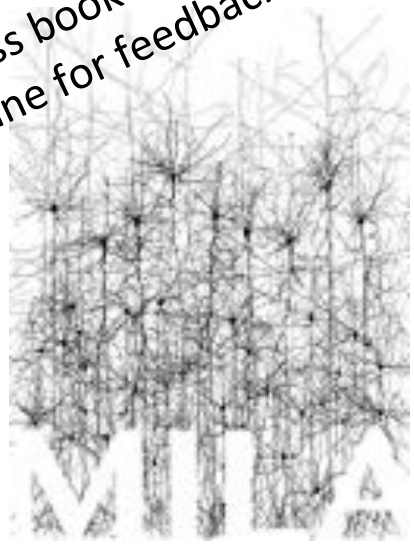
Deep Learning

Yoshua Bengio

November 6, 2015

ACPR'2015, Kuala Lumpur

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

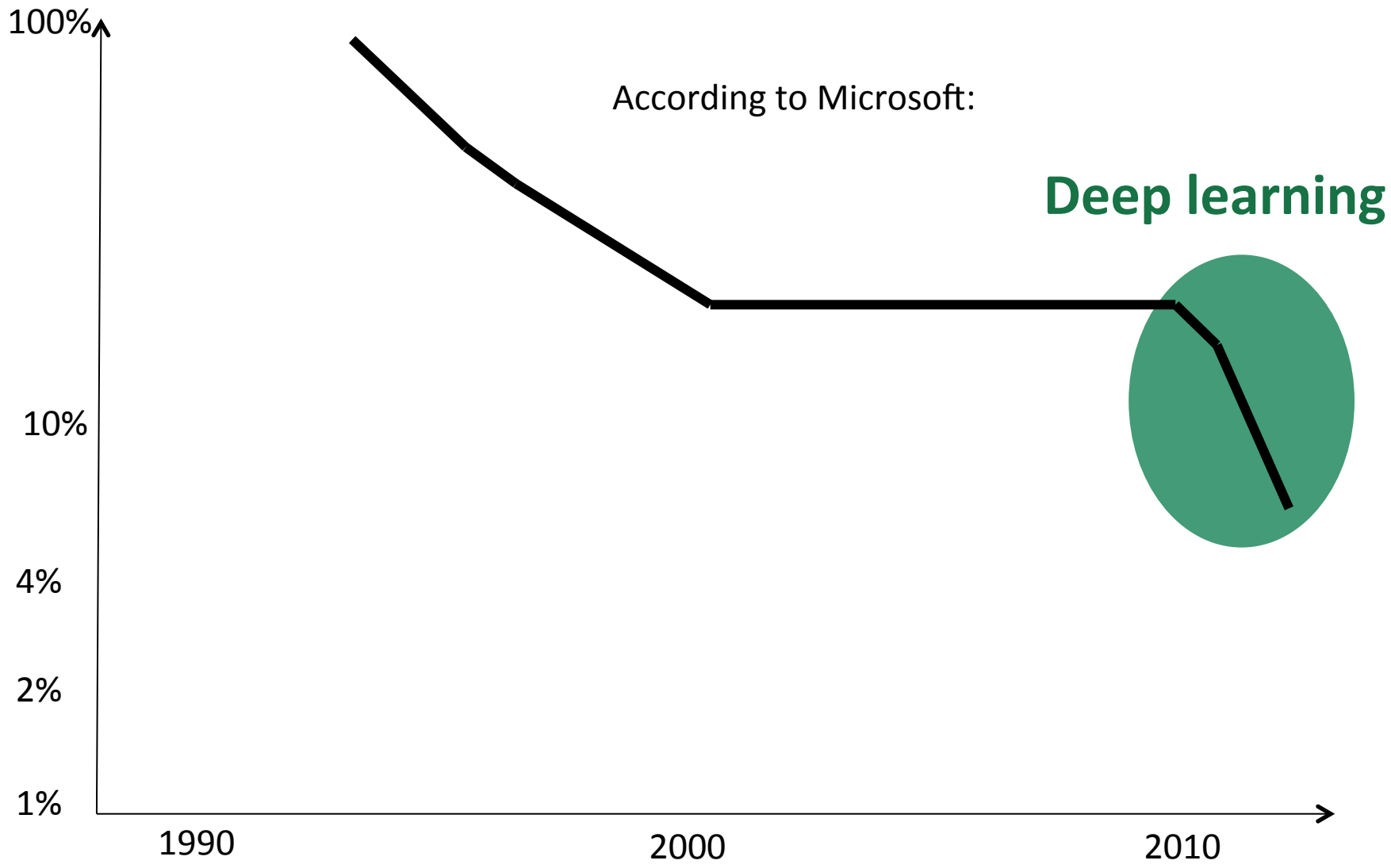
Initial Breakthrough in 2006

Canadian initiative: CIFAR

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

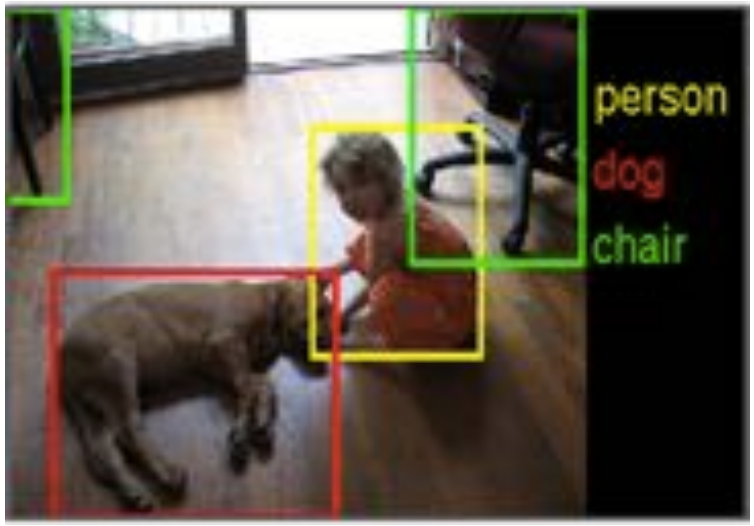
- Unsupervised feature learners:
  - RBMs
  - Auto-encoder variants
  - Sparse coding variants

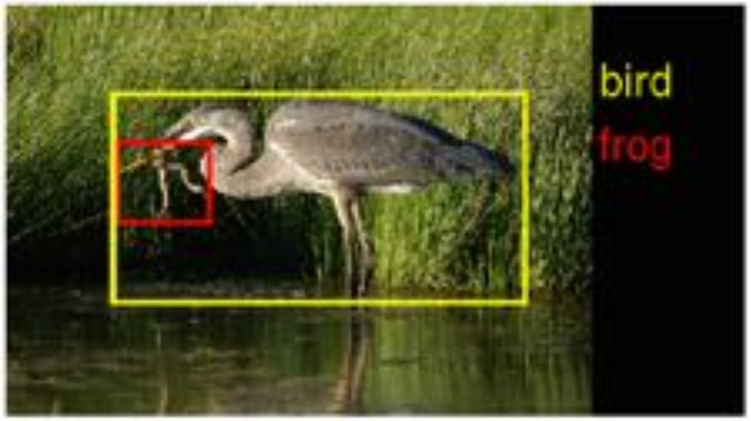
2010-2012: Breakthrough in speech recognition → in Androids by 2012



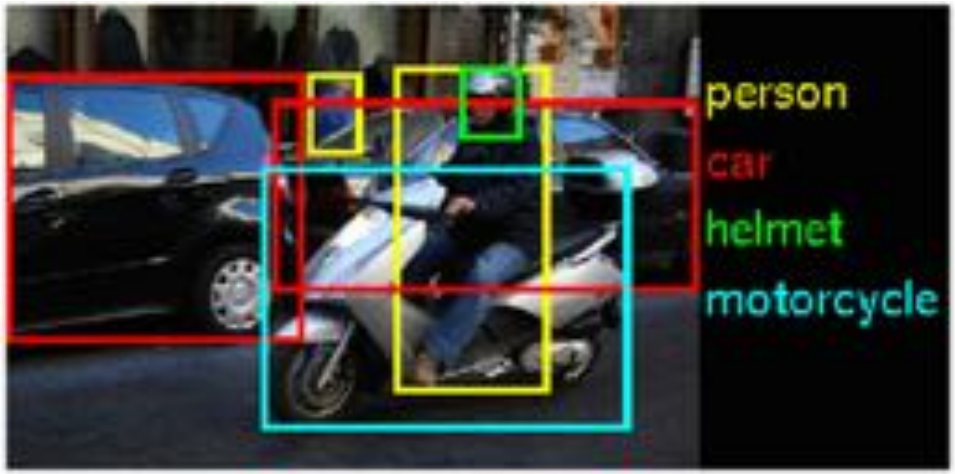
Breakthrough in computer vision: 2012-2015

