Joris M. Mooij
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
j.m.mooij@uva.nl

Inaugural Lecture

Causality: from data

On Thursday 13 October 2022, Joris Mooij delivered his inaugural lecture upon acceptance of the position of professor in Mathematical Statistics at the University of Amsterdam. In many scientific and societal issues the relationship between causes and their effects is an important topic. Mooij explains why both causal models and experimental research remain necessary to answer causal questions (despite recent breakthroughs in machine learning).

In this lecture, I will take you on a tour through what I consider one of the most fascinating scientific disciplines: causality. We will consider a diverse range of scientific questions from a variety of fields. For example:

- *Does smoking cause lung cancer?* There is widespread agreement nowadays that it does, but this question was the topic of a huge debate in the sixties of the last century, involving famous statisticians like Ronald Fisher and Jerzy Neyman.
- *Does chocolate consumption increase cognitive abilities?* In other words: do you get smarter, if you eat lots of chocolate? This (perhaps more innocent) question is still the subject of scientific debate as of today, and the available evidence seems inconclusive.
- *Does the new COVID-19 vaccine protect better against hospitalization?* Pharmaceutical companies have updated their vaccines to protect against the new variants of the SARS-CoV-2 virus. To decide which vaccine to use, it is important to know whether someone who got a booster with the new vaccine has a lower probability to end up in hospital because of COVID-19 compared to someone who got a booster with the old vaccine instead.
- *Does knocking out gene X change the activity of gene Y?* A cell of a typical living organism has thousands of genes, which control the bio-chemical 'machinery' of the cell. Not all genes are active at the same time, and the activity of a gene is regulated by the activity of other genes. If a researcher disables one specific gene in a so-called 'knock-out' experiment, this may result in a change in the activity of one or multiple other genes. Understanding these *gene regulatory mechanisms* is one of the quests in biology.

You might wonder: what do all these questions have in common? The answer is that they are all of the form: *what will happen to Y if action X is performed?*

Similarly, many policy decisions and questions of societal relevance are of a causal nature. For example:

- *How will the revenue of a company change if it increases the price of a product?* A higher price typically means that less products will be sold. Does this decrease in volume compensate for the increased pricing?
- *How much will inflation in the Netherlands decrease if the European Central Bank increases the interest rate by 1 percent point?* An important question, but hard to answer reliably, because of the complex nature of our macroeconomy.
- *Do female graduate students applying for college have lower admission chances than male graduate students?* This appeared to be the case at UC Berkeley (California) in 1973. However, it turned out to be a statistical paradox. Upon closer investigation of the available data by statisticians, it appeared that there was no evidence for gender bias.
- *Would changing gender increase the chance of graduating cum laude for female PhD candidates?* This is a similar type of question, but closer to home, and still relevant today. The Dutch newspaper *NRC* discovered in 2018 that at many Dutch universities, the fraction of male PhD candidates that graduate *cum laude* ('with distinction') is about twice as high as that of female PhD candidates. Is this gender bias, or a naïve (and possibly incorrect) interpretation of the data, like in the UC Berkeley case?

Again, note that all these questions are of the same general form. They all concern to what extent a cause X influences its (potential) effect Y.

Joris Mooij

to science

Historical remarks

For many centuries, humans have been trying to understand the universe by thinking in terms of causes and effects. The importance of this way of thinking can hardly be overestimated. Let me take you on a brief excursion through history to show how various philosophers and scientists thought about causality. While Judea Pearl goes all the way back to Adam and Eve in his recent *Book of Why* [16], I will start with one of the ancient Greek philosophers, Democritus (ca. 460–370 b.C.). Democritus is mostly known for his formulation of an atomic theory of the universe. He clearly appreciated the importance of causality when he wrote:

> "I would rather discover one true cause than gain the kingdom of Persia."

But what does it mean to "discover one true cause"? Philosophers, amongst them David Hume (1711–1776), have been struggling with this question. Hume wrote [7]:

> "Thus we remember to have seen that species of object we call *flame*, and to have felt that species of sensation we call *heat*. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one *cause* and the other *effect*, and infer the existence of the one from that of the other."

While this definition of causality contains many appropriate elements, one could criticize it as being overly simplistic. For example, does the rooster's crow really cause the sun to rise? And does the barometer needle really cause rain? Furthermore, is the "constant conjunction in all past instances" really necessary? If we say "smoking causes lung cancer", we do not necessarily mean that *everyone* who smokes will get lung cancer.

Difficulties like these have led some to throw the towel in the ring. One of them was Bertrand Russell (1872–1970), a famous logician and philosopher, and one of the authors of the *Principia Mathematica* [19] (an attempt to derive all of mathematics by pure logic from a small set of axioms). Russell proposed to abandon the concept of causality completely. He wrote [17]:

> "All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word 'cause' never occurs. The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm."

The difficulties of formally defining the notion of causality also led Karl Pearson (1857–1936), one of the founders of statistics — still well-known today from the *correlation coefficient* named after him — to propose to get rid of the notion. Pearson wrote [14]:

> "Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of even modern science, namely, the category of cause and effect."

You may have heard the slogan 'correlation is not causation'. Pearson's point of view was that one should *only* consider correlation. It is striking that even today, I teach BSc mathematics students how to calculate Pearson's correlation coefficient in an introductory statistics course, but I do not teach them anything about causality. And this is not just me: this appears to be typical for most statistics courses taught at most universities in most countries. I am convinced that this is a missed opportunity!

In the last decades, though, things have changed quite dramatically. If we fastforward to today, we see that causality is a thriving and growing scientific discipline. It is scattered across fields, and has seen important contributions from epidemiology, econometrics, genetics, machine learning, statistics, artificial intelligence, computer science, and more.[1] For example, the consultancy company Gartner recently included Causal AI in its *Hype Cycle for Emerging Technologies* [5], because they expect that it will deliver "a high degree of competitive advantage over the next 5 to 10 years."

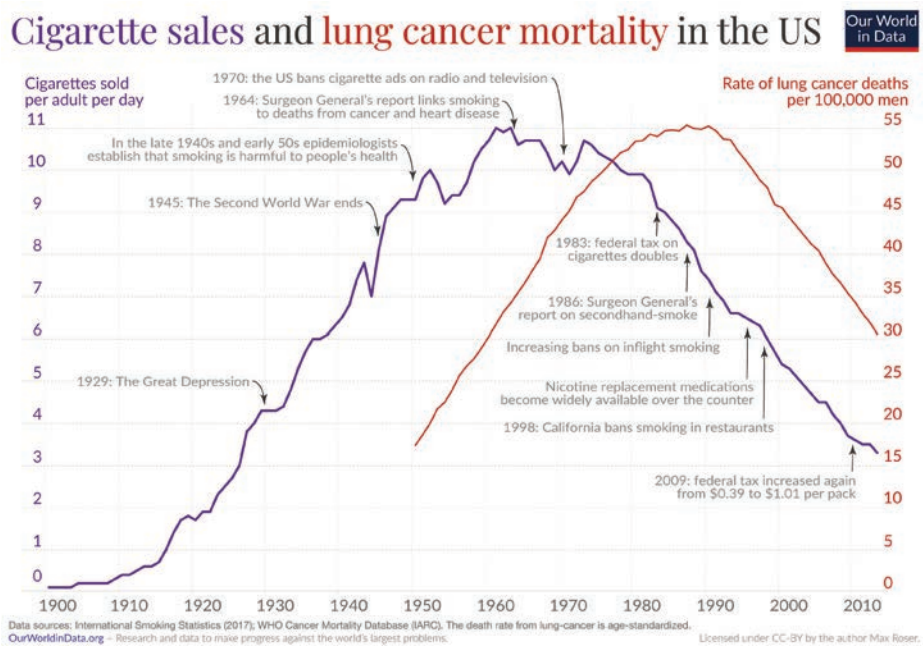


Figure 1 There is a clear (time-lagged) correlation between cigarette sales and lung cancer mortality in the US.[2]

Causation or correlation?

Let us revisit the causal question "Does smoking cause lung cancer?" One might wonder: what evidence exists that smoking causes lung cancer? I illustrated one piece of evidence in Figure 1: if we plot the number of cigarettes sold per adult per day in the United States (in purple), and the rate of lung cancer deaths for men in the US (in red), both over a period of several decades, we observe a striking similarity in the shapes of these curves (although the lung cancer deaths occur with a delay of about 25 years).

However, following in the footsteps of for example Ronald Fisher, you may argue that this is not a very convincing proof. Indeed, this could also be an example of a so-called *spurious* correlation, that is, a correlation without causation. Two examples of such spurious correlations are shown in Figure 2. If we plot US spending on science, space and technology over the years, and the number of suicides by hanging, strangulation and suffocation, we observe a striking similarity between the two curves. It appears quite implausible, though, that there is any causal relation between the two. The divorce rate in Maine (a state in the US) over the years also shows a strikingly similar pattern as the per capita consumption of margarine. Would one take this as evidence for a causal relation between the two?

