Causal analysis framework for recommendation

Peng Wu

Prof. Xiao-Hua Zhou's Team

Beijing International Center for Mathematical Research,
Peking University

January 23, 2022

招聘| 北京大学周晓华课题组博士后招聘启事

北京大学公共卫生学院生物统计系 2022-01-22 10:36

北京大学周晓华课题组博士后招聘启事

信息项目	信息内容
标　题	北京大学周晓华课题组北京国际数学研究中心和生物统计系联合博士后招聘启事
招聘岗位	博士后2人

Link: https://mp.weixin.qq.com/s/IMZNTsB_1EkY1zZ0fLb0Tw

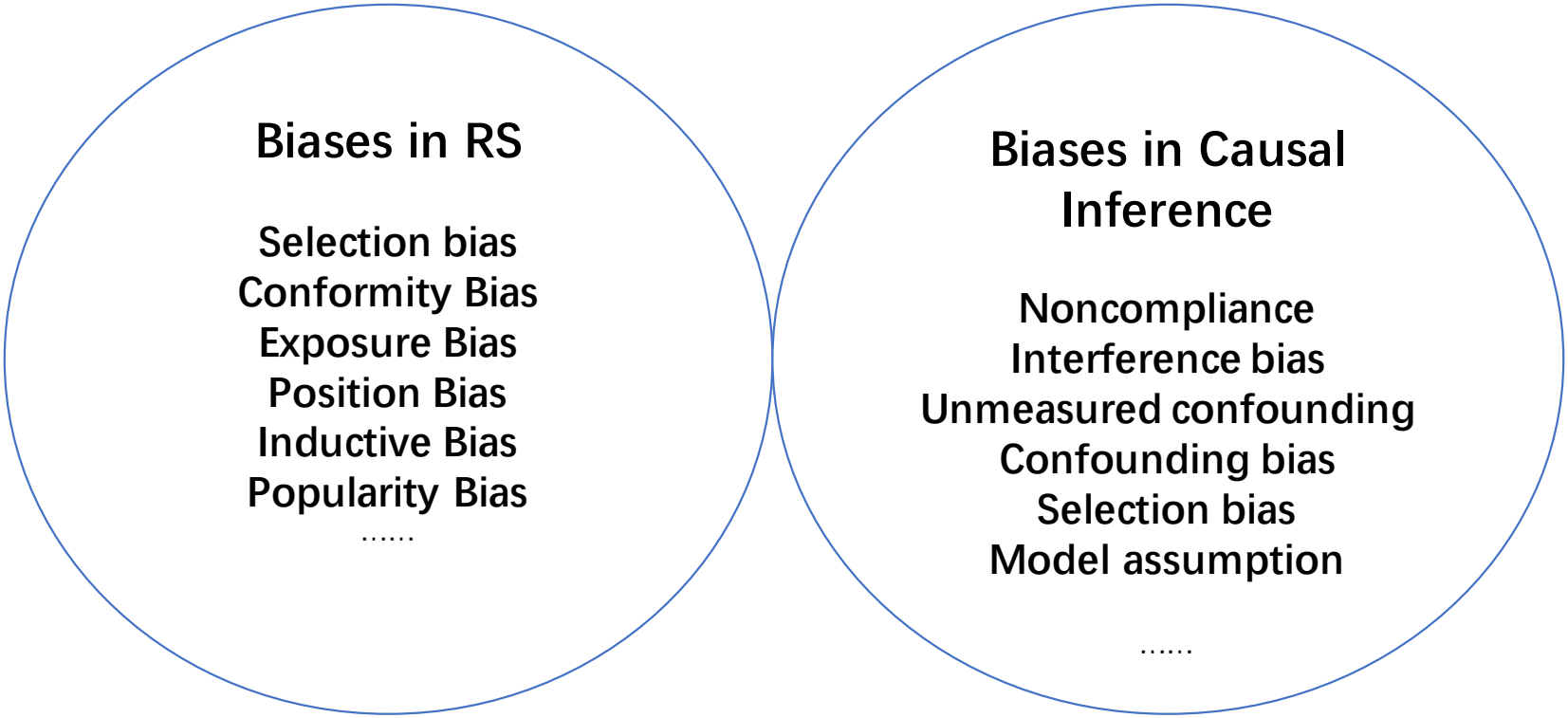
Research: Causal inference, Causal recommendation, Causal AI, Intelligent causal medicine.

合作导师周晓华教授，北京大学讲席教授，国家海外高层次人才引进计划创新人才长期项目入选者，北京大学国家药品医疗器械监管科学研究院副院长，北京大学公共卫生学院生物统计系主任，国际知名生物统计学家，北京大学北京国际数学研究中心生物统计及生物信息实验室主任，北京大数据研究院中医大数据中心主任。先后担任美国印第安纳大学生物统计助理教授，副教授，美国华盛顿大学生物统计系教授，美国国家阿尔茨海默病协调中心副主任，国际生物统计学会中国分会(IBS-China)理事长，中国现场统计研究会生物医学统计学会会长，美国科学促进会会士，美国统计学会会士，国际统计研究院会士。美国联邦政府食品和药物管理局（FDA）医疗器械和放射健康顾问委员会委员。曾荣获美国联邦政府退伍军人事务部授予的研究生涯科学家奖、中国国家自然科学基金委海外杰青，中国教育部高层次文教专家、中国教育部海外名师等荣誉称号，获美国统计学会贝叶斯分析科学分会及国际贝叶斯统计科学学会Mitchell奖，SCIENCE CHINA-Mathematics评选的年度优秀论文奖。在国际顶尖的统计和生物统计期刊J. R. Statist. Soc. B, JASA、Biometrika、Ann. Statist, Biometrics, Stat. Med，Science Advance等发表SCI学术论文270余篇，其中170余篇为第一或通信作者。先后任生物统计学顶尖期刊，Statistics in Medicine，Journal of American Statistical Association-Theory and Method副编辑，也是国际生物统计学会中国分会会刊，Biostatistics & Epidemiology主编。

1.1 Motivation

- The introduction of causal techniques into recommender systems (RS) has brought great development to this field and has gradually become a trend.
- On one hand, the existing causal methods in RS lack a clear causal and mathematical formalization on the scientific questions of interest. Many confusions need to be clarified: what exactly is being estimated, for what purpose, in which scenario, by which technique, and under what plausible assumptions.
- On the other hand, technically speaking, the existence of various biases is the main obstacle to drawing causal conclusions from observed data. Yet, formal definitions of the biases in RS are still not clear, which leads to difficulty in discussing theoretical properties and limitations of various debiasing approaches.
- Both of the limitations greatly hinder the development of RS.

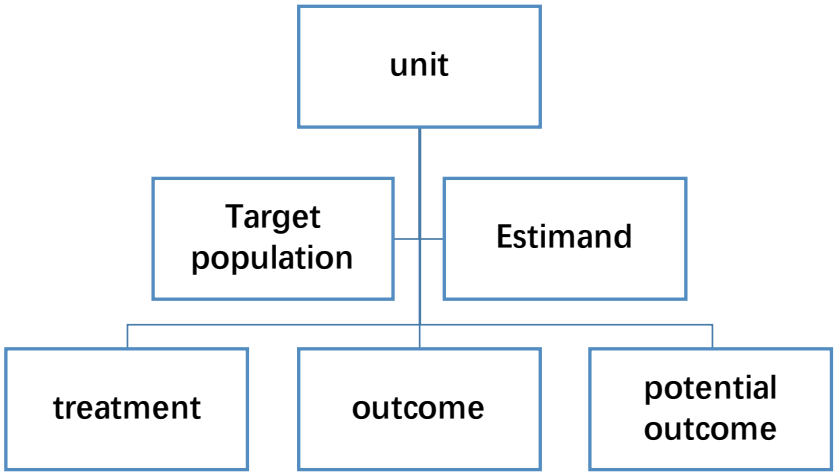
1.2 Goal



- Formalize different tasks/scenarios in RS using causal framework.
- Provide formal definitions of various biases in RS.

1.3 Causal framework

- Potential outcome framework (PO).



Identification: a set of assumptions. mathematical tools.

- Structural causal model (SCM).
  - structural equation model;
  - causal graph.

Identification: causal graph; other assumptions. do-calculus.

