Hewlett Packard Enterprise

Hewlett Packard Labs

Characterization and Benchmarking of Deep Learning

Natalia Vassilieva, PhD
Sr. Research Manager

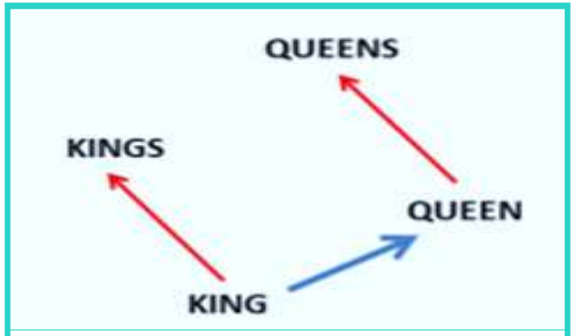
Deep learning applications

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
- Simultaneous interpretation



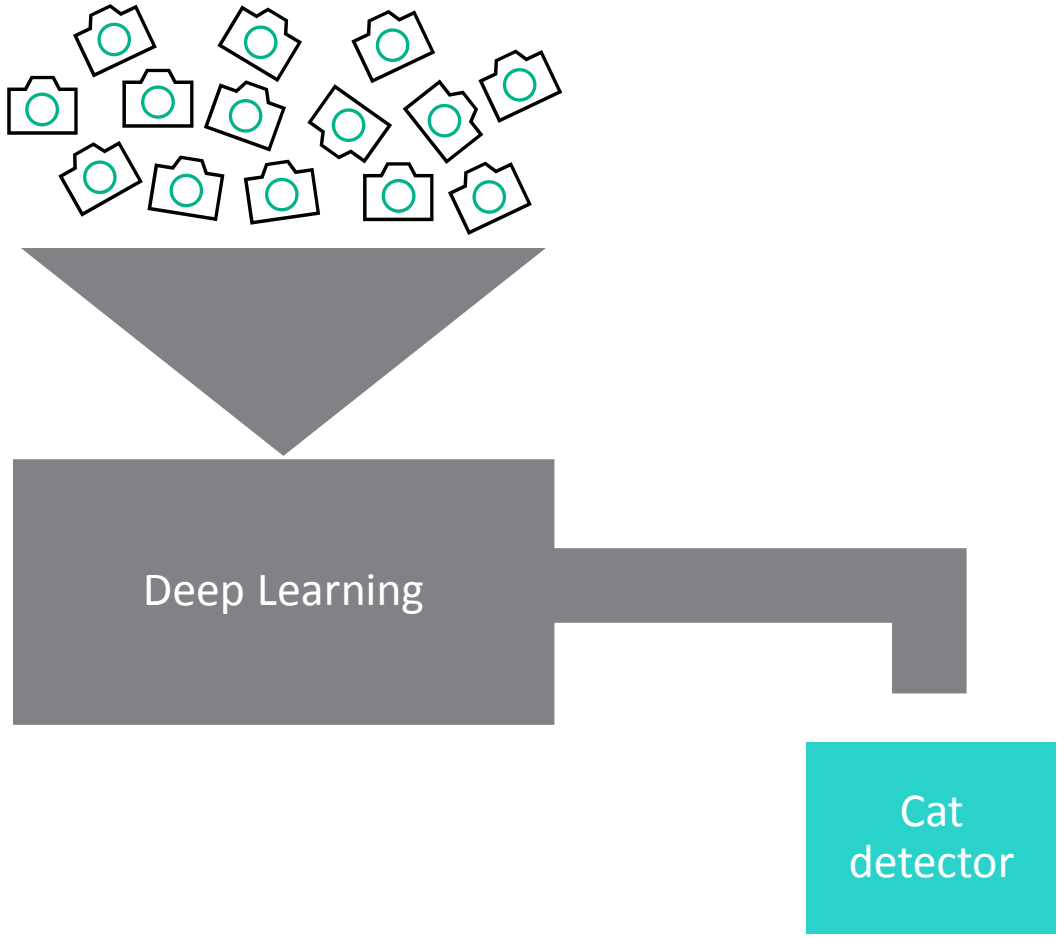
Text

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

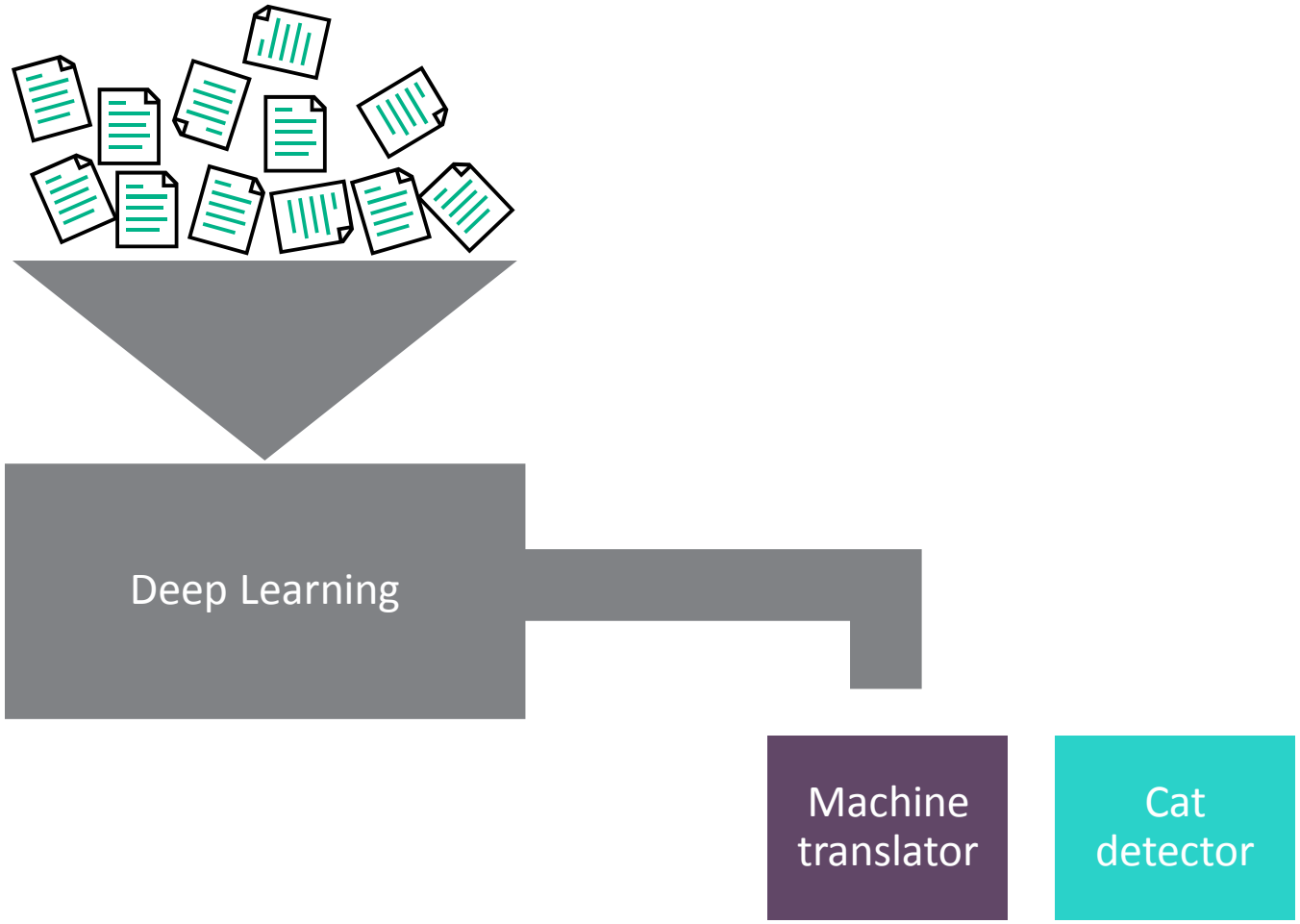
Other

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

Is Deep Learning a “universal algorithm”?



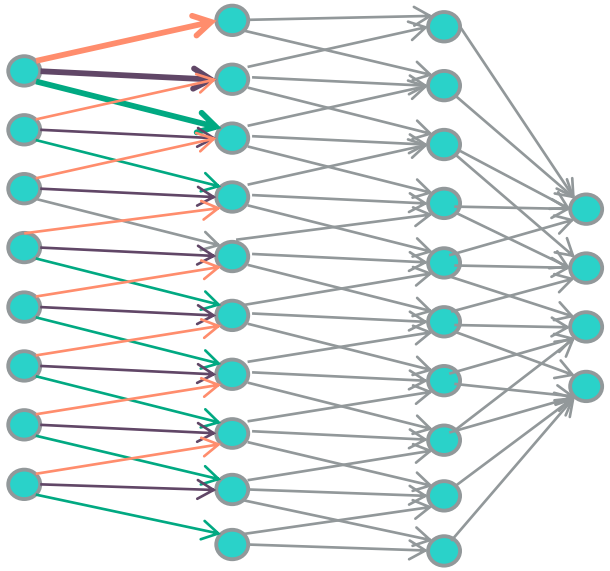
Is Deep Learning a “universal algorithm”?



