Deep Learning

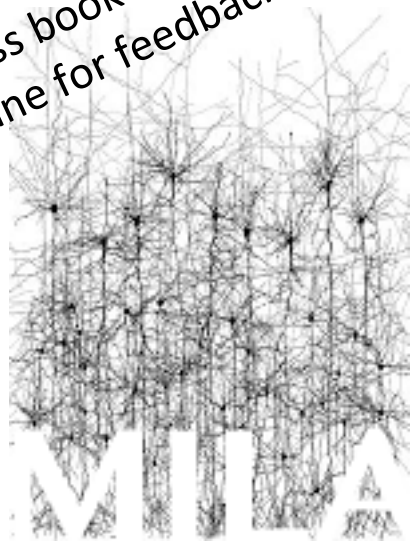
Yoshua Bengio

July 20, 2015

Lisbon Open Data Meetup

Université de Montréal

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, some in machine translation

Ongoing Progress: Natural Language Understanding

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

Xu et al, to appear 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.

Why is Deep Learning Working so Well?

Machine Learning, AI & No Free Lunch

- Three key ingredients for ML towards AI

  1. Lots & lots of data

  2. Very flexible models

  3. Powerful priors that can defeat the curse of dimensionality

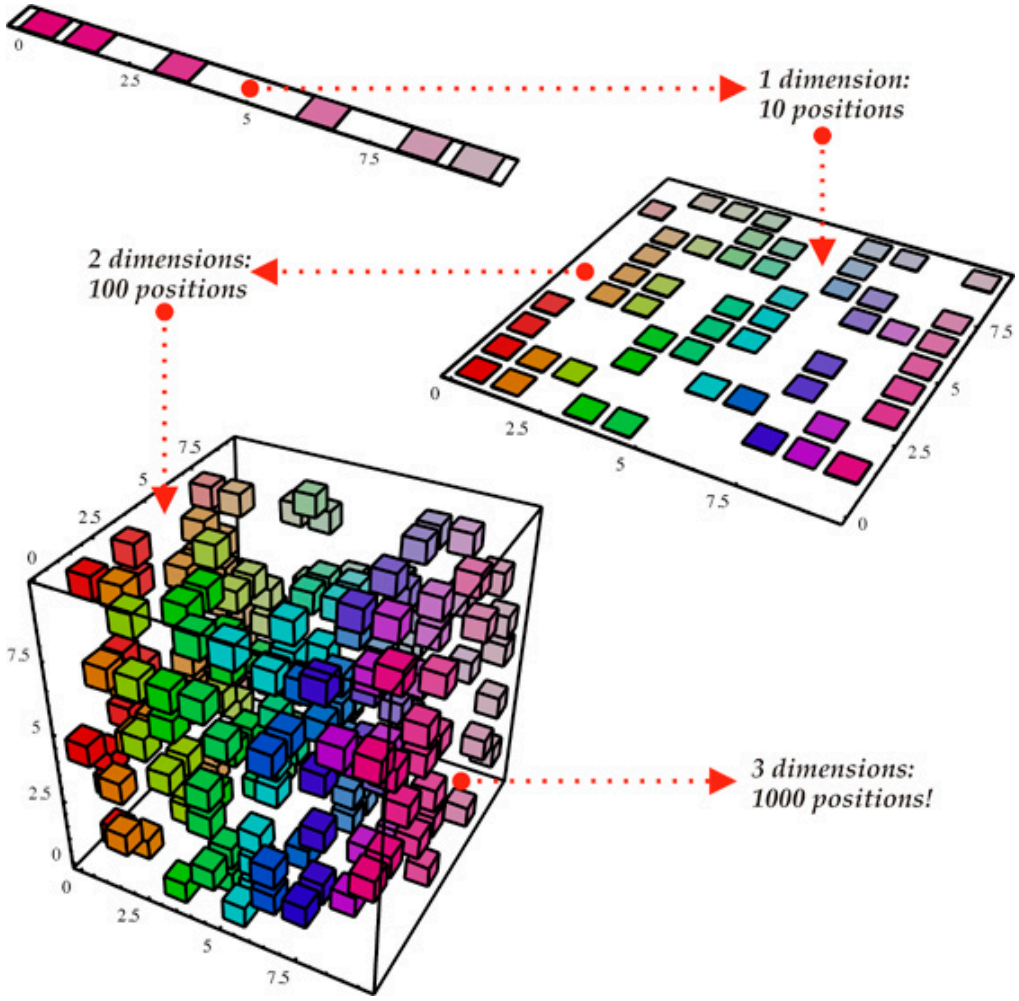
Ultimate Goals

- **AI**
- Needs **knowledge**
- Needs **learning**
  (involves priors + ***optimization****/search*)
- 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



