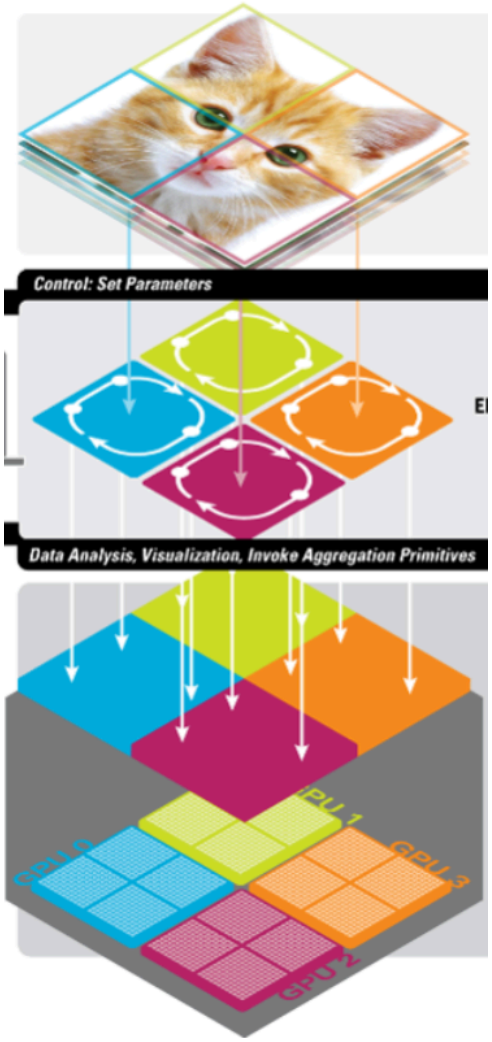
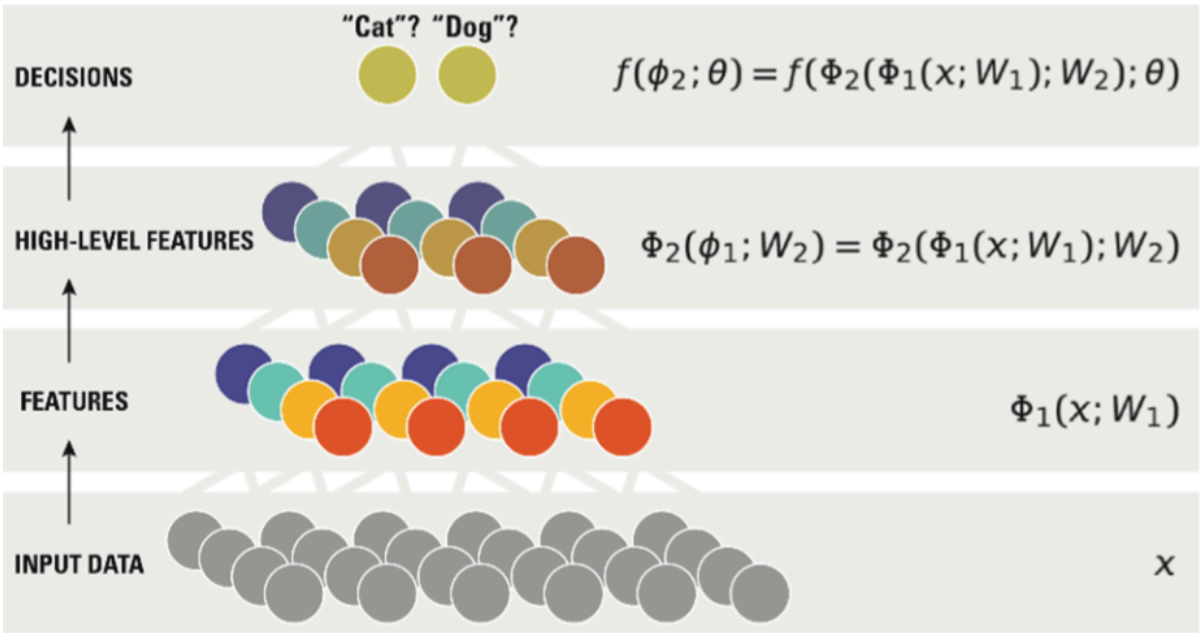
Deep Learning

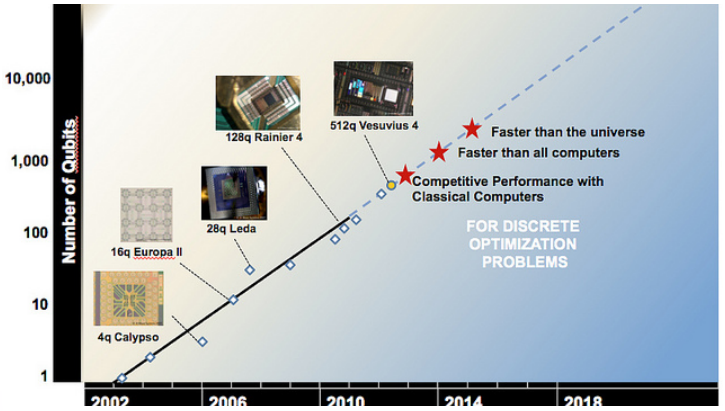
Piotr Luszczek

March 28, 2018

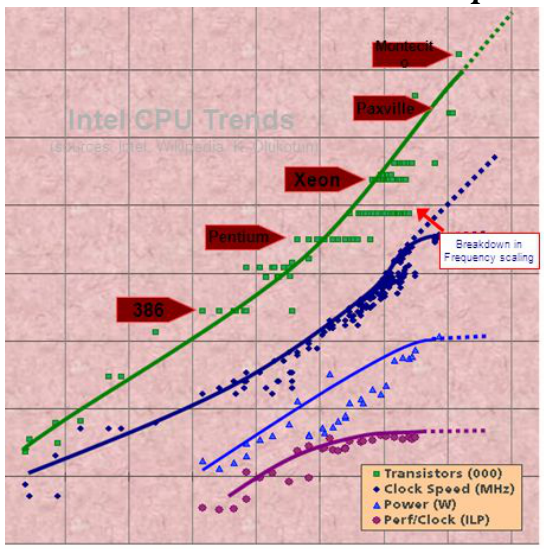




Computing Landscape: Current and Future



$\frac{kB}{T}$



quantum

Performance

∞

1

10 Hz

Frequency

∞

The Backdrop of Consciousness: Brain Statistics

- Neurons, synapses, axons, functional units, cortices
  - Communications of ACM, 54(8), 2011
  - Human neurons: 10^{10}, synapses: 10^{15} @10 Hz
  - Humans use more 10% of the brain (sorry Luc Besson's "Lucy")
- Part of brain devoted for visual processing
- Retina size and resolution
  - 10k by 10k
  - 200ppi

	Mouse	Rat	Cat	Monkey	Human
Neurons	$16 \cdot 10^6$	$55 \cdot 10^6$	$763 \cdot 10^6$	$2 \cdot 10^9$	$20 \cdot 10^9$
Synapses	$128 \cdot 10^9$	$442 \cdot 10^9$	$6.1 \cdot 10^{12}$	$16 \cdot 10^{12}$	$200 \cdot 10^{12}$

- Visual processing speed vs. textual processing speed of the brain
  - 60,000
  - Should a text DNN be so much slower than image DNN?
  - Other cortices: auditory, cerebral, olfactory, prefrontal, visual

Based on 1 Data Point...

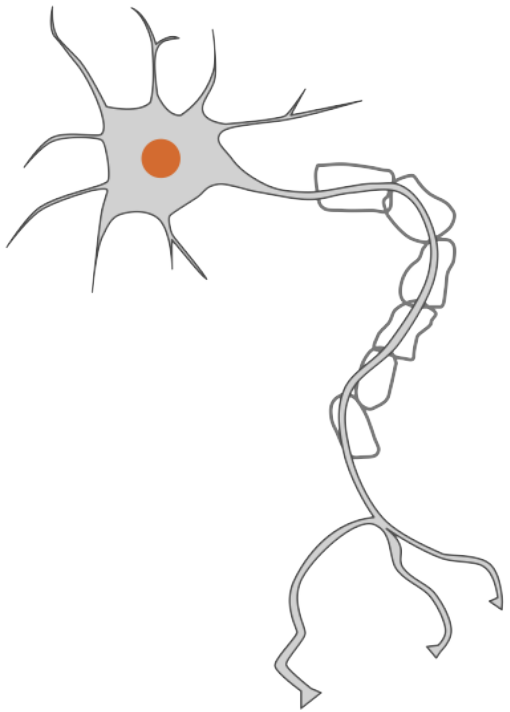
- Humans are the best examples that we have of a system that can solve a variety of complex problems at the bleeding edge of AI and Machine Learning
  - Learning languages
  - Learning causal relationships
  - Learning from small amount of data
  - Doing science: observe, structure, generalize, model, test
- Does brain do Bayesian inference or we model the brain as making Bayesian inference?

Brain as an Everyday Predictor

- Baseball
  - Guinness book of records in speed
  - 60 ms delay to process oncoming ball
  - No time to analyze as the ball flies, must predict
  - From batter perspective a fast ball "disappears"
- Hockey
  - "Go where the puck will be, not where it is"
- Illusions
  - Visual
    - Brain is skewed towards motion
    - Brain is skewed towards parallax
  - Magical
    - Focus and attention
- Reaction time: milliseconds

Brain Structure

- Low level
  - Axons (ports)
  - Neurons (everything: cell body, nucleus, connectivity)
  - Dendrites (detectors)
  - Synapses (connections) and Myelin
- High level
  - Cortices
  - Hippocampus
  - Eyes
  - Stems
- Information media
  - Electricity
  - Chemistry (sodium and potassium ions)
  - Mechanical



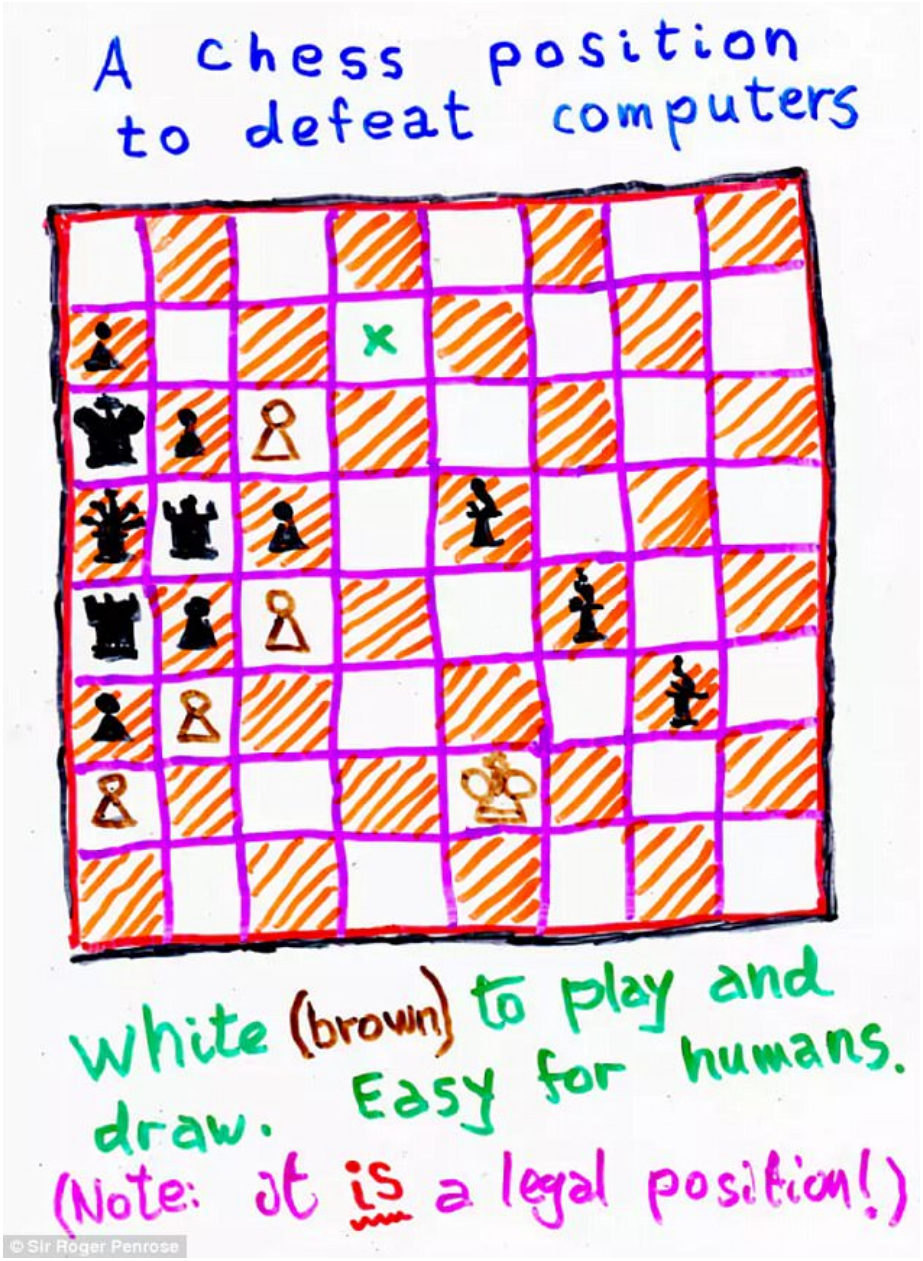
Neuroscience versus Deep Learning

- Deep Learning is a hierarchical processing model which was inspired by neuroscience
  - Jeff Hawkins
- Major differences
  - Not a biological model
  - Two types of synapses: positive and negative
  - Only precise synaptic waves
  - Lack of dendritic processing (no active dendrites)
  - Brain is trained differently

