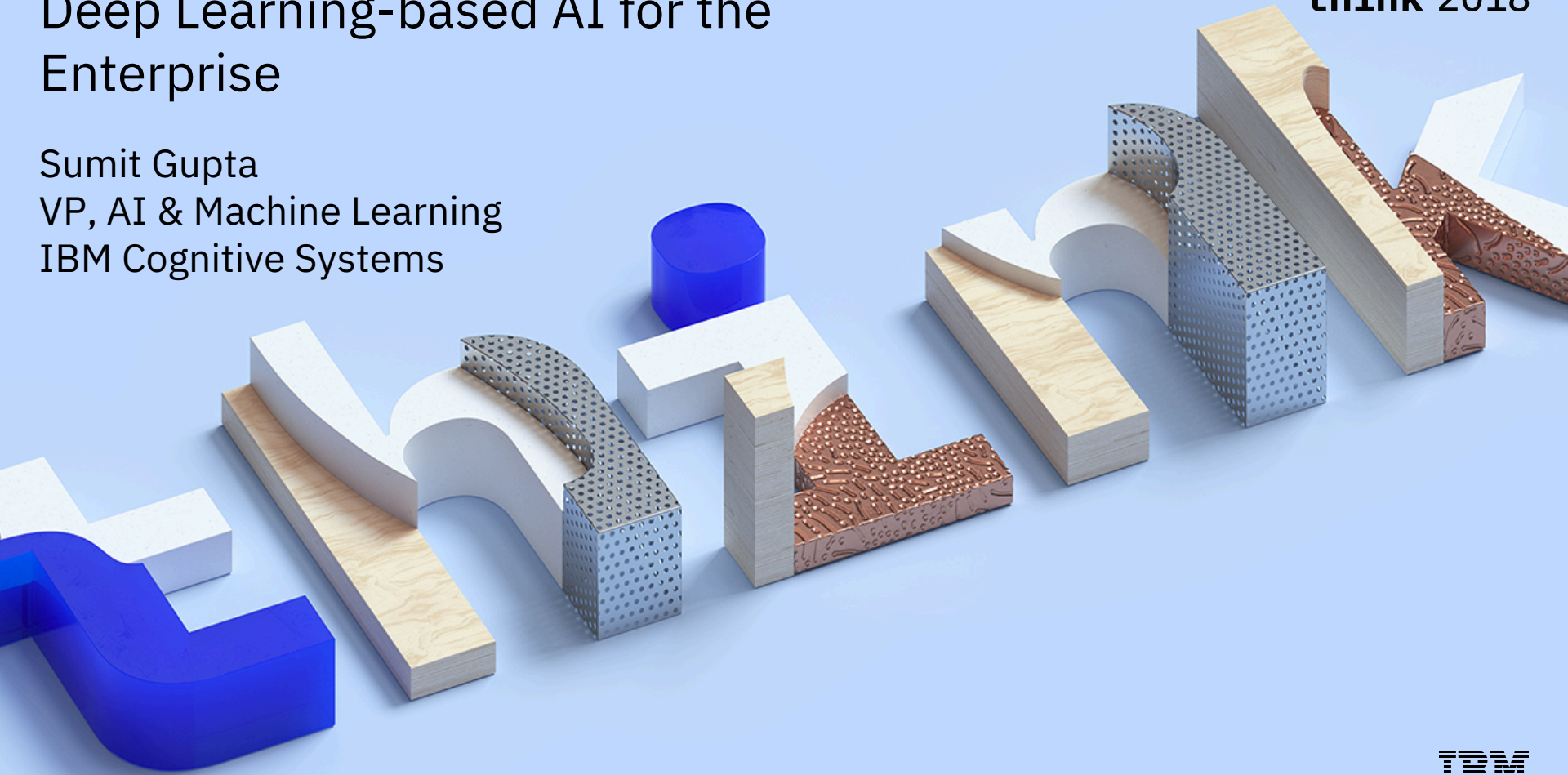
Deep Learning-based AI for the Enterprise

