

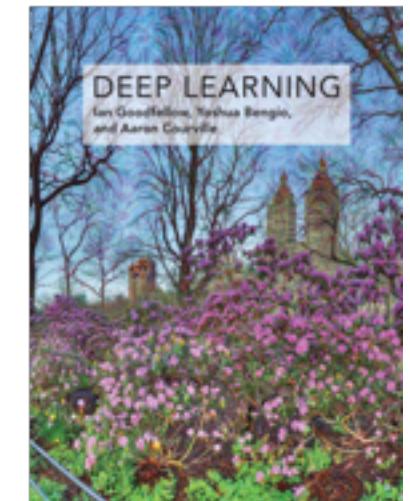
From Deep Learning of Disentangled Representations to Higher-Level Cognition

Yoshua Bengio, October 24th, 2017

University of Amsterdam



PLUG: Deep Learning, MIT Press book is out,
chapters will remain online



Still Far from Human-Level AI

- Industrial successes mostly based on **supervised** learning



- Learning superficial clues, not generalizing well enough outside of training contexts, easy to fool trained networks:
 - Current models cheat by picking on surface regularities

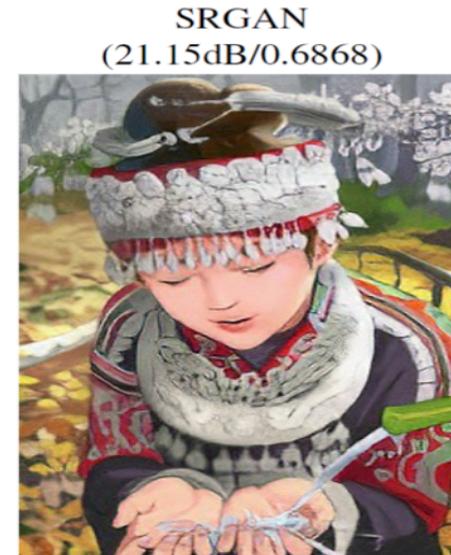
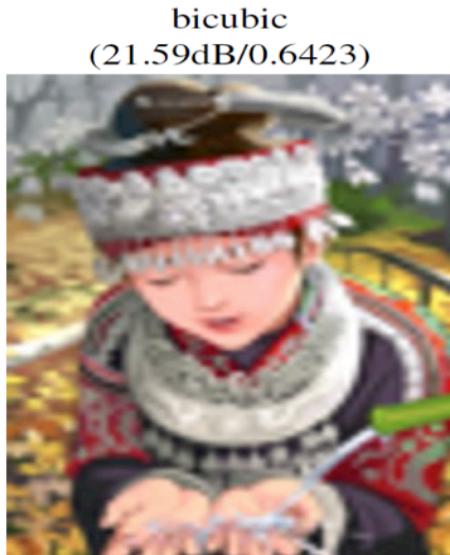
The need for predictive causal modeling: rare & dangerous states

- Example: autonomous vehicles in near-accident situations
- Current supervised learning may not handle well these cases because they are too rare (not enough data)
- It would be even worse with current RL (statistical inefficiency)
- Long-term objective: develop better predictive models of the world able to **generalize in completely unseen scenarios**, but it does not seem reasonable to model the sequence of future states in all their details
- Human drivers: no need to die a thousand deaths

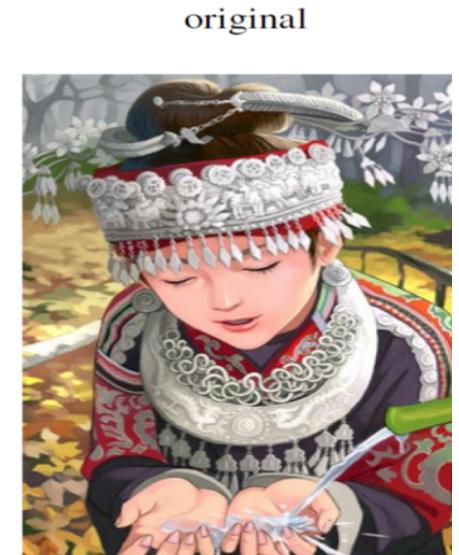


Deep Generative Models & GANs

- Simulation
- Filling-in missing data
- Generating text from speech, speech from text, image 2 image
- Inpainting, denoising, **super-resolution**
- Art, etc.



