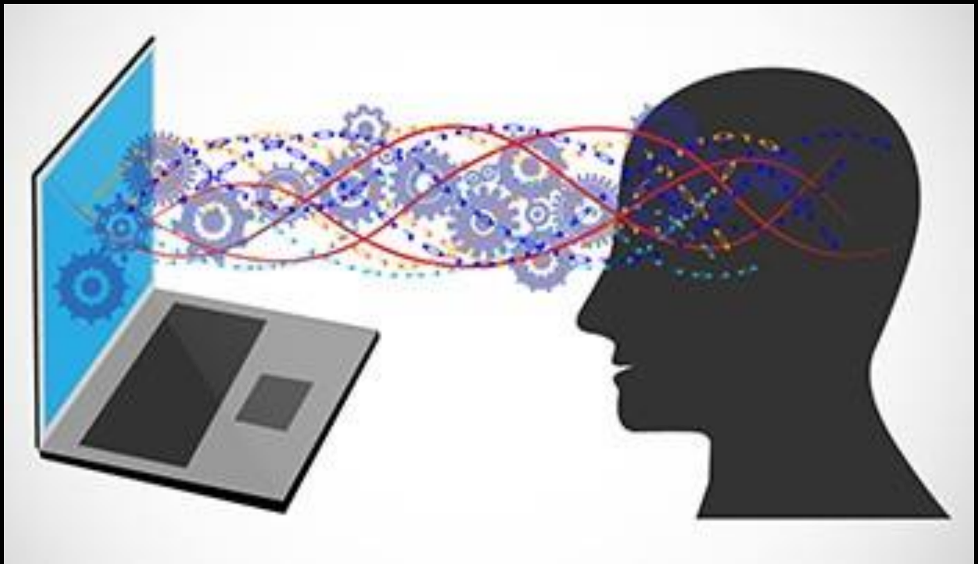
Disentangled learning

AI & Knowledge

- Putting knowledge into computers
- Much knowledge is intuitive, uncommunicable



Still Far from Human-Level AI

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

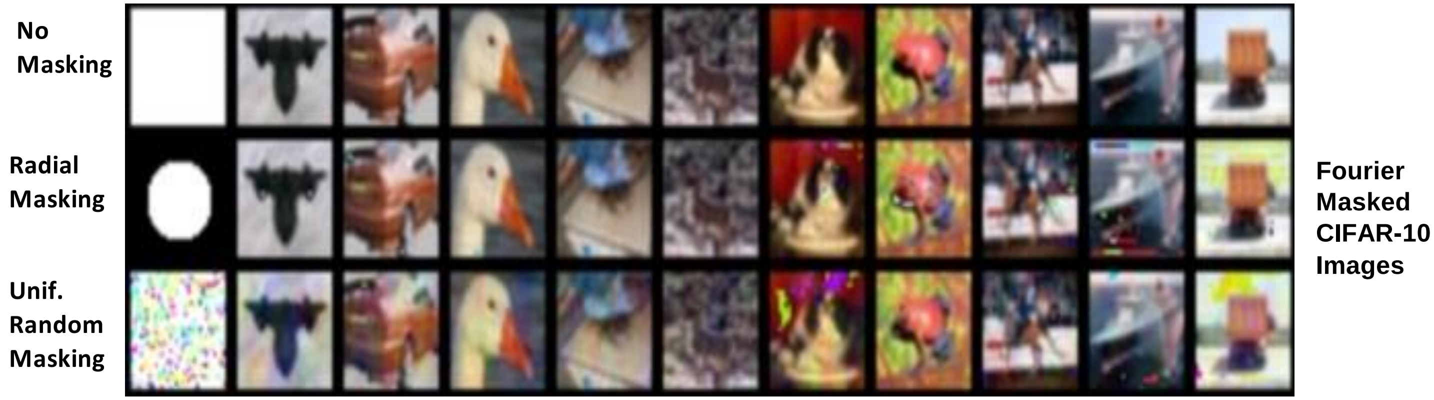
- Learning superficial clues, not generalizing well enough outside of training contexts, easy to fool trained networks:
  - Current models cheat by picking on surface regularities

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*.

