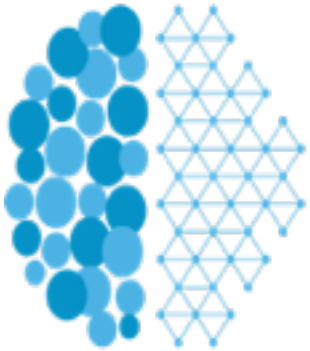
Bridging the gap between brains, cognition and deep learning

Yoshua Bengio, November 19th, 2017

Montreal AI & Neuroscience Workshop



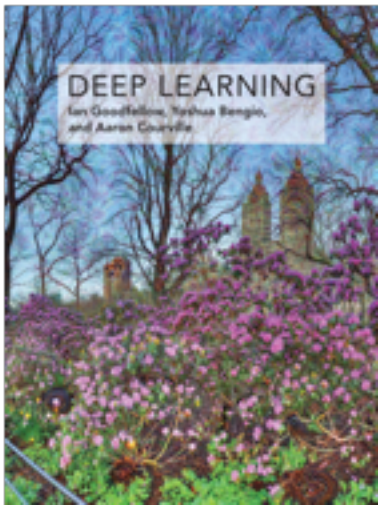
MILA

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

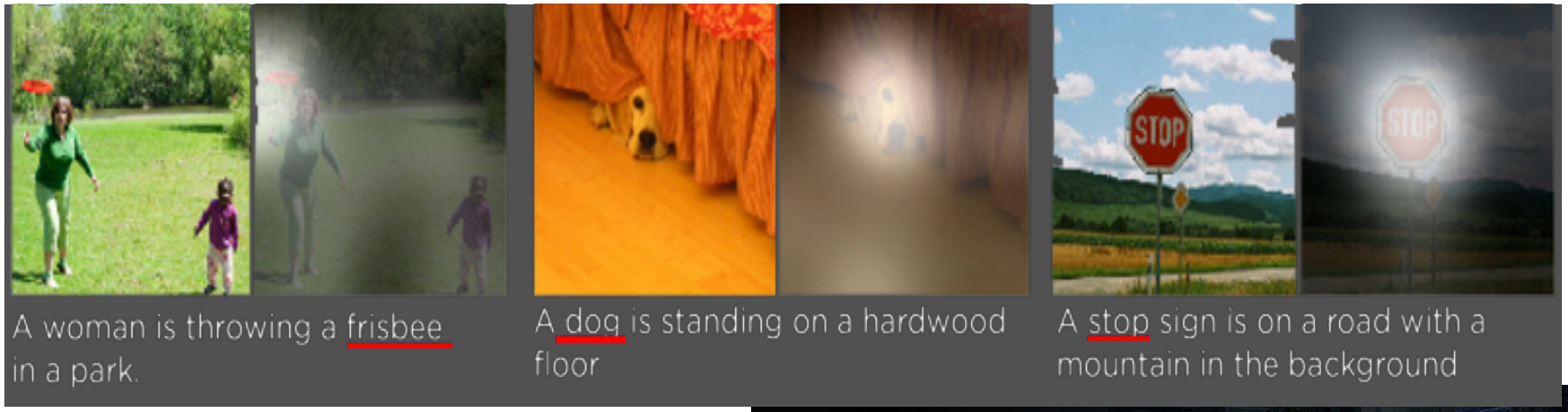
Université de Montréal

IVADO

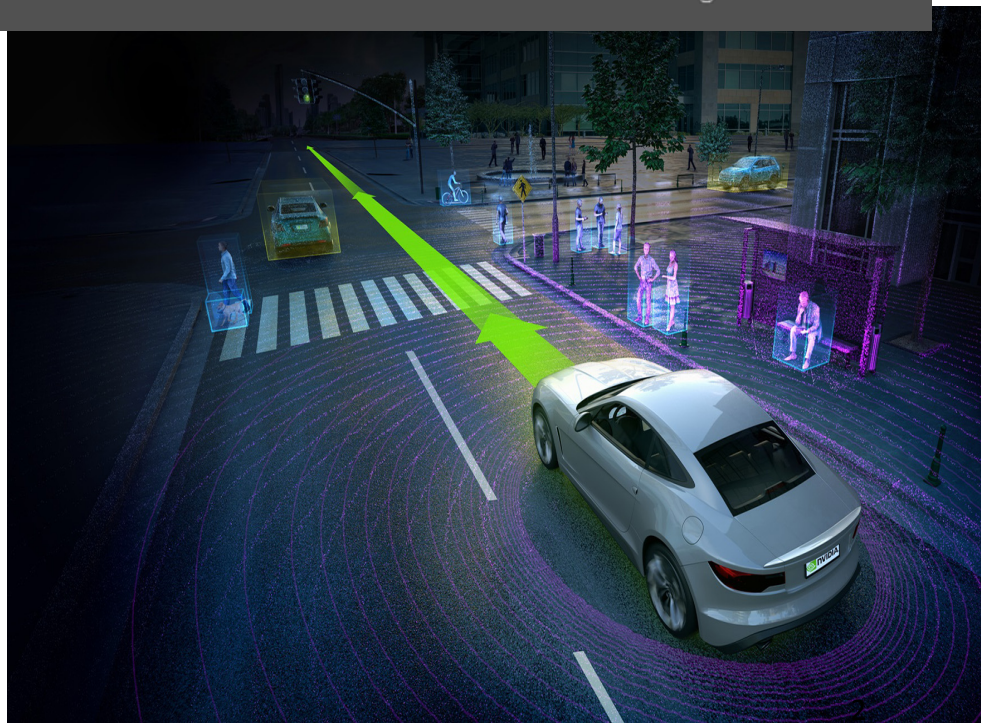
PLUG: **Deep Learning**, MIT Press book is out, chapters will remain online



Deep Learning → AI Breakthroughs



Computers have made huge strides in perception, manipulating language, playing games, reasoning, ...



What Deep Learning Owes to Connectionism from the 80's

Iteratively learning distributed representations through a composition of neurally inspired simple operations towards a justifiable training objective that forces the learner to capture the relevant statistical structure of the data.

