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, Department of Electrical and Computer Engineering, Computer Science & School of Data Science, University of Virginia, USA

P. RICHARD HAHN, Department of Mathematics and Statistics, Arizona State University

HUAN LIU, Computer Science and Engineering, Arizona State University

This work considers the question of how convenient access to copious data impacts our ability to learn causal effects and relations. In what ways is learning causality in the era of big data different from – or the same as – the traditional one? To answer this question, this survey provides a comprehensive and structured review of both traditional and frontier methods in learning causality and relations along with the connections between causality and machine learning. This work points out on a case-by-case basis how big data facilitates, complicates, or motivates each approach.

CCS Concepts: • **Computing methodologies** → *Artificial intelligence*; • **Mathematics of computing** → *Probability and statistics*; **Probability and statistics**;

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 differences between statistical associations and causations. For example, when the temperatures are hot, the owner of an ice cream shop may 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 cause

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, Department of Electrical and Computer Engineering, Computer Science & School of Data Science, University of Virginia, USA, Charlottesville, VA, 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.

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 [114]. Historically, learning causality has been studied in a myriad of high-impact domains including education [36, 63, 65, 86], medical science [101, 109], economics [74], epidemiology [64, 99, 127], meteorology [38], and environmental health [94]. 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 [63]. Taking the *randomized controlled trials* as an example [32], 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 group and the un-treated (control) groups are equivalent in all relevant respects, while ruling out the influence of any other factors. Then, the impact of the drug on some health outcome — say, the duration of a migraine — 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 era of big data. As an example, consider that the possibility of unmeasured confounders might be mitigated by the fact that a large number of features can be measured. So, first, we aim to answer causal questions using big data. For example, do positive Yelp[1] reviews drive customers to restaurants, or do they merely reflect popularity? This causal question can be addressed with data from an extensive database of Yelp. Second, answering causal questions with big data leads to some unique new problems. For example, public databases or data collected via web crawling or application program interfaces (APIs) are unprecedentedly large, we have little intuition about what types of bias a dataset can suffer from — the more plentiful data makes it 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 (e.g., prediction). Perhaps the most notable example is the high-dimensionality of modern data [92], such as text data [73].

Efforts have been made to the intersection between big data and causality. Examples include but are not limited to those discussed in [8, 39, 56, 104, 148]. The goal of this survey is to provide a comprehensive and structured review of both traditional and frontier methods in learning causality, as well as discussion about some open problems. We do not assume that a target audience may be familiar with learning causality.

1.1 Overview and Organization

Broadly, machine learning tasks are either *predictive* or *descriptive* in nature. But beyond that we may want to understand something *causal*, imagining 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 specific variables (features or the label) change if we manipulate the value of another specific 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* and *causal discovery* questions, respectively [48, 119]. For learning causal effects, 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 for which we observe the response as *the outcome*, respectively. One typical example is that *how much do hot temperatures raise ice cream sales*. For learning causal relations, researchers attempt to determining whether there exists a causal relationship between a variable and another. In our temperature and

---

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

ice cream example, it is clear that ice cream sales do not cause high temperatures, but in other examples it may not be clear. For example, we may be interested in investigating the question like *whether a genetic disposition towards cancer should be responsible for individuals taking up smoking?*

In this survey, we aim to provide a comprehensive review on how to learn causality from massive data. Below, we present an outline of the topics that are covered in this survey. First, in Section 2, we introduce the preliminaries of learning causality from data for both causal inference or causal discovery. We focus on the two formal frameworks, namely *structural causal models* [112] and the *potential outcome framework* [108, 131]. Section 3 focuses on the methods that are developed for the problem of learning causal effects (causal inference). Based on different types of data, these methods fall into three categories: methods for data without and with unobserved confounders, and advanced methods for big data. In Section 4, the widely used methods for learning causal relations are discussed. After introducing traditional methods, we describe advanced methods addressing special challenges in big data. Afterwards, in Section 5, we discuss recent research that connects learning causality and machine learning. We examine how the research in three subareas, supervised and semi-supervised learning, domain adaptation and reinforcement learning, can be connected to learning causality.

1.2 Data for Learning Causality

We discuss data for learning causality[2]. A comprehensive introduction of such data can be found in [24]. Although *interventional data* and a mixture of interventional and observational data are used in the literature [61, 79, 143, 161], we focus on *observational data* in this survey to take advantage of pervasive big data. In observational data, the value of a variable is determined by its causes. In contrast, in interventional data, there exists at least one variable whose value is set through intervention. For example, to study the causal effect of Yelp rating on customer flows of restaurants, we can either use existing records (observational data) or collect interventional data via manipulating the ratings of restaurants.

The ground truth used to train or evaluate causal learning methods often includes causal effects or causal relations. Ground truth of average causal effects can be obtained via randomized experiments. For example, the average effect of a new feature in a recommendation system is often estimated through an A/B test [161]. However, it is often not possible to collect ground truth of individual causal effects via randomized experiments as it requires knowledge of the *counterfactual* [113]. In practice, such ground truth is often acquired through simulations based on domain knowledge [77, 97]. Ground truth of causal relations is often obtained through prior knowledge (e.g., certain mutations of genes can cause diseases).

Data for Learning Causal Effects. Here, we review the types of data, the problems that can be studied if the data is given, and the methods that can provide practical solutions. We list three types of data for learning causal effects. First, a standard dataset for learning causal effects (X, t, y) includes feature matrix X, a vector of treatments t and outcomes y. We are particularly interested in the causal effect of one variable t (treatment) on another variable y (outcome). For the second type, there is auxiliary information about inter-dependence or interference between units such as links or temporal inter-dependencies between different data units, represented by a matrix A. Examples of this type of data include attributed networks [93, 165], time series [40], and marked temporal point process [54]. Moreover, the third type of data often comes with unobserved confounders, we need the help of special causal variables, including the *instrumental variable (IV)*, the *mediator*, or the *running variable*. These variables are defined by causal knowledge, thus specific methods can be applied for learning causal effect for such types of data.

---

[2]The data and algorithm indexes for learning causality are covered in the Appendix

Problems	Data	Example Datasets	Methods
Learning causal effects	Datasets contain features, treatment and outcome (X, t, y).	IHDP, Twins , Jobs	Regression adjustment, Propensity score, Covariate balancing, Machine learning based
	Datasets with features, treatment, outcome and special variable(s): (X, t, y, z)	1980 Census Extract, CPS Extract	IV methods, Front-door criterion, RDD, Machine learning based
Learning causal relations.	Multivariate data with causal relations, denoted by X with a causal graph G, including bivariate data with causal direction.	Abscisic Acid Signaling Network , Weblogs SIDO	Constraint-based, Score-based methods, Algorithms for FCMs.
	Multivariate time series $\{[x_{,1}(l), \dots, x_{,J}(l)]\}_{l=1}^{L}$ with a causal graph G	PROMO	

Table 1. Overview of this work in terms of the problems, data and methods.

Data for Learning Causal Relations. We describe two types of data for learning causal relations (causal discovery). The first type is the multivariate data X along with a ground truth causal graph G for evaluation, with which we learn the causal graph. A special case is the bivariate data and the task reduces to distinguishing the cause from the effect [104]. The causal graph is often defined by prior knowledge and could be incomplete. The second type of data for learning causal relations is the multivariate time series data which also comes with a ground truth causal graph. The task is to learn causal relations among the variables [50]. Although the ground-truth causal graph is often unique [50, 119], many methods output a set of candidate causal graphs.

1.3 Previous Work and Contributions

There are a number of other comprehensive surveys in the area of causal learning. Pearl [112] aims to convey the fundamental theory of causality based on the structural causal models. Gelman [48] provides high-level opinions about the existing formal frameworks and problems for causal learning. Mooji et al. [104] focus on learning causal relations for bivariate data. Spirtes and Zhang [146] summarize methods for learning causal relations on both i.i.d. and time series data with a focus on several semi-parametric score based methods. Athey and Imbens [10] describe 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

We consider a study of how Yelp ratings influence potential restaurant customers [3]. Yelp is a website where customers can share their reviews of a certain goods and services. 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 assume that these variables are binary. A restaurant receives the treatment $t = 1$ if its rating is above some threshold; otherwise, it is under control treatment $t = 0$. For the outcome, $y = 1$ means a restaurant is completely booked and $y = 0$ means it is not.

Table 2. Nomenclature

Nomenclature		
Terminology	Alternatives	Explanation
causality	causal relation, causation	causal relation between variables
causal effect		the strength of a causal relation
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 relations	causal discovery causal learning causal search	inferring 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

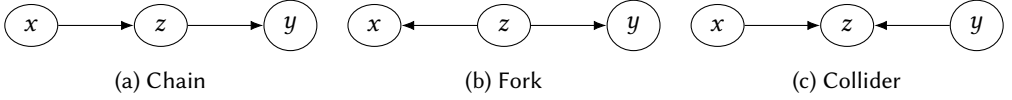(a) Chain   (b) Fork   (c) Collider

Fig. 1. Three typical DAGs for conditional independence

2 PRELIMINARIES

Here, we present the preliminaries for two fundamental frameworks: structural causal models and the potential outcome framework. To formulate causal knowledge, we need *causal models*. A causal model is a mathematical abstraction that quantitatively describes the causal relations between variables. First, causal assumptions or prior causal knowledge can be represented by an incomplete causal model. Then, what is missing can be learned from data. The two most well-known causal models are the structural causal models (SCMs) [113] and the potential outcome framework [108, 131]. They are considered as the foundations because they enable a consistent representation of prior causal knowledge, assumptions, and estimates.
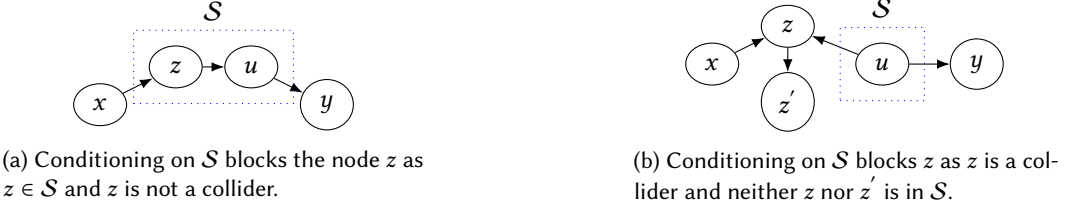
We present the terminologies and notations that are used throughout this survey. Table 2 displays key nomenclature. In this survey, a lowercase letter, say x, denotes a specific value of a corresponding random variable (or RV), X. Bold lowercase letters denote vectors or sets (e.g., $\boldsymbol{x}$) and bold uppercase letters signify matrices (e.g., $\boldsymbol{X}$). Calligraphic uppercase letters such can signify special sets such as sets of nodes $\mathcal{V}$ and edges $\mathcal{E}$ in a graph G. $\boldsymbol{X}$ and $\boldsymbol{x}_i$ present features for all instances and that for the i-th instance, respectively. Without specification, the subscripts denote the instance and the dimension. For example, $\boldsymbol{x}_i$ denotes features of the i-th instance and $x_{*,j}$ signifies the j-th feature. The letter t will be used to denote the treatment variable; in this work, it is often assumed to be binary and univariate. The letter y denotes the outcome variable. We use the subscript and superscript of y to signify the instance and the treatment it corresponds to. When the treatment is binary, y_i^1 denotes the outcome when the instance i is treated ($t_i = 1$). The letter τ denotes various treatment effects, defined as a change in the outcome variable for different levels of treatment.

2.1 Structural Causal Models

Structural causal models (SCMs) provide a comprehensive theory of causality [112]. An SCM often consists of two components: the *causal graph* (causal diagram) and the *structural equations*.
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.

Definition 1. 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 of x on y.*

(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 of z being blocked by conditioning on $\mathcal{S}$

(a) An SCM without intervention.

(b) An SCM under the intervention $do(t)$.

Fig. 3. SCMs without and under intervention $do(t')$ for the Yelp example, where x, t and y denote restaurant category, Yelp rating and customer flow.

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, as is common in the field, 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 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 *dependency-separation* (d-separation) based on the definition of *blocked* path. Fig. 1 shows three typical DAGs. 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. When z is a *collider* node (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 a collider (Fig. 1c), conditioning on z introduces an association between x and y, i.e., $x \perp\!\!\!\perp y$ and $x \not\perp\!\!\!\perp y|z$. Generally, we say conditioning on a set of nodes $\mathcal{Z}$ blocks a path p iff there exists at least one node $z \in \mathcal{Z}$ in the path p.

Definition 2. Blocked. *We say a node z is blocked by conditioning on a set of nodes $\mathcal{S}$ if one of the two conditions is satisfied: (1) $z \in \mathcal{S}$ and z is not a collider node (Fig. 2a); (2) z is a collider node, $z \notin \mathcal{S}$ and no descendant of z is in $\mathcal{S}$ (Fig. 2b).*

With this definition, we say a set of variables $\mathcal{S}$ d-separates two variables x and y iff $\mathcal{S}$ blocks all paths between them. d-separation plays a crucial role in explaining causal concepts. The *Causal Markovian condition* is often assumed in SCMs, which means we can factorize the joint distribution represented by a *Markovian* SCM of variables $\mathcal{V} = \{x_{*,1}, ..., x_{*,J}\}$ with:

$$P(x_{*,1}, ..., x_{*,J}) = \prod_j P(x_{*,j}|\boldsymbol{Pa}_{*,j}, \epsilon_{*,j}), \tag{1}$$

where $\boldsymbol{Pa}_{*,j}$ denotes the set of parent variables of $x_{*,j}$, each of which has an arrow in $x_{*,j}$. Moreover, $\epsilon_{*,j}$ is the noise which represents the causal effect of unobserved variables on $x_{*,j}$. Here, we introduce the key concepts of learning causality through a toy SCM which embeds causal knowledge for the Yelp example [3]. In Fig. 3a, there are three random variables, i.e., the restaurant category x (*confounder*), Yelp star rating t (treatment) and customer flow y (outcome). The three directed edges

represent the three causal effects: (1) Restaurant category influences its Yelp rating. For example, the average rating of fast food restaurants is lower than that of high-end restaurants. (2) Restaurant category also influences its customer flow. For example, the average customer flow of fast food restaurant is higher than that of high-end restaurants. (3) Yelp rating of a restaurant influences its customer flow.

