Hewlett Packard Enterprise

Deep Learning – a cookbook view

or "Comparative Analysis of Different Deep Learning Solutions "

The evolution of artificial intelligence

Small data sets

Early artificial intelligence
- **ENIAC** Heralded the Giant Brain"; used for WW II ballistics
- Industrial robots

Statistical and mathematical models applied to solve problems

1940 – 1980

Massive structured data sets

Machine learning
- **Deep Blue** Beating World Chess Champion Kasparov
- **DARPA Challenge** Autonomous vehicle drove 132 miles

Advanced Analytics and Heuristic

1990 – 2000s

Massive unstructured big data

Deep Learning
- Unsupervised training
- Generic code
- Pattern recognition

Systems can
- Observe
- Test
- Refine

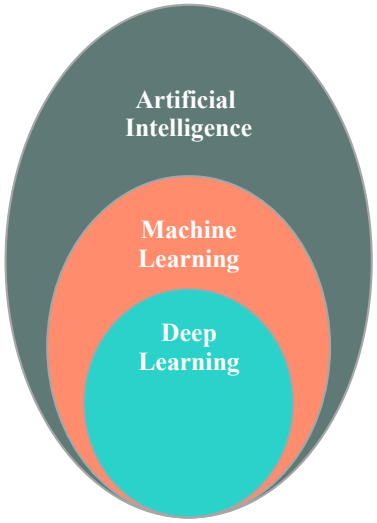
Successes
- **AlphaGO** First Computer GO program to beat a human
- **Deep Face** Facial verification
- **Libratus** AI Poker App
- **Digital virtual assistants** Siri
- **Google** Self-driving cars

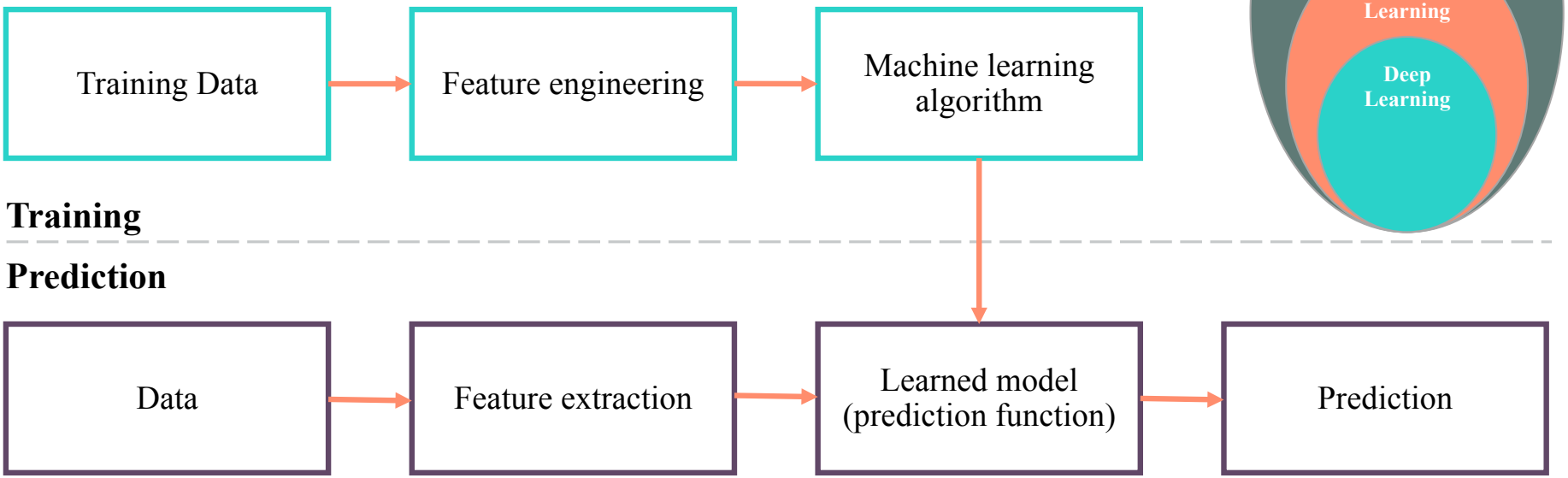
Predictive models defined by machines based neural networks

