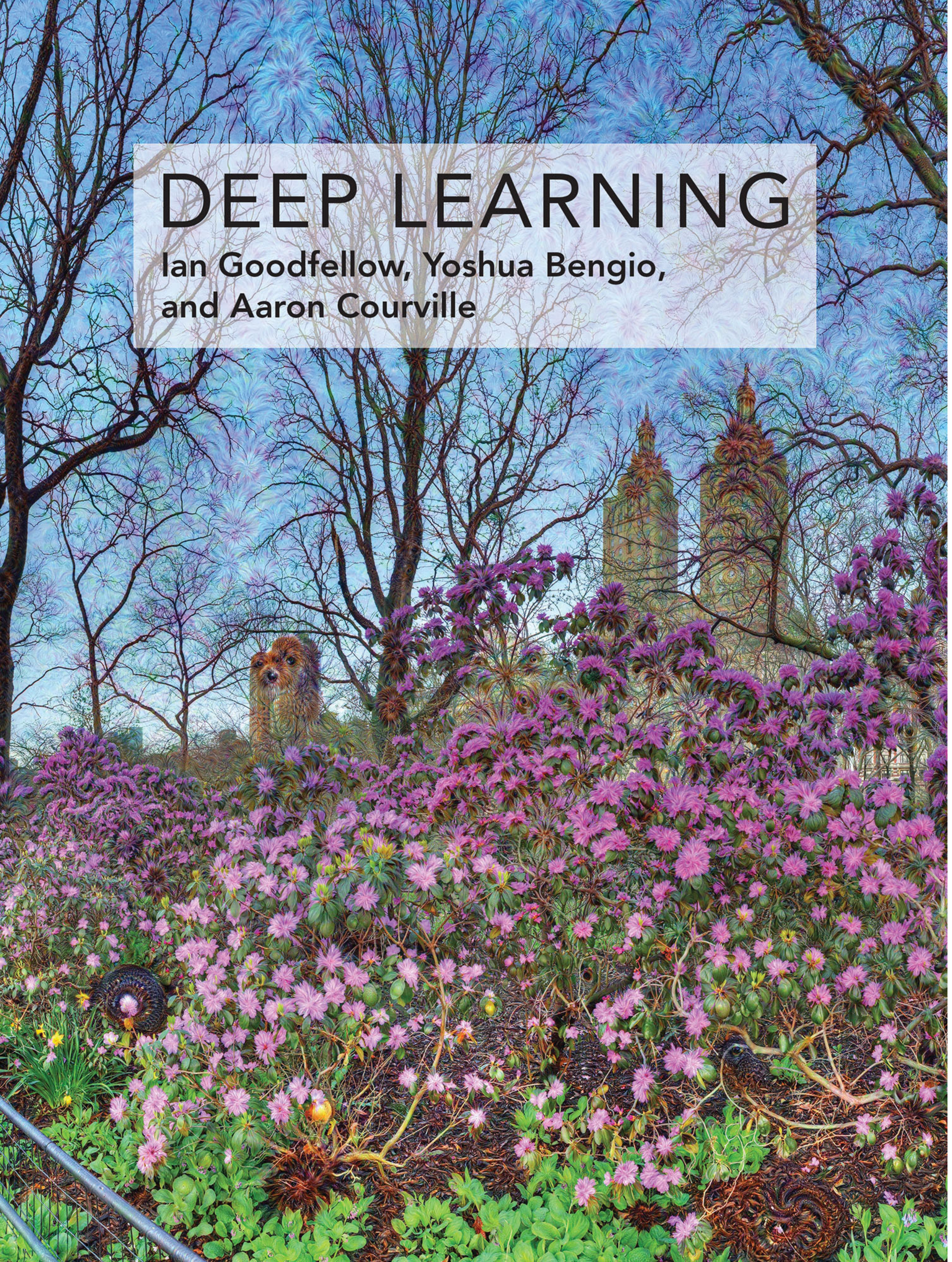

DEEP LEARNING

Ian Goodfellow, Yoshua Bengio, and Aaron Courville

TOWARDS COMPOSITIONAL UNDERSTANDING OF THE WORLD BY AGENT-BASED DEEP LEARNING

YOSHUA BENGIO

NeurIPS'2019 Workshop on Context and Compositionality in Biological and Artificial Neural Networks
December 14th, 2019, Vancouver BC

Mila

Université de Montréal

CIFAR
CANADIAN
INSTITUTE
FOR
ADVANCED
RESEARCH

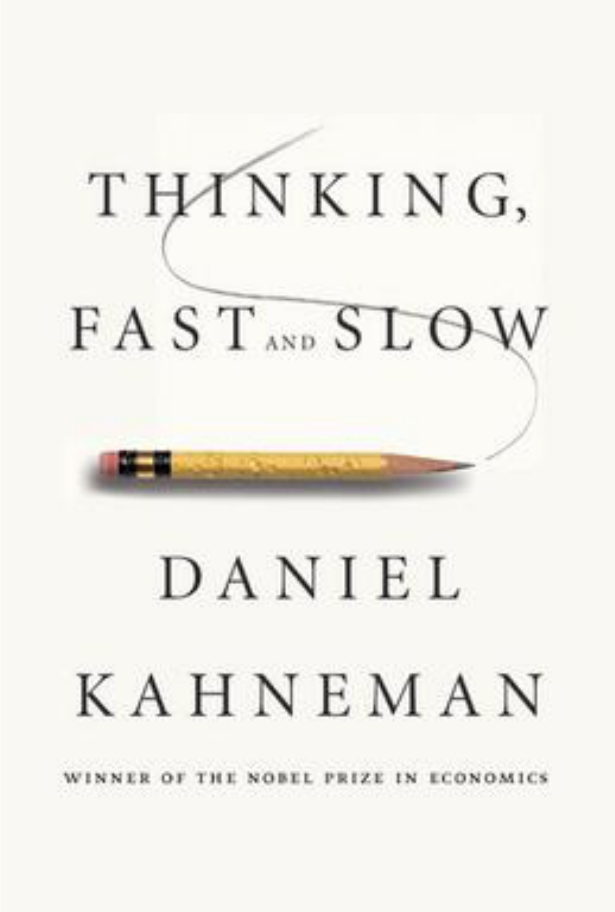
ICRA
INSTITUT
CANADIEN
DE
RECHERCHES
AVANCÉES

SYSTEM 1 VS. SYSTEM 2 COGNITION

2 systems (and categories of cognitive tasks):

System 1

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



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

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 → system 2**
  - *High-level semantic representations*
  - *Compositionality*
  - *Causality*
- **Agent perspective:**
  - *Better world models*
  - ***Causality***
  - *Knowledge-seeking*
- **Connections between all 3 above!**

DEALING WITH CHANGES IN DISTRIBUTION

AGENT LEARNING NEEDS **OOD GENERALIZATION**

Agents face non-stationarities

Changes in distribution due to

- their actions
- 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

COMPOSITIONALITY HELPS IID AND OOD GENERALIZATION

Different forms of compositionality
each with different exponential advantages

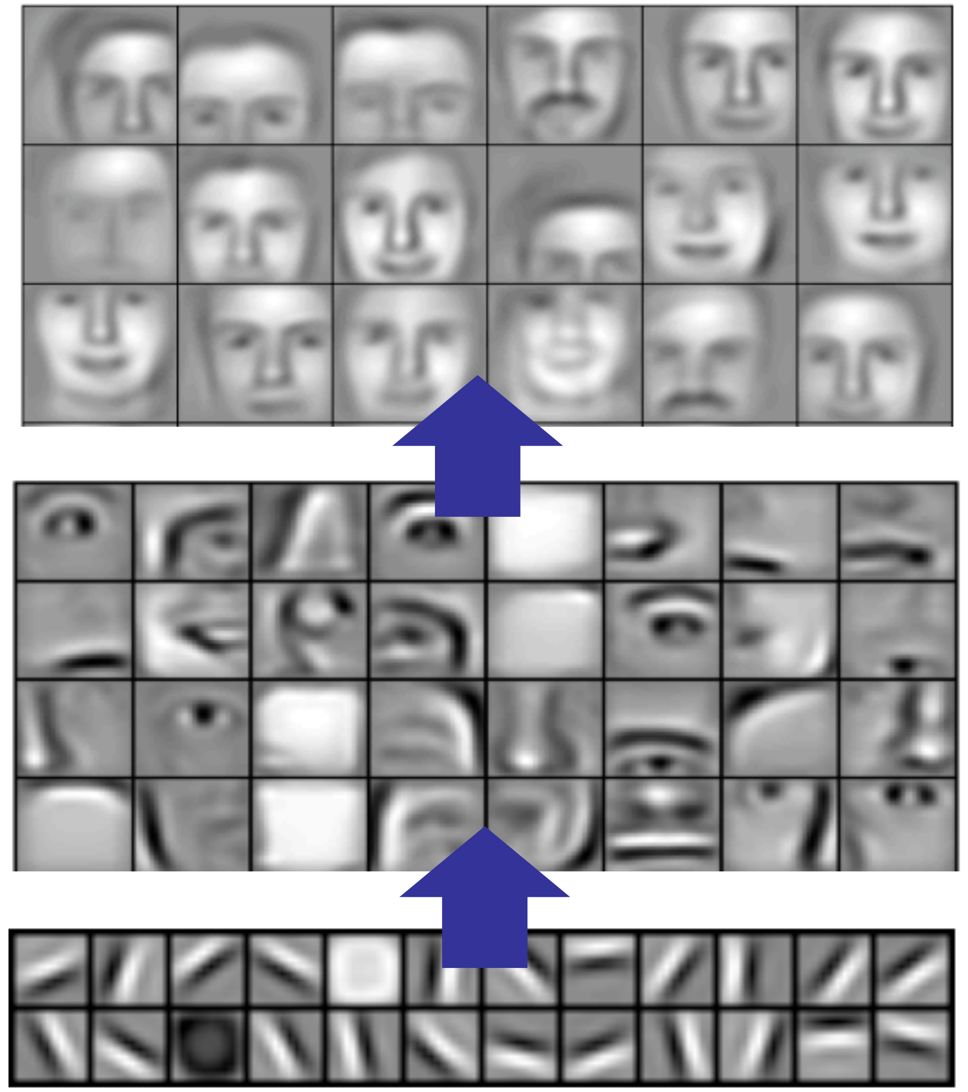
- Distributed representations

  (Pascanu et al ICLR 2014)

- Composition of layers in deep nets

  (Montufar et al NeurIPS 2014)

- **Systematic generalization in language, analogies, abstract reasoning? TBD**

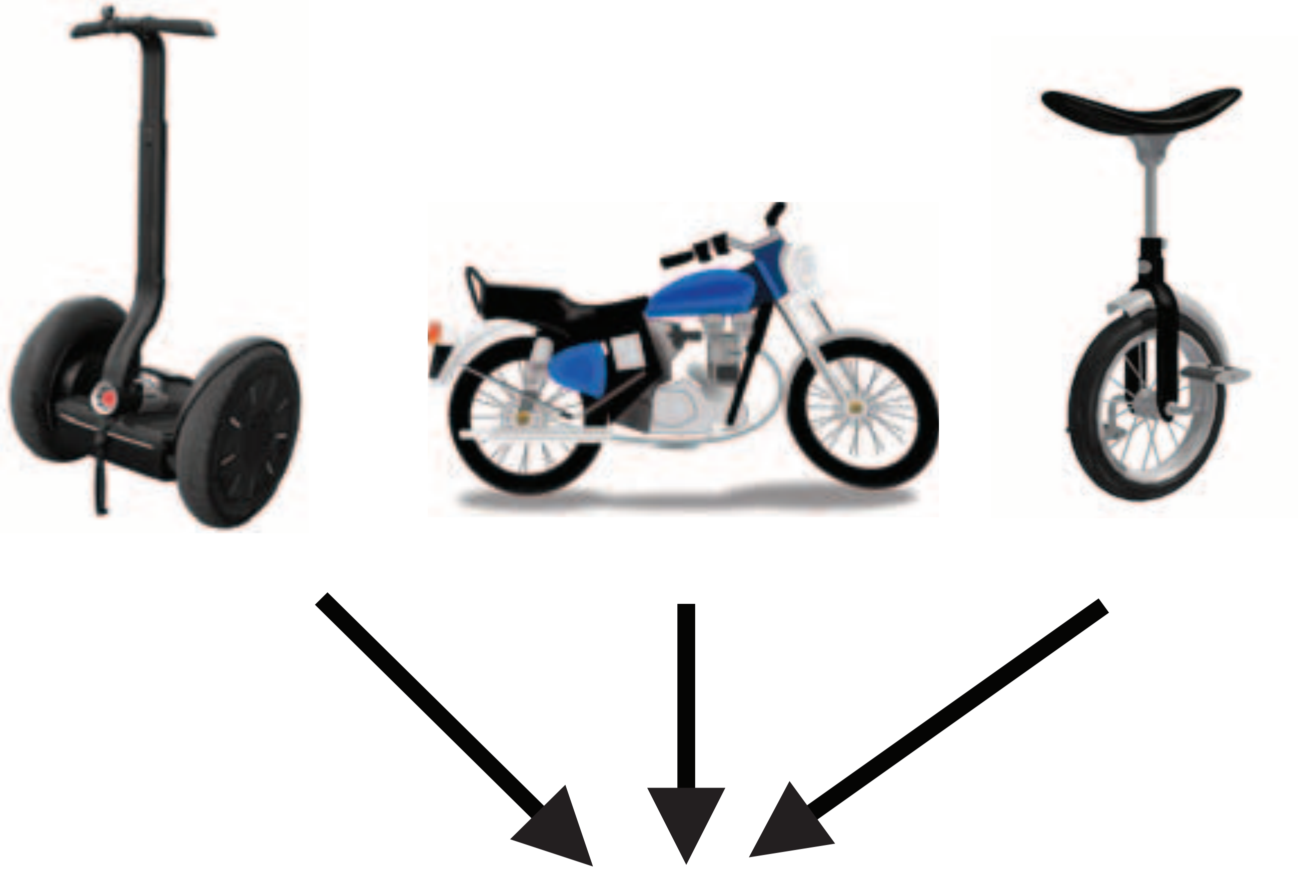


(Lee, Grosse, Ranganath & Ng, ICML 2009)

