
Deep Learning for AI

AGI'2014
Québec, Canada

Yoshua Bengio

August 2, 2014

Université 
de Montréal



Ultimate Goal

- **Understand the principles giving rise to intelligence**

Focus

- **Learning: mathematical and computational principles allowing one to learn from examples in order to acquire knowledge**

Breakthrough

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

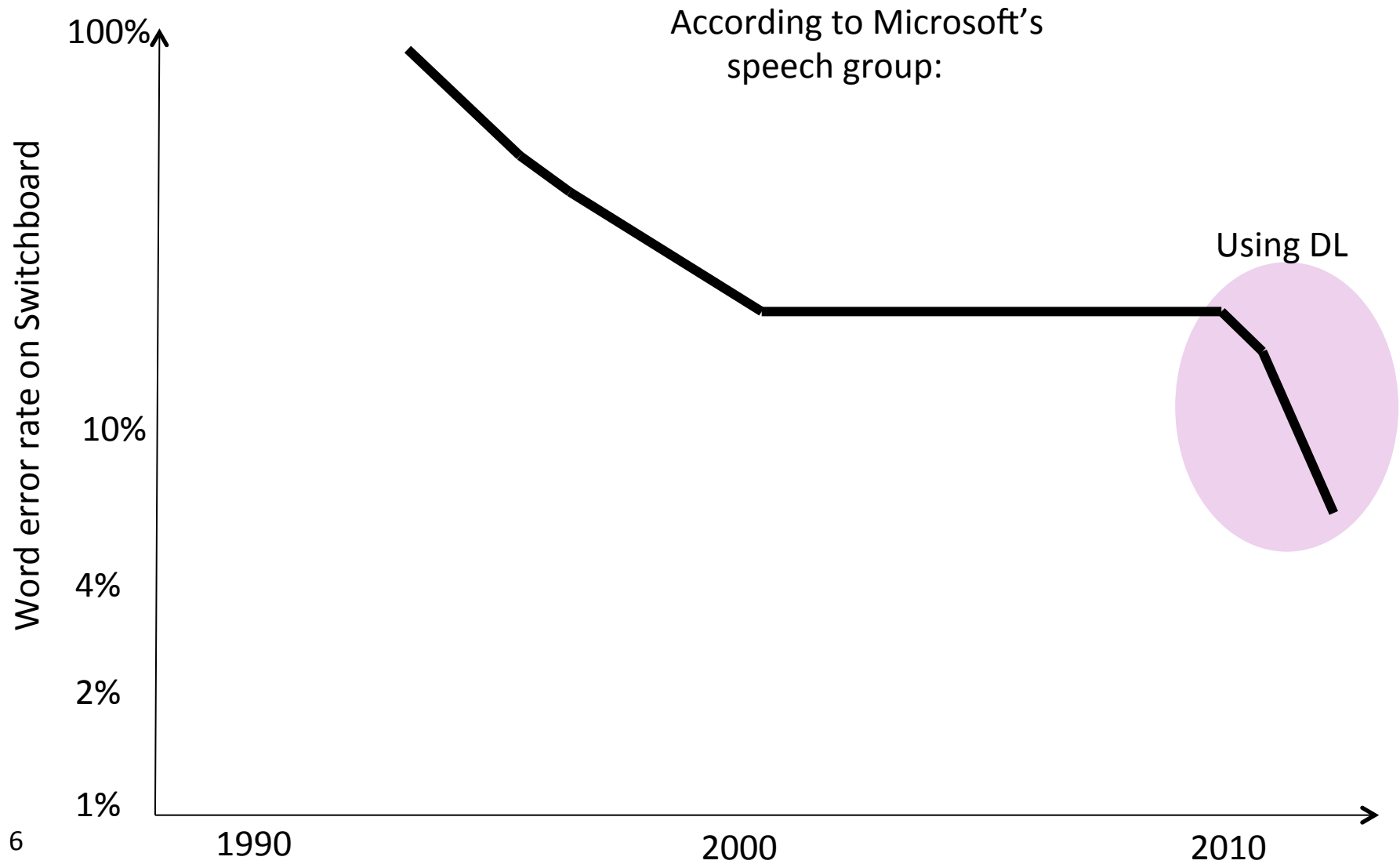
Impact

Deep learning has revolutionized

- **Speech recognition**
- **Object recognition**

More coming, including other areas of computer vision, NLP, dialogue, reinforcement learning...

The dramatic impact of Deep Learning on Speech Recognition



Object Recognition Breakthrough



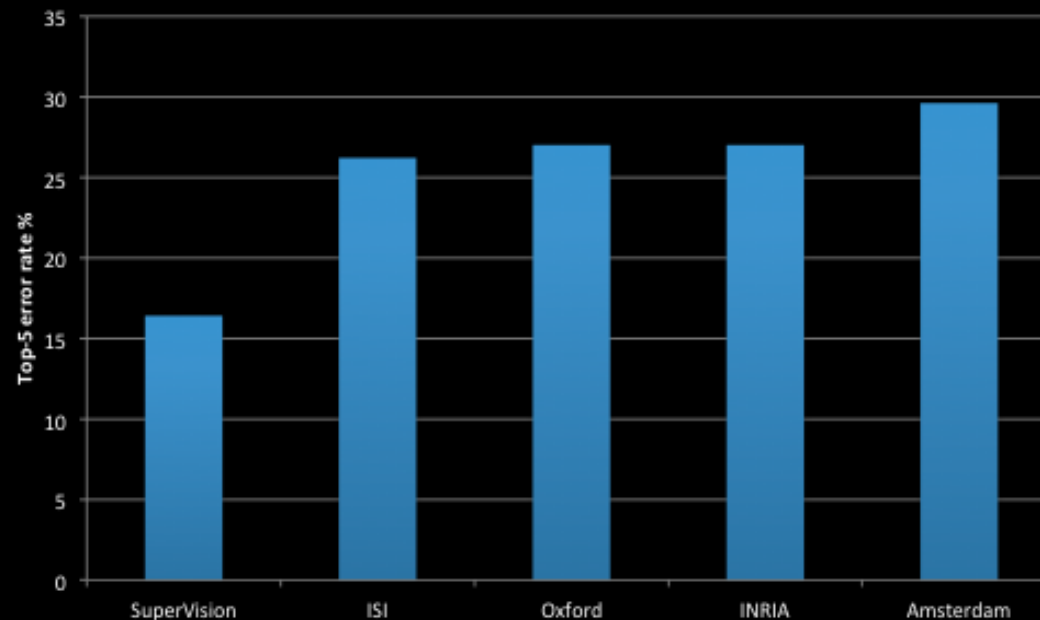
ImagetNet Breakthrough

- Achieves state-of-the-art on many object recognition tasks.

See: deeplearning.cs.toronto.edu

ImageNet Classification 2012

- Krizhevsky et al. -- 16.4% error (top-5)
- Next best (non-convnet) – 26.2% error



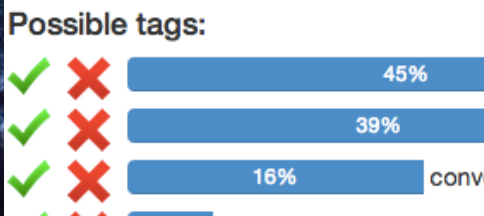
Slide from Rob Fergus, NIPS tutorial, 2012

Object Recognition Works

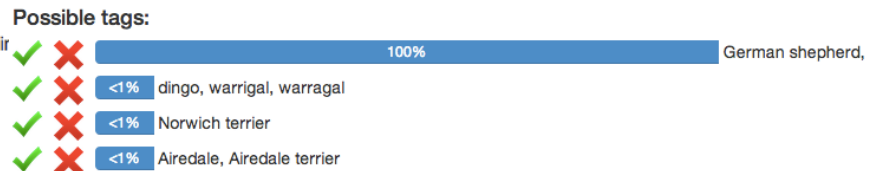
- Try it at <http://deeplearning.cs.toronto.edu>



Possible tags:
 ✓ X
 ✓ X
 ✓ X
 ✓ X

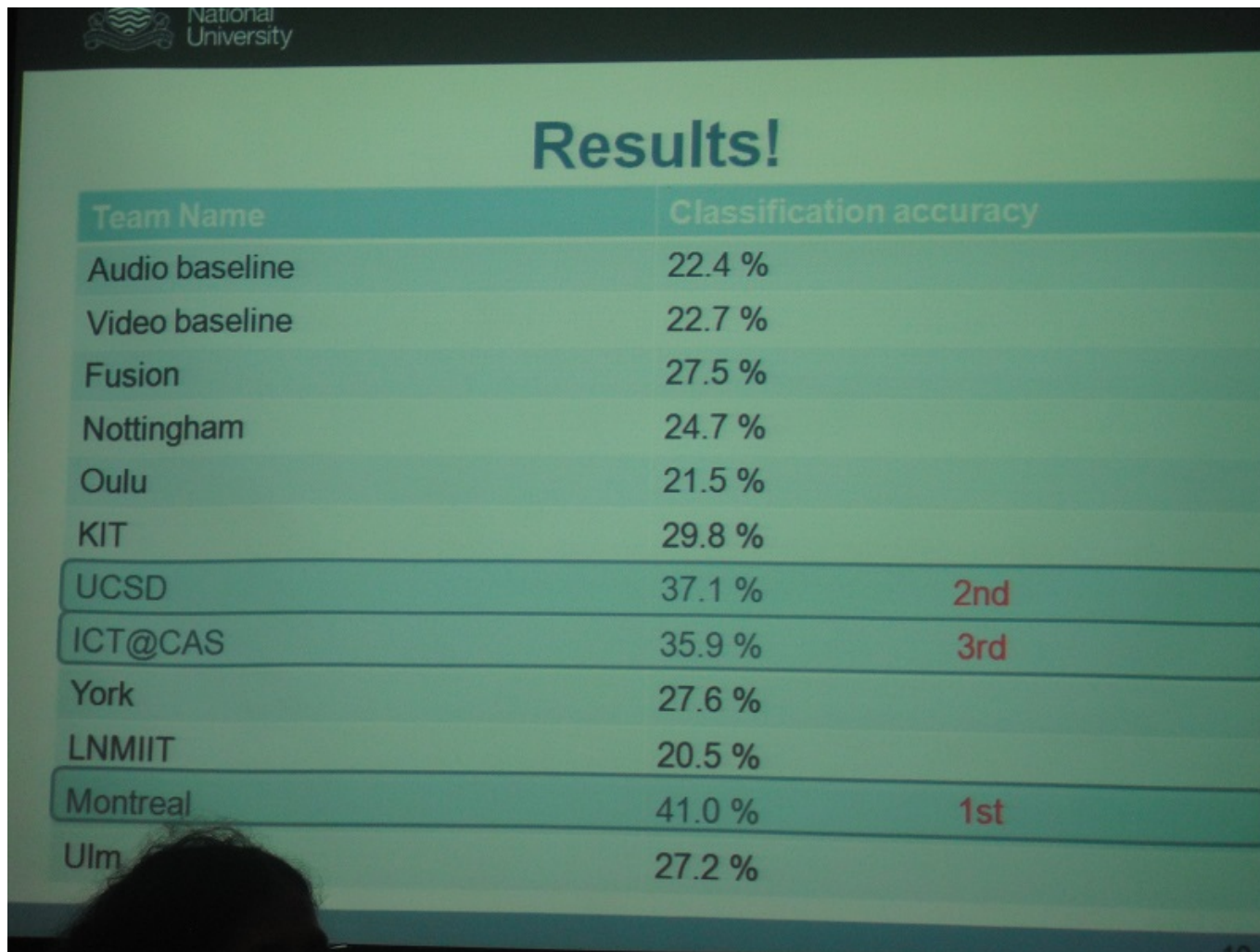


Pos
 ✓
 ✓
 ✓
 ✓



Montreal Deep Nets Win Emotion Recognition in the Wild Challenge

Predict emotional expression from video (using images + audio)



National University

Results!