Machine Learning and AI Applications

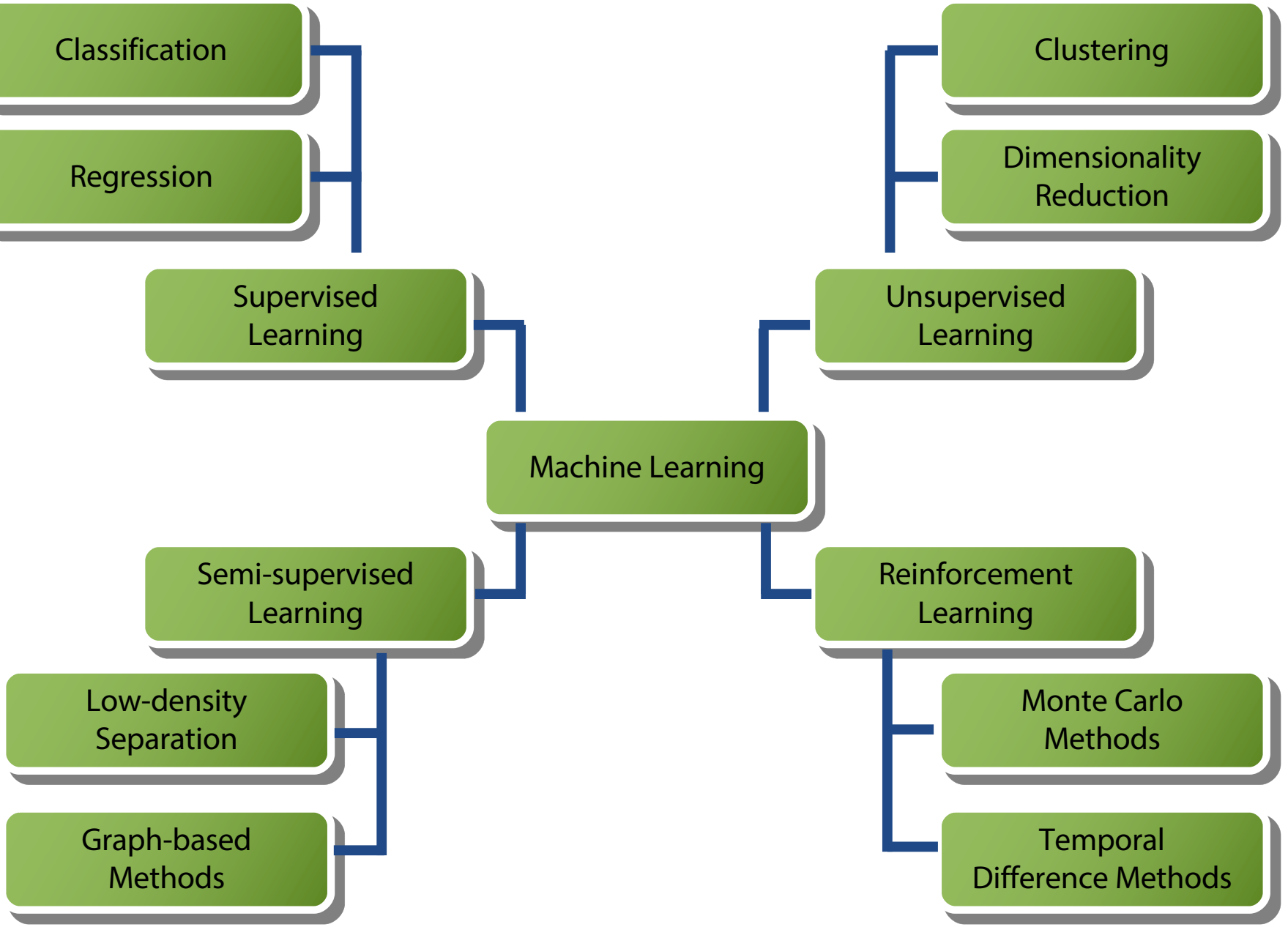
- Object recognition and identification
  - Tagging
  - Outlining
  - Focusing
  - Describing
- Language recognition and translation
  - Sentiment analysis
- Embodied cognition

AI and Humans: Chess Problem by Roger Penrose



Decision tree versus intuition

An ML Taxonomy



Recent Deep Neural Network Advances

- Image classification: 1-5% error
  - Started at above 40% error
- Deep Mind (now Google and Alphabet)
  - Atari video games
  - Playing GO board game (Alpha GO)
- Chess learning and playing at master level
  - Deep Learning Machine Teaches Itself Chess in 72 Hours, Plays at International Master Level
- Language translation and transcription
  - Recurrent Neural Networks
- Learning about learning
  - Generative Adversarial Neural Networks
    - Alpha 0
    - Style transfer

(Hasty) Generalizations: Methods / Performance

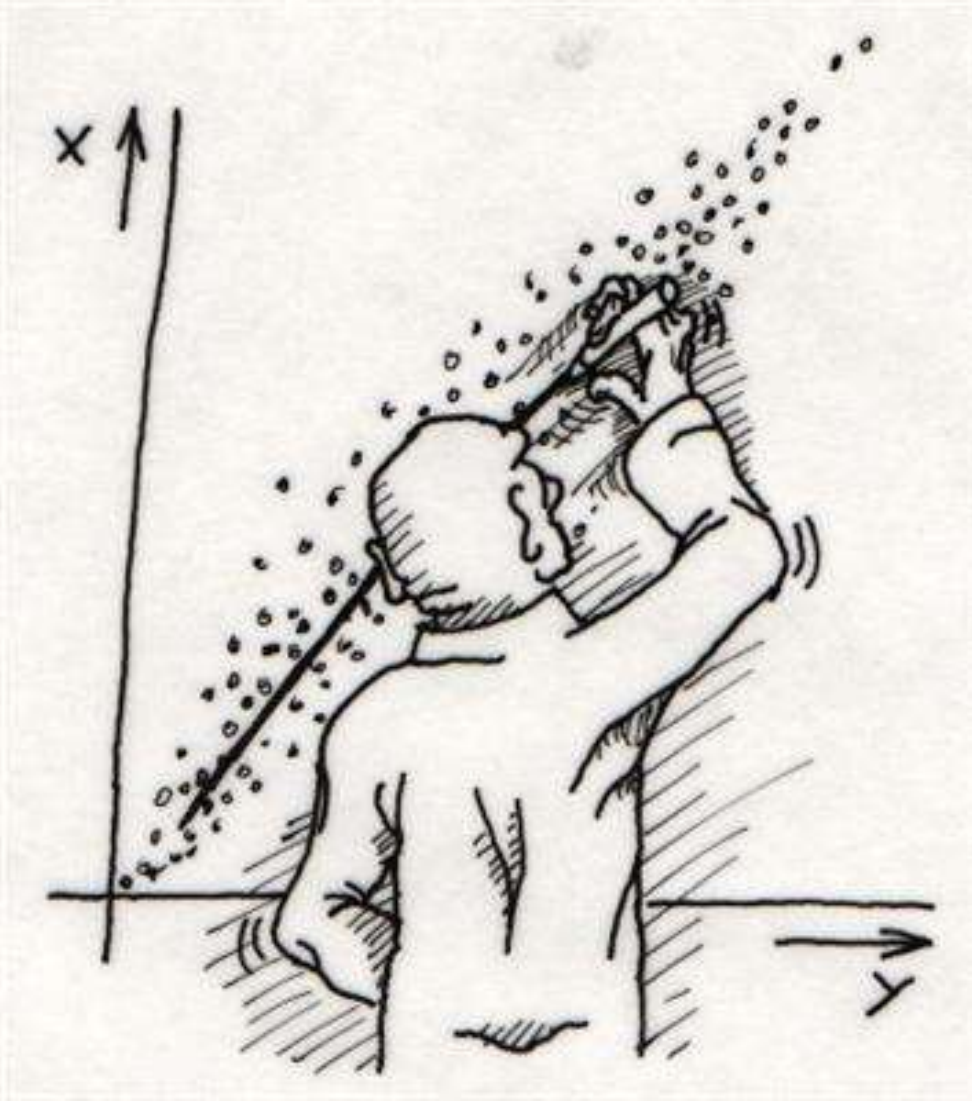
- Dense matrix methods
  - Based on matrix-matrix multiply
- Subspace projection and iterative methods
  - Based on matrix-vector multiply
- Deep learning (and other ML methods)
  - Based on inner- and dot-products:
    - Logistic regression
    - Perceptron
    - SVM
    - “Kernel trick”
    - …

Rosenblatt's Mark I Perceptron (1957)

Source: Arvin Calspan Advanced Technology Center; Hecht-Nielsen, R. Neurocomputing (Reading, Mass.: Addison-Wesley, 1990).

http://rutherfordjournal.org/article040101.html

Remember Linear Regression?



Making a curve fit.

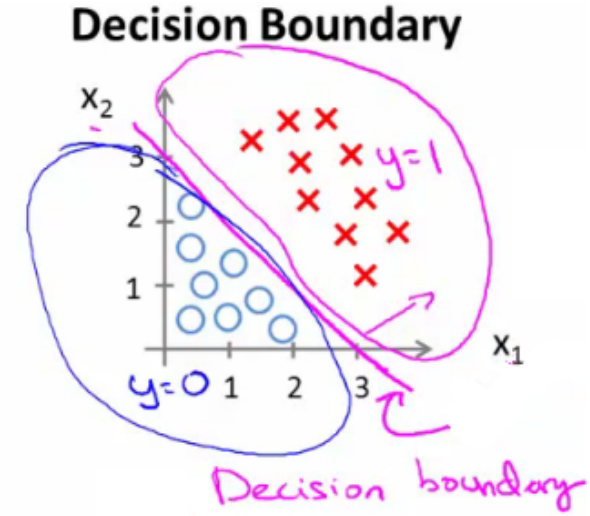
© 1998 G. Meixner

$$A = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$B = \frac{\sum y_i - A\left(\sum x_i\right)}{n}$$

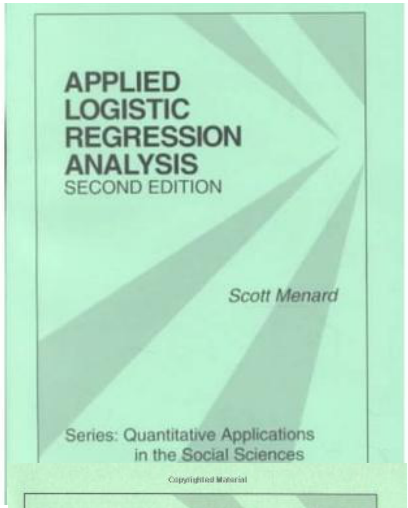
http://www.vias.org/science_cartoons/regression.html

