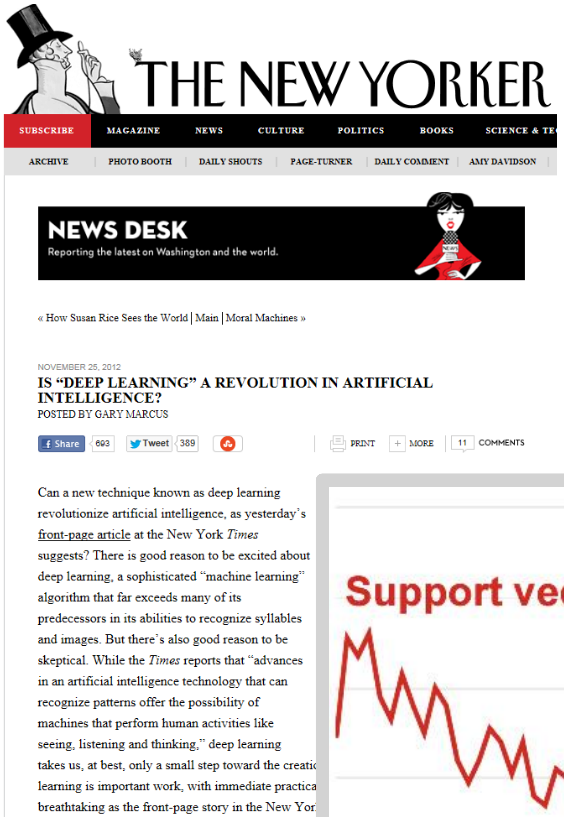
Deep Learning for Visual Perception

Sven Behnke

University of Bonn
Computer Science Institute VI
Autonomous Intelligent Systems

Much Interest in Deep Learning


THE NEW YORKER

SUBSCRIBE | MAGAZINE | NEWS | CULTURE | POLITICS | BOOKS | SCIENCE & TE

ARCHIVE | PHOTO BOOTH | DAILY SHOUTS | PAGE-TURNER | DAILY COMMENT | AMY DAVIDSON

NEWS DESK
Reporting the latest on Washington and the world.

« How Susan Rice Sees the World | Main | Moral Machines »

NOVEMBER 25, 2012

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

POSTED BY GARY MARCUS

Share 693 Tweet 389 PRINT MORE 11 COMMENTS

Can a new technique known as deep learning revolutionize artificial intelligence, as yesterday's front-page article at the New York *Times* suggests? There is good reason to be excited about deep learning, a sophisticated "machine learning" algorithm that far exceeds many of its predecessors in its abilities to recognize syllables and images. But there's also good reason to be skeptical. While the *Times* reports that "advances in an artificial intelligence technology that can recognize patterns offer the possibility of machines that perform human activities like seeing, listening and thinking," deep learning takes us, at best, only a small step toward the creati learning is important work, with immediate practica breathtaking as the front-page story in the New Yor


The Register®

Data Centre | Cloud | Software | Networks | Security | Policy | Business | Jobs | Hardware | Science | Bootno

Financial News | Small Biz | CIO | Media

BUSINESS

Baidu muscles in on Google's turf with Silicon Valley deep learning lab

Chinese search giant beds down next to Apple in Cupertino

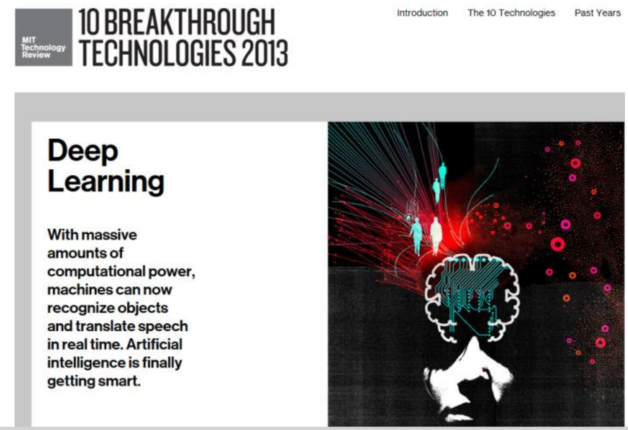
By Phil Muncaster, 15th April 2013 Follow 3,371 followers

Win a Samsung 40-inch LED HDTV with The Reg and HP!

1

RELATED

Chinese search giant Baidu has opened the doors to a new research facility in Google's back yard where it's hoping to tap the local talent to consolidate early mover advantage in the burgeoning field of "deep learning".

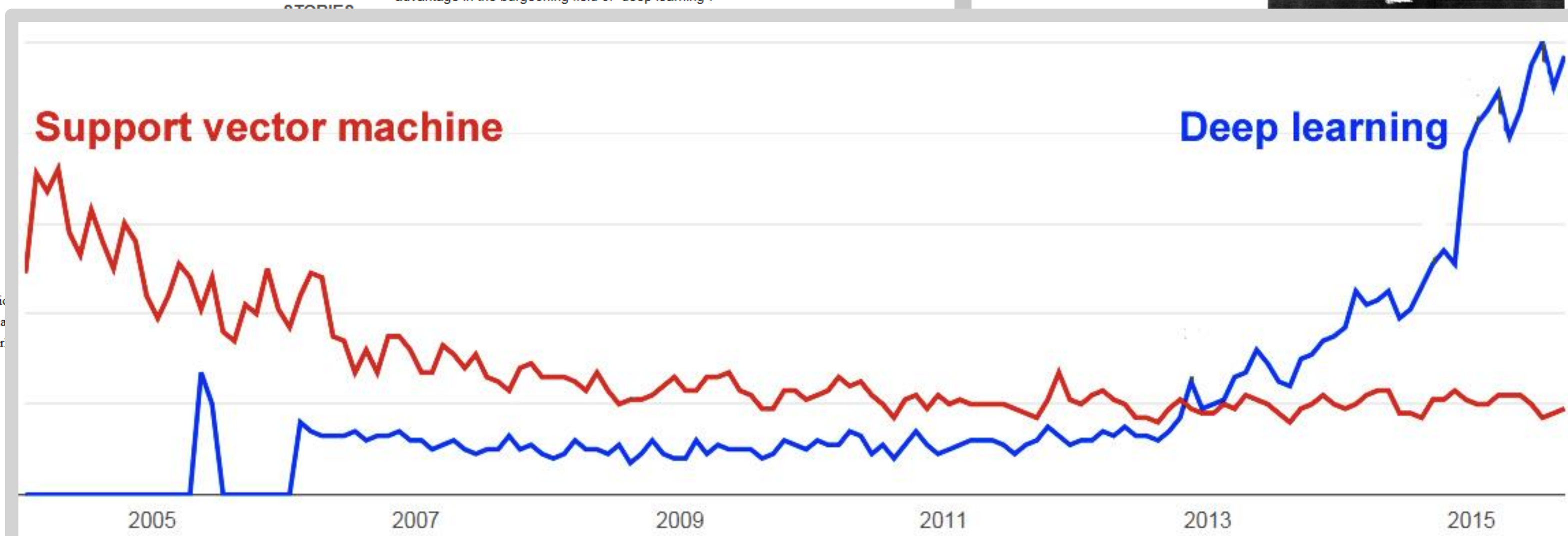

MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013

Introduction | The 10 Technologies | Past Years

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.



[Google Trends]

Industry Acquisitions and Hirings

- Google
  - DNNresearch (Geoffrey Hinton)
  - DeepMind (Demis Hassabis)
- Baidu
  - Andrew Ng
- Facebook
  - Yann LeCun
- Microsoft
  - Li Deng

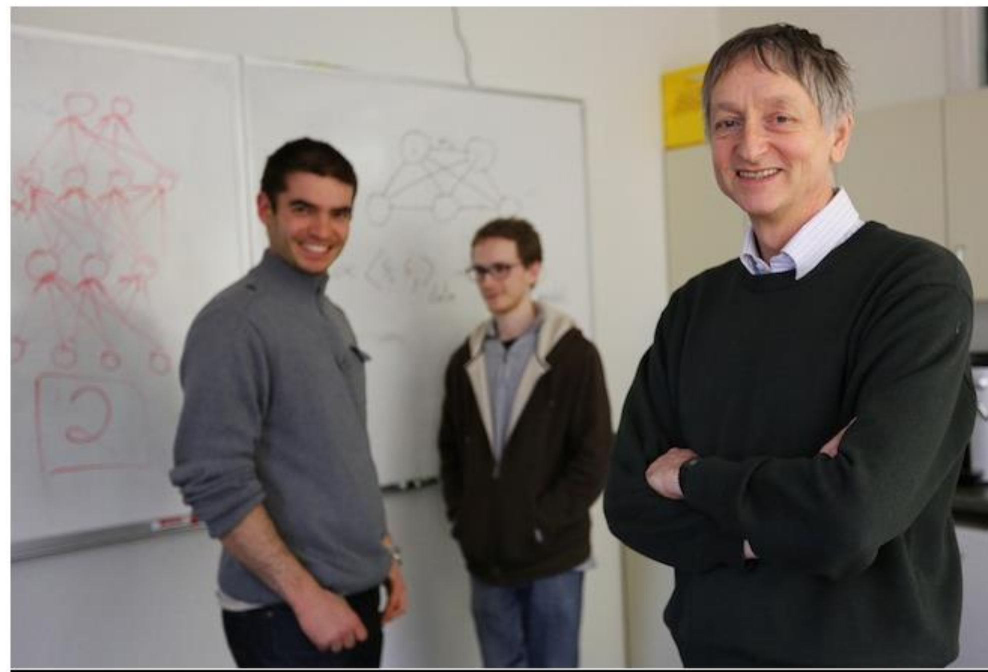
WIRED GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN

Google Hires Brains that Helped Supercharge Machine Learning

BY ROBERT MCMILLAN 03.13.13 6:30 AM

Follow @bobmcmillan

Share 672 Tweet 272 +1 144 in 63



Geoffrey Hinton (right) Alex Krizhevsky, and Ilya Sutskever (left) will do machine learning work at Google. *Photo: U of T*

Special Issues and Meetings

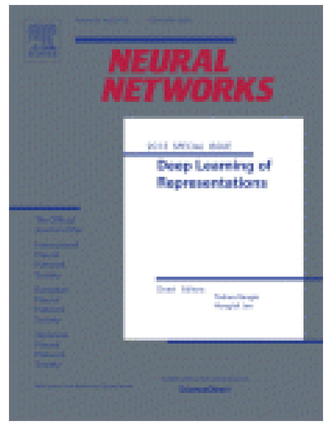
- Special issues of many journals (PAMI, NN)

IEEE TRANSACTIONS ON
PATTERN ANALYSIS AND MACHINE INTELLIGENCE
CELEBRATING 35 YEARS
A publication of the IEEE Computer Society

AUGUST 2013 VOLUME 35 NUMBER 8 ITPIDJ (ISSN 0162-8828)

SPECIAL SECTION ON LEARNING DEEP ARCHITECTURES

Guest Editors' Introduction: Special Section on Learning Deep Architectures
S. Bengio, L. Deng, H. Larochelle, H. Lee, and R. Salakhutdinov 1795
Representation Learning: A Review and New Perspectives
Y. Bengio, A. Courville, and P. Vincent 1798
Learning to Relate Images
R. Memisevic 1829
Deep Hierarchies in the Primate Visual Cortex: What Can We Learn for Computer Vision?
N. Krüger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. Rodriguez-Sanchez, and L. Wiskott 1847
Invariant Scattering Convolution Networks
J. Bruna and S. Mallat 1872
Deep Learning with Hierarchical Convolutional Factor Analysis
B. Chen, G. Polatkan, G. Sapiro, D. Blei, D. Dunson, and L. Carin 1887
Scaling Up Spike-and-Slab Models for Unsupervised Feature Learning
I.J. Goodfellow, A. Courville, and Y. Bengio 1902
Learning Hierarchical Features for Scene Labeling
C. Farabet, C. Couprie, L. Najman, and Y. Lecun 1915
Stacked Autoencoders for Unsupervised Feature Learning and Multiple Organ Detection in a Pilot Study Using 4D Patient Data
H.-C. Shin, M.R. Orton, D.J. Collins, S.J. Doran, and M.O. Leach 1930



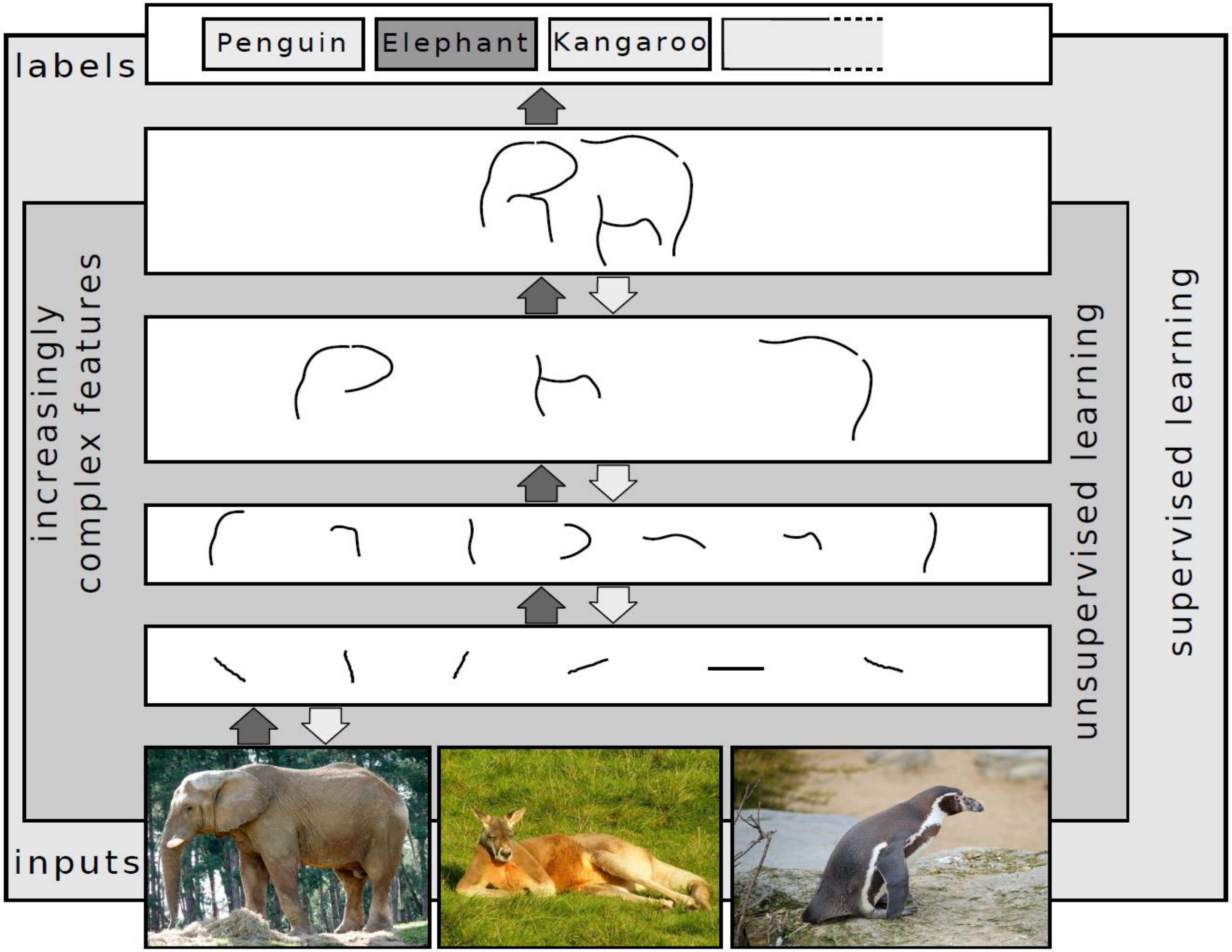
- Specialized workshops at major machine learning conferences (NIPS, ICLM)
- Representation Learning Conference (ICLR)
- Deep Learning Summits (RE.WORK, NVidia)

Deep Learning Definition

- Deep learning is a set of algorithms in machine learning that attempt to **learn layered models of inputs**, commonly neural networks.
- The layers in such models correspond to **distinct levels of concepts**, where
  - higher-level concepts are defined from lower-level ones, and
  - the same lower-level concepts can help to define many higher-level concepts.

[Bengio 2009]

Layered Representations

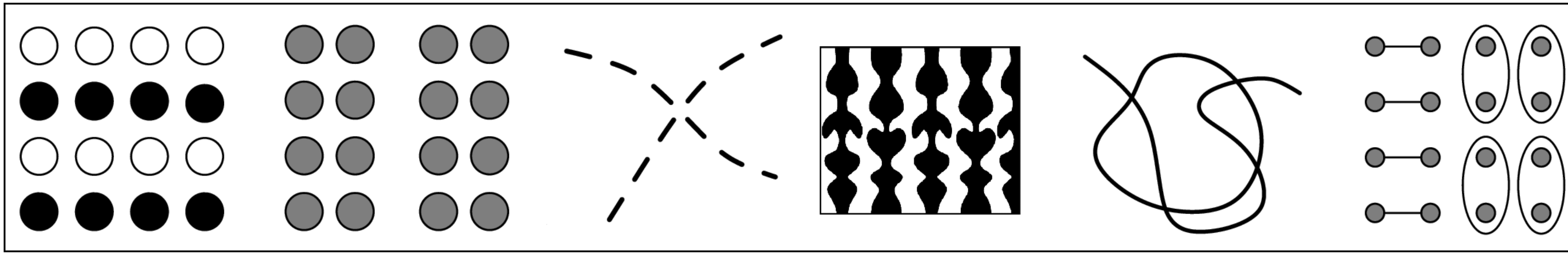


[Schulz and Behnke, KI 2012]

