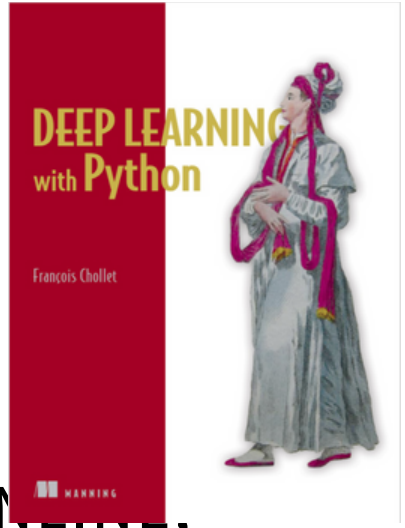
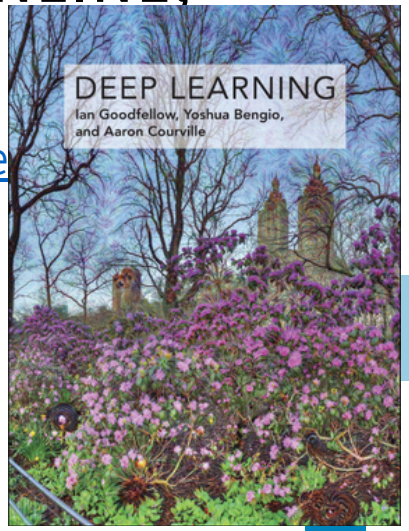
Lecture 6: Introduction to deep learning

Practical deep learning

Further resources

- This course is largely “inspired by”: “Deep Learning with Python” by François Chollet
- Recommended textbook: “Deep learning” by Goodfellow, Bengio, Courville
- LOTS OF FURTHER MATERIAL AVAILABLE ONLINE, E.G.:
  http://cs231n.stanford.edu/ http://course.fast.ai/
  https://developers.google.com/machine-learning/crash-course
  www.nvidia.com/dlilabs http://introtodeeplearning.com/
  https://github.com/oxford-cs-deepnlp-2017/lectures,
  https://jalammar.github.io/
- Academic courses





What is artificial intelligence?

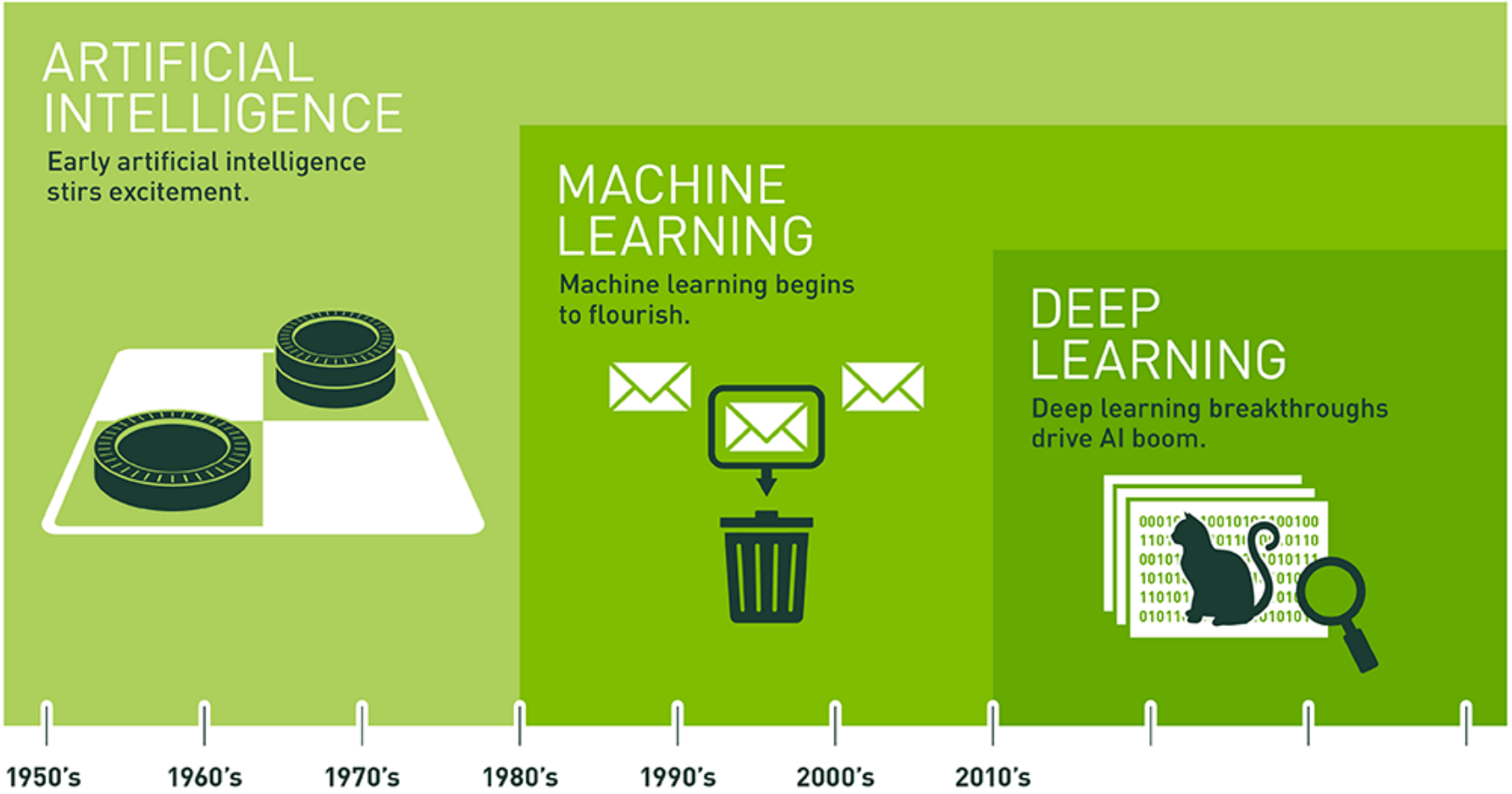
Artificial intelligence is the ability of a computer to perform tasks commonly associated with intelligent beings.

What is machine learning?

Machine learning is the study of algorithms that learn from examples and experience instead of relying on hard-coded rules and make predictions on new data.

What is deep learning?

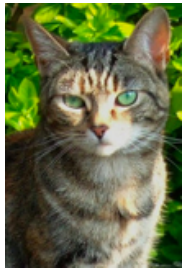
Deep learning is a subfield of machine learning focusing on learning data representations as successive layers of increasingly meaningful representations.