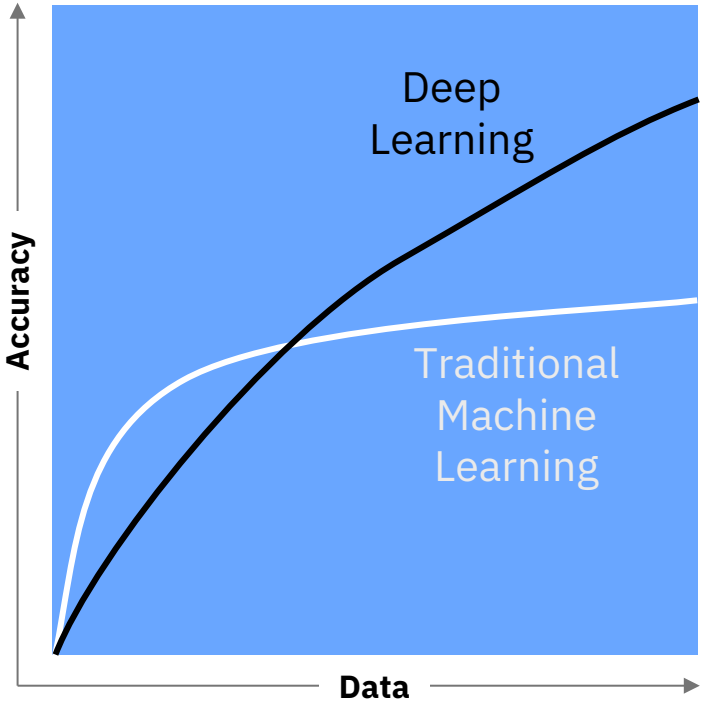
Sumit Gupta
VP, AI & Machine Learning
IBM Cognitive Systems

