Causal Reinforcement Learning

Motivation, Concepts, Challenges, and Applications

陆超超

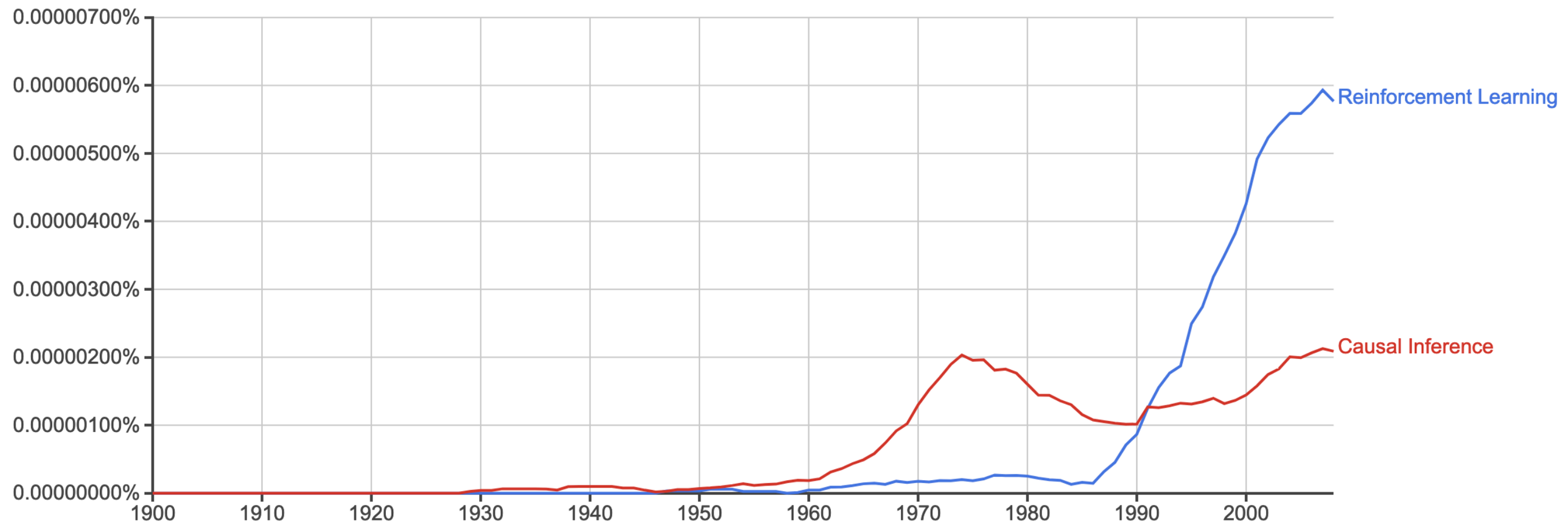
Department of Engineering
University of Cambridge
Cambridge, UK

Department of Empirical Inference
Max Planck Institute for Intelligent Systems
Tübingen, Germany

集智学园 | 因果科学与CausalAI读书会 | 29 Nov 2020

Google Books Ngram Viewer

Ngrams not found: Causal Reinforcement Learning



Google Books Ngram Viewer

Ngrams not found: Causal Reinforcement Learning

The Bible Times

The Bible, for example, tells us that just a few hours after tasting from the tree of knowledge, Adam is already an expert in ***causal arguments****.*

When God asks: "Did you eat from that tree?"

This is what Adam replies: "The woman whom you gave to be with me, She handed me the fruit from the tree; and I ate."

Eve is just as skilful: "The serpent deceived me, and I ate."

The thing to notice about this story is that God did not ask for ***explanation****, only for the* ***facts*** *– it was Adam who felt the need to explain. The message is clear:* ***causal explanation*** *is a man-made concept.*

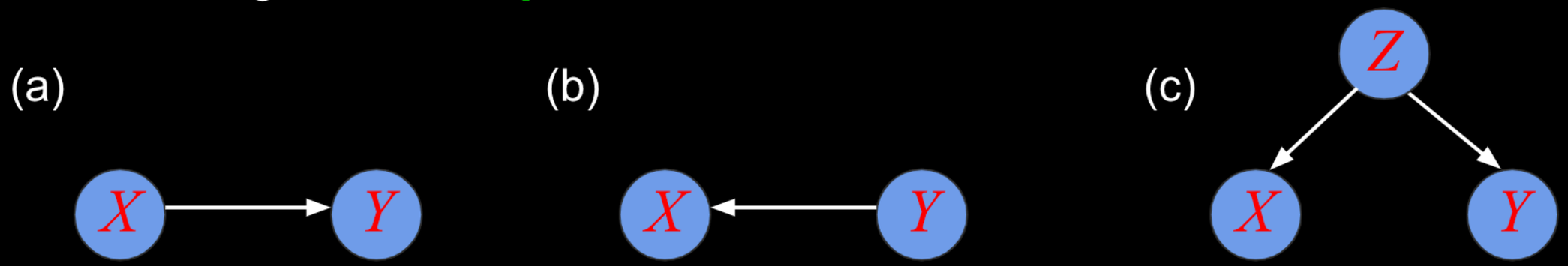
Judea Pearl. *Causality*, 2009.

Recap on Causal Inference

Association vs. Causation

Principle of Common Cause [Reichenbach, 1991]

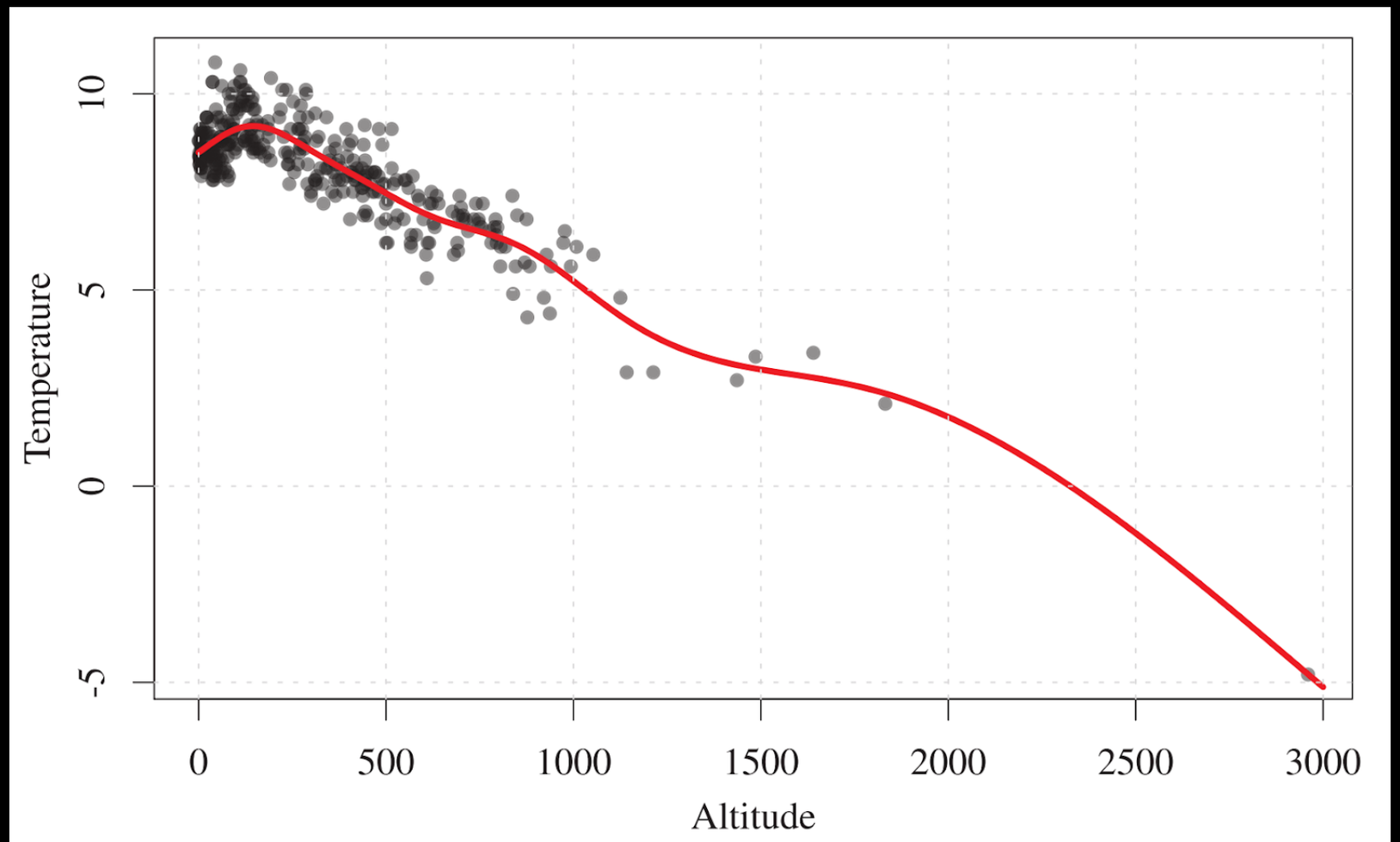
If two random variables X and Y are **statistically dependent**, then one of the following **causal explanations** must hold:



Causation has two obvious advantages:

1) Predict what would happen if some variables are *intervened*.
2) Predict the outcomes of cases that you *never observed before*.

Independent Causal Mechanism



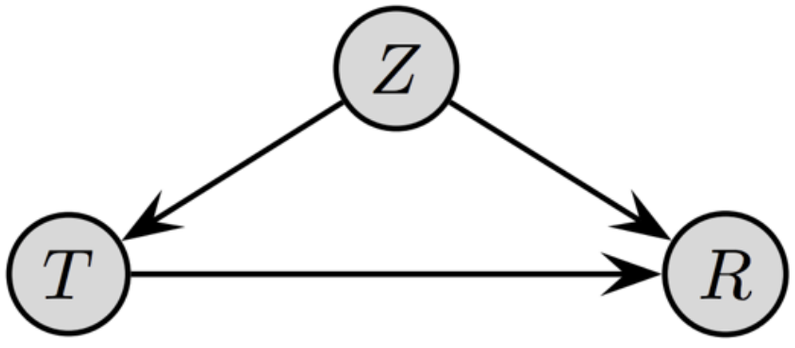
Credit: Elements of Causal Inference

$$p(a,t) = p(a|t)p(t) \qquad T \to A$$
$$= p(t|a)p(a) \qquad A \to T$$

Confounder

	Overall	Patients with small stones	Patients with large stones
Treatment *a*: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment *b*: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

Credit: Elements of Causal Inference

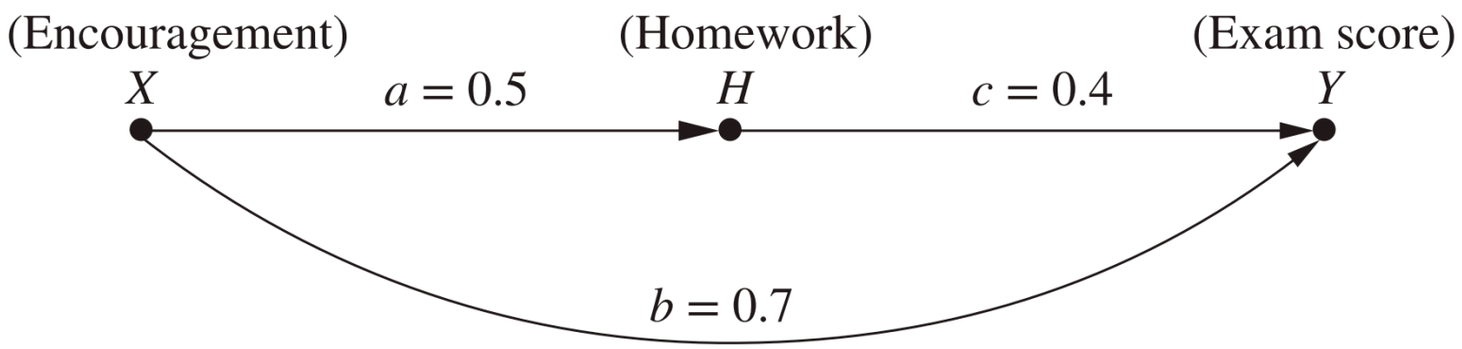


$$P(R=1 \mid \mathbf{do}(T=1)) = \sum_{z=\{0,1\}} P(R=1 \mid T=1, Z=z)P(Z=z)$$

One World vs. Two Worlds

Counterfactuals

Population vs. Individual



$$X = U_X$$

$$H = a \cdot X + U_H$$

$$Y = b \cdot X + c \cdot H + U_Y$$

Let us consider a student named Joe, for whom we measure X = 0.5, H = 1, and Y = 1.5. Suppose we wish to answer the following query: What would Joe's score have been had he doubled his study time?

$$U_X = 0.5,$$
$$U_H = 1 - 0.5 \cdot 0.5 = 0.75, \text{ and}$$
$$U_Y = 1.5 - 0.7 \cdot 0.5 - 0.4 \cdot 1 = 0.75.$$

$$\begin{aligned} Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) &= 0.5 \cdot 0.7 + 2.0 \cdot 0.4 + 0.75 \\ &= 1.90 \end{aligned}$$

Judea Pearl. *Causality*, 2009.

A Simple Taxonomy

Model	Predict in i.i.d. setting	Predict under changing distr. or intervention	Answer counterfactual questions	Obtain physical insight	Learn from data
Mechanistic/ physical, e.g., Sec. 2.3	yes	yes	yes	yes	?
Structural causal model, e.g., Sec. 6.2	yes	yes	yes	?	?
Causal graphical model, e.g., Sec. 6.5.2	yes	yes	no	?	?
Statistical model, e.g., Sec. 1.2	yes	no	no	no	yes

Credit: Elements of Causal Inference

Identification

- Identification in **Causal Reasoning**

Interventional prob. ⟶ Observational prob.

- Identification in **Causal Learning**

Uniqueness of Causal Orientation

- Identification in **Latent Confounder Models**

Uniqueness of Causal Strength

Recap on RL

Reinforcement Learning (RL)

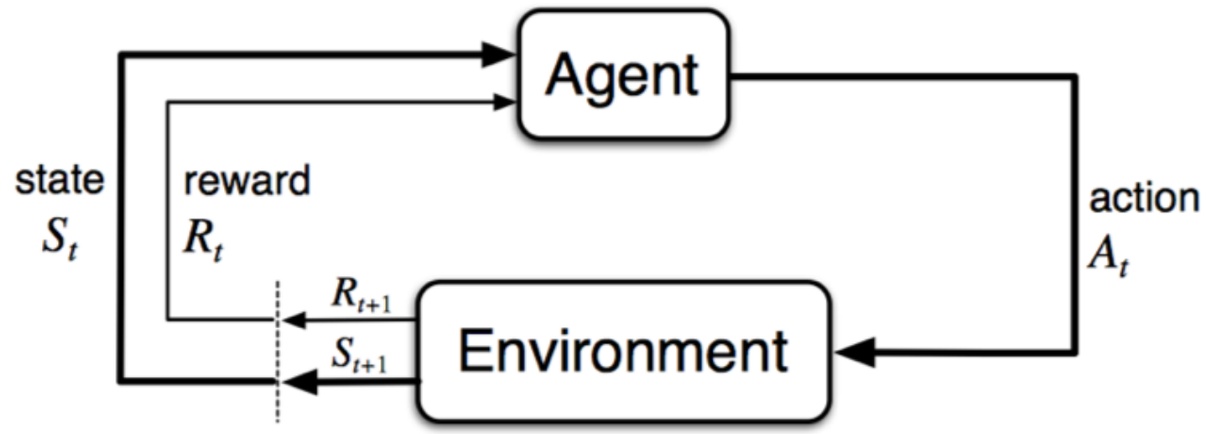


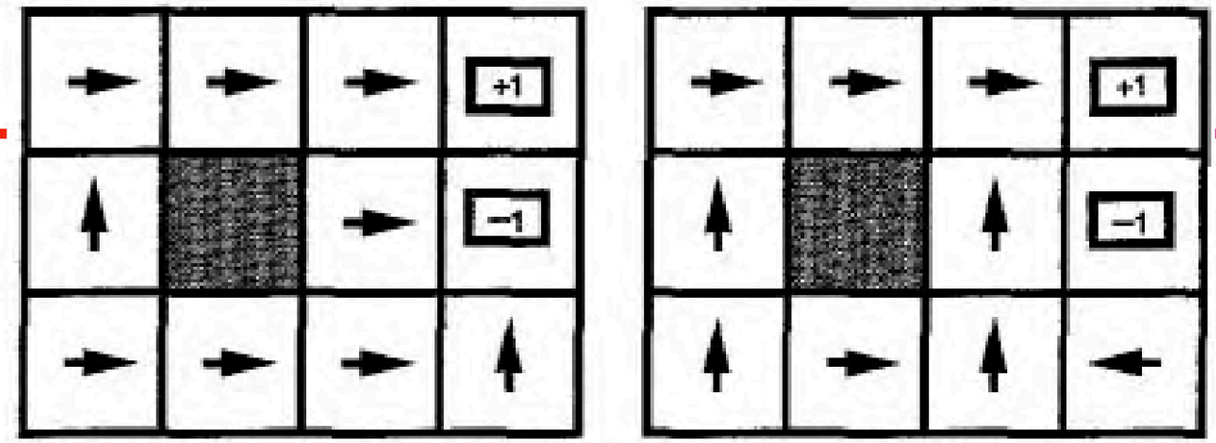
Figure 1: The agent-environment feedback loop [Sutton and Barto, 1998]

Hypothesis 1 (**The Reward Hypothesis**). *That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward).*

Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*, 2018.

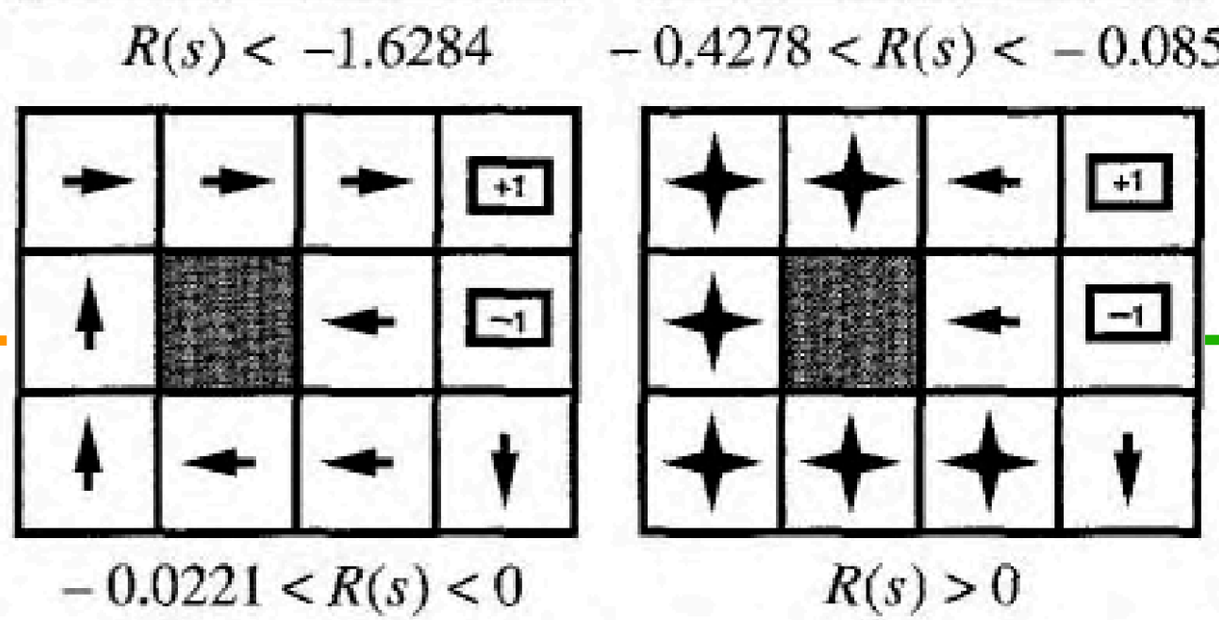
My Favourite

"Life is so painful that the agent heads straight for the nearest exit, even if the exit is worth -1."



