# Trustworthy Recommender Systems

Wenqi Fan[1], Xiangyu Zhao[2], Lin Wang[1], Xiao Chen[1], Jingtong Gao[2], Qidong Liu[2], Shijie Wang[1]

[1]The Hong Kong Polytechnic University

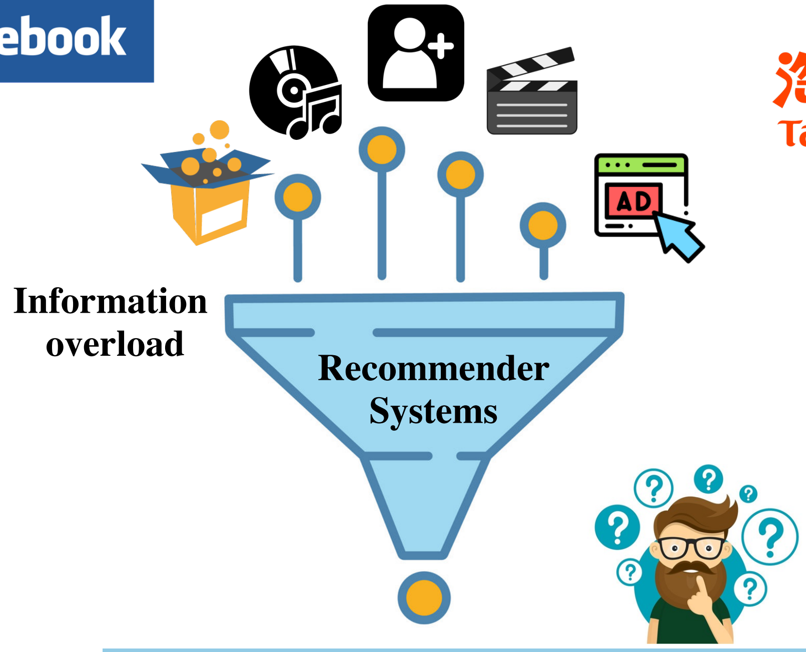
[2]City University of Hong Kong

**Website (Slides)**: https://advanced-recommender-systems.github.io/trustworthiness-tutorial/

Survey: A Comprehensive Survey on Trustworthy Recommender Systems, arXiv:2209.10117, 2022.

# Recommender Systems

**Age of Information Explosion**


**Recommend item X to user**

**Items** can be: Products, Friends, News, Movies, Videos, etc.

# Recommender Systems

**Recommendation has been widely applied in online services:**

- **E-commerce**, Content Sharing, Social Networking ...

amazon

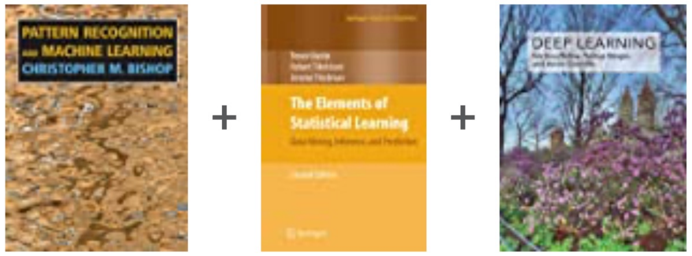
ebay

淘宝网 Taobao.com

JD.COM

**Product Recommendation**

Frequently bought together




A + B + C

Total price: $208.9

Add all three to Cart

Add all three to List

# Recommender Systems

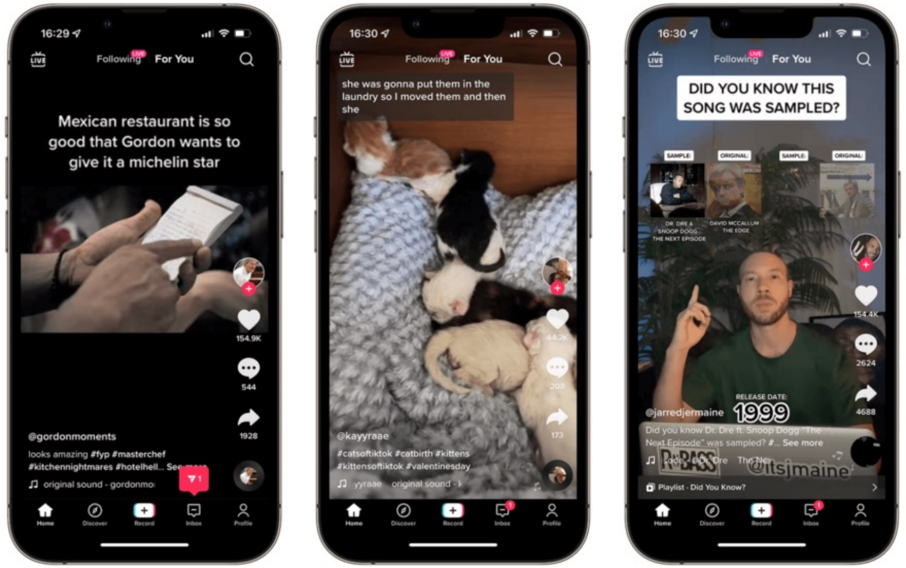
**Recommendation has been widely applied in online services:**

- E-commerce, **Content Sharing**, Social Networking ...

**News/Video/Image Recommendation**

**Top 10 Global Breakthrough Technologies in 2021**

TikTok's recommendation algorithm

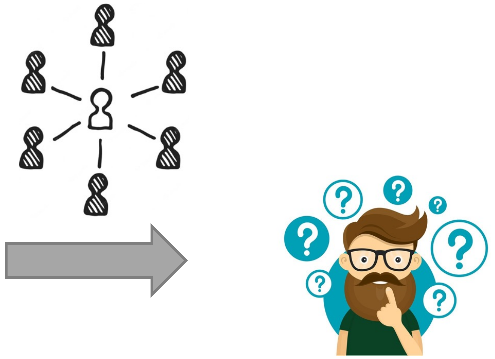


# Recommender Systems

**Recommendation has been widely applied in online services:**

- E-commerce, Content Sharing, **Social Networking** ...

facebook 新浪微博 weibo.com LinkedIn



## Social Recommendations

WeChat

Like what you read?
Share with your friends!

**Subscriptions (訂閱號信息)**

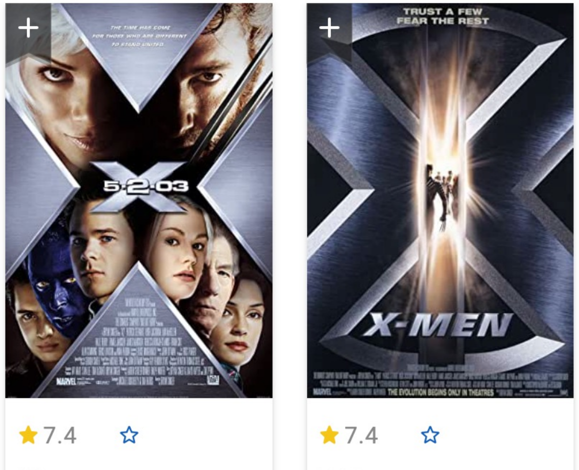
**Read by 9 friends**


**Top Stories（看一看）**
**Wow (朋友在看)**

# Recommender System is Everywhere

Business

Healthcare




Entertainment

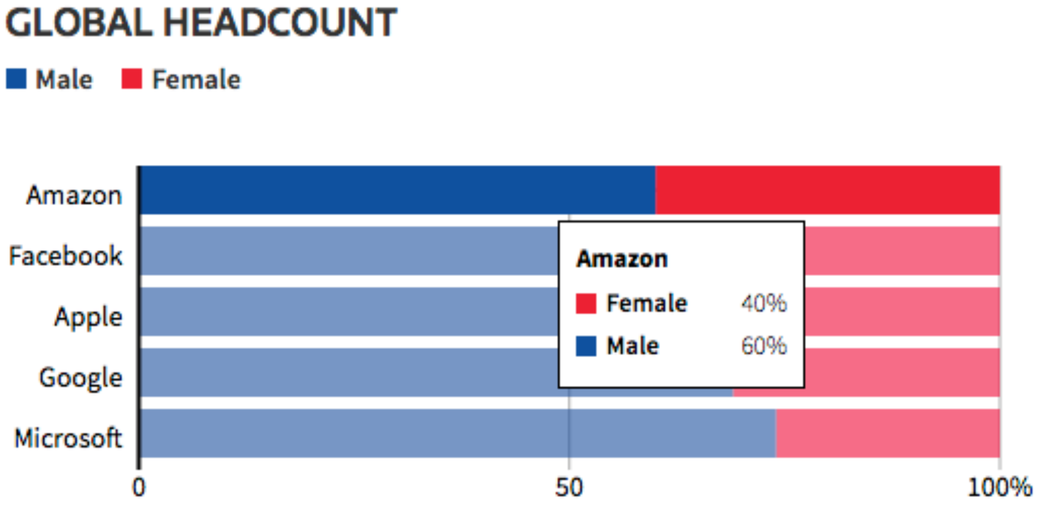
Education

# The Good and The Bad

**The Good** ⟷ **<u>The Bad</u>**

# Discrimination & Fairness Issue

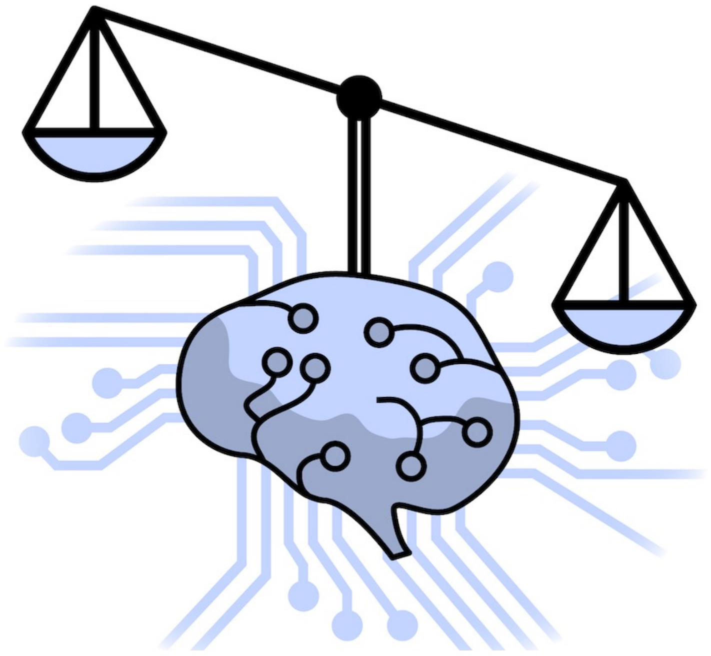
Job recommendation
(Lambrecht et al., 2019)



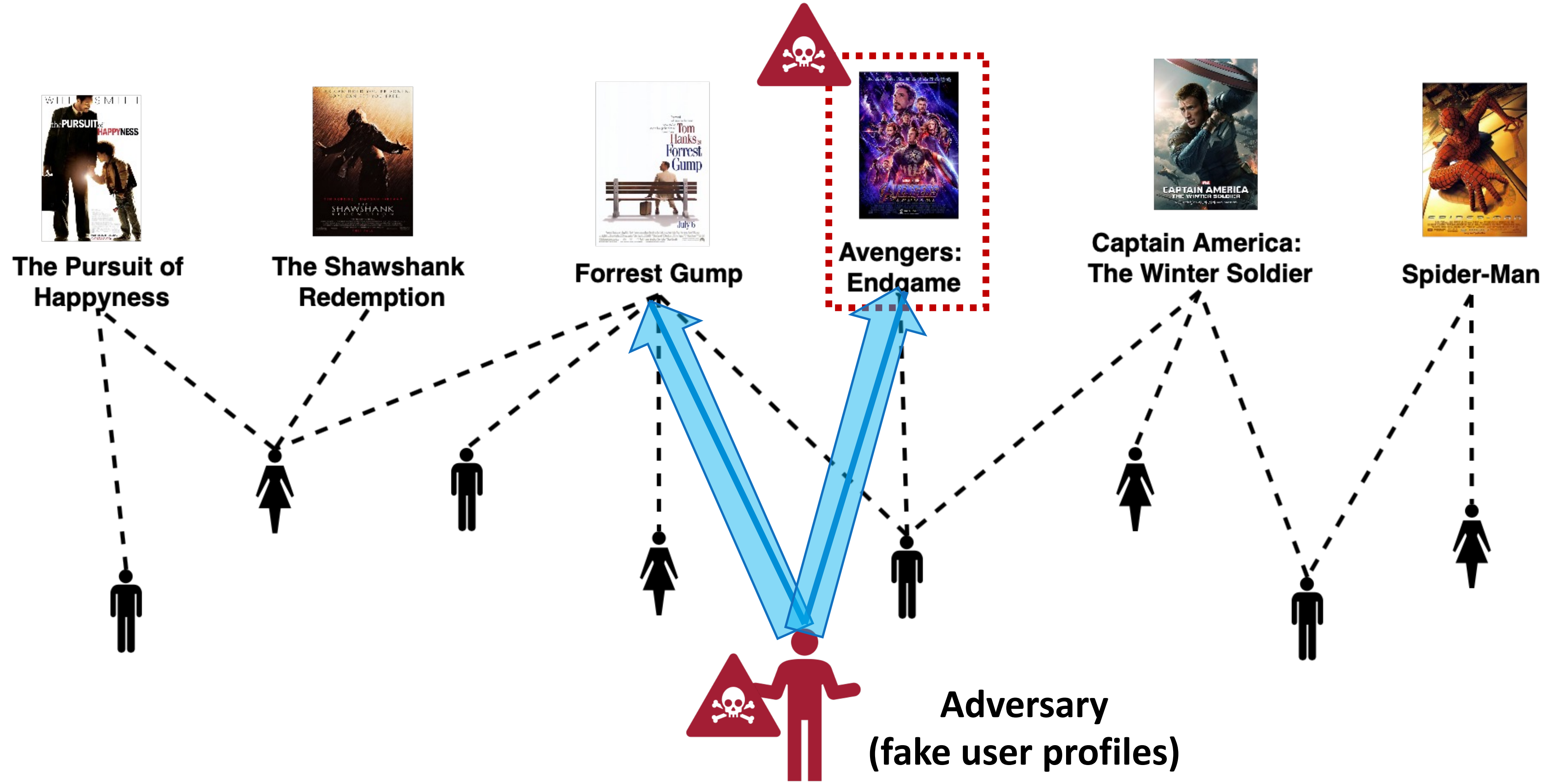

Lambrecht, et al. "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads." *2019*.
Bias and Debias in Recommender System: A Survey and Future Directions, 2021.

# Non-discrimination & Fairness

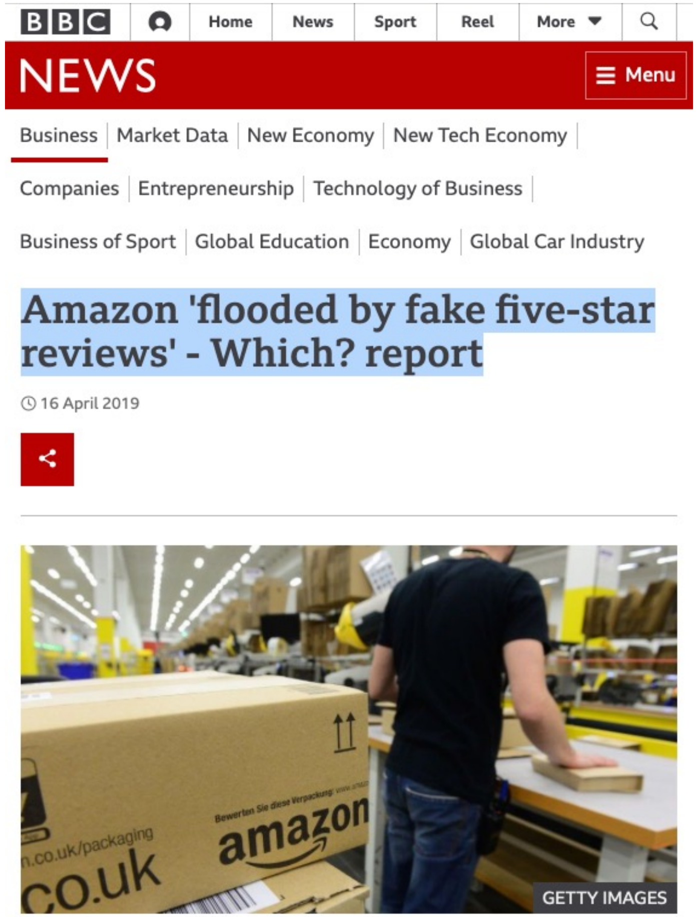
- A recommender system should avoid discriminatory behaviors in human-machine interaction.
- A recommender system should ensure fairness in decision-making.



# Safety & Robustness Issue



# Attacks can happen in Recommender Systems


BBC | Home | News | Sport | Reel | More

NEWS ☰ Menu

Business | Market Data | New Economy | New Tech Economy |

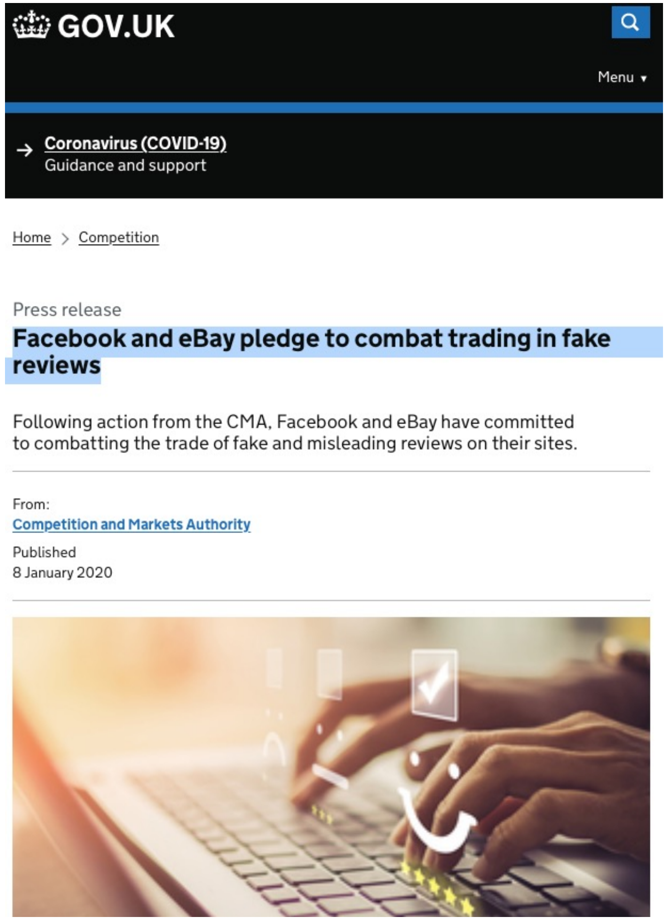
Companies | Entrepreneurship | Technology of Business |

Business of Sport | Global Education | Economy | Global Car Industry

**Amazon 'flooded by fake five-star reviews' - Which? report**

16 April 2019




GOV.UK

Menu

Coronavirus (COVID-19)
Guidance and support

Home > Competition

Press release

**Facebook and eBay pledge to combat trading in fake reviews**

Following action from the CMA, Facebook and eBay have committed to combatting the trade of fake and misleading reviews on their sites.

From:
Competition and Markets Authority

Published
8 January 2020

**"More than three-quarters of people are influenced by reviews when they shop online."**

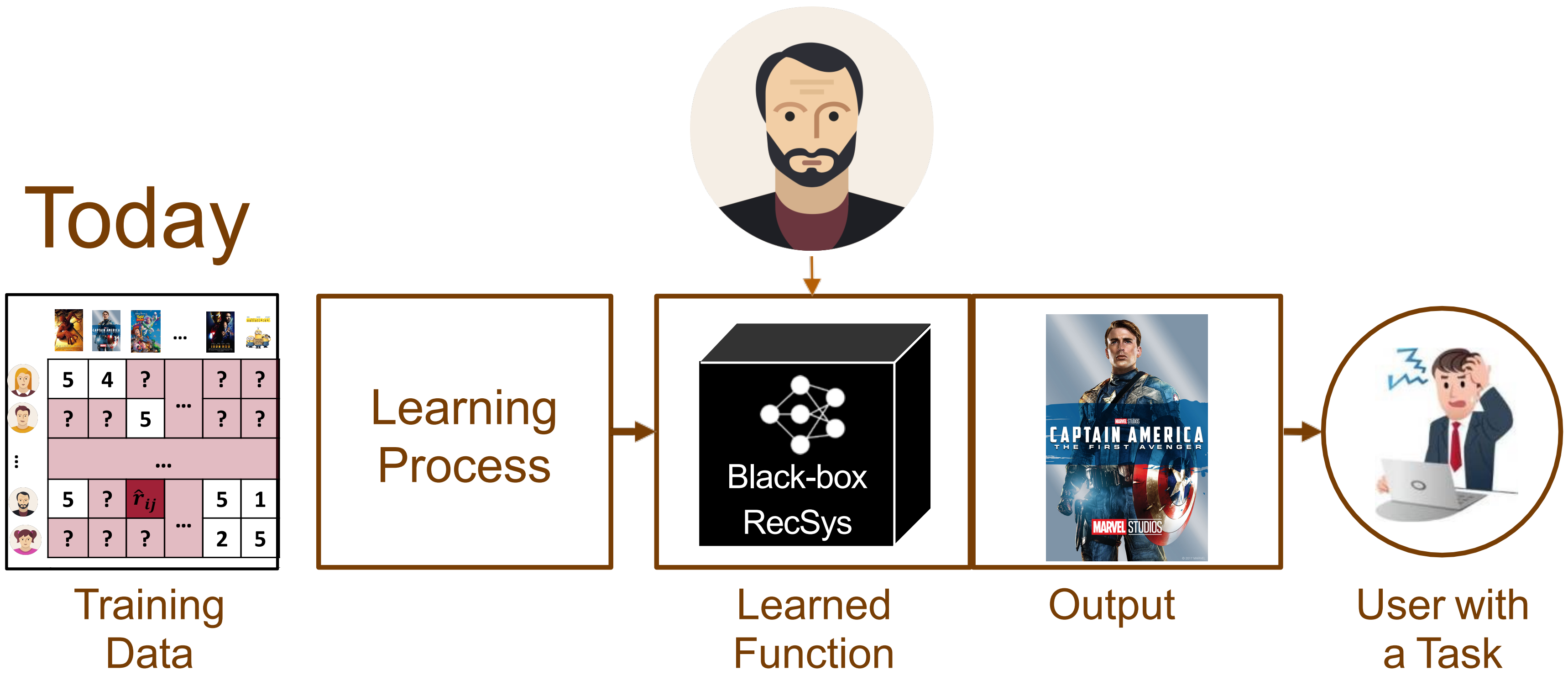
**Understand system's vulnerability and how attacks can be performed**

**Defend against potential adversarial attacks**

"The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry", Information Systems Research, 2016
https://www.bbc.com/news/business-47941181
https://www.gov.uk/government/news/facebook-and-ebay-pledge-to-combat-trading-in-fake-reviews

# Black-box Issue

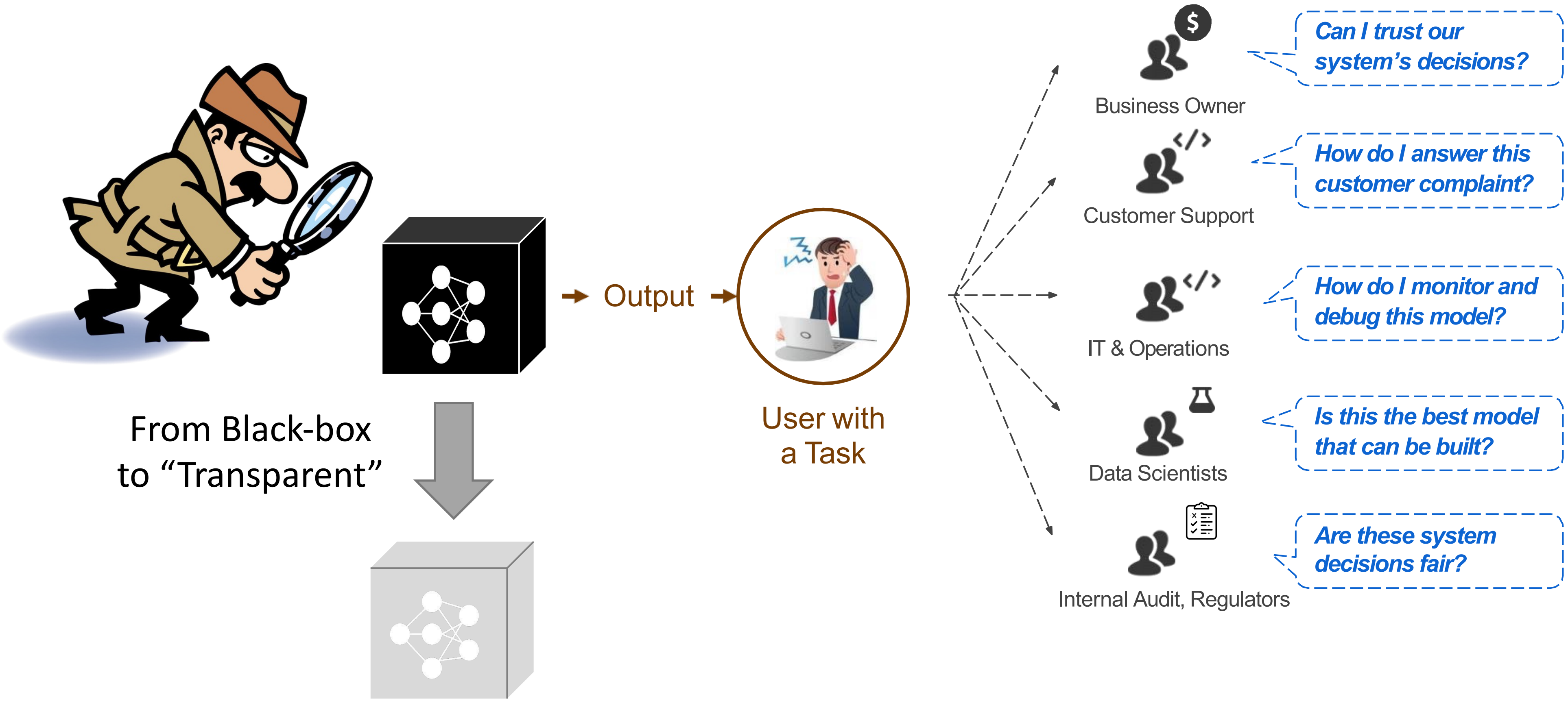
How recommender systems work?




- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

# Explainability

## Black-box system creates confusion and doubt




The Need for Explainable Recommendation

Yongfeng Zhang, et.al, Explainable Recommendation: A Survey and New Perspectives, 2020.

# Privacy Issue


- The success of recommender systems heavily relies on data that might contain private and sensitive information.
- Can we still take the advantages of data while effectively protecting the privacy?

# Environmental Issue


**GPU Power Consumption Comparison**

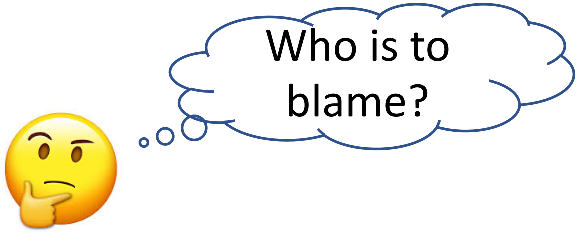
| Dataset | XDL | DLRM | FAE |
|---|---|---|---|
| Criteo Kaggle | 61.83W | 58.91W | 55.81W |
| Alibaba | 56.39W | 60.21W | 56.62W |
| Criteo Terabyte | 59.71W | 62.47W | 57.03W |
| Avazu | 60.2W | 58.03W | 56.4W |

Estimated carbon emissions from training common recommendation models

Accelerating recommendation system training by leveraging popular choices, VLDB, 2021.

# Auditability & Accountability

**Violent movie**




A clear responsibility distribution, which focuses on who should take the responsibility for what impact of recommender systems.

# Auditability & Accountability

- Five roles in Recommender Systems

System Designers

System Deployers

Decision Makers

End Users

System Auditors

It is necessary to determine the roles and the corresponding responsibility of different parties in the function of a recommender system.

# Interactions Among Different Dimensions


**How do these SIX dimensions influence each other?**

There exist both **accordance** and the **conflicts** among the six dimensions.

# Trustworthy Recommender Systems


**"A Comprehensive Survey on Trustworthy Recommender Systems", arXiv:2209.10117, 2022.**

# A Survey on The Computational Perspective

## A Comprehensive Survey on Trustworthy Recommender Systems

WENQI FAN, The Hong Kong Polytechnic University, Hong Kong
XIANGYU ZHAO*, City University of Hong Kong, Hong Kong
XIAO CHEN, The Hong Kong Polytechnic University, Hong Kong
JINGRAN SU, The Hong Kong Polytechnic University, Hong Kong
JINGTONG GAO, City University of Hong Kong, Hong Kong
LIN WANG, The Hong Kong Polytechnic University, Hong Kong
QIDONG LIU, City University of Hong Kong, Hong Kong
YIQI WANG, Michigan State University, USA
HAN XU, Michigan State University, USA
LEI CHEN, The Hong Kong University of Science and Technology, Hong Kong
QING LI, The Hong Kong Polytechnic University, Hong Kong


https://arxiv.org/abs/2209.10117

**WWW'2023**
**Tutorial**
**Website (Slides)**

https://advanced-recommender-systems.github.io/trustworthiness-tutorial/

# Trustworthy Recommender Systems

**Introduction**

Wenqi Fan

**Non-discrimination & Fairness**

Xiao Chen

**Safety & Robustness**

Shijie Wang

**Explainability**

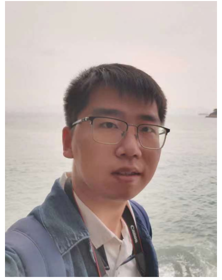
Jingtong Gao

**Privacy**

Lin Wang

**Environmental Well-being**

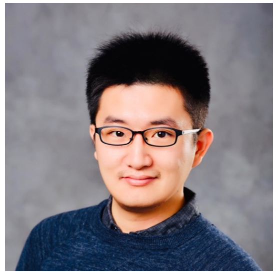
**Accountability & Auditability**



Qidong Liu

**Dimension Interactions**

**Future Directions**



Xiangyu Zhao

# Trustworthy Recommender Systems

**Introduction**

Wenqi Fan

**Non-discrimination & Fairness**

Xiao Chen

**Safety & Robustness**

Shijie Wang

**Explainability**

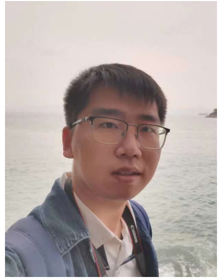
Jingtong Gao

**Privacy**

Lin Wang

**Environmental Well-being**

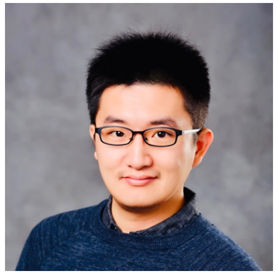
**Accountability & Auditability**



Qidong Liu

**Dimension Interactions**

**Future Directions**



Xiangyu Zhao

# Contents

**CONCEPTS AND TAXONOMY**

METHODOLOGY

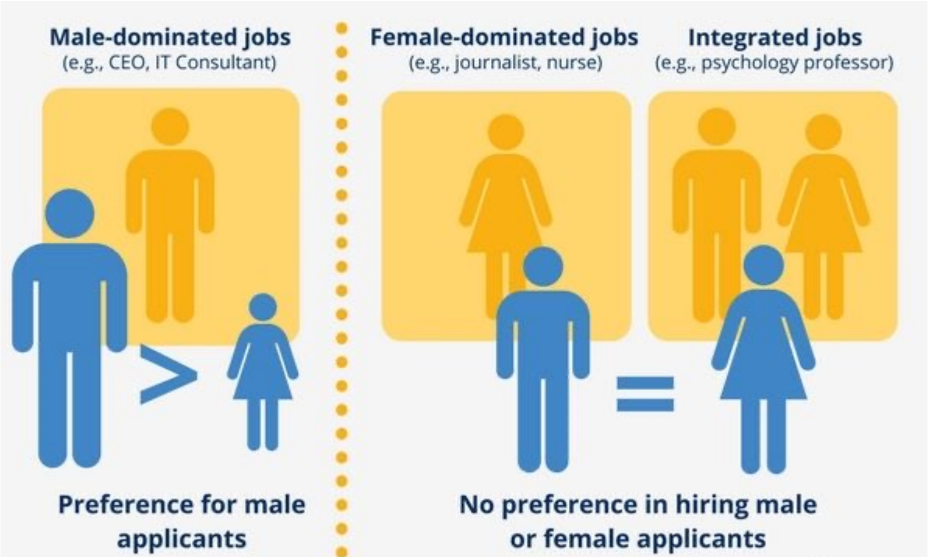
APPLICATIONS

SURVEYS AND TOOLS

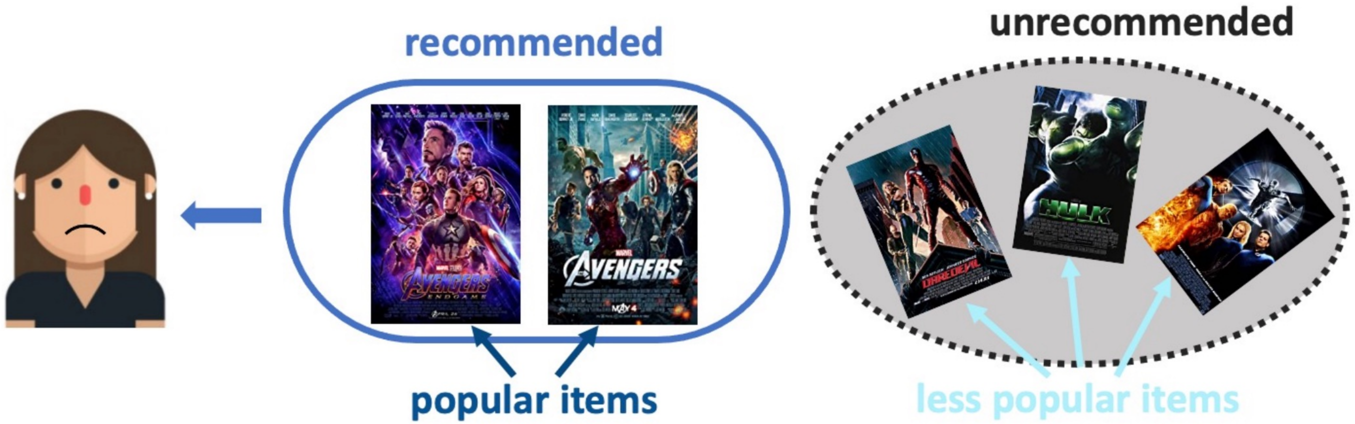
FUTURE DIRECTIONS

# Potential discrimination and bias in RecSys

- Recommender Systems make unfair decisions for specific user/item groups




**Gender Discriminatory Bias [1]**

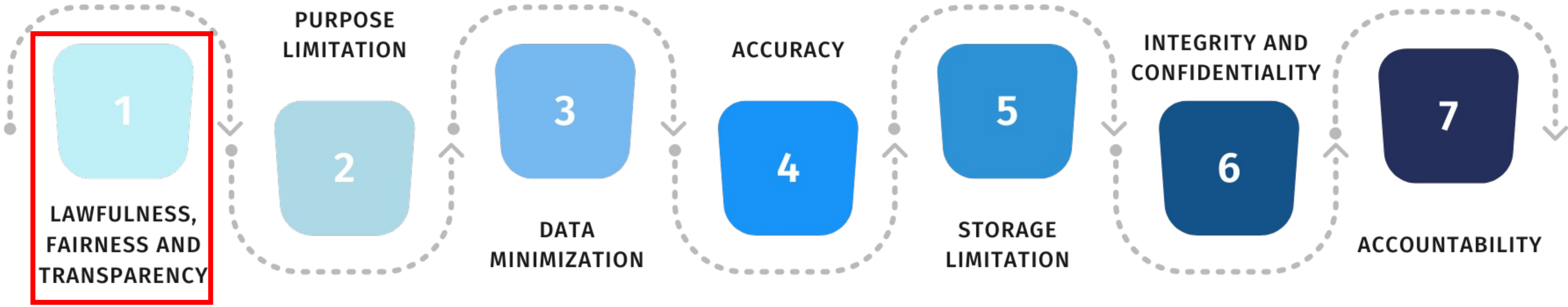



**Popularity Bias [2]**

[1] Lambrecht, et al. "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads." 2019.
[2] Abdollahpouri, et al. "Popularity bias in ranking and recommendation." 2019.

