Towards Causal Reinforcement Learning (CRL)

Elias Bareinboim

Causal Artificial Intelligence Lab

Columbia University

(@eliasbareinboim)

Slides: https://crl.causalai.net

ICML, 2020

JOINT WORK WITH CAUSAL AI LAB & COLLABORATORS

Yotam Alexander (Columbia)

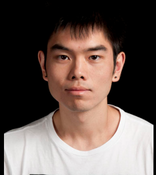
Juan Correa (Columbia)

Kai-Zhan Lee (Columbia)

Sanghack Lee (Columbia)

Adele Ribeiro (Columbia)

Kevin Xia (Columbia)



Junzhe Zhang (Columbia)

Amin Jaber (Purdue)

Chris Jeong (Purdue)

Yonghan Jung (Purdue)

Daniel Kumor (Purdue)

Judea Pearl (UCLA)

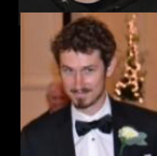
Carlos Cinelli (UCLA)

Andrew Forney (UCLA)

Brian Chen (Brex)

Jin Tian (Iowa State)

Duligur Ibeling (Stanford)



Thomas Icard (Stanford)

Murat Kocaoglu (IBM)

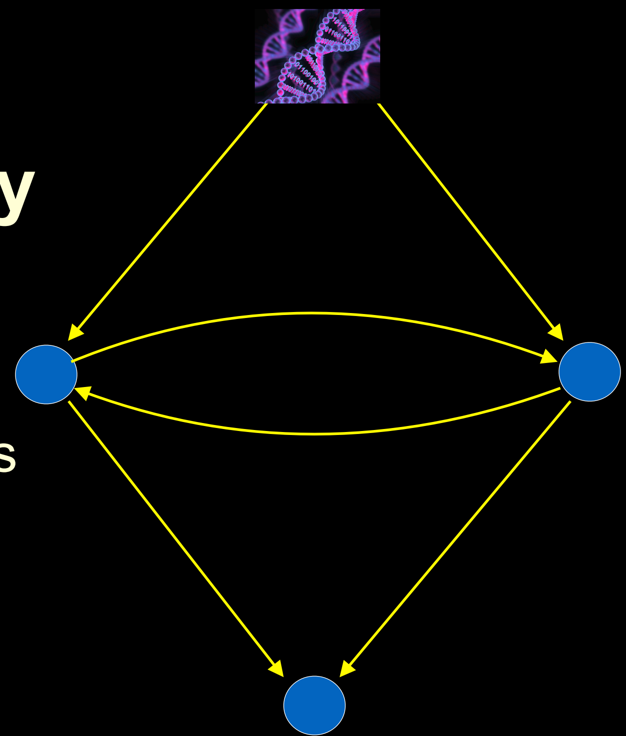
Karthikeyan Shanmugam (IBM)

Jiji Zhang (Lingnan University)

Paul Hünermund (Copenhagen)

CausalAI Lab

Structural Causal Models



1. Explainability

(Effect identification and decomposition, Bias Analysis and Fairness, Robustness and Generalizability)

2. Decision-Making

(Reinforcement Learning, Randomized Controlled Trials, Personalized Decision-Making)

3. Applications, Education, Software

Data Science:
Principled ("scientific") inferences from large data collections.

AI-ML:
Principles and tools for designing robust and adaptable learning systems.

What is Causal RL?

- Reinforcement Learning (RL) is awesome at handling sample complexity and credit assignment.
- Causal Inference (CI) is great at leveraging structural invariances across settings and conditions.
- Can we have the best of both worlds? Yes!

Simple solution: $\boxed{\text{Causal RL = CI + RL}}$

Our goal: Provide a cohesive framework that takes advantage of the capabilities of both formalisms (from first principles), and that allows us to develop the next generation of AI systems.

Outline

- Part 1. Foundations of CRL (60’)
  - Intro to Structural Causal Models, Pearl Causal Hierarchy (PCH), Causal Hierarchy Theorem (CHT)
  - Current RL & CI methods through CRL Lens
- Part 2. New Challenges and Opportunities of Causal Reinforcement Learning (60’)

Goal: Introduce the main ideas, principles, and tasks.

Not focused on the implementation details.

For a more detailed discussion, see: NeurIPS’15, PNAS’16, ICML’17, IJCAI’17, NeurIPS-18, AAAI-19, UAI-19, NeurIPS-19, ICML-20 … + new CRL survey.

Resources: https://crl.causalai.net

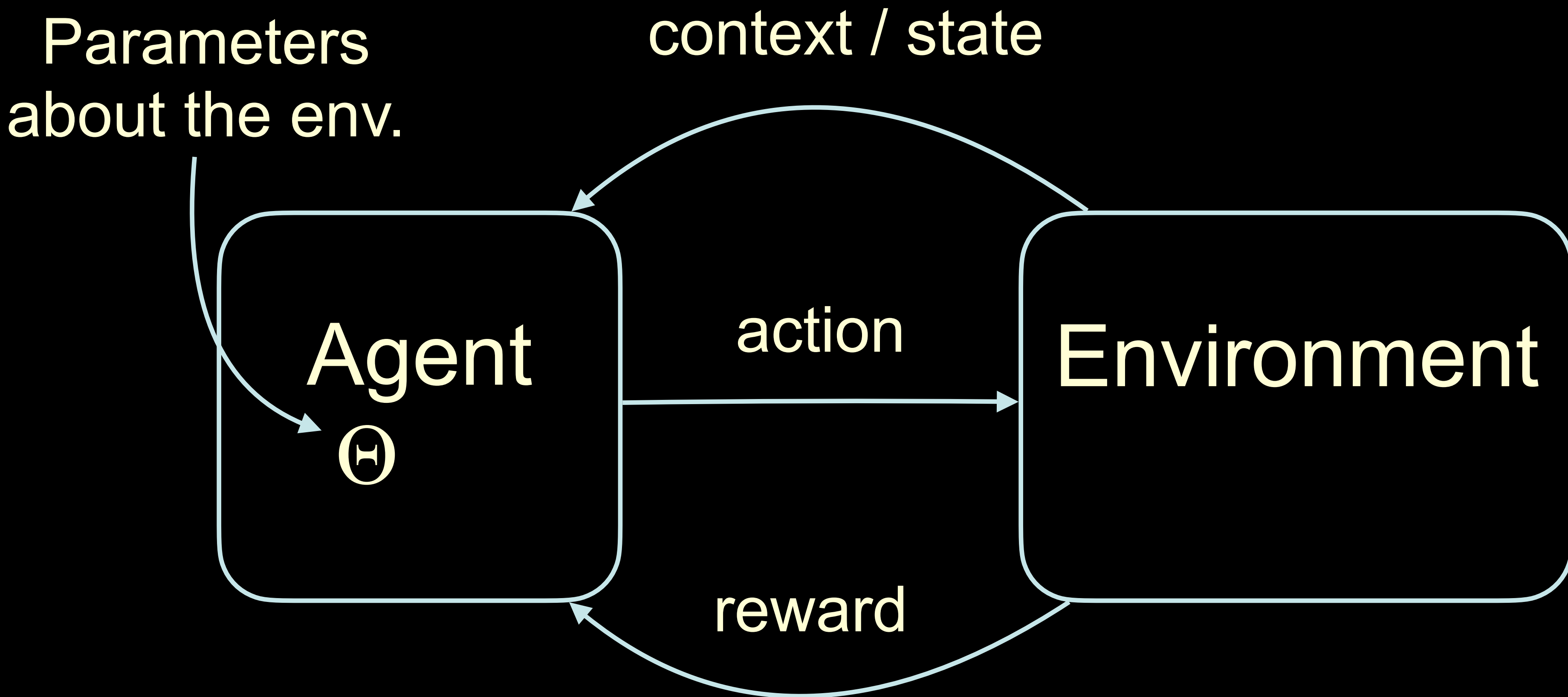
PRELUDE:
REINFORCEMENT LEARNING

What’s Reinforcement Learning?

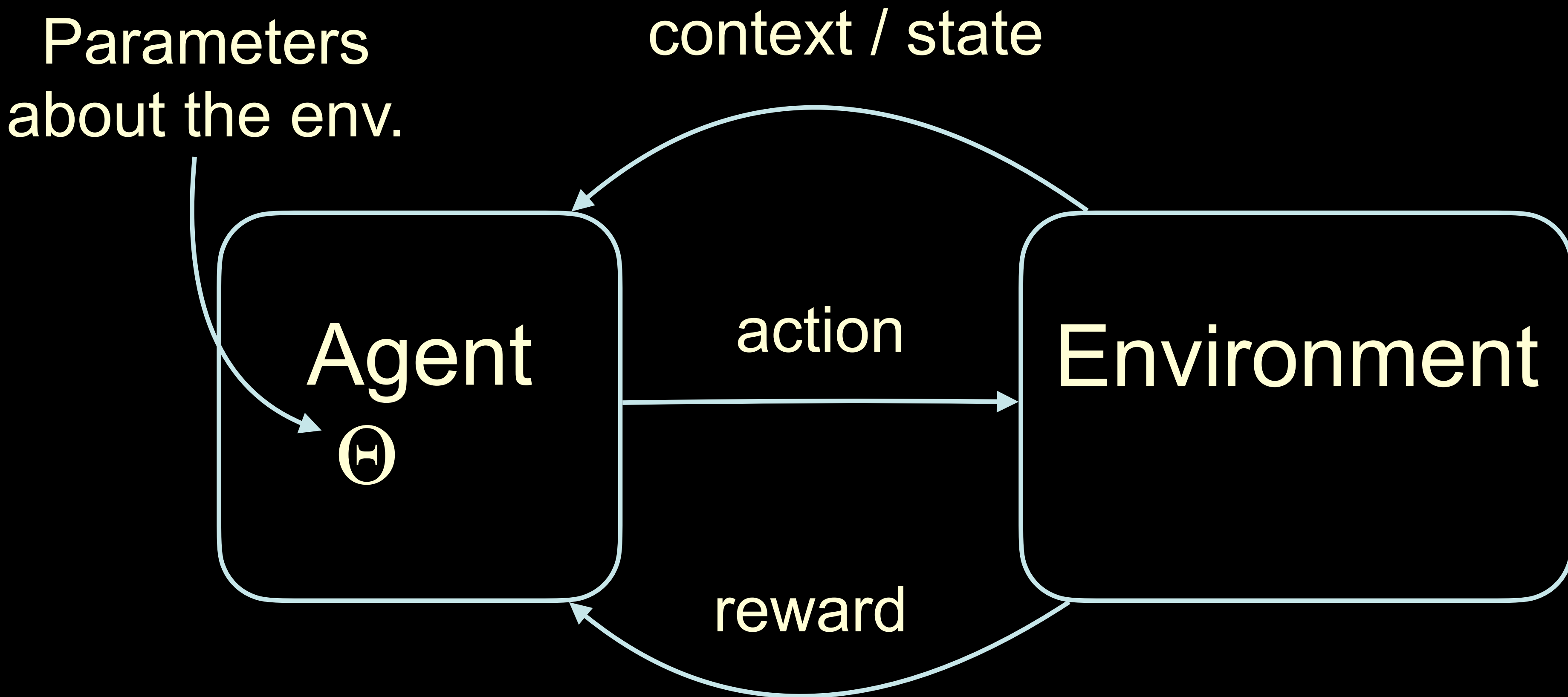
- Goal-oriented learning -- how to maximize a numerical reward signal.
- Learning about, from, and while interacting with an external environment.
- Adaptive learning -- each action is tailored for the evolving covariates and actions’ history.

(Learning without having a full specification of the system; versus planning/programming)

RL - Big Picture

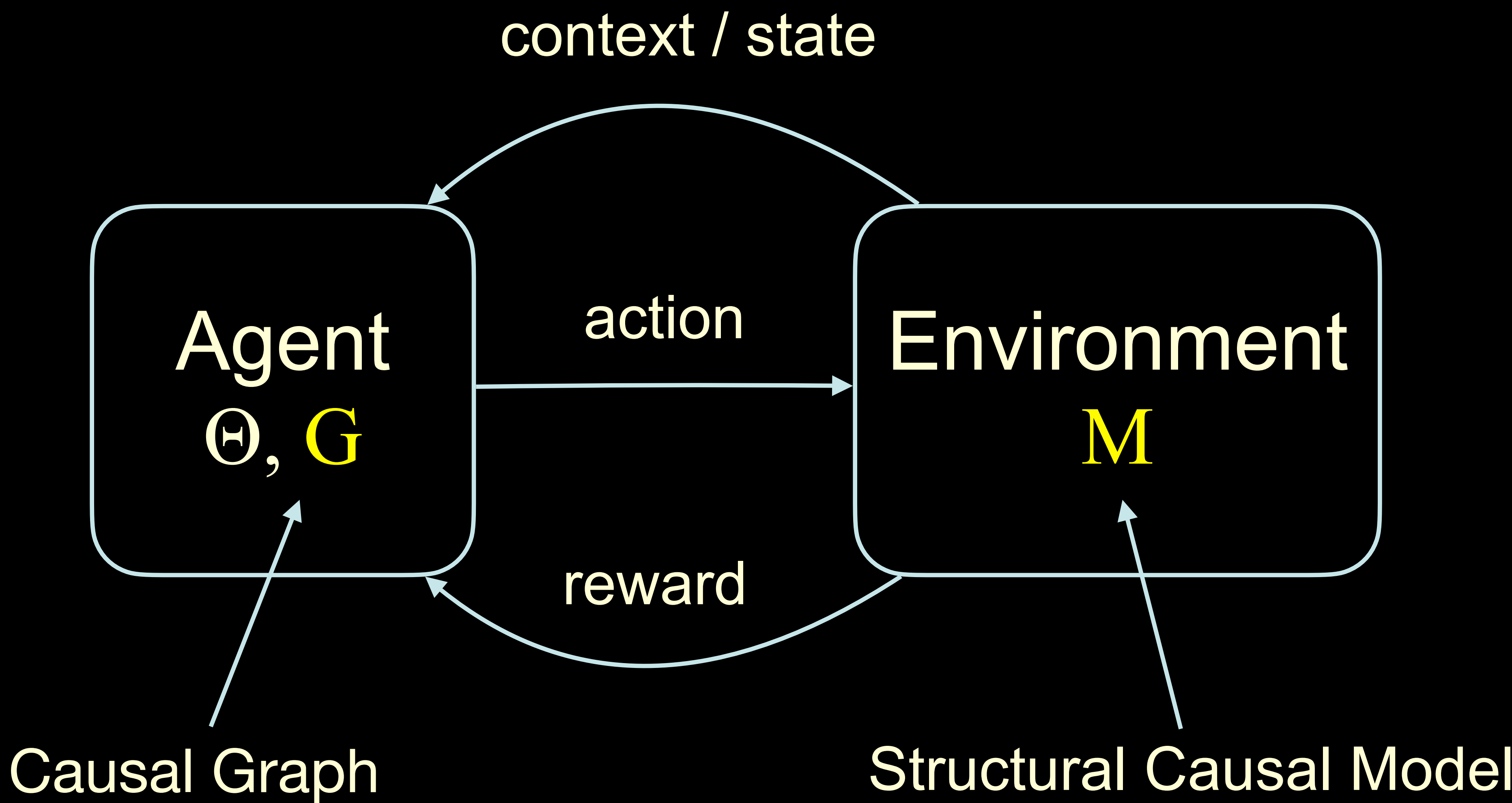


RL - Big Picture

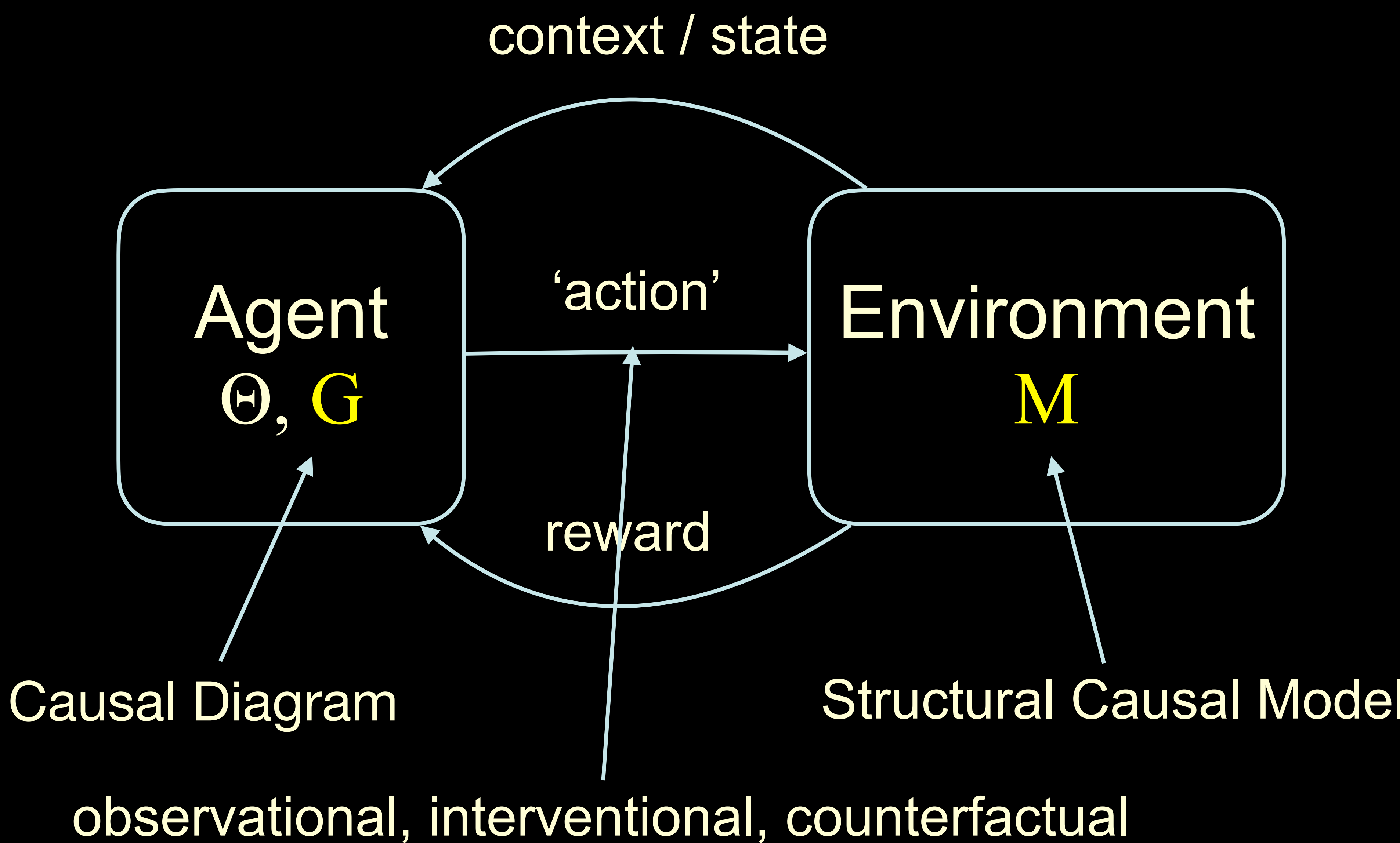


- Receive feedback in the form of rewards.
- Agent's utility is defined by the reward function.
- Must (learn to) act so as to maximize expected rewards.

Causal RL - Big Picture

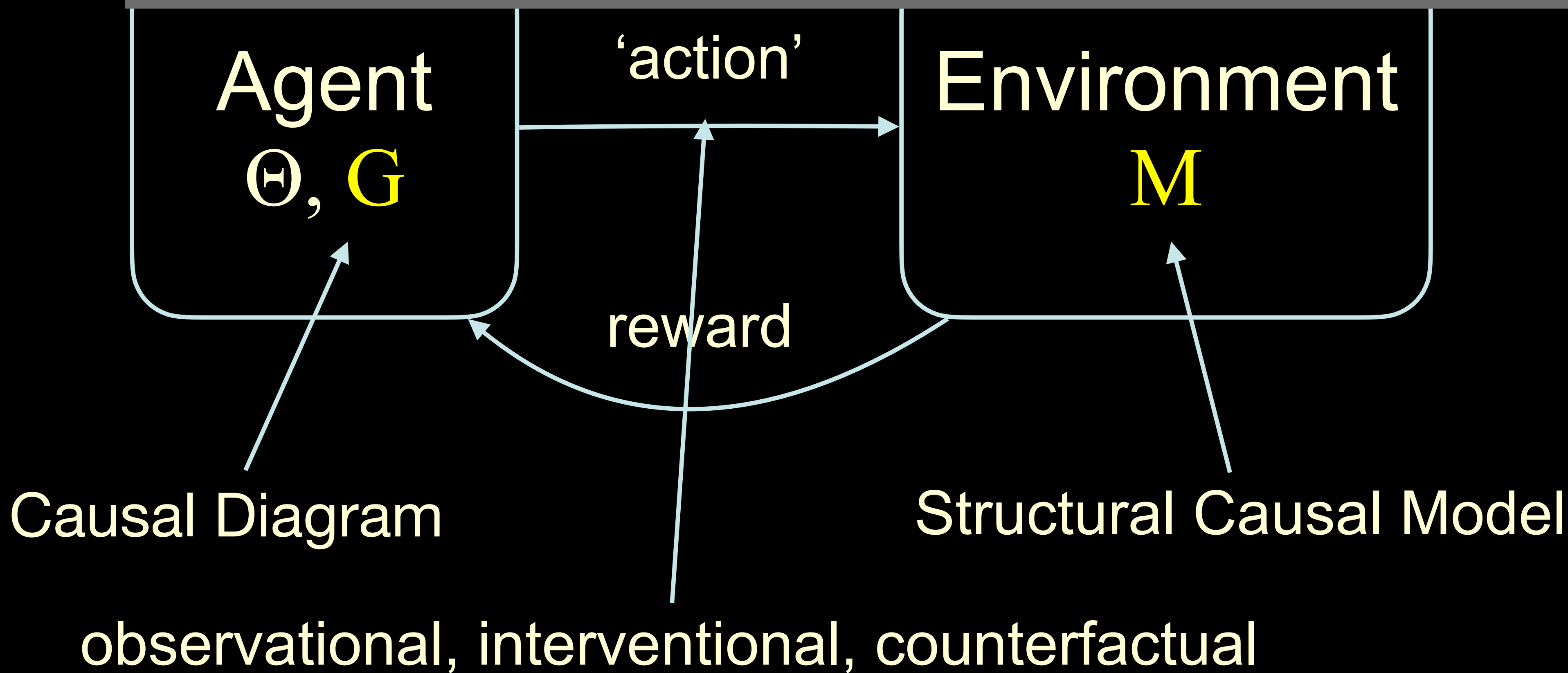


Causal RL - Big Picture



Two key observations (RL → CRL):

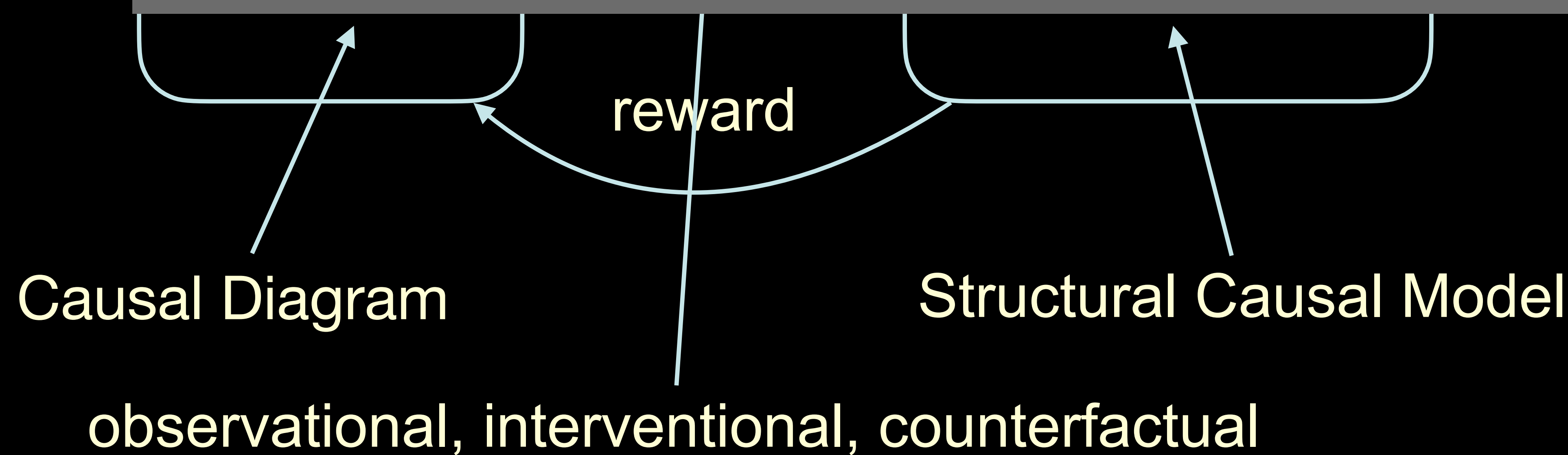
1. The environment and the agent will be tied thr. the pair SCM M & causal graph G.
2. We'll define different types of "actions", or interactions, to avoid ambiguity (PCH).



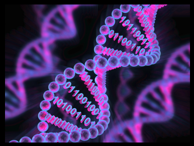
Two key observations (RL → CRL):

1. **The environment and the agent will be tied thr. the pair SCM M & causal graph G.**
2. **We'll define different types of "actions", or interactions, to avoid ambiguity (PCH).**

Let's define and understand
(1) the pair <M, G>, and (2) the PCH.



STRUCTURAL CAUSAL MODELS & CAUSAL GRAPHS

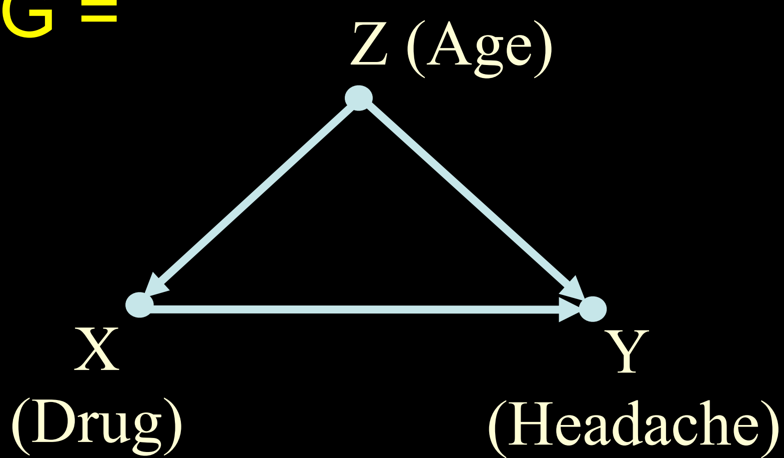


SCM -- REPRESENTING THE DATA GENERATING MODEL

- Processes

$$\text{Drug} \leftarrow f_D\,(\text{Age}, U_D)$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

G =



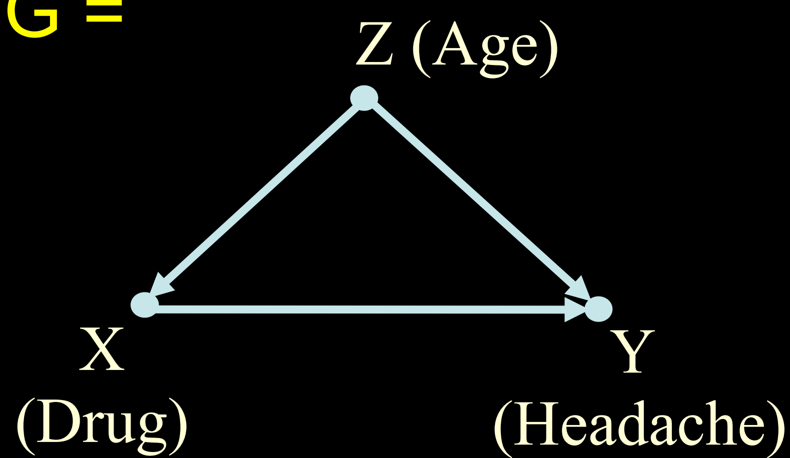
P(Z, X, Y)
(observational)

SCM -- REPRESENTING THE DATA GENERATING MODEL

- Processes

$$\text{Drug} \leftarrow f_D\,(\text{Age}, U_D)$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

G =



P(Z, X, Y)
(observational)

- Intervention

$$\text{Drug} \leftarrow \text{Yes}$$

SCM -- REPRESENTING THE DATA GENERATING MODEL

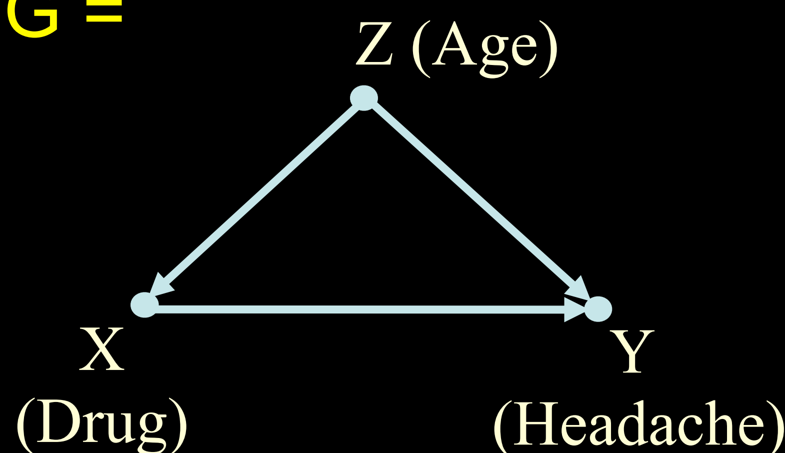
- Processes

$$\text{Drug} \leftarrow f_D(\text{Age}, U_D)$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

- Intervention

$$\text{Drug} \leftarrow \text{Yes}$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

G =



P(Z, X, Y)
(observational)

SCM -- REPRESENTING THE DATA GENERATING MODEL

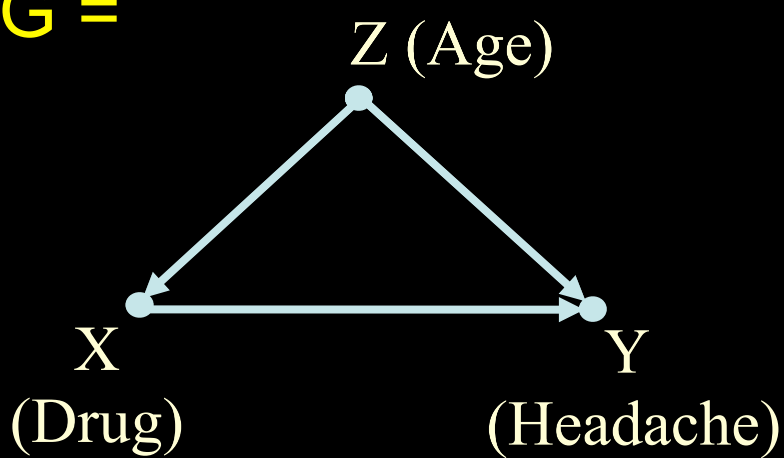
- Processes

$$\text{Drug} \leftarrow f_D(\text{Age}, U_D)$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

- Intervention

$$\text{Drug} \leftarrow \text{rand}()$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

G =



P(Z, X, Y)
(observational)

SCM -- REPRESENTING THE DATA GENERATING MODEL

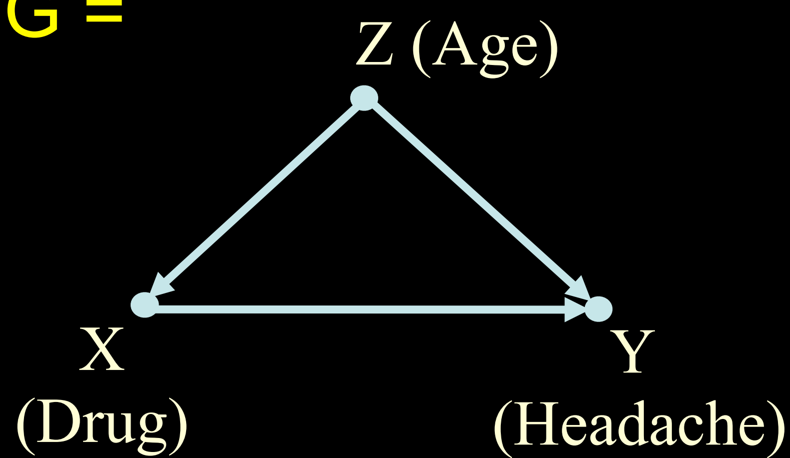
- Processes

$$\text{Drug} \leftarrow f_D(\text{Age}, U_D)$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

- Intervention

$$\text{Drug} \leftarrow \Pi(\text{Age})$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

G =



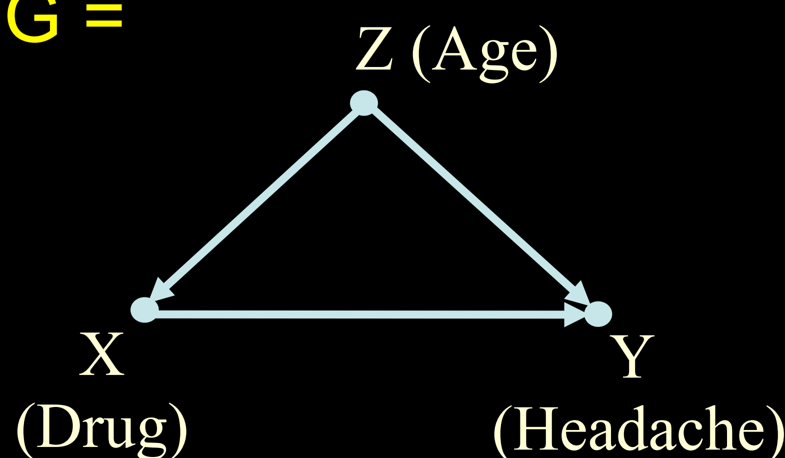
P(Z, X, Y)
(observational)

SCM -- REPRESENTING THE DATA GENERATING MODEL

- Processes

$$\text{Drug} \leftarrow f_D(\text{Age}, U_D)$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

G =



P(Z, X, Y)
(observational)

- Intervention

$$\text{Drug} \leftarrow \Pi(\text{Age})$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