"Life is quite unpleasant; the agent takes the shortest route to the +1 state and is willing to risk falling into the -1 state by accident."

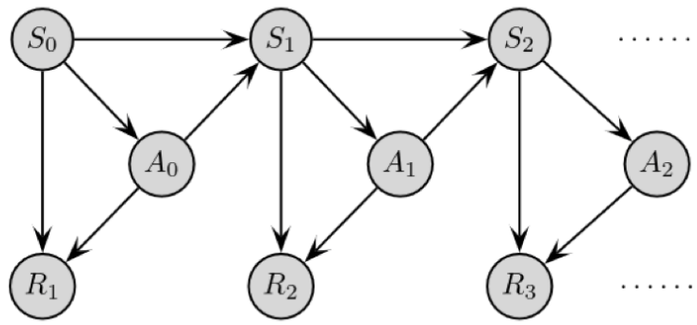
"When life is only slightly dreary, the optimal policy takes no risks at all, even though this means banging its head against the wall quite a few times."



"Life is positively enjoyable and the agent avoids both exits.."

Artificial Intelligence: A Modern Approach
Stuart J. Russell and Peter Norvig. 1995

Markov Decision Processes (MDPs)



Formally, MDPs can be defined by:

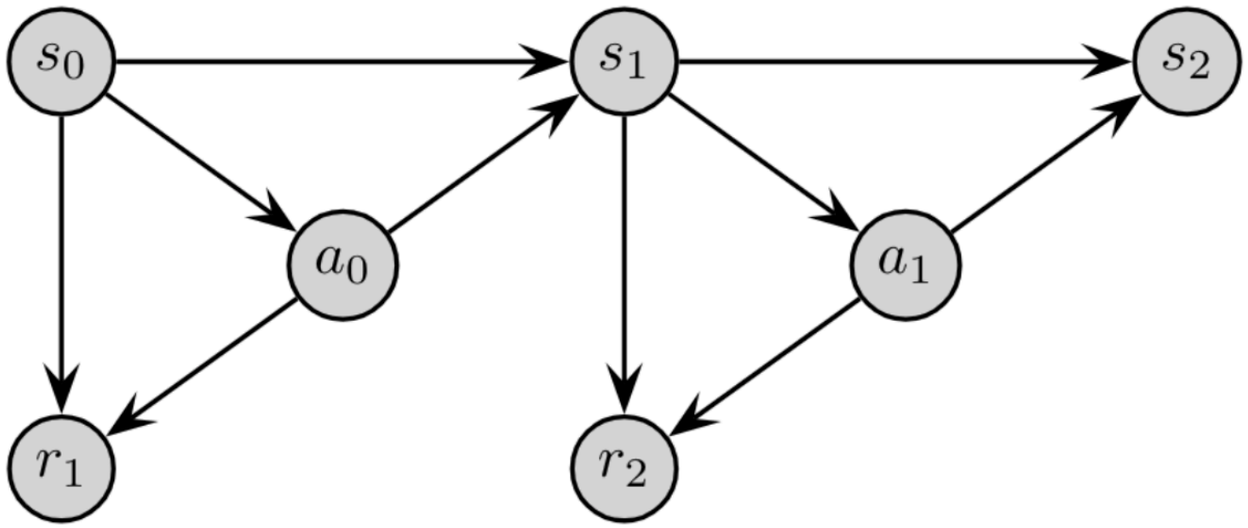
- $\mathcal{S}$, a set of states. $\mathcal{S}$ can be continuous $\mathbb{R}^{D_\mathcal{S}}$, or discrete $\{s_1, \ldots, s_{N_\mathcal{S}}\}$
- $\mathcal{A}$, a set of actions. $\mathcal{A}$ can be continuous $\mathbb{R}^{D_\mathcal{A}}$, or discrete $\{a_1, \ldots, a_{N_\mathcal{A}}\}$;
- $\mathcal{R}$, a set of rewards. $\mathcal{R}$ can be continuous $\mathbb{R}$, or discrete $\{r_1, \ldots, r_{N_\mathcal{R}}\}$;
- p: $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, a state-transition probability function, defining the probability of an agent who executes action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ resulting in a transition to sate $s' \in \mathcal{S}$, i.e., $p(s'|s, a)$.
- r: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$, an expected reward received when executing action $a \in \mathcal{A}$ from state $s \in \mathcal{S}$, that is,

$$r(s, a) \doteq \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a]; \tag{2.2}$$

- $0 \leq \gamma \leq 1$, a discount factor determining the present value of future rewards.

Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*, 2018.

MDPs



$$p(s_t|s_{t-1}, a_{t-1})$$

$$p(a_t|s_t)$$

$$p(r_t|s_{t-1}, a_{t-1})$$

MDPs

Definition 1 (**Markov Decision Process**). *A Markov decision process is a tuple* $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ *such that*

$$p(s', r|s, a) = Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\} \tag{2}$$

where $S_t \in \mathcal{S}$ *(state space),* $A_t \in A$ *(action space),* $R_t \in \mathcal{R}$ *(reward space) and* p *defines the dynamics of the process.*

Definition 2 (**Discounted Returns**). *Discounted Return is defined as the total sum of rewards following a time step* t *until the end of the sequence of rewards discounted by a factor* γ *at each time step*

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \tag{3}$$
$$G_t = R_{t+1} + \gamma G_{t+1}$$

where $R_i \in \mathcal{R}\ \forall i$ *and* $\gamma \in [0, 1]$.

Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*, 2018.

Value Functions

Definition 3 (**Policy**). *A policy is defined as the probability distribution of actions at a given states.*

$$\pi(A_t = a \mid S_t = s) \; \forall S_t \in \mathcal{S} \tag{4}$$

where $A_t \in \mathcal{A}(s)$ *is the state specific action space.*

Definition 4 (**State Value Function**). *Value function of a state* s *under policy* π *is defined as the expected return when starting in state* s *and following a policy* π *to take actions*

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[G_t \mid S_t = s\right] \; \forall s \in \mathcal{S} \tag{5}$$

Definition 5 (**Action Value Function**). *Value function of a state* s *and action* a *under policy* π *is defined as the expected return when starting in state* s*, taking action* a *and following a policy* π *to take actions further.*

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}\left[G_t \mid S_t = s, A_t = a\right] \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s) \tag{6}$$

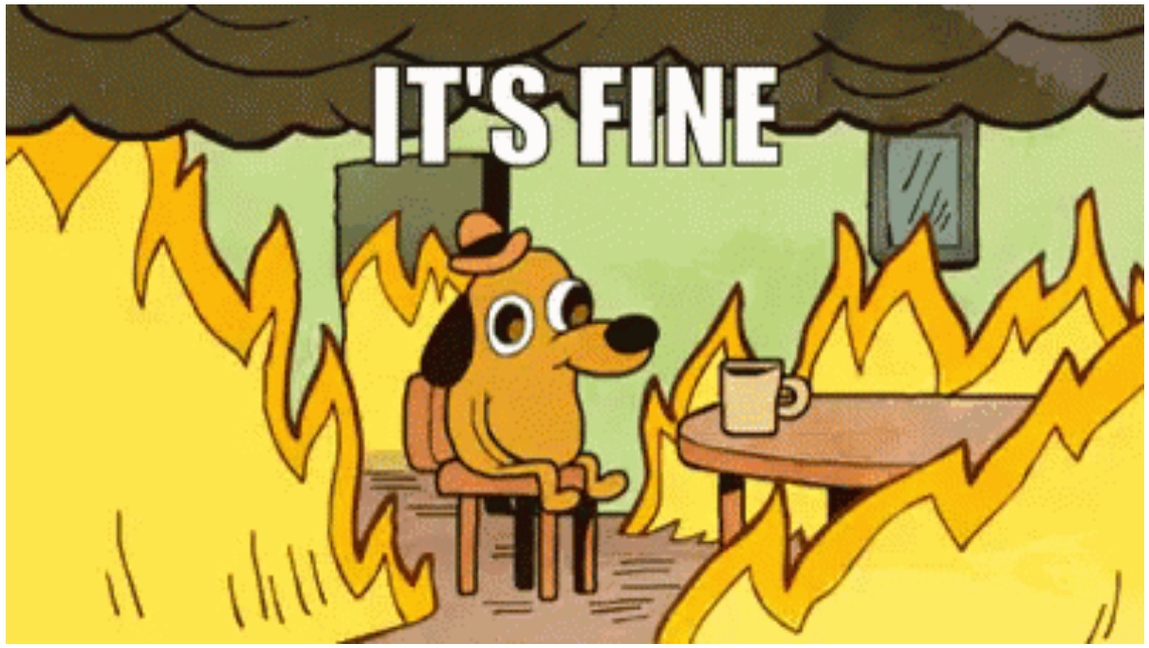
Richard Sutton and Andrew Barto. ***Reinforcement Learning: An Introduction***, 2018.

Terminology Comparisons

- Model-based versus Model-free
- On-policy versus Off-policy
- Online versus Offline/Batch Setting/Observational Setting
- Inverse RL (IRL) versus Imitation Learning (IL)
- Offline RL versus Imitation Learning (IL)
- POMDPs versus Predictive State Representations (PSRs)

What’s Wrong with RL?

Reinforcement Learning never worked, and 'deep' only helped a bit.



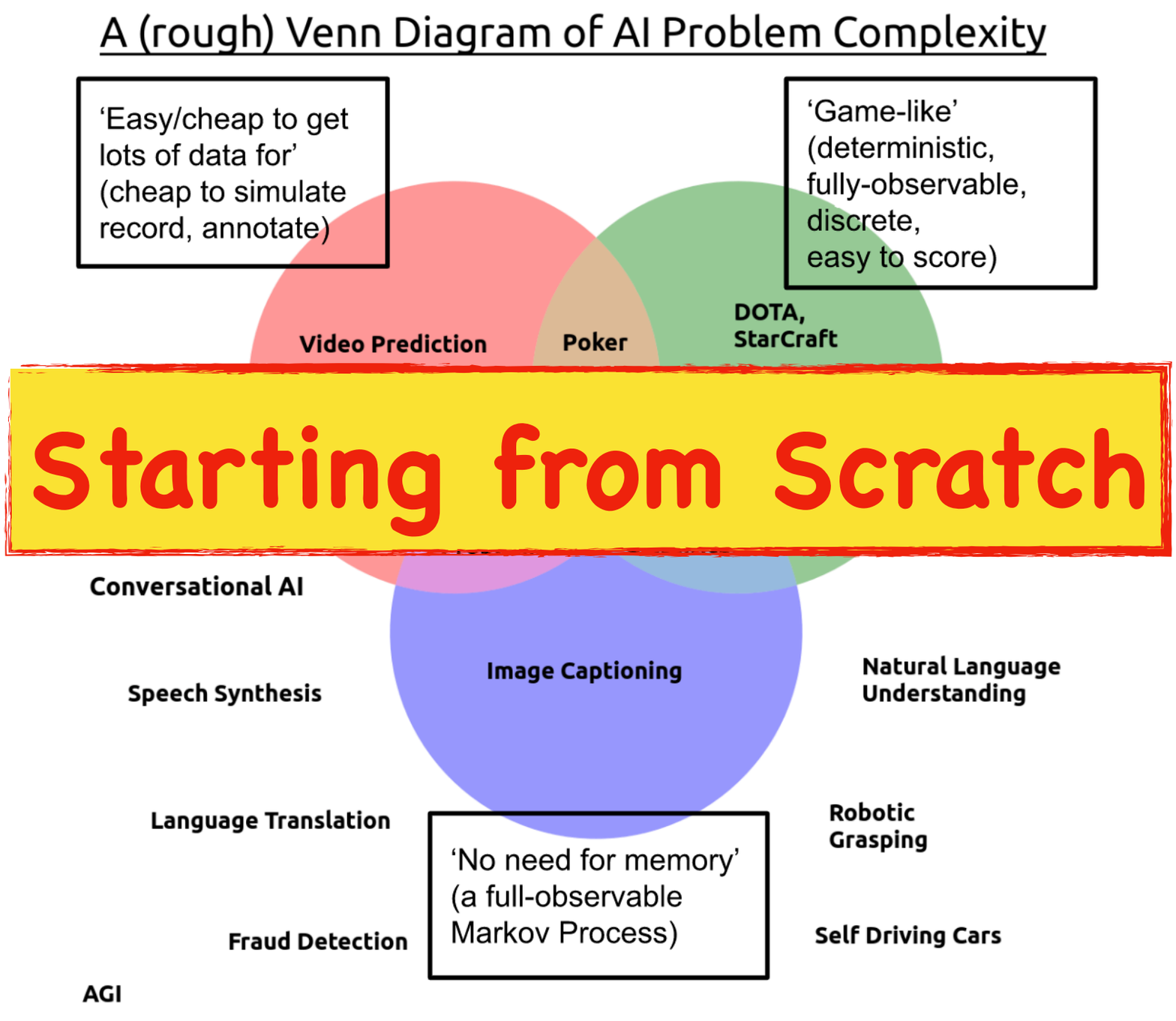
RL researchers all the time

Legit RL research request

Exploration and Long Term Credit Assignment

Himanshu Sahni's Blog

RL's Fundamental Flaw



Andrey Kurenkov's Blog

RL is a Cherry

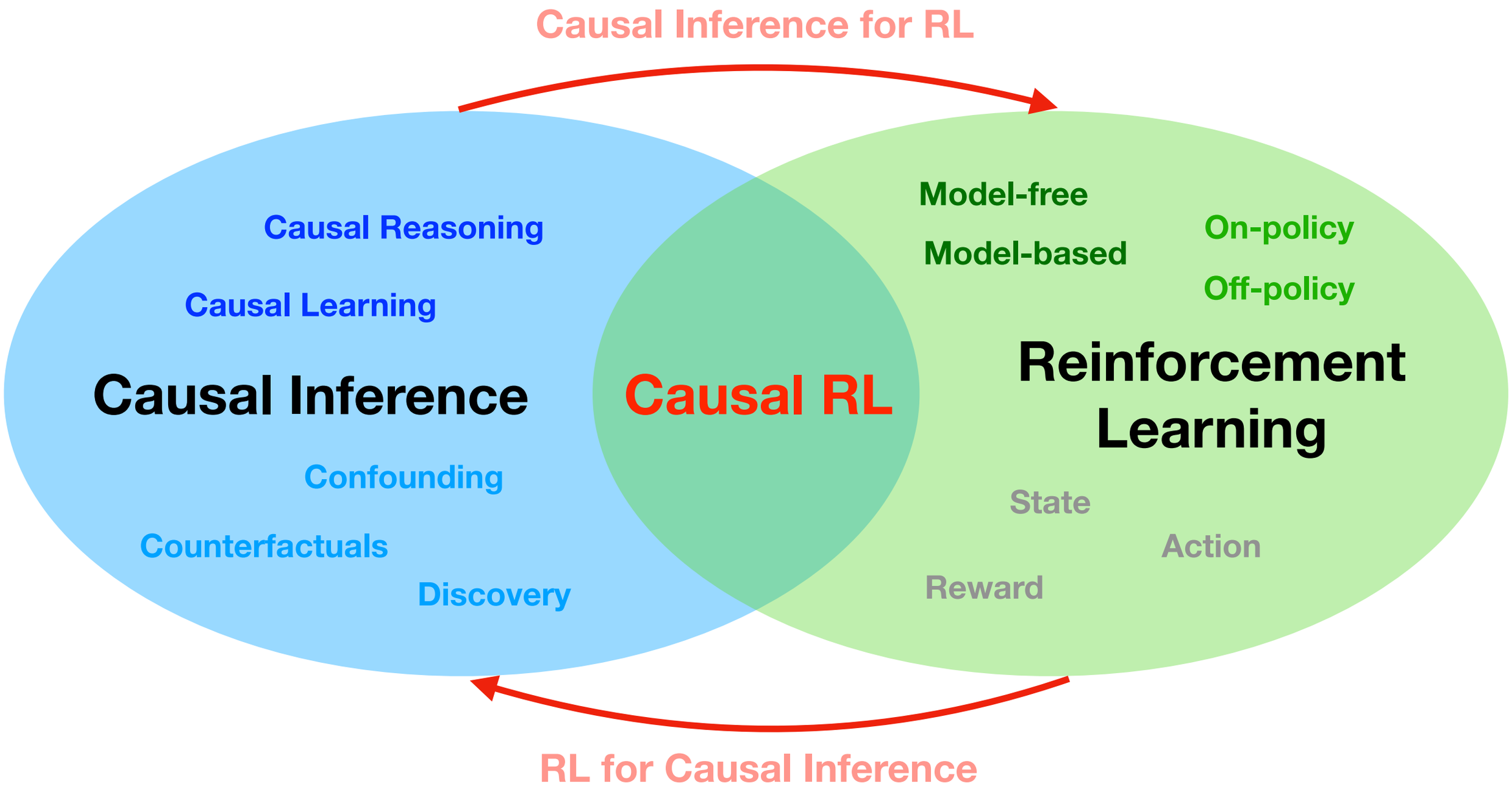
How Much Information is the Machine Given during Learning?