- 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

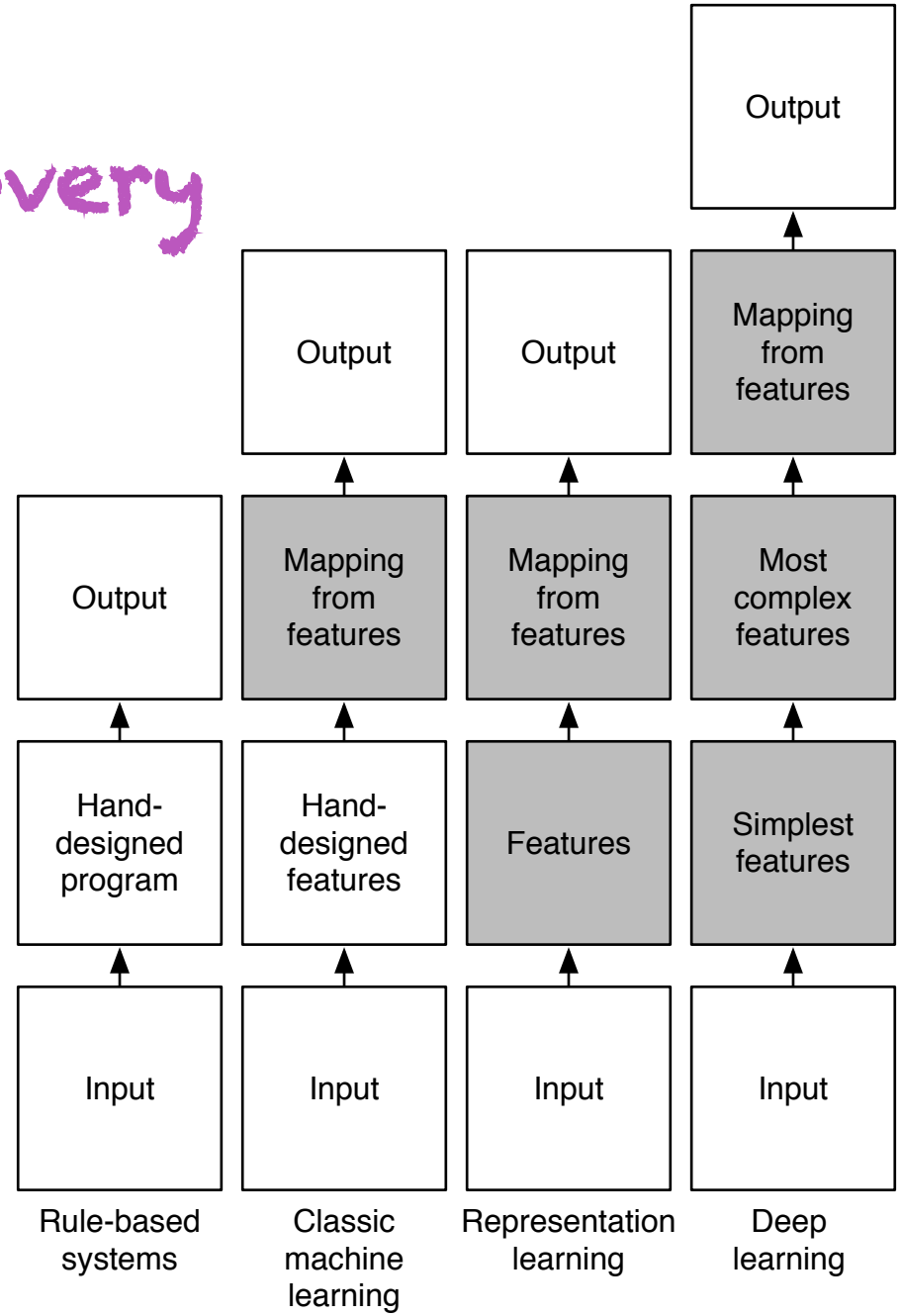
Google Acquires Artificial Intelligence Startup DeepMind For More Than $500M

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

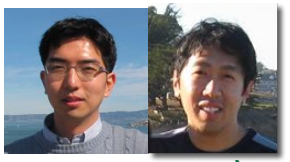
IT Companies are Racing into Deep Learning

Why is Deep Learning Working so Well?

Automating Feature Discovery



Learning multiple levels of representation

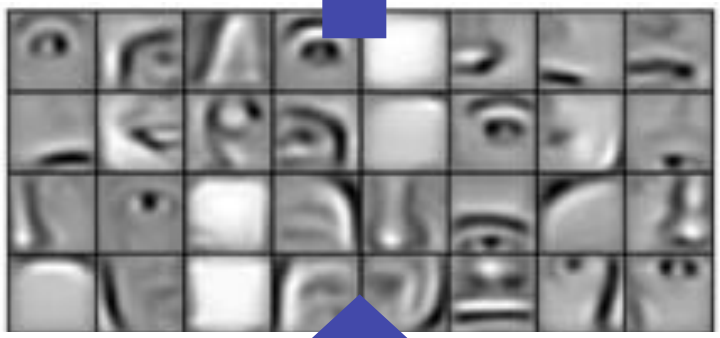


(Lee, Largman, Pham & Ng, NIPS 2009)
(Lee, Grosse, Ranganath & Ng, ICML 2009)

Successive model layers learn deeper intermediate representations

Layer 3

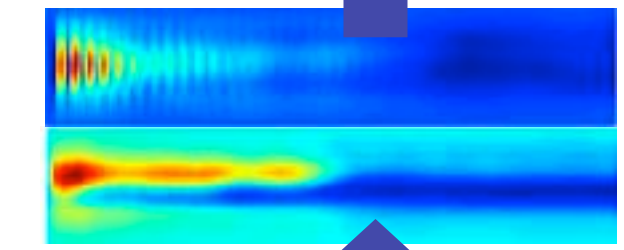
Parts combine to form objects

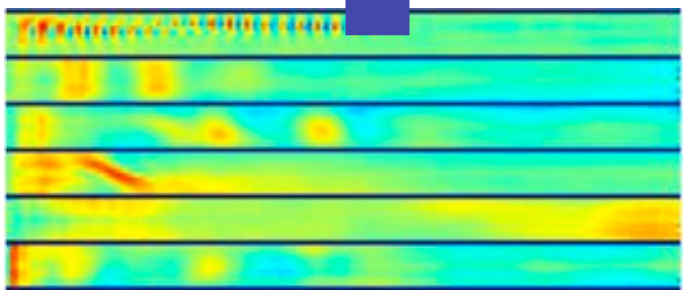


Layer 2

Layer 1

High-level linguistic representations





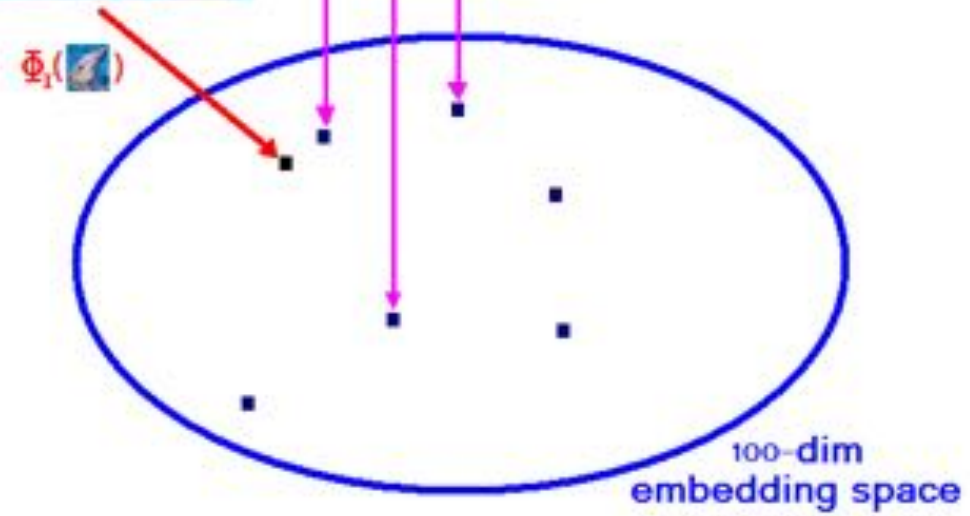
Prior: underlying factors & concepts compactly expressed w/ multiple levels of abstraction

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.