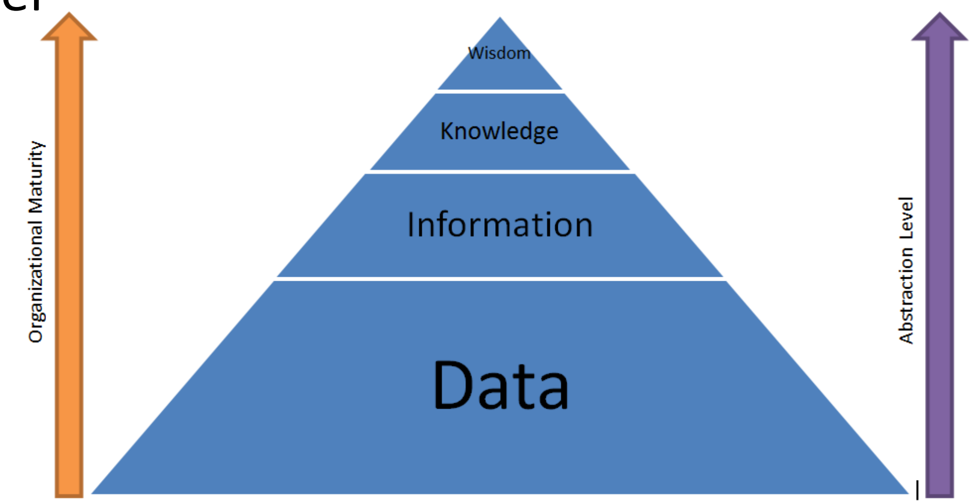
- From the perspective of learning high level abstractions, Fourier image statistics can be *superficial* regularities, not changing object category, but changing them leads CNNs to make mistakes



Learning Multiple Levels of Abstraction

(Bengio & LeCun 2007)

- 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



New concern:
Also disentangle the computation (modules) and the hypothesized causal mechanisms

Mila

Beyond the iid assumption

- The assumption that the test data is from the same distribution as the training data is too strong, and it is often violated in practice, leading to poor out-of-distribution generalization.

- I propose to consider relaxed assumptions: the test data was generated under the same causal dynamics, but from different initial conditions (which may be unlikely under the training distribution).