Y. LeCun

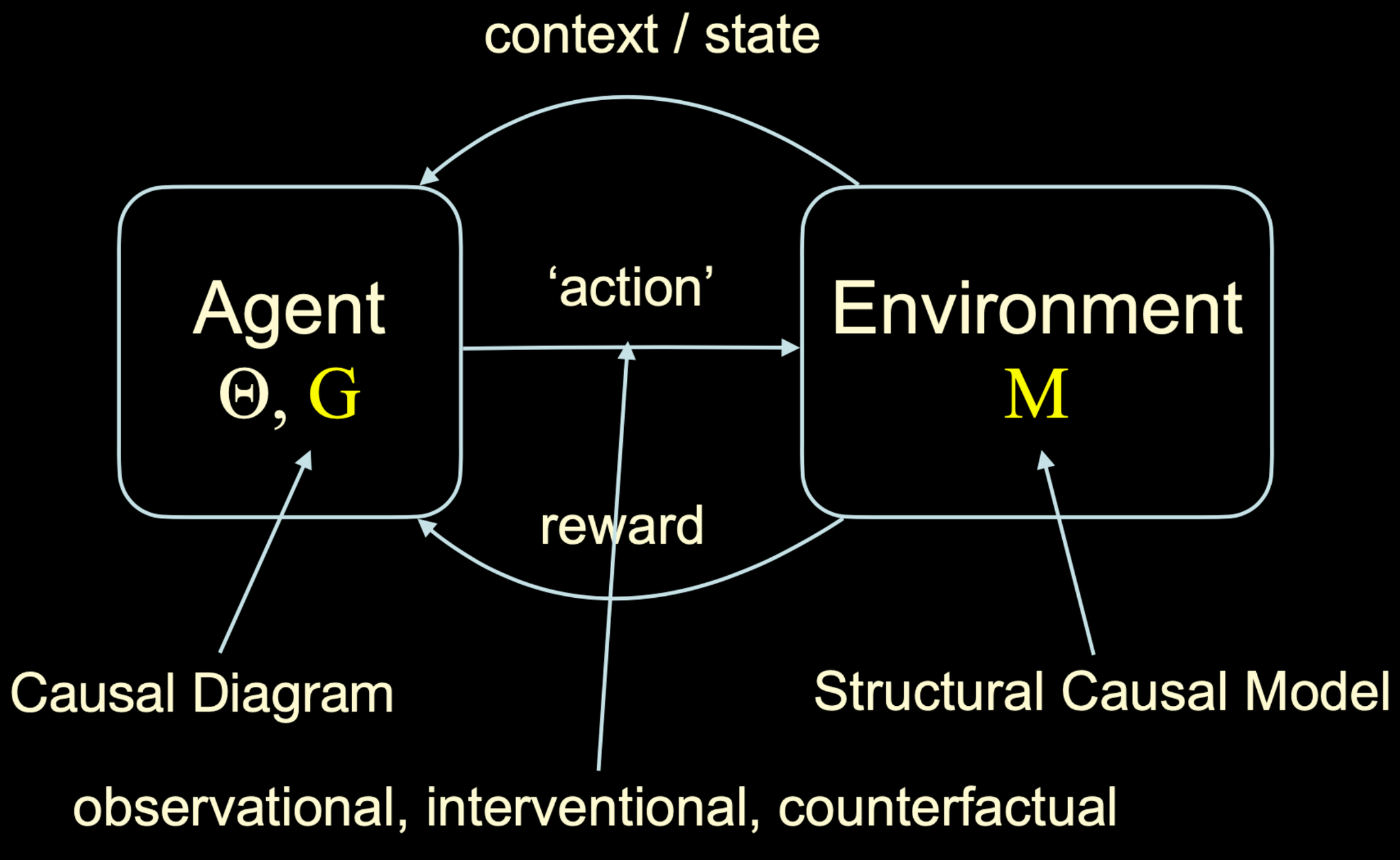
- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

What is Causal RL?

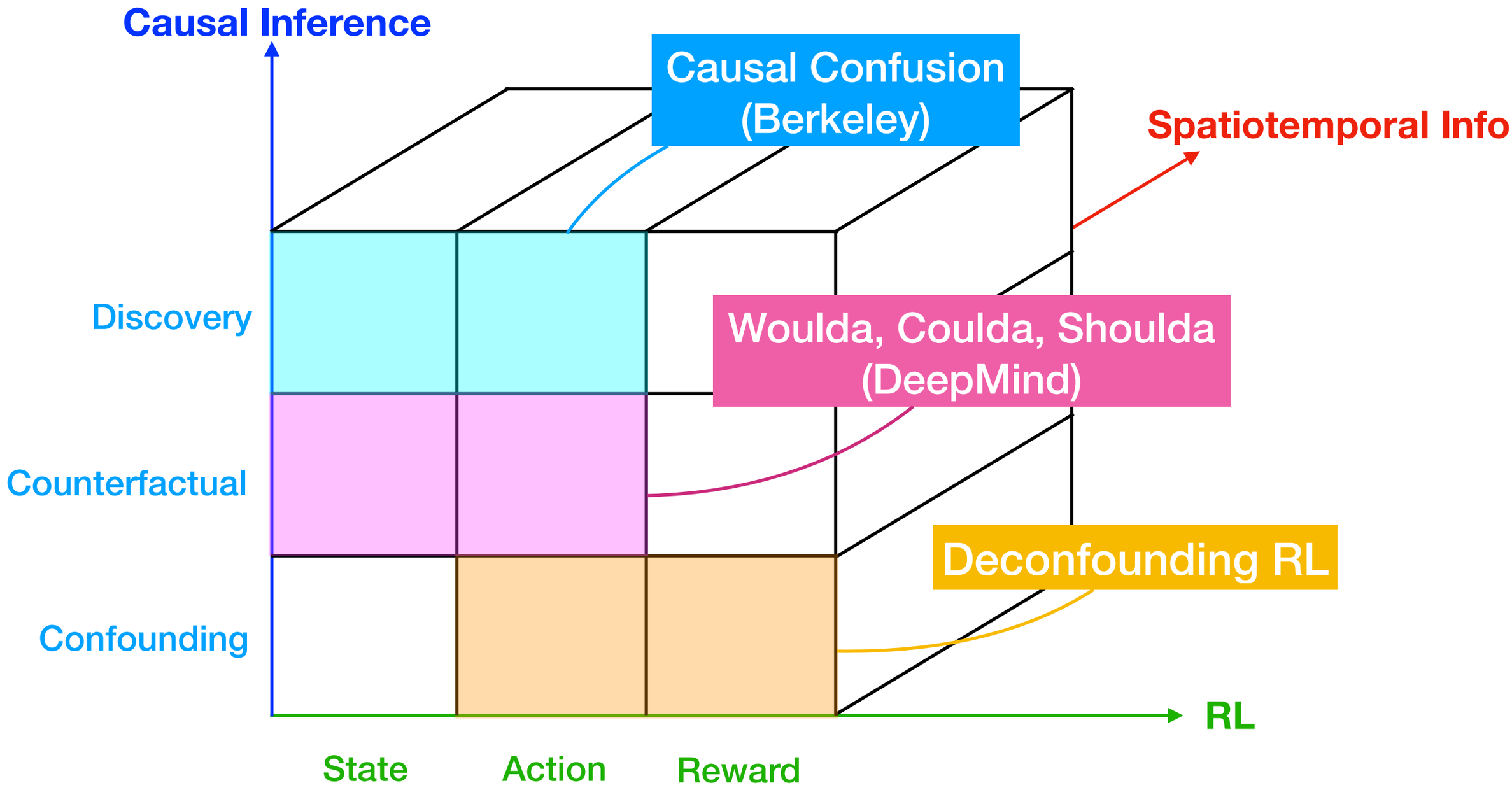
Causal RL I


Causal Inference for RL
Causal Reasoning
Causal Learning
Causal Inference
Confounding
Counterfactuals
Discovery
Causal RL
Model-free
Model-based
On-policy
Off-policy
Reinforcement Learning
State
Action
Reward
RL for Causal Inference

Causal RL II [Credit to Elias]



Causal RL III — Example



Why is Causal RL?

Why from RL

Is RL an exercise in causal inference? Of course! Albeit a restricted one. By deploying interventions in training, RL allows us to infer consequences of those interventions, but ONLY those interventions. A causal model is needed to go BEYOND, i.e., to actions not used in training.

The relation between RL and causal inference has been a topic of some debate. It can be resolved, I believe, by understanding the limits of each.

Question 1: why is RL on the original high-dimensional Atari games harder than on downsampled versions?
Question 2: why is RL easier if we permute the replayed data?

RL is closer to causality research than the machine learning mainstream in that it sometimes effectively directly estimates do-probabilities (on-policy learning). However, as soon as off-policy learning is considered, in particular in the batch (or observational) setting, issues of causality become subtle.

Why from Natural Science

AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019 | Science & Environment

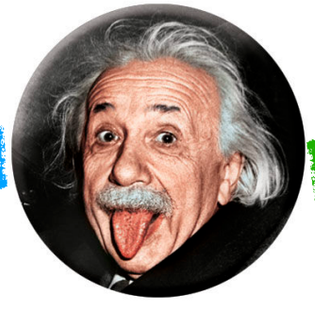
Reproducibility Crisis

Flawed Patterns

Experiments

Theory

Symmetry



Why from Cognition

*Humans **summarise rules or experience** from their **interaction** with nature and then **exploit this to improve their adaptation** in the next exploration.*

*What **Causal RL** does is exactly to **mimic human behaviours,** i.e., **learning causal relations** from an agent that **communicates** with the environment and then **optimising its policy based on the learned causal structures**.*

"*Our grasp of the world — the way we mirror its causal structure — is at the mercy of the inferential tools we have in the brain.*"

— JAKOB HOHWY

"*Play is the answer to how anything new comes about.*"

— JEAN PIAGET

"*All reasonings concerning matter of fact seem to be founded on the relation of cause and effect. By means of that relation alone we can go beyond the evidence of our memory and senses.*"

— DAVID HUME

The Debate on AGI

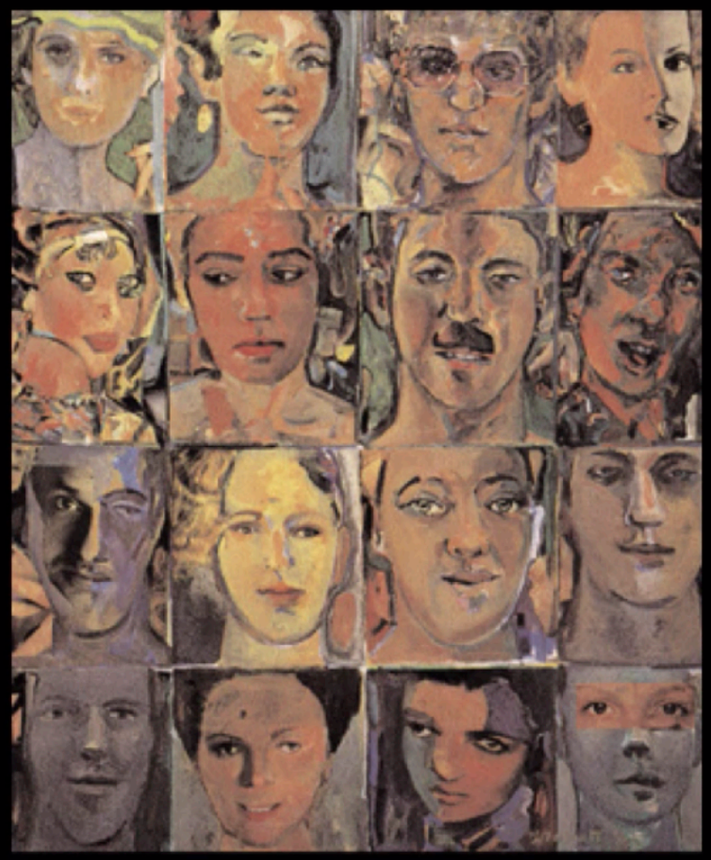
Does AI Need More Innate Machinery?

Marcus and LeCun in Complete Agreement on Seven Points

October 2017

- AI is still in its infancy
- Machine learning is fundamentally necessary for reaching strong AI
- Deep learning is a powerful technique for machine learning
- Deep learning is not sufficient on its own for cognition
- [model-free] Reinforcement learning is not the answer, either
- AI systems still need better internal forward models
- Commonsense reasoning remains fundamentally unsolved

Some basics that evolution might have endowed humans with



- Representations of objects
- Structured, algebraic representations
- Operations over variables
- A type-token distinction
- A capacity to represent sets, locations, paths tracjectories, obstacles and enduring individuals
- A way of representing the affordances of objects
- Spatiotemporal contiguity / conservation of mass
- Causality
- Translational invariance
- Capacity for cost-benefit analysis

Questions

- **"All of these AI systems we see, none of them is 'real' AI**

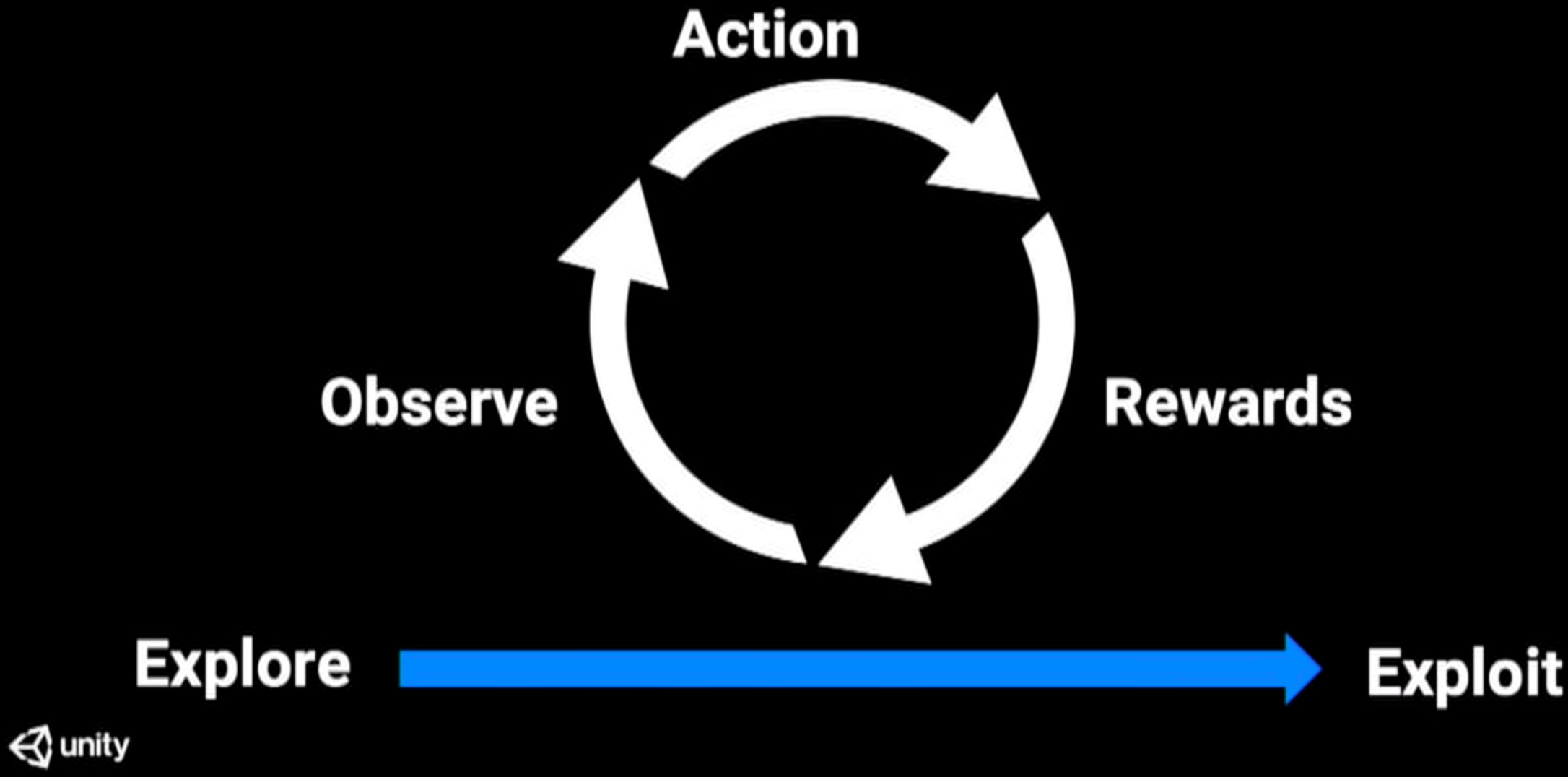
  – Josh Tennenbaum at CCN 2017

  - I agree (Josh and I start our talks the same way).

- **The brain learns with an efficiency that none of our machine learning methods can match.**
  - Our supervised learning systems require large numbers of example
  - Our reinforcement learning systems require millions of trials
  - that's why we don't have robots that as agile as a cat or a rat
  - that's why we don't have dialog systems that have common sense
- **What is missing?**

  Learning paradigms that build (predictive) models of the world through observation and action.

Nature's Learning Method: Reinforcement



GOTO 2018 • On the Road to Artificial General Intelligence • Danny Lange

Key Concepts in Causal RL

State, Action, and Reward

Reward	state, action $\rightarrow$ reward
Transition	state, action $\rightarrow$ state
Hidden state	hidden state $\rightarrow$ observation

Table 1: Summary of causal relationships in reinforcement learning.

Policy is NOT a causal relationship.

Samuel J. Gershman. ***Reinforcement Learning and Causal Models*****, 2015**

Soft vs Hard Intervention

Conditional Actions

$$P(Y|\mathbf{do}(X=g(Z))) = \sum_Z P(Y|\mathbf{do}(X=g(Z)),Z)P(Z|\mathbf{do}(X=g(Z)))$$
$$= \sum_Z P(Y|\mathbf{do}(X),Z)|_{X=g(Z)}P(Z)$$
$$= \mathbb{E}_Z[P(Y|\mathbf{do}(X),Z)|_{X=g(Z)}]$$

Stochastic Policies

$$P(Y|\mathbf{do}(\pi(X|Z))) = \sum_X \sum_Z P(Y|\mathbf{do}(X),Z)P(X|Z)P(Z)$$

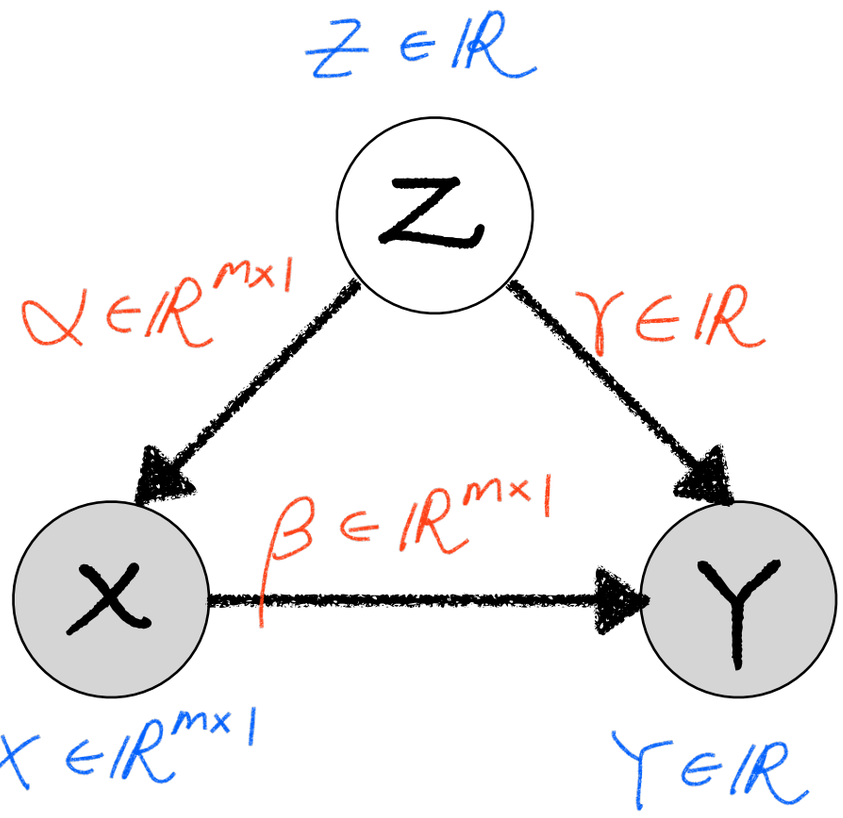
Judea Pearl. *Causality*, 2009.

