UNIVERSITAT DE BARCELONA

CAUSAL INFERENCE AND MACHINE LEARNING

About the course

01 **Introduction**

Observational and Interventional Distributions

02 **Potential Outcomes**

Fundamental Problem of Causal Inference

03 **Causal Graphs**

Do Calculus

04 **Estimand-based Estimation**

Metalearners

05 **Estimand-agnostic Estimation**

Counterfactuals

06 **Causal Machine Learning**

Supervised and Reinforcement Learning

07 **Practical Causal Inference**

Exercises

The relationship between causality and artificial intelligence can be seen from two points of view: how causality can help solve some of the current problems of AI and how causal inference can leverage machine learning techniques. In this course we will review the two points of view with special emphasis on examples and practical cases.

Universitat de Barcelona

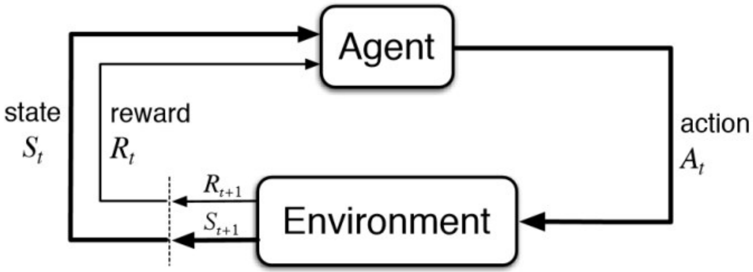
Causal Machine Learning

Reinforcement Learning

Enrique Mora
enrique.mora@es.nestle.com
UNIVERSITAT DE BARCELONA

Causal ML. Reinforcement Learning

t	discrete time step
T	final time step of an episode t
A_t	action at time t
$\mathbf{S}_t$	state at time t
L_t	regret/loss at time t
R_t	reward at time t
$\mathcal{R}$	return
π	policy (decision-making rule)
$\pi(a \mid s)$	probability of taking action a in state s
s, s'	true states
x, x'	observed states
$v_\pi(s)$	value of state s under policy π (expected return)
$q_\pi(s, a)$	value of taking action a in state s under policy π
τ	trajectory, i.e., $\tau = \{x_t, a_t, x_{t+1}\}_{t=1}^{T}$



•

Causality + RL = Causal RL

RL: focused on building algorithms to *maximise rewards*

- *Using synthetic data simulators*
- *Able to generate large amounts of data*

Causality: focused on the *identifiability* and *inferences* based on given *causal structure*

- *Typically given a limited-size observation dataset*
- *From an unknown environment and policy*
- *We cannot interact with the environment online*

Off-line RL: learning optimal policies from a dataset generated from an unobserved policy.

- *Learn from observational data*
- *Without access to the environment*

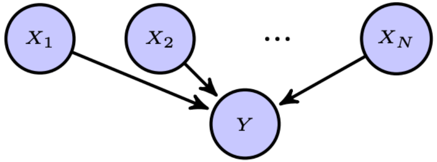
Causal RL

Problem	Output	Benefits over non-causal RL
Causal Bandits	$\hat{\pi} = \arg\min_{\pi \in \Pi} L_n(\pi)$	Optimal simple regret guarantees
Model-Based RL	$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell\left(\boldsymbol{\theta}, (R_{t+1}, S_{t+1})\right)$	Deconfounding
Multi-Environment RL	$\hat{\pi} = \arg\max_{\pi \in \Pi} \mathbb{E}_{c \sim p(c)} \left[\mathcal{R}\left(\pi, \mathcal{M}^c\right)\right]$	Interpretable task embeddings, systematic generalization
Off-Policy Policy Evaluation	$\hat{v}_\pi(s) = \mathbb{E}_{\boldsymbol{x} \sim d_0} \left[\sum_{t=0}^{T-1} \gamma^t r_t \mid \boldsymbol{x}_0 = \boldsymbol{x}\right]$	Deconfounding
Imitation Learning	$\hat{\pi} = \arg\min_{\pi \in \Pi} \mathbb{E}_{\boldsymbol{x} \sim d_{\pi^*}} \left[\ell\left(\boldsymbol{x}, \pi, \pi^*\left(\boldsymbol{x}\right)\right)\right]$	Deconfounding
Credit Assignment	$\mathcal{M}_{a_t \to r_{t+k}}$ or $\mathcal{M}_{a_t \to s_{t+1}}$ or $\mathcal{M}_{a_t^i \to a_t^j}$	Intrinsic reward, Data-efficiency
Counterfactual Data Augmentation	$\tilde{\tau} = \{\tilde{\boldsymbol{x}}_t, \tilde{a}_t, \tilde{\boldsymbol{x}}_{t+1}\}_{t=1}^T$	Data-efficiency
Agent Incentives	Incentive criteria and measures	Avoiding unintended harmful behavior

