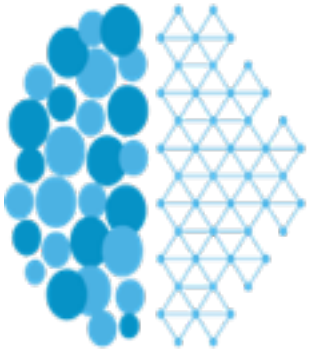
Deep Learning and Cognition

Yoshua Bengio

ReWork Deep Learning Summit

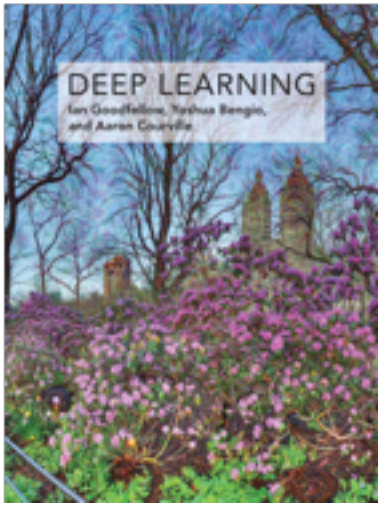
Montreal, October 10th, 2017



MILA

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

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



Université de Montréal

IVADO

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

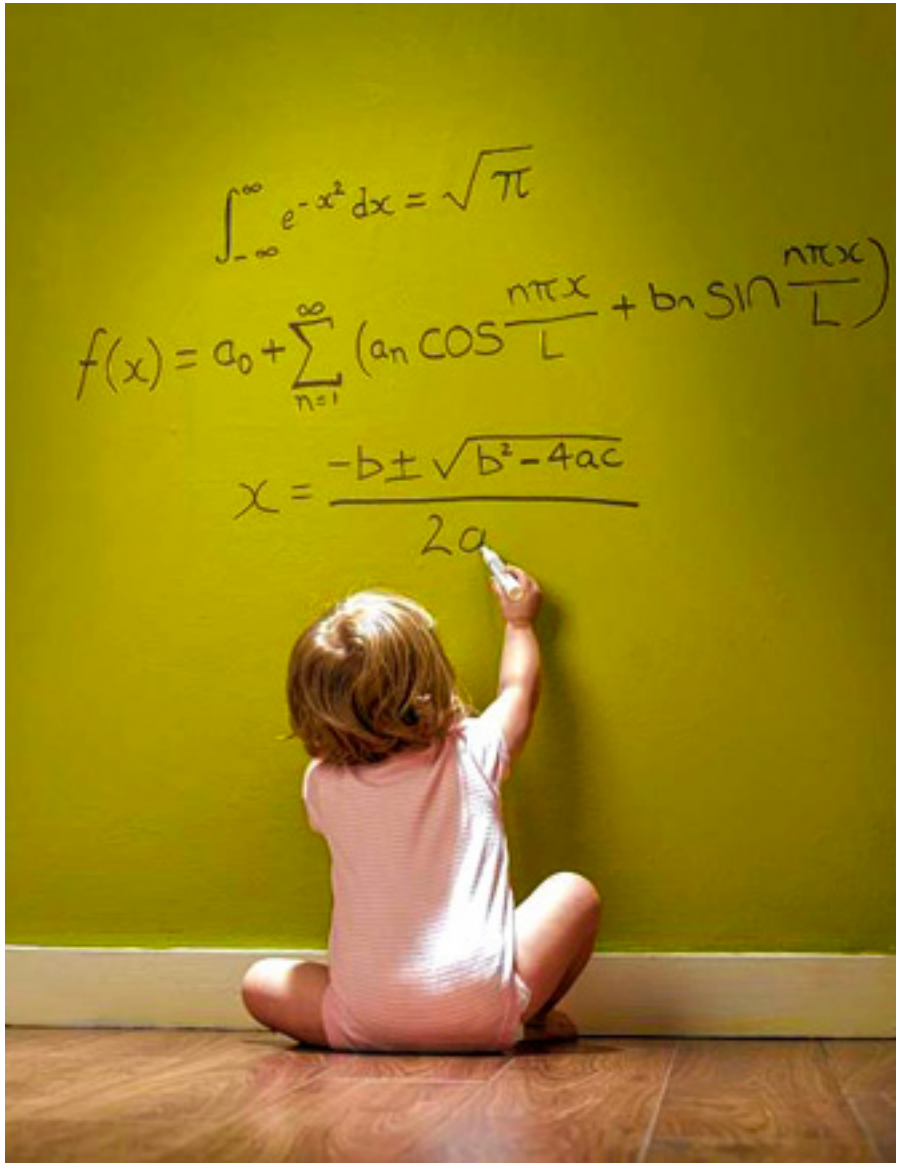
What's Missing?