# Why Need Fairness in RecSys: From the Ethics Perspective

- 7 principles of EU GDPR regulation



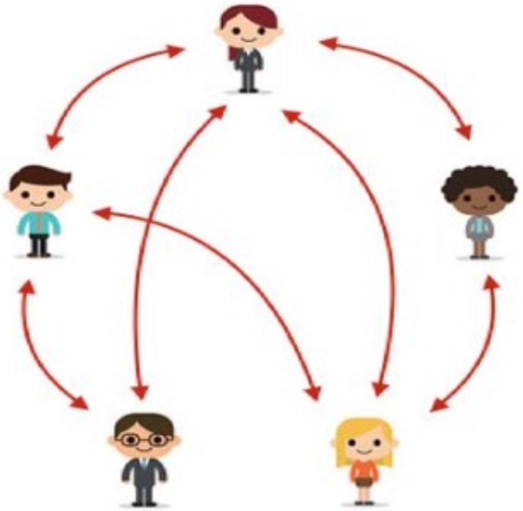

Fairness often couples with other responsible AI perspectives (e.g., explainability).

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/lawfulness-fairness-and-transparency

# Why Need Fairness in RecSys: From the Utility Perspective

- Fair exposure opportunity guarantees the sustainable development of the RecSys platform

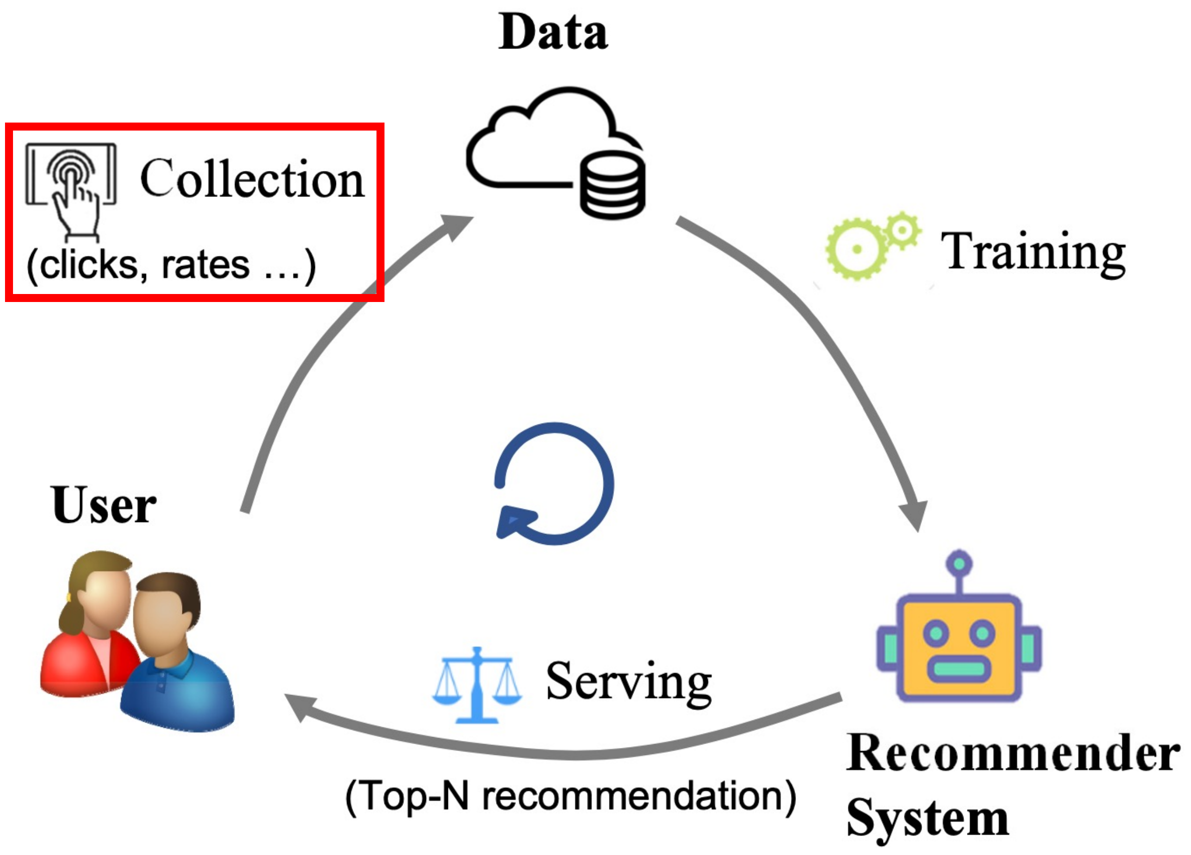

Big retailors vs. Small retailors
in the e-commerce system



Star accounts vs. Grassroot accounts
in the social recommendation system

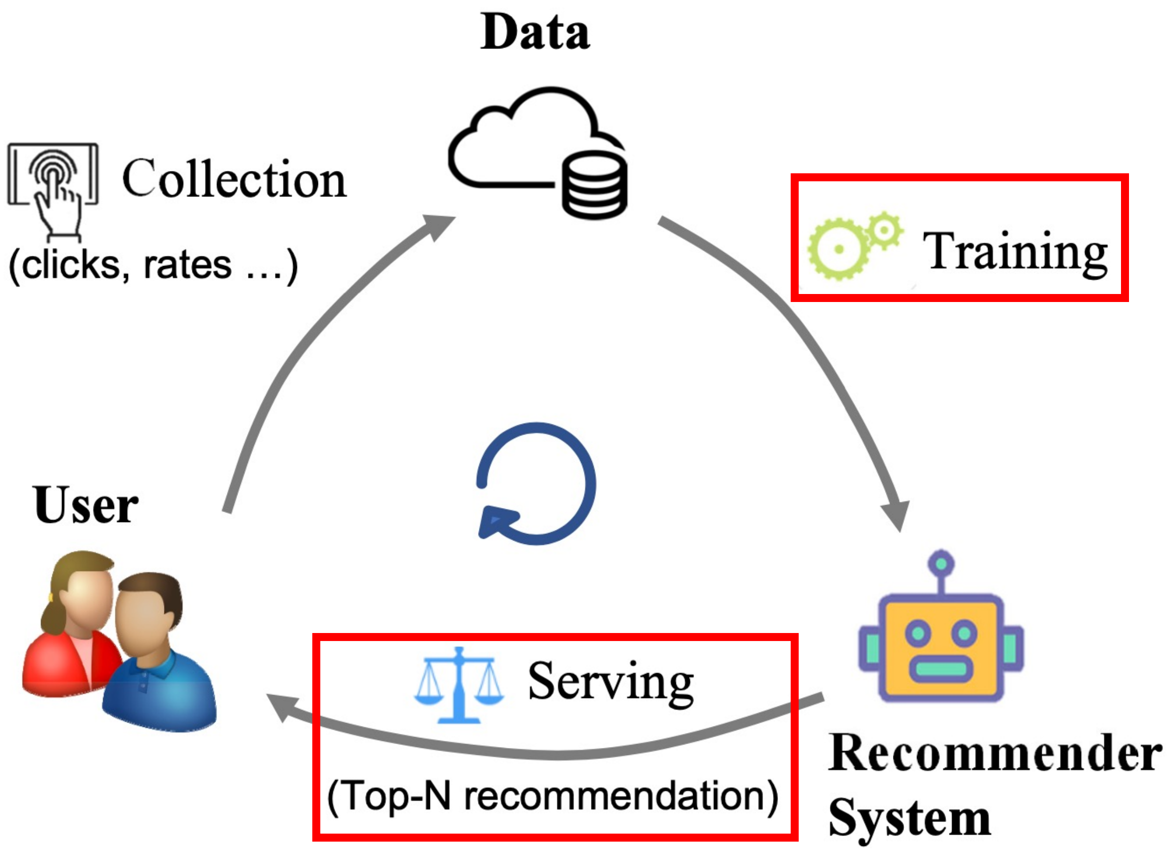
# Sources of Bias

- **Data bias**
  - **Selection Bias:**
    selecting rating behavior of users
  - **Exposure Bias:**
    unobserved interactions may not fully represent the disliked items of users
  - **Conformity Bias:**
    users behave similarly to other group members
  - **Position Bias:**
    the higher positions on a recommendation list tends to receive more interaction



# Sources of Bias

- Data bias
  - Selection Bias
  - Exposure Bias
  - Conformity Bias
  - Position Bias
- **Model and result bias**
  - **Popularity Bias:**
    popular items are over-recommended compared to what their popularity warrant



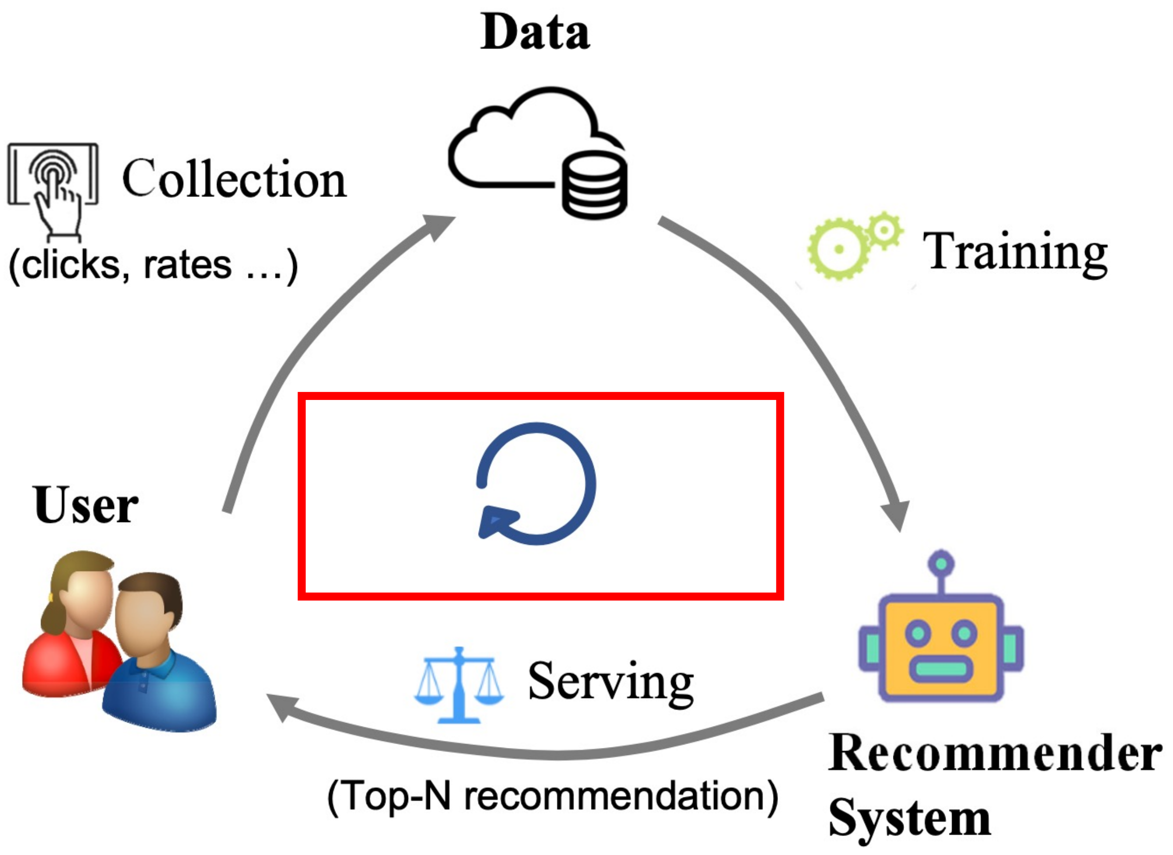

Chen, et al. "Bias and debias in recommender system: A survey and future directions." *TOIS* 2023.

# Sources of Bias

- Data bias
  - Selection Bias
  - Exposure Bias
  - Conformity Bias
  - Position Bias
- Model and result bias
  - Popularity Bias
- **Feedback loop bias**
  - **Reinforced RS Feedback Loop Bias:**

    Unfair recommendations would influence users' behaviors in the online serving process

    Biased user behavior data enlarges model discrimination



# Fairness Definition

- **Procedural Fairness:** procedural justice in decision-making processes

- **Outcome Fairness:** fair outcome performance

User Fairness vs. Item Fairness

Group Fairness vs. Individual Fairness

Causal Fairness vs. Associative Fairness

Static Fairness vs. Dynamic Fairness

# Fairness Evaluation Metrics

- **Absolute Difference (AD):** group-wise utility difference

$$AD = |u(G_0) - u(G_1)|$$

- **Variance:** performance dispersion at the group/individual-level

$$\text{Variance} = \frac{1}{|\mathcal{V}|^2} \sum_{v_i \neq v_j} \left(u(v_i) - u(v_j)\right)^2$$

- **Min-Max Difference:** the difference between the maximum and the minimum score value of all allocated utilities
- **Entropy**
- **KL-Divergence …**

# Contents

CONCEPTS AND TAXONOMY

**METHODOLOGY**

APPLICATIONS

SURVEYS AND TOOLS

FUTURE DIRECTIONS

# Method category

| Pre-processing | In-processing | Post-processing |
|---|---|---|
| Transform the data to remove the data bias before training | Modify the learning algorithms to remove discrimination during the model training process | Perform post-processing by evaluating a holdout set that was not involved during model training |

# Pre-processing methods

- **Resampling**

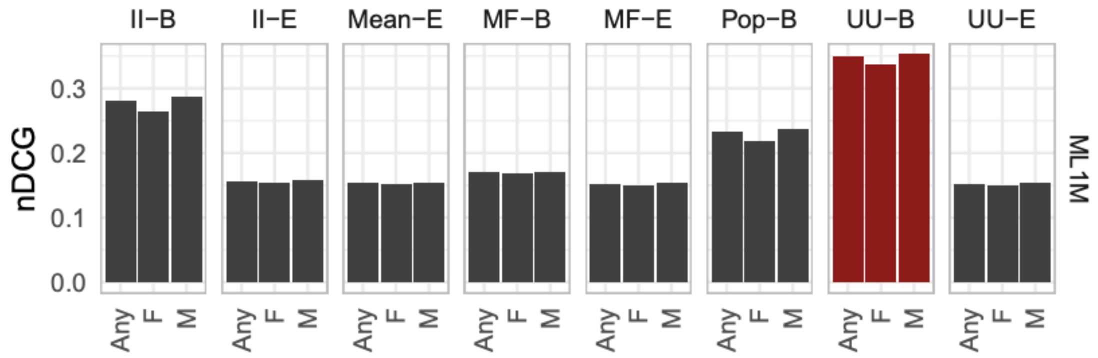
  Rebalance the dataset distribution w.r.t the sensitive attribute

- **Data Augmentation**

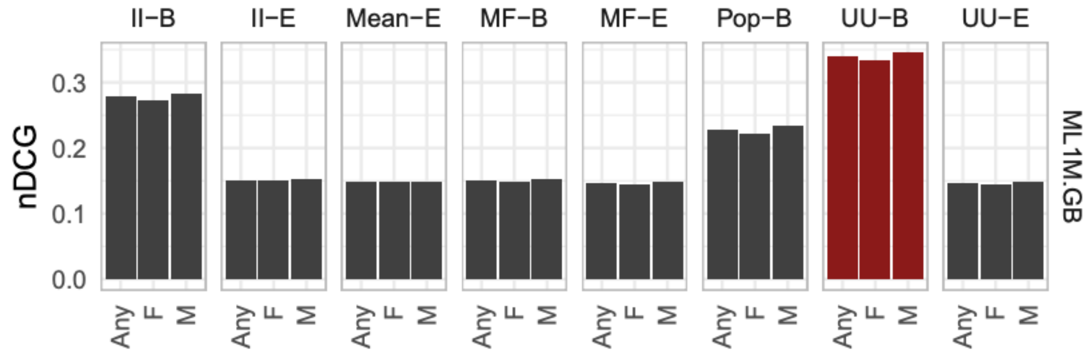
  Generating additional data for promoting the fairness of recommender systems

# Pre-processing method (Resampling)

**Idea:** Different demographic groups obtain different utilities due to imbalanced data distribution. Balance the ratio of various user groups via a re-sampling strategy.




statistically-significant differences between gender groups

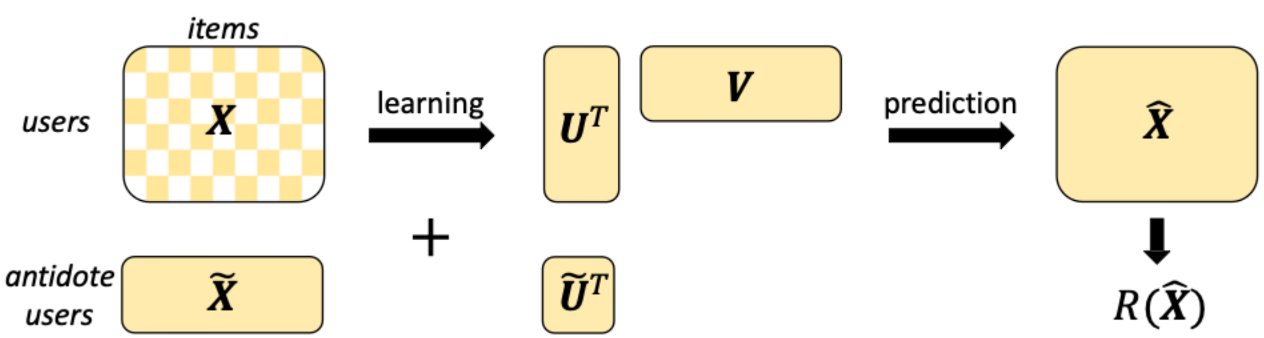


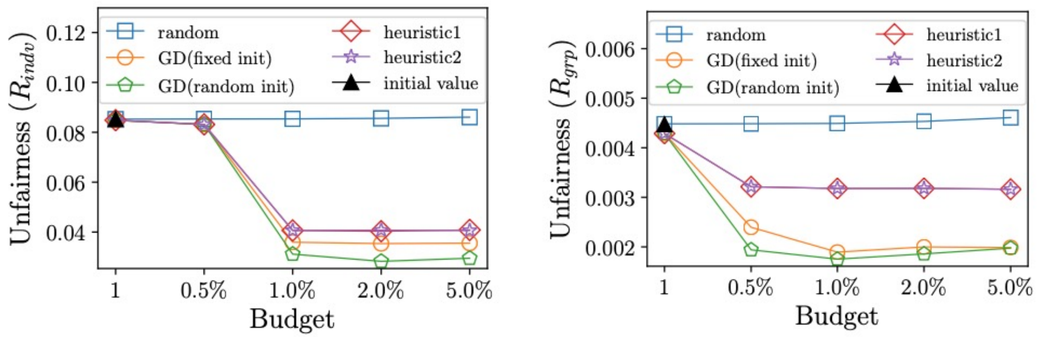

results on gender-balanced dataset

All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. ICFAT 2018.

# Pre-processing method (Adding Antidote Data)

**Idea:** Improving the social desirability of recommender system outputs by adding more "antidote" data to the input.




**(a) Individual fairness** **(b) Group fairness**

**Matrix Factorization:** $$\arg\min_{\mathbf{U},\mathbf{V}} \; ||P_\Omega(\mathbf{X}-\mathbf{U}^\intercal\mathbf{V})||_F^2 + \lambda(||\mathbf{U}||_F^2 + ||\mathbf{V}||_F^2)$$

**Objectives:** $$\arg\min_{\tilde{\mathbf{X}}\in\mathbb{M}} \; R(\hat{\mathbf{X}}(\Theta(\mathbf{X};\tilde{\mathbf{X}})))$$

*fairness objective* *antidote data*

Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems. WSDM 19

# Summary of Pre-processing methods

Flexibility, decoupled with the recommender systems

Performance gains might be degraded by the following steps

# In-processing method

- **Regularization and constrained optimization**
- **Adversary Learning**
- **Causal graph**
- **Reinforcement Learning**
- **Others**

# In-processing method (Regularization)

**Idea:** propose four new metrics that address different forms of unfairness. These metrics can be optimized by adding fairness terms to the learning objective [1].

$$U_{abs} = \frac{1}{n}\sum_{i=1}^{n}\Big|\,|E_{adv}[y]_i - E_{adv}[r]_i| - |E_{\neg adv}[y]_i - E_{\neg adv}[r]_i|\,\Big|,$$

$$\min_{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{u},\boldsymbol{v}} \; J(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{u},\boldsymbol{v}) + U \;.$$

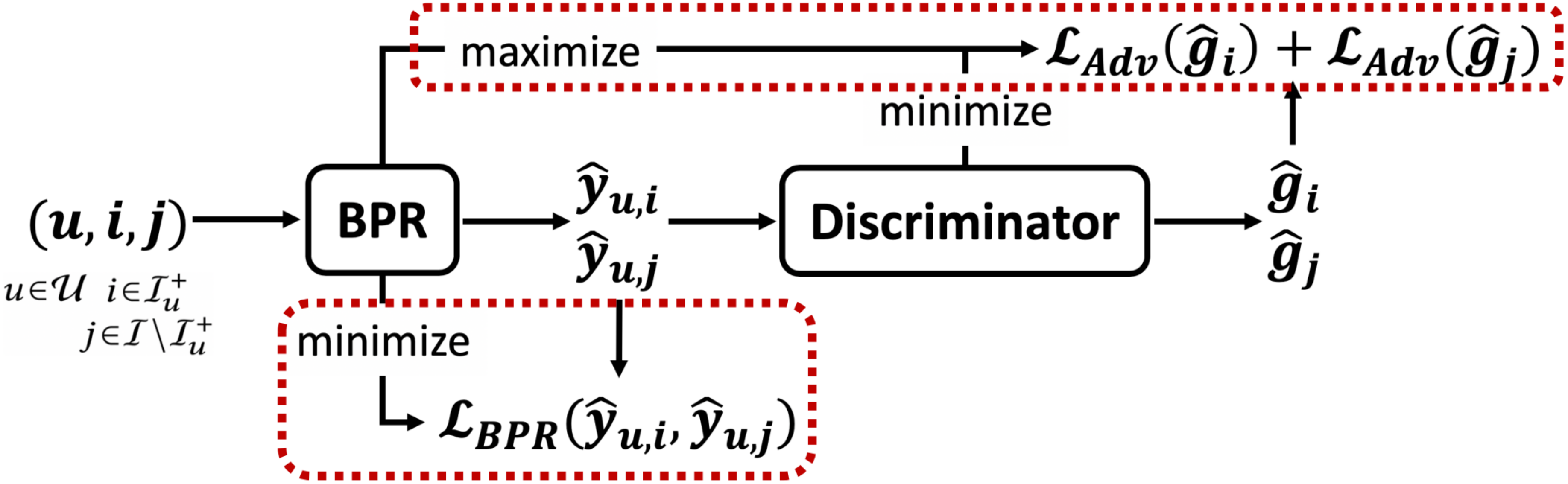
**Idea:** a novel pairwise regularizer for pairwise ranking fairness [2].

$$\min_{\theta}\left(\sum_{(\mathbf{q},j,y,z)\in\mathcal{D}} \mathcal{L}_{rec}\left(f_\theta\left(\mathbf{q},\mathbf{v}_j\right),(y,z)\right)\right) + |\mathrm{Corr}_{\mathcal{P}}(A,B)|,$$

[1] Beyond Parity: Fairness Objectives for Collaborative Filtering. NeurIPS17
[2] Fairness in recommendation ranking through pairwise comparisons. KDD19

# In-processing method (Adversary Learning)

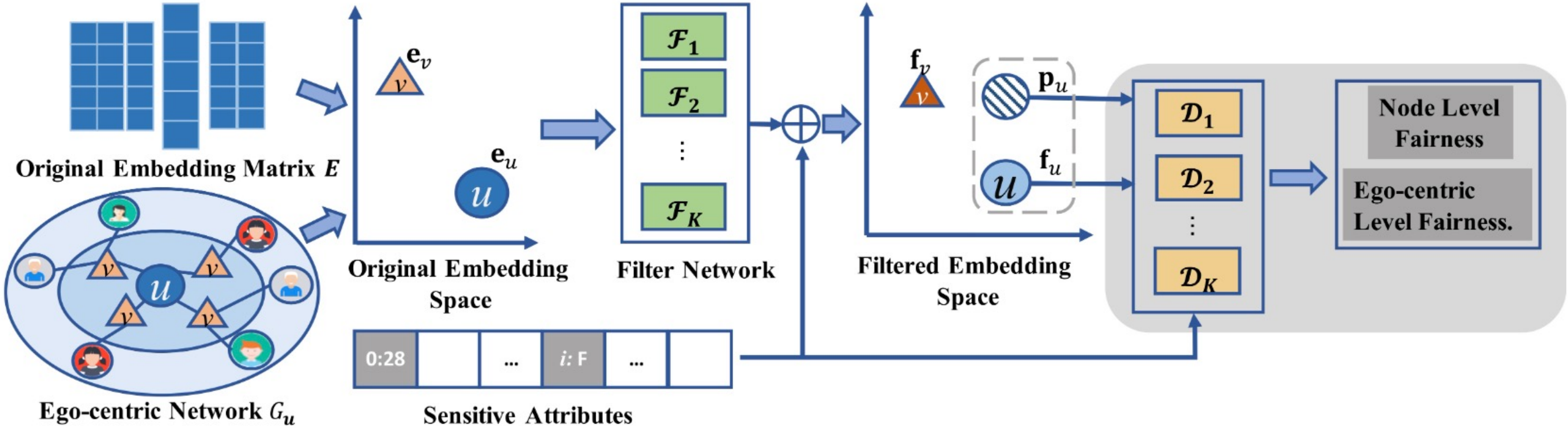
**Idea:** decouple the predicted score with the group attribute.
normalize the score distribution for each user to align predicted score with ranking position.




Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. SIGIR20

# In-processing method (Adversary Learning)

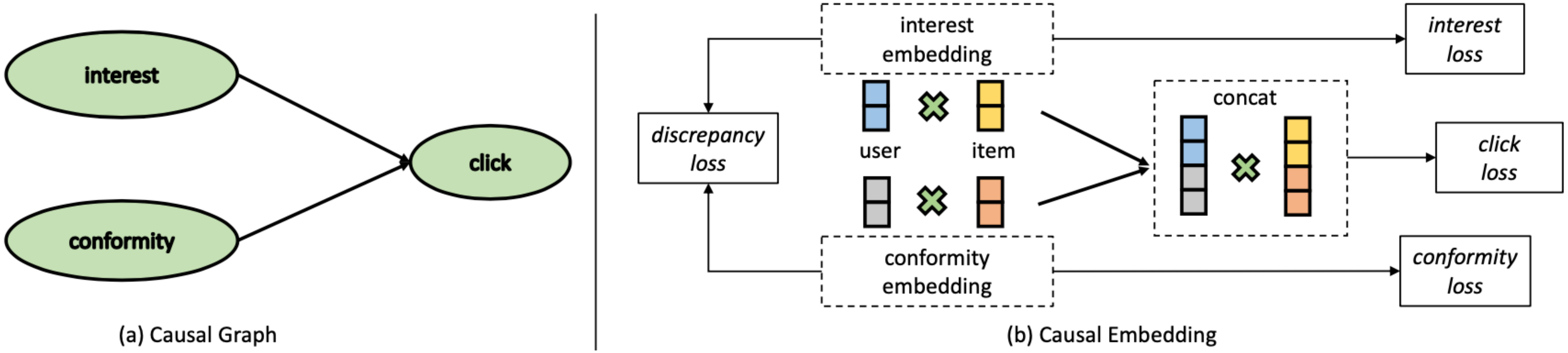
**Idea:** propose a graph-based perspective for fairness-aware representation learning of any recommendation models**.** Adversarial learning of a user-centric graph.




Learning Fair Representations for Recommendation: A Graph-based Perspective WWW21

# In-processing method (Causal Graph)

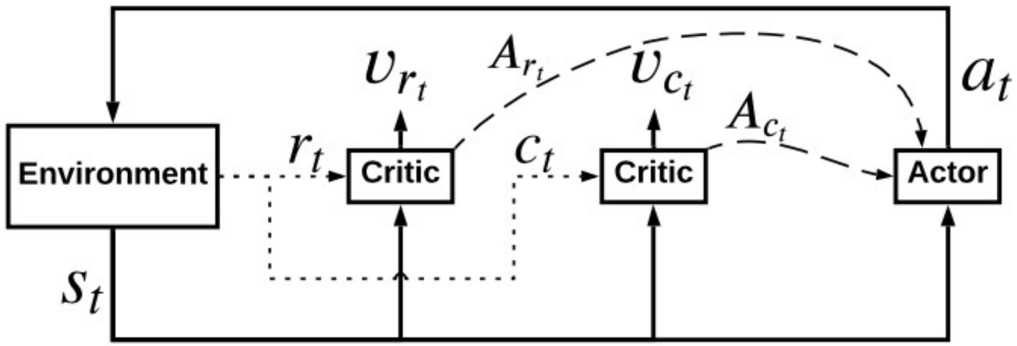
**Idea:** Disentangling Interest and Conformity with Causal Embedding (DICE). Separate embeddings are adopted to capture the two causes, and are trained with cause-specific data.