Challenges in Causal Inference for RL

Chaochao Lu

Cambridge Machine Learning Group

Latent Confounding



$$
\begin{aligned}
Z &:= \epsilon_Z \\
X &:= \alpha' Z + \epsilon_X \\
Y &:= \beta' X + \gamma Z + \epsilon_Y
\end{aligned}
$$

$\epsilon_W \sim N(0, \sigma_W^2)$ for $W \in \{X, Y, Z\}$

$\epsilon_X \in \mathbb{R}^{m\times 1}$, $\epsilon_Z, \epsilon_Y \in \mathbb{R}$

$$
\Sigma_{XYZ} = \begin{pmatrix} \Sigma_{ZZ} & \Sigma_{ZX} & \Sigma_{ZY} \\ \Sigma_{XZ} & \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YZ} & \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}
$$

$\mathbb{R}^{m\times m}$

$\mathbb{R}^{m\times 1}$

$\mathbb{R}^{1\times 1}$

$$
\begin{aligned}
\Sigma_{XX} &= \alpha\alpha'\sigma_Z^2 + \operatorname{diag}(\sigma_X^2) \\
\Sigma_{XY} &= \Sigma_{XX}\beta + \gamma\sigma_Z^2\alpha \\
\Sigma_{YY} &= (\beta'\alpha + \gamma)^2\sigma_Z^2 + \beta' \operatorname{diag}(\sigma_X^2)\beta + \sigma_Y^2
\end{aligned}
$$

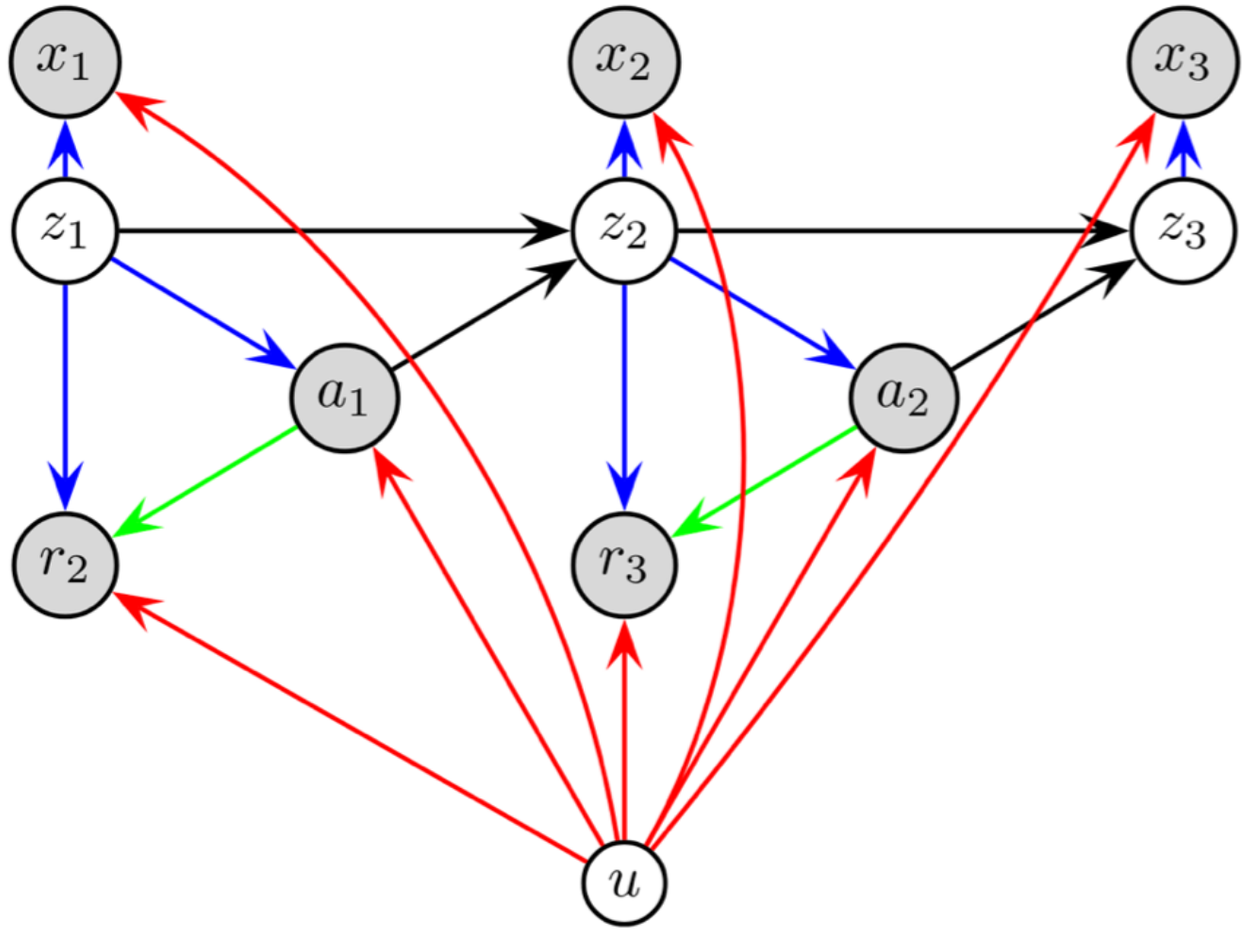
When $m > 3$

$$
(\alpha_1, \beta_1, \gamma_1, \sigma_{Z,1}^2, \sigma_{X,1}^2, \sigma_{Y,1}^2) \neq (\alpha, \beta, \gamma, \sigma_Z^2, \sigma_X^2, \sigma_Y^2)
$$

$3m + 3$ unknown parameters

[Alex D'Amour's Blog]

Deconfounding RL



$$p(r_{t+1} \mid z_t, a_t) \rightarrow p(r_{t+1} \mid z_t, do(a_t))$$

Lu et al. ***Deconfounding Reinforcement Learning in Observational Settings***, 2018

Counterfactuals

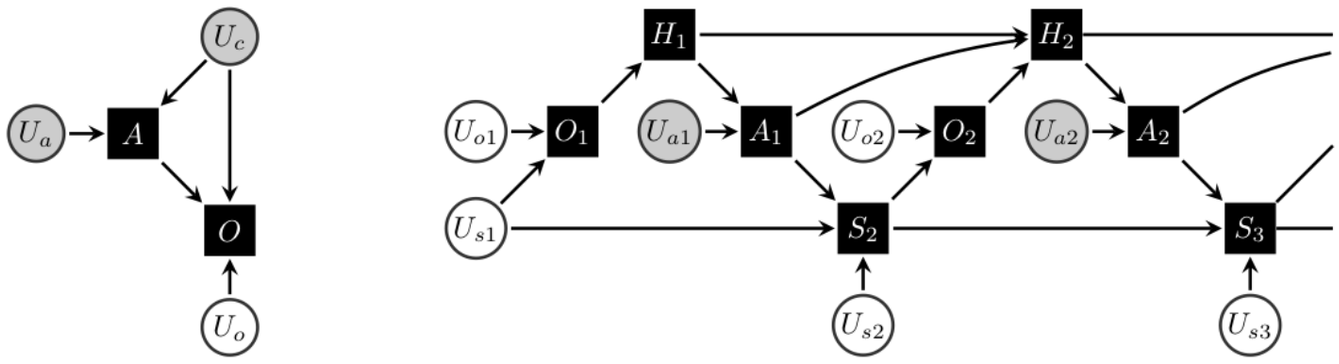
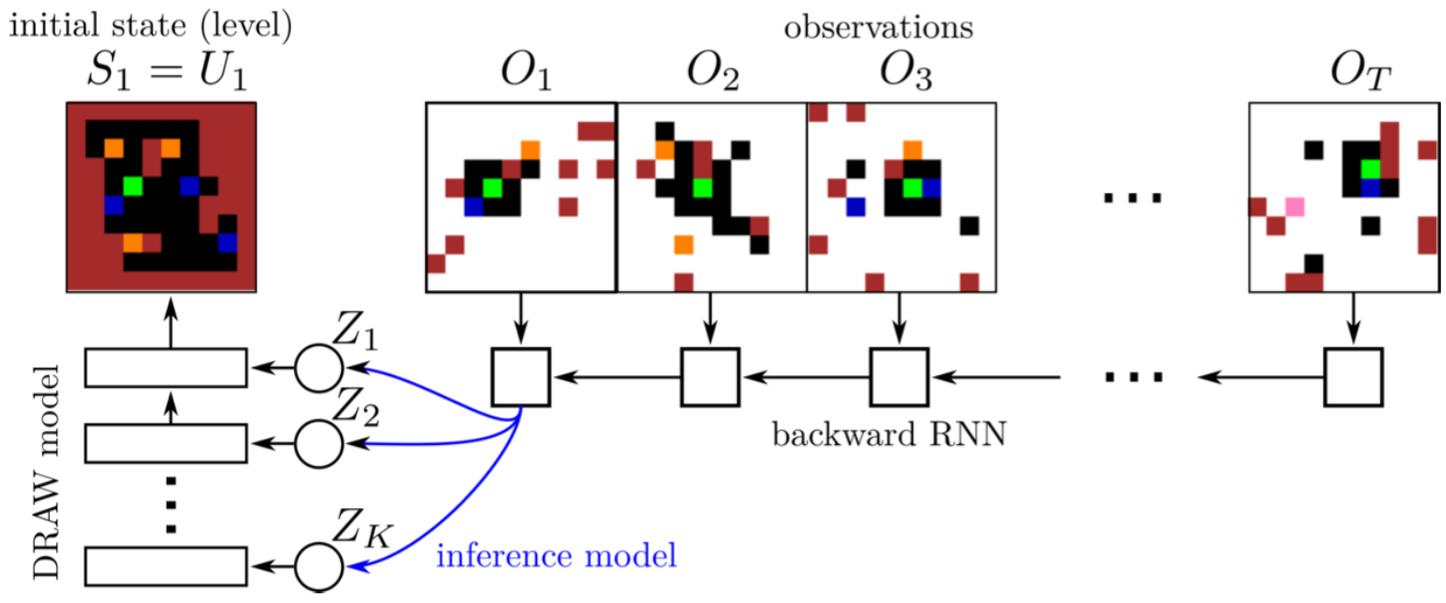
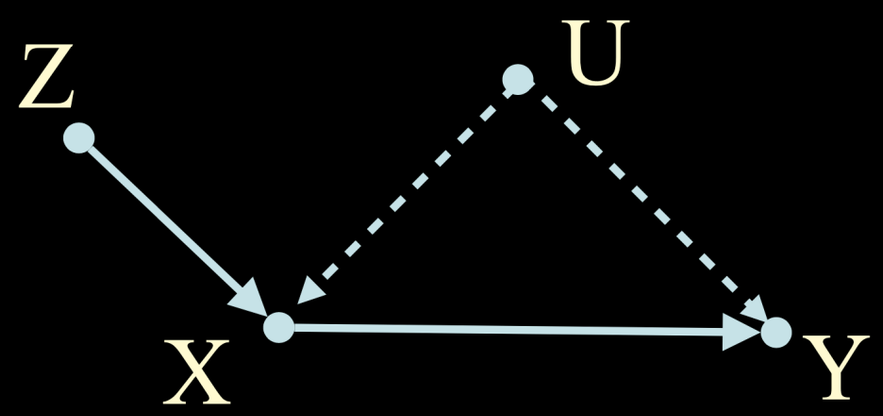


Figure 1: Structural causal models (SCMs) model environments using random variables U (circles, 'scenarios'), that summarize immutable aspects, some of which are observed (grey), some not (white). These are fed into deterministic functions f_i (black squares) that approximate causal mechanisms. **Left:** SCM for a contextual bandit with context U_c, action A, feedback O and scenario U_o. **Right:** SCM for a POMDP, with initial state $U_{s1} = S_1$, states S_t and histories H_t. The mechanism that generates the actions A_t is the policy π.



Buesing et al. *Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search*, 2019

Where to Intervene and to See



$$E[\,Y \mid do(x,z)\,] = E[\,Y \mid do(x)\,]$$

$\because$ $(Y \perp Z \mid X)$ in $G_{\overline{X},\underline{Z}}$ (Rule 3 of do-calculus)

Implication: prefer playing do(X) to playing do(X, Z).

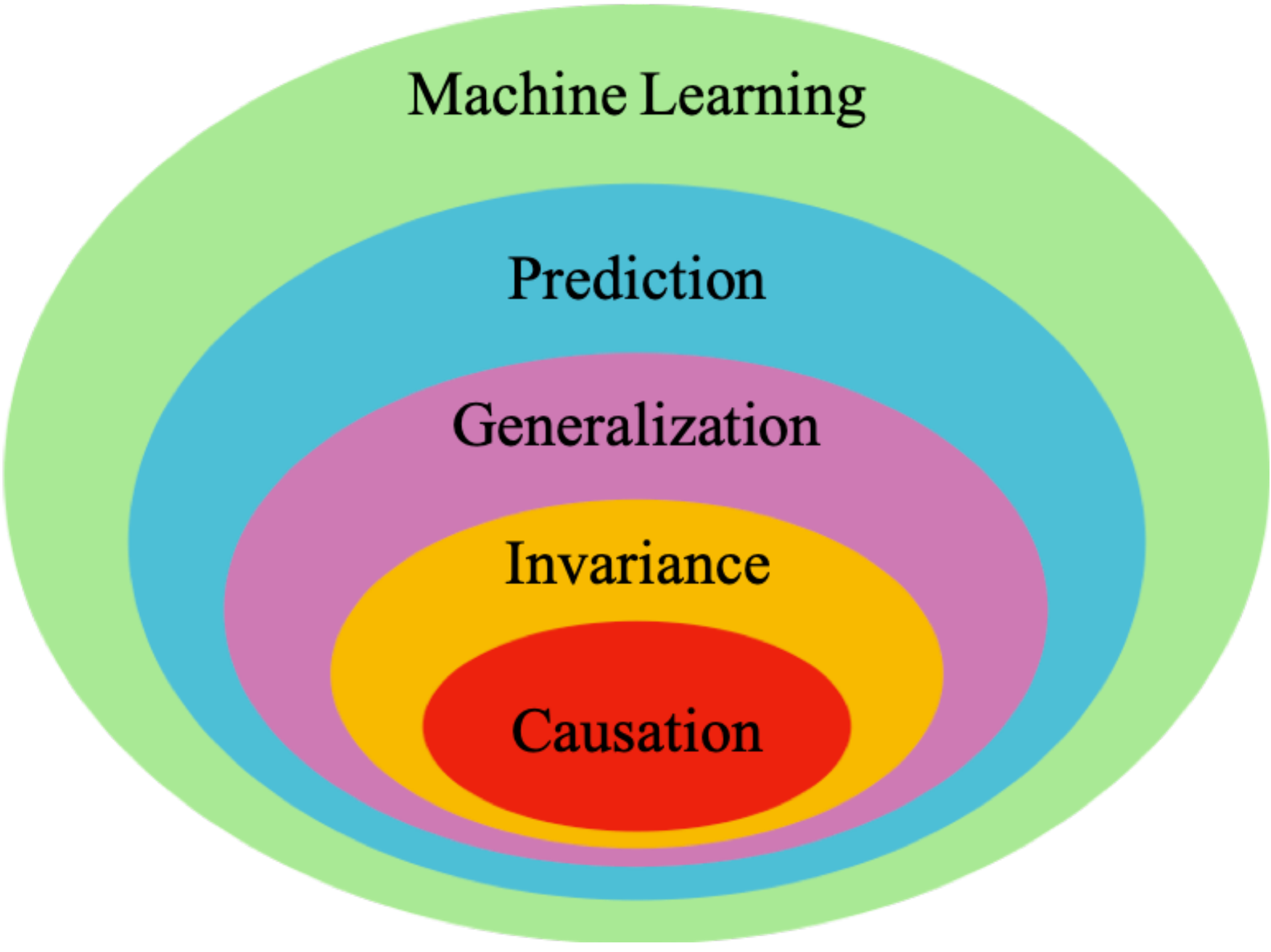
$$E[Y] = \sum_z E[Y|do(z)]\,P(z)$$
$$\leq \sum_z E[Y|do(z^*)]\,P(z)$$
$$= E[Y|do(z^*)] \qquad z^* \equiv \text{argmax}_z\, E[Y|do(z)]$$

$$\therefore E[Y] \leq E[Y|do(z^*)]$$

Implication: playing do(Z) should be preferred to playing do().

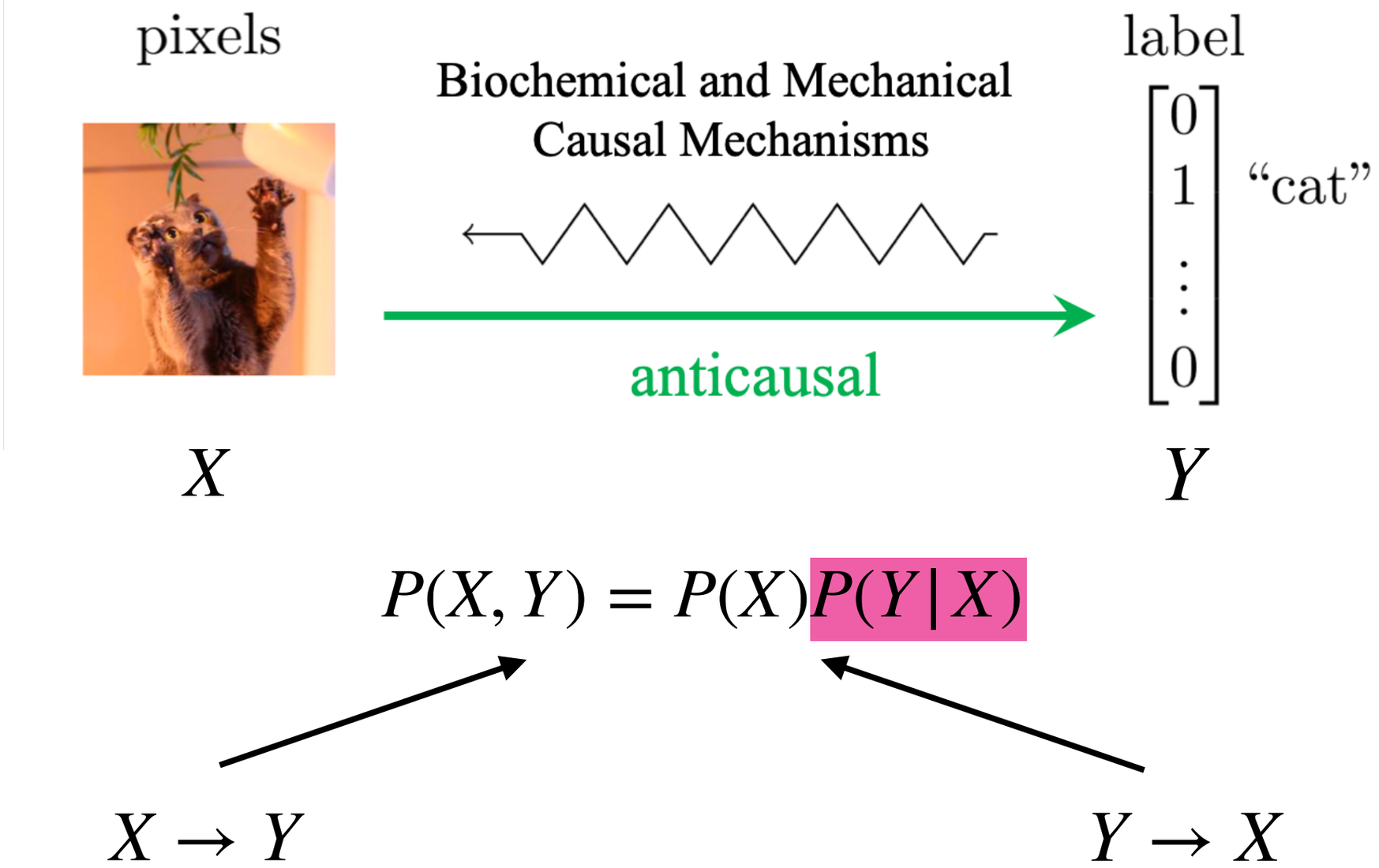
Lee et al. ***Structural Causal Bandits: Where to Intervene***, 2018