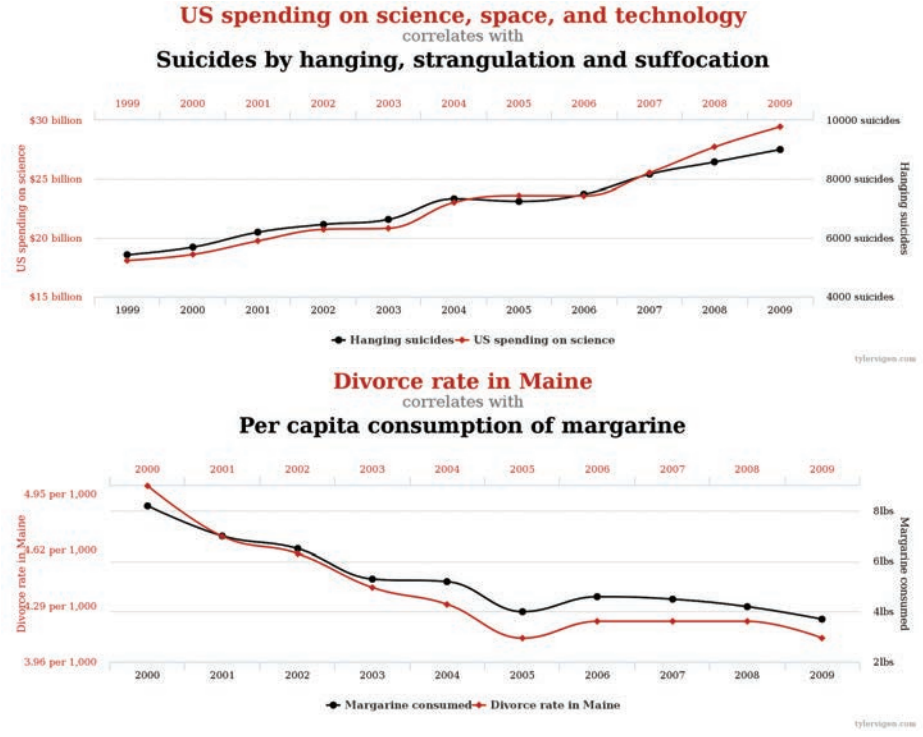


Figure 2 Two examples of correlations between time-series that appear to be spurious.[3]

I took these examples from a website, where you can find many more.[4] The surprising nature of these examples is due to selection bias: the website was created with the help of an algorithm that searches for short pieces of highly correlated time-series in a large database containing many different time-series. Because of the multiple-testing issue, these strong correlations may actually not be statistically significant (indeed, even if you search for such patterns in random data, you will eventually find them).

So, if causation is not correlation, then what is it? Giving a precise definition is not straightforward. It is perhaps as challenging as defining other elementary notions like 'space' and 'time'. I provide here a simplified definition that contains the basic gist (note, though, that it is still so vague that no mathematician or statistician would be satisfied with it!).

First, consider *deterministic* systems, that is, systems in which chance plays no role. Suppose that variables X and Y describe part of the system's state.

Definition 1. We say that X *causes* Y if a minimal external intervention on the system that sets the value of X may change the possible values of Y.

For example, consider a bicycle. If we take for X the position of the break lever, and for Y the rotation angle of the wheels, then X causes Y (since pulling the break lever prevents the wheels from rotating). However, rotating the wheels does not change the possible positions that the break lever can have, so Y does not cause X.

For *stochastic* systems, where chance plays a role, we just need a small change in the definition. As we noted before: not everyone who smokes will get lung cancer.

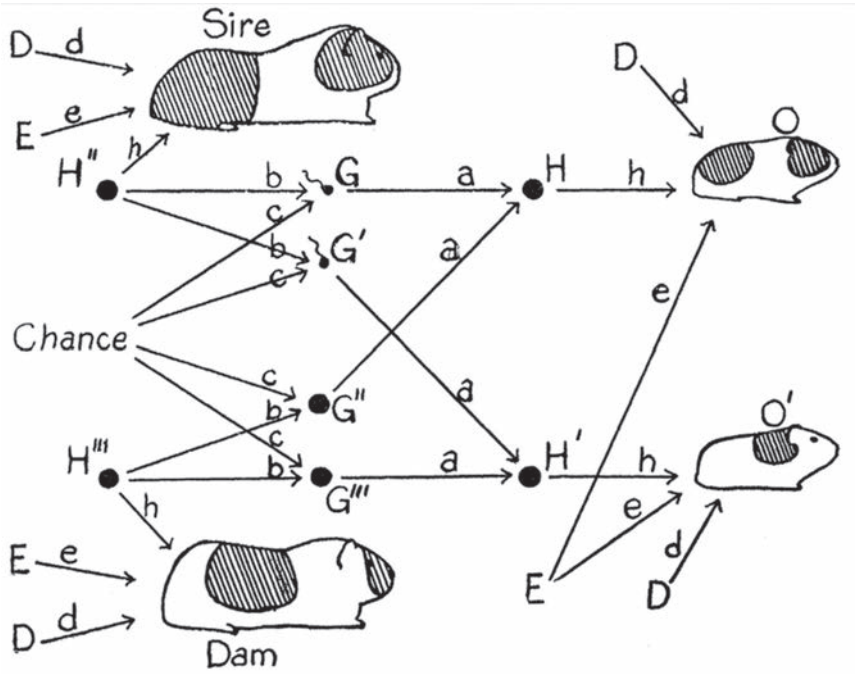


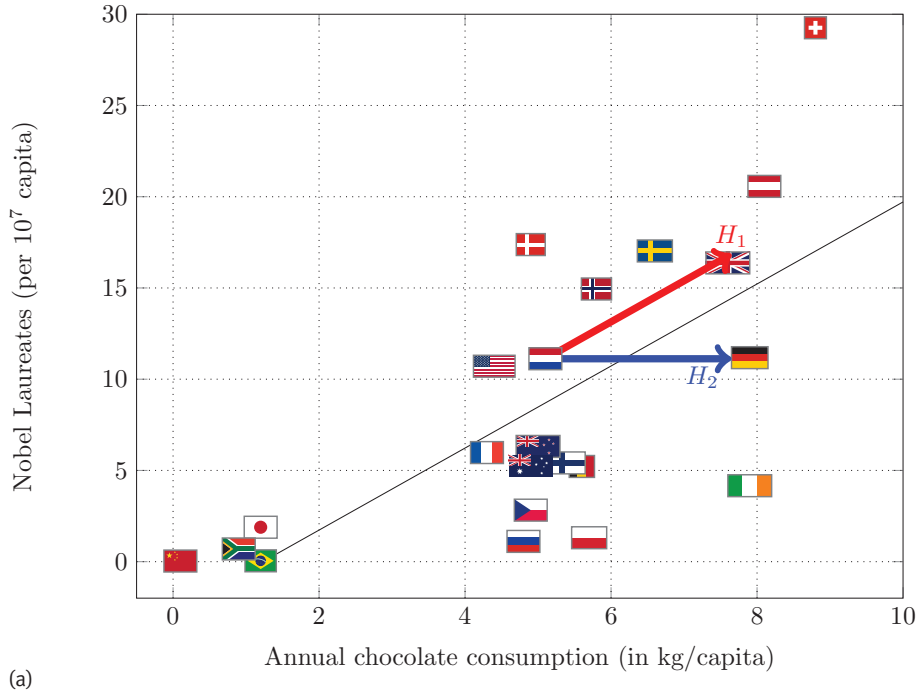
Figure 3 'Path diagram' in [20] expressing causal relations between fur patterns and various genetic and environmental factors. This is perhaps the first instance of a graphical causal model.

Therefore, we replace 'the possible values of' by 'the probability of'. This leads to the following definition for stochastic systems.

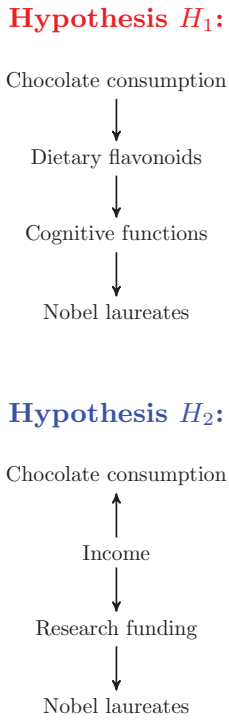
Definition 2. We say that X *causes* Y if a minimal external intervention on the system that sets the value of X may change the probability distribution of Y.

I would like to illustrate this by using the 'path diagram' in Figure 3, used by geneticist Sewall Wright in 1920 to communicate a causal model of how the fur pattern of guinea pigs is determined by various genetic and environmental factors [20]. 'Chance' is explicitly represented here, and plays a role in how the genetic constitution of the offspring depends on that of the parents. If we take for X the genetic constitution of the parent individuals, and for Y the fur patterns of their children, then X causes Y in a non-deterministic way, according to the above definition applied to Wright's model.

