A Survey of Learning Causality with Data: Problems and Methods

RUOCHENG GUO, Computer Science and Engineering, Arizona State University
LU CHENG, Computer Science and Engineering, Arizona State University
JUNDONG LI, Computer Science and Engineering, Arizona State University
P. RICHARD HAHN, Department of Mathematics and Statistics, Arizona State University
HUAN LIU, Computer Science and Engineering, Arizona State University

The era of big data provides researchers with convenient access to copious data. However, we often have little knowledge of such data. The increasing prevalence of massive data is challenging the traditional methods of learning causality because they were developed for the cases with limited amount of data and strong prior causal knowledge. This survey aims to close the gap between big data and learning causality with a comprehensive and structured review of both traditional and frontier methods followed by a discussion about some open problems of learning causality. We begin with preliminaries of learning causality. Then we categorize and revisit methods of learning causality for the typical problems and data types. After that, we discuss the connections between learning causality and machine learning. At the end, some open problems are presented to show the great potential of learning causality with data.

CCS Concepts: • **Mathematics of computing** → *Causal networks*; • **Information systems** → *Data mining*; • **Computing methodologies** → *Machine learning*;

ACM Reference Format:
Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2010. A Survey of Learning Causality with Data: Problems and Methods. *ACM Trans. Web* 9, 4, Article 39 (March 2010), 36 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Causality is a generic relationship between an effect and the cause that gives rise to it. It is hard to define, and we often only know intuitively about causes and effects. Because it rained, the streets were wet. Because the student did not study, he did poorly on the exam. Because the oven was hot, the cheese melted on the pizza. When it comes to learning causality with data, we need to be aware of the difference between statistical associations and causations. For example, when the temperatures are hot, the owner of an ice cream shop might observe high electric bills and also high sales. Accordingly, she would observe a strong association between the electric bill and the sales figures, but the electric bill was not *causing* the high sales — leaving the lights on in the shop over night would have no impact on sales. In this case, the outside temperature is the common

Authors' addresses: Ruocheng Guo, Computer Science and Engineering, Arizona State University, Tempe, AZ, 85281, rguo12@asu.edu; Lu Cheng, Computer Science and Engineering, Arizona State University, Tempe, AZ, 85281, lcheng35@asu.edu; Jundong Li, Computer Science and Engineering, Arizona State University, Tempe, AZ, 85281, jundongl@asu.edu; P. Richard Hahn, Department of Mathematics and Statistics, Arizona State University, Tempe, AZ, 85281, prhahn@asu.edu; Huan Liu, Computer Science and Engineering, Arizona State University, Tempe, AZ, 85281, huan.liu@asu.edu.

1559-1131/2010/3-ART39 $15.00
https://doi.org/0000001.0000001

cause of both the high electric bill and the high sales numbers, and we say that it is a *confounder* of the causality of the electricity usage on the ice cream sales.

The ability to learn causality is considered as a significant component of human-level intelligence and can serve as the foundation of AI [102]. Historically, learning causality has been studied in myriad of high-impact domains including education [35, 57, 59, 75], medical science [89, 97], economics [66], epidemiology [58, 87, 115], meteorology [36] and environmental health [81]. Limited by the amount of data, solid prior causal knowledge was necessary for learning causality. Researchers performed studies on data collected through carefully designed experiments where solid prior causal knowledge is of vital importance [57]. To take a prototype example of *randomized controlled trials* [31], to study the efficacy of a drug, a patient would be randomly assigned to take the drug or not, which would guarantee that — *on average* — the treated and the un-treated (control) group are equivalent in all relevant respects, ruling out the influence of any other factors. Then, the impact of the drug on some health outcome — say, the duration of a migraine headache — can be measured by comparing the average outcome of the two groups.

The purpose of this survey is to consider what new possibilities and challenges arise for learning about causality in the present era of big data, taken here to mean both the availability of massive datasets. As an example, consider that the possibility of unmeasured confounders — might be mitigated now as vastly many more features can be measured. So, on one hand, it becomes possible for researchers to answer interesting causal questions with the help of big data. For instance, do positive Yelp[1] reviews drive customers to restaurants, or do they merely reflect popularity but not influence it? This causal question can be addressed by data from an extensive database maintained by Yelp. On the other hand, answering causal questions with big data leads to some unique new problems. For example, though public databases or data collected through web crawling or application program interfaces (APIs) are unprecedentedly large, we have very little intuition about what types of bias the dataset can suffer from — the data are more plentiful, but also perhaps more mysterious and, therefore, harder to model responsibly. At the same time, causal investigation is made more challenging by the same fundamental statistical difficulties that big data poses for other learning tasks (such as prediction). Perhaps the most notable example of this is the high-dimensionality of modern data [79], such as text data [65].

Many begin to investigate this intersection between big data and causal inquiry. Notable examples include but are not limited to [9, 37, 84, 92, 108]. In this survey, we aim instead to catalogue the different types of data that are available in this era and to provide an overview of the existing methods that attempt to answer causal questions using those data. As part of this effort, we will review the two primary formal frameworks for studying causality as well as the basic methodologies for learning causality, that underly more advanced techniques designed for big data.

1.1 Overview and Organization

Broadly, machine learning tasks are either *predictive* or *descriptive* in nature. But beyond prediction and description we may want to understand something *causal*, imaging that we were able to modify some variables and rerun the data-generating process. These types of questions can also take two (related) forms: 1) How much would some variables (features or the label) change if we manipulate the value of another variable? and 2) By modifying the value of which variables could we change the value of another variable? These questions are referred to as *causal inference* questions and *causal discovery* questions, respectively [45, 106]. For learning causal effects (causal inference), we investigate to what extent manipulating the value of a potential cause would influence a possible effect. Following the literature, we call the variable to be manipulated as *treatment* and the variable

[1]https://www.yelp.com/

for which we observe the response as *the outcome*, respectively. How much do hot temperatures raise ice cream sales, for example. For learning causal relationships (causal discovery), researchers attempt to determining whether there exists a causal relationship between a variable and another. In our temperature and ice cream example, it is clear that ice cream sales do not cause high temperatures, but in other examples it may not be clear. To take an extreme example, might it be that a genetic disposition towards cancer was responsible for individuals taking up smoking?

In this survey, we aim to start from the data perspective and provide a comprehensive review on how to learn causality from the massive data that we are often encountered with nowadays. Below, we present an outline of the topics that are covered in this survey. First, in Section 2, we introduce the preliminaries of learning about causality from data, which we will shorten to *learning causality* to encompass either causal inference or causal discovery. We mainly focus on the two most important formal frameworks, namely the *structural causal models* [100] and the *potential outcome framework* [96, 119]. Next, in Section 3 and 4 we go over the most common methodologies for learning causality from data. Specifically, in these two sections, the methods for both learning causal effects and relationships are categorized by the types of data they can handle. Section 3 focuses on the methods that are developed for the problem of learning causal effects (causal inference). Based on what data type they work on, these methods fall into three categories: methods for i.i.d data, non-i.i.d. data with the back-door criterion and the data without it. Then, in Section 4, the widely used methods for learning causal relationships are discussed. According to the data type, we first cover the methods for discovering causal relationships between variables in i.i.d. data. Then, we describe those methods of learning causal relationships that can tackle the inter-dependencies in time series data. Afterwards, in Section 5, we aim to provide an aspect of how some previous work narrowed the gap between learning causality and machine learning. Especially, we go over how the research in three subareas of machine learning, namely supervised and semi-supervised learning, domain adaptation and reinforcement learning can be connected to learning causality.

1.2 Data for Learning Causality

In this subsection, we discuss data and methods that are used for learning causal effects and relationships[2]. We start with the data types and methods for learning causal effects and then cover those for learning causal relationships.

Data for Learning Causal Effects. Here, we provide an overview of the types of data for learning causality, the problems that can be studied if the data is given, and the methods that can provide practical solutions. We introduce three types of data that can be applied to study learning causal effect. First, a standard dataset for learning causal effects $(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{y})$ includes a matrix of features $\boldsymbol{X}$ which is considered to provide enough information about the instances (satisfies the back-door criterion, see Section 2), a vector of treatments $\boldsymbol{d}$ and outcomes $\boldsymbol{y}$. With such a representation, this type of data is similar to what is often used for supervised learning. The only difference is that we are particularly interested in the causal effect of one feature D on the label or another feature as the outcome Y. For the second type, in addition to those that are present in the first type, there is auxiliary information about inter-dependence or interference between units such as links or temporal inter-dependencies between different data units (samples), represented by a matrix $\boldsymbol{A}$. Some special cases of this type of data can be attributed networks [80, 145], time series [38] and marked temporal point process [53]. Moreover, when there are unobserved confounders in the third type, we need the help of special causal variables, including the *instrumental variable (IV)*, the *mediator* and the *running variable*. These special variables are defined by typical causal knowledge, thus specific methods can be applied for learning causal effect for such types of data (see Section 3).

[2]The data index and algorithm index for learning causality are described in Appendix

Data for Learning Causal Relationships. We also describe two types of data for the study of (learning causal relationships) causal discovery. The first type of data is the conventional attribute-value data along with the known causal relationships, represented as $(\boldsymbol{X}, \boldsymbol{y})$, or $(\boldsymbol{X})$ and a ground truth causal graph for evaluation. With this type of data, we aim to learn the causal graph. A special case of this type of data is the bivariate data and the task reduces to distinguishing the cause from the effect [92]. The causal graph is usually defined by prior knowledge and could be incomplete. The second type of data for learning causal relationships is the multivariate time series data which also comes with a causal graph as the ground truth. We aim to learn causal relationships between different variables [48].

1.3 Previous Work and Contributions

There are a number of other comprehensive surveys in the area of causal learning. Pearl [100] aimed to convey the fundamental theory of causality based on the structural causal models. Gelman [45] provided high-level opinions about the existing formal frameworks and problems for causal learning. Mooji et al. [92] focused on learning causal relationships for bivariate data. Spirtes and Zhang [131] summarized methods for learning causal relationships on both i.i.d. and time series data but they focus on several semi-parametric score based methods. Athey and Imbens [11] described decision trees and ensemble machine learning models for learning causal effects.

Different from previous work, this survey is structured around various data types, and what sorts of causal questions can be addressed with them. Specifically, we describe what types of data can be used for the study of causality, what are the problems that can be solved for each type of data and how they can be solved. In doing so, we aim to provide a bridge between the areas of machine learning, data mining and causal learning in terms of terminologies, data, problems and methods.

1.4 Running Example

As a running example, we consider a study of how Yelp ratings influence potential restaurant customers [5]. Yelp is a website where consumers can share their reviews of a certain goods and services, including restaurants. Each review includes an integer rating from 1 to 5 stars. For our purposes, the Yelp rating is our *treatment* variable and the number of customers (in some well-defined period) is the *outcome* variable. For simplicity, we will consider these variables to be binary. A restaurant receives an active treatment $D = 1$ if its rating is positive (above some threshold); otherwise, it is under a control treatment $D = 0$. For the outcome, $Y = 1$ means a restaurant is completely booked and $Y = 0$ means it is not.

2 PRELIMINARIES FOR LEARNING CAUSALITY WITH DATA

Different from previous work, to build a solid technical background to tackle the challenges of learning causality with massive data, we present the preliminaries for both structural causal models and the potential outcome framework. First, we need serious representations for causal knowledge, which are often referred to as the *causal models*. We define the causal model as follows:

Definition 1. Causal Model. *A causal model is a mathematical abstract that quantitatively describes the causal relationships between variables.*

"No causes in, no causes out", the famous quote from Cartwright [22] summarizes the procedure of learning causality with data. First, causal assumptions or prior causal knowledge is often represented by an incomplete causal model. Then, the information missing from the causal model is filled in through learning from data. In the literature of learning causality, various formal frameworks have been developed for mathematical formulation of causal models. The two most widespread ones are the structural causal models (SCMs) [101] and the potential outcome framework [96, 119].

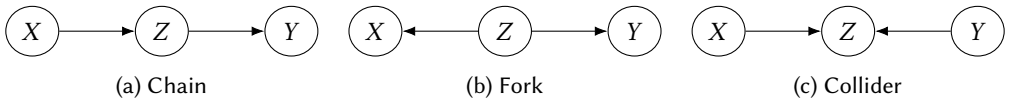
Table 1. Nomenclature

Nomenclature		
Terminology	Alternatives	Explanation
causality	causal relationship, causation	causal relationship between variables
causal effect		the strength of a causal relationship
instance	unit, sample, example	an independent unit of the population
features	covariates, observables pre-treatment variables	variables describing instances
learning causal effects	(forward) causal inference (forward) causal reasoning	identification and estimation of causal effects
learning causal relationships	causal discovery causal learning causal search	inference of possible causal graphs from data
causal graph	causal diagram	a graph with variables as nodes and causality as edges
confounder	confounding variable	a variable causally influences both treatment and outcome

These two formal frameworks are considered as the foundation of causality because they enable a consistent representation of prior causal knowledge, assumptions, and estimates such that we can start from the knowns (knowledge and assumptions) to learn the unknowns. As in the problem of *learning causal relationships*, we often start without any causal knowledge and aim to learn them from data only. We can consider the problem of learning causal relationships as learning causal models from data. We will discuss more details about this in Section 4.

We present the terminologies and notations that are used throughout this survey. We show a nomenclature for this survey in Table 1. In this survey, the a lowercase letter such as x, d or y denotes a specific value of the corresponding random variable. The bold lowercase letters such as $\boldsymbol{x}$ represent vectors. The uppercase letters such as X signify random variables. The Calligraphic uppercase letters such as $\mathcal{X}$ denote a set of random variables. The bold uppercase letters like $\boldsymbol{X}$ stand for a matrix. $\mathcal{U}$ signifies the set of instances, samples, examples or units, each instance in $\mathcal{U}$ is represented by u. $\mathcal{X}$ denotes the set of random variables for features, covariates, observations or pre-treatment variables. The matrix and vector forms of $\mathcal{X}$, namely $\boldsymbol{X}$ and $\boldsymbol{x}_u$, present features for all instances and that for the instance u, respectively. Without specification, the subscripts of the random variables of a set of features $\mathcal{X}$ denote the instance and the dimension, respectively. For example, $X_{u,j}$ signifies the random variable for the j-th feature for instance u. $\mathcal{S}$ denotes a subset of features that blocks all the *back-door path* (defined later in this Section). They are also referred to as the *sufficient set* or the *admissible set*. D denotes the treatment variable, in this work, it is often assumed to be binary and univariate. Y is referred to as the outcome variable, in this work, it is assumed to be univariate. We use the subscript and superscript of Y to represent the instance and the treatment it corresponds to. For example, when the treatment is binary, Y_u^1 denotes the outcome when instance u is under treatment $D_u = 1$. Under the binary treatment, $Y_u^{D_u}$ denotes the observed (factual) outcome while $Y_u^{1-D_u}$ represents the counterfactual outcome of instance u. τ denotes a certain type of treatment effect.

Next, we introduce the structural causal models (SCMs) with definitions of key concepts for learning causal effects followed by the potential outcome framework.



(a) Chain (b) Fork (c) Collider

Fig. 1. Three typical DAGs for conditional independence

2.1 Structural Causal Models

SCMs are developed toward a comprehensive theory of causation [100]. A causal model formulated by SCM consists of two components: the *causal graph* (causal diagram) and the *structural equations*[3].

Causal Graphs. A causal graph forms a special class of Bayesian network with edges representing the causal effect. Thus, it inherits the well defined conditional independence criteria of Bayesian network. Here, we formally define causal graph.

