Causality-inspired Recommendation:

Robustness, Transparency and Fairness

Xiangmeng Wang

Leader: Data Science and Machine Intelligence Lab
School of Computer Science, Advanced Analytics Institute
University of Technology Sydney

Outline

- Part I: Introduction
  - Trustworthy Recommendation
  - Three-layer hierarchy: robustness, transparency and fairness
  - Causal learning theory
  - Causal learning for Trustworthy Recommendation
- Part II: Featured Research
  - Causal learning approaches
  - Featured research on causality-inspired Trustworthy Recommendation
- Part III: Future work

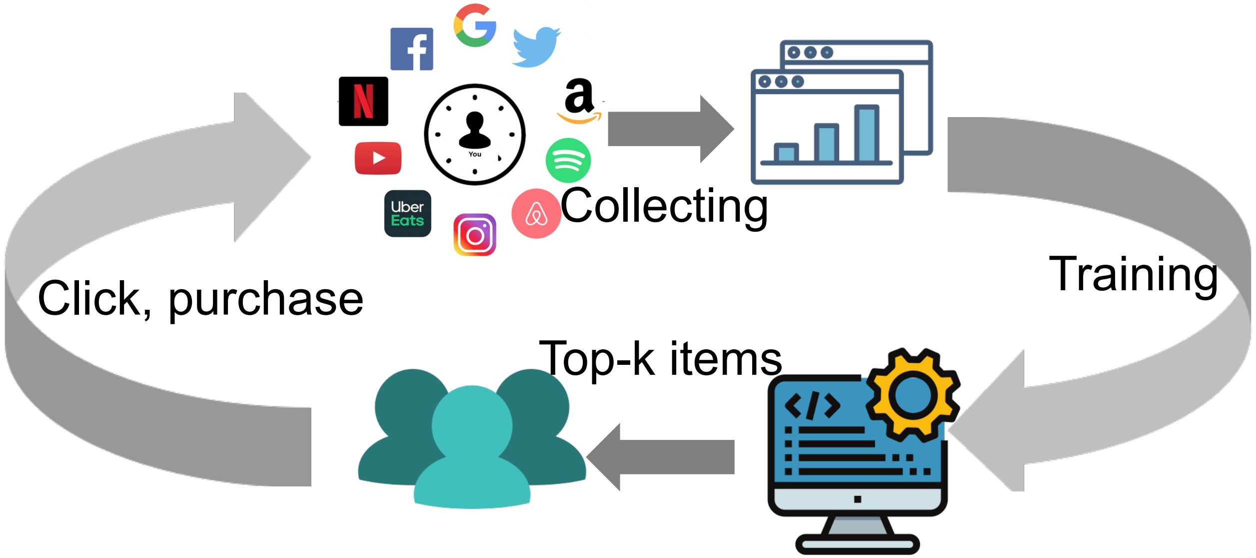
Part I: Introduction

Recommender system (RecSys)

An **information filtering** technique, which provides users with information that he/she may be interested in.

- User Satisfaction
- Advertising Profits

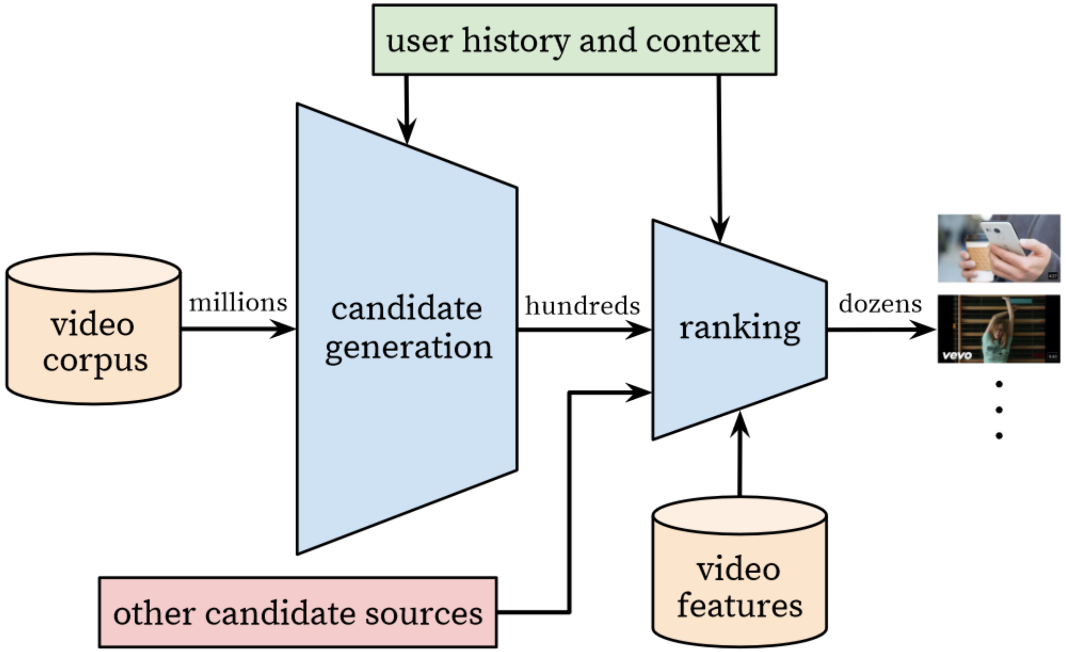
Purpose



Recommendation Loop

Problem formulation

- **Input**: Items (e.g. video corpus), user-item interactions (e.g., user view history and content) or other data source (e.g., user/video features)
- **Output**: A few items (e.g. videos) are filtered or ranked and then show them to the users.
- **Evaluation**: system utility (e.g., ranking accuracy)



RQ:

Whether the model makes accurate predictions?

Classic models

Basic assumption: Minimize the gap between historical feedback (observational) and prediction

1	**2**	?
4	?	8
0	?	6

observational data

$\approx$

1	**2**	**9**
4	**5**	8
0	**6**	6

model prediction

- Collaborative filtering
  - Latent factor models
- Shallow representation
  - Matrix factorization
  - Factorization machine
- Deep representation
  - Neural collaborative filtering
  - Graph neural representation

- **Data driven:** The model performance is highly depend on the quality of observational data.
- Consider **utility such as model accuracy** only

Shortcomings of classic models

Classic RecSys models are data-driven, and they consider utility, such as model accuracy only, cause:

- **Unrobustness:**
  - Data bias, data missing and data noise cause unrobust model training
  - The model may be affected by hidden factors (e.g., social media)
- **Lack explainability**
  - Classic RecSys retain black-box nature.
  - User feedback usually entangles users' real interests, hard to generate post-hoc explanations
  - Does not consider explanation evaluation
- **Fairness**
  - Data may contain sensitive information such as user genders
  - Does not consider fairness evaluation

Trustworthy Recommender Systems

Aims to **competent** RecSys that incorporates the core aspects of trustworthiness such as explainability, fairness, robustness, privacy and controllability.

Robustness

Explainability

Fairness

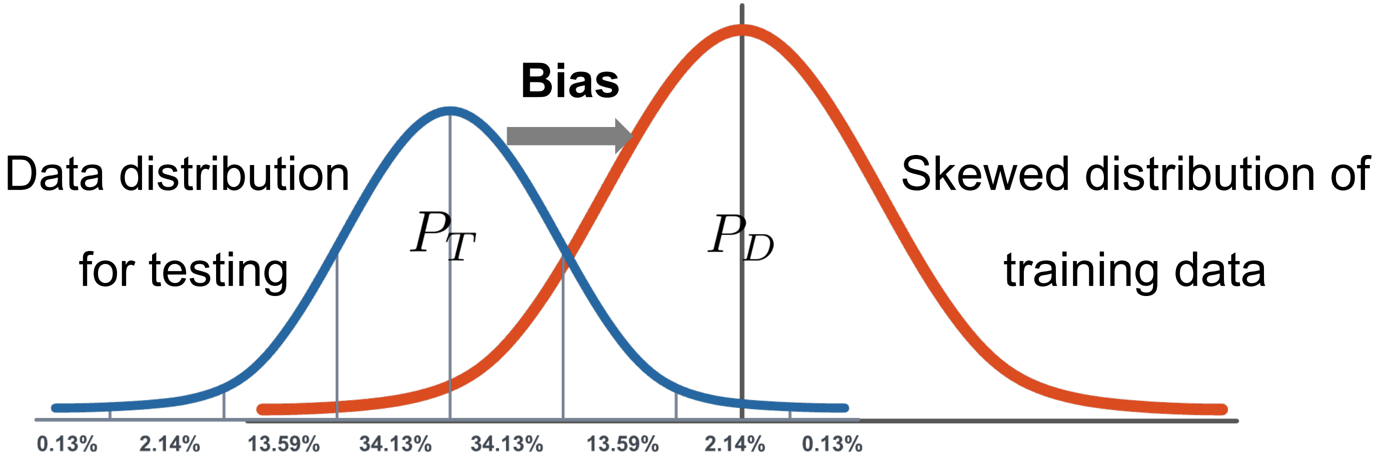
- Improve system responsibility
- Gain trust from users
- Promote recommender systems for social good

Three-layer hierarchy to trustworthy RecSys

Trustworthy AI: A Computational Perspective, ArXiv: 2107.06641, 2021.
Tutorial: https://sites.google.com/msu.edu/trustworthy-ai/

Robustness Issue

Data bias: the distribution of observational data is different from the ideal data distribution (experimental).

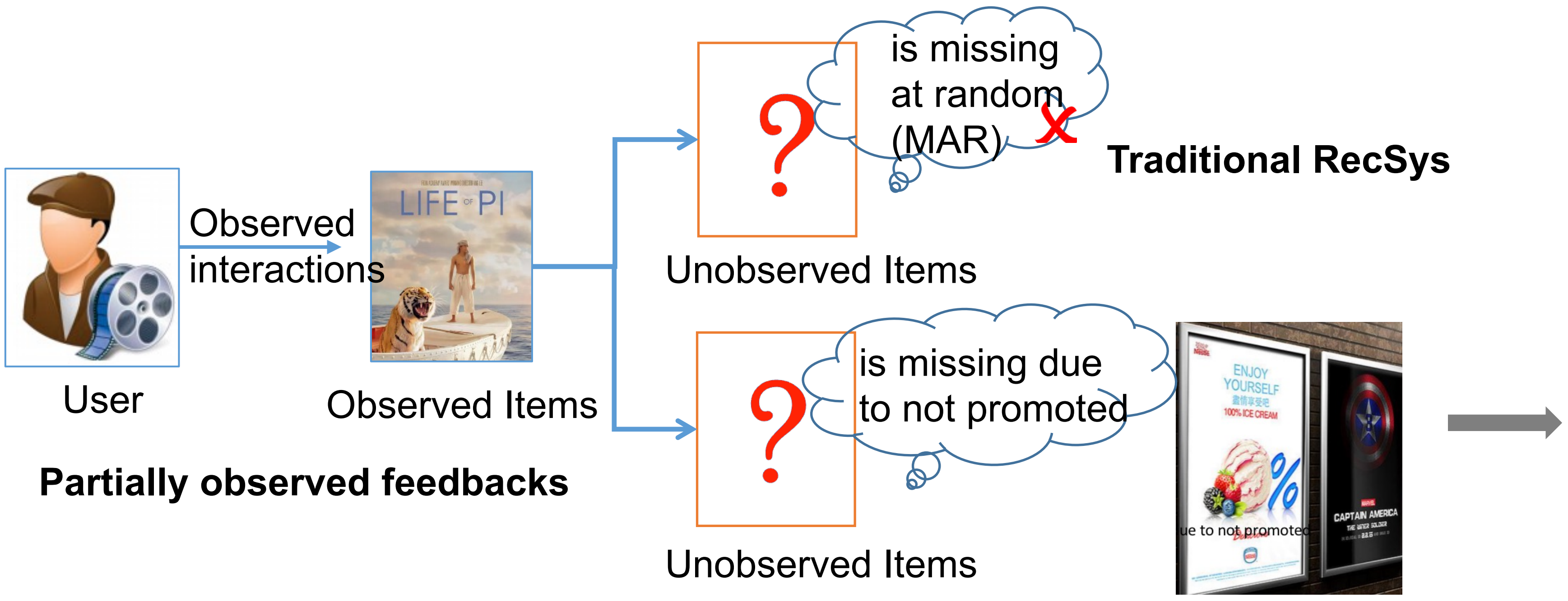


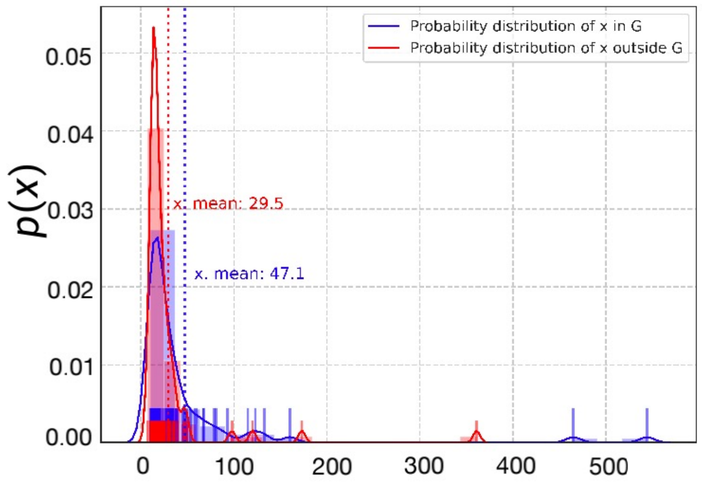
- **Data bias is everywhere:**
  - Biased data collection
    - e.g., uneven exposure of items
  - User give wrong feedback to items
    - e.g., user conformity