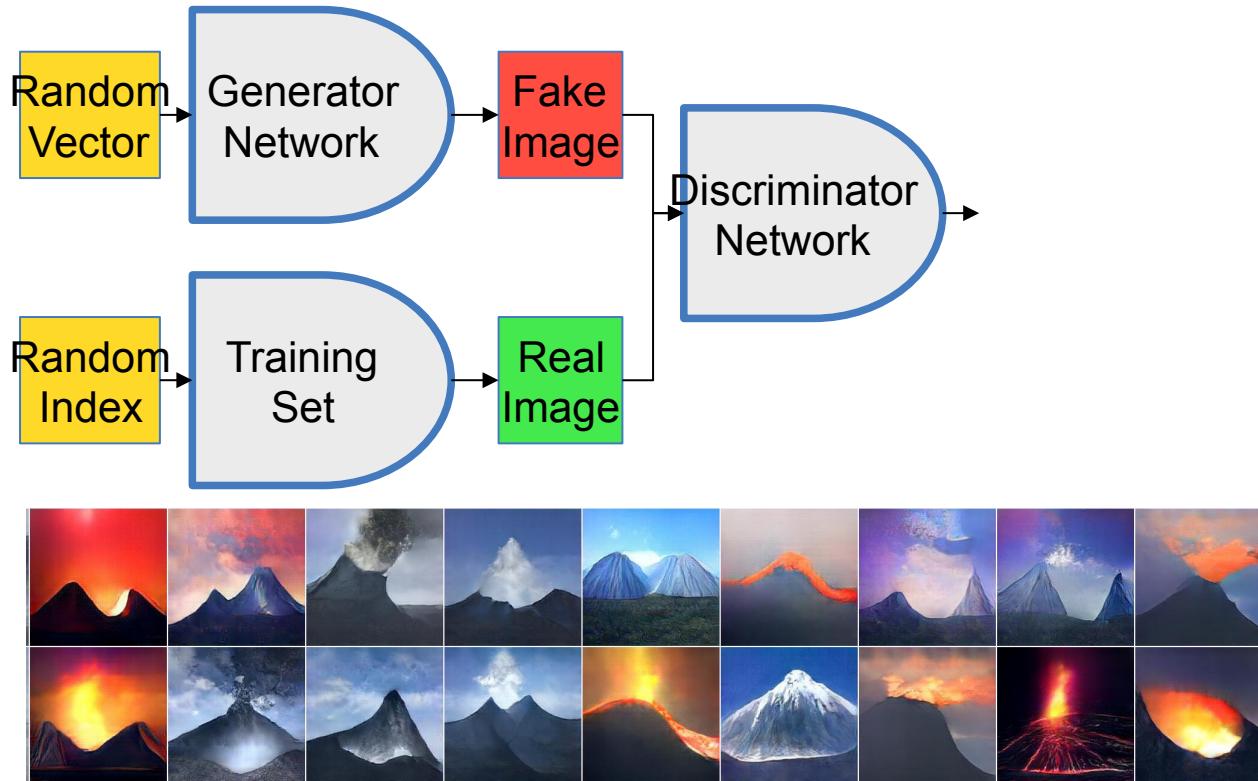
Ledig et al. 2017



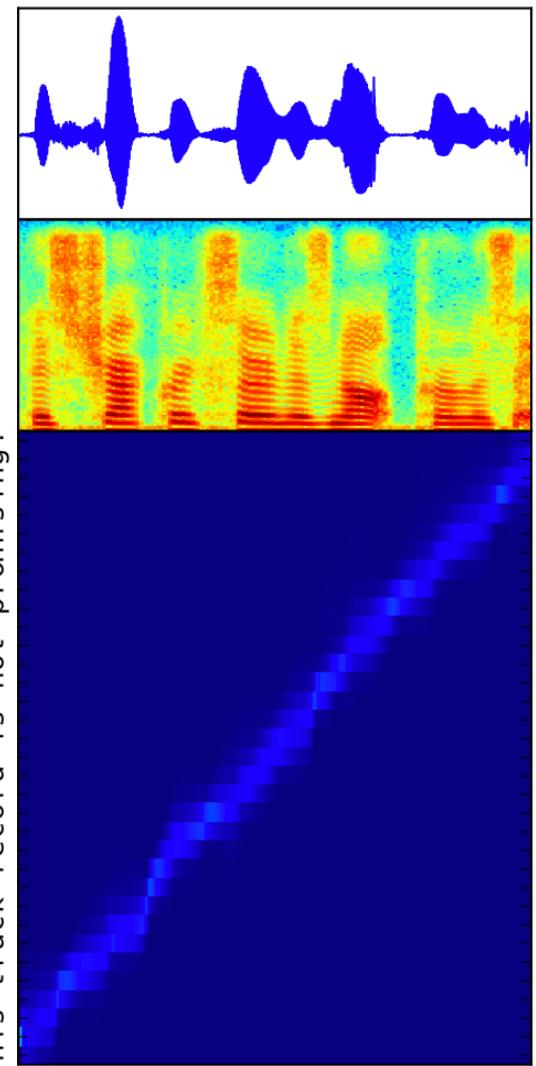
Deep Unsupervised Learning Takes off with GANs (NIPS'2014)



- Progress in **unsupervised generative neural nets** allows them to synthesize a diversity of images, sounds and text imitating unlabeled images, sounds or text



This track record is not promising.



Text 2 Image, Colorization

This bird is red and brown in color, with a stubby beak



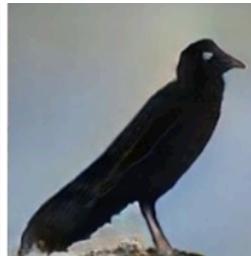
The bird is short and stubby with yellow on its body



A bird with a medium orange bill white body gray wings and webbed feet



This small black bird has a short, slightly curved bill and long legs



A small bird with varying shades of brown with white under the eyes



A small yellow bird with a black crown and a short black pointed beak



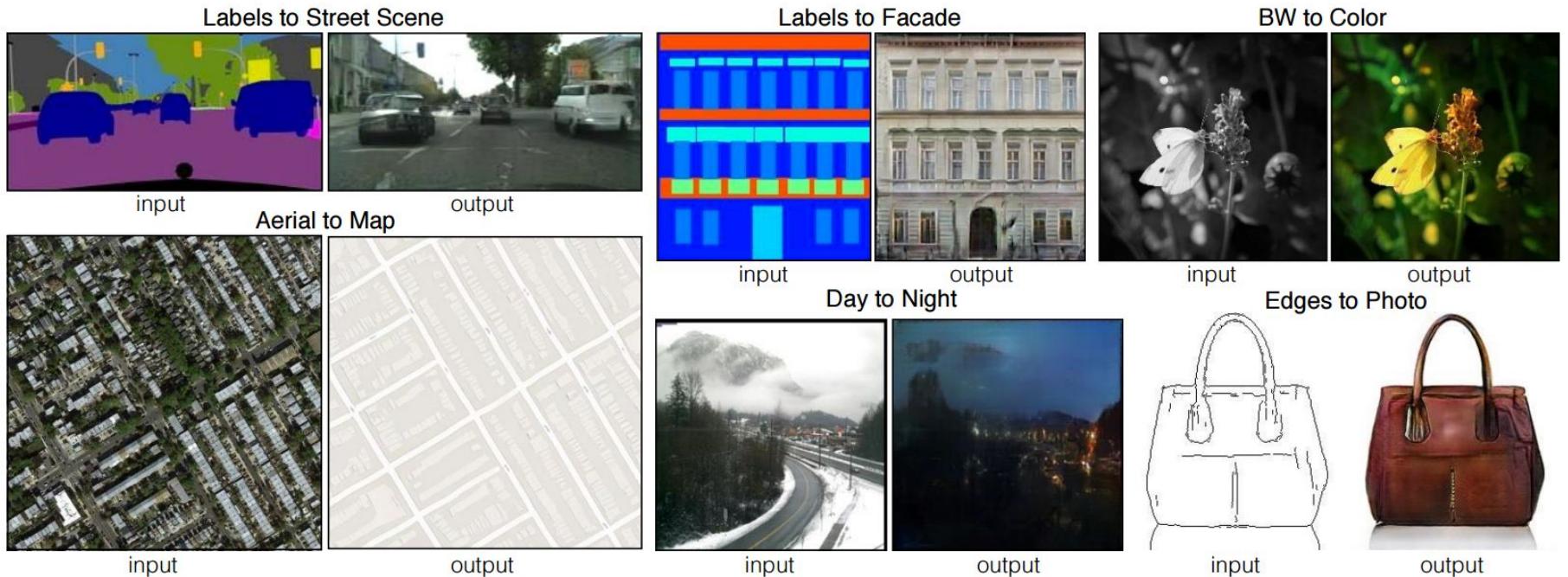
This small bird has a white breast, light grey head, and black wings and tail



