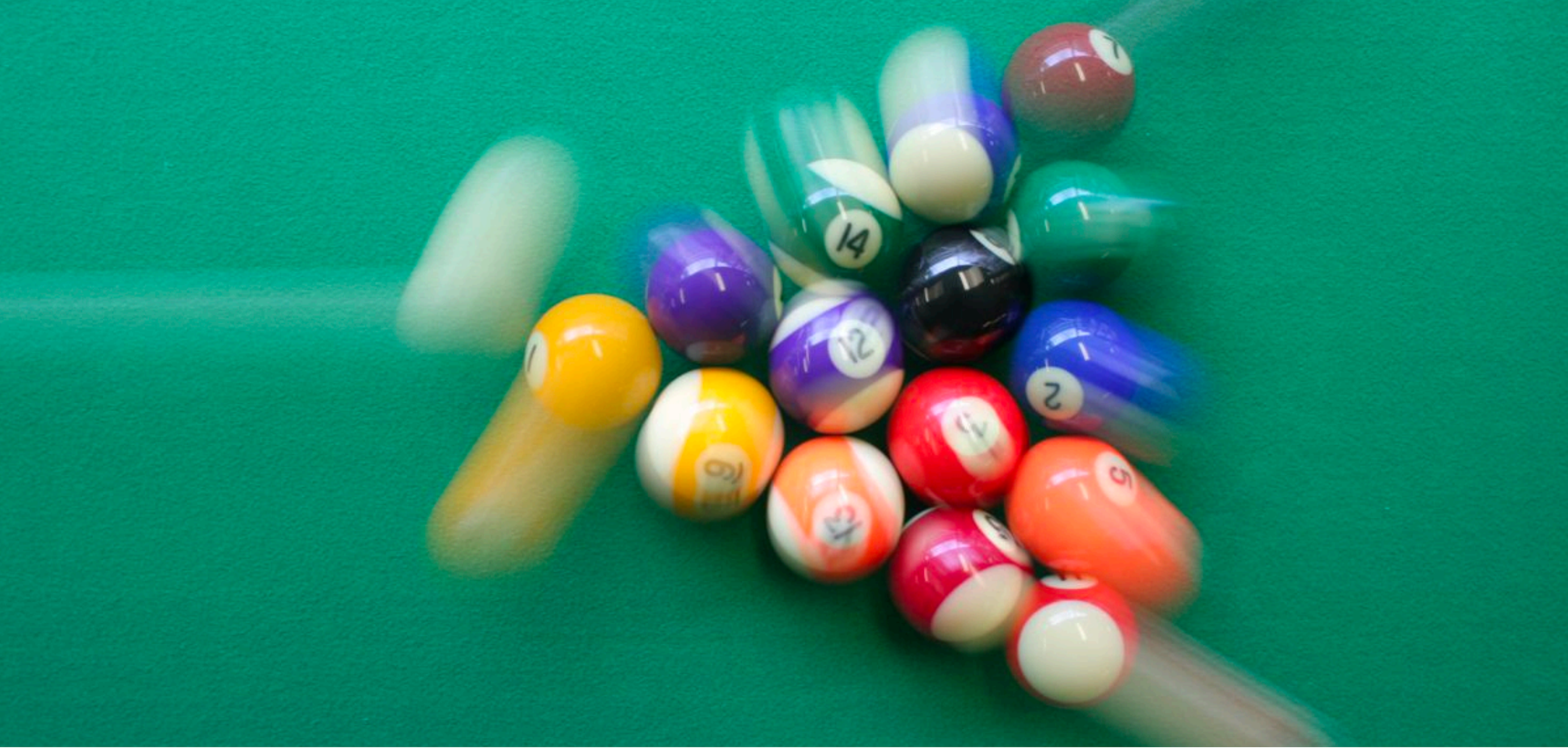


Causal Data Science

Jordi Vitrià

UNIVERSITAT DE BARCELONA

Postgraduate Course in Data Science and Big Data

Causal Data Science

Data science tasks:

- **Description** is using data to provide a quantitative summary of certain features of the world. Descriptive tasks include, for example, computing the proportion of individuals with diabetes in a large healthcare database and representing social networks in a community.

- **Prediction** (or association) is using data to map some features of the world (the inputs) to other features of the world (the outputs).

- **Counterfactual prediction** is using data to predict certain features of the world if the world had been different, as is required in **causal inference** applications. An example of causal inference is the estimation of the mortality rate that would have been observed if all individuals in a study population had received screening for colorectal cancer vs. if they had not received screening.

Miguel A. Hernán, John Hsu, Brian Healy. Data science is science's second chance to get causal inference right: A classification of data science tasks. arXiv:1804.10846v5

- Some methodologists have referred to the causal inference task as "**explanation**", but this is a somewhat misleading term because causal effects may be quantified while remaining unexplained.
- Statistical inference is often required in all three tasks. For example, one might want to add 95% confidence intervals around descriptive, predictive, or causal estimates involving samples of target populations.

Miguel A. Hernán, John Hsu, Brian Healy. Data science is science's second chance to get causal inference right: A classification of data science tasks. arXiv:1804.10846v5

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 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!

How to answer a WHY question...

https://www.youtube.com/watch?v=Q1lL-hXO27Q

Causal Data Science

Difficult questions for a data scientist:

- Should I add this variable to my model?
- Why does this counter-intuitive variable show up as a predictive one?
- Why does this variable suddenly become insignificant if I add another variable?
- Why does the direction of the correlation being opposite to what you think?
- Why is the correlation zero when I thought it would be higher?
- Why does the direction of relationship reverse when I dis-aggregate the data into sub-population?

To predict something in the future, there are two ways:

- I know that when I see X, I will see Y (Association)
- I know that X causes Y (Causality)

Both ways can predict. Both ways might yield similar model performance.

Miguel A. Hernán, John Hsu, Brian Healy. Data science is science's second chance to get causal inference right: A classification of data science tasks. arXiv:1804.10846v5

Prediction vs Causal Inference

Data science has excelled at commercial applications, such as shopping and movie recommendations, credit rating, stock trading algorithms, and advertisement placement. All these applications of data science have one thing in common: **they are predictive, not causal**.

Mapping observed inputs to observed outputs is a natural candidate for automated data analysis because this task only requires 1) a large dataset with inputs and outputs, 2) an algorithm that establishes a mapping between inputs and outputs, and 3) a metric to assess the performance of the mapping, often based on a gold standard.

The role of expert knowledge is the key difference between prediction and causal inference tasks. 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. Answering a causal question typically requires a combination of data, analytics, and expert causal knowledge.

Miguel A. Hernán, John Hsu, Brian Healy. Data science is science's second chance to get causal inference right: A classification of data science tasks. arXiv:1804.10846v5

Example


Pel meu futur fill, des d'avui mateix
Respiro Salut
CSB Consorci Sanitari de Barcelona
Agència de Salut Pública
Generalitat de Catalunya
Departament de Salut
Pla Director d'Oncologia
embaràs sense fum

Example

Suppose we want to use a large health records database to **predict infant mortality** (the output) using clinical and lifestyle factors collected during pregnancy (the inputs).

We have just applied our expert knowledge to decide what the output and inputs are, and to select a particular database.

At this point of the process our expert knowledge will not be needed any more: **an algorithm can provide a mapping between inputs and outputs at least as good as any mapping we could propose and, in many cases, astoundingly better**.

But now suppose we want to use the same health records database to **determine the causal effect of maternal smoking during pregnancy on the risk of infant mortality**.

Miguel A. Hernán, John Hsu, Brian Healy. Data science is science's second chance to get causal inference right: A classification of data science tasks. arXiv:1804.10846v5

Example

A key problem is **confounding**: pregnant women who do and do not smoke differ in many characteristics (e.g., alcohol consumption, diet, access to adequate prenatal care) that affect the risk of infant mortality.

Therefore, a causal analysis needs to identify and **adjust for those confounding** factors which, by definition, are associated with both maternal smoking and infant mortality.

In a controlled clinical trial, the way of adjusting is to ensure that the study groups do not differ with respect to possible confounders by matching the two comparison groups.

Miguel A. Hernán, John Hsu, Brian Healy. Data science is science's second chance to get causal inference right: A classification of data science tasks. arXiv:1804.10846v5

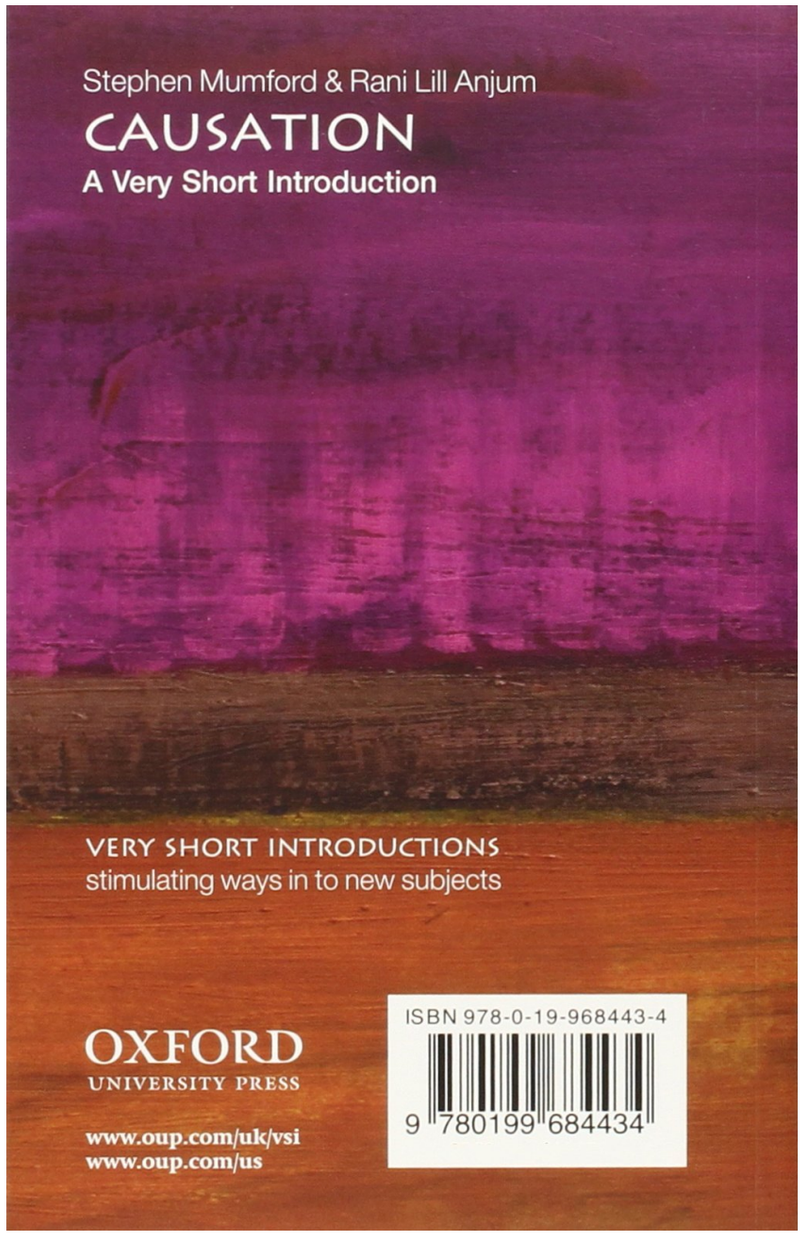
Example

However, not all factors associated with maternal smoking and infant mortality are confounders that should be adjusted for.

For example, birthweight is strongly associated with both maternal smoking and infant mortality, but adjustment for birthweight induces bias because birthweight is a risk factor that is itself causally affected by maternal smoking.

Miguel A. Hernán, John Hsu, Brian Healy. Data science is science's second chance to get causal inference right: A classification of data science tasks. arXiv:1804.10846v5

Philosophy



https://twitter.com/EpiEllie/status/1041369711659962370

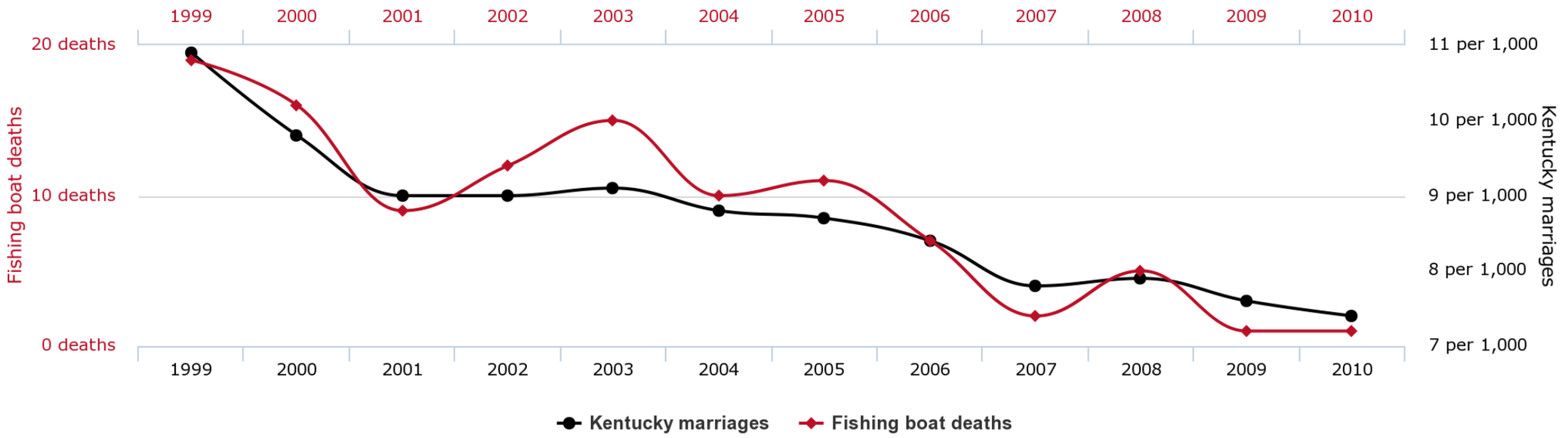
Correlation and Causation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

People who drowned after falling out of a fishing boat

correlates with

Marriage rate in Kentucky



http://tylervigen.com/spurious-correlations

Correlation and Causation

Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

◆ INTRODUCTION ◆

Introductory statistics textbooks routinely warn of the dangers of confusing correlation with causation, pointing out that while a high correlation coefficient is indicative of (linear) association, it cannot be taken as a measure of causation. Such warnings are typically accompanied by illustrative examples, such as the correlation between the reading skills of children and their shoe size, or the apparent relationship between educational level and unemployment (see e.g. Freedman *et al.* 1998). However, such examples are often either trivially explained via an obvious confounder (e.g. age, in the case of reading age and shoe size) or are not obviously cases of mere association (e.g. educational level may indeed be at least partly responsible for time spent unemployed). In what follows, I give an example based on genuine data of an association which is clearly ludicrous, but which cannot be so easily dismissed as non-causal via an obvious confounder.

