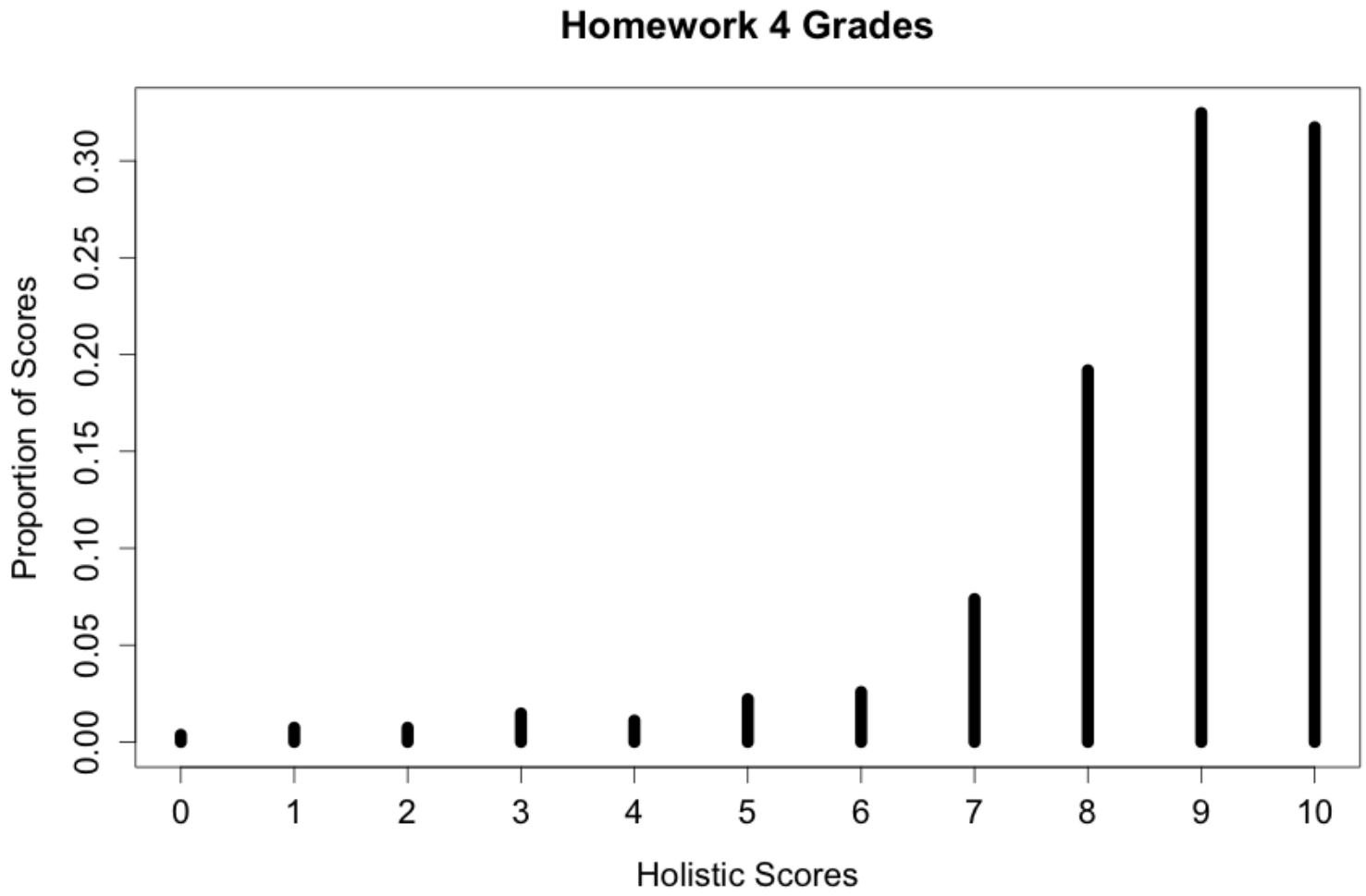
Announcements



Deep Learning

What is Deep learning?

- buzzword

Reinventing Social Media: Deep Learning, Predictive Marketing, And Image Recognition Will Change Everything

COOPER SMITH
FEB. 16, 2014, 6:02 PM | 6,821

CULTURE
→Animals, Civil Liberties, Tech, Top Stories

Why Facebook, Google, and the NSA Want Computers That Learn Like Humans

Deep learning could transform artificial intelligence. It could also get pretty creepy.

—By ***Dana Liebelson*** | September/October 2014 Issue

Like Share 454 Tweet 490 Email 71

Scientists See Promise in Deep-Learning Programs

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

BY GARY MARCUS

Share Tweet g+1

Deep Learning - The Biggest Data Science Breakthrough of the Decade

One of the last deep learning professors to remain completely in academia.

(http://www.reddit.com/r/MachineLearning/comments/1ysry1/ama_yoshua_bengio)

What is Deep learning?

- buzzword
- state-of-the art / breakthrough in machine learning

Automatic speech recognition

Method	PER (%)
Randomly Initialized RNN	26.1
Bayesian Triphone GMM-HMM	25.6
Monophone Randomly Initialized DNN	23.4
Monophone DBN-DNN	22.4
Triphone GMM-HMM with BMMI Training	21.7
Monophone DBN-DNN on fbank	20.7
Convolutional DNN	20.0

Image classification

Type	Classifier	Preprocessing	Error rate (%)
Linear classifier	Pairwise linear classifier	Deskewing	7.6[9]
K-Nearest Neighbors	K-NN with non-linear deformation (P2DHMDM)	Shiftable edges	0.52[14]
Boosted Stumps	Product of stumps on Haar features	Haar features	0.87[15]
Non-Linear Classifier	40 PCA + quadratic classifier	None	3.3[9]
Support vector machine	Virtual SVM, deg-9 poly, 2-pixel jittered	Deskewing	0.56[16]
Neural network	6-layer NN 784-2500-2000-1500-1000-500-10 (on GPU), with elastic distortions	None	0.35[17]
Convolutional neural network	Committee of 35 conv. net, 1-20-P-40-P-150-10, with elastic distortions	Width normalizations	0.23[8]

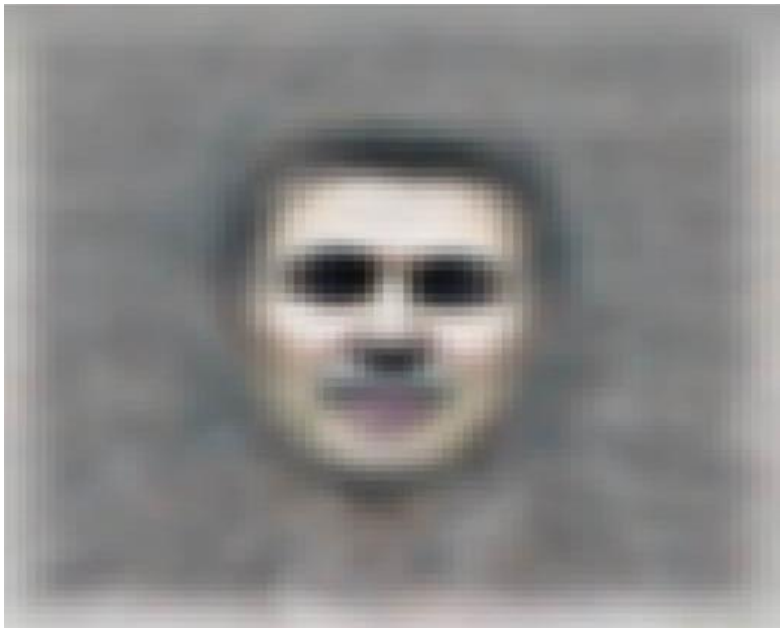
http://en.wikipedia.org/wiki/Deep_learning

Dan Claudiu Cireșan

- **Competitions (all methods use my NN framework)**
- NEW! First place at Assessment of Mitosis Detection Algorithms, MICCAI 2013 Grand Challenge, Nagoya, Japan (with Alessandro Giusti).
- NEW! Best score on test set from Chinese Handwriting Recognition Competition; task: offline characters, ICDAR 2013, Dallas, US - details in Multi-Column Deep Neural Networks for Offline Handwritten Chinese Character Classification - IDSIA Technical Report, August 2013.
- First place at Mitosis Detection in Breast Cancer Histological Images, ICPR 2012, Tsukuba, Japan (with Alessandro Giusti).
- First place at Segmentation of neuronal structures in EM stacks challenge - ISBI 2012, Barcelona, Spain (with Alessandro Giusti). We were the only team with better than human pixel level segmentation performance.
- First place at Offline Chinese Character Recognition (task1: "Offline Chinese Character Recognition") at ICDAR 2011, Beijing, China (with Ueli Meier).
- First place at The German Traffic Sign Recognition Benchmark (both phases) at IJCNN 2011, San Jose, US (with Ueli Meier and Jonathan Masci). We were the only team with better than human performance.

http://people.idsia.ch/~ciresan/

How Many Computers to Identify a Cat? 16,000



http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?pagewanted=all

More Successful Applications

- Microsoft uses DL for speech rec. service (audio video indexing), based on Hinton/Toronto's DBNs (Mohamed et al 2011)
- Google uses DL in its Google Goggles service, using Ng/Stanford DL systems
- NYT talks about these: http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?_r=1
- Substantially beating SOTA in language modeling (perplexity from 140 to 102 on Broadcast News) for speech recognition (WSJ WER from 16.9% to 14.4%) (Mikolov et al 2011) and translation (+1.8 BLEU) (Schwenk 2012)
- SENNA: Unsup. pre-training + multi-task DL reaches SOTA on POS, NER, SRL, chunking, parsing, with >10x better speed & memory (Collobert et al 2011)
- Recursive nets surpass SOTA in paraphrasing (Socher et al 2011)
- Denoising AEs substantially beat SOTA in sentiment analysis (Glorot et al 2011)
- Contractive AEs SOTA in knowledge-free MNIST (.8% err) (Rifai et al NIPS 2011)
- Le Cun/NYU's stacked PSDs most accurate & fastest in pedestrian detection and DL in top 2 winning entries of German road sign recognition competition

Slide from Yoshua Bengio

Training a recurrent net to predict the next character

- Ilya Sutskever used 5 million strings of 100 characters each, taken from Wikipedia. For each string he starts predicting at the 11th character.
- It takes a month on a GPU board to get a really good model. It needs very big mini-batches.
- Ilya's best model is about equal to the state of the art for character prediction, but works in a very different way from the best other models.
  - It can balance quotes and brackets over long distances.

