sbbd:03

Chapter 3

Causality for Data Scientists: Discovery, Causal Inference, and Applications in Machine Learning

Gustavo Ferreira Viegas de Oliveira[1], Fabrício Aguiar Silva[1], Marcus Henrique Soares Mendes[1]

[1]NESPeD Lab, Universidade Federal de Viçosa (UFV), Florestal, MG, Brasil

{gustavo.viegas, fabricio.asilva, marcus.mendes}@ufv.br

Abstract

Machine Learning (ML) models excel at identifying patterns in data, yet they are fundamentally limited in answering causal questions such as "What happens if I intervene?" or "Which features actually drive this outcome?" This tutorial chapter introduces the foundations of computational causality for data scientists and database researchers. Starting from Simpson Paradox and Pearl Causal Hierarchy, we develop Directed Acyclic Graphs (DAGs), the do-operator, d-separation, and the backdoor and frontdoor identification criteria. We then present the two complementary formalisms: Structural Causal Models (SCMs) and the Potential Outcomes framework, with emphasis on the estimands ATE and CATE. The chapter covers Causal Discovery algorithms for learning causal structures from observational data, and Causal Inference methods for effect estimation and validation. Finally, we connect these foundations to practical ML applications: causal graph-guided feature selection and counterfactual explainability. Practical exercises accompanying this chapter employ public datasets and open-source Python libraries.

3.1. Introduction

Machine Learning (ML) models have achieved impressive predictive performance across many domains, yet they carry a structural limitation that becomes critical in decision-making contexts: they learn statistical associations from observational data rather than causal mechanisms. A recommendation system that observes users buying rain boots after viewing umbrella listings learns a genuine correlation. But if it concludes that displaying more umbrellas will drive more boot purchases, it mistakes a shared cause (rainy weather) for a directed effect. Acting on such a spurious association allocates resources toward an intervention that will not achieve its intended result.

This limitation is not incidental and cannot be overcome simply by collecting more data or fitting more expressive models. As Pearl [33] argues, answering questions about interventions and counterfactuals requires qualitative assumptions about how the world generates data. These assumptions cannot be extracted from observations alone; they must be encoded in a causal model. Causality provides the formal language for expressing such assumptions, determining what can be identified from the available data, and translating statistical patterns into actionable knowledge.

3.1.1. Motivating Example: Simpson's Paradox

Consider a hospital comparing two treatments, A and B, for a condition that can present as either mild or severe. Aggregate records indicate that patients who received treatment A had an overall mortality rate of 16%, compared with 19% for those who received treatment B. This appears to support treatment A. However, examining each severity subgroup separately reverses the conclusion: treatment B achieves lower mortality among mild cases (10% versus 15%) and also among severe cases (20% versus 30%). This inversion is the Simpson's Paradox [45].

The resolution lies in the causal structure of the data. Treatment B is scarcer and tends to be allocated to patients in severe condition. Since severe condition independently raises mortality, condition severity acts as a *confounder*: a common cause of both treatment assignment and the outcome. The aggregate comparison conflates the causal effect of treatment with this confounding path. Stratifying by severity removes the confounding influence and reveals the true picture: treatment B is superior in every subgroup.

What makes this example instructive is that the correct conclusion depends entirely on the causal graph, not on the numbers alone. If severity causes treatment assignment, stratifying by severity is the appropriate analysis. If instead treatment assignment were to cause the observed severity (for example, if waiting for the scarcer treatment B allows the condition to worsen), the graph would change and so would the analysis [32]. Two datasets with identical marginals can require opposite decisions depending on the underlying causal mechanism. Statistics alone cannot distinguish these scenarios.

3.1.2. Why Causality Matters for Data Science and Databases

Operational databases and data warehouses record the outcomes of real-world processes governed by complex causal mechanisms: clinical interventions, marketing campaigns, pricing decisions, sensor-driven control systems. Practitioners routinely extract features from these records and train ML models to support future decisions. Without causal reasoning, three systematic problems arise.

1. Confounded feature associations: Features correlated with the outcome through confounding paths rather than causal paths produce models that perform well in cross-validation but fail to generalize when the confounding structure changes across time or deployment environment [41, 34].

2. Inability to reason about interventions: When a business analyst asks what will happen to customer retention if the price is reduced by 10%, the question is inherently interventional. It requires evaluating $P(Y \mid do(X))$, the distribution of retention under a forced price change, rather than $P(Y \mid X)$, the conditional distribution when price hap-

pens to take a certain value in historical records. These two quantities differ whenever confounders are present [32].

3. Opaque and misleading explanations: Attribution methods that distribute credit over correlated input features may assign importance to variables that are symptoms or correlates of the true cause rather than causes themselves. Counterfactual explanations, which identify the minimal input change that would alter a prediction, are by nature causal and produce insights that are more directly actionable for stakeholders [29].

Causal methods address all three problems by grounding the analysis in an explicit model of the data-generating process, providing formal criteria for when causal quantities are identifiable from observational data, and offering practical tools for estimation and validation.

3.1.3. Chapter Organization

The chapter is organized in seven main sections that build progressively from conceptual foundations to practical methods.

Section 3.2 establishes the conceptual groundwork: why statistical association is insufficient for causal conclusions, what it means to intervene in a system, and Pearl's three-level Causal Hierarchy that situates machine learning within the broader landscape of causal reasoning.

Section 3.3 develops the graphical language of causal models: Directed Acyclic Graphs, the structural roles of confounders, mediators, and colliders, d-separation, Markov equivalence classes, and the backdoor and frontdoor identification criteria.

Section 3.4 introduces the two formal frameworks that give mathematical content to the graphical intuitions: Structural Causal Models, with structural equations and the graph-surgery interpretation of interventions, and the Potential Outcomes framework, with the identification assumptions that enable estimation from observational data.

Section 3.5 presents algorithms for learning causal structure from data, covering constraint-based, score-based, and functional approaches, and the practical considerations that guide algorithm selection.

Section 3.6 addresses the estimation of causal effects: identification via the do-calculus, estimation strategies from regression adjustment to propensity score methods and meta-learners, and the validation of causal conclusions through systematic refutation tests.

Section 3.7 connects causal foundations to practical machine learning, including graph-guided feature selection and engineering for robust models, and counterfactual-based explanations for interpretable predictions.

Section 3.8 surveys the open-source ecosystem of causal analysis software and benchmark datasets, providing a practical starting point for applying the methods covered in this chapter.

3.1.4. Code and Practical Resources

The practical materials accompanying this chapter are available at the following repository: `https://github.com/gfviegas/causality-for-data-scientists-course`. There, readers will find a collection of Jupyter notebooks providing step-by-step implementations of the principal methods discussed in each section, using publicly available datasets including the Sachs protein signaling network [39]. The notebooks use modern tools such as `causal-learn` [55] for causal discovery and `DoWhy` [43] for causal inference and refutation, and are designed to run in Google Colab without requiring local installation. In addition to the primary exercises, the repository includes curated pointers to external implementations, supplementary datasets, and further reading for practitioners who wish to explore individual topics in greater depth.

3.2. Causality and the Causal Hierarchy

Reasoning causally requires more than observing patterns in data. This section establishes the conceptual groundwork: why statistical association is insufficient for causal conclusions, what it means to intervene in a system rather than merely observe it, and how Pearl's Causal Hierarchy situates these questions within a unified framework. The formal graphical and algebraic tools for expressing and computing causal relationships are developed in the sections that follow.

3.2.1. Causation vs. Correlation

A fundamental principle in causal reasoning is that statistical association does not imply causation. Two variables may be strongly associated without one causing the other. It is worth noting a terminological precision: *correlation* technically refers only to linear statistical dependence, whereas *association* is the broader concept covering any form of statistical dependence. We use the term association throughout, as it is the more general concept relevant to causal reasoning [32].

Associations arise from two qualitatively different sources: *causal association*, which flows along directed causal paths, and *confounding association*, which flows along paths that share a common cause. The ice cream and drowning example is a classic illustration: sales of ice cream and the number of drowning incidents are positively associated, yet neither causes the other. Both are driven by a shared cause, warm summer weather. A policy intervention based on this association, such as restricting ice cream sales, would have no effect on drowning rates. This is the fundamental danger of acting on associations without understanding their causal origin.