Definition 2. Causal Graph. *A causal graph $G = (\mathcal{V}, \mathcal{E})$ is a directed graph that describes the causal effects between variables, where $\mathcal{V}$ is the node set and $\mathcal{E}$ the edge set. In a causal graph, each node represents a random variable including the treatment, the outcome, other observed and unobserved variables. A directed edge $X \rightarrow Y$ denotes a causal effect from X to Y.*

A *path* is a sequence of directed edges and a *directed path* is a path whose edges point to the same direction. In this work, we only consider *directed acyclic graphs* (DAGs) where no directed path starts and terminates at the same node. Given a SCM, the conditional independence embedded in its causal graph provides sufficient information for us to confirm whether it satisfies the criteria such that we can apply certain causal inference methods. To understand the conditional independence, here, we briefly review the concept of *d-separation* based on the definition of *blocked* path. Fig. 1 shows three typical DAGs in explaining the conditional independence. In the *chain* (Fig. 1a), X causally affects Y through its influence on Z. In the *fork* (Fig. 1b), Z is the common cause of both X and Y. In this case, X is associated with Y but there is no causation between them. In the *collider* (see Fig. 1c), both X and Y cause Z but there is no causal effect or association between X and Y. In the chain and fork, the path between X and Y is blocked if we condition on Z, which can be denoted as $X \perp\!\!\!\perp Y|Z$. Contrarily, in the case of a collider (Fig. 1c), conditioning on Z introduces an association between X and Y, which can be represented by $X \perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Y|Z$. Generally, conditioning on a set of nodes $\mathcal{S} \subset \mathcal{V}$ blocks a path p iff there at least exists a node Z in p that is blocked. Here, we formally define a blocked node in causal graphs.

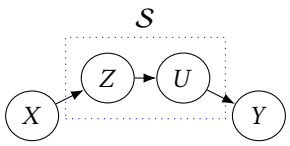
Definition 3. Blocking. *We say a node Z is blocked by conditioning on a set of nodes $\mathcal{S}$ if either one of the two following conditions is satisfied: $Z \in \mathcal{S}$ and Z is not a collider (Fig. 2a); or Z is a collider and neither Z nor any descendant of Z is in $\mathcal{S}$ (Fig. 2b).*

With this definition, we say a set of nodes $\mathcal{S}$ d-separates two variables X and Y iff $\mathcal{S}$ blocks all paths between them. The concept of d-separation plays a crucial role in explaining causal concepts which are introduced later in this subsection.

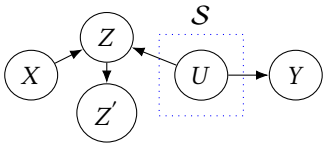
Similar to Bayesian networks, the assumption of the *causal Markovian condition* is often applied to SCMs, which means that with this assumption, we can factorize the joint distribution represented by a *Markovian* SCM of variables $\mathcal{V} = \{X_{,1}, ..., X_{,J}\}$ with:

$$P(X_{,1}, ..., X_{,J}) = \prod_{j=1}^{J} P(X_{,j}|\mathcal{P}_{,j}, E_{,j}), \tag{1}$$

[3]The terminology structural equation model was used to denote linear equations with causal effect as the coefficient for the treatment variable. However, recently, researchers start to use structural equations to refer to non-linear and even non-parametric equations with a more generalized definition of causal effect [107**?**].

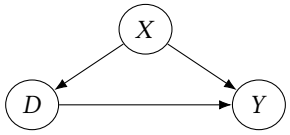


(a) Conditioning on $\mathcal{S}$ blocks the node Z as $Z \in \mathcal{S}$ and Z is not a collider.

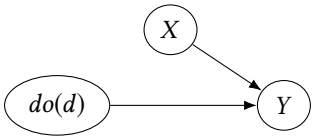


(b) Conditioning on $\mathcal{S}$ blocks Z as Z is a collider and neither Z nor $Z^{'}$ is in $\mathcal{S}$.

Fig. 2. Examples for blocked variable Z



(a) A SCM without intervention.



(b) A SCM under the intervention $do(d)$.

Fig. 3. SCMs without and under intervention $do(d)$ for the Yelp example: both SCMs include three variables: restaurant category X, Yelp rating D and customer flow Y.

where $\mathcal{P}_{,j}$ denotes the set of parent variables of $X_{,j}$, each of which has an arrow in $X_{,j}$. Moreover, the notation $E_{,j}$ stands for the noise term which represents the causal effect of unobserved variables on $X_{,j}$. In this work, we focus on Markovian SCMs. Here, we introduce the key concepts of learning causality through a toy SCM which embeds prior causal knowledge for the Yelp example [5]. In Fig. 3a, there are three random variables, namely X, D and Y, which stand for the restaurant category (*confounder*), Yelp star rating (treatment) and customer flow (outcome). With the three directed edges, this causal graph embeds the knowledge of the three causal effects:

(1) the category of a restaurant influences its Yelp rating. For example, the average rating of fast food restaurants is lower than that of Ramen restaurants.
(2) the category of a restaurant also influences its customer flow. For example, the average customer flow of fast food restaurant is higher than that of Ramen restaurants.
(3) Yelp rating of a restaurant influences its customer flow, which is a common understanding.

Structural Equations. Given a causal graph, a set of equations called structural equations can be developed to specify the causal effects represented by the directed edges in the graph. We can use a set of non-parametric structural equations as the representation for three causal effects embedded in the causal graph (Fig. 3a). Specifically, this associated structural equation model can be written down as a set of equations below, where each equation corresponds to one edge in the graph:

$$\begin{aligned} X &= f_X(E_X) \\ D &= f_D(X, E_D) \\ Y &= f_Y(X, D, E_Y). \end{aligned} \tag{2}$$

In Eq. 2, E_X, E_D and E_Y denote the random variables for the noise of each observed random variable. They are often assumed to be *exogenous*, which means there is no edge that points to them. As a result, they are independent of each other. Semantically, the noise terms represent the causal effect of unobserved variables on the variable on the LHS. It worths mentioning that for each equation, we assume that the variables on the RHS influences those on the LHS, not the other way around. For example, the second equation represents a data-generating process which examines

the values of X and E_D and then assigns the variable D with the value $f_D(X, E_D)$ [101]. Rewriting this equation in a different order as $X = f_D^{-1}(D, E_D)$ can be misleading as it embeds the causal knowledge that Yelp rating causally influences restaurant type. Thus, the direction of the causal effect is flipped, which is not desired. The structural equation (Eq. 2) provides a quantitative way to represent *intervention* on a certain variable of the corresponding causal graph (Fig. 3a). In SCMs, *do-calculus* [101] is developed to define intervention. In particular, the do-calculus introduces a new mathematical operator $do(d)$, which denotes the simulation of a physical intervention which sets the value of the random variable D to d. The notation of $do(d)$ leads to a formal expression of the corresponding interventional distributions as follows:

Definition 4. Interventional Distribution (Post-intervention Distribution). *The interventional distribution or post-intervention distribution $P(Y|do(x))$ denotes the distribution of the variable Y when we rerun the modified data-generation process where the value of variable X is set to x.*

For example, for the causal graph in Fig. 3a, the post-intervention distribution $P(Y|do(d))$ refers to the distribution of customer flow Y as if the rating D is set to d by intervention, where all the arrows into D are removed, as shown in Fig. 3b. The structural equations associated with Fig. 3b under the intervention on the treatment variable, denoted by $do(d)$, can be written as:

$$\begin{aligned} X &= f_X(E_X) \\ D &= d \\ Y &= f_Y(X, d, E_Y), \end{aligned} \tag{3}$$

which formulates the interventional distribution as $P(Y|do(d)) = f_Y(X, d, E_Y)$. Then, when it comes to the causal effect of D on Y, in the language of SCMs, the problem of calculating causal effects can be translated into queries about the interventional distribution $P(Y|do(d))$ with d different values. Implicitly, with SCMs, we assume that the variables follow the same causal relationships shown by a certain SCM for each instance in the population. Therefore, SCMs enable us to define *average treatment effect* (ATE). For the running example, the ATE of Yelp rating on the customer flow can be defined as a function:

$$ATE(d_+, d_-) = \mathbb{E}[Y|do(d_+)] - \mathbb{E}[Y|do(d_-)], d_+ > d_-, \tag{4}$$

where d_+ and d_- refer to the ratings that are considered as positive and negative, respectively. In many cases, the treatment variable is binary ($D \in \{0, 1\}$), thus the ATE reduces to a value:

$$ATE = \mathbb{E}[Y|do(1)] - \mathbb{E}[Y|do(0)] \tag{5}$$

instead of a function. However, the gap between an interventional distribution and the relevant probability (e.g., $P(Y|do(d))$ and $P(Y|D)$) impedes us from calculating ATE. In the literature of *learning causality*, we call this gap as the *confounding bias*. We present the formal definition of confounding bias with do-calculus and SCMs.

Definition 5. Confounding Bias. *Given variables Y, X, confounding bias exists for causal effect $X \rightarrow Y$ iff the statistical association is not always the same as the corresponding interventional distribution, namely $P(Y|X) \neq P(Y|do(x))$.*

Confounding bias often results from the existence of *back-door path* (e.g., the path $D \leftarrow X \rightarrow Y$ in Fig. 3a)[4]. Its formal definition is as follows:

[4]Confounding bias can exist even when there is no back-door path. One example is a type of selection bias [17], when the causal graph is $D \rightarrow Z \leftarrow X \rightarrow Y$ and the dataset is collected only for instances with $Z_u = 1$, then within this dataset, the estimated statistical association $P(Y|D)$ can be non-zero although we know that there is no causal effect $D \rightarrow Y$.

Definition 6. Back-door Path. *Given a pair of treatment and outcome variables* (D, Y)*, we say a path connecting* D *and* Y *is a back-door path for this pair* (D, Y) *iff it satisfies two conditions: first, it is not a directed path; second, it is not blocked (it has no collider).*

An example of back-door path is the path $D \leftarrow X \rightarrow Y$ in Fig. 3a. With the definition of back-door path, we can give the formal definition for a *confounder* or *confounding variable* as:

Definition 7. Confounder (Confounding Variable). *Given a pair of treatment and outcome variables* (D, Y)*, we say a variable* $Z \notin \{D, Y\}$ *is a confounder or confounding variable iff it is the central node of a fork and it is on a back-door path for* (D, Y)*.*

In particular, in the running example, the probability distribution $P(Y|D)$ results from a mixture of the causal effect $P(Y|do(d))$ and the statistical associations produced by the back-door path $D \leftarrow X \rightarrow Y$, where X is the confounder. Note that neither $X \rightarrow D$ nor $X \rightarrow Y$ is the causal effect we want to estimate. Estimating the causal effects we care about from observational data requires methodologies that can eliminate confounding bias. The procedure of removing confounding bias is referred to as *causal identification*. Causal identification can be formally defined as:

Definition 8. Causal Identification. *We say a causal effect is identified iff the hypothetical distribution (e.g., interventional distribution) that defines the causal effect is formulated as a function of probability distributions.*

In other words, for causal identification, we need to block the back-door paths that reflect other irrelevant causal effects. Intuitively, a way to eliminate confounding bias is to estimate the causal effect within subpopulations where the instances are homogeneous w.r.t. confounding variables [100]. This corresponds to *adjustment* on variables that satisfy the *back-door criterion* for causal identification [101]. With the definition of a blocked path, we present a formal definition of the back-door criterion.

Definition 9. Back-door Criterion. *Given a treatment-outcome pair* (D, Y)*, a set of features* $\mathcal{S}$ *satisfies the back-door criterion for* (D, Y) *iff when every variable in* $\mathcal{S}$ *is blocked, all back-door paths are blocked for the pair* (D, Y)*.*

A set of variables that satisfies the back-door criterion is referred to as a *admissible set* or a *sufficient set*. For the running example, we are interested in the causal effect of Yelp star rating on the customer flow ($D \rightarrow Y$) or equivalently the interventional distribution $P(Y|do(d))$. So for causal identification, we aim to figure out a set of features that satisfies the back-door criterion for the treatment-outcome pair (D, Y). For example, if restaurant category X is the only confounder for the causal effect of Yelp rating on customer flow, then $\mathcal{S} = \{X\}$ satisfies the back-door criterion. In the literature of learning causality with SCMs, there are two types of data w.r.t. the back-door criterion for causal inference. In the first type, we assume that the whole set or a subsets of the features $\mathcal{S}$ satisfies the back-door criterion such that by making adjustment on $\mathcal{S}$, $P(Y|do(d))$ can be identified. We will introduce methods for learning causal effects with data of this type in Section 3.1. In the second type, other criteria are used to identify causal effects even without the back-door criterion.

Beyond do-calculus. Although do-calculus plays an important role in the language of SCMs, it has some limitations. The limitation mainly comes from the assumption of do-calculus that the variables of all instance follow the same causal relationships. This implies that it is difficult to formulate individual-level hypothetical distributions with do-calculus in SCMs. Let us consider the running example, even if we could hack Yelp and replace the rounded star rating with the true average rating for all restaurants, we still cannot answer questions such as what would the customer flow for restaurant C be if we had increased its rating by 0.5 star without changing the ratings of others? In [102], Pearl refers to the hypothetical distributions for such cases which

cannot be identified through interventions as *counterfactuals*. Naturally, do-calculus, the formal representation of hypothetical intervention, cannot help us formulate counterfactuals within the language of SCMs. Therefore, besides do-calculus, Pearl [100] introduced a new set of notations. For example, $P(y_d|y', d')$ denotes the probability of $Y = y$ if D had been observed with value d, given the fact that we observe $Y = y', D = d'$ in the data. In the running example, for a restaurant with rating d' and customer flow y', the counterfactual probability $P(y_d|y', d')$ stands for how likely the restaurant's customer flow would be y if we had observed its rating as d.

2.2 Potential Outcome Framework

The potential outcome framework [96, 119] or the Neyman-Rubin Causal Model, proposed by Neyman and then made well-known by Rubin, has been widely adopted in many applied scientific research [11, 66]. The potential outcome framework is mainly applied to learning causal effect as it corresponds to a given treatment-outcome pair (D, Y). Here, we start this part with a formal definition of the *potential outcome*:

Definition 10. Potential Outcome. *Given the treatment variable D, the population being studied U and the outcome variable Y, the potential outcome of the instance $u \in U$, Y_u^d, is defined as the value of Y would have taken if the value of D for instance u had been set to d.*

Following this definition, the main challenge of causal inference is the fact that we can only observe one potential outcome for a given instance. With the definition of potential outcome, we can define the *individual treatment effect* (ITE) as the difference between two potential outcomes of a certain instance under two different levels of treatment. This also makes the potential outcome framework a natural fit for formulating causality of subpopulations. In a vast majority of applied studies based on the potential outcome framework, researchers simplify the treatment variable to be binary ($D \in \{0, 1\}$) without loss of generality, where $D = 1$ and $D = 0$ mean that an instance is assigned to the treatment and control group, respectively.