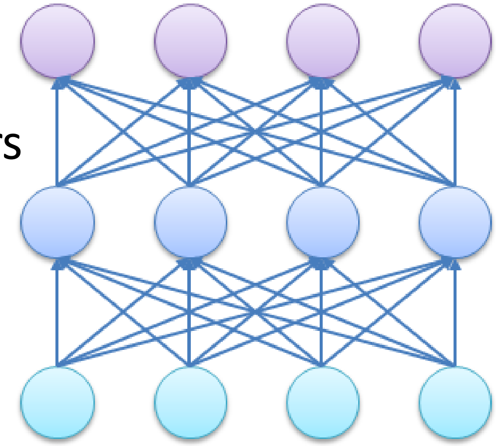
Stochastic dynamical system

Initial conditions

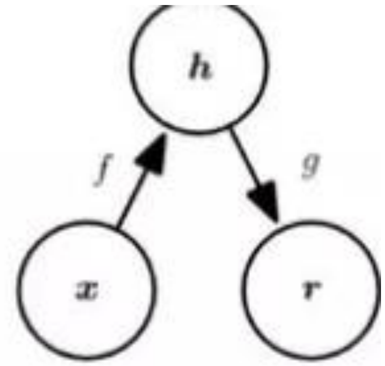
Observed data

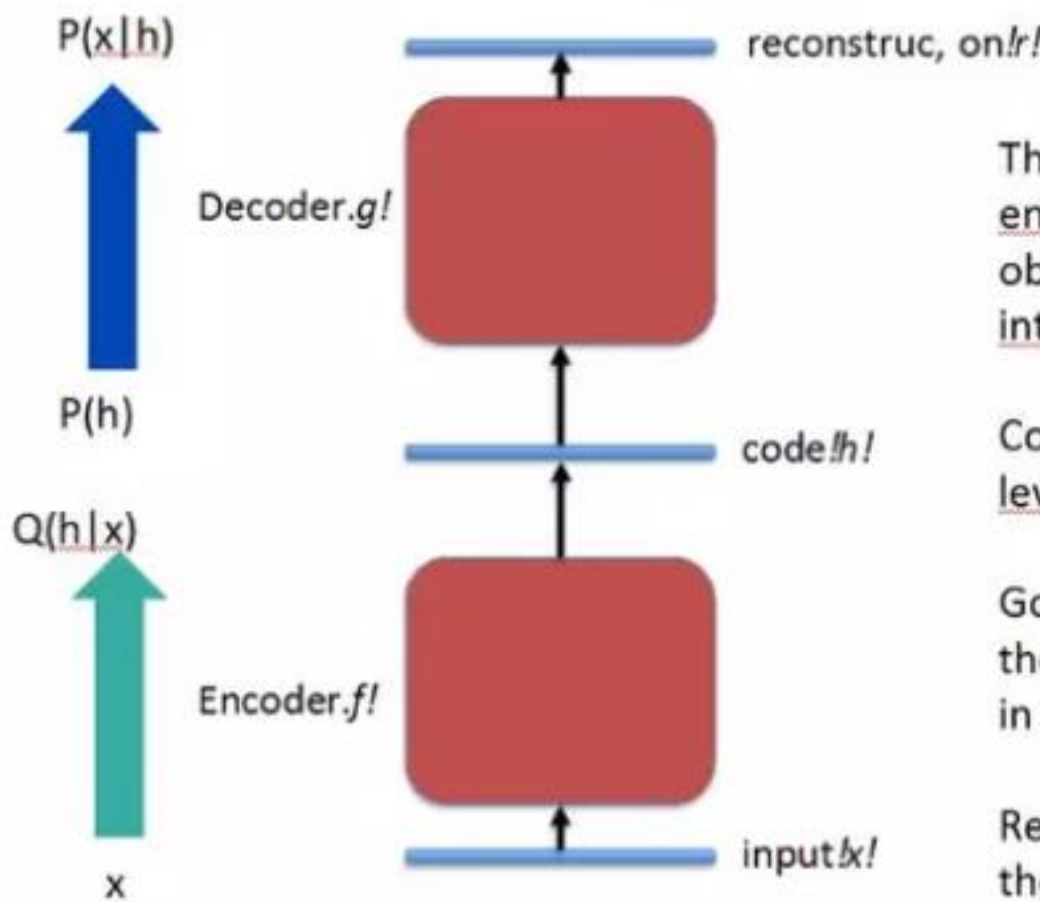
How to Discover Good Disentangled Representations

- How to discover abstractions?
- What is a good representation? ***(Bengio et al 2013)***
- ***Dependencies are simple in the right representation***
- 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*



Auto-Encoders: 2-way map





There are many variants of auto-encoders, with different training objectives and probabilistic interpretations.

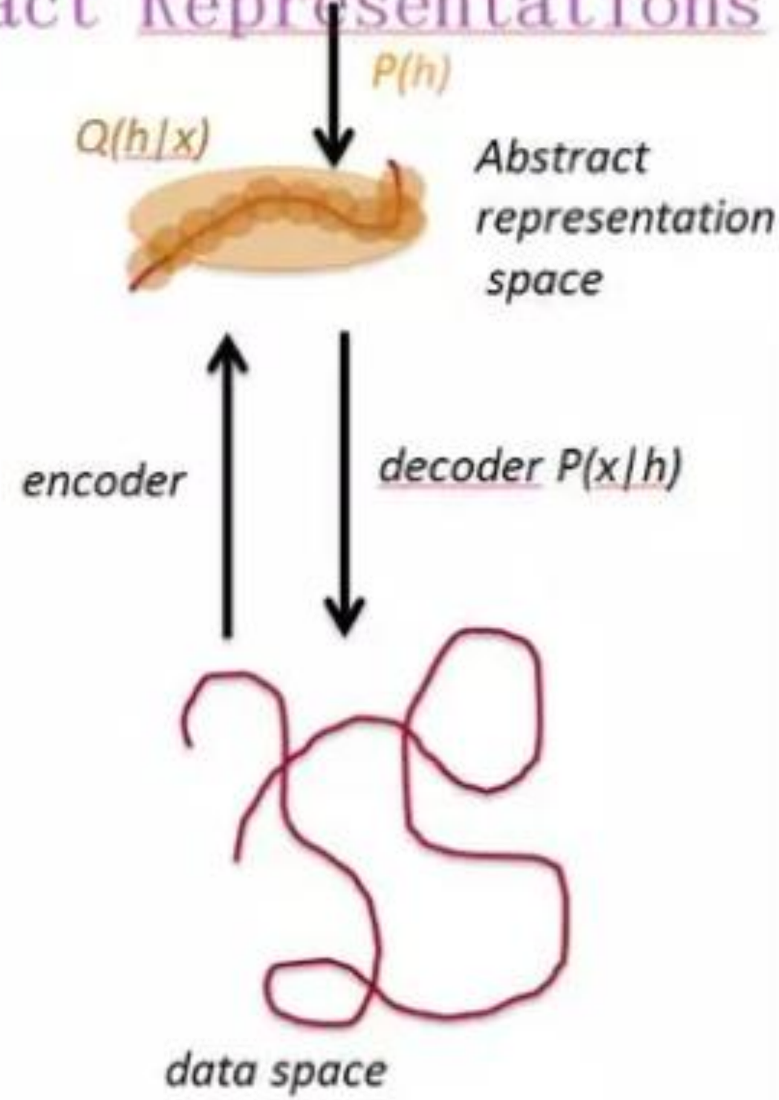
Code h is meant to be a higher-level representation of input x.