Structural Equations. Given a causal graph along with a set of structural equations, we can specify the causal effects signified by the directed edges. A set of non-parametric structural equations can quantify the three causal effects shown in the causal graph in Fig. 3a as:

$$x = f_x(\epsilon_x), \ t = f_t(x, \epsilon_t), \ y = f_y(x, t, \epsilon_y). \tag{2}$$

In Eq. 2, ϵ_x, ϵ_t and ϵ_y denote the "noise" of the observed variables, concieved as *exogenous* or mutually independent sources of unmeasured variation. The noise terms represent the causal effect of unobserved variables on the variable on the LHS. Note that for each equation, we assume that the variables on the RHS influences those on the LHS, not the other way around. Rewriting this equation in a different order as $x = f_t^{-1}(t, \epsilon_t)$ can be misleading as it implies that Yelp rating causally influences restaurant type. The structural equations (Eq. 2) provide a quantitative way to represent *intervention* on a variable of the corresponding causal graph (Fig. 3a). The *do-calculus* [113] of Pearl was proposed to define intervention in SCMs. Specifically, the do-calculus introduces a new operator $do(t')$, which denotes the intervention of setting the value of the variable t to t'. The notation of $do(t)$ leads to a formal expression of the interventional distributions:

Definition 3. Interventional Distribution (Post-intervention Distribution). *The interventional 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 $P(y|do(t))$ refers to the distribution of customer flow y as if the rating t is set to t' by intervention, where all the arrows into t are removed, as shown in Fig. 3b. The structural equations associated with Fig. 3b under the intervention on the treatment variable, denoted by $do(t')$, can be written as:

$$x = f_x(\epsilon_x), \ t = t', \ y = f_y(x, t, \epsilon_y), \tag{3}$$

which formulates the interventional distribution as $P(y|do(t')) = f_y(x, t, \epsilon_y)$. Then, when it comes to the causal effect of t on y, in the language of SCMs, the problem of calculating causal effects can be translated into queries about the interventional distribution $P(y|do(t))$ with different t. Implicitly, we assume that the variables follow the same causal relations of a SCM for each instance. Hence, SCMs enable us to define *average treatment effect* (ATE). For the running example, the ATE of Yelp rating might be defined as:

$$\tau(t, c) = \mathbb{E}[y|do(t)] - \mathbb{E}[y|do(c)], t > c, \tag{4}$$

where t and c refer to the ratings that are considered as positive and negative, respectively. In many cases, the treatment variable is binary, thus the ATE reduces to a value $\mathbb{E}[y|do(1)] - \mathbb{E}[y|do(0)]$. It is crucial to note that $P(y|do(t))$ and $P(y|t)$ are not the same, which makes calculating ATEs impossible. This gap can give rise to *confounding bias*, which results if one estimates treatment effects using $P(y|t)$ where $P(y|do(t))$ is in fact required. The existence of a *back-door path* is a common source of confounding, rendering $P(y|do(t))$ and $P(y|t))$ distinct. An example of is the path $t \leftarrow x \rightarrow y$ in Fig. 3a. Observe also that randomized treatment assignment directly avoids back-door paths, side-stepping confounding bias. We present the formal definitions of confounding bias, back-door path and confounder with do-calculus and SCMs.

Definition 4. Confounding Bias. *Given variables x, y, confounding bias exists for causal effect $x \to y$ iff the probabilistic distribution representing the statistical association is not always equivalent to the interventional distribution, i.e., $P(y|x) \neq P(y|do(x))$.*

Definition 5. Back-door Path. *Given a pair of treatment and outcome variables (t, y), we say a path connecting t and y is a back-door path for (t, y) iff it satisfies that (1) it is not a directed path; and (2) it is not blocked (it has no collider).*

Definition 6. Confounder (Confounding Variable). *Given a pair of treatment and outcome variables (t, y), we say a variable $z \notin \{t, y\}$ is a confounder iff it is the central node of a fork and it is on a back-door path of (t, y).*

In particular, in the running example, the probability distribution $P(y|t)$ results from a mixture of the causal effect $P(y|do(t))$ and the statistical associations produced by the back-door path $t \leftarrow x \to y$, where x is the confounder. To obtain unbiased estimate of causal effects from observational data requires eliminating confounding bias, a procedure referred to as *causal identification.*

Definition 7. 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 over observables.*

A common way to identify causal effects in SCMs is to block the back-door paths that reflect other irrelevant causal effects. A way to eliminate confounding bias is to estimate the causal effect within subpopulations where the instances are homogeneous w.r.t. confounding variables [112]. This corresponds to *adjustment* on variables that satisfy the *back-door criterion* for causal identification [113]. Now we present a formal definition of the back-door criterion.

Definition 8. Back-door Criterion. *Given a treatment-outcome pair (t, y), a set of features $\boldsymbol{x}$ satisfies the back-door criterion of (t, y) iff conditioning on $\boldsymbol{x}$ can block all back-door paths of (t, y).*

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

Confounding bias without back-door path. Confounding bias may exist without back-door paths. An example is a type of selection bias [18], when the causal graph is $t \to z \leftarrow x \to y$ and the dataset is collected only for instances with $z_i = 1$, then within this dataset, the estimated statistical association $P(y|t)$ can be non-zero although we know that there is no causal effect $t \to y$. Selection bias can also result from adjustment on certain variables (e.g., colliders or descendants of the outcome variable). Without knowing the complete graph, Entner et al. [41] provide a set of rules that are sufficient to decide whether a set of variables satisfy the back door criterion, or that t actually has no effect on y. This implies that there is a middle ground between hoping to adjust for all and only the right stuff, and trying to learn the entire causal graph.

Beyond do-calculus. Do-calculus has some limitations, which mainly come from the i.i.d. assumption [126]. This implies that it is difficult to formulate individual-level hypothetical distributions

with do-calculus. Let us consider the running example, even if we could hack Yelp and let it show median rating instead of average rating, we still cannot answer questions such as what would the customer flow for a restaurant be if we had increased its rating by 0.5 star without changing the ratings of others? In [114], Pearl refers to the hypothetical distributions which cannot be identified through interventions as *counterfactuals.* Do-calculus, which formally represents hypothetical interventions, cannot formulate counterfactuals in SCMs. Therefore, besides do-calculus, Pearl [112] introduced a new set of notations. For example, $P(y^t|y', t')$ denotes the probability of the outcome y would be if the observed treatment value is t, given the fact that we observe y', t' in the data. In the running example, for a restaurant with rating t' and customer flow y', the counterfactual probability $P(y^t|y', t')$ is the distribution of the customer flow if we had observed its rating as t.

2.2 Potential Outcome Framework

The potential outcome framework [108, 131] is widely used by practitioners to learn causal effects as it is defined w.r.t. a given treatment-outcome pair (t, y). A *potential outcome* is defined as:

Definition 9. Potential Outcome. *Given the treatment and outcome t, y, the potential outcome of instance i, y_i^t, is the outcome that would have been observed if the instance i had received treatment t.*

This framework allows a straightforward articulation of the basic challenge of causal inference [67]: only one potential outcome can be observed for each instance. Using potential outcomes it is possible to define the *individual treatment effect* (ITE) as the difference between potential outcomes of a certain instance under two different treatments. ITE can be extended to ATE on arbitrary populations. Practitioners often assume binary treatment ($t \in \{0, 1\}$), where $t = 1$ ($t = 0$) mean that an instance is under treatment (control). The formal definition of ITE is:

Definition 10. Individual Treatment Effect. *Assuming binary treatment, given an instance i and its potential outcomes y_i^t, the individual treatment effect is defined as $\tau_i = y_i^1 - y_i^0$.*

Based on ITE, the ATE of the target population and other subpopulation average treatment effects such as *conditional average treatment effect* (CATE) can be defined. 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. Given ITEs, ATE can be formulated as the expectation of ITEs over the whole population $i = 1, ..., n$ as:

$$\tau = \mathbb{E}_i[\tau_i] = \mathbb{E}_i[y_i^1 - y_i^0] = \frac{1}{n} \sum_{i=1}^{n} (y_i^1 - y_i^0), \tag{5}$$

The ATE on subpopulations is often of interest. An example is the *conditional average treatment effect* (CATE) of instances with the same features, i.e., $\tau(\boldsymbol{x}) = \mathbb{E}_{i:\boldsymbol{x}_i = \boldsymbol{x}}[\tau_i]$.

Using the potential outcome framework to estimate treatment effects. Similar to how $P(y|do(t))$ and $P(y|t)$ are distinct in the SCM framework, $P(y^t)$ and $P(y|t = 1)$ are not the same within the potential outcomes framework. First, the most fundamental assumption that is commonly used to facilitate estimation is *the stable unit treatment value assumption* (SUTVA), which can be broken down into two conditions: *well-defined treatment levels* and *no interference*. The condition of well-defined treatment indicates that given two different instances $i \neq j$, if the values of their treatment variable are equivalent, then they receive the same treatment. 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 $y_i^t = y_i^{t_i}$, where $t \in \{0, 1\}^n$ denotes the vector of treatments for all instances. Although the condition of no interference is often assumed, there are cases when the inter-dependence between instances matters [122, 151]. The second assumption, *consistency*, means that the observed outcome is independent of the how

the treatment is assigned. Finally, a commonly invoked condition is the unconfoundedness (a.k.a. ignorability), which posits that the set of confounding variables $\mathcal{S}$ can be measured. Starting from this point, we use $\boldsymbol{s}$ to denote the vector of confounding variables. *Unconfoundedness* means that the values of the potential outcomes are independent of the observed treatment, given the set of confounding variables. Unconfoundedness is defined as: $y_i^1, y_i^0 \perp\!\!\!\perp t_i | \boldsymbol{s}$, where $\boldsymbol{s}$ denotes a vector of confounders, each element of which is a feature that causally influences both the treatment t_i and the outcome y_i^t. We can see that this is also an assumption defined at the individual level. Unconfoundedness directly leads to causal identification as Pearl [112] showed that, given Eq. 2.2, $\mathcal{S}$ always satisfies the back-door criterion of (t, y). That is, under unconfoundedness we have $P(y^1 | t, \boldsymbol{s}) = P(y | t, \boldsymbol{s})$. A further condition $P(t = 1 | \boldsymbol{x}) \in (0, 1)$ if $P(\boldsymbol{x}) > 0$ is often invoked, which, when combined with ignorability, is referred to as *strong ignorability*.

Comparing SCMs and potential outcomes. The two formal frameworks are logically equivalent, which means an assumption in one can always be translated to its counterpart in the other [112]. There are also some differences between them. In the potential outcome framework, the causal effects of the variables other than the treatment and the special variables such as instrumental variable (see Section 3.2.1) are not defined. This is a strength of this framework as we can model the interesting causal effects without knowing the complete causal graph [2]. While in SCMs, we are able to study the causal effect of any variable. Therefore, SCMs are often preferred when learning causal relations among a set of variables [2]. Conversely, if the goal is narrowly to estimate a given treatment effect, developing estimators can be more straightforward using the potential outcomes framework. The reader may draw their own conclusions after consulting Section 3.

3 LEARNING CAUSAL EFFECTS

In this section, we introduce methods for learning causal effects. We aim to understand how to quantify causal effects in a data-driven way. We first introduce the problem statement and evaluation metrics. Next, we review three categories of methods: those with and without unconfoundedness and advanced methods for big data. We define the problem of learning causal effects.

Definition 11. Learning Causal Effects *Given n instances, $[(\boldsymbol{x}_1, t_1, y_1), ..., (\boldsymbol{x}_n, t_n, y_n)]$, learning causal effects quantifies how the outcome y is expected to change if we modify the treatment from c to t, which can be defined as $\mathbb{E}[y|t] - \mathbb{E}[y|c]$, where t and c denote a treatment and the control.*

Depending on the application, we 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 or an individual. Among all types of treatment effects, the ATE is often interesting when it comes to making decision on whether a treatment should be introduced to a population. Furthermore, in SCMs and do-calculus, the 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 types of treatment effects. In terms of evaluation, regression error metrics such as mean absolute error (MAE) can be used to evaluate models for learning ATE. Given the ground truth τ and the inferred ATE $\hat{\tau}$, the MAE on ATE is:

$$\epsilon_{MAE\_ATE} = |\tau - \hat{\tau}|. \tag{6}$$

However, when the population consists of 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 homogeneous group. In many cases, without knowledge about the affiliation of groups, an assumption we can make is that each subpopulation is defined by different feature values. Thus, we can learn a function to map the features that define a subpopulation

to its estimated ATE. With this assumption, given a certain value of features $\boldsymbol{x}$ and binary treatment t, the CATE is a function of $\boldsymbol{x}$ and is defined as:

$$\tau(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}, t = 1] - \mathbb{E}[y|\boldsymbol{x}, t = 0]. \tag{7}$$

In this case, we assume that only the features and the treatment are two factors that determine the outcome. The target is to learn a function $\hat{\tau}$ to estimate CATE. Empirically, with cross-validation, we can evaluate the quality of the learned function $\hat{\tau}(\boldsymbol{x})$ based on the mean squared error (MSE):

$$\epsilon_{PEHE} = \frac{1}{n} \sum_{i=1}^{n} (y_i^1 - y_i^0 - \hat{\tau}(\boldsymbol{x}_i))^2, \tag{8}$$

which is often referred to as *precision in estimation of heterogeneous effect* (PEHE). It is also adopted for evaluating estimated individual treatment effects (ITE) [65, 77, 97, 138]. Note that PEHE is the mean squared error of the estimated ITEs.

3.1 Traditional Methods without Unobserved Confounders

To simplify the problem, it is commonly assumed (or hoped) that all confounders are among the observed features. In SCMs, this is equivalent to assuming that conditioning on some subset of observed features, denoted by $\boldsymbol{s}$, can block all the back-door paths. *Adjustment* eliminates confounding bias based on the subset of features $\boldsymbol{x}$. We introduce three families of adjustments: regression adjustment, propensity score methods and covariate balancing. We assume binary treatment $t \in \{0, 1\}$ and adopt the language of generalized structural equation introduced in Section 2. The causal graph embedding the assumption for such methods is shown in Fig. 4.
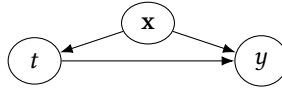

Fig. 4. A causal graph for the unconfoundedness assumption which is used for learning causal effects.

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 label and the features. As discussed in Section 2, to learn causal effects, we are interested in interventional distributions and counterfactuals which cannot be directly estimated from data. Following the potential outcome framework, we infer the counterfactual outcomes $y_i^{1-t_i}$ based on the features $\boldsymbol{x}$ and the treatment t. Roughly speaking, there are two types of regression adjustment. One is to fit a single function to estimate $P(y|\boldsymbol{x}, t)$. It is enough for inferring ITE because there would be no confounding bias by conditioning on $\boldsymbol{x}$, i.e., $P(y|t, \boldsymbol{x}) = P(y|do(t), \boldsymbol{x})$. So we can infer the counterfactual outcome as $\hat{y}_i^{1-t_i} = \mathbb{E}(y_i|1 - t_i, \boldsymbol{x}_i)$. The second is to fit a model for each potential outcome, i.e. $P^1(y|\boldsymbol{x}) = P(y|t = 1, \boldsymbol{x})$ and $P^0(y|\boldsymbol{x}) = P(y|t = 0, \boldsymbol{x})$. Then we can estimate ATE by:

$$\hat{\tau} = \left[\sum_{i=1}^{n} (\hat{y}_i^1 - \hat{y}_i^0) \right] / n, \tag{9}$$

where we estimate $\hat{y}_i^t$ by the model $\mathbb{E}(y|t, \boldsymbol{x}_i)$.

