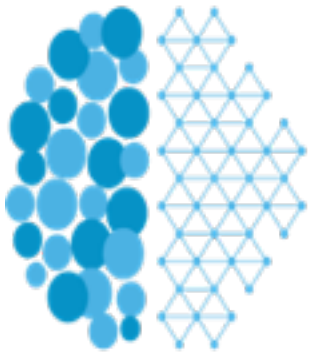
Towards disentangling underlying explanatory factors

Yoshua Bengio

July 13th, 2018

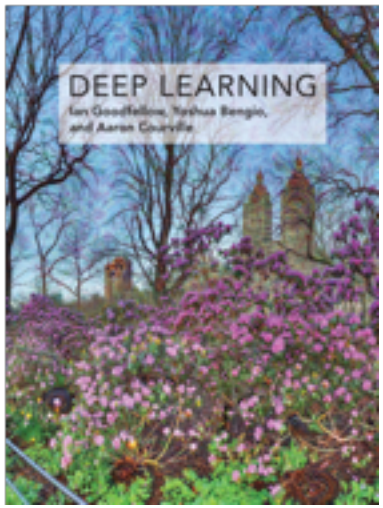
ICML'2018 Workshop on Learning with Limited Labels



MILA

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

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



Université de Montréal

IVADO

Generalizing Beyond i.i.d. Data

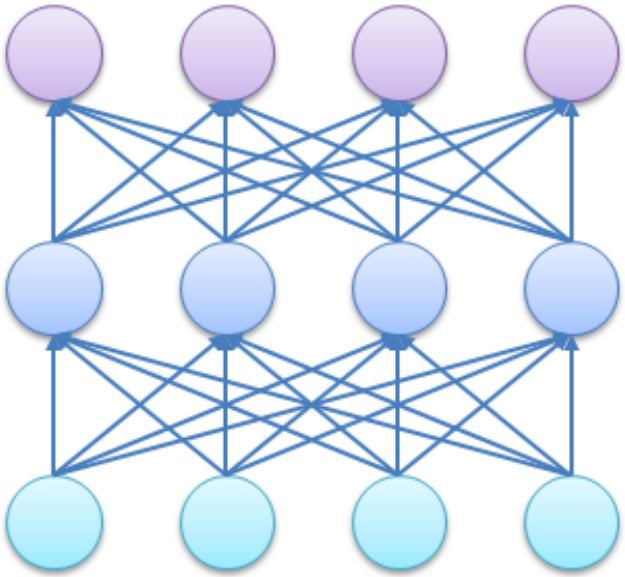
- Current ML theory is strongly dependent on the iid assumption
- Real-life applications often require generalizations in regimes not seen during training

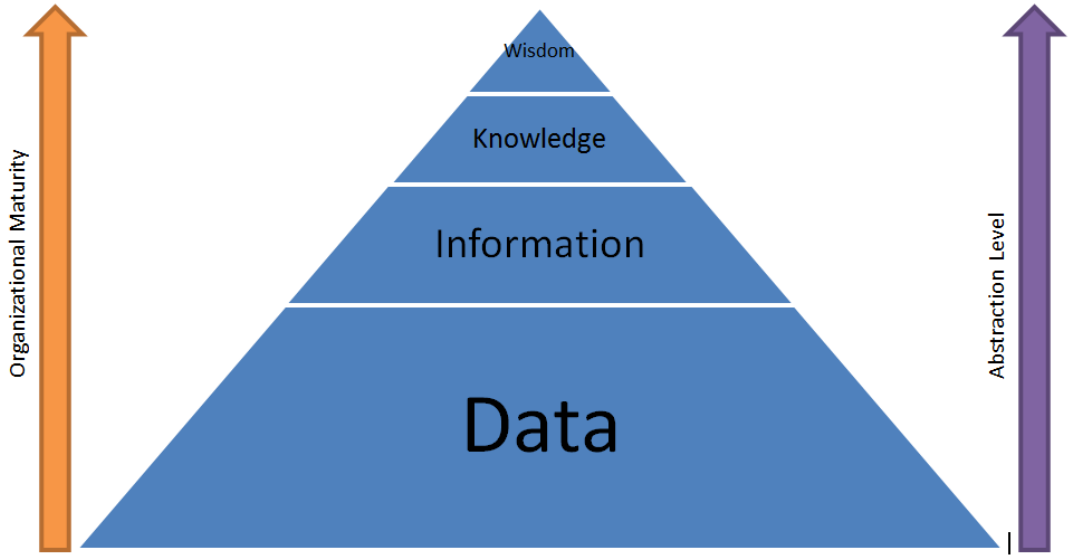
- Humans can project themselves in situations they have never been (e.g. imagine being on another planet, or going through exceptional events like in many movies)
- **Key to success: understanding explanatory/ causal factors and mechanisms**