Disentangling user interest and conformity for recommendation with causal embedding. WWW21.

# In-processing method (Reinforcement Learning)

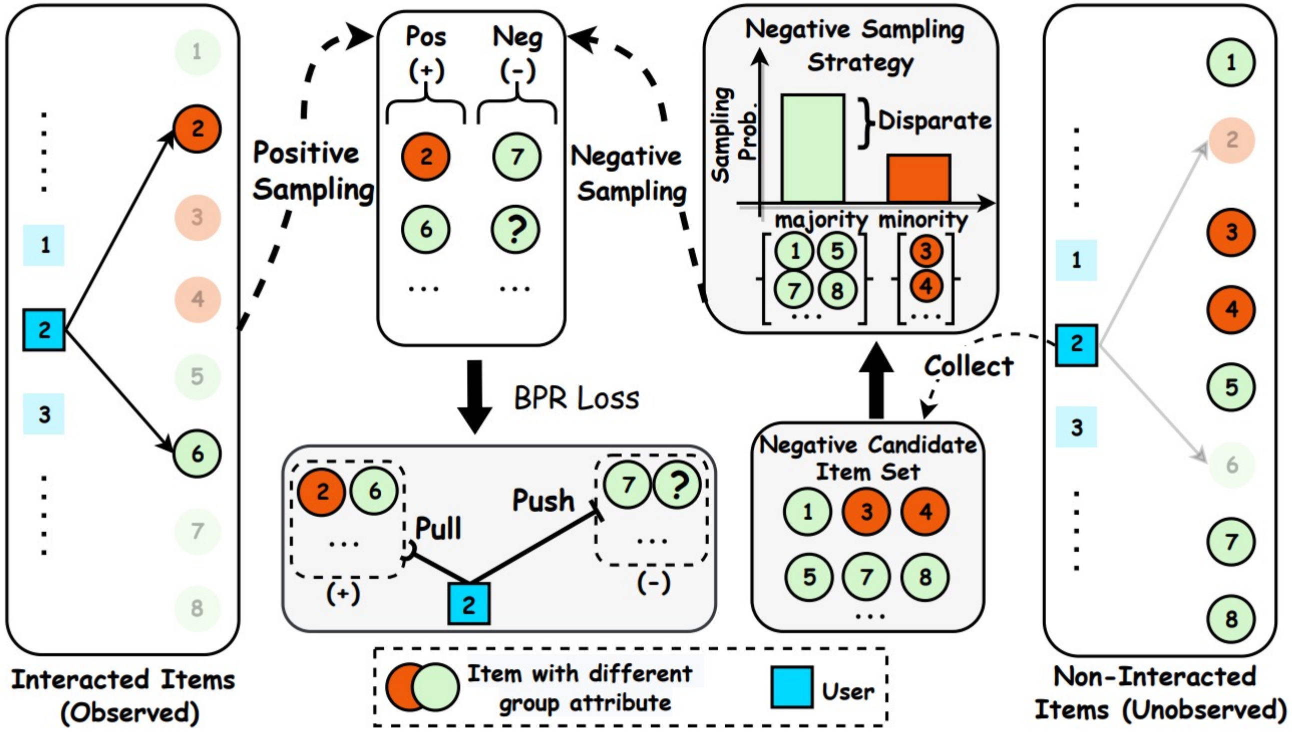
**Idea:** propose a fairness-constrained reinforcement learning algorithm, which models the recommendation problem as a Constrained Markov Decision Process (CMDP). Dynamically adjust the recommendation policy for the fairness requirement.




Towards Long-term Fairness in Recommendation. WSDM21.

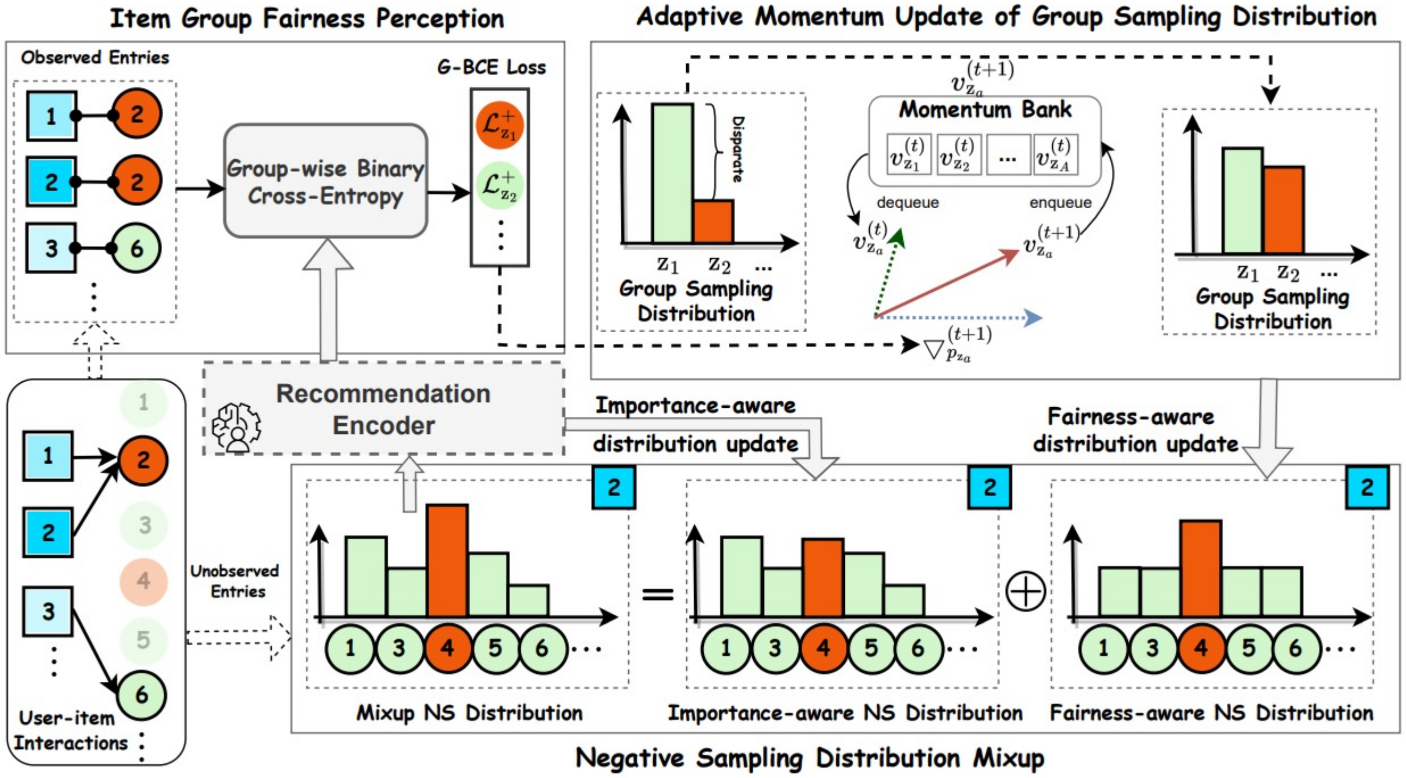
# In-processing method (Negative Sampling)

- **Observation:** the majority item group obtains low (biased) prediction scores via the BPR loss (group-wise performance disparity)



# In-processing method (Negative Sampling)

- **Idea:** adjust the negative sampling distribution (group-wise) adaptively in the training process for meeting the item group fairness objective




Fairly Adaptive Negative Sampling for recommendations. WWW 23

# In-processing method (Negative Sampling)

- Bi-level Optimization of FairNeg

The optimization of the group-wise negative sampling distribution is nested within the recommendation model parameters optimization

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \mathcal{L}_{\text{Recall-Disp}}(\Theta_{\boldsymbol{p}}) := \sum_{\mathrm{z_a} \in Z} \left| \mathcal{L}^+_{\mathrm{z}_a} - \frac{1}{|A|} \Sigma_{\mathrm{z} \in Z} \mathcal{L}^+_{\mathrm{z}} \right|,$$

$$\Theta^*_{\boldsymbol{p}} = \arg\min_{\Theta} \mathcal{L}_{\text{utility}}(\Theta, \boldsymbol{p}) := -\sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{V}^+_u, j \in \mathcal{V}^-_u} \mathcal{L}_{\text{BPR}}(u, i, j; \Theta, \boldsymbol{p}),$$

- Updating Group Sampling Distribution

(1) Group-wise gradient calculation

$$\nabla^{(t)}_{p_{\mathrm{z}_a}} := \mathcal{L}^{+(t)}_{\mathrm{z}_a} - \frac{1}{|A|} \Sigma_{\mathrm{z} \in Z} \mathcal{L}^{+(t)}_{\mathrm{z}},$$

(2) Adaptive momentum update

$$v^{(t+1)}_{\mathrm{z}_a} = \gamma v^{(t)}_{\mathrm{z}_a} + \alpha \cdot \nabla^{(t+1)}_{p_{\mathrm{z}_a}},$$

$$p^{(t+1)}_{\mathrm{z}_a} = p^{(t)}_{\mathrm{z}_a} - v^{(t+1)}_{\mathrm{z}_a},$$

Fairly Adaptive Negative Sampling for recommendations. WWW 23

# Summary of In-processing methods

Substantial fairness improvements

Fairness and utility trade-off

Resource-intensive

# Post-processing method

- **Slot-wise reranking**
- **Global-wise reranking**
- **User-wise reranking**

# Slot-wise Re-ranking

**Idea:** propose a personalized re-ranking algorithm to achieve a fair microlending RS.

A combination of personalization score and a fairness term.

$$\max_{v \in R(u)} \underbrace{(1-\lambda)P(v \mid u)}_{\text{personalization}} + \underbrace{\lambda \sum_{c} P(\mathcal{V}_c) \mathbb{1}_{\{v \in \mathcal{V}_c\}} \prod_{i \in S(u)} \mathbb{1}_{\{i \notin \mathcal{V}_c\}}}_{\text{fairness}},$$

# User-wise Re-ranking

**Idea:** formulate fairness constraints on rankings in terms of exposure allocation. Find rankings that maximize the utility for the user while provably satisfying a specific notion of fairness.

$$
\begin{aligned}
\mathbf{P} = \operatorname{argmax}_{\mathbf{P}} \ & \mathbf{u}^T \mathbf{P} \mathbf{v} && \text{(expected utility)} \\
\text{s.t. } & \mathbb{1}^T \mathbf{P} = \mathbb{1}^T && \text{(sum of probabilities for each position)} \\
& \mathbf{P} \mathbb{1} = \mathbb{1} && \text{(sum of probabilities for each document)} \\
& 0 \le \mathbf{P}_{i,j} \le 1 && \text{(valid probability)} \\
& \mathbf{P} \text{ is fair} && \text{(fairness constraints)}
\end{aligned}
$$

$$\text{Exposure}(G_0|\mathbf{P}) = \text{Exposure}(G_1|\mathbf{P}) \tag{4}$$

$$\Leftrightarrow \frac{1}{|G_0|} \sum_{d_i \in G_0} \sum_{j=1}^{N} \mathbf{P}_{i,j} \mathbf{v}_j = \frac{1}{|G_1|} \sum_{d_i \in G_1} \sum_{j=1}^{N} \mathbf{P}_{i,j} \mathbf{v}_j \tag{5}$$

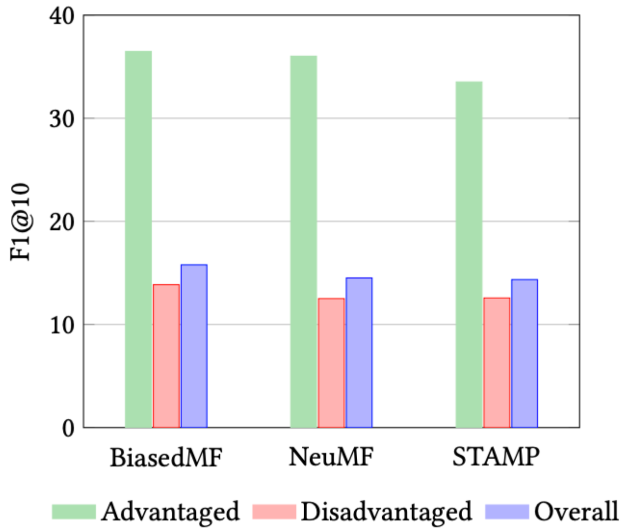
$$\Leftrightarrow \sum_{d_i \in \mathcal{D}} \sum_{j=1}^{N} \left( \frac{\mathbb{1}_{d_i \in G_0}}{|G_0|} - \frac{\mathbb{1}_{d_i \in G_1}}{|G_1|} \right) \mathbf{P}_{i,j} \mathbf{v}_j = 0 \tag{6}$$

$$\Leftrightarrow \mathbf{f}^T P \mathbf{v} = 0 \qquad \left(\text{with } \mathbf{f}_i = \frac{\mathbb{1}_{d_i \in G_0}}{|G_0|} - \frac{\mathbb{1}_{d_i \in G_1}}{|G_1|}\right)$$

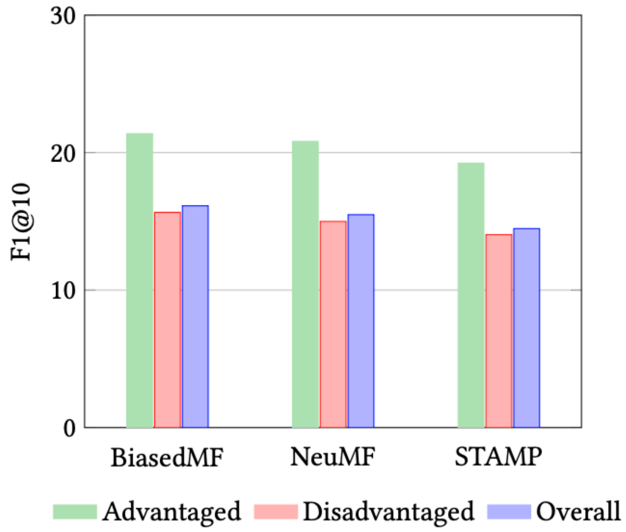
Fairness of Exposure in Rankings. KDD 18

# Global-wise Re-ranking

**Idea:** a re-ranking approach to mitigate this unfairness problem by adding constraints over evaluation metrics.




**(a) Original**




**(b) Fair Method**

$$
\begin{aligned}
\max_{\mathbf{W}_{ij}} \quad & \sum_{i=1}^{n} \sum_{j=1}^{N} \mathbf{W}_{ij} S_{i,j} \\
\text{s.t.} \quad & \mathrm{UGF}\left(Z_1, Z_2, \mathbf{W}\right) < \varepsilon \\
& \sum_{j=1}^{N} \mathbf{W}_{ij} = K, \mathbf{W}_{ij} \in \{0, 1\}
\end{aligned}
$$

# Summary of Post-processing methods

Can be applied to any recommendation systems

Constrained to unfair recommendation model outputs

# • Summary of existing methods

| **Taxonomy** | **Method type** | Related research |
|---|---|---|
| Pre-processing | Data Re-sampling | [95] |
| | Adding Antidote Data | [289] |
| In-processing | Regularization & Constrained Optimization | [26, 351, 393, 409, 461] |
| | Adversarial Learning | [33, 207, 215, 221, 285, 379, 380] |
| | Reinforcement Learning | [120, 122, 244] |
| | Causal Graph | [121, 162, 387, 452] |
| | Others | [31, 110, 167, 224] |
| Post-processing | Slot-wise Re-ranking | [124, 185, 189, 243, 262, 300, 305]<br>[306, 323, 328, 405, 419] |
| | User-wise Re-ranking | [28, 253, 304, 318] |
| | Global-wise Re-ranking | [87, 114, 219, 250, 279, 335, 384, 462] |

A Comprehensive Survey on Trustworthy Recommender Systems. Arxiv 22

# Contents

CONCEPTS AND TAXONOMY

METHODOLOGY

**APPLICATIONS**

SURVEYS AND TOOLS

FUTURE DIRECTIONS

# Applications

- **Ecommerce (Amazon, Etsy)**
- **Social Media (Twitter, LinkedIn)**
- **Content Streaming (Spotify, Youtube)**
- **Ride-hailing (Uber, Lyft)**



# Contents

CONCEPTS AND TAXONOMY

METHODOLOGY

APPLICATIONS

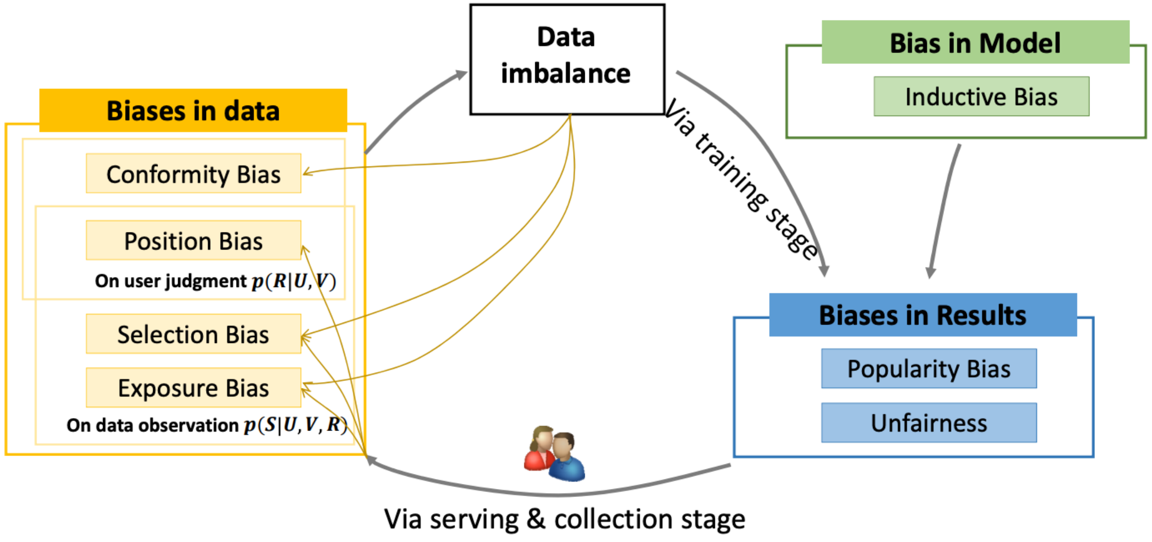
**SURVEYS AND TOOLS**

FUTURE DIRECTIONS

# Surveys

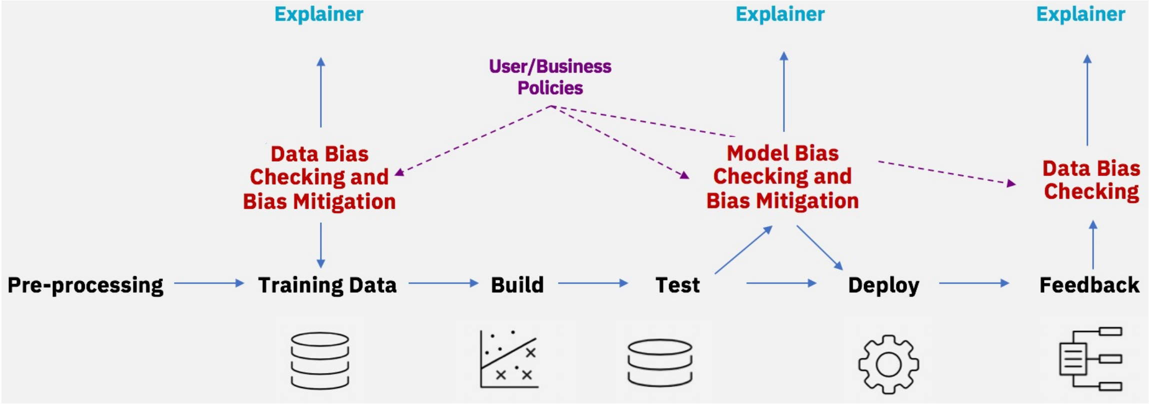
- TOIS 23' Bias and Debias in Recommender System: A Survey and Future Directions

- Arxiv 22' A Comprehensive Survey on Trustworthy Recommender Systems

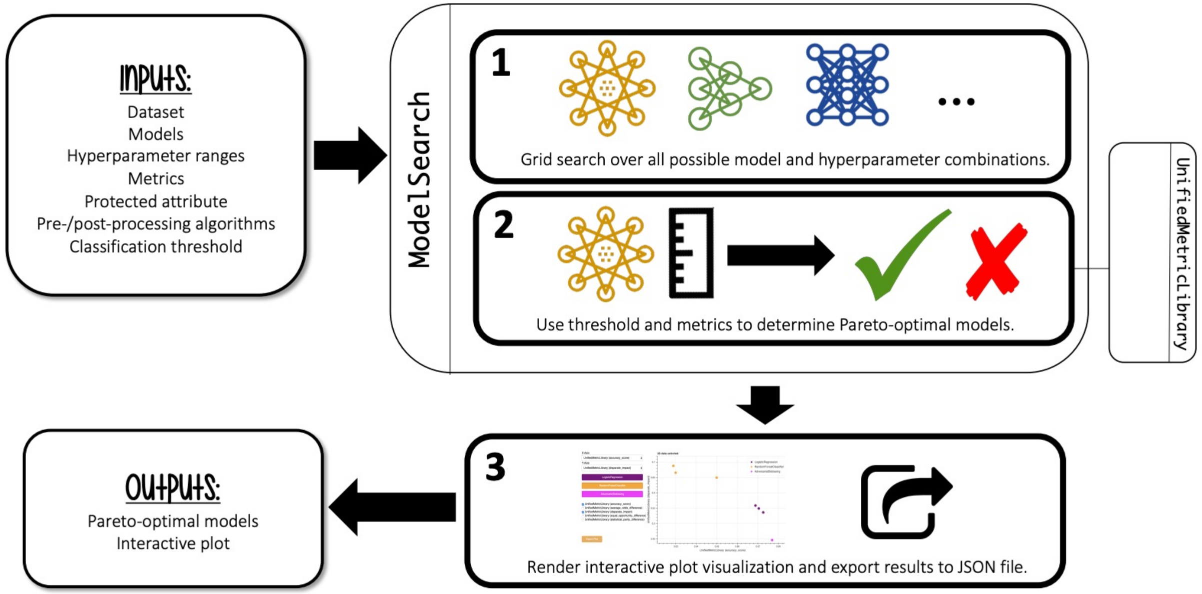



# Tools

- **IBM Fairness 360**




- **Fairkit-learn**



# Contents

# Future Directions

- **Consensus on Fairness Definition**
- **Fairness-Utility tradeoff**
- **Fairness-aware algorithm design**
- **Better evaluation**

# Trustworthy Recommender Systems

Introduction

Wenqi Fan

Non-discrimination & Fairness

Xiao Chen

Safety & Robustness

Shijie Wang

Explainability

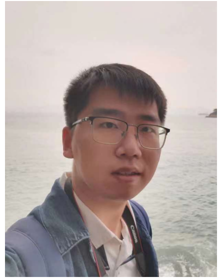
Jingtong Gao

Privacy

Lin Wang

Environmental Well-being

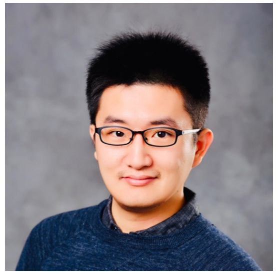
Accountability & Auditability



Qidong Liu

Dimension Interactions

Future Directions



Xiangyu Zhao

# Real World Attacks in Recommender Systems

**DIGITAL LIVING** | JULY 26, 2022

## Amazon's War on Fake Reviews

*By Matt Stieb, Intelligencer staff writer*


Photo-Illustration: Intelligencer; Photos: Getty Images/Amazon

**BUSINESS**

## How merchants use Facebook to flood Amazon with fake reviews

By Elizabeth Dwoskin and Craig Timberg

April 23, 2018 at 1:26 p.m. EDT

An Amazon distribution center in Madrid, shown in November. (Emilion Naranjo/EPA-EFE/Shutterstock)

https://nymag.com/intelligencer/2022/07/amazon-fake-reviews-can-they-be-stopped.html
https://www.washingtonpost.com/business/economy/how-merchants-secretly-use-facebook-to-flood-amazon-with-fake-reviews/2018/04/23/5dad1e30-4392-11e8-8569-26fda6b404c7_story.html

# Safety and Robustness

"A decision aid, no matter how sophisticated or 'intelligent' it may be, may be rejected by a decision maker who does not trust it, and so its potential benefits to system performance will be lost."

—Bonnie M. Muir, psychologist at University of Toronto

# Safety and Robustness

By examining Adversarial Robustness,
we expect the recommender system to:

- Be reliable, secure and stable

# Outline

Concepts and Taxonomy

Adversarial Attack

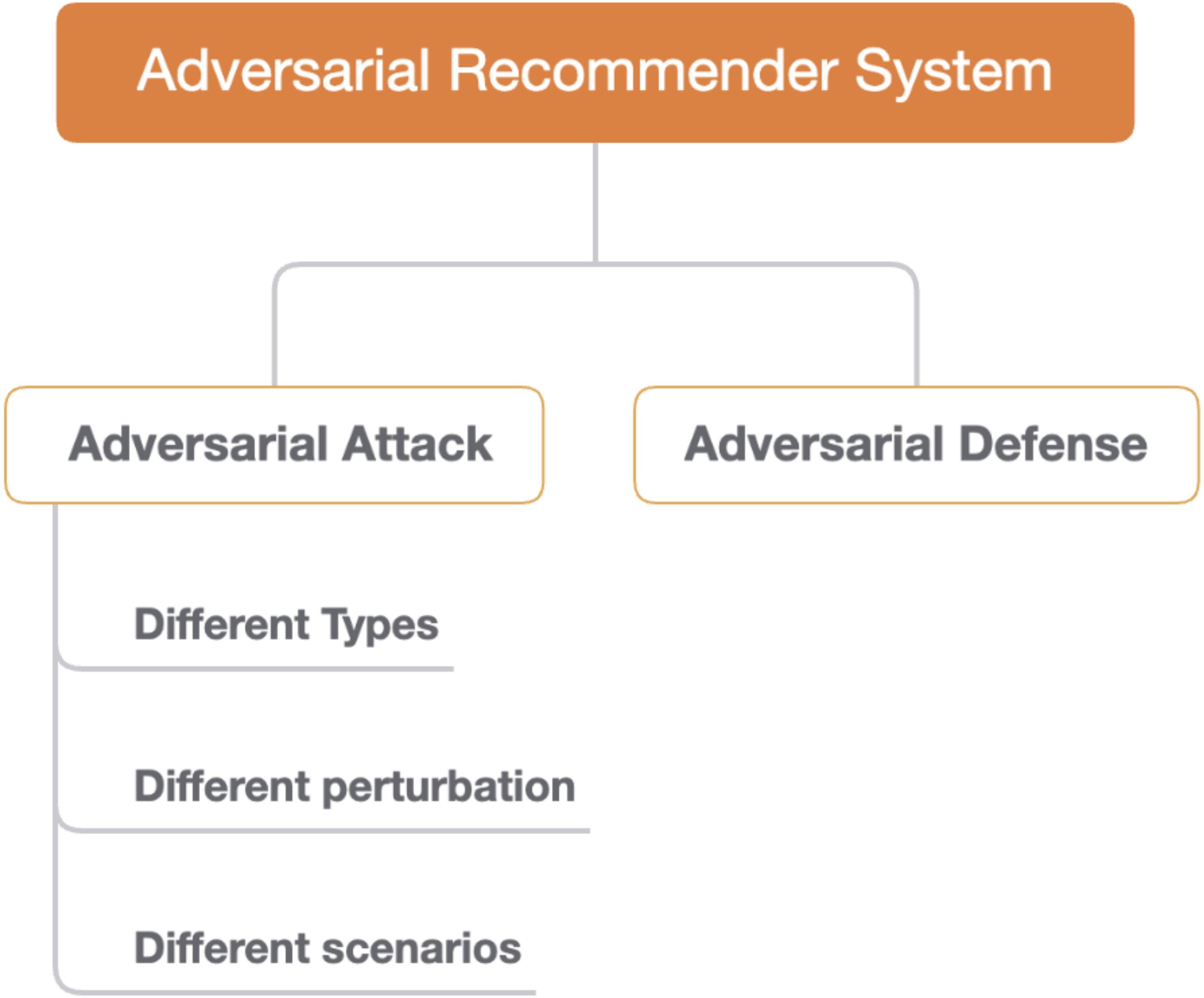
Adversarial Defense

Application

Adversarial Learning Surveys and Tools

Future directions

# Taxonomy

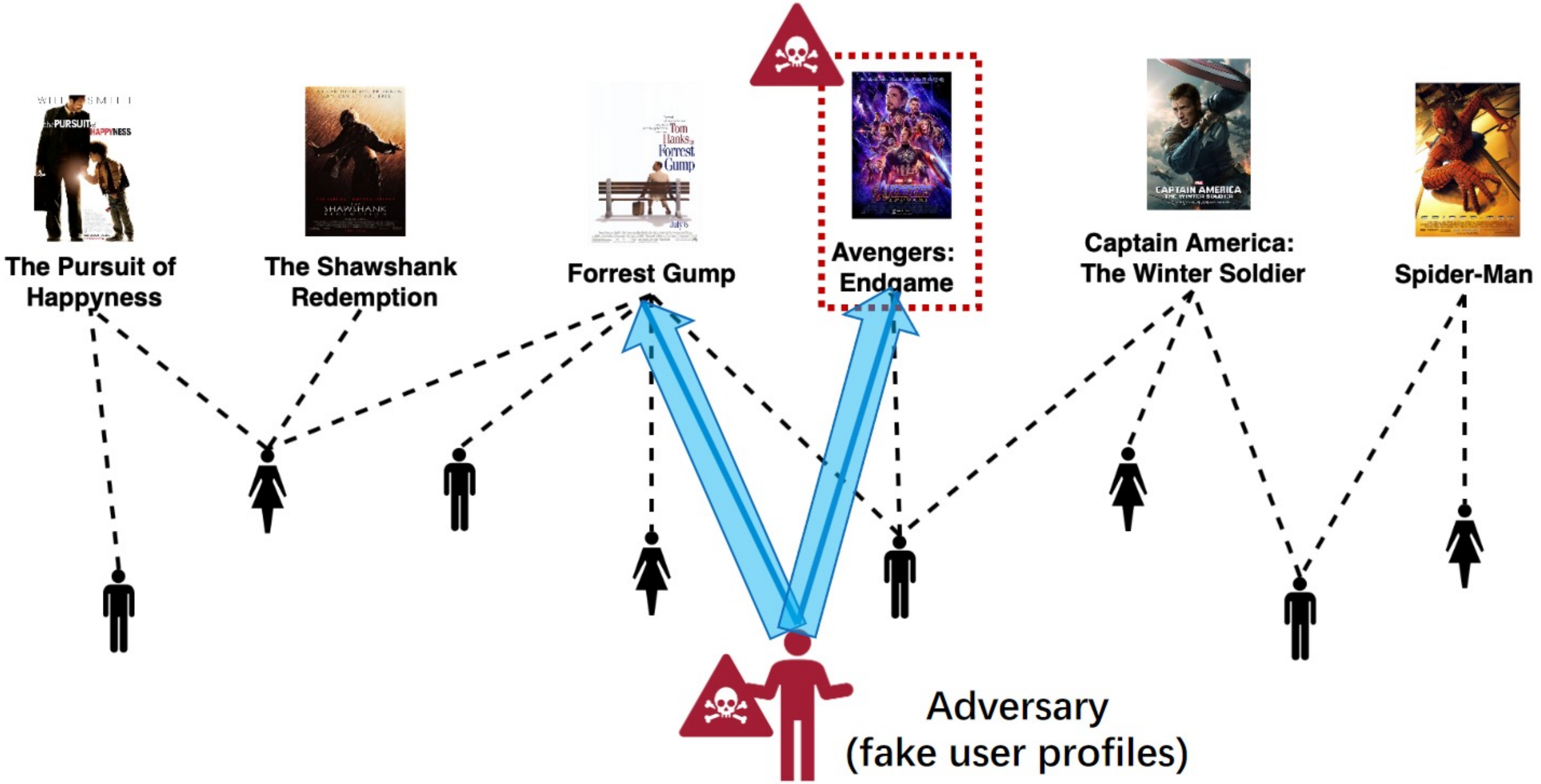

Adversarial Recommender System
Adversarial Attack
Adversarial Defense
Different Types
Different perturbation
Different scenarios

# Adversarial Attack

- Poisoning Attacks vs. Evasion Attacks
  - They happen in training phase/ happen in test/inference phase
- White-box attacks vs. Grey-box attacks vs. Black-box attacks
  - They have all knowledge of the recommender system / have partial knowledge/ have no knowledge or limit knowledge
- Targeted Attacks vs. Untargeted Attacks
  - They aim to promote/demote a set of target items/ aim to degrade a recommendation system's overall performance

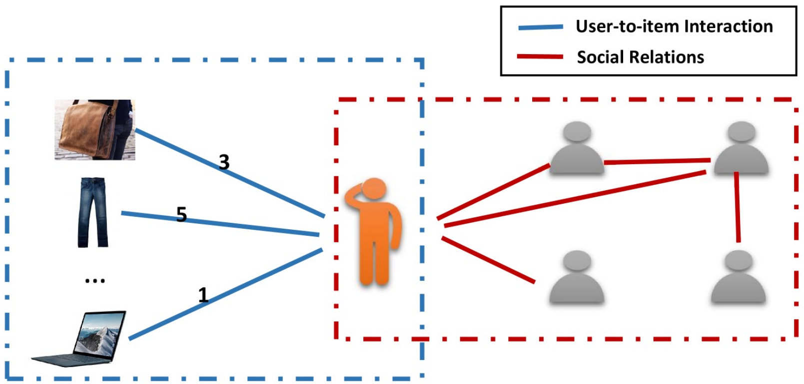
# Adversarial in Different Perturbation

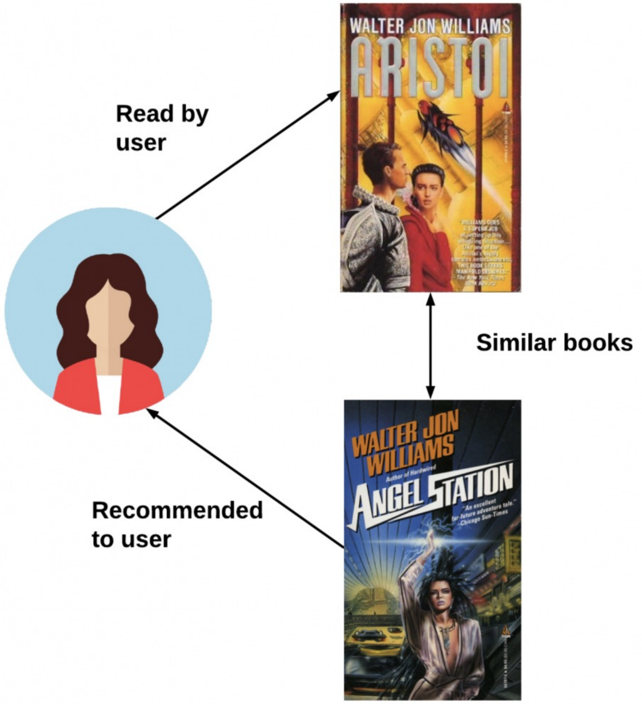
- Adding fake user profiles into user-item interactions, modifying user attributes information, adding social relations, etc

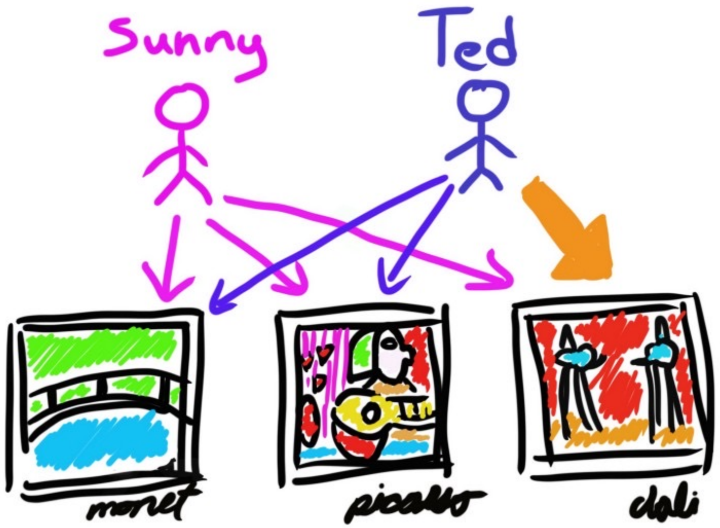


# Adversarial in Different Scenarios

- Collaborative Filtering Recommender System
- Social Recommender System
- Content-based Recommender System
- . . .