σ-calculus (Correa & Bareinboim 2020)

SCM -- REPRESENTING THE DATA GENERATING MODEL

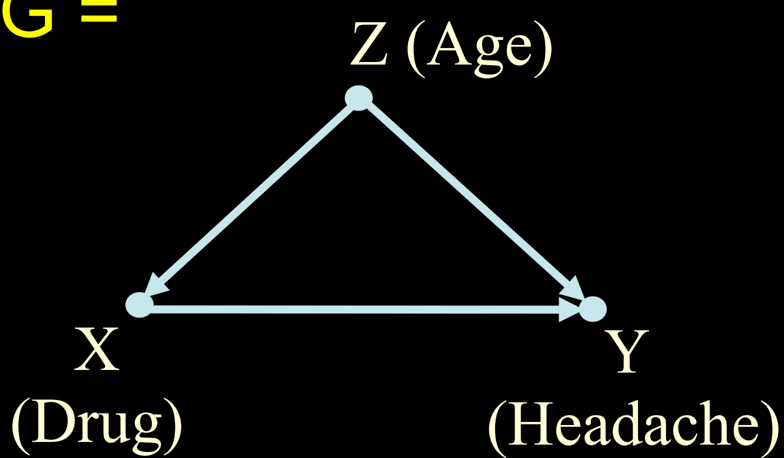
- Processes

$$\text{Drug} \leftarrow f_D\,(\text{Age}, U_D)$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

- Intervention

$$\text{Drug} \leftarrow \Pi(\text{Age})$$
$$\text{Headache} \leftarrow f_H\,(\text{Drug}, \text{Age}, U_H)$$

G =



P(Z, X, Y)
(observational)

SCM -- REPRESENTING THE DATA GENERATING MODEL

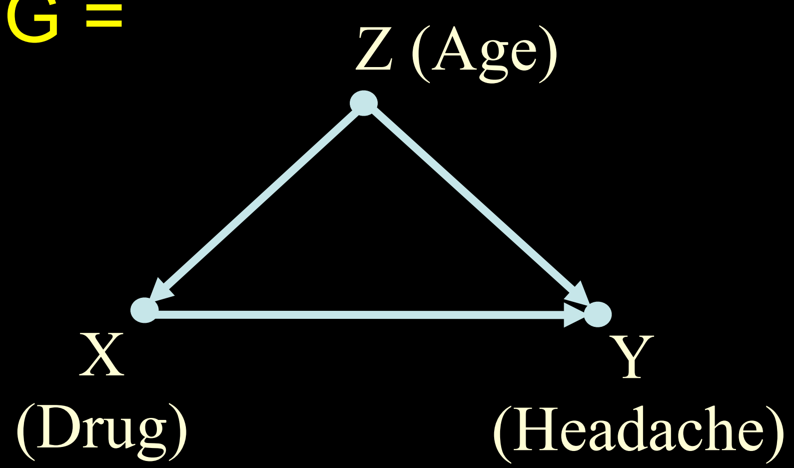
- Processes

$$\text{Drug} \leftarrow f_D(\text{Age}, U_D)$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

- Intervention

$$\text{Drug} \leftarrow \text{Yes}$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

G =



P(Z, X, Y)
(observational)

SCM -- REPRESENTING THE DATA GENERATING MODEL

- Processes

Drug $\leftarrow f_D$ (Age, U_D)

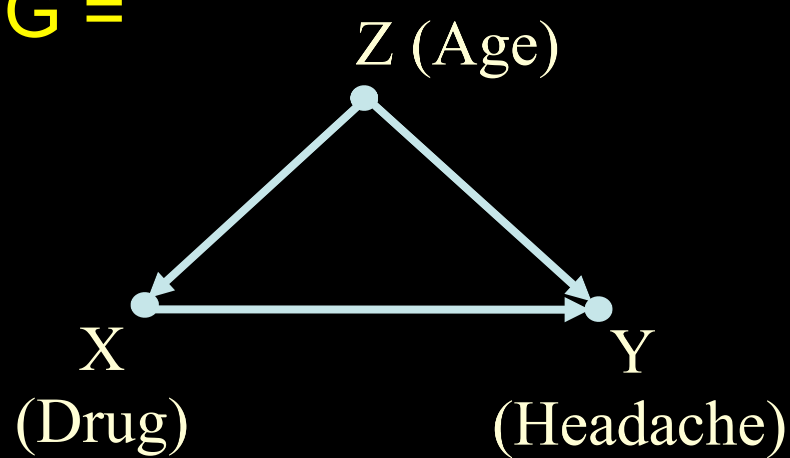
Headache $\leftarrow f_H$(Drug, Age, U_H)

- Intervention

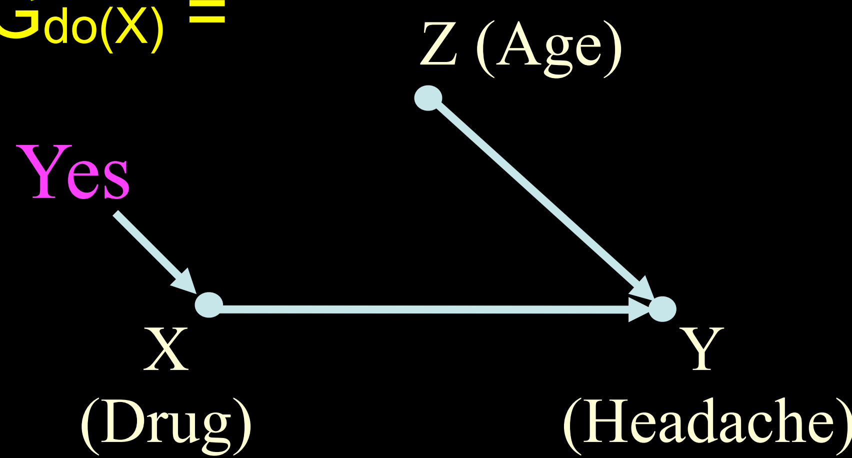
Drug $\leftarrow$ Yes

Headache $\leftarrow f_H$(Drug, Age, U_H)

$G =$



$G_{do(X)} =$



$P(Z, X, Y)$

(observational)

$P(Z_{x=yes}, Y_{x=yes}) = P(Z, Y \mid do(X = Yes))$

(counterfactuals) (interventional)

SCM -- REPRESENTING THE DATA GENERATING MODEL

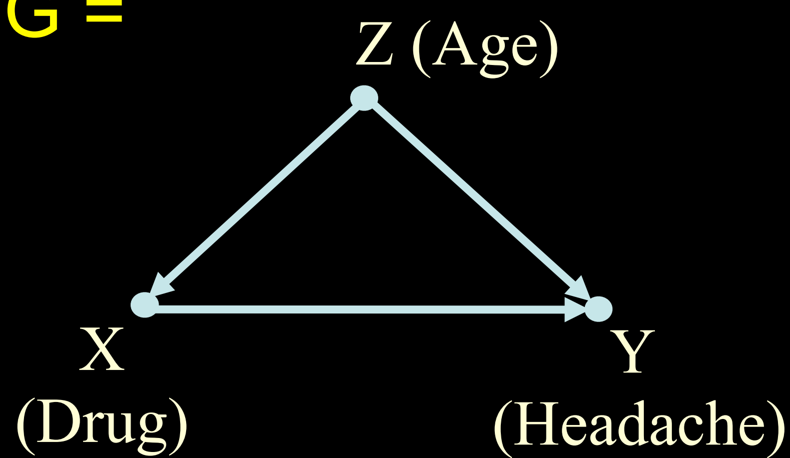
- Processes

Drug $\leftarrow f_D$ (Age, U_D)
Headache $\leftarrow f_H$(Drug, Age, U_H)

- Intervention

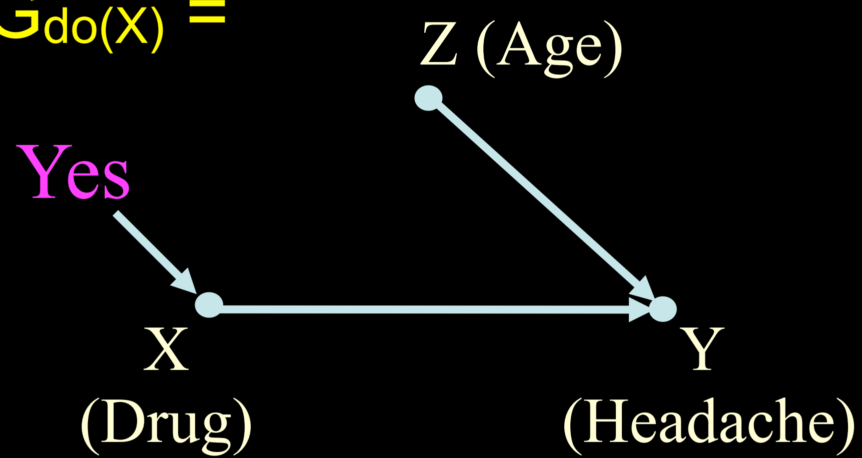
Drug $\leftarrow$ Yes
Headache $\leftarrow f_H$(Drug, Age, U_H)

G =



P(Z, X, Y)
(observational)

$G_{do(X)}$ =



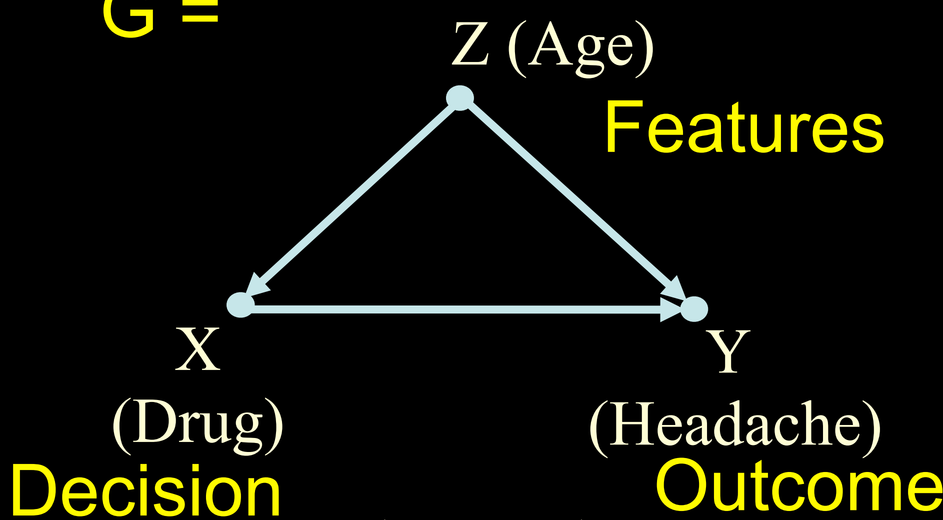
P(Z, Y | do(X = Yes))
(interventional)

SCM -- REPRESENTING THE DATA GENERATING MODEL

- Processes

$$\text{Drug} \leftarrow f_D\,(\text{Age}, U_D)$$
$$\text{Headache} \leftarrow f_H(\text{Drug}, \text{Age}, U_H)$$

G =

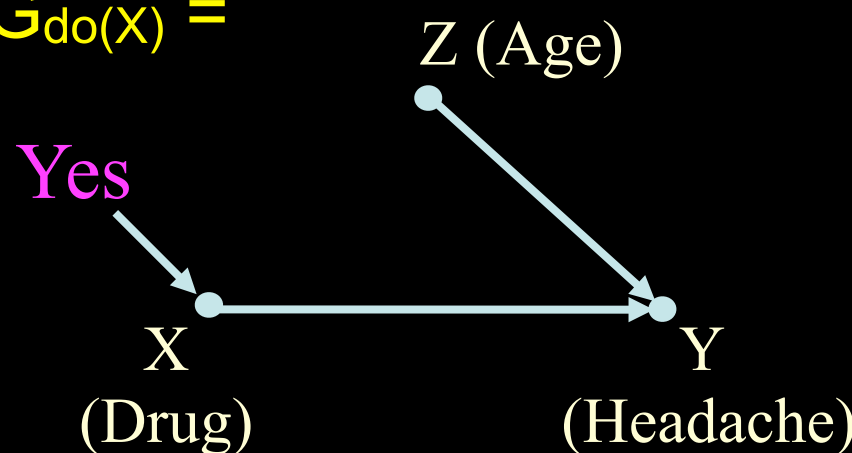


$P(Z, X, Y)$
(observational)

- Intervention

$$\text{Drug} \leftarrow \text{Yes}$$
$$\text{Headache} \leftarrow f_H\,(\text{Drug}, \text{Age}, U_H)$$

$G_{do(X)}$ =



$P(Z, Y \mid do(X = \text{Yes}))$
(interventional)

SCM -- REPRESENTING THE DATA GENERATING MODEL

- Processes

Drug $\leftarrow f_D(\text{Age}, U_D)$

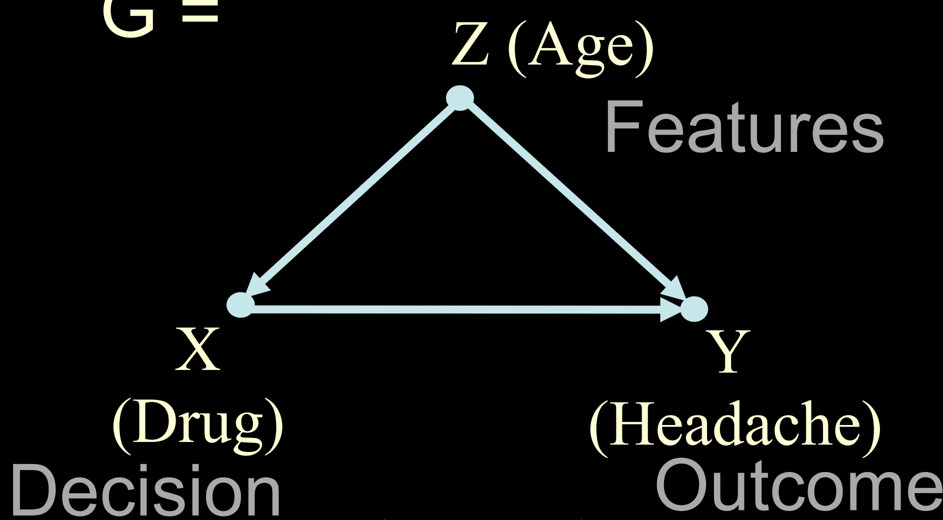
Headache $\leftarrow f_H(\text{Drug}, \text{Age}, U_H)$

- Intervention

Drug $\leftarrow$ Yes

Headache $\leftarrow f_H(\text{Drug}, \text{Age}, U_H)$

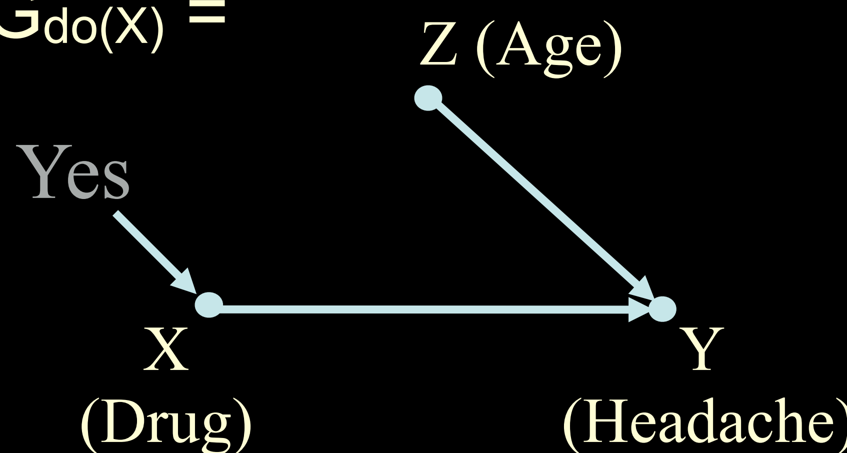
G =



P(Z, X, Y)

(observational)

$G_{do(X)}$ =



P(Z, Y | do(X = Yes))

(interventional)

SCM -- REPRESENTING THE DATA GENERATING MODEL

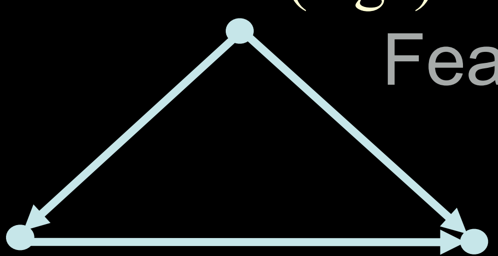
- Processes

Drug $\leftarrow f_D$ (Age, U_D)
Headache $\leftarrow f_H$(Drug, Age, U_H)

- Intervention

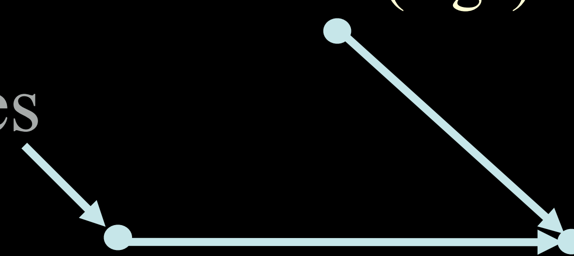
Drug $\leftarrow$ Yes
Headache $\leftarrow f_H$(Drug, Age, U_H)

G =



P(Z, X, Y)

$G_{do(X)}$ =



P(Z, Y | do(X = Yes))

Seeing → Doing

STRUCTURAL CAUSAL MODELS

Definition: A structural causal model M (or, data generating model) is a tuple $(V, U, F, P(u))$, where

- $V = \{V_1,\ldots,V_n\}$ are endogenous variables,
- $U = \{U_1,\ldots,U_m\}$ are exogenous variables,
- $F = \{f_1,\ldots, f_n\}$ are functions determining V,
  for each Vi, $V_i \leftarrow f_i(Pa_i, U_i)$, where $Pa_i \subset V$, $U_i \subset U$.
- $P(u)$ is a distribution over U.

(Axiomatic characterization [Halpern, Galles, Pearl, 1998].)

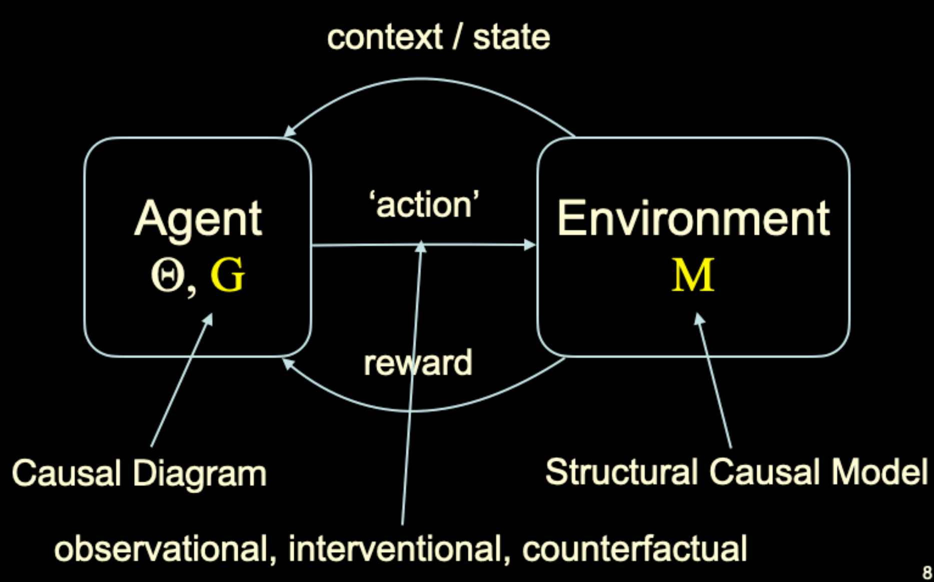
Prop. SCM M implies Pearl Causal Hierarchy (PCH).

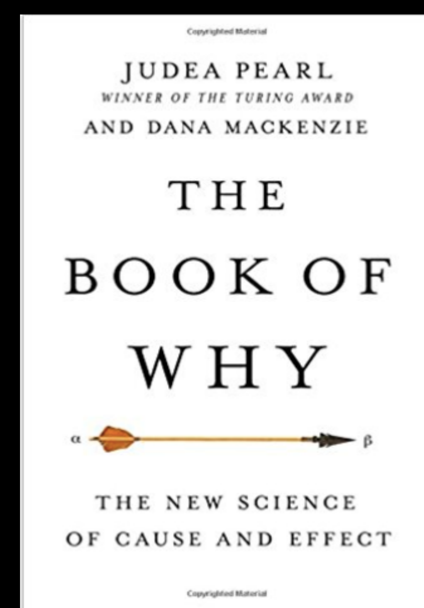
PEARL CAUSAL HIERARCHY (PCH)

PEARL CAUSAL HIERARCHY (PCH)

(LADDER OF CAUSATION)

Causal RL - Big Picture



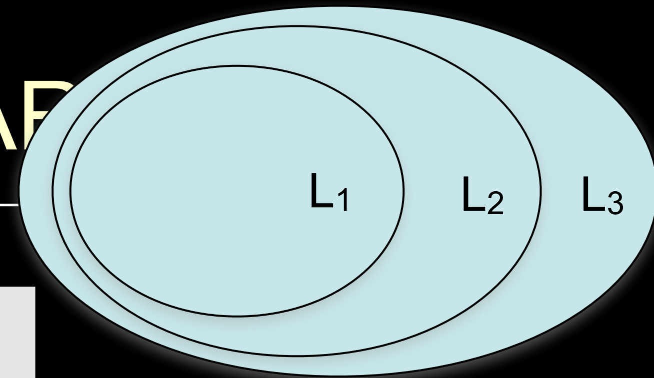


SCM → PEARL CAUSAL HIERARCHY (PCH)

	Layer (Symbol)	Typical Activity	Typical Question	Examples
L_1 👀	Associational $P(y \mid x)$	Seeing ML - (Un)Supervised DT, Bayes net, Regression, NN	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?
L_2 💪	Interventional $P(y \mid do(x), c)$	Doing ML - Reinforcement Causal BN, MDP	What if? What if I do X?	What if I take aspirin, will my headache be cured?
L_3 🧠	Counterfactual $P(y_x \mid x', y')$	Imagination, Introspection	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?

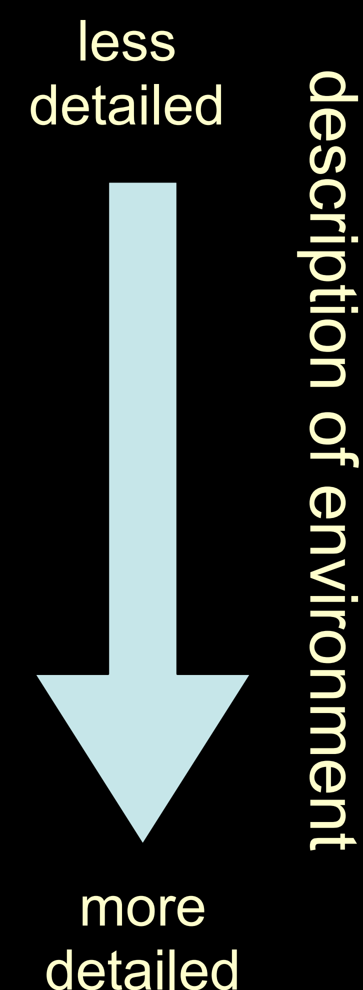
Structural Causal Model

SCM → PEARL CAUSAL HIERAR



	Layer (Symbol)	Typical Activity	Typical Question
L_1	Associational $P(y \mid x)$	Seeing ML - (Un)Supervised DT, Bayes net, Regression, NN	What is? How would seeing X char my belief in Y‘
L_2	Interventional $P(y \mid do(x), c)$	Doing ML - Reinforcement Causal BN, MDP	What if? What if I do X‘
L_3	Counterfactual $P(y_x \mid x', y')$	Imagination, Introspection	Why? What if I had acted different

Structural Causal Model



CAUSAL HIERARCHY THEOREM

[Bareinboim, Correa, Ibeling, Icard, 2020]

Given that an SCM M → PCH, we can show the following:

Theorem (CHT). With respect to Lebesgue measure over (a suitable encoding of L_3-equivalence classes of) SCMs, the subset in which any PCH ‘collapse’ is measure zero.

Informally, for almost any SCM (i.e., almost any possible environment), the PCH does not collapse, i.e., the layers of the hierarchy remains distinct.

Corollary. To answer question at Layer i (about a certain interaction), one needs knowledge at layer i or higher.

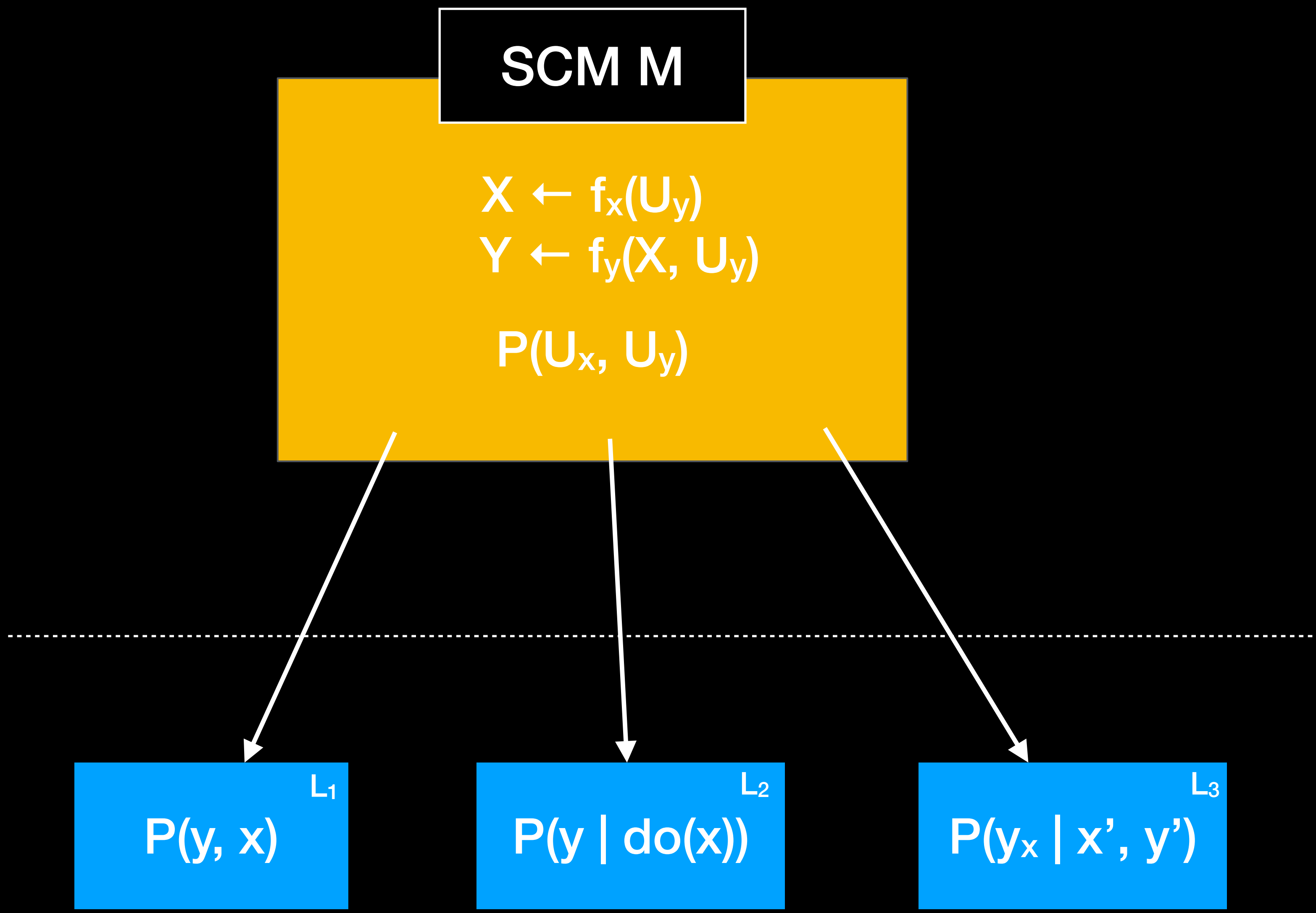
WHY IS CAUSAL INFERENCE "NON-TRIVIAL"? SCMs ARE ALMOST NEVER OBSERVED

SCM M

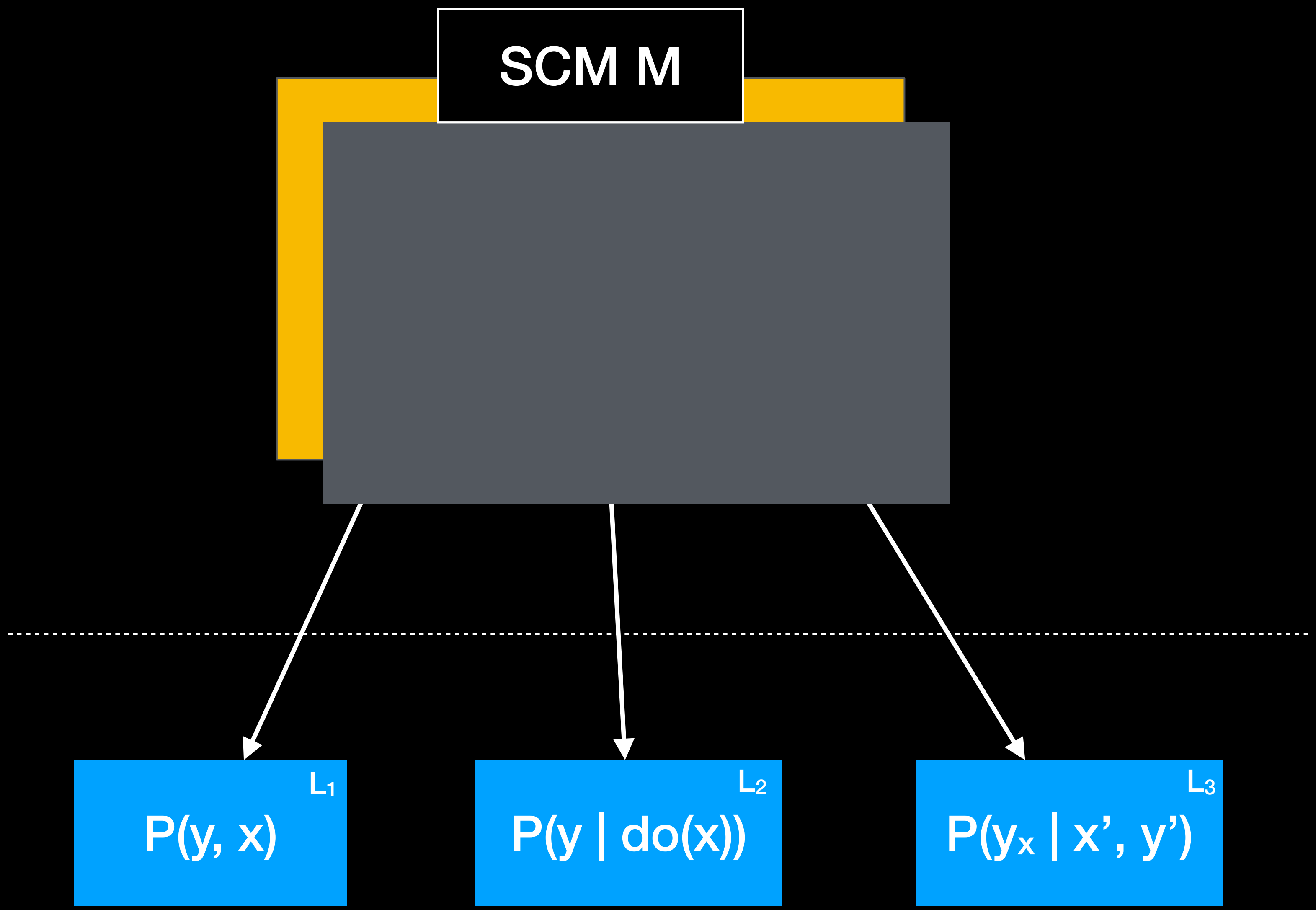
$X \leftarrow f_x(U_y)$
$Y \leftarrow f_y(X, U_y)$