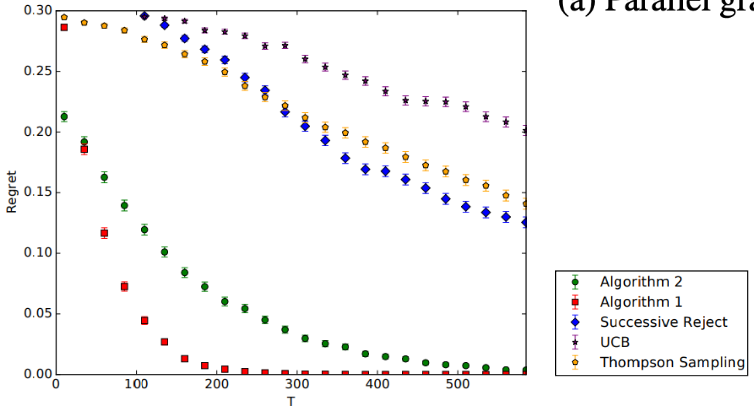
Causal Bandits

Algorithm 1 Parallel Bandit Algorithm

Input: Total rounds T and N.
for $t \in 1, \dots, T/2$ **do**
  Perform empty intervention $do()$
  Observe $\boldsymbol{X}_t$ and Y_t
for $a = do(X_i = x) \in \mathcal{A}$ **do**
  Count times $X_i = x$ seen: $T_a = \sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = x\}$
  Estimate reward: $\hat{\mu}_a = \frac{1}{T_a} \sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = x\} Y_t$
  Estimate probabilities: $\hat{p}_a = \frac{2T_a}{T}$, $\hat{q}_i = \hat{p}_{do(X_i=1)}$
Compute $\hat{m} = m(\hat{\boldsymbol{q}})$ and $A = \left\{a \in \mathcal{A} \colon \hat{p}_a \le \frac{1}{\hat{m}}\right\}$.
Let $T_A := \frac{T}{2|A|}$ be times to sample each $a \in A$.
for $a = do(X_i = x) \in A$ **do**
  for $t \in 1, \dots, T_A$ **do**
    Intervene with a and observe Y_t
  Re-estimate $\hat{\mu}_a = \frac{1}{T_A} \sum_{t=1}^{T_A} Y_t$
return estimated optimal $\hat{a}_T^* \in \arg\max_{a \in \mathcal{A}} \hat{\mu}_a$



(a) Parallel graph



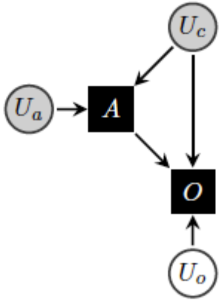
"before the agent takes the next action, it observes further samples for all non-intervened variables"

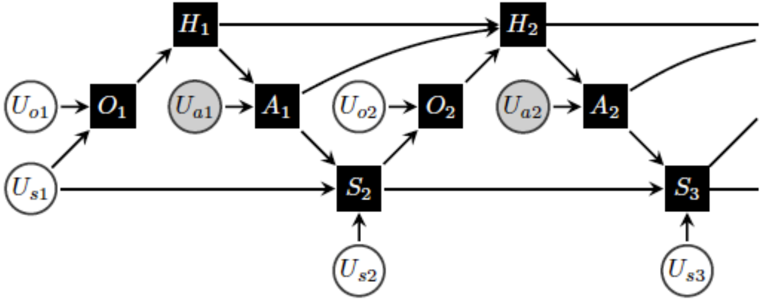
[Source]

