

# Deep Learning

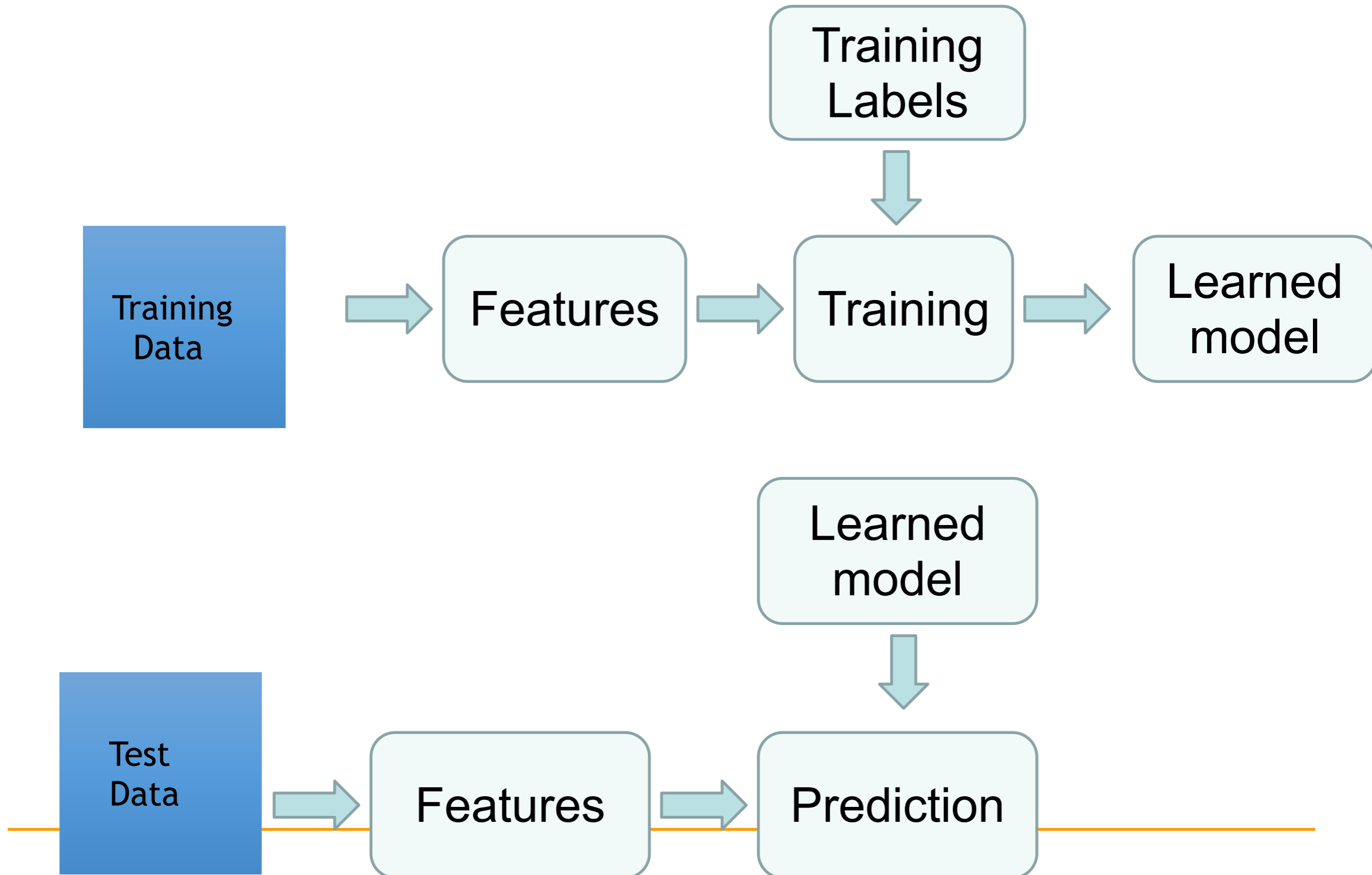
Hardware and frameworks

# Machine Learning Hardware

- CPUs
- GPUs
- FPGAs
- *Other accelerators*



# Machine Learning Project



# Machine Learning hardware

- *Speed up each block of the pipeline  
for example, matrix-matrix multiplication, convolution*
- *Data or memory paths for machine learning work  
example: caching*
- *Application-specific functional units*

