Spotify TR/ML Conference, Sept 24th, 2018

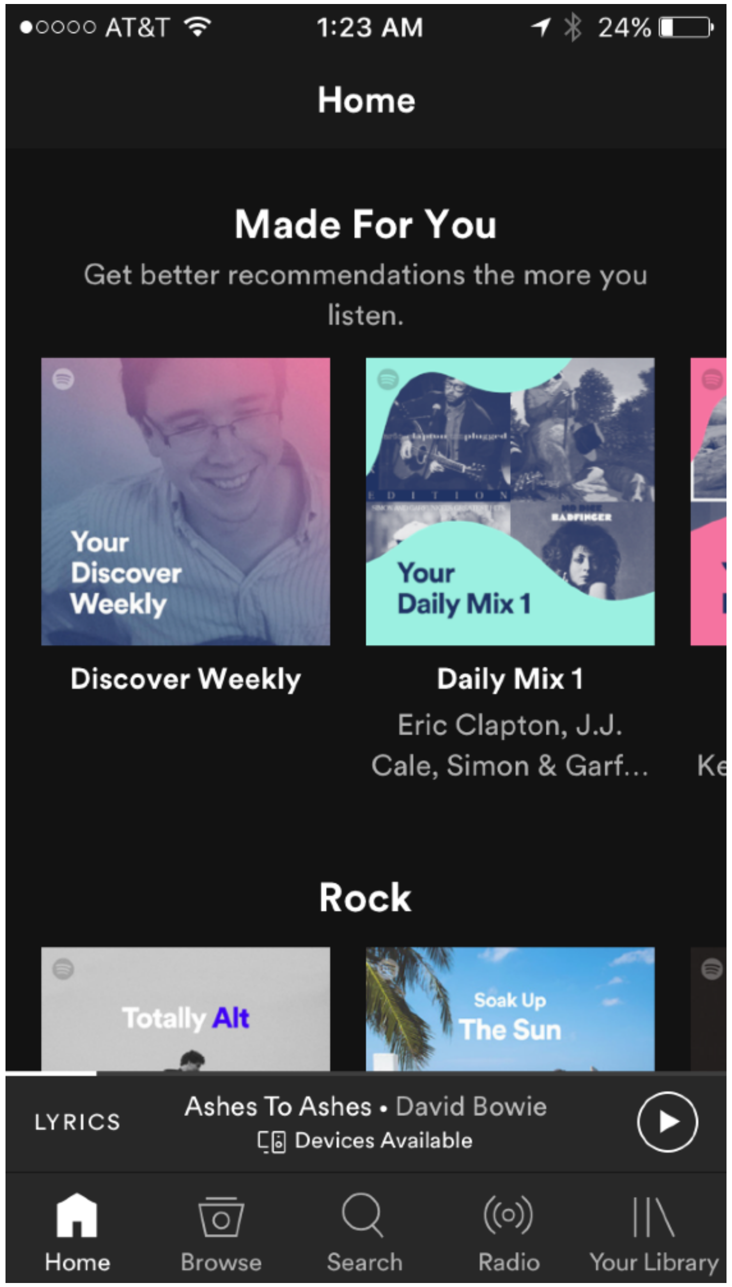
Explore, Exploit, and Explain: Personalizing Explainable Recommendations with Bandits

James McInerney, Ben Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, Rishabh Mehrotra

Spotify

email: jamesm@spotify.com

Research question: how to explore-exploit over explainable recommendations?



- e.g. home page of Spotify
- items arranged into shelves, each shelf has an explanation for the associated recommendation

Outline

1. Pareto principle for content producers
2. a causal diagnosis of filter bubbles in recommendation
3. contextual bandits for recommendation
4. explained recommendations
5. introducing Bart (bandits for recommendations as treatments)
6. offline and online experiments on homepage data
7. conclusions & future work

A small number of producers dominate consumption in culture

producer popularity

leads to leads to

exposure to new consumers

A small number of producers dominate consumption in culture

e.g. musicians, authors, actors

producer popularity

leads to

leads to

exposure to new consumers

A small number of producers dominate consumption in culture

e.g. musicians, authors, actors

producer popularity

leads to

leads to

exposure to new consumers

e.g. word of mouth, media coverage

A small number of producers dominate consumption in culture

e.g. musicians, authors, actors

producer popularity

leads to

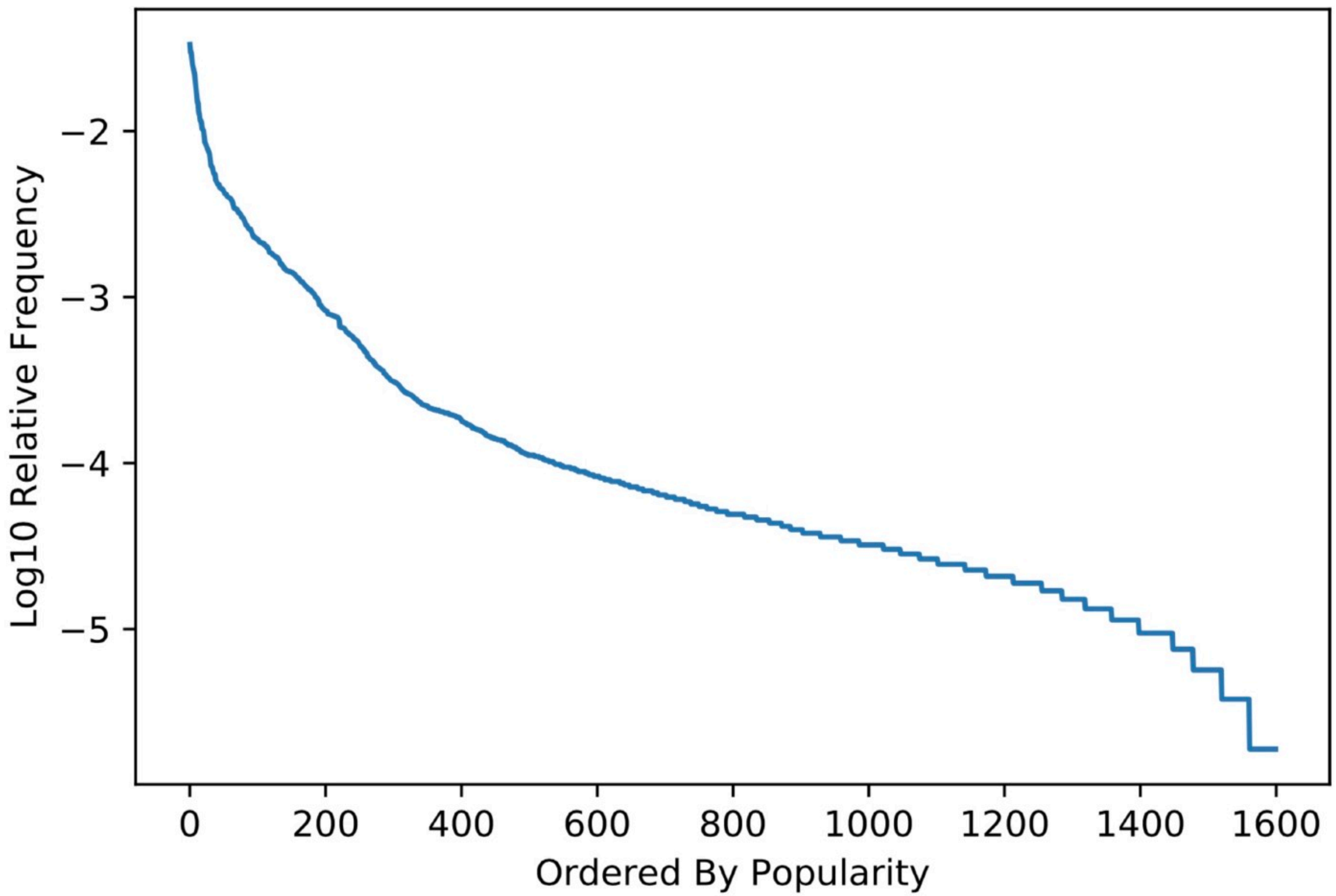
leads to

exposure to new consumers

e.g. word of mouth, media coverage

- known as the Matthew effect or Pareto principle [Juran, 1937]

A small number of producers dominate consumption in culture



Collaborative filtering perpetuates the Pareto principle

e.g. matrix factorization

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^{\top}$$

$\#\text{users} \times \#\text{items}$

logged feedback data

$\#\text{users} \times K$ $\#\text{items} \times K$

latent variables fitted
to logged feedback data

Collaborative filtering perpetuates the Pareto principle

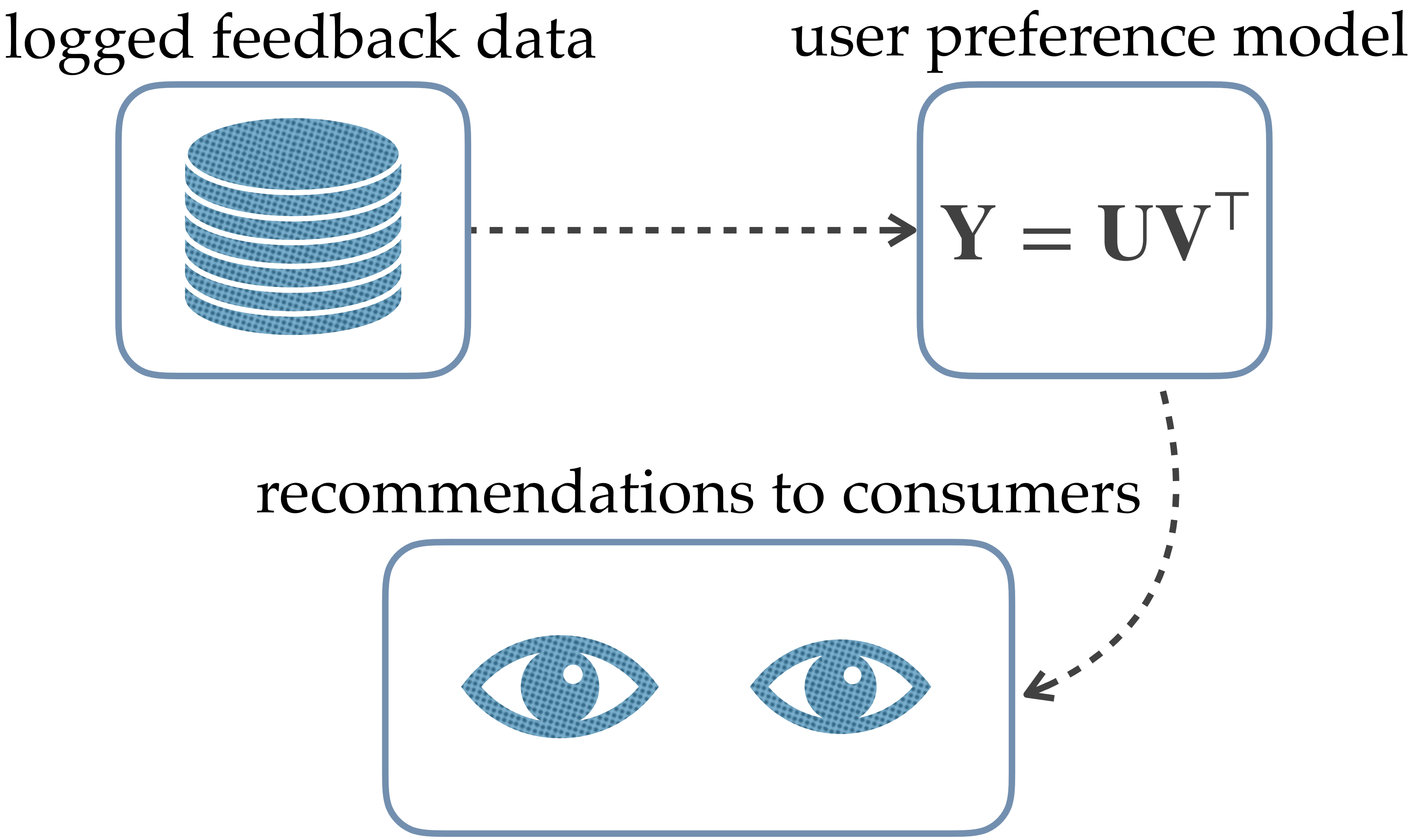
e.g. matrix factorization

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^{\top}$$

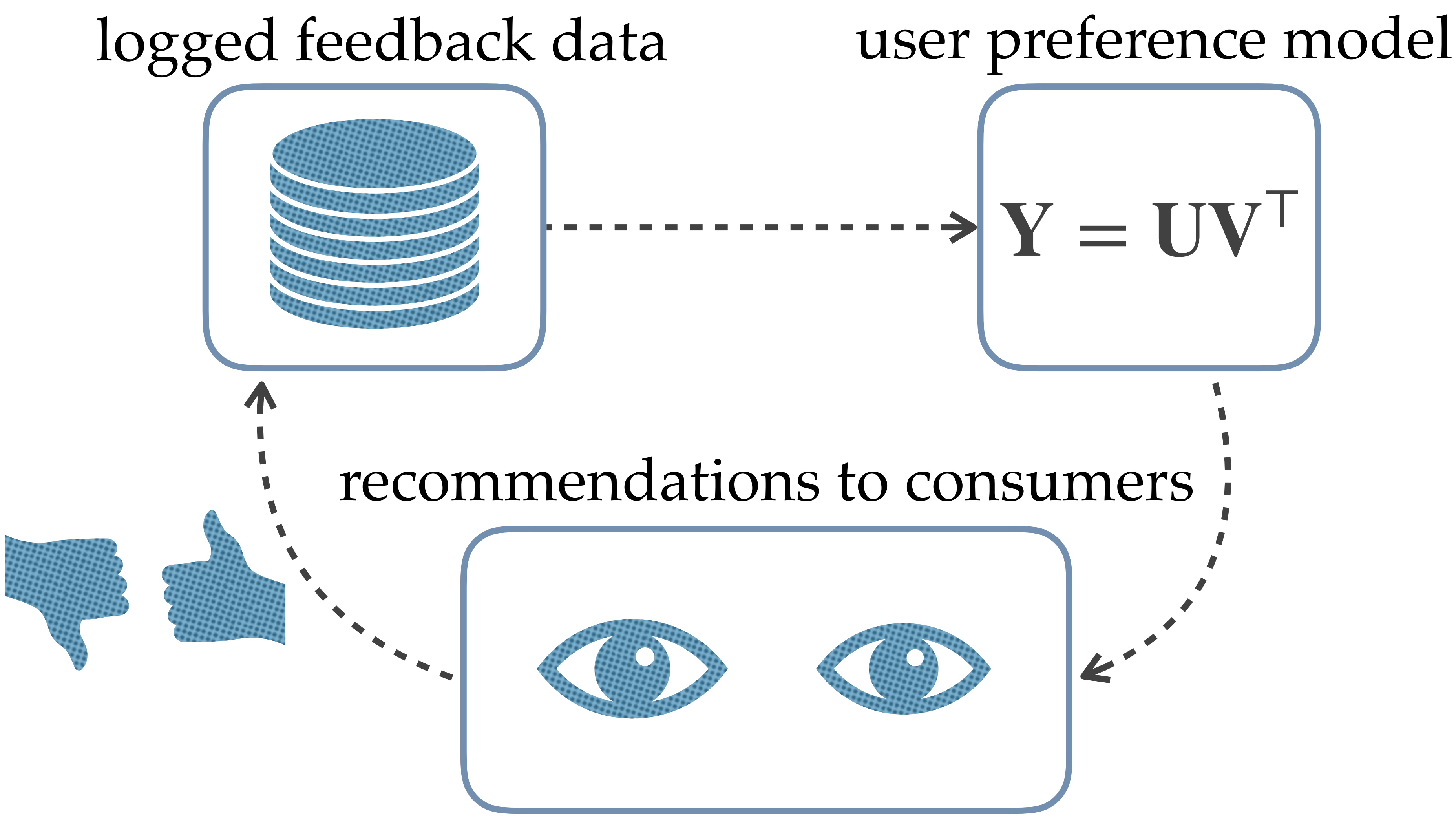
$$\underbrace{\#\text{users} \times \#\text{items}}_{\text{logged feedback data}} \qquad \underbrace{\#\text{users} \times K \quad \#\text{items} \times K}_{\text{latent variables fitted to logged feedback data}}$$

- in general: collaborative filtering engines use implicit feedback data from users to learn a model of user preferences

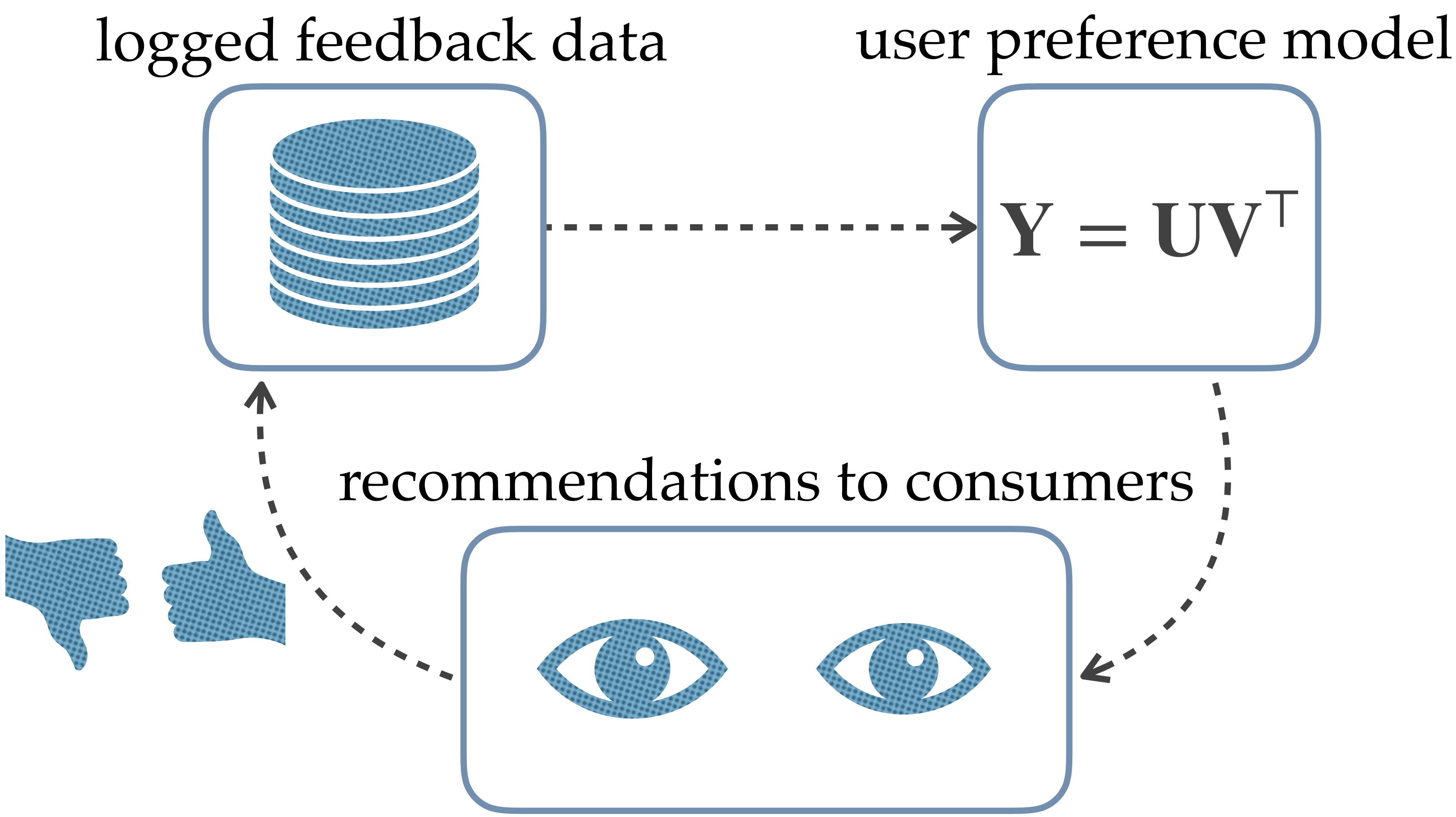
Collaborative filtering perpetuates the Pareto principle



Collaborative filtering perpetuates the Pareto principle



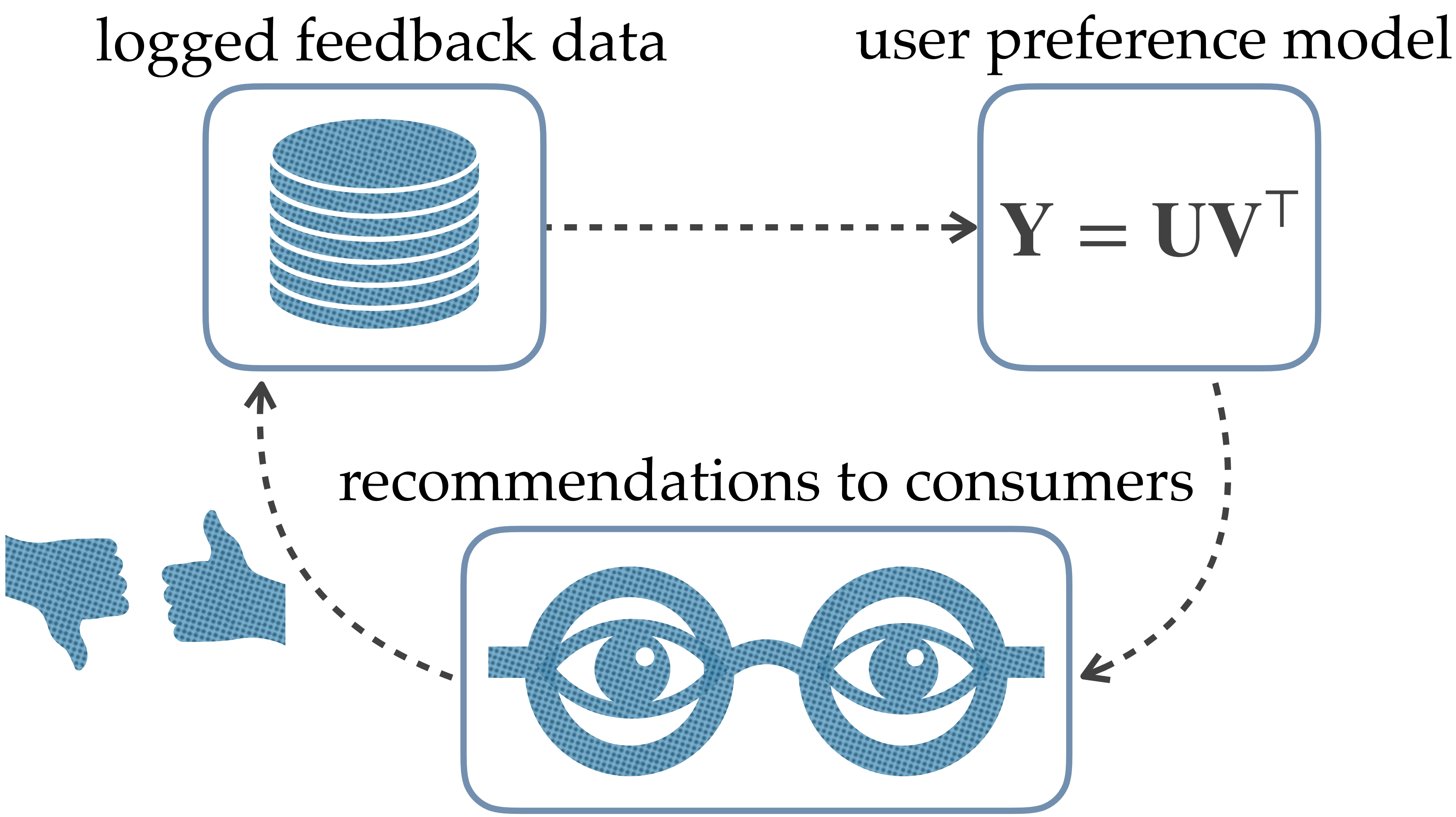
Collaborative filtering perpetuates the Pareto principle



"How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility" (Chaney et al. 2017)

"Modeling User Exposure in Recommendation" (Liang et al. 2016)

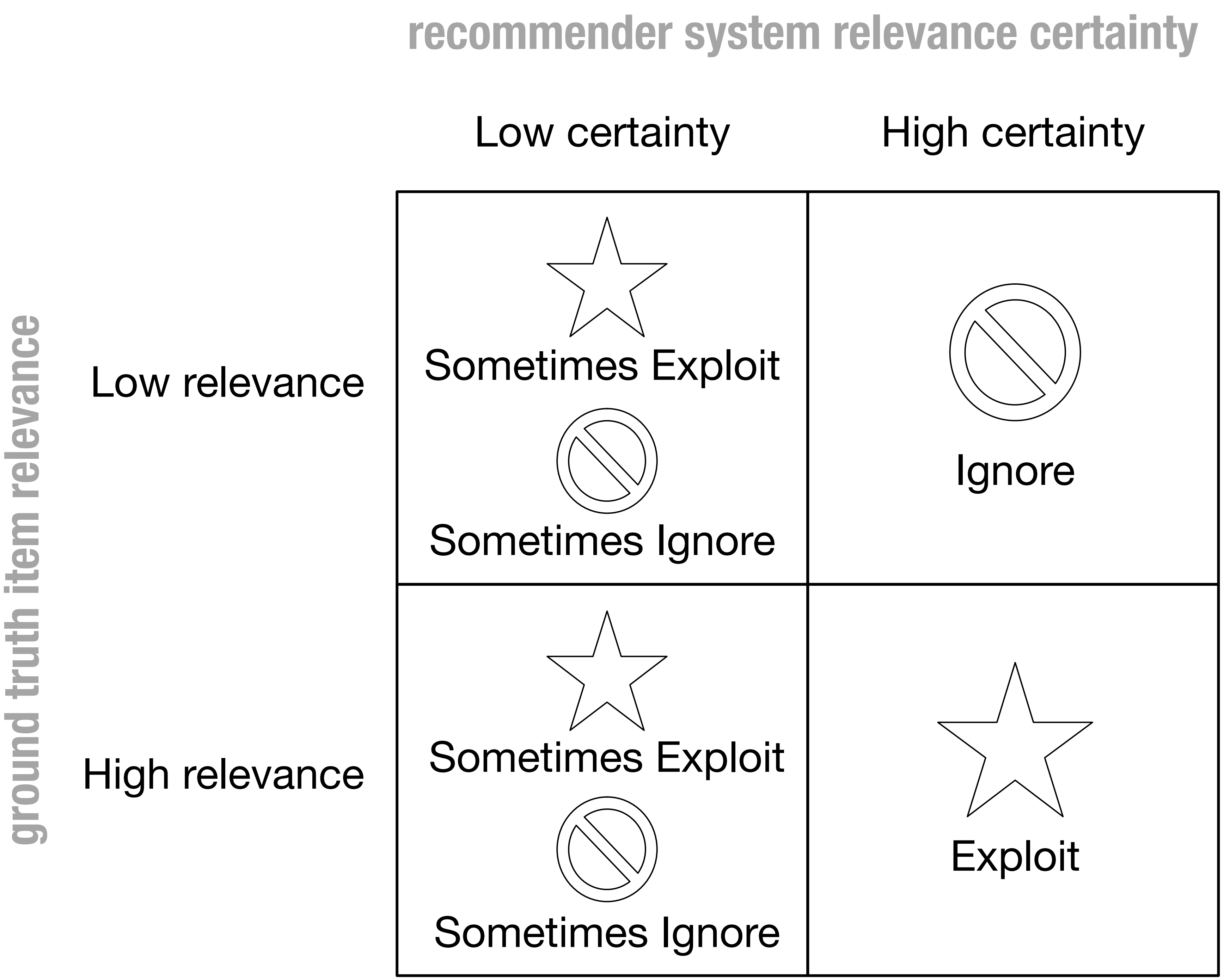
Collaborative filtering perpetuates the Pareto principle