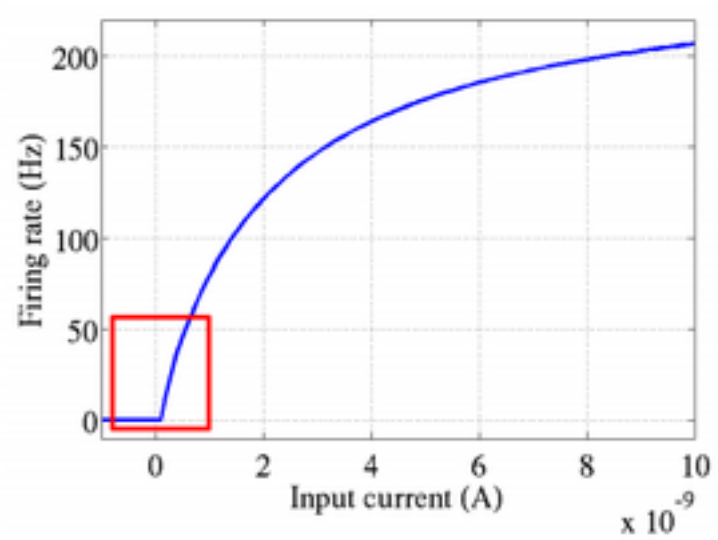
Canada's lead in deep learning

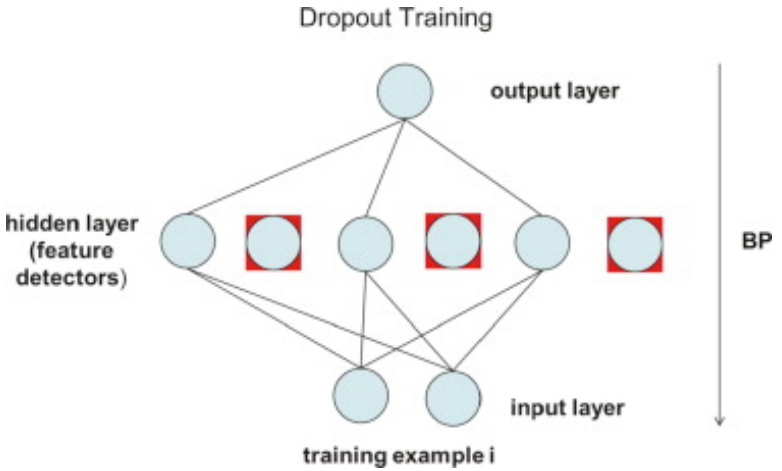
Thanks to investments in basic research at a time when the topic was not fashionable

A Canadian-led trio initiated the deep learning AI revolution

What's New with Neural Nets in 21st Century?

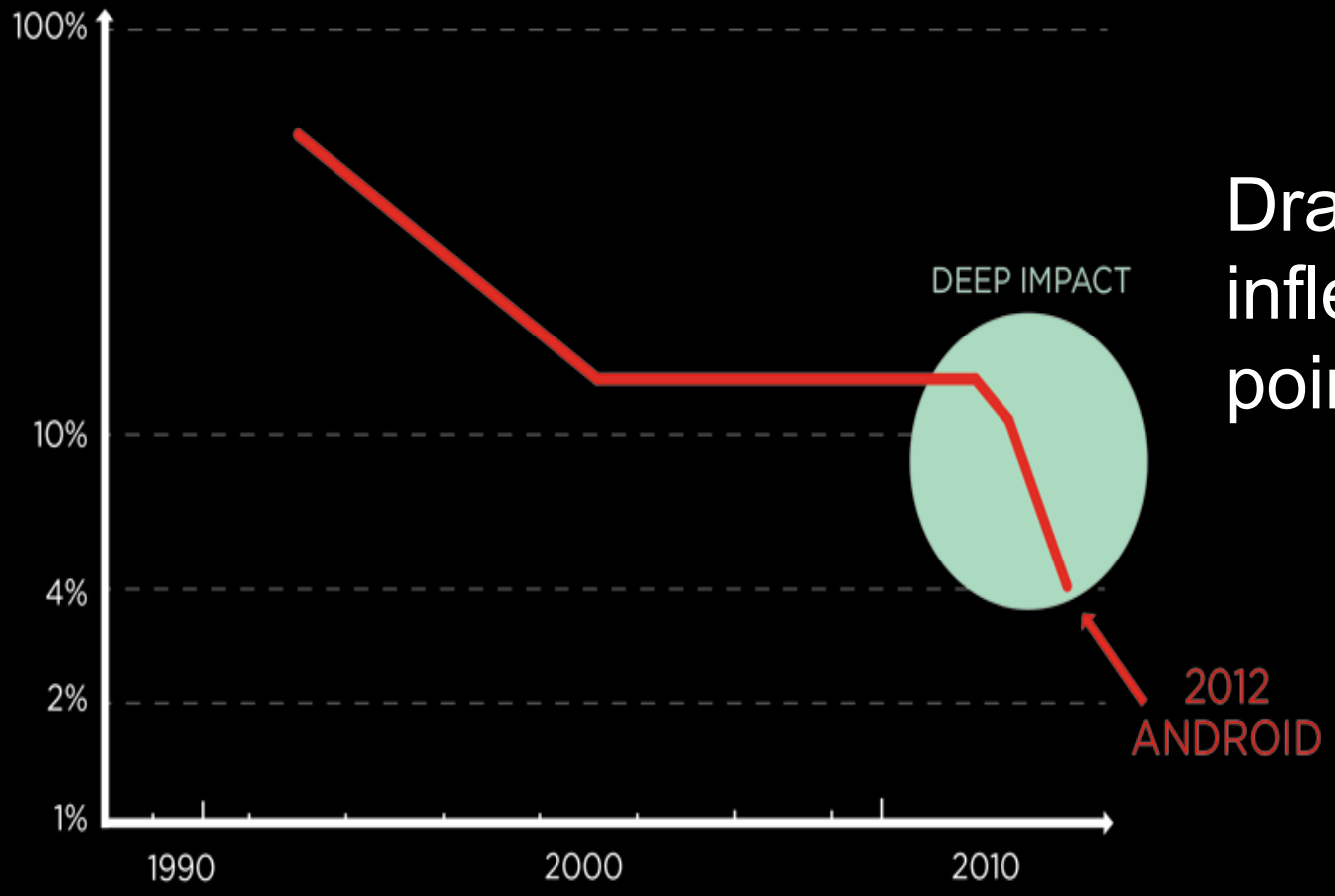
- Ability to train deeper nets!
- Biologically inspired ReLUs instead of sigmoids, enable training much deeper nets by backprop (Glorot & Bengio AISTATS 2011)





- Some forms of noise (e.g. **spiking-like** dropout) are powerful regularizers yielding superior generalization abilities
- Attention!
- Generative neural networks, deep reinforcement learning

