Scaling up Deep Learning towards AI

Yoshua Bengio

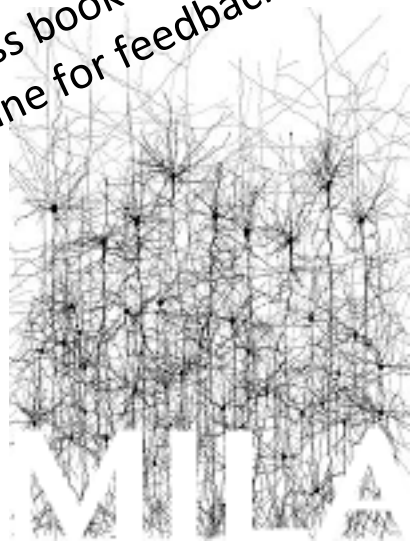
October 13, 2015

IBM Cognitive Colloquium, San Francisco

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

Université de Montréal

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



Breakthrough

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

Impact

Deep learning has revolutionized

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

More on the way, including other areas of computer vision, NLP, dialogue, reinforcement learning, robotics, control…

Challenges to Scale towards AI

- Computational challenge
- Reasoning, natural language understanding and knowledge representation
- Large-scale unsupervised learning

Computational Scaling

- Recent breakthroughs in speech, object recognition and NLP hinged on faster computing, GPUs, and large datasets

- In speech, vision and NLP applications we tend to find that

*as **Ilya Sutskever** would say*

BIGGER IS BETTER

Because deep learning is

EASY TO REGULARIZE while

it is MORE DIFFICULT TO AVOID UNDERFITTING

Scaling up computation: we still have a long way to go in raw computational power

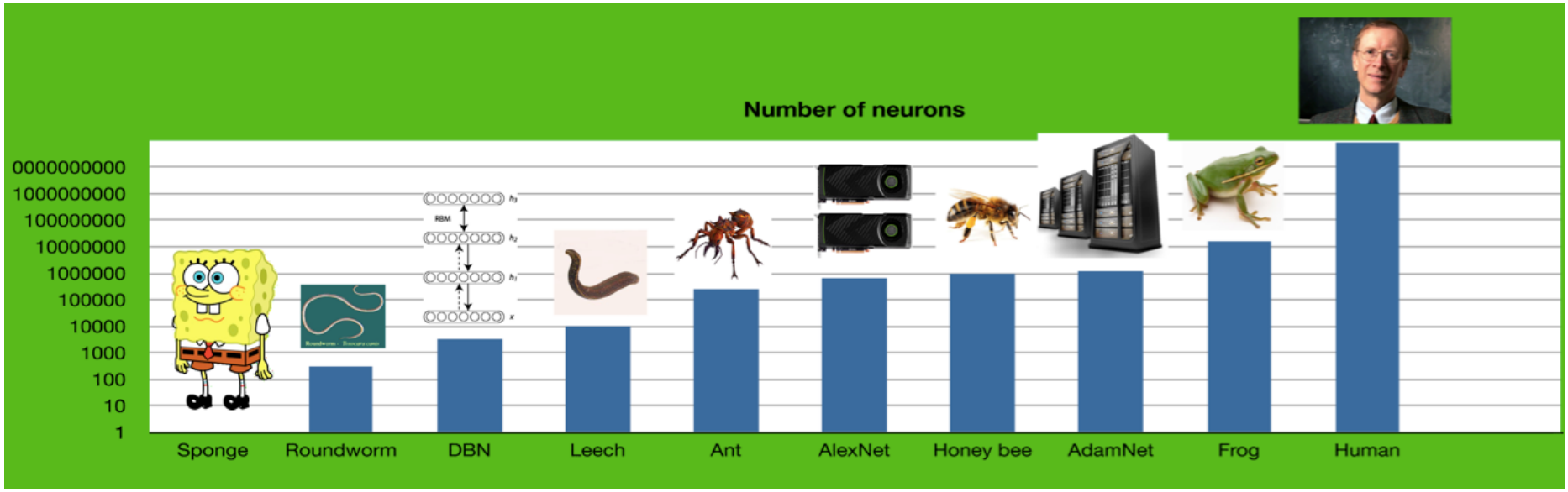
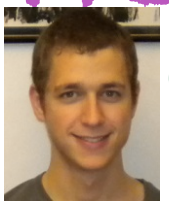


