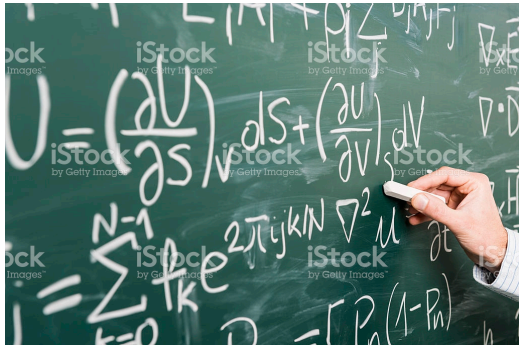
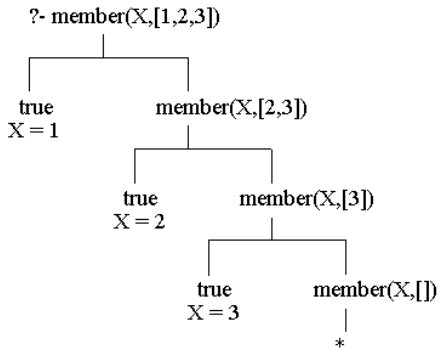

Mila
Deep Learning for AI
Yoshua Bengio
CAIANELLO AWARD LECTURE
VIETRI SUL MARE, 28 JUNE 2019
CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH
ICRA
INSTITUT CANADIEN DE RECHERCHES AVANCÉES

Neural Networks & AI: Underlying Assumption

- There are principles giving rise to intelligence (machine, human or animal) via learning, simple enough that they can be described compactly, similarly to the laws of physics, i.e., our intelligence is not just the result of a huge bag of tricks and pieces of knowledge, but of general mechanisms to acquire knowledge.

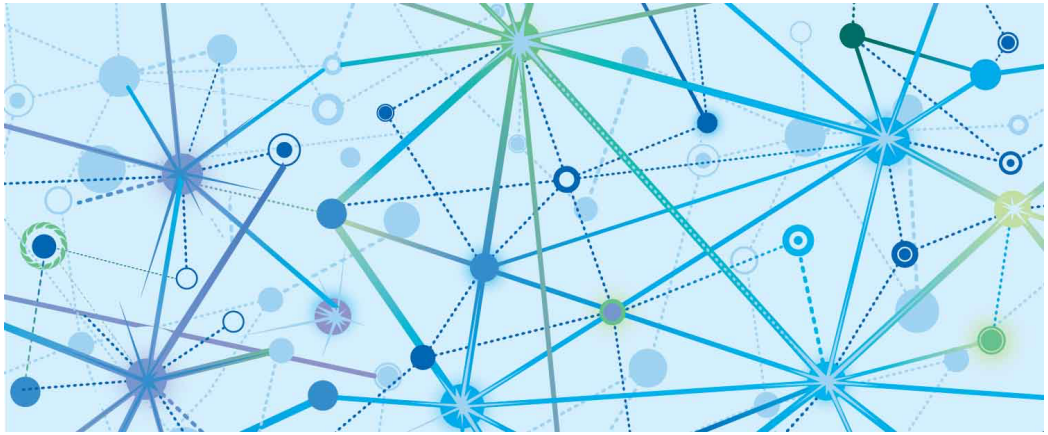


The Machine Learning approach to AI



- **Classical AI, rule-based, symbolic**
  - knowledge is provided by humans
    - but intuitive knowledge (e.g. much of common sense) not communicable
  - machines only do inference
  - no learning, adaptation
  - insufficient handling of uncertainty
  - not grounded in low-level perception and action
- **Machine learning tries to fix these problems**
  - succeeded to a great extent
  - higher-level (conscious) cognition still seems out of reach

The Neural Net Approach to AI



- **Brain-inspired**
- Synergy of a large number of simple adaptive computational units
- Focus on **distributed representations**
  - **E.g. word representations** (Bengio et al NIPS'2000)
- View intelligence as arising of combining
  - an objective function
  - an approximate optimizer (learning rule)
  - an initial architecture / parametrization
- End-to-end learning (all the pieces of the puzzle adapt to help each other)

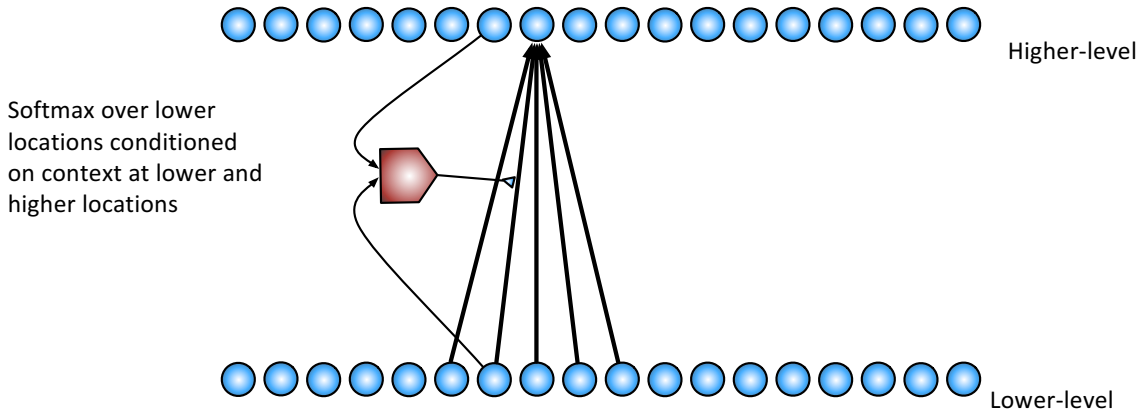
What is Deep Learning about?

