Responsible ML for Real-World Search and Recommender Systems

A Multistakeholder Perspective

Ashudeep Singh

Applied Scientist, Pinterest

mail@ashudeepsingh.com

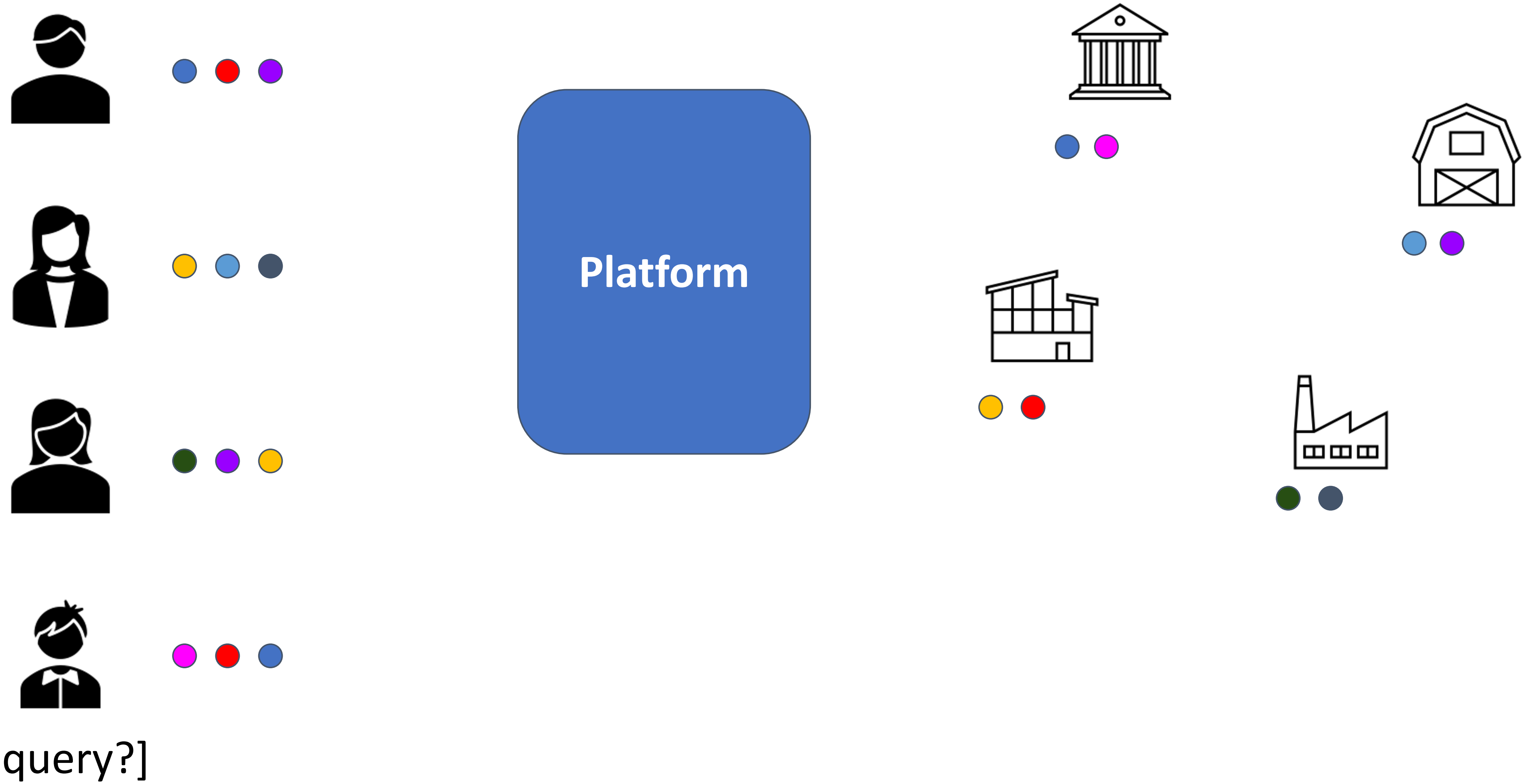
About Me

- Applied Scientist at Pinterest
- Past:
  - PhD in Computer Science from Cornell University
  - Visiting researcher, intern at Google Brain, Microsoft Research, Facebook.
  - Bachelors in Computer Science and Engineering from IIT Kanpur in India.

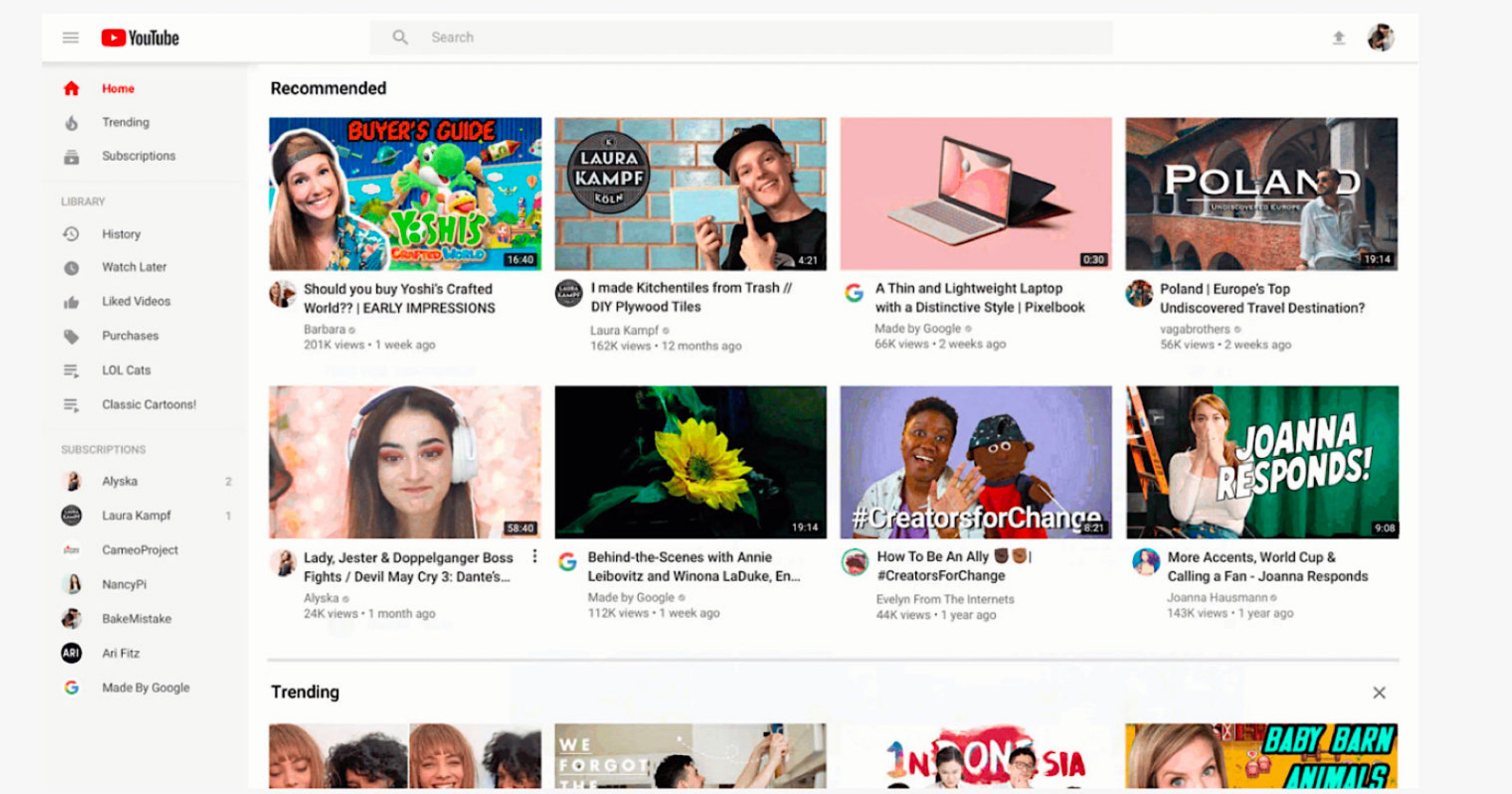
Research Interests:

- Recommender systems and Search
- Machine learning from human interactions
- Fairness and Responsible Machine Learning

