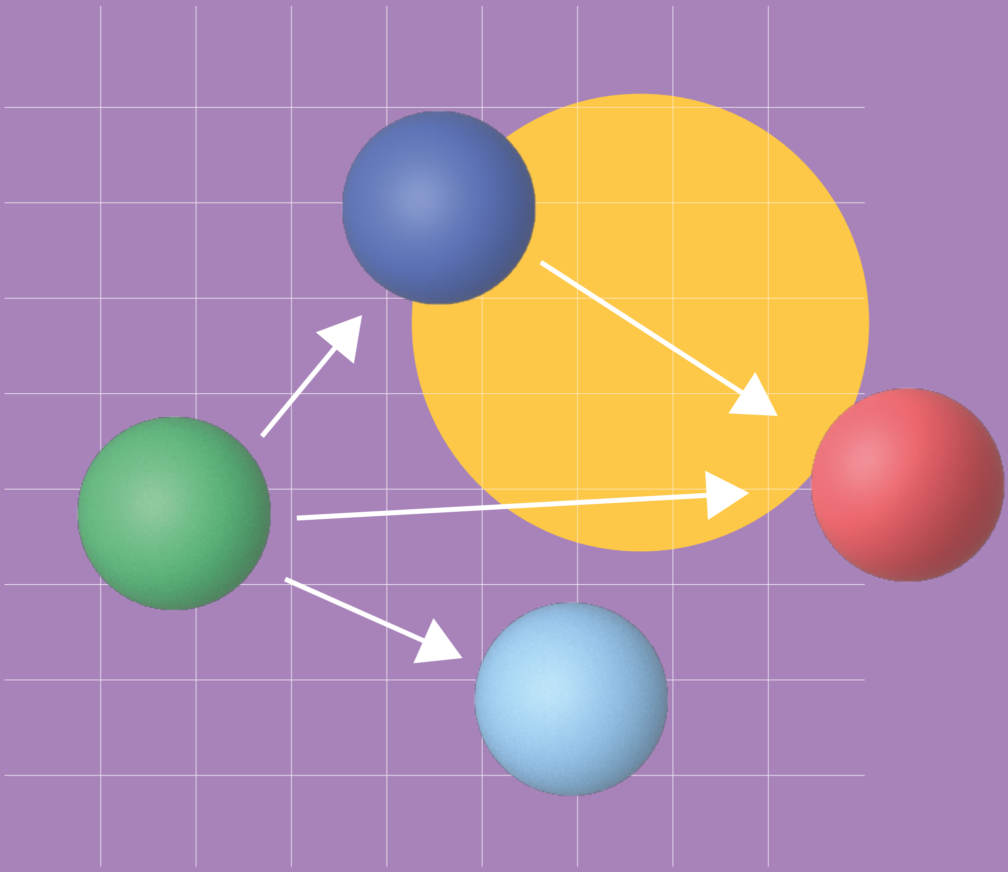
EBOOK

How to Move Beyond ML Predictions: An Introduction to Causal Inference



About the Authors

Léo Dreyfus-Schmidt is the VP of Research at Dataiku, where he leads the ML research team on their mission to reduce scientific uncertainty of ML topics and, ultimately, craft generic and robust solutions for enterprises. His topics of interest range from data shift, domain adaptation, and actionable uncertainty estimation to causal ML.

Jean-Yves Gérardy is a Research Scientist at Dataiku, where he works on enabling data scientists and data analysts to use causal inference to answer business questions. He has a background in economics and a keen interest in econometrics methods.

Table of Contents

Click any heading to navigate directly to the section.

Introduction

We are all inquisitive about alternative scenarios, perhaps it's simply in our nature. For example, one may wonder what their salary would be had they chosen a different major in college; another person may daydream about what the world would look like if there were no conflicts among nations; finally, a more practical question someone may ask is whether or not they would have missed their train if they had left home five minutes earlier.

Those "what if'' questions not only dominate our day-to-day lives — they also underpin our entire judiciary system. A defendant will most likely be indicted if it can be established that no harm would have come to the plaintiff had the defendant not done what they did.

Businesses are also naturally interested in these questions. For example, a gym owner might want to know whether a given customer would be more likely to renew their yearly subscription if they were offered a discount or subjected to a commercial. A phone manufacturer would certainly be interested in knowing whether or not increasing the price of its lead product would yield greater profits.

All of these questions are inquiries about the causal effect of an action or treatment (**T**) on some outcome of interest (**Y**). The scientific discipline used to answer that type of question from data is called **Causal Inference (CI).**

CI is arguably more difficult to comprehend than traditional machine learning (ML). It requires more constraining assumptions and uses different modeling techniques. Attempting to infer causal effects using the standard ML toolbox on observational data (i.e., data where the treatment variable is not under the control of researchers) is likely to give misleading results with catastrophic outcomes.

In this ebook, we will go over the idiosyncrasies of CI, and illustrate the danger of using regular ML to infer causal effects. As we will see, with ML models and under unsatisfied causal assumptions, one measures a correlation between **T** and **Y** which is rarely a causal effect. Correlation does not imply causation, as the adage goes.

