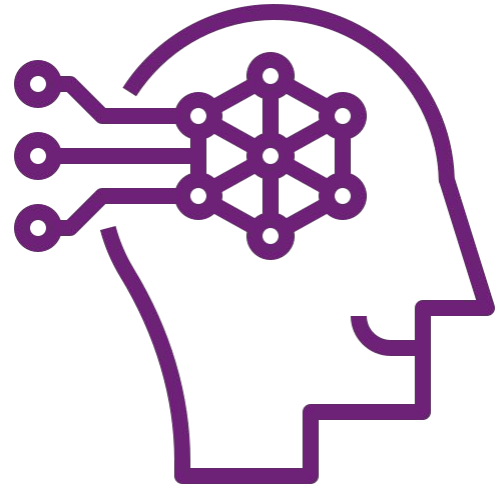


Causality and the Machine Learning

limitations for
predictive tasks

Luis Moneda



About me

Work

- Data Science Manager at Nubank

Education

- MSc Computer Science student (IME-USP)
- Bachelor in Computer Engineering (Poli-USP)
- Bachelor in Economics (FEA-USP)

Twitter - @lgmoneda

LinkedIn

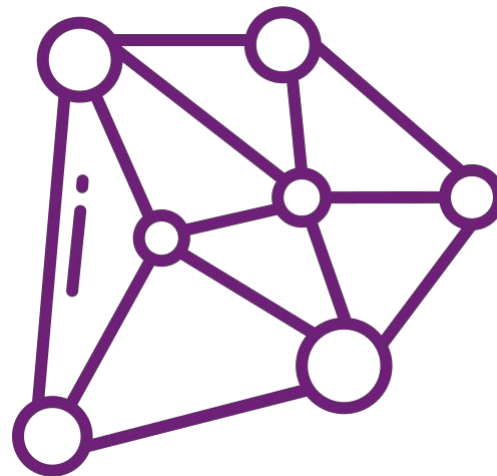
E-mail

Blog: lgmoneda.github.io/

Datasets (Kaggle)

Outline

- 1 - Data Science Tasks
- 2 - Empirical Risk Minimization
- 3 - ML limitations
- 4 - Independent Causal Mechanism Principle
- 5 - Invariant Causal Prediction (ICP)
- 6 - Invariant Risk Minimization (IRM)
- 7 - Reading recommendations



Data Science Tasks

Data Science Tasks

Description	Prediction	Causal Inference
<ul style="list-style-type: none">- Computing proportions- Aggregation metrics- Clustering- Visualizations	Mapping inputs (X) to output y.	Using data to calculate certain feature of the world if the world had been different: counterfactual prediction.

Data Science Tasks - Examples

Description	Prediction	Causal Inference
<ul style="list-style-type: none">- What proportion of women aged 60-80 years had a stroke last year?	<p>What is the probability of having a stroke next year for women with certain characteristics?</p>	<p>Will taking the drug A reduce, on average, the risk of stroke in women with certain characteristics?</p>

Data Science Tasks - Confusion Matrix

		What you want		
		Description	Prediction	Causal Inference
Approach you're using	Description	You're able to provide a snapshot of your data.	Nice benchmark, poor performance.	Misleading results, bad decisions.
	Prediction	Why predict if you have the actual?	Low error predictions.	Biased estimations.
	Causal Inference	Cost ineffective, harder, overkill.	Cost ineffective, not the best performance.	Unbiased estimation of actions' effects.

Data Science Tasks - Confusion Matrix

		What you want		
		Description	Prediction	Causal Inference
Approach you're using	Description	You're able to provide a snapshot of your data.	Nice benchmark, poor performance.	Misleading results, bad decisions.
	Prediction	Why predict if you have the actual?	Low error predictions.	Biased estimations.
	Causal Inference	Cost ineffective, harder, overkill.	Cost ineffective, not the best performance.	Unbiased estimation of actions' effects.

Empirical Risk Minimization

Supervised Learning summarized

$$X \xrightarrow{f} y$$

- Statistical Learning theory
- Empirical Risk Minimization
- Independently identically distributed (iid)
- We want to predict things nicely, we don't care about what is the f
- If your model predicts well a random subset of X , it is generalizing

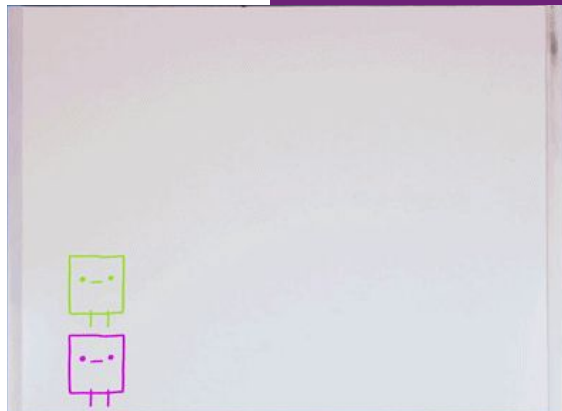
Prediction

Most of successful applications today in DS are merely predictive!

Why?

