Explainable Data Science

Jordi Vitrià

UNIVERSITAT DE BARCELONA

jordi.vitria@ub.edu

Departament de Matemàtiques i Informàtica

I am a Full Professor at the Mathematics & Computer Science Department, **Universitat de Barcelona**. I am the Director of the **Master in Fundamental Principles of Data Science** at UB. I am the leader of the **DataScience@UB** group, whose objective is to promote research & technology transfer in the areas of data analytics, machine learning and AI.

---

http://datascience.barcelona/
http://algorismes.github.io

Chapter I

Explanations

How to answer a WHY question…

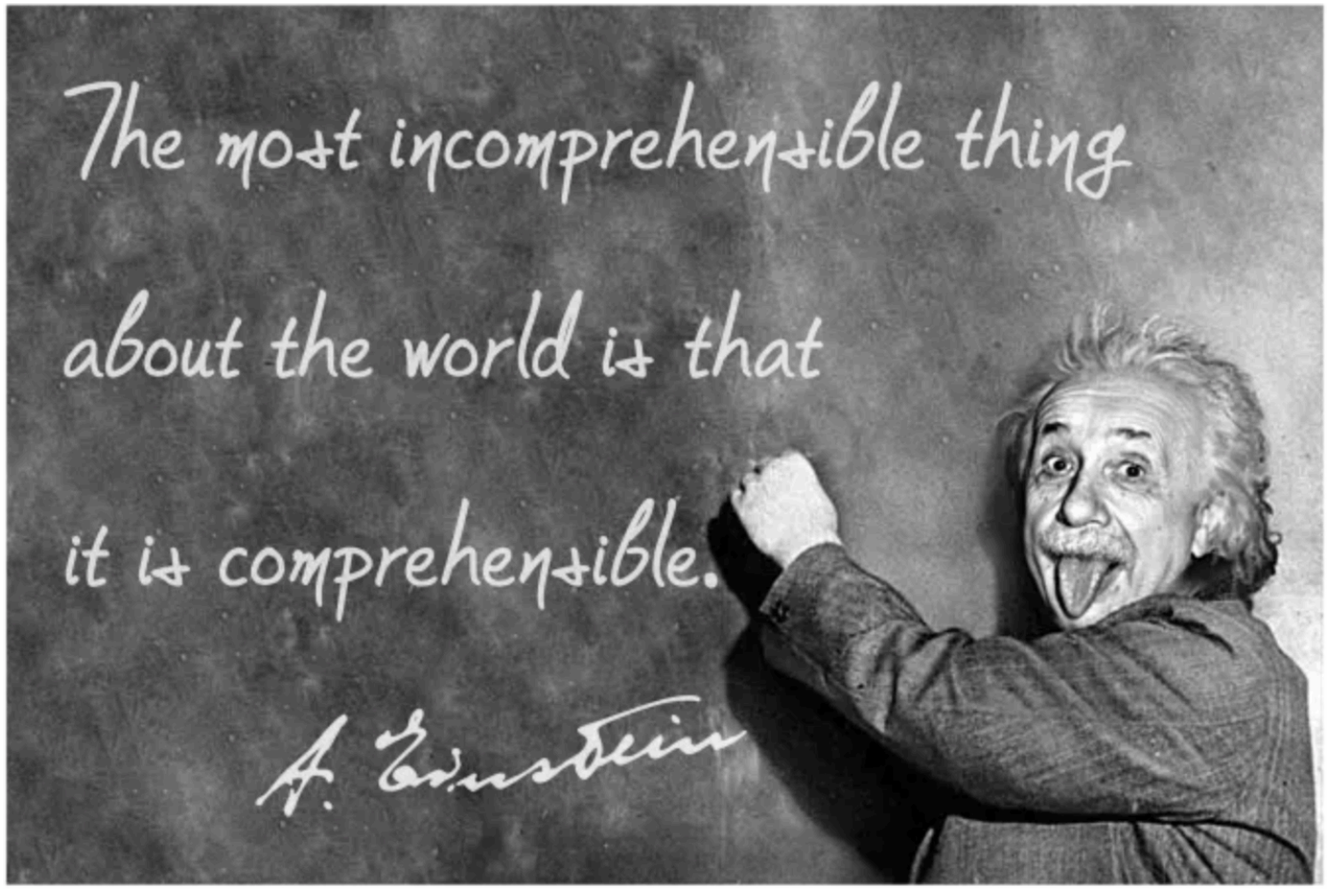
Richard Feynman

https://www.youtube.com/watch?v=Q1IL-hXO27Q

What is an **explanation**?
An answer to a why question…

The problem of Infinite Regress
(ancient Greek dialogue):

DMITRI: If Atlas holds up the world, who holds up Atlas?
TASSO: Atlas stands up in the back of a turtle.
DMITRI: What does the turtle stands up?
TASSO: Another turtle.
DMITRI: And what does *that* turtle stands up?
TASSO: My dear Dimitri, it's turtles all the way down!



An explanation is the **answer to a why-question** (Miller 2017).

- Why did not the treatment work on the patient?
- Why was my loan rejected?

Miller, Tim. 2017. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” arXiv Preprint arXiv:1706.07269.

A good explanation is:

- **Contrastive**. Humans usually do not ask why a certain prediction was made, but **why this prediction was made *instead of another prediction***. The solution for the automated creation of contrastive explanations might also involve finding prototypes or archetypes in the data.

- **Selected**. People do not expect explanations that cover the actual and complete list of causes of an event. We are used to selecting **one or two causes** from a variety of possible **causes** as **THE** explanation.

- **Social**. The social context determines the content and nature of the explanations. Getting the social part of the machine learning model right depends entirely on your specific application.

Miller, Tim. 2017. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” arXiv Preprint arXiv:1706.07269.

A good explanation is:

- **Focused on the abnormal**. People focus more on causes that had a small probability but nevertheless happened.

- **Thruthful**. The explanation should predict the event as truthfully as possible, which in machine learning is sometimes called **fidelity**.

- **Consistent** with prior beliefs of the explainee. This is difficult to integrate into machine learning!

- **General and probable**. A cause that can explain many events is very general and could be considered a good explanation. Generality can easily be measured by the feature's support, which is the number of instances to which the explanation applies divided by the total number of instances.

Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." arXiv Preprint arXiv:1706.07269.

Chapter II

Algorithmic Decisions

Derrière l'algorithme de Parcoursup, un choix idéologique

La répartition des étudiants entre les universités et les filières est un problème complexe puisqu'elle s'effectue sur base d'un conflit massif entre l'offre et la demande : on dénombre plus de 880.000 candidats pour un total (à raison de 10 vœux possibles par candidat) de quelques 7.000.000 de vœux de formation *[810.000 ont in fine validé leurs voeux, NDLR]*. La résolution d'un tel conflit n'est plus sérieusement envisageable humainement. Dès lors qu'un algorithme travaille à cette mise en relation n'est pas à remettre en question. La vraie question est celle de l'objectif assigné à l'algorithme et des choix qu'il doit exécuter.

Cette décision politique et idéologique se lit dans la formule algorithmique même de Parcoursup. Cet algorithme, dont l'objectif est de mettre en relation deux objets, d'un côté des établissements, de l'autre des étudiants, est en effet inspiré par le célèbre algorithme de Gale et **Shapley**, repris par Alvin Roth, prix Nobel d'économie en 2012. Il relève au fond d'un vieux problème économique que l'on appelle l'appariement stable.

https://www.nouvelobs.com/education/20180713.OBS9643/derriere-l-algorithme-de-parcoursup-un-choix-ideologique.html

The **stable marriage problem** has been stated as follows:

Given *n* men and *n* women, where each person has ranked all members of the opposite sex in order of preference, marry the men and women together such that there are no two people of opposite sex who would both rather have each other than their current partners. When there are no such pairs of people, the set of marriages is deemed **stable**.

There was once a time when humans made important decisions about other humans.

