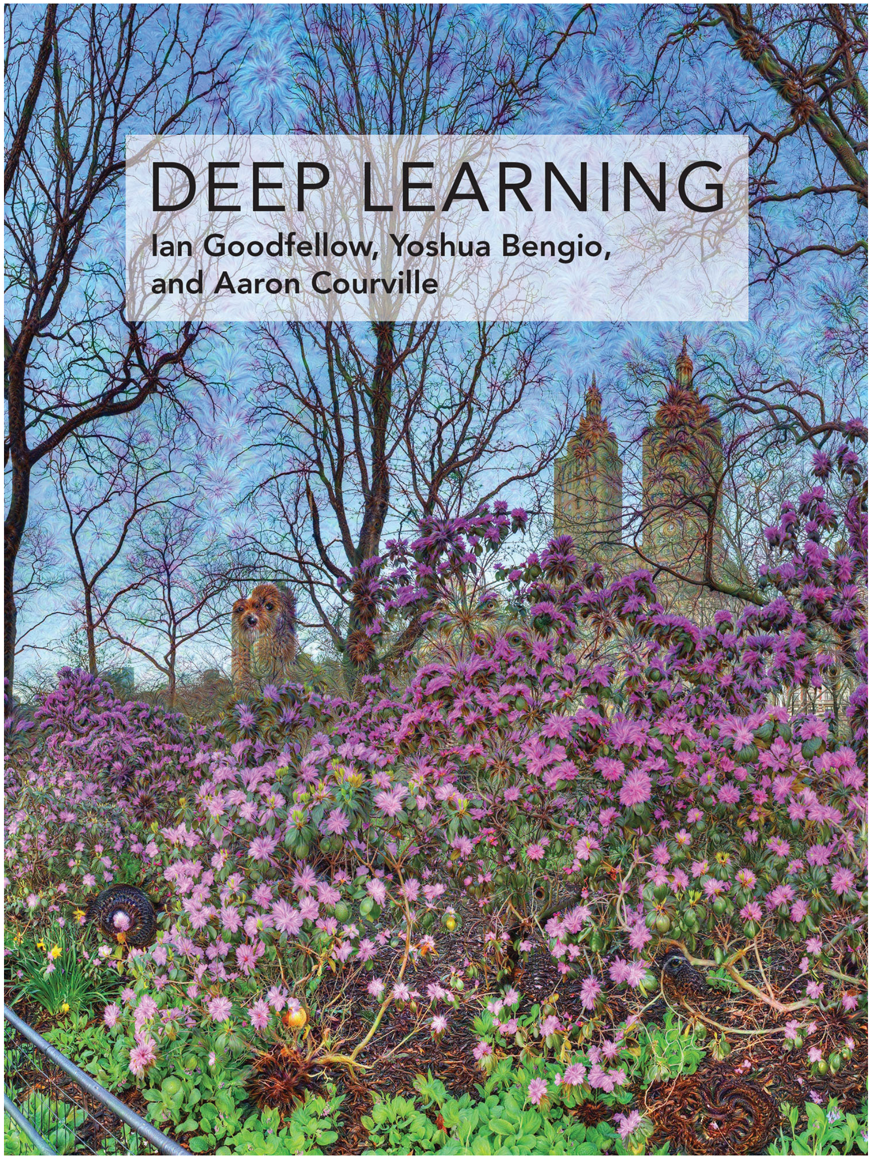


DEEP LEARNING FOR SYSTEM 2 PROCESSING

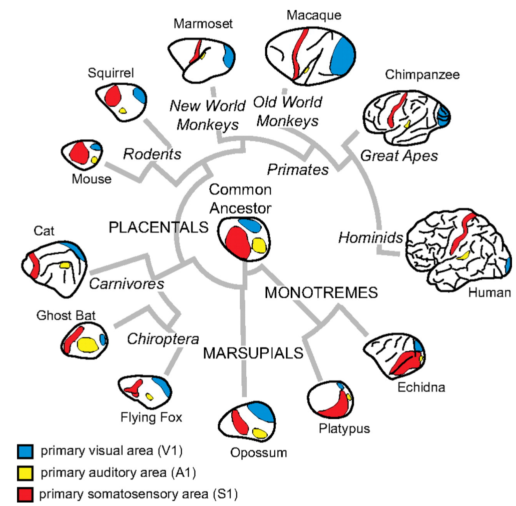
YOSHUA BENGIO

AAAI'2019 Invited Talk
February 9th, 2020, New York City

Mila
Université de Montréal
CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH
ICRA
INSTITUT CANADIEN DE RECHERCHES AVANCÉES