Machine Learning, AI & No Free Lunch

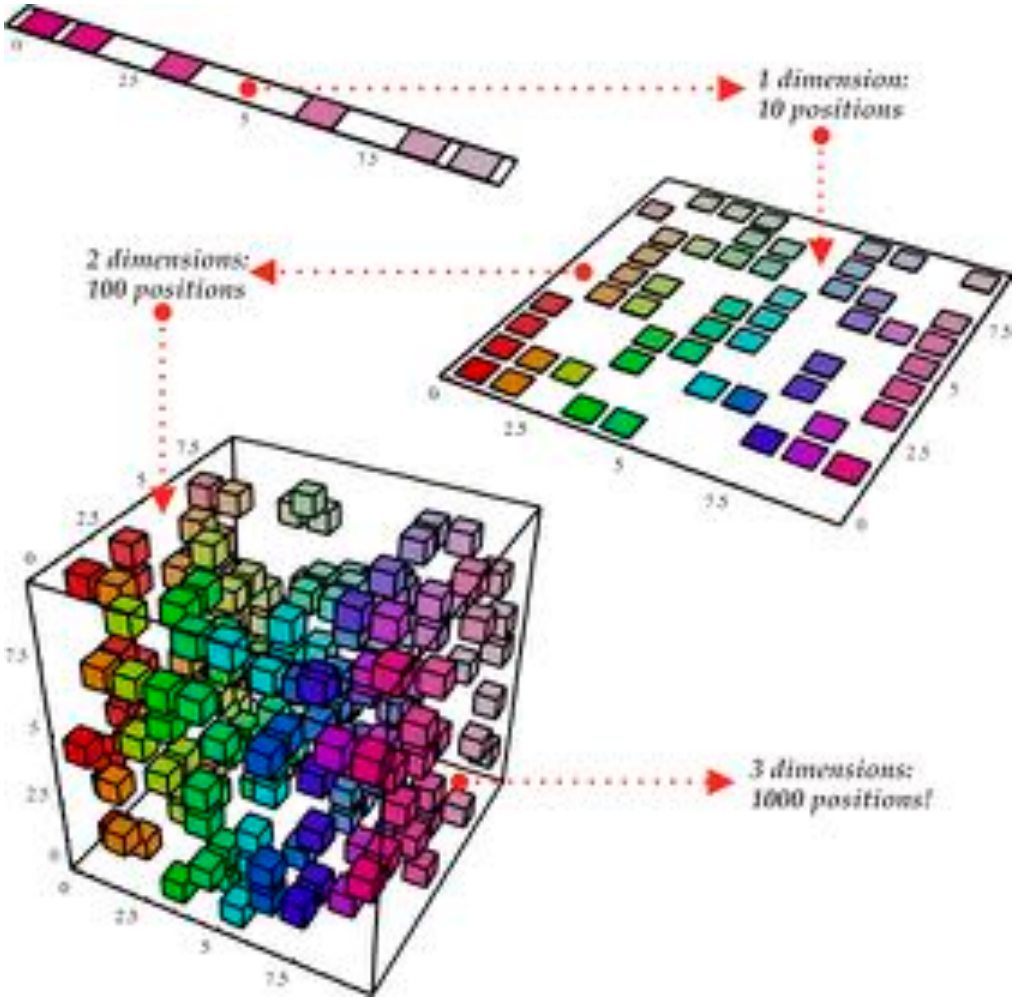
- Four key ingredients for ML towards AI

  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

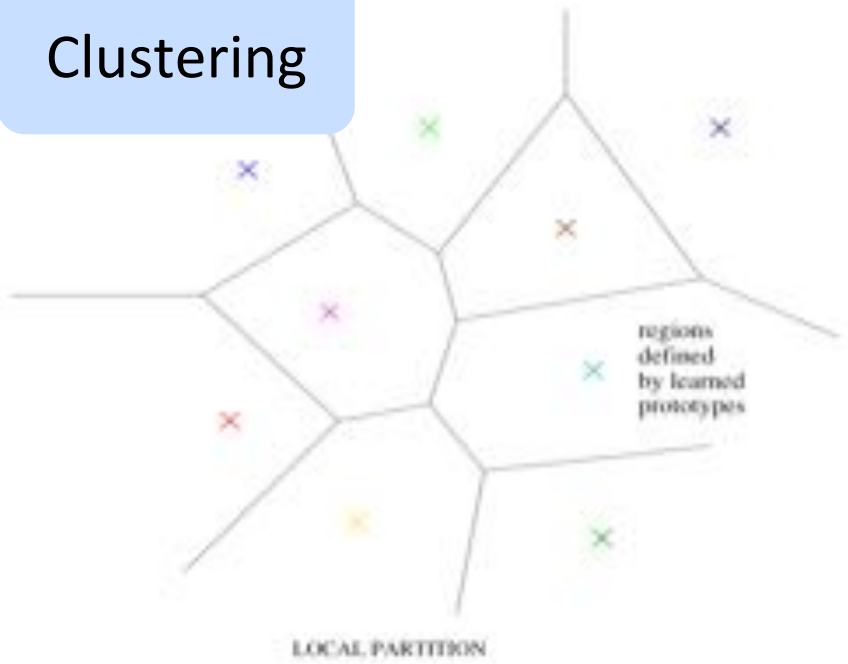
Distributed representations / embeddings: feature learning

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**

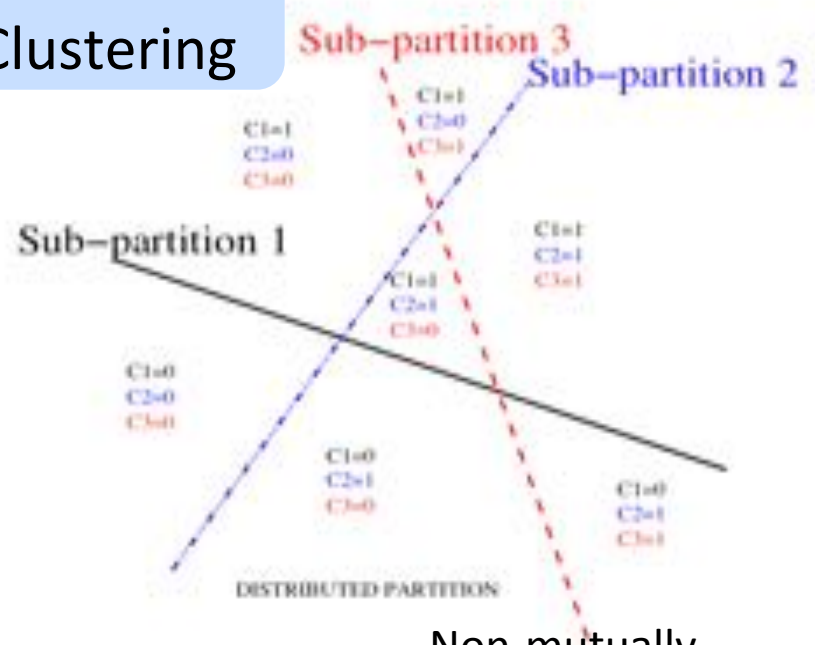
→ No non-trivial generalization to regions without examples

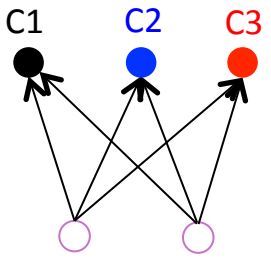
The need for distributed representations

- 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**

Multi-Clustering





Non-mutually exclusive features/ attributes create a combinatorially large set of distinguiable configurations

The Depth Prior can be Exponentially Advantageous

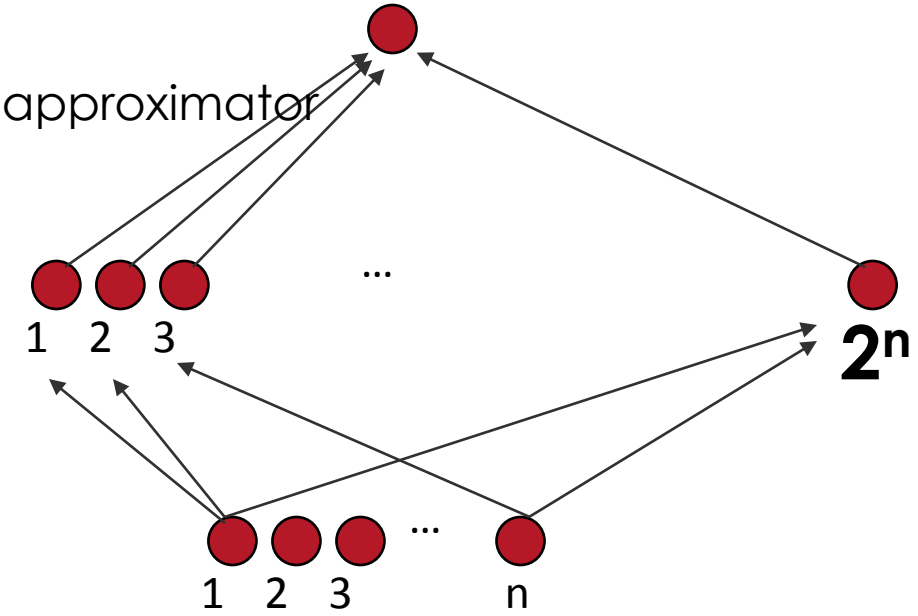
Theoretical arguments:

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

RBMs & auto-encoders = universal approximator

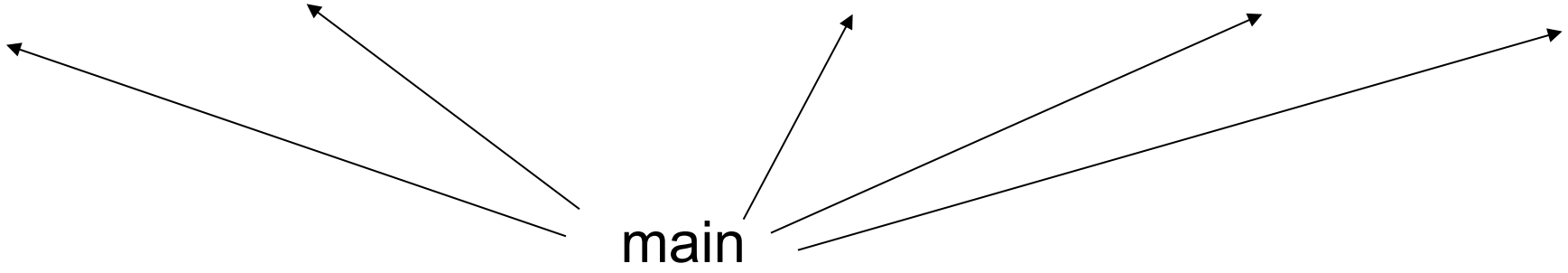
Theorems on advantage of depth:
(Hastad et al 86 & 91, Bengio et al 2007, Bengio & Delalleau 2011, Braverman 2011, Pascanu et al 2014, Montufar et al **NIPS 2014**)

