **Nvidia Caffe** and more come pre-packaged and ready to go.

Client success stories

Whether you need to implement machine learning or high performance computing, both hardware and software are essential. These companies made the smart choice and observed immediate benefits.

2X performance increase

A company specialized in Enterprise Content Management needed to implement a cognitive solution in order to tackle the wave of unstructured data.

IBM Power Systems Deep Learning infrastructure delivered a x2 increase in performance over

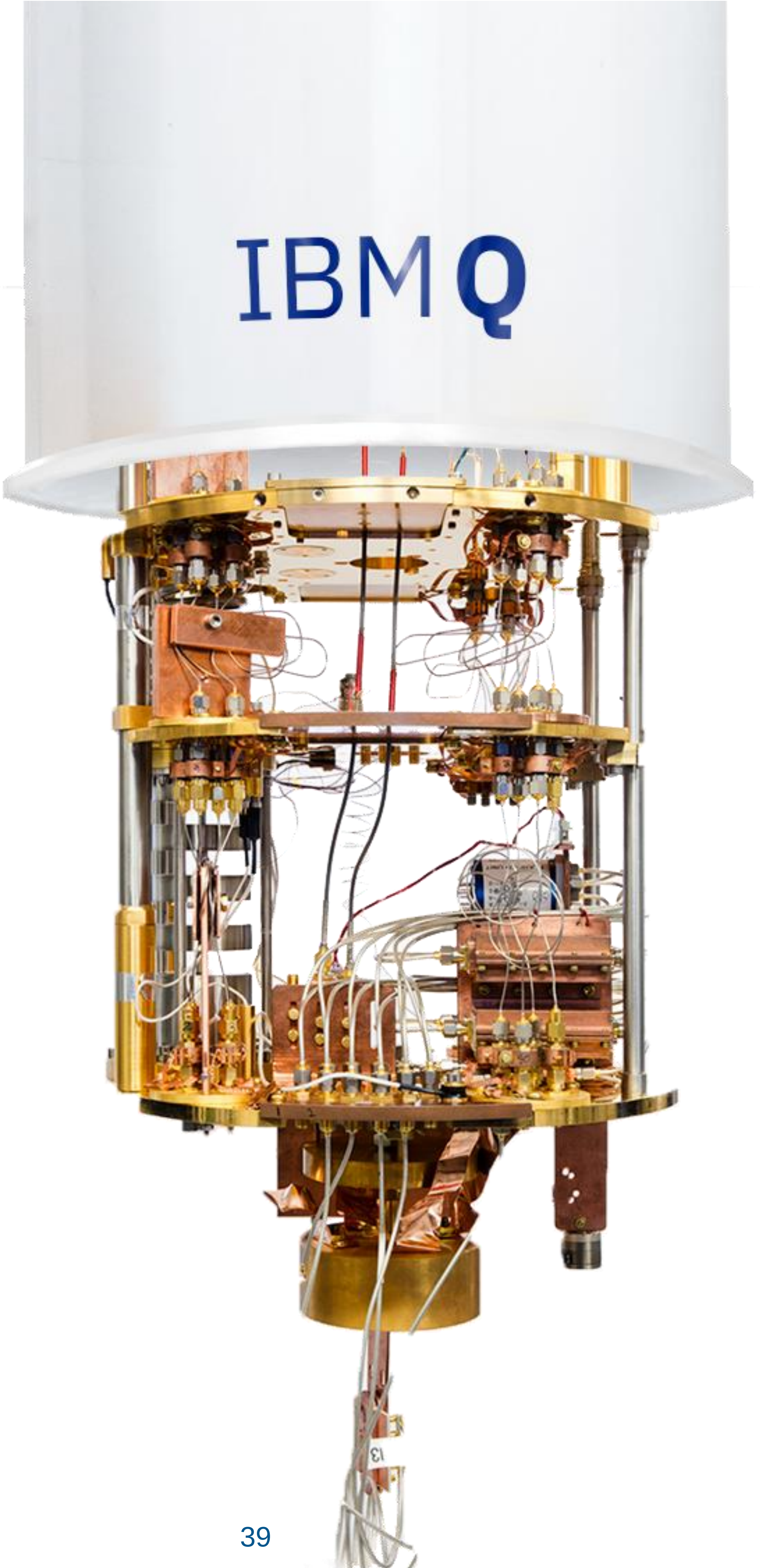
3x performance with 1/3 the cores

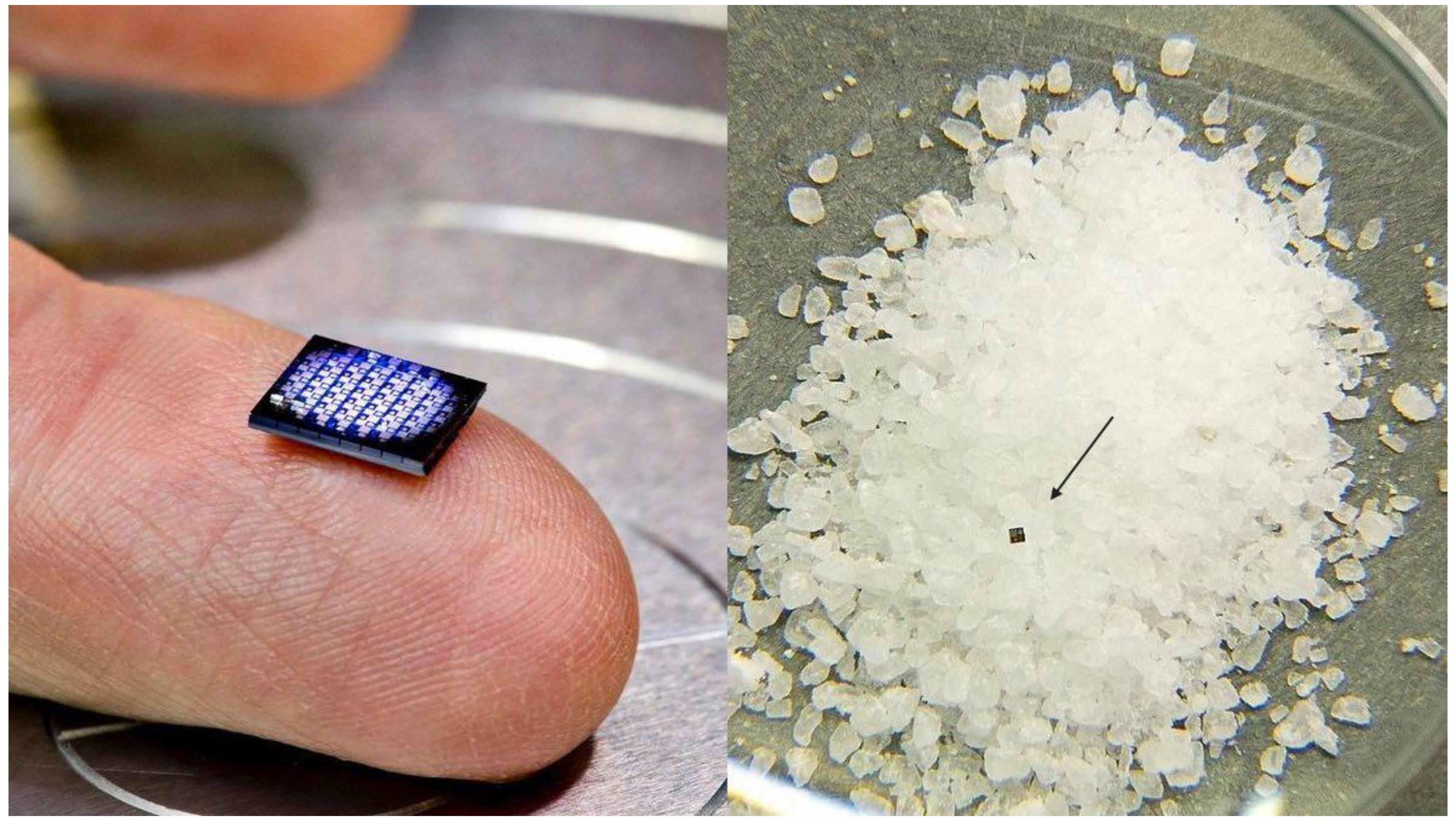
An university research center needed to make sense of huge amounts of genomic data a the old HPC infrastructure was simply not cutting it.