Then we can formally define that the ITE becomes as a unique value as:

Definition 11. Individual Treatment Effect. *Assuming binary treatment, given an instance $u \in U$ and its potential outcomes Y_u^1 and Y_u^0, the individual treatment effect is defined as:*

$$\tau_u = Y_u^1 - Y_u^0. \tag{6}$$

With this definition of ITE under the potential outcome framework (Eq. 6), we can extend it to ATE over the whole population being studied and other subpopulation average causal effects such as *conditional average treatment effect* (CATE). Earlier in this section, we have already defined ATE with do-calculus, here we show that ATE can also be formulated in the potential outcome framework. Formally, given the definition of ITE, we can formulate ATE as the expectation of ITE over the whole population U as:

$$\tau = \mathbb{E}_{u \in U}[\tau_u] = \mathbb{E}_{u \in U}[Y_u^1 - Y_u^0] = \frac{1}{|U|} \sum_{u \in U} (Y_u^1 - Y_u^0), \tag{7}$$

where $|U|$ denotes the population size. The average treatment effects on subpopulations can also be interesting for some studies. An example is the CATE of instances with a certain $\boldsymbol{X}$ $\tau(\boldsymbol{X}) = \mathbb{E}_{u \in \{v | v \in U \cap \boldsymbol{X}_v = \boldsymbol{X}\}}[\tau_u]$.

Three assumptions of the potential outcome framework. There are three assumptions we have make in formulating ITE from potential outcomes: *the stable unit (instance) treatment value assumption* (SUTVA), *consistency* and *ignorability (unconfoundedness)*. One can break down SUTVA into two distinct conditions: *well-defined treatment levels* and *no interference*. The condition of well-defined treatment indicates that given two different instances $u, u' \in U, u \neq u'$, if the values

of their treatment variable are equivalent ($D_u = D_{u'}$), then the two instances u and u' receive exactly the same treatment. In the running example, this condition can be interpreted in a way that being rated as 4-star means the same for different restaurants. The condition of no interference signifies that the potential outcomes of an instance is independent of what treatments the other units receive, which can be formally expressed as:

$$Y_u^{\boldsymbol{D}} = Y_u^{D_u}, \tag{8}$$

where $\boldsymbol{D} \in \{0,1\}^{|U|\times 1}$ denotes the vector of treatment variables for all the instances. Using the toy example, we can explain the condition of no interference as no matter what rating restaurant A receives, it would not influence the customer flow of restaurant B. The SUTVA assumption is in accordance with the implicit assumption of SCMs, where the same SCM describes the causal relationships for all instance and there is no interference between any pair of instances [102]. Although the condition of no interference applies to the most of situations for causal inference, there are real-world applications where the inter-dependence between instances cannot be ignored. Two examples are the *spillover effect* and the *treatment entanglement*, where instances causally influence each other [10, 111, 136]. The second assumption, consistency, means that the value of potential outcomes would not change no matter how the treatment is observed or assigned through an intervention. In the running example, with consistency, we assume that for a restaurant u, if the observed customer flow is $y_u^{d_u}$, then no matter how the rating of a restaurant d_u is observed or set by intervention, the value of $y_u^{d_u}$ is the same. Finally, with ignorability (unconfoundedness), we assume that all the confounding variables are observed and reliably measured by a set of features $\mathcal{S}$ for each instance $u \in U$. Using the language of SCMs, this means $\mathcal{S}$ satisfies the back-door criterion, we can ensure all the back-door paths are blocked. In other words, there does not exist any back-door path via unobserved confounders [132]. *Ignorability* means that the values of the potential outcomes are independent of the observed treatment, given the values of a certain set of confounding variables. Mathematically, ignorability can be formulated as:

$$Y_u^1, Y_u^0 \perp\!\!\!\perp D_u | \mathcal{S}, \tag{9}$$

where $\mathcal{S}$ denotes a set of confounding variables, namely a subset of features that describes the instance u and causally influences the values of both the treatment D_u and outcome Y_u. From the notation, we can see that this is also an assumption defined at the individual level. Ignorability plays a core role in the potential outcome framework as it directly leads to causal identification [102]. An extra condition $P(D_u = 1|\boldsymbol{X}_u) \in (0,1)$ is usually added to ignorability to make it *strong ignorability*. In Section 3, we will introduce how strong ignorability leads to causal identification in methods for causal inference such as *propensity score matching*.

3 CAUSAL INFERENCE: LEARNING CAUSAL EFFECTS

We have discussed about the formal frameworks in Section 2. They are important for learning causal effects (causal inference) in terms of formulating prior knowledge or assumptions about the causal relationship between variables with causal graph. In this section, we introduce methods for learning causal effect. In particular, we aim to understand how to quantify a certain causal effect in a data-driven way. We formally define the problem of learning causal effect below:

Definition 12. Learning Causal Effects (Causal Inference) *Given multivariate data for N instances, $[(\boldsymbol{x}_1, d_1, y_1), ..., (\boldsymbol{x}_N, d_N, y_N)]$, learning causal effect quantifies how the outcome variable Y is expected to change if we modify the treatment variable from $D = d$ to $D = d'$, and it can be mathematically written as $\mathbb{E}[Y|d'] - \mathbb{E}[Y|d]$.*

A vast majority of existing work about learning causality focuses on the problem of learning causal effects. Depending on what the problem is, we may care about the causal effect for different populations. It can be the whole population, a known subpopulation that is defined by some conditions, an unknown subpopulation which we need to detect, or an individual. Amongst all kinds of treatment effects, the average treatment effect (ATE) is often the most interesting one when it comes to the decision on whether a certain treatment should be carried out on the population. Furthermore, in the language of SCM and do-calculus, identification of ATE only requires to query interventional distributions but not counterfactuals. This means that ATE is often easier to identify and estimate than other subpopulation treatment effects. We can apply the prevalent regression error metrics in evaluating models for learning ATE. For example, we can measure the absolute error on ATE. Given the ground truth τ and the inferred ATE $\hat{\tau}$, the absolute error on ATE is:

$$\epsilon_{ABS\_ATE} = |\tau - \hat{\tau}|. \tag{10}$$

Similarly, we can apply the same metrics for learning average treatment effects over subpopulations.

However, when the population consists of considerable heterogeneous groups, ATE can be misleading. For example, Yelp rating may matter much more for restaurants in big cities than those in small towns. Therefore, ATE can be spurious as an average of heterogeneous causal effects. In contrast, the average should be taken within each individual group. In many cases, without knowledge about the ground truths for subpopulations, one assumption we can make is that the subpopulations are defined by features. Instead of treating ATE as a constant, for the study of heterogeneous subpopulation causal effects, this assumption allows us to learn a function which maps the features that define a subpopulation to its estimated ATE. With this assumption, we can define the average treatment effect of instances with the same features $\boldsymbol{X} = \boldsymbol{x}$ as the conditional average treatment effect:

Definition 13. Conditional Average Treatment Effect (CATE) *Given a certain value of features $\boldsymbol{x}$ and binary treatment D, the CATE is a function of $\boldsymbol{x}$ and is defined as:*

$$\tau(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, D = 1] - \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, D = 0]. \tag{11}$$

In this case, we do not care about which instance u is, but assume only the features $\boldsymbol{X}$ and the treatment D are two factors that determine the outcome Y. The target is to learn a function $\hat{\tau}$ to approximate the CATE, which minimizes the expected squared error in the estimated CATE:

$$\mathbb{E}[(\hat{\tau}(\boldsymbol{X}) - \tau(\boldsymbol{X}))^2]. \tag{12}$$

Empirically, with cross-validation, we can evaluate the quality of the learned function $\hat{\tau}(\boldsymbol{X})$ based on the mean squared error (MSE) on the set of instances (U):

$$\epsilon_{PEHE} = \frac{1}{|U|} \sum_{u \in U} (Y_u^1 - Y_u^0 - \hat{\tau}(\boldsymbol{X}_u))^2. \tag{13}$$

This metric is also referred to as *precision in estimation of heterogeneous effect* (PEHE). It is also adopted for evaluating estimated individual treatment effects (ITE) [59, 69, 84, 124].

3.1 Learning Causal Effects on Data with Back-Door Criterion

In many real-world problems, based on the prior knowledge of causal relationships between variables, researchers assume that conditioning on the complete set or some subsets of features can block all the back-door paths. This assumption allows us to solve the problem of learning causal effects through *adjustment* (conditioning) on a set of variables $\mathcal{S}$ that blocks all the back-door paths. Adjustment refers to methods that eliminate confounding bias based on $\mathcal{S}$. There are two classic implementations of adjustment: regression adjustment and propensity score methods.

In this Section, we use $\mathcal{S}$ to denote a set of features which blocks all the back-door path or satisfies strong ignorability. Without loss of generality, we assume binary treatment $D \in \{0, 1\}$. We adopt the language of generalized structural equation introduced in Section 2. The causal graph embedding the causal assumption for this type of methods is shown in Fig. 4.

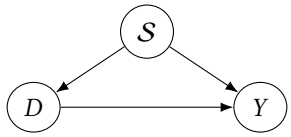


Fig. 4. A causal graph for data with the back-door criterion.

3.1.1 Regression Adjustment. In supervised learning, we fit a function to estimate the probability distribution $P(Y|\boldsymbol{X})$ where Y and $\boldsymbol{X}$ denote the observed label and the features. In Section 2, we already discussed that, for learning causal effect, we are interested in interventional distributions and counterfactuals which cannot be directly estimated from data. From a practicable perspective, using the terminologies of the potential outcome framework, we aim to infer the counterfactual outcomes Y^{1-D_u} based on the features $\mathcal{S}$ and the treatment D. There are two types of regression adjustment. First, we can fit a single function to estimate $P(Y|\mathcal{S}, D)$. This is enough for estimation of ITE for each instance because $\mathcal{S}$ is a sufficient set, which means by conditioning on $\mathcal{S}$, there is no confounding bias, namely $P(Y|D = d, \mathcal{S} = s) = P(Y|do(d), \mathcal{S} = s)$. So we can infer the counterfactual outcome as $\hat{Y}_u^{1-D_u} = \mathbb{E}(Y_u|1 - D_u, \mathcal{S}_u)$. Alternatively, in similar ways, it is also possible to fit two separate functions for the estimation of two potential outcomes, i.e. $P_1(Y|\mathcal{S}) = P(Y|D = 1, \mathcal{S})$ and $P_0(Y|\mathcal{S}) = P(Y|D = 0, \mathcal{S})$. Then we can estimate ATE by:

$$\hat{ATE} = \frac{\sum_{u \in \mathcal{U}} \hat{Y}_u^1 - \hat{Y}_u^0}{|\mathcal{U}|}, \tag{14}$$

where the factual outcome is observed $\hat{Y}_u^{D_u} = Y_u$ and the counterfactual outcome is estimated by the model $\hat{Y}_u^{1-D_u} = E(Y|1 - D_u, \mathcal{S}_u)$. In short, with regression adjustment we first estimate the counterfactual outcomes with regression models. Then, with the counterfactual outcomes, ITE can be calculated as the difference between two different outcomes.

3.1.2 Propensity Score Methods. Propensity score methods can be considered as a special case of matching methods which has been studied by statisticians [94]. Matching methods divide instances into strata and treat each stratum as a randomized controlled trial (RCT). Based on this assumption, ATE is automatically identified and can be estimated by the naïve estimator within each stratum. To make matching methods work, we assume *perfect stratification* that each group defined by a sufficient set $\mathcal{S}$ only contains indistinguishable instances except the observed treatment D and the potential outcomes Y [94]. Formally, perfect stratification can be written as:

$$\begin{aligned} E_{u \in \mathcal{U}}[Y_u^1 | D_u = 1, f_S(\mathcal{S})] &= E_{u \in \mathcal{U}}[Y_u^1 | D_u = 0, f_S(\mathcal{S})] \\ E_{u \in \mathcal{U}}[Y_u^0 | D_u = 1, f_S(\mathcal{S})] &= E_{u \in \mathcal{U}}[Y_u^0 | D_u = 0, f_S(\mathcal{S})], \end{aligned} \tag{15}$$

where function $f_S(\mathcal{S})$ outputs a continuous value based on which instances are stratified into homogeneous groups. The two equations in Eq. 15 can be interpreted as: given the group affiliation, the expected values of the potential outcomes do not change with the value of the treatment variable. This is equivalent to ignorable treatment assignment in each stratum defined by $f_S(\mathcal{S})$. The parameterization of the function $f_S(\mathcal{S})$ can be flexible. For example, it may only take a subset of $\mathcal{S}$ as the input. But we need to be careful when there exists a group which only contains instances

with $D = 1$ or $D = 0$. In this case we cannot estimate ATE in this stratum. This issue is referred to as *the lack of overlap*. To deal with this problem, *matching as weighting* methods are proposed. The most widely adopted matching as weighting methods specify the function $f_S(\mathcal{S})$ in estimating *propensity score* $P(D|\mathcal{S})$. Propensity score means the probability of receiving treatment. This results from the assumption that $\mathcal{S}$ predicts the treatment assignment. A natural question arises: *Does propensity score help*? Propensity score helps in the sense that it squeezes the space of features into one dimension to avoid the possible data sparseness issue. Following Rosenbaum and Rubin [118], we can estimate the propensity score by training a classifier which predicts how likely an instance is going to receive treatment given its features. A widely-adopted parameterization of $P(D|\mathcal{S})$ is the logistic regression function which outputs a number between 0 and 1. It can be written as:

$$P(D|\mathcal{S}; \boldsymbol{w}, w_0) = \frac{1}{1 + \exp(-\boldsymbol{w}^T \boldsymbol{s} - w_0)}, \tag{16}$$

where $\boldsymbol{s}$ is the vector form of the sufficient set $\mathcal{S}$. The same as supervised learning, we can estimate the parameters $\boldsymbol{w}$ and w_0 through minimization of the negative log-likelihood:

$$\arg\min_{\boldsymbol{w}, w_0} - \sum_{u \in U} \log P(D_u|\mathcal{S}_u). \tag{17}$$

Austin [12] categorized propensity score methods into four classes: *propensity score matching* (PSM), *propensity score stratification*, *inverse probability of treatment weighting* (IPTW) and *adjustment based on propensity score*. Here we focus on the PSM and IPTW because propensity score stratification is a natural extension of PSM, and adjustment based on propensity score is a combination of regression adjustment and other propensity score methods.