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

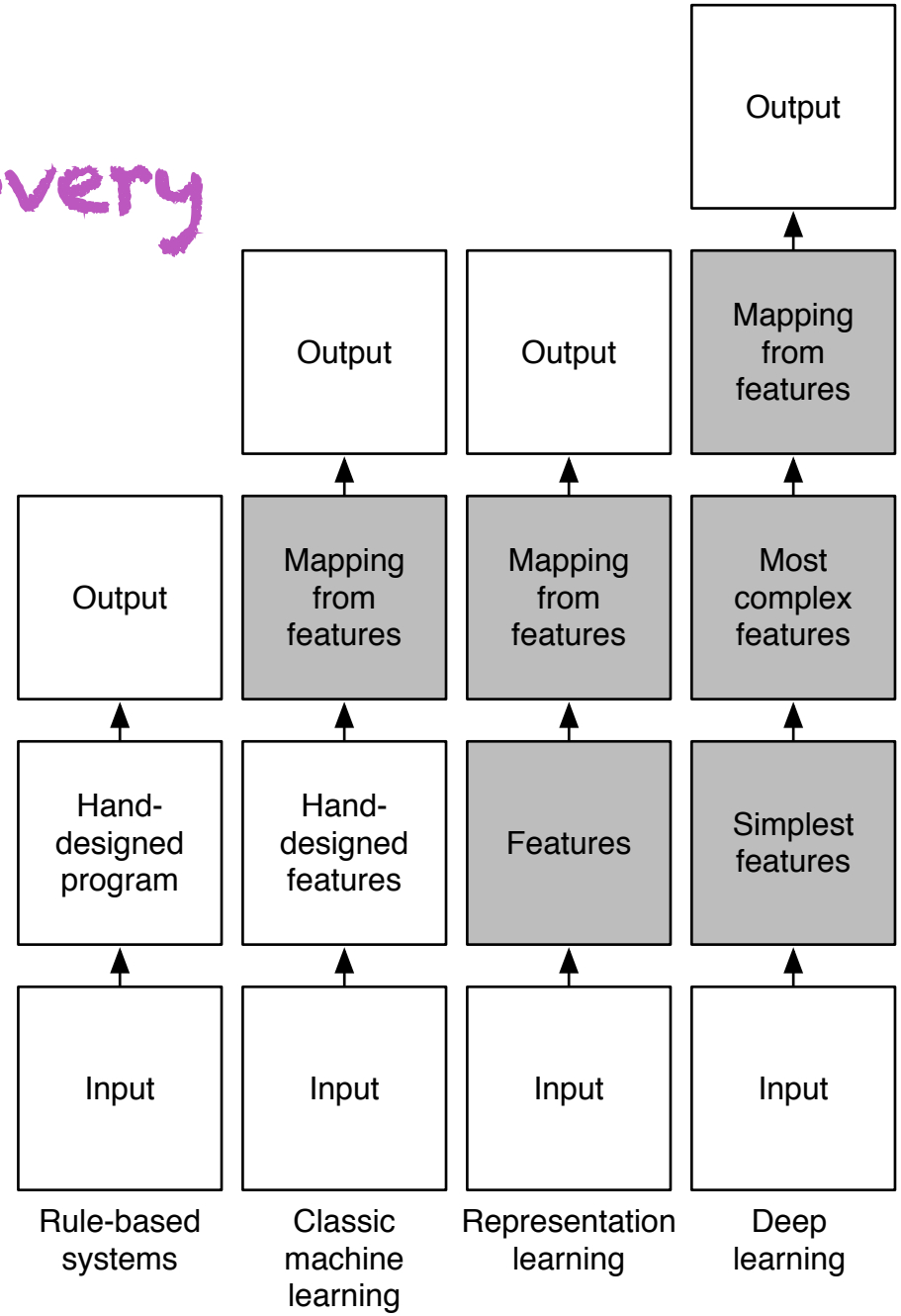
Exploiting compositionality gives an exponential gain in representational power

Distributed representations / embeddings: feature learning

Deep architecture: multiple levels of feature learning

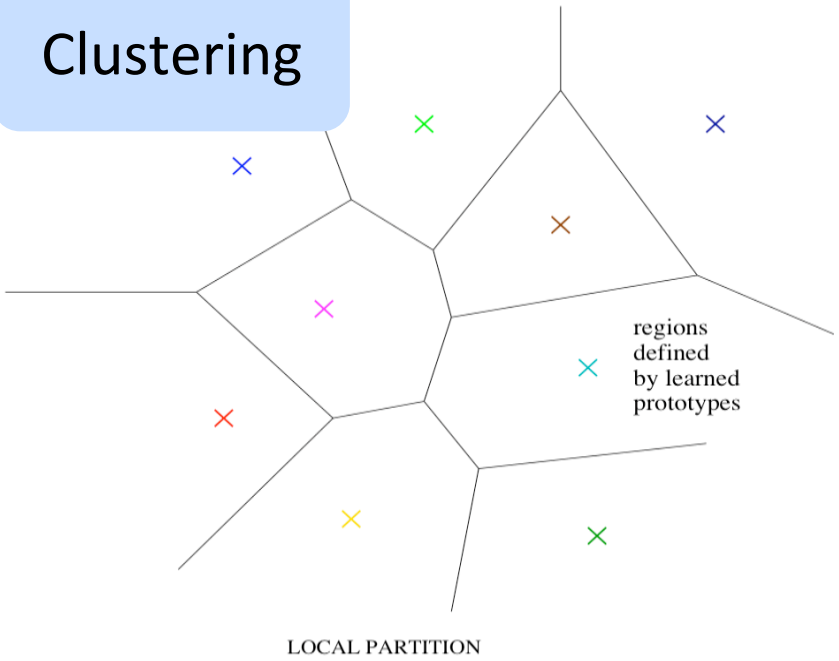
Prior: compositionality is useful to describe the world around us efficiently

Automating Feature Discovery



Non-distributed representations

Clustering



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

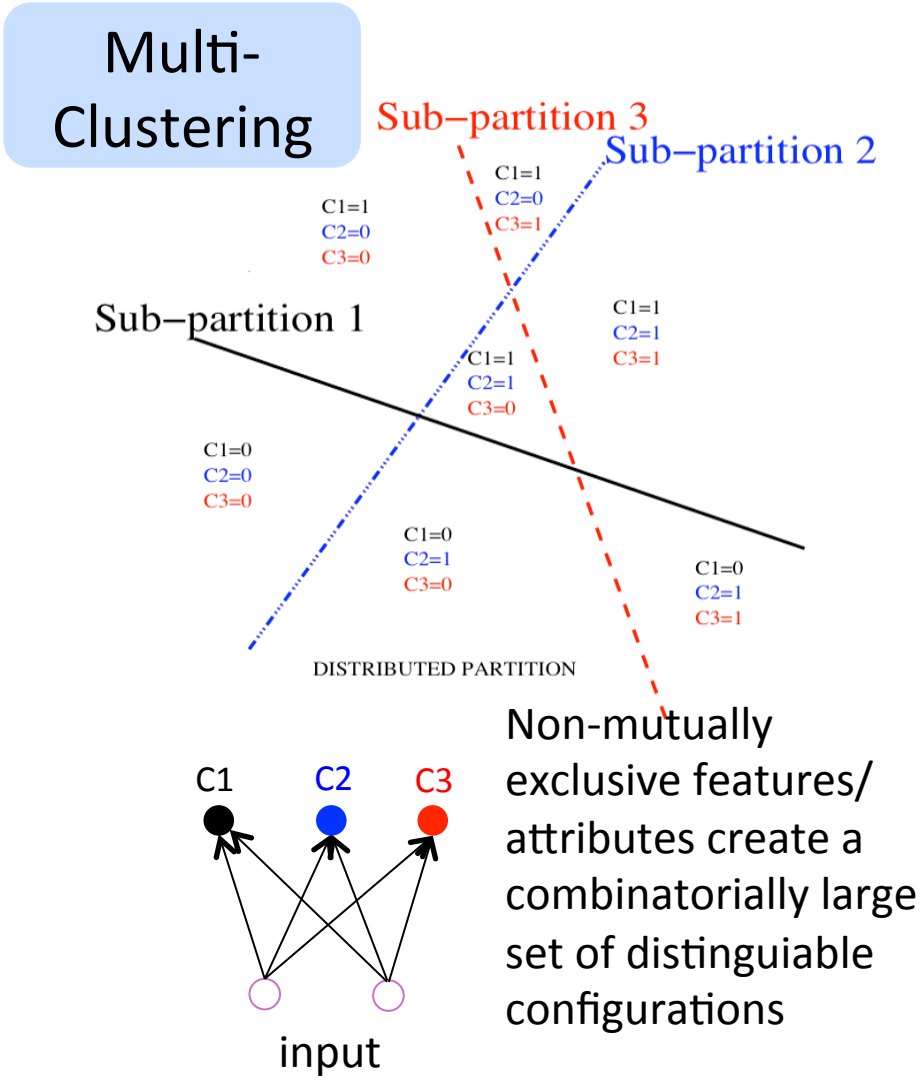
- Parameters for each distinguishable region
- **# of distinguishable regions is linear in # of parameters**