[Zhang et al. 2017](#)

[Lucy Li](#)

Image 2 Image



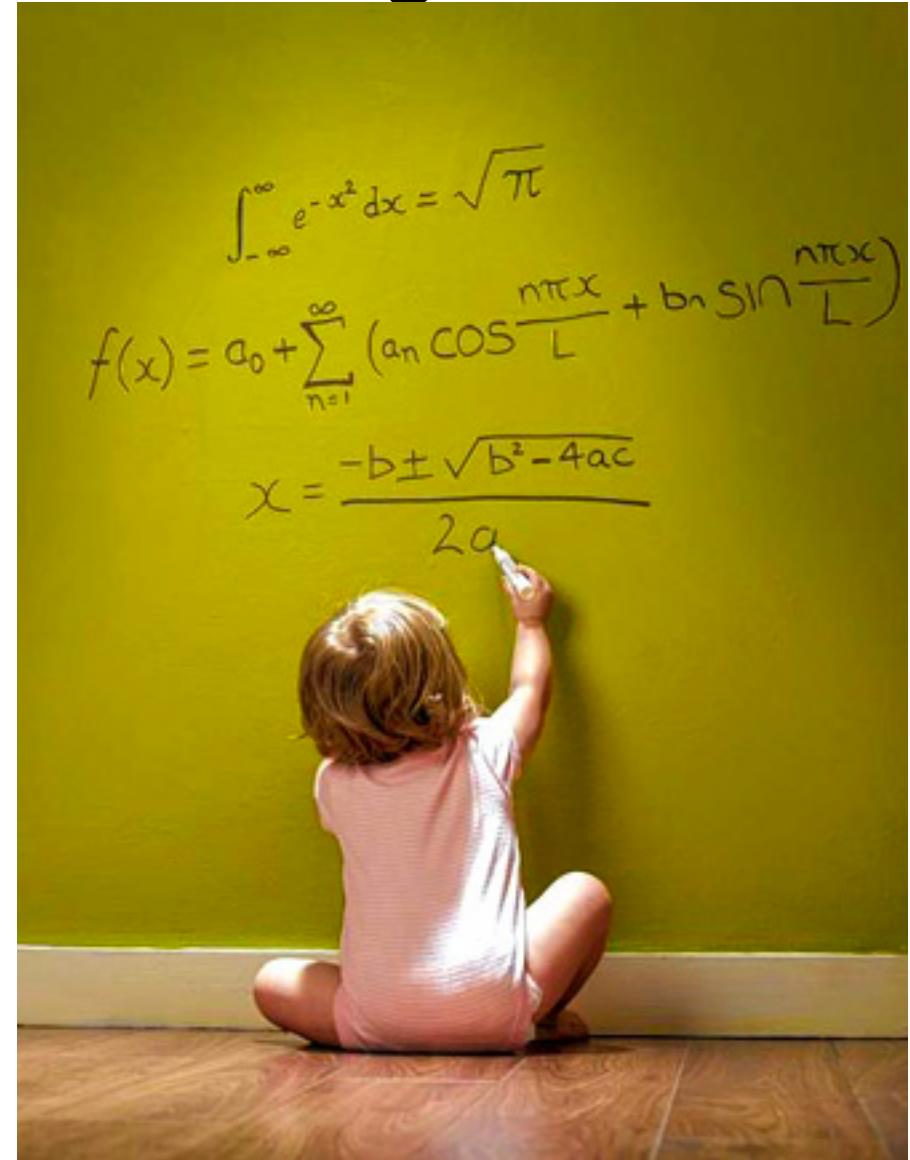
[Isola et al. 2016](#)

What's Missing?

Deep
Understanding

Humans outperform machines at autonomous learning

- Humans are very good at unsupervised learning, e.g. a 2 year old knows intuitive physics
- Babies construct an approximate but sufficiently reliable model of physics, how do they manage that?
Note that **they interact with the world**, not just observe it.