Personalized rankings

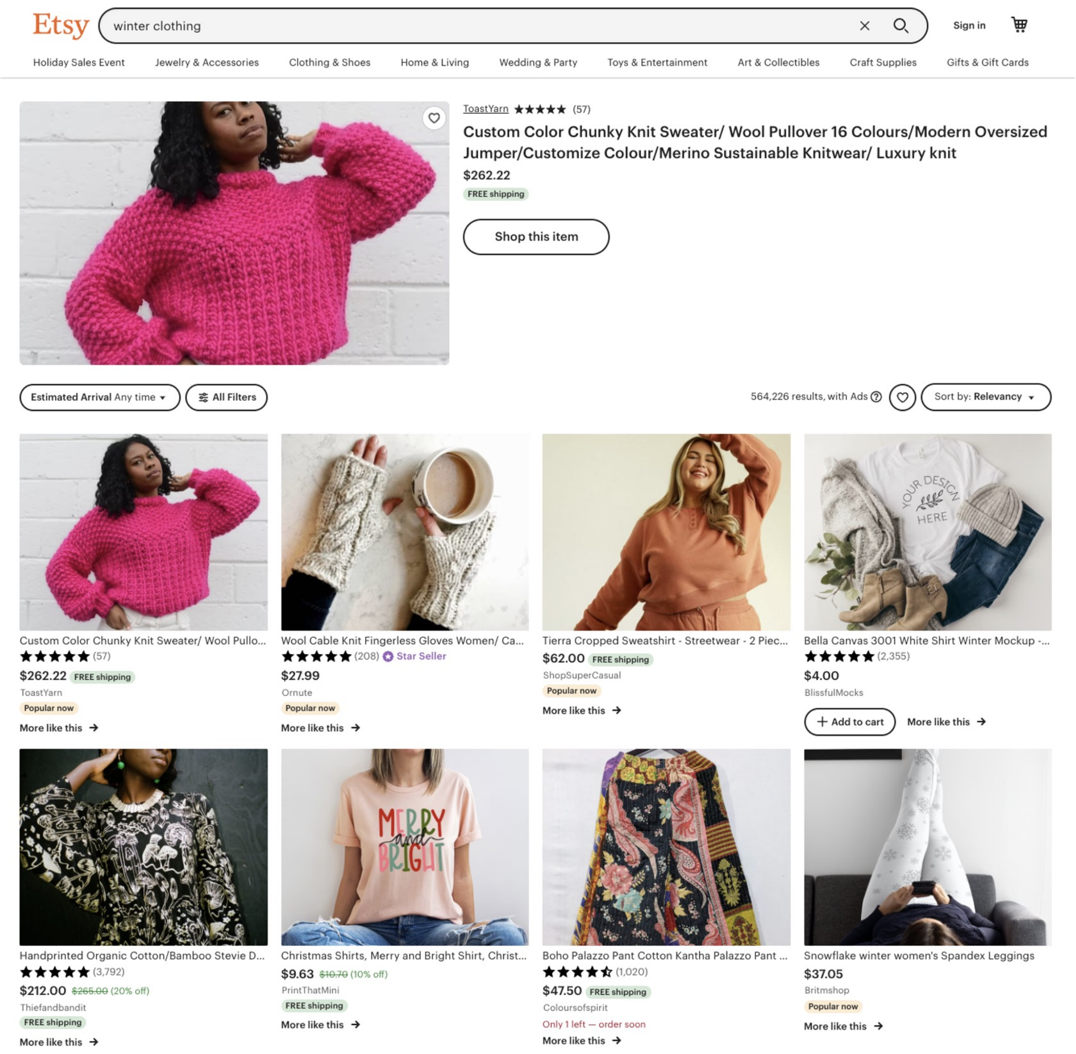


Entertainment

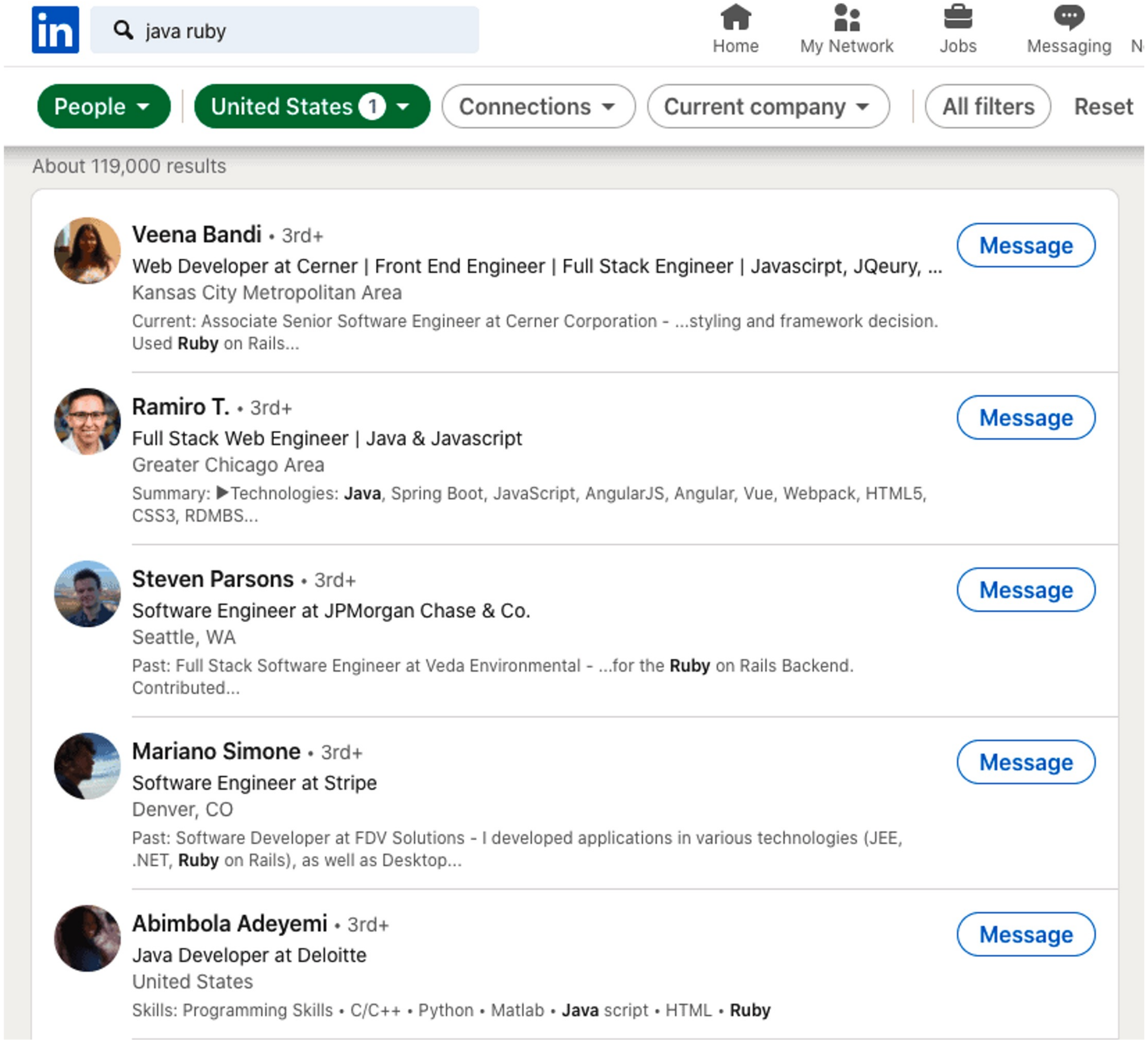


Social media

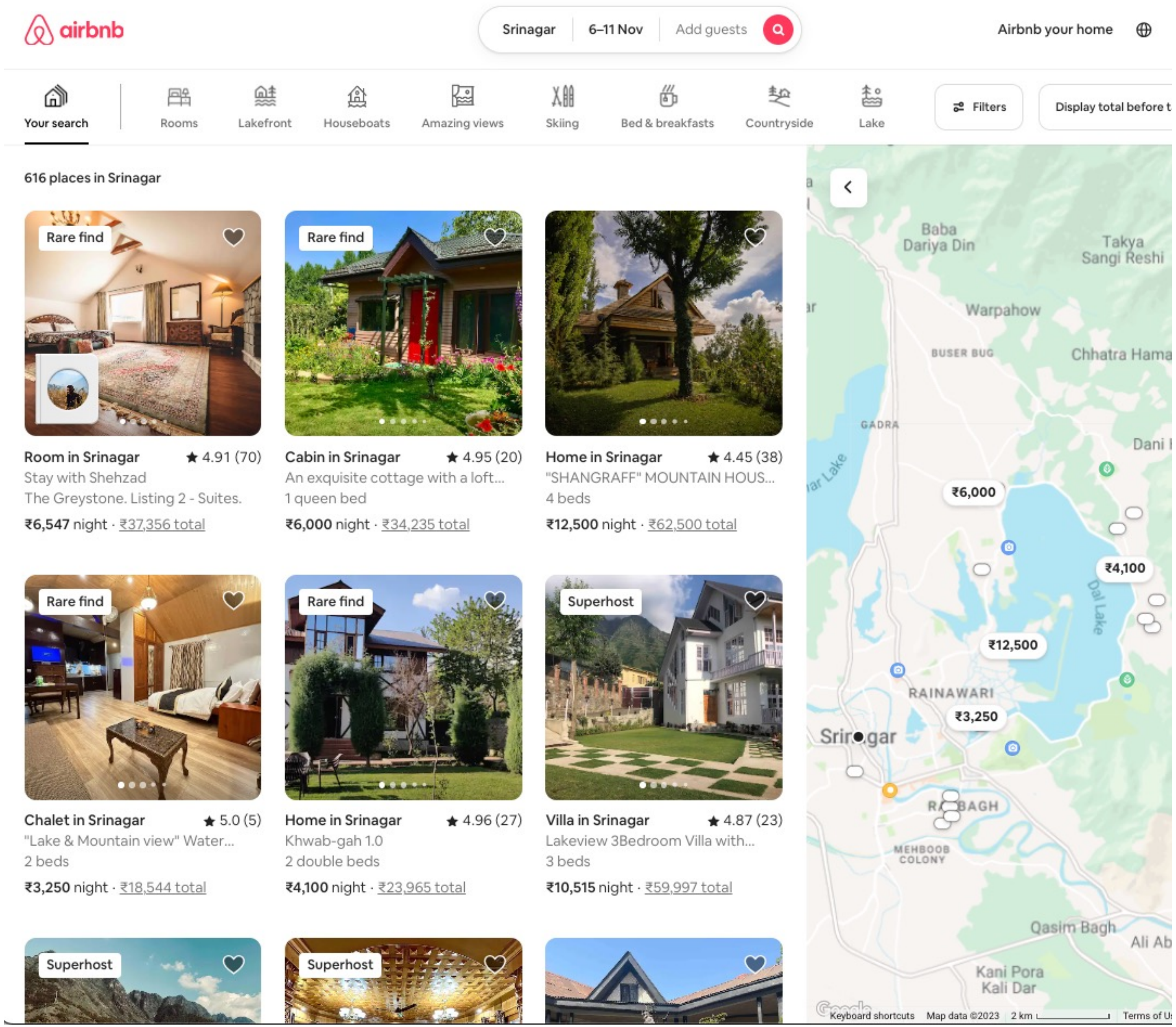
Shopping


Etsy
winter clothing
Sign in
Holiday Sales Event
Jewelry & Accessories
Clothing & Shoes
Home & Living
Wedding & Party
Toys & Entertainment
Art & Collectibles
Craft Supplies
Gifts & Gift Cards
ToastYarn (57)
Custom Color Chunky Knit Sweater/ Wool Pullover 16 Colours/Modern Oversized Jumper/Customize Colour/Merino Sustainable Knitwear/ Luxury knit
$262.22
FREE shipping
Shop this item
Estimated Arrival Any time
All Filters
564,226 results, with Ads
Sort by: Relevancy
Custom Color Chunky Knit Sweater/ Wool Pullo...
(57)
$262.22 FREE shipping
ToastYarn
Popular now
More like this
Wool Cable Knit Fingerless Gloves Women/ Ca...
(208) Star Seller
$27.99
Ornute
Popular now
More like this
Tierra Cropped Sweatshirt - Streetwear - 2 Piec...
$62.00 FREE shipping
ShopSuperCasual
Popular now
More like this
Bella Canvas 3001 White Shirt Winter Mockup -...
(2,355)
$4.00
BlissfulMocks
Add to cart
More like this
Handprinted Organic Cotton/Bamboo Stevie D...
(3,792)
$212.00 $265.00 (20% off)
Thiefandbandit
FREE shipping
More like this
Christmas Shirts, Merry and Bright Shirt, Christ...
$9.63 $10.70 (10% off)
PrintThatMini
FREE shipping
More like this
Boho Palazzo Pant Cotton Kantha Palazzo Pant ...
(1,020)
$47.50 FREE shipping
Coloursofspirit
Only 1 left — order soon
More like this
Snowflake winter women's Spandex Leggings
$37.05
Britmshop
Popular now
More like this

Employment


java ruby
Home
My Network
Jobs
Messaging
People
United States 1
Connections
Current company
All filters
Reset
About 119,000 results
Veena Bandi • 3rd+