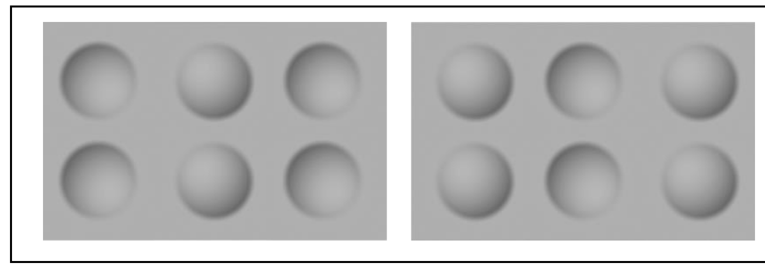
Performance of the Human Visual System

Psychophysics

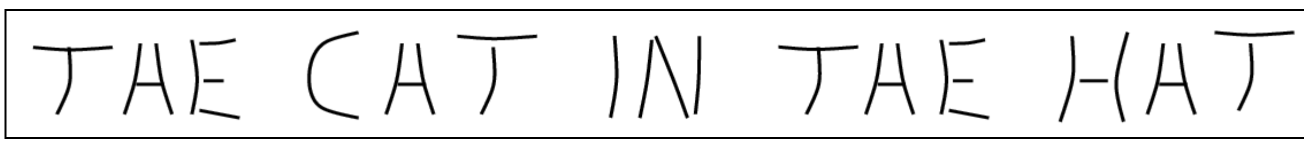
- Gestalt principles



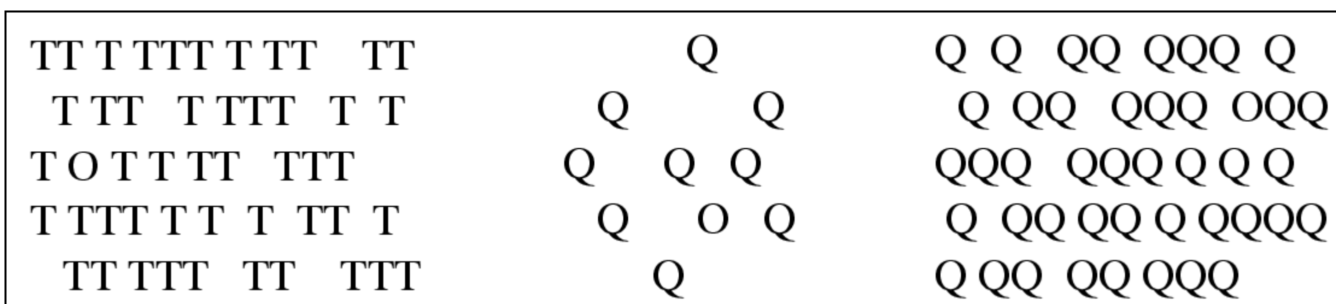
- Heuristics



- Context

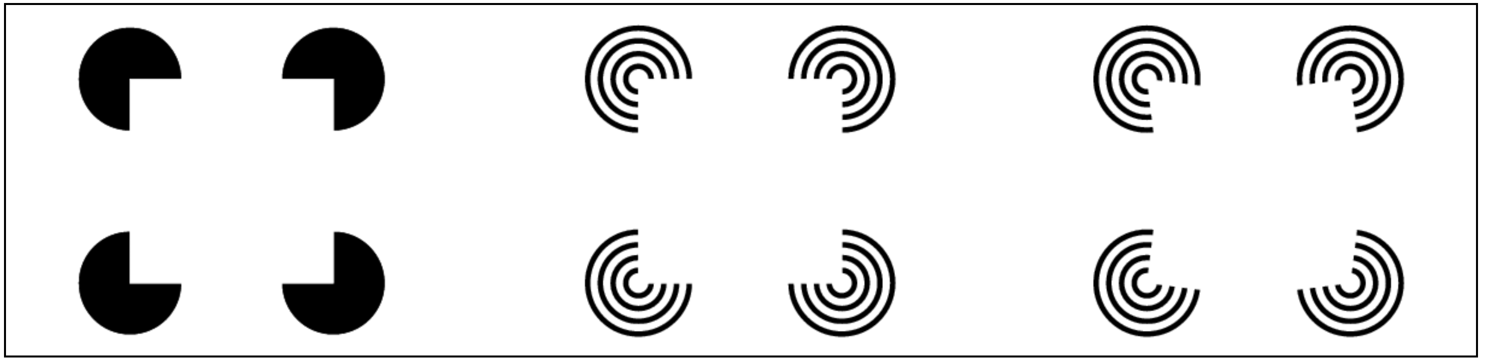


- Attention



Visual Illusions

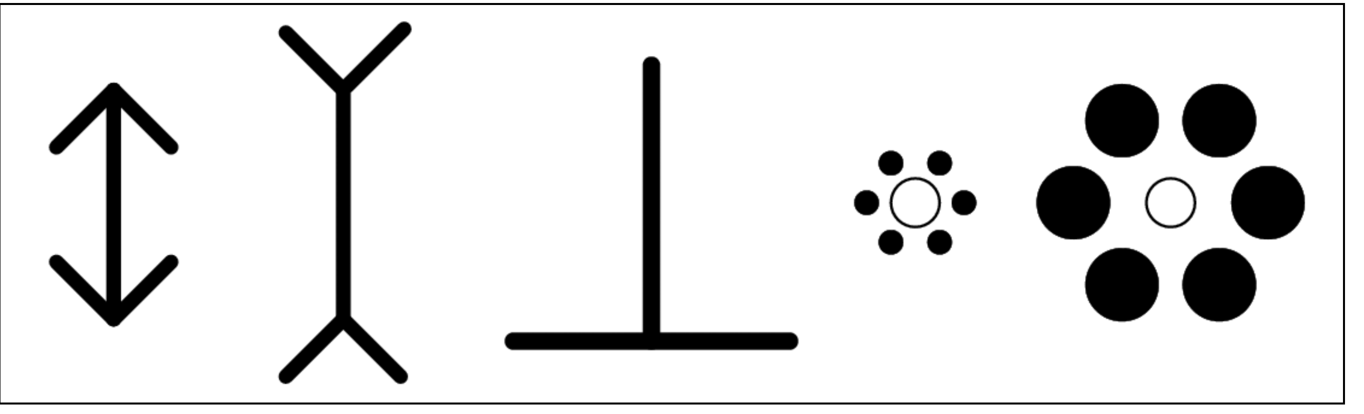
Kanizsa Figures



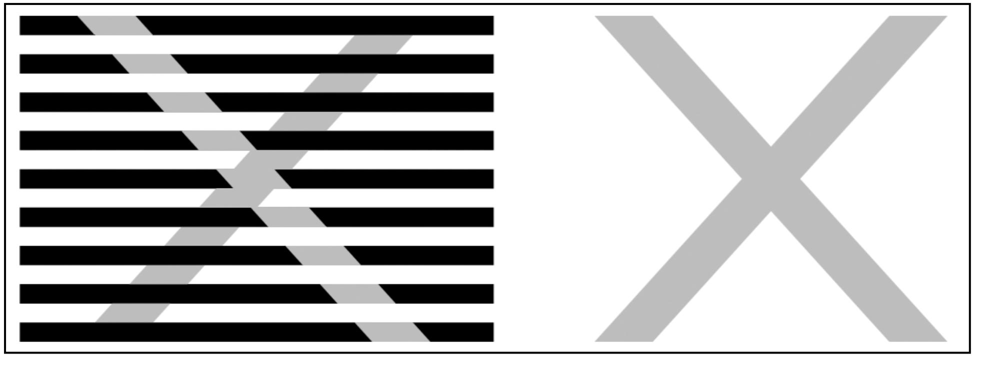
Müller-Lyer

horizontal/ vertical

Ebbinghaus-Titchener



Munker-White

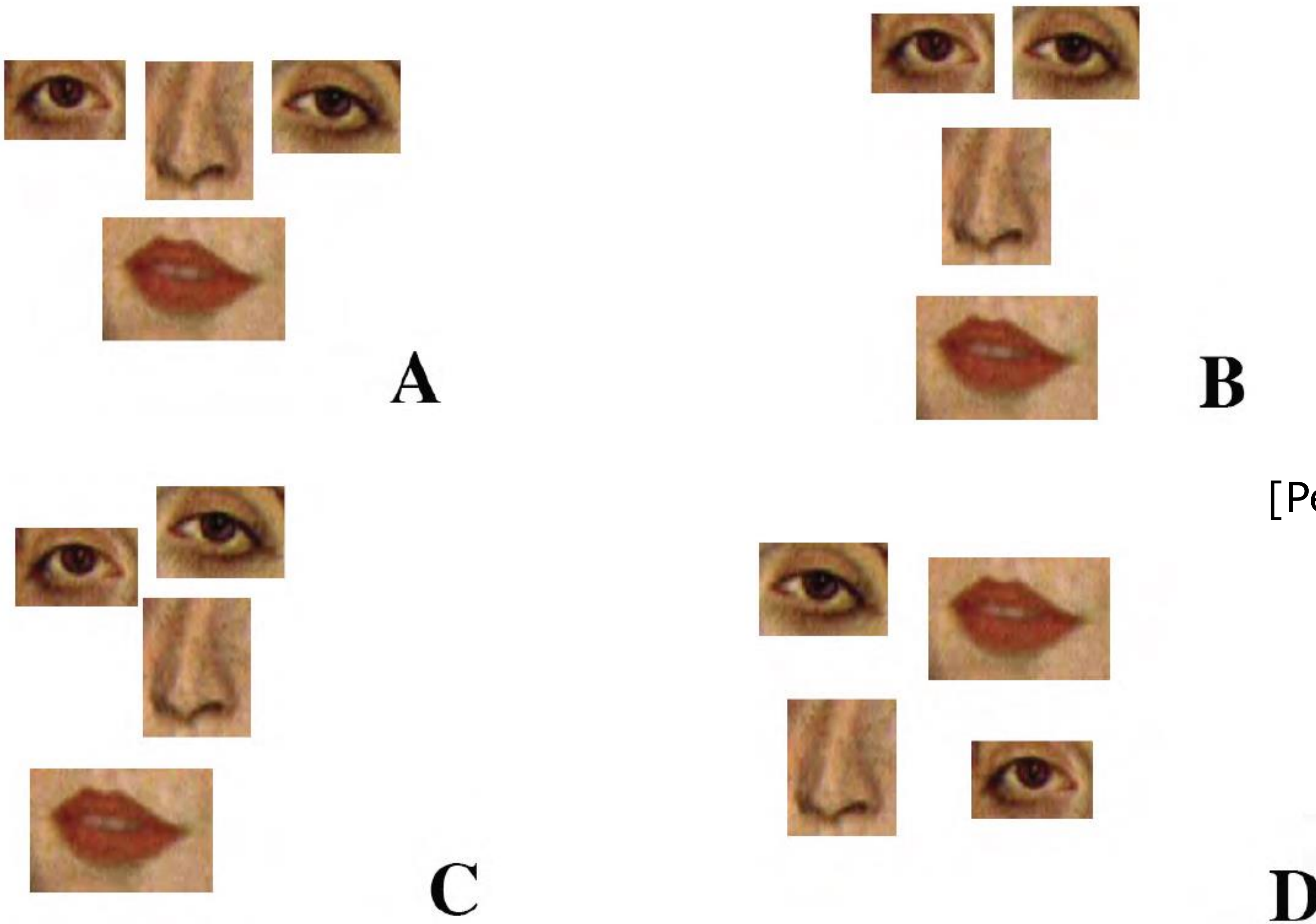


Observations

In the world around us it mostly holds that:

- Neighboring things have something to do with each other
  - Spatially
  - Temporally
- There is hierarchical structure
  - Objects consist of parts
  - Parts are composed of components, …

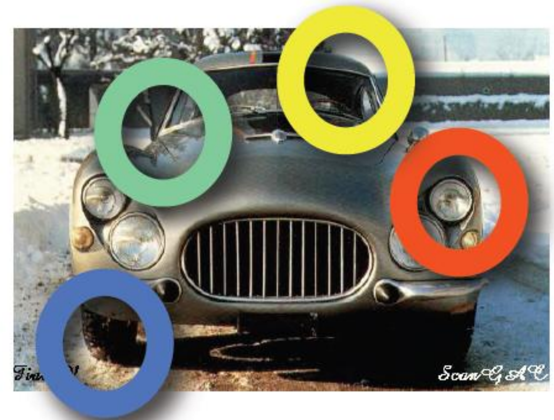
Spatial Arrangement of Facial Parts



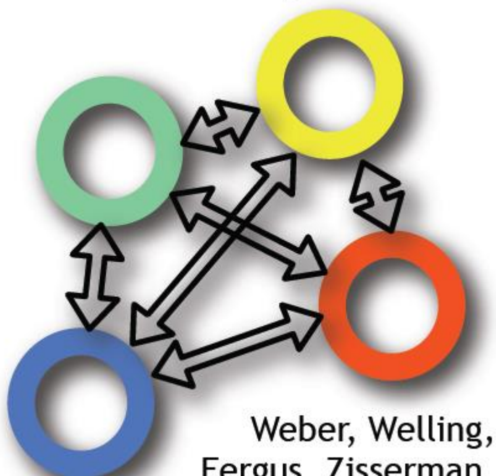
[Perona]

Face Perception

Horizontal and Vertical Dependencies

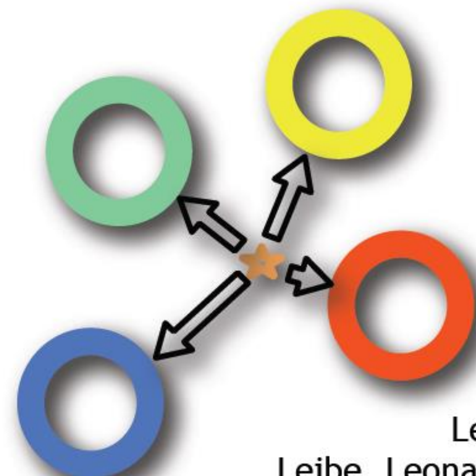


Constellation Model:
Fully connected shape model



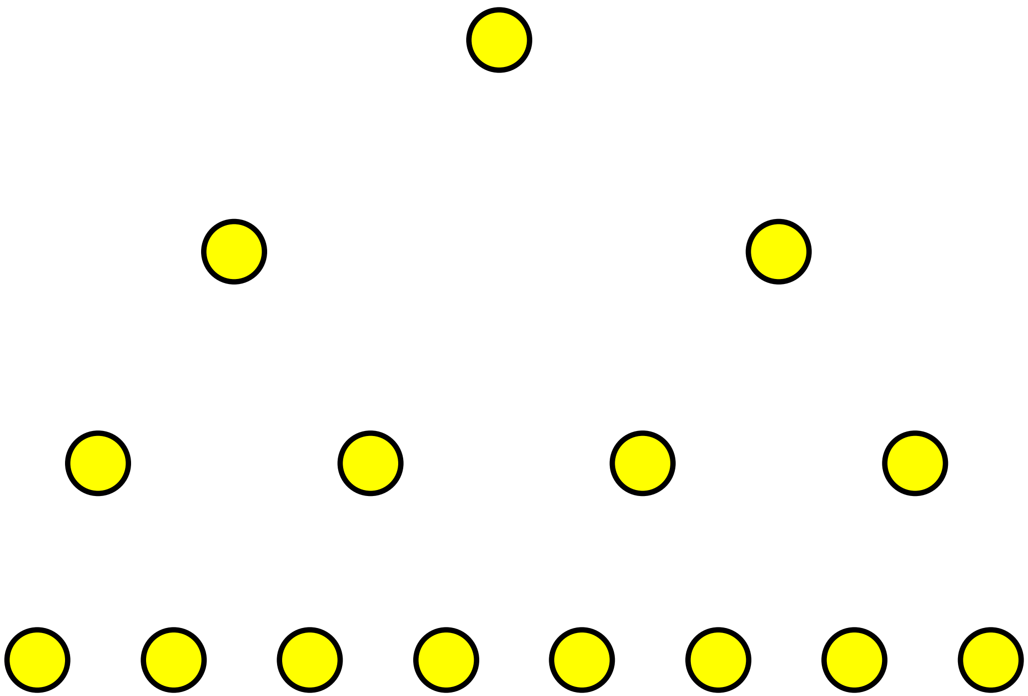
Weber, Welling, Perona ’00
Fergus, Zisserman, Perona ’03

Implicit Shape Model:
Star-Model w.r.t. Reference Point



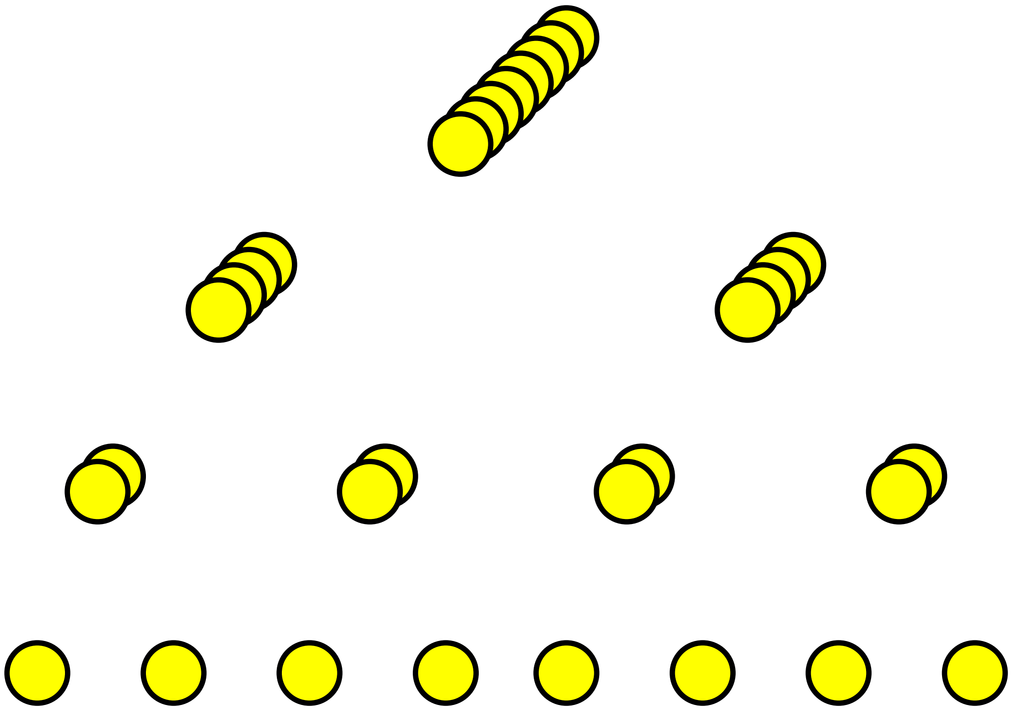
Leibe, Schiele ’03
Leibe, Leonardis, Schiele ’04

Multi-Scale Representation



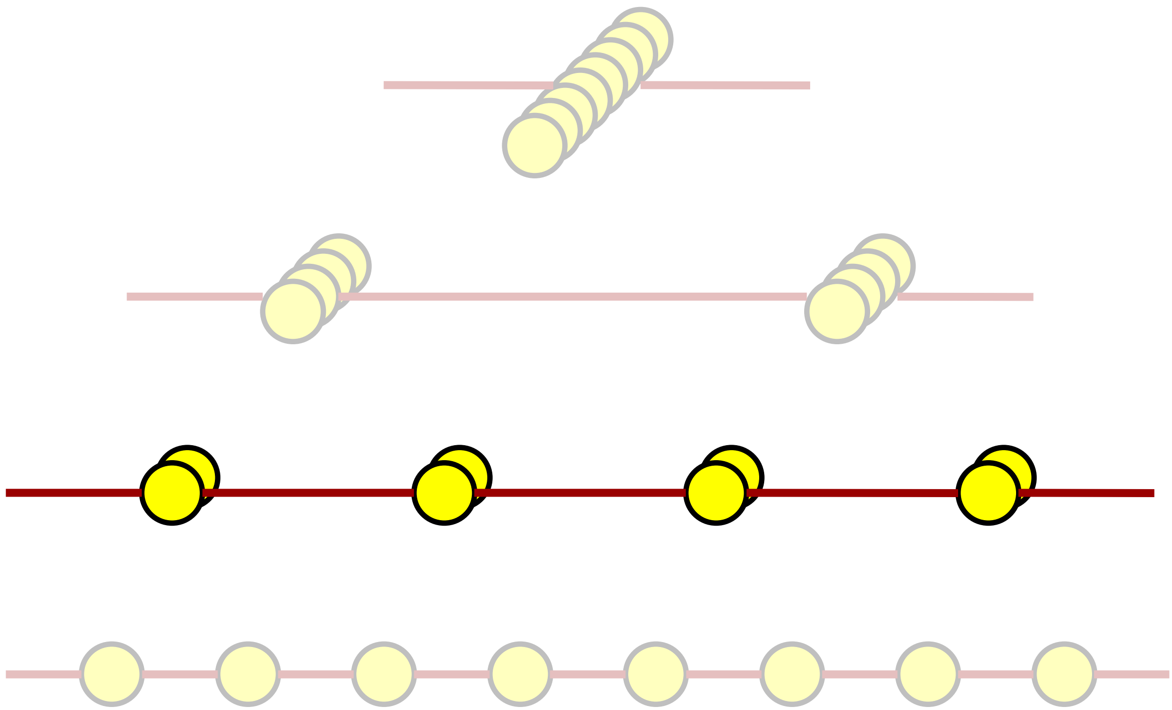
- Image pyramids are not expressive enough

Increasing Number of Features with Decreasing Resolution



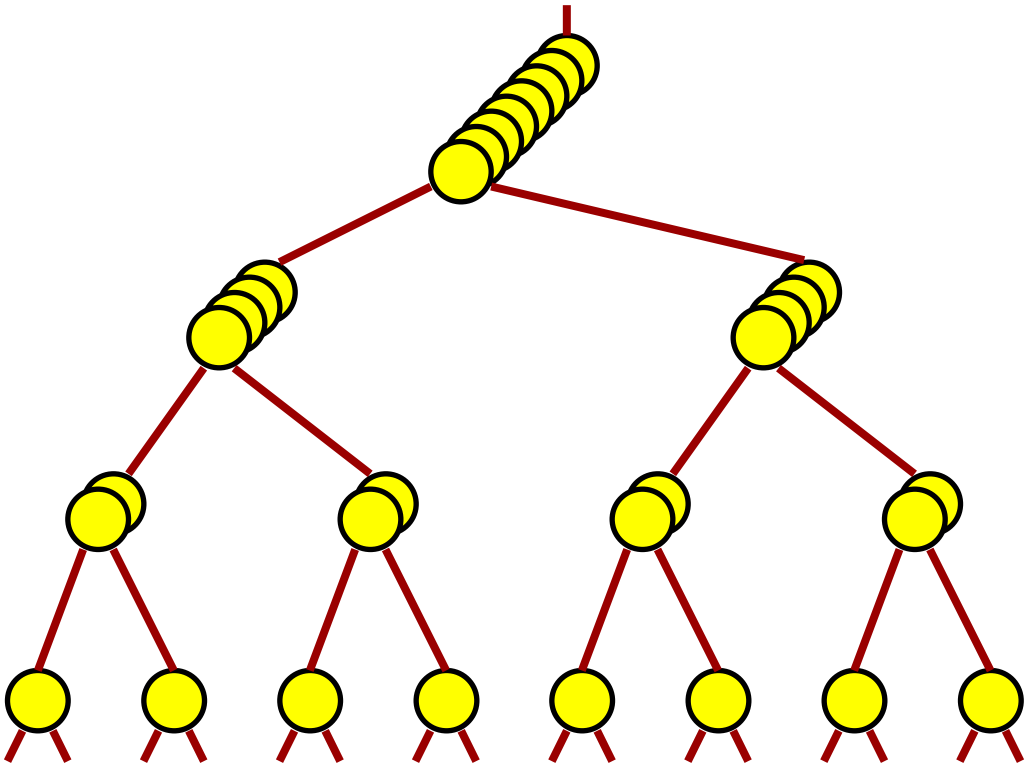
- Rich representations also in the higher layers

Modeling Horizontal Dependencies



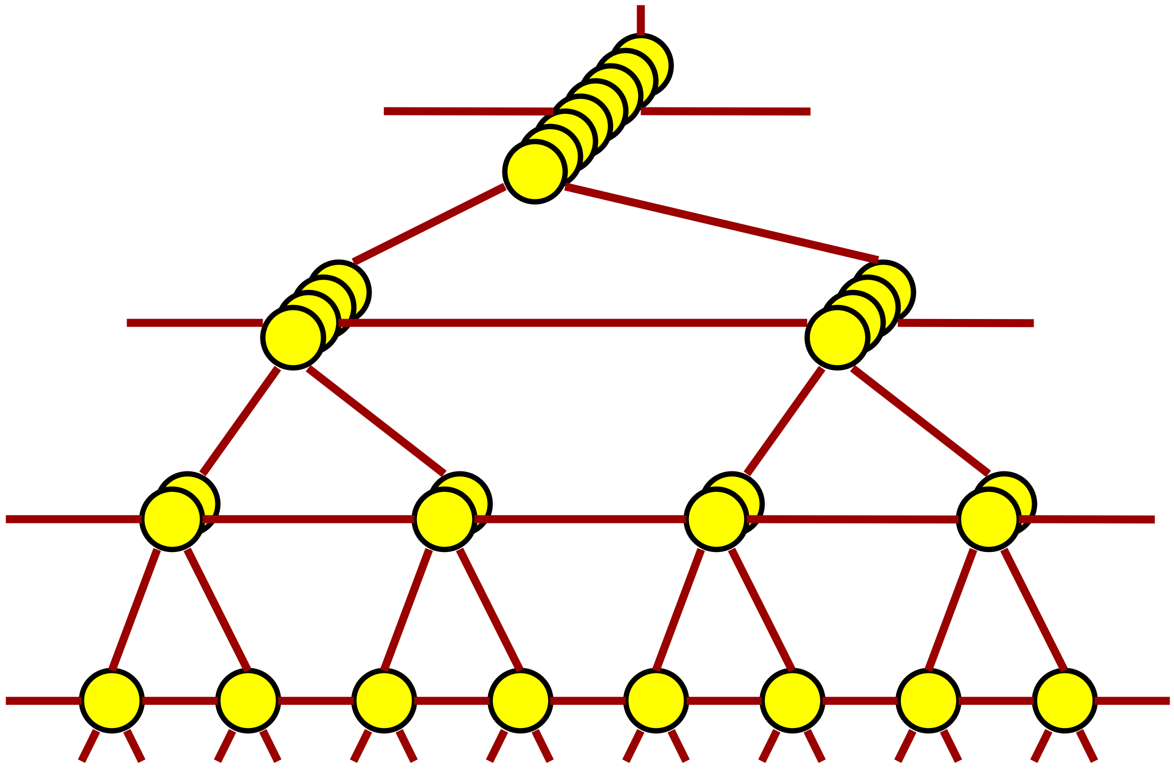
- 1D: HMM, Kalman Filter, Particle Filter
- 2D: Markov Random Fields
- Decision for level of description problematic
- Ignores vertical dependencies, flat models do not scale