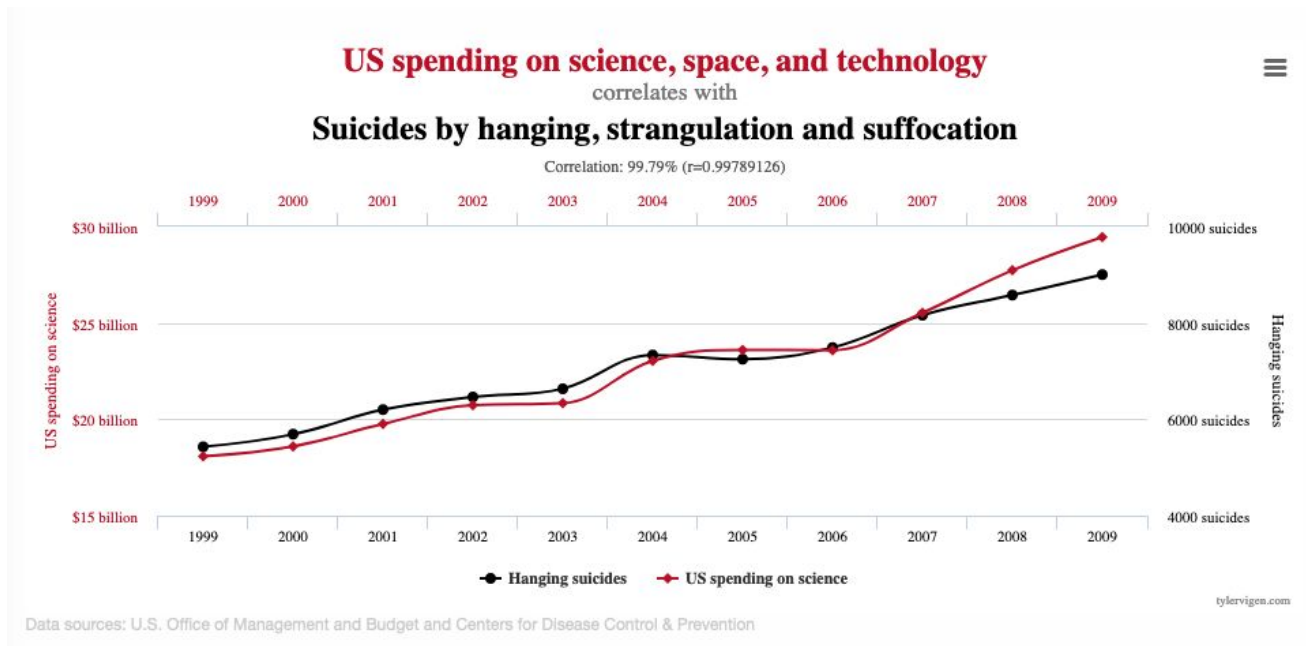
- 1) A large dataset with inputs and outputs;
- 2) An algorithm that establishes a mapping between inputs and outputs;
- 3) A validation strategy following the learning paradigm
- 4) A metric to assess the performance of the mapping!

All the information required is in the data!

