Parallel Machine Learning and Artificial Intelligence

Dr. Handan Liu

h.liu@northeastern.edu

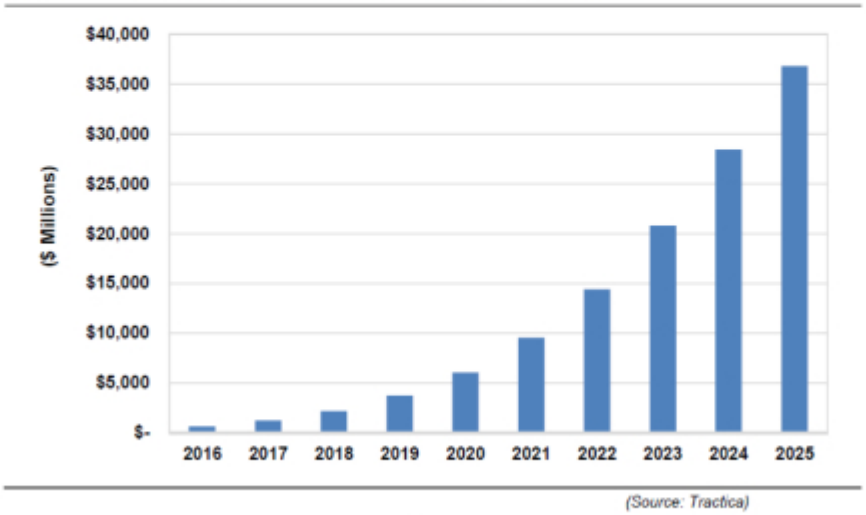
Northeastern University

Introduction to High Performance Deep Learning

The Bright Future of Deep Learning

Market for Artificial Intelligence Projected to Hit $36 Billion by 2025 from 2016
But now they forecast: by 2025 AI software revenues alone will reach near $100 billion globally.

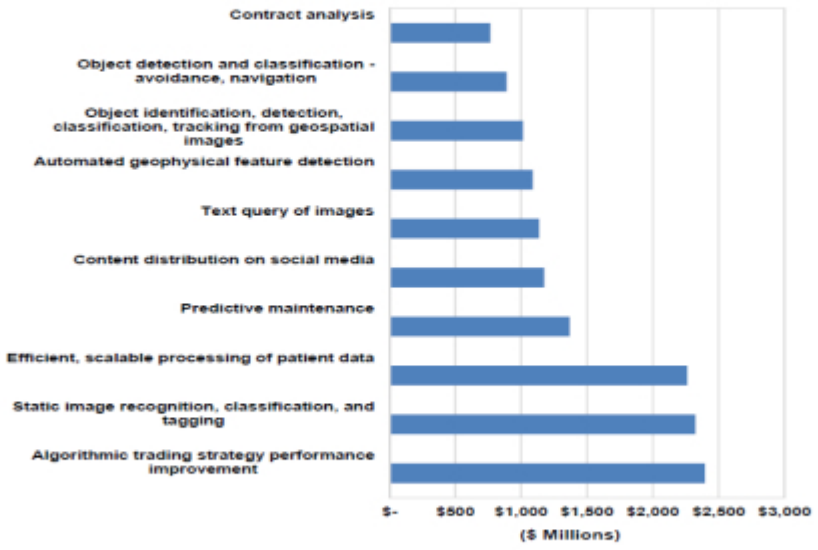
Chart 1.1 Artificial Intelligence Revenue, World Markets: 2016-2025