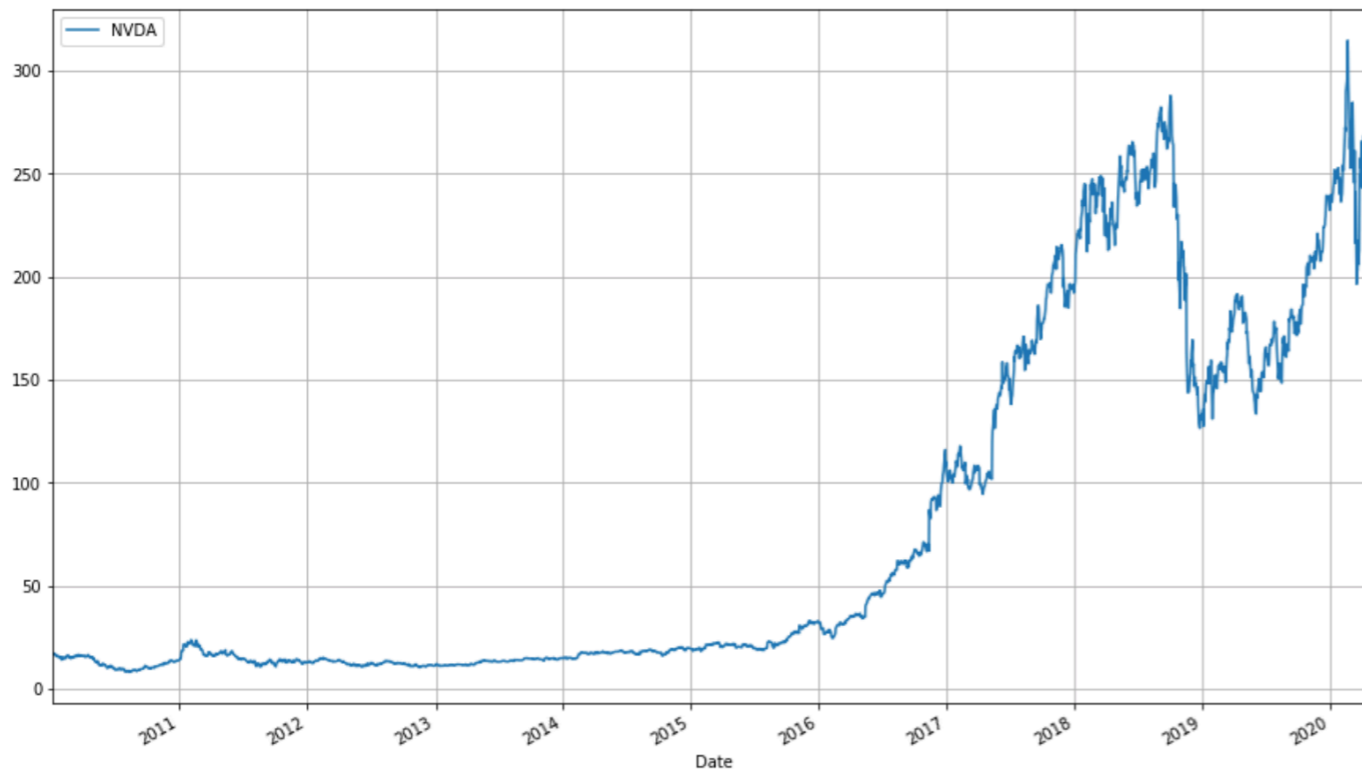
---

# Processing

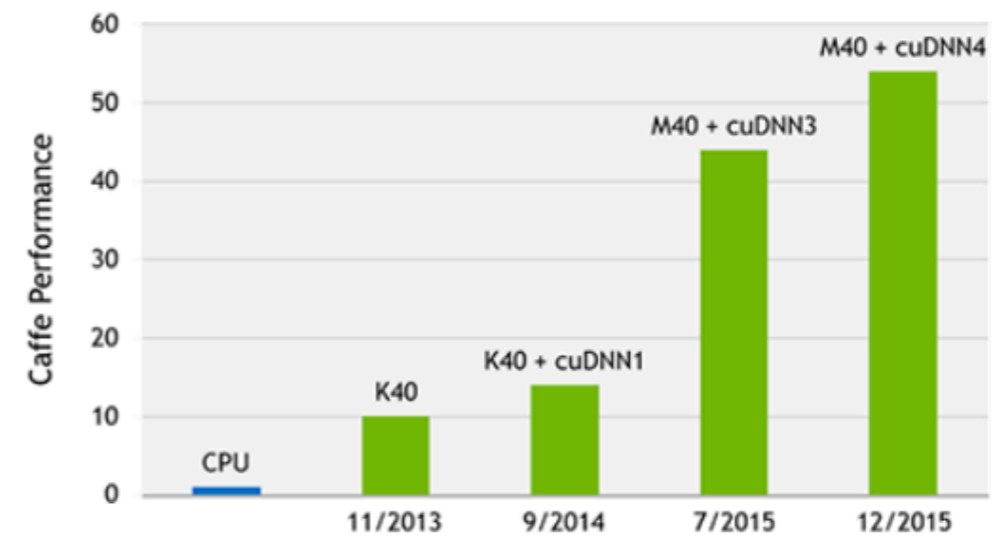
- *CPU is good at executing few complex operations.*
  - *In ML most of the processing involves matrix multiplication.*
  - *Lots of small calculations.*
  - *GPU is well suited for those kind of computations.*
-

# Processing

- *GPU utilizes parallel architecture.*
- *It is very good at handling many sets of very simple instructions.*