1.4 Causal analysis framework

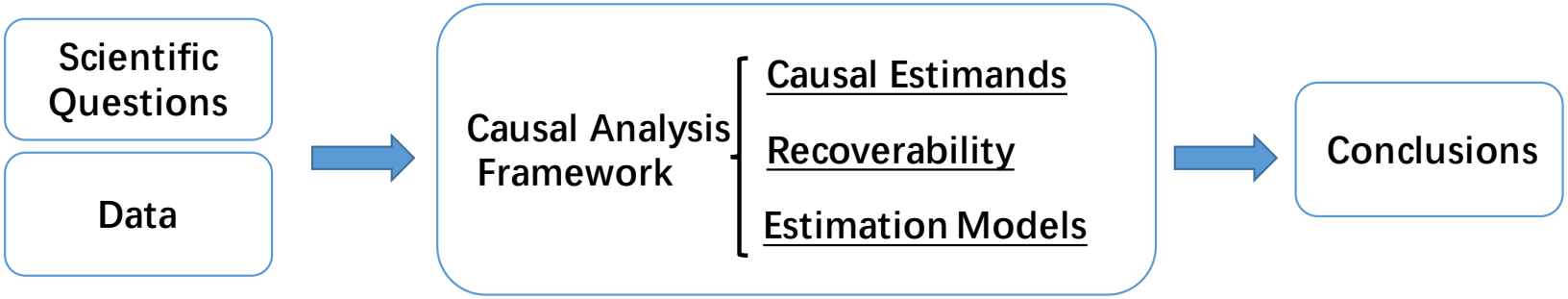
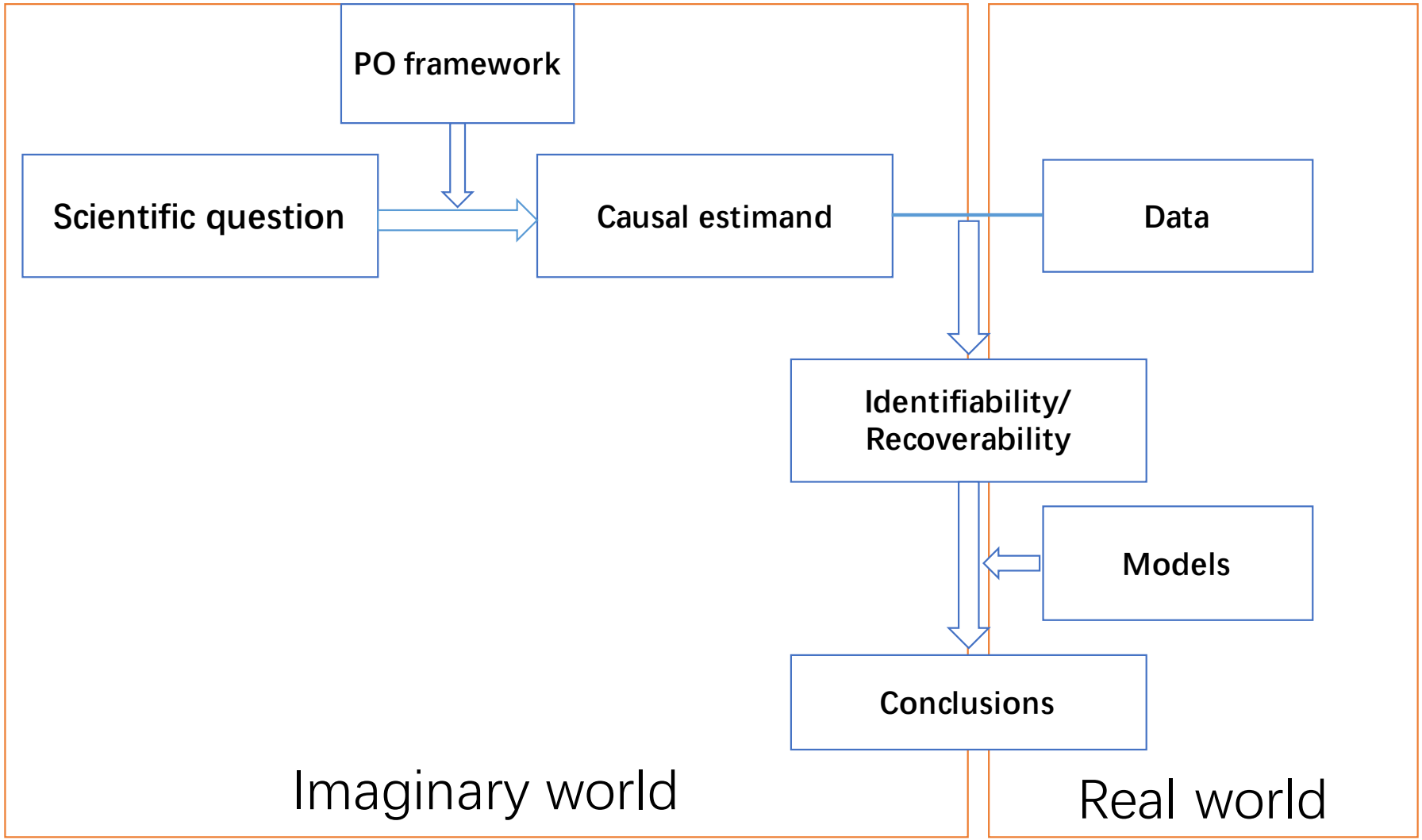


Figure 1: A unified workflow.

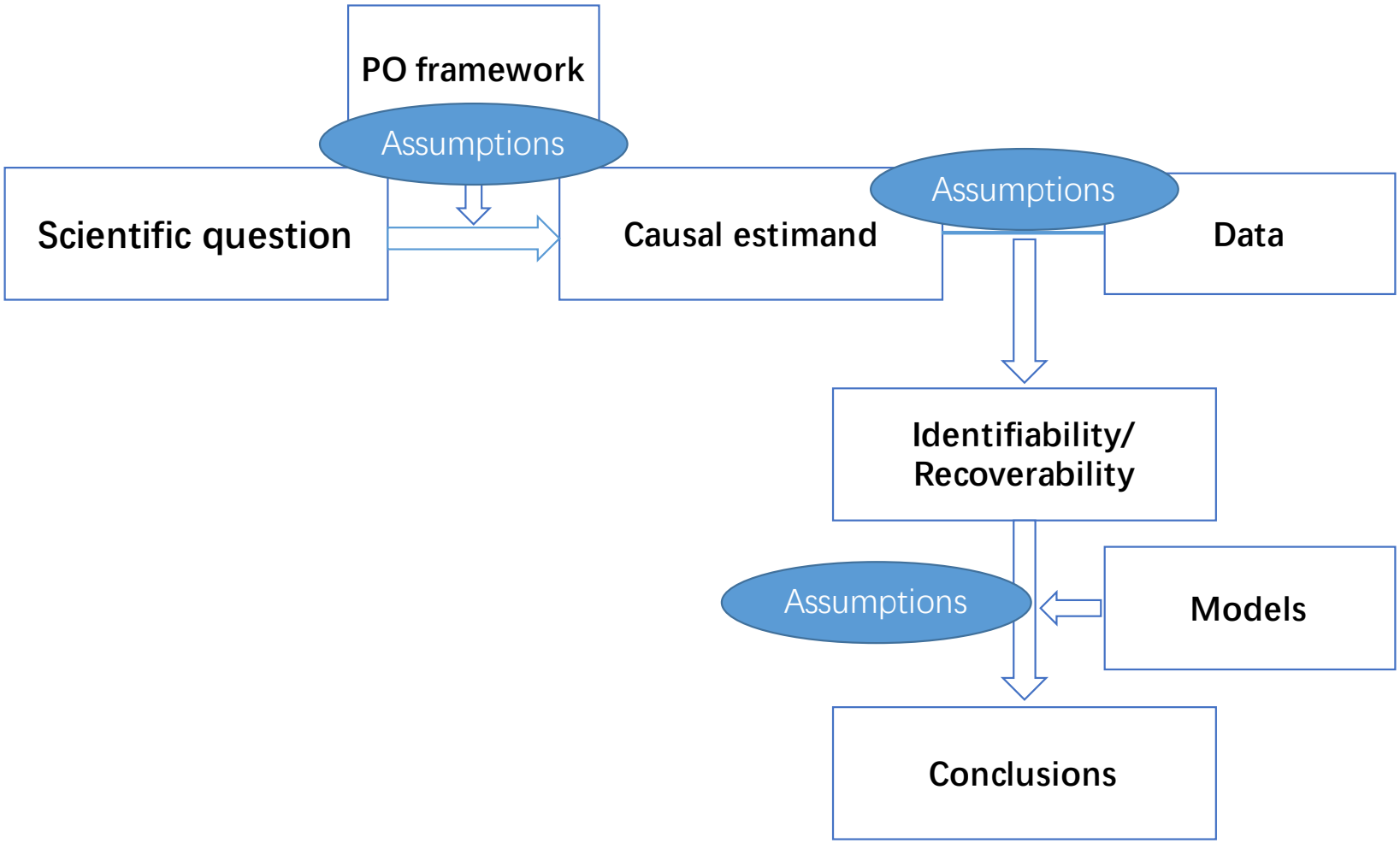
A unified workflow of investigating causal problems consists of three steps:

1. Define a causal estimand to answer the scientific question.
2. Discuss the recoverability of the estimand given the data.
3. Build models to obtain the consistent estimator of the estimand.

1.4 Causal analysis framework



1.4 Causal analysis framework



1.5 Overview of Main conclusions

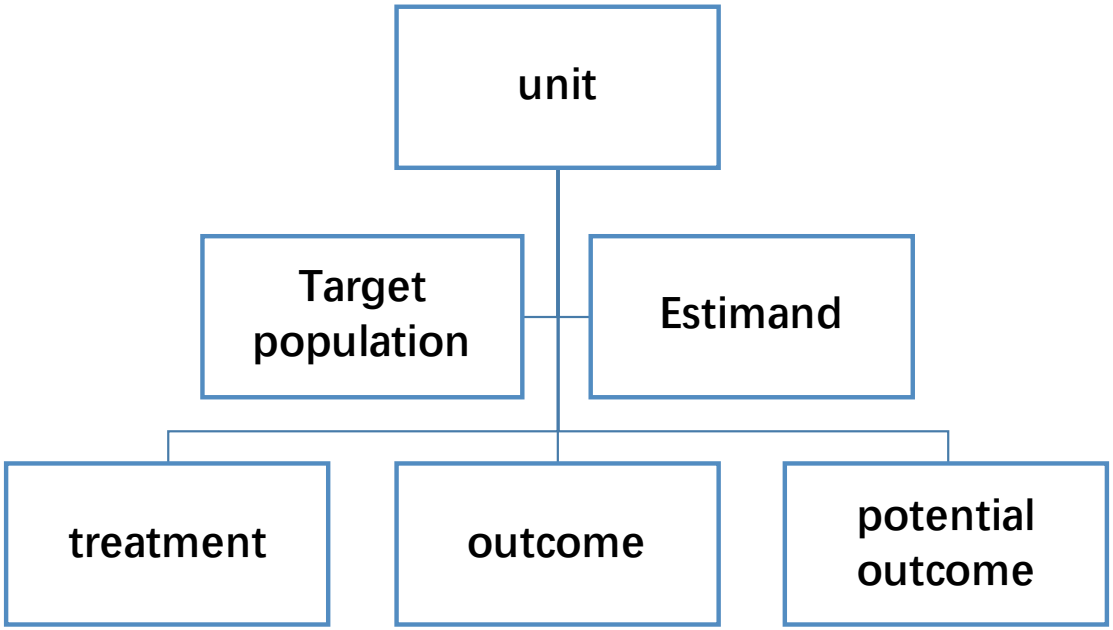
Table 1: New perspective of biases in RS.

