How Can AI Discover Cause and Effect?

causaLens

info@causaLens.com
causalens.com

How Can AI Discover Cause and Effect?

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

Until recently, discovering cause-and-effect relationships involved conducting a carefully controlled experiment or else relying on human intuition. Breakthroughs in science and technology have opened up new ways to search for causes.

We explain how Causal AI autonomously finds causes, using "causal discovery algorithms", and how it also boosts experimentation and human intuition.

What is causal discovery?

Let's take a toy example. Suppose we have a dataset with the following variables: **ice cream sales**, **shark attacks**, **temperature**, and **wind speed**.

Let's assume the following story holds true. Hot weather in the summer months causes people both to buy ice cream and go swimming and surfing, leading to more shark attacks. Another meteorological phenomenon, wind, draws sharks to near-shore areas also resulting in attacks.

These are claims about cause and effect. We can summarize them in a causal graph (Figure 1).

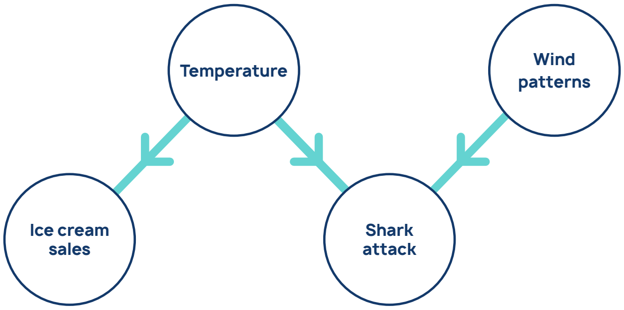


Figure 1. A causal graph for our example. Temperature is a common cause of both ice cream sales and shark attacks, which are also driven by coastal wind patterns.

The aim of causal discovery is to pick out the true causal structure, like the structure in the diagram above, from the data.

[1] We colloquially use the word "correlation" to talk about relationships between variables, but strictly speaking correlation is a measure of linear statistical dependence. We'll use the term "dependence" here to refer to any data pattern, linear or otherwise.

How causal knowledge is acquired

Causal knowledge can be acquired in three mutually complementary ways:

- through **experimentation** (ideally randomized controlled trials)
- via **human expertise** and intuitions
- and, with **causal discovery** algorithms

Experimentation is the gold standard for discovering causes, but many experiments are unethical or impossible to conduct. For instance, we can't throw people to the sharks or change the weather in order to observe the impact on ice cream sales.

We return to these three methodologies below, after focusing on causal discovery technology.

Correlation is not causation

Causal discovery from observational data is hard because correlation does not imply causation.

For example, shark attacks and ice cream sales are correlated or statistically dependent on each other (Figure 2). But there is no direct causal relationship between them — a third factor, the weather, is driving the correlation. This phenomenon is called confounding.

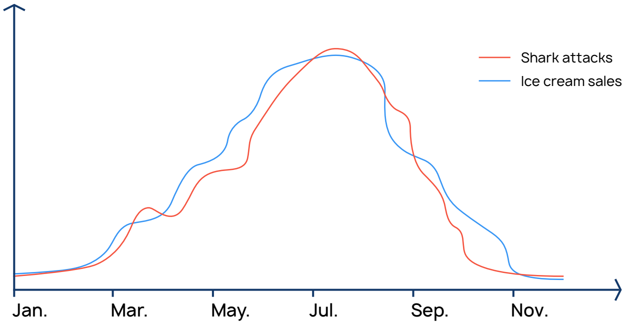


Figure 2. Correlation doesn't imply causation.

How Can AI Discover Cause and Effect?

If causes could be directly read off correlations, then causal discovery would be easy and current machine learning algorithms would already have solved it. But as it is, causal discovery requires specialized technology, that can dig beneath correlations to the underlying data-generating process.

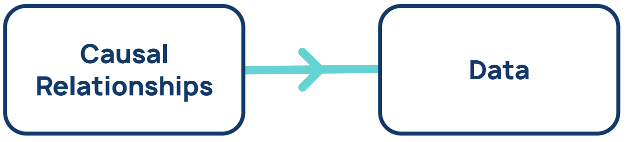


Figure 3. Causal relationships generate correlations found in data. Causal AI discovers these underlying causal relationships, whereas machine learning just analyses correlations.

Finding evidence for causation in data

Causal discovery algorithms can find clues for causal relationships in observational data. *Conditional independence* relationships are a key piece of evidence that many algorithms search for. Let's unpack this concept.