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.

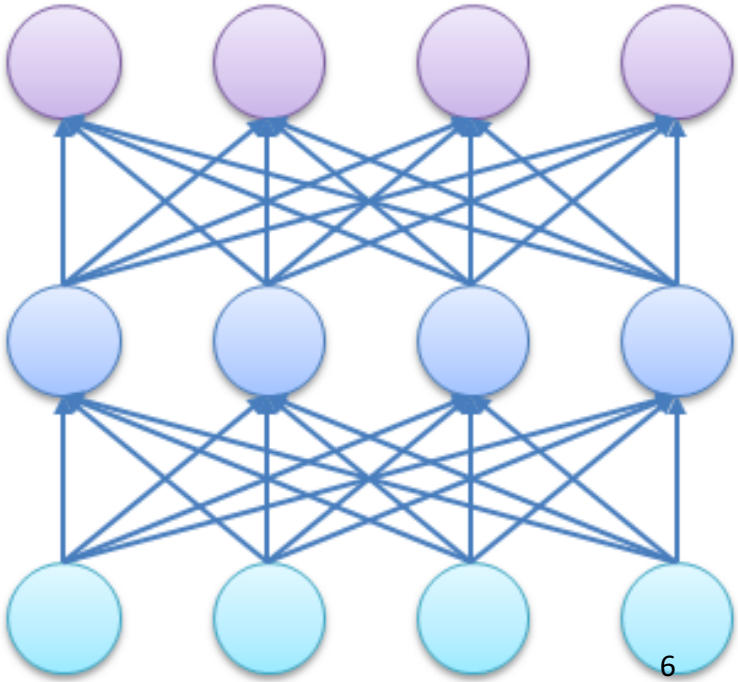


What's Missing?

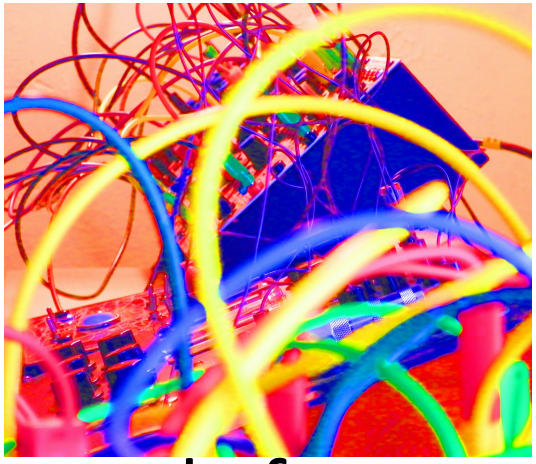
Abstract Representations

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*



Invariance and Disentangling



- Invariant features
- Which invariances?
- Alternative: learning to disentangle factors
- Good disentangling →
  avoid the curse of dimensionality,
  easy to select relevant factors →
  fast adaptation to new tasks, 0-shot learning

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

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, $\phi \in \{1, .., N\}$, define *selectivity*:

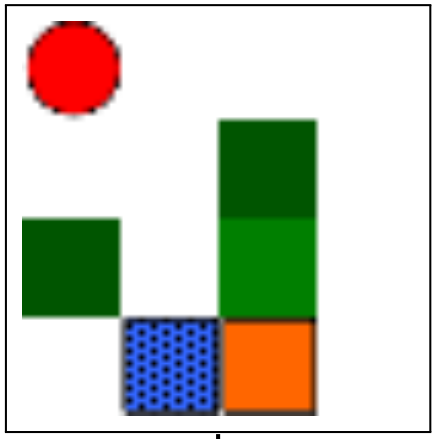
$$\sum_{k=1}^{N} \mathbb{E}_{(s_t, a_t, s_{t+1})} \left[\pi_k(a_t|s_t) \frac{f_k(s_{t+1}) - f_k(s_t)}{\sum_{k'} |f_{k'}(s_{t+1}) - f_{k'}(s_t)|} \right]$$

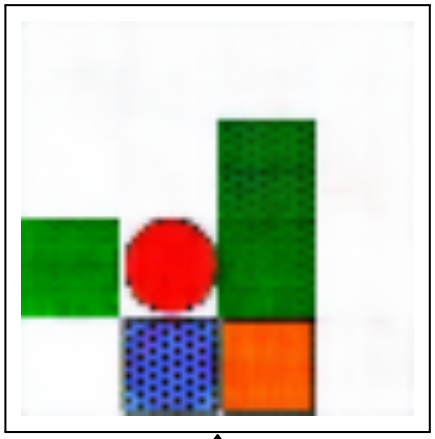
- Optimize both policy π_k and representation f_k to minimize

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

Predict the effect of actions in attribute space

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



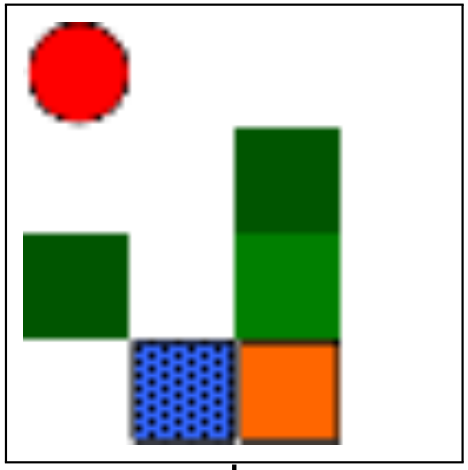


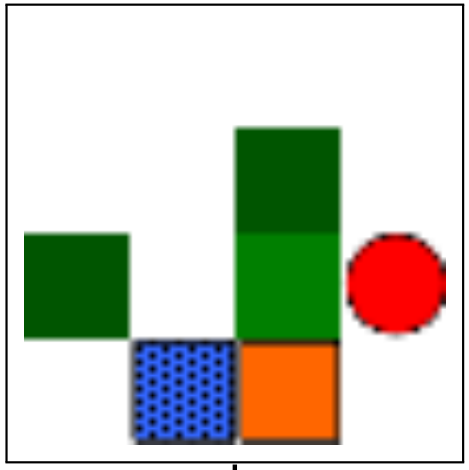
Encoder

Decoder

$$\underbrace{h}_{(0.4,\ 13.1)} \longrightarrow \underbrace{\hat{h}'}_{(-4.6,\ -1.9)} = h + \underbrace{dh_{right}}_{(5,\ -5)} + \underbrace{2 \cdot dh_{down}}_{(-10,\ -10)}$$

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





Encoder

Encoder

h_1

h_2

$(0.4,\ 13.1)$

$(5.9,\ -11.6)$

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

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.