think 2018

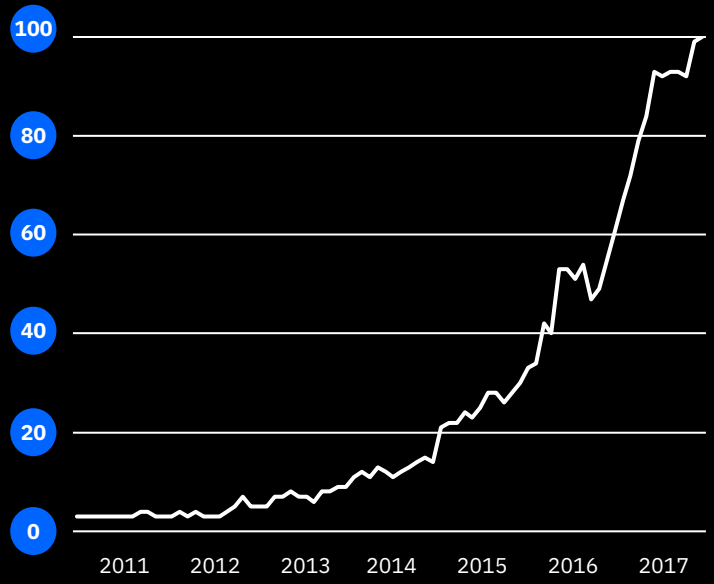


IBM

Deep Learning Has Revolutionized Machine Learning

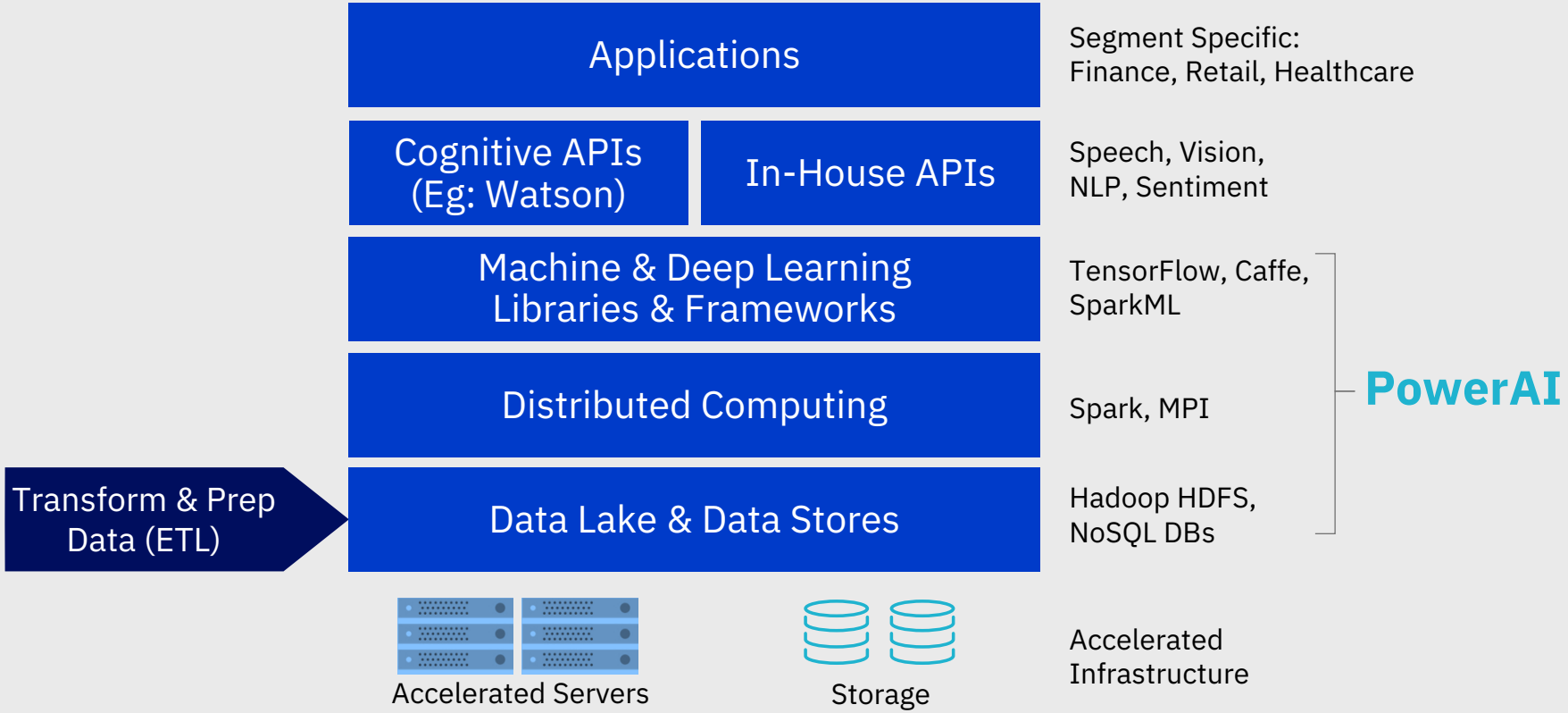


of Searches for Deep Learning from 2011 to 2017



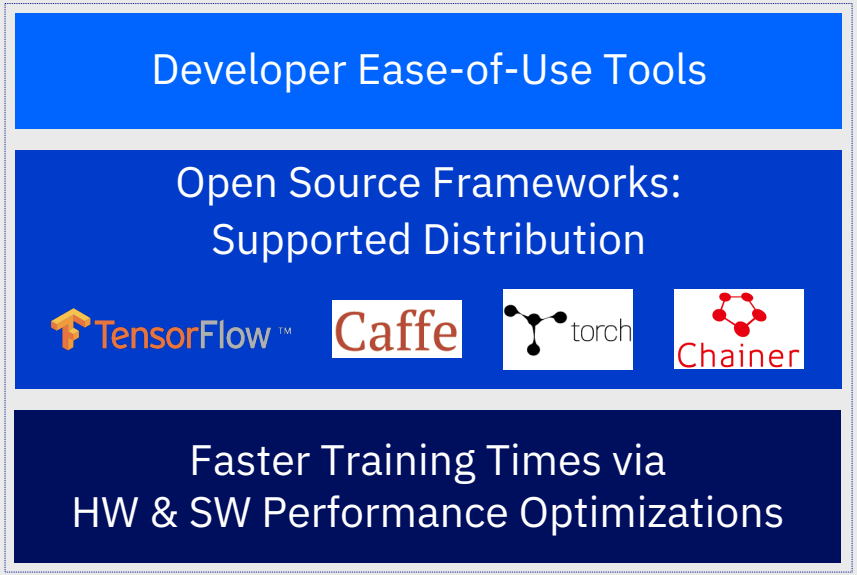
Source: Google Trends. Search term “Deep Learning”