Good auto-encoders preserve the most important information in x.

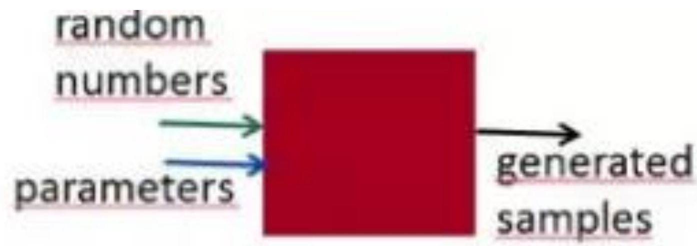
Reconstruction error measures the loss in information.

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



Generative Models



- One way to demonstrate that a learner understands the data distribution is to ask it to generate new examples from it

- New face images generated by a GAN variant called BEGAN using a training set of face images.

System 1 vs System 2 Cognition

Two systems (and categories of cognitive tasks):

- **System 1**
  - Intuitive, fast heuristic, UNCONSCIOUS, non-linguistic
  - *What current **deep learning** does quite well*
- **System 2**
  - Slow, logical, sequential, CONSCIOUS, linguistic, algorithmic
  - *What **classical symbolic AI** was trying to do*
- **Grounded language learning**: combine both language learning and world modeling

Mila

The Consciousness Prior

Bengio 2017, arXiv:1709.08568

- Focus on **representation learning** and one aspect of consciousness:
- Conscious thoughts are very low-dimensional objects compared to the full state of the (unconscious) brain = analogous to a sentence or a rule in rule-based systems
- Yet they have unexpected predictive value or usefulness

  à strong constraint or prior on the underlying representation

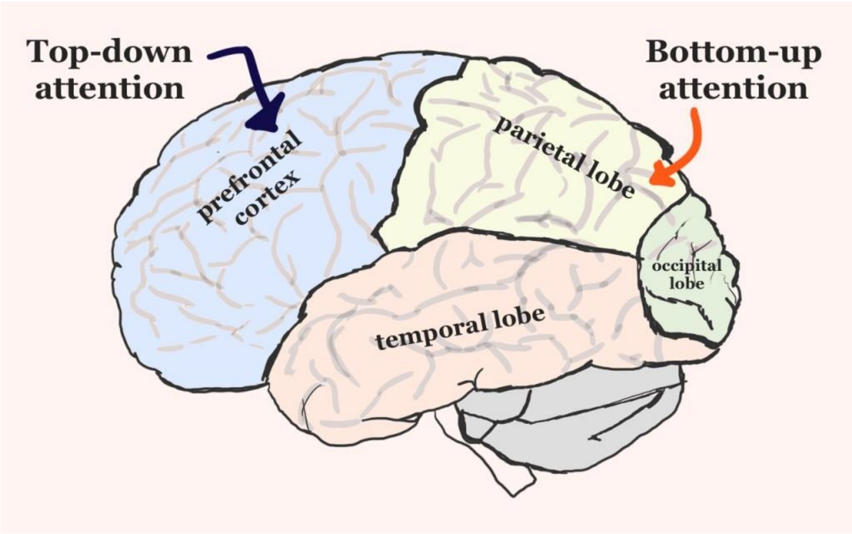
    - **Thought**: composition of few selected factors / concepts at the highest level of abstraction of our brain
    - Richer than but closely associated with short verbal expression such as a **sentence** or phrase, a **rule** or **fact** (link to classical symbolic AI & knowledge representation)
    - Variables in rule ó features in representation space
    - Rules ó causal mechanisms