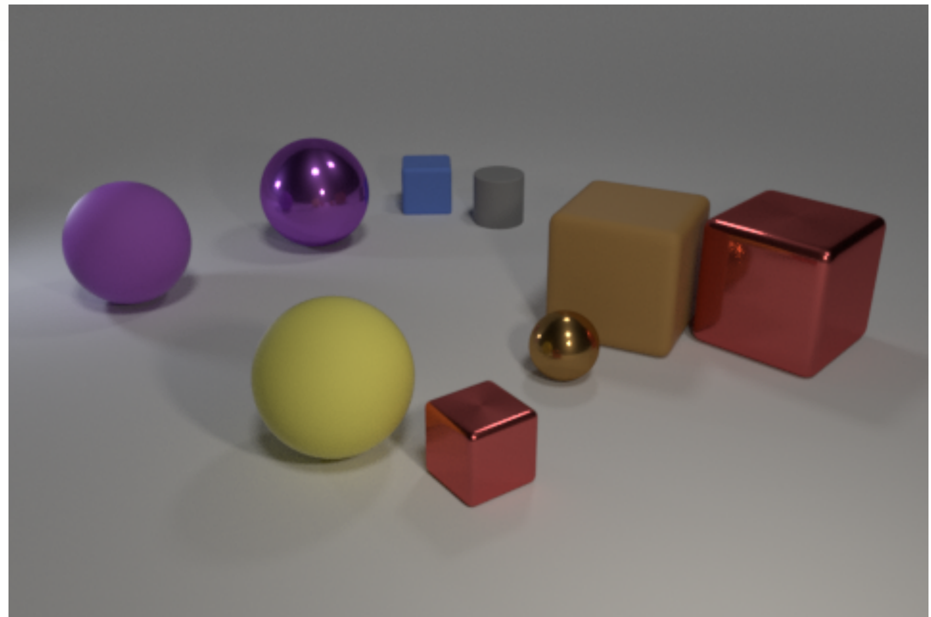
SYSTEMATIC GENERALIZATION

- 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 & Baroni 2017)
(Bahdanau et al & Courville ICLR 2019)
CLOSURE: ongoing work by Bahdanau et al & Courville on CLEVR

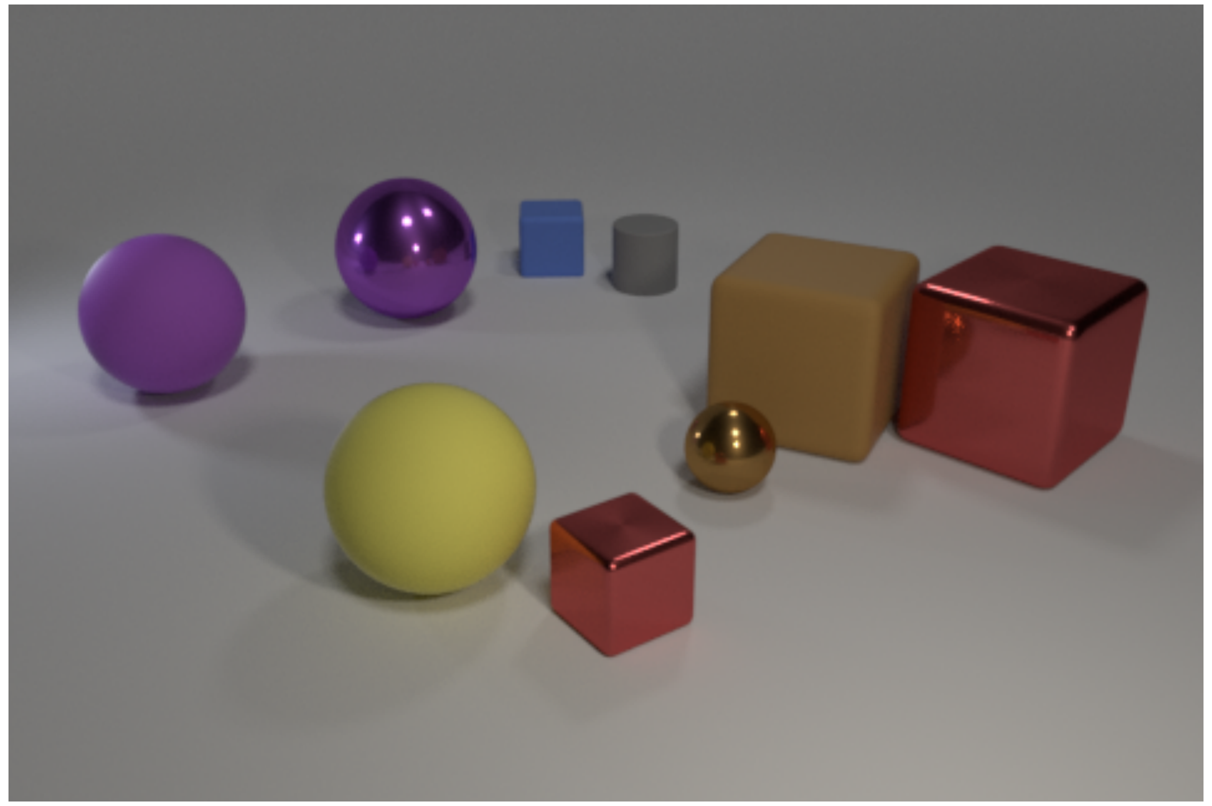


(Lake et al 2015)



CLOSURE: Known Referring Expressions in Novel Contexts

CLOSURE: Assessing Systematic Generalization of CLEVR Models Bahdanau et al, ArXiV)



Q1 (CLEVR): There is another cube that is the same size as the brown cube; what is its color?

a matching referring expression

Q2 (CLEVR): There is a thing that is in front of the yellow thing; does it have the same color as cylinder?

a comparison question

Q3 (CLOSURE):There is a rubber object that is the same size as the gray cylinder; does it have the same color as the tiny shiny block?

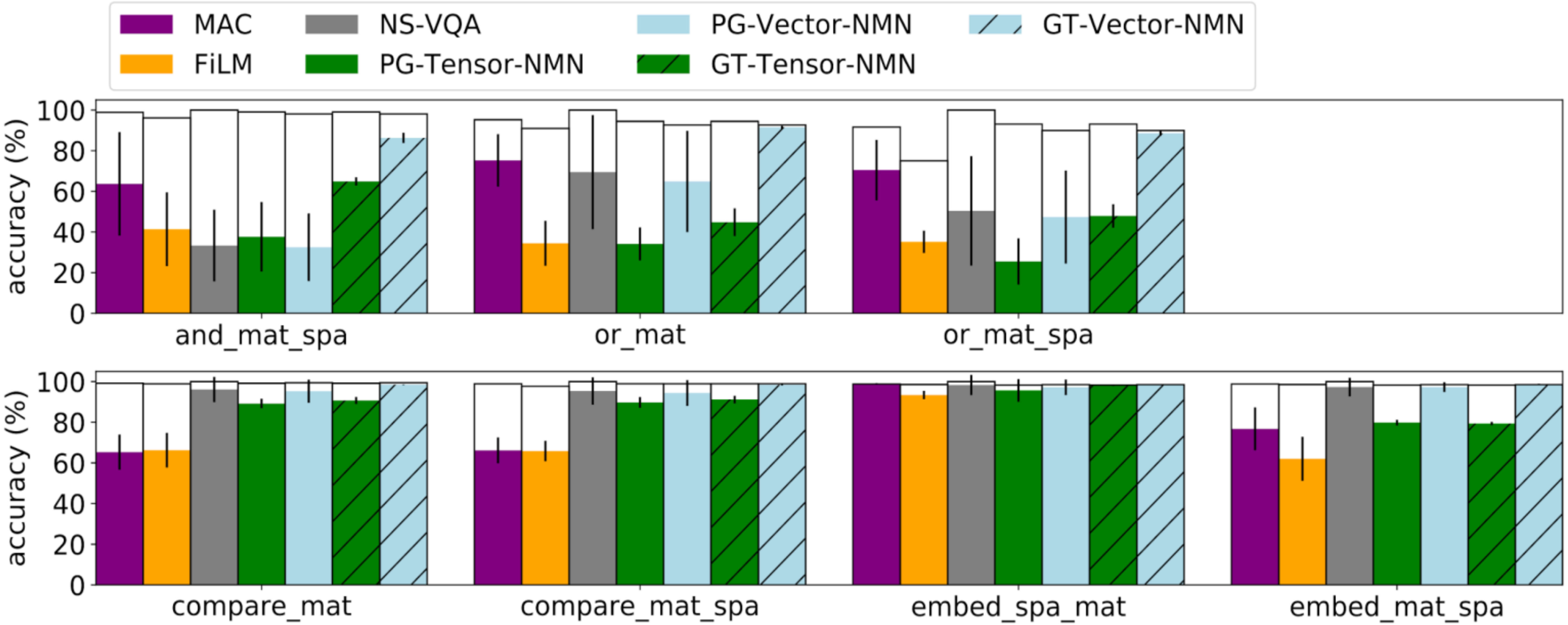
NEW: a comparison question with a matching referring expression

7 CLOSURE Tests

- matching REs and embedded complex REs (2 tests)
  - *Is there a cylinder that is the same material as the object to the left of the blue thing?*
- matching REs and comparison questions (2 tests)
  - *There is another cube that is the same material as the gray cube; does it have the same size as the metal thing to the right of the tiny gray cube?*
- matching REs and logical operations (3 tests)
  - *What is the color of the thing that is to the left of the red cylinder and is the same size as the red block?*

CLEVR models struggle on CLOSURE questions

- end-to-end models (FiLM & MAC) struggle on 6 out of 7 tests
- seq2seq program generator (NS-VQA) struggles on the logical tests
- (surprise!) tensor-valued neural module networks (Tensor-NMN) fair badly even when connected in ground-truth layouts (our new Vector-NMN fares better)



more experiments (including few-shot) in the paper!

CONTRAST WITH **THE SYMBOLIC AI PROGRAM**

Avoid pitfalls of classical AI rule-based symbol-manipulation

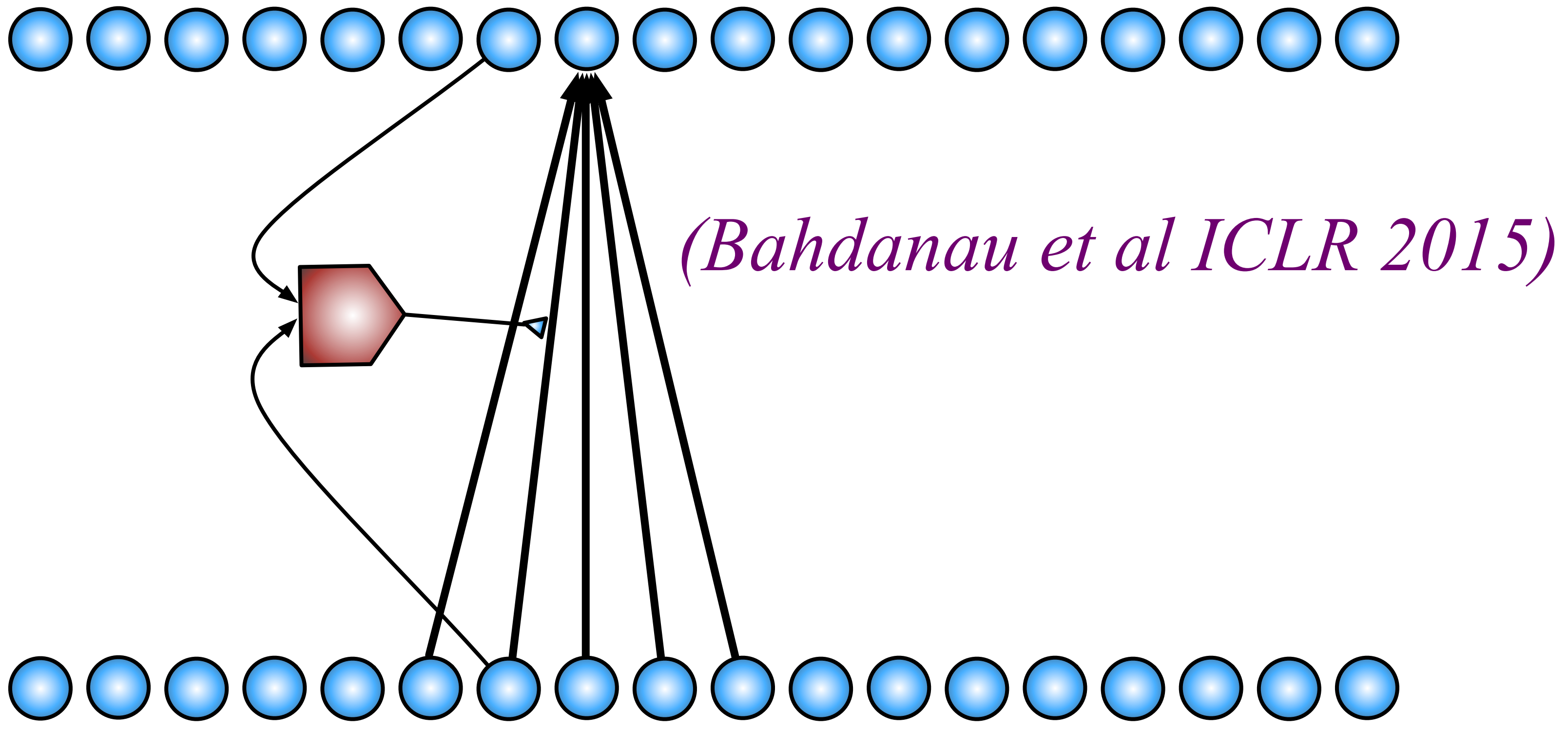
- Need efficient large-scale learning
- Need semantic grounding in system 1
- Need distributed representations for generalization
- Need efficient = trained search (also system 1)
- Need uncertainty handling

But want

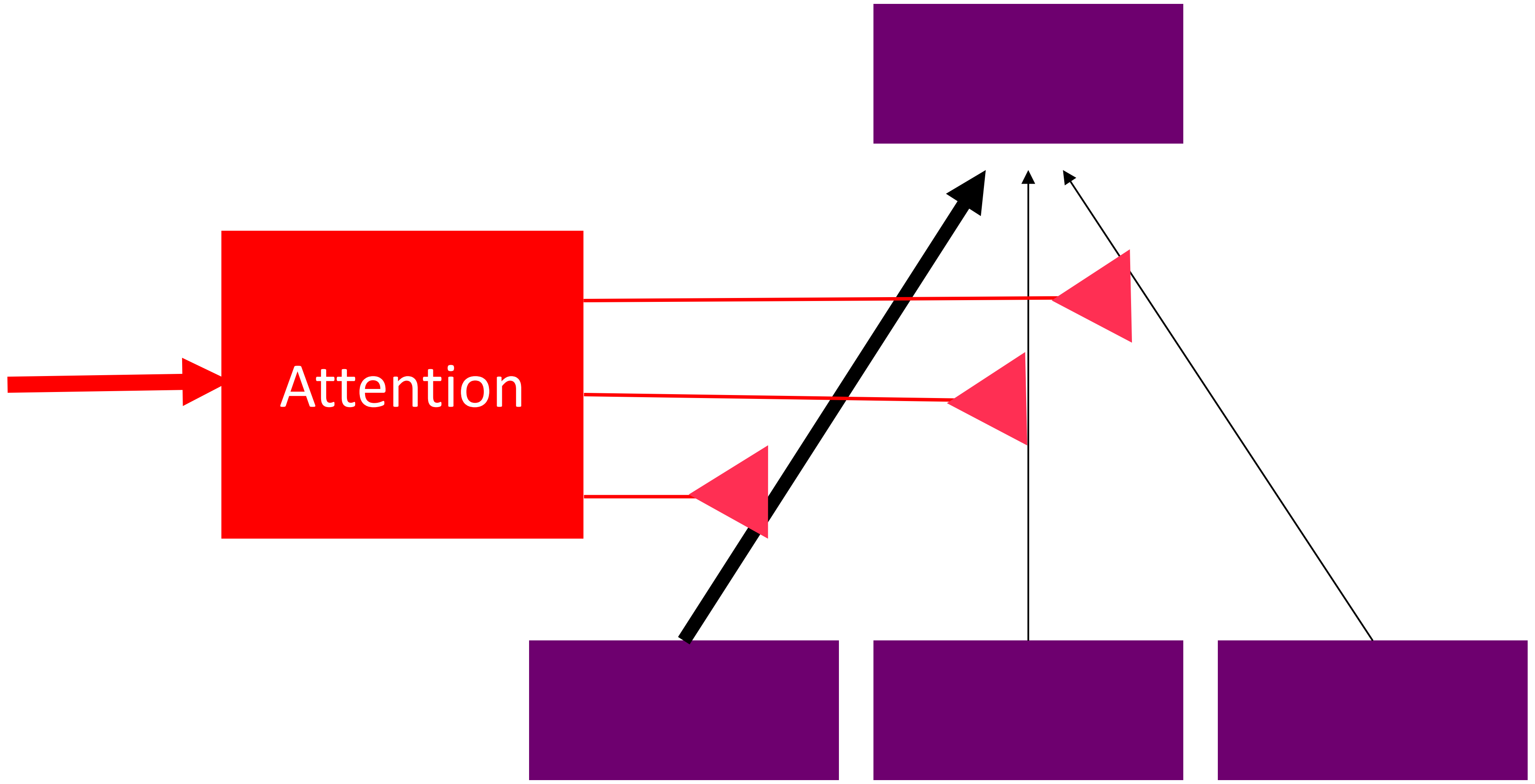
- Systematic generalization
- Factorizing knowledge in small exchangeable pieces
- Manipulating variables, instances, references & indirection

SYSTEM 2 BASICS: ATTENTION AND CONSCIOUSNESS

CORE INGREDIENT FOR CONSCIOUSNESS: **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)*
- Self-attention: SOTA in NLP (transformers)



MEMORY ACCESS & VANISHING GRADIENT - REMINDING AND CREDIT ASSIGNMENT

Humans selectively recall memories that are relevant to the current behavior

This creates a link between arbitrarily far past and the present

Automatic reminding:

- Triggered by contextual features.
- Can serve a useful computational role in ongoing cognition.
- Can be used for credit assignment to past events?