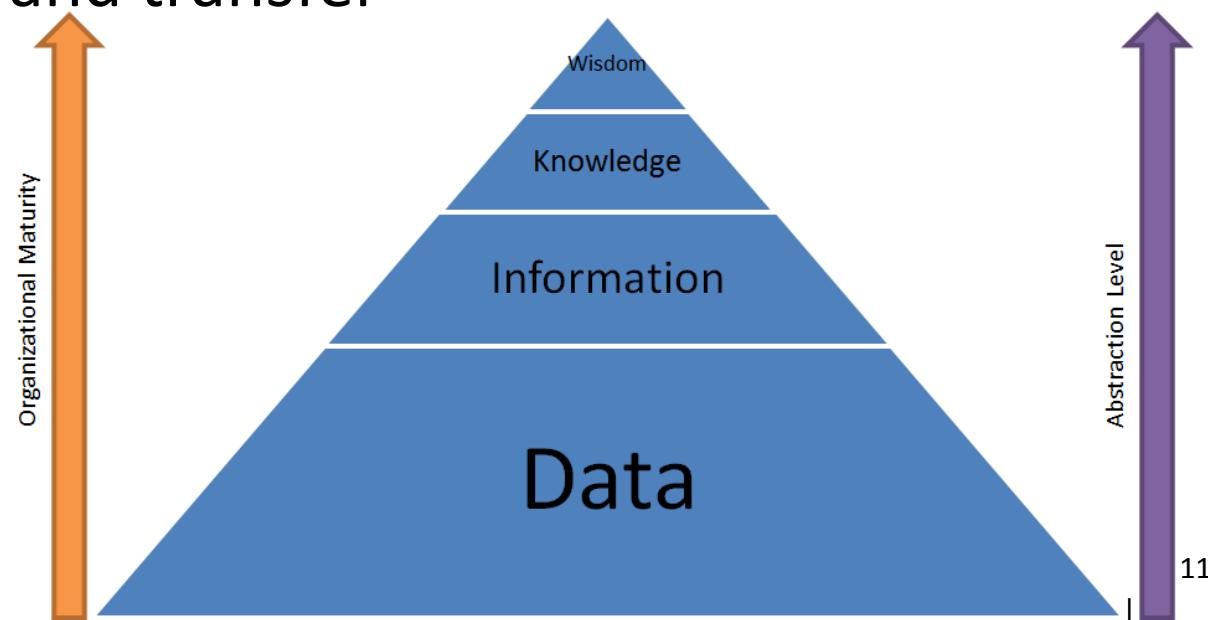


What's Missing?

Abstract
Representations

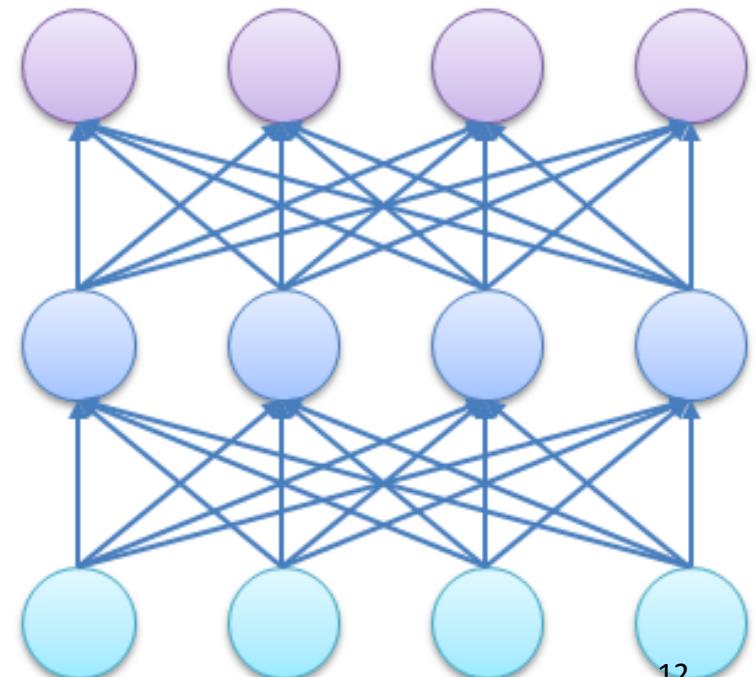
Learning Multiple Levels of Abstraction

- 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



How to Discover Good Representations

- How to discover abstractions?
- What is a good representation?
- Need clues to **disentangle** the underlying factors
 - Spatial & temporal scales
 - Marginal independence
 - *Controllable factors*



Acting to Guide Representation Learning & Disentangling



- Some factors (e.g. objects) correspond to ‘independently controllable’ aspects of the world
- *Can only be discovered by acting in the world*

Independently Controllable Factors

(Emmanuel Bengio, Valentin Thomas, Joelle Pineau, Doina Precup, Yoshua Bengio, 2017)

(Valentin Thomas, Jules Pondard, Emmanuel Bengio, Marc Sarfati, Philippe Beaudoin, Marie-Jean Meurs, Joelle Pineau, Doina Precup, Yoshua Bengio, 2017)

- Jointly train for each aspect (factor)
 - A policy π_k (which tries to selectively change just that factor)
 - A representation (which maps state to value of factor) f_k

Discrete case, $\phi \in \{1, \dots, N\}$, define *selectivity*:

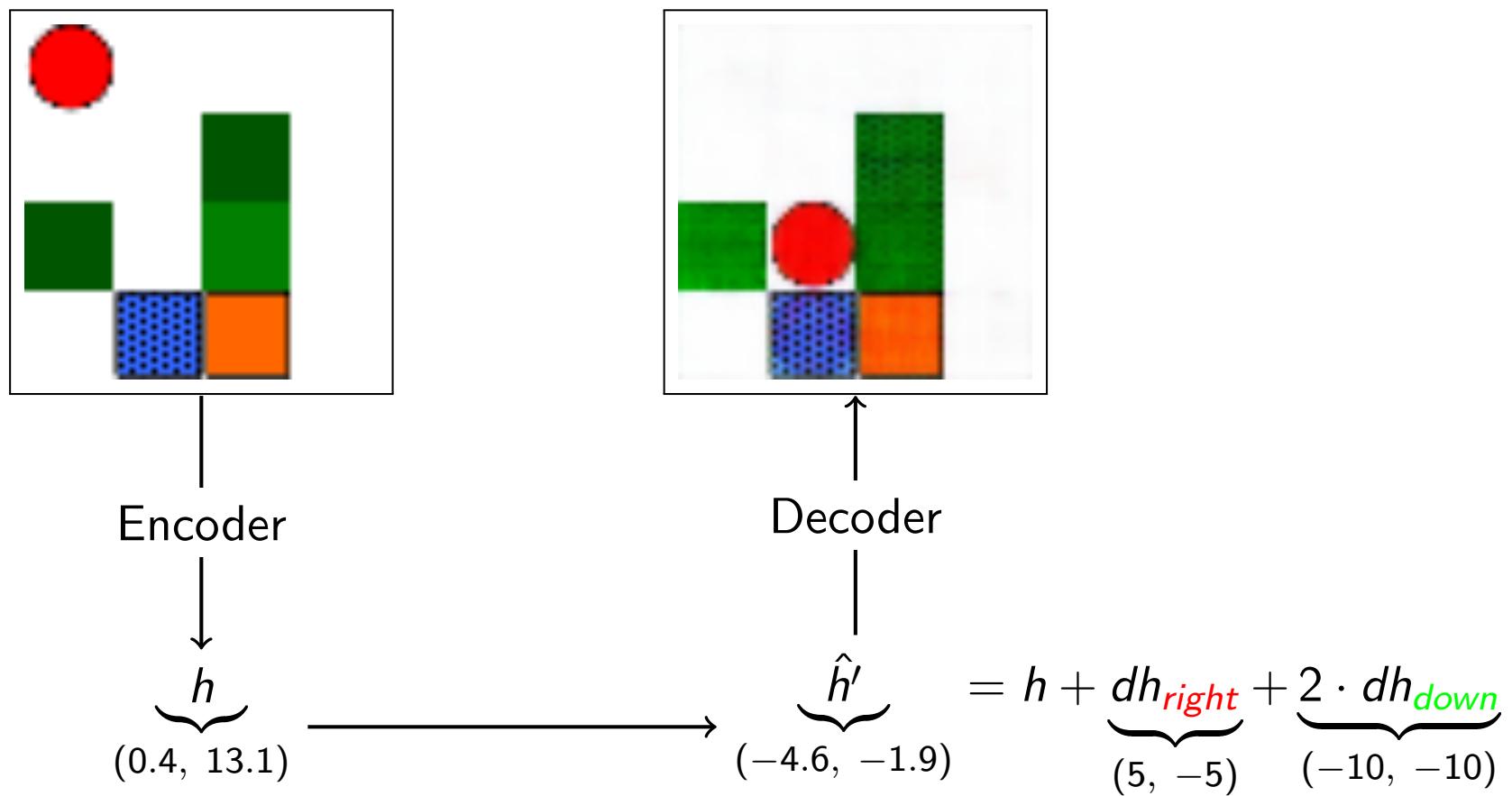
$$\sum_{k=1}^N \mathbb{E}_{(s_t, a_t, s_{t+1})} \left[\pi_k(a_t | s_t) \frac{f_k(s_{t+1}) - f_k(s_t)}{\sum_{k'} |f_{k'}(s_{t+1}) - f_{k'}(s_t)|} \right]$$