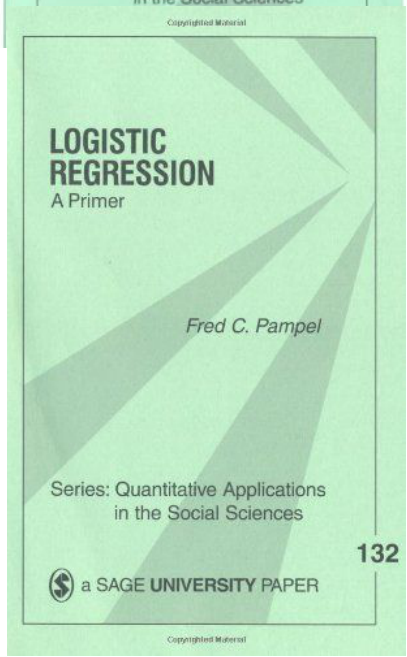
How About Logistic Regression?



$$log\frac{p(x)}{1-p(x)} = Ax + B$$

$$p(x) = \frac{1}{1+e^{-(Ax+B)}}$$



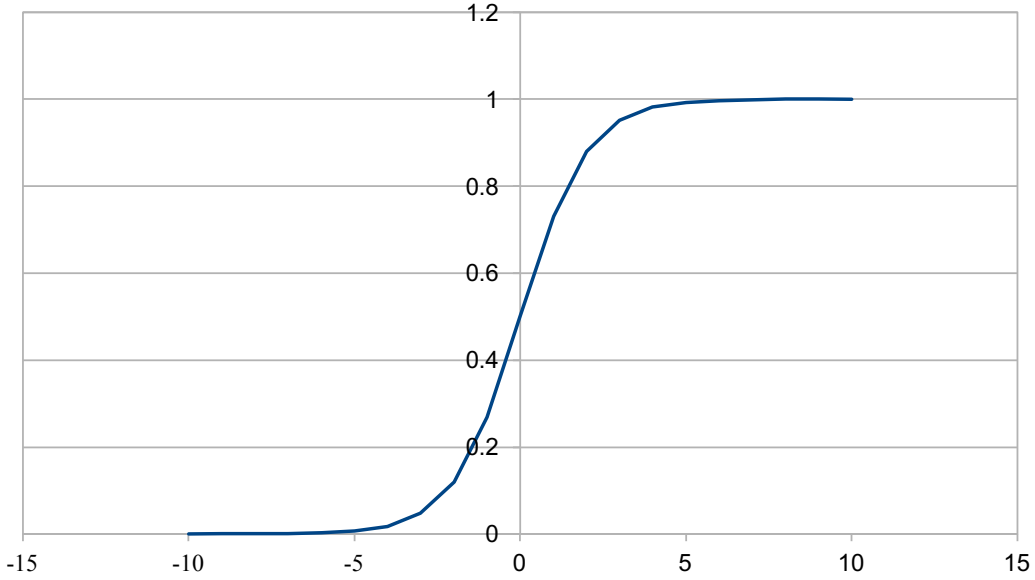


Issues:

- Linear classifier
- Linear separability and convergence
- Negative example:

XOR

O	X
X	O



Logistical Regression as a Binary Classifier

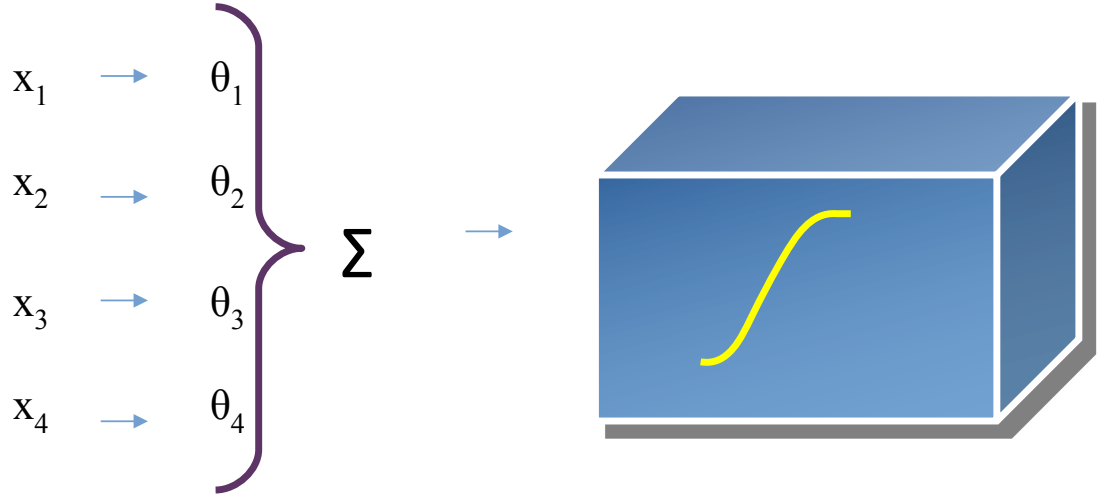
$$f(\vec{x},\vec{\theta}) \equiv \sigma(\theta^T x) = \frac{1}{1+exp(-\theta^T x)}$$

Probability that f(x,θ) is 1

$$L(\vec{\theta}) = -\sum_{i}^{m}\{y_i \equiv 1\}log[f(x_i,\vec{\theta})] + \{y_i \equiv 0\}log[1-f(x_i,\vec{\theta})]$$

Find θ that minimizes objective function L(θ):

- Predicts positives
- Does not predict negatives



Finding Optimal Weights

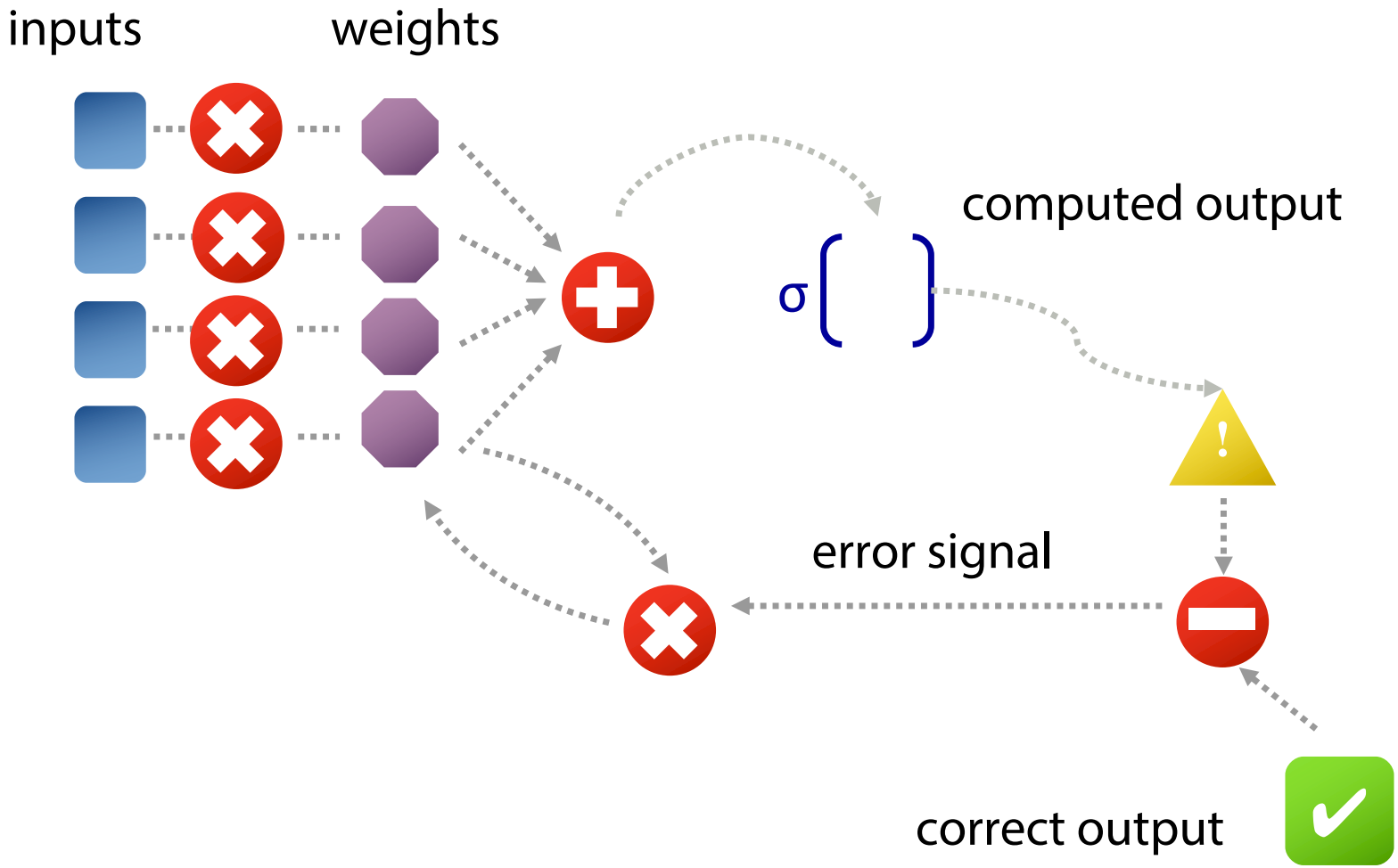
Find θ that minimizes objective function L(θ)

$$\nabla_\theta L(\vec{\theta}) = -\sum_i^m x_i \cdot \left(y_i - f(x_i, \theta)\right)$$

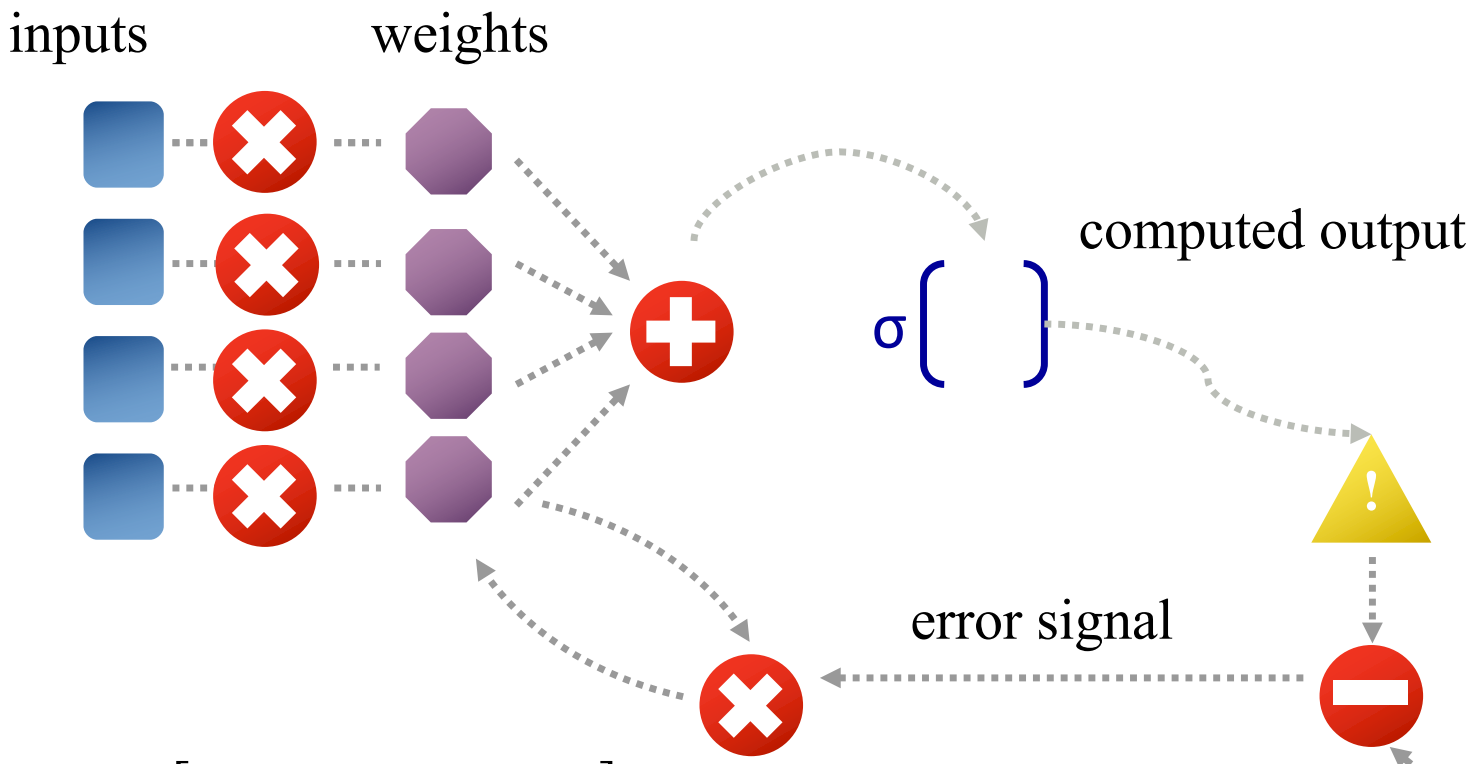
$$\vec{\theta}^{(k+1)} = \vec{\theta}^{(k)} - \eta \nabla_\theta L(\vec{\theta})$$