A Real-World Case: Vaccine Effectiveness and Simpson's Paradox. Simpson's Paradox [45] occurs when a trend visible in aggregate data reverses when the data are stratified by a subgroup variable. It is not merely a theoretical curiosity: it appears routinely in public health, economics, and clinical research whenever a confounding variable is overlooked.

Table 3.1 presents COVID-19 Delta variant case data from the United Kingdom.[1]

[1]Delta variant COVID-19 surveillance data from Public Health England, compiled by OpenIntro: `https://www.openintro.org/data/index.php?data=simpsons_paradox_covid`.

At first glance, the aggregate numbers are startling: vaccinated individuals show a death rate of 0.41%, more than twice the 0.17% rate among unvaccinated individuals. A naive reading would conclude that vaccination increases mortality risk, a conclusion that fueled genuine misinformation during the pandemic.

Table 3.1. COVID-19 Delta variant death rates by vaccination status and age group (UK). The aggregate comparison reverses when the data are stratified by age, the confounding variable.

Age Group	Vaccination Status	Death Rate (%)
Aggregate	Unvaccinated	0.167
Aggregate	Vaccinated	0.411
Under 50	Unvaccinated	0.033
Under 50	Vaccinated	0.023
50 and over	Unvaccinated	5.959
50 and over	Vaccinated	1.684

Stratifying by age group completely reverses the picture. Among individuals under 50, the vaccinated death rate (0.023%) is lower than the unvaccinated rate (0.033%). Among those aged 50 and over, the vaccinated death rate (1.684%) is dramatically lower than the unvaccinated rate (5.959%). The vaccine is protective in every subgroup.

The paradox arises because age is a *confounder*: it affects both vaccination uptake and mortality risk simultaneously. During the Delta wave, older individuals were prioritized for vaccination, so the vaccinated group was disproportionately drawn from the high-risk age stratum. This inflates the overall death rate of the vaccinated group without reflecting any harmful effect of the vaccine. Figure 3.1 illustrates this causal structure as a Directed Acyclic Graph (DAG, formally introduced in Section 3.3).

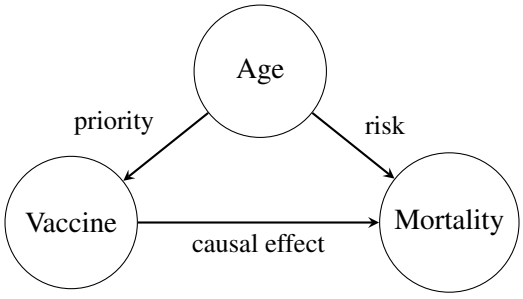


Figure 3.1. Causal graph for the COVID-19 vaccine example.

The aggregate statistic $P(\text{death} \mid \text{vaccinated}) > P(\text{death} \mid \text{unvaccinated})$ captures all association flowing between vaccine status and mortality, including the confounding path through age. The causal effect of vaccination, $P(\text{death} \mid do(\text{vaccinated}))$, isolates only the directed arrow from vaccine to mortality by conceptually removing the influence of the age-driven allocation. This distinction is the core problem that causal reasoning is designed to address: the same data can yield opposite conclusions depending on whether the analysis is associational or causal.

3.2.2. Interventions and the Do-Calculus

The *do-calculus* of Pearl provides a formal framework for reasoning about interventions [32]. The core distinction is between observational and interventional distributions: $P(Y|X)$ represents the probability of Y given that we observe X, while $P(Y|do(X))$ represents the probability of Y if we intervene to set X to a specific value.

In observational data, $P(Y|X)$ reflects correlations that may result from multiple causal pathways, including confounding. In contrast, $P(Y|do(X))$ isolates the causal effect of X on Y by simulating an intervention that eliminates incoming causal influences to X [32]. For example, $P(\text{recovery}|\text{took medicine})$ indicates correlation, potentially confounded by disease severity, while $P(\text{recovery}|do(\text{took medicine}))$ represents the actual causal effect. Crucially, this distinction cannot be resolved by collecting more data: answering interventional questions requires causal assumptions that go beyond the observational distribution.

3.2.3. The Causal Hierarchy

Pearl [33] formalizes the different types of causal questions through the *Three Layer Causal Hierarchy*, also known as the *Ladder of Causation*, which categorizes causal queries into three distinct levels of increasing complexity, as illustrated in Figure 3.2. The first level, *Association*, corresponds to observing and finding correlations in data, captured by conditional probabilities $P(Y|X)$. This is the domain where traditional machine learning excels. The second level, *Intervention*, involves predicting the effects of deliberate actions, expressed through the do-operator as $P(Y|do(X))$. This level requires understanding causal mechanisms beyond mere correlations. The third level, *Counterfactuals*, addresses hypothetical scenarios about what would have happened under different circumstances, enabling retrospective reasoning and explanation.

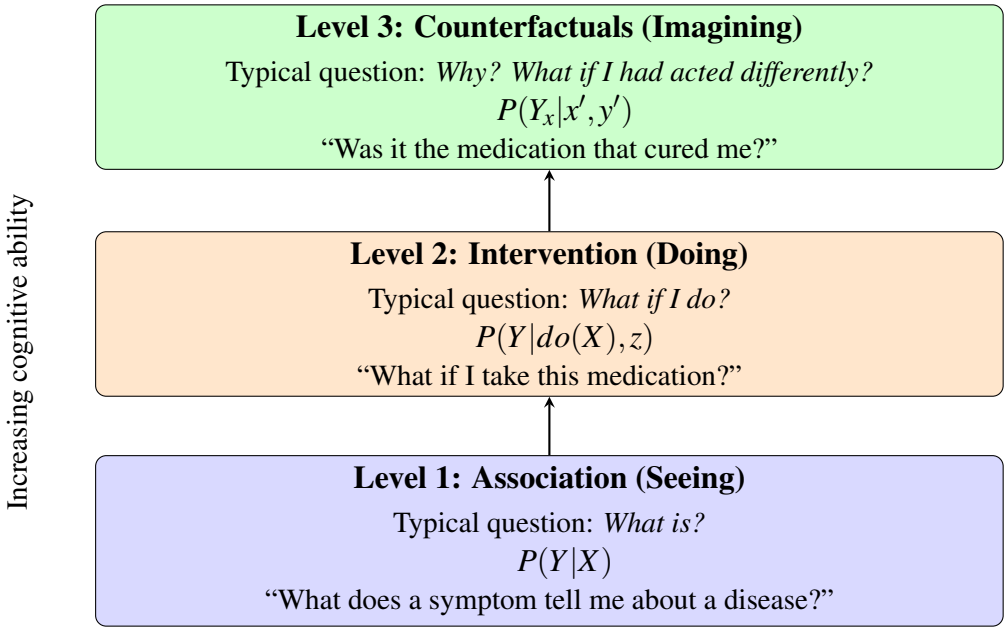


Figure 3.2. The Three Layer Causal Hierarchy (Ladder of Causation) [33].

A key insight from this hierarchy is that data alone, regardless of quantity, cannot answer questions at higher levels without additional causal assumptions [33]. Traditional ML operates almost exclusively at Level 1: given enough data, it learns predictive associations with high accuracy. Yet a model trained to predict hospital readmission from electronic health records cannot, without further assumptions, tell a clinician what will happen if a patient is prescribed a new drug. Answering that question requires Level 2 reasoning. Explaining why a particular patient experienced a certain outcome, or what would have changed had the treatment been different, requires Level 3. The frameworks and algorithms introduced in the remainder of this chapter provide the tools for reasoning at all three levels.

Causal models overcome fundamental limitations of correlation-based approaches: they remain robust to distribution shifts when causal mechanisms are stable [41, 34], support interventional reasoning to predict the effects of actions [32], and provide interpretable explanations consistent with human reasoning [29]. These attributes are critical in high-stakes domains such as healthcare, economics, and policy evaluation [20, 18].

3.3. Graphical Causal Models

Graphical models provide a visual and mathematical language for representing causal relationships and reasoning about conditional independence. This section introduces the key concepts essential for causal discovery and inference.