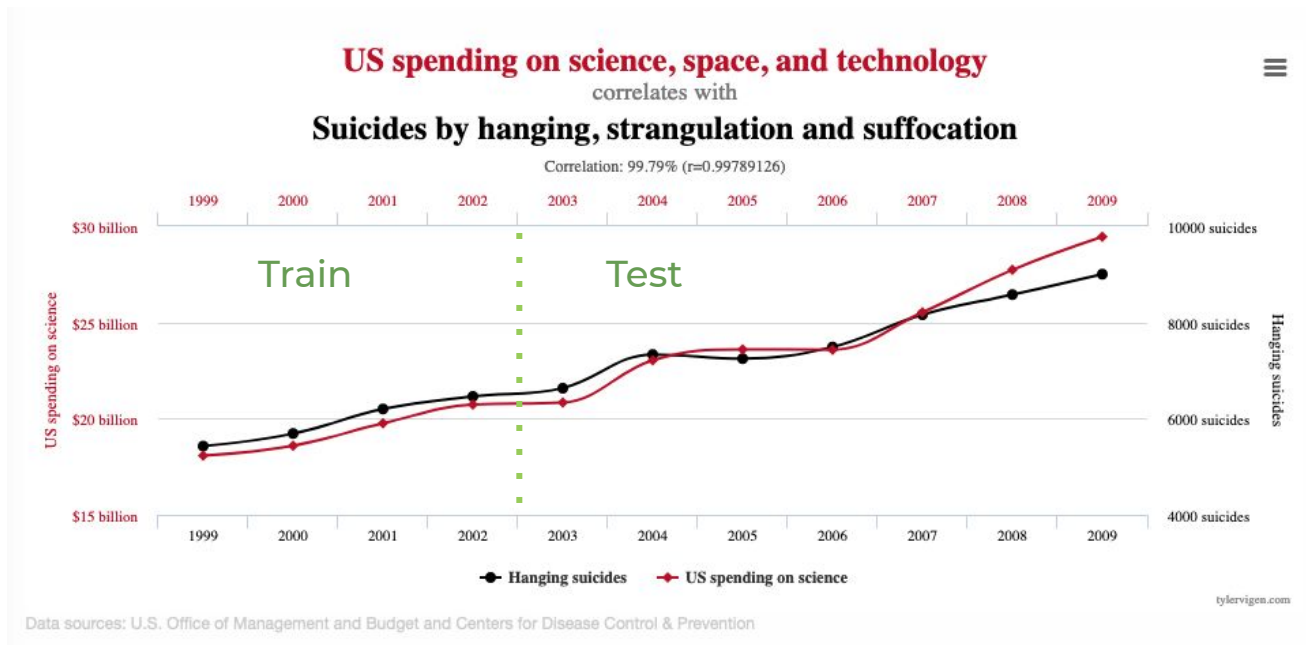


Predictive Machine Learning Limitations

Spurious correlation



Spurious correlation



Spurious correlation

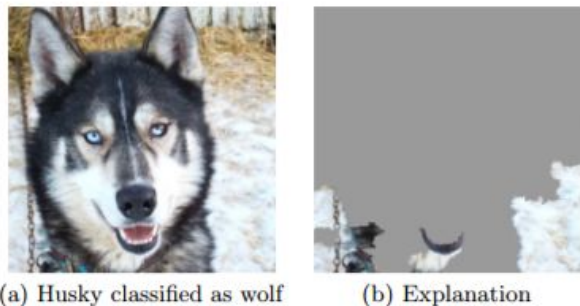


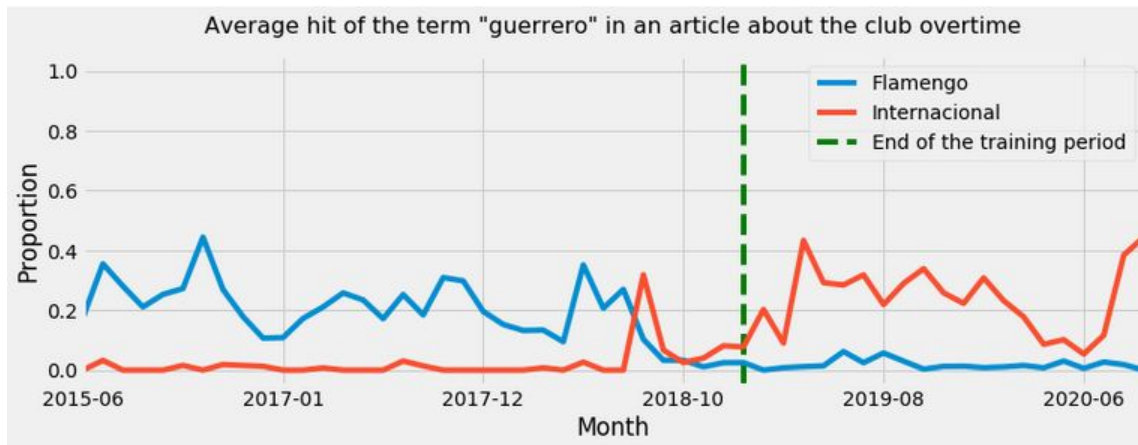
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Ribeiro et al. 2016. "Why Should I Trust You?"
Explaining the Predictions of Any Classifier. KDD16.

Spurious correlation



Spurious correlation

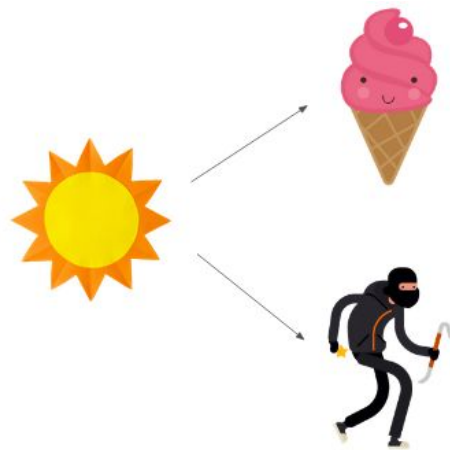
Spuriousness has a long story...

Pearson (1897): "Causation is correlation, except when correlation is spurious, when correlation is not causation."

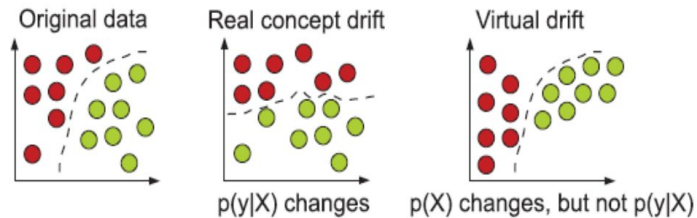
Yule (1926): "...if we had or could have experience of the two variables over a much longer period of time we could not find any appreciable correlation between them"

Reichenbach (1956): Common cause principle.