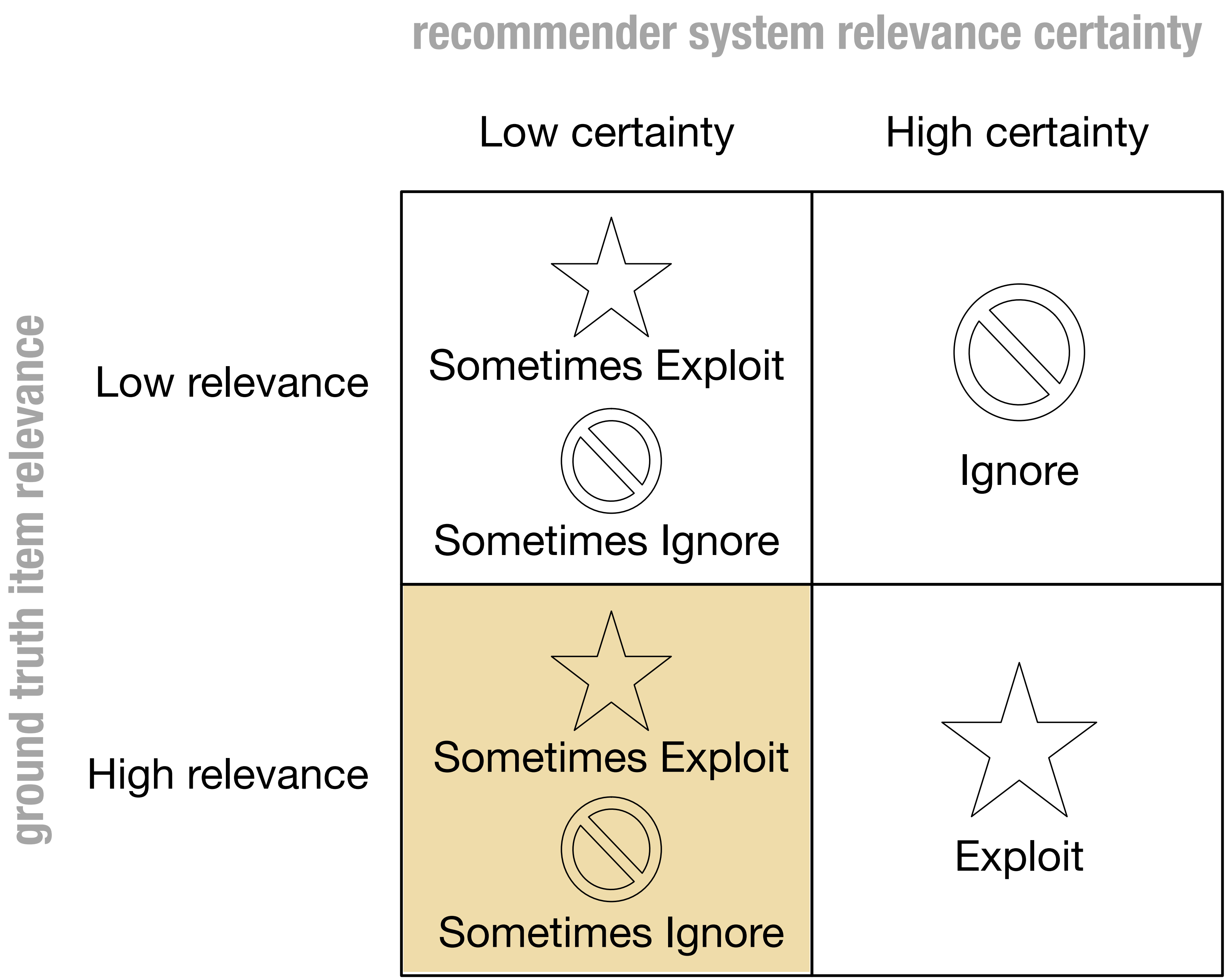
"How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility" (Chaney et al. 2017)

"Modeling User Exposure in Recommendation" (Liang et al. 2016)

Standard collaborative filtering methods are limited because they can only exploit or ignore

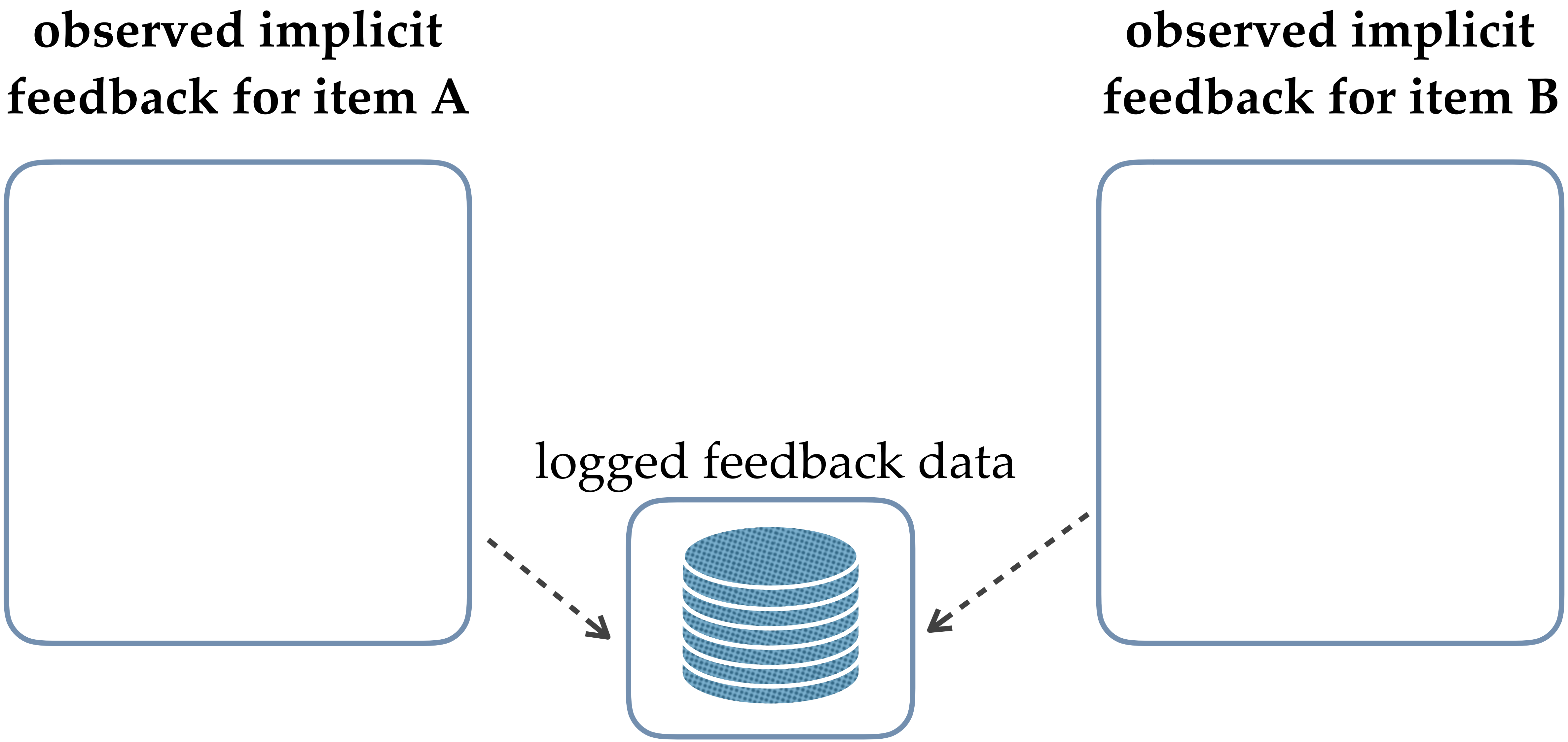


Standard collaborative filtering methods are limited because they can only exploit or ignore



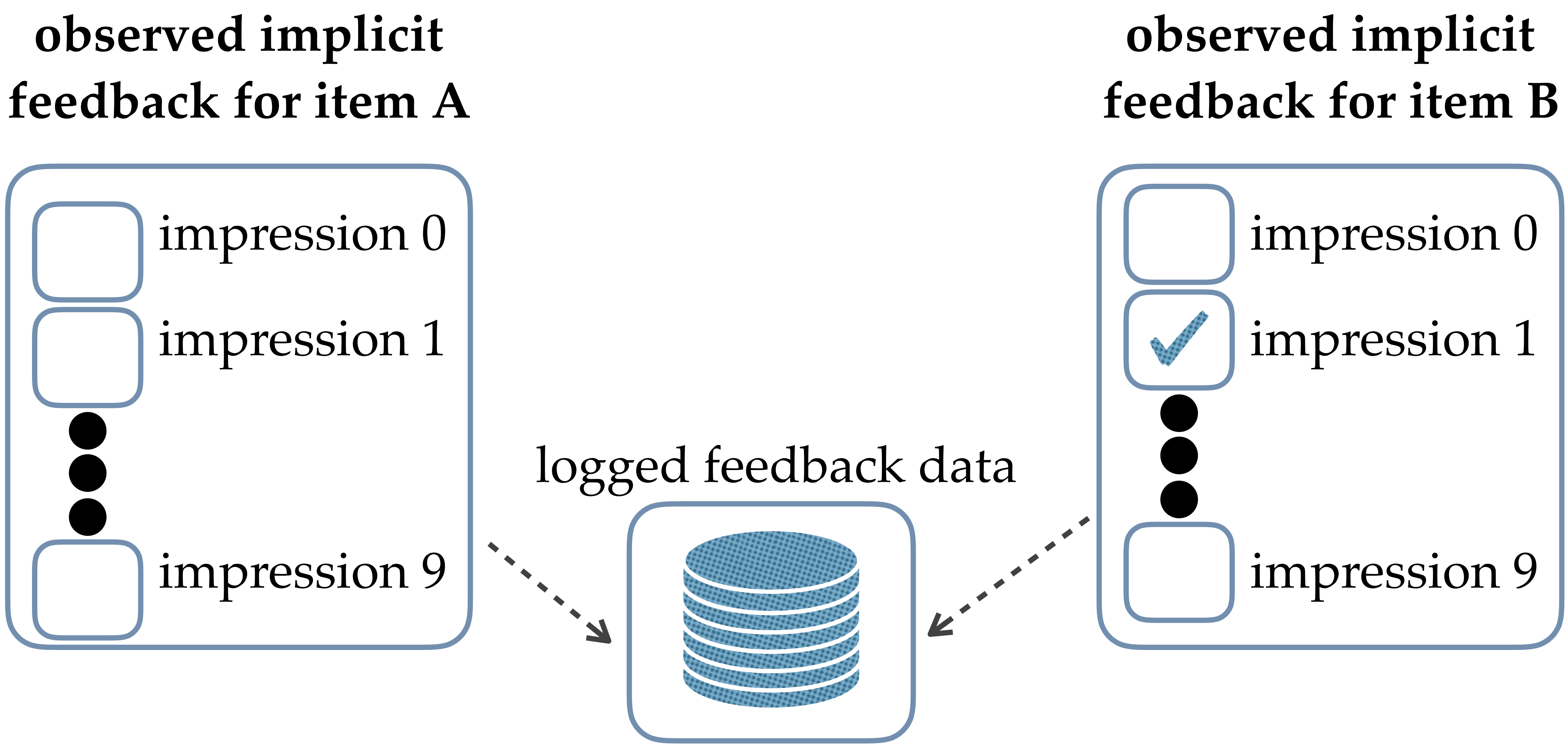
Standard collaborative filtering methods are limited in that they can only exploit or ignore

- e.g. two items, A and B, with the same click rate = 0.1



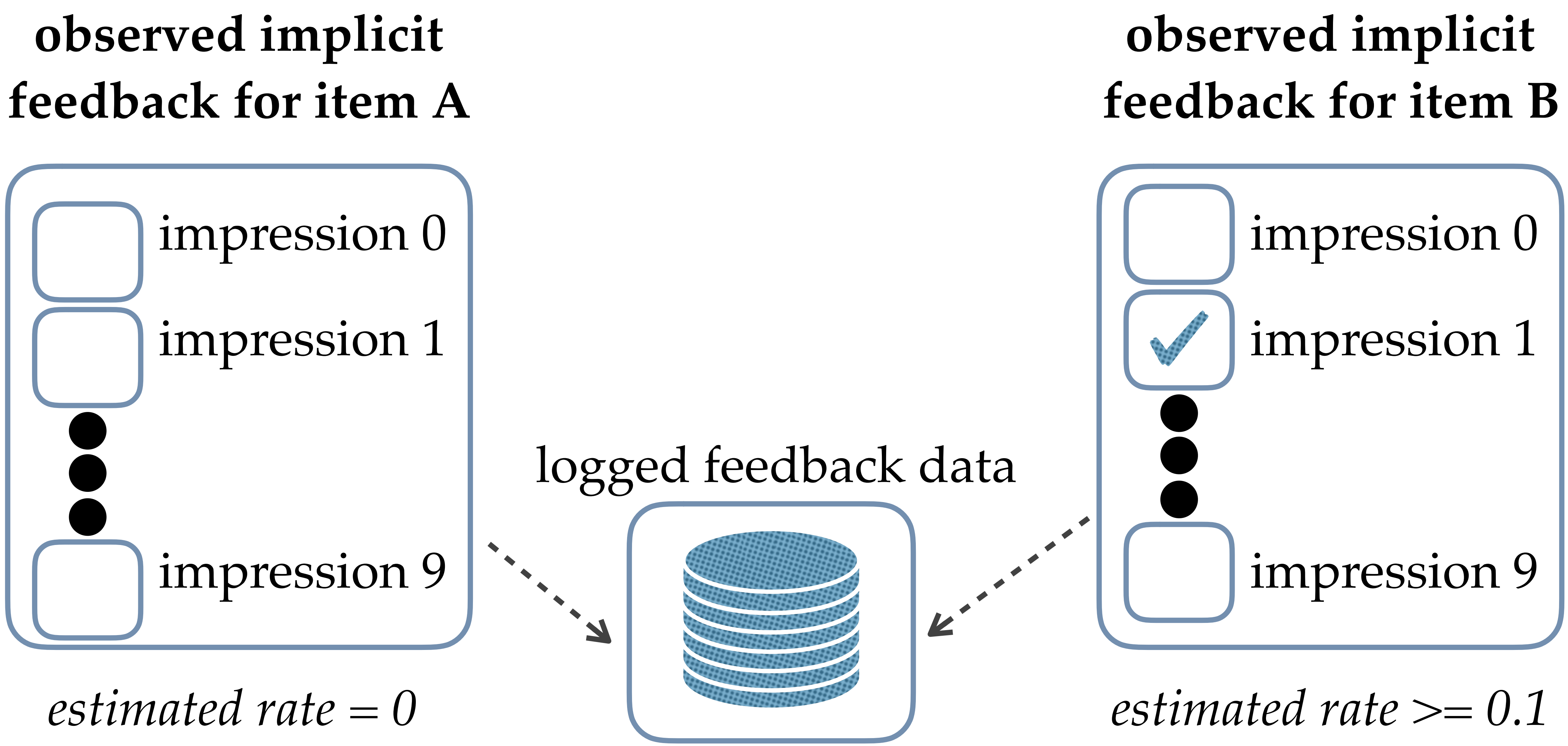
Standard collaborative filtering methods are limited in that they can only exploit or ignore

- e.g. two items, A and B, with the same click rate = 0.1



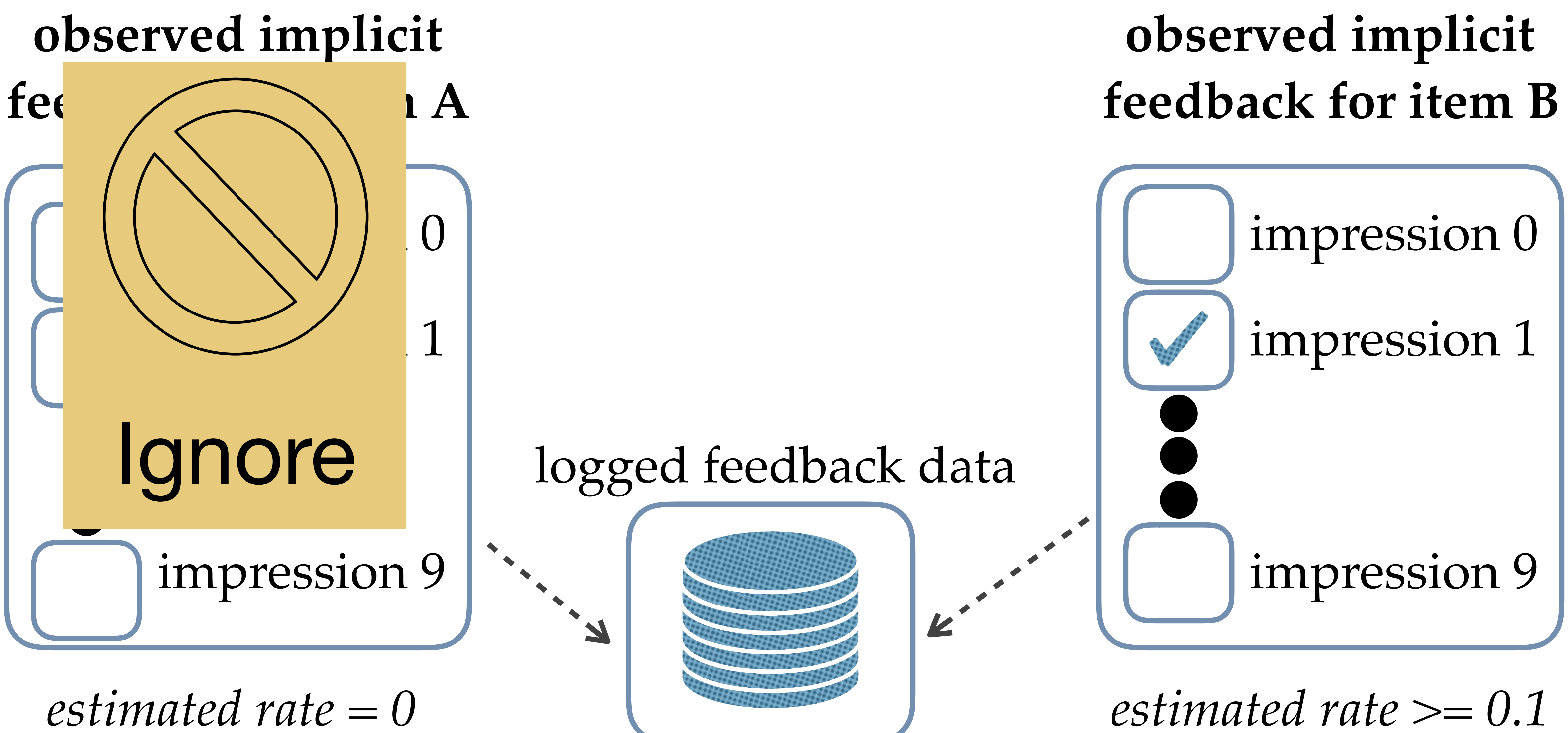
Standard collaborative filtering methods are limited in that they can only <u>exploit</u> or <u>ignore</u>

- e.g. two items, A and B, with the same click rate = 0.1



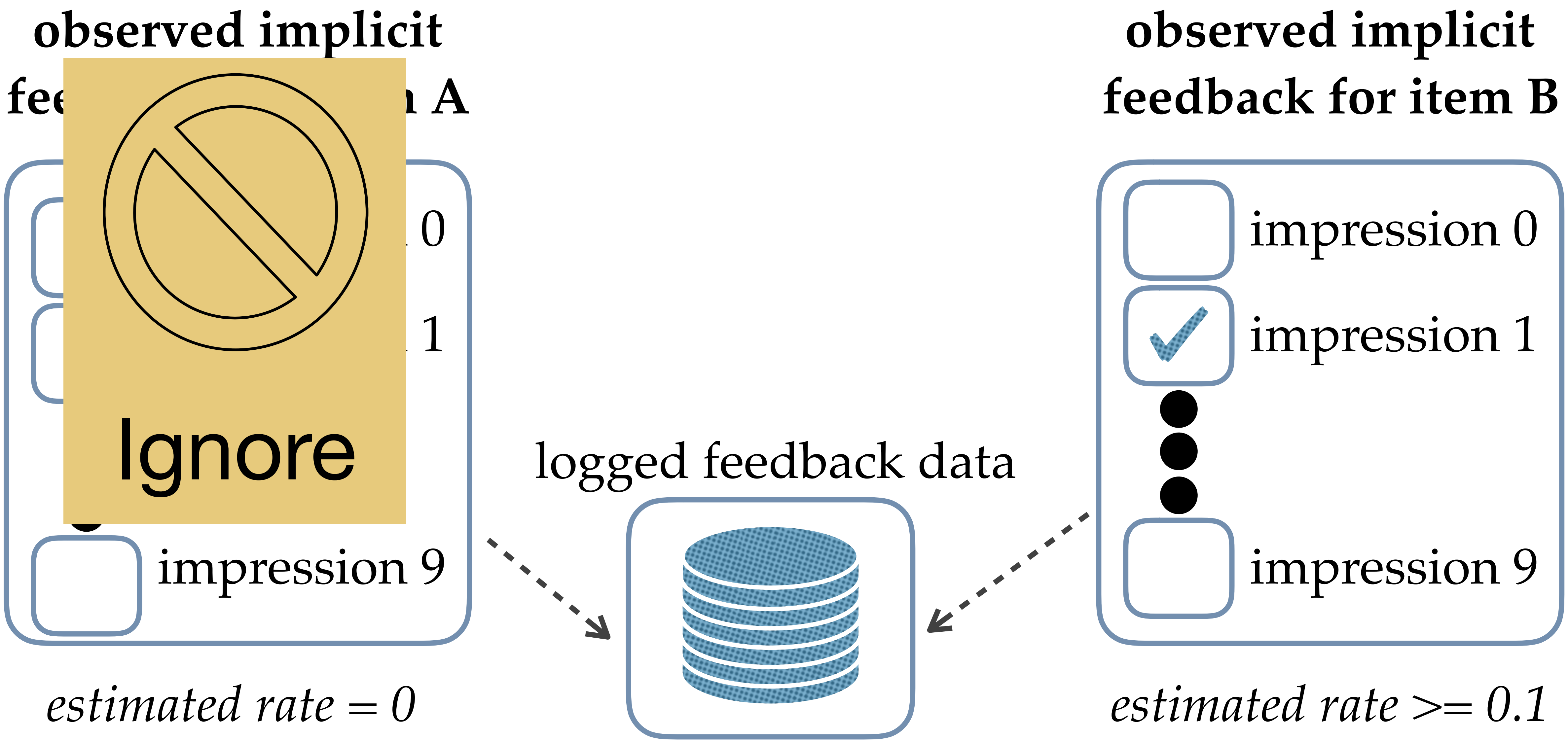
Standard collaborative filtering methods are limited in that they can only exploit or ignore

- e.g. two items, A and B, with the same click rate = 0.1



Standard collaborative filtering methods are limited in that they can only exploit or ignore

- e.g. two items, A and B, with the same click rate = 0.1



- the estimated performance will be identical only 31.3% of the time

Randomized controlled trials

Charles Sanders Peirce

"At the beginning [...] the pack was well shuffled, and, the operator and subject having taken their places, the operator was governed by the color of the successive cards in choosing whether he should first diminish the weight and then increase it, or vice versa."

On Small Differences in Sensation,
C. S. Peirce & J. Jastrow (1885)

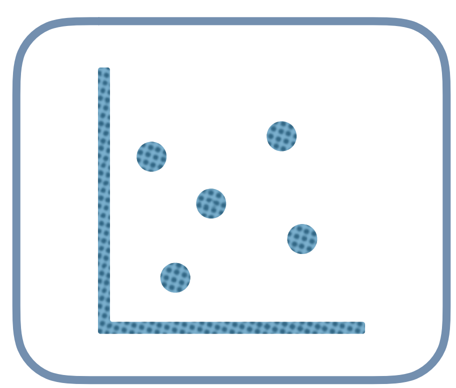
Randomized controlled trials

Charles Sanders Peirce

"At the beginning […] the pack was well shuffled, and, the operator and subject having taken their places, the operator was governed by the color of the successive cards in choosing whether he should first diminish the weight and then increase it, or vice versa."

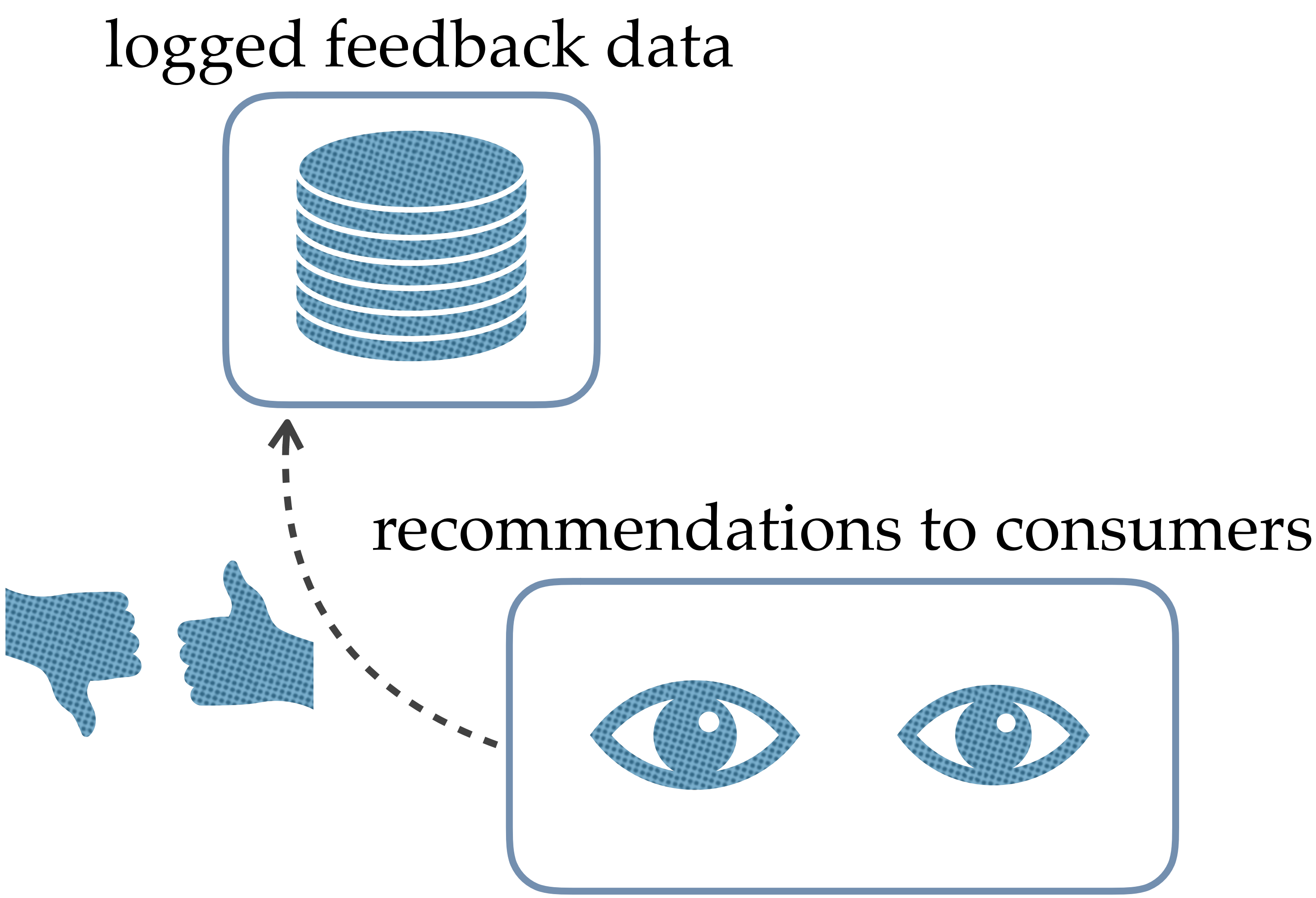
On Small Differences in Sensation,
C. S. Peirce & J. Jastrow (1885)

In recommendation:

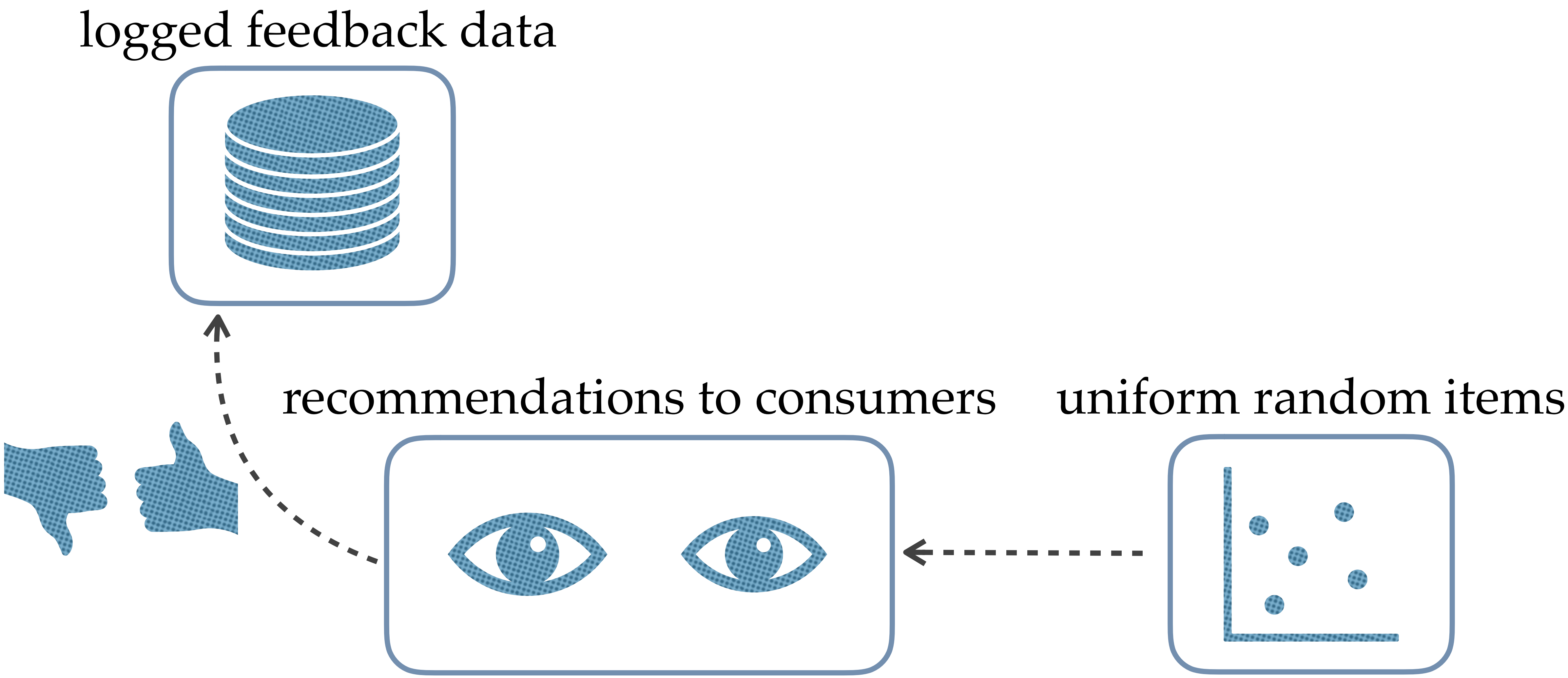


uniform random items

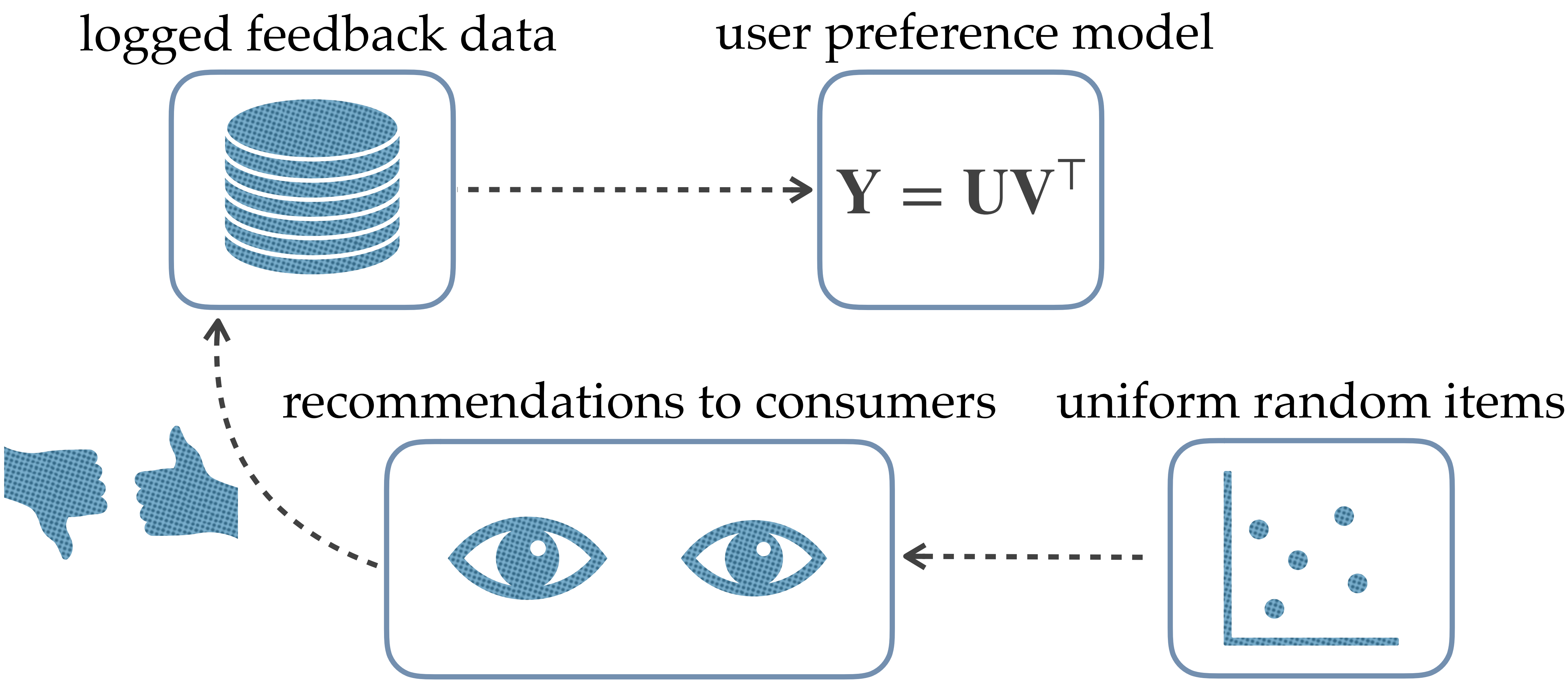
Let's restart from the basic ideal of randomized controlled trials



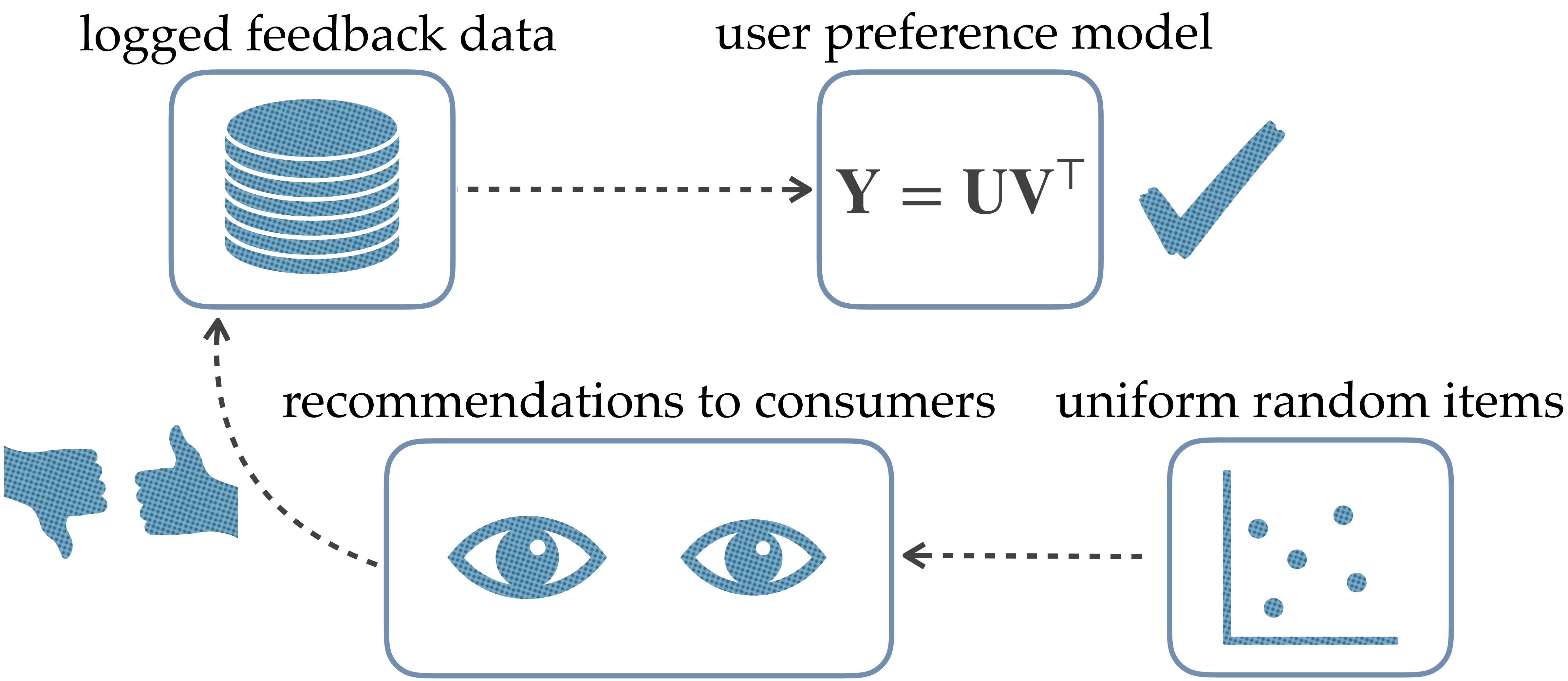
Let's restart from the basic ideal of randomized controlled trials



Let's restart from the basic ideal of randomized controlled trials



Let's restart from the basic ideal of randomized controlled trials



Let's restart from the basic ideal of randomized controlled trials

✓ =

Let's restart from the basic ideal of randomized controlled trials

$$\checkmark = \mathbb{E}_{X,A\sim\text{Uniform}(\mathscr{A}),Y}[\log p_\theta(Y\,|\,A, X)]$$

Let's restart from the basic ideal of randomized controlled trials

"choose a model and train it on data how you like"

$$\checkmark = \mathbb{E}_{X,A\sim\mathrm{Uniform}(\mathscr{A}),Y}[\log p_\theta(Y \mid A, X)]$$

Let's restart from the basic ideal of randomized controlled trials

$$\checkmark \quad = \quad \overbrace{\mathbb{E}_{X,A\sim\mathrm{Uniform}(\mathscr{A}),Y}}^{\text{“train on the right data”}}[\log p_\theta(Y \mid A, X)]$$

Let's restart from the basic ideal of randomized controlled trials

"train on the right data"

$$\checkmark \quad = \quad \mathbb{E}_{X,A\sim \text{Uniform}(\mathscr{A}),Y}[\log p_\theta(Y \,|\, A, X)]$$

random item recommended

set of all items

model parameters

context

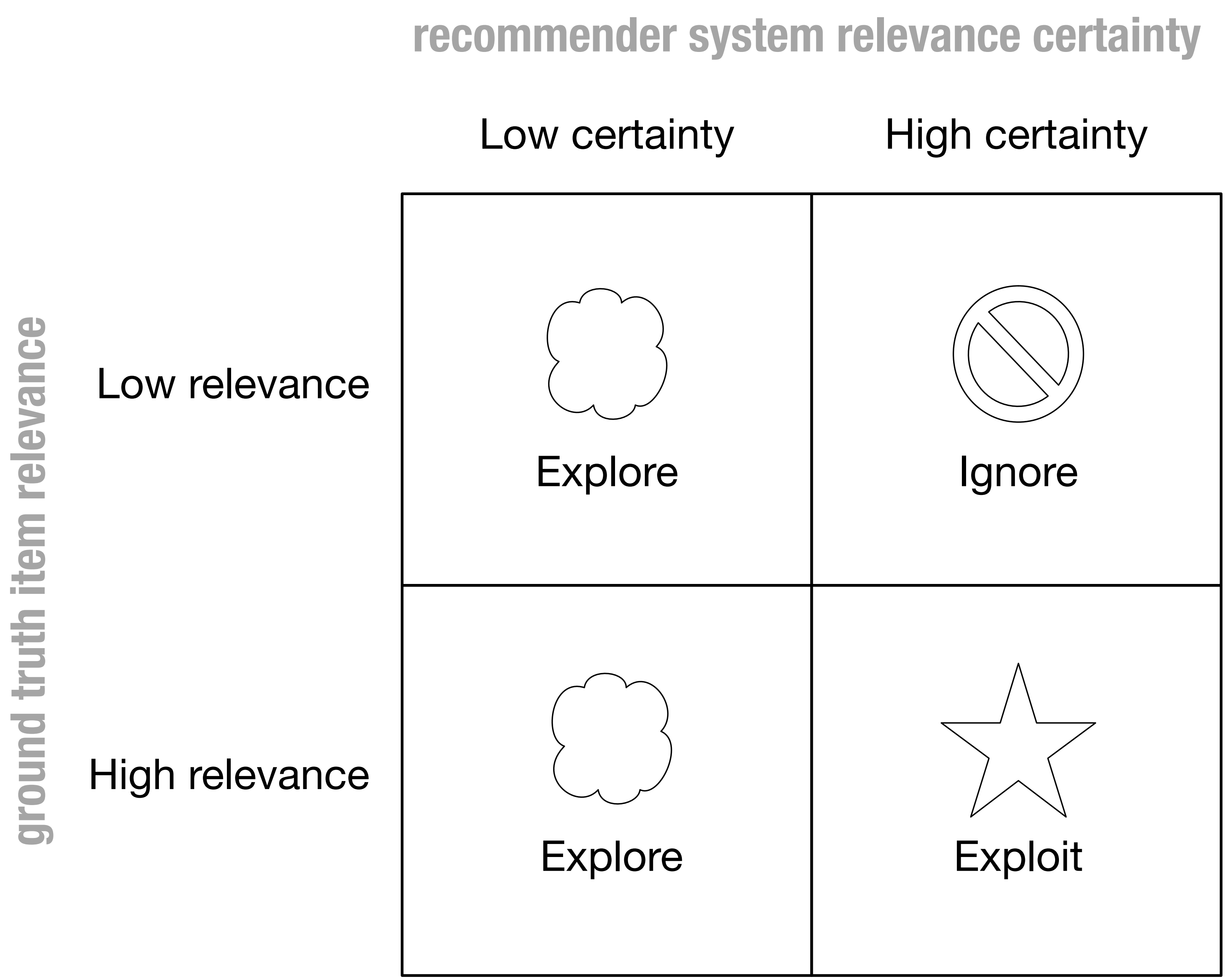
But we don't want to just recommend random stuff all the time ⚡⚡⚡

But we don't want to just recommend random stuff all the time ⚡⚡⚡

- Enter <u>exploration-exploitation</u> [Sutton & Barto, 1998]

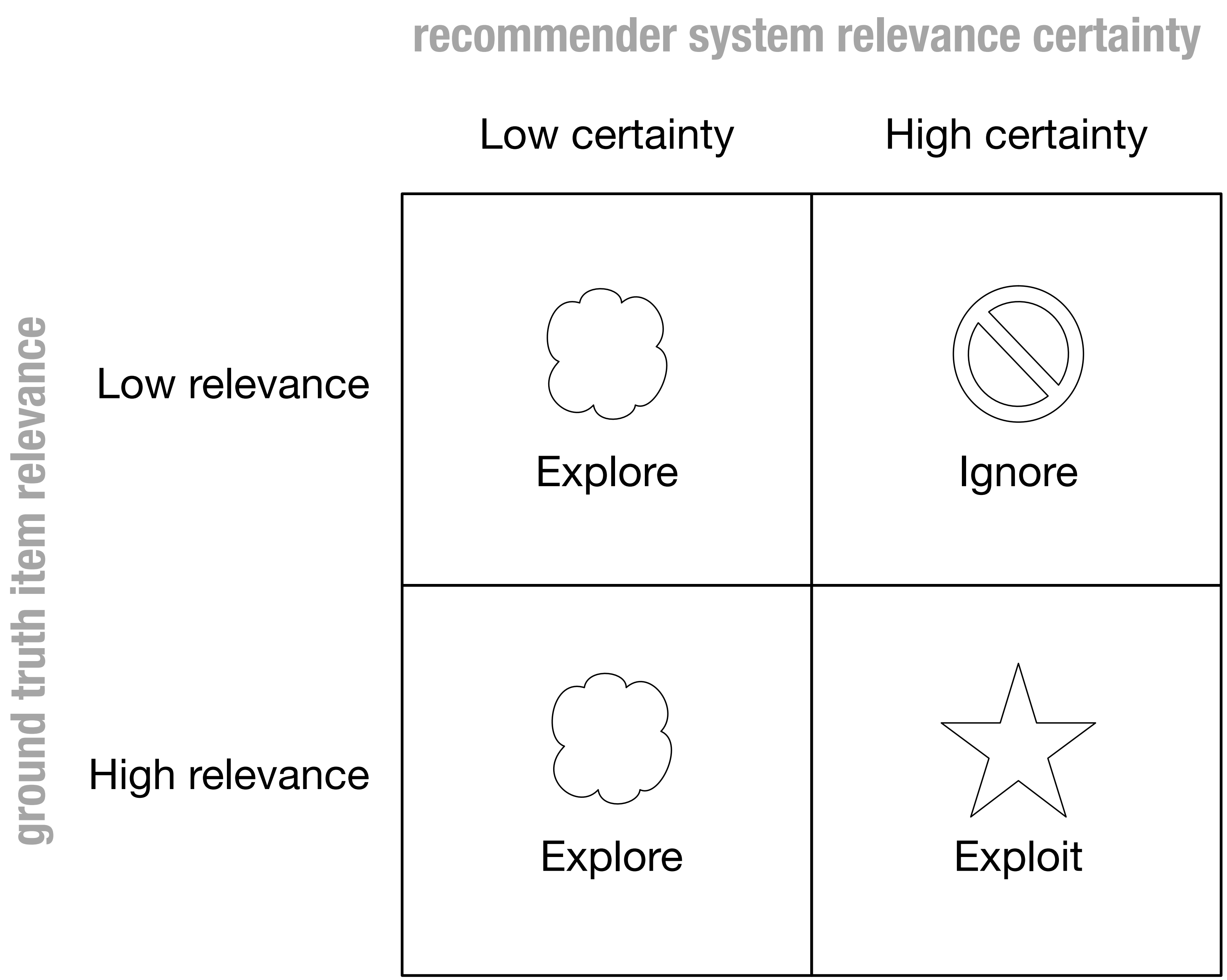
But we don't want to just recommend random stuff all the time ⚡⚡⚡

- Enter <u>exploration-exploitation</u> [Sutton & Barto, 1998]



But we don't want to just recommend random stuff all the time ⚡⚡⚡

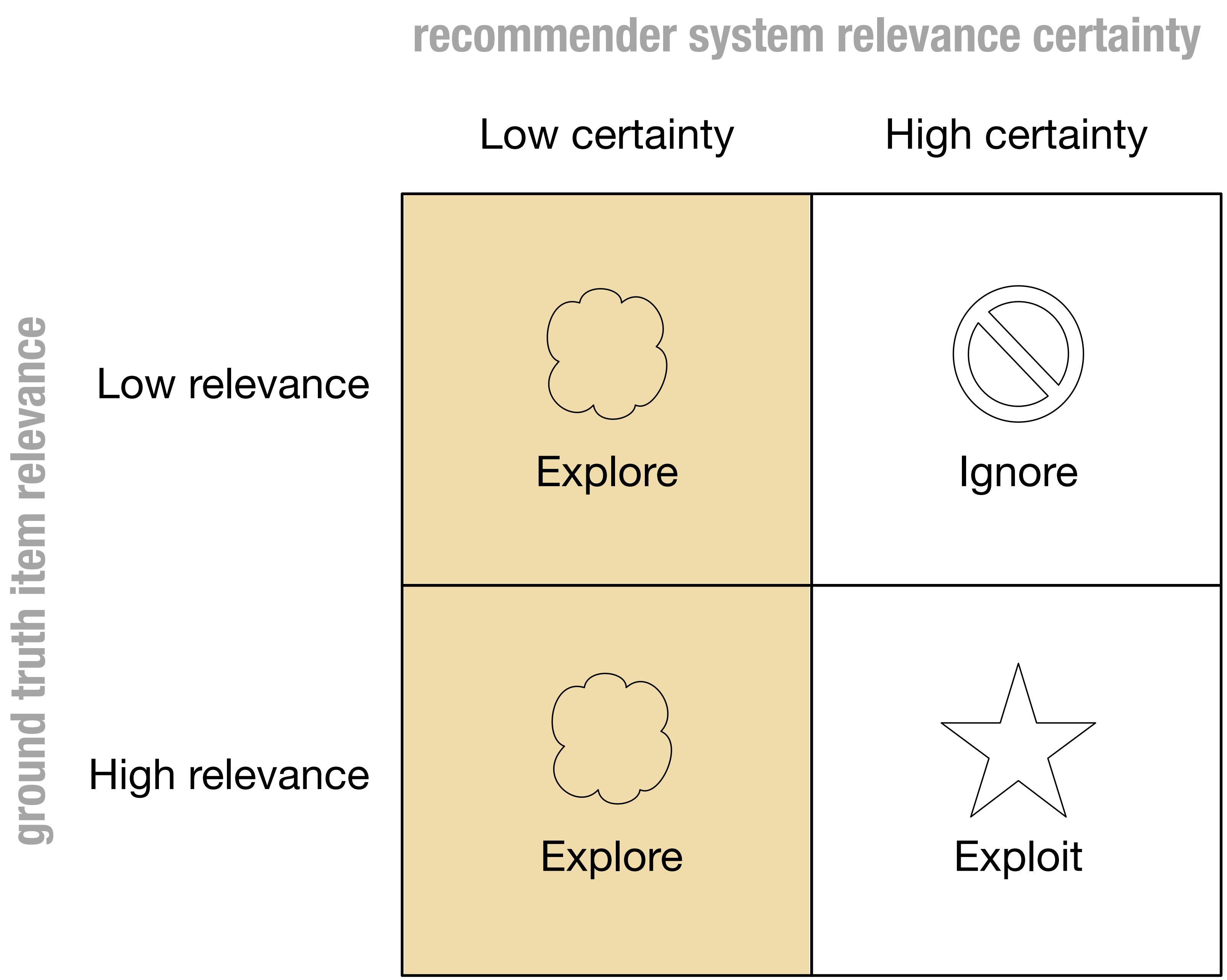
- Enter <u>exploration-exploitation</u> [Sutton & Barto, 1998]



- When the recommender is certain it has a bad item, it ignores it.
- When the recommender is certain it has a good item, it recommends it.

But we don’t want to just recommend random stuff all the time ⚡⚡⚡

- Enter exploration-exploitation [Sutton & Barto, 1998]



- When the recommender is certain it has a bad item, it ignores it.
- When the recommender is certain it has a good item, it recommends it.

How to balance exploration and exploitation?

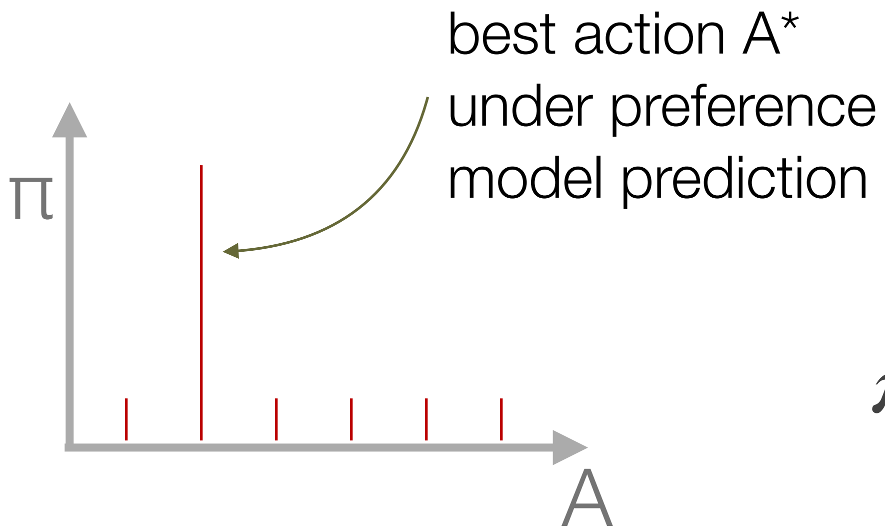
How to balance exploration and exploitation?

- the central question of contextual multi-armed bandits
- standard methods include epsilon-greedy, Thompson sampling, and upper confidence bounds

How to balance exploration and exploitation?

- the central question of contextual multi-armed bandits
- standard methods include epsilon-greedy, Thompson sampling, and upper confidence bounds

ε-greedy



$$\pi(A \mid X) = \begin{cases} (1-\epsilon) + \frac{\epsilon}{|\mathscr{A}|} & \text{when } A = A^* \\ \frac{\epsilon}{|\mathscr{A}|} & \text{otherwise} \end{cases}$$

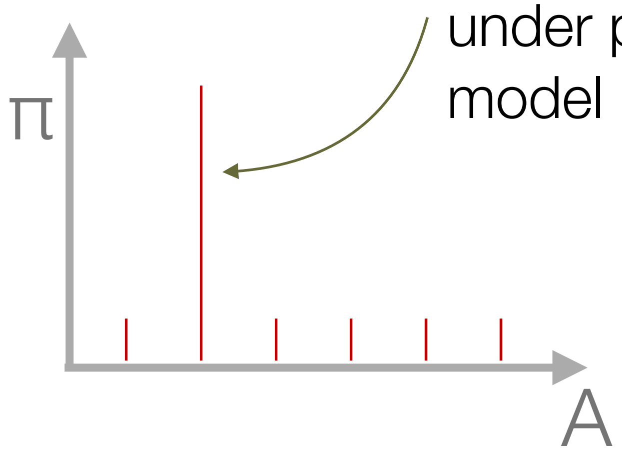
How to balance exploration and exploitation?