$P(U_x, U_y)$

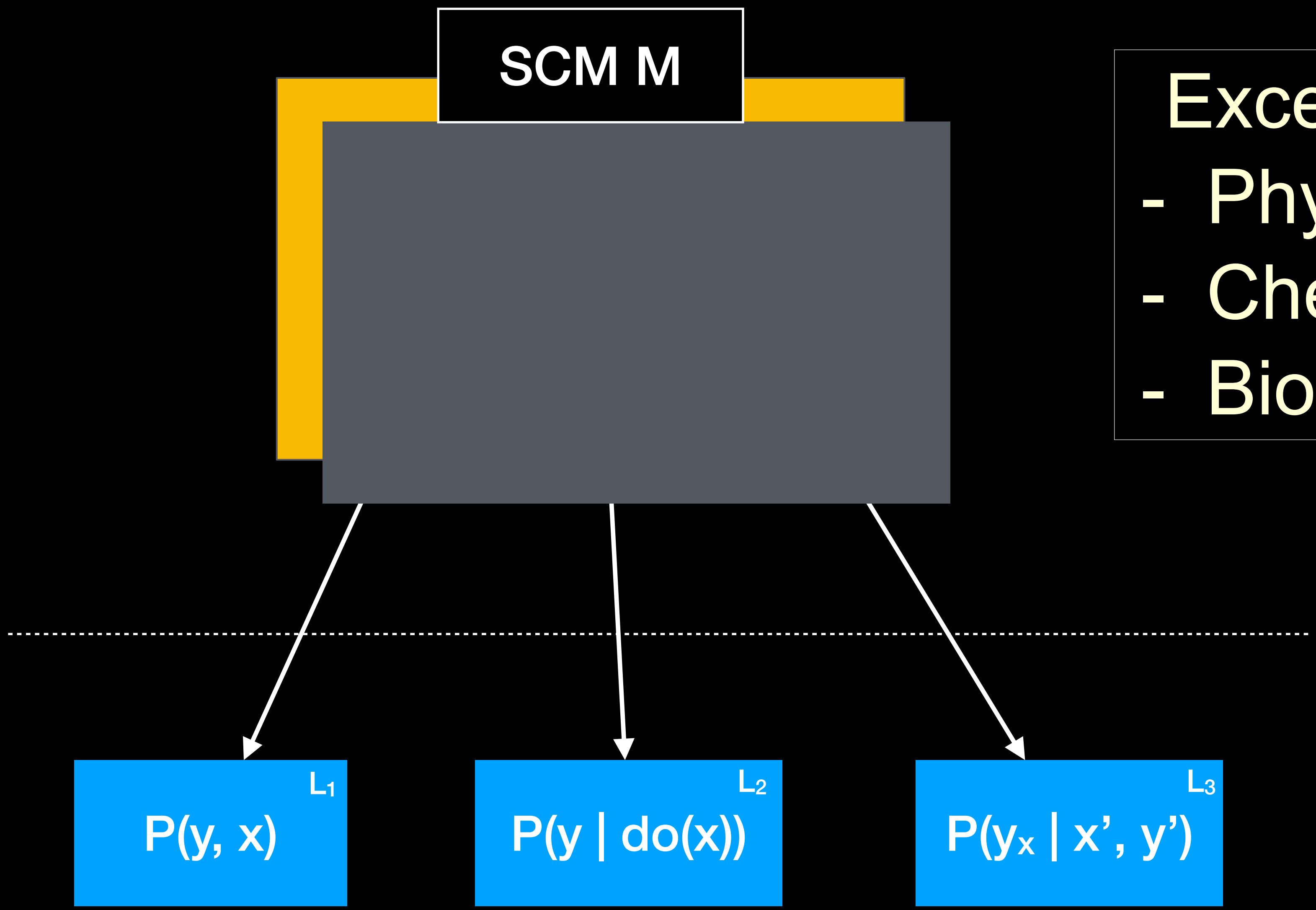
WHY IS CAUSAL INFERENCE "NON-TRIVIAL"? SCMs ARE ALMOST NEVER OBSERVED



WHY IS CAUSAL INFERENCE “NON-TRIVIAL”? SCMs ARE ALMOST NEVER OBSERVED

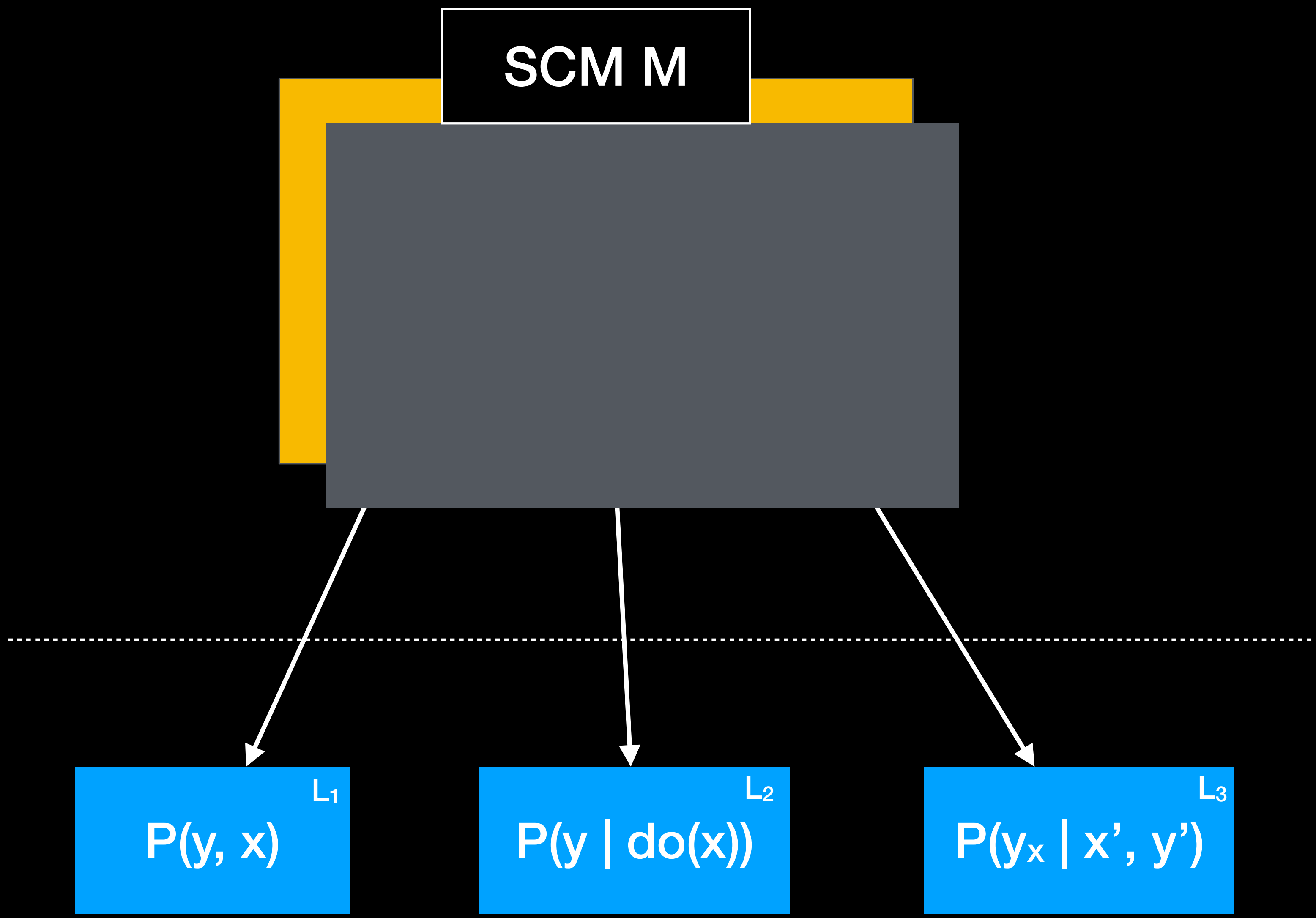


WHY IS CAUSAL INFERENCE "NON-TRIVIAL"? SCMs ARE ALMOST NEVER OBSERVED



Exceptions:
- Physics
- Chemistry
- Biology

WHY IS CAUSAL INFERENCE “NON-TRIVIAL”? SCMs ARE ALMOST NEVER OBSERVED



WHY IS CAUSAL INFERENCE "NON-TRIVIAL"? SCMs ARE ALMOST NEVER OBSERVED

SCM M

L_1

$P(y, x)$

L_2

$P(y \mid do(x))$

L_3

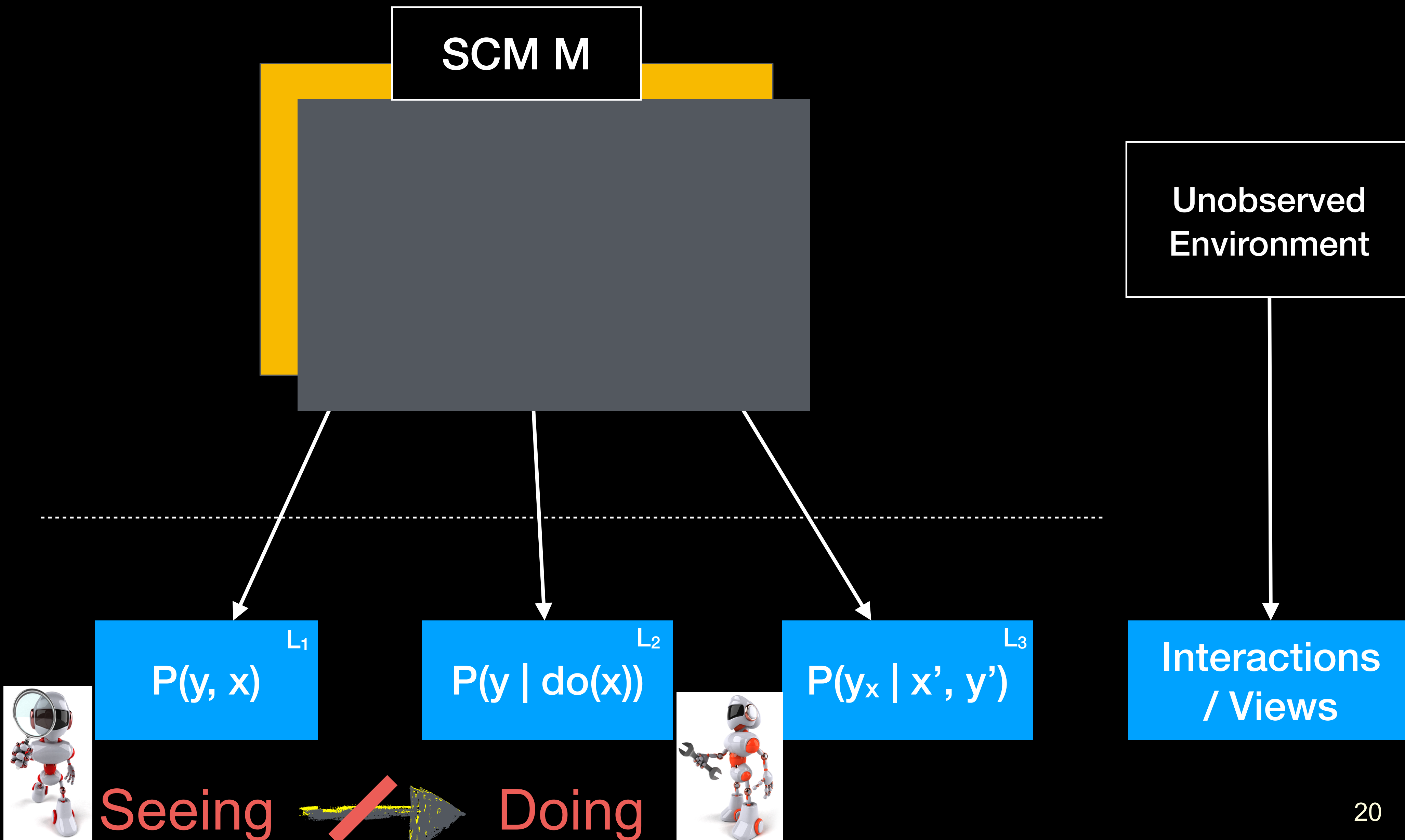
$P(y_x \mid x', y')$

?

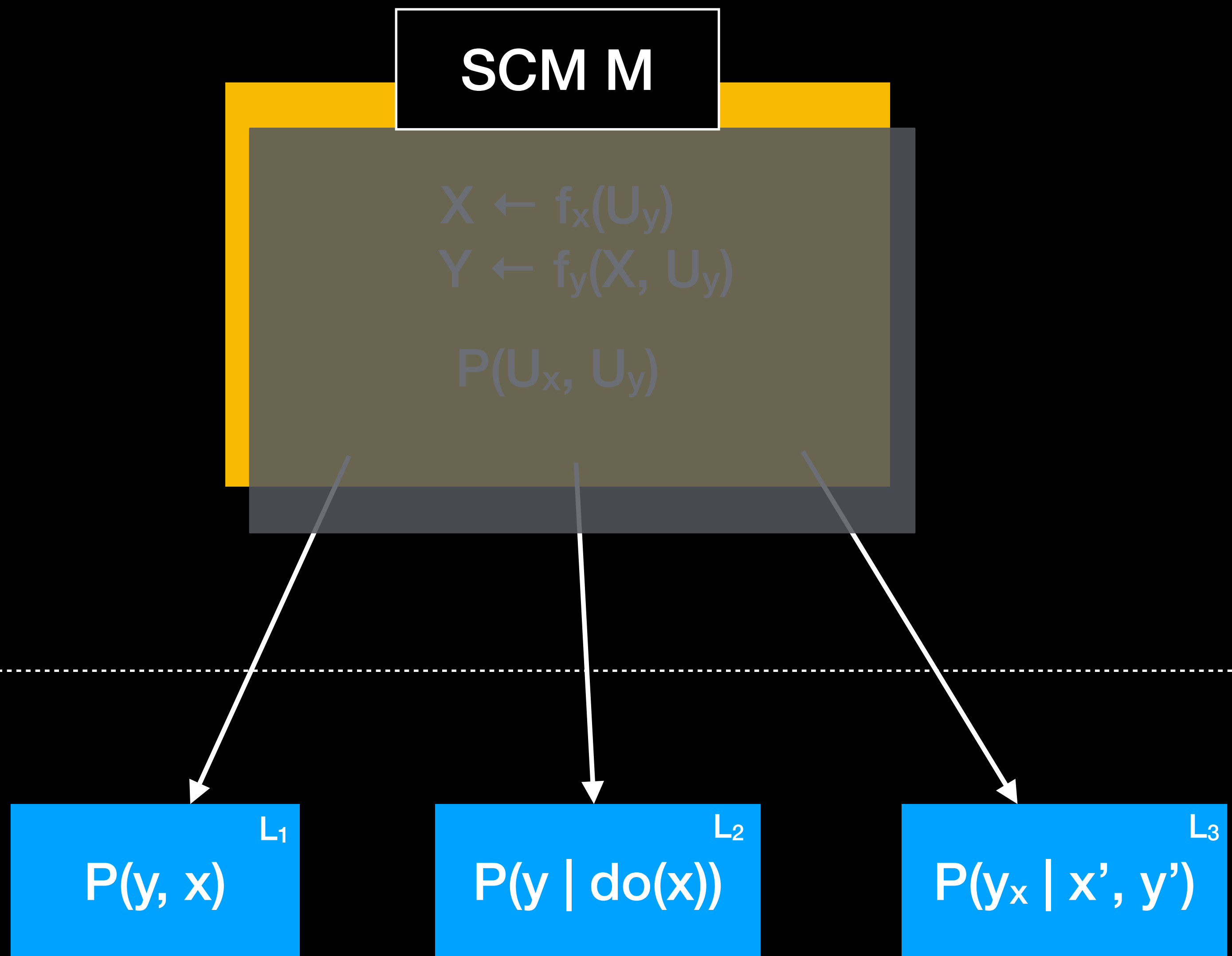
Seeing

Doing

WHY IS CAUSAL INFERENCE "NON-TRIVIAL"? SCMs ARE ALMOST NEVER OBSERVED

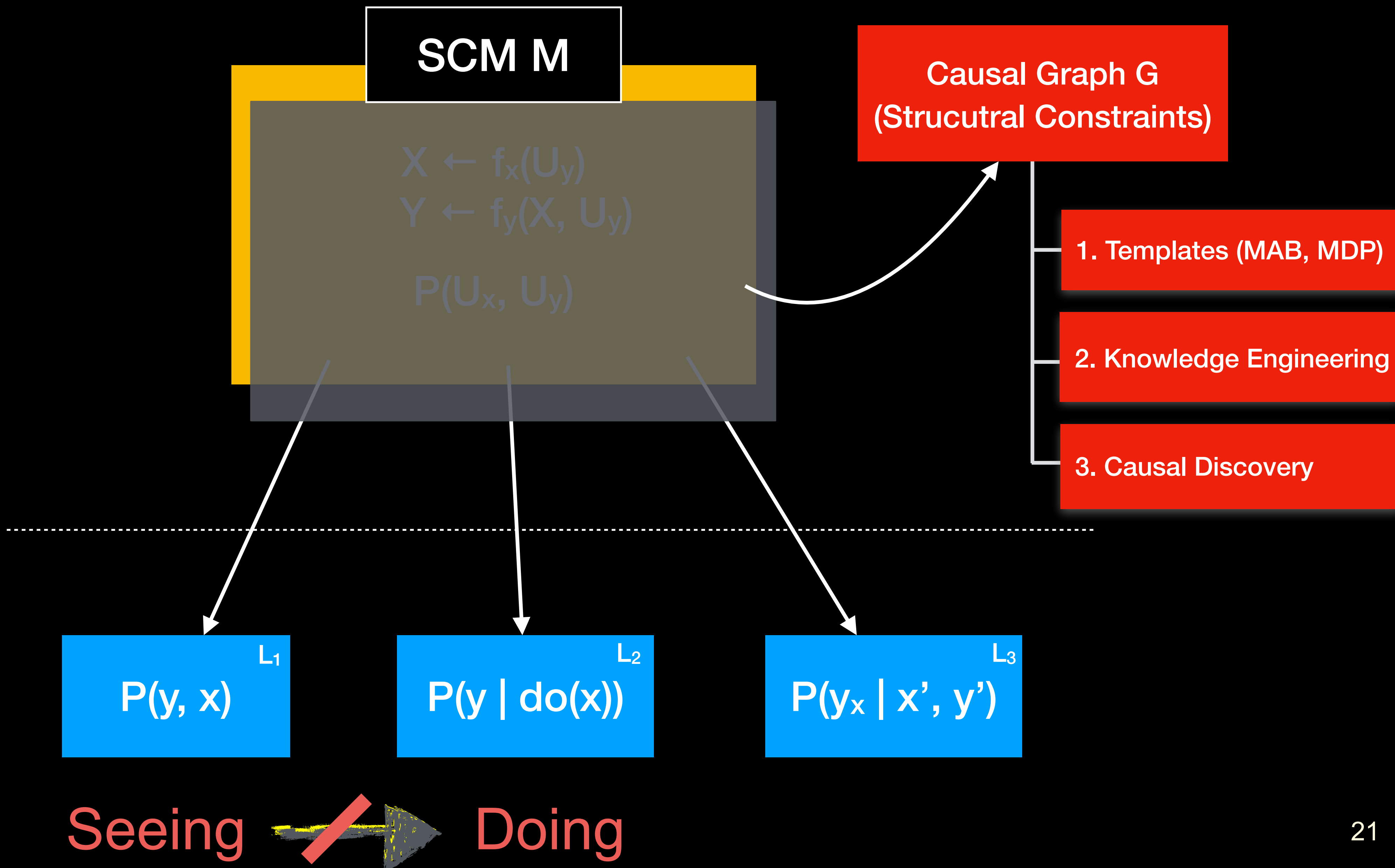


ENCODING STRUCTURAL CONSTRAINTS — CLASSES OF CAUSAL GRAPHS



Seeing $\not\rightarrow$ Doing

ENCODING STRUCTURAL CONSTRAINTS — CLASSES OF CAUSAL GRAPHS



KEY POINTS (SO FAR)

- The environment (mechanisms) can be modeled as an SCM
  - SCM M (specific environment) is rarely observable
- Still, each SCM M can be probed through qualitatively different types of interactions (distributions) -- the PCH -- i.e.:
  - L_1: Observational
  - L_2: Interventional
  - L_3: Counterfactual
- CHT (Causal Hierarchy Thm.): For almost any SCM, lower layers (say, L_i) underdetermines higher layers (L_{i+1}).
  - This delimits what an agent can infer based on the different types of interactions (and data) it has with the environment;
  - For instance, from passively observing the environment (L_1), it cannot infer how to act (L_2).
  - From intervening in the environment (L_2), it can't infer how things would have been had she acted differently (L_3).
- Causal Graph G is a surrogate of the invariances of the SCM M.

CURRENT METHODS IN RL & CI THROUGH CRL LENS

REINFORCEMENT LEARNING AND CAUSAL INFERENCE

Goal: Learn a policy ∏ s.t. sequence of actions ∏(.) = $(X_1, X_2\ldots, X_n)$ maximizes reward $E_{\prod}[Y \mid do(X)]$.

Current strategies found in the literature (circa 2020):

1. Online learning
   - Agent performs experiments herself
   - Input: experiments $\{(do(X_i), Y_i)\}$; Learned: $P(Y \mid do(X))$

Offline

2. Off-policy learning
   - Agent learns from other agents' actions
   - Input: samples $\{(do(X_i), Y_i)\}$; Learned: $P(Y \mid do(X))$
3. Do-calculus learning
   - Agent observes other agents acting
   - Input: samples $\{(X_i, Y_i)\}$, G; Learned: $P(Y \mid do(X))$

REINFORCEMENT LEARNING AND CAUSAL INFERENCE

Goal: Learn a policy $\prod$ s.t. sequence of actions $\prod(.) = (X_1, X_2\ldots, X_n)$ maximizes reward $E_{\prod}[Y \mid do(X)]$.

Current strategies found in the literature (circa 2020):

1. Online learning (→ 💪)
   - Agent performs experiments herself
   - Input: experiments $\{(do(X_i), Y_i)\}$; Learned: $P(Y \mid do(X))$
2. Off-policy learning (💪 → 💪)
   - Agent learns from other agents' actions
   - Input: samples $\{(do(X_i), Y_i)\}$; Learned: $P(Y \mid do(X))$
3. Do-calculus learning (👀 → 💪)
   - Agent observes other agents acting
   - Input: samples $\{(X_i, Y_i)\}$, G; Learned: $P(Y \mid do(X))$

Offline

REINFORCEMENT LEARNING AND CAUSAL INFERENCE

Goal: Learn a policy $\prod$ s.t. sequence of actions $\prod(.) = (X_1, X_2\ldots, X_n)$ maximizes reward $E_{\prod}[Y \mid do(X)]$.

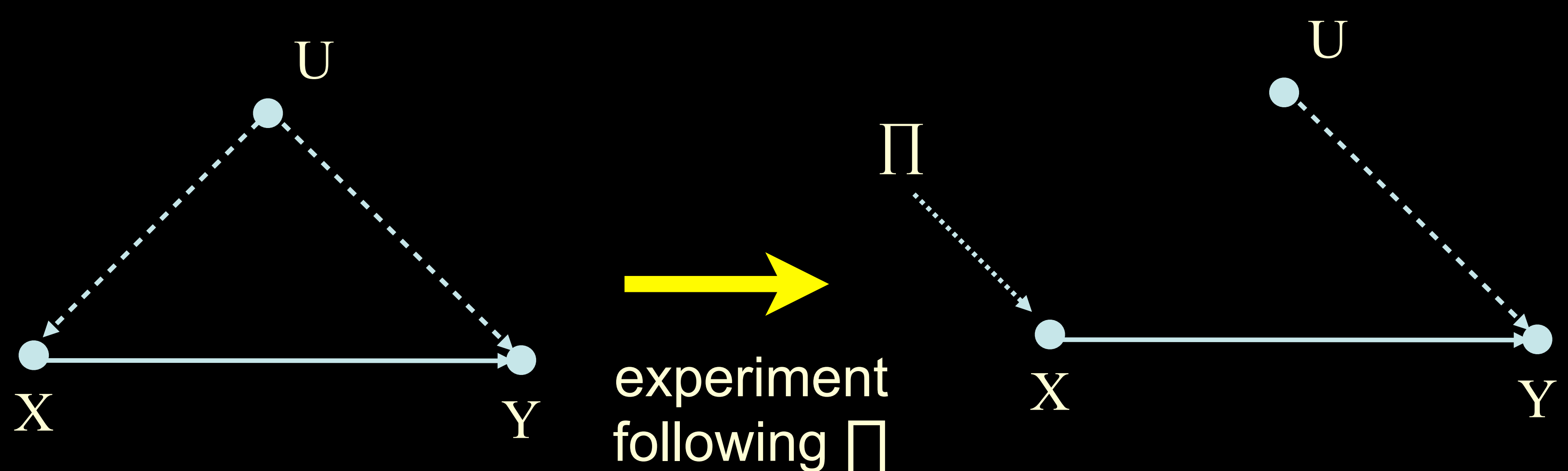
Current strategies found in the literature (circa 2020):

1. Online learning ($\rightarrow$ do(x))
   - Agent performs experiments herself
   - Input: experiments $\{(do(X_i), Y_i)\}$; Learned: $P(Y \mid do(X))$
2. Off-policy learning (do(x) $\rightarrow$ do(x))
   - Agent learns from other agents' actions
   - Input: samples $\{(do(X_i), Y_i)\}$; Learned: $P(Y \mid do(X))$
3. Do-calculus learning (see(v) $\rightarrow$ do(x))
   - Agent observes other agents acting
   - Input: samples $\{(X_i, Y_i)\}$, G; Learned: $P(Y \mid do(X))$

Offline

1. ONLINE LEARNING

- Finding x* is immediate once E[Y | do(X)] is learned.
- E[Y | do(X)] can be estimated through randomized experiments or adaptive strategies.
  - Pros: Robust against unobserved confounders (UCs)
  - Cons: Experiments can be expensive or impossible

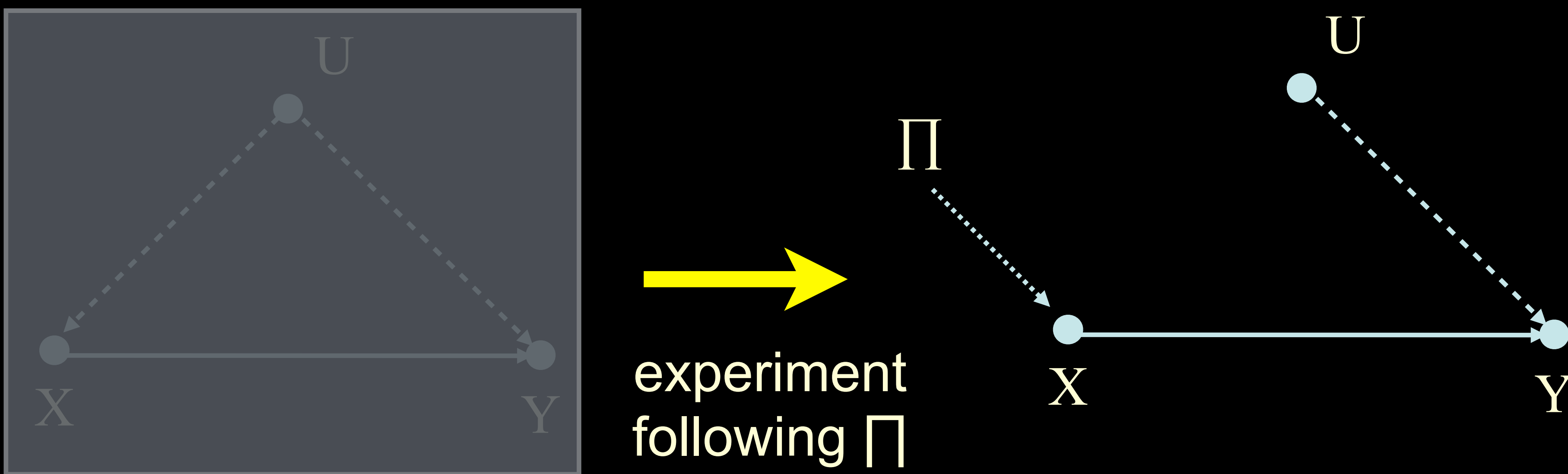


Pre-randomization (passive world)

Post-randomization (active)

1. ONLINE LEARNING

- Finding x* is immediate once E[Y | do(X)] is learned.
- E[Y | do(X)] can be estimated through randomized experiments or adaptive strategies.
  - Pros: Robust against unobserved confounders (UCs)
  - Cons: Experiments can be expensive or impossible



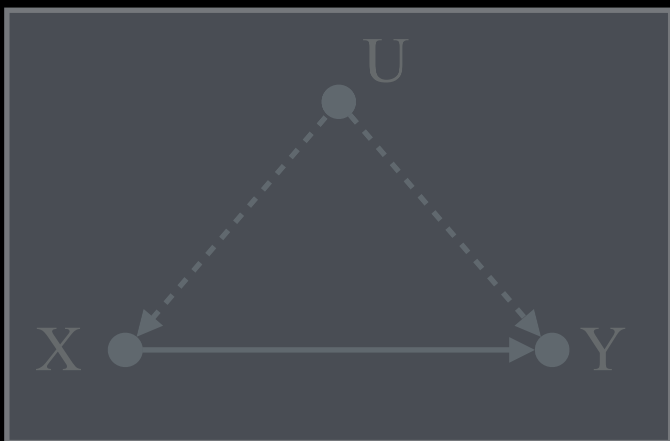
Pre-randomization (passive world)

Post-randomization (active)

* More details: [Fisher, 1936; Auer et al., 2002; Jaksch et al., 2010; Lattimore et al., 2016].

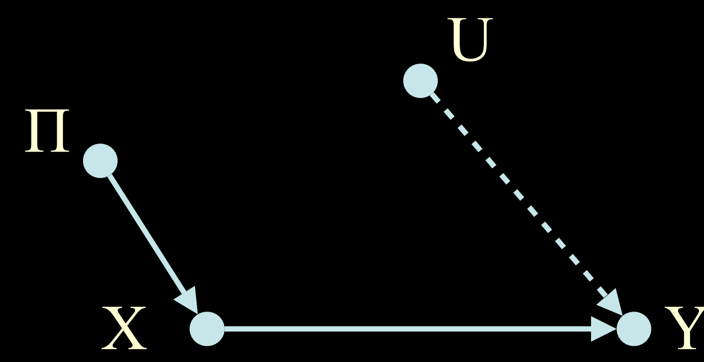
1. ONLINE LEARNING

(interventional learning)



Pre-randomization world (passive)

experiment following ∏



under do(X)

Agent

no data

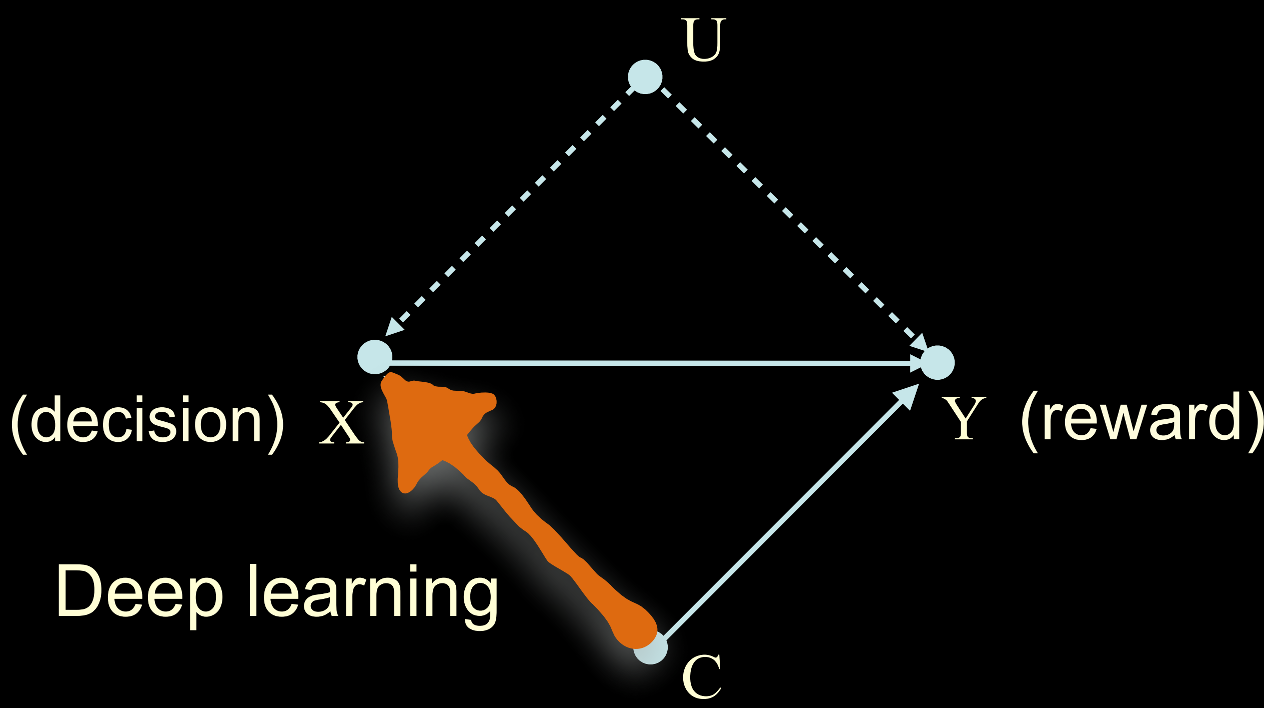
$E_{x\sim\pi}[Y \mid do(x)]$

do(x_0) do(x_1) … do(x_0)

* Online learning can be improved thr. causal machinery [ZB, ICML'20].

NOTE: COVARIATE-SPECIFIC CAUSAL EFFECTS (CONTEXTUAL)

- Model can be augmented to accommodate set of observed covariates C (also known as context); U is the set of (remaining) unobserved confounders (UCs).

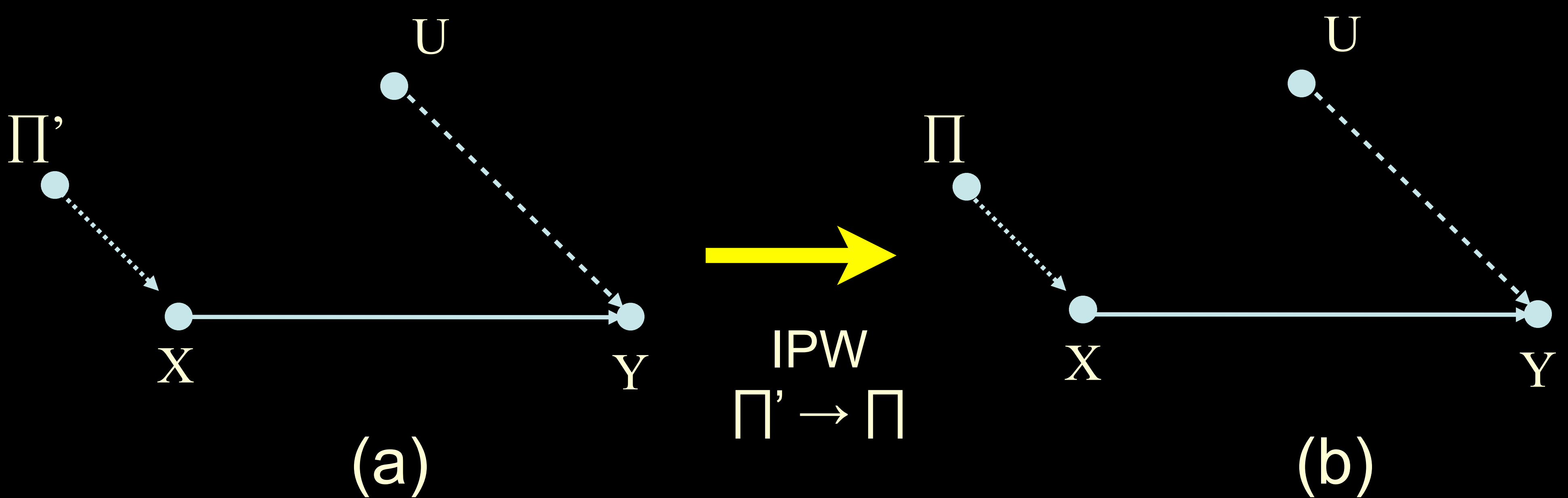


» Challenge: high-dimensional C

- Goal: learn a policy ∏(c) so as to optimize based on the c-specific causal effect, P(Y | do(X), C = c).