Modeling Vertical Dependencies



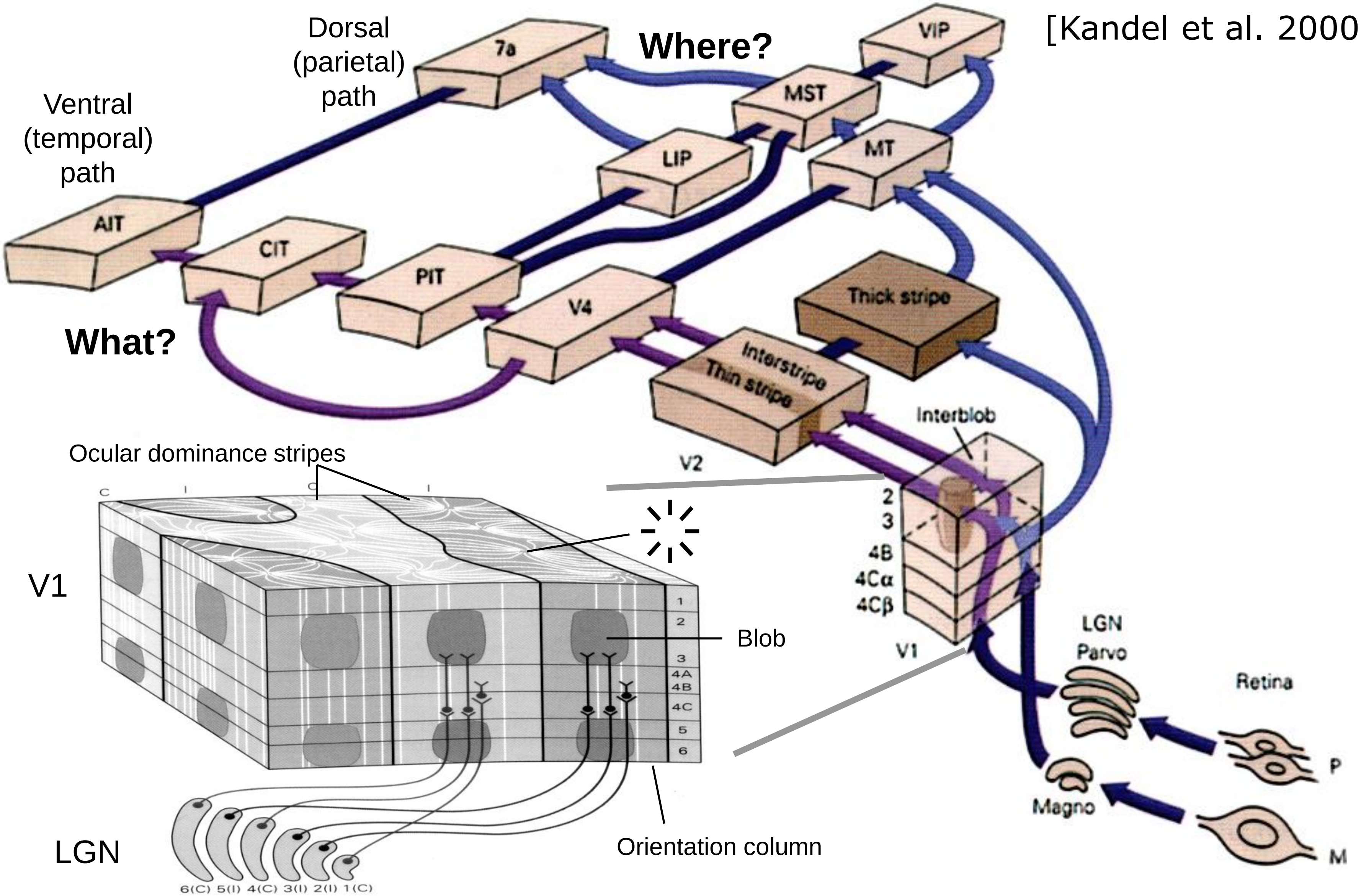
- Structure graphs, etc.
- Ignores horizontal dependencies

Horizontal and vertical Dependencies

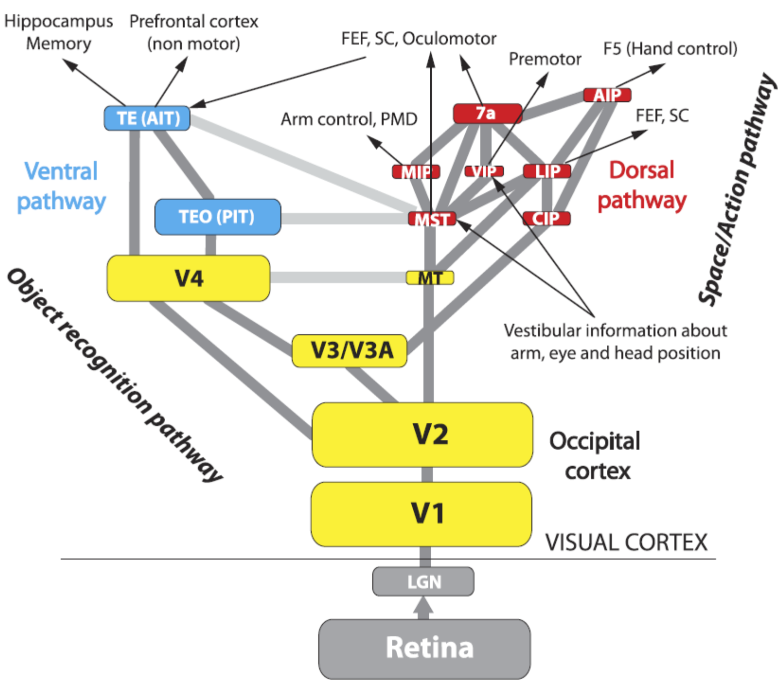


- Problem: Cycles make exact inference impossible
- Idea: Use approximate inference

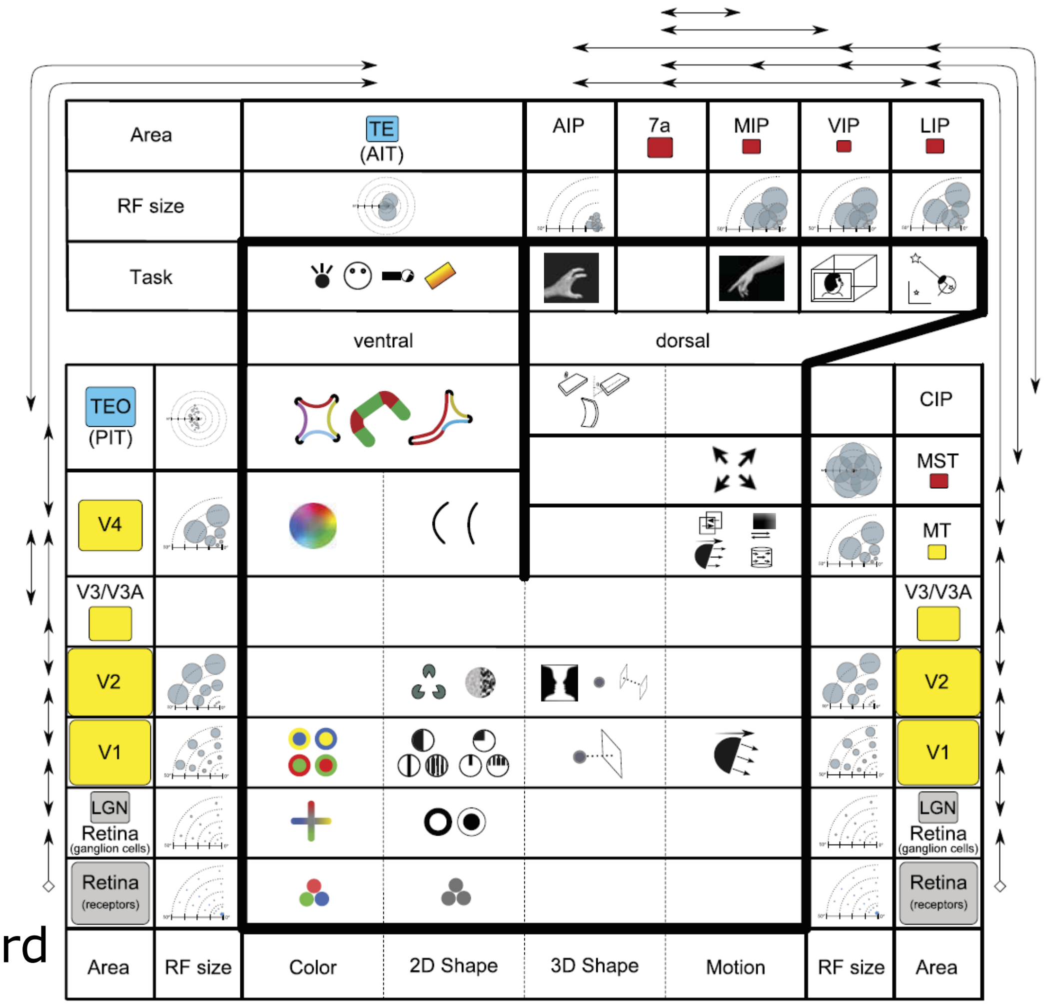
Human Visual System



Visual Processing Hierarchy



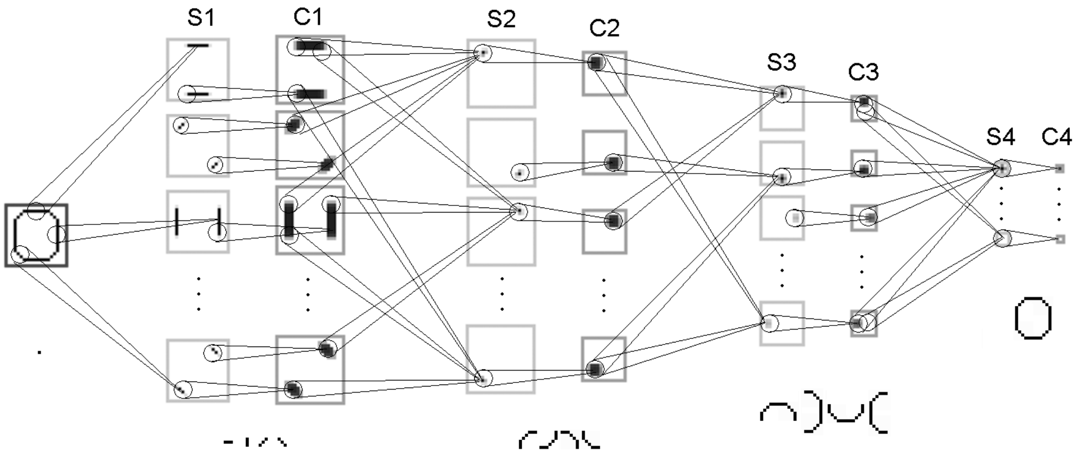
- Increasing complexity
- Increasing invariance
- All connections bidirectional
- More feedback than feed forward
- Lateral connections important



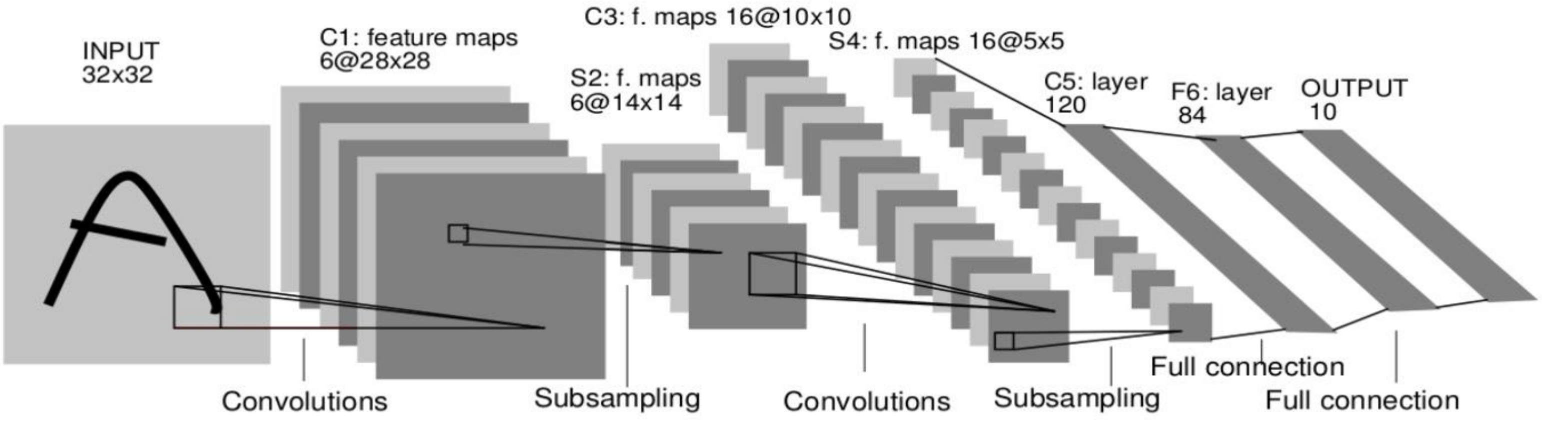
[Krüger et al., TPAMI 2013]

Feed-Forward Models

- Neocognitron: Fukushima 1980

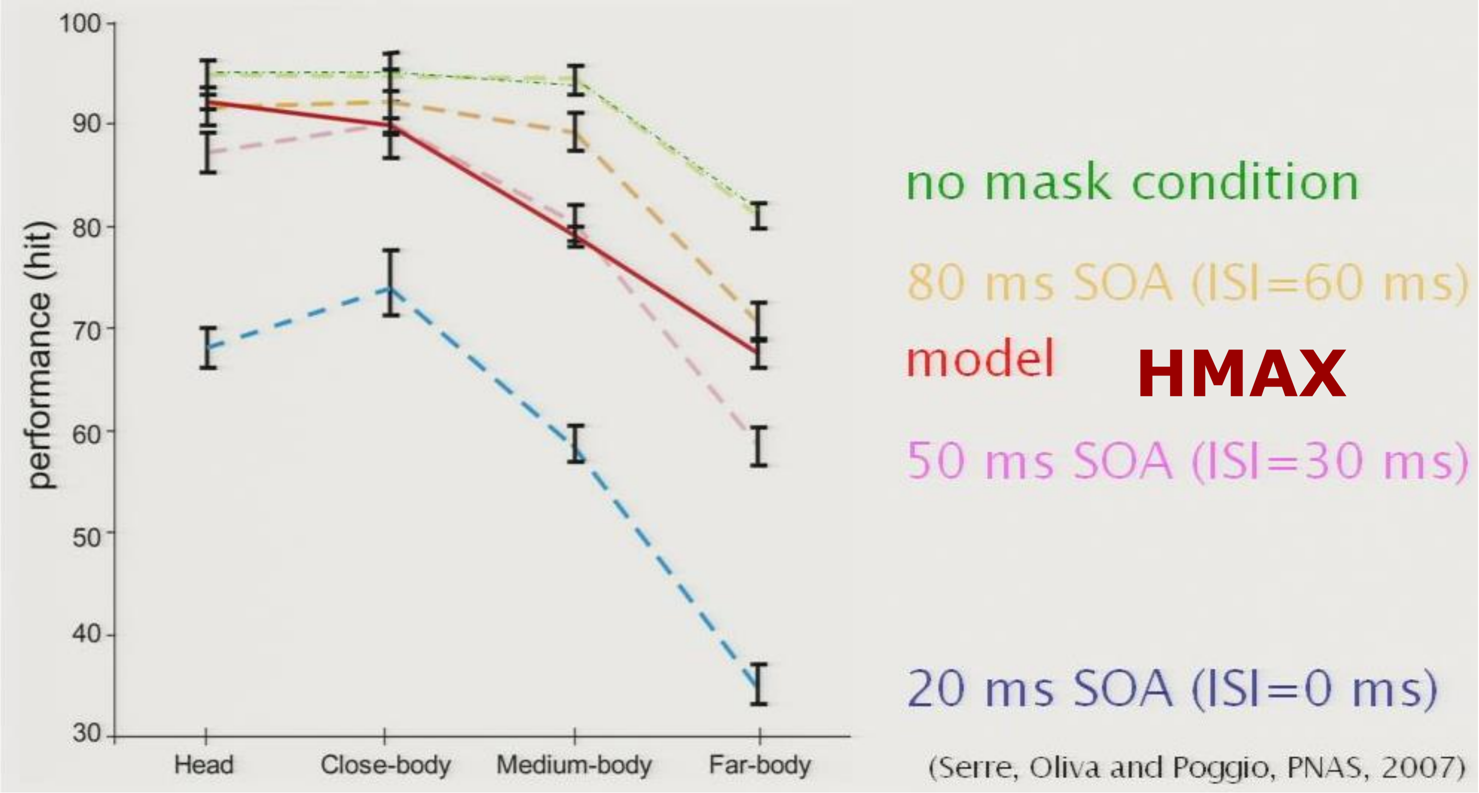


- Supervised training of convolutional networks: LeCun 1989



Feed-forward Models Cannot Explain Human Performance

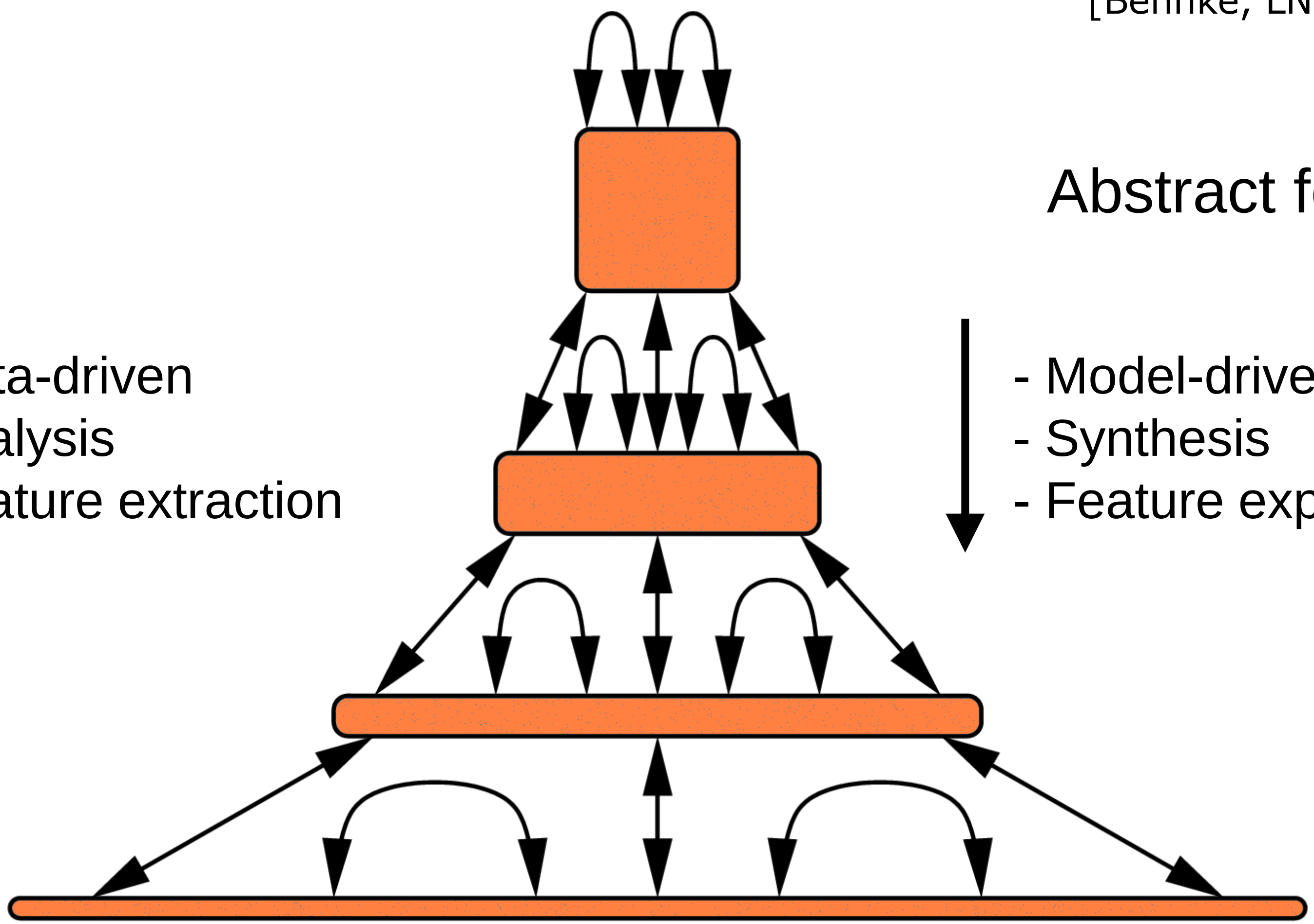
- Performance increases with observation time



Neural Abstraction Pyramid

[Behnke, LNCS 2766, 2003]

Abstract features



- Data-driven
- Analysis
- Feature extraction

- Model-driven
- Synthesis
- Feature expansion

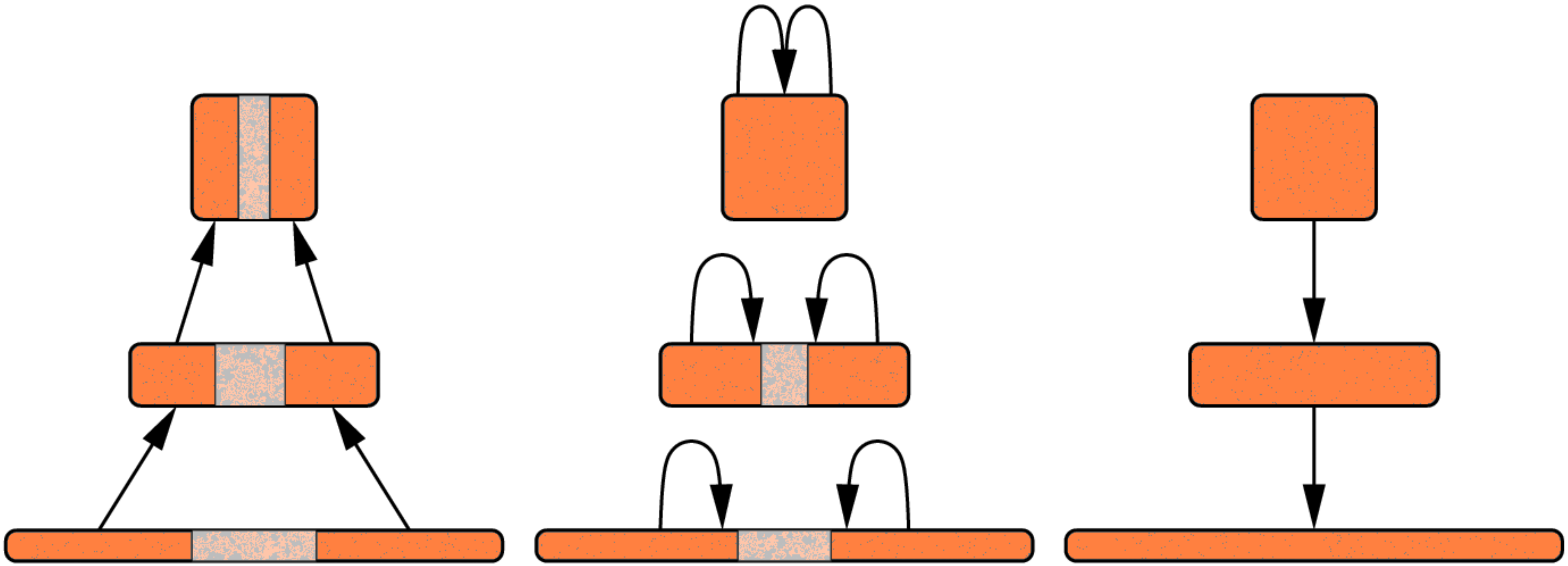
Signals

- Grouping - Competition - Completion

Iterative Interpretation

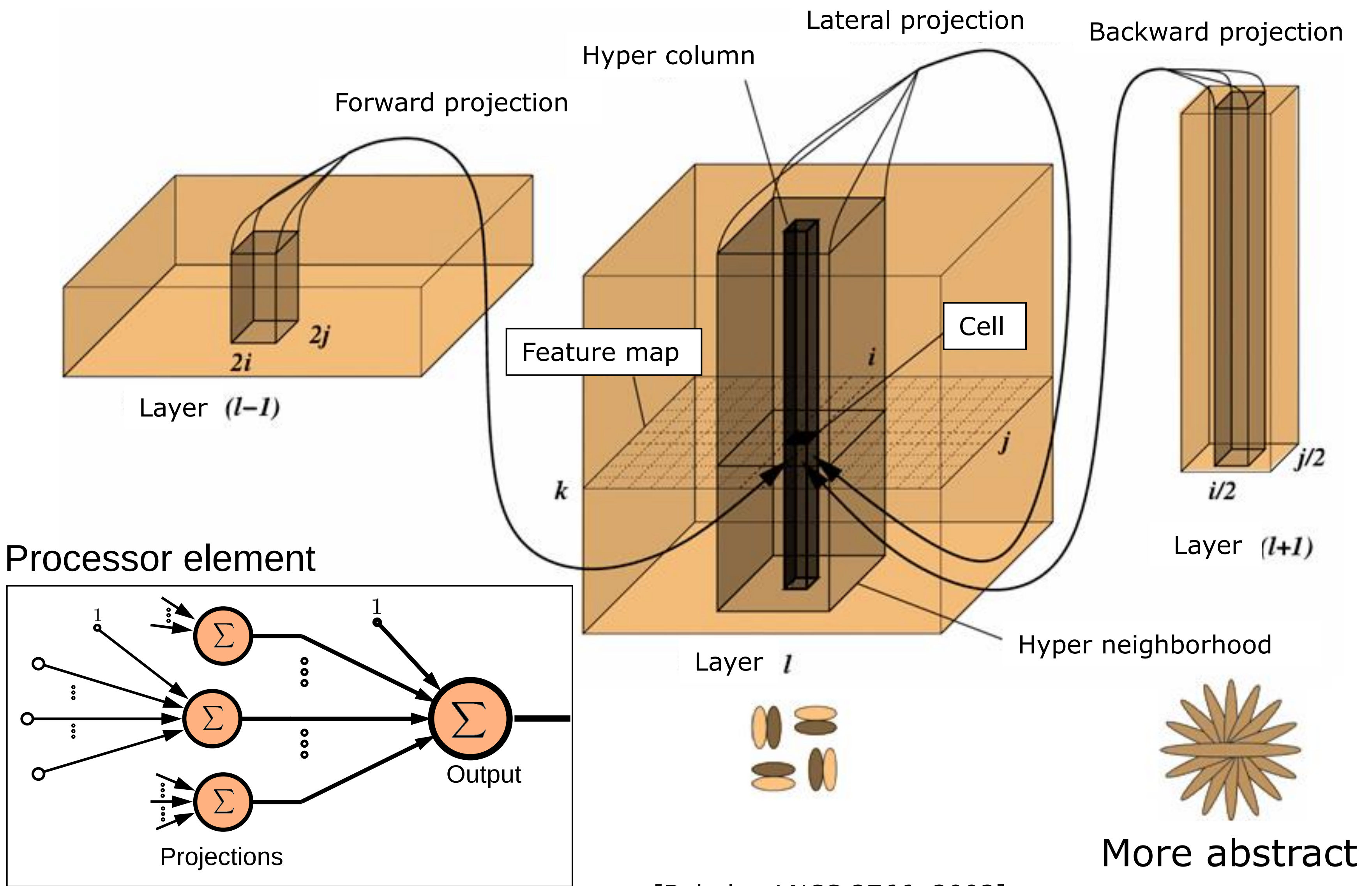
[Behnke, LNCS 2766, 2003]