Team Name	Classification accuracy	
Audio baseline	22.4 %	
Video baseline	22.7 %	
Fusion	27.5 %	
Nottingham	24.7 %	
Oulu	21.5 %	
KIT	29.8 %	
UCSD	37.1 %	2nd
ICT@CAS	35.9 %	3rd
York	27.6 %	
LNMIIT	20.5 %	
Montreal	41.0 %	1st
Ulm	27.2 %	

Dec. 9, 2013

10 BREAKTHROUGH TECHNOLOGIES 2013

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.



Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous.



Prenatal DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child?



Adv

Ma

Ske
prin
wor
mar
the
tech
jet p

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain

Smart Watches

Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely

Big Phc

Coll
ana
from
pho

Deep Learning in the News



EXCLUSIVE

Facebook, Google in 'Deep Learning' Arms Race

Yann LeCun, an NYU artificial intelligence researcher who now works for Facebook. Photo: Josh Valcarcel/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 \(@catherineshu\)](#)

Challenges

(Bengio, arxiv 1305.0445 Deep learning of representations: looking forward)

- **Unsupervised Learning, Structured outputs & Reinforcement Learning**
 - *Intractable* computations with latent variables
 - Key to more adaptable models and complex output decisions
- **Scaling up**
 - **Computation & Optimization**
- **Marrying Deep Learning & Reasoning**

Potential Outcome: AI

- **Computers that can**
 - **see and hear**
 - **understand natural language**
 - **understand human behavior**
- **Better understanding of human & animal intelligence**
- **Personnal assistants, self-driving cars...**

Technical Goals Hierarchy

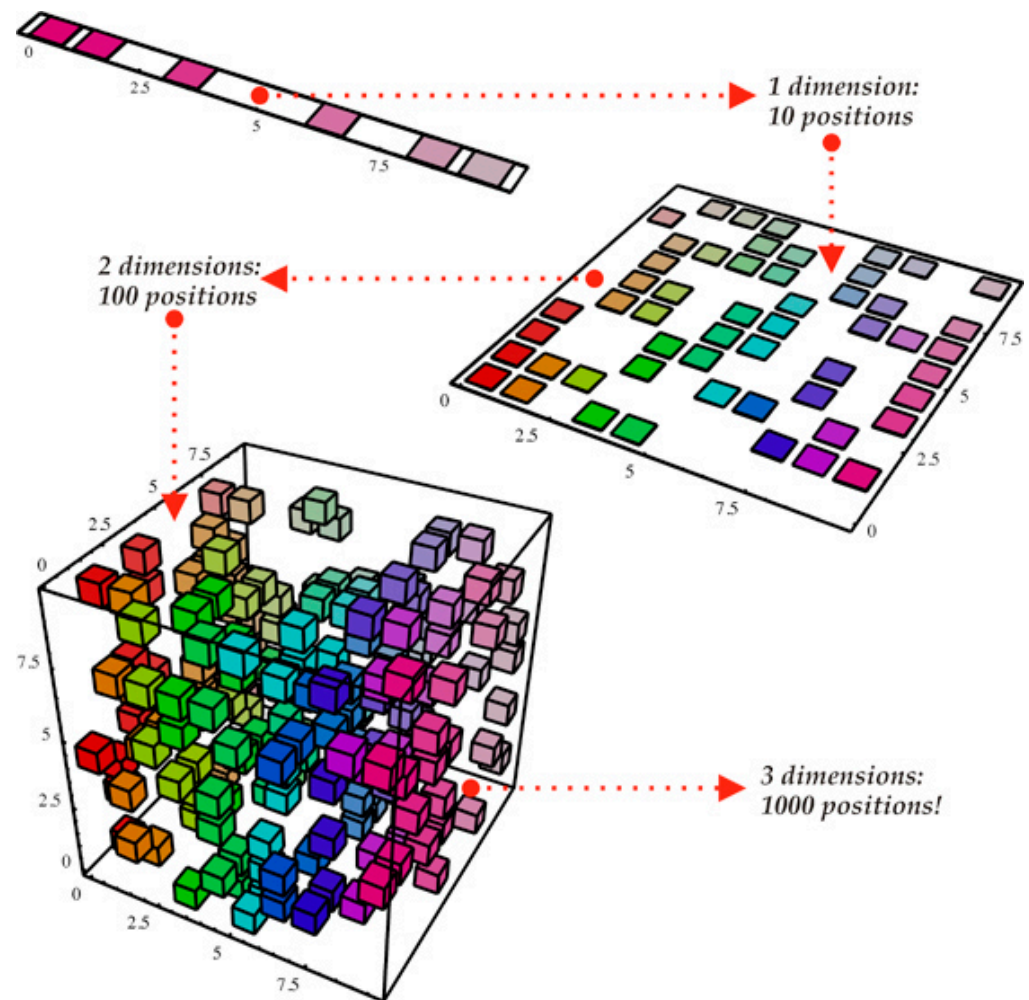
To reach AI:

- Needs **knowledge**
- Needs **learning**
(involves priors + *optimization/search* + *efficient computation*)
- Needs **generalization**
(guessing where probability mass concentrates)
- Needs ways to fight the curse of dimensionality
(exponentially many configurations of the variables to consider)
- Needs disentangling the underlying explanatory factors
(making sense of the data)

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

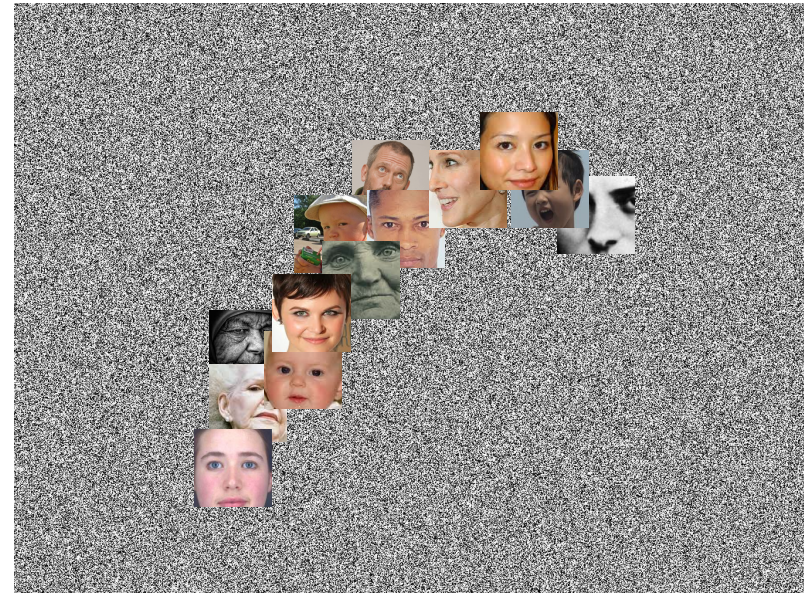
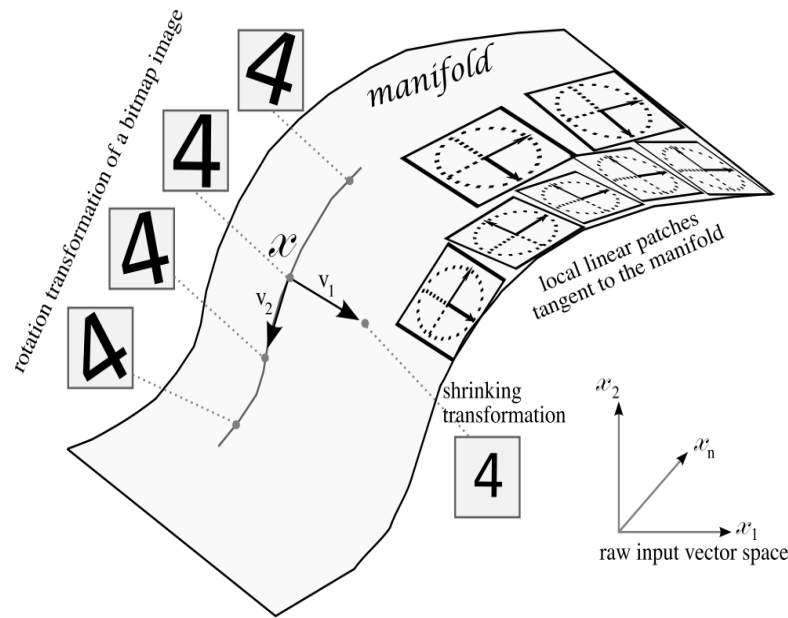
To generalize locally,
need representative
examples for all
relevant variations!

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



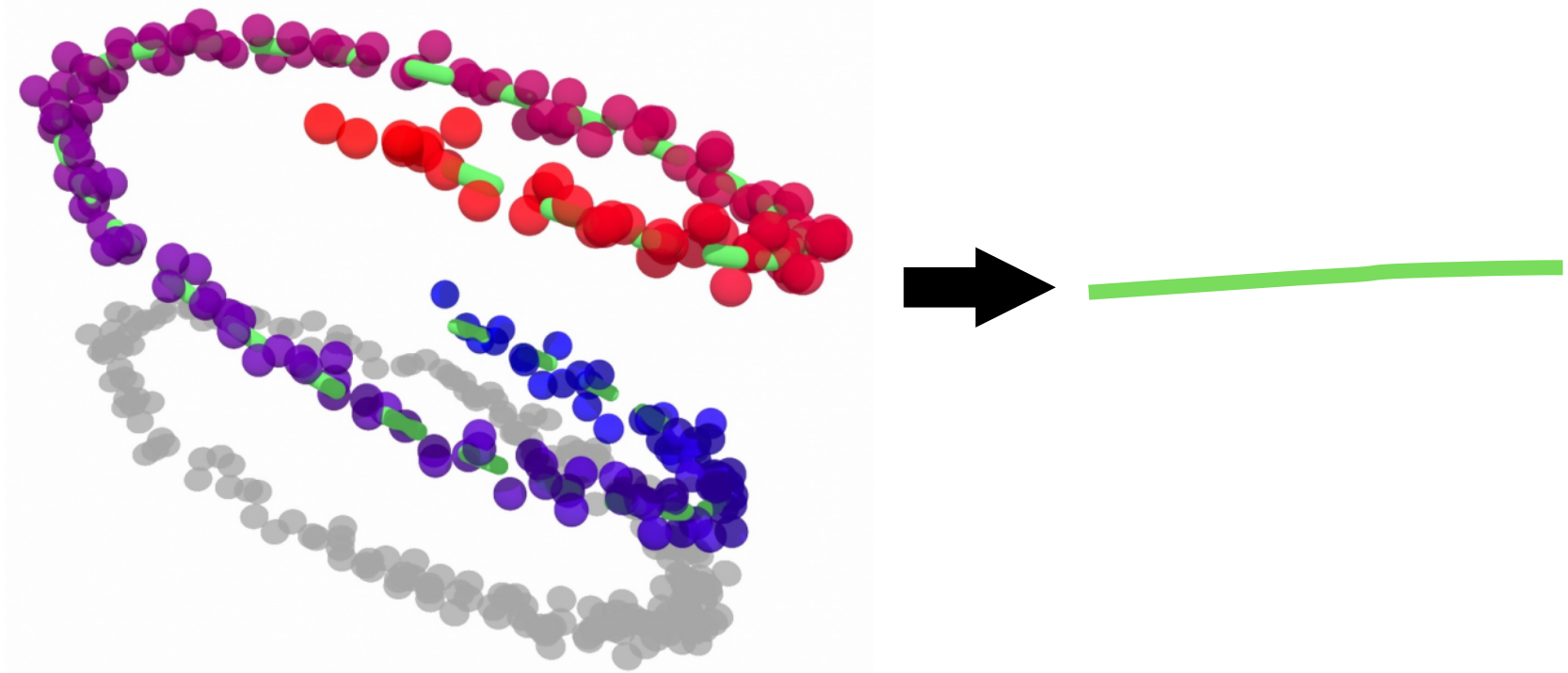
For AI Tasks: Manifold structure

- examples **concentrate** near a lower dimensional “manifold”
- **Evidence: most input configurations are unlikely**