Today

Traditional machine learning

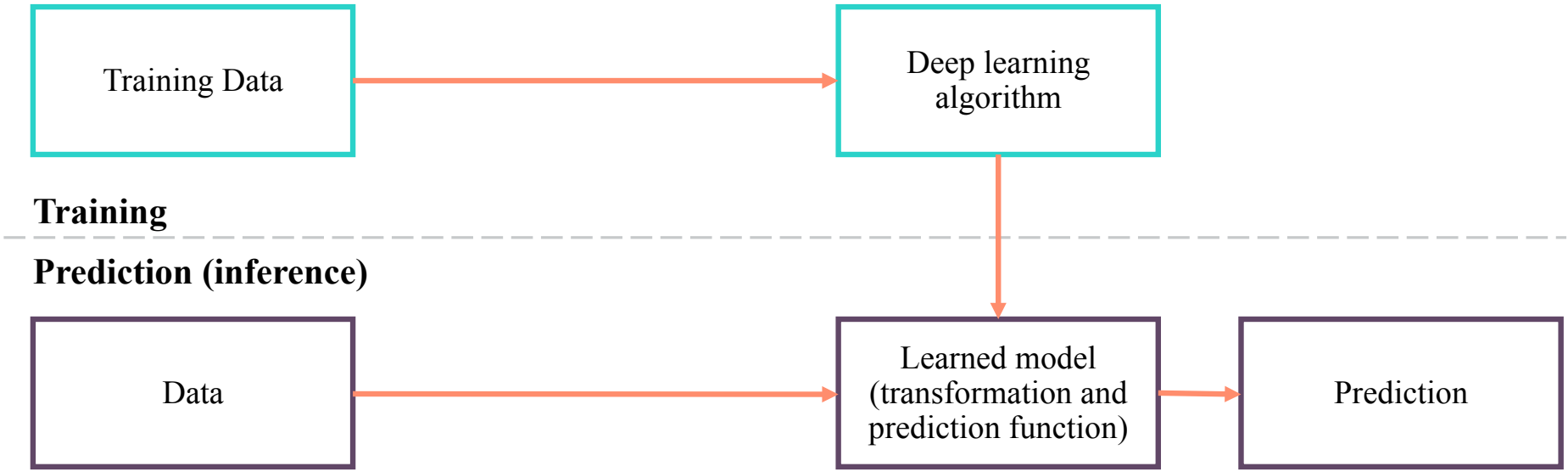
Requires feature engineering





Deep learning

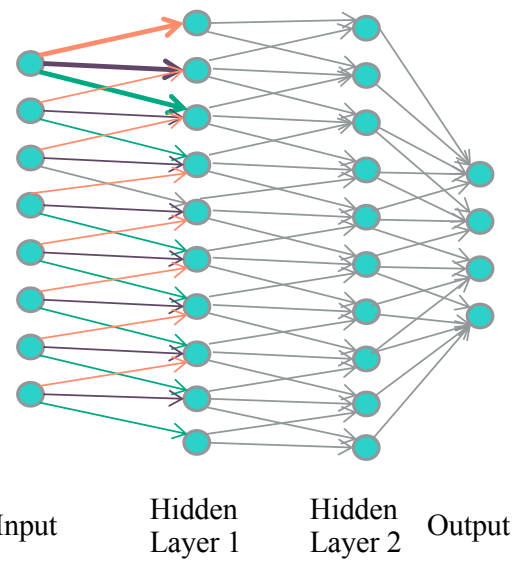
Efficient data representations, no more feature engineering