Slide from G. Hinton

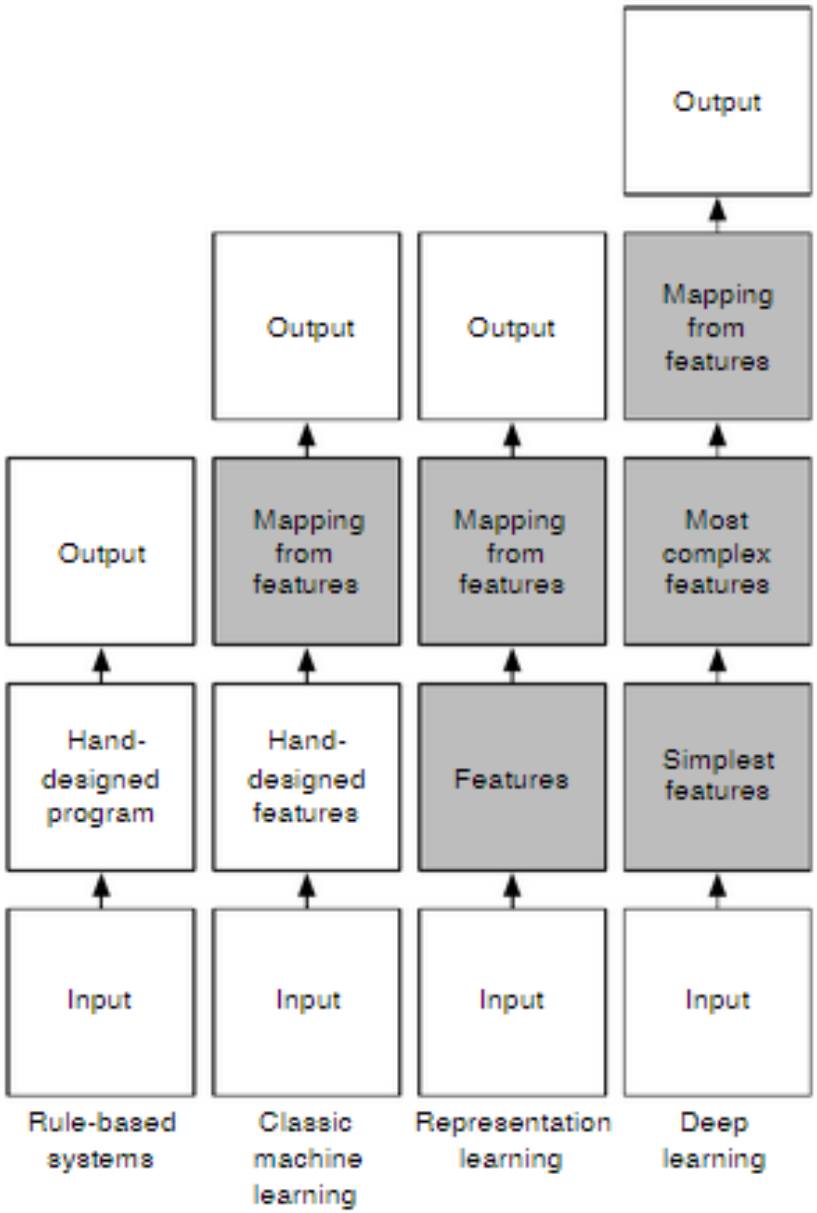
Some text generated by Ilya Sutskever's recurrent neural network one character at a time

In 1974 Northern Denver had been overshadowed by CNL, and several Irish intelligence agencies in the Mediterranean region. However, on the Victoria, Kings Hebrew stated that Charles decided to escape during an alliance.

Slide from G. Hinton

The meaning of life

- https://www.youtube.com/watch?v=vShMxxqtDDs



http://www.iro.umontreal.ca/~bengioy/dlbook/intro.html

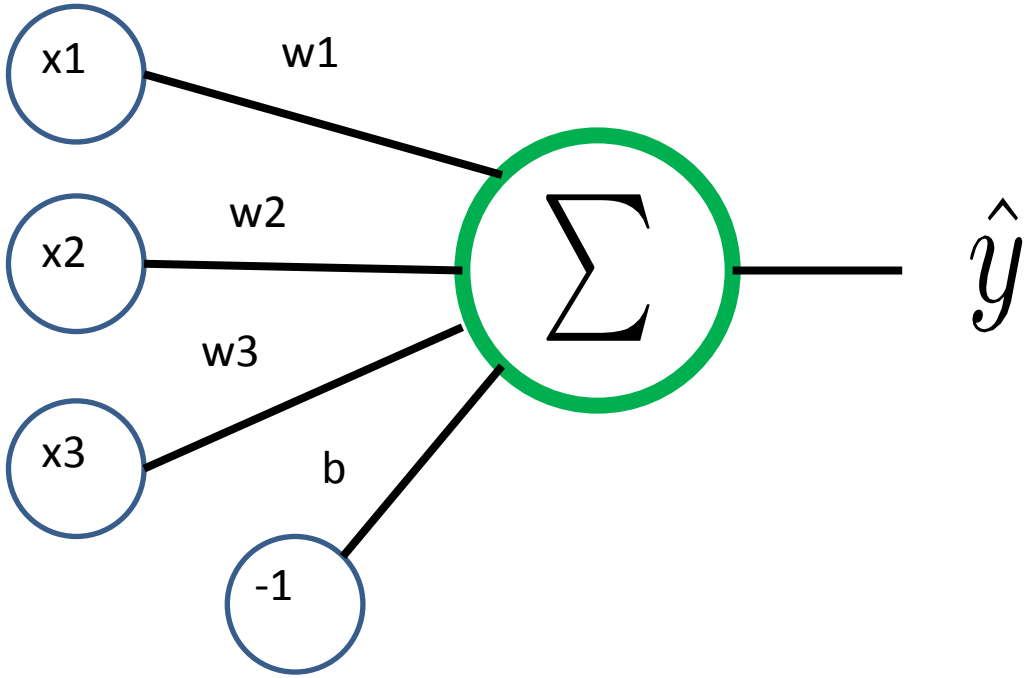
What is Deep learning?

- buzzword
- state-of-the art in machine learning
- nothing new

Deep Learning Techniques

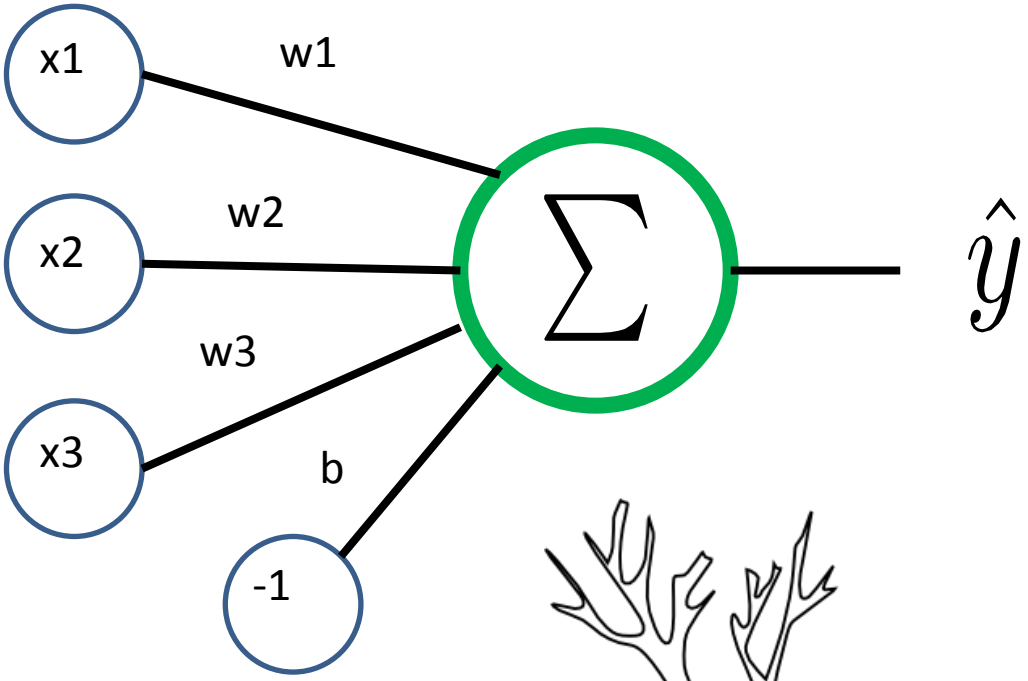
- Artifical neural network
  - Introduced in the 60s
- Convolutional neural network
  - Introduced in the 80s
- Boltzman machine
  - Introduces in the 80s