2010-2012: breakthrough in speech recognition



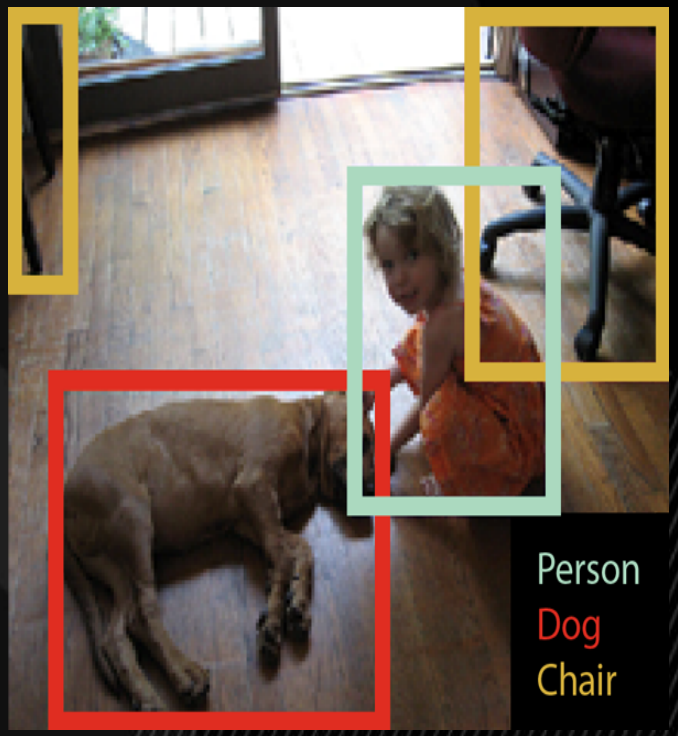
Dramatic inflection point

Source: Microsoft

2012-2015: breakthrough in computer vision

2015: ***human-level performance***

Ability to process unstructured data: text, images, signals, web

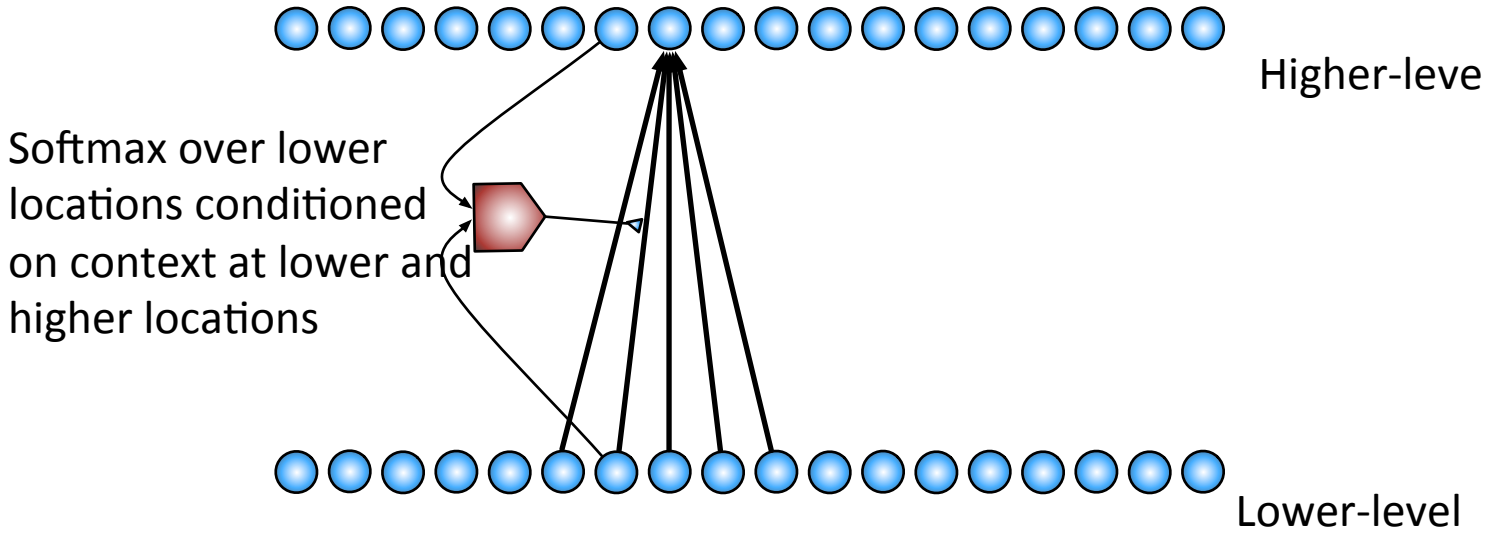



March 2016:
World Go Champion
Beaten by Machine

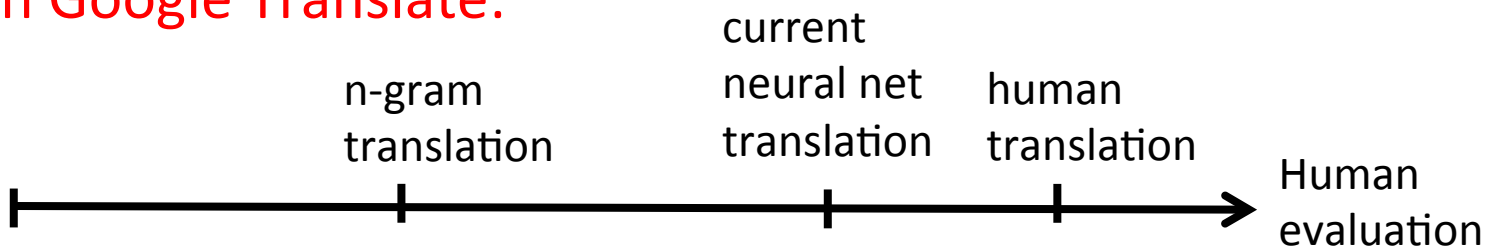
The Attention Revolution in Deep Learning

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

Neural Machine Translation (ICLR'2015)



- Now in Google Translate:



Still Far from Human-Level AI

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

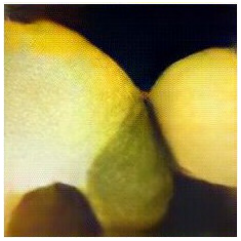
- Learning superficial clues, not generalizing well enough outside of training contexts, easy to fool trained networks:
  - Current models cheat by picking on surface regularities

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)