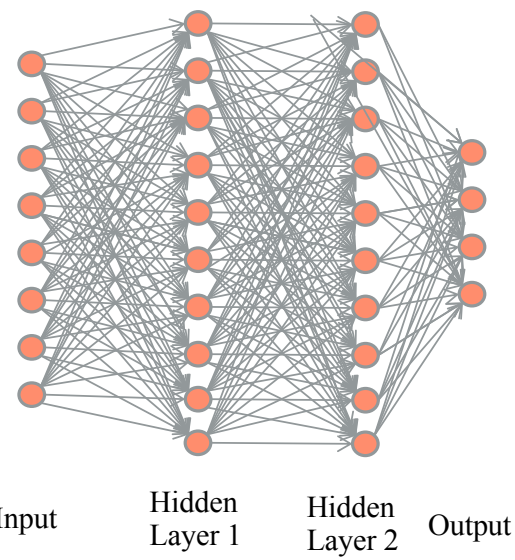
Types of artificial neural networks

Topology to fit data characteristics

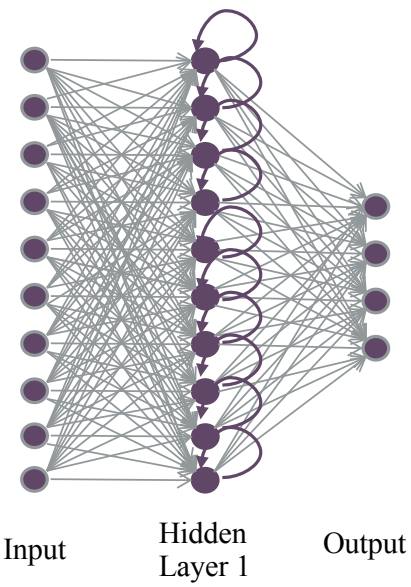
Convolutional:
Images



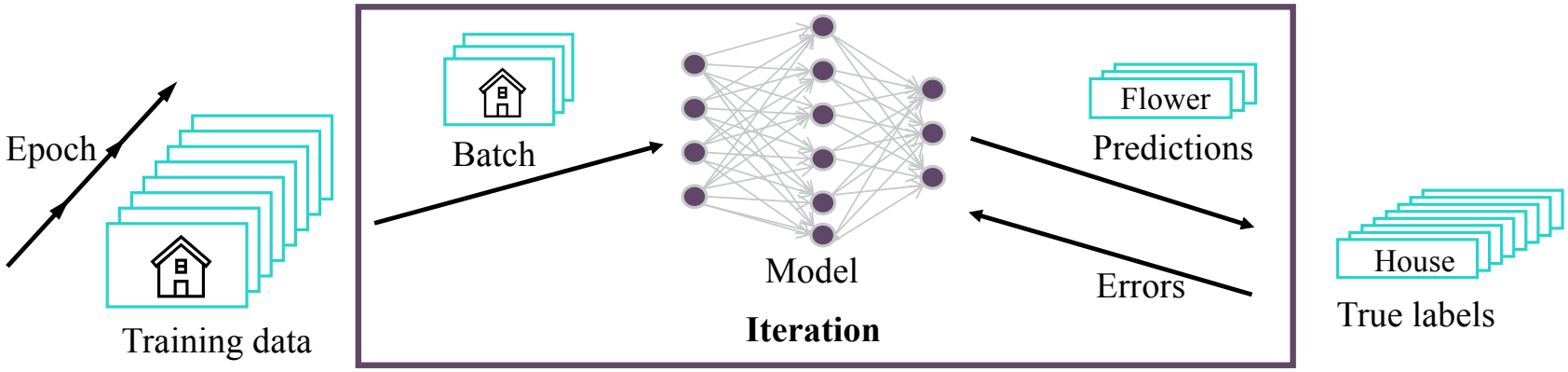
Fully connected:
Speech, text, sensor

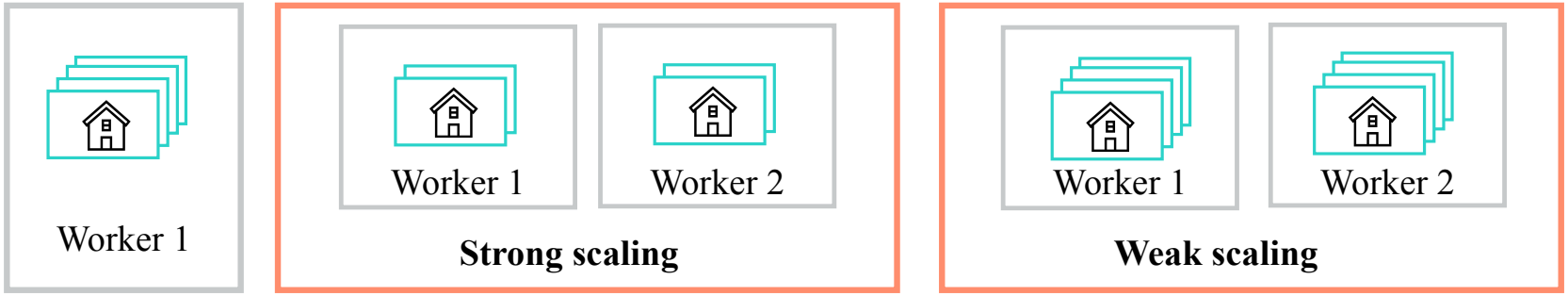


Recurrent:
Speech, text, sensor



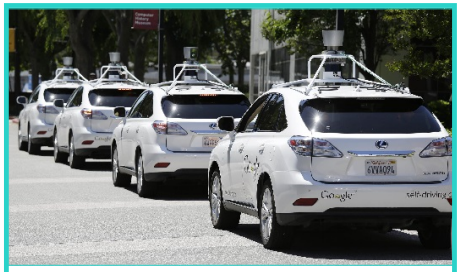
Terminology





Why deep learning?