My starting point is the familiar folk tale that babies are delivered by storks. The origins of this connection are believed to lie partly in the association between storks and the concept of women as bringers of life, and also in the bird's feeding habits, which were once regarded as a search for embryonic life in water (Cooper 1992). The legend lives on to this day, with neonate-bearing storks being a regular feature of greetings cards celebrating births.

While it is (I trust) obvious that the legend is complete nonsense, it is legitimate to ask precisely how one might set about refuting it scientifically. If one were approaching the question in the same way that many other links are investigated (e.g. suspected links between diet and cancer risk), one may well decide to carry out a correlational study, to see if the number of storks in a country bears a simple relationship to the number of human births in that country. Although the presence of a statistically significant degree of correlation cannot be taken to imply causation, its absence would certainly constitute evidence against a simple relationship. This possibility can quickly be investigated in the present case using standard hypothesis testing, with the null hypothesis being the absence of any correlation between the number of storks and the number of live births in a particular country. This I now proceed to do.

◆ TESTING THE STORK-BIRTH RELATIONSHIP ◆

The white stork (*Ciconia ciconia*) is a surprisingly common bird in many parts of Europe, and data on the number of breeding pairs are available for 17 European countries (Harbard 1999, pers. comm.); the latest figures, covering the period from 1980 to 1990, are given in table 1, along with demographic data taken from Britannica Yearbook for 1990.

Plotting the number of stork pairs against the number of births in each of the 17 countries, one can discern signs of a possible correlation between the two (see figure 1).

The existence of this correlation is confirmed by performing a linear regression of the annual number of births in each country (the final column in table 1) against the number of breeding pairs of white storks (column 3). This leads to a correlation coefficient of $r = 0.62$, whose statistical significance can be gauged using the standard t-test, where $t = r \cdot \sqrt{[(n-2)/(1-r^2)]}$ and n is the sample size. In our case, $n = 17$ so that $t = 3.06$, which for $(n-2) = 15$ degrees of freedom leads to a p-value of 0.008.

Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

◆ ANALYSIS ◆

What are we to make of this result, which points to a highly statistically significant degree of correlation between stork populations and birth rates? The correlation coefficient is not particularly high, but according to its *p*-value, there is only a 1 in 125 chance of obtaining at least as impressive a value *assuming* the null hypothesis of no correlation were true. Yet as with any *p*-value (and contrary to what unwary users of them believe),

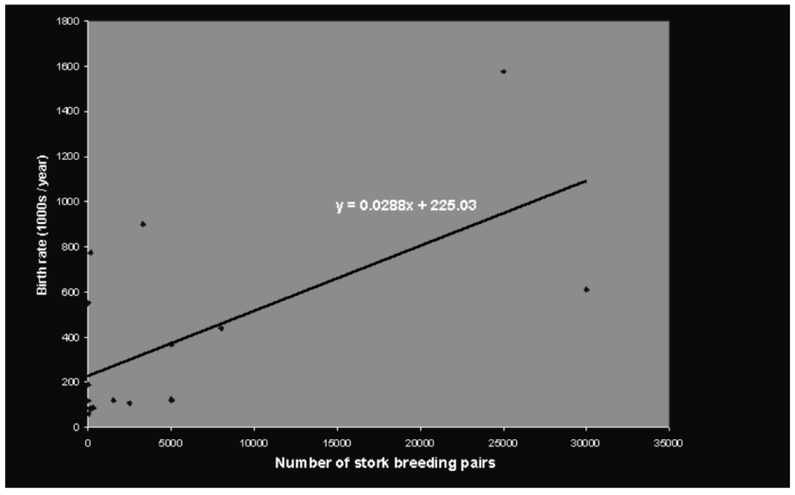


Fig 1. How the number of human births varies with stork populations in 17 European countries.

Correlation and Causation

Correlation is a statistical technique which tells us how strongly the pair of variables are linearly related and change together.

Causation takes a step further than **correlation**.

Causation says any change in the value of one variable will **cause** a change in the value of another variable, which means one variable makes other to happen. It is also referred as cause and effect.

CORRELATION **CAN** IMPLY CAUSATION (but not often)

CAUSATION **DOES NOT IMPLY** (LINEAR) CORRELATION

(Causation implies high mutual information)

Simpson's Paradox

Example:

Since 2000, the median US wage has risen about 1%, adjusted for inflation.

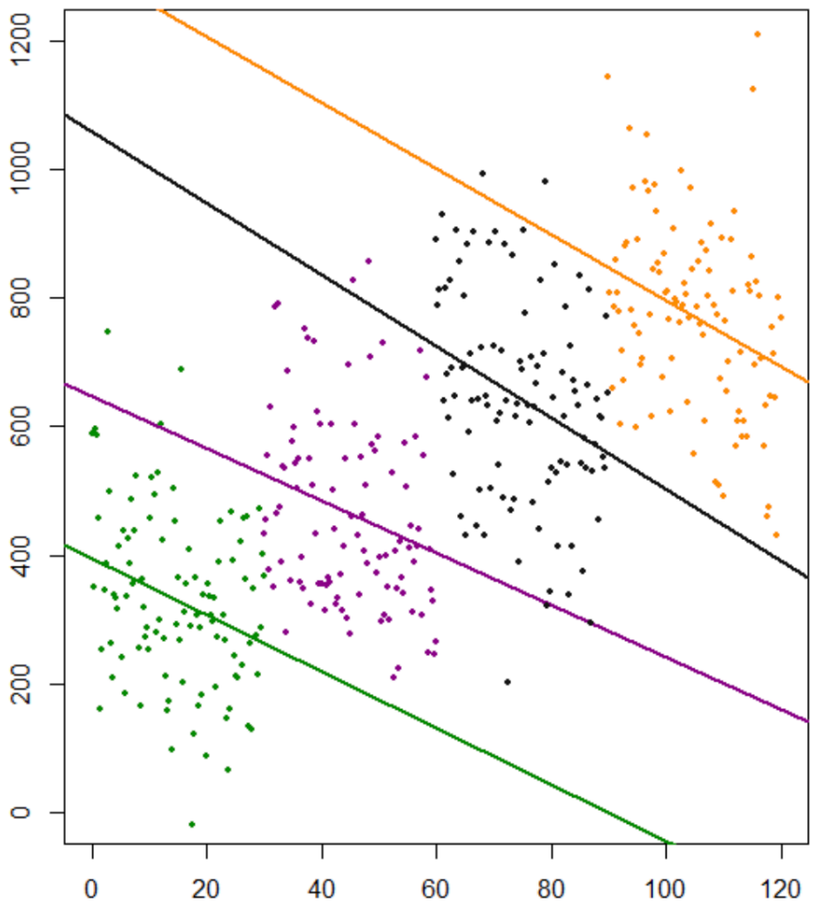
But over the same period, the median wage for:

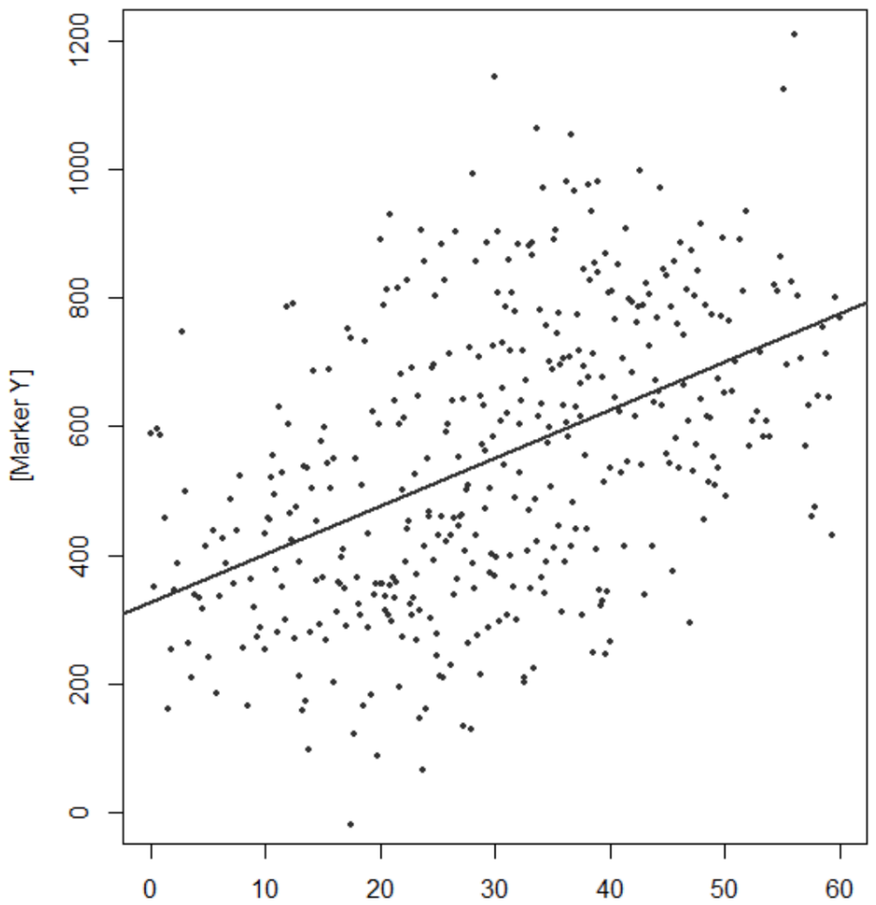
- high school dropouts,
- high school graduates with no college education,
- people with some college education, and
- people with Bachelor's or higher degrees

have all decreased.

In other words, within every educational subgroup, the median wage is lower now than it was in 2000.

https://blog.revolutionanalytics.com/2013/07/a-great-example-of-simpsons-paradox.html





Explanation:

There are now many more college graduates (who get higher-paying jobs) than there were in 2000, but wages for college graduates collectively have fallen at a much slower rate (down 1.2%) than for those of lower educational attainment (whose wages have fallen precipitously, down 7.9% for high school dropouts). The growth in the proportion of college graduates swamps the wage decline for specific groups.

Simpson’s Paradox

Example:

Suppose you're suffering from kidney stones and go to see your doctor. The doctor tells you two treatments are available, treatment A and treatment B. You ask which treatment works better, and the doctor says "*Well, a study found that treatment A has a higher probability of success than treatment B*."

http://michaelnielsen.org/reinventing_explanation/index.html

Simpson’s Paradox

Example:

You start to say "*I'll take treatment A, thanks!*", when your doctor interrupts: "*But the same study also looked to see which treatment worked better, depending on whether patients had large kidney stones or small kidney stones.*"

You say "*Well, do I have large kidney stones or small kidney stones*"? As you speak the doctor interrupts again, looking sheepish, and says "*Actually, it doesn't matter. You see, they found that treatment B has a higher probability of success than treatment A, regardless of whether you have large or small kidney stones.*"

http://michaelnielsen.org/reinventing_explanation/index.html

Simpson’s Paradox

Example:

It sounds impossible. But it's true: an actual study was done in which treatment TB was found to work with higher probability than treatment TA, for both large and small kidney stones, despite the fact that treatment TA works with higher overall probability than Treatment TB.

Here's the numbers from the study:

C. R. Charig, D. R. Webb, S. R. Payne, O. E. Wickham (March 1986)	**Treatment TA helps**	**Treatment TB helps**
Large kidney stones	69% (55 / 80)	**73%** (192 / 263)
Small kidney stones	87% (234 / 270)	**93%** (81 / 87)
All patients	**83%** (289 / 350)	78% (273 / 350)

http://michaelnielsen.org/reinventing_explanation/index.html

Simpson’s Paradox

The practical significance of Simpson's paradox surfaces in **decision making situations** where it poses the following dilemma:

Which data should we consult in choosing an action, the aggregated or the partitioned?

Simpson’s Paradox

In the Kidney Stone example above, it is clear that if one is diagnosed with "Small Stones" or "Large Stones" the data for the respective subpopulation should be consulted and Treatment TB would be preferred to Treatment TA.

But what if a patient is not diagnosed, and the size of the stone is not known; would it be appropriate to consult the aggregated data and administer Treatment TA?

This would stand contrary to common sense; a treatment that is preferred both under one condition and under its negation should also be preferred when the condition is unknown.

Simpson’s Paradox

On the other hand, if the partitioned data is to be preferred a priori, what prevents one from partitioning the data into arbitrary sub-categories (say based on eye color or post-treatment pain) artificially constructed to yield wrong choices of treatments?

It can be shown that, indeed, in many cases it is the aggregated, not the partitioned data that gives the correct choice of action. Worse yet, given the same figures, one should sometimes follow the partitioned and sometimes the aggregated data, depending on the **story behind the data**, with each story dictating its own choice.

If a test reveals a patient has kidney stones but gives me no information about their size. Which treatment do I recommend? Is there any accepted resolution to this problem?

Solving Simpson's Paradox by using Causal Reasoning

There are three ways to assess the existence of a **causal relationship** between two variables, X and Z:

$$X \rightarrow Z$$

The easiest way is an **intervention**: You randomly force X to have different values and you measure Z. This is what we do in randomized clinical trial or in an A/B Test.

This is not always feasible (because of ethical or economic reasons)

Solving Simpson's Paradox by using Causal Reasoning

The second way is the **front door method**. You want to show that X acts on Z via Y, i.e., $X \rightarrow Y \rightarrow Z$.

If you assume that Y is potentially caused by X but has no other causes, and you can measure that Y is correlated with X and Z is correlated with Y, then you can conclude causality must be flowing via Y.

For example: X is smoking, Z is cancer, Y is tar accumulation. Tar can only come from smoking, and it correlates with both smoking and cancer. Therefore, smoking causes cancer via tar.

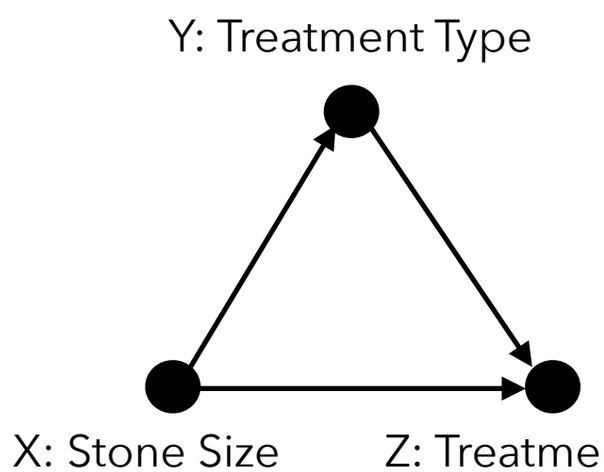
Solving Simpson's Paradox by using Causal Reasoning

The third way is the **back door method**. You want to show that X and Z aren't correlated because of a "back door", e.g. common cause, i.e., $X \leftarrow Y \rightarrow Z$.