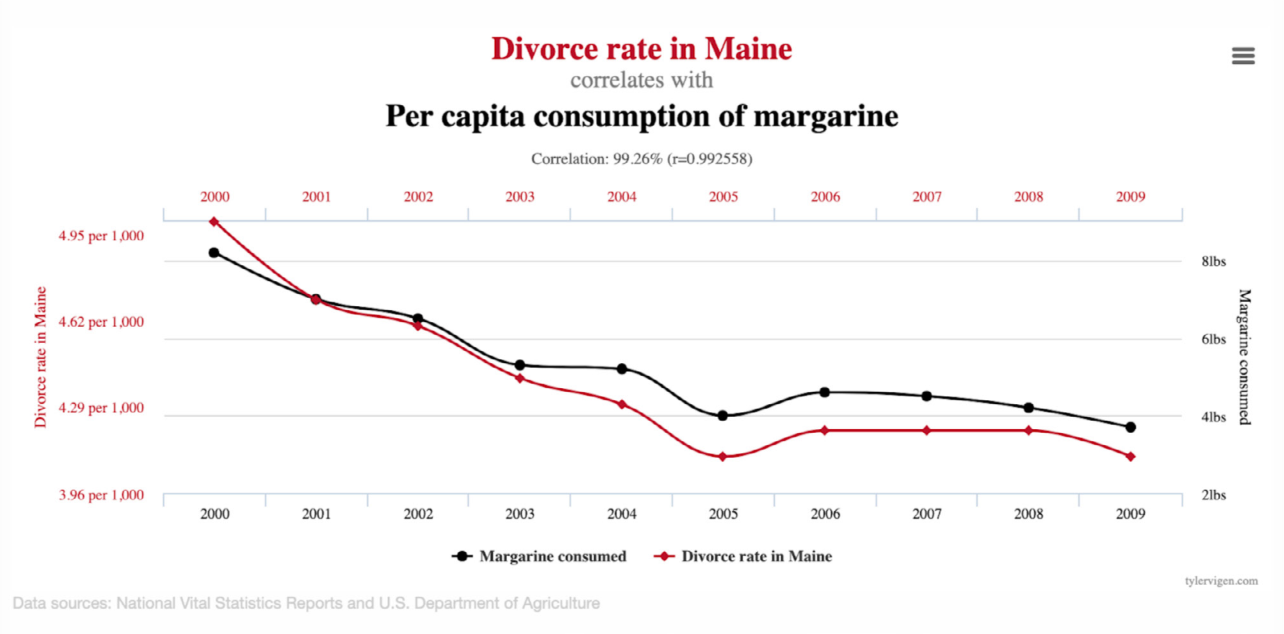
Correlation vs. Causation

Since the distinction between correlation and causation is at the heart of CI, let's spend a little time on those two terms. Suppose we have two events: event A and event B. Correlation means that we observe one of the following scenarios:

- A occurs then B occurs,
- B occurs then A occurs,
- A and B occur simultaneously.

Now, the human mind has a tendency to create dependence patterns when observing correlations, especially when an event occurs before another. For example, A then B could be wrongly interpreted as A **causes** B, that is "B could not have happened without A."

The way our minds make a causal leap when exposed to a simple correlation is made obvious in spurious correlation examples. That's what makes them amusing.

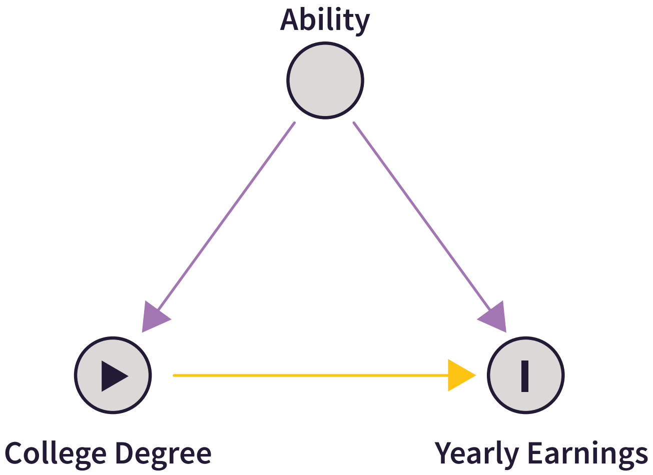


In the above chart, if we recognize the correlation between the two events, we know that one could not have caused the other. We intuit that, had one event not happened (e.g., a different divorce rate in Maine on a given year), the other event would have been left unchanged (per capita consumption of margarine).

While obvious in those spurious correlation examples, we are more easily tricked when correlations appeal to our deeply rooted beliefs. For example, if looking at census (observational) data, we find that college graduates earn, on average, an additional $20,000 every year compared to their non-college graduate counterparts, we may hastily conclude that a college degree causes an extra $20,000 in yearly salary.

While it is certainly true that education increases job prospects, is it true that a college degree causes a $20,000 bump in average earnings? We'll make the case that the true causal effect is probably lower than this measured correlation. To see this, let's invoke **Reichenback's Common Cause Principle** that says that if two events are correlated, then either one causes the other or both are caused by a common event. This common event is referred to as a **confounder**.