Applications

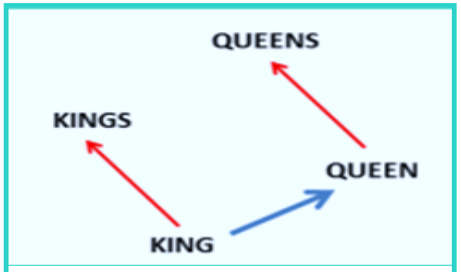


Vision

- Search & information extraction
- Security/Video surveillance
- Self-driving cars
- Medical imaging
- Robotics

Speech

- Interactive voice response (IVR) systems
- Voice interfaces (Mobile, Cars, Gaming, Home)
- Security (speaker identification)
- Health care
- People with disabilities



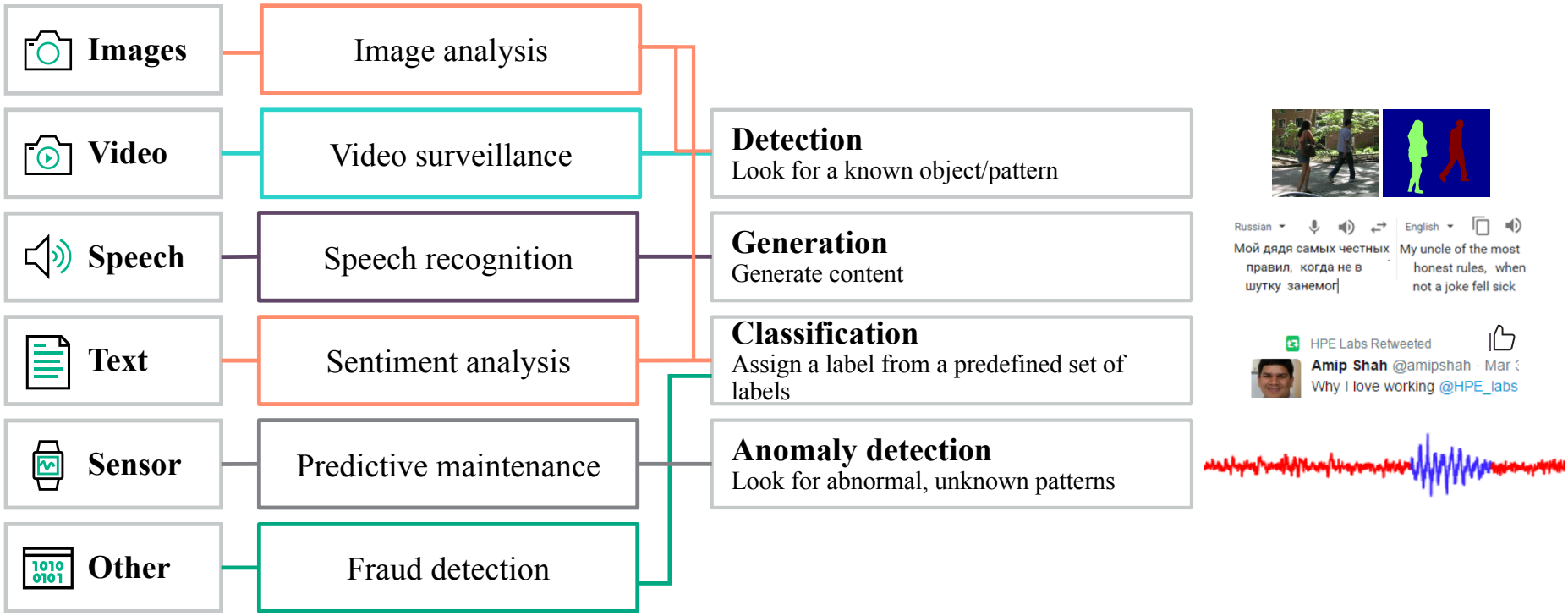
Text

- Search and ranking
- Sentiment analysis
- Machine translation
- Question answering

Other

- Recommendation engines
- Advertising
- Fraud detection
- AI challenges
- Drug discovery
- Sensor data analysis
- Diagnostic support

Applications break down



How an individual customer's AI evolves



Explore
How can AI help me?

Do things better
- Product development
- Customer experience
- Productivity
- Employee experience

Do new things
- New disruptions

Experiment
How can I get started?

Boundary constraints
(regulations, etc.)

Data
Data model? Location?
How to **create** a model?
- Homegrown solution or open source?
- Simple ML or scalable DL?

Design
How to design and deploy the PoC?
- On-prem, cloud?
- How to think about inference

Performance
What is the best config to run?
How to tune the model to improve accuracy?

Scale up and Optimize
How can I scale and optimize?