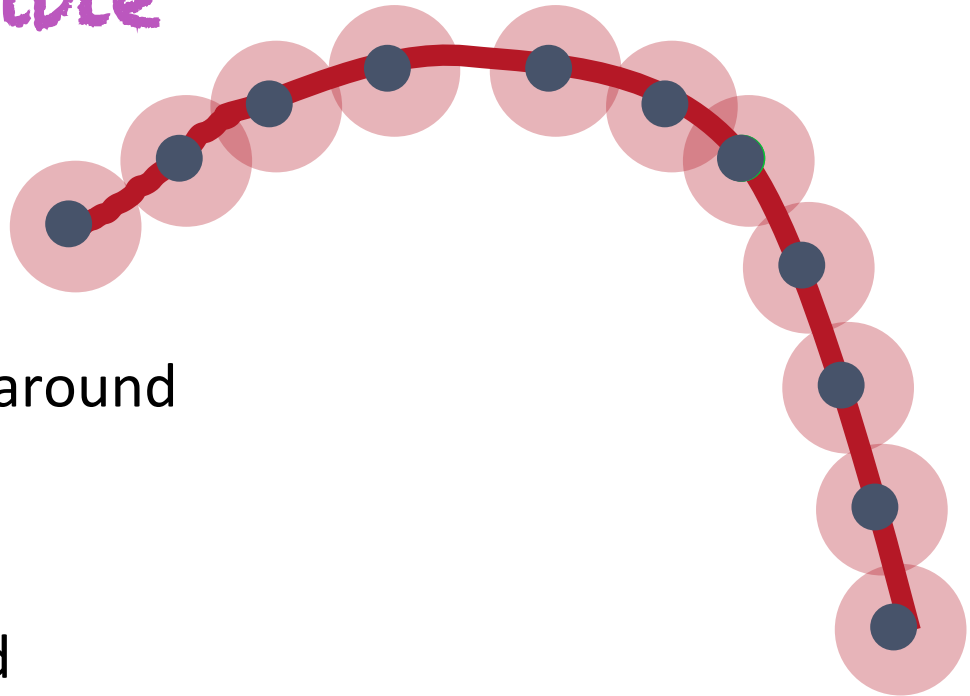
Geometrical view on machine Learning

- Generalization: guessing **where** *probability* mass concentrates
- Challenge: the curse of dimensionality (exponentially many configurations of the variables to consider)
- Representation Learning: mapping to a new space, unfolding



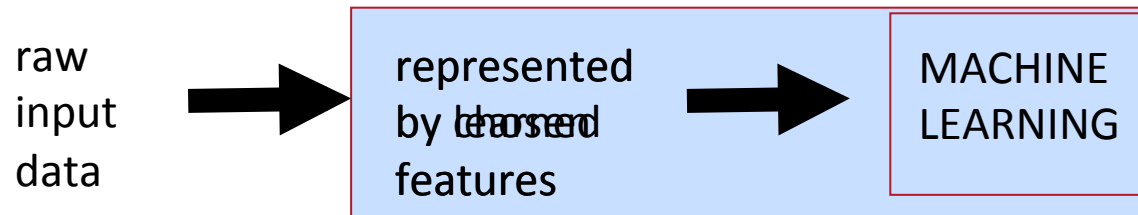
Putting Probability Mass where Structure is Plausible

- Empirical distribution: mass at training examples
- Smoothness: spread mass around
- Insufficient
- Guess some 'structure' and generalize accordingly
- Equivalent to guessing a good representation in which distance is meaningful and relationships are simple, linear

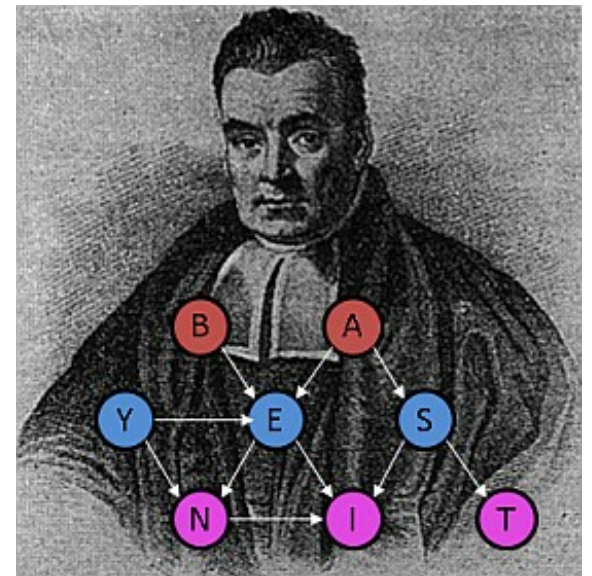


Representation Learning

- Good **features** essential for successful ML: 90% of effort



- Handcrafting features vs learning them
- Good representation?
- **guesses**
the features / factors / causes



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

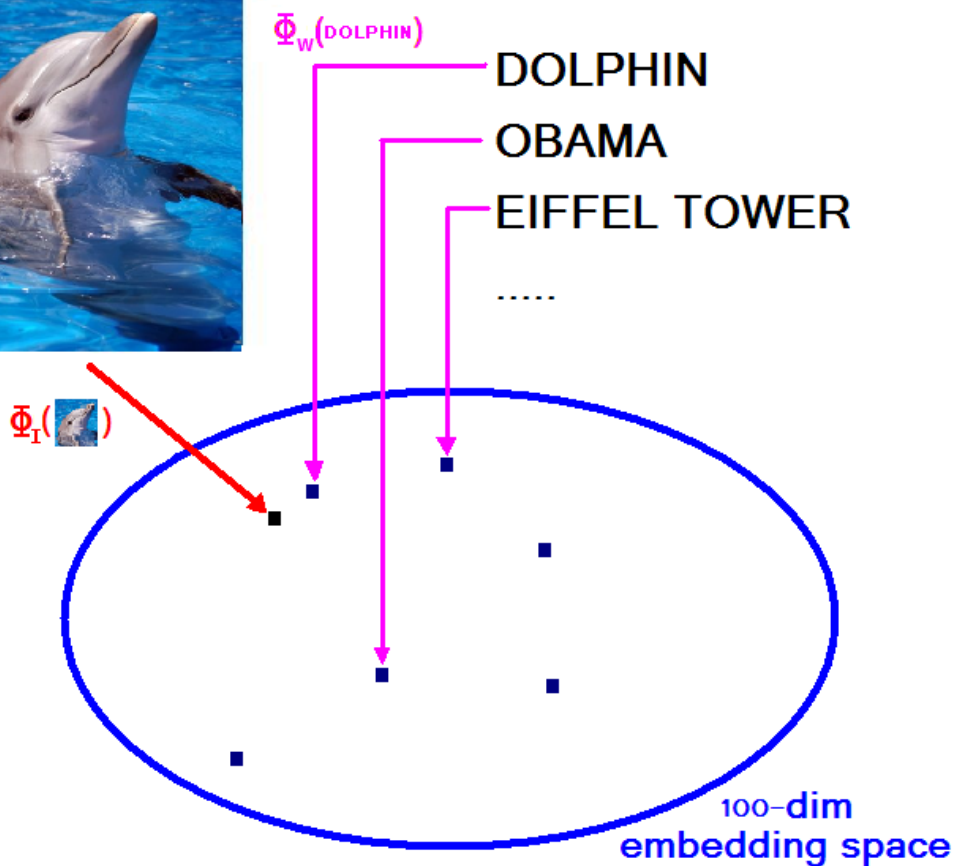


Google:

S. Bengio, J.
Weston & N.
Usunier



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



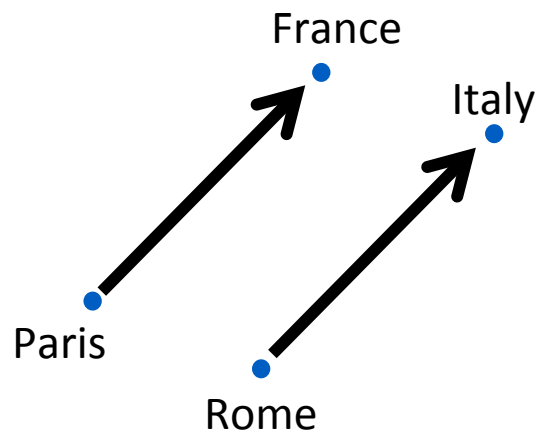
Learn $\Phi_I(\cdot)$ and $\Phi_W(\cdot)$ to optimize precision@k.

Following up on (Bengio et al NIPS'2000) Neural word embeddings - visualization



Analogical Representations for Free (Mikolov et al, ICLR 2013)

- Semantic relations appear as linear relationships in the space of learned representations
- King – Queen \approx Man – Woman
- Paris – France + Italy \approx Rome



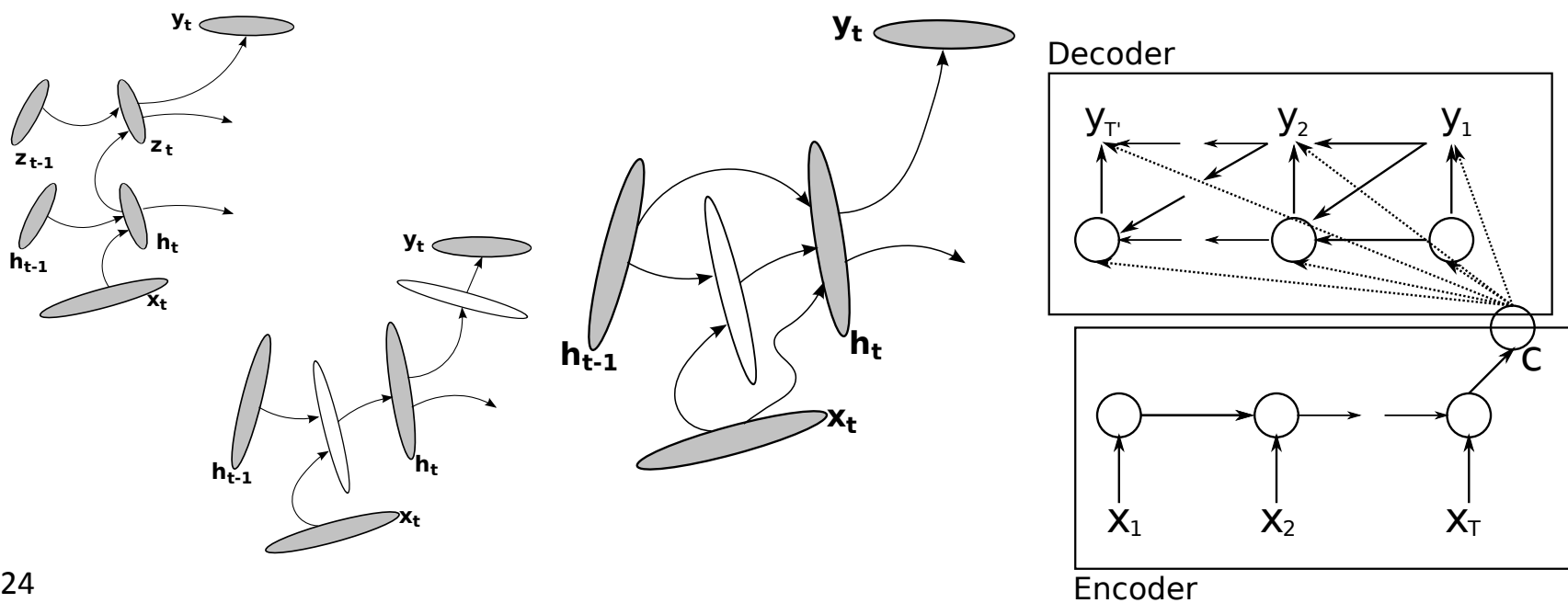
Breakthroughs in Machine Translation

- (Cho et al, EMNLP 2014) Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

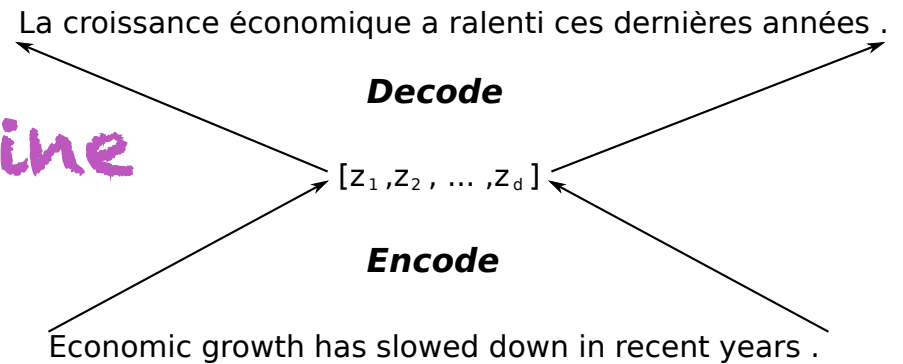
Almost 2 BLEU points improvement for English-French