Courtesy:
https://omdia.tech.informa.com/topic-pages/artificial-intelligence

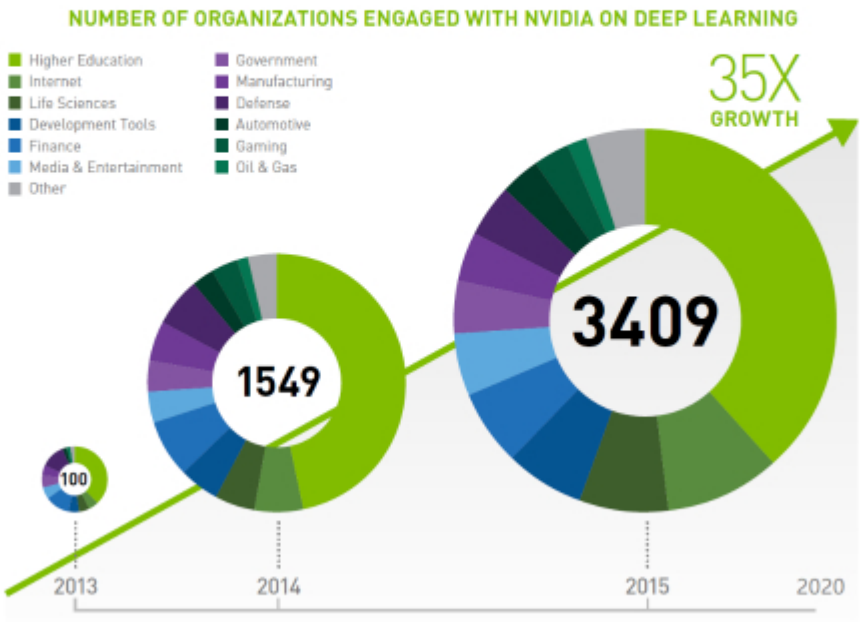
Current and Future Use Cases of Deep Learning

Chart 1.2 Artificial Intelligence Revenue, Top 10 Use Cases, World Markets: 2025



(Source: Tractica)

The Rise of GPU-based Deep Learning



Deep Learning can transform science

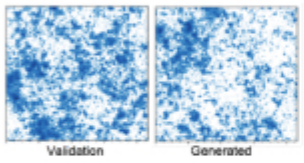
- Deep neural networks have powerful capabilities for science
  - Automatically learn patterns from high-dimensional data
  - Encode inductive biases, symmetries

- Some emerging promising application areas
  - Analysis of large scientific datasets
  - Accelerate expensive simulations
  - Real time control and design of experiments

Deep Learning science examples

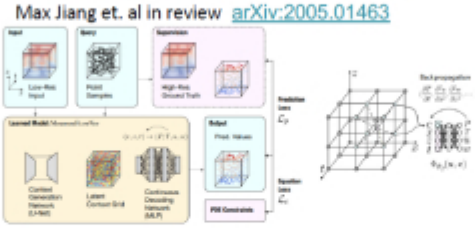
CosmoGAN for simulations

Mustafa et. al Comput. Astrophys. 6, 1 arXiv:1706.02390



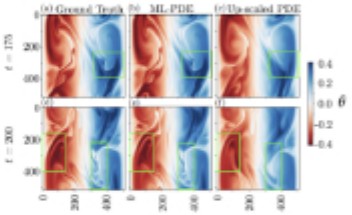
Mesh-free space-time super-resolution

Max Jiang et. al in review arXiv:2005.01463



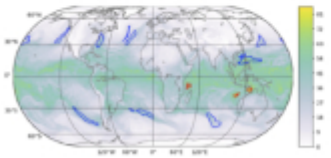
Using ML to Augment Coarse-Grid Computational Fluid Dynamics Simulations