3.1.2 Propensity Score Methods. Propensity score methods can be considered as a special case of matching methods [106]. Matching methods divide instances into strata and treat each stratum as a randomized controlled trial (RCT). Based on this assumption, ATE is identified and can be estimated by the naïve estimator within each stratum. In matching methods, we assume *perfect*

stratification, which means that (1) each group is defined by a set of features $\boldsymbol{x}$; (2) instances in a group are indistinguishable except the treatment and the potential outcomes [106]. Formally, perfect stratification means $\mathbb{E}[y_i^t|t_i = 1, f(\boldsymbol{x})] = \mathbb{E}[y_i^t|t_i = 0, f(\boldsymbol{x})], t \in \{0, 1\}$. Function $f(\boldsymbol{x})$ outputs a continuous value and we stratify instances into groups based on $f(\boldsymbol{x})$. This equation can be interpreted as: given the group affiliation, the expected values of the potential outcomes do not change with the observed treatment. This is equivalent to the unconfoundedness assumption in each stratum defined by $f(\boldsymbol{x})$ whose parameterization can be flexible.

But when there exists a group which only contains instances with $t = 1$ or $t = 0$, where we cannot estimate ATE in such a stratum with the naïve estimator. This issue is referred to as a *lack of overlap*. Note that the strong ignorability condition includes $P(t = 1|\boldsymbol{x}) \in (0, 1)$, thus ensuring overlap in the large sample limit. To mitigate the lack of overlap, *matching as weighting* methods are proposed, side-stepping the need for perfect matchings, which are difficult to achieve in practice. The most widely adopted methods define the function $f(\boldsymbol{x})$ as an estimator of the *propensity score* $P(t|\boldsymbol{x})$. Although the features and treatment assignments are fixed given observational data, we assume that observed treatment is assigned by sampling from the true propensity score: $t_i \sim P(t_i|\boldsymbol{x}_i)$. That is, the propensity score is the probability of receiving treatment, given the features. The main advantage of the propensity score, in contrast to perfect matching, is that it is a sufficient dimension reduction [130] in the sense that strata defined purely in terms of the propensity score will permit unconfounded causal inference. Moreover, as a single continuous quantity, it is possible to develop estimators based on weighted averages rather than discrete stratification.

Of course, in practice the propensity score is not known beforehand, but must be estimated. Fortunately, supervised learning provides many methods for estimating the propensity score by training a classifier to predict whether an instance would be treated, given its features. It is common to estimate $P(t|\boldsymbol{s})$ by logistic regression. Despite its popularity, note that the validity of a linear logistic model can be suspected and nonparametric alternatives are readily available.

Propensity score methods can be categorized into four classes [12]: *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 as propensity score stratification is an extension of PSM, and adjustment based on propensity score is a combination of regression adjustment and propensity score methods.

Propensity Score Matching (PSM). PSM matches a treated (controlled) instance to a set of controlled (treated) instances with similar propensity scores. For example, in *Greedy One-to-one Matching* [53], for each instance i, we find an instance j with the most similar propensity score to i in the other treatment group. Once the instances are matched, we can estimate ATE as:

$$\hat{\tau} = \left[\sum_{i:t_i=1} (y_i - y_j) + \sum_{i:t_i=0} (y_j - y_i) \right] / n. \tag{10}$$

Besides the Greedy One-to-one PSM, there are many other PSM methods. The difference comes what methods we use to match instances. Readers can check [12] for various PSM methods. Stratification on propensity scores is an extension of PSM. Having propensity score estimated, we can stratify instances based on the predefined thresholds on propensity scores or the number of strata.

Thus, stratum-specific ATE can be calculated by the naïve estimator. Specifically, ATE is calculated as the weighted average over all strata:

$$\hat{\tau} = \sum_j |U_j|(\frac{1}{|U_j^1|} \sum_{i \in U_j^1} y_i - \frac{1}{|U_j^0|} \sum_{i \in U_j^0} y_i)/\sum_j |U_j|, \tag{11}$$

where U_j, U_j^1 and U_j^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 be used to account for the difference between instances in the same stratum [12, 74, 98].

Inverse Probability of Treatment Weighting (IPTW). IPTW [66] is a covariate balancing method. Intuitively, we can weight instances based on their propensity scores to synthesize a RCT [12]. A common way to define the sample weight w_i is by:

$$w_i = \frac{t_i}{P(t_i|\mathbf{x}_i)} + \frac{1 - t_i}{1 - P(t_i|\mathbf{x}_i)}. \tag{12}$$

With Eq. 12, we can find that for a treated instance i and a controlled instance j, $w_i = \frac{1}{P(t_i|\mathbf{x}_i)}$ and $w_j = \frac{1}{1-P(t_j|\mathbf{x}_j)}$. So the weight refers to the inverse probability of receiving the observed treatment (control). For example, if we observe 10 instances with $\mathbf{x}_i = \mathbf{x}$ and only one of them is treated. Then we estimate the propensity score $P(t = 1|\mathbf{x})$ as 0.1. To synthesize a RCT, we need to balance the two treatment groups by weighting the treated instance 9 times as the instances under control, which is done by Eq. 12. Then we can calculate a weighted average of factual outcomes for the treatment and control groups:

$$\hat{\tau} = \frac{1}{n^1} \sum_{i:t_i=1} w_i y_i - \frac{1}{n^0} \sum_{i:t_i=0} w_i y_i, \tag{13}$$

where n^1, n^0 denote the number of instances under treatment and control. 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. 13. Regression adjustment can also be applied to the weighted dataset to reduce the residual of the synthetic RCT [76]. Instances with propensity score close to 1 or 0 may suffer from an extremely large weight. In [64], Hernan proposes to stabilize weights to handle this issue in IPTW.

Doubly Robust Estimation (DRE). Funk et al. [46] propose DRE as a combination of a regression adjustment $\mathbb{E}[y|t, \mathbf{x}]$ and another method that estimates the propensity score $\mathbb{E}[t|\mathbf{x}]$. In fact, 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 these two models as:

$$\hat{y}_i^1 = \frac{y_i t_i}{\hat{P}(t_i|\mathbf{x}_i)} - \frac{\tilde{y}_i^1(t_i - \hat{P}(t_i|\mathbf{x}_i))}{\hat{P}(t_i|\mathbf{x}_i)} \quad , \quad \hat{y}_i^0 = \frac{y_i(1 - t_i)}{1 - \hat{P}(t_i|\mathbf{x}_i)} - \frac{\tilde{y}_i^0(t_i - \hat{P}(t_i|\mathbf{x}_i))}{1 - \hat{P}(t_i|\mathbf{x}_i)} \tag{14}$$

where $\tilde{y}_i^{t_i}$ denotes the estimated potential outcomes of the instance i with regression adjustment $\mathbb{E}[y|t, \mathbf{x}]$ and $\hat{P}(t_i|\mathbf{x}_i)$ is the estimated propensity score for the instance i. Taking a closer look at Eq. 14, we can find that the regression adjustment model is applied to the estimation of counterfactual outcomes as: $\hat{y}_i^{1-t_i} = \tilde{y}_i^{1-t_i}$, while more complicated, a mixture of the regression adjustment of propensity score models is developed for the factual outcomes. Then we can estimate ATE by taking the average over the estimated ITE for all the instances as in Eq. 9.

Targeted Maximum Likelihood Estimator (TMLE) [154] is a more generalized method than DRE. ATE can be inferred with TMLE as: $\frac{1}{n} \sum_{i=1}^n Q_n^*(1, \mathbf{x}_i) - Q_n^*(0, \mathbf{x}_i)$. To obtain $Q_n^*(t, \mathbf{x})$, there are three steps: (1) A model $Q_n^0(t, \mathbf{x})$ is fitted to estimate the factual outcomes with the features and the treatment. (2) A model $g(t = 1, \mathbf{x})$ is fitted for propensity scores $P(t = 1|\mathbf{x})$. (3) Given Q_n^0 and $g(t, \mathbf{x})$, another model is fitted to minimize the mean squared error (MSE) on the factual outcomes. Assuming $y \in [0, 1]$, this is done by learning a new estimator $\bar{Q}_n^*(t, \mathbf{x})$ with parameters $\hat{\epsilon}_0$ and $\hat{\epsilon}_1$:

$$\bar{Q}_n^*(t, \mathbf{x}) = \text{expit}\left[Q_n^0(t, \mathbf{x})/(1 - Q_n^0(t, \mathbf{x})) + \hat{\epsilon}_0 H_0(t, \mathbf{x}) + \hat{\epsilon}_1 H_1(t, \mathbf{x})\right], \tag{15}$$

where $\text{expit}(a) = \frac{1}{1+\exp(-a)}$, $H_0(t, \boldsymbol{x}) = -\frac{\mathbb{1}(t=0)}{g(t=0|\boldsymbol{x})}$ and, $H_1(t, \boldsymbol{x}) = \frac{\mathbb{1}(t=1)}{g(1|W)}$.

3.1.3 Covariate Balancing.
Besides reweighting samples with propensity scores, the *covariate balancing* methods learn sample weights through regression [84].

Entropy Balancing (EB). Hainmueller [60] proposes EB, a preprocessing method for covariate balancing. The goal is to learn sample weights of the instances under control such that the moments of the two groups are matched. The weights are learned by minimizing the objective:

$$\arg\min_{w_i} H(\boldsymbol{w}) = \sum_{i:t_i=0} d(w_i) \ s.t. \ \sum_{i:t_i=0} w_i c_{ri}(\boldsymbol{x}_i) = m_r \ with \ r = 1, ..., R, \tag{16}$$

where $\sum_{i:t_i=0} w_i = 1$; $w_i \geq 0, \forall i \in \{i|t_i = 0\}$. $d(\cdot)$ is a distance metric (e.g., KL divergence $d(w_i) = w_i \log(w_i/q_i)$) measuring the distance between the learned weights $\boldsymbol{w}$ and base weights $\boldsymbol{q}, q_i \geq 0$, and $\sum_i q_i = 1$. We can use uniform weights $q_i = 1/n^0$, where n^0 denotes the number of instances under control. $\sum_{i:t_i=0} w_i c_{ri}(\boldsymbol{x}_i) = m_r$ refers to a set of R balance constraints where $c_{ri}(\boldsymbol{x}_i)$ is specified as a moment function for the control group and m_r denotes the counterpart of the treatment group. For example, when $c_{ri}(\boldsymbol{x}_i) = (x_i^j)^r$, then $\sum_{i:t_i=0} w_i c_{ri}(\boldsymbol{x}_i)$ denotes the reweighted r-th moment of the feature x^j for the control, and therefore, m_r would contain the r-th order moment of a feature x^j from the treatment group. Compared to other balancing methods, EB allows a large set of constraints such as moments of feature distributions and interactions. Different from the matching methods, EB keeps weights close to the base weights to prevent information loss.

Approximate Residual Balancing (ARB). ARB [11] combines balancing weights with a regularized regression adjustment for learning ATE from high-dimensional data. ARB consists of three steps. First, the sample weights $\boldsymbol{w}$ are learned as:

$$\arg\min_{\boldsymbol{w}} (1 - \xi)||\boldsymbol{w}||_2^2 + \xi||\frac{1}{n^1} \sum_{i:t_i=1} \boldsymbol{x}_i - X_{i:t_i=0}^T \boldsymbol{w}||_\infty^2 \ s.t. \ \sum_{i:t_i=0} w_i = 1 \ and \ w_i \in [0, (n^0)^{-2/3}], \tag{17}$$

where $X_{i:t_i=0}$ denotes the feature matrix for the control group. Then a regularized linear regression adjustment model with parameters $\boldsymbol{\beta}$ fitted as:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i:t_i=0} (y_i - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \lambda((1 - \alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1), \tag{18}$$

where λ and α are hyperparameters controlling the strength of regularization. At the end, we can estimate ATE as $\hat{\tau} = \frac{1}{n^1} \sum_{i:t_i=1} y_i - (\frac{1}{n^1} \sum_{i:t_i=1} \boldsymbol{x}_i^T\boldsymbol{\beta} + \sum_{i:t_i=0} w_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}))$. Compared to EB [60], ARB handles sparseness of high-dimensional data with lasso and elastic net [149].

Covariate Balancing Propensity Score (CBPS). CBPS [72], a method robust to misspecification of propensity score model, is proposed to model propensity scores and balance covariate simultaneously. Assuming the propensity score model $f(\boldsymbol{x})$ with parameters $\boldsymbol{\beta}$, the efficient Generalized Method of Moments (GMM) estimator is used to learn $\boldsymbol{\beta}$:

$$\arg\min_{\boldsymbol{\beta}} \left[\frac{1}{n} \sum_i g(t_i, \boldsymbol{x}_i)\right]^T \Sigma(\boldsymbol{t}, X)^{-1} \left[\frac{1}{n} \sum_i g(t_i, \boldsymbol{x}_i)\right], \tag{19}$$

$g(t_i, \boldsymbol{x}_i) = \left(\frac{t_i f'(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)} - \frac{(1-t_i)f'(\boldsymbol{x}_i)}{1-f(\boldsymbol{x}_i)}, \frac{(t_i-f(\boldsymbol{x}_i))f'(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)(1-f(\boldsymbol{x}_i))}\right)^T$ are the moment conditions. They are derived from the first order condition of minimizing $-\sum_{i=1}^n t_i \log f(\boldsymbol{x}_i) + (1-t_i) \log(1-f(\boldsymbol{x}_i))$, the MLE estimator by which we learn $\boldsymbol{\beta}$. Similar to EB [60], CBPS combines two methods: covariate balancing and IPTW. Compared to EB, CBPS models propensity scores.
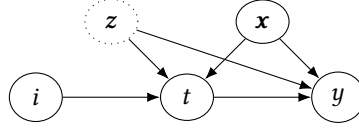
Fig. 5. A causal graph of a valid instrumental variable (i) when there are unobserved confounders (z). The binary exogenous variable i stands for whether a customer submits a review. The restaurant type (x) is an observed confounder and z is a set of unobserved confounders.