In our example, it is likely that college degrees cause higher earnings, but it is also plausible that an unobserved confounder, individual ability, causes both an increase in one's probability of getting a college degree and in one's earnings later in life. This unobserved ability inflates the true and unobserved causal effect of a college degree on earnings. In statistical parlance, we say that our causal estimate suffers from an **Omitted Variable Bias.**



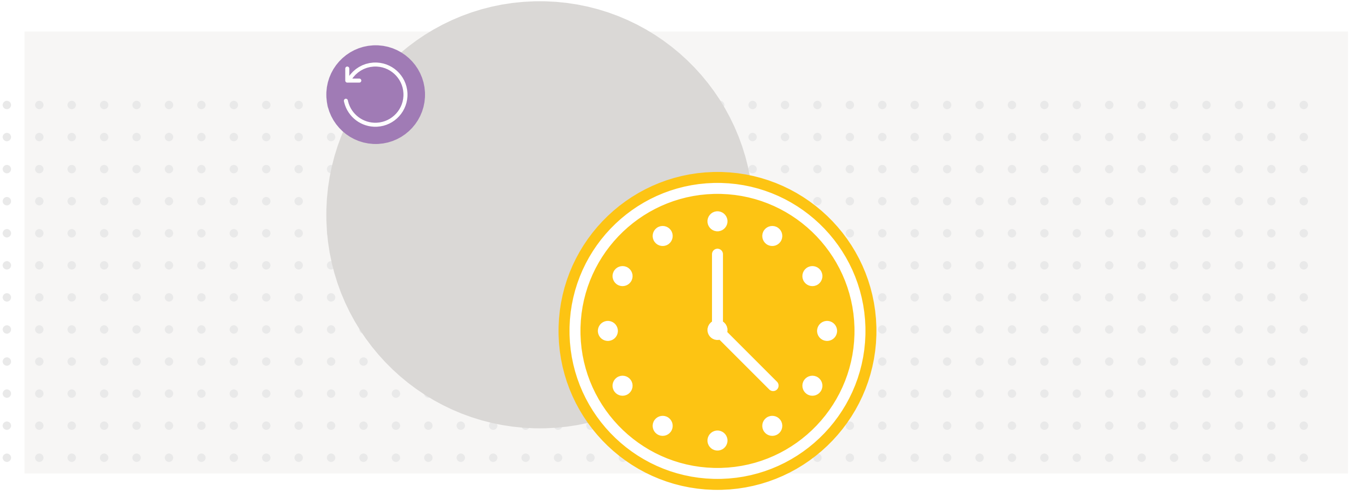
WHAT DOES THIS GRAPH MEAN?

The above *causal* graph summarizes the assumptions we're making about the data. An arrow going from node A to node B means that A causes B. To discover the true effect of *College Degree on Yearly Earnings*, one would have to somehow break the link between *Ability* and one of the other two variables

ESTIMATING COUNTERFACTUALS OR THE NEED FOR A TIME MACHINE

Ideally, to measure the true causal effect of *College Degree on Yearly Earnings*, we would need to travel back in time, take each individual from the census survey, and magically change their *College Degree* status. Then, years later, we would measure their earnings and compare them to their actual earnings.

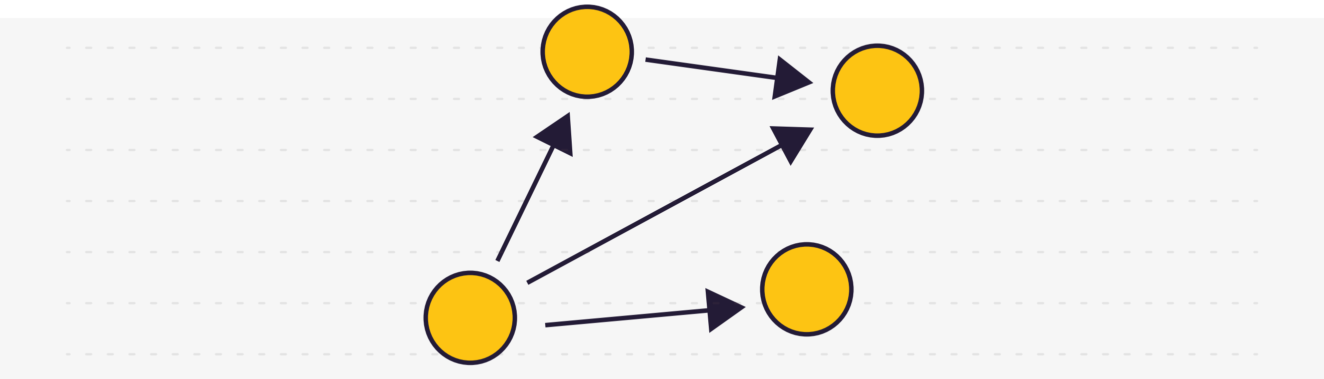
If we could measure individual earnings with and without a college degree, we would essentially do away with the influence of unobserved ability on having a college degree. Unfortunately, measuring an outcome under different scenarios for the same individual is not possible. This impossibility is known as the **Fundamental Problem of Causal Inference**, which we will introduce in more detail shortly.



In the case of education, policy makers need to have an accurate measure of the effect of college degrees on earnings, as this could help decide whether or not a policy aimed at encouraging people to get more education is worth the cost.

Now that we understand the distinction between correlation and causation, let's introduce it in one of CI's most popular applications in the field of marketing: uplift modeling. This will also allow us to introduce technicalities and causal jargon as gently as possible, as well as highlight how CI can sometimes recover causal effects, even without a time machine.