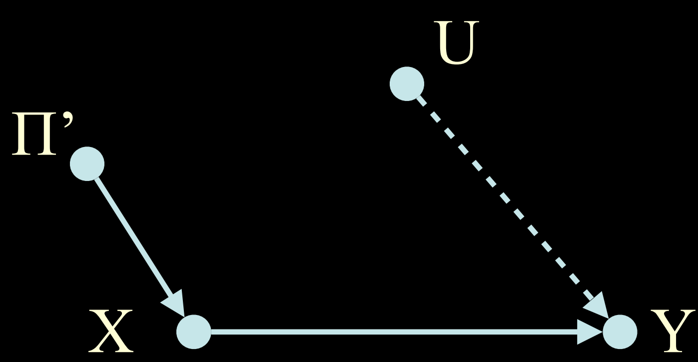
2. OFF-POLICY LEARNING

- E[Y | do(X)] can be estimated through experiments conducted by other agents and different policies.
  - Pros: no experiments need to be conducted
  - Cons: rely on assumptions that (a_1) same variables were randomized and (a_2) context matches (e.g., C = {}).

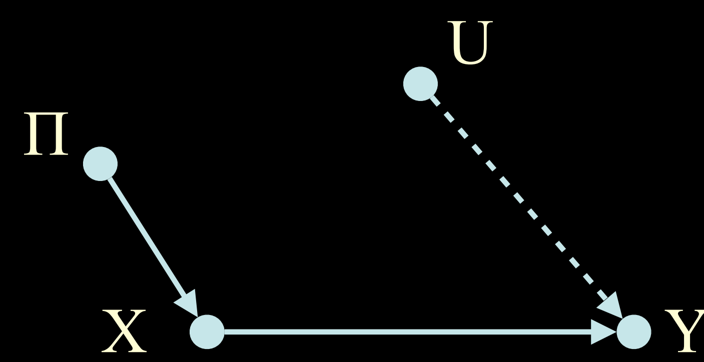


* More details: [Watkins & Dayan, 1992; Dudik et al., 2011; Jiang & Li, 2016].

2. OFF-POLICY LEARNING



under do(X)



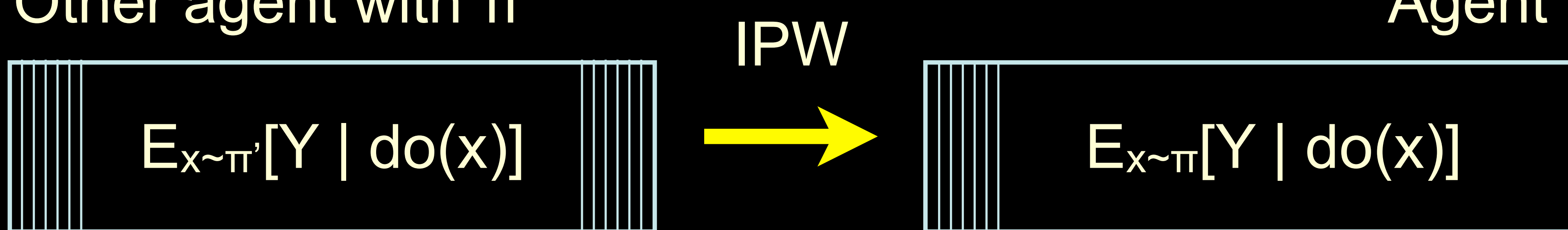
under do(X)

$$P_{\Pi}(y \mid do(x)) = \sum_{x,c} P_{\Pi'}(y, x, c)\frac{P_{\Pi}(x \mid c)}{P_{\Pi'}(x \mid c)}$$

A lot of work here since the variance may blow up…

Other agent with π’

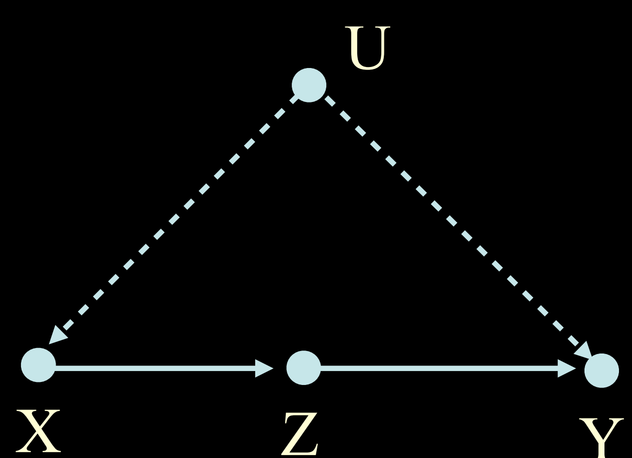
Agent



$do(x_0)$ $do(x_1)$ … $do(x_0)$

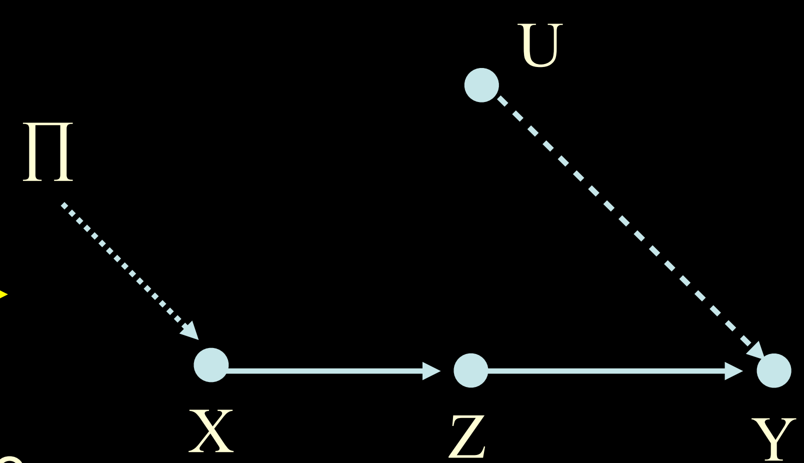
3. DO-CALCULUS LEARNING*

- E[Y | do(X)] can be estimated from non-experimental data (also called *natural / behavioral regime*)
  - Pros: estimation is feasible even when context is unknown and experimental variables do not match (i.e., off-policy assumptions are violated).
  - Cons: Results are contingent on the model; for weak models, effect is not uniquely computable (not ID).



Passive-world data-collection

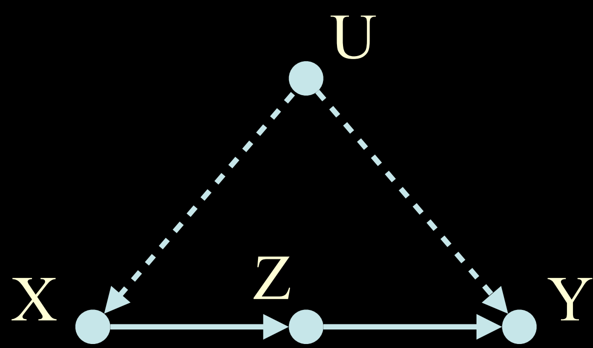
do-calc inference



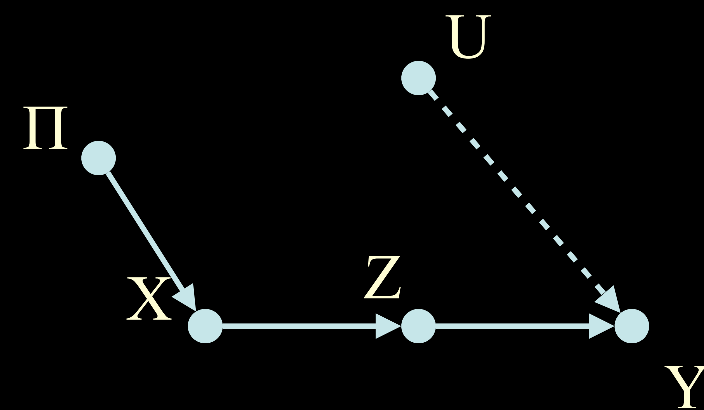
Do-world (Post-interventional)

* For details, see data-fusion survey [Bareinboim & Pearl, PNAS'2016].

3. DO-CALCULUS LEARNING



Causal Graph G



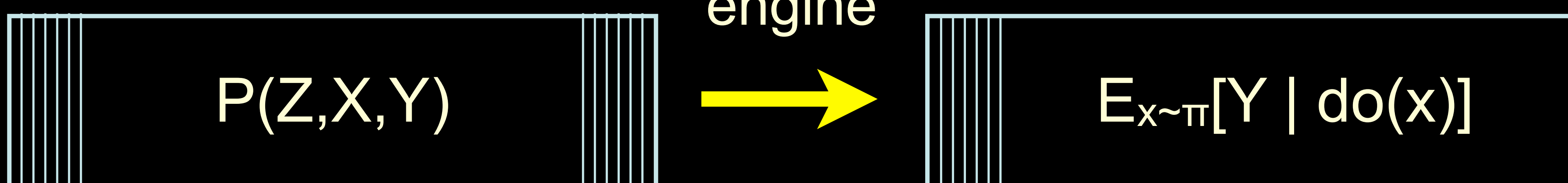
under hypothetical do(X)

$\Sigma_z P(z|x) \Sigma_{x'} P(y|x',z) P(x')$

do-calc inference engine

Observation

Agent



obs, obs, … obs

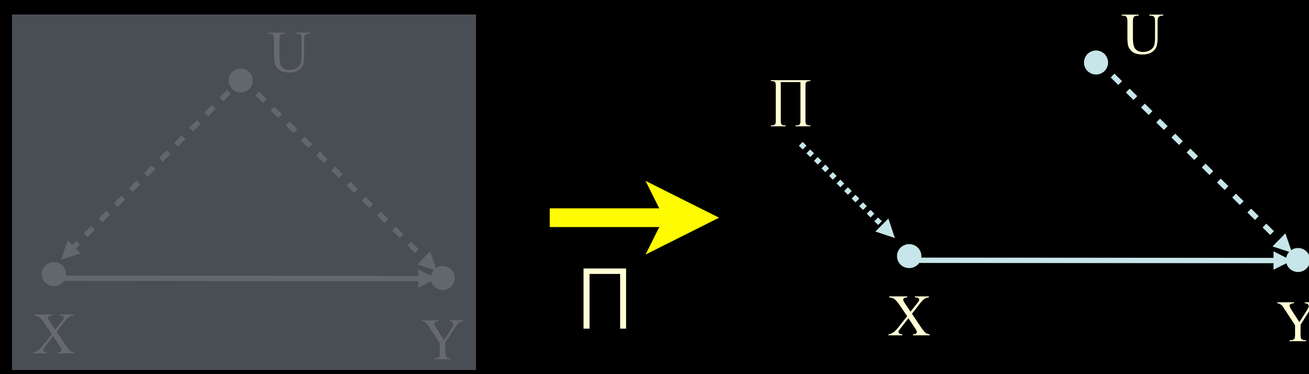
do(z), obs … do(w)

* For a more general treatment, see (LCB, UAI'19)

SUMMARY RL-CAUSAL (CIRCA 2020)

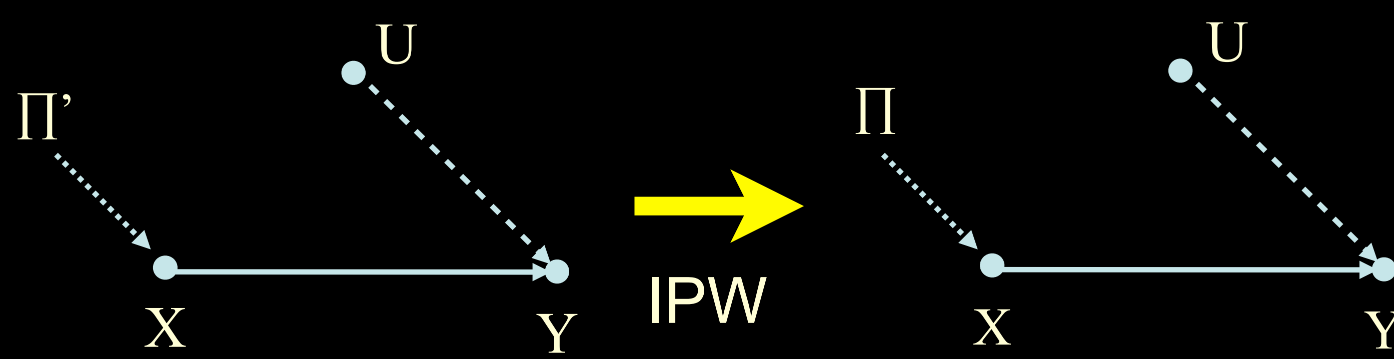
1. Online

(→ $do_{\pi}(x)$)



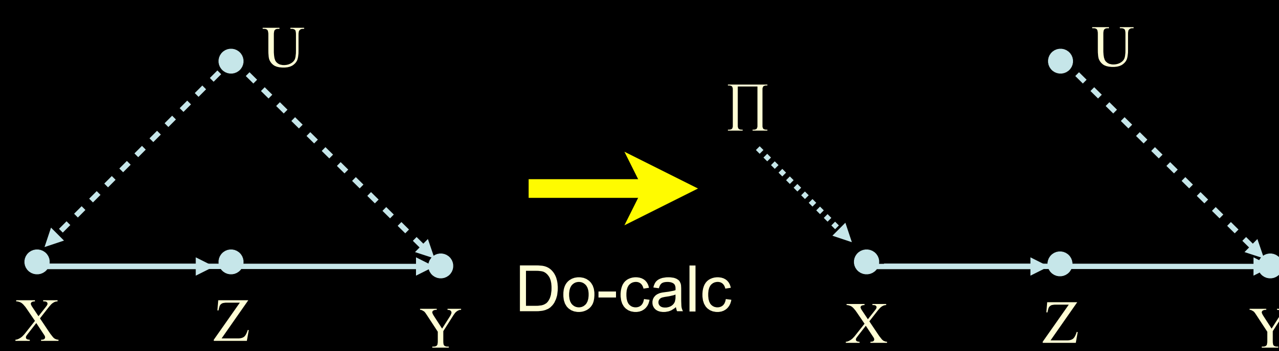
2. Off-policy

($do_{\pi'}(x)$ → $do_{\pi}(x)$)



3. Do-calculus

(see(.) → $do_{\pi}(x)$)



Do these strategies always work?

IS LEARNING IN INTERACTIVE SYSTEMS ESSENTIALLY DONE?

IF NOT, WHAT IS MISSING?

TOWARDS
CAUSAL REINFORCEMENT LEARNING

🧘‍♀️

CRL NEW CHALLENGES & LEARNING OPPORTUNITIES (I)

Task 1 (IJCAI'17, NeurIPS'19, ICML'20)

Generalized Policy Learning

(combining online + offline learning)

Task 2 (NeurIPS'18, AAAI'19)

When and where to intervene?

(refining the policy space)

Task 3 (NeurIPS'15, ICML'17)

Counterfactual Decision-Making

(changing optimization function based on intentionality, free will, and autonomy)

CRL NEW CHALLENGES & LEARNING OPPORTUNITIES (II)

Task 4 (NeurIPS'14, PNAS'16, UAI'19, AAAI'20)

Generalizability & robustness of causal claims (transportability & structural invariances)

Task 5 (NeurIPS'17, ICML'18, NeurIPS'19)

Learning causal model by combining observations (L_1) and experiments (L_2)

Task 6 (R-66 @CausalAI)

Causal Imitation Learning

TASK 1.
GENERALIZED POLICY LEARNING
(Combining Online and Offline Learning)

TASK 1.
GENERALIZED POLICY LEARNING

(Combining Online and Offline Learning)

Junzhe Zhang

CRL-TASK 1. GENERALIZED POLICY LEARNING (GPL)

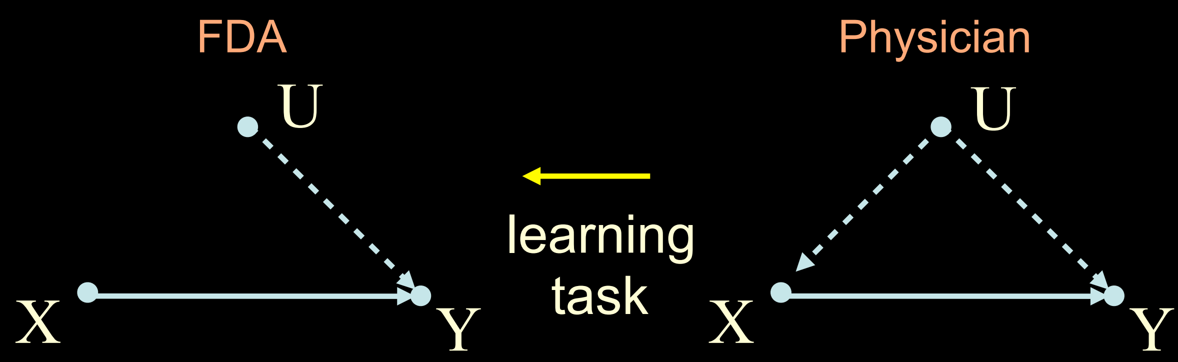
- Online learning is usually undesirable due to financial, technical, or ethical constraints. In general, one wants to leverage data collected under different conditions to speed up learning, without having to start from scratch.
- On the other hand, the conditions required by offline learning are not always satisfied in many practical, real world settings.
- In this task, we move towards realistic learning scenarios where these modalities come together, including when the most traditional, and provably necessary, assumptions do not hold.

GENERALIZED POLICY LEARNING

- Off-policy a_2 ✕
- Do-calc ID ✕
- Online ?

Task 1. Input: $P(x, y)$, learn: $P(y \mid do(x))$.

- Robotics: learning by demonstration when the teacher can observe a richer context (e.g., more accurate sensors).
- Medical: optimal experimental design from observational data.

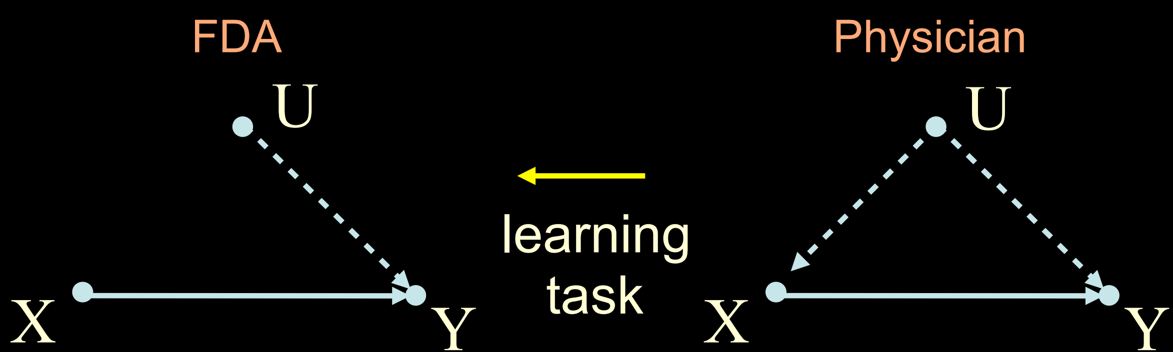


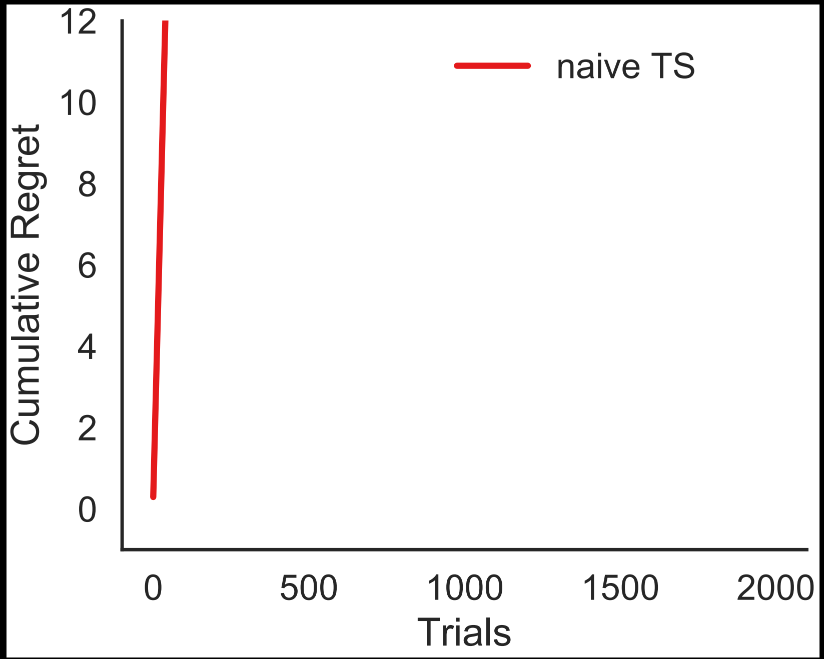
GENERALIZED POLICY LEARNING

- Off-policy a_2 ✗
- Do-calc ID ✗
- Online ?

Task 1. Input: $P(x, y)$, learn: $P(y \mid do(x))$.

- Robotics: learning by demonstration when the teacher can observe a richer context (e.g., more accurate sensors).
- Medical: optimal experimental design from observational data.





Let's ignore their differences, and pretend that physician and FDA are exchangeable — call "naive TS".
In other words, "naive TS" attempts to use observational data as prior.

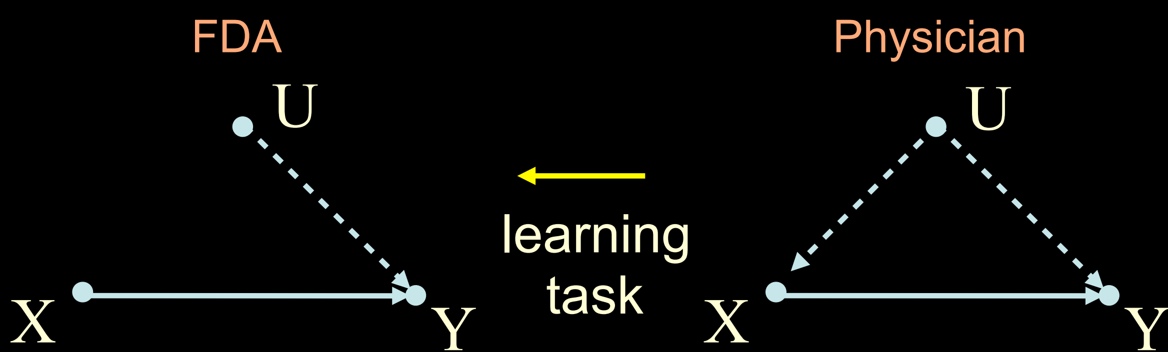
Traditional TS means ignoring the observational data.

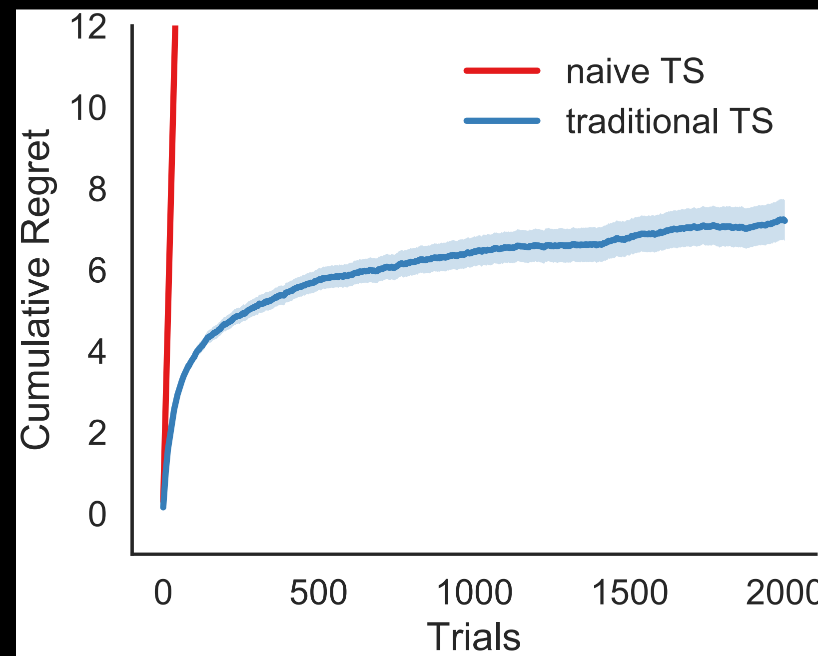
GENERALIZED POLICY LEARNING

- Off-policy a_2 ✗
- Do-calc ID ✗
- Online ?

Task 1. Input: $P(x, y)$, learn: $P(y \mid do(x))$.

- Robotics: learning by demonstration when the teacher can observe a richer context (e.g., more accurate sensors).
- Medical: optimal experimental design from observational data.





Let's ignore their differences, and pretend that physician and FDA are exchangeable — call "naive TS".

In other words, "naive TS" attempts to use observational data as prior.

Traditional TS means ignoring the observational data.

How could this be happening?!

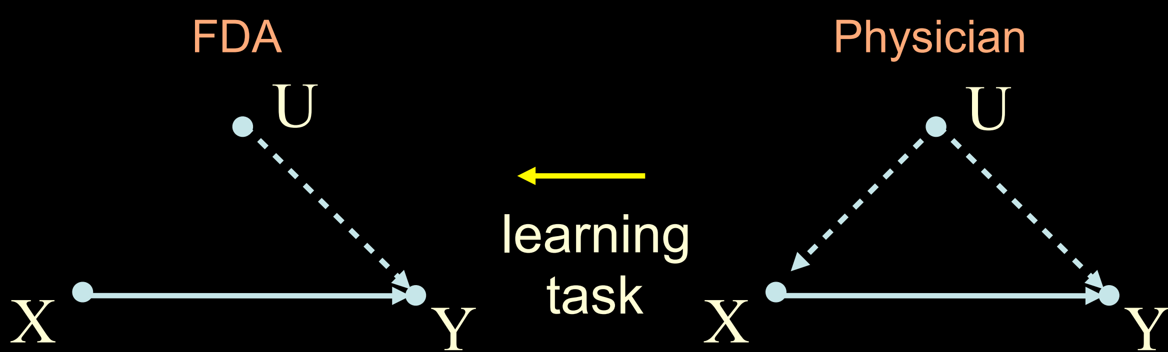
Could more data be hurting?

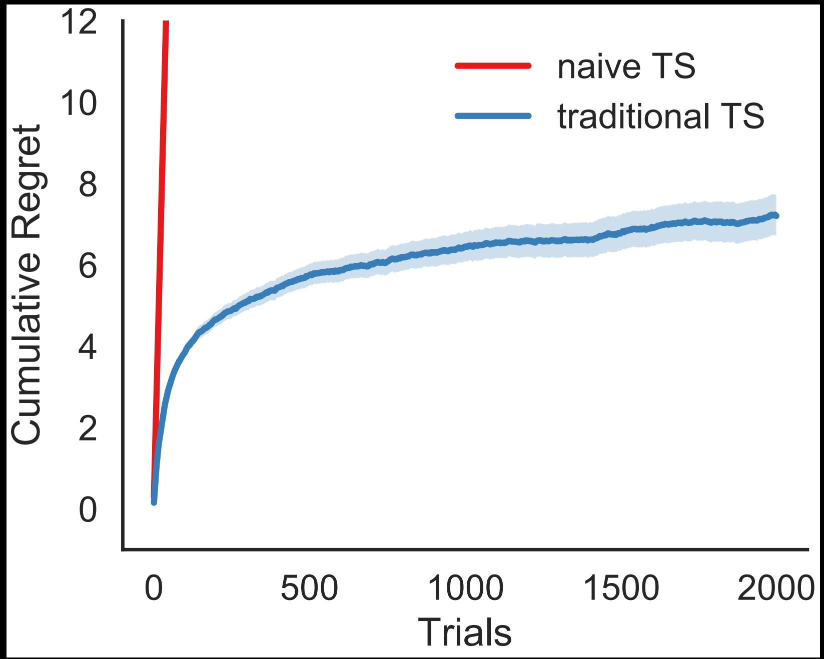
GENERALIZED POLICY LEARNING

- Off-policy a_2 ✗
- Do-calc ID ✗
- Online ?

Task 1. Input: $P(x, y)$, learn: $P(y \mid do(x))$.

- Robotics: learning by demonstration when the teacher can observe a richer context (e.g., more accurate sensors).
- Medical: optimal experimental design from observational data.





Let's ignore their differences, and pretend that physician and FDA are exchangeable — call "naive TS".

In other words, "naive TS" attempts to use observational data as prior.

Traditional TS means ignoring the observational data.

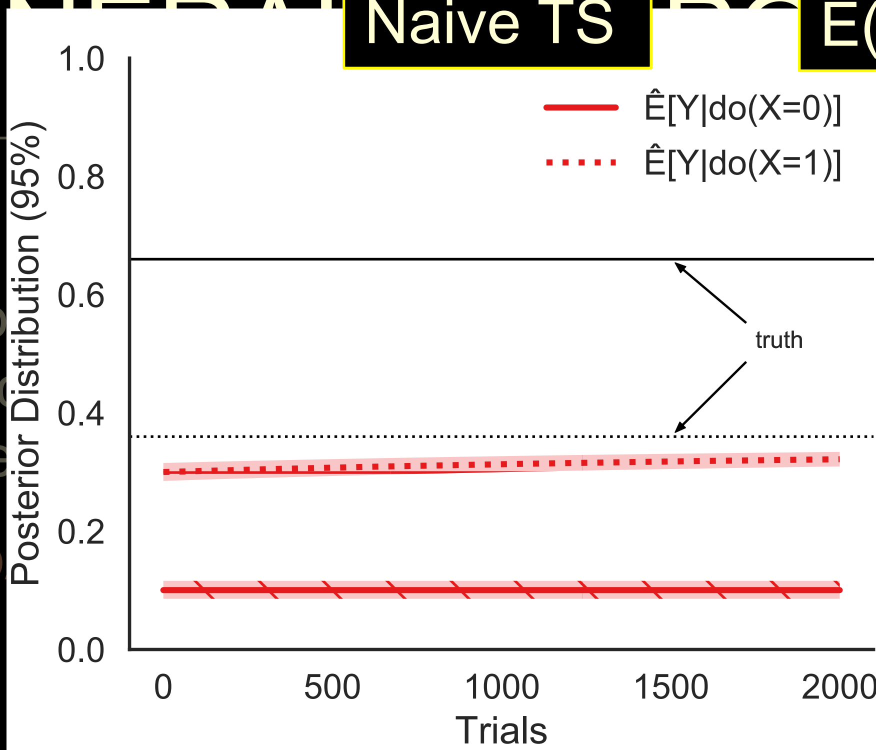
How could this be happening?!

Could more data be hurting?

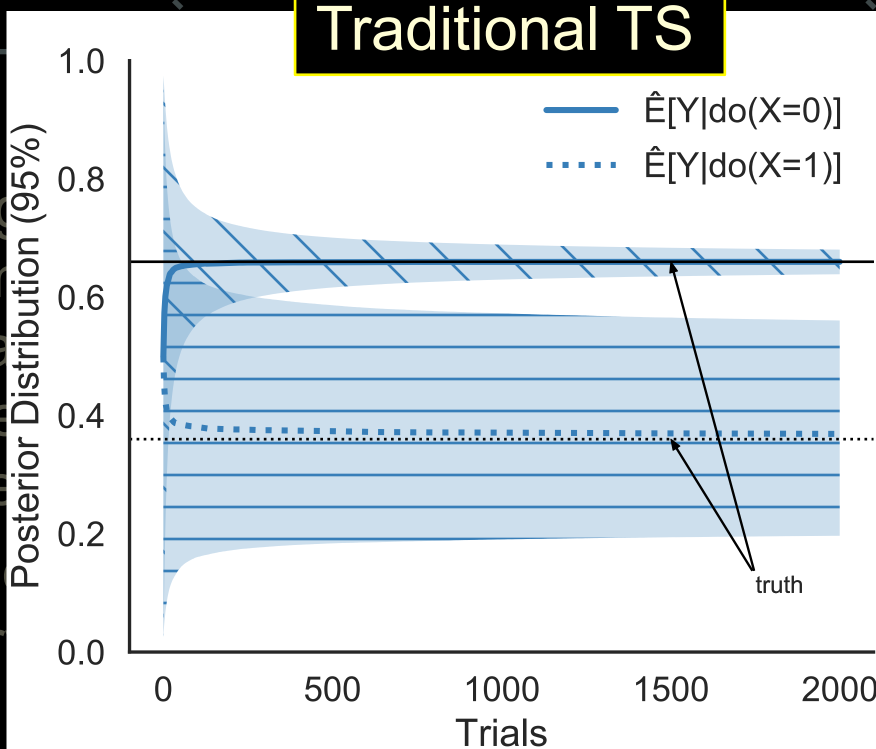
$E(Y \mid X = 0) < E(Y \mid X = 1)$
$E(Y \mid do(X = 0)) > E(Y \mid do(X = 1))$

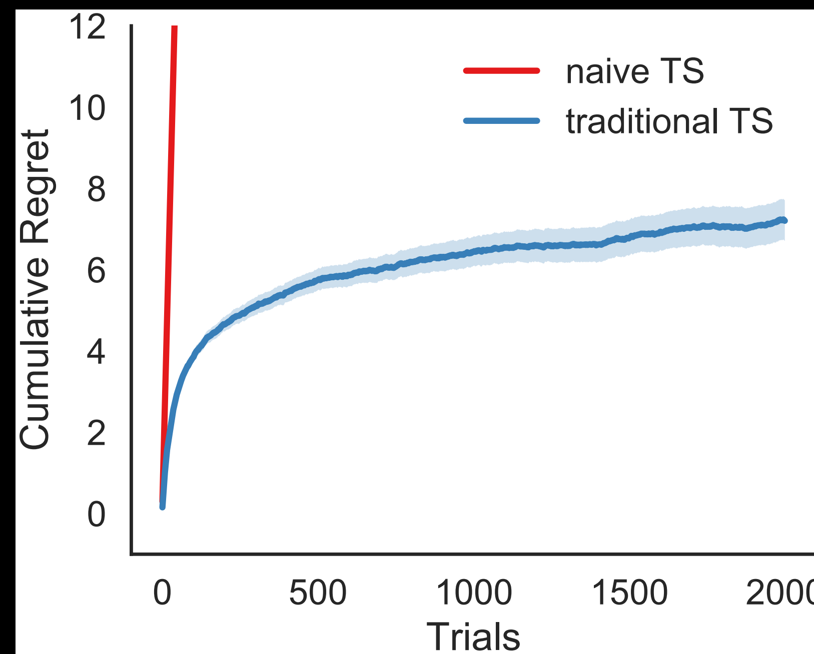
Why is naive-TS doing so badly?