Provisioning for inference

Infrastructure scale up
- Training
- Inference
- On-prem / cloud / hybrid

Data management
- Between edge and core
- Security
- Updates
- Regulations
- Tracing

Hewlett Packard Enterprise

Key IT challenges are constraining deep learning adoption

Limited knowledge, resources and capabilities

How to get started?

"I need simple, infrastructure and software capabilities to rapidly and efficiently support deep learning app development."

Immature, sub-optimal foundation

How to go to production?

"I could use more expert advice and tailored solutions for migrating and integrating apps in a production environment."

Inability to scale and integrate

How to optimize?

"I need help integrating the latest technologies into my deep learning environment to accelerate actionable insights."

Lack of technology integration capabilities

What about AI consumers ?

Do it yourself

Current wave of AI / Machine Learning is core to their business. All in-house

Google, Baidu, Facebook, Microsoft, Apple, etc.

How do I do it ?

Could benefit from better data science, machine learning, but it is not historically their core-competency

Banks, advertisers, healthcare, manufacturing, food, automotive, etc.

Not ready for an ASIC. Don't know what they need exactly. Many still developing on CPUs. Can't use solutions that can't be verified or understood

I know better

Super-Experts – current wave is woefully inadequate

Government – DoD, DoE, NSA, NASA, etc.

Begging for higher performance ASICs. Know exactly what they want to do. Strong technology pull.

Hewlett Packard Enterprise

Where to start ?

Recommend DL stack by vertical application

Verticals	Voice interfaces	Social media	Manufacturing	Oil & gas	Connected cars
Data type	Speech	Images	Video	Sensor data	
Data	Small	Moderate	Large		
Typical layers	Convolutional	Fully-connected	Recurrent	… ← **Neural Network sits here**	
Frameworks	TensorFlow	Caffe 2	CNTK	Torch	…
Infrastructure	x86	GPUs	FPGAs	TPU ?	…

Hewlett Packard Enterprise

Neural Network : Popular Networks

Network	Model size (# params)	Model size (MB)	GFLOPs (forward pass)
AlexNet	60,965,224	233 MB	0.7
GoogleNet	6,998,552	27 MB	1.6
VGG-16	138,357,544	528 MB	15.5
VGG-19	143,667,240	548 MB	19.6
ResNet50	25,610,269	98 MB	3.9
ResNet101	44,654,608	170 MB	7.6
ResNet152	60,344,387	230 MB	11.3

Today's scale

Model size, data size, compute requirements

Application	Model	Training data	FLOPs per epoch
Vision	$1.7 * 10^9$ ~6.8 GB	$14*10^6$ images ~2.5 TB (256x256) ~10 TB (512x512)	$6*1.7*10^9*14*10^6$ $\sim 1.4*10^{17}$
Speech	$60 * 10^6$ ~240 MB	100K hours of audio $\sim 34*10^9$ frames ~50 TB	$6*60*10^6*34*10^9$ $\sim 1.2*10^{19}$
Text	$6.5 * 10^6$ ~260 MB	$856*10^6$ words	$6*6.5*10^6*856*10^6$ $\sim 3.3*10^{16}$
Signals	$1.2 * 10^6$ ~4.8 MB	$3*10^6$ frames	$6*1.2*3*10^6*3*10^6$ $6.5*10^{13}$

Today's hardware

Model size, data size, compute requirements

Application	Model	Training data	FLOPs per epoch
Vision	$1.7 * 10^9$ ~6.8 GB	$14*10^6$ images ~2.5 TB (256x256) ~10 TB (512x512)	$6*1.7*10^9*14*10^6$ $\sim 1.4*10^{17}$

1 epoch per hour:
~39 TFLOPS

Today's hardware:

Google TPU2: 180 TFLOPS Tensor ops (FP16 ??)

NVIDIA Tesla V100: 15 TFLOPS SP (30 TFLOPS FP16 , 120 TFLOPS Tensor ops), 12 GB memory

NVIDIA Tesla P100: 10.6 TFLOPS SP, 16 GB memory

NVIDIA Tesla K40: 4.29 TFLOPS SP, 12 GB memory

NVIDIA Tesla K80: 5.6 TFLOPS SP (8.74 TFLOPS SP with GPU boost), 24 GB memory

INTEL Xeon Phi: 2.4 TFLOPS SP

Superdome X: ~21 TFLOPS SP, 24 TB memory

So what to recommend?