A Business Application: Uplift Modeling



Customer acquisition and retention is key to profitability for businesses selling any goods or services. That's why virtually all B2Cs are concerned with churn prevention.

Data-rich companies typically train ML models on historical customer data to predict a churn target. The model-predicted probabilities are used to select a follow-up action (whether promotions, discounts, emails, or phone calls) designed to entice the customer to stay.

Traditionally, businesses rank their customers by predicted probability of churn and take action only on those customers that fall into certain probability ranges. With the systemization of AI and the natural abundance of data, churn modeling has become one of the most standard AI use cases.

Although churn models help predict whether or not a customer is at risk of churning, they provide no indication as to whether or not that customer would react favorably to the follow-up action. It would be a waste of time and money to target customers that would not react positively to the action.

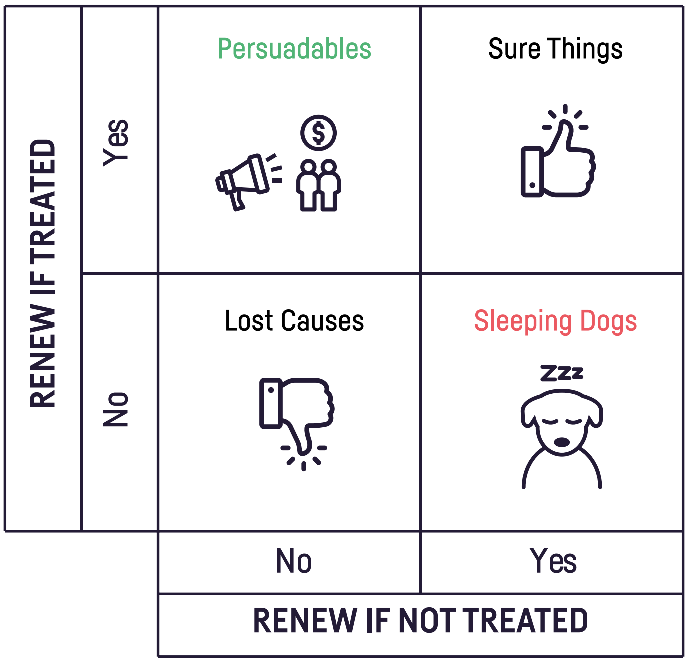
To know how a given customer would react to a follow-up action, it would be useful to know whether or not that customer would renew when she is "treated" with a marketing action (**T_i = 1**) as well as when she is left "untreated" (**T_i=0**). In other words, for customer or **unit i**, we're interested in measuring the **Individual Treatment Effect (ITE)** or **Causal Effect τ_i** defined as:

$$\tau_i = Y_i(1) - Y_i(0)$$

Where **$Y_i(T_i)$** stands for the **potential outcome** of individual i under treatment **T_i; Y_i(1)** is thus an indicator for subscription renewal when i receives the treatment, and **Y_i(0)** the same outcome when **i** does not receive it.

The combinations of those different potential outcomes give rise to four customer types:

- **Sure Things:** These customers will renew with and without treatment: $\boldsymbol{\tau_i = 1 - 1 = 0}$
- **Lost Causes:** These customers will churn with and without treatment: $\boldsymbol{\tau_i = 0 - 0 = 0}$
- **Sleeping Dogs:** These customers will churn if treated and renew if not treated: $\boldsymbol{\tau_i = 0 - 1 = -1}$
- **Persuadables:** These customers will renew if treated and churn if not treated: $\boldsymbol{\tau_i = 1 - 0 = 1}$



RENEW IF TREATED			
	Yes	Persuadables	Sure Things
	No	Lost Causes	Sleeping Dogs
		No	Yes
		RENEW IF NOT TREATED	

In the case of uplift modeling, CI is used to make predictions about the effect of the treatment on the renewal outcome at the individual level to ultimately focus on only targeting the Persuadables.

The Fundamental Problem of CI

Looking at the uplift modeling use case, the astute reader will have noticed that only one potential outcome is observed for a given individual, the other being a **counterfactual**. We cannot hope to observe the ITE let alone train an ML model to predict it. This is again the **fundamental problem of CI** that we hinted at earlier.

The next best thing we could hope to predict is the **Conditional Average Treatment Effect (CATE)**, defined as follows:

$$\tau(x) = \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]$$

Where $\mathbf{X_i}$ represents a vector of pre-treatment features with observed value **x**. In our uplift modeling example, it could be any features that describe customer behavior. This amounts to measuring the expected effect of the treatment among individuals that share the same characteristics features **x**.

It may seem that the fundamental problem of CI is also affecting the CATE. Indeed, it is a causal term (statisticians would call it a *causal estimand*) because it still depends on potential outcomes.

Thankfully, under some crucial but untestable assumptions on the data (**Conditional Ignorability, Consistency**, and **No-Interference** introduced below), one can replace those potential outcomes with some variables that could theoretically be observed in the data. In other words, one can transform those causal estimands into a statistical expression (or statistical estimand for our statistician friends). This property is called identifiability. For example, **identifiability** allows us the re-write the CATE as:

$$\tau(x) = \mathbb{E}[Y_i|X_i = x, T_i = 1] - \mathbb{E}[Y_i|X_i = x, T_i = 0]$$

Which is now a difference in the conditional expected outcome.