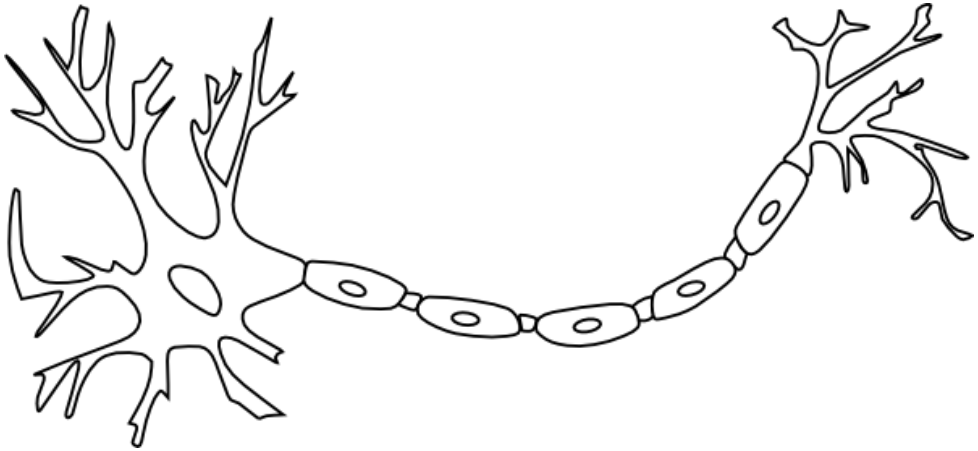
Perceptron



$$w^T x + b = 0$$

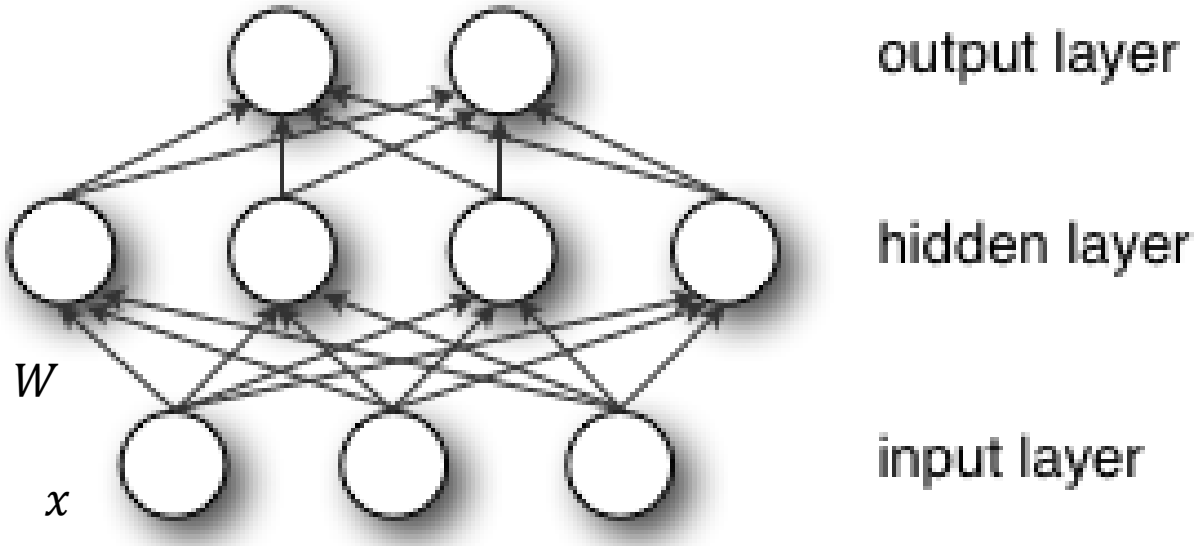
Perceptron Fun Facts





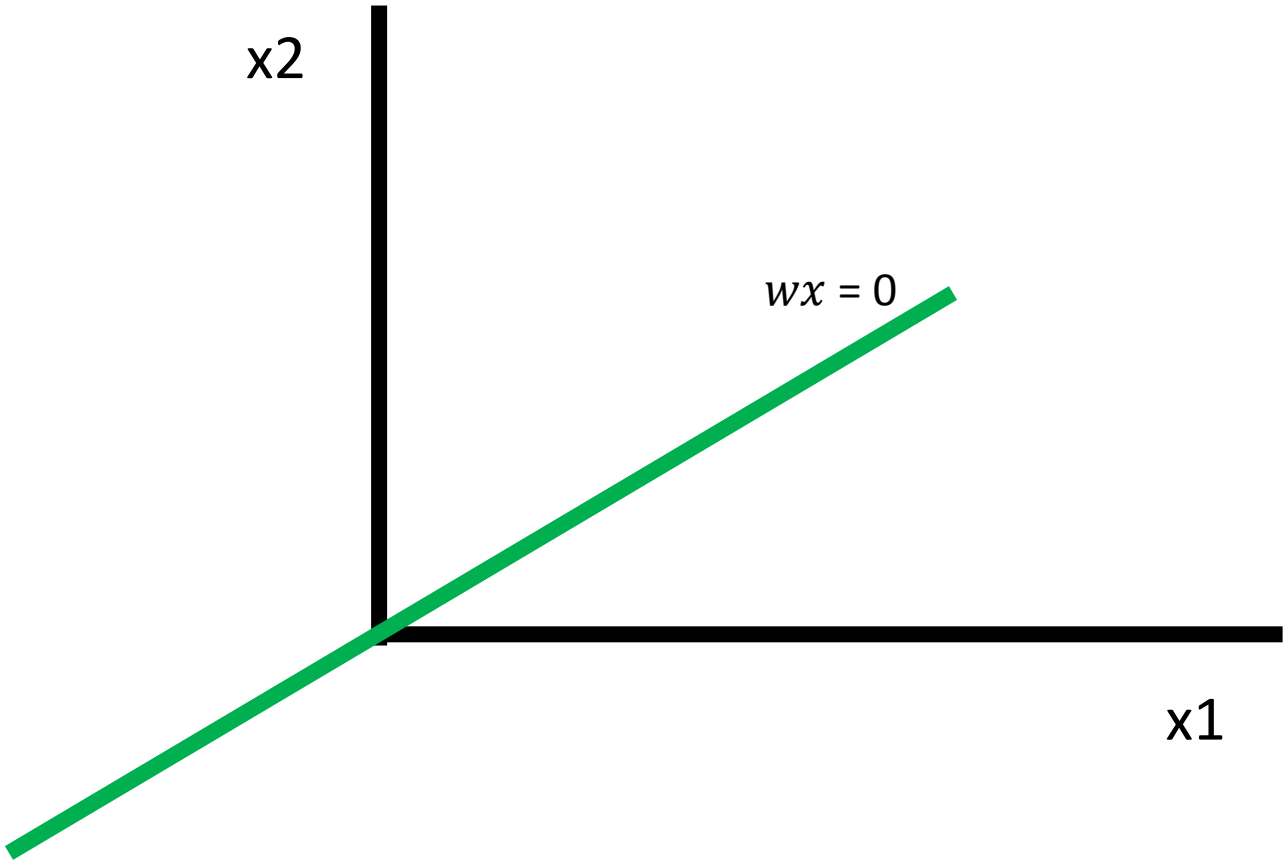
Multi-Layer Perceptron

$$s\left(b^{(1)} + W^{(1)}x\right)$$

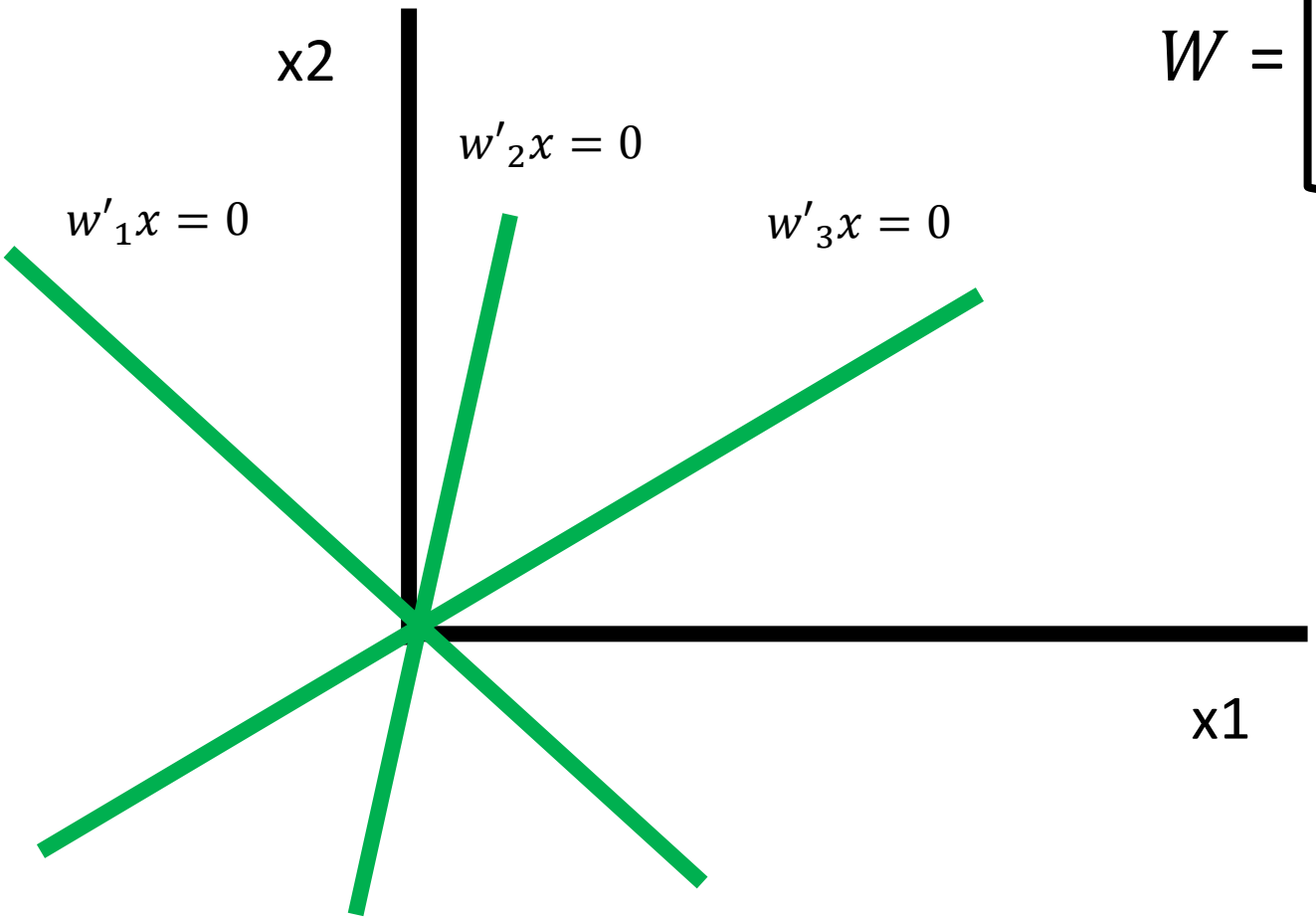


http://deeplearning.net/tutorial/mlp.html

Perceptron



Multi-Perceptron

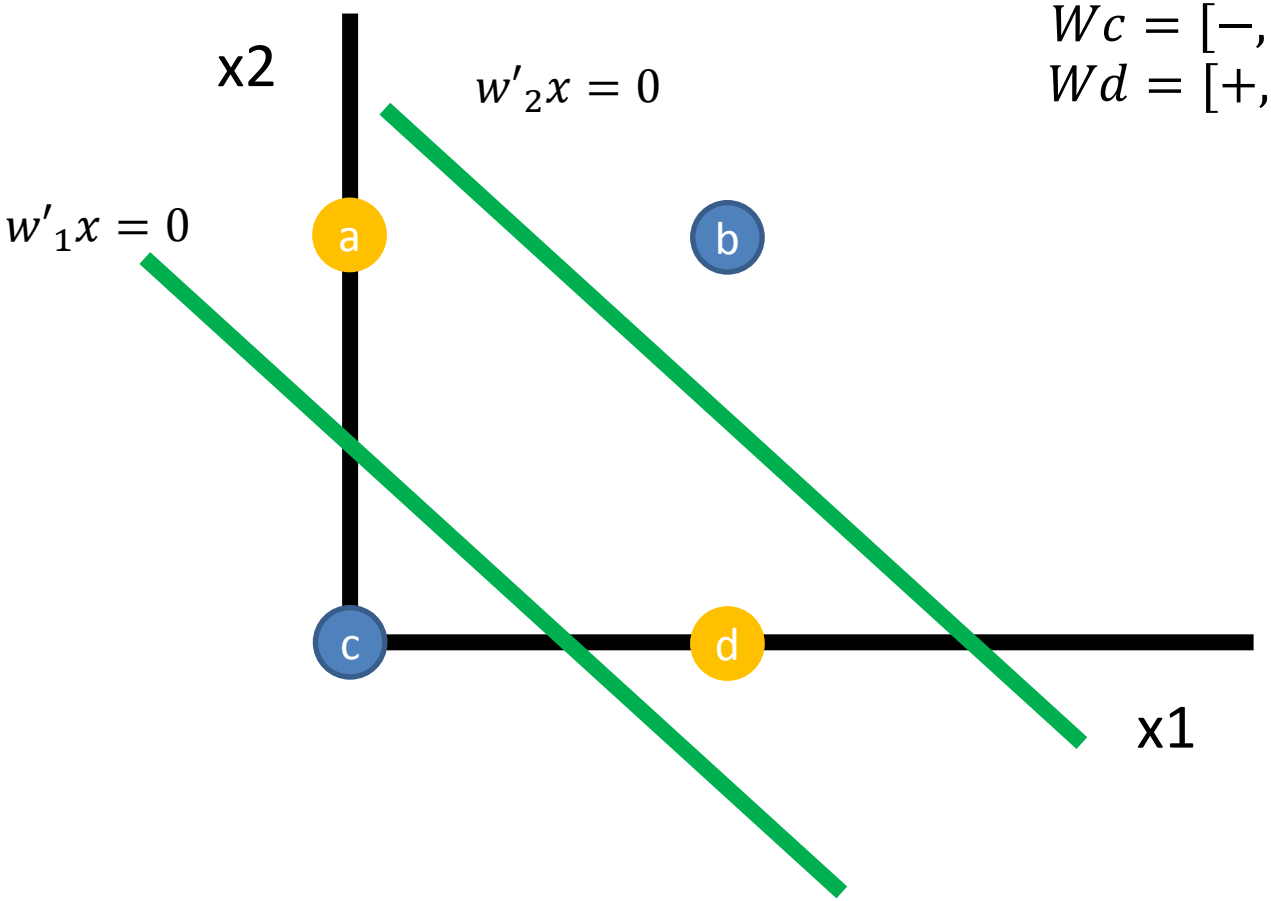


$$W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{bmatrix}$$