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

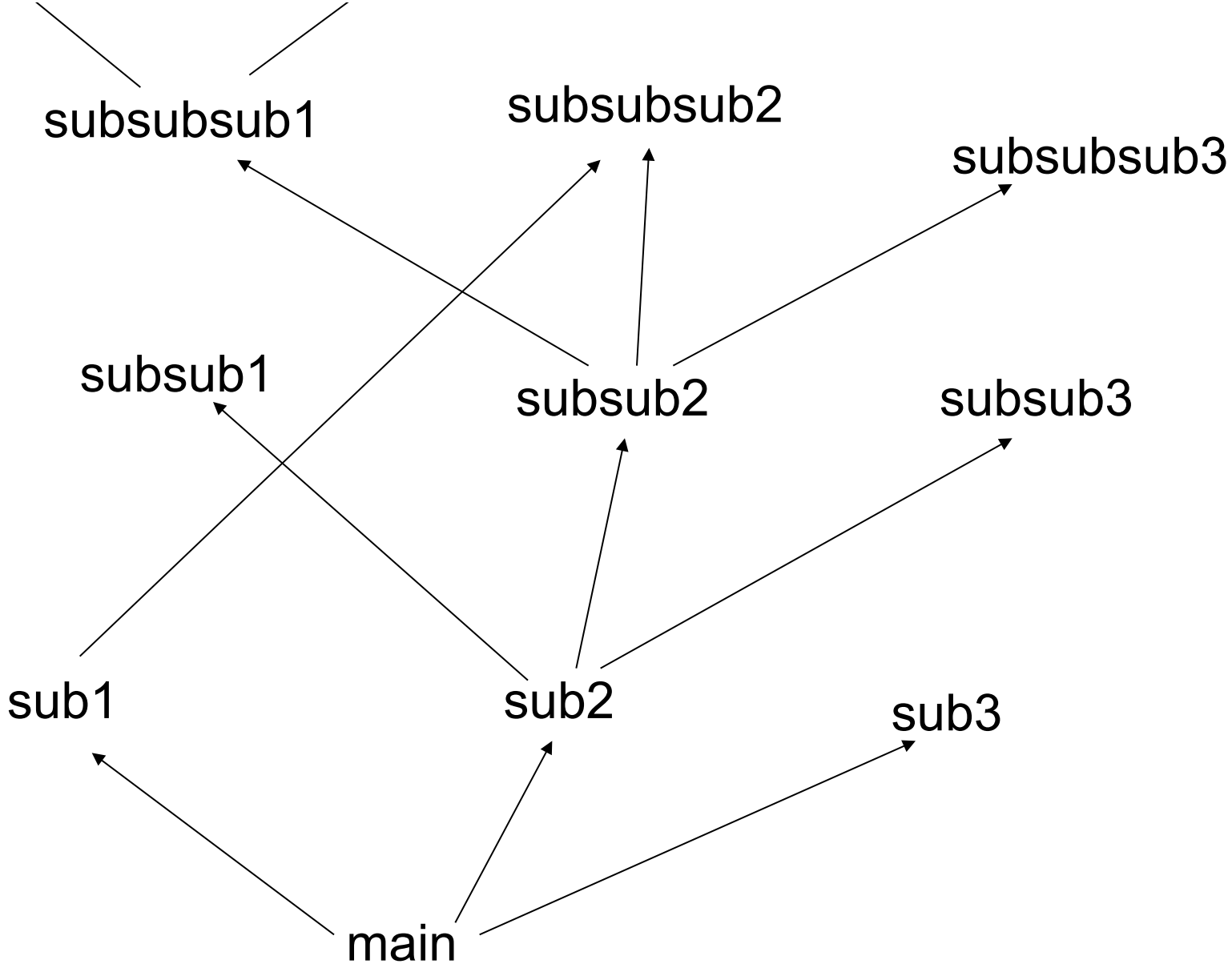


subroutine1 includes
subsub1 code and
subsub2 code and
subsubsub1 code

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



“Shallow” computer program



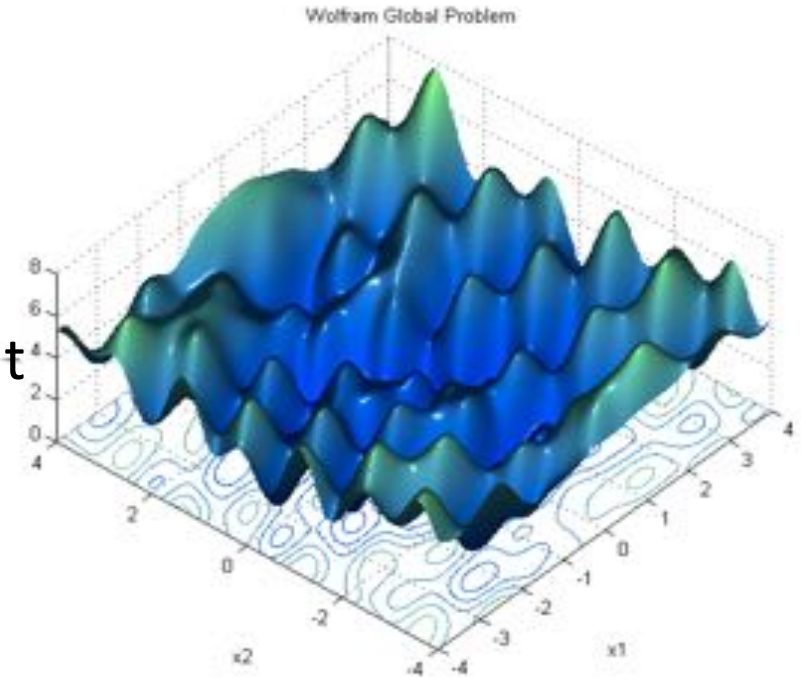
“Deep” computer program

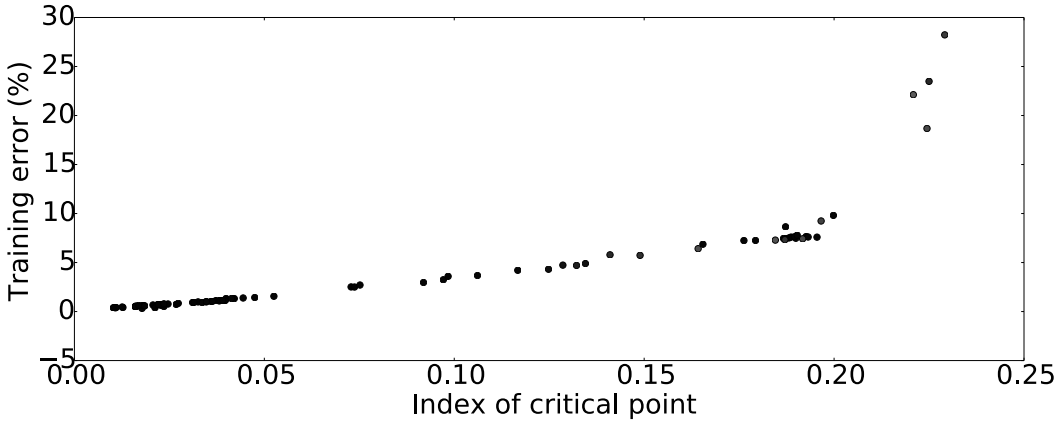
A Myth is Being Debunked: Local Minima in Neural Nets → Convexity is not needed

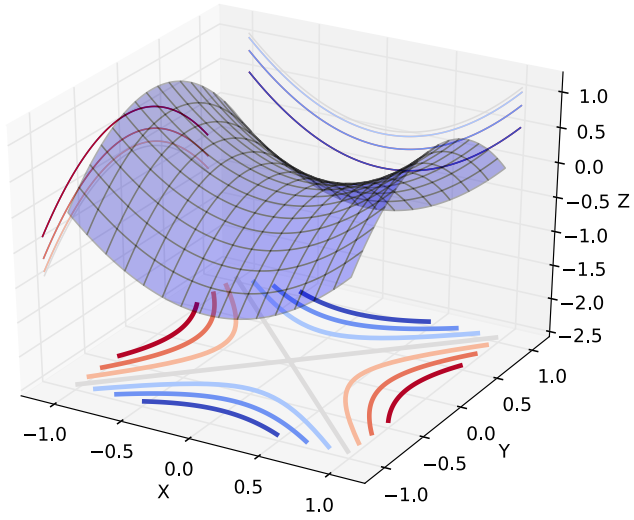
- (Pascanu, Dauphin, Ganguli, Bengio, arXiv May 2014): *On the saddle point problem for non-convex optimization*
- (Dauphin, Pascanu, Gulcehre, Cho, Ganguli, Bengio, NIPS' 2014): *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*
- (Choromanska, Henaff, Mathieu, 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)