Assign credit through only a few past states, instead of all states:

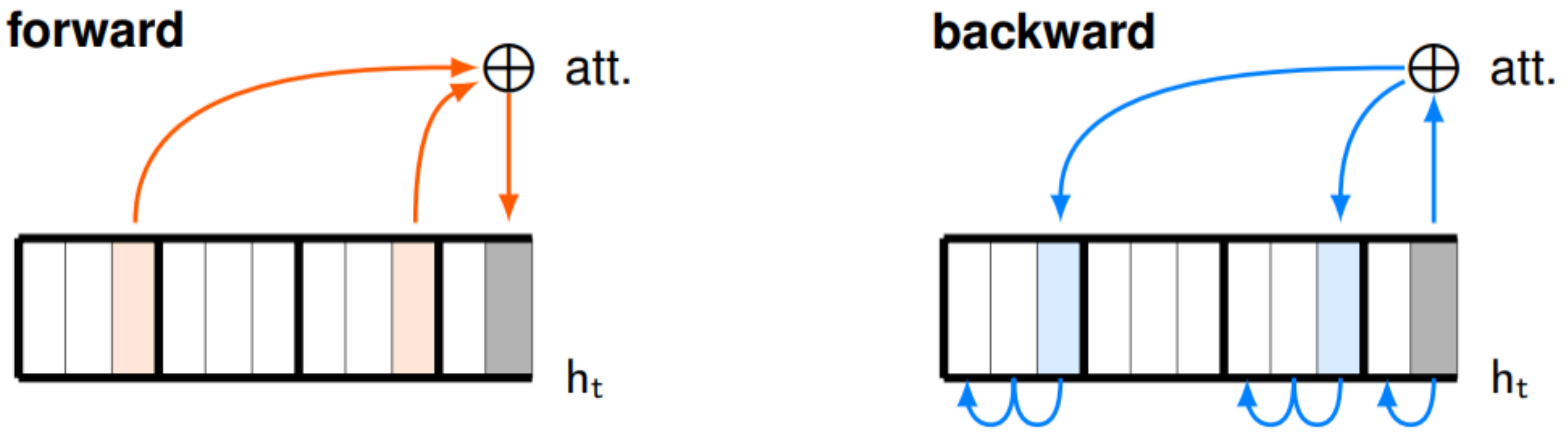
- Sparse, local credit assignment.
- How to pick the states to assign credit to?

Sparse Attentive Backtracking: attention on the past

Rosemary Ke, Anirudh Goyal, Olexa Bilaniuk, Jonathan Binas, Mike Mozer, Yoshua Bengio,

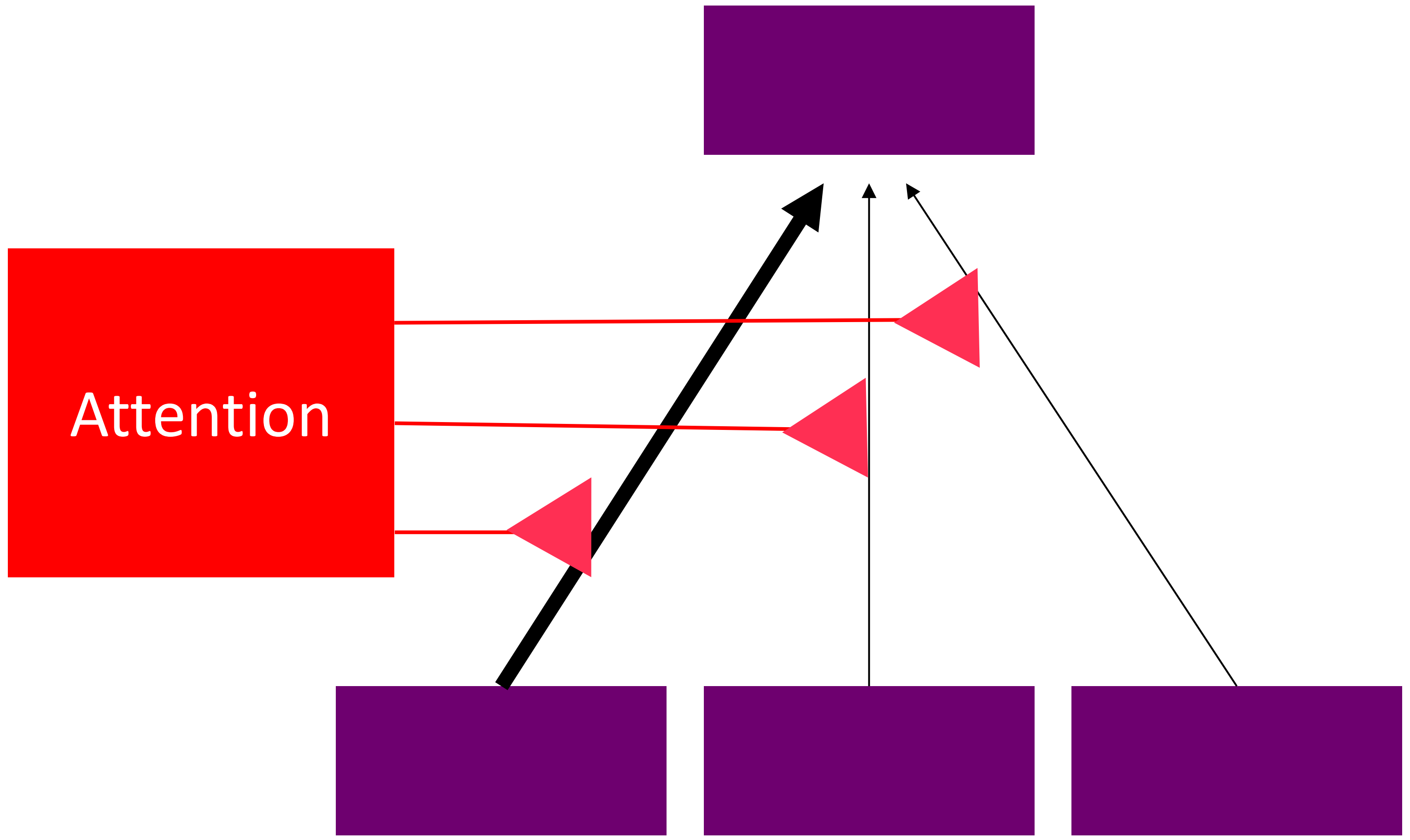
NeurIPS 2018

The attention mechanism of the associative memory picks up past memories which match (associate with) the current state.



➔ Bypass the vanishing gradient problem and capture long-term dependencies
Ongoing work with G. Lajoie, G. Kerg, B. Kanuparthi

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)

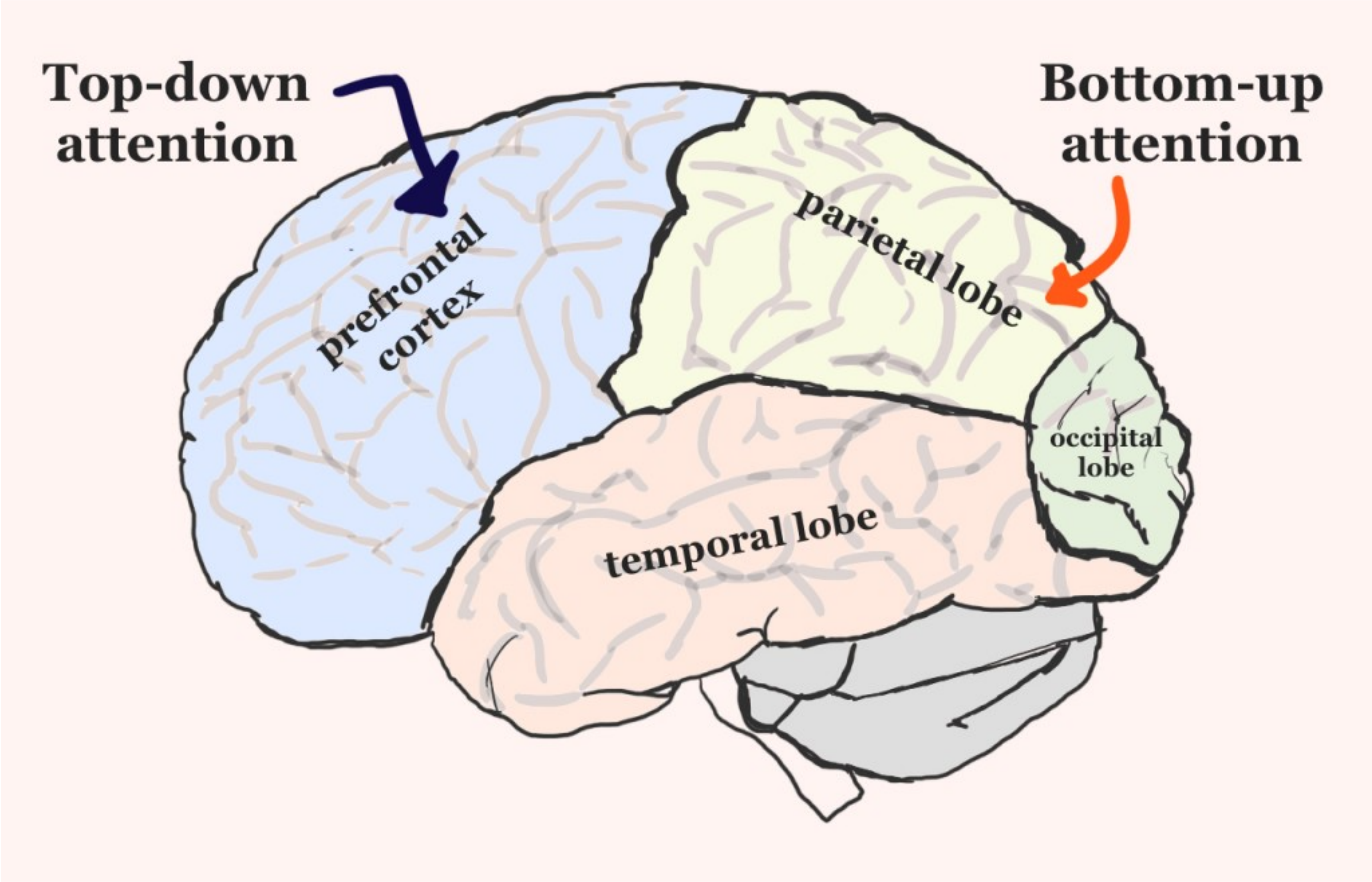
FROM ATTENTION TO **CONSCIOUSNESS**

C-word not taboo anymore in cognitive neuroscience

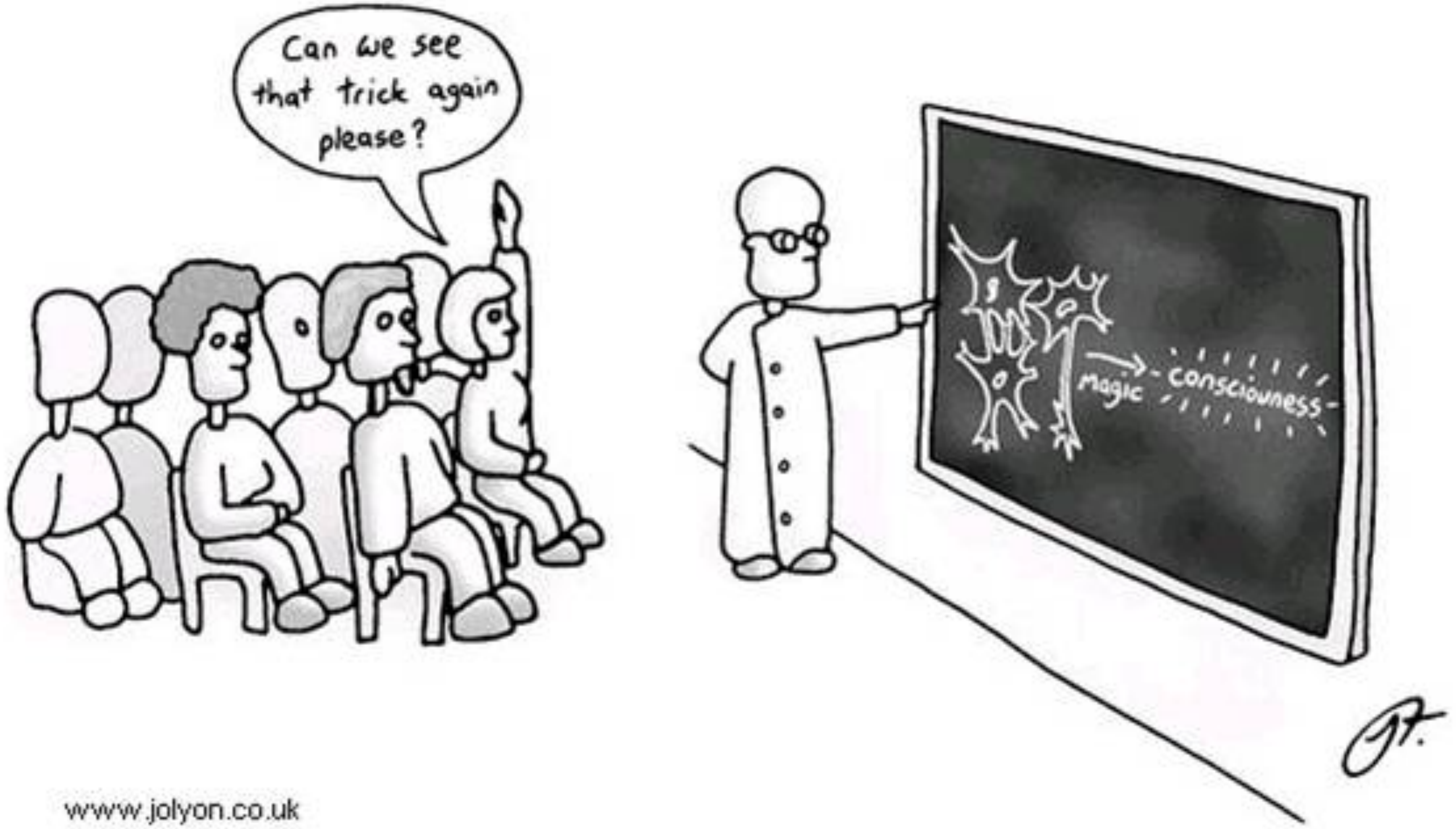
Global Workspace Theory

(Baars 1988++, Dehaene 2003++)