NO-FREE-LUNCH THEOREM, INDUCTIVE BIASES & HUMAN-LEVEL AI

- **No-free-lunch theorem →** there is no completely general intelligence, some inductive biases / priors are necessary
- **Generality & discoverability:** simpler less specialized priors are however more likely to be discovered by evolution and applicable to a broader set of contexts
- **Deep learning** already incorporates human-inspired priors
  - *Computation as composition of simpler pieces, neurons in layers, layers over layers (Pascanu et al ICLR 2014; Montufar et al NeurIPS 2014)*
  - *More powerful priors can bring up to an exponential advantage in sample complexity*



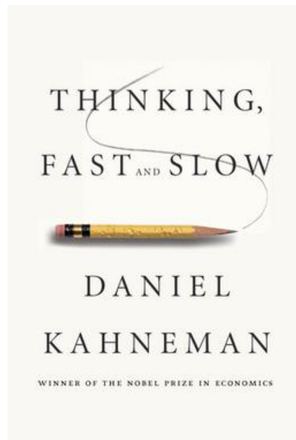
SYSTEM 1 VS. SYSTEM 2 COGNITION

2 systems (and categories of cognitive tasks):

Manipulates high-level / semantic concepts, which can be recombined combinatorially

System 1

- Intuitive, fast, **UNCONSCIOUS**, non-linguistic, habitual
- Current DL



System 2

- Slow, logical, sequential, **CONSCIOUS**, linguistic, algorithmic, planning, reasoning
- Future DL

MISSING TO EXTEND DEEP LEARNING TO REACH **HUMAN-LEVEL AI**

- **Out-of-distribution generalization & transfer**
- **Higher-level cognition: system 1 $\rightarrow$ system 2**
  - *High-level semantic representations*
  - *Compositionality*
  - *Causality*
- **Agent perspective:**
  - *Better world models*
  - ***Causality***
  - *Knowledge-seeking*
- **Connections between all 3 above!**

HYPOTHESES FOR **CONSCIOUS PROCESSING BY AGENTS, SYSTEMATIC GENERALIZATION**

- *Sparse factor graph in space of high-level semantic variables*
- *Semantic variables are causal: agents, intentions, controllable objects*
- Shared 'rules' across instance tuples (arguments)
- *Distributional changes due to localized causal interventions (in semantic space)*
- Meaning (e.g. grounded by an encoder) is stable & robust wrt changes in distribution
- Credit assignment is only over short causal chains

Proposal: what may be the evolutionary advantage of system 2 processing?

DEALING WITH CHANGES IN DISTRIBUTION

Mila

AGENT LEARNING NEEDS **OOD GENERALIZATION**

Agents face non-stationarities

Changes in distribution due to

- their actions
- ***ESPECIALLY:***
  actions of other agents
- different places, times, sensors, actuators, goals, policies, etc.

Multi-agent systems: many changes in distribution
Ood generalization needed for continual learning

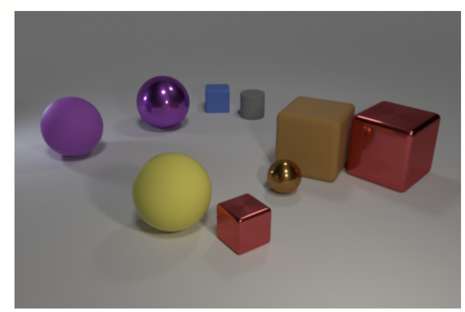
Mila

SYSTEMATIC GENERALIZATION

- Studied in linguistics
- **Dynamically recombine existing concepts**
- Even when new combinations have 0 probability under training distribution
  - E.g. Science fiction scenarios
  - E.g. Driving in an unknown city
- Not very successful with current DL

(Lake et al 2015)

(Lake & Baroni 2017)
(Bahdanau et al & Courville ICLR 2019)
CLOSURE: (Bahdanau et al & Courville arXiv:1912.05783) on CLEVR



Mila

CONTRAST WITH **THE SYMBOLIC AI PROGRAM**

Avoid pitfalls of classical AI rule-based symbol-manipulation

- Need efficient large-scale learning
- Need semantic grounding in system 1
- Need distributed representations for generalization
- Need efficient = trained search (also system 1)
- Need uncertainty handling