3.2 Traditional Methods with Unobserved Confounders

In many real-world problems of learning causal effects, there exist unobserved confounders. In these cases, the assumption of unconfoundedness is not satisfied. In the language of SCMs, this means we are not able to block back-door path by conditioning on the features. Therefore, a family of methods are developed to handle this situation. The intuition is to utilize alternative information. Here, we focus on three most popular methods for learning causal effects with unobserved confounders: *instrumental variable methods*, *front-door criterion* (or "identification by enumeration of mechanism"), and *regression discontinuity design*.

3.2.1 Instrumental Variable Methods. Instrumental variables enable us to learn causal effects with unobserved confounders, which are defined as:

Definition 12. Instrumental Variable *Given an observed variable i, features $\mathbf{x}$, the treatment t and the outcome y, we say i is a valid instrumental variable (IV) for the causal effect of $t \rightarrow y$ iff i satisfies: (1) $i \not\perp\!\!\!\perp t|\mathbf{x}$, and (2) $i \perp\!\!\!\perp y|\mathbf{x}, do(t)$ [6].*

This means a valid IV causally influences the outcome only through affecting the treatment. In SCMs, the first condition means there is an edge $i \rightarrow t$ or a non-empty set of collider(s) $\mathbf{x}$ s.t. $i \rightarrow t \leftarrow \mathbf{x}$ where $\mathbf{x}$ denotes the features or a subset of features. The second condition requires that $i \rightarrow t \rightarrow y$ is the only path that starts from i and ends at y. Thus, blocking t makes $i \perp\!\!\!\perp y$. This implies the *exclusive restriction* that there must not exist direct edge $i \rightarrow y$ or path $i \rightarrow \mathbf{x}' \rightarrow y$ where $\mathbf{x}' \subseteq \mathbf{x}$. Mathematically, for all t and $i \neq j$, this can also be denoted by $y(do(i), t) = y(do(j), t)$.

In the running example, if we only observe one confounder - the restaurant type (x), while the other confounder (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 i causally influences t and it can only causally affect y through its influence on t. With a valid IV, we identify the causal effect $t \rightarrow y$ if both the interventional distributions - $P(t|do(i))$ and $P(y|do(i))$ are identifiable.

A Linear SCM for an IV Estimator. Here, we show an IV estimator with a linear SCM. If we also assume that the observed and unobserved confounders $\mathbf{x}$ and $\mathbf{u}$ come with a zero mean, we can write down the structural equations for the causal graph in Fig. 5 as:

$$t = \alpha_i i + \boldsymbol{\alpha}_z^T \mathbf{z} + \boldsymbol{\alpha}_x^T \mathbf{x} + \alpha_0 + \epsilon_t, \quad y = \tau t + \boldsymbol{\beta}_z^T \mathbf{z} + \boldsymbol{\beta}_x^T \mathbf{x} + \beta_0 + \epsilon_y, \tag{20}$$

where ϵ_t and ϵ_y are Gaussian noise terms with zero mean. By substituting t in the second equation with the RHS of the first equation in Eq. 20, we get:

$$y = \tau \alpha_i i + (\tau \alpha_z + \beta_z)^T \mathbf{z} + (\tau \alpha_x + \beta_x)^T \mathbf{x} + \gamma_0 + \eta, \tag{21}$$

where $\gamma_0 = \tau \alpha_0 + \beta_0$, $\eta = \tau \epsilon_d + \epsilon_y$. Then estimator for the ATE (τ) is obtained:

$$\hat{\tau} = (\mathbb{E}[y|i] - \mathbb{E}[y|i']) / (\mathbb{E}[t|i] - \mathbb{E}[t|i']). \tag{22}$$

Here, we rely on the following assumptions: linear structural equations, homogeneous treatment effect, valid IV, zero-mean additive noise, and unobserved confounders. What if some of them do not
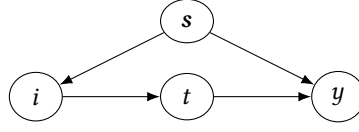
Fig. 6. Assuming that we observe all confounders $\boldsymbol{x}$ as s for the causal effect of the IV, whether a customer writes a review on customer flow y, 2SLS can estimate the treatment effect of rating on customer flow ($t \rightarrow y$).

hold? Can this estimator work under some conditions? In [6], Angrist et al. showed that the ratio estimator (Eq. 22) identifies ATE when either the effect of i on t or that of i on y is homogeneous. **An IV Estimator under the potential outcome framework.** The potential outcome framework formulates the ITE of the IV i on the outcome y as:

$$y_j(i_k = 1, t_l(i_k = 1)) - y_j(i_k = 0, t_l(i_j = 0)), \qquad (23)$$

where $y_j(i_k, t_l)$ and $t_l(i_k)$ are the value of y and t by setting the value of the k-th IV to i_k. We also assume the IVs are binary. With the two conditions in Definition 12, we know that i affects y via its influence on t, so we remove i_j that explicitly influences the value of y_j and reduces Eq. 23 to:

$$[y_j^1 P(t_j = 1|i_j = 1) + y_j^0 P(t_j = 0|i_j = 1)] - [y_j^1 P(t_j = 1|i_j = 0) + y_j^0 P(t_j = 0|i_j = 0)]$$
$$= (y_j^1 - y_j^0)(P(t_j = 1|i_j = 1) - P(t_j = 1|i_j = 0)). \qquad (24)$$

We obtain the ratio estimation (Eq. 22) again when we assume homogeneous treatment effect of i on t or that of i on y.

Two-stage Least Square (2SLS). As the IV estimator in Eq. 22 is restrictive, we may have to control a set of features $\boldsymbol{x}$ to block the back-door paths between the IV and the outcome so that the IV can be valid. These cases make it difficult or infeasible to use the estimator in Eq. 22. So we introduce 2SLS [5]. Fig. 6 shows an example for such cases where $\boldsymbol{x}$ 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 $i \leftarrow \boldsymbol{x} \rightarrow y$ has to be blocked. Besides, we may have multiple IVs for each treatment and multiple treatments. Assuming there is a set of treatments $\boldsymbol{t}$ and each treatment $t_{,j}$ has a set of IVs $\boldsymbol{i}_{,j}$. In 2SLS, two regressions are performed to learn the causal effects $(t_{,1} \rightarrow y), ..., (t_{,j} \rightarrow y), ...$: (1) we fit a function $\hat{t}_{,j} = f_{t,j}(\boldsymbol{i}_{,j}, \boldsymbol{x}_{,j})$ for each treatment variable $t_{,j}$. (2) we learn a function $y = g(\hat{\boldsymbol{t}}, \boldsymbol{x})$ where $\hat{\boldsymbol{t}}$ signifies the set of treatments. Then the coefficient on $D_{,j}$ is a consistent estimate of the ATE of the j-th treatment $t_{,j}$ on y. The intuition of 2SLS follows how we find a valid IV. In the first stage, we estimate how much a certain treatment $t_{,j}$ changes if we modify the relevant IV $\boldsymbol{i}_{,j}$. In the second stage, we see how the changes in $t_{,j}$ caused by $\boldsymbol{i}_{,j}$ would influence y.

For practical studies in big data, Kang et al. [80] show identification of causal effects is possible if more than 50% of the IVs are valid. They also discuss conditions that further allow identification with more than 50% invalid IVs without knowing which IVs are valid.

3.2.2 Front-door Criterion. The front-door criterion [111] enables us to learn causal effects $t \rightarrow y$ with unobserved confounders. With the front-door criterion we condition on a set of variables $\boldsymbol{m}$ which satisfies the following three conditions: (1) $\boldsymbol{m}$ blocks all the directed paths from t to y, (2) there are no unblocked back-door paths from t to $\boldsymbol{m}$, (3) t blocks all the back-door paths from $\boldsymbol{m}$ to y. In other words, we say that the set of variables $\boldsymbol{m}$ *mediates* the causal effect of t on y. From the first condition, we decompose $t \rightarrow y$ to a product of $t \rightarrow \boldsymbol{m}$ and $\boldsymbol{m} \rightarrow y$ as: $P(y|do(d)) = \int_{\mathcal{M}} P(y|do(\boldsymbol{m}))P(\boldsymbol{m}|do(d))d\boldsymbol{m}$. The second condition means there is no confounding

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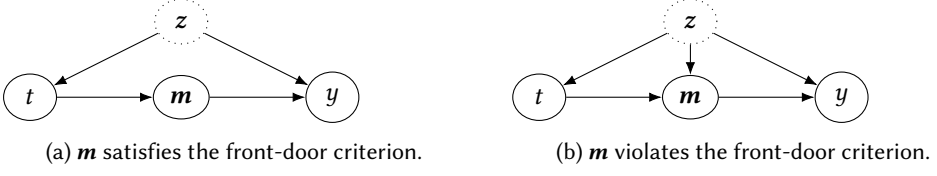
(b) $\boldsymbol{m}$ violates the front-door criterion.

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

bias for the causal effect $t \rightarrow \boldsymbol{m}$: $P(\boldsymbol{m}|do(d)) = P(\boldsymbol{m}|d)$. The third condition infers $P(y|do(\boldsymbol{m}))$ by:

$$P(y|do(\boldsymbol{m})) = \int_{\mathcal{T}} P(y|t, \boldsymbol{m})P(t)dt. \tag{25}$$

Then the interventional distribution corresponding to $t \rightarrow y$ can be identified as:

$$P(y|do(d)) = \int_{\mathcal{M}} P(\boldsymbol{m}|d) \sum_{t \in \mathcal{T}} P(y|t, \boldsymbol{m})P(t). \tag{26}$$

We can estimate the probabilities on the RHS of Eq. 26 from observational data. For example, we can let the set of variables $\boldsymbol{m}$ be the ranking of a restaurant in the search results. When the ranking is decided by the Yelp rating, $(\boldsymbol{z} \perp\!\!\!\perp \boldsymbol{x}|t, y)$, $\boldsymbol{m}$ satisfies the front-door criterion (Fig. 7a). However, when the ranking $\boldsymbol{m}$ is affected by both the rating t and confounders $\boldsymbol{z}$ (e.g. the restaurant category), then $\boldsymbol{m}$ 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. Sometimes, treatment assignments 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 [3]. Yelp shows the rating of a restaurant rounded to the nearest half star. For example, restaurant i with average rating 3.26 and restaurant j 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) [3, 21] is that the restaurants with average rating close to the cutoff $r_0 = 3.25$ are homogeneous w.r.t. the confounders. Therefore, what can make a difference in their factual outcomes is the treatment. In other words, the treatments are randomly assigned to such restaurants, which leads to the identification of the ATE. In Sharp RDD, we assume that the observed outcome is a function of the running variable as:

$$y_i = f(r_i) + \tau t_i + \epsilon_i = f(r_i) + \tau \mathbb{1}(r_i \geq r_0) + \epsilon_{yi}, \tag{27}$$

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 [49] pointed out that high-order polynomials can be misleading in RDD. In the Yelp study, the fact that customers' decision on which restaurant to go solely relies on the Yelp rating supports this assumption. 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) [22]. The *Fuzzy RDD* method [7, 21] is developed to handle the cases when cut-offs on the running variable are not strictly implemented. For example, users may see the real average rating when they look into details of the restaurants and find out that the two restaurants i and j are not that different in terms of rating. Similar to the propensity

score methods, Fuzzy RDD assumes the existence of a stochastic treatment assignment process $P(t|r)$. But $P(t|r)$ is also assumed to be discontinuous. The structural equations for Fuzzy RDD is:

$$y_i f(r_i) + \tau t_i + \epsilon_{yi} = f(r_i) + \pi_2 \mathbb{1}(r_i > r_0) + \epsilon'_{yi}, \ t_i = g(r_i) + \pi_1 \mathbb{1}(r_i > r_0) + \epsilon_{ti} \qquad (28)$$

where $\tau = \frac{\pi_2}{\pi_1}$ is the ATE we want to estimate, ϵ_{yi}, ϵ'_{yi} and ϵ_{ti} are the noise terms. As τ is a division between the causal effects $\mathbb{1}(r > r_0) \rightarrow t$ and $\mathbb{1}(r > r_0) \rightarrow y$, Fuzzy RDD can be considered as an IV method where the discontinuous variable $\mathbb{1}(r > r_0)$ plays the role of IV. A practical guide of RDD can be found in [23].

3.3 Advanced Methods for Learning Causal Effects from Big Data

The success of machine learning inspires advanced methods for learning causal effects with big data. We cover two types of methods: learning causal effects with neural networks and ensembles. **Learning Causal Effects with Neural Networks.** A straightforward way to learn causal effects with neural networks is to learn representations for features. To study the causal effect of forming a group on receiving a loan in a microfinance platform, GloVe [116] and Recurrent Neural Networks (RNN) [103] embed text features into a low-dimensional space. Then outcomes are inferred by fitting a function $f(\mathbf{h}, t)$ on factual outcomes y. In [120], Pham and Shen propose to apply neural networks to estimate the probability distributions such as $\hat{P}(y|t, \mathbf{x})$ and $\hat{P}(t|\mathbf{x})$.

Moreover, a series of work learns representations of confounders instead of relying on observed features. The 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 allows us to go beyond the unconfoundedness assumption. With specific deep learning models such as the *Balancing Counterfactual Regression* [77], the *TARnet* [138], and the *Causal Effect Variational Autoencoder* (CEVAE) [97], we can learn representations $\mathbf{z}_i$ of each instance i based on $(\mathbf{x}_i, d_i, y_i)$. Here, we introduce the most recent method, namely the CEVAE, which represents advances along this line.