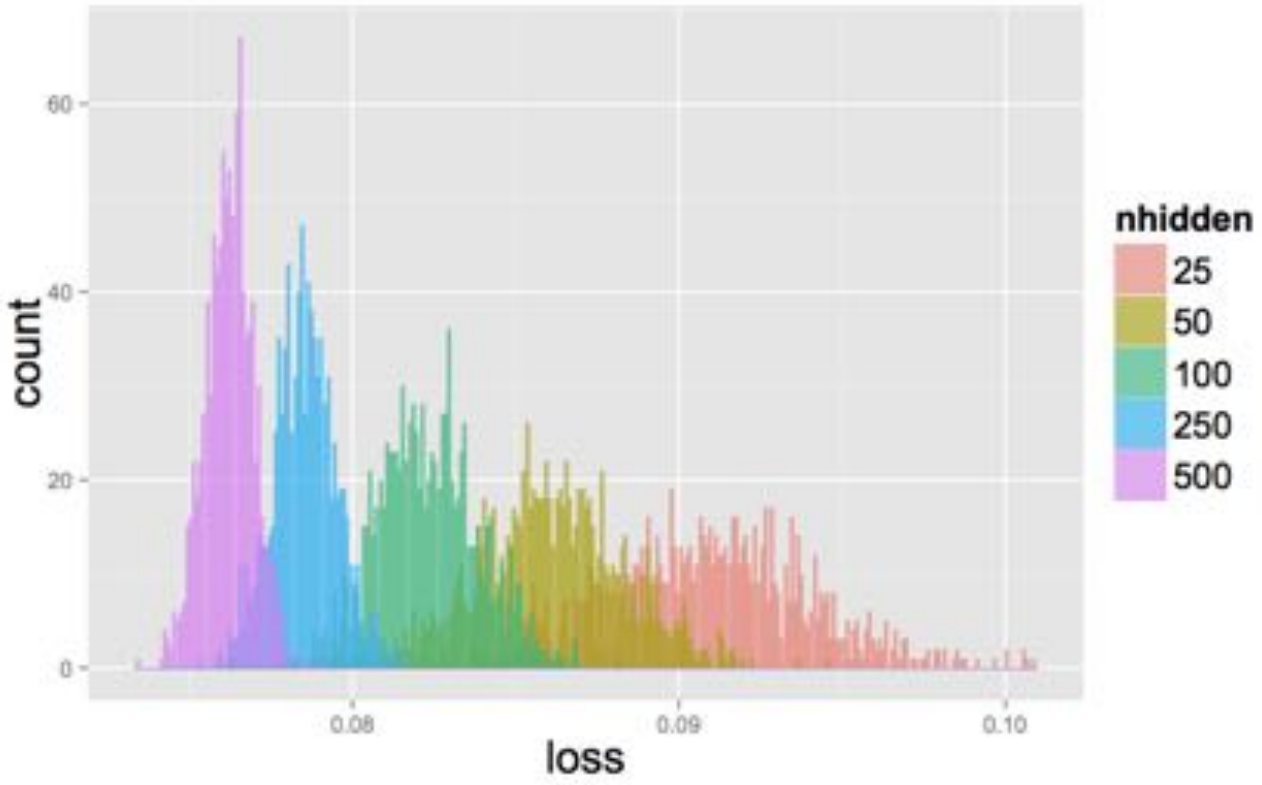


Low Index Critical Points

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



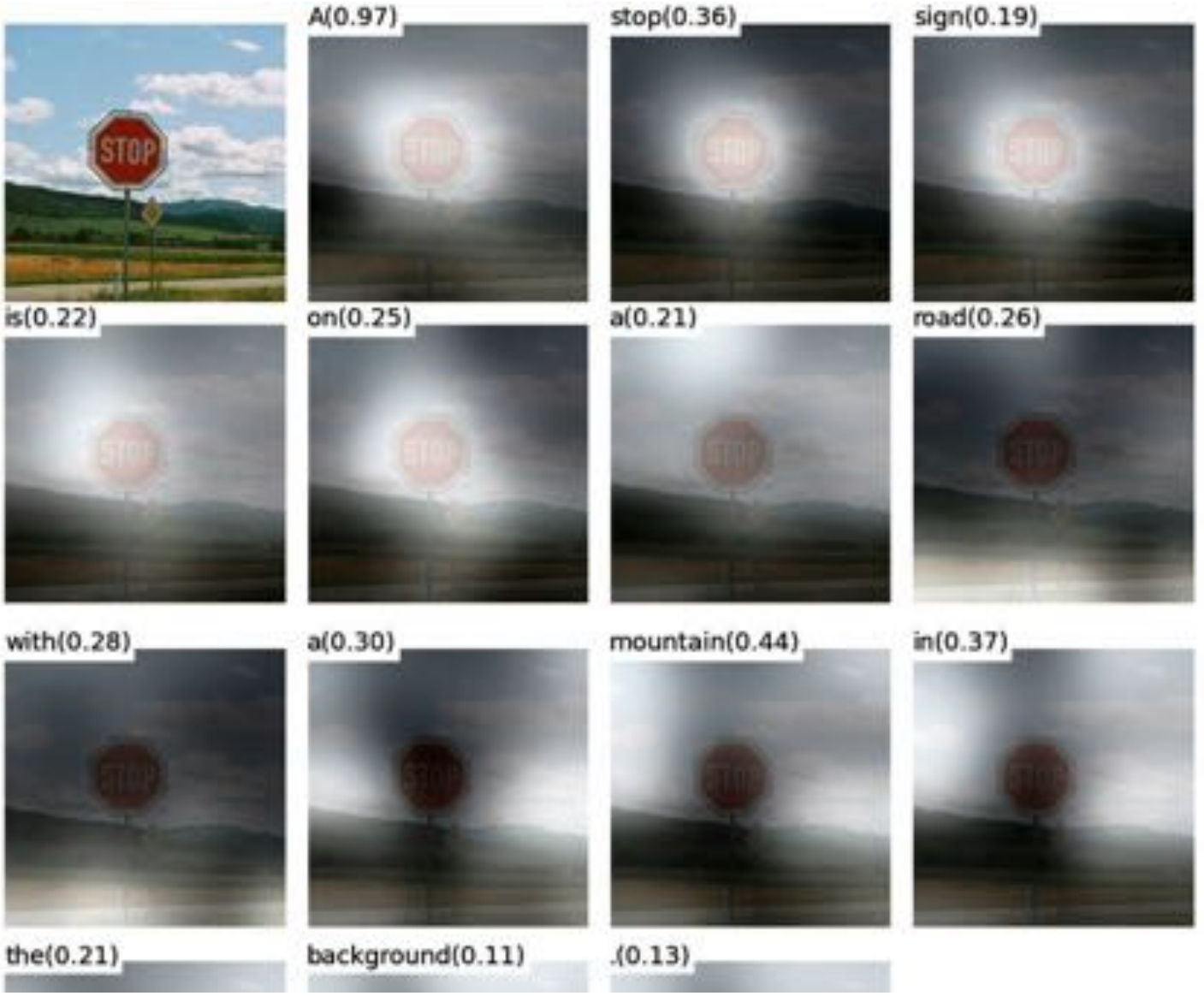
Deep Learning: Beyond Pattern Recognition, towards AI

- 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)

Speaking about what one sees

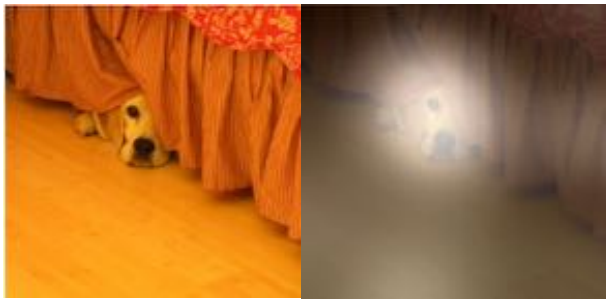


Ongoing Progress: Combining Vision and Natural Language Understanding

- 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.