- the central question of contextual multi-armed bandits
- standard methods include epsilon-greedy, Thompson sampling, and upper confidence bounds

ε-greedy

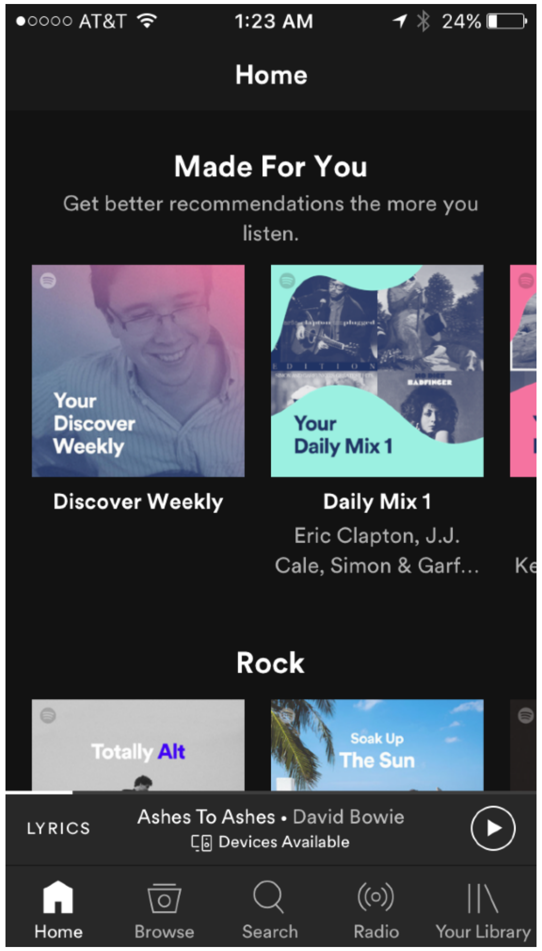
exploration parameter (when fixed -> crude exploitation; can also decay over time)

best action A* under preference model prediction

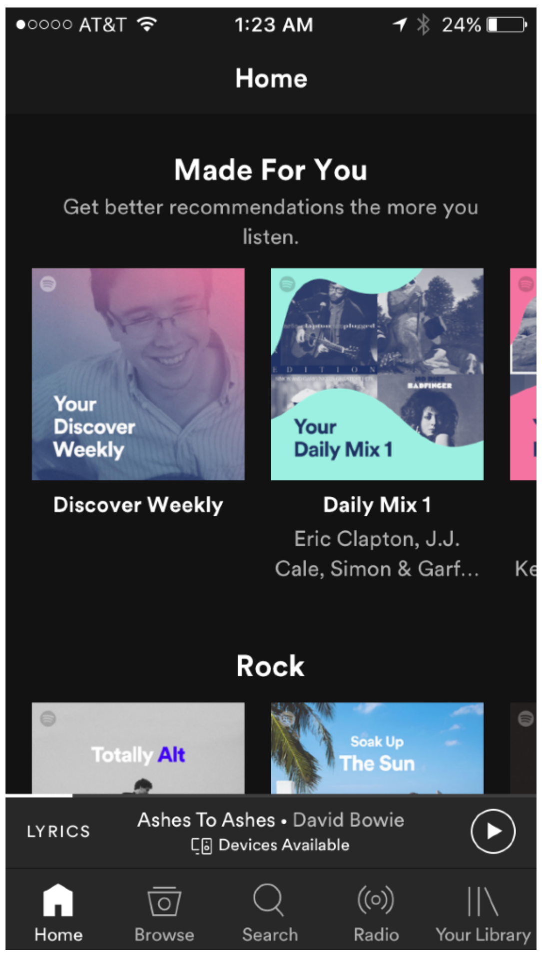


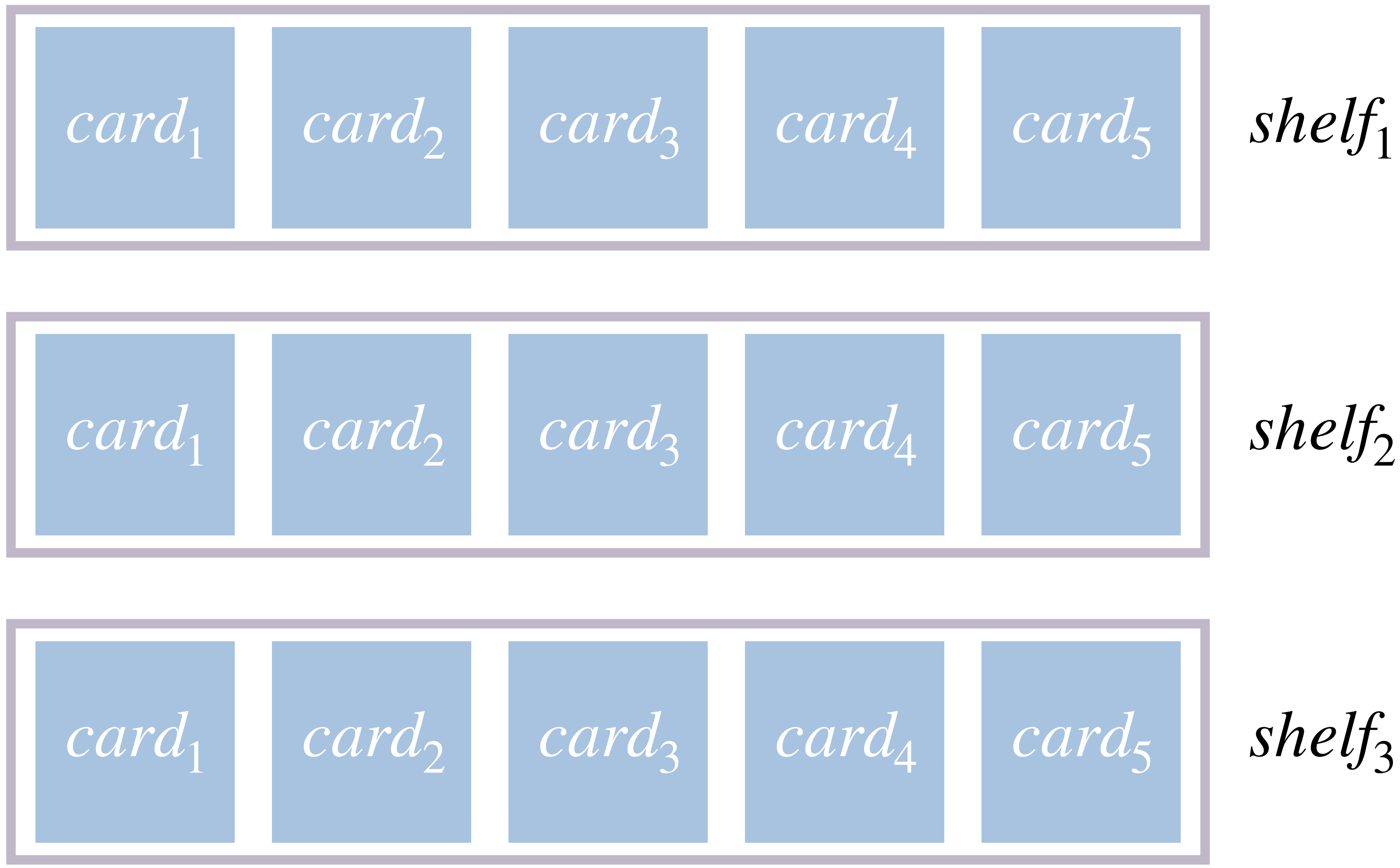
$$\pi(A \mid X) = \begin{cases} (1-\epsilon) + \frac{\epsilon}{|\mathscr{A}|} & \text{when } A = A^* \\ \frac{\epsilon}{|\mathscr{A}|} & \text{otherwise} \end{cases}$$

Research question: how to explore-exploit over explainable recommendations?

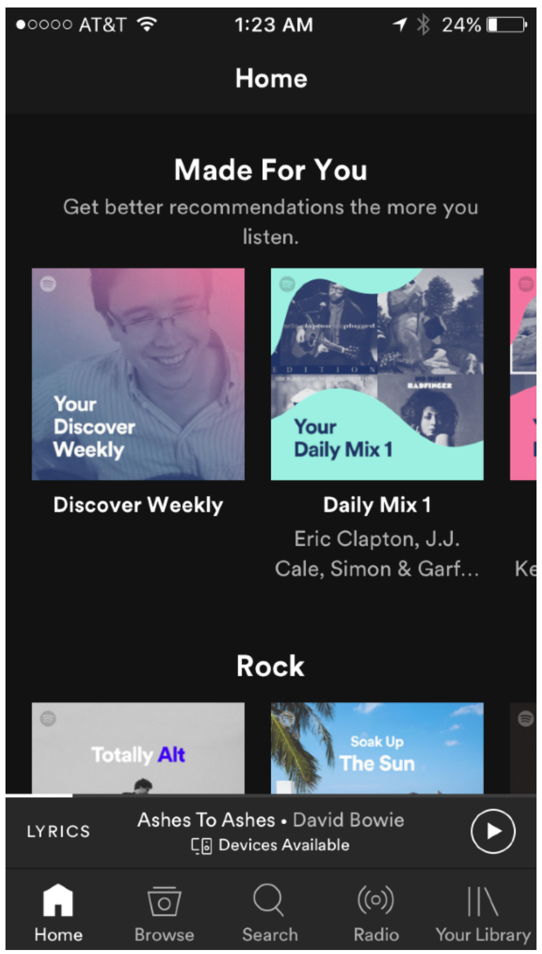


Research question: how to explore-exploit over explainable recommendations?

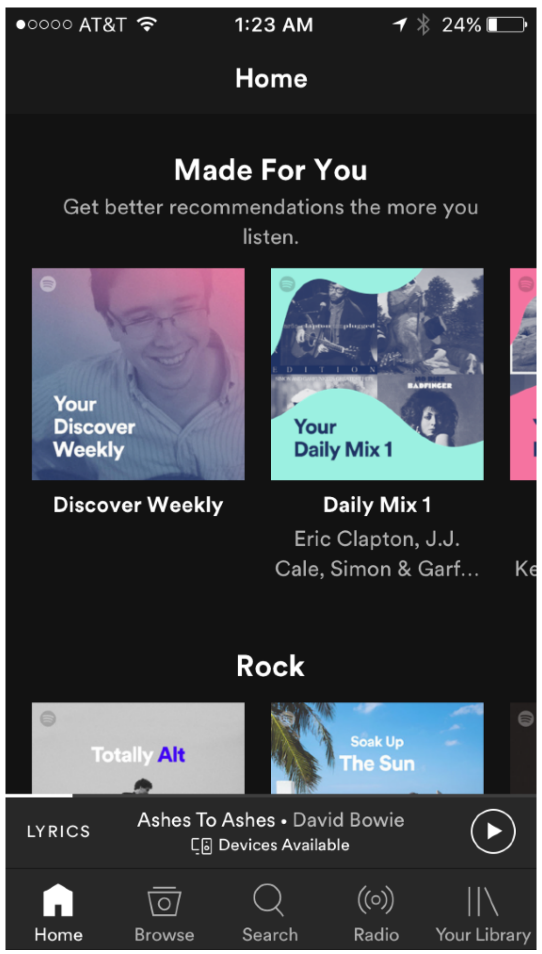




Research question: how to explore-exploit over explainable recommendations?

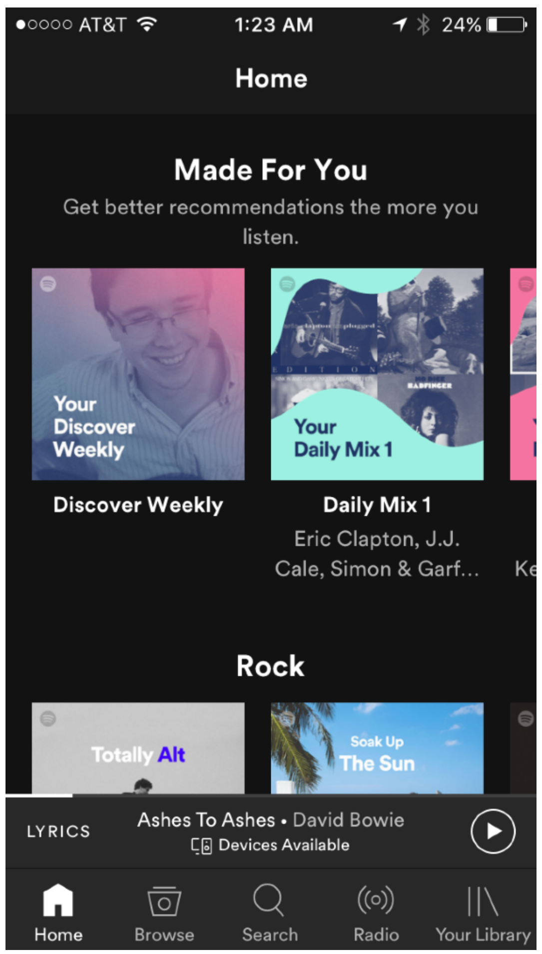


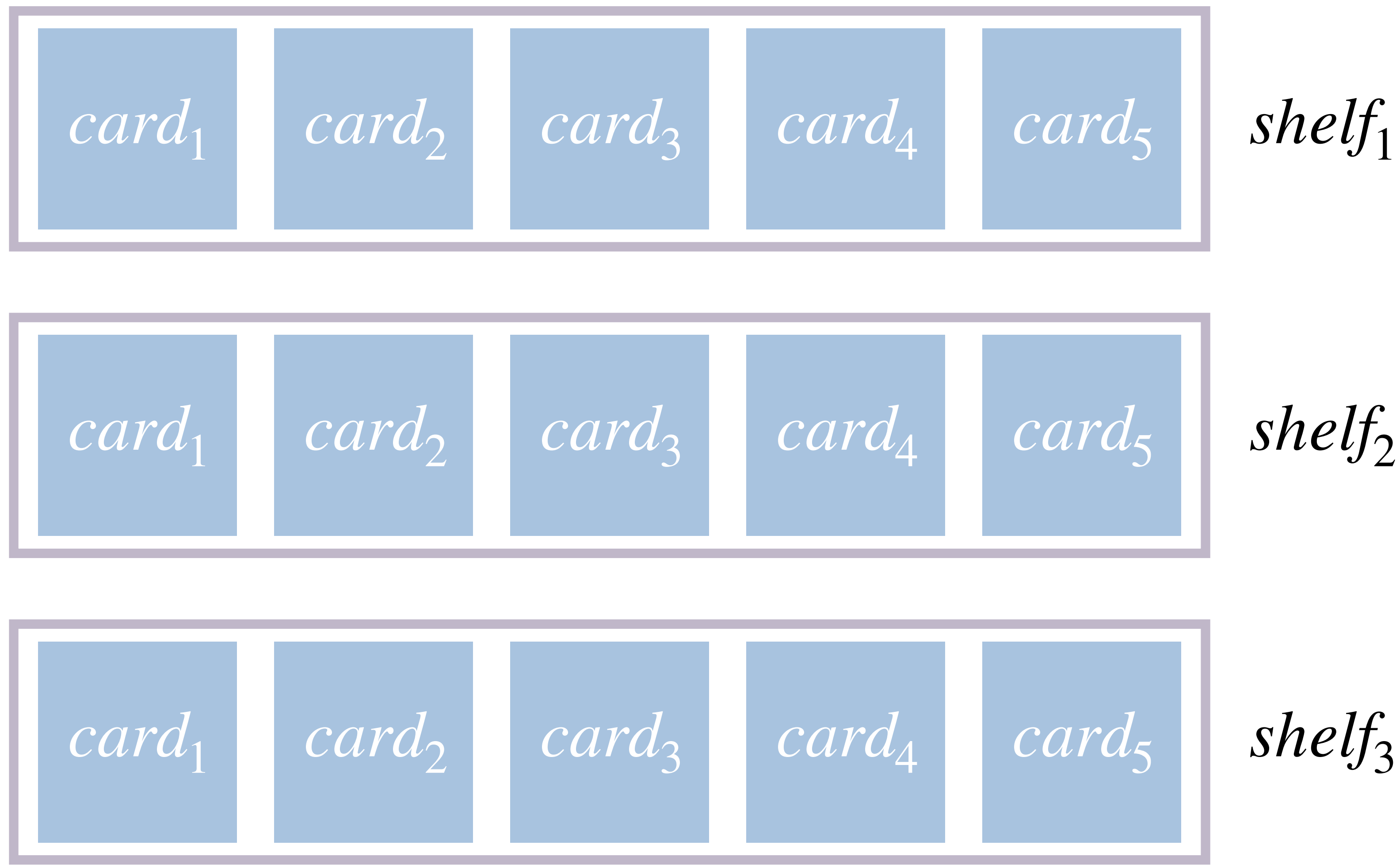
Research question: how to explore-exploit over explainable recommendations?



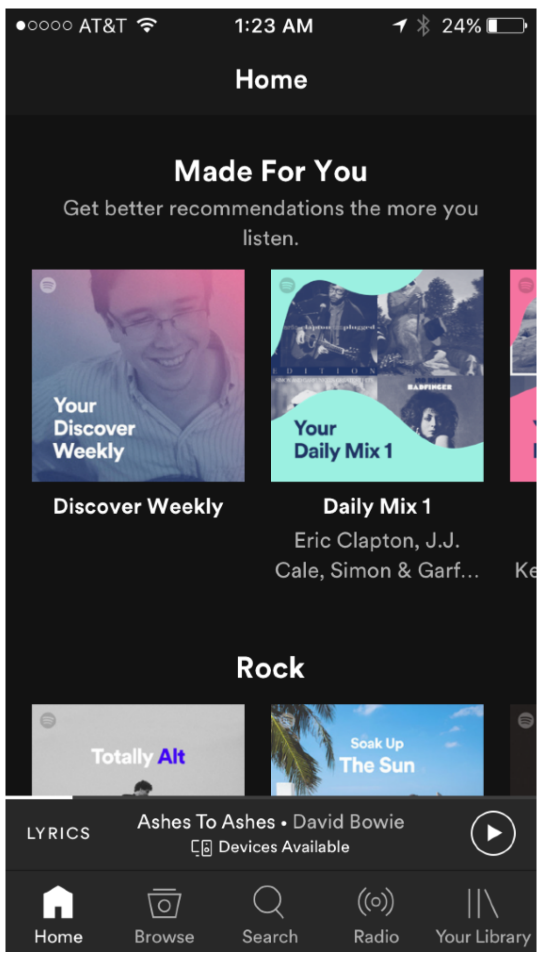
$shelf_1$

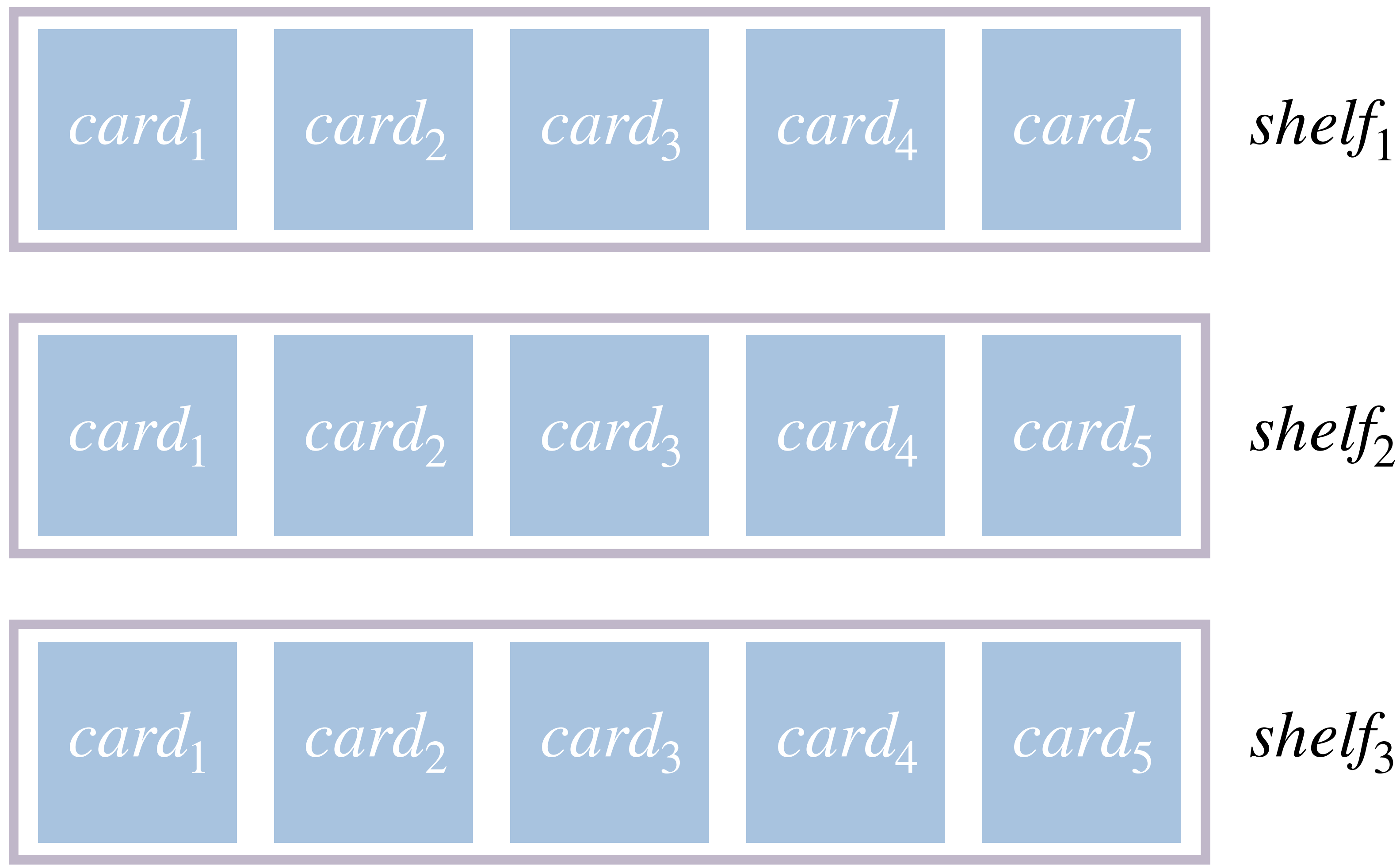
Research question: how to explore-exploit over explainable recommendations?





Research question: how to explore-exploit over explainable recommendations?





naively, the bandit has to try every possible combination of item and explanation many times before being able to exploit the best combinations

Bart

- Bart (bandits for recommendations as treatments) consists of:
  - a user preference model conditioned on the context
  - a ranking procedure + propensities
  - a training procedure

For details, see our new publication "Explore, Exploit, Explain" at RecSys
www.jamesmc.com/s/BartRecSys.pdf

Bart

- Bart (bandits for recommendations as treatments) consists of:
  - a user preference model conditioned on the context
  - a ranking procedure + propensities
  - a training procedure

factorization machine capturing interactions between features in a parameter efficient manner [Rendle, 2010]

For details, see our new publication "Explore, Exploit, Explain" at RecSys
www.jamesmc.com/s/BartRecSys.pdf

Bart

- Bart (bandits for recommendations as treatments) consists of:
  - a user preference model conditioned on the context
  - a ranking procedure + propensities
  - a training procedure

anything we know about the user and item, including region, age group, recent listening patterns, time of day

factorization machine capturing interactions between features in a parameter efficient manner [Rendle, 2010]

For details, see our new publication “Explore, Exploit, Explain” at RecSys
www.jamesmc.com/s/BartRecSys.pdf

Bart

- Bart (bandits for recommendations as treatments) consists of:
  - a user preference model conditioned on the context
  - a ranking procedure + propensities
  - a training procedure

anything we know about the user and item, including region, age group, recent listening patterns, time of day

counterfactual maximum likelihood [Joachims & Swaminathan, 2016]

factorization machine capturing interactions between features in a parameter efficient manner [Rendle, 2010]

For details, see our new publication "Explore, Exploit, Explain" at RecSys
www.jamesmc.com/s/BartRecSys.pdf

Bart

- Bart (bandits for recommendations as treatments) consists of:
  - a user preference model conditioned on the context
  - a ranking procedure + propensities
  - a training procedure

anything we know about the user and item, including region, age group, recent listening patterns, time of day

factorization machine capturing interactions between features in a parameter efficient manner [Rendle, 2010]

counterfactual maximum likelihood [Joachims & Swaminathan, 2016]

For details, see our new publication "Explore, Exploit, Explain" at RecSys
www.jamesmc.com/s/BartRecSys.pdf

Ranking Procedure

Let's make our lives easy: aim to train user preference model on logged impressions assumed independent given context.

impression_id	card_id	shelf_id	context	streamed?
0	101	0	Stockholm	No
1	3	0	Stockholm	Yes
2	45	1	Stockholm	No
3	99	1	New York	No
4	11	0	New York	Yes

Ranking Procedure

Let's make our lives easy: aim to train user preference model on logged impressions assumed independent given context.

impression_id	card_id	shelf_id	context	streamed?
0	101	0	Stockholm	No
1	3	0	Stockholm	Yes
2	45	1	Stockholm	No
3	99	1	New York	No
4	11	0	New York	Yes

What set of bandit assumptions lead to this procedure?