	Assumptions	Biases in causal inference	Biases in RS
Define causal estimands	SUTVA(a)	undefined	position bias
	SUTVA(b)	interference bias	conformity bias
Recoverability	consistency	noncompliance	undefined
	positivity	undefined	exposure bias
	exchangeability	confounding bias	popularity bias
	conditional exchangeability	hidden confounding bias	undefined
	random sampling	selection bias	user selection bias, exposure bias
Model	model specification	model mis-specification	inductive bias

Significance:

- It provides an opportunity to apply the existing causal inference methods to RS. For example, the non-compliance problem and interference bias have been intensively studied in causal inference literature.
- In addition, for the unique characteristics of RS, we expect that a series of new methods will be developed by weakening or substituting the assumptions.

2.1 Key elements in PO framework



Remarks:

1. PO does not involve the data collected and the model adopted;
2. It also does not specify the relationships among treatment, feature, and outcome.

2.2 Unit

The unit is the most fine-grained research subject.

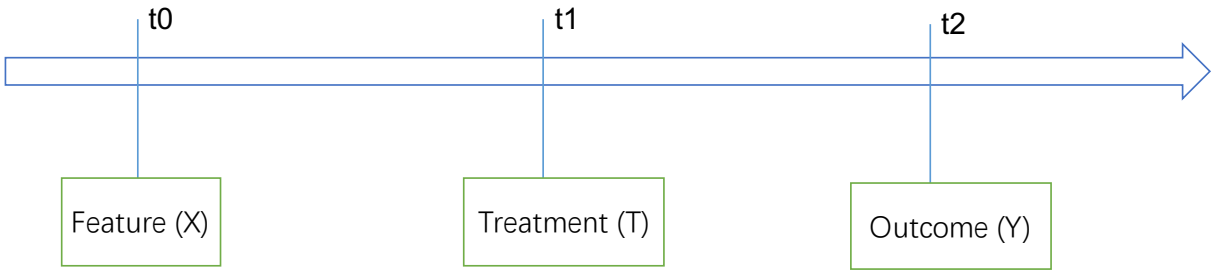
- A clear explanation of it is very important to define the causal estimand, particularly in the field of RS.
- In RS, a unit usually corresponds to a user-item pair; sometimes it is a user.

The variety and vagueness of the unit stem from the fact that RS involves two entangled populations: users and items.

2.3 Treatment, covariate and outcome

For each unit, we have a treatment, an outcome, and possibly a feature vector.

- A treatment $T \in \mathcal{T}$, that is performed at a well-defined time.
- A feature X, measured at a well-defined time before treatment.
- An outcome Y, measured at a well-defined time after treatment.



If different units are measured at different times, they would not be comparable.

2.4 Potential outcome

However, T, X and Y are still not enough to define a causal estimand.

Definition (Potential outcome)

A potential (or counterfactual) outcome $Y(t)$ for $t \in \mathcal{T}$, which is the outcome that would be observed if T had been set to t.

The most fine-grained causal effect, i.e., individualized causal effect, is defined at the unit level. For example, consider the case of binary treatment, namely $\mathcal{T} = \{0, 1\}$. For m-th unit,

$$\text{ITE} = Y_m(1) - Y_m(0).$$

In practice, the individualized causal effect often refers to the feature-specific causal effect, defined by

$$\mathbb{E}(Y(1) - Y(0) \mid X = x).$$

2.5 Target population

To clarify the population of interest, we need to specify a target population.

Definition (Target population)

Target population is the population that we want to make an inference on.

We denote $\mathbb{P}$ and $\mathbb{E}$ as the distribution and expectation on the target population. In RS, the target population is usually the population consisting of all user-item pairs, or all users, or all items, which depends on the specific scientific question.

2.6 Causal estimand

Definition (Causal estimand)

Causal estimand is a functional of the joint distribution of treatment, feature and potential outcomes on the target population, providing a recipe for answering the scientific question of interest from any hypothetical data whenever it is available.

Remarks:

- The definition of the causal estimand does not involve the data collected and the model adopted.
- It also doest not involve the relationship between X, T and Y. In other word, when defining causal estimand, it needn't distinguish confounder, collider, instrument variable .....

Have we made any assumptions so far?

2.7 The SUTVA assumption

Usually, the stable unit treatment value assumption (SUTVA) is necessary to ensure the well-definedness of potential outcome $Y(t)$.

Definition (SUTVA, Assumption 1)

(a) no-multiple-versions-of-treatment, only a single version of the treatment and a single version of the control;

(b) no-interference, the potential outcomes of units are not affected by the treatment status of the other individuals in the population.

2.8 No-multiple-versions-of-treatment and position bias

Definition (Position Bias (in implicit feedback data))

Position bias happens as users tend to interact with items in higher position of the recommendation list regardless of the items' actual relevance so that the interacted items might not be highly relevant.

Insight 1: The position bias can be seen as a violation of no multiple versions of treatment assumption.

Example: In the task of click-through rate prediction, suppose a unit is a user-item pair. Define $Y_{u,i}(1)$ as the click behavior if the item i is exposed to the user u. Then $Y_{u,i}(1)$ will rely on the position of exposure and multiple versions of treatment occur.

2.8 No-multiple-versions-of-treatment and position bias

- One method is to redefine each version of treatment as a different treatment.
- Redefining each version of treatment as a different treatment may not always be possible or desirable.

Formal definition: Define $Y_m(t, k^t)$ be the potential outcome if T is set to value t by means k^t, where $k^t \in K^t = \{1, ..., n^t\}$. The the multiple version of treatment assumption is defined as

$$Y_m(t, k) = Y_m(t, k'), \forall m, t, \text{ and } k, k \in K^t. \tag{1}$$

If (1) holds, then $Y_m(t) = Y_m(t, k), \forall k \in K^t$, the potential outcome is well-defined. The no-multiple-versions-of-treatment also referred to as treatment-variation-irrelevance.

2.9 Interference and conformity bias

Definition (Conformity Bias (explicit feedback data))

Conformity bias happens as users tend to rate similarly to the others in a group, even if doing so goes against their own judgment, making the rating values do not always signify user true preference.

Insight 2: The conformity bias can be seen as a violation of "interference" assumption.

2.9 Interference and conformity bias

If there is no interference, we can define the potential outcomes

$$Y_m(1), Y_m(0).$$

Each unit has only two potential outcomes.

In the presence of interference, the potential outcome is defined by

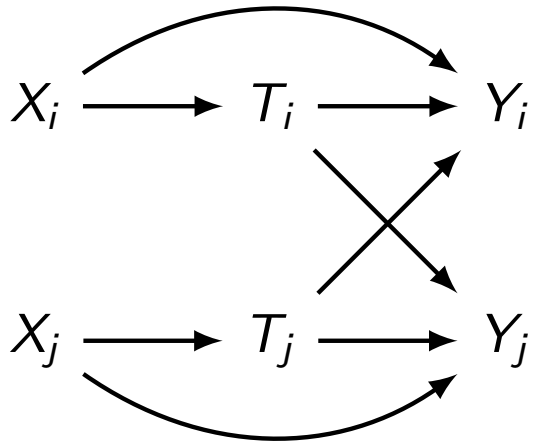
$$Y_m(\boldsymbol{T}),$$

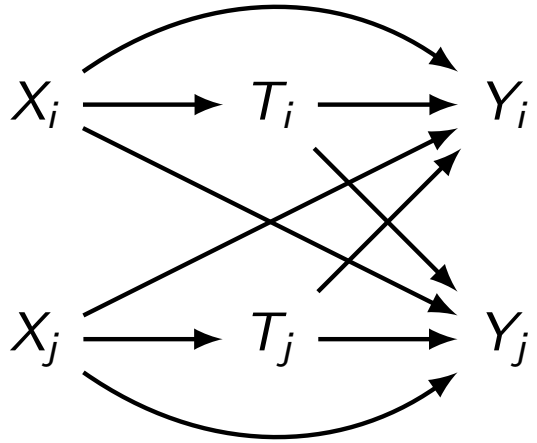
where $\boldsymbol{T} = (T_1, ..., T_N)$. Each unit has 2^N potential outcomes, N is the sample size.

Formal definition of direct interference:

$$Y_m(t_m, \boldsymbol{t}_{-m}) = Y_m(t_m, \boldsymbol{t}'_{-m}), \quad \forall \boldsymbol{t}_{-m}, \boldsymbol{t}'_{-m}$$

2.9 Interference and conformity bias





2.9 Interference and conformity bias

Consider a (backdoor) SCM

$$\begin{cases} X_i = f_X(\epsilon_{X_i}) \\ T_i = f_T(X_i) + \epsilon_{T_i}, \\ Y_i = f_Y(T_i, X_i) + \epsilon_{Y_i}. \end{cases} \tag{2}$$

If there exists interference, let $\boldsymbol{X} = (X_1, ..., X_N)$. The SCM describing the data generation process may be given as follows

$$\begin{cases} X_i = f_X(\epsilon_{X_i}) \\ T_i = f_Z(\boldsymbol{X}) + \epsilon_{T_i}, \\ Y_i = f_Y(\boldsymbol{X}, \boldsymbol{T}) + \epsilon_{Y_i}. \end{cases} \tag{3}$$