Propensity Score Matching (PSM). PSM is the approach to match a treated (controlled) instance to a set of controlled (treated) instances with similar estimated propensity scores. The most common approach is *Greedy One-to-one Matching* [52] for its computational efficiency and effectiveness. For each instance u, we find a matched instance u' for u by:

$$\arg\min_{u' \in \{v | v \in U \setminus u \cap D_v = 1 - D_u\}} dist(u, u'), \tag{18}$$

where $dist(\cdot, \cdot)$ is a distance metric measuring how similar two instances are. For matching in the space or subspace of the observed variables, we can use Euclidean distance $dist(u, u') = ||\boldsymbol{s}'_u - \boldsymbol{s}'_{u'}||_2$ where $\boldsymbol{s}'$ denotes the vector form of the selected variables $S' \subseteq S$ for matching. But for PSM the distance is often calculated based on the absolute difference of propensity scores as $dist(u, u') = |P(D|\mathcal{S}_u) - P(D|\mathcal{S}_{u'})|$. Once the instances are matched, we can estimate ATE by taking the average over the difference between observed outcomes from the treated group and the controlled group:

$$\hat{\tau} = \frac{\sum_{u \in U^1}(Y_u - Y_{u'}) + \sum_{u \in U^0}(Y_{u'} - Y_u)}{|U|}. \tag{19}$$

One-to-one PSM is based on the assumption that within each matched pair, the treatment assignment is ignorable. Besides the Greedy One-to-one PSM, there are many other PSM methods. The difference comes from whether we use Eq. 18 or other method to match instances. A widely used alternative to Greedy One-to-one PSM is the *Nearest Neighbor PSM*, where we match a treated instance with its nearest neighbors in terms of distance in the feature space, $dist(\mathbb{S}_u, \mathbb{S}_{u'})$. We either specify the number of nearest neighbors or the maximum acceptable distance to obtain the set of nearest neighbors. Then a random instance is selected from the nearest neighbors to form a matched pair. Finally, ATE is estimated by Eq. 19. We can also do many-to-one PSM, where a treated instance is matched to Q controlled instances. In [91], Ming and Rosenbaum showed that making Q a variable can substantially reduce bias. In [52], Gu and Rosenbaum carried out a systematic comparison

between various PSM methods. Stratification on propensity score is a natural extension of PSM. After propensity score is estimated, we can stratify instances based on the predefined thresholds of propensity scores or the number of strata. For example, in the clinical study, instances are often divided into five equal-size groups based on their propensity scores [12]. Thus, stratum-specific ATE can be directly calculated by the naïve estimator for instances in the same stratum. Then we take a weighted average over all strata to calculate the ATE:

$$\hat{\tau} = \frac{\sum_{i=1}^{N} |U_i| (\frac{1}{|U_i^1|} \sum_{u \in U_i^1} Y_u - \frac{1}{|U_i^0|} \sum_{u \in U_i^0} Y_u)}{\sum_{i=1}^{N} |U_i|}, \tag{20}$$

where U_i, U_i^1 and U_i^0 denote the set of instances, treated instances and controlled instances in the i-th stratum, respectively. A combination of regression adjustment and propensity score stratification can also be used to account for the difference between instances in the same stratum [12, 66, 86].

Inverse Probability of Treatment Weighting (IPTW). The IPTW [60] strategy is introduced to overcome the problem of data sparseness in the subspace of features. The intuition for IPTW is that we can weight instances based on their propensity scores such that we synthesize a RCT based on the data [12]. One common way to define the weight w_u for a instance u is by:

$$w_u = \frac{D_u}{P(D_u|\mathcal{S}_u)} + \frac{1 - D_u}{1 - P(D_u|\mathcal{S}_u)}. \tag{21}$$

With Eq. 21, we can find that for the treated instance u and the controlled instance u', $w_u = \frac{1}{P(D_u|\mathcal{S}_u)}$ and $w_{u'} = \frac{1}{1-P(D_{u'}|\mathcal{S}_{u'})}$. So the weight represents the inverse probability of receiving treatment for the treated and that of not receiving treatment for the controlled instances. Then we can calculate a weighted average of factual outcomes for the treatment and control groups:

$$\hat{\tau} = \frac{1}{|U^1|} \sum_{u \in U^1} w_u Y_u - \frac{1}{|U^0|} \sum_{u \in U^0} w_u Y_u = \frac{1}{|U^1|} \sum_{u \in U^1} \frac{Y_u}{P(D_u|\mathcal{S}_u)} - \frac{1}{|U^0|} \sum_{u \in U^0} \frac{Y_u}{1 - P(D_u|\mathcal{S}_u)}. \tag{22}$$

This is based on the idea that weighting the instances with inverse probability makes a synthetic RCT. Hence, a naïve estimator can be applied to estimate the ATE as in Eq. 22. Regression adjustment can also be applied to the weighted dataset to reduce the residual of the synthetic RCT [68]. Instances with propensity score close to 1 or 0 may suffer from an extremely large weight. In [58], Hernan proposed to stabilize weights to handle this issue in IPTW. In addition to estimating ATE, Morgan [93] showed that IPTW naturally lead to unbiased and consistent estimators for the average treatment effect of the treated instances (ATT) and that of the controlled instances (ATC). Besides reweighting samples with propensity scores, the *confounding balancing* methods learn sample weights through regression [74].

Doubly Robust Estimation (DRE). Funk et al. [43] proposed DRE as a combination of a regression adjustment model $\mathbb{E}[Y|D, \boldsymbol{X}]$ and another that estimates the propensity score $\mathbb{E}[D|\boldsymbol{X}]$. The robustness of this model is based on that only one of the two underlying models needs to be correctly specified to make it an unbiased and consistent estimator of ATE. In particular, a DRE model estimates individual-level potential outcomes based on the two models as:

$$\hat{Y}_u^1 = \frac{Y_u D_u}{\hat{P}(D_u|\boldsymbol{X}_u)} - \frac{\tilde{Y}_u^1 (D_u - \hat{P}(D_u|\boldsymbol{X}_u))}{\hat{P}(D_u|\boldsymbol{X}_u)} \quad , \quad \hat{Y}_u^0 = \frac{Y_u (1 - D_u)}{1 - \hat{P}(D_u|\boldsymbol{X}_u)} - \frac{\tilde{Y}_u^0 (D_u - \hat{P}(D_u|\boldsymbol{X}_u))}{1 - \hat{P}(D_u|\boldsymbol{X}_u)} \tag{23}$$

where $\tilde{Y}_u^{D_u}$ denotes the estimated potential outcomes for the instance u with the regression adjustment model $\mathbb{E}[Y|D, \boldsymbol{X}]$ and $\hat{P}(D_u|\boldsymbol{X}_u)$ is the estimated propensity score for the instance u. Taking a closer look at Eq. 23, we can find that the regression adjustment model is directly applied for the estimation of counterfactual outcomes as: $\hat{Y}_u^{1-D_u} = \tilde{Y}_u^{1-D_u}$, while more complicated, a mixture of

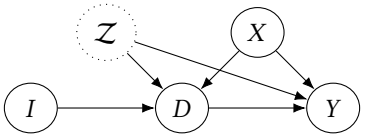


Fig. 5. A causal graph of a valid instrumental variable (I) when there are unobserved confounders ($\mathcal{Z}$). The binary exogenous variable I stands for whether a customer submits a review. The restaurant type (X) is an observed confounder and $\mathcal{Z}$ is a set of unobserved confounders.

the regression adjustment of propensity score models is designed for the factual outcomes. Then we can estimate ATE by taking the average over the estimated ITE for all the instances as in Eq. 14. A more generalized version of DRE , the *Targeted Maximum Likelihood Estimator* is proposed in [137].

3.2 Learning Causal Effects with Unobserved Confounders

In many real-world problems of learning causal effects, there exist unobserved confounders. The prior knowledge may imply that some confounding variables are not measurable or it may not be enough to confirm the causal relationship between all observed variables. In these cases, we are not able to block back-door path by simply conditioning on observed variables. Or, in the language of the potential outcome framework, we say the treatment assignment is not ignorable. Therefore, a family of methods are developed to handle this situation. The intuition is to find alternative information to complete this task. Here, we focus on three most popular methods for learning causal effects from data with unobserved confounders: *instrumental variable methods*, *front-door criterion* and *regression discontinuity design*.

3.2.1 Instrumental Variable Methods. There is a type of variable called *instrumental variables* with which causal effects can be learned. We provide the definition below:

Definition 14. Instrumental Variable *Given an observed variable I, other observed variables $\mathcal{X}'$, the treatment D and the outcome Y, we say variable I is a valid instrumental variable (IV) for the causal effect of $D \to Y$ iff I satisfies two conditions [7]: $I \not\perp\!\!\!\perp D|\mathcal{S}$ and $I \perp\!\!\!\perp Y|\mathcal{S}, do(d)$.*

In other words, a valid IV I causally influences the outcome Y only through affecting the treatment D. In the language of SCM, the first condition means that there is an edge $I \to D$ or a non-empty set of collider(s) $\mathcal{X}'$ s.t. $I \to \mathcal{X}' \leftarrow D$ where $\mathcal{X}' \subseteq X$. The second condition requires that the path $I \to D \to Y$ is the only path that starts from I and ends at Y. Thus, blocking D makes I and Y independent. This implies the *exclusive restriction* that there must not exist the direct edge $I \to Y$ or path $I \to \mathcal{X}' \to Y$ where $\mathcal{X}' \subseteq \mathcal{X}$. Mathematically, for all D and $i \neq i'$, this can also be denoted by:

$$Y(do(i), D) = Y(do(i'), D). \tag{24}$$

In the running example, if we only observe one of the confounders, the restaurant type (X), while the other confounders ($\mathcal{Z}$) remain unobserved. By assuming that whether a customer submits a review (I) is an exogenous random variable, then it is a valid IV (Fig. 5). This is because whether a customer writes a review (I) causally influences the average rating (D) and it can only causally affect the customer flow (Y) through its influence on D. With a valid IV, we identify the causal effect $D \to Y$ if both the interventional distributions - $P(D|do(i))$ and $P(Y|do(i))$ are identifiable.

A Linear SCM for IV Estimator. If we assume that the causal effect of D on Y is a constant τ, then we only need an *isolated source of variation* in D and Y to identify and estimate the causal effect τ [94]. This means that we only require this source of variation to explain part of the variance in D which also influences Y through D. According to the two conditions listed above, a valid IV I

naturally satisfies this requirement. We follow the language of SCM and assume that the structural model is linear. If we also assume that the observed and unobserved confounders $\mathcal{X}$ and U come with zero mean. We can write down the structural equations for the causal graph in Fig. 5 below.

$$\begin{aligned} D &= \alpha_i I + \boldsymbol{\alpha}_u^T \boldsymbol{z} + \boldsymbol{\alpha}_x^T \boldsymbol{x} + \alpha_0 + \epsilon_d \\ Y &= \tau D + \boldsymbol{\beta}_u^T \boldsymbol{z} + \boldsymbol{\beta}_x^T \boldsymbol{x} + \beta_0 + \epsilon_y, \end{aligned} \tag{25}$$

where $\boldsymbol{z}$ and $\boldsymbol{x}$ are the vector form of the set of random variables $\mathcal{Z}$ and $\mathcal{X}$, respectively; ϵ_d and ϵ_y are Gaussian noise terms with zero mean. By substituting D in the second equation with the RHS of the first equation in Eq. 25, we get:

$$Y = \tau\alpha_i I + (\tau\alpha_u + \beta_u)^T \boldsymbol{z} + (\tau\alpha_x + \beta_x)^T \boldsymbol{x} + \gamma_0 + \eta, \tag{26}$$

where $\gamma_0 = \tau\alpha_0 + \beta_0$ and $\eta = \tau\epsilon_d + \epsilon_y$. Then it is not difficult to figure out an estimator for the average treatment effect (τ):

$$\hat{\tau} = \frac{\mathbb{E}[Y|I = i] - \mathbb{E}[Y|I = i']}{\mathbb{E}[D|I = i] - \mathbb{E}[D|I = i']}. \tag{27}$$

Here, we rely on a bunch of assumptions to accomplish the task of learning causal effects: linear structural equations, valid IV, zero-mean additive noise and zero-mean unobserved confounders. What if some of them are not satisfied by the dataset we are interested in? For example, the causal relationship is non-linear. In the following example, with the potential outcome framework we show this estimator works in the non-linear case.

An IV Estimator under the potential outcome framework. The language of the potential outcome framework formulates the individual causal effect of the IV I on the outcome Y:

$$Y_u(I_u = 1, D_u(I_u = 1)) - Y_u(I_u = 0, D_u(I_u = 0)), \tag{28}$$

where we assume the IV I is binary. With the *exclusion restriction*, we know that I only affects Y through its influence on D, so we remove I_u as an explicit variable which influences the value of Y_u and reduces Eq. 28 to:

$$\begin{aligned} &Y_u(D_u(I_u = 1)) - Y_u(D_u(I_u = 0)) \\ &= [Y_u^1 P(D_u = 1|I_u = 1) + Y_u^0 P(D_u = 0|I_u = 1)] - [Y_u^1 P(D_u = 1|I_u = 0) + Y_u^0 P(D_u = 0|I_u = 0)] \\ &= (Y_u^1 - Y_u^0)(P(D_u = 1|I_u = 1) - P(D_u = 1|I_u = 0)). \end{aligned} \tag{29}$$

So by taking the expectation over the whole population we obtain the same estimator as in Eq. 27. This implies that this estimator works even when the structural equations are non-linear. In fact, IV can work in non-linear and even non-parametric models [82]. The difficulty mainly lies in computing the influence of I on Y, which is represented by the interventional distribution $P(Y|do(i))$:

$$\mathbb{E}[Y|do(i)] = \int_{d'} \mathbb{E}[Y|do(d')]P(d'|do(i))dd'. \tag{30}$$

You can refer to [82] for the heuristics that approximate the integral on the RHS of Eq. 30.

Two-stage Least Square (2SLS). As we have seen, the IV estimator in Eq. 27 is restrictive. In some cases, we might have to control a set of observed variables $\mathcal{S}$ to block the undesired back-door paths between the IV and the outcome variable so that the IV can be valid. These situations make it difficult or infeasible to apply the estimator in Eq. 27. We introduce 2SLS [6], a more general IV method. Fig. 6 shows an example for such a case where $\mathcal{S}$ denotes the set of confounders (e.g., whether a coupon can be found on Yelp for the restaurant) for the causal effect of whether a customer makes a review on the customer flow $I \rightarrow Y$. To make I valid, the back-door path $I \leftarrow \mathcal{S} \rightarrow Y$ has to be blocked. Besides, we may have more than one IVs for each treatment and

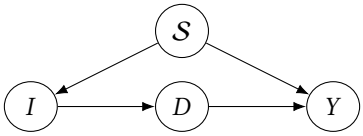


Fig. 6. Assuming that we can observe all the confounders $\mathcal{S}$ for the causal effect of whether a customer writes a review (I) on customer flow ($I \rightarrow Y$), we can perform 2SLS to estimate the causal effect of Yelp ratings on customer flow ($D \rightarrow Y$).

multiple treatment variables for the outcome. Assuming there is a set of treatment variables $\mathcal{D}$ and each treatment variable $D_{,j}$ has a set of instrumental variables $\mathcal{I}_{,j}$. In 2SLS, two regressions are carried out for learning the causal effects $(D_{,1} \rightarrow Y), ..., (D_{,j} \rightarrow Y), ...$:

(1) First stage: fit a function $\hat{D}_{,j} = f_{D_{,j}}(\mathcal{I}_{,j}, \mathcal{S}_{,j})$ for each treatment variable $D_{,j}$.
(2) Second stage: fit a function $Y = f_Y(\hat{\mathcal{D}}, \mathcal{S})$ where $\hat{\mathcal{D}}$ signifies the set of treatment variables. Then the coefficient on $D_{,j}$ is a consistent estimate of the average treatment effect (ATE) of the j-th treatment $D_{,j}$ on Y.

The intuition of 2SLS follows how we find a valid IV. In the first stage, we estimate how much a certain treatment $D_{,j}$ changes if we modify the relevant IV $I_{,j}$. In the second stage, we see how the changes in $D_{,j}$ caused by $I_{,j}$ would influence Y.

3.2.2 Front-door Criterion. The front-door criterion [99] enables us to learn a certain causal effect $D \rightarrow Y$ with unobserved confounders. With front-door criterion we condition on a set of variables $\mathcal{M}$ which satisfies the following three conditions:

(1) $\mathcal{M}$ blocks all the directed paths from D to Y.
(2) There are no unblocked back-door paths from D to $\mathcal{M}$; and
(3) D blocks all the back-door paths from $\mathcal{M}$ to Y.