Jaideep Pathak et. al in review arXiv:2010.00072



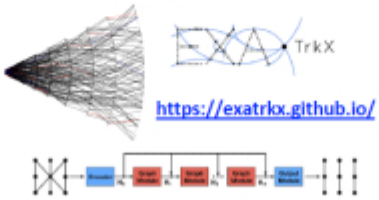
Exascale DL for climate analytics

Thorsten Kurth et. al arXiv:1810.01993



GraphNN for LHC Tracking

PI: Paolo Calafiura



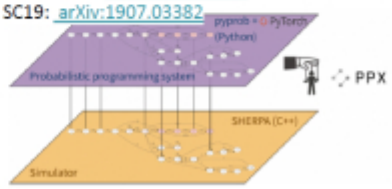
https://exatrkx.github.io/

Etalumis: Probabilistic Programming for Scientific Simulators at Scale

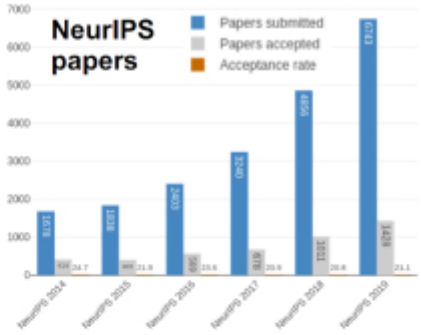
Atılım Baydin et. al

NeurIPS19: arXiv:1807.07706

SC19: arXiv:1907.03382



Adoption is on the rise



The scientific communities are enthusiastic

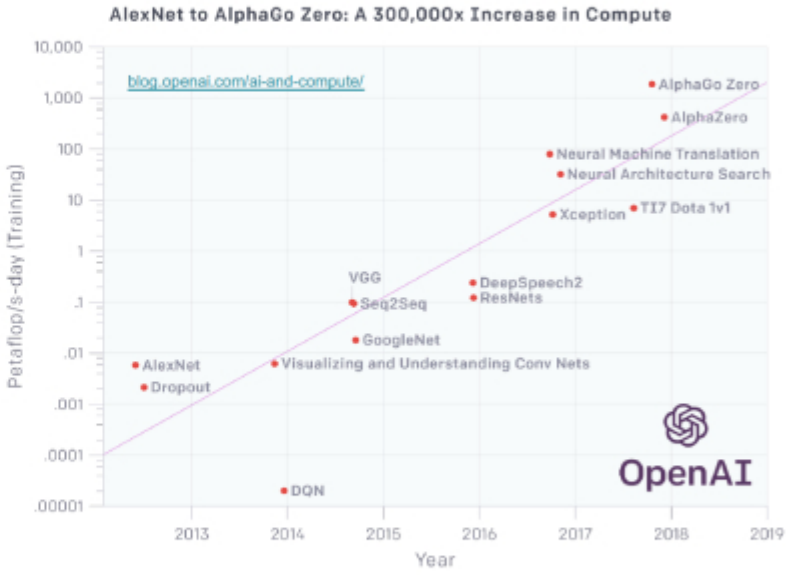
- Growing number of studies, papers
- Growing presence at ML+science conferences
- Recognition of achievements with awards:
  2018 Turing Award, 2018 Gordon Bell

The DOE is investing heavily

- Several funding calls in AI+science
- AI4Science townhall series, >1000 attendees
- AI4Science 300 page report

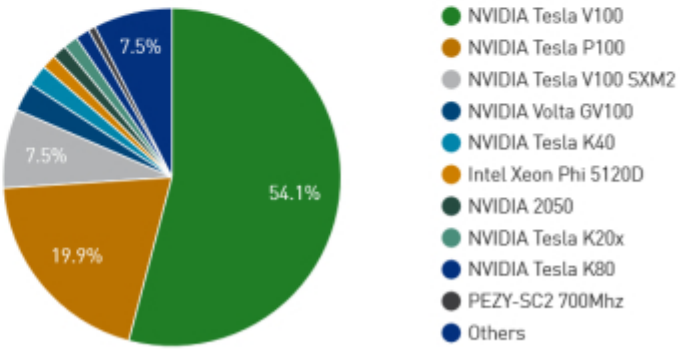
Growing computing needs

- More complex tasks, bigger models => more compute
- A single GPU just doesn't cut it for many DL problems now
- HPC systems are powerful resources to meet this demand



Deep Learning, Many-cores, and HPC

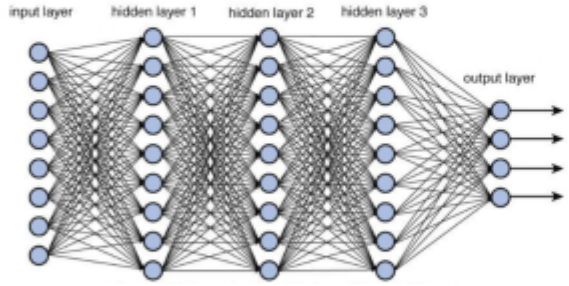
- In the High Performance Computing (HPC) area
  - ~90% in Top500 HPC systems use NVIDIA GPUs (Nov 2019)
  - CUDA-Aware Message Passing Interface (MPI)
  - NVIDIA Tesla and Volta architecture
    - Dedicated DL super-computers