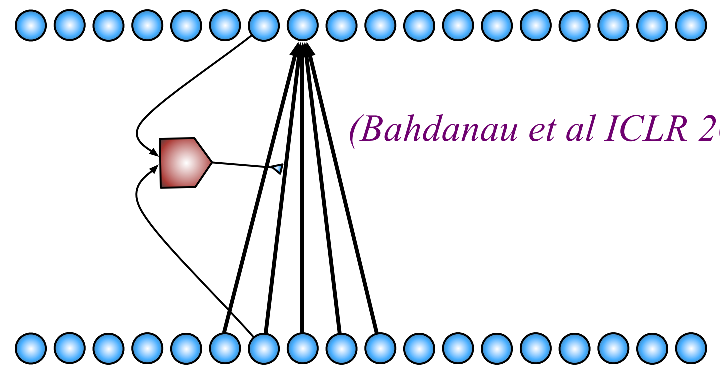
But want

- Systematic generalization
- Factorizing knowledge in small exchangeable pieces
- Manipulating variables, instances, references & indirection

Mila

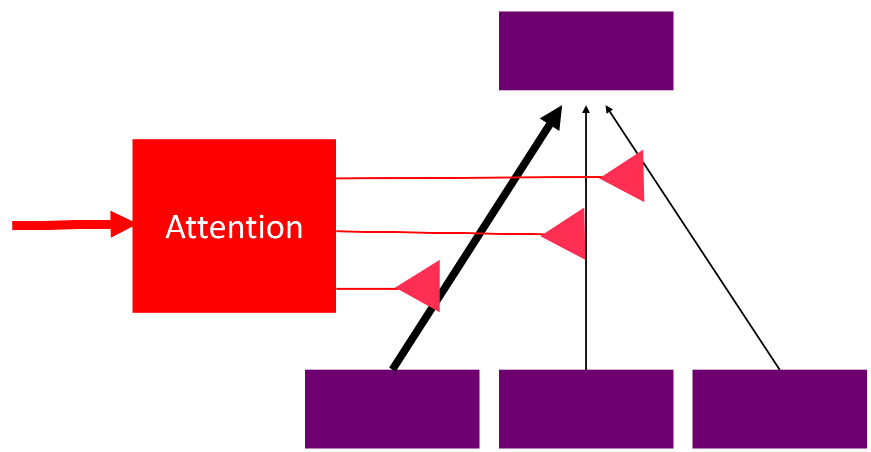
SYSTEM 2 BASICS: ATTENTION AND CONSCIOUS PROCESSING

CORE INGREDIENT FOR CONSCIOUS PROCESSING: **ATTENTION**



(Bahdanau et al ICLR 2015)

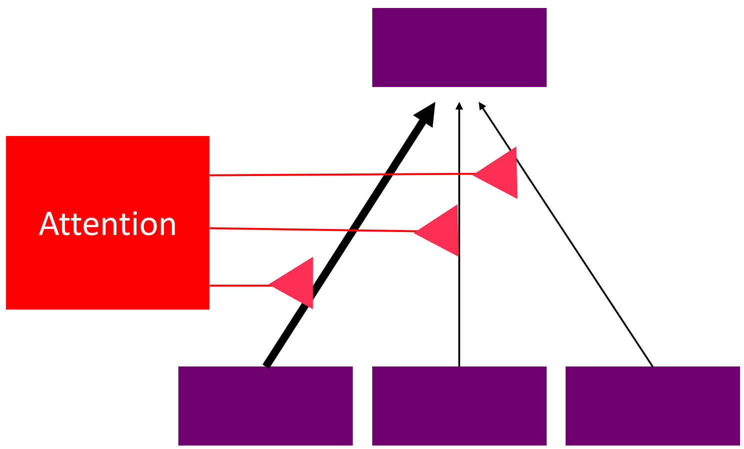
- **Focus** on a one or a few elements at a time
- **Content-based soft attention** is convenient, can backprop to *learn where to attend*
- Attention is an **internal action**, needs a **learned attention policy** *(Egger et al 2019)*



- Operating on unordered SETS of (key, value) pairs
- SOTA in NLP

Mila

FROM ATTENTION TO **INDIRECTION**



- Attention = dynamic connection
- Receiver gets the selected value
- Value of what? From where?

  → Also send ‘name’ (or key) of sender
- Keep track of 'named’ objects: indirection
- Manipulate sets of objects (transformers)