Graph neural networks for social recommendation, Fan et al 2019.
https://thingsolver.com/introduction-to-recommender-systems/

# Adversarial Defenses

- Perturbations Detection vs. Adversarial Training

  - It is to identify perturbations data and remove them/ enhances the robustness of recommender systems

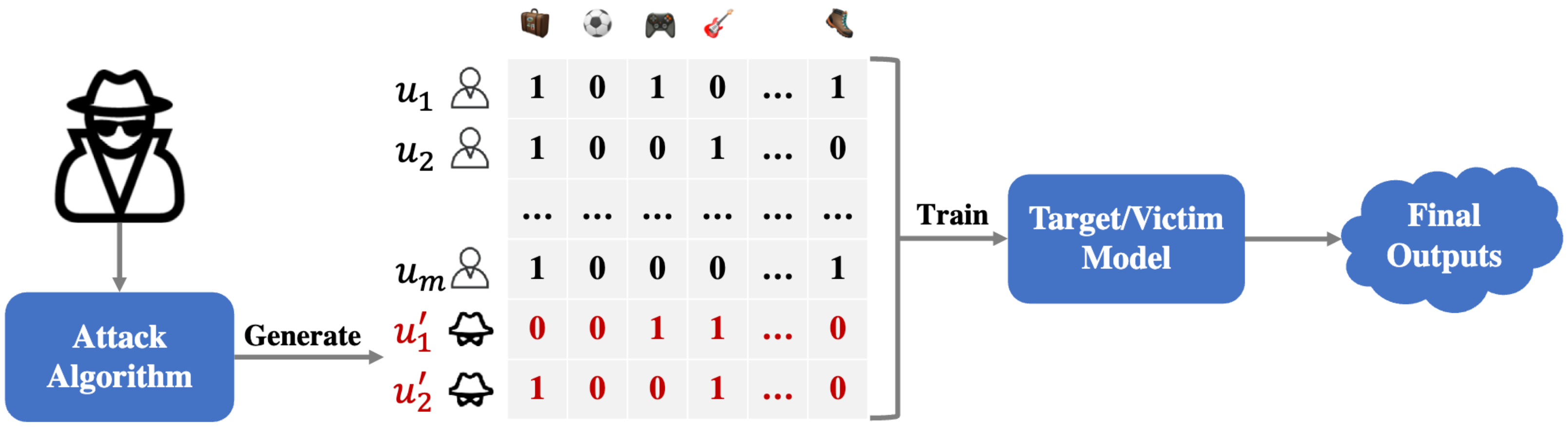
# Outline

- Concepts and Taxonomy
- Adversarial Attack
- Adversarial Defense
- Application
- Adversarial Learning Surveys and Tools
- Future directions

# Adversarial Attack for Recommender System

- A Unified Formulation of Poisoning Attack

$$\min_{\widehat{U}} \mathcal{L}_{adv}(\theta^*), \qquad \text{s.t.} \qquad \theta^* = \arg\min_{\theta}(\mathcal{L}_{rec}(\boldsymbol{R}, \boldsymbol{O}_\theta) + \mathcal{L}_{rec}(\widehat{\boldsymbol{R}}, \boldsymbol{O}_\theta))$$



# Heuristic Attack

- Heuristic Attack Method

  - It assigns high scores to target items
  - Give a low score to random others
  - It interacts with some popular items
  - Include random attack, average attack, bandwagon attack, and segment attack
  - . . .

# Heuristic Attack

# Heuristic Attack

- Random Attack
  - Attacker's Goal: promote certain items availability of being recommended

high scores to target item

| | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 |
|---|---|---|---|---|---|---|
| **User1** | 4 | 3 | 4 | - | 3 | 4 |
| **User2** | 5 | 5 | 1 | 4 | 1 | 3 |
| **User3** | 1 | 5 | 2 | 5 | 4 | 2 |
| **User4** | 5 | 1 | 5 | 3 | - | 5 |
| **User5** | 3 | 5 | 4 | 4 | 1 | 0 |
| **User6** | - | 5 | 5 | 4 | - | 2 |
| **Attacker1** | 1 | - | 1 | 1 | 5 | - |
| **Attacker2** | - | 1 | 1 | 1 | 5 | - |

low score to random others

Shilling Recommender Systems for Fun and Profit, WWW 2004.

# Heuristic Attack

- Average Attack

high scores to target item

| | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 |
|---|---|---|---|---|---|---|
| **User1** | 4 | 3 | 4 | - | 3 | 4 |
| **User2** | 5 | 5 | 1 | 4 | 1 | 3 |
| **User3** | 1 | 5 | 2 | 5 | 4 | 2 |
| **User4** | 5 | 1 | 5 | 3 | - | 5 |
| **User5** | 3 | 5 | 4 | 4 | 1 | 0 |
| **User6** | - | 5 | 5 | 4 | - | 2 |
| **Attacker1** | 3 | 4 | 3 | 4 | 5 | - |
| **Attacker2** | 3 | 4 | 3 | 4 | 5 | - |

average score to random others

Shilling Recommender Systems for Fun and Profit, WWW 2004.

# Heuristic Attack

- Bandwagon attack

popular item

target item

| | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 |
|---|---|---|---|---|---|---|
| **User1** | - | 4 | 4 | - | 3 | - |
| **User2** | - | 5 | 1 | - | 1 | 3 |
| **User3** | 1 | 4 | 2 | 1 | 4 | - |
| **User4** | - | 4 | 5 | - | - | - |
| **User5** | - | 5 | 4 | - | 1 | - |
| **User6** | - | 5 | 5 | - | - | - |
| **Attacker1** | - | 4 | 4 | - | 5 | - |
| **Attacker2** | - | 4 | 4 | - | 5 | - |

Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness, TOIT 2007.

# Heuristic Attack

- Segment attack

Similar item

target item

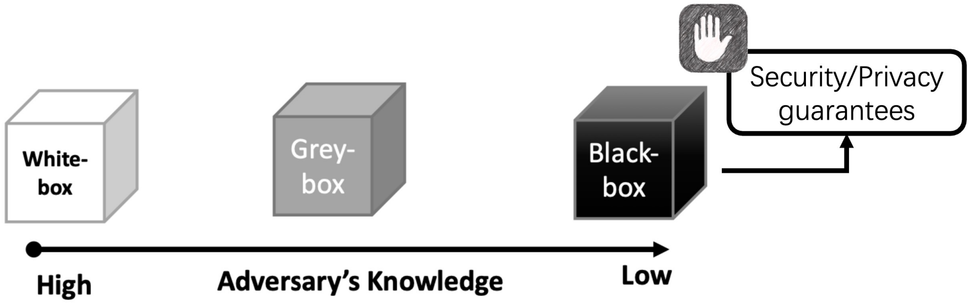
| | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 |
|---|---|---|---|---|---|---|
| **User1** | 4 | 3 | 4 | - | 3 | 4 |
| **User2** | 5 | 5 | 1 | 4 | 1 | 3 |
| **User3** | 1 | 5 | 2 | 5 | 4 | 2 |
| **User4** | 5 | 1 | 5 | 3 | - | 5 |
| **User5** | 3 | 5 | 4 | 4 | 1 | 0 |
| **User6** | - | 5 | 5 | 4 | - | 2 |
| **Attacker1** | 1 | 4 | 4 | 1 | 5 | - |
| **Attacker2** | - | 4 | 4 | 1 | 5 | - |

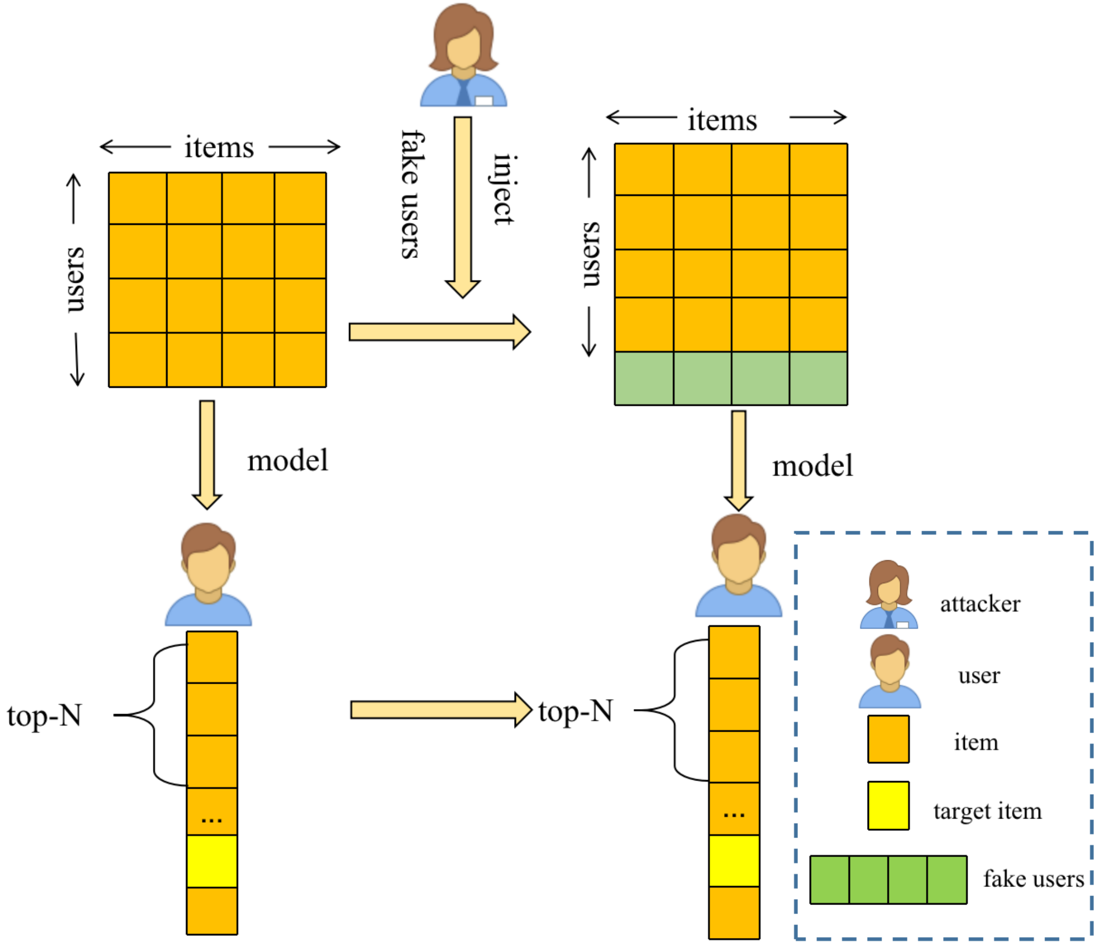
Segment-based injection attacks against collaborative filtering recommender systems, ICDM 2005.

# Gradient-based Attack

- Gradient-based Methods

  - White-Box Attack: Optimization

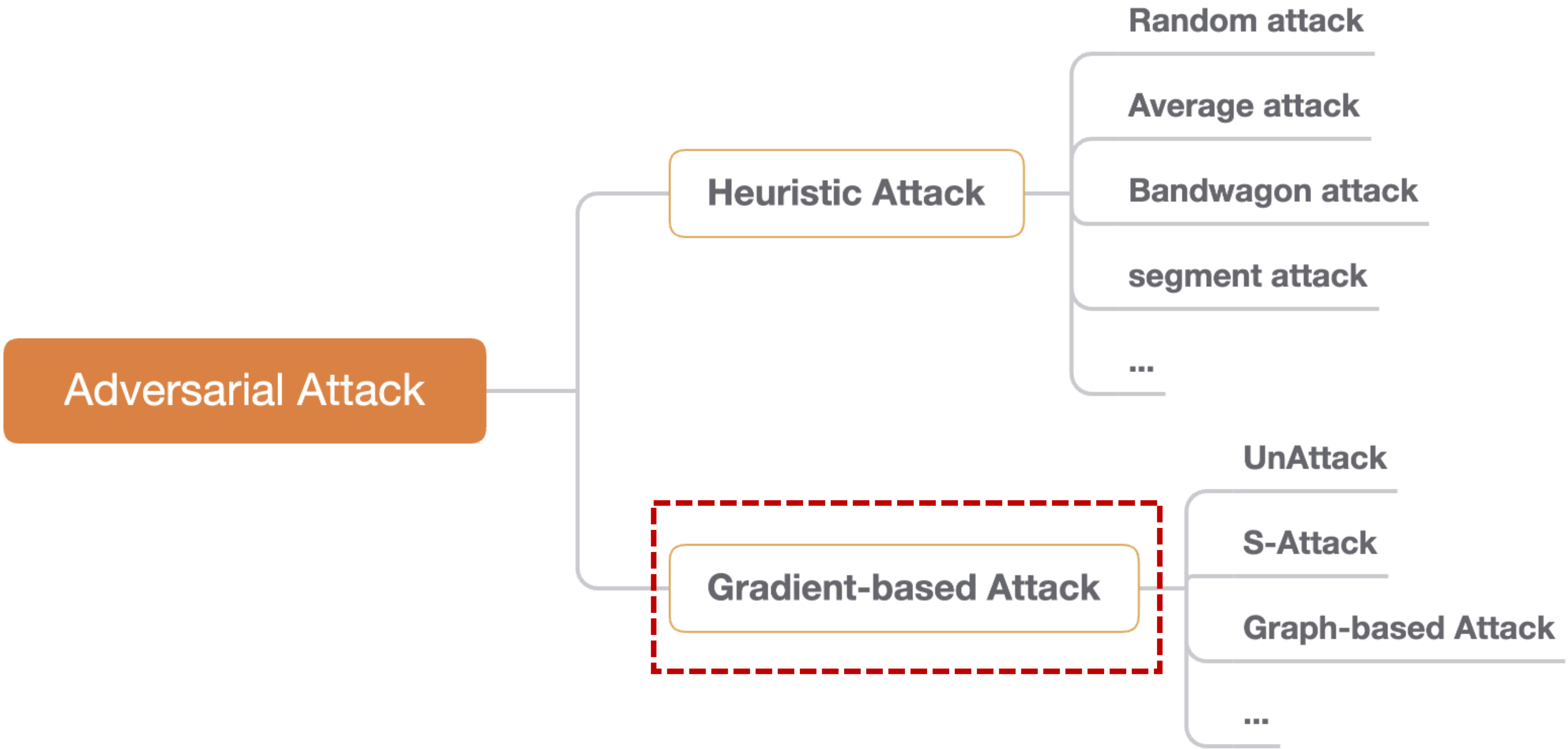



$$\min_{\widehat{U}} \mathcal{L}_{adv}(\theta^*), \quad \text{s.t.} \quad \theta^* = \arg\min_{\theta}(\mathcal{L}_{rec}(\mathbf{R}, \mathbf{O}_\theta) + \mathcal{L}_{rec}(\widehat{\mathbf{R}}, \mathbf{O}_\theta))$$

Data poisoning attacks on neighborhood-based recommender systems, ETT 2019.

# Gradient-based Attack



# UNAttack

- UNAttack
  - Optimize the ratings of fake users one by one rather than for all m fake users at the same time
  - Borrow the strategy from the ranking problem to construct pairwise loss function

$$p_{ui} = \sum_{v \in S(u,K) \cap U_i^{+})} s_{uv} X_{vi}$$

$$loss_1 = \sum_{v \in S(u,K)} \sigma(\boxed{s_{uv}} - s_{uf})$$

$$loss_2 = \sum_{i \in L_u} \sigma(\boxed{p_{ui}} - p_{ut})$$

$$loss_u = (1-\lambda) loss_1 + \lambda loss_2$$

$$Minimize(F(X_f) = loss)$$

$$s.t.\ |X_f| \le z,$$

$$X_{fi} \in \{0, 1, \ldots, r_{max}\}$$

$$loss = \sum_{u \in U_t^-} loss_u$$

$$loss = \sum_{u \in U_t^-} loss_u$$

Make the fake user be in the top-K nearest neighbours of user, which can be expressed as $s_{uf} > s_{uv}$.

Data poisoning attacks on neighborhood-based recommender systems, ETT 2019.

# UNAttack

- UNAttack
  - Choosing the optimal filler-items for fake users

$$X_f^{(t)} = Project(X_f^{(t-1)} - \eta \frac{\partial F(X_f)}{\partial X_f})$$

where $Project(x)$ is the project function that cuts each $X_{fi}$ into the range $[0,1,..r_{max}]$.

$$\frac{\partial F(X_f)}{\partial X_f} = \sum_{u \in U_t^-} (1-\lambda)\frac{\partial loss_1}{\partial X_f} + \lambda \frac{\partial loss_2}{\partial X_f}$$

Gradient

$$\frac{\partial (loss_1)}{\partial X_f} = \sum_{v \in S(u,k)} \frac{\partial \sigma(Q)}{\partial Q}\boxed{(\frac{\partial s_{uv}}{\partial X_f} - \frac{\partial s_{uf}}{\partial X_f})}$$

$$\frac{\partial (loss_2)}{\partial X_f} = \sum_{i \in L_u} \sum_{v \in W} \frac{\partial \sigma(P)}{\partial P}(\frac{\partial s_{uv} X_{vi}}{\partial X_f} - \frac{\partial s_{uf} X_{ft}}{\partial X_f})$$

similarity

$$\frac{\partial s_{uf}}{\partial X_f} = \frac{X_u}{\|X_u\|\|X_f\|} - \frac{X_u X_f}{\|X_u\|\|X_f\|}\frac{X_f}{\|X_f\|^2}$$

Data poisoning attacks on neighborhood-based recommender systems, ETT 2019.

# UNAttack

- UNAttack

Give the target items the maximum ratings.

Inspired by the ranking problem, all items will be ranked according to $X_{fi}$, and top-z items with the highest values will be chosen as the filler-items.

**Algorithm 1.** UNAttack

**Input**: Matrix $R_{m \times n}$
**Parameter**: $\lambda, K, N, z, j$
**Output**: $j$ fake users

```
for each fake user f do
    Solve the problem in Equation 6 with current rating matrix R to get X_f
    Let X_ft = r_max
    Select z items with highest value in X_fi as filler items.
    For each filler-items j, X_fj ~ N(μ_j, σ_j^2)
    R_{m×n} = R_{m×n} ∪ X_f
end for
```

The rating score assigned to each filler-item is drawn from a normal distribution of the normal users' rating data of this item.

# S-Attack

- Attack matrix factorization based recommender systems
  - Attacker's Goal: promote certain items availability of being recommended
  - Attacker's knowledge: fully (partial) observable dataset
  - Challenge:
    - User ratings are discrete
    - Excessive number of users

$$\arg\min_{X,Y} \sum_{(u,i)\in\mathcal{E}} \left(r_{ui} - \boldsymbol{x}_u^\top \boldsymbol{y}_i\right)^2 + \lambda\left(\sum_u \|\boldsymbol{x}_u\|_2^2 + \sum_i \|\boldsymbol{y}_i\|_2^2\right)$$

$$\begin{aligned} \max\ & h(t) \\ \text{s.t. } & |\Omega_v| \le n+1, && \forall v \in \mathcal{M}, \\ & r_{vi} \in \{0, 1, \cdots, r_{max}\}, && \forall v \in \mathcal{M}, \forall i \in \Omega_v. \end{aligned}$$

Influence Function based Data Poisoning Attacks to Top-N Recommender Systems, WWW 2020.

# S-Attack

- Step 1: Optimize one by one
- Step 2: Relax the discrete ratings to continuous

$$\boldsymbol{w}_v = [w_{vi}, i \in \Omega_v]^\top$$

$$r_{vi} \in \{0, 1, \cdots, r_{max}\} \Rightarrow w_{vi} \in [0, r_{max}] \Rightarrow w_{vi} \in \{0, 1, \cdots, r_{max}\}$$

Discrete | Continues | Discrete

Influence Function based Data Poisoning Attacks to Top-N Recommender Systems, WWW 2020.

# S-Attack

- Step 3: Approximating the Hit Ratio
- Step 4: Determining the Set of Influential Users

$$\min_{\boldsymbol{w}_v} \mathcal{L}_{\mathcal{U}}(\boldsymbol{w}_v) = \sum_{u \in \mathcal{U}} \sum_{i \in \underbrace{\boxed{\Gamma_u}}_{\text{Top-k list}}} g(\hat{r}_{ui} - \hat{r}_{ut}) + \eta \|\boldsymbol{w}_v\|_1$$

$$\text{s.t. } w_{vi} \in [0, r_{max}],$$

$$\min_{\boldsymbol{w}_v} \mathcal{L}_{\overbrace{\boxed{\mathcal{S}}}^{\text{Influential Users}}}(\boldsymbol{w}_v) = \sum_{u \in \mathcal{S}} \sum_{i \in \Gamma_u} g(\hat{r}_{ui} - \hat{r}_{ut}) + \eta \|\boldsymbol{w}_v\|_1$$

$$\text{s.t. } w_{vi} \in [0, r_{max}].$$

Influence Function based Data Poisoning Attacks to Top-N Recommender Systems, WWW 2020.

# Graph-Based Attack

- Attack graph-based recommender systems
  - Attack using random walk algorithm

Random walk:

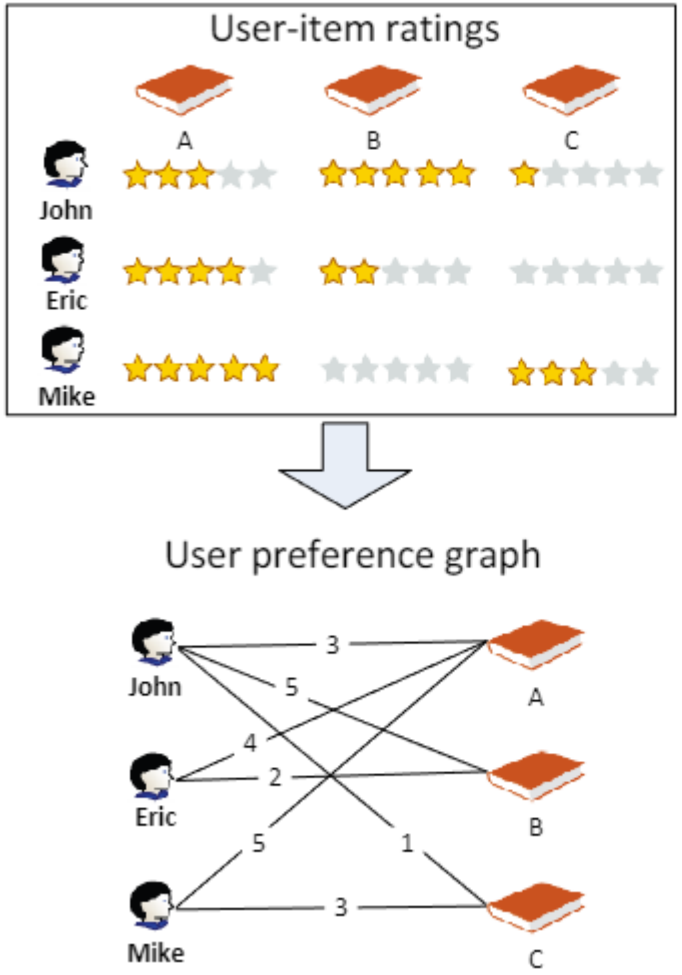
$$p_u = (1 - \alpha) \cdot Q \cdot p_u + \alpha \cdot e_u$$

$$Q_{xy} = \begin{cases} \frac{r_{xy}}{\sum_{z \in \Gamma_x} r_{xz}} & \text{if } (x, y) \in E \\ 0 & \text{otherwise} \end{cases}$$

Loss function:

$$l_u = \sum_{i \in L_u} g(p_{ui} - p_{ut})$$

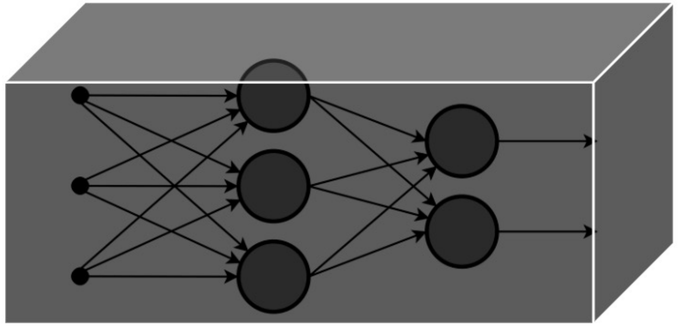
$$g(x) = \frac{1}{1 + \exp(-x/b)}$$




Poisoning Attacks to Graph-Based Recommender Systems, ACSAC 2018.

