

# Deep Learning of Representations

**Yoshua Bengio**

U. Montreal

October 21<sup>st</sup>, 2013

Masterclass lecture, UCL Center for Computational  
Statistics and Machine Learning

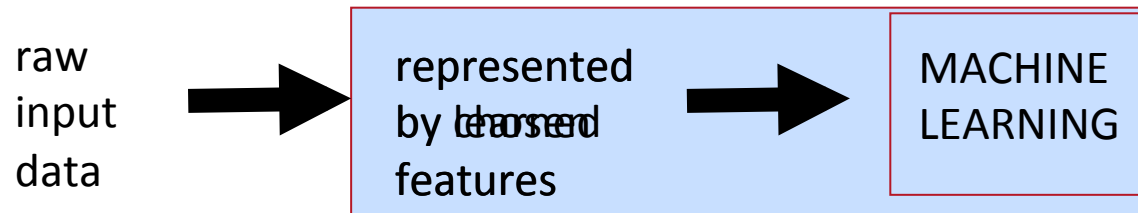


# Ultimate Goals

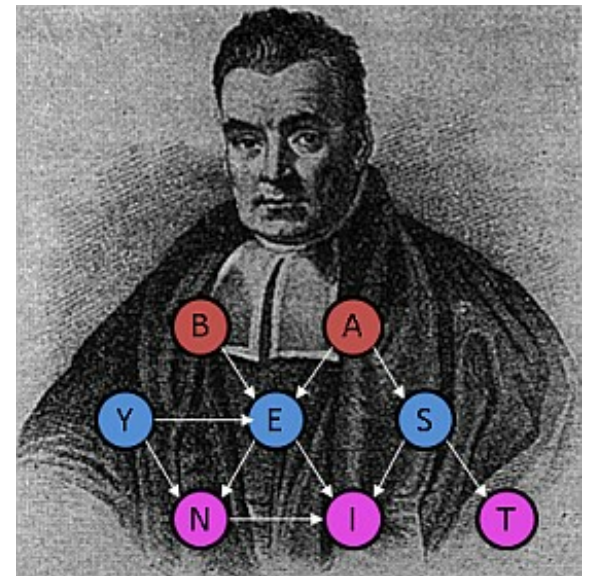
- **AI**
- Needs **knowledge**
- Needs **learning**  
(involves priors + *optimization/search*)
- Needs **generalization**  
(guessing where probability mass concentrates)
- Needs ways to fight the curse of dimensionality  
(exponentially many configurations of the variables to consider)
- Needs disentangling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

# 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)



$\Phi_I(\cdot)$

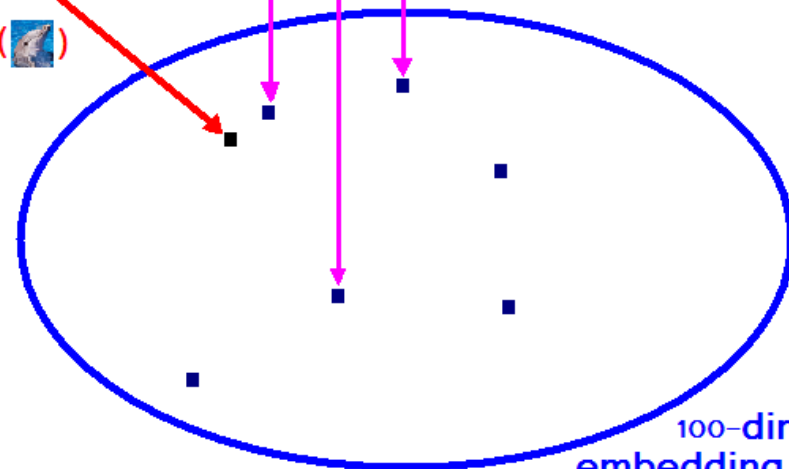
$\Phi_W(\text{DOLPHIN})$

DOLPHIN

OBAMA

EIFFEL TOWER

.....



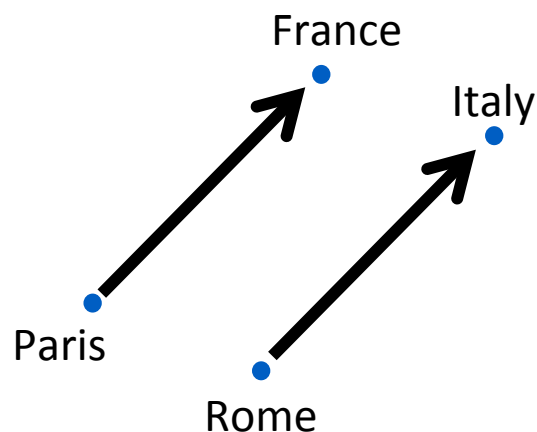
Learn  $\Phi_I(\cdot)$  and  $\Phi_W(\cdot)$  to optimize precision@k.

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



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

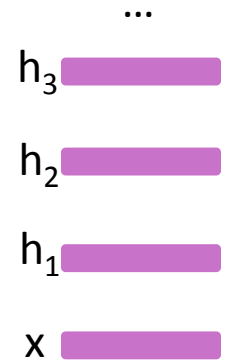
- Semantic relations appear as linear relationships in the space of learned representations
- King – Queen  $\approx$  Man – Woman
- Paris – France + Italy  $\approx$  Rome



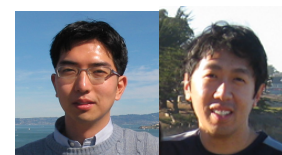
# Deep Representation Learning

Learn multiple levels of representation of increasing complexity/abstraction

- theory: exponential gain
- brains are deep
- cognition is compositional
- Better mixing (Bengio et al, ICML 2013)
- **They work! SOTA on industrial-scale AI tasks (object recognition, speech recognition, language modeling, music modeling)**