- Interpret most obvious parts first



- Use partial interpretation as context to resolve local ambiguities

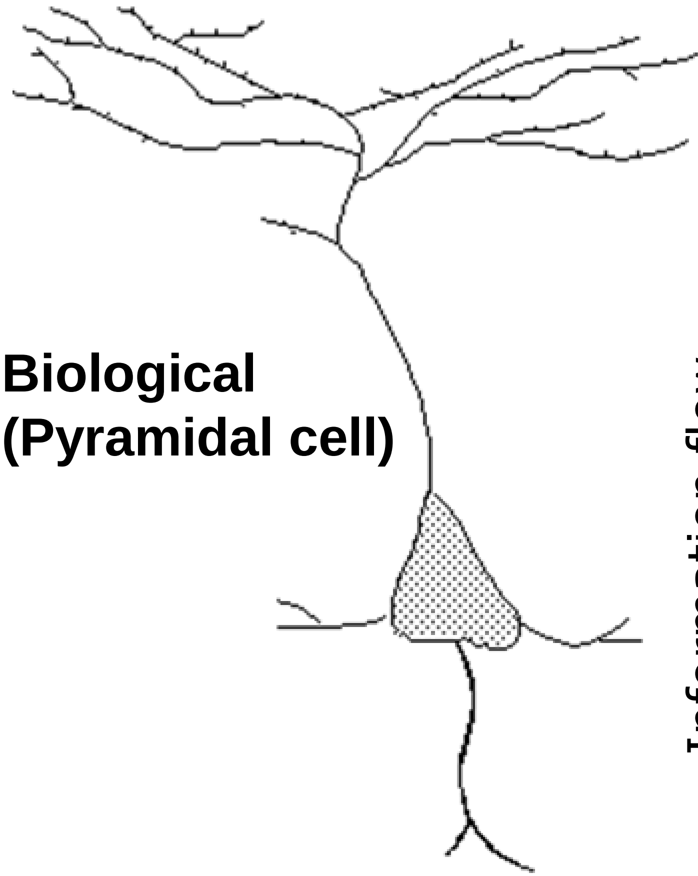
Local Recurrent Connectivity



[Behnke, LNCS 2766, 2003]

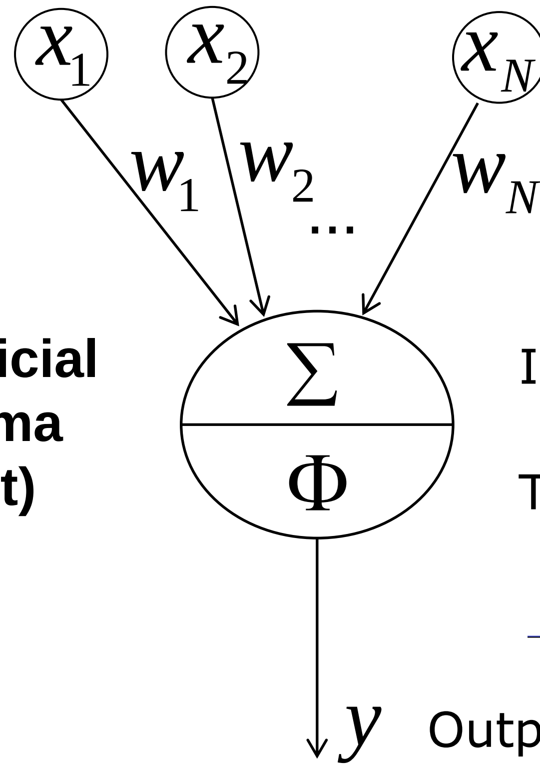
Biological vs. Artificial Neurons

Biological (Pyramidal cell)



Information flow

Artificial (Sigma Unit)

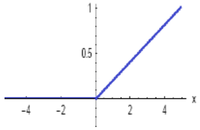


Inputs

Weights

Integration function

Transfer function



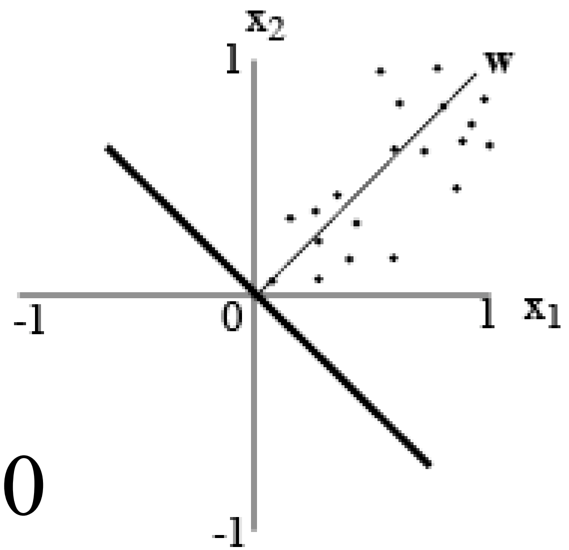
Output

$$y = \Phi\left(\sum_{j}^{N} w_j x_j\right)$$

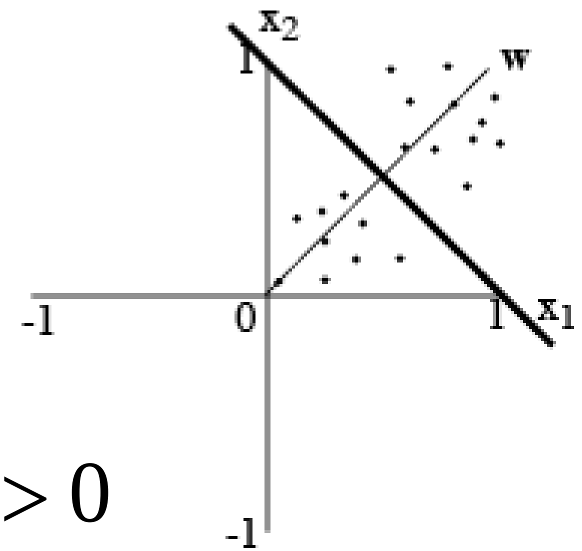
Separation of Input Patterns

- Dot product $\mathbf{w} \cdot \mathbf{x}$ separates the input space into two regions: one with value >=0 and one with value <0
- Separation is a line, defined by the weights and bias ϑ

$$w_1 x_1 + w_2 x_2 - \vartheta = 0 \qquad x_2 = \frac{\vartheta}{w_2} - \frac{w_1}{w_2} x_1$$

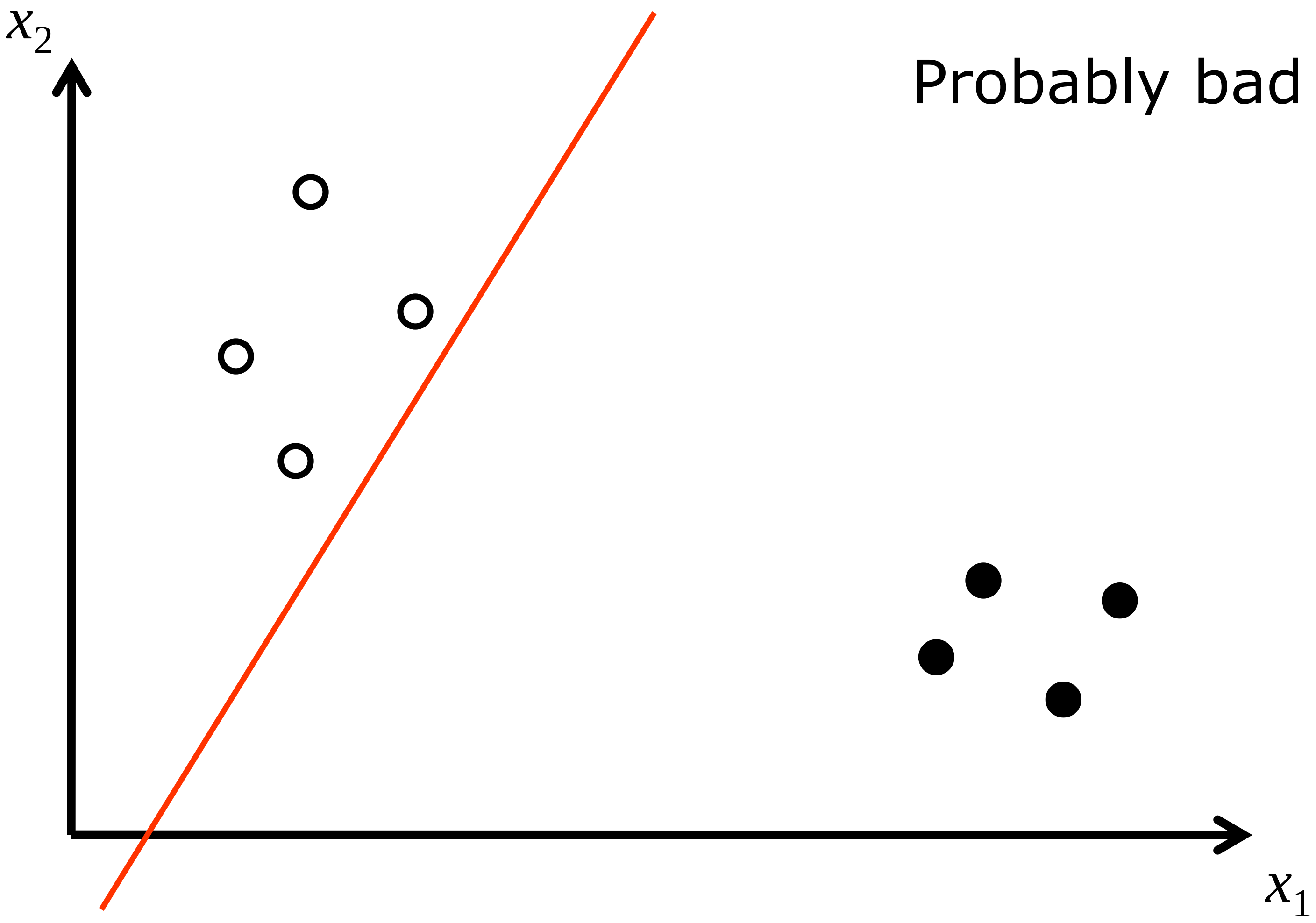


$\vartheta = 0$



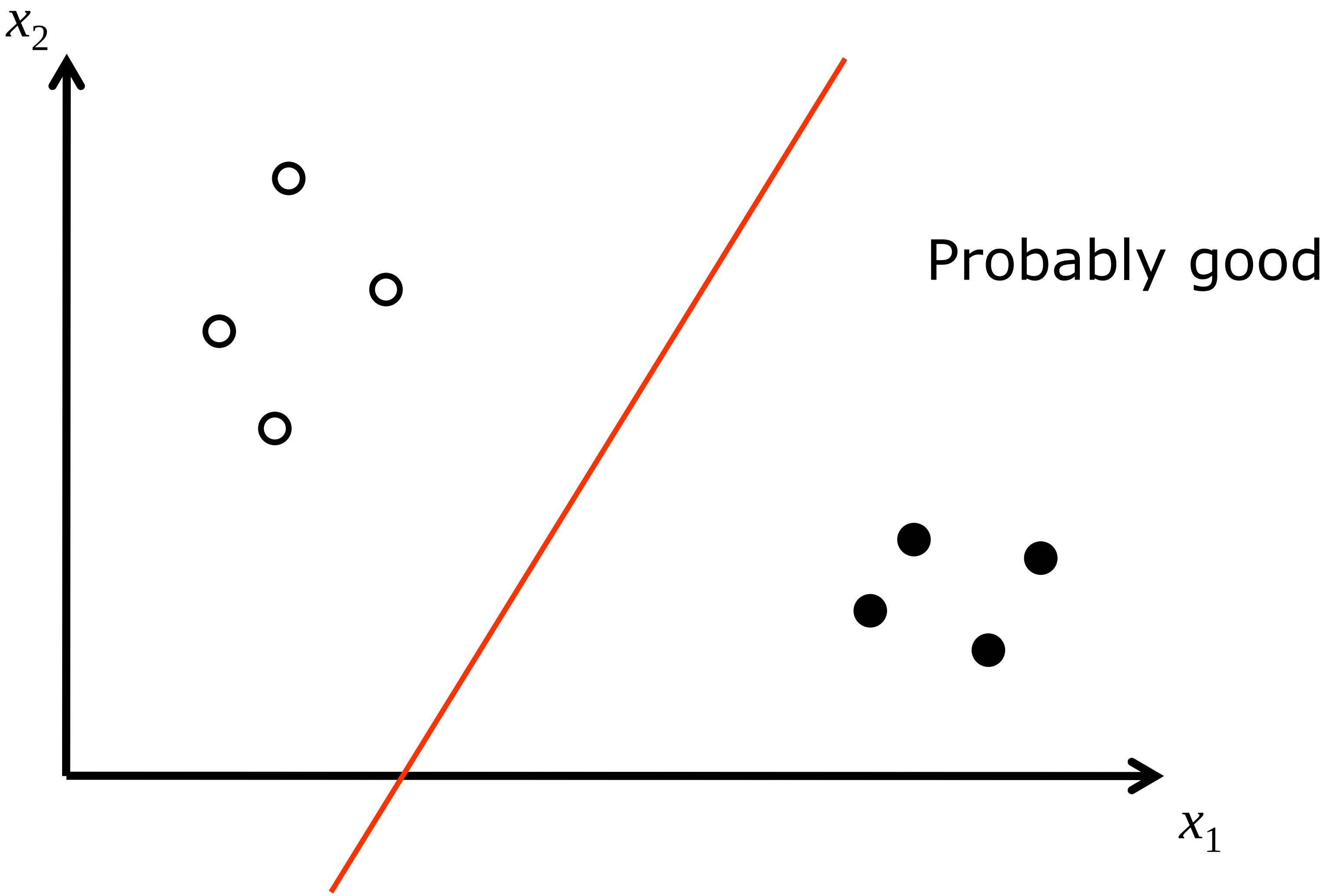
$\vartheta > 0$

Generalization

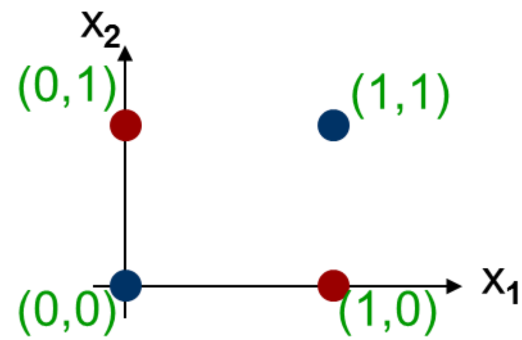


Probably bad

Generalization

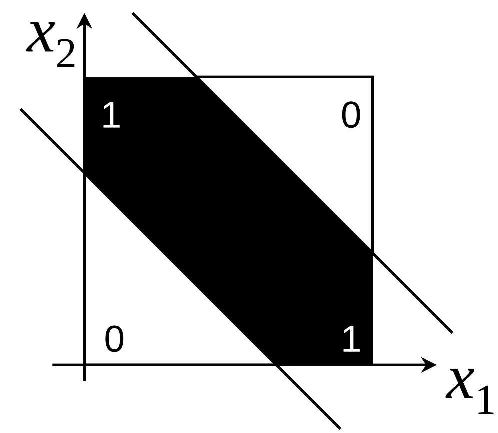


XOR Problem

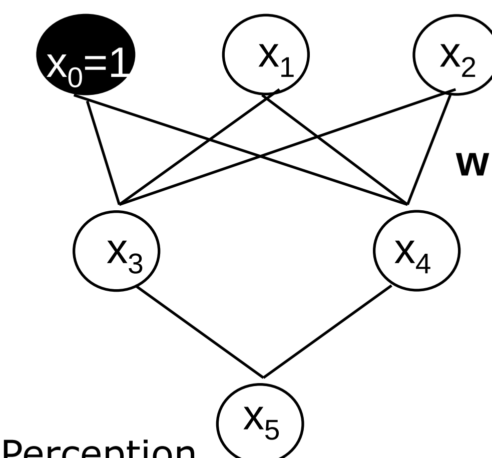


- Boolean XOR function is not linearly separable

- If we could use two hyper planes, we could separate one class from both sides



- This can be accomplished by a Multi-Layer Perceptron



- Problem:
  How to train multiple layers?

Backpropagation of Error

- Forward propagation of activity
- Backward propagation of error gradient
- Weight update by gradient descent

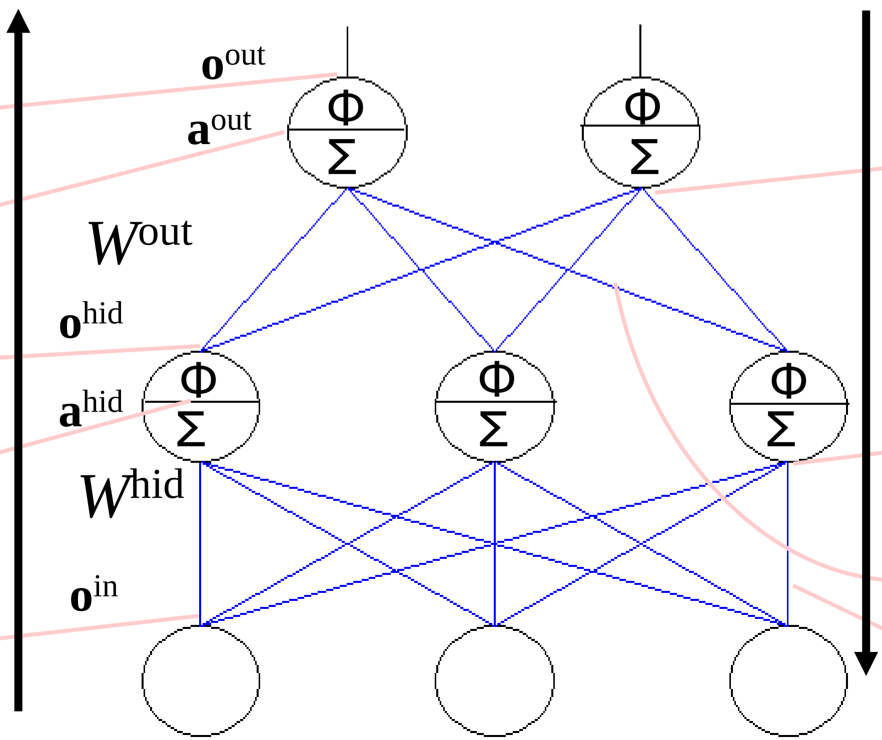
$$5)\ o_k^{out} = \Phi(a_k^{out})$$

$$4)\ a_k^{out} = \sum_{j=1}^{m} w_{jk}^{out} o_j^{hid}$$

$$3)\ o_j^{hid} = \Phi(a_j^{hid})$$

$$2)\ a_j^{hid} = \sum_{i=1}^{l} w_{ij}^{hid} o_i^{in}$$

$$1)\ o_i^{in} = x_i$$



Error

$$6)\ \delta_k^{\mathrm{out}} = \Phi'\left(a_k^{\mathrm{out}}\right)\left(o_k^{\mathrm{out}} - t_k\right)$$

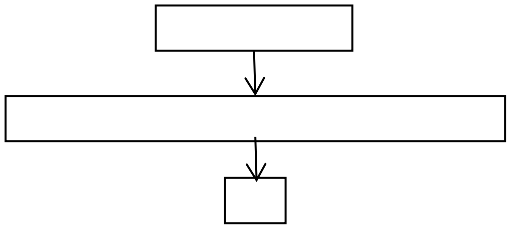
$$7)\ \delta_j^{\mathrm{hid}} = \Phi'\left(a_j^{\mathrm{hid}}\right)\sum_{k=1}^{n} w_{jk}^{\mathrm{out}} \delta_k^{\mathrm{out}}$$

$$8)\ \Delta w_{jk}^{out} = -\eta o_j \delta_k;$$

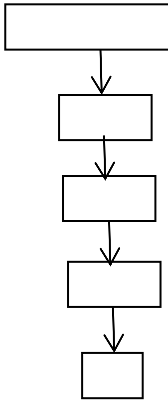
$$\Delta w_{ij}^{hid} = -\eta o_i \delta_j;$$

Flat vs. Deep Networks

- A neural network with a **single hidden layer** that is wide enough can compute any function (Cybenko, 1989)
  - Certain functions, like parity, may require **exponentially** many hidden units (in the number of inputs)