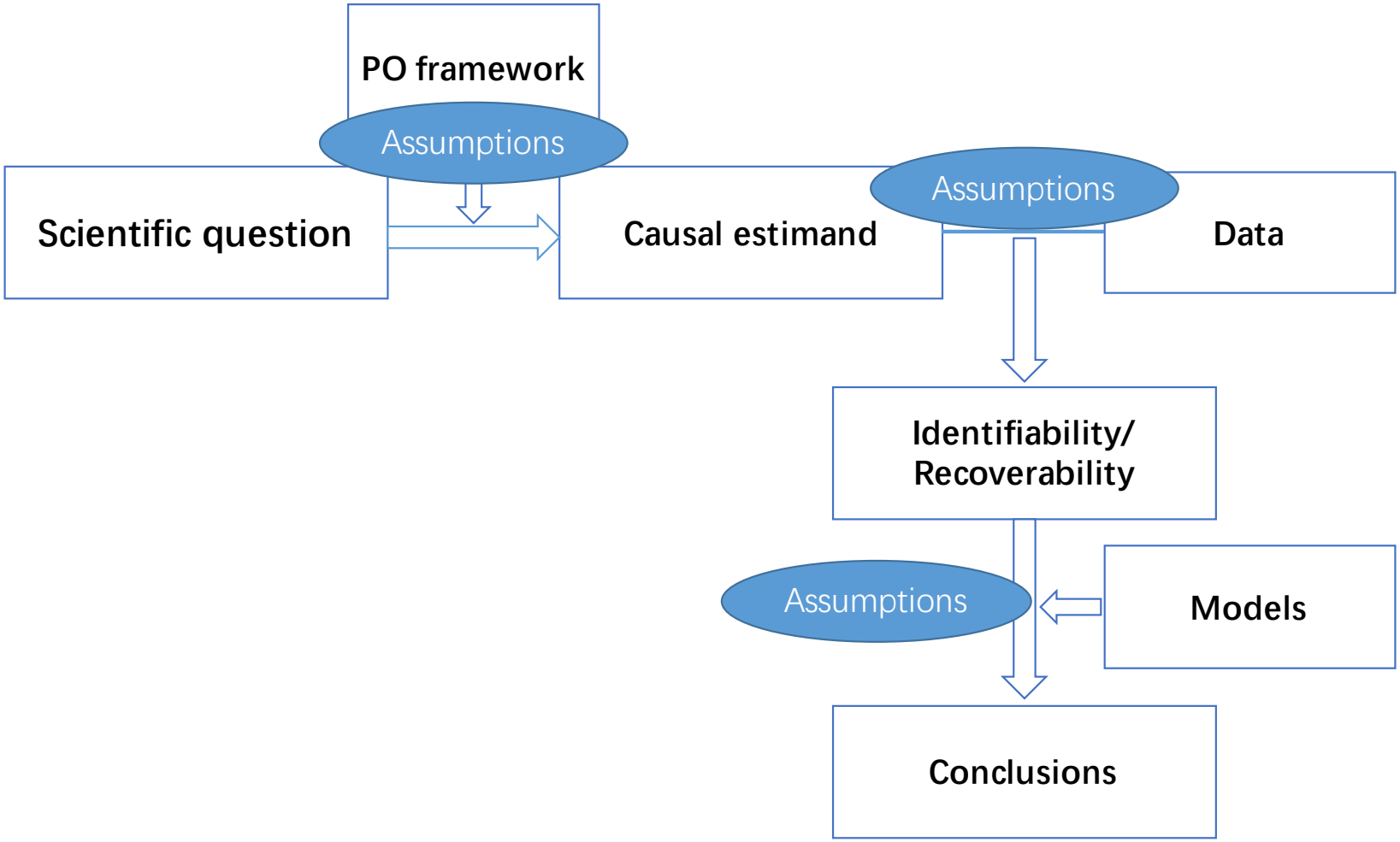


Figure 2: Causal analysis framework

Hereafter, we maintain the SUTVA assumption.

2.10 From scientific question to causal estimand

Significance: Through formalizing the scientific question into a causal estimand, we can answer the following questions: what exactly is being estimated and for what purpose.

The workflow of translating a scientific problem into a meaningful causal estimand is summarized as follows.

1. Define the unit.
2. Define the treatment, feature, outcome and potential outcomes corresponding to the scientific question under study.
3. Define the target population.
4. Define the causal estimand to answer the scientific question.

2.11 Examples

Example. (Binary treatment) An example is advertisement. A unit is a user-item pair, the target population consists of all user-item pairs, and the outcome $Y_{u,i}$ is the indicator of a click event, i.e., $Y_{u,i} = 1$ if user u clicks item i, $Y_{u,i} = 0$ otherwise. The treatment $T_{u,i} = 1$ if item i is exposed to user u, $T_{u,i} = 0$ otherwise, the potential outcomes $Y_{u,i}(1)$ and $Y_{u,i}(0)$ denote the indicator of click event if the item is/isn't exposed to the user u. The estimand of interest is $\mu_1(x)$ denoting the CTR, or $\tau(x)$ denoting the uplift of CTR.

For general treatment, define

$$\mu_t(x) = \mathbb{E}[Y(t) \mid T = t], \ t \in \mathcal{T}, \tag{4}$$

and for binary treatment, define

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x], \tag{5}$$

which are the common estimands in RS.

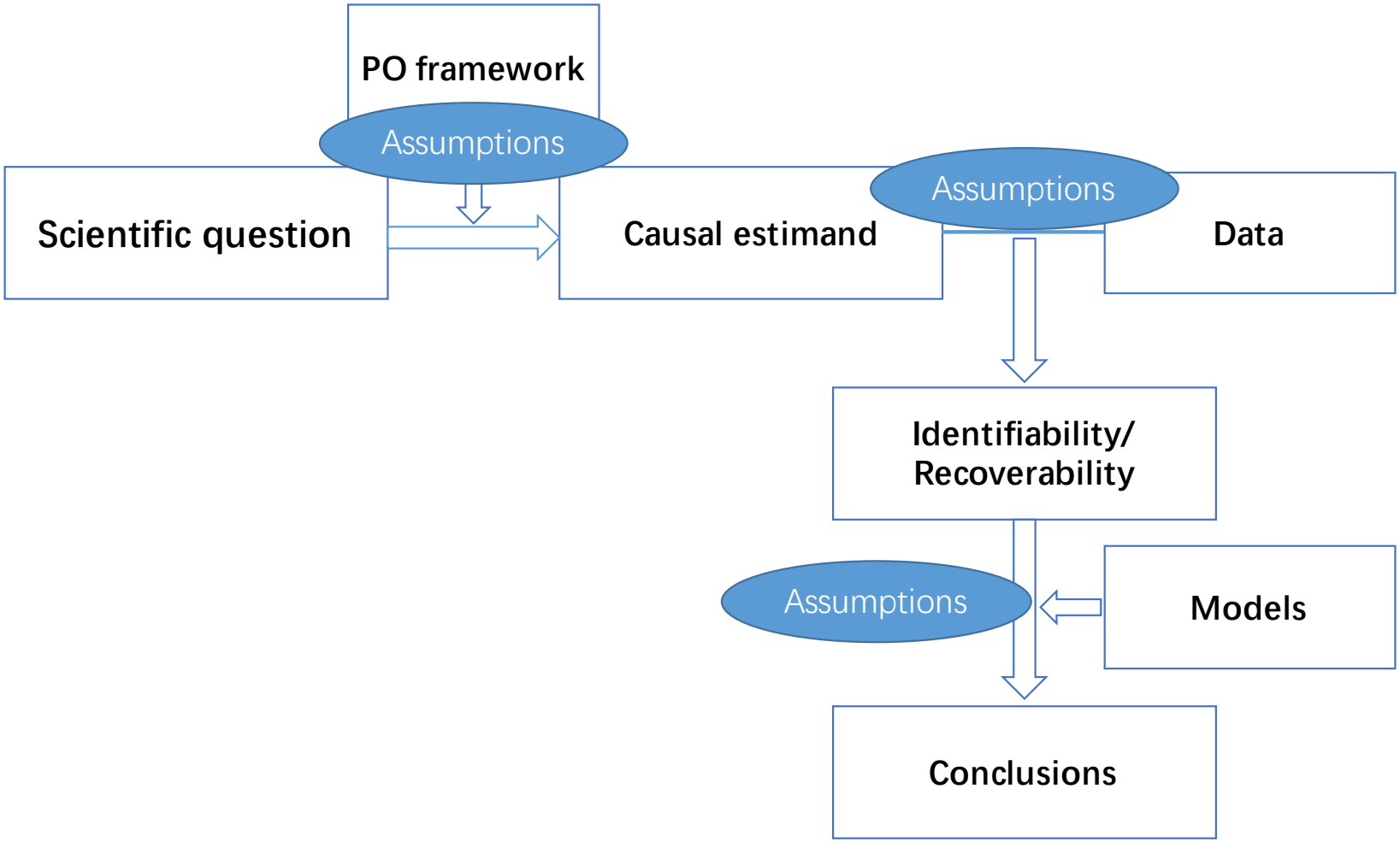


Figure 3: Causal analysis framework

3.1 Definition of recoverability

Definition (Recoverability of target quantity Q)

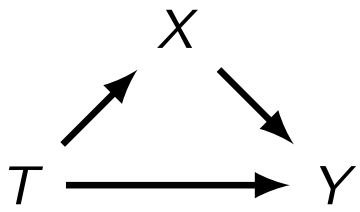
Let $\mathcal{A}$ denote the set of assumptions about the data generation process and let Q be any functional of the underlying distribution $\mathbb{P}(X, T, \{Y(t), t \in \mathcal{T}\})$. Q is recoverable if there exists a procedure that computes a consistent estimator of Q for all strictly positive observed-data distributions.

Significance:

- Explicitly presenting the recoverability assumptions underlying the debiasing approaches.
- Providing a desirable perspective to evaluate the debiasing methods by assessing the assumptions and provides an opportunity to develop new approaches by weakening the assumptions.