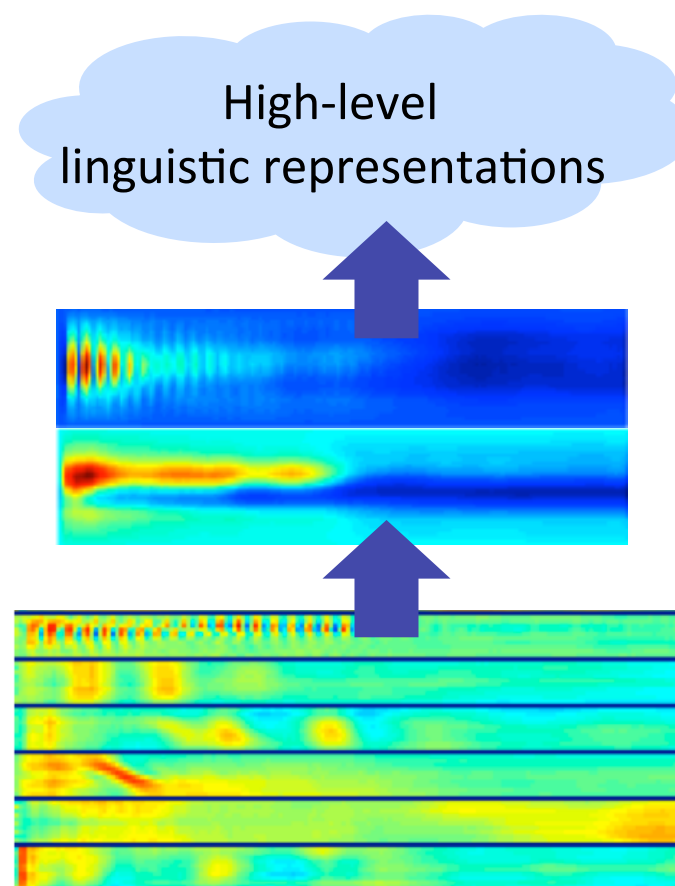
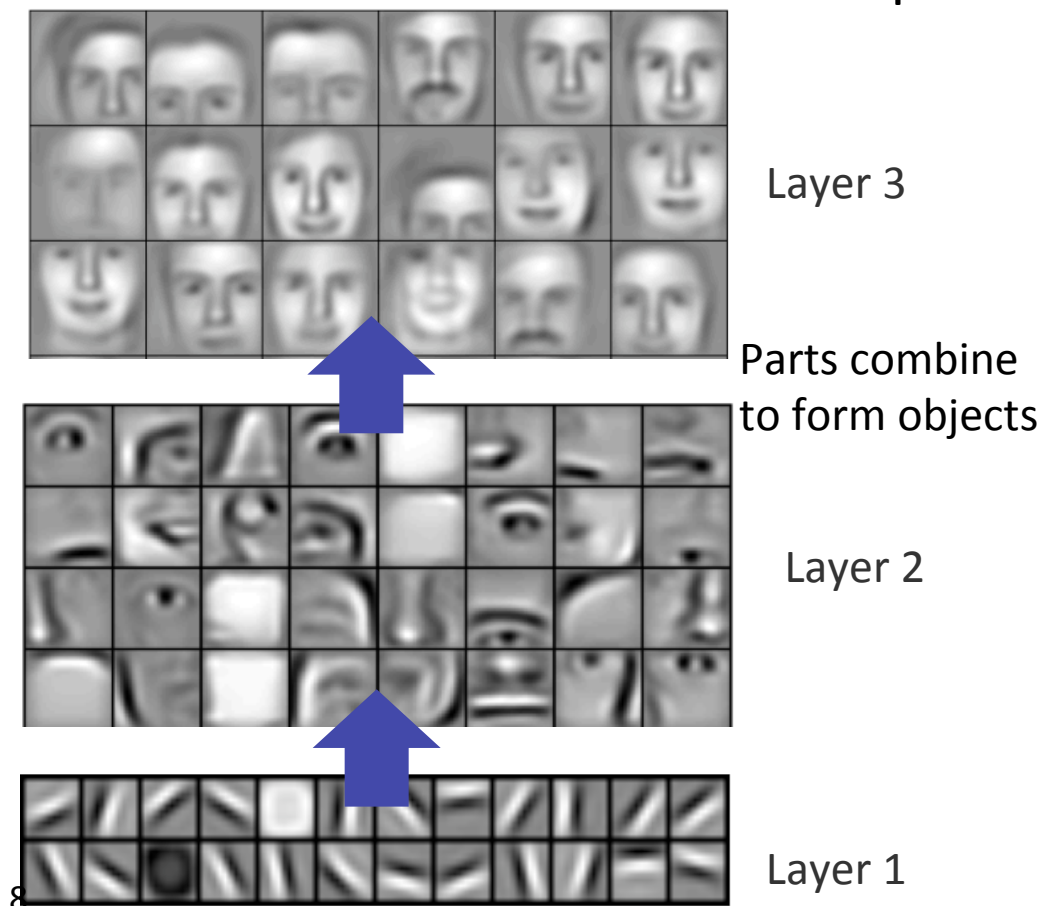
# Learning multiple levels of representation



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

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

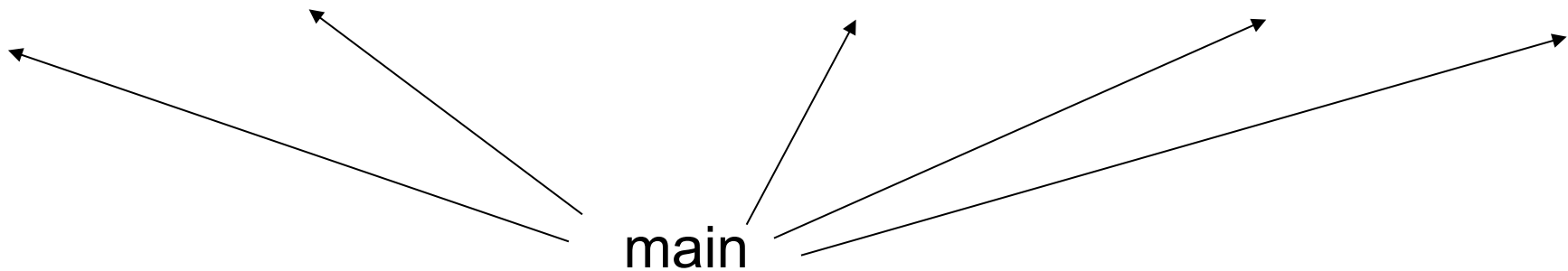
Successive model layers learn deeper intermediate representations