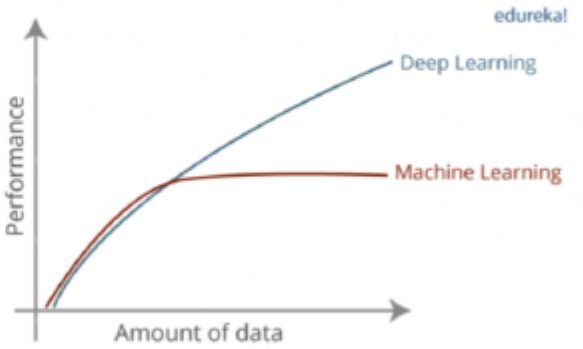


Source: https://www.top500.org

Deep Learning is a powerful set of tools

- Fueling an AI revolution, powering many recent technologies.
- Powered by deep neural networks

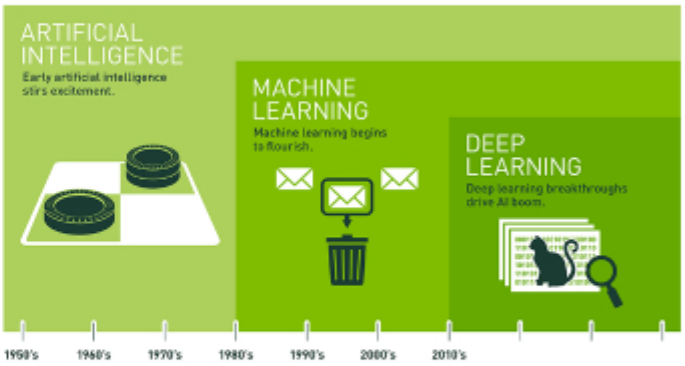




- Driven by the rise of GPUs and availability of large, curated datasets

Deep Learning (DL)

- Deep learning is a subset of AI and machine learning that uses multi-layered artificial neural networks to deliver state-of-the-art accuracy in tasks such as object detection, speech recognition, language translation and others.
- With NVIDIA GPU-accelerated deep learning frameworks, researchers and data scientists can significantly speed up deep learning training



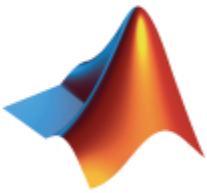
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

NVIDIA AI Platform

- Developing AI applications start with training deep neural networks with large datasets.
- GPU-accelerated **deep learning frameworks**
  - offer flexibility to design and
  - train custom deep neural networks and
  - provide interfaces to commonly-used programming languages such as Python and C/C++.
- NVidia provides **deep Learning SDK** high-performance libraries that implement building block APIs for implementing training and inference directly into their apps.

Deep Learning Frameworks

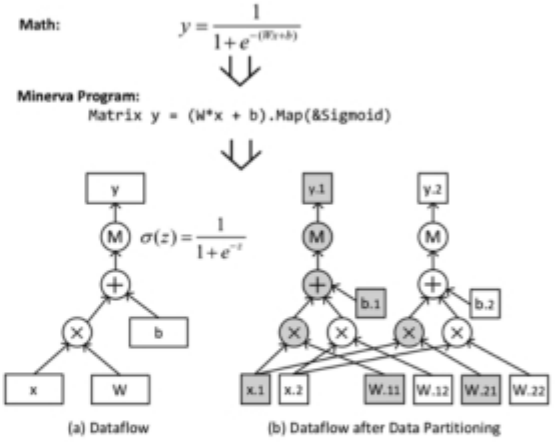
- Top Deep Learning frameworks (2020)
  - PyTorch
  - MXNet
  - TensorFlow
  - MATLAB
  - NVIDIA Caffe
  - Chainer
  - PaddlePaddle