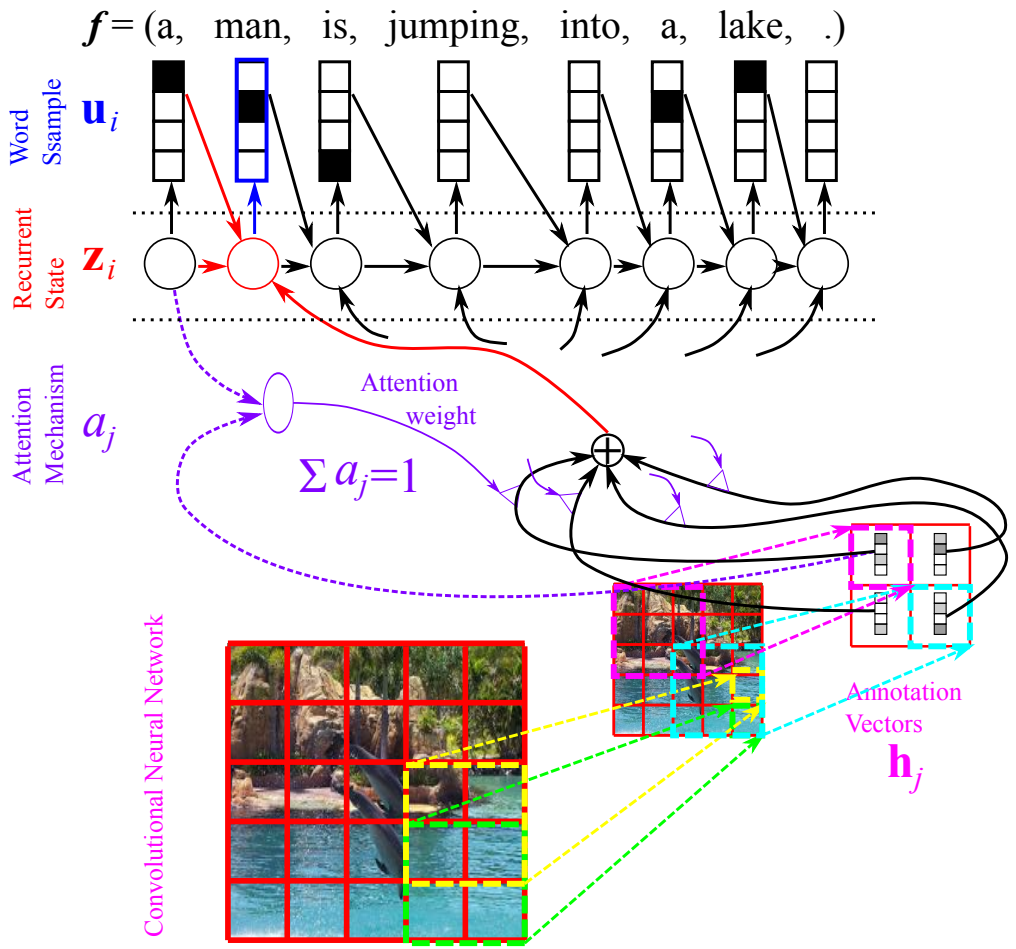
A stop sign is on a road with a mountain in the background.

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

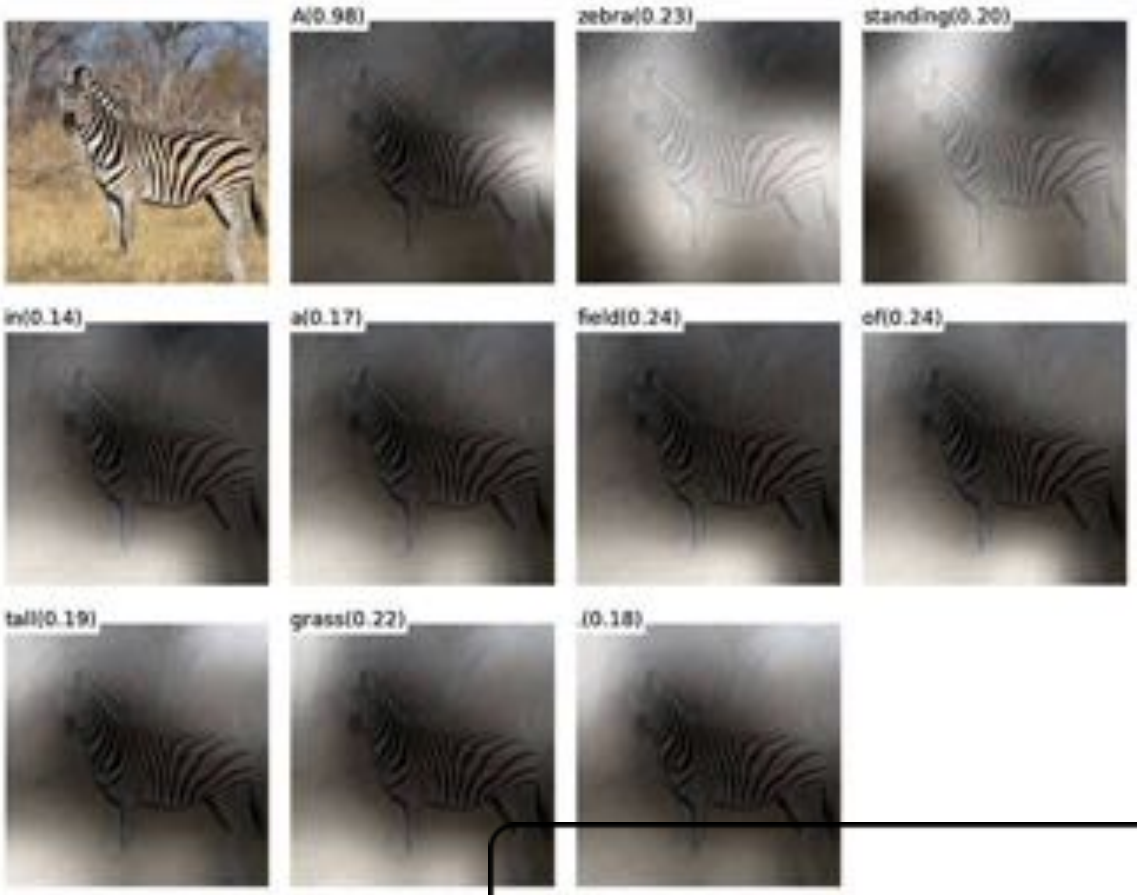
A group of people sitting on a boat in the water.

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

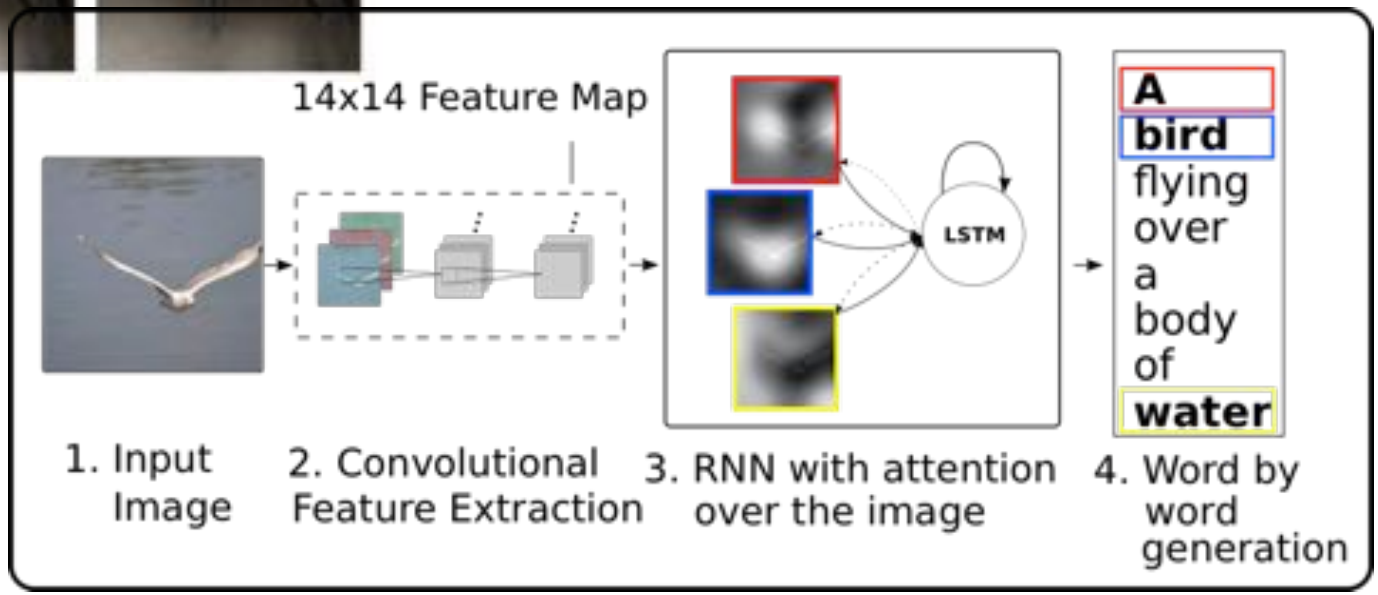
Image-to-Text: Caption Generation



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



Paying Attention to Selected Parts of the Image While Uttering Words



Still some way to go...

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and
a hat on a skateboard.

A person is standing on a beach
with a surfboard.

A woman is sitting at a table
with a large pizza.

A man is talking on his cell phone
while another man watches.