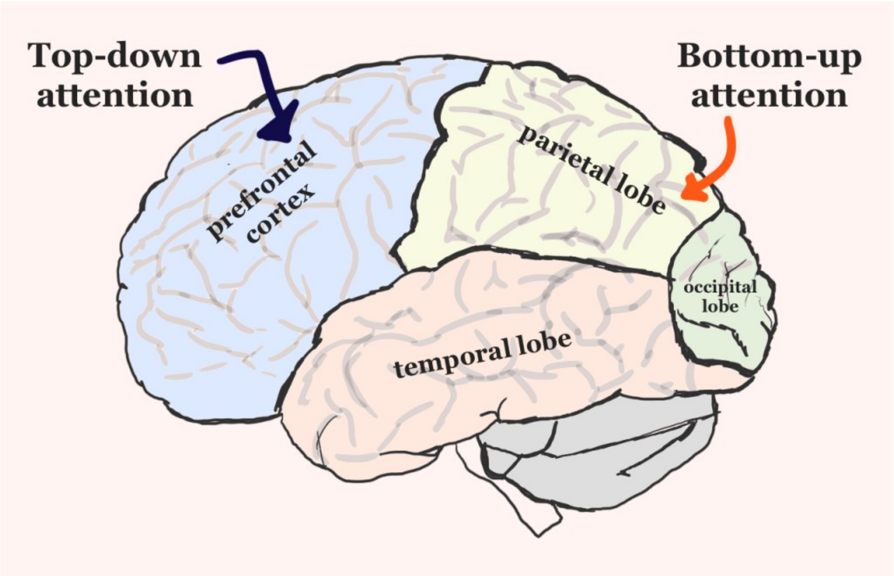
FROM ATTENTION TO **CONSCIOUSNESS**

C-word not taboo anymore in cognitive neuroscience

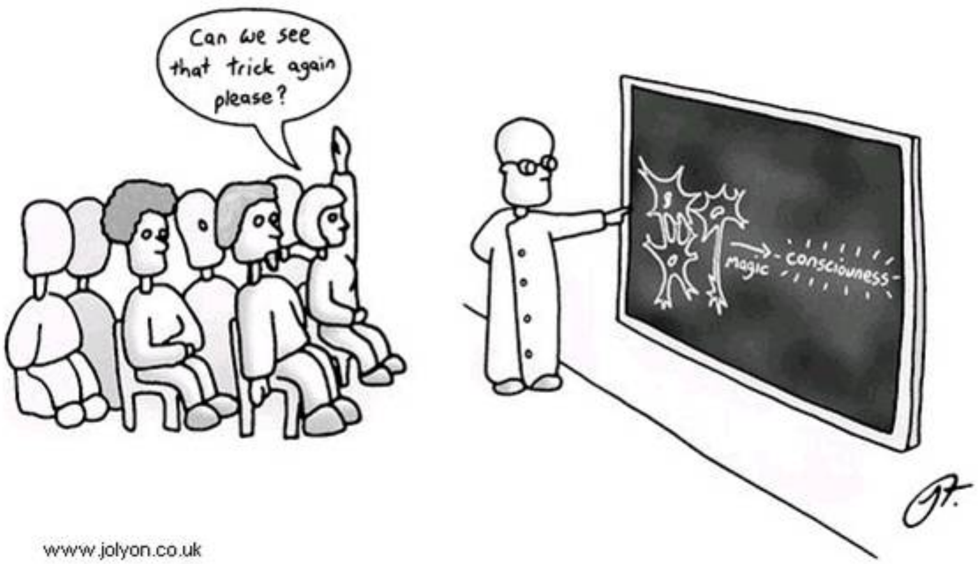
Global Workspace Theory

(Baars 1988++, Dehaene 2003++)

- Bottleneck of conscious processing
  - *WHY A BOTTLENECK?*
- Selected item is broadcast, stored in short-term memory, conditions perception and action
- System 2-like sequential processing, conscious reasoning & planning & imagination



ML FOR CONSCIOUSNESS & CONSCIOUSNESS FOR ML



- Formalize and test **specific hypothesized functionalities of consciousness**
- Get the magic out of consciousness
- Understand evolutionary advantage of consciousness: computational and statistical (e.g. systematic generalization)
- Provide these advantages to learning agents

Mila

THOUGHTS, CONSCIOUSNESS, LANGUAGE

- Consciousness: from humans reporting
- High-level representations ⟺ language
- High-level concepts: meaning anchored in low-level perception and action → **tie system 1 & 2**
- Grounded high-level concepts

  → better natural language understanding

- **Grounded language learning**

  e.g. BabyAI: *(Chevalier-Boisvert and al ICLR 2019)*

WHY A CONSCIOUSNESS BOTTLENECK?