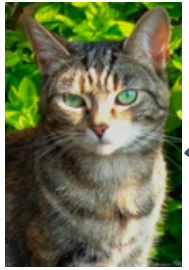
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

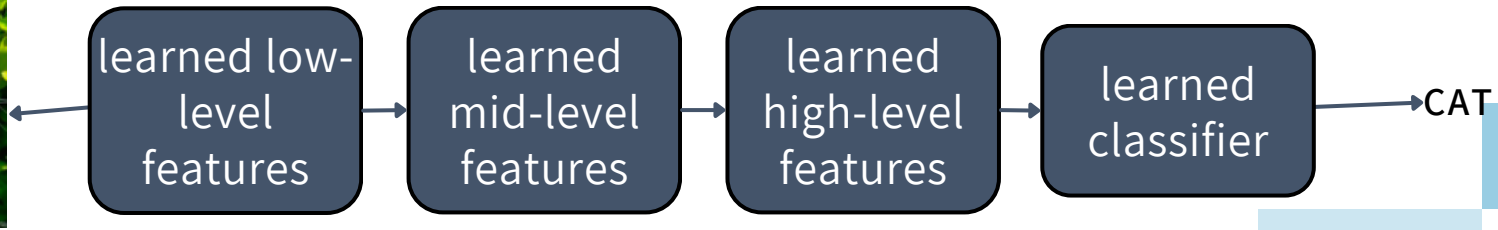
Image from https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

“Traditional” machine learning:

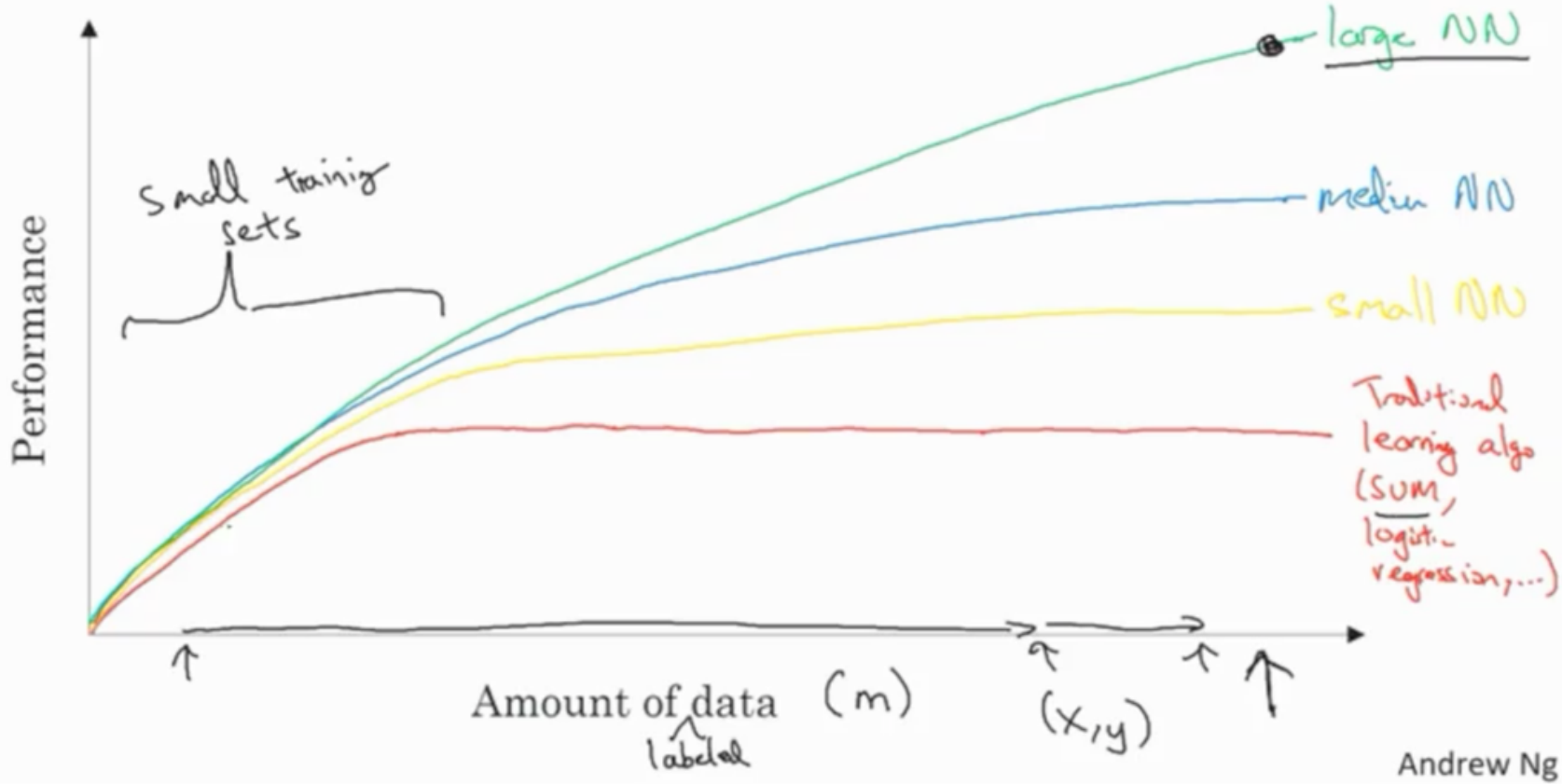


Deep, “end-to-end” learning:





Scale drives deep learning progress



Andrew Ng

Table 1: Major milestones that will be covered in this paper

Year	Contributer	Contribution
300 BC	Aristotle	introduced Associationism, started the history of human's attempt to understand brain.
1873	Alexander Bain	introduced Neural Groupings as the earliest models of neural network, inspired Hebbian Learning Rule.
1943	McCulloch & Pitts	introduced MCP Model, which is considered as the ancestor of Artificial Neural Model.
1949	Donald Hebb	considered as the father of neural networks, introduced Hebbian Learning Rule, which lays the foundation of modern neural network.
1958	Frank Rosenblatt	introduced the first perceptron, which highly resembles modern perceptron.
1974	Paul Werbos	introduced Backpropagation
1980	Teuvo Kohonen	introduced Self Organizing Map
	Kunihiko Fukushima	introduced Neocogitron, which inspired Convolutional Neural Network
1982	John Hopfield	introduced Hopfield Network
1985	Hilton & Sejnowski	introduced Boltzmann Machine
1986	Paul Smolensky	introduced Harmonium, which is later known as Restricted Boltzmann Machine
	Michael I. Jordan	defined and introduced Recurrent Neural Network
1990	Yann LeCun	introduced LeNet, showed the possibility of deep neural networks in practice
1997	Schuster & Paliwal	introduced Bidirectional Recurrent Neural Network
	Hochreiter & Schmidhuber	introduced LSTM, solved the problem of vanishing gradient in recurrent neural networks
2006	Geoffrey Hinton	introduced Deep Belief Networks, also introduced layer-wise pretraining technique, opened current deep learning era.
2009	Salakhutdinov & Hinton	introduced Deep Boltzmann Machines
2012	Geoffrey Hinton	introduced Dropout, an efficient way of training neural networks

