MILA

IVADO
INSTITUTE FOR DATA VALORISATION

DEEP LEARNING

Yoshua Bengio

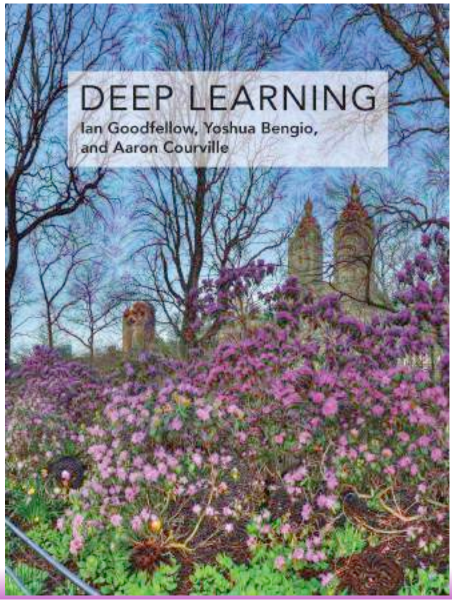
13 December 2016

ICDM’2016, Barcelona

CIFAR
CANADIAN
INSTITUTE
FOR
ADVANCED
RESEARCH

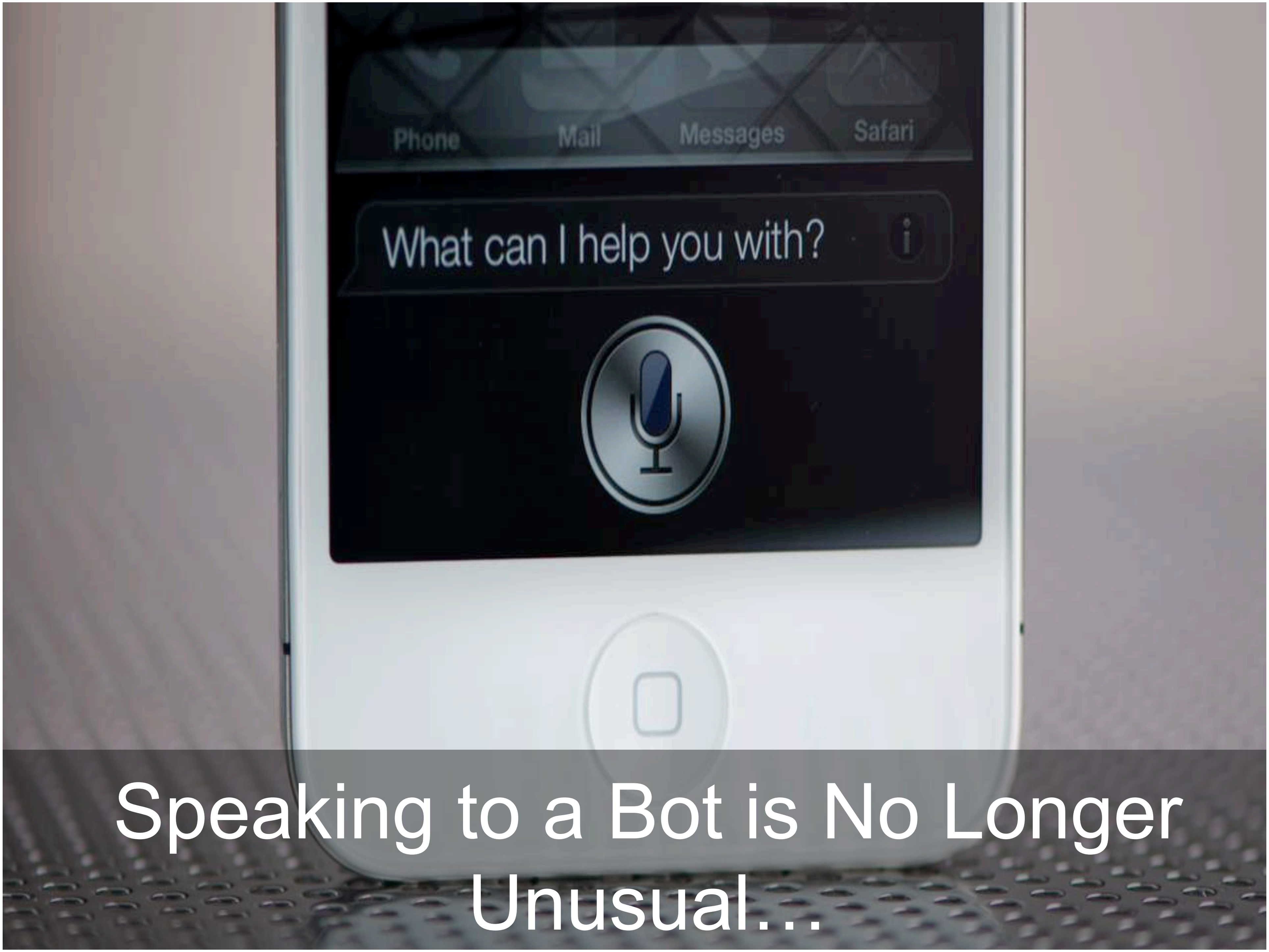
Université de Montréal

PLUG: **Deep Learning**, MIT Press book on sale, chapters online for feedback



Cars are now driving themselves…

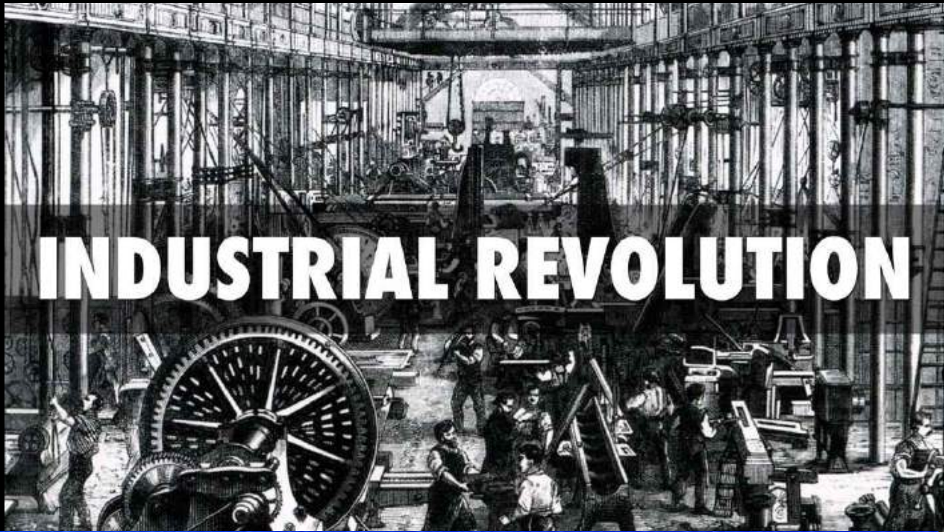
(far from perfectly, though)



Speaking to a Bot is No Longer Unusual…

March 2016:
World Go Champion Beaten by Machine

AI: The Upcoming Industrial Revolution



First industrial revolution:

- Machines extending humans' **mechanical power**

Upcoming industrial revolution:

- Machines extending humans' **cognitive power**
  - From the digital economy to the AI economy
  - Predicted growth at least 25%/yr
  - All sectors of the economy

A new revolution seems to be in the work after the industrial revolution.

Devices are becoming intelligent.

And Deep Learning is at the epicenter of this revolution.

Breakthrough in deep learning

A Canadian-led trio at CIFAR initiated the deep learning AI revolution

- Fundamental breakthrough in 2006:

first successful recipe for training a deep supervised neural network

- Second major advance in 2011, with rectifiers
- Breakthroughs in applications since then

YOSHUA BENGIO
Montreal

CIFAR

GEOFF HINTON
Toronto

Google

YANN LECUN
New York

Facebook

AI Needs Knowledge

- Failure of classical AI: a lot of knowledge is not formalized, expressed with words
- Solution: computer gets knowledge from data, learns from examples

MACHINE LEARNING

Machine Learning, AI & No Free Lunch

- Five key ingredients for ML towards AI
  1. Lots & lots of data
  2. Very flexible models
  3. Enough computing power
  4. Computationally efficient inference
  5. **Powerful priors that can defeat the curse of dimensionality**

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

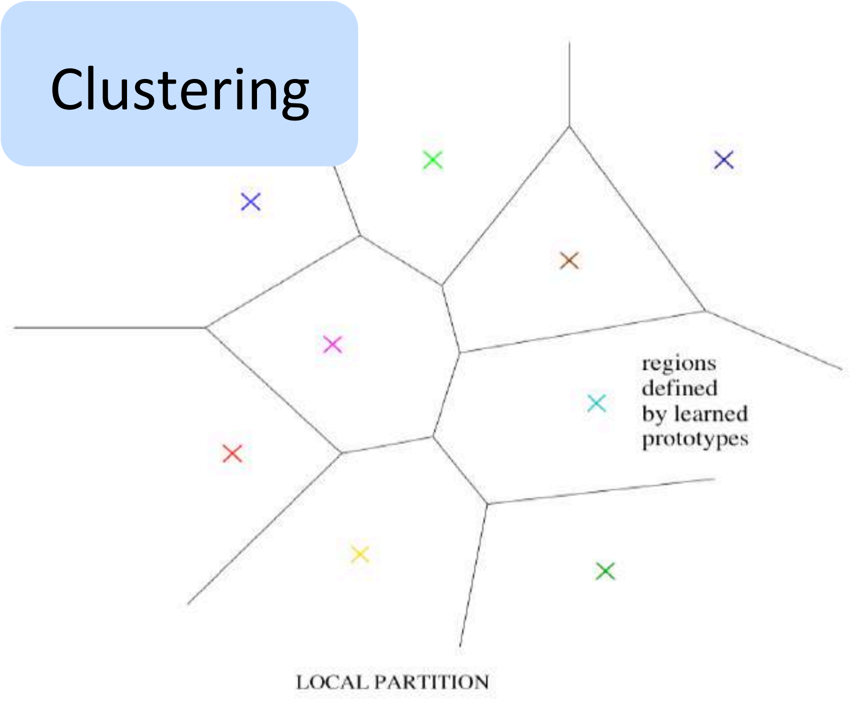
Exploiting compositionality gives an exponential gain in representational power

Distributed representations / embeddings: feature learning

Deep architecture: multiple levels of feature learning

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

Non-distributed representations



- 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**

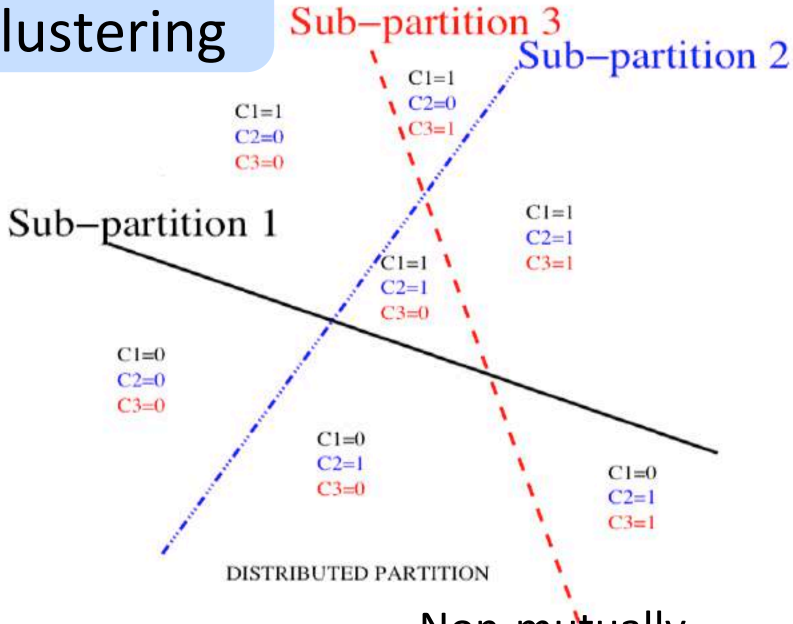
→ No non-trivial generalization to regions without examples

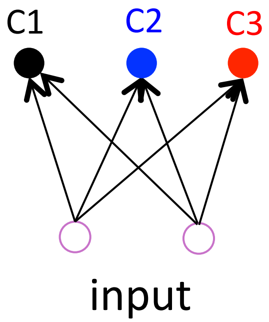
The need for distributed representations