Distribution difference between P_T and P_D → Risk discrepancy between $\hat{L}_T(f)$ and $L(f)$ → Suboptimal accuracy $f^* \neq f$

Robustness Issue

Data missing: unobserved user-item feedback cannot be collected



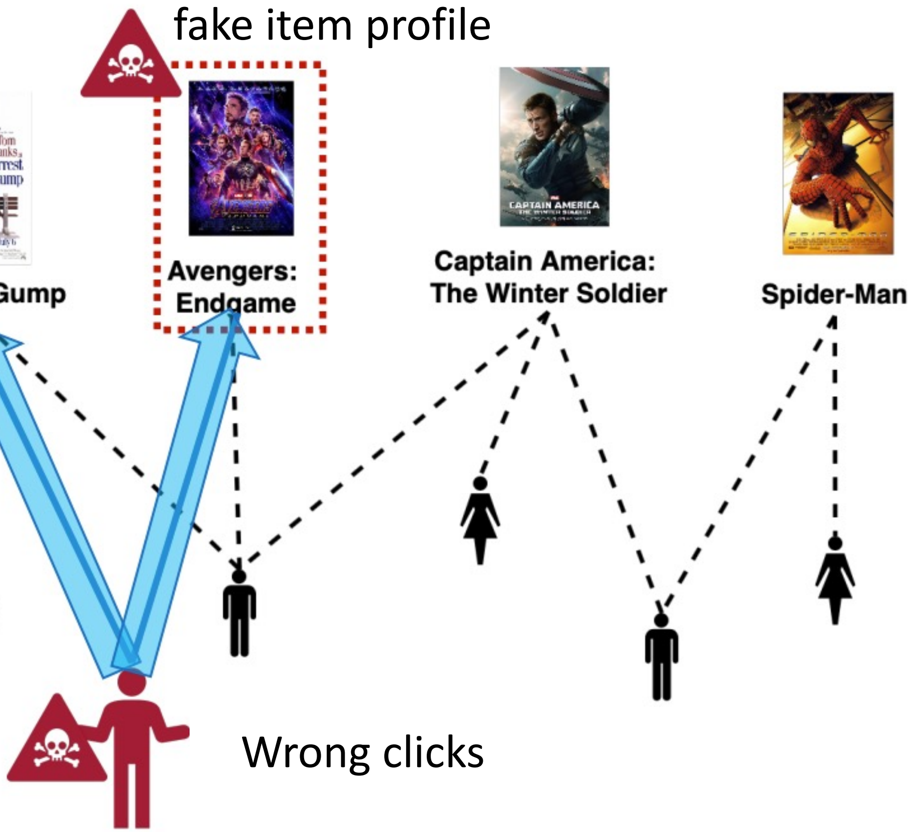


(b) Distribution on Epinions.

- Data missing causes uneven item exposure
- The trained model will further deprive the exposure of unexposed items
  - i.e, the poor gets poorer phenomenon

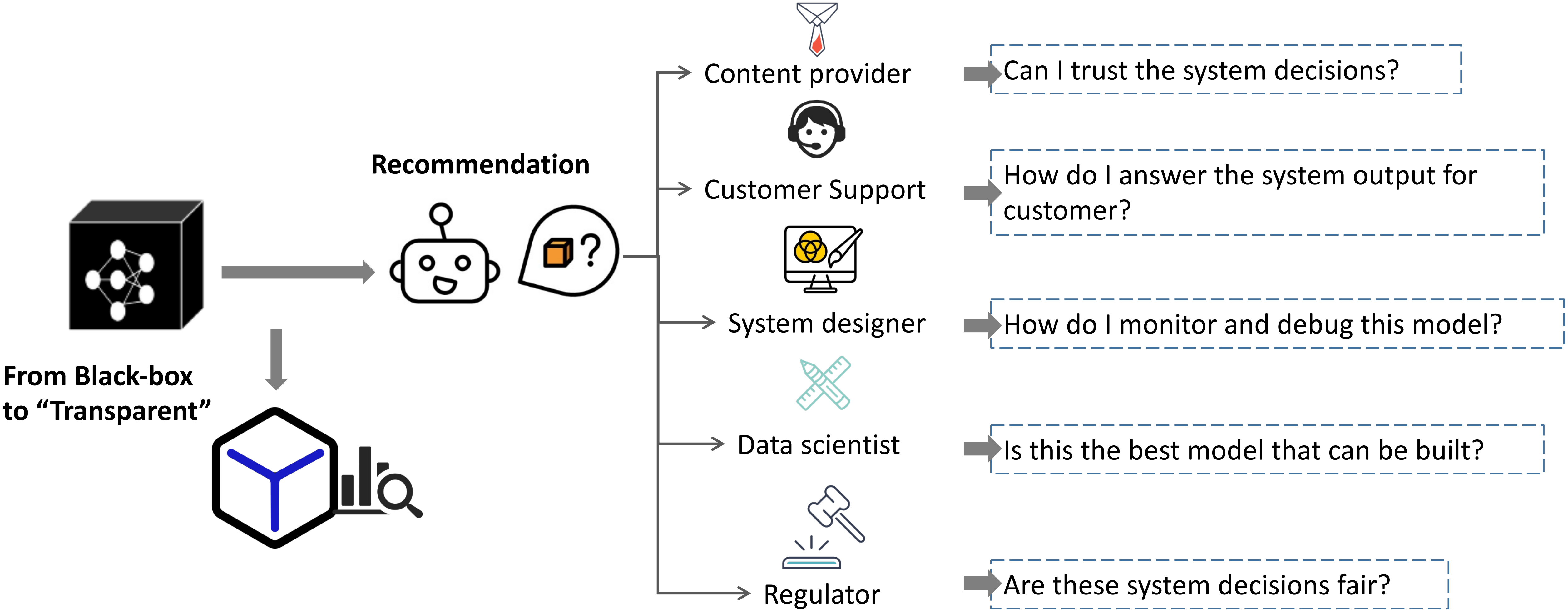
Robustness Issue

Data noise: observed user feedback or context information may be noisy, not reflecting the actual satisfaction of user



Explainability Issue

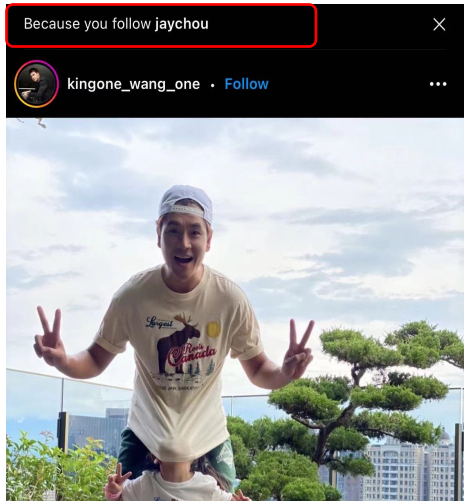
Black-box recommendation model creates confusion and doubt

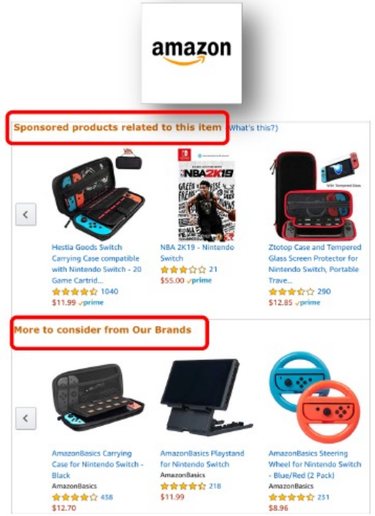


Explainability Issue

User persuadableness

- provide personalized recommendations complemented with explanations to answer: Why such items are recommended to you?





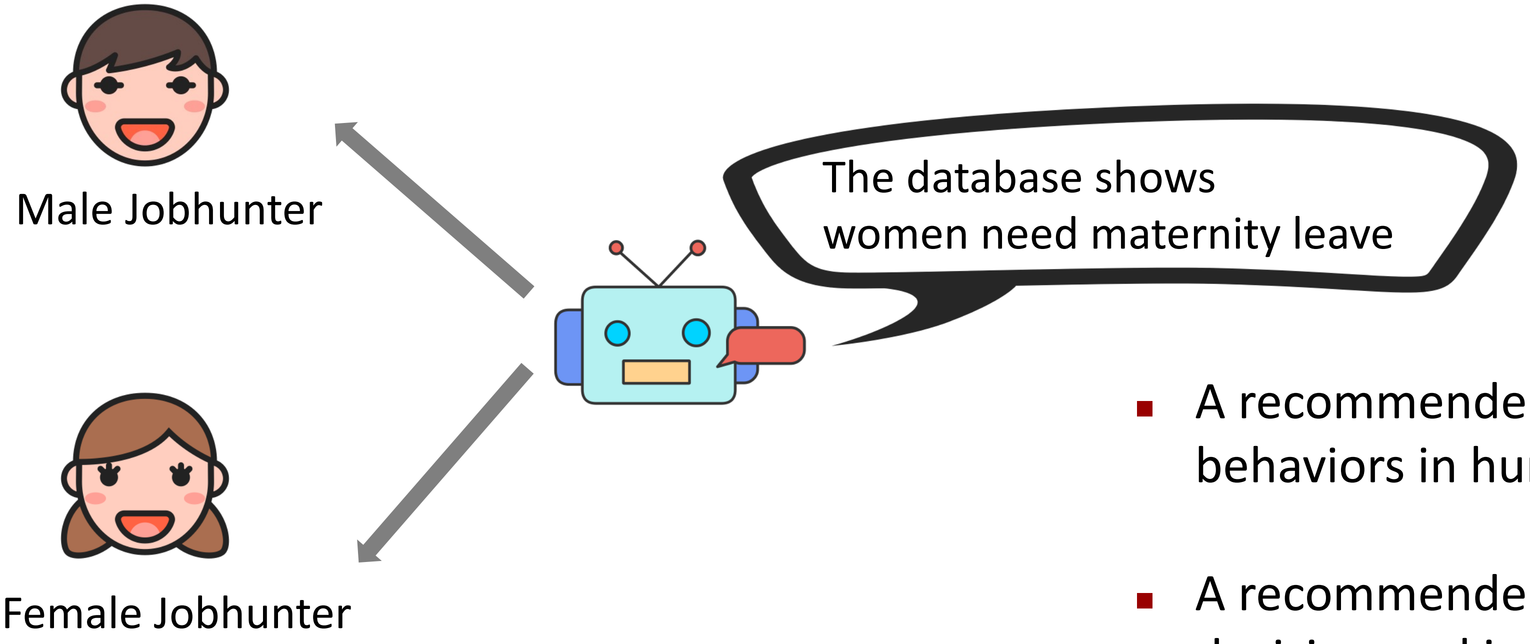
- Win users' trust in recommender systems
- Improve recommendation persuasiveness

Model diagnostics

- help system developer understand what can be done to improve the model

Fairness Issue

Refer to **unfair allocations of recommended items**, caused by e.g., gender discrimination



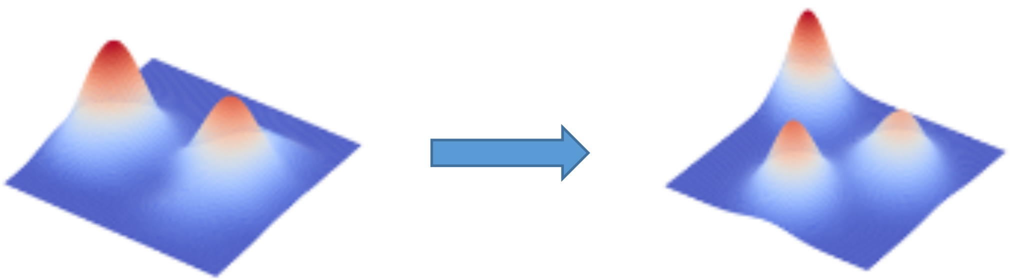
- A recommender system should avoid discriminatory behaviors in human-machine interaction.
- A recommender system should ensure fairness in decision-making.