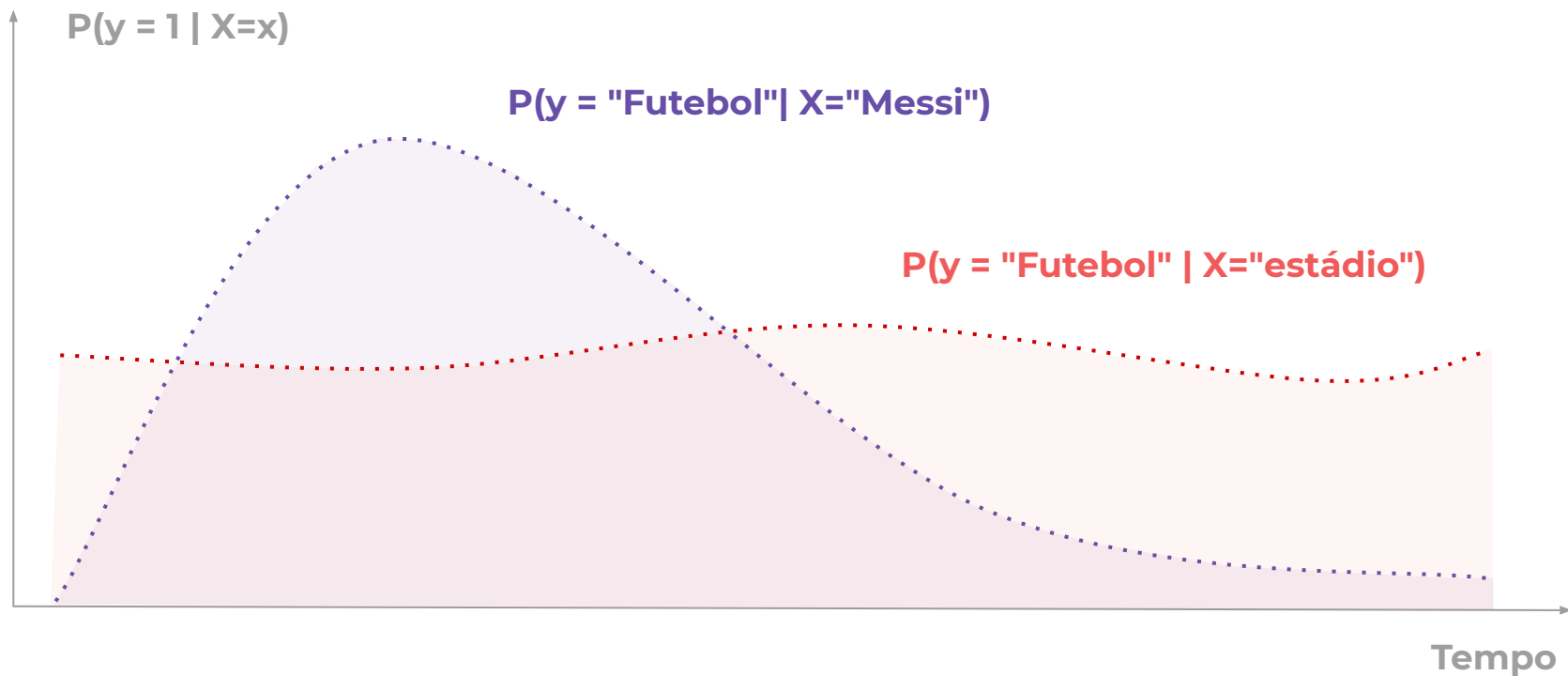
Pearl (1990s): Spuriousness when there's a backdoor path in the causal graph.



Dataset shift



Concept drift



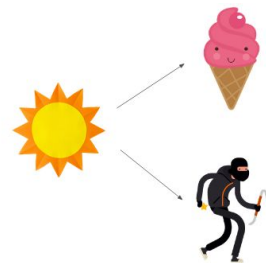
Model underspecification

If you want to predict ice cream consumption with the following decomposition:

$$P(\text{Ice Cream} \mid \text{Weather, Crime})P(\text{Crime} \mid \text{Weather})P(\text{Weather})$$

You get unnecessarily exposed to model degradation due to:

- Virtual / covariate shift in $P(\text{Weather})$
- Real / concept drift in $P(\text{Crime} \mid \text{Weather})$

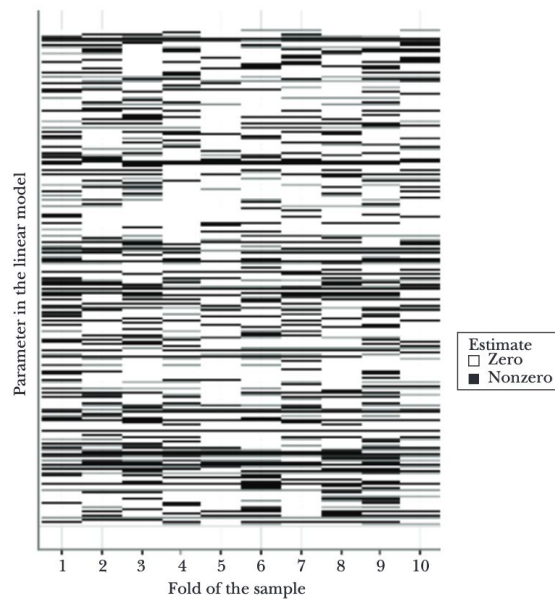


While the real model is simply: $P(\text{Ice Cream} \mid \text{Weather})P(\text{Weather})$. The only change you're exposed is a concept drift for $P(\text{Ice Cream} \mid \text{Weather})$.

"An ML pipeline is underspecified when it can return many predictors with equivalently strong held-out performance in the training domain."

Model underspecification

Selected Coefficients (Nonzero Estimates) across Ten LASSO Regressions





**Why is it interesting to have a
graphical probabilistic models
background?**

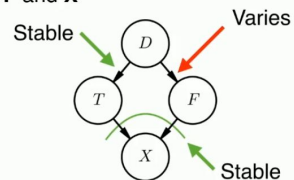
Many suggested solutions use PGM language

Famous researchers are turning their attention to causality also, like **Yoshua Bengio** and **Bernhard Scholkopf**.