Need to disentangle both

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



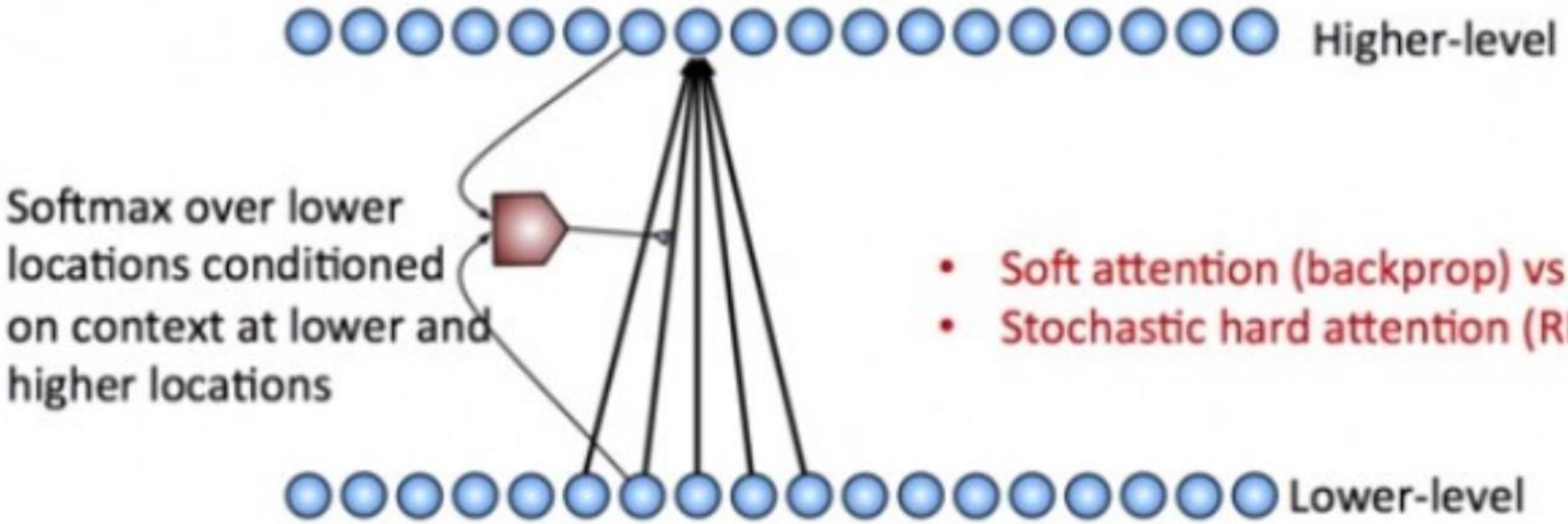
- Different from sparse auto-encoders: controller chooses focus, conditionally

Mila

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

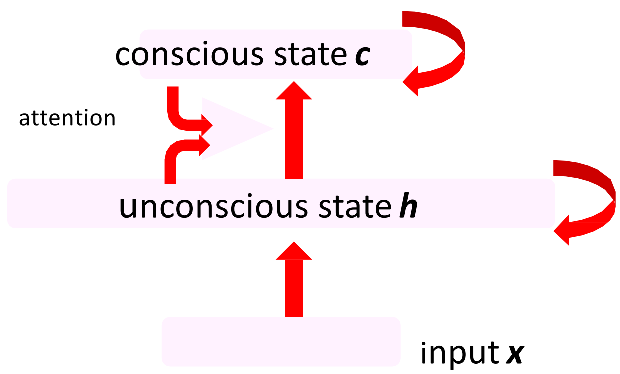


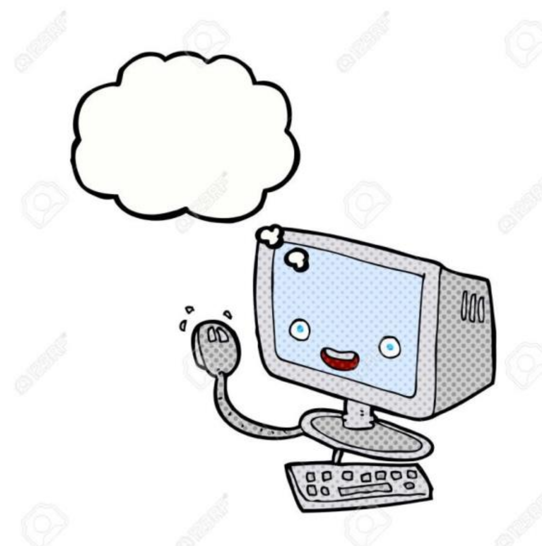
- Soft attention (backprop) vs
- Stochastic hard attention (RL)

- Impact of soft-attention: not just machine translation, also reasoning & memory, handling data structures, etc.