→ 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**



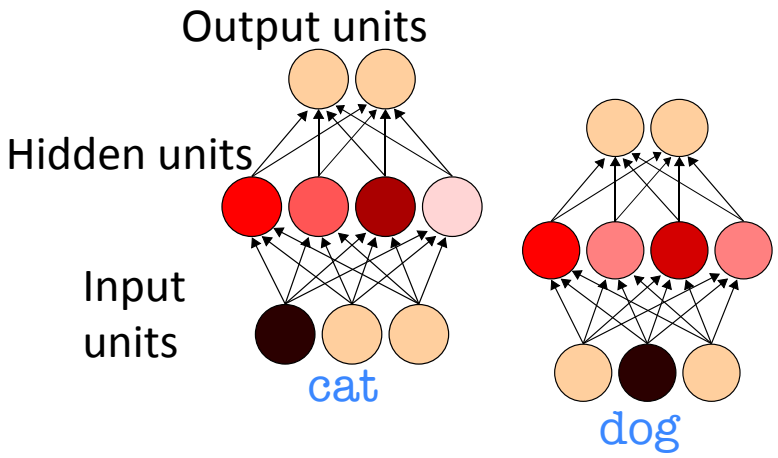
Classical Symbolic AI vs Representation Learning

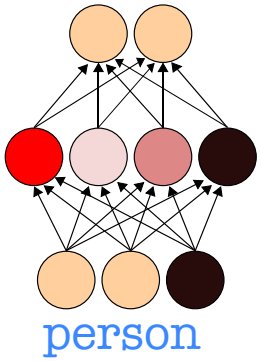
- Two symbols are equally far from each other
- Concepts are not represented by symbols in our brain, but by patterns of activation

  (Connectionism, 1980's)