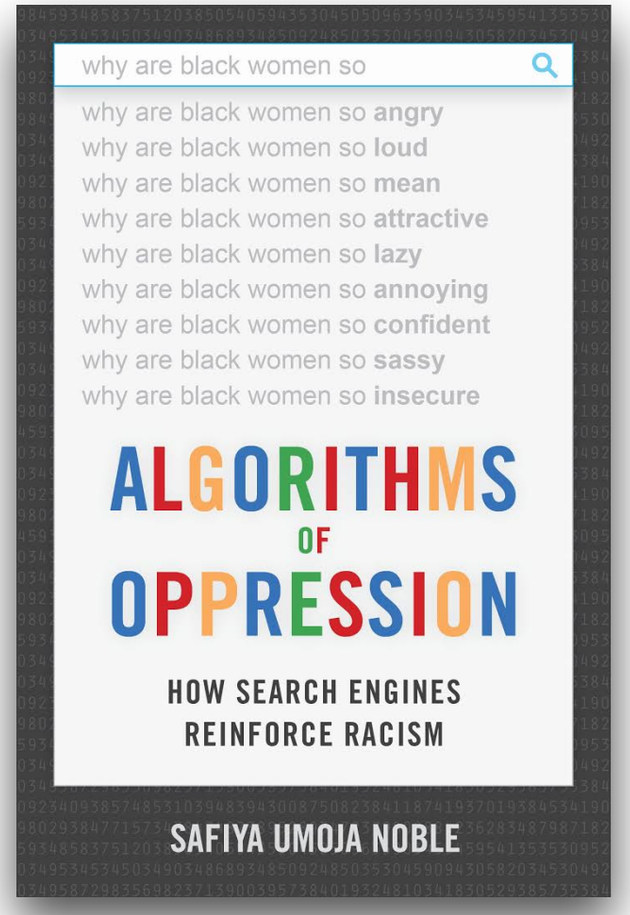
You went to your bank manager, in person, to ask for a loan.
A human hiring committee decided which candidate would get a job.
And judges even determined what a guilty offender's sentence would be!

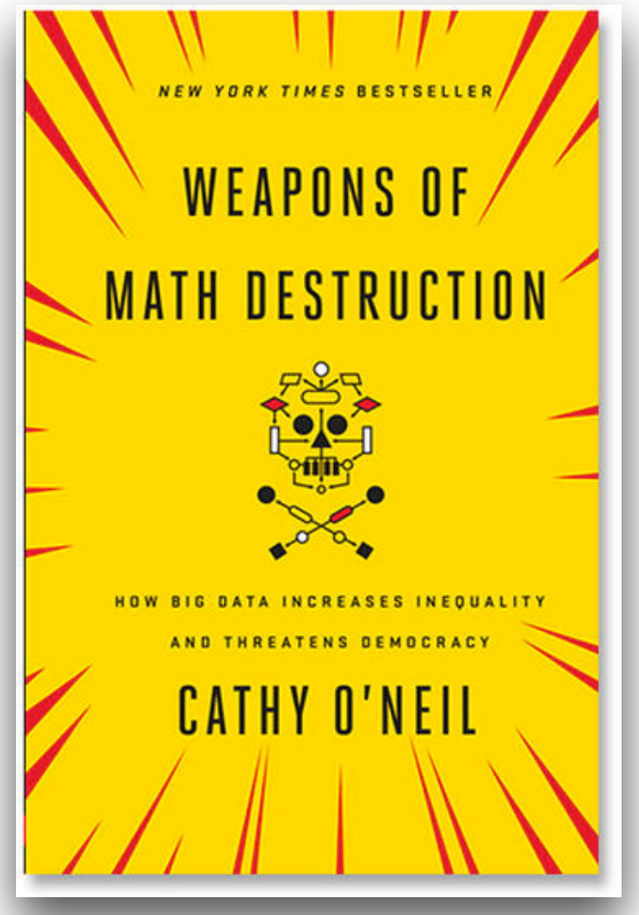
Now, many of those decisions are made by AIs.

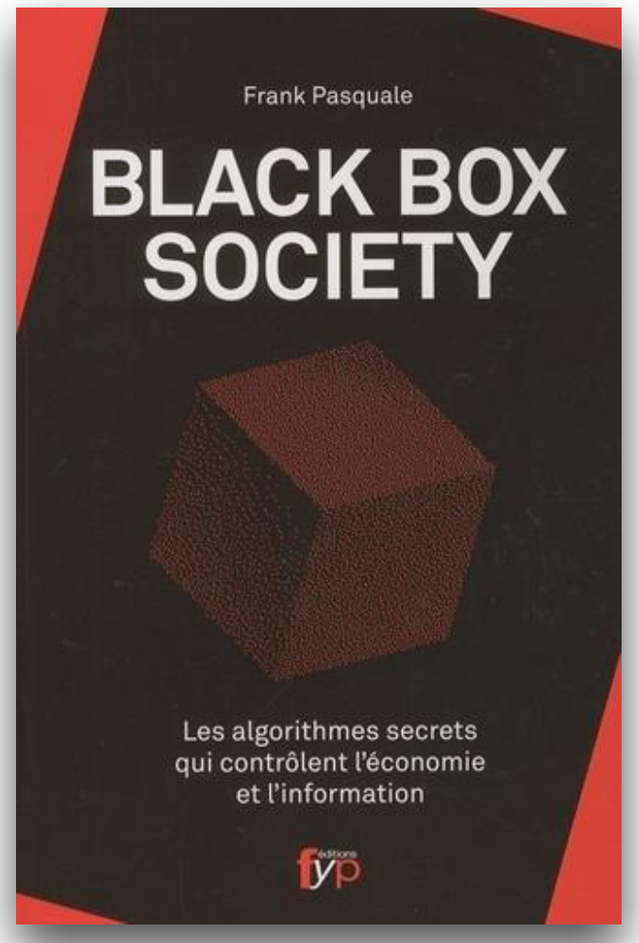
There are a couple of problems with AIs making these decisions. First of all, the algorithms the machines use are often proprietary, meaning they aren't available to the general public or other computer scientists to examine. Second, and perhaps more troubling, AIs often interpret data in ways more complex than even their programmers can understand. In those cases, nobody could explain how the "black box" in an AI works, even if they wanted to.

CBC Radio · January 19, 2019
https://www.cbc.ca/radio/spark/422-1.4982026/asking-why-instead-of-how-could-better-explain-ai-decisions-1.4982038

It is difficult to assess the outcome of ML black boxes







There are unknown consequences at a societal level.

The explanation depends on several factors.
There is **no best explanation**.

Model	Data	Human	Task
Simple Complex Black Box	Numeric Binary Categorical Text Images	Owner Programmer Analyst Supervisor Operator Executor Decision-subject Data-subject	Local/Global Generic/Accurate Low/High Stakes

We can look for answers to very different **questions**:

- How does the trained model make predictions?
- How do parts of the model affect predictions?
- Is it working as intended?
- Is it doing sensible predictions?
- Are decisions conforming to regulation?
- Why did the model make a certain prediction for an instance?
- Why did the model make specific predictions for a group of instances?
- Am I treated fairly?
- What could I do differently to get a favourable outcome next time?

It is a difficult problem!

The New York Times

YOUR MONEY ADVISER

In California, Gender Can No Longer Be Considered in Setting Car Insurance Rates

By Ann Carrns

Jan. 18, 2019

Minh Uong/The New York Times

California joined about a half-dozen states this month in banning the use of a person's gender when assessing risk factors for car insurance, a change that could potentially alter rates for scores of drivers across the state.

Delip Rao @deliprao · 18 h

Not only does the gender variable has to be decorrelated with each of the other other individual factors, but also has to be decorrelated to any arbitrary function of any arbitrary subset of the other facts. For anyone wondering, this perfect decorrelation is next to impossible.

Tradueix el tuit

1 4

Delip Rao @deliprao · 18 h

Why caution must be exercised in understanding/accepting such things, or we are simply living in an illusion of progress as opposed to actual progress.