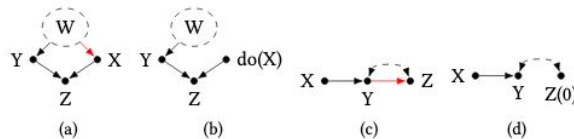
Data Generating Process

- Goal: Diagnose **T** from **F** and **X**

T: Pneumonia
D: Department
F: Style features
X: Lung X-ray



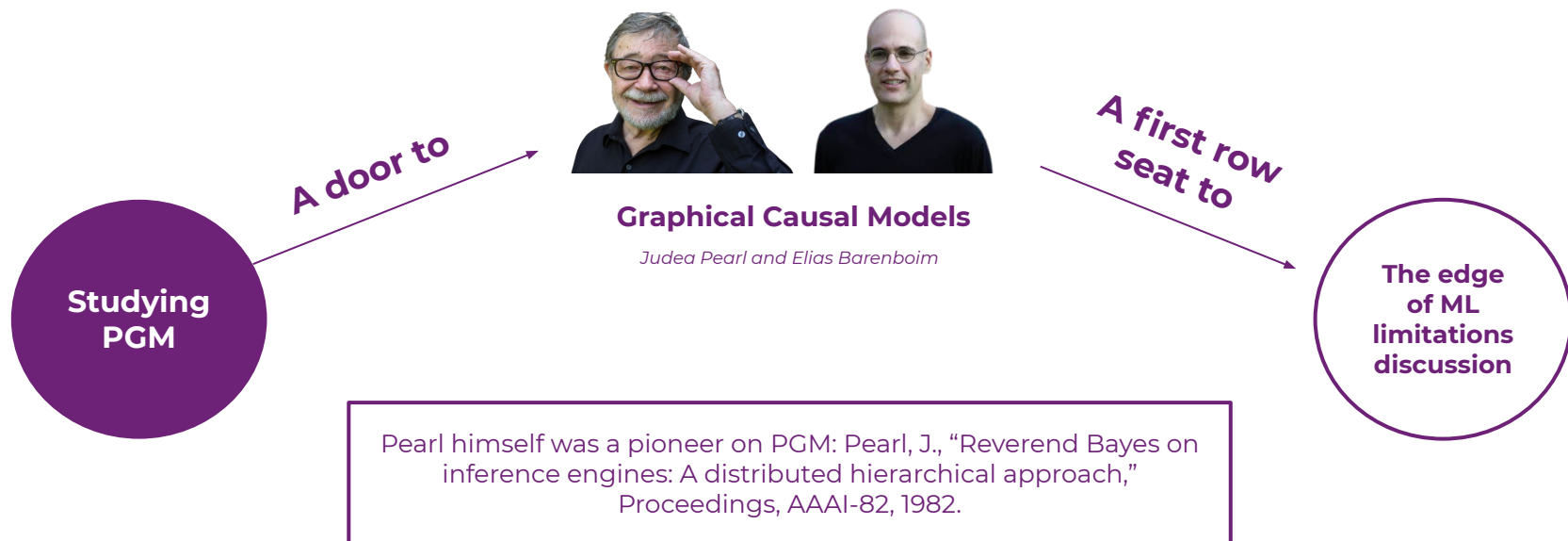
- Some of these mechanisms will be **stable** across environments, others are **unstable** and more likely to change
 - Ex: Effect of pneumonia and style on X-ray image does not change.
 - Ex: Protocols/preferences for style features differ from department to department or even technician to technician



The images link to the papers.

PGM is the base of

Graphical Causal Models



How to deal with causal questions?

Causal - Core Concepts, Notation

W : Treatment assignment

X_i : Features / Characteristics

Y : Observed outcome

Y^1 : Outcome that would be observed if treated

Y^0 : Outcome that would be observed if not treated

Causal - Core concepts

Potential outcome

The outcome we would see under each possible treatment option (Y^n).

Counterfactual

Slightly different than potential outcomes, but often used interchangeably.

What would have happened had the action been different?