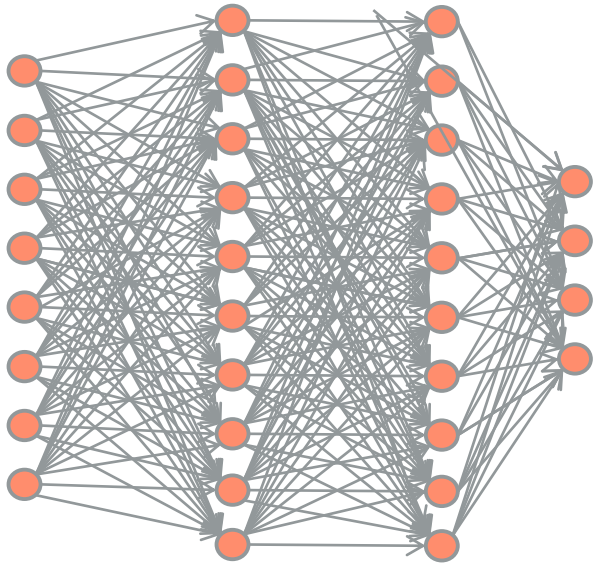
Types of artificial neural networks

Topology to fit data characteristics

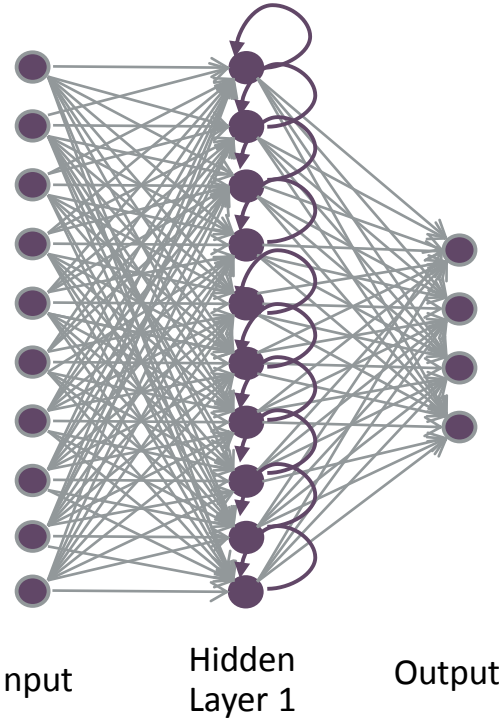
Convolutional:
Images



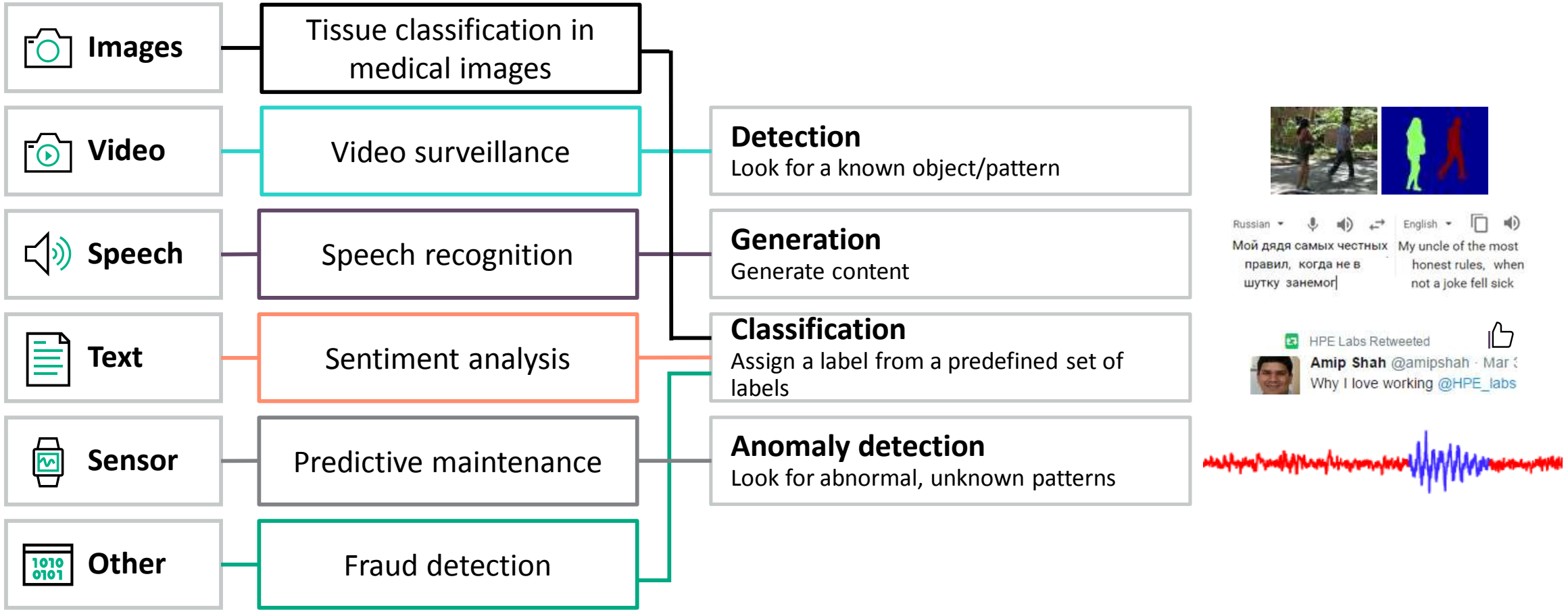
Fully connected:
Speech, text, sensor



Recurrent:
Speech, text, sensor



Applications break down



One size does NOT fit all

Application

Data type

Data size

Model (topology of artificial neural network):

- How many layers
- How many neurons per layer
- Connections between neurons (types of layers)

Deep learning ecosystem

How to pick the right hardware/software stack?

Popular models

Name	Type	Model size (# params)	Model size (MB)	GFLOPs (forward pass)
AlexNet	CNN	60,965,224	233 MB	0.7
GoogleNet	CNN	6,998,552	27 MB	1.6
VGG-16	CNN	138,357,544	528 MB	15.5
VGG-19	CNN	143,667,240	548 MB	19.6
ResNet50	CNN	25,610,269	98 MB	3.9
ResNet101	CNN	44,654,608	170 MB	7.6
ResNet152	CNN	60,344,387	230 MB	11.3
Eng Acoustic Model	RNN	34,678,784	132 MB	0.035