Since you have assumed a causal model, you merely need to block all of the paths (by observing variables and **conditioning** on them) that evidence can flow up from X and down to Z. It's a bit tricky to block these paths, but there is a clear algorithm that lets you know which variables you have to observe to block these paths.

Solving Simpson's Paradox by using Causal Reasoning

In our problem, **kidney stone size** X and **treatment type** Y are both causes of success Z. X may be a cause of Y if other doctors are assigning treatment based on kidney stone size. Clearly there are no other causal relationships between X, Y, and Z. Y comes after X so it cannot be its cause. Similarly Z comes after X and Y. **Since X is a common cause (confounding variable), it should be measured**.



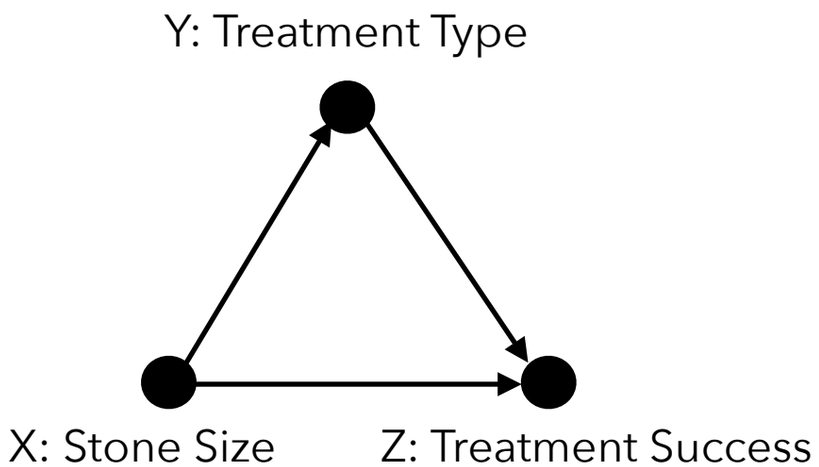
Causal Diagram

Solving Simpson's Paradox by using Causal Reasoning

X (stone size) is confounder of Y (treatment type) and Z (treatment success). For an unbiased estimate of of the effect of Y on Z we must adjust the confounder.

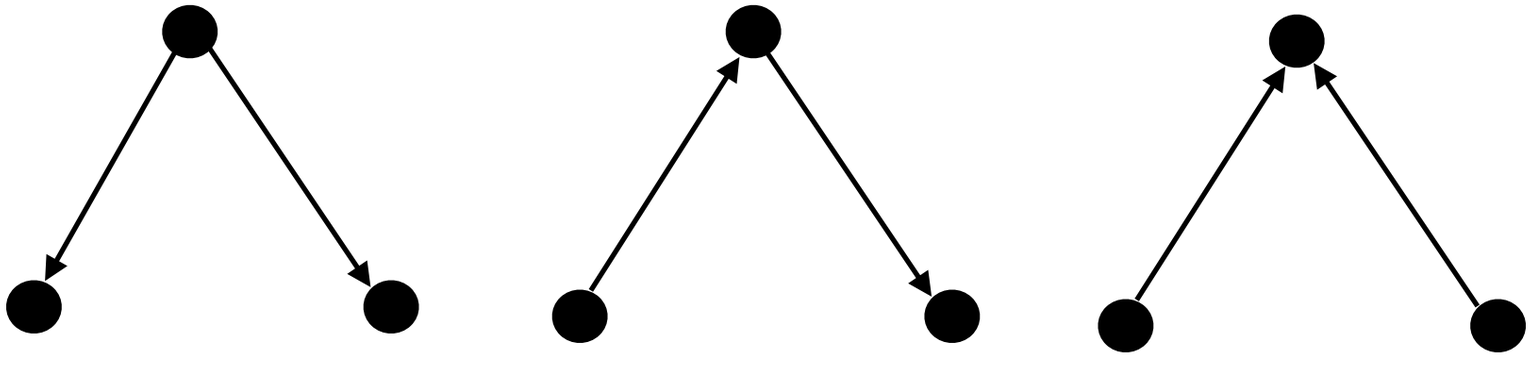
We can do it by looking at the data for X separately, then taking the average.

Solving Simpson's Paradox by using Causal Reasoning



	Treatment TA helps	Treatment TB helps
Large kidney stones	69% (55 / 80)	**73%** (192 / 263)
Small kidney stones	87% (234 / 270)	**93%** (81 / 87)
Average	78%	**83%**

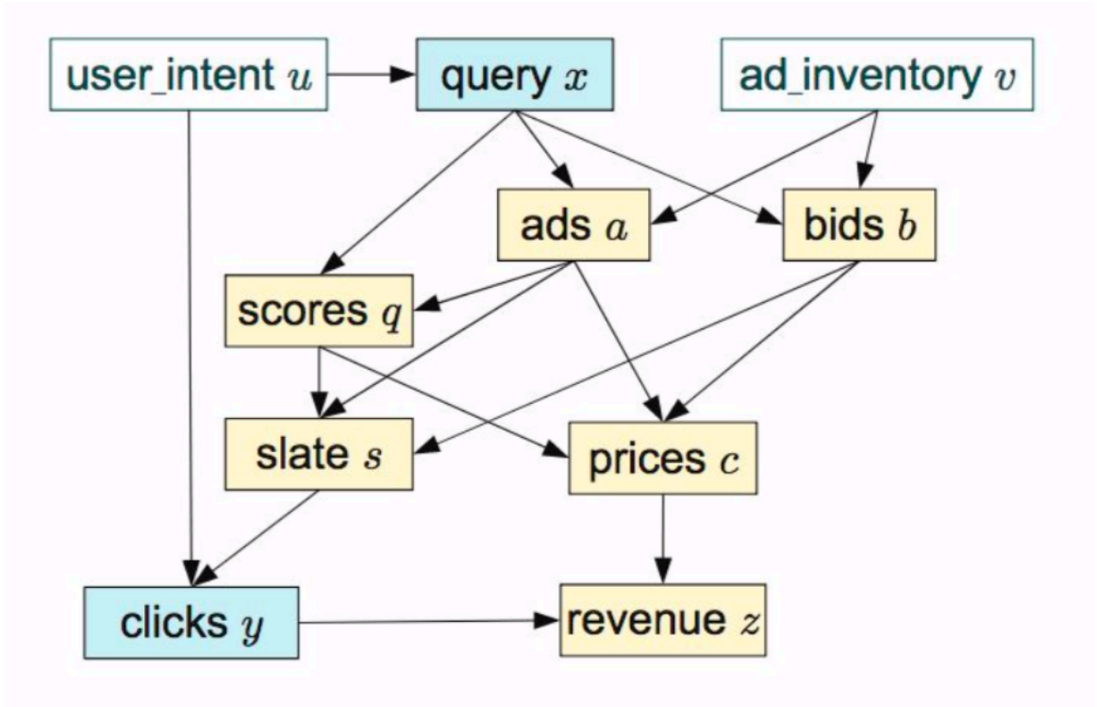
Other situations



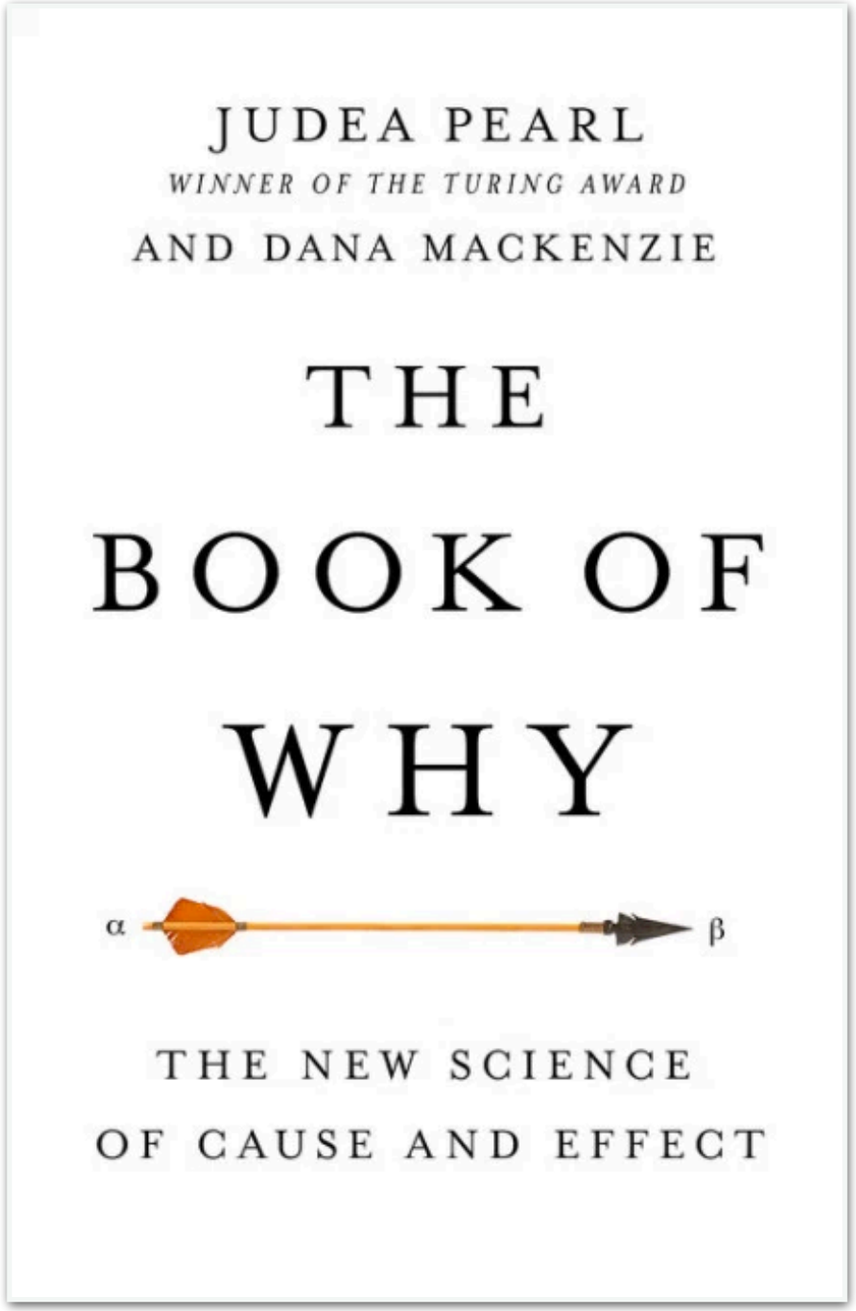
Confounder

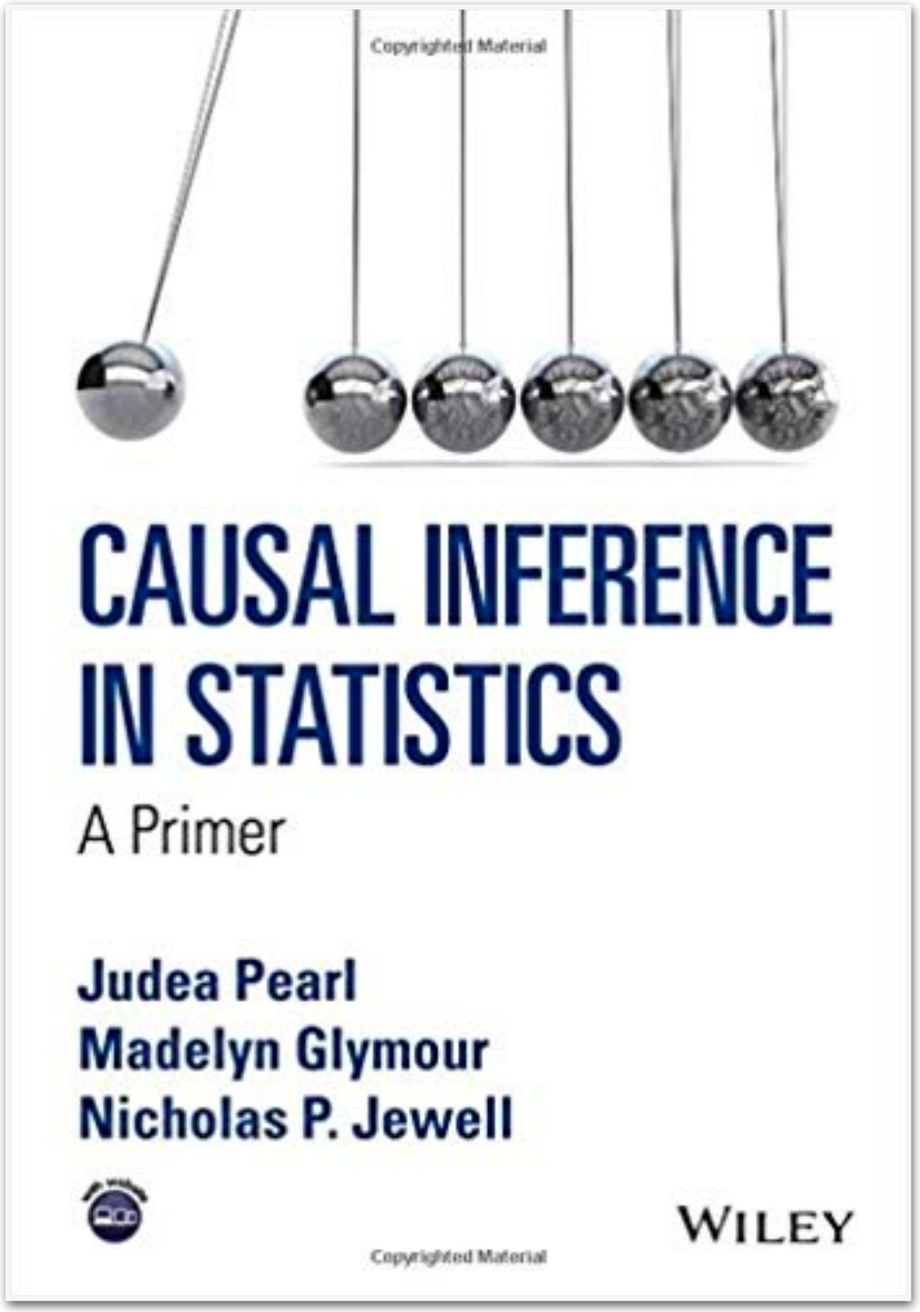
Mediator

Collider



Bottou, Léon, et al. "Counterfactual reasoning and learning systems: the example of computational advertising." *The Journal of Machine Learning Research* 14.1 (2013): 3207-3260.