Learning Multiple Levels of Abstraction

(Bengio & LeCun 2007)

- The big payoff of deep learning is to facilitate learning higher levels of abstraction
- Higher-level abstractions can **disentangle the factors of variation**, which allows much easier generalization and transfer





Invariance and Disentangling

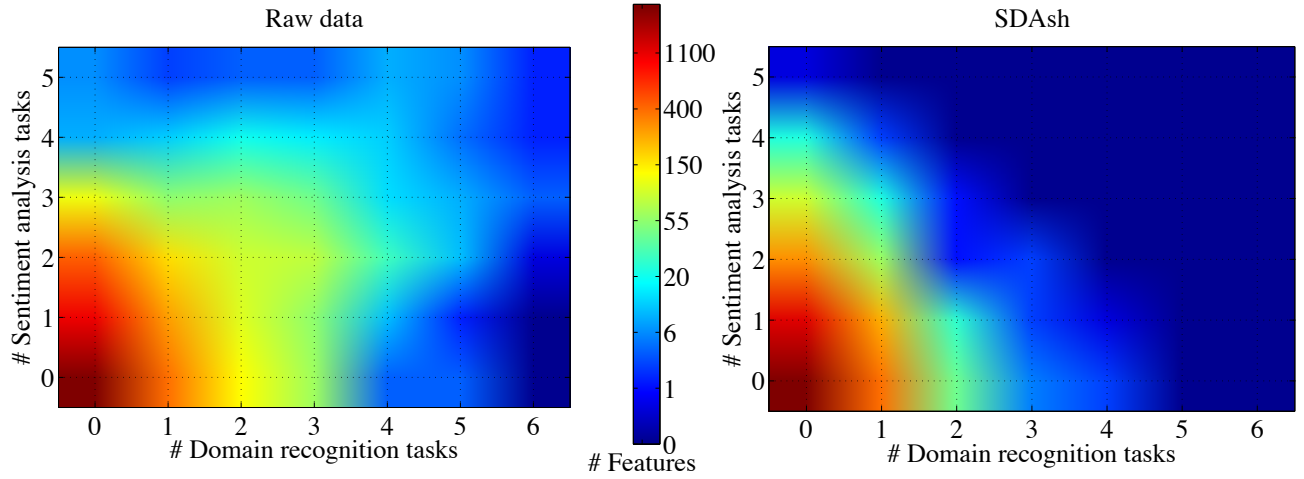
- Invariant features
- Which invariances?
- Alternative: learning to disentangle factor

- Good disentangling →
  avoid the curse of dimensionality:

Dependencies are "simple" when the data is projected in the right abstract space

Disentangling from denoising objective (Glorot, Bordes & Bengio ICML 2011)