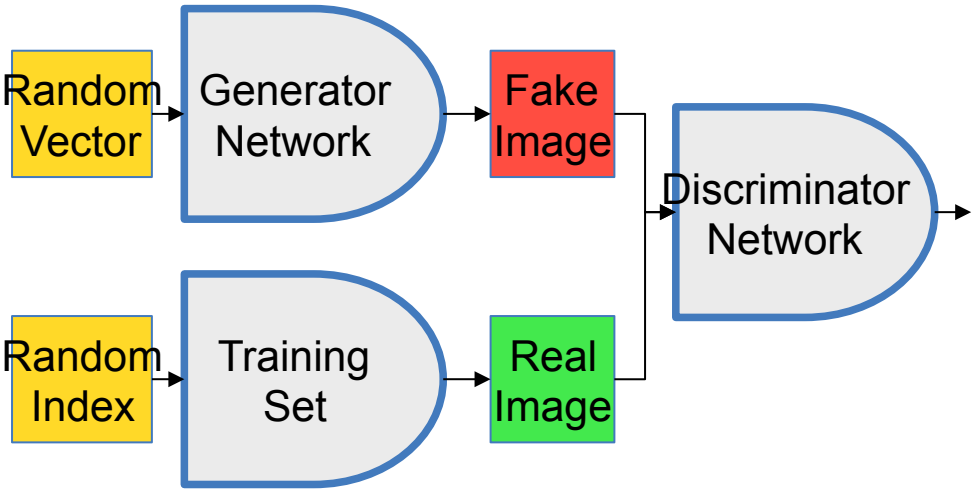
- It would be even worse with current RL (statistical inefficiency)
- Long-term objective: develop better predictive models of the world able to **generalize in completely unseen scenarios**, but it does not seem reasonable to model the sequence of future states in all their details
- Human drivers: no need to die a thousand deaths

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



Predict a multi-modal future

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



(Karras et al 2017)

(Nguyen et al 2016)

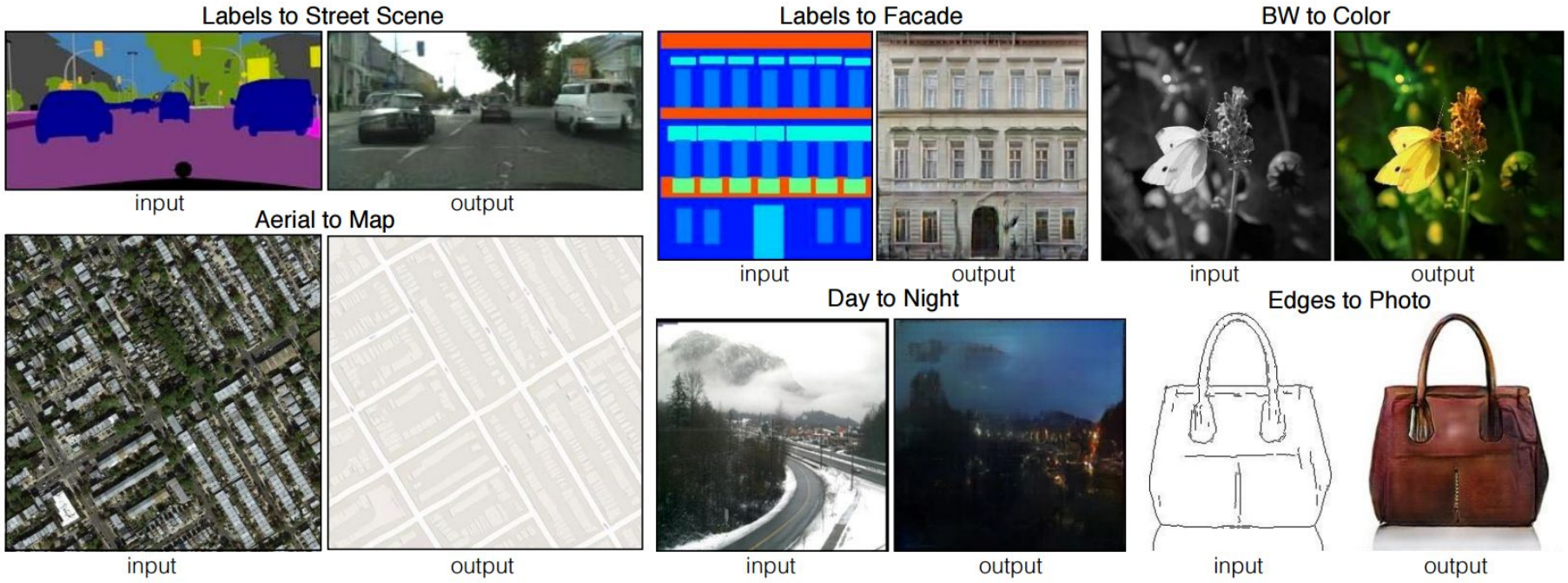
Text 2 Image, Colorization

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

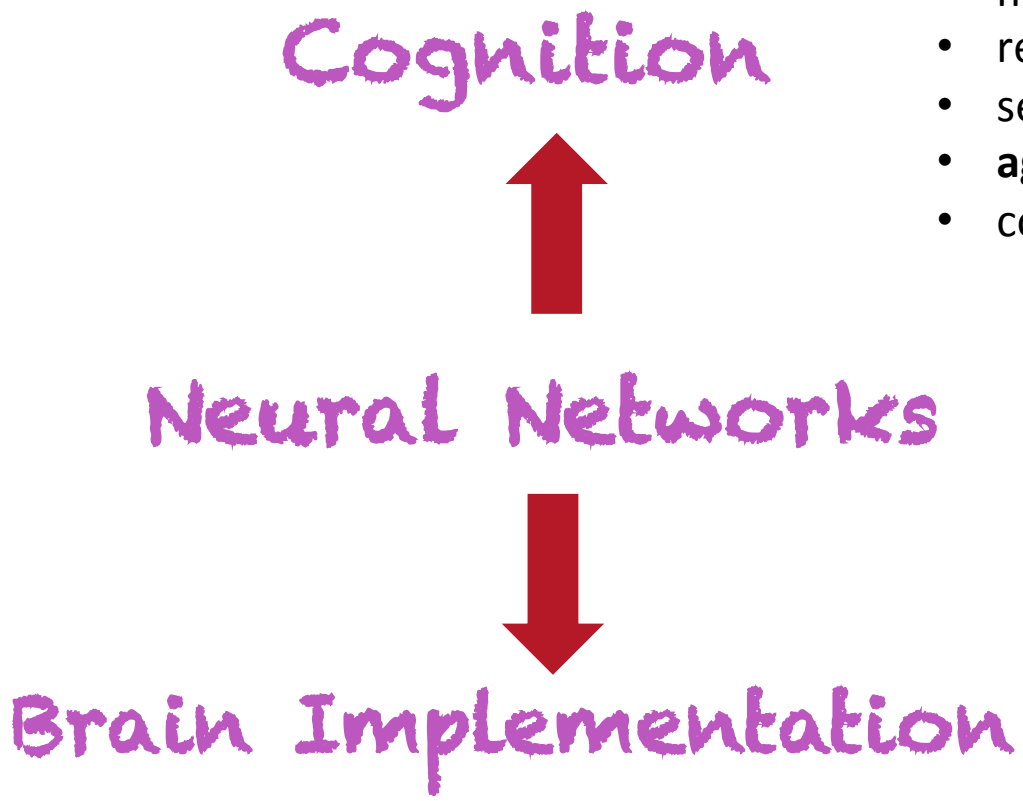
Can the conceptual advances behind deep learning successes help us figure out the big picture of how the brain learns complex behavior?