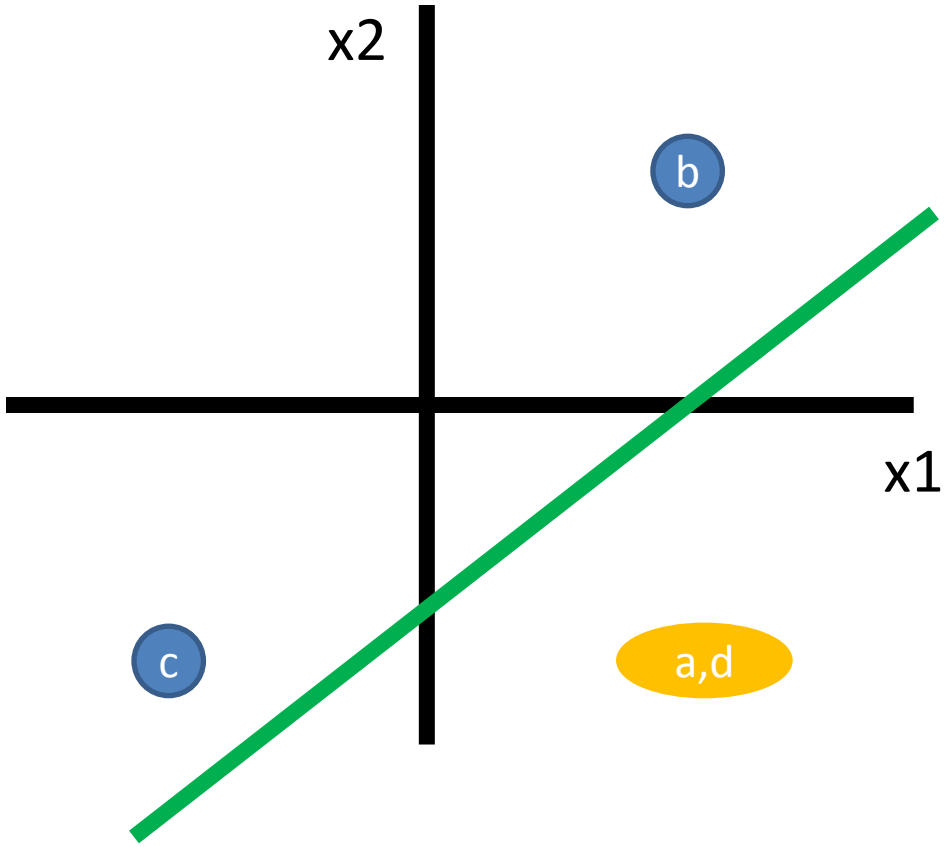
$$Wx = ?$$

Xor Problem

$Wa = [+, -]$
$Wb = [+, +]$
$Wc = [-, -]$
$Wd = [+, -]$



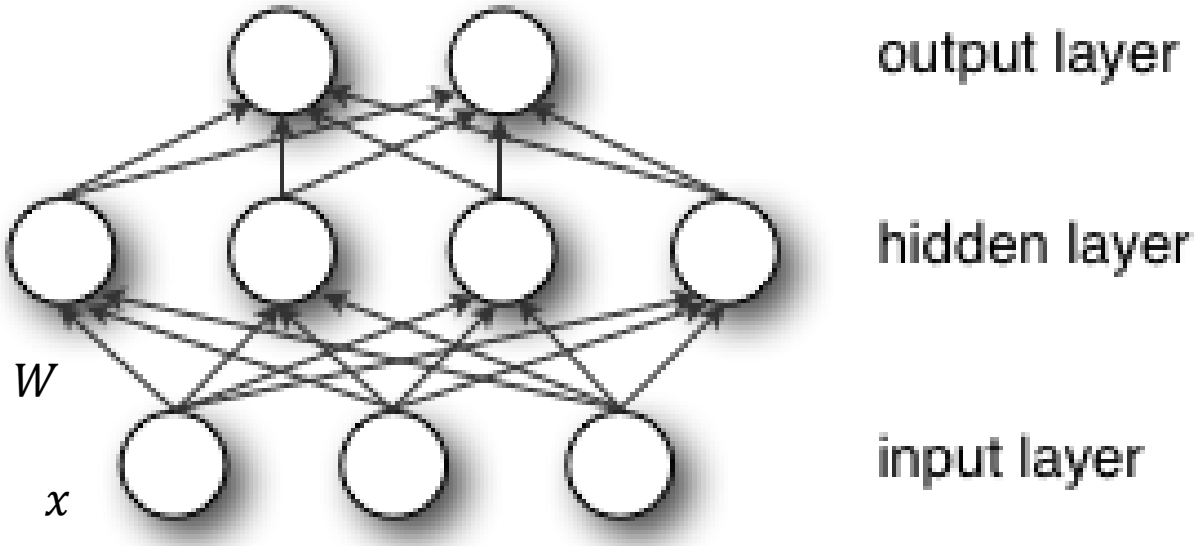
Xor Problem



$$Wa = [+, -]$$
$$Wb = [+, +]$$
$$Wc = [-, -]$$
$$Wd = [+, -]$$

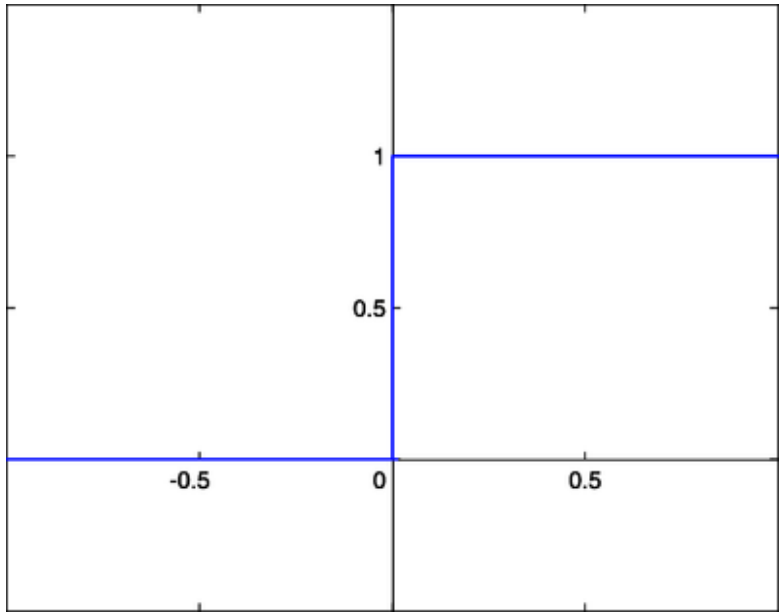
Multi-Layer Perceptron

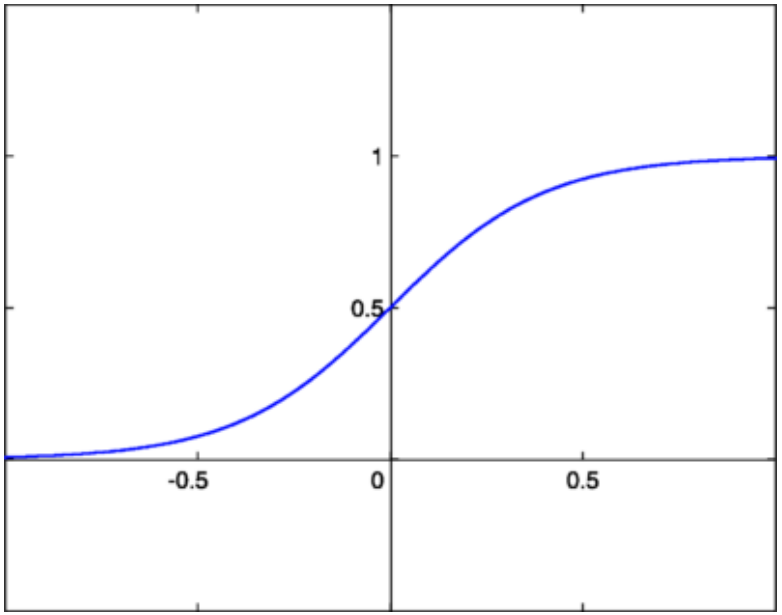
$$s\left(b^{(1)} + W^{(1)}x\right)$$



http://deeplearning.net/tutorial/mlp.html

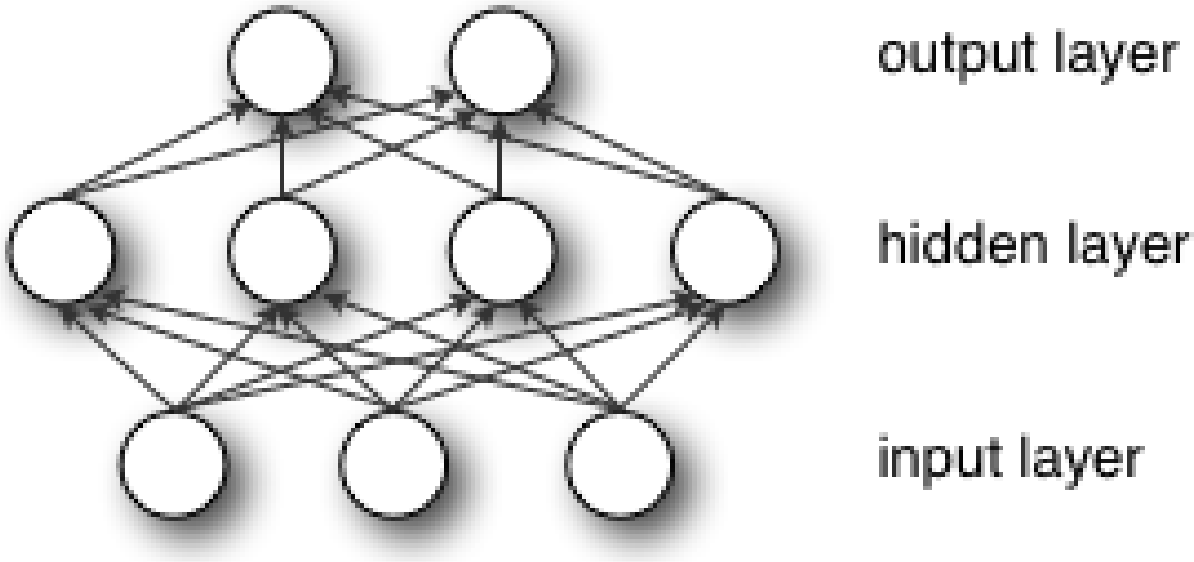