Tradueix el tuit

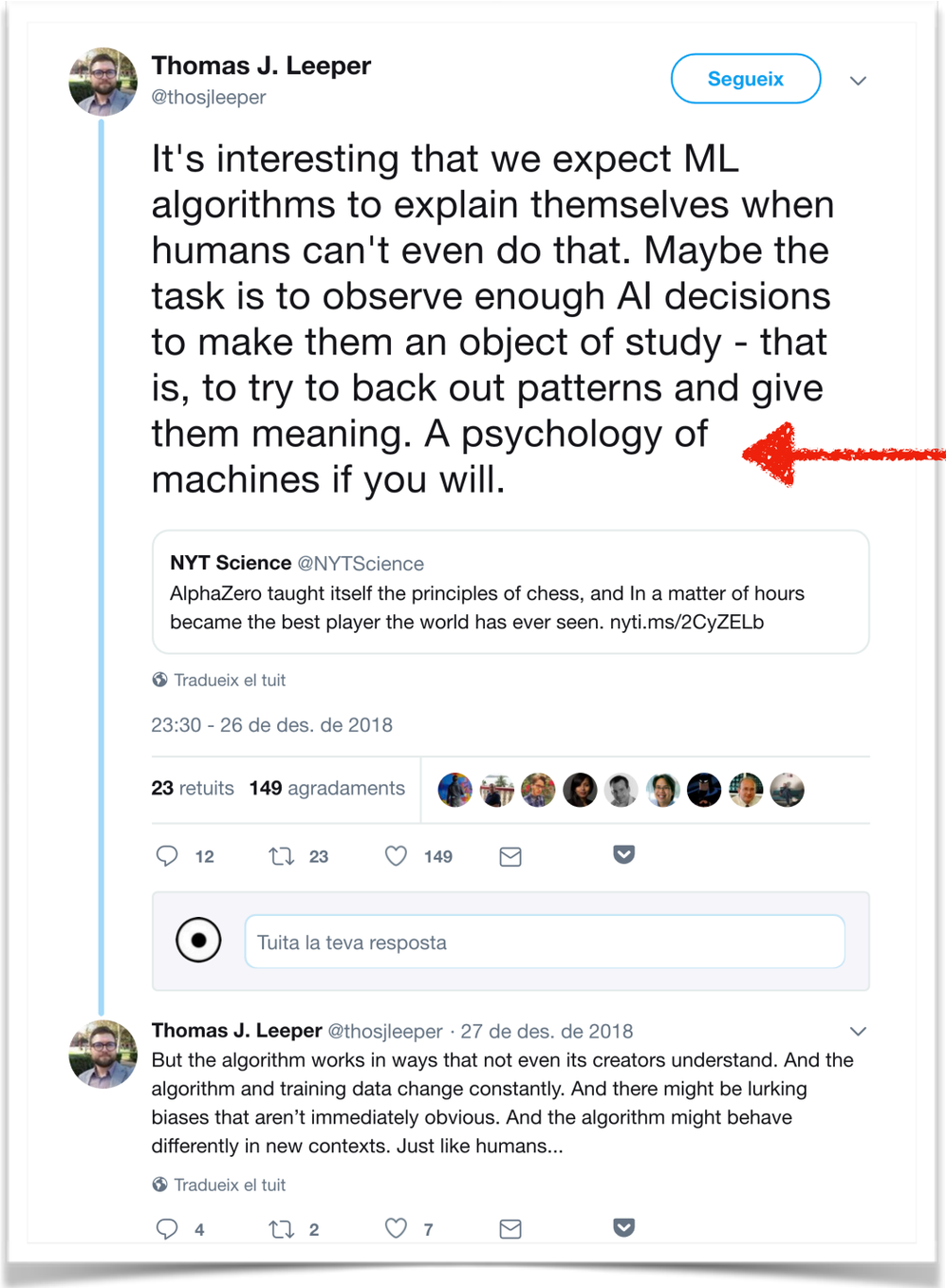
1 1 7

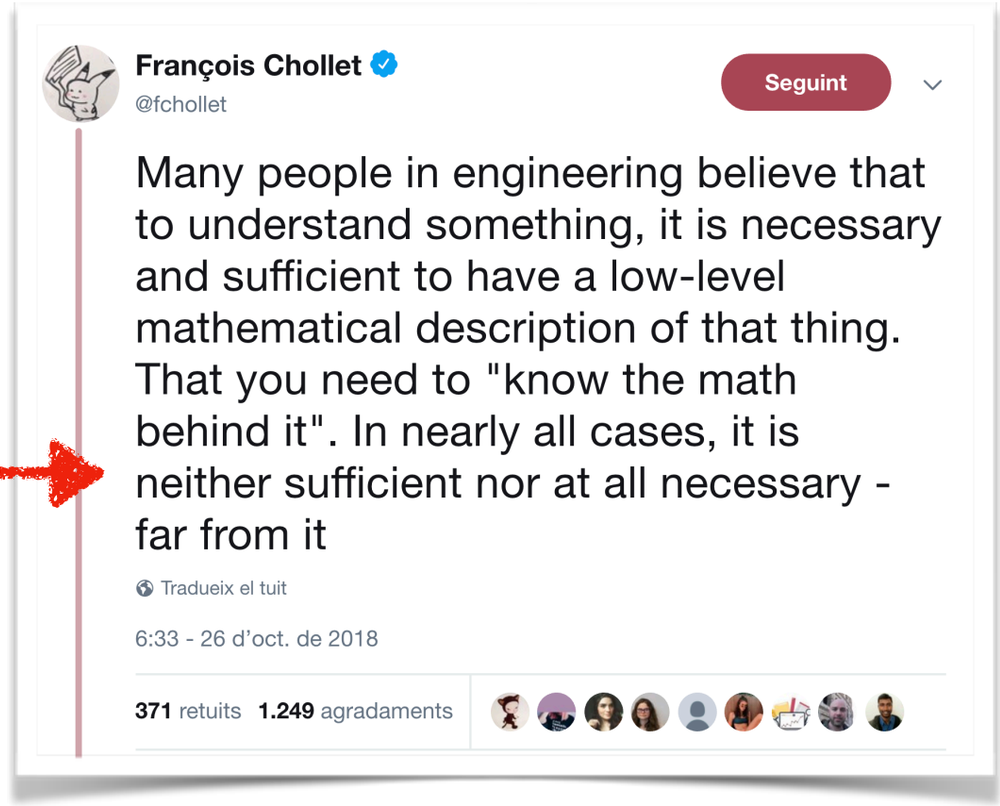
Delip Rao @deliprao · 17 h

So, if this perfect decorrelation is impossible, what is one to do? 1) Always be vigilant, and 2) Always measure disparate outcomes wrt to the sensitive variables. Remember that justice should not be blind, and that statistical fairness should be the aim than absolute fairness.

Tradueix el tuit

3 6


Thomas J. Leeper
@thosjleeper
Segueix
It's interesting that we expect ML algorithms to explain themselves when humans can't even do that. Maybe the task is to observe enough AI decisions to make them an object of study - that is, to try to back out patterns and give them meaning. A psychology of machines if you will.
NYT Science @NYTScience
AlphaZero taught itself the principles of chess, and In a matter of hours became the best player the world has ever seen. nyti.ms/2CyZELb
Tradueix el tuit
23:30 - 26 de des. de 2018
23 retuits 149 agradaments
12 23 149
Tuita la teva resposta
Thomas J. Leeper @thosjleeper · 27 de des. de 2018
But the algorithm works in ways that not even its creators understand. And the algorithm and training data change constantly. And there might be lurking biases that aren't immediately obvious. And the algorithm might behave differently in new contexts. Just like humans...
Tradueix el tuit
4 2 7


François Chollet
@fchollet
Seguint
Many people in engineering believe that to understand something, it is necessary and sufficient to have a low-level mathematical description of that thing. That you need to "know the math behind it". In nearly all cases, it is neither sufficient nor at all necessary - far from it
Tradueix el tuit
6:33 - 26 d'oct. de 2018
371 retuits 1.249 agradaments

If you can ensure that the **machine learning model can explain decisions**, you can check the following traits more easily:

- **Fairness**: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups. An interpretable model can tell you why it has decided that a certain person should not get a loan, and it becomes easier for a human to judge whether the decision is based on a learned demographic (e.g. racial) bias.
- **Privacy**: Ensuring that sensitive information in the data is protected.
- **Reliability** or Robustness: Ensuring that small changes in the input do not lead to large changes in the prediction.
- **Causality**: Check that only causal relationships are picked up.
- **Trust**: It is easier for humans to trust a system that explains its decisions compared to a black box.

We don’t need interpretability if the model has no significant impact, the problem is well studied (f.e. OCR) or if this enable people to manipulate or to game a critical system.