Causal learning v.s. Correlation learning

Classic data-driven models:

- Data-driven models may infer spurious correlations which would not reflect user true preference and are not interpretable.

$$P_D(u, i) \approx P_T(u, i)$$

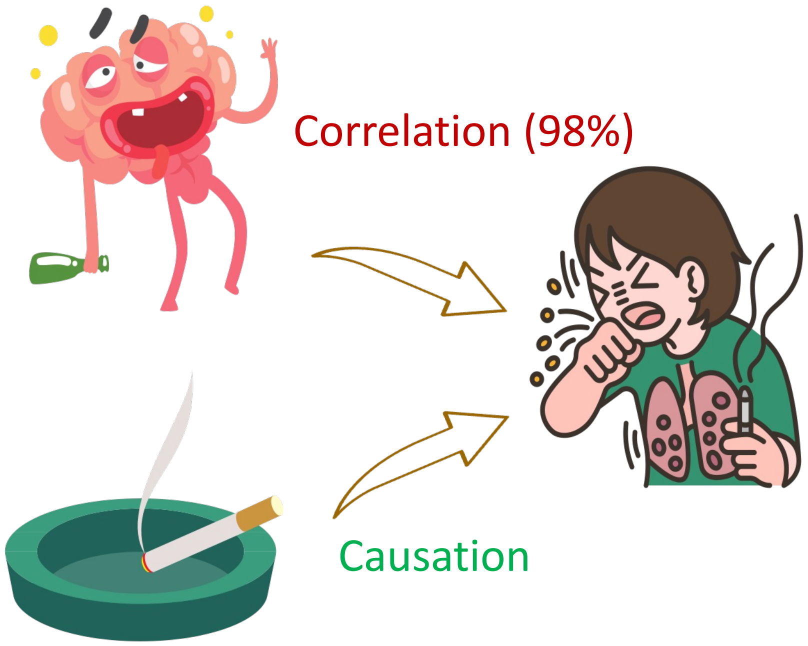


True preference distribution on testing data (stable causation)

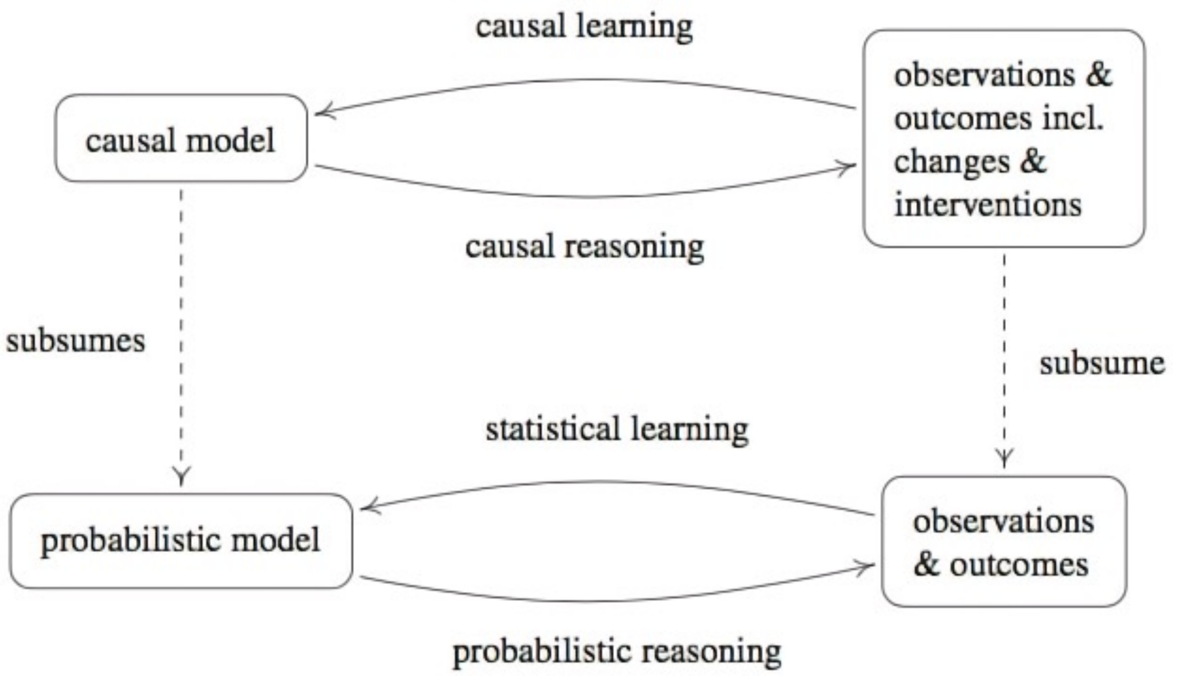
Skewed preference distribution on training data (spurious correlation)

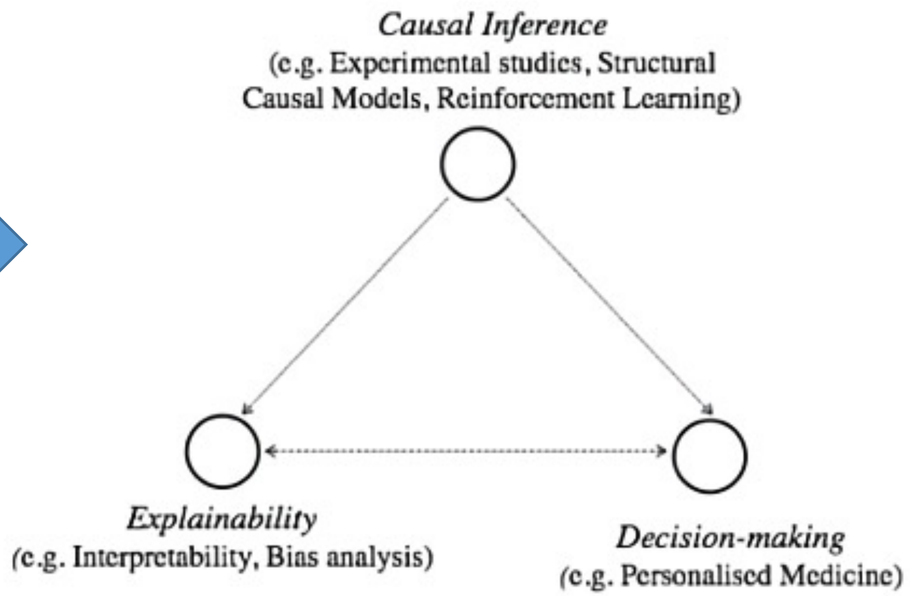
Causal learning models:

- Relationships where an intervention in one variable (cause) contributes to a change in another variable (effect).



Causal learning v.s. Correlation learning



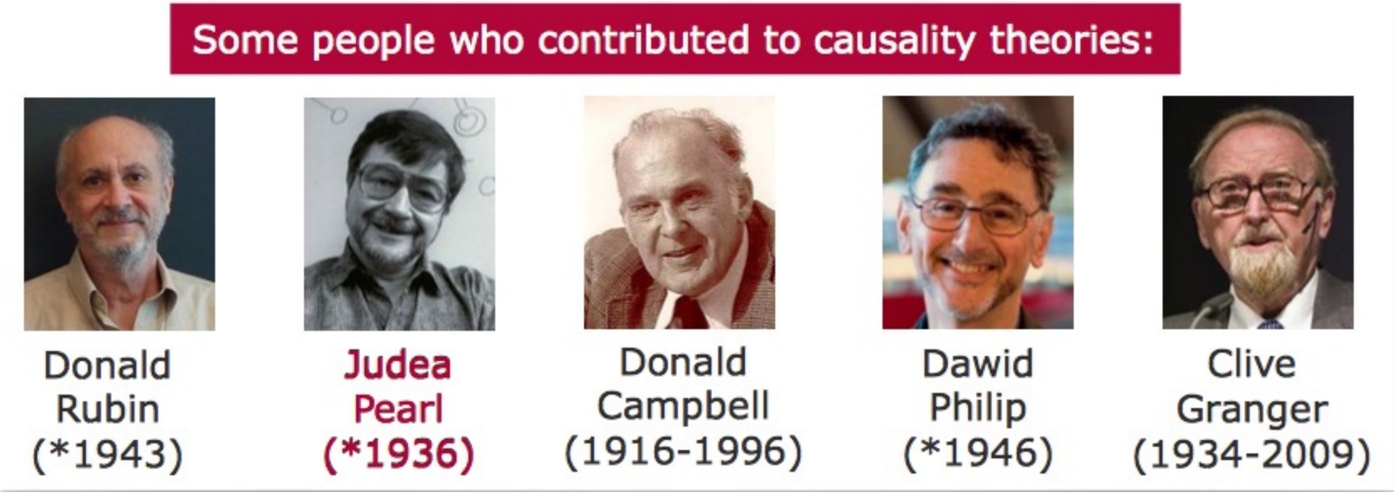


Causal reasoning & probabilistic reasoning

three pillars of causal inference

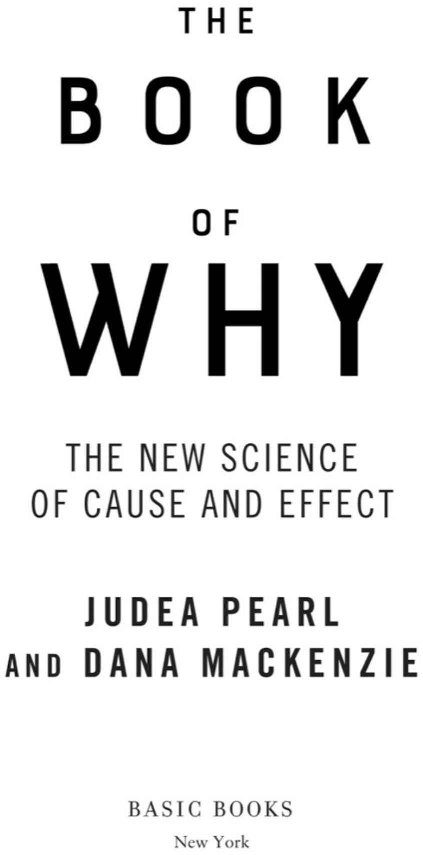
Causal learning

Background



- Causality theory helps to decide when, and how, causation can be inferred from domain knowledge and data.
- The basis of a causality theory is causal model that provides a language to encode causal relationships

Causal learning



THE

BOOK

OF

WHY

THE NEW SCIENCE
OF CAUSE AND EFFECT

JUDEA PEARL
AND DANA MACKENZIE

BASIC BOOKS
New York

ACM Turing Award 2011:
"For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning."

Causal learning

Causal inference is driven by applications and is at the core of statistics (the science of using information discovered from collecting, organising, and studying numbers)

- **Many origins of causal inference**
  - Biology and genetics;
  - Agriculture;
  - Epidemiology, public health, and medicine;
  - Economics, education, psychology, and other social sciences;
  - Artificial intelligence and computer science;
  - Management and business.

Causal learning

What does causal learning bring?

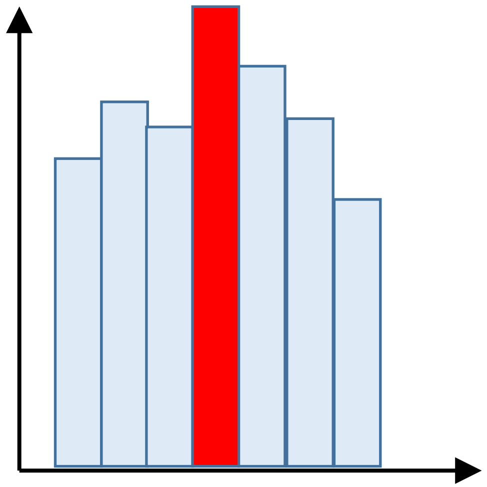
Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing *X* change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing, Intervening	What if? What if I do *X*?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it *X* that caused *Y*? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past two years?

Causal learning

1 **Description**

What is there?

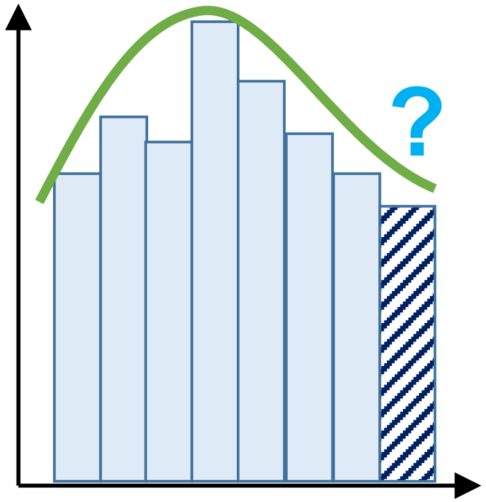
e.g., what month do items sell the most?



2 **Prediction**

What will happen?

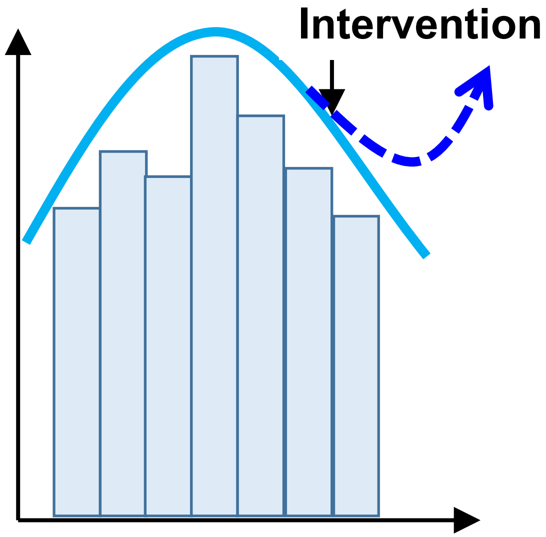
e.g., how many items will we sell in next month



3 **Causal Inference**

What would happen?