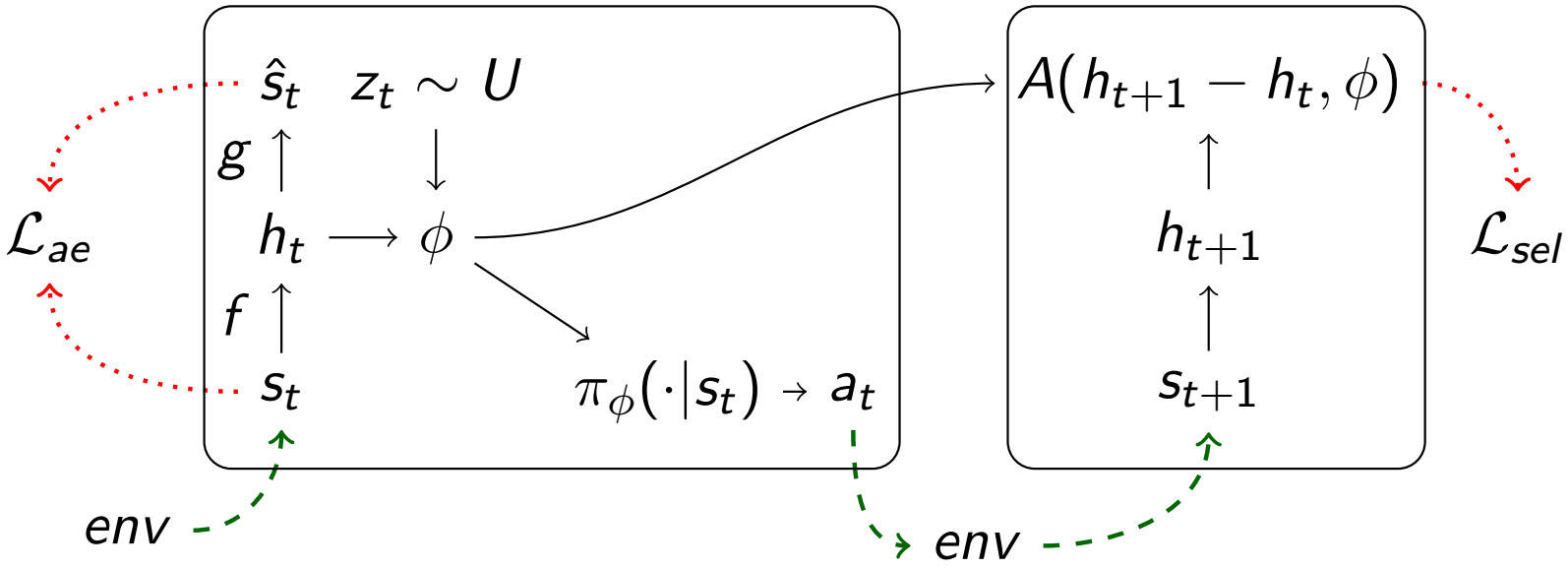
$\phi = G(h, z) \in \mathbb{R}^n$ is now **generated** from $h = f(s)$, $z \sim \mathcal{N}(0, 1)$:

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

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(acoustics) so much worse than modeling P(acoustics | phonemes) P(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 in abstract space rather than in the pixel space**

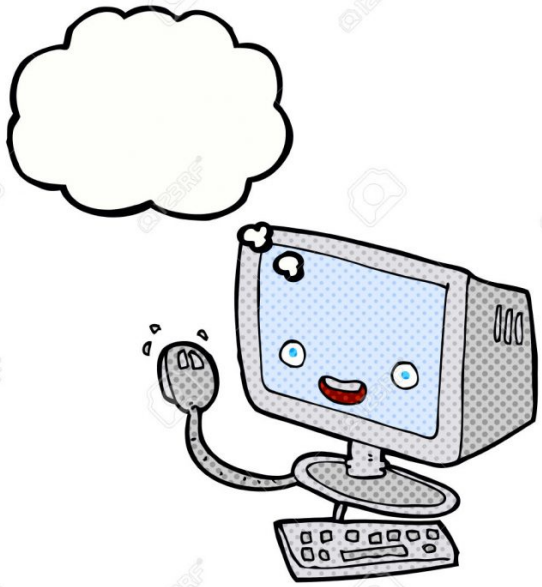
The Consciousness Prior

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)