Identifiability is half the battle. Next comes **estimability**, or the ability to use statistical tools to come up with a value for the CATE using data. This step requires an extra assumption (**Common Support**) which happens to be testable. There are many causal models in the CI toolbox to estimate causal effects. We will introduce those models below.

Now, we'll return to our uplift application. A causal model will provide an estimate for the CATE. Since $\mathbf{Y}_i$ is a binary variable, the CATE is a difference in probability of conversion, and is therefore bound between -1 and 1.

With a good causal model, persuadable customers should have a predicted CATE approaching 1. On the other hand, "Sleeping Dogs" should have their prediction approach -1. Lastly, "Sure Things'' and "Lost Causes'' should have a predicted CATE close to 0. Based on that information, we can now target individuals by descending order of predicted uplift.

How Does Causal ML Differ From Traditional ML?

CI differs from traditional ML with respect to three aspects:

1. The assumptions made on the data
2. The models and their output
3. The metrics used to evaluate the performance of those models

DEEPER DIVE: ASSUMPTIONS FOR CAUSAL ML

Of the three aspects listed above, the assumptions are the most important. If unsatisfied, we're left with biased and misleading causal predictions, no matter how well our causal model is performing.

The following assumptions are crucial for the **identifiability** and **estimability** of causal effects.

CONDITIONAL IGNORABILITY

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \mid X_i$$

This assumption — also known as unconfoundedness — states that conditional on the features $\mathbf{X_i}$, the treatment $\mathbf{T_i}$ is independent of either potential outcome. We're not used to thinking in terms of potential outcomes, so this assumption might be difficult to unpack. A more intuitive way to explain this assumption is to say all the covariates $\mathbf{X_i}$ affecting both the treatment variable $\mathbf{Y_i}$ and the outcome variable $\mathbf{Y_i}$ were measured and can be used in the causal model. In other words, there are no unmeasured confounders.

Why Does This Assumption Matter?

As we saw in the education use case, if an unobserved covariate (ability) is positively correlated with both the treatment variable (College Degree) and the outcome variable (Salary), the causal effect of the treatment on the outcome variable will be biased (i.e., the treatment effect is not identified).

Another example is when a medical treatment is more readily given to people with a lower, unmeasured chance of recovery. The effect of treatment on recovery will be biased downwards. Another identifiability failure example that we will explore in greater detail in the next section is when a business offers discount vouchers only to their customers that are more likely to generate more revenue with the voucher. Here, we would simply overestimate the effect of the discount.

While crucial, this assumption *cannot be tested.* That's the reason why many businesses resort to experiments, where each individual is either assigned the treatment or the control at random. This randomization is a radical way of ensuring that $\mathbf{T_i}$ is not the consequence of any unobserved common cause of $\mathbf{T_i}$ and $\mathbf{Y_i}$. Note that experiments require researchers to set aside a fraction of individuals and test on them. Some of these individuals will very likely receive a treatment whose effect is detrimental to the business but this ensures reliable estimation of the causal effect.

Randomization does not necessarily mean that each individual is given a 50% chance of receiving the treatment. The field of experimental design has developed more sophisticated and efficient randomization methods whose exposition would be beyond the scope of this ebook.

For example, in an experiment where individuals are accrued by batch, researchers or an assignment algorithm can adjust probabilities based on the treatment effect measured from previous batches. Increasing treatment probability of predicted persuadables all the while decreasing that of other predicted profiles ensures that we are not "wasting" too many individuals for experimental purposes.

COMMON SUPPORT

Common support is a property of the data where every individual has a non-zero probability of being assigned any treatment.

$$0 < P(T_i = 1 \mid X_i = x) < 1 \ for \ all \ x$$

Where $\mathbf{P(T_i|X_i)}$ is **i**'s probability or the propensity of being treated.

Why Does This Assumption Matter?

A violation of this assumption means that for a subset of the data (characterized by $\mathbf{X_i{=}x}$), either everyone receives the treatment or everyone receives the control. In either case, we're missing a comparison group within the people characterized by $\mathbf{X_i{=}x}$, and causal models will need to extrapolate CATE predictions from the predictions of not so similar units. This is an estimability issue.

Outside experimental data, this propensity is not known, and therefore needs to be estimated using standard supervised ML (i.e., training a model of $\mathbf{T_i}$ as a function of $\mathbf{X_i}$).

Having a good estimate of $\mathbf{P(T_i \mid X_i)}$ (i.e., what we call a **propensity score model**) is of paramount importance for two reasons.

First, it is the basis for checking whether or not the common support assumption holds and potentially identifying violating subpopulations in the data. Once those subpopulations are identified, we may try to “fix” the common support assumption using two common strategies. The first way is to simply recognize that CATE predictions for those subpopulations are extrapolated (read unreliable) and need to be discarded.

The second option is to remove features from our analysis, if possible (say if they fully characterize the violating subpopulation). The more features we use, the finer our data can be segmented into subpopulations. This inevitably increases the chances of picking up a violation of the common support assumption in some of those finer subpopulations. Unfortunately, this last strategy elevates the risk of introducing omitted variable biases in the CATE estimates. As we remove features, we may indeed remove confounders.

Second, some classes of causal models rely on propensity score estimates to provide more robust predictions. Any errors in those estimates may compound into the CATE estimates.

