
When Should Reinforcement Learning Use Causal Reasoning

Oliver Schulte
School of Computing Science
Sports Analytics Group
Simon Fraser University

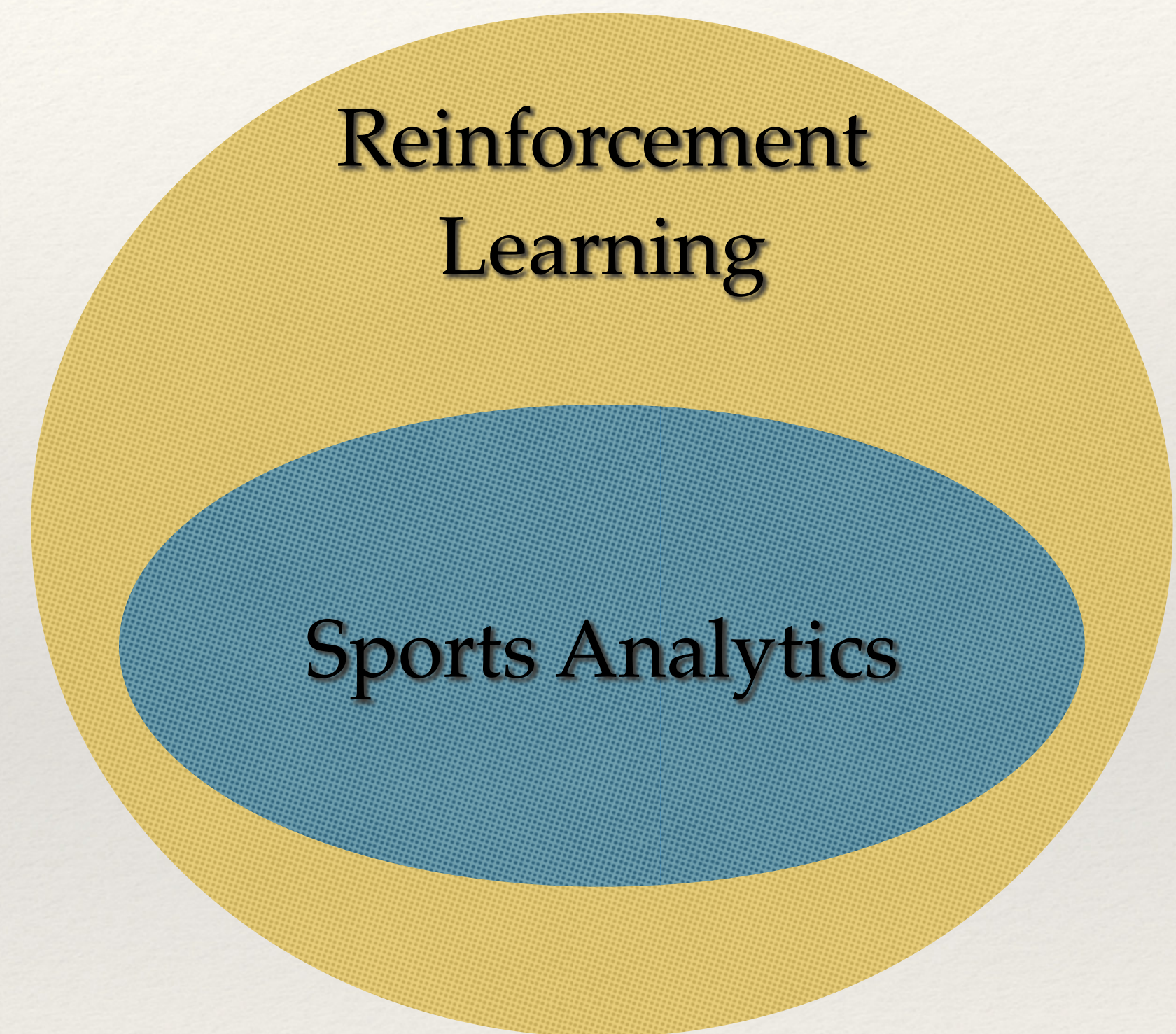


Pascal Poupart
Cheriton School of Computer Science
University of Waterloo

Thank you to Mathias and
Arynn for organizing!

Update/Advertising: Reinforcement Learning and Sports Analytics

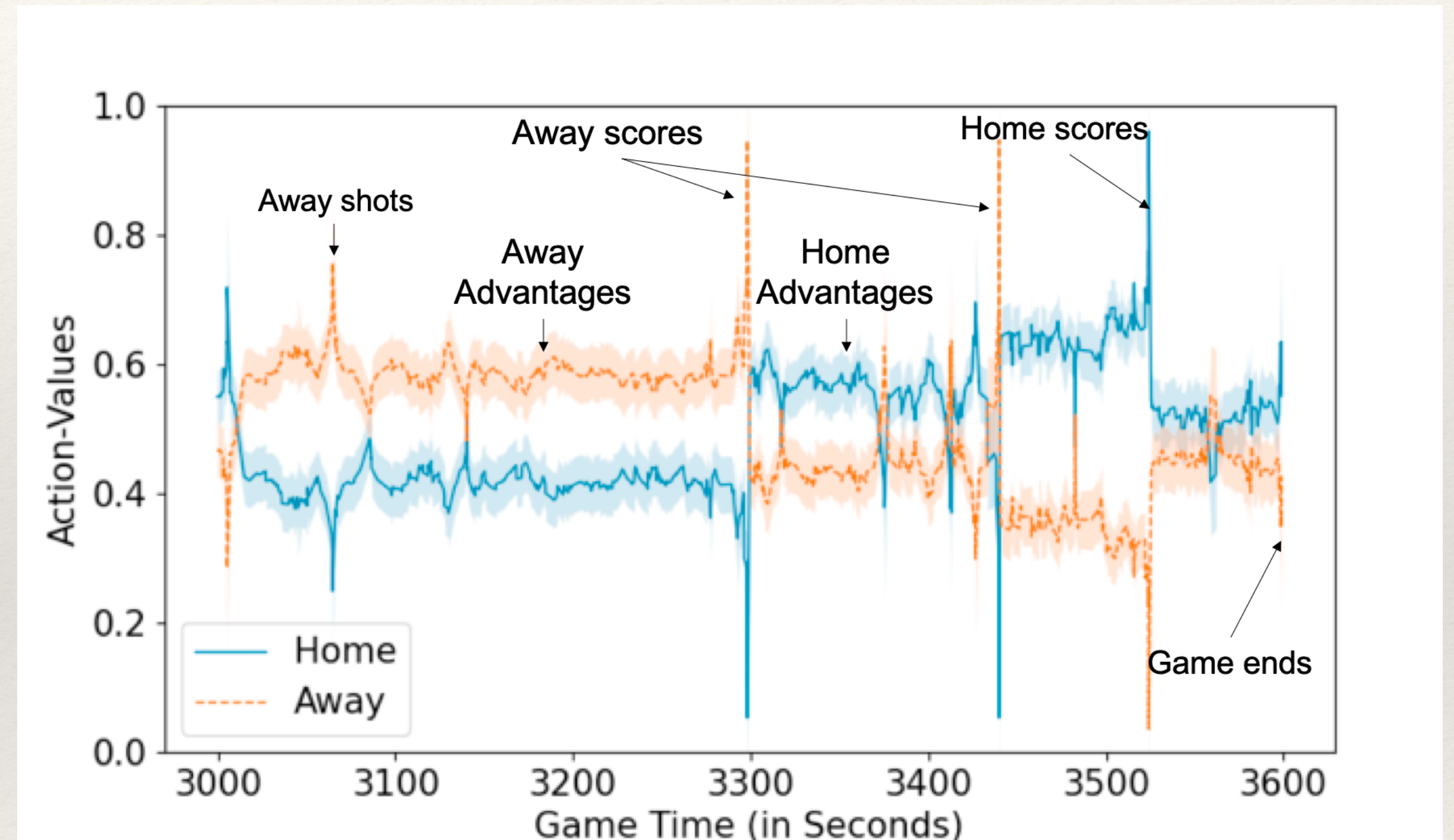
- Research on ML and causal modelling for structured data: networks, graphs, event logs
- Since 2015: Applying reinforcement learning to *sports analytics*
- Collaboration with Sportlogiq from Montreal
- Big SL data set: 1M+ Events in 1 Season



RL for Sports: Value Functions and Player Ranking

- General Idea: learn a value function for the National Hockey League (off-line)
 - Use distributional RL to capture uncertainty (standard deviations)
- Action values \rightarrow player ranking