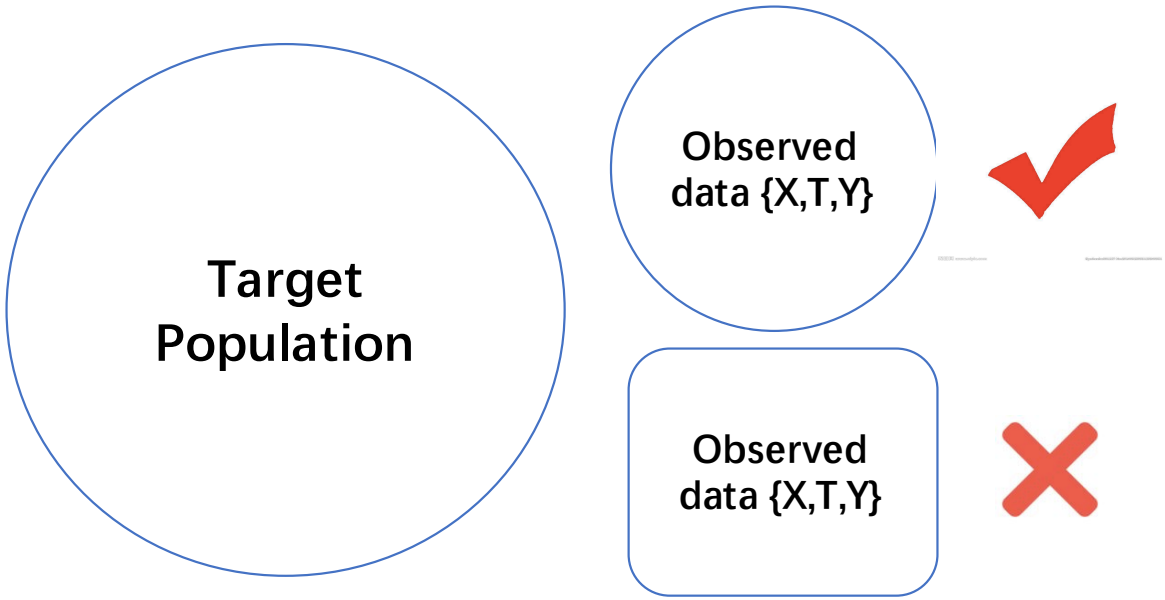
3.2 Common assumptions for recoverability

- *Assumption 1* (SUTVA): (a) no multiple versions of treatment; (b) no-interference.
- *Assumption 2* (Consistency): $Y(t) = \sum_{t^* \in \mathcal{T}} I(t^* = t) Y$ for any $t \in \mathcal{T}$.
- *Assumption 3* (Positivity): $\mathbb{P}(T = t \mid X = x) > 0$ for any t and x.
- *Assumption 4* (Conditional exchangeability): $Y(t) \perp T \mid X$, for any $t \in \mathcal{T}$. A stronger version is exchangeability: $Y(t) \perp T$, for any $t \in \mathcal{T}$.



- *Assumption 5* (Random sampling): $\mathbb{P}(x, t, y) = \mathbb{P}_{\mathcal{O}}(x, t, y)$, where $\mathbb{P}$ represents the target population distribution and $\mathbb{P}_{\mathcal{O}}$ represents the observed sample distribution.

3.3 Random sampling



Remark: Random sampling assumption is defined with observed data. (It does not involve potential outcome)

3.4 Example: backdoor criterion

For example, if $\mathbb{E}[Y(t) \mid X = x]$ is of interest, we can reformulate it as

$$\mathbb{E}[Y(t)|X = x] = \mathbb{E}[Y(t)|X = x, T = t] = \mathbb{E}[Y|X = x, T = t], \tag{6}$$

- the first identity relies on the positivity and conditional exchangeability assumptions
- the second identity requires the consistency assumption.
- By random sampling assumption, $\mathbb{E}[Y \mid X = x, T = t]$ can be estimated consistently from the observed data directly.

3.5 Consistency

The consistency assumption implies that

$$Y_m(t) = Y_m \quad \text{if } T_m = t.$$

It links the potential outcomes in the hypothetical world to the observed outcomes in reality.

real world			interventional world	
T_m	X_m	Y_m	$Y_m(0)$	$Y_m(1)$
0	✓	✓	✓	
0	✓	✓	✓	
0	✓	✓	✓	
1	✓	✓		✓
1	✓	✓		✓
1	✓	✓		✓

4.1 Positivity and exposure bias

- *Assumption 3* (Positivity): $\mathbb{P}(T = t \mid X = x) > 0$ for any t and x.

Definition (Exposure Bias (implicit feedback data))

Exposure bias happens as users can only be exposed to a part of specific items so that unobserved interactions do not always represent negative preference.

Insight 3: Exposure bias can be viewed as a violation of the positivity assumption.

4.2 Exchangeability and confounding bias

- *Assumption 4* (Conditional exchangeability): $Y(t) \perp T \mid X$, for any $t \in \mathcal{T}$.
  A stronger version is exchangeability: $Y(t) \perp T$, for any $t \in \mathcal{T}$.

Definition (Confounding bias)

Confounding bias refers to the association (T and Y) created due to the presence of factors affecting both the treatment and the outcome, i.e., $\exists t \in \mathcal{T}$, $Y(t) \not\perp T$.
Usually it will lead to

$$\mathbb{E}[Y(t)] \neq \mathbb{E}[Y(t)|T = t].$$

4.3 Random sampling and selection bias

- *Assumption 5* (Random sampling): $\mathbb{P}(x, t, y) = \mathbb{P}_{\mathcal{O}}(x, t, y)$, where $\mathbb{P}$ represents the target population distribution and $\mathbb{P}_{\mathcal{O}}$ represents the observed sample distribution.

Definition (Selection bias)

Selection bias means that the sample distribution is different from that of target population, i.e.,

$$\mathbb{P}(x, t, y) \neq \mathbb{P}_{\mathcal{O}}(x, t, y).$$

4.4 Differences between confounding bias and selection bias

1. Confounding bias cannot be eliminated as the sample size increases.
2. Selection bias abounds in RS.
   - model selection bias: the system aims to recommend items that the user may like by filtering out items with low predicted ratings.
   - user selection bias: users tend to rate recommended items that he likes and rarely rates recommended items that he dislikes.
3. Like the case of confounding bias, biased estimates will be produced regardless of the number of samples collected.
4. Conceptually, the bias arising from selection differs fundamentally from the one due to confounding.
   - Selection bias comes from the systematic bias during the collection of units into the sample. A well-designed sampling procedure can reduce selection bias.
   - In contrast, confounding bias stems from the systematic bias inherently determined by the causal mechanism (relations) among features, treatment, and outcome, irrespective of the data collection process.

4.4 Differences between confounding bias and selection bias

- Randomization of treatment assignment can eliminate the effect of (unmeasured) confounding bias, but cannot remove the influence of selection bias
- Selection bias comes from the data collection process and hence always involves missing data. It is noteworthy that missing data is not a causal problem and does not involve potential outcomes, while confounding bias is defined with potential outcomes.
- The missing data problem can be regarded as a selection bias if we restrict the analysis to non-missing units.
- Confounding bias can also be treated as a missing data problem. For example, consider the case of binary treatment, then for each unit, only a potential outcome ($Y(1)$ or $Y(0)$) can be observed and the other can be regarded as missing.

4.4 Differences between confounding bias and selection bias

Table 2: Missing outcome data, selection bias

O_{ui}	X_{ui}	Y_{ui}
1	✓	✓
1	✓	✓
0	✓	
0	✓	

Table 3: Binary treatment, confounding bias

T_{ui}	X_{ui}	Y_{ui}	$Y_{ui}(0)$	$Y_{ui}(1)$
0	✓	✓	✓	
0	✓	✓	✓	
1	✓	✓		✓
1	✓	✓		✓

4.4 Differences between confounding bias and selection bias: an example

Example: A unit is a user, the target population is all users, O_u, T_u and Y_u are indicators of exposure, click, and conversion of user u on the advertising. We treat T_u as treatment and define potential outcome $Y_u(1)$.

Table 4: Binary treatment, confounding bias

O_u	T_u	X_u	Y_u	$Y_u(0)$	$Y_u(1)$
1	0	✓	✓	✓	
1	0	✓	✓	✓	
1	1	✓	✓		✓
1	1	✓	✓		✓
0	0	✓			
0	0	✓			
0	1	✓			
0	1	✓			

- Both selection bias and confounding exist.
- If $Y_u(1)$ is of interest, can we regard it as a missing data problem?

4.5 Main conclusions

Definition (User selection Bias in RS)

Selection Bias happens as users are free to choose which items to rate, so that the observed ratings are not a representative sample of all ratings.

Definition (Inductive Bias in RS)

Inductive bias denotes the assumptions made by the model to better learn the target function and to generalize beyond training data.

4.5 Main conclusions

Table 5: New perspective of biases in RS.

	Assumptions	Biases in causal inference	Biases in RS
Define causal estimands	SUTVA(a)	undefined	position bias
	SUTVA(b)	interference bias	conformity bias
Recoverability	consistency	noncompliance	undefined
	positivity	undefined	exposure bias
	exchangeability	confounding bias	popularity bias
	conditional exchangeability	hidden confounding bias	undefined
	random sampling	selection bias	user selection bias, exposure bias
Model	model specification	model mis-specification	inductive bias

Significance:

- It provides an opportunity to apply the existing causal inference methods to RS. For example, the non-compliance problem and interference bias have been intensively studied in causal inference literature.
- In addition, for the unique characteristics of RS, we expect that a series of new methods will be developed by weakening or substituting the assumptions.