The Learning Mechanism is a Compact and Abstract Explanation of the Brain

Similar to the laws of physics: e.g. we consider **understanding** the physical world, mostly by having figured out the laws of physics, not just by describing its consequences (the immense complexity of describing the physical world)

Successful learning framework (architecture, optimizer, objective) is a compact abstract explanation, much more so than the actual detailed neuron-by-neuron functions performed by a trained brain

ML validation: can learn complex tasks
Neuroscience validation: matches biology at some level



Attention:

- vectors → data structures
- memory access, one-shot memorization
- reasoning
- semantics & language
- **agency & causality**
- consciousness

- **biological backprop**
- dropout & spikes
- multi-module architecture

Deep Learning & Neuroscience: Still a Large Gap

- **Backprop** and the ability to jointly train multiple layers is the workhorse of current deep learning successes. END-TO-END TRAINING OF DEEP COMPUTATIONS ROCKS. **Backprop is the building block behind modern unsupervised (generative) learning and RL**. But has been deemed not biologically plausible.
  - how to propagate gradients? linear neurons? separate net?
  - what is the role of feedback connections? lateral connections?
  - How to **efficiently** train a stochastic continuous-time dynamical system wrt a **global** objective?
    - *Random perturbation-based methods do not scale, BP does beautifully*

From Deep Learning to Neuroscience
Propagation of Error Signals

Deep Learning

Backpropagation:
Requires a special computational path for the propagation of error derivatives backward in the network.

Neuroscience

Hypothesis: ↔ *no empirical evidence yet*
Error signals are encoded in ds/dt.
No need for a special computational path.

This idea was first proposed by Hinton & McClelland:
"Recirculation algorithm" (1987)

Equilibrium Propagation

(Scellier & Bengio 2017, Frontiers in Neuroscience)

Backpropagation

Free Phase

-network relaxes to fixed point
-read prediction at the outputs

Forward Pass

-read prediction at the outputs

Weakly Clamped Phase

-nudge outputs towards targets
-error signals (back)propagate
-network relaxes to new nearby fixed point

Backward Pass

-compare prediction/target
-compute error derivatives

requires:
-special computational circuit
-special kind of computation

Equilibrium Propagation Theorem

(Scellier & Bengio, Bridging the Gap Between Energy-Based Models and Backpropagation, *Frontiers in Neuroscience, 2017)*

- Gradient on the objective function (cost at equilibrium) can be estimated by a ONE-DIMENSIONAL finite-difference

$$\frac{d}{d\theta} J_\beta^\delta(\theta, \mathrm{v}) = \lim_{\xi \to 0} \frac{1}{\xi} \left(\frac{\partial F}{\partial \theta} \left(\theta, \beta + \xi\delta, s_{\theta,\mathrm{v}}^{\beta+\xi\delta}, \mathrm{v} \right) - \frac{\partial F}{\partial \theta} \left(\theta, \beta, s_{\theta,\mathrm{v}}^{\beta}, \mathrm{v} \right) \right)$$

Small nudging

Sufficient statistic after nudging

Sufficient statistic before nudging

Stochastic version:

$$\frac{d}{d\theta} \widetilde{J}_\beta^\delta(\theta, \mathrm{v}) = \lim_{\xi \to 0} \frac{1}{\xi} \left(\mathbb{E}_{\theta,\mathrm{v}}^{\beta+\xi\delta} \left[\frac{\partial F}{\partial \theta} (\theta, \beta + \xi\delta, s, \mathrm{v}) \right] - \mathbb{E}_{\theta,\mathrm{v}}^{\beta} \left[\frac{\partial F}{\partial \theta} (\theta, \beta, s, \mathrm{v}) \right] \right)$$

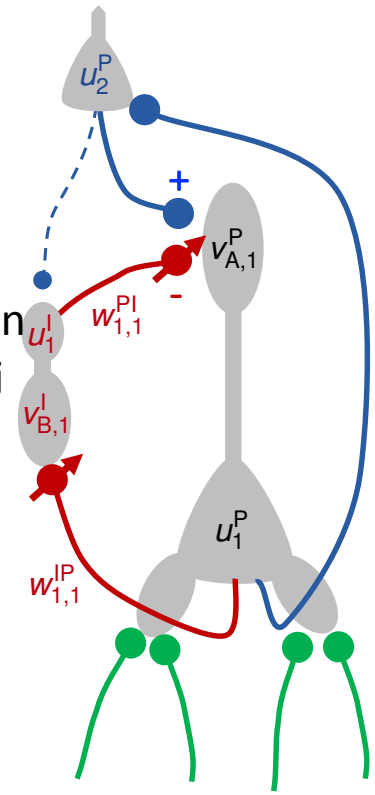
→ **Gives rise to Hebbian / anti-Hebbian updates with Hopfield net energy fn**

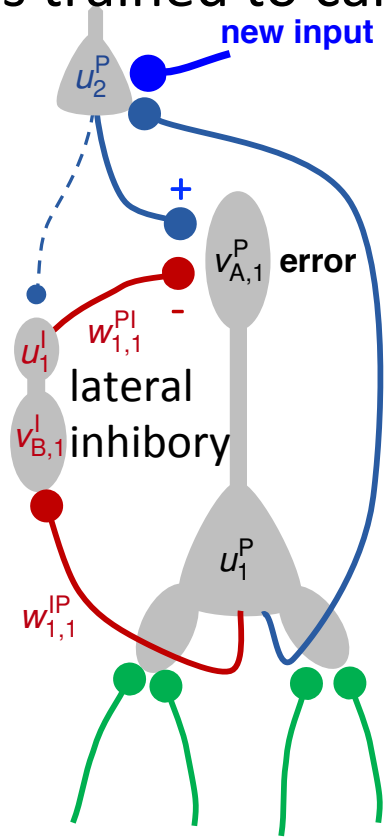
Lateral Feedback Interneurons May Solve the Linear Feedback Puzzle


with Walter Senn & Joao Sacramento

Building on *Urbanczik & Senn 2014*

- Lateral feedback via interneurons imitates feedforward path, their lateral projections trained to cancel top-down feedback





Basal dendrites: bottom-up

Apical dendrites: top-down feedback minus mirror unit's cancellation.