Ranking procedure

Ranking procedure

Assumptions of shelf browsing model

Horizontal scrolling

Ranking procedure

Assumptions of shelf browsing model

Horizontal scrolling

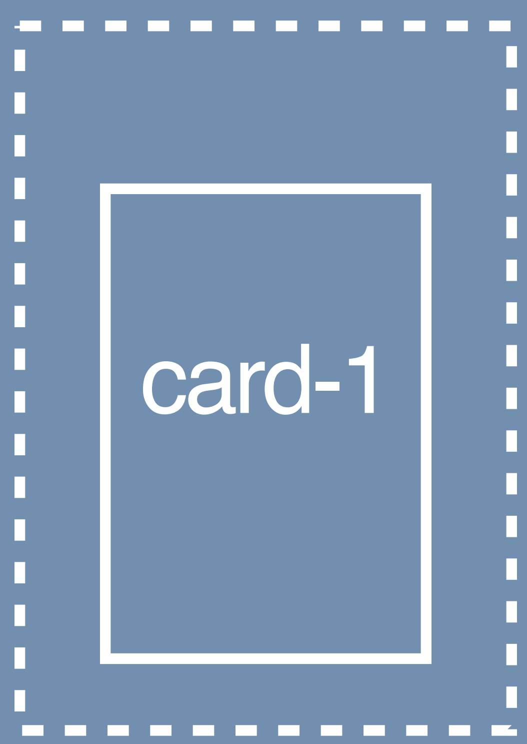
User awareness

Ranking procedure

Assumptions of shelf browsing model

Horizontal scrolling

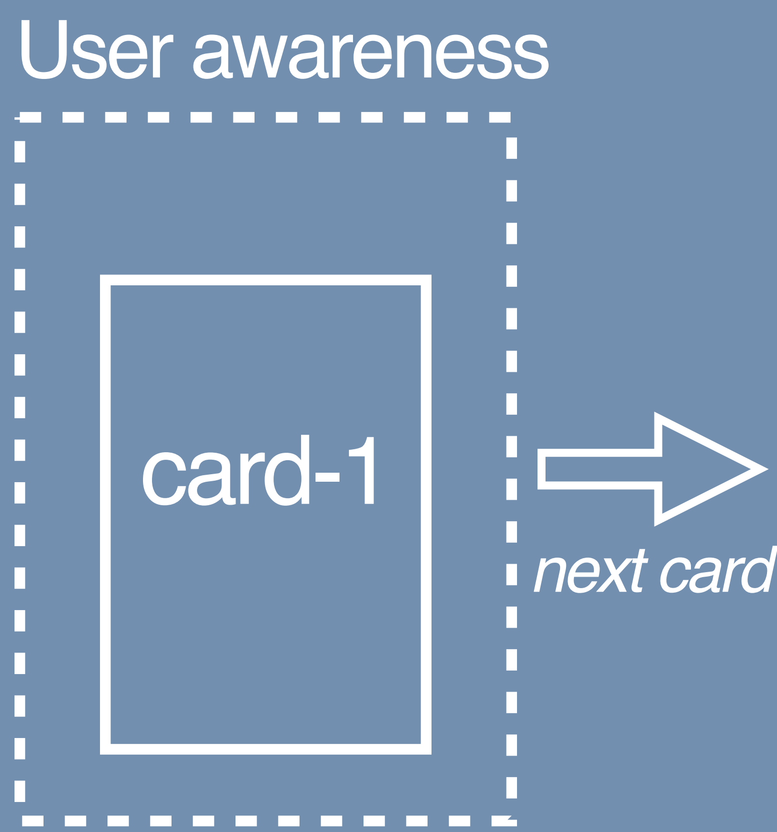
User awareness



Ranking procedure

Assumptions of shelf browsing model

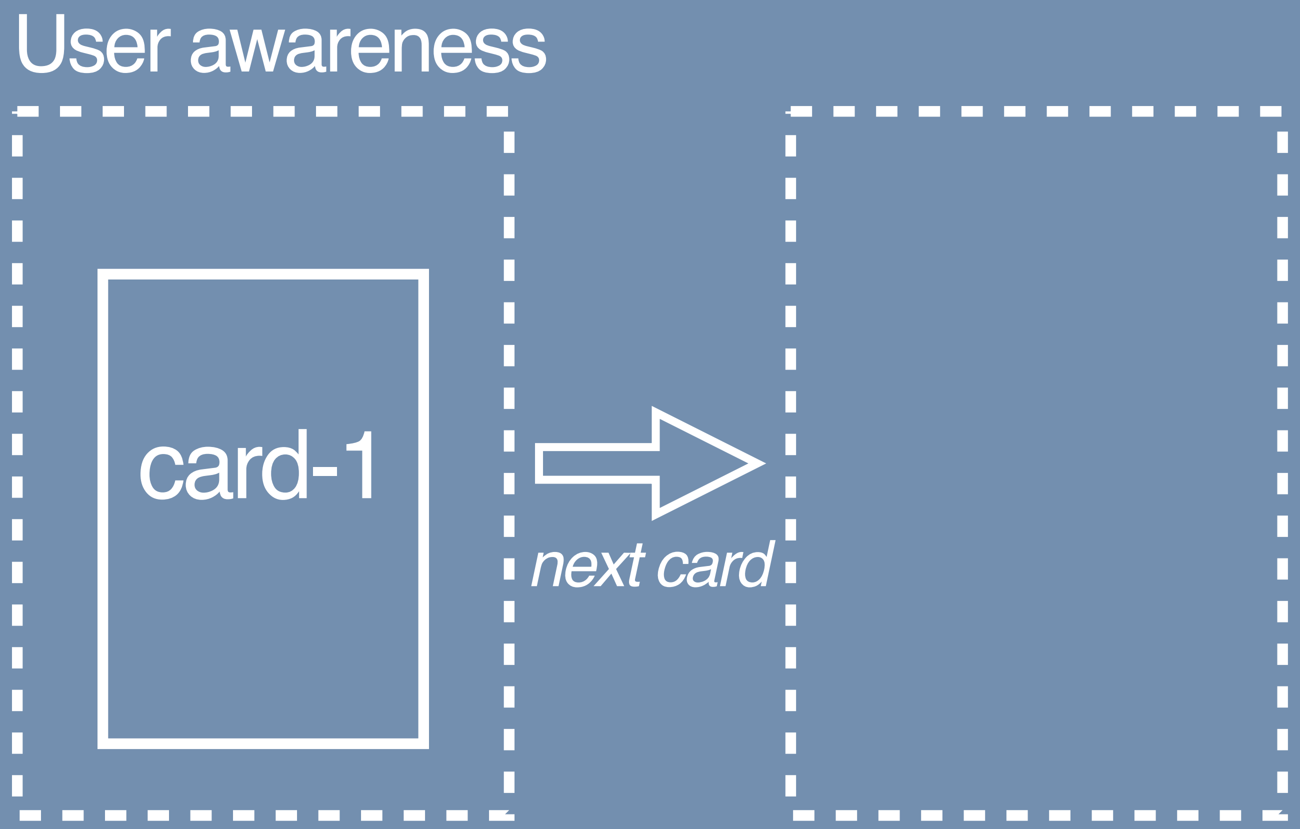
Horizontal scrolling



Ranking procedure

Assumptions of shelf browsing model

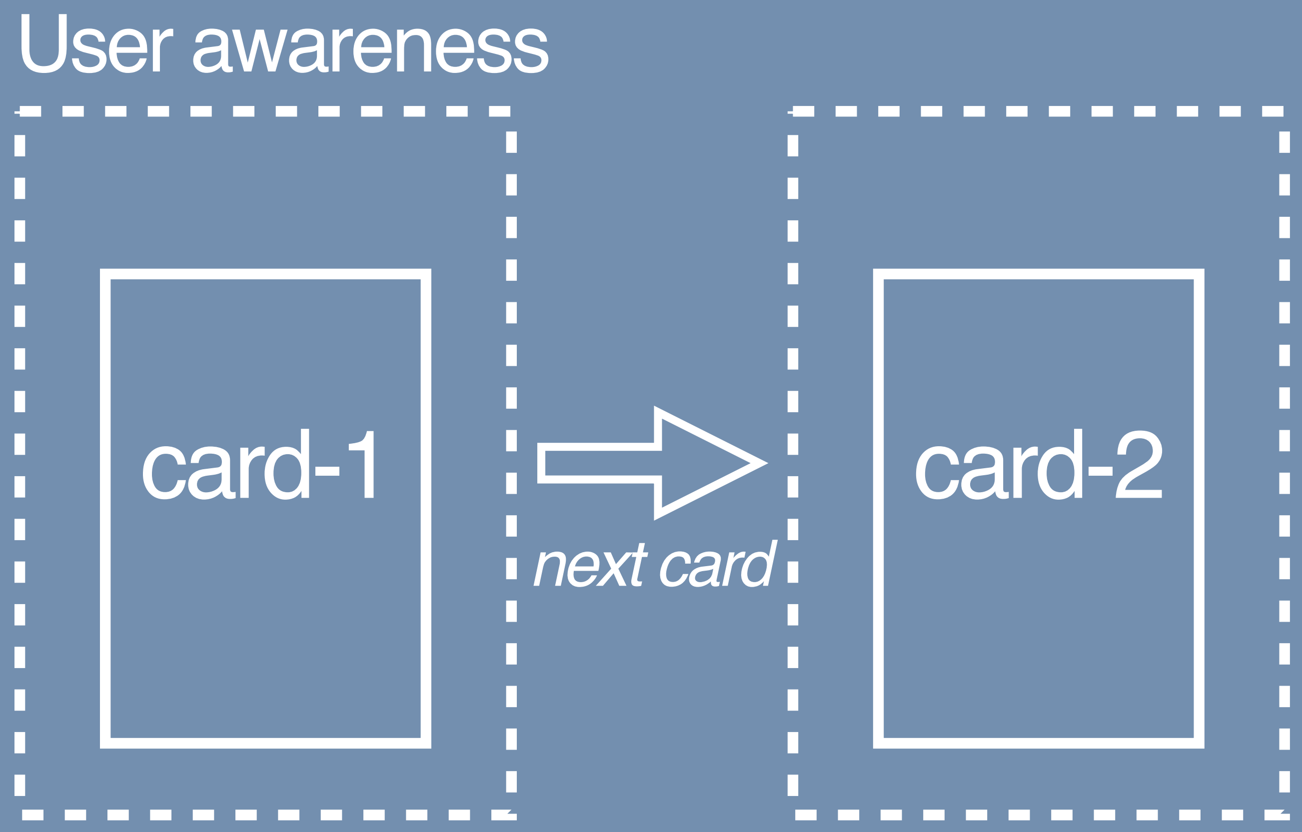
Horizontal scrolling



Ranking procedure

Assumptions of shelf browsing model

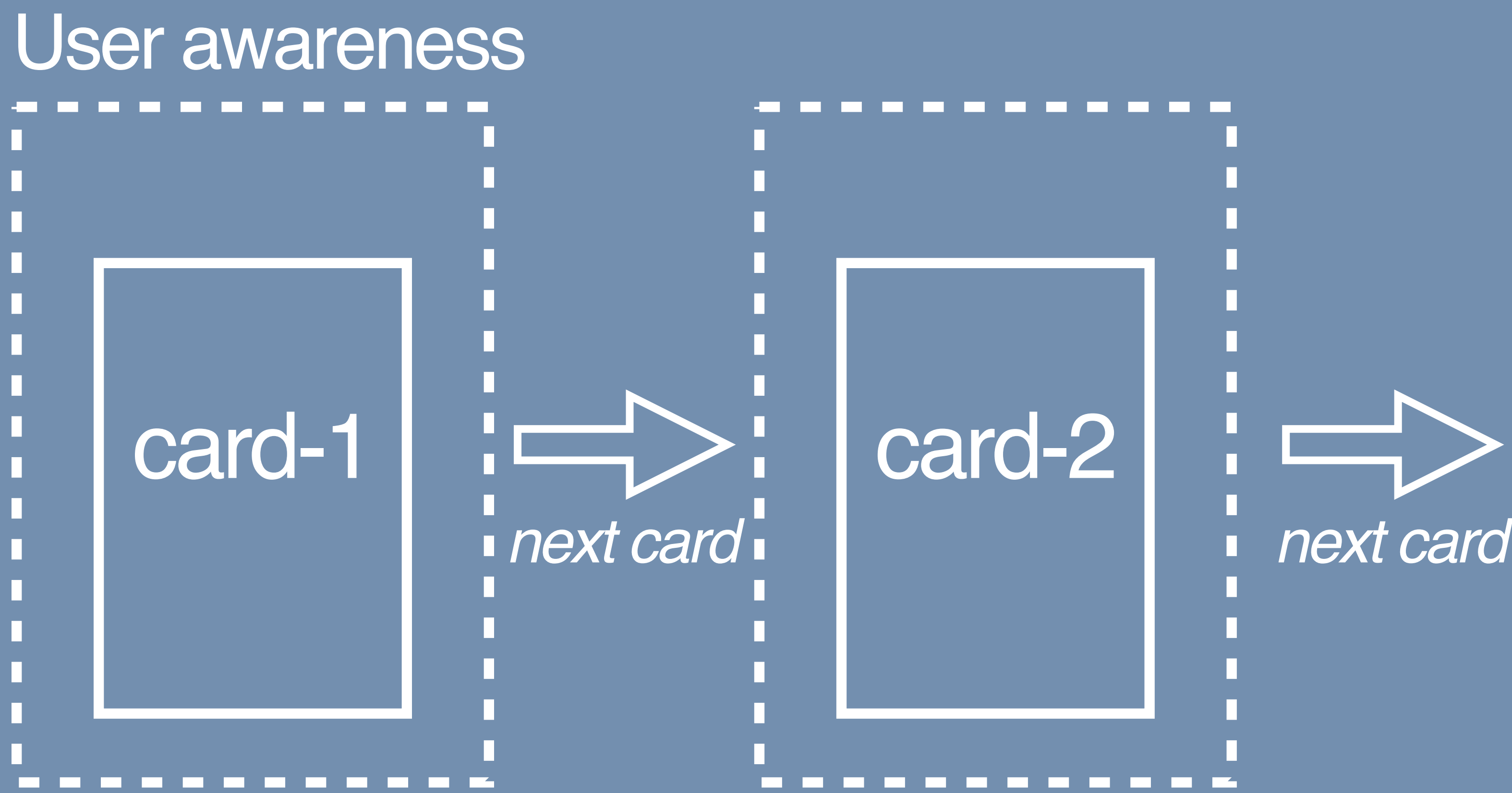
Horizontal scrolling



Ranking procedure

Assumptions of shelf browsing model

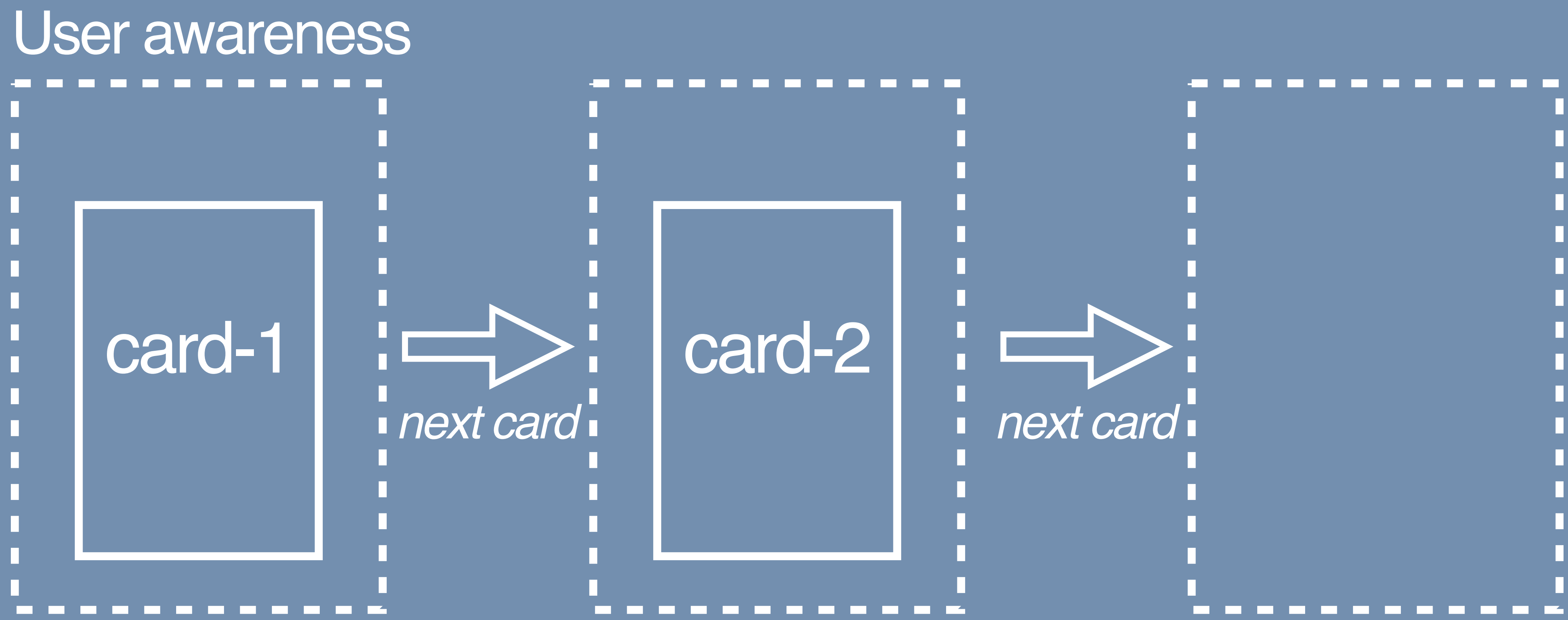
Horizontal scrolling



Ranking procedure

Assumptions of shelf browsing model

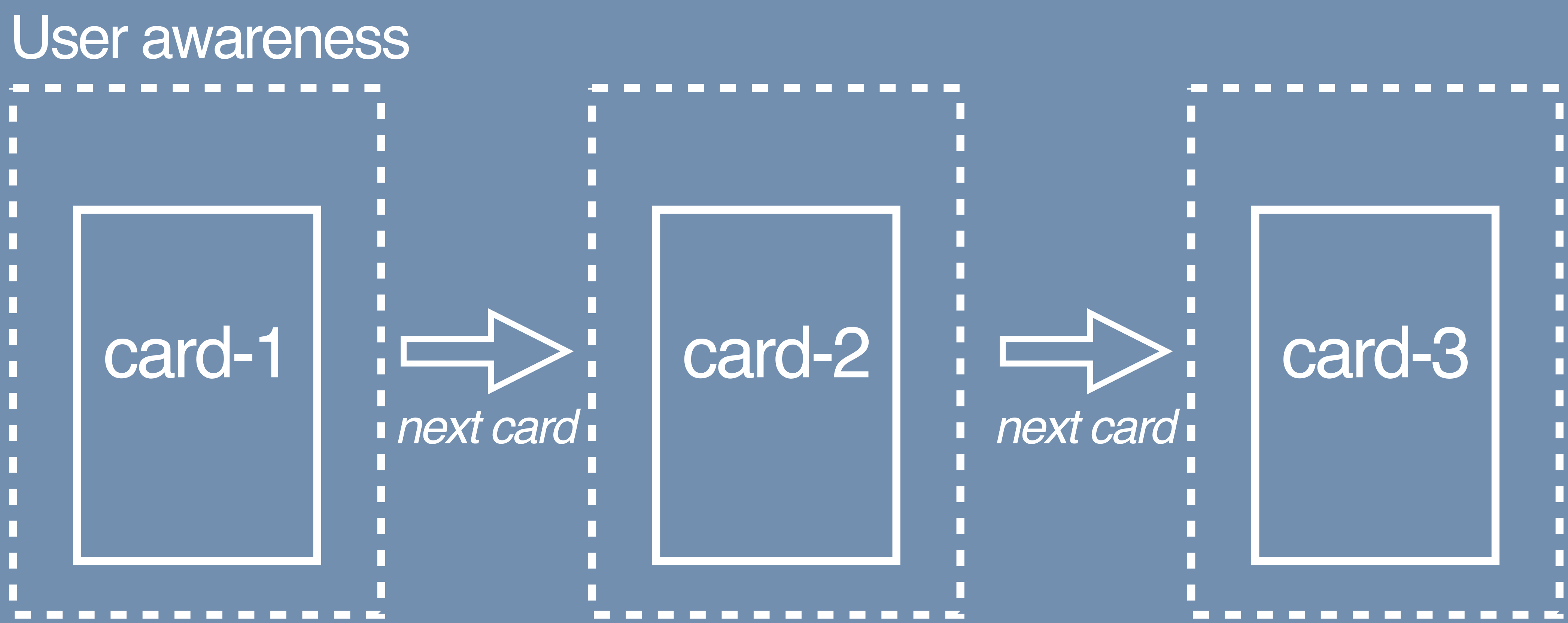
Horizontal scrolling



Ranking procedure

Assumptions of shelf browsing model

Horizontal scrolling



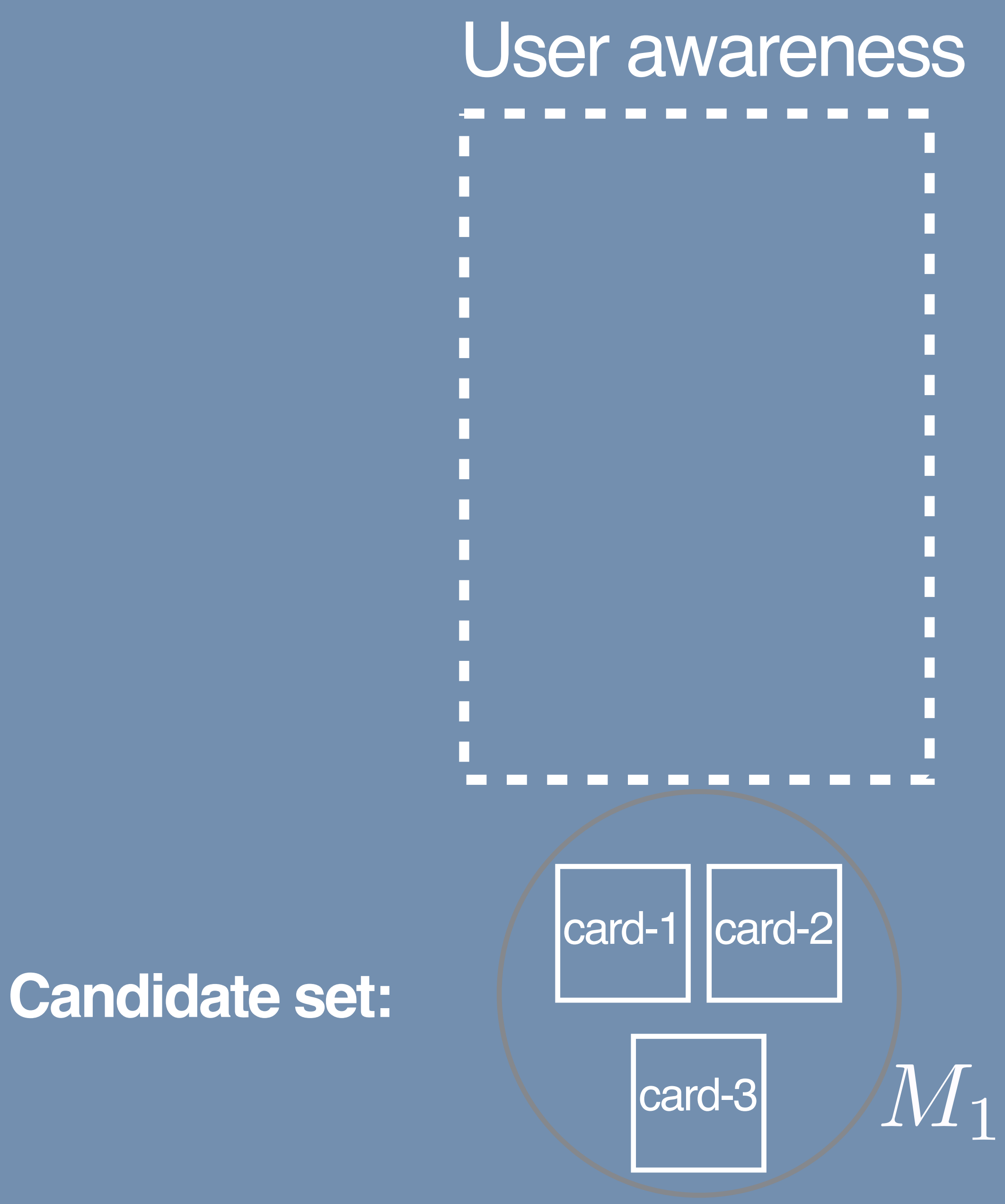
Ranking procedure with bandit

Horizontal scrolling

User awareness

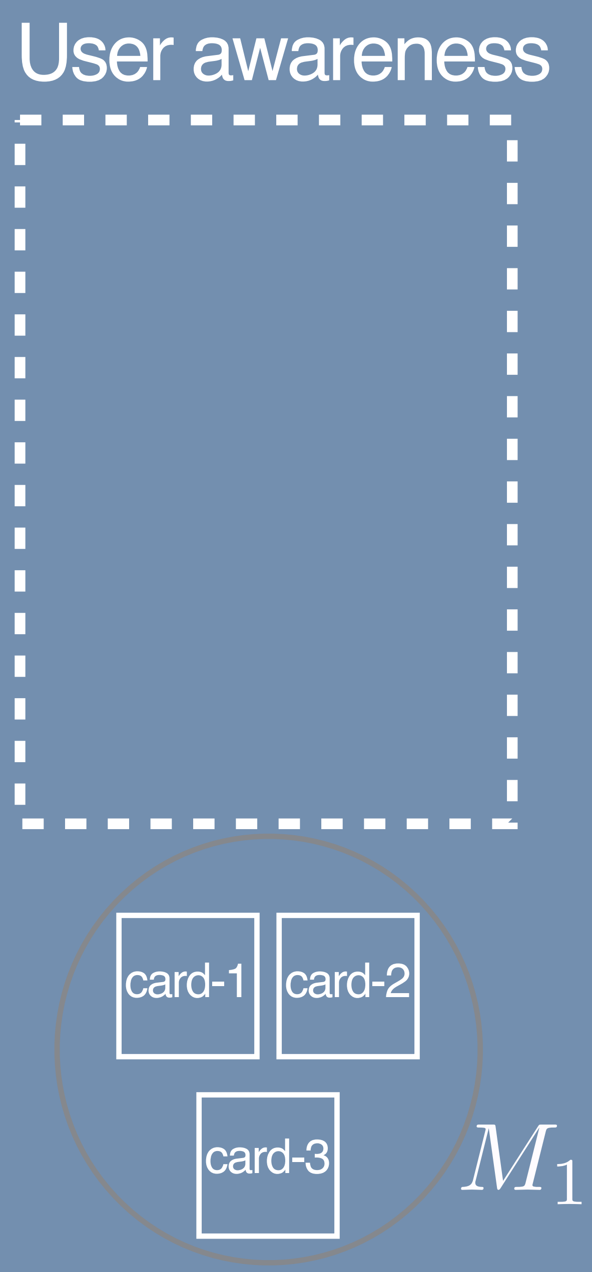
Ranking procedure with bandit

Horizontal scrolling



Ranking procedure with bandit

Horizontal scrolling

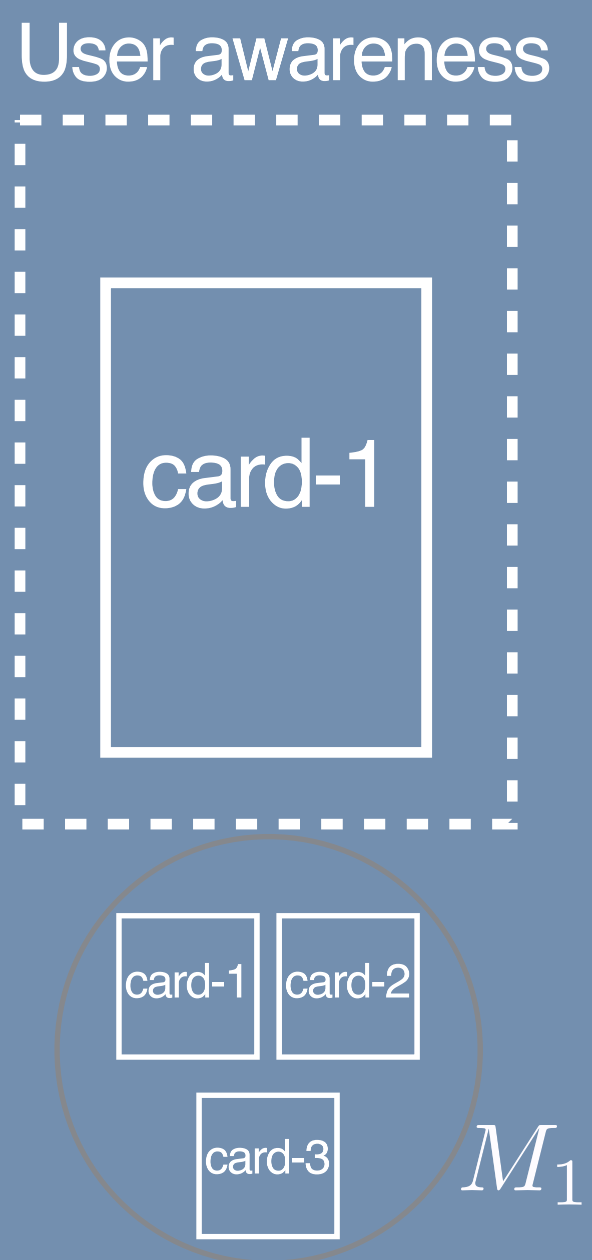


Candidate set:

Action select: $\mathrm{card}_1 \sim \pi_{s,r}(M_1)$

Ranking procedure with bandit

Horizontal scrolling

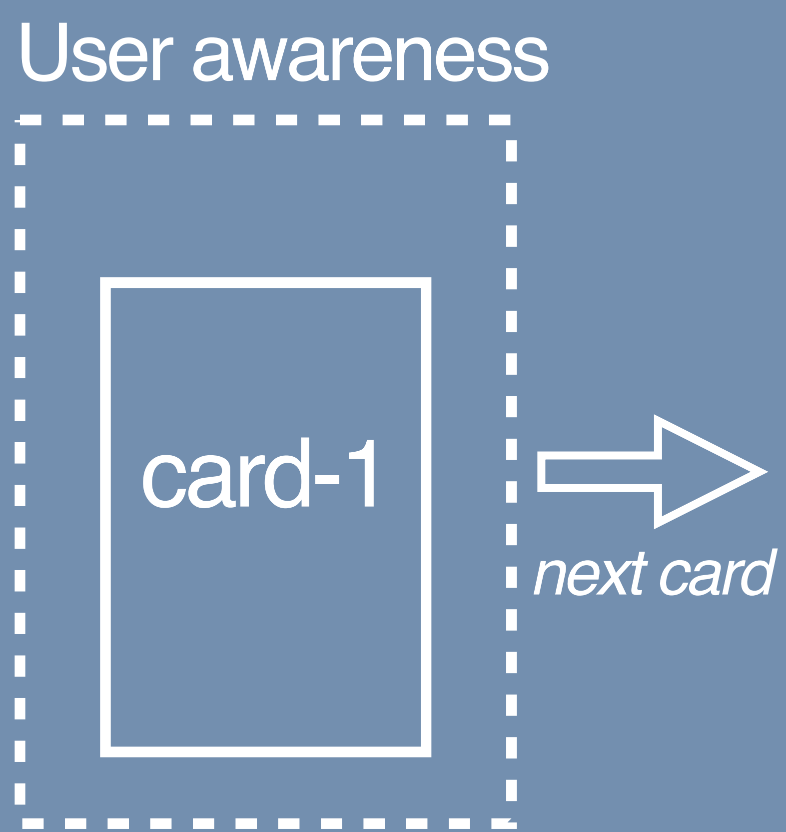


Candidate set:

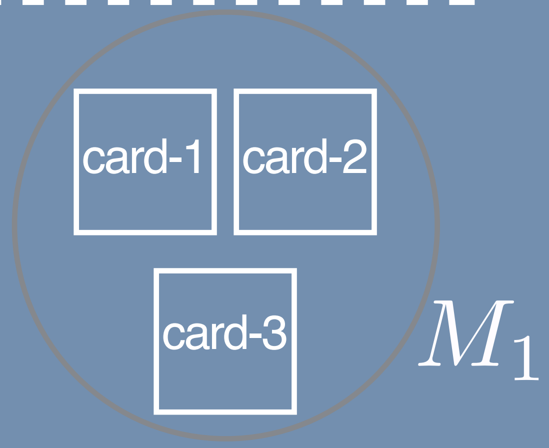
Action select: $\mathrm{card}_1 \sim \pi_{s,r}(M_1)$