Causal Representation



Chaochao Lu. *The Agnostic Hypothesis: A Unifying View of Machine Learning*, 2020

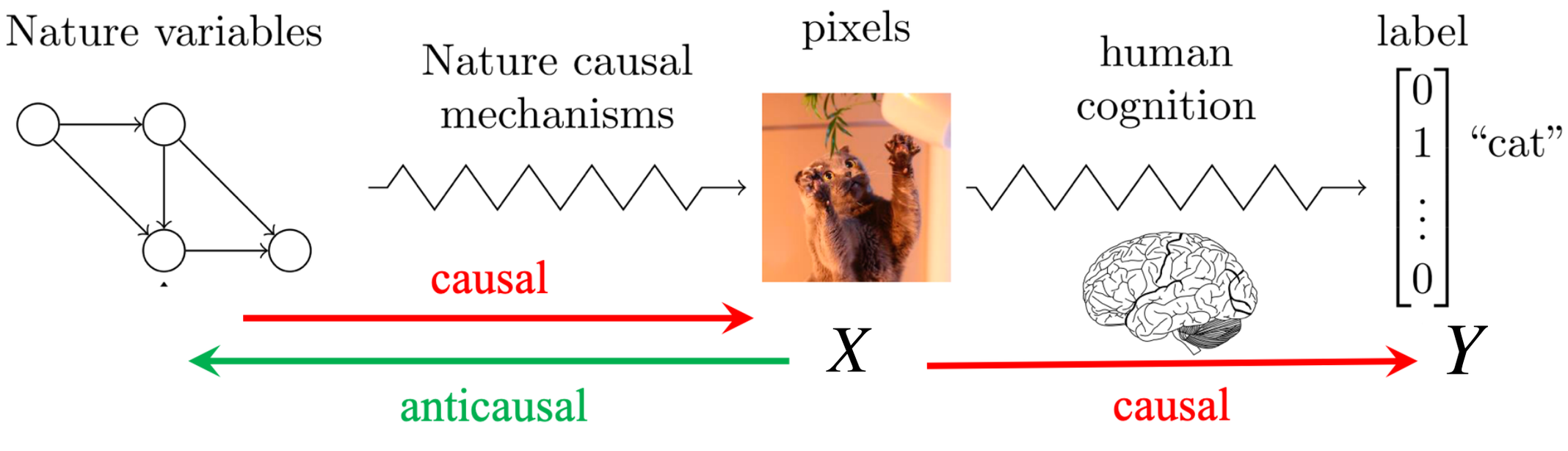
Opinion 1: Anticausal



Schölkopf et al. *On Causal and Anticausal Learning*, 2012

Chaochao Lu. *Is Image Classification a Causal Problem?*, 2020

Opinion 2: Causal



- In the causal direction, Nature variables (e.g., colour, light, angle, animal, etc.) produce images through nature causal mechanisms.
- In the anticausal direction, we attempt to disentangle the underlying causal factors of variation behind images (i.e., Nature variables).
- Disentanglement vs. Inference
- Hierarchy of Nature Variables vs. Occam's Razor

- predict human annotations from images in order to imitate the cognitive process (i.e., humans produce labels by following a causal and cognitive process after observing images.)
- $P(Y \mid X)$ should be stable across environments or domains. Hence, empirical risk minimisation (ERM) should work quite well.

Arjovsky et al. *Invariant Risk Minimization*, 2019

Chaochao Lu. *Is Image Classification a Causal Problem?*, 2020

The Agnostic Hypothesis

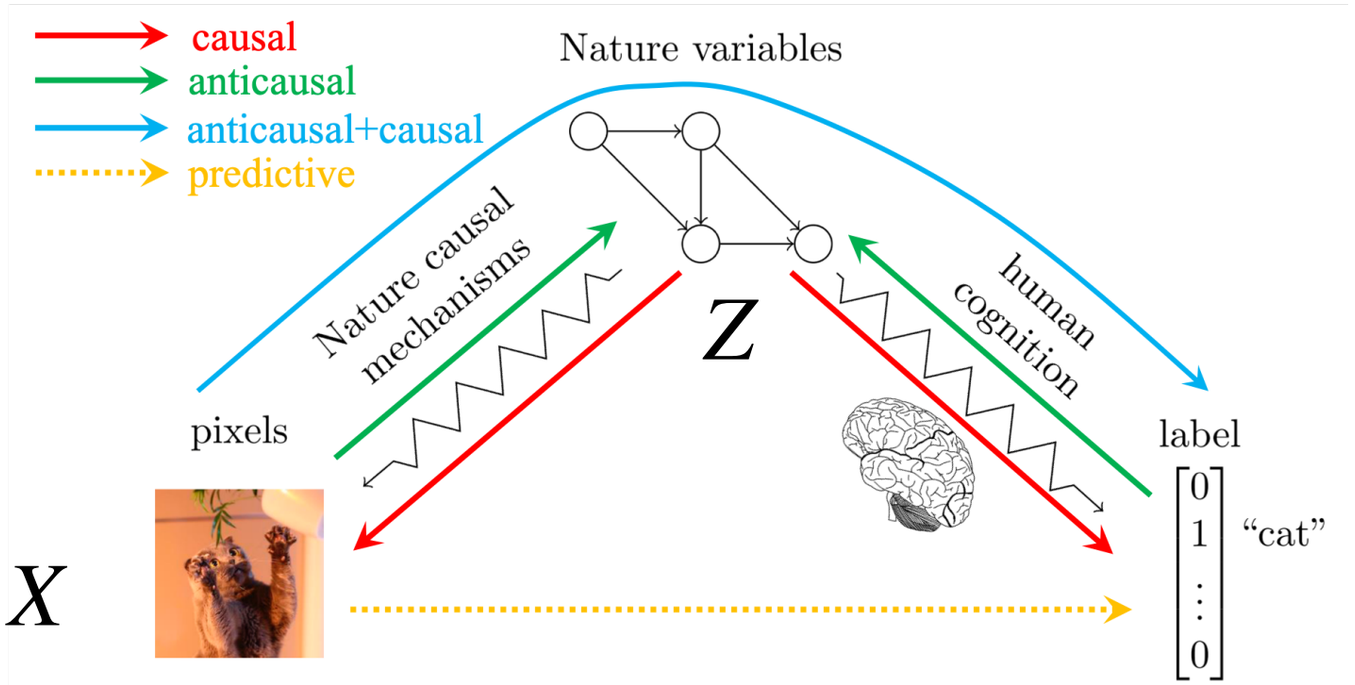


image and label as two different representation spaces

$X \leftarrow Z \rightarrow Y$

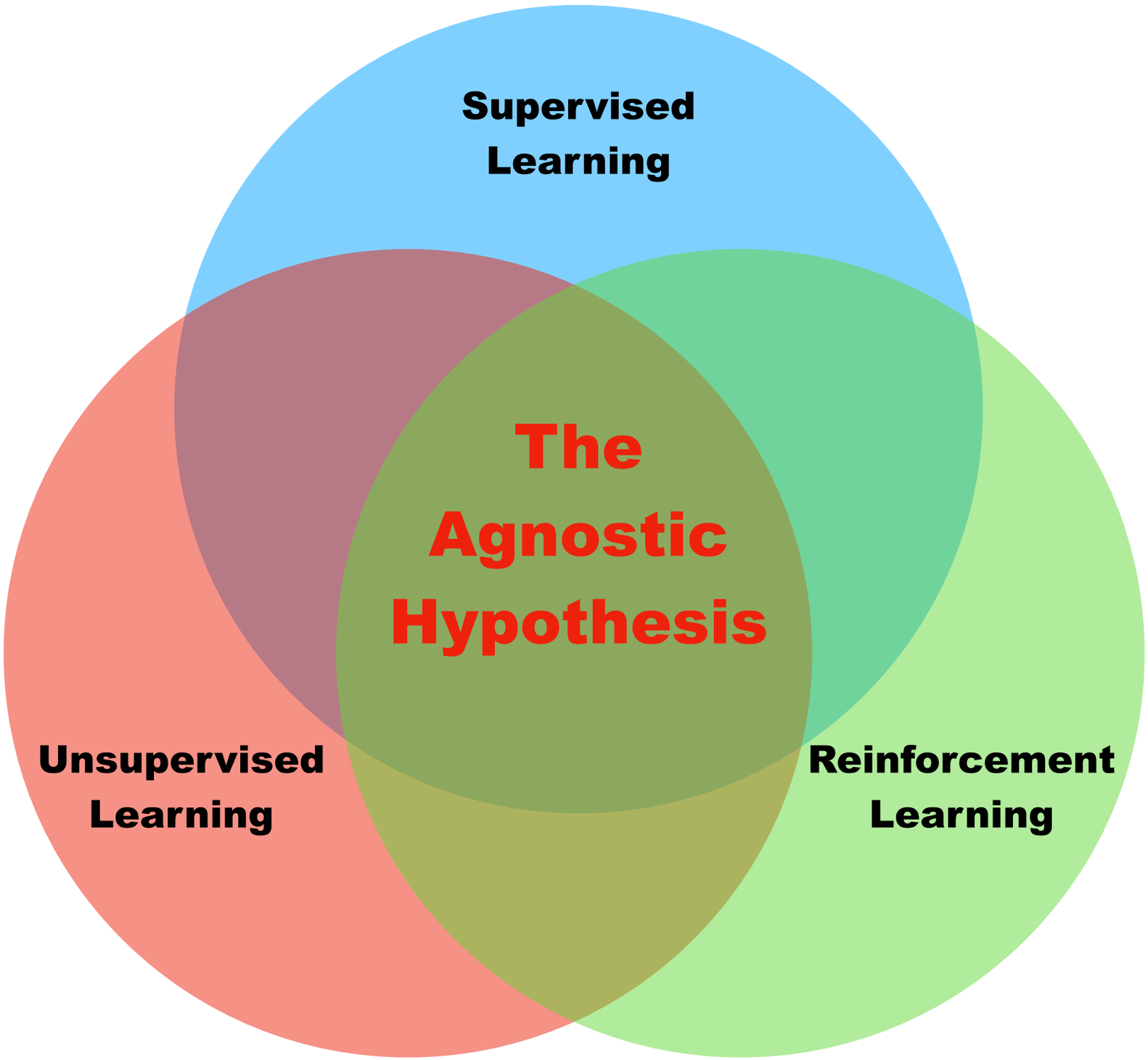
a feature extractor and a classifier

- **Plato's Theory of Forms** (Credit to Hannes)
- **Manipulability Theory**
- **Principle of Common Cause**
- **Theory of Linguistics** (Credit to Rebecca)

Chaochao Lu. ***Is Image Classification a Causal Problem?***, 2020

Chaochao Lu. ***The Agnostic Hypothesis: A Unifying View of Machine Learning***, 2020

Connections to ML



Chaochao Lu. *The Agnostic Hypothesis: A Unifying View of Machine Learning*, 2020

Independently Controllable Features

(Bengio et al. 2017)

Autoencoder

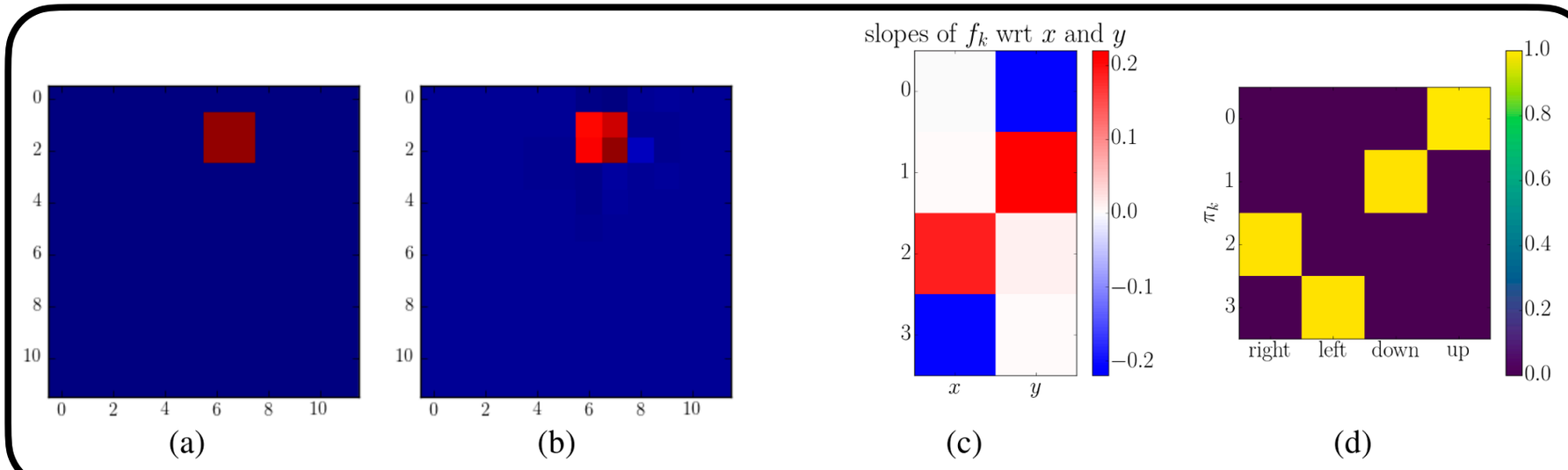
$$\min_{\theta} \frac{1}{2}\|x - g(f(x))\|_2^2$$

Selectivity

$$sel(s,a,k) = \mathbb{E}_{s' \sim \mathcal{P}^a_{ss'}} \left[\frac{|f_k(s') - f_k(s)|}{\sum_{k'} |f_{k'}(s') - f_{k'}(s)|} \right]$$

Objective

$$\underbrace{\mathbb{E}_s[\tfrac{1}{2}||s - g(f(s))||_2^2]}_{\text{reconstruction error}} - \lambda \underbrace{\sum_k \mathbb{E}_s[\sum_a \pi_k(a|s) \log sel(s,a,k)]}_{\text{disentanglement objective}}$$



Challenges in RL for Causal Inference

Intractability

d	Number of DAGs with d nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733202876993030479964123523138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	33277190122710759173617757331126112588358307625842190258354677350 5

Peters et al. *Elements of Causal Inference.* 2017

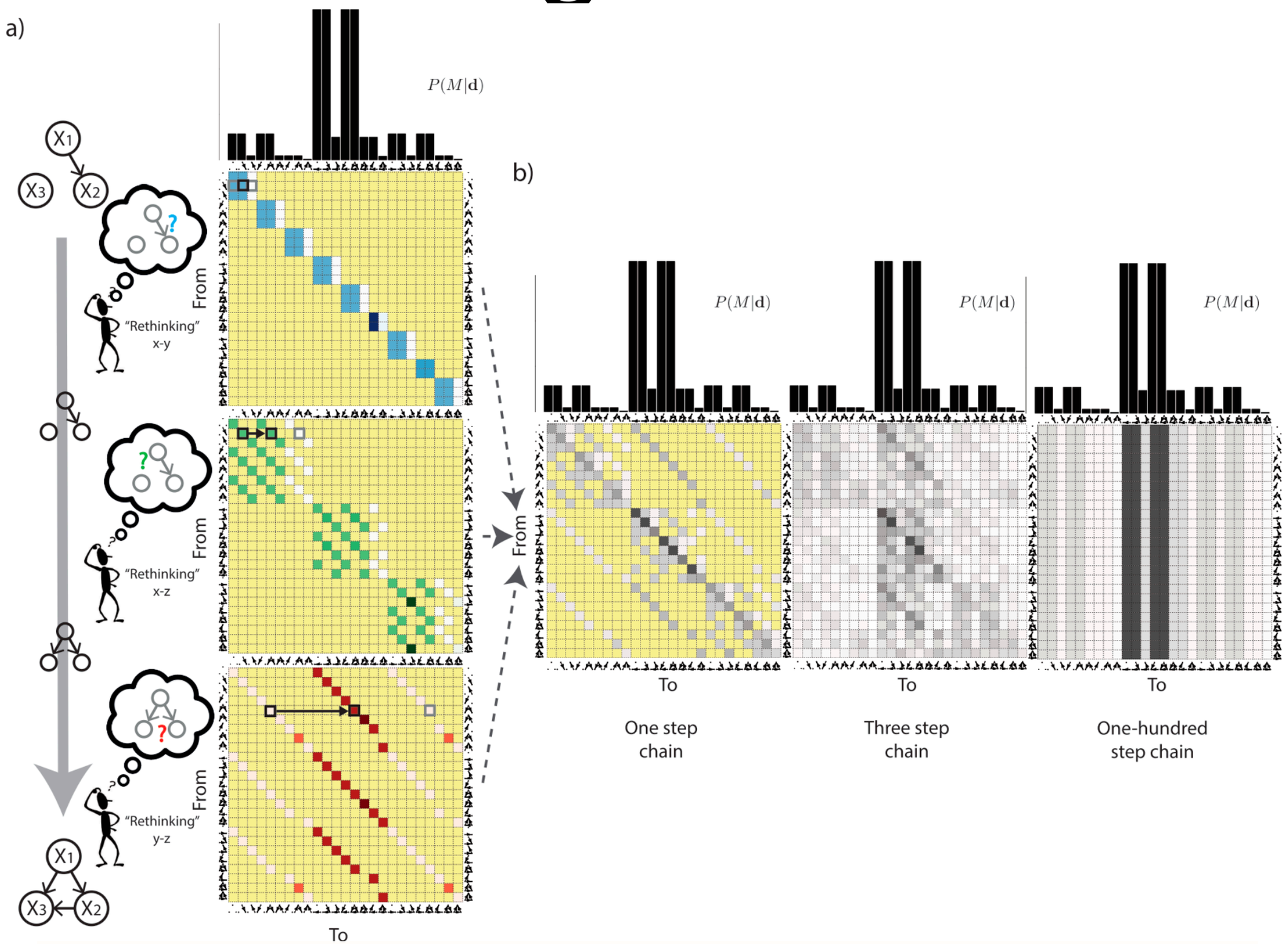
Neurath's Ship

> "[Learners] *are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support. In this way, by using the old beams and driftwood the ship can be shaped entirely anew, but only by gradual reconstruction.*"
>
> — WILLARD V. O. QUINE

Inference + Intervention

Neil Bramley. *Constructing the world: Active causal learning in cognition.* 2017