- Bottleneck of conscious processing
- 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



www.jolyon.co.uk

- 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

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)*

Grounded Language Learning

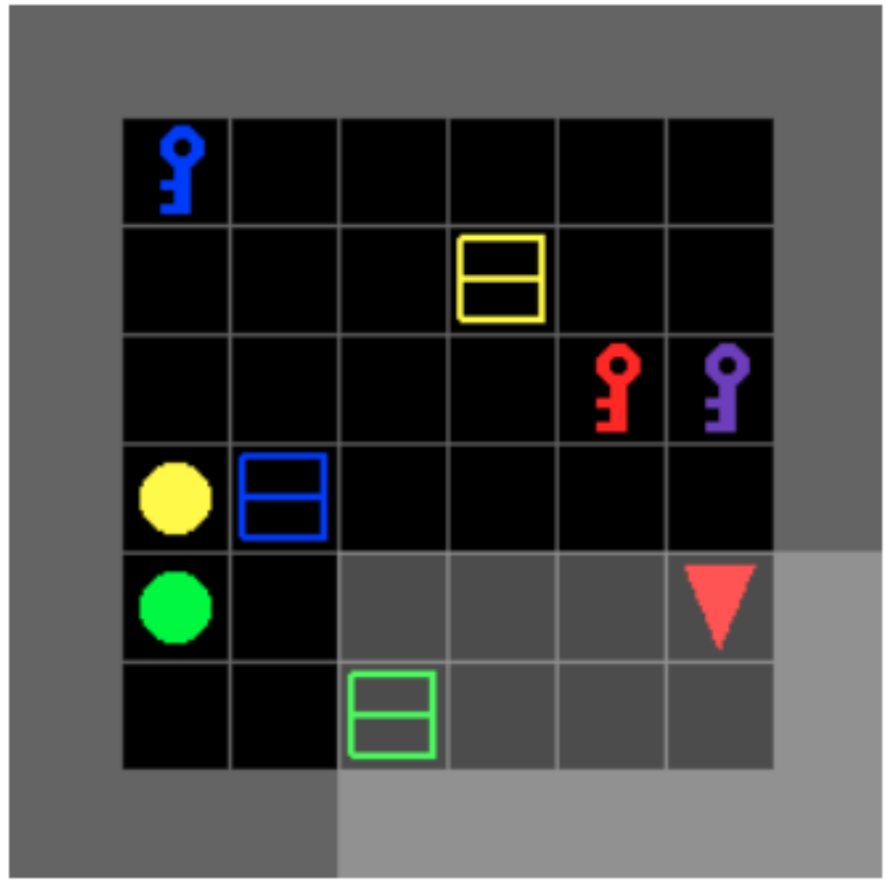
BABY AI PLATFORM *Chevalier-Boisvert et al & Bengio ICLR 2019*

Purpose: simulate language learning from a human and study data efficiency

Comprises:

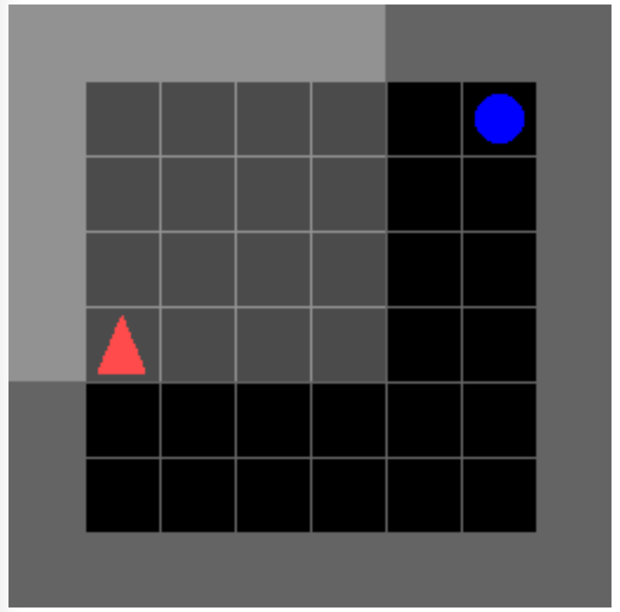
- a gridworld with partial observability (Minigrid)
- a compositional natural-looking Baby language with over 10^19 instructions
- 19 levels of increasing difficulty
- a heuristic stack-based expert that can solve all levels

github.com/mila-udem/babyai

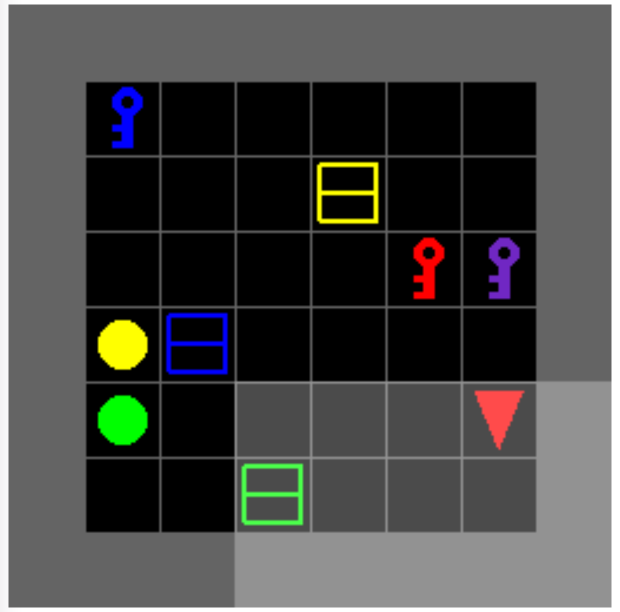


(b) PutNextLocal: "put the blue key next to the green ball"

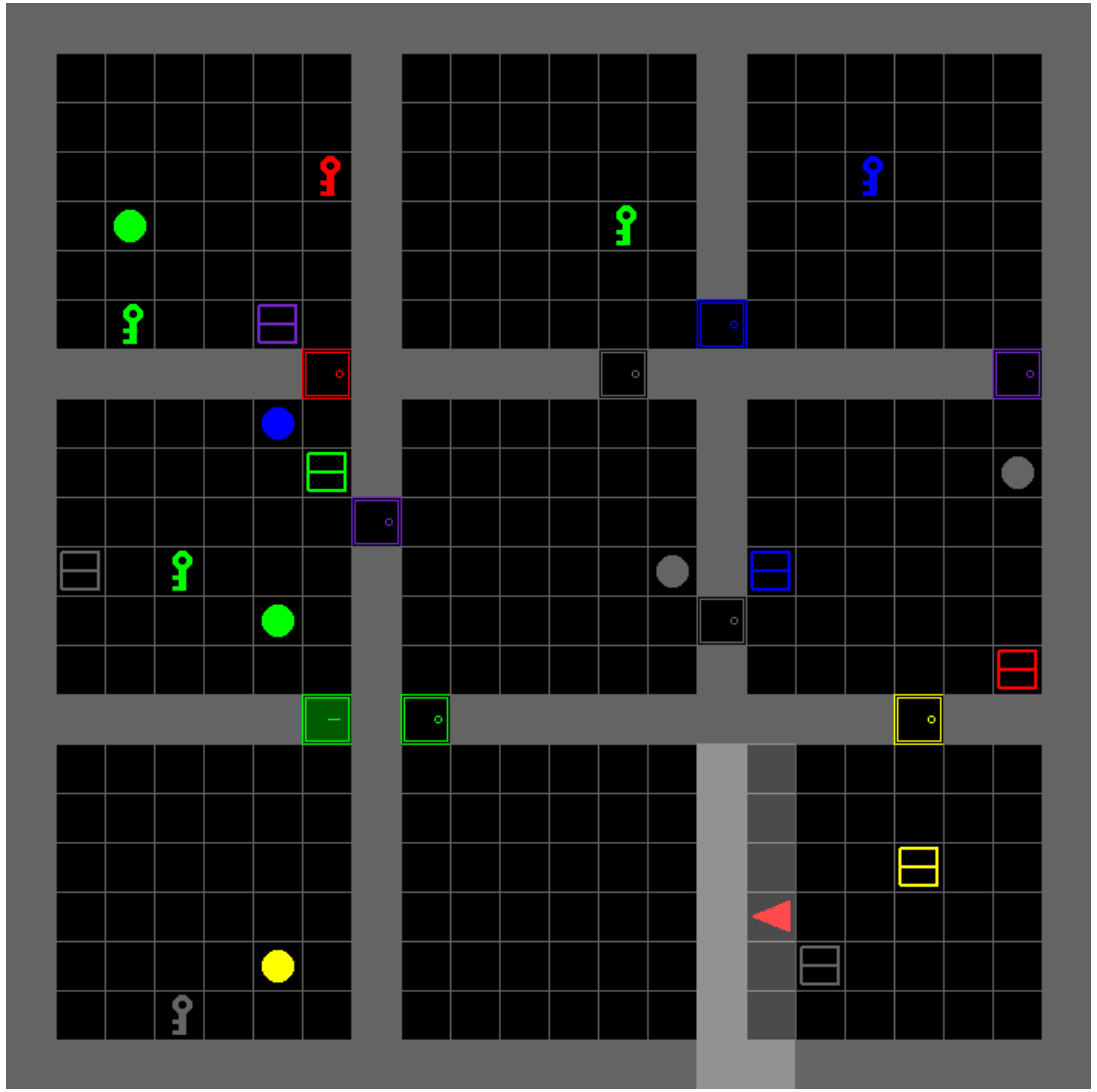
Early Steps in Baby AI Project



(a) GoToObj: "go to the blue ball"



(b) PutNextLocal: "put the blue key next to the green ball"



(c) BossLevel: "pick up the grey box behind you, then go to the grey key and open a door". Note that the green door near the bottom left needs to be unlocked with a green key, but this is not explicitly stated in the instruction.

- Designing and training experts for each level, which can serve as teachers and evaluators for the Baby AI learners
- Partially observable, 2-D grid, instructions about objects, locations, actions

go to the red ball

open the door on your left

put a ball next to the blue door

open the yellow door and go to the key behind you

put a ball next to a purple door after you put a blue box next to a grey box and pick up the purple box

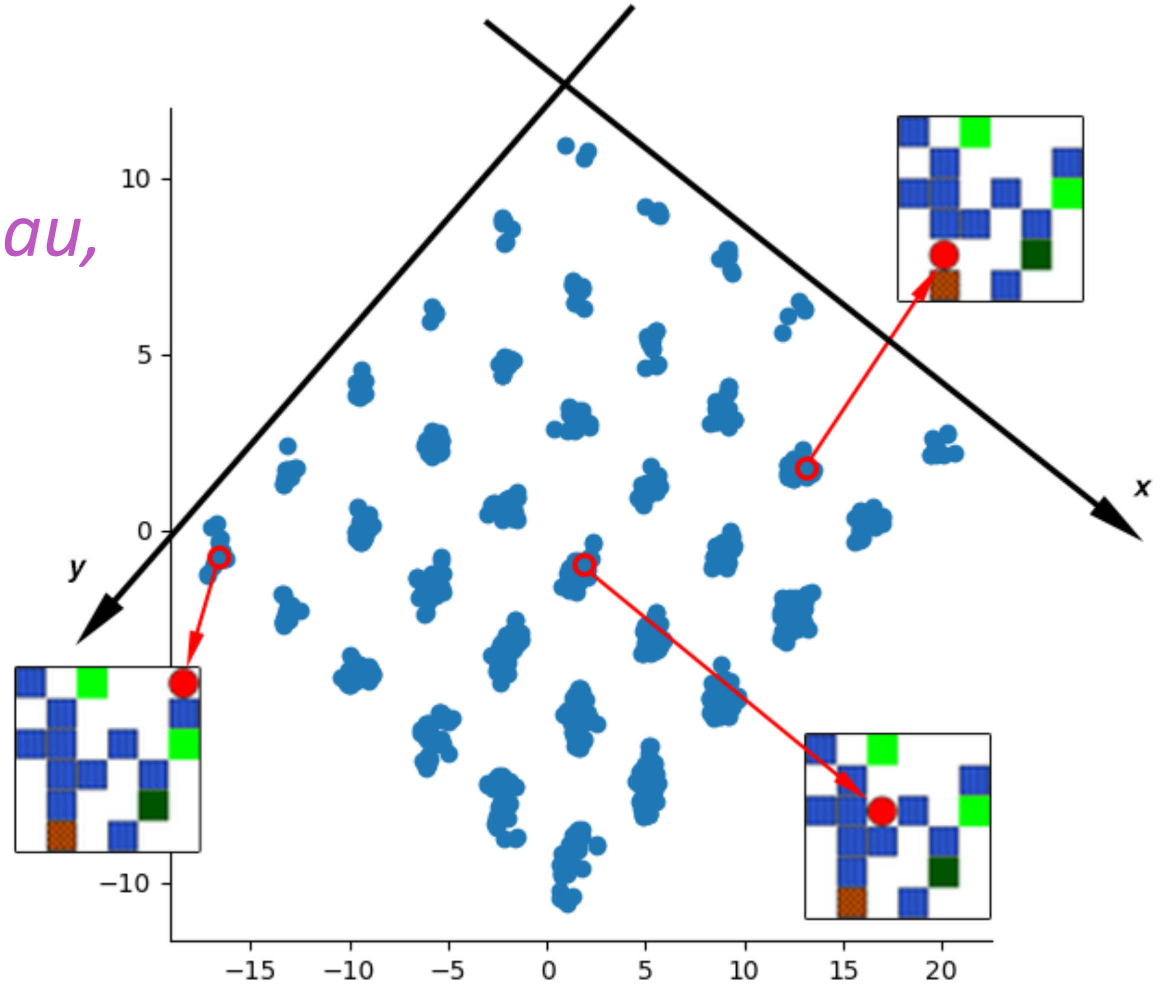
Mila

AFFORDANCES, OPTIONS, EXPLORATION & CONTROLLABLE FACTORS

- Affordances: concepts / aspects of the environment which can be changed by the agent
- Temporal abstractions: options, super-actions, macros or procedures, which can be composed to form more complex procedures *(Sutton, Precup & Singh 1999)*
- Controllable factors: jointly learn a set of (policy, factor) such that the policy can control the factor and maximize mutual information between policies and factors *(Bengio, Thomas, Pineau, Precup & Bengio 2017)*

The handles on a tea set provide an obvious affordance for holding.