Courtesy of https://developer.nvidia.com/deep-learning-frameworks

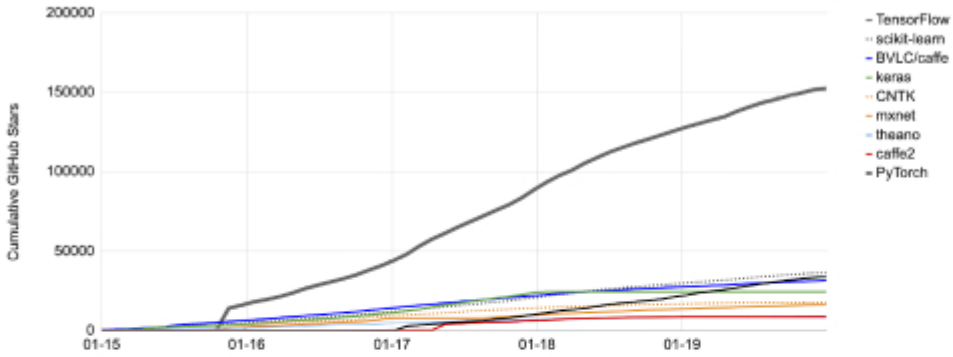
Why do we need DL frameworks?

- Deep Learning frameworks have emerged
  - hide most of the *annoying ☺ mathematics*
  - focus on the *design* of neural networks
- Distributed DL frameworks are being designed
  - We have saturated the peak potential of a single GPU/CPU/KNL
  - Parallel (multiple processing units in a single node) and/or Distributed (usually involves multiple nodes) frameworks are emerging
- Distributed frameworks are being developed along two directions
  - The HPC Eco-system: MPI-based Deep Learning
  - Enterprise Eco-system: BigData-based Deep Learning

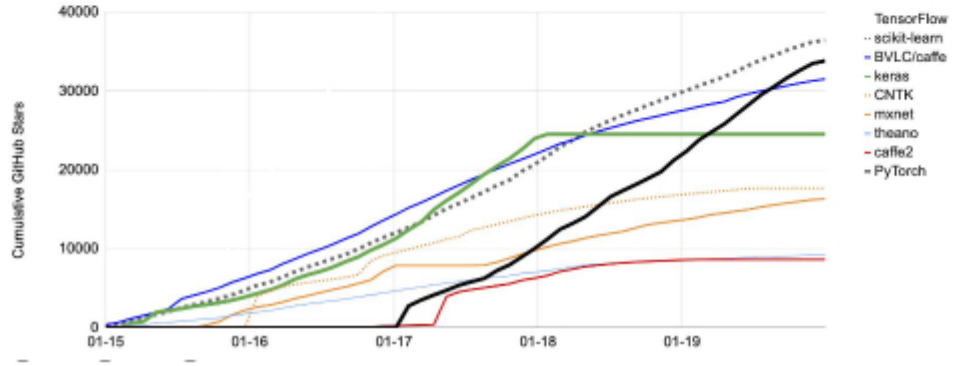


DL Frameworks and GitHub Statistics

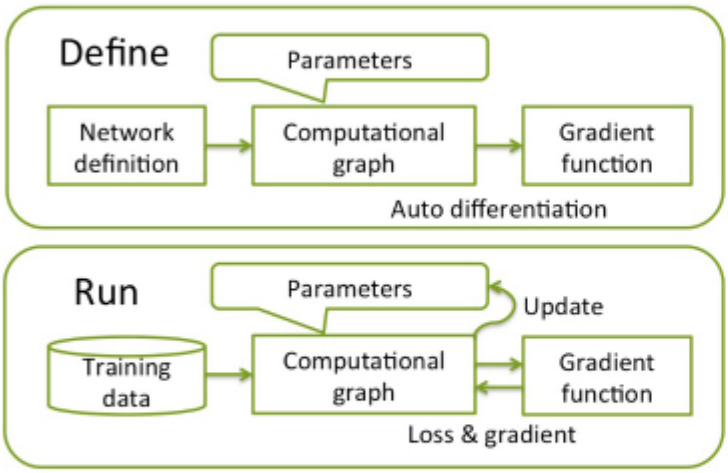
Cumulative GitHub stars by AI library (2015—2019)
Source: Github, 2019.