NO INTERFERENCE

For each unit **i**

$$Y_i(T_1, \ldots, T_{1-1}, T_i, T_i, \ldots, T_N) = Y_i(T_i)$$

In plain English, this assumption states that for each unit **i**, the treatment status of others does not impact **i**’s potential outcome. Rather, **i**’s potential outcome is only a function of their own treatment.

An example of a violation of this assumption may arise when trying to measure the effect of vaccination on some individual health outcome. An unvaccinated person may still benefit from the vaccination of others through herd immunity.

Social interactions are also a threat to the No Interference assumption. If the treatment is some kind of training offered at random to some people in a community, treated people may share some or all of their new knowledge with the people in the control group. In this case, our treatment effect estimate will be biased toward 0, as the control group somewhat catches up with the treatment group.

It is impossible to test this assumption: Data scientists together with subject matter experts must make a judgment call.

CONSISTENCY

Consistency says that when $\mathbf{T_i = t}$, then $\mathbf{Y_i = Y_i(t)}$

In other words, unit **i**'s outcome under observed treatment **t** is exactly its observed outcome. This assumption means that we are not allowing for unrecorded variation in treatment.

Businesses should make sure not to introduce unrecorded variability in treatment. This could occur when some customers either received discount A or discount B lumped under a single treatment indicator in the data. In that case, it is possible that for some **i**, **Yi(A)** differs from **Yi (B)**, meaning that for some versions of treatment the assumption breaks. Lack of Consistency poses a threat to identifiability.

Biases or Why Assumptions Matter

To drive the point home on identifiability, let us explore a business use case and show that relying on the ML toolbox without heeding causal assumptions may have catastrophic consequences. To see this, let's consider a clothing company that has multiple stores across the U.S. As part of a marketing campaign, the company gives each shop manager a quota of discount vouchers that they can give to their affiliated customers.

A few months later, once the campaign is over and purchasing behaviors are recorded, the company's analytics team collects the campaign data to assess the effect of vouchers on the company's revenue. Their ultimate goal is to learn from that campaign to design the next best possible voucher campaign.

The analysts start by looking at the difference in the average revenue (**Y_i**) from the two groups of customers — those who received a voucher (**Ti=1**) and those who did not (**T_i=0**). In other words, the analytics team directly estimates the statistical estimand of the ATE:

$$\mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0]$$

Using a difference in means:

$$\frac{\sum_i Y_i T_i}{\sum_i T_i} - \frac{\sum_i Y_i(1 - T_i)}{\sum_i (1 - T_i)}$$

Next, the team might want to know the effect of treatment at a more granular level. That is, given customer characteristics **x** (e.g., age, purchase history etc), how beneficial was giving a voucher to them? This amounts to estimating the CATE at **x**. Knowing the CATE is necessary to design more targeted campaigns down the road. One way to go about this would be to train an ML model to predict **Yi** as a function of **Xi** and **Ti** effectively creating an estimator for the conditional expectation **𝔼[Yi|Xi,Ti]**. Let's call that estimator **m(Yi(Xi, Ti))**. The CATE at **x** is given by:

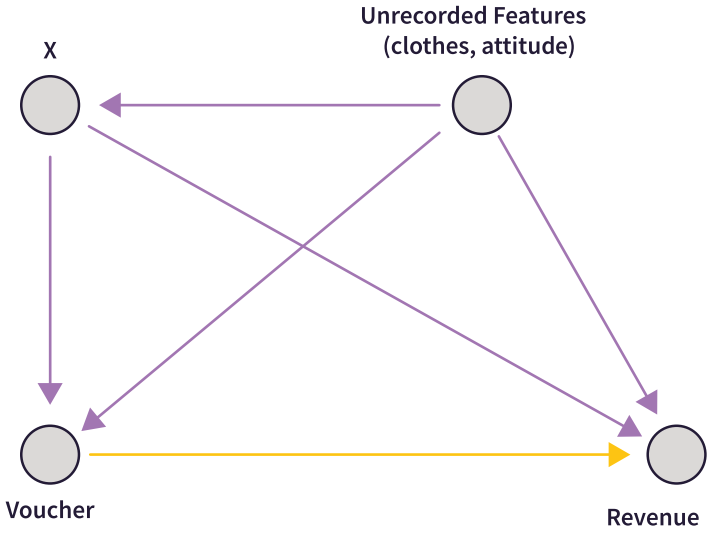
$$m(Y_i(X_i = x, T_i = 1)) - m(Y_i(X_i = x, T_i = 0))$$

DOES THE ANALYTICS TEAM'S METHOD CORRECTLY ESTIMATE THE VOUCHER EFFECT?

The estimation strategies to measure $\mathbb{E}\mathbf{[Yi|Ti]}$ and $\mathbb{E}\mathbf{[Yi|Xi,Ti]}$ in the ATE and CATE are not wrong. The problem resides in the assumption that the analysts implicitly made at the identification level and which is unlikely to be satisfied. The assumption is that the statistical estimands $\mathbb{E}\mathbf{[Yi|Ti]}$ and $\mathbb{E}\mathbf{[Yi|Xi,Ti]}$ identify the causal estimands $\mathbb{E}[\mathbf{Yi(Ti)]}$ in the ATE and $\mathbb{E}[\mathbf{Yi(Ti)|Xi]}$ in the CATE, respectively.