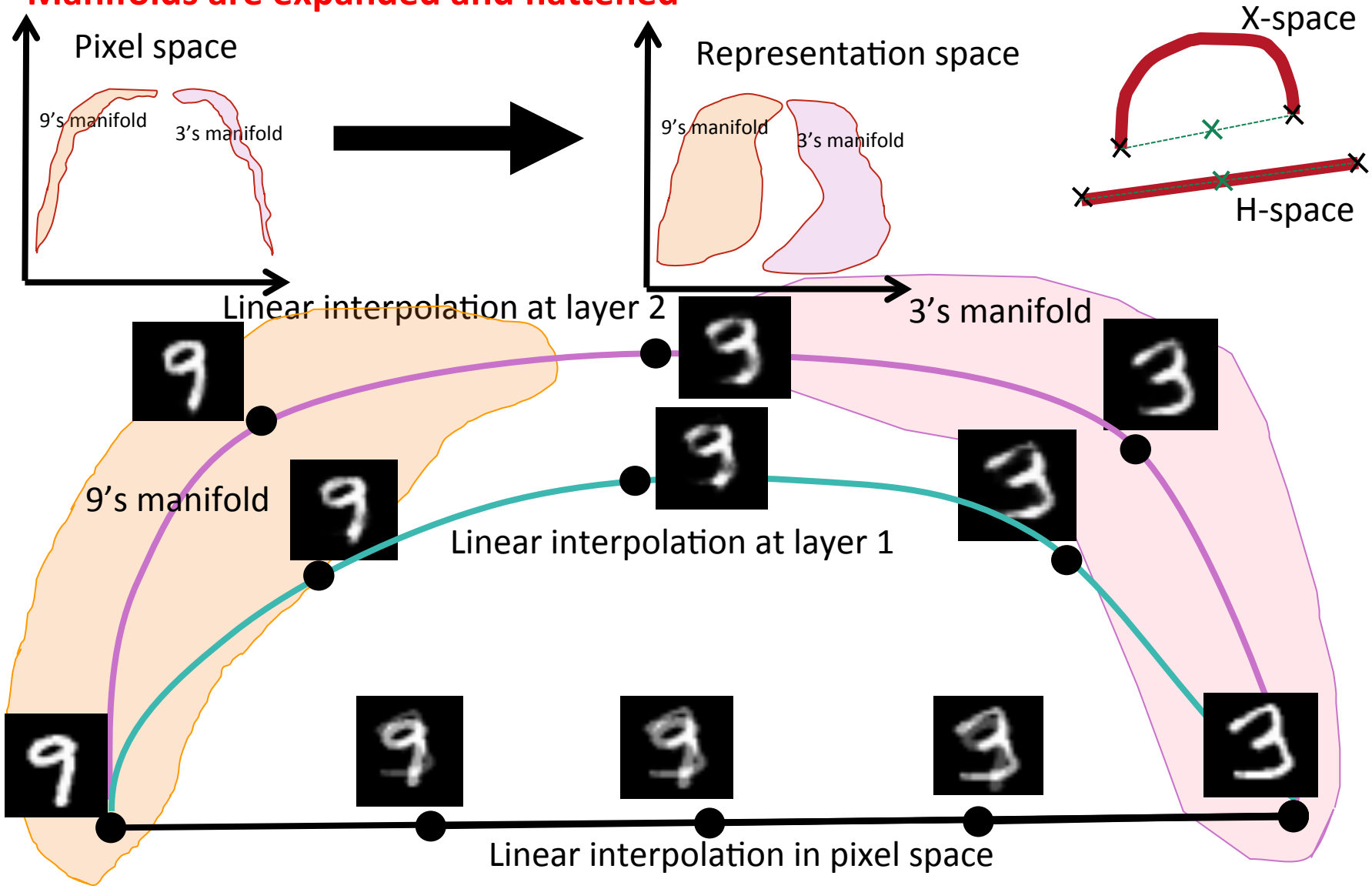
- Early deep learning research already is looking for possible disentangling arising from unsupervised learning of representations
- Experiments on stacked denoising auto-encoders with ReLUs, on BoW text classification
- Features tend to specialize to either sentiment or domain



Space-Filling in Representation-Space

(Bengio et al ICML 2013)

- **Deeper representations → abstractions → disentangling**
- **Manifolds are expanded and flattened**