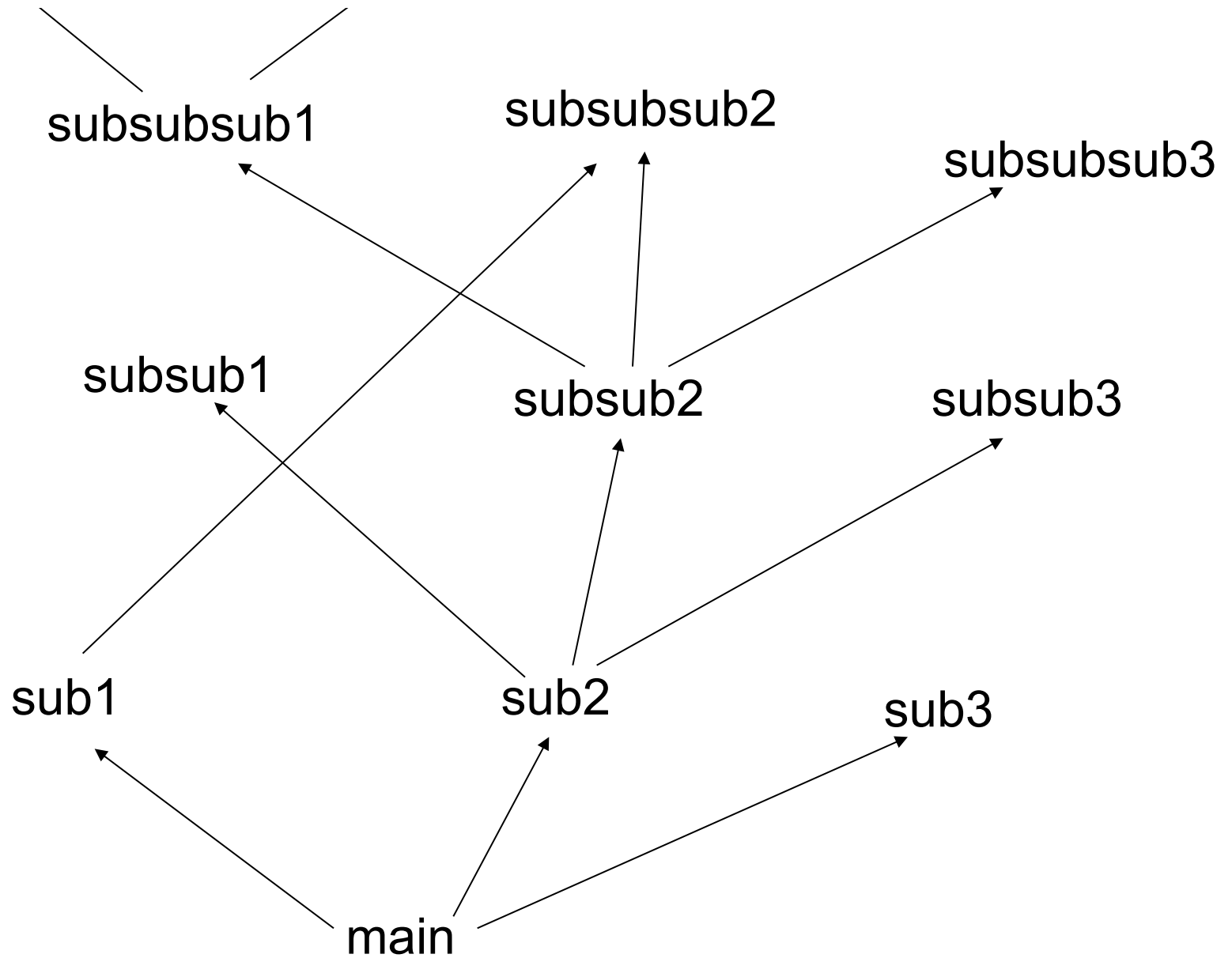
**Prior: underlying factors & concepts compactly expressed w/ multiple levels of abstraction**

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**

# Deep Architectures are More Expressive

Theoretical arguments:

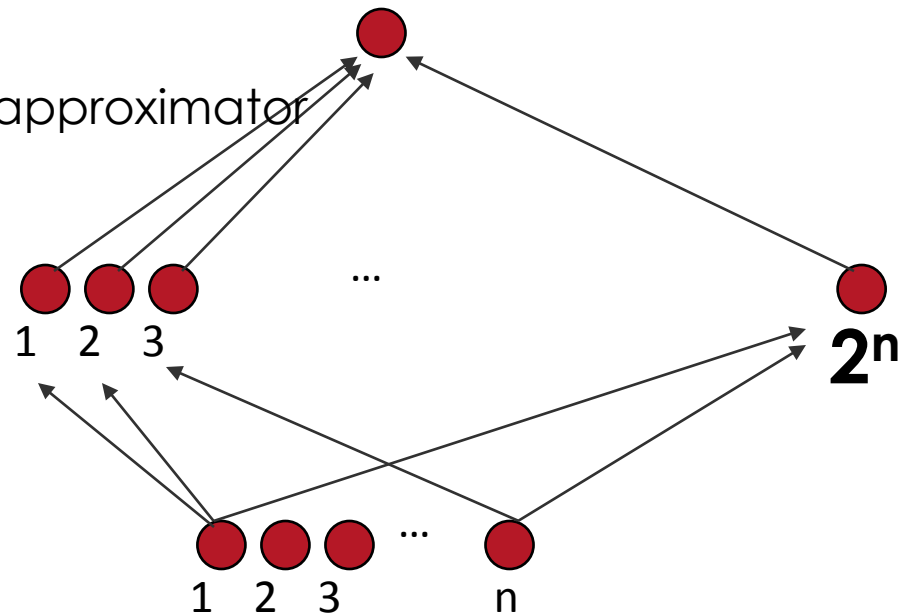
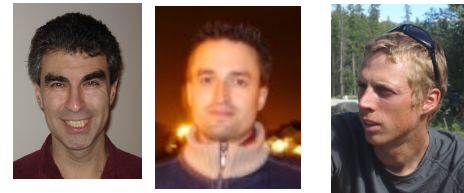
2 layers of {  
Logic gates  
Formal neurons  
RBF units

= 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)