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

The ladder of knowledge

J.Pearl proposes a "**ladder of knowledge**" with three levels:

- **Association**: How would seeing X change my belief in Y? → Seeing smoke raises the probability of fire.

- **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 my expected lifespan change if I become a vegetarian?

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

The ladder of knowledge

- The first level, dealing with associations, is studied using the rules from **probability theory** and can be learned from data using **statistical methods**.

- The second level deals with **interventions**. To assess the effect of interventions, one either has to perform a suitable **experiment** (which might be expensive or even not possible) or one has to be able to **reason about the causal structure** about the variables of the system.

- The final level is even more difficult to model, as it deals with reality as it could be if circumstances were different. By definition, there is no data available, nor could we ever perform experiments. In order to make statements about such hypothetical situations, we need an intricate understanding of the system and how everything is linked together.

Example

Credit: https://michielstock.github.io/causality/

- Suppose that we want to model soil, plants, insects and birds of a type of ecosystem. To keep things simple, these four properties are represented by binary stochastic variables soil (S), plants (P), insects (I) and birds (B).

- This means that there are only two states for every variable. In this case, the soil can be poor (S=0) or rich (S=1) in nutrients and we model the presence (X=1) or absence (X=0) of one species of plants, insects and birds, respectively.

- We are interested in how the presence of the plant species influences the presence of the insect species.

Example

Credit: https://michielstock.github.io/causality/

Collecting data

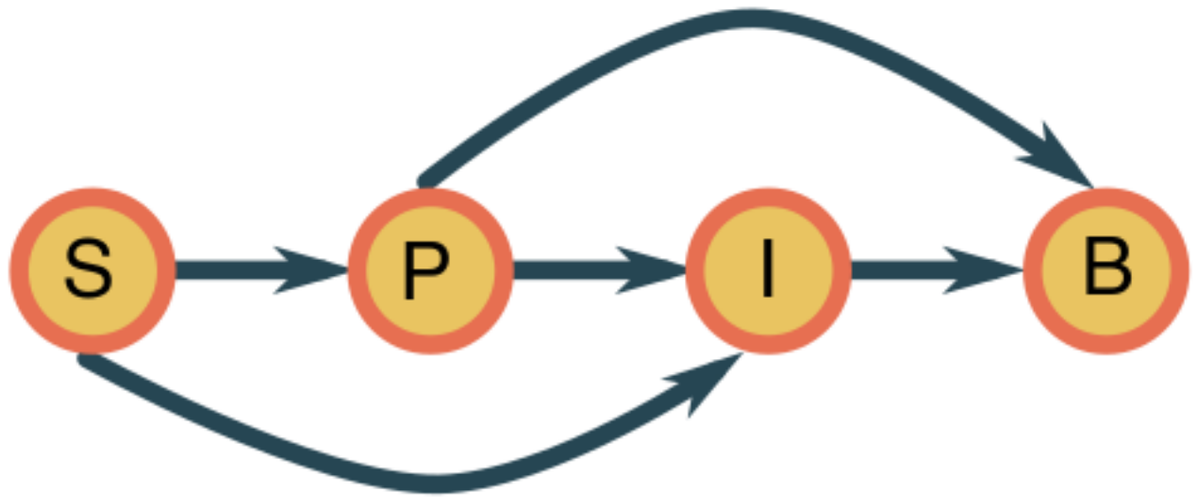
Soil	Plants	Insects	Birds
0	1	1	1
1	1	0	1
...	...	...	...
0	0	0	0

In more realistic situations, a we might use a much more richer description of the variables, such as a real value or even a vectorial description of for example which kinds of plants are present in which quantity.

Example

Credit: https://michielstock.github.io/causality/

The causal diagram



This means the joint probability distribution factorizes as:

$$\Pr(S, P, I, B) = \Pr(S)\Pr(P \mid S)\Pr(I \mid S, P)\Pr(B \mid P, I)$$

This is expert knowledge!

Example

Credit: https://michielstock.github.io/causality/

Association

Given that I observe a plot with the plant species present, are there likely birds?

This question corresponds to the standard conditional probability that B=1 given P=1. This can be stated and computed as:

$$\Pr(B = 1 \mid P = 1) = \frac{\Pr(B = 1 \cap P = 1)}{\Pr(P = 1)}$$

The two quantities in the fraction can readily be estimated from the table of data. If the number of variables increases we have to combat the curse of dimensionality using more sophisticated parametric methods: logistic regression, support vector machines, random forests and the like.

Example

Credit: https://michielstock.github.io/causality/

Causation

What happens if we seed some of the plants, forcing P=1. Does this have a large effect on the probability of B=1? This question can be denoted as:

$$\Pr(B = 1 \mid do(P) = 1)$$

where $do(P)=1$ means that **we set** P to 1.

Pearl et al have showed that there is a way for removing this operator from the query and reformulate it using only standard probability expressions. This means that we can estimate such effects from the data - provided that we have a causal diagram!

Example

Credit: https://michielstock.github.io/causality/

Counterfactuals

A certain plot has neither plants nor birds, would there be birds if plants were present?

This query is formalized as the counterfactual outcome:

$$\Pr(B = 1 \mid S = s, P = 0, I = i, P^* = 1)$$

indicating the probability of having birds if there were plants present (indicated by *P**).

To compute the above quantity, one needs additional rules to translate the expression in classical probability terms. If you are lucky, it is possible to obtain an expression for the counterfactual distribution which can be estimated from data.

Microsoft / **dowhy**

Watch 52 | Star 614 | Fork 64

Code | Issues 3 | Pull requests 0 | Projects 0 | Wiki | Insights

DoWhy is a Python library that makes it easy to estimate causal effects. DoWhy is based on a unified language for causal inference, combining causal graphical models and potential outcomes frameworks. http://causalinference.gitlab.io/dowhy

causal-inference | python3 | graphical-models

92 commits | **1** branch | **0** releases | **18** contributors | MIT

Branch: **master** | New pull request | Create new file | Upload files | Find file | Clone or download

amit-sharma Merge pull request #29 from akelleh/add_gformula_estimator ... Latest commit `66b1fa3` 2 hours ago

docs	consistent naming for examples	3 days ago
dowhy	Merge pull request #29 from akelleh/add_gformula_estimator	2 hours ago
tests	added tests for all except RD estimator	3 days ago
.gitignore	Initial commit	8 months ago
LICENSE	Initial commit	8 months ago
Makefile	added initial code	7 months ago
README.rst	Merge branch 'master' into GrammarEnhancements	3 months ago
requirements.txt	Add pydot to requirements.txt	3 months ago
setup.py	Merge pull request #14 from kazigk/master	4 months ago

Causality and conditional probabilities

inFERENCe

posts on machine learning, statistics, opinions on things I'm reading in the space

Home

May 24, 2018

ML beyond Curve Fitting: An Intro to Causal Inference and do-Calculus

You might have come across **Judea Pearl's new book**, and a **related interview** which was widely shared in my social bubble. In the interview, Pearl dismisses most of what we do in ML as curve fitting. While I believe that's an overstatement (conveniently ignores RL for example), it's a nice reminder that most productive debates are often triggered by controversial or outright arrogant comments. Calling machine learning alchemy was a great recent example. After reading the article, I decided to look into his famous do-calculus and the topic causal inference once *again*.

Again, because this happened to me semi-periodically. I first learned do-calculus in a (very unpopular but advanced) undergraduate course Bayesian networks. Since then, I have re-encountered it every 2-3 years in various contexts, but somehow it never really struck a chord. I always just thought "this stuff is difficult and/or impractical" and eventually forgot about it and moved on. I never realized how fundamental this stuff was, until now.

This time around, I think I fully grasped the significance of causal reasoning and I turned into a full-on believer. I know I'm late to the game but I almost think it's basic hygiene for people working with data and conditional probabilities to understand the basics of this toolkit, and I feel embarrassed for completely ignoring this throughout my career.

In this post I'll try to explain the basics, and convince you why you should think about this, too. If you work on deep learning, that's an even better reason to understand this. Pearl's comments may be unhelpful if interpreted as contrasting deep learning with causal inference. Rather, you should interpret it as highlighting causal inference as a huge, relatively underexplored, application of deep learning. Don't get discouraged by causal diagrams looking a lot like Bayesian networks (not a coincidence seeing they were both pioneered by Pearl) they don't compete with, they complement deep learning.

To set things up, let's say we have i.i.d. data sampled from some joint $p(x, y, z, \ldots)$. Let's assume we have lots of data and the best tools (say, deep networks) to fully estimate this joint distribution, or any property, conditional or marginal distribution thereof. In other words, let's assume p is known and tractable. Say we are ultimately interested in how variable y behaves given x. At a high level, one can ask this question in two ways:

- observational $p(y|x)$: What is the distribution of Y given that I **observe** variable X takes value x. This is what we usually estimate in supervised machine learning. It is a conditional distribution which can be calculated from $p(x, y, z, \ldots)$ as a ratio of two of its marginals. $p(y|x) = \frac{p(x,y)}{p(x)}$. We're all very familiar with this object and also know how to estimate this from data.

- interventional $p(y|do(x))$: What is the distribution of Y if I were to **set** the value of X to x. This describes the distribution of Y I would observe if I intervened in the data generating process by artificially forcing the variable X to take value x, but otherwise simulating the rest of the variables according to the original process that generated the data. (note that the data generating procedure is NOT the same as the joint distribution $p(x, y, z, \ldots)$ and this is an important detail).

Aren't they the same thing?

No. $p(y|do(x))$ and $p(y|x)$ are not generally the same, and you can verify this with several simple thought experiments. Say, Y is the pressure in my espresso machine's boiler which ranges roughly between 0 and 1.1 bar depending on how long it's been turned on. Let X be the reading of the built-in barometer. Let's say we jointly observe X and Y at random times. Assuming the barometer functions properly $p(y|x)$ should be a unimodal distribution centered around x, with randomness due to measurement noise. However, $p(y|do(x))$ won't actually depend on the value of x and is generally the same as $p(y)$, the marginal distribution of boiler pressure. This is because artificially setting my barometer to a value (say, by moving the needle) won't actually cause the pressure in the tank to go up or down.

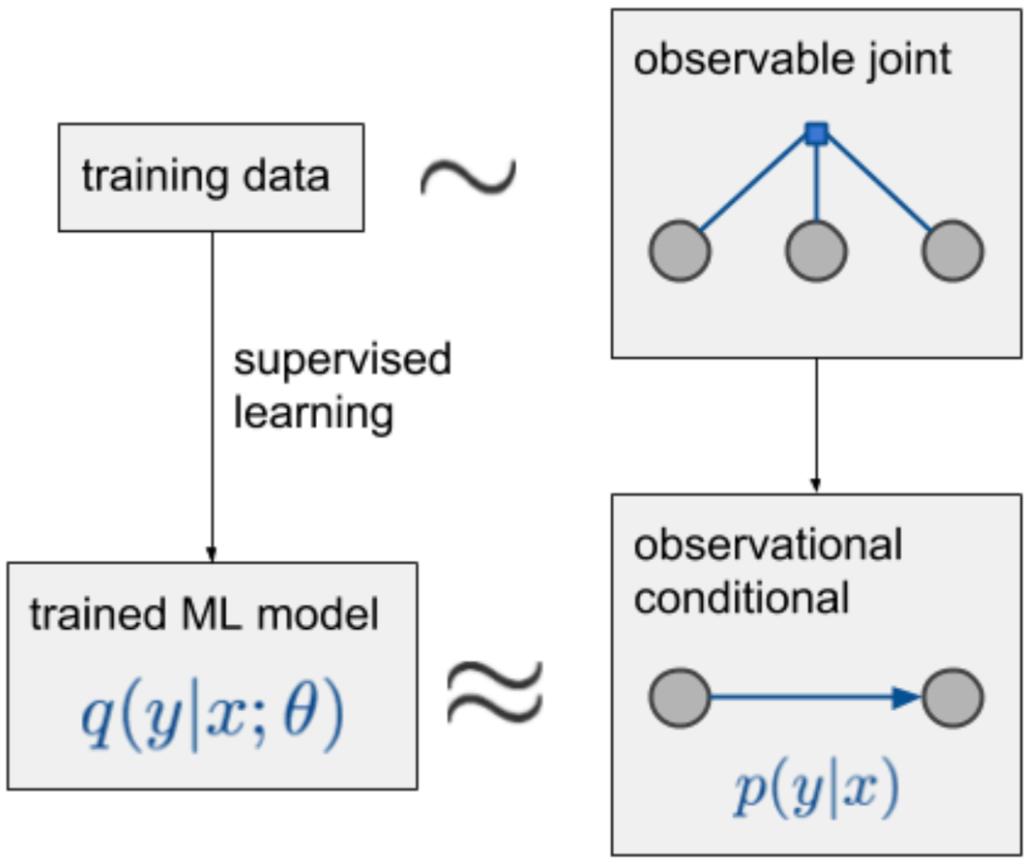
Which one do I want?

Depending on the application you want to solve, you should seek to estimate one of these conditionals. If your ultimate goal is diagnosis or forecasting (i.e. observing a naturally occurring x and inferring the probable values of y) you want the observational conditional $p(y|x)$. This is what we already do in supervised learning, this is what Judea Pearl called curve fitting. This is all good for a range of important applications such as classification, image segmentation, super-resolution, voice transcription, machine translation, and many more.

In applications where you ultimately want to control or choose x based on the conditional you estimated, you should seek to estimate $p(y|do(x))$ instead. For example, if x is a medical treatment and y is the outcome, you are not merely interested in observing a naturally occurring treatment x and predicting the outcome, we want to *proactively choose* the treatment x given our understanding of how it effects the outcome y. Similar situations occur in system identification, control and online recommender systems.