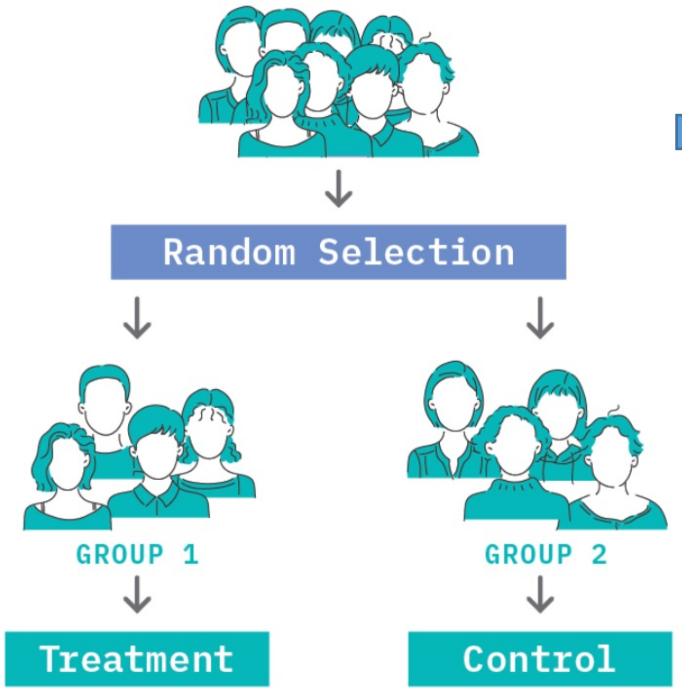
e.g., how much more items we sell if we run more google ads



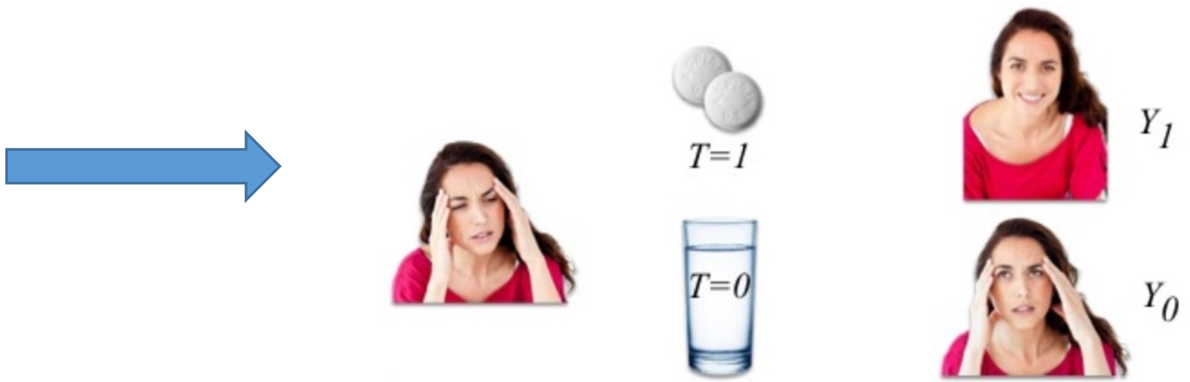
Causal learning

Intervention

- Assess the causal effect of some potential cause (e.g. an action, or event) on some outcomes



Randomized Controlled Trial



Causal effect

- Individual level: individual treatment effect (ITE) on the outcome for instance is the difference between its two potential outcomes

$$\tau_u = Y_u^1 - Y_u^0$$

- Population level: average treatment effect (ATE) is computed over the whole population

$$\tau_{ATE} = \mathbb{E}_{u \in U}[\tau_u] = \mathbb{E}_{u \in U}[y|do(1)] - \mathbb{E}_{u \in U}[y|do(0)]$$

Causal learning

Example

- ITE:

Instance: Bob

Treatment: $D = \begin{cases} 1 & \text{milk} \\ 0 & \text{no milk} \end{cases}$

Observed outcome: Y^1: asleep at 5: 00 am

Causal effect of milk: Y^1 | Y^0=?

Comparison

- ATE :

$$\tau_{ATE} = \mathbb{E}_{u \in U}[\tau_u] = \mathbb{E}_{u \in U}[Y_u^1 - Y_u^0]$$

  - ATE only requires to query interventional distributions but not counterfactuals

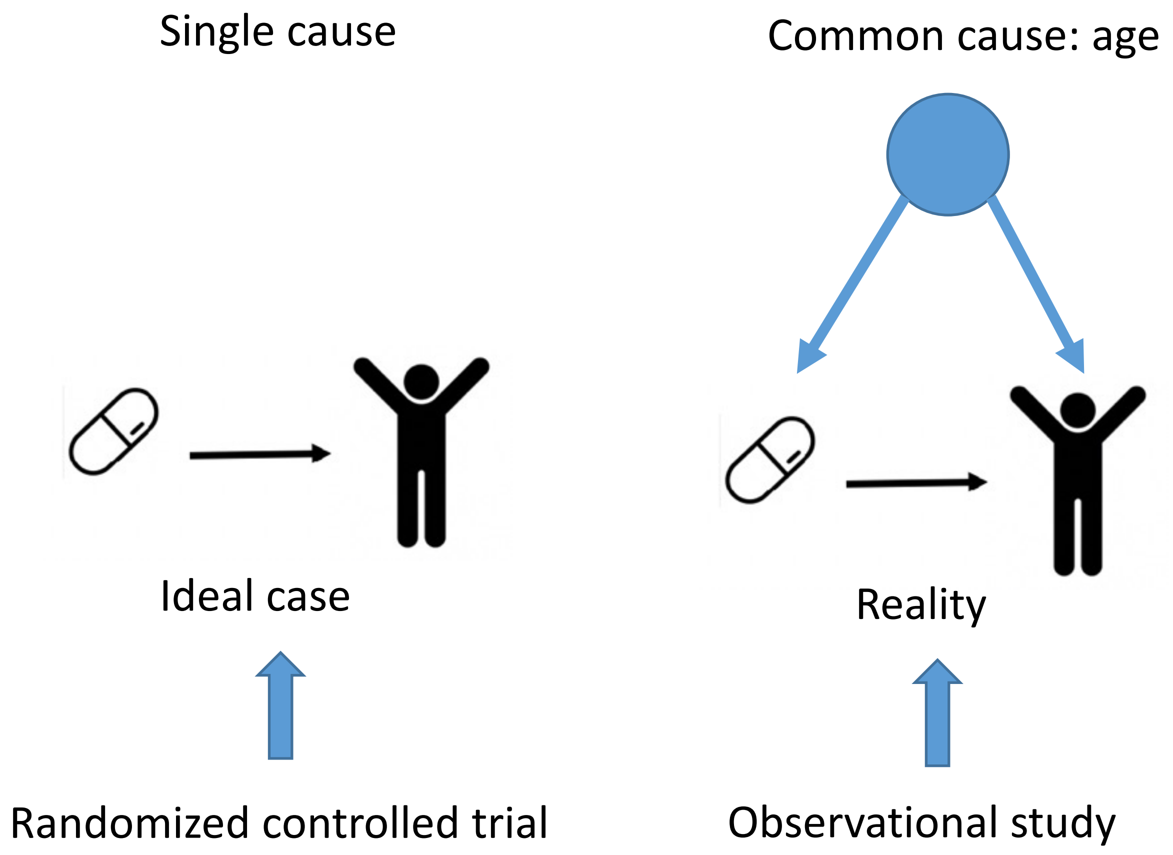
Causal learning

Confounder

- The assignment is not random in observational study (real-world senario)

- Randomized Controlled Trial
  - Randomly assign the control/treated
  - Gold-standard for studying causal learning
  - Time consuming and more ethical concerns

- Observational study
  - Assignment is NOT random
  - Confounding bias is presented

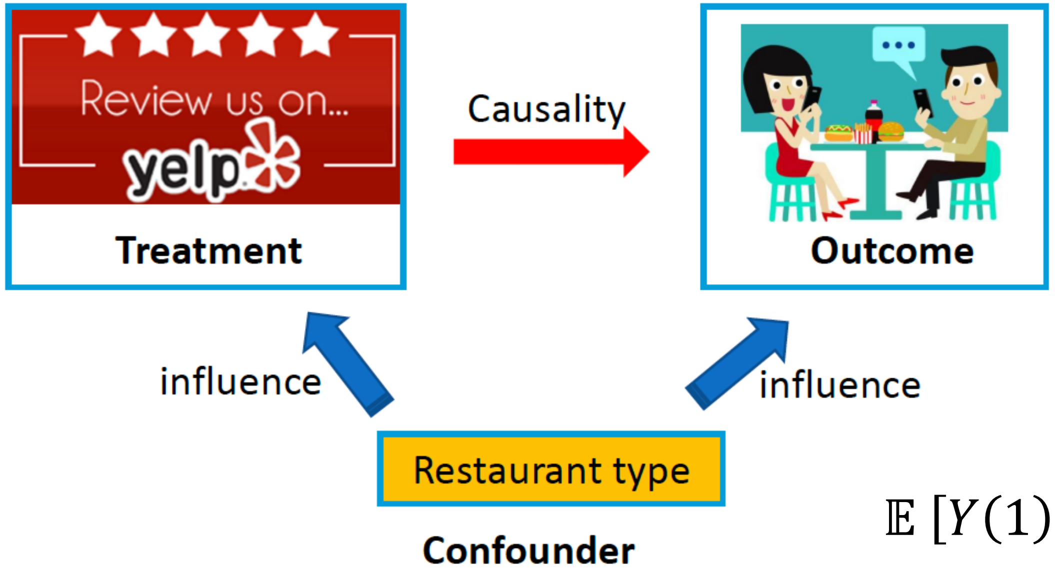


Causal learning

Confounder

- Notation
  - Treatment: the variable to be manipulated
  - Outcome: the variable that can be observed with some responses
  - **Confounder**: the variable influences both treatment and outcome

- Example



e.g., the number of customers

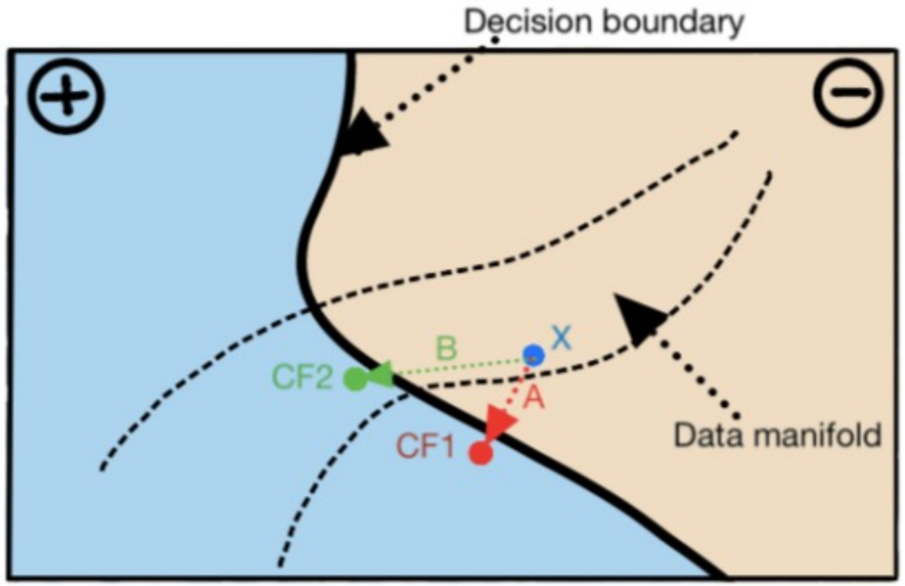
$$\mathbb{E}\left[Y(1)\right] - \mathbb{E}[Y(0)] \neq \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$$

Causal learning

Counterfactual

- Answers the "what if" question: e.g., what would the expected value of the demand Q have been if we were set the price at $\mathrm{P} = p_1$?

- Example:

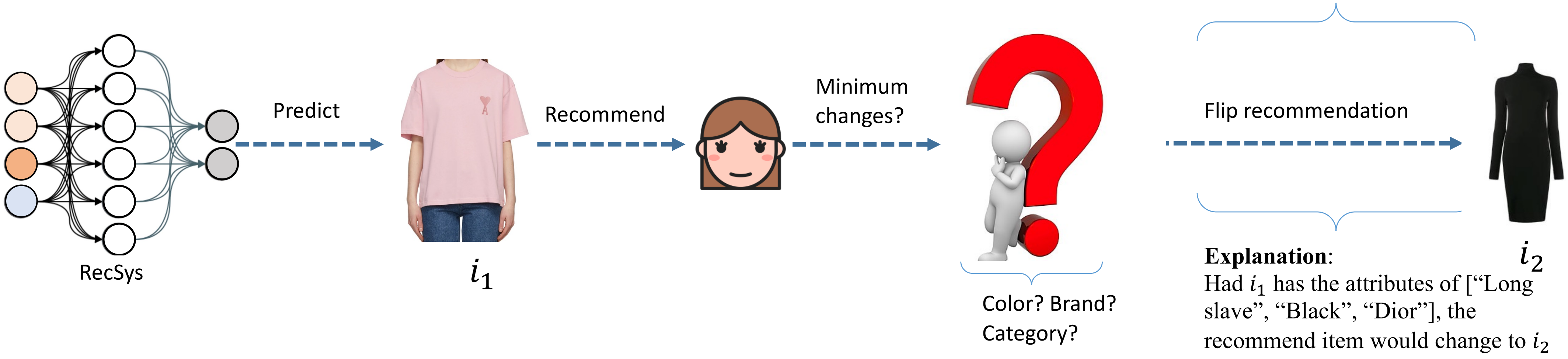


[Counterfactual explanation]
A minimal set of influential factors that, if applied, flip the model decision.