Figure: Ian Goodfellow

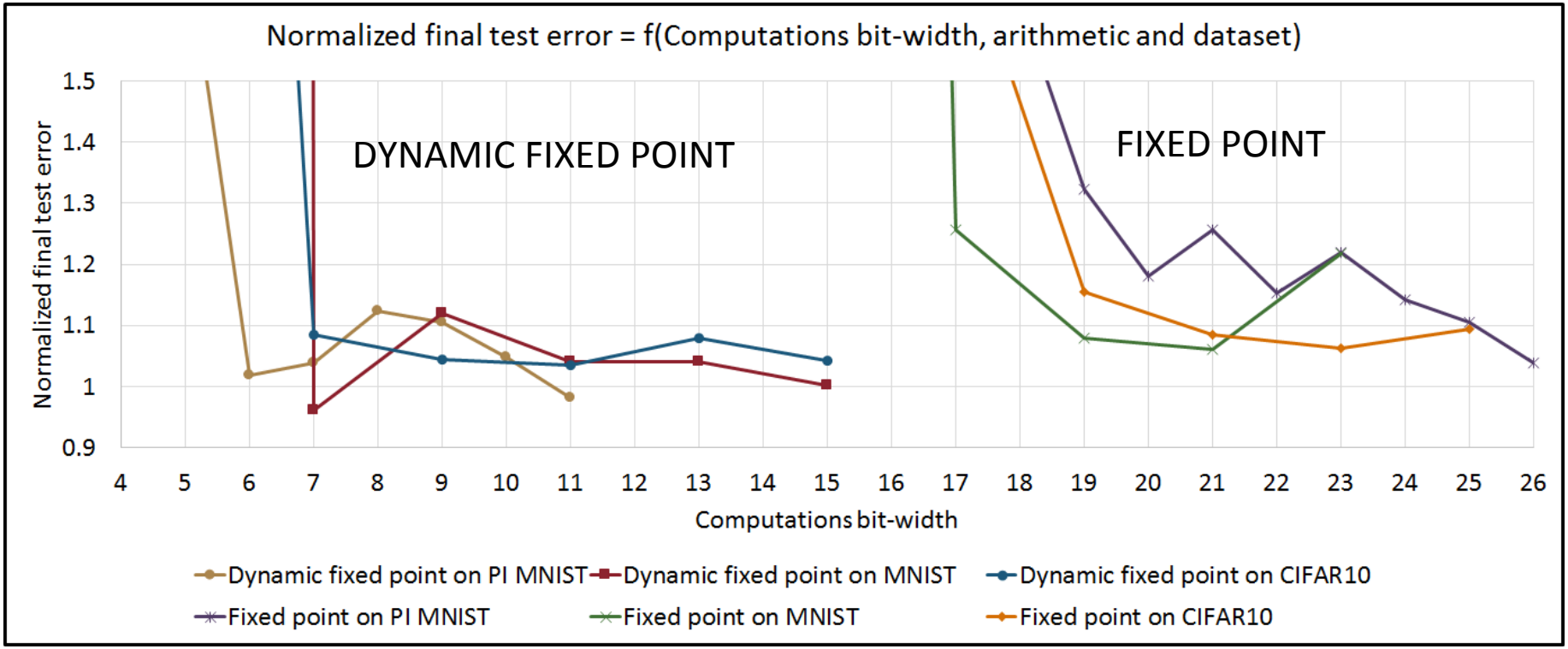
Low-Precision Training of Deep Nets



Courbariaux, Bengio & David, ICLR 2015 workshop

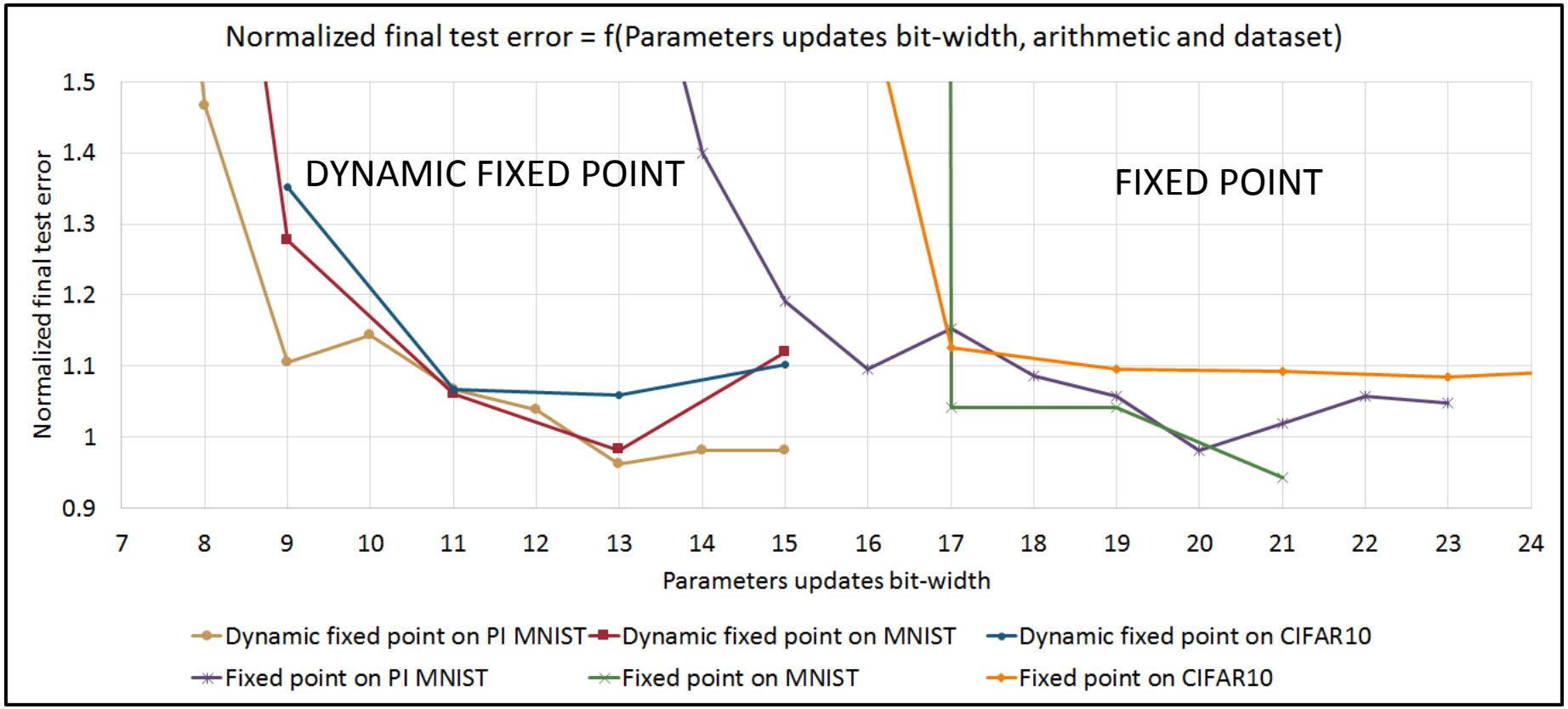
- See *(Gupta et al, arXiv Feb. 2015)* for a recent review
- Previous work showed that it was possible to quantize weights **after training** to obtain low-precision implementations of trained deep nets (8 bits or even less if you retrain and keep high precision at top layer)
- This work is about **training** with low precision
- How many bits are required? 12

Number of bits for computations



10 bits were selected, with dynamic fixed point

Number of bits for updating and storing weights



12 bits were selected

NIPS'2015: Single-Bit Weights

BitConnect, Courbariaux, David & Bengio, NIPS'2015

- Using stochastic rounding and 16-bit precision operations, we are able to train deep nets with weights quantized to 1 bit, i.e., we can get rid of most (2/3) multiplications

- This could have a drastic impact on hardware implementations, especially for low-power devices…

Results on MNIST

Method	Validation error rate (%)	Test error rate (%)
No regularizer	1.21 ± 0.04	1.30 ± 0.04
BinaryConnect (det.)	1.17 ± 0.02	1.29 ± 0.08
BinaryConnect (stoch.)	1.12 ± 0.03	1.18 ± 0.04
50% Dropout	0.94 ± 0.04	1.01 ± 0.04
Maxout networks [26]		0.94
Deep L2-SVM [27]		0.87