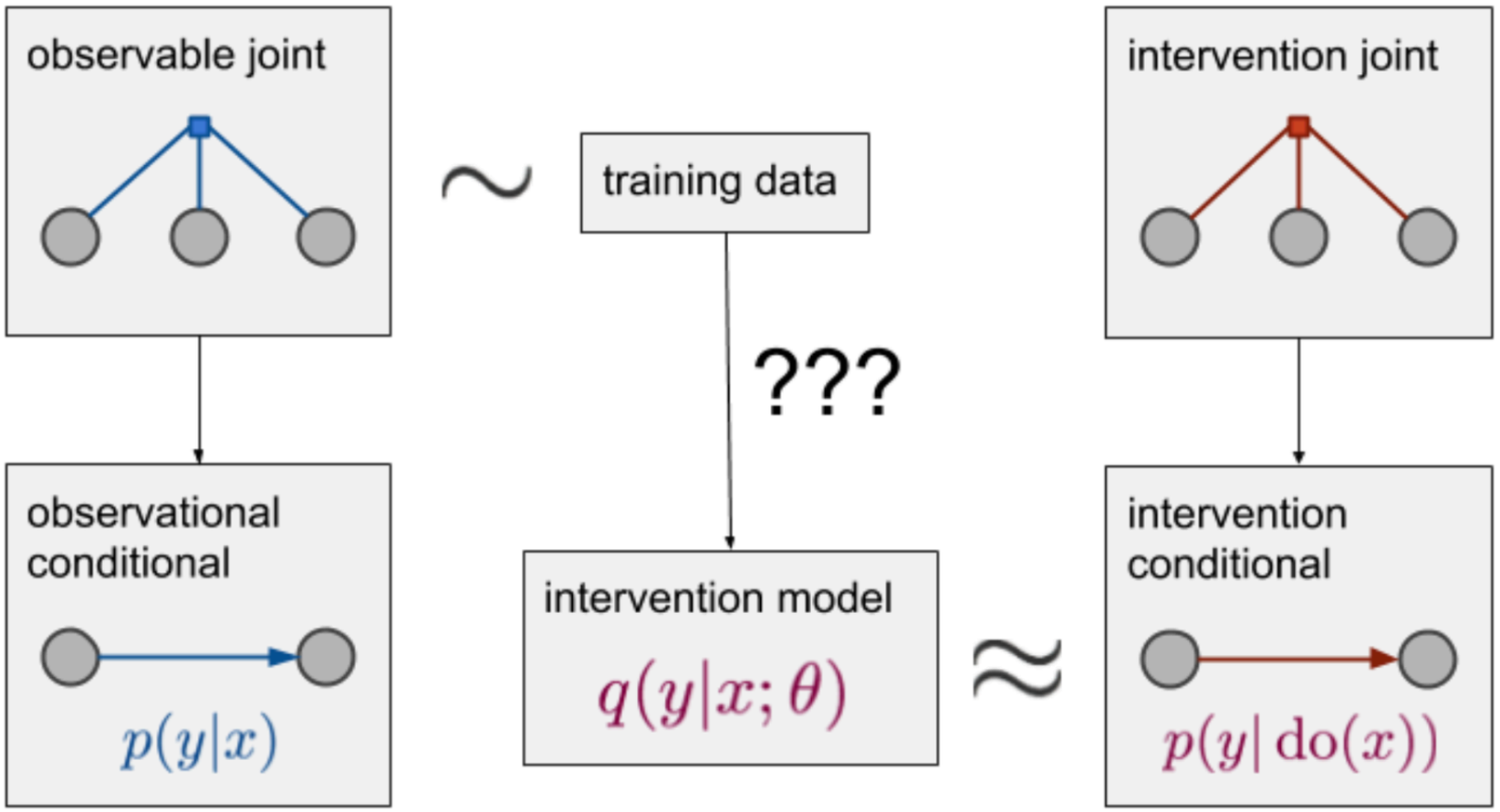
How are all these things related?

Let's start with a diagram that shows what's going on if we only care about $p(y|x)$, i.e. the simple supervised learning case:



Let's say we observe 3 variables, x, z, y, in this order. Data is sampled i.i.d. from some observable joint distribution over 3 variables, denoted by the blue factor graph labelled 'observable joint'. If you don't know what a factor graph is, it's not important, the circles represent random variables, the little square represents a joint distribution of the variables it's connected to. We are interested in predicting y from x, and say that z is a third variable which we do not want to infer but we can also measure (I included this for completeness). The observational conditional $p(y|x)$ is calculated from this joint via simple conditioning. From the training data we can build a model $q(y|x;\theta)$ to approximate this conditional, for example using a deep net minimizing cross-entropy or whatever.

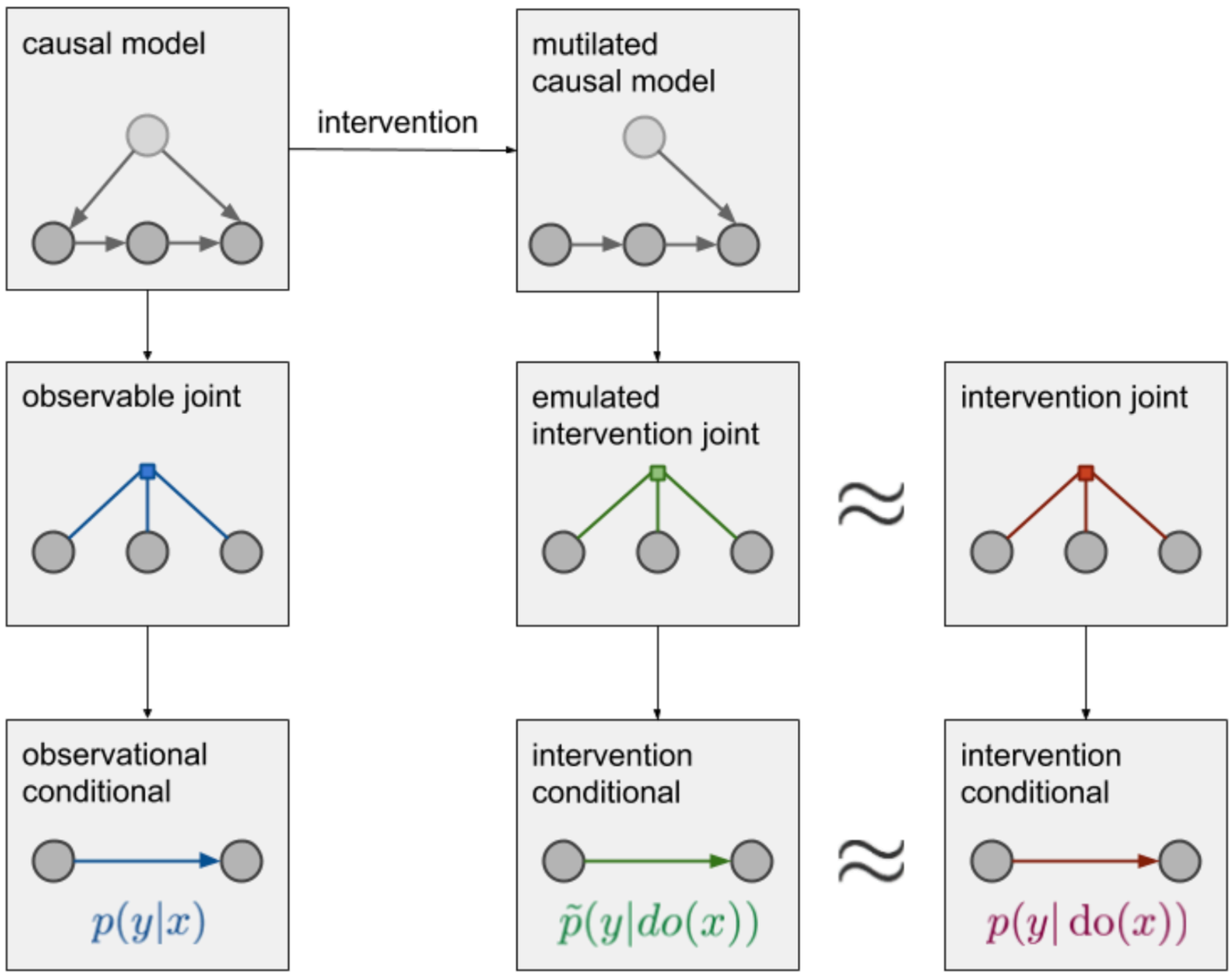
Now, what if we're actually interested in $p(y|do(x))$ rather than $p(y|x)$? This is what it looks like:



So, we still have the blue observed joint and data is still sampled from this joint. However, the object we wish to estimate is on the bottom right, the red intervention conditional $p(y|do(x))$. This is related to the intervention joint which is denoted by the red factor graph above it. It's a joint distribution over the same domain as p but it's a different distribution. If we could sample from this red distribution (e.g. actually run a randomized controlled trial where we get to pick x), the problem would be solved by simple supervised learning. We could generate data from the red joint, and estimate a model directly from there. However, we assume this is not possible, and all we have is data sampled from the blue joint. We have to see if we can somehow estimate the red conditional $p(y|do(x))$ from the blue joint.

Causal models

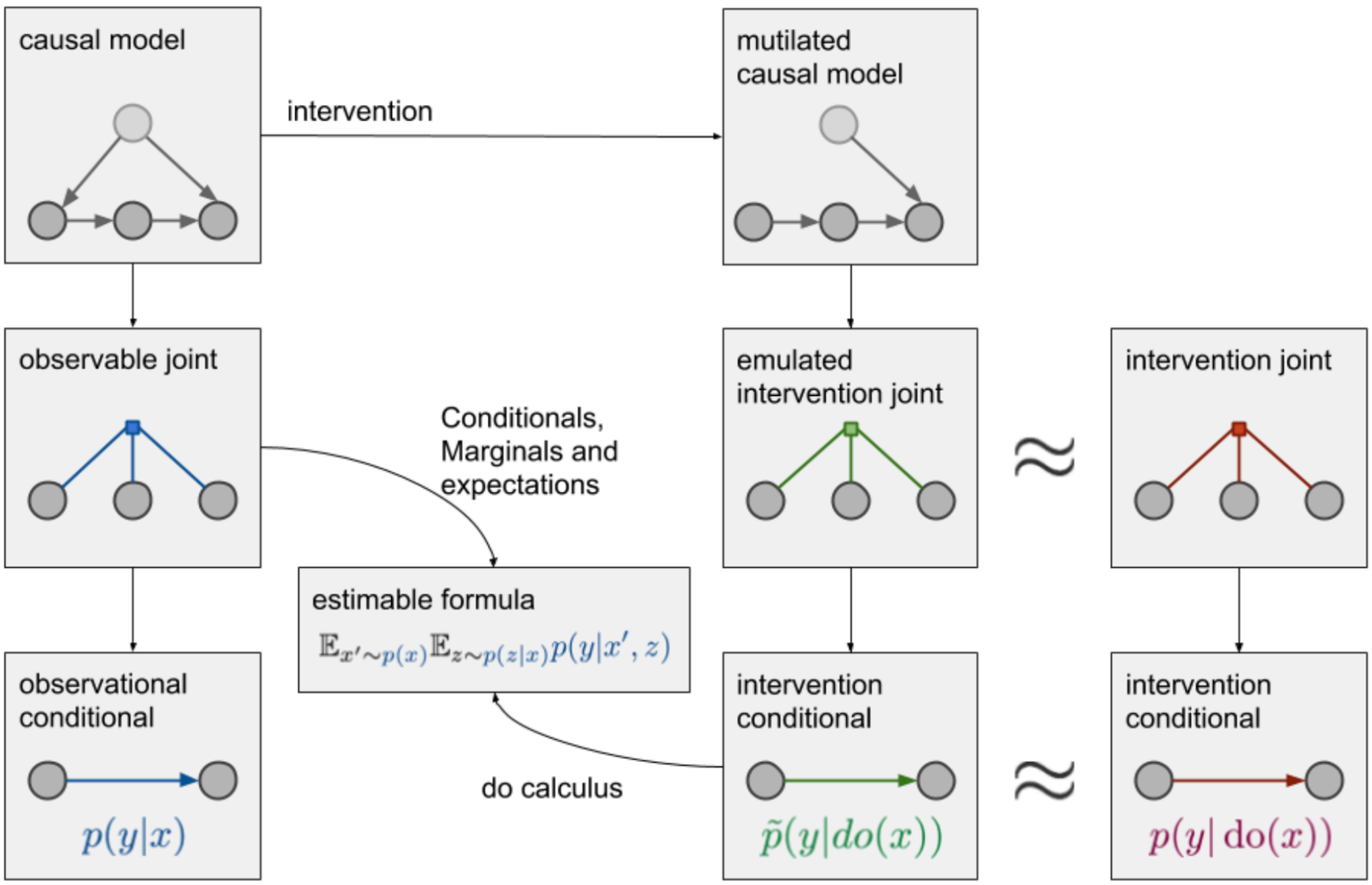
If we want to establish a connection between the blue and the red joints, *we must* introduce additional assumptions about the causal structure of the data generating mechanism. The only way we can make predictions about how our distribution changes as a consequence of an interaction is if we know how the variables are causally related. This information about causal relationships is not captured in the joint distribution alone. We have to introduce something more expressive than that. Here is how what this looks like:



Do-calculus

Now the question is, how can we say anything about the green conditional when we only have data from the blue distribution. We are in a better situation than before as 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 green conditional distribution until we can express it in terms of various marginals, conditionals and expectations under the blue distribution. Do-calculus extends our toolkit of working with conditional probability distributions with four additional rules we can apply to conditional distributions with the do operators in them. These rules take into account properties of the causal diagram. The details can't be compressed into a single blog post, but here is **an introductory paper on them.**.

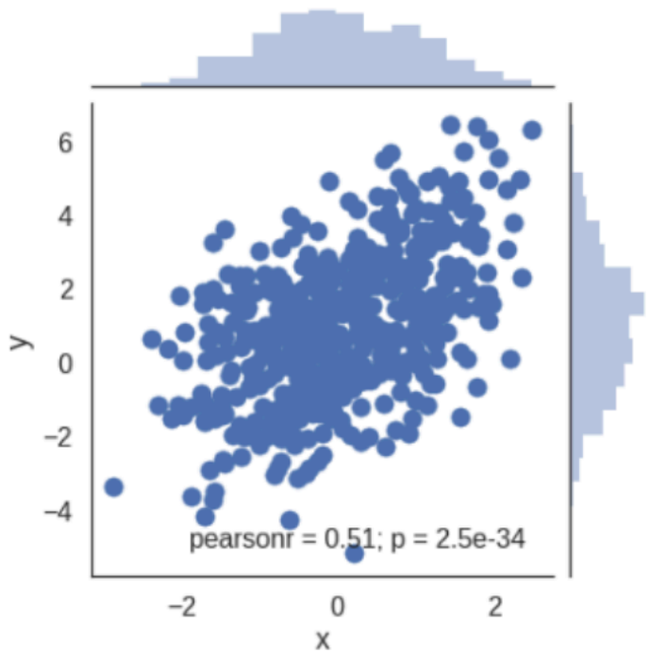
Ideally, as a result of a do-calculus derivation you end up with an equivalent formula for $\tilde{p}(y|do(x))$ which no longer has any do operators in them, so you estimate it from observational data alone. If this is the case we say that the causal query $\tilde{p}(y|do(x))$ is *identifiable*. Conversely, if this is not possible, no matter how hard we try applying do-calculus, we call the causal query *non-identifiable*, which means that we won't be able to estimate it from the data we have. The diagram below summarizes this causal inference machinery in its full glory.