Naive TS



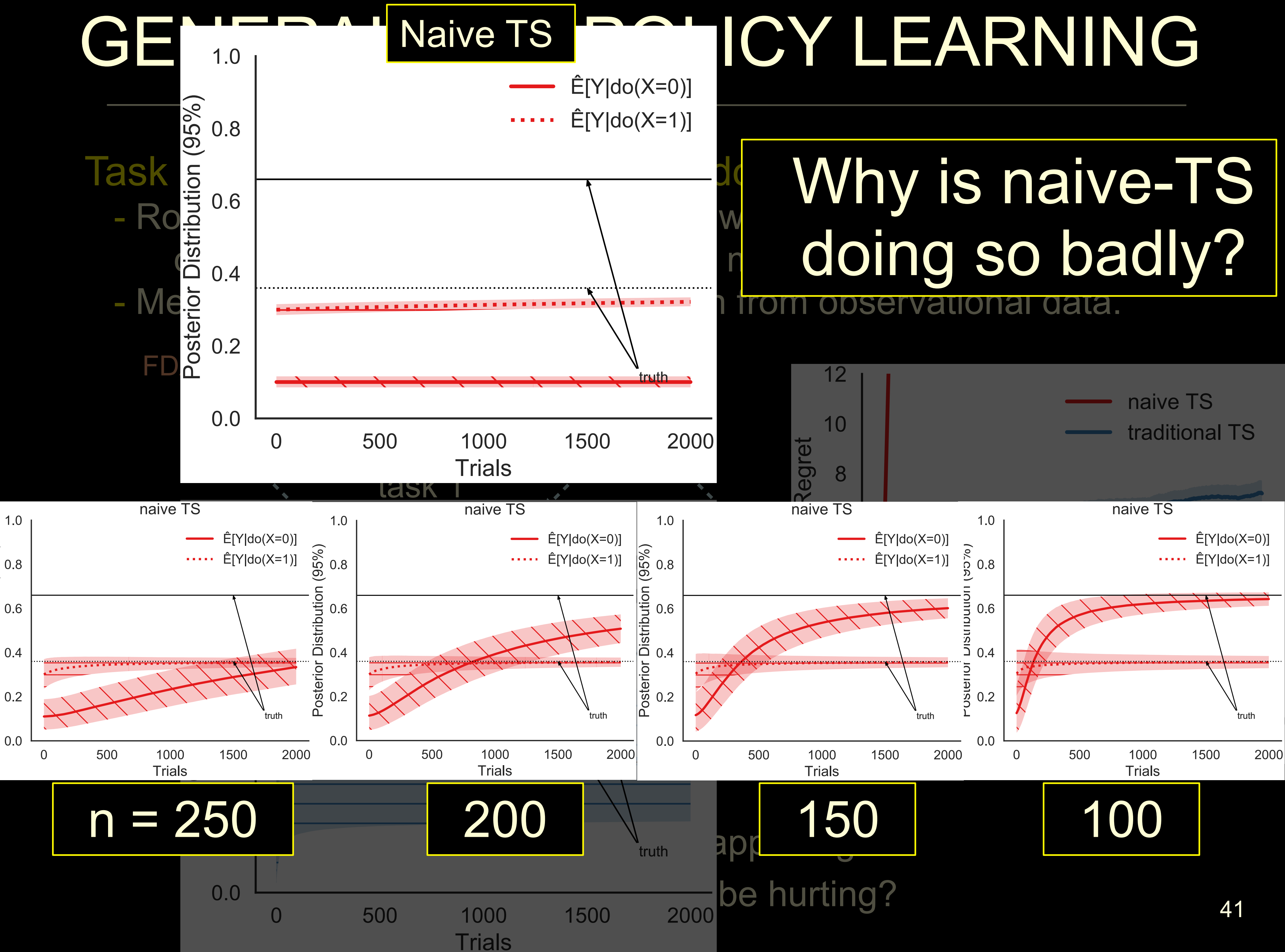
Traditional TS





How could this be happening?!

Could more data be hurting?


GEN ICY LEARNING
Naive TS
Why is naive-TS doing so badly?
Posterior Distribution (95%)
Ê[Y|do(X=0)]
Ê[Y|do(X=1)]
truth
Trials
naive TS
traditional TS
Regret
naive TS
n = 250
200
150
100
from observational data.
be hurting?

Structural Explanation for Naive-TS's behavior -- The Challenge of Non-Identifiability

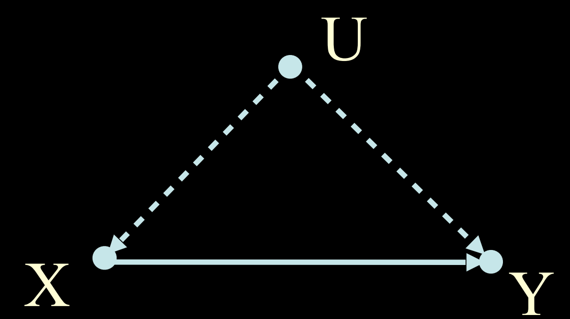
- SCM M (Unobserved)

X=U

P(U=0)=0.3

P[Y|X,U]	U=0	U=1
X=0	0.1	0.9
X=1	0.5	0.3

- Causal Graph G



- Distributions

$E(Y \mid X = 0) < E(Y \mid X = 1)$

$E(Y \mid do(X = 0)) > E(Y \mid do(X = 1))$

	E[Y|do(X)]	E[Y|X]
X=0	0.66	0.1
X=1	0.36	0.3
	L_2	L_1

Structural Explanation for Naive-TS's behavior -- The Challenge of Non-Identifiability

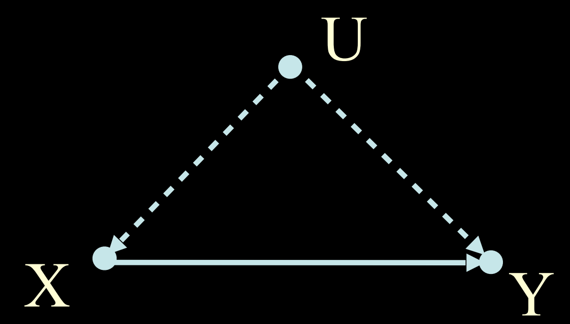
- SCM M (Unobserved)

X=U

P(U=0)=0.3

P[Y|X,U]	U=0	U=1
X=0	0.1	0.9
X=1	0.5	0.3

- Causal Graph G



- Distributions

E(Y | X = 0) < E(Y | X = 1)
E(Y | do(X = 0)) > E(Y | do(X = 1))

	E[Y|do(X)]	E[Y|X]
X=0	0.66	0.1
X=1	0.36	0.3
	L_2	L_1

Structural Explanation for Naive-TS's behavior -- The Challenge of Non-Identifiability

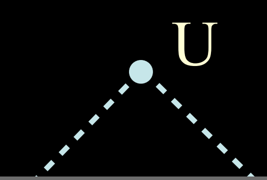
- SCM M (Unobserved)

X=U

P(U=0)=0.3

P[Y|X,U]	U=0	U=1
X=0	0.1	0.9
X=1		

- Causal Graph G



X=1 is looking quite good, should I do() it?

- Distributions

$E(Y \mid X = 0) < E(Y \mid X = 1)$

$E(Y \mid do(X = 0)) > E(Y \mid do(X = 1))$

	E[Y|do(X)]	E[Y|X]
X=0	0.66	0.1
X=1	0.36	0.3
	L_2	L_1

Structural Explanation for Naive-TS's behavior -- The Challenge of Non-Identifiability

- SCM M (Unobserved)
- Causal Graph G

Questions (more general):

1. How do I know this pattern is not present in my data? Don't know :(
2. Does this then imply that I should throw away all the data not collected by me (the agent) and learn from scratch? Hopefully not…
3. After all, is there any useful information in the obs. data? Yes!

Let's try to understand how to leverage confounded data…

	L_2	L_1
X=0	0.66	0.1
X=1	0.36	0.39

Step 1. Extracting Causal Information from Confounded Observations

Solution: Bounding $E[Y \mid do(x)]$ from observations $P(x,y)$.

Theorem. Given observations coming from any distribution $P(x,y)$, the average causal effect $E[Y \mid do(x)]$ is bounded in $[l_x, h_x]$, where

$$l_x = E[Y \mid x]\, P(x) \quad \text{and} \quad h_x = l_x + 1 - P(x).$$

- Linear Program formulation in other causal graphs (non-parametric SCMs): [Balke & Pearl, 1996; Zhang and Bareinboim, IJCAI'17]
- Incorporating parametric knowledge: [Kallus & Zhou, 2018; Namkoong et al., 2020]
- Sequential treatments in longitudinal settings: [Zhang & Bareinboim, NeurIPS'19; ICML'20]

Step 2. Incorporating Bounds into Learning (e.g., Causal Thompson Sampling)

Input: prior parameters $\boldsymbol{\alpha}$, β,
causal bounds $[l_x, h_x]$ for each arm x.

/* $[l_x, h_x]$ are computed from confounded observations */

Initialization: $S_x=0$, $F_x=0$ for each arm x

For t = 1, …, T do
For each x do
Repeat
Draw $\theta_x \sim Beta(S_x+\boldsymbol{\alpha}, F_x+\beta)$.
Until $\theta_x \in [l_x, h_x]$
End
Play $do(x_t)$ where $X_t = \text{argmax}_x\ \theta_x$.
Observed Y_t and update F_{xt} and S_{xt}.
End

/* Causal bounds are ascertained thr. a rejection procedure. */

GE… …ICY LEARNING

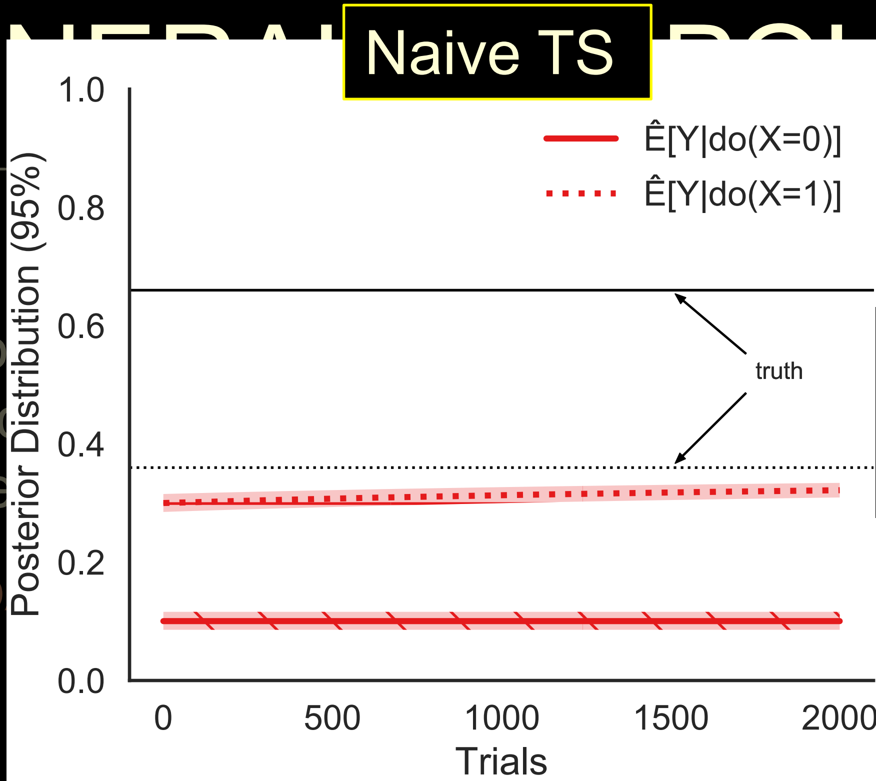
Task … do(x)).

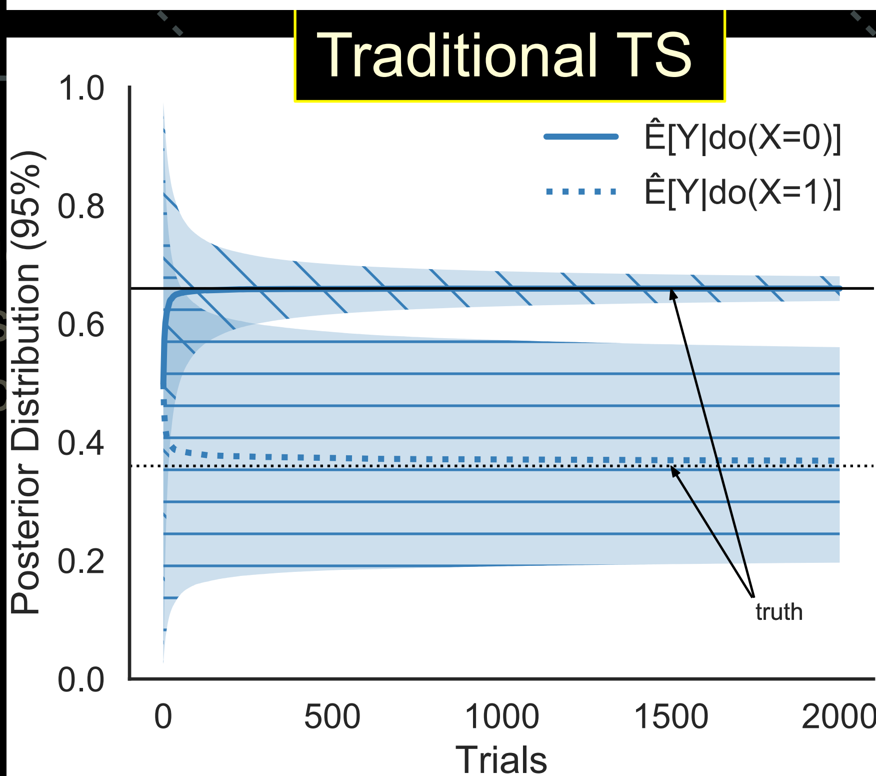
- do-calc ID

- Ro…

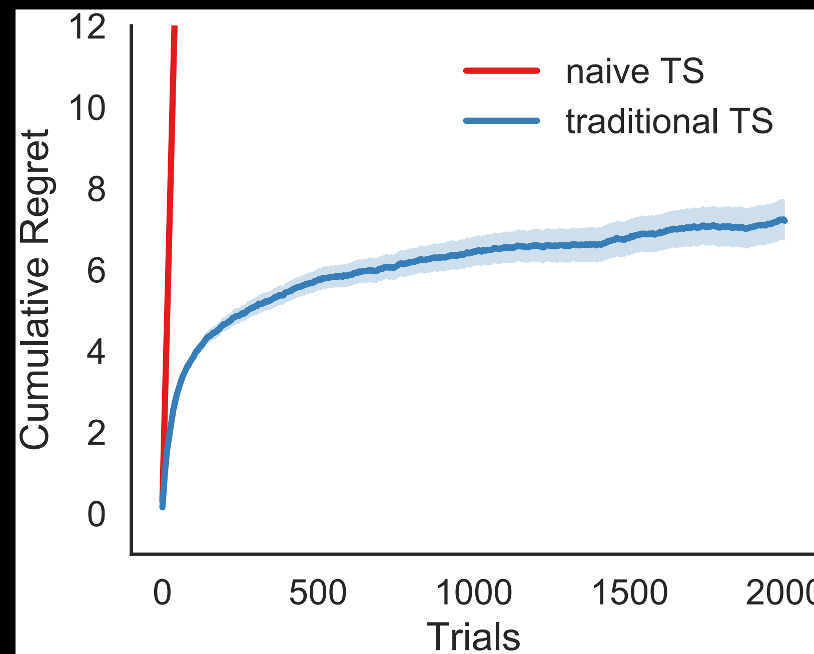
- Me…

FD…





Can we do better using the causal bounds?



X … Y

Let's i… …g
that s… …e
interc…

…appening?!

…be hurting?

GE ICY LEARNING

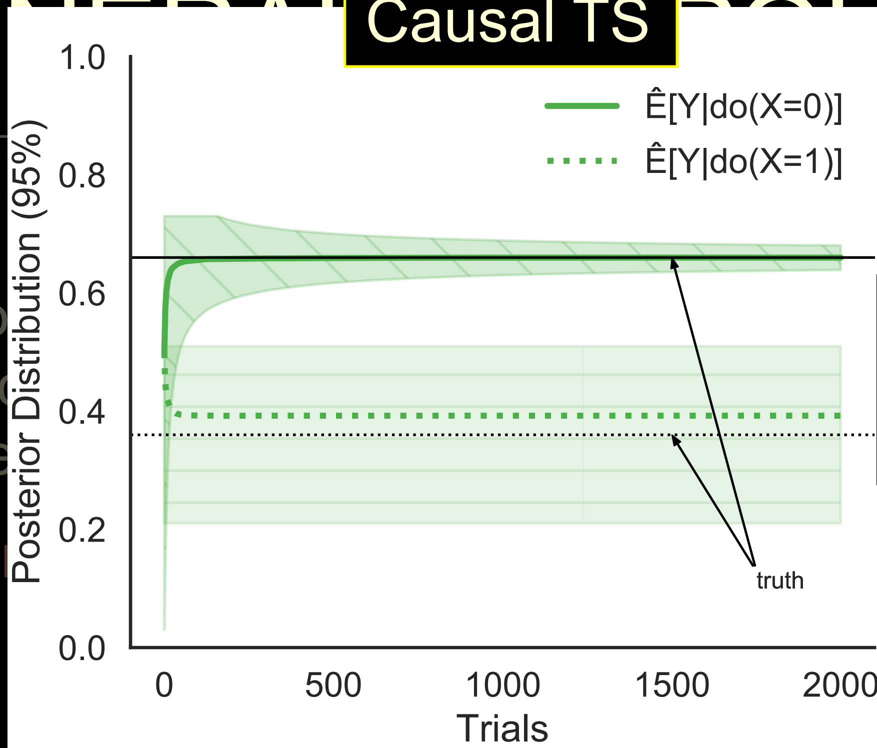
Task do(x)).

- do-calc ID

- Ro

- Me

Causal TS

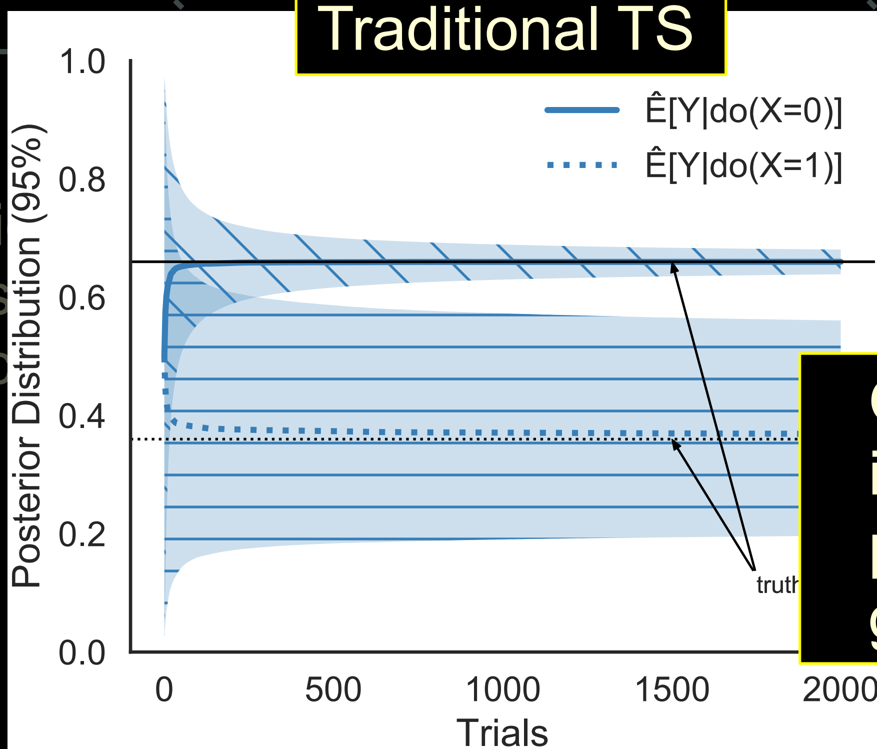


Can we do better using the causal bounds?

X

Y

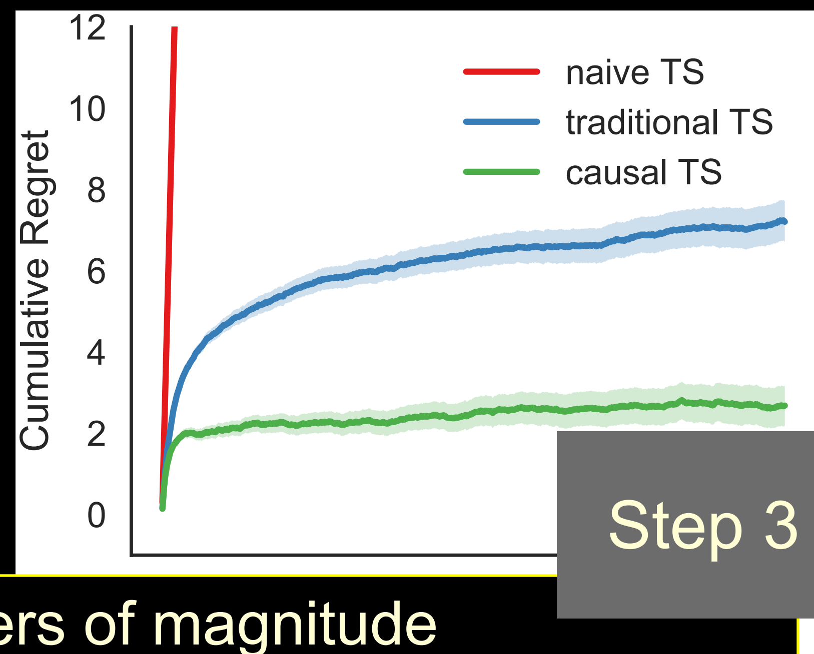
Traditional TS



Let's i

that s

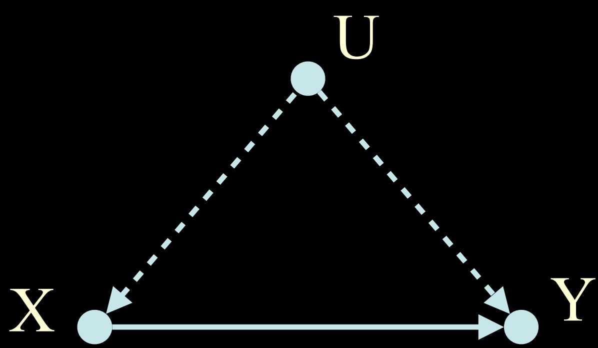
interc



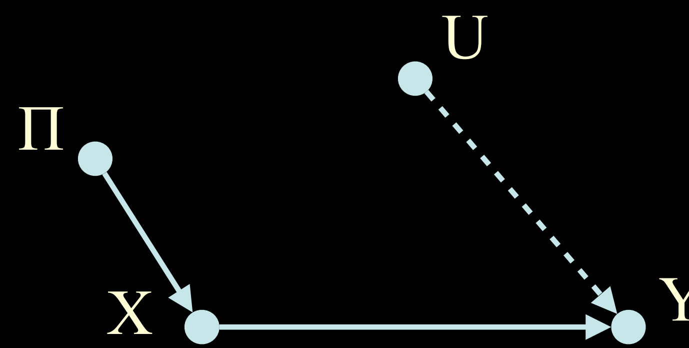
Step 3

Orders of magnitude improvement can be achieved in practice, and can be proved in general settings (ZB, IJCAI'17).

GENERALIZED POLICY LEARNING -- BIG PICTURE

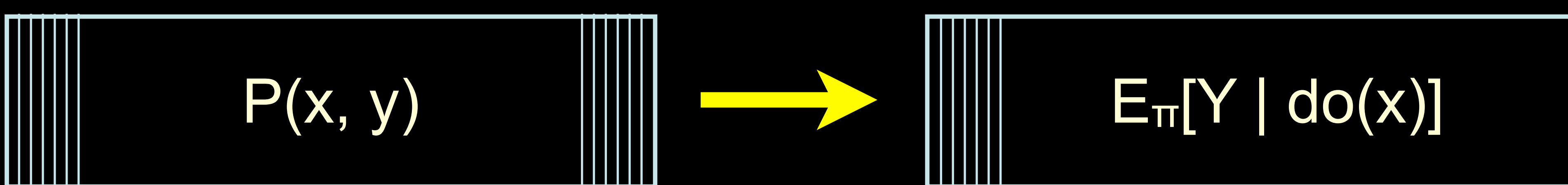


Causal graph G



under do(X)

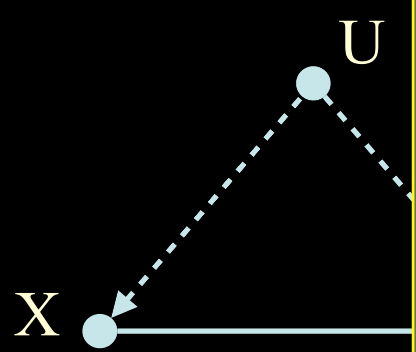
Observation GPL-bounding Agent



obs, obs … obs

$do(x_0)$ $do(x_1)$ … $do(x_0)$

GENERALIZED POLICY LEARNING -- BIG PICTURE



Causal gra

Observatio

P(x, y)

obs, obs … obs

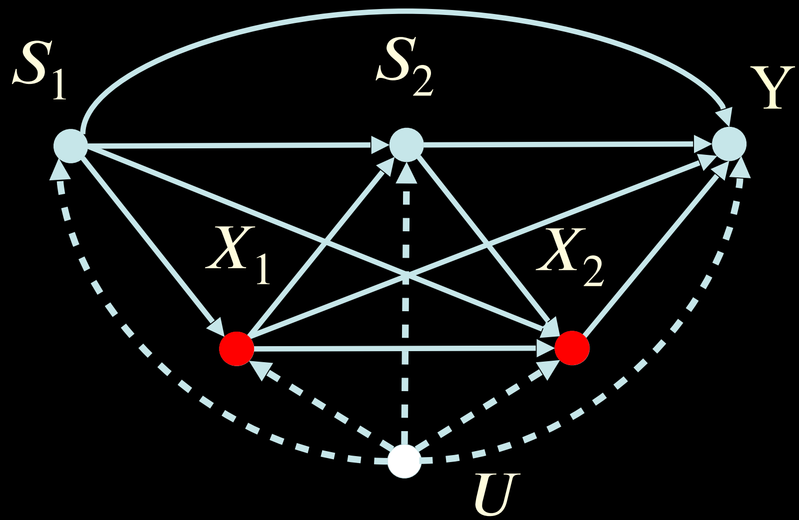
do(x_0) do(x_1) … do(x_0)

SUMMARY (GPL Template):

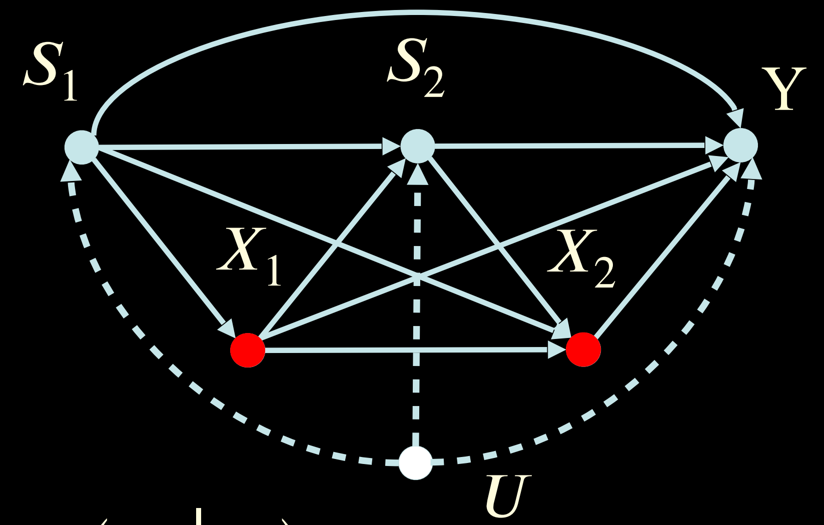
1. If policy is *identifiable* from offline methods, return optimal one through Do-calculus/IPW.
2. Extract causal information from obs. data, and compose causal bounds based on the available structural assumptions (on G & M).
3. Offline + Online: Incorporate causal bounds into online allocation procedure.
4. Prove regret bounds (Theory).

NEW RESULT: GPL FOR DYNAMIC TREATMENT REGIMES

- DTRs is a popular model for sequential treatment in medical domains [Murphy, 2003; Moodie et al., 2007]:



Observational

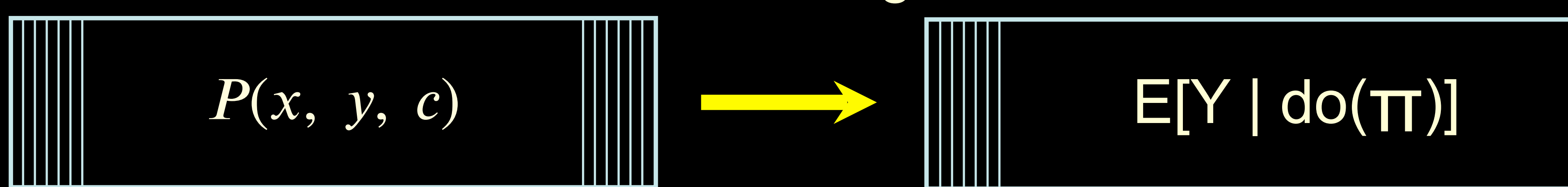


$$do\left(x_1 \,\middle|\, s_1\right),\ do(x_2 \mid x_1, s_1, s_2)$$

Other agent with π

GPL-bounding

Agent

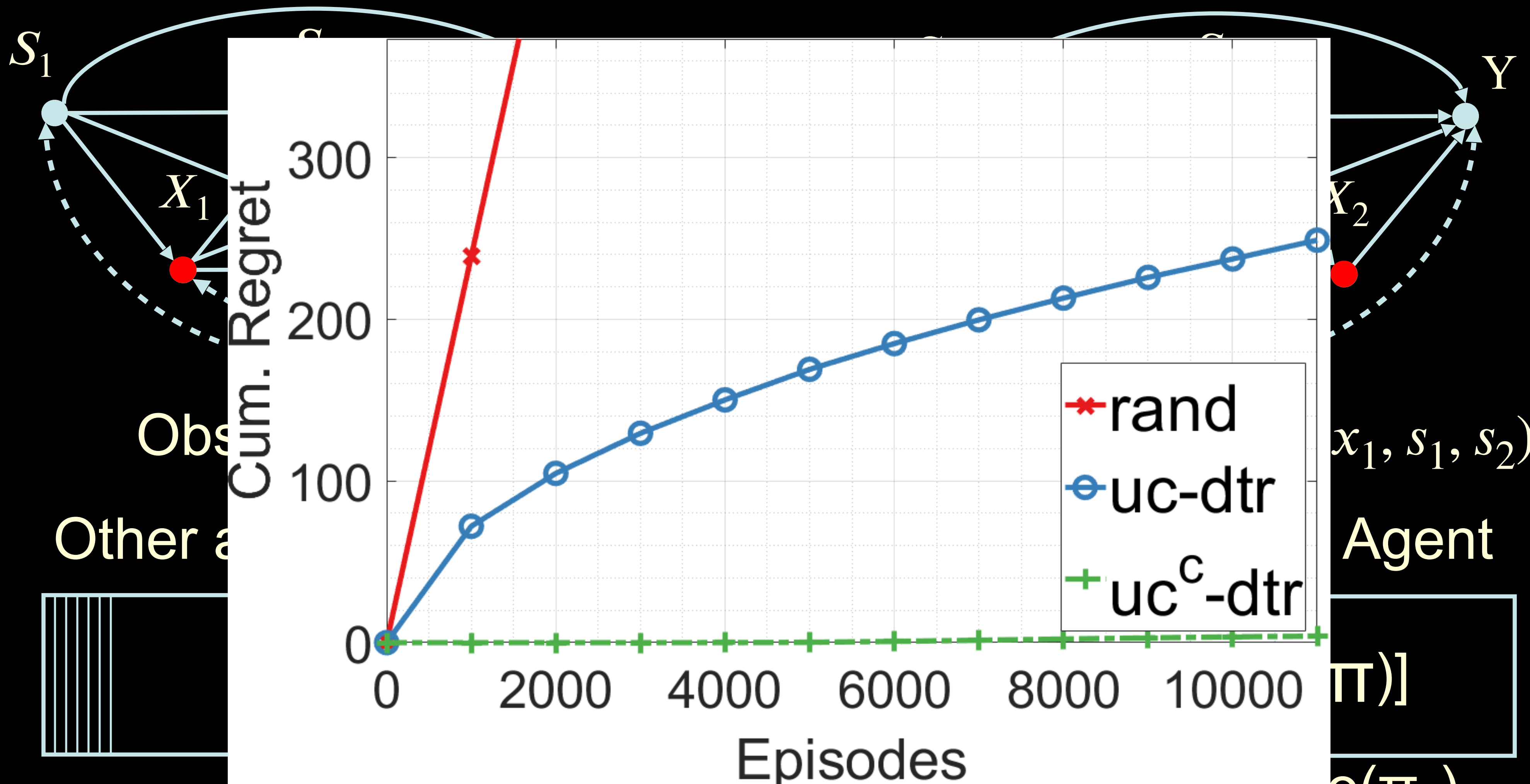


obs, obs, … obs

do(π_0) … do(π_1)

NEW RESULT: GPL FOR DYNAMIC TREATMENT REGIMES

- DTRs is a popular model for sequential treatment in medical domains [Murphy, 2003; Moodie et al., 2007]:



Obs … $x_1, s_1, s_2)$

Other a … Agent

… π)]

obs, obs, … obs … do(π_0) … do(π_1)

* For details, see [Zhang & Bareinboim, NeurIPS'19; ICML'20].