- Optimize both policy π_k and representation f_k to minimize

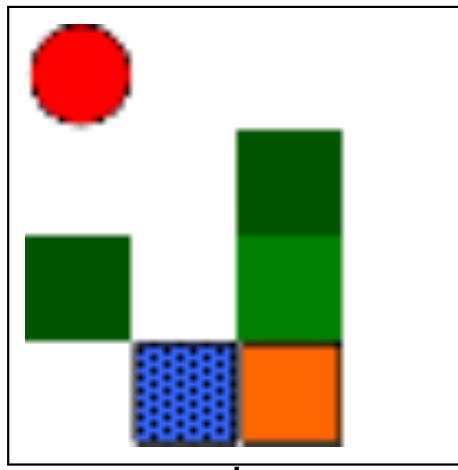
$$\underbrace{\mathbb{E}_s[\frac{1}{2} \|s - g(f(s))\|_2^2]}_{\text{reconstruction error}} - \lambda \underbrace{\sum_k \mathbb{E}_s[\sum_a \pi_k(a | s) \log sel(s, a, k)]}_{\text{disentanglement objective}}$$

Predict the effect of actions in attribute space

Given initial state and set of actions, predict new attribute values and the corresponding reconstructed images



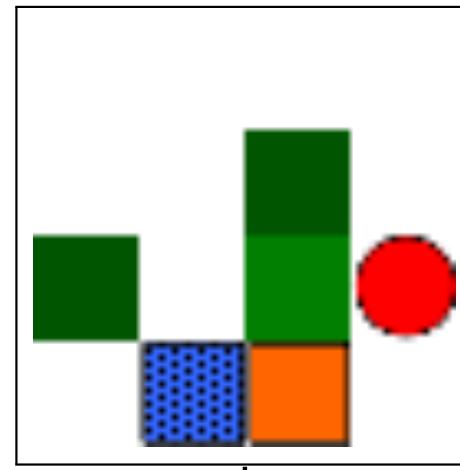
Given two states, recover the causal actions leading from one to the other



Encoder



$\underbrace{h_1}_{(0.4, 13.1)}$



Encoder



$\underbrace{h_2}_{(5.9, -11.6)}$

$$dh = (5.5, -24.8) \approx 2 \cdot dh_{\text{down}} + 3 \cdot dh_{\text{right}}$$

Continuous Set of Attributes: Attribute Embeddings = variable name

Principle

We map controllable factors to embeddings ϕ instead of coordinates k (**one** policy network). Discovers by itself the relevant number of features.

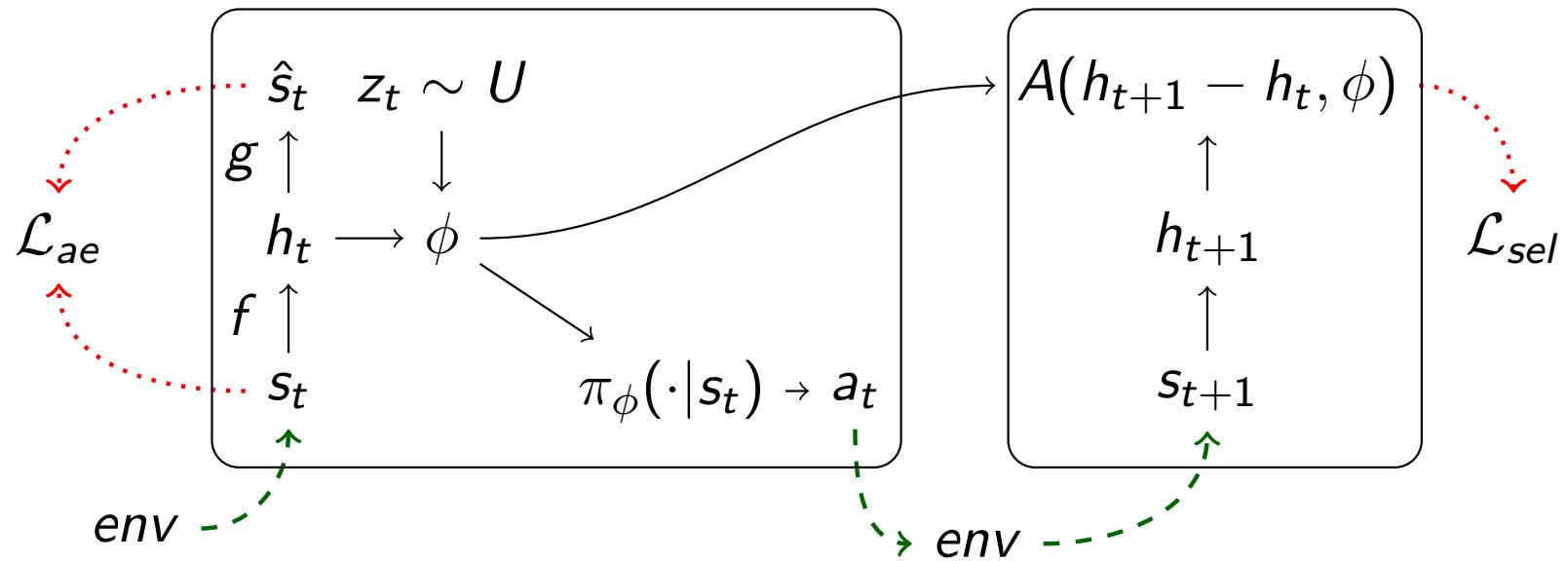
$\phi = G(h, z) \in \mathbb{R}^n$ is now **generated** from $h = f(s)$, $z \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}_{(s_t, a_t, s_{t+1})} \mathbb{E}_\phi \left[\pi_\phi(a_t | s_t) \frac{A(f(s_{t+1}) - f(s_t), \phi)}{\mathbb{E}_{\phi' = G(h_t, z')} [|A(f(s_{t+1}) - f(s_t), \phi')|]} \right]$$