Causal learning

Counterfactual

- Application in Explainable RecSys

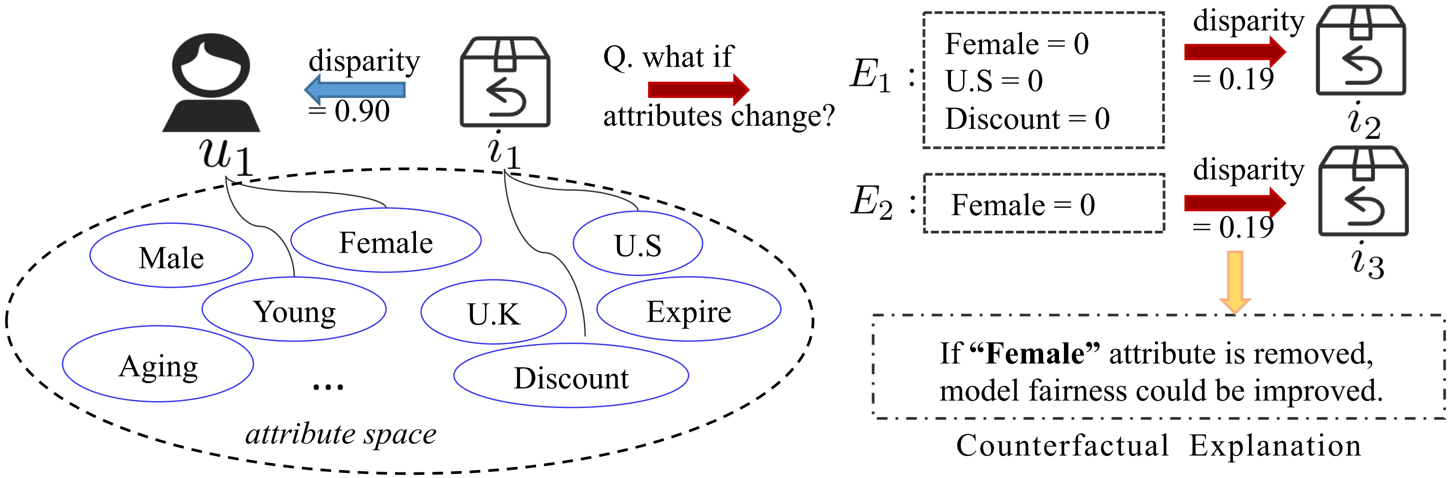


Explainable Recommendation

Causal learning

Counterfactual

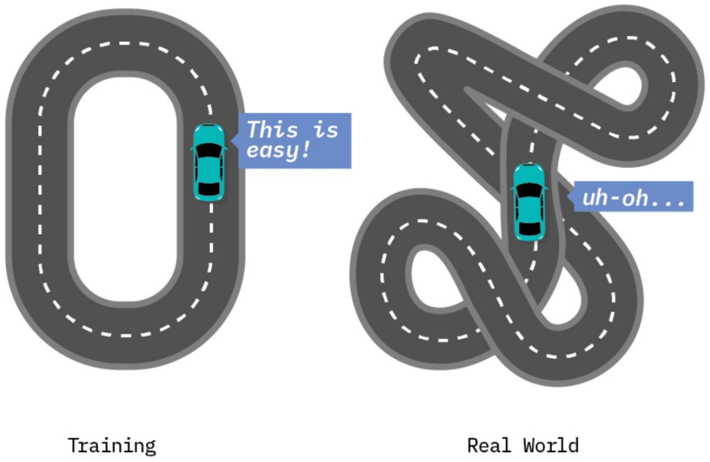
- Application in Trustworthy RecSys

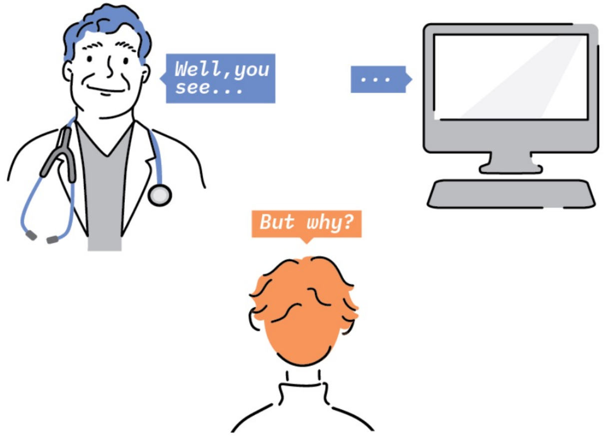


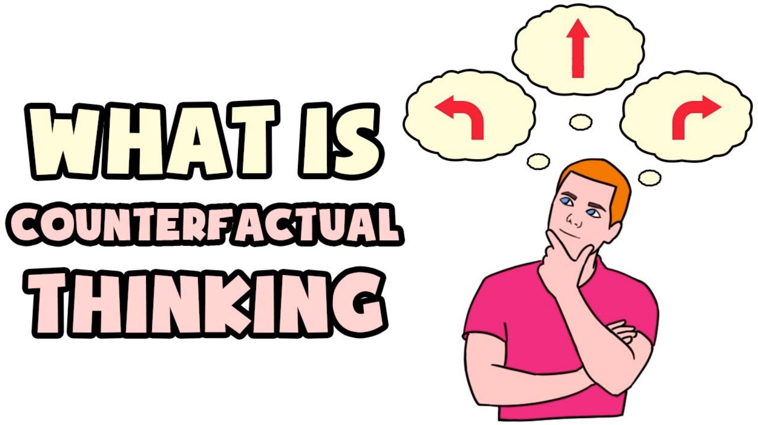
Fairness diagnostics

Causal learning for Trustworthy RecSys

Why causal learning







Deconfounding for robustness

Counterfactual reasoning for explainability

Counterfactual reasoning for fairness

Part II: Featured Research

Causal learning approaches

For observational studies, we need a definition of causality that does **not hinge on (explicit) randomisation**

Pioneers in causal inference have come up with three definitions/languages:

- **Stuctual Causal Model (SCM) -** Judea Pearl
- **Potential Outcome Framework (RCM) -** Donald Rubin

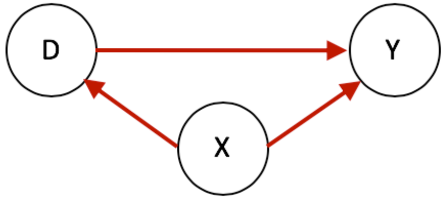
Stuctual Causal Model (Pearl's SCM)

Structural equation

- Each function represents a causal process

$$X = f_X(E_X)$$
$$D = f_D(X, E_D)$$
$$Y = f_Y(X, D, E_Y).$$

Structural equation



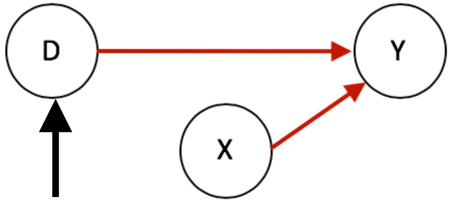
Causal graph

Causal graph

- A directed acyclic graph
- Error terms are jointly independent

Interventional and counterfactual logic

- An intervention on variable D by $do(D)$
- New graph is generated by removing all edges from parents to x_i
- Causal effect computation

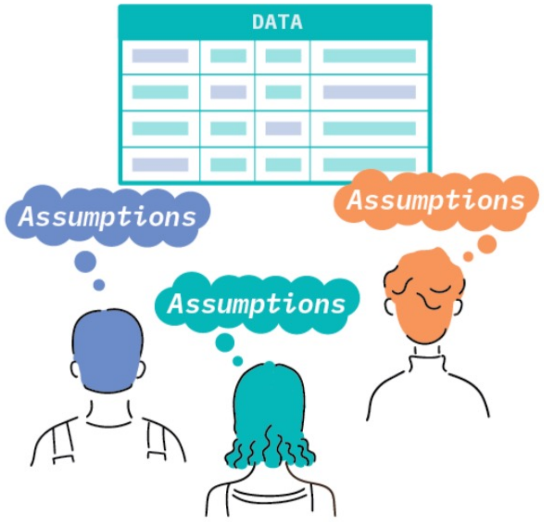


intervention

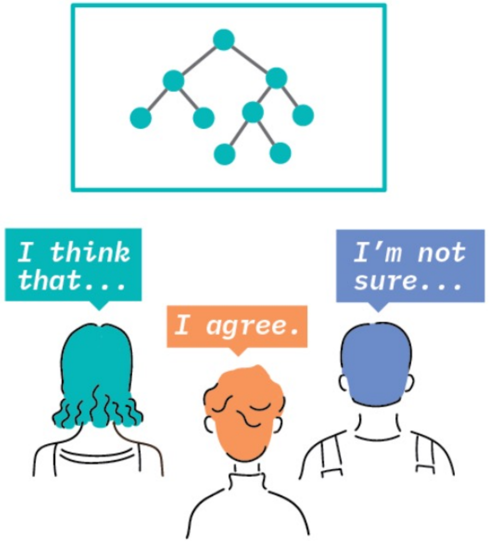
Stuctual Causal Model (Pearl’s SCM)

Causal graph

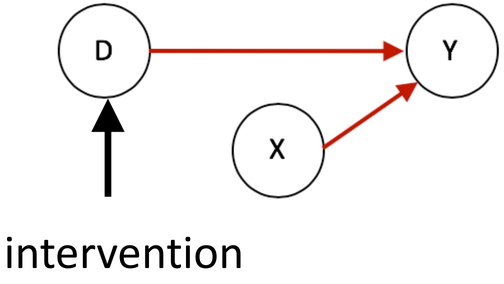
- Is developed based on assumptions
- Deconfounding: blocks bad effects from confounders (causal identification)



We all make causal assumptions when looking at data.



Making a causal graph makes those assumptions explicit.



- Control the confounder
- True causal effect:

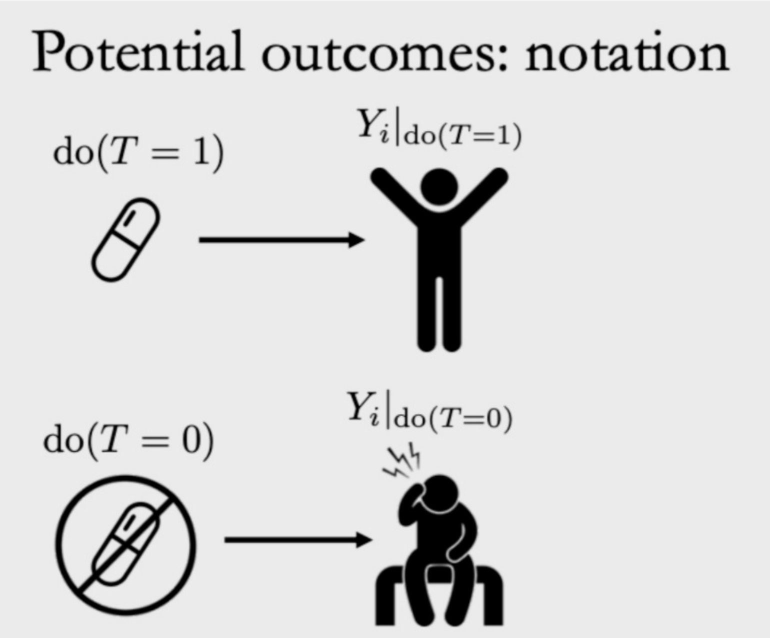
$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y \mid T = 1, C] - \mathbb{E}[Y \mid T = 0, C]$$

Potential Outcome Framework (Rubin Causal Model)

Potential outcome

- Definition: Given the treatment and outcome t, y, if the instance i is under treatment t, the potential outcome of instance is y_i^t

- Aims to directly model ITE or ATE:



$$\text{ITE:} \quad \tau i = y_i^1 - y_i^0$$

$$\text{ATE:} \quad \tau = E_i\,[\tau_i] = E_i[y_i^1 - y_i^0]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0)$$

RCM works under

- The stable unit treatment value assumption (SUTVA)
- Consistency
- Ignorability (unconfoundedness)

SCM v.s. RCM

- SCMs and RCMs are essentially interchangeable and **equivalent to each other**

- In the RCM, causal effects of variables other than treatment and instrumental variables are not defined.
  - We can model causal effects of interest **without knowing the complete causal graph.**

- RCM requires strong assumptions, such as unconfoundedness
  - **Cannot be applied to deconfounding learning.**

- In SCM, causal effects of any variable can be studied.
  - When studying **causal relationships between arbitrary sets of variables**, SCM is often the preferred approach.