- 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

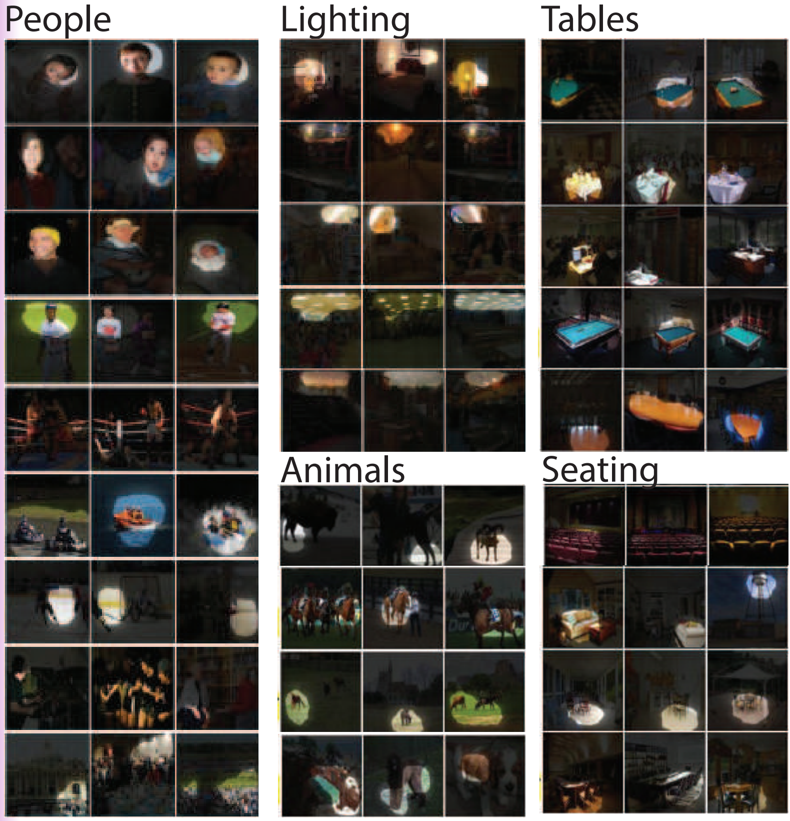


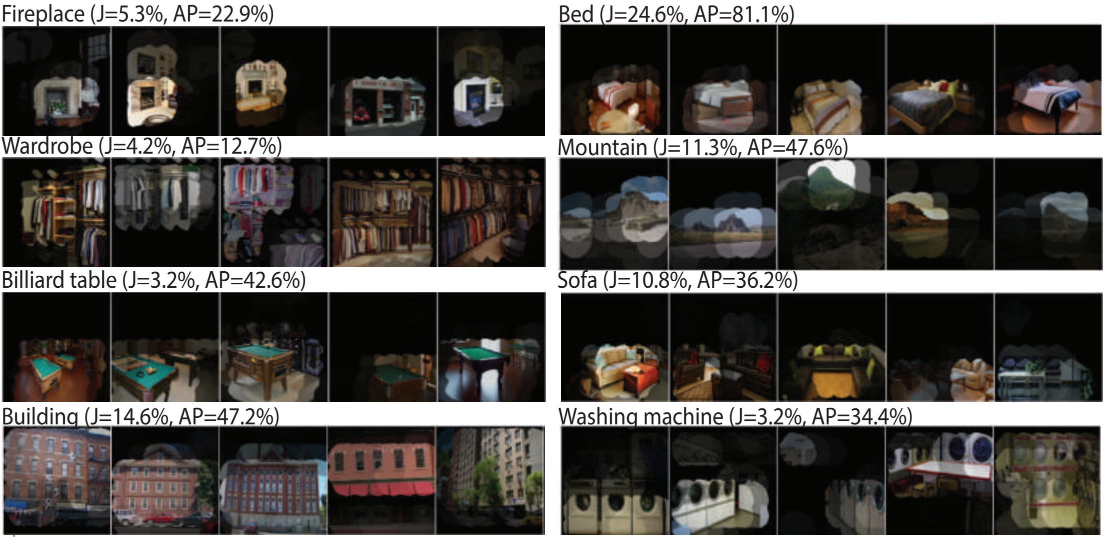


Non-mutually exclusive features/attributes create a combinatorially large set of distinguiable configurations

Hidden Units Discover Semantically Meaningful Concepts

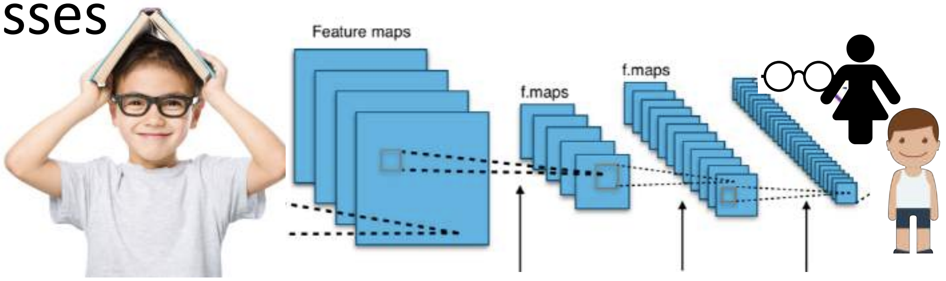
- *Zhou et al & Torralba, arXiv1412.6856* , ICLR 2015
- *Network trained to recognize places, not objects*




Fireplace (J=5.3%, AP=22.9%)
Bed (J=24.6%, AP=81.1%)
Wardrobe (J=4.2%, AP=12.7%)
Mountain (J=11.3%, AP=47.6%)
Billiard table (J=3.2%, AP=42.6%)
Sofa (J=10.8%, AP=36.2%)
Building (J=14.6%, AP=47.2%)
Washing machine (J=3.2%, AP=34.4%)

Each feature can be discovered without the need for seeing the exponentially large number of configurations of the other features

- Consider a network whose hidden units discover the following features:
  - Person wears glasses
  - Person is female
  - Person is a child
  - Etc.



If each of *n* feature requires *O(k)* parameters, need *O(nk)* examples

Non-parametric methods would require $O(n^d)$ examples

The Depth Prior can be Exponentially Advantageous

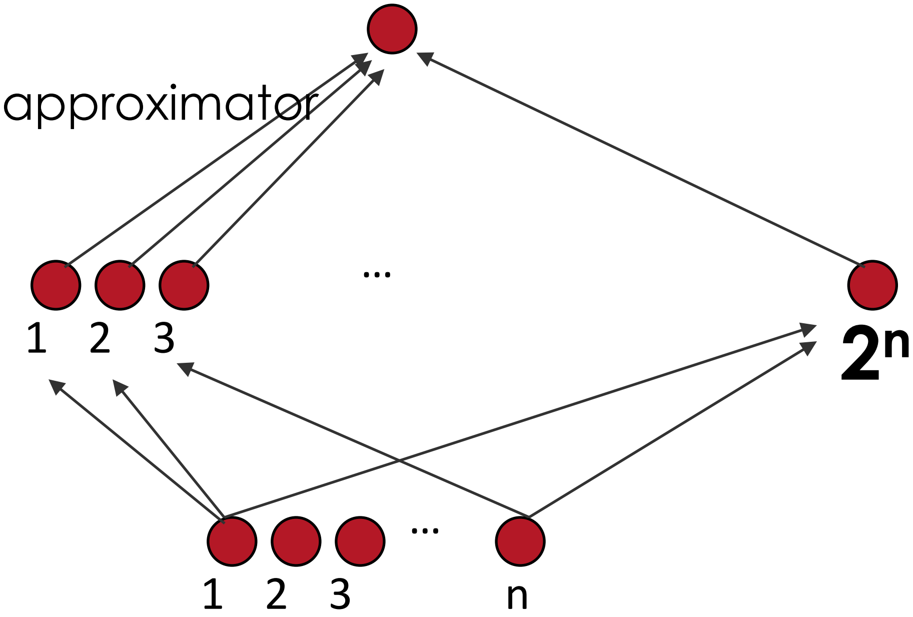
Theoretical arguments:

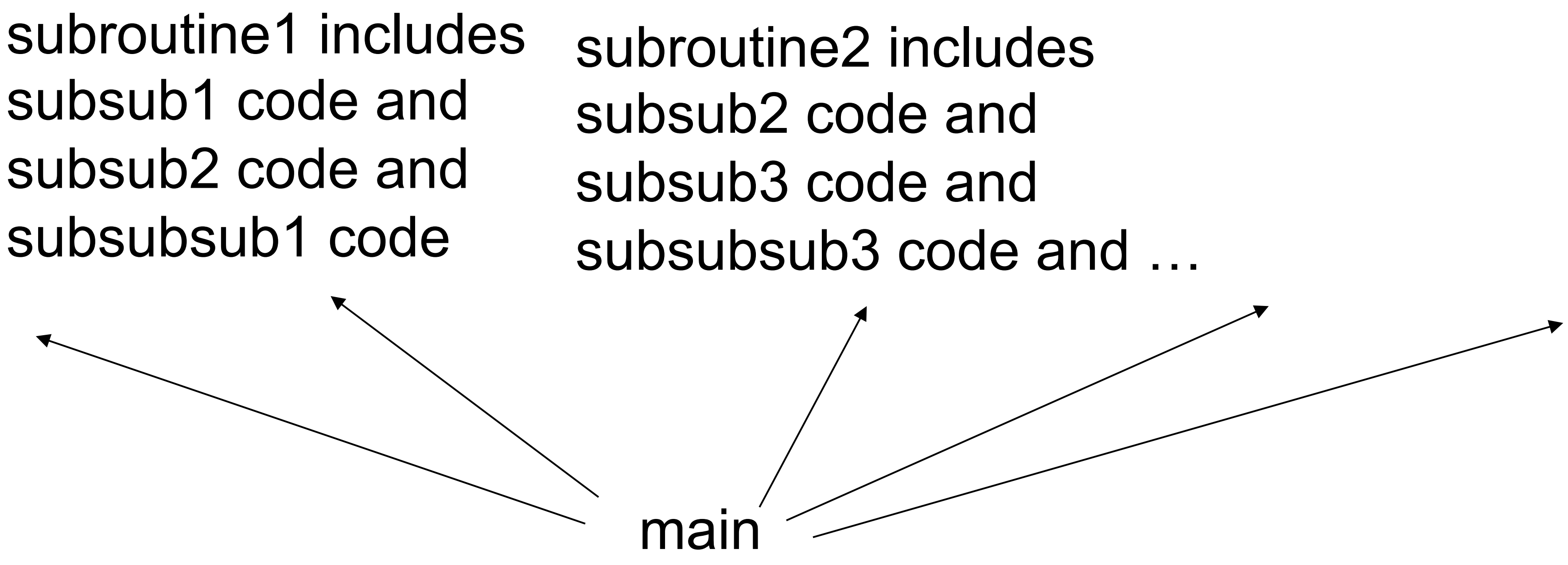
2 layers of { Logic gates / Formal neurons / RBF units } = universal approximator