TASK 2.
WHEN AND WHERE TO INTERVENE?
(Refining the policy space)

Sanghack Lee

CRL-TASK 2.
WHEN AND WHERE TO INTERVENE?

- In general, it's assumed throughout the literature a policy space such that actions are fixed a priori (e.g., a set $X = \{X_1, \ldots, X_k\}$), and intervening is usually assumed to lead to positive outcomes.

- Our goal here is to understand when interventions are required, or if they may lead to unintended consequences (e.g., side effects).

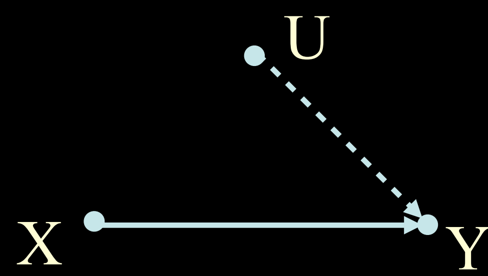
when / if

- In the case interventions may be needed, we would like to understand what should be changed in the underlying environment so as to bring a desired state of affairs about (e.g., maybe $do(X_1, X_3, X_7)$ instead of $do(X_1, X_2, X_3, \ldots, X_7)$).

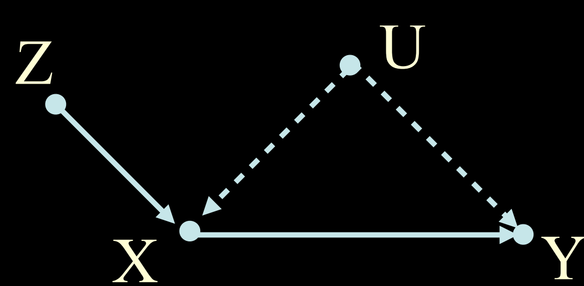
where

UNDERSTANDING THE POLICY SPACE

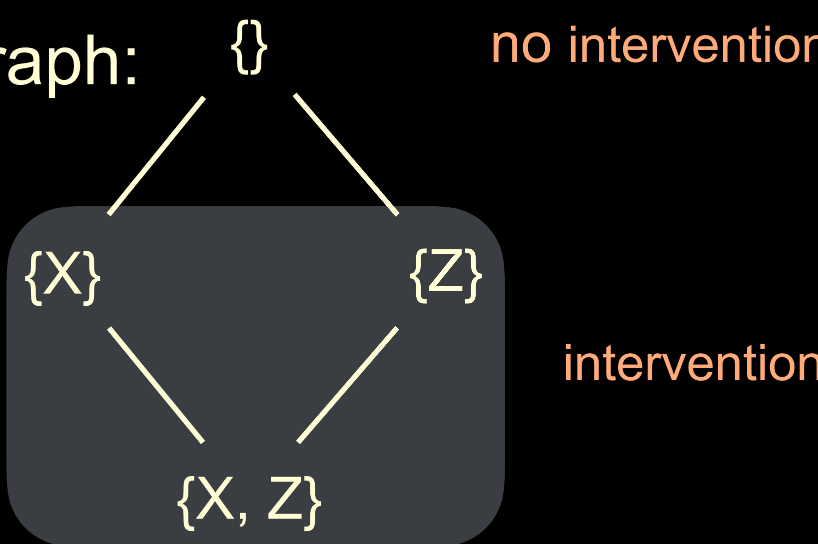
- Consider the causal graph of a bandit model:



- Our goal is to optimize Y (e.g., keep it high as much as possible), and we are not a priori committed to intervening on any specific variable, or intervening at all.
- Consider now the 3-var causal graph:



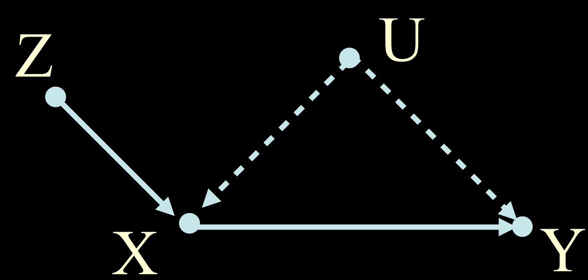
causal graph G



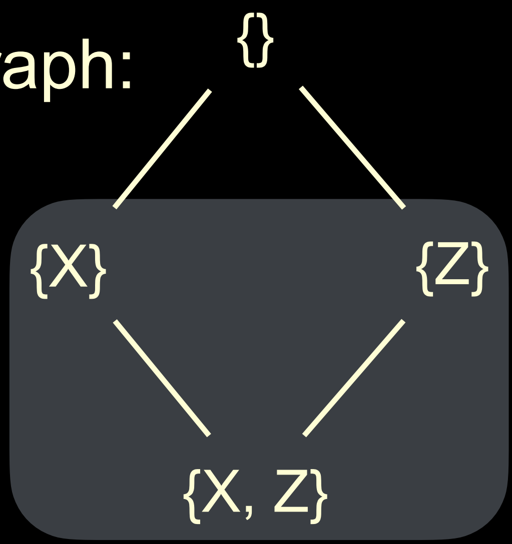
policy space

UNDERSTANDING THE POLICY SPACE

- Our goal is to optimize Y (e.g., keep it high as much as possible), and we are not a priori committed to intervening on any specific variable, or intervening at all.
- Consider now the 3-var causal graph:



causal graph G

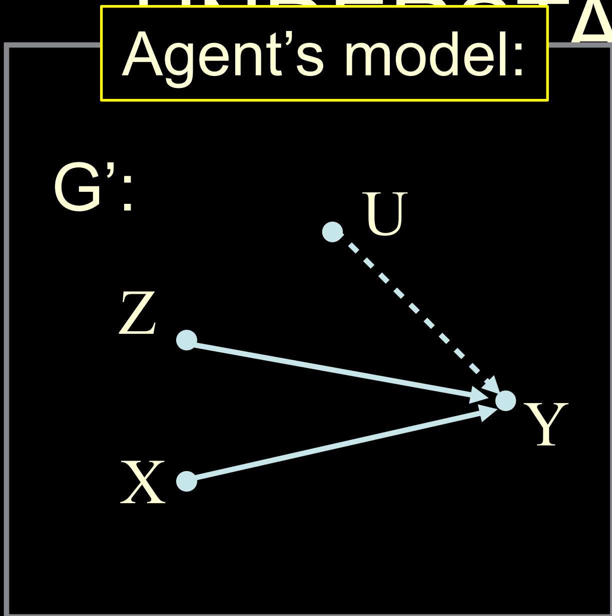


policy space

- Causal-insensitive strategy: Ignore the causal structure G, take {X, Z} as one larger variable, and search based on

$$\mathrm{argmax}_{xz}\, E[Y \mid do(X = x, Z = z)]$$

UNDERSTANDING THE POLICY SPACE

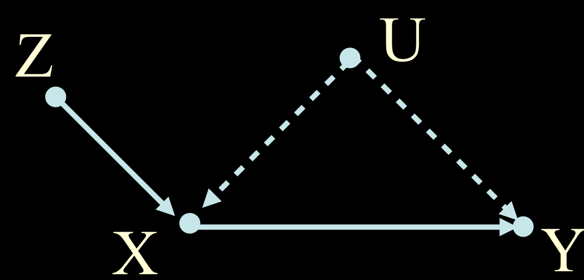


optimize Y (e

we are not a

variable, or

the 3-var cau



causal graph G

- Causal-insensitive strateg
  take {X, Z} as one larger v

$$\text{argmax}_{xz}\, E[Y \mid do(X = x, Z = z)]$$

Key observations:

1. Note that the implicit causal graph in the agent's mind (*G'*), which follows from standard optimization procedure, is different than *G*.

2. The true causal model *G* encodes constraints of the underlying environment (SCM M).

Question -- Despite what is in the agent's mind (or optimization function), it's still the case that it will be evaluated by the SCM M. Is then being oblivious to the pair <G, M> okay? Can't we just do more interventions?

Meaning, more do(X=x, Z=z), and things will eventually converge?

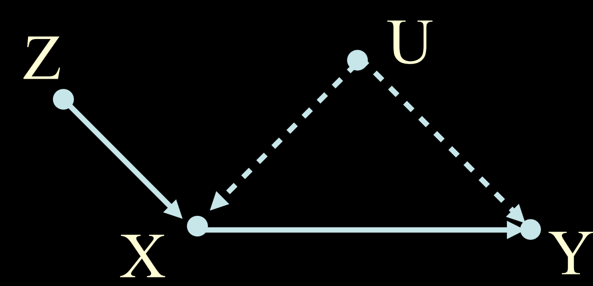
THE CAUSAL STRUCTURE CANNOT BE DISMISSED

- SCM M (Unobserved)

$$Z \leftarrow U_z$$
$$X \leftarrow Z \oplus U$$
$$Y \leftarrow X \oplus U$$

$$P(U=1) = P(U_z=1) = 0.5$$

- Causal Graph G



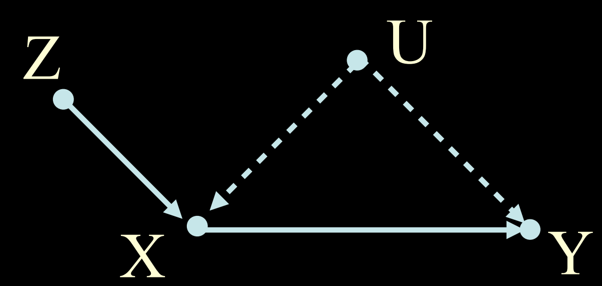
THE CAUSAL STRUCTURE CANNOT BE DISMISSED

- SCM M (Unobserved)

$$Z \leftarrow U_z$$
$$X \leftarrow Z \oplus U$$
$$Y \leftarrow X \oplus U$$

$$P(U=1) = P(U_z=1) = 0.5$$

- Causal Graph G



$$E[Y|\, do(X)] = E[Y|\, do(X,Z)] = 0.5$$

$$E[Y|\, do(Z)] = (Z \oplus U) \oplus U = Z$$

So, if do(Z=1),

$$E[Y \mid do(Z = 1)] = 1$$

THE CAUSAL STRUCTURE CANNOT BE DISMISSED

- SCM M (Unobserved)

$Z \leftarrow U_z$

X

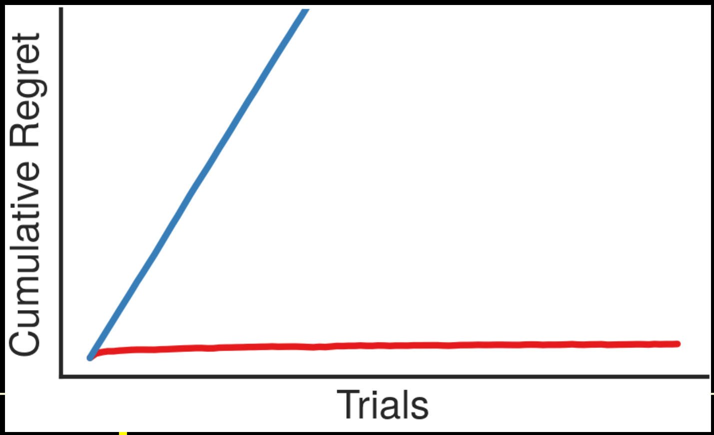
Y

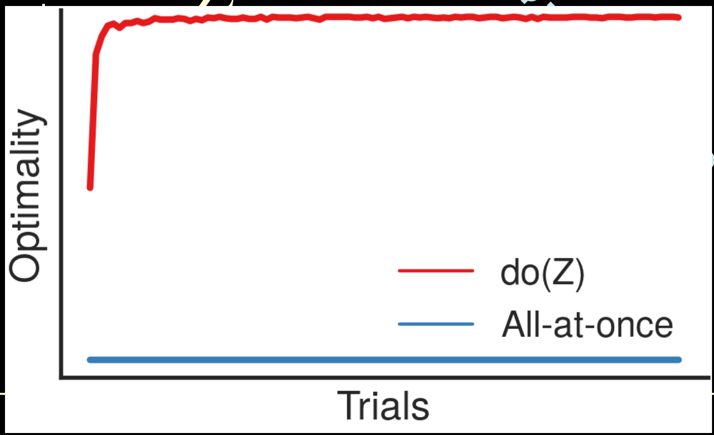
P(

- Causal Graph G

Z U

Y





E[Y|

E[Y|

- A causal insensitive strategy (i.e., "all-at-once", do(X,Z)) will not pick up the do(Z)-intervention, and will never converge!
- A naive, "all-subsets" strategy works since it includes do(Z=1)

So, if do(Z=1),

$E[Y \mid do(Z = 1)] = 1$

THE CAUSAL STRUCTURE CANNOT BE DISMISSED

- SCM M (Unobserved)

$Z \leftarrow U_Z$

X

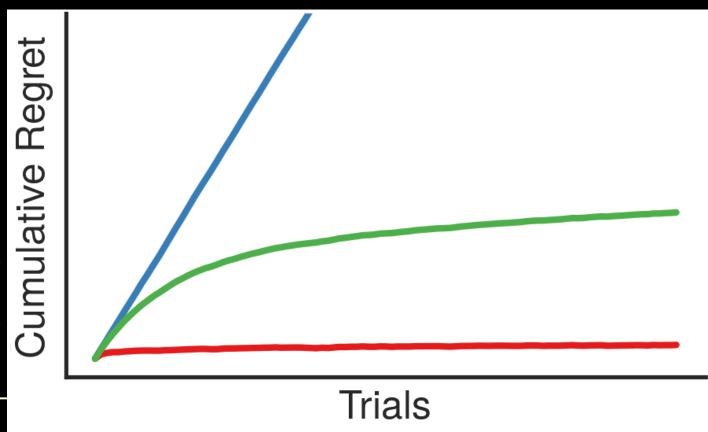
Y

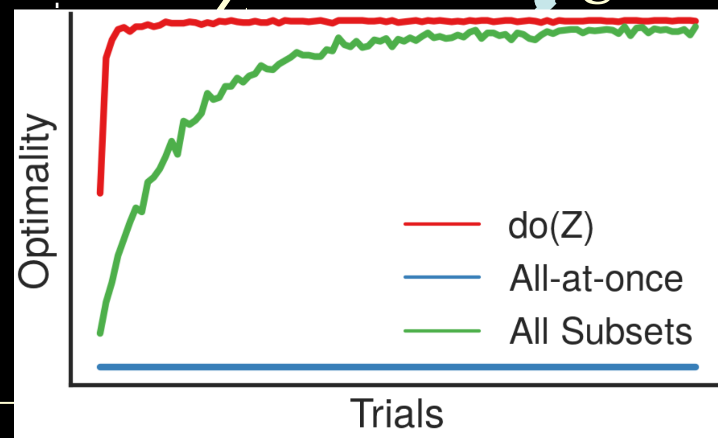
P(

- Causal Graph G

Z U

Y





E[Y|

E[Y|

- A causal insensitive strategy (i.e., “all-at-once”, do(X,Z)) will not pick up the do(Z)-intervention, and will never converge!
- A naive, “all-subsets” strategy works since it includes do(Z=1)

So, if do(Z=1),

$E[Y \mid do(Z = 1)] = 1$

THE CAUSAL STRUCTURE CANNOT BE DISMISSED

- SCM M (Unobserved)

$Z \leftarrow U_z$

X

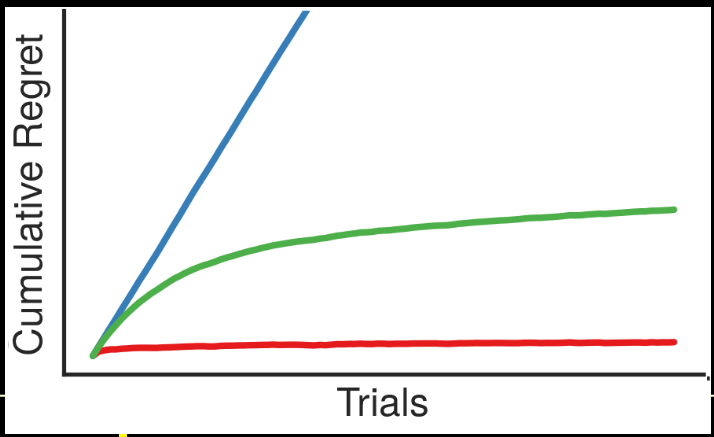
Y

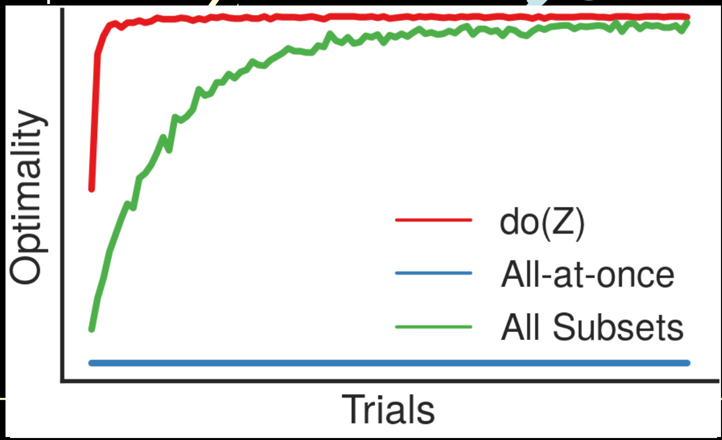
P(

- Causal Graph G

Z U

Y





E[Y|

E[Y|

So,

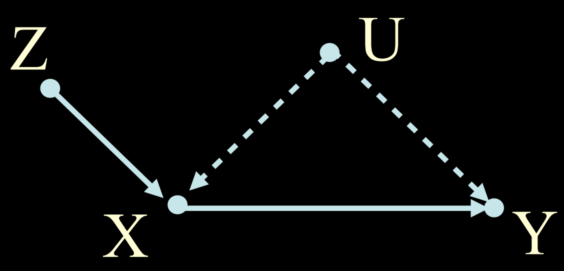
E[Y

- A causal insensitive strategy (i.e., "all-at-once", do(X,Z)) will not pick up the do(Z)-intervention, and will never converge!
- A naive, "all-subsets" strategy works since it includes do(Z=1)

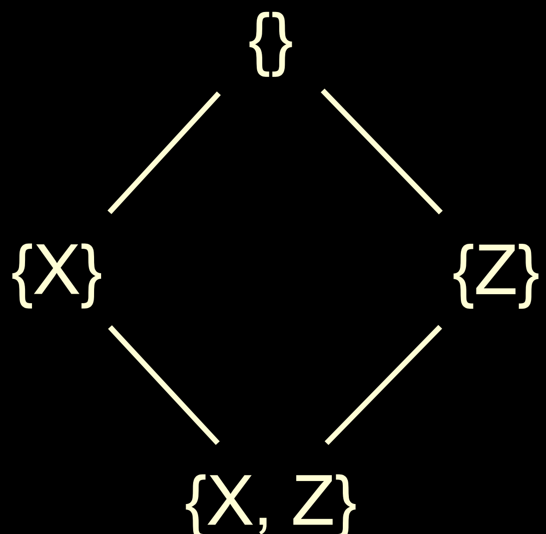
Can we do better than these two naive strategies?

POLICY SPACE (EXAMPLE)

Causal graph G



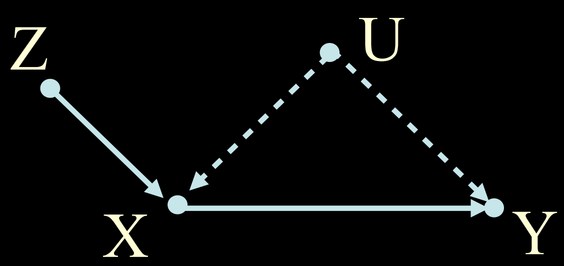
Policy space



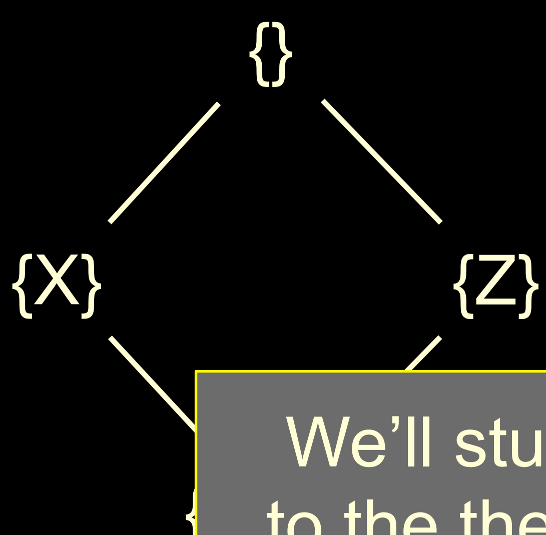
Intervention Sets (IS)	**Actions**
{}	do()
do(X)	do(X=0) do(X=1)
do(Z)	do(Z=0) do(Z=1)
do(X,Z)	do(X=0,Z=0) do(X=0,Z=1) do(X=1,Z=0) do(X=1,Z=1)

POLICY SPACE (EXAMPLE)

Causal graph G



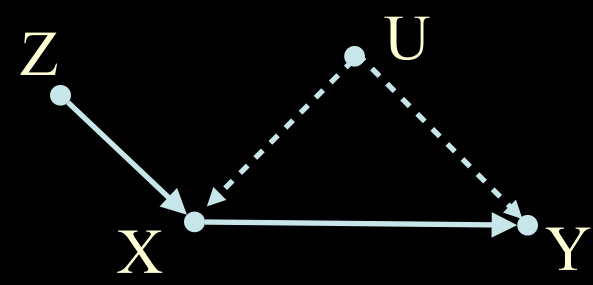
Policy space



Intervention Sets (IS)	Actions
{}	do()
do(X)	do(X=0) do(X=1)
do(Z)	do(Z=0) do(Z=1)
do(X,Z)	do(X=0,Z=0)

We'll study properties of the policy space with respect to the the topological constraints imposed by M in G.

PROPERTY 1: INTERVENTIONAL EQUIVALENCE



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

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

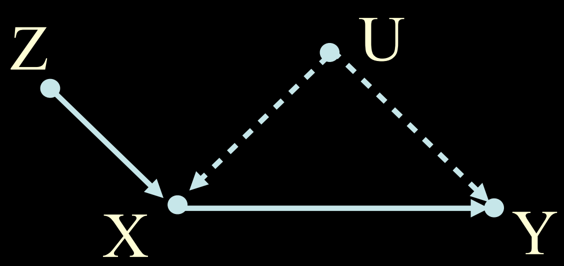
Implication: prefer playing do(X) to playing do(X, Z).

Definition (Minimal Intervention Set, MIS)

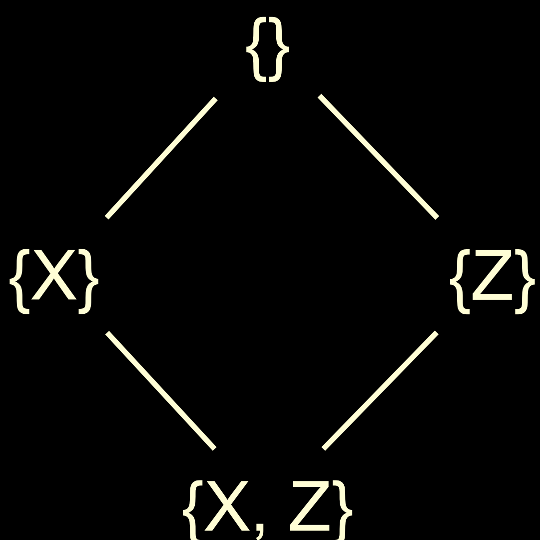
Given $\langle G, Y \rangle$, a set of variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$ is said to be a *minimal intervention set* if there is no $\mathbf{X}' \subset \mathbf{X}$ such that $E[Y \mid do(\mathbf{x}')] = E[Y \mid do(\mathbf{x})]$ for every SCM conforming to G where $\mathbf{x}'$ is consistent with $\mathbf{x}$.

PROPERTY 1: MIS (EXAMPLE)

Causal graph G



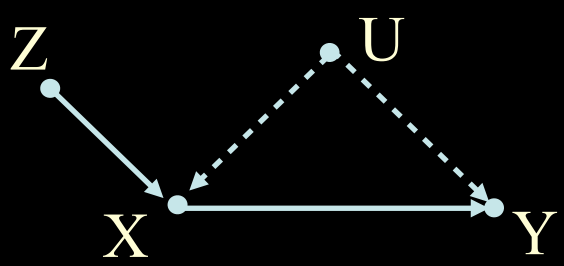
Policy space



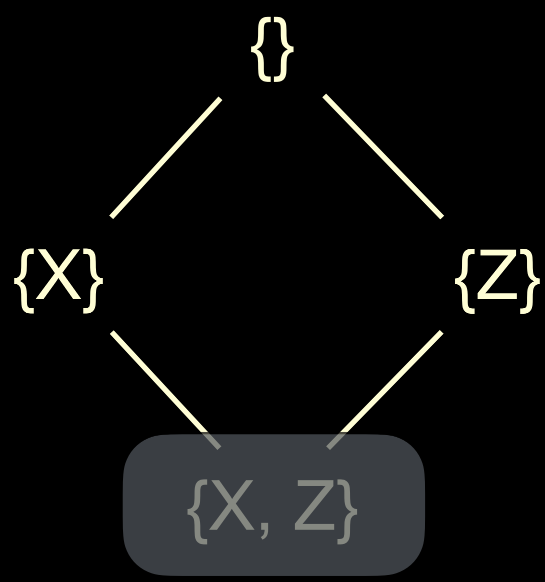
Intervention Sets (IS)	**Actions**	**MIS**
{}	do()	
do(X)	do(X=0) do(X=1)	
do(Z)	do(Z=0) do(Z=1)	
do(X,Z)	do(X=0,Z=0) do(X=0,Z=1) do(X=1,Z=0) do(X=1,Z=1)	

PROPERTY 1: MIS (EXAMPLE)

Causal graph G

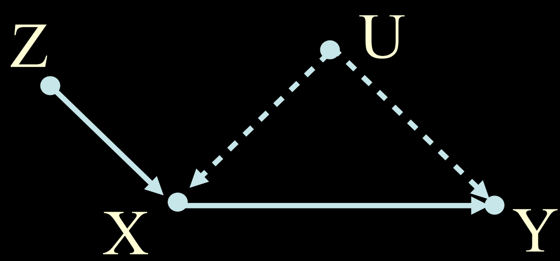


Policy space



Intervention Sets (IS)	Actions	MIS
{}	do()	✔
do(X)	do(X=0) do(X=1)	✔
do(Z)	do(Z=0) do(Z=1)	✔
do(X,Z)	do(X=0,Z=0) do(X=0,Z=1) do(X=1,Z=0) do(X=1,Z=1)	✗

PROPERTY 2: PARTIAL-ORDEREDNESS



$$\begin{aligned} E[Y] &= \sum_z E[Y|do(z)]\, P(z) \\ &\leq \sum_z E[Y|do(z^*)]\, P(z) \\ &= E[Y|do(z^*)] \end{aligned}$$

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

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

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

Definition (Possibly-Optimal MIS, POMIS)

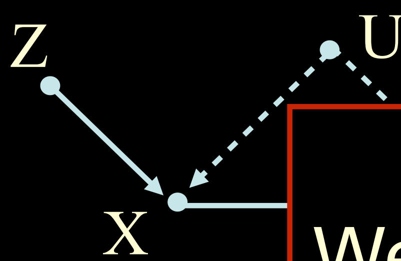
Given <*G*,Y>, let $\mathbf{X} \in$MISs. $\mathbf{X}$ is said to be a possibly-optimal MIS if there exists a SCM M conforming to *G* such that

$$\max_{\mathbf{x}} E[Y \mid do(\mathbf{X}=\mathbf{x})] > \max_{\mathbf{W} \in \text{MIS} \setminus \{\mathbf{X}\}} E[Y \mid do(\mathbf{W}=\mathbf{w})]$$

PROPERTY 2: PARTIAL-ORDEREDNESS

$$E[Y] = \sum_z E[Y|do(z)]\, P(z)$$

$$\leq \sum_z E[Y|do(z^*)]\, P(z)$$



E[Y|do(z)]

We provide a complete characterization of POMIS & algorithm that enumerates all POMISs given a causal graph G.

Implica ng do().

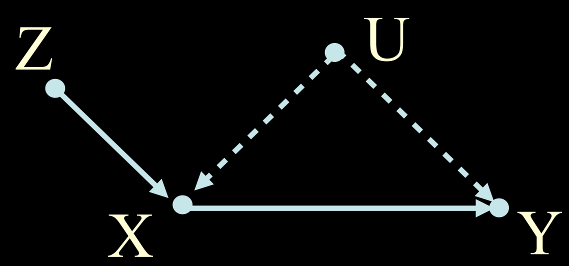
Definition (Possibly-Optimal MIS, POMIS)

Given <*G*,Y>, let $\mathbf{X} \in$ MISs. $\mathbf{X}$ is said to be a possibly-optimal MIS if there exists a SCM M conforming to *G* such that

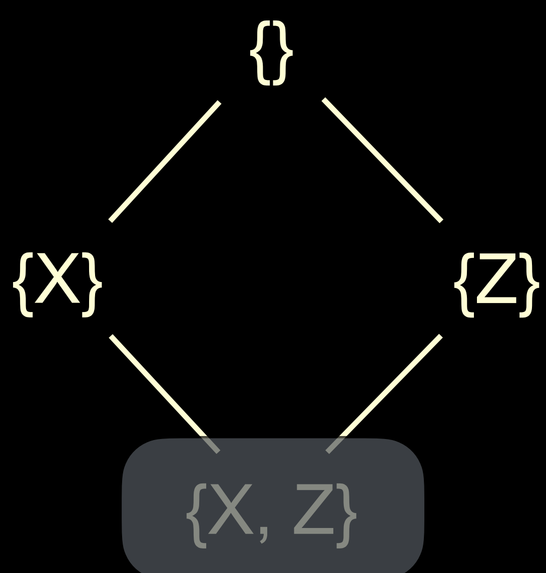
$$\max_{\mathbf{x}} E[Y \mid do(\mathbf{X}=\mathbf{x})] > \max_{\mathbf{W} \in \text{MIS} \setminus \{\mathbf{X}\}} E[Y \mid do(\mathbf{W}=\mathbf{w})]$$

PROPERTY 2: POMIS (EXAMPLE)

Causal graph G



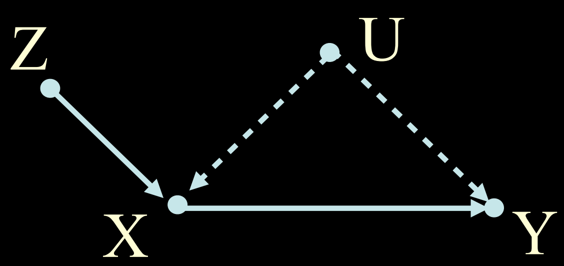
Policy space



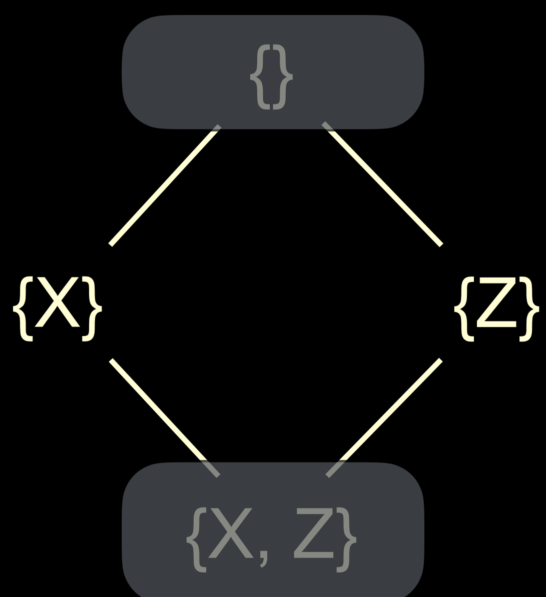
intervention sets	actions	MIS	POMIS
{}	do()	✔	
do(X)	do(X=0) do(X=1)	✔	
do(Z)	do(Z=0) do(Z=1)	✔	
do(X,Z)	do(X=0,Z=0) do(X=0,Z=1) do(X=1,Z=0) do(X=1,Z=1)	✗	

PROPERTY 2: POMIS (EXAMPLE)

Causal graph G



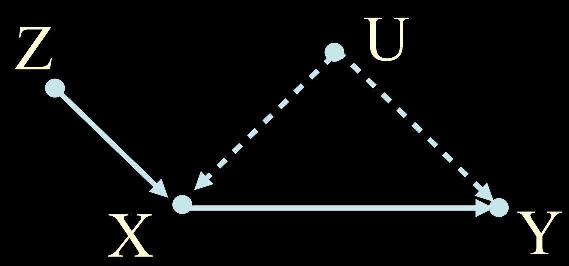
Policy space



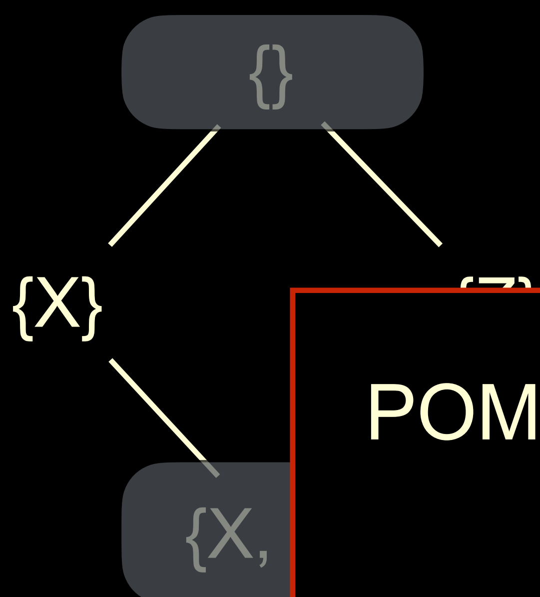
intervention sets	actions	MIS	POMIS
{}	do()	✔	✗
do(X)	do(X=0) do(X=1)	✔	✔
do(Z)	do(Z=0) do(Z=1)	✔	✔
do(X,Z)	do(X=0,Z=0) do(X=0,Z=1) do(X=1,Z=0) do(X=1,Z=1)	✗	✗

PROPERTY 2: POMIS (EXAMPLE)

Causal graph G



Policy space



intervention sets	actions	MIS	POMIS
{}	do()	✔	✗
do(X)	do(X=0) do(X=1)	✔	✔
do(Z)	do(Z=0) do(Z=1)	✔	✔
do(X,Z)	do(X=0,Z=0)		

POMIS share the reward mechanism (SCM) & POMIS' arms are dependent.