- Stochastic Gradient Descent is a generic method
  - Used in Adaline, Perceptron, K-Means, SVM, Lasso
- Considers empirical risk vs. expected risk
- Related to Randomized Kaczmarz

Backpropagation: Single Layer



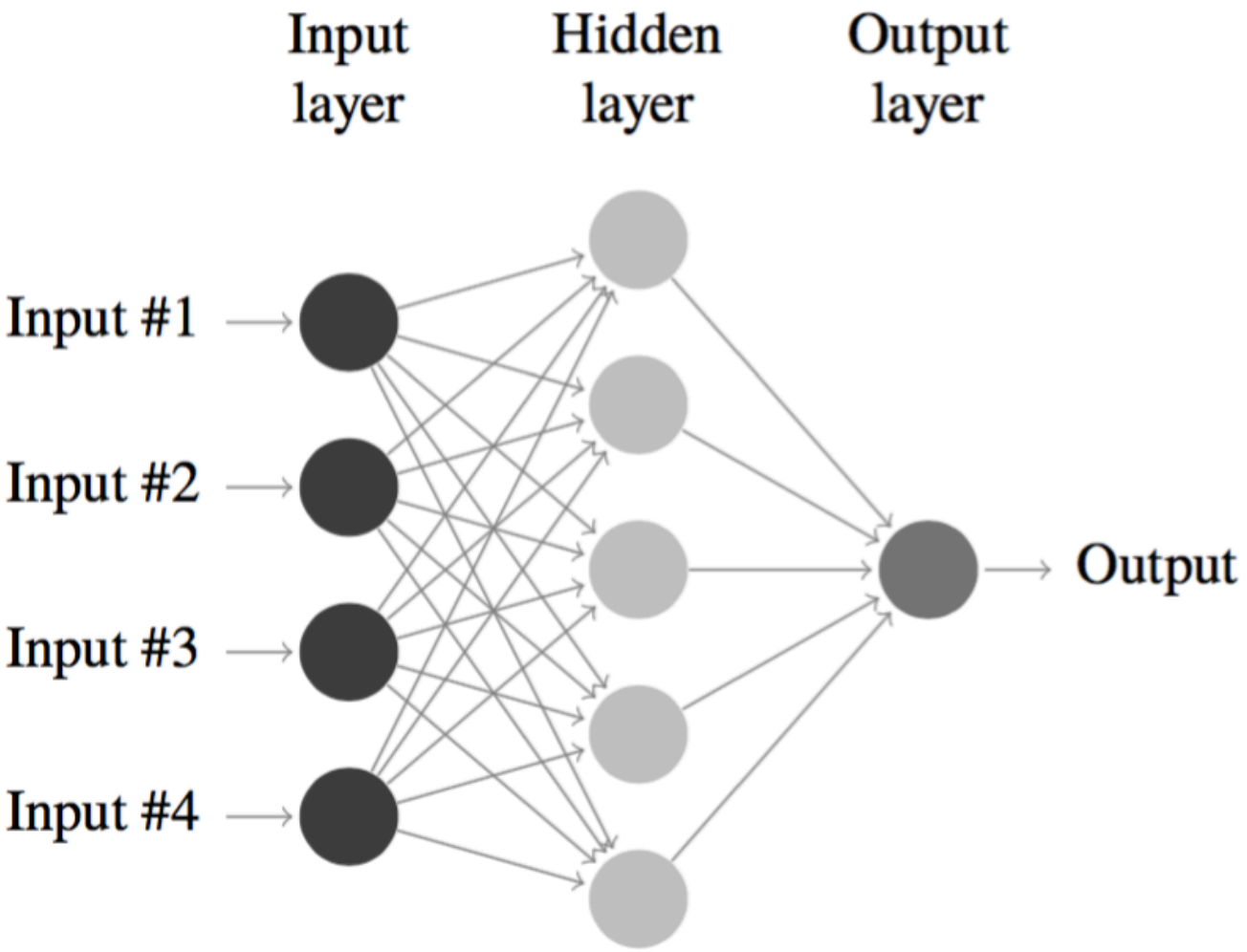
Backpropagation: Hidden Layer



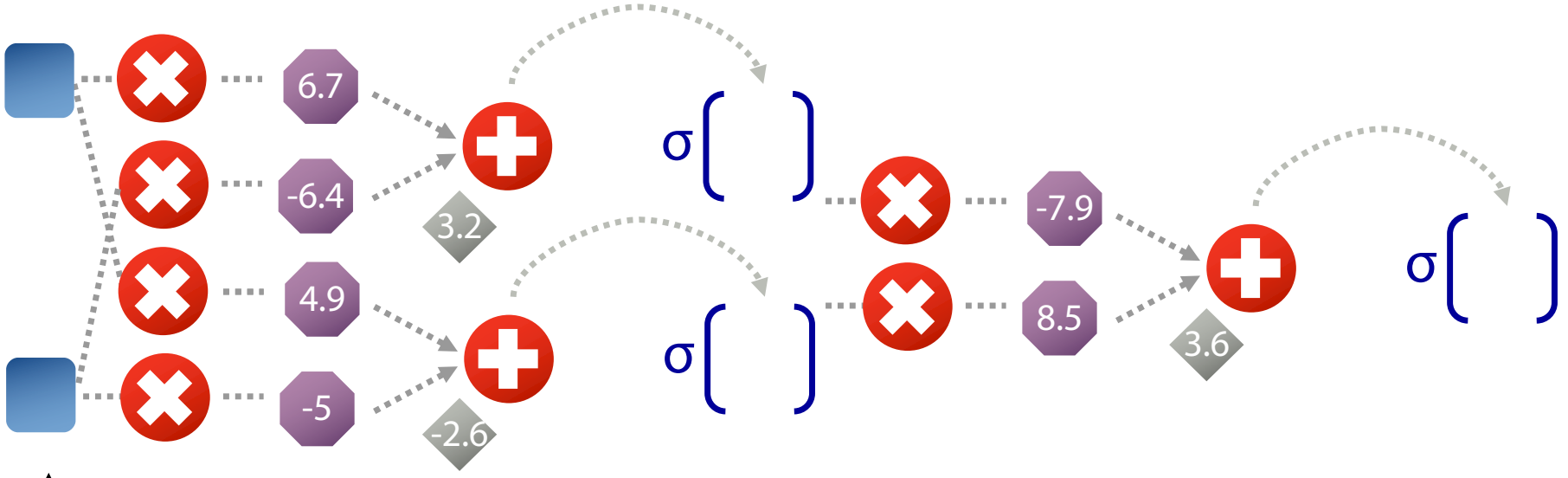
$$w^{(k+1)} = w^{(k)} - \eta \left[\sum_{i,j}^{m} w_{ij} \left(y^{(k+1)} - y^{(k)} \right) \right] \frac{df(z)}{dz}$$

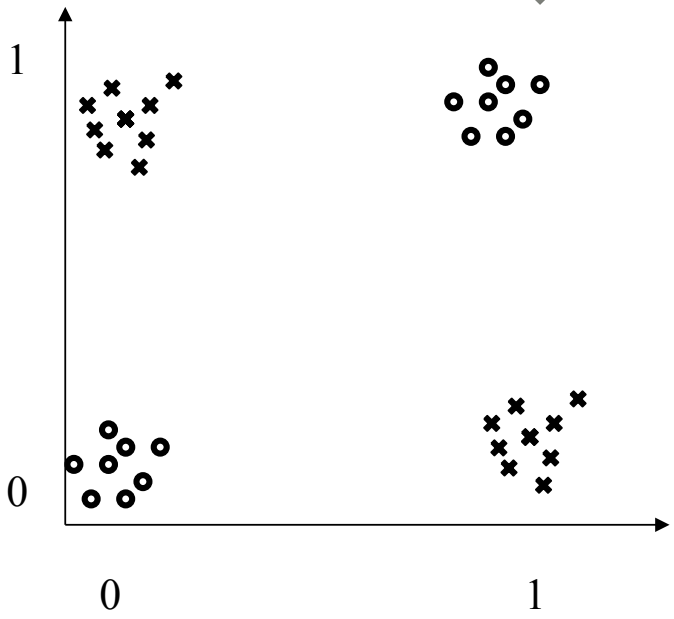
http://home.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.htm

Stacking Layers



Example: XOR





Input 1	Input 2	Output
0	0	0.03
0	1	0.96
1	0	0.97
1	1	0.03

Connection to SVM

- Suitable for classification tasks
- Separation with hyperplanes
- Minimizes loss function
  - Loss function may have similar structure
- Based on dot-product

From Pixels to Labels

177	153	118	91	85	100	124	145
151	124	93	77	86	115	148	168
115	93	78	83	108	145	177	191
88	79	84	104	136	168	190	197
82	85	103	127	152	170	180	182
91	101	120	138	150	157	159	159
103	114	127	136	140	140	140	141
111	119	126	130	130	129	128	130

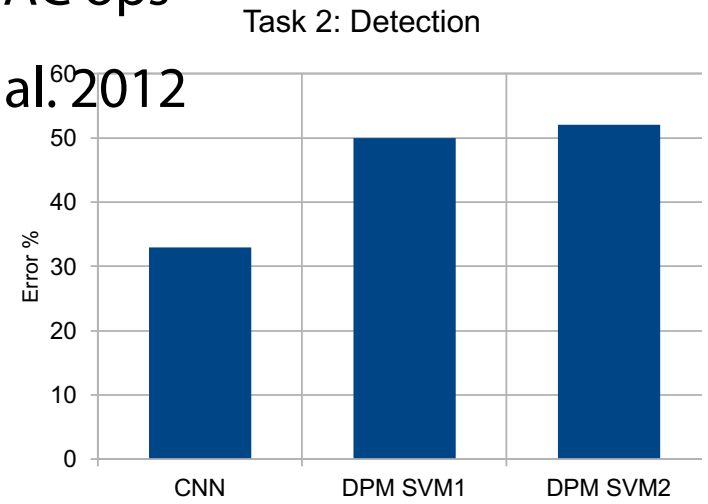
177
151
115
88
82
91
103
111
153
124
93
79
85
101
114
...