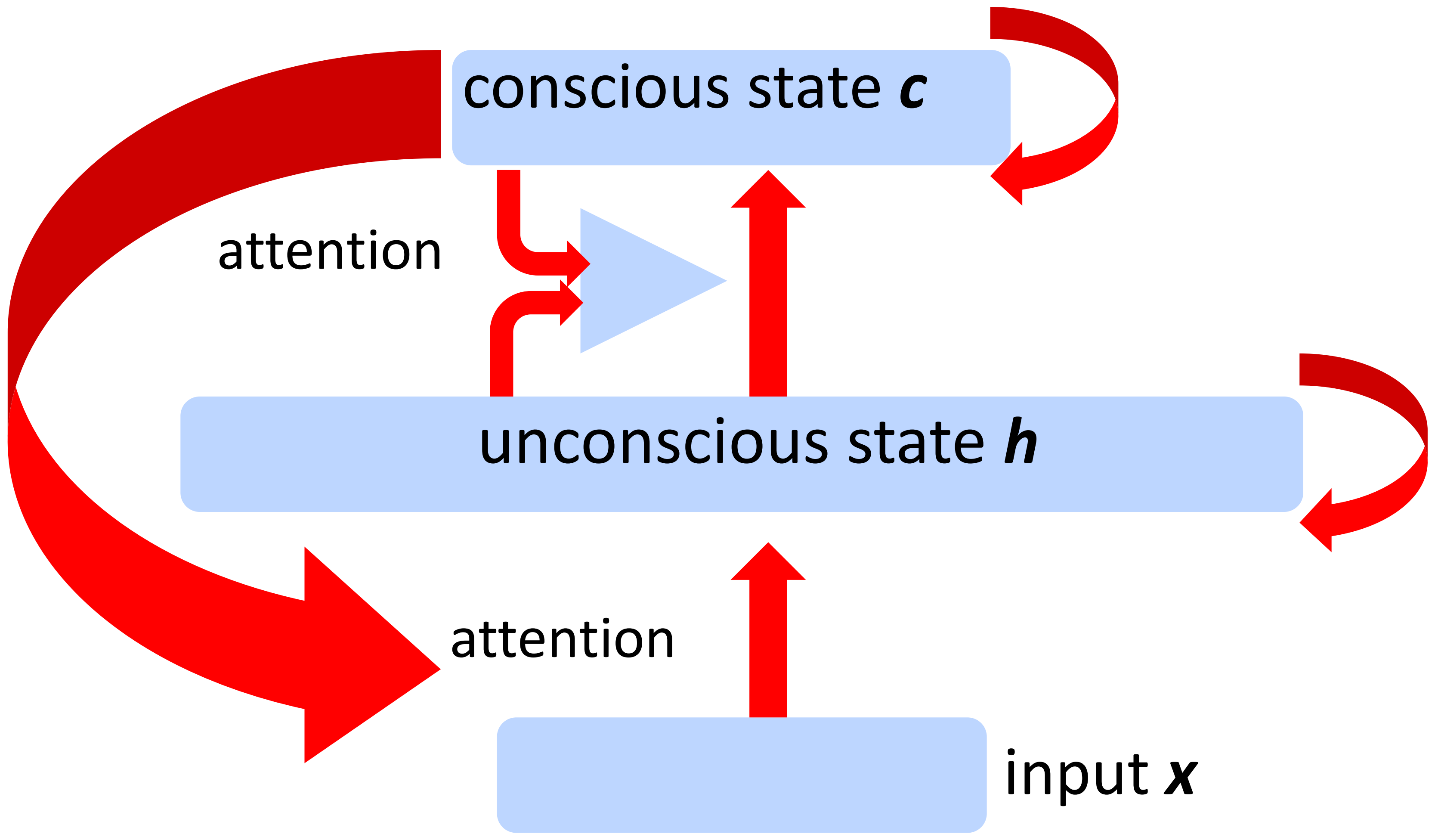
(Gibson, 1979)



THE CONSCIOUSNESS PRIOR: SPARSE FACTOR GRAPH

CONSCIOUSNESS PRIOR

Bengio 2017, arXiv:1709.08568

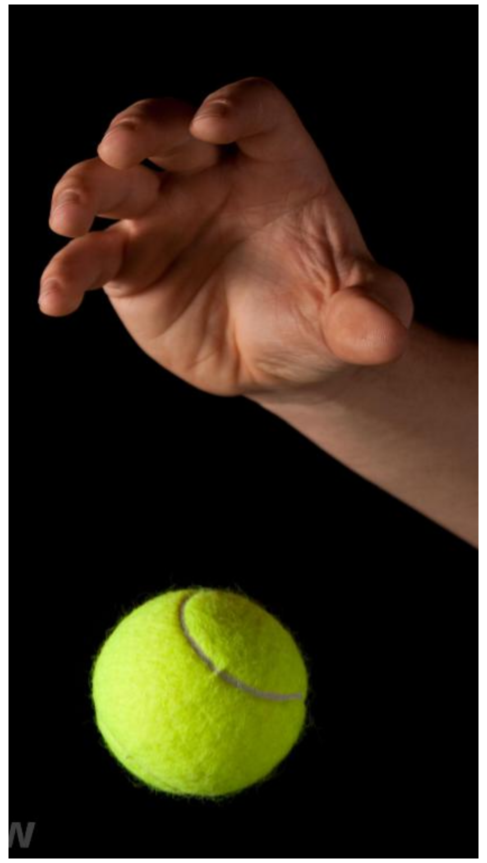


Different kinds of attention in the brain

- **Attention: to form conscious state, thought**
- **A thought is a low-dimensional object**, few selected aspects of the unconscious state
- Need 2 high-level states:
  - Large unconscious state
  - Tiny conscious state
- Part of inference mechanism wrt joint distribution of high-level variables

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

Bengio 2017, arXiv:1709.08568



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

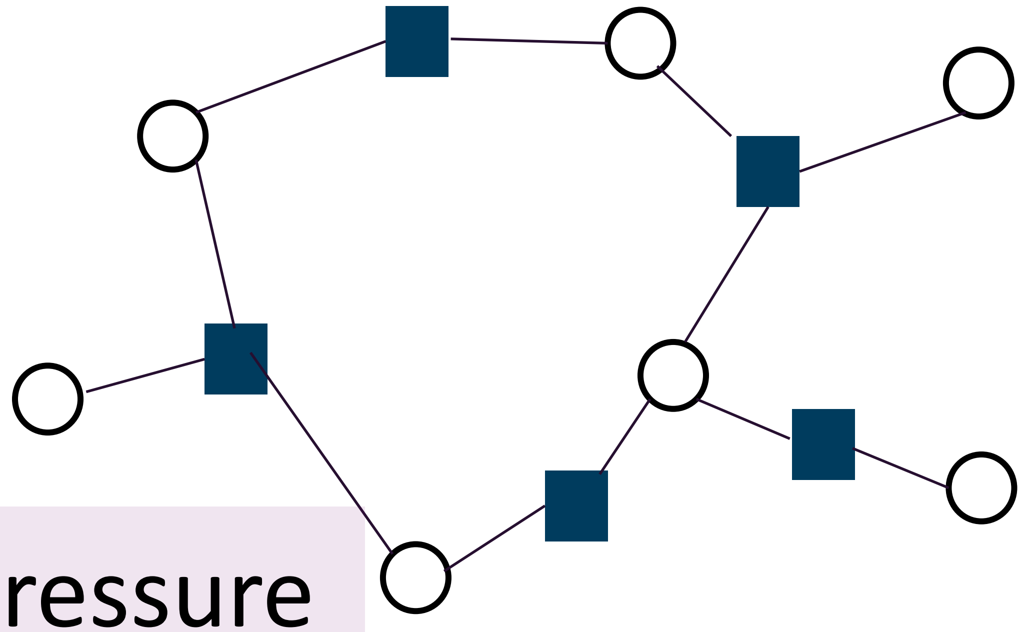
  we can predict some given very few others

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

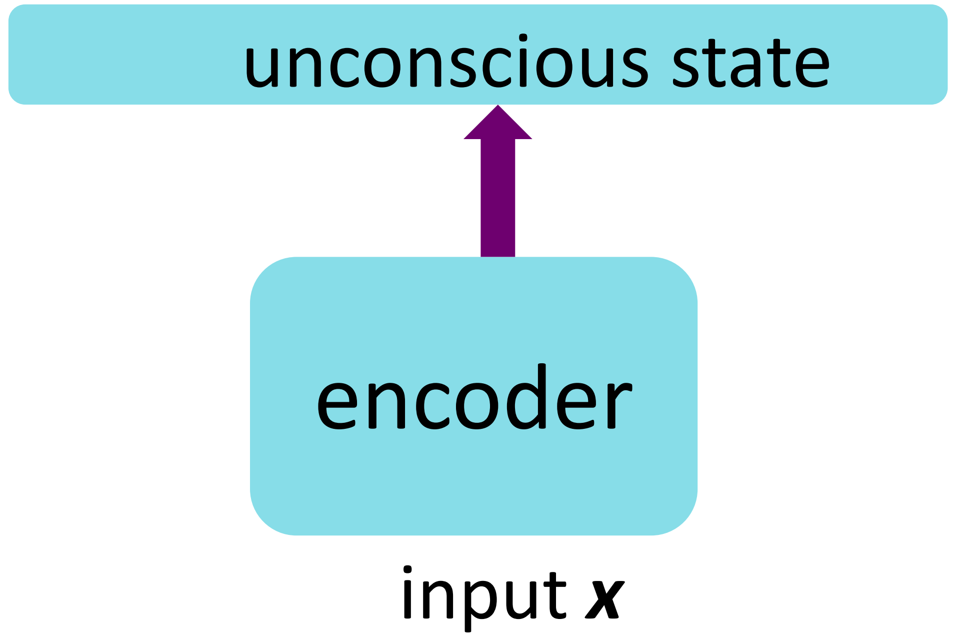
- **Disentangled factors** != marginally independent,

  e.g. ball & hand

- **Prior**: sparse factor graph joint distribution between high-level variables, consistent with inference mechanism which looks at just a few variables at a time.



Prior puts pressure on encoder



LOCALIZED CHANGE HYPOTHESIS

WHAT CAUSES CHANGES IN DISTRIBUTION?

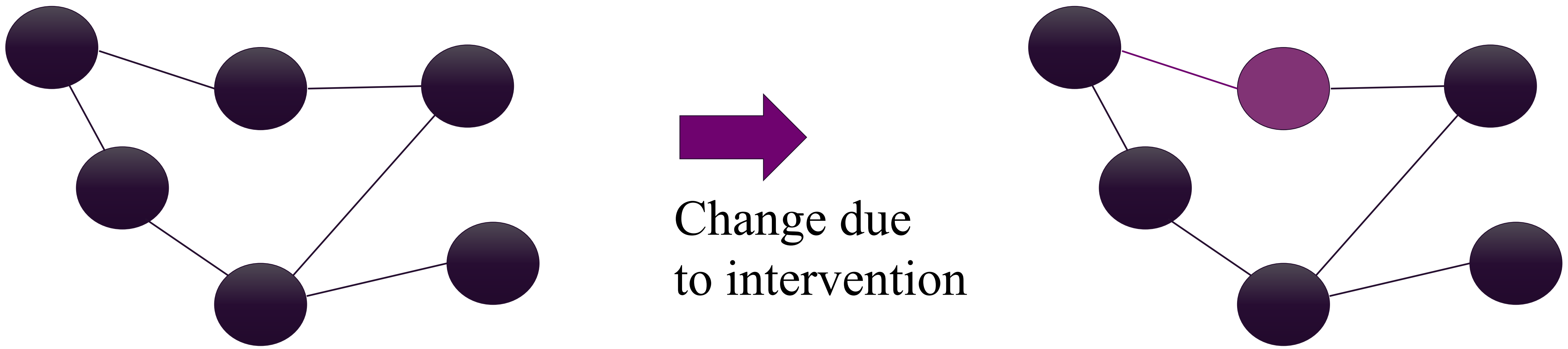
Underlying physics: actions are localized in space and time.

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**



Change due to intervention

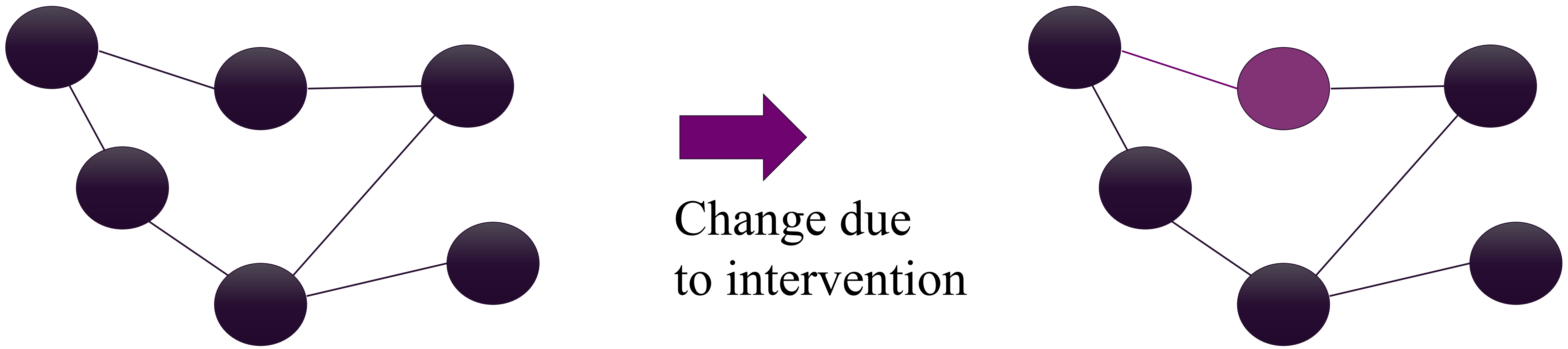
COUNTING ARGUMENT:
LOCALIZED CHANGE→OOD TRANSFER