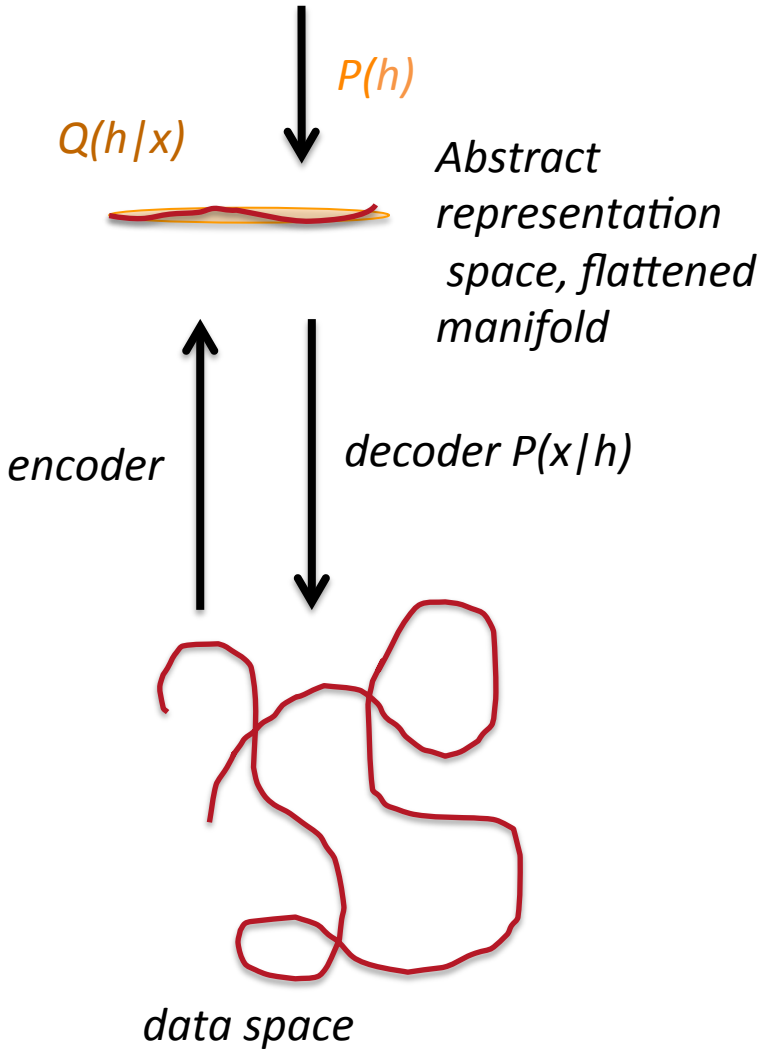
Interpolating in Latent Space

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

Radford et al 2016

Latent Variables and Abstract Representations

- Encoder/decoder view: maps between low & high-levels
- Encoder does inference: interpret the data at the abstract level
- Decoder can generate new configurations
- Encoder flattens and disentangles the data manifold