Two variables are *independent* if there's no relationship between them — knowing the value of one variable tells us nothing about the other one. Tesla's share price is independent of shark attacks, for instance.

Conditional independence builds on this concept:

★ Variables X, Y are *conditionally independent* of one another given a third variable (or set of variables) C, if knowing the value of C renders X and Y independent.

Shark attacks and ice cream sales are conditionally independent given the weather, because if we already know what the weather is then the incidence of shark attacks tell us absolutely nothing new about ice cream sales.

Why does conditional independence help with causal discovery?
Intuitively, testing for conditional independence is a little bit like running a controlled experiment. In an experiment, we try to isolate causal effects by controlling the environment and then modulating a variable that we're interested in. Causal discovery software doesn't have the luxury of actually controlling the environment as an experimenter would. But conditioning on all the background factors is the next best thing. (Our causaLake module does this automatically by implementing causal discovery algorithms on millions of variables across vast data lakes).

A classic causal discovery algorithm

Let's give a high-level overview of one causal discovery algorithm, to give a flavour of how they work.

"Constraint-based" algorithms are a type of causal discovery algorithm. They use conditional independence relationships as constraints, and construct causal structures that respect those constraints.

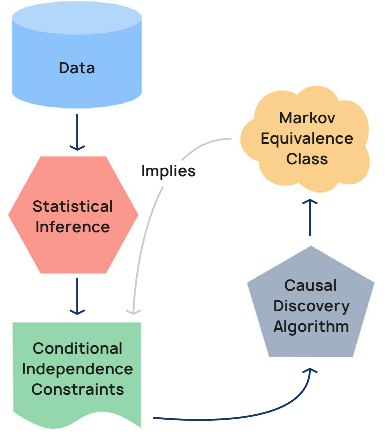
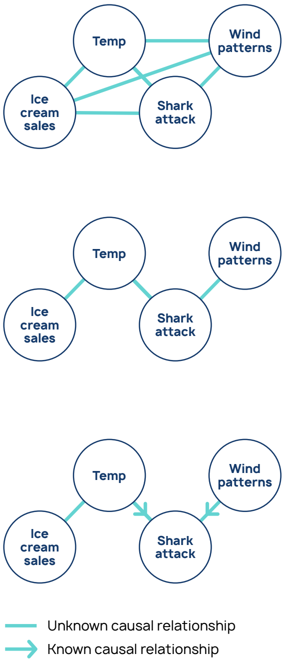


Figure 4. Workflow for constraint-based causal discovery. The AI establishes conditional independence relationships in the data. Causal discovery algorithms then discover the class of causal models that respect the conditional independence constraints. The algorithm outputs a set of candidate causal structures that are compatible with the data, called a "Markov Equivalence Class".

The "PC algorithm" is a classic constraint-based algorithm.



i. Start by assuming all variables are related somehow, and we don't know in which direction.

ii. Conduct a series of conditional independence tests. Look for independencies, and remove edges accordingly. Then look for independencies conditional on one other variable, and remove arrows. Repeat this procedure, incrementing the number of variables you're conditioning on.

iii. Orient any "colliders" (causal structures with the form C1 → E ← C2) — these have a distinctive signature in data, and so they can be leveraged to orient the edges. Propagate edge orientations by following the logic that the remaining edges are not colliders.

Figure 5. A worked example of the PC algorithm.

info@causaLens.com
causalens.com

Applying causal discovery in the real world is challenging

Causal discovery algorithms (like PC) are a giant conceptual leap forward in AI over conventional machine learning. But unfortunately the classic algorithms, many of which are available via open source software, all have limitations in real-world applications.

Unrealistic assumptions
Many popular causal discovery algorithms make excessively strong assumptions in most use cases. For instance, the PC algorithm assumes that there are no confounders lying outside of our data, which is often false. This can lead to inaccurate causal models.

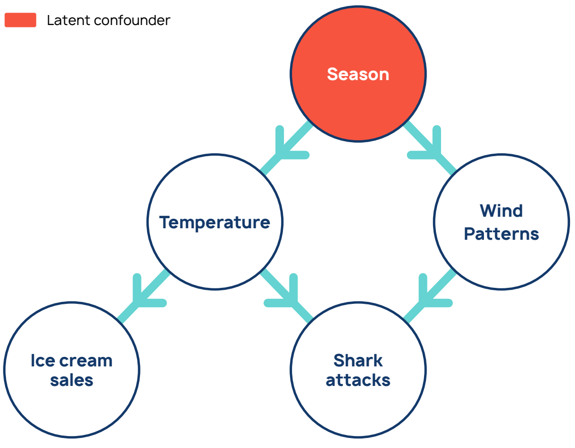


Figure 6. Some standard algorithms assume that there are no unobserved confounders, but in reality there usually are. For instance, the season (which is outside our example dataset) drives both temperature and wind patterns.