Software

Caffe, theano, torch, KALDI, Caffe2, Microsoft CNTK, mxnet, TensorFlow, Chainer

Hardware

NVIDIA, intel, ARM, AMD, Google

Building performance models

Alex Net

GoogleNet

VGG-16, VGG -19

ResNet 50, 101,152

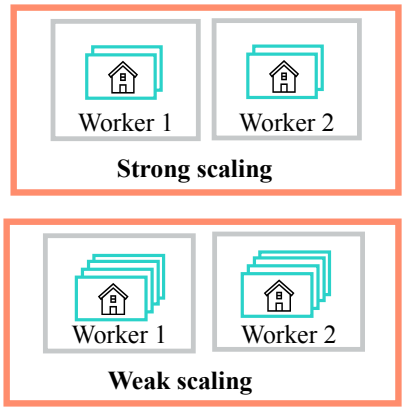
Eng Acoustic Model

TensorFlow

Caffe 2

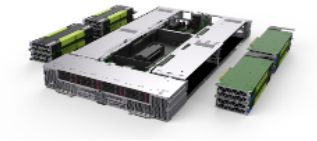
Tensor RT

BVLC Caffe



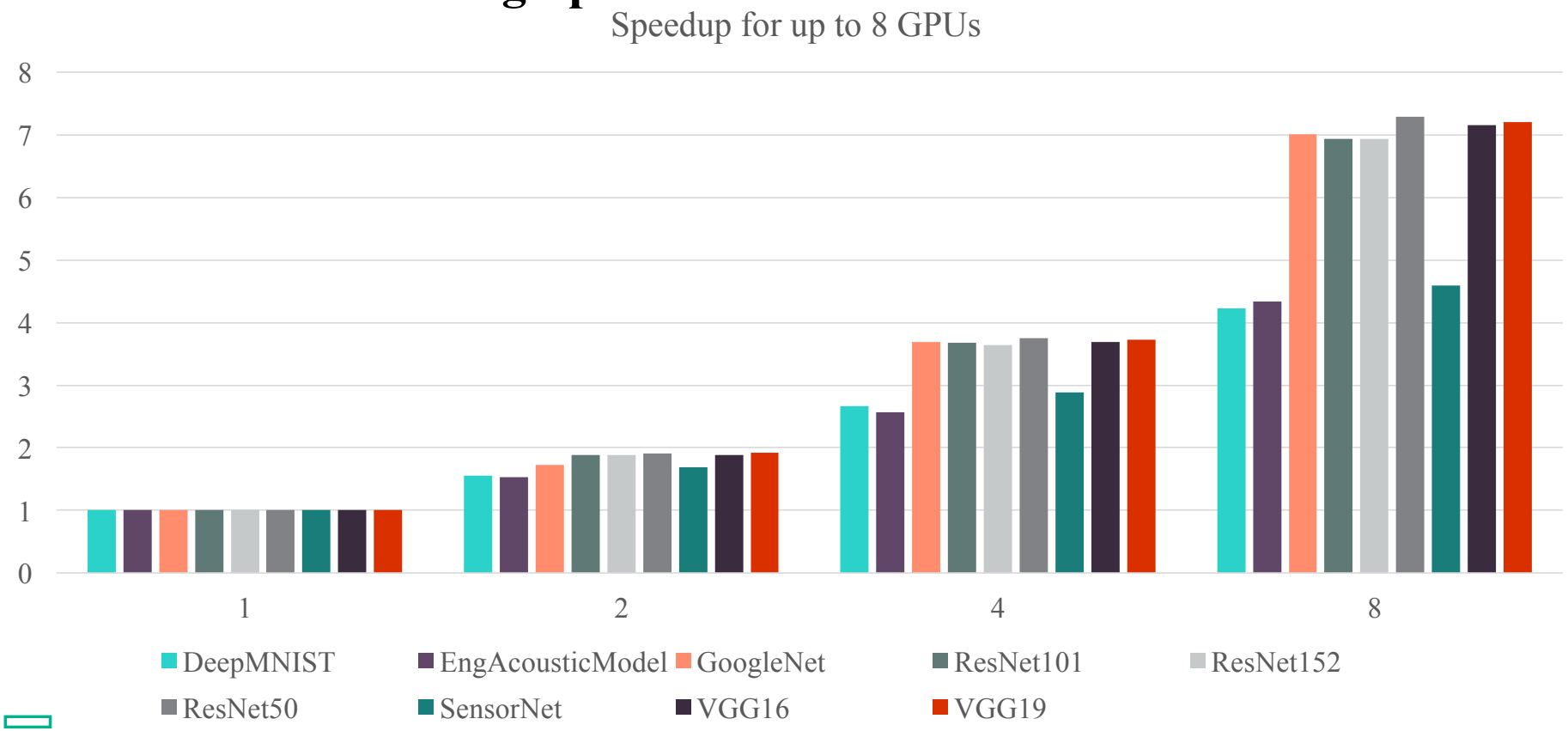
Hardware

Scalable, automated real-time intelligence



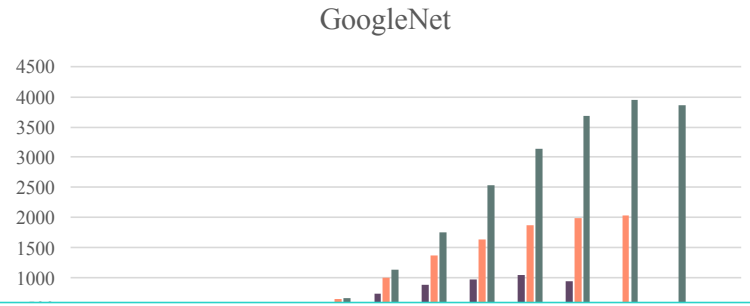
Populated with 8 GPUs

TensorFlow – Weak Scaling – Training – Different models perfromance in Tensor Flow . Scaling up to 8 GPUs



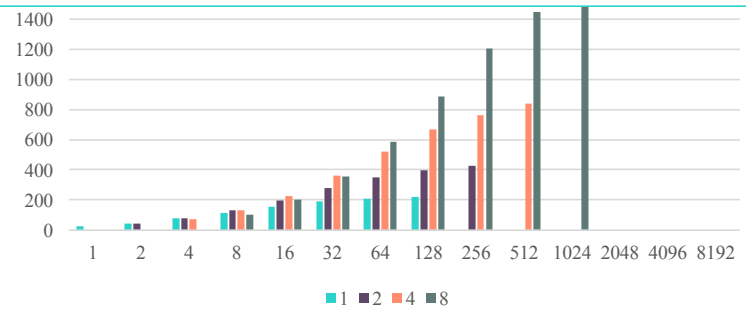
Hewlett Packard Enterprise

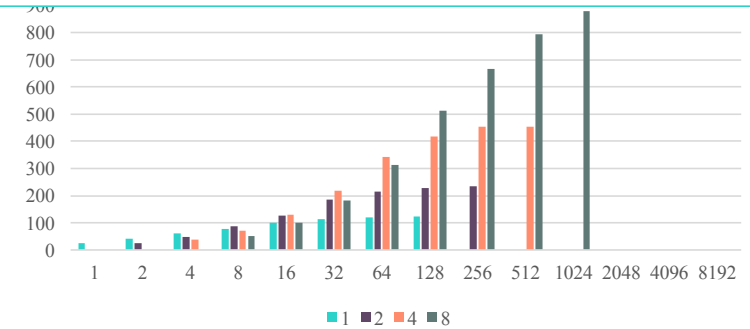
TensorFlow - Inference (Inferences per Second) - Different Models witth different Batch numbers



HOW TO ANALYZE ALL THE DIFFERENT NUMBERS .

AS WE ADD MORE OPTIONS and MORE TECHNOLOGIES IT WOULD BE IMPOSSIBLE TO USE