Getting Rid of the Remaining Multiplications

- The main remaining multiplications (1/3) are due to the weight update of the form

$$\Delta W_{ij} \propto \frac{\partial C}{\partial a_i} h_j$$

- It can be eliminated by quantizing h_j to powers of 2 (Simard & Graf 1994): the multiplication becomes a shift. Similarly for the learning rate.
- The quantization can also be stochastic, to preserve the expected value of the update:

$$E[\tilde{h}_j] = h_j$$

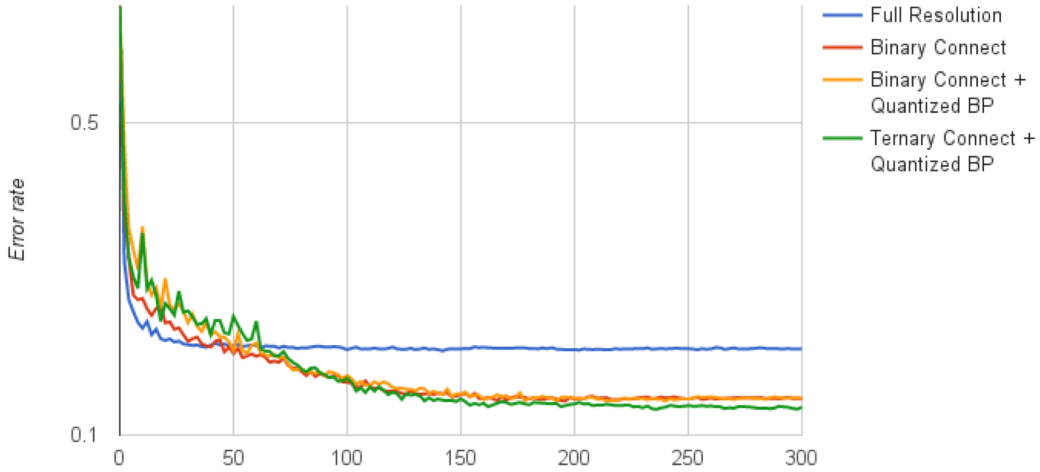
Neural Networks with Few Multiplications

- ArXiv paper, 2015: Lin, Courbariaux, Memisevic & Bengio

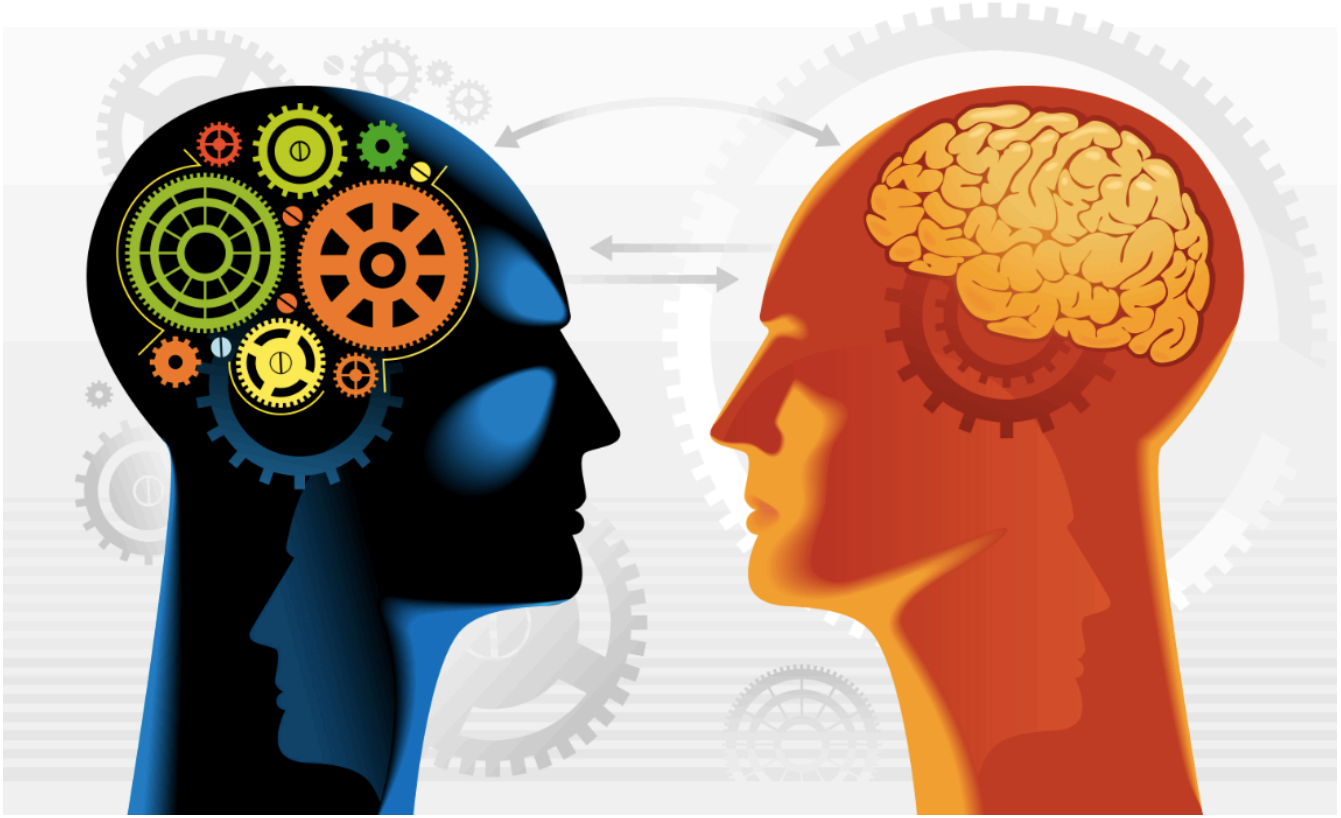
	Full precision	Binary connect	Binary connect + Quantized backprop	Ternary connect + Quantized backprop
MNIST	1.33%	1.23%	1.29%	1.15%
CIFAR10	15.64%	12.04%	12.08%	12.01%
SVHN	2.85%	2.47%	2.48%	2.42%

Works!