To see why this is not true, remember that store managers had the discretion over which customer to give a voucher to. Based on factors that were not recorded in the data, managers could have chosen to give the voucher only to customers that would bring more sales to their stores. For example, managers or their staff could have had a chat with the customer at the register before deciding to give them a voucher. The way the customer dresses probably didn't go unnoticed as an indicator of socioeconomic status and could have driven the decision to give them the voucher.

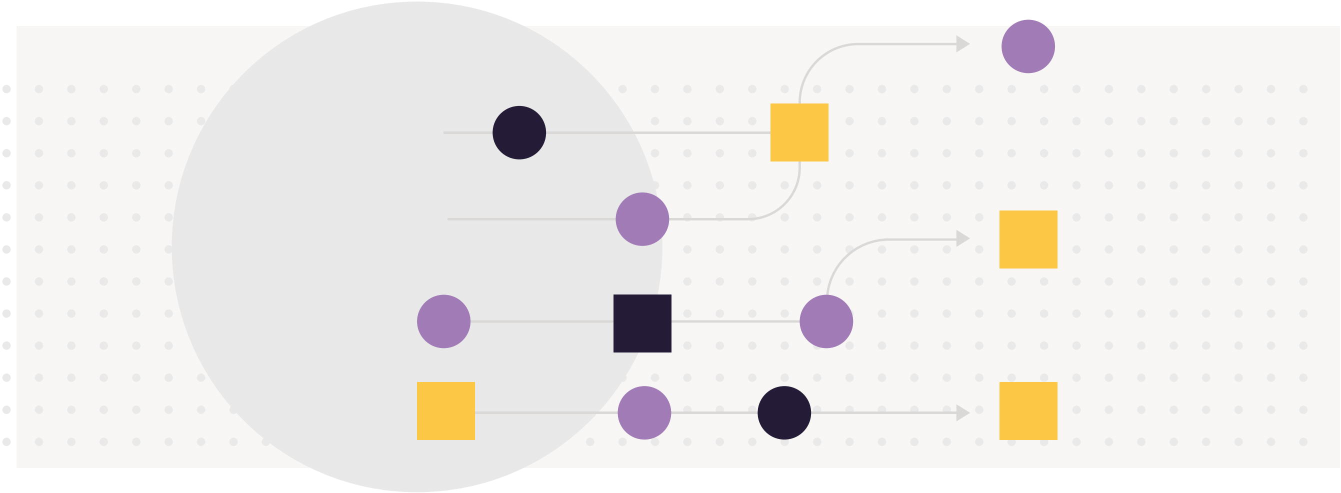
The vouchers were given to whoever was thought to generate the most revenue with them. This means that the causal effect estimate of the voucher on a customer characterized by **x** is inflated. In statistical parlance, the estimate is said to be *biased*. This upward bias could mislead future marketing decisions. For example, if the company decided to systematically offer discount vouchers to those **x** customers, they might realize that they're missing their revenue target in the long run.



The above graph shows that the unrecorded features are a common cause of both the outcome variable, *Revenue*, and the treatment variable, *Voucher*. The conditional ignorability assumption does not hold, which is a likely outcome when using only observational data.

The best way to evaluate the effect vouchers would be to give them at random, thereby severing any relation between unobserved variables and treatment assignment.

Causal Models



While traditional ML models are trained to predict an observed outcome **Yi** as a function of **Xi**, causal ML models are trained to predict an unobserved causal effect of **Ti** on **Yi** as a function of **Xi**: causal ML models return CATE predictions.

There are two main families of causal models or estimators:

Indirect Estimators combine one or more regular ML models trained to predict the observed outcome (**Yi**) as a function of **Xi** and **Ti** . The predicted CATE in unit i is computed by subtracting the predicted outcome without treatment (**Ti=0**) from the predicted outcome with treatment (**Ti=1**).

In the previous section, the analytics team trained an ML model, **m(Yi(Xi, Ti))**, to predict **Yi** as a function of Xi and Ti. They computed the predicted CATE as the difference between two predicted outcomes, with and without treatment. This causal model is known as an **S-Learner** (S as in single) and is one of the most intuitive Indirect Estimators.

Direct Estimators model the effect of the treatment directly without the need for predicted intermediary observed outcome models. This is achieved by either modifying the observed outcome variable (**Yi**) or by modifying ML algorithms (e.g., using a different splitting criterion in a decision tree).

Causal Metrics

In supervised ML, the performance of a model is typically assessed on a held-out dataset by comparing the model predictions to some observed ground truth. Metrics such as RMS2 or MAPE for regression problems or ROC AUC for classification problems are typically used.

CATE predictions are not straightforward to evaluate as we are never observing the true causal effect in any unit. Remember that we can never observe counterfactuals (the Fundamental Problem of CI strikes again). As a result, causal metrics will only be good for ranking different causal models — they do not provide any objective measure for the quality of the CATE predictions. This is to be contrasted with supervised ML predictions where an RMSE very close to 0 means that the model predictions are perfect (barring any leakage issue).

Ultimately, all causal metrics try to evaluate the quality of a causal model by asking the following question: **Is the difference in expected outcome between treated and control groups greater for observations whose predicted CATE is larger?**

Two main metrics are available: The QINI Coefficient and its cousin the Area Under the Uplift Curve (AUUC).