- Rebranding and extending Neural Networks (ca 2005-2007)
- New methods to enable training of deeper networks
  - (Bengio et al NIPS'2006): pre-training stacks of auto-encoders before supervised training, greedy supervised and unsupervised pre-training
  - (Glorot & Bengio AISTATS'2010): initialization with near 1 e-values Jacobians
  - (Glorot & Bengio AISTATS'2011): importance of ReLU for training deep nets
- Beyond pattern recognition:
  - Progress in deep unsupervised models, generative models
    - (Vincent et al & Bengio 2008)++: denoising auto-encoders, self-supervised objectives
    - (Goodfellow et al & Bengio 2014): GANs = generative adversarial networks
  - Attention mechanisms & operating on arbitrary data structures
  - Meta-learning (Bengio & Bengio 1991; many more in last 2 years)

The Attention Revolution in Deep Learning

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

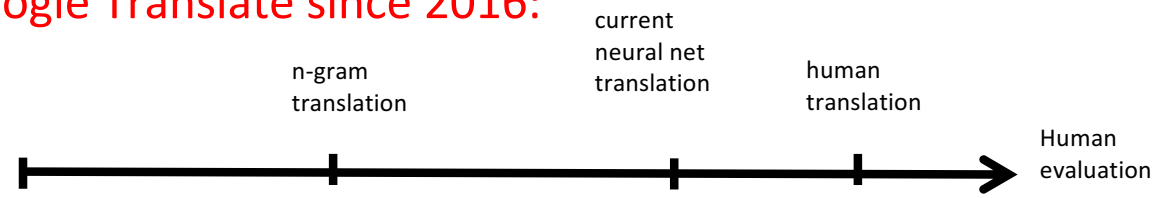
Neural Machine Translation (ICLR'2015)



Attention enables:

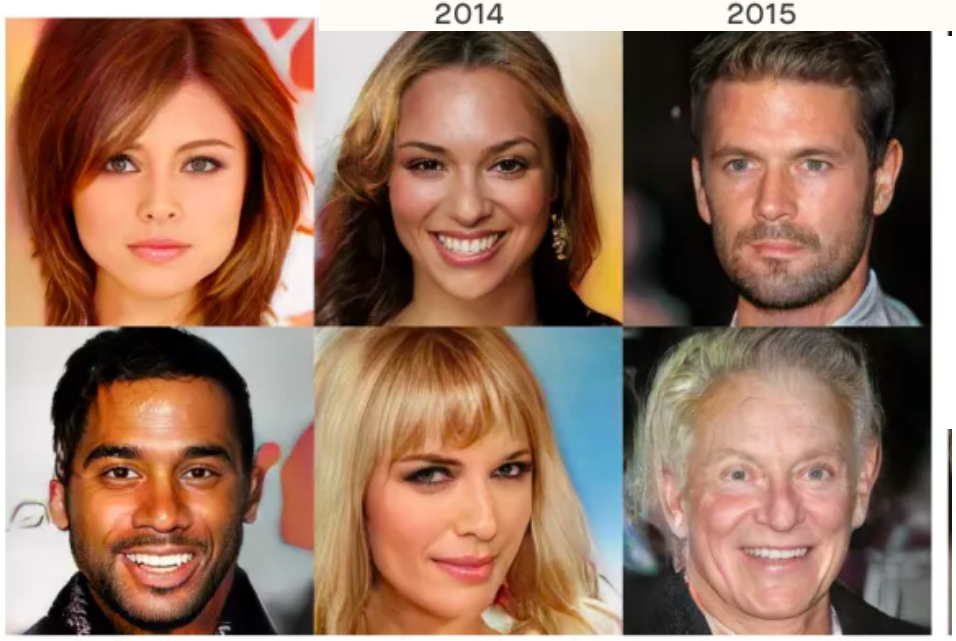
- Differentiable memory access
- Operating on sets
- Long-term dependencies
- Self-attention, transformers, SOTA
- Consciousness

- In Google Translate since 2016:



Generative Adversarial Networks

Goodfellow et al & Bengio NIPS 2014



Xu et al 2018, AttnGAN

System 1 vs System 2 Cognition

Two systems (and categories of cognitive tasks):

- **System 1**
  - Cortex-like (state controller and representations)
  - intuitive, fast heuristic, UNCONSCIOUS, non-linguistic
  - what current DL does quite well
- **System 2**
  - Hippocampus (memory) + prefrontal cortex
  - slow, logical, sequential, CONSCIOUS, linguistic, algorithmic
  - what classical symbolic AI was trying to do
- **Grounded language learning**: combine both systems