- (Devlin et al, ACL 2014) Fast and Robust Neural Network Joint Models for Statistical Machine Translation

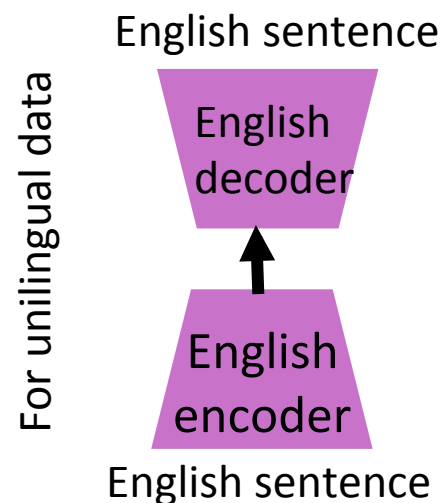
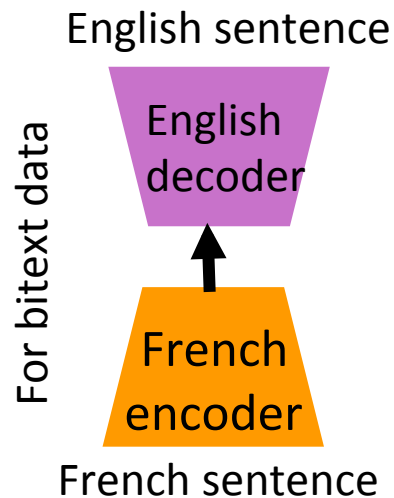
Best paper award, 6 BLEU points improvement for Arabic-English



Encoder-Decoder Framework for Machine Translation



- One encoder and one decoder per language
- Universal intermediate representation
- $\text{Encode}(\text{French}) \rightarrow \text{Decode}(\text{English}) = \text{translation model}$
- $\text{Encode}(\text{English}) \rightarrow \text{Decode}(\text{English}) = \text{language model}$
- Parametrization grows linearly with # languages, not quadratic



Learning multiple levels of representation

There is theoretical and empirical evidence in favor of multiple levels of representation

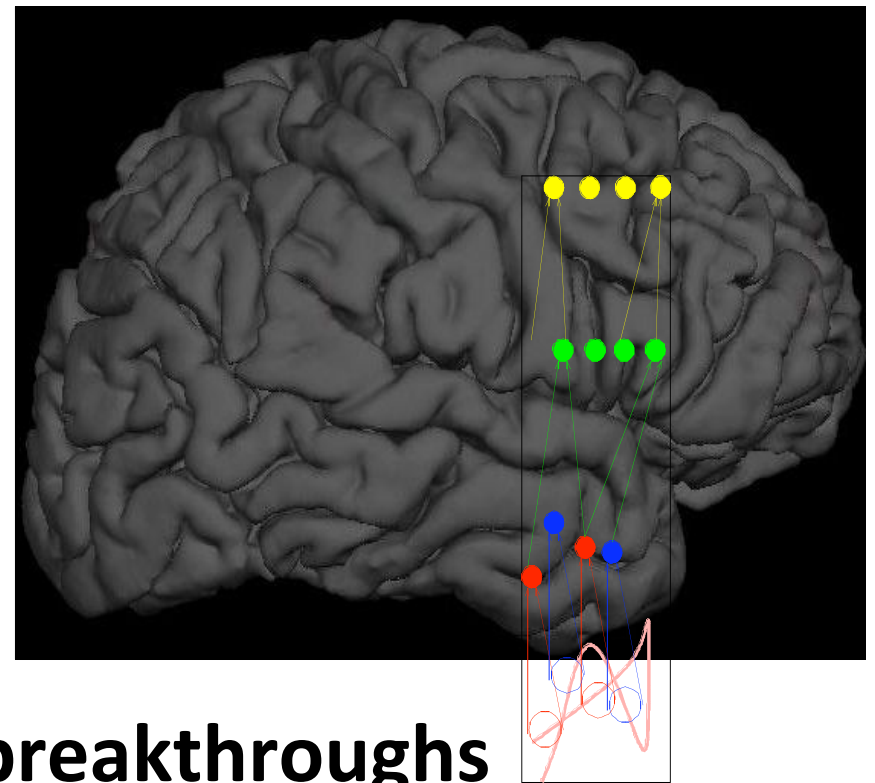
Exponential gain for some families of functions

Biologically inspired learning

Brain has a deep architecture

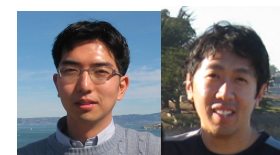
Cortex seems to have a generic learning algorithm

Humans first learn simpler concepts and compose them



It works! Speech + vision breakthroughs

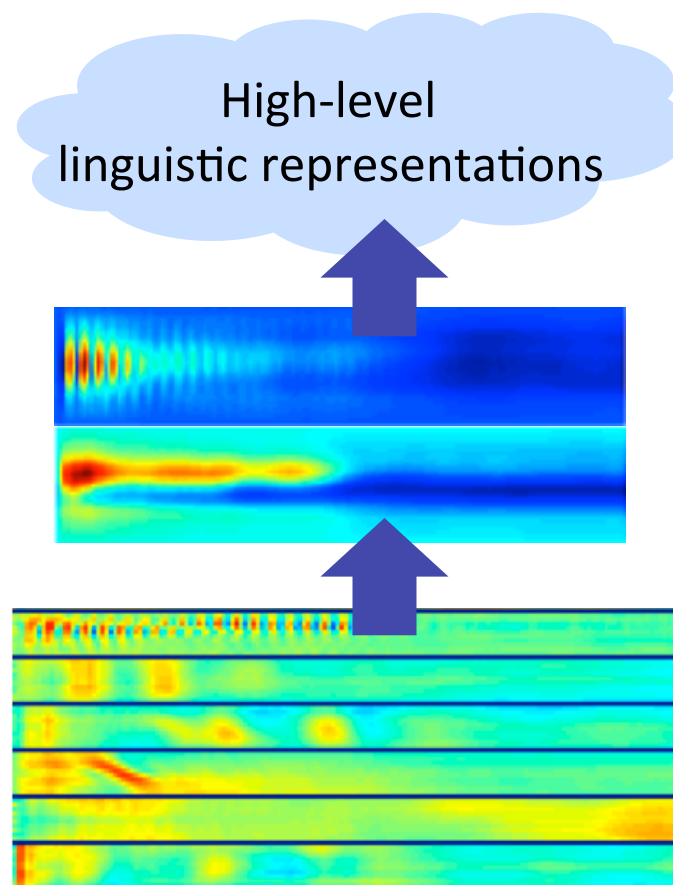
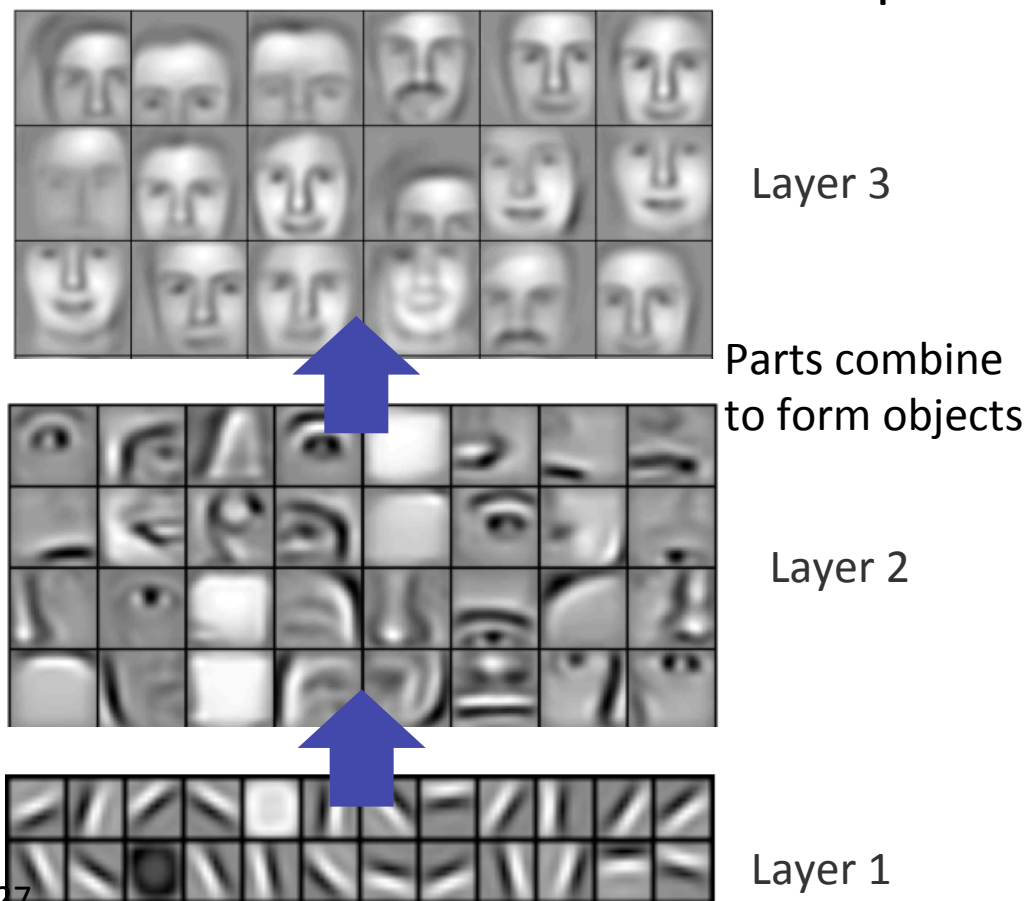
Learning multiple Levels of representation



(Lee, Largman, Pham & Ng, NIPS 2009)

(Lee, Grosse, Ranganath & Ng, ICML 2009)

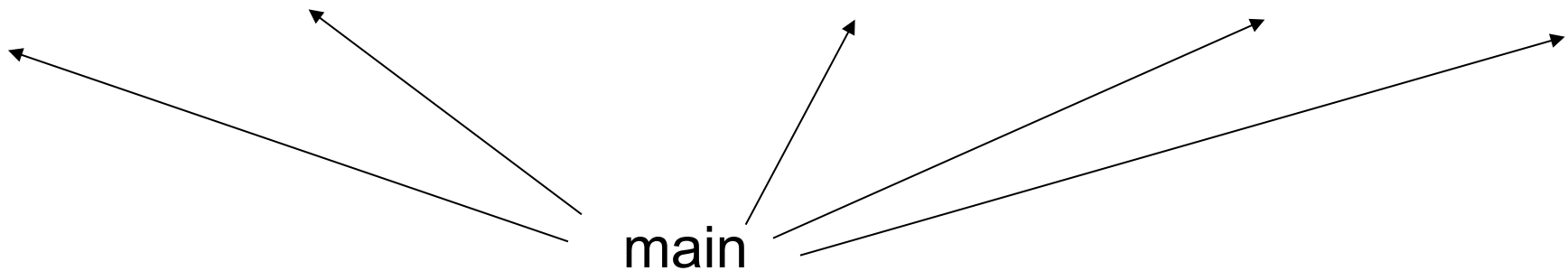
Successive model layers learn deeper intermediate representations