AI Infrastructure Stack



PowerAI

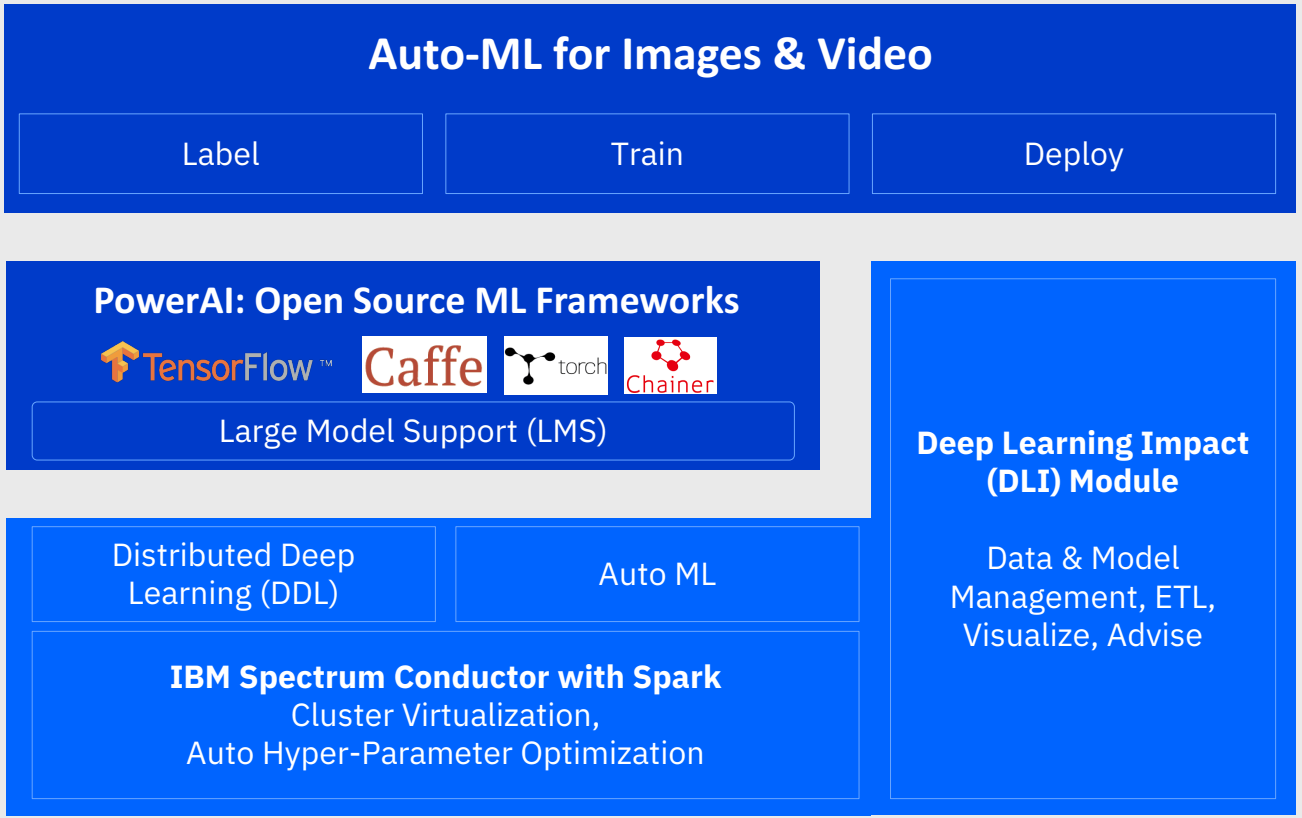
Integrated & Supported AI Platform
Higher Productivity for Data Scientists
Enable non-Data Scientists to use AI