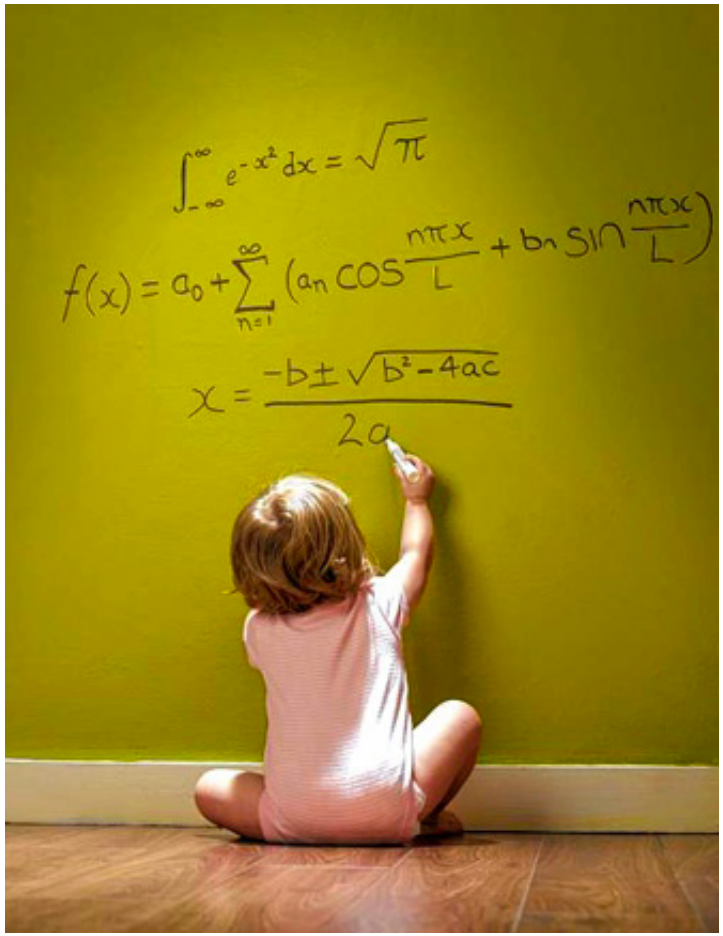
Still Far from Human-Level AI

- Industrial successes mostly based on **supervised** learning requiring lots of human-labeled data implicitly defining the relevant high-level abstractions.
- Learning relatively superficial clues, sometimes not generalizing well outside of training contexts, easy to fool trained networks:

Mila

Humans outperform machines at unsupervised 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.



Learning « How the world ticks »

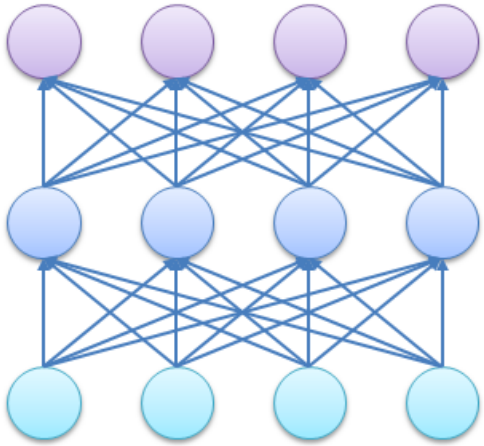
- So long as our machine learning models « cheat » by relying only on superficial statistical regularities, they remain vulnerable to out-of-distribution examples

- 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 (**not necessarily statistically independent**), such as
  - Spatial & temporal scales
  - Marginal independence
  - Simple dependencies between factors
    - *Consciousness prior*
  - Causal / mechanism independence
    - *Controllable factors*

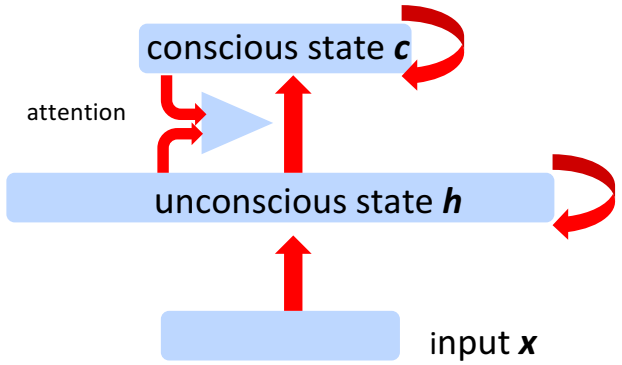


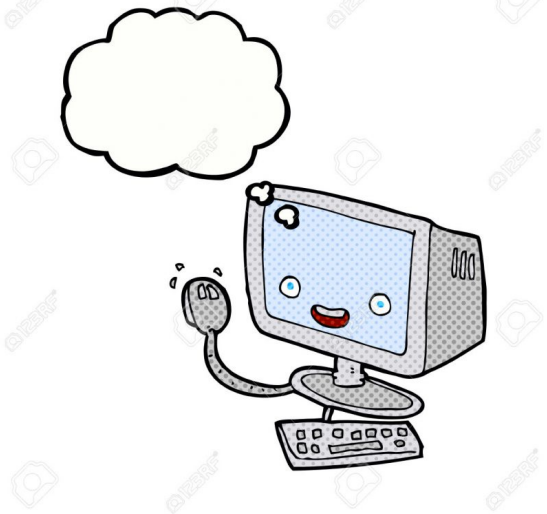
The Consciousness Prior

Bengio 2017, arXiv:1709.08568

- 2 levels of representation:
  - High-dimensional abstract representation space (all known concepts and factors) ***h***

    (not necessarily independent, but with sparse dependencies)
  - Low-dimensional conscious thought ***c***, extracted from ***h***





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

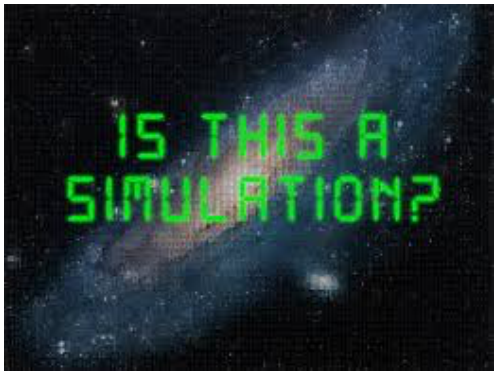
Acting to Guide Representation Learning & Disentangling

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

- **Some factors (e.g. objects) correspond to 'independently controllable' aspects of the world**
  - Corresponds to maximizing mutual information between intentions (goal-conditioned policies) and changes in the state (trajectories), conditioned on the current state.
- *Can only be discovered by acting in the world*
  - *Control linked to notion of objects & agents*
  - *Causal but agent-specific & subjective: affordances*

Current Model-Free RL too Statistically Inefficient: Combine Model-Based and Model-Free RL

- Simulate possible futures (given current state and actions) in order to train the policy (which can act quickly, without having to perform expensive planning)
- Need a good generative model of how agents cause changes in the world (effects)
- Better to generate future abstract states rather than future perceptions



Deep Learning Objective: discover causal representation

- What are the right representations? Causal variables explaining the data
- How to discover them?
- How to discover their causal relationship, the causal graph?