With the recent advances in variational inference for deep latent variable models, Louizos et al. [97] propose the CEVAE. The CEVAE consists of the inference network and the model network. The inference network is the encoder. Given an instance $(\mathbf{x}_i, t_i, y_i)$, the encoder learns a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ from which we can sample its latent representation $\mathbf{z}_i$. Then, the model network is the decoder that reconstructs the data from the latent representation. The two neural networks are shown in Fig. 8. Those variational distributions ($q(\cdot)$) approximate the corresponding infeasible posterior distributions. Similar to the VAE [81] 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:

$$\mathcal{L} = \sum_i E_{q(z_i|\mathbf{x}_i, t_i, y_i)}[\log P(\mathbf{x}_i, t_i|\mathbf{z}_i) + \log P(y_i|t_i, \mathbf{z}_i) + \log P(\mathbf{z}_i) - \log q(\mathbf{z}_i|\mathbf{x}_i, t_i, y_i)]. \qquad (29)$$

The main difference between the CEVAE and the regular VAE is that, in CEVAE, there is a data point, $(\hat{y}_i^t, t_i, \hat{\mathbf{x}}_i)$ reconstructed for each combination of instance and treatment (i, t), which enables the inference of counterfactual outcomes once the neural networks in Fig. 8 are trained. The comparisons in [97, 138] with three benchmark datasets (i.e., IHDP, Twins, and Jobs) show that representation learning methods [97, 138] are the state-of-the-art for learning causal effects. **Learning Heterogeneous Causal Effects with Ensembles.** Ensemble models achieve the state-of-the-art performance in many supervised learning problems. With ensemble models, we can 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. In [65], Hill proposes to apply Bayesian Additive Trees (BART) [28] to estimate CATE. In particular, BART takes the features and the treatment as input and output the

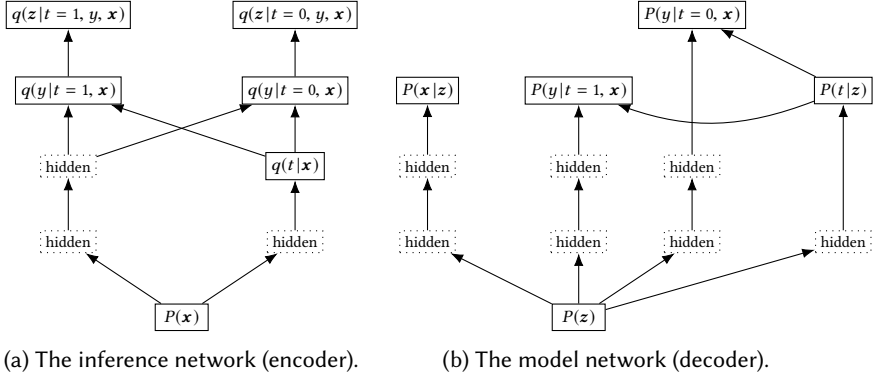(a) The inference network (encoder). (b) The model network (decoder).

Fig. 8. The neural network structures of CEVAE. The parameters, i.e., mean and variance, of each variational distribution $q(\cdot)$, are outputs of the neural network layers below it.
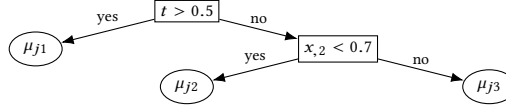

Fig. 9. A subtree $g(\boldsymbol{x}, t)$ in BART

distribution of potential outcomes as $f(\boldsymbol{x}, t) = \mathbb{E}[y|t, \boldsymbol{x}]$, which returns the sum of the outputs of Q Bayesian regression trees as $f(\boldsymbol{x}, t) = \sum_{j=1}^{Q} g_j(\boldsymbol{x}, t)$. Thus, we can estimate the CATE for given $\boldsymbol{x}$ as $\hat{\tau}(\boldsymbol{x}) = f(\boldsymbol{x}, 1) - f(\boldsymbol{x}, 0)$. Each subtree is defined by the tree structure and a set of b leaf nodes $\{\mu_{j1}, ..., \mu_{jb}\}$. An example of a BART subtree is shown in Fig. 9, where each interior node (rectangle) sends an instance to one of its children. The k-th node of the j-th subtree has a parameter μ_{jk}, i.e., the mean outcome of the instances classified to this node. BART has several advantages [59, 65]: (1) it is good at capturing non-linearity and discontinuity, (2) it needs little hyperparameter tuning and (3) it infers posterior distribution of outcomes, which allows uncertainty quantification. Hahn et al. [59] propose to handle the problem regularization-induced confounding (RIC) with BART [58]. RIC happens when the potential outcomes heavily depend on the features rather than the treatment.

In [156], Wager and Athey propose the Causal Forest which outputs asymptotically normal and consistent estimation of CATE. Each tree in the causal forest 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 $\boldsymbol{x}$. Then leaf 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 $\hat{\tau}_j(\boldsymbol{x}) = \frac{1}{|U_l^1|} \sum_{i \in U_l^1} Y_i - \frac{1}{|U_l^0|} \sum_{i \in U_l^0} Y_i$, where $U_l^t = \{i|t_i = t, L_j(\boldsymbol{x}_i) = l\}$ refers to the subset of instances that are sent to the l-th leaf of the j-th subtree whose treatment is t. Then the causal forest simply outputs the average of the CATE estimation from the J trees as $\hat{\tau}(\boldsymbol{x}) = \frac{1}{J} \sum_j \hat{\tau}_j(\boldsymbol{x})$. Note that there are studies dealing with the case where heterogeneous subpopulations cannot be identified by features such as *principle stratification* [44, 155].

Sometimes, an instance's treatment or outcome depends on other instances. For example, the customer flow of restaurants in the same area may amount to a constant. Besides features, treatments, and outcomes, information such as networks, temporal sequential structures can be utilized to capture such dependencies. Learning causal effects with non-i.i.d data can be done by modeling interference [122] and disentangling instances via i.i.d. representations [55–57].

(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. 10. Two exemplary causal graphs that belong to an equivalence class

While this survey focuses on observational studies, A/B tests play a crucial role in tasks like estimating ATEs, exploring design space and attribute effects to causes for practical decision making [15]. Bakshy et al. [15] developed a domain-specific language, PlanOut, for deploying Internet-scale A/B tests. For A/B tests, big data raises unique challenges. Taddy et al. [148] pointed out the problems that impede using standard tools for inferring ATEs in heavy-tailed distributions (e.g., Internet transaction data): slower learning rate, invalid Gaussian assumption and the failure of nonparametric bootstrap estimators on sampling uncertainty about the mean. They propose a semi-parametric model for the data generating process of heavy-tailed data to address these issues.

4 CAUSAL DISCOVERY: LEARNING CAUSAL RELATIONS

In this section, we start with the problem statement and evaluation metrics. Then a review of traditional causal discovery methods followed by methods for causal discovery from big data. In learning causal relations (causal discovery), we examine whether a causal relation exists.

Definition 13. Learning Causal Relations. *Given J variables, $\{x_{,j}\}_{j=1}^{J}$, we aim to determine whether the j-th variable $x_{,j}$ changes if we modify the j'-th variable $x_{,j'}$ for all $j \neq j'$.*

In the running example, learning causal relations enable us to answer the questions such as: *Do features such as location causally affect the customer flow? Is location a confounder for the causal effect of Yelp rating on customer flow?* To achieve this, we postulate that causality can be detected amongst statistical dependencies [119, 133]. An algorithm solving this problem learns a set of causal graphs as candidates [144]. To evaluate the learned causal relations, 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 different causal graphs.

Definition 14. Equivalence Class. *We say that two causal graphs G and G' belong to the same equivalence class iff each conditional independence that G has is also implied by G' and vise versa.*

Figs. 10a and 10b show two causal graphs that belong to the same equivalence class. They share the same set of conditional independence $\{x_{,2} \perp\!\!\!\perp x_{,3}|x_{,1}\}$.
Evaluation Metrics. Here, we briefly introduce the evaluation metrics adopted by the field of learning causal relations. The metrics can be categorized into two types: (1) the distances between the learned causal graph and the ground truth and (2) the accuracy of discovered causal relations. For the first category, a comprehensive study of metrics comparing learned causal graphs (Bayesian networks) can be found in [35]. In [26], Chickering et al. counted the number of learned causal graphs G' that have the ground truth G as a subgraph. The structural Hamming distance (SHD) has been widely adopted in [117, 153]. SHD is defined as the number of edits (adding, removing or reversing an edge) that have to be made to the learned graph G' for it to become the ground truth G. In [142], the distance between two graphs are measured by the Frobenius norm of the difference between their adjacency matrices. The second type of metrics are introduced based on the fact that the discovery of an adjacency relation and an arrowhead can be treated as a binary classification problem. In [4], the precision, recall (true positive rate) and false positive rate for both adjacencies

(discovered neighbors) and arrowheads (direction of discovered causal relations) are defined as:

$$precision = \frac{TP}{TP + FP}, \ recall \ (tpr) = \frac{TP}{TP + FN}, \ fpr = \frac{FP}{TN + FP} \tag{30}$$

The areas under the precision-recall and fpr-tpr (ROC) curves are widely used [13, 153].

4.1 Traditional Methods for Learning Causal Relations

Following [100], we review three families of algorithms: constraint-based, score-based, and those based on functional causal models. The first two rely on statistical tests to discover candidate causal graphs, and the third learns causal relations by estimating coefficients in structural equations.

Constraint-based (CB) Algorithms learn a set of causal graphs that satisfy the conditional independence embedded in the data. Statistical tests are utilized to verify if a candidate graph satisfies the independence based on the *faithfulness* assumption [144]:

Definition 15. Faithfulness. *Conditional independence between a pair of variables, $x_{,j} \perp\!\!\!\perp x_{,j'}|z$ for $x_{,j} \neq x_{,j'}, z \subseteq x \setminus \{x_{,j}, x_{,j'}\}$, can be estimated from a dataset X iff z d-separates $x_{,j}$ and $x_{,j'}$ in the causal graph $G = (\mathcal{V}, \mathcal{E})$ which defines the data-generation process of X.*

Faithfulness means the statistical dependence between variables estimated from the data does not violate the independence defined by any causal graph which generates the data [144]. The main challenge is the computational cost as the number of possible causal graphs is super-exponential to the number of variables. Hence, existing algorithms focus on reducing the number of tests.

The Peter-Clark (PC) Algorithm. The PC algorithm [144] works in a two-step fashion. First, it learns an undirected (*skeleton graph*) from data. Then, it detects the directions of the edges to return an equivalent class of causal graphs. It starts with a fully connected graph and the depth $q = 0$. Then for each ordered pair of connected variables $(x_{,j}, x_{,j'})$, it tests if the conditional independence $x_{,j} \perp\!\!\!\perp x_{,j'}|\tilde{z}$ is satisfied for each $\tilde{z} \subseteq \mathcal{N}(x_{,j})$ or $\tilde{z} \subseteq \mathcal{N}(x_{,j'})$ of size q, where $\mathcal{N}(\cdot)$ is the set of neighbors. If the conditional independence holds, it removes the edge $(x_{,j}, x_{,j'})$ and saves $\tilde{z}$ as the separating set of $(x_{,j}, x_{,j'})$. Once all such edges are removed, the depth q increases by 1 and this process continues till the number of neighbors for each variable is less than q. In the second step, we decide the directions of edges. We first determine *v-structures*. For a triple $(x_{,j}, x_{,j'}, x_{,j''})$ with no edge between $x_{,j}$ and $x_{,j''}$, we make it a v-structure $x_{,j} \rightarrow x_{,j'} \leftarrow x_{,j''}$ iff $x_{,j'} \notin \tilde{z}$, where $\tilde{z}$ denotes saved separating set of $x_{,j}$ and $x_{,j''}$. Note that no new v-structures would be created as a result of edge orientation. Then the remaining undirected edges are oriented following the three rules: (1) we orient $x_{,j} - x_{,j'}$ to $x_{,j} \rightarrow x_{,j'}$ if there exists an edge $x_{,j''} \rightarrow x_{,j'}$ and $x_{,j''}$ and $x_{,j}$ are not neighbors. (2) we orient $x_{,j} - x_{,j'}$ to $x_{,j} \rightarrow x_{,j'}$ if there is a chain $x_{,j} \rightarrow x_{,j''} \rightarrow x_{,j'}$. (3) we orient $x_{,j} - x_{,j'}$ to $x_{,j} \rightarrow x_{,j'}$ if there are two chains $x_{,j} - x_{,k} \rightarrow x_{,j'}$ and $x_{,j} - x_{,l} \rightarrow x_{,j'}$.

Other CB algorithms include the IC algorithm [113] and their variants [78, 88]. However, most standard statistical tests require Gaussian or multinomial distributions. To overcome these restrictions, novel conditional independence tests are proposed to cover other families of distributions [45, 124, 136, 164]. To take unobserved confounders into consideration, algorithms such as FCI (fast causal inference) and its extensions [31, 145] are proposed to search through an extended space of causal graphs. Moreover, to go beyond observational data, Kocaoglu et al. [82] considered the problem of designing a set of interventions with minimum cost to uniquely identify any causal graph from the given skeleton. They show the problem can be solved in polynomial time.

CB algorithms can also be applied to non-i.i.d. data such as time series. For example, the FCI algorithm has been adapted for time series [29, 42]. *Time series models with independent noise* (TiMINo) [118], a robust algorithm based on non-linear independent tests, can avoid discovering false relations with a misspecified model. TiMINo takes time series as input and outputs a DAG or remains undecided.

There are two main drawbacks of this family of algorithms. First, the faithfulness assumption can be violated. For example, with limited samples, independence tests may even contradict each other. Second, the causal direction between two variables may remain unknown.

Score-based (SB) Algorithms. To relax the faithfulness assumption, SB algorithms replace conditional independence tests with the goodness of fit tests. SB algorithms learn causal graphs by maximizing the scoring criterion $S(X, G')$ which returns the score of the causal graph G' given data X. Intuitively, low scores should be assigned to the graphs which embed incorrect conditional independence. For goodness of fit tests, two components need to be specified: the structural equations and the score function. First, we consider the structural equations. 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 [134] is the most widely adopted metric $S(X, G') = \log P(X|\hat{\boldsymbol{\theta}}, G') - \frac{J}{2} \log(n)$, where $\hat{\boldsymbol{\theta}}$ is the MLE of the parameters, J denotes the number of variables 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 [129], 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 [62] assumes Dirichlet prior on parameters for the multinomial distributions of variables. With the two components fixed, the 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. Searching for the causal graph with maximal score, also known as structural learning is both NP-hard and NP-complete [25, 27]. It is not computationally feasible to score all possible causal graphs exhaustively. Therefore, heuristics such as GES [26] and its extension, Fast GES (FGES) [123] are proposed to reach a locally optimal solution. When it comes to interventional data, Wang et al. [157] propose algorithms to learn causal relations when a mixture of interventional and observational data is given, which are non-parametric and handle non-Gaussian data well.

