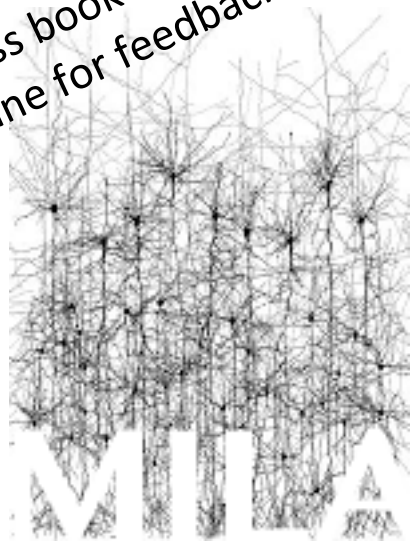
Deep Learning

Yoshua Bengio

September 28, 2015

ICIP'2015, Quebec City

PLUG: **Deep Learning**, MIT Press book in preparation, draft chapters online for feedback



Breakthrough

- **Deep Learning: machine learning algorithms based on learning multiple levels of representation / abstraction.**

Amazing improvements in error rate in object recognition, object detection, speech recognition, and more recently, in natural language processing / understanding

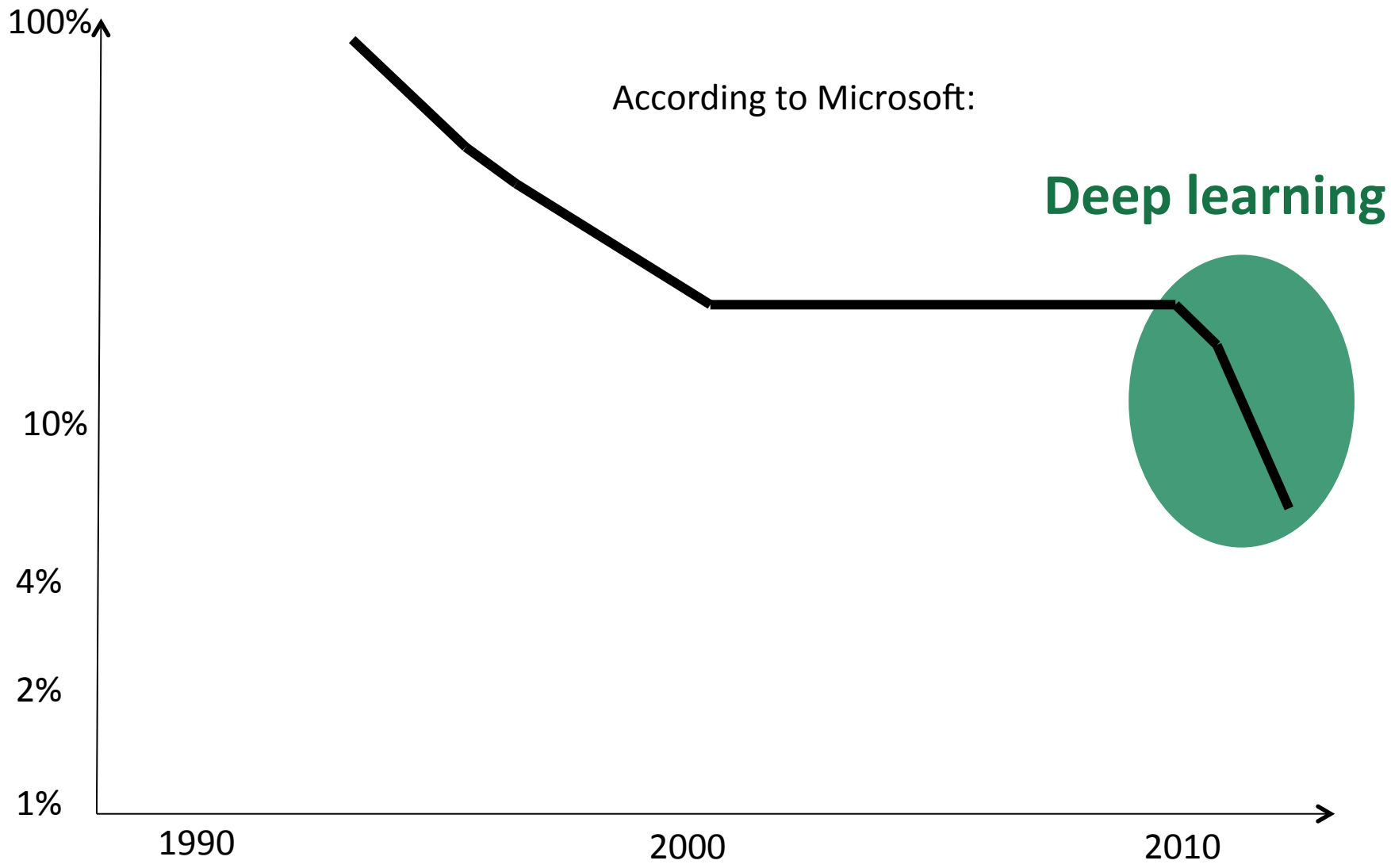
Initial Breakthrough in 2006

Canadian initiative: CIFAR

- Ability to train deep architectures by using layer-wise unsupervised learning, whereas previous purely supervised attempts had failed

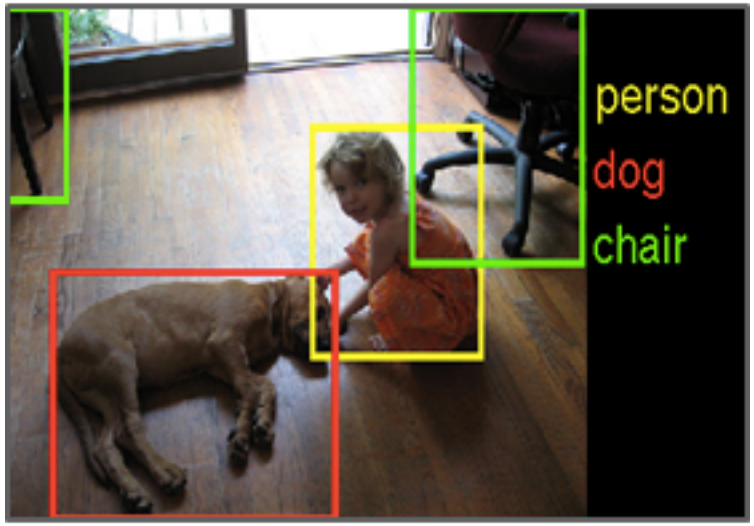
- Unsupervised feature learners:
  - RBMs
  - Auto-encoder variants
  - Sparse coding variants

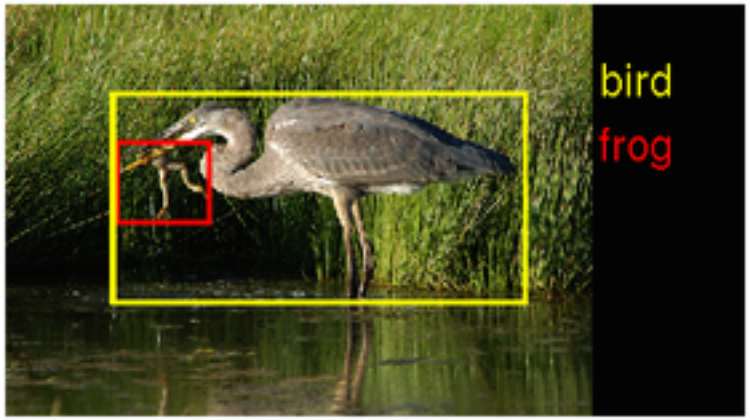
2010-2012: Breakthrough in speech recognition → in Androids by 2012



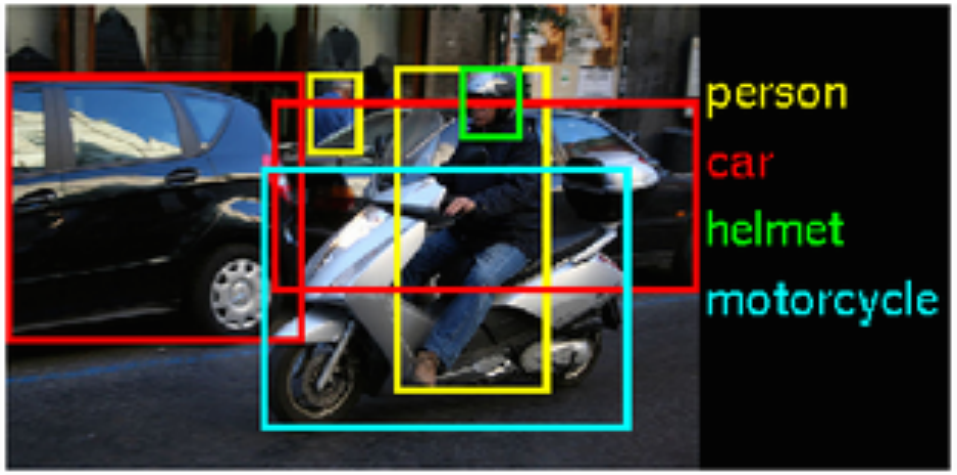
Breakthrough in computer vision: 2012-2015

- GPUs + 10x more data





- 1000 object categories,
- Facebook: millions of faces
- 2015: human-level performance



Deep Learning in the News

The New York Tim

Monday, June 25, 2012 Last Update: 11:50 PM ET

EXCLUSIVE

Facebook, Google in 'Deep Learning' Arms Race

Yann LeCun, an NYU artificial intelligence researcher who now works for Facebook. *Photo: Josh Valcarcel/WIRED*

WIRED

NEWS BULLETIN

Google Beat Facebook for DeepMind

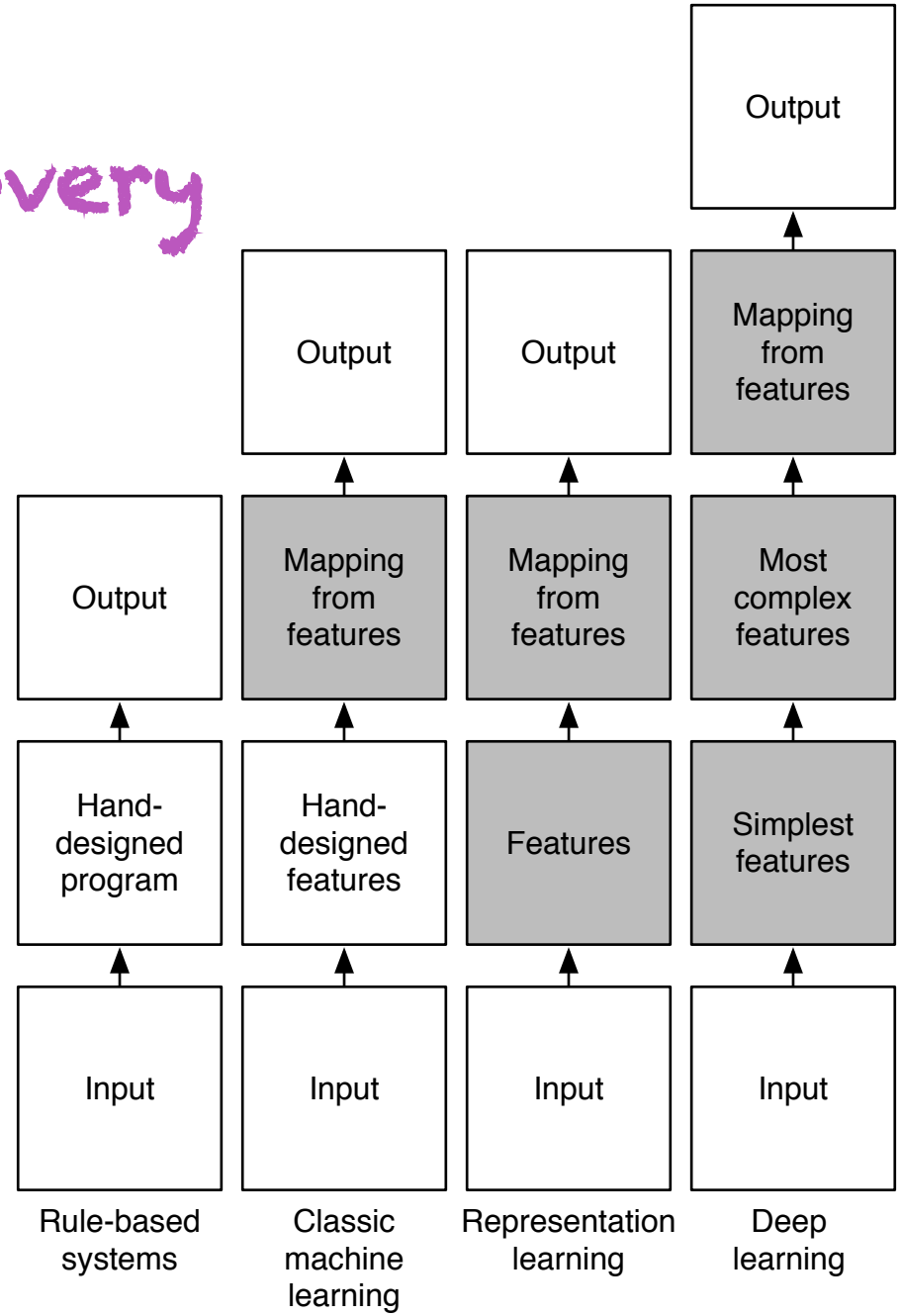
Google Acquires Artificial Intelligence Startup DeepMind For More Than $500M