Ranking procedure with bandit

Horizontal scrolling



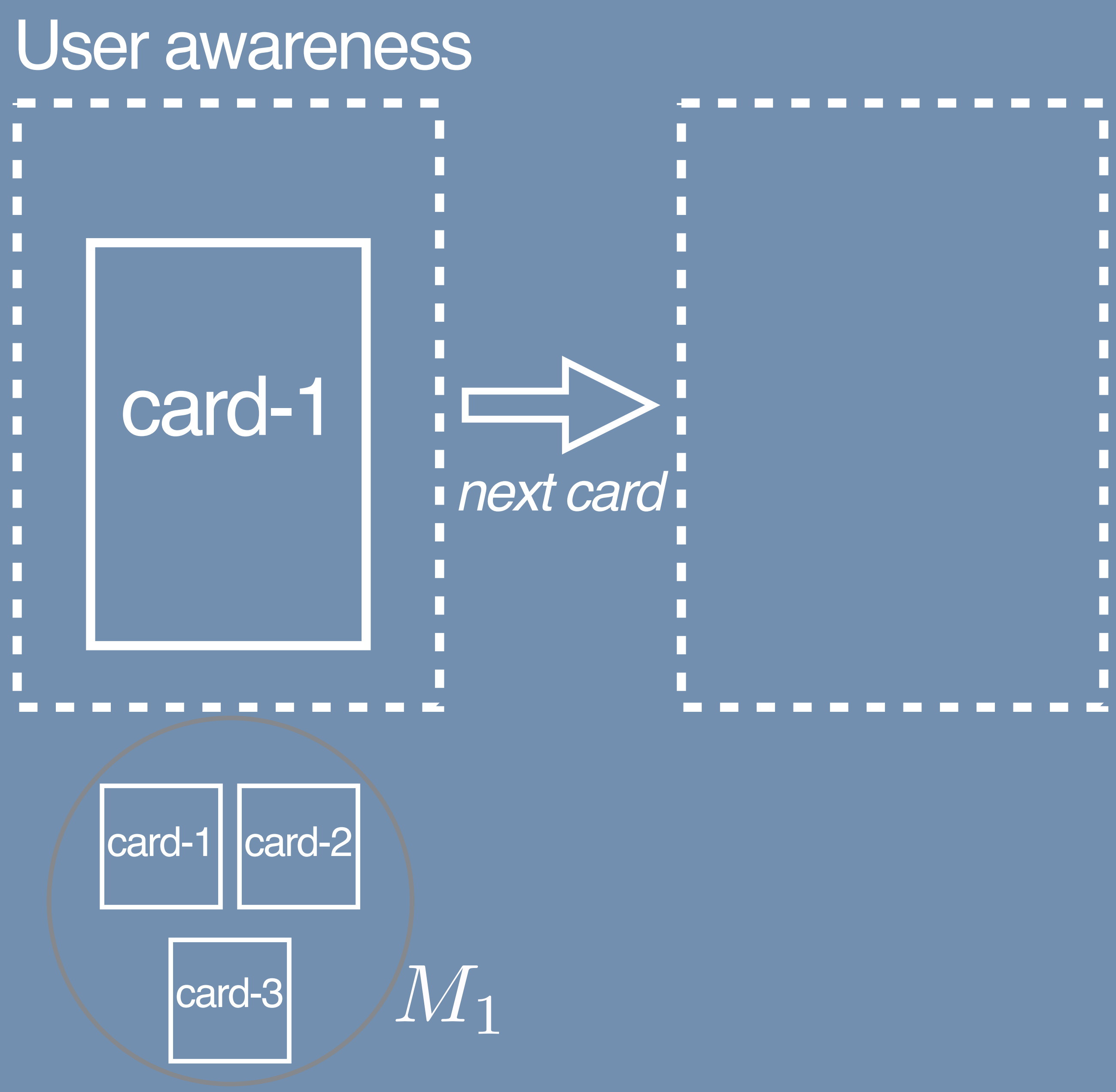
Candidate set:



Action select: $\mathrm{card}_1 \sim \pi_{s,r}(M_1)$

Ranking procedure with bandit

Horizontal scrolling



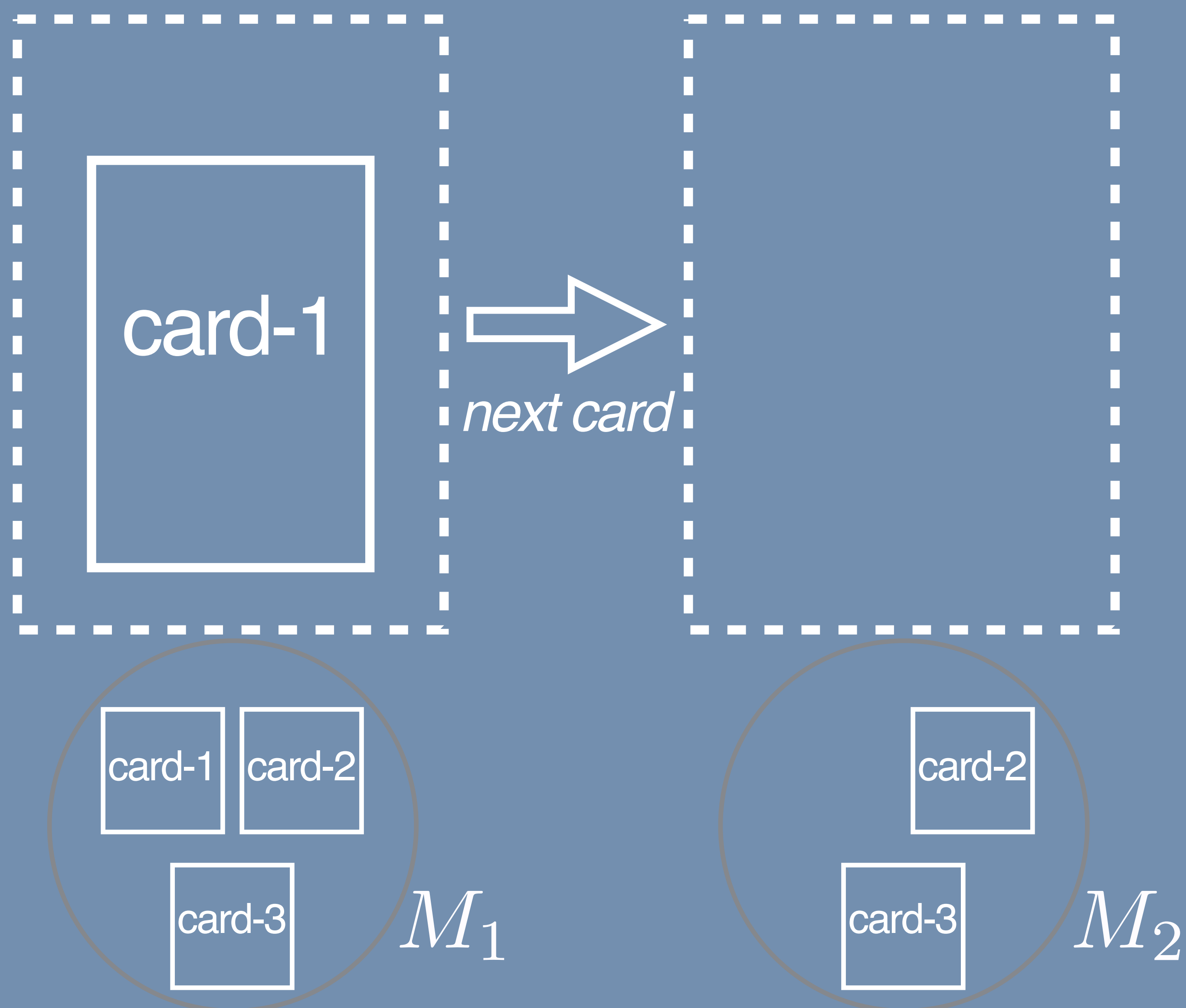
Candidate set:

Action select: $\mathrm{card}_1 \sim \pi_{s,r}(M_1)$

Ranking procedure with bandit

Horizontal scrolling

User awareness

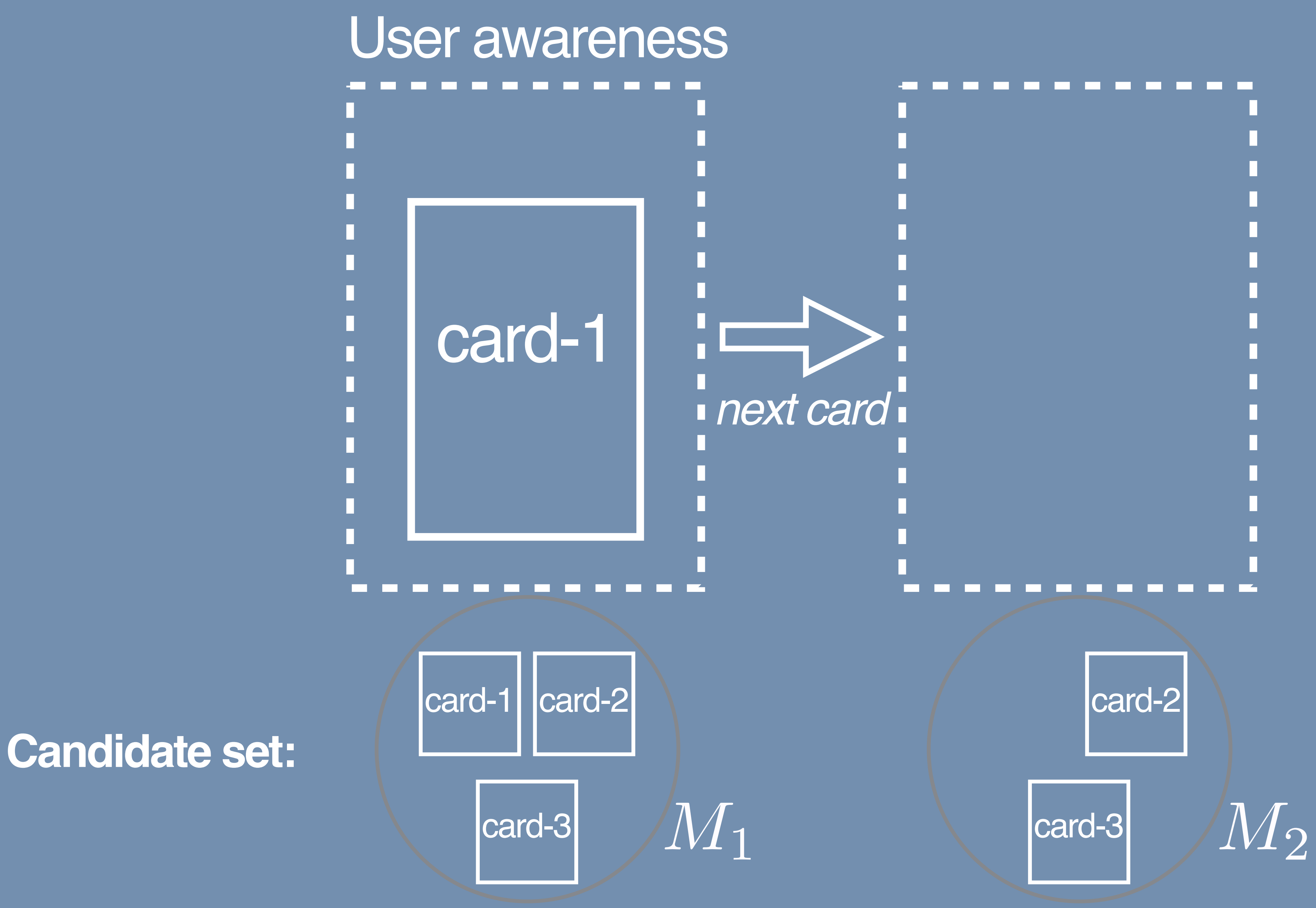


Candidate set:

Action select: $\text{card}_1 \sim \pi_{s,r}(M_1)$

Ranking procedure with bandit

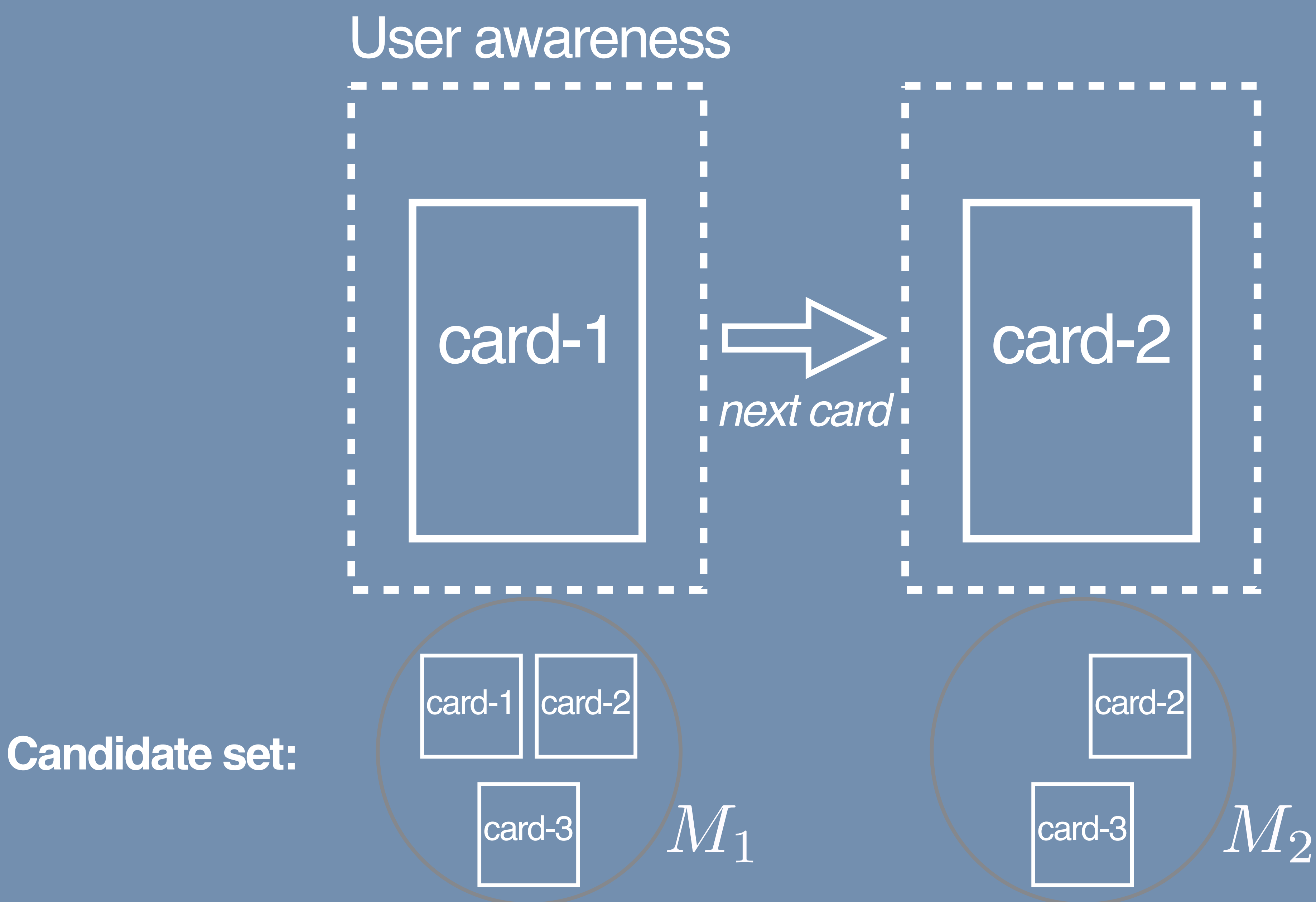
Horizontal scrolling



Action select: $\text{card}_1 \sim \pi_{s,r}(M_1)$ $\text{card}_2 \sim \pi_{s,r}(M_2)$

Ranking procedure with bandit

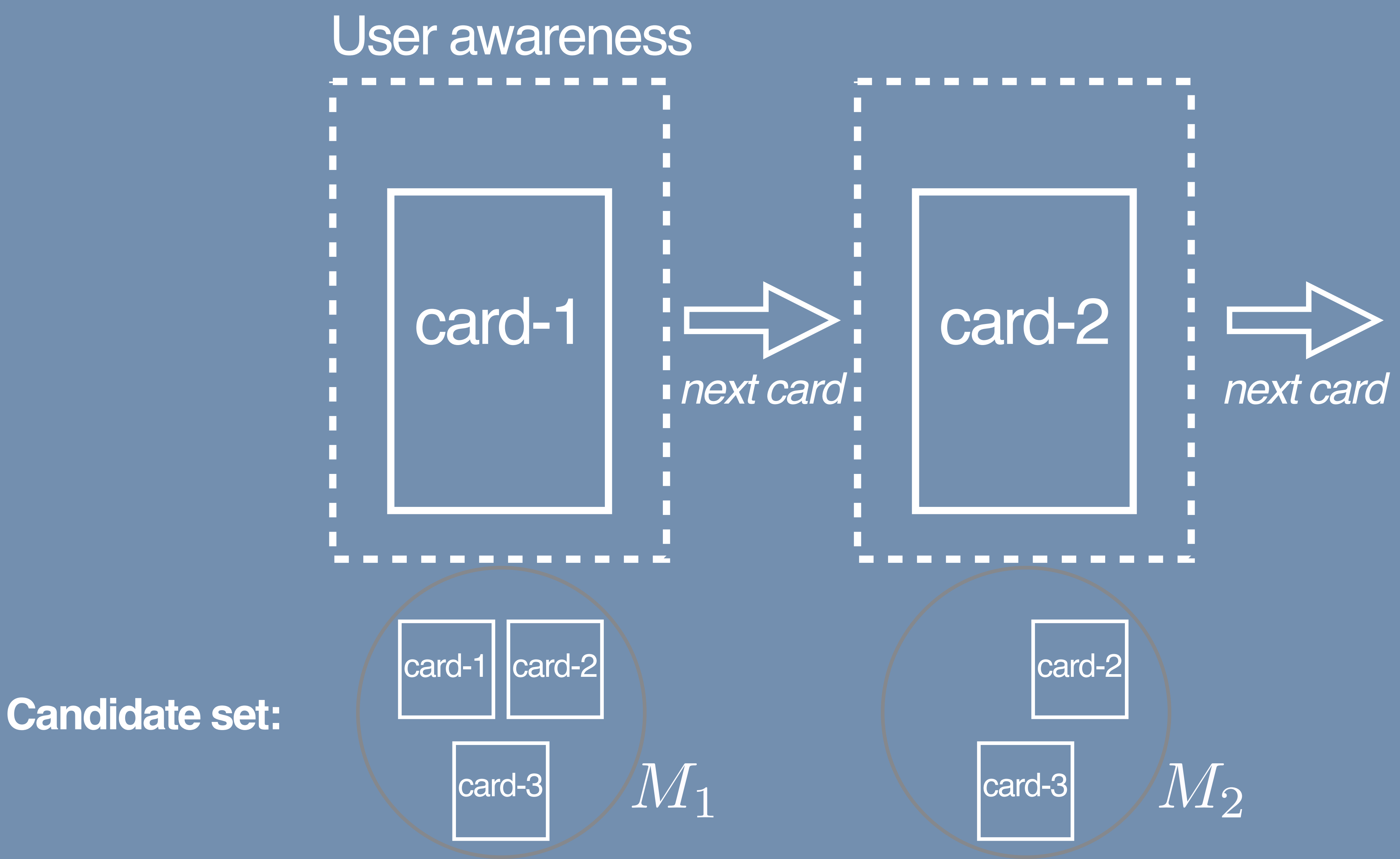
Horizontal scrolling



Action select: $\text{card}_1 \sim \pi_{s,r}(M_1)$ $\text{card}_2 \sim \pi_{s,r}(M_2)$

Ranking procedure with bandit

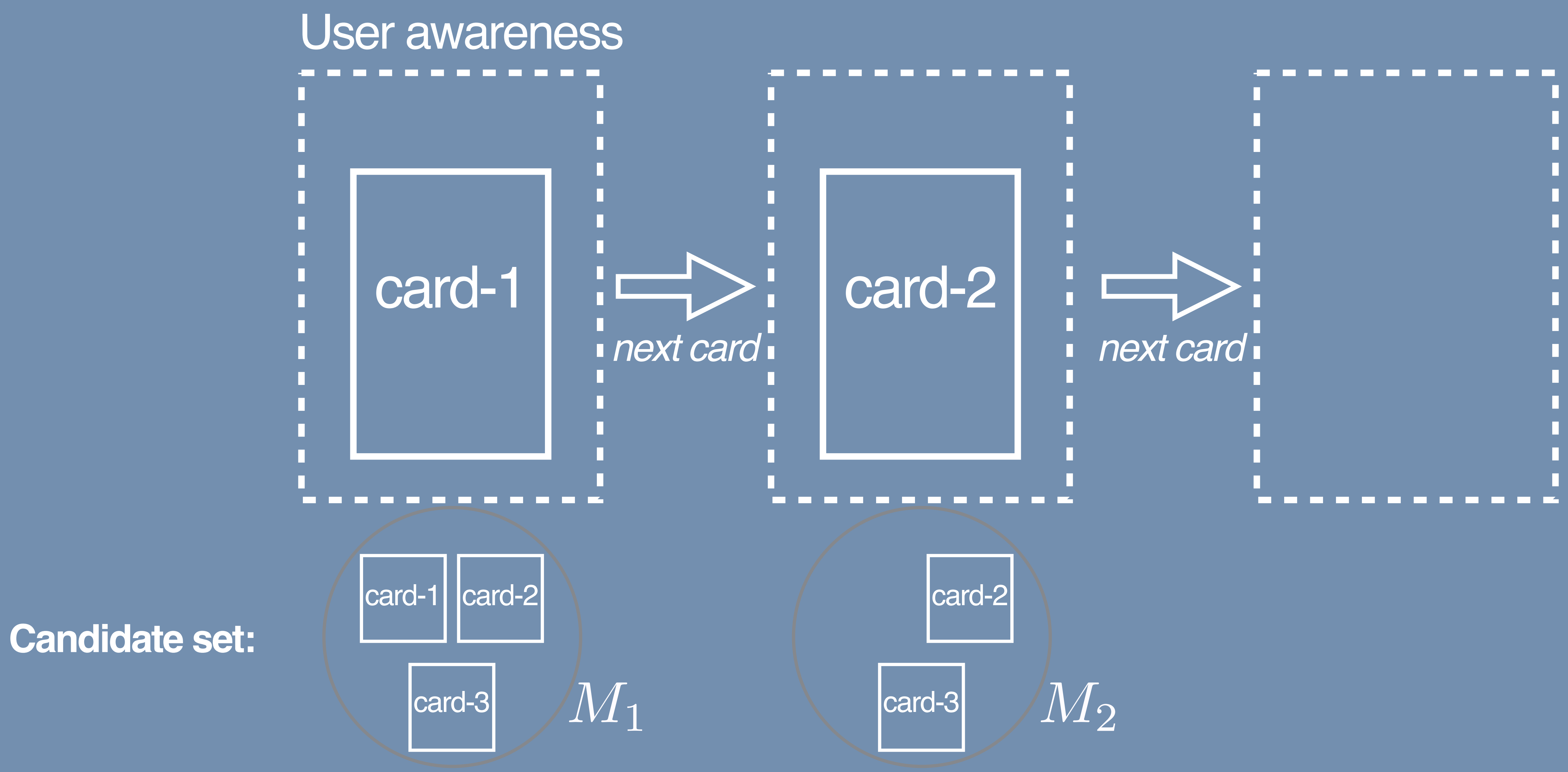
Horizontal scrolling



Action select: $\text{card}_1 \sim \pi_{s,r}(M_1)$ $\text{card}_2 \sim \pi_{s,r}(M_2)$

Ranking procedure with bandit

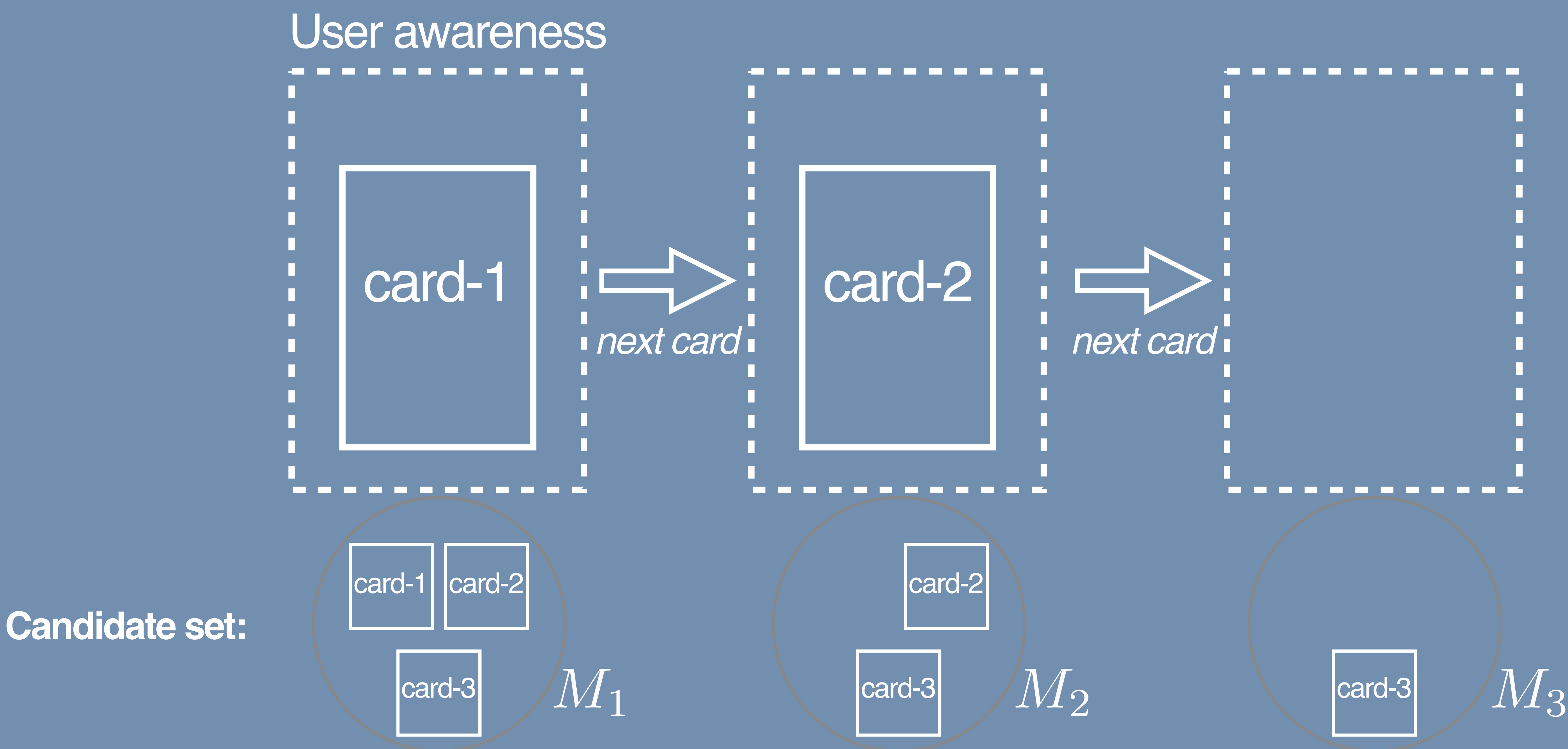
Horizontal scrolling



Action select: $\text{card}_1 \sim \pi_{s,r}(M_1)$ $\text{card}_2 \sim \pi_{s,r}(M_2)$

Ranking procedure with bandit

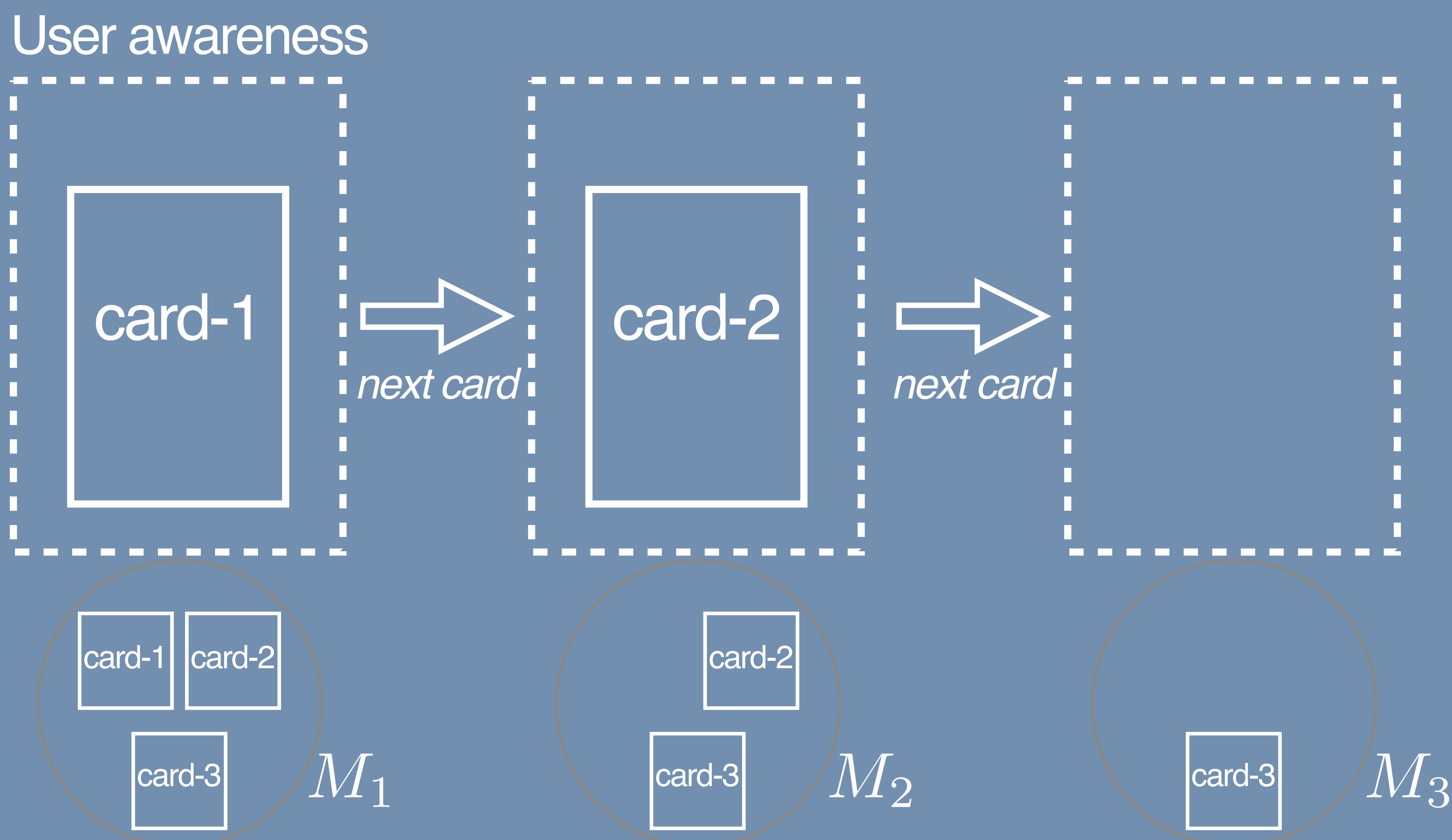
Horizontal scrolling



Action select: $\mathrm{card}_1 \sim \pi_{s,r}(M_1)$ $\mathrm{card}_2 \sim \pi_{s,r}(M_2)$