Web Developer at Cerner | Front End Engineer | Full Stack Engineer | Javascirpt, JQeury, ...
Kansas City Metropolitan Area
Current: Associate Senior Software Engineer at Cerner Corporation - ...styling and framework decision. Used Ruby on Rails...
Message
Ramiro T. • 3rd+
Full Stack Web Engineer | Java & Javascript
Greater Chicago Area
Summary: ▶Technologies: Java, Spring Boot, JavaScript, AngularJS, Angular, Vue, Webpack, HTML5, CSS3, RDMBS...
Message
Steven Parsons • 3rd+
Software Engineer at JPMorgan Chase & Co.
Seattle, WA
Past: Full Stack Software Engineer at Veda Environmental - ...for the Ruby on Rails Backend. Contributed...
Message
Mariano Simone • 3rd+
Software Engineer at Stripe
Denver, CO
Past: Software Developer at FDV Solutions - I developed applications in various technologies (JEE, .NET, Ruby on Rails), as well as Desktop...
Message
Abimbola Adeyemi • 3rd+
Java Developer at Deloitte
United States
Skills: Programming Skills • C/C++ • Python • Matlab • Java script • HTML • Ruby
Message

Rental Properties



What recommender system do you use the most?

A common approach

Predict relevance *r(i, j)* of item *j to* user *i*

For user *i*, show items in descending order of *r(i, j)*

This has been the subject of debate for decades (e.g., Robertson, 1977)

But in practice, it's the still the dominant approach

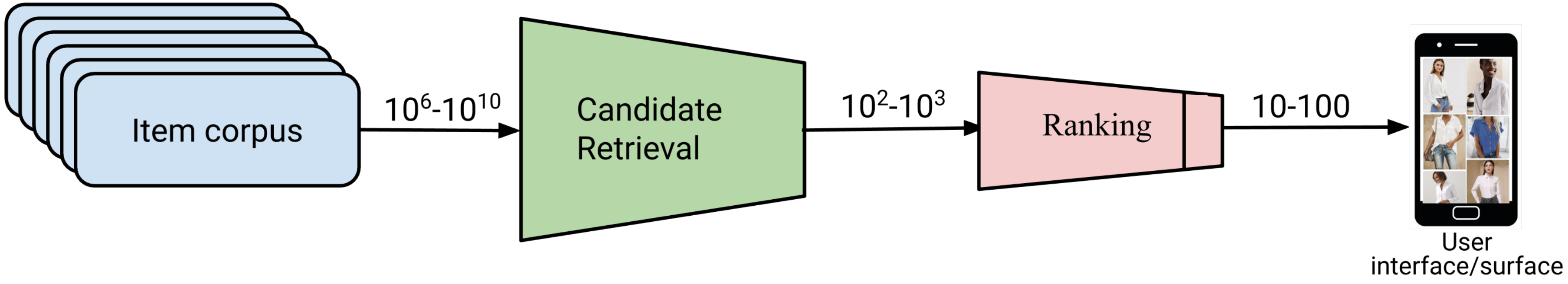
Key questions

1. How do we measure “relevance”?
    a. Is it single-dimensional? Independent across items?
    b. How do we get good data on it?
2. If we had a good measure of relevance, how should we use it?
    a. What constraints are there?
    b. Is descending-order ranking sufficient?
    c. How do we practically make such platforms work?