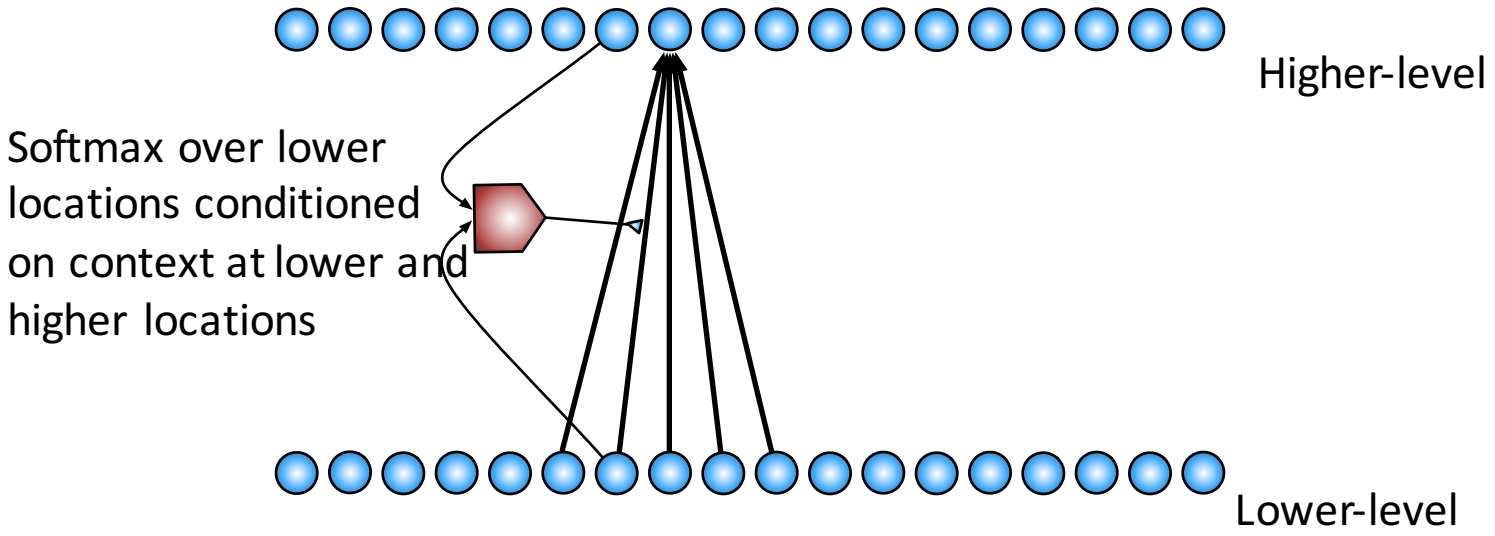
How to select a few relevant abstract concepts in a thought?

Content-based Attention

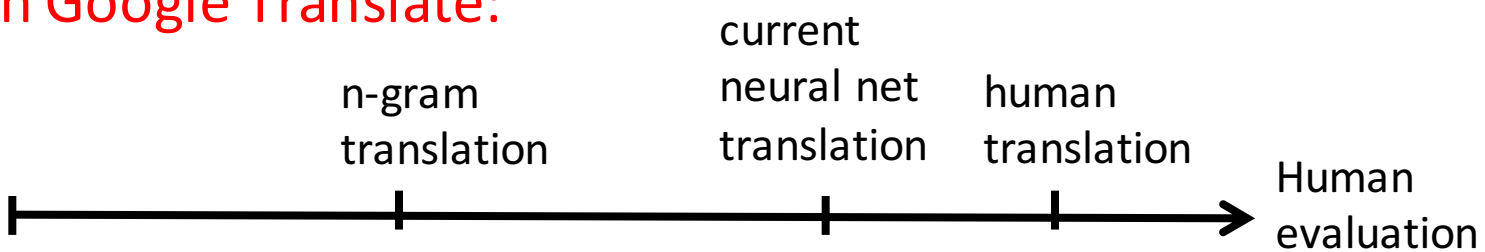
The Attention Revolution in Deep Learning

- **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)