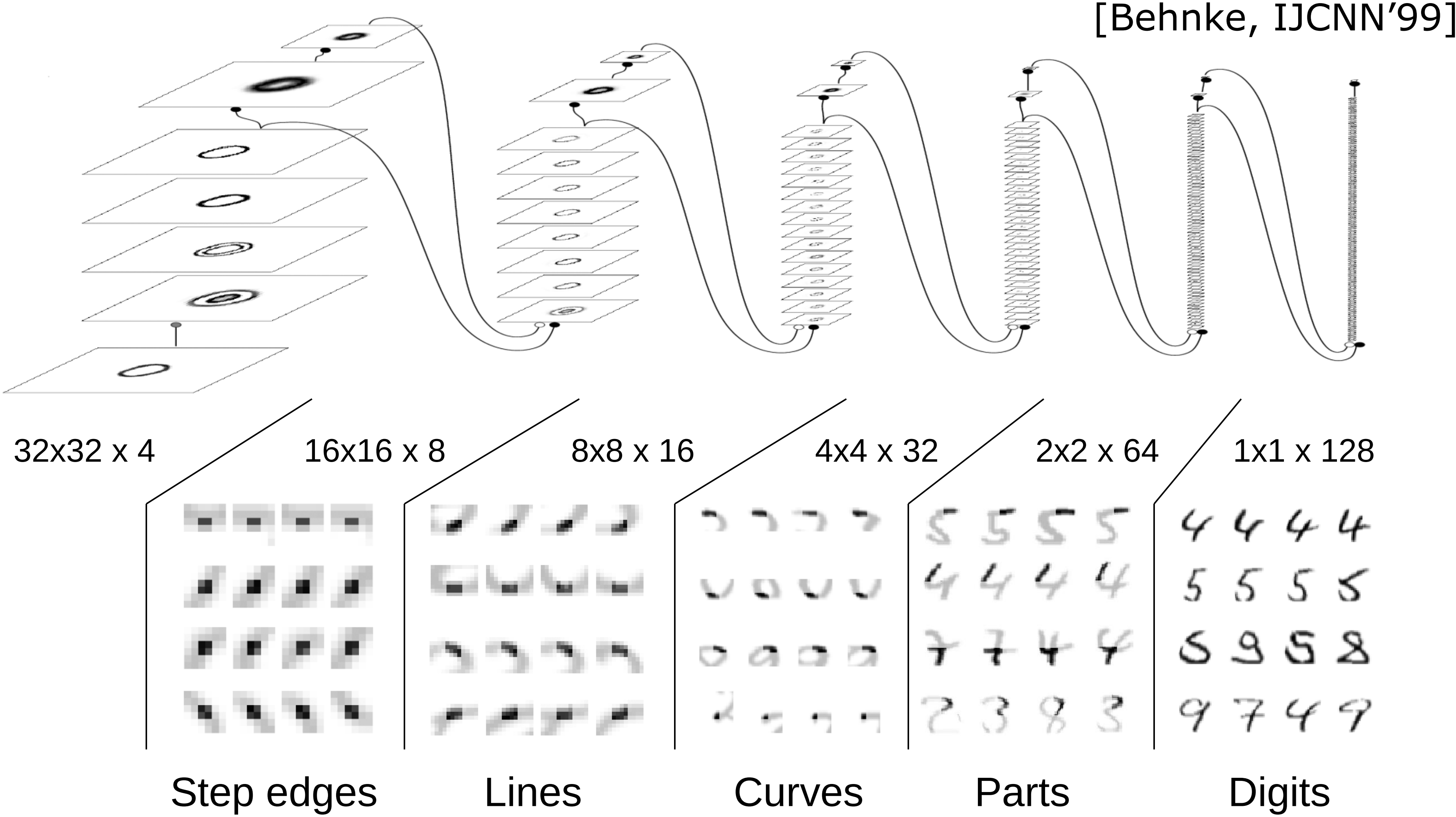


- **Deep networks** (with multiple hidden layers) may be exponentially more efficient
  - Parity example: Compute carry bit sequentially



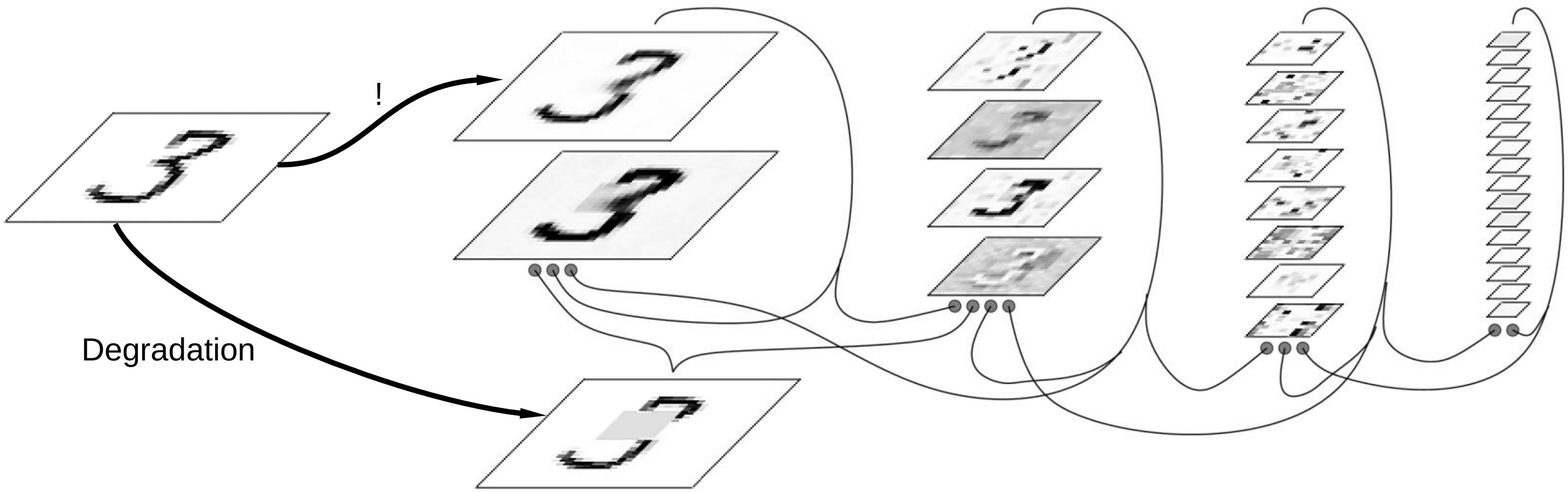
Learning a Feature Hierarchy

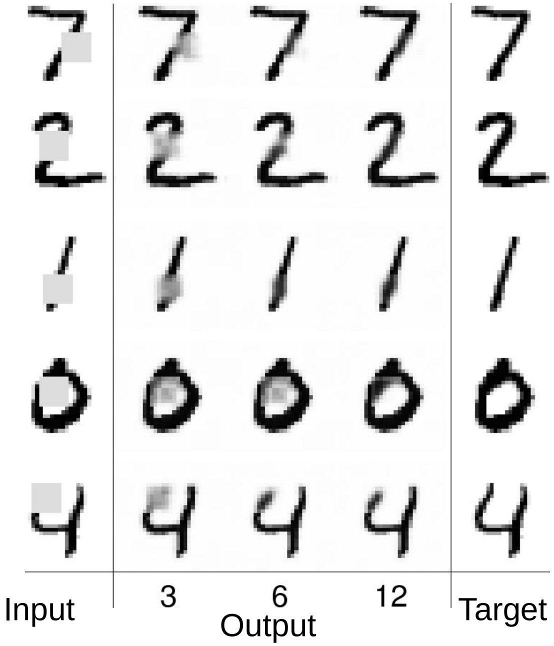
[Behnke, IJCNN'99]

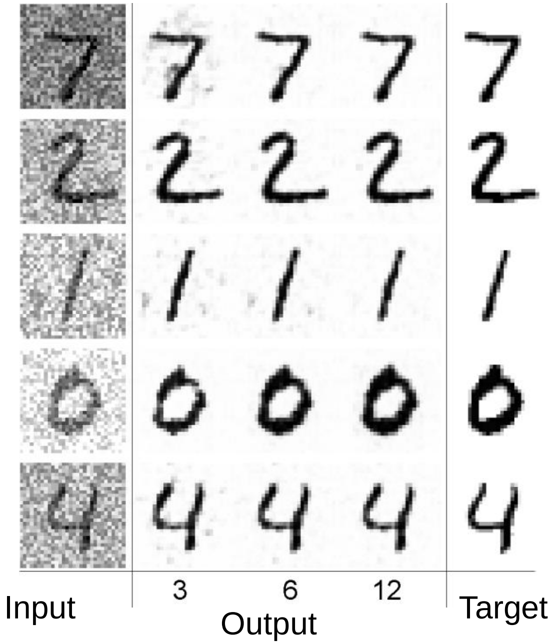


Digit Reconstruction

[Behnke, IJCAI'01]

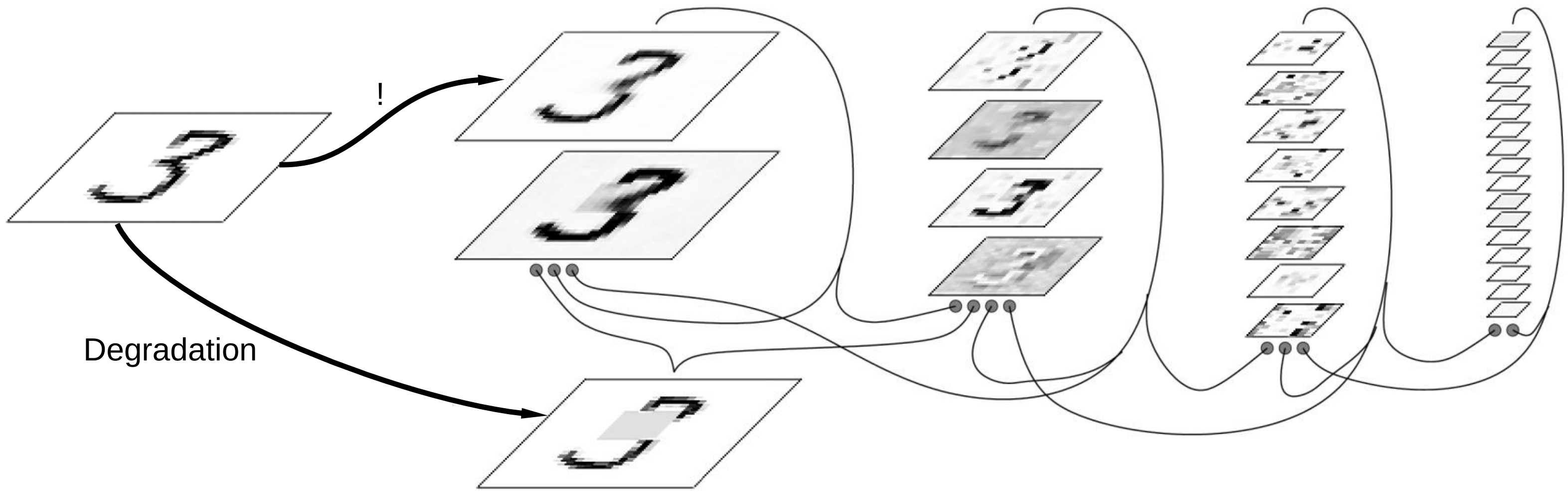


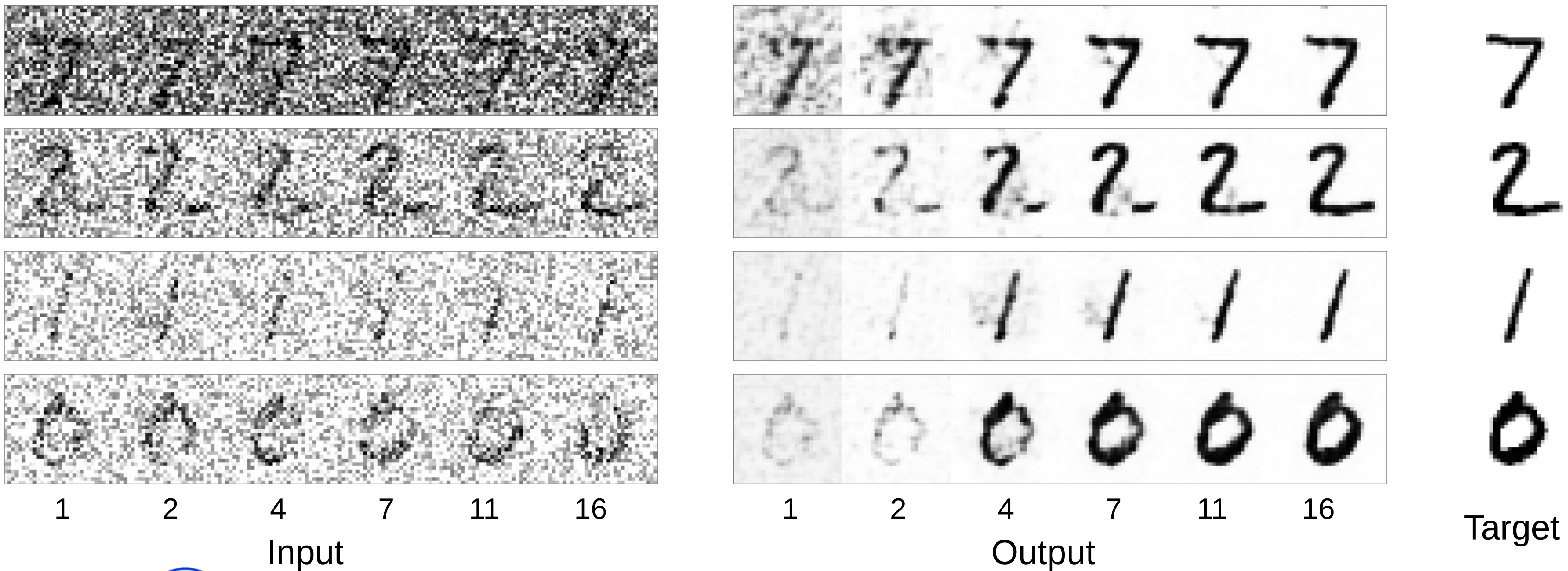




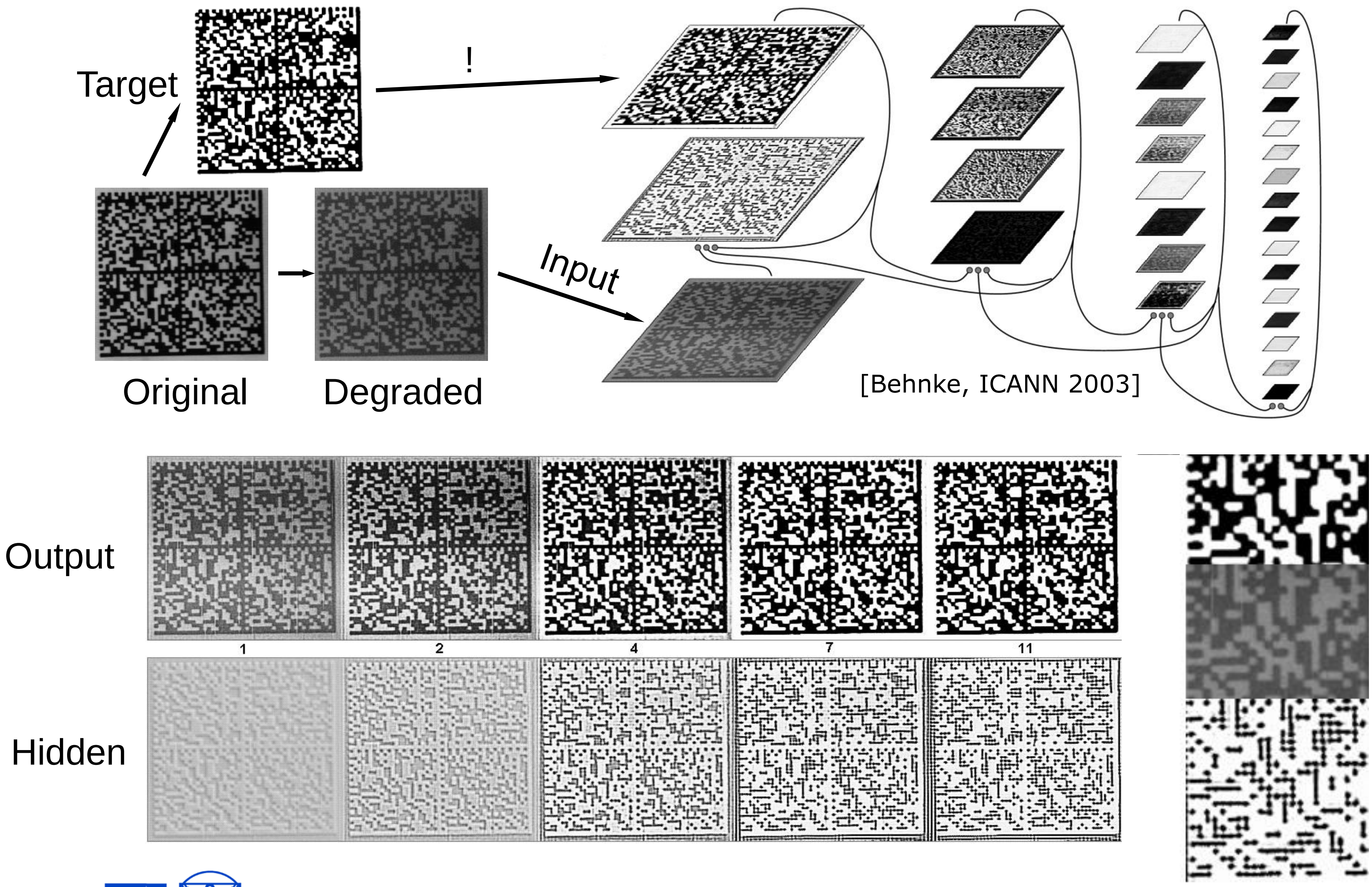
Digit Reconstruction

[Behnke, IJCAI'01]





Binarization of Matrix Codes



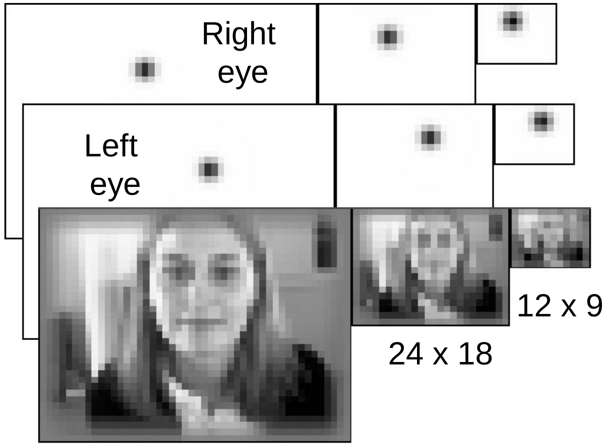
Face Localization

[Behnke, KES'03]

- BioID data set:
  - 1521 images
  - 23 persons

- Encode eye positions with blobs

384 x 288

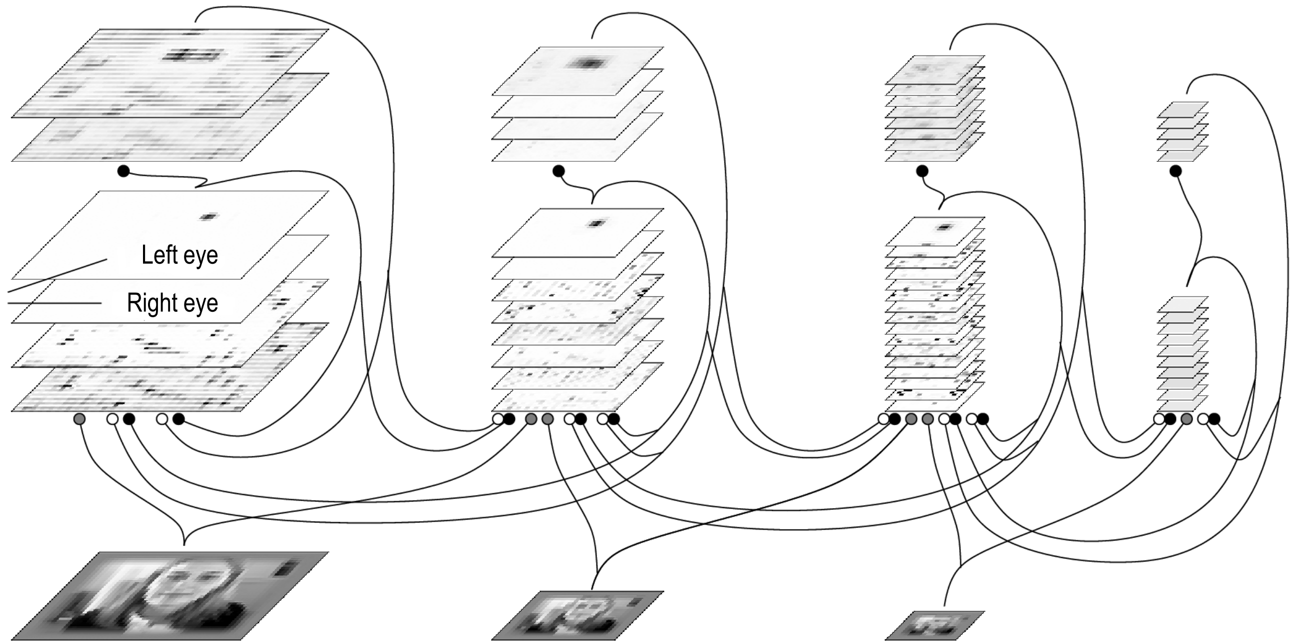


Face Localization

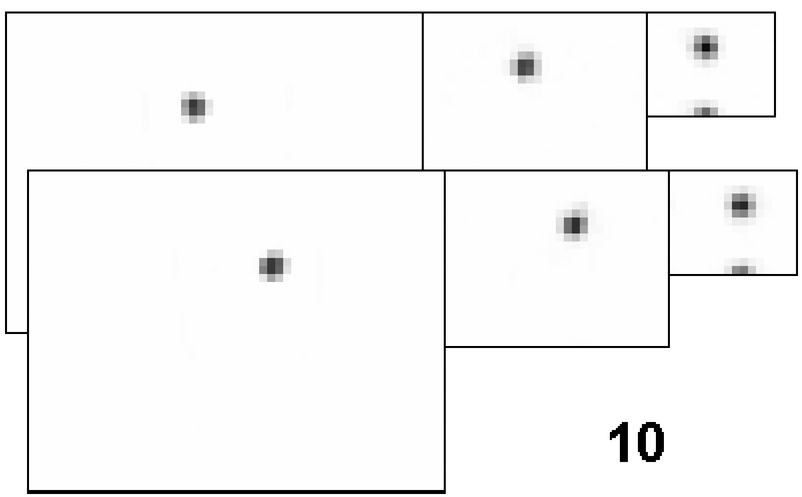
[Behnke, KES’03]