3.3.1. Directed Acyclic Graphs

A Directed Acyclic Graph (DAG) is a graph consisting of nodes connected by directed edges with no directed cycles [32]. In causal modeling, nodes represent variables and directed edges represent direct causal relationships. For an edge $X \to Y$, X is a *parent* of Y, and Y is a *child* of X. The *ancestors* of a node are all nodes from which a directed path leads to it; *descendants* are nodes reachable via directed paths.

In causal DAGs, an edge $X \to Y$ indicates that X is a direct cause of Y. This representation encodes the *Causal Markov Condition*: each variable is independent of its non-descendants conditional on its parents [46]. This condition allows the joint probability distribution P over all variables to be factorized as a product of conditional distributions, as shown in Equation 1, where $P(X_i|\text{Parents}(X_i))$ denotes the conditional probability of variable X_i given the values of its parent nodes in the graph. This factorization captures how each variable depends only on its direct causes, enabling efficient representation and computation of complex multivariate distributions.

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \tag{1}$$

The *faithfulness assumption* states that the only conditional independencies in the distribution are those entailed by the Markov condition, with no additional independencies from parameter cancellations [46]. This assumption is crucial for causal discovery algorithms that infer structure from conditional independence tests.

The *Markov blanket* of a node X is the minimal set of nodes that renders X inde-

pendent of all others when conditioned upon. In a DAG, it consists of the parents of X, the children of X, and the other parents of those children [46].

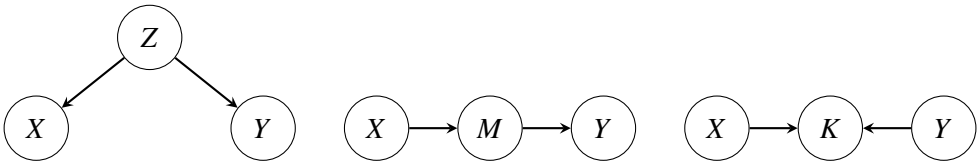
3.3.2. Confounders, Mediators, and Colliders

The role a variable plays in a causal graph determines how it affects inference. Three roles are particularly important, each corresponding to one of the fundamental path structures in a DAG.

A **confounder** is a common cause of both the treatment X and the outcome Y, represented by a fork ($X \leftarrow Z \rightarrow Y$). Because Z drives both variables, it creates a spurious association between X and Y that does not reflect a direct causal effect. The vaccine-age-mortality graph introduced in Section 3.2 is a canonical example: age confounds the apparent relationship between vaccination and mortality.

A **mediator** lies on a directed causal path from X to Y, represented by a chain ($X \rightarrow M \rightarrow Y$). The mediator M transmits part or all of the causal effect of X on the outcome. Conditioning on a mediator blocks the indirect pathway through it, which is appropriate when estimating the *direct* effect of X on Y but inappropriate when estimating the *total* effect.

A **collider** is a variable caused by two others, represented by a v-structure ($X \rightarrow K \leftarrow Y$). Unlike confounders and mediators, a collider does not transmit association between X and Y by default: without conditioning, the two parents are independent along this path. However, conditioning on a collider (or on any of its descendants) opens the path and creates a spurious dependence between X and Y. This counterintuitive behavior is the mechanism behind selection bias and the Berkson paradox.



(a) Confounder $X \leftarrow Z \rightarrow Y$ **(b) Mediator** $X \rightarrow M \rightarrow Y$ **(c) Collider** $X \rightarrow K \leftarrow Y$

Figure 3.3. The three structural roles of a variable in a causal DAG.

Figure 3.3 illustrates the three roles. The distinction between them has direct practical consequences: properly adjusting for a confounder removes bias, conditioning on a mediator can block the effect being estimated, and conditioning on a collider introduces bias where none existed. The d-separation criterion, introduced next, provides the formal machinery for determining which paths are active or blocked given any set of conditioned variables.

3.3.3. d-Separation

Directional Separation (d-Separation) is a graphical criterion for determining conditional independence [32]. Two sets of nodes X and Y are d-separated by a set Z if all paths between them are blocked by Z. Whether a path is blocked depends on the structural role

of each intermediate node W, determined by the direction of arrows on either side of it. Figure 3.4 illustrates the three fundamental path structures.

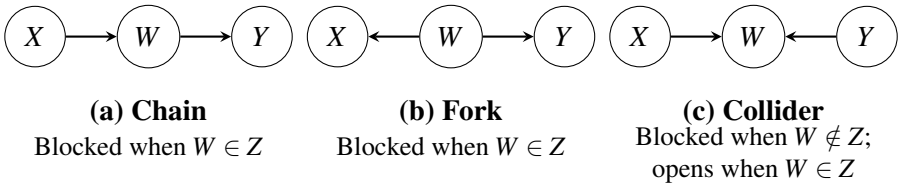


Figure 3.4. The three fundamental path structures in a causal DAG.

In a **chain** or a **fork**, the path between X and Y is open when $W \notin Z$ and blocked when $W \in Z$. The two structures behave identically with respect to blocking because both involve W as a non-collider on the path. In a **collider**, the logic inverts: the path is blocked by default when neither W nor any descendant of W is in Z; conditioning on W opens the path, creating an association between X and Y that was not previously present.

The collider structure is particularly consequential: it is the mechanism behind selection bias, the Berkson paradox, and M-bias in observational studies [32]. A well-known example is the observation that among hospitalized patients (who are conditioned on by the act of sampling), two unrelated conditions appear negatively correlated simply because patients with neither condition are less likely to be hospitalized. d-Separation is foundational to constraint-based causal discovery algorithms, which test conditional independencies to infer causal structure [46].

3.3.4. Markov Equivalence

Multiple DAGs can encode the same conditional independence relationships. Such DAGs are *Markov equivalent* [46]. Figure 3.5 shows the canonical three-node example: the chain ($X \to Y \to Z$), the reversed chain ($X \leftarrow Y \leftarrow Z$), and the fork ($X \leftarrow Y \to Z$) all encode that $X \perp Z \mid Y$ and are therefore observationally indistinguishable. The v-structure ($X \to Y \leftarrow Z$), by contrast, encodes a different independence structure ($X \perp Z$ marginally, but $X \not\perp Z \mid Y$) and belongs to a distinct equivalence class. This means that from observational data alone, causal discovery algorithms can only identify an equivalence class of graphs, not a unique DAG.

A Markov equivalence class is represented by a *Completed Partially Directed Acyclic Graph (CPDAG)*, containing directed edges for orientations shared by all equivalent DAGs and undirected edges for those that vary [46]. When hidden confounders are possible, more general representations are needed: *Maximal Ancestral Graphs (MAGs)* and *Partial Ancestral Graphs (PAGs)* represent equivalence classes under latent variable models [53].

3.3.5. Paths and Adjustment

A *causal path* from X to Y follows edge directions; the total causal effect is transmitted through all such paths. A *backdoor path* starts with an edge into X (pointing toward X) and represents confounding, providing alternative explanations for the X-Y association that do not reflect causation [32].

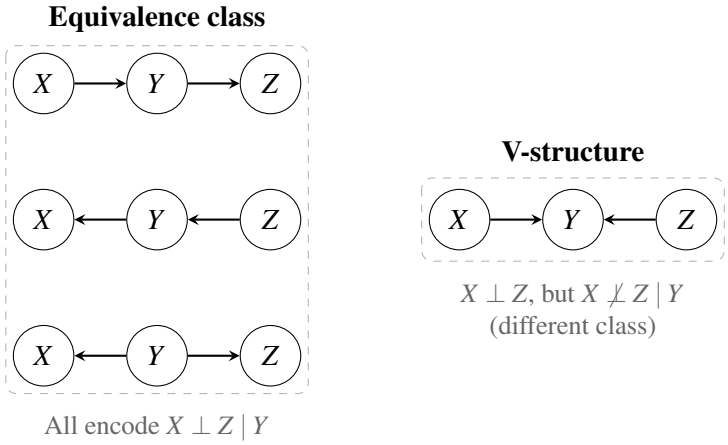


Figure 3.5. Markov equivalence classes in three-node DAGs.