Posted Jan 26, 2014 by Catherine Shu (@catherineshu)

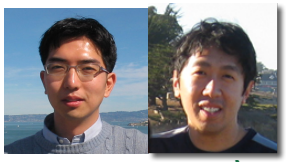
IT Companies are Racing into Deep Learning

Why is Deep Learning Working so Well?

Automating Feature Discovery

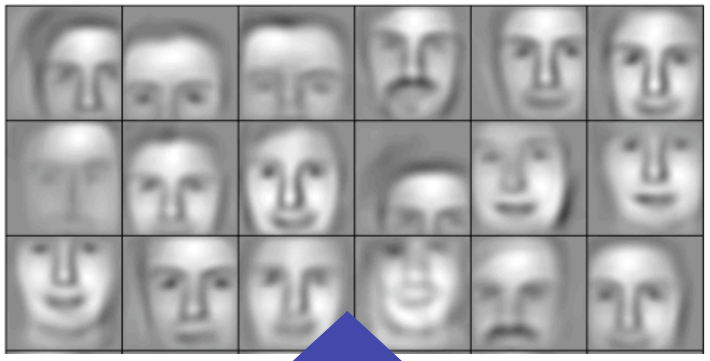


Learning multiple levels of representation



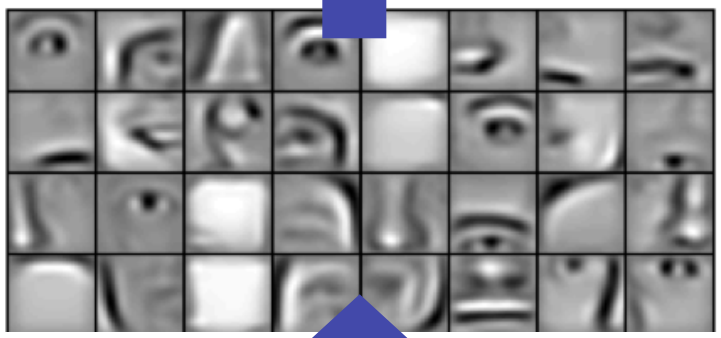
(Lee, Largman, Pham & Ng, NIPS 2009)
(Lee, Grosse, Ranganath & Ng, ICML 2009)

Successive model layers learn deeper intermediate representations



Layer 3

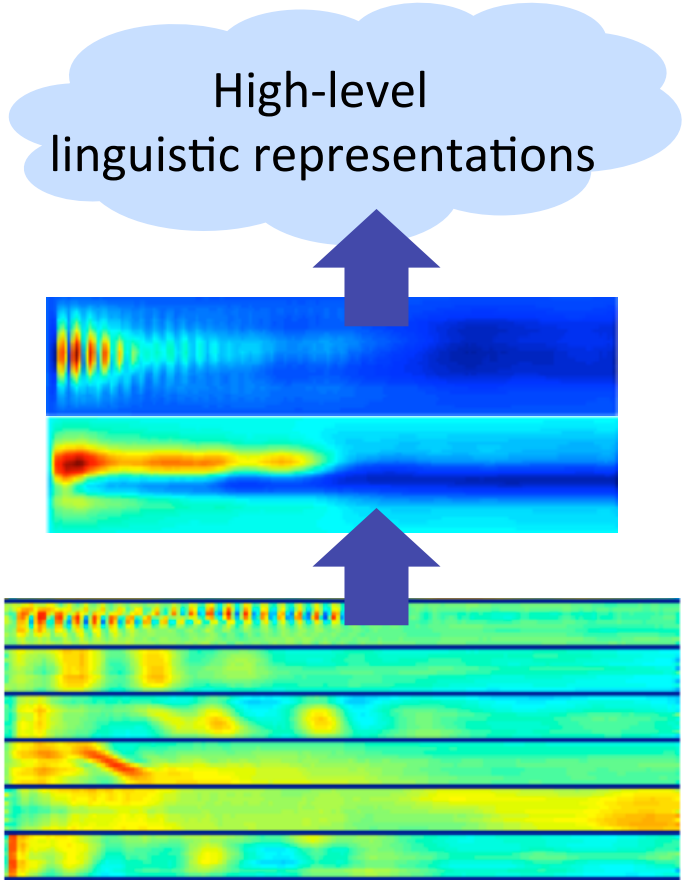
Parts combine to form objects



Layer 2

Layer 1

High-level linguistic representations



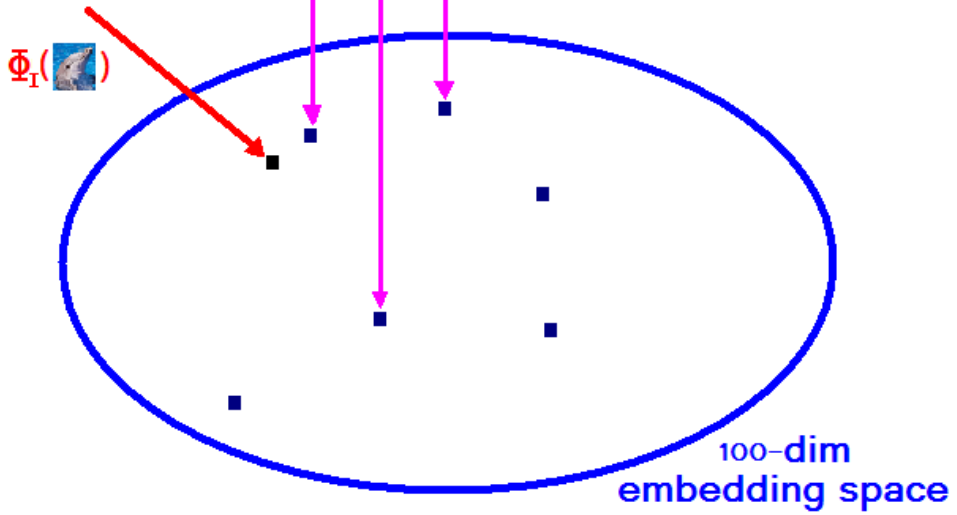
Prior: underlying factors & concepts compactly expressed w/ multiple levels of abstraction

Google Image Search:
Different object types represented in the same space

Google:

S. Bengio, J. Weston & N. Usunier

(IJCAI 2011, NIPS'2010, JMLR 2010, MLJ 2010)



Learn $\Phi_I(\cdot)$ and $\Phi_W(\cdot)$ to optimize precision@k.

Why is deep learning working so well?

Machine Learning, AI & No Free Lunch

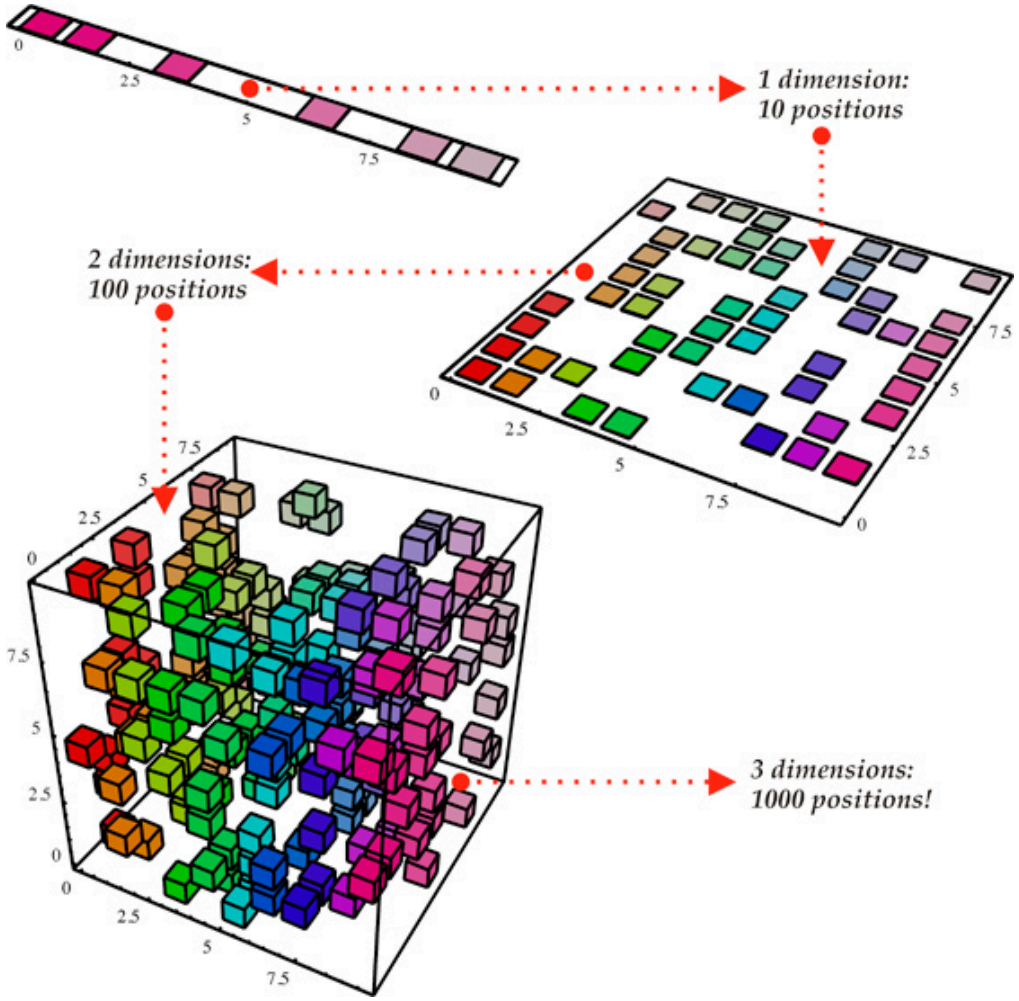
- Four key ingredients for ML towards AI

  1. Lots & lots of data
  2. Very flexible models
  3. Enough computing power
  4. Powerful priors that can defeat the curse of dimensionality

ML 101. What We Are Fighting Against: The Curse of Dimensionality

To generalize locally, need representative examples for all relevant variations!