As a remark to the statisticians in the audience: note that this definition shows that the standard framework of classical statistics is too narrow and needs to be extended. Indeed, a statistical model describes the joint distribution of X and Y, but not how the distribution of Y changes when we intervene on X. Two such extensions have become popular for modeling stochasticity and causality. The *potential outcome framework* considers jointly defined random variables for each hypothetical intervention (so-called 'potential outcomes'). The other framework models how the probability distribution of the system's variables depends on interventions (thus essentially treating interventions as parameters of the distribution).[5] In both frameworks, graphs can be used to represent causal relations and independence relations between the variables. While there has been a heated debate on which of the two frameworks is superior, the differences are actually minor.[6]



(a)



(b)

Figure 4 (a) There is a significant correlation (with Pearson correlation coefficient $R = 0.69$, and p-value 0.0004) between chocolate consumption and Nobel laureates across countries.[7] (b) Two causal hypotheses that might explain the observed correlation between chocolate consumption and Nobel laureates. For both hypotheses, I indicated with colored arrows in (a) the expected effects of an intervention in which the Dutch government increases the chocolate consumption in the Netherlands by 50%.

Chocolate consumption and Nobel laureates

Now that we have some idea of how causation differs from correlation, let us return to the question "Does chocolate consumption increase cognitive abilities?" In a publication in the prestigious *New England Journal of Medicine*, an interesting observation was reported: the average annual chocolate consumption is significantly correlated with the number of Nobel laureates per capita [10]. The graph in Figure 4(a) visualizes the data. Chocolate consumption is on the horizontal axis, and the fraction of Nobel laureates is on the vertical axis. We indeed observe a correlation, as the data fall roughly on a straight line, with a Pearson correlation coefficient of about 0.7. In particular, we can read off that an average Dutch inhabitant eats 5 kg of chocolate a year. This is about two thirds of the average chocolate consumption of an inhabitant of the United Kingdom. What would happen if the Dutch ate 50% more chocolate? Would the number of Dutch Nobel laureates also increase by 50%? And thus, would it be a good idea for the University of Amsterdam to provide free chocolate for all staff members and students? As we shall see, the answer depends on the causal relations between these two variables.

Messerli provided the following possible explanation of the observed correlation [10], that I will refer to as hypothesis H_1. Chocolate contains certain chemical substances known as 'dietary flavonoids'. There is some evidence in animal studies that dietary flavonoids may have positive effects on brain regions involved with memory and learning. Messerli speculates that the consumption of chocolate may lead to an increased uptake of dietary flavonoids in the brain, which may improve the cognitive functions of the brain, eventually leading to more Nobel laureates. I have illustrated this hypothesis by means of a causal graph in Figure 4(b), where the arrows indicate a direct causal relationship of one variable on another, just like in the 'path diagram' of Wright that we just saw.

But we can also consider an alternative theory, hypothesis H_2: the average income in a country determines how much money the inhabitants have for buying chocolate, and also how much tax payers' money is used to fund scientific research, eventually leading to Nobel laureates. In this hypothesis, the variable 'income' is a common cause of our two variables of interest (chocolate consumption and Nobel laureates), and is called a *confounder* for that reason. This confounder may explain the observed correlation of chocolate consumption and Nobel laureates, even if there is no direct causal relation between the two.

According to these hypotheses, what would happen if the government intervened to increase the chocolate consumption in the Netherlands by 50%? I illustrated this in Figure 4(a). Under hypothesis H_1 (red arrow), chocolate consumption causes Nobel laureates, and therefore, we would expect the number of Nobel laureates to go up as well, perhaps even to the level of that of the UK. In contrast, under hypothesis H_2 (blue arrow), chocolate consumption does *not* cause Nobel laureates, hence we would *not* expect any change, and the Netherlands would probably end up somewhere near Germany.

We may conclude that predictions of the consequences of actions can depend in a very sensitive way on the underlying causal relationships between the variables. This also means that using off-the-shelve machine learning tools for supervised learning (including deep neural networks) to make such predictions may be a bad idea.

Randomized controlled trials

One way to investigate whether chocolate consumption indeed improves cognitive abilities is to setup a so-called *randomized controlled trial*. This approach to estimating causal effects is called a 'gold standard', as it is considered to be the most reliable method. In our case, it would work as follows (see also Figure 5). We first select a sample of individuals (say, a representative sample of inhabitants of the Netherlands). Then, by flipping a coin we divide the individuals into two groups, the intervention group and the control group. On all individuals in the intervention group, we enforce a diet containing large amounts of chocolate. The eating habits of the individuals in the control group are left unchanged. We sustain the diets for several years, and then measure the cognitive abilities of all individuals. Finally, we compare the outcomes in the two groups. If we see that the individuals in the intervention group have become significantly smarter on average than those in the control group, we conclude that there is a causal relationship, and can estimate the size of the causal effect.

The approach of randomized controlled trials was popularized by Ronald Fisher, and nowadays provides the pillar of 'evidence-based' medicine. Also known as 'A/B-testing', it is used extensively by big tech companies to optimize their business algorithms.

The example also points out some of the limitations of randomized controlled trials. First, there are logistic aspects. We need a sufficiently large sample size to arrive at statistically significant conclusions. If the outcome of interest is a rare event (for example, winning a Nobel price), the number of participants in the trial needs to be huge. And how exactly would we enforce this chocolate diet in practice? Another issue is ethics. This is the reason that no randomized controlled trial has yet been performed on humans to investigate whether smoking causes lung cancer: it would simply not be

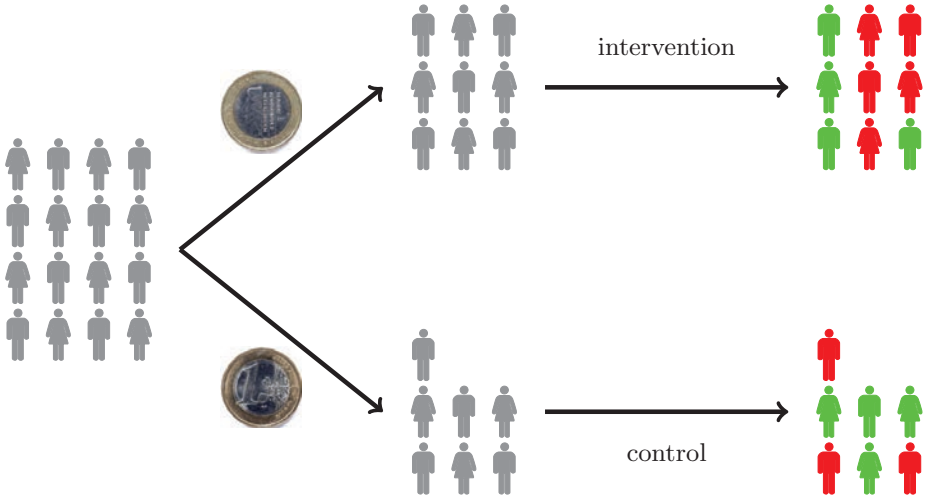


Figure 5. Schematic illustration of a randomized controlled trial.

ethical to force subjects in the treatment group to smoke for many years, given that we expect many of them to develop lung cancer as a result of the experiment. Another limitation lies in the inclusion criteria. In medical studies, subjects are often healthy young males, whereas one might be interested more specifically in the effect of treatment on elderly diseased females. To which extent can the conclusions of a randomized controlled trial be extrapolated to other subpopulations?

So, do we have alternatives? I don't believe that a randomized controlled trial has ever been performed in astronomy, for instance. Yet, astronomers rely on purely *observational* data to arrive at a causal understanding of the universe.[8] In medicine, could we perhaps make use of routinely collected electronic health records to estimate the causal effects of treatments on health outcomes?

Naïve attempts to use observational data (rather than experimental data) to infer causal effects can fail horribly. A striking illustration of this is related to a phenomenon known as *Simpson's paradox*. This paradox can be thought of as a statistical analogue to an optical illusion: we appear to see something that cannot exist.

Simpson's paradox

I will attempt to visualize Simpson's paradox for you. Consider the following hypothetical scenario. Suppose someone has collected data from electronic patient records concerning COVID-19 vaccinations. The question at stake is which of two vaccine types (A or B) is most effective at preventing hospitalization because of COVID-19. I will denote the treatment variable with X, and the outcome variable with Y. Suppose for the moment that there are two possible treatments (A and B, corresponding to the two vaccines), and two possible outcomes: positive and negative (where negative corresponds to hospitalization within six months after vaccination).