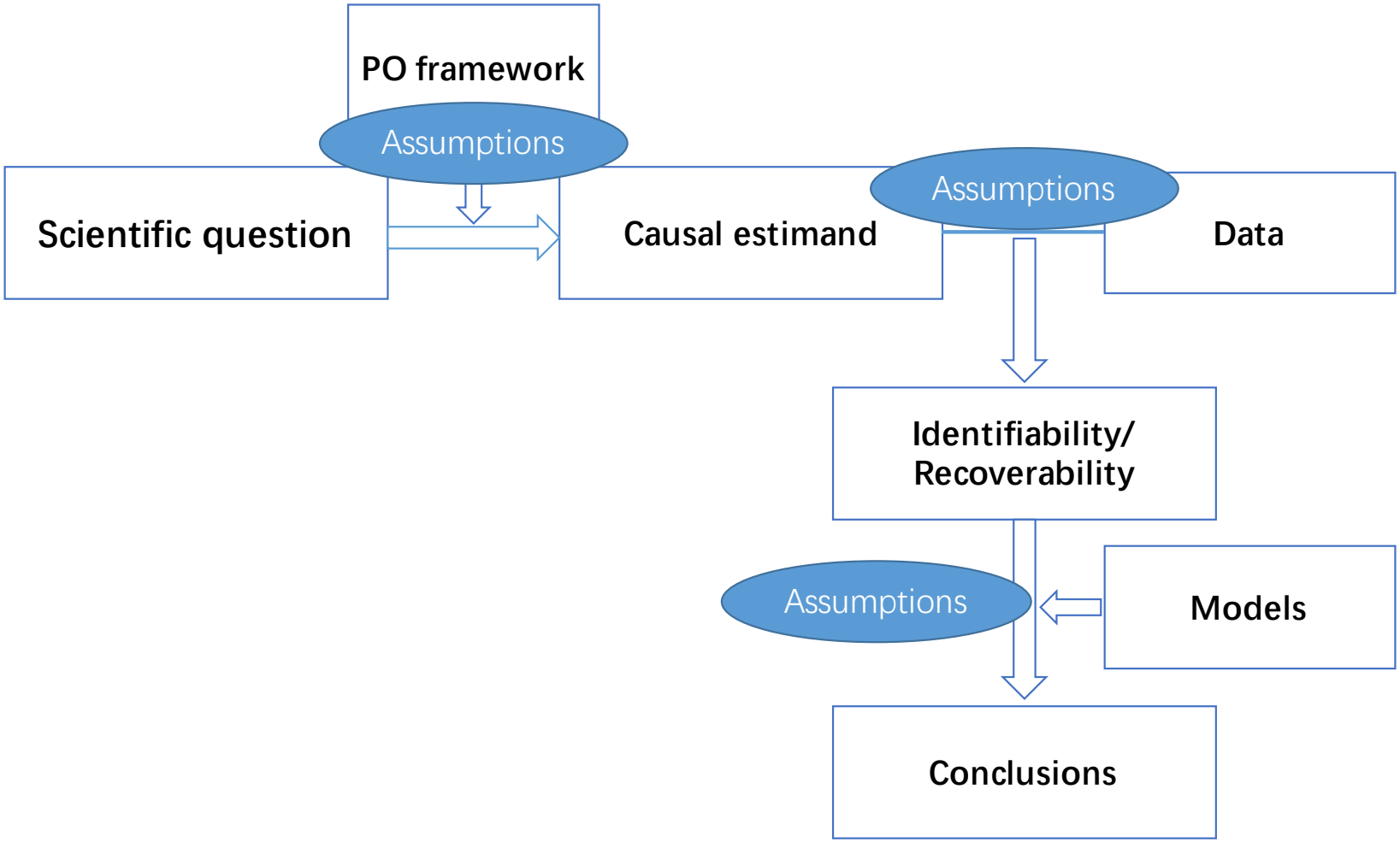


Figure 4: Causal analysis framework

Notations

For general treatment, define

$$\mu_t(x) = \mathbb{E}[Y(t) \mid X = x], \ t \in \mathcal{T},$$

and for binary treatment, define

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x],$$

which are the common estimands in RS.

5.1 Overview of typical scenarios

Table 6: Correspondence between scenarios and biases

Scenarios	Estimands	Biases
Missing outcome data	$\mu_1(x)$	selection bias
Single treatment variable	$\mu_1(x)$, $\tau(x)$	confounding bias
Compliance	$\tilde{\tau}(x)$	non-compliance and confounding bias
Policy Learning	$\pi(t\|x)$	confounding bias
Data Fusion	$\mu_1(x)$, $\tau(x)$	hidden confounding bias

5.2 Scenario 1: Missing outcome data

Essentially, missing data is not a causal problem and it doesn't involve potential outcomes.

However, missing data problem is closely related to causal inference. Specifically, if we regard the observing indicator O as the treatment, and define $Y(1)$ as the outcome if all the units could be observed. Here we use $Y(1)$ instead of Y is to underline that the outcome is part of observable. Then the most common estimand is $\mu_1(x)$ and the main challenge is selection bias.

Table 7: Data structure of scenarios 1.

O_{ui}	X_{ui}	Y_{ui}
1	✓	✓
1	✓	✓
0	✓	
0	✓	

Example 1: Movie rating websites.

5.3 Scenario 2: potential outcomes with single treatment

Scenario 2 discusses the case of a single treatment variable, which is the most popular situation in RS. The data types of treatment can be binary, categorical, or continuous variables. For illustration, assume that there are K treatment levels, i.e., $\mathcal{T} = \{0, 1, ..., K-1\}$.

Table 8: Data structure of scenario 2.

T_{ui}	X_{ui}	Y_{ui}	$Y_{ui}(0)$	$Y_{ui}(1)$	$\cdots$	$Y_{u,i}(K-1)$
0	✓	✓	✓			
0	✓	✓	✓			
1	✓	✓		✓		
1	✓	✓		✓		
⋮	⋮	⋮			⋮	
$K-1$	✓	✓				✓
$K-1$	✓	✓				✓

Example 2: CTR prediction, CVR prediction, CTCVR prediction, uplift modeling.

5.3 Estimation of $\mu_1(x)$: missing outcome

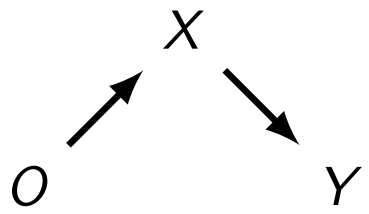
- *Assumption 6* (model specification of the target estimand): $\mu_1(x) = f_\phi(x)$,

Then we specify a loss function $L(y(1), f_\phi(x))$, which is computable only when $O = 1$. Define

$$\pi(x) = \mathbb{P}(O = 1|X = x), g(x) = \mathbb{E}[L(Y(1), f_\phi(X)) \mid X = x].$$

- *Assumption 7* (model specification of nuisance parameters): (a) Propensity score model specification: $\pi(x) = \pi_\beta(x)$; (b) Error imputation model specification: $g(x) = g_\theta(x)$.

5.3 Estimation of $\mu_1(x)$: missing outcome



For brevity, let $L_{u,i}(\phi) = L(Y_{u,i}(1), f_\phi(X_{u,i}))$. The loss function of IPS method is given as

$$\mathcal{L}_{IPS}(\phi;\beta) = \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \frac{O_{u,i}L_{u,i}(\phi)}{\pi_\beta(X_{u,i})}.$$

Then under Assumptions 3-4 and 7(a), we have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{IPS}(\phi;\beta)] &= \mathbb{E}\left[\frac{O_{u,i}L_{u,i}(\phi)}{\pi(X_{u,i})}\right] = \mathbb{E}\left[\mathbb{E}\left\{\frac{O_{u,i}L_{u,i}(\phi)}{\pi_\beta(X_{u,i})} \mid X_{u,i}\right\}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}(O_{u,i}|X_{u,i})\cdot\mathbb{E}(L_{u,i}(\phi)|X_{u,i})}{\pi_\beta(X_{u,i})}\right] \\
&= \mathbb{E}[\mathbb{E}[L_{u,i}(\phi)|X_{u,i})] = \mathbb{E}[L_{u,i}(\phi)].
\end{aligned}$$

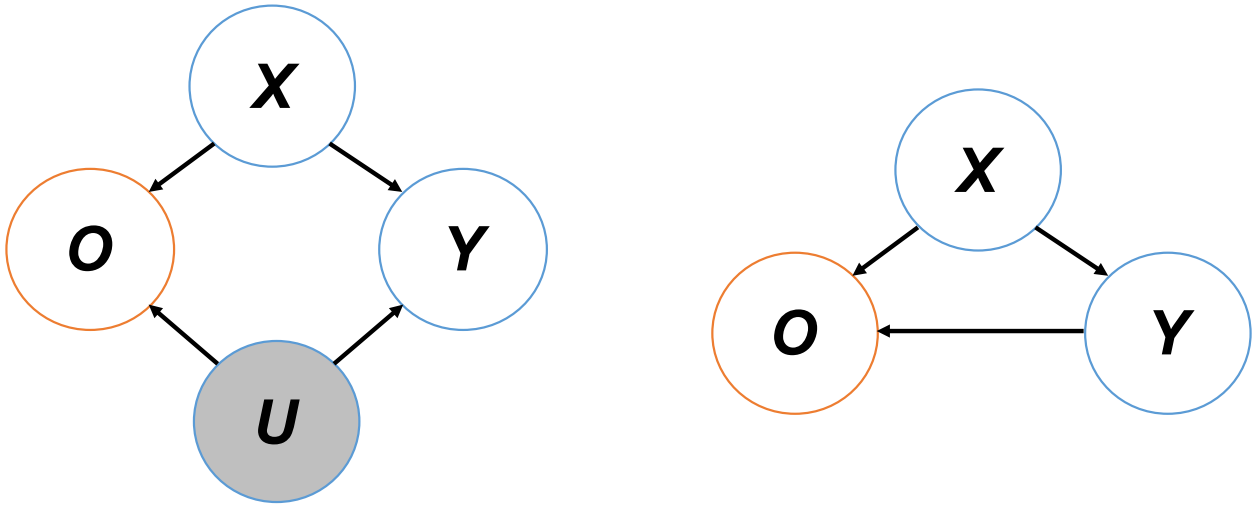


Figure 5: Cases where IPS/DR method is invalid

It violates Assumption 4.