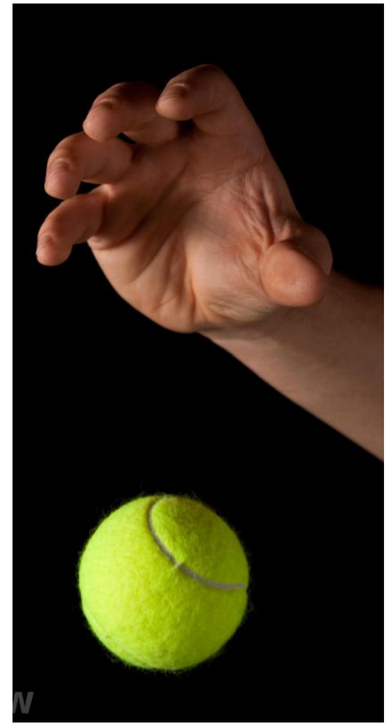
THE CONSCIOUSNESS PRIOR

= SPARSE FACTOR GRAPH

Mila

CONSCIOUSNESS **PRIOR**
➔ **SPARSE FACTOR GRAPH**

Bengio 2017, arXiv:1709.08568



- Property of **high-level variables which we manipulate with language**:

  we can predict some given very few others

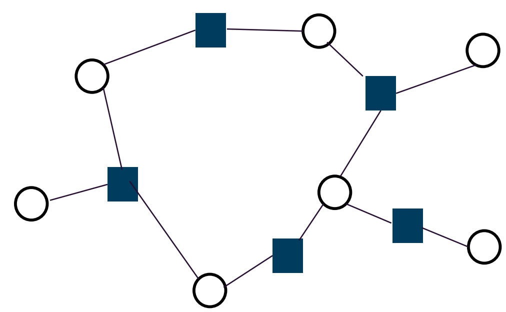
  - E.g. "if I drop the ball, it will fall on the ground"

- **Disentangled factors** != marginally independent,

  e.g. ball & hand

- **Prior**: sparse factor graph joint distribution between high-level variables

- Inference involves few variables at a time, selected by **attention mechanism** and memory retrieval

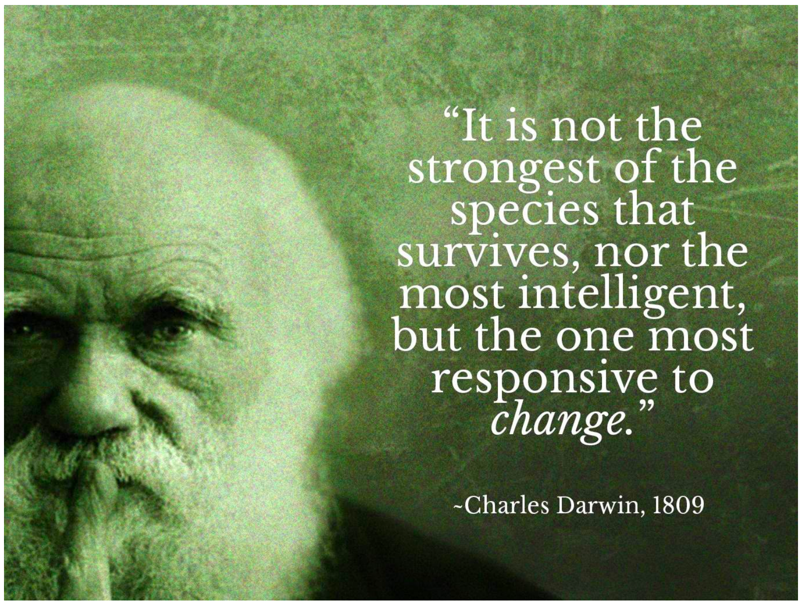


Mila

META-LEARNING: END-TO-END OOD GENERALIZATION, *SPARSE CHANGE PRIOR*

Mila

META-LEARNING FOR TRAINING TOWARDS OOD GENERALIZATION



- Meta-learning or learning to learn
  (Bengio et al 1991; Schmidhuber 1992)
  - Backprop through inner loop or REINFORCE-like estimators
- Bi-level optimization
  - Inner loop (may optimize something) → outer loss
  - Outer loop: optimizes E[outer loss] (over tasks, environments)
- E.g.
  - Evolution ∘ individual learning
  - Lifetime learning ∘ fast adaptation to new environments
- Multiple time-scales of learning
- **End-to-end learning to generalize ood + fast transfer**

Mila

WHAT **CAUSES** CHANGES IN DISTRIBUTION?

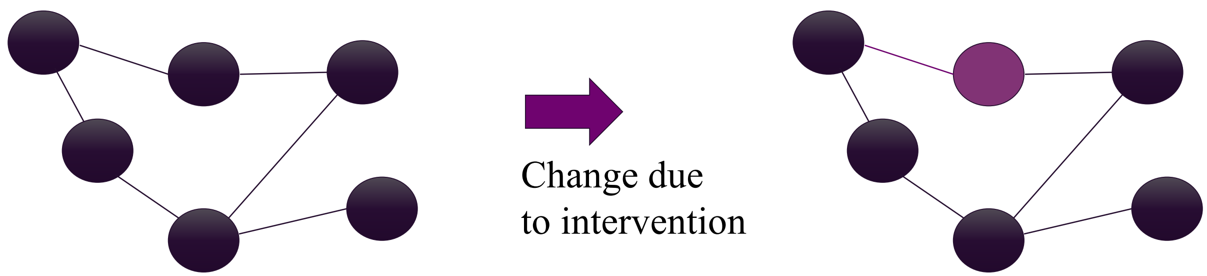
Underlying physics: actions are localized in space and time.