Our researches on causality-inspired Recommendation

Bias Handling for Recommendation Robustness

- Selection bias mitigation in Social Recommendation
- Distribution shift in Reinforcement learning based-Recommendation

Explainable Recommendation

- Semantics-Aware Intent Learning - Explain users' intents with item semantics
- Counterfactual explanation for Recommendation

Fairness-aware Recommendation

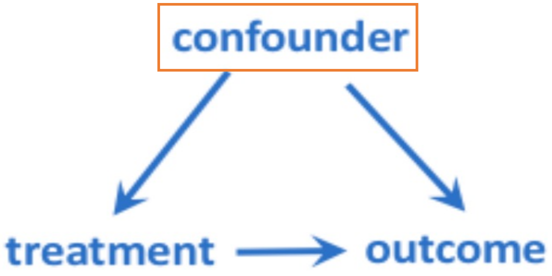
- Counterfactual explanation for Fairness

Selection bias mitigation in Social Recommendation

Be Causal: De-biasing Confounding in Recommendation

ACM Transactions on Knowledge Discovery from Data

- Data missing causes selection bias
  - In real-world social recommendations, the unobserved items are missing not at random (MNAR)
    - e.g., Users tend to watch movies watched by their friends
  - The MNAR results selection bias, which is attributed to the presence of confounders (social network)

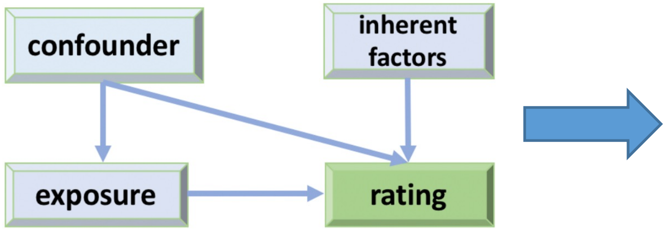


Selection bias mitigation in Social Recommendation

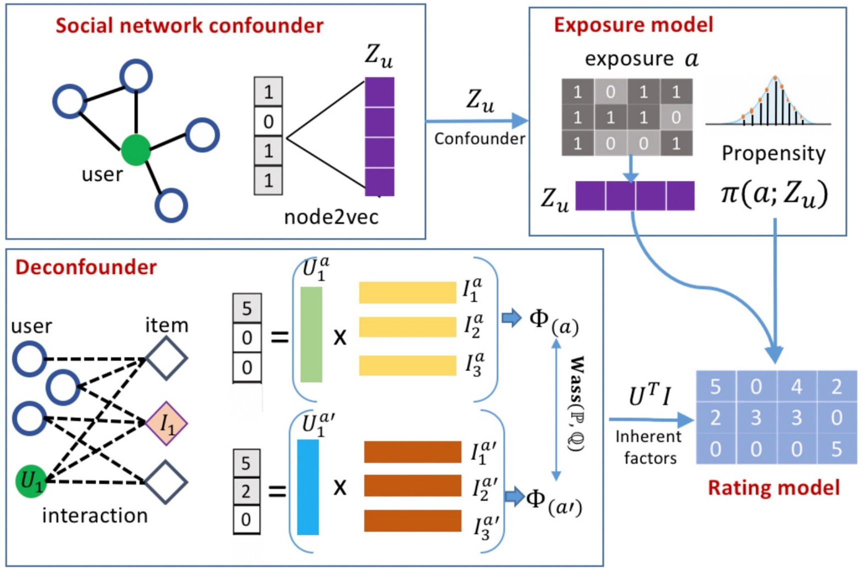
Be Causal: De-biasing Confounding in Recommendation

ACM Transactions on Knowledge Discovery from Data

- Causal graph-based model framework



Designed Causal graph



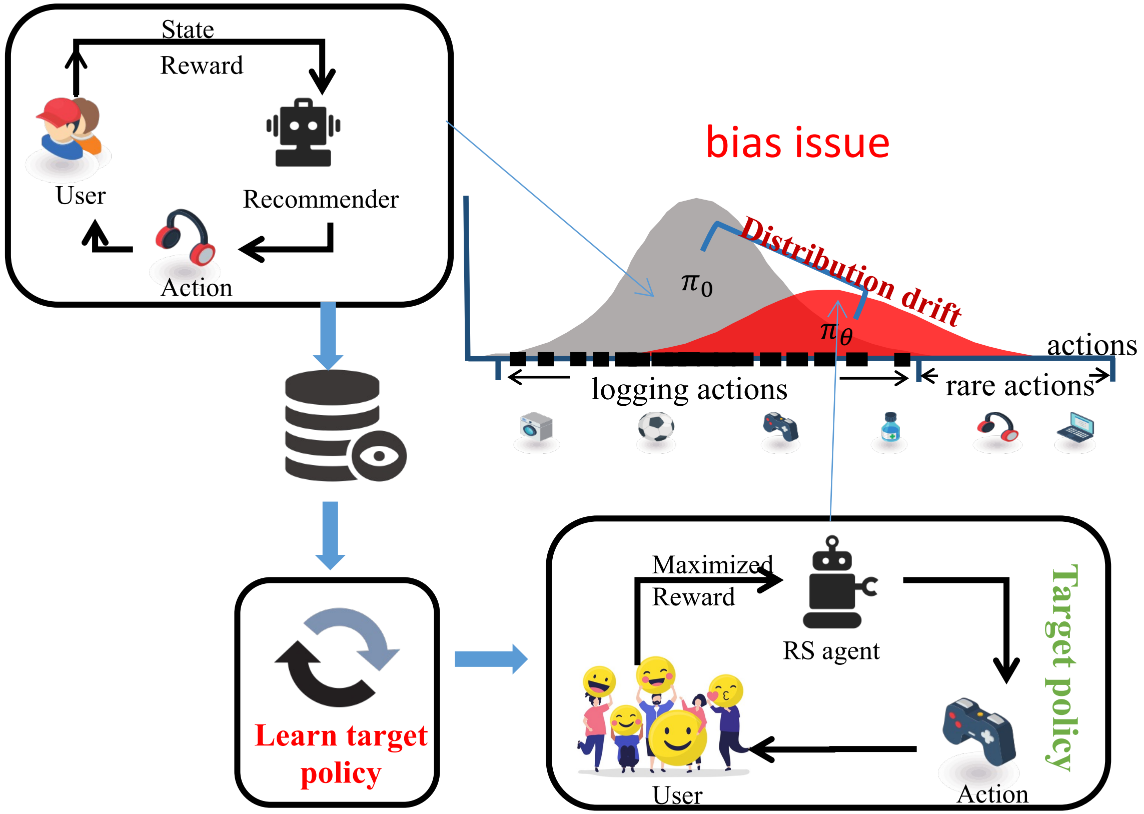
Model framework

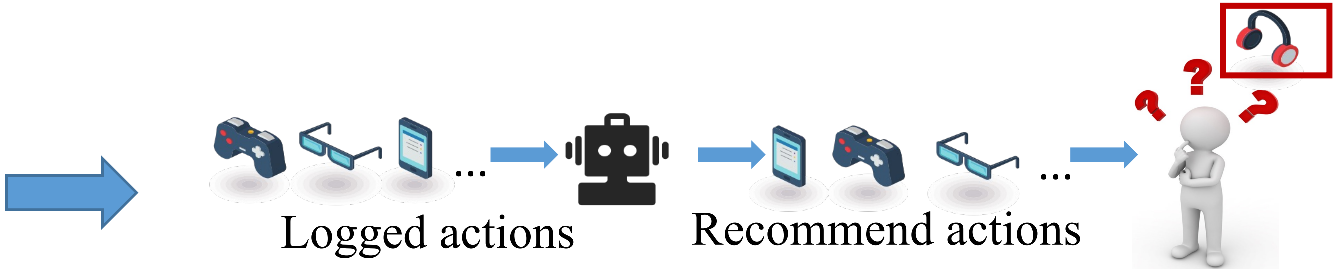
- Quantify social confounders with **Social network confounder model**
- Build the exposure mechanism with **Exposure model**
- Learn balanced representation independent of exposure with **Deconfounder model**
- Using balanced representation for **Rating prediction**

Distribution shift in RL based-Recommendation

Off-policy Learning over Heterogeneous Information for Recommendation

- Off-policy learning suffers the bias issue caused by the policy distribution shift



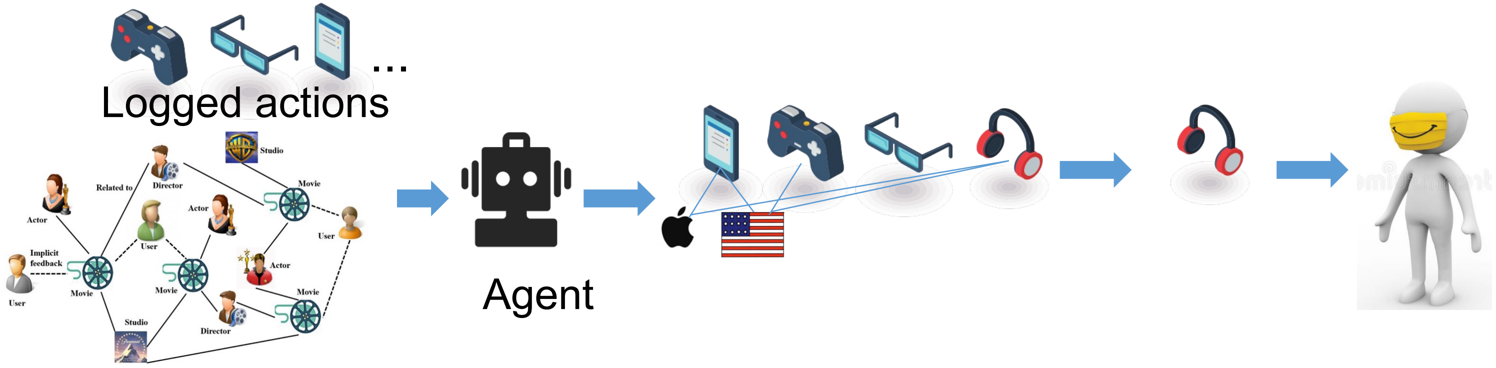


Poor gets poorer phenomenon

Distribution shift in RL based-Recommendation

Off-policy Learning over Heterogeneous Information for Recommendation

- **Real-world context information** could be useful to augment partially observed data and infer users' potential preference

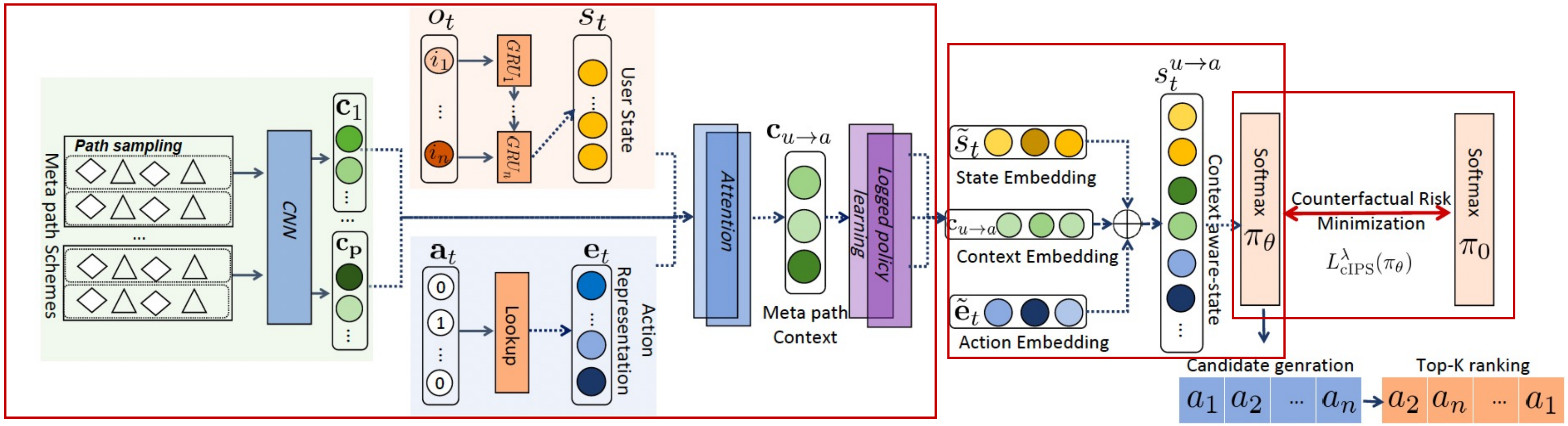


Heterogeneous Information Network

- **Counterfactual Risk Minimization** to answer how much reward would be received if a new policy had been deployed, instead of the original policy

Distribution shift in RL based-Recommendation

Off-policy Learning over Heterogeneous Information for Recommendation



We design three steps for the HIN-enhanced off-policy learning

- **Co-attentive** state, action and context **representation learning**
- **HIN-augmented policy learning** through aggregating context-aware state representation
- **Counterfactual Risk Minimization** to correct the discrepancy between target policy and logging policy