In other words, we say that the set of variables $\mathcal{M}$ *mediates* the causal effect of D on Y. To fully illustrate the front-door criterion, we explain these conditions step by step. From the first condition, we decompose $D \rightarrow Y$ to a product of $D \rightarrow \mathcal{M}$ and $\mathcal{M} \rightarrow Y$ as:

$$P(Y|do(d)) = \int_{\boldsymbol{m}} P(Y|do(\boldsymbol{m}))P(\boldsymbol{m}|do(d))d\boldsymbol{m}, \tag{31}$$

where $\boldsymbol{m}$ denotes a vector of values for the mediators $\mathcal{M}$. The second condition means that there is no confounding bias for the causal effect of D on $\mathcal{M}$:

$$P(\boldsymbol{m}|do(d)) = P(\boldsymbol{m}|d). \tag{32}$$

The third condition allows us to infer $P(Y|do(\boldsymbol{m}))$ by:

$$P(Y|do(\boldsymbol{m})) = \int_{d'} P(Y|d', \boldsymbol{m})P(d'). \tag{33}$$

Then the interventional distribution corresponding to $D \rightarrow Y$ can be identified as:

$$P(Y|do(d)) = \int_{\boldsymbol{m}} P(\boldsymbol{m}|d) \sum_{d'} P(Y|d', \boldsymbol{m})P(d'). \tag{34}$$

Machine learning or traditional statistical techniques can be applied to estimate the probabilities on the RHS of Eq. 34 from observational data. In the running example, assuming that the set of variables $\mathcal{M}$ represents the ranking of a restaurant in the search results shown by Yelp. When the ranking is only decided by the Yelp rating, $(\mathcal{Z} \perp\!\!\!\perp \mathcal{M}|D, Y)$, $\mathcal{M}$ satisfies the front-door criterion

(a) $\mathcal{M}$ satisfies the front-door criterion. (b) $\mathcal{M}$ violates the front-door criterion.

Fig. 7. Two causal graphs where $\mathcal{M}$ satisfies and violates the front-door criterion

(Fig. 7a). However, when the ranking $\mathcal{M}$ is also determined by both the rating D and confounders $\mathcal{Z}$ (e.g. the restaurant category), then $\mathcal{M}$ is not a valid set of mediators (Fig. 7b). Different from the back-door criterion, the front-door criterion enables us to learn causal effects when some confounders are unobserved.

3.2.3 Regression Discontinuity Design. In some real-world problems, the treatment assignment may only depend on the value of a special feature, which is the *running variable* R. For example, the treatment is determined by whether its running variable is greater than a cut-off value r_0. The study of the causal effect of Yelp star rating R on the customer flow Y is a perfect example for such a case [5]. Yelp shows the rating of a restaurant rounded to the nearest half star in the search results. For example, restaurant u with average rating 3.26 and restaurant u' with 3.24 would be shown with 3.5 and 3.0 stars. Based on this fact, we can say $r_0 = 3.25$ is a cut-off which defines the treatment variable. Then for a restaurant with average rating $R \in [3, 3.5]$, we say it receives treatment ($D = 1$) when its rounded star rating is greater than its average rating ($R \geq r_0$). Otherwise, we say a restaurant is under control ($D = 0$). The intuition for Sharp Regression Discontinuity Design (Sharp RDD) [5, 20] is that the restaurants with average rating close to the cutoff $r_0 = 3.25$ are homogeneous w.r.t. the set of confounders. Therefore, the only thing can make a difference in their observed outcomes is the treatment. In other words, the treatments are randomly assigned to such restaurants, which lead to identification of the ATE. Formally, in Sharp RDD, we assume that the observed outcome is a function of the running variable as:

$$Y_u = f(R_u) + \tau D_u + \epsilon_u = f(R_u) + \tau \mathbb{1}(R_u \geq r_0) + \epsilon_{yu}, \tag{35}$$

where $f(\cdot)$ is a function which is continuous at r_0, τ is the ATE and ϵ_{yu} denotes the noise term. The choice of function $f(\cdot)$ can be flexible. But the risk of misspecification of $f(\cdot)$ exists. For example, Gelman and Imbens [46] pointed out that high-order polynomials can be misleading in RDD. In the Yelp study, this assumption can be supported by the argument that many customers' decision about which restaurant to go solely relies on the Yelp rating. In Fig. 8, we show a visualization of Sharp RDD for the running example study with synthetic data.

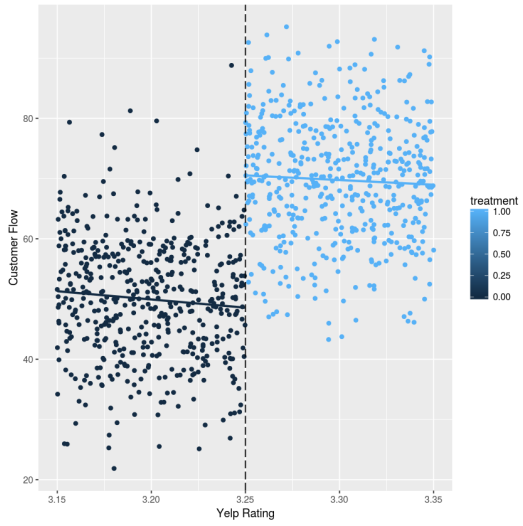
For many other real-world problems, however, it is not always the case where we can obtain a perfect cutoff value like the Yelp rating $r_0 = 3.25$ (stars) and the minimum drinking age $r_0 = 21$ (years old) [21]. The *Fuzzy RDD* method [8, 20] is developed to allow a more flexible relationship between the cutoff value and the treatment assignment.

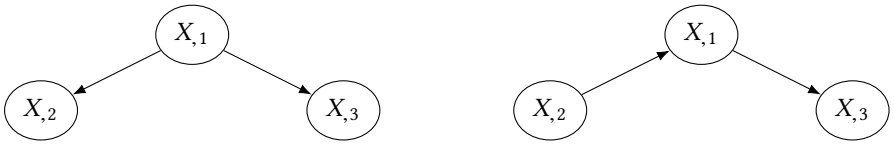
4 CAUSAL DISCOVERY: LEARNING CAUSAL RELATIONSHIPS

In the problem of learning causal relationship (a.k.a. causal discovery), it examines whether a certain set of causal relationships exists. We define the problem of learning causal relationships as:

Definition 15. Learning Causal Relationships. *Given N instances, $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$, to learn the causal relationships between any two variables, we aim to determine whether the j-th variable $X_{,j}$ would change if we modify the j'-th variable $X_{,j'}$ for all $j, j' = 1, ..., J$ and $j \neq j'$.*

Fig. 8. Visualization of a linear Sharp RDD model: it is fitted to some synthetic data generated from Eq. 35 with a linear $f(\cdot)$ for the Yelp study.





(a) A graph has $X_{,2} \perp\!\!\!\perp X_{,3}|X_{,1}$ (b) A graph has $X_{,2} \perp\!\!\!\perp X_{,3}|X_{,1}$

Fig. 9. Two example causal graphs that belong to an equivalence class

In the running example, learning causal relationships enable us to answer the questions such as: *Do other features such as location, menu prices, online ordering service also causally affect the customer flow? Is location a confounder for the causal effect of Yelp rating on customer flow?*

To learn causal relationships from data, researchers postulate that causality can be detected at least amongst statistical dependencies [107, 120]. An algorithm for learning causal relationships learns a family of causal graphs as candidates for the "gold standard" causal graph G. To evaluate the learned causal relationships, we often compare each of the learned causal graphs $\hat{G}$ with the ground-truth G. The concept of the *equivalence class* is important for comparing causal graphs.

Definition 16. Equivalence Class. *We say that two causal graphs G and G' belong to the same equivalence class iff every conditional independence that G has is also implied by G' and vise versa.*

Fig. 9a and 9b show two example causal graphs which belong to the same equivalence class. Both of them have the same set of conditional independence $\{X_{,2} \perp\!\!\!\perp X_{,3}|X_{,1}\}$.

A simple metric is the number of learned graphs that are equivalent to the ground truth G. In some previous work, we can also count the number of learned causal graphs G' that have G as a subgraph [26]. In [128], Shimizu et al. measured the distance between the adjacency matrix of a

learned causal graph ($\hat{\boldsymbol{A}}$) and that of the ground truth ($\boldsymbol{A}$) by the Frobenius norm:

$$dist(G, \hat{G}) = ||\boldsymbol{A} - \hat{\boldsymbol{A}}||_F = \sqrt{trace\left\{(\boldsymbol{A} - \hat{\boldsymbol{A}})^T(\boldsymbol{A} - \hat{\boldsymbol{A}})\right\}}, \tag{36}$$

where $trace(\boldsymbol{X})$ is the sum of diagonal elements of the squared matrix $\boldsymbol{X}$.

4.1 Learning Causal Relationships with i.i.d. Data

The Constraint-based Algorithms. This class of algorithms aims to learn a set of causal graphs which satisfy the conditional independences embedded in the data. These algorithms carry out statistical tests to verify if a candidate causal graph fits all the independences based on the *faithfulness* assumption [129] which can be formulated as:

Definition 17. Faithfulness. *Conditional independence between a pair of variables, $X_{,j} \perp\!\!\!\perp X_{,j'}|\mathcal{Z}$ for $X_{,j} \neq X_{,j'}, \mathcal{Z} \subseteq \mathcal{X} \setminus \{X_{,j}, X_{,j'}\}$, can be estimated from a dataset $\boldsymbol{X}$ iff $\mathcal{Z}$ d-separates $X_{,j}$ and $X_{,j'}$ in the causal graph $G = (\mathcal{V}, \mathcal{E})$ which defines the data-generation process for $\boldsymbol{X}$.*

Under this assumption, if data for the running example is generated according to the causal graph shown in Fig. 4, Yelp rating is a dependent of the customer flow ($D \not\perp\!\!\!\perp Y$). The challenge is mainly the computational expense. The number of possible causal graphs (DAGs) is super-exponential to the number of variables. Therefore, algorithms are proposed to reduce the number of tests.

The Peter-Clark (PC) Algorithm. The PC algorithm [129] is an example for the constraint-based algorithms. The PC algorithm works in a two-step fashion. First, it learns an undirected Bayesian network (*skeleton graph*) from data. In the second step, it detects the directions of the edges to return an equivalent class of causal graphs [77]. The first step starts with a fully connected Bayesian network and depth $q = 0$. Then for each ordered pair of connected variables $(X_{,j}, X_{,j'})$, it tests if the conditional independence $X_{,j} \perp\!\!\!\perp X_{,j'}|\tilde{\mathcal{Z}}$ is satisfied for each set $\tilde{\mathcal{Z}} \subseteq \mathcal{X} \setminus \{X_{,j}, X_{,j'}\}$ of size q. If the conditional independence is satisfied, it removes the edge $(X_{,j}, X_{,j'})$ and saves $\tilde{\mathcal{Z}}$ as the separating set of $(X_{,j}, X_{,j'})$. Once all such edges are removed, the depth is increased by one and this process continues till the number of neighbors for each variable is less than q. For the second step, please refer to [129] for the details of how the edge directions are determined in the PC algorithm.

Other similar constraint-based algorithms include the IC algorithm [101], and their variants [70, 77]. However, most standard statistical tests require Gaussian or multinomial distributions. To overcome these restrictions, novel conditional independence tests are proposed for more flexible distributions [42, 113, 122, 144]. To take unobserved confounders into consideration, algorithms are introduced to search through an extended space of causal graphs such as FCI and its extensions [30, 130]. But there are two main drawbacks of this family of algorithms. First, the faithfulness assumption can be too strong for data with limited sample size where independence tests may even contradict each other. Second, it cannot tell causal direction between two variables.

Score-based Algorithms. To relax the faithfulness assumption, another line of work replaces conditional independence tests with the goodness of fit tests. In particular, score-based algorithms learn causal graphs by solving the optimization problem below:

$$\hat{G} = \underset{G' \; DAG \; over \; \mathcal{X}}{\arg\max} \; Q(\boldsymbol{X}, G'), \tag{37}$$

where $Q(\boldsymbol{X}, G')$ returns the score of the causal graph G' given data $\boldsymbol{X}$ for the set of variables $\mathcal{X}$. Intuitively, low scores should be assigned to the causal graphs which embed incorrect conditional independence. For goodness of fit tests, two components need to be specified: the parametrization of structural equations and the score function that evaluates the possible causal graphs. The first component is the parameterization of structural equations. Conventionally, structural equations

are often assumed to be linear with additive Gaussian noise [26], which introduces parameters $\boldsymbol{\theta}$. Each structural equation describes how a variable is causally influenced by its parent variables and a noise term. The second component is a score function which maps a candidate causal graph to a scalar based given a certain parameterization of structural equations. The Bayesian Information Criterion (BIC) score [121] is the most widely adopted metric:

$$Q(\boldsymbol{X}, G') = \log P(\boldsymbol{X}|\hat{\boldsymbol{\theta}}, G') - \frac{J}{2}\log(N), \tag{38}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimation of the model parameters, J denotes the number of variable (dimension) and N signifies the number of instances. BIC score prefers causal graphs that can maximize the likelihood of observing the data with regularization on the number of parameters and the sample size. In [117], a similar score function is proposed based on maximum likelihood estimation with a different regularizer. Moreover, from the Bayesian perspective, with priors over causal graph structure and parameters, posteriors can be used to define scores. For example, Bayesian Dirichlet score [56] assumes Dirichlet prior on parameters for the multinomial distributions of variables. With the two components fixed, score of a certain causal graph for a given dataset is well defined. Then we focus on searching for the causal graphs which provide the best score for a given dataset. Generally speaking, searching for the causal graph with maximal score, also known as structural learning is both NP-hard and NP-complete [25, 27]. So, computationally, it is not feasible to score all possible causal graphs exhaustively. Therefore, heuristics such as Greedy Equivalence Search (GES) [26] and its extension, Fast GES (FGES) [112] are proposed to reach a locally optimal solution. Please refer to [114] for the TETRAD toolbox and this type of algorithms for learning causal relationships.

Algorithms for Non-Gaussian Models. Shimizu et al. [127] proposed the Linear Non-Gaussian Acyclic Model (LiNGAM) which enables us to distinct cause from effect between a pair of variables. In the matrix form, the LiNGAM model can be written as:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{39}$$

where $\boldsymbol{x}$, $\boldsymbol{A}$ and $\boldsymbol{\epsilon}$ denote the vector of features for an instance, the adjacency matrix for the causal graph (*connection strength matrix* [126]) and the vector of noise for an instance, respectively. Columns of both $\boldsymbol{x}$ and $\boldsymbol{A}$ are sorted according to the *causal order* ($k(j)$) of each variable, respectively. In the LiNGAM model, the task of learning causal relationships turns into estimating a strictly lower triangle matrix $\boldsymbol{A}$ which can determine a unique causal order $k(j)$ for each variable $X_{,j}$. We assume that the structural equations of the causal graph in Fig. **??** can be specified by the LiNGAM with a lower triangle matrix $\boldsymbol{A}$ below:

$$\begin{bmatrix} s \\ d \\ y \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1.2 & 0 & 0 \\ 0.8 & 1.3 & 0 \end{bmatrix} \begin{bmatrix} s \\ d \\ y \end{bmatrix} + \begin{bmatrix} \epsilon_s \\ \epsilon_d \\ \epsilon_y \end{bmatrix}, \tag{40}$$

then the causal order of the three variables s, d, y is 1, 2 and 3, respectively. For learning causal relationships in the LiNGAM model, Shimizu et al. proposed the ICA-LiNGAM algorithm [127]. In ICA-LiNGAM, the matrix $\boldsymbol{A}$ is estimated in two steps. First, we can rewrite Eq. 39 as:

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{\epsilon}, \tag{41}$$

where $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{A})^{-1}$. As each dimension of $\boldsymbol{\epsilon}$ is assumed to be independent and follows non-Gaussian distributions, Eq. 41 defines the ICA model for the LiNGAM. Thus we can apply ICA to obtain an estimate of $\boldsymbol{B}$. Then an initial estimate of $\boldsymbol{A}$, namely $\boldsymbol{A}'$, is computed based on $\boldsymbol{B}^{-1}$. Finally, to estimate the causal order $k(j)$ for each $X_{,j}$, permutations are applied to $\boldsymbol{A}'$ to obtain an estimate of $\boldsymbol{A}$ which is as close to a strictly lower triangle matrix as possible. A main downfall of ICA-LiNGAM

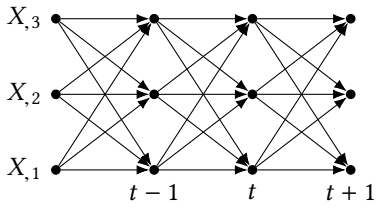


Fig. 10. An example of a chain causal graph for time series

is that ICA algorithms may converge to local optima. To guarantee the convergence to the global optima in a fixed number of steps, Shimizu et al. proposed the DirectLiNGAM algorithm [128], which also determines $\boldsymbol{A}$ through estimating the causal ordering of variables $k(j)$.