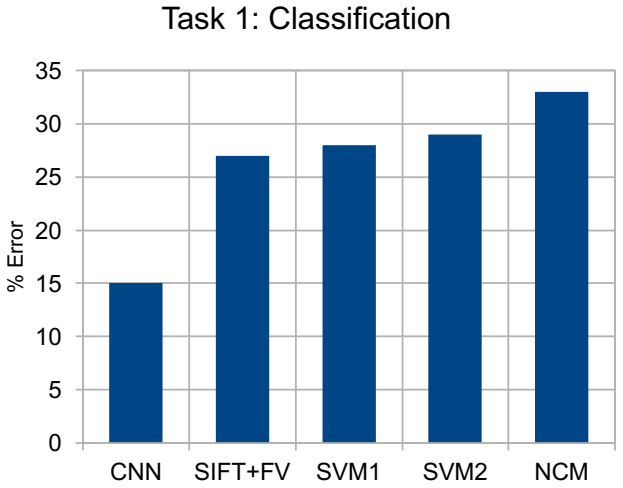
dog

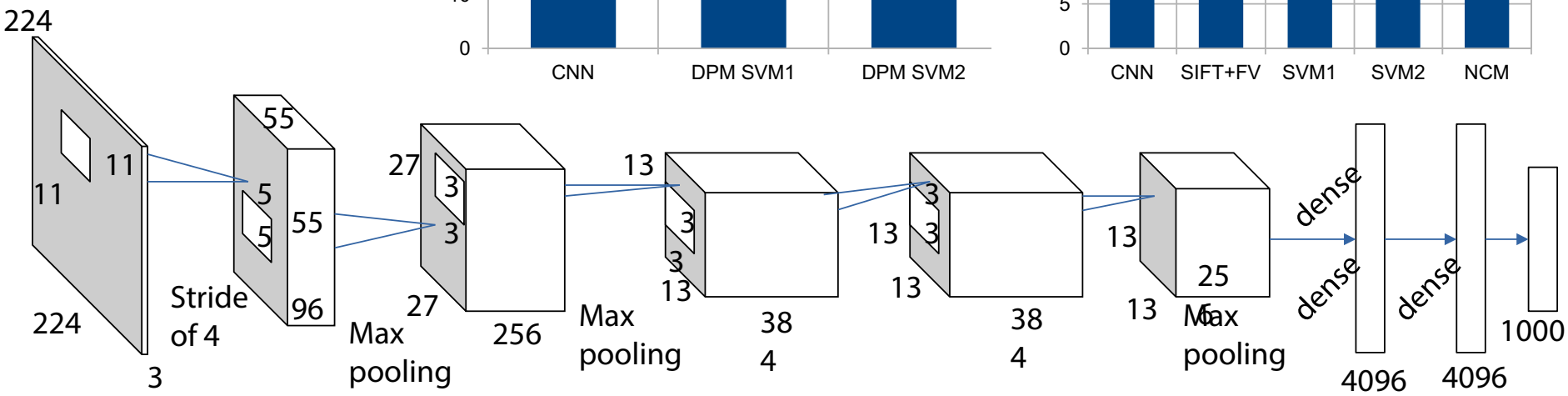
http://upload.wikimedia.org/wikipedia/commons/thumb/c/c4/Abandoned_Dog.jpg/128px-Abandoned_Dog.jpg

AlexNet: Convolutional Neural Network

- Won ImageNet 2012 LSVRC
  - 60 million parameters
  - 832 million MAC ops
  - Krizhevsky et al. 2012



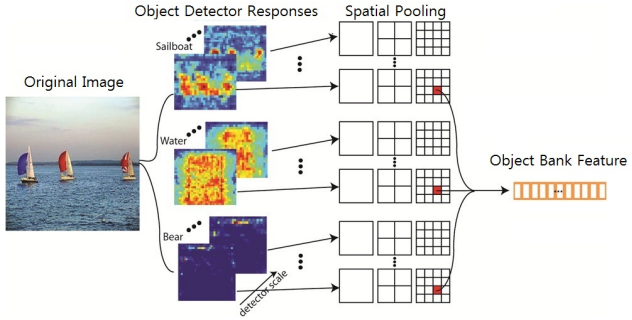




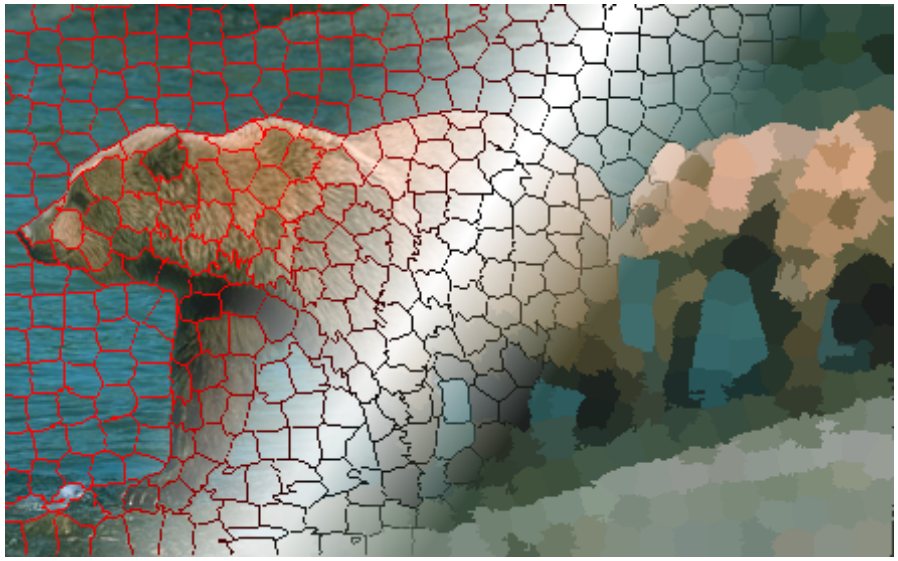
http://www.slideshare.net/yuhuang/deep-learning-for-image-denoising-superresolution-27435126

From Features to Labels (application specific)

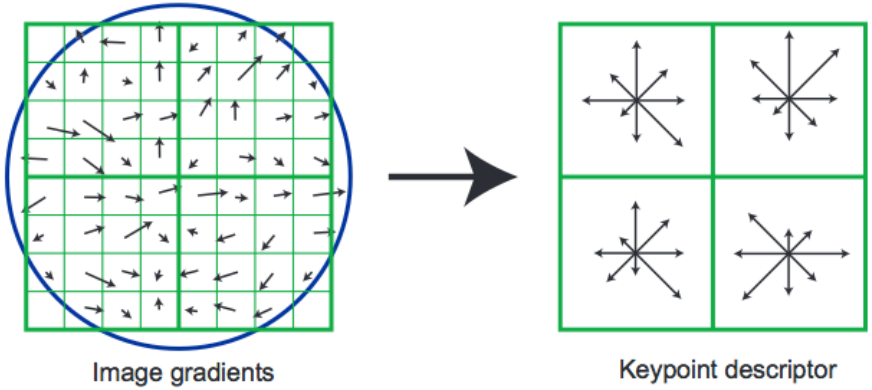
http://vision.stanford.edu/projects/objectbank/

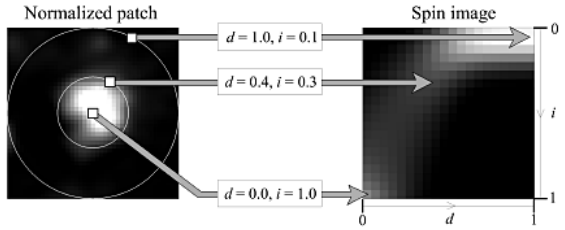


https://www.tnt.uni-hannover.de/project/superpixels/



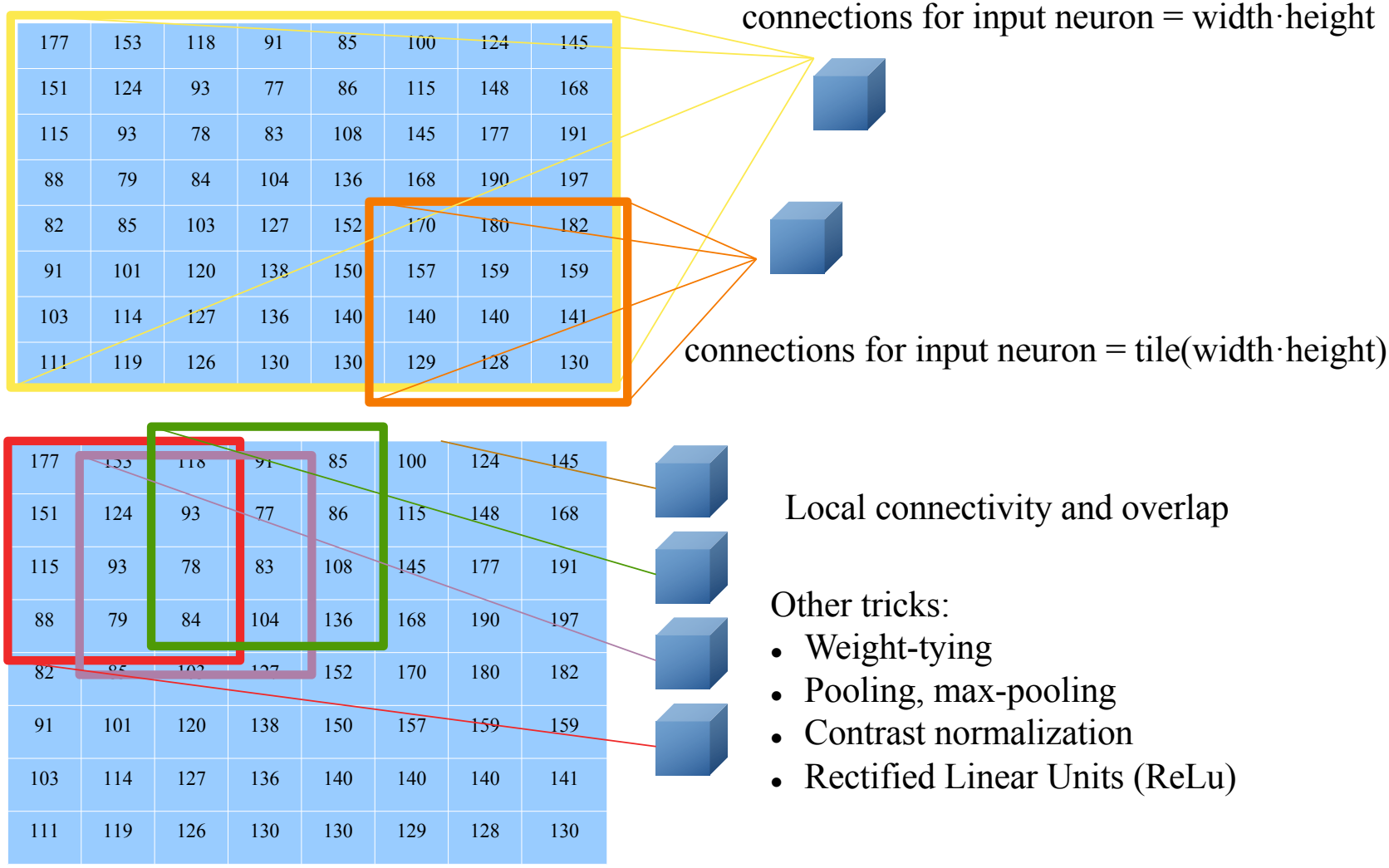
http://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf





http://doi.ieeecomputersociety.org/cms/Computer.org/dl/trans/tp/2005/08/figures/i12654.gif