PowerAI Vision

PowerAI

PowerAI
Tech Preview

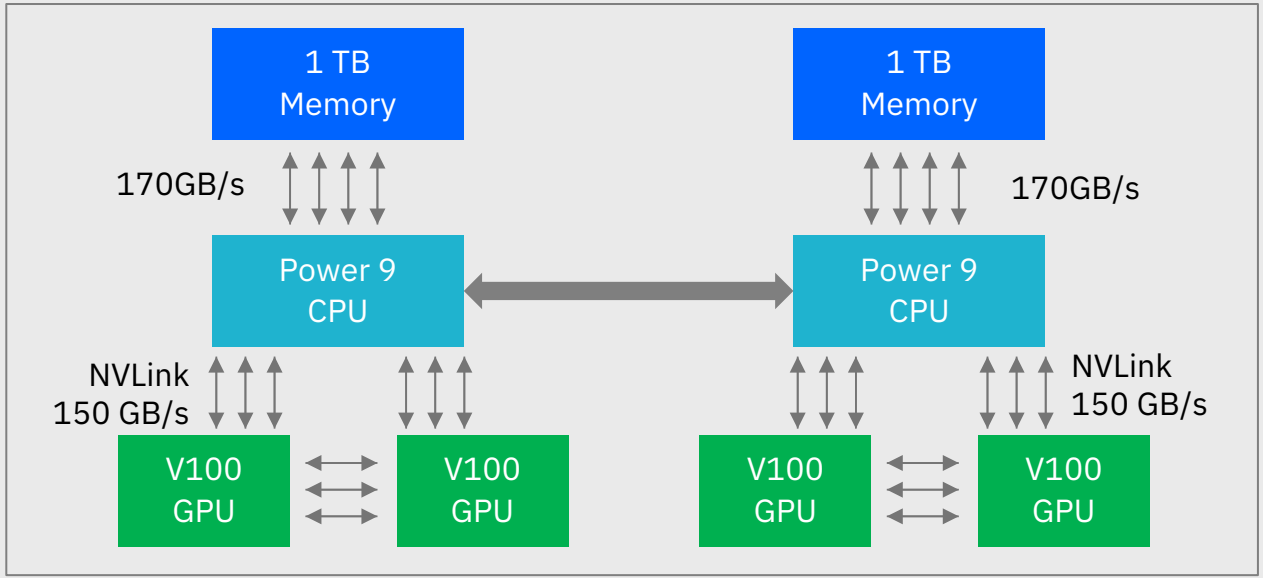


5x Faster Data Communication with Unique CPU-GPU NVLink High-Speed Connection

Store Large Models in System Memory

Fast Transfer via NVLink

Operate on One Layer at a Time



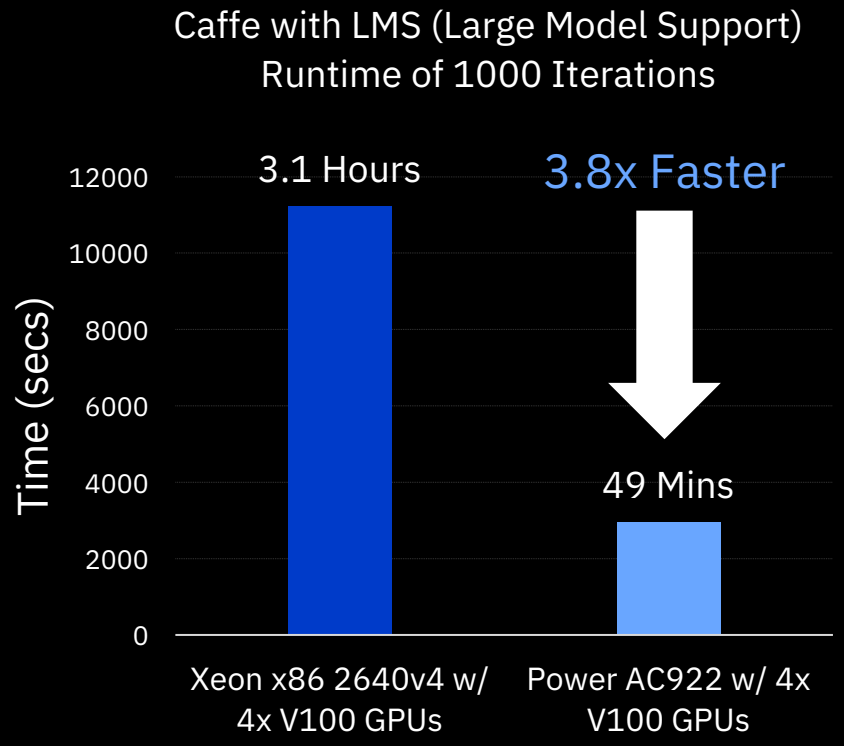
IBM AC922 Power System

Deep Learning Server (4-GPU Config)

Large AI Models Train ~4 Times Faster

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

Detailed Benchmark Information in Back



GoogleNet model on Enlarged ImageNet Dataset (2240x2240)

Distributed Deep Learning (DDL)