Greedy Equivalence Search (GES). Here we introduce GES as an example of SB algorithms. In [26], assuming discrete variables, the BDeu criterion is used:

$$S_{BDeu}(G', X) = \log \prod_{j=1}^{J} 0.001^{(r_j-1)q_j} \prod_{k=1}^{q_j} \frac{\Gamma(10/q_j)}{\Gamma(10/q_j + N_{jk})} \prod_{l=1}^{r_j} \frac{\Gamma(10/(r_i q_i) + N_{jkl})}{\Gamma(10/(r_j q_j))}, \quad (31)$$

where r_j and q_j signify the number of configurations of variable $x_{,j}$ and parent set $\boldsymbol{Pa}_j$ specified by the graph G'. $\Gamma(n) = (n-1)!$ is the Gamma function. N_{jkl} denotes the number of records for which $x_{,j} = k$ and $\boldsymbol{Pa}_j$ is in the k-th configuration and $N_{jk} = \sum_l N_{jkl}$. After initialized with the equivalent class of DAG models with no edges, two-stage greedy search is performed. First, a greedy search is performed only to insert edges. The insertion operator $Insert(x_{,j}, x_{,j'}, z)$ takes three inputs: $x_{,j}$ and $x_{,j'}$ are non-adjacent nodes in the current graph, z denotes any subset of $x_{,j'}$'s neighbors that are not adjacent to $x_{,j}$. The insertion operator modifies the graph by (1) adding the edge $x_{,j} \to x_{,j'}$ and (2) directing the previous undirected edge $z - x_{,j'}$ as $z \to x_{,j'}$. It is worth mentioning that undirected edges can result from the equivalent class of graphs. As a greedy algorithm, in each iteration, for the current graph, we find the triple $x_{,j}, x_{,j'}, z$ leading to the best score (Eq. 31) and perform the insert operator until a local maximum is reached. Then, the second greedy search is performed initialized with the local optimum of the previous phase, only to delete edges. The delete operator, $Delete(x_{,j}, x_{,j'}, z)$ takes two adjacent nodes $x_{,j}$ and $x_{,j'}$ with edge $x_{,j} - x_{,j'}$ or $x_{,j} \to x_{,j'}$

and z denoting any subset of neighbors of $x_{,j'}$ which are also adjacent to $x_{,j}$. For each iteration, given the current graph, the triple $x_{,j}, x_{,j'}, z$ with the highest score is selected to update the graph with the delete operator. GES terminates when the local maximum is reached in the second phase.

Hybrid algorithms [153, 159] that exploit principled ways to combine CB and SB algorithms have also attracted considerable attention. The MMHC algorithm [153] is proposed toward enough scalability for thousands of variables. First, it learns the skeleton of a causal graph using the Max-Min Parents and Children (MMPC) algorithm [152], which is similar to the CB algorithms. Then, it orients the edges with Bayesian scoring hill climbing search, which is similar to SB algorithms.

Algorithms based on Functional Causal Models (FCMs). In FCMs, a variable $x_{,j}$ can be written as a function of its directed causes $\boldsymbol{Pa}_j$ and some noise term ϵ_j as $x_{,j} = f(\boldsymbol{Pa}_j, \epsilon_j)$. Different from the two families of methods mentioned above, with FCMs, we are able to distinguish between different DAGs from the same equivalent class. Here, we adopt Linear Non-Gaussian Acyclic Model (LiNGAM) [141] as the FCM to introduce algorithms. The LiNGAM model can be written as: $\boldsymbol{x} = \boldsymbol{Ax} + \boldsymbol{\epsilon}$, where $\boldsymbol{x}$, $\boldsymbol{A}$ and $\boldsymbol{\epsilon}$ denote the vector of variables, the adjacency matrix of the causal graph [140], and the vector of noise, 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 relations turns into estimating a strictly lower triangle matrix $\boldsymbol{A}$ which determines a unique causal order $k(j)$ for each variable $x_{,j}$. For example, if a FCM can be specified by a LiNGAM as:

$$\begin{bmatrix} s \\ d \\ y \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1.2 & 0 & 0 \\ 0.8 & 1.3 & 0 \end{bmatrix} \begin{bmatrix} s \\ d \\ y \end{bmatrix} + \begin{bmatrix} \epsilon_s \\ \epsilon_d \\ \epsilon_y \end{bmatrix}, \tag{32}$$

then the causal order of the three variables s, d, y is 1, 2 and 3, respectively.

ICA-LiNGAM. Based on independent component analysis (ICA) [70], the ICA-LiNGAM algorithm [141] is proposed to learn causal relations with the LiNGAM model, with which we estimate the matrix $\boldsymbol{A}$. First, we can rewrite Eq. 4.1 as $\boldsymbol{x} = \boldsymbol{B\epsilon}$, where $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{A})^{-1}$. As each dimension of $\boldsymbol{\epsilon}$ is assumed to be independent and non-Gaussian, it defines the ICA model for the LiNGAM. Thus we can apply ICA to estimate $\boldsymbol{B}$. Given data $\boldsymbol{X}$ of the variables $\boldsymbol{x}$, we use ICA algorithm [70] to obtain the decomposition $\boldsymbol{X} = \boldsymbol{BS}$. We can learn $\boldsymbol{W} = \boldsymbol{B}^{-1}$ by maximizing the objective:

$$\sum_j J_G(\boldsymbol{w}_j) \ s.t. \ \mathbb{E}[(\boldsymbol{w}_k^T \boldsymbol{x})(\boldsymbol{w}_l^T \boldsymbol{x})] = \delta_{kl}, \tag{33}$$

where $J_G(\boldsymbol{w}_i) = \{\mathbb{E}[G(\boldsymbol{w}_i^T \boldsymbol{x})] - \mathbb{E}[G(\boldsymbol{v})]\}^2$, G can be any nonquadratic function (e.g., $G(y) = y^4$). $\boldsymbol{v}$ denotes samples from a normal distribution $\mathcal{N}(0, 1)$ and δ_{kl} is the magnitude of dependence between the two variables. Then an initial estimate of $\boldsymbol{A}$, namely $\boldsymbol{A}'$, is computed based on $\boldsymbol{W}$ as $\boldsymbol{A}' = \boldsymbol{I} - \tilde{\boldsymbol{W}}'$. $\tilde{\boldsymbol{W}}'$ is obtained by dividing each row of $\tilde{\boldsymbol{W}}$ by the corresponding diagonal element. $\tilde{\boldsymbol{W}}$ is calculated by finding the unique permutation of rows of $\boldsymbol{W}$ which is nonzero on the diagonal. 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 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. propose the DirectLiNGAM algorithm [142], 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 relations between variables and the distribution of noise [68, 69]. 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 [69, 119]. One step further, Post-nonlinear Models expand the functional space with non-linear relations between the variables and the noise [163].

Those algorithms for FCMs (e.g., ICA-LiNGAM) can also be adapted to handle time series data. For example, an auto-regressive LiNGAM is proposed to learn causal relations from time series [71].

4.2 Learning Causal Relations from Big Data

The challenges raised by big data for learning causal relations include (1) handling high-dimensional data and (2) dealing with large-scale mixed data.

Causal Discovery from High-dimensional Data. There are two questions to answer before applying a causal discovery algorithm to high-dimensional data: (i) Is it theoretically consistent in the high-dimensional setting? (ii) Is it scalable for high-dimensional data? Traditional CB algorithms (e.g., the PC algorithm [144] and its variants [13]) have been shown to be applicable to high-dimensional datasets with thousands of variables in terms of both consistency and scalability [47, 89]. Following [13], in [30], Colombo and Maathuis point out that outputs of CB methods that use the PC algorithm for obtaining the skeleton depend on the order in which variables are given. This can be a pronounced issue leading to highly variable results in high-dimensional settings. Therefore, they propose a new skeleton search algorithm which can be widely used in a series of CB algorithms including the PC, FCI, RFCI and CCD algorithms [31, 125, 144].

The case is different for SB algorithms as the traditional work leaves the consistency and scalability in the high dimension setting as open problems [107]. In [123], Ramsey et al. examine two modifications to address the scalability problem of the GES algorithm [26]: (1) FGES and (2) FGES with Markov Blanket Search (FGES-MB). FGES-MB modifies the first stage of GES by only adding edges found by two-hop markov blanket search on each variable [123]. Their experiments show the FGES algorithm with BIC score and FGES-MB can maintain high precision and good recall when the causal graph has one million Gaussian variables with average degree 2. Nandy et al. [107] find that we can theoretically achieve the consistency in high dimensional settings when adaptive restrictions are added to the GES algorithm [26] on the search space. The proposed variant of GES, ARGES, restricts the choice of variable z in the first stage of the GES algorithm to those form v-structures $(x_{,j}, x_{,j'}, z)$. Then theoretically, they show that ARGES is consistent in several sparse high-dimensional settings. Experimental results show ARGES scales to 400 samples of 2,400 variables and outperforms the baselines (MMHC [153], PC [144] and GES [26] etc.) measured by the ROC curve. When it comes to hybrid algorithms, as we mention in Section 4.1, the MMHC algorithm [153] can scale to thousands of variables.

Learning Causal Relations from Mixed Data. Big data often contains both continuous and discrete variables, which is referred to as mixed data. Mixed data poses challenges for learning causal relations as continuous and discrete variables often require different independence tests (for CB methods) and goodness of fit tests (for SB methods). Raghu et al. [121] performed an extensive empirical study on the CB and SB methods for learning causal relations with mixed data. To adapt CB algorithms to mixed data, a straightforward solution is to consider the independence tests that handle mixed data such as a multinomial logistic regression and a conditional Gaussian test (CG). In the experiments, the authors focused on the small sample high dimension setting, which often fits the case of biomedical data. Empirical results show that using these independence tests for mixed data can significantly boost the recall of discovered causal relations when there are 100 samples for 100 variables. In [135], authors find that modeling the joint distribution of both continuous and discrete variables with mixed graphical models (MGMs) [91] as a preprocessing step can improve CB and SB methods with mixed data. Note that the joint distribution implies a graph structure over these heterogeneous variables.

For scalable causal discovery on high-dimensional large-scale mixed data, Andrew et al. [4] derived the degenerate Gaussian (DG) score and proved its consistency. With the four metrics: the

precision and recall of arrowheads and those of adjacencies, results show that DG outperformed CG [121], MGM [91] and copula PC [33] on synthetic datasets with 500 variables and 1,000 samples.

At the end, we point out that most of the existing algorithms share a limitation: they can only be applied to discovering the causal relations of variables whose samples have been observed in the training data. Developing models that can be generalized to new variables remain an open problem.

5 CONNECTIONS TO MACHINE LEARNING

This section covers connections between learning causality and supervised and semi-supervised learning, domain adaptation, and reinforcement learning. We explore two aspects: How can causal knowledge improve machine learning? How can machine learning help learning causality?

5.1 Supervised Learning and Semi-supervised Learning

Supervised Learning. From a data perspective, some problems of learning causality can be reduced to supervised learning or semi-supervised learning problems.

Here, we discuss how supervised learning algorithms can help learning causality. The problem of learning causal relations can be transformed into as a prediction problem once we label the data with causal relations. In particular, suppose we are given labeled training data of the form $(c_1, a_1), ..., (c_N, a_N)$ where each c_j is an i.i.d. dataset $c_j = (X_1, y_1), ..., (X_{N_j}, y_{N_j})$ sampled from a joint distribution $P_j(x, y)$ and each dataset has an additional label $a_j \in (\rightarrow, \leftarrow)$ describing whether the dataset is *causal* $x \rightarrow y$ or *anti-causal* $y \rightarrow x$. Anti-causal means that the label y is the cause of the features x. The main challenge here is to obtain the label of causal direction. For some datasets, the causal relations are naturally revealed [96]. In addition, we can leverage the knowledge that a dataset is causal or anti-causal to improve supervised learning models. Causal regularization [14, 139] is proposed to learn more interpretable and generalizable models. In [14], a causal regularizer guides predictive models towards learning causal relations between features and labels. It is assumed that, besides the predictive model, there is a classifier $c^i = P(x^i$ does not cause $y)$ outputs whether a feature x^i causes the label y. Then the objective function of a predictive model with the causal regularizer is formulated as:

$$\arg\min_{w} \frac{1}{n} \sum_{j=1}^{n} \mathcal{L}(x_j, y_j | w) + \lambda \sum_{i=1}^{m} c^i |w^i|, \tag{34}$$

where $\mathcal{L}$ denotes the loss function of the predictive model with parameters w. Intuitively, the lower the probability of a feature to be a cause, the more penalty will be added to its corresponding weight, which eventually encourages the model to pay more attention to those features that are more likely to be causes of the label. In [83], the following causal regularizer is proposed to set each feature as the treatment and learn sample weights such that the distribution of the two groups can be balanced w.r.t. to each treatment (feature):

$$\sum_{j=1}^{m} || \frac{X_{,-j}^T (w \odot I_{,j})}{w^T I_{,j}} - \frac{X_{,-j}^T (w \odot (e - I_{,j}))}{w^T (e - I_{,j})} ||_2^2, \tag{35}$$

where $w \in \mathbb{R}^n$ signifies the sample weights, e denotes the $n \times 1$ vector with all elements equal to 1, $X_{,j}$ and $X_{,-j}$ are the j-th column of the feature matrix and the matrix of remaining features, $I_{i,j}$ refers to the treatment status of the i-th instance when the j-th feature is set as the treatment. The authors added a constraint to the original loss function of a logistic regression model to ensure the value of this causal regularizer is not greater than a predefined hyperparameter $\gamma \in \mathbb{R}^+$. Doing this can help identify causal features and construct robust predictive model across different domains.

Semi-Supervised Learning (SSL). A machine learning problem can be either causal or *anti-causal* [133]. For example, in hand written digits recognition [90], which digit to write is first determined, then the digit would be represented as a matrix of pixel values. Such causal structure has implications for SSL. In SSL, the target is to improve the quality of estimated $P(y|\boldsymbol{x})$ with additional unlabeled instances which can provide information of the marginal distribution $P(\boldsymbol{x})$. We can first consider the cases when semi-supervised learning would fail. For example, when $p(\boldsymbol{x})$ is a uniform distribution, observing more unlabeled instances provides no information about $P(y|\boldsymbol{x})$. In contrast, in the case of representation learning, if $\boldsymbol{h}$ contains hidden causes of the features $\boldsymbol{x}$, and the label y is one of the cause of $\boldsymbol{x}$, then predicting y from $\boldsymbol{h}$ is likely to be easy [51]. Specifically, the true data-generating process implies $\boldsymbol{h}$ is a parent of $\boldsymbol{x}$, and thus, $p(\boldsymbol{h}, \boldsymbol{x}) = P(\boldsymbol{h})P(\boldsymbol{x}|\boldsymbol{h})$. So the marginal distribution is $P(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{h}}[P(\boldsymbol{x}|\boldsymbol{h})]$. With $P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{p(\boldsymbol{x})}$, we know $P(\boldsymbol{x})$ directly affects $P(y|\boldsymbol{x})$. Therefore, the causal structure of $P(\boldsymbol{x})$ can help the prediction of $P(y|\boldsymbol{x})$, which achieves the target of SSL. However, the number of causes can be extremely large. For example, a positive Yelp review can result from good service, delicious food, cheap price, or decent restaurant environment. Brute force solutions are not feasible as it is often impossible to capture most of the causes. So, we need to figure out what causes to encode for a certain target y. Criteria such as mean squared error on reconstructed features are used to train autoencoders and generative models, which assumes that a latent variable is salient iff it affects the value of most features. However, there can be tasks where the label is only associated with few causes. For example, to predict the customer flow of truck drivers in fast food restaurants near highway, few hidden causes may be useful. Therefore, the criteria need to be adaptable in accordance with the task. Generative Adversarial Networks (GAN) [52] are proposed to address this issue for images. GAN can adapt its criteria s.t. the latent variables (e.g., ears of human in human head images) that only affect the value of few features can also be learned as representations. Making optimal decisions on which causes we learn representation for is still an open problem.