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



Change due
to intervention

META-LEARNING KNOWLEDGE REPRESENTATION FOR GOOD OOD PERFORMANCE

- Use ood generalization as training objective

- Good decomposition / knowledge representation ➔ good ood performance

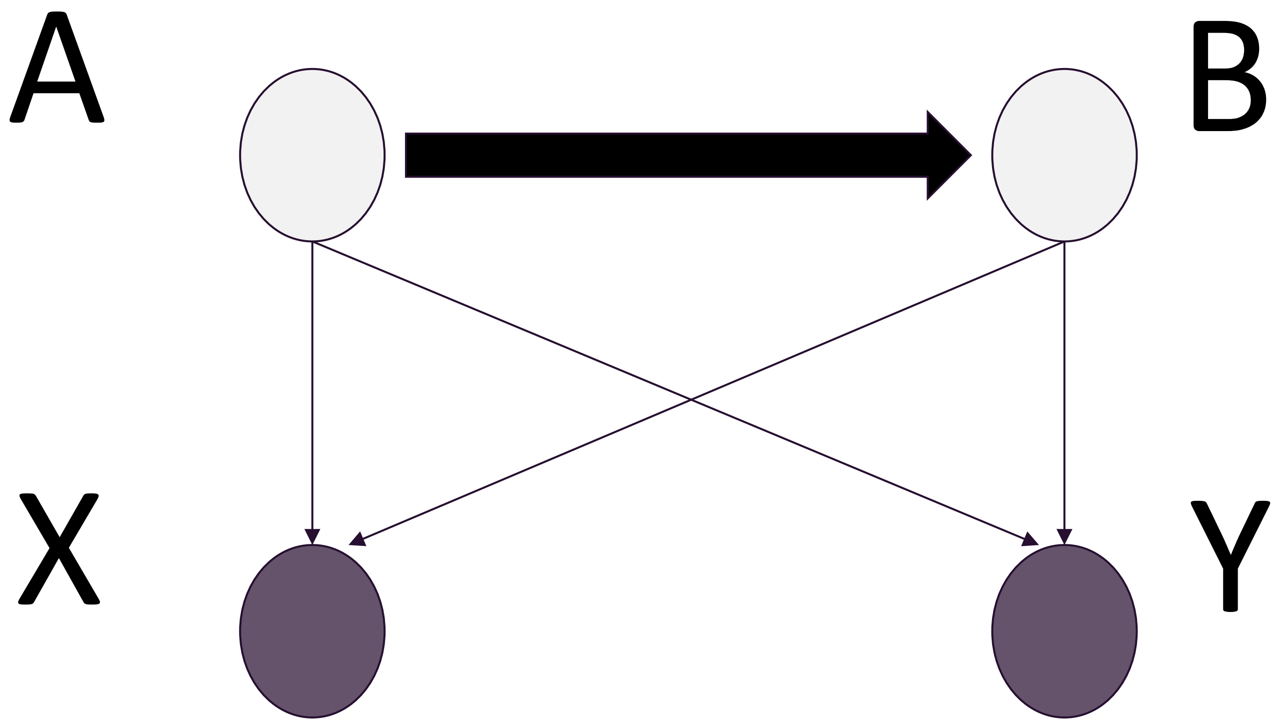
- Good ood performance = training signal for factorizing knowledge

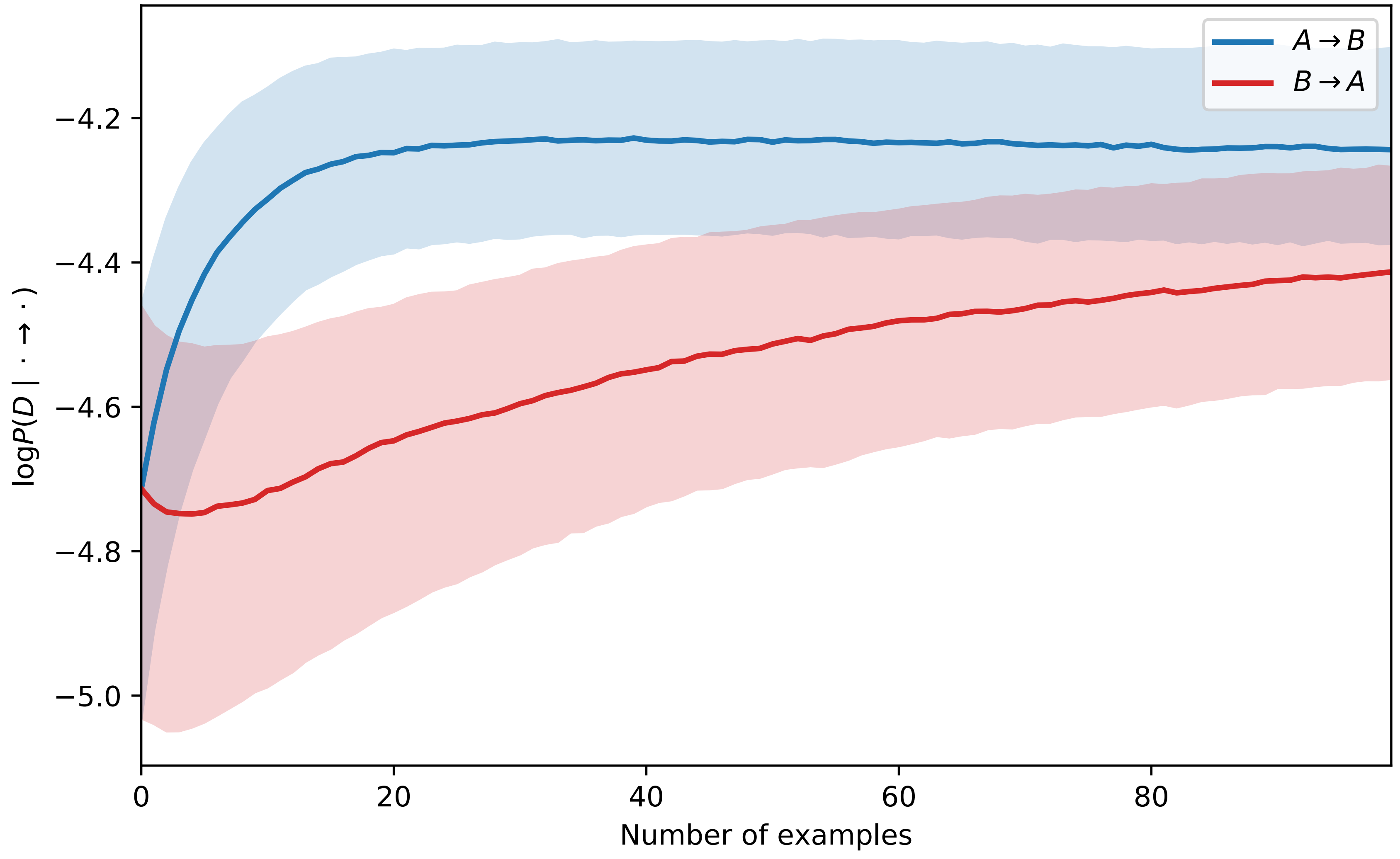
EXAMPLE: DISCOVERING CAUSE AND EFFECT
= **HOW TO FACTORIZE A JOINT DISTRIBUTION?**

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

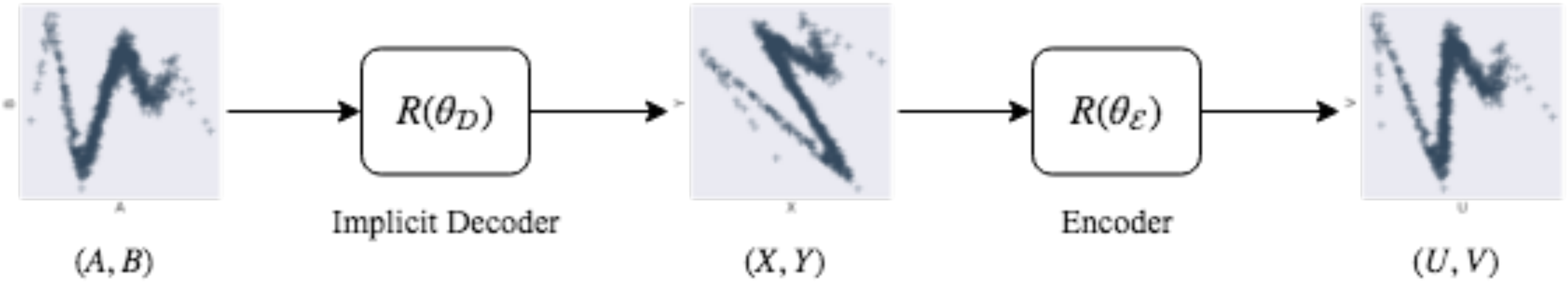




A NOVEL APPROACH TO CAUSALITY: **DISENTANGLING THE CAUSES**

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

- 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

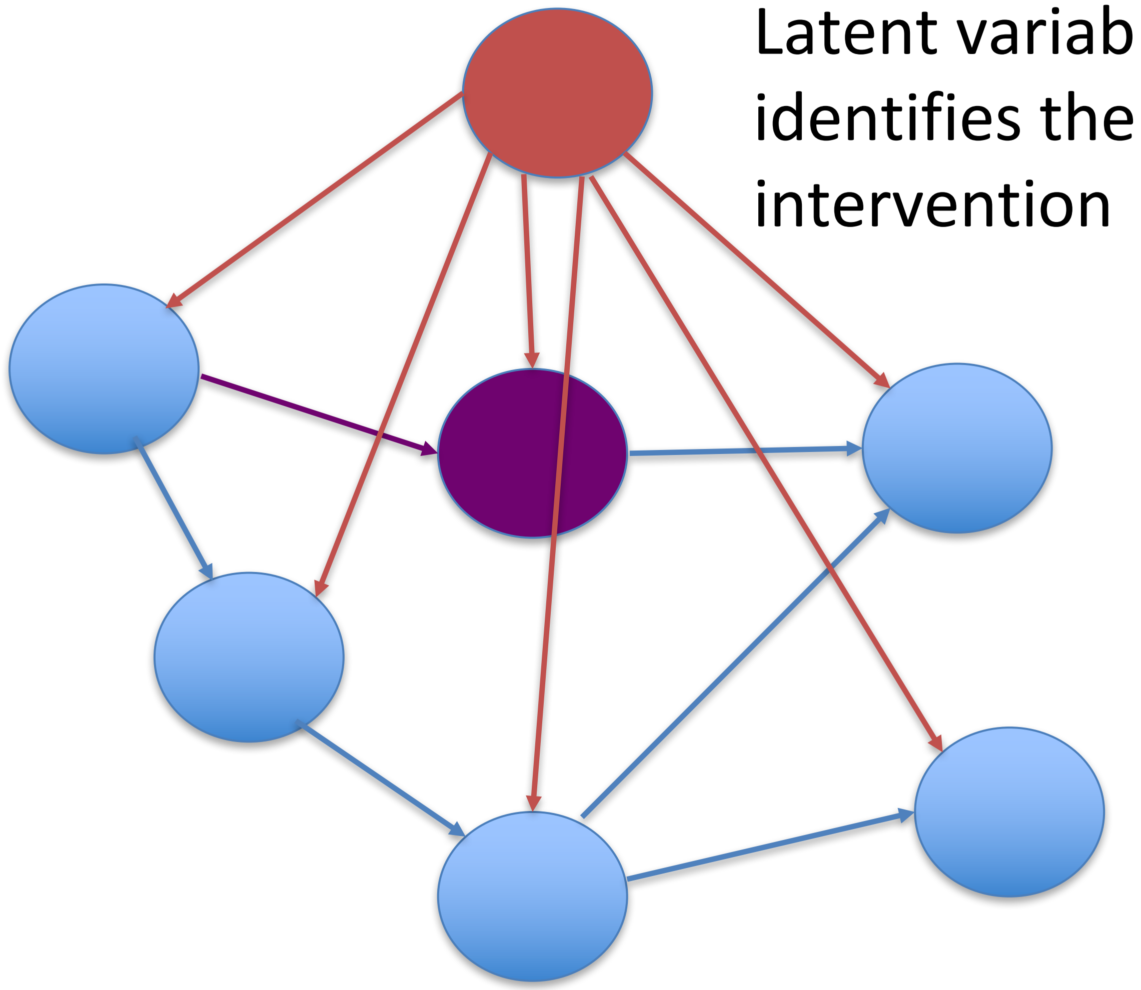


Experiments successful in 2-D with simple linear mappings, Bengio et al 2019.

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



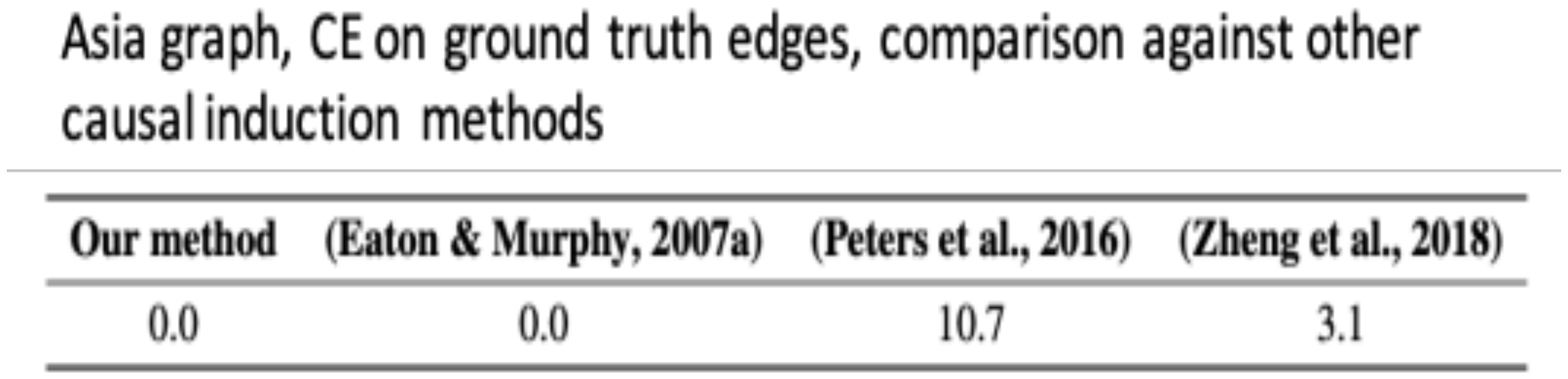
Latent variable identifies the intervention

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*

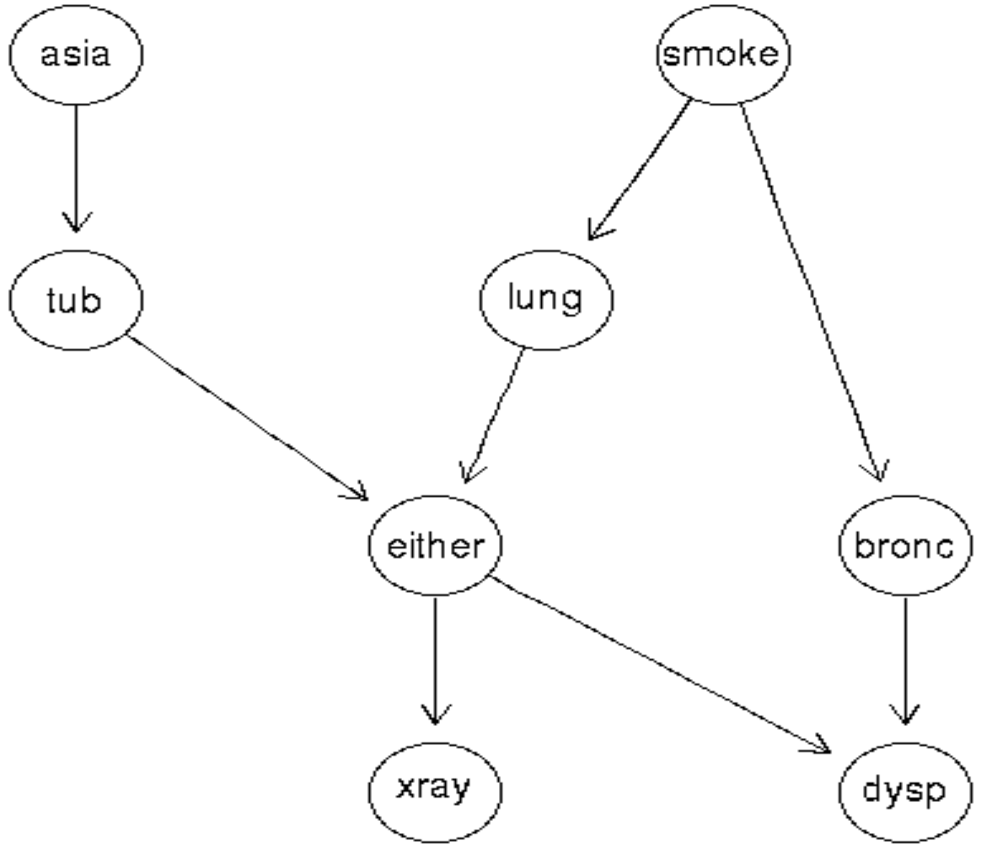
- Learning small causal graphs, avoid exponential explosion of # of graphs by parametrizing factorized distribution over graphs