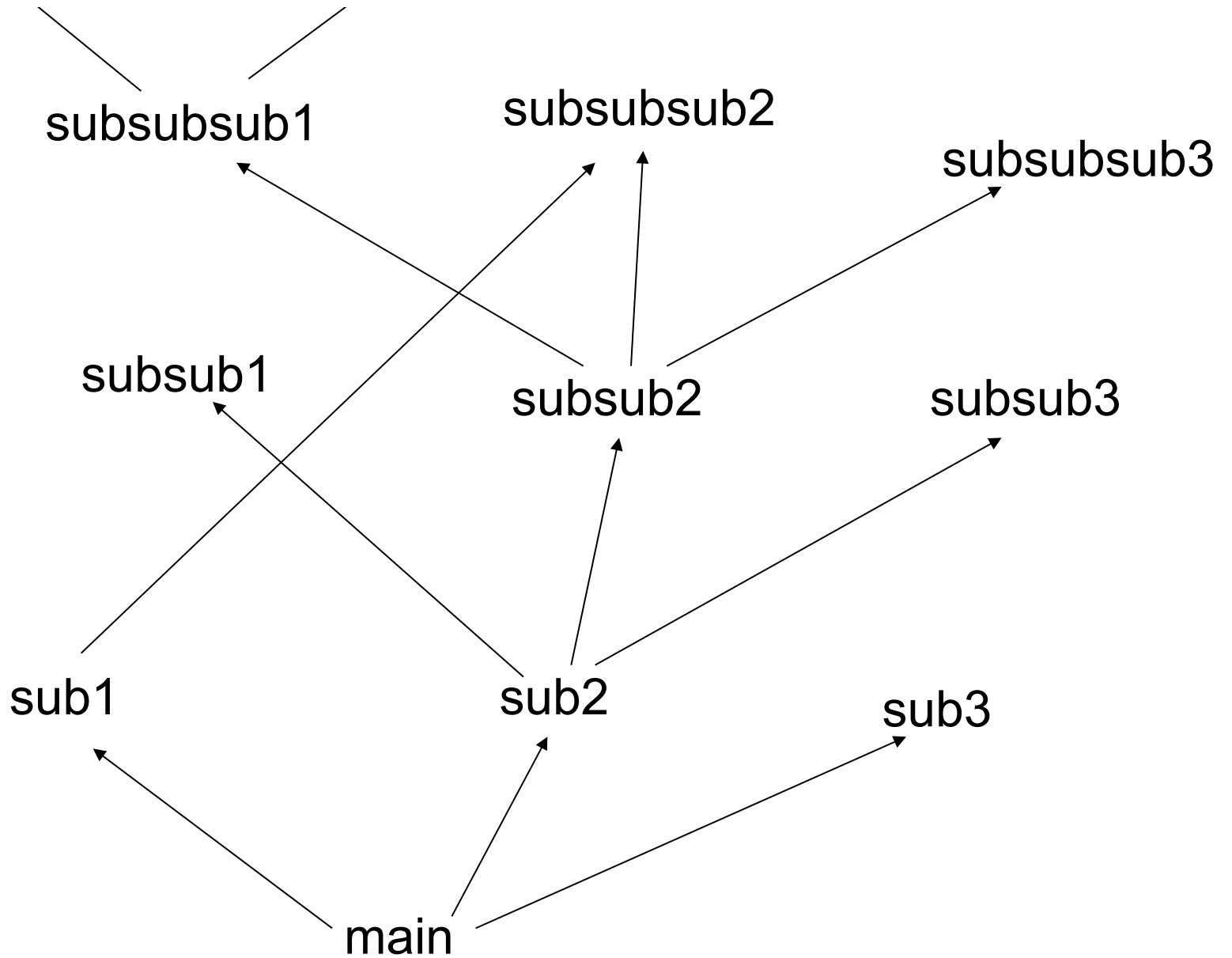
Prior: underlying factors & concepts compactly expressed w/ multiple levels of abstraction

subroutine1 includes
subsub1 code and
subsub2 code and
subsubsub1 code

subroutine2 includes
subsub2 code and
subsub3 code and
subsubsub3 code and ...



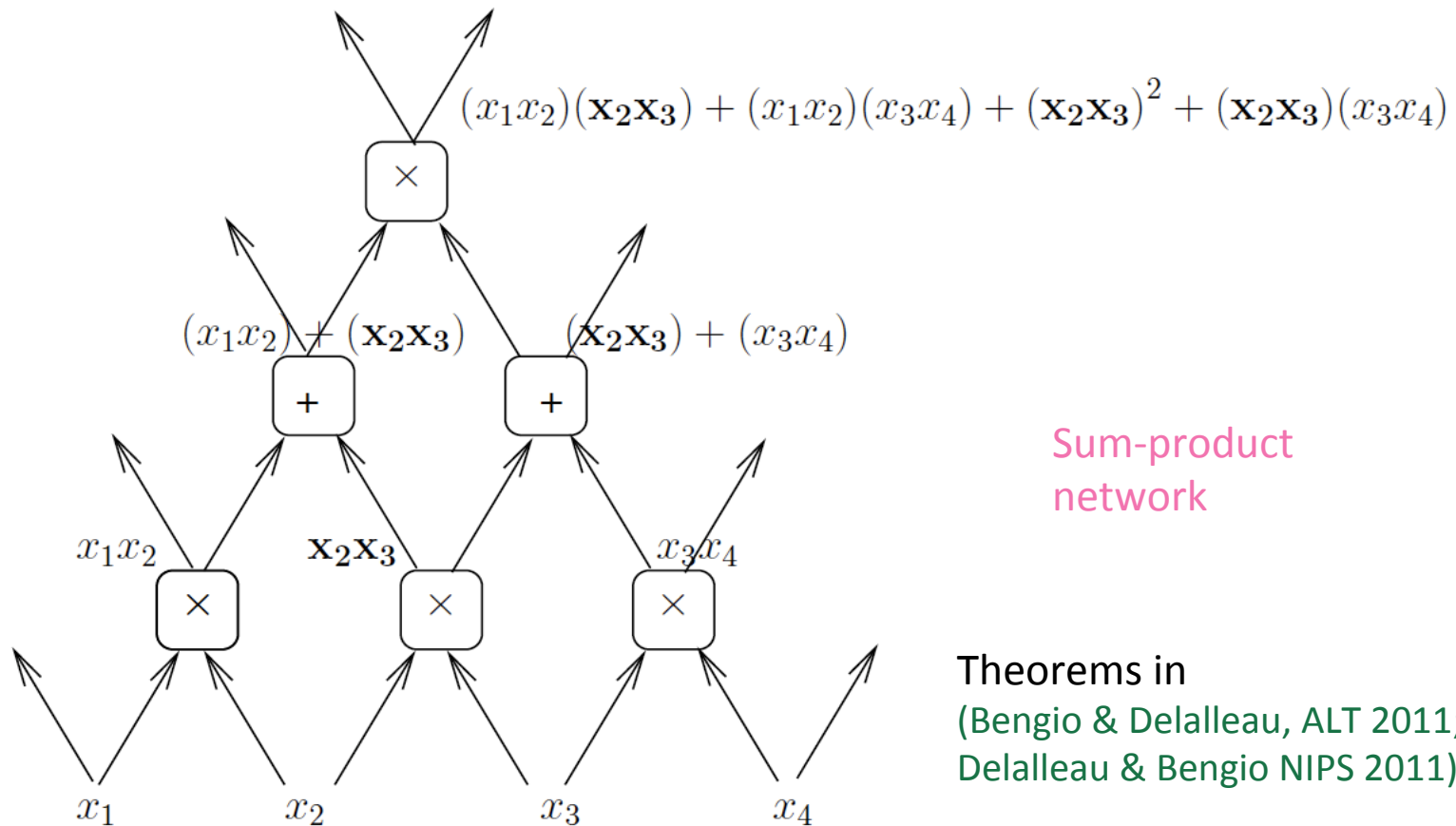
“Shallow” computer program



“Deep” computer program

Sharing Components in a Deep Architecture

Polynomial expressed with shared components: advantage of depth may grow exponentially



Deep Architectures are More Expressive



Theoretical arguments:

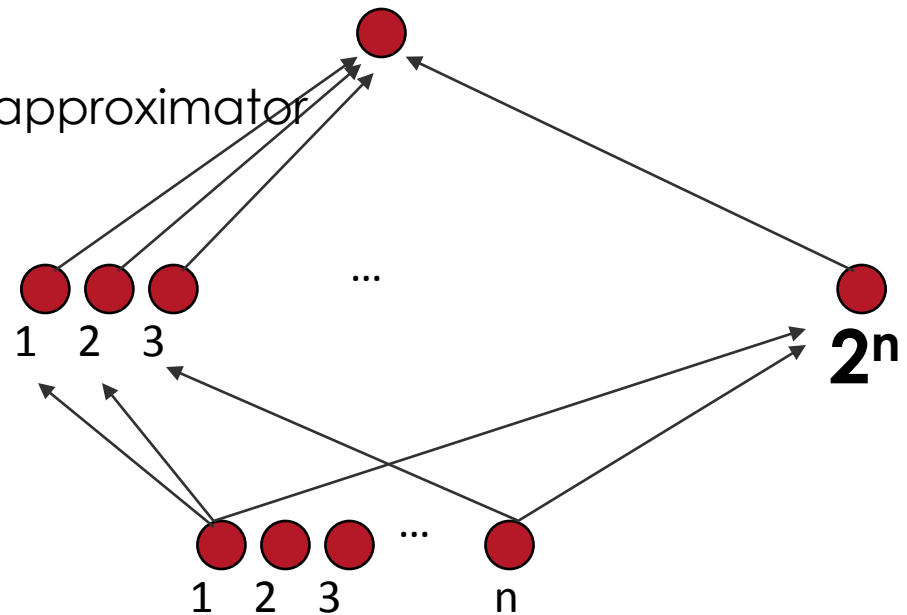
2 layers of $\left\{ \begin{array}{l} \text{Logic gates} \\ \text{Formal neurons} \\ \text{RBF units} \end{array} \right.$ = universal approximator

RBM's & 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)

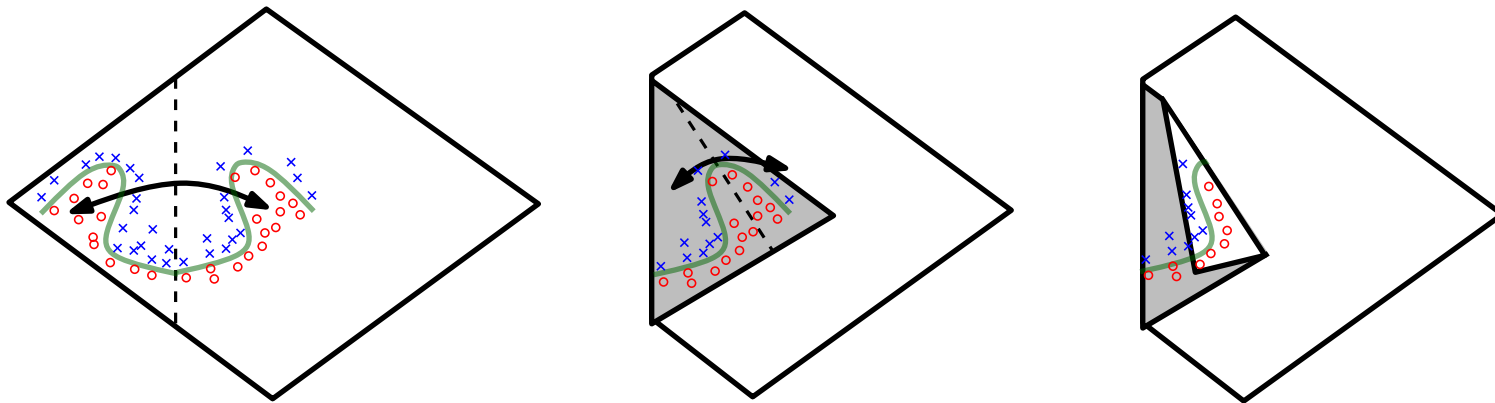
Some functions compactly represented with k layers may require exponential size with 2 layers



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

(Pascanu, Montufar, Cho & Bengio; ICLR 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:



Major Breakthrough in 2006

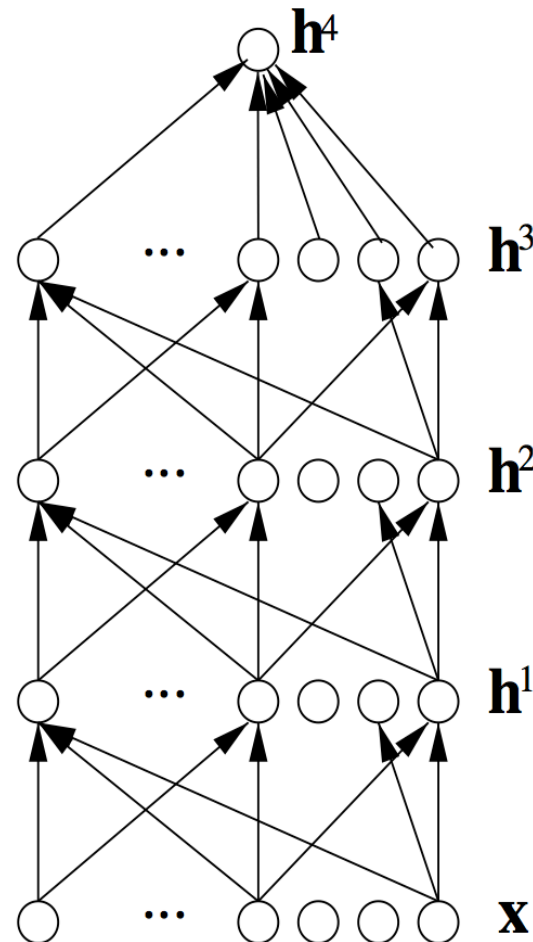


- Ability to train deep architectures by using layer-wise unsupervised learning, whereas previous purely supervised attempts had failed
- Unsupervised feature learners:
 - RBMs
 - Auto-encoder variants
 - Sparse coding variants

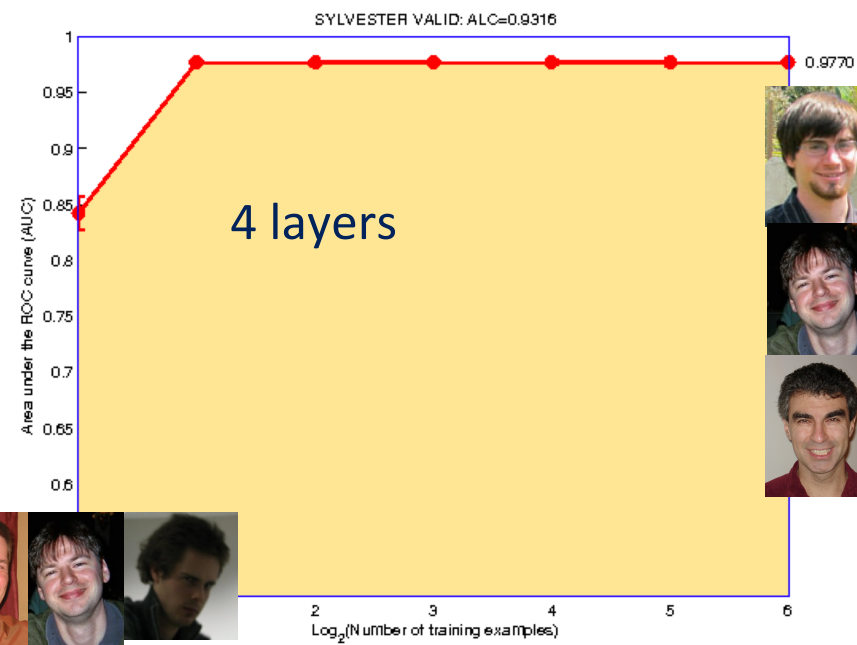
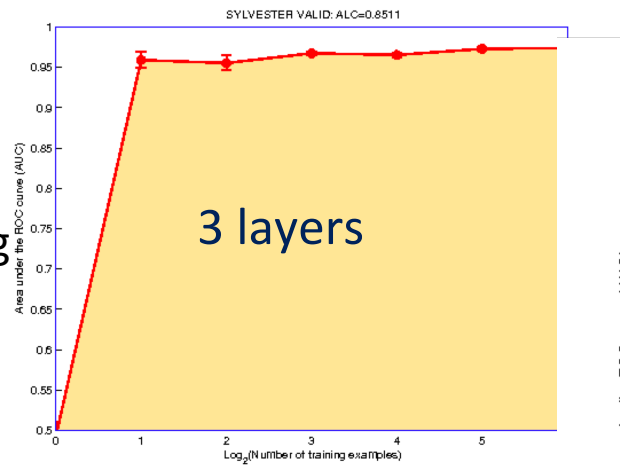
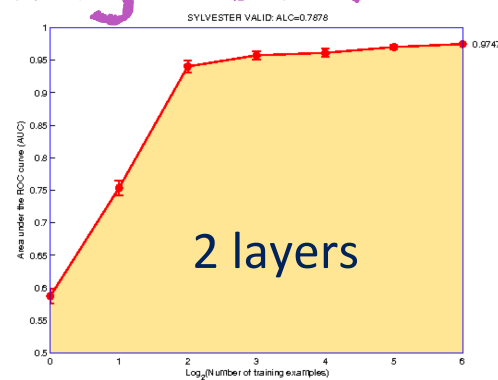
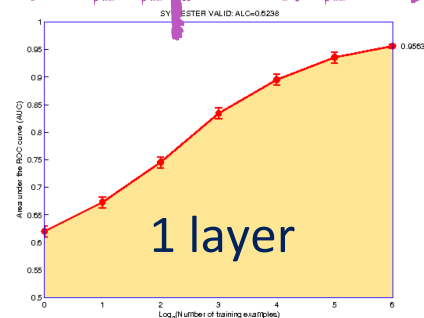
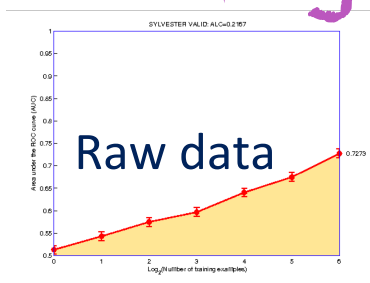


Deep Supervised Neural Nets

- Now can train them even without unsupervised pre-training:
better initialization and non-linearities (rectifiers, maxout), generalize well with large labeled sets and regularizers (dropout)
- **Unsupervised pre-training:**
rare classes, transfer, smaller labeled sets, or as extra regularizer.



Unsupervised and Transfer Learning Challenge + Transfer Learning Challenge: Deep Learning 1st Place



NIPS'2011
Transfer Learning
Challenge
Paper:
ICML'2012

ICML'2011
workshop on
Unsup. &
Transfer Learning



Is there any hope to
generalize non-locally?

Yes! Need good priors!

Depth prior: Abstraction

Bypassing the curse

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

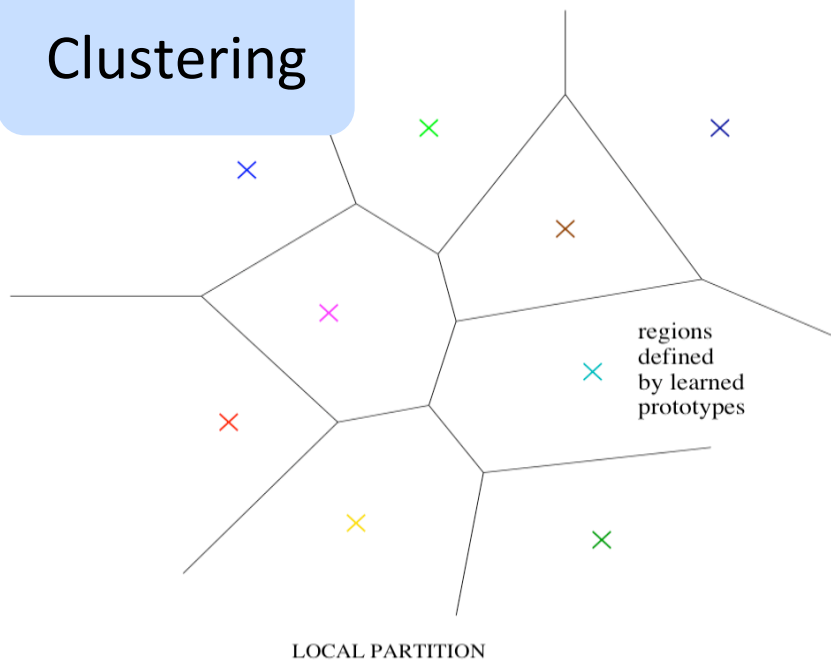
Distributed representations / embeddings: **feature learning**

Deep architecture: **multiple levels of feature learning**

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

Non-distributed representations

Clustering



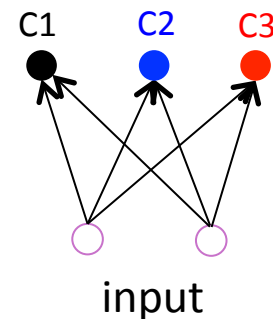
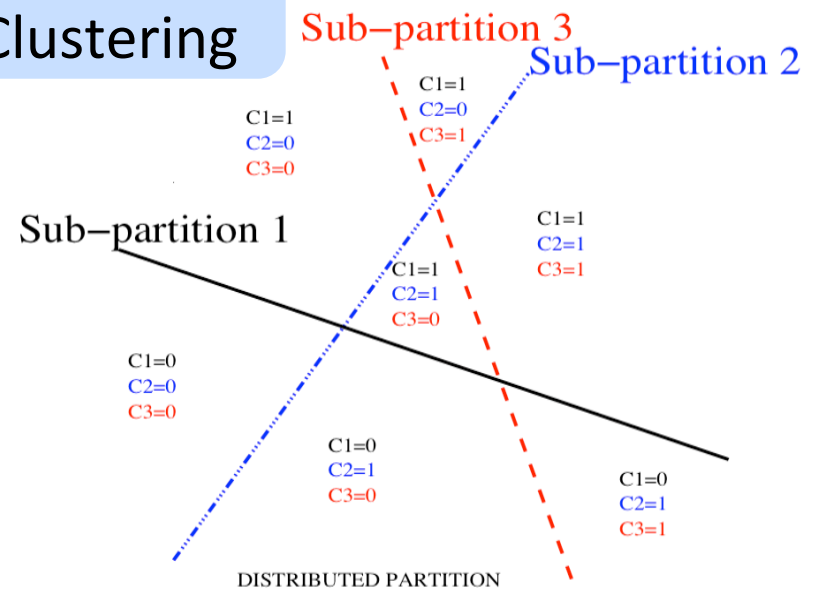
- Clustering, Nearest-Neighbors, RBF SVMs, local non-parametric density estimation & prediction, decision trees, etc.
- Parameters for each distinguishable region
- **# of distinguishable regions is linear in # of parameters**