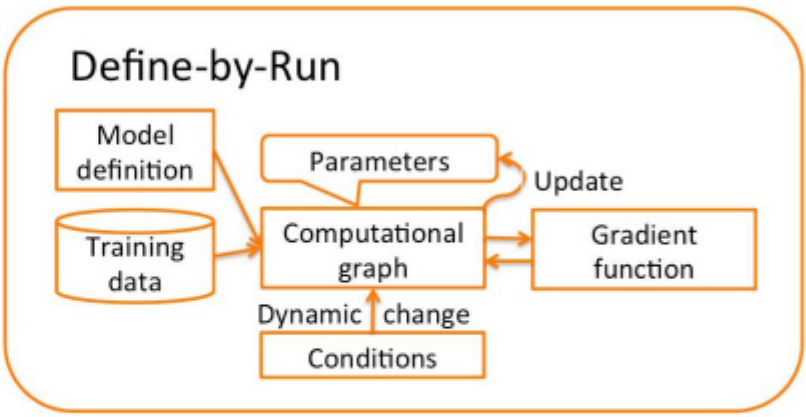


Cumulative GitHub stars by AI library, not including TensorFlow (2015—2019)
Source: Github, 2019.



Define-by-run frameworks vs. Define-and-run?





- Define-and-run: TensorFlow, Caffe, Torch, Theano, and others
- Define-by-run
  - PyTorch, MXNet and Chainer
  - TensorFlow 2 with Eager Execution (Define-by-run) mode

Torch/PyTorch

- Torch was written in Lua
- Adoption wasn't wide-spread
- PyTorch is a Python adaptation of Torch
- Biggest support by Facebook
- Key selling point is ease of expression and "define-by-run" approach

Refer to: http://pytorch.org

MXNet

- Apache MXNet is an open-source deep learning framework, used to train, and deploy deep neural networks.
- It is scalable, allowing for fast model training, and supports a flexible programming model and multiple programming languages.
- The MXNet library is portable and can scale to multiple GPUs and multiple machines.

Refer to: https://mxnet.apache.org/versions/1.8.0/

Google TensorFlow

- The most widely used framework open-sourced by Google
- Runs on almost all execution platforms available (CPU, GPU, TPU, Mobile, etc.)
- Very flexible but performance has been an issue
- Certain Python peculiarities like variable_scope, etc.

Refer to:: https://www.tensorflow.org/

Caffe/Caffe2/NVCaffe

- Yangqing Jia (BVLC)
  - Author of Caffe and Caffe2 (Facebook)
- The framework has a modular C++ backend
- C++ and Python frontends
- Caffe is a single-node but multi-GPU framework
- Caffe2 is now a part of PyTorch.
- NVCaffe is an NVIDIA-maintained fork of BVLC Caffe tuned for NVIDIA GPUs, particularly in multi-GPU configurations.

Microsoft Cognitive Toolkit (CNTK)

- Formerly CNTK, now called the Cognitive Toolkit
- C++ backend
- C++ and Python frontend
- ASGD, SGD, and several others choices for Solvers/Optimizers
- Constantly evolving support for multiple platforms
- Performance has always been the “key feature”