# Black-Box Attack

- Black-Box Attack

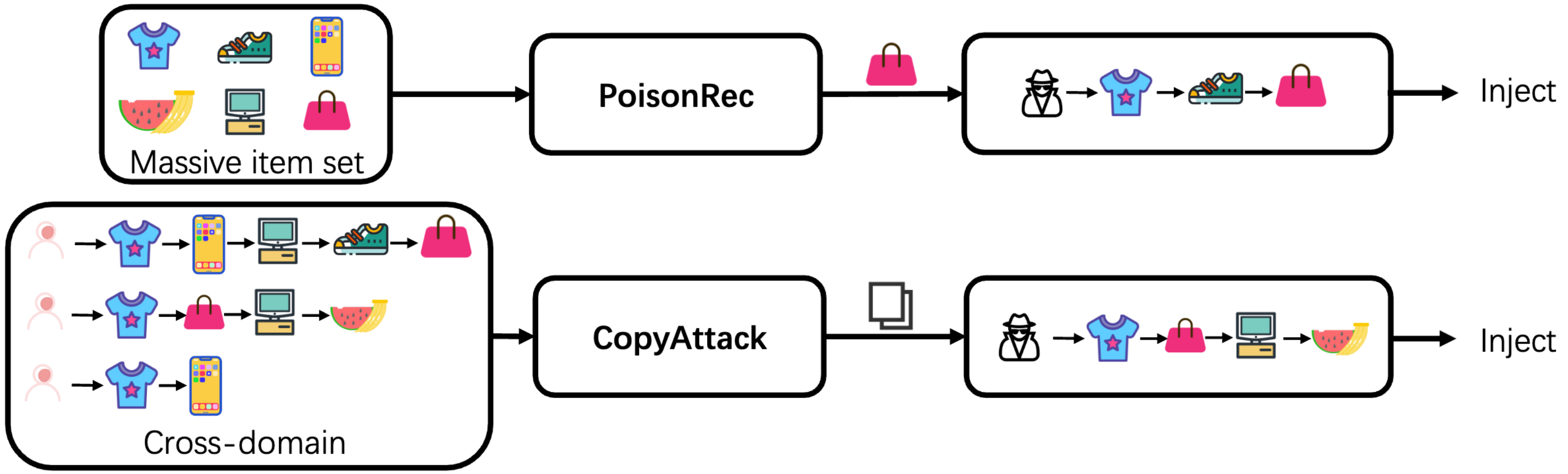
**Perturbed Data** →  → **Promote/Demote Target Item**

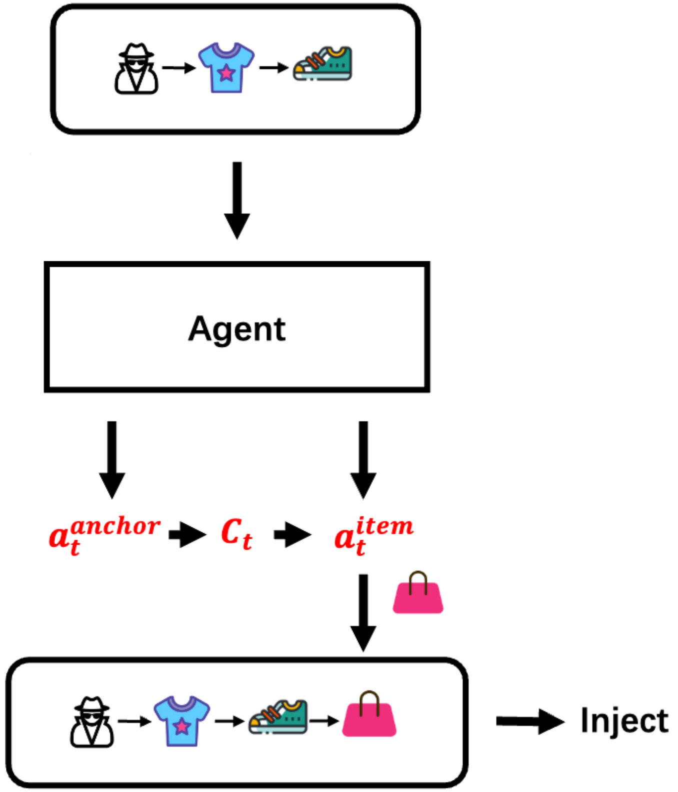
# Reinforcement Learning-based Attack

- Challenges in existing attacking methods:
  - Model structure, parameters and training data are unknown
  - Unable to get user-item interactions
  - Black-box setting
    - Reinforcement Learning (RL) -- Query Feedback (Reward)

# Reinforcement Learning-based Attack

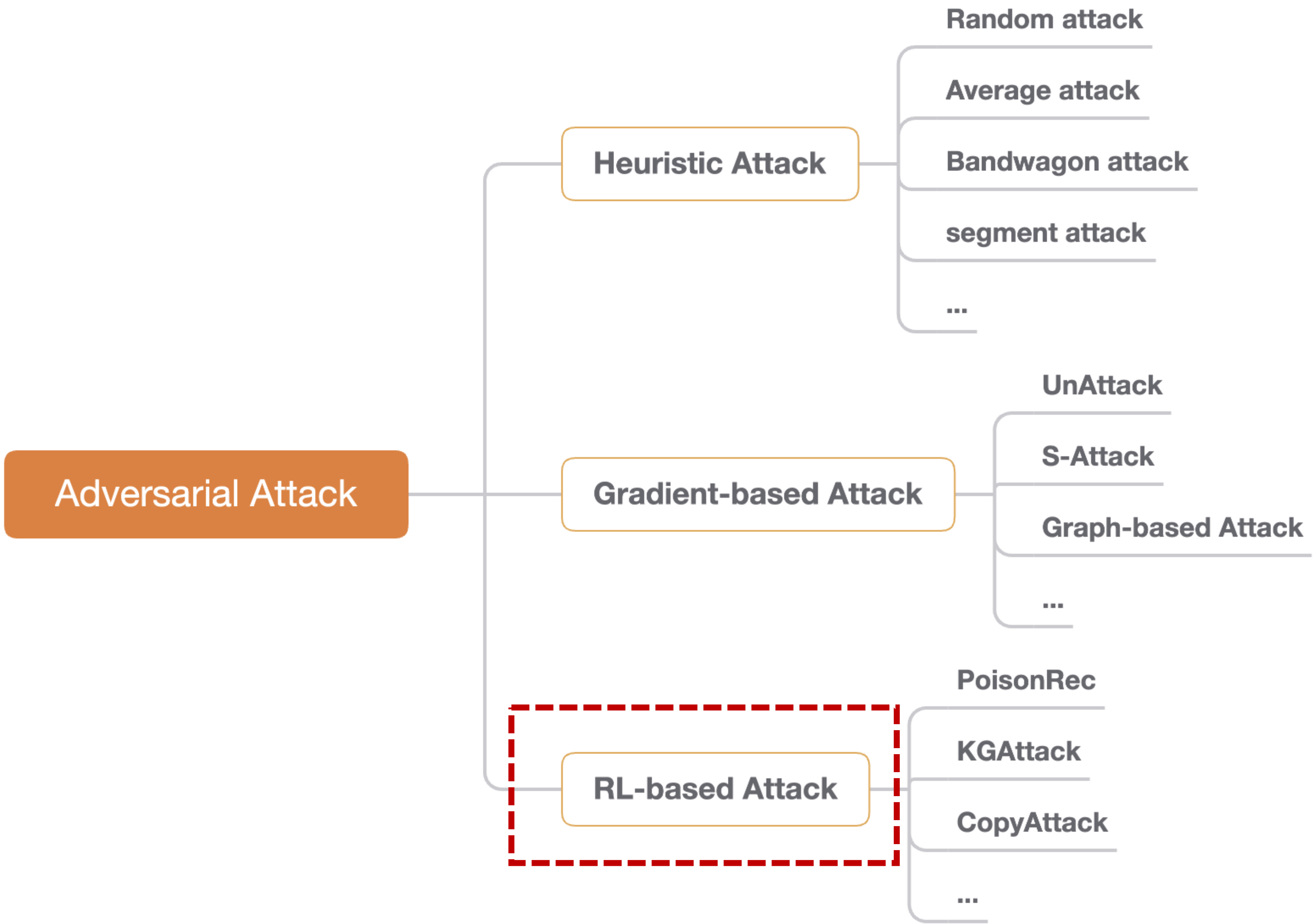
- Reinforcement Learning-based Methods
  - PoisonRec
  - KGAttack
  - CopyAttack




An Adaptive Data Poisoning Framework for Attacking Black-box Recommender Systems, ICDE 2020.
Attacking Black-box Recommendations via Copying Cross-domain User Profiles, ICDE 2021
Knowledge-enhanced Black-box Attacks for Recommendations, KDD 2022

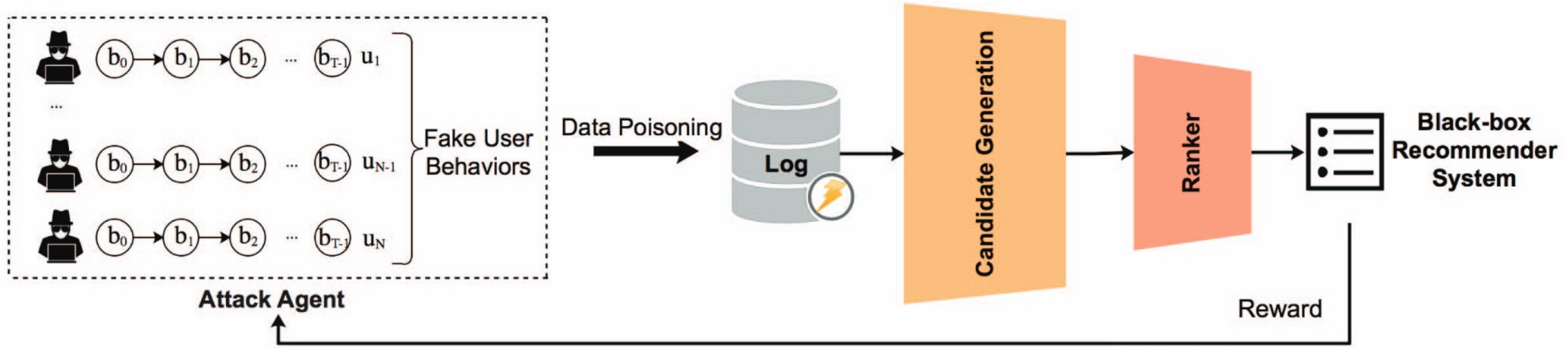
# Reinforcement Learning-based Attack



# PoisonRec

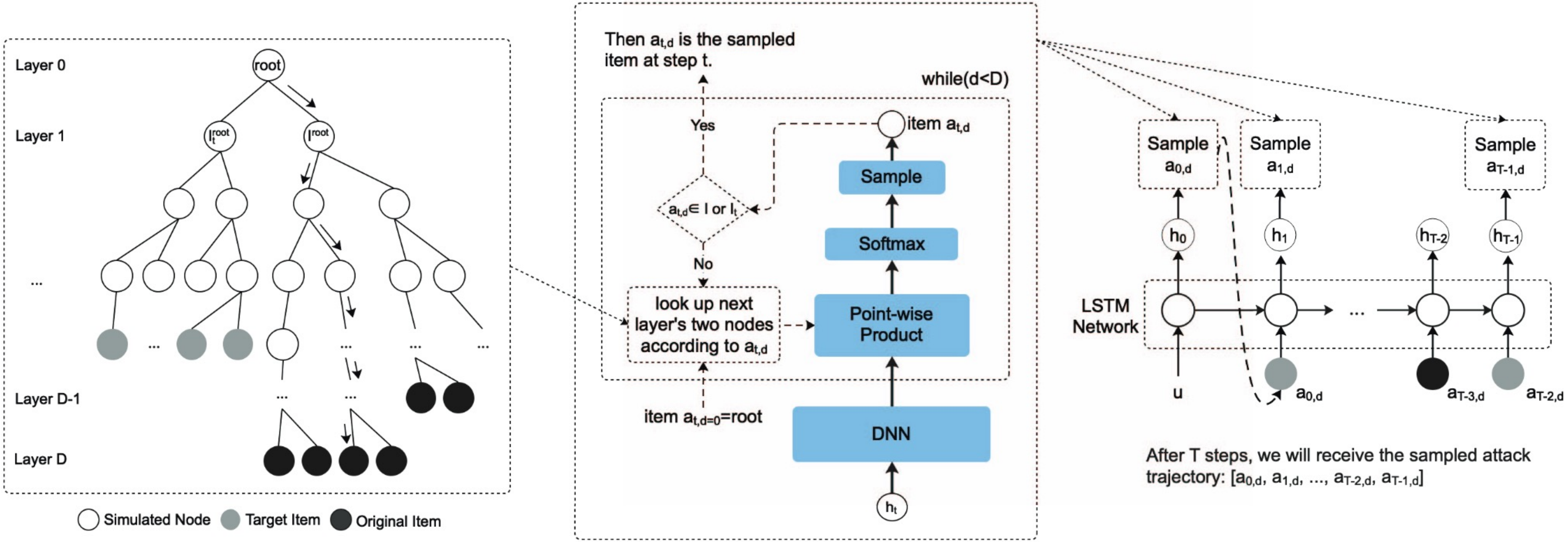
- Target: $$RecNum = \sum_u |L_u \cap I_t|$$

- DNN + PPO



# PoisonRec

- Introduce (Biased Complete Binary Tree) BCBT to reduce action space




(a) The biased complete binary tree, BCBT

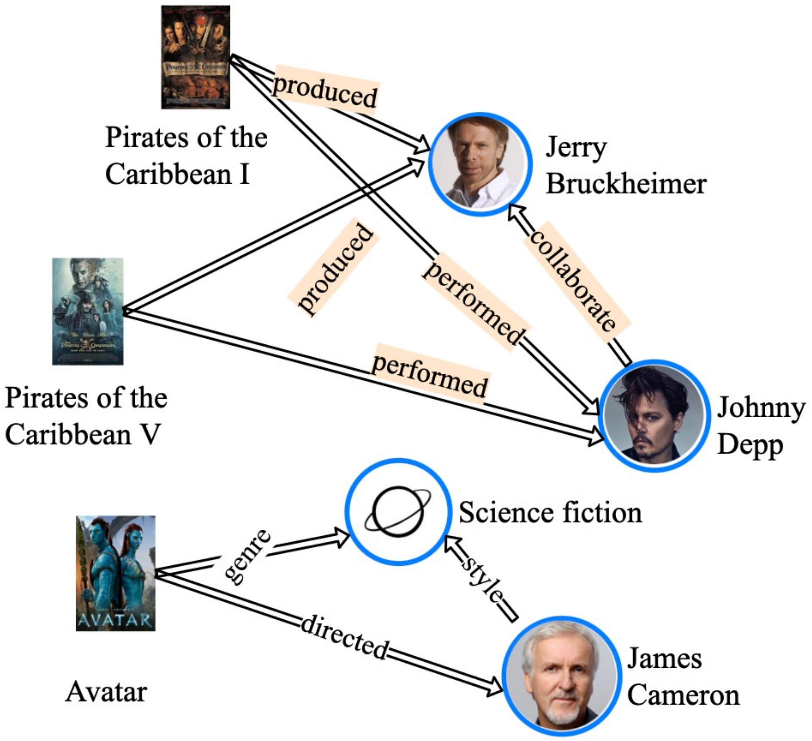
(b) The sampling process on BCBT

(c) The sampling process for a complete attack trajectory.

An Adaptive Data Poisoning Framework for Attacking Black-box Recommender Systems, ICDE 2020.

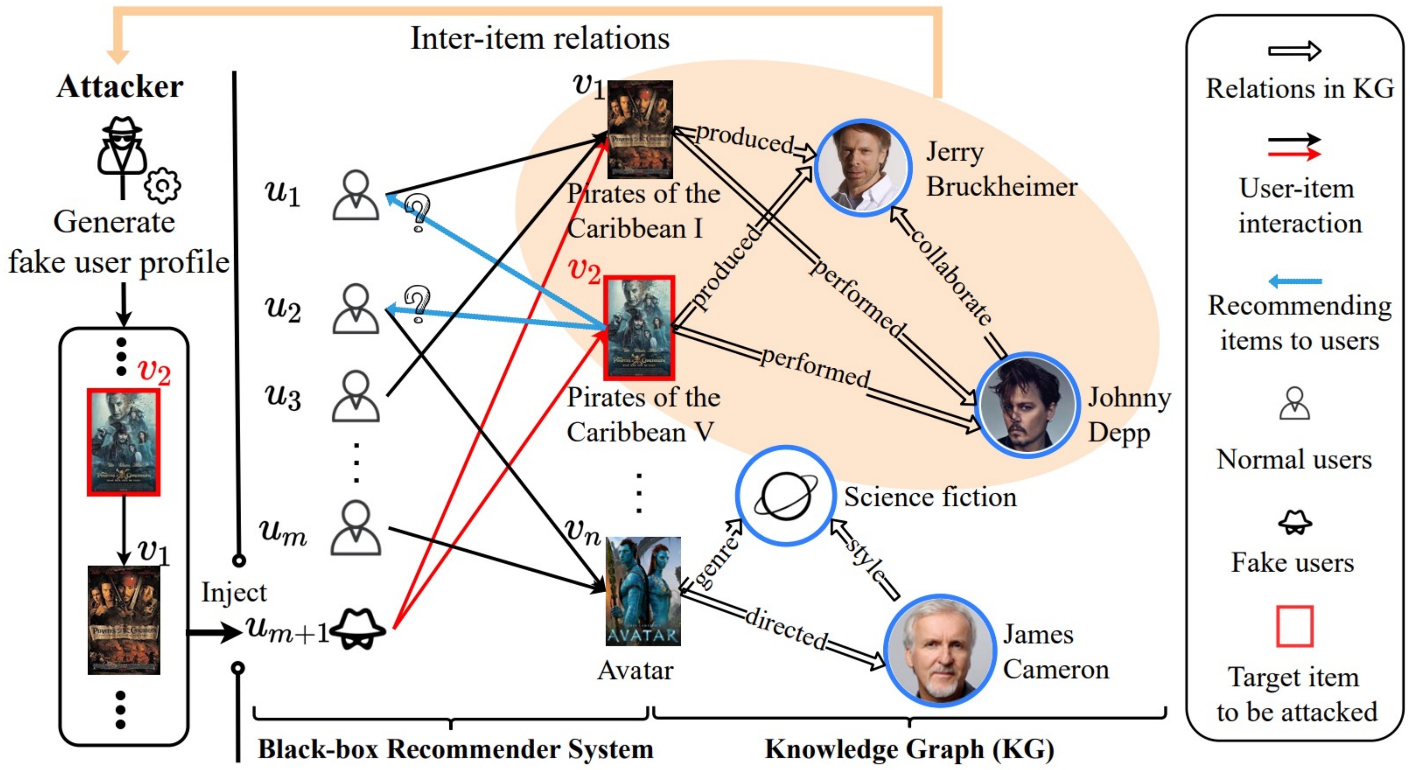
# KGAttack

- Side-information: Knowledge Graph (KG)
  - Rich auxiliary knowledge: relations among items and real-world entities
  - The underlying relationships between Target items and other items



# KGAttack

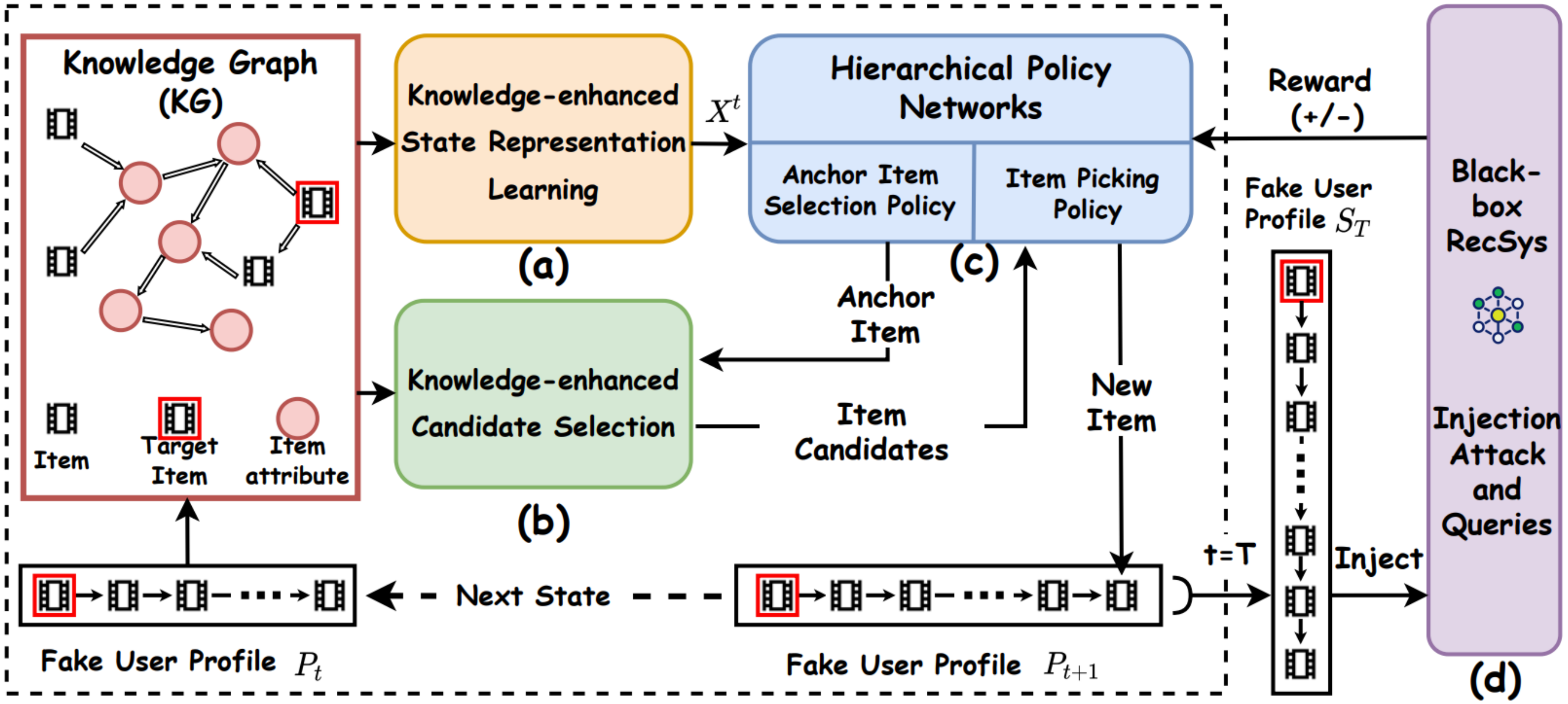
- Employs the KG to enhance the generation of fake user profiles from the massive item sets




Knowledge-enhanced Black-box Attacks for Recommendations, KDD 2022

# KGAttack

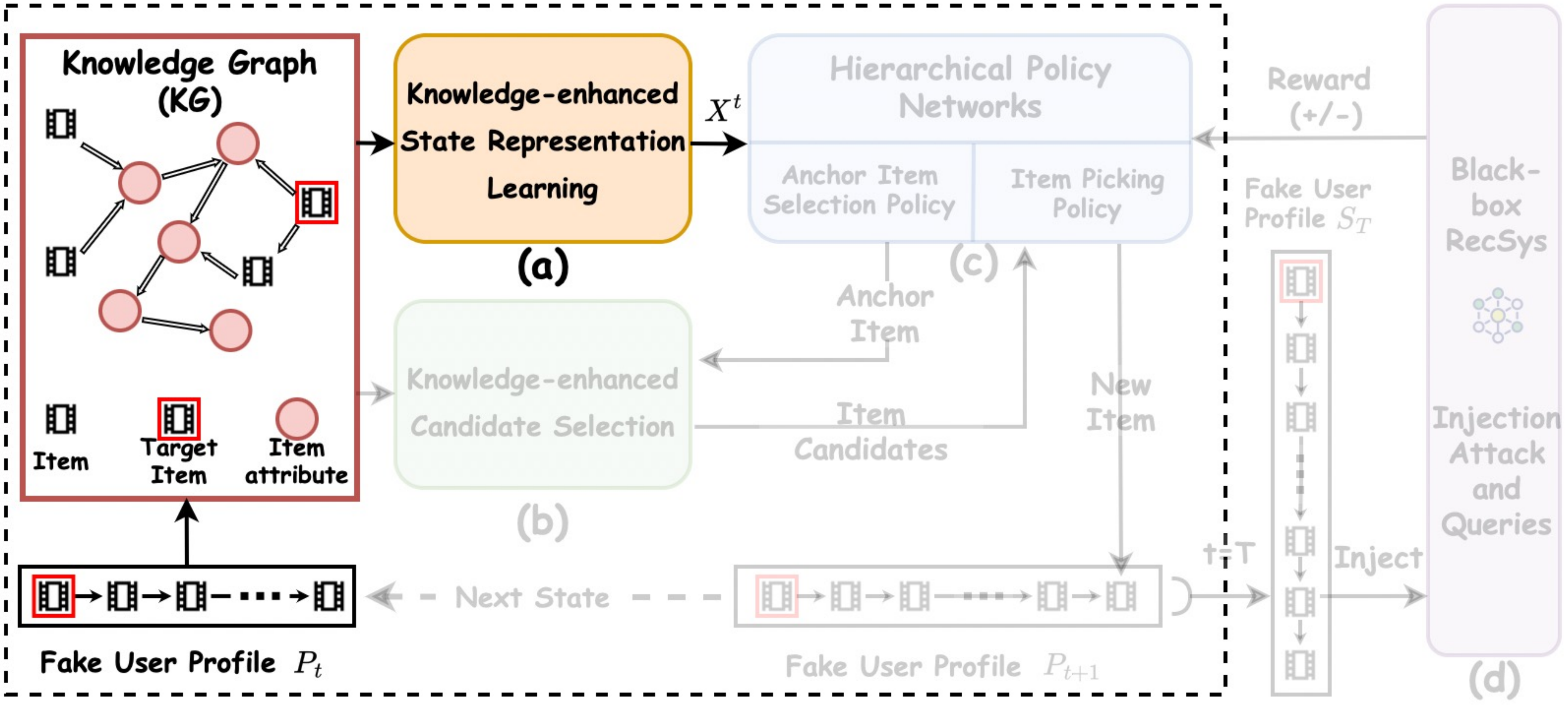
- Using KG to enhance the representation of state
- RL agent, generate user profiles




Knowledge-enhanced Black-box Attacks for Recommendations, KDD 2022

# KGAttack

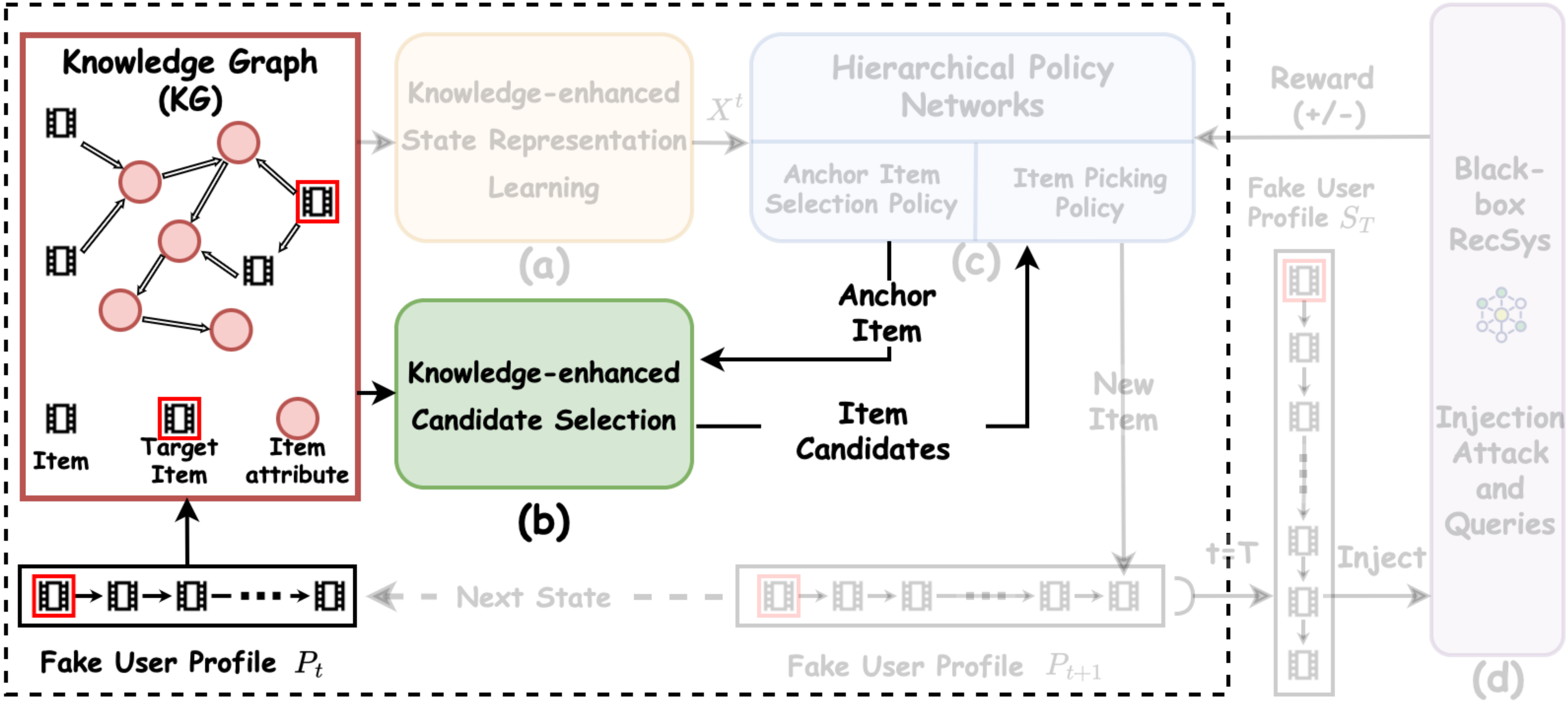
- (a): Using KG to enhance the representation of state




Knowledge-enhanced Black-box Attacks for Recommendations, KDD 2022

# KGAttack

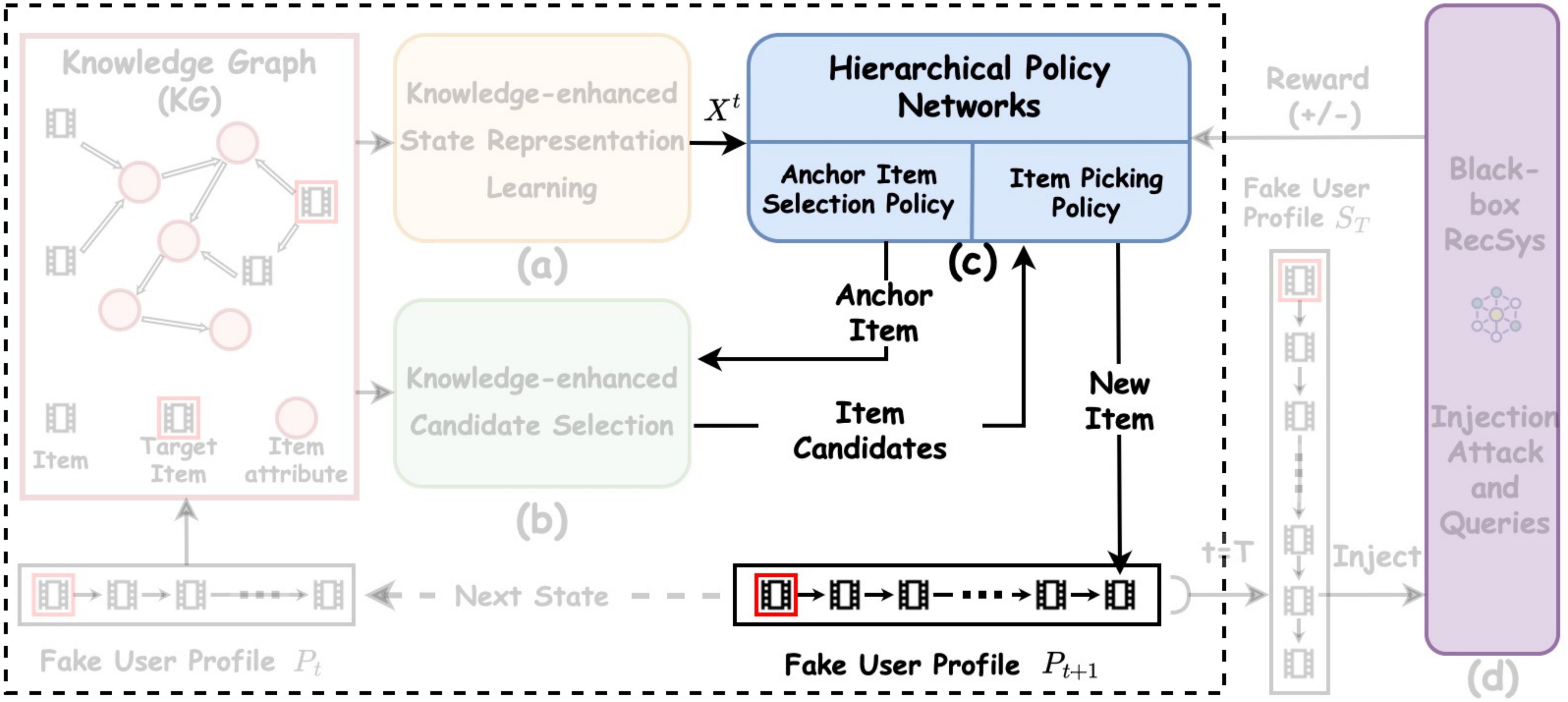
- (b): Using KG to localize relevant item candidates




Knowledge-enhanced Black-box Attacks for Recommendations, KDD 2022

# KGAttack

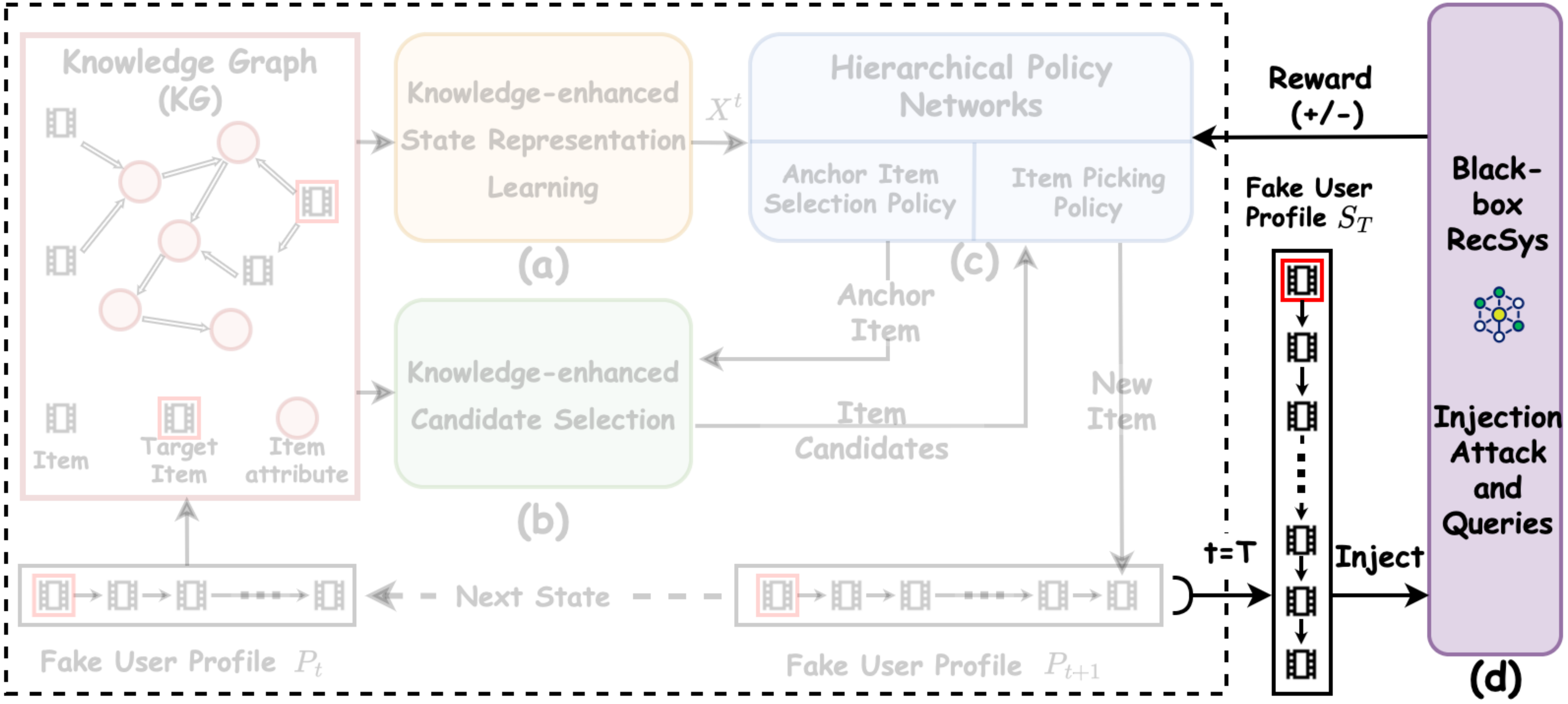
- (c): Using KG to localize relevant item candidates