Step vs Sigmoid Activation





$$s(x) = \frac{1}{1 + e^{-cx}}$$

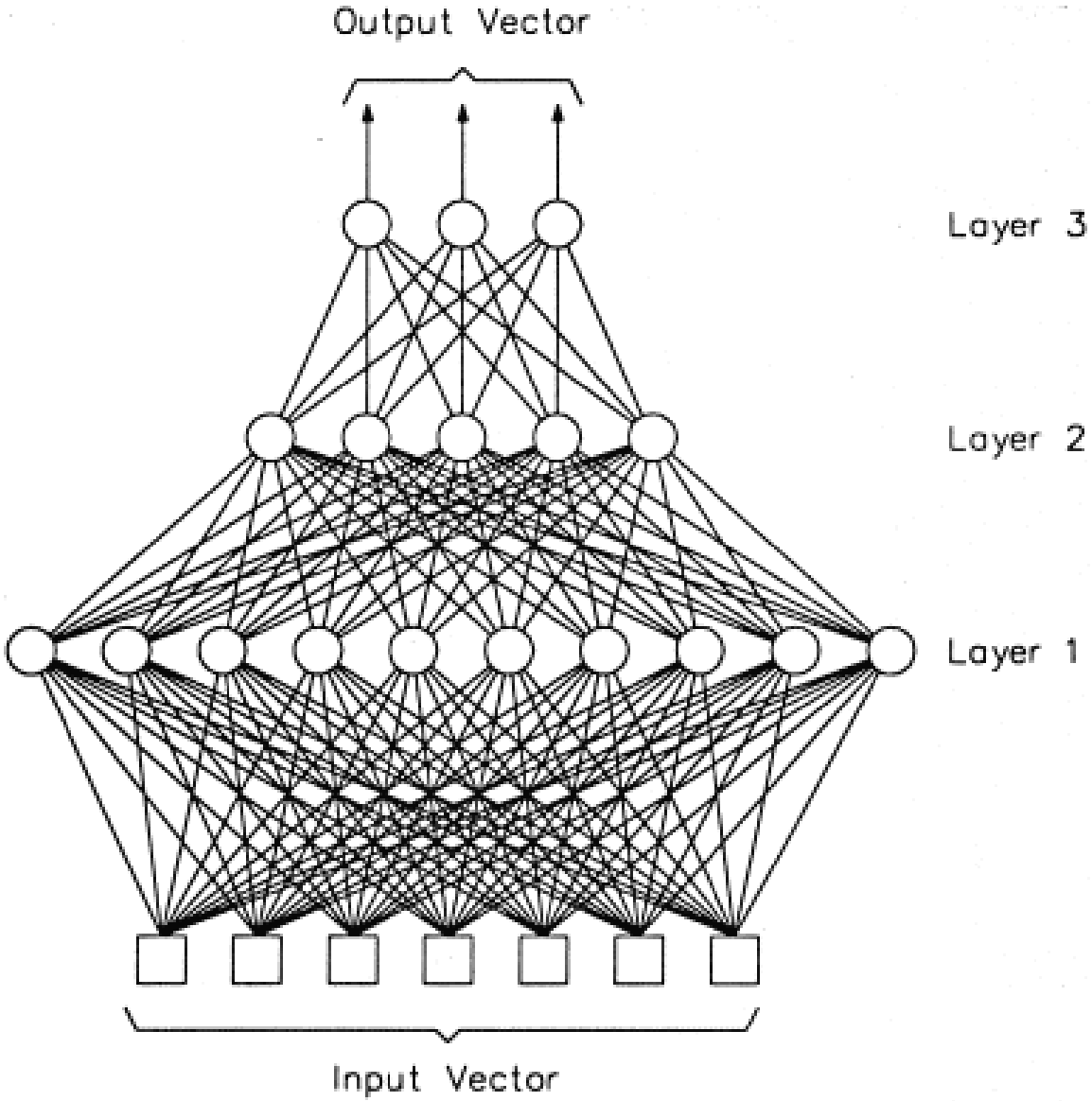
Multi-Layer Perceptron



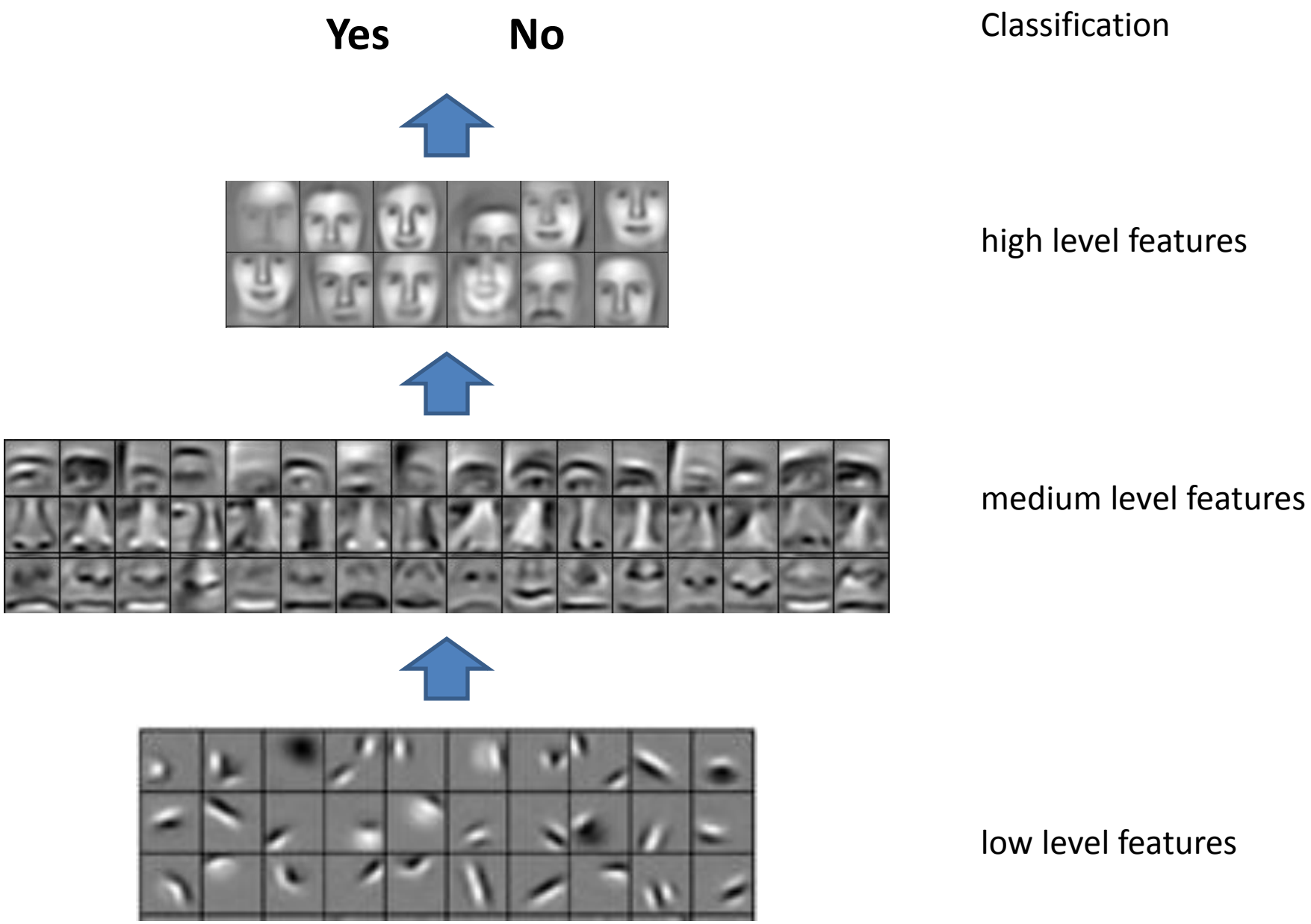
$$f(x) = G(b^{(2)} + W^{(2)}\left(s\left(b^{(1)} + W^{(1)}x\right)\right))$$

G: logistic function, softmax for multiclass

http://deeplearning.net/tutorial/mlp.html



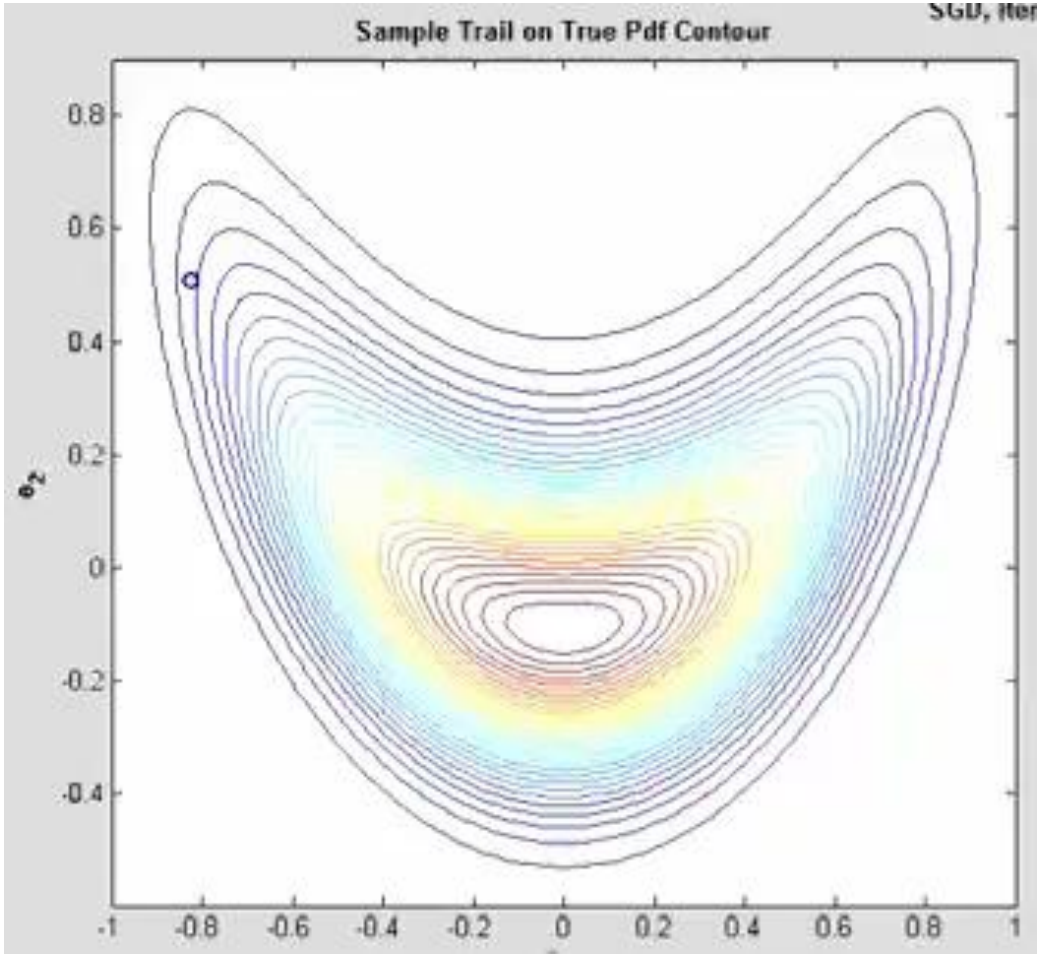
http://www.statistics4u.com/fundstat_eng/img/hl_multil_perceptron.png