The Next Frontier: Reasoning and Question Answering

- Currently working on artificial tasks, with memory networks:

Joe went to the garden then Fred picked up the milk; Joe moved to the bathroom and Fred dropped the milk, and then Dan moved to the living_room.
Where is Dan? A: living room I believe
Where is Joe? A: the bathroom
Fred moved to the bedroom and Joe went to the kitchen then Joe took the milk there and Dan journeyed to the bedroom; Joe discarded the milk.
Where is the milk now ? A: the milk is in the kitchen
Where is Dan now? A: I think he is in the bedroom
Joe took the milk there, after that Mike travelled to the office, then Joe went to the living_room, next Dan went back to the kitchen and Joe travelled to the office.
Where is Joe now? A: I think Joe is in the office

From "Memory Networks", Weston et al. ICLR 2015;
"End-to-end memory networks", Sukhbatar et al. NIPS'2015

Sam walks into the kitchen.	Brian is a lion.	Mary journeyed to the den.
Sam picks up an apple.	Julius is a lion.	Mary went back to the kitchen.
Sam walks into the bedroom.	Julius is white.	John journeyed to the bedroom.
Sam drops the apple.	Bernhard is green.	Mary discarded the milk.
Q: Where is the apple?	Q: What color is Brian?	Q: Where was the milk before the den?
A. Bedroom	A. White	A. Hallway

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?

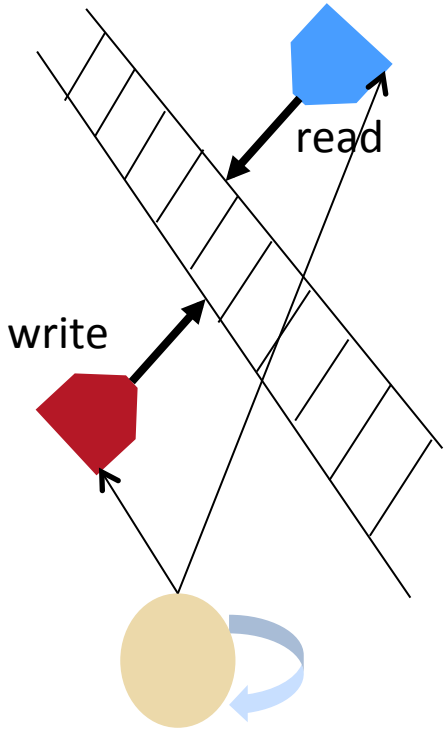
How many slices of pizza are there?
Is this a vegetarian pizza?

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

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

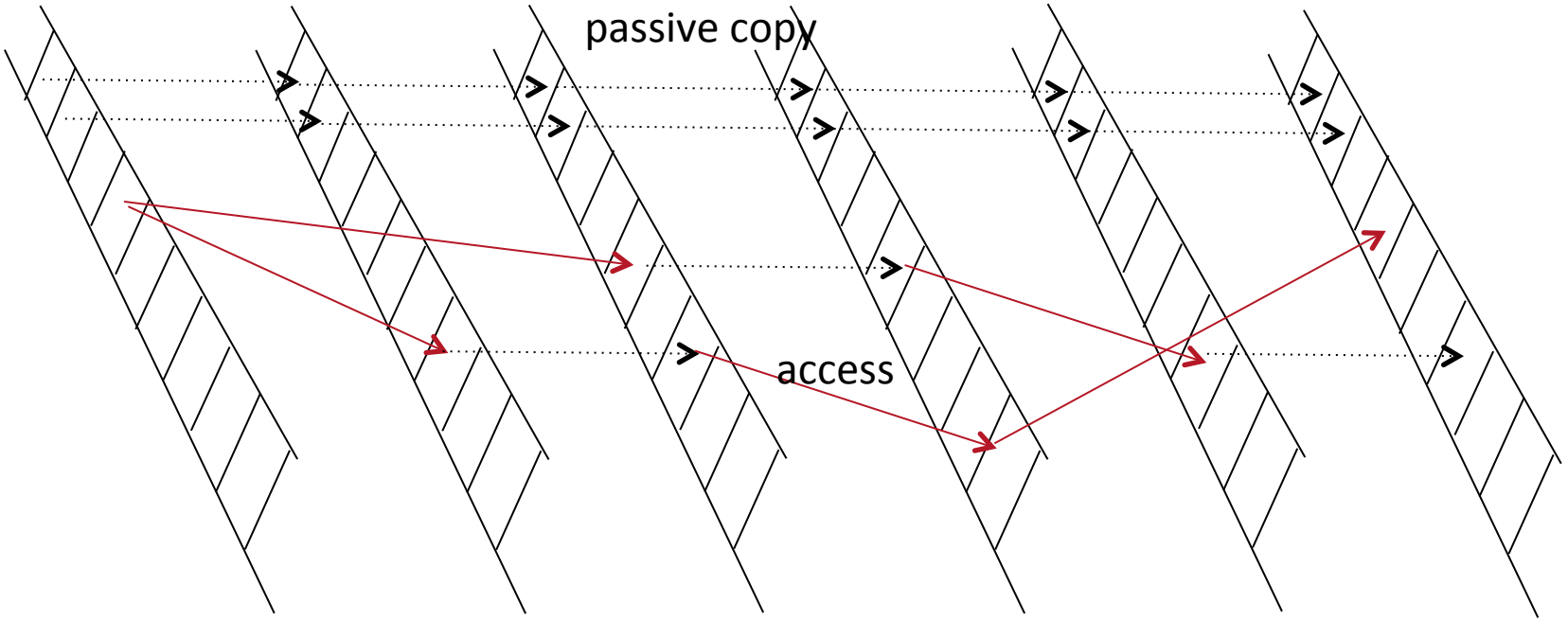
Attention Mechanisms for Memory Access Enable Reasoning

- 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



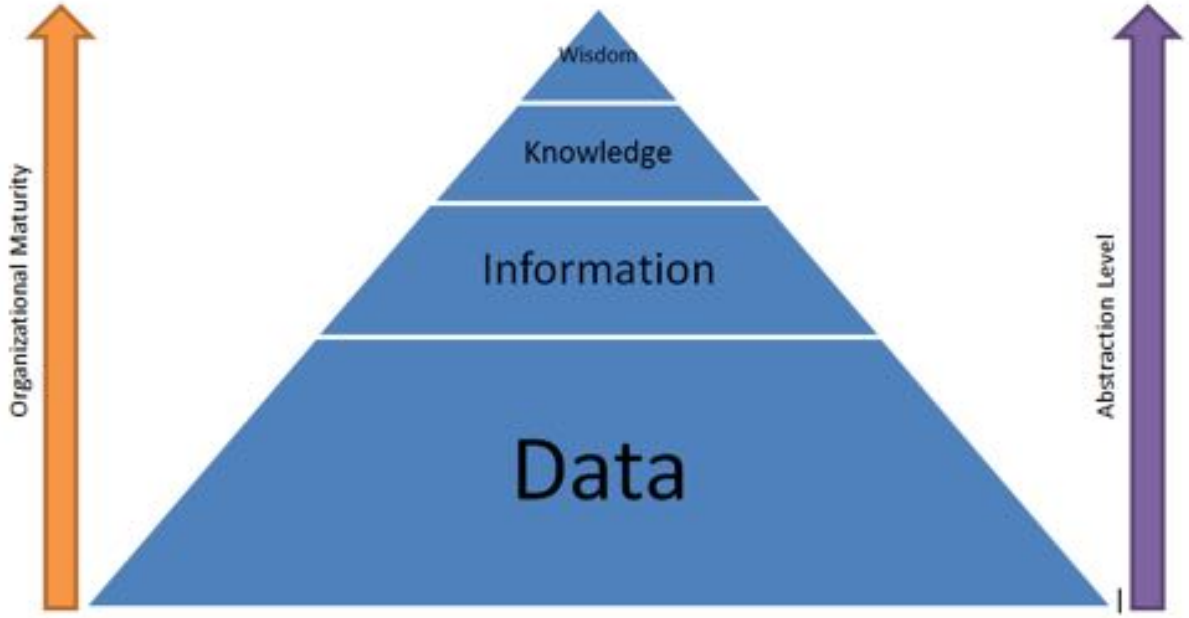
Why does it work? Pushing off the Curse of Long-Term Dependencies

- Whereas LSTM memories always decay exponentially (even if slowly), a mental state stored in an external memory can stay for arbitrarily long durations, until overwritten.



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

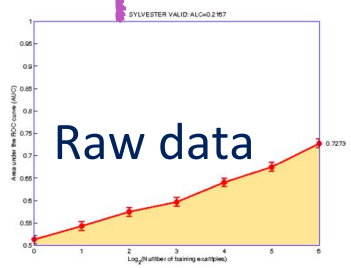


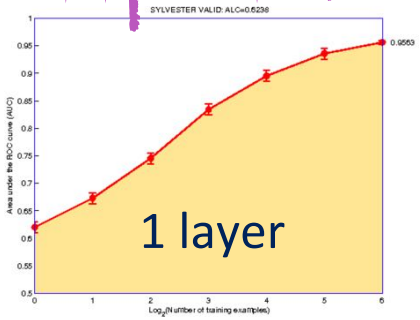
How do humans generalize from very few examples?