In total, the data concerns 10 000 cases of COVID-19. In Table 1 (columns '♂ + ♀') we see for instance that of the 5000 individuals that had vaccine B, 2000 ended up in hospital, but 3000 did not.[9] This means that 60% of the individuals that had vaccine B had a positive outcome. Of the individuals that had vaccine A, only 50% had a positive outcome. I visualized these

	♂ + ♀		♂		♀	
	$Y=+$	$Y=-$	$Y=+$	$Y=-$	$Y=+$	$Y=-$
$X=\mathrm{A}$	2500	2500	1500	2250	1000	250
$X=\mathrm{B}$	3000	2000	375	875	2625	1125

Table 1 Data for the fictitious example of Simpson's paradox. Treatment X (type of COVID-19 vaccine) can take values A, B. Outcome Y (hospitalization after COVID-19 infection) can take values +,−. The aggregated data can be split into gender-specific subgroups.

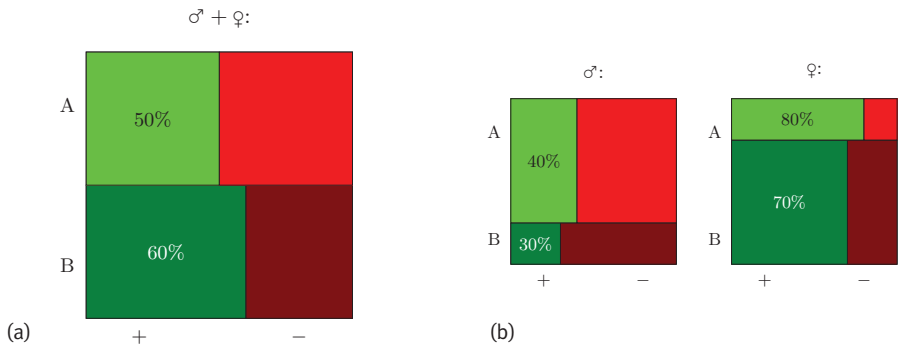


Figure 6 Visualizations of the fractions of positive outcomes corresponding to the data in Table 1; (a) aggregated over genders; (b) gender-specific. Area is proportional to counts.

fractions in Figure 6(a). Which of the two vaccines would you prefer, based on this data?

A closer look at the data reveals something remarkable. The genders of these 10 000 individuals were also recorded. For simplicity of exposition, assume that all of them are either male or female. I provided the counts for both genders separately in Table 1 (columns '♂' and '♀'). For example, of the 3000 individuals that got vaccine B and had a positive outcome, only 375 were male, and 2625 were female. You can check for yourselves that all the numbers add up. Let us look at the fractions, which I visualized in Figure 6(b). For males, those that got vaccine A had 40% probability on a positive outcome, versus only 30% for vaccine B. For females, 80% of those that got vaccine A had a positive outcome, versus only 70% for vaccine B. So, based on this additional information, which of the two vaccines would you now prefer?

Before I proceed with explaining how I would answer this question, let me emphasize the paradoxical nature of these numbers. It is intuitively clear that it cannot be the case that vaccine A is best for men *and* best for women, but worst overall. Indeed, if these data came from a randomized controlled trial, such a paradox could not happen. The paradox stems from an incorrect interpretation of the *correlations* between treatment and outcome as *causal relations*, the very mistake that Karl Pearson warned us against!

To understand this better, we consider different causal hypotheses that might apply in our case. I have illustrated them as causal graphs in Figure 7. Under the hypothesis in (a), gender (Z) causes both vaccine type (X) and hospitalization outcome (Y). One can prove that under this causal hypothesis, the data implies that vaccine A should be preferred. Another hypothesis, the one in (b), assumes an unobserved common cause L of X and Z, which explains the observed correlation between these two variables. Also in this case, one can prove that vaccine A should

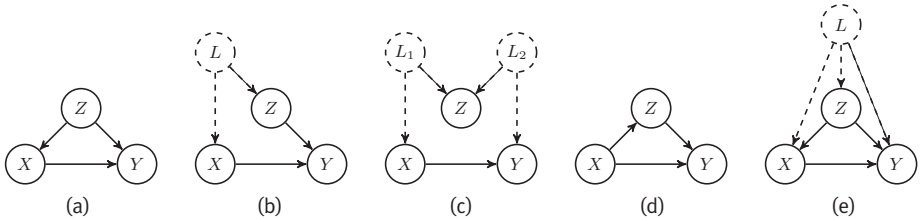


Figure 7 Different possible causal hypotheses for the data in Table 1 (X is treatment, Z is gender, Y is outcome, L is an unobserved confounder). One can show that for (a) and (b), treatment A is better than B, while for (c) and (d), the opposite holds. For (e) it is unclear which treatment is better.

be preferred. But, there also exist causal hypotheses for which one can prove that vaccine B should be preferred! Two such hypotheses are shown in (c) and (d). So we see that the right answer to the question which vaccine to prefer depends not only on the data, but also on our causal assumptions!

However, we can still rule out some of these hypotheses. Indeed, we may assume that treatment does not affect gender (I have heard many rumours and conspiracy theories about what undesired effects COVID-19 vaccines may have, but not yet that they transform males into females or vice versa!). That assumption allows us to rule out hypothesis (d). Furthermore, according to basic biology, gender is determined at conception by the chromosomes in the sperm cell that enters the egg cell and is an independent random event (like flipping a coin). Therefore, it is hard to think of any possible common cause of gender and treatment. That assumption allows us to rule out hypotheses (b) and (c).

Does this then lead us to the final conclusion that one should prefer vaccine A? Not yet! Indeed, we cannot easily rule out the possibility of another variable that causes treatment and outcome, as in hypothesis (e). For example, 'age' could be such a variable. Different vaccines have been used for different age groups, and elderly people generally have a higher risk of ending up in hospital with a virus infection. But including 'age' might lead to another instance of Simpson's paradox![10]

So, is there then nothing we can conclude from the data? One concrete answer to this question is provided by the *natural bounds* [12]. If we cannot rule out unobserved confounders, the best we can do is to calculate lower and upper bounds on the causal effect. For the mathematicians in the audience, I show here the theorem and the proof. Though conceptually advanced, the derivation itself is straightforward.

Theorem 1 [12]. *If treatment X, outcome Y and observed confounder Z are discrete variables, then in the presence of additional unobserved confounding (of X, Y and Z) we can bound:*

$$\begin{aligned} &p(X=x, Y=y \mid Z=z) \\ &\quad \leq p(Y(x)=y \mid Z=z) \\ &\quad \leq p(X=x, Y=y \mid Z=z) \\ &\qquad + 1 - p(X=x \mid Z=z). \end{aligned}$$

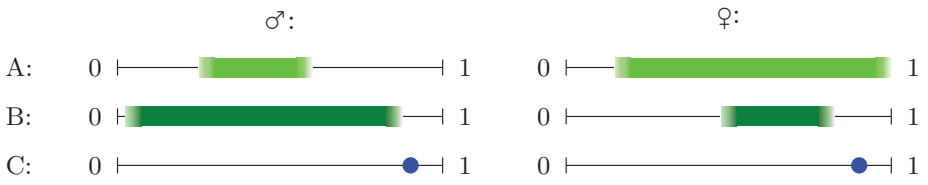


Figure 8 The probability of a positive outcome under enforced treatment x for gender z, $p(Y(x) \mid Z=z)$, must lie within the corresponding intervals (obtained by applying Theorem 1 to the data in Table 1). For a third possible treatment C, the probability of a positive outcome under enforced treatment is assumed to be precisely known (blue dots).

Proof. Using consistency $(Y = Y(X))$ and elementary probability theory:

$$\begin{aligned} &p(X=x, Y=y \mid Z=z) \\ &\quad = p(X=x, Y(x)=y \mid Z=z) \\ &\quad \leq p(Y(x)=y \mid Z=z) \\ &\quad = p(X=x, Y(x)=y \mid Z=z) \\ &\qquad + p(X \neq x, Y(x)=y \mid Z=z) \\ &\quad \leq p(X=x, Y=y \mid Z=z) \\ &\qquad + p(X \neq x \mid Z=z) \\ &\quad = p(X=x, Y=y \mid Z=z) \\ &\qquad + 1 - p(X=x \mid Z=z) \end{aligned}$$

for all z with $p(Z=z) > 0$. □

I have visualized these natural bounds for our data in Figure 8. The fact that the lower and upper bounds have to be estimated from a finite sample of individuals yields additional (statistical) uncertainty, which I have visualized here by making the bounds fuzzy. Note, however, that in this case most uncertainty is 'causal' (due to uncertainty about the causal relations) rather than 'statistical' (due to extrapolating from a finite sample to the entire population). In other words, most uncertainty would *not* go away as we collect more and more samples. If we gave all males vaccine A, they would have a probability for a positive outcome between $\sim 30\%$ and $\sim 55\%$ (in the left light-green bar in Figure 8). If, instead, we gave all males vaccine B, that probability must lie between $\sim 7.5\%$ and $\sim 82.5\%$ (in the left dark-green bar). Since we only know that these probabilities must fall within these two ranges, and these ranges overlap, we cannot conclude which vaccine is to be preferred for males (based on the data, and our causal hypothesis). For females, the situation is similar, although the probability ranges differ.