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



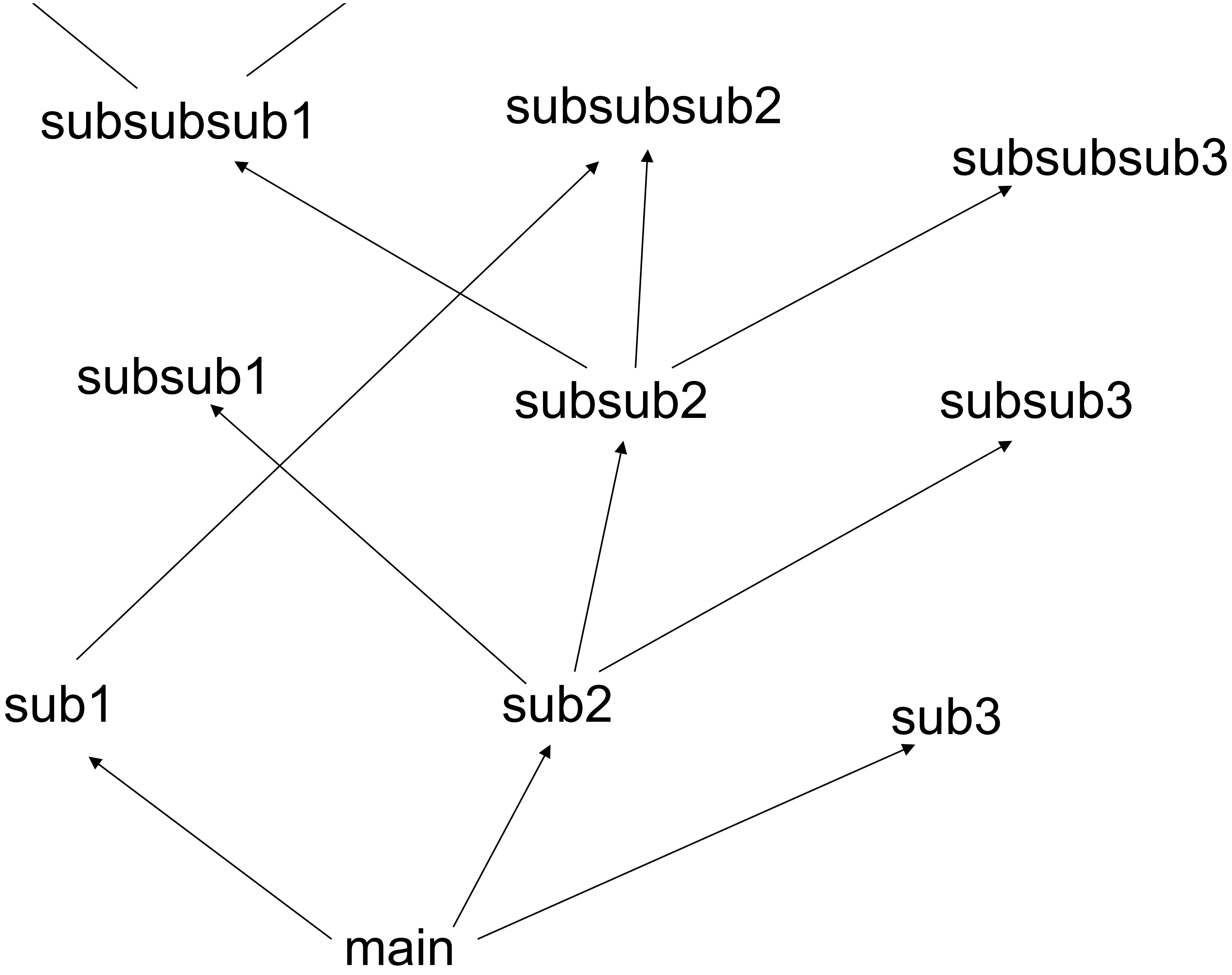


subroutine1 includes
subsub1 code and
subsub2 code and
subsubsub1 code

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

main

“Shallow” computer program



“Deep” computer program

Exponential advantage of depth

- Expressiveness of deep networks with piecewise linear activation functions: exponential advantage for depth *(Montufar et al, NIPS 2014)*
- Number of pieces distinguished for a network with depth *L* and n_i units per layer is at least

$$\left(\prod_{i=1}^{L-1} \left\lfloor \frac{n_i}{n_0} \right\rfloor^{n_0}\right) \sum_{j=0}^{n_0} \binom{n_L}{j}$$

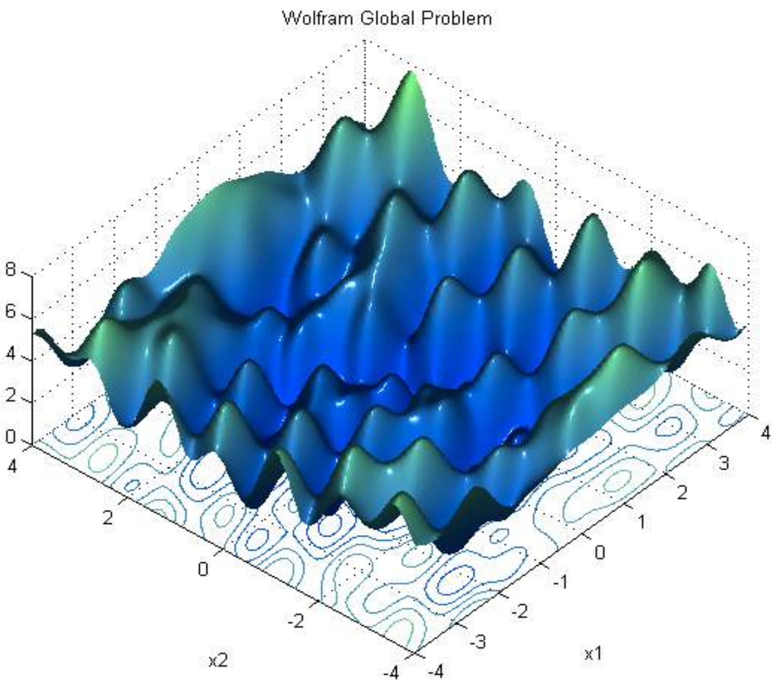
or, if hidden layers have width *n* and input has size n_0

$$\Omega\left(\left({}^{n}\!/_{n_0} \right)^{(L-1)n_0} n^{n_0} \right)$$

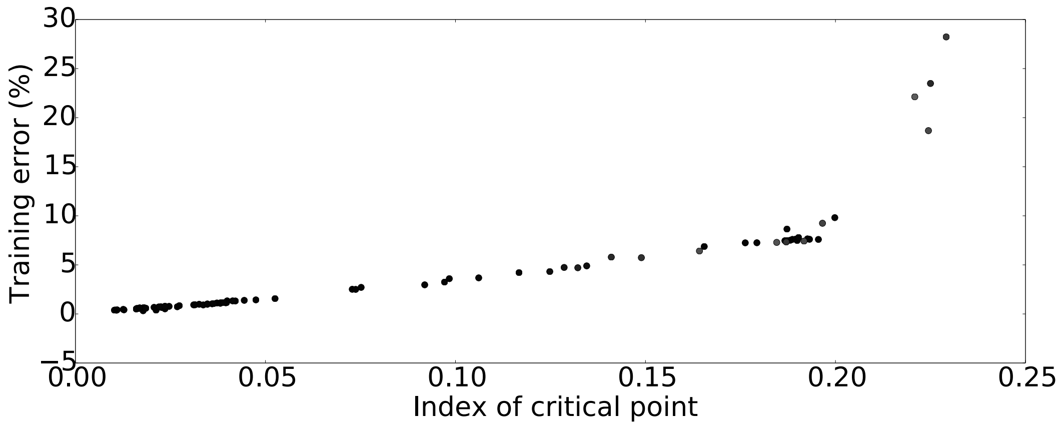
A Myth is Being Debunked: Local Minima in Neural Nets → Convexity is not needed

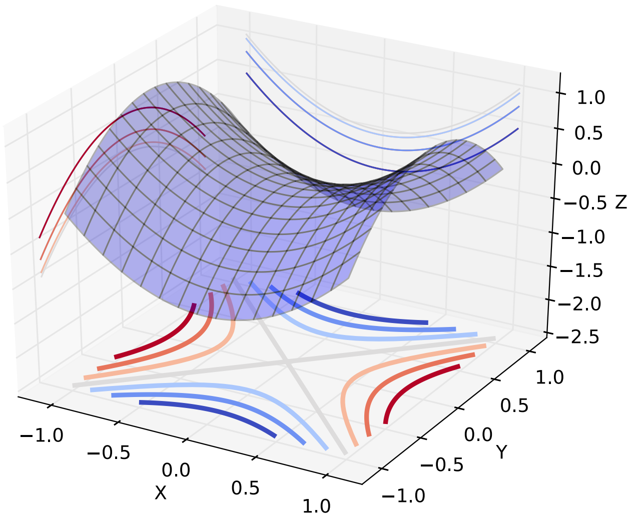
- (Pascanu, Dauphin, Ganguli, Bengio, arXiv May 2014): *On the saddle point problem for non-convex optimization*
- (Dauphin, Pascanu, Gulcehre, Cho, Ganguli, Bengio, NIPS' 2014): *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*
- (Choromanska, Henaff, Mathieu, Ben Arous & LeCun AISTATS 2015): *The Loss Surface of Multilayer Nets*

Saddle Points

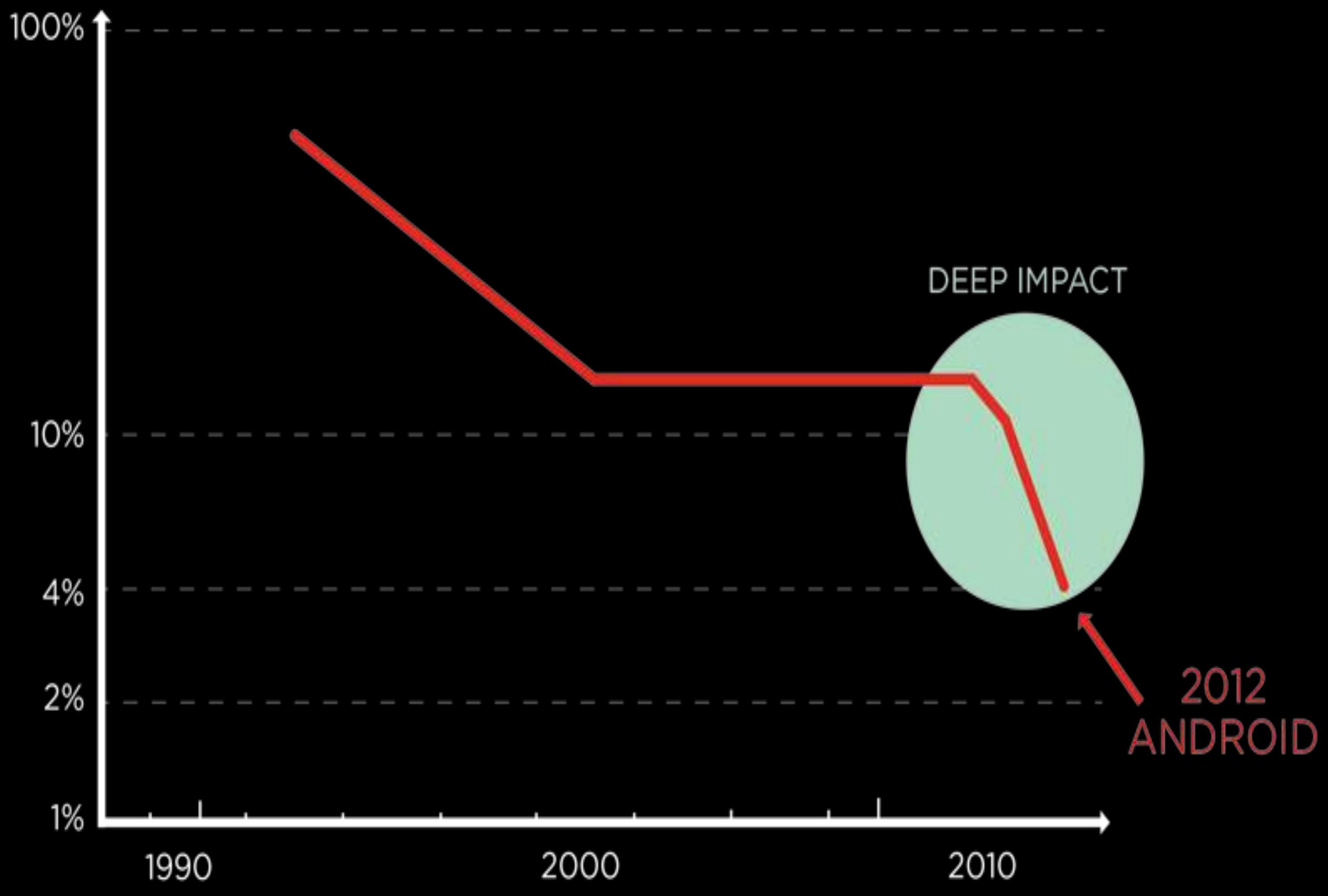


- Local minima dominate in low-D, but saddle points dominate in high-D
- Most local minima are close to the bottom (global minimum error)