How much the **value** of property ϕ changed relatively to other properties.

Continuous Attributes Model: Computational Graph

Total objective = autoencoder loss + selectivity loss



- Off-policy A2C/PG method on a batch of ϕ

What's Wrong with our Unsupervised Training Objectives?

They are in pixel
space rather than
abstract space

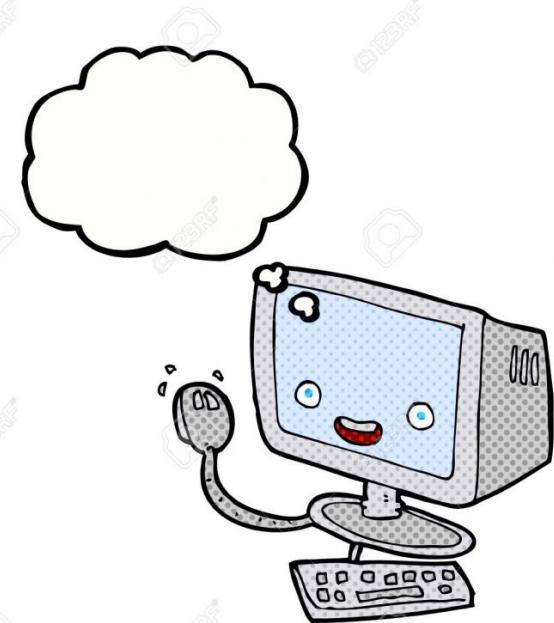
Abstraction Challenge for Unsupervised Learning

- Why is modeling $P(\text{acoustics})$ so much worse than modeling $P(\text{acoustics} \mid \text{phonemes}) P(\text{phonemes})$?
- Why are our current models not able to figure out phonemes AND model their distribution separately?
- May have to do with the different time scales and objective function at the wrong level of abstraction:
 - **log-likelihood focuses most of its value on the vast majority of bits characterizing the acoustic details (instead of the higher-level linguistic structure)**
 - **it would be good to just predict the future in abstract space rather than in the pixel space**

The Consciousness Prior

Bengio 2017, arXiv:1709.08568

- Conscious thoughts are very low-dimensional objects compared to the full state of the (unconscious) brain
- Yet they have unexpected predictive value or usefulness
 - strong constraint or prior on the underlying representation
- **Thought**: composition of few selected factors / concepts at the highest level of abstraction of our brain
- Richer than but closely associated with short verbal expression such as a **sentence** or phrase, a **rule** or **fact** (link to classical symbolic AI & knowledge representation)



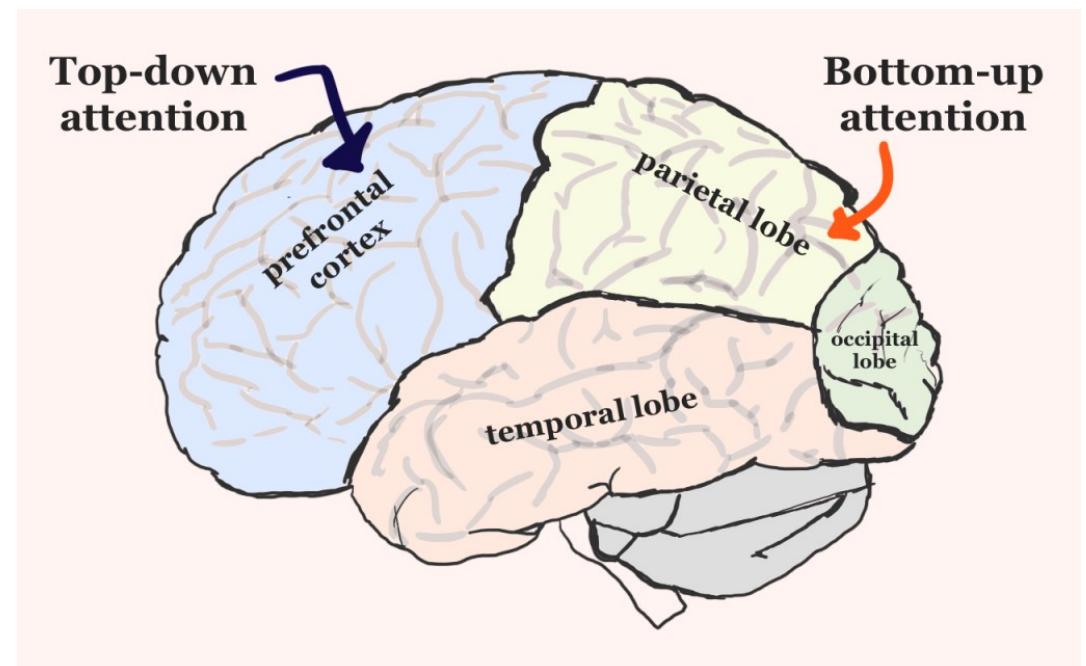
How to select a few
relevant abstract
concepts in a
thought?