By switching to IBM Power Systems they observed a 3x performance increase, all the w

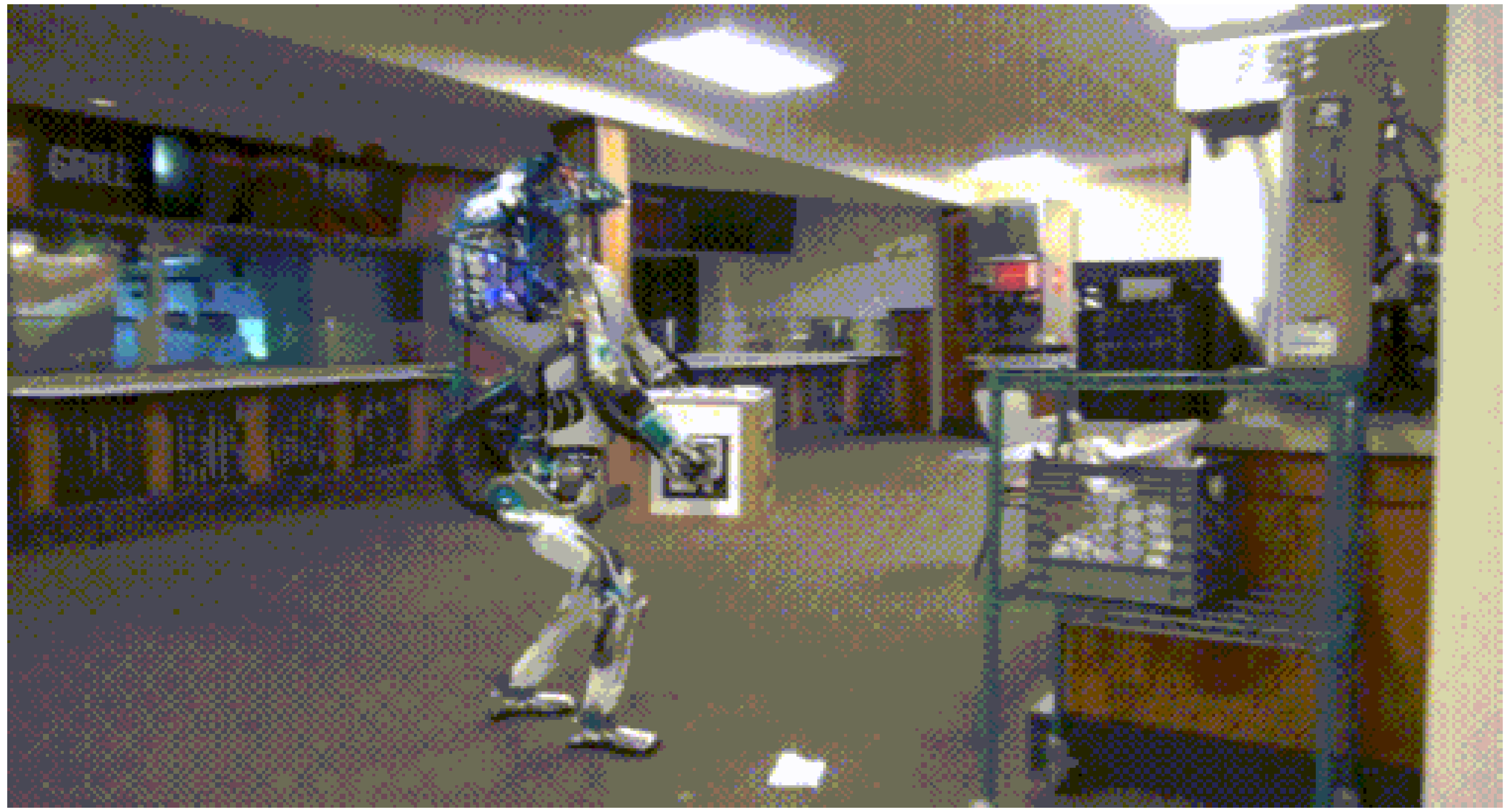
and NVIDIA deployment on Nimbix is the first of its kind in the public cloud and is one most advanced computing platforms available on de