Classical solution: hope for a smooth enough target function, or make it smooth by handcrafting good features / kernel



Bypassing the curse of dimensionality

We need to build compositionality into our ML models

Just as human languages exploit compositionality to give representations and meanings to complex ideas

Exploiting compositionality gives an exponential gain in representational power

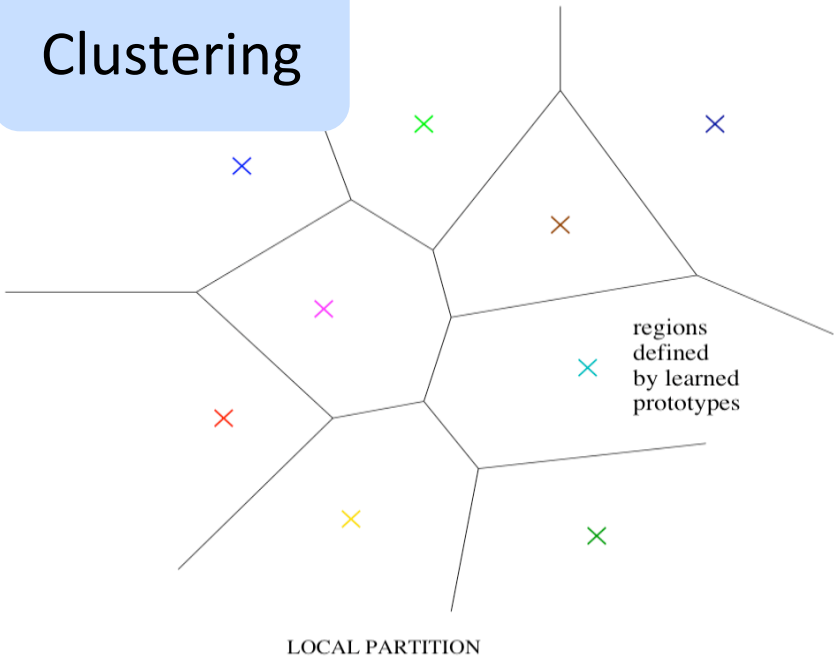
Distributed representations / embeddings: feature learning

Deep architecture: multiple levels of feature learning

Prior: compositionality is useful to describe the world around us efficiently

Non-distributed representations

Clustering



- Clustering, n-grams, Nearest-Neighbors, RBF SVMs, local non-parametric density estimation & prediction, decision trees, etc.

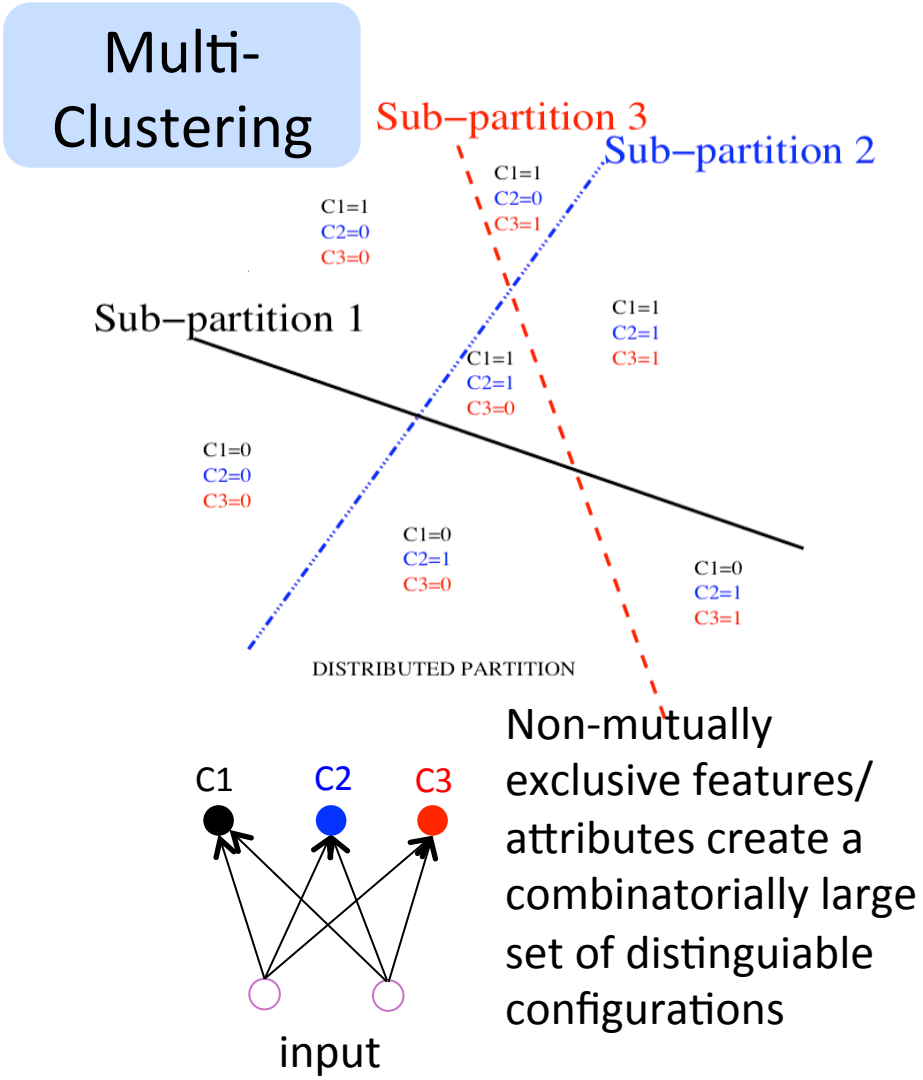
- Parameters for each distinguishable region
- **# of distinguishable regions is linear in # of parameters**

→ No non-trivial generalization to regions without examples

The need for distributed representations

- Factor models, PCA, RBMs, Neural Nets, Sparse Coding, Deep Learning, etc.

- Each parameter influences many regions, not just local neighbors
- **# of distinguishable regions grows almost exponentially with # of parameters**
- **GENERALIZE NON-LOCALLY TO NEVER-SEEN REGIONS**



Summary of Some New Theory Results

- Expressiveness of deep networks with piecewise linear activation functions: exponential advantage for depth

  (Montufar et al NIPS 2014)

- Theoretical and empirical evidence against bad local minima

  (Dauphin et al NIPS 2014)

- Manifold & probabilistic interpretations of auto-encoders
  - Estimating the gradient of the energy function *(Alain & Bengio ICLR 2013)*
  - Sampling via Markov chain *(Bengio et al NIPS 2013)*
  - Variational auto-encoder breakthrough *(Gregor et al arXiv 2015)*

The Depth Prior can be Exponentially Advantageous

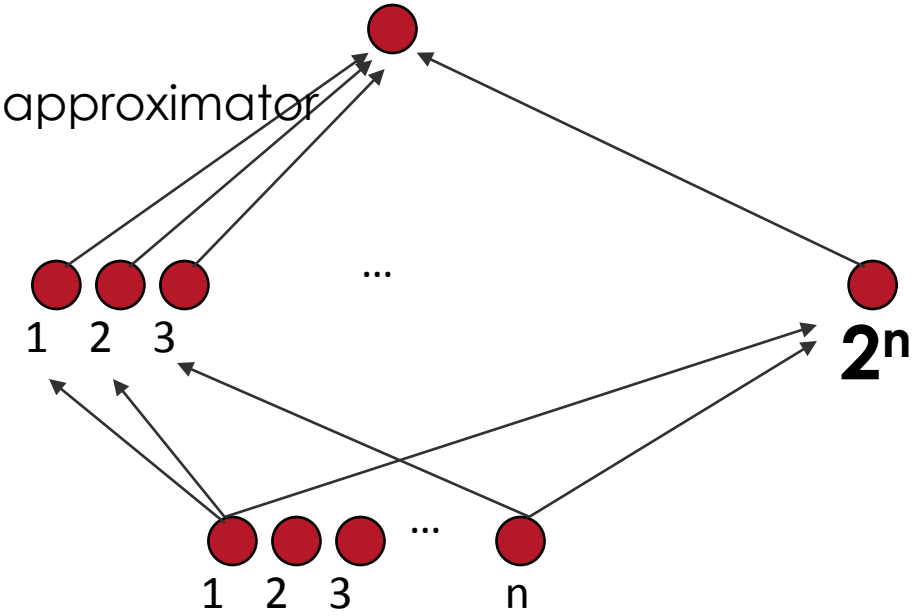
Theoretical arguments:

2 layers of { Logic gates / Formal neurons / RBF units } = universal approximator

RBMs & auto-encoders = universal approximator

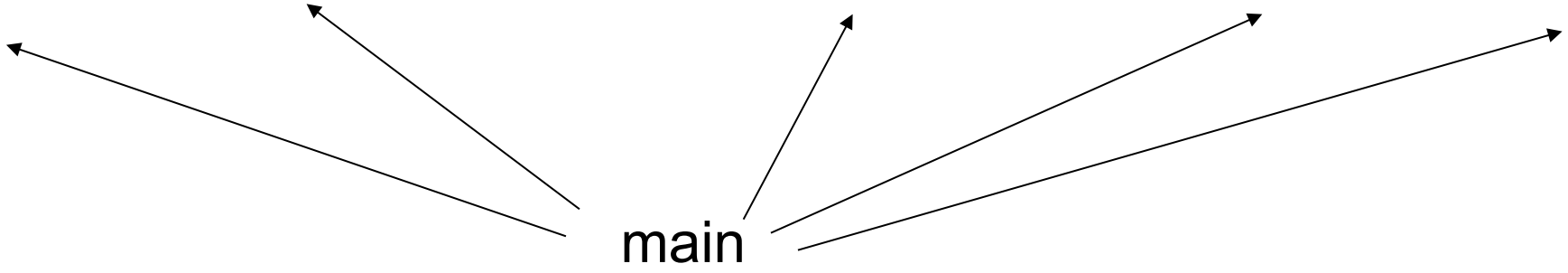
Theorems on advantage of depth:
(Hastad et al 86 & 91, Bengio et al 2007, Bengio & Delalleau 2011, Braverman 2011, Pascanu et al 2014, Montufar et al **NIPS 2014**)

Some functions compactly represented with k layers may require exponential size with 2 layers