Name	GIM	Assists	Goals	Points	Team	Salary
Taylor Hall	96.40	39	26	65	EDM	\$6,000,000
Joe Pavelski	94.56	40	38	78	SJS	\$6,000,000
Johnny Gaudreau	94.51	48	30	78	CGY	\$925,000
Anze Kopitar	94.10	49	25	74	LAK	\$7,700,000
Erik Karlsson	92.41	66	16	82	OTT	\$7,000,000
Patrice Bergeron	92.06	36	32	68	BOS	\$8,750,000
Mark Scheifele	90.67	32	29	61	WPG	\$832,500
Sidney Crosby	90.21	49	36	85	PIT	\$12,000,000
Claude Giroux	89.64	45	22	67	PHI	\$9,000,000
Dustin Byfuglien	89.46	34	19	53	WPG	\$6,000,000
Jamie Benn	88.38	48	41	89	DAL	\$5,750,000
Patrick Kane	87.81	60	46	106	CHI	\$13,800,000
Mark Stone	86.42	38	23	61	OTT	\$2,250,000
Blake Wheeler	85.83	52	26	78	WPG	\$5,800,000
Tyler Toffoli	83.25	27	31	58	DAL	\$2,600,000
Charlie Coyle	81.50	21	21	42	MIN	\$1,900,000
Tyson Barrie	81.46	36	13	49	COL	\$3,200,000
Jonathan Toews	80.92	30	28	58	CHI	\$13,800,000
Sean Monahan	80.92	36	27	63	CGY	\$925,000
Vladimir Tarasenko	80.68	34	40	74	STL	\$8,000,000



Leafs@Flyers March 2019

Liu, G., Luo, Y., Schulte, O. and Poupart, P. (2022),
Uncertainty-Aware Reinforcement Learning for Risk-Sensitive Player Evaluation in Sports,
Neurips Proceedings pp. 20218--20231.

Causality and Reinforcement Learning: Overview

A Match Made in Heaven?

Goals of Talk

- Foundations talk
 - No experiments
 - Instead definitions, examples, theorems
- Motivation
 - Goal 1: Explain when and how causal models can help RL
 - Goal 2: Connect causal modelling and RL communities.
Short tutorial on causal concepts for RL researchers
- Long tutorial: Bareinboim et al (2020), Survey Deng et al. (2023)
- also section E in Schölkopf, B., Locatello, F., Bauer, S., Rosemary Nan Ke, Kalchbrenner, N., Goyal, A. and Bengio, Y. (2021)

Bareinboim, E. et al. (2020), 'Towards Causal Reinforcement Learning', ICML Tutorial

Z. Deng, J. Jiang, G. Long, and C. Zhang (2023), "Causal Reinforcement Learning: A Survey," arXiv preprint

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A. and Bengio, Y. (2021),

'Toward causal representation learning', *Proceedings of the IEEE* **109**(5), 612--634.

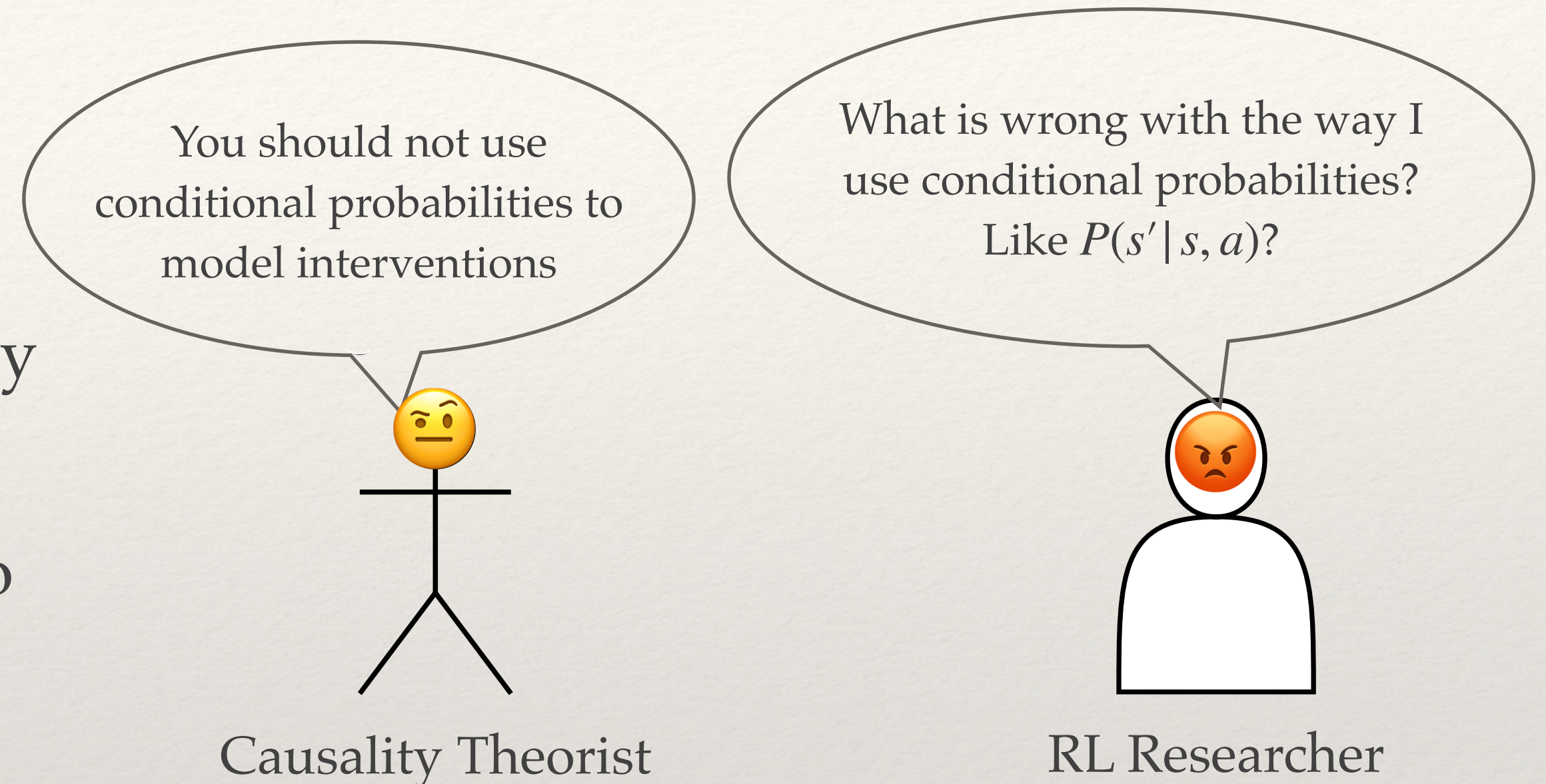
Causality and RL: Common Ground

- Both fields model the **effects of actions**
- *Time* is important in dynamic RL process models
- Temporal information is very useful in causal learning because
causes precede their effects in time
- The future cannot cause the past