Slows down training a bit but improves results by regularizing

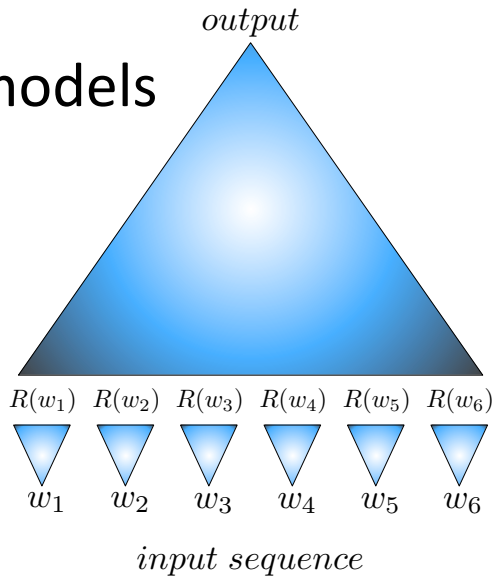


The Language Understanding Challenge

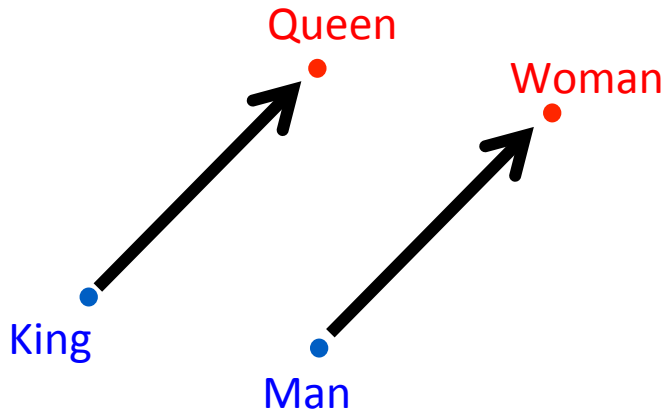


Learning Word Semantics: a Success

- Bengio et al 2000 introduced neural language models and word vectors (word embeddings)

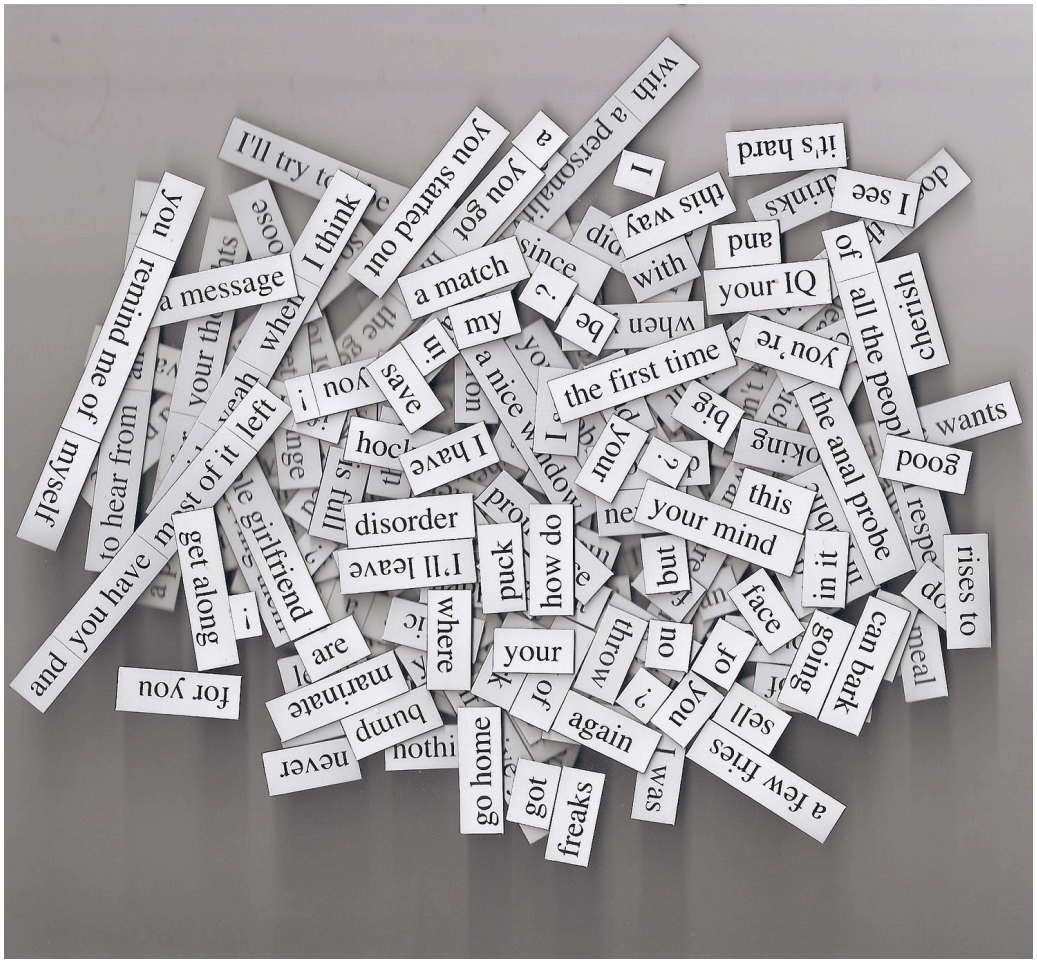


Mikolov showed these embeddings capture analogical relations:



The Next Challenge: Rich Semantic Representations for Word Sequences

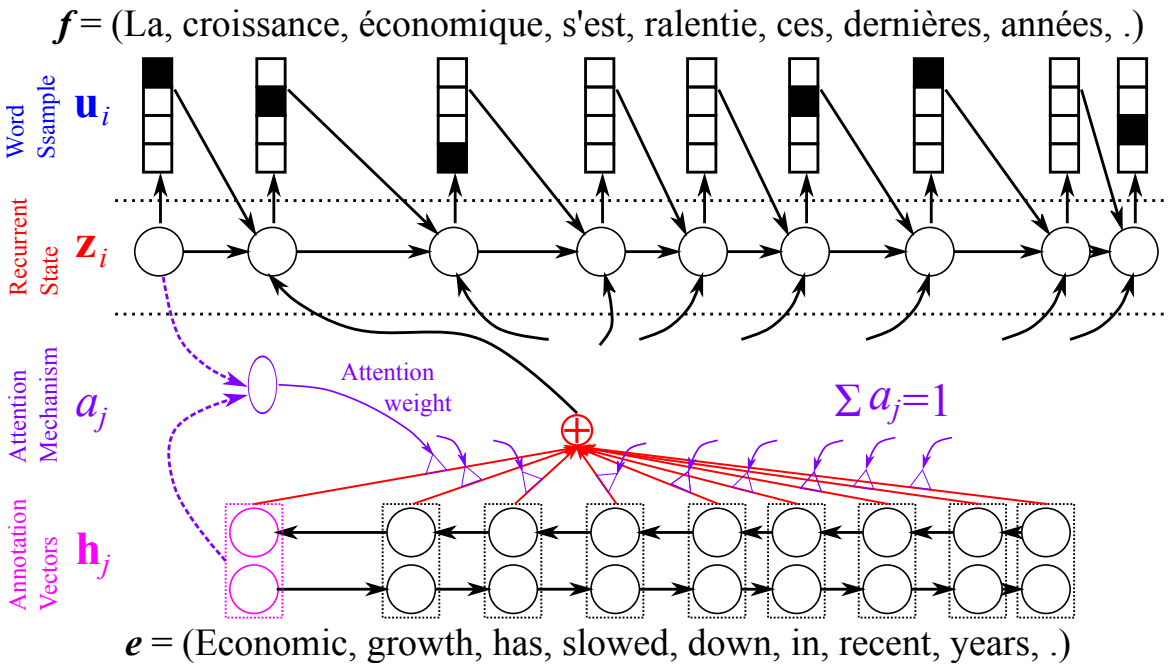
- First challenge: machine translation
- Second challenge: document understanding and question answering



Attention-Based Neural Machine Translation

Related to earlier Graves 2013 for generating handwriting

- (Bahdanau, Cho & Bengio, arXiv sept. 2014)
- (Jean, Cho, Memisevic & Bengio, arXiv dec. 2014)



End-to-End Machine Translation with Recurrent Nets and Attention Mechanism

- Reached the state-of-the-art in one year, from scratch

(a) **English→French (WMT-14)**

	NMT(A)	Google	P-SMT
NMT	32.68	30.6*	**37.03**$^\bullet$
+Cand	33.28	–	
+UNK	33.99	32.7°	
+Ens	**36.71**	**36.9**$^\circ$	

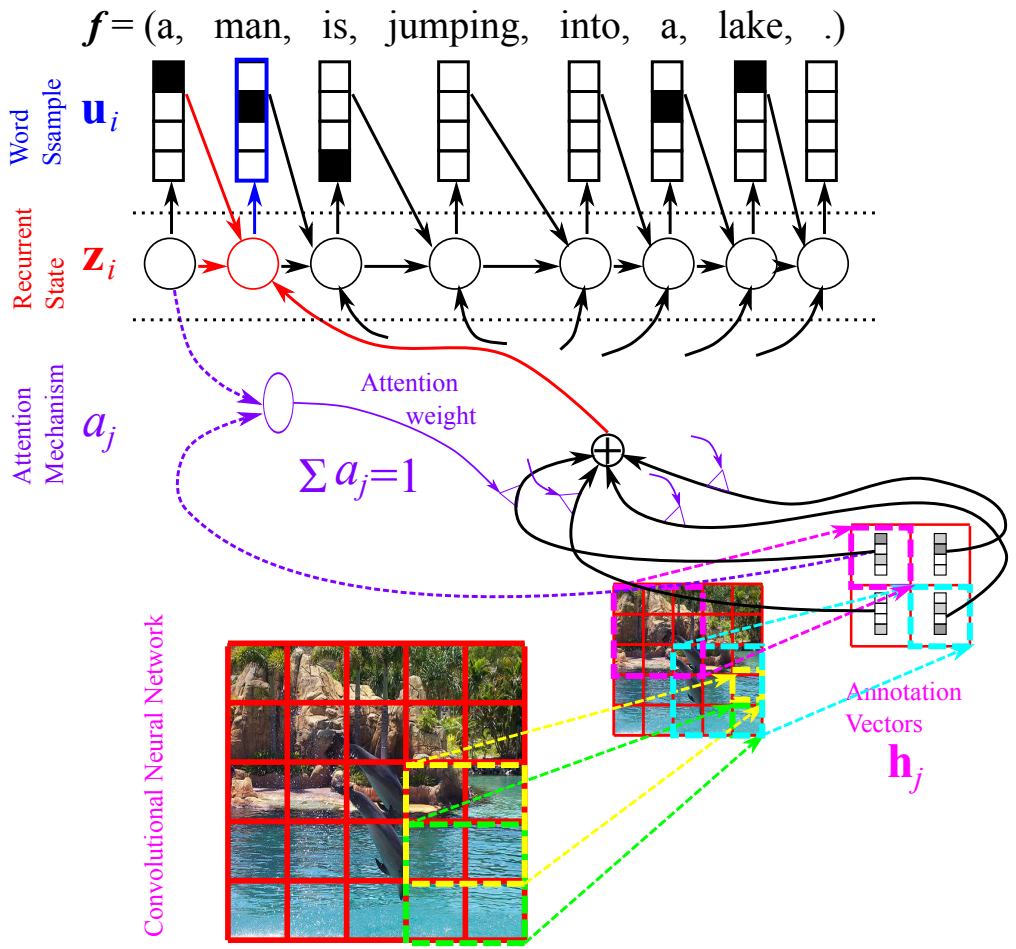
(b) **English→German (WMT-15)**

Model	Note
24.8	Neural MT
24.0	U.Edinburgh, Syntactic SMT
23.6	LIMSI/KIT
22.8	U.Edinburgh, Phrase SMT
22.7	KIT, Phrase SMT

(c) **English→Czech (WMT-15)**

Model	Note
18.3	Neural MT
18.2	JHU, SMT+LM+OSM+Sparse
17.6	CU, Phrase SMT
17.4	U.Edinburgh, Phrase SMT
16.1	U.Edinburgh, Syntactic SMT

Image-to-Text: Caption Generation



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

The Good

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.