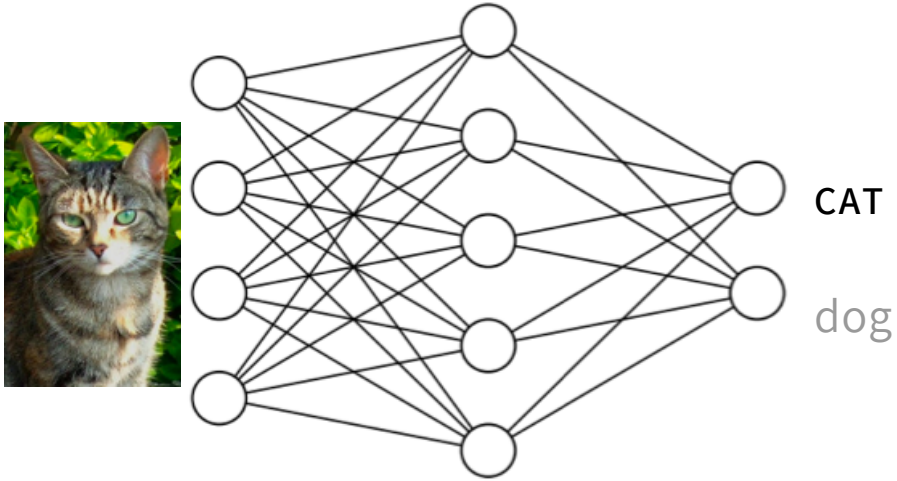
FROM: WANG & RAJ: ON THE ORIGIN OF DEEP LEARNING (2017)

MAIN TYPES OF MACHINE LEARNING

Main types of machine learning

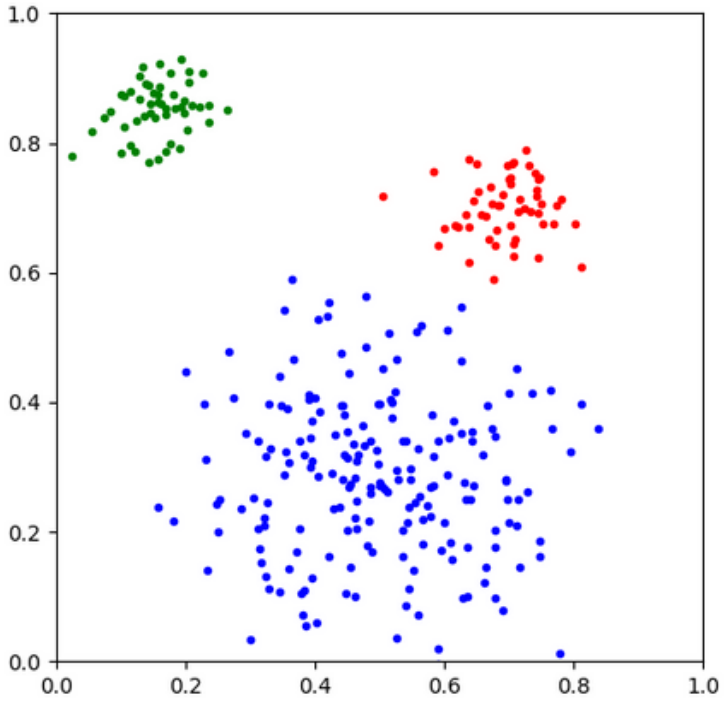
- **Supervised learning**

- UNSUPERVISED LEARNING
- SELF-SUPERVISED LEARNING
- REINFORCEMENT LEARNING



MAIN TYPES OF MACHINE LEARNING

- Supervised learning
- UNSUPERVISED LEARNING
- Self-supervised learning
- Reinforcement learning



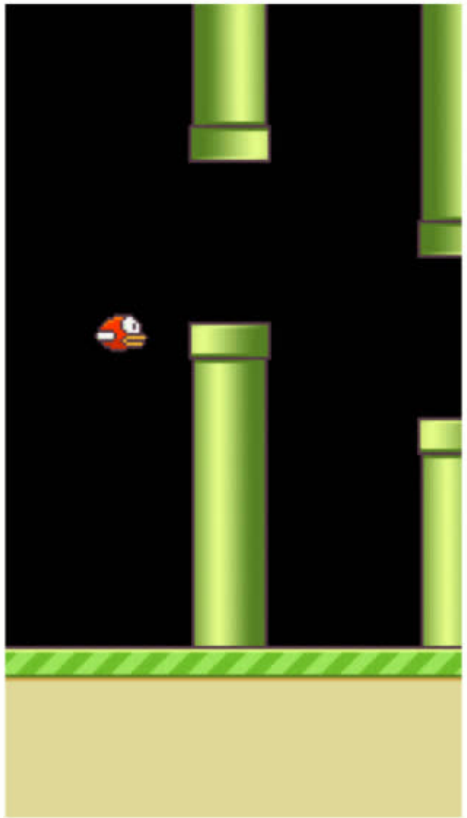
MAIN TYPES OF MACHINE LEARNING

- SUPERVISED LEARNING
- UNSUPERVISED LEARNING
- **Self-supervised learning**
- REINFORCEMENT LEARNING

Image from https://arxiv.org/abs/1710.10196

Main types of machine learning

- SUPERVISED LEARNING
- UNSUPERVISED LEARNING
- SELF-SUPERVISED LEARNING
- **Reinforcement learning**

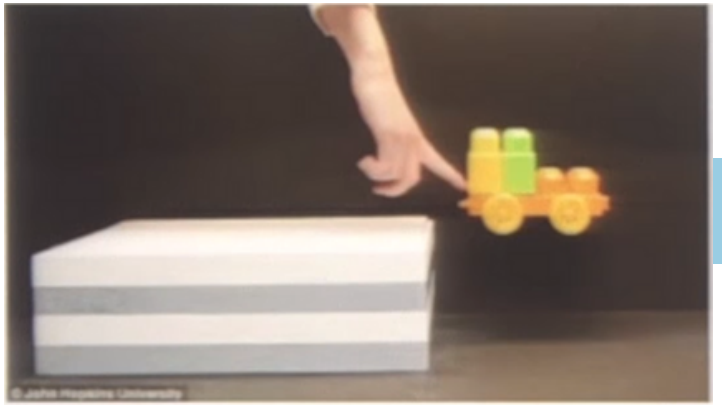


Animation from https://yanpanlau.github.io/2016/07/10/FlappyBird-Keras.html

FUNDAMENTALS OF MACHINE LEARNING

Data

- HUMANS LEARN BY OBSERVATION AND UNSUPERVISED LEARNING
  - MODEL OF THE WORLD / COMMON SENSE REASONING
- MACHINE LEARNING NEEDS LOTS OF (LABELED) DATA TO COMPENSATE