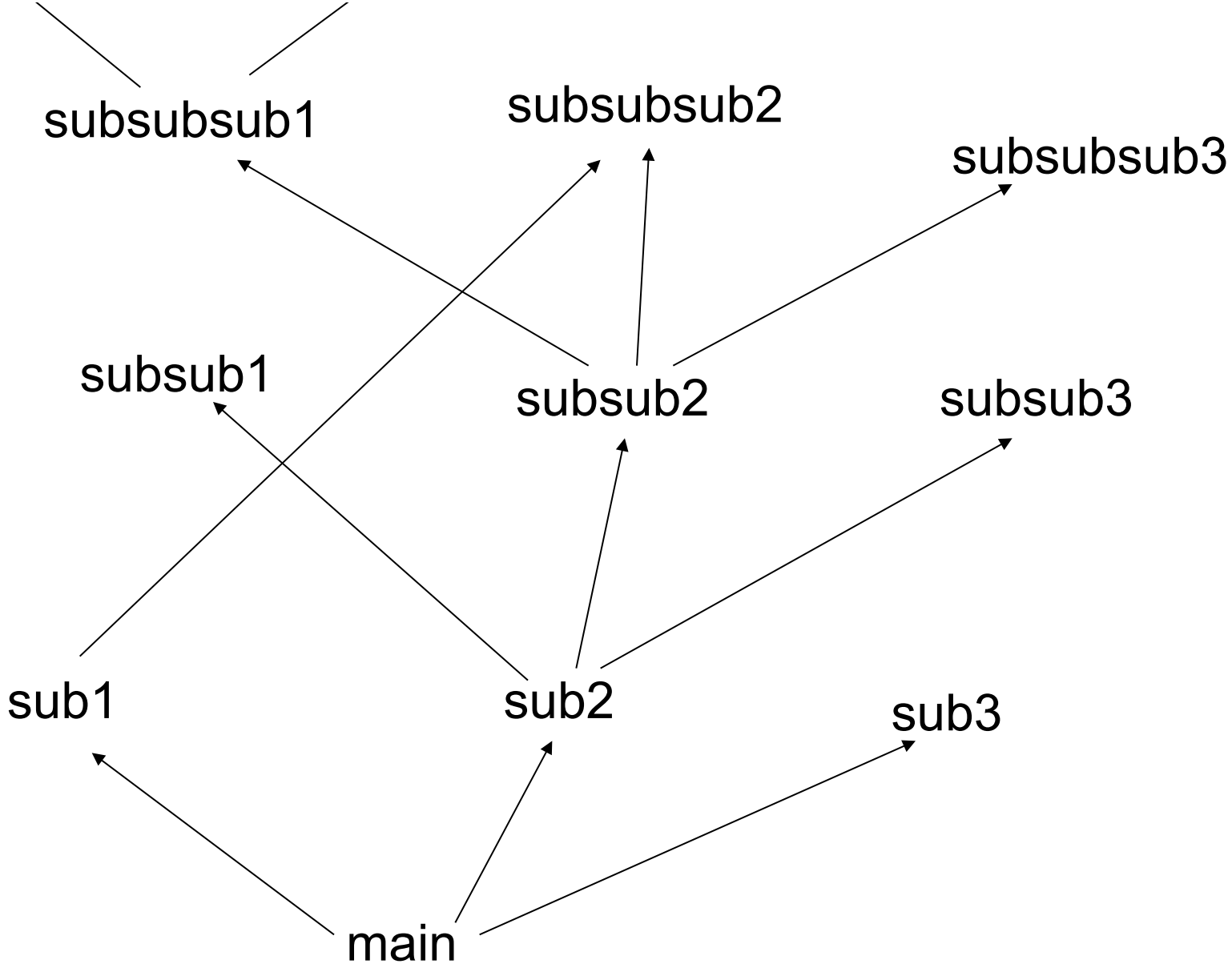


subroutine1 includes
subsub1 code and
subsub2 code and
subsubsub1 code

subroutine2 includes
subsub2 code and
subsub3 code and
subsubsub3 code and …



“Shallow” computer program



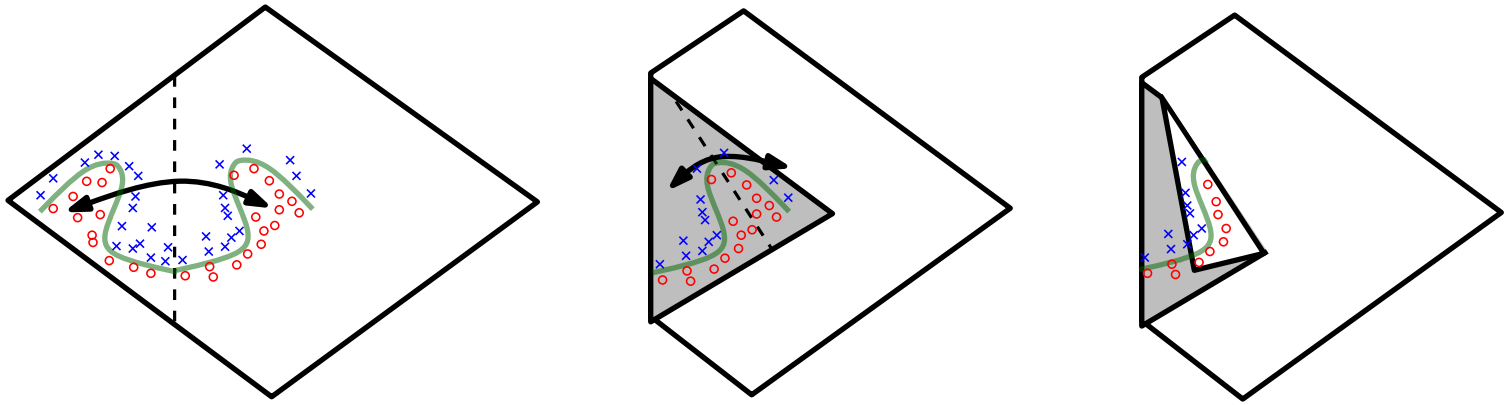
“Deep” computer program

New theoretical result: Expressiveness of deep nets with piecewise-linear activation fns

(Pascanu, Montufar, Cho & Bengio; ICLR 2014)

(Montufar, Pascanu, Cho & Bengio; NIPS 2014)

Deeper nets with rectifier/maxout units are exponentially more expressive than shallow ones (1 hidden layer) because they can split the input space in many more (not-independent) linear regions, with constraints, e.g., with abs units, each unit creates mirror responses, folding the input space:

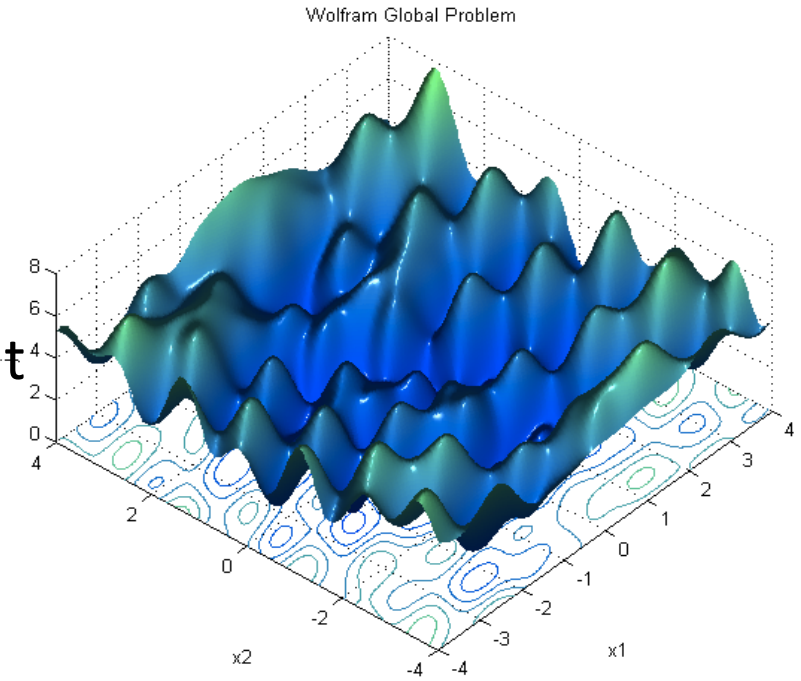


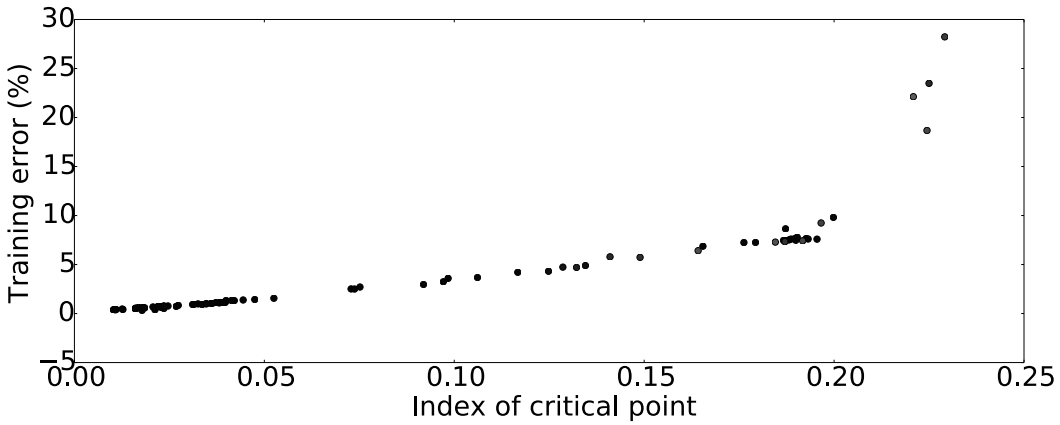
A Myth is Being Debunked: Local Minima in Neural Nets → Convexity is not needed

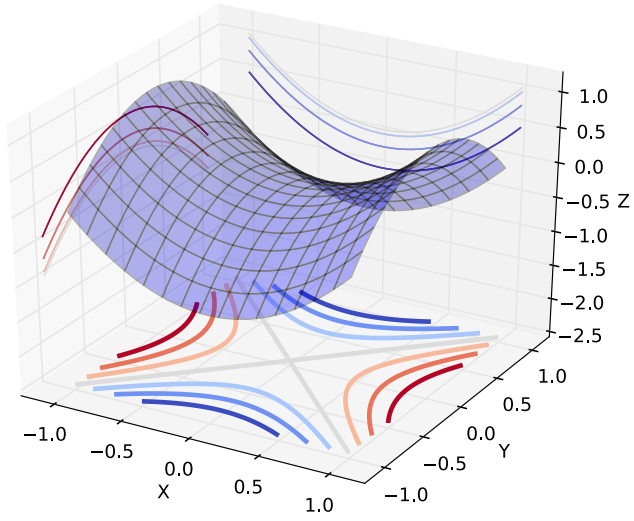
- (Pascanu, Dauphin, Ganguli, Bengio, arXiv May 2014): *On the saddle point problem for non-convex optimization*
- (Dauphin, Pascanu, Gulcehre, Cho, Ganguli, Bengio, NIPS' 2014): *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*
- (Choromanska, Henaff, Mathieu, Ben Arous & LeCun 2014): *The Loss Surface of Multilayer Nets*

Saddle Points

- Local minima dominate in low-D, but saddle points dominate in high-D
- Most local minima are close to the bottom (global minimum error)