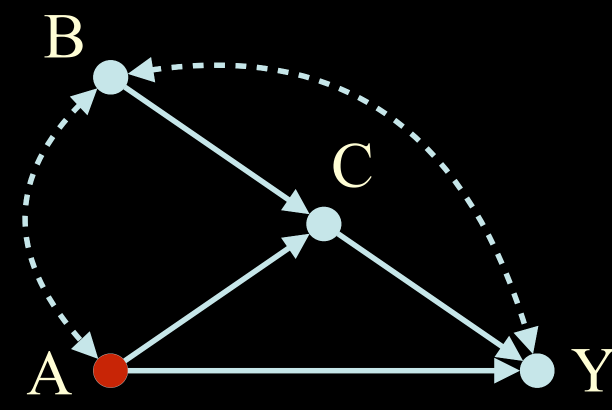
PROPERTY 3: ARMS' QUANTITATIVE RELATIONSHIPS

- Goal: infer an arm's expected reward from other arms' data,

$$P(y|do(\mathbf{x})) \leftarrow \{ P(\mathbf{V} \mid do(\mathbf{Z})) \}_{Z \in \mathbf{POMIS} \setminus \{\mathbf{X}\}}$$

- New ID algorithm (z²ID) to find a matching POMIS, that can borrow some additional data.

- Example



Given POMISs {}, {B}, and {C}:

$$P(y) = \sum_{a,b,c} P_b(c|a) P_c(a, b, y)$$

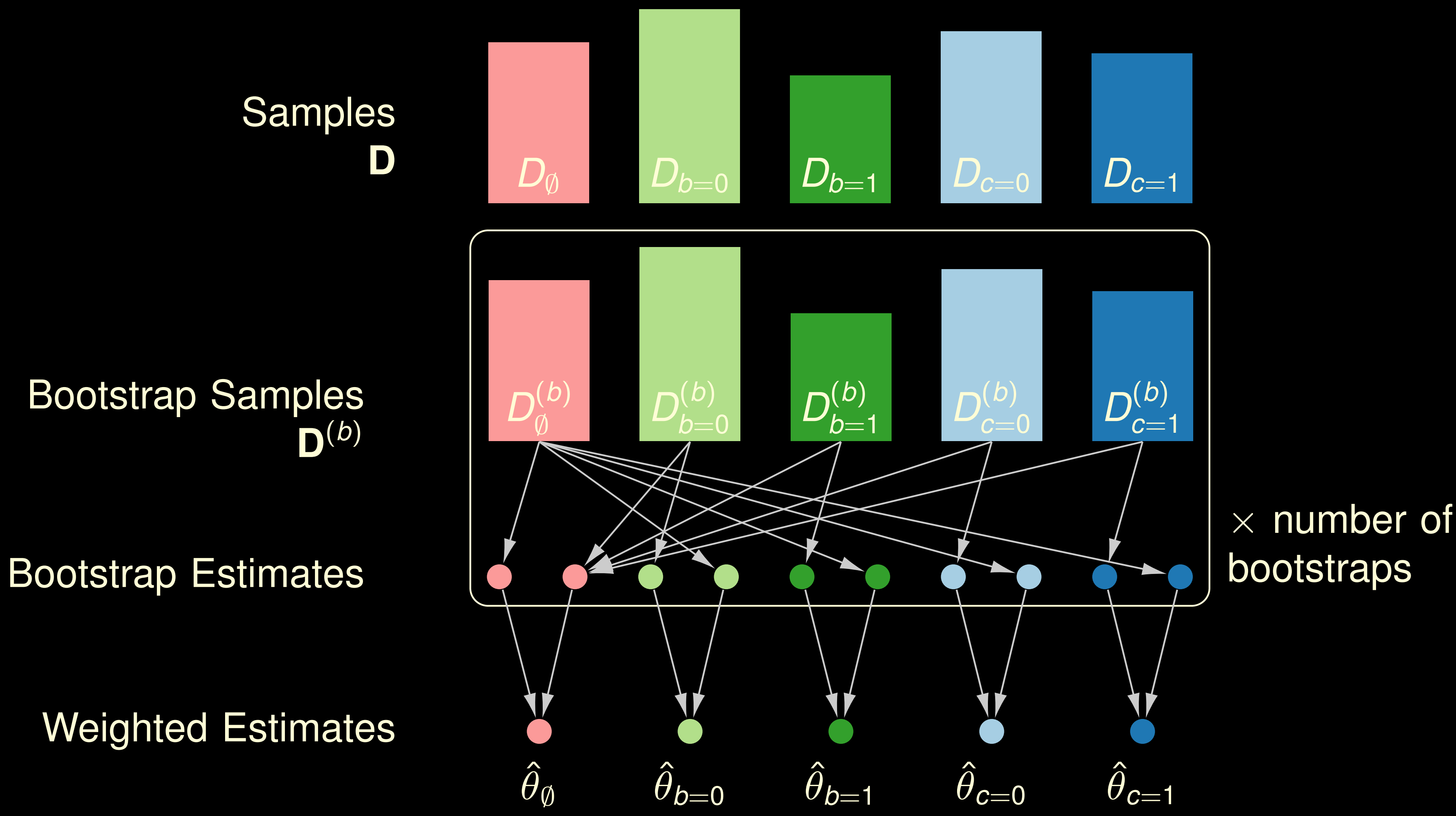
$$P_b(y) = \sum_{a,c} P(c|a, b) \sum_{b'} P(y|a, b', c) P(a, b')$$

$$P_c(y) = \sum_{a,b} P(y|a, b, c) P(a, b)$$

$$P_c(y) = \sum_{a} P_b(y|a, c) P_b(a)$$

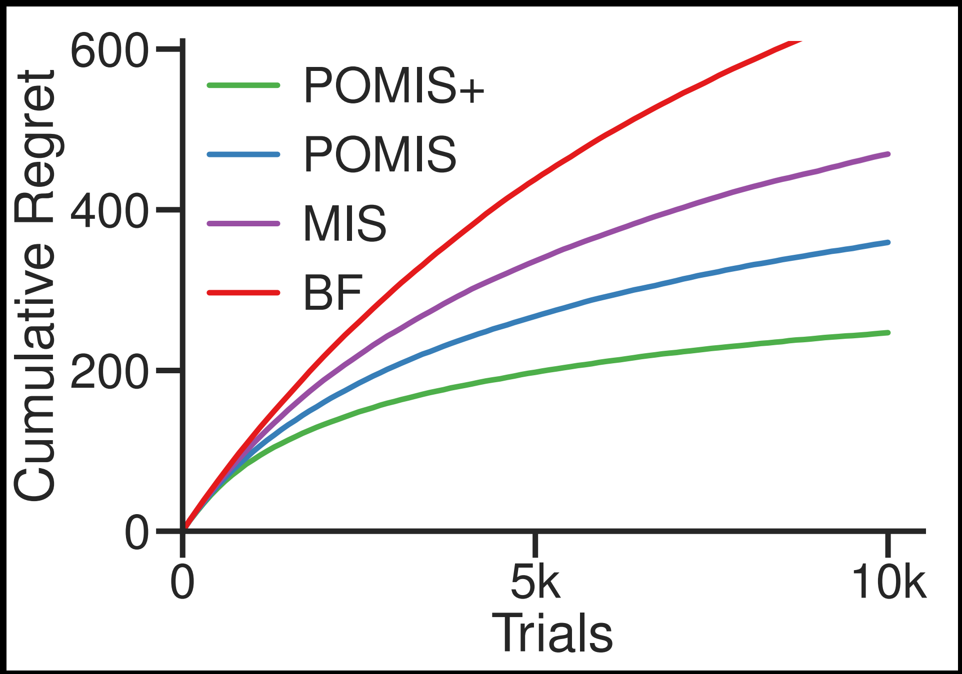
PROPERTY 3: ARMS' QUANTITATIVE RELATIONSHIPS

- Make the most of data — Minimum Variance Weighting

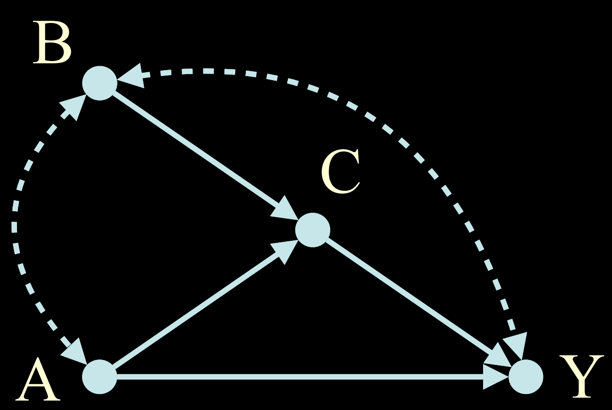


WHEN AND WHERE TO INTERVENE -- ALGORITHMS & EXPERIMENTS

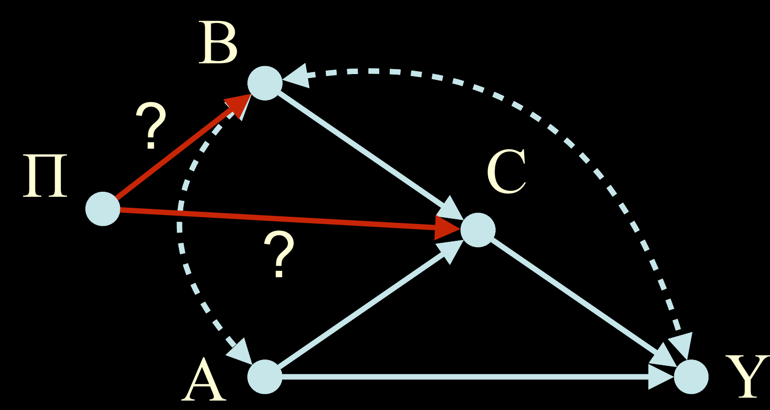
- We embed these results into TS/UCB solvers:
  - z^2-TS: posterior distributions for expected rewards → adjust 'posterior distributions' reflecting all used data
  - z^2-kl-UCB: upper confidence bounds for expected rewards → adjust 'upper bounds' by taking account samples from other arms
- Performance: POMIS+ ≥ POMIS ≥ MIS ≥ Brute-force



WHEN & WHERE TO INTERVENE -- BIG PICTURE

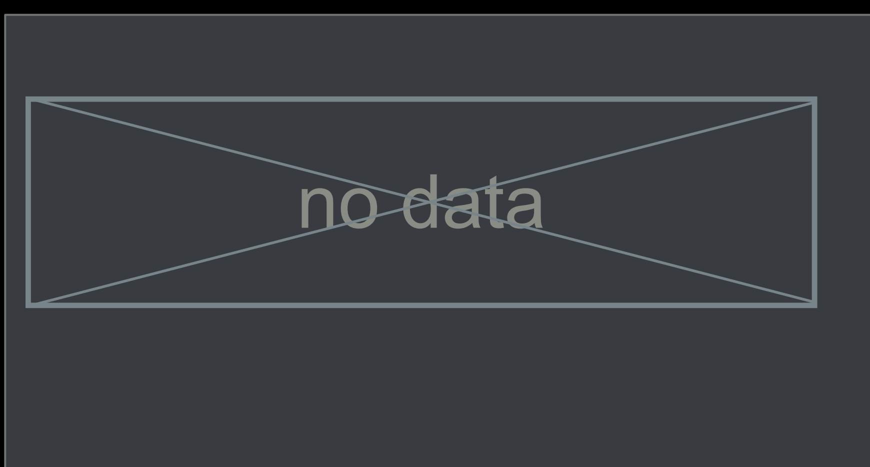


Causal Graph G



under do(x)

POMIS,
formulas



Agent

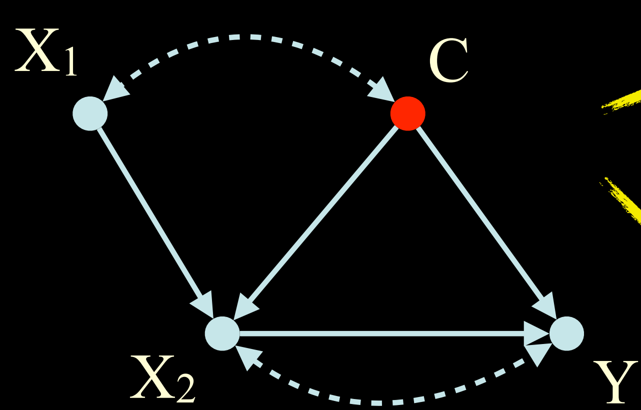
$E_{\mathbf{x}\sim\pi}[Y \mid do(\mathbf{x})]$?

obs(), do(b), …. do(c)

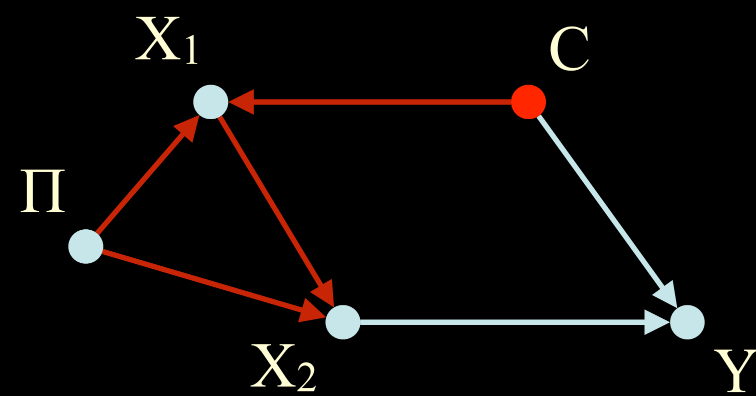
NEW RESULT: WHERE TO INTERVENE & WHAT TO SEE

- In addition to deciding where to intervene, agents also need to decide where to look…

Causal Graph G

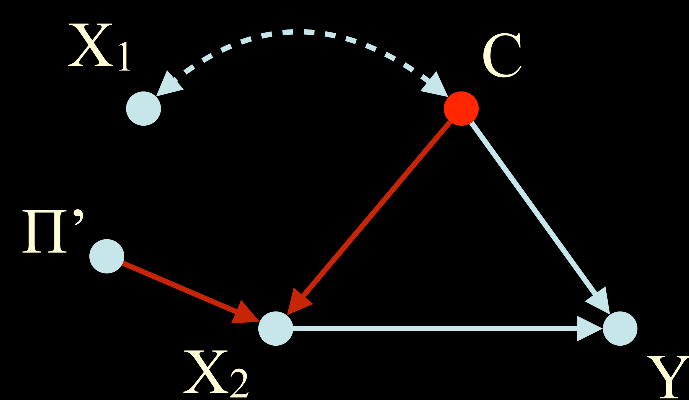


Additional Context C



$\text{do}(x_1|c), \text{do}(x_2|x_1)$

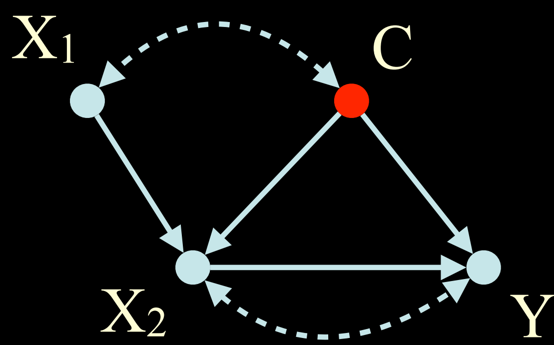
⋮



$\text{do}(x_2|c)$

* both C and X_1 can become a context

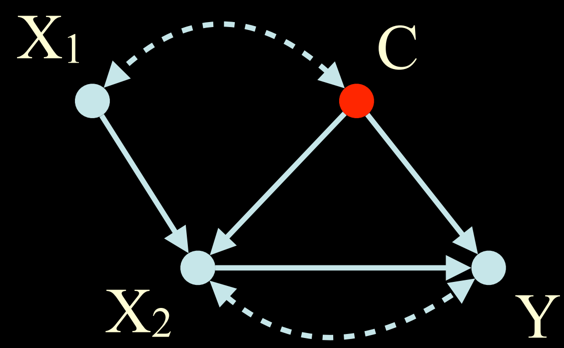
WHERE TO INTERVENE & WHAT TO SEE — POLICY SPACE



Causal Graph G

$\{\}$	$\{X_2\}$			
$do()$	$do(x_2)$	$do(x_2 \mid c)$	$do(x_2 \mid x_1)$	$do(x_2 \mid c, x_1)$
$\{X_1\}$	$\{X_1, X_2\}$			
$do(x_1)$	$do(x_1)$, $do(x_2)$	$do(x_1)$, $do(x_2 \mid c)$	$do(x_1)$, $do(x_2 \mid x_1)$	$do(x_1)$, $do(x_2 \mid c, x_1)$
$do(x_1 \mid c)$	$do(x_1 \mid c)$, $do(x_2)$	$do(x_1 \mid c)$, $do(x_2 \mid c)$	$do(x_1 \mid c)$, $do(x_2 \mid x_1)$	$do(x_1 \mid c)$, $do(x_2 \mid c, x_1)$

WHERE TO INTERVENE & WHAT TO SEE — POLICY SPACE



Policies with the same maximum expected rewards

$do()$

$do(x_1)$

$do(x_1|c)$

$do(x_2)$

$do(x_1), do(x_2)$

$do(x_1|c), do(x_2)$

$do(x_1), do(x_2|x_1)$

$do(x_2|x_1)$

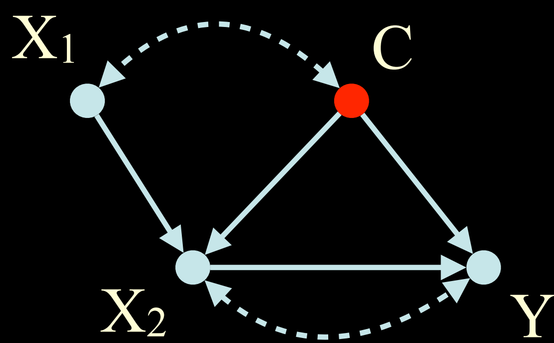
$do(x_1|c), do(x_2|x_1)$

$do(x_2|c)$ $do(x_2|c,x_1)$

$do(x_1), do(x_2|c)$ $do(x_1), do(x_2|c,x_1)$

$do(x_1|c), do(x_2|c)$ $do(x_1|c), do(x_2|c,x_1)$

WHERE TO INTERVENE & WHAT TO SEE — POLICY SPACE



1. minimal policy among reward-equivalent policies

$do()$

$do(x_1)$

$do(x_1|c)$

$do(x_2)$

$do(x_1), do(x_2)$

$do(x_1|c), do(x_2)$

$do(x_1), do(x_2|x_1)$

$do(x_2|x_1)$

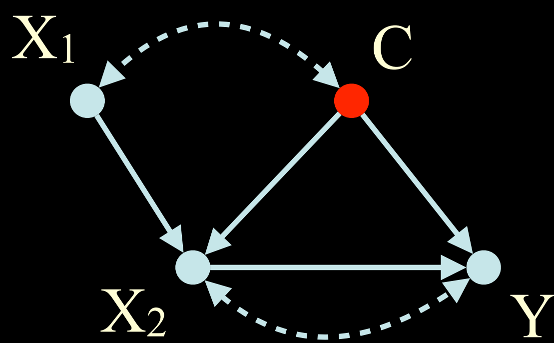
$do(x_1|c), do(x_2|x_1)$

$do(x_2|c)$ $do(x_2|c,x_1)$

$do(x_1), do(x_2|c)$ $do(x_1), do(x_2|c,x_1)$

$do(x_1|c), do(x_2|c)$ $do(x_1|c), do(x_2|c,x_1)$

WHERE TO INTERVENE & WHAT TO SEE — POLICY SPACE



1. minimal policy among

Partial-orders among policies wrt maximum expected rewards

do()

do(x_1)

do(x_1|c)

do(x_2)

do(x_1), do(x_2)

do(x_1|c), do(x_2)

do(x_1), do(x_2|x_1)

$\leq$

do(x_2|x_1)

do(x_1|c), do(x_2|x_1)

do(x_2|c)

do(x_2|c,x_1)

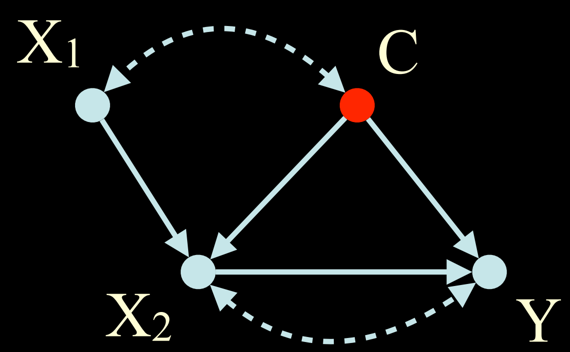
do(x_1), do(x_2|c)

do(x_1), do(x_2|c,x_1)

do(x_1|c), do(x_2|c)

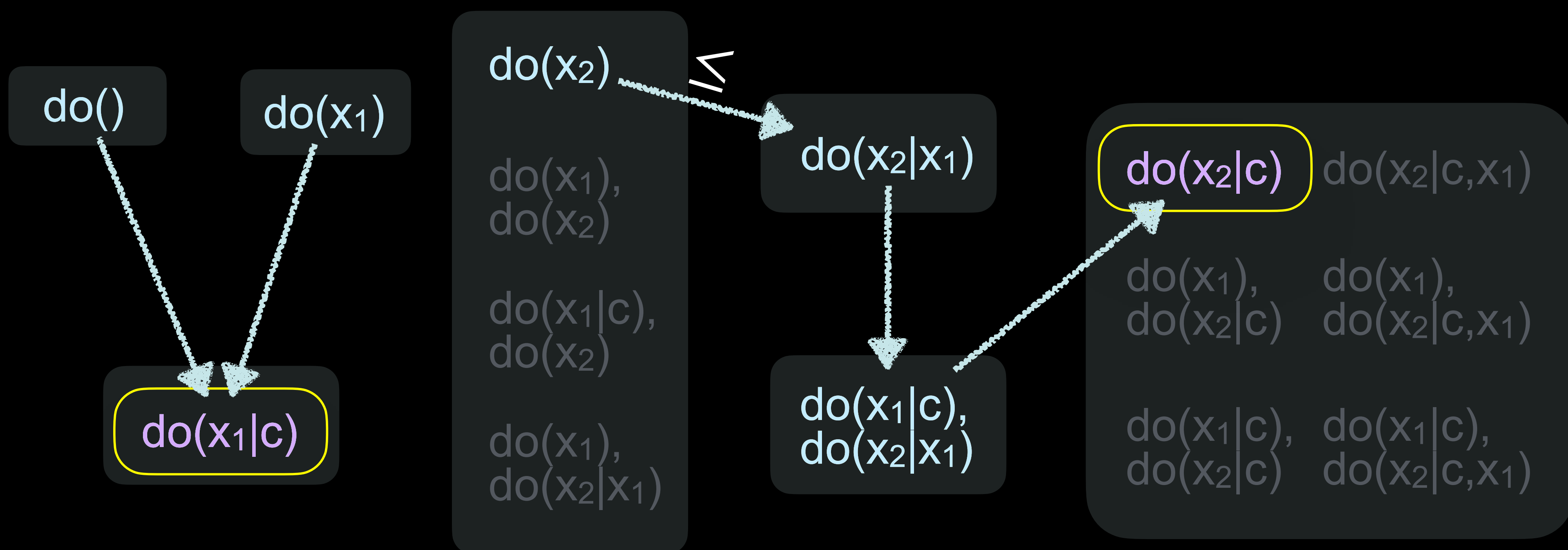
do(x_1|c), do(x_2|c,x_1)

WHERE TO INTERVENE & WHAT TO SEE — POLICY SPACE



1. minimal policy among reward-equivalent policies

2. possibly-optimal policies among min. policies.



* For details, see [R-63 @CausalAI].

TASK 3.
COUNTERFACTUAL DECISION-MAKING
(Intentionality, Free Will, Autonomy)

Andrew Forney

Judea Pearl

CRL-TASK 3. COUNTERFACTUAL DECISION-MAKING

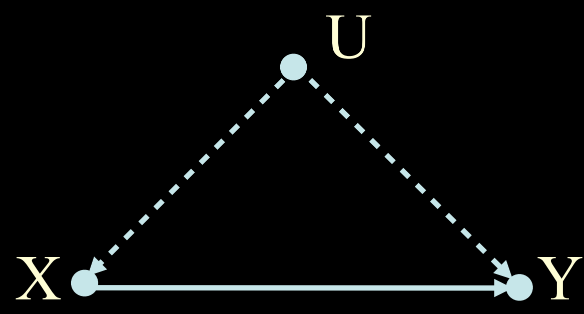
- Agents act in a reflexive manner, without considering the reasons (or causes) for behaving in a particular way. Whenever this is the case, they can be exploited without never realizing.
- This is a general phenomenon in online learning whenever the agent optimizes by Fisherian rand./ the do-distribution (incl. all known RL settings).
- Our goal is to endow agents with the capability of performing counterfactual reasoning (taking their own intent into account), which leads to a more refined notion of regret & a new OPT function.

COUNTERFACTUAL DECISION-MAKING

Question:

How should one select the treatment (x^*) to a particular unit $U=u$ so as to maximize expected reward (Y)?

What if we have observational data? Experimental data?

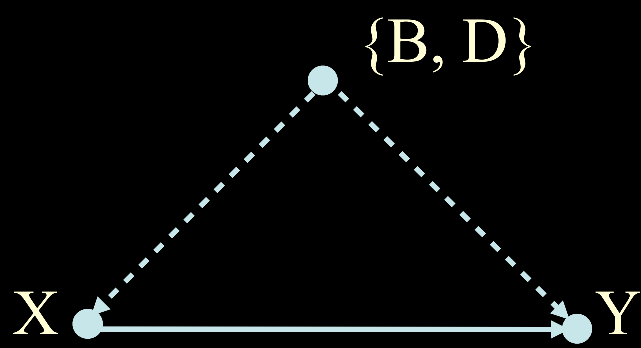


Applications:

» Robotics
» Medical Treatment
» Job Training Program

GREEDY CASINO.
INDIVIDUAL VERSUS POPULATION-LEVEL DECISIONS

Goal: Find a strategy (∏) so as to minimize cumulative regret.



X = type of the machine (x_0, x_1)
Y = reward (y_0, y_1)
B = blinking machine (b_0, b_1)
D = drunkenness level (d_0, d_1)

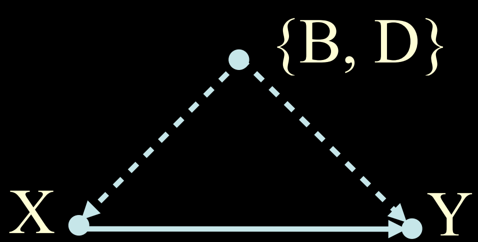
- Regulations: payout has to be ≥ 0.3.
- Casino learns how customers operates and decides to set the payout structure as follows (using ML):

E [y_1 | X, B, D]	D = 0		D = 1	
	B = 0	B = 1	B = 0	B = 1
X = x_1	0.10	0.50	0.40	0.20
X = x_0	0.50	0.10	0.20	0.40

GREEDY CASINO.
INDIVIDUAL VERSUS POPULATION-LEVEL DECISIONS

- Casino's model: $f_X(B, D)$, $P(B)$, $P(D)$,

$E[y_1 \mid X, B, D]$	D = 0		D = 1	
	B = 0	B = 1	B = 0	B = 1
$X = x_1$	0.10	0.50	0.40	0.20
$X = x_0$	0.50	0.10	0.20	0.40



random sample (L_1)

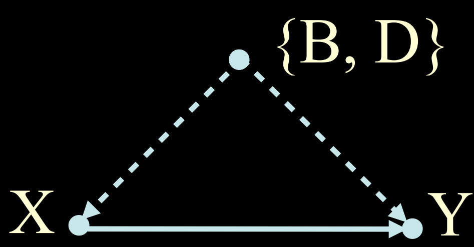
D_1 $E(y_1 \mid X = x_0) = 0.15$

$E(y_1 \mid X = x_1) = 0.15$

GREEDY CASINO.
INDIVIDUAL VERSUS POPULATION-LEVEL DECISIONS

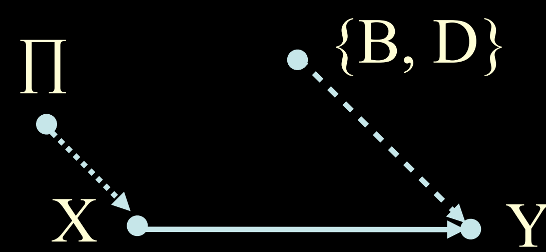
- Casino's model: $f_X(B, D)$, $P(B)$, $P(D)$,

$E[y_1 \mid X, B, D]$	D = 0		D = 1	
	B = 0	B = 1	B = 0	B = 1
$X = x_1$	0.10	0.50	0.40	0.20
$X = x_0$	0.50	0.10	0.20	0.40



random sample (L_1)

random experiment (L_2)

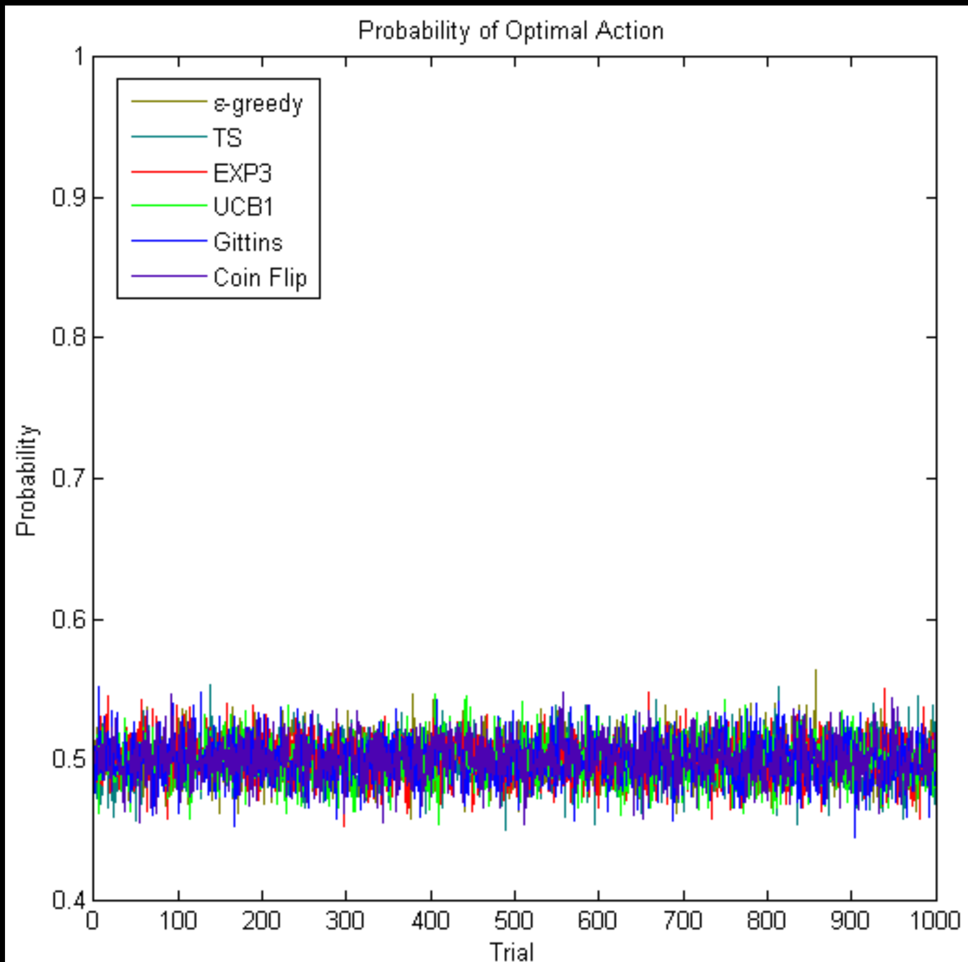


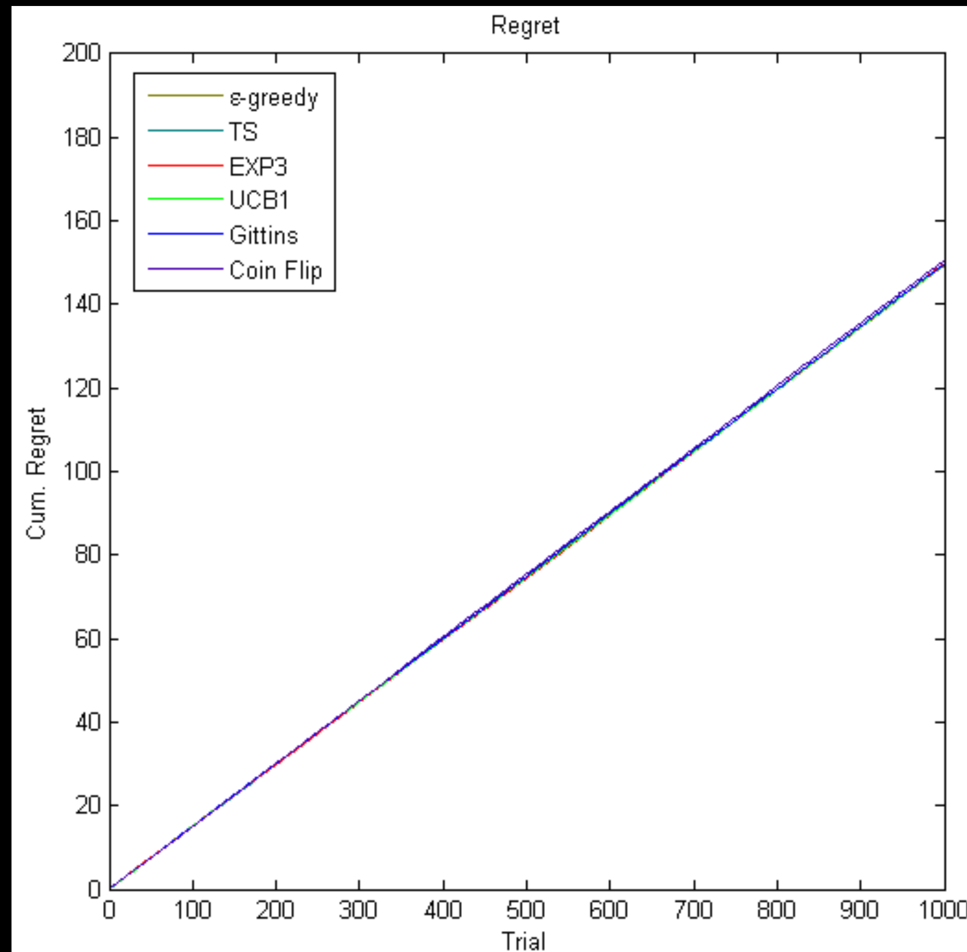
D_1 $E(y_1 \mid X = x_0) = 0.15$
$E(y_1 \mid X = x_1) = 0.15$

D_2 $E(y_1 \mid do(X = x_0)) = 0.30$
$E(y_1 \mid do(X = x_1)) = 0.30$

GREEDY CASINO. INDIVIDUAL VERSUS POPULATION-LEVEL DECISIONS

- Attempt 1. ML (ϵ-greedy, Thompson Sampling, UCB, EXP3).