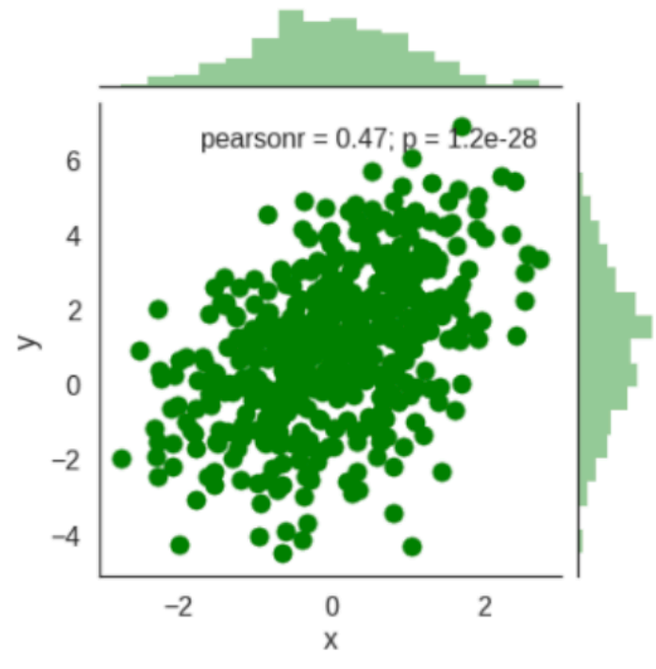
The new panel called "estimable formula" shows the equivalent expression for $\tilde{p}(y|do(x))$ obtained as a result of the derivation including several do-calculus rules. Notice how the variable z which is completely irrelevant if you only care about $p(y|x)$ is now needed to perform causal inference. If we can't observe z we can still do supervised learning, but we won't be able to answer causal inference queries $p(y|do(x))$.

Causal Inference 2: Illustrating Interventions via a Toy Example

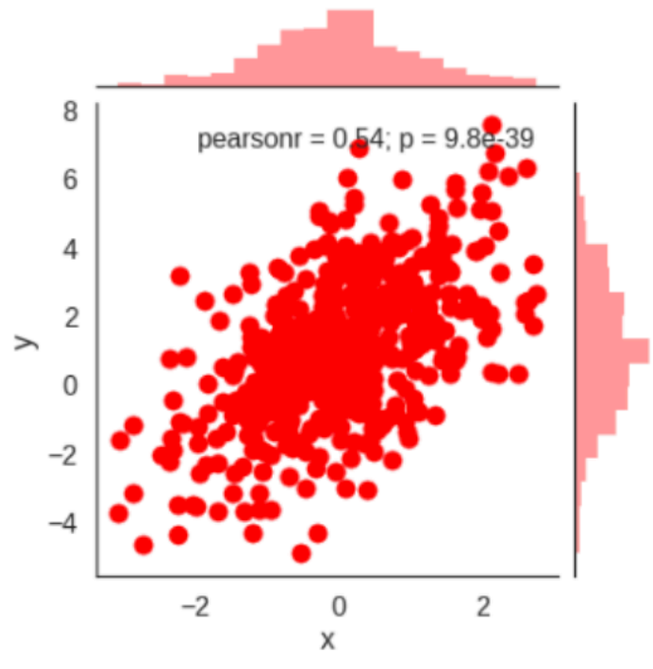
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```




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




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



```
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
```

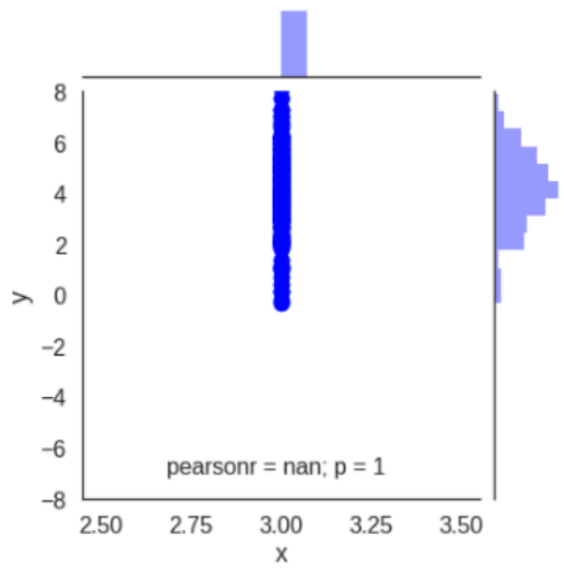
```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```

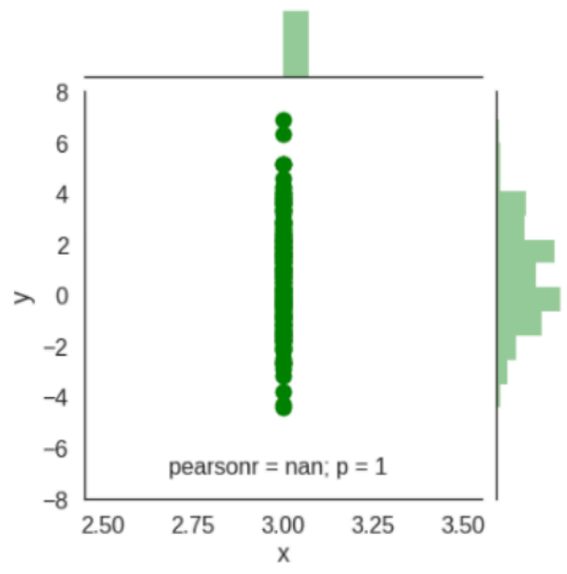
We can now run the scripts in this hacked interpreter and see how the intervention changes the distribution of x and y:

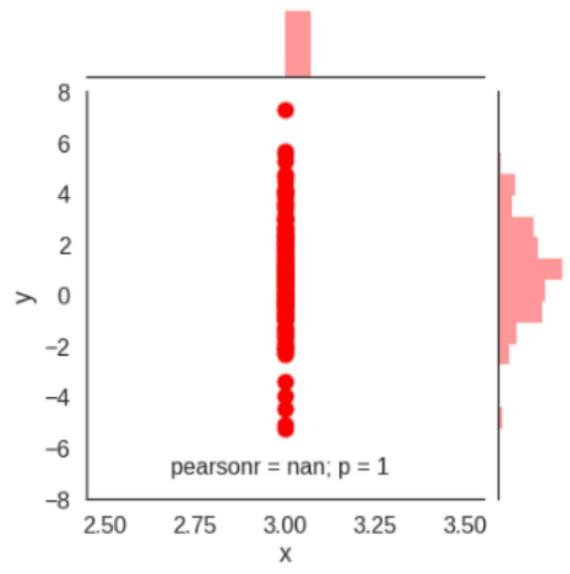
```
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3
```

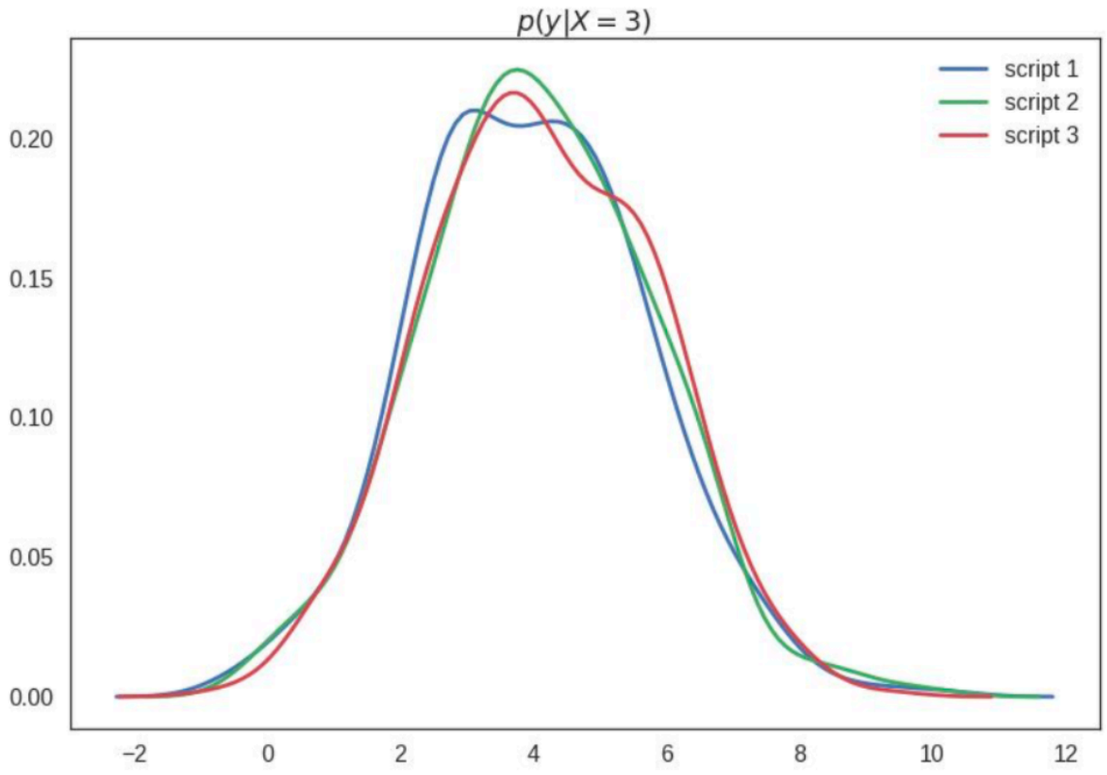
```
y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3
```

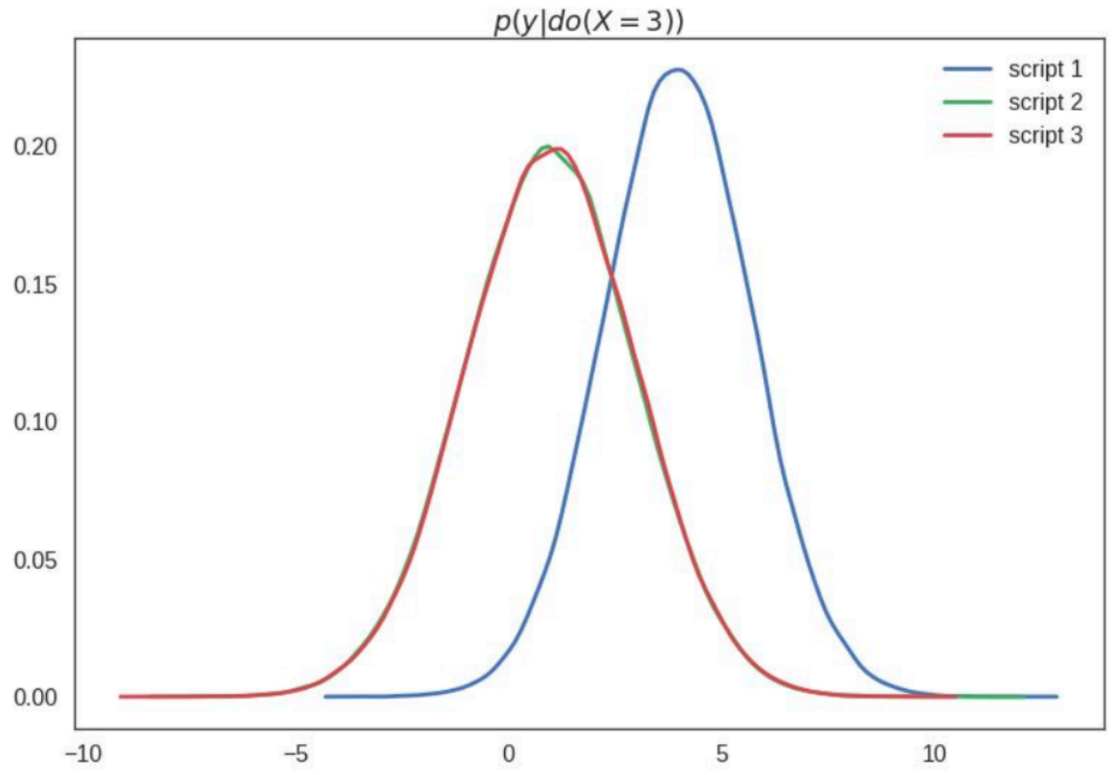
```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```







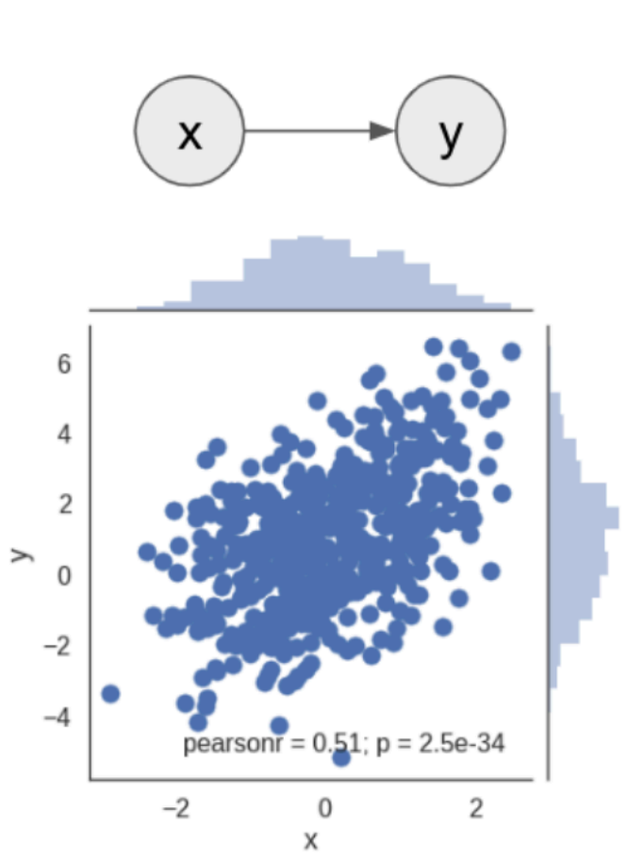

$p(y|X = 3)$
script 1
script 2
script 3
0.20
0.15
0.10
0.05
0.00
−2
0
2
4
6
8
10
12

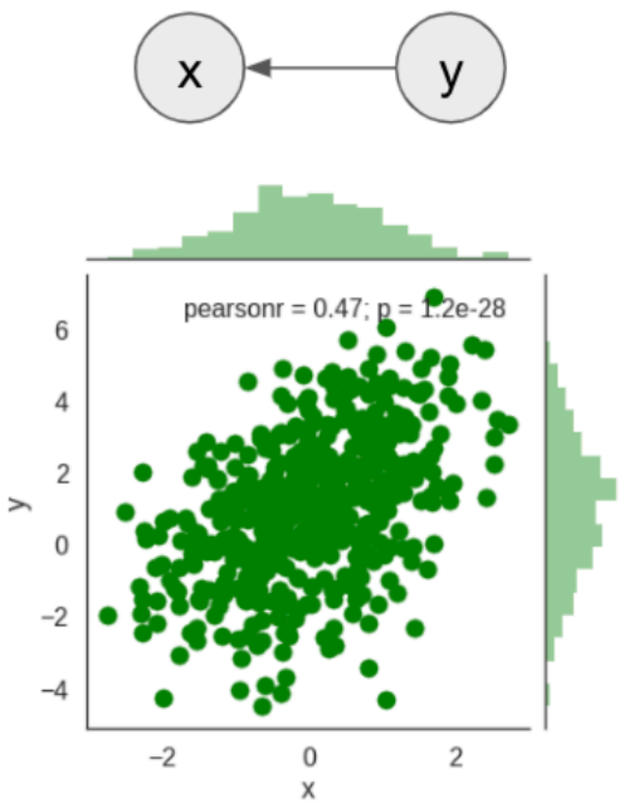

$p(y|do(X = 3))$
script 1
script 2
script 3
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
10