The Backdoor Criterion. A set of variables Z satisfies the backdoor criterion relative to a treatment X and outcome Y if: (1) no node in Z is a descendant of X, and (2) Z blocks every backdoor path from X to Y. When such a set exists, the causal effect of X on Y is identifiable and can be computed via the backdoor adjustment formula, as Equation 2 shows.

$$P(Y|do(X)) = \sum_{z} P(Y|X,Z=z) \cdot P(Z=z) \tag{2}$$

This formula adjusts for confounding by stratifying on Z: for each value of the confounders, it computes the conditional probability of Y given X, then averages over the distribution of Z. The key insight is that conditioning on a valid adjustment set Z blocks all non-causal paths while leaving causal paths open [32].

The Frontdoor Criterion. When no valid backdoor adjustment set exists (because all confounders are unmeasured), the frontdoor criterion offers an alternative identification strategy. A set of variables M satisfies the frontdoor criterion relative to X and Y if: (1) M intercepts all directed paths from X to Y, (2) there is no unblocked backdoor path from X to M, and (3) all backdoor paths from M to Y are blocked by X. When these conditions hold, the causal effect is as shown in Equation 3.

$$P(Y|do(X)) = \sum_{m} P(M=m|X) \sum_{x'} P(Y|M=m,X=x') \cdot P(X=x') \tag{3}$$

The frontdoor formula works by decomposing the causal effect into two steps: the effect of X on the mediator M, and the effect of M on Y (adjusting for X as a confounder of the $M \rightarrow Y$ relationship). This criterion is particularly useful when a mediator M fully transmits the effect of X on Y [32].

3.4. Formal Causal Frameworks

The graphical language introduced in Section 3.3 provides a visual and structural representation of causal relationships. Two complementary algebraic frameworks translate

these graphs into mathematical objects suitable for precise reasoning and estimation: Structural Causal Models (SCMs), which specify the generative mechanism of the data, and the Potential Outcomes framework, which defines the causal quantities practitioners seek to estimate. Understanding both is essential for moving from graphical intuitions to formal identification and inference.

3.4.1. Structural Causal Models

Structural Causal Models (SCMs), formalized by Pearl [32], represent a data-generating process through a set of structural equations and an associated DAG. An SCM $\mathscr{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$ consists of:

- **Exogenous variables U**: background factors not caused by any other variable in the model, representing unmeasured noise.
- **Endogenous variables V**: observed variables, each assigned a structural equation.
- **Structural equations F**: a function $V_i := f_i(\text{Pa}(V_i), U_i)$ for each $V_i \in \mathbf{V}$, where $\text{Pa}(V_i)$ are the parents of V_i in the associated DAG and U_i is a noise term.

The asymmetric assignment operator $:=$ (rather than $=$) emphasizes that the equation is a generative mechanism: changing the right-hand side variables changes V_i, but not vice versa.

Concrete example. Consider the COVID-19 vaccine scenario from Section 3.2. A simple SCM over three variables might be:

$$A := U_A, \tag{4}$$

$$V := \mathbb{1}[A > 50] \cdot U_V^{(1)} + \mathbb{1}[A \leq 50] \cdot U_V^{(2)}, \tag{5}$$

$$M := \alpha V + \beta A + U_M. \tag{6}$$

Here A (age) is exogenous; V (vaccination) depends on age and random noise; M (mortality) depends on both V and A, with noise U_M. The associated DAG has edges $A \to V$, $A \to M$, and $V \to M$, exactly as shown in Figure 3.1.

Interventions as graph surgery. The key operation in SCMs is the *do*-intervention. Setting $do(V = 1)$ removes all incoming edges to V and replaces Equation 5 with the constant $V := 1$. The resulting *mutilated model* $\mathscr{M}_{do(V=1)}$ has only two remaining equations: $A := U_A$ and $M := \alpha \cdot 1 + \beta A + U_M$. In this mutilated model A no longer influences V (since there is no equation for V containing A), so the confounding path $V \leftarrow A \to M$ is severed. Computing the distribution of M in the mutilated model gives $P(M \mid do(V = 1))$, the true causal effect of universal vaccination, uncontaminated by age-based selection.

This graph-surgery interpretation provides the semantic foundation for the backdoor and frontdoor criteria: those criteria identify which observational quantities equal the interventional distribution without needing to explicitly mutilate the graph.

3.4.2. The Potential Outcomes Framework

The Potential Outcomes Framework, also known as the Rubin Causal Model [38], approaches causality from the perspective of counterfactual comparisons. For each unit i

and treatment value t, there exists a *potential outcome* $Y_i(t)$: the value of the outcome that unit i would exhibit if assigned treatment t, regardless of the treatment they actually received.

The individual causal effect for unit i is $\tau_i = Y_i(1) - Y_i(0)$: the difference between the outcome under treatment and the outcome under control. In the vaccine example, $Y_i(1)$ is the mortality indicator for individual i if vaccinated, and $Y_i(0)$ is the same indicator if unvaccinated. The key causal estimands are:

- **Average Treatment Effect (ATE)**: $\tau = \mathbb{E}[Y(1) - Y(0)]$, the average effect across the full population.
- **Average Treatment Effect on the Treated (ATT)**: $\mathbb{E}[Y(1) - Y(0) \mid T = 1]$, the average effect among treated units.
- **Conditional Average Treatment Effect (CATE)**: $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, the effect conditional on covariates, enabling personalized estimates.

The two frameworks are complementary: SCMs provide the graphical language for encoding and reasoning about causal structure, while the Potential Outcomes framework excels at defining estimands and guiding estimation strategies. Modern tools such as DoWhy [43] integrate both perspectives within a single pipeline.

3.4.3. The Fundamental Problem of Causal Inference

A central challenge in causal reasoning is what Holland [19] calls the *Fundamental Problem of Causal Inference*: for any individual unit, only one potential outcome is ever observed. If a patient receives treatment ($T = 1$), one observes $Y_i(1)$ but never $Y_i(0)$, the outcome that would have occurred under control. Conversely, if the patient receives control ($T = 0$), one observes $Y_i(0)$ but never $Y_i(1)$. The individual causal effect $\tau_i = Y_i(1) - Y_i(0)$ is therefore fundamentally unidentifiable for any single unit from observational data alone.

This missing data perspective has important consequences. Because the individual causal effect cannot be directly estimated, causal inference typically targets population-level estimands such as the ATE. Population-level identification becomes possible when the treatment assignment mechanism satisfies sufficient conditions. The most important is *unconfoundedness* (also called conditional exchangeability or ignorability): treatment assignment is independent of potential outcomes conditional on observed covariates, $Y(0), Y(1) \perp T \mid X$ [38, 20]. Intuitively, within any stratum defined by X, treated and untreated units are comparable in expectation.

Two further conditions are required. *Positivity* (or overlap) requires that every covariate stratum has a nonzero probability of receiving either treatment value: $0 < P(T = 1 \mid X = x) < 1$ for all x with $P(X = x) > 0$. Without positivity, some subgroups are never observed under one treatment, making comparison impossible [20]. *Consistency* requires that the observed outcome for a unit equals its potential outcome under the received treatment: $T = t \Rightarrow Y = Y(t)$. Combined with the *no interference* assumption (each unit's outcome depends only on its own treatment), consistency constitutes the Stable Unit Treatment Value Assumption (SUTVA) [19].

Under unconfoundedness, positivity, and SUTVA, the ATE is identifiable from observational data through the adjustment formula:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X\big[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\big]. \tag{7}$$

This result illustrates the two-step process that governs the rest of the chapter: *identification* translates a causal estimand into a statistical estimand expressible in terms of the observed distribution; *estimation* then computes a numerical value from finite data. The causal graph provides the formal basis for verifying whether identification assumptions such as unconfoundedness are plausible in a given application, connecting the SCM perspective directly to the Potential Outcomes framework.

3.5. Causal Discovery

Causal Discovery (CD) aims to learn causal structures from data without relying solely on domain knowledge [46]. Given observational data, the goal is to infer a causal graph (or equivalence class) that represents the underlying data-generating process. This section introduces the problem, key assumptions, and main algorithmic approaches.

3.5.1. The Causal Discovery Problem

The Causal Discovery problem can be stated as follows: given a dataset of observations over variables $\mathbf{V} = \{V_1, \ldots, V_n\}$, infer the causal DAG G that generated the data. Due to Markov equivalence (subsection 3.3.4), observational data alone typically identify only an equivalence class represented by a CPDAG, not a unique DAG [46].