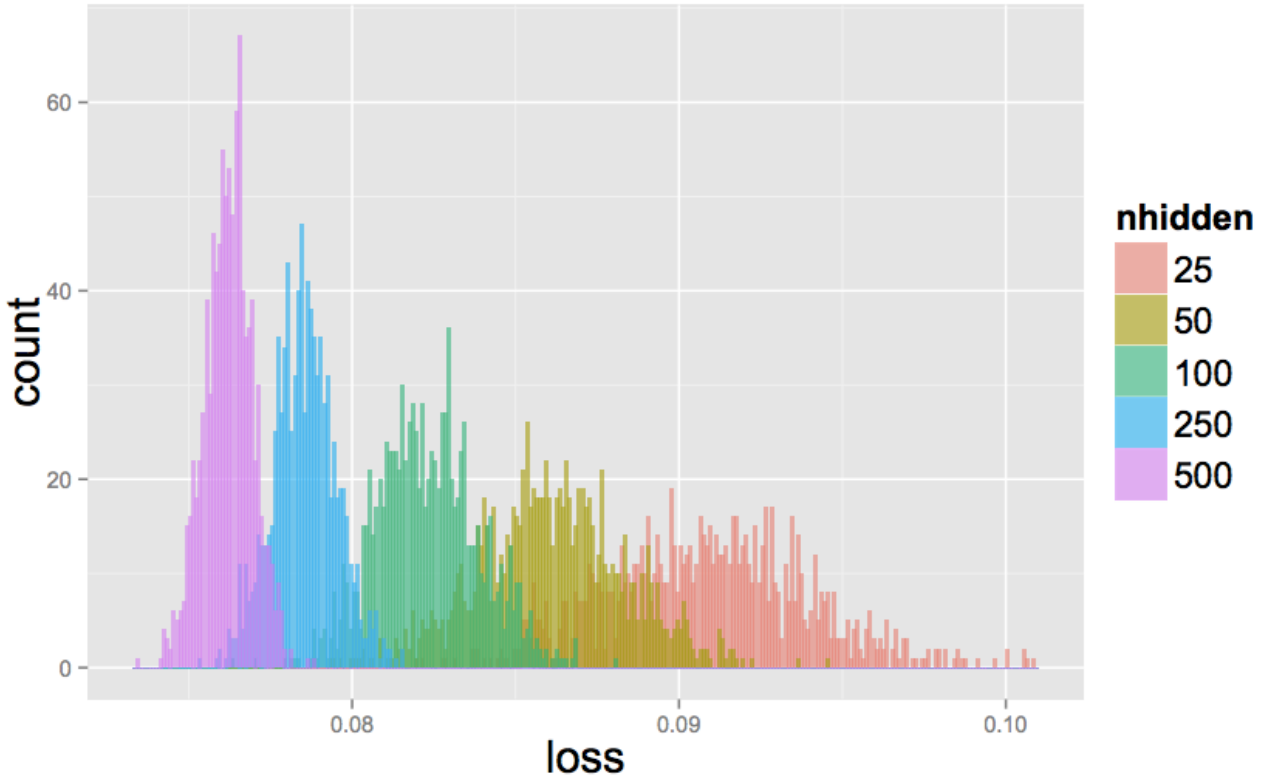


Low Index Critical Points

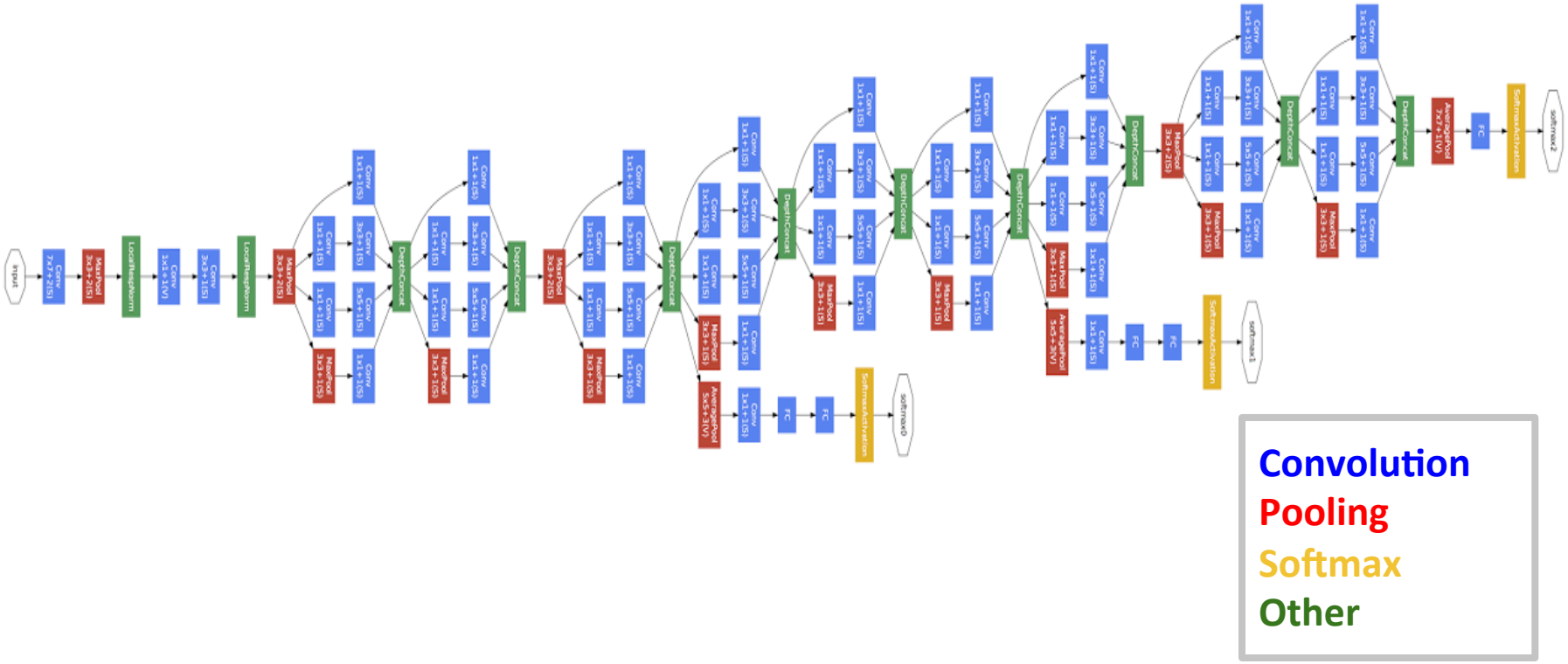
Choromanska et al & LeCun 2014, 'The Loss Surface of Multilayer Nets'

Shows that deep rectifier nets are analogous to spherical spin-glass models

The low-index critical points of large models concentrate in a band just above the global minimum

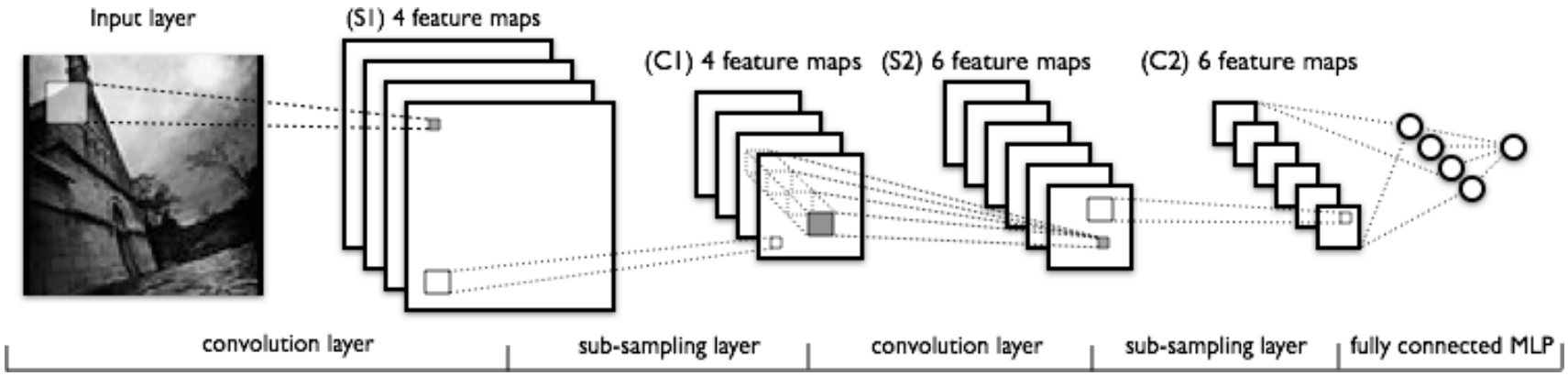


GoogLeNet: 22 layers, intermediate targets



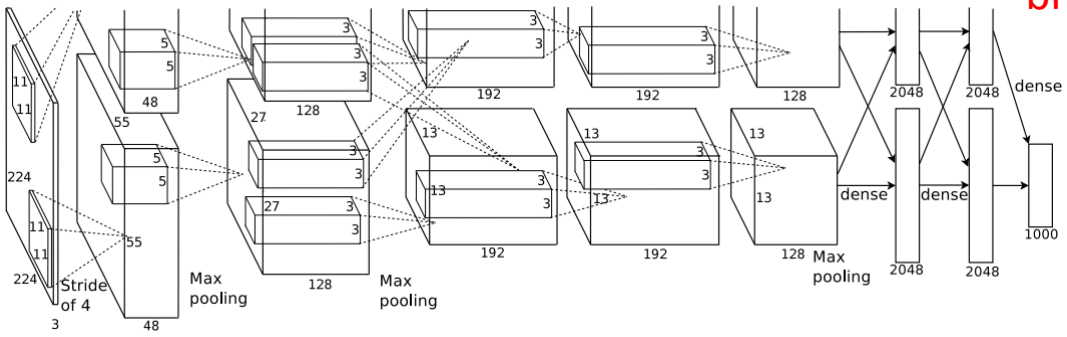
Alternating convolutions & pooling

- Inspired by visual cortex, idea from Fukushima's Neocognitron, combined with back-prop and developped by **LeCun** since 1989



- Increasing number of features, decreasing spatial resolution
- Top layers are fully connected

Krizhevsky, Sutskever & Hinton 2012
breakthrough in object recognition



Deep Learning: Beyond Pattern Recognition, towards AI

- Many researchers believed that neural nets could at best be good at pattern recognition
- And they are really good at it!

- But many more ingredients needed towards AI. Recent progress:

  - REASONING: with extensions of recurrent neural networks
    - Memory networks & Neural Turing Machine
  - PLANNING & REINFORCEMENT LEARNING: DeepMind (Atari game playing) & Berkeley (Robotic control)

Ongoing Progress: Combining Vision and Natural Language Understanding

- Recurrent nets generating credible sentences, even better if conditionally:
  - Machine translation
  - Image 2 text

Xu et al, ICML'2015

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

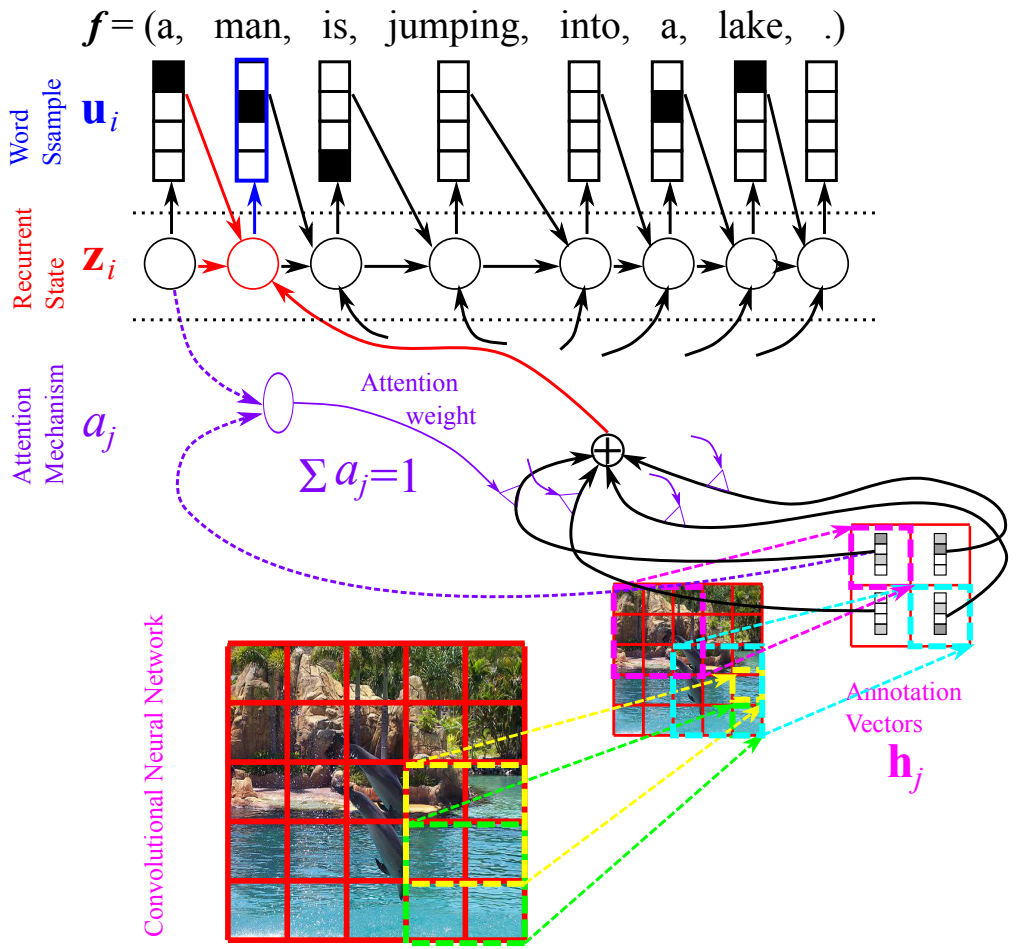
A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