Recently, Additive Noise Models (ANMs) are proposed to relax the linear restriction on the relationships between variables and the distribution of noise [61, 62]. ANMs also help reduce the search space of causal graph as data normally does not admit two ANMs with conflicts in directions of causal effects [62, 107]. One step further, Post-nonlinear Models expand the functional space with non-linear relationships between the variables and the noise [143].

4.2 Learning Causal Relationships with Time Series Data

Time series is an important non-i.i.d. representation of data for many influential applications such as speech recognition [50], sleep stage classification [23], and quantitative trading [4]. As causality can be implied by the arrow of time [109], it can be defined in various ways for time series data. From a practical viewpoint, many researchers studied *Granger causality*[5] [49] as the approximation of real causality as it does not require a pre-defined causal model [14, 15, 38, 85]. In the matrix form, the linear Granger causality can be written as:

$$\boldsymbol{X}(t) = \boldsymbol{A}\boldsymbol{X}(t-1) + \boldsymbol{\epsilon}(t-1), \tag{42}$$

where the matrix $\boldsymbol{A}$ contains the temporal causal relations. Although Granger causality is merely temporal constrained statistical association, under the faithfulness assumption, Granger causality is a necessary condition for causation under the faithfulness assumption [107]. At the same time, some work tried to connect time series to the real causality. In [38], Eichler had a comprehensive discussion about how to define causality for time series data. For a guide of data preparation for learning causal relationships in time series data, please refer to [88]. We can represent a time series as a chain causal graph (Fig. 10) so that algorithms for i.i.d. data can be adapted to learn causal relationships in time series.

Constraint-based Algorithms for Time Series. Similar to the constraint-based algorithms for learning causal relationship in i.i.d. data. This class of algorithms for learning causal relationship in time series are based on the statistical independence tests. Researchers [29, 39] proposed to adapt FCI algorithm for time series. Peters et al. [105] proposed a more robust algorithm based on non-linear independent tests, known as TiMINo, which can avoid discovering incorrect causal relationships under model misspecification.

Algorithms for Non-Gaussian Time Series Models. Those algorithms for non-Gaussian SCMs can also be adapted for learning causal relationships in time series. They are also based on the asymmetry in cause-effect pairs in these models. For example, an auto-regressive LiNGAM is proposed to learn causal relationships [64].

[5]Here, the term Granger causality is also used to refer to its nonlinear variants.

With these well established algorithms, there are challenges of learning causal relationships in time series from the data perspective. Two issues can happen in the data collection process: the subsample problem and hidden time series. The subsample problem of time series data refers to the situation that only a low-resolution version of the original time series is available, for example, we only observe the time series every k time steps [131]. The key assumption for the recently proposed methods [47, 63, 110] is that there exists a *true timescale* or *causal frequency* at which we can discover proper causal graph structures with the highest confidence [33]. Furthermore, for learning causal relationship in time series data, hidden time series acts like unobserved confounders in i.i.d. data. Confounding bias can lead to wrong causal conclusions [131]. Geiger et al. [44] showed that causal relationships can be discovered under confounding bias with several assumptions.

5 CONNECTIONS TO MACHINE LEARNING

Machine learning techniques seek to infer the properties of probabilistic distributions of the observed data. In contrast, learning causality cares about the underlying data-generating process. The key to bridge them is to let machine learning models go beyond statistical associations [107]. In this section, we discuss the connections between learning causality and several machine learning problems: supervised and semi-supervised learning, domain adaptation and reinforcement learning. In particular, we explore two aspects: how causal knowledge can improve prediction accuracy? How machine learning approaches can help answer causal questions?

5.1 Supervised Learning and Semi-supervised Learning

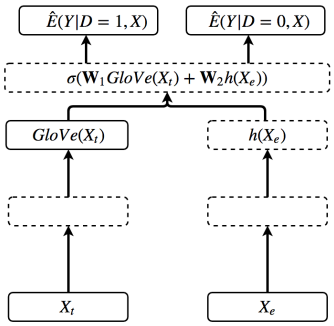
Supervised Learning. We can connect learning causality to supervised learning in two aspects: first, advanced methods can be leveraged for learning causality; second, from a data perspective, we can transform some problems of learning causality into supervised learning problem.

5.1.1 Advanced Methods for Learning Causal Effects. Based on the success of machine learning techniques, more advanced methods have been developed for learning causal effects. Here we cover several categories of widely used and recently proposed methods for learning causal effects: *improved traditional methods with neural networks*, *representation learning for confounders*, *learning heterogeneous causal effects with sparse models* and *ensembles*.

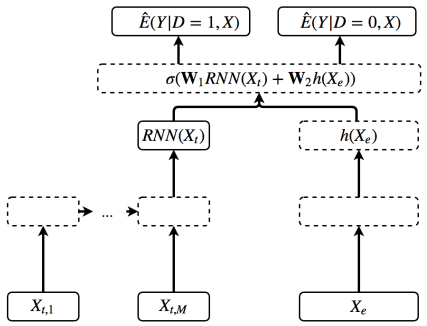
Learning Causal Effects with Neural Networks. Neural networks have shown their strength in predictive tasks for big data in various fields [50, 72, 90]. The most straightforward way to help learning causal effects with neural networks is to learn representations for features. Specially, in the study of the causal effect of forming a group on receiving loan in the microfinance platform Kiva [6], GloVe [103] and Recurrent Neural Networks (RNN) [90] are used to embed text features into a low-dimensional space. Here we denote the text features by $\boldsymbol{X}_t$ and the other features by $\boldsymbol{X}_e$ and assume the vocabulary size is M. We use σ as nonlinear functions such as ReLU [95]. Fig. 11 shows two neural networks which first learn representation for text features $\boldsymbol{X}_t$ and the other features $\boldsymbol{X}_e$ separately, and then concatenate them to infer the potential outcomes. In Fig. 11a and Fig. 11b, GloVe and RNN are applied to learn representations for the text features, respectively. In [108], Pham and Shen also proposed to improve the traditional methods by applying neural networks to estimate the probability distributions such as $\hat{P}(Y|D, \mathcal{X})$ and $\hat{P}(D|\mathcal{X})$.

Representation Learning for Confounders. A series of recent work for learning causal effects leverages representation learning for confounders instead of observed features. The core assumption is that we can learn representations for the confounders, which are considered to be a better approximation of the confounders than the features. It frees us from the assumption that there is no unobserved confounder. With specifically designed deep learning models such as the *Balancing*

[6] https://www.kiva.org/



(a) A neural network learns representations for the text features $\boldsymbol{X}_t$ and the other features $\boldsymbol{X}_e$ with GloVe and MLP, respectively.



(b) A neural network learns representations for the text features $\boldsymbol{X}_t$ and the other features $\boldsymbol{X}_e$ with RNN and MLP, respectively.

Fig. 11. Two neural networks learning representations for features $\boldsymbol{X}$. Representations for the text features $\boldsymbol{X}_t$ and the other features $\boldsymbol{X}_e$ are learned separately and concatenated to infer potential outcomes.

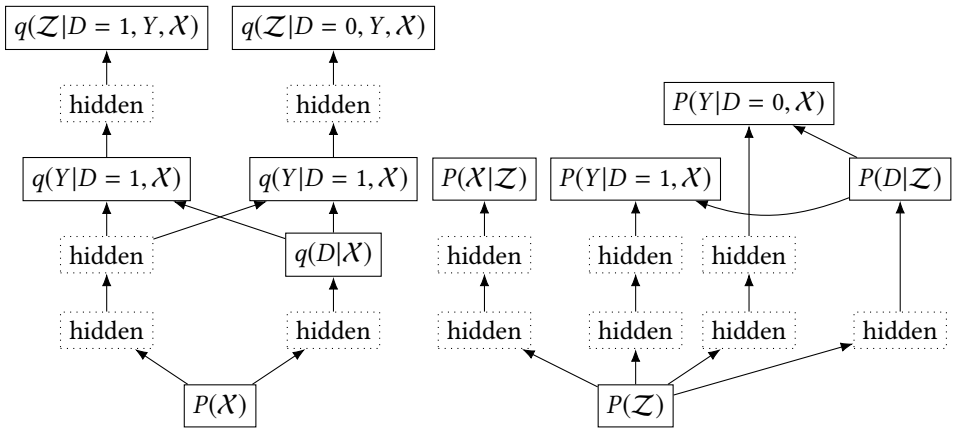
Counterfactual Regression [69], the *TARnet* [124] and the *Causal Effect Variational Autoencoder* (CEVAE) [84], we are able to learn a representation $\boldsymbol{z}_u$ parameterized by neural networks for each instance u based on $(\boldsymbol{x}_u, d_u, y_u)$. Here, we introduce the most recent method, namely the CEVAE, which represents advances in this directions of research.

With the recent advances in variational inference for deep latent variable models, Louizos et al. [84] proposed the CEVAE which learns a Gaussian representation for the confounders. The CEVAE consists of the inference network and the model network. The inference network is the encoder where we start with the observed data $(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{y})$ and embed it with latent representation $\boldsymbol{Z}$ as samples from a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_Z, \Sigma_Z)$. Then, the model network plays the role of decoder and reconstruct the data from the latent representation. The two neural networks are shown in Fig. 12. Those $q(\cdot)$ distributions are approximations for the posterior distributions parameterized by neural networks. Similar to the VAE [71] for predictive tasks, the CEVAE is trained through minimizing the KL divergence between the data and its reconstruction. So the loss function is formulated as:

$$\mathcal{L} = \sum_{u \in U} E_{q(\boldsymbol{z}_u|\boldsymbol{x}_u, d_u, y_u)}[\log P(\boldsymbol{x}_u, d_u|\boldsymbol{z}_u) + \log P(y_u|d_u, \boldsymbol{z}_u) + \log P(\boldsymbol{z}_u) - \log q(\boldsymbol{z}_u|\boldsymbol{x}_u, d_u, y_u)]. \quad (43)$$

Different from other representation learning methods, in the CEVAE, through minimizing the loss function in Eq. 43, we learn the Gaussian representation for $\boldsymbol{Z}$ and the function that maps $\boldsymbol{Z}$ to the counterfactual outcomes at the same time. The main difference between the CEVAE and the regular VAE is that, in CEVAE, there are two data points, $(\hat{y}_u^{d_u}, d_u, \hat{\boldsymbol{x}}_u)$ and $(\hat{y}_u^{1-d_u}, 1-d_u, \hat{\boldsymbol{x}}_u)$, reconstructed for each instance u. While the regular VAE is unsupervised, only reconstructing the features $\mathcal{X}$. Thus, counterfactual outcomes can be inferred once the neural networks in Fig. 12 are trained.

Learning Heterogeneous Causal Effects with Sparse Models. Different from many other studies, in the study of the efficacy for social programs and medical treatments by Imai and Ratkovic [65], there are many different strategies (treatments) available for each instance. Therefore, in this case, it is important to select the most effective treatment, design optimal treatment regime for each subgroup, test whether there exists heterogeneous subpopulations and adapt estimation of causal effects from one subpopulation to another. They formed the learning causal effects problem on such data as a variable selection problem, and proposed a regression adjustment model based



(a) The inference network (encoder). (b) The model network (decoder).

Fig. 12. The neural network structures of CEVAE. For each approximated distribution $q(\cdot)$, the corresponding hidden layer outputs necessary parameters for it. We learn a Gaussian distribution for the representation of confounders ($\mathcal{Z}$) from $\mathcal{X}$, D and Y.

on SVM [32] where sparsity is imposed by lasso regularization [135]. Formally, the model can be learned through the optimization problem:

$$\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}} = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})}{\arg\min} \sum_{u \in U} d_u |1 - y_u(\mu + \boldsymbol{\beta}^T \boldsymbol{z}_u + \boldsymbol{\gamma}^T \boldsymbol{v}_u)|^2 + \lambda_z ||\boldsymbol{\beta}||_1 + \lambda_v ||\boldsymbol{\gamma}||_1, \tag{44}$$

where $D_u \in \{0, 1, ..., K\}$ is the multi-valued treatment variable, the outcome variable $Y_u \in \{0, 1\}$ is binary, $\boldsymbol{z}_u$ and $\boldsymbol{v}_u$ are the vector form of two disjoint subsets of features $Z_u, V_u \subset X_u$, λ_v and λ_z are the hyperparameters that control the sparsity of the learned model. Other similar work on learning sparse models for ATE estimation includes [134, 140].

Learning Heterogeneous Causal Effects with Ensembles. Ensemble models can provide comparable performance to deep learning models in the supervised learning problems. With ensemble models, we train a series of weak classifiers on random subsamples of data (i.e., Bootstrapping) and make predictions by aggregating their outputs (i.e., Bagging). Variants of ensemble models are developed toward learning causal effects for either ATE or CATE. In [59], Hill proposed to apply Bayesian Additive Trees (BART) [28] for the estimation of CATE. In particular, for learning CATE, BART takes the treatment and features $(D, \mathcal{X})$ as input and output outcomes Y. BART is trained to estimate the distribution of outcomes based on treatment and features as $f(D, \mathcal{X}) = \mathbb{E}[Y|D, \mathcal{X}]$, where the function $f(D, \mathcal{X})$ returns the sum of the outputs of Q Bayesian subtrees $g(D, \mathcal{X}; T_j, M_j)$ as:

$$f(D, \mathcal{X}) = \sum_{j=1}^{Q} g(D, \mathcal{X}; T_j, M_j). \tag{45}$$

Then we can estimate the CATE for given $\mathcal{X} = \boldsymbol{x}$ as $\hat{\tau}(\boldsymbol{x}) = f(1, \boldsymbol{x}) - f(0, \boldsymbol{x})$. Each subtree is defined by the tree structure T_j and a set of b leaf nodes $M_j = \{\mu_{j1}, ..., \mu_{jb}\}$. An example of the subtree structure of BART is shown in Fig. 13, where each interior node (rectangle) acts like a binary classifier to send each input instance to its left or right child. Each leaf node has a parameter μ_{jk} which denotes the mean outcome of the instances classified to the k-th node of the j-th subtree. To justify the usage of BART, in [55, 59], authors mentioned that BART has shown to have several advantages over other models: BART is good at capture of non-linearities and discontinuity, it frees

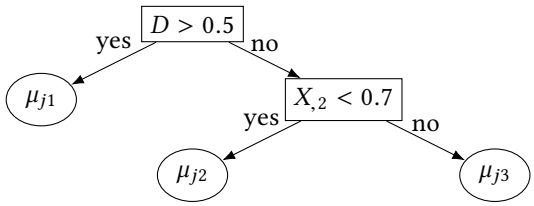


Fig. 13. A subtree $g(D, \mathcal{X}, T_j, M_j)$ in BART

researchers from hyperparameter tuning and it outputs posterior distribution of outcomes, which allows uncertainty quantification. In [55], Hahn et al. extended the usage of BART for the estimation of CATE and heterogeneity under significant regularization-induced confounding (RIC) [54]. RIC happens when $\mathbb{E}[Y|D, \mathcal{X}]$ heavily depends on the features $\mathcal{X}$ rather than the treatment D.