Yes
No
Classification
high level features
medium level features
low level features

DNNs are hard to train

- backpropagation – gradient descent
- many local minima
- prone to overfitting
- many parameters to tune
- SLOW

Stochastic Gradient Decent



https://www.youtube.com/watch?v=HvLJUsEc6dw

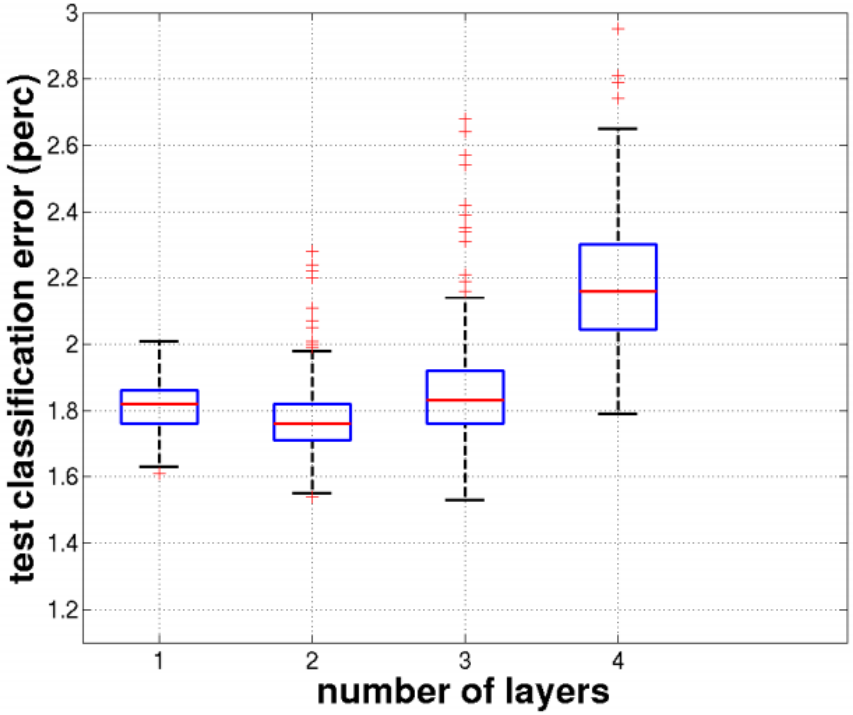
Development

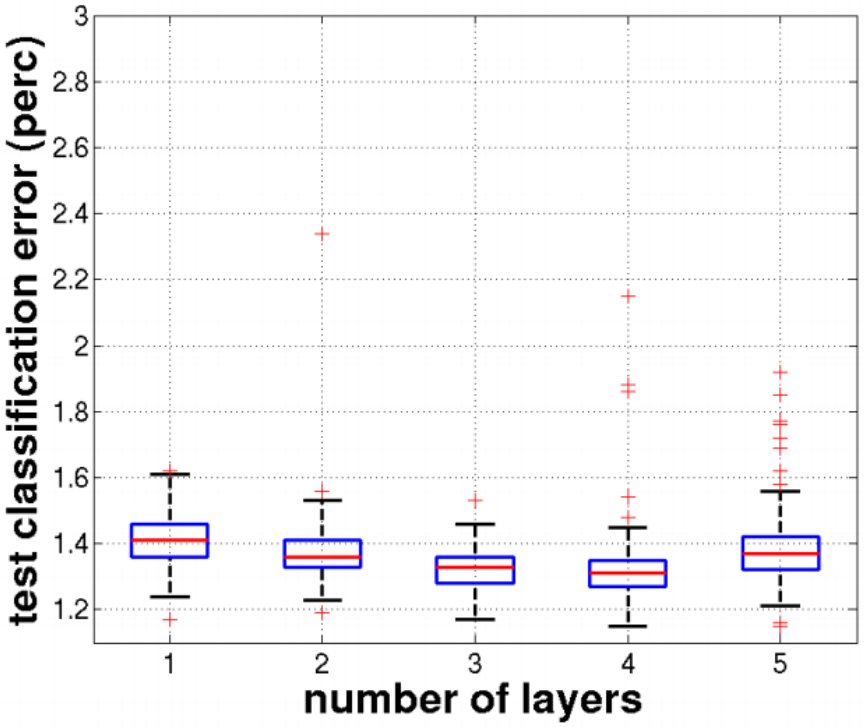
- Computers got faster!
- Data got bigger.
- Initialization got better.

2006 Breakthrough

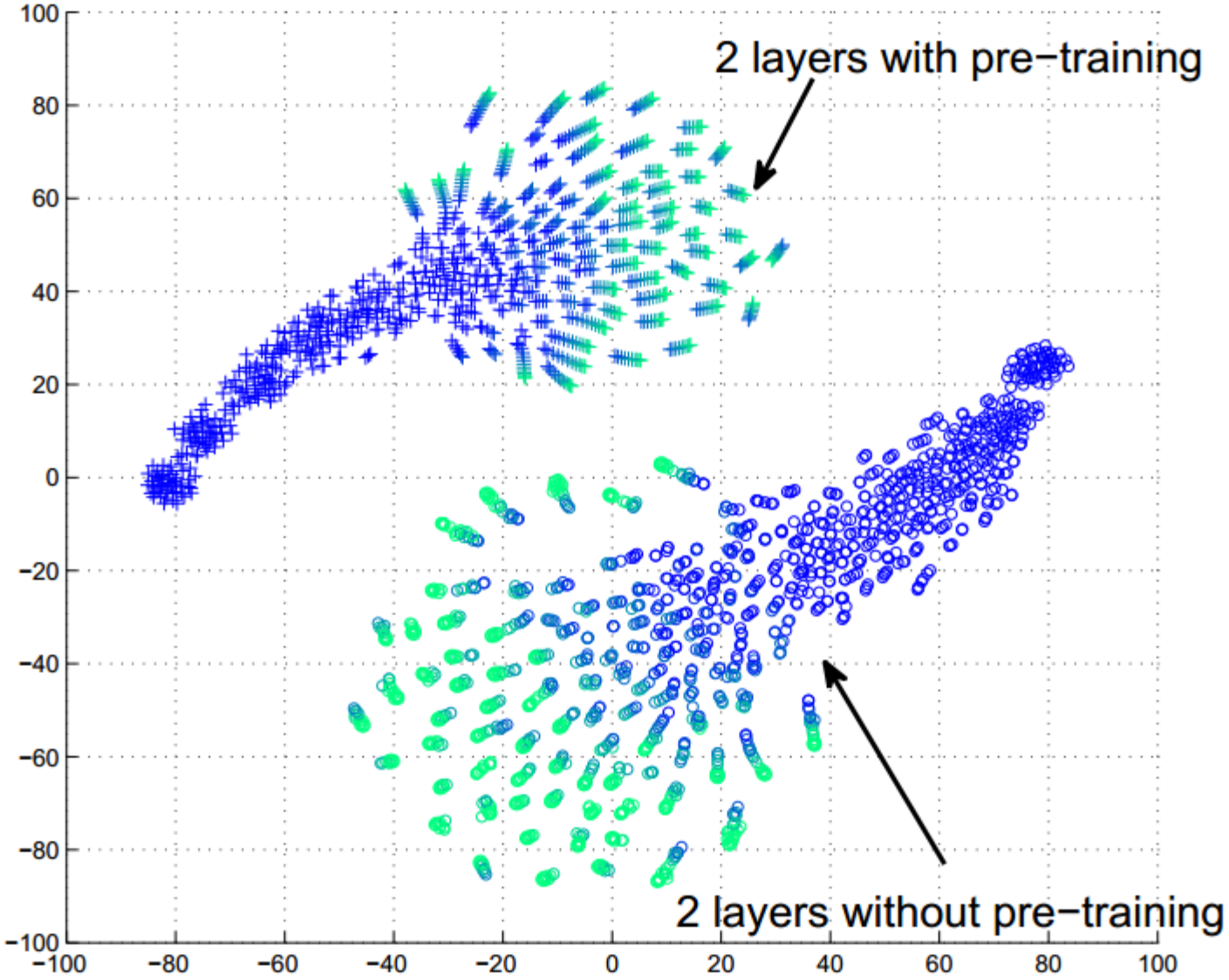
- Ability to train deep architectures by using layer-wise unsupervised learning, whereas previous purely supervised attempts had failed
- Unsupervised feature learners:
  - RBMs
  - Auto-endocer variants
  - Sparse coding variants