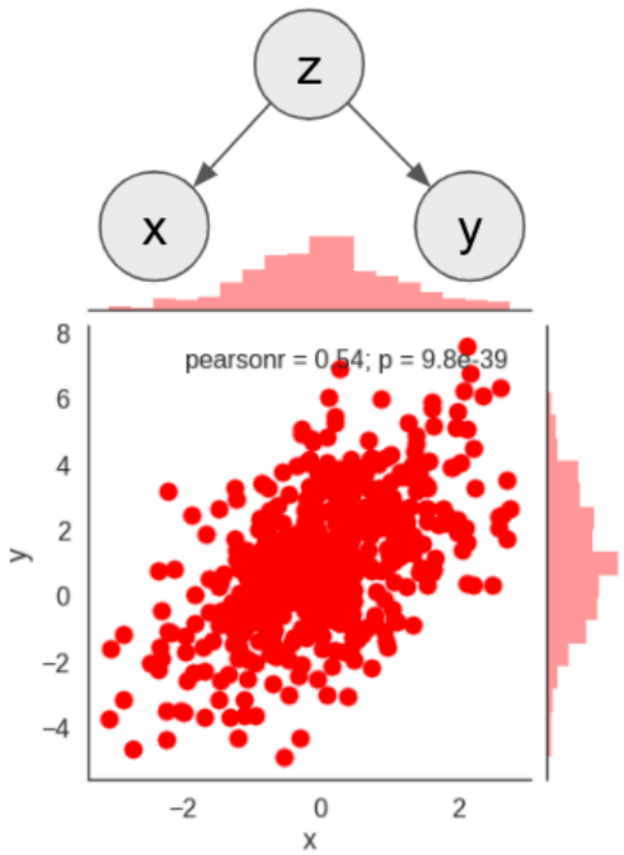
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

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

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

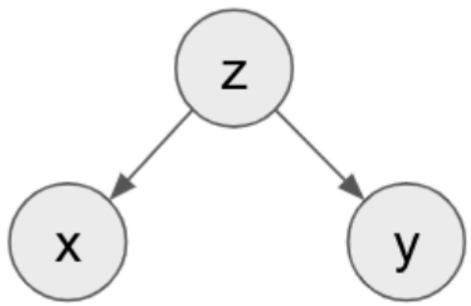






$$P(y|do(X)) = p(y|x)$$

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



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

Counterfactuals

Let me first point out that *counterfactual* is one of those overloaded words. You can use it, like Judea Pearl, to talk about a very specific definition of counterfactuals: a probablilistic answer to a "what would have happened if" question (I will give concrete examples below). Others use the terms like *counterfactual machine learning* or *counterfactual reasoning* more liberally to refer to broad sets of techniques that have anything to do with causal analysis. In this post, I am going to focus on the narrow Pearlian definition of counterfactuals. As promised, I will start with a few examples:

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

conditionied on four sets of things:

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

Given that Alice did not get promoted in her job, and given that she is a woman, and given everything else we can observe about her circumstances and performance, what is the probability of her getting a promotion if she was a man instead?

Again, the main reason for asking this question is to establish to what degree being a woman is directly responsible for the observed outcome. Note that this is an individual notion of fairness, unlike the aggregate assessment of whether the promotion process is fair or unfair statistically speaking. It may be that the promotion system is pretty fair overall, but in the particular case of Alice unfair discrimination took place.

> A counterfactual question is about a specific datapoint, in this case Alice.

Another weird thing to note about this counterfactual is that the intervention (Alice's gender magically changing to male) is not something we could ever implement or experiment with in practice.

Example 3: My beard and my PhD

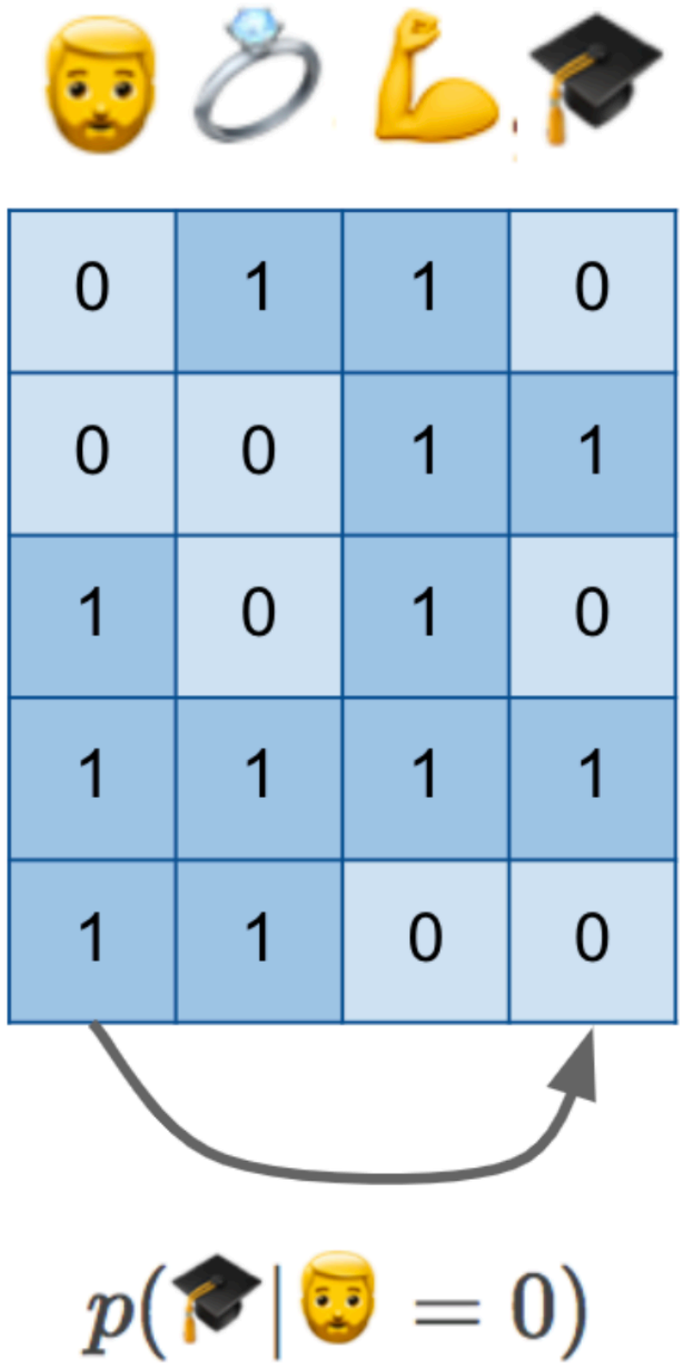
Here's the example I used in my talk, and will use throughout this post: I want to understand to what degree having a beard contributed to getting a PhD:

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.

Before I start describing how to express this as a probability, let's first think about what we intuitively expect the answer to be? In the grand scheme of things, my beard probably was not a major contributing factor to getting a PhD. I would have pursued PhD studies, and probably completed my degree, even if something would have prevented me to keep my beard. So

> We expect the answer to this counterfactual to be a high probability, something close to 1.

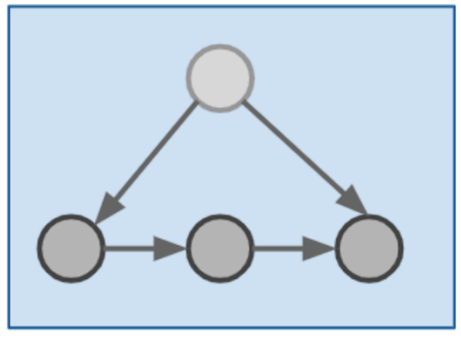
Let's start with the simplest thing one can do to attempt to answer my counterfactual question: collect some data about individuals, whether they have beards, whether they have PhDs, whether they are married, whether they are fit, etc. Here's a cartoon illustration of such dataset:



Intervention queries

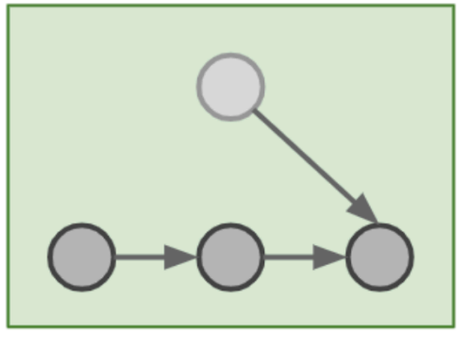
We've learned in the previous two posts that if we want to reason about interventions, we have to express a different conditional distribution, $p(🎓|do(👨 = 0))$. We also know that in order to reason about this distribution, we need more than just a dataset, we also need a causal diagram. Let's add a few things to our figure:

observed, factual



0	1	1	0
0	0	1	1
1	0	1	0
1	1	1	1
1	1	0	0

imagined, counterfactual



0	1	1	0
0	0	1	1
0	0	1	0
0	1	0	1
0	1	0	0

$p(🎓|do(👨 = 0))$

The causal diagram lets us reason about the distribution of data in an alternative world, a parallel universe if you like, in which everyone is somehow magically prevented to grow a beard. You can imagine sampling a dataset from this distribution, shown in the green table. We can measure the association between PhD degrees and beards in this green distribution, which is precisely what $p(🎓|do(👨 = 0))$ means. As shown by the arrow below the tables, $p(🎓|do(👨 = 0))$ is about predicting columns of the green dataset from other columns of the green dataset.

Can $p(🎓|do(👨 = 0))$ express the counterfactual probability we seek? Well, remember that we expected that I would have obtained a PhD degree with a high probability even without a beard. However, $p(🎓|do(👨 = 0))$ talks about a the PhD of a random individual after a no-beard intervention. If you take a random person off the street, shave their beard if they have one, it is not very likely that your intervention will cause them to get a PhD with a high probability. Not to mention that your intervention has no effect on most women and men without beards. We intuitively expect $p(🎓|do(👨 = 0))$ to be close to the base-rate of PhD degrees $p(🎓)$, which is apparently **somewhere around 1-3%**.

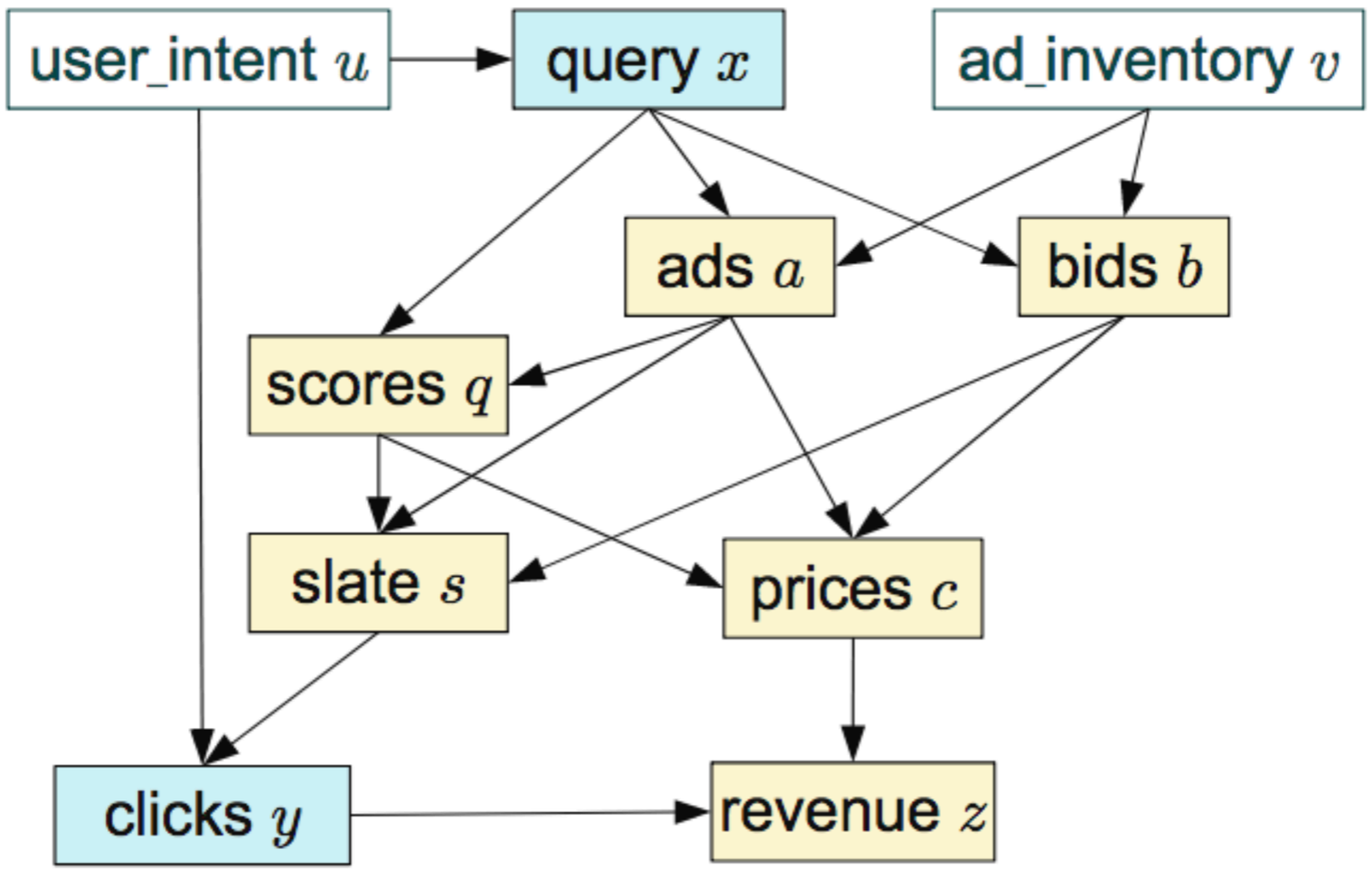
> $p(🎓|do(👨 = 0))$ talks about a randomly sampled individual, while a counterfactual talks about a specific individual

Counterfactuals are "personalized" in the sense that you'd expect the answer to change if you substitute a different person in there. My father has a mustache, (let's classify that as a type of beard for pedagogical purposes), but he does not have a PhD degree. I expect that preventing him to grow a mustache would not have made him any more likely to obtain a PhD. So his counterfactual probability would be a probability close to 0.