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

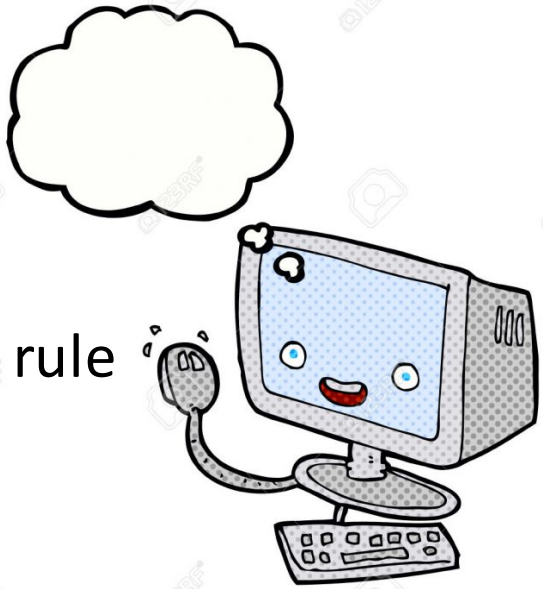
Attention weights

Predicted value

Earlier conscious state

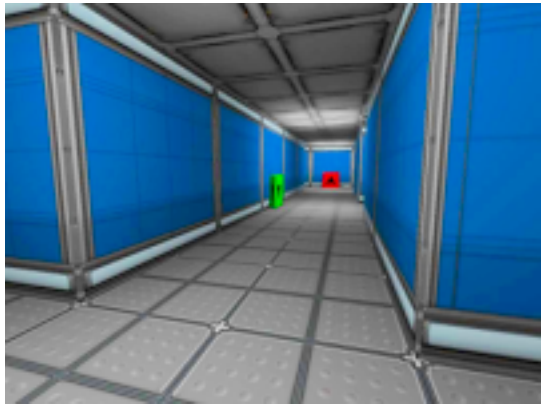
- How to train the attention mechanism which selects which variables to predict?
- (predicted variables, conditioning variables) = rule

Connection to classical symbolic AI



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**

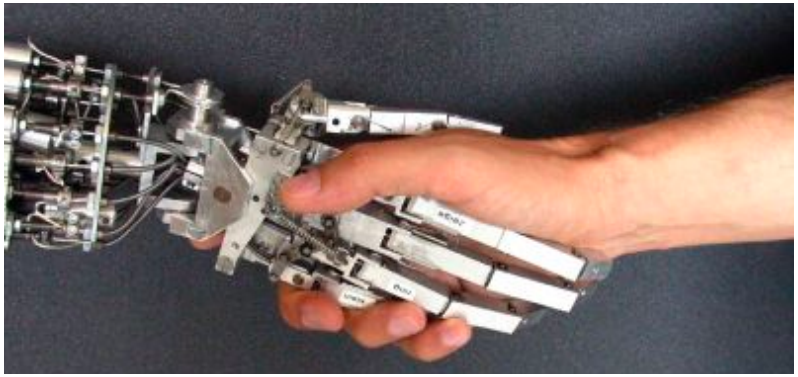


- Naturally gives rise to the notion of

  objects, **attributes** & **agents**

- Natural language & consciousness prior: other clue about abstract representations
- Unsupervised RL research, performed in simulated environments