DNN Design Space

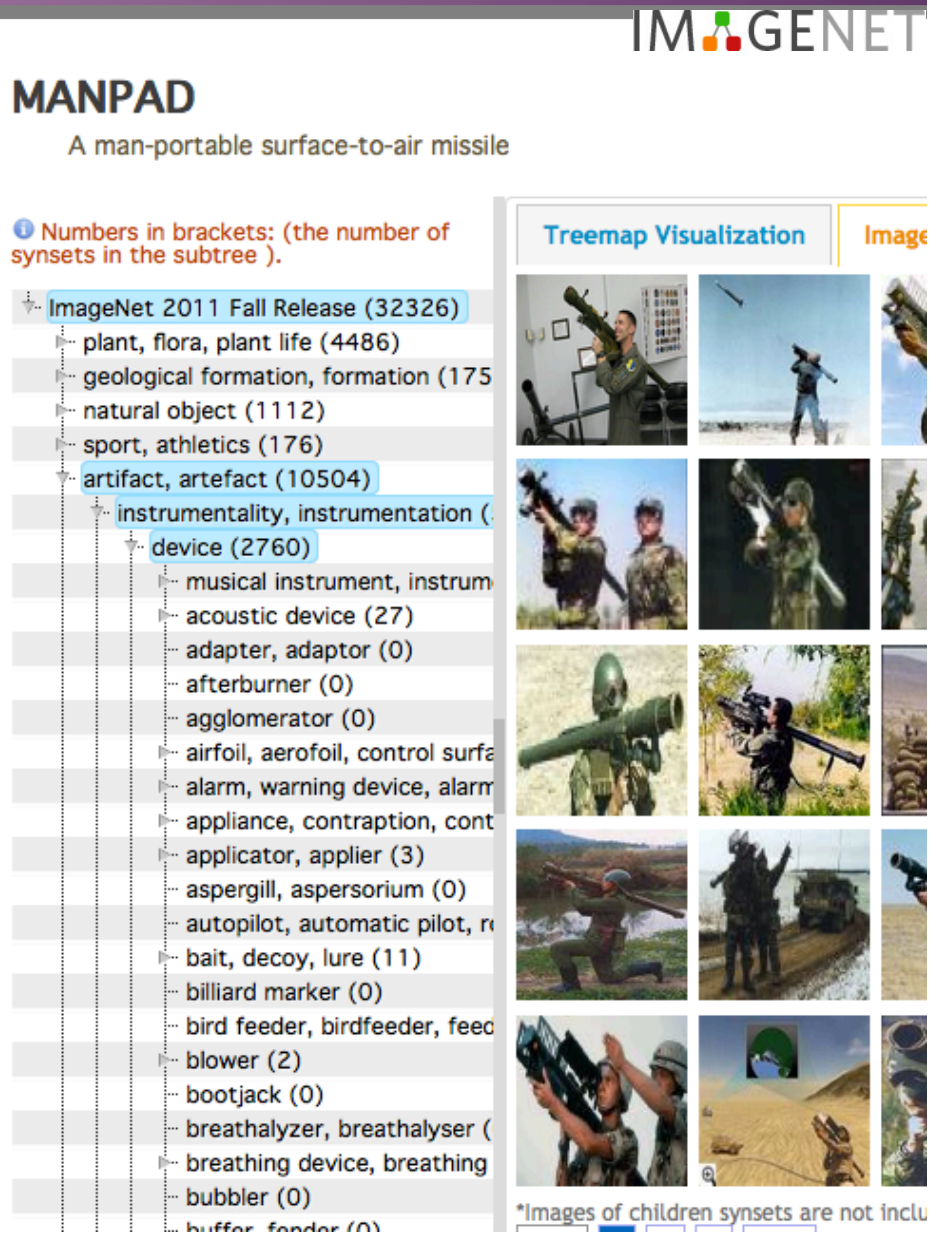


Training with Backpropagation: Points of Interest

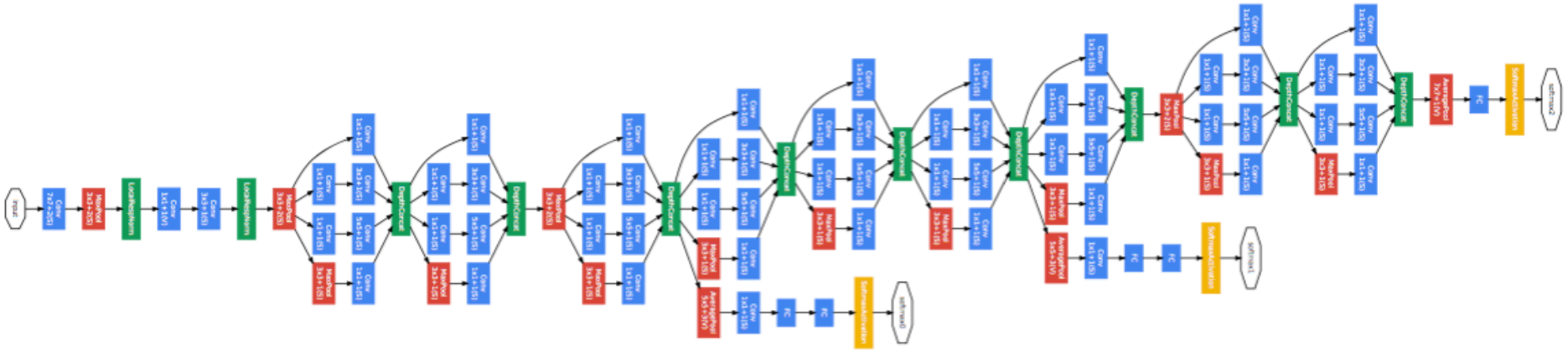
- We need multiple layers for complicated problems
  - Convexity, separability are not required
- Problems
  - Diminishing gradients
  - Unbound weights
    - Normalize
  - Overfitting
    - Addressed with regularization
  - Sensitivity to training set
    - Who ordered your training images (on disk)?
  - Training time vs. recall time
    - Weeks, days, hours vs. millisecond

ImageNet Collection of Classified Images

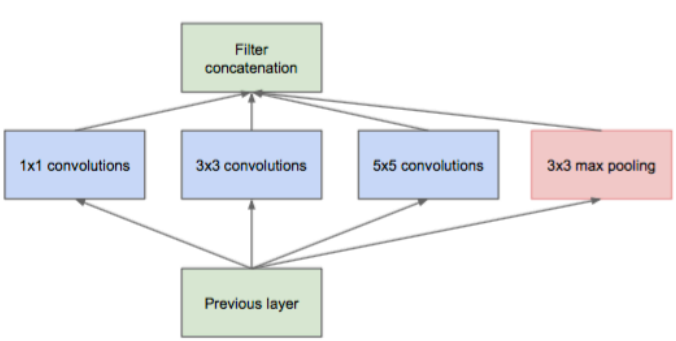
- 15 million images
  - Tagged with concepts (synonym sets = synsets)
- Some images have
  - SIFT features
  - Bounding box annotations
- Availability
  - Links to images, API
  - Full download for educational purposes
- Competition
  - ImageNet Large Scale Visual Recognition Challenge
- There are precision/recall numbers for human testers

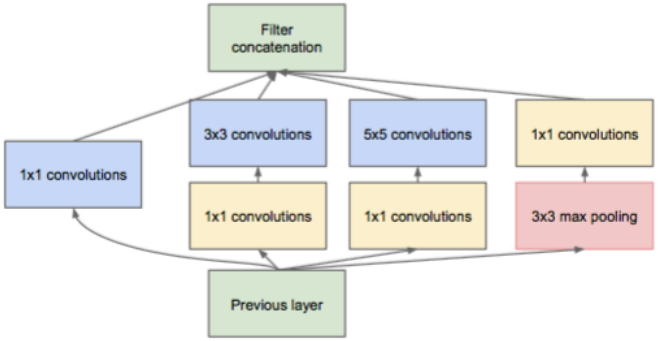


GoogLeNet (2014)



Inception Module





DNN Software Space

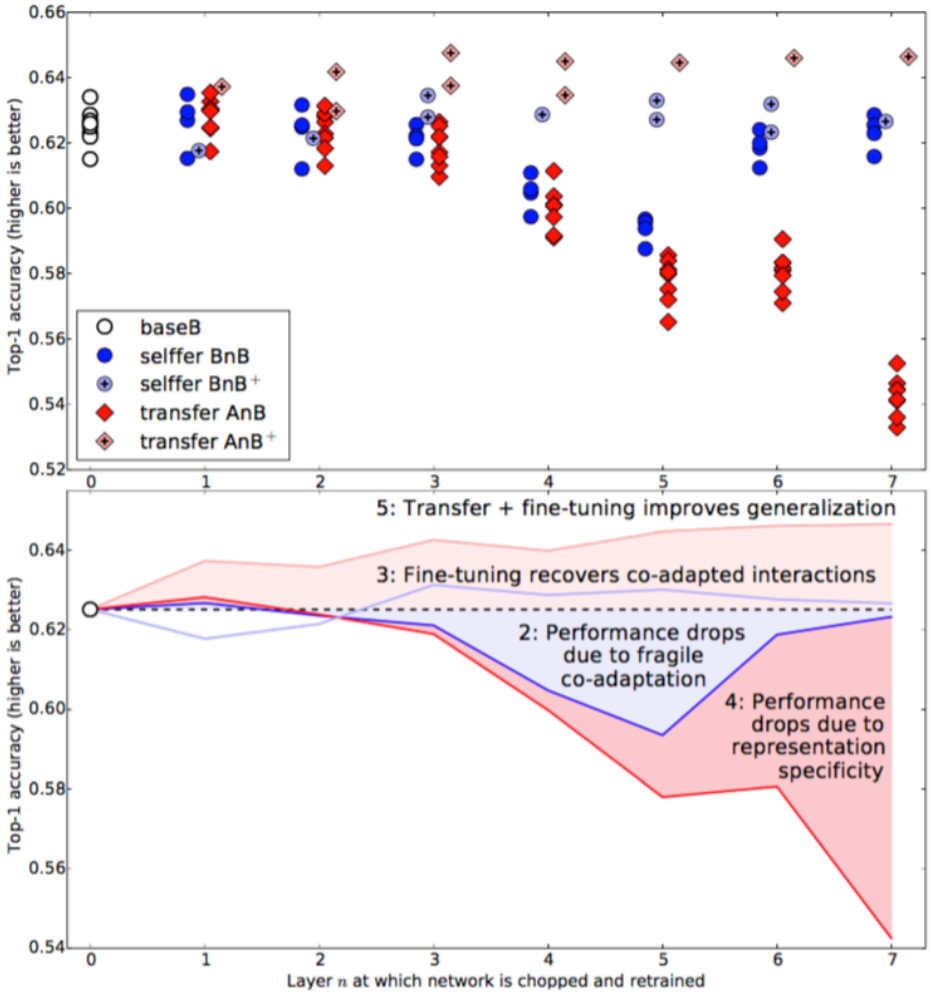
- Torch (based on LuaJIT): Facebook, Google, Twitter, NYU, ...
- Caffe (define network with JSON config file)
  - UC Berkeley demo.caffe.berkeleyvision.org
- Theano (high level: mathematical, Python, code generation)
  - University of Montreal
- cuDNN
  - NVIDIA, cuBLAS : Linear Algebra as cuDNN : Deep Learning
- MaxDNN, neon
  - Based on MaxAs (PasAs), Ebay and Nervana Systems (now part of Intel)
- Facebook Torch plugin
  - FFT-based convolution (fbfft faster than cuFFT), frequency space
  - Recently open sourced, competitive performance to cuDNN
- Others: ConvNet, ANNarchy, ...

Caffe Model Zoo

- Model-Zoo on GitHub
- Finetuning on Flickr Style
- Fully Convolutional Networks for Semantic Segmentation
- CaffeNet fine-tuned for Oxford flowers dataset
- CNN Models for Salient Object Subitizing
- Deep Learning of Binary Hash Codes for Fast Image Retrieval
- Scene Recognition
- Age and Gender Classification
- Car model classification
- Real-time semantic segmentation architecture for scene understanding
- Holistically-Nested Edge Detection
- Translating Videos to Natural Language
- Face CNN descriptor
- Yearbook Photo Dating
- Emotion Recognition in the Wild
- Facial Landmark Detection

Transferable Features between Layers

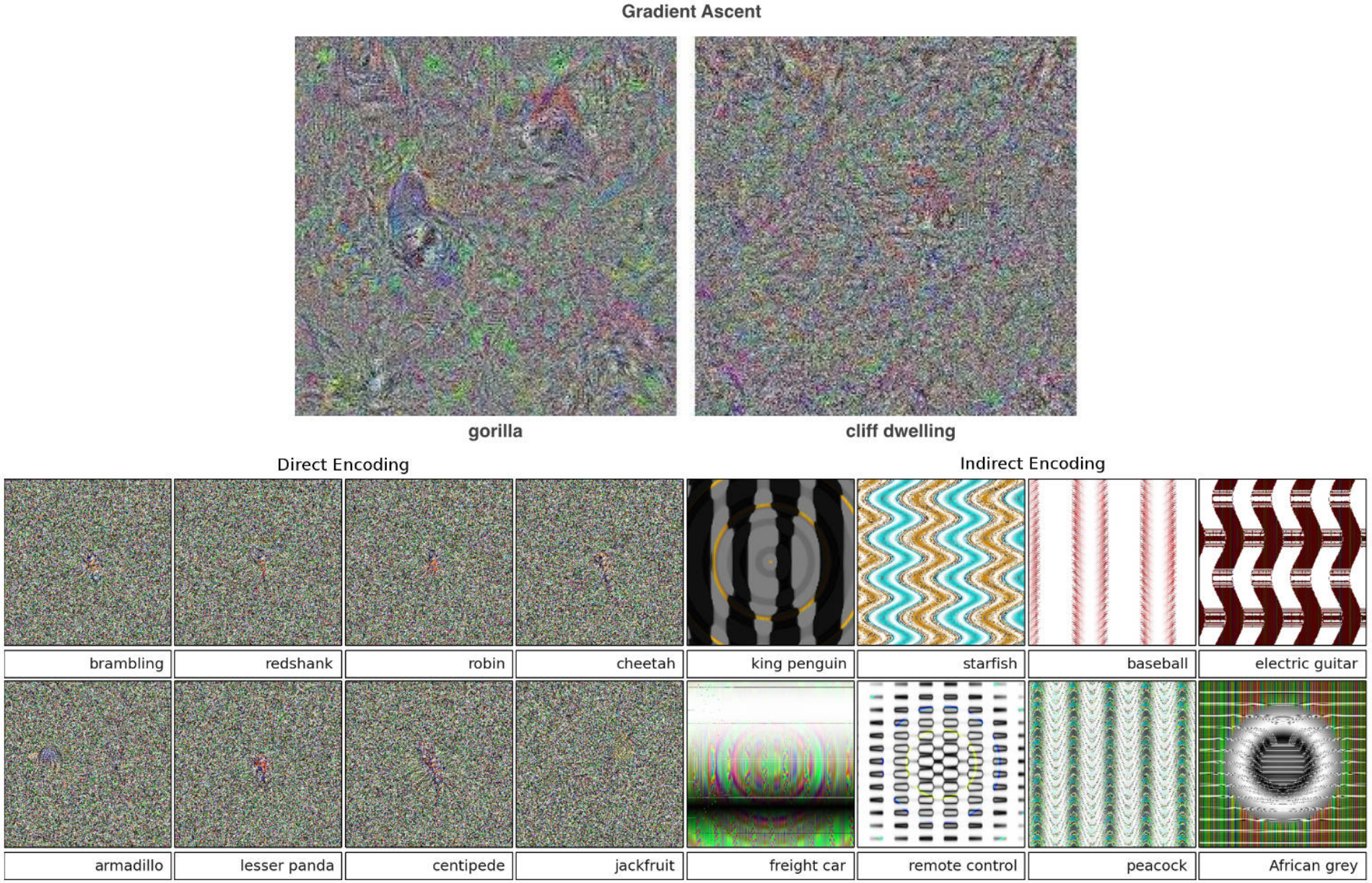
- Training takes a week
  - Produces weights
- Recall takes milliseconds
- Can training data be reused?
- Is swapping of layers OK?
- Pretraining