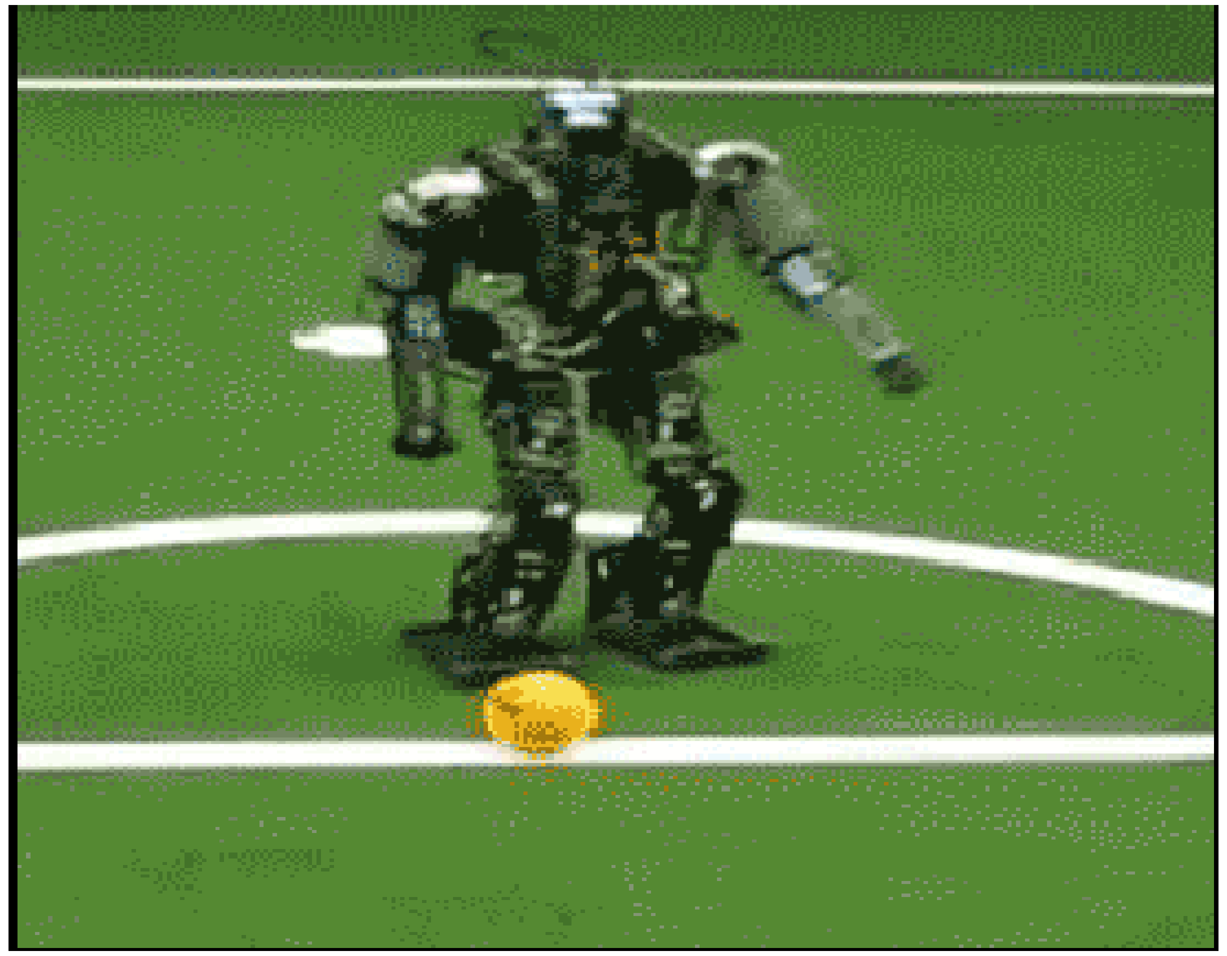
So, What’s really next ?