RL	Causal Model
Action	Intervention / Treatment
Reward	Response

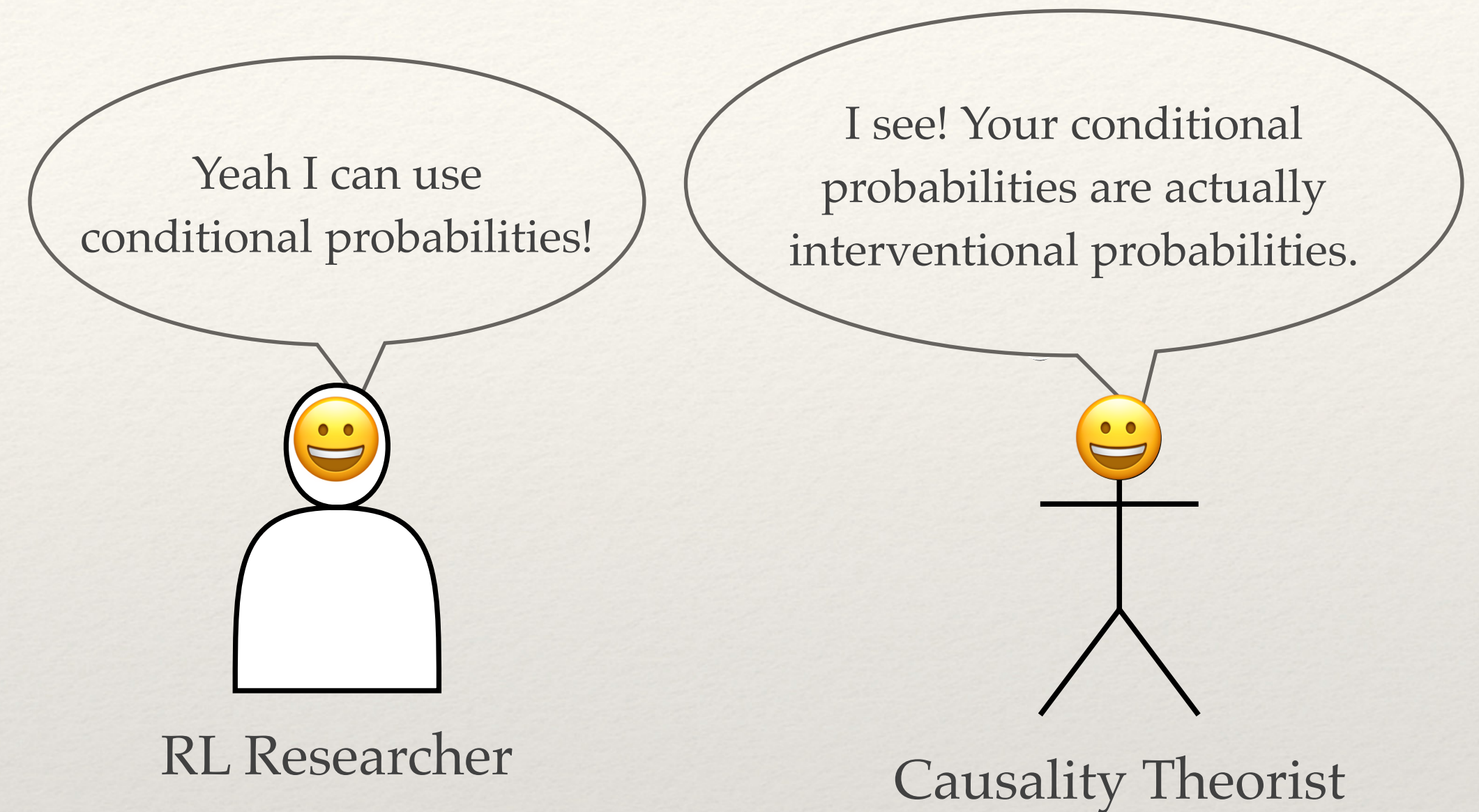
Causality and RL: Differences

- Starting Point of causality theory:
Causation \neq Correlation
- ➔ Conditional Probability \neq Interventional Probability
aka *Causal Effect*
- Example: Doctor visits correlate with illness but do not make you sick (Pearl 2000)
- $P(\text{ill} \mid \text{visit}) \gg P(\text{ill}) = P(\text{ill} \mid \text{do}(\text{visit}))$
 - ↑
Condition
on action
 - ↑
Intervene
to send patient to doctor

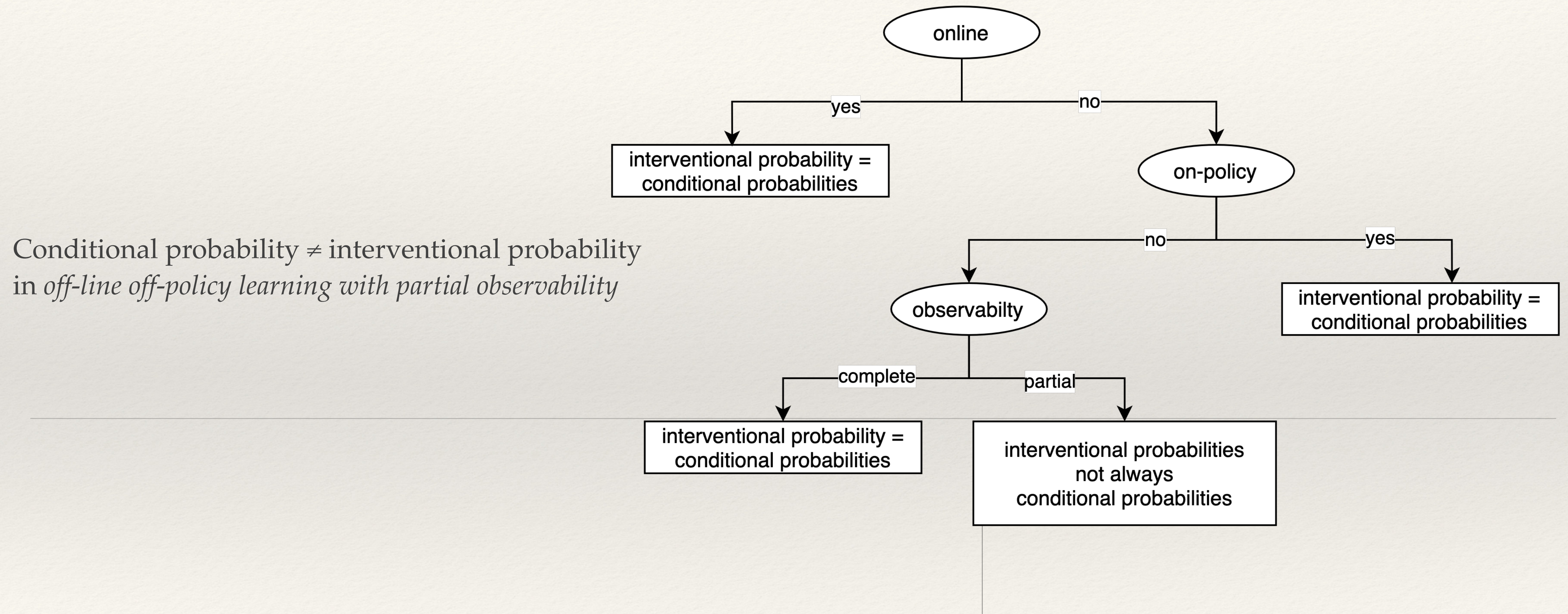


Conditional vs. Interventional Probabilities

- Schölkopf, Nan Ke, Goyal, ... Bengio (2021): “[RL] sometimes effectively directly estimates do-probabilities. E.g., on-policy learning estimates do-probabilities for the interventions specified by the policy.”
- Can we formalize and prove this claim?
 - E.g. $P(s' | s, a) = P(s' | s, do(a))$
- Is this equivalence true only for on-policy learning?



Correlation =? Causation in Different RL settings



Pearl's Ladder of Causation

- Different kinds of queries of increasing complexity
- Illustrated using queries about rewards

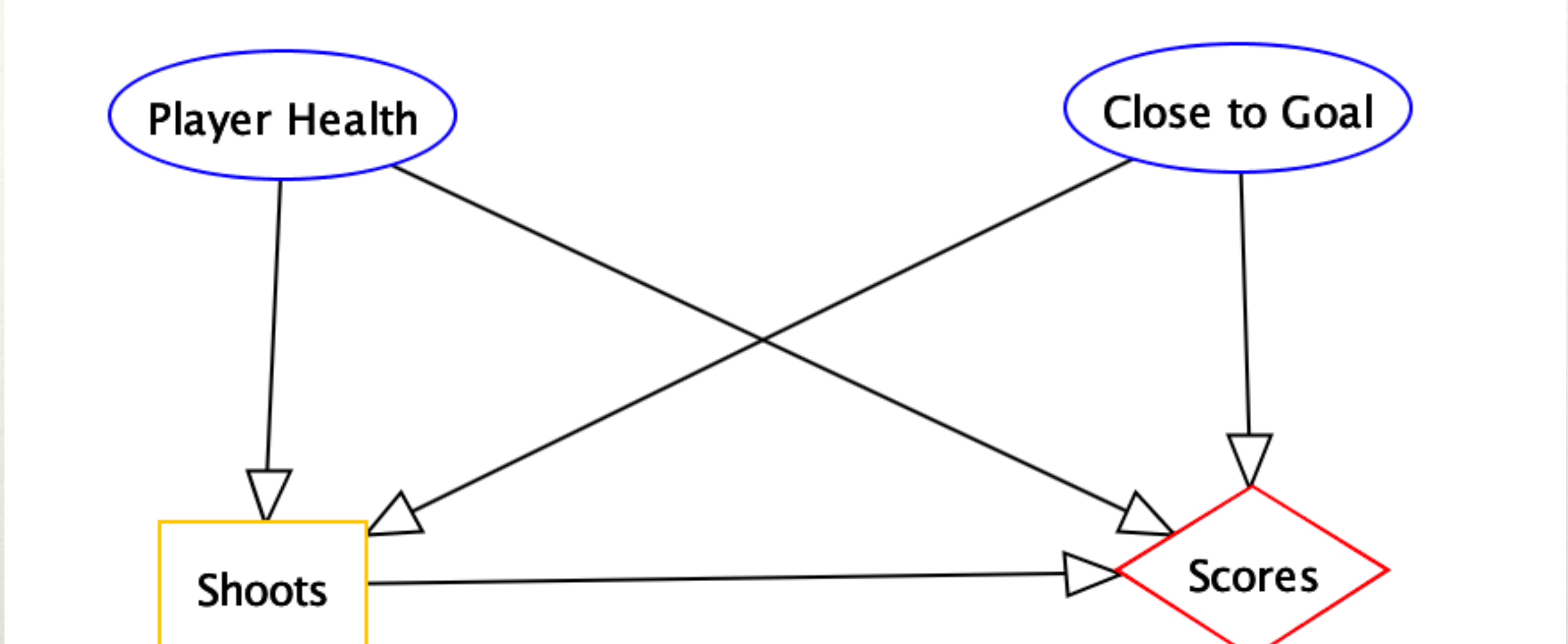
Level	Notation	Paraphrase	Example
Association/ Observation	$P(r \mid s, a)$	What reward follows after an agent chooses a ?	How often does a shot lead to a goal?
Intervention/ Action	$P(r \mid s, do(a))$	If I chose a , what will my reward be?	If I take a shot, will I score a goal?
Counterfactual	$P(r_a \mid s, b, r)$	How would my reward change if I had chosen a instead of b ?	I failed to score. What if I had taken a shot instead of making a pass?