TextCNN	CNN	151,690	0.6 MB	0.009

Popular models

Name	Type	Model size (# params)	Model size (MB)	GFLOPs (forward pass)
AlexNet	**CNN**	**60,965,224**	**233 MB**	**0.7**
GoogleNet	CNN	6,998,552	27 MB	1.6
VGG-16	CNN	138,357,544	528 MB	15.5
VGG-19	CNN	143,667,240	548 MB	19.6
ResNet50	CNN	25,610,269	98 MB	3.9
ResNet101	CNN	44,654,608	170 MB	7.6
ResNet152	**CNN**	**60,344,387**	**230 MB**	**11.3**
Eng Acoustic Model	RNN	34,678,784	132 MB	0.035
TextCNN	CNN	151,690	0.6 MB	0.009

Compute requirements

Name	Type	Model size (# params)	Model size (MB)	GFLOPs (forward pass)
ResNet152	**CNN**	**60,344,387**	**230 MB**	**11.3**

Training data: 14M images (ImageNet)

FLOPs per epoch: $3 * 11.3 * 10^9 * 14 * 10^6 \approx 5 * 10^{17}$

1 epoch per hour: ~140 TFLOPS

Today's hardware:

Google TPU2: 180 TFLOPS Tensor ops

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

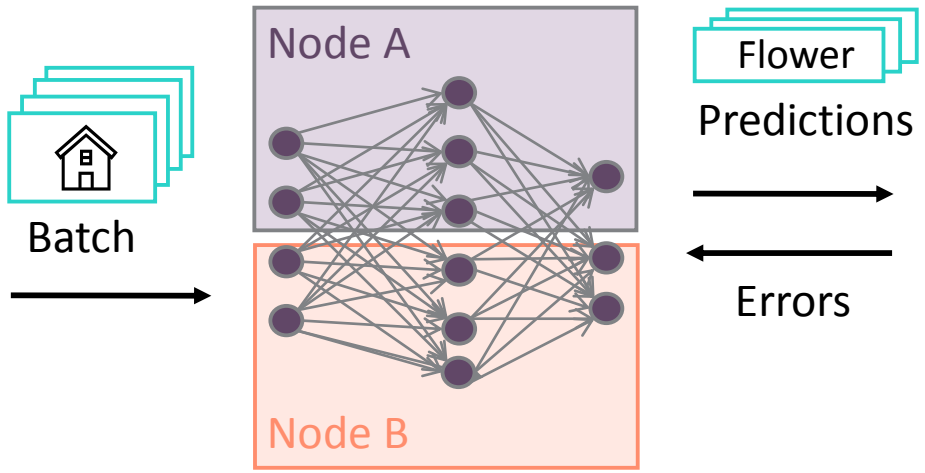
NVIDIA Tesla K40: 4.29 TFLOPS SP, 12 GB memory