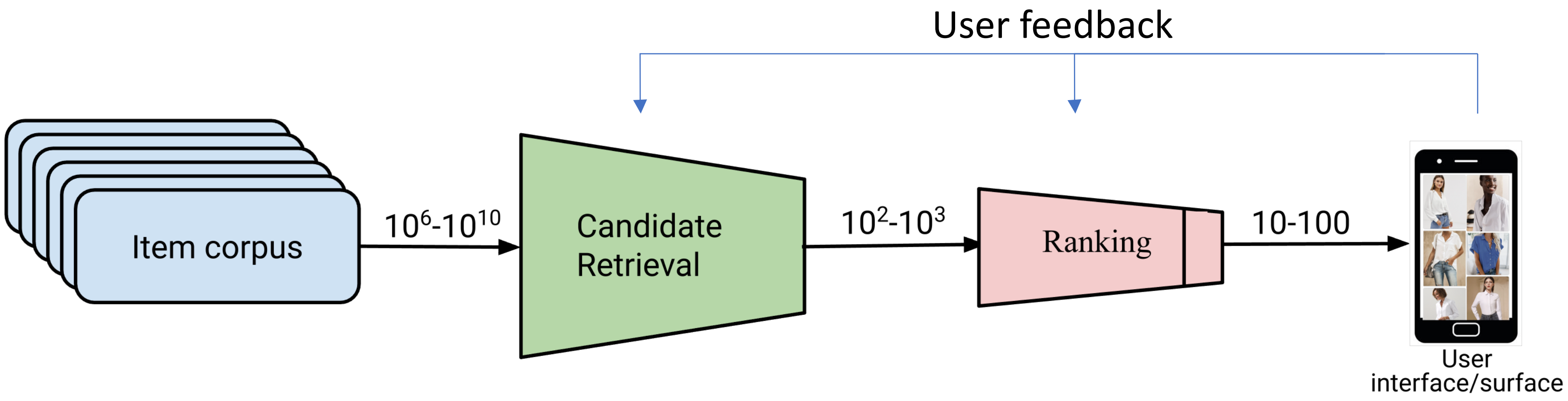
User-Centric Optimization

- Serve the user most relevant items that provide value
- How do we measure relevance and value?
- Proxy: **user engagement**
    - Is this the right proxy?

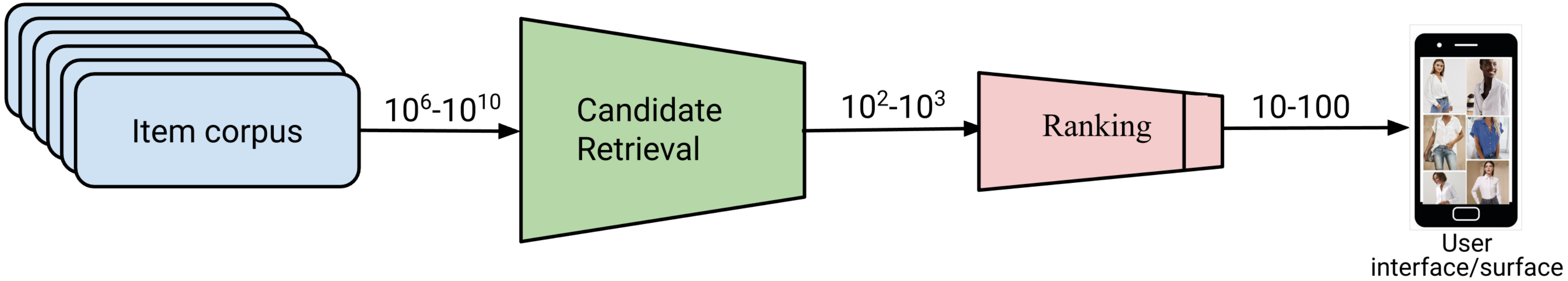
Practical Recommender Systems: Overview



Practical Recommender Systems: Overview



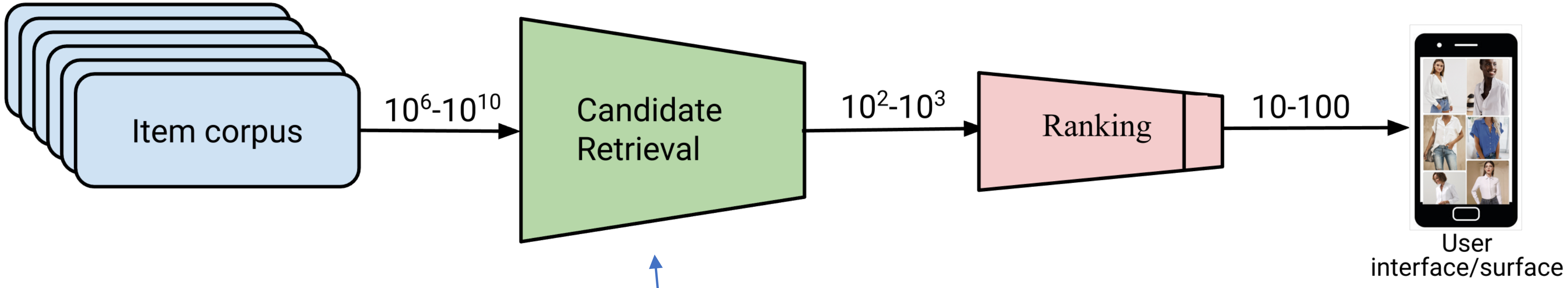
Practical Recommender Systems: Overview



Which of the two steps requires:

(a) Lower latency (higher throughput)?
(b) Higher precision?
(c) Higher recall?

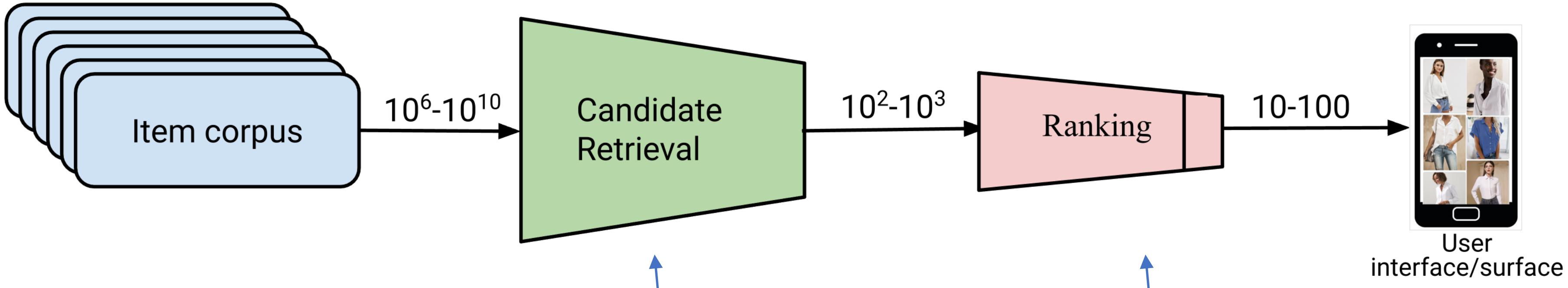
Practical Recommender Systems: Overview



Low latency, High recall.

e.g., Nearest Neighbor search on embeddings, Collaborative Filtering.

Practical Recommender Systems: Overview



Low latency, High recall.

e.g., Nearest Neighbor search on embeddings, Collaborative Filtering.

High precision, can afford more computation per item.

e.g. Learning-to-rank.

System level view

- Example algorithms at the two stages:
  - Retrieval: e.g., Representation learning → Nearest neighbor search on vector embeddings
  - Ranking: e.g., Learning-to-rank
- Both learnt from user feedback

Stage 1: Candidate Retrieval

- Collaborative filtering

Collaborative Filtering

- Collaborative filtering uses similarities between users and items simultaneously to provide recommendations, i.e.,
    - recommend an item to user A based on the interests of a “similar” user B.
- Common method: Matrix Factorization of the user-item rating matrix.

Given a dataset of user item ratings: $Y_{u,i}$,

Find a user and item embedding matrix (U and V), so that the $U^T V$ is as close to the ratings matrix.

Collaborative Filtering

- Collaborative filtering uses similarities between users and items simultaneously to provide recommendations, i.e.,
  - recommend an item to user A based on the interests of a "similar" user B.
- Common method: Matrix Factorization of the user-item rating matrix.