The problem is challenging for several reasons: the search space of possible graphs grows super-exponentially with the number of variables; finite samples provide imperfect estimates of conditional independencies; and the presence of hidden confounders can make the true structure fundamentally unidentifiable from observational data alone [46, 16].

3.5.2. Key Assumptions

Causal Discovery algorithms rely on assumptions that enable structure learning from statistical patterns:

Causal Markov Condition. Each variable is independent of its non-descendants given its parents. This links the graph structure to the probability distribution, allowing algorithms to use conditional independence tests as constraints on possible structures [46].

Faithfulness. All conditional independencies in the data are entailed by the Markov condition applied to the true graph. This rules out "unfaithful" distributions in which independence arises from parameter cancellations rather than from graphical structure. Without faithfulness, multiple non-equivalent graphs could produce identical independence patterns [46].

Causal Sufficiency. All common causes of measured variables are also measured, meaning there are no hidden confounders. When this assumption is violated, algorithms such as FCI that output PAGs rather than CPDAGs are required [53].

3.5.3. Algorithmic Approaches

Causal Discovery algorithms fall into three main families and hybrid approaches, each exploiting different properties of the data:

Constraint-Based Methods use conditional independence tests to eliminate edges and orient them based on patterns of dependence [46]. The PC algorithm [23] is the foundational approach: it starts with a complete graph, removes edges between conditionally independent variables, and applies orientation rules to identify v-structures and propagate orientations. Variants include FCI for settings with hidden confounders [53]. These methods are statistically consistent under the Markov and faithfulness assumptions but sensitive to errors in independence tests.

Score-Based Methods search for the graph that maximizes a scoring function balancing fit and complexity [9]. The Greedy Equivalence Search (GES) algorithm operates directly on equivalence classes, using forward and backward phases to add and remove edges while optimizing a score such as Bayesian Information Criterion (BIC). Score-based methods avoid multiple testing issues but face computational challenges in large search spaces.

Functional Methods exploit asymmetries in the functional form of causal relationships. Linear Non-Gaussian Acyclic Model (LiNGAM) [44] assumes linear relationships with non-Gaussian noise, using independent component analysis to identify the unique causal ordering. These methods can identify the full DAG (not just the equivalence class) when their functional assumptions hold, but they are sensitive to assumption violations.

Hybrid Methods combine elements of the above approaches. For example, some algorithms use constraint-based methods to restrict the search space before applying score-based optimization [16].

Table 3.2 presents a representative, though not exhaustive, selection of CD algorithms from each family, including their supported data types (C = Continuous, D = Discrete/Categorical) and whether they assume Gaussian-distributed data. The field of CD is an active area of research, with new algorithms being proposed regularly [16]. Each algorithm embodies different theoretical assumptions and computational trade-offs, making algorithm selection inherently problem-dependent.

3.5.4. Practical Considerations

Selecting a CD algorithm depends on domain characteristics: data size and dimensionality, expected graph density, availability of interventional data, and whether hidden confounders are plausible. No single algorithm dominates across all settings [16]. In practice, researchers often apply multiple algorithms and seek consensus, or use domain knowledge to constrain the search space. The output of CD (typically a CPDAG or PAG) serves as input to Causal Inference methods that estimate the magnitude of causal effects.

3.6. Causal Inference

While Causal Discovery learns the structure of causal relationships, Causal Inference (CI) estimates the magnitude of causal effects [32]. Given a causal graph (known or

Table 3.2. Representative Causal Discovery algorithms.

Algorithm	Full Name	Supported Data Types	Gaussian Assumption	Family
PC [23]	Peter-Clark	C, D	No	Constraint
GS [27]	Grow-Shrink	C, D	No	Constraint
IAMB [48]	Incremental Assoc. Markov Blanket	C, D	No	Constraint
Fast-IAMB [50]	Fast Markov Blanket	C, D	No	Constraint
Inter-IAMB [50]	Interleaved Markov Blanket	C, D	No	Constraint
MMPC [49]	Max-Min Parents and Children	C, D	No	Constraint
GES [9]	Greedy Equivalence Search	C, D	Yes	Score
GIES [30]	Greedy Interv. Equivalence Search	C, D	Yes	Score
CCDr [1]	Concave penalized Coord. Descent	C	Yes	Score
BES [52]	Bayesian Equivalence Search	C, D	Yes	Score
CGNN [15]	Causal Generative Neural Networks	C	No	Score
SAM [22]	Structural Agnostic Model	C	No	Score
LiNGAM [44]	Linear Non-Gaussian Acyclic Model	C	No*	Functional
CAM [7]	Causal Additive Models	C	Yes	Functional
GRaSP [25]	Greedy Relax. of Sparsest Perm.	C, D	No	Hybrid

*LiNGAM assumes linear relationships with non-Gaussian noise.

discovered), causal inference answers questions such as: "What is the effect of treatment X on outcome Y?" This section covers identification, estimation, validation, and the current landscape of software tools and benchmark datasets.

3.6.1. Identification

A causal effect is *identifiable* if it can be uniquely determined from the observational distribution given the causal graph [32]. Identification precedes estimation: before computing a causal effect from data, one must establish that the effect is theoretically recoverable.

The *backdoor criterion* (subsection 3.3.5) is the most common identification strategy. When a valid adjustment set exists, the causal effect $P(Y|do(X))$ can be computed by adjusting for confounders. The *frontdoor criterion* applies when no backdoor adjustment is possible, but a mediating variable satisfies certain conditions [32].

The *do-calculus* of Pearl provides three rules for manipulating interventional distributions, and is complete for identification: if an effect is identifiable, *do-calculus* can derive the identifying formula [32]. In practice, graphical criteria like backdoor and frontdoor cover most applications.

3.6.2. Estimation

Once identification is established, estimation computes the causal effect numerically from finite data. Common causal estimands include:

- **Average Treatment Effect (ATE)**: $\tau = \mathbb{E}[Y(1) - Y(0)]$, the average effect across the full population.
- **Average Treatment Effect on the Treated (ATT)**: $\mathbb{E}[Y(1) - Y(0) \mid T = 1]$, the

average effect restricted to units that received treatment.

- **Conditional Average Treatment Effect (CATE)**: $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, the effect conditional on covariates, enabling personalized effect estimates.

Regression adjustment (outcome modeling). The most direct approach is to model the conditional outcome $\mu(t,x) = \mathbb{E}[Y \mid T = t, X = x]$ and estimate the ATE as:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mu}(1, x_i) - \hat{\mu}(0, x_i) \right]. \tag{8}$$

Any supervised learning model may serve as the outcome model $\hat{\mu}$. When a single model is fit on all units using the treatment indicator as a feature, this approach is called the S-learner (single learner) [16]. Because the treatment indicator is just one feature among many, flexible models may underweight it, biasing the treatment effect estimate toward zero. The T-learner (two learners) addresses this by fitting separate outcome models $\hat{\mu}_1$ and $\hat{\mu}_0$ on the treated and control subgroups respectively, and estimating $\tau(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$. The T-learner naturally produces CATE estimates but may suffer in low-overlap regions where one group is underrepresented [20].

Propensity score methods. The *propensity score* $e(x) = P(T = 1 \mid X = x)$ is the probability of receiving treatment given covariates. Rosenbaum and Rubin showed that if treatment is unconfounded given X, it is also unconfounded given $e(X)$. This balancing property reduces high-dimensional covariate adjustment to a one-dimensional problem. In *matching*, each treated unit is paired with one or more control units with similar propensity scores, and the treatment effect is estimated from matched pairs. In *Inverse Probability Weighting* (IPW), each unit is reweighted by the inverse probability of receiving its actual treatment, creating a pseudo-population in which treatment assignment is independent of covariates:

$$\hat{\tau}^{\text{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{T_i Y_i}{\hat{e}(x_i)} - \frac{(1 - T_i) Y_i}{1 - \hat{e}(x_i)} \right]. \tag{9}$$

In practice, the propensity score is estimated from data using logistic regression or a flexible classifier [20]. IPW is sensitive to extreme propensity score values (near 0 or 1), where a small number of units receive very large weights; stabilized or trimmed versions are often used in practice.