Chapter III

Causality

Jordi Vitrià
Universitat de Barcelona

Data Science Scenario

Data science tasks (inferring answers to different kinds of questions):

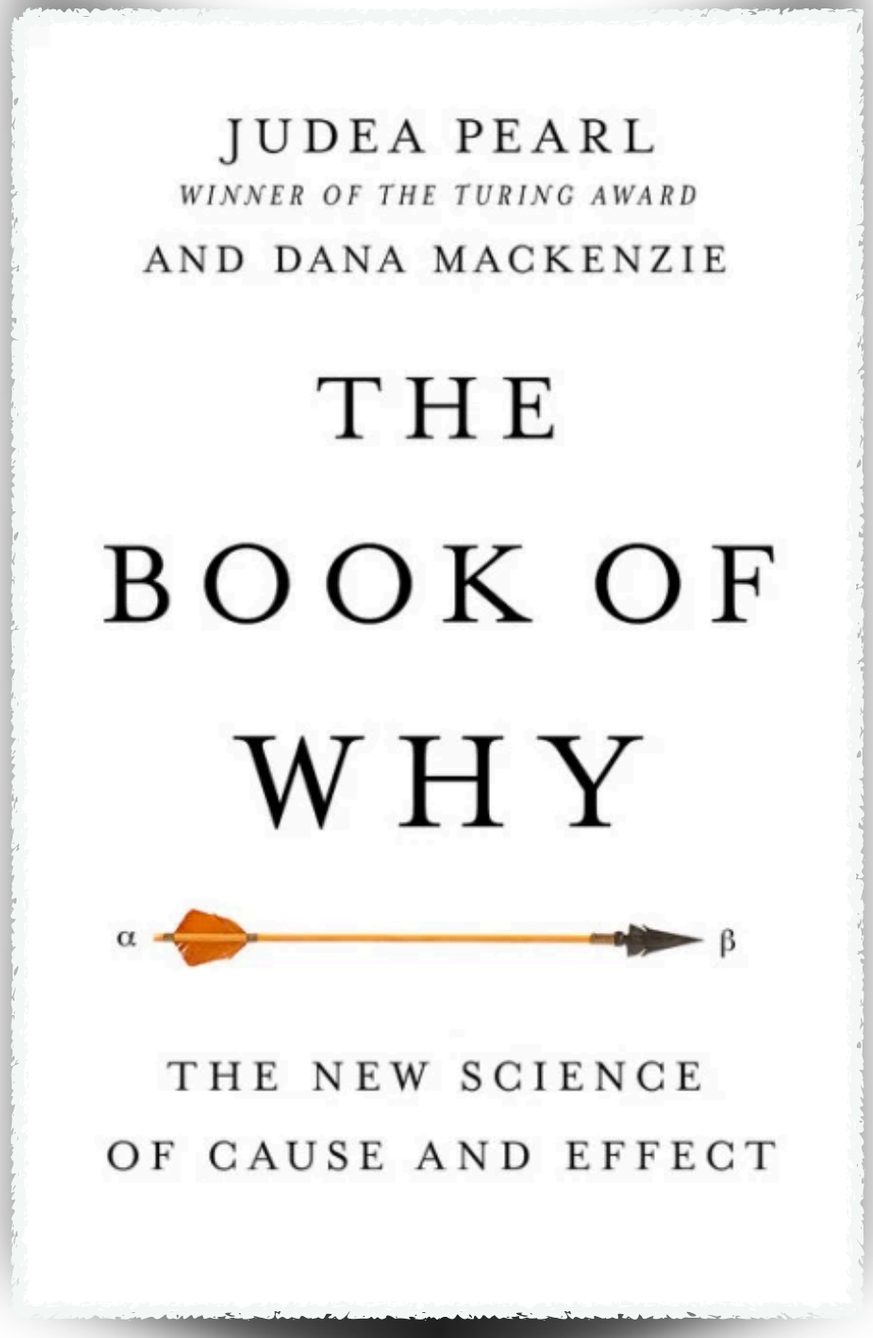
- **Description** is using data to provide a quantitative summary of certain features of the world. → What is the mean value of X?

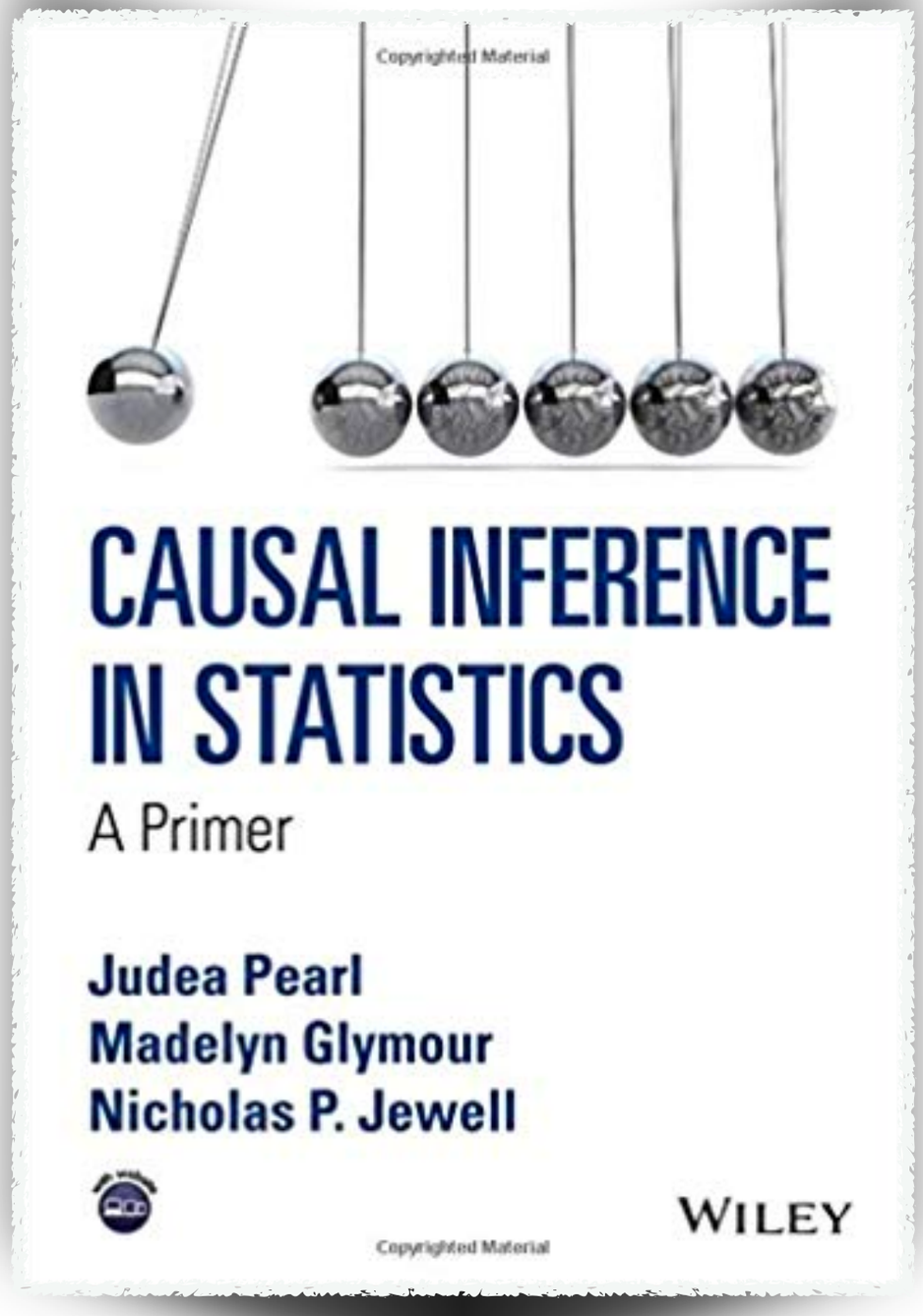
- **Prediction** (or **association**) is using data to map some features of the world (the inputs) to other features of the world (the outputs). → How would seeing X change my belief in Y?

- **Causation**: Measuring the causal influence of a variable X in another variable Y, while excluding any influences on Y not actually due to the causal effect of X, and being able to guess what the effect will be if one performs an action. → How would expected lifespan change if more people become vegetarian?

Intervention

- **Counterfactuals**: Being able to reason about hypothetical situations, things that *could* happen. → Would my grandfather still be alive if he did not smoke?


JUDEA PEARL
WINNER OF THE TURING AWARD
AND DANA MACKENZIE
THE
BOOK OF
WHY
α
β
THE NEW SCIENCE
OF CAUSE AND EFFECT


Copyrighted Material
CAUSAL INFERENCE
IN STATISTICS
A Primer
Judea Pearl
Madelyn Glymour
Nicholas P. Jewell
WILEY
Copyrighted Material

Counterfactuals: David Blei's election example

Given that Hilary Clinton did not win the 2016 presidential election, and given that she did not visit Michigan 3 days before the election, and given everything else we know about the circumstances of the election, what can we say about the probability of Hilary Clinton winning the election, had she visited Michigan 3 days before the election?

Let's try to unpack this. We are are interested in the **probability** that:

- she *hypothetically* wins the election

conditioned on four sets of things:

- she lost the election
- she did not visit Michigan
- any other relevant an observable facts
- she *hypothetically* visits Michigan

Why would quantifying this probability be useful? Mainly for credit assignment.

Causal inference tasks requiere expert knowledge. Data is not sufficient.

Answering a causal question typically requires a combination of data, analytics, and expert causal knowledge.