Unsupervised Pretraining



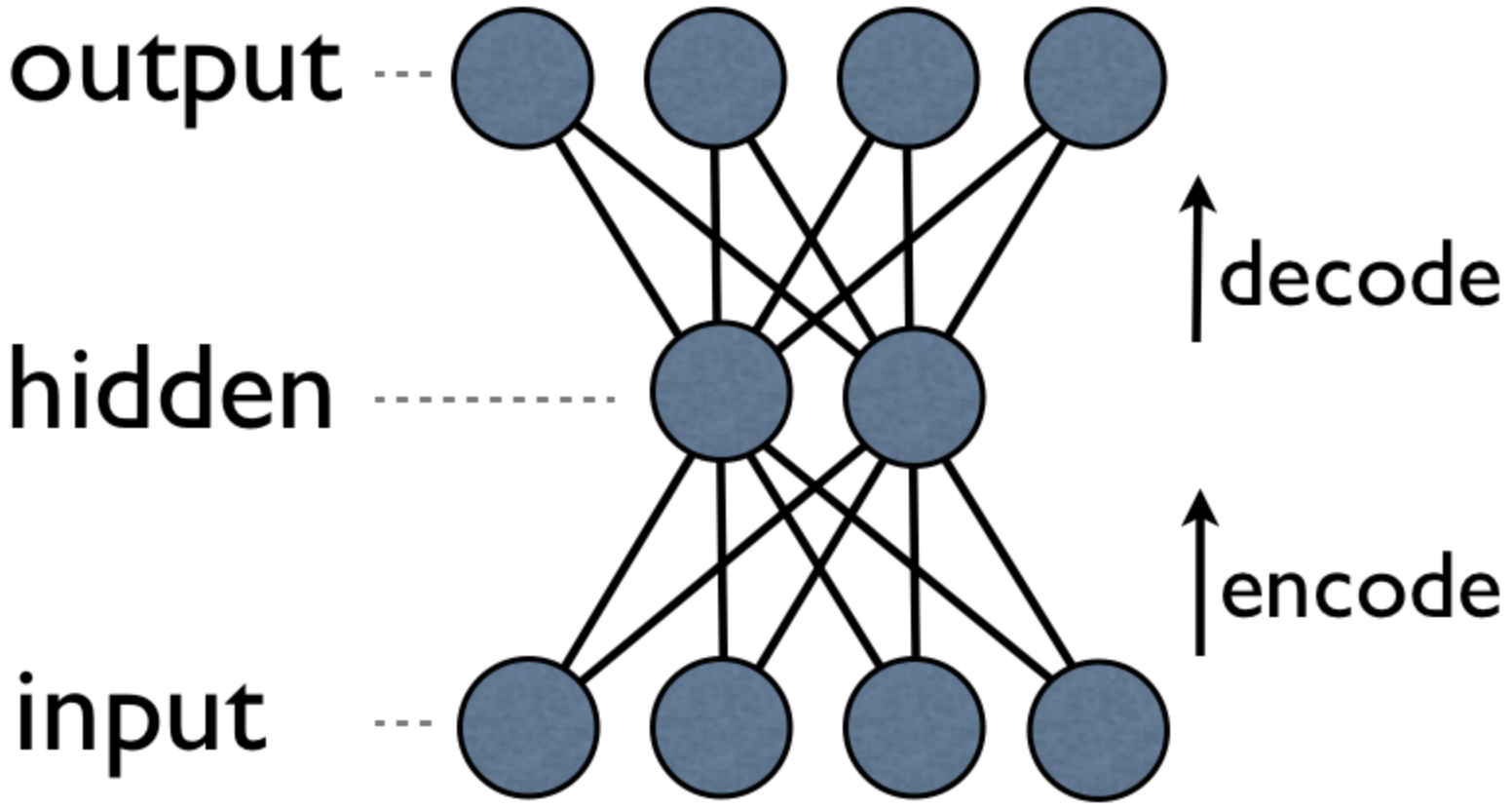


http://jmlr.org/papers/volume11/erhan10a/erhan10a.pdf



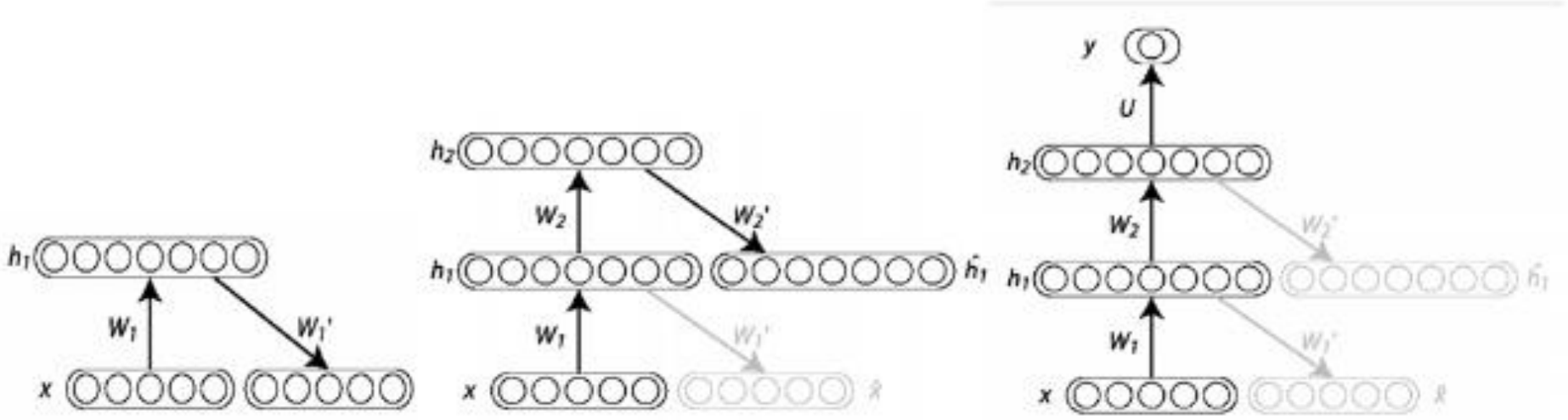
http://jmlr.org/papers/volume11/erhan10a/erhan10a.pdf

Autoencoder



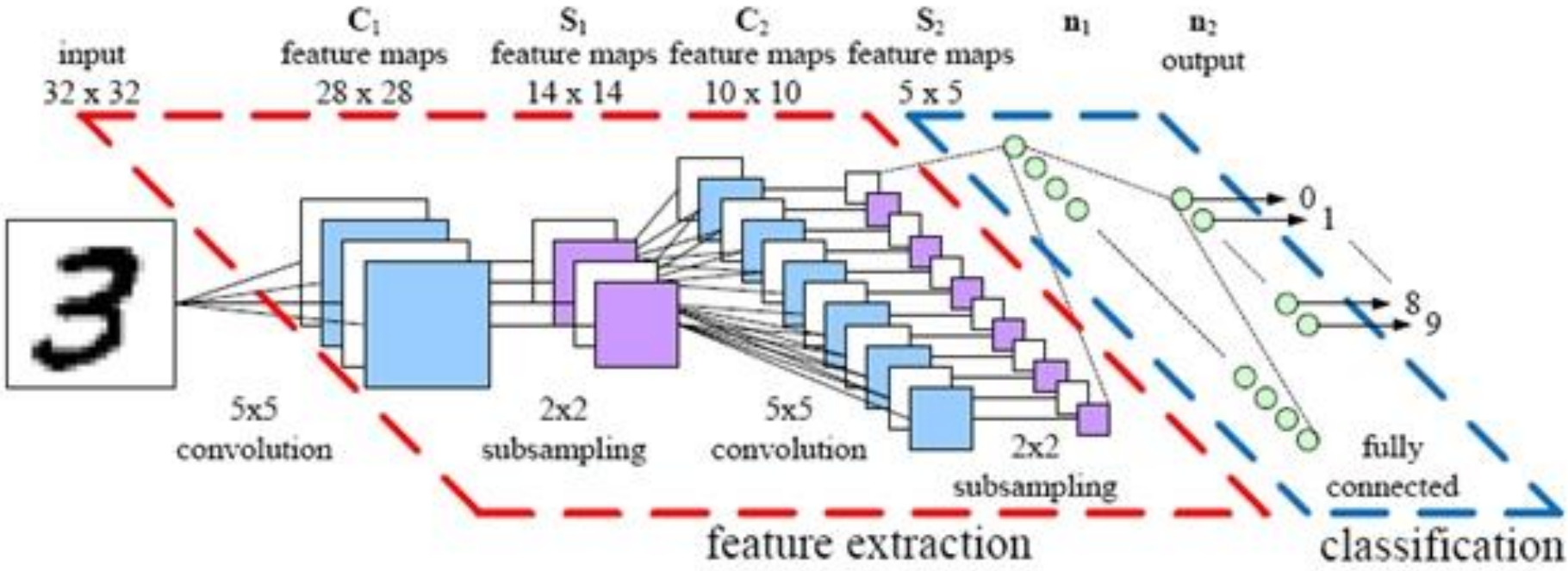
Pretraining: Stacked Denoising Auto-encoder