Content-based
Attention

On the Relation between Abstraction and Attention

- Attention allows to focus on a few elements out of a large set
- Soft-attention allows this process to be trainable with gradient-based optimization and backprop

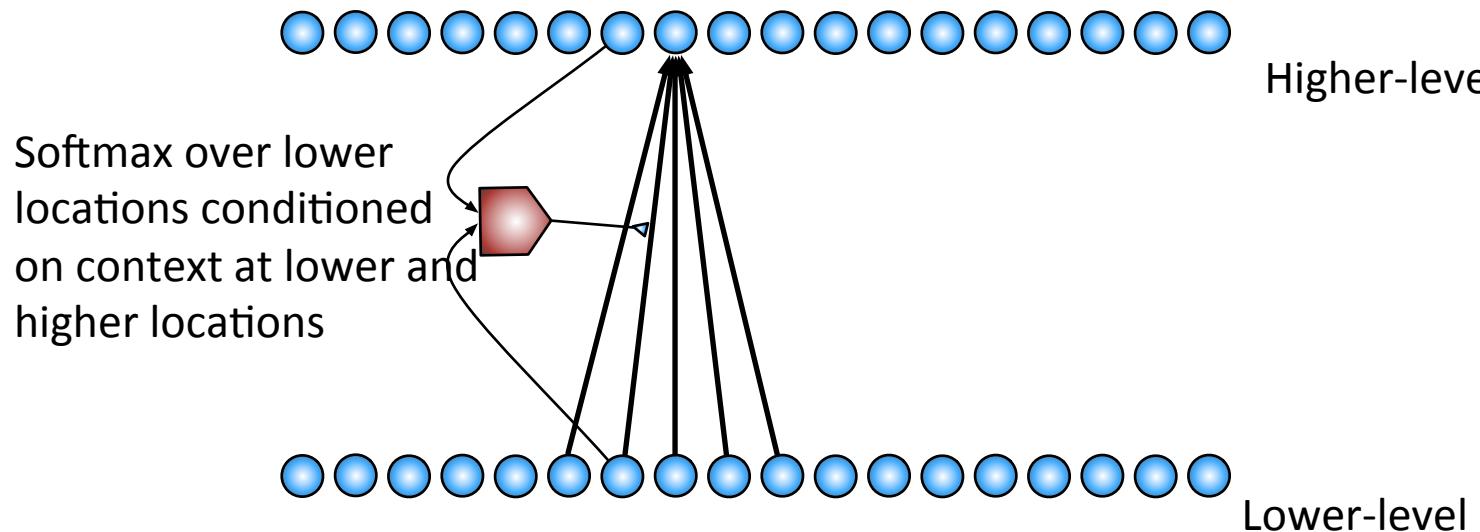
Attention focuses on a few appropriate abstract or concrete elements of mental representation



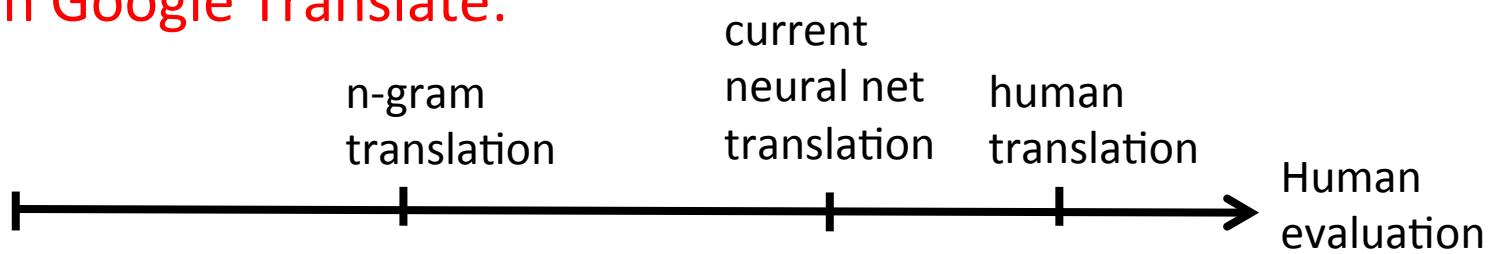
The Attention Revolution in Deep Learning

- **Attention mechanisms exploit GATING units**, have unlocked a breakthrough in machine translation:

Neural Machine Translation (ICLR'2015)



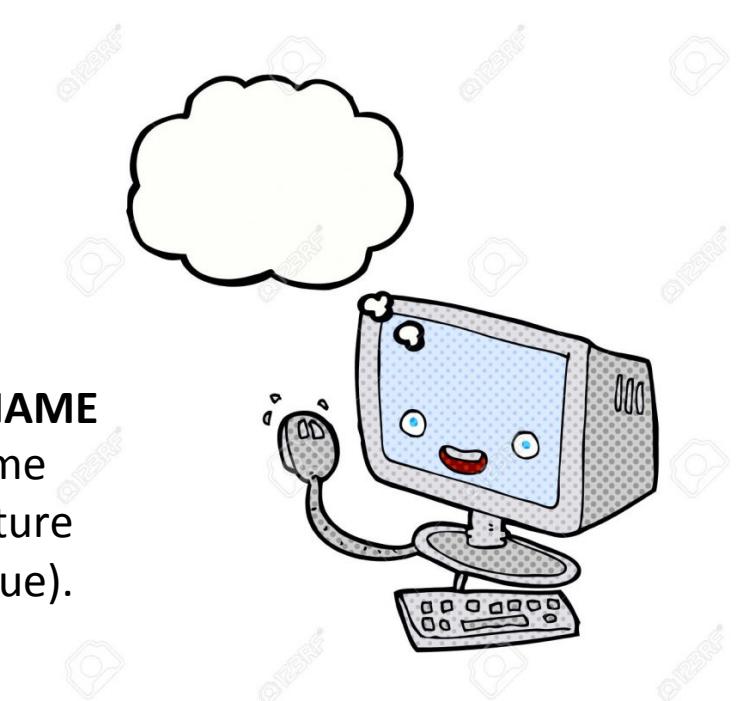
- Now in Google Translate:



The Consciousness Prior

Bengio 2017, arXiv:1709.08568

- 2 levels of representation:
 - High-dimensional abstract representation space (all known concepts and factors) h
 - Low-dimensional conscious thought c , extracted from h
- Example: c is a prediction about some future event, involves current variables and their values, and a prediction about a future variable
- Predictor needs to **refer to a predicted variable by NAME** (e.g. embedding) so as to be able to separate the name from the value and recover the prediction when a future event makes the variable observed (at a different value).