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

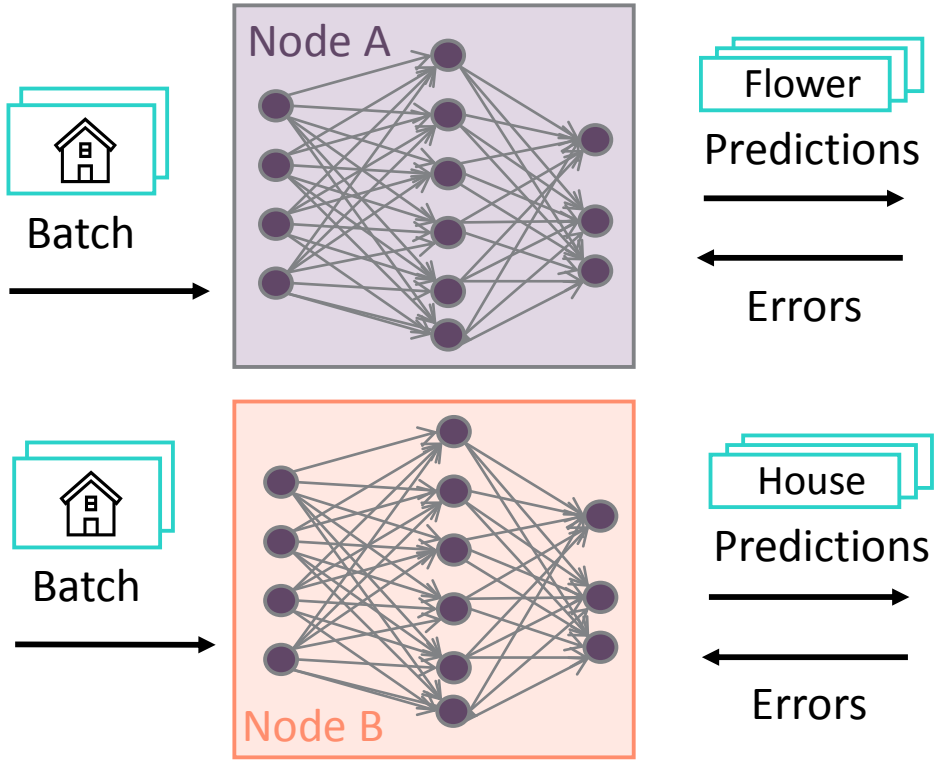
INTEL Xeon Phi: 2.4 TFLOPS SP

Distributed training

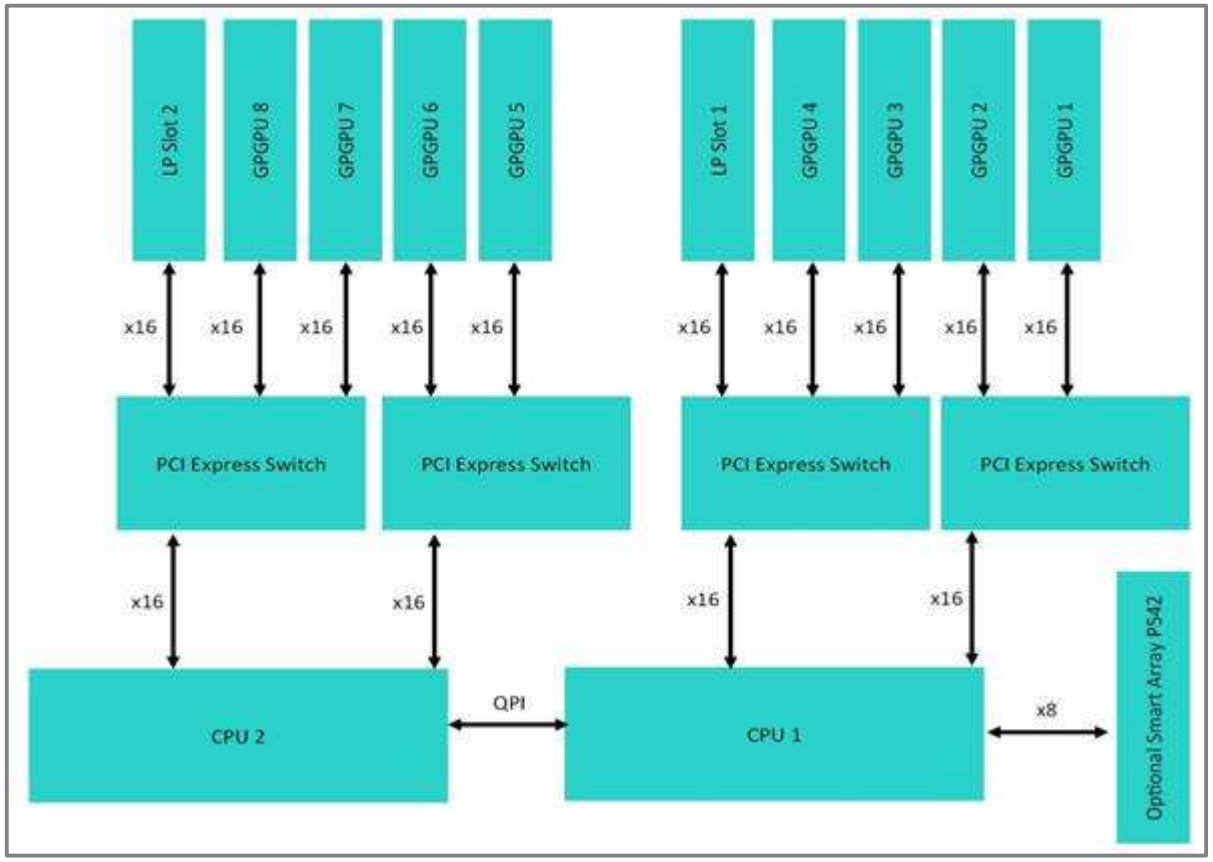
Model parallelism

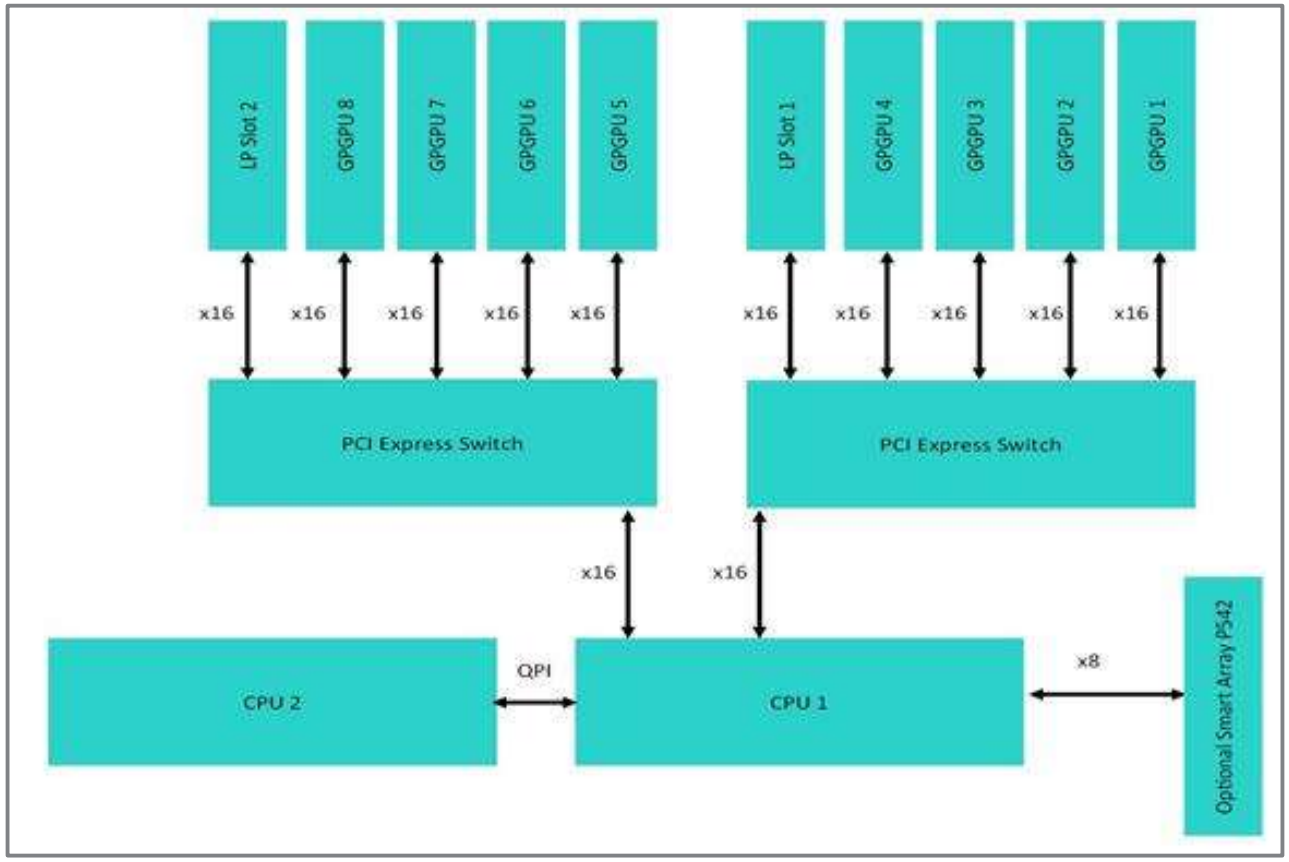


Data parallelism