Distribution shift in RL based-Recommendation

HIN-augmented policy learning

- Context-aware state, action representation learning (Attention mechanism):

$$\boldsymbol{\beta}_t^u = Relu\left(\mathbf{W}_u s_t + \mathbf{W}_{u\to a}\mathbf{c}_{u\to a} + \mathbf{b}_u\right)$$
$$\boldsymbol{\beta}_t^a = Relu\left(\mathbf{W}_a \mathbf{e}_t + \mathbf{W}_{u\to a}\mathbf{c}_{u\to a} + \mathbf{b}_a\right)$$

$$\tilde{s}_t = \boldsymbol{\beta}_t^u \odot s_t$$
$$\tilde{\mathbf{e}}_t = \boldsymbol{\beta}_t^a \odot \mathbf{e}_t$$

- Context-aware policy learning:

$$s_t^{u\to a} = \tilde{s}_t \oplus \mathbf{c}_{u\to a} \oplus \tilde{\mathbf{e}}_t \qquad \pi_\theta(a_t \mid s_t^{u\to a}) = \frac{\exp\left(\mathbf{e}_{t+1}^\top s_t^{u\to a}\right)}{\sum_{a_t\in\mathcal{A}_t}\exp\left(\mathbf{e}_t{}^\top s_t^{u\to a}\right)}$$

CRM-based unbiased optimization (cIPS estimator):

$$L_{\text{cIPS}}^{\lambda}(\pi_\theta) = \frac{1}{T}\sum_{t=1}^{T}(r_t - \lambda_t)\min\left\{\frac{\pi_\theta\left(a_t \mid s_t^{u\to a}\right)}{\pi_0\left(a_t \mid s_t^{u\to a}\right)}, c\right\}$$

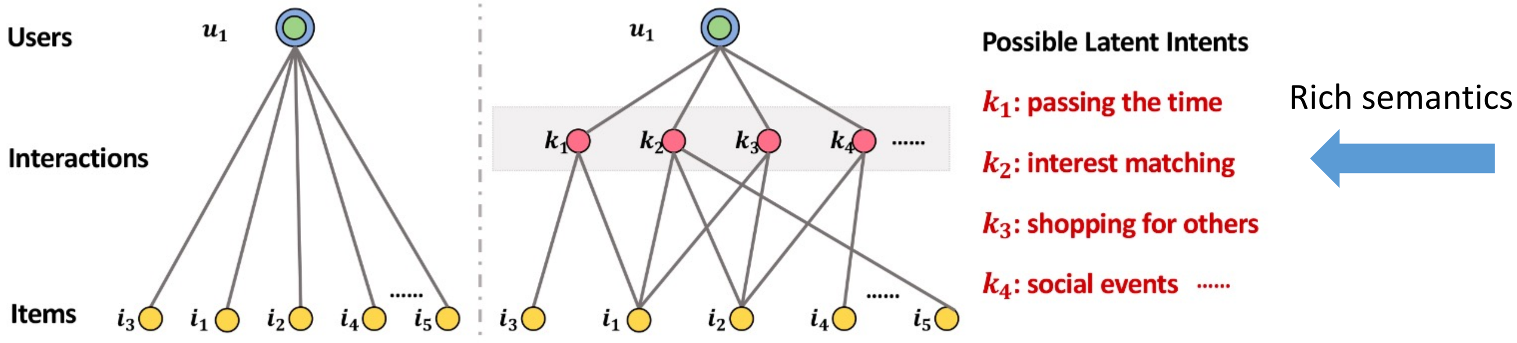
$$\begin{aligned} R(\pi_\theta) &= \mathbb{E}_{\pi_\theta}\left[\gamma^t L_{\text{cIPS}}^{\lambda}(\pi_\theta)\right] \\ &= \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T}\gamma^t\left(r(s_t^{u\to a}, a_t) - \lambda_t\right)\min\left\{\frac{\pi_\theta\left(a_t \mid s_t^{u\to a}\right)}{\pi_0\left(a_t \mid s_t^{u\to a}\right)}, c\right\}\right] \end{aligned}$$

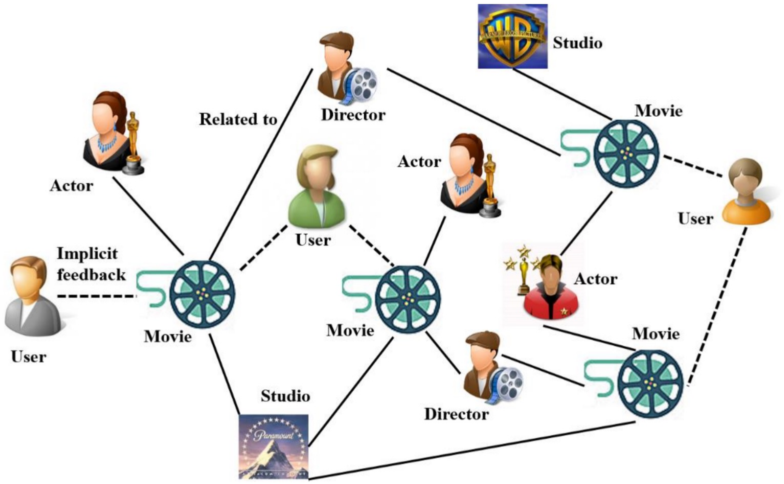
Semantics-Aware Intent Learning

Causal Disentanglement for Semantics-Aware Intent Learning

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

- Disentangle users' true interests
- Explain users’ intents by item semantics (contexual information)





Heterogenous Information Network (HIN)

Semantics-Aware Intent Learning

Causal Disentanglement for Semantics-Aware Intent Learning

IEEE TRANSACTIONS ON
KNOWLEDGE AND DATA ENGINEERING

Challenge

- The complexity in heterogeneous information display skewed distributions, thus can bias the user preference and prediction score

Contribution

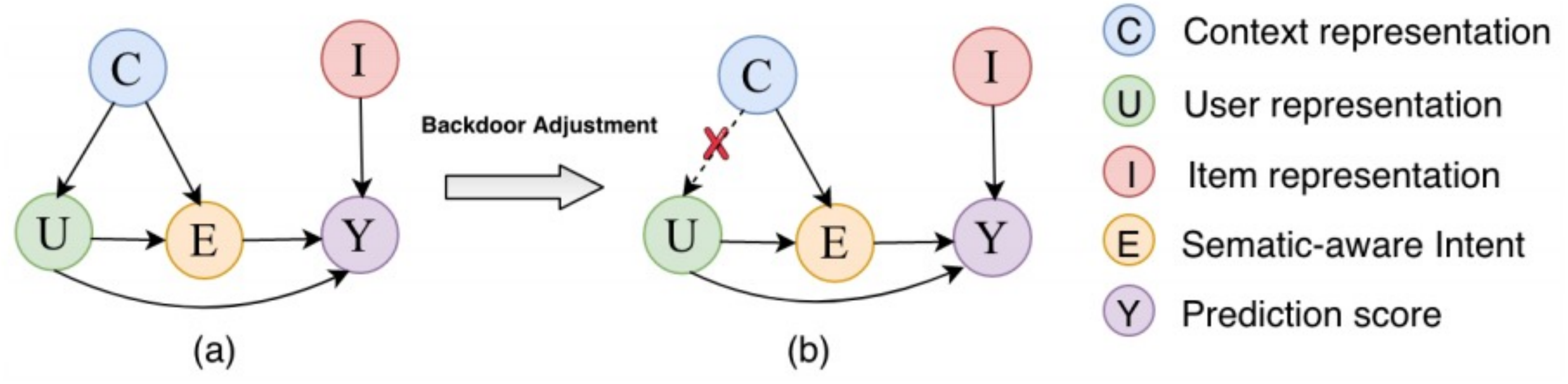
- Provides semantics to user intents (Interpretability)
- Debias bias stemmed from heterogenous information network (Robustness)

Semantics-Aware Intent Learning

Causal Disentanglement for Semantics-Aware Intent Learning

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

The SCM model for disentangling learning



- Context information in C is the confounder since it is the common cause for user information U and E
- Backdoor adjustment can block the path from C to U, thus can remove the confounding bias

Semantics-Aware Intent Learning

Causal Disentanglement for Semantics-Aware Intent Learning

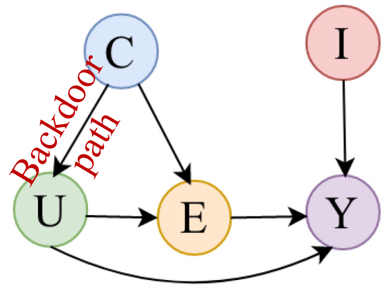
IEEE TRANSACTIONS ON
KNOWLEDGE AND DATA ENGINEERING

Backdoor adjustment

- Backdoor criterion

  Definition. A set of variables W satisfies the backdoor criterion relative to T and Y if :

  1. W blocks all backdoor paths from T to Y
  2. W does not contain any descendants of T



C satisfies Backdoor criterion: C blocks backdoor path from U (treatment) to Y (outcome)

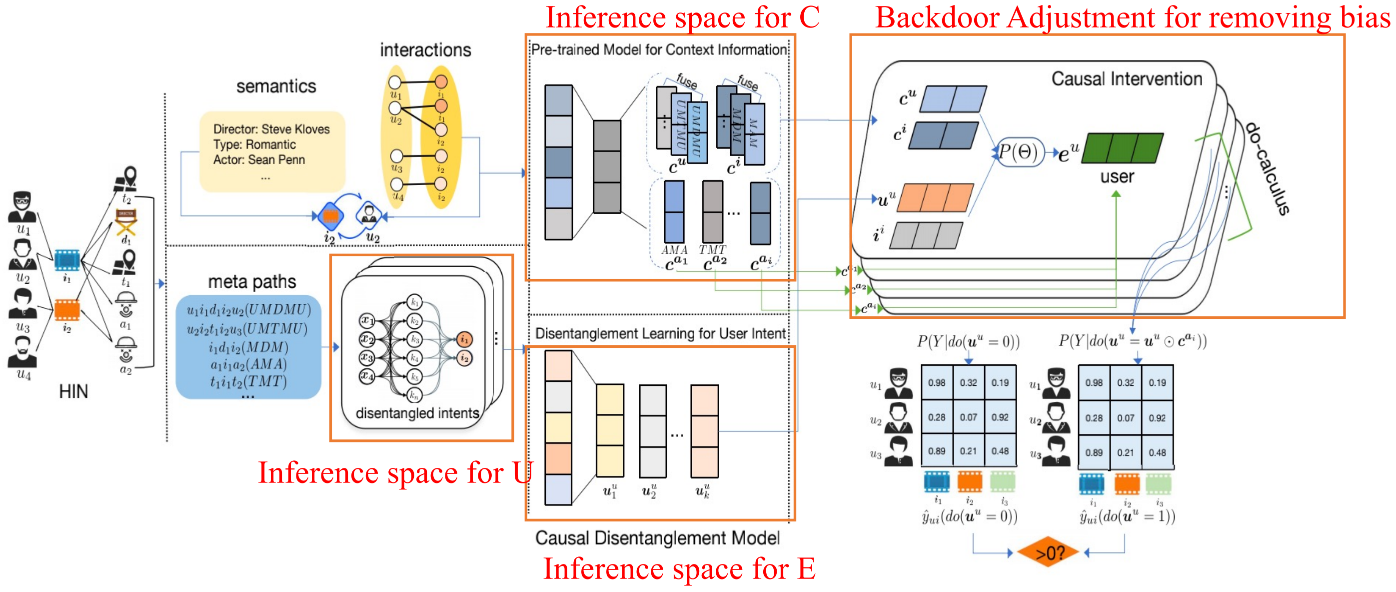
- Backdoor adjustment via do-operator:
  - As C satisfies the backdoor criterion, the do-operator $P(y \mid do(u))$ is the true causality of U on Y, equal to blocking path C$\longrightarrow$U

Semantics-Aware Intent Learning

Causal Disentanglement for Semantics-Aware Intent Learning

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

Framework



We design two steps for the unbiased semantic-aware user intents learning

- **Semantic aware user intents learning:** Learn semantic aware representation E with HIN information
- Fine-tune E with **Causal intervention** for easing the bias stemmed from HIN

Counterfactual Explanation for Recommendation

Reinforced Path Reasoning for Counterfactual Explainable Recommendation

- Bridge the gap of generating item attribute-based counterfactual explanations from Knowledge Graphs (KGs)