Ranking procedure with bandit

Horizontal scrolling

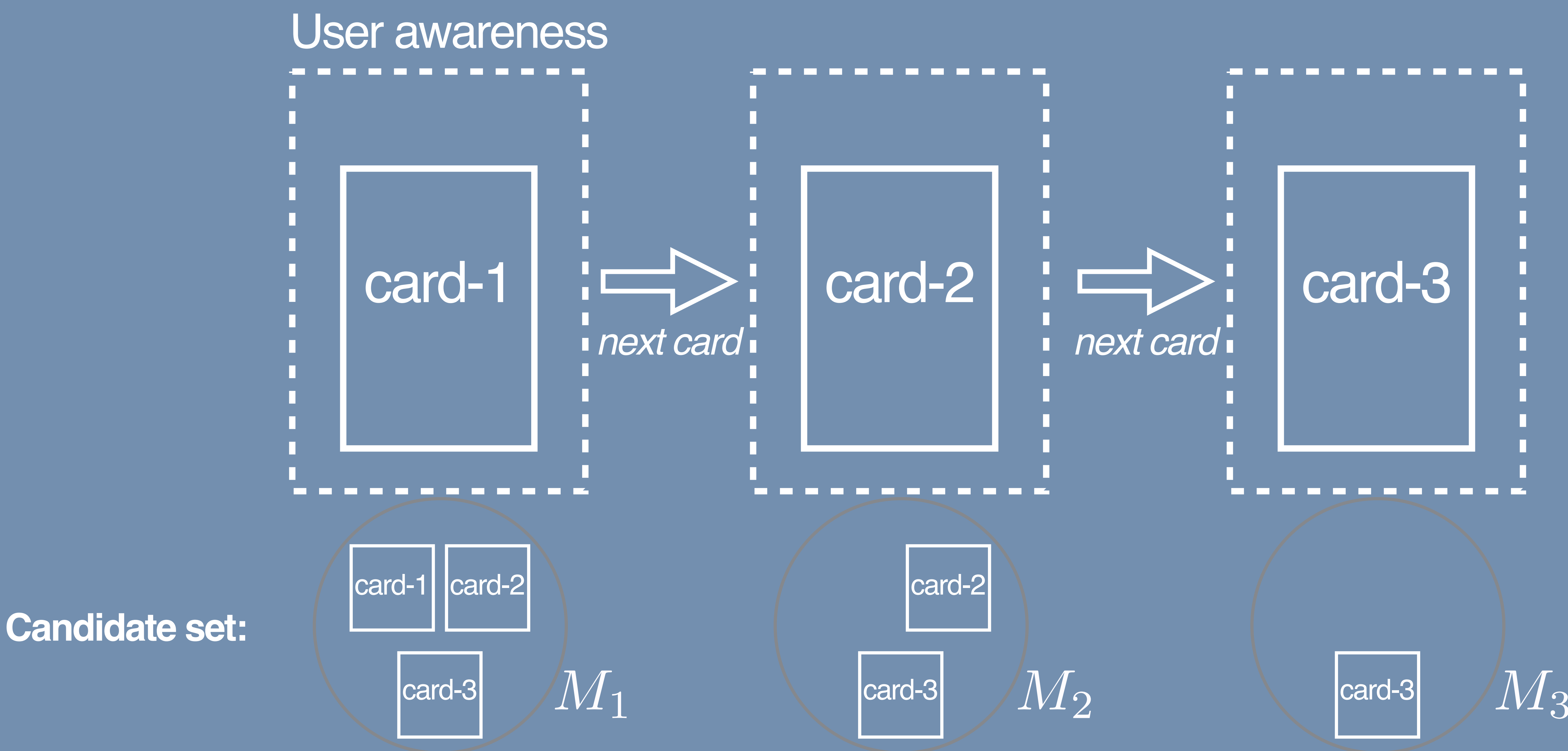


Candidate set:

Action select: $\text{card}_1 \sim \pi_{s,r}(M_1)$ $\text{card}_2 \sim \pi_{s,r}(M_2)$ $\text{card}_3 \sim \pi_{s,r}(M_3)$

Ranking procedure with bandit

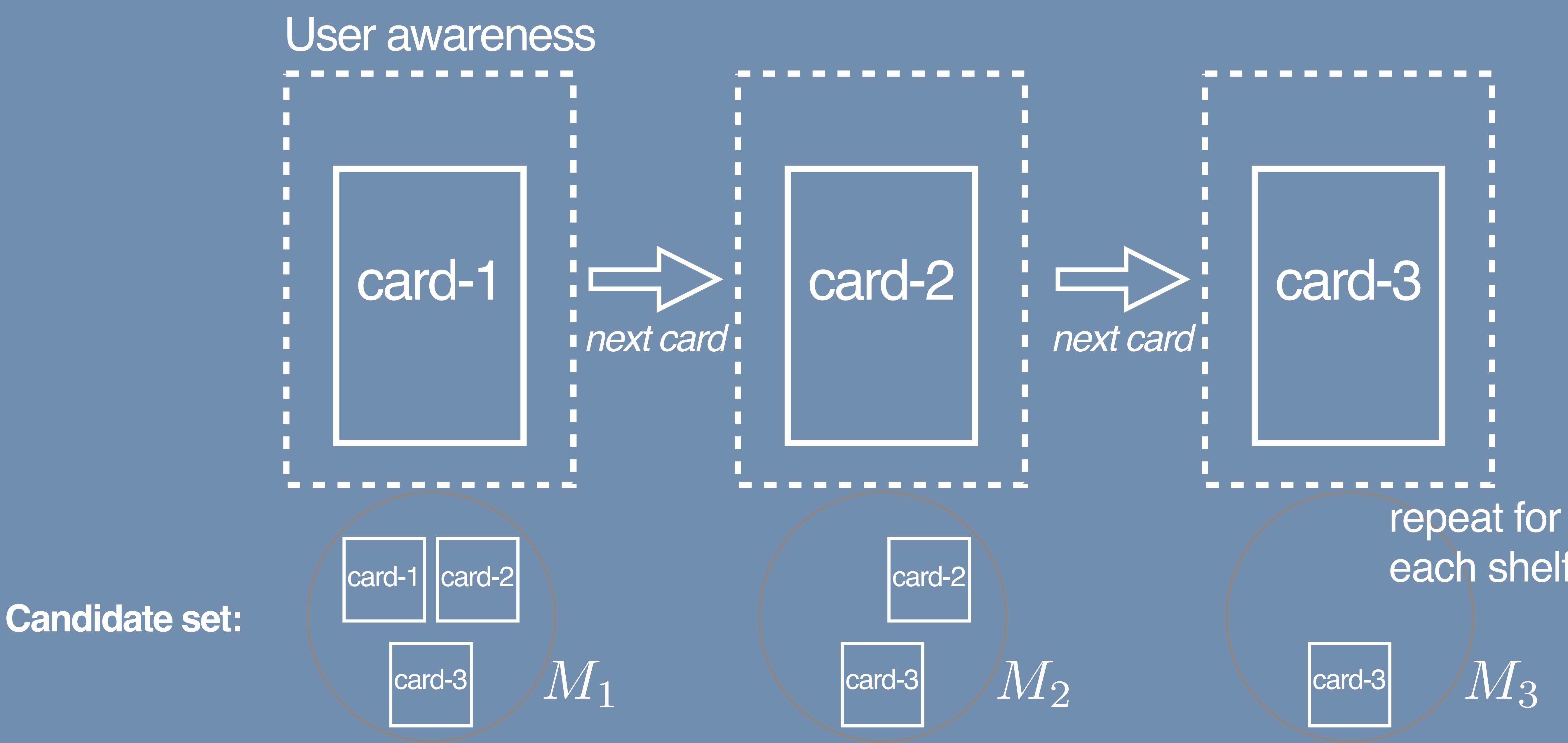
Horizontal scrolling



Action select: $\mathrm{card}_1 \sim \pi_{s,r}(M_1)$ $\mathrm{card}_2 \sim \pi_{s,r}(M_2)$ $\mathrm{card}_3 \sim \pi_{s,r}(M_3)$

Ranking procedure with bandit

Horizontal scrolling



Action select: $\mathrm{card}_1 \sim \pi_{s,r}(M_1)$ $\mathrm{card}_2 \sim \pi_{s,r}(M_2)$ $\mathrm{card}_3 \sim \pi_{s,r}(M_3)$