Output

Input

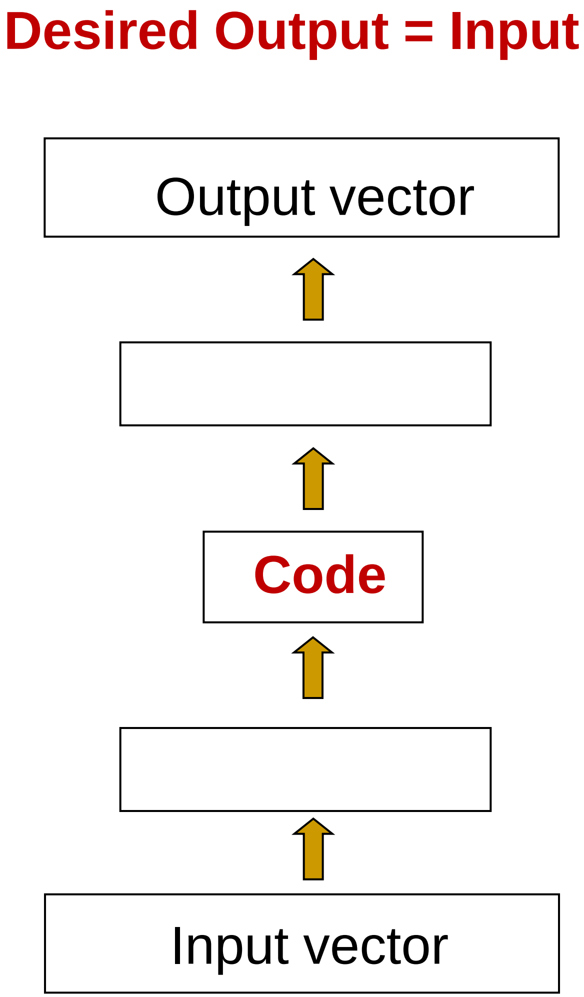


Output



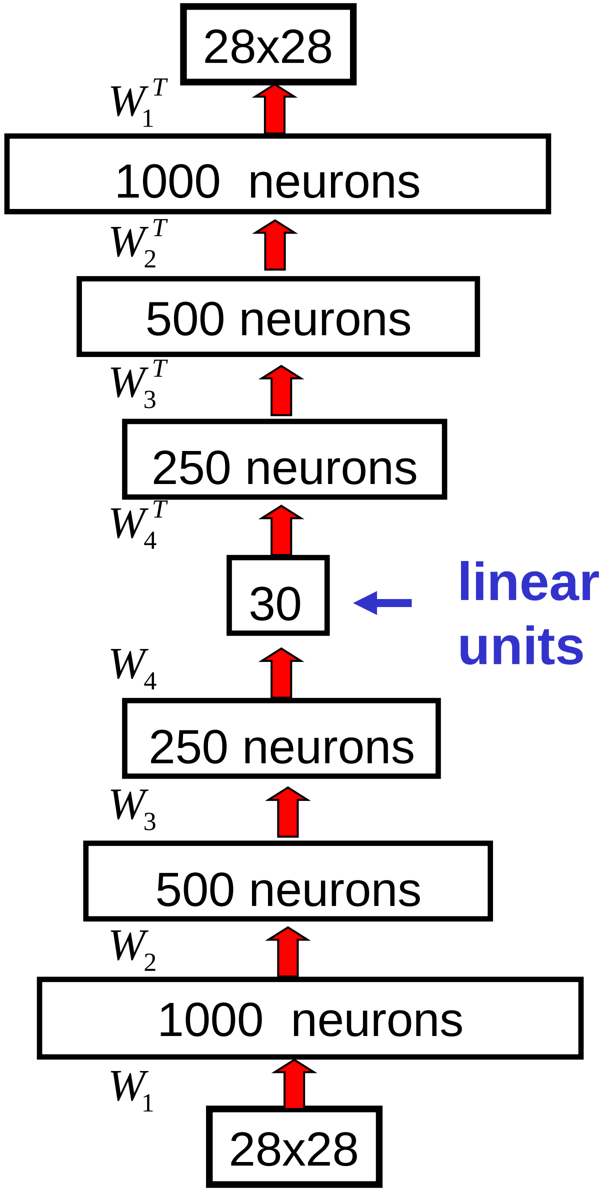
Auto-Encoder

- Try to push input through a bottleneck
- Activities of hidden units form an efficient code
  - There is no space for redundancy in the bottleneck
- Extracts frequently independent features (factorial code)



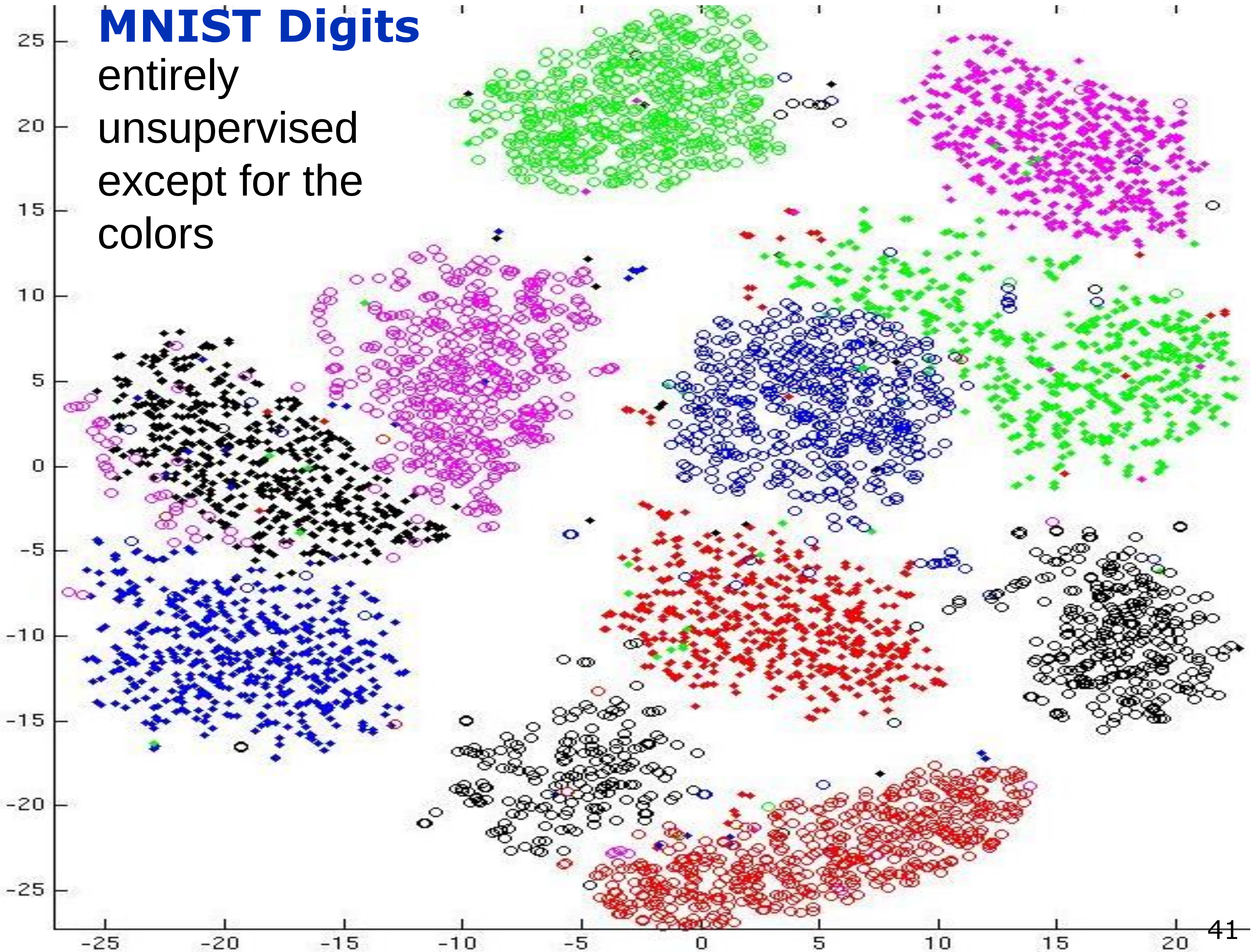
Deep Autoencoders
(Hinton & Salakhutdinov, 2006)

- Multi-layer autoencoders for non-linear dimensionality reduction
- Difficult to optimize deep autoencoders using backpropagation
- Greedy, layer wise training
- Unrolling
- Supervised fine-tuning



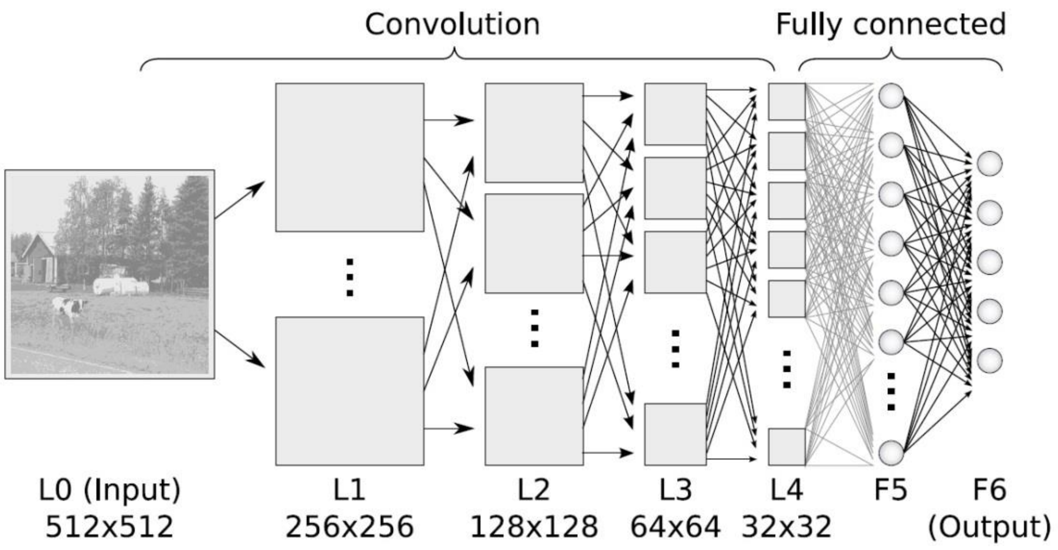
MNIST Digits

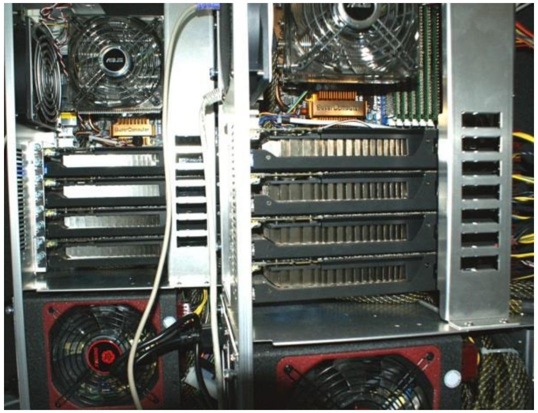
entirely unsupervised except for the colors

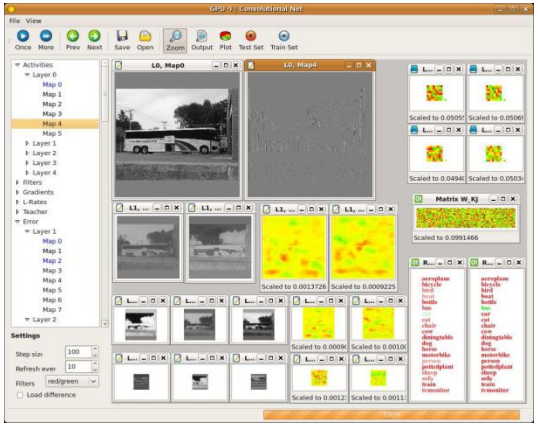


GPU Implementations (CUDA)

- Affordable parallel computers
- General-purpose programming
- Convolutional [Scherer & Behnke, 2009]



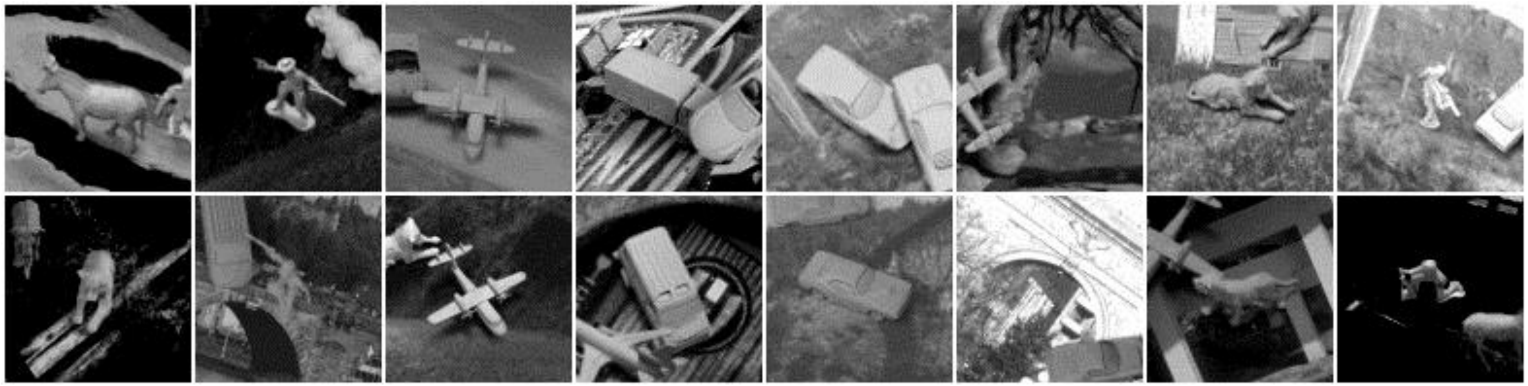




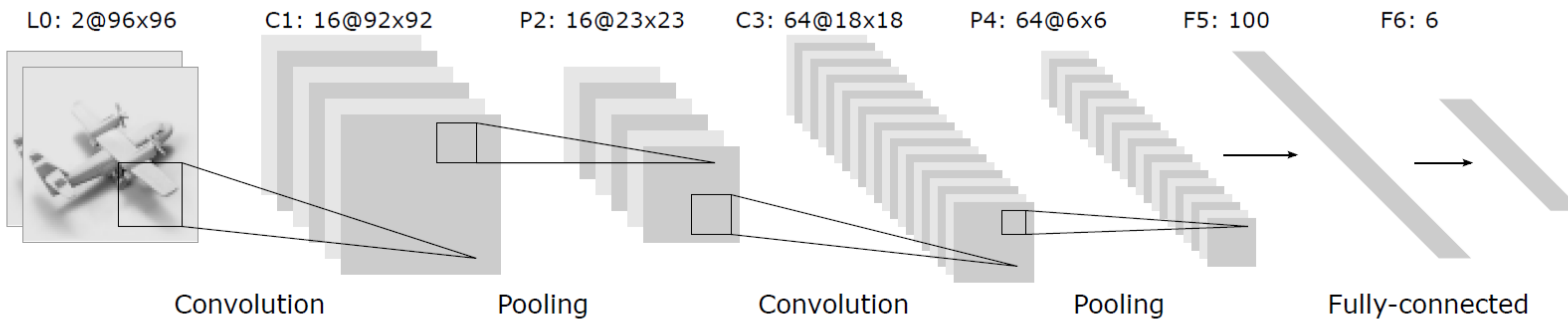
- Local connectivity [Uetz & Behnke, 2009]

Image Categorization: NORB

- 10 categories, jittered-cluttered



- **Max-Pooling**, cross-entropy training

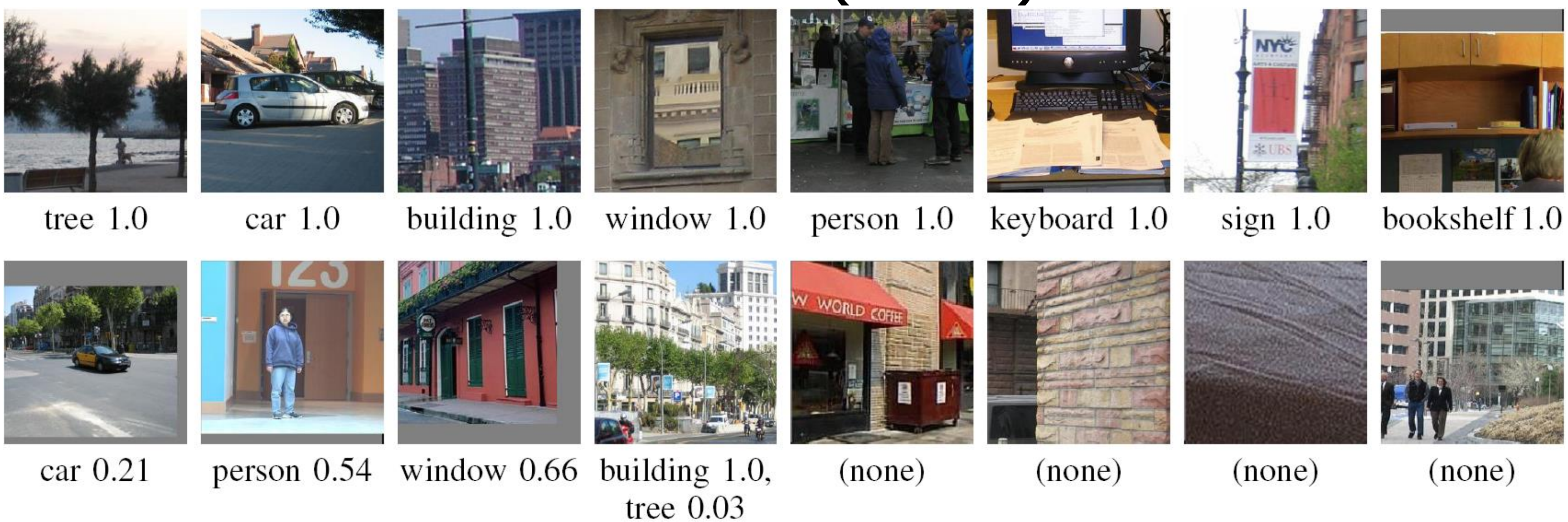


- Test error: 5,6% (LeNet7: 7.8%)

[Scherer, Müller, Behnke, ICANN'10]

Image Categorization: LabelMe

- 50,000 color images (256x256)
- 12 classes + clutter (50%)

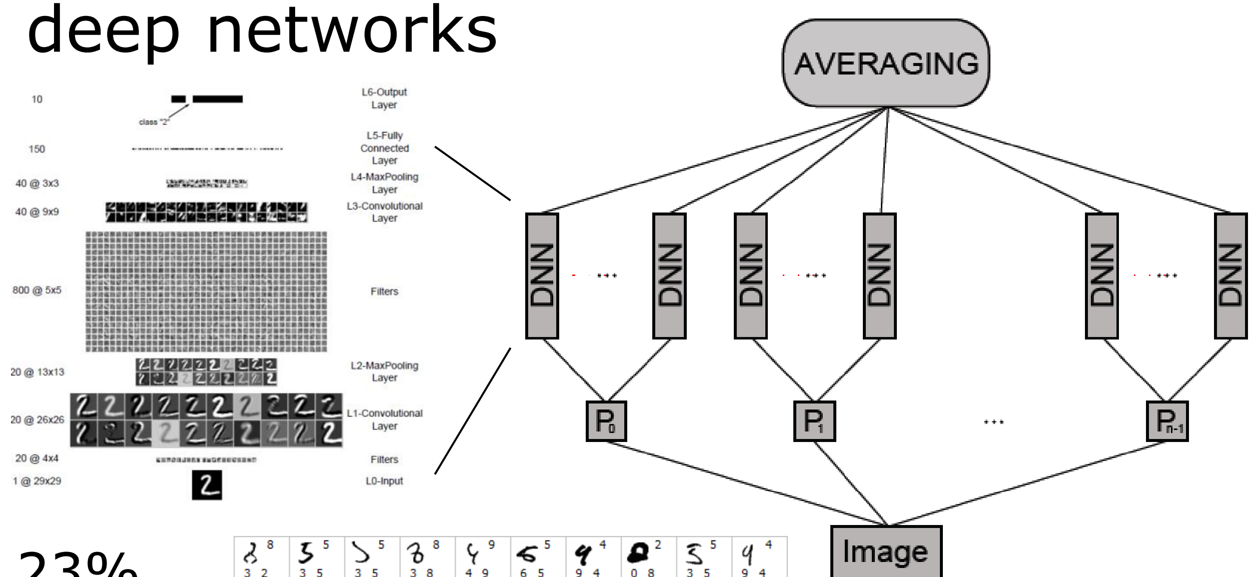


- Error TRN: 3.77%; TST: 16.27%
- Recall: 1,356 images/s

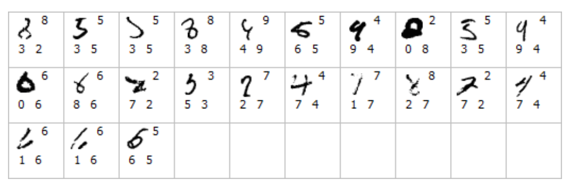
[Uetz, Behnke, ICIS2009]

Multi-Column Deep Convolutional Networks

- Different preprocessings
- Trained with distortions
- Bagging deep networks



- MNIST: 0.23%
- NORB: 2.7%
- CIFAR10: 11.2%
- Traffic signs: 0.54% test error



[Ciresan et al. CVPR 2012]

ImageNet Challenge

- 1.2 million images
- 1000 categories, no overlap
- Subset of 11 million images from 15.000+ categories
- Hierarchical category structure (WordNet)

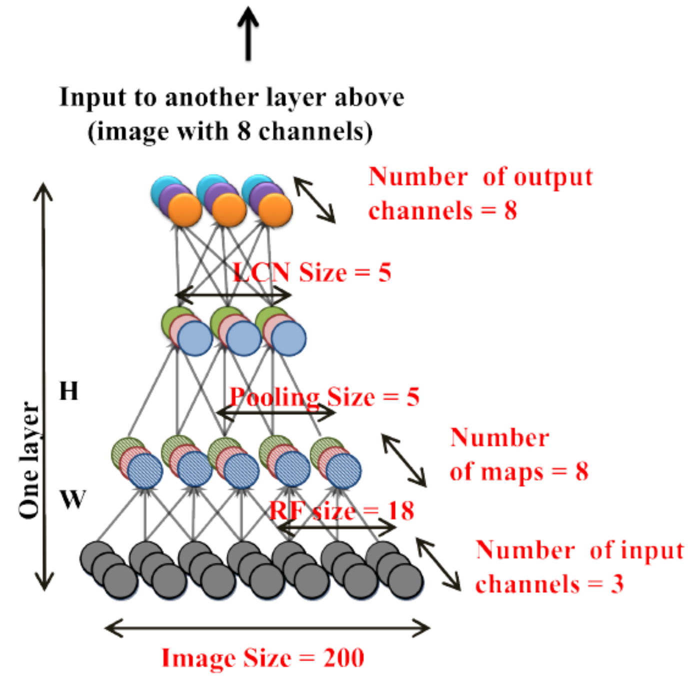
Golf cart (motor vehicle, self-propelled vehicle, wheeled vehicle, … Egyptian cat (domestic cat, domestic animal, animal)

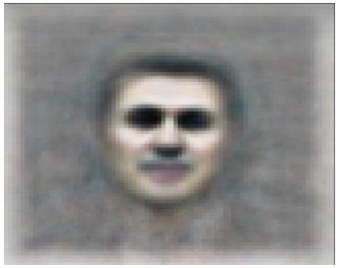
- Task: recognize object category
- Low penalty for extra detections
- Hierarchical error computation

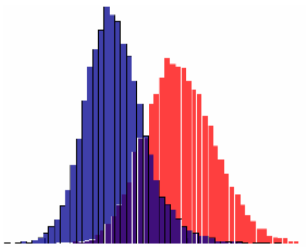
Large Unsupervised Feature Learning

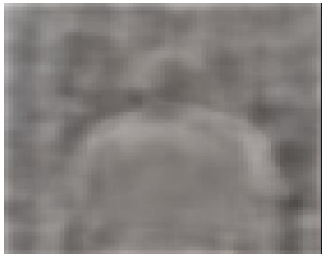
- 9 layer model
- Locally connected
- Sparse auto-encoder
- L2 pooling
- Local contrast normalization
- 1 billion connections
- Trained on 10 million images
- Unsupervised learned detectors