Refer to: https://docs.microsoft.com/en-us/cognitive-toolkit/

Other Popular DL Frameworks…

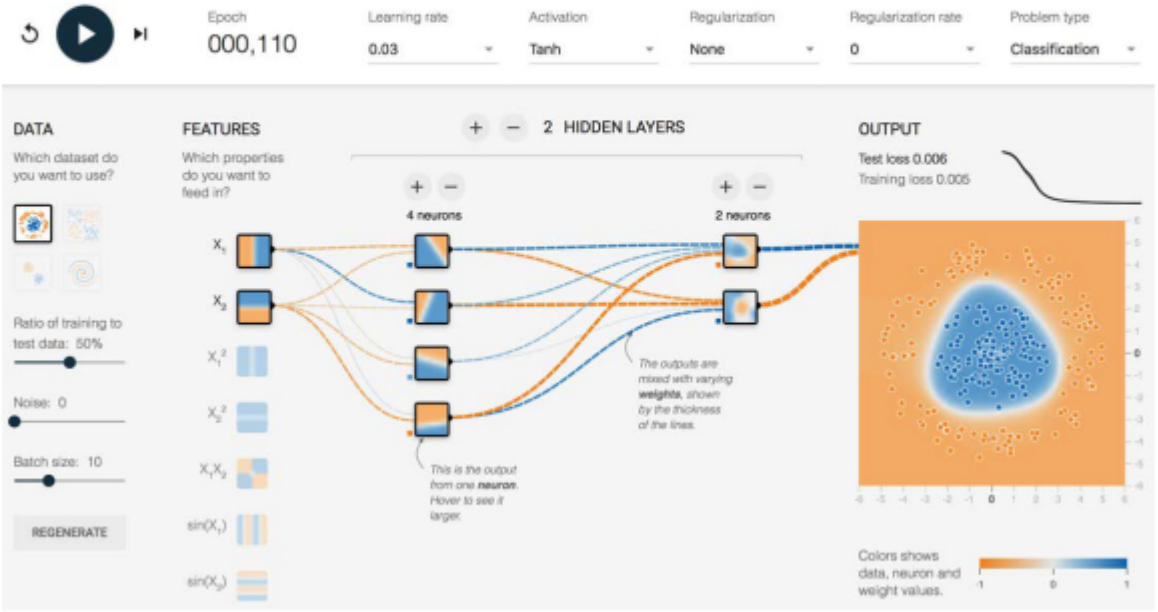
- Chainer - https://chainer.org/
- PaddlePaddle - https://github.com/PaddlePaddle/Paddle
- Keras - https://keras.io
- ……

So where do we run our DL framework?

- Early (2014) frameworks used a single fast GPU
  - As DNNs became larger, faster and better GPUs became available
  - At the same time, parallel (multi-GPU) training gained traction as well
- Today
  - Parallel training on multiple GPUs is being supported by most frameworks
  - Distributed (multiple nodes) training is still upcoming
    - ✔ A lot of fragmentation in the efforts (MPI, Big-Data, NCCL, Gloo, etc.)
  - On the other hand, DL has made its way to Mobile and Web too!
    - ✔ Smartphones -- OK Google, Siri, Cortana, Alexa, etc.
    - ✔ DrivePX – the computer that drives NVIDIA's self-driving car
    - ✔ Google announced Deeplearn.js (a DL framework in a web-browser)
    - ✔ TensorFlow playground -- http://playground.tensorflow.org/

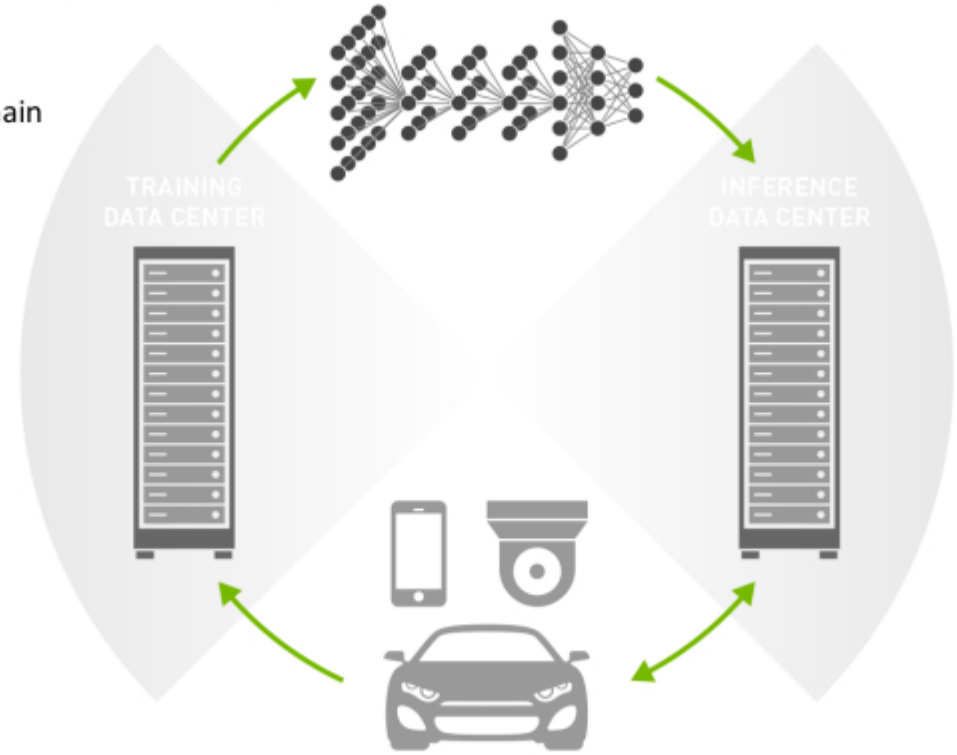
TensorFlow playground (Quick Demo)