Mila

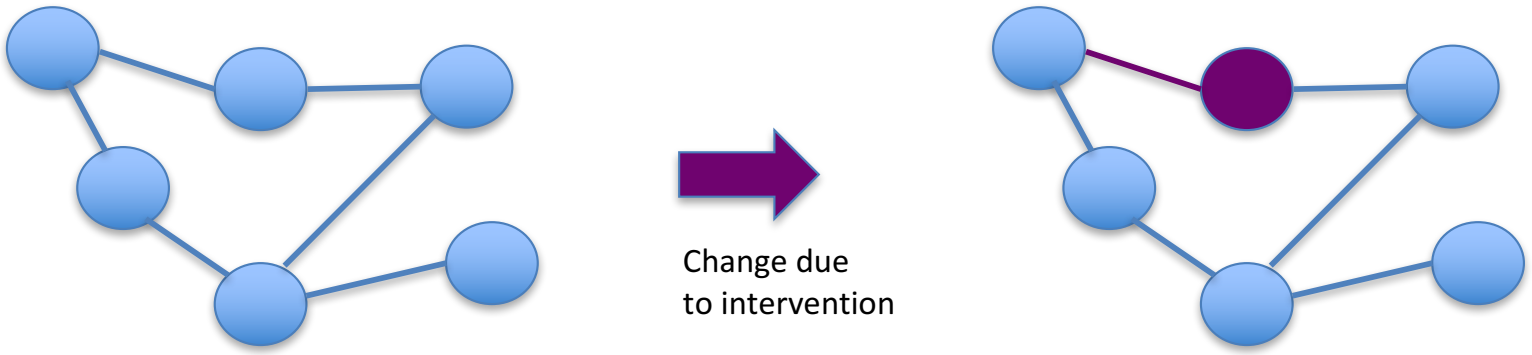
Disentangling: Factoring out aspects of the acquired knowledge

- How to disentangle the unobserved explanatory variables?
- How to separate the dependencies between these variables into separate easily re-usable pieces?
- How to modularize procedural knowledge into easily re-usable pieces? (options etc)
- How to modularize knowledge for easier re-use & adaptation, good transfer?

Mila

Separating Knowledge in Small Pieces

- Pieces which can be re-used combinatorially
- Pieces which are stable vs nonstationary, subject to interventions



Mila

Missing from Current ML: Understanding & Generalization Beyond the Training Distribution

- Learning theory only deals with generalization within the same distribution
- Models learn but do not generalize well (or have high sample complexity when adapting) to modified distributions, non-stationarities, etc.
- Poor reuse, poor modularization of knowledge

Beyond iid: Hypotheses about how the environment changes
Independent Mechanisms and the Small Change Hypothesis

- Independent mechanisms:
  - changing one mechanism does not change the others *(Peters, Janzig & Scholkopf 2017)*
- Small change:
  - Non-stationarities, changes in distribution, involve few mechanisms (e.g. the result of a single-variable intervention)

Mila

What if we had the right causal structure?

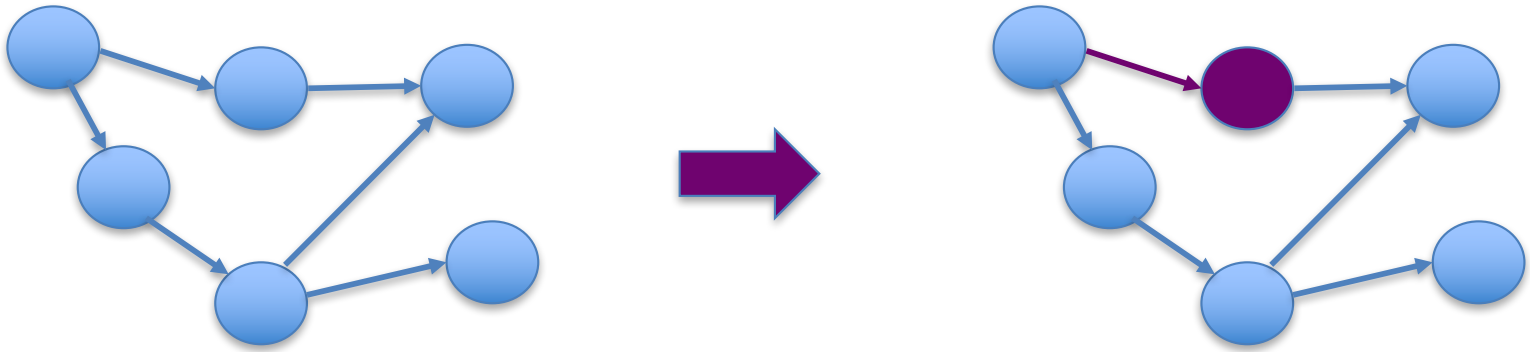
CLAIM: Under the hypothesis of independent mechanisms and small changes across different distributions:

- **smaller sample complexity to recover from a distribution change**
  - E.g. for transfer learning, agent learning, domain adaptation, etc.