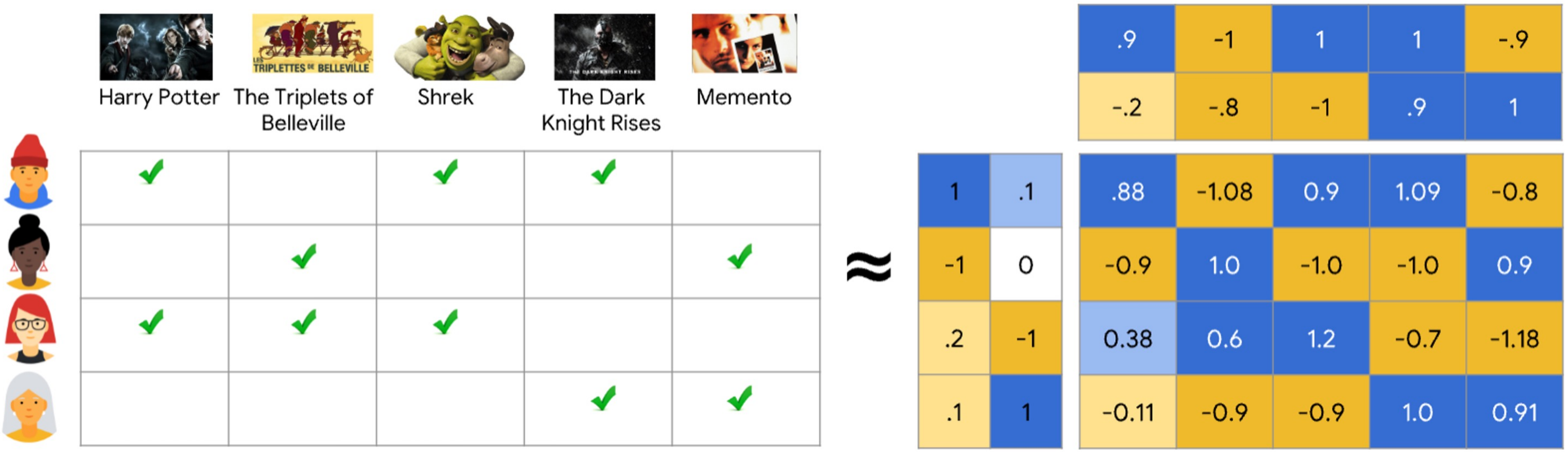
Given a dataset of user item ratings: $Y_{u,i}$,

Find a user and item embedding matrix (U and V), so that the $U^T V$ is as close to the ratings matrix.

$$V = \begin{bmatrix} .9 & -1 & 1 & 1 & -.9 \\ -.2 & -.8 & -1 & .9 & 1 \end{bmatrix}$$

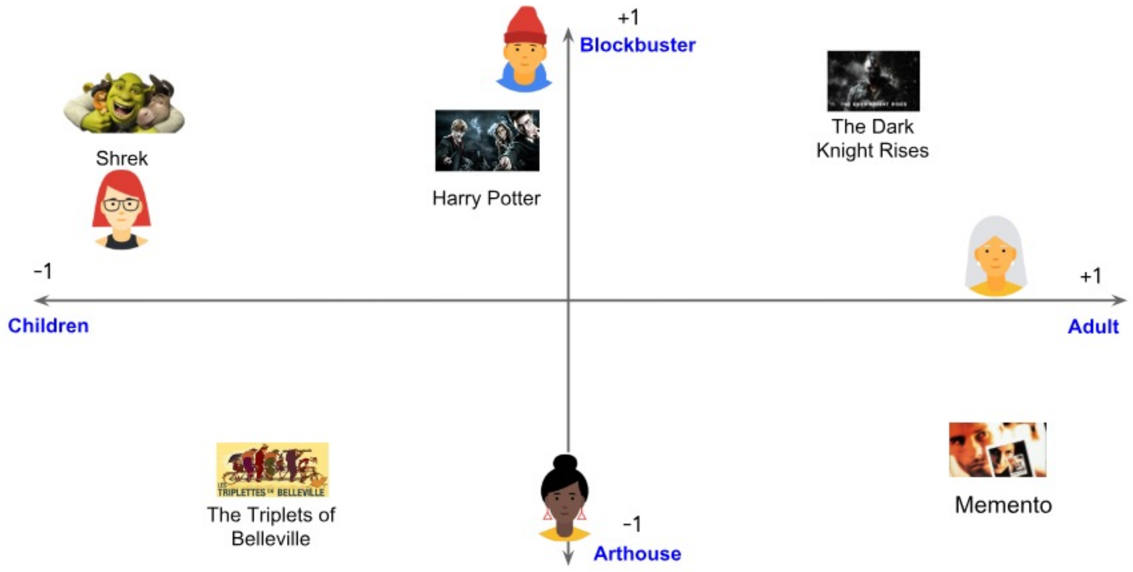
$$\approx \begin{bmatrix} 1 & .1 \\ -1 & 0 \\ .2 & -1 \\ .1 & 1 \end{bmatrix} \quad \begin{bmatrix} .88 & -1.08 & 0.9 & 1.09 & -0.8 \\ -0.9 & 1.0 & -1.0 & -1.0 & 0.9 \\ 0.38 & 0.6 & 1.2 & -0.7 & -1.18 \\ -0.11 & -0.9 & -0.9 & 1.0 & 0.91 \end{bmatrix}$$

Source: https://developers.google.com/machine-learning/recommendation/collaborative/basics



Collaborative Filtering

An illustration



Stage 1: Candidate Retrieval

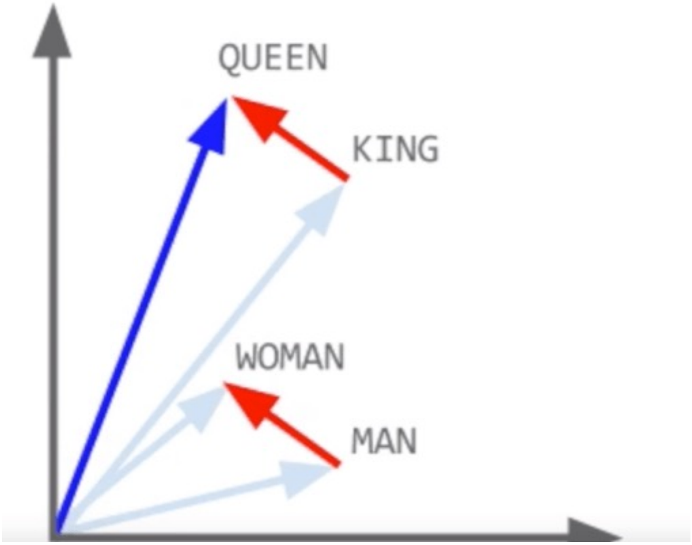
- Collaborative filtering
  - Output of the training process: a vector representation of all users and all items.
  - Serving time: Find the top item vectors that match the user vector.

Stage 1: Candidate Retrieval

- Collaborative filtering
  - Output of the training process: a vector representation of all users and all items.
  - Serving time: Find the top item vectors that match the user vector.

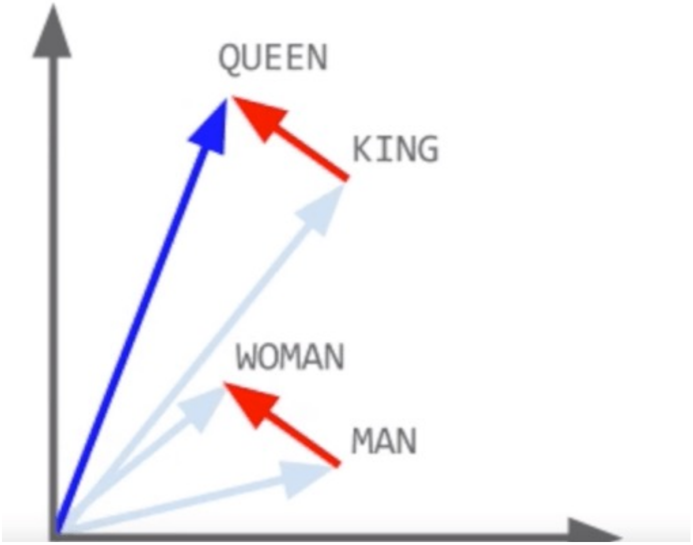
- More recently: several other techniques use neural networks, latent models, etc. to learn this vector representation to make retrieval fast and easy