And the Bad

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.

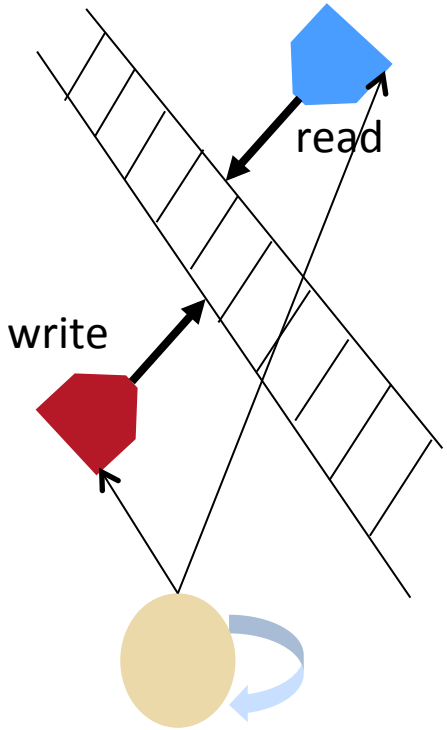
A person is standing on a beach
with a surfboard.

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

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

Attention Mechanisms for Memory Access

- 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



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

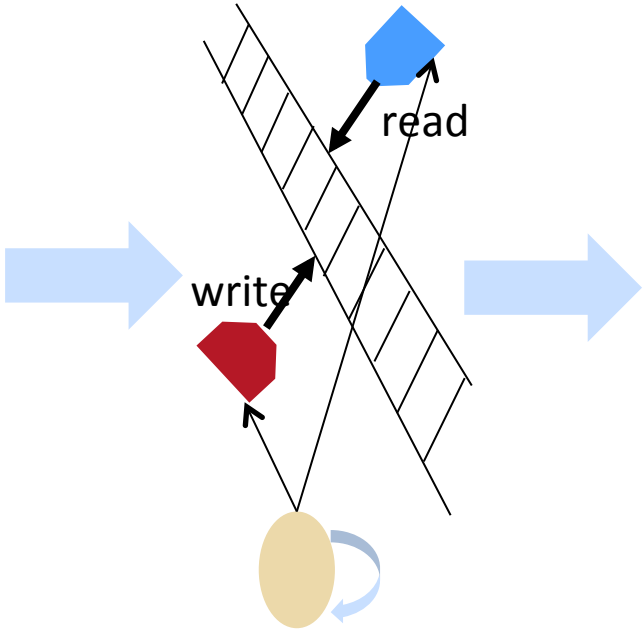
From "Memory Networks", Weston et al. ICLR 2015;
"End-to-end memory networks", Sukhbatar et al. NIPS'2015

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

Ongoing Project: Knowledge Extraction

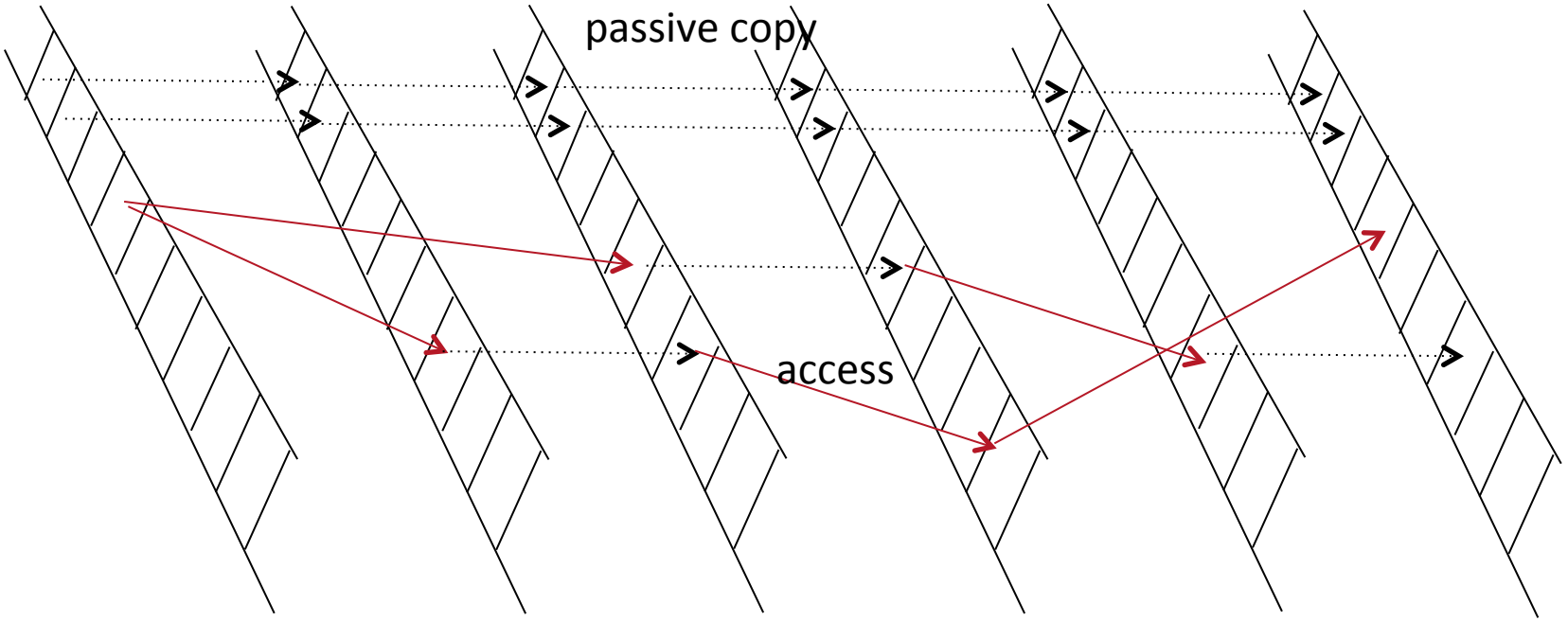
- Learn to fill the memory network from natural language descriptions of facts
- Force the neural net to understand language
- Extract knowledge from documents into a usable form