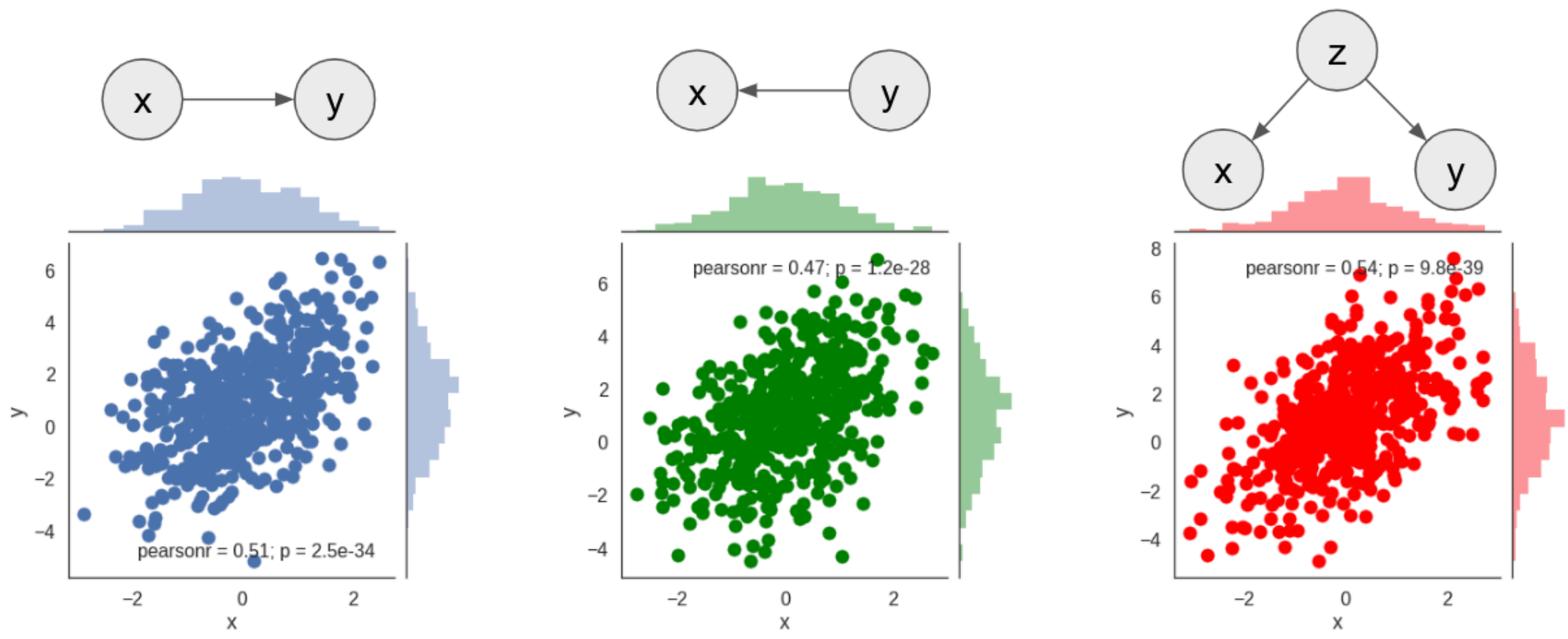
Causal inference tasks require expert knowledge not only to specify the question (the causal effect of what treatment on what outcome) and to identify/generate relevant data sources, but also to describe the **causal structure of the system** under study.