Does this mean that the observational data is useless? Not at all! Suppose there is a third vaccine C, for which a randomized controlled trial has been done, which has a probability of a positive outcome of 90%. Then we know that vaccine C is to be preferred over both vaccines A and B for males, and that it is preferred over vaccine B for females. So, in a randomized controlled trial we only need to compare vaccines A and C for females (which is obviously more efficient than setting up a large randomized controlled trial to compare all three vaccines for both genders).

Estimating causal effects from observational data is done routinely by medical researchers. What I don't understand, though, is why these bounds are typically not reported. Instead, researchers usually only report confidence intervals for point estimates, making the strong, typically untestable (and likely wrong), assumption that there is no unobserved confounding between treatment and outcome.

Gender bias at UC Berkeley?

The numbers in the previous example of Simpson's paradox were made up. But one also encounters this paradox in real life. A famous case concerned the admission of graduate students at University College Berkeley [1]. In 1973, university officials noticed in pooled data that while $\sim 45\%$ of all male applicants were admitted, only $\sim 30\%$ of all female applicants were admitted.[11] This being a statistically significant difference (with an astronomically small p-value of 10^{-22}), the officials called for a closer investigation as they were anxious the university might get sued for gender discrimination. Statisticians looking into the data noticed that the difference in acceptance probability mostly disappeared when looking at the level of individual departments, while some departments even showed a slight bias *in favor* of females. What would you conclude: does this provide evidence of gender bias, or not?

It is again helpful to consider possible causal hypotheses, three of which are illustrated in Figure 9. The simplest hypothesis is (a): gender (Z) causes department choice (M), which then influences admission (Y) — but without an unfair *direct* effect of gender on admission. Hypothesis (b) and (c) add an unobserved confounder

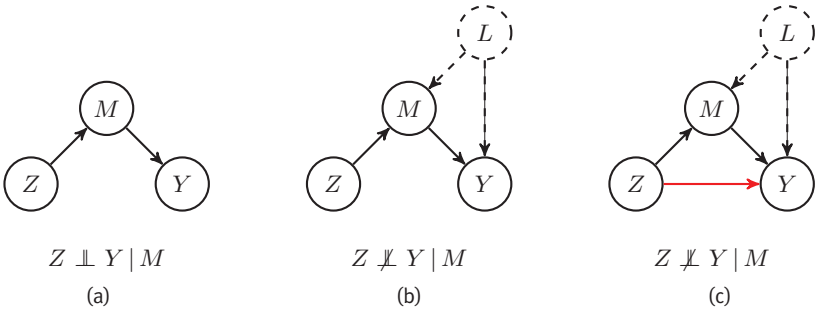


Figure 9. Three possible causal hypotheses for the Berkeley admission data (Z: gender; Y: admission; M: department choice; L: latent confounder). (c) is considered unfair, whereas (a) and (b) are considered fair. (a) implies a conditional independence between Z and Y given M, whereas for (b) and (c) this would generically not be the case.

(L) of department choice and admission. An example of such a confounder could be someone's math skills, which may influence both their department choice (as someone with bad math skills might be more likely to apply at a literature department than at an engineering department) and their admission (if math skills are taken into account in the selection procedure). Hypothesis (c), finally, incorporates the possibility of unfair gender bias as a direct effect of gender on admission (the arrow marked in red).

For (a) one can prove that admission rates are the same for males and females within each department, while for (b) and (c) this is not necessarily the case.[12] Testing for this conditional independence (of Z and Y given M) in the data, we find that hypothesis (a) is almost compatible with the data, but there is still some evidence of a weak conditional dependence of Z and Y given M.[13] If we decide to therefore reject hypothesis (a), we can proceed by performing a statistical test whether the data would favor hypothesis (c) over (b).[14] It turns out that the data is perfectly compatible with hypothesis (b), which describes a fair selection process, and contains no evidence that hypothesis (c) would be more likely. This means that based on the data itself, and our causal hypotheses, we *cannot* conclude that there must have been gender bias in the selection process.[15]

Gender bias at Dutch universities?

In the Netherlands, about 5% of the PhD students that graduate, do so *cum laude* ('with distinction'). Naïvely, one would expect that this percentage should be the same for male and female PhD students. However, in 2018, the Dutch newspaper *NRC* reported that at most Dutch universities, the fraction of male PhD students that graduate *cum laude* is about twice as large as the fraction of female PhD students that do so [3]:

> "Aan alle Nederlandse universiteiten hadden mannen de afgelopen jaren meer kans om *cum laude* te promoveren dan vrouwen. De criteria voor *cum laude* promoveren zijn niet objectief gedefinieerd, dus er is volop ruimte voor genderbias."

The University of Amsterdam is no exception: 6.1% of the male PhD candidates graduated cum laude in 2010–2017, versus 3.7% of the female PhD candidates according to the newspaper article.

This story starts out analogous to that of the UC Berkeley admission case, but I do not know how it unfolds. Could this be an instance of Simpson's paradox, and would the difference disappear if we conditioned on, say, faculty? Or is this perhaps a real case of gender bias? I do not know, but it certainly deserves closer investigation![16]

My research

After this introduction to the field of causality, I would like to say a few words about my own research. I have tried to summarize the research I have been doing over the past 15 years on a single slide. The result is shown in Figure 10.

It all starts with research questions related to *modeling*: in other words, what are appropriate and useful mathematical representations of causality in various types of systems.

Once the modeling framework has been established, the next stage consists of developing mathematical results that aid causal *reasoning*, for example, Markov properties, or the natural bounds that we have seen before. Some challenging aspects in these first two stages are feedback loops (or 'causal cycles', more on those in a minute), modeling and reasoning with data from different domains (for example, combining observational and interventional data), and modeling dynamical systems (where variables become time-dependent). Generality and complexity both increase when not ruling out unobserved confounding and selection bias.

The next stage, 'problem solving', is where most of the hard work has to be done: in order to answer causal questions, we develop statistical algorithms

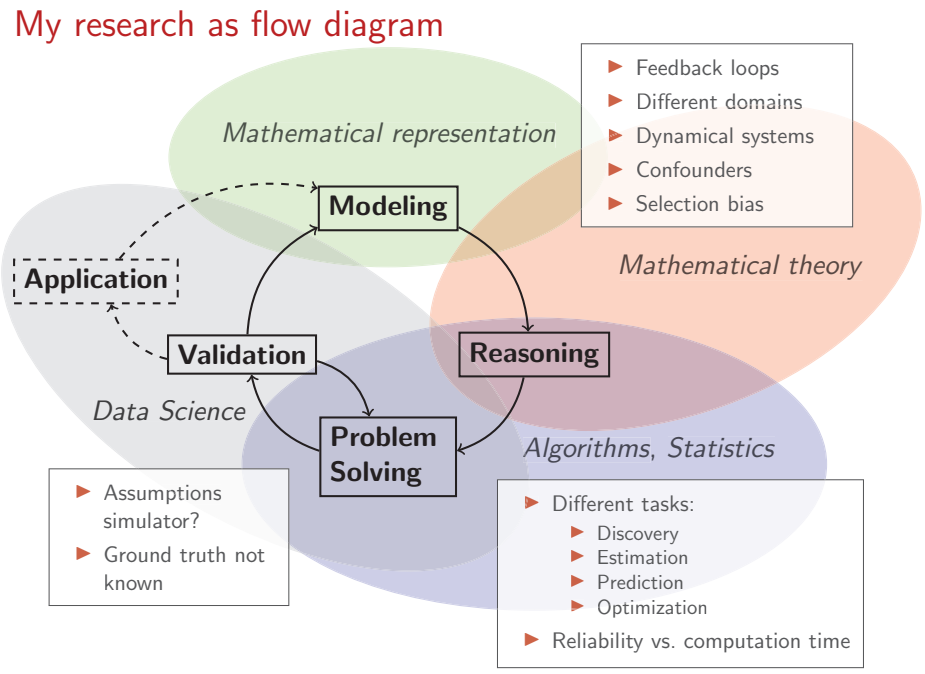


Figure 10 Summary of my research over the past 15 years.

that estimate quantities of interest from data. We improve existing algorithms to become more generally applicable, or more efficient in terms of statistical power or computation time. We attempt to give theoretical guarantees for our algorithms: we try to prove that with a high probability, the algorithms give the correct answer for sufficiently large data sets, under suitable assumptions. Here, one has to distinguish different tasks: discovering causal relations, estimation of model parameters, predicting results of actions, or optimizing a reward. Each problem can often be solved in many different ways, yielding different trade-offs between statistical power, computation time and generality.