Knowledge-enhanced Black-box Attacks for Recommendations, KDD 2022

# KGAttack

- (d): Injection attacks and query




Knowledge-enhanced Black-box Attacks for Recommendations, KDD 2022

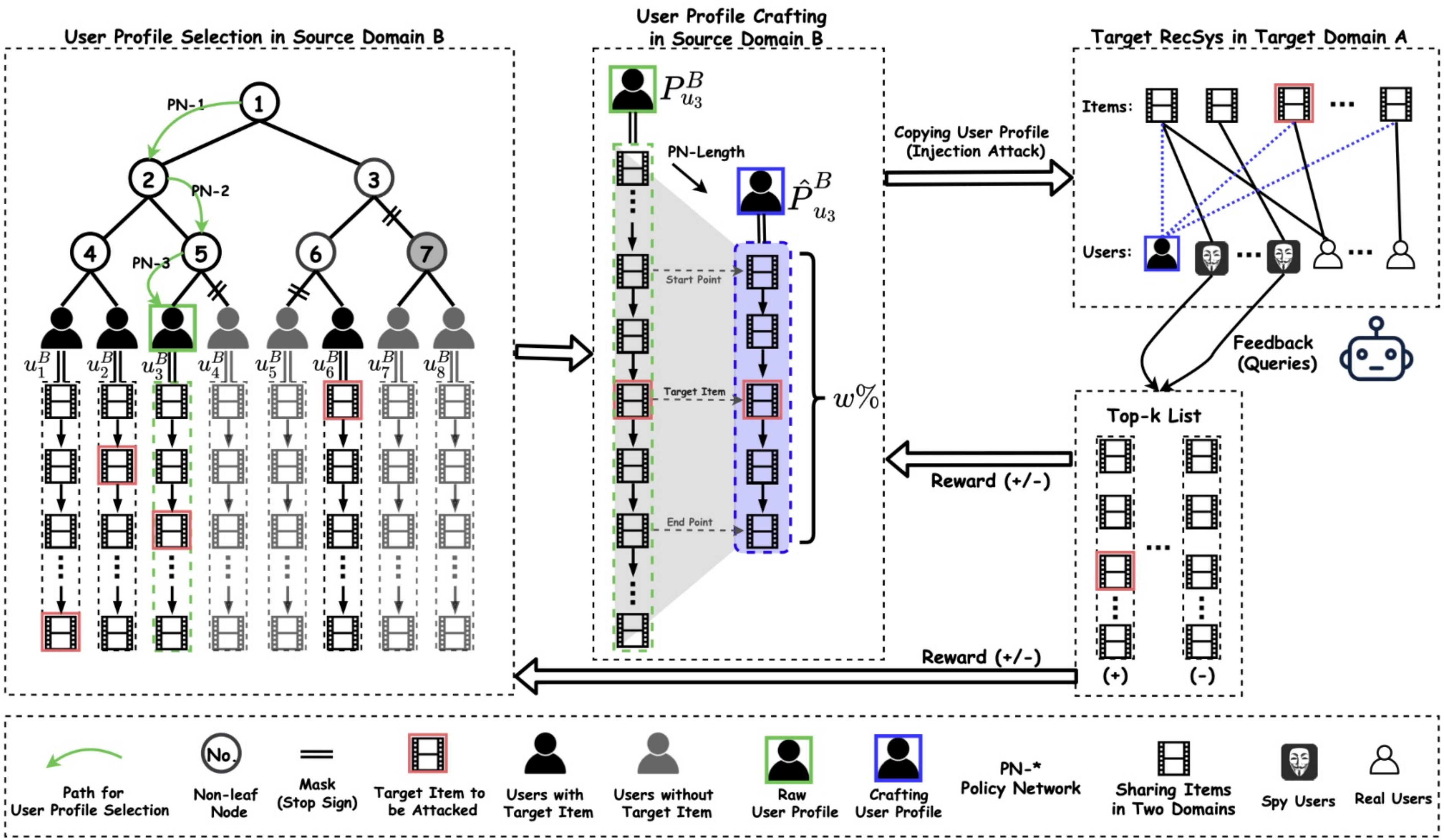
# CopyAttack

- Cross-domain Information
  - Share a lot of items
  - Users from these platforms with similar functionalities also share similar behavior patterns/preferences


Taobao JD.com


Attacking Black-box Recommendations via Copying Cross-domain User Profiles, ICDE 2021

# CopyAttack




Attacking Black-box Recommendations via Copying Cross-domain User Profiles, ICDE 2021

# CopyAttack

- User Profile Selection
  - Construct hierarchical clustering tree
  - Masking Mechanism - specific target items
  - Hierarchical-structure Policy Gradient

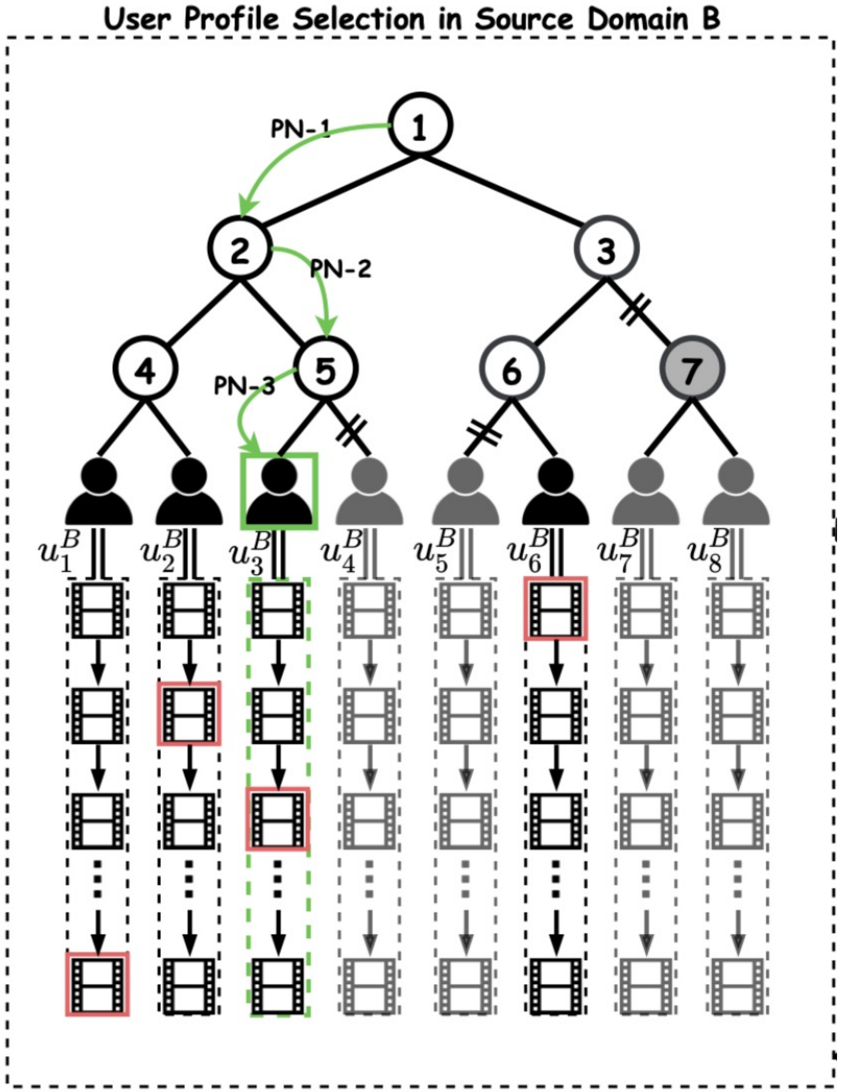
$$a_t^u = \left\{a_{[t,1]}^u, a_{[t,2]}^u, \ldots, a_{[t,d]}^u\right\}$$

$$p^u(a_t^u \mid s_t^u) = \prod_d^d p_d^u(a_t^u \mid \cdot, s_t^u)$$
$$= p_d^u\left(a_{[t,d]}^u \mid s_t^u\right) \cdot p_{d-1}^u\left(a_{[t,d-1]}^u \mid s_t^u\right) \cdots p_1^u\left(a_{[t,1]}^u \mid s_t^u\right)$$

$$\mathbf{x}_{v_*} = RNN\left(\mathcal{U}_t^{B \to A}\right)$$

$$p_i^u(\cdot \mid s_t^u) = \text{softmax}\left(MLP\left(\left[\mathbf{q}_{v_*}^B \oplus \mathbf{x}_{v_*}\right] \mid \theta_i^u\right)\right)$$

Time Complexity: $\mathcal{O}(|\mathcal{U}^B|) \longrightarrow \mathcal{O}\left(d \times |\mathcal{U}^B|^{1/d}\right)$



# CopyAttack

- User Profile Crafting
  - Clipping operation to craft the raw user profiles

$$W = \{10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$$

  - Sequential patterns (forward/backward)

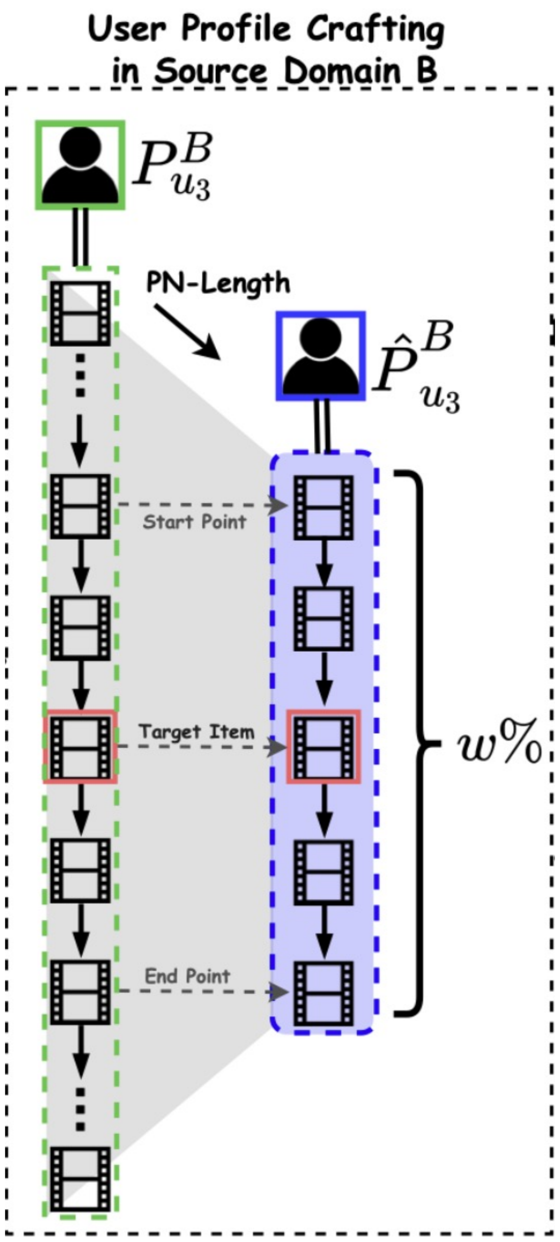
Example:

w = 50%

$$P^B_{u_i} = \{v_1 \rightarrow v_2 \boxed{\rightarrow v_3 \rightarrow v_4 \rightarrow v_{5*} \rightarrow v_6 \rightarrow v_7} \rightarrow v_8 \rightarrow v_9 \rightarrow v_{10}\}$$

$$\hat{P}^B_{u_i} = \{v_3 \rightarrow v_4 \rightarrow v_{5*} \rightarrow v_6 \rightarrow v_7\}$$

$$p^l(\cdot \mid s^l_t) = \mathrm{softmax}\big(MLP([\mathbf{p}^B_i \oplus \mathbf{q}^B_{v_*}] \mid \theta^l)\big)$$




Attacking Black-box Recommendations via Copying Cross-domain User Profiles, ICDE 2021

# Outline

- Concepts and Taxonomy
- Adversarial Attack
- **Adversarial Defense**
- Application
- Adversarial Learning Surveys and Tools
- Future directions

# Detection

- Exceptions and outliers in the recommendation system
  - Discrepancies between user’s ratings and item’s average ratings
  - Spectrum-based features of series rate values of each user
  - Cluster instances
  - User behaviors
  - The process of learning users and items representations
  - The distribution of normal users’ behaviors over a partial dataset
  - . . .

# Detection


Adversarial Defense
Detection
DegSim and RDMA
PPu and Du
TSGR, RSF, and TBR
...

# Detection

- Detection of shilling attacks in online recommender systems
  - Detecting Process:
    - Extract the supposed characteristics, DegSim and RDMA

Degree of similarity with Top Neighbors:

$$\mathrm{Degsim_u} = \frac{\sum_{\mathrm{v}=1}^{\mathrm{k}} \mathrm{W_{u,v}}}{\mathrm{k}}$$

Rating Deviation from Mean Agreement:

$$RDMA_j = \frac{\sum_{i=0}^{N_j} \frac{|r_{i,j} - Avg_i|}{NR_i}}{N_j}$$

Preventing shilling attacks in online recommender systems, WIDM 2005

# Detection

- Detection of shilling attacks via selecting patterns analysis
  - Detecting Process:
    - Extract the supposed characteristics, popularity profile and popularity distribution

A set of item popularity values of rated items: $PP_u = (d_{u,1}, d_{u,2}, \ldots, d_{u,N_u})$

Popularity distribution: $D_u = (p_{u,1}, p_{u,2}, \ldots, p_{u,d_{max}})$

Shilling Attack Detection in Recommender Systems via Selecting Patterns Analysis, IEICE 2016

# Detection

- Detection of trust shilling attacks in recommender systems
  - Detecting Process:
    - Extract the supposed characteristics, TSGR, RSF, and TBR

$$\boxed{TSGR_i} = \frac{tg_i \cap tr_i}{tg_i \cup tr_i}$$

User $i$'s trust similarity between trust givers and trust receivers

$$\boxed{PTBR_u} = \frac{pn_u}{N_u}$$

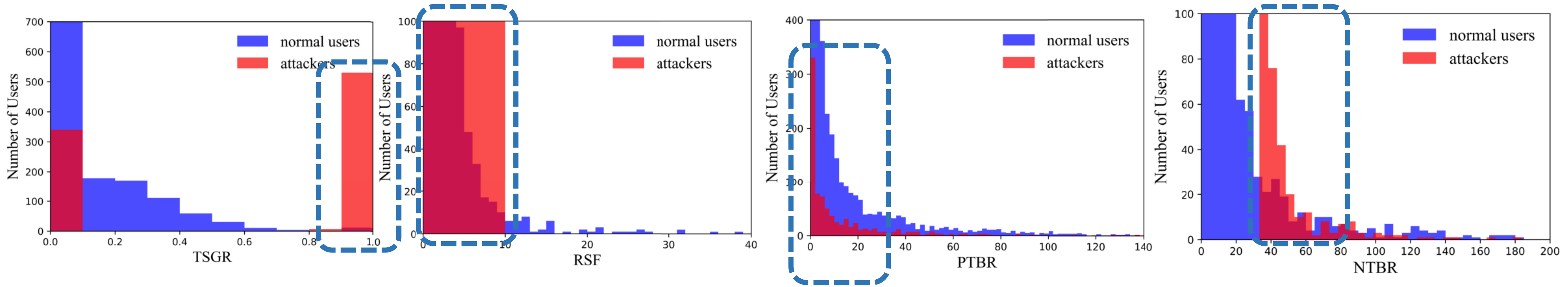
Positive Trust Behavior Ratio

$$\boxed{NTBR_u} = \frac{nn_u}{N_u}$$

Negative Trust Behavior Ratio

# Detection

- Normal vs. attackers distributions for each feature:



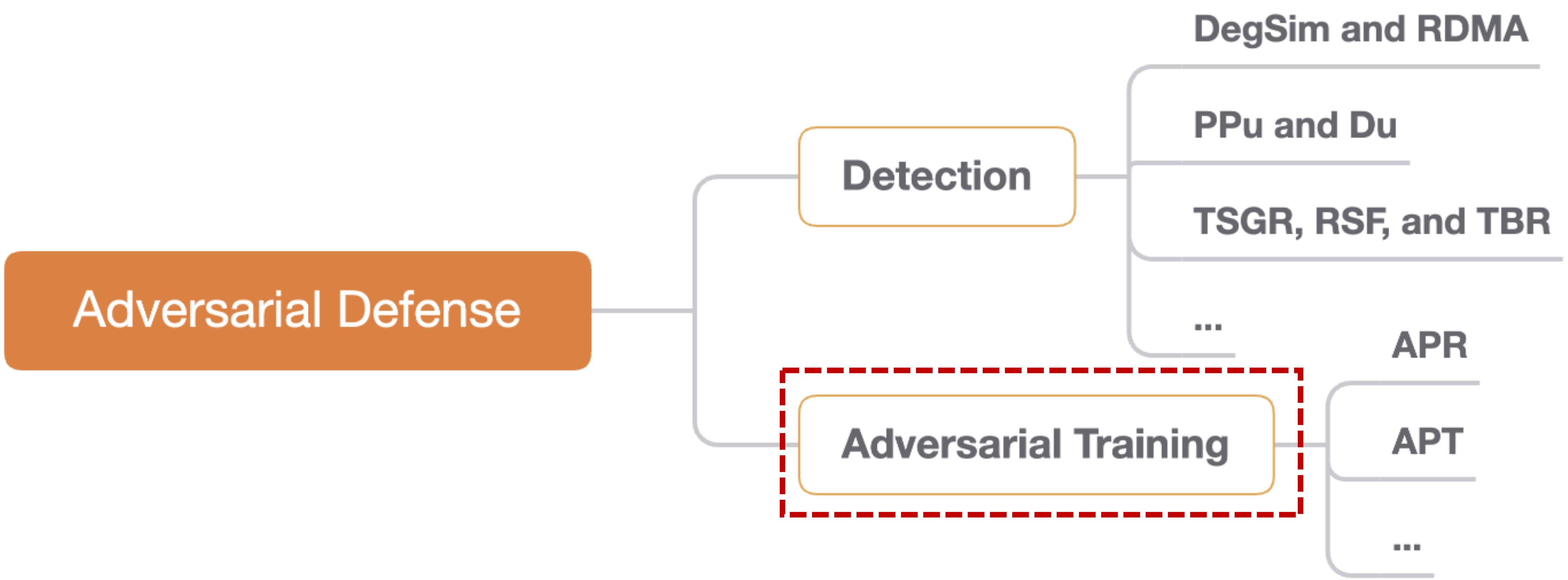

Detection of Trust Shilling Attacks in Recommender Systems, IEICE 2022

# Adversarial Training

- Adversarial training contains two alternating processes:
  - Generating perturbations that can confuse a recommendation model
  - Training the recommendation model along with generated perturbations

$$\min_{\theta} \max_{\eta} \mathcal{L}(\mathcal{X} + \eta, \theta)$$

# Adversarial Training



# Adversarial Training

- Adversarial Personalized Ranking (APR)

Optimization objectives against noise:

$$\Delta_{adv} = \arg \max_{\Delta, ||\Delta|| \leq \epsilon} L_{BPR}(\mathcal{D}|\hat{\Theta} + \Delta)$$

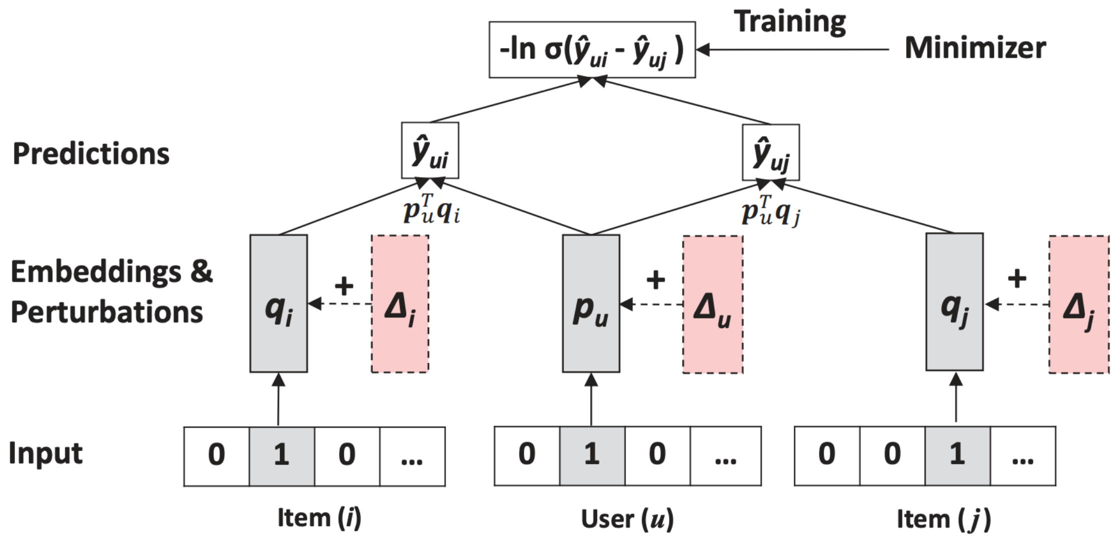
Adversarial Personalized Ranking (APR):

$$\mathrm{L}_{\mathrm{APR}}(\mathcal{D} \mid \Theta) = \mathrm{L}_{\mathrm{BPR}}(\mathcal{D} \mid \Theta) + \lambda \mathrm{L}_{\mathrm{BPR}}(\mathcal{D} \mid \Theta + \Delta_{\mathrm{adv}})$$

$$\text{where } \Delta_{\mathrm{adv}} = \arg \max_{\Delta, \|\Delta\| \leq \epsilon} \mathrm{L}_{\mathrm{BPR}}(\mathcal{D} \mid \hat{\Theta} + \Delta)$$

The training process of APR:

$$\Theta^*, \Delta^* = \arg \min_{\Theta} \max_{\Delta, ||\Delta|| \leq \epsilon} L_{BPR}(\mathcal{D}|\Theta) + \lambda L_{BPR}(\mathcal{D}|\Theta + \Delta)$$



# Adversarial Training

- Adversarial poisoning training (APT)

$$\min_{\theta_R} \min_{\boxed{\mathcal{D}^*}, |\mathcal{D}^*|=n^*} \mathcal{L}(\mathcal{D} \cup \mathcal{D}^*, \theta_R)$$

$D^* = \{r_1^*, \ldots, r_{n*}^*\}$ is a set of $n*$ fake users dedicated to minimizing the empirical risk.

**Algorithm 1:** Adversarial Poisoning Training

**Input:** The epochs of training $T$, pre-training $T_{pre}$, and poisoning interval $T_{inter}$.

```
Randomly initialize the user set D* defined in Definition 3.1.   ①
for T_pre epochs do
    Do standard training on the dataset D;   ②
end
D' = D;
for T − T_pre epochs do
    for per T_inter epochs do
        Calculate the influence vector I according to Eq. 5;   ③
        for each ERM user in D* do
            Select m* items in Φ with probability
            exp(−tI_i) / Σ_{j∈Φ} exp(−tI_j) and rate the selected items with
            normal distribution (μ_i + r+, σ_i) at random;   ④
        end
        D' = D ∪ D*;   ⑤
    end
    Do standard training on the dataset D';
end
```

# Summary

- **Adversarial Recommender System**
  - Adversarial Attack
    - **Heuristic Attack**
      - Random attack
      - Average attack
      - Bandwagon attack
      - segment attack
      - ...
    - **Gradient-based Attack**
      - UnAttack
      - S-Attack
      - Graph-based Attack
      - ...
    - **RL-based Attack**
      - PoisonRec
      - KGAttack
      - CopyAttack
      - ...
  - Adversarial Defense
    - **Detection**
      - DegSim and RDMA
      - PPu and Du
      - TSGR, RSF, and TBR
      - ...
    - **Adversarial Training**
      - APR
      - APT
      - ...

# Outline

- Concepts and Taxonomy
- Adversarial Attack
- Adversarial Defense
- Application
- Adversarial Learning Surveys and Tools
- Future directions

# Application

- The application of adversarial training can help improve the trustworthiness and reliability of recommendation systems in various domains, including:
  - E-health recommendation
  - E-commercial recommendation
  - . . .

# Outline

1. Concepts and Taxonomy
2. Adversarial Attack
3. Adversarial Defense
4. Application
5. Adversarial Learning Surveys and Tools
6. Future directions

# Adversarial Learning Surveys

- Attack:
  - Zhang, Fuguo. "A survey of shilling attacks in collaborative filtering recommender systems." 2009
  - Gunes, Ihsan, et al. "Shilling attacks against recommender systems: A comprehensive survey." 2014
  - Si, Mingdan, and Qingshan Li. "Shilling attacks against collaborative recommender systems: a review." 2020

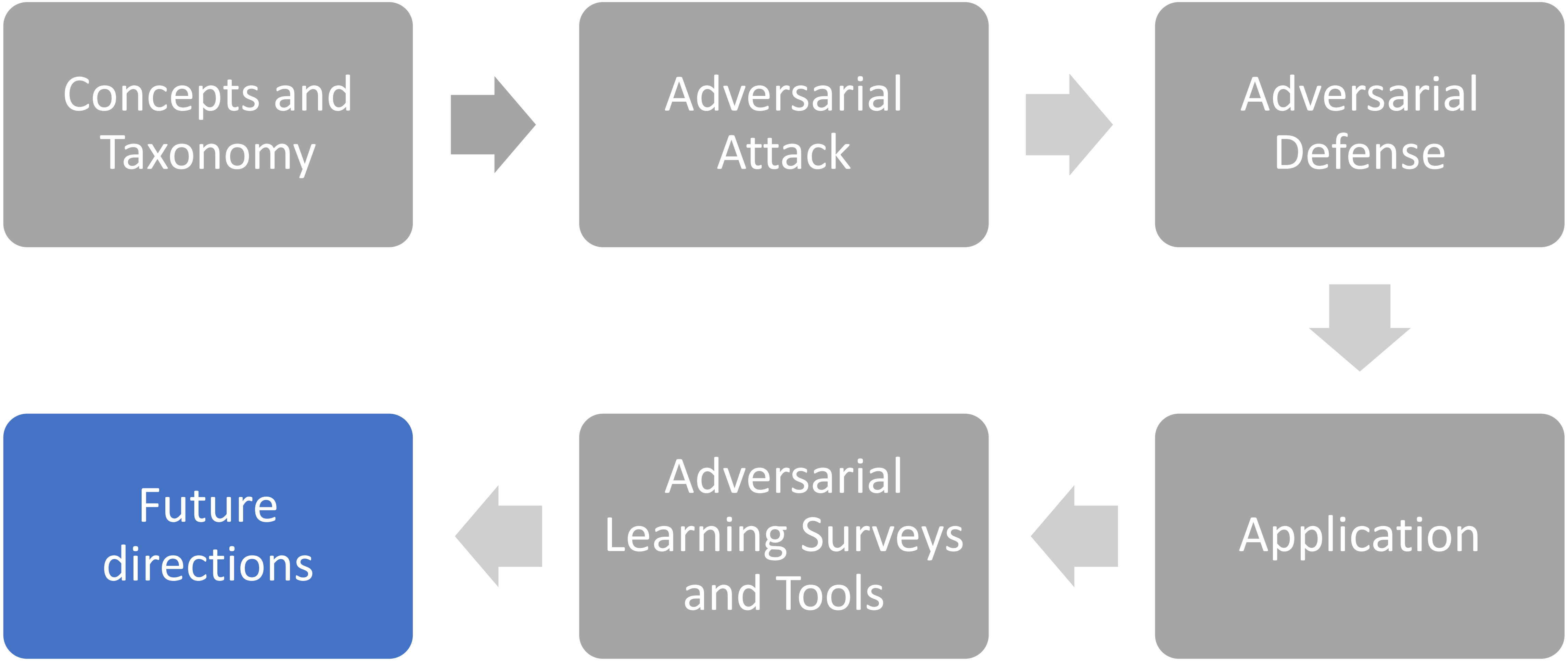
- Adversarial recommender systems:
  - Truong, Anh, Negar Kiyavash, and Seyed Rasoul Etesami. "Adversarial machine learning: The case of recommendation systems." 2018
  - Deldjoo, Yashar, Tommaso Di Noia, and Felice Antonio Merra. "A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks." 2021

# Adversarial Learning Tools

- RGRecSys (Ovaisi et al., 2022)

RGRecSys: A Toolkit for Robustness Evaluation of Recommender Systems, Ovaisi et al 2022.