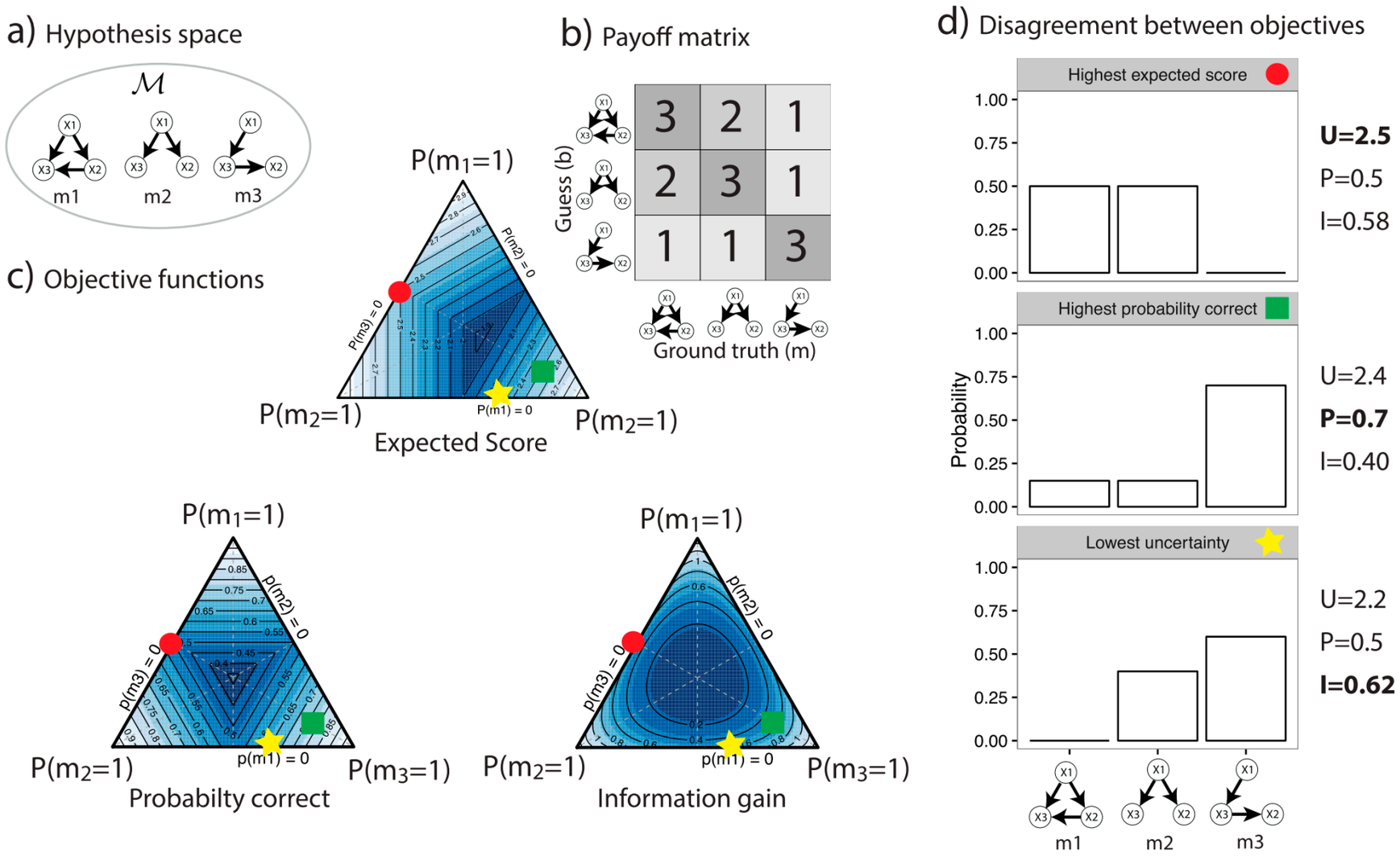
Inferring the World



Neil Bramley. *Constructing the world: Active causal learning in cognition.* 2017

Intervening the World

$$\arg\max_{\mathbf{c}\in\mathcal{C}} \mathop{\mathbb{E}}_{\mathbf{d}'\in\mathcal{D}_{\mathbf{c}}} \left[V(M|\mathbf{d}', D^{t-1}, \mathbf{w}; C^{t-1}, \mathbf{c})\right]$$



Neil Bramley. Constructing the world: Active causal learning in cognition. 2017

Connections to Machine Learning

Causal RL in Transfer Learning

Causal RL in Meta RL

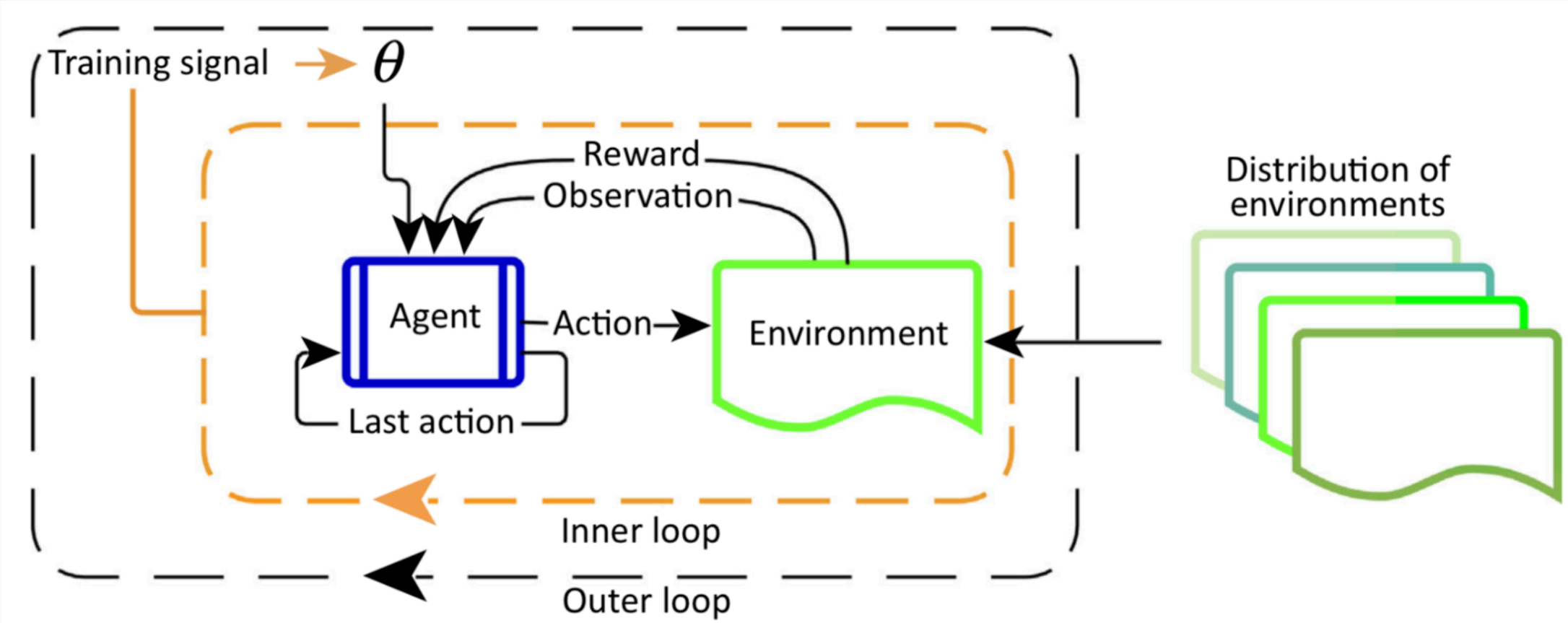


Fig. 2. Illustration of meta-RL, containing two optimization loops. The outer loop samples a new environment in every iteration and adjusts parameters that determine the agent's behavior. In the inner loop, the agent interacts with the environment and optimizes for the maximal reward. (Image source: Botvinick, et al. 2019

Causal RL in Multi-Agent RL

Challenge I: Joint Action Space

concerning result by [Lowe et al., 2017] shows that for a simple setting of binary actions, the probability of taking a gradient step in the correct direction decreases exponentially with the number of agents. Formally

$$Pr\left[\langle \hat{\nabla} J, \nabla J \rangle > 0\right] \propto 0.5^N \tag{26}$$

where the agent's policy is initialzed to an uninformed policy s.t. $\pi(a = 1|s) = 0.5$, N is the number of agents and $\hat{\nabla} J$ is the gradient estimate from a single sample.

Multi-Agent Reinforcement Learning: A Report on Challenges and Approaches.
(Sanyam Kapoor. 2018)

Causal RL in Multi-Agent RL

Challenge II: Common Knowledge of Rationality

Pastine et al. *Introducing Game Theory*, 2017

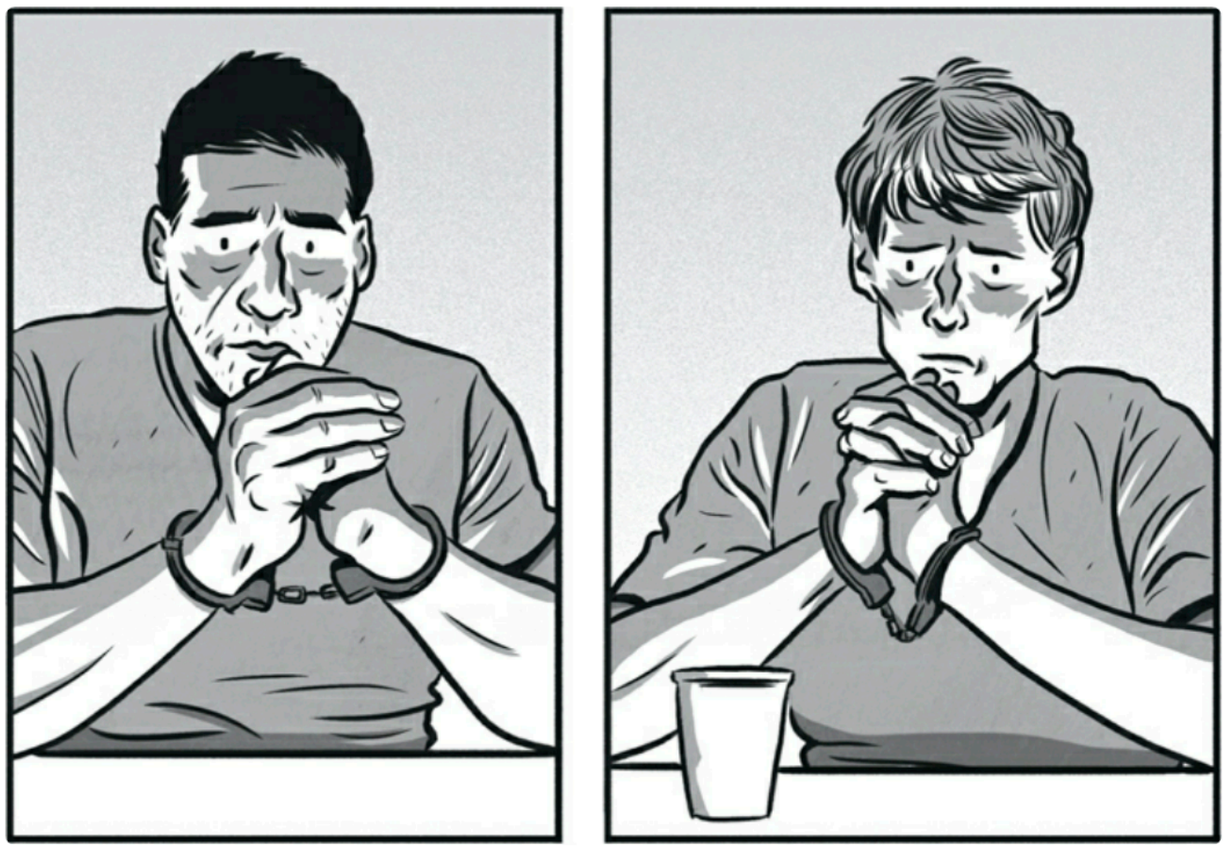
Common knowledge of rationality is a more subtle requirement. Not only do we both have to be rational, but I have to know that you are rational. I also need a second level of knowledge: I have to know that you know that I am rational. I need a third level of knowledge as well: I have to know that you know that I know that you know I am rational. And so on to deeper and deeper levels. Common knowledge of rationality requires that we are able to continue this chain of knowledge indefinitely.

Foerster et al. Counterfactual Multi-Agent Policy Gradients, 2017

Jaques et al. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning, 2019

Causal RL in Multi-Agent RL

Challenge III: Game-Theoretic Effect



	Ben: Silent	Ben: Confess
Silent	A:-1, B:-1	A:-15, B:0
Confess	A:0, B:-15	A:-10, B:-10

Challenge IV: Non-Markovian Nature of Environments

Pastine et al. *Introducing Game Theory*, 2017

Multi-Agent Reinforcement Learning: A Report on Challenges and Approaches.
(Sanyam Kapoor. 2018)

Potential Applications

Computer Vision I

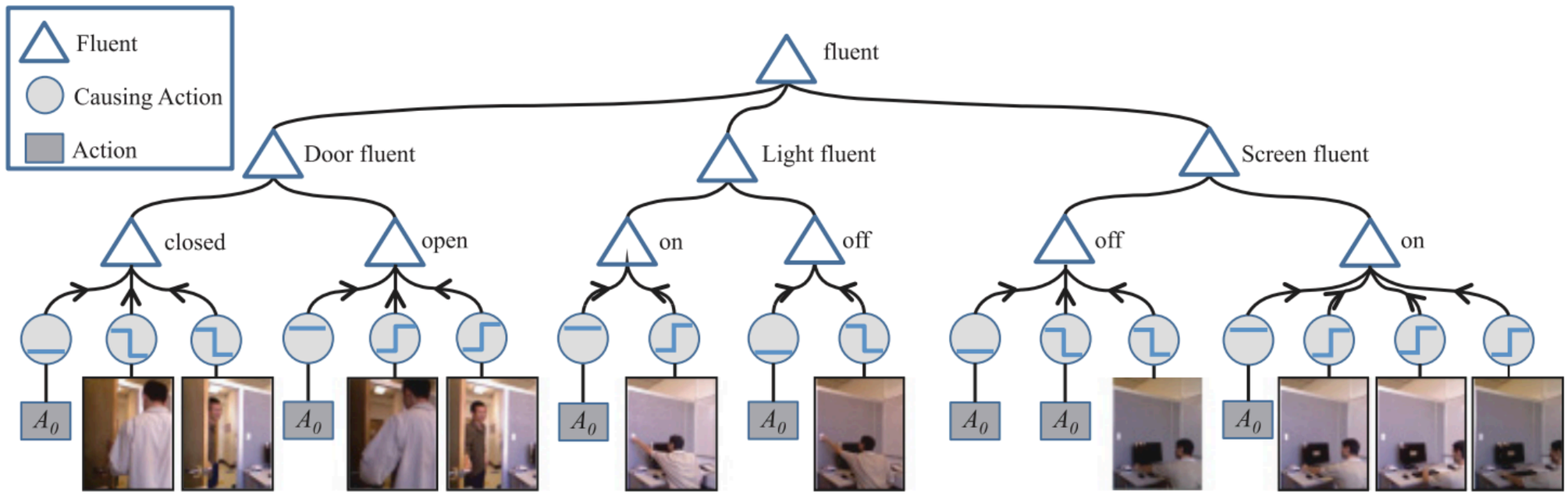


Fig. 8. A Causal And-Or Graph for door status, light status, and screen status. Action A_0 represents nonaction (a lack of state-changing agent action). Nonaction is also used to explain the change of the monitor status to off when the screensaver activates. Arrows point from causes to effects, and undirected lines show deterministic definition.

Learning Perceptual Causality from Video

(Fire et al. 2015)

Computer Vision II

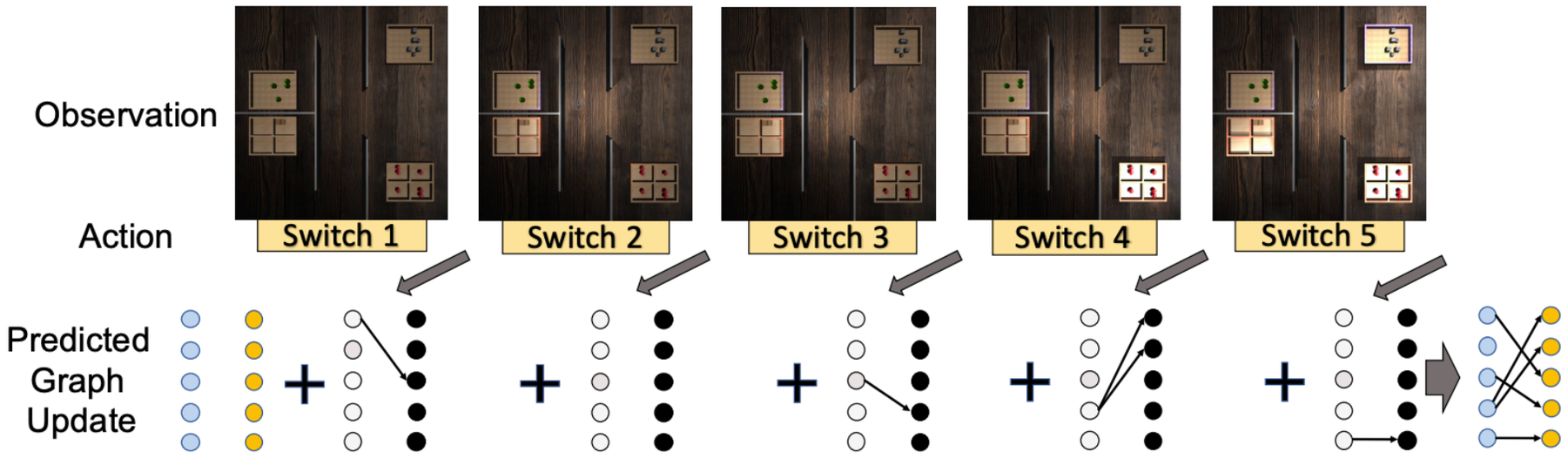


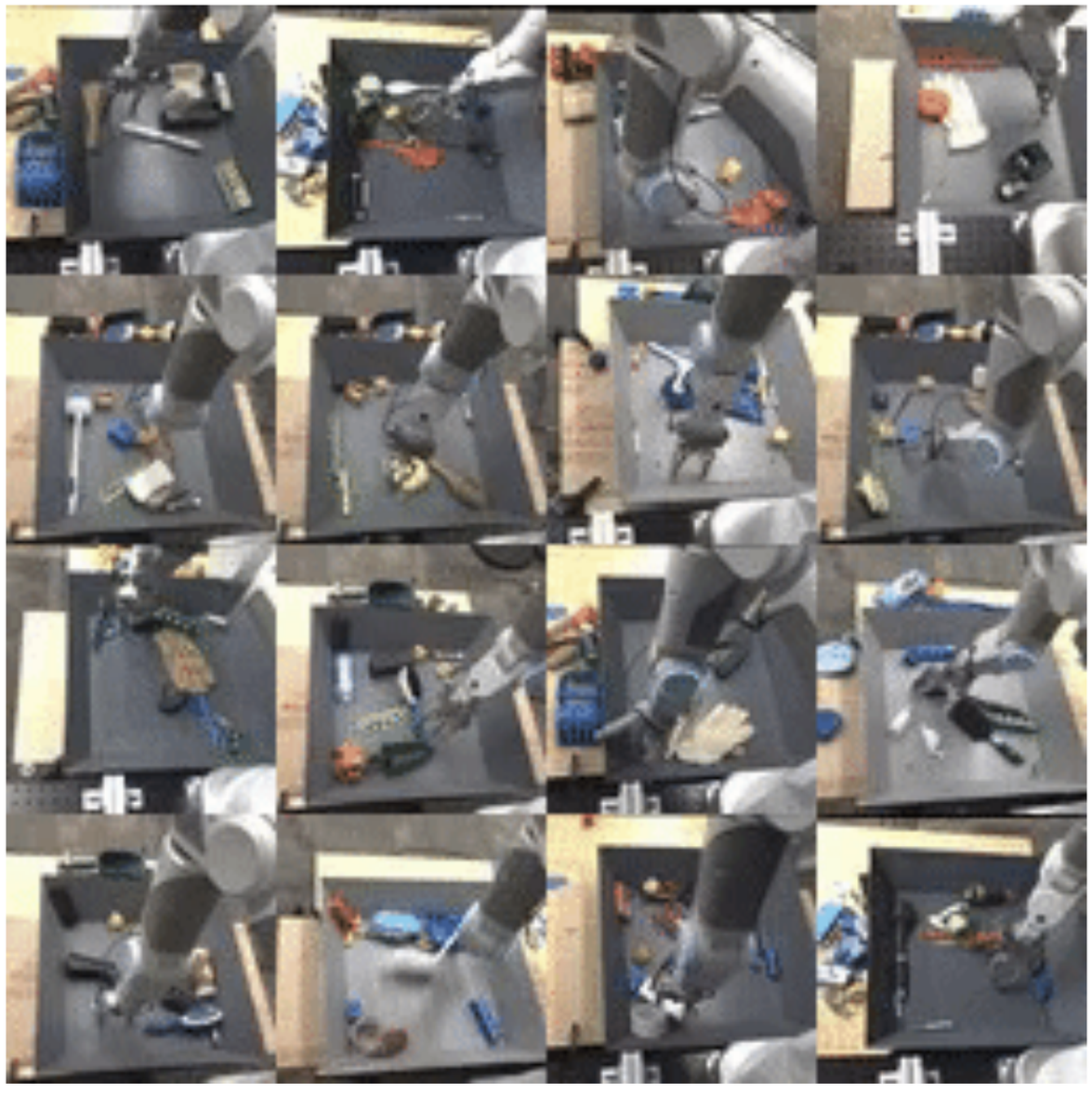
Figure 6: **Sample of Causal Induction**. Here we show an example of our Iterative Causal Induction Model for 5 switches, in the "One-to-Many" case. Given the trajectory of actions and images of the scene, the model needs to reason about which lights were turned on, and how what update this implies in the graph. In this example, the first observed action turns on one of the switches, and the model makes the corresponding update to the graph. The next switch does not change the lighting so the model outputs no update to the graph. The next action sees one light go on, and updates the corresponding switch. The next action turns on two lights, and the graph is updated to reflect this. Lastly, since one light remains unaccounted for, the model knows to add that edge to the graph. Note: The edges and updates are soft updates, but the model learns to predict close to exactly 1 for edges and exactly 0 for non-edges.