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



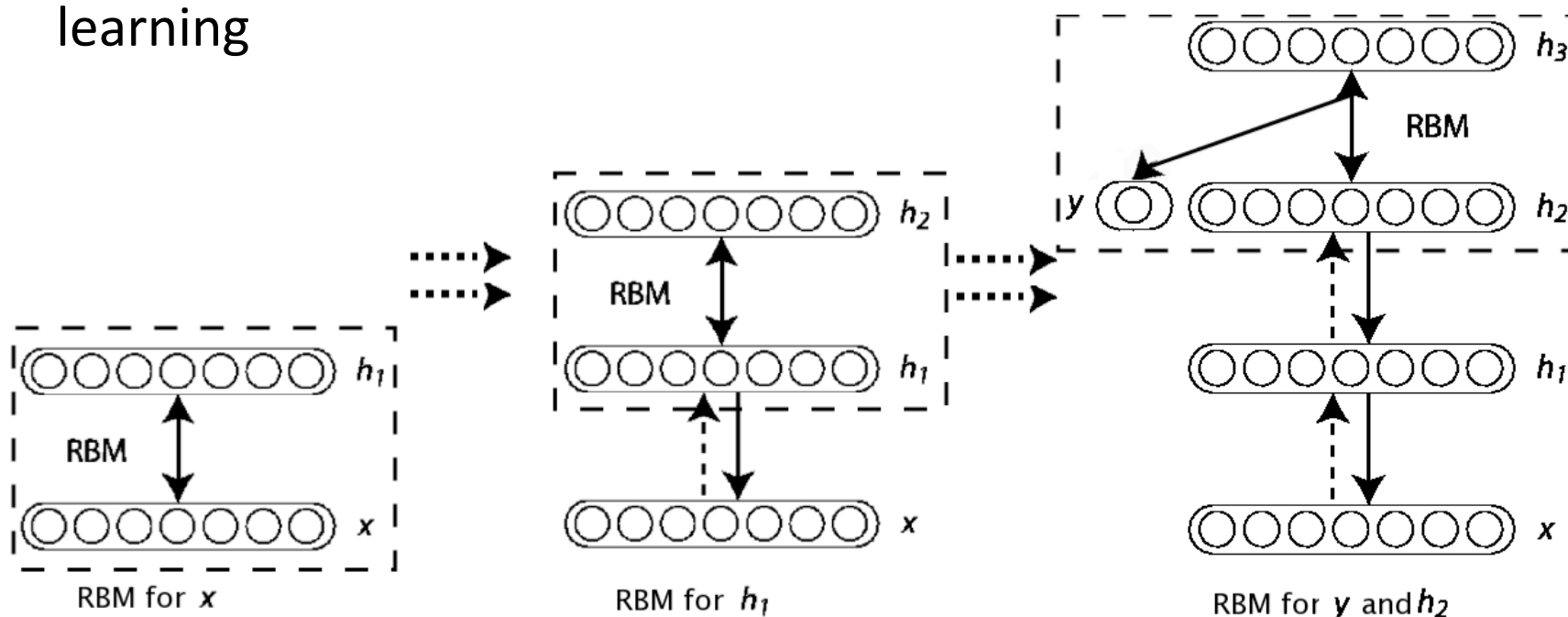
# Breakthrough in 2006

- 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



# Stacking Single-Layer Learners

- One of the big ideas from 2006: layer-wise unsupervised feature learning

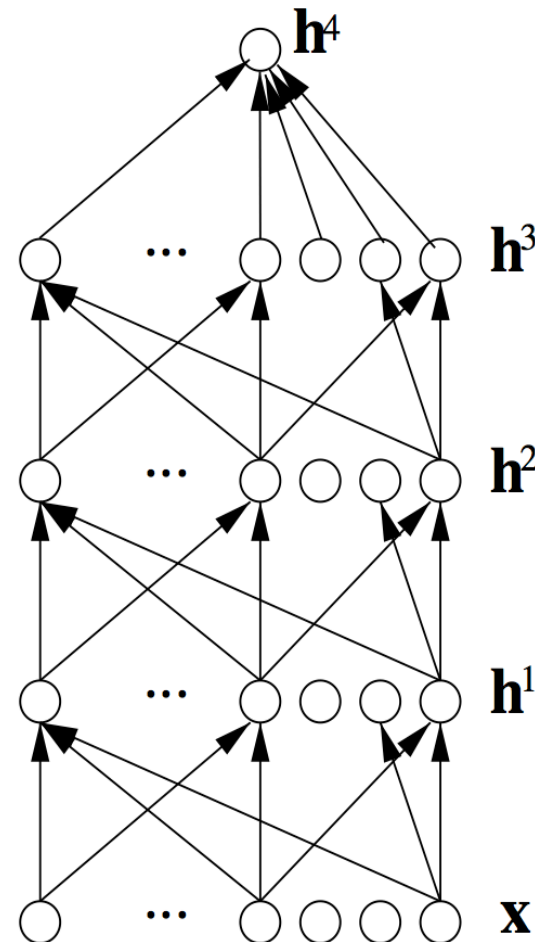


Stacking Restricted Boltzmann Machines (RBM)  $\rightarrow$  Deep Belief Network (DBN)

Stacking regularized auto-encoders  $\rightarrow$  deep neural nets

# 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.



# Deep Learning in the News



Researcher Dreams Up Machines  
That Learn Without Humans  
06.27.13

Yoshua Bengio. Image: C

## The New York Times

Scientists See Promise in  
Deep-Learning Programs

John Markoff

November 23, 2012

THE GLOBE AND MAIL

CANADA'S NATIONAL NEWSPAPER • FOUNDED 1844

Google taps U  
of T professor  
to teach  
context to  
computers  
03.11.13

15



The Man Behind the Google Brain: Andrew Ng  
and the Quest for the New AI

BY DANIELA HERNANDEZ 05.07.13 6:30 AM

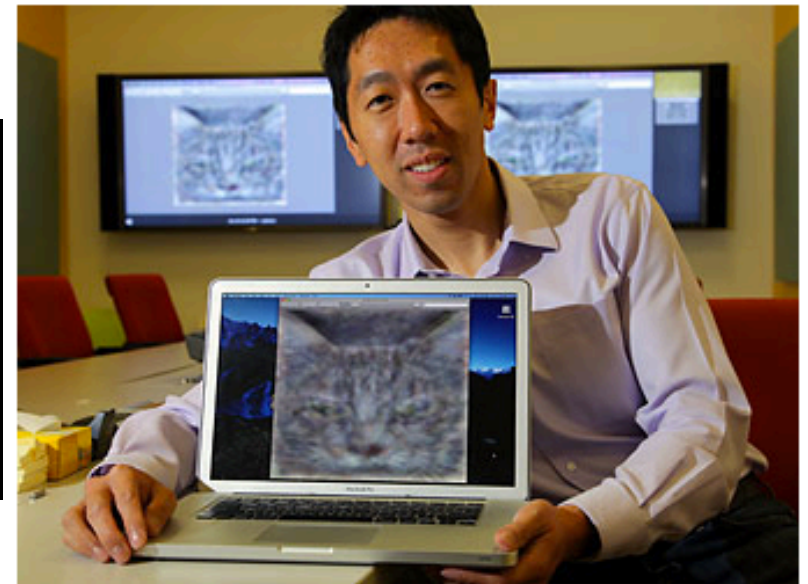
## The New York Times

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

DIGITAL SUBSCRIPTION: 4 WEEKS

ING DIRECT

Follow Us 



Jim Wilson/The New York Times

### Despite Itself, a Simulated Brain Seeks Cats

By JOHN MARKOFF 12 minutes ago

A Google research team, led by Andrew Y. Ng, above, and Jeff Dean, created a neural network of 16,000 processors that reflected human obsession with Internet felines.