# Outline


Concepts and Taxonomy
Adversarial Attack
Adversarial Defense
Application
Adversarial Learning Surveys and Tools
Future directions

# Future Directions

- Investigate vulnerability of different recommender systems
- Generate adversarial perturbations on user-item interactions for adversarial robust training
- Address open problems and challenges in robustness in recommendation

# Trustworthy Recommender Systems

**Introduction**

Wenqi Fan

**Non-discrimination & Fairness**

Xiao Chen

**Safety & Robustness**

Shijie Wang

**Explainability**

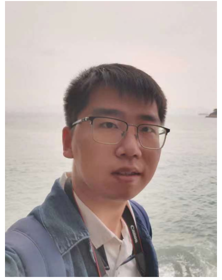
Jingtong Gao

**Privacy**

Lin Wang

**Environmental Well-being**

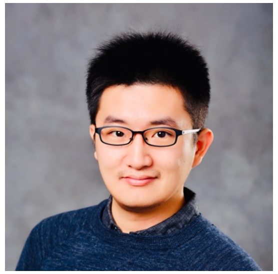
**Accountability & Auditability**



Qidong Liu

**Dimension Interactions**

**Future Directions**



Xiangyu Zhao

# Trustworthy Recommender Systems

**Introduction**

Wenqi Fan

**Non-discrimination & Fairness**

Xiao Chen

**Safety & Robustness**

Shijie Wang

**Explainability**

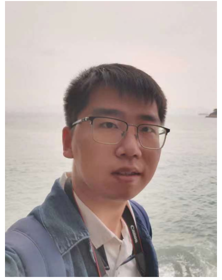
Jingtong Gao

**Privacy**

Lin Wang

**Environmental Well-being**

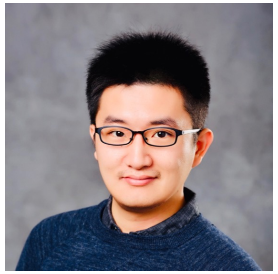
**Accountability & Auditability**



Qidong Liu

**Dimension Interactions**

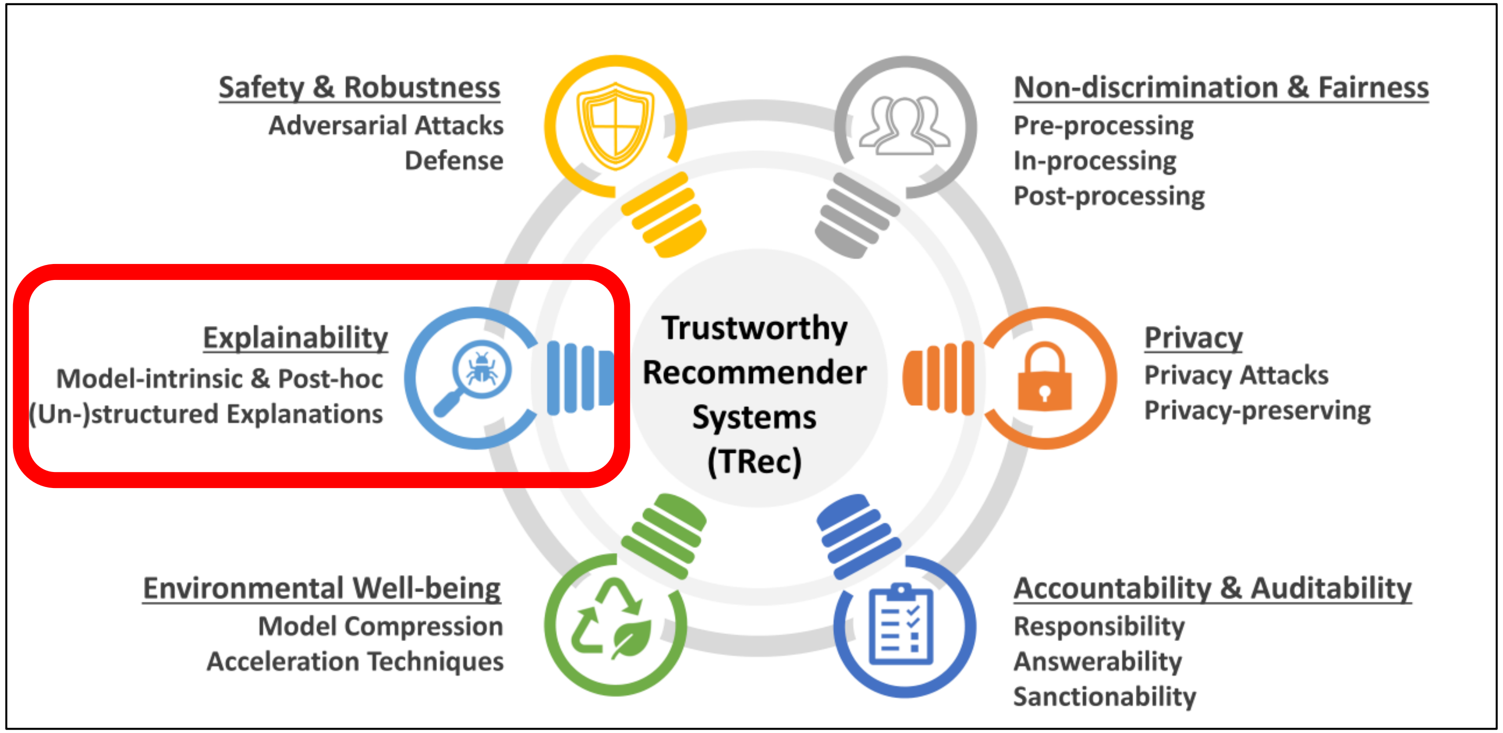
**Future Directions**



Xiangyu Zhao

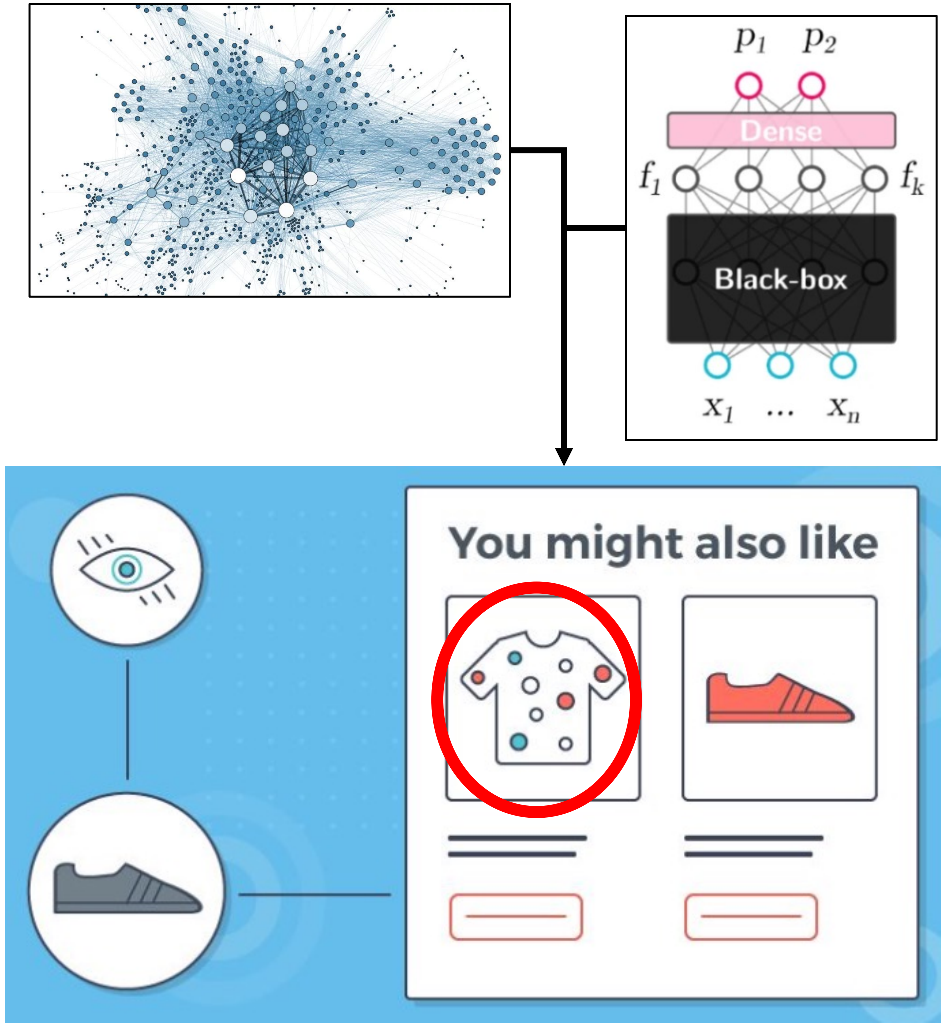
# Explainability

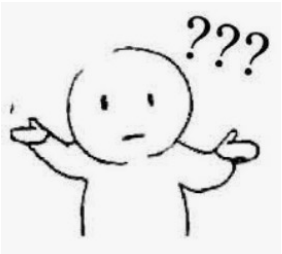
- **What's explainability in Rec, or to say explainable recommendations?**
  - It refers to the recommendation algorithms focusing on providing explanation for recommendation results



# Explainability

- **Why do we need explainability in a trustworthy Rec system?**
  - Complicated modeling & Black-box module:

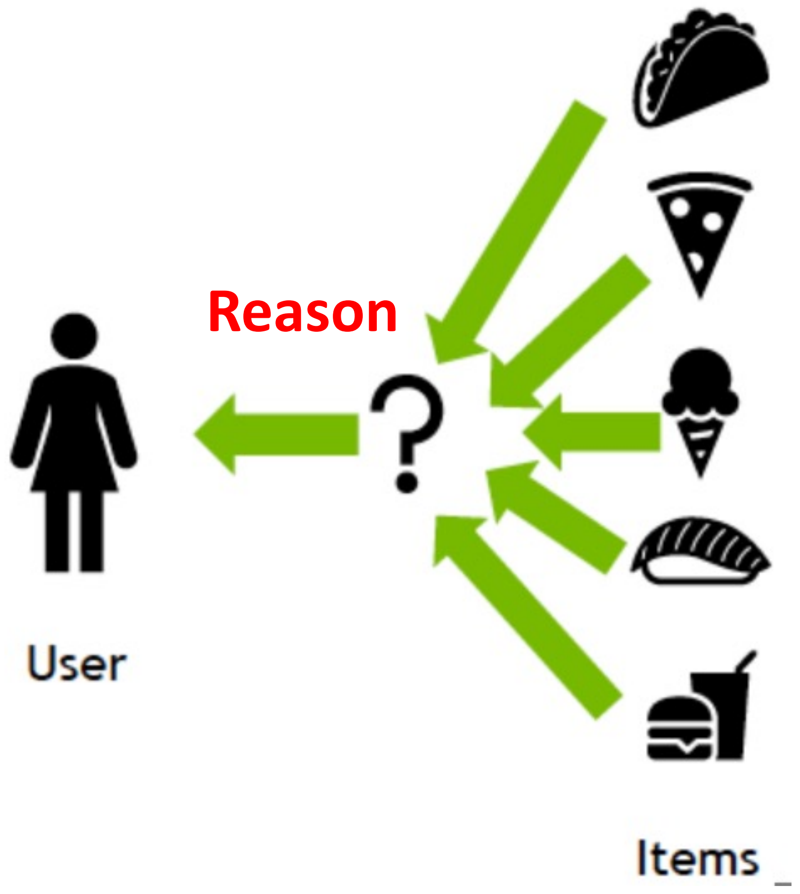



  - Why would you recommend this to me?
  - Similar style, same brand,
    or just a mis-recommendation?

# Concepts

- **The ability to explain or to present in understandable terms to a human**



# Explainability

**METHODS**

EVALUATIONS

APPLICATIONS

FUTURE DIRECTIONS

# Taxonomy

- **How to produce explanations: model-intrinsic based (mostly used) or post-hoc**

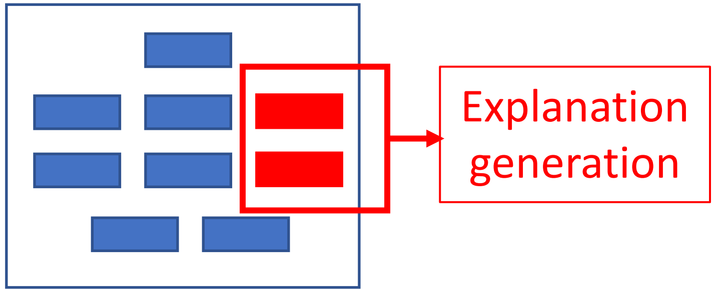
- **How the explanations are presented: structured or unstructured**

| | **Model-intrinsic based** | **Post-Hoc** | ***Characteristics*** |
|---|---|---|---|
| **Structured** | [48, 114, 364, 389, 390, 396] | [280, 319] | Logical, Visible |
| **Unstructured** | [63, 64, 291] | [211, 315, 338] | Diversified, Fragmented |
| ***Focus*** | Model's reasoning process | Instances' relationship | - |

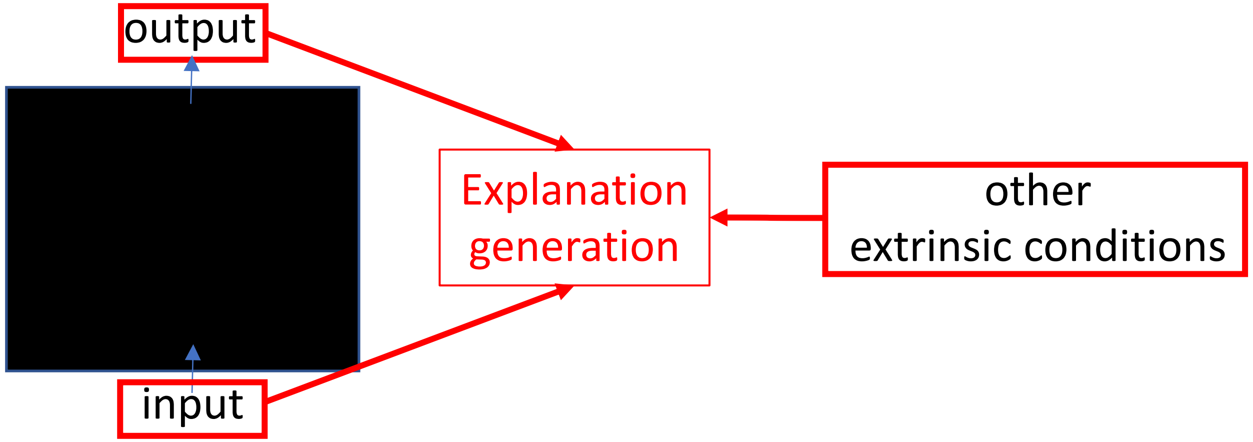
Note: Since some studies construct models from multiple perspectives at the same time, these different classifications are not completely antithetical

# Taxonomy

- **The first criteria: How to produce explanations**
  - Model-intrinsic based methods: seek to derive explanations from the **intrinsic structure** of the model

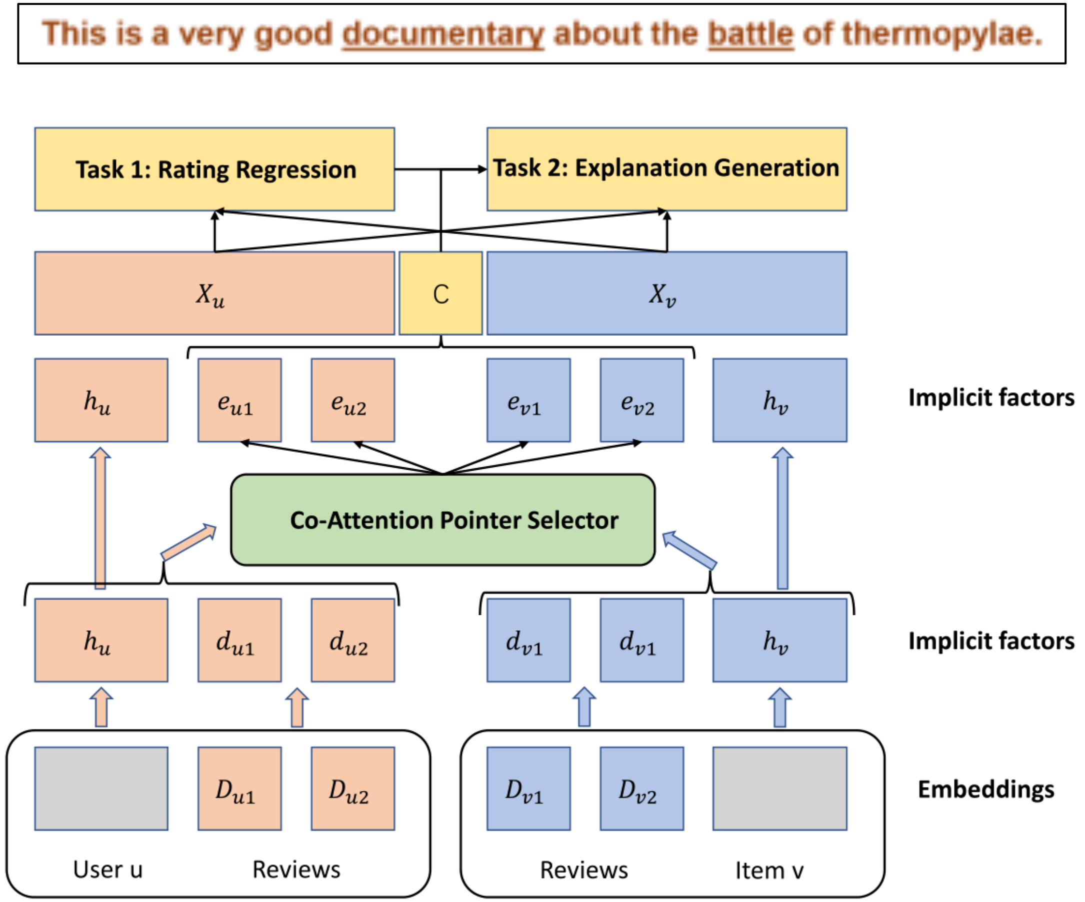



  - Post-hoc methods: provide explanations based only on the inputs, outputs and extrinsic conditions of the model



# Model-intrinsic based methods

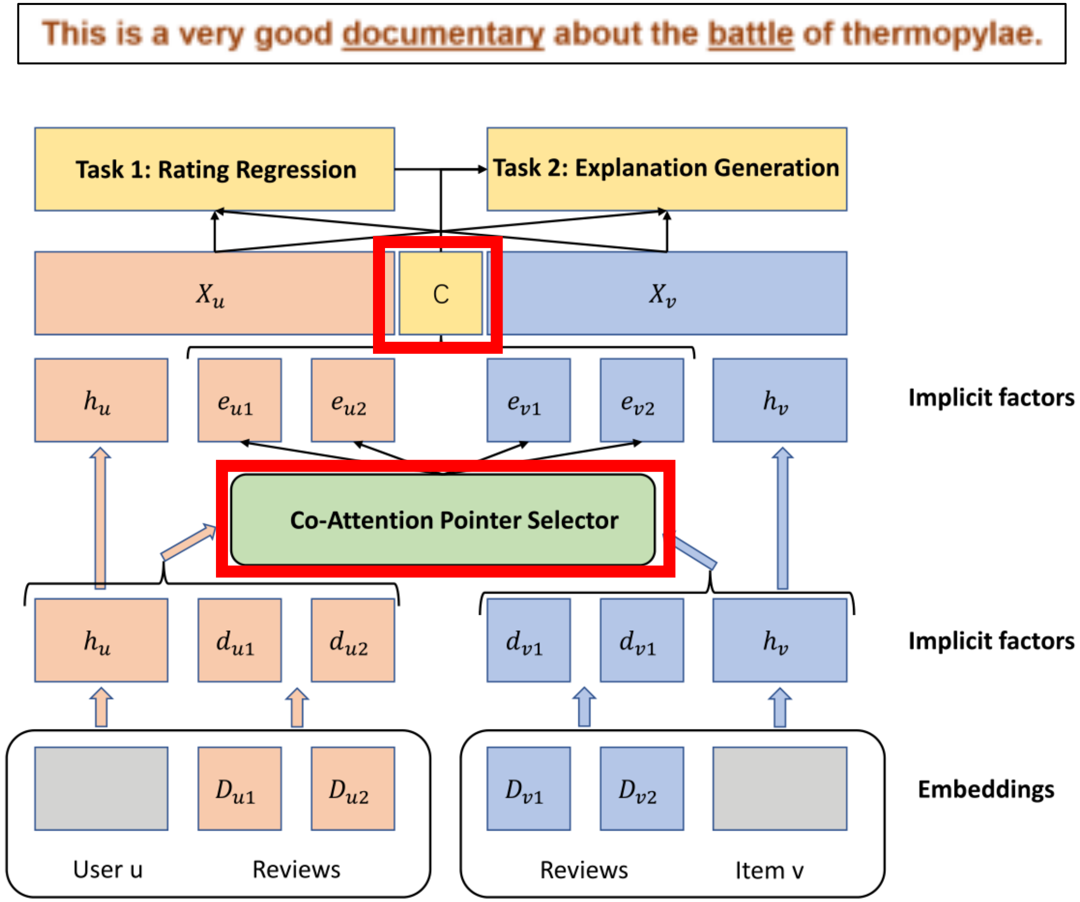
- **CAML**
  - The explanation is one of the major tasks and modeling goals
  - Only effective for the embedded models and cannot simply be reused in other models




[1] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-Attentive Multi-Task Learning for Explainable Recommendation.. In IJCAI. 2137–2143.

# Model-intrinsic based methods

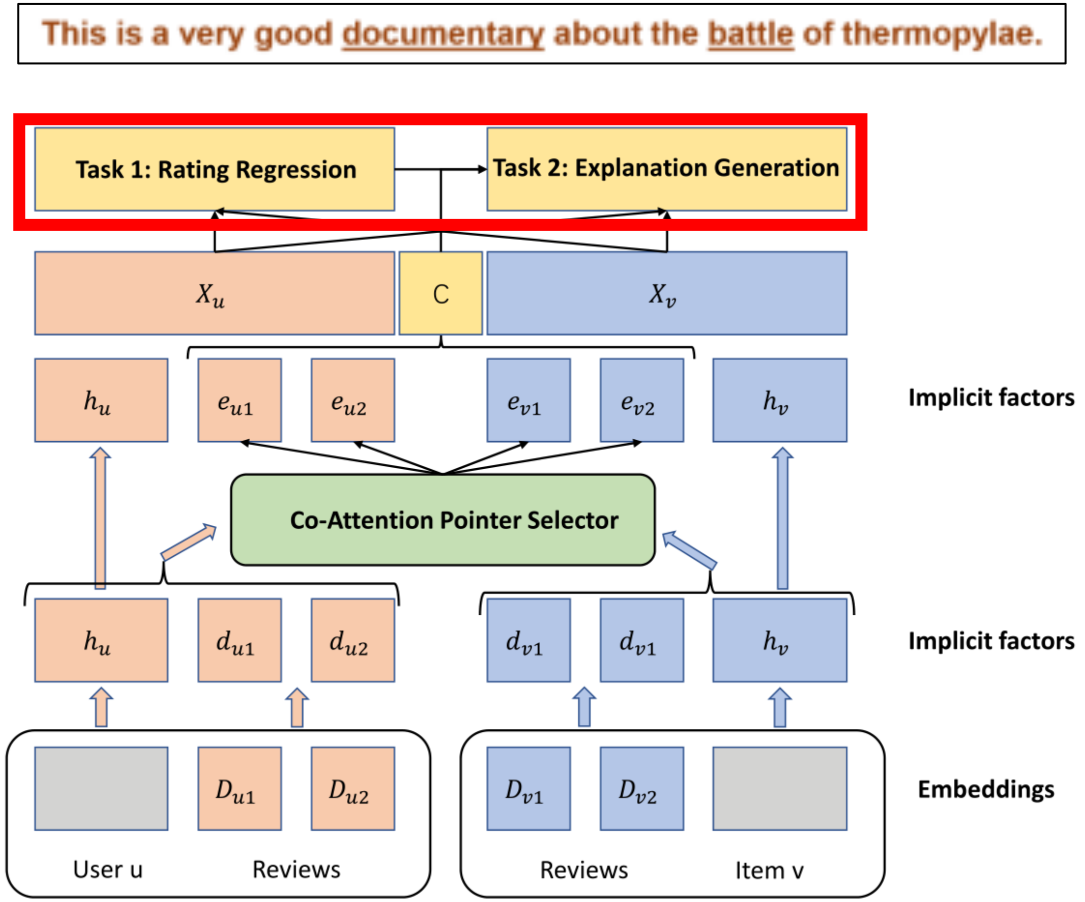
- **CAML**
  - The explanation is one of the major tasks and modeling goals
  - Only effective for the embedded models and cannot simply be reused in other models




[1] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-Attentive Multi-Task Learning for Explainable Recommendation.. In IJCAI. 2137–2143.

# Model-intrinsic based methods

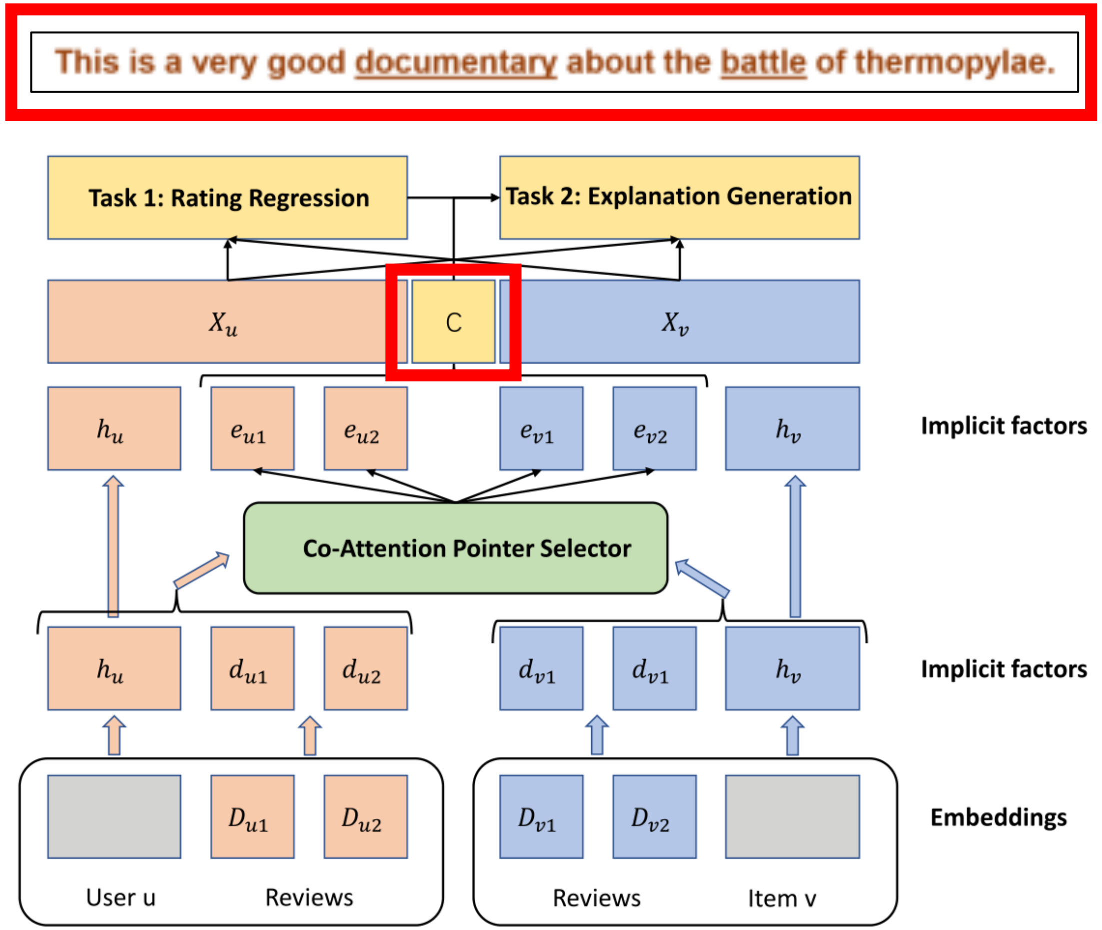
- **CAML**
  - The explanation is one of the major tasks and modeling goals
  - Only effective for the embedded models and cannot simply be reused in other models




[1] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-Attentive Multi-Task Learning for Explainable Recommendation.. In IJCAI. 2137–2143.

# Model-intrinsic based methods

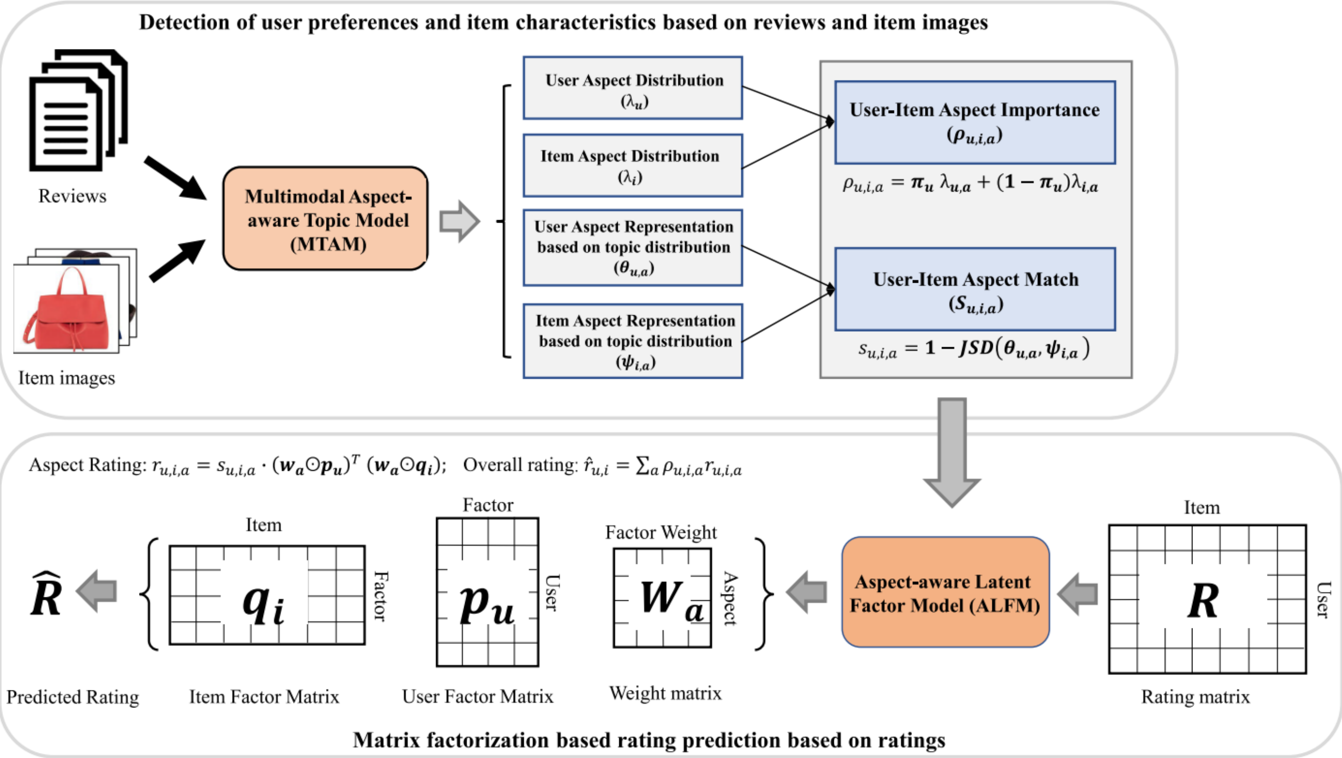
- **CAML**
  - The explanation is one of the major tasks and modeling goals
  - Only effective for the embedded models and cannot simply be reused in other models




[1] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-Attentive Multi-Task Learning for Explainable Recommendation.. In IJCAI. 2137–2143.

# Model-intrinsic based methods

- **MMALFM**




| | | |
|---|---|---|
| User_2397 | Food | sauce, fried, bread, fresh, huge, flavor, shrimp, dessert, dish |
| | Ambience | nice, bar, atmosphere, location, friendly, inside, decor, staff, music |
| | Price | expensive, high, cheap, pricey, decent, pay, reasonable, priced, deal |
| | Service | table, server, friendly, minutes, nice, staff, asked, make, seated |
| | Misc. | never, give, restaurant, times, stars, friends, night, places, dinner |
| Item_137 | Food | sauce, salad, fries, dish, cheese, dishes, burger, fresh, crab |
| | Ambience | bar, atmosphere, patio, area, inside, wine, small, cool, decor |
| | Price | price, worth, prices, better, bit, meal, sauce, dishes, quality |
| | Service | table, bar, friendly, wait, server, staff, minutes, beer, atmosphere |
| | Misc. | eat, dinner, Vegas, experience, wait, friends, times, never, give |
| Item_673 | Food | nigiri, sake, tempura, shrimp, sauce, items, poke, crab, chef |
| | Ambience | atmosphere, friendly, bar, staff, inside, area, spot, monta, feel |
| | Price | price, worth, prices, nigiri, sake, tempura, items, lunch, special |
| | Service | service, table, server, friendly, minutes, staff, nice, asked, seated |
| | Misc. | restaurant, times, give, favorite, night, places, stars, friends, Vegas |

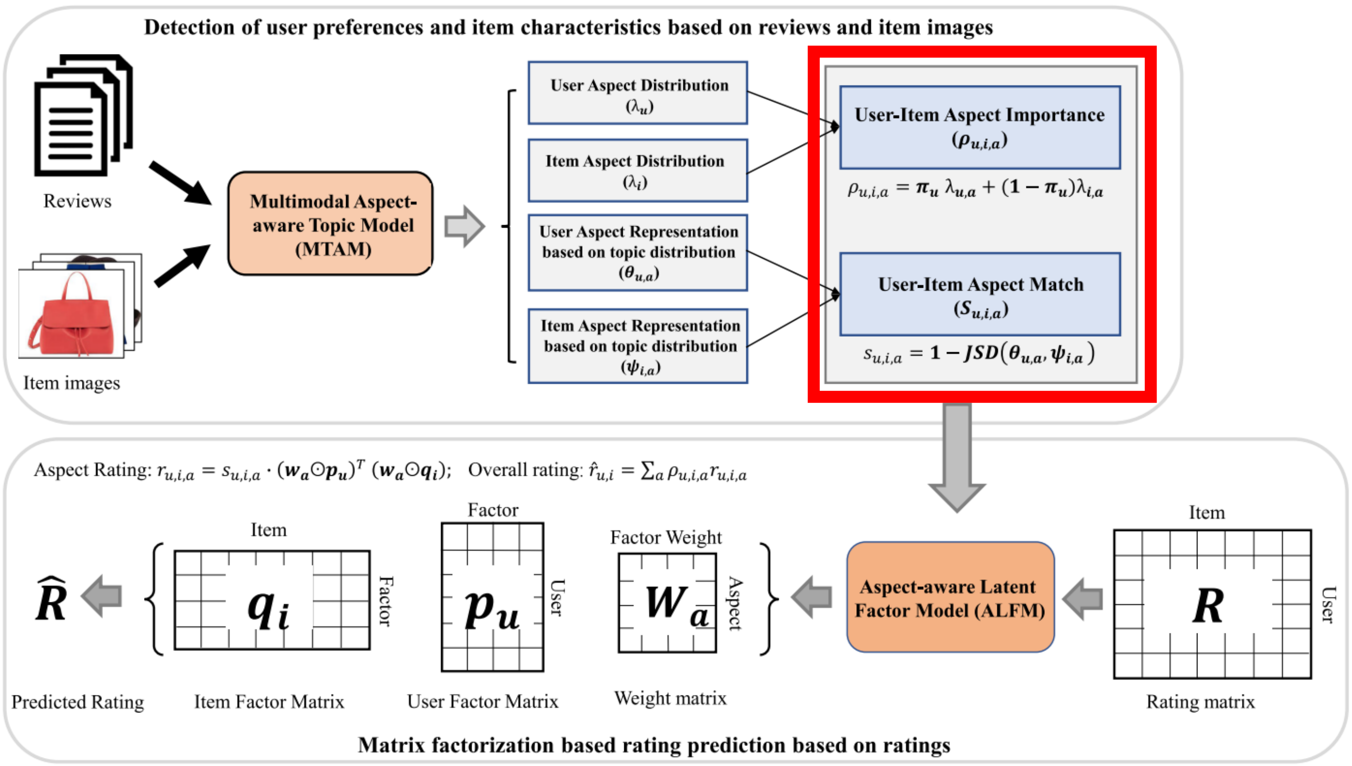
Table 6. Interpretation for Why the "User 2397" Rated "Item 137" and "Item 673" with 5 and 2, Respectively

| Item | Aspect | Food | Ambience | Price | Service | Misc. |
|---|---|---|---|---|---|---|
| Item_137 | Importance | 0.3815 | 0.1034 | 0.0723 | 0.2038 | 0.2390 |
| | Matching | 0.5672 | 0.4523 | 0.5329 | 0.6021 | 0.7138 |
| | Polarity | + | + | − | + | + |
| Item_673 | Importance | 0.3726 | 0.0794 | 0.0853 | 0.2076 | 0.2551 |
| | Matching | 0.1813 | 0.6535 | 0.4512 | 0.6018 | 0.7093 |
| | Polarity | − | − | + | + | − |


[1] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. ACM Transactions on Information Systems (TOIS) 37, 2 (2019), 1–28.