As the British say, "the proof of the pudding is in the eating". The next, and perhaps most important, stage is to *validate* how well the algorithms work on real data, by establishing benchmarks. This is the part of the field that is still most in its infancy, especially compared with other parts of AI (like computer vision and natural language processing), where the availability of good benchmarks has led to rapid and impressive developments. It is easy to run computer simulations to get a feeling for how well our algorithms perform on finite data sets, but how realistic are such simulations? When using real data, we often face the problem that the ground truth is not known, and we cannot assess how reliable our algorithms are. In my experience, results on real data often turn out to be disappointing, which then motivates going back to the 'problem solving' stage, and trying harder to design an efficient algorithm, or to go back entirely to the 'modeling' stage.

I would love to tell you more about several of my research projects, but time is limited. Therefore, I just picked two topics that I am especially passionate about.

Feedback loops

The research I am most proud of is our work on *feedback loops*. Many causality researchers have, for a long time, considered feedback loops as exotic, complicated, and some even went as far as to question their very existence. The typical mindset of many causality researchers regarding causal cycles seemed to be: an intriguing possibility, but unlikely to be relevant in the real world.

However, once you start looking for feedback loops, you will see them everywhere. In other fields, no one questions their existence (see also Figure 11). As an example, take climate science. Citing a report of the United Nations Environment Programme [13]:

> "Part of the uncertainty around future climates relates to important feedbacks between different parts of the climate system: air temperatures, ice and snow albedo (reflection of the sun's rays), and clouds."

The word feedback appears 43 times in this 238-page report! Or, in biology. I cite from an editorial of the American Society of Clinical Oncology daily news [9]:

> "Feedback mechanisms may be critical to allow cells to achieve the fine balance between dysregulated signaling and uncontrolled cell proliferation (a hallmark of cancer) as well as the capacity to switch pathways on or off when needed for physiologic purposes."

Or, take econometry. In simple models of an ideal economic market, a feedback loop determines the equilibrium price of a commodity trading good: the price is determined by supply and demand, while supply and demand both depend on price. This is why it is difficult to predict, for example, the gas price.

In a series of papers that my collaborators and I wrote over the past 10 years, we extended the popular causal modeling framework of path diagrams (pioneered by Wright one century ago) to incorporate feedback loops, and we generalized causal reasoning theory and algorithms to allow for feedback loops.

Causal discovery

A research topic that I am also passionate about is *causal discovery*, that is, inferring the presence or absence of causal relations from data. An intriguing alternative to randomized controlled trials is to exploit conditional independences. About three decades ago, causality researchers realized that certain statistical patterns in data can be considered as `fingerprints' of the causal relationships between the variables. These conditional independence patterns can be probed with statistical tests, and under certain assumptions one can draw some conclusions on what the causal relationships between the variables must have been. I illustrated the inference process in Figure 12. Interestingly, this idea works even when allowing for unobserved confounders and selection bias (and, as we have shown more recently, also causal cycles).

Based on this principle, many causal discovery algorithms have been devel-

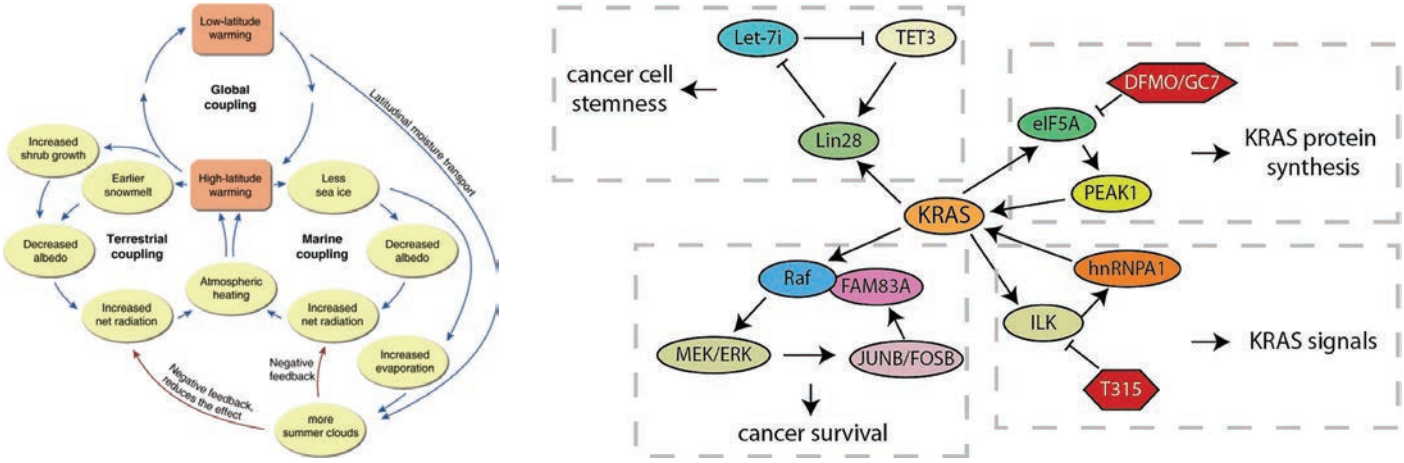
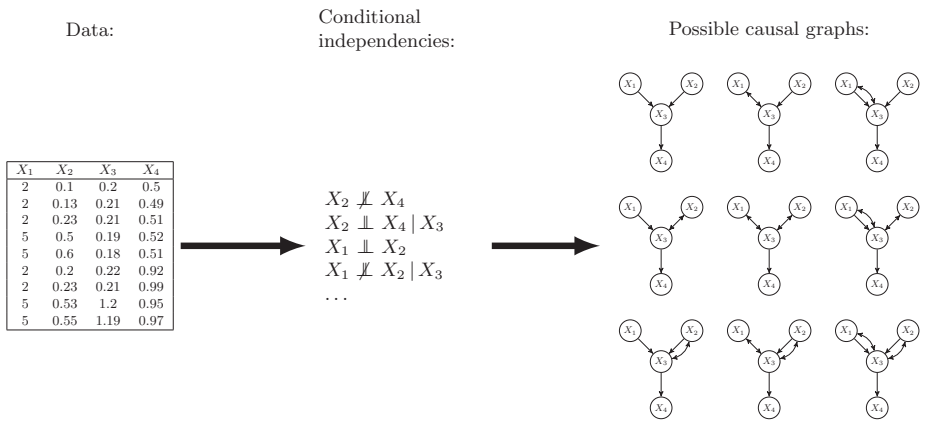


Figure 11 Feedback loops occur in many different systems, for example in climate science[17] (left) and in biology[18] (right).



X_1	X_2	X_3	X_4
2	0.1	0.2	0.5
2	0.13	0.21	0.49
2	0.23	0.21	0.51
5	0.5	0.19	0.52
5	0.6	0.18	0.51
2	0.2	0.22	0.92
2	0.23	0.21	0.99
5	0.53	1.2	0.95
5	0.55	1.19	0.97

Figure 12 Constraint-based causal discovery algorithms infer possible causal structures from conditional independences (certain statistical patterns that can be tested in data).

oped. We can prove that these algorithms work under suitable assumptions. We can demonstrate empirically that they work in computer simulations. But do they work in the real world? Answering this question turned out to be surprisingly hard.

One of our attempts involves yeast. Yeast is a pretty useful organism: you can use it to bake bread, to brew beer, or, as we did, to validate causal discovery algorithms. We made use of a large-scale gene expression data set for yeast, measured in a huge experimental effort by researchers from UMC Utrecht and Utrecht University [8]. The expression levels of about 6000 yeast genes were measured under many different experimental conditions, including almost 1500 single gene knockout interventions. This gigantic randomized controlled trial allows us to estimate the gene regulatory network. The resulting causal graph (with almost 6000 variables!) is depicted in Figure 13, and as you can see, it looks pretty complicated (while it is not even complete!).

The challenge we took on was to predict from this data which gene expression levels change if we knock-out a certain gene *without actually using the data for that knock-out experiment*. This challenge, at first sight, appears similar to reading off from the data in Figure 4 whether chocolate consumption causes Nobel laureates. I have visualized the expression data for two examples of correlated gene pairs in Figure 14. On the horizontal axis we have the expression of one gene, and on the vertical axis, the expression of another gene. The data points correspond with relative gene expression levels in colonies of yeast cells. The blue points are the observational data, and the red points are interventional data corresponding with different single gene knockouts. The causal discovery algorithm is solving a challenging task: it has to decide from the data what will happen with the expression of the gene on the vertical axis if we reduce the expression of the gene on the horizontal axis by knocking it out. There are millions of such gene pairs that we can consider!