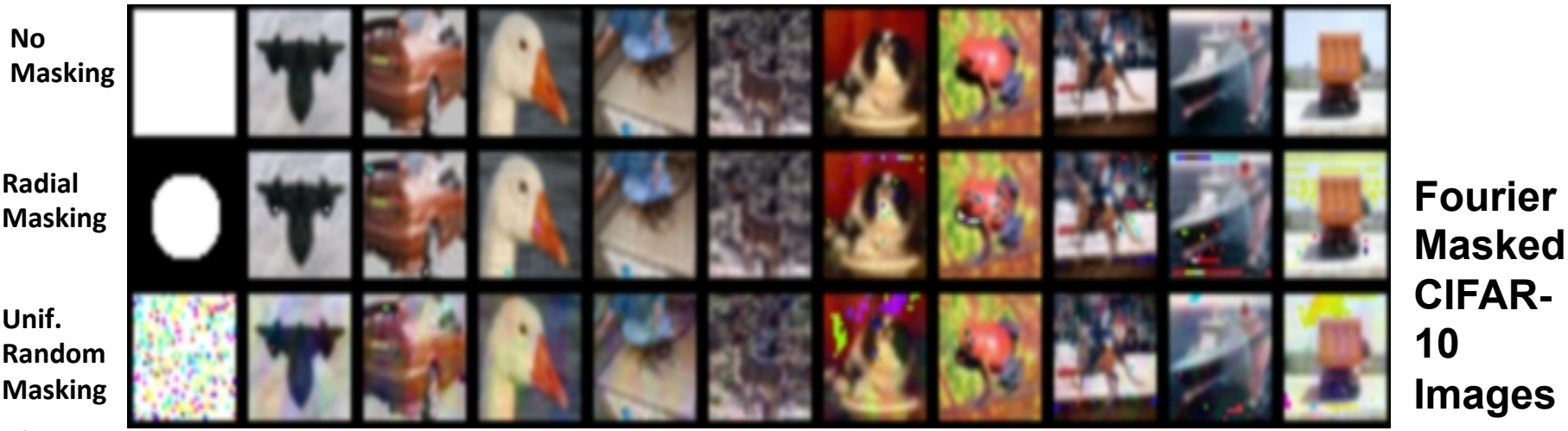


Measuring the Tendency of CNNs to Learn Surface Statistical Regularities

Jason Jo and Yoshua Bengio 2017, arXiv:1711.11561

- **Hypothesis**: *Deep CNNs have a tendency to learn superficial statistical regularities in the dataset rather than high level abstract concepts*.

- From the perspective of learning high level abstractions, Fourier image statistics can be *superficial* regularities, not changing object category

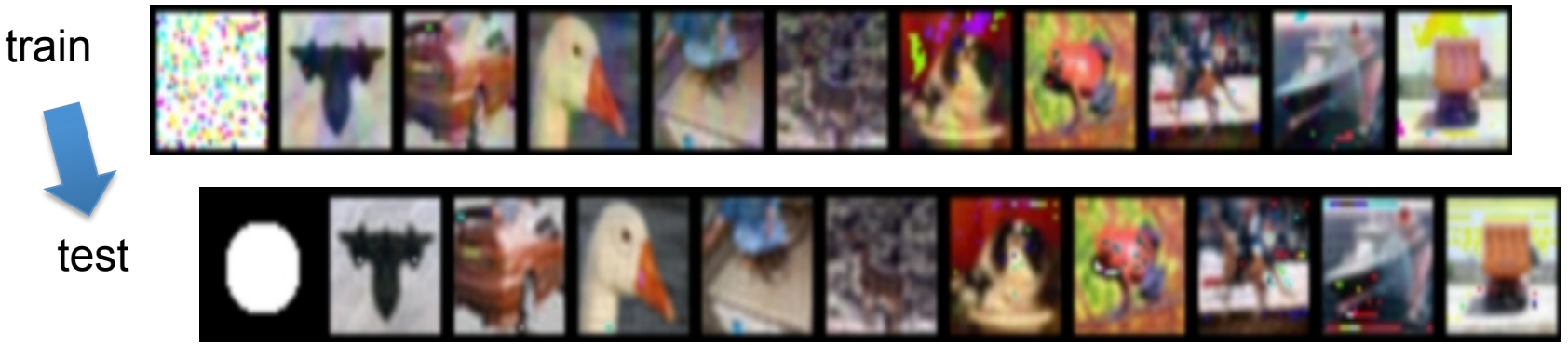


Measuring the Tendency of CNNs to Learn Surface Statistical Regularities

Jason Jo and Yoshua Bengio 2017, arXiv:1711.11561

- Different Fourier filters, same high level abstractions (objects) but different surface statistical regularities (Fourier image statistics).

- Experiment: Train on one training set and evaluate the test sets.
- A generalization gap: max difference in test accuracies



- Large generalization gap: CNN exploits too much of low level regularities, as opposed to learning the abstract high level concepts.

What's Missing with Deep Learning?