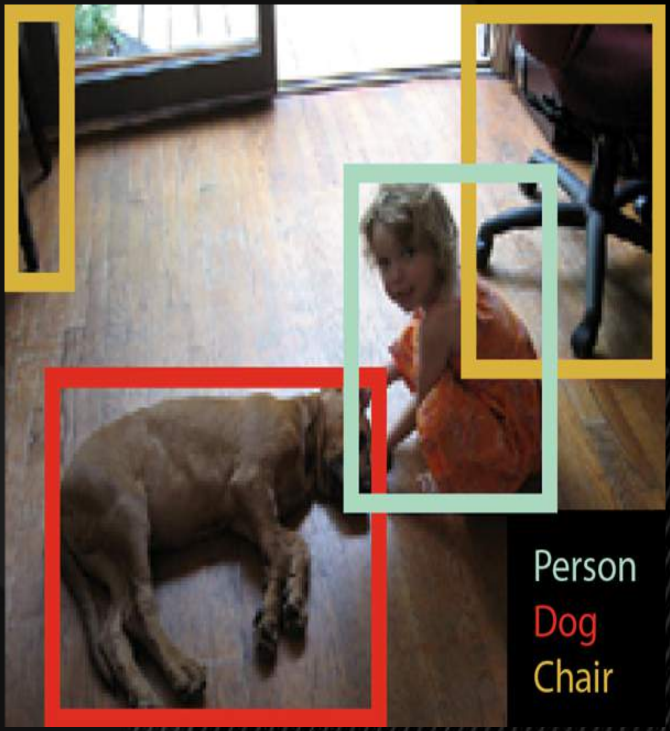
2010-2012: breakthrough in speech recognition



Source: Microsoft

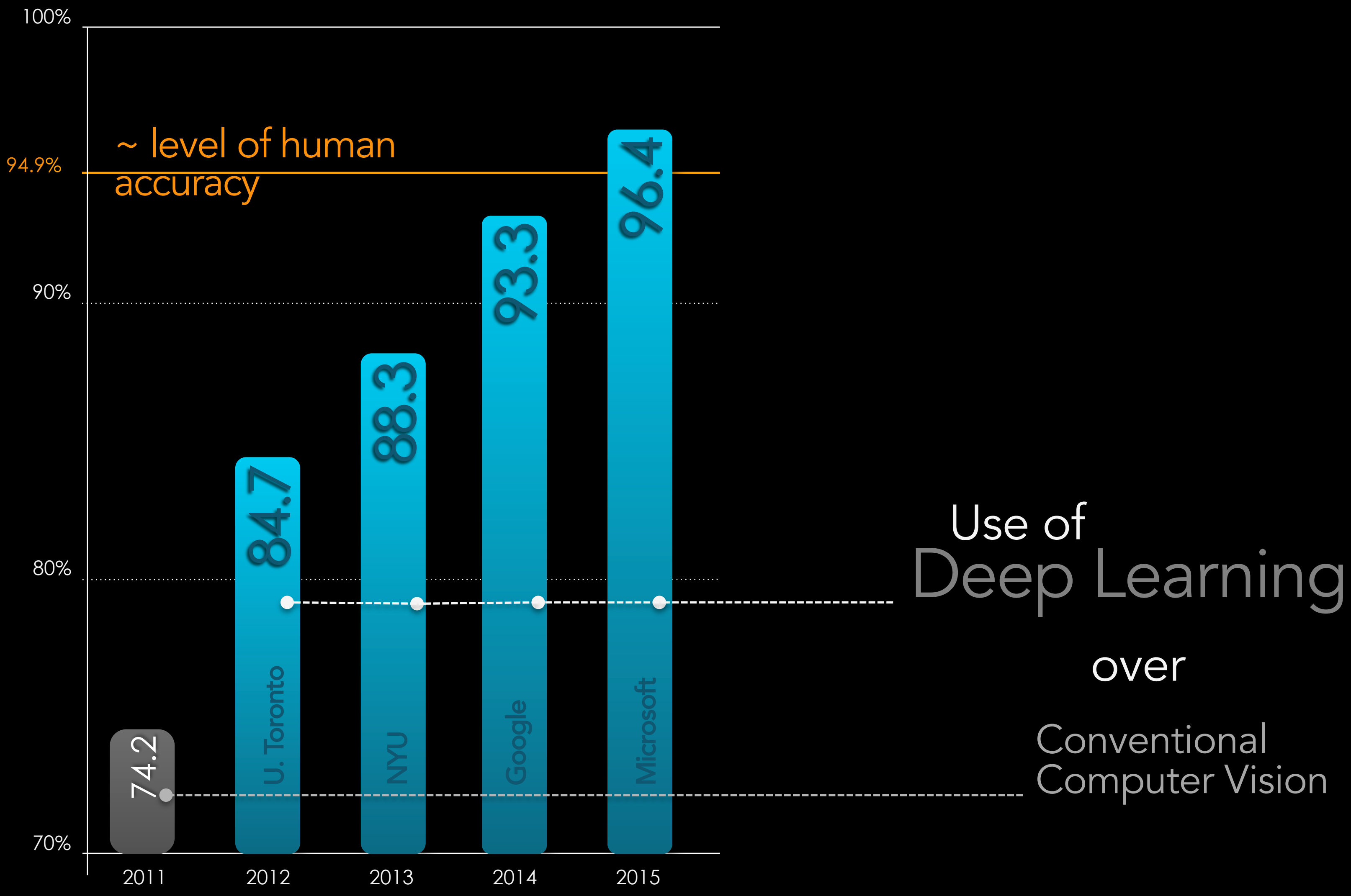
2012-2015: breakthrough in computer vision

- Graphics Processing Units (GPUs) + 10x more data
- 1,000 object categories,
- Facebook: millions of faces
- ***2015: human-level performance***



ImageNet Accuracy Still Improving

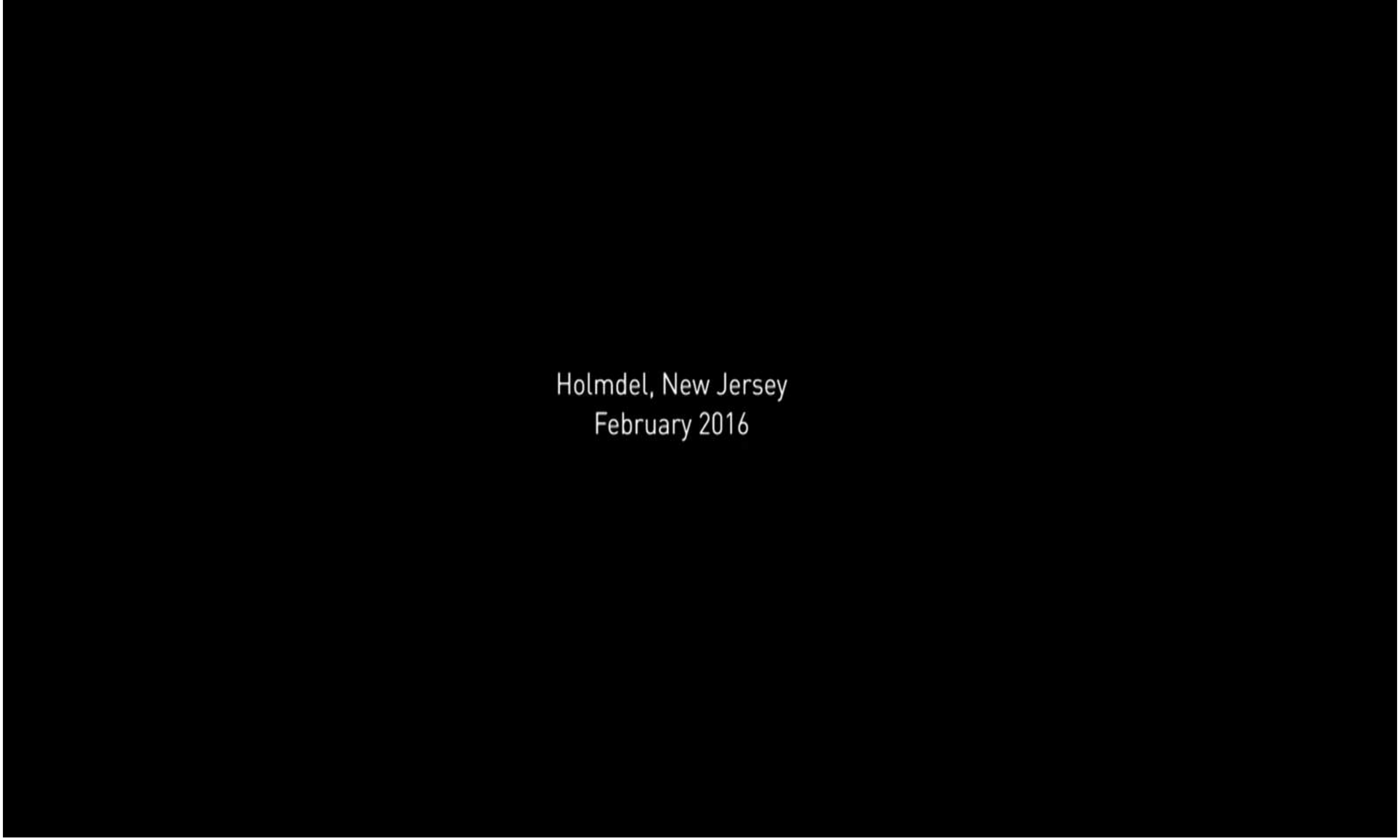
Top-5 Classification task



IT companies are racing into deep learning

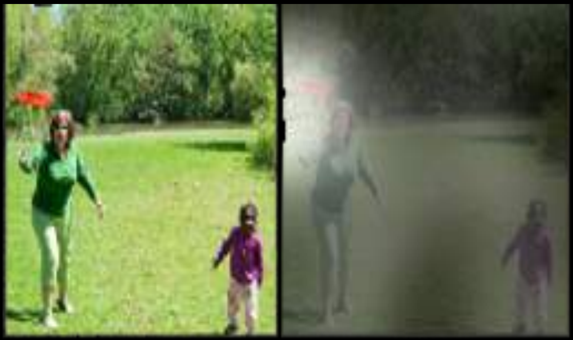
IBM
Google
SAMSUNG
NUANCE
YAHOO!
QUALCOMM
Microsoft
Baidu 百度
OpenAI
amazon
intel
Alibaba.com

From computer vision to self-driving cars: 2016

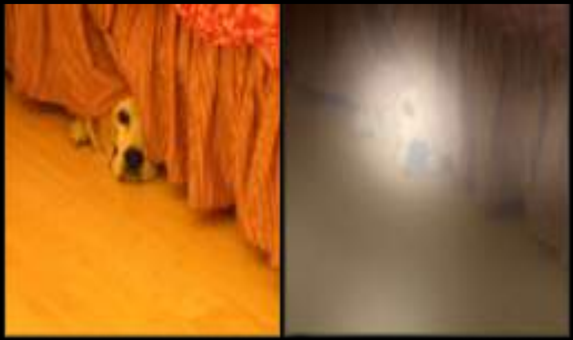


Holmdel, New Jersey
February 2016

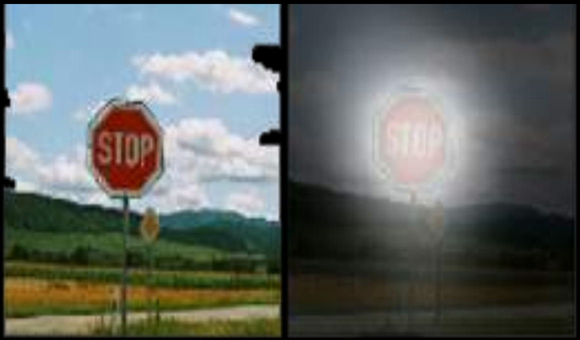
Ongoing progress: combining vision and natural language understanding



A woman is throwing a frisbee in a park.