- Stacking Auto-Encoders

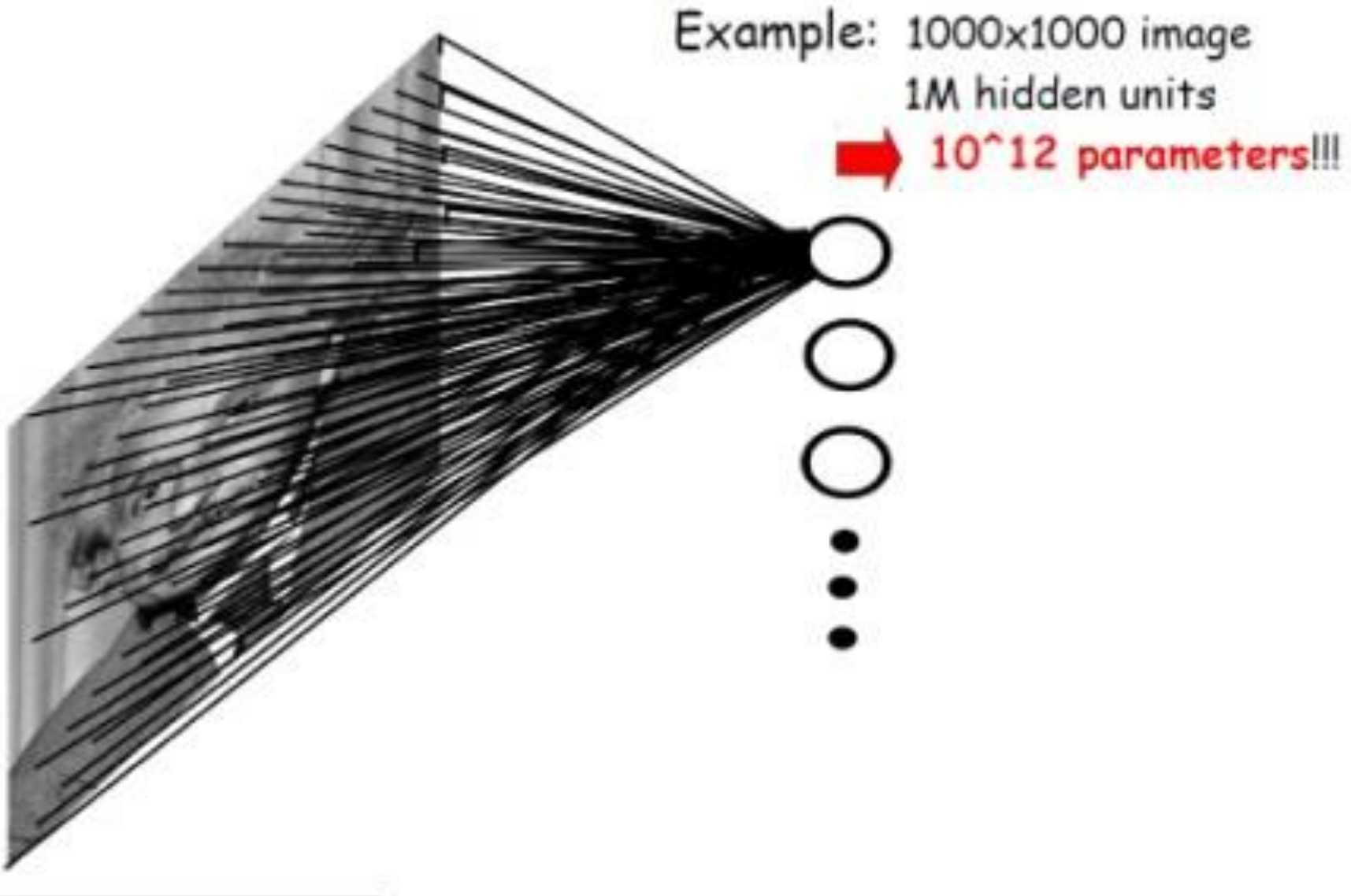


from: Bengio ICML 2009

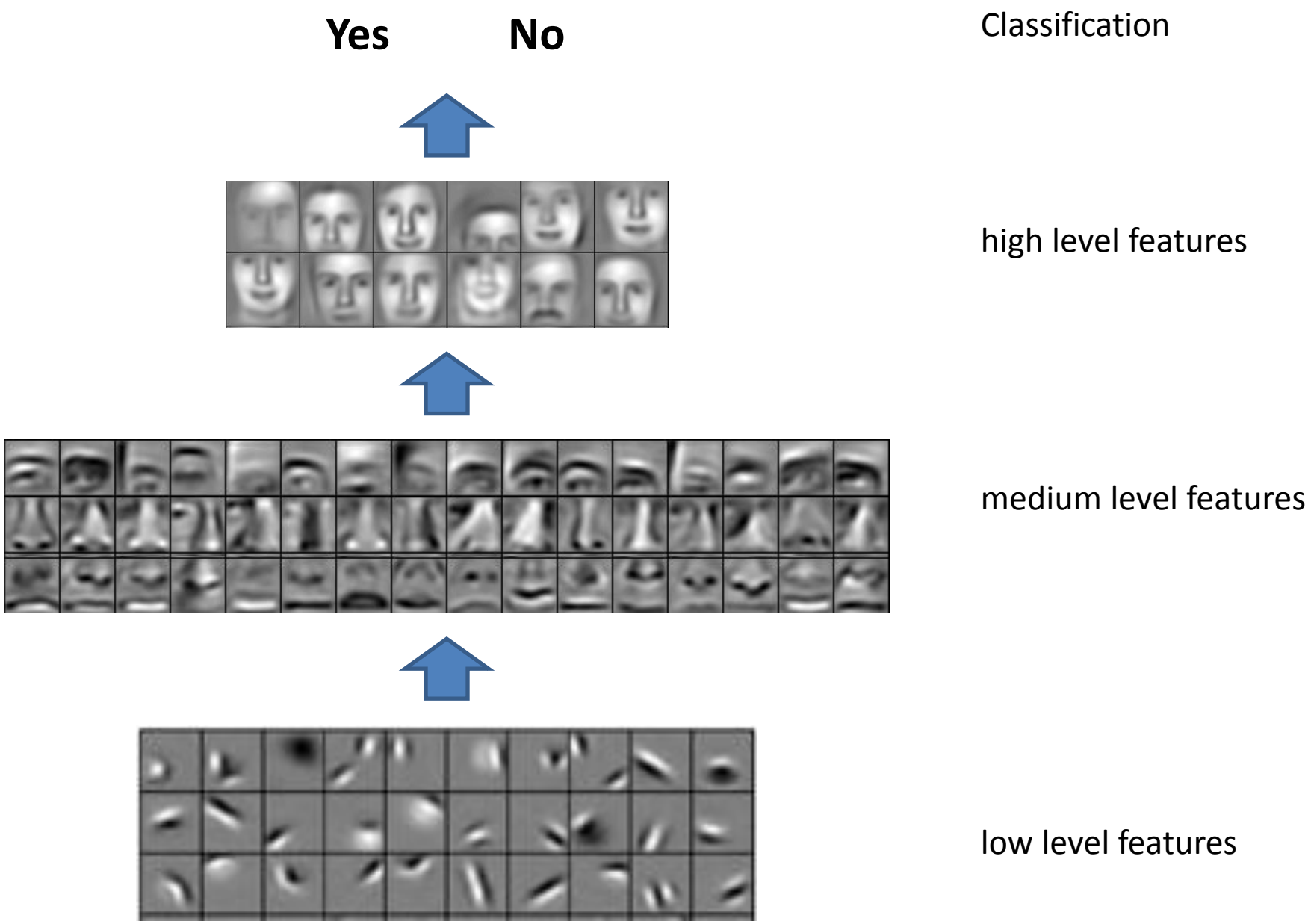
Convolutional Network



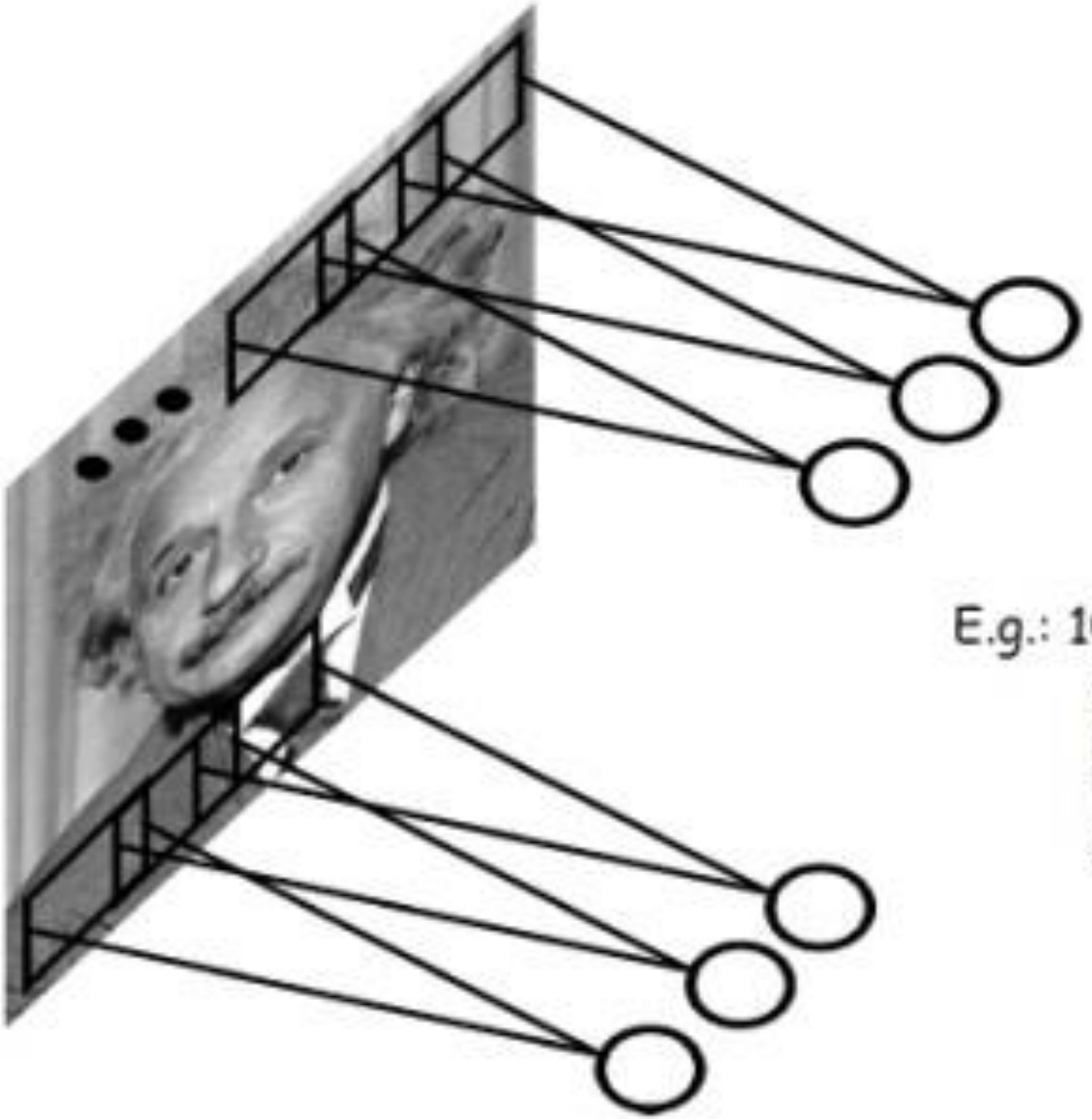
FULLY CONNECTED NEURAL NET



http://www.amolgmanurkar.com/classifyST1usingCNN.html


Yes
No
Classification
high level features
medium level features
low level features

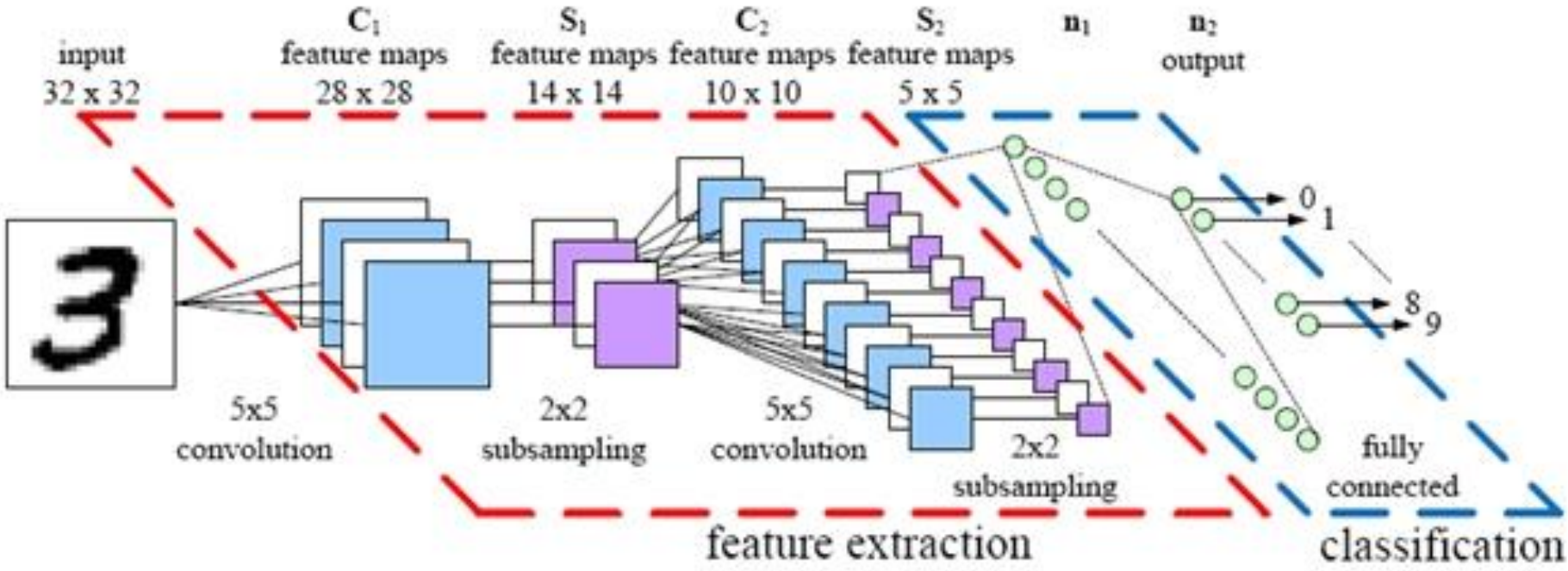
CONVOLUTIONAL NET



E.g.: 1000x1000 image
100 Filters
Filter size: 10x10
10K parameters

http://www.amolgmahurkar.com/classifySTLusingCNN.html

Convolutional Network



Nice Demo

- http://places.csail.mit.edu/demo.html

Resources

- http://www.deeplearning.net/tutorial/

- http://www.iro.umontreal.ca/~bengioy/dlbook/