Geoffrey Hinton

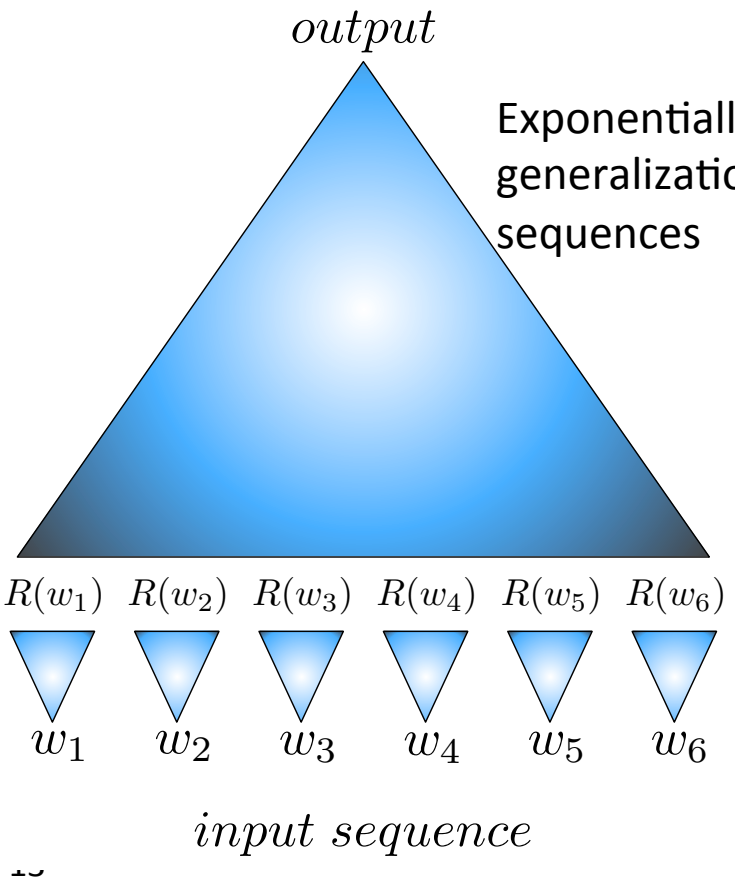

David Rumelhart



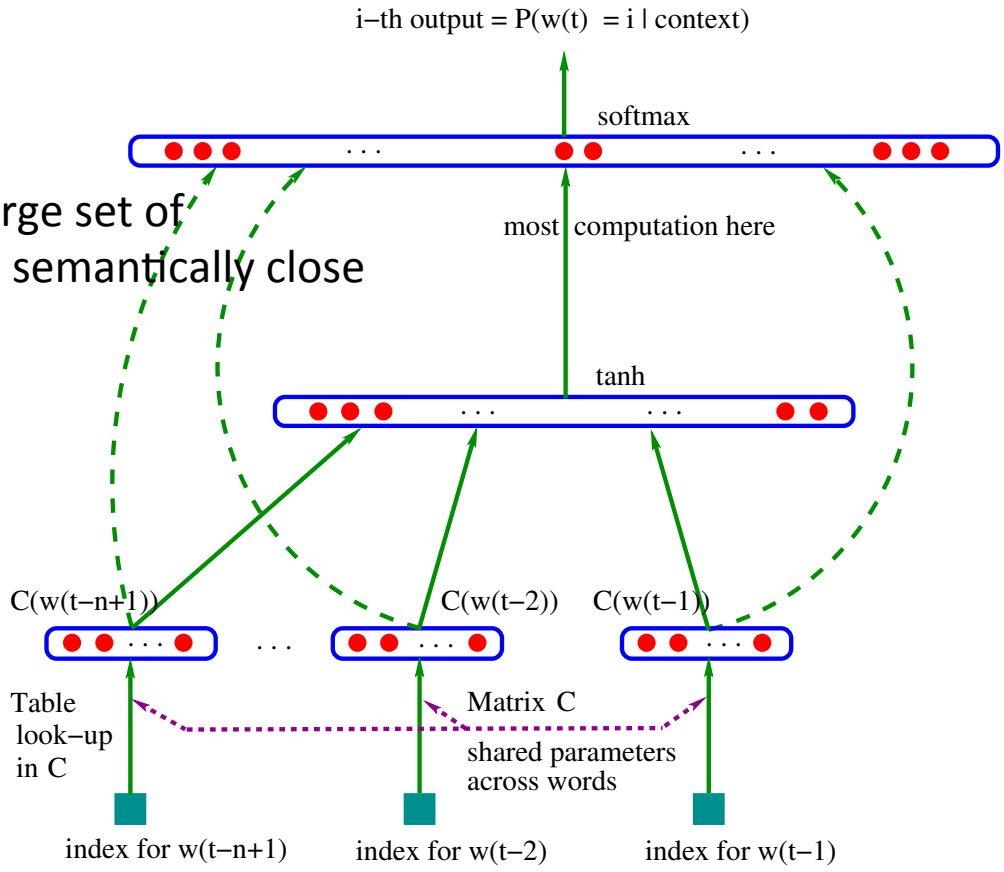


Neural Language Models: fighting one exponential by another one!

- (Bengio et al NIPS'2000)



Exponentially large set of generalizations: semantically close sequences

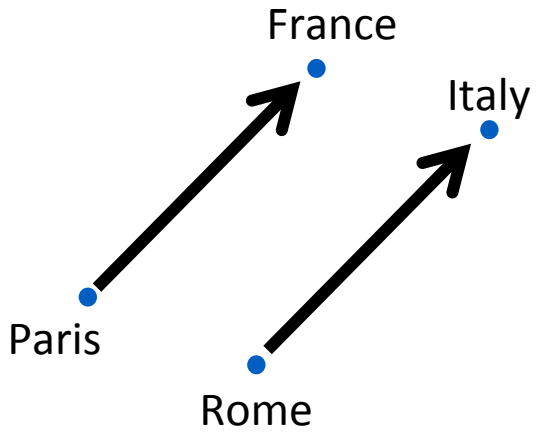


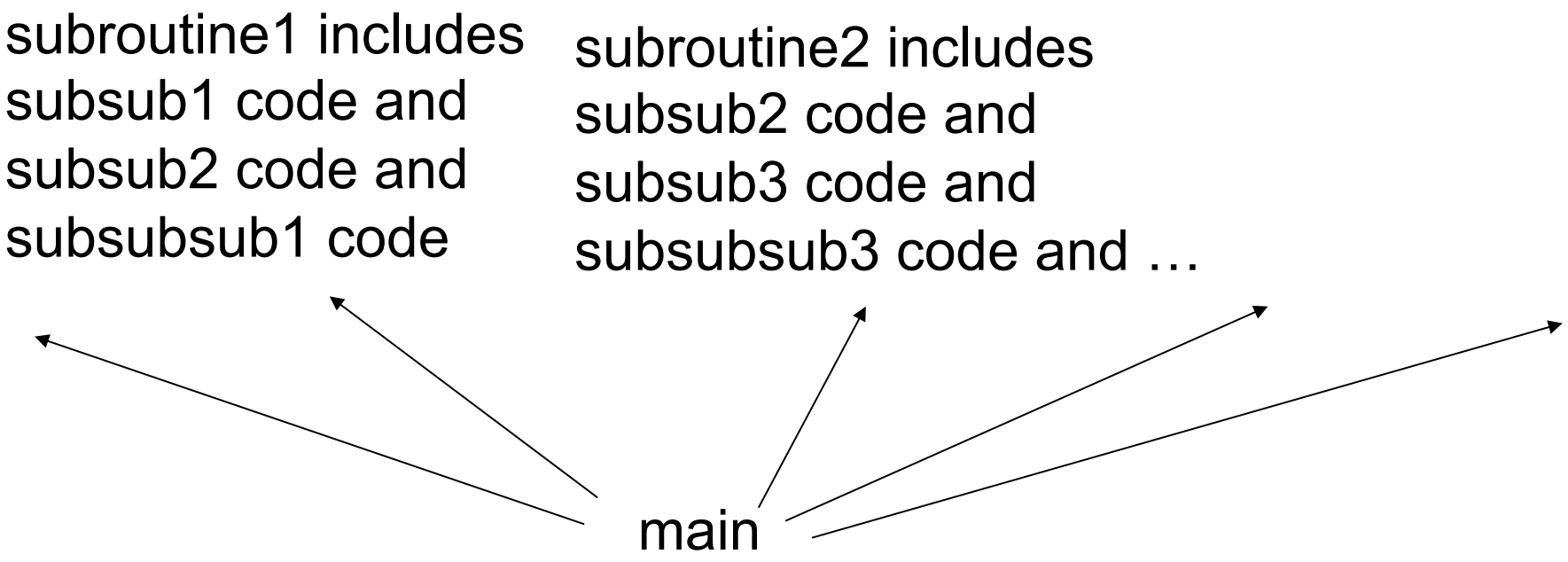
Exponentially large set of possible contexts