Deep Understanding

Learning « How the world ticks »

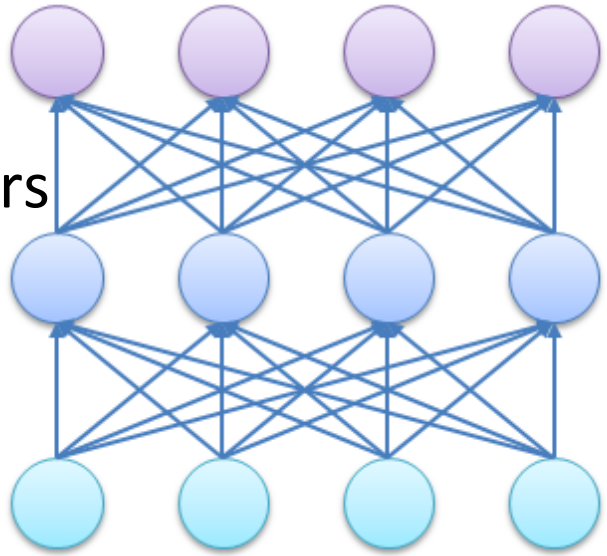
- So long as our machine learning models « cheat » by relying only on superficial statistical regularities, they remain vulnerable to out-of-distribution examples

- Humans generalize better than other animals thanks to a more accurate internal model of the **underlying causal relationships**

- To predict future situations (e.g., the effect of planned actions) far from anything seen before while involving known concepts, an essential component of reasoning, intelligence and science

How to Discover Good Disentangled Representations

- How to discover abstractions?
- What is a good representation? ***(Bengio et al 2013)***
- Need clues (= priors) to help **disentangle** the underlying factors, such as
  - Spatial & temporal scales
  - Marginal independence
  - Simple dependencies between factors
    - *Consciousness prior*
  - Causal / mechanism independence
    - *Controllable factors*



Acting to Guide Representation Learning & Disentangling

(E. Bengio et al, 2017; V. Thomas et al, 2017)

- **Some factors (e.g. objects) correspond to 'independently controllable' aspects of the world**
- *Can only be discovered by acting in the world*
  - *Control linked to notion of objects & agents*
  - *Causal but agent-specific & subjective: affordances*

Abstraction Challenge for Unsupervised Learning

- Why is modeling P(acoustics) so much worse than modeling P(acoustics | phonemes) P(phonemes)?

- Wrong level of abstraction?
  - **many more entropy bits in acoustic details then linguistic content**

  → predict the future in in abstract space instead: non-trivial

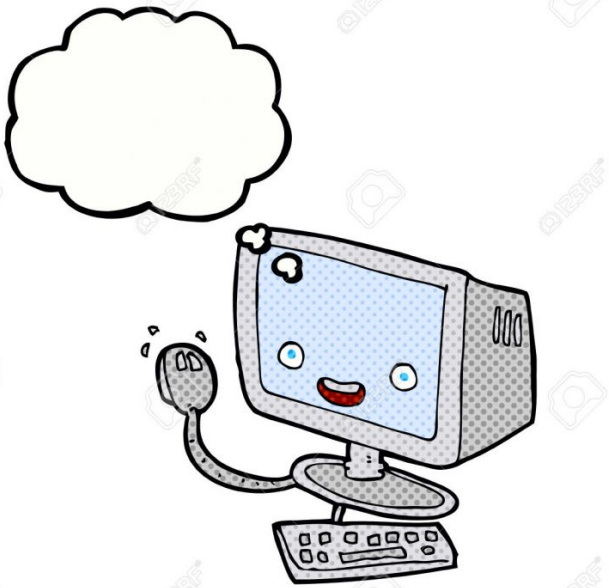
The Consciousness Prior

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 (key/value) 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)