Janzing and Schölkopf [75] consider a special case of SSL: let $y = f(x)$ and $x, y \in [0, 1]$, where f is an unknown anti-causal model. Given $n-1$ labeled instances $\{(x_i, y_i)\}_{i=1}^{n-1}$, we seek to infer the label $y_n = f(x_n)$ of an unlabeled instance x_n. In this setting, it is proved that SSL outperforms supervised learning when $P(x)$ and f are dependent and a certain independence between $P(y)$ and $g = f^{-1}$ holds. The independence $P(y) \perp\!\!\!\perp g$ is assumed and can be defined as $Cov[P(y), \log g'] = 0$, where g' denotes the derivative of g and $\log g', P(x) \in [0, 1]$. Given that, it can be shown that $Cov[P(x), \log f'] > 0$, which means $P(x)$ contains information of the function f we aim to learn. In addition, the fact that SSL only works in the anti-causal direction can help learn causal relations [137]. As shown above, if the problem is anti-causal, we expect that better knowledge of $P(x)$ helps prediction of $P(y|x)$ as they are dependent. In contrast, if it is a causal problem, then knowing $P(x)$ barely helps us learn $P(y|x)$. Therefore, comparing the errors of estimations on $P(y|x)$ and $P(x|y)$ enables us to determine the direction of causality. In particular, Gaussian process (GP) regression models are trained to estimate $P(x|y)$ as:

$$P(x|y, \boldsymbol{y}^*) = \int_{\mathcal{Z}, \theta} P(z, \theta, x|\boldsymbol{y}^*, y)dzd\theta \approx \int_{\mathcal{Z}, \theta} P(x|y, \boldsymbol{y}^*, z, \theta)P(z, \theta|\boldsymbol{y}^*)dzd\theta, \qquad (36)$$

where z signifies the latent variables and $\boldsymbol{y}^* = (y_1, ..., y_{n-1})$ are the observed data points. $\mathcal{Z}$ and θ are the set of possible values of the latent variables and the model parameters, respectively. The first factor $P(x|y, \boldsymbol{y}^*, z, \theta)$ is the supervised GP regression and the second factor $P(z, \theta|\boldsymbol{y}^*)$ denotes the posterior distribution over z and θ given observed labels $\boldsymbol{y}^*$. Assuming uniform priors for z and θ, using Bayes's rule $p(z, \theta|\boldsymbol{y}^*) = \frac{p(\boldsymbol{y}^*|z, \theta)p(z)p(\theta)}{p(\boldsymbol{y}^*)} \propto p(\boldsymbol{y}^*|z, \theta)$ which is parameterized by a Gaussian distribution defined by GP-LVM [150]. Thus, we can estimate $P(x|y, \boldsymbol{y}^*) = \frac{1}{m} \sum_i p(x|y, \boldsymbol{y}^*, z^i, \theta^i)$

with m MCMC samples from $p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}^*)$. In a similar way, we can find that $p(x|y, \boldsymbol{y}^*, \boldsymbol{z}^i, \boldsymbol{\theta}^i)$ is also proportional to a Gaussian distribution defined by GP-LVM. Thus, we can estimate $P(x|y)$ as mentioned above and $P(y|x)$ by repeating the procedure with x and y swapped. Finally, log likelihood of the two estimates reveals the causal direction. However, it is an open problem to scale such approaches (regression for causal discovery) to high-dimensional, noisy or non-i.i.d. data.

There exists a special type of SSL regarding learning causal effects. Given massive data on $(\boldsymbol{x}, t)$ but small samples with observed outcomes y, Hahn et al. [59] highlighted that information of $P(\boldsymbol{x}, t)$ can be brought to help the estimation of $P(y|\boldsymbol{x}, t)$.

5.2 Domain Adaptation

Domain adaptation [19, 34] studies how to adapt machine learning models trained in some domains to the others. One application of domain adaptation is to improve prediction accuracy when we have plenty of labeled data from the source domain (e.g., Yelp review) but not for the target domain (e.g., reviews from another website). Domain adaptation is related to learning causality by *invariant prediction in different domains* [117], assuming that causal relations do not change across domains. Given a target variable y^e and m predictor variables $\boldsymbol{x}^e = (x_1^e, ..., x_m^e)$ from different domains $e \in \{1, ..., E\}$, the goal is to predict the value of y. Invariant prediction assumes that the conditional $P(y|\boldsymbol{Pa}_y)$ is consistent for all domains, where $\boldsymbol{Pa}_y$ is a set of direct causes of y. Formally,

$$P(y^e|\boldsymbol{Pa}_y^e) = P(y^f|\boldsymbol{Pa}_y^f). \tag{37}$$

The assumption is valid when the distributions are induced by an underlying SCM and the different domains correspond to different interventional distributions where y is not under intervention. Then we can conclude that (1) invariant prediction is achieved and (2) $\mathcal{Z}^*$ is the set of estimated causes of the target variable y. In [117], the authors propose a method to estimate $\boldsymbol{Pa}_y$. Assuming that the collection $\mathcal{S}$ consists of all subsets S of features that result in *invariant prediction*, satisfying $P(y^e|x_S^e) = P(y^f|x_S^f)$, the variables appearing in each such set S form the estimated causes of the label $\boldsymbol{Pa}_y$. Finally, a valid subset S^* that achieves the best performance in the source domains is selected as the features for cross domain prediction. This is because the selected subset is guaranteed to be optimal in terms of domain generalization error. Due to the independent mechanism assumption [119, 128], the selected subset is also robust against arbitrary changes of marginal distribution of predictors in the target domains. Similar results for domain generalization have been obtained through a global balancing approach [83] and a causal feature selection method [110]. They are based on sample re-weighting. In [83], the proposed model, Deep Global Balancing Regression (DGBR), leverages an auto-encoder model to map data into a latent space before reweighting instead of directly reweighting the original samples [110]. We summarize the usage of the low-dimensional representations of DGBR in two ways: (1) They are used in the global balancing regularizer, where each variable is successively set as the treatment variable. Then we balance all the variables via learning global sample weights. (2) We can predict outcomes based on the representations using regularized regression. The causal regularizer of DGBR is:

$$\sum_{j=1}^{p} \left\| \frac{\phi(\boldsymbol{X}_{,-j})^T(\boldsymbol{w} \odot \boldsymbol{X}_{,j})}{\boldsymbol{w}^T \boldsymbol{X}_{,j}} - \frac{\phi(\boldsymbol{X}_{,-j})^T(\boldsymbol{w} \odot (\boldsymbol{e} - \boldsymbol{X}_{,j}))}{\boldsymbol{w}^T(\boldsymbol{e} - \boldsymbol{X}_{,j})} \right\|_2^2, \tag{38}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ signifies the sample weights, $\boldsymbol{e}$ denotes the $n \times 1$ vector with all elements equal to 1, $\boldsymbol{X}_{,j}$ and $\boldsymbol{X}_{,-j}$ are the j-th column of the feature matrix and the matrix of remaining features. In addition, the objective function is a weighted loss of logistic regression along with a constraint to limit the value of this causal regularizer not greater than a predefined positive hyperparameter. While for prediction under concept drift [158], where Eq. 37 is violated but the marginal distributions

of the predictors remain the same, one may allow apriori causal knowledge to guide the learning process and circumvent the discrepancies between the source and target domains [115], a.k.a. *causal transportability*. The study of transportability seeks to identify conditions under which causal knowledge learned from experiments can be reused in different domains with observational data only. A formal definition of causal transportability can be referred to [115]. In [17], the authors further provide a necessary and sufficient condition to decide, given assumptions about differences between the source and target domains, whether transportability is feasible.

5.3 Reinforcement Learning

Reinforcement learning (RL) [147] is studied for solving sequential decision-making problems. The key variables in RL are the action a, the state z, and the reward y. When an agent performs an action, it reaches the next state and receives a reward. The *Markov decision process* is often adopted, where the next state z_{t+1} depends on the current state z_t and action a_t and the reward of the next state y_{t+1} is determined by z_t, z_{t+1} and a_t. A RL model learns a *policy* $\pi(a_t, z_t) = P(a_t|z_t)$ which determines which action to take given the current state. 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 observes the customer flow. Then the owner will make decisions for the next time step based on whether the customer flow increases or not.

Unobserved Confounders in RL. Unobserved confounders raise issues of learning policies for RL models such as multi-armed bandits (MAB) [16]. Without knowing the causal model, MAB algorithms can perform as badly as randomly taking an action in each time step. Specifically, the Causal Thompson Sampling algorithm [16] is proposed to handle unobserved confounders in MAB problems. The reward distributions of the arms that are not preferred by the current policy 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 [119].

Unbiased Reward Prediction. Given trajectories (actions, states and rewards) of an observed policy, we can utilize causal inference methods to predict rewards for another policy, especially for Episodic RL (ERL) problems. ERL is a subclass of RL where the state is reset to the default value after a finite number of actions. ERL helps decision-making in many applications such as card games and advertisement placement [20]. One popular approach leverages IPTW for predicting reward of ERL models. In IPTW, a treatment refers to an action and the policy-specific propensity score is defined as the probability to perform the selected action given the observed state. Particularly, given trajectories produced by running an observed policy π L times $[(a_1(1), z_1(1)), (a_2(1), z_2(1)), ...], ..., [(a_1(n), z_1(n)), (a_2(n), z_2(n)), ...]$, we can estimate the expected sum of rewards of a policy $\tilde{\pi}$ with IPTW as:

$$\hat{\xi} := \frac{1}{L} \sum_{l=1}^{L} y(l) \frac{\prod_{k=1}^{K} \tilde{\pi}(a_k(l)|z_k(l))}{\prod_{k=1}^{K} \pi(a_k(l)|z_k(l))}, \tag{39}$$

where K is the number of time steps in each episode. Improved variants are proposed in [20].

RL with Auxiliary Causal Knowledge. There is a line of work to improve RL models with causal knowledge as side information [87, 160]. Here, we use the Causal Bandit (CB) problem [87] as an example. In this problem, given J binary variables $x_{,1}, ..., x_{,J}$ and their causal graph but not the causal mechanisms $x_{,j} = f(\boldsymbol{Pa}_j, \epsilon_j)$, we aim to find the intervention that is most likely to set a specified variable $x_{,k}$ to 1. An intervention is defined as a vector $\boldsymbol{a} \in \{*, 0, 1\}^J$, where
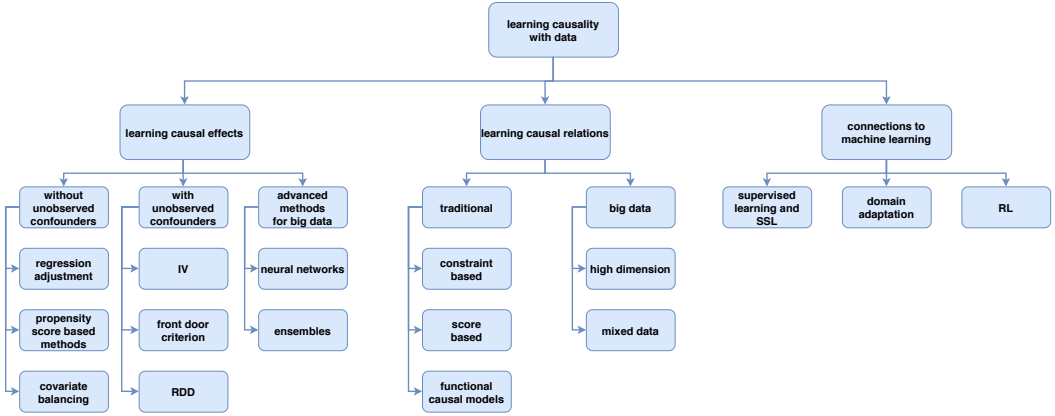
Fig. 11. Learning causality with data: a summary of the survey

$a_j \neq *$ means $x_{,j}$ is set to a_j. For exploration, a CB algorithm learns $\mu(\boldsymbol{a}) = P(x_{,k} = 1|do(\boldsymbol{a}))$, the probability that the target $x_{,k}$ is set to 1 by $\boldsymbol{a}$. For exploitation, a CB algorithm minimizes the regret $R = \mu(\boldsymbol{a}^*) - \mathbb{E}[\mu(\hat{\boldsymbol{a}})]$, where $\boldsymbol{a}^*$ is the optimal action and $\hat{\boldsymbol{a}}$ denotes the algorithm's selection. In [87], the parallel bandit (PB) algorithm is proposed to solve this problem with guarantee to outperform non-causal MAB algorithms. Given total rounds L, in first $L/2$ rounds, the PB algorithm collects observational data by doing intervention $\boldsymbol{a} = [*, ..., *]$. Then it analyzes the observational data for each intervention $\boldsymbol{a} = do(x_{,j} = x)$ to estimate the reward as $\hat{\mu}(\boldsymbol{a}) = \frac{1}{L_a} \sum_{l=1}^{L/2} \mathbb{1}(x_{l,j} = x_{,j}(\boldsymbol{a}))$ and probabilities as $\hat{p}_{\boldsymbol{a}} = \frac{2L_a}{L}, \hat{q}_j = \hat{p}_{do(x_{,j}=1)}$, where $L_a = \sum_{l=1}^{L/2} \mathbb{1}(x_{l,j} = x)$ denotes the number of times we observe what $\boldsymbol{a}$ could have done in the observational data. Next, we create the set of rarely observed actions as $\mathcal{A}' = \{\boldsymbol{a}|\hat{p}_{\boldsymbol{a}} \leq \frac{1}{\hat{m}}\}$, where $\hat{m}$ is a threshold defined by the vector of interventional distributions $\boldsymbol{q}$. Then we uniformly sample $\boldsymbol{a} \in \mathcal{A}'$ and observe the value of $x_{,k}$. At the end, we compute $\mathbb{E}[x_{,k}]$ of resulting each action as the estimated reward $\hat{\mu}(\boldsymbol{a})$ and select the one with the largest $\hat{\mu}(\boldsymbol{a})$. Other work bridging RL and causality includes causal approaches for transfer learning in RL models [162] and data-fusion for reinforcement learners [43].

At the end, we summarize advantages and disadvantages of causal machine learning. The advantages of machine learning with causality include: (1) invariant prediction under environment changes [83, 110, 117], (2) model generality and interpretability [14, 139], and (3) performance improvement with theoretical guarantee [87, 160]. On the other hand, causal machine learning mainly faces the challenges of insufficient amount of data. Causal machine learning algorithms may require ground truth of counterfactuals [77, 138] or interventional data [87, 160] for training or evaluation, which can be difficult to collect.