Hypothesis to replace iid assumption:

changes = consequence of an intervention on few causes or mechanisms

Extends the hypothesis of (informationally) Independent Mechanisms *(Scholkopf et al 2012)*

➔ **local inference or adaptation in the right model**



Mila

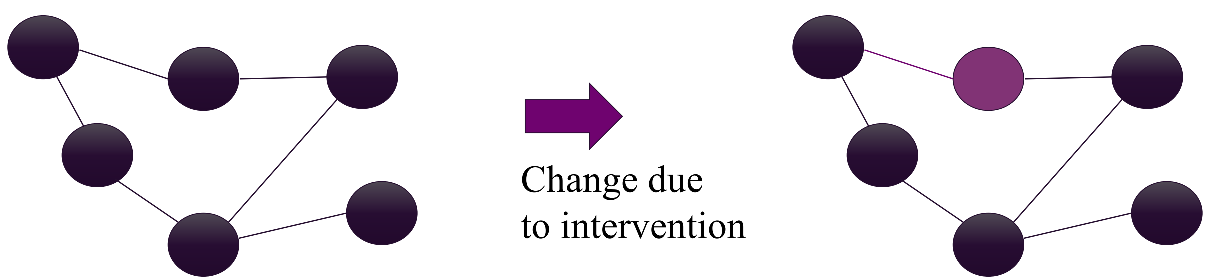
COUNTING ARGUMENT: **LOCALIZED CHANGE→OOD TRANSFER**

Good representation of variables and mechanisms + localized change hypothesis

→ few bits need to be accounted for (by inference or adaptation)

→ few observations (of modified distribution) are required

→ good ood generalization/fast transfer/small ood sample complexity



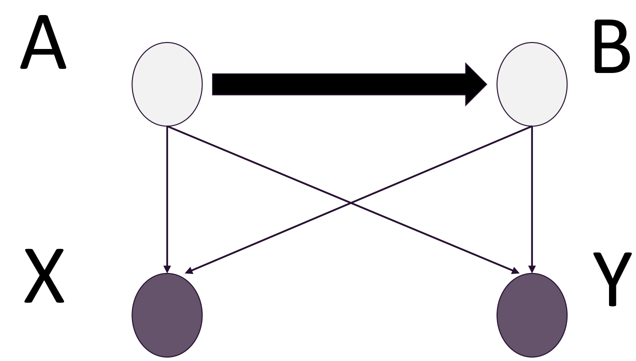
META-LEARNING KNOWLEDGE REPRESENTATION FOR GOOD OOD PERFORMANCE

- Use ood generalization as training objective
- Good decomposition / knowledge representation ➔ good ood performance
- Good ood performance = training signal for factorizing knowledge

Mila

EXAMPLE: DISCOVERING CAUSE AND EFFECT
= **HOW TO FACTORIZE A JOINT DISTRIBUTION?**

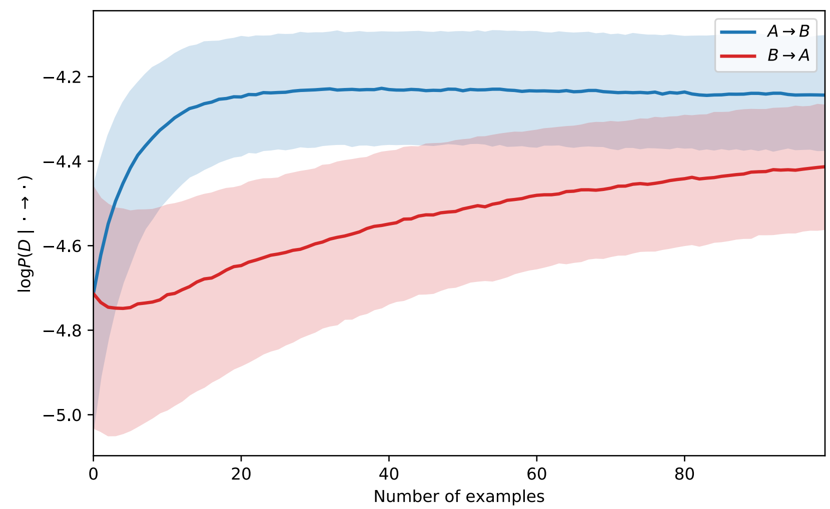
A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms



- Learning whether A causes B or vice-versa
- Learning to disentangle (A,B) from observed (X,Y)
- Exploit changes in distribution and speed of adaptation to guess causal direction

Bengio et al 2019 arXiv:1901.10912

- *Ongoing work: theory proving when the correct model converges faster by online SGD*



Mila

EXAMPLE: DISCOVERING CAUSE AND EFFECT
= HOW TO FACTORIZE A JOINT DISTRIBUTION?