Off-Policy Policy Evaluation

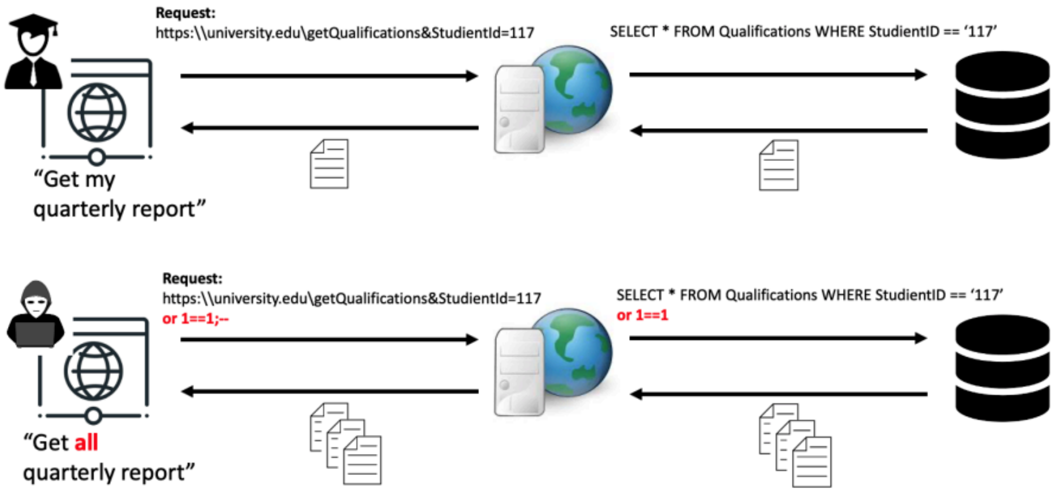
Off-policy policy evaluation (OPPE) is a problem in reinforcement learning where the goal is to evaluate a given policy (evaluation policy) using data generated by a different one (behaviour policy).

The off-policy evaluation problem is challenging because the data generated by the behaviour policy may not be representative of the target policy, leading to bias in the estimates.





Example. SQL Injection Attack



Example. RL as Hacker

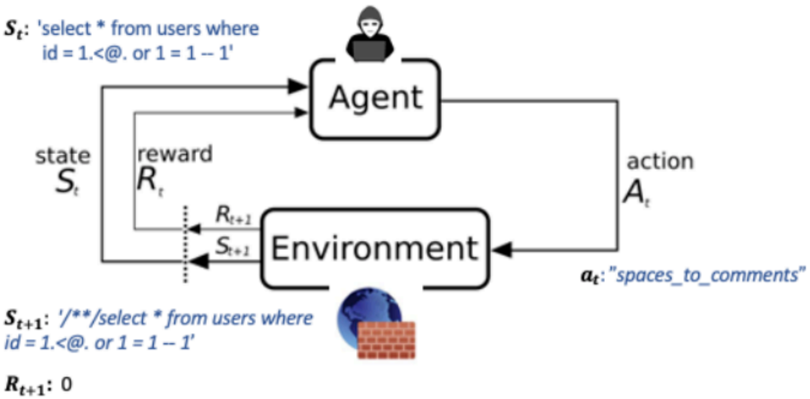
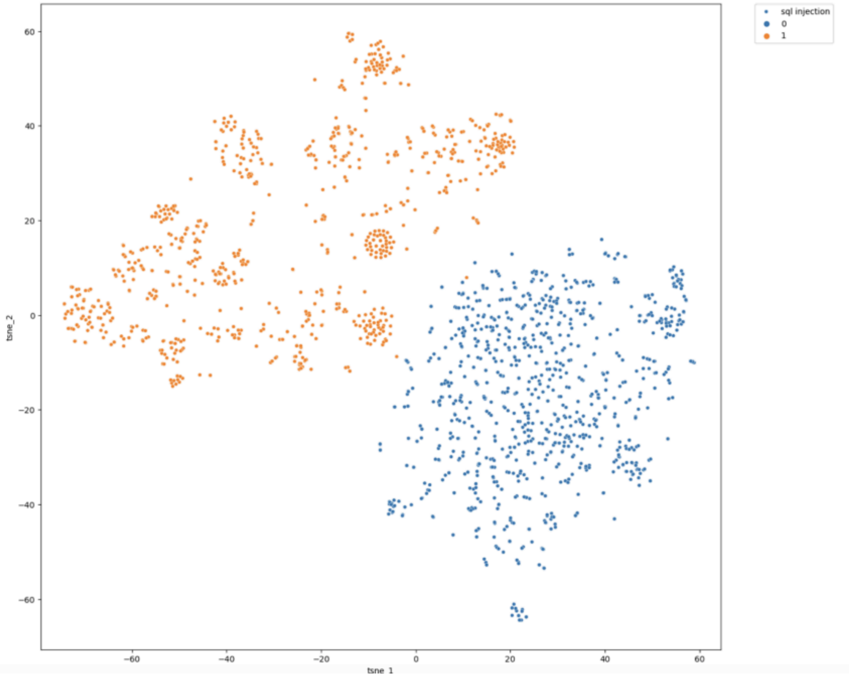