6 CONCLUSIONS AND SOME OPEN PROBLEMS

Recent attempts have been made to solve the problem of learning causality with more data and less knowledge than traditional studies. Although existing efforts may not directly address learning causality with big data, they build the foundation of data-driven studies for learning causal effects and relations. Another highlight of this work is the connections between 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. 11 shows a summary of the contents covered in this survey. SCMs and the potential outcome framework 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 relations (causal discovery) with data. The methods for learning causal effects with three types of data are presented: (1) i.i.d. data and (2) non-i.i.d. data where we assume that the unconfoundedness is satisfied, and (3) 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. We discuss how we can connect learning causality to machine learning methods that solve the three families of problems: supervised and semi-supervised learning, domain adaptation and reinforcement learning. We describe the connections by answering: (1) how learning causality yields better prediction in the machine learning problem and (2) how machine learning techniques can be applied for learning causality?

Existing research demonstrates how to learn causality with data paves the way for more work toward addressing the challenges of big data. From the data perspective, we present some open problems to achieve the great potential of learning causality with data:

- Many problems in learning causality from observational data remain open. Potential problems include handling (1) anomalies [1, 37], (2) treatment entanglement [151], (3) complex treatments [95] (e.g., image and text as treatment), (4) temporal observations [102, 132].
- Using causal knowledge to improve machine learning algorithms remains an open area. Potential research directions include: (1) causal interpretation of black box deep learning algorithms [105], (2) learning causality-aware models for robustness [9] and fairness [85].

APPENDIX

To facilitate development, evaluation and comparison of methods for learning causality, we create the open source data index [24] (https://github.com/rguo12/awesome-causality-data) and algorithm index (https://github.com/rguo12/awesome-causality-algorithms). They are categorized by the problem and the type of data.

ACKNOWLEDGEMENT

REFERENCES

[1] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: a survey. *DMKD* 29, 3 (2015), 626–688.

[2] Dionissi Aliprantis. 2015. A distinction between causal effects in Structural and Rubin Causal Models. (2015).

[3] 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.

[4] Bryan Andrews, Joseph Ramsey, and Gregory F Cooper. 2019. Learning High-dimensional Directed Acyclic Graphs with Mixed Data-types. *PMLR* 104 (2019), 4.

[5] 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.

[6] 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.

[7] 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.

[8] Sinan Aral and Christos Nicolaides. 2017. Exercise contagion in a global social network. *Nature communications* 8 (2017), 14753.

[9] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[10] Susan Athey and Guido W Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *stat* 1050, 5 (2015).

[11] Susan Athey, Guido W Imbens, and Stefan Wager. 2018. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Series B (Stat. Methodol.)* 80, 4 (2018), 597–623.

[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] Davide Bacciu, Terence A Etchells, Paulo JG Lisboa, and Joe Whittaker. 2013. Efficient identification of independence networks using mutual information. *Comput. Stat.* 28, 2 (2013), 621–646.

[14] Mohammad Taha Bahadori, Krzysztof Chalupka, Edward Choi, Robert Chen, Walter F Stewart, and Jimeng Sun. 2017. Causal regularization. *arXiv preprint arXiv:1702.02604* (2017).

[15] Eytan Bakshy, Dean Eckles, and Michael S Bernstein. 2014. Designing and deploying online field experiments. In *WWW*. ACM, 283–292.

[16] Elias Bareinboim, Andrew Forney, and Judea Pearl. 2015. Bandits with unobserved confounders: A causal approach. In *NeurIPS*. 1342–1350.

[17] Elias Bareinboim and Judea Pearl. 2012. Transportability of causal effects: Completeness results. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[18] Elias Bareinboim and Jin Tian. 2015. Recovering Causal Effects from Selection Bias.. In *AAAI*. 3475–3481.

[19] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP*. 120–128.

[20] 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.

[21] Donald T Campbell. 1969. Reforms as experiments. *Am. Psychol.* 24, 4 (1969), 409.

[22] 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.

[23] Matias D Cattaneo, Nicolás Idrobo, and Rocío Titiunik. 2017. A practical introduction to regression discontinuity designs. *Cambridge Elements: Quantitative and Computational Methods for Social Science-Cambridge University Press I* (2017).

[24] Lu Cheng, Raha Moraffah, Ruocheng Guo, K.S. Candan, Adrienne Raglin, and Huan Liu. 2019. A Practical Data Repository for Causal Learning with Big Data. In *Bench*.

[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 and Marloes H Maathuis. 2014. Order-independent constraint-based causal structure learning. *JMLR* 15, 1 (2014), 3741–3782.

[31] 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.

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

[33] Ruifei Cui, Perry Groot, and Tom Heskes. 2016. Copula PC algorithm for causal discovery from mixed data. In *ECMLPKDD*. Springer, 377–392.

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

[35] Martijn de Jongh and Marek J Druzdzel. 2009. A comparison of structural distance measures for causal Bayesian network models. *Recent Advances in Intelligent Information Systems, Challenging Problems of Science, Computer Science series* (2009), 443–456.

[36] 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.

[37] Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. 2019. Deep anomaly detection on attributed networks. In *SDM*. SIAM, 594–602.

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

[39] 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.
[40] Michael Eichler. 2012. Causal inference in time series analysis. *Causality: Statistical perspectives and applications*
      (2012), 327–354.
[41] Doris Entner, Patrik Hoyer, and Peter Spirtes. 2013. Data-driven covariate selection for nonparametric estimation of
      causal effects. In *AISTATS*. 256–264.
[42] Doris Entner and Patrik O Hoyer. 2010. On causal discovery from time series data using FCI. *PGM* (2010), 121–128.
[43] Andrew Forney, Judea Pearl, and Elias Bareinboim. 2017. Counterfactual Data-Fusion for Online Reinforcement
      Learners. In *ICML*. 1156–1164.
[44] Constantine E Frangakis and Donald B Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58, 1 (2002),
      21–29.
[45] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2008. Kernel measures of conditional
      dependence. In *NeurIPS*. 489–496.
[46] 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.
[47] Bin Gao and Yuehua Cui. 2015. Learning directed acyclic graphical structures with genetical genomics data. *Bioinformatics* 31, 24 (2015), 3953–3960.
[48] Andrew Gelman. 2011. Causality and statistical learning. *AJS* 117, 3 (2011), 955–966.
[49] 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.
[50] Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. 2017. Causal discovery from
      temporally aggregated time series. In *UAI*, Vol. 2017.
[51] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
[52] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and
      Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*. 2672–2680.
[53] 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.
[54] Ruocheng Guo, Jundong Li, and Huan Liu. 2018. INITIATOR: Noise-contrastive Estimation for Marked Temporal
      Point Process.. In *IJCAI*. 2191–2197.
[55] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Counterfactual Evaluation of Treatment Assignment Functions with
      Networked Observational Data. In *SDM*. SIAM, 271–279.
[56] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Learning Individual Causal Effects from Networked Observational
      Data. In *WSDM*. ACM, 232–240.
[57] Ruocheng Guo, Yichuan Li, Jundong Li, K. Selǧuk Candan, Adrienne Raglin, and Huan Liu. 2020. IGNITE: A
      Minimax Game Toward Learning Individual Treatment Effects from Networked Observational Data.. In *IJCAI*.
[58] 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.
[59] 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).
[60] Jens Hainmueller. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced
      samples in observational studies. *Political Analysis* 20, 1 (2012), 25–46.
[61] Alain Hauser and Peter Bühlmann. 2015. Jointly interventional and observational data: estimation of interventional
      Markov equivalence classes of directed acyclic graphs. *J. R. Stat. Soc. Series B (Stat. Methodol.)* 77, 1 (2015), 291–318.
[62] 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.
[63] David Heckerman, Christopher Meek, and Gregory Cooper. 2006. A Bayesian approach to causal discovery. In
      Innovations in Machine Learning. Springer, 1–28.
[64] 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.
[65] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph Stat.* 20, 1 (2011),
      217–240.
[66] 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.
[67] Paul W Holland. 1986. Statistics and causal inference. *J. Amer. Statist. Assoc.* 81, 396 (1986), 945–960.
[68] 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).

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

[70] Aapo Hyvärinen and Erkki Oja. 2000. Independent component analysis: algorithms and applications. *Neural networks* 13, 4-5 (2000), 411–430.

[71] 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.

[72] Kosuke Imai and Marc Ratkovic. 2014. Covariate balancing propensity score. *J. R. Stat. Soc. Series B (Stat. Methodol.)* 76, 1 (2014), 243–263.

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

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

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

[76] 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.

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

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

[79] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. 2018. Removing hidden confounding by experimental grounding. In *NeurIPS*. 10888–10897.

[80] Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. 2016. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Amer. Statist. Assoc.* 111, 513 (2016), 132–144.

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

[82] Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. 2017. Cost-optimal learning of causal graphs. In *ICML*. JMLR. org, 1875–1884.

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

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

[85] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *NeurIPS*. 4066–4076.

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

[87] Finnian Lattimore, Tor Lattimore, and Mark D Reid. 2016. Causal bandits: Learning good interventions via causal inference. In *NeurIPS*. 1181–1189.

[88] 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).

[89] Thuc Duy Le, Lin Liu, Anna Tsykin, Gregory J Goodall, Bing Liu, Bing-Yu Sun, and Jiuyong Li. 2013. Inferring microRNA–mRNA causal regulatory relationships from expression data. *Bioinformatics* 29, 6 (2013), 765–771.

[90] 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 *NeurIPS*. 396–404.

[91] Jason D Lee and Trevor J Hastie. 2015. Learning the structure of mixed graphical models. *J. Comput. Graph. Stat.* 24, 1 (2015), 230–253.

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

[93] Jundong Li, Ruocheng Guo, Chenghao Liu, and Huan Liu. 2019. Adaptive unsupervised feature selection on attributed networks. In *SIGKDD*. ACM, 92–100.

[94] 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.

[95] Yichuan Li, Ruocheng Guo, Weiying Wang, and Huan Liu. 2019. Causal Learning in Question Quality Improvement. In *BenchCouncil*.

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

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

[98] 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.

[99] HernÃąn MA and Robins JM. forthcoming. *Causal Inference.* CRC Boca Raton, FL:.

[100] Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass* 13, 1 (2018), e12470.

[101] Subramani Mani and Gregory F Cooper. 2000. Causal discovery from medical textual data.. In *Proceedings of the AMIA Symposium.* 542.

[102] Ericsson Marin, Ruocheng Guo, and Paulo Shakarian. 2017. Temporal analysis of influence to predict usersâĂŹ adoption in online social networks. In *SBP.* Springer, 254–261.

[103] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH.*

[104] 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.

[105] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Ragliny, and Huan Liu. 2020. Causal Interpretability for Machine Learning–Problems, Methods and Evaluation. *arXiv preprint arXiv:2003.03934* (2020).

[106] Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference.* Cambridge University Press.

[107] Preetam Nandy, Alain Hauser, Marloes H Maathuis, et al. 2018. High-dimensional consistency in score-based and hybrid structure learning. *Ann. Stat.* 46, 6A (2018), 3151–3183.

[108] Jersey Neyman. 1923. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10 (1923), 1–51.

[109] 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.

[110] Michael J Paul. 2017. Feature Selection as Causal Inference: Experiments with Text Classification. In *CoNLL.* 163–172.

[111] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

[112] Judea Pearl. 2009. Causal inference in statistics: An overview *. *Stat. Surv.* 3 (2009), 96–146.

[113] Judea Pearl. 2009. *Causality.* Cambridge university press.

[114] Judea Pearl. 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016* (2018).

[115] Judea Pearl and Elias Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence.*

[116] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP.* 1532–1543.

[117] 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.

[118] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2013. Causal inference on time series using restricted structural equation models. In *NeurIPS.* 154–162.

[119] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms.* MIT press.

[120] 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).

[121] Vineet K Raghu, Allen Poon, and Panayiotis V Benos. 2018. Evaluation of causal structure learning methods on mixed data types. *PMLR* 92 (2018), 48.

[122] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. Linked Causal Variational Autoencoder for Inferring Paired Spillover Effects. In *CIKM.* ACM, 1679–1682.

[123] 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.

[124] Joseph D Ramsey. 2014. A scalable conditional independence test for nonlinear, non-Gaussian data. *arXiv preprint arXiv:1401.5031* (2014).

[125] Thomas Richardson. 1996. A discovery algorithm for directed cyclic graphs. In *UAI.* Morgan Kaufmann Publishers Inc., 454–461.

[126] Thomas S Richardson and James M Robins. [n. d.]. Single World Intervention Graphs: A Primer. ([n. d.]).

[127] James M Robins, Miguel Angel Hernan, and Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology.

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

[129] 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.

[130] 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.

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

[132] Soumajyoti Sarkar, Ruocheng Guo, and Paulo Shakarian. 2019. Using network motifs to characterize temporal network evolution leading to diffusion inhibition. *SNAM* 9, 1 (2019), 14.

[133] 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).

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

[135] Andrew J Sedgewick, Ivy Shi, Rory M Donovan, and Panayiotis V Benos. 2016. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinform.* 17, 5 (2016), S175.

[136] 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.

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

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

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

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

[141] 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.

[142] 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.

[143] Ricardo Silva. 2016. Observational-interventional priors for dose-response learning. In *NeurIPS*. 1561–1569.

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

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

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

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

[148] Matt Taddy, Hedibert Freitas Lopes, and Matt Gardner. 2016. Scalable semiparametric inference for the means of heavy-tailed distributions. *arXiv preprint arXiv:1602.08066* (2016).

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

[150] Michalis Titsias and Neil D Lawrence. 2010. Bayesian Gaussian process latent variable model. In *AISTATS*. 844–851.

[151] 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).

[152] Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. 2003. Algorithms for large scale Markov blanket discovery.. In *FLAIRS conference*, Vol. 2. 376–380.

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

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

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

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

[157] Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. 2017. Permutation-based causal inference algorithms with interventions. In *NeurIPS*. 5822–5831.

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

[159] Man Leung Wong, Shing Yan Lee, and Kwong Sak Leung. 2002. A hybrid approach to discover Bayesian networks from databases using evolutionary programming. In *ICDM*. IEEE, 498–505.

[160] Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. 2018. Causal bandits with propagating inference. *arXiv preprint arXiv:1806.02252* (2018).

[161] Xuan Yin and Liangjie Hong. 2019. The Identification and Estimation of Direct and Indirect Effects in A/B Tests through Causal Mediation Analysis. In *SIGKDD*. ACM.

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

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

[164] 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).

[165] Dawei Zhou, Jingrui He, Hongxia Yang, and Wei Fan. 2018. Sparc: Self-paced network representation for few-shot rare category characterization. In *SIGKDD*. ACM, 2807–2816.