Yosinski et al. 2014

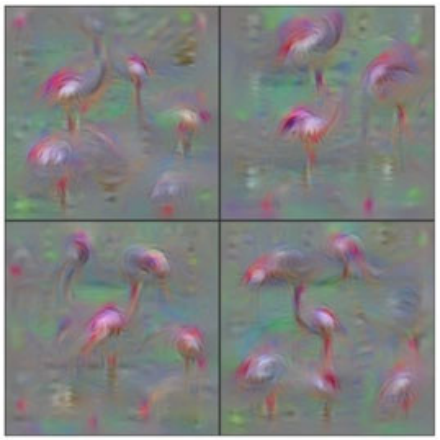
Chiwawa or Muffin?

Fooling DNNs

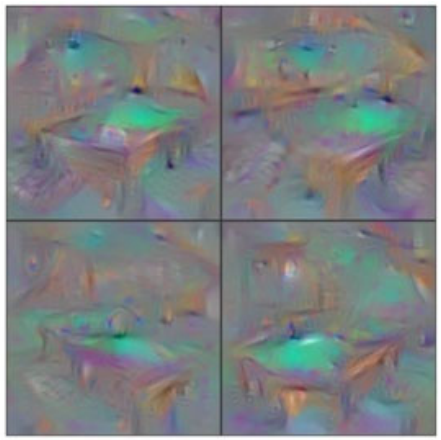


Yosinski et al. 2015

What Neurons "want to see"?

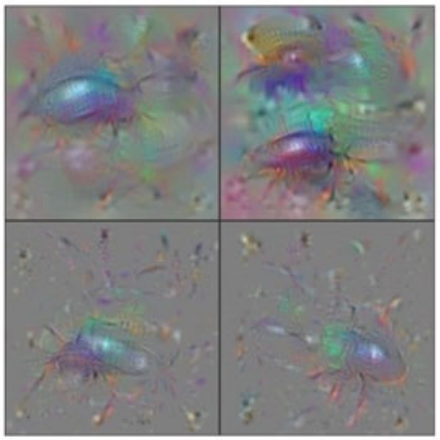


Flamingo

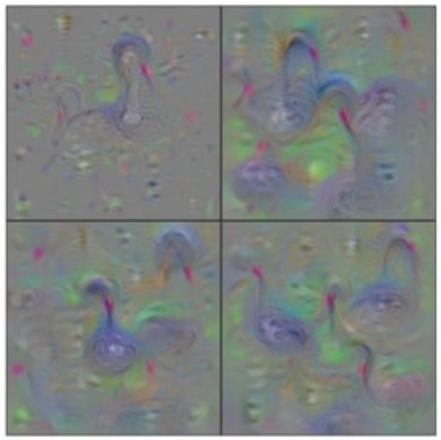


Billiard Table

School Bus



Ground Beetle

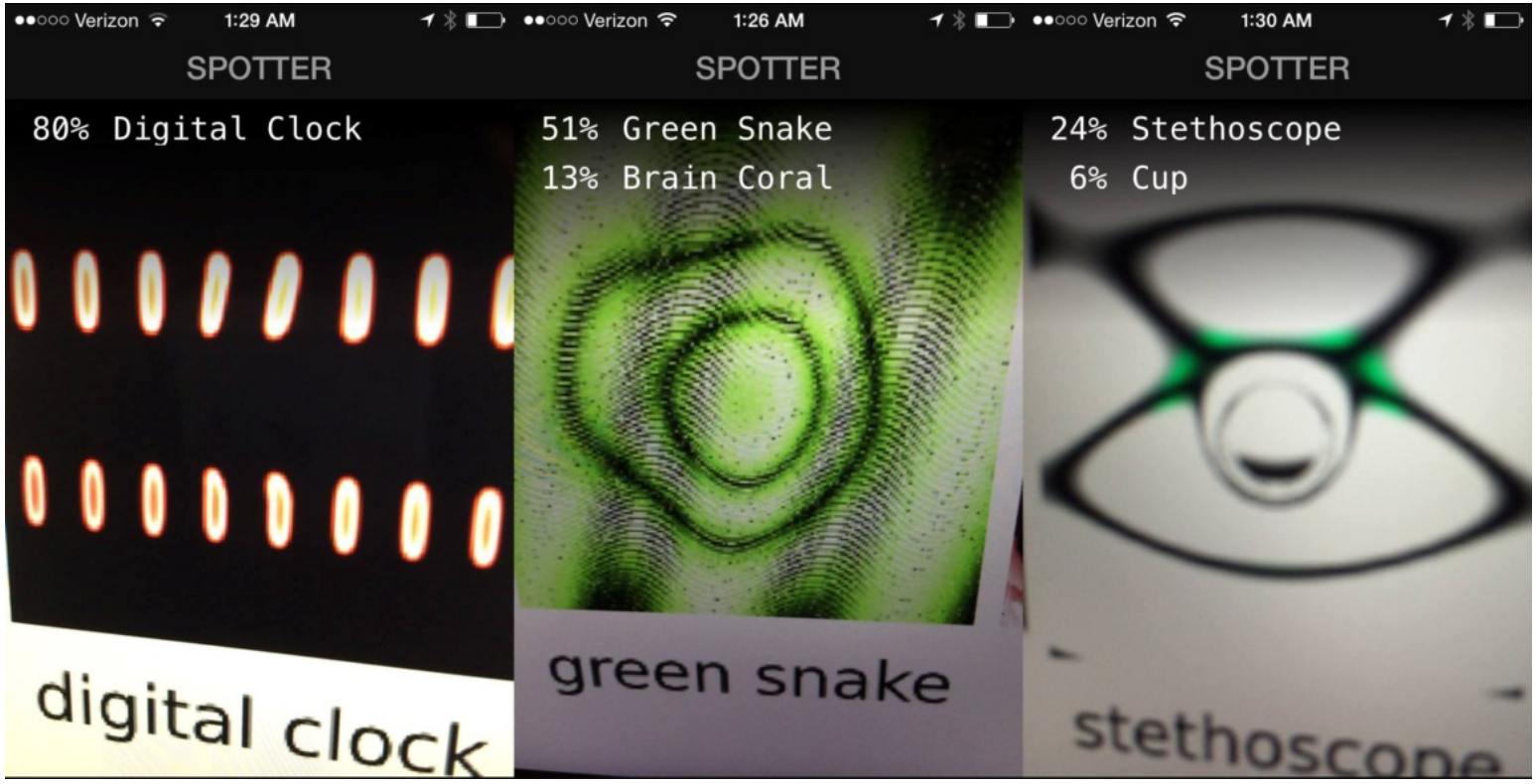


Black Swan

Tricycle

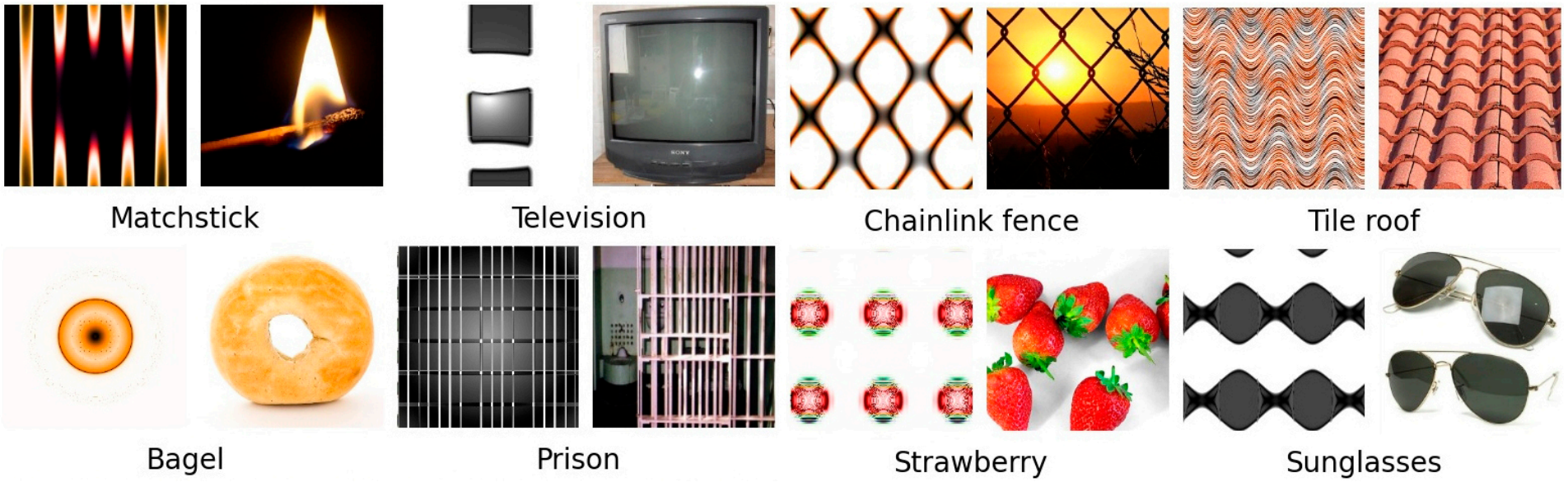
Yosinski et al. 2015

Fooling iPhone DNN



Yosinski et al.