Outline

1. Background I: Formal definition of do-probabilities
2. Background II: Confounded MDPs and off-policy evaluation (OPE)
3. Proposition: In online RL, conditional probabilities = interventional probabilities
4. Background III: Formal definition of counterfactual probabilities
5. Illustration: Even in online RL, (hindsight) counterfactual probabilities \neq conditional probabilities

Background I: Interventional Probabilities

Causal Models and Interventional Probabilities

- Do-probabilities are defined with respect to a *causal model*
 - Could be a causal Bayesian network or a structural causal model (function-based)
 - We start with causal BNs for simplicity / visualization
 - Specifically influence diagrams aka decision networks (Russell and Norvig 2010)
 - Demo
- 
- ```
graph TD; PH([Player Health]) --> S[Shoots]; PH --> Ss{Scores}; CTG([Close to Goal]) --> S; CTG --> Ss; S --> Ss;
```
- The diagram is an influence diagram with four nodes: 'Player Health' (blue oval), 'Close to Goal' (blue oval), 'Shoots' (yellow rectangle), and 'Scores' (red diamond). Arrows indicate dependencies: 'Player Health' points to 'Shoots' and 'Scores'; 'Close to Goal' points to 'Shoots' and 'Scores'; 'Shoots' points to 'Scores'.
- Reward: player scores iff shoots, is healthy, close to goal, goalie is not healthy
  - Policy: player shoots iff healthy and close to goal

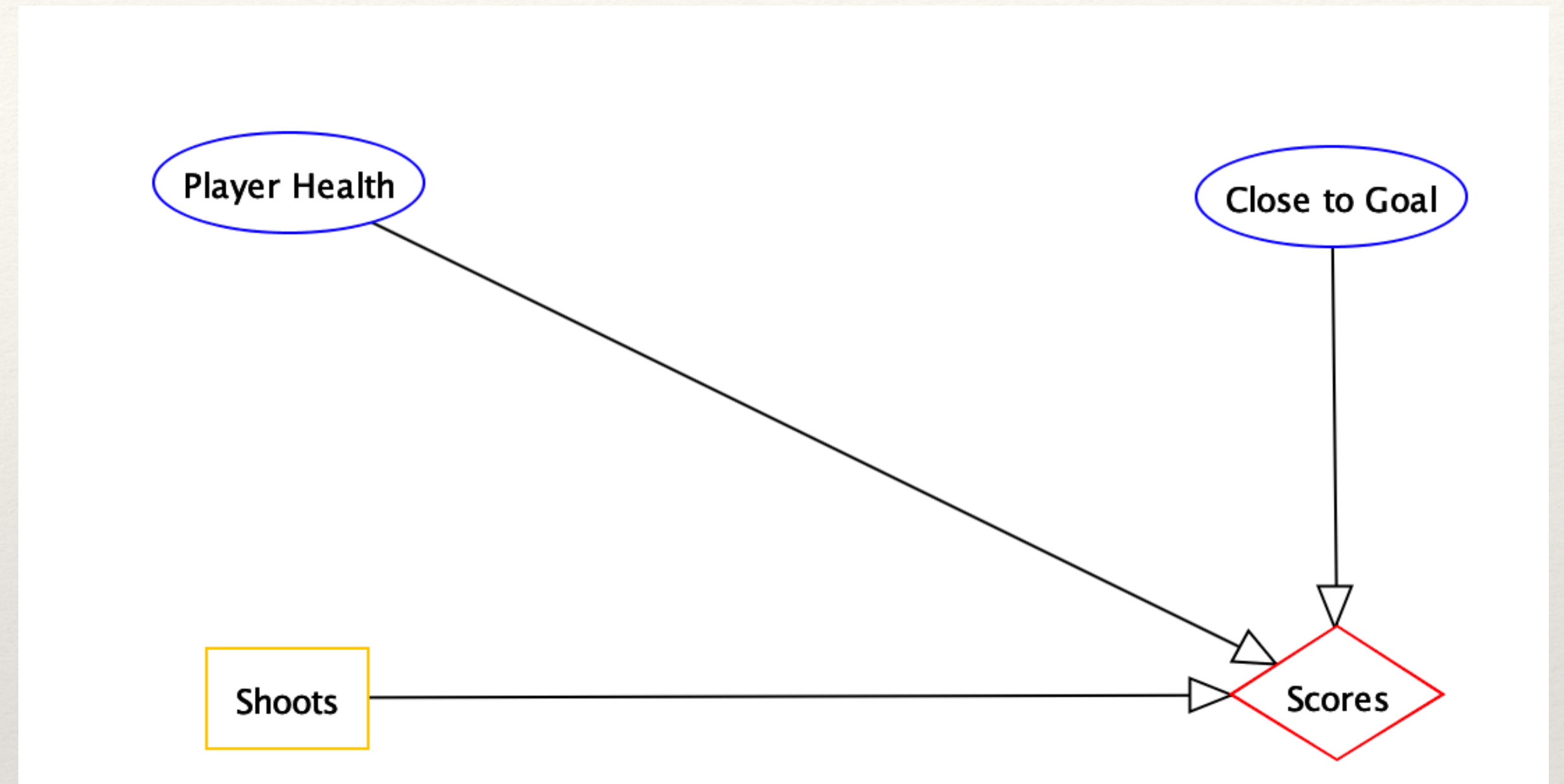


# Truncation Semantics for Do-Operator

- To evaluate  $P(Y|X, do(A = a))$ :
  1. Eliminate all links into  $A$
  2. Assign prior  $p(a) = 1$  to  $A$
  3. Evaluate  $P(Y|X)$  in the **truncated model**

➔ blocks inferences *from effect to cause*

- ➔  $P(Scores = T | Shoots = T, ClosetoGoal = T) = 1/2$   
 $< P(Scores = T | do(Shoots) = T, ClosetoGoal = T) = 1/4$





# Observability and Intervention

- Lemma: Suppose that  $X \supseteq \text{Parents}_A$ . Then  $P(Y | X, A) = P(Y | X, do(A))$ .
- Intuition: If we observe the parents and the child, it does not matter whether the parents are disconnected from the child.
- Example:  $P(\text{Scores} = T | \text{Shoots} = T, \text{CloseToGoal} = T, \text{PlayerHealth} = T) = 1/2$   
 $P(\text{Scores} = T | do(\text{Shoots}) = T, \text{CloseToGoal} = T, \text{PlayerHealth} = T) = 1/2$
- Significance: If the causes of an action are observable, then conditional probabilities = interventional probabilities



# Background II: Confounded MDPs and OPE



# On-policy and Online learning



- **Behavioral policy**  $\pi_\beta$ :  
route the agent has taken in the past
  - Generates data
- **Evaluation policy**  $\pi$ :  
alternative route to be evaluated