Causal structure of a system

```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```



The joint distributions p(x,y) of these three causal models are indistinguishable.

Source: https://www.inference.vc/

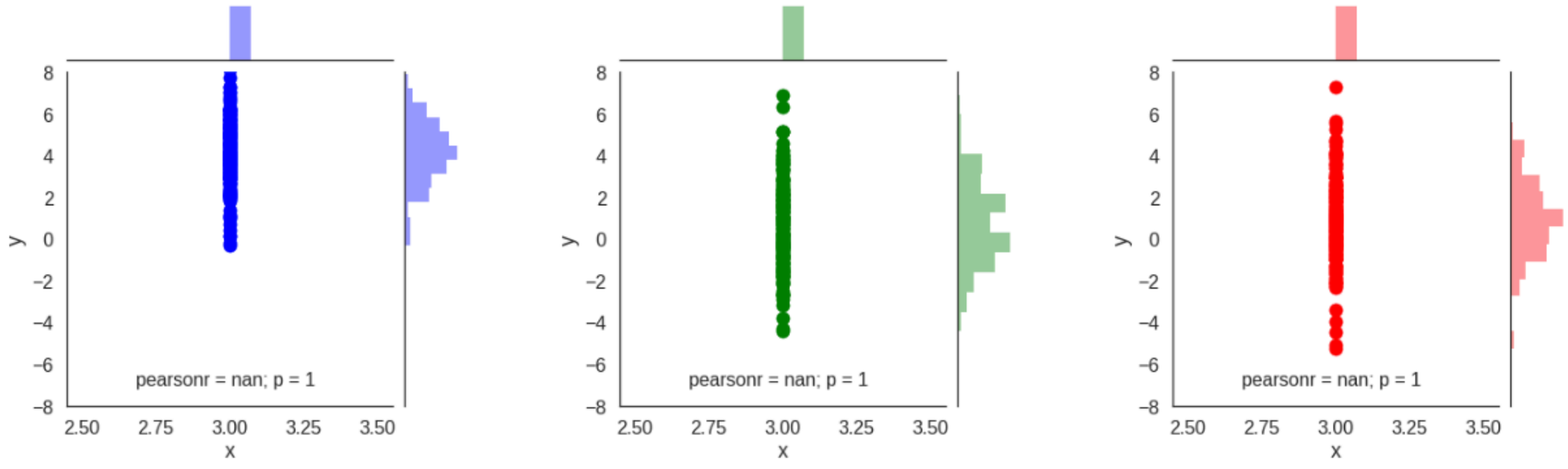
Causal structure of a system

Intervention x=3

x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3

y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3

z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3



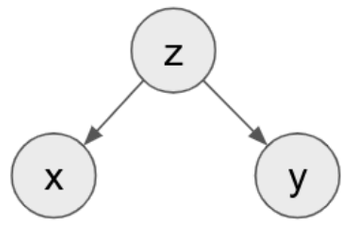
But their marginals p(y/x) are different if there is an intervention!

Source: https://www.inference.vc/

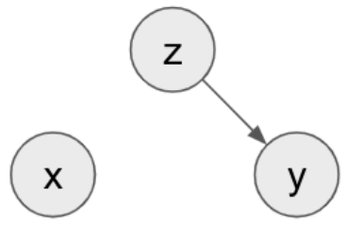
Causal structure of a system

$P(y|do(X)) = p(y|x)$

$P(y|do(X)) = p(y)$



$P(y|do(X)) = p(y)$



Do calculus:

Now the question is, **how can we say anything about the intervened conditional when we only have data from the observed distribution p(x,y)**. We have the causal model relating the two. To cut a long story short, this is what the so-called *do-calculus* is for.

Do-calculus allows us to massage the intervened conditional distribution until we can express it in terms of various marginals, conditionals and expectations under the observed distribution.

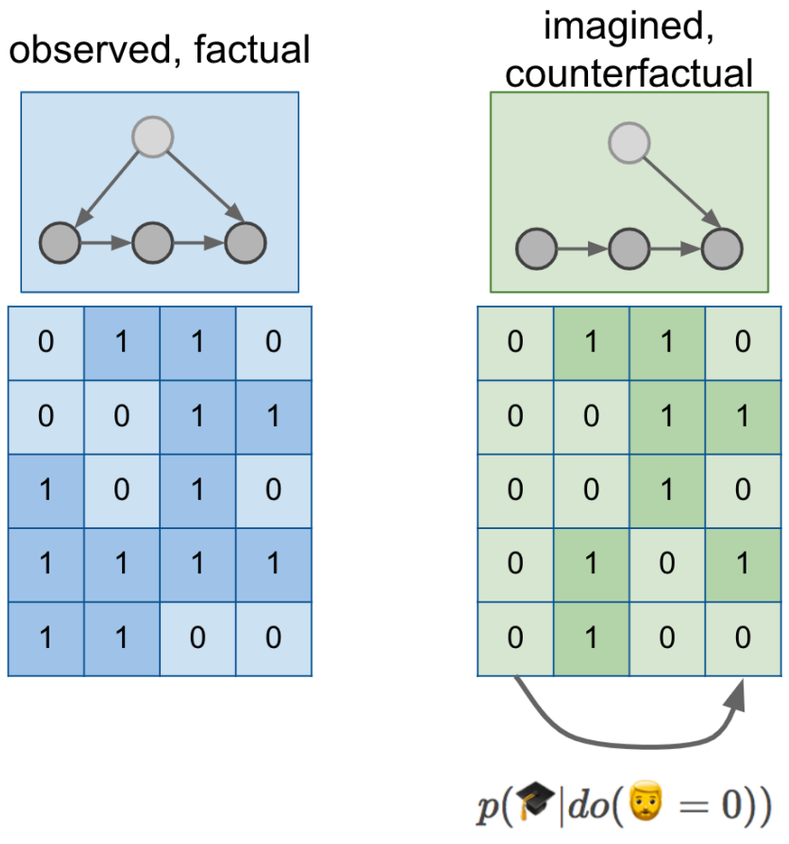
Source: https://www.inference.vc/

Counterfactual Queries

Queries

Let's now suppose that I want an answer to this question: *Given that I have a beard, and that I have a PhD degree, and everything else we know about me, with what probability would I have obtained a PhD degree, had I never grown a beard?*

Let's consider the scenario for **interventional queries**:



This is not an interventional query. This is a counterfactual query!

Source: https://www.inference.vc/

Counterfactual Queries

p(🎓|*do*(👨=0)) talks about a **randomly sampled individual**, while a **counterfactual talks about a specific individual**!

To get an answer to our question we have to step beyond causal graphs and introduce another concept: **structural equation models (SEM)**.

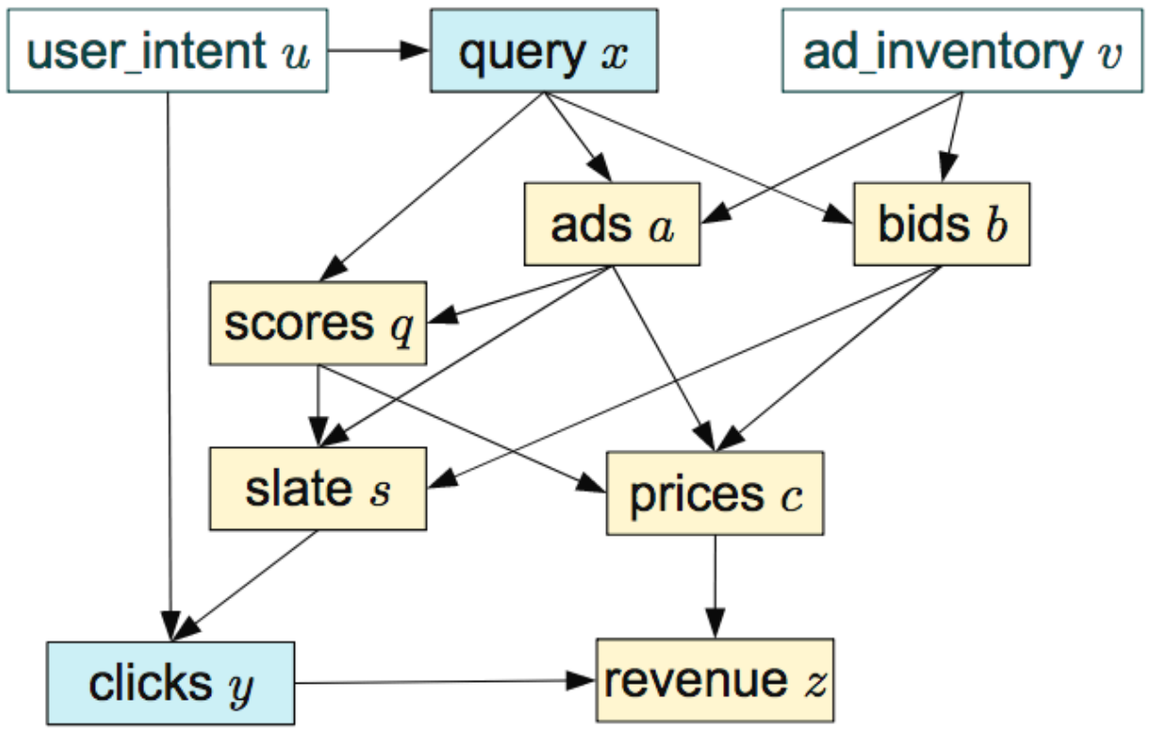
Journal of Machine Learning Research 14 (2013) 3207-3260 Submitted 9/12; Revised 3/13; Published 11/13

Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising

Léon Bottou LEON@BOTTOU.ORG
Microsoft
1 Microsoft Way
Redmond, WA 98052, USA

Jonas Peters* PETERS@STAT.MATH.ETHZ.CH
Max Planck Institute