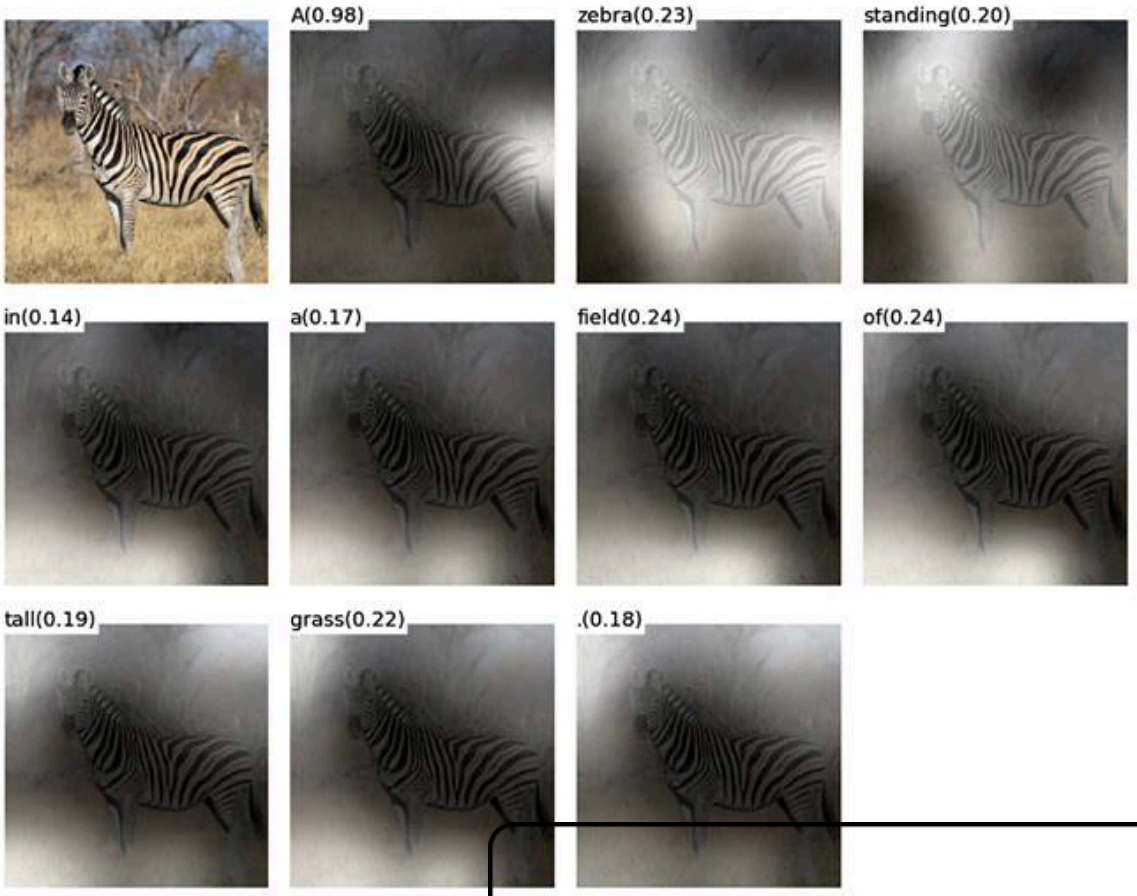
A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

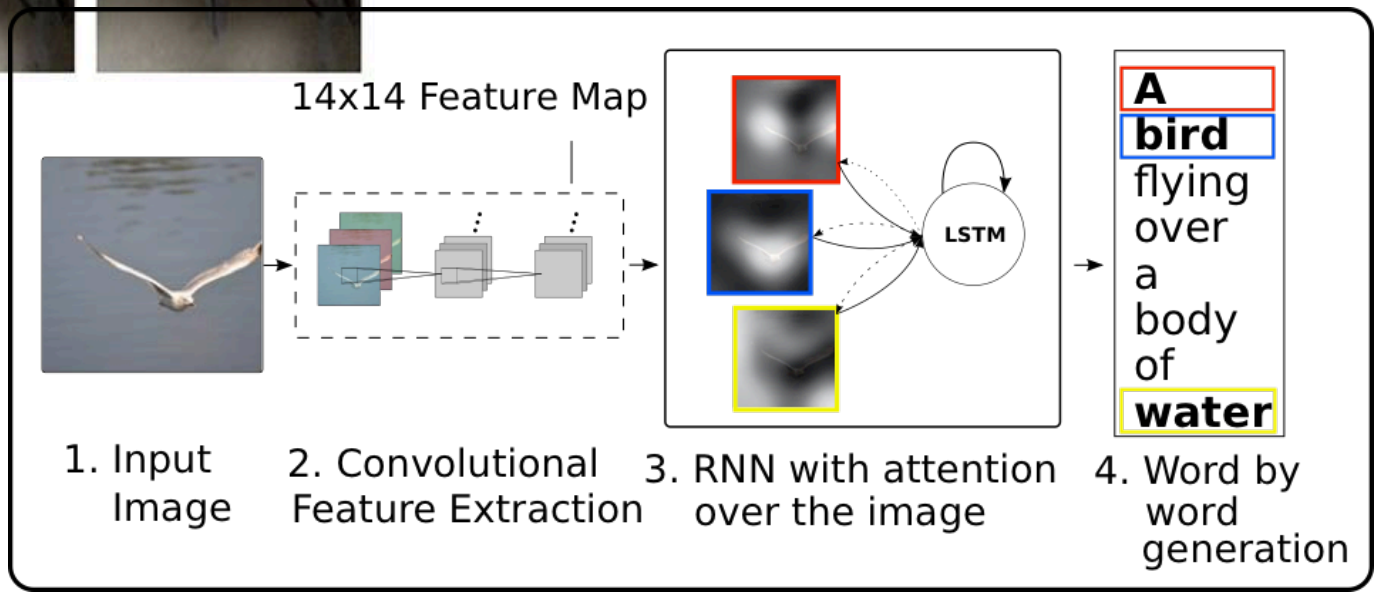
Image-to-Text: Caption Generation



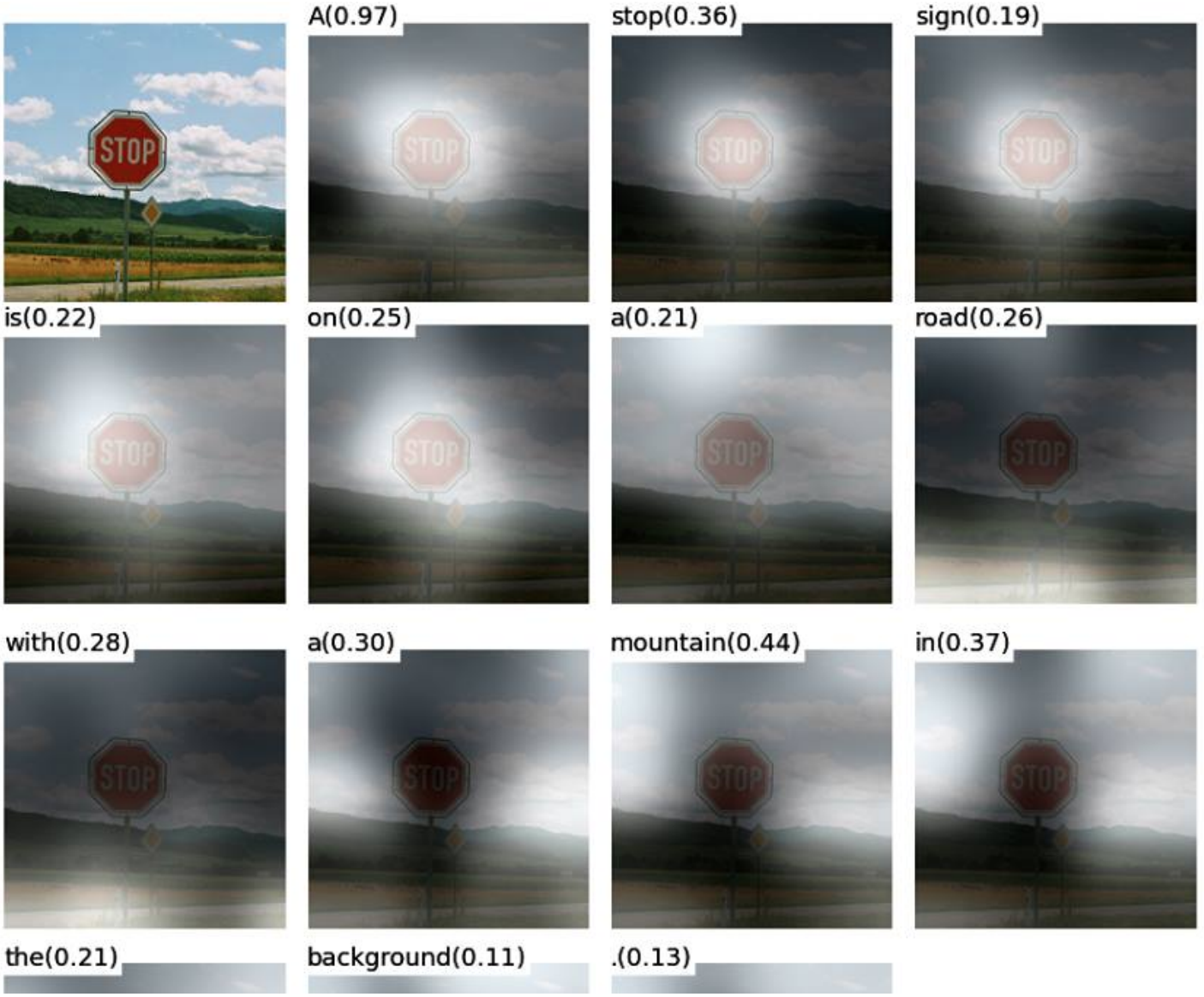
(Xu et al., 2015), (Yao et al., 2015)