In [139], Wager and Athey proposed an ensemble model, the Causal Forest. It can output asymptotically normal and consistent estimation of CATE. Each tree in a causal forest model partitions the original covariate space recursively into subspaces such that each subspace is represented by a leaf. Function $L_j(\boldsymbol{x})$ returns which leaf of the j-th causal tree in the forest a certain instance belongs to, given its features $\mathcal{X} = \boldsymbol{x}$. Then each subspace of the j-th tree is considered as a RCT such that the CATE of a given $\boldsymbol{x}$ is identified and can be estimated by:

$$\hat{\tau}_j(\boldsymbol{x}) = \frac{1}{|U_l^1|} \sum_{u \in U_l^1} Y_u - \frac{1}{|U_l^0|} \sum_{u \in U_l^0} Y_u, \tag{46}$$

where $U_l^d = \left\{u | L_j(\boldsymbol{x}_u) = l, D_u = d\right\}$ refers to the subset of instances that belongs to the l-th leaf whose treatment is d. Then the causal forest simply outputs the average of the CATE estimation from the J trees as $\hat{\tau}(\boldsymbol{x}) = \frac{1}{J} \sum_j \hat{\tau}_j(\boldsymbol{x})$. It worths mentioning that there are studies dealing with the case where heterogeneous subpopulations cannot be identified by features. In particular, *principle stratification* [41, 138] is proposed to deal with such cases.

Methods for Non-i.i.d. Data In some applications, an instance's treatment or outcome can depend on those of other instances. For example, the customer flow of restaurants in the same area may amount to a constant. Therefore, in addition to the features, treatments and outcomes, there can be additional information describing the inter-dependence between instances. As we mentioned in Section 1, such inter-dependence can be a network of instances, time series of instances or temporal point process. Possible solutions for learning causal effects on such types of data include explicit modeling the interference [111] and disentangle instances with i.i.d. representations [51]. Besides this step, the rest is very similar to the methods introduced earlier in this subsection for i.i.d. data. For example, Rakesh et al. [111] developed a variant of CEVAE [84] for modeling a type of interference, namely the spillover effect. Learning causal effects on non-i.i.d. is still an open problem for data with other types of interference such as temporal inter-dependence.

5.1.2 Learning Causality as Classification. Now we discuss how supervised machine learning techniques can help us for learning causal effects and relationships. In addition, a recent proposed idea is to turn the problem of learning causal relationships into prediction from a data perspective. It turns out that the problem of learning causal relationships can be transformed into as a classical prediction problem once we label the data with causal relationships. In particular, suppose we are given labeled training data of the form $(c_1, a_1), ..., (c_N, a_N)$ where each c_i is a i.i.d. dataset $c_i = (\boldsymbol{X}_1, \boldsymbol{Y}_1), ..., (\boldsymbol{X}_{N_i}, \boldsymbol{Y}_{N_i})$ sampled from a joint distribution $P_i(\mathcal{X}, Y)$ and each dataset has an additional label $a_i \in (\rightarrow, \leftarrow)$ describing whether the dataset C_i corresponds to $\mathcal{X} \rightarrow Y$ or $Y \rightarrow \mathcal{X}$. The main

challenge of this research direction is to obtain the label of causal direction. For some special datasets, the causal relationships is naturally revealed [83]. On the other hand, we can leverage causal knowledge in various ways to improve the prediction performance of supervised learning models. A new approach, causal regularization [13, 125] is proposed to learn more interpretable and generalizable models.

Semi-supervised Learning (SSL). A machine learning problem can be either causal or *anti-causal* [120]. Anti-causal means that the label Y is the cause of the features $\mathcal{X}$. For example, in the handwritten digit recognition problem [78], the writer first determines which digit to write (Y). Then, the digit is written and is represented as a matrix of pixel values ($\mathcal{X}$). Therefore, it is an anti-causal problem [107]. Such an underlying causal structure between the features and the label has implications for many machine learning problems, especially for SSL. In SSL, the target is to improve the estimation of $P(Y|\mathcal{X})$ with additional unlabeled instances $\boldsymbol{x}_{n+1}, ..., \boldsymbol{x}_{n+m}$. These instances can provide some information of the marginal distribution $P(\mathcal{X})$. Janzing and Schölkopf [67] have proved that SSL indeed outperforms supervised learning in the anti-causal direction whenever $P(\mathcal{X})$ and $P(Y|\mathcal{X})$ are independent. The fact that SSL can only work in the anti-causal direction can help learn causal relationships [123].

5.2 Domain Adaptation

Domain adaptation [18, 34] studies how to adapt machine learning models trained in some domains to the others. One real-world application of domain adaptation is to improve prediction accuracy when we have plenty of labeled data for a source domain (e.g., Yelp review) but not for the target domain (e.g., Tripadvisor[7] review). Domain adaptation is naturally related to learning causality [120] by *invariant prediction in different domains*. Suppose we have observed data with a target variable Y^e and j predictor variables $\mathcal{X}^e = (X^e_{\cdot 1}, ..., X^e_{\cdot J})$ from different domains $e \in \xi = \{1, ..., E\}$ and the target is to predict the value of Y. Invariant prediction assumes that the conditional probability distribution $P(Y^e|\mathcal{X}^e_{\cdot \mathcal{Z}^*})$ is consistent for all domains for a subset of features $\mathcal{Z}^* \subseteq \{1, ..., J\}$. Formally, for all $e, f \in \xi$ and $\mathcal{Z}^*$, invariant prediction refers to:

$$P(Y^e|\mathcal{X}^e_{\cdot \mathcal{Z}^* = \mathcal{X}_{\mathcal{Z}^*}}) = P(Y^f|\mathcal{X}^f_{\cdot \mathcal{Z}^* = \mathcal{X}_{\mathcal{Z}^*}}). \tag{47}$$

If data for each domain is produced by an intervention on a predictor variable based on a underlying SCM, then we can make two conclusions: invariant prediction is satisfied and $\mathcal{Z}^*$ is the set of direct causes of the target variable Y. Based on this, for prediction, we can choose a valid subset $\mathcal{Z}^*$ that achieves the best performance in the training domains, as it is guaranteed to be optimal for domain generalization even when the marginal distribution of predictors are arbitrary in the test domains [107, 116]. Similar results for domain generalization have been obtained through a global balancing approach [73] and a causal feature selection method [98]. While for prediction under concept drift [141], where Eq. 47 is violated but the marginal distribution of predictors remain, causal knowledge also helps [76]. In addition, for causal discovery, one can detect causes for the target variable through searching for valid subsets $\mathcal{Z}^*$ that satisfy Eq. 47 over available domains. A variant [104] outputs the conjunction of the valid subsets as the set of causes for the target variable.

5.3 Reinforcement Learning

Reinforcement learning (RL) [133] is studied for solving sequential decision-making problems. The three key variables in RL are the action A, the state (of the world) Z and the reward Y. When an agent performs an action based on the current state, it will reach the next state and receive a reward. To describe the relationships between these variables, researchers adopt the *Markov decision process*,

[7] https://www.tripadvisor.com/

which means the next state Z_{t+1} depends only on the current state Z_t and action A_t and the reward of the next state Y_{t+1} is determined only by Z_t, Z_{t+1} and A_t. A RL model learns a *strategy* or *policy* $\pi(a, z) = P(A_t = a|Z_t = z)$ which determines which action to take given the current state Z_t. The objective is to maximize the sum of the rewards. In the running example, we can assume that the state Z_t represents the location of a restaurant, the action A_t can be moving to a certain place or staying at the same place and the reward is the customer flow Y. In each time step, the restaurant owner decides which action to take and then observe the customer flow. Then the owner will make decision for the next time step based on whether the customer flow increases or not. One key issue is the unobserved confounders for multi-armed bandits (MAB) [16]. Without knowing the underlying causal relationships, traditional MAB algorithms fail to outperform the strategy that randomly picks an action in each time step. Specifically, the Causal Thompson Sampling algorithm [16] is proposed to deal with unobserved confounders in MAB problems. The reward distributions of those arms that are not preferred by the current strategy can also be estimated through hypothetical interventions on the action (choice of arm). By doing this we can avoid confounding bias in estimating the causal effect of choosing an arm on the expected reward. To connect causality with RL, we view a strategy or a policy in RL as an intervention [107]. Given rewards of an observed strategy, we can utilize causal inference methods to predict rewards for another strategy, especially for Episodic RL (ERL) problems. ERL is a subclass of RL where the state is reset after a finite number of actions. ERL helps decision-making in a wide range of real-world applications such as the card game Blackjack and the advertisement placement [19]. One popular approach leverages IPTW (see Section 3) for reward prediction. In this approach, a treatment refers to an action and the strategy-specific propensity score is defined as the probability to perform the observed action given the observed state. Particularly, sequences of actions, states and rewards produced by running an observed strategy π are recorded in the data as $[(A_1(1), Z_1(1)), (A_2(1), Z_2(1)), \ldots], [(A_1(2), Z_1(2)), (A_2(2), Z_2(2)), \ldots], \ldots$. Then we can estimate the expected sum of rewards of a strategy $\tilde{\pi}$ with IPTW on the rewards observed for strategy π:

$$\hat{\xi}_{n,ERL} := \frac{1}{n} \sum_{i=1}^{n} Y(i) \frac{\prod_{j=1}^{K} \tilde{\pi}(A_j(i)|Z_j(i))}{\prod_{j=1}^{K} \pi(A_j(i)|Z_j(i))}, \tag{48}$$

where n and K denote the number of repentance of the strategy π in data and the number of time steps in each episode. In [19], an improved variant of this approach is also mentioned. Other recent work bridging RL and causality includes causal approaches for transfer learning in RL models [142] and data-fusion for online reinforcement learners [40].

6 CONCLUSIONS AND SOME OPEN PROBLEMS

Different from previous surveys, this work aims to solve the problem of learning causality under the big data setting where we have more data and less knowledge than the traditional causal studies. Although the methods in the existing literature may not directly address learning causality for such cases, they build the foundation of data-driven approaches for both learning causal effects and relationships. Another idea highlighted in this work is the connections between learning causality and machine learning. We aim to demonstrate that it is possible to leverage the connections between them in achieving better solutions for both causal and predictive problems. Moreover, machine learning models can benefit from exploiting learned causal knowledge in Section 5.

Fig. 14 shows a summary of all the contents covered in this survey. These two frameworks enable us to formulate problems of learning causality with mathematical languages. Then, we cover the two types of problems: learning causal effect (causal inference) and relationships (causal discovery) with data. The methods for learning causal effects with three types of data are presented: i.i.d. and

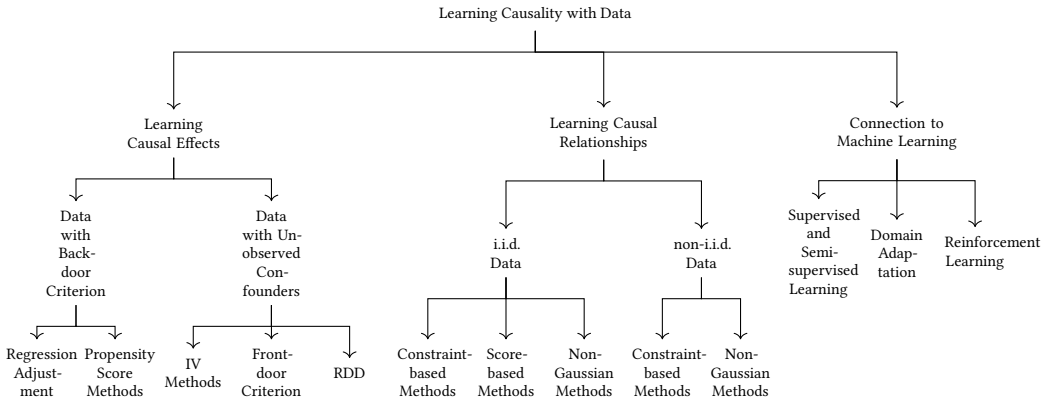


Fig. 14. Learning causality with data: a summary of the survey

non-i.i.d. data with back-door criterion satisfied and data with unobserved confounders. Next, we discuss how to learn causal relationships from two types of data: i.i.d. data and time series data. Finally, we discuss the connections between learning causality and machine learning. In particular, we discuss how we can help learning causality with methods for solving the three families of machine learning problems: supervised and semi-supervised learning, domain adaptation and reinforcement learning. We describe the connections from two perspectives: how learning causality yields better prediction in the machine learning problem? How machine learning techniques can be applied for learning causality?

Even existing research solved some problems of learning causality with data, more work needs to be done toward addressing the challenges that come with big data. From the data perspective, we present some open problems to review the great potential of learning causality with data:

- Study of heterogeneous groups: A dataset can come with heterogeneous groups. Though existing work addressed this problem by showing difference between groups [3] in terms of causal effects. But more efficient methods are needed for massive data. An extreme case of heterogeneous groups are anomalies. Although anomaly detection has been well studied in machine learning and data mining [1, 2, 24], detection of instances that are irregular in causal effects and relationships is still an open problem.
- Learning causality with imbalanced data: For example, a dataset for learning causal effect can come with very few treated units but much more controlled units. The problem of learning causality for such data remains to be solved.
- Learning causality with complex variables: An example of this type is the problem of learning causal effect of taking courses on employment. The treatment variable, namely the courses can be taken, can be formulated as a knowledge graph. So the problem turns into learning the causal effect of knowledge graphs on employment. The complex variable can also be the outcome or other variables and multiple complex variables can appear in the same problem.

APPENDIX

To facilitate development, evaluation and comparison of methods for learning causality, we introduce the open source data and algorithm index.

Data Index for Learning Causality We develop the open source data index for learning causality. It is available at Github (https://github.com/rguo12/awesome-causality-data). The datasets are categorized by the problem and the type of data.
Algorithm Index for Learning Causality The open source algorithm index for learning causality is at Github (https://github.com/rguo12/awesome-causality-algorithms). This index lists the algorithms mentioned in this survey. We also group the algorithms by the problem and the type of data.