3x







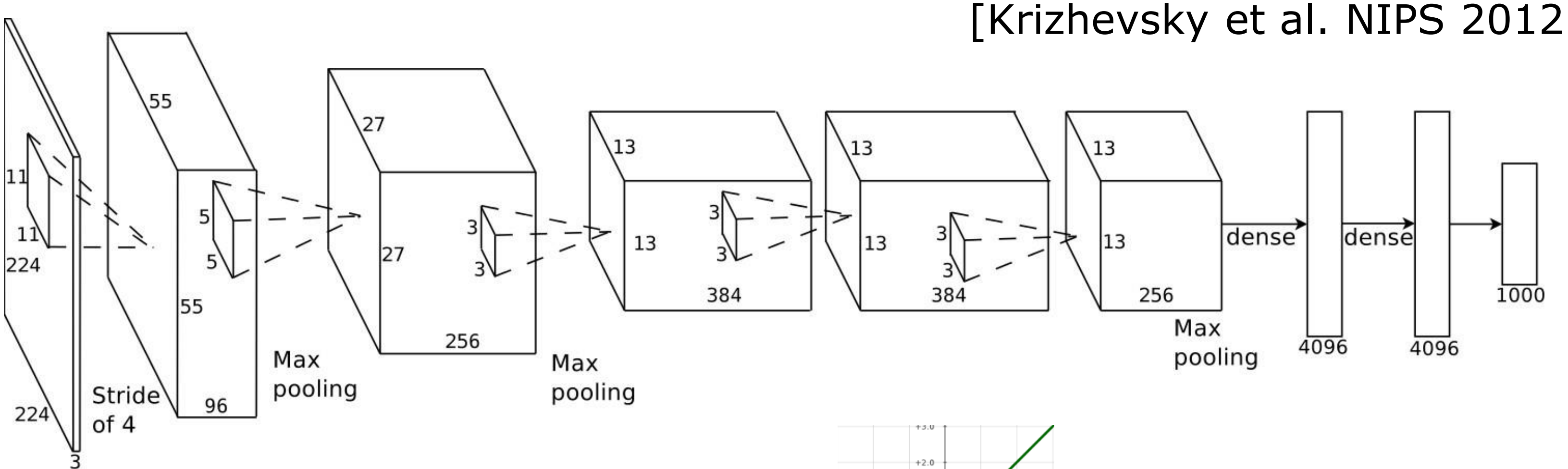


- Supervised ImageNet 2011 results (14M images, 22K categories): 15.8%

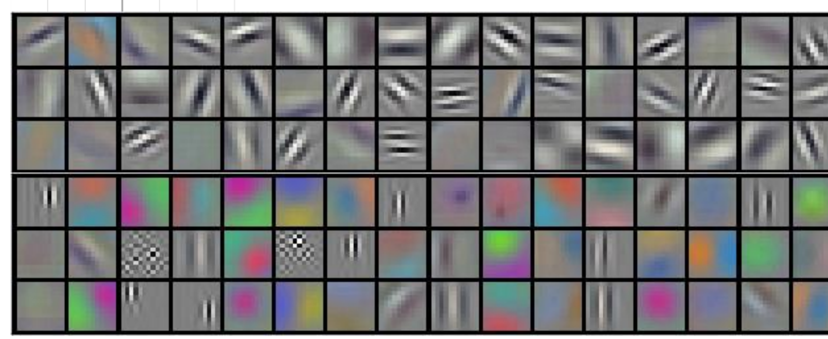
[Le et al. 2012]

Large Convolutional Network

[Krizhevsky et al. NIPS 2012]

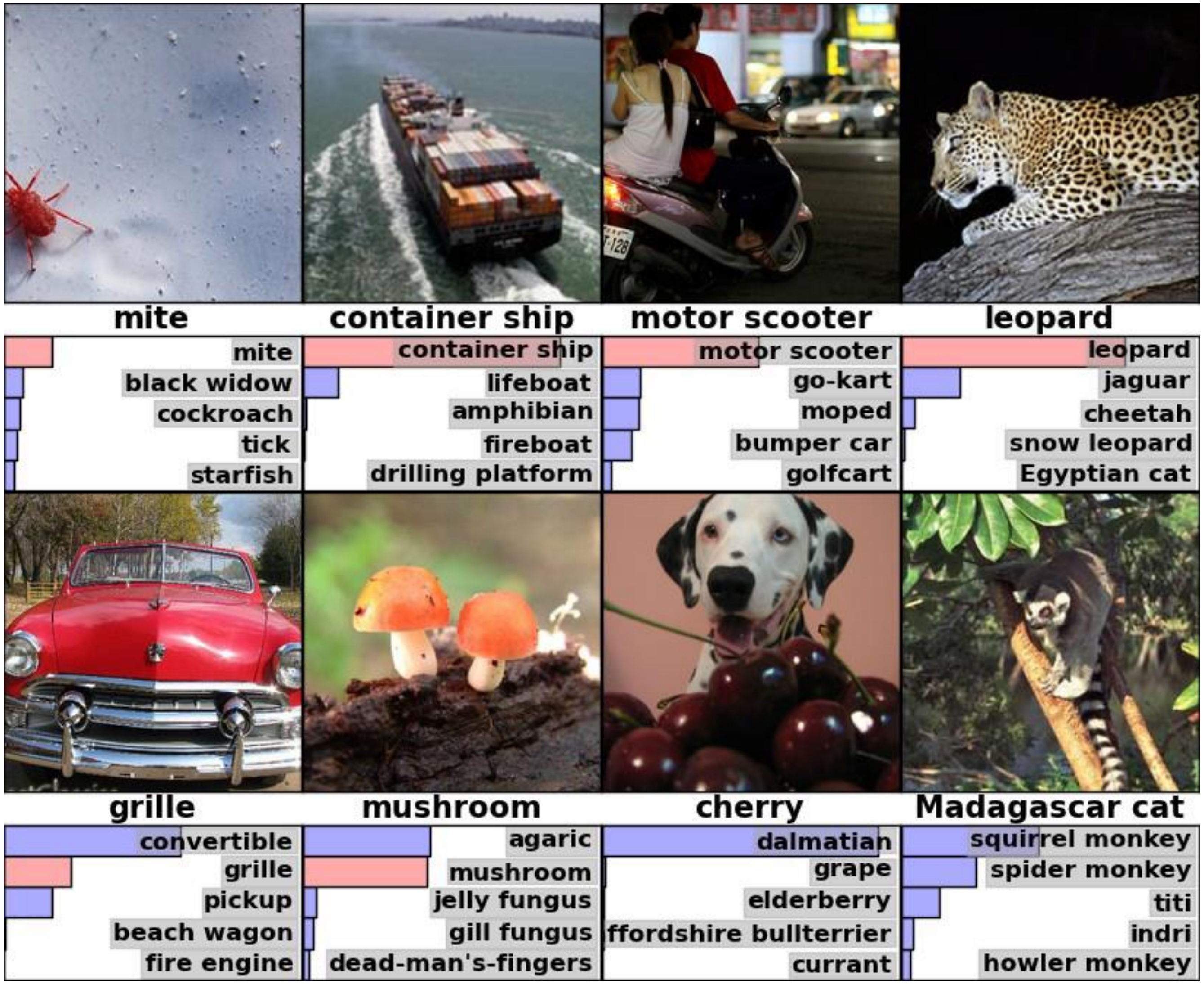


- Rectifying transfer functions
- 650,000 neurons
- 60,000,000 parameters
- 630,000,000 connections
- Trained using dropout and data augmentation
- Testing 10 sub-images
- ILSVRC-2012: top-5 error 15.3%



96 learned low-level filters

Validation Classification



[Krizhevsky et al. NIPS 2012]

Surpassing Human Performance

GT: horse cart
1: horse cart
2: minibus
3: oxcart
4: stretcher
5: half track

GT: birdhouse
1: birdhouse
2: sliding door
3: window screen
4: mailbox
5: pot

GT: forklift
1: forklift
2: garbage truck
3: tow truck
4: trailer truck
5: go-kart

GT: letter opener
1: drumstick
2: candle
3: wooden spoon
4: spatula
5: ladle

GT: coucal
1: coucal
2: indigo bunting
3: lorikeet
4: walking stick
5: custard apple

GT: komondor
1: komondor
2: patio
3: llama
4: mobile home
5: Old English sheepdog

GT: yellow lady's slipper
1: yellow lady's slipper
2: slug
3: hen-of-the-woods
4: stinkhorn
5: coral fungus

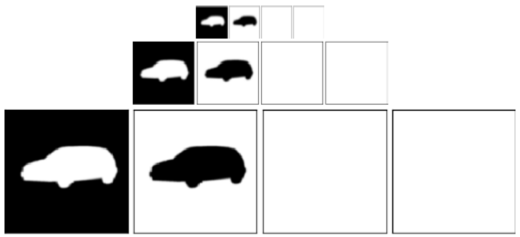
GT: spotlight
1: grand piano
2: folding chair
3: rocking chair
4: dining table
5: upright piano

[He et al. 2015]

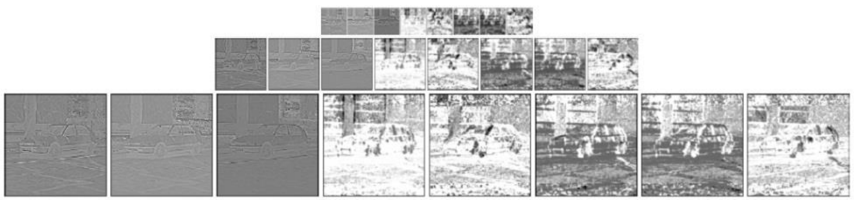
Object-class Segmentation

[Schulz, Behnke 2012]

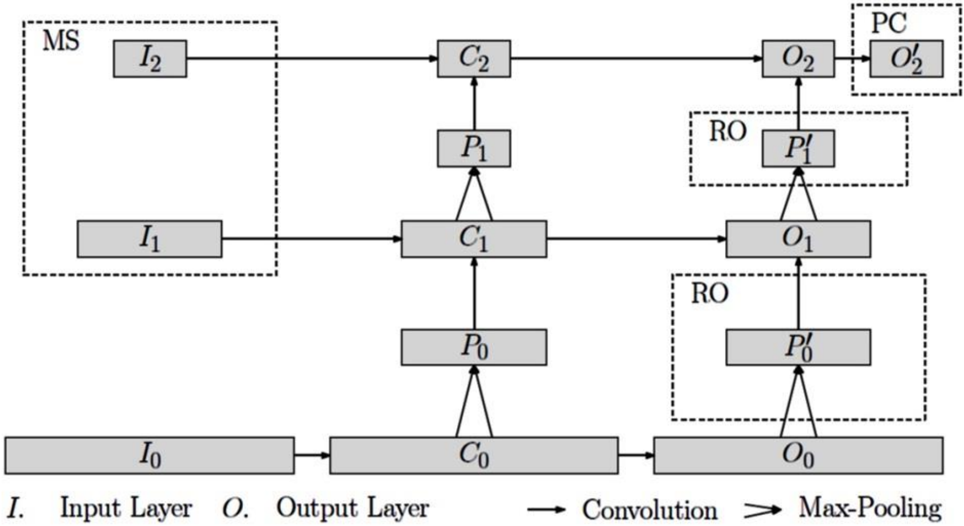
- Class annotation per pixel

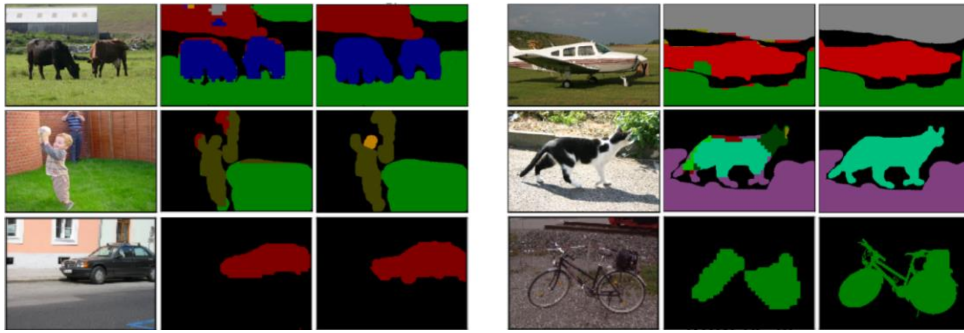


- Multi-scale input channels



- Evaluated on MSRC-9/21 and INRIA Graz-02 data sets





Input Output Truth Input Output Truth

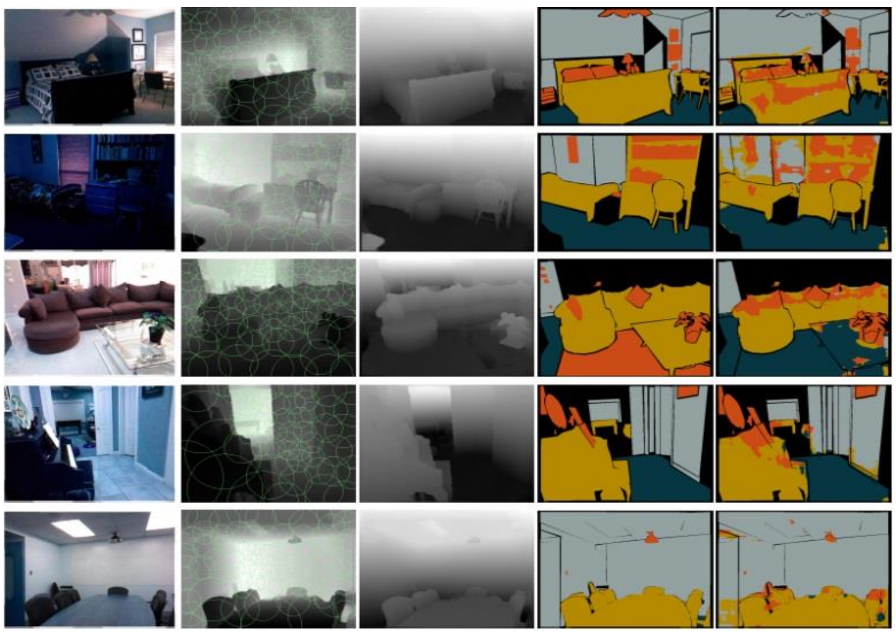
Object Detection in Images

- Bounding box annotation
- Structured loss that directly maximizes overlap of the prediction with ground truth bounding boxes
- Evaluated on two of the Pascal VOC 2007 classes

[Schulz, Behnke, ICANN 2014]

RGB-D Object-Class Segmentation

- Kinect-like sensors provide dense depth
- Scale input according to depth, compute pixel height



RGB Depth Height Truth Output

NYU Depth V2

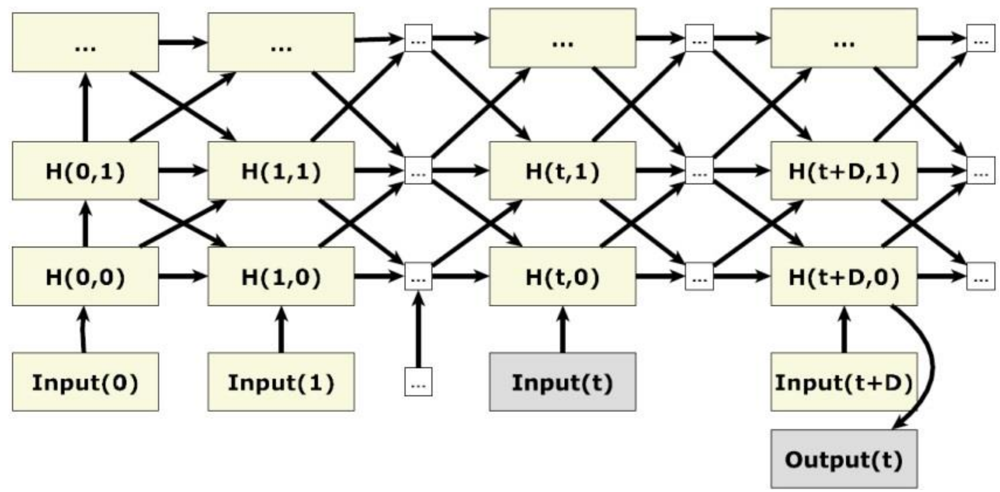
Method	floor	struct	furnit	prop	Class Avg.	Pixel Acc.
CW	84.6	70.3	58.7	52.9	66.6	65.4
CW+DN	87.7	70.8	57.0	53.6	67.3	65.5
CW+H	78.4	74.5	55.6	62.7	67.8	66.5
CW+DN+H	93.7	72.5	61.7	55.5	70.9	70.5
CW+DN+H+SP	91.8	74.1	59.4	63.4	72.2	71.9
CW+DN+H+CRF	93.5	80.2	66.4	54.9	**73.7**	**73.4**
Müller et al.[8]	94.9	78.9	71.1	42.7	71.9	72.3
Random Forest [8]	90.8	81.6	67.9	19.9	65.1	68.3
Couprie et al.[9]	87.3	86.1	45.3	35.5	63.6	64.5
Höft et al.[10]	77.9	65.4	55.9	49.9	62.3	62.0
Silberman [12]	68	59	70	42	59.7	58.6

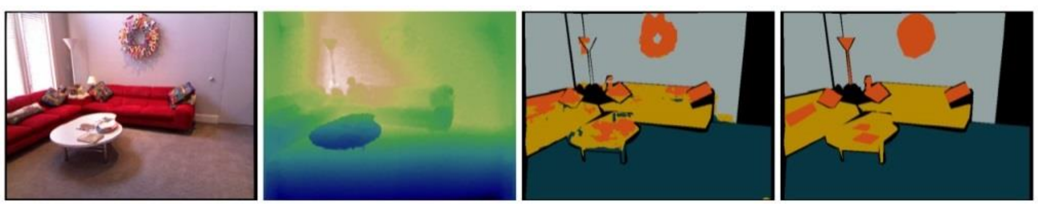
CW is covering windows, H is height above ground, DN is depth normalized patch sizes. SP is averaged within superpixels and SVM-reweighted. CRF is a conditional random field over superpixels [8]. Structure class numbers are optimized for class accuracy.

[Schulz, Höft, Behnke, ESANN 2015]

Neural Abstraction Pyramid for RGB-D Video Object-class Segmentation

- NYU Depth V2 contains RGB-D video sequences
- Recursive computation is efficient for temporal integration





RGB Depth Output Truth

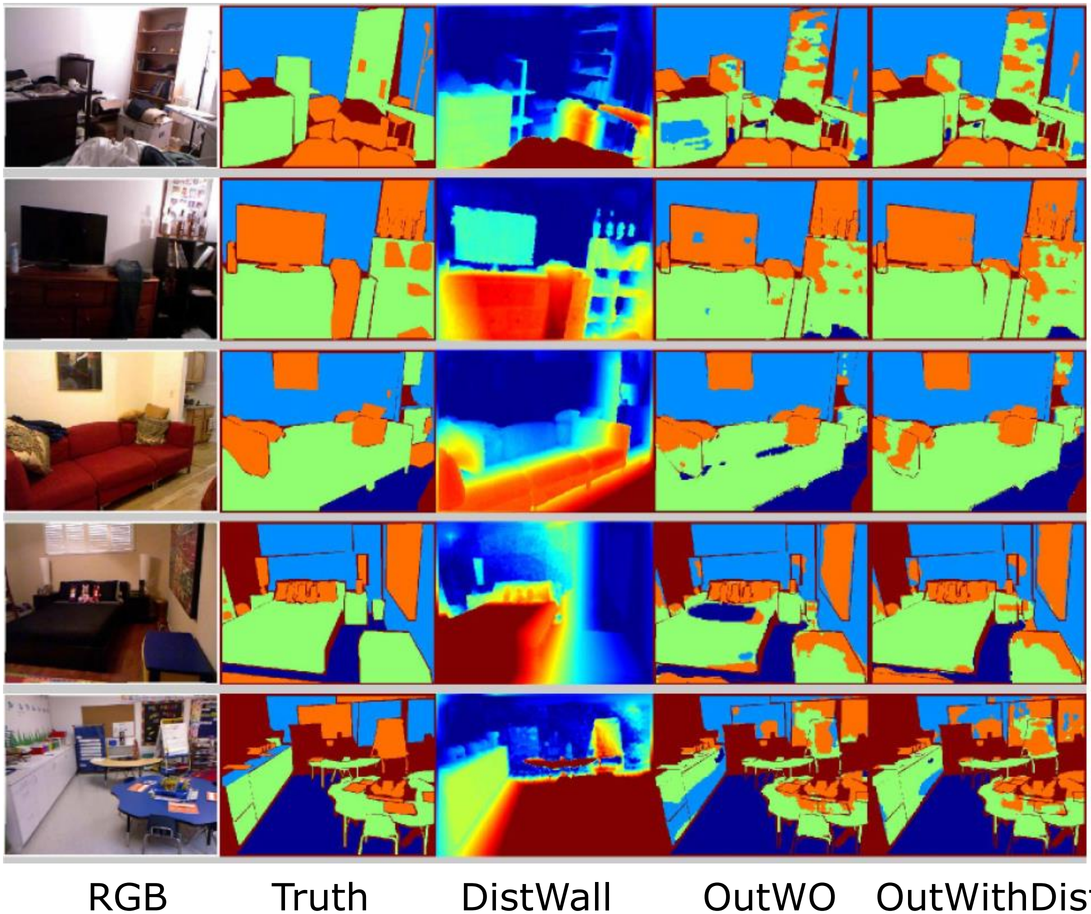
Method	Class Accuracies (%)				Average (%)	
	ground	struct	furnit	prop	Class	Pixel
Höft *et al.* [19]	77.9	65.4	55.9	49.9	62.0	61.1
Unidirectional + MS	73.4	66.8	**60.3**	49.2	62.4	63.1
Schulz *et al.* [20] (no height)	87.7	70.8	57.0	53.6	67.3	65.5
Unidirectional + SW	**90.0**	**76.3**	52.1	**61.2**	**69.9**	**67.5**

[Pavel, Schulz, Behnke, IJCNN 2015]