Mila

Small Change in the Right Space

Distribution change: only one or a few mechanisms change



Before: eyes open

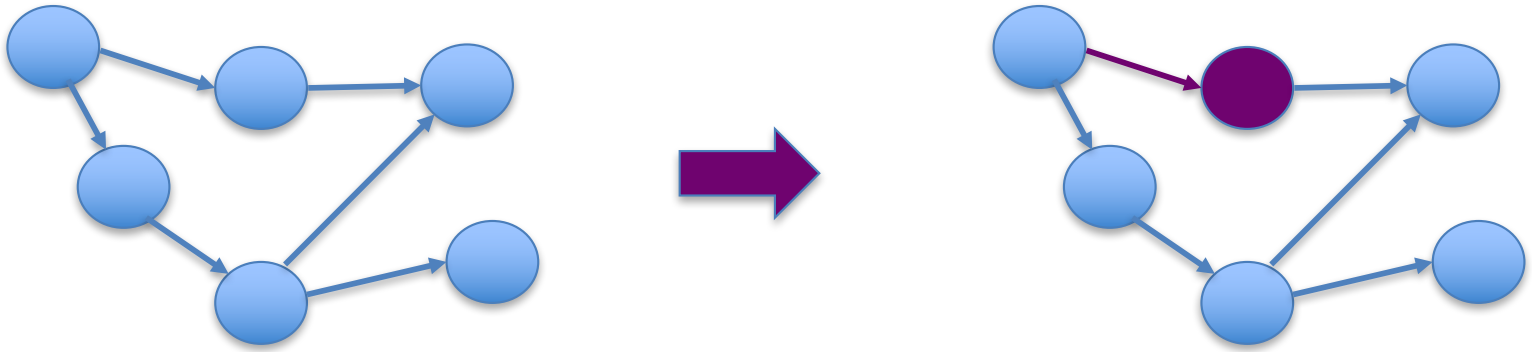
After: eyes closed,
totally different in pixel space,
small change in object space

Under the right parametrization, few parameters need to change after an intervention

Mila

Small Change → Small Sample Complexity

Few parameters need to change → small L2 change → *few examples needed to recover from the change*



Under the right parametrization → fast adaptation to interventions

Mila

Simple Running Example

- Consider two r.v. A, B, with A cause of B.
- Correct causal model decomposes
  - P(A,B) = P(A) P(B|A)

- Consider 2 distributions P_1 and P_2, only P(A) changes
- If we first train on P_1 and we have the right decomposition, adapting on P_2 is fast because

$$E_{P(B|A)}[\frac{\partial \log P_\theta(B|A)}{\partial \theta}] \approx 0 \text{ when } P_\theta(B|A) \approx P(B|A)$$