Meta-learners for CATE. The X-learner addresses the common practical scenario where the treated and control groups differ substantially in size [16]. It proceeds in two stages. In the first stage, separate outcome models $\hat{\mu}_1$ and $\hat{\mu}_0$ are fit on each group (as in the T-learner). In the second stage, imputed treatment effects are computed for each unit: $\tilde{\tau}_i^{(1)} = Y_i - \hat{\mu}_0(x_i)$ for treated units and $\tilde{\tau}_i^{(0)} = \hat{\mu}_1(x_i) - Y_i$ for control units. Two CATE models $\hat{\tau}_1$ and $\hat{\tau}_0$ are then fit on these imputed effects and combined via the propensity score: $\hat{\tau}(x) = \hat{e}(x)\,\hat{\tau}_0(x) + (1 - \hat{e}(x))\,\hat{\tau}_1(x)$. The propensity weighting places more trust on the group that is better represented in the data, yielding more data-efficient estimates in unbalanced settings.

Doubly robust estimators. Doubly robust (DR) estimators combine an outcome model $\hat{\mu}$ and a propensity model $\hat{e}$, resulting in estimates that are consistent if *either* model is correctly specified:

$$\hat{\tau}^{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}(1,x_i) - \hat{\mu}(0,x_i) + T_i\frac{Y_i - \hat{\mu}(1,x_i)}{\hat{e}(x_i)} - (1-T_i)\frac{Y_i - \hat{\mu}(0,x_i)}{1-\hat{e}(x_i)}\right]. \tag{10}$$

This redundancy makes DR estimators attractive when the correct functional form of either the outcome or the treatment model is uncertain [20]. Modern variants such as targeted maximum likelihood estimation (TMLE) and debiased machine learning further improve finite-sample behavior through cross-fitting and regularization. The DoWhy library [43] and EconML [47] provide implementations of these estimators alongside the graphical identification tools described in Section 3.3.

3.6.3. Validation and Refutation

Unlike purely predictive models that can be validated through held-out test sets, causal estimates fundamentally depend on assumptions that cannot be verified from observational data alone [32, 20]. The most critical of these is the assumption of no unmeasured confounding (also known as ignorability or exchangeability), which asserts that all variables affecting both treatment and outcome have been observed and properly adjusted for [20]. Since the counterfactual outcomes (i.e., what would have happened under alternative treatments) are never observed, there is no direct way to test whether these assumptions hold in practice [19]. Consequently, validation in causal inference focuses on assessing the robustness and plausibility of conclusions rather than directly verifying correctness.

Placebo tests apply the estimation procedure to outcomes or treatments known to have no causal relationship; detecting an effect indicates bias. **Sensitivity analysis** examines how estimates change under hypothetical violations of assumptions, quantifying robustness to unmeasured confounding [20]. **Refutation tests** in frameworks like DoWhy include adding random common causes, replacing treatment with a placebo, or testing on data subsets [43].

3.7. Applications in Machine Learning

The theoretical apparatus developed in the preceding sections is not merely of academic interest. Causal graphs, identification criteria, and effect estimators translate directly into practical improvements in how ML models are built, deployed, and explained. This section discusses two of the most consequential application areas: graph-guided feature selection for constructing robust predictive models, and counterfactual explanations for producing interpretable and actionable predictions. It closes with a brief description of the integrated causal analysis pipeline that connects discovery, inference, and deployment.

3.7.1. Causal Feature Selection and Engineering

Standard feature selection methods based on correlation or mutual information identify variables statistically associated with the outcome without distinguishing causal associations from spurious ones. A spurious association arises because both the feature and the outcome share a common cause (a confounder), not because the feature influences the

outcome directly. Such features may improve predictive accuracy in one environment but fail entirely when the confounding structure changes, a phenomenon known as distribution shift [41, 34].

Causal feature selection addresses this problem by using the structure of the causal graph to identify the variables that are stably predictive of the outcome across different environments [51]. The core concept is the *Markov blanket* of a target variable Y: the minimal set of variables that renders Y conditionally independent of all others. In a causal DAG, the Markov blanket of Y consists of its direct causes (parents), its direct effects (children), and the other direct causes of those effects (co-parents). A model trained exclusively on the Markov blanket of Y captures all predictive information about Y encoded in the graph, without incorporating noise from variables that are only marginally associated through indirect confounding paths.

In practice, causal graph structure guides several selection decisions. First, variables that are descendants of the treatment variable T should generally be excluded from adjustment sets, as conditioning on them can block causal paths (if they are mediators) or induce spurious associations (if they are colliders on a backdoor path). Second, including pure instruments, variables that affect T but have no direct path to Y, inflates the variance of propensity score estimators without reducing confounding bias [32]. Third, the backdoor criterion provides a systematic procedure for identifying a minimal adjustment set: the smallest subset of covariates that blocks all confounding paths between treatment and outcome, yielding unbiased estimates with the smallest possible variance penalty.

Beyond robustness to distribution shift, causally informed feature selection reduces dimensionality by discarding variables whose predictive content is entirely attributable to confounding that has already been accounted for. This can improve computational efficiency and reduce overfitting in small-sample settings, while also producing more interpretable models whose features have a coherent causal role [51].

Causal reasoning can also guide *feature engineering* rather than feature pruning alone. Instead of selecting a subset of existing variables, a practitioner can construct new composite features that encode hypothesized causal mechanisms more directly than any individual raw variable. For example, if the causal graph suggests that call failure rate and complaint frequency are both direct causes of customer dissatisfaction, which in turn drives churn, a composite *complaint severity* index that aggregates both variables may better reflect the underlying causal path than either variable alone. More broadly, behavioral summaries, satisfaction indices, and value segmentation variables derived from causal parent sets can enrich the feature space while remaining grounded in causal understanding. This approach expands the dataset with causally meaningful representations rather than reducing it, and is therefore complementary to selection.

A concrete demonstration of both approaches is provided by [13], which applies causality-driven feature engineering to telecommunications customer churn prediction. Using the Causal-Nest framework [12] to identify causal structure in the data, the study constructs 24 new features organized into categories including customer behavior patterns, satisfaction indices, value segmentation, and direct causal constructs derived from hypothesized causal relationships with churn. Combined with oversampling of the causally enriched feature set to address class imbalance, this approach achieves significant

improvements in minority-class F1-score for ensemble classifiers over correlation-based baselines, illustrating how causality can guide not just which features to keep but which new representations to create.

3.7.2. Counterfactual Explanations

Explainability methods for ML models are broadly divided into attribution-based methods, which assign importance scores to input features, and counterfactual methods, which identify the minimal change to the input that would alter the model's prediction. Techniques such as SHAP and LIME belong to the first category: they answer the question "which features contributed most to this prediction?" Counterfactual methods answer a qualitatively different question: "what would need to be different about this input for the outcome to change?"

From a causal perspective, the attribution question is ambiguous whenever input features are correlated. A high importance score on a feature may reflect its direct causal influence on the outcome, or merely its correlation with a true cause that was not included in the model. Counterfactual questions, by contrast, are naturally framed at the third level of Pearl's Causal Hierarchy [33]: they require imagining a world where the input had been different, which demands a causal model of the data-generating process.

A *counterfactual explanation* for a prediction $\hat{y} = f(x)$ is a minimally perturbed input x' such that $f(x') \neq \hat{y}$. For a loan rejection, such an explanation might read: "had your annual income been \$4,000 higher and your outstanding balance \$2,000 lower, the application would have been approved." This type of explanation is actionable (it suggests specific changes), plausible (the suggested values are within a realistic range), and consistent with causal intuitions about the decision process.

When a causal graph is available, counterfactual generation can be constrained to be causally valid. A naive counterfactual might suggest changing a variable (say, job title) without accounting for the downstream variables that would also change as a result (income, savings rate). A causally grounded generator, operating through the structural equations of the SCM, propagates the suggested change through the graph, ensuring that the proposed counterfactual input corresponds to a coherent, feasible alternative [33]. Open-source tools such as DiCE implement diverse counterfactual generation and can be extended with causal constraints. The connection between counterfactual explanations, algorithmic recourse, and causal fairness is an active research area at the intersection of causality, decision theory, and responsible artificial intelligence.

3.7.3. The Integrated Causal Pipeline

The sections of this chapter each correspond to a stage of a complete causal analysis pipeline. In practice, this pipeline proceeds as follows.