# Model-intrinsic based methods

- **MMALFM**




| | | |
|---|---|---|
| User_2397 | Food | sauce, fried, bread, fresh, huge, flavor, shrimp, dessert, dish |
| | Ambience | nice, bar, atmosphere, location, friendly, inside, decor, staff, music |
| | Price | expensive, high, cheap, pricey, decent, pay, reasonable, priced, deal |
| | Service | table, server, friendly, minutes, nice, staff, asked, make, seated |
| | Misc. | never, give, restaurant, times, stars, friends, night, places, dinner |
| Item_137 | Food | sauce, salad, fries, dish, cheese, dishes, burger, fresh, crab |
| | Ambience | bar, atmosphere, patio, area, inside, wine, small, cool, decor |
| | Price | price, worth, prices, better, bit, meal, sauce, dishes, quality |
| | Service | table, bar, friendly, wait, server, staff, minutes, beer, atmosphere |
| | Misc. | eat, dinner, Vegas, experience, wait, friends, times, never, give |
| Item_673 | Food | nigiri, sake, tempura, shrimp, sauce, items, poke, crab, chef |
| | Ambience | atmosphere, friendly, bar, staff, inside, area, spot, monta, feel |
| | Price | price, worth, prices, nigiri, sake, tempura, items, lunch, special |
| | Service | service, table, server, friendly, minutes, staff, nice, asked, seated |
| | Misc. | restaurant, times, give, favorite, night, places, stars, friends, Vegas |

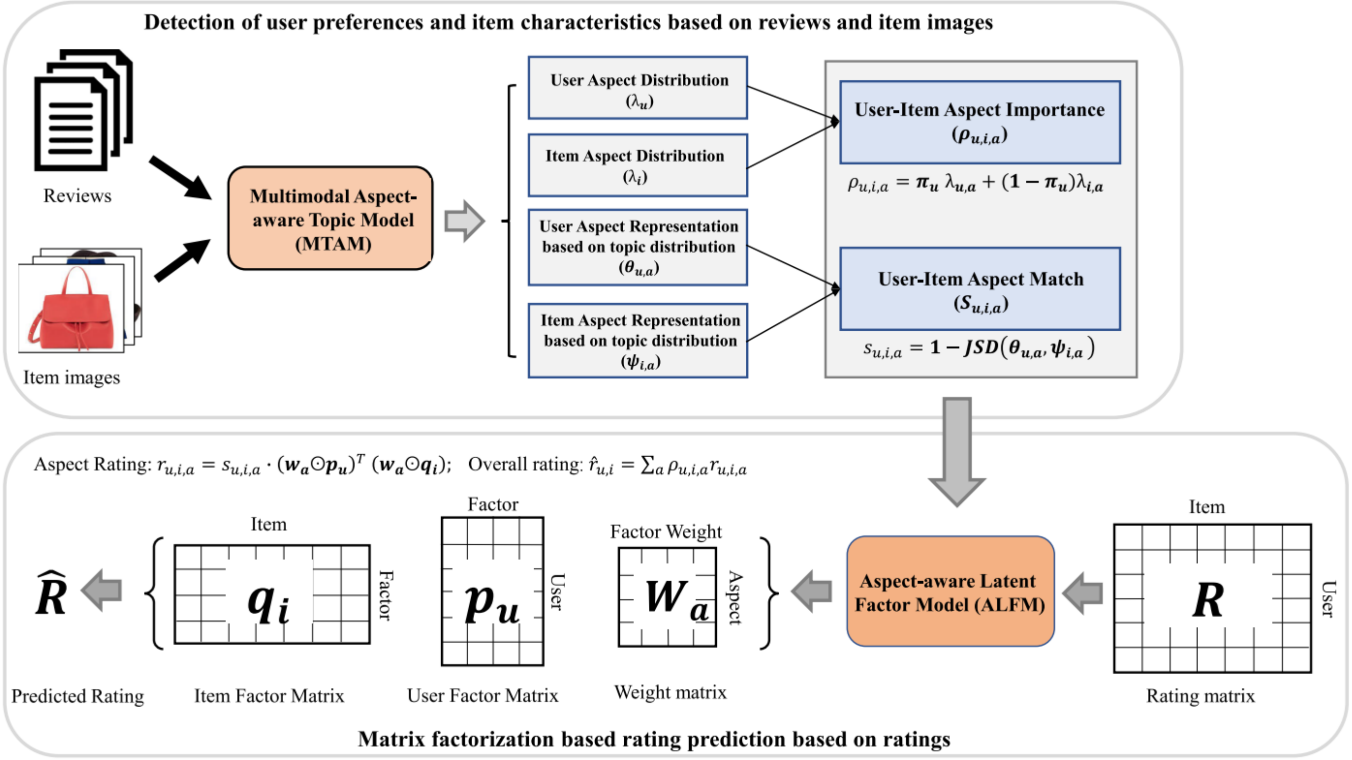
Table 6. Interpretation for Why the "User 2397" Rated "Item 137" and "Item 673" with 5 and 2, Respectively

| Item | Aspect | Food | Ambience | Price | Service | Misc. |
|---|---|---|---|---|---|---|
| Item_137 | Importance | 0.3815 | 0.1034 | 0.0723 | 0.2038 | 0.2390 |
| | Matching | 0.5672 | 0.4523 | 0.5329 | 0.6021 | 0.7138 |
| | Polarity | + | + | − | + | + |
| Item_673 | Importance | 0.3726 | 0.0794 | 0.0853 | 0.2076 | 0.2551 |
| | Matching | 0.1813 | 0.6535 | 0.4512 | 0.6018 | 0.7093 |
| | Polarity | − | − | + | + | − |

[1] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. ACM Transactions on Information Systems (TOIS) 37, 2 (2019), 1–28.

# Model-intrinsic based methods

- **MMALFM**




| | | |
|---|---|---|
| User_2397 | Food | sauce, fried, bread, fresh, huge, flavor, shrimp, dessert, dish |
| | Ambience | nice, bar, atmosphere, location, friendly, inside, decor, staff, music |
| | Price | expensive, high, cheap, pricey, decent, pay, reasonable, priced, deal |
| | Service | table, server, friendly, minutes, nice, staff, asked, make, seated |
| | Misc. | never, give, restaurant, times, stars, friends, night, places, dinner |
| Item_137 | Food | sauce, salad, fries, dish, cheese, dishes, burger, fresh, crab |
| | Ambience | bar, atmosphere, patio, area, inside, wine, small, cool, decor |
| | Price | price, worth, prices, better, bit, meal, sauce, dishes, quality |
| | Service | table, bar, friendly, wait, server, staff, minutes, beer, atmosphere |
| | Misc. | eat, dinner, Vegas, experience, wait, friends, times, never, give |
| Item_673 | Food | nigiri, sake, tempura, shrimp, sauce, items, poke, crab, chef |
| | Ambience | atmosphere, friendly, bar, staff, inside, area, spot, monta, feel |
| | Price | price, worth, prices, nigiri, sake, tempura, items, lunch, special |
| | Service | service, table, server, friendly, minutes, staff, nice, asked, seated |
| | Misc. | restaurant, times, give, favorite, night, places, stars, friends, Vegas |

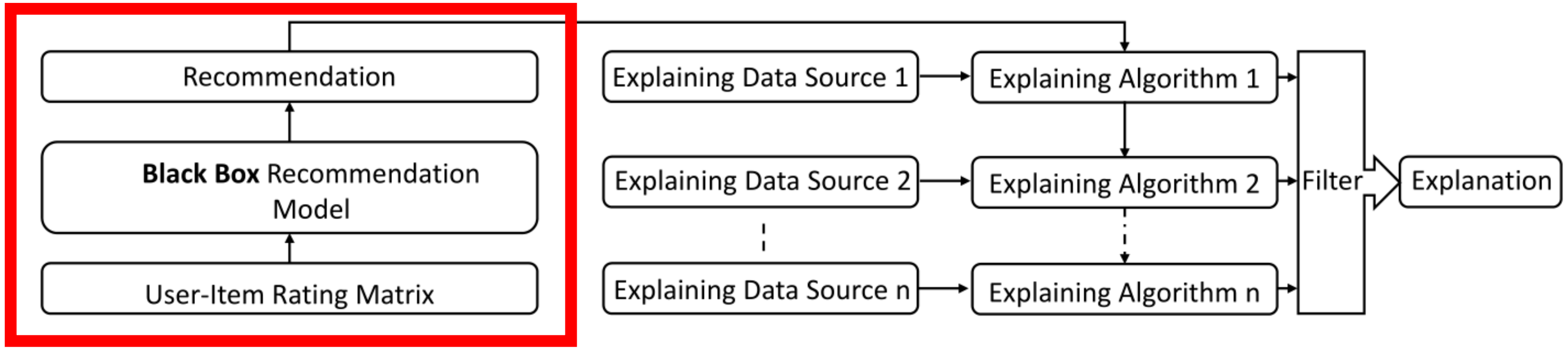
Table 6. Interpretation for Why the "User 2397" Rated "Item 137" and "Item 673" with 5 and 2, Respectively

| Item | Aspect | Food | Ambience | Price | Service | Misc. |
|---|---|---|---|---|---|---|
| Item_137 | Importance | 0.3815 | 0.1034 | 0.0723 | 0.2038 | 0.2390 |
| | Matching | 0.5672 | 0.4523 | 0.5329 | 0.6021 | 0.7138 |
| | Polarity | + | + | − | + | + |
| Item_673 | Importance | 0.3726 | 0.0794 | 0.0853 | 0.2076 | 0.2551 |
| | Matching | 0.1813 | 0.6535 | 0.4512 | 0.6018 | 0.7093 |
| | Polarity | − | − | + | + | − |


[1] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. ACM Transactions on Information Systems (TOIS) 37, 2 (2019), 1–28.

# Post-hoc methods

- **An example from Shmaryahu et al.**
  - It generates explanations directly from the recommendation and explaining data source

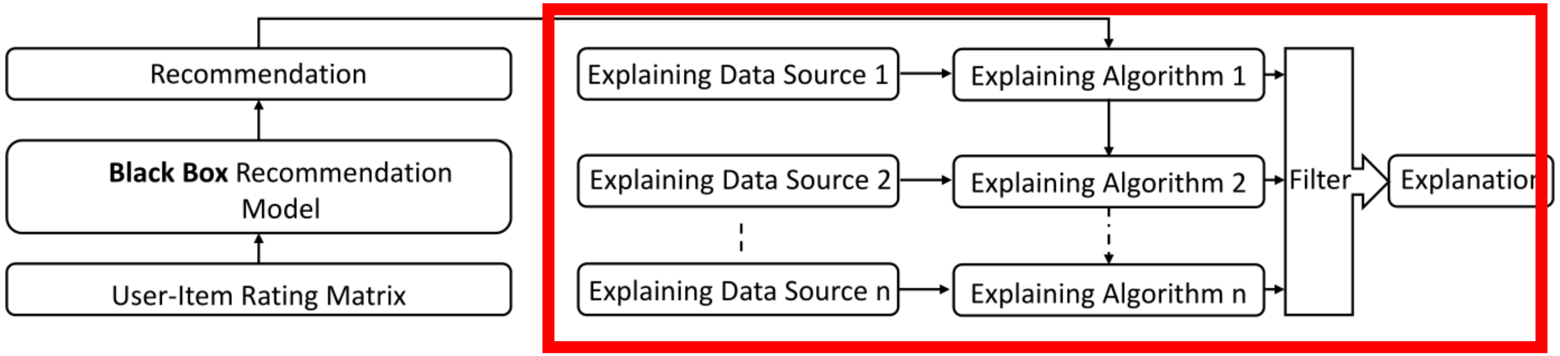




[1] Dorin Shmaryahu, Guy Shani, and Bracha Shapira. 2020. Post-hoc Explanations for Complex Model Recommendations using Simple Methods. In IntRS@ RecSys. 26–36.

# Post-hoc methods

- **An example from Shmaryahu et al.**
  - It generates explanations directly from the recommendation and explaining data source



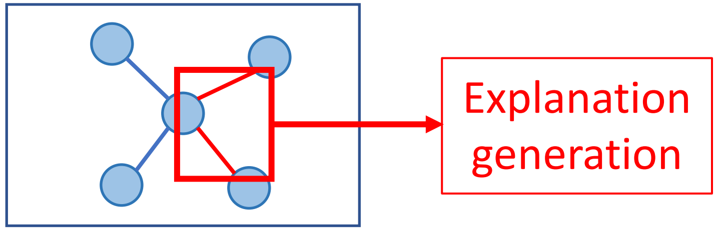



[1] Dorin Shmaryahu, Guy Shani, and Bracha Shapira. 2020. Post-hoc Explanations for Complex Model Recommendations using Simple Methods. In IntRS@ RecSys. 26–36.

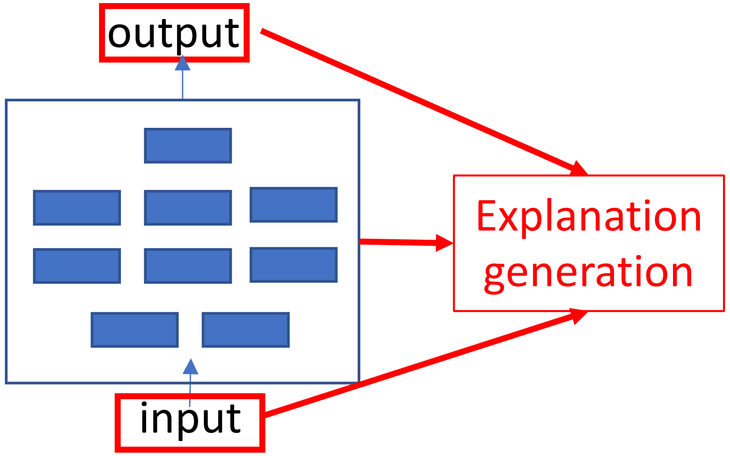
# Taxonomy

- **The second criteria: How the explanations are presented**
  - Structured methods: present explanations in the form of **logical reasoning** based on some particular structures, such as a graph, or a knowledge graph

A layer in a model

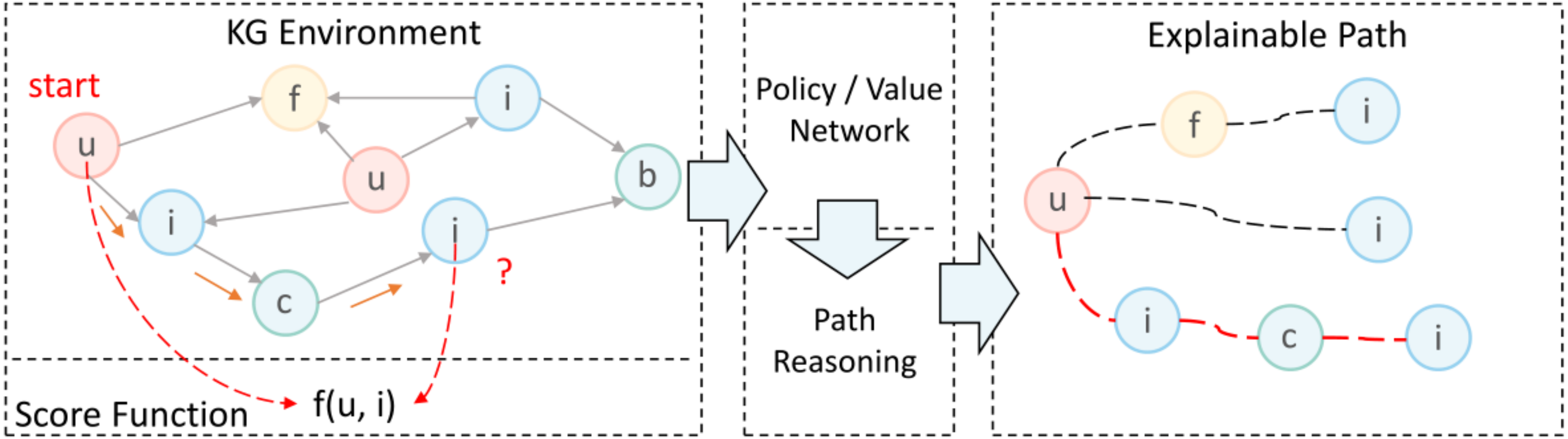



  - Unstructured methods: provide explanations based on the inputs, outputs and models, do not rely on, or explicitly rely on logical reasoning



# Structured methods

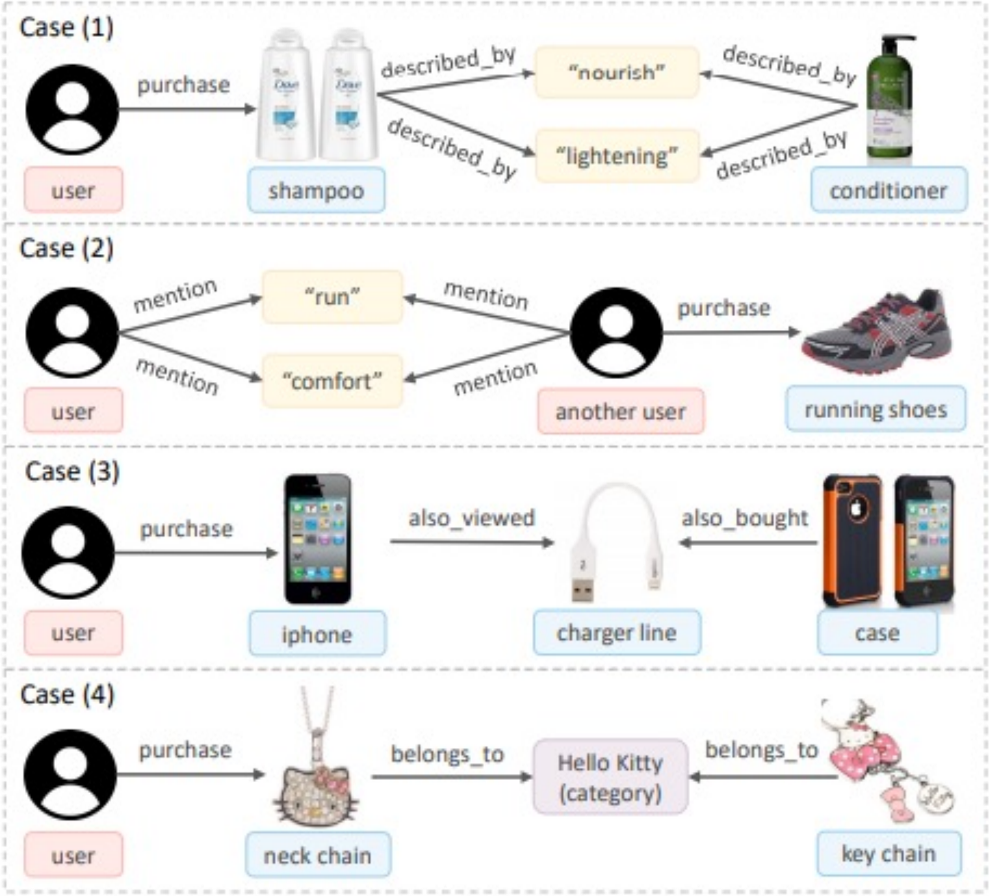
- **PGPR**
  - An explanation path graph generated with knowledge graph
  - Path definition: $p_k(e_0, e_k) = \left\{ e_0 \overset{r_1}{\leftrightarrow} e_1 \overset{r_2}{\leftrightarrow} \cdots \overset{r_k}{\leftrightarrow} e_k \right\}$




[1] Yikun Xian, Zuohui Fu, Shan Muthukrishnan, Gerard De Melo, and Yongfeng Zhang. 2019. Reinforcement knowledge graph reasoning for explainable recommendation. In Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 285–294.

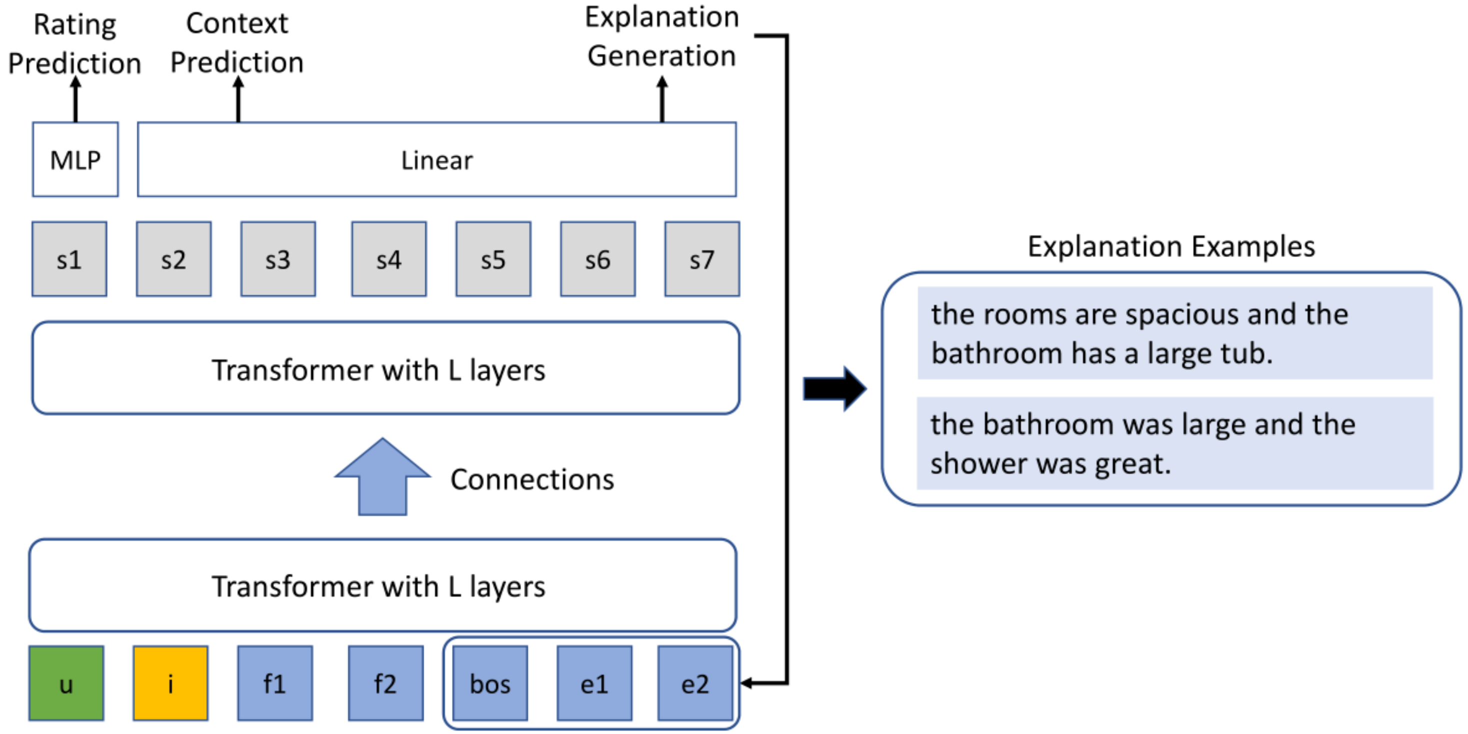
# Structured methods

- **PGPR**
  - Explanation path



# Unstructured methods

- **PETER**
  - Generate explanation sentence word by word
  - The final explanation is a sentence based on probability, not the sole reason deduced according to deterministic rules or structures




[1] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. arXiv preprint arXiv:2105.11601 (2021).

# Unstructured methods

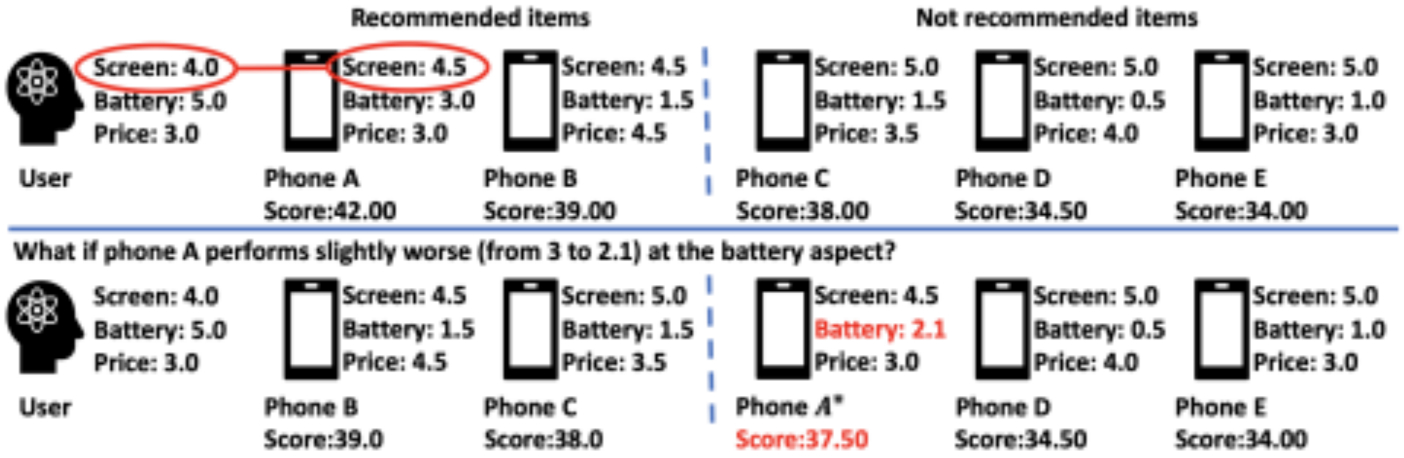
- **CountER**
  - It tries to use small changes in item aspects to reverse the decision

```
If the item had been slightly worse on [aspect(s)],
then it will not be recommended.
```

minimize Explanation Complexity
s.t., Explanation is Strong Enough

Matching-based:

Counterfactual reasoning:




[1] untao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual explainable recommendation. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management.

# Unstructured methods

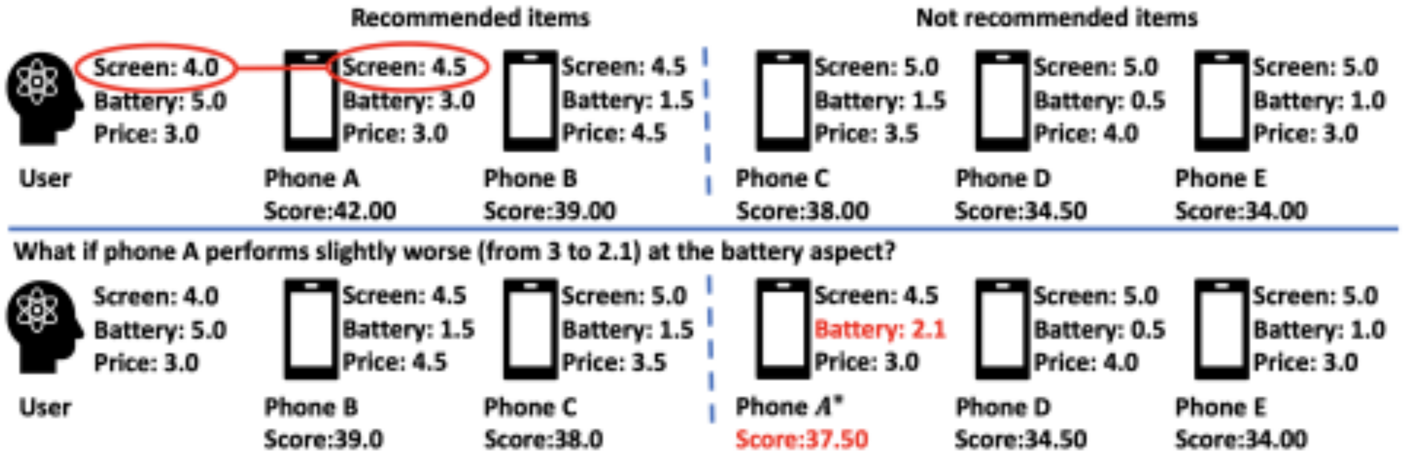
- **CountER**
  - It tries to use small changes in item aspects to reverse the decision

```
If the item had been slightly worse on [aspect(s)],
then it will not be recommended.
```

minimize Explanation Complexity
s.t., Explanation is Strong Enough

Matching-based:

Counterfactual reasoning:




[1] untao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual explainable recommendation. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management.

# Explainability

METHODS

**EVALUATIONS**

APPLICATIONS

FUTURE DIRECTIONS

# Taxonomy of research on evaluations

- **Evaluation perspectives**
  - Effectiveness
  - Transparency
  - Scrutability

- **Evaluation form**
  - Quantitative metrics
  - Case study
  - Real-world performance
  - Ablation Study

# Taxonomy of Evaluation

- **Evaluation perspectives**
  - Effectiveness
  - Transparency
  - Scrutability

| Evaluation perspective | Evaluation criteria | Related research |
|---|---|---|
| Effectiveness | Whether the explanations are useful to users? (e.g. Decision making, Recommendation results) | [8, 58, 337] |
| Transparency | Whether the explanations can reveal the working principles of the model? | [18, 144, 225] |
| Scrutability | Whether the explanations contribute to the prediction of the model? | [327, 347, 362] |

Reference: Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In Recommender systems handbook. Springer, 479–510.

# Taxonomy of Evaluation

- **Evaluation form**
  - **Quantitative:** ROUGE score, BLEU, USR, FMR...
  - **Case study:** Whether the explanation conforms to human logic
  - **Real-world performance:** The practical effects of the explanation
  - **Ablation study:** How algorithmic modules provide explanations and how these modules enhance the recommendation model

| Evaluation form | Corresponding perspectives | Related research |
|---|---|---|
| Quantitative metrics | Effectiveness; Scrutability | [337, 338] |
| Case study | Effectiveness; Transparency | [225, 362, 396] |
| Real-world performance | Effectiveness; Scrutability; Transparency | [58, 347, 392] |
| Ablation Study | Effectiveness; Transparency | [64, 211, 327] |

# Explainability

METHODS

EVALUATIONS

**APPLICATIONS**

FUTURE DIRECTIONS

# E-commercial Recommendation

# Social Media

# Explainability

METHODS

EVALUATIONS

APPLICATIONS

**FUTURE DIRECTIONS**

# Natural Language Generation

- **Templated based (now)**

I recommend Iron Man to you because you've seen The Avengers

- **Full paragraph interpretation generation (currently exist but their effectiveness has yet to improve)**

Since you've seen movies like The Avengers, and your recent interest is in the TV series, we recommend something similar for you: Agents of S.H.I.E.L.D.

# Explainable recommendations in more fields


Academic
Support

Explainable
recommendations

Medical
Care

Education

Etc.

# Summary

- **Concept of explainability in Rec**
  - The ability to explain or to present in understandable terms to a human
- **Taxonomy of methods**
  - How to produce explanations: model-intrinsic based (mostly used) or post-hoc
  - How the explanations are presented: structured or unstructured
- **Taxonomy of evaluations**
  - Evaluation perspectives: Effectiveness, Transparency, Scrutability
  - Evaluation forms: Quantitative, Case study, Real-world performance, Ablation study
- **Application**
  - E-commercial Recommendation
  - Social Media
- **Future directions**
  - Natural Language Generation for Explanation
  - Explainable recommendations in more fields