Power of Representation learning

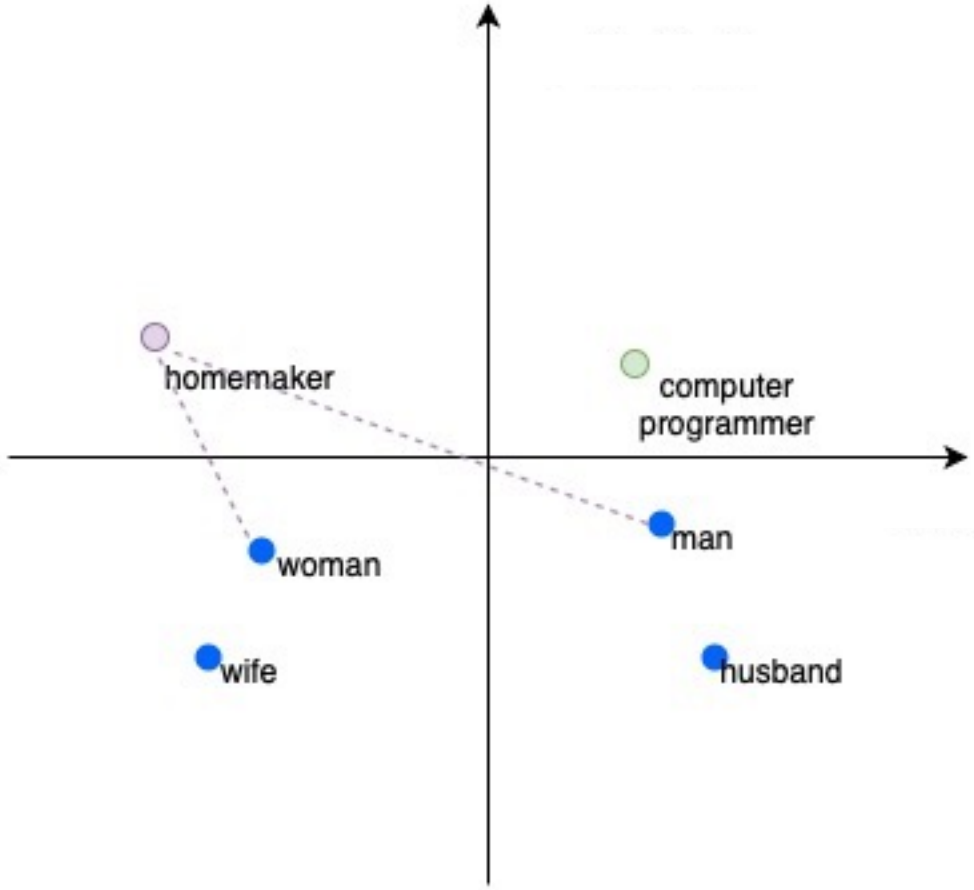


$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

Bias in ~~Power of~~ Representation learning



$$\overrightarrow{\mathrm{man}} - \overrightarrow{\mathrm{woman}} \approx \overrightarrow{\mathrm{king}} - \overrightarrow{\mathrm{queen}}$$



$$\overrightarrow{\mathrm{man}} - \overrightarrow{\mathrm{woman}} \approx \overrightarrow{\mathrm{computer\ programmer}} - \overrightarrow{\mathrm{homemaker}}.$$

Google image search

until a few years ago….

BBC Sign in | Home | News | Sport | Reel | Worklife | Travel

NEWS

Home | War in Ukraine | Climate | Video | World | Asia | UK | Business | Tech | Science

Newsbeat

Google Image search for CEO has Barbie as first female result

16 April 2015

We had to scroll down the page to before this picture of Barbie appears

Google image search

until a few years ago….

BBC Sign in | Home | News | Sport | Reel | Worklife | Travel

NEWS

Home | War in Ukraine | Climate | Video | World | Asia | UK | Business | Tech | Science

Newsbeat

Google Image search for CEO has Barbie as first female result

16 April 2015

We had to scroll down the page to before this picture of Barbie appears

Discussion point: What are the possible causes?

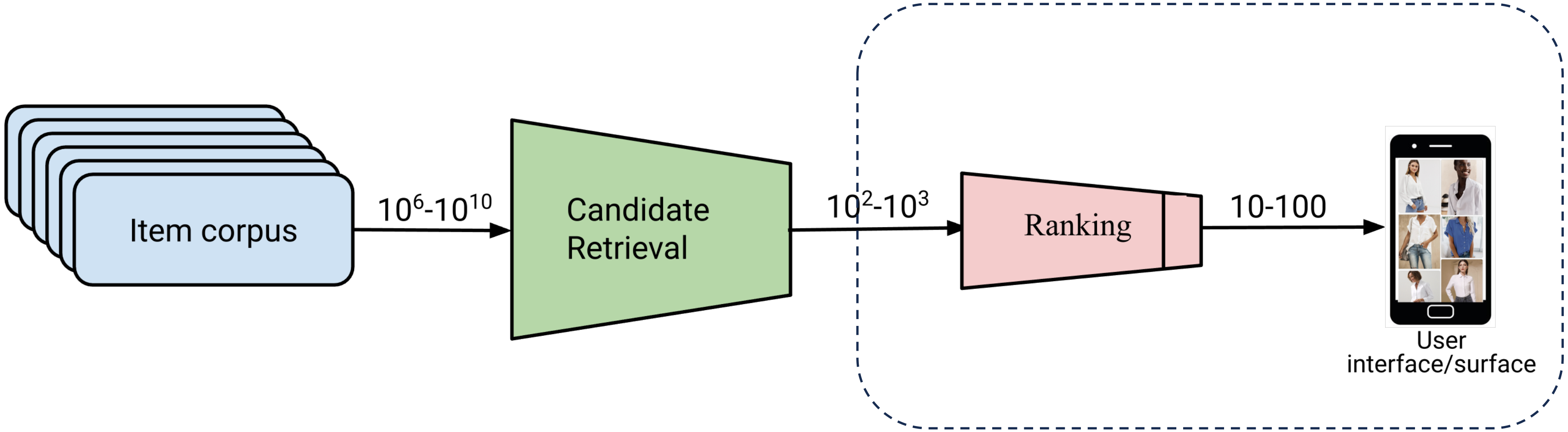
Percentage of women in the top 100 Google image search results for telemarketers: 64%
Percentage of U.S. telemarketers who are women: 50%

Google image search results for "construction worker"

Google image search results for "female construction worker"

Source: https://www.washington.edu/news/2022/02/16/googles-ceo-image-search-gender-bias-hasnt-really-been-fixed/

Practical Recommender Systems: Overview



For the next section, we will focus solely on ranking problems…

Probability Ranking Principle (PRP)

Robertson (1977):

- "if a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request,
- where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose,
- the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data."

THE PROBABILITY RANKING PRINCIPLE IN IR

S. E. ROBERTSON
School of Library, Archive, and Information Studies, University College London

The principle that, for optimal retrieval, documents should be ranked in order of the probability of relevance or usefulness has been brought into question by Cooper. It is shown that the principle can be justified under certain assumptions, but that in cases where these assumptions do not hold, the principle is not valid. The major problem appears to lie in the way the principle considers each document independently of the rest. The nature of the information on the basis of which the system decides whether or not to retrieve the documents determines whether the document-by-document approach is valid.

PRP in a two-sided system

- In two-sided markets, PRP might be inadequate since it does not explicitly consider the item-side utility.

- Examples:
  - Job Candidate Ranking
    - Amplifies existing societal biases.

Job Candidate Ranking Example

Position		x	P(interview)
1		A_1	50.99%
2		A_2	50.98%
3		A_3	50.97%
…	…		…
101		B_1	49.99%
102		B_2	49.98%
103		B_3	49.97%
…	…		…