5.3 Estimation of $\mu_1(x)$: missing outcome

For DR approach, the loss function is

$$\mathcal{L}_{DR}(\phi;\beta,\theta) = \frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}}\left[g_\theta(X_{u,i}) + \frac{O_{u,i}\{L_{u,i}(\phi) - g_\theta(X_{u,i})\}}{\pi_\beta(X_{u,i})}\right].$$

Similarly, under Assumptions 3-4, it follows that

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{DR}(\phi;\beta,\theta)] &= \mathbb{E}\left[L_{u,i}(\phi) + \frac{\{O_{u,i} - \pi_\beta(X_{u,i})\}\cdot\{L_{u,i}(\phi) - g_\theta(X_{u,i})\}}{\pi_\beta(X_{u,i})}\right] \\ &= \mathbb{E}\left[L_{u,i}(\phi) + \frac{\{\pi(X_{u,i}) - \pi_\beta(X_{u,i})\}\{g(X_{u,i}) - g_\theta(X_{u,i})\}}{\pi_\beta(X_{u,i})}\right],\end{aligned}$$

from which we can see that $\mathbb{E}[\mathcal{L}_{DR}(\phi;\beta,\theta)] = \mathbb{E}[L_{u,i}(\phi)]$ if either Assumption 7(a) or Assumption 7(b) holds, which is the property of doubly robust.

5.3 Estimation of $\mu_1(x)$: missing outcome

Under Assumptions 2-4, let $\mu(x) = \mathbb{E}[Y|X = x]$, $\pi(x) = \mathbb{P}(T = 1|X = x)$, $\mu_t(x) = \mathbb{E}[Y(t)|X = x] = \mathbb{E}[Y|X = x, T = t]$ for $t = 0, 1$. Many methods, including S-learner, T-learner, U-learner, R-learner, X-learner, IPW-learner and DR-learner, are based on the following equations.

$$
\begin{aligned}
\tau(x) &= \mu_1(x) - \mu_0(x) \\
&= \mathbb{E}(Y(1) - \mu_0(X)|X = x) = \mathbb{E}(\mu_1(X) - Y(0) \mid X = x) \\
&= \mathbb{E}\Big\{ \frac{TY}{\pi(X)} - \frac{(1-T)Y}{1-\pi(X)} | X = x \Big\} \\
&= \mathbb{E}\Big\{ \frac{Y - \mu(X)}{T - \pi(X)} | X = x \Big\} \\
&= \mathbb{E}\Big\{ \frac{T\{Y - \mu_1(X)\}}{\pi(X)} - \frac{(1-T)\{Y - \mu_0(X)\}}{1-\pi(X)} + \mu_1(X) - \mu_0(X) | X = x \Big\}.
\end{aligned}
$$

5.4 Scenario 3: a combination of biased data and uniform data

Let $\mathcal{D}_{\mathcal{B}}$ and $\mathcal{D}_{\mathcal{U}}$ are the set of user-item pairs for the biased dataset and unbiased dataset, respectively. We assume the sample size of the unbiased dataset is much smaller than that of the biased dataset, i.e., $|\mathcal{D}_{\mathcal{U}}| << |\mathcal{D}_{\mathcal{B}}|$, since it is costly to collect unbiased samples through uniform policy.

Characters of biased and unbiased data

- Biased data: large sample size; it is inevitable to suffer from the problem of unmeasured confounders.
- Unbiased data: small sample size; no (unmeasured) confounding bias; it is a gold standard for evaluating the deibasing approaches.

5.4 Scenario 3: a combination of biased data and uniform data

For most tasks in RS, both of the biased and the unbiased data have the same data structure given as

$$\begin{cases} \text{Biased data} & \{(X_{u,i}, T_{u,i}, Y_{u,i}) : (u,i) \in \mathcal{D}_{\mathcal{B}}\}, \\ \text{Unbiased data} & \{(X_{u,i}, T_{u,i}, Y_{u,i}) : (u,i) \in \mathcal{D}_{\mathcal{U}}\}. \end{cases}$$

For illustration, we assume the treatment is binary and $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$ is the target quantity.

5.4 Scenario 3: a combination of biased data and uniform data

Setting 1: no hidden confounders in the biased data

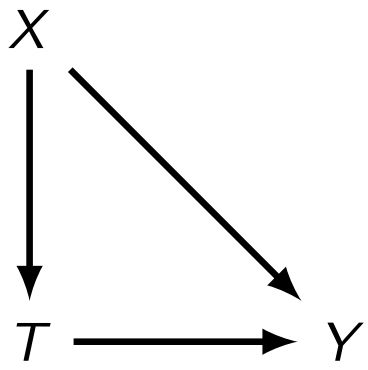


Figure 6: Biased data

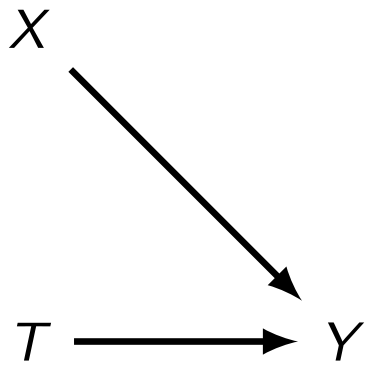


Figure 7: Unbiased data

5.4 Scenario 3: a combination of biased data and uniform data

In this setting, we assume the conditional exchangeability (Assumption 4) holds in the biased data, which implies that X includes all confounders (or blocks every backdoor path between T and Y) and conditional on X is enough to control the confounding bias. In this case, $\mu_1(x)$ is recoverable with the biased data solely under Assumptions 1-4, as discussed in section 3.

A natural question is: whether the unbiased data is helpful to improve the quality of recommendations? Intuitively, the unbiased data provides a better way to evaluate the resulting recommendation model, and hence it may give a better optimizing direction for training the model parameters.

5.4 Scenario 3: a combination of biased data and uniform data

Setting 2: hidden confounders exist in the biased data

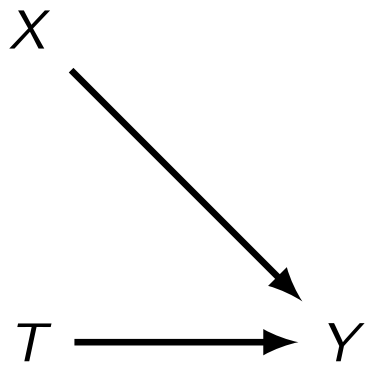


Figure 8: Unbiased data

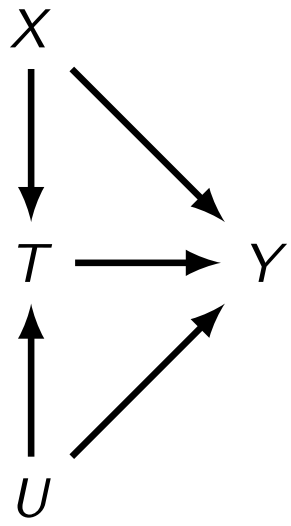


Figure 9: Biased data

where U is unmeasured confounder.

Example: I select a movie to watch, one reason is that a friend tells me it is a good movie. But the current recommender system is hard to collect this kind of confounders.

5.4 Scenario 3: a combination of biased data and uniform data

If there exist hidden confounders, then the conditional exchangeability assumption would be violated in the biased data. Without loss of generality, we consider the task of CTR prediction and focus on the causal estimand $\mu_1(x)$. Define

$$w(x) = \mathbb{E}[Y_{u,i} \mid X_{u,i} = x, T_{u,i} = 1].$$

When there exist some hidden confounders, $\mu_1(x_{u,i}) \neq w(x_{u,i})$. Their difference $\eta(x) = \mu_1(x) - w(x)$ reflects the effect resulted from the unmeasured confounders. Since we can estimate $w(x)$ directly, it is sufficient to estimate the control function $\eta(x)$.

- *Assumption 8* (model specification on the control function) $\eta(x) = \eta_\gamma(x)$.

5.4 Scenario 3: a combination of biased data and uniform data

Thus, an estimation strategy can be given as: first, estimate $w(x_{u,i})$ using $\mathcal{D}_{\mathcal{B}}$; second, estimate $\eta_\gamma(x_{u,i})$ using $\mathcal{D}_{\mathcal{U}}$, for example, minimizing

$$\arg\min_{\gamma} \sum_{(u,i)\in\mathcal{D}_{\mathcal{U}}} (r_{u,i} - \hat{w}(x_{u,i}) - \eta_\gamma(x_{u,i}))^2,$$

where $\hat{w}(x_{u,i})$ is an estimate of $w(x_{u,i})$. If we denote $\hat{\eta}(x_{u,i}) = \eta_{\hat{\gamma}}(x_{u,i})$, then the final estimate of $\mu_1(x_{u,i})$ is $\hat{w}(x_{u,i}) + \hat{\eta}(x_{u,i})$. Here, the model adopted for the control function should be simple; otherwise, it will suffer the problem of overfitting.

5.5 Scenario 4: Policy learning

The recommendation problem also can be as a policy learning problem.

Definition (Policy)

A policy π is a map from the space $\mathcal{X}$ of feature to a probability distribution over the treatment space $\mathcal{T}$. Specifically, $\pi(t|x)$ satisfies that $\sum_{t\in\mathcal{T}} \pi(t|x) = 1$ for any $x \in \mathcal{X}$.

Policy learning seeks to find the optimal policy that maximizes the policy value, which is defined as follows.

Definition (Policy value)

Policy value $V(\pi)$ refers to the expectation of the reward under the policy π, i.e.,

$$V(\pi) = \mathbb{E}\left[\sum_{t\in\mathcal{T}} \pi(t|X)Y(t)\right] = \mathbb{E}\left[\sum_{t\in\mathcal{T}} \pi(t|X)\mu_t(X)\right].$$

The best possible policy is $\pi^* = \arg\max_{\pi\in\Pi} V(\pi)$, where Π is the space consisting of all possible policies.