Neural word embeddings - visualization
Directions = Learned Attributes

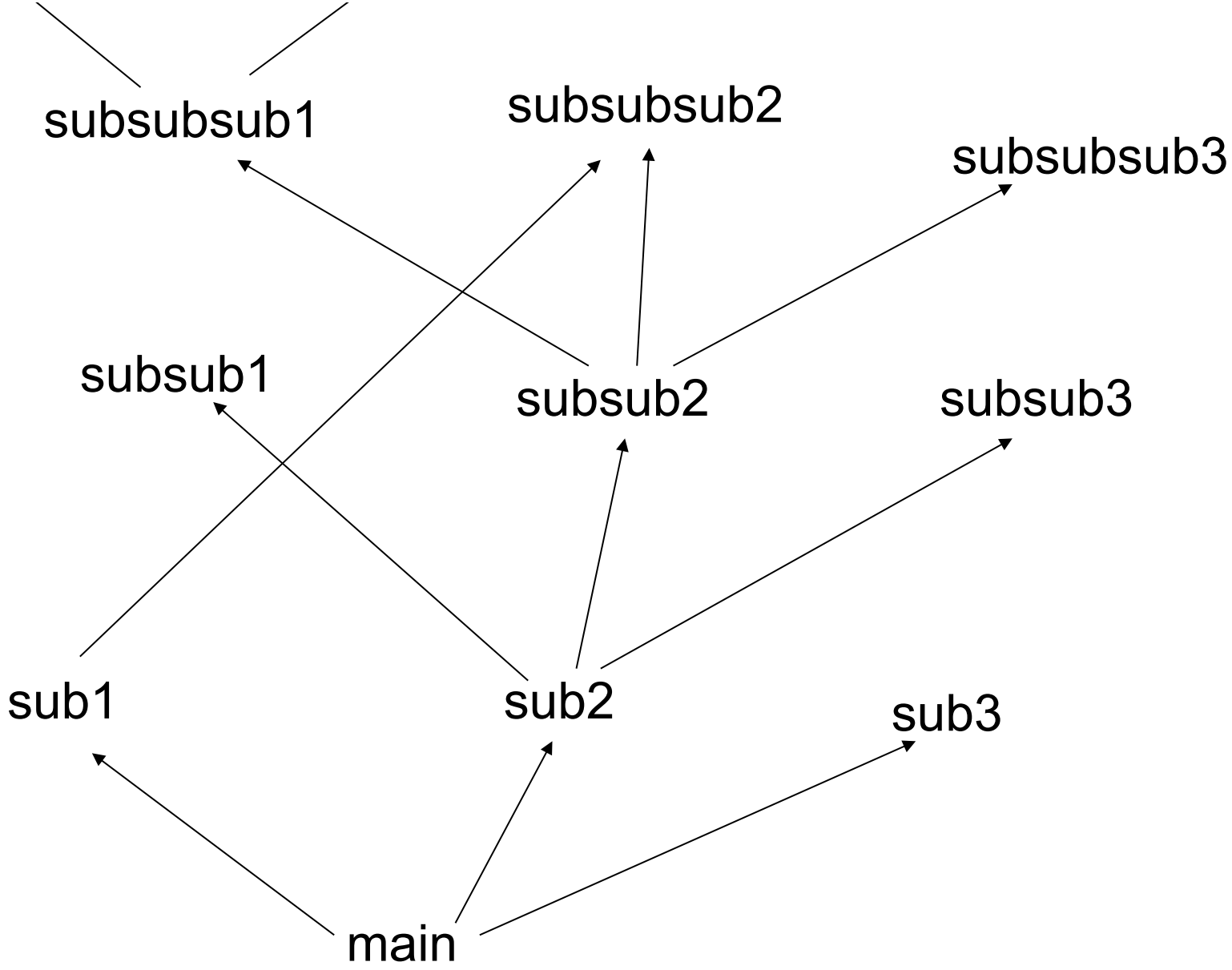
Analogical Representations for Free (Mikolov et al, ICLR 2013)

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





“Shallow” computer program



“Deep” computer program

The Depth Prior can be Exponentially Advantageous

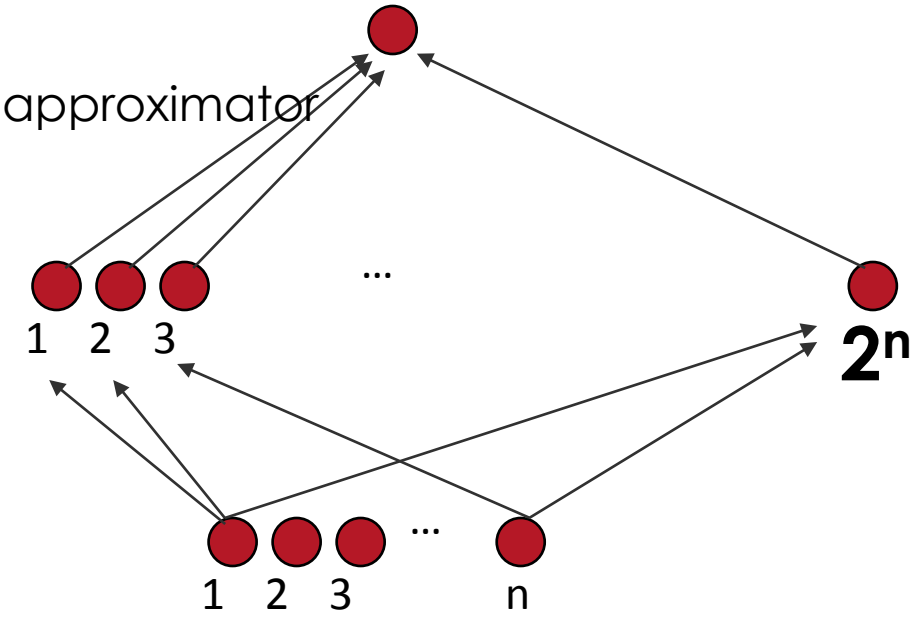
Theoretical arguments:

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

RBMs & auto-encoders = universal approximator

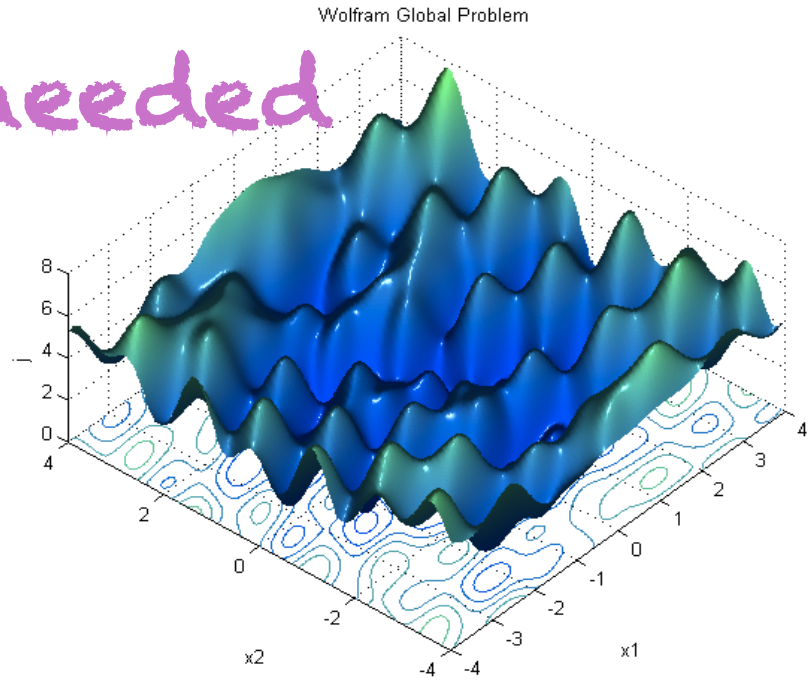
Theorems on advantage of depth:
(Hastad et al 86 & 91, Bengio et al 2007, Bengio & Delalleau 2011, Braverman 2011, Pascanu et al 2014, Montufar et al **NIPS 2014**)