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***



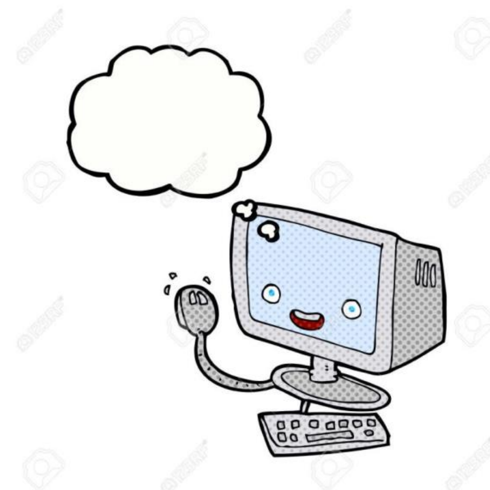


- ***c*** includes names (keys) and values of factors

Mila

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 representations** (*Becker & Hinton 1992*), **between intentions (policies) and changes in representations** (affordances, independently controllable factors)
  - *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*



Mila

Where We Are: Still Far Away

- All industrial successes are based on pure supervised learning
- Still learning superficial clues that do not generalize well outside of training contexts and make it easy to fool trained networks:
  - Current models cheat by picking on surface regularities, e.g., background greenery → animal is present
- Still unable to do a good job of learning higher-level abstractions at multiple time scales, deal with very long-term dependencies
- Still relying heavily on smooth differentiable predictors (using backprop)

Progress and Obstacles in Deep Unsupervised Generative Models

- Humans are very good at unsupervised learning, e.g. 2 year old know intuitive physics
- RBMs and DBMs: obstacle probably due to gradient estimator relying on good mixing of MCMC (which gets worse as training progresses because distribution becomes sharper)
- Autogressive models (NADE, MADE, PixelRNN, PixelCNN, WaveNet): easier to train but no latent variables
- VAEs and GANs: the current frontier, hard to train, still unsatisfactory in terms of extracting abstraction

Early Days of GAN Samples

MNIST

TFD

CIFAR-10 (fully connected)

CIFAR-10 (convolutional)

Convolutional GANs

(Radford et al, arXiv 1511.06343)

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

Challenges of Training GANs

- Training can be unstable and diverge 训练过程不稳定，容易发散
  对超参和训练细节敏感
- Sensitive to hyper-parameters and details of training
- Mode collapse: almost same image generated many times
  模式单调，总是生成相同图像
- Missing modes: subtypes of the data are absent
- Difficult to handle many underlying categories without providing them during training (as in supervised training)
- Difficult to monitor progress
- No accepted quantitative measure of quality
- **But GANs can work amazingly well!** GANs问题虽多，但效果出奇的好！
- So more than a dozen variants have been proposed to address some of these issues, lots of research ongoing
- **See Ian Goodfellow's NIPS tutorial**

What's Missing

- More autonomous learning, **unsupervised learning**
- Discovering the **underlying causal factors**
- Model-based RL which extends to completely new situations by **unrolling powerful predictive models which can help reason about rarely observed dangerous states**
- Sufficient **computational power** for models large enough to capture human-level knowledge
- Autonomously discovering **multiple time scales to handle very long-term dependencies**
- Actually **understanding language** (also solves generating), requiring enough world knowledge / commonsense
- Large-scale **knowledge representation** allowing one-shot learning as well as discovering new abstractions and explanations by 'compiling' previous observations

Acting to Guide Representation Learning

- What is a good latent representation?
- The notion of disentangling the underlying factors of representation is not specific enough
- New on-going research: **appropriate factors each correspond to 'independently controllable' aspects of the world**
- Can only be discovered by acting in the world
- Some factors deduced by analogy (e.g. the sun) as caused by imagined (or imaginary) agents