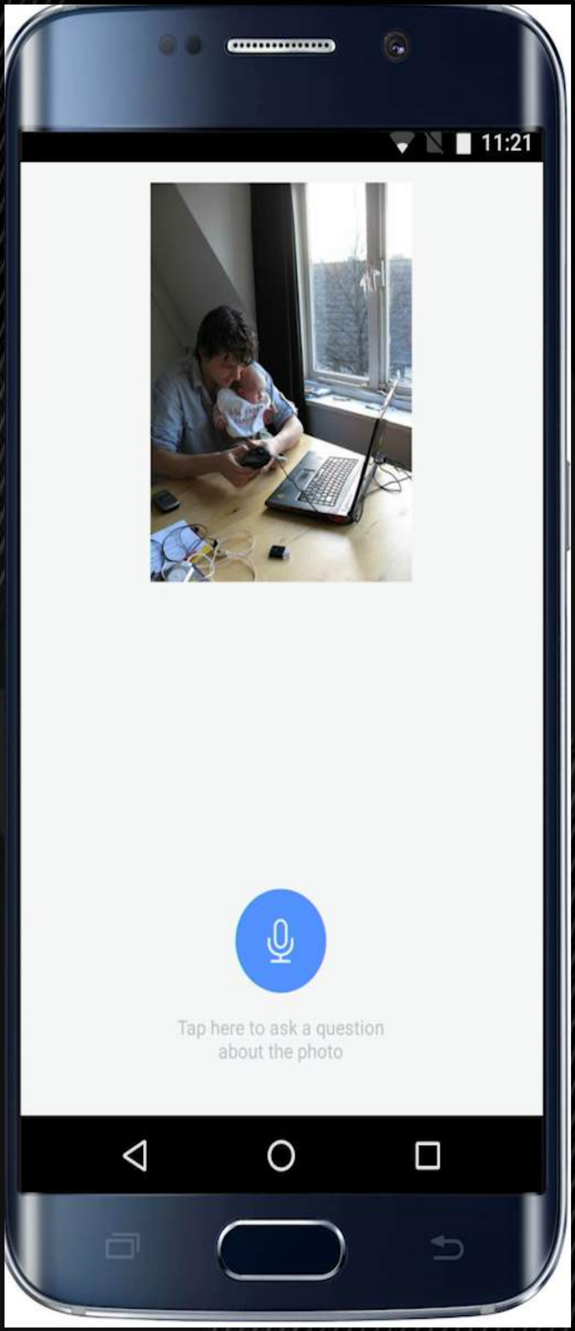
A dog is standing on a hardwood floor



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

With a lot more data…

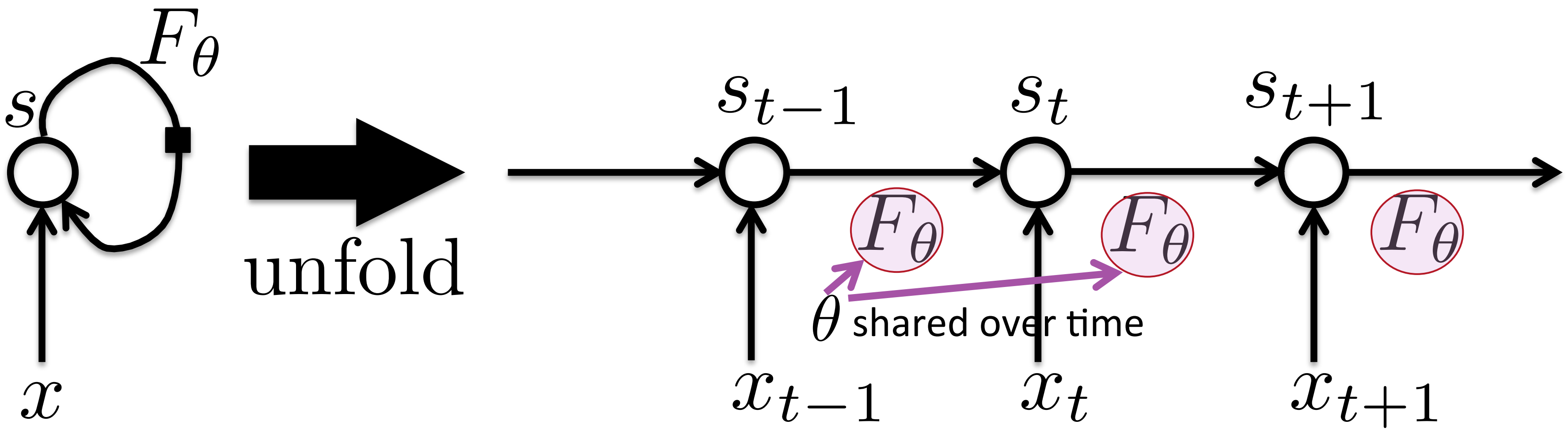
visual question answering



Recurrent Neural Networks

- Selectively summarize an input sequence in a fixed-size state vector via a recursive update

$$s_t = F_\theta(s_{t-1}, x_t)$$



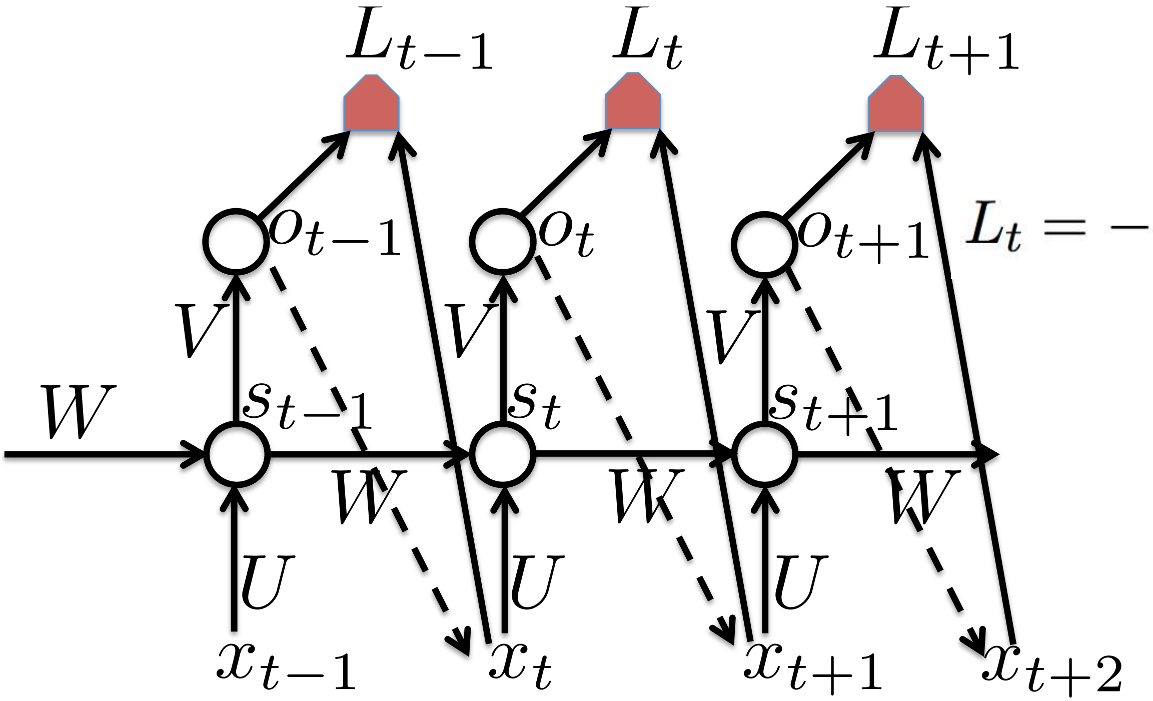
$$s_t = G_t(x_t, x_{t-1}, x_{t-2}, \ldots, x_2, x_1)$$

➔ **Generalizes naturally to new lengths not seen during training**

Generative RNNs

- An RNN can represent a fully-connected **directed generative model**: every variable predicted from all previous ones.

$$P(\mathbf{x}) = P(x_1, \ldots x_T) = \prod_{t=1}^{T} P(x_t | x_{t-1}, x_{t-2}, \ldots x_1)$$

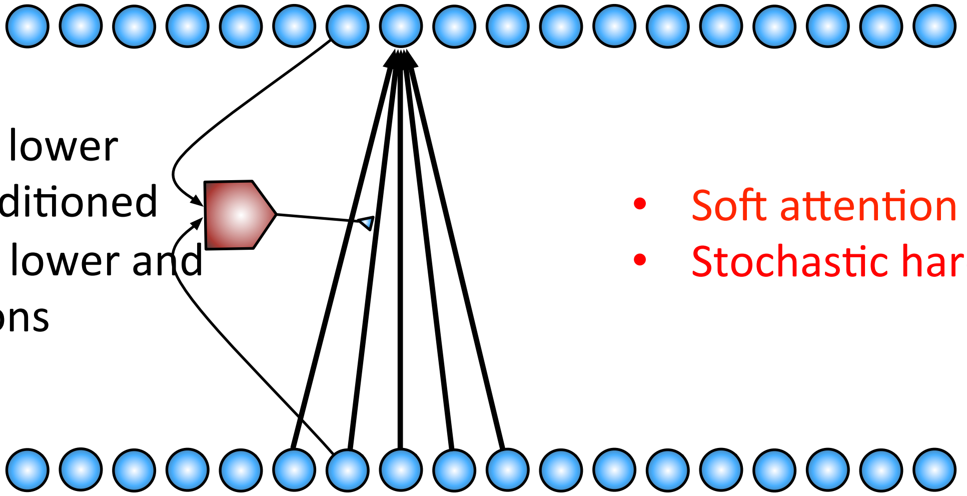


$$L_t = -\log P(x_t | x_{t-1}, x_{t-2}, \ldots x_1)$$

Attention Mechanism for Deep Learning

(Bahdanau, Cho & Bengio, ICLR 2015; Jean et al ACL 2015; Jean et al WMT 2015; Xu et al ICML 2015; Chorowski et al NIPS 2015; Firat, Cho & Bengio 2016)

- Consider an input (or intermediate) sequence or image
- Consider an upper level representation, which can choose « where to look », by assigning a weight or probability to each input position, as produced by an MLP, applied at each position



Higher-level

Softmax over lower locations conditioned on context at lower and higher locations

Lower-level

- Soft attention (backprop) vs
- Stochastic hard attention (RL)

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

(Bahdanau et al ICLR 2015, Jean et al ACL 2015, Gulcehre et al 2015, Firat et al 2016)

- 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

Google-Scale NMT Success

(Wu et al & Dean, Nature, 2016)

- After beating the classical phrase-based MT on the academic benchmarks, there remained the question: will it work on the very large scale datasets like used for Google Translate?
- Distributed training, very large model ensemble
- Not only does it work in terms of BLEU but it makes a killing in terms of human evaluation on Google Translate data