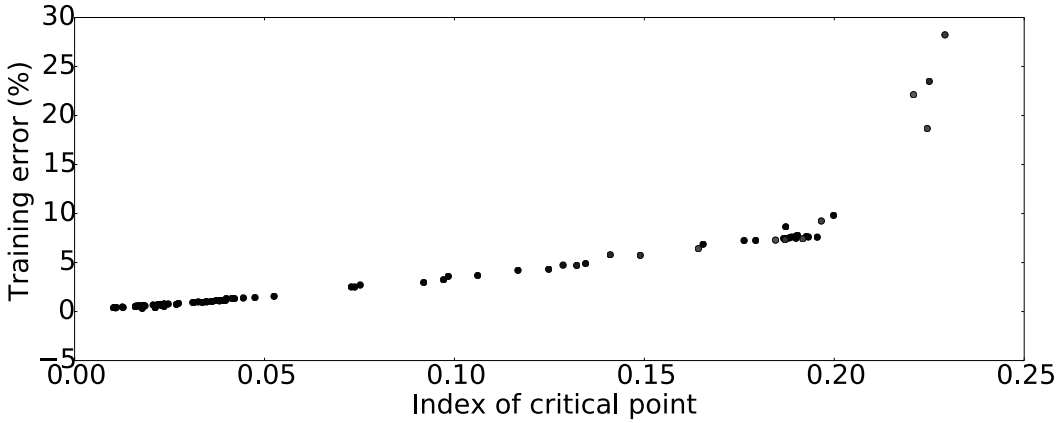
Some functions compactly represented with k layers may require exponential size with 2 layers

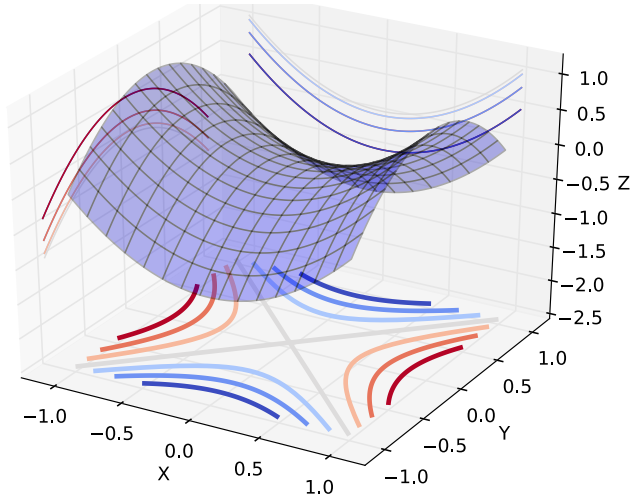


A Myth is Being Debunked: Local Minima in Neural Nets
→ Convexity is not needed

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



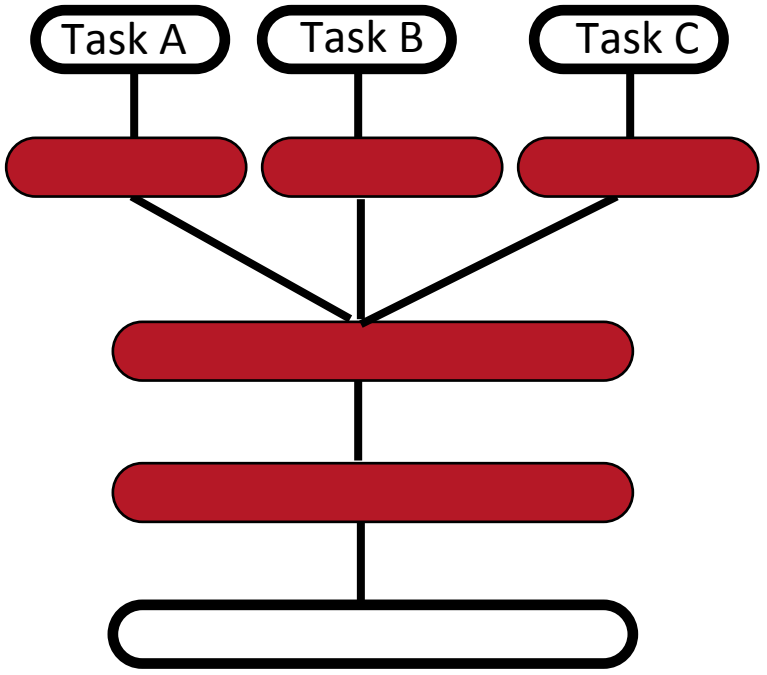




Multi-Task Learning

- Generalizing better to new tasks (tens of thousands!) is crucial to approach AI
- Deep architectures learn good intermediate representations that can be shared across tasks

  (Collobert & Weston ICML 2008, Bengio et al AISTATS 2011)
- Good representations that disentangle underlying factors of variation make sense for many tasks because **each task concerns a subset of the factors**