Learning Neural Causal Models from Unknown Interventions *Ke et al 2019 arXiv:1910.01075*

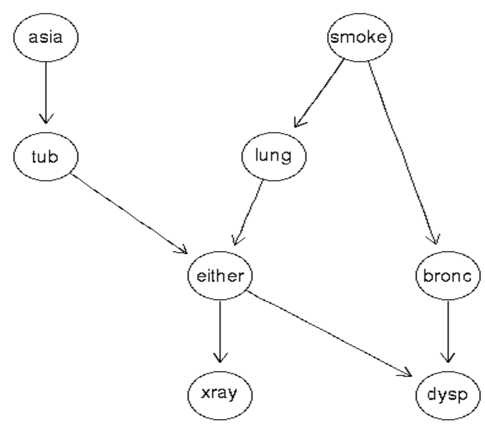
- Learning small causal graphs, avoid exponential explosion of # of graphs by parametrizing factorized distribution over graphs

- With enough observations of changes in distribution: perfect recovery of the causal graph without knowing the intervention; converges faster on sparser graphs

- Inference over the intervention:
  faster causal discovery

Asia graph, CE on ground truth edges, comparison against other causal induction methods

Our method	(Eaton & Murphy, 2007a)	(Peters et al., 2016)	(Zheng et al., 2018)
0.0	0.0	10.7	3.1



Mila

Consequence of the consciousness prior (sparse factor graph):

OPERATING ON SETS OF POINTABLE OBJECTS WITH DYNAMICALLY RECOMBINED MODULES

Mila

RIMS: MODULARIZE COMPUTATION AND OPERATE ON SETS OF NAMED AND TYPED OBJECTS

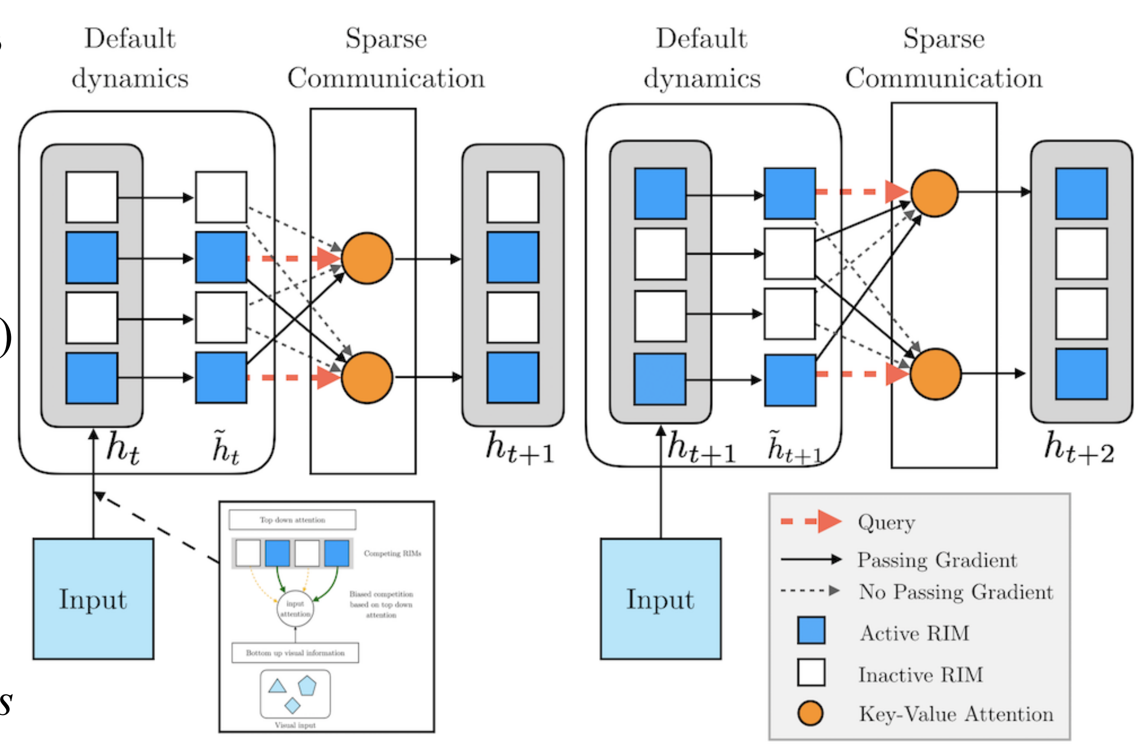
Recurrent Independent Mechanisms

Goyal et al 2019, arXiv:1909.10893

Multiple recurrent sparsely interacting modules, each with their own dynamics, with object (key/value pairs) input/outputs selected by multi-head attention