# 10 BREAKTHROUGH TECHNOLOGIES 2013

Intr

## 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

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

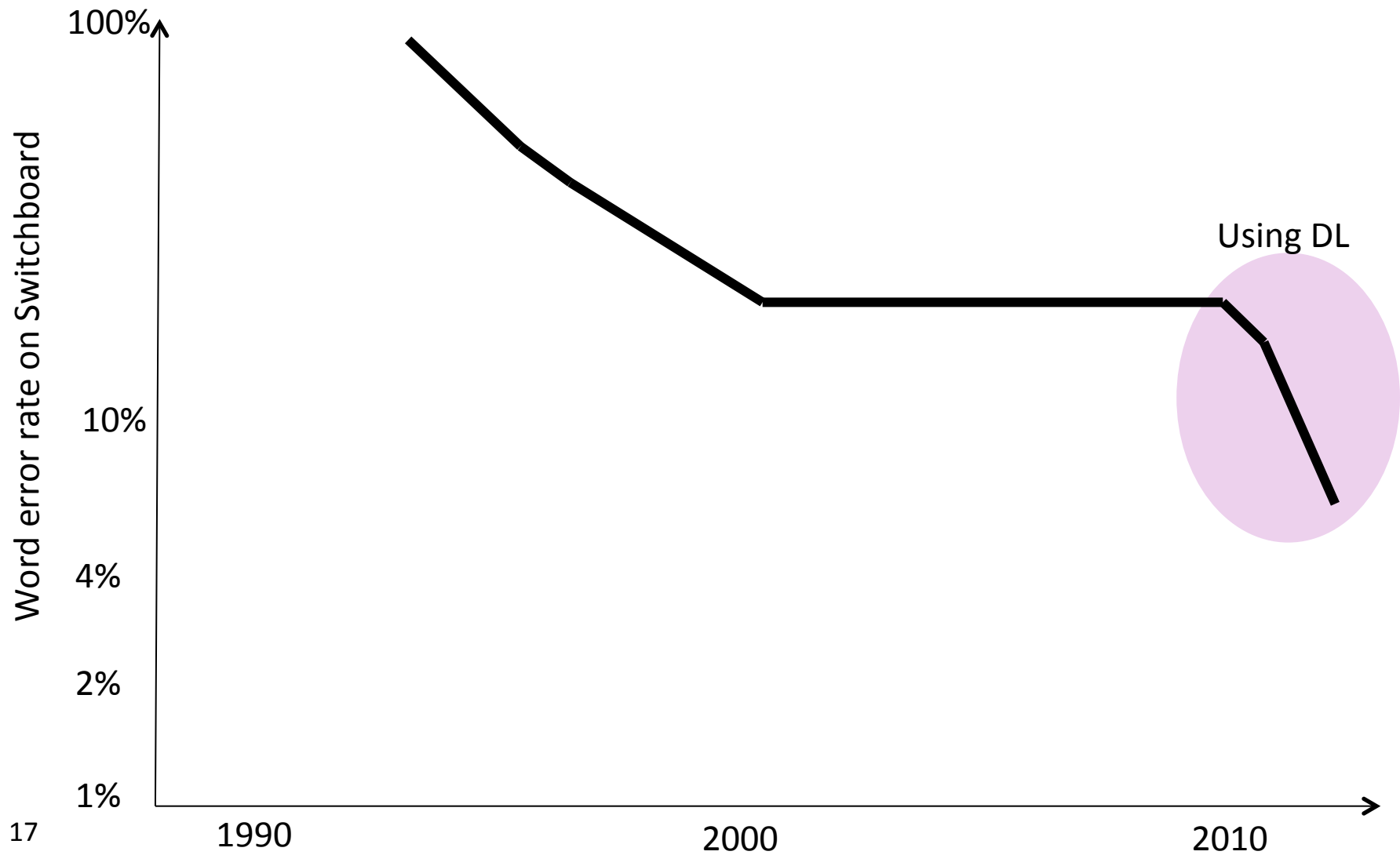
## Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely

## Big Ph

Coll  
ana  
from  
pho

# The dramatic impact of Deep Learning on Speech Recognition



# Some Applications of DL

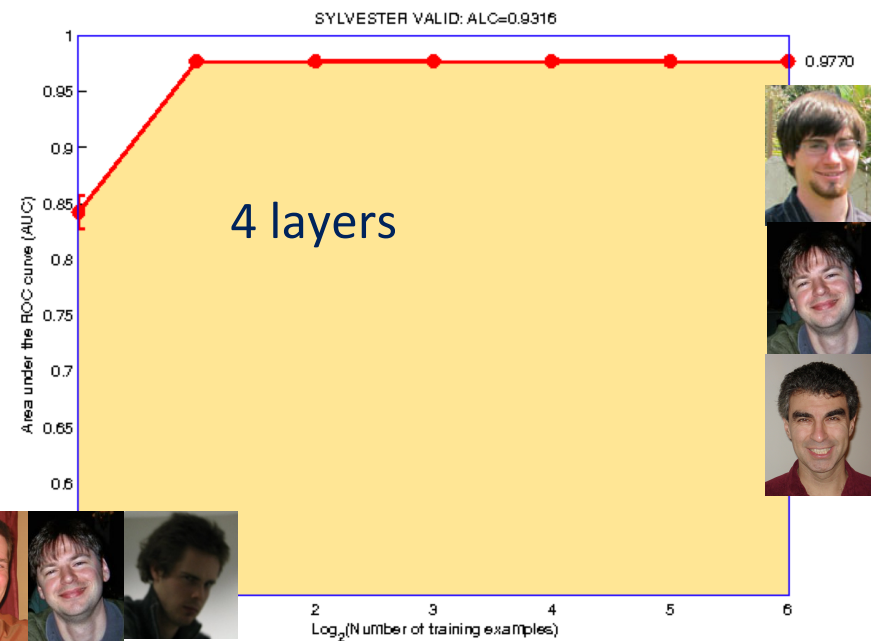
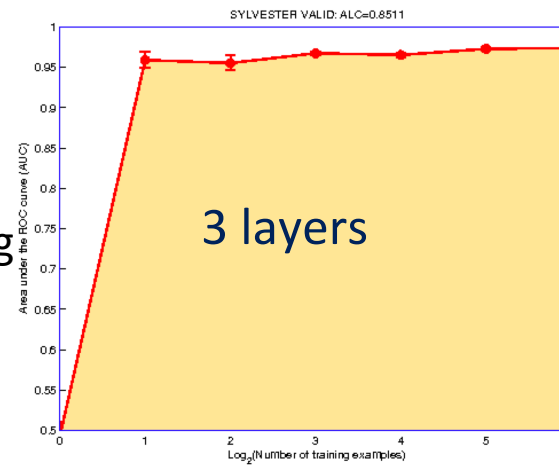
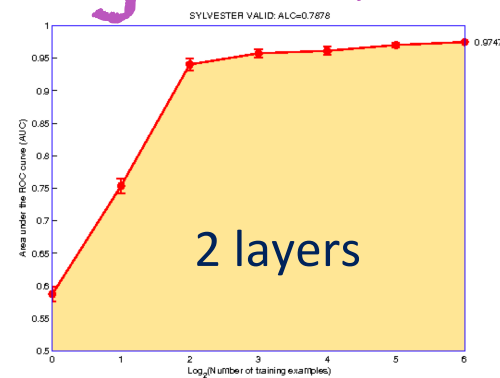
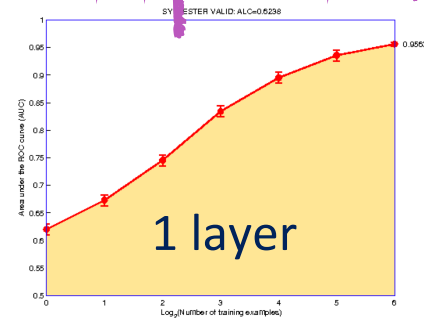
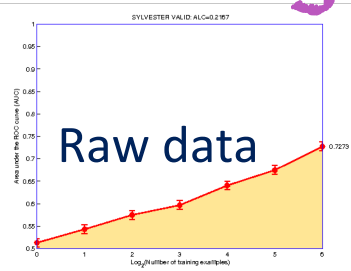
- **Language Modeling** (Speech Recognition, Machine Translation)
- **Acoustic Modeling** (**speech recognition**, music modeling)
- **NLP syntactic/semantic tagging** (Part-Of-Speech, chunking, Named Entity Recognition, Semantic Role Labeling, Parsing)
- **NLP applications**: sentiment analysis, paraphrasing, question-answering, Word-Sense Disambiguation
- **Object recognition in images**: photo search and image search: handwriting recognition, document analysis, handwriting synthesis, **superhuman traffic sign classification**, street view house numbers, **emotion detection from facial images**, roads from satellites.
- **Personalization**/recommendation/fraud/ads
- **Molecular properties**: QSAR, quantum calculations