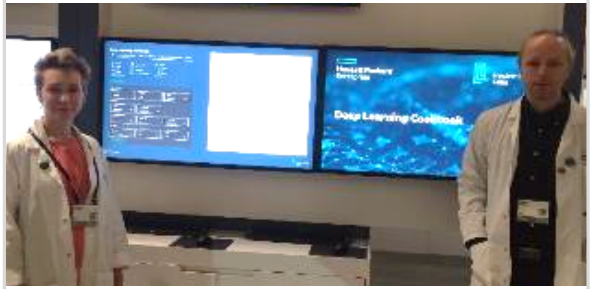


Hewlett Packard Enterprise

HPE demystifies deep learning for faster intelligence across all organizations

New IT expertise, blueprints and technologies to **get started, scale, integrate and optimize**

Get started rapidly:
Develop deep learning models



IT expertise and solutions
to "get started" with deep learning models

Expertise
- Rapid technology selection guides
- State of the art training

Solutions
- Integrated purpose-built solutions
- Out of the box solutions

Scale and Integrate: Deliver attractive returns

Proven blueprints and services
for "scalable" production deployments

Proven Blueprints
- Reference Architectures
- Innovation labs for best practices

Services
- Deploy, integrate and support
- Flexible, on-demand capacity

Optimize Environment:
Enhance competitive advantage

Technology integration capabilities
to maximize performance

Integration capabilities
- Enhanced global Centers of Excellence
- Next gen technology integration