- Inference over the intervention:
  faster causal discovery

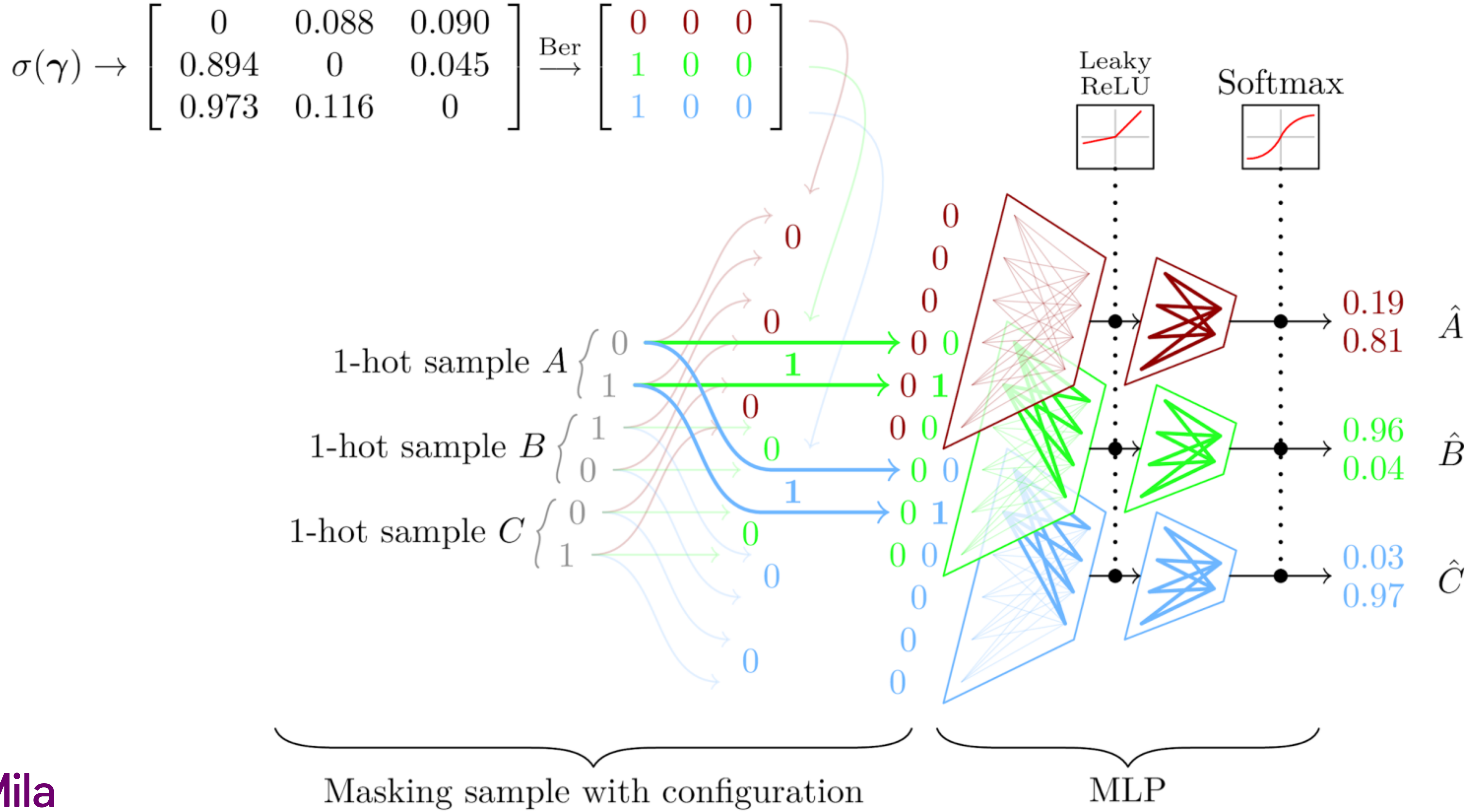


Asia graph, CE on ground truth edges, comparison against other causal induction methods

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



MULTIVARIATE CATEGORICAL MLP CONDITIONALS



$$\sigma(\gamma) \rightarrow \begin{bmatrix} 0 & 0.088 & 0.090 \\ 0.894 & 0 & 0.045 \\ 0.973 & 0.116 & 0 \end{bmatrix} \xrightarrow{\text{Ber}} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

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

OPERATING ON SETS OF POINTABLE OBJECTS WITH DYNAMICALLY RECOMBINED MODULES

RECURRENT INDEPENDENT MECHANISMS

Goyal et al 2019, arXiv:1909.10893

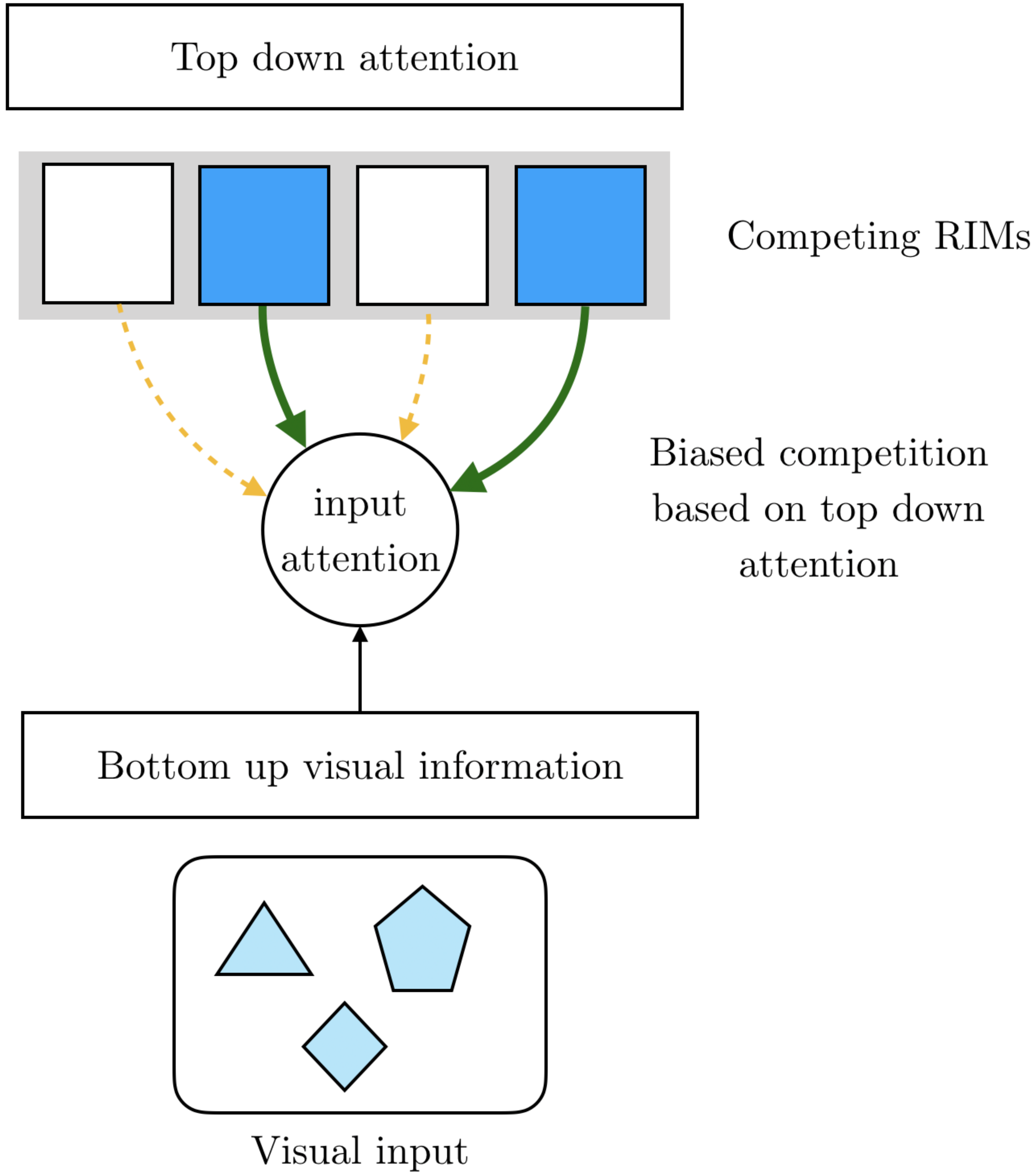
- Recurrent Neural Network with multiple modules which remain independent *by default* and only communicate with attention.

- Additionally, only some fraction are allowed to update their recurrent state on each time step.

- This separation is very hard for most RNNs to achieve.
  - In an LSTM, would require (k-1)^2/k^2 parameters to be zero.

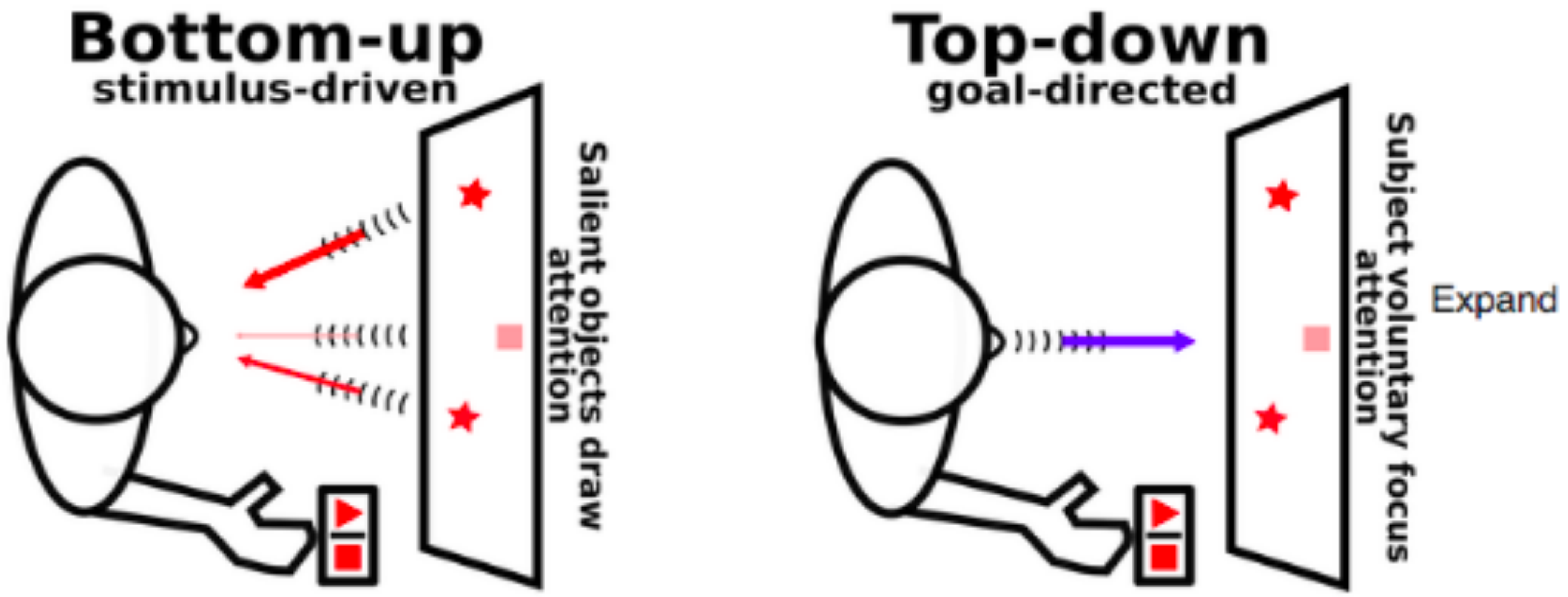
RECURRENT INDEPENDENT MECHANISMS

- Each module only attends to selected part of input.
- Robust to distribution shift.
- Robust to distractors.


Top down attention
Competing RIMs
input attention
Biased competition based on top down attention
Bottom up visual information
Visual input

RECURRENT INDEPENDENT MECHANISMS

- Data dependent activation of mechanisms
- Active mechanisms communicate with other mechanisms
- Inactive mechanisms follow the default dynamics


Bottom-up
stimulus-driven
Salient objects draw attention
Top-down
goal-directed
Subject voluntary focus attention
Expand