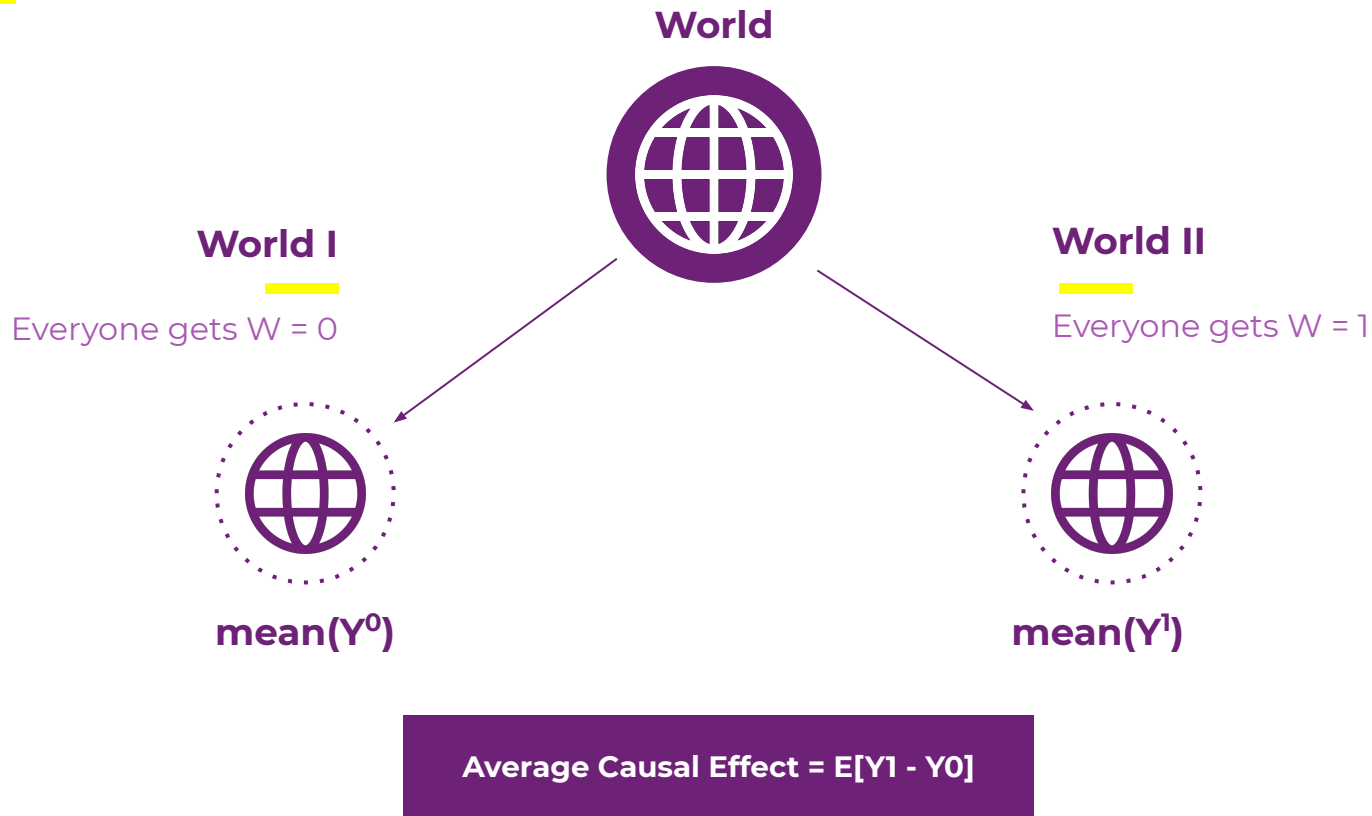
Before treatment decision is made, any outcome is a potential outcome: Y^1 or Y^0 .

After treatment there's an observed outcome Y^A and a counterfactual one Y^{1-A} .

Confounding

Anything that can impact both W and Y .

Causal - Causal Effect



Causal - Randomized Controlled Trial (RCT)

It's almost like having two new worlds!

- Golden standard;
- Solves all our problems!
- It has its own challenges, but once solved the results are robust;
- People in academia are used to do it.



The challenge

If random testing is a way to avoid all the difficulties of estimating causal effect, why do we even bother?

- It may not be **ethical**
- It can be **costly**

The challenge is estimating causal effect using either just **observational data** or using it with some random test data.

The clash of the worlds

Causality is at the center of most ML criticism

A ctrl+f for "causal" in a couple of popular papers about ML limitations

- [Underspecification Presents Challenges for Credibility in Modern Machine Learning](#) (17 matches)
- [Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift](#) (2 matches)
- [Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence](#) (4 matches)
- [Invariant Risk Minimization](#) (6 matches)

Generalization as "Out of Distribution generalization" or "Domain generalization"

Independent Causal Mechanism

Causal invariance

A is a city's altitude, **T** is the average year temperature and we have a sample for a couple of countries.

$$\begin{aligned} p(a, t) &= p(a \mid t)p(t) & \mathbf{T} &\longrightarrow \mathbf{A} \\ &= p(t \mid a)p(a) & \mathbf{A} &\longrightarrow \mathbf{T} \end{aligned}$$

How would we know the right direction?

Causal invariance

Now we add an identifier to each country:

$$p^{Brazil}(a, t) = p^{Brazil}(t \mid a)p^{Brazil}(a)$$

$$p^{Germany}(a, t) = p^{Germany}(t \mid a)p^{Germany}(a)$$

Hypothesis: physics is invariant on different contexts:

$$p^{Brazil}(t \mid a) = p^{Germany}(t \mid a) = p(t \mid a)$$

Causal invariance - method

Assuming an Additive Noise Model:

$$Y = f(X) + N_Y$$

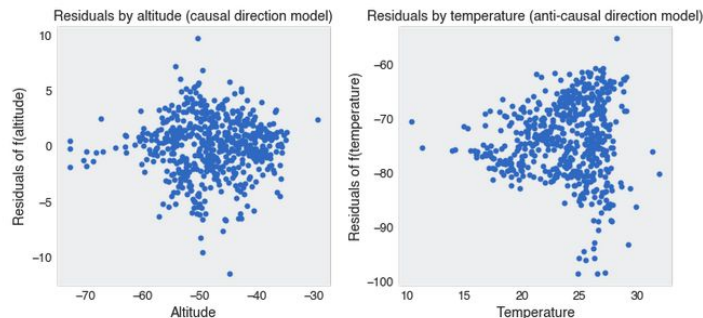
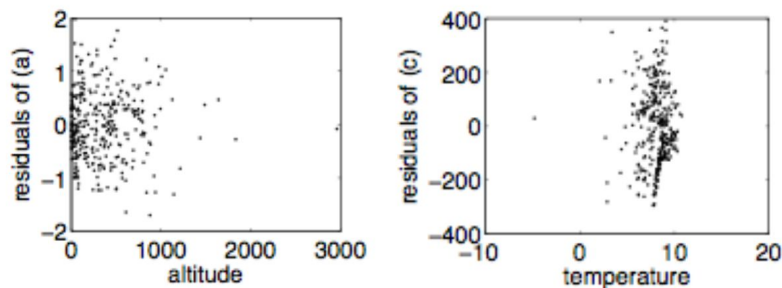
Where Y and the noise N are independent.