A key difference with the data in Figure 4 is that we have an additional variable: the experiment type (observational or interventional). With a conditional independence test, we can use this additional information to decide whether an observed correlation between gene expression levels implies a causal relationship of one on the other. Figure 14(a) shows a case in which the causal discovery algorithm correctly identified a causal relation: the expression of gene YPL273W causes the expression of gene YMR321C.[20] In Figure 14(b) we see a case where a causal relation found by the causal discovery algorithm is incorrect: there, knocking out YPL154C does not significantly change the expression of YDR032C.

With collaborators of ETH Zürich we validated how well causal discovery algorithms perform in this challenge [11]. Here I will focus on LCD, one of the simplest constraint-based causal discovery algorithms [4]. Out of more than 19 million candidate cause-effect gene pairs, I preselected roughly 1000 using an algorithm called L_2Boosting, and of those 88 were selected by the causal discovery algorithm LCD. For 9 of those, the data provides the ground truth causal effect. One can see clearly in

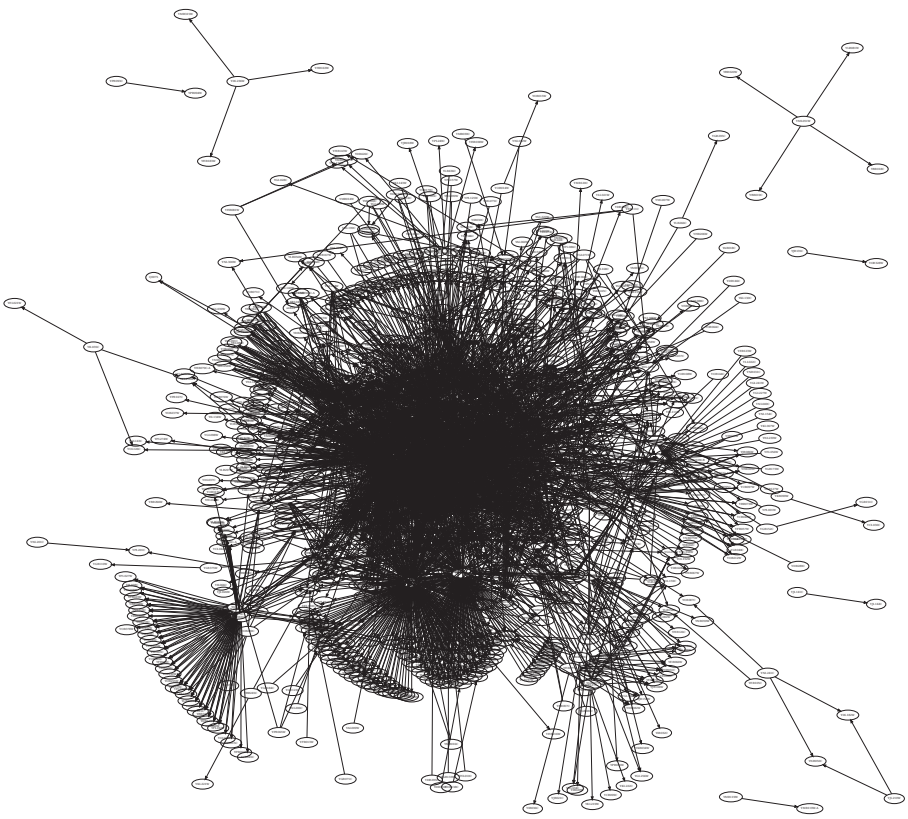


Figure 13 Estimated subgraph of the gene regulatory network for yeast, based on experimental data from [8].[19]

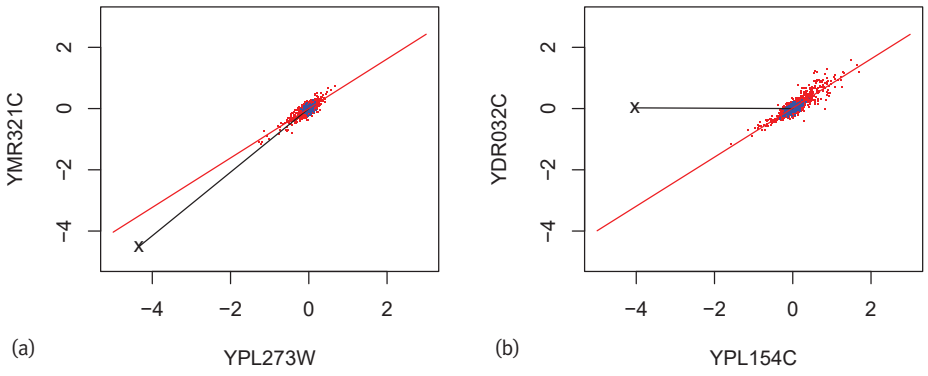


Figure 14 Examples of correlated gene pairs in the yeast gene expression data from [8]. Observational data are in blue, interventional (some gene knockout) in red, and the black cross shows the result of knocking out the gene on the horizontal axis. (a) is causal (YPL273W causes YMR321C), while (b) is not (YPL154C does not cause YDR032C).

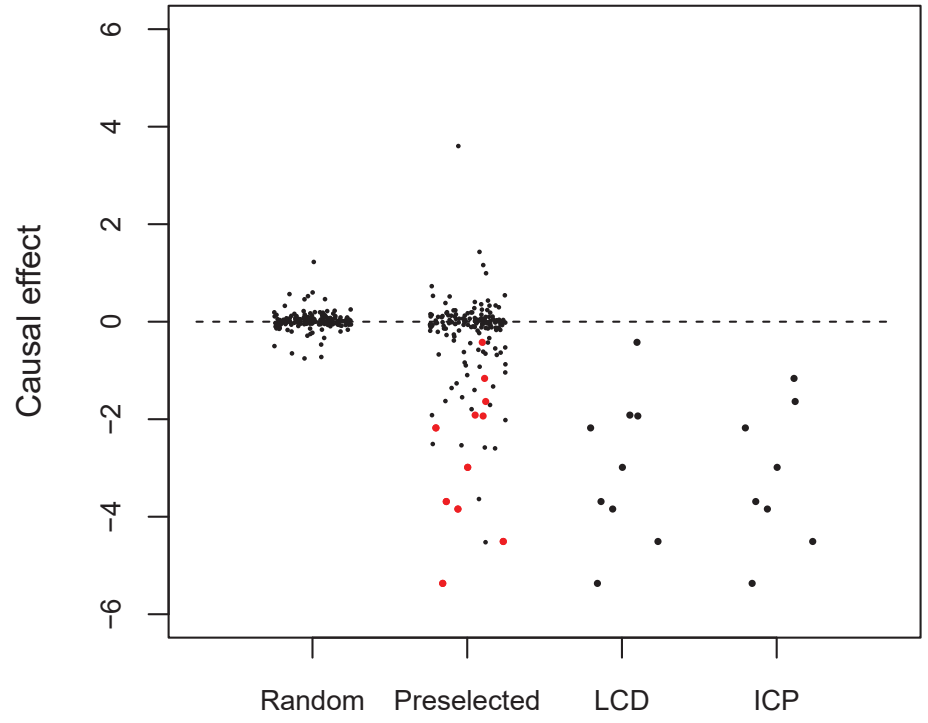


Figure 15 Validation of causal discovery algorithms with the knockout data from [8]. Random: randomly selected gene pairs (baseline); Preselected: gene pairs preselected using L_2Boosting; LCD/ICP: further selected using LCD/ICP.

Figure 15 that these 9 gene pairs indeed correspond with a nonzero causal effect, and hence, a true gene regulatory relation. To my knowledge, this was the first convincing proof that causal discovery based on conditional independences can indeed work in practice.

Applied research

While the core of my research has always been theoretical, I also enjoy more applied research projects where we try to put the theory to use. I will mention a few examples. I am involved in the AI4Science project of the Faculty of Science of the University of Amsterdam, where we collaborate with biologists to apply and further develop causal discovery methods to improve the understanding of gene regulation in a bacterium, *bacillus subtilis*. Microsoft Research sponsors one of my PhD students to study how causality can be used for developing personalised medication and treatment strategies. The Mercury Machine Learning Lab, one of the labs of the Innovation Centre for Artificial Intelligence, is funded by booking.com. I am involved in this lab as co-director and as a supervisor of two PhD students and a postdoc. The aim of the lab is to carry out fundamental scientific research on learning from controlled sources, for example for taking informed business decisions. It should come as no surprise that causality is considered to be of essential importance. These collaborations are great opportunities to see some of the algorithms being deployed in practice.

Ik heb gezegd.

Acknowledgments

I thank my colleagues of the Korteweg-de Vries Institute for Mathematics, and in particular Eric Opdam, Michel Mandjes, and the others who played a role in my transition from the Informatics Institute. I hope that I will successfully follow in the footsteps of my precursors on this chair. I am indebted to those who funded my research, amongst them NWO, ERC, various universities and companies, but also all European tax payers. I have had the privilege of doing research with many different people, and I would like to thank them all. In particular, many thanks to my former and current PhD students Tineke Blom, Stephan Bongers, Philip Versteeg, Noud de Kroon, Teodora Pandeva, Philip Boeken and Leihao Chen, and my former postdocs Thijs van Ommen, Patrick Forré and Sara Magliacane, for the interesting discussions and great collaborations.