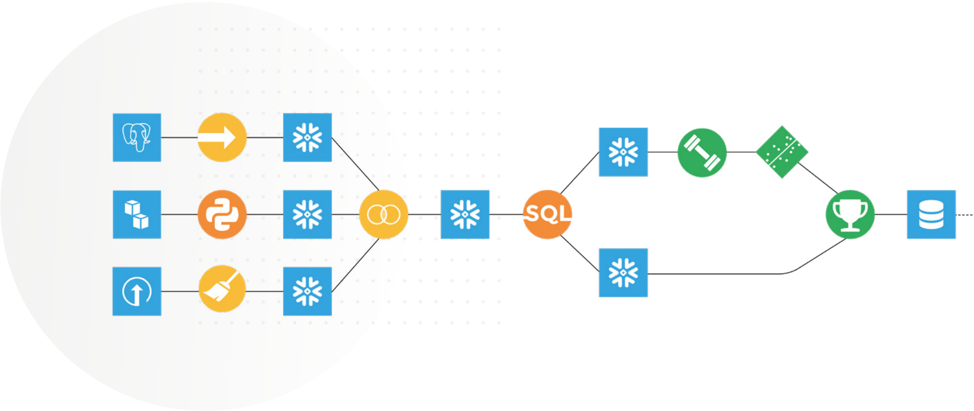
With Great Power Comes Great Responsibility

CI is a powerful tool for addressing powerful questions. Business owners can use CI to predict how a given customer would react to an action or alternative scenario. If you made it this far, we don't think we would need to further convince you that this kind of knowledge is invaluable.

Yet, at the risk of sounding like a broken record, we must once again caution that CI is only as good as its mostly untestable assumptions. Unlike traditional ML, where competent data scientists can relatively easily make sure predictions are accurate, CI is a total shot in the dark if we've the slightest doubt about the validity of the assumptions.

When using observational data, building a graph often comes as the best way of seeing the different causal or correlational relations between variables (observed or unobserved). This task requires strong expert knowledge and can quickly become overwhelming. In most cases, we would argue that it's difficult to be 100% confident that we are not omitting one or more confounders of the treatment and outcome variable. We're risking ending up with biased causal predictions. As explained earlier, nothing beats a well-designed experiment.

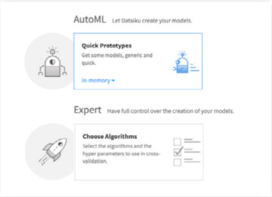
Everyday AI, Extraordinary People



Elastic Architecture Built for the Cloud

Google Cloud · snowflake · Microsoft Azure · aws

Machine Learning

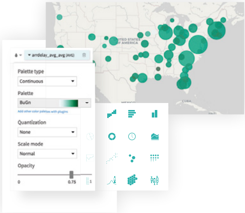

AutoML Let Dataiku create your models.

Quick Prototypes
Get some models, generic and quick.
In-memory ▾

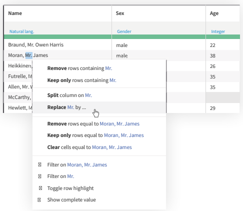
Expert Have full control over the creation of your models.

Choose Algorithms
Select the algorithms and the hyper parameters to use in cross-validation.

Visualization



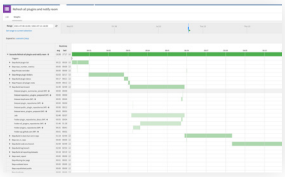
Data Preparation



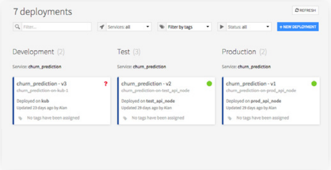
Name	Sex	Age
Natural lang.	Gender	Integer
Braund, Mr. Owen Harris	male	22
Moran, Mr. James	male	38
Heikkinen,		26
Futrelle, M		35
Allen, Mr. W		35
McCarthy,		
Hewlett, M		29

Remove rows containing Mr.
Keep only rows containing Mr.
Split column on Mr.
Replace Mr. by ...
Remove rows equal to Moran, Mr. James
Keep only rows equal to Moran, Mr. James
Clear cells equal to Moran, Mr. James
Filter on Moran, Mr. James
Filter on Mr.
Toggle row highlight
Show complete value

DataOps



Governance & MLOps


7 deployments

Development (2)	Test (3)	Production (2)
Service: churn_prediction	Service: churn_prediction	Service: churn_prediction
churn_prediction - v3 churn_prediction-on-kub-1 Deployed on kub Updated 23 days ago by Alan No tags have been assigned	churn_prediction - v2 churn_prediction-on-test_api_node Deployed on test_api_node Updated 29 days ago by Alan No tags have been assigned	churn_prediction - v1 churn_prediction-on-prod_api_node Deployed on prod_api_node Updated 29 days ago by Alan No tags have been assigned

Applications

450+
CUSTOMERS

45,000+
ACTIVE USERS

Dataiku is the world's leading platform for Everyday AI, systemizing the use of data for exceptional business results. Organizations that use Dataiku elevate their people (whether technical and working in code or on the business side and low- or no-code) to extraordinary, arming them with the ability to make better day-to-day decisions with data.

data
iku