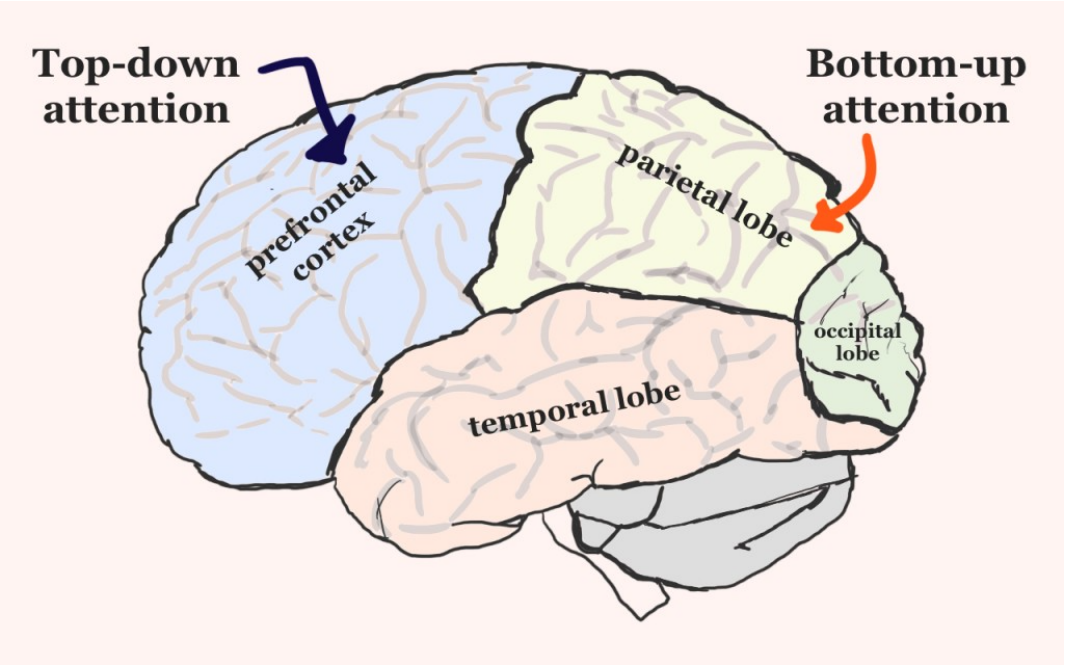
How to select a few relevant abstract concepts making a thought?

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

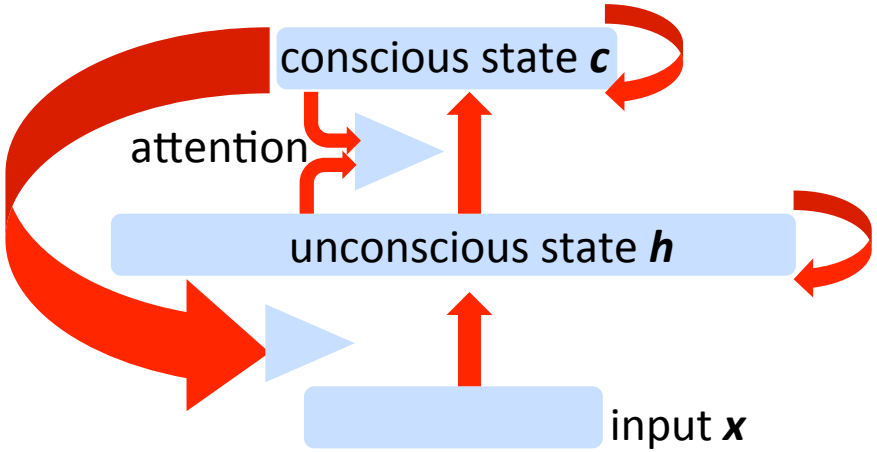
Attention focuses on a few appropriate abstract or concrete elements of mental representation



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***



- ***c*** includes names (keys) and values of fac

Disentangling up to Linear Projection

- My old view of disentangling: each dimension of the representation = one 'nameable' (semantic) factor
- Potential problem: the number of 'nameable' factors is limited by the number of units, and brains don't use a completely localized representation for named things
- My current view of disentangling: it is enough that a linear projection exist to 'classify' or 'predict' any of the factors
- The 'number' of potential 'nameable' factors is now exponentially larger (e.g. subsets of dimensions, weights of these projections)

The Consciousness Prior

Bengio 2017, arXiv:1709.08568

- Conscious prediction over attended variables *A* (soft attention)