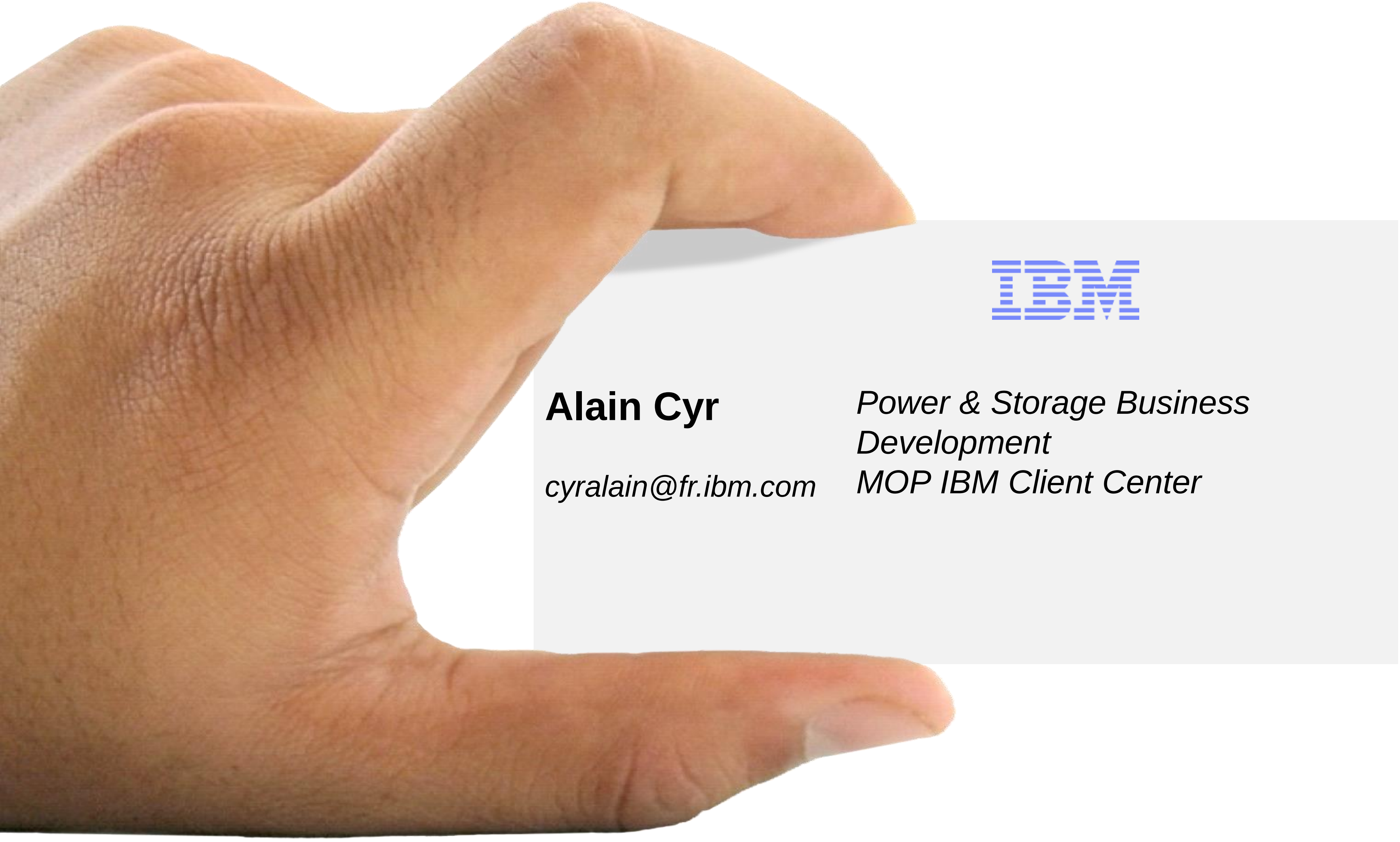


Back to what's going on now ! No singularity yet





Questions ?



IBM

Alain Cyr

cyralain@fr.ibm.com

Power & Storage Business Development
MOP IBM Client Center