* Bandits minimize short-term regret based on the do()-distribution.

GREEDY CASINO: CAN WE DO BETTER?

GREEDY CASINO: CAN WE DO BETTER?

- Attempt 2. Counterfactual randomization
- RDC (Regret Decision Criterion):

$$X^* = \arg\max_x E(Y_{X = x_1} \mid X = x_0)$$

$$X^* = \arg\max_x E(Y \mid do(X = x))$$

- This should be read as the counterfactual sentence:

  “Expected value of Y had X been x_1, given that $X = x_0$?”
  (Also known as Effect of Treatment on the Treated.)

GREEDY CASINO: CAN WE DO BETTER?

- Attempt 2. Counterfactual randomization
- RDC (Regret Decision Criterion):

$$X^* = \arg\max_x E(Y_{X=x_1} \mid X = x_0)$$

$$X^* = \arg\max_x E(Y \mid do(X = x)) = E(Y_{X=x})$$

*Also called counterfactual, but too weak (L_2), we'll just call do().

- This should be read as the counterfactual sentence:

"Expected value of Y had X been x_1, given that $X = x_0$?"
(Also known as Effect of Treatment on the Treated.)

GREEDY CASINO: CAN WE DO BETTER?

- Attempt 2. Counterfactual randomization
- RDC (Regret Decision Criterion):

$$X^* = \arg\max_x E(Y_{X = x_1} \mid X = x_0)$$

- This should be read as the counterfactual sentence:

  "Expected value of Y had X been x_1, given that $X = x_0$?"
  (Also known as Effect of Treatment on the Treated.)

- General counterfactuals are difficult (or impossible) to evaluate from data (even experimentally), except for some special conditions (e.g., binary treatment, backdoor admissibility, unconfoundedness) (Pearl, 2000, Ch. 9).

COUNTERFACTUAL DECISION-MAKING

- RDC (Regret Decision Criterion):

$$X^* = \text{argmax}_x \, E(Y_{X = x_1} \mid X = x_0)$$

- Evaluating RDC-type expressions:
  - Note that the agent is about to play machine x_0, which means that (the unknown) $f_X(b, d)$ evaluated to x_0.

COUNTERFACTUAL DECISION-MAKING

- RDC (Regret Decision Criterion):

$$X^* = \text{argmax}_x \, E(Y_{X = x_1} \mid X = x_0)$$

- Evaluating RDC-type expressions:
  - Note that the agent is about to play machine x_0, which means that (the unknown) $f_X(b, d)$ evaluated to x_0.
  - Pause, interrupting decision flow, and wonder:

"I am about to play x_0, would I be better off going with my intuition (x_0) or against it (x_1)?"

COUNTERFACTUAL DECISION-MAKING

- RDC (Regret Decision Criterion):

$$X^* = \text{argmax}_x \; E(Y_{X = x_1} \mid X = x_0)$$

- Evaluating RDC-type expressions:
  - Note that the agent is about to play machine x_0, which means that (the unknown) $f_X(b, d)$ evaluated to x_0.
  - Pause, interrupting decision flow, and wonder:

"I am about to play x_0, would I be better off going with my intuition (x_0) or against it (x_1)?"

Note. If at step 2, we …

COUNTERFACTUAL DECISION-MAKING

- RDC (Regret Decision Criterion):

$$X^* = \text{argmax}_x\, E(Y_{X = x_1} \mid X = x_0)$$

- Evaluating RDC-type expressions:
  - Note that the agent is about to play machine x_0, which means that (the unknown) $f_X(b, d)$ evaluated to x_0.
  - Pause, interrupting decision flow, and wonder:

"I am about to play x_0, would I be better off going with my intuition (x_0) or against it (x_1)?"

Note. If at step 2, we …

- do not interrupt, allowing $X = x_0 \rightarrow P(x_0, y)$.

COUNTERFACTUAL DECISION-MAKING

- RDC (Regret Decision Criterion):

$$X^* = \text{argmax}_x \, E(Y_{X = x_1} \mid X = x_0)$$

- Evaluating RDC-type expressions:
  - Note that the agent is about to play machine x_0, which means that (the unknown) $f_X(b, d)$ evaluated to x_0.
  - Pause, interrupting decision flow, and wonder:

"I am about to play x_0, would I be better off going with my intuition (x_0) or against it (x_1)?"

Note. If at step 2, we …

- do not interrupt, allowing $X = x_0 \rightarrow P(x_0, y)$.
- do interrupt and make $X = \text{rand}() = x_1 \rightarrow P(y \mid do(x_1))$.

COUNTERFACTUAL DECISION-MAKING

- RDC (Regret Decision Criterion):

$$X^* = \operatorname{argmax}_x E(Y_{X = x_1} \mid X = x_0)$$

- Evaluating RDC-type expressions:
  - Note that the agent is about to play machine x_0, which means that (the unknown) $f_X(b, d)$ evaluated to x_0.
  - Pause, interrupting decision flow, and wonder:

"I am about to play x_0, would I be better off going with my intuition (x_0) or against it (x_1)?"

Note. If at step 2, we …

- do not interrupt, allowing $X = x_0 \rightarrow P(x_0, y)$.
- do interrupt and make $X = rand() = x_1 \rightarrow P(y \mid do(x_1))$.

COUNTERFACTUAL DECISION-MAKING

- RDC (Regret Decision Criterion):

$$X^* = \text{argmax}_x \, E(Y_{X = x_1} \mid X = x_0)$$

- Evaluating RDC-type expressions:
  - Note that the agent is about to play machine x_0, which means that (the unknown) $f_X(b, d)$ evaluated to x_0.
  - Pause, interrupting decision flow, and wonder:

"I am about to play x_0, would I be better off going with my intuition (x_0) or against it (x_1)?"

Note. If at step 2, we …

- do not interrupt, allowing $X = x_0 \rightarrow P(x_0, y)$.
- do interrupt and make $X = \text{rand}() = x_1 \rightarrow P(y \mid do(x_1))$.
- do interrupt and make $X = \text{rand}() = x_1 \mid x_0 \rightarrow P(Y_{x_1} \mid x_0)$.

COUNTERFACTUAL DECISION-MAKING

- RDC (Regret Decision Criterion):

$$X^* = \text{argmax}_x\, E(Y_{X = x_1} \mid X = x_0)$$

- Evaluating RDC-type expressions:
  - Note that the agent is about to play machine x_0, which means that (the unknown) $f_X(b, d)$ evaluated to x_0.
  - Pause, interrupting decision flow, and wonder:

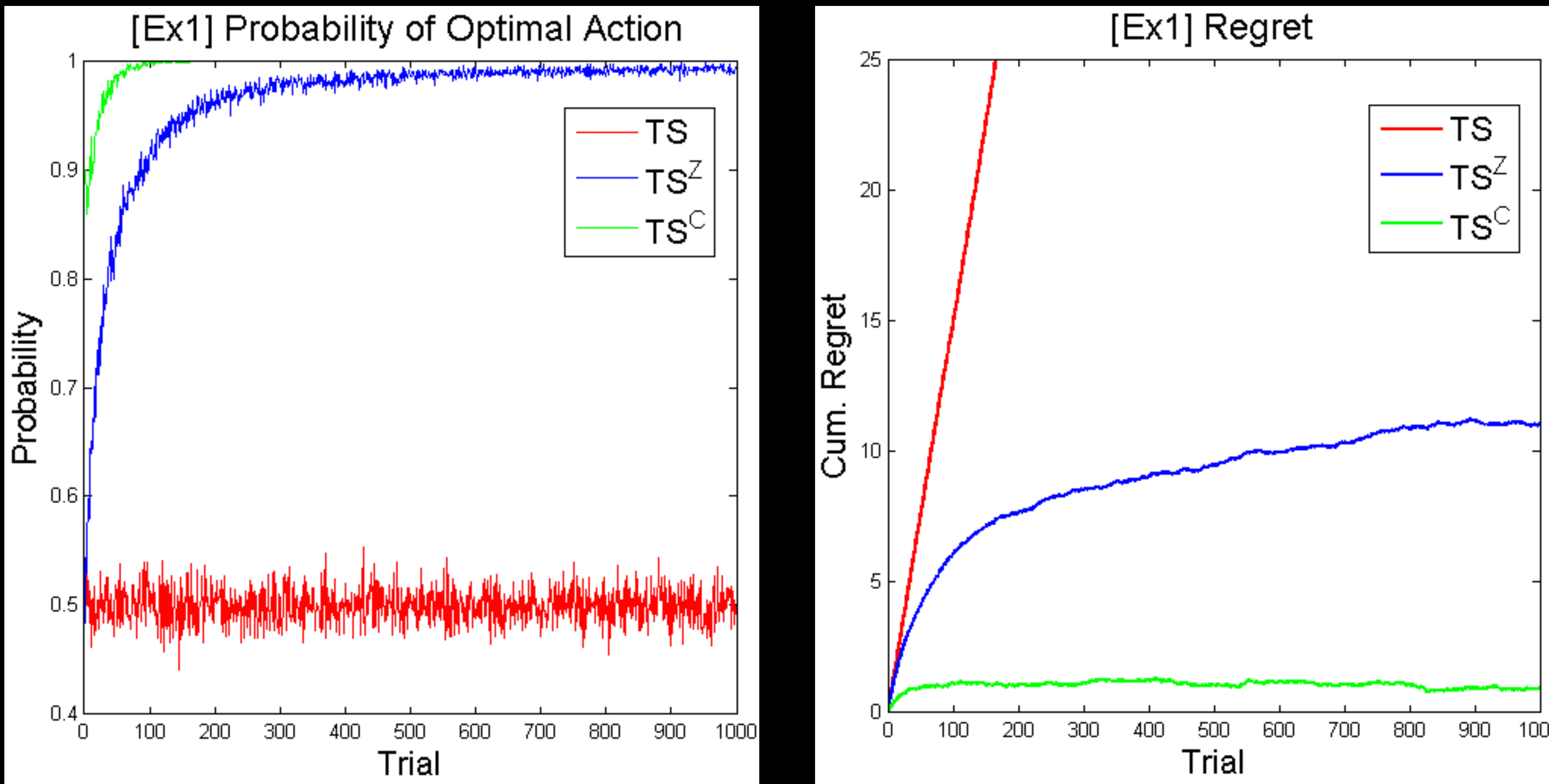
"I am about to play x_0, would I be better off going with my intuition (x_0) or against it (x_1)?"

Note. If at step 2, we …

- do not interrupt, allowing $X = x_0 \rightarrow P(x_0, y)$. EDT
- do interrupt and make $X = \text{rand}() = x_1 \rightarrow P(y \mid do(x_1))$. CDT
- do interrupt and make $X = \text{rand}() = x_1 \mid x_0 \rightarrow P(Y_{x_1} \mid x_0)$. RDT

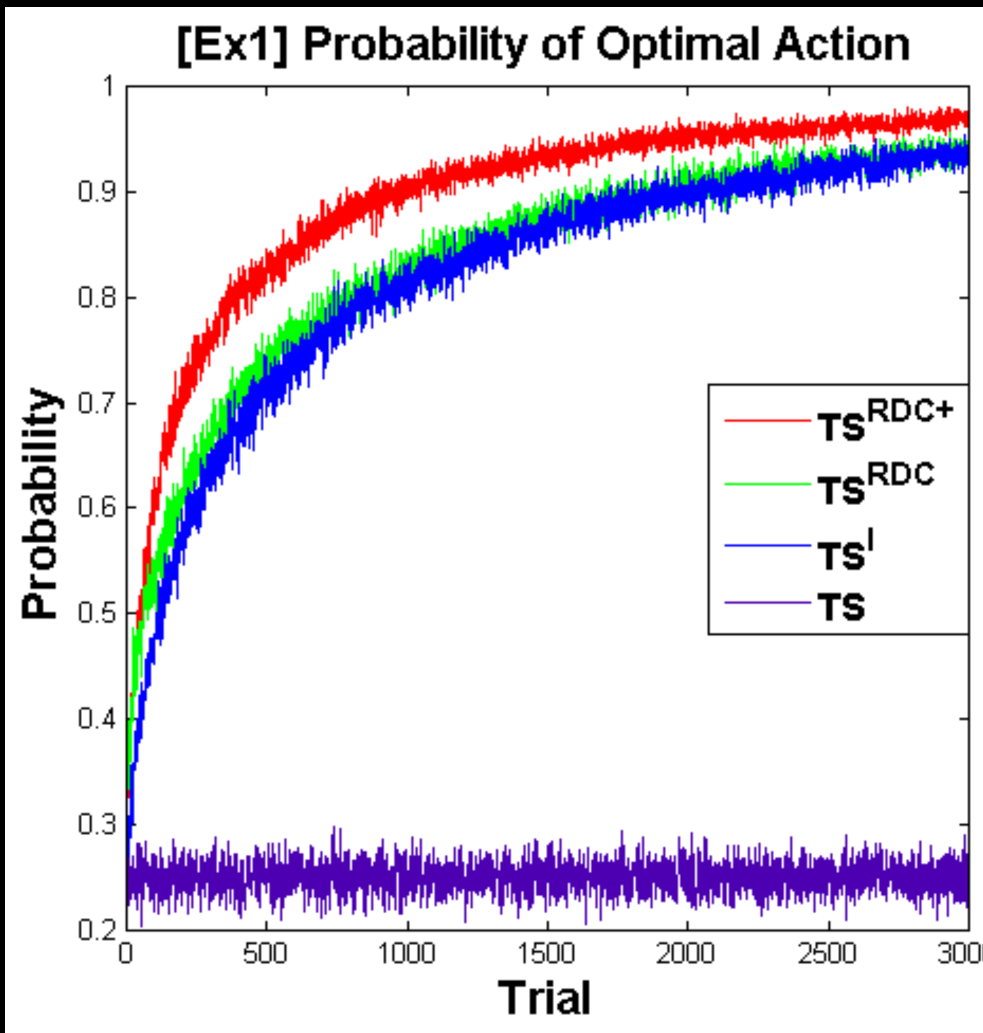
REGRET DECISION CRITERION: EXPERIMENTAL RESULTS

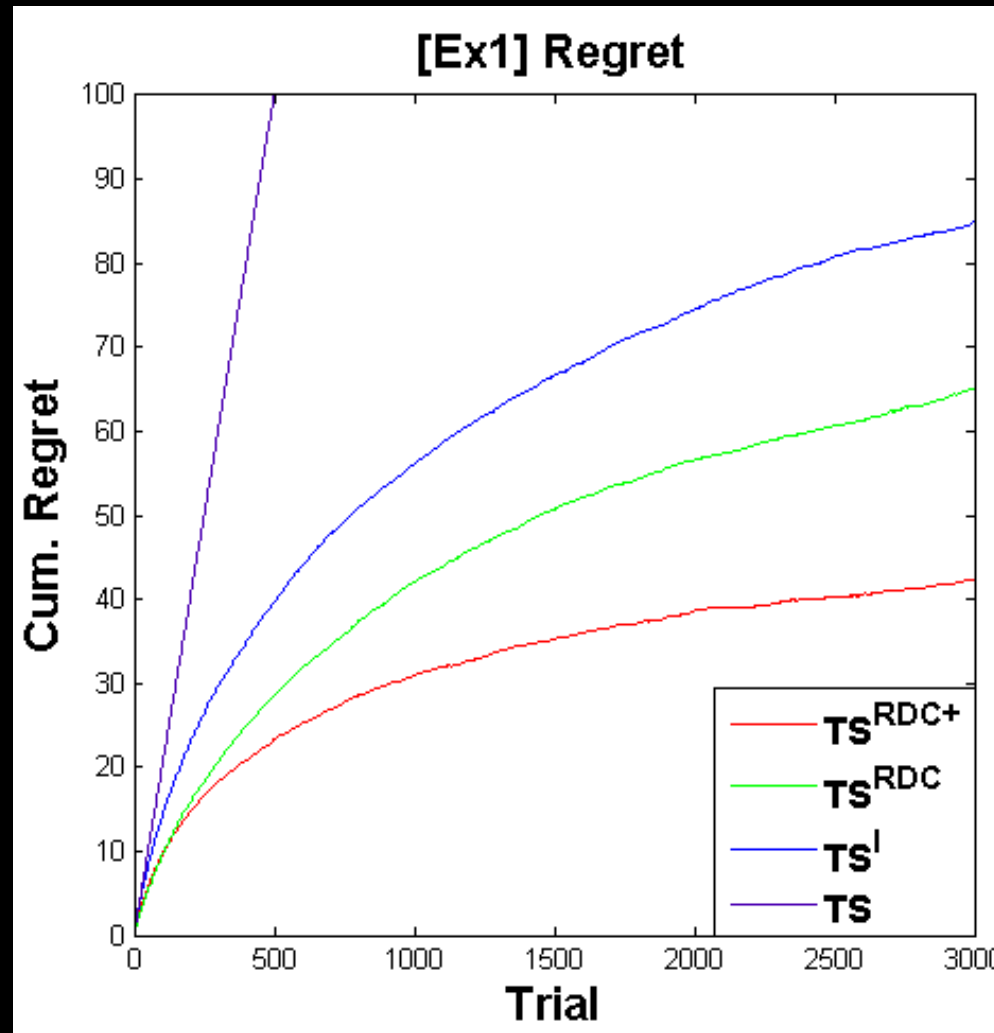
- Greedy Casino Parametrization



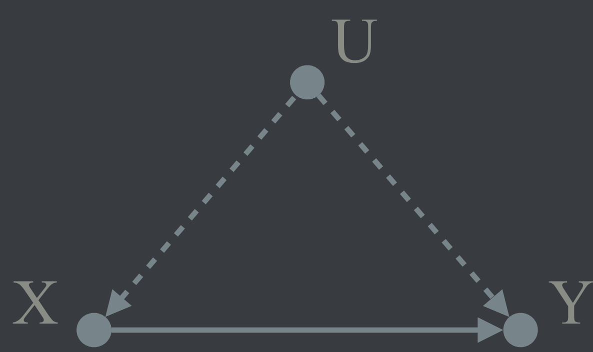
REGRET DECISION CRITERION: EXPERIMENTAL RESULTS

- What if the experimental distribution is available (4-arm case)?



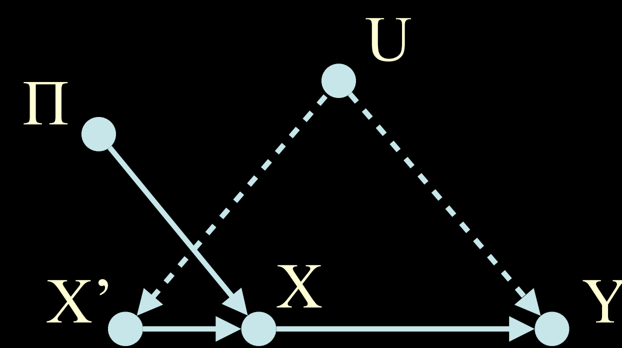


TASK 3. COUNTERFACTUAL LEARNING



the environment

no data



under ctf. randomization

Agent

$E_{x\sim\pi}[Y_x \mid x']$?

$Y_{x_0}|x_1,\ Y_{x_1}|x_1,\ \ldots,\ Y_{x_1}|x_0$

APPLICATION: HUMAN-AI COLLABORATION
(CAN HUMANS BE OUT OF THE LOOP?*)

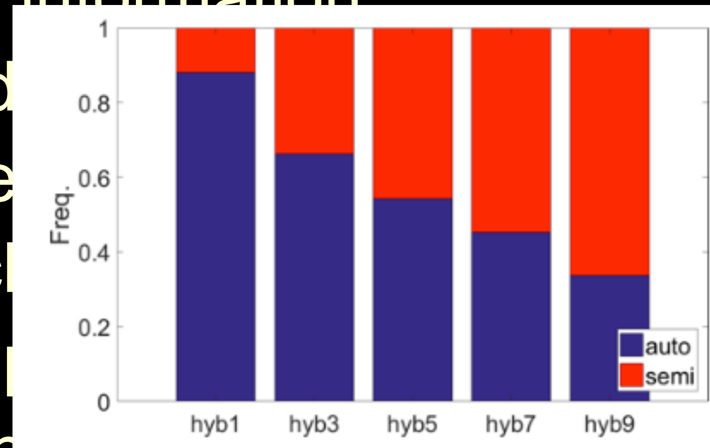
- Observation from the RDC, if $E[Y_x|x'] = E[Y|do(x)]$ → the human's intuition has no value of information.
- In words, the human expert could be replaced without sacrificing the performance of the system, at least in principle full autonomy can be achieved.
- Contribution: New Markovian properties (L_2, L_3) that establishes whether an agent can be autonomous.

Env. Model	Optimality		Autonomy
	A_{exp}	A_{ctf}	
MDPUC^-	✓	✓	✓
MDPUC	✗	✓	✗
MDPUC^+	✗	✓	✗
DSCM^-	✓	✓	✓
DSCM	✗	✓	✗

* For details, see [R-64 @CausalAI].

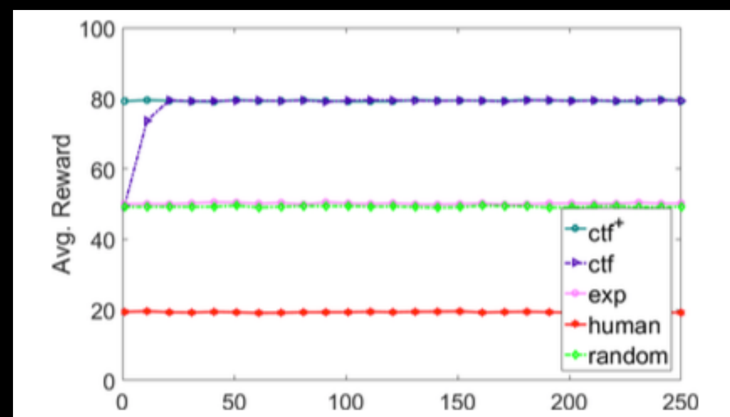
APPLICATION: HUMAN-AI COLLABORATION (CAN HUMANS BE OUT OF THE LOOP?*)

- Observation from the RDC, if $E[Y_x|x'] = E[Y|do(x)] \rightarrow$ the human's intuition has no value of information
- In words, the human expert could sacrificing the performance of the principle full autonomy can be ac
- Contribution: New Markovian pro establishes whether an agent ca



(c) Hybrid Planning

Env. Model	Optimality		Autonomy
	A_{exp}	A_{ctf}	
MDPUC$^-$	✓	✓	✓
MDPUC	✗	✓	✗
MDPUC$^+$	✗	✓	✗
DSCM$^-$	✓	✓	✓
DSCM	✗	✓	✗



(a) Online Learning

* For details, see [R-64 @CausalAI].

SUMMARY CRL TASKS

CRL CAPABILITIES (I)

1. Generalized Policy Learning (on+offline) 👀 + 💪

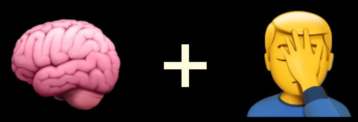
- Online learning is too costly and learning from scratch is usually impractical. Still, the assumptions of offline learning are rarely satisfied in practice.
- Goal: Move towards more realistic learning scenarios where the two modalities come together, extracting as much causal information as possible from confounded data, and using it in the most efficient way.

2. When and where to intervene? 🧭 → 💪

- Agents usually have a fixed policy space (actions), and intervening is usually assumed as beneficial.
- Goal: Understand when interventions are needed and whenever this is the case, what should be changed in the system to bring about the desired outcome.

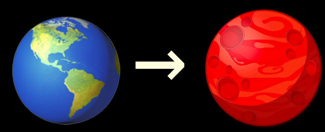
CRL CAPABILITIES (II)

3. Counterfactual Decision-Making (intentionality, regret & free will)



- Agents act in a reflexive manner, without considering the reasons (causes) for behaving in a certain way.
- Goal: Endow agents with the capability of taken their own intent into account, which will lead to a new notion of regret based on counterfactual randomization.

4. Generalizable and Robust Decision-Making (transportability & structural invariances)



- The knowledge acquired by an agent is usually circumscribed to the domain where it was deployed.
- Goal: Allow agents to extrapolate knowledge, making more robust and generalizable claims by leveraging the causal invariances shared across environments.

CRL CAPABILITIES (III)

5. Learning Causal Models by Combining Observations & Experimentation 🔬 + ⚒️

- Agents have a fixed causal model, constructed from templates or from background knowledge.
- Goal: Allow agents to systematically combine the observations and interventions it’s already collecting to construct an equivalence class of causal models.

6. Causal Imitation Learning 🙈 → 🙉

- Mimicking is one of the common ways of learning. Whenever the demonstrator has a different causal model, imitating may lead to disastrous side effects.
- Goal: Understand the conditions so that imitation by behavioral cloning is valid and leads to faster learning. Otherwise, introduce more refined imitation modalities.

CRL CAPABILITIES (III)

CRL (CHEAT SHEET)

1. Generalized Policy Learning (on+offline)
   Combining L_1 + L_2 interactions to learn policy Π.

2. When and where to intervene?
   Identifying subset of L_2 and optimize the policy space.

3. Counterfactual Decision-Making
   Optimization function based on L_3 counterfactual & random.

4. Generalizability and Robustness
   Generalizing from training environment (SCM M) to SCM M*.

5. Learning Causal Model G
   Combining L_1 + L_2 interactions to learn G (of M).

6. Causal Imitation Learning
   Learning L_2 -policy based on partially observable L_1-data (expert).

behavioral cloning is valid and leads to faster learning. Otherwise, introduce more refined imitation modalities.

CONCLUSIONS

- CI & RL are fundamentally intertwined and novel learning opportunities emerge when this connection is fully realized.
  - The structural invariances encoded in the causal graph (w.r.t. SCM *M*) can be leveraged and combined with RL allocation procedures leading to robust learning.
  - Still, failure to acknowledge distinct invariances of the environment (*M*) almost always leads to poor decision-making.
- CRL opens up a new family of learning problems that were neither acknowledged nor understood before, including the combination of online & offline learning (GPL), when/where to intervene, counterfactual decision-making, generalizability across environments, to cite a few.
- Program: Develop a principled framework for designing causal AI systems integrating [observational, experimental, counterfactual] data, modes of reasoning, knowledge.
  - Leads to a natural treatment to human-like explainability and rational decision-making.

CONCLUSIONS

- CI & RL are fundamentally intertwined and novel learning opportunities emerge when this connection is fully realized.
  - The structural invariances encoded in the causal graph (w.r.t. SCM *M*) can be leveraged and combined with RL allocation procedures leading to robust learning.
  - Still, failure to acknowledge distinct invariances of the environment (*M*) almost always leads to poor decision-making.
- CR… ere neit… e con… e to intervene, counterfactual decision-making, generalizability across environments, to cite a few.

Resources: https://crl.causalai.net

- Program: Develop a principled framework for designing causal AI systems integrating [observational, experimental, counterfactual] data, modes of reasoning, knowledge.
  - Leads to a natural treatment to human-like explainability and rational decision-making.

THANK YOU!

REFERENCES

[F 1935] Fisher, R. A. *The Design of Experiments*. Oliver and Boyd 1935.

[WD 1992] Watkins, C., Dayan, P. *Q-Learning.* Machine Learning volume 8. 1992.

[BP 1994] Balke, A., Pearl, J. *Counterfactual Probabilities: Computational Methods, Bounds, and Applications* In Proceedings of the Conference on Uncertainty in Artificial Intelligence 1994.

[SB 1998] R. Sutton, A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[P 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge Press, 2000.

[ACF 2002] Auer, P., Cesa-Bianchi, N., Fischer, P. *Finite-time Analysis of the Multiarmed Bandit Problem* Machine Learning volume 47. 2002.

[JOA 2010] Jaksch, T., Ortner, R., Auer, P. *Near-optimal Regret Bounds for Reinforcement Learning*. Journal of Machine Learning Research 11. 2010.

[DLL 2011] Dudik, M., Langford, J., Li, L. *Doubly robust policy evaluation and learning*. In Proceedings of 28th International Conference on Machine Learning. 2011.

[BP 2014] E. Bareinboim, J. Pearl. *Transportability from Multiple Environments with Limited Experiments: Completeness Results.* In Proceedings of the 27th Annual Conference on Neural Information Processing Systems, 2014.

REFERENCES

[BFP 2015] E. Bareinboim, A. Forney, J. Pearl. *Bandits with Unobserved Confounders: A Causal Approach.* In Proceedings of the 28th Annual Conference on Neural Information Processing Systems, 2015.

[BP 2016] E. Bareinboim, J. Pearl. *Causal inference and the data-fusion problem.* Proceedings of the National Academy of Sciences, v. 113 (27), pp. 7345-7352, 2016.

[JL 2016] Jiang, N., Li, L. *Doubly robust off-policy value evaluation for reinforcement learning*. In Proceedings of the 33rd International Conference on Machine Learning. 2016.

[ZB 2016] J. Zhang, E. Bareinboim. *Markov Decision Processes with Unobserved Confounders: A Causal Approach.* CausalAI Lab, Technical Report (R-23), 2016.

[FPB 2017] A. Forney, J. Pearl, E. Bareinboim. *Counterfactual Data-Fusion for Online Reinforcement Learners.* In Proceedings of the 34th International Conference on Machine Learning, 2017.

[KSB 2017] M. Kocaoglu, K. Shanmugam, E. Bareinboim. *Experimental Design for Learning Causal Graphs with Latent Variables.* In Proceedings of the 31st Annual Conference on Neural Information Processing Systems, 2017.

[ZB 2017] J. Zhang, E. Bareinboim. *Transfer Learning in Multi-Armed Bandits: A Causal Approach.* In Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017.

REFERENCES

[GSKB 2018] Ghassami, A., Salehkaleybar, S., Kiyavash, N., Bareinboim, E. *Budgeted Experiment Design for Causal Structure Learning.* In Proceedings of the 35th International Conference on Machine Learning. 2018.

[KZ 2018] Kallus, N., Zhou, A. *Confounding-robust policy improvement.* In Advances in Neural Information Processing Systems 2018.

[LB 2018] S. Lee, E. Bareinboim. *Structural Causal Bandits: Where to Intervene?* In Proceedings of the 32nd Annual Conference on Neural Information Processing Systems, 2018.

[PM 2018] J. Pearl, D. Mackenzie. *The book of why: The new science of causal and effect*. Basic Books, 2018.

[FB 2019] A. Forney, E. Bareinboim. *Counterfactual Randomization: Rescuing Experimental Studies from Obscured Confounding.* In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019.

[KJSB 2019] Kocaoglu, M., Jaber, A., Shanmugam, K., Bareinboim, E. *Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions.* In Proceedings of the 33rd Annual Conference on Neural Information Processing Systems. 2019.

[LB 2019] S. Lee, E. Bareinboim. *Structural Causal Bandits with Non-manipulable Variables.* In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019.

REFERENCES

[LCB 2019] S. Lee, J. Correa, E. Bareinboim. *General Identifiability with Arbitrary Surrogate Experiments* In Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence, 2019.

[ZB 2019] Zhang, J., Bareinboim, E. *Near-Optimal Reinforcement Learning in Dynamic Treatment Regimes.* In Advances in Neural Information Processing Systems 2019.

[BCII 2020] Bareinboim, E, Correa, J, Ibeling, D, Icard, T. *On Pearl's Hierarchy and the Foundations of Causal Inference.* In "Probabilistic and Causal Inference: The Works of Judea Pearl" (ACM Special Turing Series). 2020.

[BLZ 2020] Bareinboim, E, Lee, S, Zhang, J. *An Introduction to Causal Reinforcement Learning.* Columbia CausalAI Laboratory, Technical Report (R-65). 2020.

[CB 2020] Correa, J, Bareinboim, E. *Transportability of Soft Effects: Completeness Results.* Columbia CausalAI Laboratory, Technical Report (R-68). 2020.

[JKSB 2020] Jaber, A, Kocaoglu, M, Shanmugam, K, Bareinboim, E. *Causal Discovery from Soft Interventions with Unknown Targets: Characterization & Learning.* Columbia CausalAI Laboratory, Technical Report (R-67). 2020.

[JTB 2020] Jung, Y, Tian, J, Bareinboim, E. *Learning Causal Effects via Empirical Risk Minimization.* Columbia CausalAI Laboratory, Technical Report (R-62). 2020.

REFERENCES

[LB 2020] Lee, S, Bareinboim, E. *Characterizing Optimal Mixed Policies: Where to Intervene, What to Observe.* Columbia CausalAI Laboratory, Technical Report (R-63). 2020.

[NKYB 2020] Namkoong, H., Keramati, R.,Yadlowsky, S., Brunskill, E. *Off-policy Policy Evaluation For Sequential Decisions Under Unobserved Confounding.* arXiv:2003.05623. 2020.

[ZB 2020a] Zhang, J., Bareinboim, E. *Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach.* In Proceedings of the 37th International Conference on Machine Learning. 2020.

[ZB 2020b] Zhang, J, Bareinboim, E. *Bounding Causal Effects on Continuous Outcomes.* Columbia CausalAI Laboratory, Technical Report (R-61). 2020.

[ZB 2020c] Zhang, J, Bareinboim, E. *Can Humans Be Out of the Loop?* Columbia CausalAI Laboratory, Technical Report (R-64). 2020.

[ZKB 2020] Zhang, J, Kumor, D, Bareinboim, E. *Causal Imitation Learning with Unobserved Confounders.* Columbia CausalAI Laboratory, Technical Report (R-66). 2020.