High Exposure

Position Bias

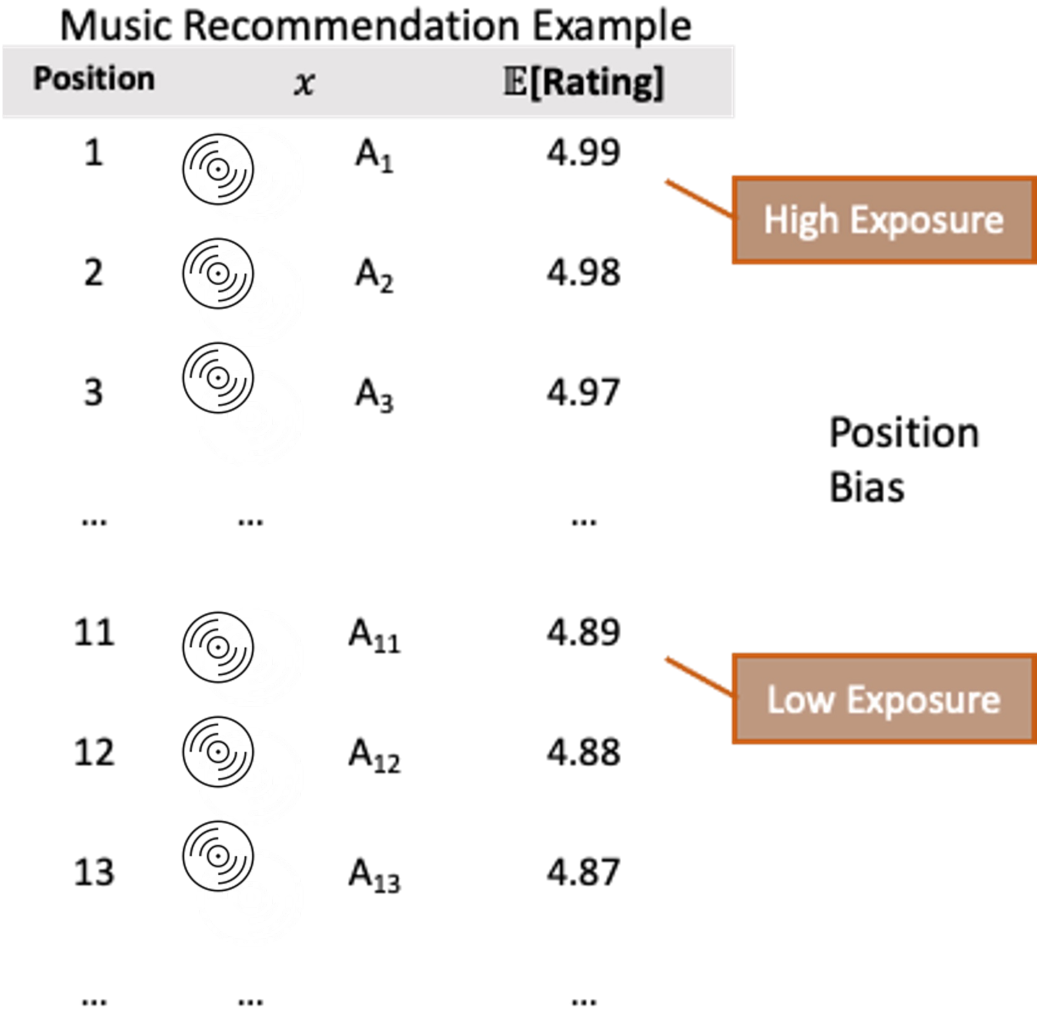
Low Exposure

[Singh & Joachims 2018, Biega et al. 2018]

PRP in a two-sided system

- In two-sided markets, PRP might be inadequate since it does not explicitly consider the item-side utility.

- Examples:
  - Job Candidate Ranking
    - Amplifies existing societal biases.
  - Music Recommendation
    - Winner-takes-all!



[Singh & Joachims 2018, Biega et al. 2018]

PRP in a two-sided system

- In two-sided markets, PRP might be inadequate since it does not explicitly consider the item-side utility.

- Examples:
  - Job Candidate Ranking
    - Amplifies existing societal biases.
  - Music Recommendation
    - Winner-takes-all!
  - News Ranking
    - Polarization of the platform.

News Ranking Example

Position	x	$P(read)$
1	R R_1	50.99%
2	R R_2	50.98%
3	R R_3	50.97%
…	…	…
101	T T_1	49.99%
102	T T_2	49.98%
103	T T_3	49.97%
…	…	…

High Exposure

Position Bias

Low Exposure

[Singh & Joachims 2018, Biega et al. 2018]

In online platforms,

Exposure → Opportunity

Hence,

Fairness → Fair Allocation of Exposure

Position-based Model of Exposure

Exposure e_k is the probability a user observes the item at position k.

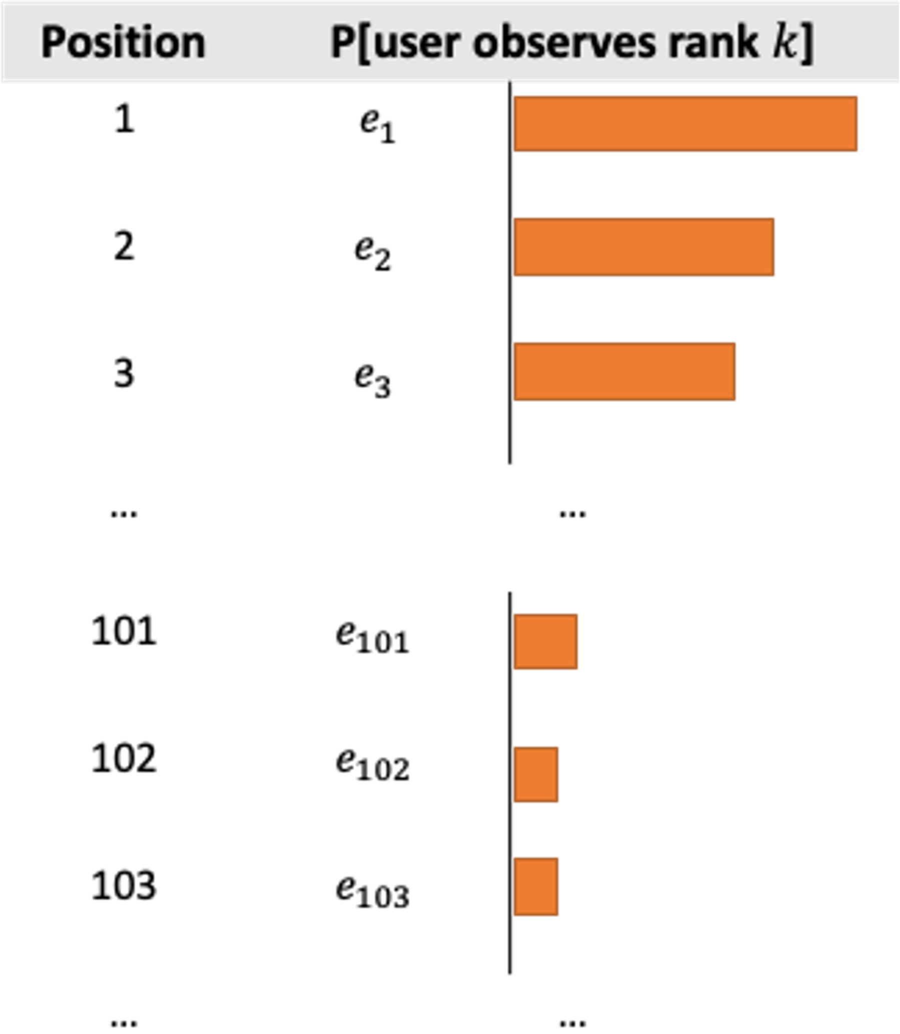
Exposure of a group of items (e.g., seller, artist, etc.)

$$Exp(G|y) = \sum_{y(k) \in G} e_k$$

Other user-click models: Cascading click model (CCM), etc. *[Chuklin et al. 2015]*

How to estimate?

- Eye tracking *[Joachims et al. 2007]*
- Intervention studies *[Joachims et al. 2017]*
- Intervention harvesting *[Agarwal et al. 2019]*



Fairness of Exposure

Goal: Enable the explicit statement of how exposure is allocated relative to the value or merit of the items in the group.

For example: Exposure for each individual/group should be proportional to the relevance of the group.

[Singh & Joachims 2018, Biega et al. 2018]

Equal Expected Exposure

For tasks with graded relevance (e.g., movie ratings — 1 to 5, binary relevance — 0, 1), define equal expected exposure as:

No item has less or more expected exposure as compared to other items in the same relevance grade.

[Diaz et al 2019]

Disparate Exposure & Impact

Disparate exposure: Allocate exposure proportional to relevance per group

$$\text{Exposure} \propto \text{Relevance}$$

$$\frac{Exp(G_0|x)}{Exp(G_1|x)} = \frac{Rel(G_0|x)}{Rel(G_1|x)}$$

Disparate impact: Allocate expected clickthrough rate proportional to relevance per group

$$\frac{\sum_{d \in G_0} Exp(d|x)\, Rel(d|x)}{\sum_{d \in G_1} Exp(d|x)\, Rel(d|x)} = \frac{Rel(G_0|x)}{Rel(G_1|x)}$$

[Singh & Joachims, KDD 2018]

Fairness of Exposure

Objective: Given relevance scores, find a ranking that optimizes user utility while satisfying fairness of exposure constraints, e.g., exposure proportional to average relevance.