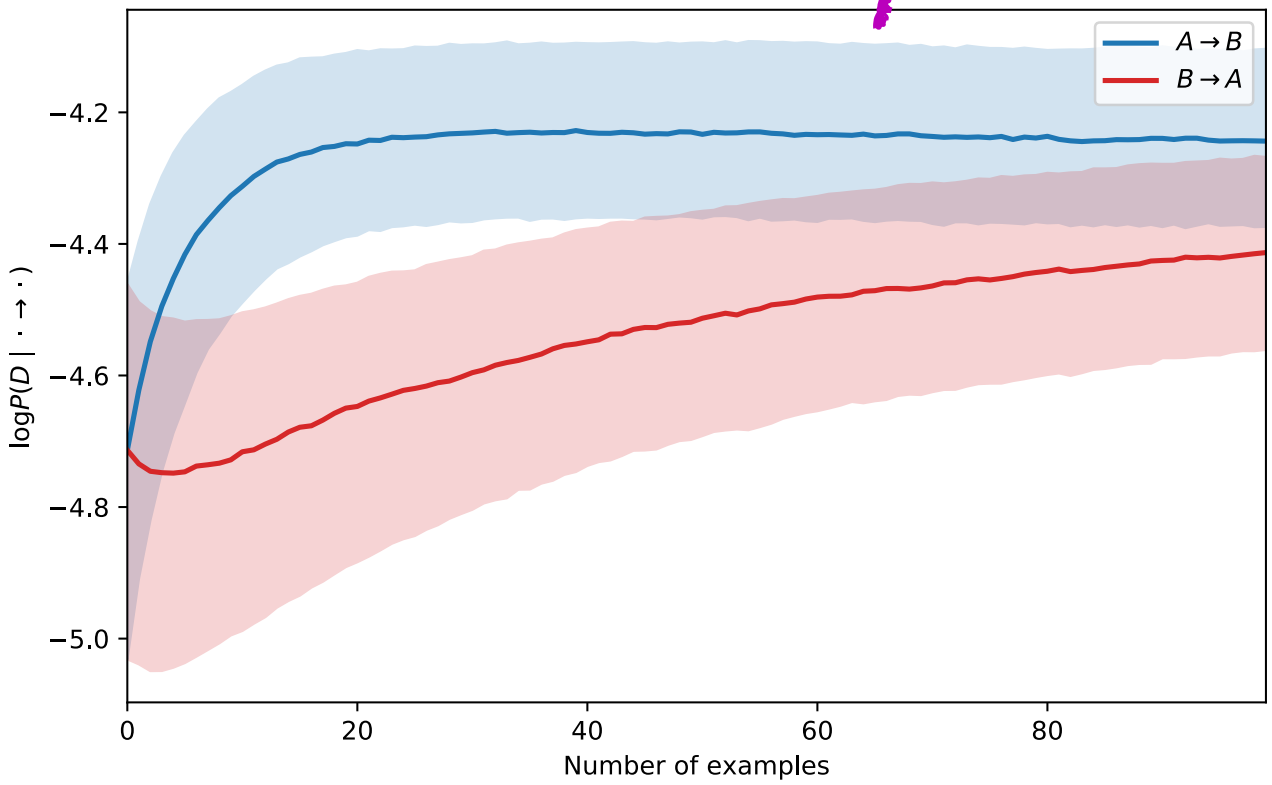
Mila

Wrong Knowledge Factorization Leads to Poor Transfer

- With the wrong factorization P(B) P(A|B), a change in P(A) influences all the modules, all the parameters
  - poor transfer: all the parameters must be adapted
- This is the normal situation with standard neural nets: every parameter participates to every relationship between all the variables
  - this causes ***catastrophic forgetting, poor transfer, difficulties with continual learning or domain adaptation, etc***

Mila

Empirical Confirmation: Correct Causal Structure Leads to Faster Adaptation



A→B is the correct causal structure: faster online adaptation to modified distribution = lower NLL regret

Mila

Turning a Hindrance into a Useful Signal

ArXiv paper, Bengio et al 2019: *A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms*

- Changes in distribution (nonstationarities in agent learning, transfer scenarios, etc) are seen as a bug in ML, a challenge
- Turn them into a feature, an asset, to help discover causal structure, or more generally to help **factorize knowledge:**
- **Tune knowledge factorization (e.g. causal structure) to maximize fast transfer**
- *"Nature does not shuffle environments, we shouldn't"*
  L. Bottou

Mila

Meta-Learning / Learning to Learn

- Generalize the idea of hyper-parameter optimization
  - Inner loop optimization (normal training), a fn of meta-params

$$\theta_t(\omega) = \text{approxmin}_\theta C(\theta, \omega, \mathcal{D}^t_{train})$$

  - Outer loop optimization (meta-training), optimize meta-params

$$\omega = \text{approxmin}_\omega \sum_t L(\theta_t(\omega), \omega, \mathcal{D}^t_{test})$$

- Meta-parameters can be the learning rule itself (Bengio & Bengio 1991; Schmidhuber 1992), learn 2 optimize
- Meta-learn an objective or reward function, or a shared encoder
- Meta-learning can be used to learn to generalize or transfer
- Can backprop through θ_t, use RL, evolution, or other tricks

Meta-Learning Objective

- Each transfer adaptation episode of length T
- Online log-likelihood for episode-wise mixture between 2 hypotheses:

$$\mathcal{R} = -\log\left[\sigma(\gamma)\mathcal{L}_{A\to B} + (1-\sigma(\gamma))\mathcal{L}_{B\to A}\right]$$

$$\mathcal{L}_{A\to B} = \prod_{t=1}^{T} P_{A\to B}(a_t, b_t\,;\,\theta_t)$$

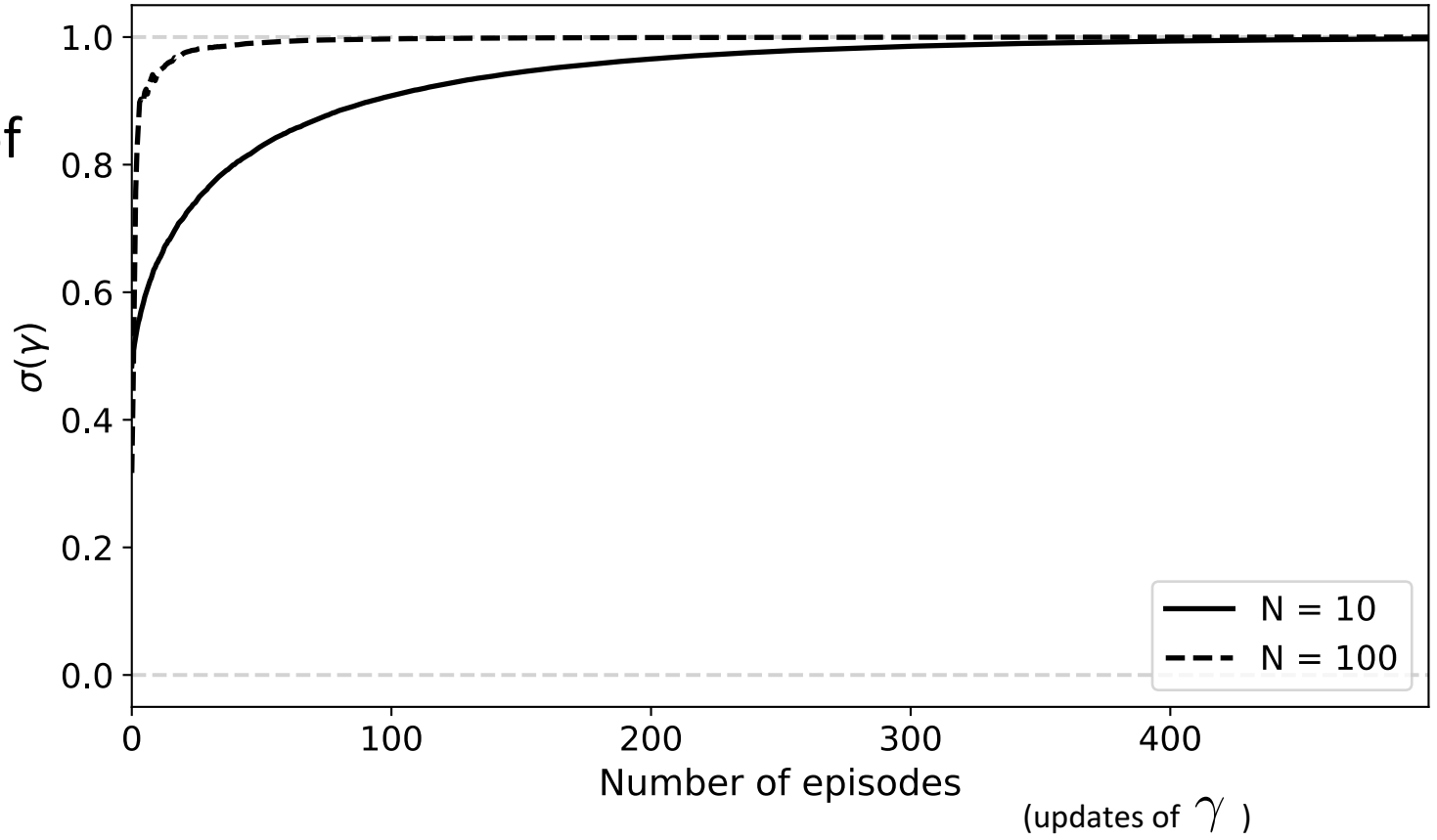
$$\mathcal{L}_{B\to A} = \prod_{t=1}^{T} P_{B\to A}(a_t, b_t\,;\,\theta_t),$$