Paying Attention to Selected Parts of the Image While Uttering Words

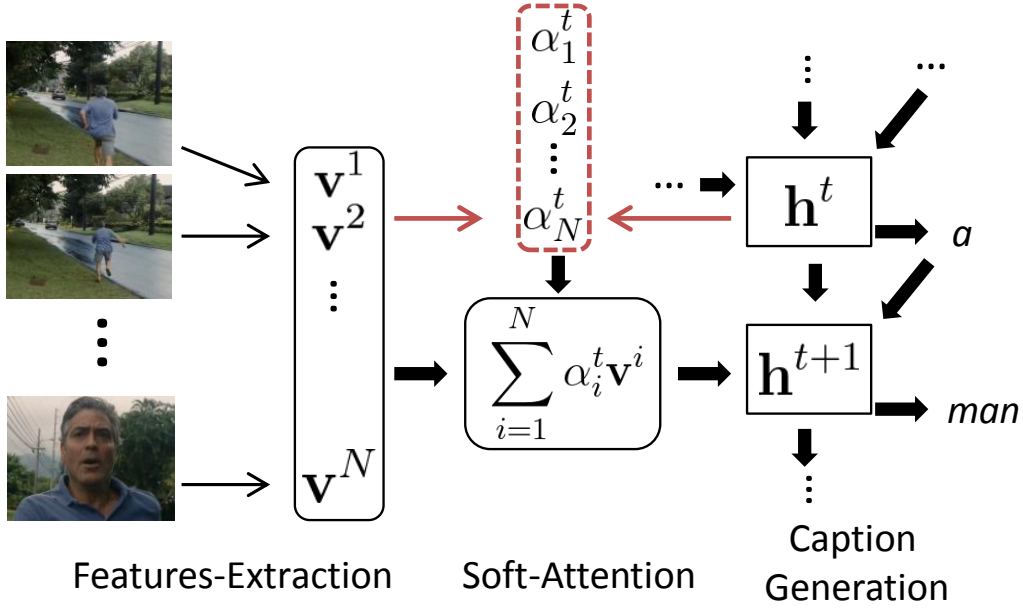


Speaking about what one sees



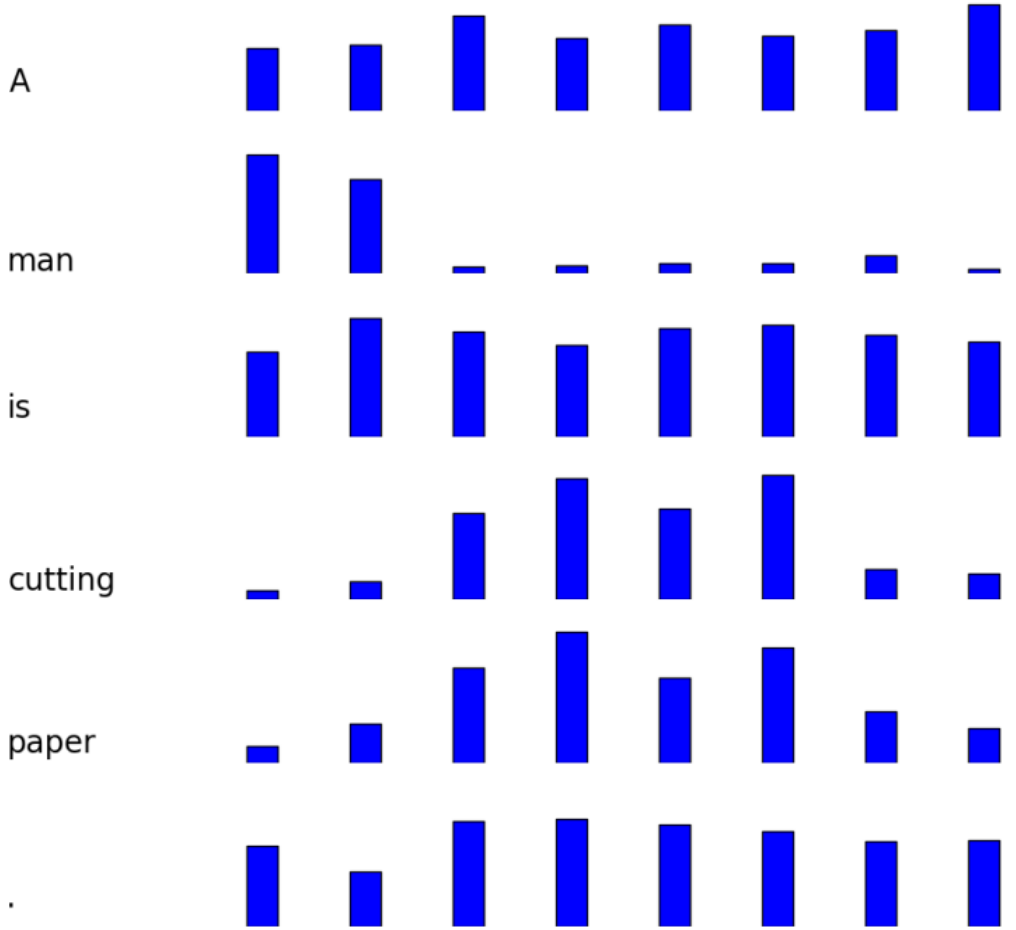
Attention through time for video caption generation

- (Yao et al arXiv 1502.08029, 2015) *Video Description Generation Incorporating Spatio-Temporal Features and a Soft-Attention Mechanism*
- Attention can be focused temporally, i.e., selecting input frames



Attention through time for video caption generation (Yao et al 2015)

- Attention is focused at appropriate frames depending on which word is generated.



Attention through time for video caption generation (Yao et al 2015)

- Soft-attention worked best in this setting

Model	Feature	Bleu					Meteor	Perplexity
		1	2	3	4	mb		
non-attention	GNet	32.0	9.2	3.4	1.2	0.3	4.43	88.28
	$\text{GNet+3DConv}_{\text{non-att}}$	33.6	10.4	4.3	1.8	0.7	5.73	84.41
soft-attention	GNet	31.0	7.7	3.0	1.2	0.3	4.05	66.63
	$\text{GNet+3DConv}_{\text{att}}$	28.2	8.2	3.1	1.3	0.7	5.6	**65.44**

Corpus:
She rushes out.
Test_sample:
The woman turns away.

Corpus:
SOMEONE sits with his arm around SOMEONE.
He nuzzles her cheek, then kisses tenderly.
Test_sample:
SOMEONE sits beside SOMEONE.

Corpus:
SOMEONE shuts the door.
Test_sample:
as he turns on his way to the door , SOMEONE turns away.

Generated captions

Beyond Object Recognition and Caption Generation

Visual Question Answering (Antol et al 2015)

I: Jane went to the hallway.
I: Mary walked to the bathroom.
I: Sandra went to the garden.
I: Daniel went back to the garden.
I: Sandra took the milk there.
Q: Where is the milk?
A: garden

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

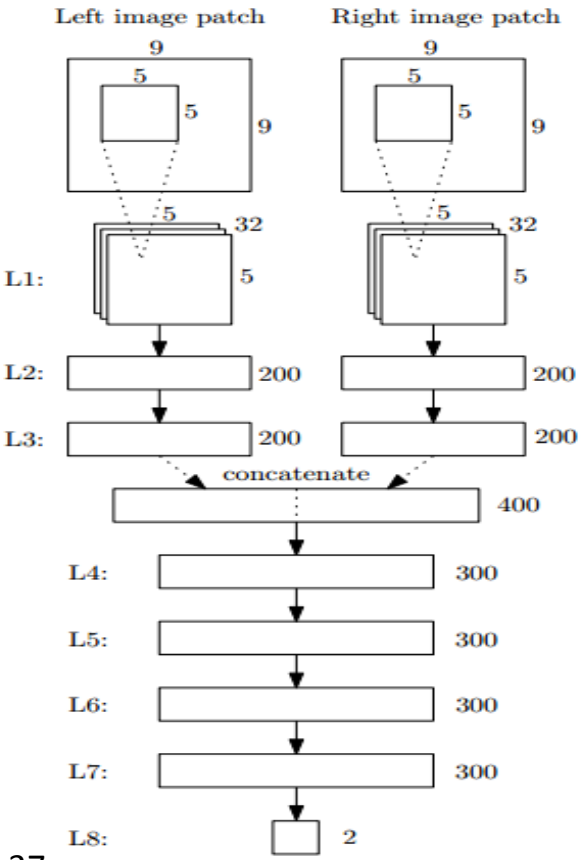
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

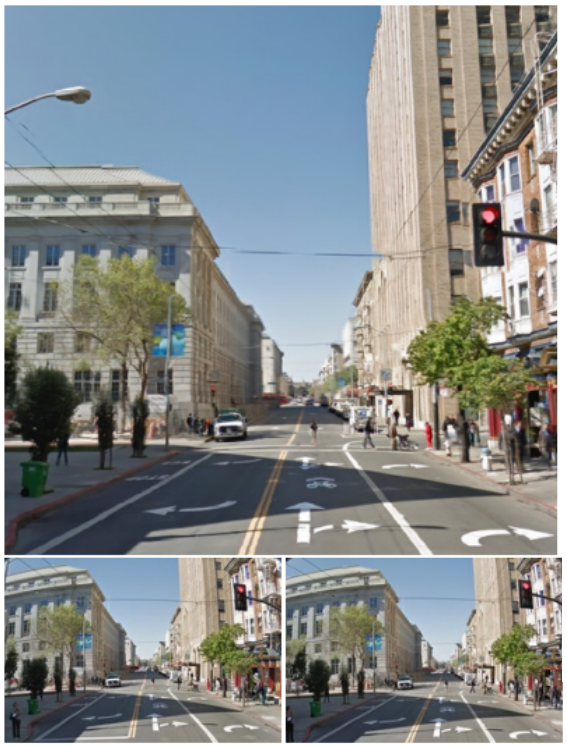
Image Processing: Depth, Motion, Odometry

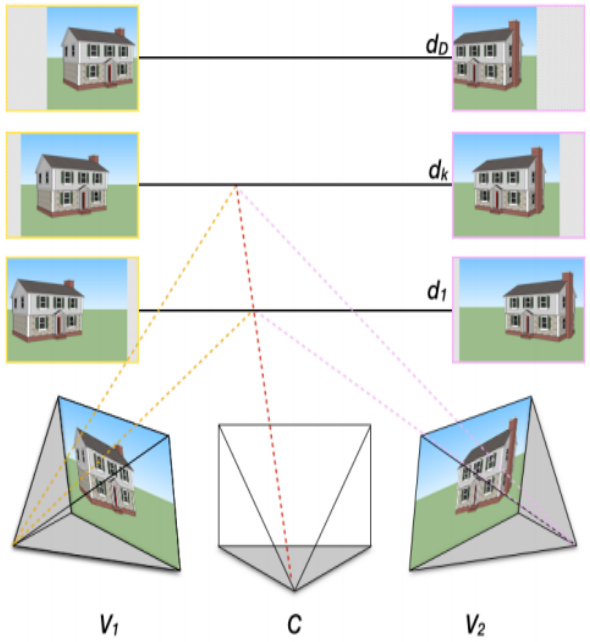
Stereo Matching

(Zbontar & LeCun 2014) best on KITTI benchmark



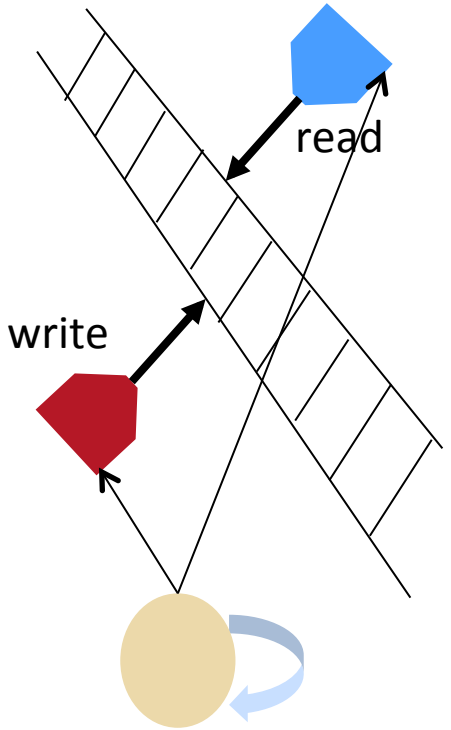
DeepStereo: Flynn, Neulander, Philbin, Snavely (2015)