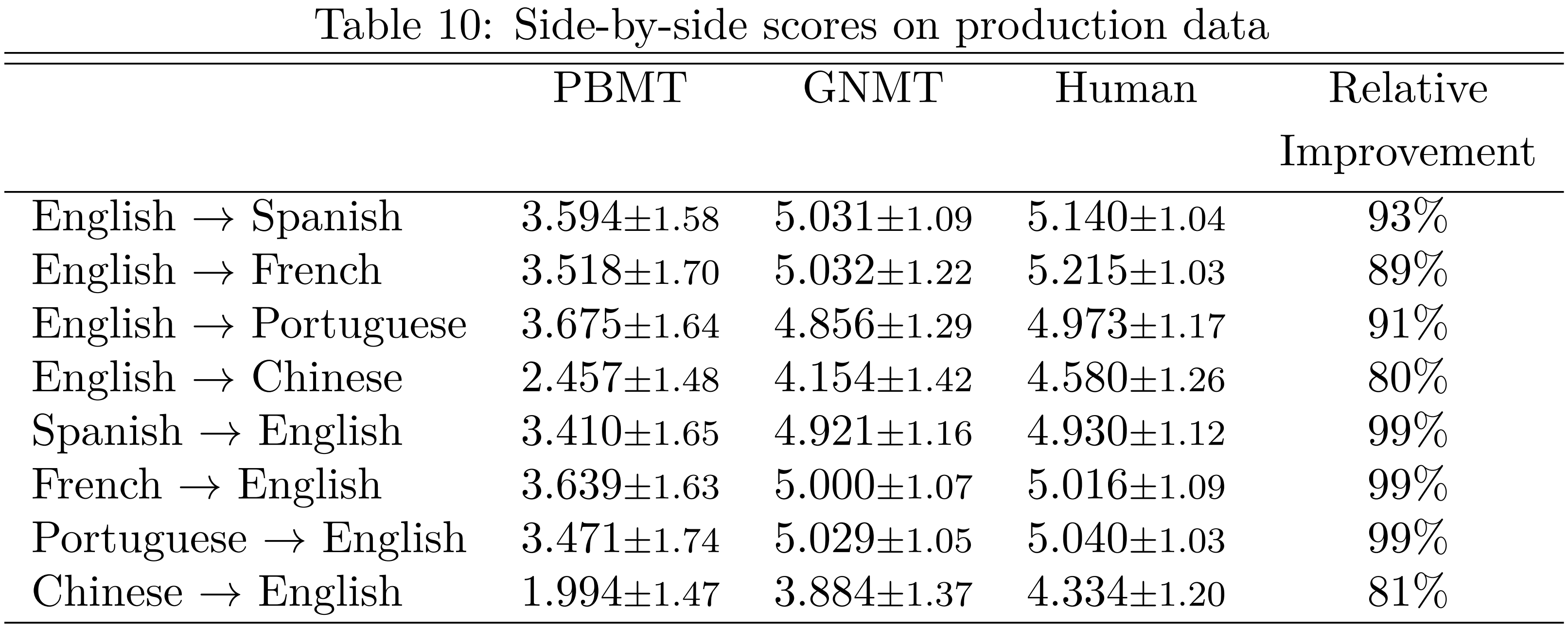


Table 10: Side-by-side scores on production data

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	3.594±1.58	5.031±1.09	5.140±1.04	93%
English → French	3.518±1.70	5.032±1.22	5.215±1.03	89%
English → Portuguese	3.675±1.64	4.856±1.29	4.973±1.17	91%
English → Chinese	2.457±1.48	4.154±1.42	4.580±1.26	80%
Spanish → English	3.410±1.65	4.921±1.16	4.930±1.12	99%
French → English	3.639±1.63	5.000±1.07	5.016±1.09	99%
Portuguese → English	3.471±1.74	5.029±1.05	5.040±1.03	99%
Chinese → English	1.994±1.47	3.884±1.37	4.334±1.20	81%

Deep Learning: Beyond Pattern Recognition, towards AI

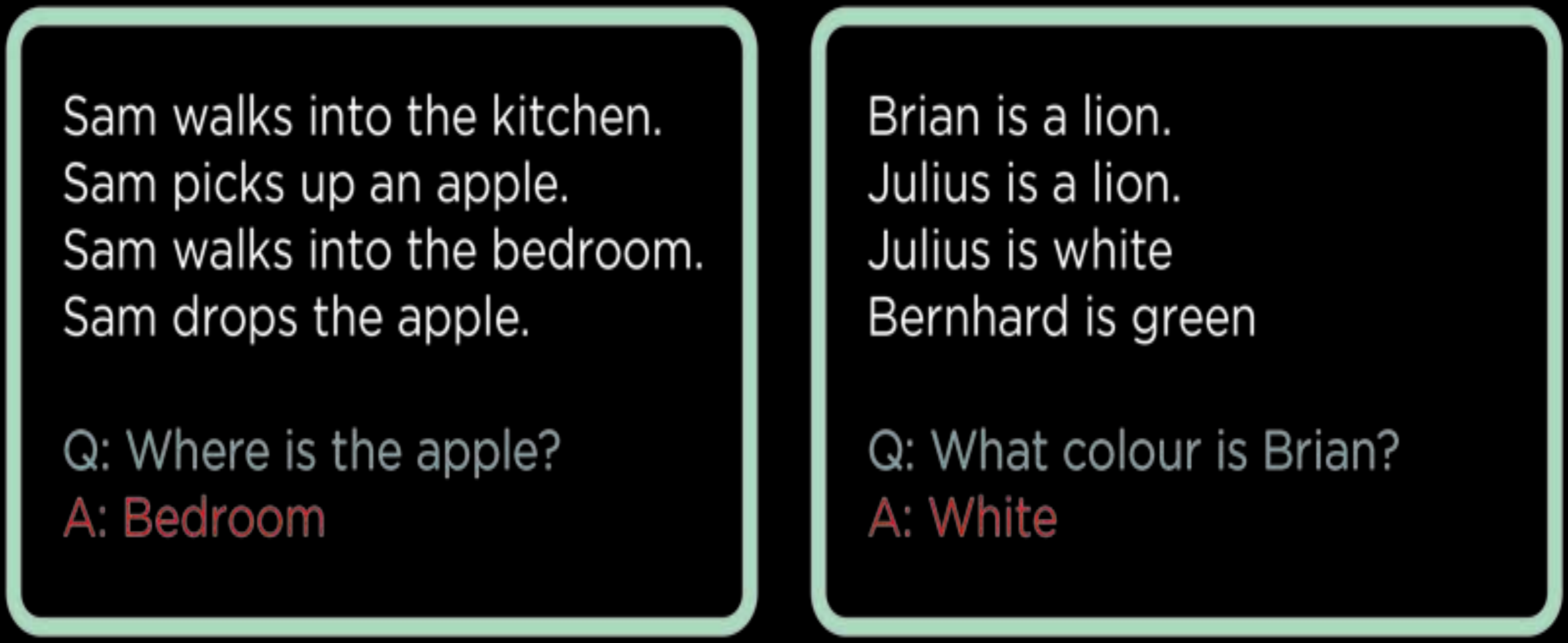
- Many researchers believed that neural nets could at best be good at pattern recognition
- And they are really good at it!

- 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 and Go game playing) & Berkeley (Robotic control)

The next frontier:
to reason and answer questions



Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.

Q: Where is the apple?
A: Bedroom

Brian is a lion.
Julius is a lion.
Julius is white
Bernhard is green

Q: What colour is Brian?
A: White

The Biggest Challenge: Unsupervised Learning & Learning Commonsense Autonomously

- 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
  - Necessary for RL without given model or domain simulator

Learning « How the world ticks »

- So long as our machine learning models « cheat » by relying only on surface statistical regularities, they remain vulnerable to out-of-distribution examples
- Humans generalize better than other animals by implicitly having a more accurate internal model of the underlying causal relationships
- This allows one to predict future situations (e.g., the effect of planned actions) that are far from anything seen before, an essential component of reasoning, intelligence and science

Invariance and Disentangling

- Invariant features

- Which invariances?

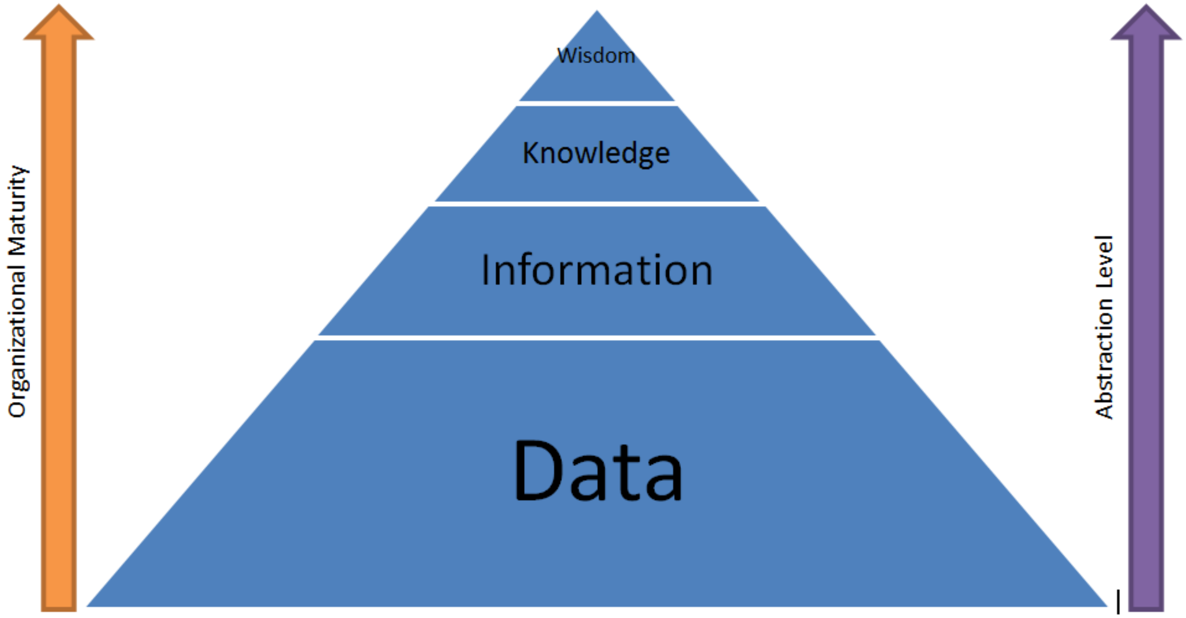
- Alternative: learning to disentangle factors