Causal Induction from Visual Observations for Goal Directed Tasks

(Nair et al. 2019)

Loving Vincent

Robotics



Video Pixel Networks
(Kalchbrenner et al. 2016)

Self-driving I

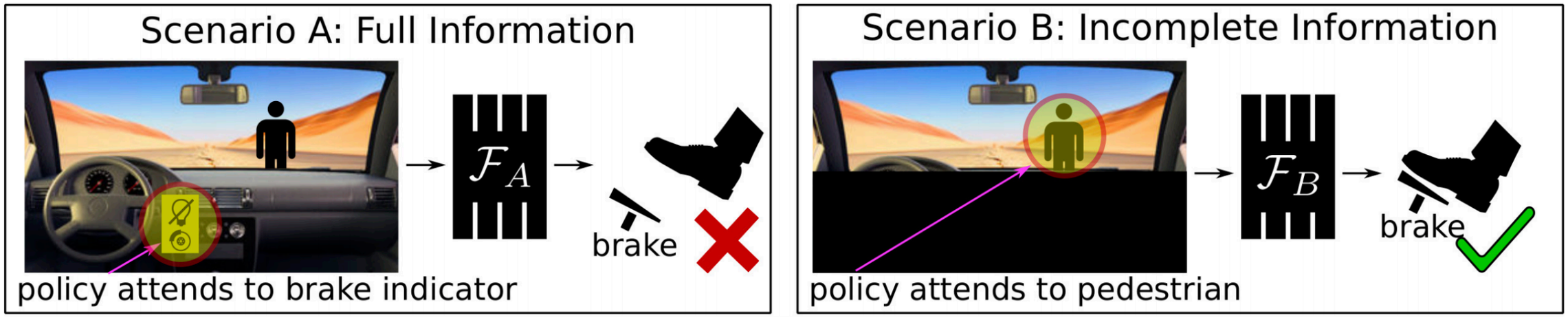


Figure 1: Causal misidentification: *more* information yields worse imitation learning performance. Model A relies on the braking indicator to decide whether to brake. Model B instead correctly attends to the pedestrian.

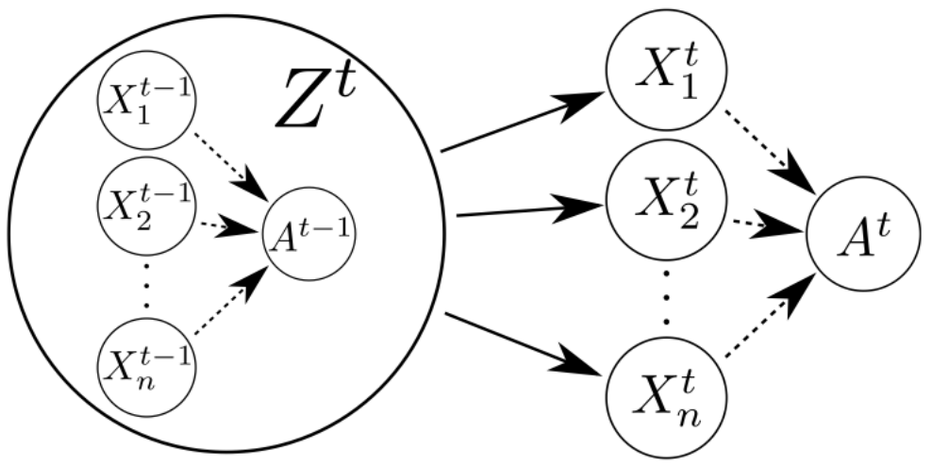


Figure 2: A graph of the underlying causal dynamics of imitation learning. Parents of a node represent its causes. State variables $\{X_i^t\}_{i=1}^n$ are fully observed.

de Hann et al. *Causal Confusion in Imitation Learning*, 2018

Self-driving II

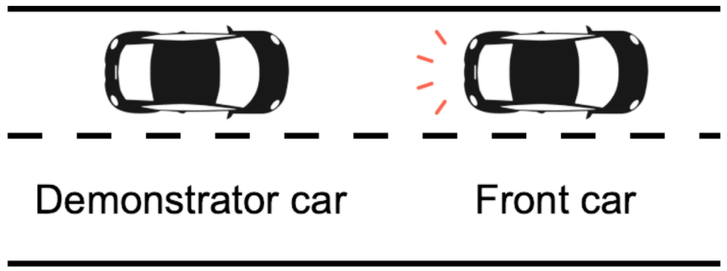
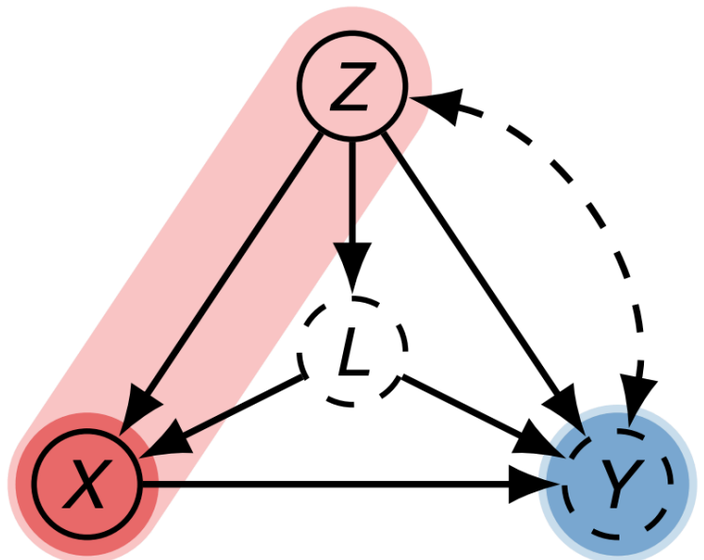


Figure 2: The tail light of the front car is unobserved in highway (aerial) drone data.

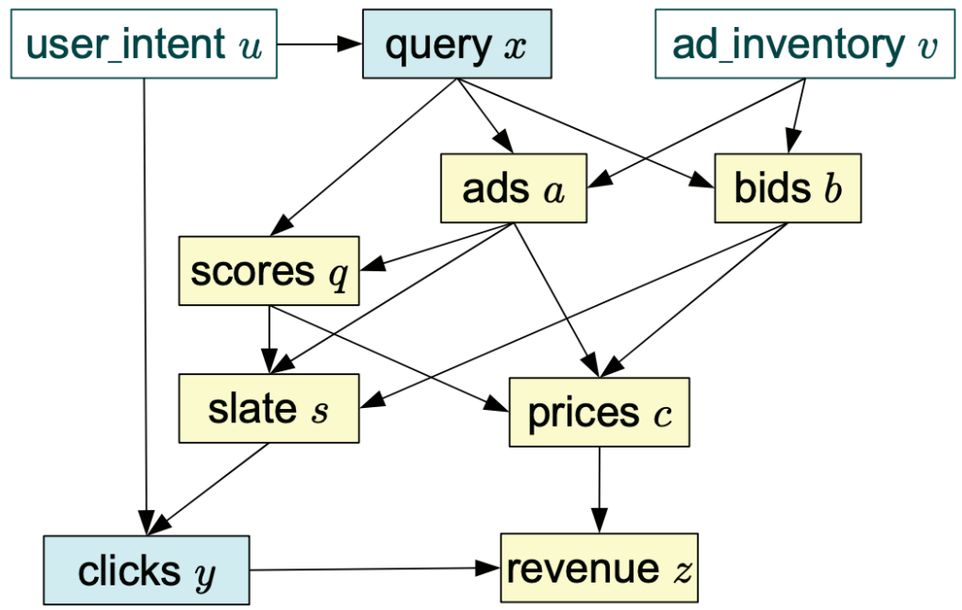


$$\mathbb{E}[Y|\mathbf{do}(\pi)] = \sum_{Y} Y \cdot P(Y|\mathbf{do}(\pi))$$

$$= \sum_{Y} Y \sum_{X,Z} P(Y|\mathbf{do}(x), Z)P(X|Z)P(Z)$$

Zhang et al. *Causal Imitation Learning with Unobserved Confounders*, 2020

Recommendation Systems



$$
\begin{aligned}
x &= f_1(u, \varepsilon_1) && \text{Query context } x \text{ from user intent } u. \\
a &= f_2(x, v, \varepsilon_2) && \text{Eligible ads } (a_i) \text{ from query } x \text{ and inventory } v. \\
b &= f_3(x, v, \varepsilon_3) && \text{Corresponding bids } (b_i). \\
q &= f_4(x, a, \varepsilon_4) && \text{Scores } (q_{i,p}, R_p) \text{ from query } x \text{ and ads } a. \\
s &= f_5(a, q, b, \varepsilon_5) && \text{Ad slate } s \text{ from eligible ads } a, \text{ scores } q \text{ and bids } b. \\
c &= f_6(a, q, b, \varepsilon_6) && \text{Corresponding click prices } c. \\
y &= f_7(s, u, \varepsilon_7) && \text{User clicks } y \text{ from ad slate } s \text{ and user intent } u. \\
z &= f_8(y, c, \varepsilon_8) && \text{Revenue } z \text{ from clicks } y \text{ and prices } c.
\end{aligned}
$$

Bottou et al. *Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising*, 2013

Healthcare/Medicine

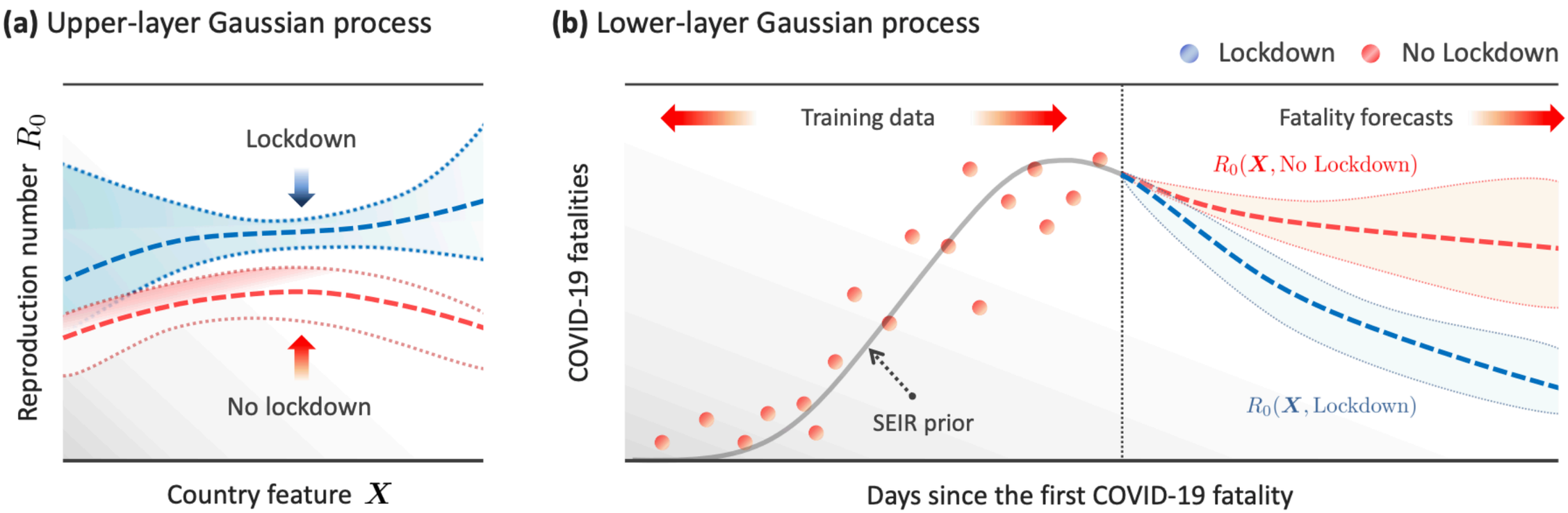


Figure 1: **Pictorial illustration of compartmental Gaussian processes.** (a) The upper-layer GP f_U maps country features and lockdown policies to a predicted R_0. Here we depict a simplified binary policy indicator (lockdown or no lockdown). (b) The lower-layer GP f_L maps time to number of COVID-19 fatalities. The mean function is an SEIR model modulated by the upper-layer GP. Projections are obtained using the GP posteriors.

When and How to Lift the Lockdown?
Global COVID-19 Scenario Analysis and Policy Assessment
using Compartmental Gaussian Processes
(Qian et al. 2020)

Conclusion & Discussion

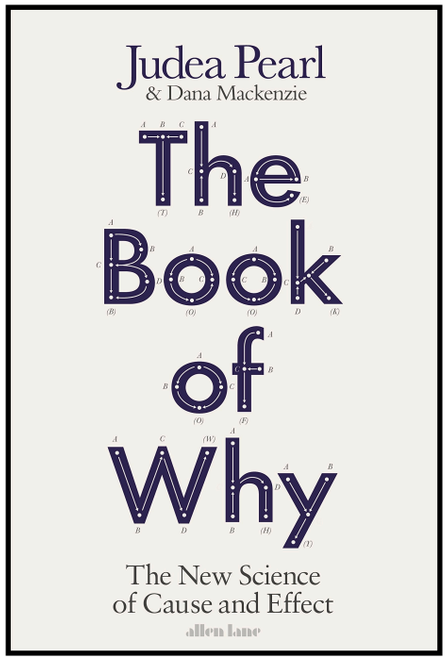
The Take-home Message

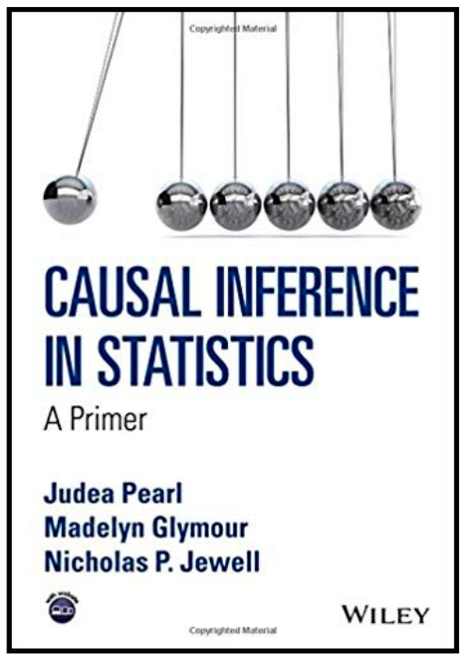
Causal RL

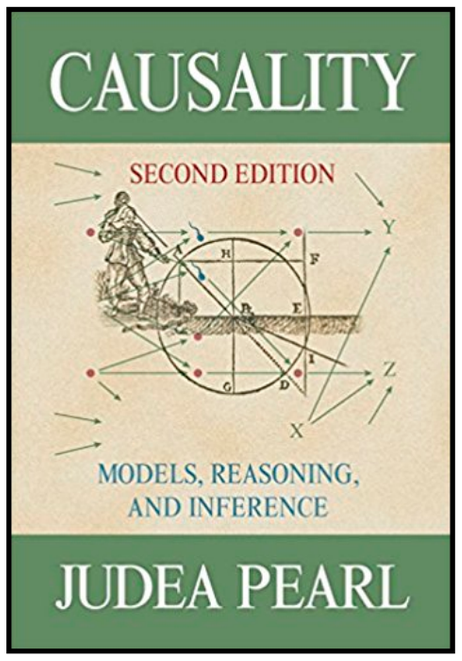
was born for

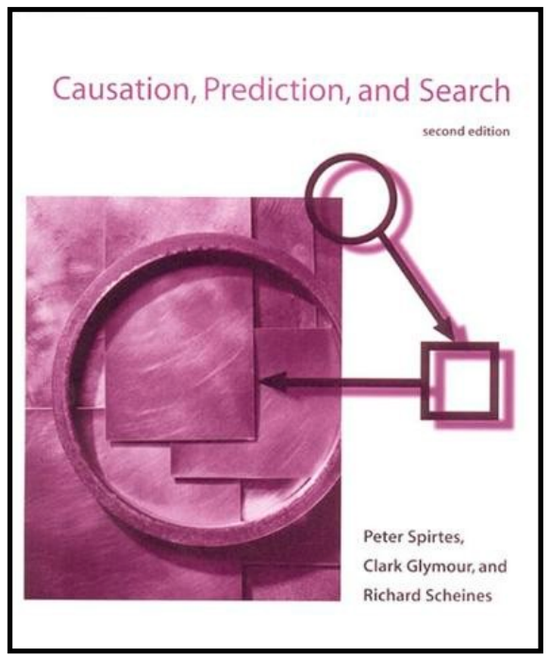
AGI.