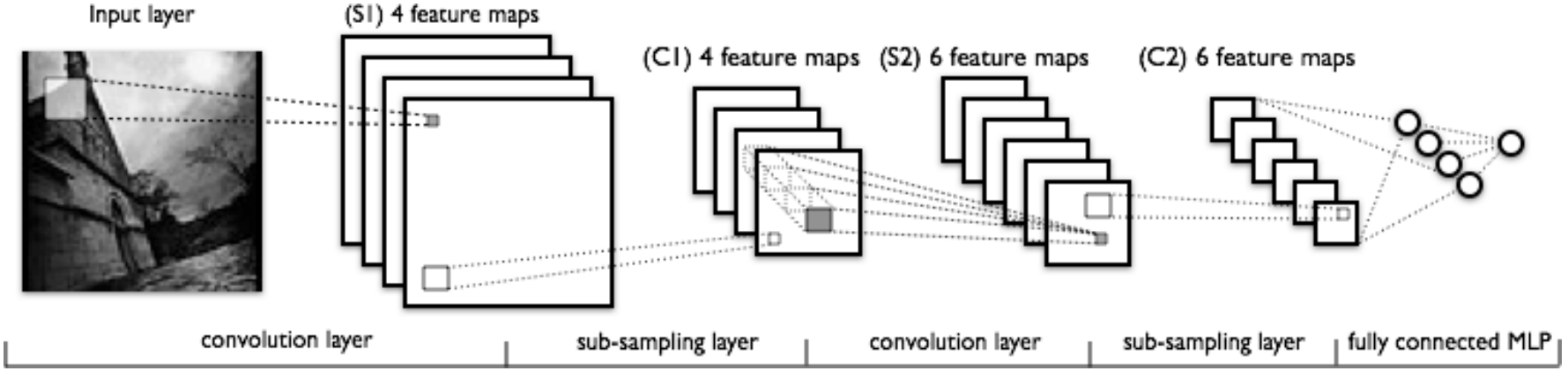


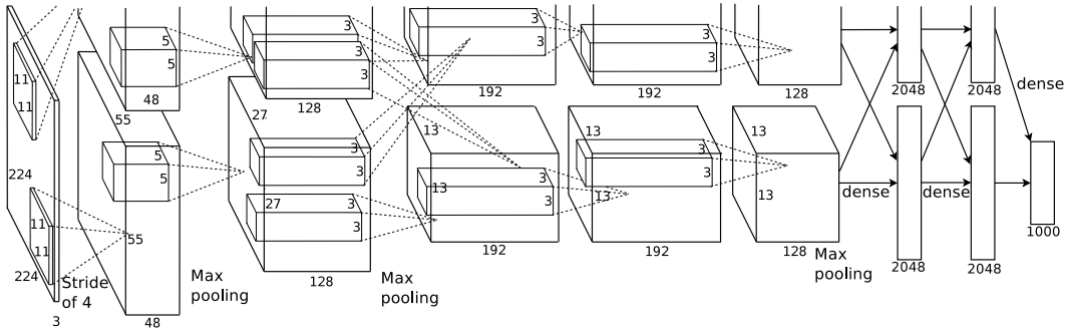
E.g. dictionary, with intermediate concepts re-used across many definitions

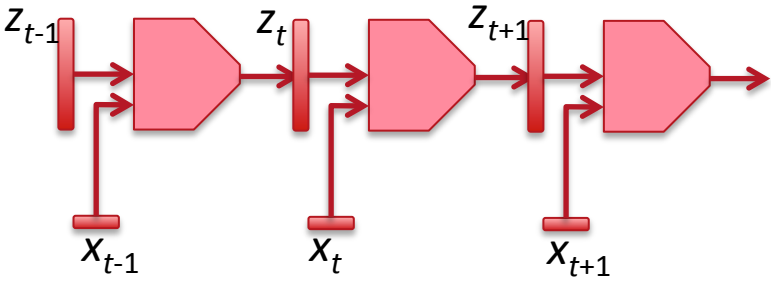
Prior: shared underlying explanatory factors between tasks

Temporal & Spatial Inputs: Convolutional & Recurrent Nets

- Local connectivity across time/space
- Sharing weights across time/space (translation equivariance)
- Pooling (translation invariance, cross-channel pooling for learned invariances)