- Good disentangling →

  avoid the curse of dimensionality

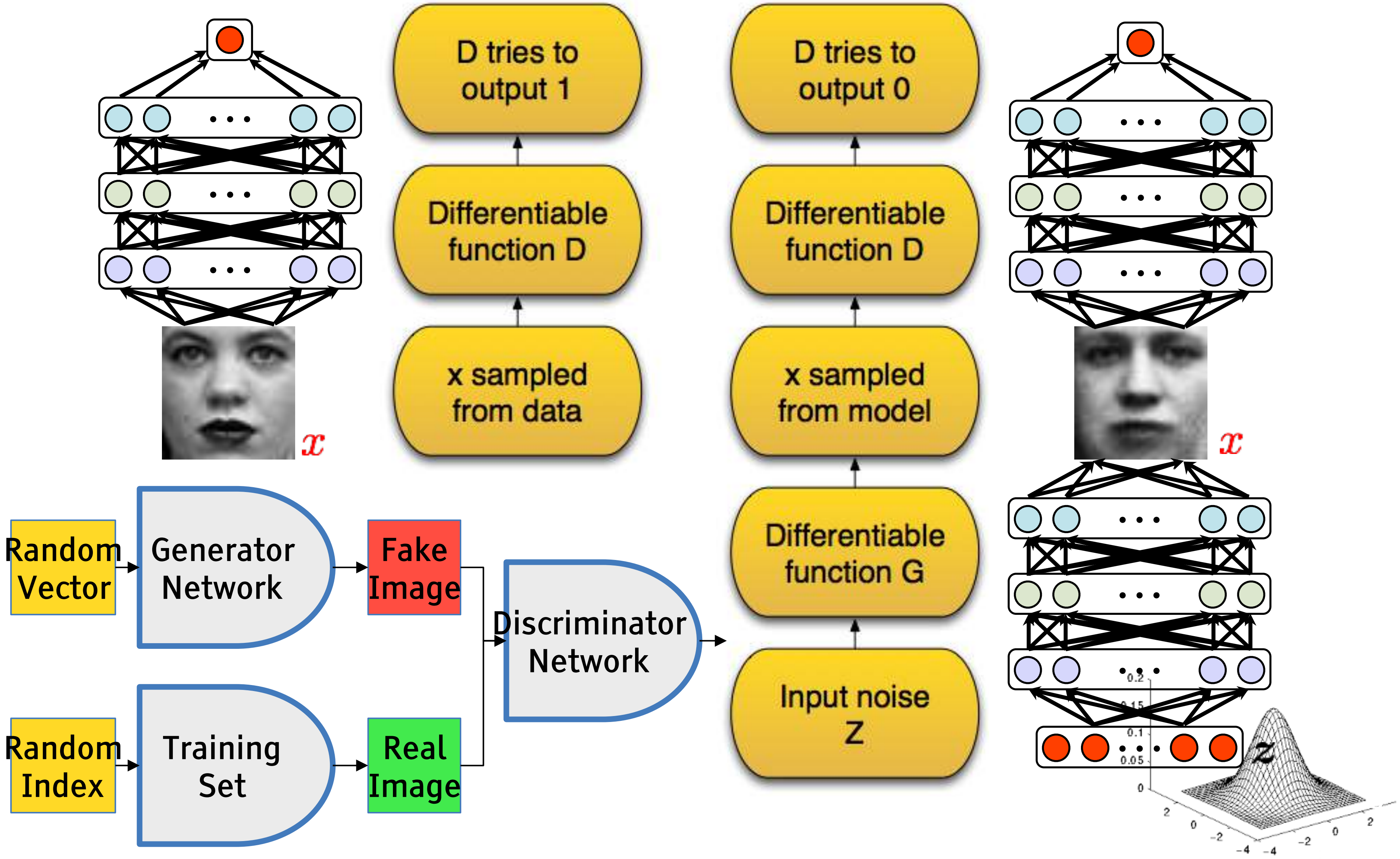
Learning Multiple Levels of Abstraction

- 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



GAN: Generative Adversarial Networks

Goodfellow et al NIPS 2014



D tries to output 1

Differentiable function D

x sampled from data

D tries to output 0

Differentiable function D

x sampled from model

Differentiable function G

Input noise Z

Random Vector

Generator Network

Fake Image

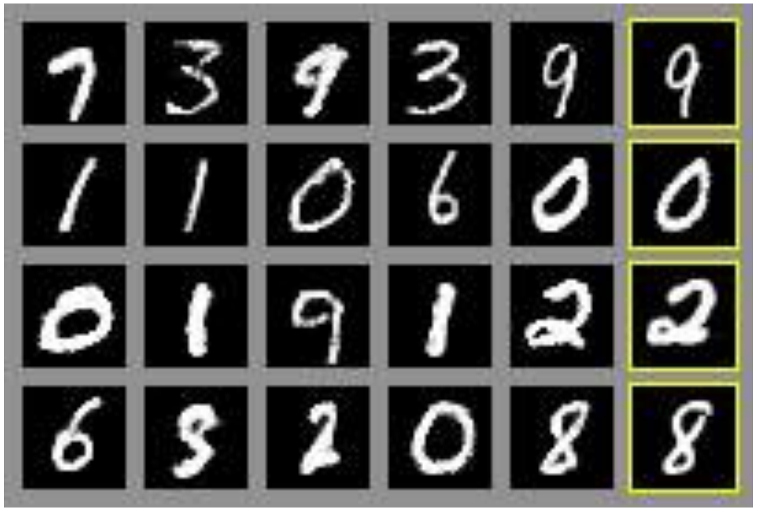
Random Index

Training Set

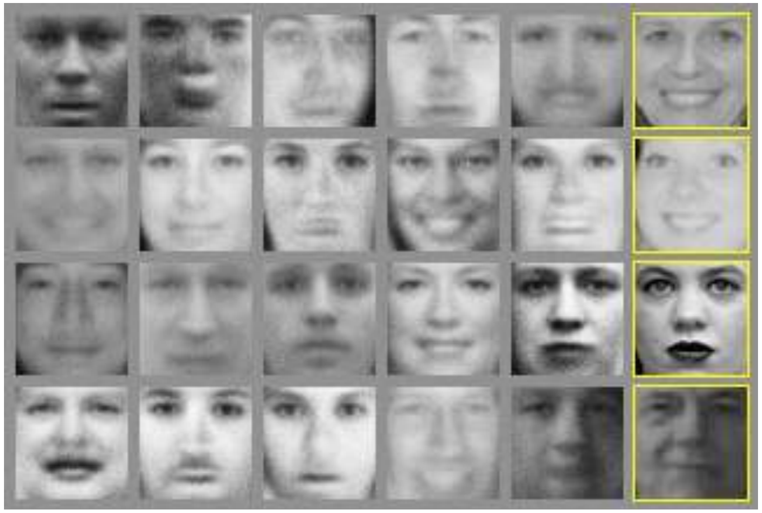
Real Image

Discriminator Network

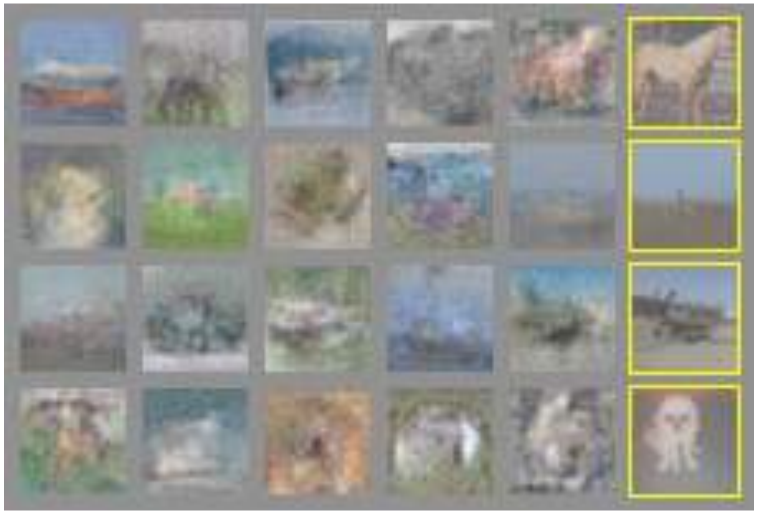
Early Days of GAN Samples



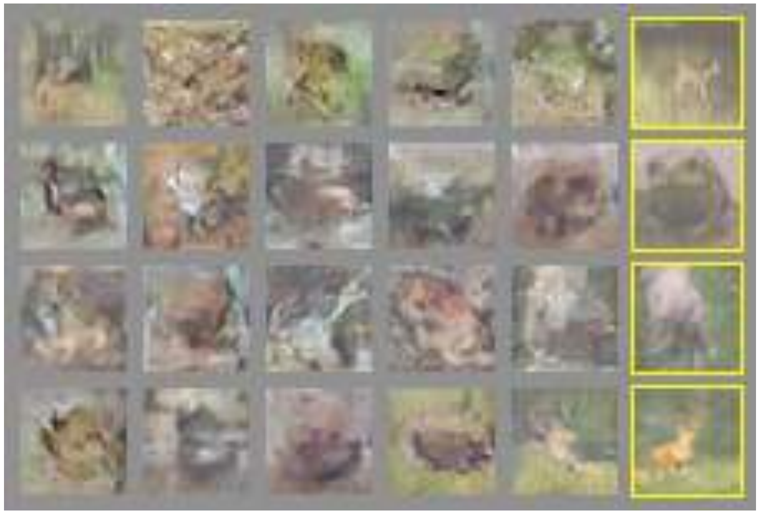
MNIST



TFD



CIFAR-10 (fully connected)

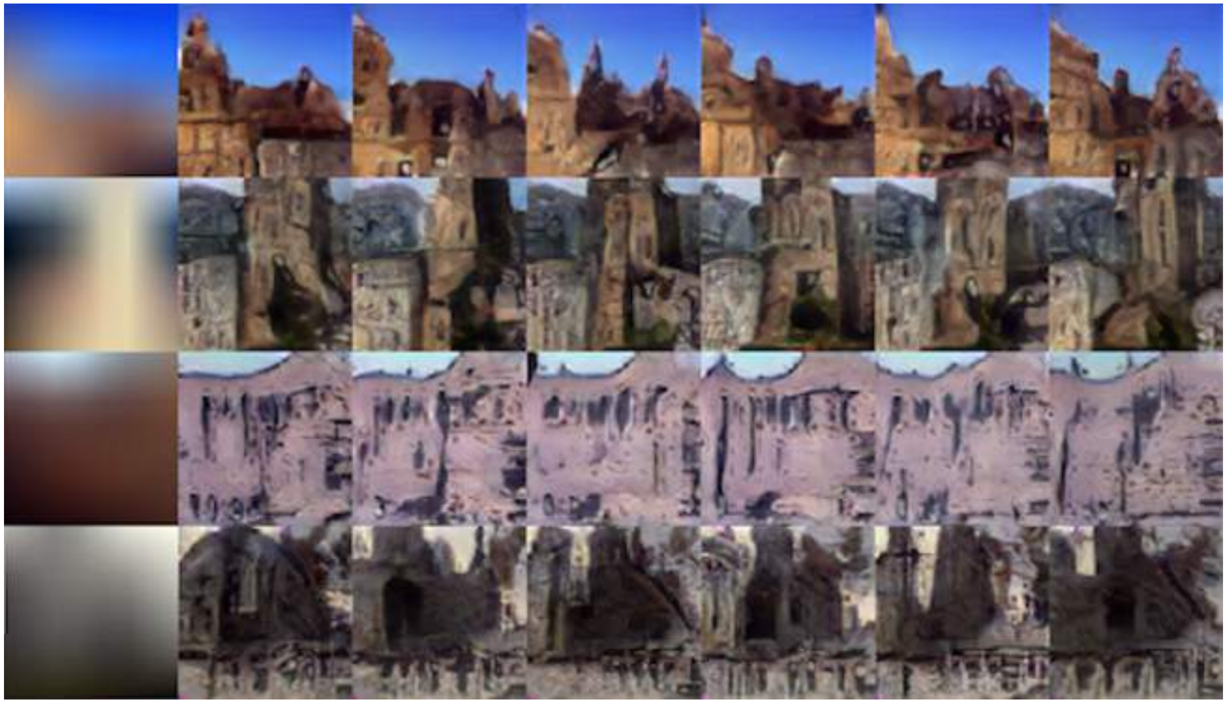


CIFAR-10 (convolutional)

LAPGAN: Visual Turing Test

(Denton et al 2015)

- 40% of samples mistaken *by humans* for real photos

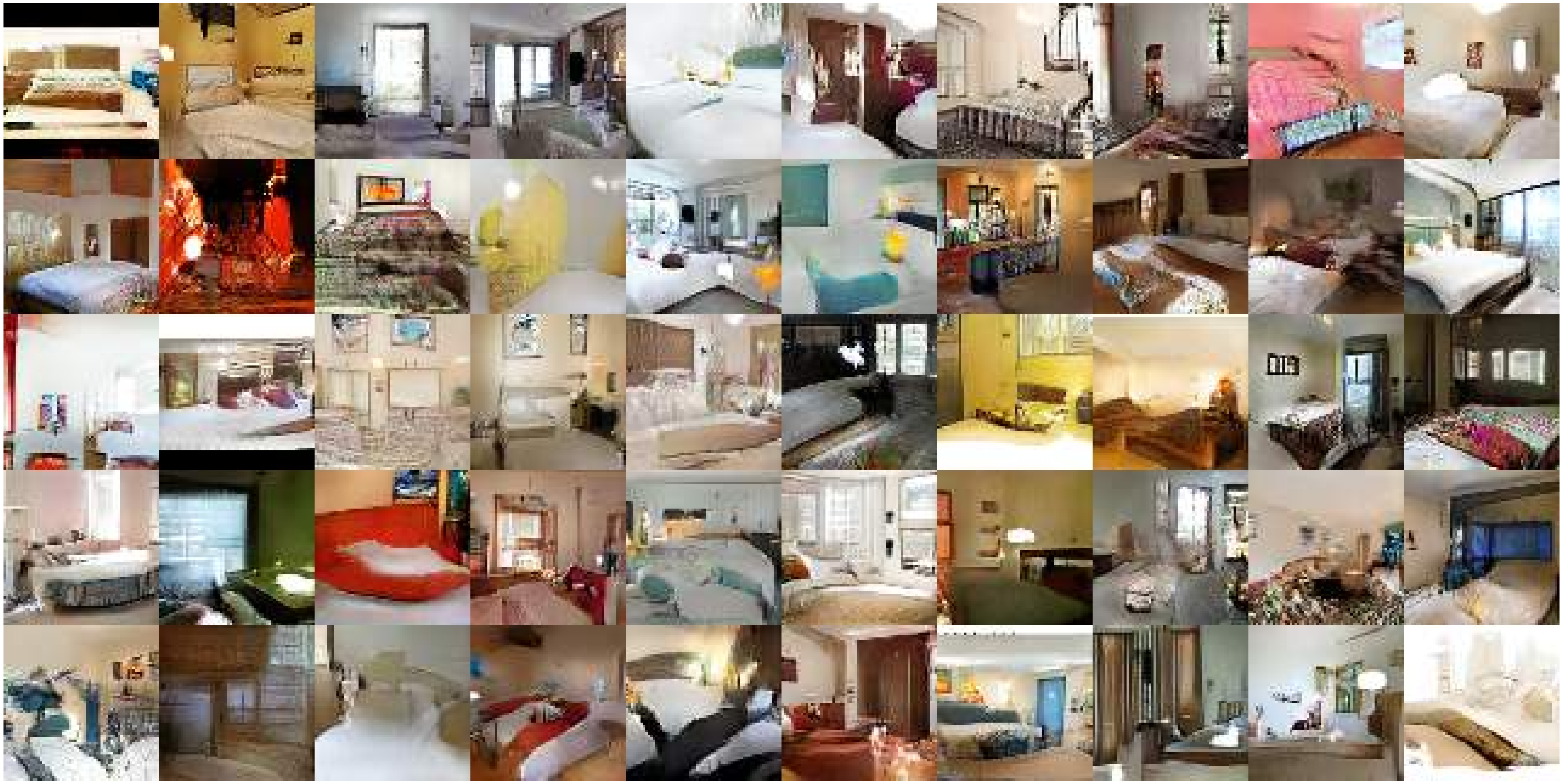


- Sharper images than max. lik. proxys (which min. KL(data|model)):
- GAN objective = compromise between KL(data|model) and KL(model|data)