- **On-policy learning**:  $\pi = \pi_\beta$   
the policy is evaluated on data that it generated
- **Online**: the agent can interact with the environment
  - e.g., take a different route
- **Offline**: the agent observes but does not act
  - e.g., access driving logs



---

# Confounded MDPS

---

- Framework for studying off-policy evaluation (OPE) in the presence of confounders (Zhang and Bareinboim 2016; Bruns-Smith ICML 2021; Kausik et al. AISTATS 2024)
  - Confounder = unobserved common cause of action and next state / reward
  - MDP with state space  $\mathbf{S} = \mathbf{O} \times \mathbf{Z}$
  - *Behavioral policy*  $\pi_\beta : \mathbf{S} \rightarrow \Delta(A)$
  - *Evaluation policy*  $\pi : \mathbf{O} \rightarrow \Delta(A)$
- ➔ The evaluation policy has access to fewer inputs (observations)

Zhang, J. and Bareinboim, E. (2016), 'Markov decision processes with unobserved confounders: A causal approach', *Tech. Rep.*  
Bruns-Smith, D. A. (2021), Model-free and model-based policy evaluation when causality is uncertain, *in* 'ICML', pp. 1116–1126.  
Kausik, C., Lu, Y., Tan, K., Makar, M., Wang, Y. and Tewari, A. (2024), Offline policy evaluation and optimization under confounding, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 1459--1467.



# Offline/Off-policy Policy Evaluation

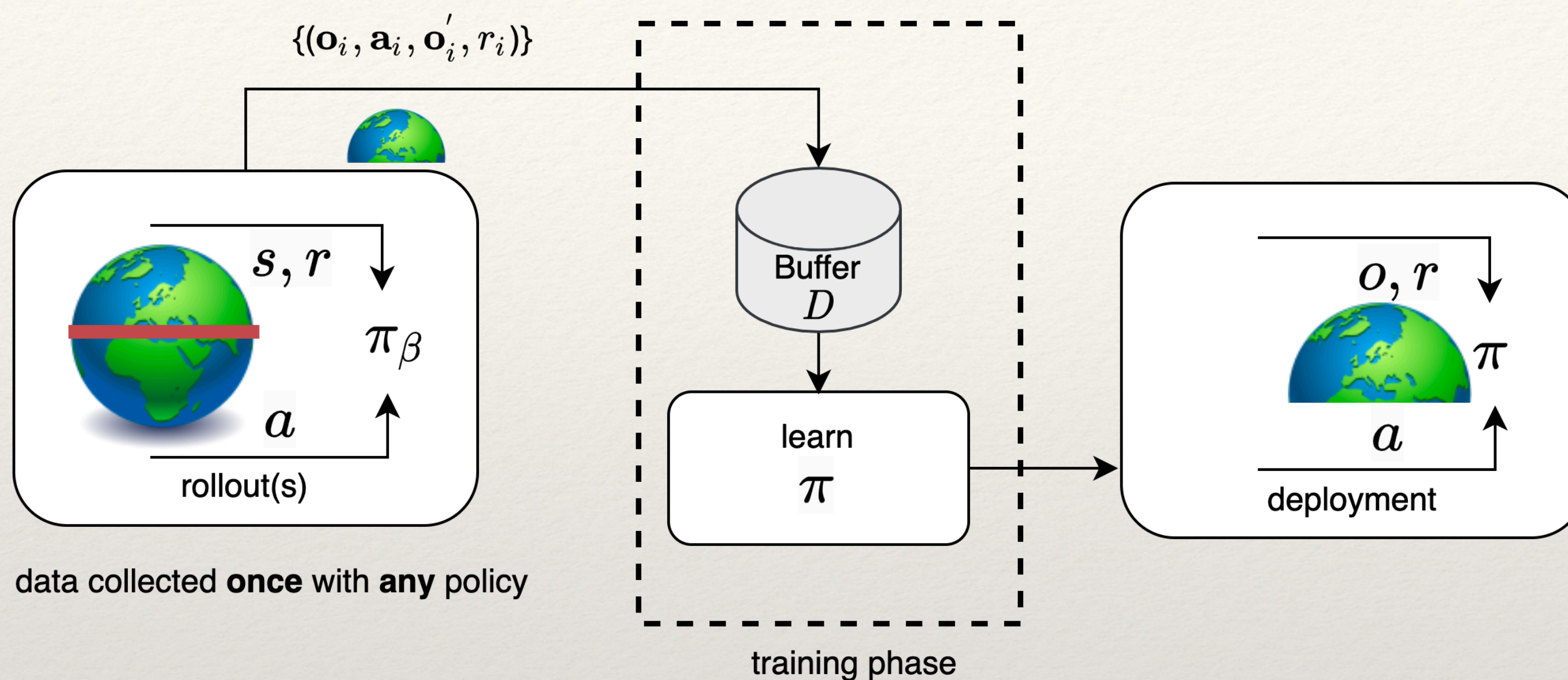


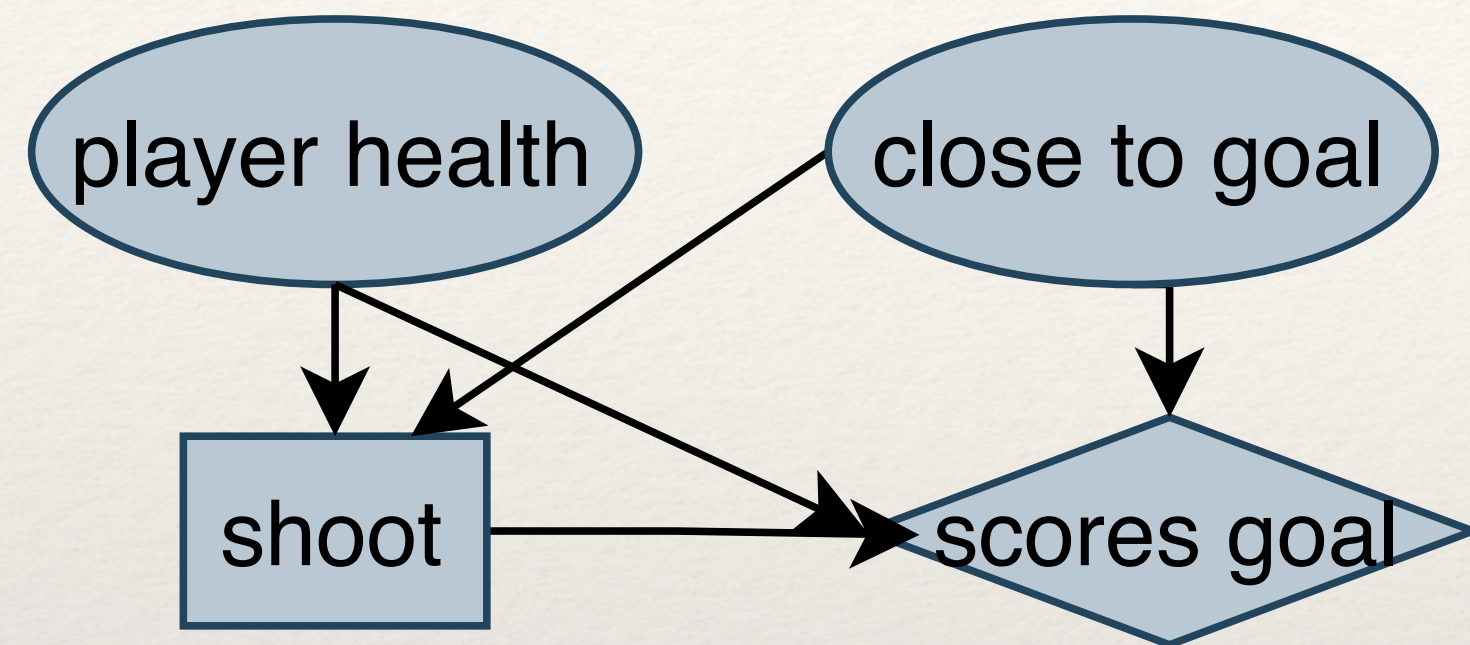
Figure adapted for  
confounded MDPs from  
Levine et al. 2020

- OPE: evaluate a learned policy from data generated by a (different) behavioral policy
- We want to evaluate the interventional value function  $Q(s, do(a))$  based on the do-operator (Zhang and Bareinboim 2020, Wang et al. 2021)