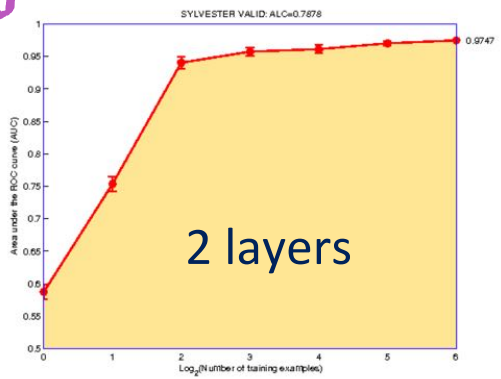
- Intelligence (good generalization) needs knowledge
- Humans **transfer** knowledge from **previous learning**:
  - Representations
  - Explanatory factors
- Previous learning from: **unlabeled data**

  \+ labels for other tasks

Unsupervised and Transfer Learning Challenge + Transfer Learning Challenge: Won by Unsupervised Deep Learning



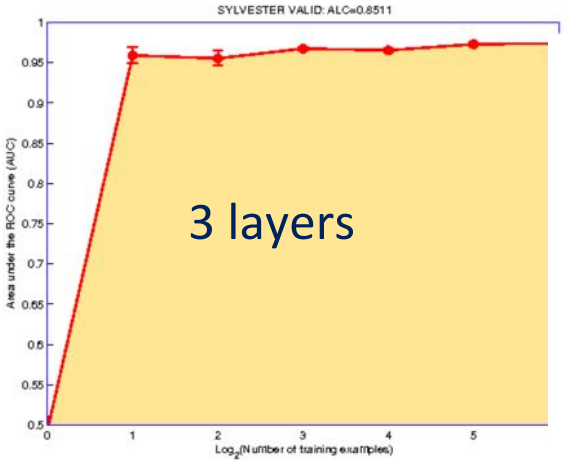


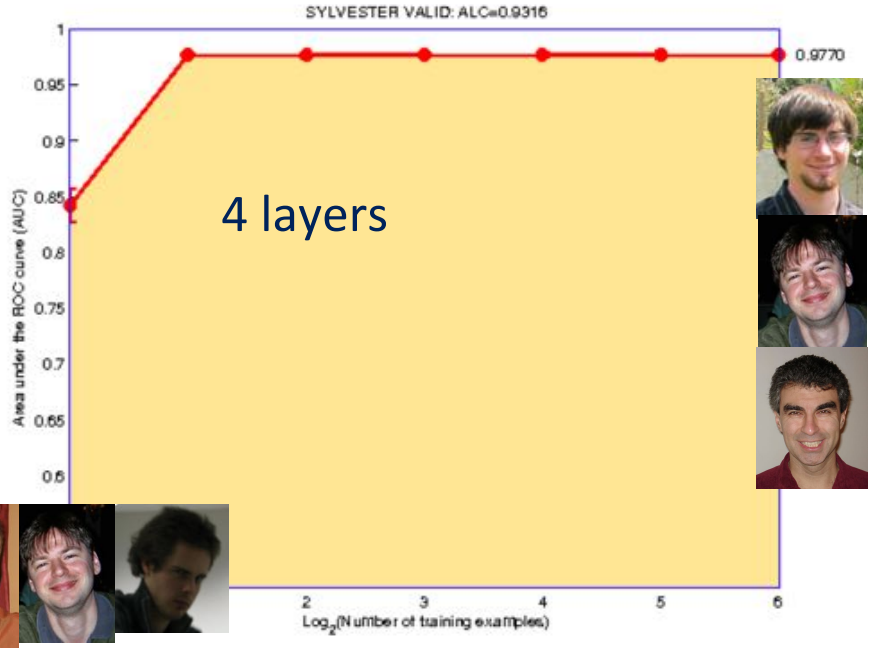


NIPS'2011 Transfer Learning Challenge

Paper: ICML'2012

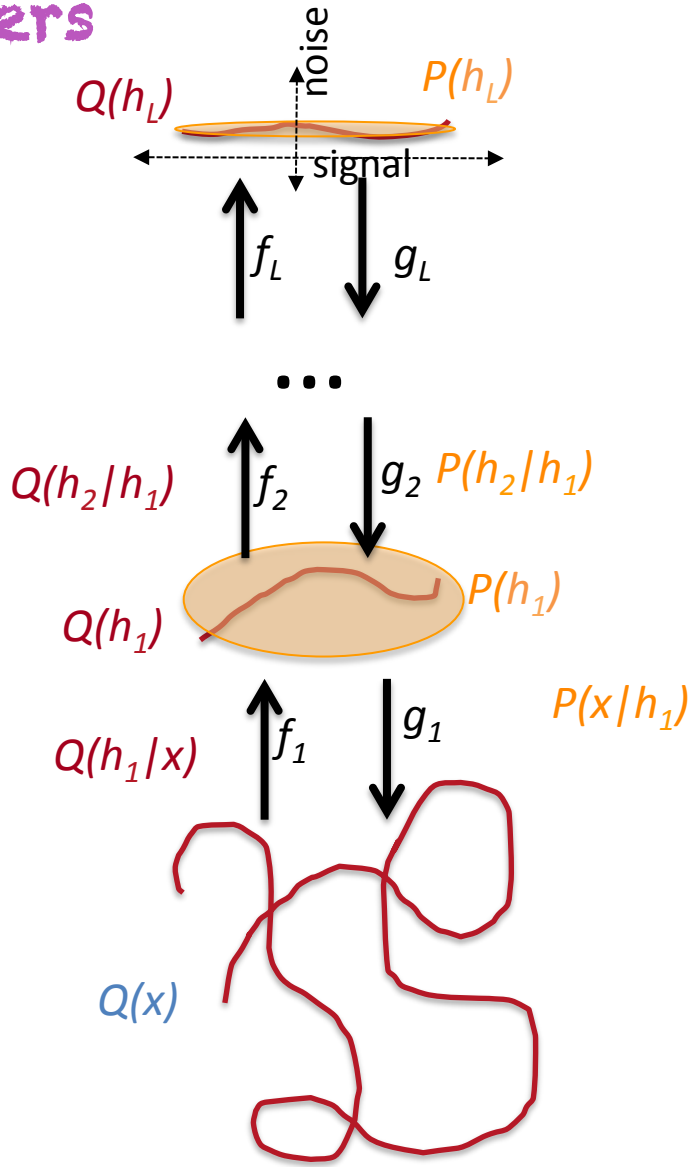
ICML'2011 workshop on Unsup. & Transfer Learning





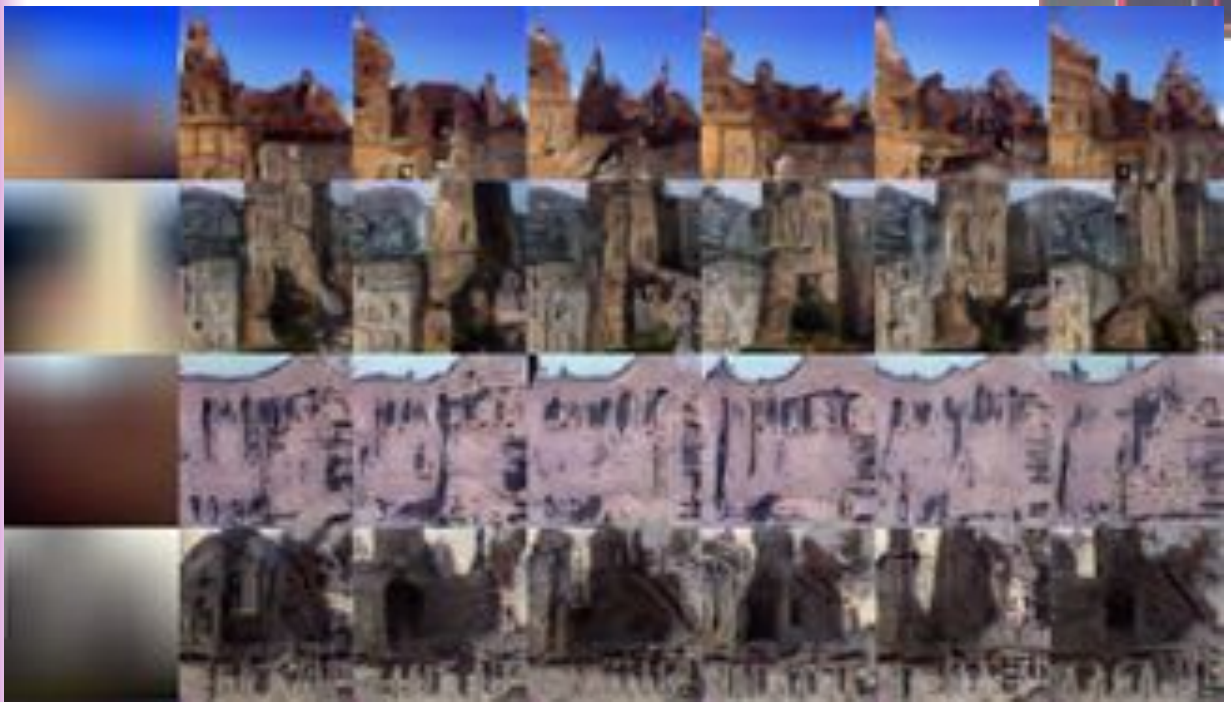
Extracting Structure By Gradual Disentangling and Manifold Unfolding & Variational Auto-Encoders

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