Deep learning training takes days to weeks

Limited scaling to multiple x86 servers

PowerAI with DDL enables scaling to 100s of servers

16 Days Down to 7 Hours
58x Faster

ResNet-101, ImageNet-22K

Near Ideal Scaling to 256 GPUs

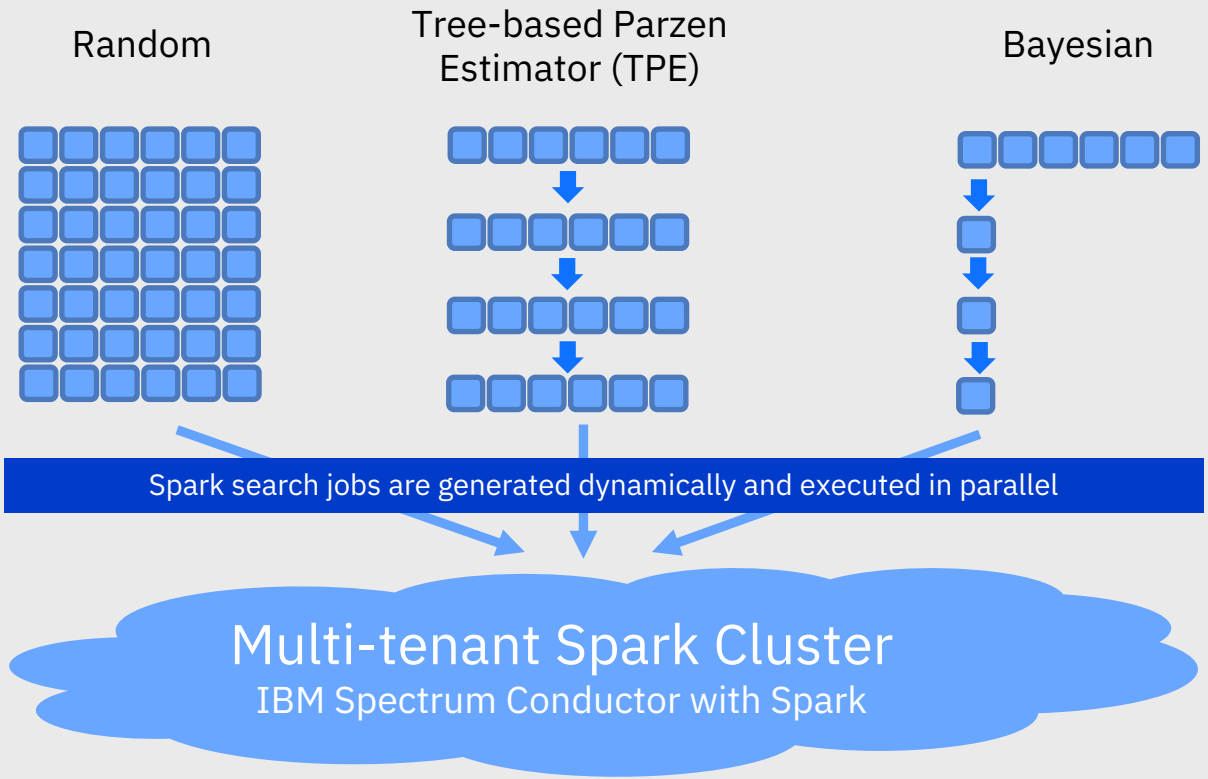
ResNet-50, ImageNet-1K

Caffe with PowerAI DDL, Running on Minsky (S822Lc) Power System

Auto Hyper-Parameter Tuning

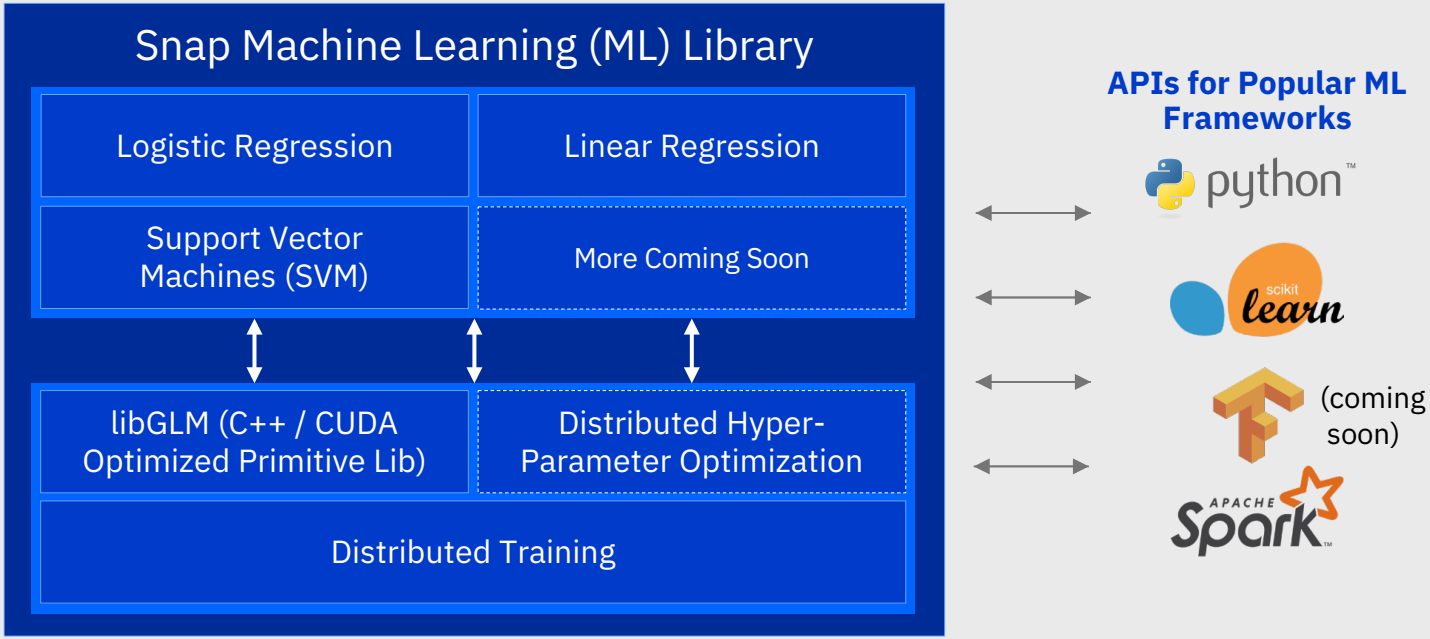
Hyper-parameters

- Learning rate
- Decay rate
- Batch size
- Optimizer
  - GradientDecedent, Adadelta, Momentum, RMSProp ...
- Momentum (for some optimizers)
- LSTM hidden unit size