Step 1 (Discovery). A causal discovery algorithm (Section 3.5) is applied to the available observational data, yielding a candidate graph or equivalence class. Domain experts review and refine the result, orienting edges that the data cannot orient and adding or removing edges based on substantive knowledge. The outcome of this step is an agreed-upon causal graph G.

Step 2 (Identification). Given G, the identification criteria of Section 3.3 determine whether the causal effect of interest is recoverable from the available data, and which adjustment set or identification formula applies.

Step 3 (Estimation). An appropriate estimator (Section 3.6) is selected based on the data characteristics (sample size, treatment balance, expected functional form) and applied to produce a numerical estimate of the causal effect.

Step 4 (Refutation). Placebo tests, sensitivity analyses, and refutation procedures assess the robustness of the estimate to violations of the identification assumptions, such as the presence of unmeasured confounders.

Step 5 (Application). The validated causal effect estimate, combined with causal feature selection and counterfactual explanations, informs downstream decisions: what interventions to apply, which features to include in a deployed model, and how to explain individual predictions to end users.

This pipeline is demonstrated end-to-end in the practical notebooks accompanying this chapter, using the Sachs protein signaling dataset [39] as a benchmark. Because the ground truth causal graph for the Sachs data is known from experimental interventions, it serves as a concrete basis for evaluating the quality of the discovery step and the correctness of the subsequent inference.

3.8. Software Tools and Datasets

The causal analysis ecosystem comprises several open-source frameworks, each addressing different stages of the causal pipeline. However, these tools remain largely fragmented, and the field lacks standardized benchmark datasets with verified ground truth.

3.8.1. Causal Discovery Tools

The landscape of CD software is characterized by a diversity of tools developed across different programming ecosystems, each with distinct strengths and limitations.

The R ecosystem hosts two foundational packages for CD: **pcalg** [28] and **bnlearn** [42]. The pcalg package implements constraint-based algorithms such as PC and FCI, along with score-based methods like GES, and provides functions for causal effect estimation via adjustment sets. The bnlearn package focuses on Bayesian network structure learning, offering constraint-based, score-based, and hybrid algorithms. Both packages are mature and well-documented, serving as reference implementations for many discovery algorithms.

However, relying on R-based backends presents practical challenges. The R package ecosystem suffers from dependency management issues, where updates to underlying packages can break existing functionality without warning. Furthermore, integrating R packages with Python-based workflows requires inter-process communication, which introduces additional complexity, performance overhead, and potential points of failure. These integration challenges are particularly problematic in production environments or when combining discovery algorithms with Python-based machine learning pipelines.

Several Python discovery packages have emerged, such as The Causal Discov-

ery Toolbox (CDT) [21] that provides a unified Python interface that wraps algorithms from multiple backends, including pcalg and bnlearn, alongside native implementations of neural network-based methods like CGNN and SAM. While CDT simplifies access to diverse algorithms, it inherits the R dependency challenges for wrapped methods. The causal-learn library [55] offers pure Python implementations of classical algorithms (PC, GES) without external dependencies, improving portability at the cost of some algorithmic coverage. Similarly, Tetrad [36] from Carnegie Mellon University (CMU) provides a comprehensive suite of discovery algorithms with both graphical and programmatic interfaces. The gCastle library [54] focuses specifically on gradient-based and differentiable causal discovery methods implemented natively in Python with deep learning frameworks.

3.8.2. Causal Inference Tools

DoWhy [43] provides a unified interface for causal modeling, identification, estimation, and refutation, integrating graphical models with potential outcomes. However, DoWhy assumes the causal graph is provided or specified by the user, offering limited integration with discovery algorithms. EconML extends DoWhy with machine learning estimators for heterogeneous treatment effects. CausalML [8] focuses on uplift modeling and CATE estimation at scale.

3.8.3. Tools Overview

Table 3.3. Comparison of causal analysis software tools. ✓ indicates the feature is supported; – indicates it is not available.

Tool	Discovery	Estimation	Refutation	UI	Parallel
CDT [21]	✓	–	–	–	–
causal-learn [55]	✓	–	–	–	–
bnlearn [42]	✓	–	–	–	–
Tetrad [36]	✓	–	–	✓	–
gCastle [54]	✓	–	–	–	–
DoWhy [43]	–	✓	✓	–	–
EconML [47]	–	✓	–	–	–
CausalML [8]	–	✓	–	–	✓
CausalNex [35]	✓	✓	–	–	–
Salesforce CausalAI [2]	✓	–	–	✓	✓
Causal-Nest [12]	✓	✓	✓	–	✓

As Table 3.3 illustrates, no existing framework provides comprehensive support across all stages of the causal analysis pipeline. Discovery tools (CDT, causal-learn, bnlearn, Tetrad, gCastle) focus on structure learning but lack estimation and refutation capabilities. Inference tools (DoWhy, EconML, CausalML) assume the causal graph is provided, offering no discovery functionality. Practitioners must manually combine these tools, handling data format conversions and ensuring consistency between discovered structures and inference assumptions. This fragmentation increases the barrier to entry and the risk of methodological errors.

Recent research has addressed this gap directly. *Causal-Nest* [12] is a framework that integrates causal discovery, graph validation, effect estimation, and refutation within a unified pipeline, exposing a web interface to lower the operational barrier for non-expert practitioners. Its evaluation spans rigorous benchmarking on datasets with known causal structure, using metrics such as the Knowledge Integrity Score (KIS) and the Refutation Pass Rate (RPR) to quantify pipeline quality, as well as a practical demonstration in a customer churn prediction scenario where the full pipeline guides feature engineering and improves model performance. Readers seeking a reference architecture for end-to-end causal analysis may consult this work; note that the full stack includes R-based discovery backends that impose heavier infrastructure requirements than the Python-native libraries recommended for the practical exercises accompanying this chapter.

3.8.4. Benchmark Datasets

A significant challenge in CD research is the scarcity of datasets with verified ground truth causal structures. Unlike supervised learning, where ground truth labels are directly observable, the true causal graph underlying observational data is rarely known with certainty [16].

The CMU Example Causal Datasets repository[2] provides one of the few curated collections of datasets for benchmarking CD algorithms. Table 3.4 summarizes representative datasets from this collection.

Table 3.4. Representative benchmark datasets for CD. Ground truth types: *Graph* indicates a known causal DAG; *Constraints* indicates temporal ordering or forbidden/required edge specifications.

Dataset	Type	Features	Instances	Ground Truth	Domain
Abalone [31]	Real	8	4,177	Constraints	Biology
Adult [3]	Real	14	48,842	Constraints	Social Science
Airfoil Self-Noise [6]	Real	6	1,503	Constraints	Aeronautics
Car Evaluation [5]	Simulated	6	1,728	Constraints	Industry
Covertype [4]	Real	54	581,012	Constraints	Biology
Sachs [39]	**Real**	**11**	**7,466**	**Graph**	**Biology**
Dry Bean [24]	Real	16	13,611	Constraints	Biology
fMRI Feedbacks [40]	Simulated	5–10	60 each	Graph	Neuroscience
HTRU2 [26]	Real	8	17,898	Constraints	Physics
Hungary Chickenpox [37]	Real	20	521	Constraints	Medicine
Myocardial Infarction [14]	Real	111	1700	Constraints	Medicine
Student Performance [10]	Real	30	649	Constraints	Social Science
Superconductivty [17]	Real	81	21,263	Constraints	Physics
Wine Quality [11]	Real	11	6,498	Constraints	Food Science

The **Sachs** dataset [39] remains the most widely used real-world benchmark with a verified ground truth graph, derived from experimental interventions on protein signaling networks. The fMRI feedback datasets provide simulated data with known ground truth for neuroscience applications [40]. For most other datasets, ground truth is specified as temporal ordering constraints or domain knowledge about forbidden and required

[2] `https://github.com/cmu-phil/example-causal-datasets`

edges, following a standardized format that encodes partial causal knowledge rather than complete graphs.

This scarcity of ground truth datasets limits rigorous evaluation of CD algorithms on real-world problems. Simulated data can verify algorithmic correctness but may not capture the complexities of real distributions. Consequently, there is a pressing need for both additional benchmark datasets with experimentally verified causal structures and integrated tools that support the full causal analysis pipeline from data to validated conclusions.