Spemannstraße 38
72076 Tübingen, Germany

Joaquin Quiñonero-Candela† JQUINONERO@GMAIL.COM
Denis X. Charles CDX@MICROSOFT.COM
D. Max Chickering DMAX@MICROSOFT.COM
Elon Portugaly ELONP@MICROSOFT.COM
Dipankar Ray DIPANRAY@MICROSOFT.COM
Patrice Simard PATRICE@MICROSOFT.COM
Ed Snelson EDSNELSO@MICROSOFT.COM
Microsoft
1 Microsoft Way
Redmond, WA 98052, USA



Counterfactual Queries

The dependencies shown by the diagram are equivalently encoded by the following set of equations:

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

For each **node** in the graph above we now have a corresponding **function** f_i. The arguments of each function are the causal **parents** of the variable it instantiates, e.g. f_1 computes x from its causal parent u, and f_2 computes a from its causal parents x and v.

In order to allow for nondeterministic relationship between the variables, we additionally allow each function f_i to take another input, ϵ_i which you can think of as a **random number.**

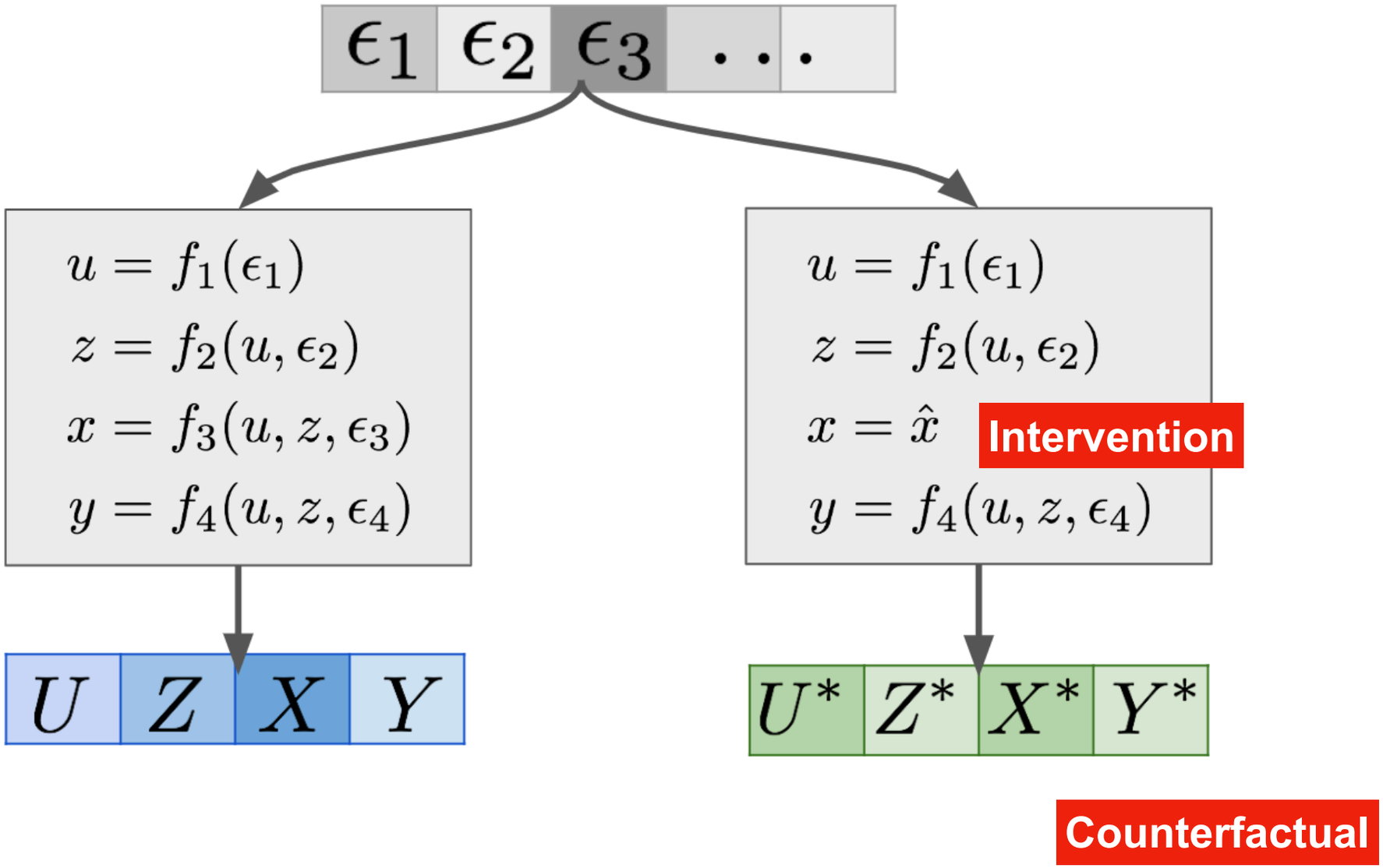
Counterfactual Queries

The structural equation model (SEM) entails the joint distribution, in that you can "**sample**" from an SEM by evaluating the functions in order, plugging in the random ϵ where needed.

In a SEM an intervention on a variable, say q, can be modeled by deleting the corresponding function, f_4, and replacing it with another function.

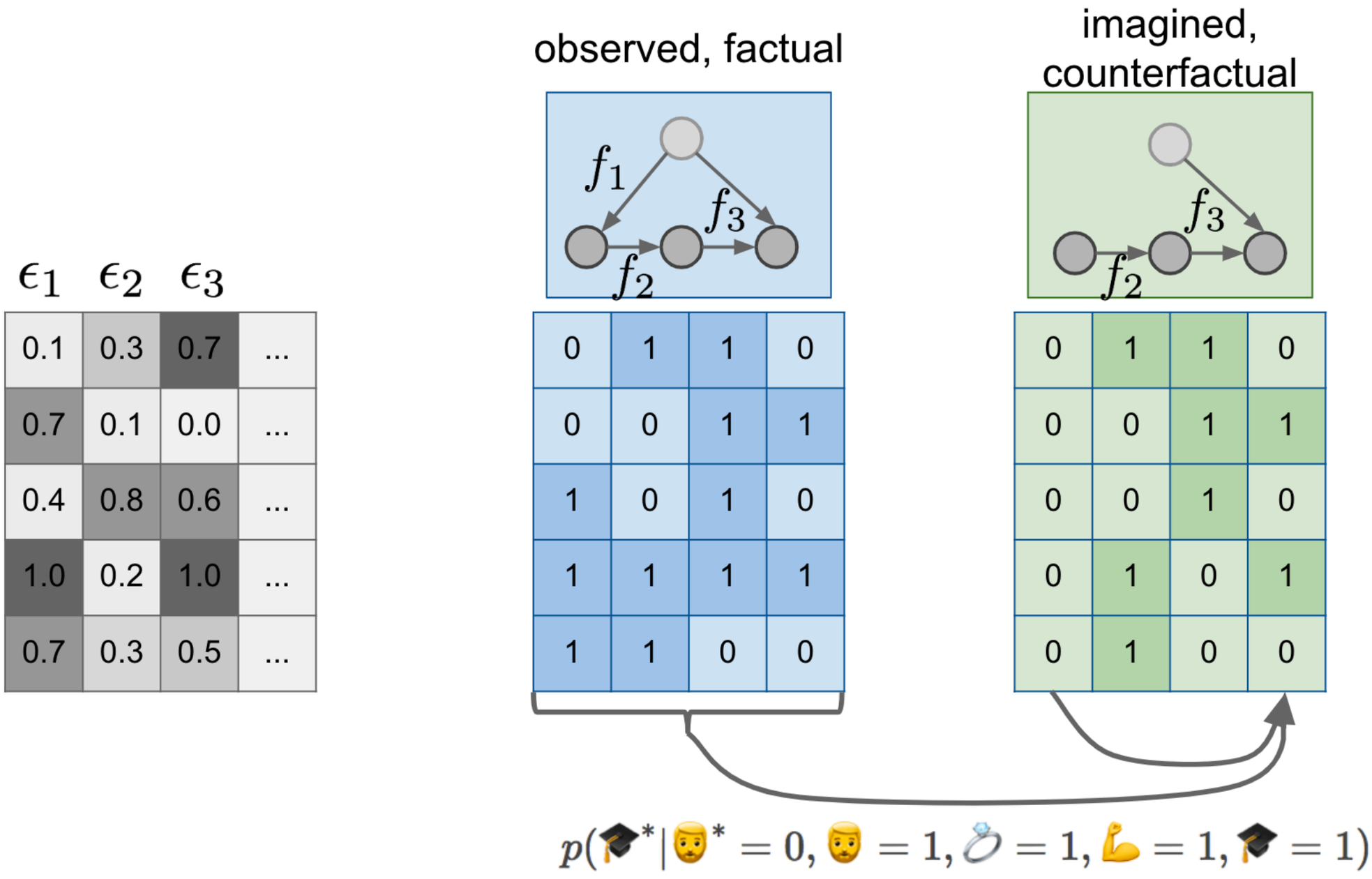
For example do$(Q = q_0)$ would correspond to a simple assignment to a constant $f_4(x, a) = q_0$.

Counterfactual Queries



Source: https://www.inference.vc/

Counterfactual Queries

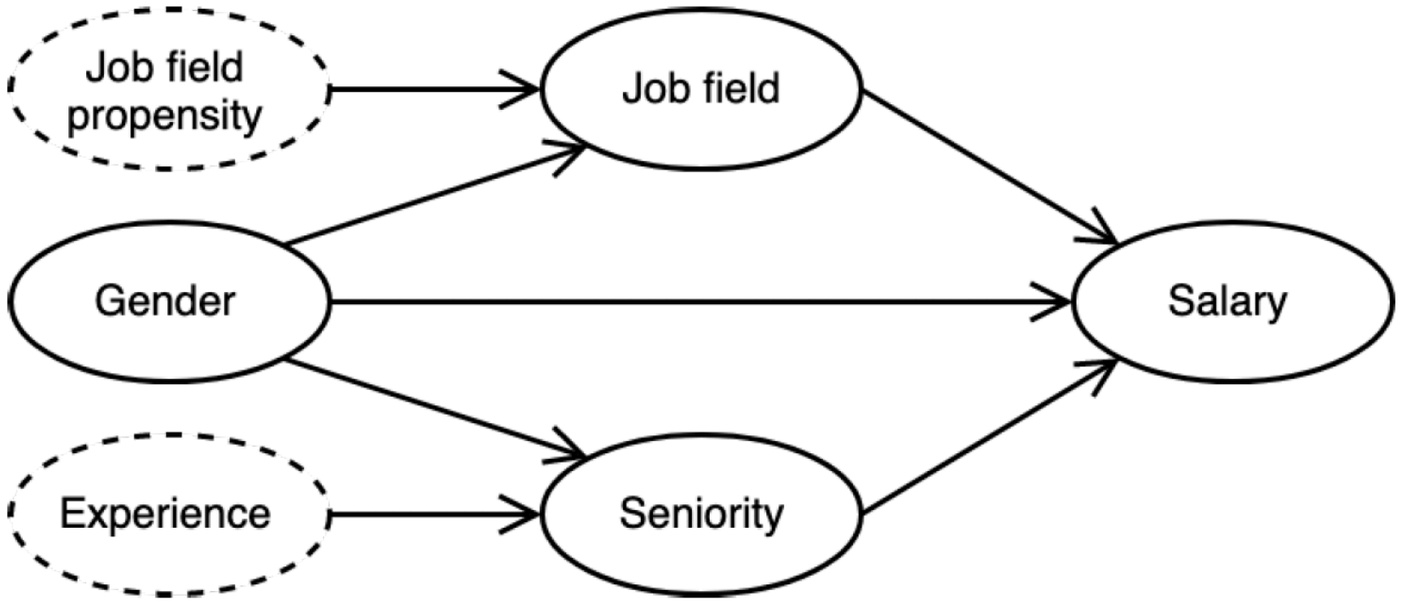


Source: https://www.inference.vc/

Chapter IV

Causal Explanations

Jordi Vitrià
Universitat de Barcelona



Model for automatic salary assignment

Why do I have this salary?

- Is gender a cause? Does it make a difference? Is my salary **fair**?
- How long do I have to work for increasing 10% my salary?
- Etc.

The problem is to find **counterfactuals** that are both close to the actual datapoint and likely to occur in the real world.

How do we build a model to support this kind of queries?

Differentaible Model

Given a Causal Graph G, we represent each node in the graph by a neural network that takes the values of its parents (its causes) and computes an outcome.