Structural Equation Models

A causal graph encodes which variables have a direct causal effect on any given node - we call these causal parents of the node. A structural equation model goes one step further to specify this dependence more explicitly: for each variable it has a function which describes the precise relationship between the value of each node the value of its causal parents.

It's easiest to illustrate this through an example: here's a causal graph of an online advertising system, taken from the excellent paper of **Bottou et al, (2013)**:

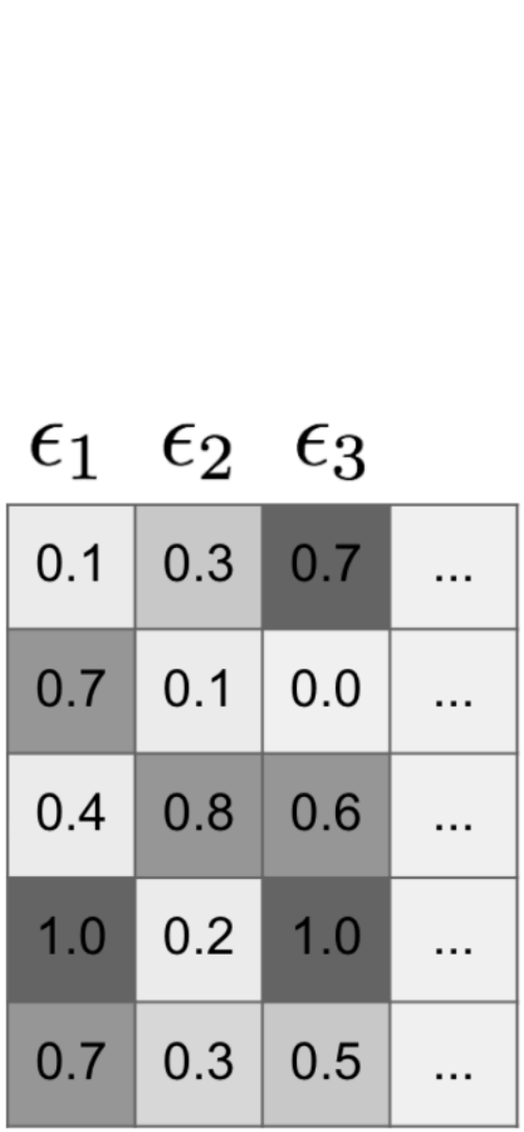


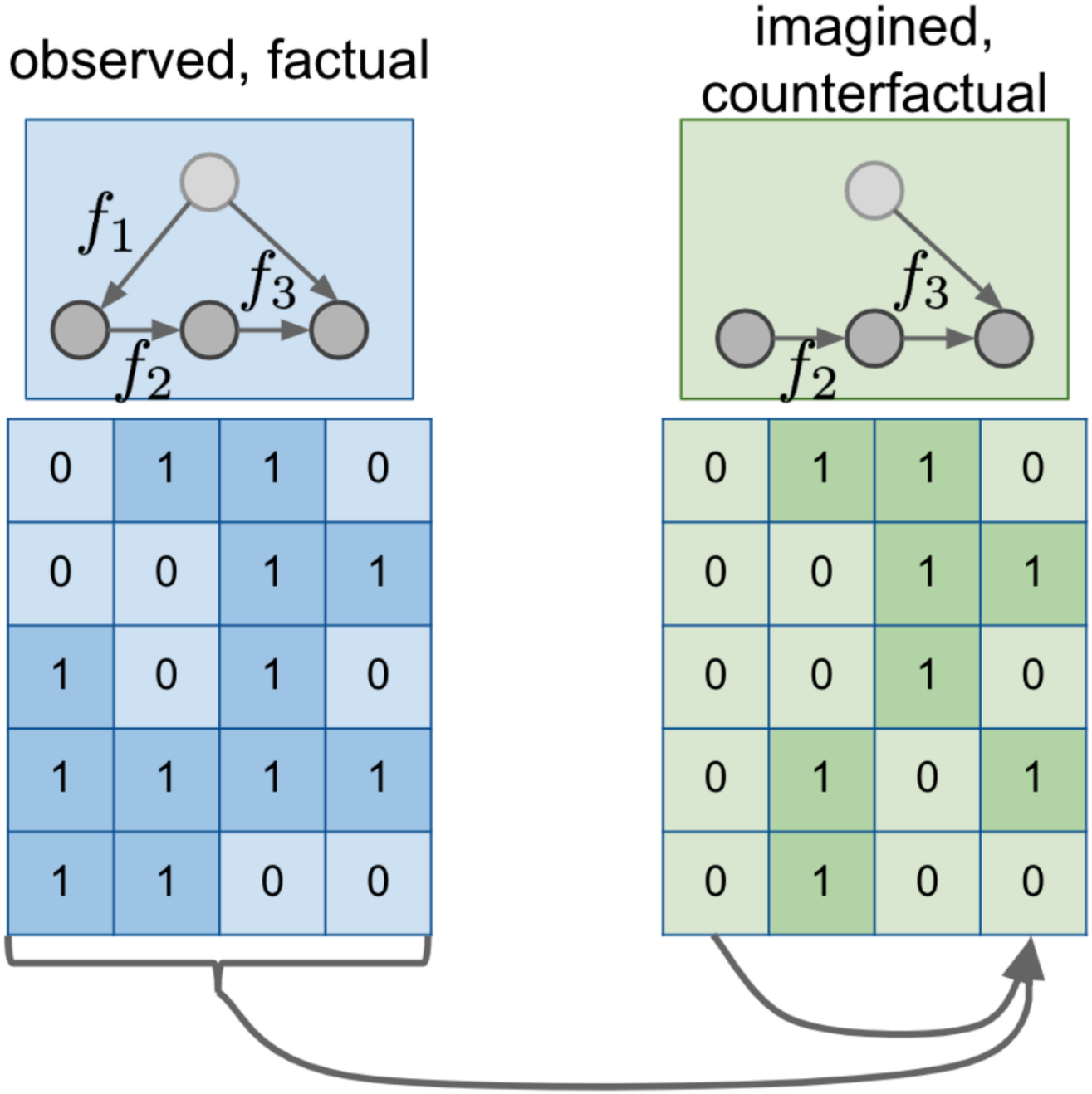
It doesn't really matter what these variables mean, if interested, read the paper. The dependencies shown by the diagram are equivalently encoded by the following set of equations:

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

Back to the beard example

Now that we know what SEMs are we can return to our example of beards and degrees. Let's add a few more things to the figure:





First change is, that instead of just a causal graph, I now assume that we model the world by a fully specified structural equation model. I signify this lazily by labelling the causal graph with the functions f_1, f_2, f_3 over the graph. Notice that the SEM of the green situation is the same as the SEM in the blue case, except that I deleted f_1 and replaced it with a constant assignment. But f_2 and f_3 are the same between the blue and the green models.

Secondly, I make the existence of the ϵ_i noise variables explicit, and show their values (it's all made up of course) in the gray table. If you feed the first row of epsilons to the blue structural equation model, you get the first blue datapoint 0110. If you feed the same epsilons to the green SEM, you get the first green datapoint (0110). If you feed the second row of epsilons to the models, you get the second rows in the blue and green tables, and so on...

Now that we established the *twin datapoint* metaphor, we can say that counterfactuals are

> making a prediction about features of the unobserved twin datapoint based on features of the observed datapoint.

Crucially, this was possible because we used the same ϵs in both the blue and the green SEM. This induces a joint distribution between variables in the observable regime, and variables in the unobserved, counterfactual regime. Columns of the green table are no longer independent of columns of the blue table. You can start predicting values in the green table using values in the blue table, as illustrated by the arrows below them.

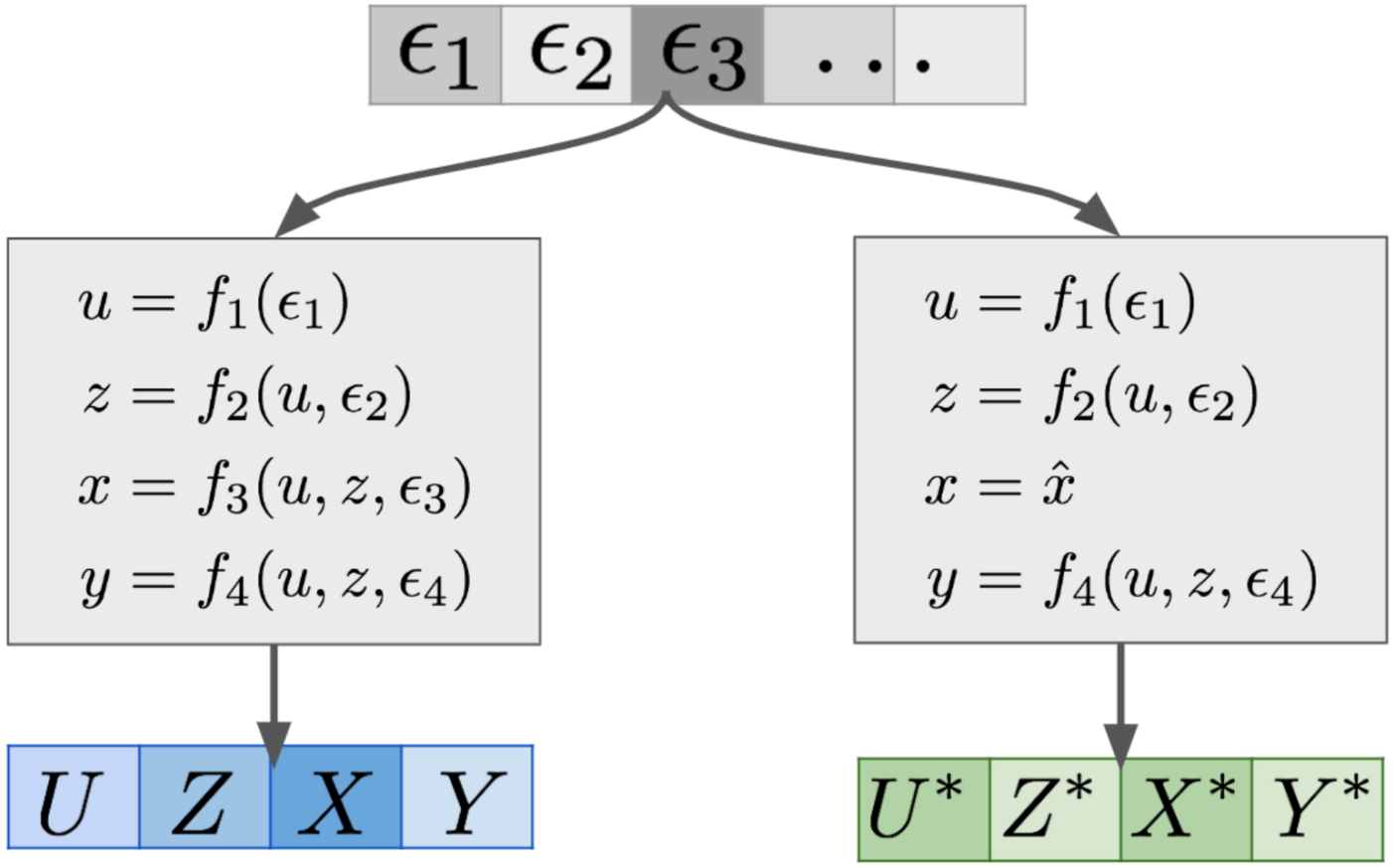
Mathematically, a counterfactual is the following conditional probability:

$$p(🎓^* | 👱^* = 0, 👱 = 1, 💍 = 1, 💪 = 1, 🎓 = 1),$$

where variables with an * are unobserved (and unobservable) variables that live in the counterfactual world, while variables without * are observable.

Looking at the data, it turns out that mirror-Ferenc, who does not have a beard, is married and has PhD, but is not quite as strong as observable Ferenc.

Here is another drawing that some of you might find more appealing, especially those who are are into GANs, VAEs and similar generative models:



A SEM is essentially a generative model of data, which uses some noise variables $\epsilon_1, \epsilon_2, \ldots$ and turns them into observations (U, Z, X, Y) in this example. This is shown in the left-hand branch of the graph above. Now if you want to make counterfactual statements under the intervention $X = \hat{x}$, you can construct a *mutilated* SEM, which is the same SEM except with f_3 deleted and replaced with the constant assignment $x = \hat{x}$. This modified SEM is shown in the right-hand branch. If you feed the ϵs into the mutilated SEM, you get another set of variables (U^*, Z^*, X^*, Y^*), shown in green. These are the features of the twin as it were. This joint generative model over (U, Z, X, Y) and (U^*, Z^*, X^*, Y^*) defines a joint distribution over the combined set of variables $(U, Z, X, Y, U^*, Z^*, X^*, Y^*)$. Therefore, now you can calculate all sorts of conditionals and marginals of this joint.

Of particular interest are these conditionals:

$$p(y^*|X^* = \hat{x}, X = x, Y = y, U = u, Z = z),$$

which is a counterfactual prediction. In reality, since $X^* = \hat{x}$ holds with a probability of 1, we can drop that conditioning.

My notation here is a bit sloppy, there are a lot of things going on implicitly under the hood, which I'm not making explicit in the notation. I'm sorry if it causes any irritation to people, I want to avoid overcomplicating things at this point. Now is a good time to point out that Pearl's notation, including do-notation is often criticized, but people use it because now it's widely adopted.

We can also express the intervention conditional $p(y|do(x))$ using this (somewhat sloppy) notation as:

$$p(y|do(X = \hat{x})) = p(y^*|X^* = \hat{x})$$

We can see that the intervention conditional only contains variables with an * so it does not require the joint distribution of $(U, Z, X, Y, U^*, Z^*, X^*, Y^*)$ only the marginal of the * variables (X^*, Z^*, X^*, Y^*). As a consequence in order to talk about $p(y|do(X = \hat{x}))$ we did not need to introduce SEMs or talk about the epsilons.

Furthermore, notice the following equality:

$$\begin{aligned} p(y|do(X = \hat{x})) &= p(y^*|X^* = \hat{x}) \\ &= \int_{x,y,u,z} p(y^*|X^* = \hat{x}, X = x, Y = y, U = u, Z = z)p(x, y, u, z)dxdydudz \\ &= \mathbb{E}_{p_{X,Y,U,Z}} p(y^*|X^* = \hat{x}, X = x, Y = y, U = u, Z = z), \end{aligned}$$

in other words, the intervention conditional $p(y|do(X = \hat{x}))$ is the average of counterfactuals over the obserevable population. This was something that I did did not realize before my MLSS tutorial, and it was pointed out to me by a student in the form of a question. In hindsight, of course this is true!