# How do humans generalize from very few examples?

- They **transfer** knowledge from previous learning:
  - **Abstract** (i.e. deep) representations
  - Explanatory factors
- Previous learning from: unlabeled data
  - + labels for other tasks

# 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

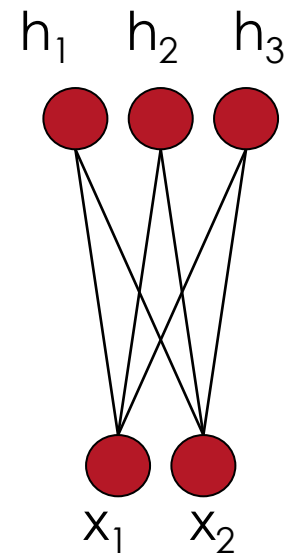


# Undirected Models: the Restricted Boltzmann Machine

[Hinton et al 2006]



- Latent (hidden) variables  $h$  model high-order dependencies
- No easy way to compute exact normalization & gradient, but MCMC approximations are used



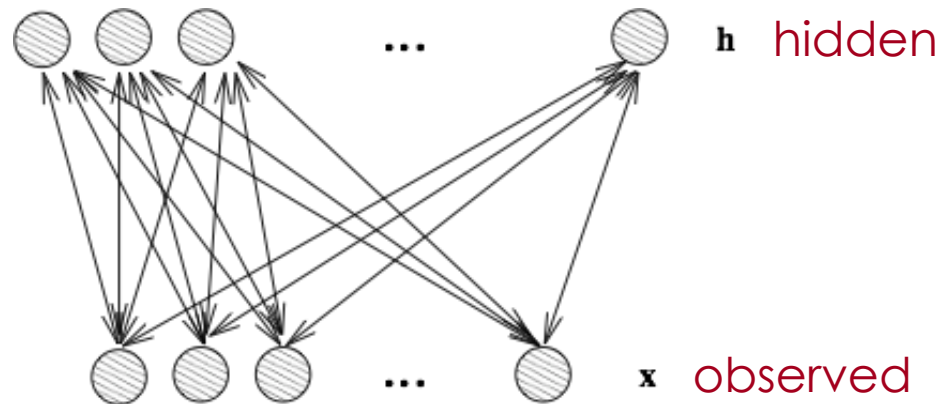
- See Bengio (2009) detailed monograph/review: *"Learning Deep Architectures for AI"*.
- See Hinton (2010) *"A practical guide to training Restricted Boltzmann Machines"*



# Restricted Boltzmann Machine (RBM)

$$P(x, h) = \frac{1}{Z} e^{b^T h + c^T x + h^T W x} = \frac{1}{Z} e^{\sum_i b_i h_i + \sum_j c_j x_j + \sum_{i,j} h_i W_{ij} x_j}$$

Needs to sample examples generated by the model during training to estimate gradient through  $Z$ , using MCMC

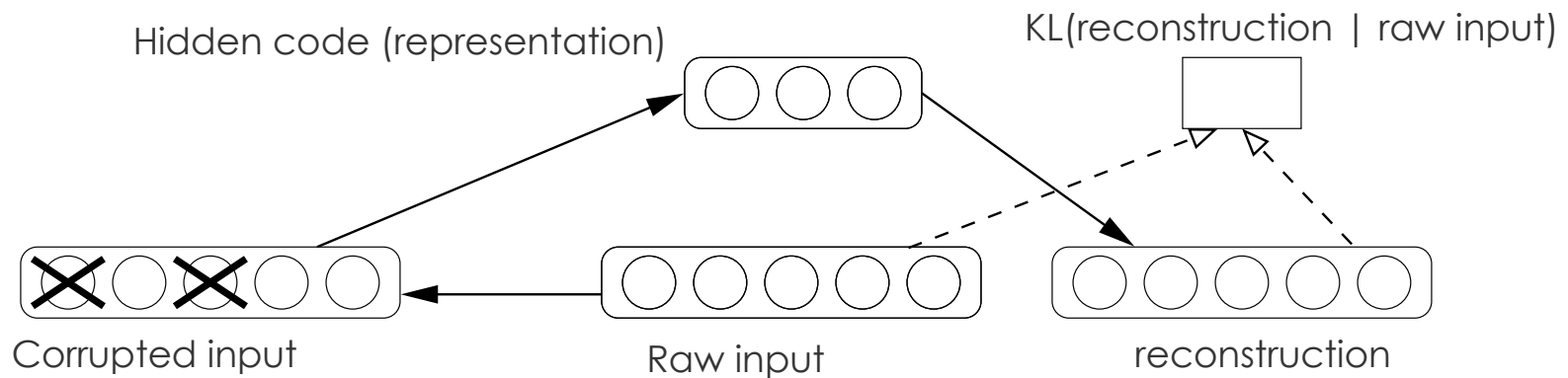


# Denoising Auto-Encoder

(Vincent et al 2008)



- Alternative building-block
  - Corrupt the input
  - Try to reconstruct the uncorrupted input



- Novel probabilistic interpretations: score matching (Vincent 2011, Alain & Bengio ICLR 2013) or as the transition kernel of a Markov chain (Bengio et al, NIPS 2013)