Data

- TENSORS: GENERALIZATION OF MATRICES
  TO N DIMENSIONS (OR RANK, ORDER, DEGREE)
  - 1D TENSOR: VECTOR
  - 2D TENSOR: MATRIX
  - 3D, 4D, 5D TENSORS
  - NUMPY.NDARRAY(SHAPE, DTYPE)
- TRAINING – VALIDATION – TEST SPLIT (+ ADVERSARIAL TEST)
- MINIBATCHES
  - SMALL SETS OF INPUT DATA USED AT A TIME
  - USUALLY PROCESSED INDEPENDENTLY

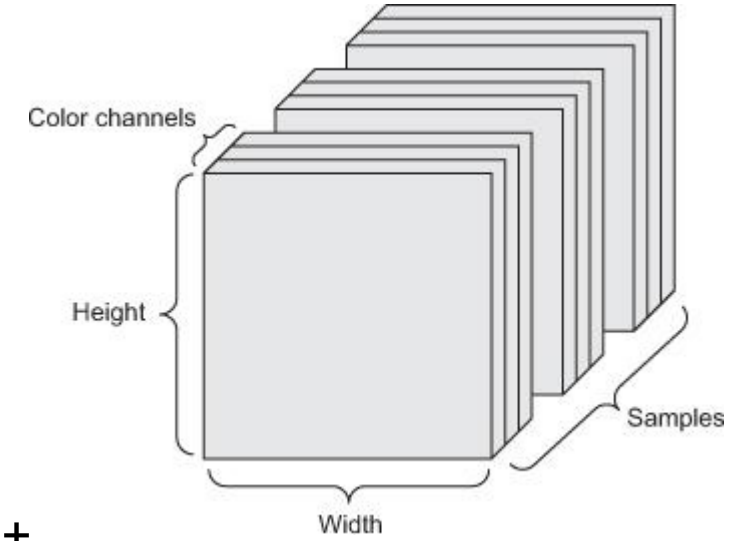
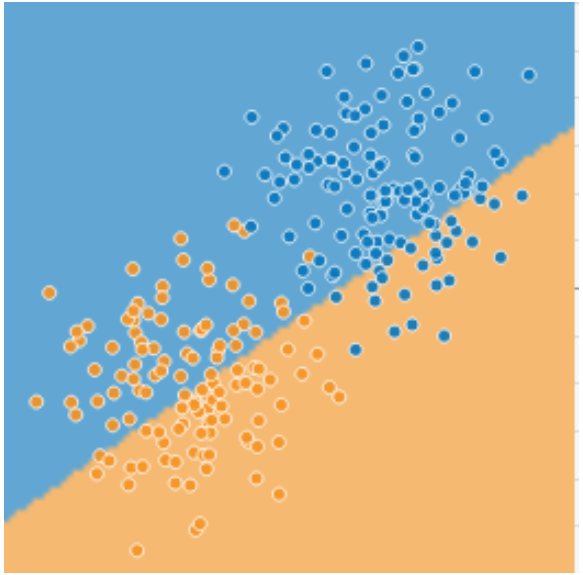
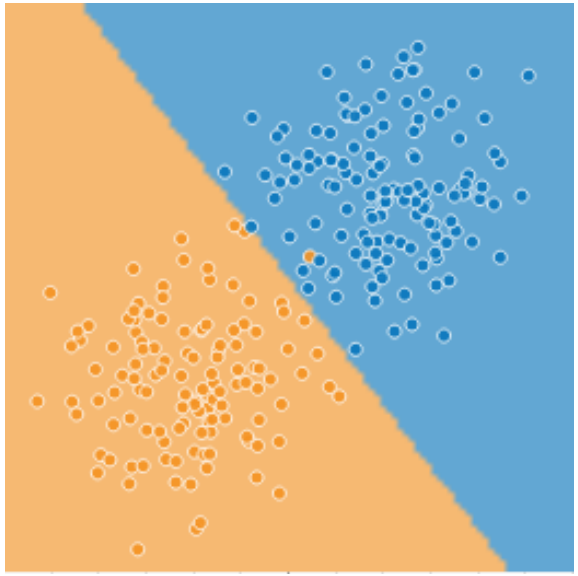
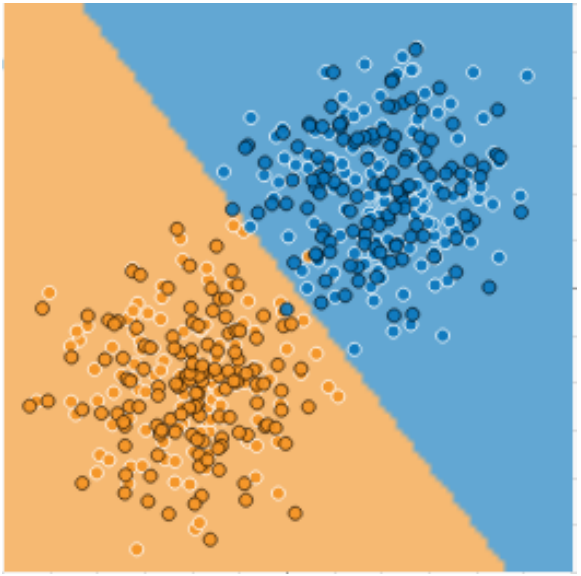


Image from: https://arxiv.org/abs/1707.08945

Model – learning/training – inference

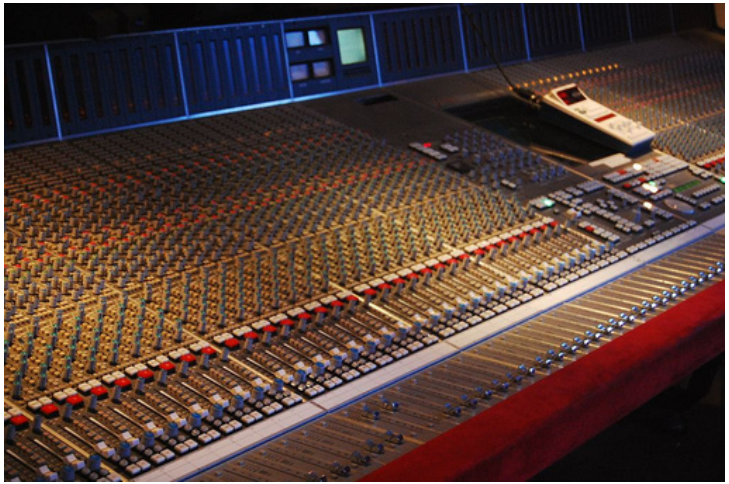




http://playground.tensorflow.org/

- $\hat{y} = f(\mathbf{x}; \theta)$
- PARAMETERS θ AND HYPERPARAMETERS

Optimization

- MATHEMATICAL OPTIMIZATION:
  "THE SELECTION OF A BEST ELEMENT (WITH
  REGARD TO SOME CRITERION) FROM SOME
  SET OF AVAILABLE ALTERNATIVES"
  (WIKIPEDIA)
- MAIN TYPES:
  FINITE-STEP, ITERATIVE, HEURISTIC
- LEARNING AS AN OPTIMIZATION PROBLEM
  - cost function:



By Rebecca Wilson (originally posted to Flickr as Vicariously) [CC BY 2.0], via Wikimedia Commons

LOSS

REGULARIZATION

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} L(f(\mathbf{x}_i;\theta), y_i) + R(\theta)$$

Optimization

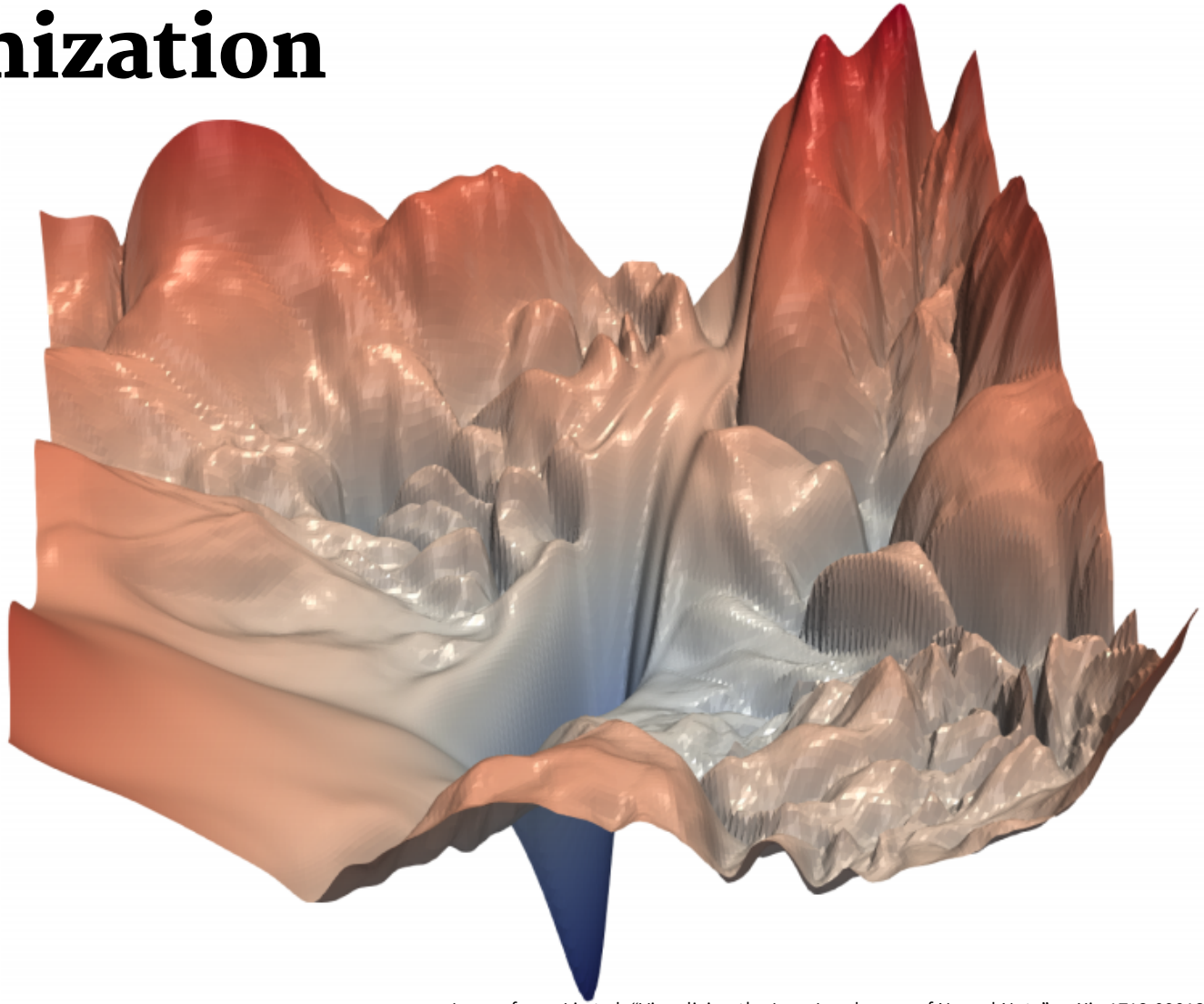


Image from: Li et al. “Visualizing the Loss Landscape of Neural Nets”, arXiv:1712.09913

Gradient descent

- DERIVATIVE AND MINIMA/MAXIMA OF FUNCTIONS
- GRADIENT: THE DERIVATIVE OF A MULTIVARIABLE FUNCTION
- GRADIENT DESCENT:

$$\theta_{t+1} = \theta_t - \alpha \frac{\partial J(\theta)}{\partial \theta}$$

- (MINI-BATCH) STOCHASTIC GRADIENT DESCENT (AND ITS VARIANTS)

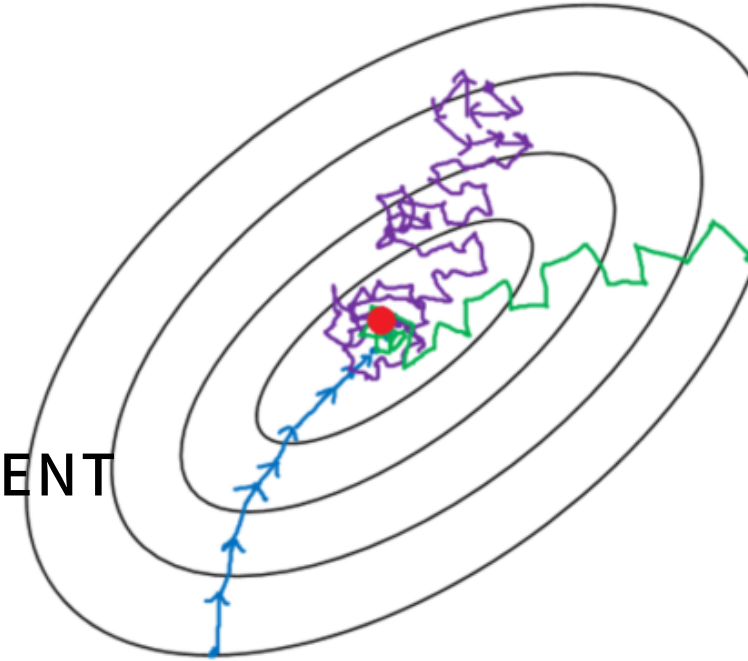
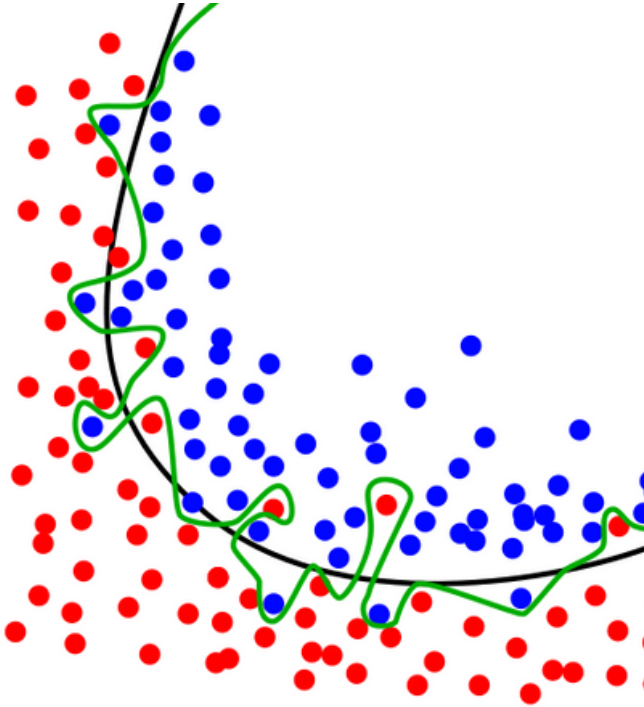