Table 1. Example of mutations

Mutation	Example
Case Swapping	admin OR 1=1# ⇒ admin oR 1=1#
Whitespace Substitution	admin OR 1=1# ⇒ admin\t\rOR\n1=1#
Comment Injection	admin OR 1=1# ⇒ admin ** \OR 1=1#

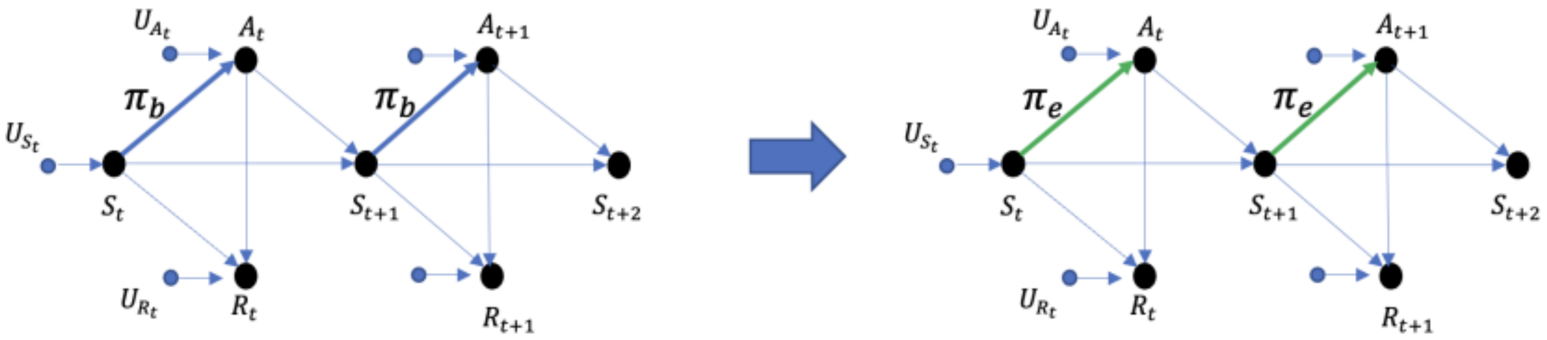
Epoch	min reward	max reward	avg. regard	avg. episode length
1	-21.40	-5.01	0.00	19.60
2	-16.36	-4.29	0.00	15.88
3	-18.75	-3.43	0.00	12.07
7	-18.75	-1.51	0.00	7.04

Example. RL as Hacker



Model	Accuracy
RoBERTa Base	99,75%
RoBERTa Fine-tuned (SQL)	99,75%
RoBERTa Custom tokenizer	98,25%

Example. Causal problem: Soft-intervention



Example. Estimators

$$V_{DM}^{\pi}(s) = E_n[\sum_{a \in A} \pi_e(a|s_0^{(i)})q(s_0^{(i)}, a; \phi)] = n^{-1} \sum_{i=0}^{N} \sum_{a \in A} \pi_e(a|s_0^{(i)})q(s_0^{(i)}, a; \phi)$$

$$V_{IS}^{\pi_e} = E_n[\omega_{0:T-1} \sum_{t=0}^{T-1} \gamma^t r_t] = n^{-1} \sum_{i=0}^{N} \omega_{0:T-1} G_i$$

$\omega_{0:T-1} = \prod_{t=0}^{T-1} \pi_e(a_t|s_t)/\pi_b(a_t|s_t)$

Example. Results

Measure	Value
Theoretical Maximum Reward	0.0
Real V_{π_e} (ppo agent)	-1.2624
Real V_{π_b} (random agent)	-6.2203
Avg. Estimated V_{π_e} (10 experiments of 200 episodes)	-4.5136
RSME estimated V_{π_e} (10 experiments of 200 episodes)	1.1320
STD estimated V_{π_e} (10 experiments of 200 episodes)	0.1580

Measure	Value
Theoretical Maximum Reward	0.0
Real V_{π_e} (ppo agent)	-1.2624
Real V_{π_b} (random agent)	-6.2203
Estimated V_{π_e} (10 experiments of 200 episodes)	1.0745
RSME estimated V_{π_e} (10 experiments of 200 episodes)	1.0987
STD estimated V_{π_e} (10 experiments of 200 episodes)	0.1423