The Future of Deep AI

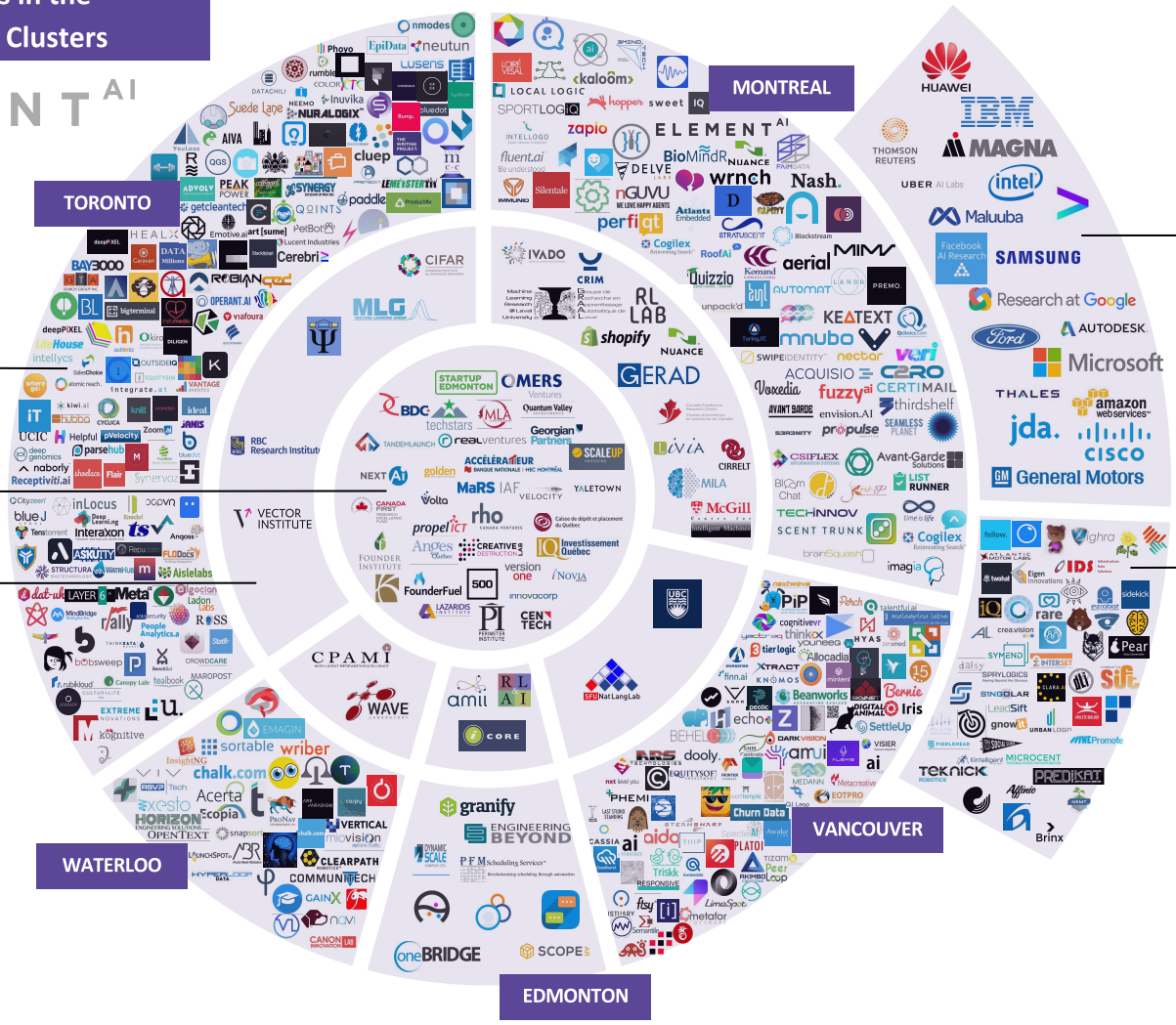


- Scientific progress is slow and continuous, but social and economic impact can be disruptive
- Many fundamental research questions are in front of us, with much uncertainty about when we will crack them, **but we will**
- Importance of continued investment in basic & exploratory AI research, for both 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

The Canadian AI Ecosystem

Top Players in the Canadian AI Clusters

ELEMENT AI



Startups & Enterprises

Incubators, accelerators & VC (Pan-Canadian)

Research Labs

International players in Canada (Pan-Canadian)

Startups & Enterprises (Outside of cluster cities)

STARTUP & ENTERPRISE COUNT PER LOCATION

195+	TORONTO
100+	VANCOUVER
90+	MONTREAL
50+	WATERLOO-KITCHENER
10+	EDMONTON
60+	ALL OTHERS

2017-06-11
For updates see jfgagne.ai

Montreal Institute for Learning Algorithms

MILA

Université de Montréal