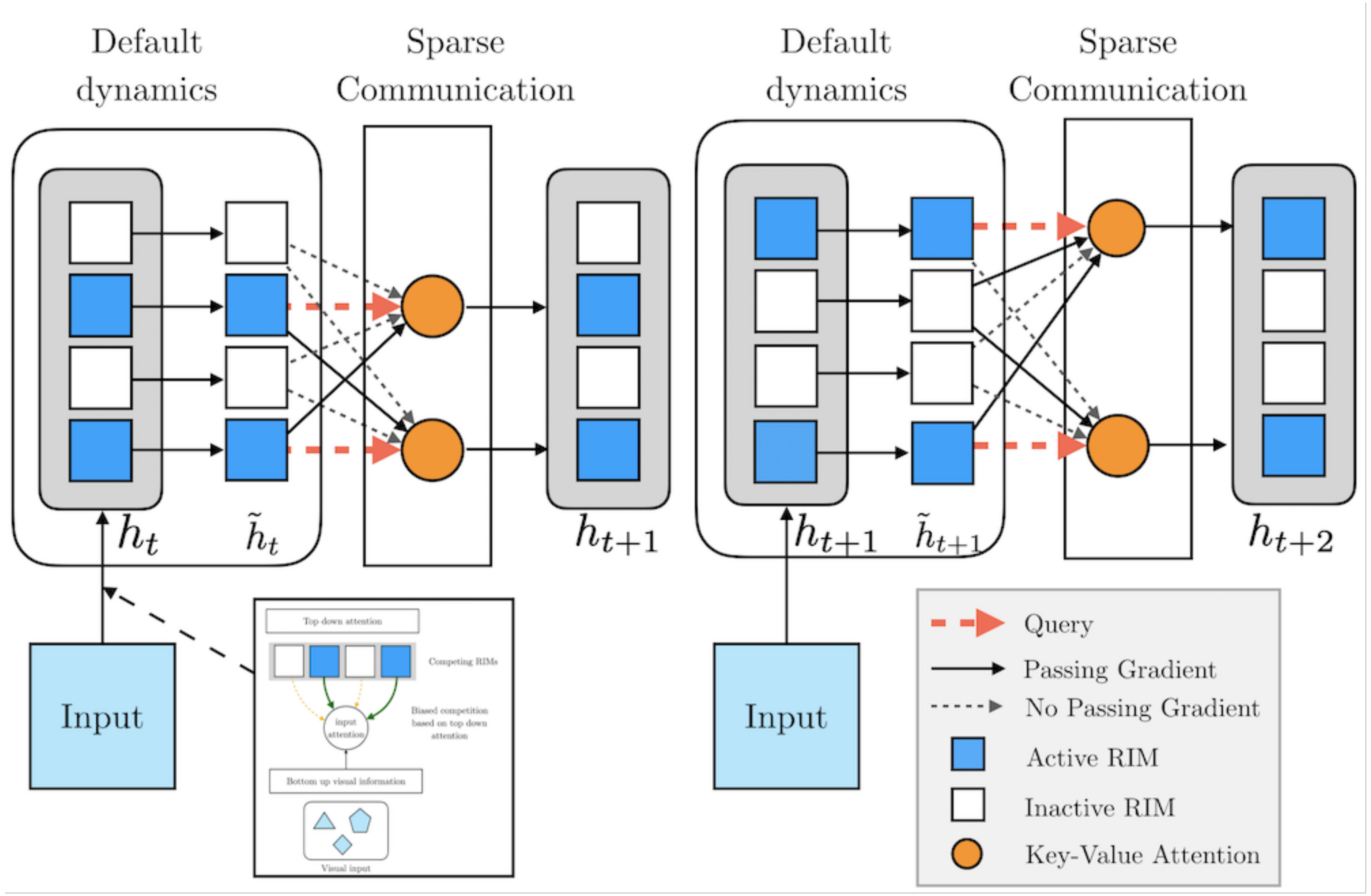
RIMS: MODULARIZE COMPUTATION AND OPERATE ON SETS OF NAMED AND TYPED OBJECTS

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

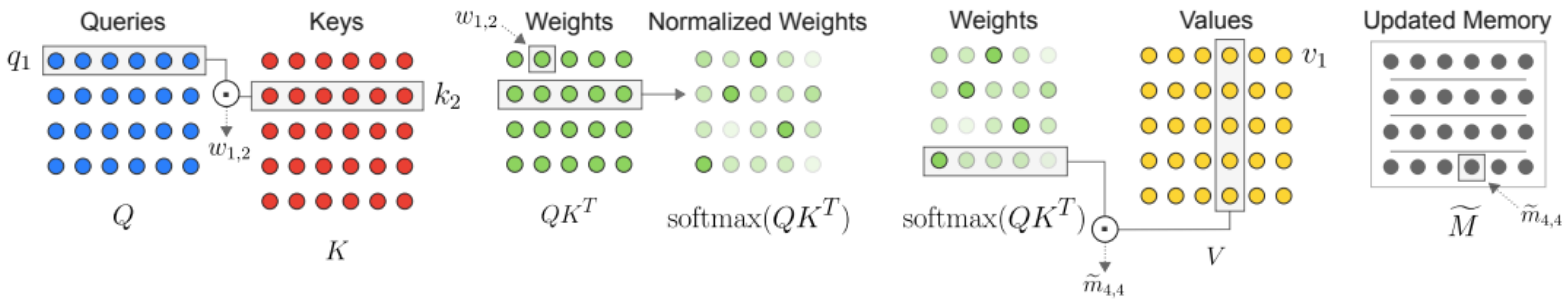
Results: better ood generalization



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

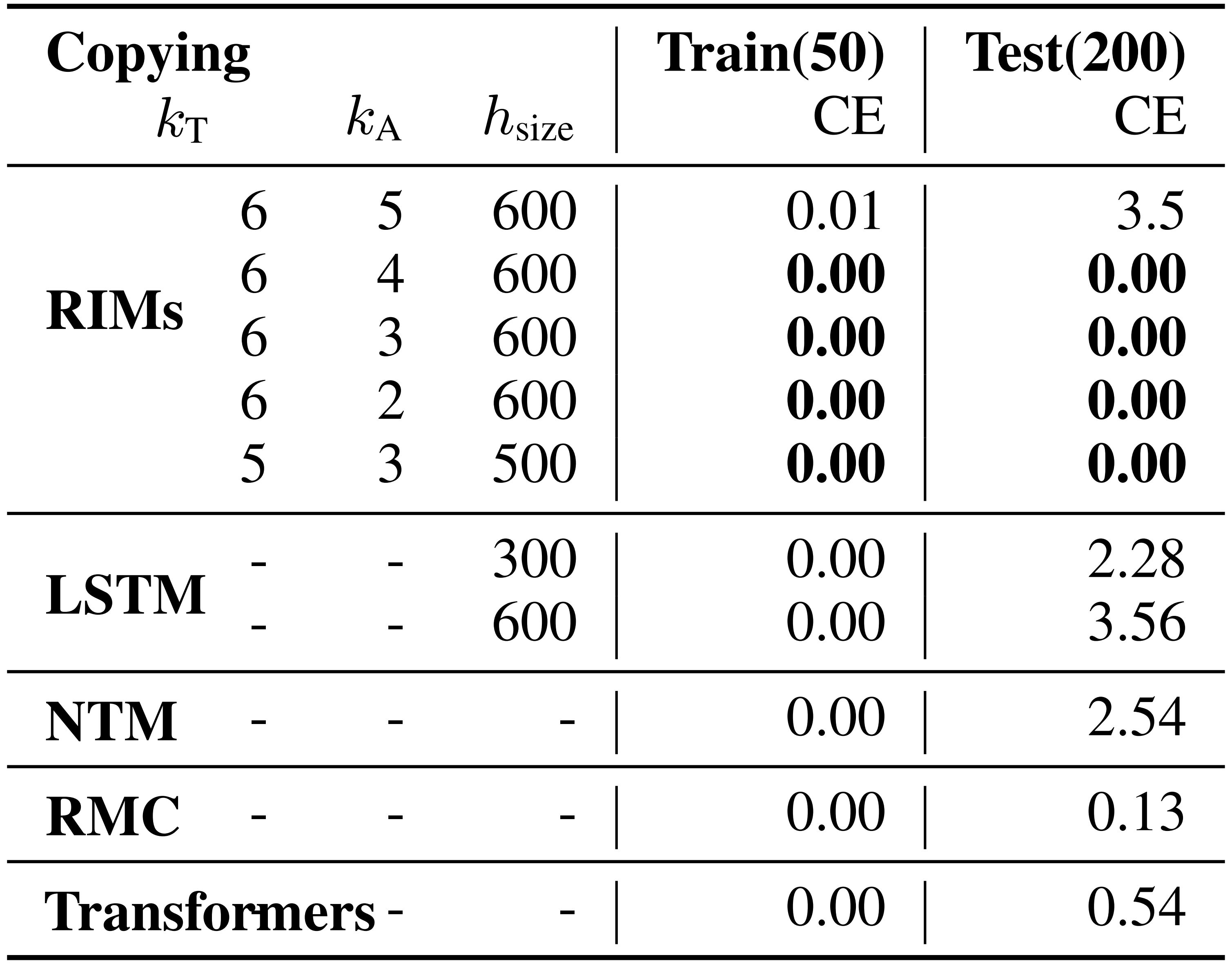
MULTI-HEADED ATTENTION & OBJECTS

RIMs and other models based on self-attention take as input and produce as outputs SETS of OBJECTS (key-value pairs) rather than vectors

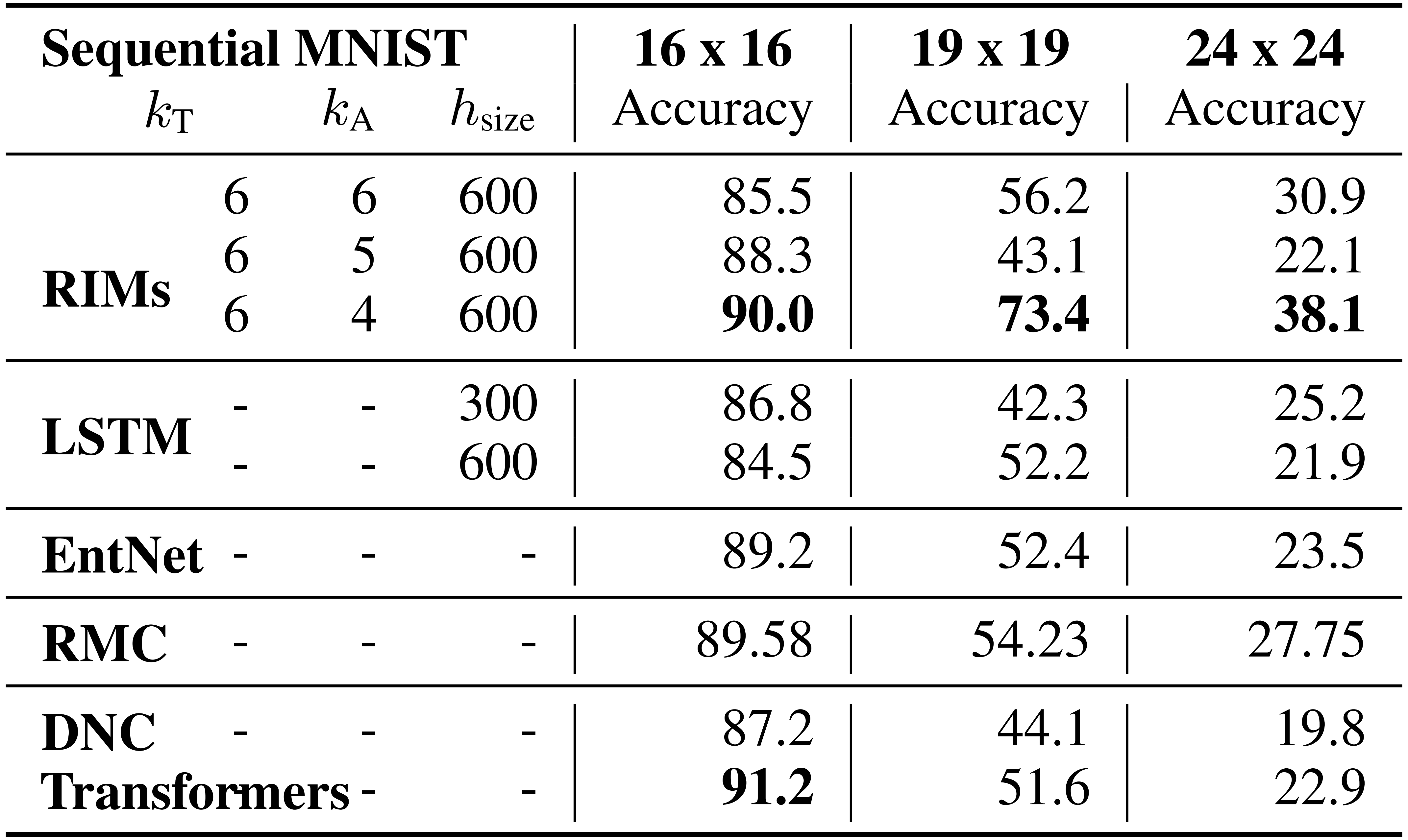


RECURRENT INDEPENDENT MECHANISMS

Goyal et al, arXiv:1909.10893



Copying				Train(50)	Test(200)
	k_T	k_A	h_{size}	CE	CE
RIMs	6	5	600	0.01	3.5
	6	4	600	**0.00**	**0.00**
	6	3	600	**0.00**	**0.00**
	6	2	600	**0.00**	**0.00**
	5	3	500	**0.00**	**0.00**
LSTM	-	-	300	0.00	2.28
	-	-	600	0.00	3.56
NTM	-	-	-	0.00	2.54
RMC	-	-	-	0.00	0.13
Transformers		-	-	0.00	0.54

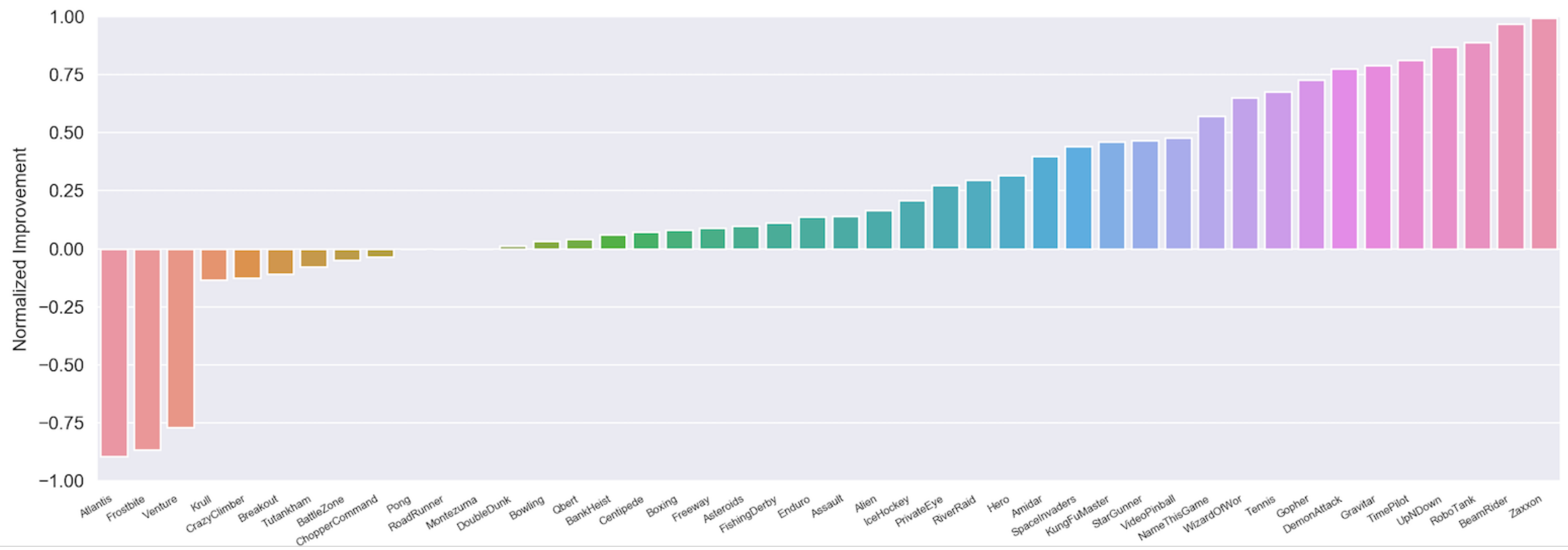


Sequential MNIST				16 x 16	19 x 19	24 x 24
	k_T	k_A	h_{size}	Accuracy	Accuracy	Accuracy
RIMs	6	6	600	85.5	56.2	30.9
	6	5	600	88.3	43.1	22.1
	6	4	600	**90.0**	**73.4**	**38.1**
LSTM	-	-	300	86.8	42.3	25.2
	-	-	600	84.5	52.2	21.9
EntNet	-	-	-	89.2	52.4	23.5
RMC	-	-	-	89.58	54.23	27.75
DNC	-	-	-	87.2	44.1	19.8
Transformers		-	-	**91.2**	51.6	22.9

RIMs generalize better than SOTA methods for sequential learning to out-of-distribution data (longer sequences, larger 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.



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 from localized causal interventions (in semantic space)
- Meaning (e.g. grounded by an encoder) stable & robust wrt changes in distribution

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

THANK YOU

Mila

Université de Montréal

McGill

Québec

CIFAR