Image from: https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3

Over- and underfitting, generalization, regularization

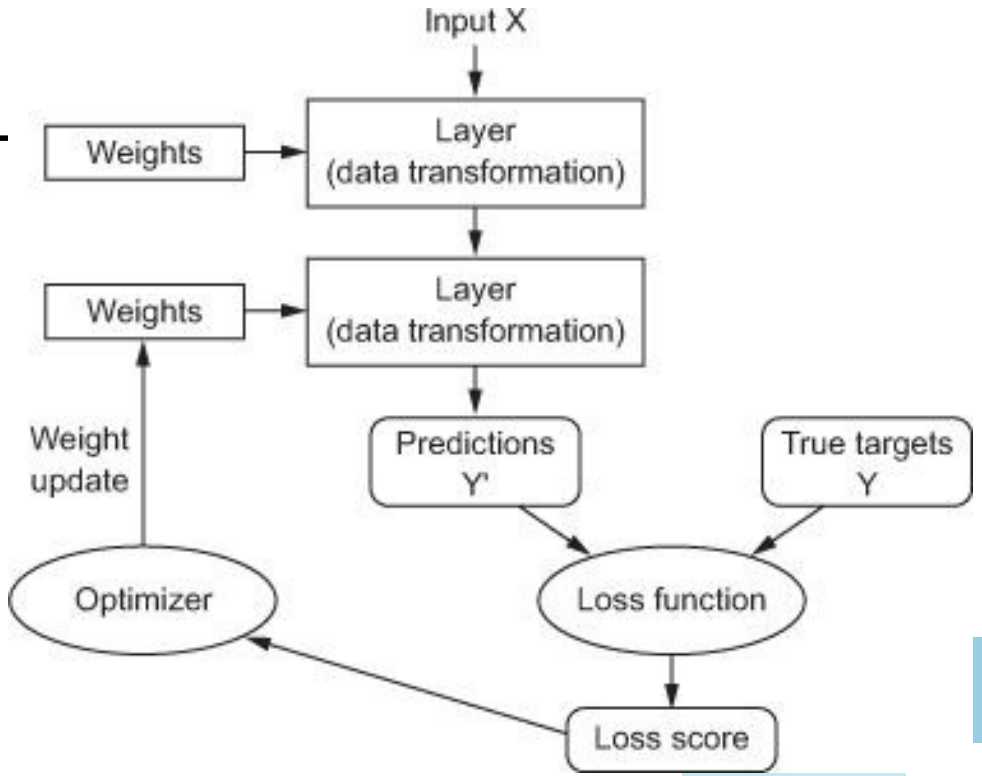
- MODELS WITH LOTS OF PARAMETERS CAN EASILY OVERFIT TO TRAINING DATA
- **Generalization: the quality of ML model is measured on new, unseen samples**
- **Regularization: any method* to prevent overfitting**
  - SIMPLICITY, SPARSITY, DROPOUT, EARLY STOPPING
  - *) OTHER THAN ADDING MORE DATA



Deep learning

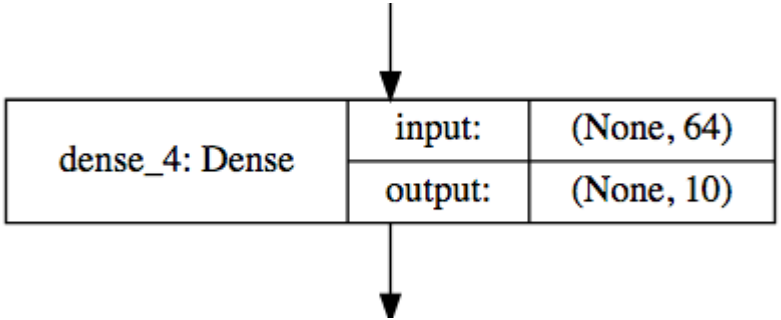
Anatomy of a deep neural network

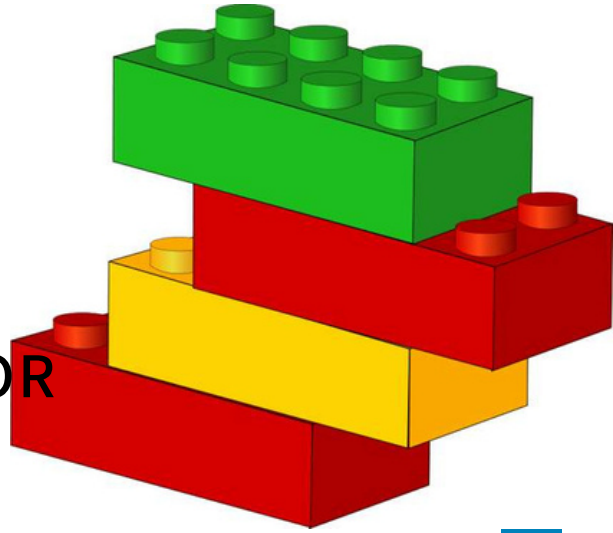
- LAYERS
- INPUT DATA AND TARGET
- LOSS FUNCTION
- OPTIMIZER



Layers

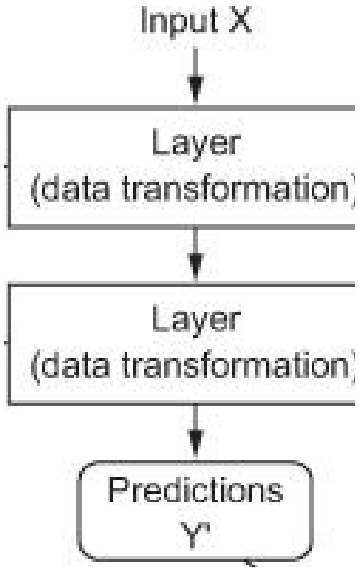
- DATA PROCESSING MODULES
- MANY DIFFERENT KINDS EXIST
  - densely connected
  - convolutional
  - recurrent
  - pooling, flattening, merging, normalization, etc.
- INPUT: ONE OR MORE TENSORS
  OUTPUT: ONE OR MORE TENSORS
- USUALLY HAVE A STATE, ENCODED AS WEIGHTS
  - learned, initially random
- WHEN COMBINED, FORM A NETWORK OR A MODEL