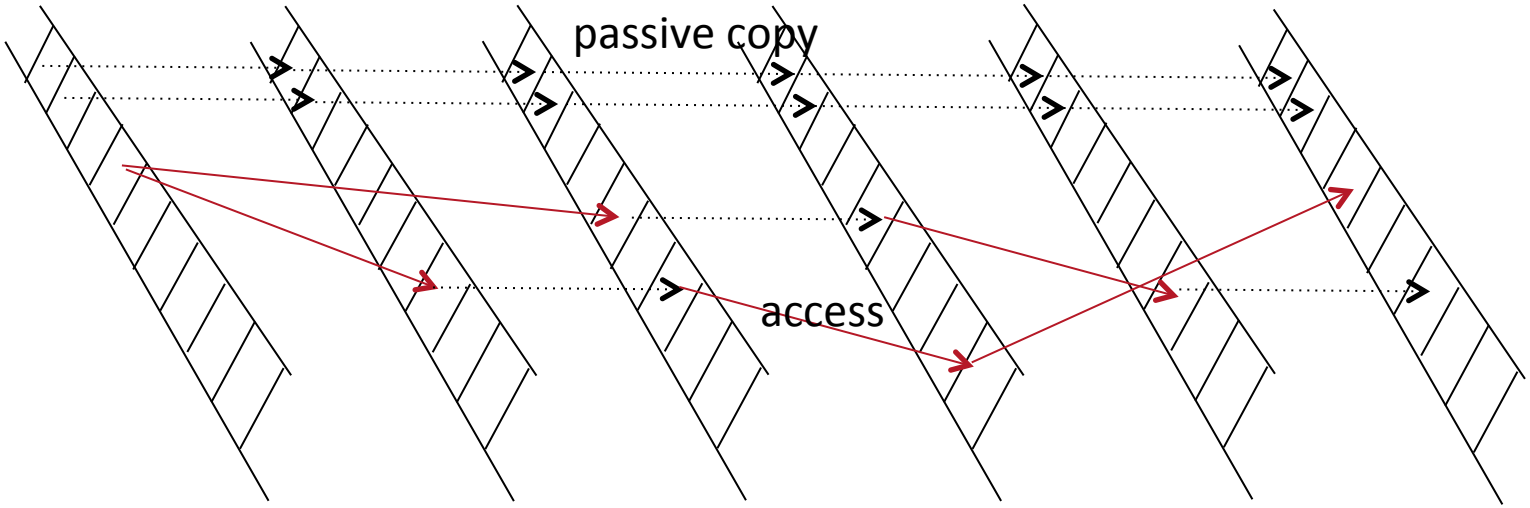
Attention Mechanisms for Memory Access enable REASONING

- Neural Turing Machines (Graves et al 2014)
- and Memory Networks (Weston et al 2014)
- Use a form of attention mechanism to control the read and write access into a memory
- The attention mechanism outputs a softmax over memory locations
- For efficiency, the softmax should be sparse (mostly 0's), e.g. maybe using a hash-table formulation.



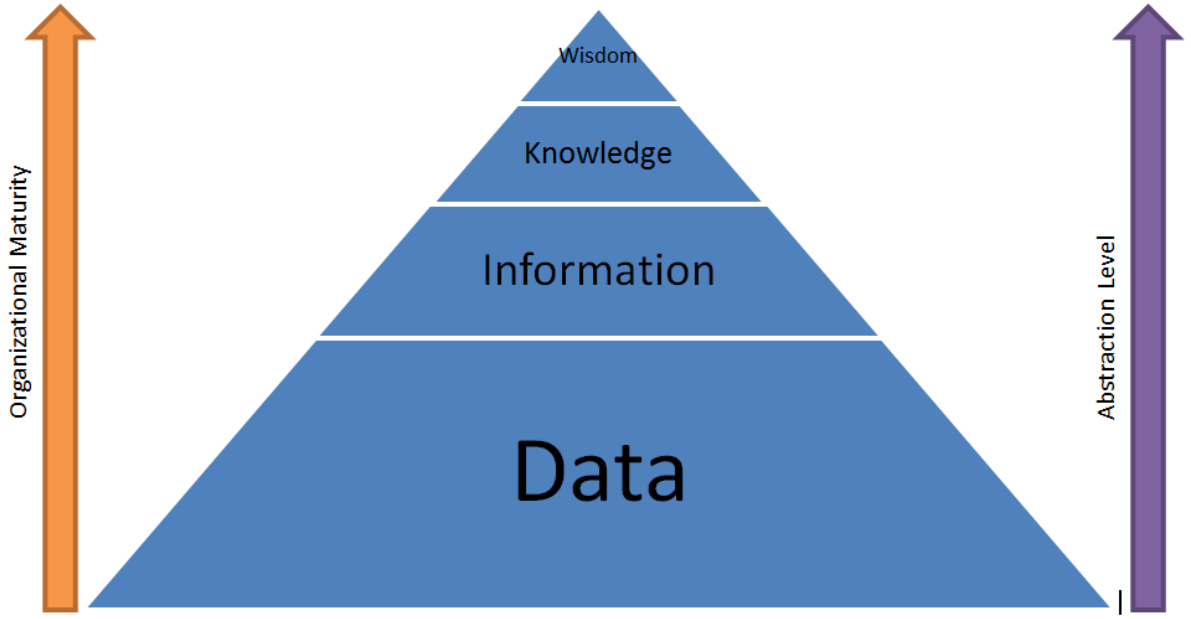
Sparse Access Memory for Long-Term Dependencies

- A mental state stored in an external memory can stay for arbitrarily long durations, until evoked for read or write
- Forgetting = vanishing gradient.
- Memory = larger state, avoiding the need for forgetting/vanishing
- Different « threads » can run in parallel if we view the memory as an associative one.



Learning Multiple Levels of Abstraction

- The big payoff of deep learning is to allow learning higher levels of abstraction
- Higher-level abstractions disentangle the factors of variation, which allows much easier generalization and transfer

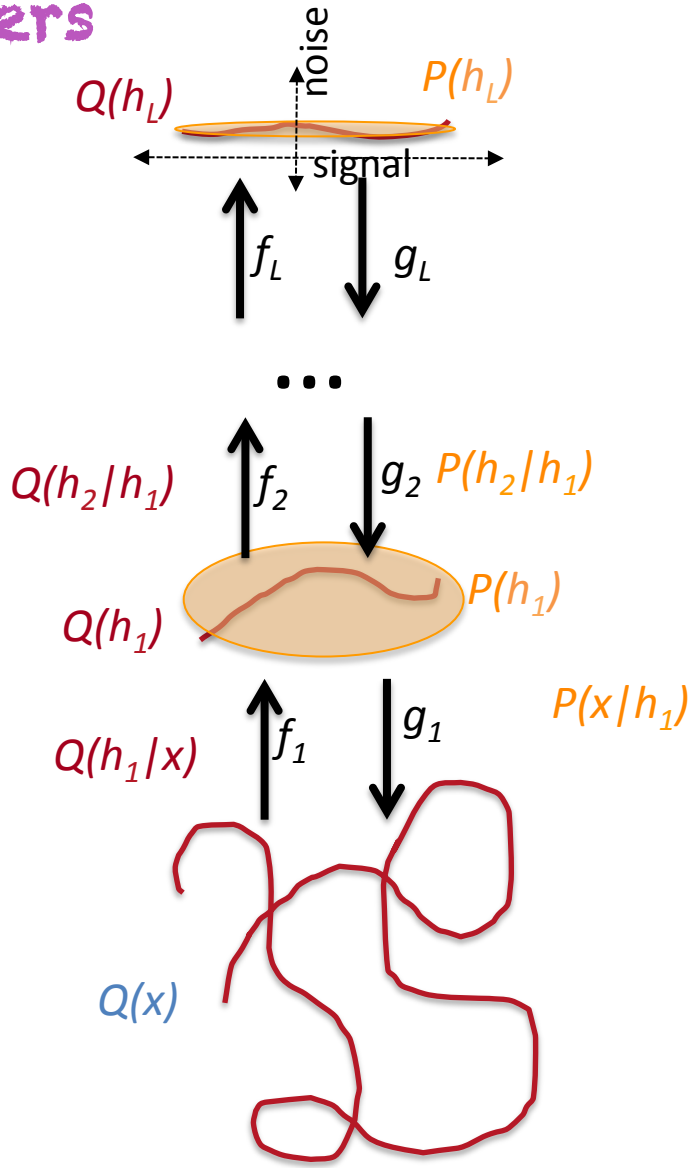


The Next Challenge: Unsupervised Learning

- Recent progress mostly in supervised DL
- Real technical challenges for unsupervised DL
- Potential benefits:
  - Exploit tons of unlabeled data
  - Answer new questions about the variables observed
  - Regularizer – transfer learning – domain adaptation
  - Easier optimization (local training signal)
  - Structured outputs

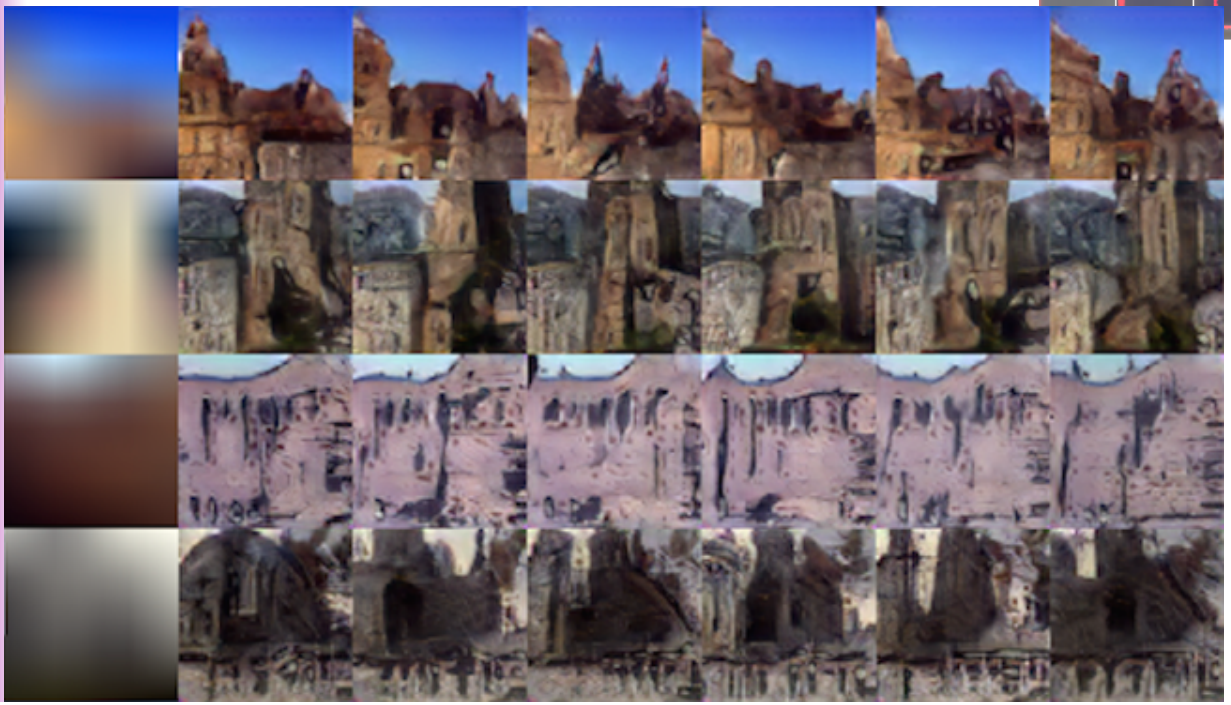
Extracting Structure By Gradual Disentangling and Manifold Unfolding & Variational Auto-Encoders

Each level transforms the data into a representation in which it is easier to model, unfolding it more, contracting the noise dimensions and mapping the signal dimensions to a factorized (uniform-like) distribution.



The Current SOTA in Generative Models of Images

DRAW, *(Gregor et al 2015)* based on variational auto-encoders *(Kingma et al ICLR'2014)*



Generative Adversarial Networks *(Goodfellow et al, NIPS'2014, Denton et al NIPS'2015)*

MILA: Montreal Institute for Learning Algorithms