→ No non-trivial generalization to regions without examples

The need for distributed representations

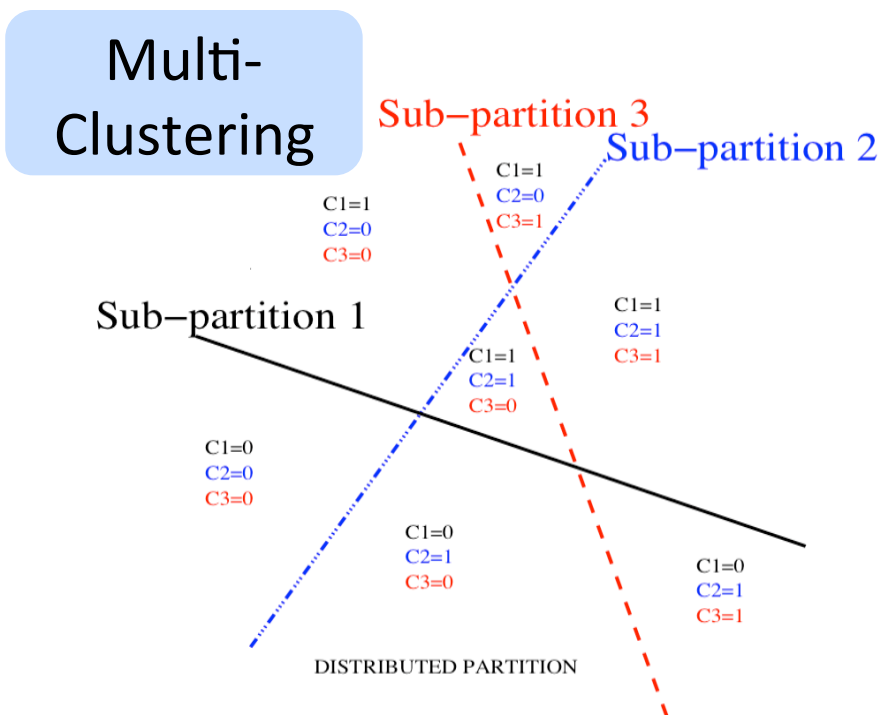
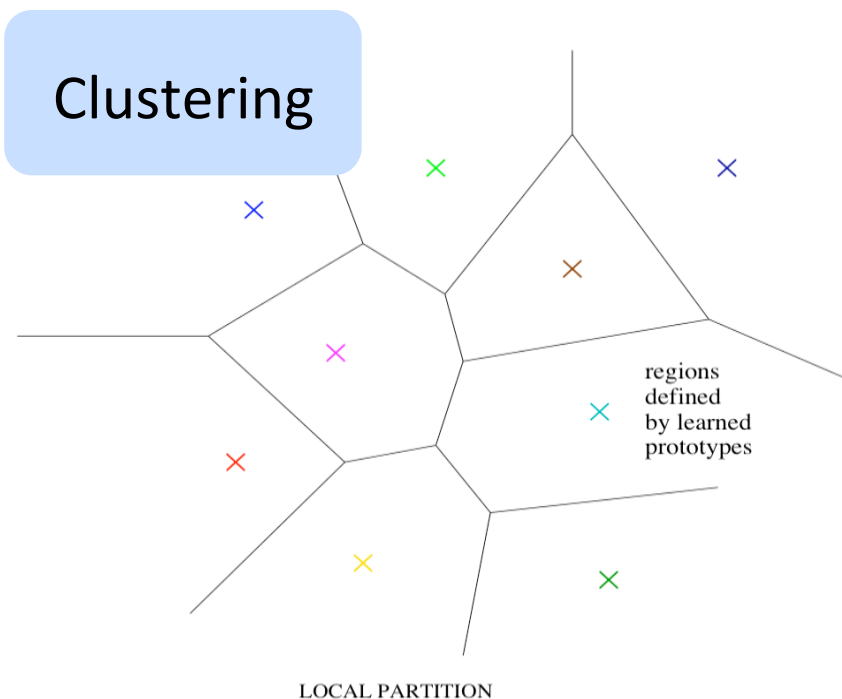
- Factor models, PCA, RBMs, Neural Nets, Sparse Coding, Deep Learning, etc.
- Each parameter influences many regions, not just local neighbors
- **# of distinguishable regions grows almost exponentially with # of parameters**
- **GENERALIZE NON-LOCALLY TO NEVER-SEEN REGIONS**

Multi-Clustering



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

The need for distributed representations



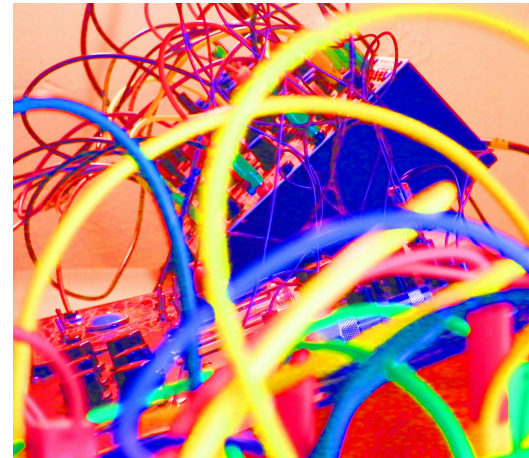
Learning a **set of features** that are not mutually exclusive can be **exponentially more statistically efficient** than having nearest-neighbor-like or clustering-like models

How do humans generalize from very few examples?

- They **transfer** knowledge from previous learning:
 - Representations
 - Explanatory factors
- Previous learning from: unlabeled data
 - + labels for other tasks
- **Prior: shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|x)$**

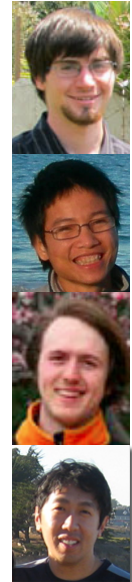
Invariance and Disentangling

- Invariant features
- Which invariances?
- Alternative: learning to disentangle factors
- Good disentangling →
 avoid the curse of dimensionality



Emergence of Disentangling

- (Goodfellow et al. 2009): sparse auto-encoders trained on images
 - some higher-level features more invariant to geometric factors of variation
- (Glorot et al. 2011): sparse rectified denoising auto-encoders trained on bags of words for sentiment analysis
 - different features specialize on different aspects (domain, sentiment)



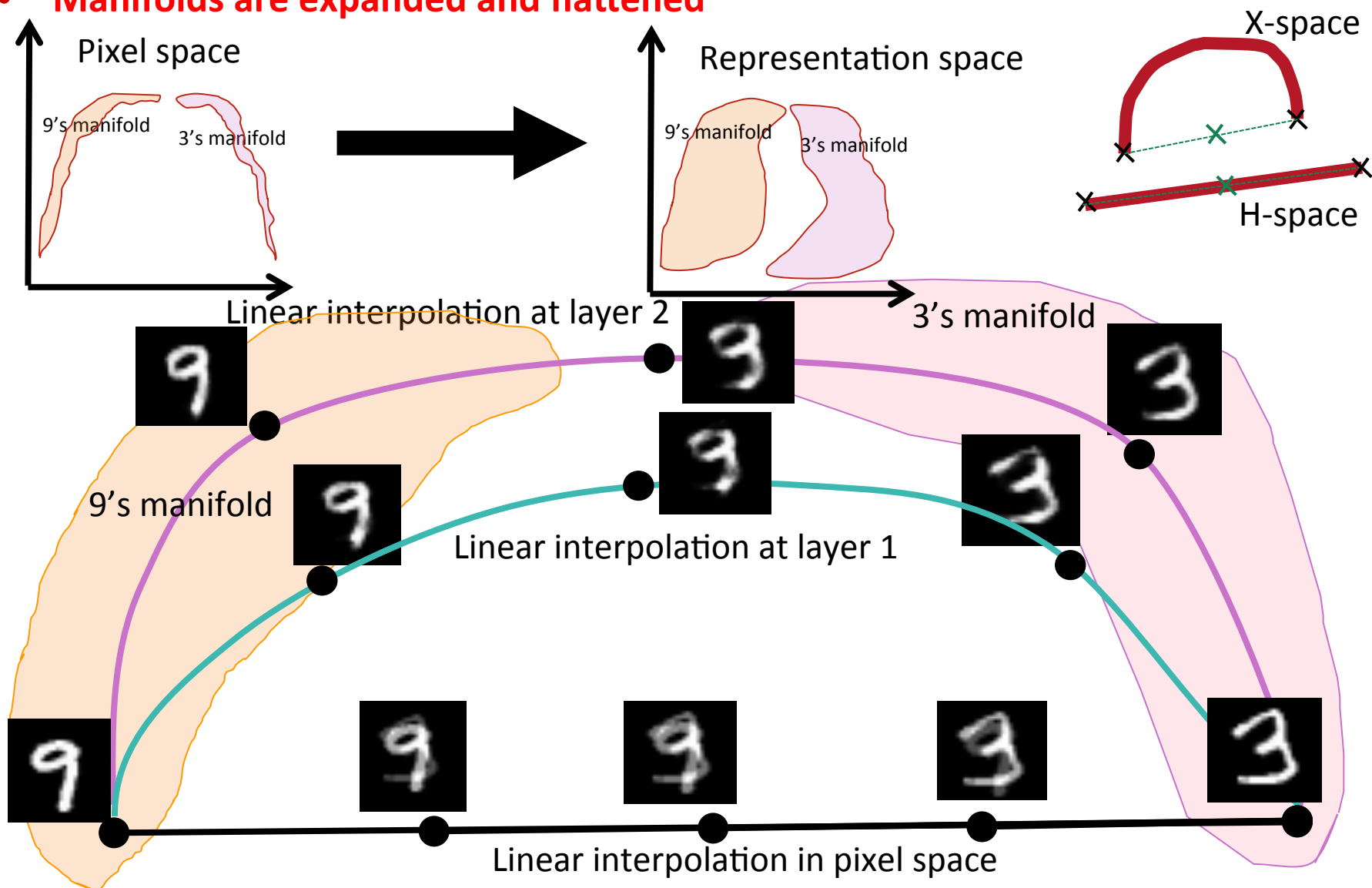
WHY?

Broad Priors as Hints to Disentangle the Factors of Variation

- *Multiple factors*: distributed representations
- Multiple levels of abstraction: *depth*
- *Semi-supervised* learning: Y is one of the factors explaining X
- *Multi-task* learning: different tasks share some factors
- *Manifold* hypothesis: probability mass concentration
- Natural *clustering*: class = manifold, well-separated manifolds
- Temporal and spatial *coherence*
- *Sparsity*: most factors irrelevant for particular X
- *Simplicity* of factor dependencies (in the right representation)

Space-Filling in Representation-Space

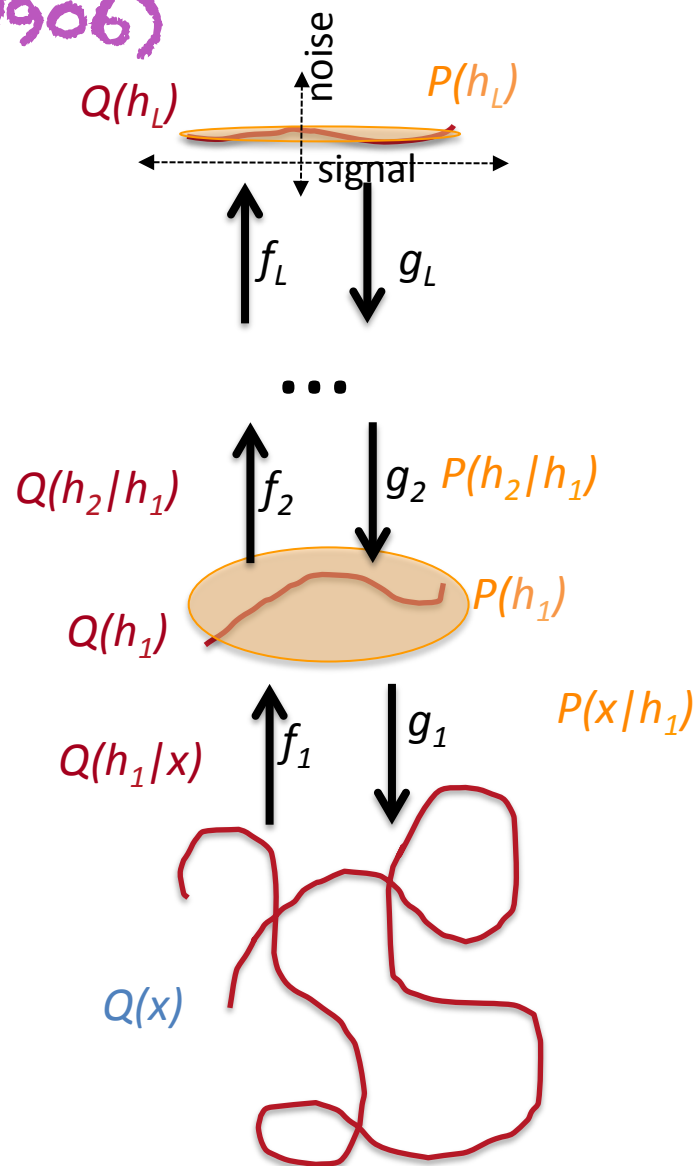
- Deeper representations \rightarrow abstractions \rightarrow disentangling
- Manifolds are expanded and flattened



Extracting Structure By Gradual Disentangling and Manifold Unfolding (Bengio 2014, arXiv 1407.7906)