WIKIPEDIA
The Free Encyclopedia

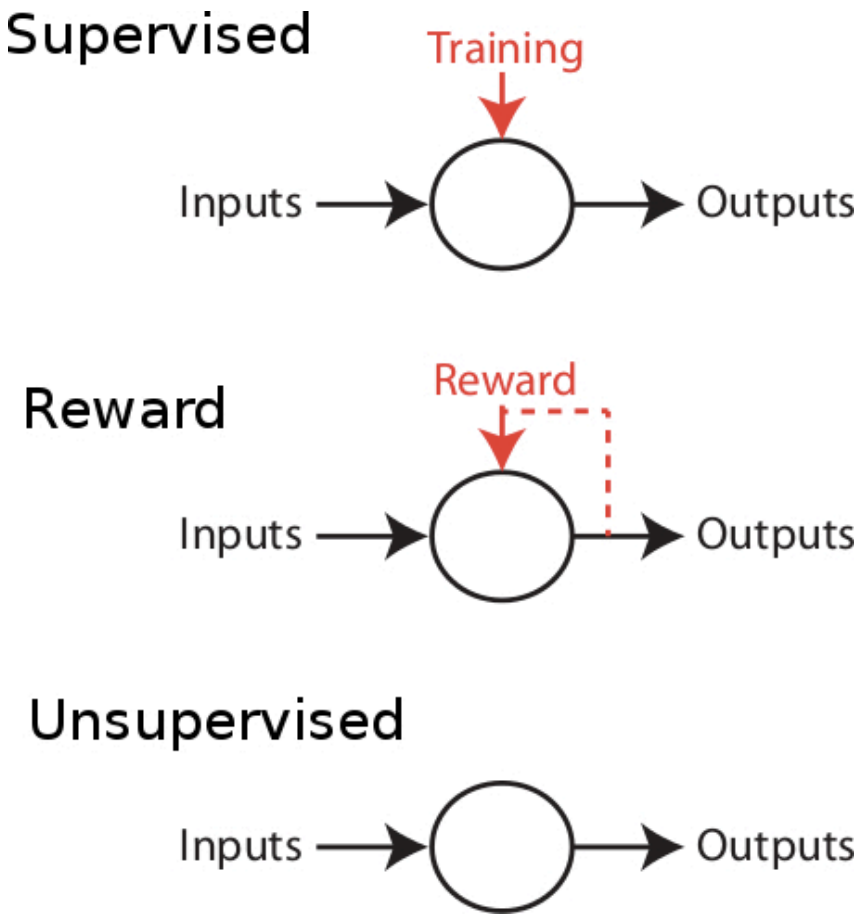


Why does it work? Pushing off the Curse of Long-Term Dependencies

- 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.



The Unsupervised Learning Challenge



Why Unsupervised Learning?

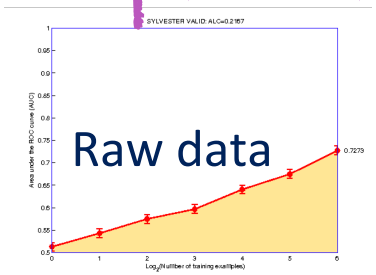
- Recent progress mostly in supervised DL
- Real technical challenges for unsupervised DL
- Potential benefits:
  - **Exploit tons of unlabeled data**
  - Answer new questions about the variables observed
  - Regularizer – transfer learning – domain adaptation
  - Easier optimization (local training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

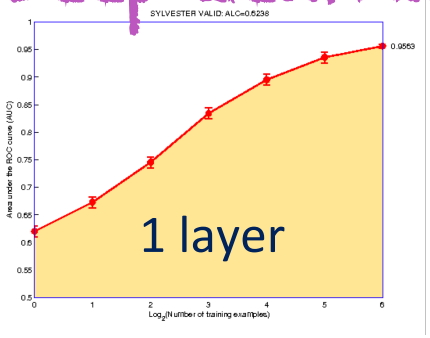
How do humans generalize from very few examples?

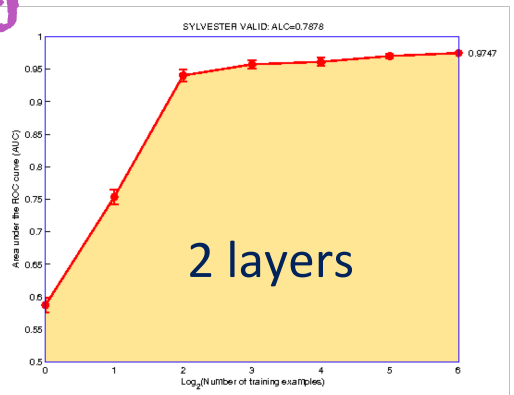
- Intelligence (good generalization) needs knowledge
- Humans **transfer** knowledge from **previous learning**:
  - Representations
  - Explanatory factors
- Previous learning from: **unlabeled data**