Computationally demanding
Many algorithms require lots of computation. For instance, the number of conditional independence tests PC has to run grows out of control as the data gets bigger. Classic "score-based" algorithms, another core approach which directly searches the space of possible causal structures, also suffer from inefficiencies. This makes most of these algorithms far too slow for many real-world use cases.

Pathological models
causaLens research has demonstrated that popular causal discovery algorithms have other problematic features. For example, we found that one score-based algorithm discovers radically different structures depending on the units that the data is denominated in. (Clearly causal structure shouldn't change depending on whether you're measuring in Fahrenheit or Celsius).

Many algorithms to choose from
There are a large number of causal discovery algorithms to choose from for any given application. Some are better suited to certain use cases than others. Selecting the right algorithm requires experience and expertise.

Can we do better?

causaLens technology implements all existing causal discovery algorithms. We have fine-tuned these algorithms to improve their performance in the field, and augmented them with our own methods. The technology autonomously applies the best methods in each situation to discover causal structure, without the need for domain experts or data scientists.

Optionally, domain experts can add their knowledge to the causal discovery process. The causaLab module is designed to integrate human knowledge from the ground up.

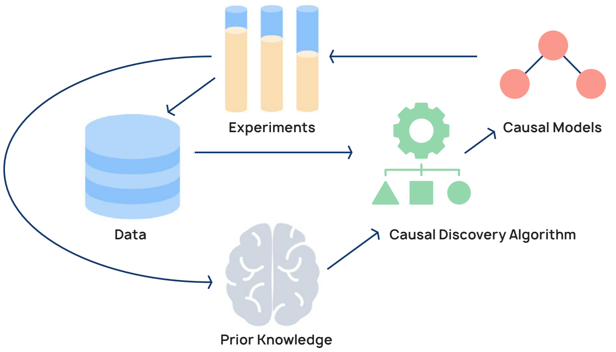


Figure 7. Experimentation, domain expertise, and causal discovery algorithms are mutually complementary methods for generating causal knowledge. Causal AI does not replace experimentation and expertise, it boosts it.

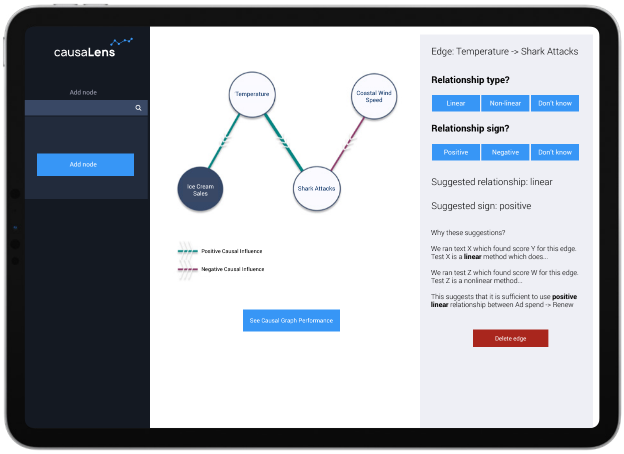


Figure 8. causaLab enables users to integrate their expertise and intuition with data-driven causal discovery algorithms. Users can select causal variables and define their causal directionality, the relationship type and sign, and how the relationship might change over time. Human input combined with causal discovery creates a powerful synergy.

The benefits of causal discovery

In healthcare, causaLens technology has accelerated discovery of protein biomarkers as a predictor for cancer by 100x. And in business, our algorithms are used by sophisticated marketing departments to identify the true drivers of customer behaviour.

AI systems become more intelligent as a result of causal discovery, and can do lots of things correlation-based machine learning algorithms can't do. This includes richer explanations, enhanced human-machine partnership, fairer algorithms, models that do not easily break, and AI systems that can meaningfully assist with decision-making.

We understand why someone might choose causal discovery over all the riches in Persia.

Further reading

For more technical information on causal discovery, please read our tech report.

To learn more about the benefits of Causal AI, check out our white papers and blog.

And if we've piqued your interest on how weather causes shark attacks, this article might interest you!