3.9. Conclusion

This chapter has provided a self-contained introduction to computational causality for data scientists, building from motivating concepts to a practical methodology for complete causal analysis. The central theme is the distinction between association and causation, and the consequences of conflating the two when decisions require reasoning about interventions and counterfactuals.

Summary. Section 3.2 established the conceptual motivation: the insufficiency of association for causal conclusions, the do-operator as the formal expression of intervention, and Pearl's Causal Hierarchy framing the three levels of causal reasoning. Section 3.3 developed the graphical language of causal models, covering confounders, mediators, and colliders, d-separation, Markov equivalence, and the backdoor and front-door identification formulas. Section 3.4 gave mathematical substance to these graphical intuitions through Structural Causal Models, with structural equations and the graph-surgery interpretation of interventions, and the Potential Outcomes framework, with the identification assumptions (unconfoundedness, positivity, SUTVA) and the adjustment formula that links causal to statistical estimands. Section 3.5 surveyed the three families of causal discovery algorithms and their assumptions, with the Sachs benchmark as a concrete reference. Section 3.6 covered effect estimation from regression adjustment and propensity score methods to meta-learners and doubly robust estimators, and emphasized systematic refutation as an indispensable complement. Section 3.7 demonstrated how the causal framework improves practical ML workflows through feature selection, causally informed feature engineering, counterfactual explanations, and the integrated five-step analysis pipeline. Section 3.8 mapped the open-source tool ecosystem and highlighted the scarcity of verified benchmark datasets as a persistent challenge for the field.

Open challenges. Several important problems remain at the frontier of the field. Causal discovery in high-dimensional settings with limited samples is still an active research area: the search space of graphs grows super-exponentially and standard independence tests lose statistical power as the number of variables increases. Principled integration of domain knowledge into the discovery process, to constrain the search space without introducing human bias, remains an open design challenge. On the inference side, sensitivity analysis methods for quantifying the impact of unmeasured confounders are still maturing, as are methods for causal inference from non-i.i.d. data such as time series, longitudinal cohorts, and network-structured observations. Finally, the scarcity of datasets with experimentally verified causal ground truth limits rigorous end-to-end evaluation of the full pipeline on real-world problems.

Outlook. The integration of causal reasoning into standard data science practice is increasingly recognized as essential for constructing systems that generalize reliably, explain their predictions transparently, and support decisions that involve deliberate interventions [41, 33]. As the tool ecosystem matures and the community develops shared infrastructure and better benchmarks, causal data science is positioned to become a standard component of the practitioner's toolkit alongside traditional statistical machine learning. The foundations presented in this chapter provide the entry point for that transition.

References

[1] Bryon Aragam and Qing Zhou. Concave penalized estimation of sparse gaussian bayesian networks. *The Journal of Machine Learning Research*, 16(1):2273–2328, 2015.

[2] Devansh Arpit, Matthew Howson, et al. Salesforce causalai library: A fast and scalable framework for causal analysis of time series and tabular data. *arXiv preprint arXiv:2301.10859*, 2023.

[3] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

[4] Jock Blackard. Covertype. UCI Machine Learning Repository, 1998.

[5] Marko Bohanec. Car Evaluation. UCI Machine Learning Repository, 1988.

[6] Thomas Brooks, D. Pope, and Michael Marcolini. Airfoil Self-Noise. UCI Machine Learning Repository, 2014.

[7] Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

[8] Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. CausalML: Python package for causal machine learning. *arXiv preprint arXiv:2002.11631*, 2020.

[9] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[10] Paulo Cortez. Student Performance. UCI Machine Learning Repository, 2014.

[11] Paulo Cortez, António Cerdeira, Fernando Almeida, et al. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. Smart Business Networks: Concepts and Empirical Evidence.

[12] Gustavo F.V. de Oliveira, Fabrício A. Silva, and Marcus H.S. Mendes. Causalnest: A framework for automated causal discovery and inference. *Knowledge-Based Systems*, 2025. *To appear*.

[13] Gustavo F.V. de Oliveira, Fabrício A. Silva, and Marcus H.S. Mendes. Causality-driven feature engineering for enhanced telecommunications churn prediction. In

Anais do XVII Congresso Brasileiro de Inteligência Computacional (CBIC 2025), pages 1–8, Belo Horizonte, MG, 2025. SBIC.

[14] S. E. Golovenkin, V. A. Shulman, D. A. Rossiev, P. A. Shesternya, S. Yu. Nikulina, Yu. V. Orlova, and V. F. Voino-Yasenetsky. Myocardial infarction complications. UCI Machine Learning Repository, 2020.

[15] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, et al. Learning functional causal models with generative neural networks. *Explainable and interpretable models in computer vision and machine learning*, pages 39–80, 2018.

[16] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys*, 53(4):1–37, 2020.

[17] Kam Hamidieh. Superconductivty Data. UCI Machine Learning Repository, 2018.

[18] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.

[19] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[20] Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[21] Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python, 2019.

[22] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, et al. Structural agnostic modeling: Adversarial learning of causal graphs. *Journal of Machine Learning Research*, 23(219):1–62, 2022.

[23] Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

[24] Murat Koklu and Ilker Ali Ozkan. Dry Bean. UCI Machine Learning Repository, 2020.

[25] Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence*, pages 1052–1062. PMLR, 2022.

[26] Robert Lyon. HTRU2. UCI Machine Learning Repository, 2015.

[27] Dimitris Margaritis et al. *Learning Bayesian network model structure from data*. PhD thesis, School of Computer Science, Carnegie Mellon University Pittsburgh, PA, USA, 2003.

[28] Markus Kalisch, Martin Mächler, Diego Colombo, et al. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

[29] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[30] Preetam Nandy, Alain Hauser, and Marloes H Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

[31] Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1995.

[32] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[33] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution, 2018.

[34] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, 2017.

[35] QuantumBlack Labs. Causalnex. `https://github.com/mckinsey/causalnex`, 2021.

[36] Joseph Ramsey and Bryan Andrews. Py-tetrad and rpy-tetrad: A new python interface with r support for tetrad causal search. In Erich Kummerfeld, Sisi Ma, Eric Rawls, and Bryan Andrews, editors, *Proceedings of the 2023 Causal Analysis Workshop Series*, volume 223 of *Proceedings of Machine Learning Research*, pages 40–51. PMLR, 14 Aug 2023.

[37] Benedek Rozemberczki. Hungarian Chickenpox Cases. UCI Machine Learning Repository, 2021.

[38] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[39] Karen Sachs, Omar Perez, Dana Pe'er, et al. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[40] Ruben Sanchez-Romero, Joseph D. Ramsey, Kun Zhang, Madelyn R. K. Glymour, Biwei Huang, and Clark Glymour. Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019.

[41] Bernhard Schölkopf. *Causality for Machine Learning*, page 765–804. Association for Computing Machinery, 1 edition, 2022.

[42] Marco Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(3):1–22, 2010.

[43] Amit Sharma and Emre Kiciman. DoWhy: An end-to-end library for causal inference. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 722–731, 2020.

[44] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, et al. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[45] Edward H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[46] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[47] Vasilis Syrgkanis, Greg Lewis, Miruna Oprescu, et al. Causal inference and machine learning in practice with econml and causalml: Industrial use cases at microsoft, tripadvisor, uber. In *Proceedings of the 27th ACM SIGKDD*, KDD '21, page 4072–4073. ACM, 2021.

[48] Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, et al. Algorithms for large scale markov blanket discovery. In *FLAIRS*, volume 2, pages 376–81, 2003.

[49] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[50] Sandeep Yaramakala and Dimitris Margaritis. Speculative markov blanket discovery for optimal feature selection. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 809–812. IEEE Computer Society, 2005.

[51] Kui Yu, Xianjie Guo, Lin Liu, et al. Causality-based feature selection: Methods and evaluations. *ACM Comput. Surv.*, 53(5), September 2020.

[52] Changhe Yuan and Brandon Malone. Learning optimal bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48:23–65, 2013.

[53] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

[54] Keli Zhang, Shengyu Zhu, Marcus Kalander, et al. gCastle: A python toolbox for causal discovery. *arXiv preprint arXiv:2111.15155*, 2021.

[55] Yujia Zheng, Biwei Huang, Wei Chen, et al. Causal-learn: Causal discovery in python, 2023.