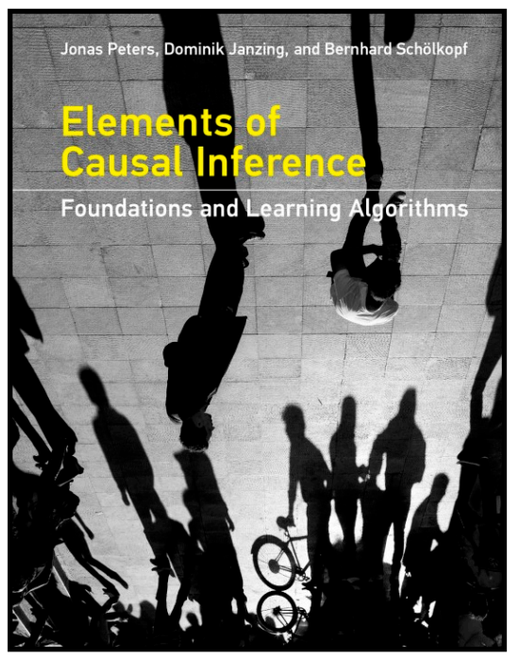
Recommendation

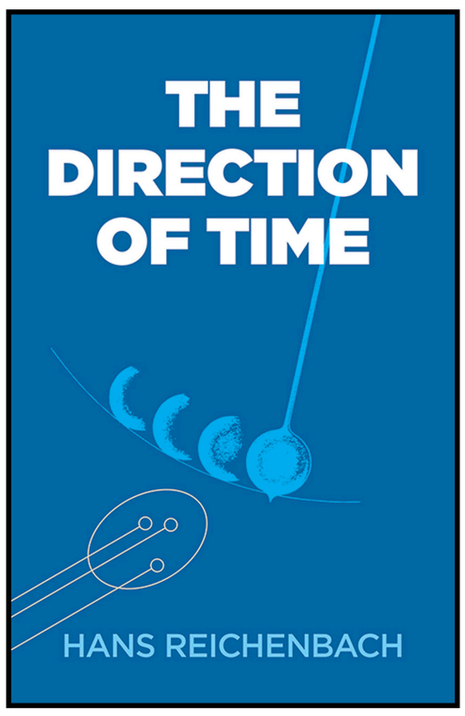


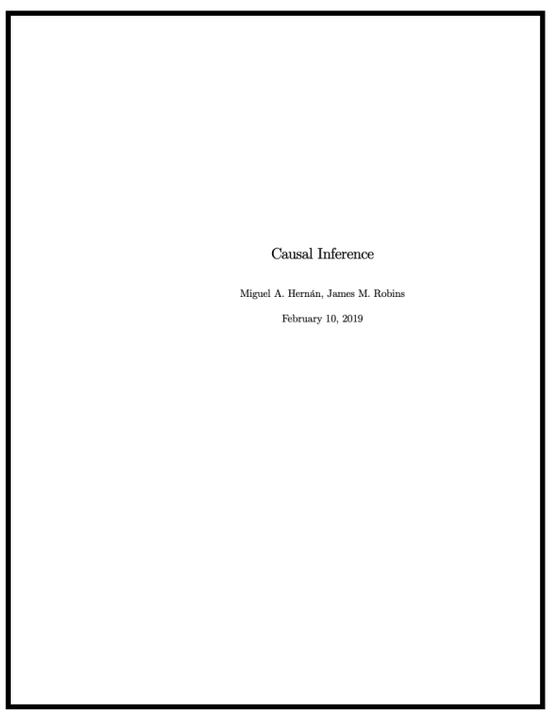


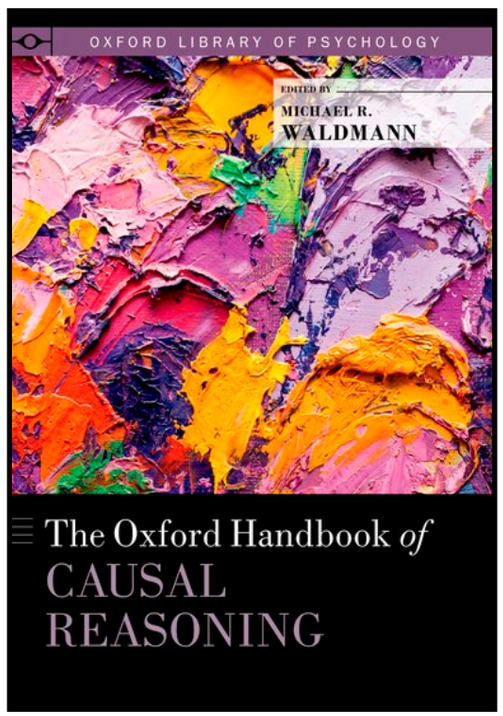


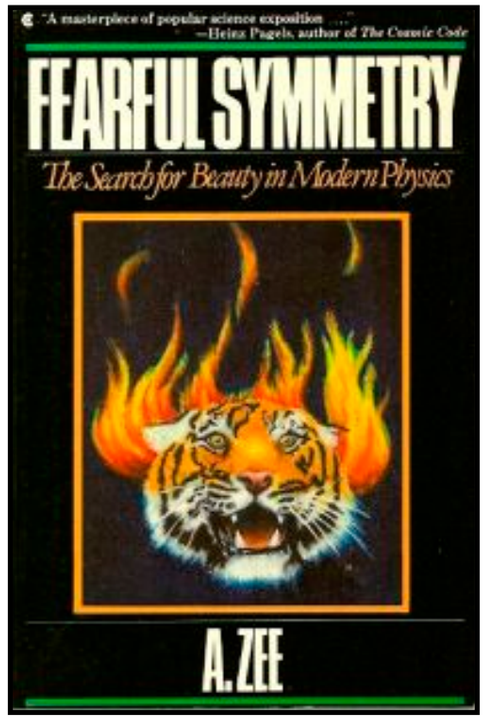












Causality for Machine Learning

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany
bs@tuebingen.mpg.de

Abstract

Graphical causal inference as pioneered by Judea Pearl arose from research on artificial intelligence (AI), and for a long time had little connection to the field of machine learning. This article discusses where links have been and should be established, introducing key concepts along the way. It argues that the hard open problems of machine learning and AI are intrinsically related to causality, and explains how the field is beginning to understand them.

arXiv:1911.10500

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

http://incompleteideas.net/IncIdeas/BitterLesson.html

“Panel”

“The essence of intelligence is while only being able to observe a world of things, try to come up with a world of ideas.” — *Vladimir Vapnik*

“Life/evolution is a process of compression.”
—Jürgen Schmidhuber

“A low-dimensional thought or conscious state is analogous to a sentence: it involves only a few variables and yet can make a statement with very high probability of being true.”
—Yoshua Bengio

ISSN 0005-1179, Automation and Remote Control, 2019, Vol. 80, No. 11, pp. 1949–1975. © *Pleiades Publishing, Ltd., 2019.*
Russian Text © *The Author(s), 2019, published in Avtomatika i Telemekhanika, 2019, No. 11, pp. 24–58.*

TOPICAL ISSUE

Complete Statistical Theory of Learning

V. N. Vapnik

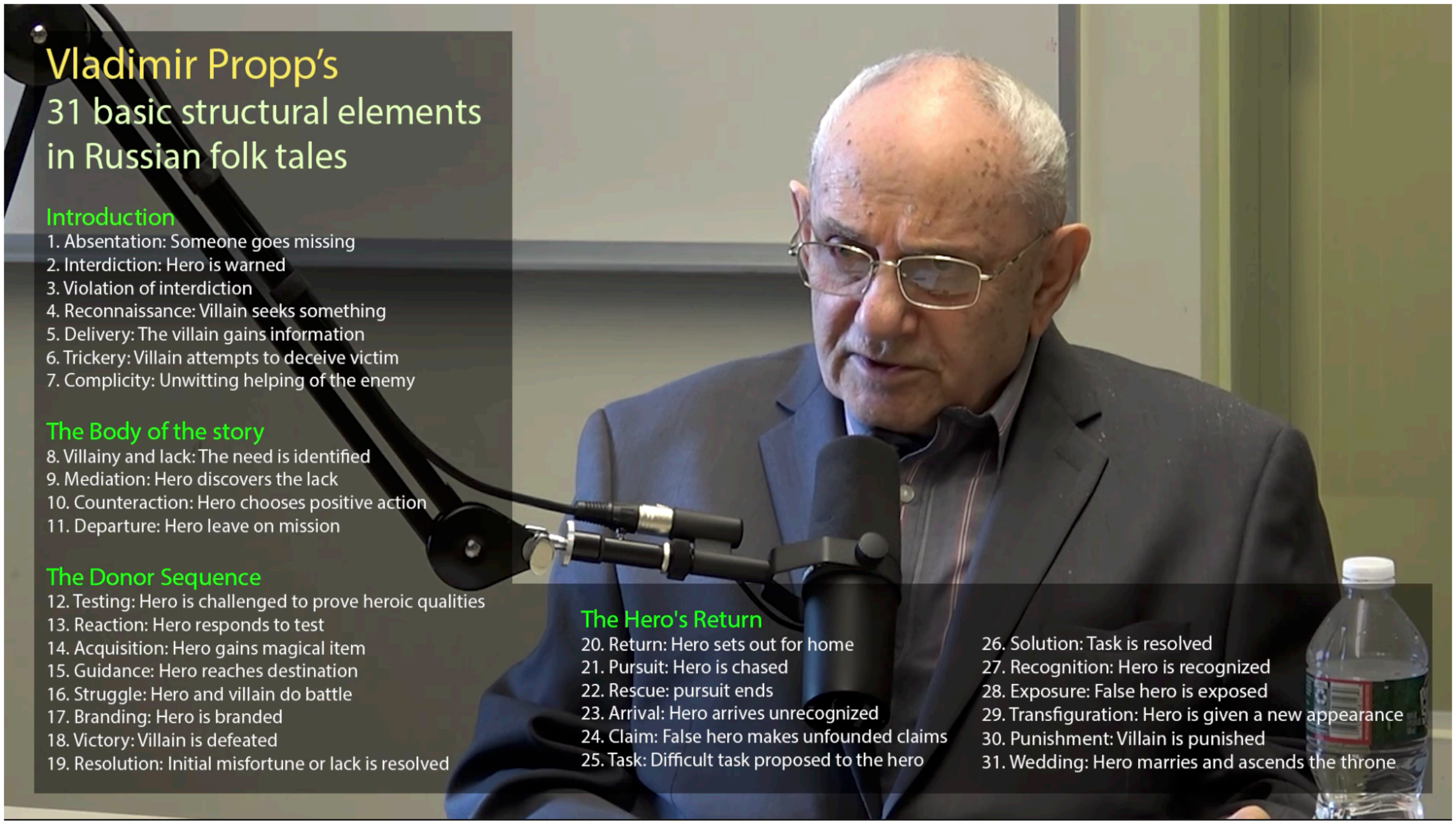
Columbia University, New York, USA
e-mail: vladimir.vapnik@gmail.com

Received July 13, 2018
Revised September 5, 2018
Accepted November 8, 2018

In the memory of outstanding scientist
and remarkable person Ja.Z. Tsypkin

Mathematical Problem versus Intelligence Problem

Vladimir Propp's Predicates



https://youtu.be/bQa7hpUpMzM

Vapnik's Challenge

Using 60,000 training examples of MNIST digit recognition problem (6,000 per/class) DNN achieved $\approx 0.5\%$ test error.

1. **Find predicates which will allow you to achieve the same level of test error using just 600 examples (60/per class).**
2. **Find a small set of basic predicates to achieve this goal.**

THE UNREASONABLE EFFECTIVENSS OF MATHEMATICS IN THE NATURAL SCIENCES

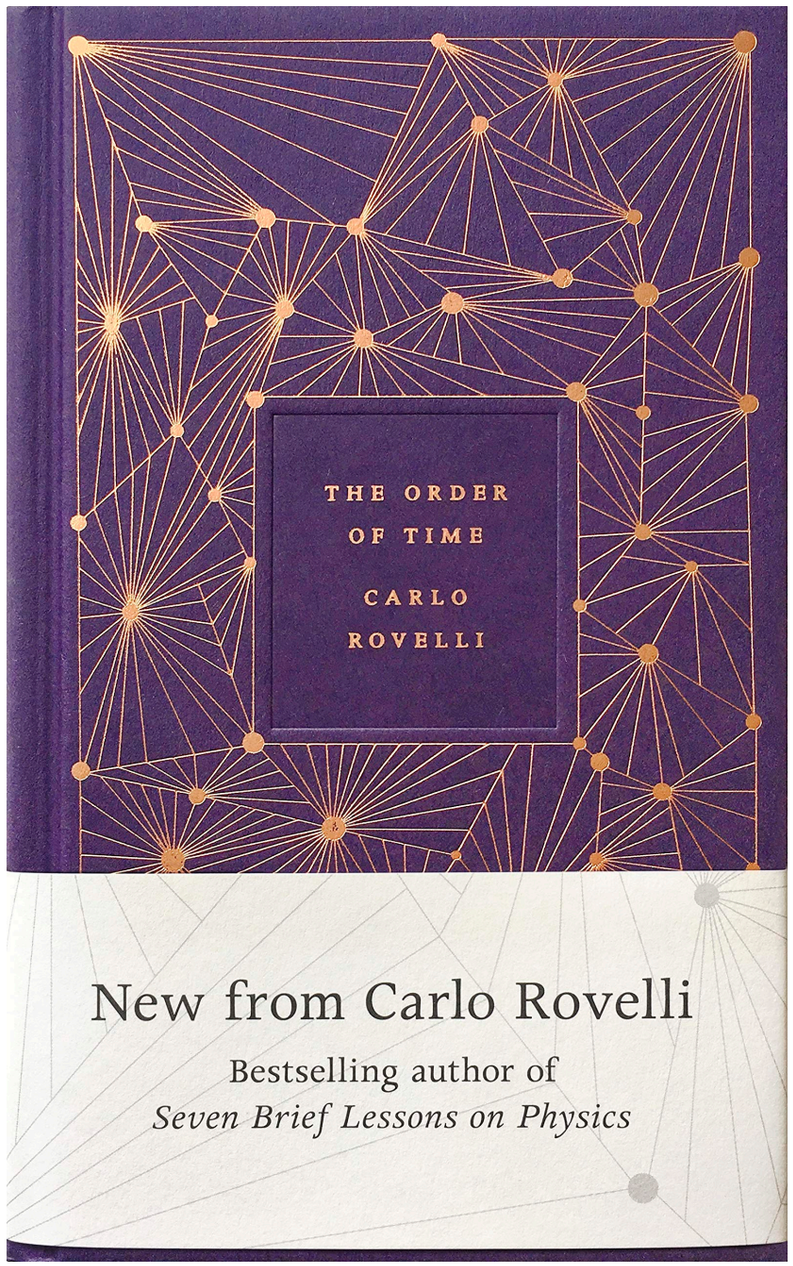
Eugene Wigner

Mathematics, rightly viewed, possesses not only truth, but supreme beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show. The true spirit of delight, the exaltation, the sense of being more than Man, which is the touchstone of the highest excellence, is to be found in mathematics as surely as in poetry.

- BERTRAND RUSSELL, Study of Mathematics

Communications in Pure and Applied Mathematics, Vol. 13, No. I (February 1960)

“Nature” vs “Nurture”



“Does time has an order?”

Behind The Scene

The
Design of Experiments

By

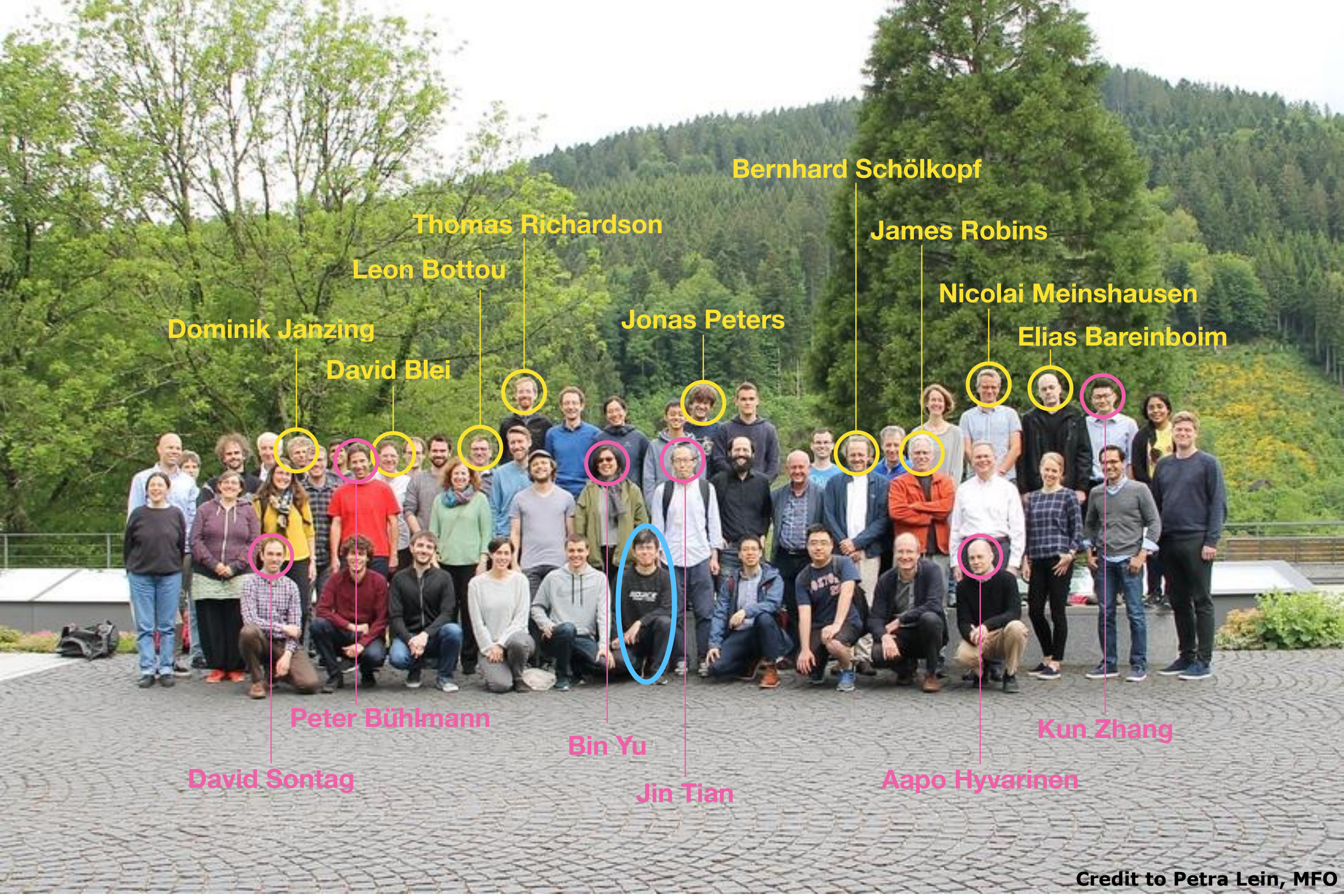
R. A. Fisher, Sc.D., F.R.S.

Formerly Fellow of Gonville and Caius College, Cambridge
Honorary Member, American Statistical Association
and American Academy of Arts and Sciences
Galton Professor, University of London

Oliver and Boyd
Edinburgh: Tweeddale Court
London: 33 Paternoster Row, E.C.
1935

Foundations and New Horizons for Causal Inference

Mathematisches Forschungsinstitut Oberwolfach, 26 May - 1 Jun, 2019



Foundations and New Horizons for Causal Inference

Mathematisches Forschungsinstitut Oberwolfach, 26 May - 1 Jun, 2019

Thank You.

Q&A

集智学园 | 因果科学与CausalAI读书会 | 29 Nov 2020