Results: better ood generalization

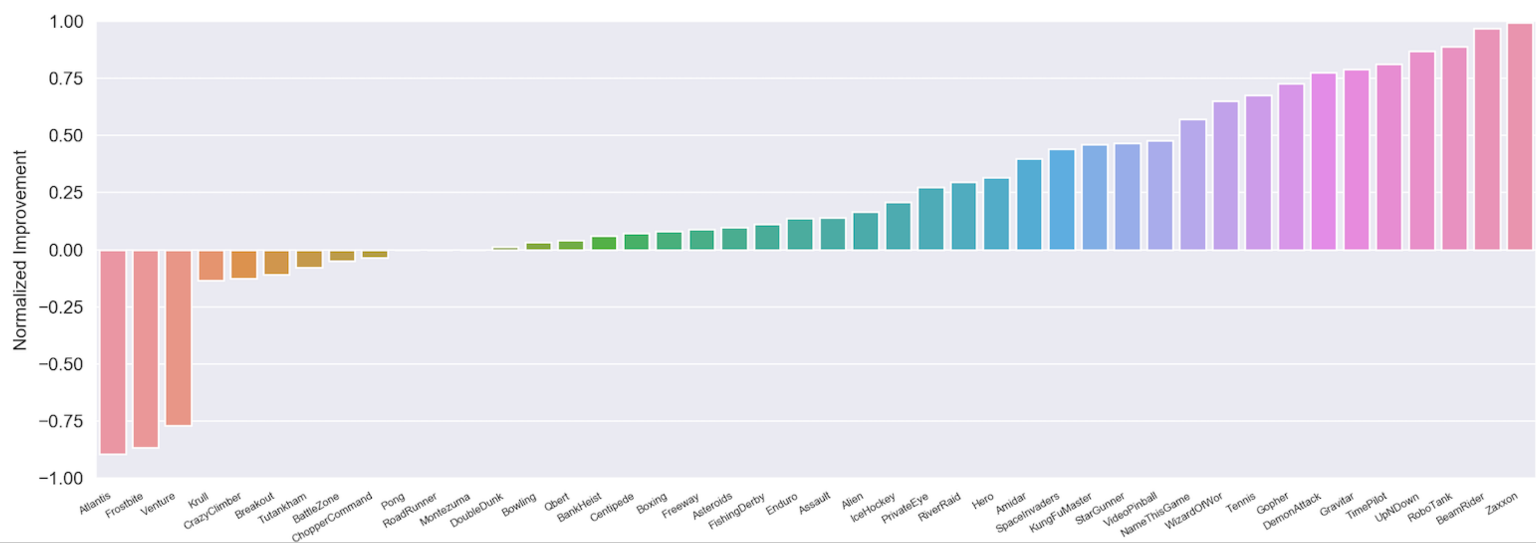
Ongoing work: hierarchy, top-down broadcasting, spatial layout of modules



Builds on rich recent litterature on object-centric representations (mostly for images)

RESULTS WITH **RECURRENT INDEPENDENT MECHANISMS**

- RIMs drop-in replacement for LSTMs in PPO baseline over all Atari games.
- Above 0 (horizontal axis) = improvement over LSTM.



Mila

HYPOTHESES FOR **CONSCIOUS PROCESSING BY AGENTS, SYSTEMATIC GENERALIZATION**

- *Sparse factor graph in space of high-level semantic variables*
- *Semantic variables are causal: agents, intentions, controllable objects*
- Shared ’rules’ across instance tuples (arguments)
- *Distributional changes due to localized causal interventions (in semantic space)*
- Meaning (e.g. grounded by an encoder) is stable & robust wrt changes in distribution
- Credit assignment is only over short causal chains

CONCLUSIONS

- After cog. neuroscience, time is ripe for ML to explore consciousness
- Could bring new priors to help systematic & ood generalization
- Could benefit cognitive neuroscience too
- Would allow to expand DL from system 1 to system 2

System 1


System 2

Mila


THANK YOU

Mila

Université de Montréal

McGill

Québec

CIFAR