Then, there's no such model that $X = g(Y) + N_x$

And X is independent of N .

Causal invariance - method

- 1 - Fit a function f as a non-linear model of X on Y (assumption of noise additive model)
- 2 - Compute the residual $N = Y - f(X)$
- 3 - Check whether N and X are statistically independent



Indeed, there's a strong dependence in the anti causal direction when we look to real data.

Invariant Causal Prediction - ICP

Invariant Causal Prediction - ICP

We're going to search for a subset of features that present a stable relationship with the target.

Hypothesis 1. (*invariant prediction*): there's a coefficient vector $\gamma^* = (\gamma_1^*, \dots, \gamma_p^*)^t$

for all $e \in \varepsilon : X^e$ there's an arbitrary distribution e

$$Y^e = \mu + X^e \gamma^* + \epsilon^e, \epsilon^e \sim F_\epsilon \quad e \quad \epsilon^e \perp X_{S^*}^e,$$

where $\mu \in \mathbb{R}$ is an intercept, ϵ^e is a random noise with zero mean, finite variance and the same distribution F_ϵ for all $e \in \varepsilon$.

Invariant Causal Prediction

For every possible subset of features:

1) Train with all data

Train a model using the data from all the context and the subset of features. Calculate the residues.

2) Test the residues

Test for every context if the residues on them have the same mean and variance than each other context. Do a joint test to decide rejecting or not that subset.

3) Final model with stable variables

In the final subset, we use the intersection of all the not rejected subsets from the previous step.

Invariant Causal Prediction

$$\begin{aligned}x_1 &\sim \mathcal{N}(0, \sigma(e)) \\y &\sim x_1 + \mathcal{N}(0, 1) \\x_2 &\sim y + \mathcal{N}(0, 1)\end{aligned}$$

Result:

$$S = \{x_1\}$$

$$\begin{aligned}x_1 &\sim \mathcal{N}(0, \sigma(e)) \\x_3 &\sim \mathcal{N}(0, \sigma(e)) \\y &\sim x_1 + 2x_3 + \mathcal{N}(0, \sigma(e)) \\x_2 &\sim y + \mathcal{N}(0, 1)\end{aligned}$$

Result:

$$S = \{x_1, x_3\}$$

Invariant Risk Minimization - IRM

Invariant Risk Minimization - IRM

The objective function is modified to reflect the preference for a model which is optimal under different contexts.

ERM

$$L_{ERM}(\beta) = R(\beta)$$

IRM

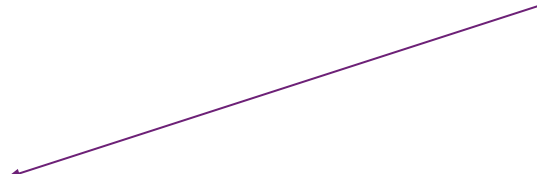
$$L_{IRM}(\Phi, \omega) = \sum_{e \in \epsilon_{tr}} R^e(\omega \circ \Phi) + \lambda \mathbb{D}(\omega, \Phi, e)$$


**"Nature does not
shuffle the data, so we
shouldn't either"**

Leon Bottou

Linear regression case

$$L_{IRM}(\Phi, \omega) = \sum_{e \in \epsilon_{tr}} R^e(\omega \circ \Phi) + \lambda \mathbb{D}(\omega, \Phi, e)$$


$$R^e(\omega \circ \Phi) = R^e(\hat{\beta}) = \frac{1}{n} (X^e \hat{\beta} - y^e)(X^e \hat{\beta} - y^e)^T$$


$$\mathbb{D}_{lin}(\omega, \Phi, e) = \|\mathbb{E}_{X^e}[(X^e \hat{\beta})^T (X^e \hat{\beta})] \omega - \mathbb{E}_{X^e, Y^e}[(X^e \hat{\beta})^T Y^e]\|^2$$

Linear regression case

Gradient descent: $\hat{\beta}_{t+1} = \hat{\beta}_t - \gamma \nabla_{\beta|\beta=\hat{\beta}_t} L(\beta)$

$$L_{IRM}(\Phi, \omega) = \sum_{e \in \varepsilon_{tr}} R^e(\omega \circ \Phi) + \lambda \mathbb{D}(\omega, \Phi, e)$$



$$-\frac{1}{N} X^T (X \hat{\beta} - y)$$



$$\frac{1}{N} (((X^e \hat{\beta})^T X^e \hat{\beta}) - ((X^e \hat{\beta})^T y)) \\ (X^e (X^e \hat{\beta})^T - ((X^e \hat{\beta})^T y^e))$$