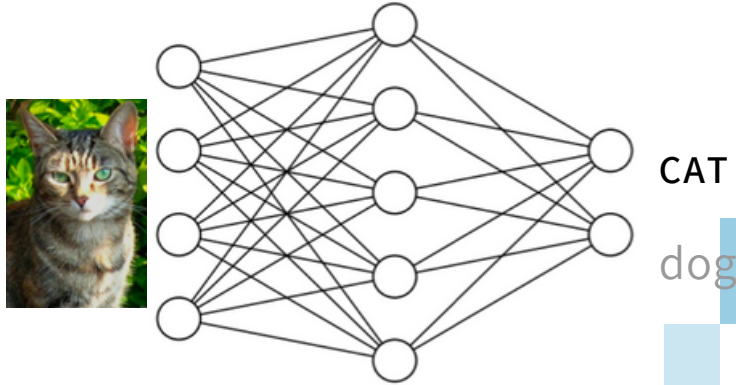




Input data and targets

- THE NETWORK MAPS THE INPUT DATA X TO PREDICTIONS Y′
- DURING TRAINING, THE PREDICTIONS Y′ ARE COMPARED TO TRUE TARGETS Y USING THE LOSS FUNCTION



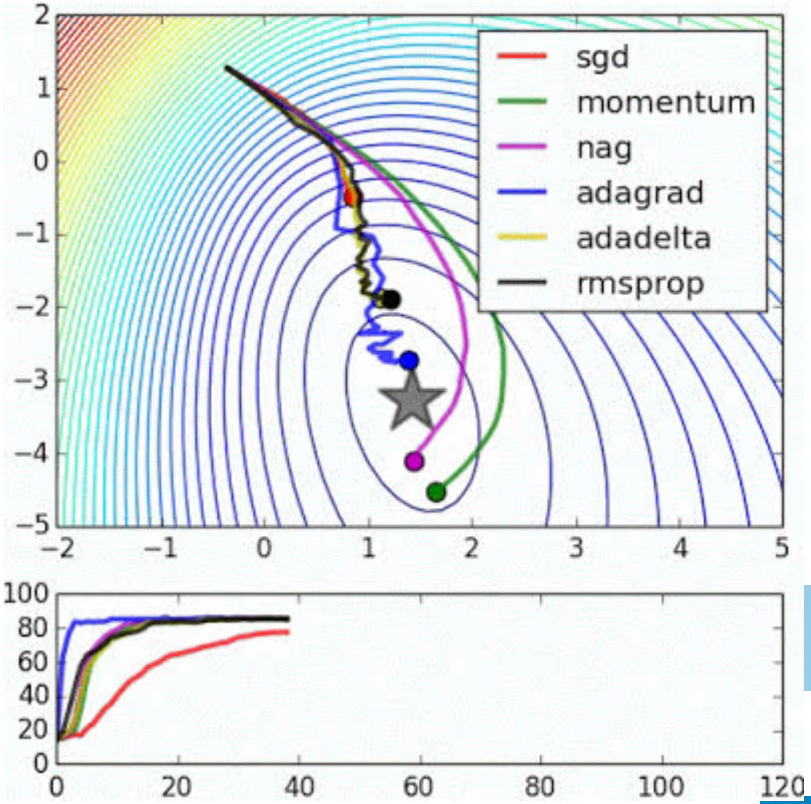


Loss function

- THE QUANTITY TO BE MINIMIZED (OPTIMIZED) DURING TRAINING
  - THE ONLY THING THE NETWORK CARES ABOUT
  - THERE MIGHT ALSO BE OTHER METRICS YOU CARE ABOUT
- COMMON TASKS HAVE “STANDARD” LOSS FUNCTIONS:
  - MEAN SQUARED ERROR FOR REGRESSION
  - BINARY CROSS-ENTROPY FOR TWO-CLASS CLASSIFICATION
  - CATEGORICAL CROSS-ENTROPY FOR MULTI-CLASS CLASSIFICATION
  - ETC.
- HTTPS://LOSSFUNCTIONS.TUMBLR.COM/