Select ideal technology configurations

with HPE Deep Learning Cookbook

"Book of recipes" for deep learning workloads

TensorFlow
Caffe2
mxnet
Caffe

- **Comprehensive tool set** based on extensive benchmarking
- **Includes** 11 workloads with 8 DL frameworks and 8 HPE hardware systems
- **Estimates workload performance** and recommends an optimal HW/SW stack for that workload

Expert advice to get you started

- **Informed decision making** - optimal hardware and software configurations
- **Eliminates the "guesswork"** - validated methodology and data
- **Improves efficiency** - detects bottlenecks in deep learning workloads

Availability of complete toolset

GitHub

- **Deep Learning Benchmarking Suite:** available on GitHub Dec 2018
- **Deep Learning Performance Analysis Tool:** planned to be released in the beginning of 2018.
- **Reference configurations:** available soon on HPE.com website

Deep Learning Cookbook helps to pick the right HW/SW stack

Benchmarking Suite

- Benchmarking scripts
- Reference models
- Performance metrics

TensorFlow Caffe2 Caffe TensorRT mxnet

Knowledgebase

Performance results

- 11 reference models
- 8 frameworks
- 8 hardware systems

Performance and scalability models

- Machine learning (SVR) to predict performance of core operations
- Analytical communication models
- Analytical models for overall performance

Reporting tool

- Performance results
- Performance prediction for arbitrary ANNs
- Scalability prediction
- Optimal HW/SW configuration for a given workload

Reference configurations

- Image classification
- Others to come

will be available externally

internal assets

Hewlett Packard Enterprise

Deep Learning Cookbook

Automatic Meeting Notes | Video Surveillance | Hospital Smart Care Unit | Custom

- [ ] Images
- [x] Videos
- [ ] Text
- [ ] Speech
- [ ] Sensor Data

- [ ] Classification
- [x] Detection
- [ ] Generation
- [ ] Anomaly Detection

- (•) Training
  - ( ) Large
  - (•) Medium
  - ( ) Small
- ( ) Inference

Recommend

Data and Model

Data size	Epochs	Model
10000000	50	VGG19

Hardware

Server	Processor unit
Apollo 6500	NVIDIA P100

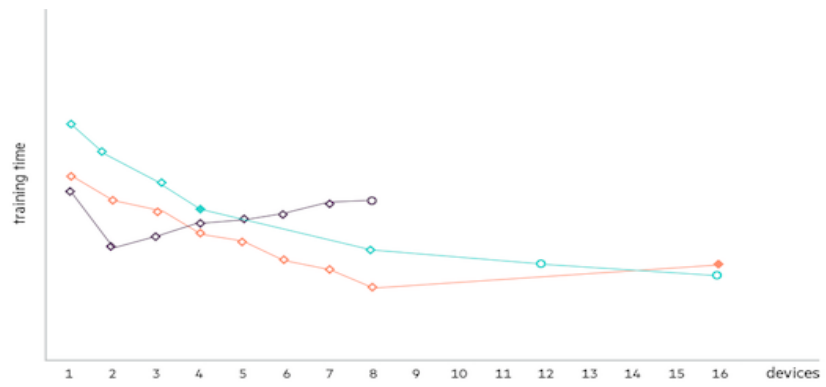
Count	Cluster size	Interconnect
8	2	InfiniBand FDR

Software

Framework	Batch size	Scaling
Caffe2	1024	strong

Add

Training performance



	Data			Hardware			Software	Time (hours)	
	Size	Epochs	Model	Server	PU		Framework		
	10000000	50	AlexNet	Apollo 6500	NVIDIA P100		Caffe2	22.5	×
				Count	Cluster size	Interconnect	Batch		
				8	2	IB	1024(strong)		
	Size	Epochs	Model	Server	PU		Framework		
	10000000	50	GoogleNet	Apollo 6500	NVIDIA P100		Caffe2	26.3	×
				Count	Cluster size	Interconnect	Batch		
				8	2	IB	1024(strong)		
	Size	Epochs	Model	Server	PU		Framework		
	10000000	50	VGG19	Apollo 6500	NVIDIA P100		Caffe2	147.4	×
				Count	Cluster size	Interconnect	Batch		
				8	2	IB	1024(strong)		

Remove all

Hewlett Packard Enterprise

Hewlett Packard Labs

Thank you

Natalia Vassilieva
nvassilieva@hpe.com

Sergey Serebryakov
sergey.serebryakov@hpe.com

Sorin Cheran
sorin.cheran@hpe.com

Bruno Monnet
bruno.monnet@hpe.com

Hewlett Packard
Enterprise