REFERENCES

[1] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. 2010. Oddball: Spotting anomalies in weighted graphs. In *PAKDD*. 410–421.
[2] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: a survey. *DMKD* 29, 3 (2015), 626–688.
[3] Nazanin Alipourfard, Peter G Fennell, and Kristina Lerman. 2018. Using Simpson's Paradox to Discover Interesting Patterns in Behavioral Data. *arXiv preprint arXiv:1805.03094* (2018).
[4] Yakov Amihud. 2002. Illiquidity and stock returns: cross-section and time-series effects. *JFM* 5, 1 (2002), 31–56.
[5] Michael Anderson and Jeremy Magruder. 2012. Learning from the Cloud: Regression Discontinuity Estimates of the Effects of an Online Review Database. *EJ* 122, October (2012), 957–989.
[6] Joshua D Angrist and Guido W Imbens. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Amer. Statist. Assoc.* 90, 430 (1995), 431–442.
[7] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* 91, 434 (1996), 444–455.
[8] Joshua D Angrist and Victor Lavy. 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Q. J. Econ.* 114, 2 (1999), 533–575.
[9] Sinan Aral and Christos Nicolaides. 2017. Exercise contagion in a global social network. *Nature communications* 8 (2017), 14753.
[10] David Arbour, Dan Garant, and David Jensen. 2016. Inferring Network Effects from Observational Data. In *KDD*. 715–724.
[11] Susan Athey and Guido W Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *stat* 1050, 5 (2015).
[12] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav. Res.* 46, 3 (2011), 399–424.
[13] Mohammad Taha Bahadori, Krzysztof Chalupka, Edward Choi, Robert Chen, Walter F Stewart, and Jimeng Sun. 2017. Causal regularization. *arXiv preprint arXiv:1702.02604* (2017).
[14] Mohammad Taha Bahadori and Yan Liu. 2012. Granger causality analysis in irregular time series. In *SDM*. 660–671.
[15] Mohammad Taha Bahadori and Yan Liu. 2013. An examination of practical granger causality inference. In *SDM*. 467–475.
[16] Elias Bareinboim, Andrew Forney, and Judea Pearl. 2015. Bandits with unobserved confounders: A causal approach. In *NIPS*. 1342–1350.
[17] Elias Bareinboim and Jin Tian. 2015. Recovering Causal Effects from Selection Bias.. In *AAAI*. 3475–3481.
[18] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP*. 120–128.
[19] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *JMLR* 14, 1 (2013), 3207–3260.
[20] Donald T Campbell. 1969. Reforms as experiments. *Am. Psychol.* 24, 4 (1969), 409.
[21] Christopher Carpenter and Carlos Dobkin. 2009. The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *AEJ: Applied Economics* 1, 1 (2009), 164–82.
[22] Nancy Cartwright et al. 1994. Nature's Capacities and their Measurement. *OUP Catalogue* (1994).
[23] Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. 2018. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans. Neural Syst. Rehabil. Eng.* (2018).
[24] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)* 41, 3 (2009), 15.

[25] David Maxwell Chickering. 1996. Learning Bayesian networks is NP-complete. In *Learning from data*. Springer, 121–130.

[26] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *JMLR* 3, Nov (2002), 507–554.

[27] David M Chickering, Dan Geiger, David Heckerman, et al. 1994. *Learning Bayesian networks is NP-hard*. Technical Report. Citeseer.

[28] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. 2010. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* 4, 1 (2010), 266–298.

[29] Tianjiao Chu and Clark Glymour. 2008. Search for additive nonlinear time series causal models. *JMLR* 9, May (2008), 967–991.

[30] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* (2012), 294–321.

[31] Thomas D Cook, Donald Thomas Campbell, and William Shadish. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston.

[32] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[33] David Danks and Sergey Plis. 2013. Learning causal structure from undersampled time series. (2013).

[34] Hal Daumé III. 2009. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815* (2009).

[35] Rajeev H Dehejia and Sadek Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* 94, 448 (1999), 1053–1062.

[36] Imme Ebert-Uphoff and Yi Deng. 2012. Causal discovery for climate research using graphical models. *JCLI* 25, 17 (2012), 5648–5665.

[37] Andrew C Eggers, Ronny Freier, Veronica Grembi, and Tommaso Nannicini. 2018. Regression discontinuity designs based on population thresholds: Pitfalls and solutions. *Am. J. Pol. Sci.* 62, 1 (2018), 210–229.

[38] Michael Eichler. 2012. Causal inference in time series analysis. *Causality: Statistical perspectives and applications* (2012), 327–354.

[39] Doris Entner and Patrik O Hoyer. 2010. On causal discovery from time series data using FCI. *PGM* (2010), 121–128.

[40] Andrew Forney, Judea Pearl, and Elias Bareinboim. 2017. Counterfactual Data-Fusion for Online Reinforcement Learners. In *ICML*. 1156–1164.

[41] Constantine E Frangakis and Donald B Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58, 1 (2002), 21–29.

[42] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2008. Kernel measures of conditional dependence. In *NIPS*. 489–496.

[43] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *Am. J. Epidemiol.* 173, 7 (2011), 761–767.

[44] Philipp Geiger, Kun Zhang, Bernhard Schoelkopf, Mingming Gong, and Dominik Janzing. 2015. Causal inference by identification of vector autoregressive processes with hidden components. In *ICML*. 1917–1925.

[45] Andrew Gelman. 2011. Causality and statistical learning. *AJS* 117, 3 (2011), 955–966.

[46] Andrew Gelman and Guido Imbens. 2018. Why high-order polynomials should not be used in regression discontinuity designs. *J. Bus. Econ. Stat.* (2018), 1–10.

[47] Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. 2015. Discovering temporal causal relations from subsampled data. In *ICML*. 1898–1906.

[48] Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. 2017. Causal discovery from temporally aggregated time series. In *UAI*, Vol. 2017.

[49] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* (1969), 424–438.

[50] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *ICASSP*. 6645–6649.

[51] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD*. 855–864.

[52] Xing Sam Gu and Paul R Rosenbaum. 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. *J. Comput. Graph Stat.* 2, 4 (1993), 405–420.

[53] Ruocheng Guo, Jundong Li, and Huan Liu. 2018. INITIATOR: Noise-contrastive Estimation for Marked Temporal Point Process.. In *IJCAI*. 2191–2197.

[54] P Richard Hahn, Carlos M Carvalho, David Puelz, Jingyu He, et al. 2018. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.* 13, 1 (2018), 163–182.

[55] P Richard Hahn, Jared S Murray, and Carlos Carvalho. 2017. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *arXiv preprint arXiv:1706.09523* (2017).

[56] David Heckerman, Dan Geiger, and David M Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20, 3 (1995), 197–243.

[57] David Heckerman, Christopher Meek, and Gregory Cooper. 2006. A Bayesian approach to causal discovery. In *Innovations in Machine Learning*. Springer, 1–28.

[58] Miguel Ángel Hernán, Babette Brumback, and James M Robins. 2000. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* (2000), 561–570.

[59] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph Stat.* 20, 1 (2011), 217–240.

[60] Keisuke Hirano, Guido W Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 4 (2003), 1161–1189.

[61] Patrik O Hoyer, Aapo Hyvarinen, Richard Scheines, Peter L Spirtes, Joseph Ramsey, Gustavo Lacerda, and Shohei Shimizu. 2012. Causal discovery of linear acyclic models with arbitrary distributions. *arXiv preprint arXiv:1206.3260* (2012).

[62] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. 2009. Nonlinear causal discovery with additive noise models. In *NIPS*. 689–696.

[63] Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. 2016. Causal Discovery from Subsampled Time Series Data by Constraint Optimization. In *PGM*. 216–227.

[64] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. 2010. Estimation of a structural vector autoregression model using non-gaussianity. *JMLR* 11, May (2010), 1709–1731.

[65] Kosuke Imai, Marc Ratkovic, et al. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7, 1 (2013), 443–470.

[66] Guido W Imbens. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* 86, 1 (2004), 4–29.

[67] Dominik Janzing and Bernhard Schölkopf. 2015. Semi-supervised interpolation in an anticausal learning scenario. *JMLR* 16 (2015), 1923–1948.

[68] Marshall M Joffe, Thomas R Ten Have, Harold I Feldman, and Stephen E Kimmel. 2004. Model selection, confounder control, and marginal structural models: review and new applications. *Am. Stat.* 58, 4 (2004), 272–279.

[69] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *ICML*. 3020–3029.

[70] Markus Kalisch and Peter Bühlmann. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *JMLR* 8, Mar (2007), 613–636.

[71] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.

[73] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. 2018. Stable Prediction across Unknown Environments. In *KDD*. 1617–1626.

[74] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. In *KDD*. 265–274.

[75] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* (1986), 604–620.

[76] Virgile Landeiro and Aron Culotta. 2016. Robust text classification in the presence of confounding bias. In *AAAI*. 186–193.

[77] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, and Huawen Liu. 2015. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *arXiv preprint arXiv:1502.02454* (2015).

[78] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. 1990. Handwritten digit recognition with a back-propagation network. In *NIPS*. 396–404.

[79] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 94.

[80] Jundong Li, Harsh Dani, Xia Hu, Jiliang Tang, Yi Chang, and Huan Liu. 2017. Attributed network embedding for learning in a dynamic environment. In *CIKM*. 387–396.

[81] Jundong Li, Osmar R Zaïane, and Alvaro Osornio-Vargas. 2014. Discovering statistically significant co-location rules in datasets with extended spatial objects. In *DaWaK*. 124–135.

[82] Qi Li and Jeffrey Scott Racine. 2007. *Nonparametric econometrics: theory and practice.* Princeton University Press.

[83] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. 2017. Discovering causal signals in images. In *CVPR 2017*.

[84] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *NIPS*. 6446–6456.

[85] Aurelie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. 2009. Grouped graphical Granger modeling methods for temporal causal modeling. In *KDD*. 577–586.

[86] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* 23, 19 (2004), 2937–2960.
[87] HernÃąn MA and Robins JM. forthcoming. *Causal Inference.* CRC Boca Raton, FL:.
[88] Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass* 13, 1 (2018), e12470.
[89] Subramani Mani and Gregory F Cooper. 2000. Causal discovery from medical textual data.. In *Proceedings of the AMIA Symposium*. 542.
[90] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*.
[91] Kewei Ming and Paul R Rosenbaum. 2000. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* 56, 1 (2000), 118–124.
[92] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *JMLR* 17, 1 (2016), 1103–1204.
[93] Stephen L Morgan and Jennifer J Todd. 2008. 6. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects. *Sociol. Methodol.* 38, 1 (2008), 231–282.
[94] Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference.* Cambridge University Press.
[95] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*. 807–814.
[96] Jersey Neyman. 1923. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10 (1923), 1–51.
[97] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. 2013. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* 381, 9875 (2013), 1371–1379.
[98] Michael J Paul. 2017. Feature Selection as Causal Inference: Experiments with Text Classification. In *CoNLL*. 163–172.
[99] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.
[100] Judea Pearl. 2009. Causal inference in statistics: An overview *. *Stat. Surv.* 3 (2009), 96–146.
[101] Judea Pearl. 2009. *Causality.* Cambridge university press.
[102] Judea Pearl. 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016* (2018).
[103] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.
[104] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. Series B Stat. Methodol.* 78, 5 (2016), 947–1012.
[105] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2013. Causal inference on time series using restricted structural equation models. In *NIPS*. 154–162.
[106] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference - Foundations and Learning Algorithms.* The MIT Press, Cambridge, MA, USA.
[107] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms.* MIT press.
[108] Thai T Pham and Yuanyuan Shen. 2017. A Deep Causal Inference Approach to Measuring the Effects of Forming Group Loans in Online Non-profit Microfinance Platform. *arXiv preprint arXiv:1706.02795* (2017).
[109] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. 2014. Seeing the arrow of time. In *CVPR*. 2035–2042.
[110] Sergey Plis, David Danks, Cynthia Freeman, and Vince Calhoun. 2015. Rate-agnostic (causal) structure learning. In *NIPS*. 3303–3311.
[111] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. Linked Causal Variational Autoencoder for Inferring Paired Spillover Effects. *arXiv preprint arXiv:1808.03333* (2018).
[112] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 2017. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics* 3, 2 (2017), 121–129.
[113] Joseph D Ramsey. 2014. A scalable conditional independence test for nonlinear, non-Gaussian data. *arXiv preprint arXiv:1401.5031* (2014).
[114] Joseph D Ramsey and Daniel Malinsky. 2016. Comparing the Performance of Graphical Structure Learning Algorithms with TETRAD. *arXiv preprint arXiv:1607.08110* (2016).
[115] James M Robins, Miguel Angel Hernan, and Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology.

[116] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. 2015. A Causal Perspective on Domain Adaptation. *stat* 1050 (2015), 19.

[117] Teemu Roos, Tomi Silander, Petri Kontkanen, and Petri Myllymaki. 2008. Bayesian network structure learning using factorized NML universal models. In *ITA Workshop, 2008*. 272–276.

[118] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[119] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 5 (1974), 688.

[120] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471* (2012).

[121] Gideon Schwarz et al. 1978. Estimating the dimension of a model. *Ann. Stat.* 6, 2 (1978), 461–464.

[122] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.* (2013), 2263–2291.

[123] Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. 2015. Inference of cause and effect with unsupervised inverse regression. In *AISTATS*. 847–855.

[124] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*. 3076–3085.

[125] Zheyan Shen, Peng Cui, Kun Kuang, and Bo Li. 2017. On Image Classification: Correlation vs Causality. *arXiv preprint arXiv:1708.06656* (2017).

[126] Shohei Shimizu. 2014. LiNGAM: non-Gaussian methods for estimating causal structures. *Behaviormetrika* 41, 1 (2014), 65–98.

[127] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A linear non-Gaussian acyclic model for causal discovery. *JMLR* 7, Oct (2006), 2003–2030.

[128] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *JMLR* 12, Apr (2011), 1225–1248.

[129] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. 2000. *Causation, prediction, and search*. MIT press.

[130] Peter Spirtes, Christopher Meek, and Thomas Richardson. 1995. Causal inference in the presence of latent variables and selection bias. In *UAI*. 499–506.

[131] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, Vol. 3. 3.

[132] Peter M Steiner, Yongnam Kim, Courtney E Hall, and Dan Su. 2017. Graphical models for quasi-experimental designs. *Sociological Methods & Research* 46, 2 (2017), 155–188.

[133] Richard S Sutton and Andrew G Barto. 1998. *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge.

[134] Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. 2014. A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* 109, 508 (2014), 1517–1532.

[135] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* (1996), 267–288.

[136] Panos Toulis, Alexander Volfovsky, and Edoardo M Airoldi. 2018. Propensity score methodology in the presence of network entanglement between treatments *. *arXiv preprint arXiv:1801.07310* (2018).

[137] Mark J Van Der Laan and Daniel Rubin. 2006. Targeted maximum likelihood learning. *Int. J. Biostat.* 2, 1 (2006).

[138] Tyler J VanderWeele. 2011. Principal stratification–uses and limitations. *Int. J. Biostat.* 7, 1 (2011), 1–14.

[139] Stefan Wager and Susan Athey. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* just-accepted (2017).

[140] Herbert I Weisberg and Victor P Pontes. 2015. Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical trials* 12, 4 (2015), 357–364.

[141] Gerhard Widmer and Miroslav Kubat. 1996. Learning in the presence of concept drift and hidden contexts. *Machine learning* 23, 1 (1996), 69–101.

[142] Junzhe Zhang and Elias Bareinboim. 2017. Transfer learning in multi-armed bandit: a causal approach. In *AMMAS*. 1778–1780.

[143] Kun Zhang and Aapo Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. In *UAI*. 647–655.

[144] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2012. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775* (2012).

[145] Dawei Zhou, Jingrui He, Hongxia Yang, and Wei Fan. 2018. SPARC: Self-Paced Network Representation for Few-Shot Rare Category Characterization. In *KDD*. 2807–2816.

Received ; revised ; accepted