Geometric and Semantic Features for RGB-D Object-class Segmentation

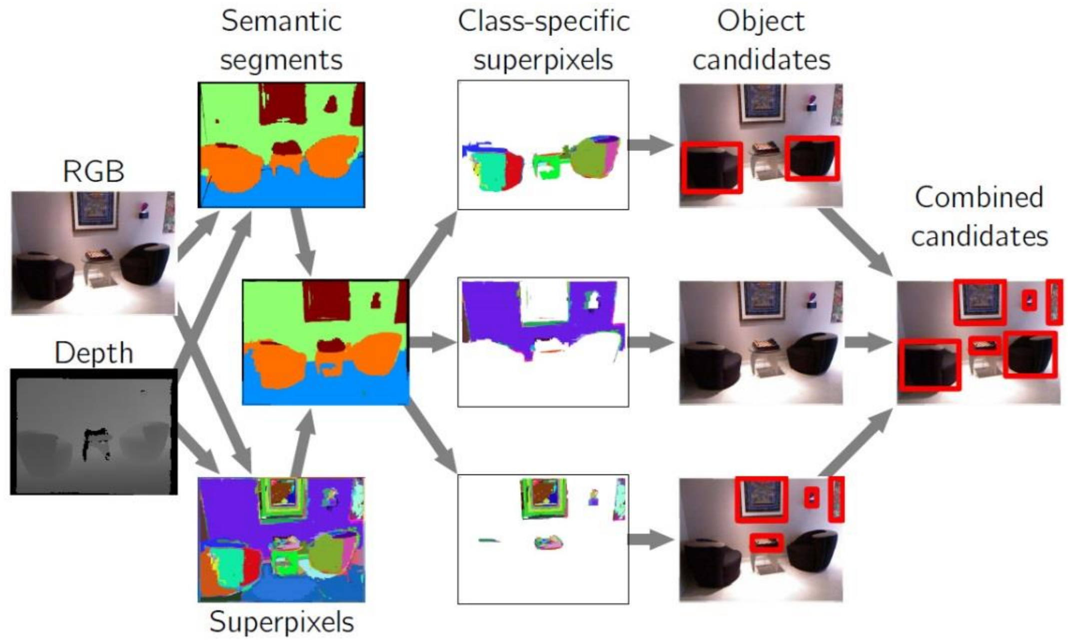
- New **geometric** feature: distance from wall
- **Semantic** features pretrained from ImageNet
- Both help significantly

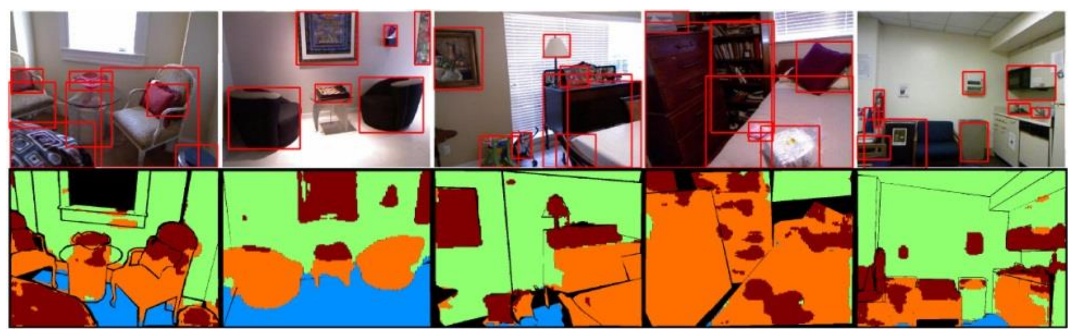
[Husain et al. under review]



Semantic Segmentation Priors for Object Discovery

- Combine bottom-up object discovery and semantic priors
- Semantic segmentation used to classify color and depth superpixels
- Higher recall, more precise object borders

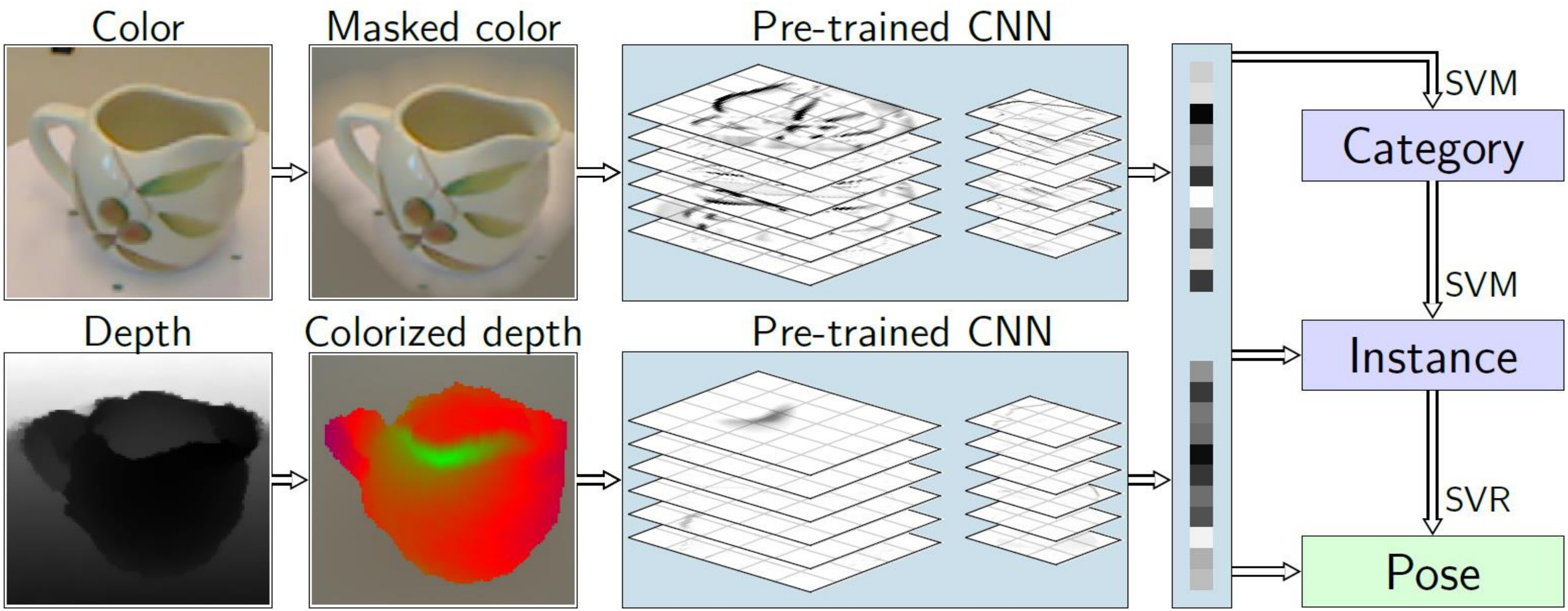




[Garcia et al. under review]

RGB-D Object Recognition and Pose Estimation

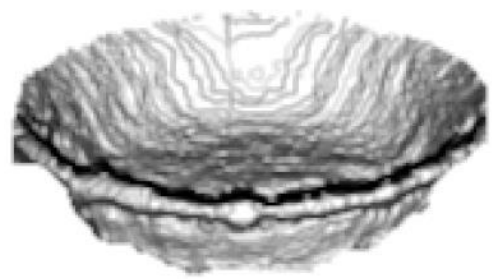
- Use pretrained features from ImageNet

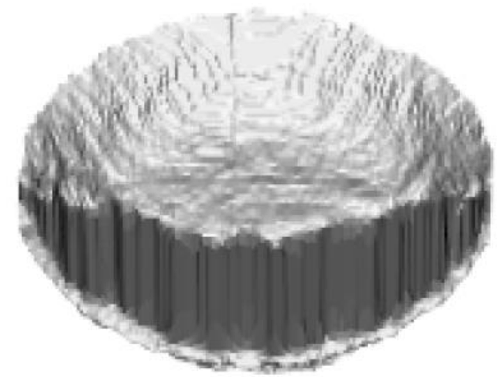


[Schwarz, Schulz, Behnke, ICRA2015]

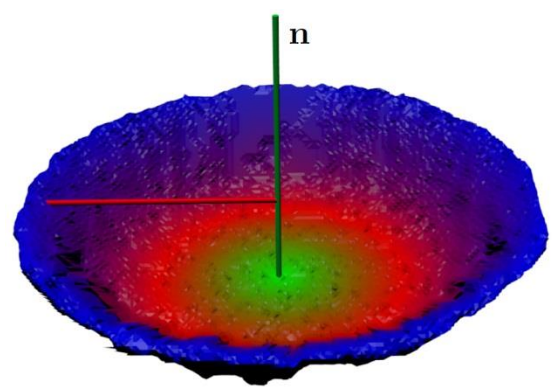
Canonical View, Colorization

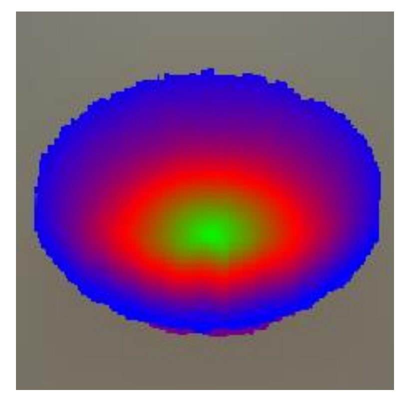
- Objects viewed from different elevation
- Render canonical view





- Colorization based on distance from center vertical

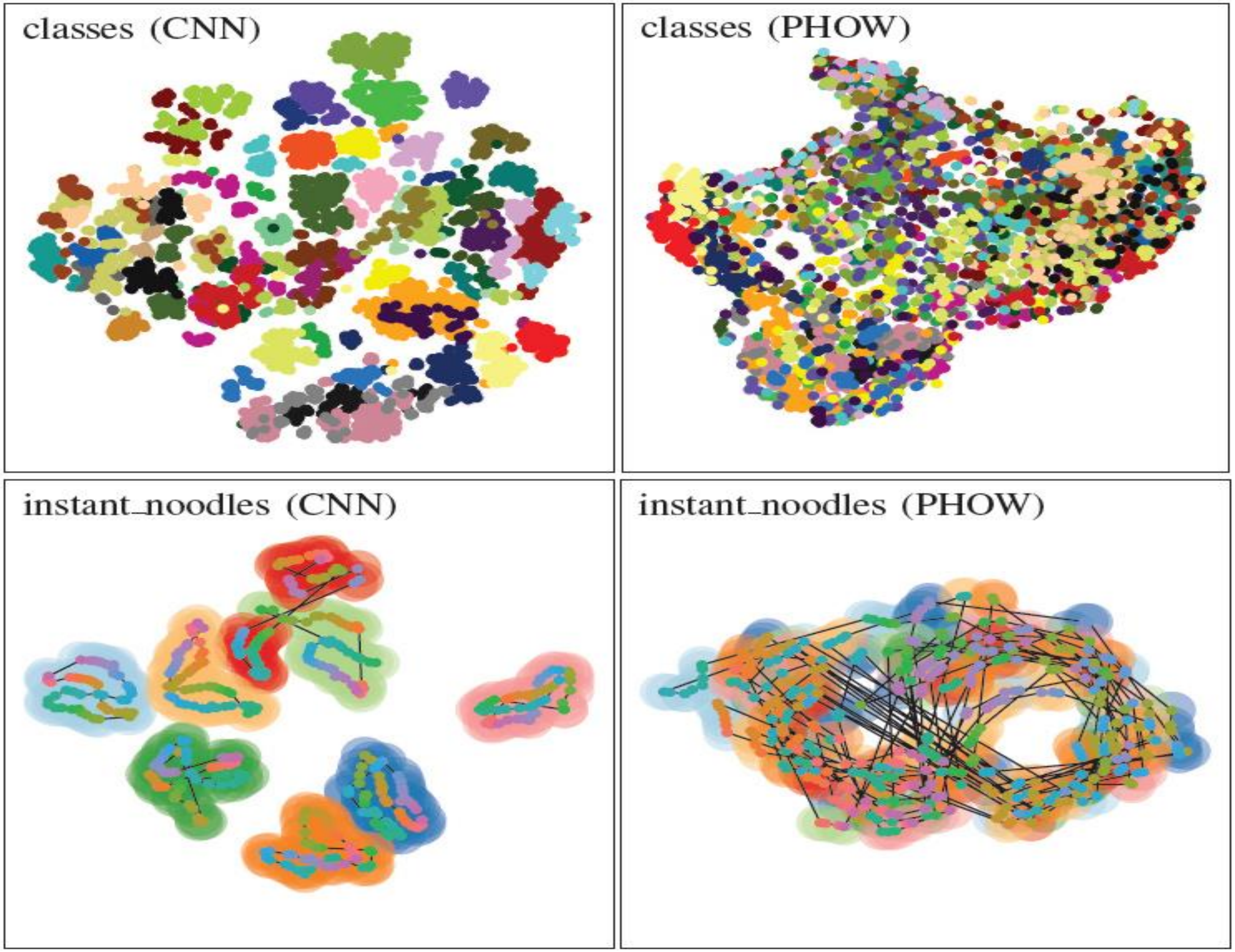




[Schwarz, Schulz, Behnke, ICRA2015]

Features Disentangle Data

- t-SNE embedding



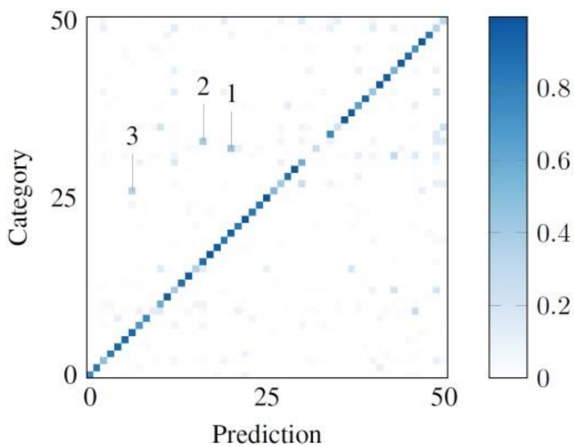
[Schwarz, Schulz, Behnke ICRA2015]

Recognition Accuracy

- Improved both category and instance recognition

	Category Accuracy (%)		Instance Accuracy (%)	
Method	RGB	RGB-D	RGB	RGB-D
Lai *et al.* [1]	74.3 ± 3.3	81.9 ± 2.8	59.3	73.9
Bo *et al.* [2]	82.4 ± 3.1	87.5 ± 2.9	**92.1**	92.8
PHOW[3]	80.2 ± 1.8	—	62.8	—
Ours	$\mathbf{83.1 \pm 2.0}$	88.3 ± 1.5	92.0	**94.1**
Ours	$\mathbf{83.1 \pm 2.0}$	$\mathbf{89.4 \pm 1.3}$	92.0	**94.1**

- Confusion



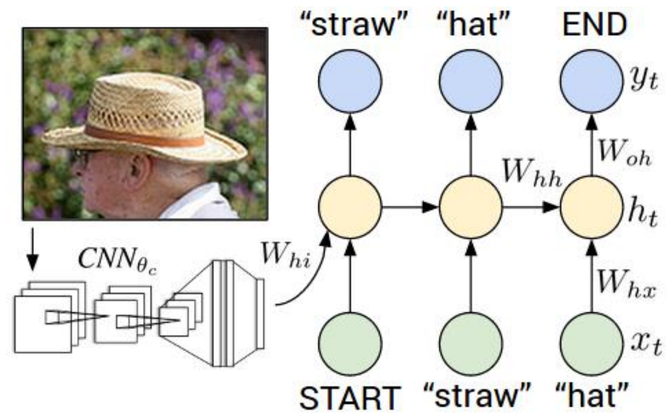
1: pitcher / coffe mug 2: peach / sponge

[Schwarz, Schulz, Behnke, ICRA2015]

Generating Image Captions

- Multimodal recurrent neural network generative model

  [Karpathy, Fei-Fei 2015]

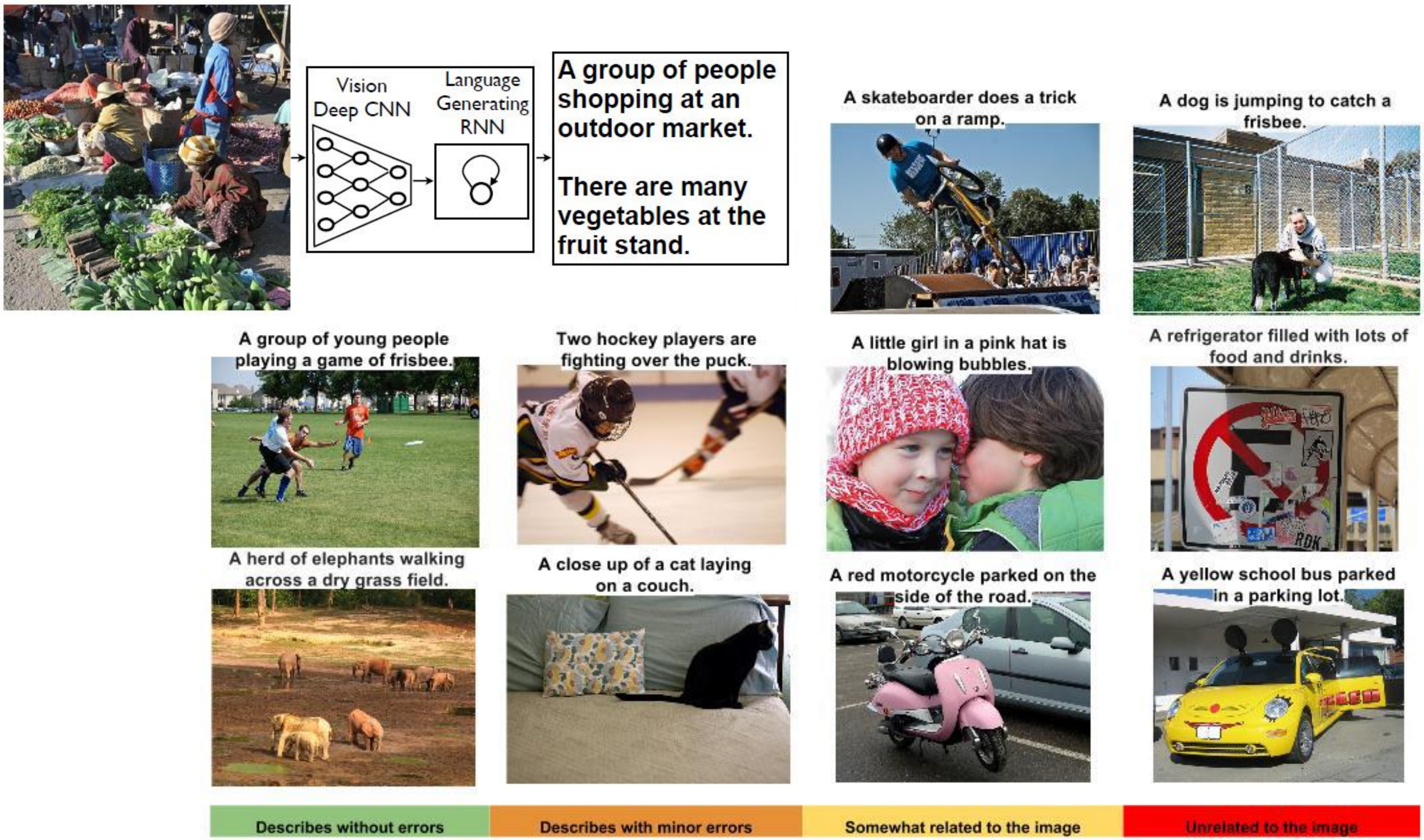


man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

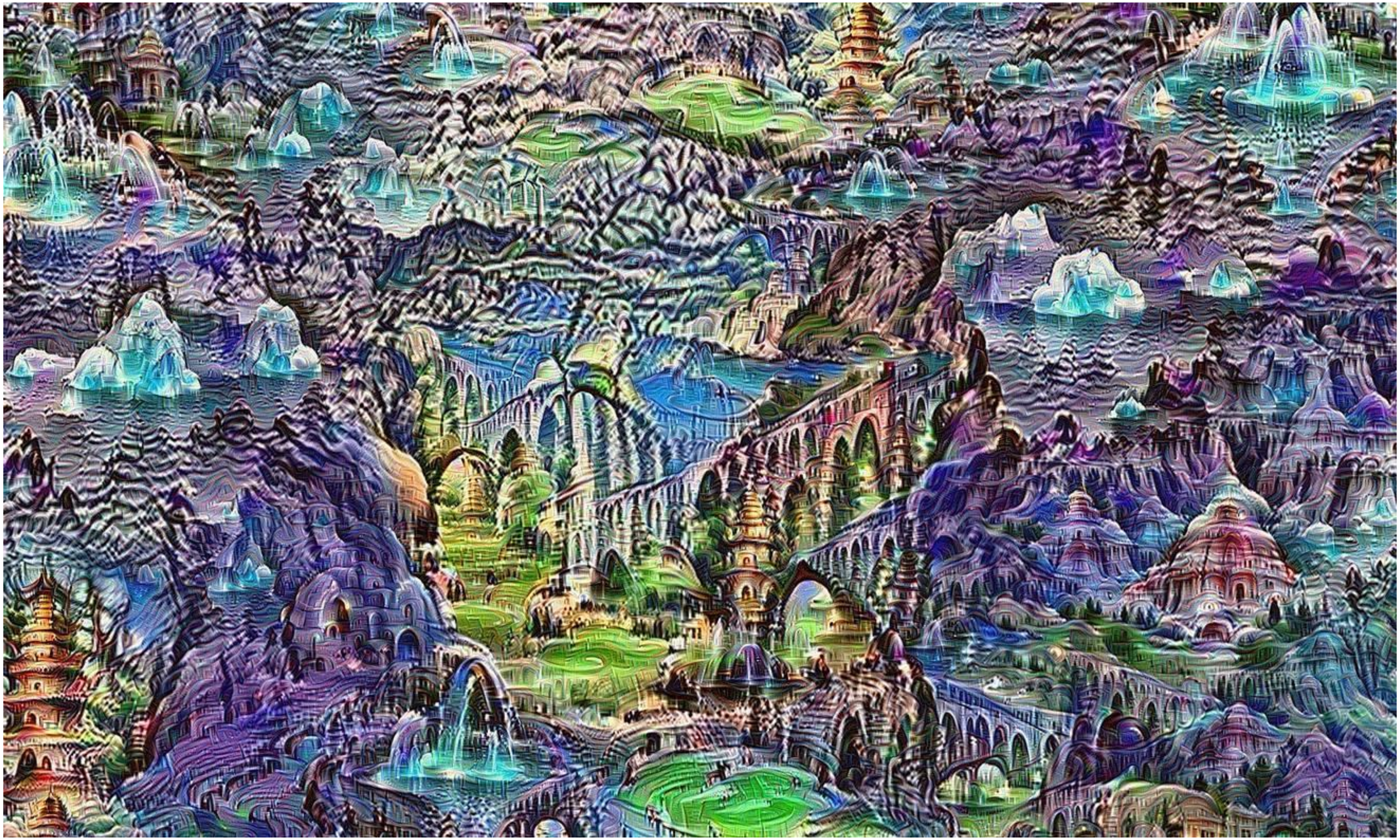
two young girls are playing with lego toy.

Generating Image Captions



[Vinyals et al. 2015]

Dreaming Deep Networks



[Mordvintsev et al 2015]

Painting Style Transfer

Original

Turner

van Gogh

Munch

[Gatys et al. 2015]

Conclusion

- Flat models do not suffice
- Jump from signal to symbols too large
- Deep learning helps here:
  - **Hierarchical, locally connected** models
  - **Non-linear** feature extraction
- **Structure** of learning machine does matter
- Proposed architectures map well to **GPUs**
- **Iterative interpretation** uses partial results as context to resolve ambiguities
- Many questions open
  - Graphical models vs. neural networks
  - Structured vs. unstructured modelling
  - Stability of recurrent networks

Presentation 1
Gregoire Montavon (TU Berlin): Deep Learning of Molecular Properties in the Chemical Compound Space

- Use deep neural networks as a non-linear function approximator in chemistry
- Targets computed by slow conventional method
- Can compute molecular properties of similar molecules quickly
- Application: Search compounds by property

Presentation 2
Takayuki Okatani (Tohoku University): Deep Learning for Material Recognition

- Material recognition is instance of image categorization
- Supervised training of deep convolutional networks
- Reaches human performance
- Seems to work different than human visual system