Snap ML

Distributed GPU-Accelerated Machine Learning Library



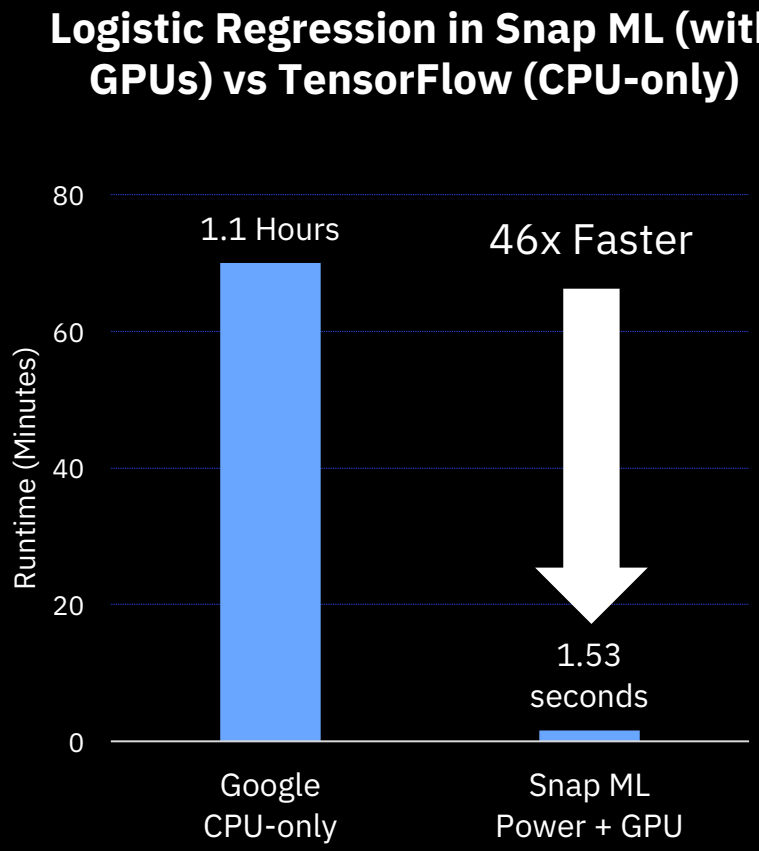
Snap ML: Training Time Goes From An Hour to Minutes

46x faster than previous record set by Google

Workload: Click-through rate prediction for advertising

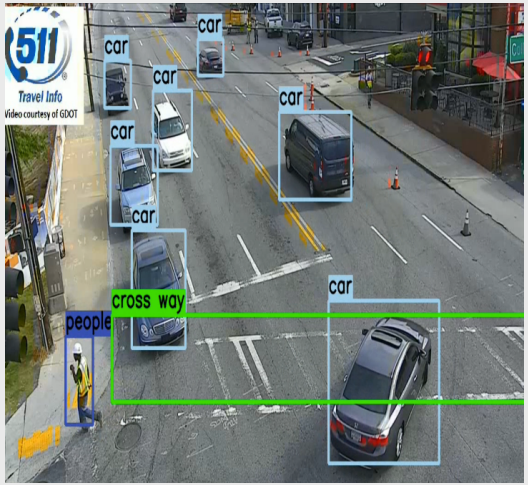
Logistic Regression Classifier in Snap ML using GPUs vs TensorFlow using CPU-only