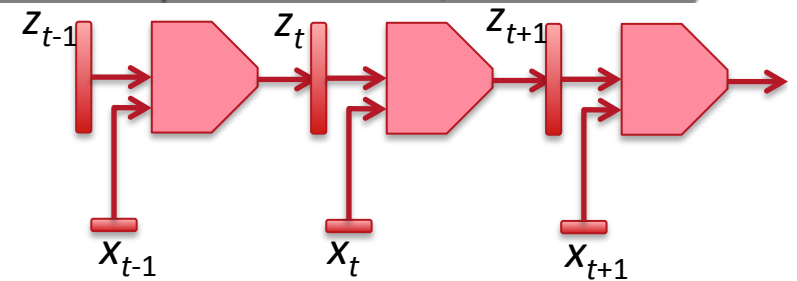
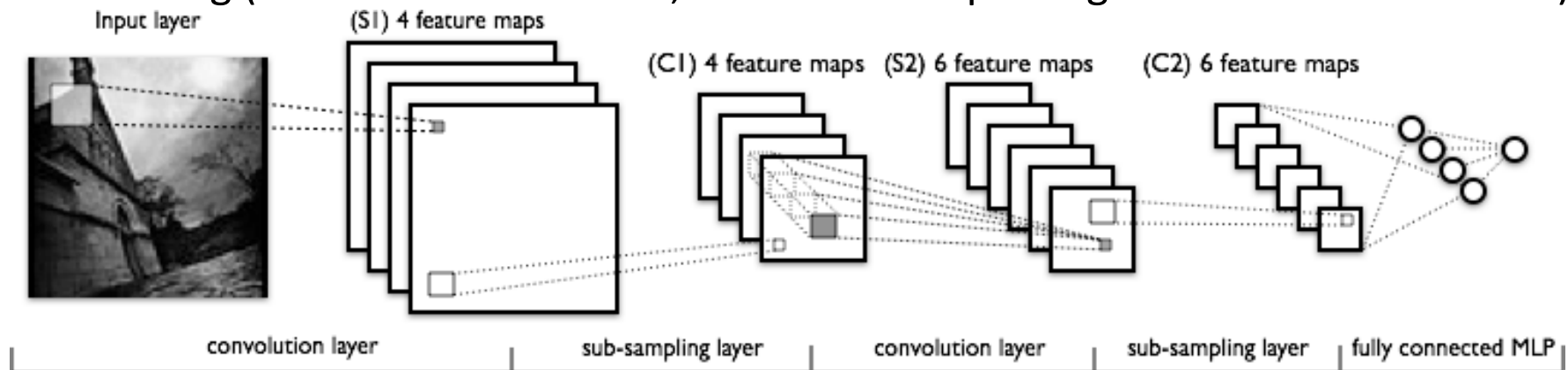
# Stochastic Neurons as Regularizer:

Improving neural networks by preventing co-adaptation of feature detectors (Hinton et al 2012, arXiv)

- **Dropouts** trick: during training multiply neuron output by random bit ( $p=0.5$ ), during test by 0.5
- Generalize denoising auto-encoders, by corrupting every layer
- Works better with rectifiers, even better with maxout (Goodfellow et al. ICML 2013)
- Equivalent to averaging over exponentially many architectures
  - Used by Krizhevsky et al to break through ImageNet SOTA
  - Also improves SOTA on CIFAR-10 (18%  $\rightarrow$  16% err)
  - Knowledge-free MNIST with DBMs (.95%  $\rightarrow$  .79% err)
  - TIMIT phoneme classification (22.7%  $\rightarrow$  19.7% err)

# A Prior for Temporal & Spatial Inputs: Convolutional & Recurrent Nets

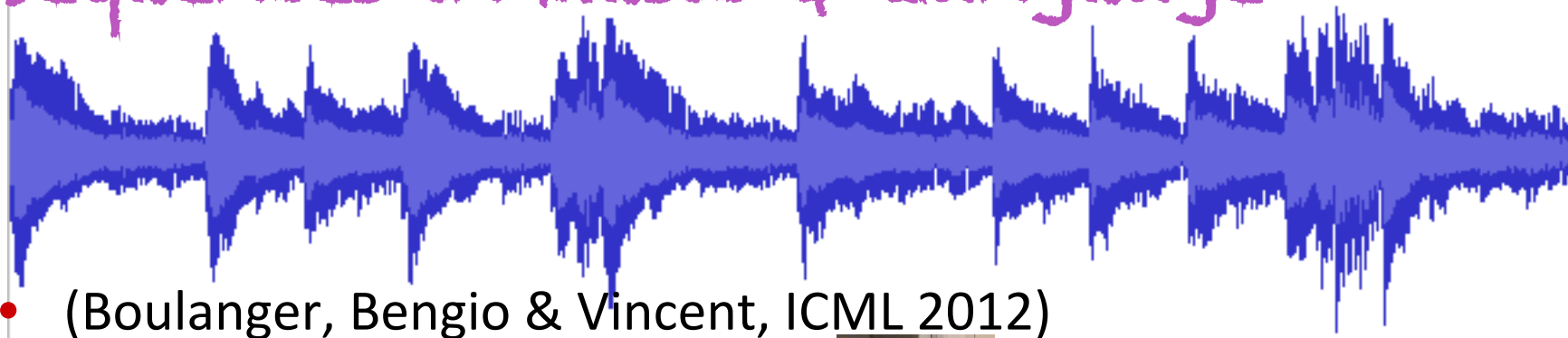
- Local connectivity across time/space
- Sharing weights across time/space (translation equivariance)
- Pooling (translation invariance, cross-channel pooling for learned invariances)



Recurrent nets (RNNs) can summarize information from the past

Bidirectional RNNs also summarize information from the future

# Deep / Recurrent Nets for Modeling Sequences in Music & Language



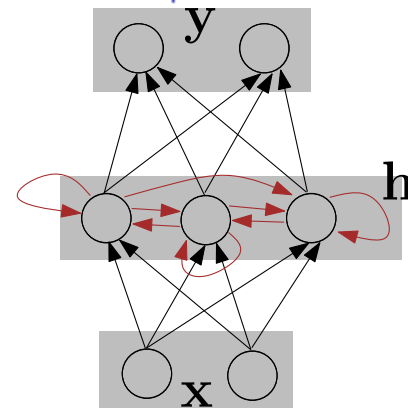
- (Boulanger, Bengio & Vincent, ICML 2012)

- Recurrent nets + RBMs

**SOTA** • Acoustics  $\rightarrow$  musical score

- (Bengio, Boulanger & Pascanu, ICASSP 2013)
  - Optimization techniques for recurrent nets
  - Symbolic sequences (music, language)

- (Pascanu, Mikolov & Bengio, ICML 2013)
  - Handling longer-term dependencies
  - Symbolic sequences (music, language)