# Example

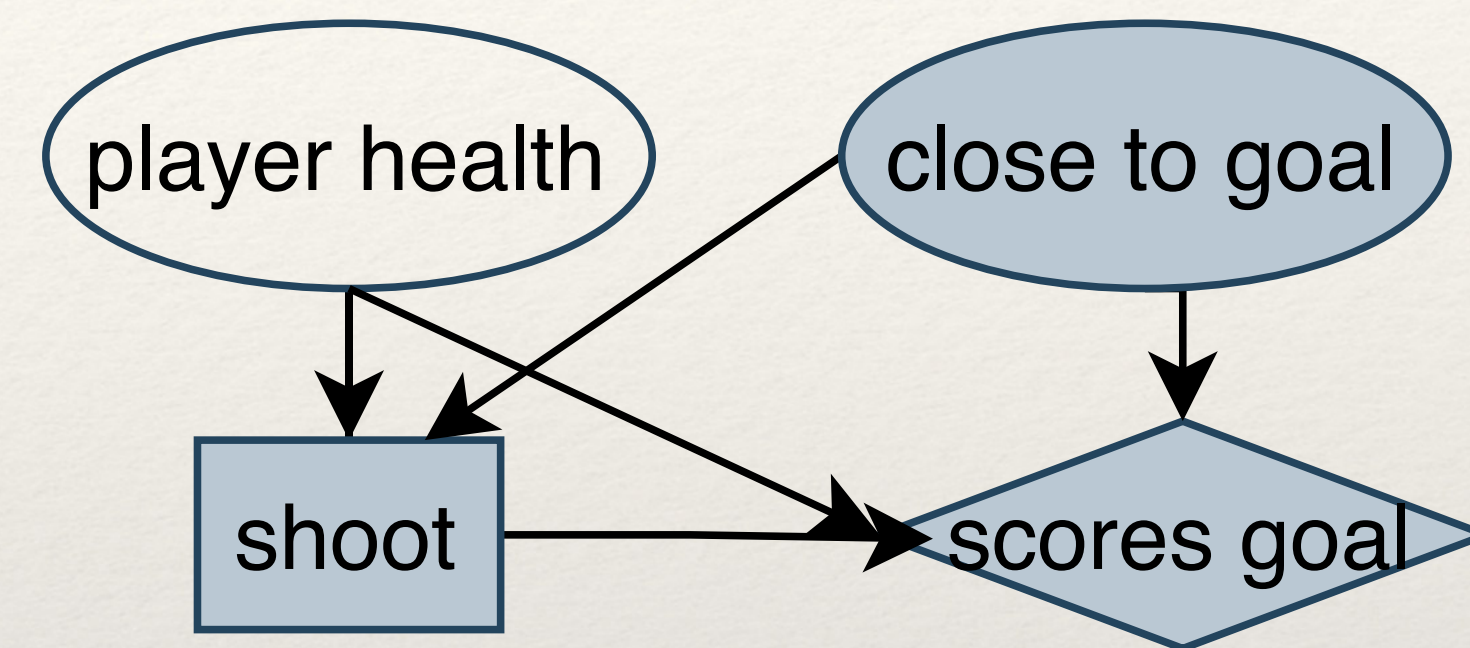
Online View



Behavioral Policy:

shoot if and only if close to goal and healthy

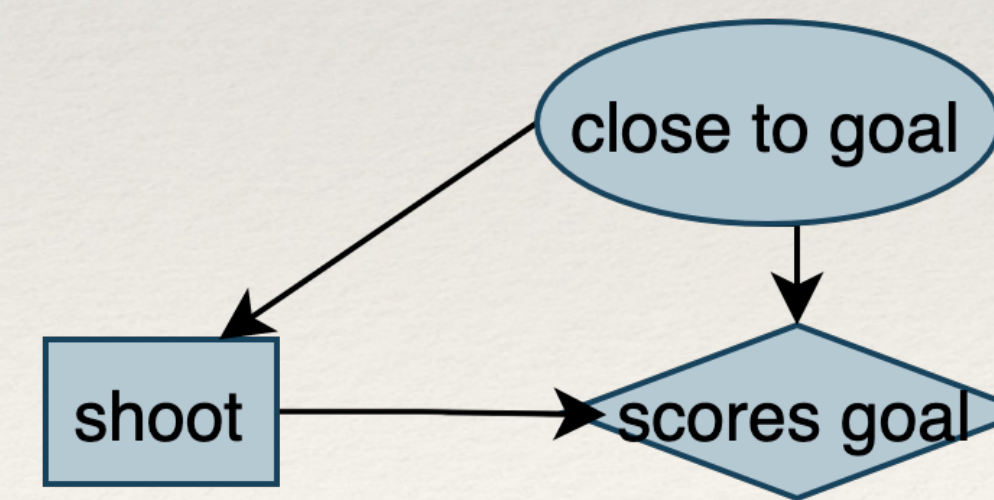
Offline View: Learning agent does not observe player health



(Marginal) Evaluation Policy:

$$\pi(\text{shoot} = T \mid \text{CloseToGoal} = T) = 1/2$$

Offline View

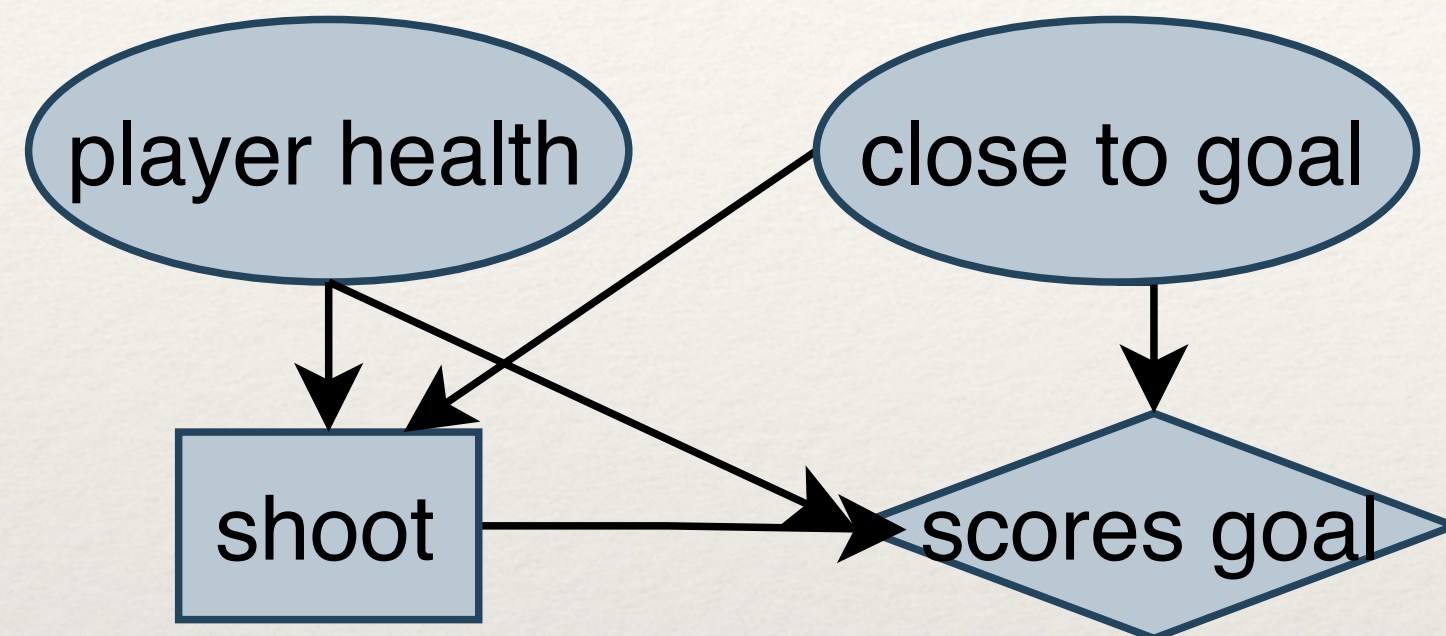


Marginal Graph  
with observable variables only



# Example: conditional rewards $\neq$ do-rewards

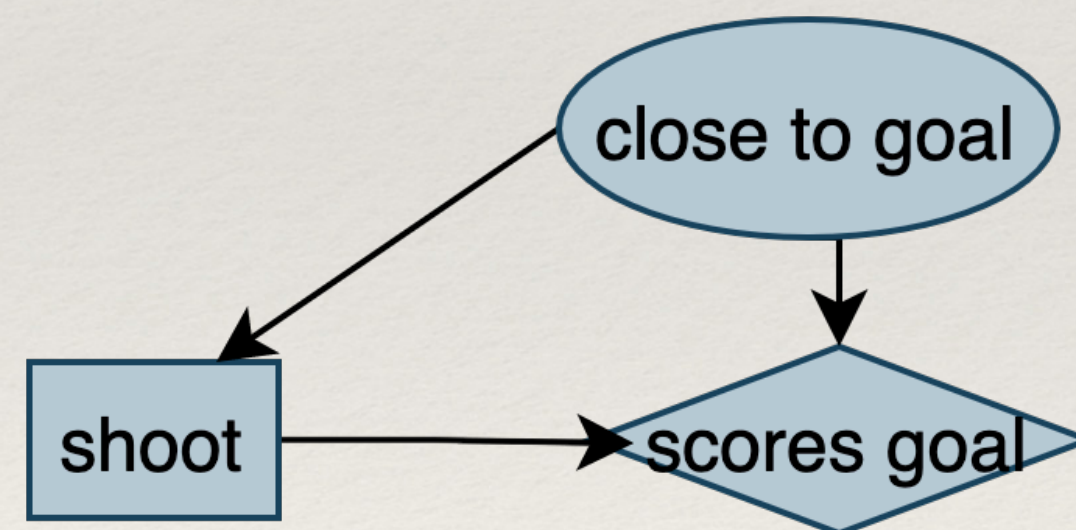
Online View



$$P(\text{Scores} = T \mid \text{Shoots} = T, \text{ClosetoGoal} = T, \text{PlayerHealth} = T) = 1/2$$

$$P(\text{Scores} = T \mid \text{do}(\text{Shoots}) = T, \text{ClosetoGoal} = T, \text{PlayerHealth} = T) = 1/2$$

Offline View



Conditional probability

$$P(\text{Scores} = T \mid \text{Shoots} = T, \text{ClosetoGoal} = T) = 1/2$$

$$< P(\text{Scores} = T \mid \text{do}(\text{Shoots}) = T, \text{ClosetoGoal} = T) = 1/4$$

True **interventional** probability

Marginal Graph  
with observable variables only



---

# Storytime

---

- Vancouver Canucks coach Rick Tocchet watches the Edmonton Oilers to learn from the best. He notices that whenever the Oilers shoot close to the goal, they score 50% of the time. So he directs the Canucks players to shoot whenever they get close. Tocchet is disappointed to find that the Canucks score only 25% of the time. “It must be that my players are worse than theirs” he thinks.
- Q: Is the coach right to blame his players?
- Answer: No. Because Tocchet did not observe the health of the Oilers players, he did not realize that they shoot only when they are healthy. His policy directs the Canucks players to shoot whether they are healthy or not, which leads to a lower success rate.



---

# Observability of Action Causes

---

- Why the difference between online and offline views?
- The key issue is whether *the causes of the behavioral agent's decisions are observable by the learning agent.*

## Proposition

Suppose that the observation signal  $\mathbf{O}$  of the learning agent includes the causes (parents) of the actions by the behavioral agent. Then

1.  $P(R | \mathbf{O}, A) = P(R | \mathbf{O}, do(A))$
2.  $P(S' | \mathbf{O}, A) = P(S' | \mathbf{O}, do(A))$
3.  $Q(\mathbf{O}, A) = Q(\mathbf{O}, do(A))$  [definitions in paper]



# Application to RL Settings

- The learning and behavioral agents are **observationally equivalent** if they share the same observation signal ( $\mathbf{O} = \mathbf{O}_\beta$ )
- Inputs are the same, policies may be different
- By proposition observation-equivalence  $\Rightarrow$  conditional probs = interventional probs

| Setting                | Observation-equivalent? | Reason                                    |
|------------------------|-------------------------|-------------------------------------------|
| On-policy              | ✓                       | Same policies                             |
| Online                 | ✓                       | Learning agent executes behavioral policy |
| Complete Observability | ✓                       | No latent variables<br>e.g., AlphaGo      |



# Counterfactuals



# The Ladder of Causation: Counterfactuals

| Level                       | Notation           | Paraphrase                                                      | Example                                                                 |
|-----------------------------|--------------------|-----------------------------------------------------------------|-------------------------------------------------------------------------|
| Association/<br>Observation | $P(r   s, a)$      | What reward follows after an agent chooses $a$ ?                | How often does a shot lead to a goal?                                   |
| Intervention/<br>Action     | $P(r   s, do(a))$  | If I chose $a$ , what will my reward be?                        | If I take a shot, will I score a goal?                                  |
| Counterfactual              | $P(r_a   s, b, r)$ | How would my reward change if I had chosen $a$ instead of $b$ ? | I failed to score. What if I had taken a shot instead of making a pass? |

- So far: Level 2
- Next: Level 3



---

# Structural Causal Models

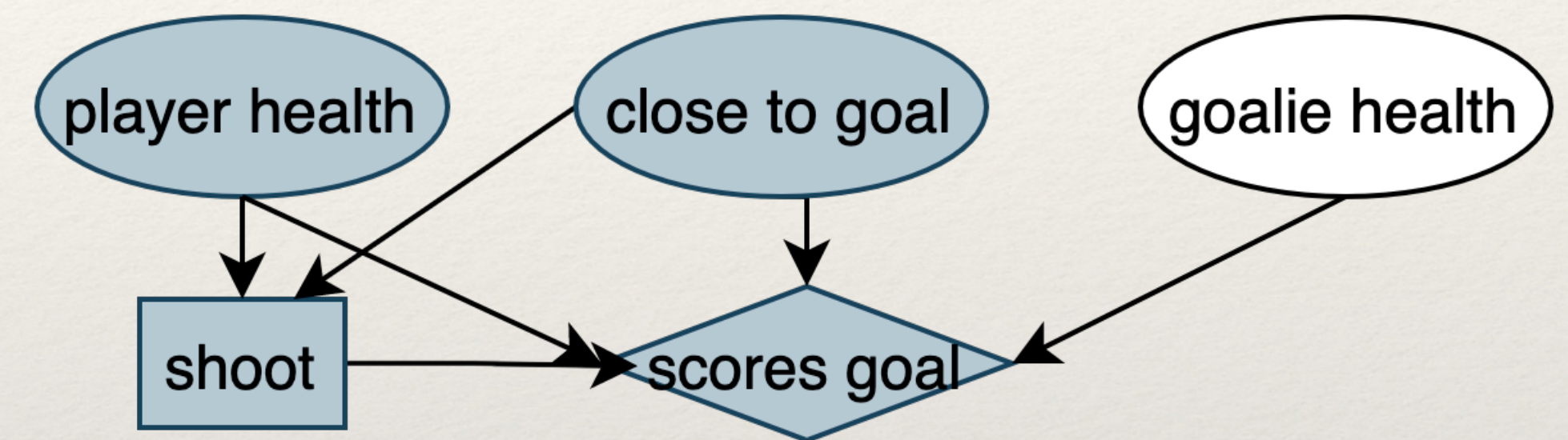
---

- Causality theory defines a formal semantics for counterfactuals such as “How would my reward change if I had chosen action  $a$  instead of  $b$ ?”
- Based on **structural causal models** (SCMs)
- An SCM  $(F, b)$  parametrizes a causal graph with
  1. *Deterministic functions*  $\text{child} = f(\text{parents})$
  2. A prior distribution  $b(U)$  over source variables  $U$
- Typically requires introducing new latent source variables (noise terms, background variables)
- Think: local decoders



# Example SCM

| Variable | Function                                 |
|----------|------------------------------------------|
| Shoot    | $SH = PH \cdot CG$                       |
| Scores   | $SC = SH \cdot PH \cdot CG \cdot (1-GH)$ |



Uniform prior over 3 source variables



---

# Evaluating Counterfactuals

---

- Compute  $P(\textcolor{red}{Y}_{\textcolor{blue}{a}^*} | X, \textcolor{red}{Y}, A)$  as follows given an SCM  $(F, b)$ 
  1. *Belief update/Abduction*: let  $b' = b(U | X, Y, A)$
  2. *Intervene/Truncate*: Remove all parents from  $A$ , set  $A = a^*$ ; defines  $F'$
  3. *Predict*: Return  $P(Y | X)$  in updated SCM  $(F', b')$
- The belief update is analogous to belief updates in a belief MDP [see paper]