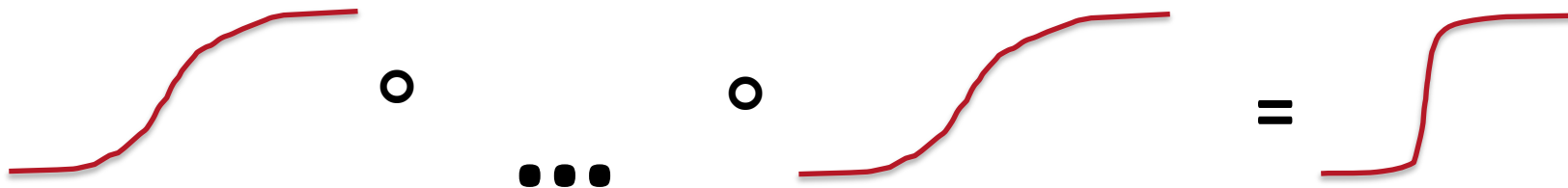
Each level transforms the data into a representation in which it is easier to model, unfolding it more, contracting the noise dimensions and mapping the signal dimensions to a factorized (uniform-like) distribution.

$$\min KL(Q(x, h) || P(x, h))$$



Issues with Back-Prop

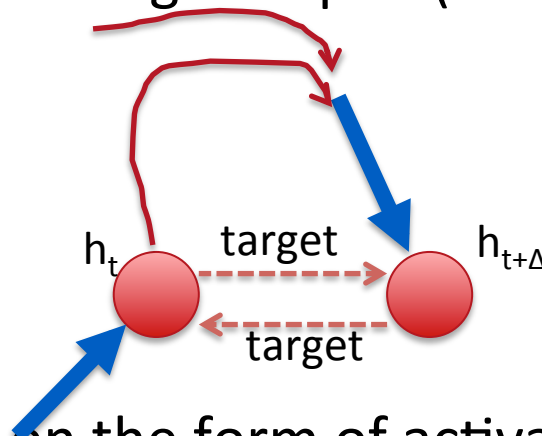
- Over very deep nets or recurrent nets with many steps, non-linearities compose and yield sharp non-linearity \rightarrow gradients vanish or explode
- Training deeper nets: harder optimization
- In the extreme of non-linearity: discrete functions, can't use back-prop
- Not biologically plausible



How Brains Might Learn Without Backprop

- Two principles:
 - The past tries to match the future: prediction
 - The future tries to match the past: reconstructionNot clear if these should be on same or different units.
- Plus: observations being clamped (not always)

Different loops =
Different lengths =
Different Δ

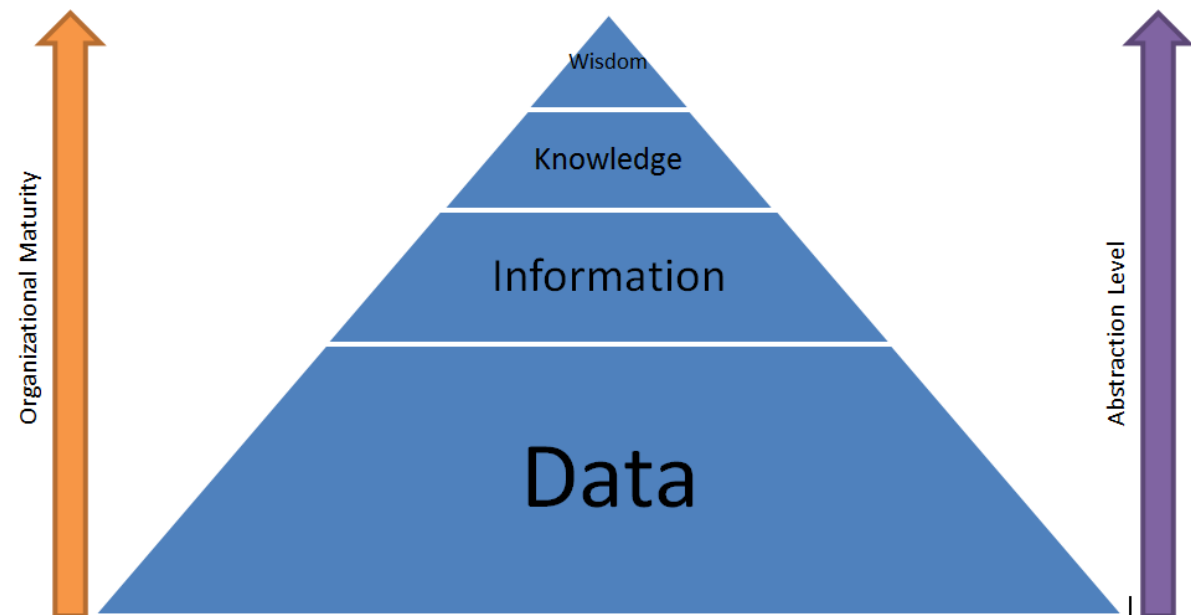


No need to store
past activations: just
average pre-synaptic
contributions with a
temporal kernel

- Does not depend on the form of activation function, tied symmetric weights, differentiability of anything, using rates vs spikes, etc.

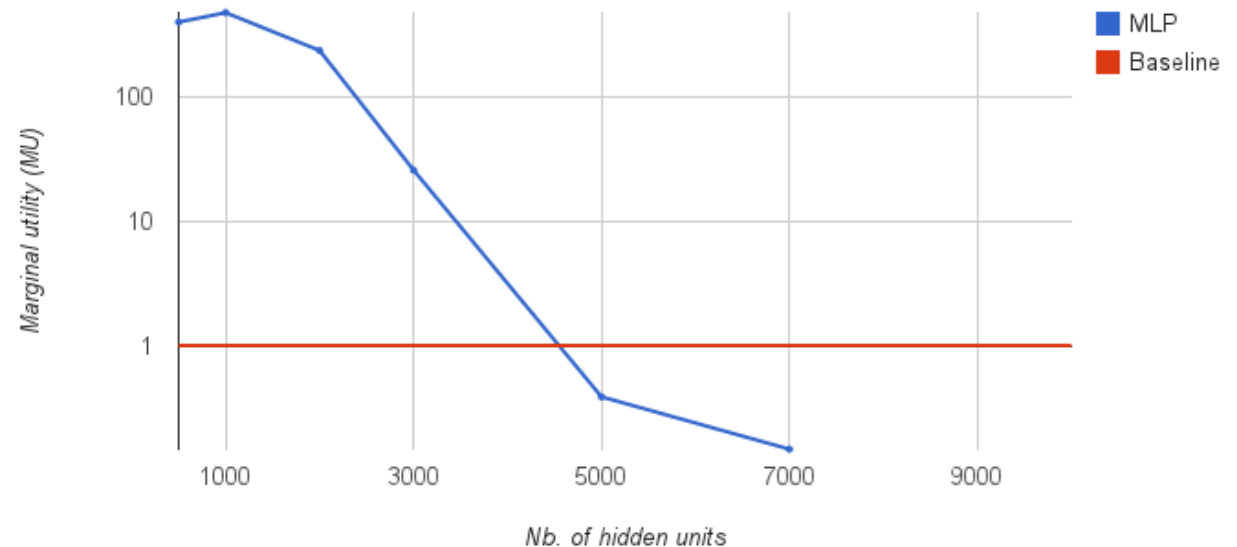
Learning Multiple Levels of Abstraction

- 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



Optimization & Underfitting

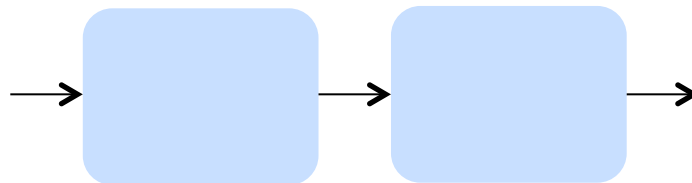
- On large datasets, major obstacle is underfitting
- **Marginal utility** of wider MLPs decreases quickly below memorization baseline



- Current limitations: local minima, ill-conditioning or else?

Guided Training, Intermediate Concepts

- In (Gulcehre & Bengio ICLR'2013) we set up a task that seems almost impossible to learn by shallow nets, deep nets, SVMs, trees, boosting etc
- Breaking the problem in two sub-problems and pre-training each module separately, then fine-tuning, nails it
- *Need prior knowledge to decompose the task*
- **Guided pre-training** allows to find much better solutions, escape effective local minima



Effective Local Minima

- It is not clear that **actual** local minima are a real issue in training deep nets
 - But initial conditions can sometimes matter a lot!
 - see evidence suggesting instead that saddle points create plateaus that act as obstacles:

Pascanu et al, On the saddle point problem for non-convex optimization, arXiv 2014

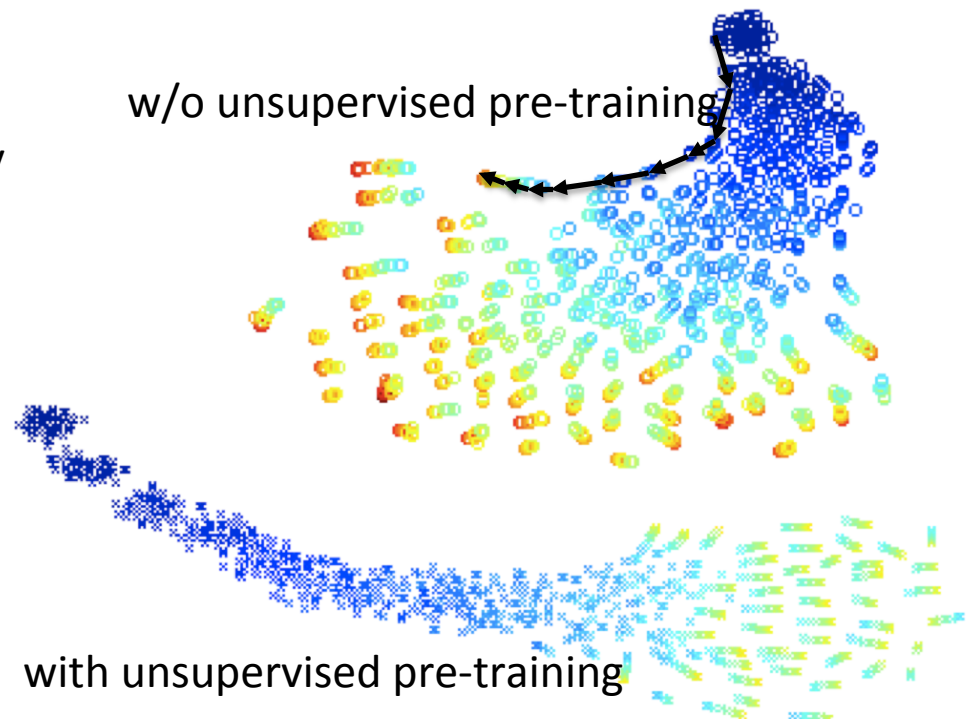
- An optimizer like the one in brains may **get stuck** → **effective local minima**

Effect of Initial Conditions in Deep Nets

- (Erhan et al 2009, JMLR)
- Supervised deep net with vs w/o unsupervised pre-training → very different minima

Neural net trajectories in function space, visualized by t-SNE

No two training trajectories end up in the same place → huge number of effective local minima



Cultural Evolution & Deep Learning

- Optimization difficulty for deeper nets, more abstract concepts
- Humans manage to bypass this difficulty thanks to culture, guidance from other humans
- The evolution of memes & culture is an effective way to explore the space of brain configurations, by divide-and-conquer:
 - Evolutionary pressure on the memes themselves, not just on their carrier

(Bengio 2013, *Evolving culture vs local minima*, ArXiv 1203.2990)

Conclusions

- Deep Learning has become a crucial machine learning tool:
 - **Int. Conf. on Learning Representation 2013 & 2014** a huge success!
Conference & workshop tracks, open to new ideas 😊
- Industrial applications (Google, IBM, Microsoft, Baidu, Facebook, and now Samsung...)
- Potential for more breakthroughs and approaching the “understanding” part of AI by
 - Scaling computation
 - Numerical optimization (better training much deeper nets, RNNs)
 - Bypass intractable marginalizations and exploit broad priors to learn more disentangled abstractions
 - Reason from incrementally added facts

LISA team: **Merci! Questions?**