  + labels for other tasks

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

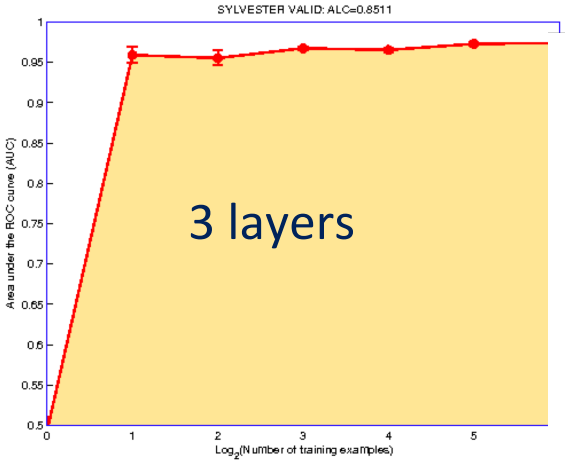


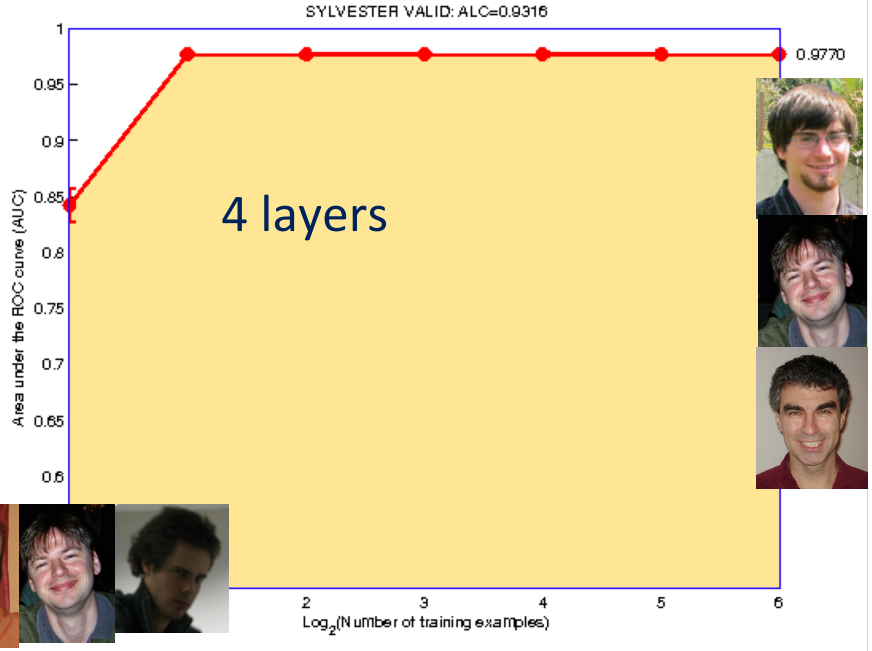




NIPS'2011 Transfer Learning Challenge Paper: ICML'2012

ICML'2011 workshop on Unsup. & Transfer Learning





Intractable (Exponential) Barriers

- Statistical curse of dimensionality:
  - Intractable number of configurations of variables, in high dimension
- Computational curse of dimensionality:
  - Intractable normalization constants
  - Intractable (non-convex) optimization?
  - Intractable inference

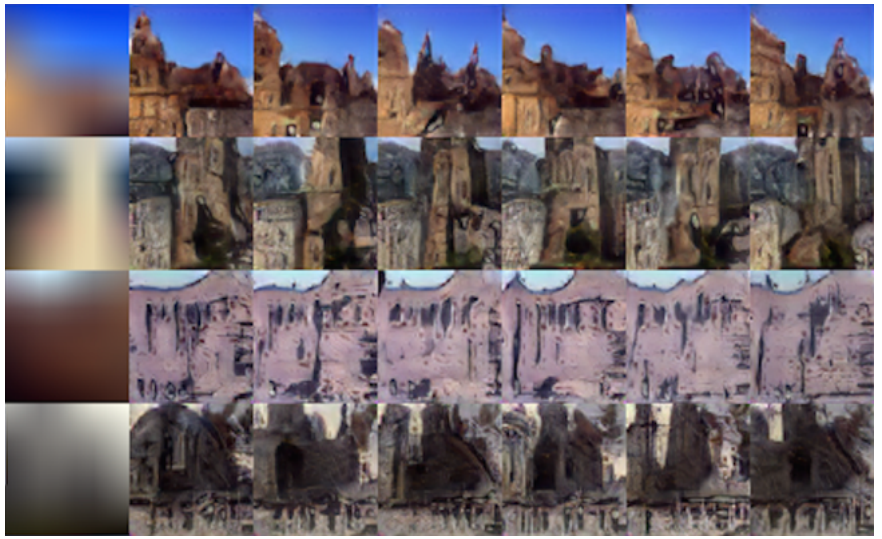
Deep Generative Learning: the Hot Frontier

- Many very different approaches being explored to bypass these intractabilities
- Exploratory mode
- Exciting area of research
- Connect to brains: bridge the gap to biology

And the gap between Boltzmann machines and Backprop (Y. Bengio)

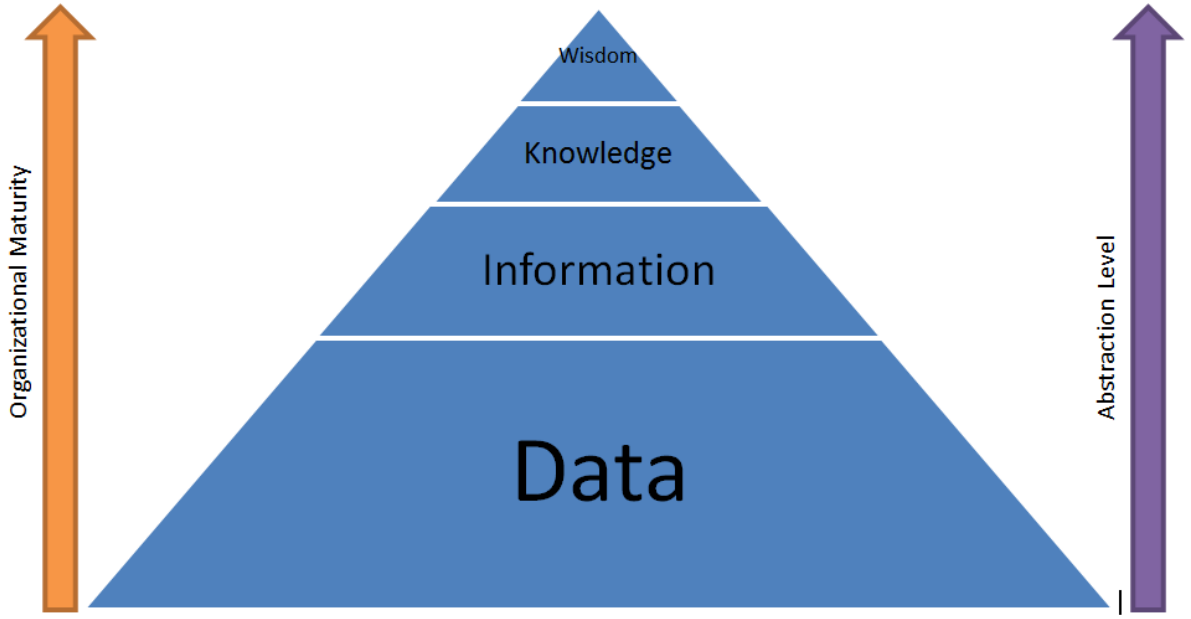
DRAW (DeepMind)

LAPGAN (NYU/Facebook)



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



MILA: Montreal Institute for Learning Algorithms