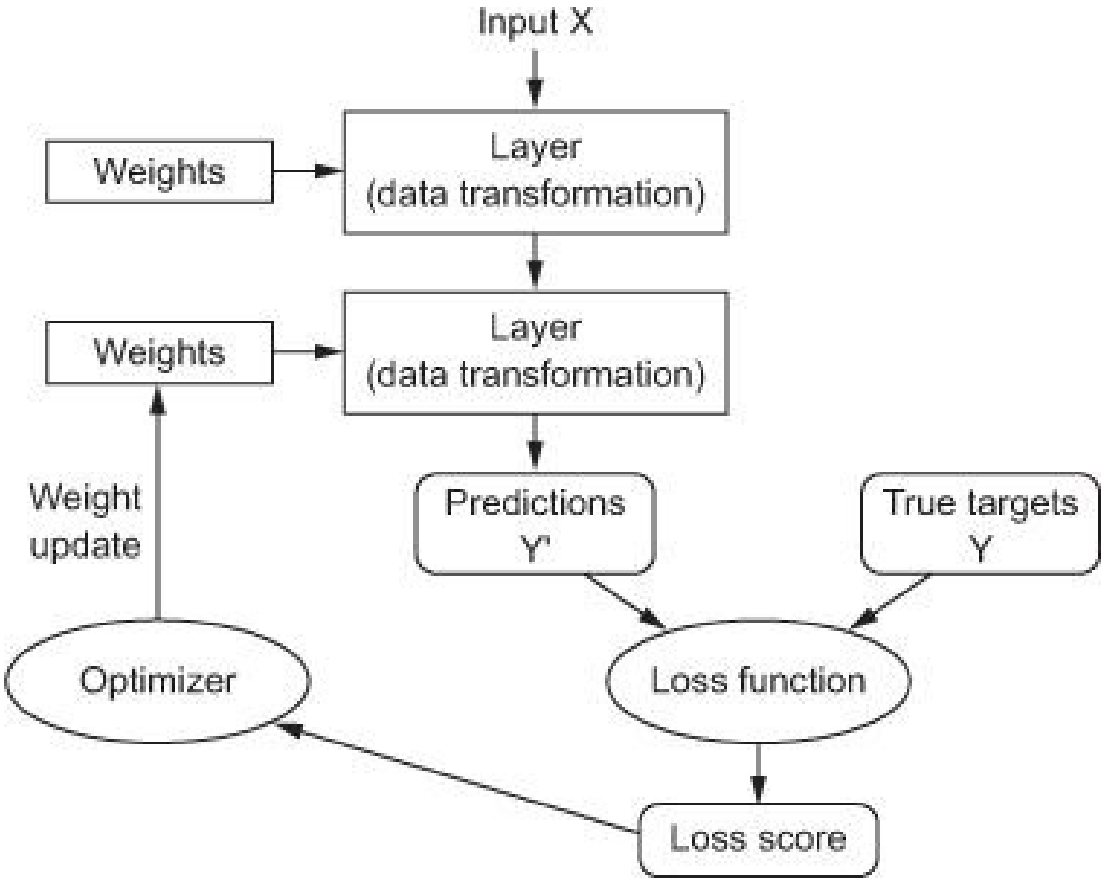
Optimizer

- HOW TO UPDATE THE WEIGHTS BASED ON THE LOSS FUNCTION
- LEARNING RATE (+SCHEDULING)
- STOCHASTIC GRADIENT DESCENT, MOMENTUM, AND THEIR VARIANTS
  - RMSPROP IS USUALLY A GOOD FIRST CHOICE
  - more info: http://ruder.io/optimizing-gradient-descent/



Animation from: https://imgur.com/s25RsOr

Anatomy of a deep neural network



Deep learning frameworks


Caffe

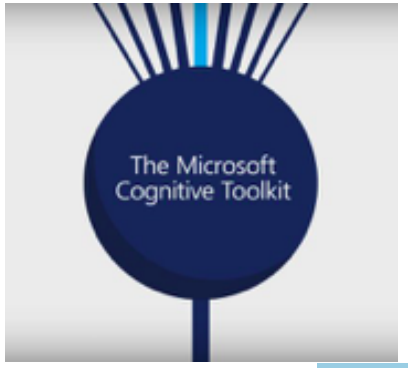

torch


Chainer

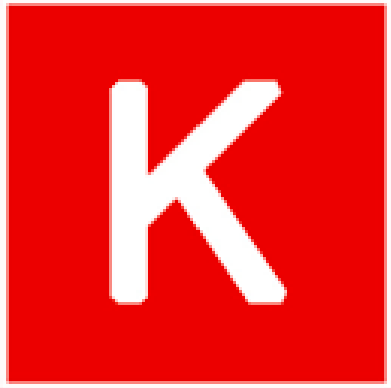

PyTorch


DEEPLEARNING4J


Caffe2


The Microsoft
Cognitive Toolkit


TensorFlow

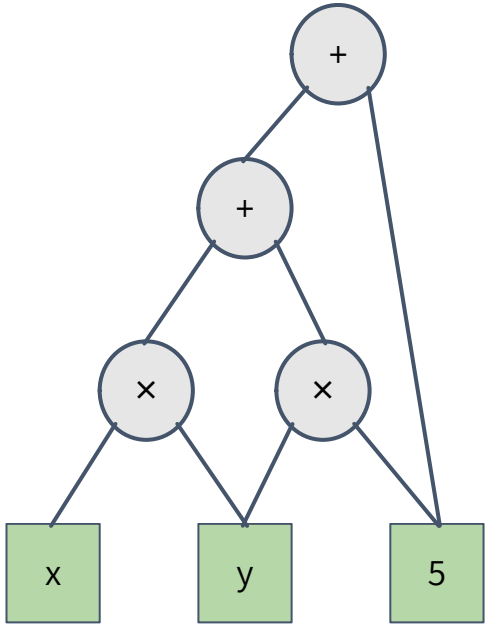


dmlc
mxnet


theano

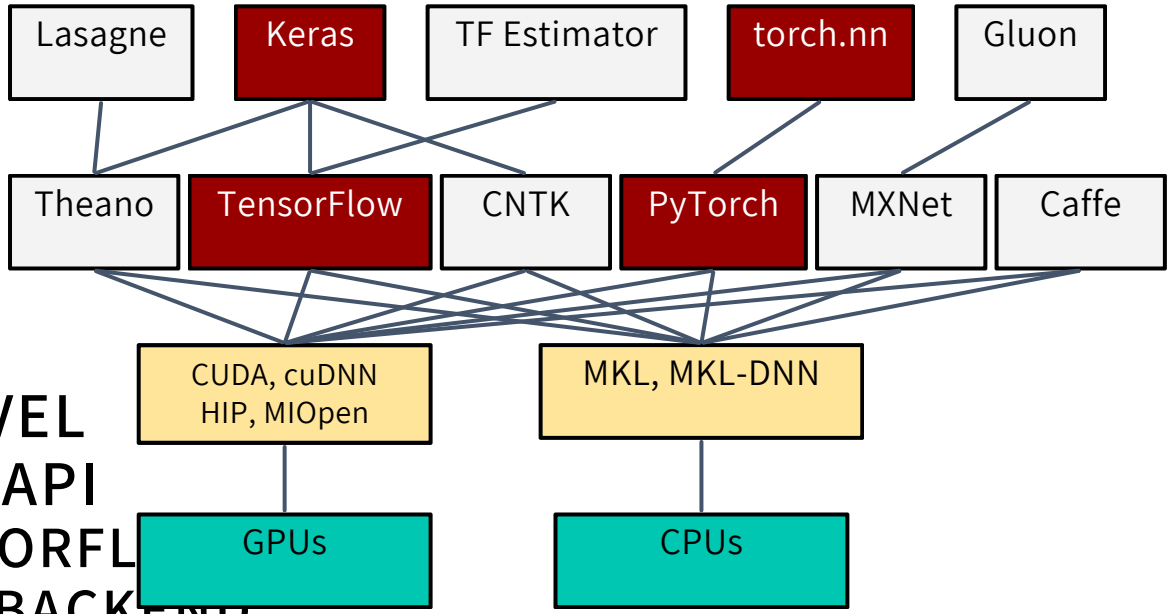
Deep learning frameworks

- ACTUALLY TOOLS FOR DEFINING STATIC OR DYNAMIC GENERAL-PURPOSE COMPUTATIONAL GRAPHS
- AUTOMATIC DIFFERENTIATION
- SEAMLESS CPU / GPU USAGE
  - MULTI-GPU, DISTRIBUTED
- PYTHON/NUMPY OR R INTERFACES
  - INSTEAD OF C, C++, CUDA OR HIP
- OPEN SOURCE



$$cy + 5y + 5$$

Deep learning frameworks



- KERAS IS A HIGH-LEVEL NEURAL NETWORKS API
  - WE WILL USE TENSORFL AS THE COMPUTE BACKEND
  - INCLUDED IN TENSORFLOW 2 AS TF.KERAS
  - HTTPS://KERAS.IO/ , HTTPS://WWW.TENSORFLOW.ORG/GUIDE/KERAS
- PYTORCH IS:
  - A GPU-BASED TENSOR LIBRARY
  - AN EFFICIENT LIBRARY FOR DYNAMIC NEURAL NETWORKS
  - HTTPS://PYTORCH.ORG/