The Consciousness Prior

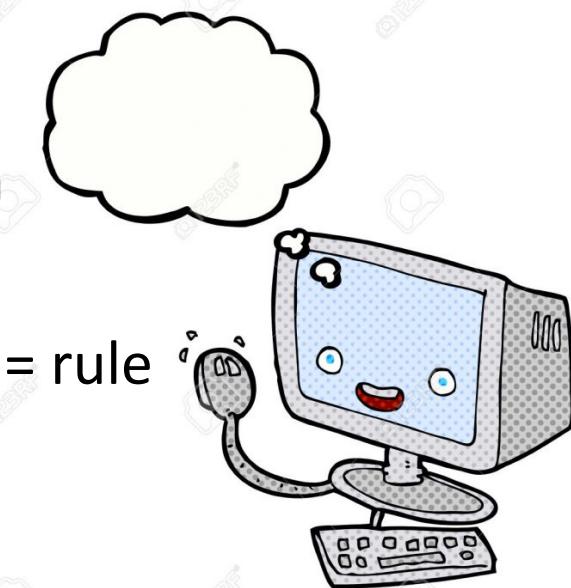
Bengio 2017, arXiv:1709.08568

- Conscious prediction over attended variables A (soft attention)

$$V = - \sum_A w_A \log P(h_{t,A} = a | c_{t-1})$$

↑
Attention weights ↑
Predicted value ↑
Earlier conscious state

- How to train the attention mechanism which selects which variables to predict?
- (predicted variables, conditioning variables) = rule
- Connection to classical symbolic AI



Representing Agents, Objects, Policies

The controllable factors idea works on a small scale: ongoing work to expand this to more complex environments

- Expand to sequences of actions and use them to define options
- Notion of objects and attributes naturally falls out
- Extension to non-static set of objects: types & instances
- Objects are groups of controllable features: by whom? Agents
 - Factors controlled by other agents; mirror neurons
- Because the set of objects may be unbounded, we need to learn to represent policies themselves, and the definition of an object is bound to the policies associated with it (for using it and changing its attributes)

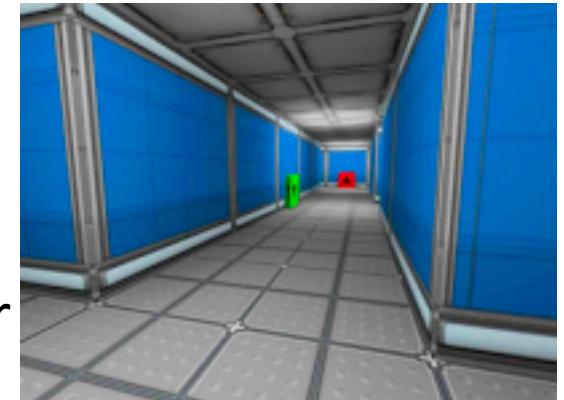
Research Program Towards Higher-Level Deep Learning

- Build a series of gradually more complex environments
- Design training frameworks for learning agents to discover the notion of objects, attributes, types, other agents, relations, the combinatorial space of policies, etc…
- Associate all these concepts to linguistic constructions: learn semantically grounded linguistic expressions.
 - Words are clues of higher-level abstractions
 - Language is an input/output modality

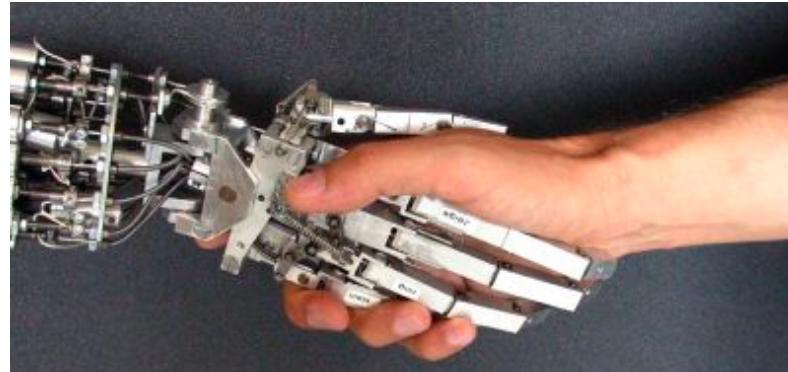
Ongoing Research: DL for AI neural nets → cognition



- Learn more abstract representations which capture causality
- **Independently controllable factors:** some abstract factors are controllable aspects of the environment, **disentangled**
- **Jointly learn conditional exploratory policies with intrinsic rewards + associated factors**
- Naturally gives rise to the notion of **objects, attributes & agents**
- Natural language & consciousness prior: other clue about abstract representations
- Unsupervised RL research, performed in simulated environments



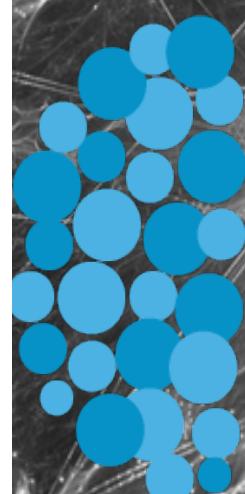
The Future of Deep AI



- Scientific progress is slow and continuous, but social and economic impact can be disruptive
- Many fundamental research questions are in front of us, with much uncertainty about when we will crack them, **but we will**
- Importance of continued investment in basic & exploratory AI research, for both practical (recruitment) short-term and long-term reasons
- Let us continue to keep the field open and fluid, be mindful of social impacts, and make sure AI will bloom **for the benefit of all**



Montreal Institute for Learning Algorithms



MILA

Université de Montréal