Classification → Generation



Nguyen et al. 2015

FP16 Hardware for DNN Training/Inference

- AMD
  - MI5, MI8, MI25 (announced)
- ARM
  - NEON VFP FP16 in V8.2-A
- Intel
  - Knights Mill (announced)
- NVIDIA Pascal and Volta
  - P100, TX1 (shipping)
  - Non-Tesla cards (Quadro, GeForce)
  - V100, DGX-1, DGX-2
    - Tensor core with 32-bit intermediates
- Supercomputers
  - TSUBAME 3 (47 Pflop/s FP16)
  - Tokyo Tech (announced)
  - Piz Daint

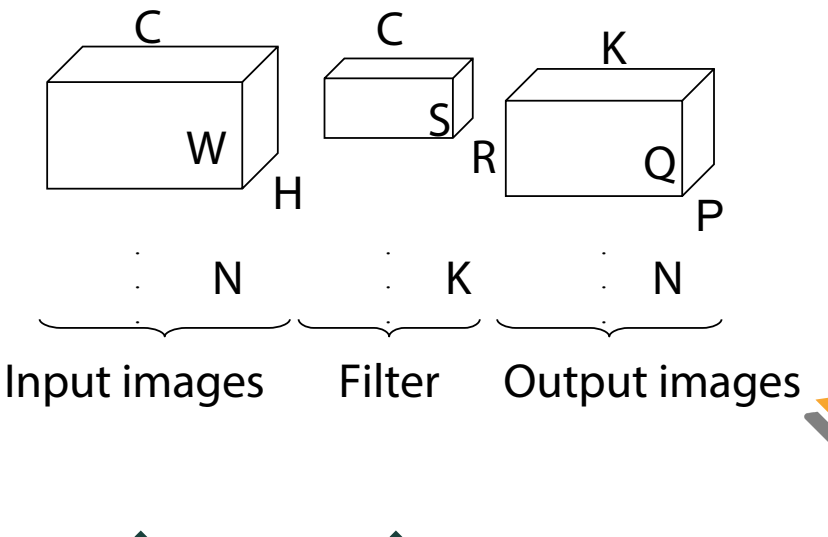
More Exotic Hardware for Deep Learning

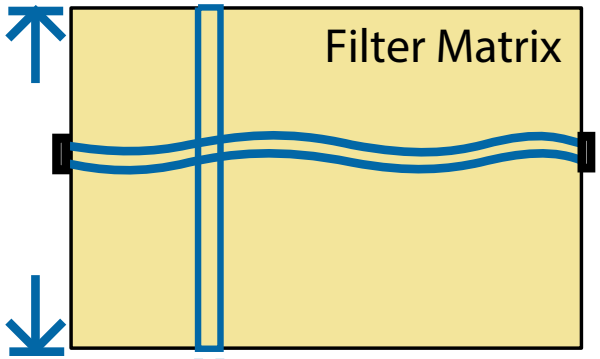
- Google
  - TPU 1: INT8 ~30 TOPS
  - TPU 2: available in the cloud
- NVIDIA
  - DRIVE PX 2: 24 DL TOPS
- Intel
  - Knights Mill

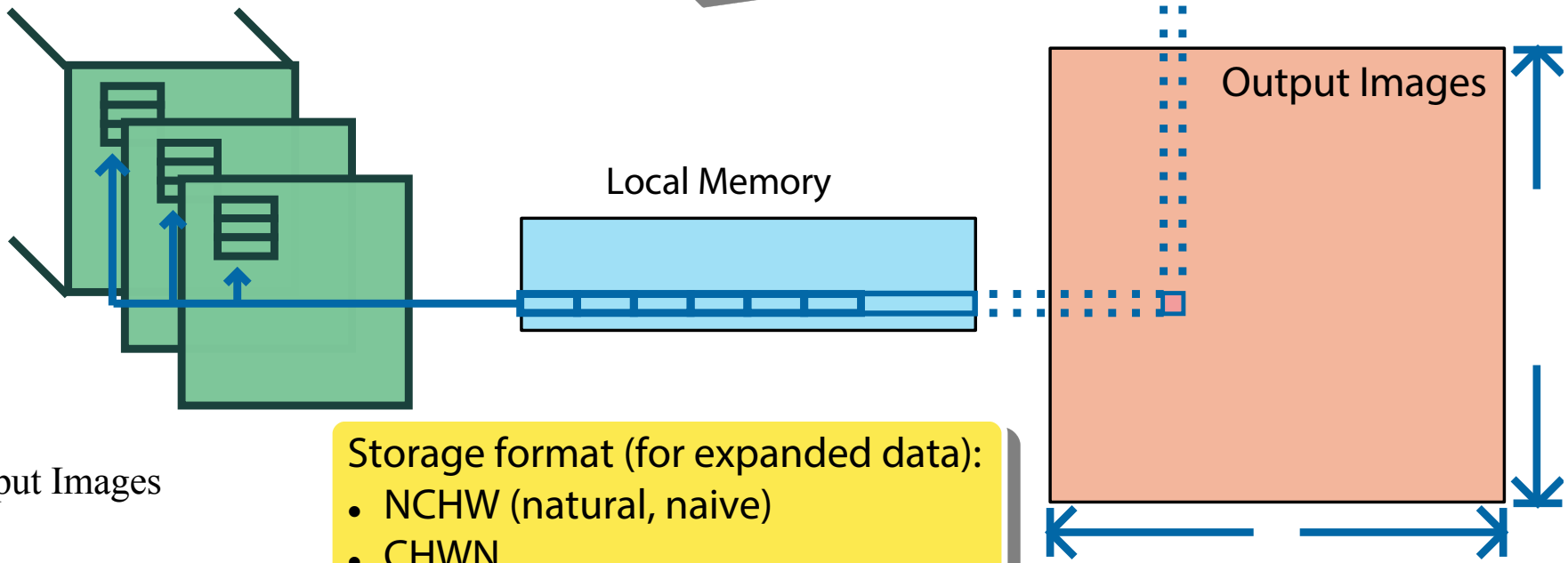
Low-Level Optimizations

- FP16: NVIDIA P100 and V100 have FP16 in hardware
  - cuDNN FP16 already had interface before the hardware shipped
    - It saves bandwidth even when computing in FP32
  - Rounding mode makes a difference (stochastic rounding)
- MaxAs, PasAs
  - Binary reverse engineering (below PTX)
  - MaxDNN and Neon (Nervana Systems, purchased by Intel)
  - Over 90% of peak performance on Deep Learning kernels
    - All major optimization methods used
      - Pipeline scheduling
      - Constant cache for computed indexes
      - Predication
- Strassen, Winograd progressions
  - Complexity theory: reduce operation count to higher performance

DNN → Tensors → SGEMM





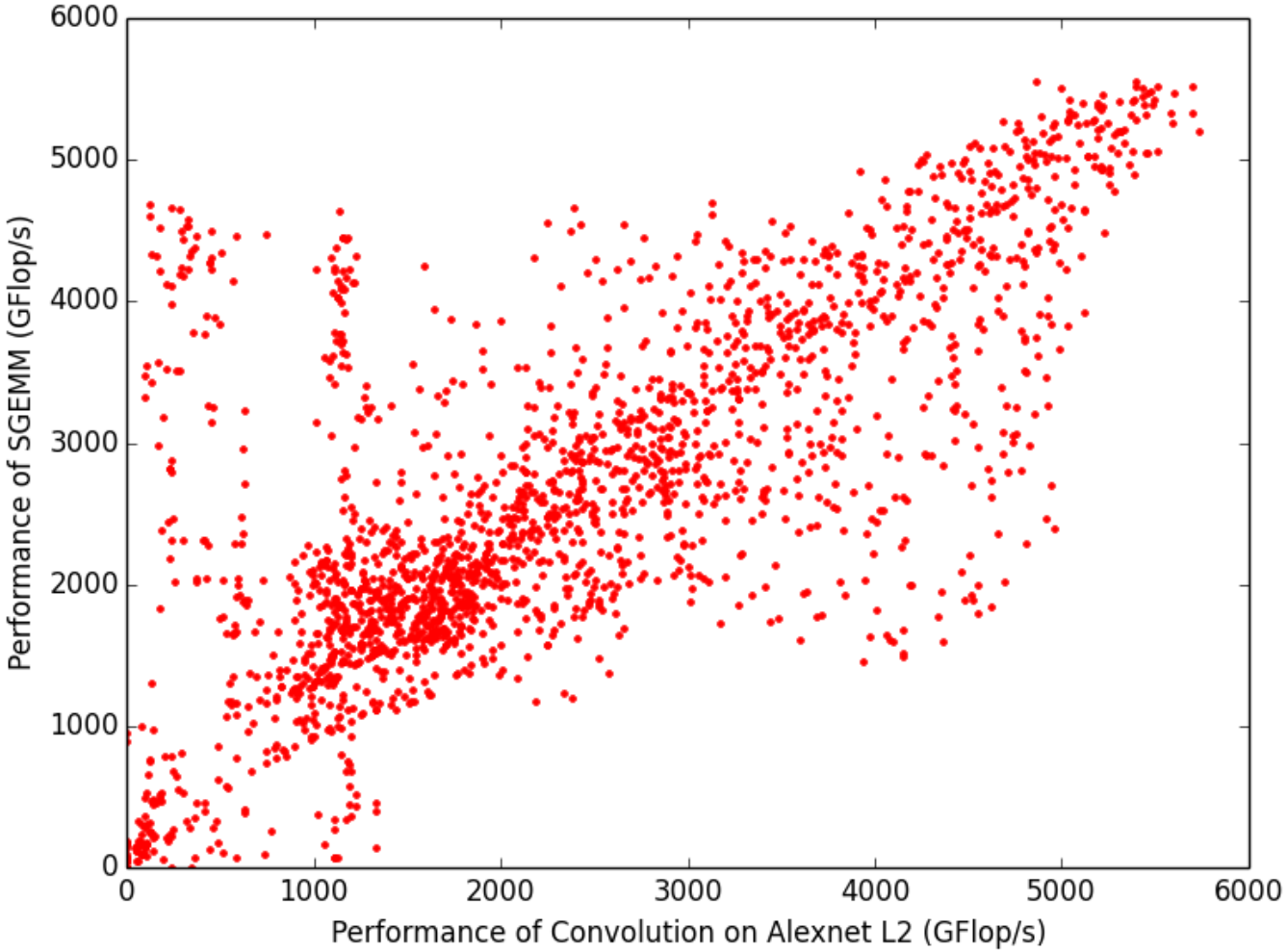


Input Images

Storage format (for expanded data):

- NCHW (natural, naive)
- CHWN

From SGEMM to DNN with Autotuning



Tensors vs. FFT

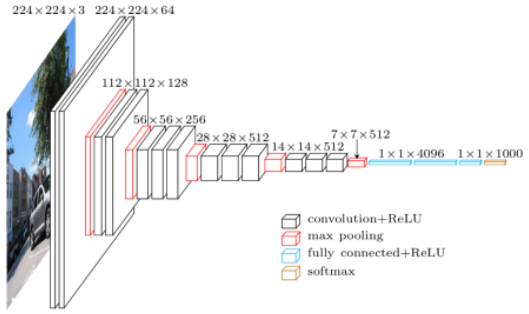
- Remember Convolution Theorem?
  - Fourier transform of convolution is a product of Fourier transforms
- FFT approach offers reduction in operation count
  - Even though the hardware efficiency is not at the same level as direct tensor computations
- The problem for FFT-based approach is the stride in convolution definition
  - This could be addressed with truncated FFTs which usually are even less efficient that regular FFTs
- This is the approach from Facebook FFT: fbfft library

Additional Resources

- www.deeplearningbook.org
  - Published by MIT Press
- Unsupervised Feature Learning and Deep Learning Tutorial
  - http://ufldl.stanford.edu/tutorial/
  - http://ufldl.stanford.edu/wiki/index.php
  - Recommended reading page
  - Code repo github.com/amaas/stanford_dl_ex
- Caffe
  - caffe.berkeleyvision.org/
  - developer.nvidia.com/cudnn
  - github.com/amd/OpenCL-caffe
  - Model Zoo
- Others
  - Google TensorFlow
  - Facebook Torch
  - Theano
  - Microsoft CNTK
  - Caffe 2 (http://caffe2.ai)
  - Keras
  - Mathematica
  - MATLAB
  - MxNet
- Machine learning
  - Kernel Machines
- Pretrained models
  - Mostly for ImageNet

Modern Networks and Data Sets

- Networks
  - VGG
    - Visual Geometry Group, Oxford
    - VGG CNN, VGG Annotator, VGG Search, VGG Face
    - VGG16, VGG19
  - ResNet
- Data Sets
  - MNIST
    - Handwritten digits
  - CIFAR10 and CIFAR100
    - 50,000 small images (32x32)
    - 10 or 100 classes of images
  - ImageNet
  - Cambridge Analytica
    - Personality types



ImageNet Training

- ResNet
  - Deep Residual Learning: https://arxiv.org/abs/1512.03385
- 1 hour on 256 GPUs
  - Accurate, Large Minibatch Stochastic Gradient Descend
  - https://arxiv.org/abs/1706.02677
- 15 minutes on 1024 GPUs
  - T. Akiba, S. Suzuki, and K. Fukuda. Extremely Large Minibatch SGD: Training ResNet-50 on ImageNet in 15 Minutes. NIPS 2017 Workshop: Deep Learning At Supercomputer Scale, 12 2017
  - https://github.com/chainer/chainermn
- Performance highlights
  - MPI Allreduce()
  - Synchronous vs. Asynchronous gradient updates
  - NVLink communication primitives: NVIDIA nccl (“nickel”)

Future Developments

- New network types
  - Capsule Networks
    - arxiv.org/abs/1710.09829
  - Compositional Pattern Producing Networks
- Automation of ML, DL, AI
  - AutoML from Google
- Active learning
  - Exploiting unlabeled data
    - Limited use of Amazon's Mechanical Turk
    - Large portion of Big Data is unlabeled
- Integration of ML and HPC
  - Scaling ML
  - Automating HPC

Homework

- Run cuDNN
- Run Caffe
- Program small NN