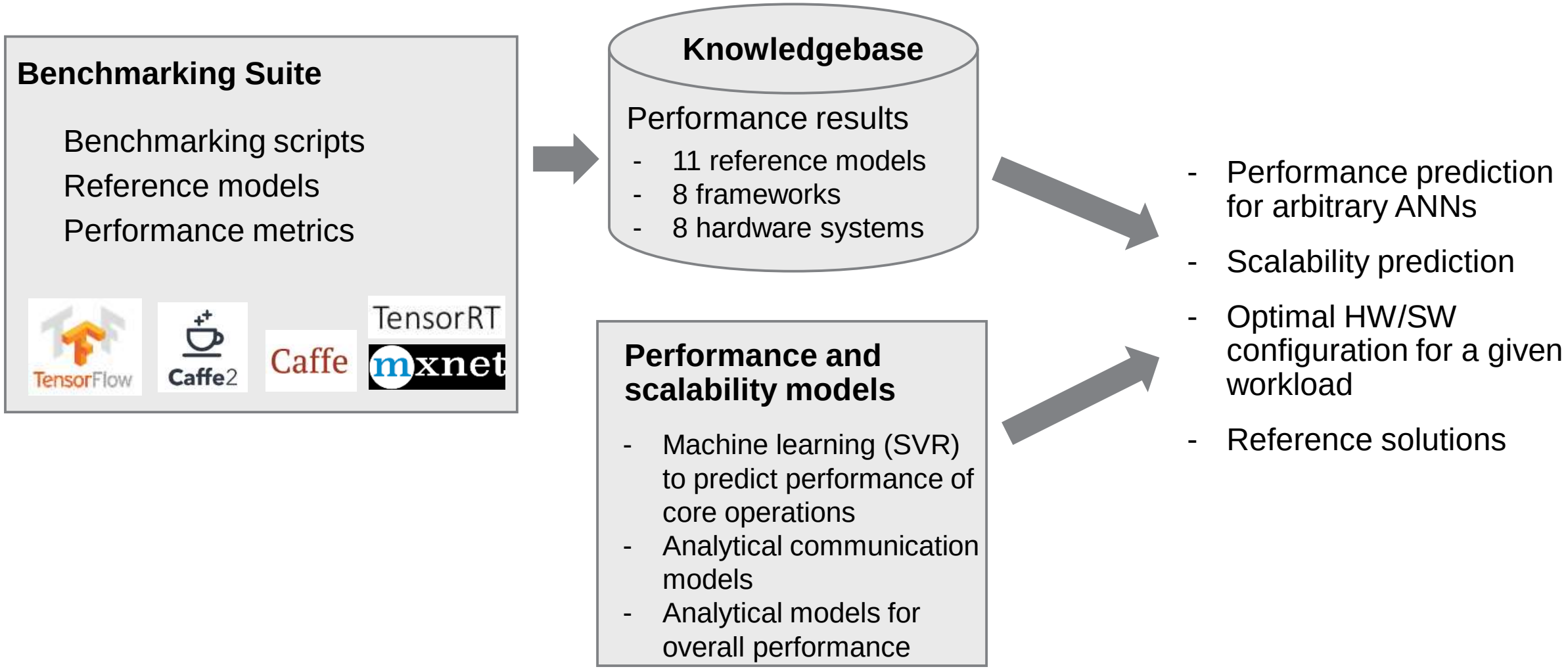


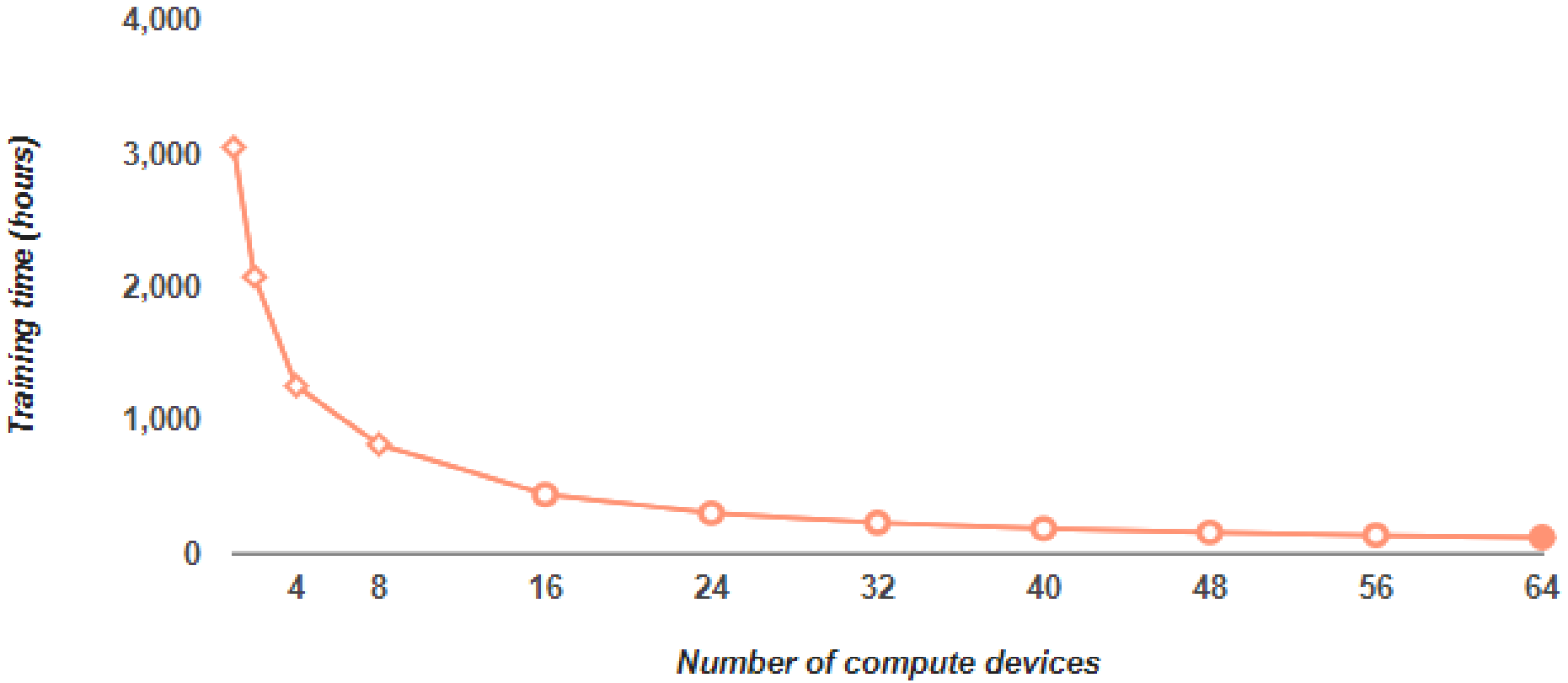
Which configuration is better?





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





Data			Hardware			Software	Time (hours)
			Server		PU	Framework	
Size	Epochs	Model	Apollo 6500		NVIDIA P100	TensorFlow	120.2
100000000	10	ResNet101	Count	Cluster size	Interconnect	Batch	
			8	8	IB	32(weak)	

Selected observations and tips

- Larger models are easier to scale (such as ResNet and VGG)
  - A single GPU can hold only small batches (the rest of memory is occupied by a model)
- Fast interconnect is more important for less compute-intensive models (FC)
- A rule of thumb: 2 CPU threads per GPU
- A rule of thumb: RAM = 2 x GPU memory x number of GPUs

Thank you

Natalia Vassilieva
nvassilieva@hpe.com