Notes

1 While this multi-disciplinarity emphasizes its scientific and practical relevance, it also bears a cost: many ideas are lost in translation. The often isolated, scientific communities have developed different (sometimes contradictory) terminology and conceptual frameworks. Even within statistics, different frameworks are being used. This is unfortunate, as it puts a considerable hurdle on scientific exchange and interaction.

2 Reproduced from https://ourworldindata.org/smoking-big-problem-in-brief, licensed under CC BY 4.0.

3 Reproduced from https://tylervigen.com/spurious-correlations, licensed under CC BY 4.0.

4 https://tylervigen.com/spurious-correlations.

5 A more precise definition is the following. Consider modeling a system with two real-valued variables X and Y, where X is the cause and Y the effect (for example, X could be the dosage of ibuprofen administered to a patient, and Y the patient's body temperature measured 60 minutes later). The potential outcomes are random variables $Y(x) : \Omega \to \mathbb{R}$ for $x \in [0, \infty)$, where $Y(x)$ is distributed as the effect after an intervention that sets X to x, and Ω is a probability space. At most one of the po-

tential outcomes can be measured (indeed, it is not possible to administer two different dosages simultaneously). Alternatively, one can define a statistical causal model that consists of the family of distributions $x \mapsto \mathbb{P}(Y \mid \mathrm{do}(x))$, with $\mathbb{P}(Y \mid \mathrm{do}(x))$ the distribution of $Y(x)$. Here, x is treated as a parameter, and we have avoided to introduce the potential outcomes as random variables jointly defined on the same probability space Ω. While the potential outcomes allow for specification of their joint distribution, the alternative approach only provides access to their marginal distributions.

6 For an account on the history of the various frameworks by some of the pioneering researchers, see *Journal of Observational Studies* 9(2), 2022.

7 I reproduced the visualization in [10], but made use of newer data from https://www.theobroma-cacao.de/wissen/wirtschaft/international/konsum on chocolate consumption in 2017, and from https://en.wikipedia.org/wiki/List_of_countries_by_Nobel_laureates_per_capita on scientific Nobel laureates until 2019. This explains deviations from Figure 1 in [10].

8 That is actually a simplification. Instead, astronomers also rely on experiments performed in our solar system, but they extrapolate the results of these experiments to the rest of the universe, assuming that the same physical laws hold everywhere.

9 Keep in mind that these numbers are not supposed to be realistic at all, the numbers were just chosen to make it easier to tell this story. I took them from [18].

10 It could be the case that for each combination of gender and age, vaccine B seems preferable, while vaccine A seems preferable when aggregating the data over all age groups.

11 For my calculations, I used the 4257 samples concerning the six largest departments, available as `UCBAdmissions` in the `R` package `datasets`. This appears to be a subset of the 12763 samples reported in [1], which may explain any discrepancy with their numbers.

12 This amounts to showing that hypothesis (a) implies the conditional independence of Z and Y given M, something that follows for example by using the d-separation criterion [15]. In (b), conditioning on department choice may create a dependence between gender and admission via the 'explaining away' phenomenon (Berkson's paradox). In (c), a direct effect of gender on admission can also lead to a conditional dependence between gender and admission given department choice.

13 Using the G-test to test for dependence between gender and admission, one finds that gender and admission are strongly correlated when aggregating over departments (p-value 4×10^{-22} for testing $Z \perp\!\!\!\perp Y$), whereas after conditioning on department only a slight correlation is left (p-value 0.0014 for testing $Z \perp\!\!\!\perp Y \mid M$).

14 This is more complicated than testing for a conditional independence, but can be done by defining 'response variables' and using linear programming to characterize the probability distributions compatible with hypothesis (b) as the convex hull of a finite number of extreme points (see also [2]), which can be translated into a finite set of inequality constraints on $p(Z, Y \mid M)$. Since the empirical distribution turns out to fall inside this convex hull, hypothesis (b) need not be rejected in favor of hypothesis (c).

15 This conclusion differs from that of [1], who conclude that the data suggest evidence of gender bias *against* males.

16 Unfortunately, the data is not publicly available.

17 Reproduced from https://www.grida.no/resources/5261 with permission of the creator (Hugo Ahlenius, UNEP/GRID-Arendal).

18 Reproduced with permission from [6], licensed under CC BY 4.0 (Creative Commens).

19 In this graph, directed edges represent an estimated subset of the direct causal effects of gene knockouts on gene expressions. I defined a (possibly indirect) causal effect as a $\log_2$ fold change of at least ± 2, according to the data from [8], and then took the transitive reduction of the corresponding directed graph to obtain a minimal set of direct causal effects.

20 Indeed, the black cross in Figure 14(a) denotes the data point that corresponds with the knockout of YPL273W. In this knockout experiment, we see that the expression of YPL273W is substantially reduced (so normally this gene is active), but also the expression of YMR321C; hence, YPL273W causes YMR321C. Interestingly, YPL273W and YMR321C turn out to be *paralogs*: these two genes contain an almost ($> 98\%$) identical subsequence of more than 300 base pairs. Therefore, I do not know whether this finding that YPL273W causes the expression of YMR321C is biologically interesting, or is just an artefact of the experimental knockout procedure.

References

1 Peter J. Bickel, Eugene A. Hammel and J. William O'Connell, Sex bias in graduate admissions: Data from Berkeley, *Science* 187 (1975), 398–403.

2 Blai Bonet, Instrumentality tests revisited, in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI 2001)*, 2001, pp. 48–55.

3 Ellen de Bruin, Helft van de promovendi is vrouw, maar cum laude krijgen ze zelden, *NRC Handelsblad*, 20 October 2018.

4 Gregory F. Cooper, A simple constraint-based algorithm for efficiently mining observational databases for causal relationships, *Data Mining and Knowledge Discovery* 1(2) (1997), 203–224.

5 Gartner, Gartner identifies key emerging technologies expanding immersive experiences, accelerating AI automation and optimizing technologist delivery, gartner.com, 10 August 2022.

6 Weigang Gu, HongZhang Shen, Lu Xie, Xiaofeng Zhang and Jianfeng Yang, The role of feedback loops in targeted therapy for pancreatic cancer, *Frontiers in Oncology* 12 (2022), 800140.

7 David Hume, *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*, John Noon, 1740.

8 Patrick Kemmeren, Katrin Sameith, Loes A. L. van de Pasch, Joris J. Benschop, Tineke L. Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W. Ko, Sebastiaan van Heesch, Mehdi M. Kashani, Giannis Ampatziadis-Michailidis, Mariel O. Brok, Nathalie A. C. H. Brabers, Anthony J. Miles, Diane Bouwmeester, Sander R. van Hooff, Harm van Bakel, Erik Sluiters, Linda V. Bakker, Berend Snel, Philip Lijnzaad, Dik van Leenen, Marian J.A. Groot Koerkamp and Frank C. P. Holstege, Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors, *Cell* 157(3) (2014), 740–752.

9 Grant A. McArthur, The RAS/RAF/MEK/ERK pathway in cancer: Combination therapies and overcoming feedback, Editorial of the *ASCO Daily News*, June 2014

10 Franz H. Messerli, Chocolate consumption, cognitive function, and Nobel laureates, *New England Journal of Medicine* 367 (2012), 1562–1564.

11 Nicolai Meinshausen, Alain Hauser, Joris M. Mooij, Jonas Peters, Philip Versteeg and Peter Bühlmann, Methods for causal inference from gene perturbation experiments and validation, *Proceedings of the National Academy of Sciences of the United States of America* 113(27) (2016), 7361–7368.

12 Charles F. Manski and Daniel S. Nagin, Bounding disagreements about treatment effects, *Sociological Methodology* 28(1) (1998), 99–137.

13 James E. Overland, John E. Walsh and Muyin Wang, Why are ice and snow changing?, in UNEP/GRID-Arendal, ed., *Global Outlook for Ice & Snow*. UNEP, 2007.

14 Karl Pearson, *The Grammar Of Science*, Adam & Charles Black, 1892.

15 Judea Pearl, *Causality: Models, Reasoning and Inference*, Cambridge University Press, 2009.

16 Judea Pearl, *The Book of Why: The New Science of Cause and Effect*, Basic Books, 2018.

17 Bertrand Russell, On the notion of cause. *Proceedings of the Aristotelian Society* 13 (1913), 1–26.

18 Larry Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, Springer, 2004.

19 Alfred N. Whitehead and Bertrand Russell, *Principia Mathematica*, Cambridge University Press, 1925–1927.

20 Sewall Wright, The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs, *Proceedings of the National Academy of Sciences* 6 (1920), 320–332.