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

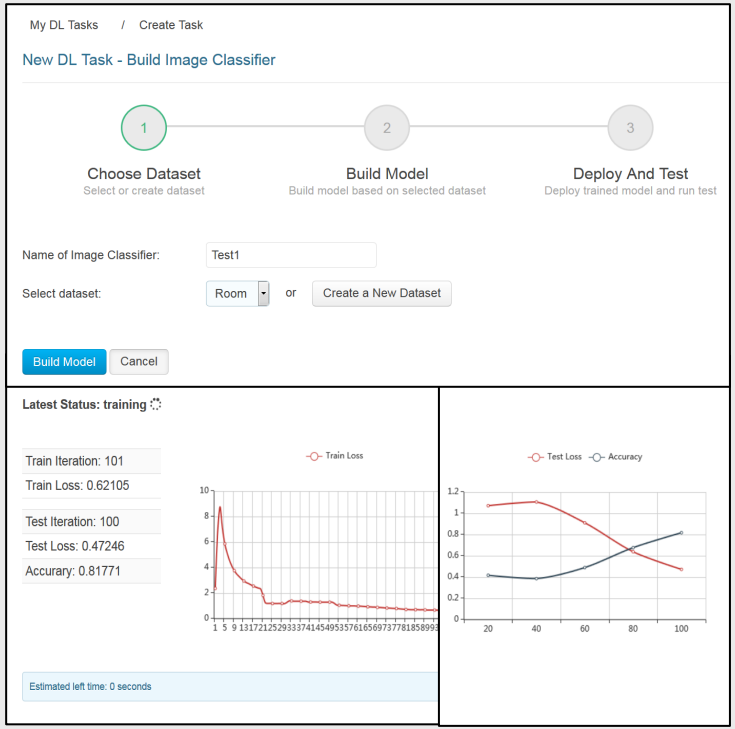


PowerAI Vision: ”Point-and-Click” AI for Images & Video

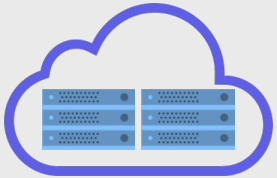
Label Image or Video Data



Auto-Train AI Model

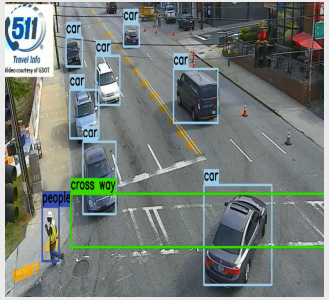


Package & Deploy AI Model



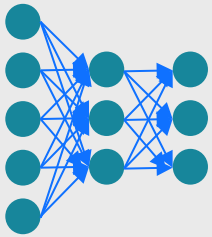
Semi-Automatic Labeling using PowerAI Vision

Manually Label



Define Labels Manually Label Some Images / Video Frames

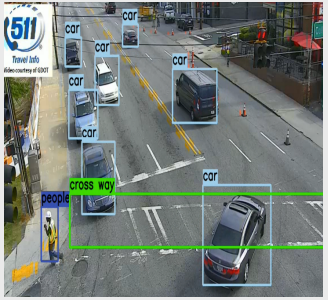
Train DL Model



Use Trained DL Model

Run Trained DL Model on Entire Input Data to Generate Labels

Correct Labels on Some Data



Manually Correct Labels on Some Data

Repeat Till Labels Achieve Desired Accuracy

AI Enterprise Use Cases

Financial Services

- Fraud detection, Credit Risk Analysis, Customer cross-sell
- Predict end of day deposit levels required
- Vision based: ID, Fraud, Insurance claims

Retailers

- Recommendation engines using POS data, E-tail image analytics
- Computer vision for shopping carts, manufacturing
- Customer Insights

Customer Service

- Understand speech, language, Q&A, enable smarter search
- Sentiment & Tone Analysis

Transportation & Industrial IoT

- Computer vision for self-driving vehicles
- Industrial / Inventory Inspection
- Predictive Maintenance

AUTOMOTIVE

ADAS, Maintenance

COMMUNICATIONS

Location-based advertising

CONSUMER GOODS

Sentiment analysis

FINANCIAL SERVICES

Risk analysis
Fraud detection

RESEARCH

Physics Modeling

MANUFACTURING

Line inspection, Defect analysis

LIFE SCIENCES

Sequence Analysis, Radiology

MEDIA/ENTERTAIN

Advertising effectiveness

CUSTOMER SERVICE

Chatbots, Helpdesk
Automated Expenses

HEALTH CARE

Patient sensors, monitoring, EHRs

OIL & GAS

Exploration, sensor analysis

RETAIL

Recommendation Engines,
Precision Marketing

TRANSPORTATION

Optimal traffic flows, Route planning

UTILITIES

Smart Meter analysis, Capacity planning

LAW & DEFENSE

Threat analysis – social media monitoring

Get Started Today with Machine & Deep Learning

IBM PowerAI

Build a Data Science Team
Your Developers Can Learn
http://cognitiveclass.ai

Identify a Low Hanging Use Case

Figure Out Data Strategy

Consider Pre-Built AI APIs

Hire Consulting Services

Get Started Today at
www.ibm.biz/poweraideveloper

Please visit one of the following Think Tanks to learn more.

12:30 PM – 12:50 PM
1:00 PM – 1:20 PM

Think Tank A
2422

Optimizing Accelerated Cognitive Systems with OpenCAPI

Think Tank B
7767

Introducing Deep Learning for Vision Analytics with IBM PowerAI Vision

Think Tank C
8060

Unlocking Shared Memory with Coherence

Think Tank D
8058

Deep Learning-Based AI for the Enterprise

Benchmark Details

Large Model Support benchmark Details

- Hardware: Power AC922; 40 cores (2 x 20c chips), POWER9 with NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU Pegas 1.0. Competitive stack: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4xTesla V100 GPU, Ubuntu 16.04.

- Chainer: IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model on Enlarged Imagenet Dataset (2240x2240) .

    - Software: Chainverv3 /LMS/Out of Core with CUDA 9 / CuDNN7 with patches found at https://github.com/cupy/cupy/pull/694 and https://github.com/chainer/chainer/pull/3762

- Caffe Results: IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2240x2240) .

    - Software: IBM Caffe with LMS Source code: https://github.ibm.com/TUNG/trlcaffe/tree/1.0-ibm-blc-bm-fix-hang+-p9collateral based on the branch "1.0-ibm-blc-bm-fix-hang+" (base for PowerAI R4) and a PR#5972 from BVLC/Caffe (for supporting cudnn7).