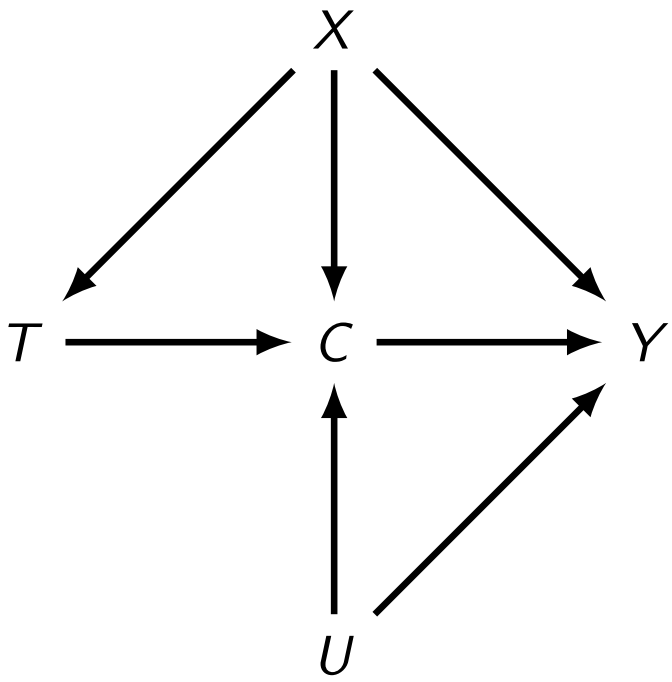
5.5 Scenario 4: Policy learning

Example. Suppose that there are a total of I items and U users.

- A unit is a user; The target population is all the users;
- the feature X_u is the attribute of user u;
- the treatment T_u has I levels, denoted as $\mathcal{T} = \{1, 2, \cdots, I\}$, where $T_u = i$ means that item i is exposed to user u;
- The reward caused by user u exposed to the item i as the potential outcomes is denoted by $Y_u(i)$.

The observed data is $\{(X_u, T_u, Y_u), u = 1, ..., U\}$. And the target quantity is the optimal policy π^*.

5.6 Scenario 5: Noncompliance



Data: $\{X_i, T_i, C_i, Y_i\}$, $i = 1, ..., n$.

- T: exposure; C: click; Y: converse;
- U: unmeasured confounders.

where C and Y are post-treatment variables.

5.6 Scenario 5: Noncompliance

Let T be the treatment, and define potential outcome

$$C(1), C(0)$$

as the potential click behaviors. Let

$$Y(t, c) = Y(t, C(t) = c)$$

be the potential conversion if $T = t$ and $C(t) = c$. Denote $Y(t) = Y(t, c)$.

- CTR (effect of T on $C(1)$):

$$\tau_1(x) = E[C(1) \mid X = x] \tag{7}$$

  Usually, $C(0) = 0$, $\tau_1(x) = E[C(1) - C(0) \mid X = x]$.

- CVR (effect of T on $Y(1)$)

$$\tau_2(x) = E[Y(1) \mid X = x] \tag{8}$$

- TCVR (effect of T on $Y(1) - Y(0)$)

$$\tau_3(x) = E[Y(1) - Y(0) \mid X = x] \tag{9}$$

5.6 Scenario 5: Noncompliance

Let C be the treatment, Post-click CVR (effect of C on $Y(1)$),

$$\tau_4(x) = E[Y(1) - Y(0) \mid X = x, C = 1]. \tag{10}$$

Summary

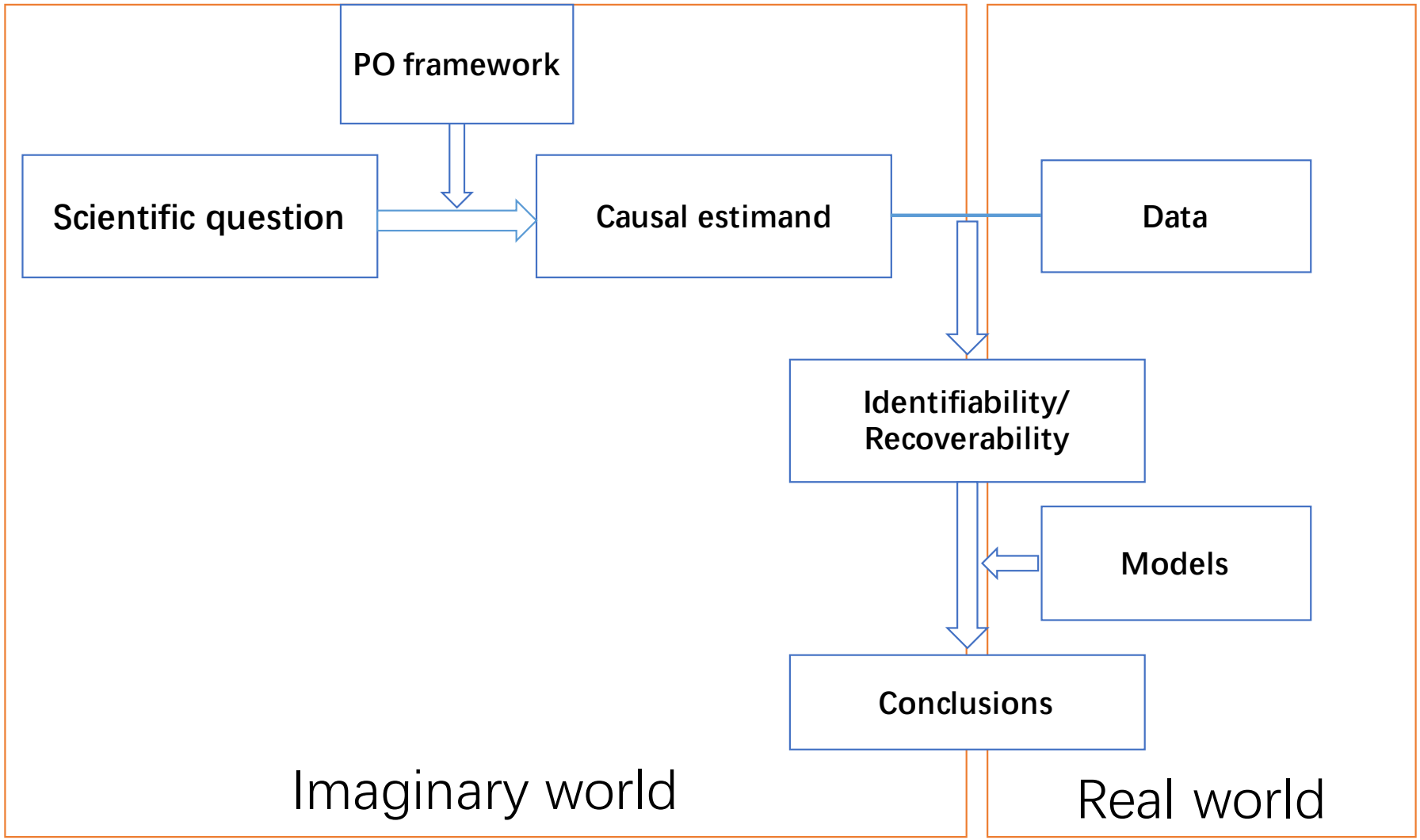


Figure 10: Causal analysis framework

Summary

- Explicating the perplexing causal concepts in RS within the potential outcomes.
- Providing a guideline of how to define, recover and estimate a causal estimand.
- Providing a new taxonomy and giving formal definitions of various biases in RS from the perspective of violating what assumptions are adopted in standard causal analysis.
- Unifying many causal problems and debiasing methods in RS into a few scenarios.
- Revealing the key assumptions underlying various debiasing approaches.

Discussion

How to determine which assumptions are violated given a scenario? When applying the proposed framework to specific problems, practitioners /researchers should

1. first determine the research goals and formulate them as causal estimands
2. and then consider the possible biases produced during data collection. If the data is missing and cannot represent the target population, that is selection bias
3. Next, we should discuss the relationships between features, treatment, and outcome. If the features influence both treatment and outcome, there exists confounding bias. Moreover, if we suspect that there are some unmeasured confounders, see data fusion, IV....
4. In addition, if we have two variables of interest measured after the treatment, it may be a non-compliance problem.

Discussion

How to verify the assumptions? As discussed throughout, we need a variety of assumptions to climb from association (data) to causality (causal conclusions). These assumptions can be divided into

- associational assumptions: e.g. Assumption 7, model specification of propensity score. It is testable in principle.
- causal assumptions: such as SUTVA, consistency, and conditional exchangeability, cannot be directly verified from data, unless one resorts to experimental control.
- In practice, whether the causal assumptions hold need to be discussed by expert's knowledge (e.g. drawing causal graphs) for each specific problem.

Discussion

Causal and RS

- Classical causal inference focus on estimation and inference. For example, through estimating causal effect to answer various scientific questions. (Based on estimating equation and score function)
- Many machine learning methods are pure prediction algorithms: e.g., deep learning models. (Based on loss function)
- Causal learning: to improve the stability, interpretability, and generalization ability. (Trade-off between prediction accuracy and other desired properties)
- Recommendation is a very interesting task:
  - Prediction accuracy is the gold standard; This implies that recommendation is a prediction problem.
  - However, classical causal inference is not designed for prediction, so it may not improve the prediction accuracy.

Discussion

- Why the prediction accuracy can be improved by using causal methods?
  - Treat the recommendation problem as an out of distribution (OOD) problem, that is, the distributions between training set and test set are different. (causal learning)
  - Many practical problems of interest in RS are essentially counterfactual (or causal) problems, such as post-view click-through rate prediction, post-click conversion rate prediction, and uplift modeling. In these cases, the potential outcome is of interest, instead of the observed outcome. A model that fits the observed outcome well may not fit the potential outcome well.
- We may consider the recommendation problem as two subproblems:

$$\textit{Prediction} + \textit{Debiasing}.$$