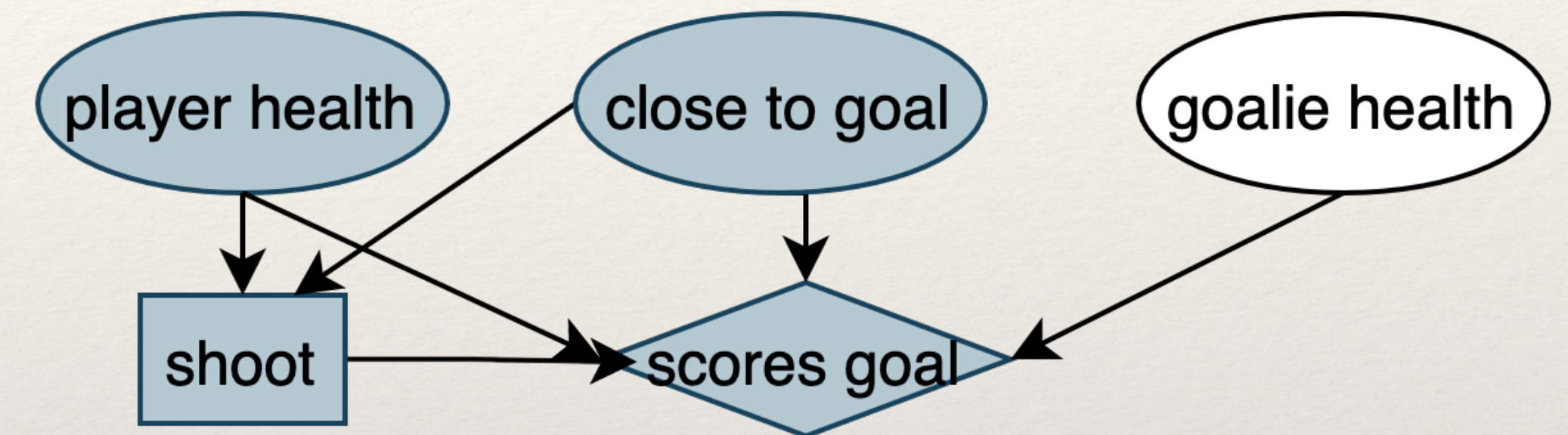


# Example Evaluation

$$P(\textcolor{red}{SC}_{\textcolor{blue}{SH}=1} \mid CG = 1, PH = 1, SH = 1, \textcolor{red}{SC} = 1)$$

| Variable | Function                                   |
|----------|--------------------------------------------|
| Shoot    | $SH = 1$                                   |
| Scores   | $SC = SH \cdot PH \cdot CG \cdot (1 - GH)$ |

$$b(GH = 0 \mid SC = 1) = 1$$



Given that the player scored, we know in hindsight that the goalie was not healthy

$$\begin{aligned} &P(SC_{SH=1} = 1 \mid CG = 1, PH = 1, SH = 1, SC = 1) \\ &= P(SC = 1 \mid CG = 1, PH = 1, GH = 0, do(SH = 1)) = 1 \end{aligned}$$



# More interesting example

Look at outcomes 1 time step ahead (Harutyunyan et al. 2019)

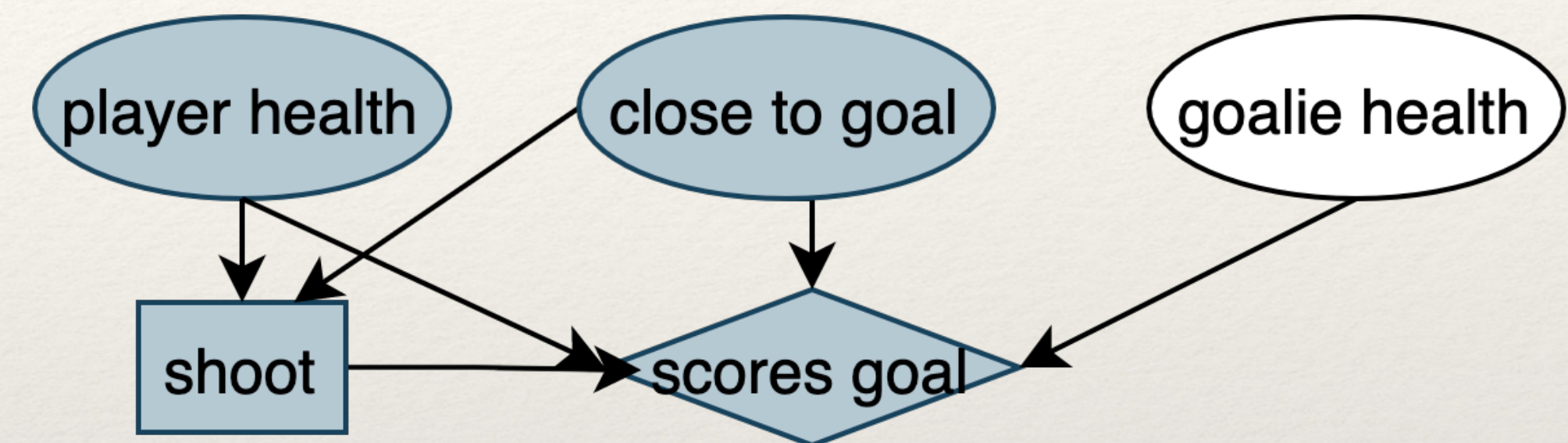
$$P(\textcolor{red}{SC}_{SH=1} \mid CG = 1, PH = 1, SH = 0, \textcolor{red}{SC}' = 1)$$

“The player did not shoot, then the team scored.  
If they had shot, would they have scored?”

$$\begin{aligned} &P(SC_{SH=1} = 1 \mid CG = 1, PH = 1, SH = 0, SC' = 1) \\ &= P(SC = 1 \mid CG = 1, PH = 1, GH = 0, do(SH = 1)) = 1 \end{aligned}$$

The player would have scored

$$b(GH = 0 \mid SC' = 1) = 1$$



Given that the team scored, we know in hindsight that the goalie was not healthy



---

# Hindsight Counterfactuals and Online Learning

---

- Hindsight counterfactuals are different from conditional probabilities even in online learning, e.g.
  - knowing that the team scored, the probability of scoring after a shot is 1.
  - Without hindsight, we do not know the state of the goalie's health, so the scoring probability is at most  $1/2$ .
- Since future outcomes are not known at decision time, it is not clear what the use case for hindsight counterfactuals is.
- Interesting suggestion (Sun et al. AAAI 2024): Use hindsight counterfactuals to generate *virtual transitions* for data augmentation.
  - Like roll-outs in model-based RL
- Insight: Both past observations **and** future observations allow us to infer a *current* latent state



---

# Related Work

---

- Related Work section in our paper discusses previous causal modelling in RL with regard to the online / offline / hybrid settings.
- Issues include state abstraction, behavioral cloning, causality-based exploration.
- Especially exciting prospect for future work: hybrid offline+online setting (Gasse et al. 2021, Bareinboim et al (2020) Tutorial)
- Can leverage large offline data to build a causal model (Geffner et al. 2022, Sun and Schulte 2023) then refine with online learning / experimentation.

Gasse, M., Grasset, D., Gaudron, G. and Oudeyer, P.-Y. (2021), 'Causal reinforcement learning using observational and interventional data', *arXiv preprint arXiv:2106.14421*.

Geffner, T., Antoran, J., Foster, A., Gong, W., Ma, C., Kiciman, E., Sharma, A., Pawlowski, N. (2022), 'Deep End-to-end Causal Inference', *arXiv preprint arXiv:2202.02195*.

Sun, X. and Schulte, O. (2023), Cause-Effect Inference in Location-Scale Noise Models: Maximum Likelihood vs. Independence Testing, *in* 'Advances in Neural Information Processing Systems'.



---

# Conclusion

---

- Basic Question: in which RL settings are conditional probabilities = interventional probabilities?
- Answer: when the learning agent can observe the causes of actions (inputs) by the behavioral agent (observational equivalence)
- ➔ Covers online learning, on-policy learning, complete observability
- Hindsight counterfactuals  $\neq$  conditional probabilities even under observational equivalence
- Related / future work on offline off-policy causal RL under partial observability



---

# Thank you for your attention

---

Arxiv paper



[https://arxiv.org/abs/  
2403.04221v1](https://arxiv.org/abs/2403.04221v1)