- To actually train a network, please visit: http://playground.tensorflow.org



Accelerating in GPU Data Center Platform

NVidia GPUs are the main driving force for faster training of DL models

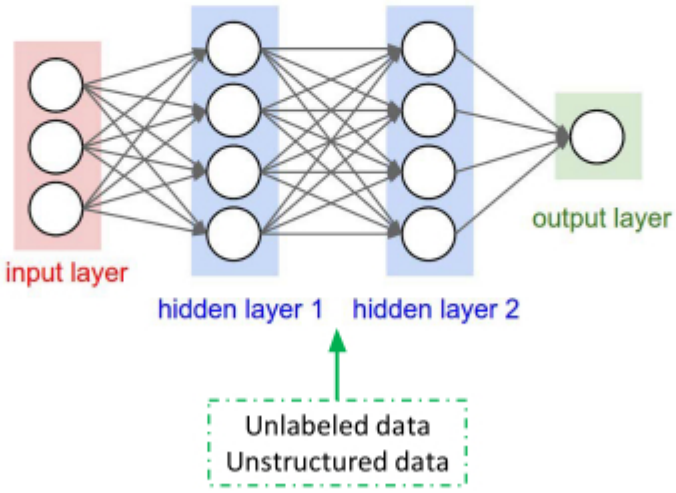


Diverse Application Areas for Deep Learning

- Vision
  - Image Classification
  - Style Transfer
  - Caption Generation
- Speech
  - Speech Recognition
  - Real-time Translation
- Text
  - Sequence Recognition and Generation
- Disease discovery
  - Cancer Detection
- Autonomous Driving
  - Combination of multiple areas like Image/Object Detection, Speech Recognition, etc.

So what is a Deep Neural Network?

- Example of a 3-layer Deep Neural Network (DNN) –(input layer is not counted)



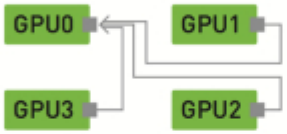
NVIDIA Deep Learning SDK

Deep Learning Primitives

cuDNN

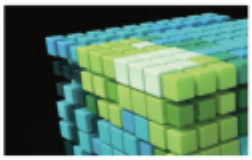
Multi-GPU Communication

NCCL



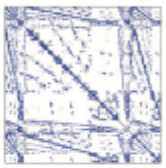
Linear Algebra

cuBLAS



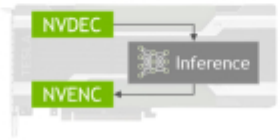
Sparse Matrix Operations

cuSPARSE



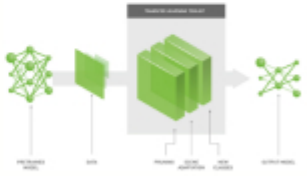
TensorRT

DeepStream SDK



Optical Flow SDK

Transfer Learning Toolkit



Deep Learning Inference Engine

Deep Learning for Video Analytics

High level SDK for tuning domain specific DNNs

Optical Flow for Video Inference

- Stay safe!
- See you next class!

Next Lecture will Continue:

GPU and CUDA