# What differences with Neural Nets of the 90's?

- Other kinds of hierarchies are possible (e.g. A. Yuille, D. McAllester )
- Bigger models
- Better training
  - **Initialization**: information flow (Jacobians e-values closer to 1)
  - **Symmetry breaking**: initialization, sparsity regularization and non-linearities (rectifier, maxout, etc.)
- Unsupervised and multi-task learning → better transfer learning
- Larger labeled sets: **the advantage increases!**
- Better regularizers (dropout, injected noise, temporal coherence)

# Deep Learning Tricks of the Trade

- Y. Bengio (2013), “Practical Recommendations for Gradient-Based Training of Deep Architectures”

*(arXiv paper or chapter of Tricks of the Trade 2013 book)*

- Unsupervised pre-training
- Stochastic gradient descent and setting learning rates
- Hyper-parameters
  - Learning rate schedule
  - Early stopping
  - Minibatches
  - Parameter initialization
  - Number of hidden units
  - L1 and L2 weight decay
  - Sparsity regularization
- Debugging
- **How to efficiently search for hyper-parameter configurations**

# Deep Learning Challenges

(Bengio, arxiv 1305.0445 Deep Learning of representations: Looking forward)

- Computational Scaling
- Optimization & Underfitting
- Intractable Marginalization, Approximate Inference & Sampling
- Disentangling Factors of Variation
- Reasoning & One-Shot Learning of Facts

Wednesday

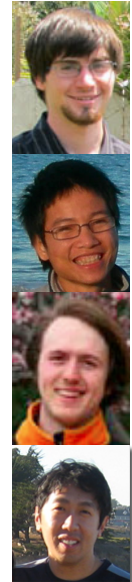
# Invariance and Disentangling

- 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?

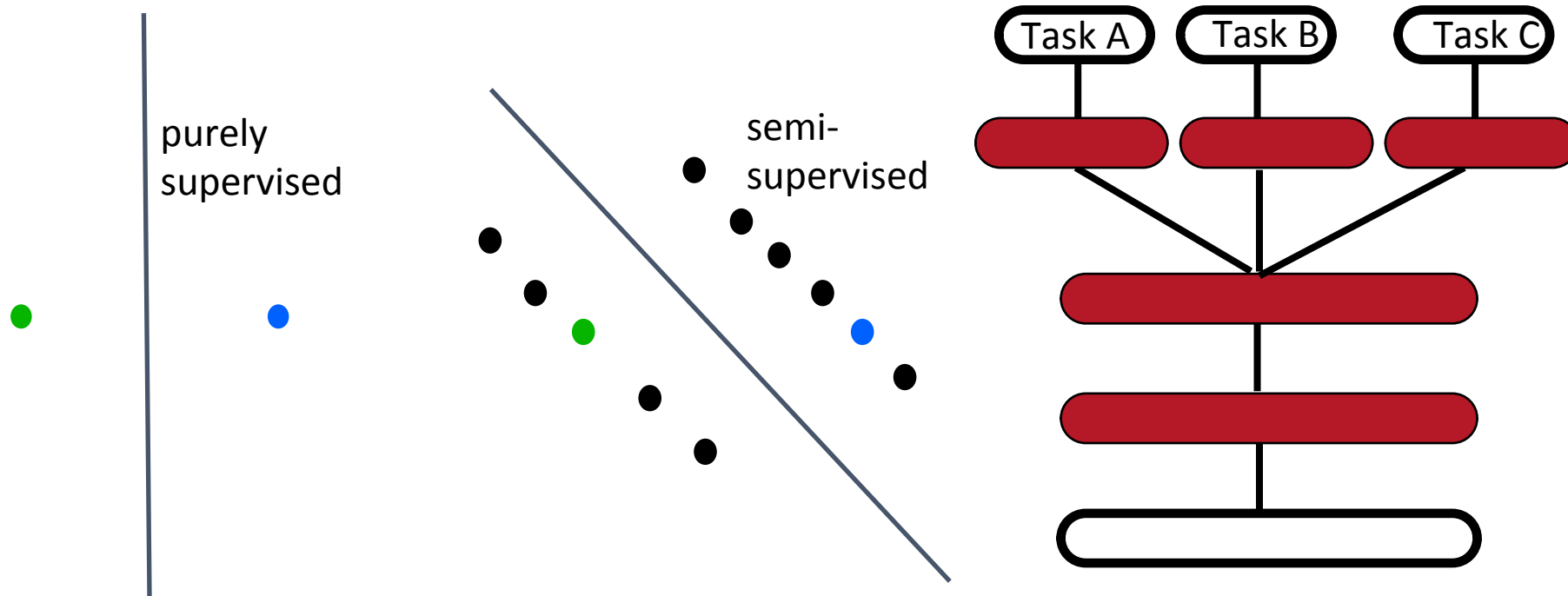
# Broad Priors as Hints to Disentangle the Factors of Variation

- 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)

tomorrow

# Sharing Factors

- Across tasks (multi-task learning), between inputs and target variables (semi-supervised learning).



# Sparse Representations

- Just add a sparsifying penalty on learned representation  
(prefer 0s in the representation)
- Information disentangling (compare to dense compression)
- More likely to be linearly separable (high-dimensional space)
- Locally low-dimensional representation = local chart
- Hi-dim. sparse = efficient variable size representation  
= data structure

Few bits of information



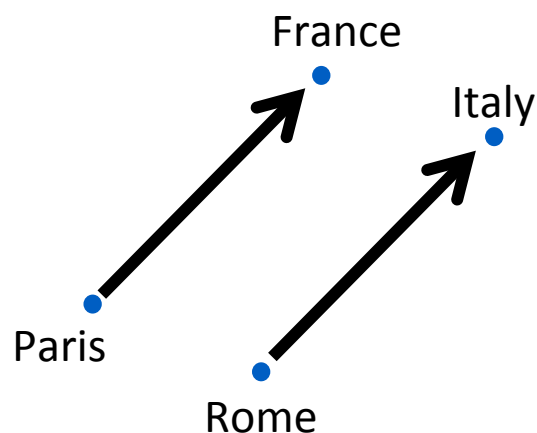
Many bits of information



**Prior: only few concepts and attributes relevant per example**

# Linear Relationships in Representation Space

- (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



Another earlier use of such a prior, for computer vision: Hinton's *capsules* (ICANN'2011)

# Conclusions

- Deep Learning has matured
  - Int. Conf. on Learning Representation 2013 a huge success!
- Industrial applications (Google, Microsoft, Baidu, Facebook, ...)
- Room for future breakthroughs:
  - Scaling computation
  - Optimization
  - Intractable marginalizations
  - more disentangled abstractions
  - Reason from incrementally added facts

LISA team: **Merci! Questions?**



LISA team: **Merci! Questions?**