Mila

Experimental Validation

Tabular parametrization of marginals and conditionals of bivariate model.

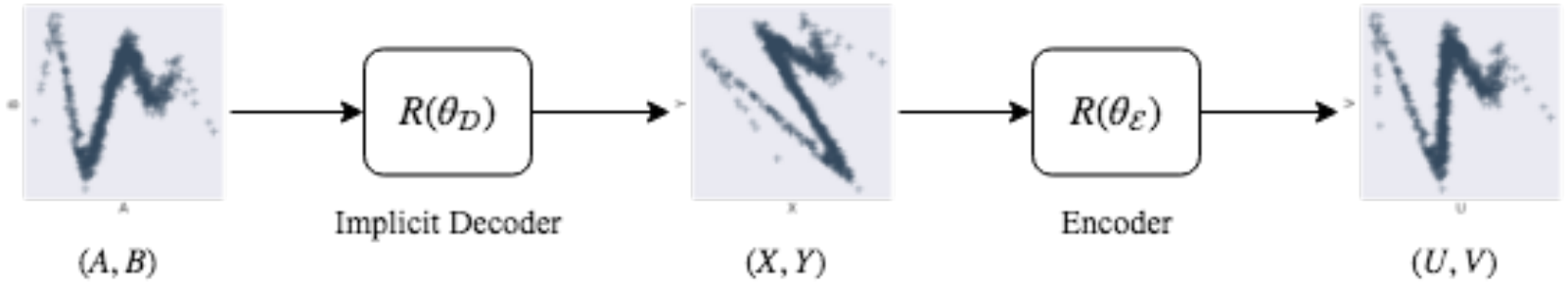
Correct causal graph can be recovered



Mila

Disentangling the Causes

- Realistic settings: causal variables are not directly observed
- Need to learn an encoder which maps raw data to causal space
- Consider both the encoder parameters and the causal graph structural parameters as meta-parameters trained together wrt proposed meta-transfer objective



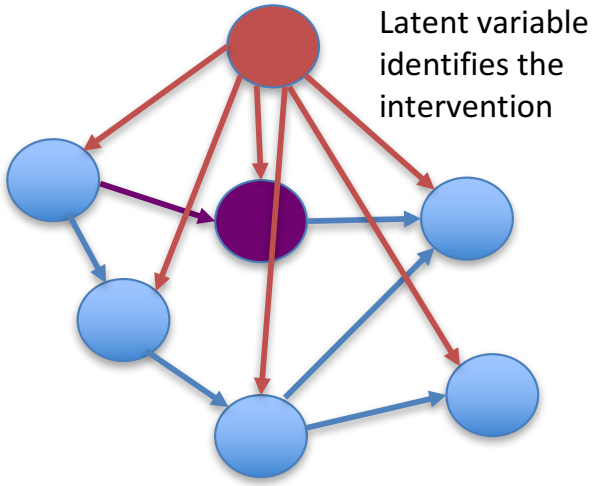
Mila

Simplest possible scenario: linear mixing (rotating decoder) and unmixing (rotating decoder)

Doing Inference on the Intervention

- To reduce the noise due to unnecessary adaptation of the unchanged modules, infer which variable was modified by the intervention: has worse relative log-likelihood after the intervention.

- This could be used to address catastrophic forgetting: infer if current distribution matches a previously seen one