With no nudging, cancellation is perfect because next layer is predictable.

With nudging, difference = backprop error signal.

Other Ongoing Efforts

- Avoiding the constraint of symmetric weights
  - although it may be approximately enforced via the learning itself, and feedback alignment suggests that backprop would work nonetheless
- Avoiding the need to wait for convergence of the dynamics before making a weight update
- Biological tests!
  - SGD in the brain, neural nudging propagation, feedback-lateral cancellation

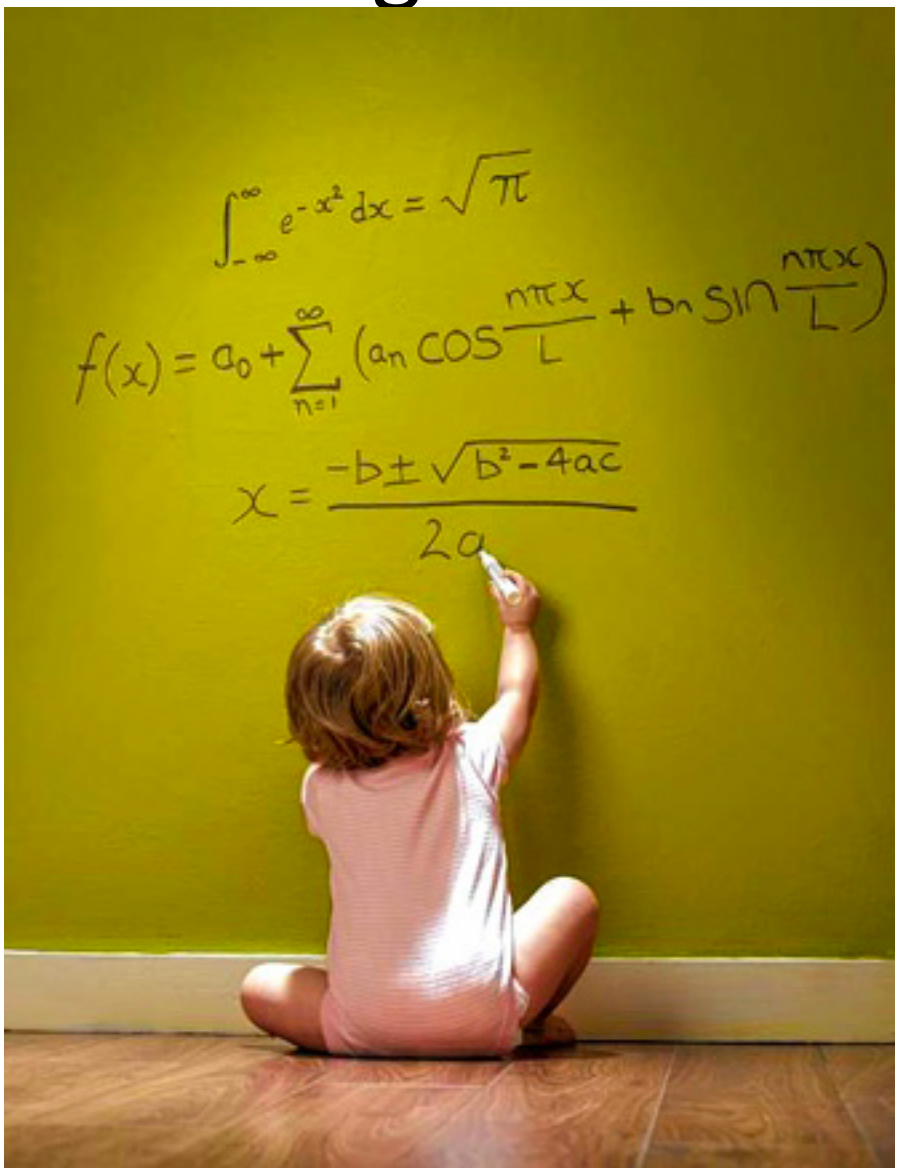
What's Missing with Deep Learning?

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.

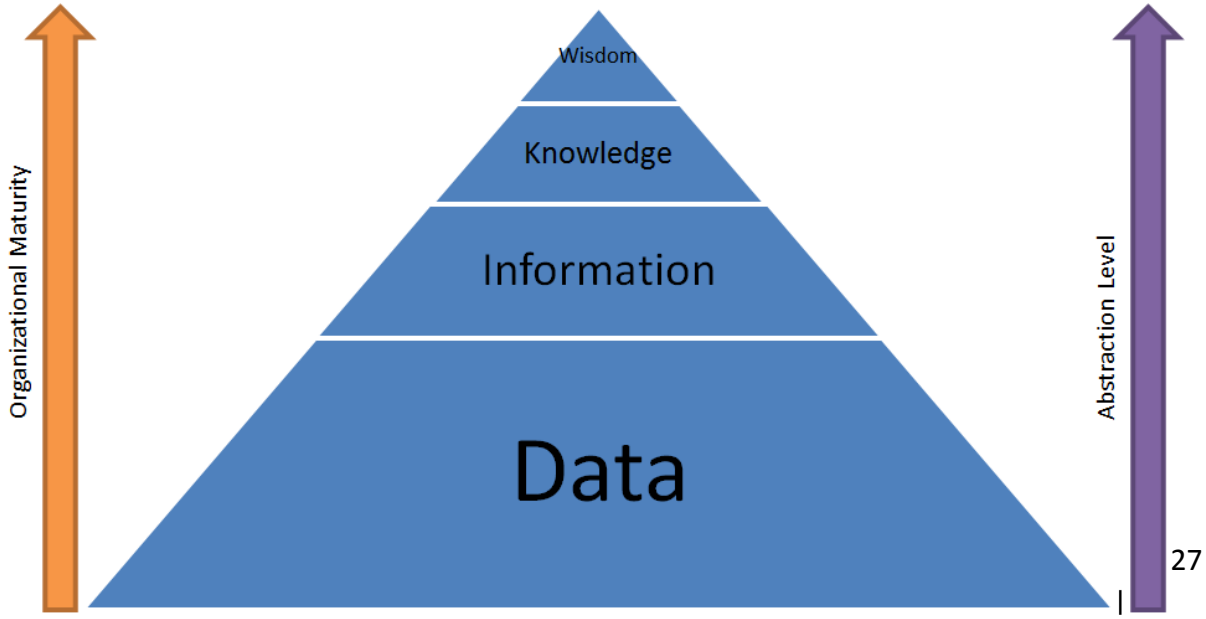


What's Missing?