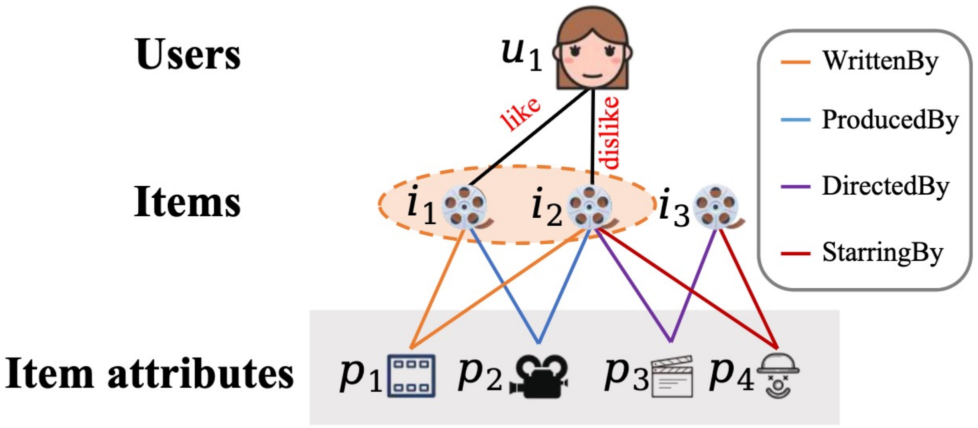


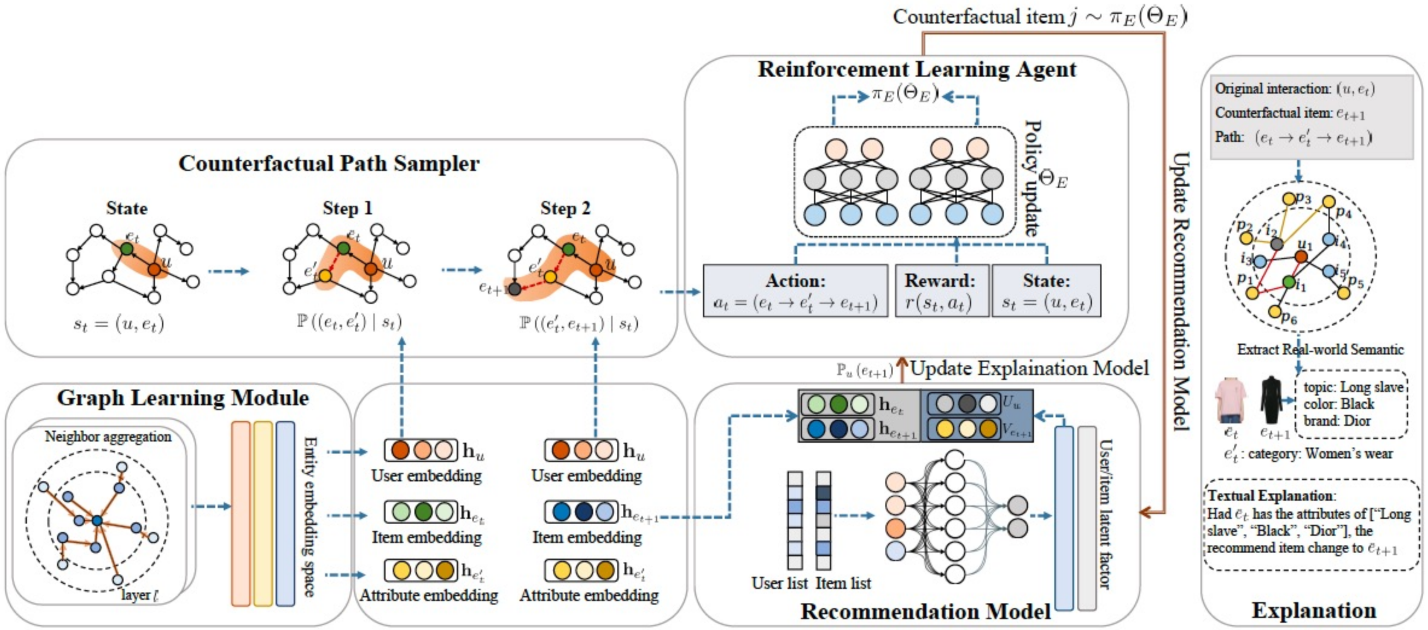
Fig. 1: Toy example of inferring item attribute-based counterfactual explanations from knowledge graphs.

[Item Attribute-based Counterfactual Explanation]
A minimal set of item attributes that, if applied, flip the recommendation decision.

Counterfactual Explanation for Recommendation

Reinforced Path Reasoning for Counterfactual Explainable Recommendation

Model framework



- **Two base models: Graph learning module** and **Recommendation model** ;
- **Counterfactual path sampler** uses entity embeddings to sample paths as actions for reinforcement learning
- **Reinforcement learning agent** learns the explanation policy by optimizing the cumulative counterfactual rewards of deployed actions from the sampler.

Counterfactual Explanation for Fairness

Counterfactual Explanation for Fairness in Recommendation

- Inferring attribute-level counterfactual explanation for fairness.
  - **Why counterfactual explanation:** Existing methods generate fairness explanations by selecting top-n features with the largest values, which may introduce pseudo-explanations (i.e., cannot find minimal explanations)

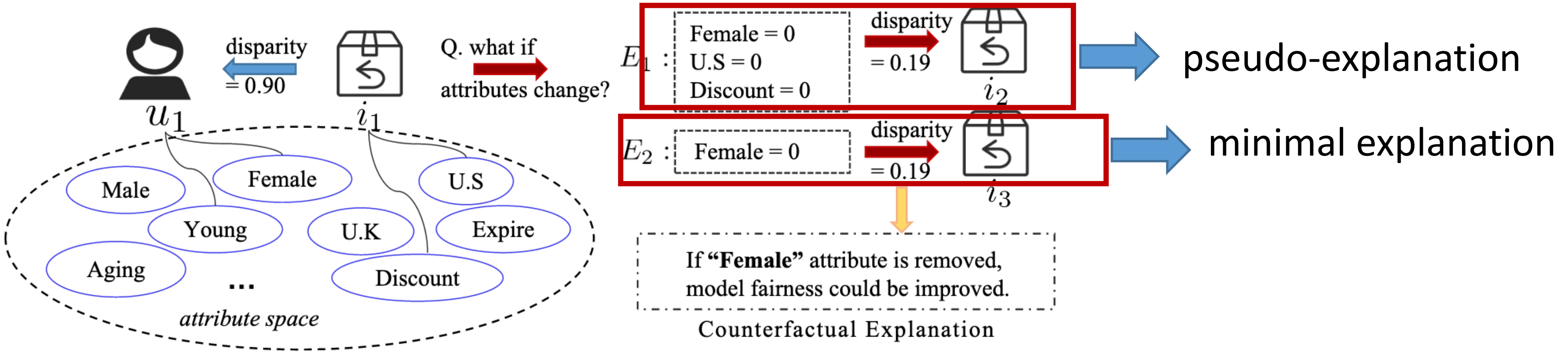


Figure 1: Toy example of inferring attribute-level counterfactual explanation for fairness.

Counterfactual Explanation for Fairness

Counterfactual Explanation for Fairness in Recommendation

Model framework overview

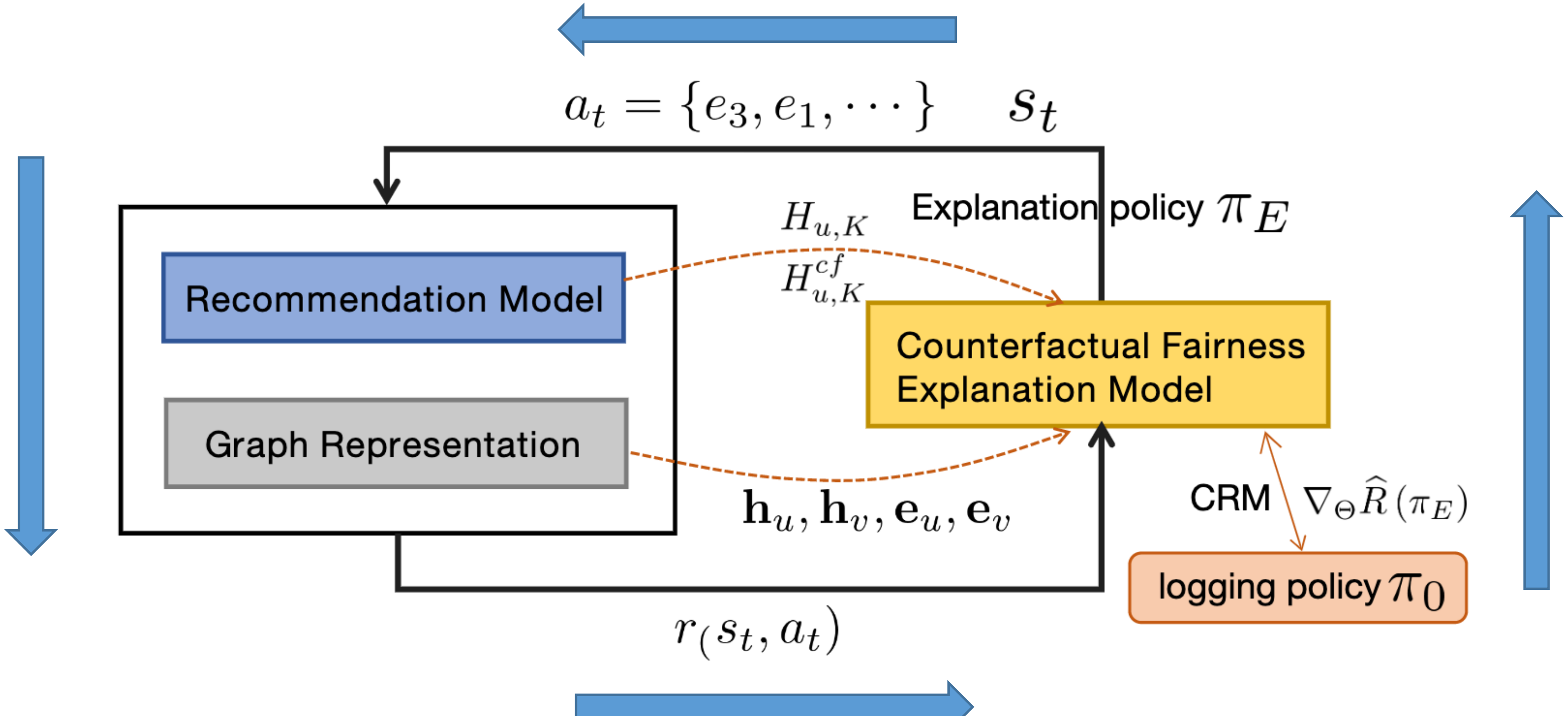
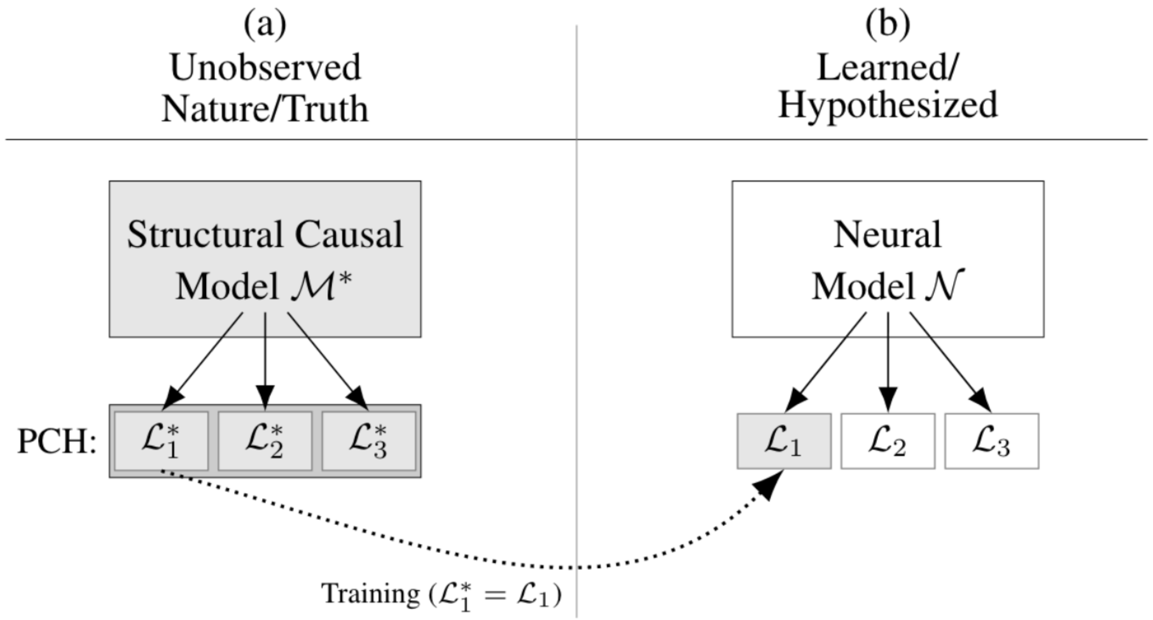


Figure 2: The proposed *CFairER* framework.

Part III: Future work

Causal-Neural Connection for Recommendation

- Future Direction I
  - Causal-Neural connection for enhancing neural networks, e.g., GCN
    - Explicitly model the causality between each of the nodes with the GCN instead of modeling the neighbor correlations
    - Complete Pearl Causal Hierarchy (PCH), i.e., "seeing" (layer 1), "doing" (2), and "imagining" (3) for causal-neural connection expressiveness



Dynamic Bias Mitigation

- Future Direction II
  - Dynamic bias
    - Biases are usually dynamic rather than static
    - Online updating of debiasing strategies