Mila

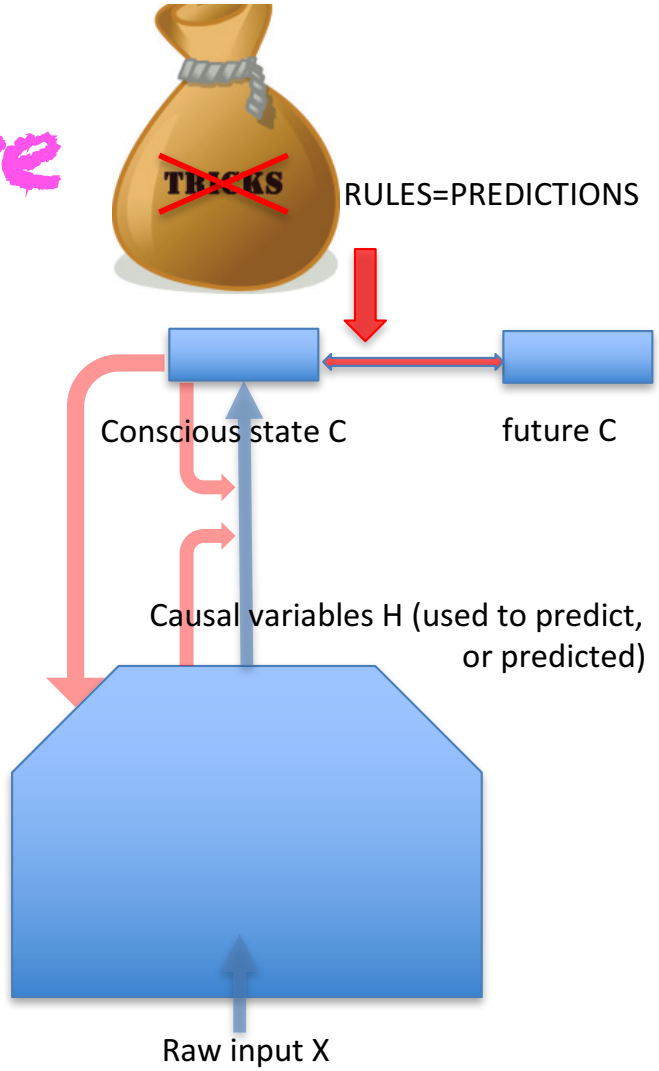
Bigger Picture

– Encoder maps sensory data to space where a few **sparse predictive rules** relate causal variables together, following the ***consciousness prior***

– ***Best graphical model assumption:***

sparse factor graph

– Need to handle unobserved state: P(H|X)

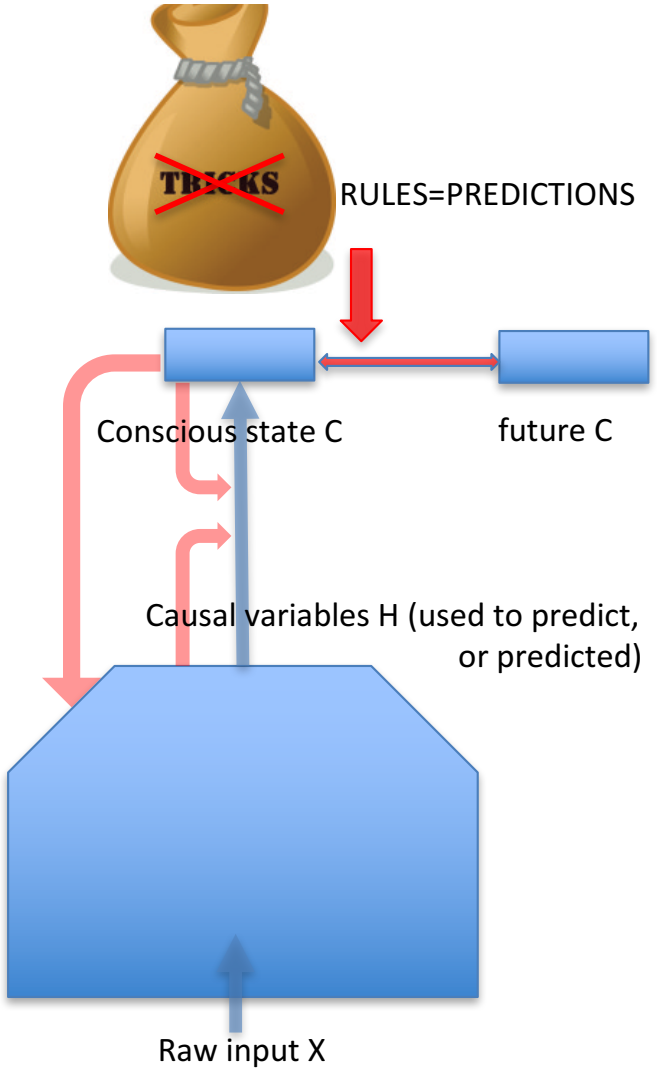


Sparse Causal Factor Graphs

$$P(V) \propto \prod_k \phi(V_{s_k})$$

where V_{s_k} is the subset of V with indices s_k

- Many rules; rule = factor = soft constraint on a small set of variables
- Directed graphical model: module = conditional distribution
- Sparse causal factor graph: module = soft constraint on a small set of variables
- Add causal edge direction to each factor



Observing Other Agents

- Can infants figure out causal structure in spite of being almost passive observers?
- Yes, if they exploit and infer the interventions made by other agents
- Our approach does not require the learner to know what the action/intervention was (but it could do inference over interventions)
- But more efficient learning if you can experiment and thus test hypotheses about cause & effect

Looking Forward

- Build a world model which captures causal effects in abstract space of causal variables, able to quickly adapt to changes in the world and generalize out-of-distribution
- Acting to acquire that knowledge (exploratory behavior)
- Bridging the gap between system 1 and system 2, old neural nets and conscious reasoning, all neural

AI for Social Good

- Beyond developing the next gadget
- AI is powerful, can be misused or bring much good
- Actionable items:
  - Favor ML applications which help the poorest countries, may help with fighting climate change, improve healthcare, education, etc.
  - ***AI Commons***: an organization in construction, which will coordinate, prioritize and channel funding for such applications

XPRIZE Mila ITU HEC PARIS CHAI Center for Human-Compatible Artificial Intelligence IEEE Standards

fondation BOTNAR PONTIFICIA ACADEMIA PRO VITA ocean INSPIRED MINDS camera culture McGovern FOUNDATION

Amir Banifatemi