Ranking procedure with bandit

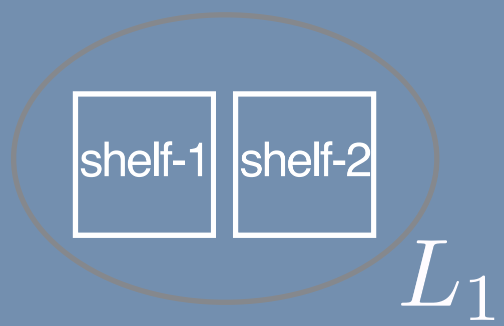
Vertical scrolling

User awareness

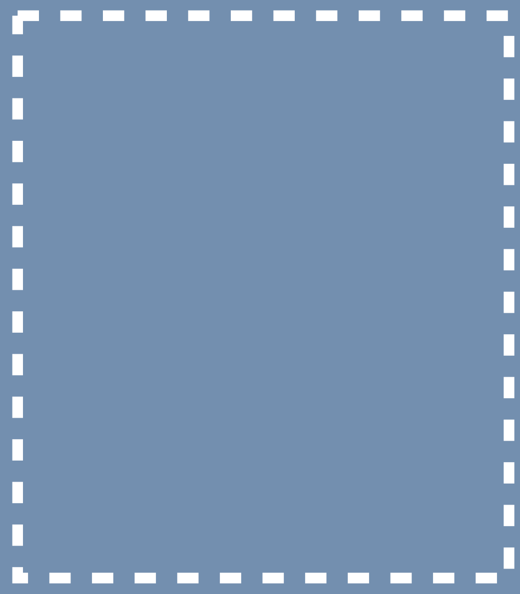
Ranking procedure with bandit

Vertical scrolling

Candidate set



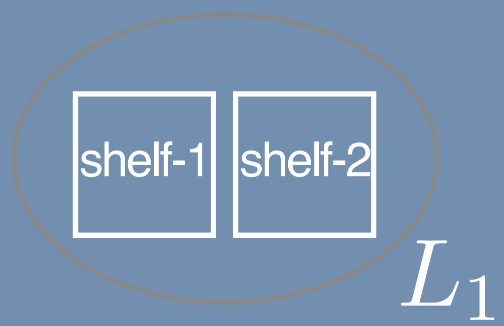
User awareness



Ranking procedure with bandit

Vertical scrolling

Candidate set



Action select

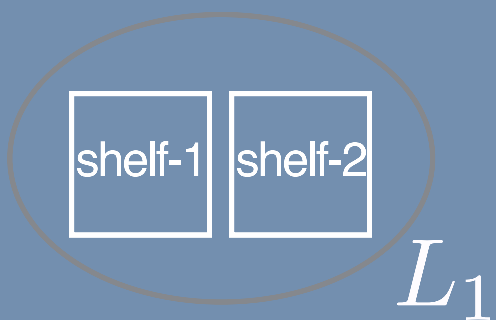
$$\text{shelf}_1 \sim \pi_{s,r'}(L_1)$$

User awareness

Ranking procedure with bandit

Vertical scrolling

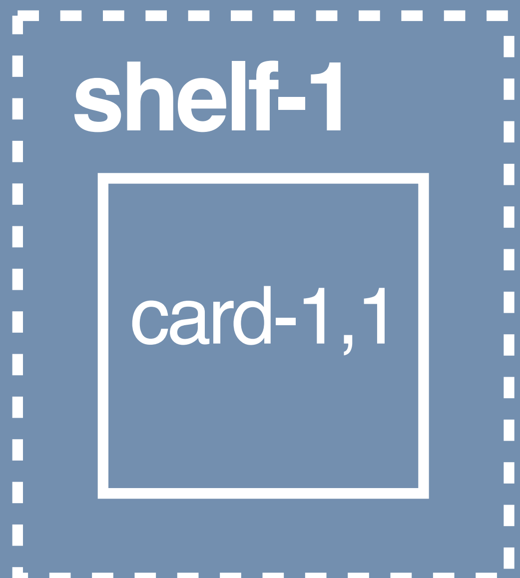
Candidate set



Action select

$$\text{shelf}_1 \sim \pi_{s,r'}(L_1)$$

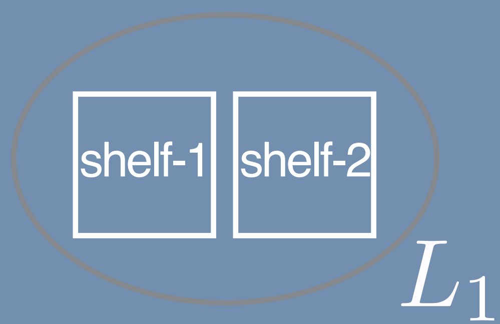
User awareness



Ranking procedure with bandit

Vertical scrolling

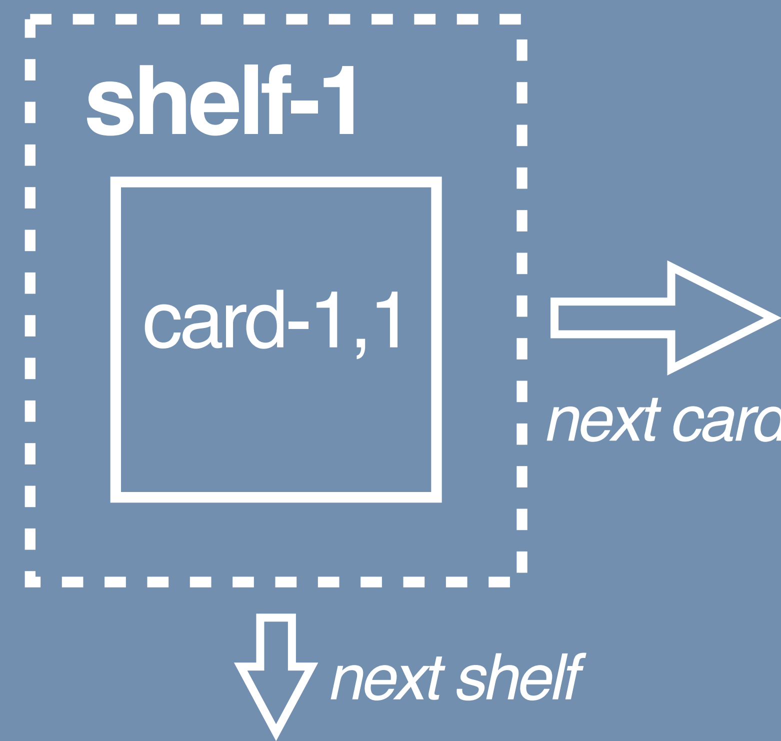
Candidate set



Action select

$$\mathrm{shelf}_1 \sim \pi_{s,r'}(L_1)$$

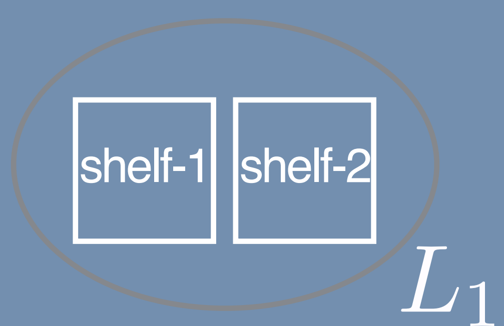
User awareness



Ranking procedure with bandit

Vertical scrolling

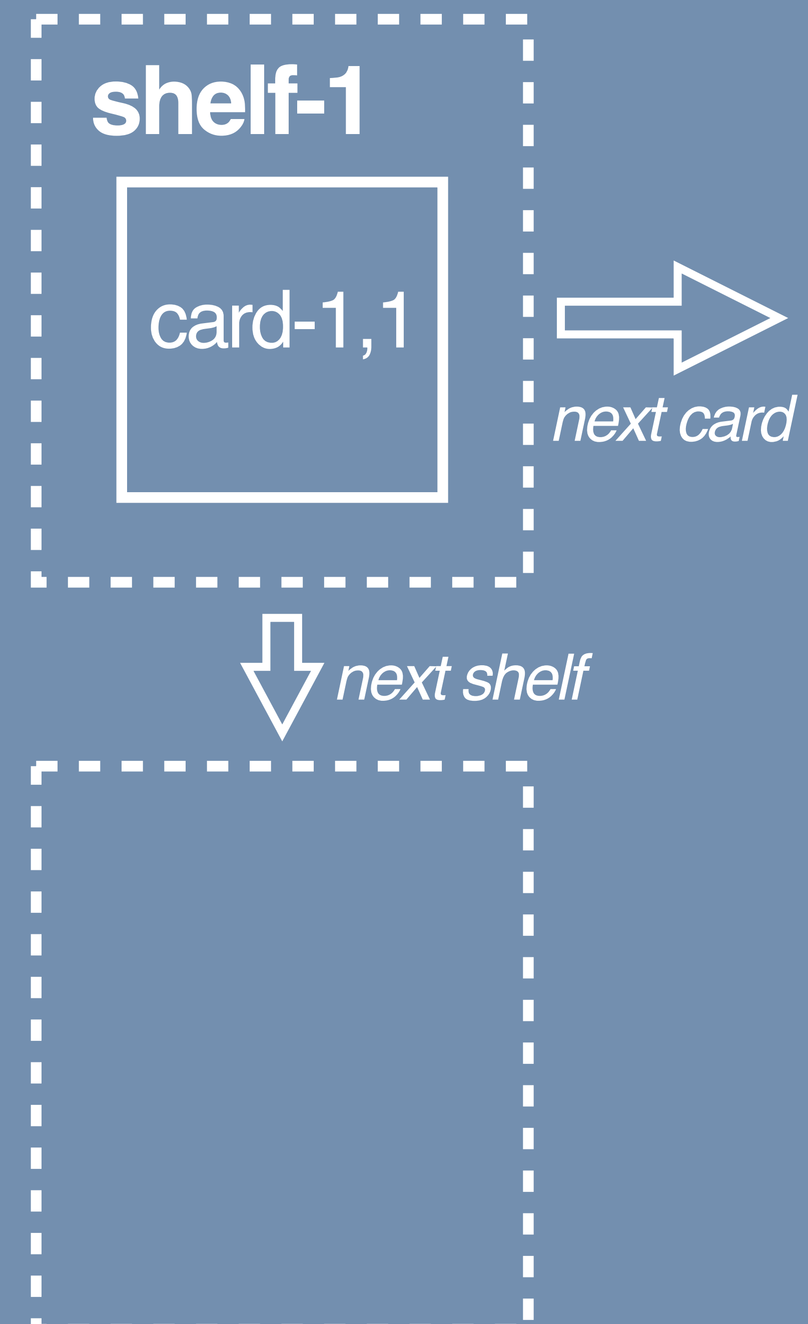
Candidate set



Action select

$$\mathrm{shelf}_1 \sim \pi_{s,r'}(L_1)$$

User awareness



Ranking procedure with bandit

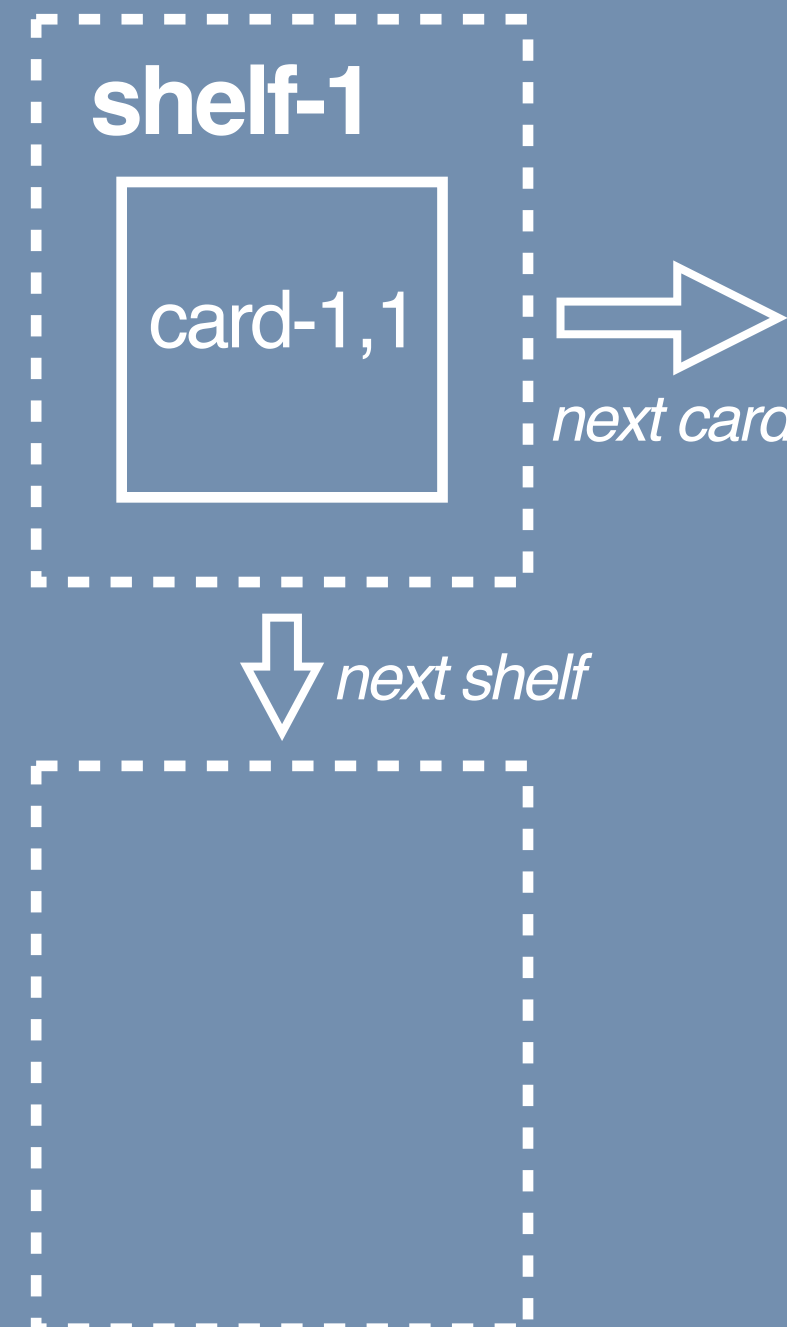
Vertical scrolling

Candidate set

Action select

$$\text{shelf}_1 \sim \pi_{s,r'}(L_1)$$

User awareness



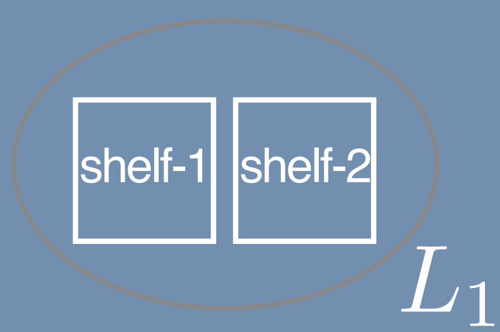
Ranking procedure with bandit

Vertical scrolling

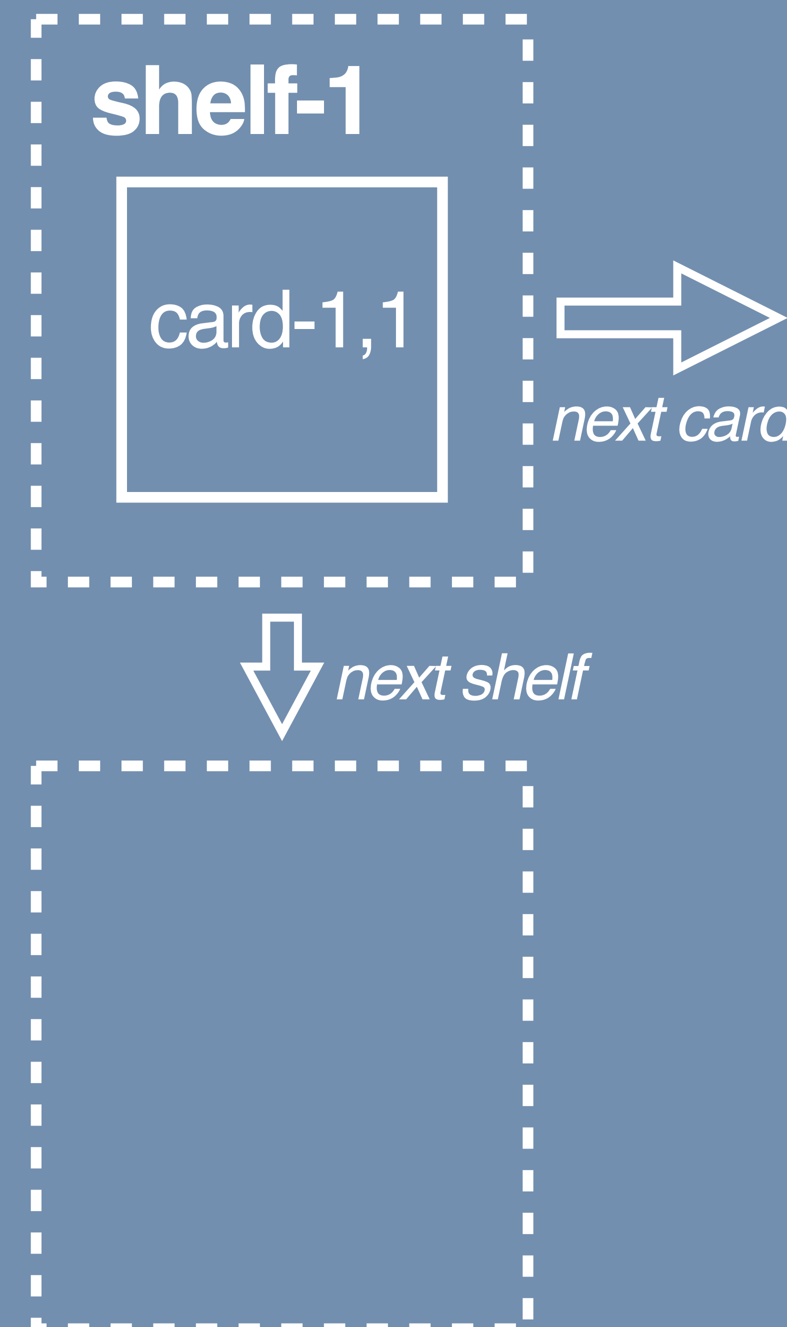
Candidate set

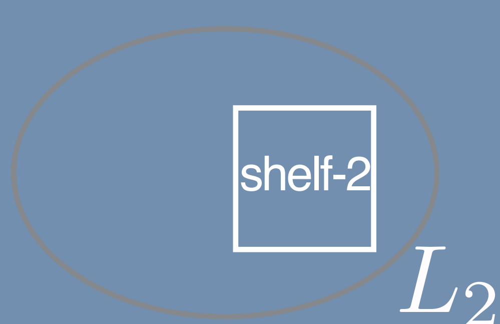
Action select

User awareness



$$\text{shelf}_1 \sim \pi_{s,r'}(L_1)$$

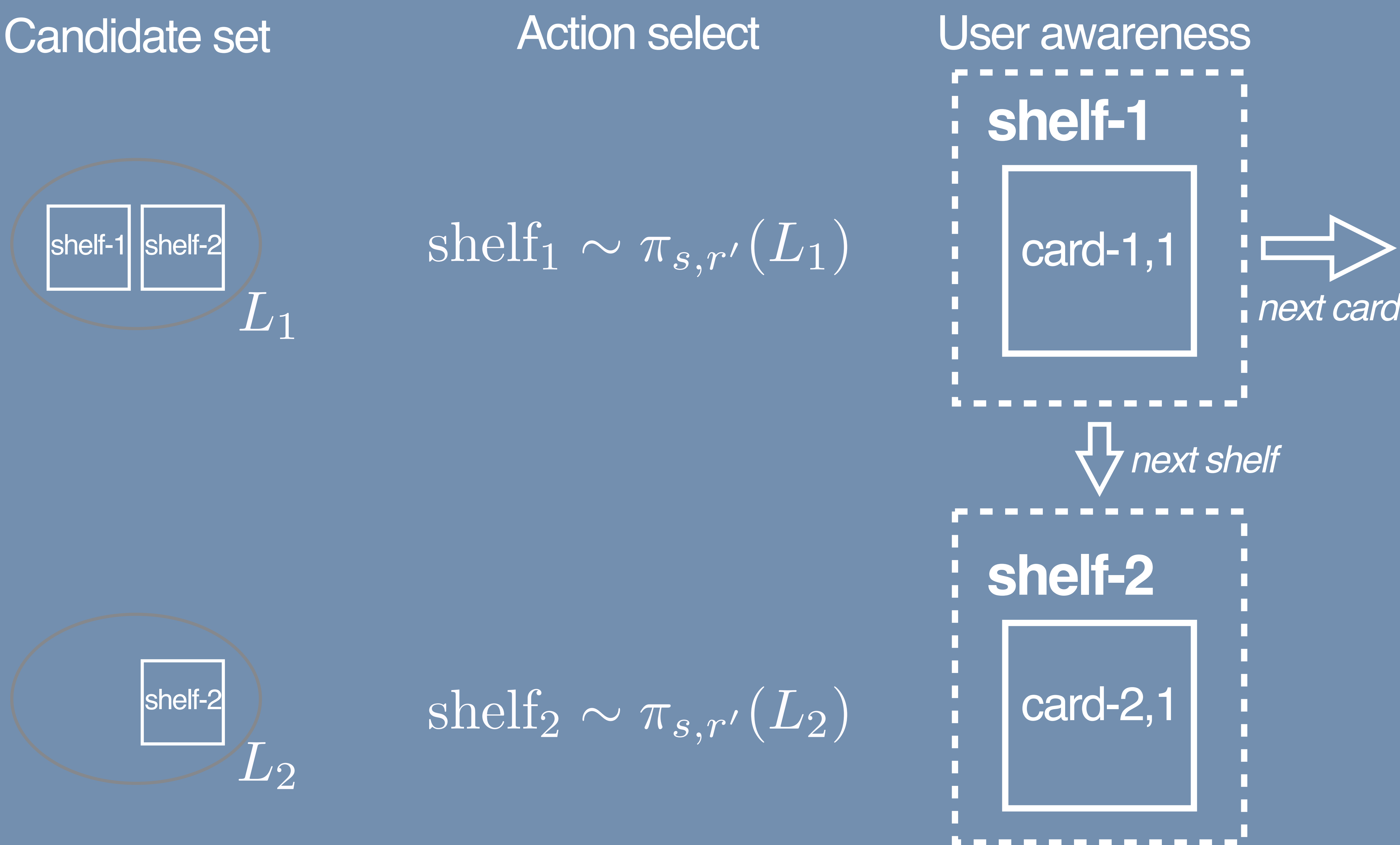




$$\text{shelf}_2 \sim \pi_{s,r'}(L_2)$$

Ranking procedure with bandit

Vertical scrolling



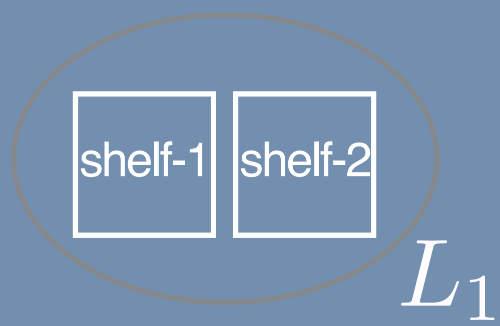
Ranking procedure with bandit

Vertical scrolling

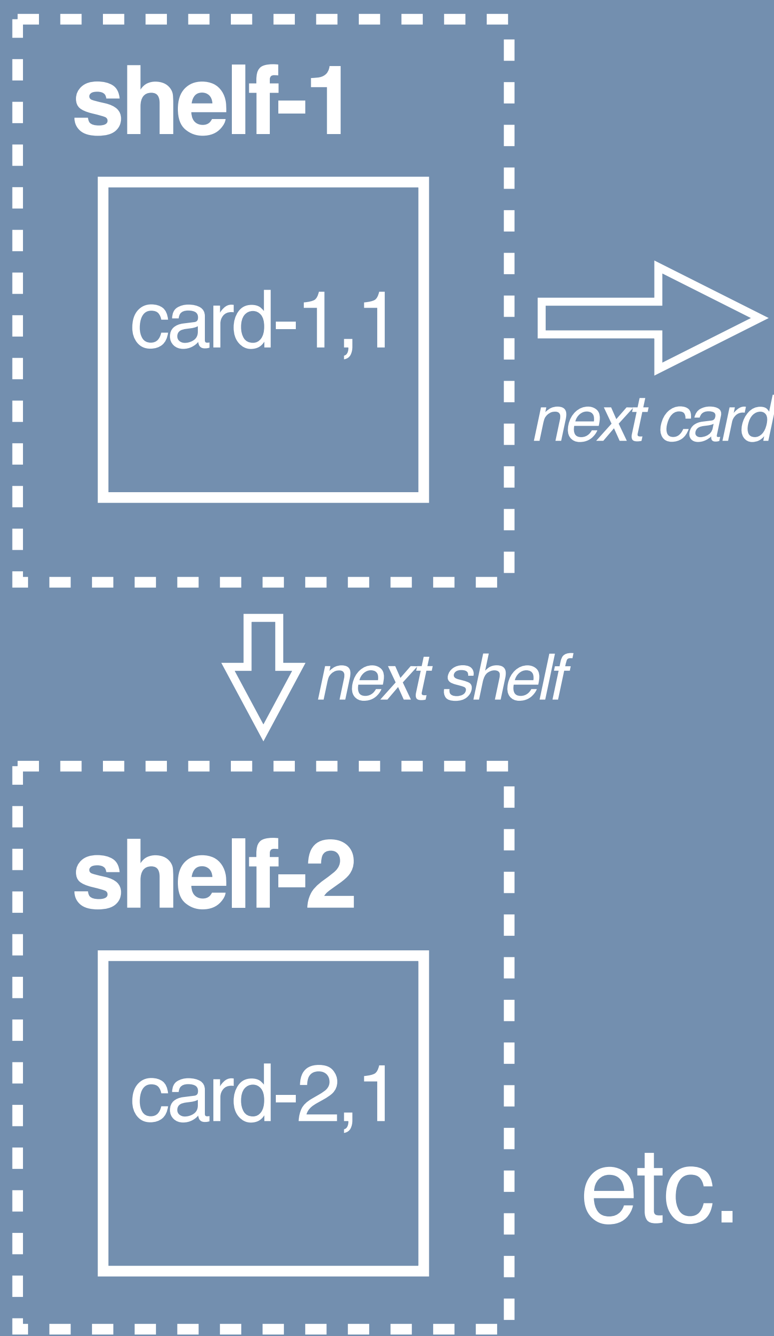
Candidate set

Action select

User awareness



$$\text{shelf}_1 \sim \pi_{s,r'}(L_1)$$

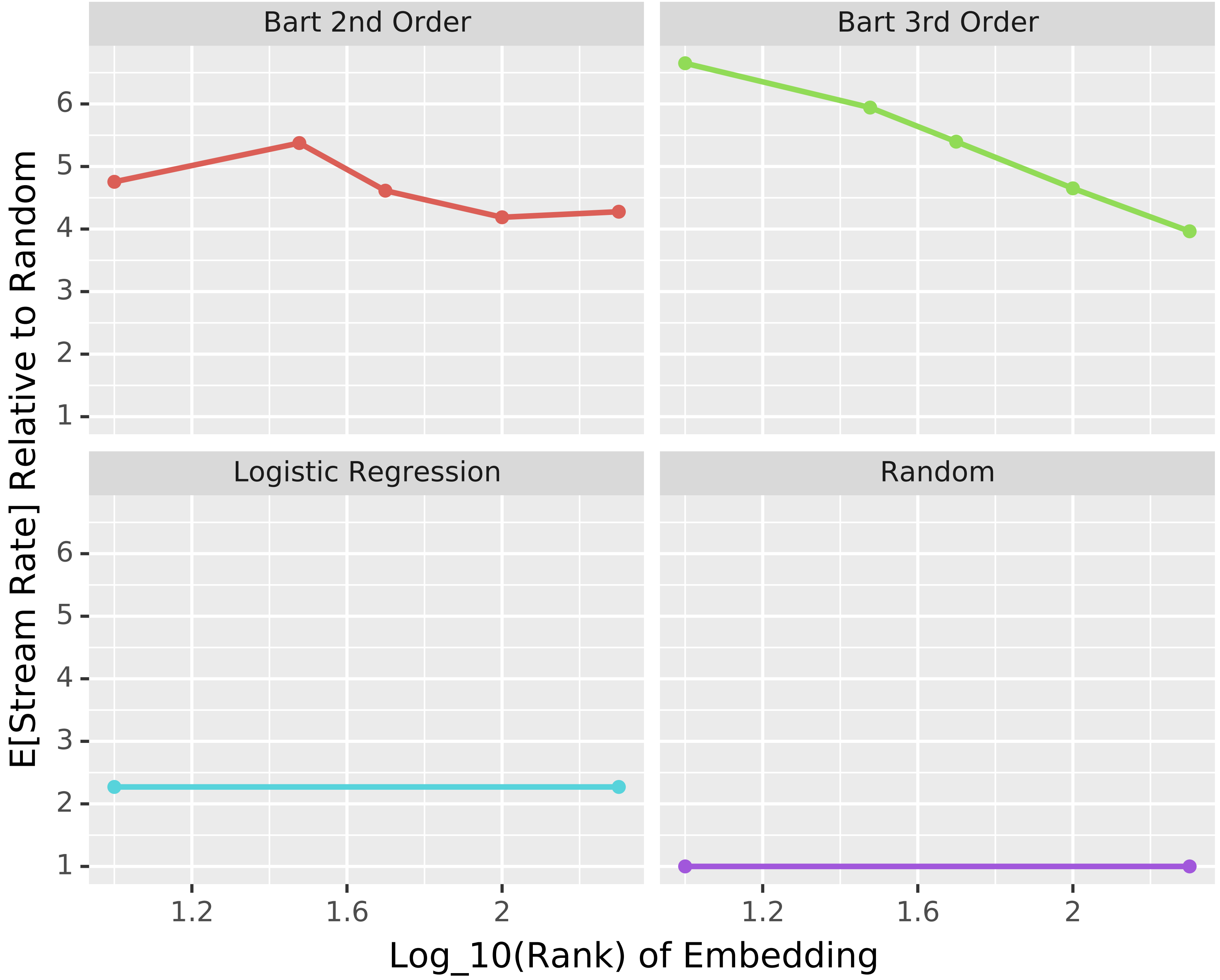


$$\text{shelf}_2 \sim \pi_{s,r'}(L_2)$$

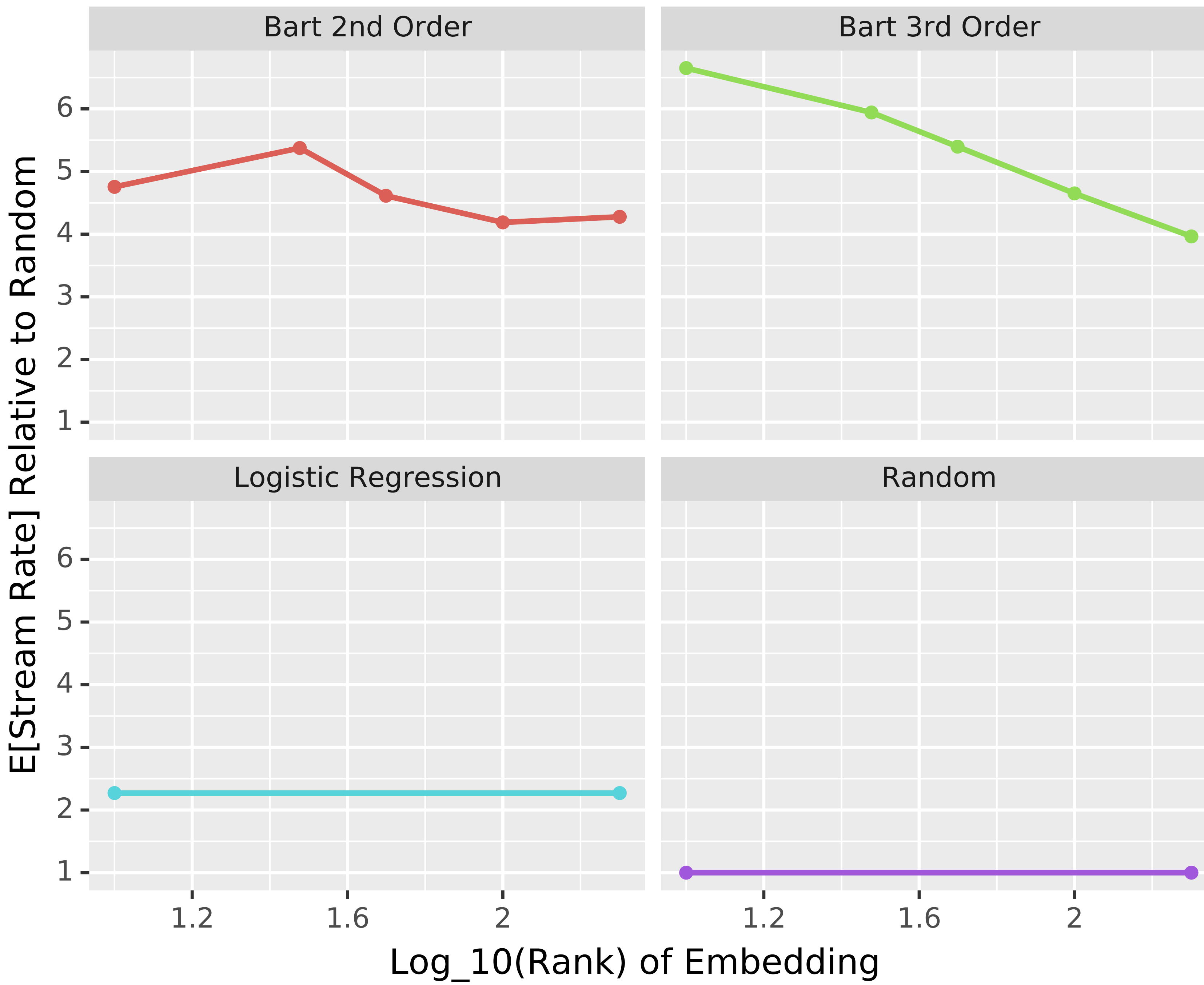
Experimental evaluation

- we collected <u>randomized</u> recommendation data
- offline experiments:
  - counterfactual estimation of A/B test performance using importance sampling reweighting
- online A/B test experiments

Offline experiments



Offline experiments



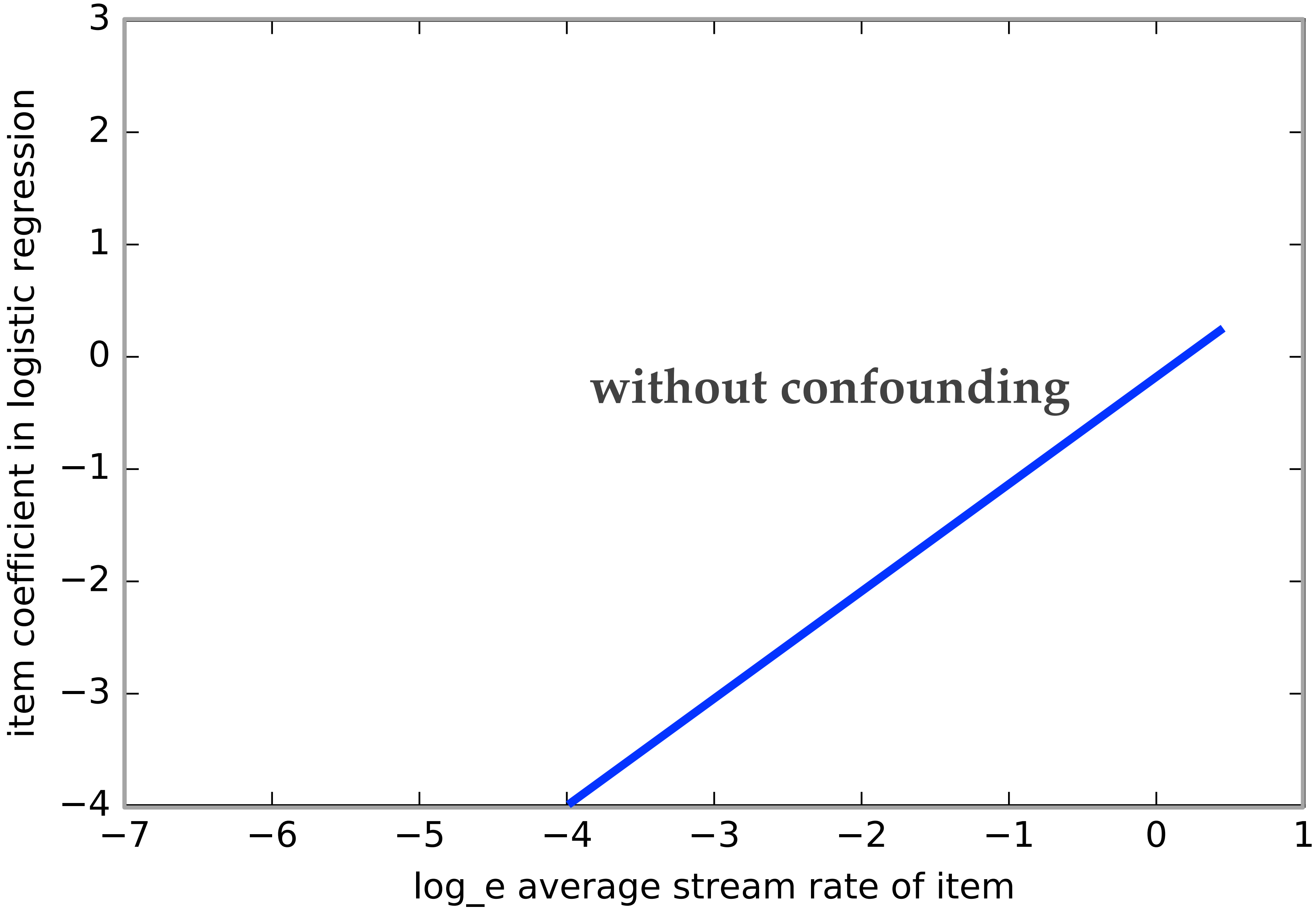
(similar conclusions as NDCG@10 for the metric)

Offline experiments

- how does the empirical stream rate of an item relate to its stream rate controlling for other factors?

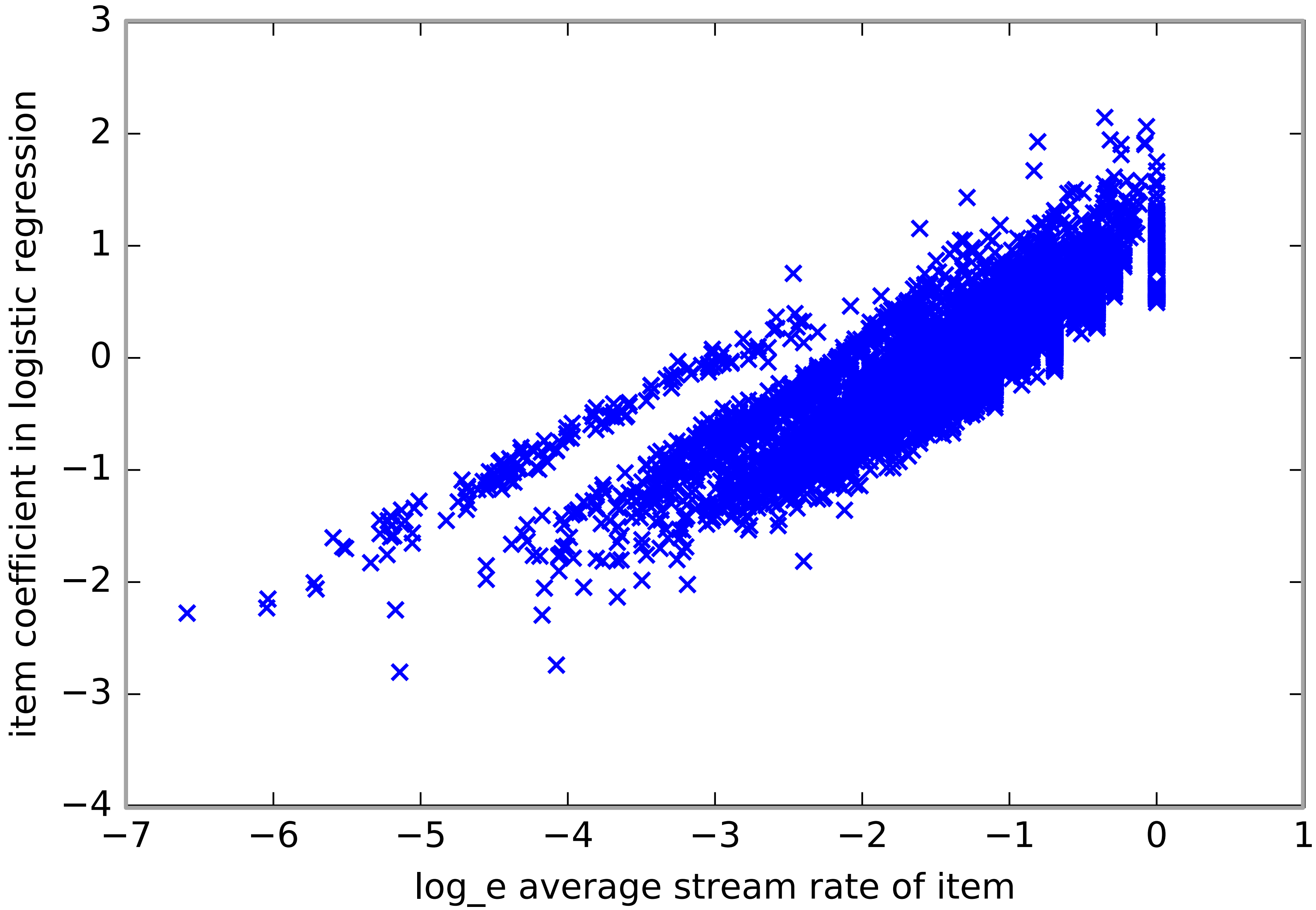
Offline experiments

- how does the empirical stream rate of an item relate to its stream rate controlling for other factors?

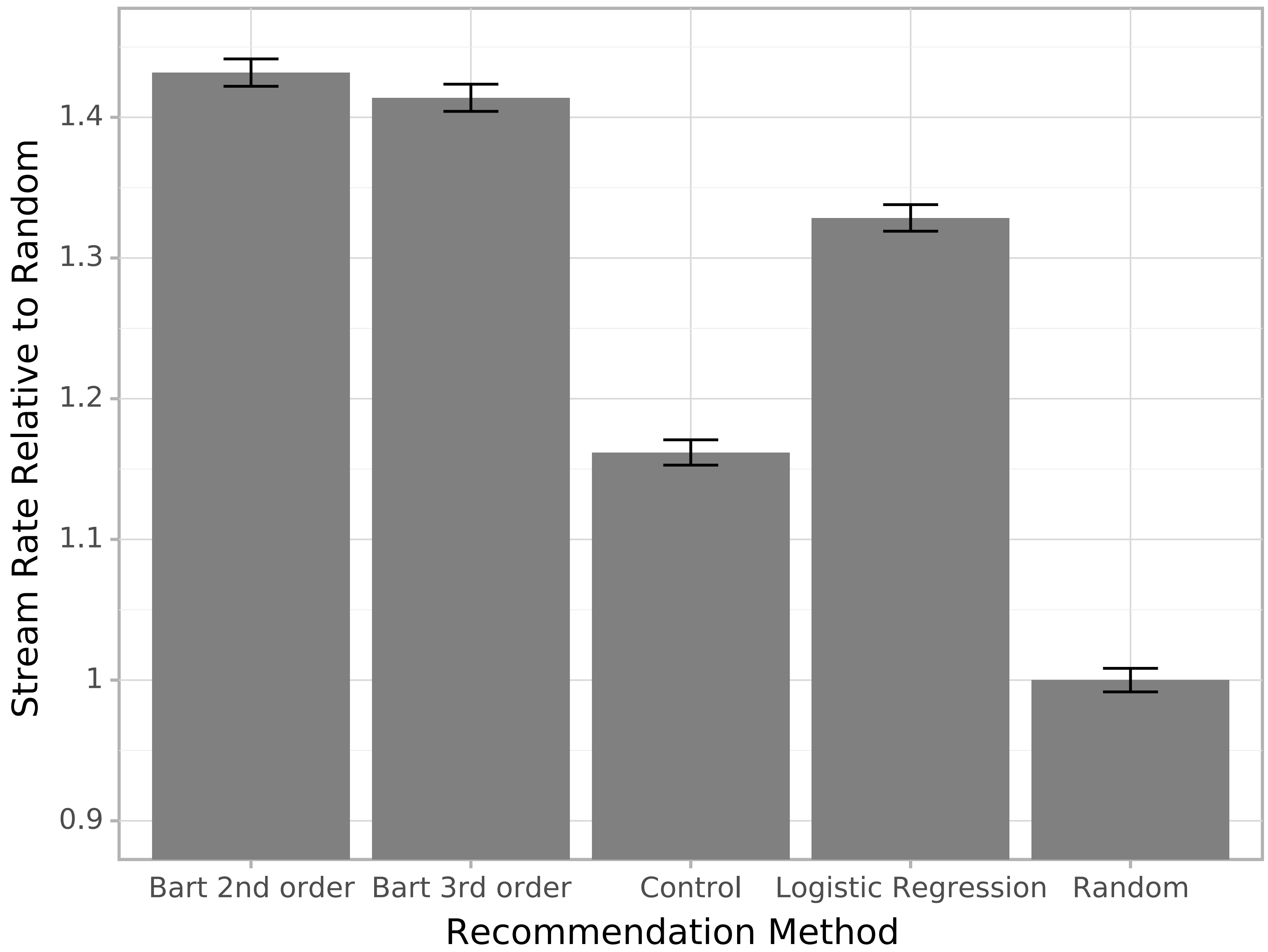


Offline experiments

- how does the empirical stream rate of an item relate to its stream rate controlling for other factors?



Online A/B test



Bart limitations and future work

- user preference model:
  - assumes independence of impression outcomes
  - attempts to estimate absolute reward, competitive pairwise model closer to how humans judge items
  - maximizes our defined reward, does it approximate user satisfaction?
- ranking model not defined to promote diversity
- exploration-exploitation over a candidate set not the full item set

Is bandits a good idea for your problem?

Things to consider:

- confounding: are you training a model using data collected with another model?
    - consider counterfactual evaluation on its own; less need to explore/exploit
- auto-confounding: are you repeatedly training a model using data generated by the same model?
    - consider counterfactual evaluation and explore/exploit

Thank you, any questions?

email: jamesm@spotify.com