Each node (X_i) is a Multilayer Perceptron (MLP) that takes as input the concatenation of its causal parent values in a single vector and, additionally, a noise signal (ϵ_i) that adds stochasticity to the node. These networks can be learned from observed data. Then, we can sample the model.

How do we build a model to support this kind of queries?

Differentaible Model

Producing counterfactual samples is possible.

Given a sample,
(1) we compute a noise signal that generates that sample, node by node, and
(2) apply it along with the desired intervention to generate a counterfactual sample.

For discrete distributions, we try noise values until a sample with the same value as the original one emerges.
For continuous distributions, we can use differentiable sampling methods that are invertible; thus, we can obtain the original noise signal with a single operation.

Explanations…

If you were a man instead of a woman, your salary would be the same (or not…).

If you work 2 more years you will get an increase of 10% in your salary…

Etc.

Chapter V

More Applications

Conterfactual Fairness

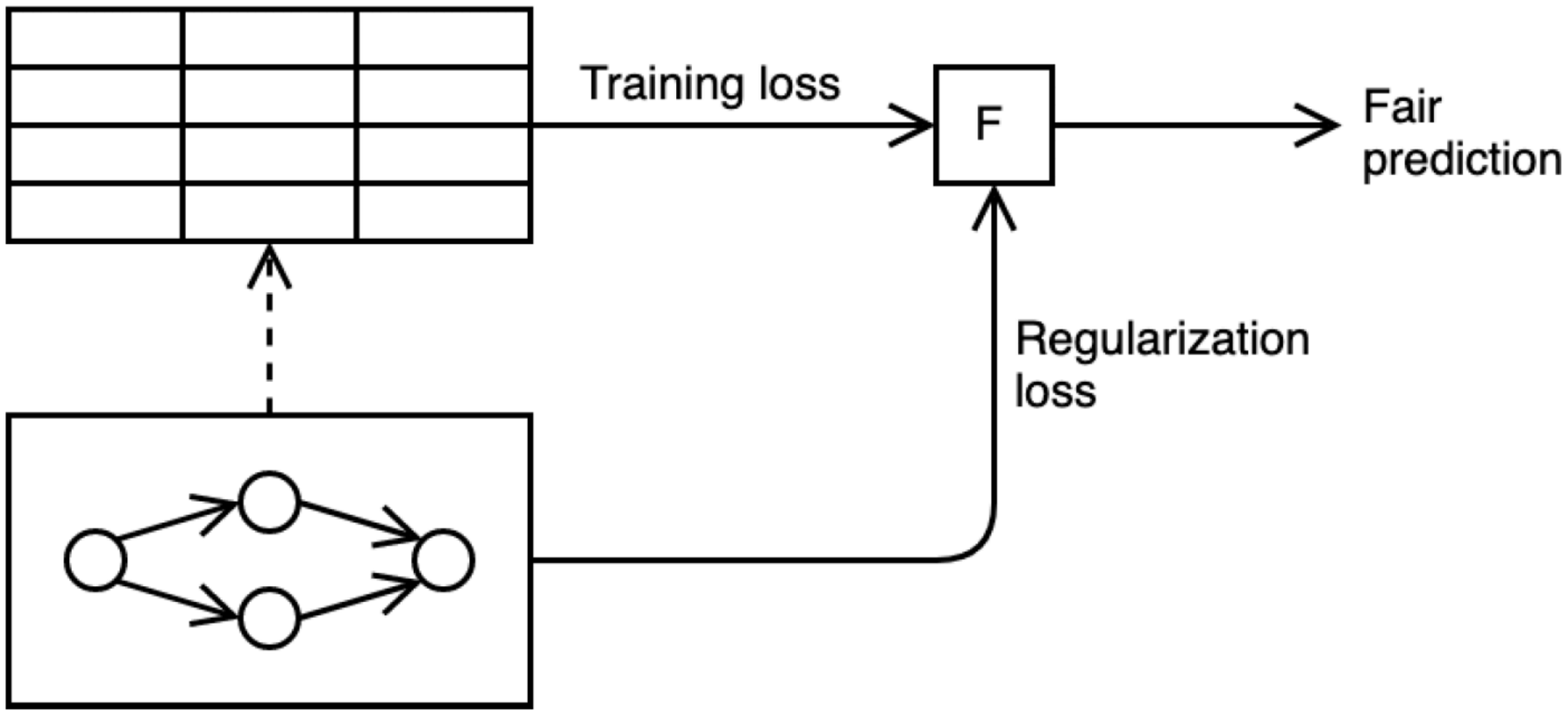
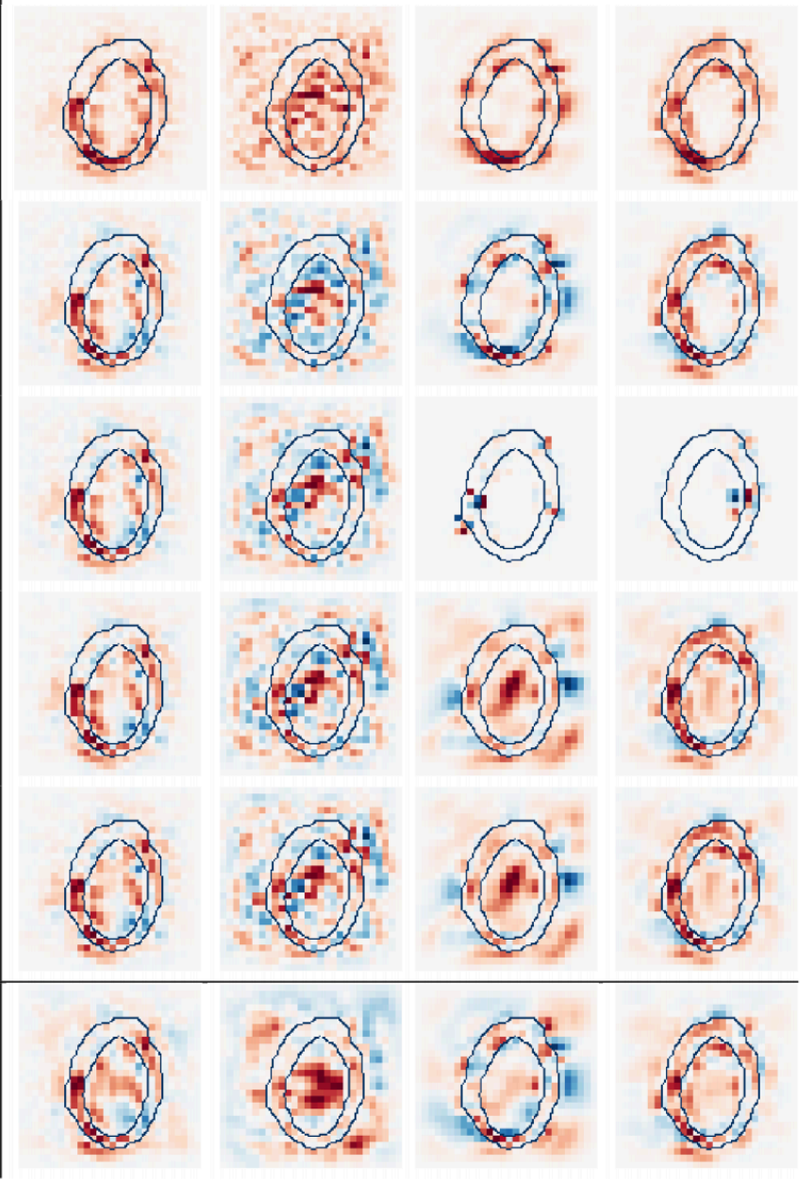


Figure 1: Counterfactual Fairness regularization diagram.

Image classification explanations

Method	Attribution $R_i^c(x)$
Saliency Maps	$\left\lvert \frac{\partial S_c(x)}{\partial x_i} \right\rvert$
Gradient * Input	$x_i \cdot \frac{\partial S_c(x)}{\partial x_i}$
ϵ-LRP	$x_i \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z)}{z}$
DeepLIFT	$(x_i - \bar{x_i}) \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z) - f(\bar{z})}{z - \bar{z}}$
Integrated Gradient	$(x_i - \bar{x_i}) \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\tilde{x})}{\partial(\tilde{x_i})}\Big\rvert_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$
Occlusion-1	$x_i \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\tilde{x})}{\partial(\tilde{x_i})}\Big\rvert_{\tilde{x}=x_{[x_i=\alpha \cdot x_i]}} d\alpha$

Example of attributions on MNIST



Explaining visual models by causal attribution

By producing contrastive explanations that are an answer to counterfactual questions such as:

> Given the fact that this face belongs to a woman, has been classified as a woman and the person does not have a beard, how would the classifier's prediction have changed had there been a beard?

https://arxiv.org/abs/1909.08891

Thank you for your attention!