Abstract Representations

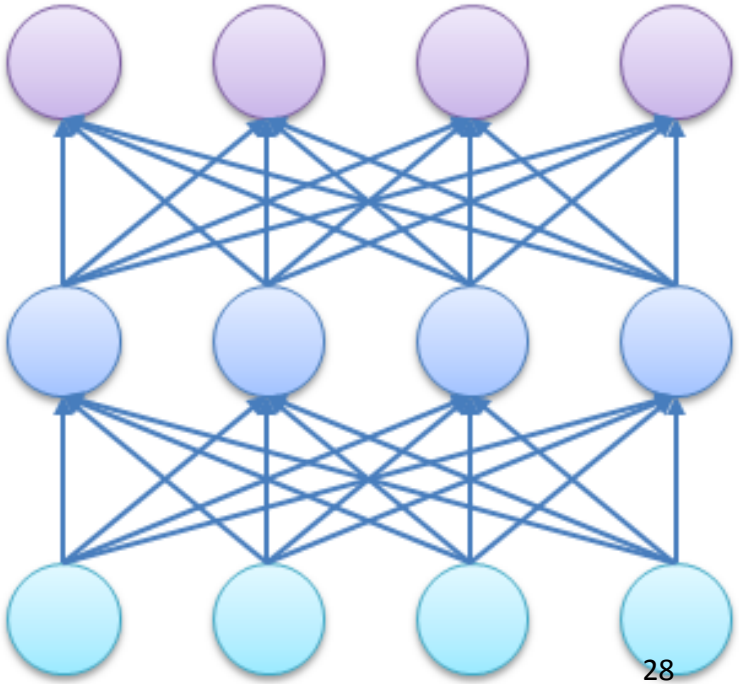
Learning Multiple Levels of Abstraction

- 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, $\phi \in \{1, .., N\}$, 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

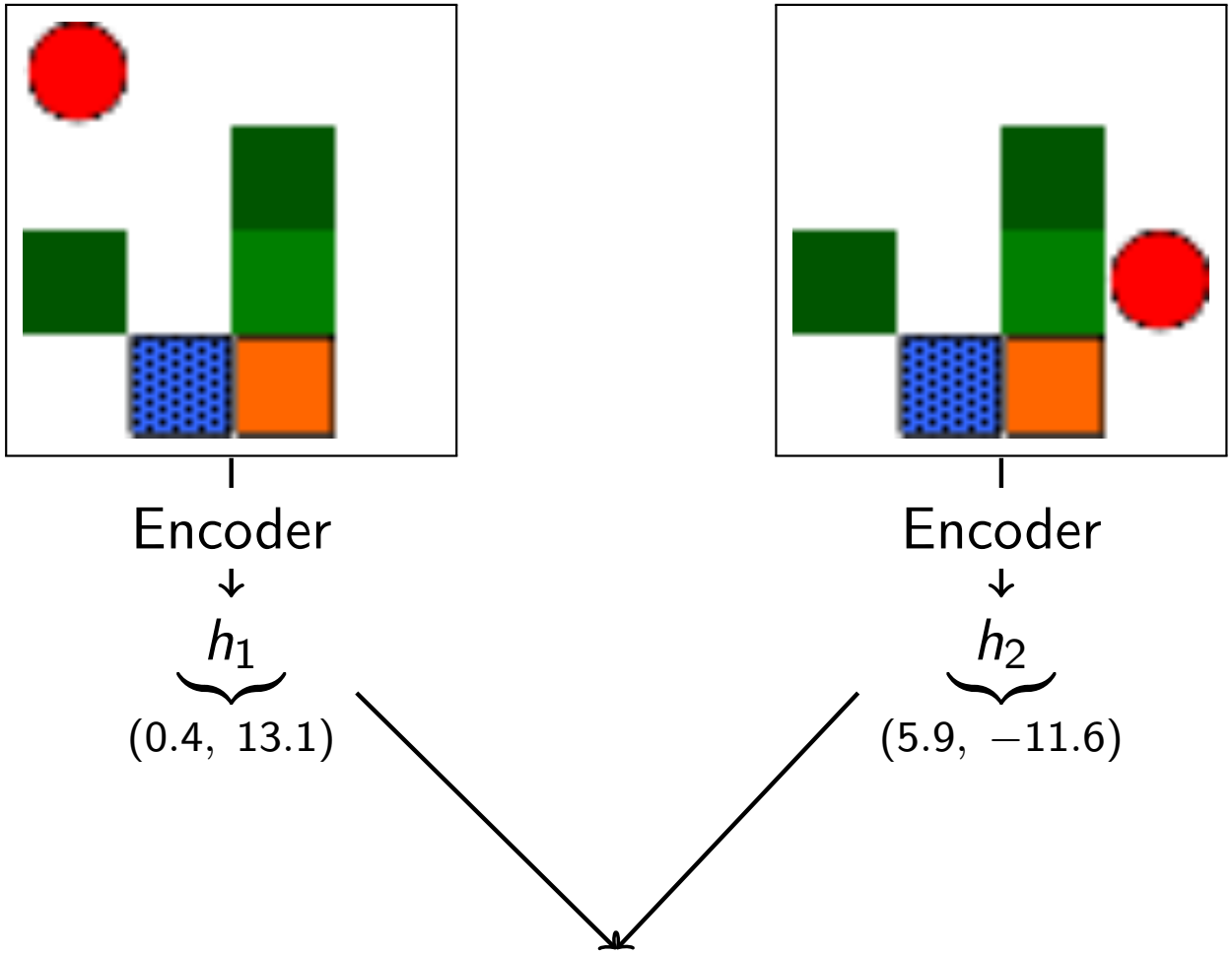
Given initial state and set of actions, predict new attribute values and the corresponding reconstructed images

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



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.