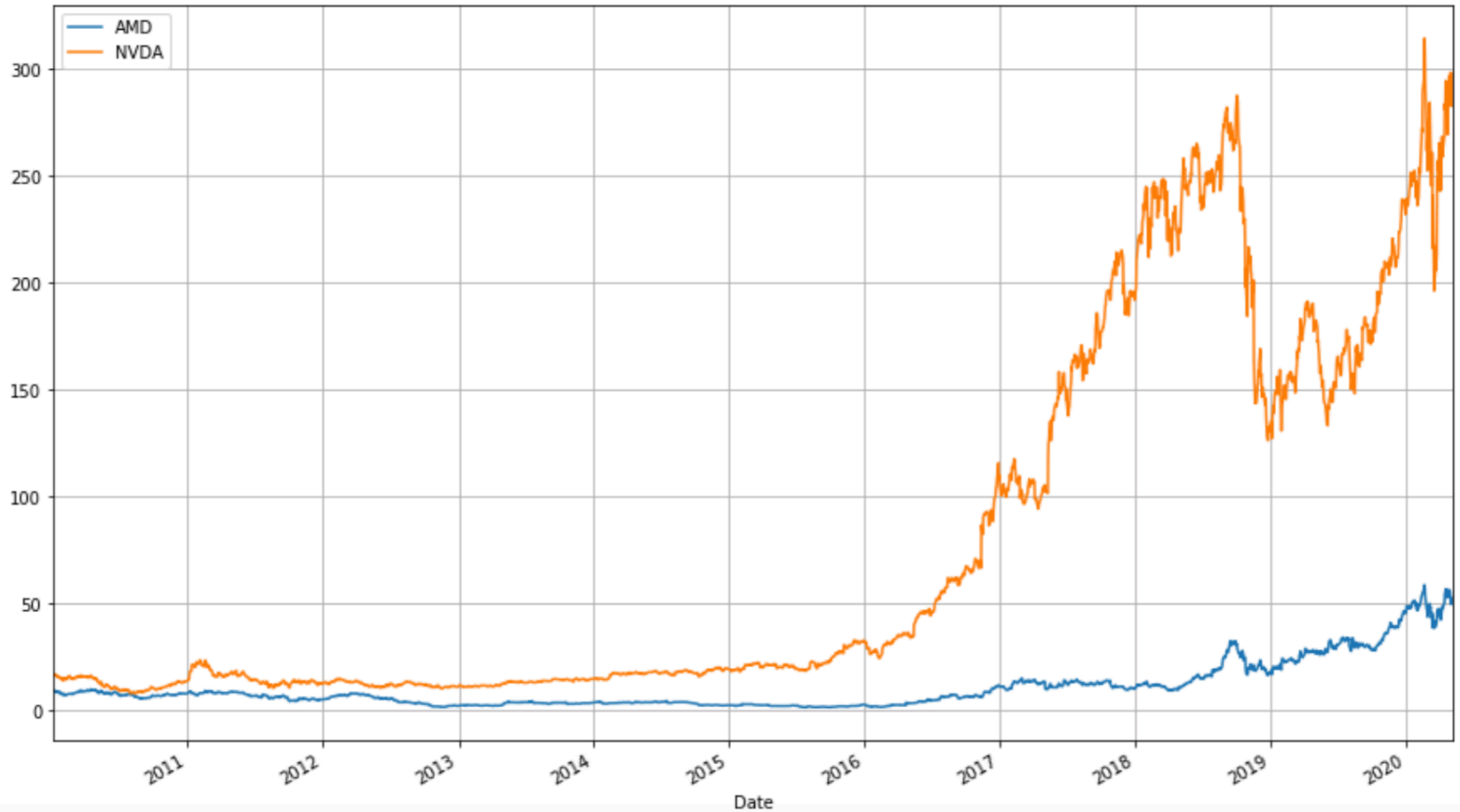


## 50X BOOST IN DEEP LEARNING IN 3 YEARS



AlexNet training throughput based on 20 iterations,  
CPU: 1x E5-2680v3 12 Core 2.5GHz, 128GB System Memory, Ubuntu 14.04

# GPU's



# CPU vs. GPU



**18 cores**



**2560 cores**



# CPU vs. GPU

	CPU i9 Xseries	GeForce GTX 1080
Cores	18 (36 threads)	2560
Clock Speed (GHz)	4.4	1.6G
Memory	Shared	8GB
Price (\$)	1799	549

# GPU programming

- CUDA
  - C-like code that runs on GPU
  - Other APIs: cuBLAS, cuFFT, cuDNN, etc
- OpenCL
  - Similar to CUDA, but runs on CPU's as well
  - usually slower

# Frameworks

- Caffe (Berkeley)
- Torch (NYU / Facebook)
- Theano (University of Montreal)

# Frameworks

- Caffe (Berkeley)
- Caffe2 (Facebook)
- Torch (NYU / Facebook)
- PyTorch (Facebook)
- Theano (University of Montreal)
- TensorFlow (Google)
- Paddle (Baidu)
- CNTK (Microsoft)
- MXNet (Amazon)

# DeepLearning Frameworks

- Computational graphs
- Gradient computation
- Run on GPU seamlessly