Convolutional GANs

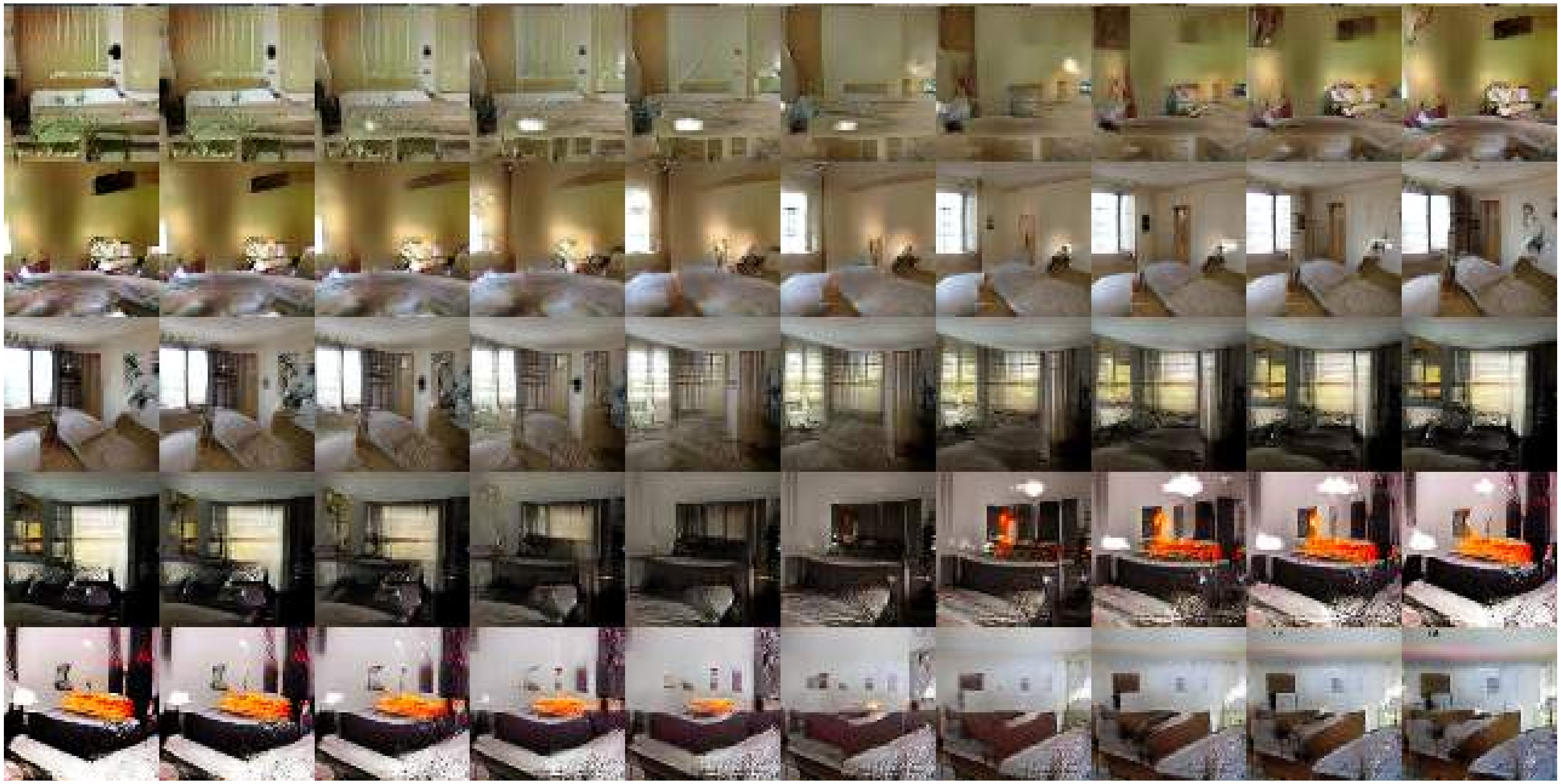
(Radford et al, arXiv 1511.06343)

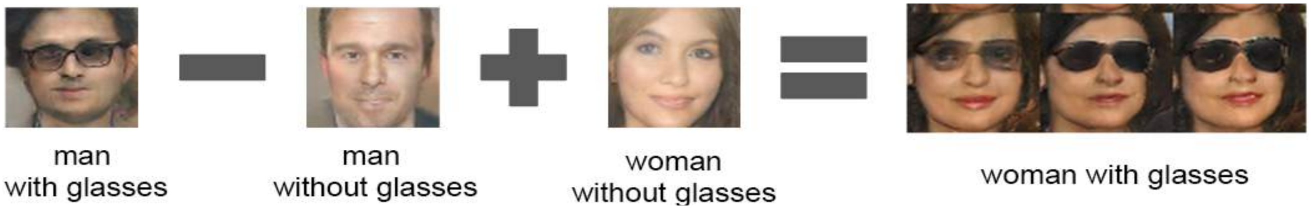
Strided convolutions, batch normalization, only convolutional layers, ReLU and leaky ReLU



GAN: Interpolating in Latent Space

If the model is good (unfolds the manifold), interpolating between latent values yields plausible images.





Combining Iterative Sampling from Denoising Auto-Encoders with GAN

Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space

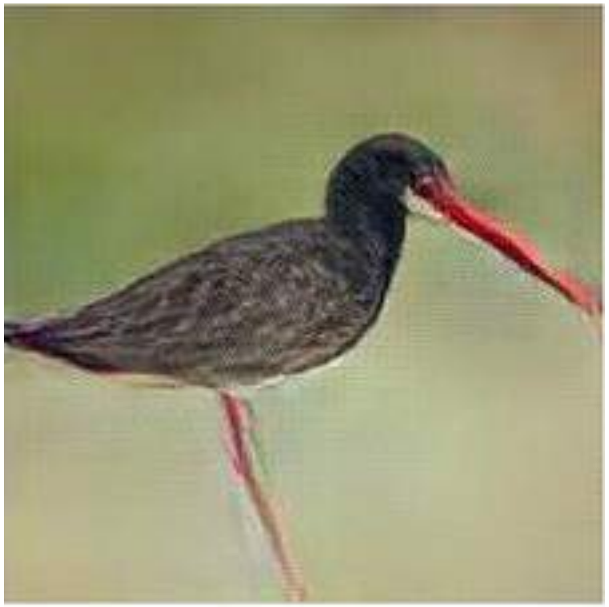
Anh Nguyen, Jason Yosinski, Yoshua Bengio, Alexey Dosovitskiy, Jeff Clune

(submitted to CVPR 2017) arXiv:1612.00005

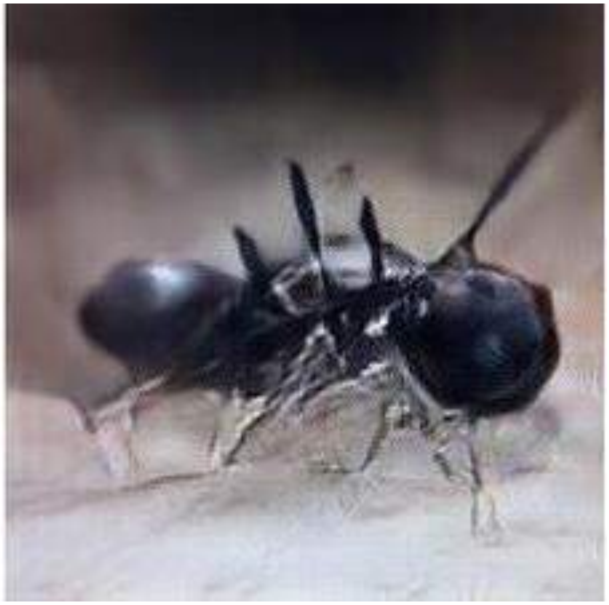
227 x 227 ImageNet GENERATED IMAGES of category Volcano

Plug & Play Generative Networks

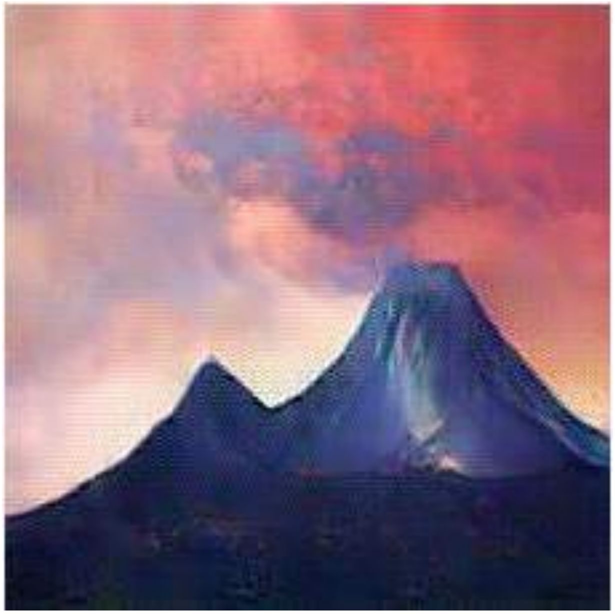
High-Resolution
Samples
227 x 227



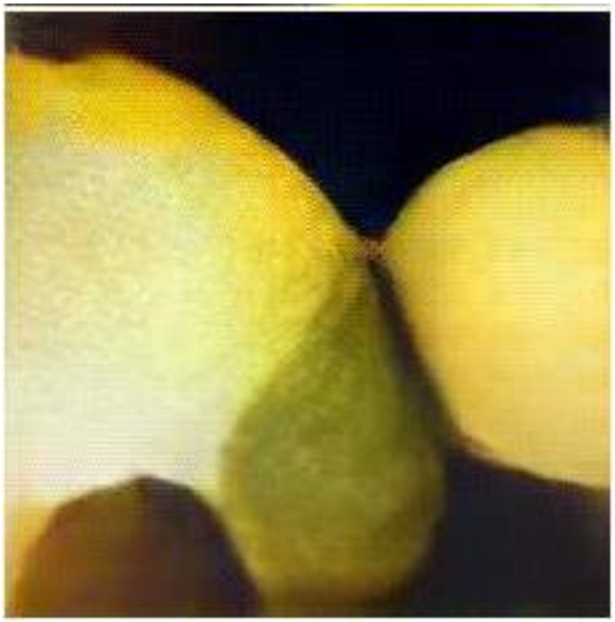
bird



ant



volcano



lemon

More Technical Challenges

- Learning long-term dependencies in recurrent neural networks
- Optimization challenge of training deep neural networks
- Taking advantage of feedback connections for attention, iterative inference & learning
- Incorporating “general knowledge” or commonsense (mostly from unsupervised learning) in RL

Applications on the horizon

Computer Interaction

Healthcare

Robotics

How to Attract the Best Researchers in Industry

- Extreme current demand for deep learning expertise, crazy salaries and acquisitions
- Not enough trained PhDs, too much industry demand
- Long-term **open research**
  - Necessary to attract and retain the strongest researchers
  - Success stories: DeepMind, FAIR, OpenAI
  - Need a pipeline & portfolio of different horizons
- Focused research: strategic, targeted choices
- Untying research org. from product-driven R&D

Open Science & Open Source

- Best deep learning researchers (even in industry) demand open science →
  - Open and early publications (arXiv)
  - Accessible open source code (github)
- Both are
  - Reputation building (attracts more scientists)
  - Reproducible science
  - Generate follow-ups, citations & impact
  - Responsible: contribute to the community

Montreal Institute for Learning Algorithms

MILA

Université de Montréal