Recurrent nets (RNNs) can summarize information from the past

Bidirectional RNNs also summarize information from the future

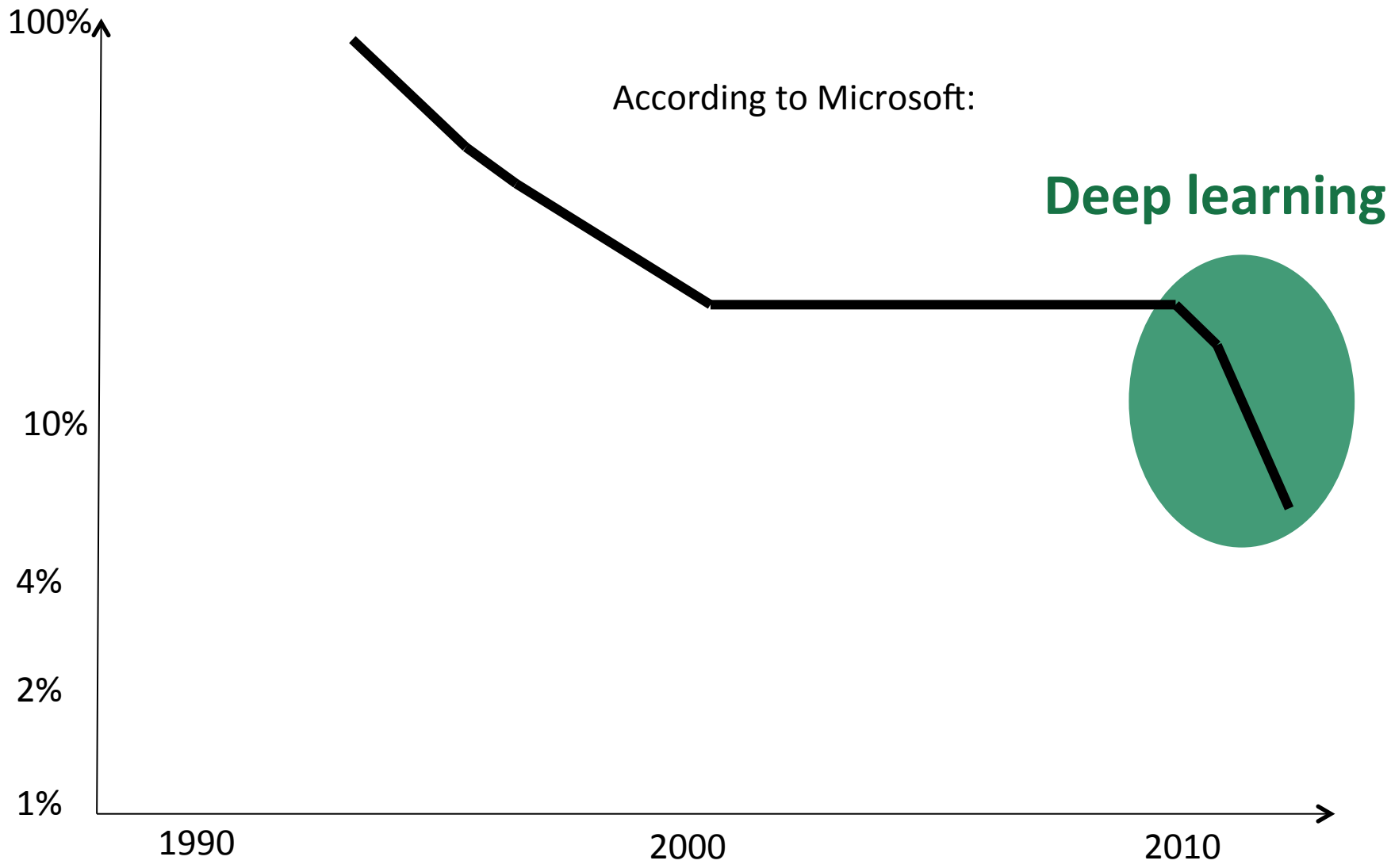
Initial Breakthrough in 2006

Canadian initiative: CIFAR

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

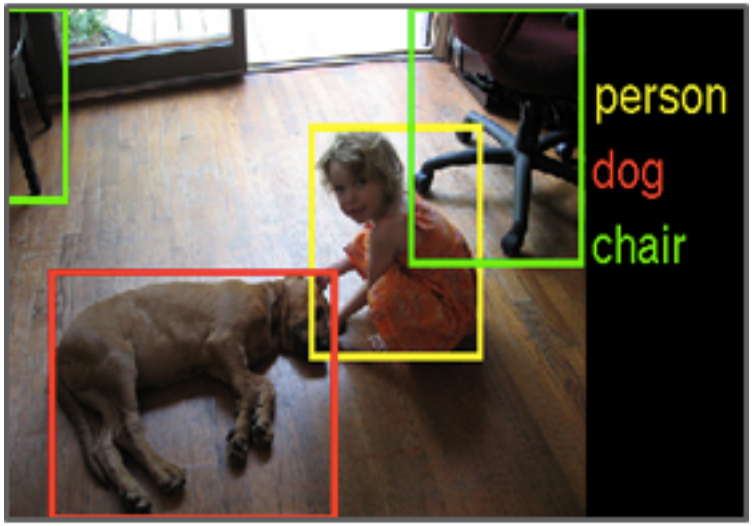
- Unsupervised feature learners:
  - RBMs
  - Auto-encoder variants
  - Sparse coding variants

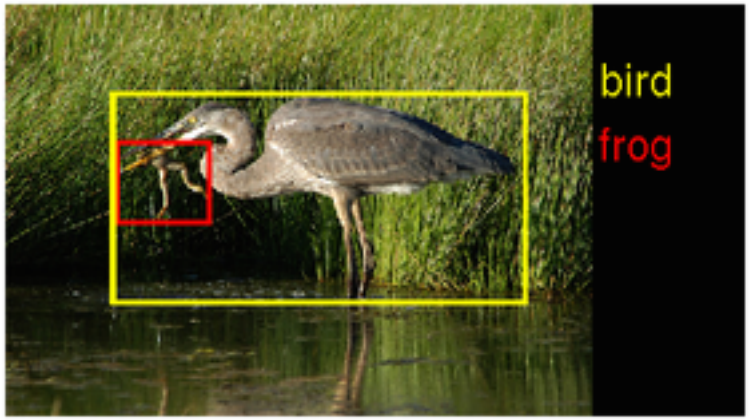
2010-2012: Breakthrough in speech recognition → in Androids by 2012



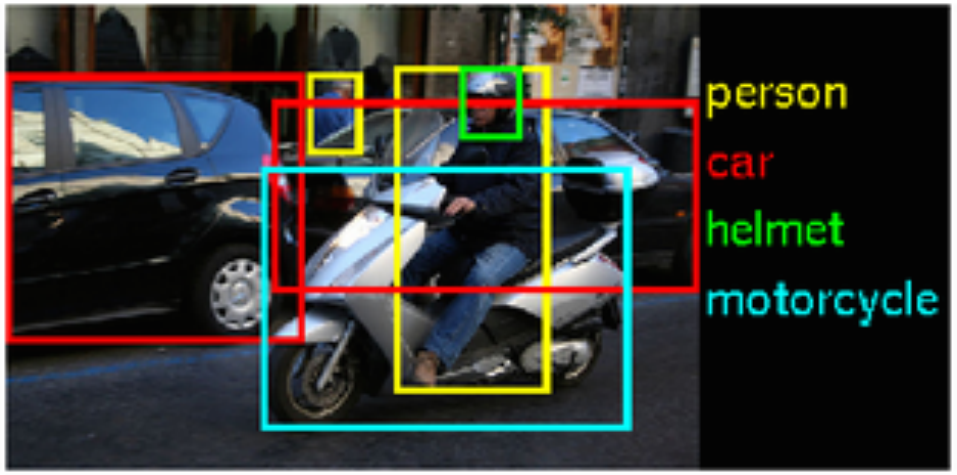
Breakthrough in computer vision: 2012-2015

- 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

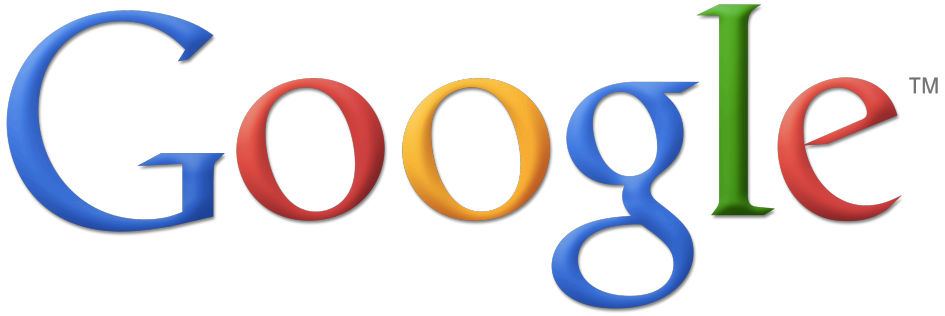
Google Beat Facebook for DeepMind

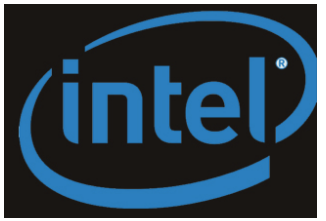
Google Acquires Artificial Intelligence Startup DeepMind For More Than $500M

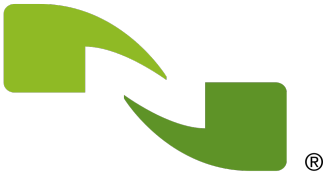
Posted Jan 26, 2014 by Catherine Shu (@catherineshu)

IT Companies are Racing into Deep Learning










Conclusions

- **Distributed representations**:
  - prior that can buy exponential gain in generalization
- **Deep composition of non-linearities**:
  - prior that can buy exponential gain in generalization
- Both yield **non-local generalization**
- Strong evidence that **local minima are not an issue, saddle points**
- **Convolutional nets** have an architecture specialized for images
- **Recurrent nets** have an archicture specialized for sequences
- **Very rapid progress in recent years!**

MILA: Montreal Institute for Learning Algorithms