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**

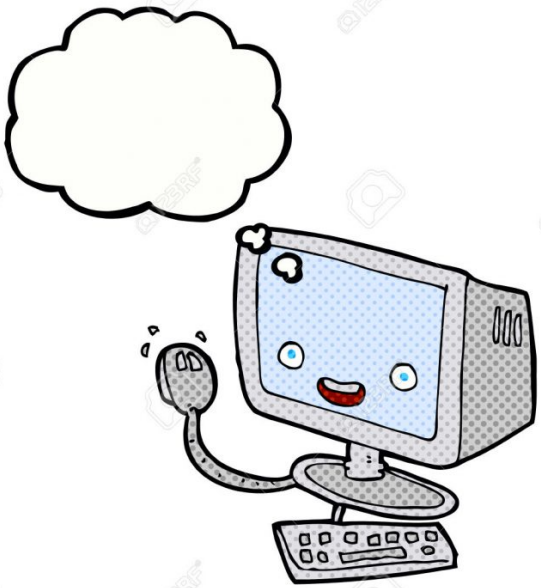
The Consciousness Prior

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)



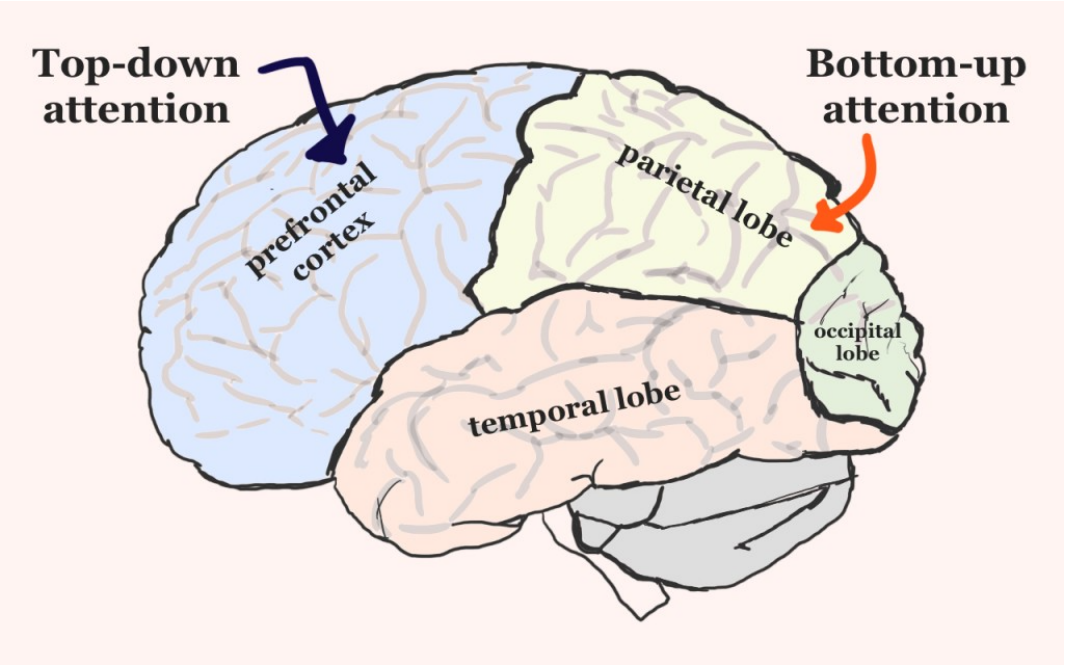
How to select a few relevant abstract concepts in a thought?

Content-based Attention

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



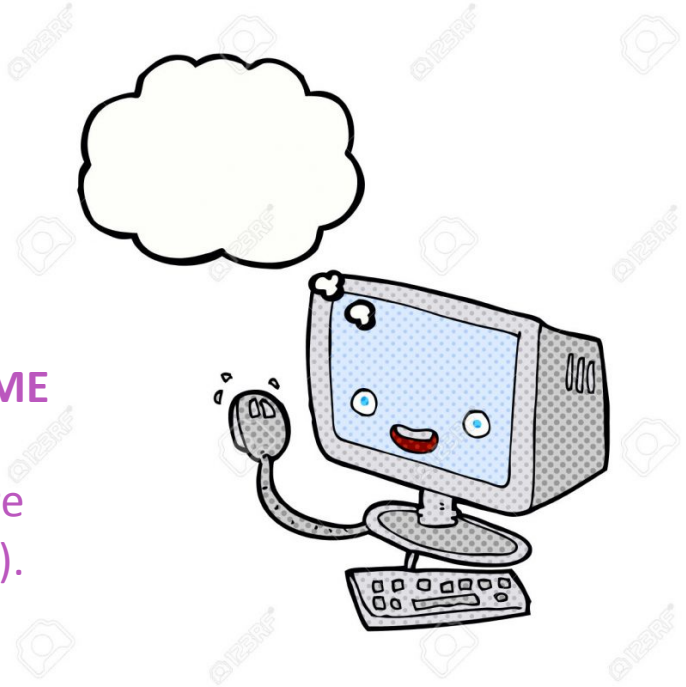
The Consciousness Prior

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)

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

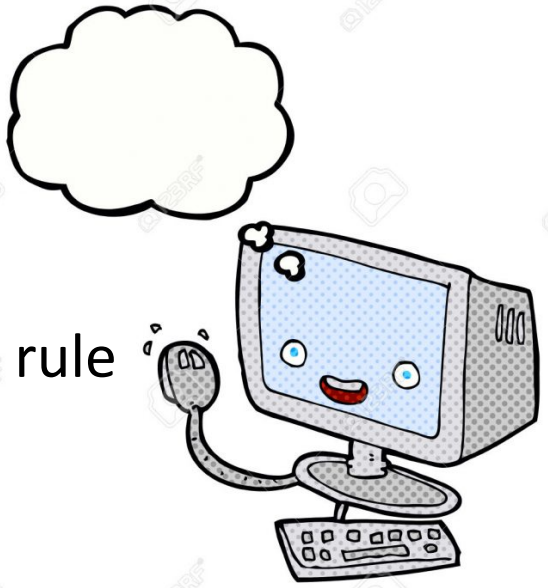
Attention weights

Predicted value

Earlier conscious state

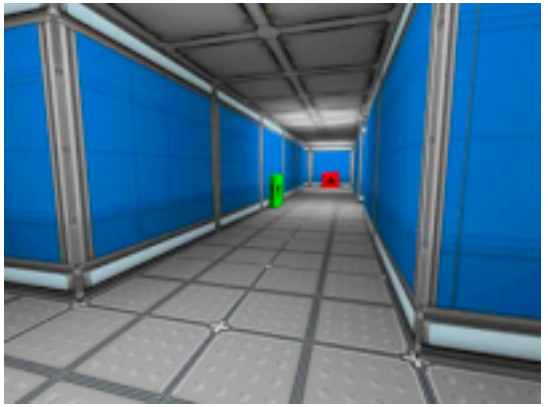
- How to train the attention mechanism which selects which variables to predict?
- (predicted variables, conditioning variables) = rule

  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**



- 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

Montreal Institute for Learning Algorithms

MILA

Université de Montréal