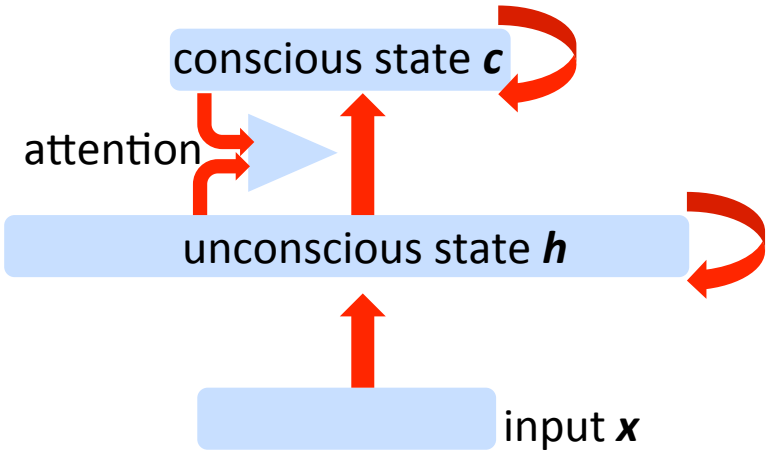
$$V = -\sum_A w_A \log P(h_{t,A} = a | c_{t-1})$$

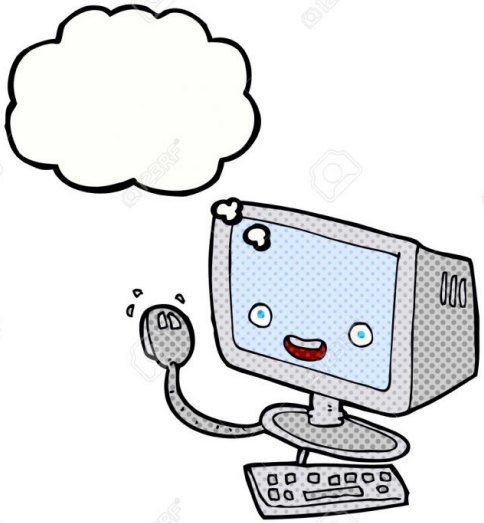
Attention weights

Factor **name**

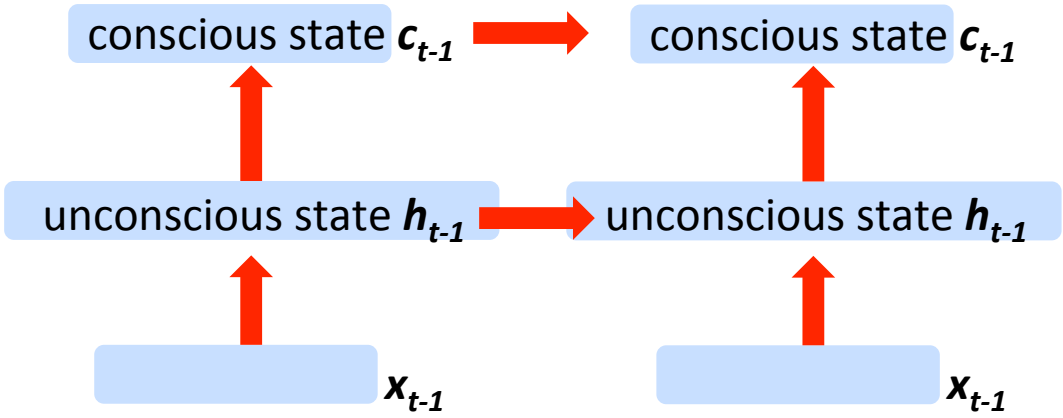
Predicted **value**

Earlier conscious state





What Training Objective?



- How to train the attention mechanism which selects which variables to predict?
  - Representation learning without reconstruction:
    - Maximize entropy of code
    - Maximize mutual information between past and future
  - *Objective function completely in abstract space, higher-level parameters model dependencies in abstract space*
  - *Usefulness of thoughts: as conditioning information for action, i.e., a particular form of planning for RL, i.e., the estimated gradient of rewards could also be used to drive learning of abstract representations*

Montreal Institute for Learning Algorithms

MILA

Université de Montréal