Implementing paper example

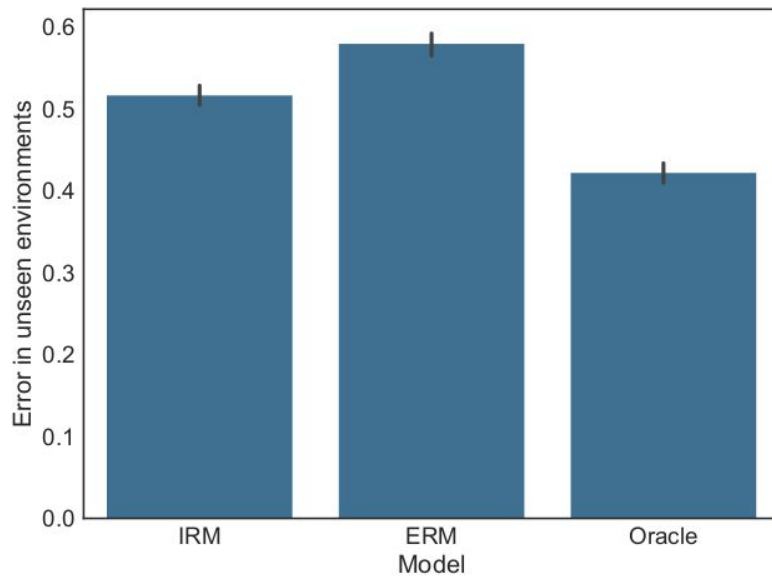
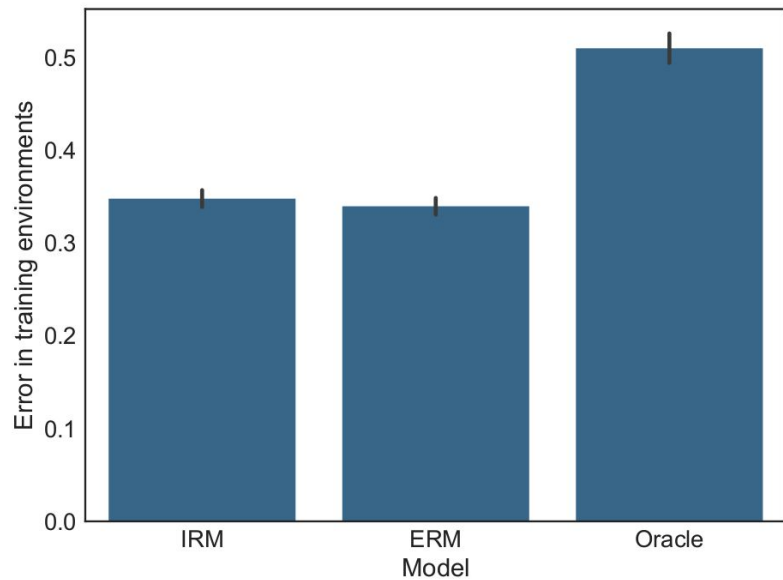
Applying the previous equations for the linear regression case and using the same set of equations used in the paper as the motivational example.

$$x_1 \sim \mathcal{N}(0, \sigma(e))$$

$$y \sim x_1 + \mathcal{N}(0, \sigma(e))$$

$$x_2 \sim y + \mathcal{N}(0, 1)$$

Results



Final considerations

Final considerations

Back to the **"Cost ineffective, not the best performance."**

It's for sure more costly, but it is the price of robustness. Dataset shift is not a sub field of ML, it is the whole field!

"...the i.i.d. assumption is the great lie of machine learning."

– Prof. Zoubin Ghahramani

Recommendations

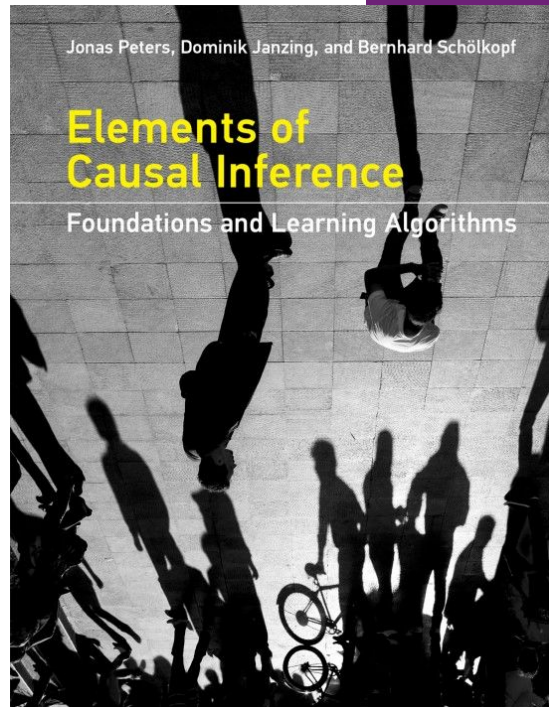
Recommendations on causal ML

Books

- Elements of Causal Inference (introduction to the research field);
- What If (practical guide for causal inference)
- Causality (focused in the graphical representation)
- Causal Inference for The Brave and True (practical and with python code)

Papers

- Invariant risk minimization
- Invariant Causal Prediction
- Causality for Machine Learning
- Toward Causal Representation Learning
- Inductive biases for deep learning of higher-level cognition



Questions?

Contact

Twitter: [@lgmoneda](#)

E-mail: lg.moneda@gmail.com