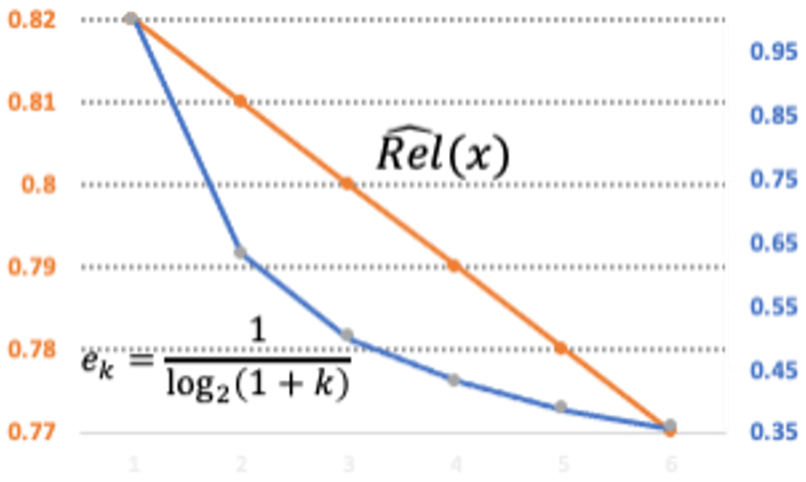
Items	$\hat{h}(x)$
A_1	0.82
A_2	0.81
A_3	0.80
B_1	0.79
B_2	0.78
B_3	0.77

$\times$

Exposure@k
e_1
e_2
e_3
e_4
e_5
e_6

Problem:

- Exposure drops off at a different rate than relevance.
- Rankings are discrete combinatorial objects.
  - Exponential solution space!



[Singh & Joachims, KDD 2018]

Key Idea 1: Stochastic Ranking Policies

- Ranking Policy

  $\pi(y|x)$ is the conditional distribution over rankings of items under query x.

y_1	y_2	y_3	y_4
A_1	A_1	A_1	B_1
A_2	B_1	A_2	A_1
A_3	A_2	B_1	B_2
B_1	B_2	A_3	A_2
B_2	A_3	B_2	B_3
B_3	B_3	B_3	A_3
0.40	0.40	0.16	0.04

Define Utility

$$U(\pi|x) = \sum_y U(y|x) \cdot \pi(y|x)$$

Define Exposure

$$Exp(d|\pi) = \sum_k e_k \cdot P(rank(d) = k \mid \pi)$$

[Singh & Joachims, KDD 2018]

Key Idea 2: Doubly Stochastic Matrices

Represent a Stochastic Ranking π as a Marginal Rank Distribution $\mathbb{P}$.

$$
\begin{array}{cc}
 & \text{Rank} \\
\text{Item} & \begin{pmatrix}
\cdot & \cdot & \cdot \\
 & \cdot & \cdot \\
\cdot & \mathbb{P}_{i,k} & \\
 & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
 & \cdot & \cdot
\end{pmatrix}
\end{array}
$$

$\mathbb{P}_{i,k}$ = Probability of item i at position k.

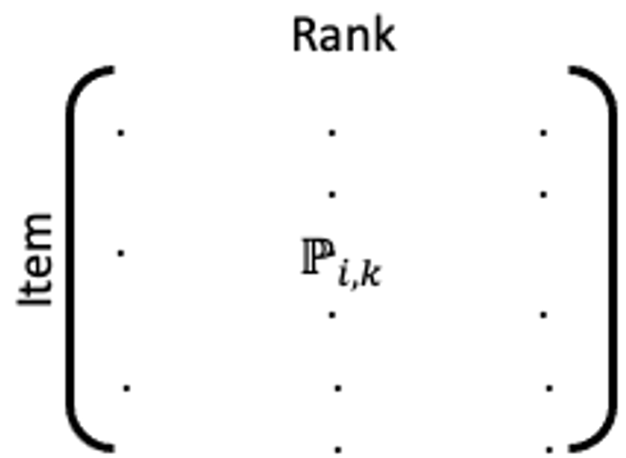
Utility (e.g., DCG, Avg Precision) and Exposure can be expressed as a Linear function of the matrix.

For example, $\text{DCG}(\mathbb{P}) = \sum_i \mu_i \sum_k \frac{\mathbb{P}_{i,k}}{\log(1+k)}$.

Optimization problem of finding $\mathbb{P}$ that optimizes utility U and satisfies fairness constraints → Linear Program

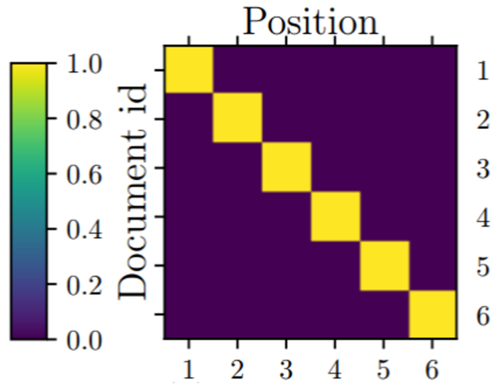
[Singh & Joachims, KDD 2018]

Key Idea 2: Doubly Stochastic Matrices



$\mathbb{P}_{i,k}$ = Probability of item i at position k.

Items	$\hat{h}(x)$
A_1	0.82
A_2	0.81
A_3	0.80
B_1	0.79
B_2	0.78
B_3	0.77



Doubly stochastic matrix representing a single ranking

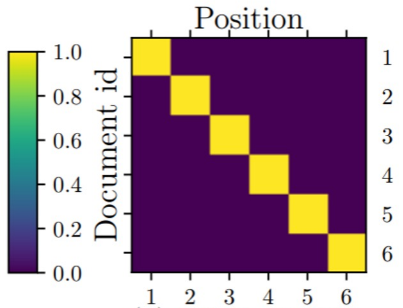
[Singh & Joachims, KDD 2018]

Example: Exposure Proportional to Relevance

Items	$\hat{h}(x)$
A_1	0.82
A_2	0.81
A_3	0.80
B_1	0.79
B_2	0.78
B_3	0.77

$\times$

Exposure@k
e_1
e_2
e_3
e_4
e_5
e_6



(a) DCG=3.8193, DTR=1.7483

Without Fairness Constraint

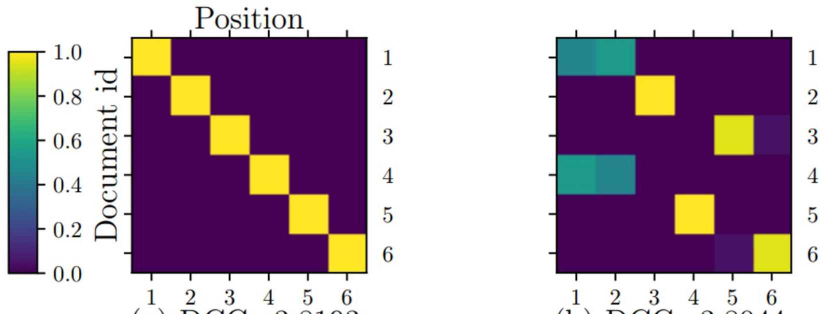
Problem setup: Maximize Utility (e.g., DCG) while fulfilling the fairness constraint (exposure proportional to relevance).

[Singh & Joachims, KDD 2018]

Example: Exposure Proportional to Relevance

Items	$\hat{h}(x)$
A_1	0.82
A_2	0.81
A_3	0.80
B_1	0.79
B_2	0.78
B_3	0.77

Exposure@k
e_1
e_2
e_3
e_4
e_5
e_6



(a) DCG=3.8193, DTR=1.7483

(b) DCG=3.8044, DTR=1.0000

Without Fairness Constraint

$\mathbb{P}_{\text{fair}}$: Proportional Exposure

Problem setup: Maximize Utility (e.g., DCG) while fulfilling the fairness constraint (exposure proportional to relevance).

Solution: Ranking Policy

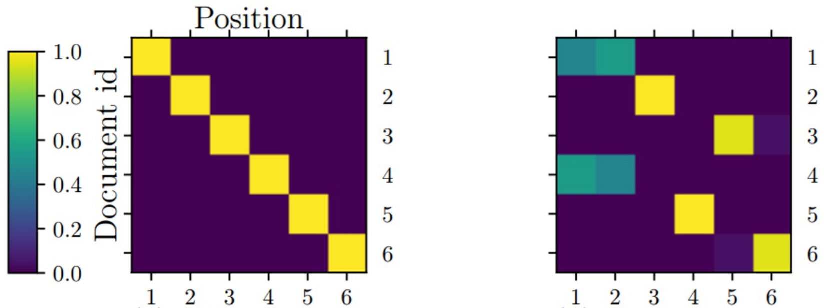
[Singh & Joachims, KDD 2018]

Example: Exposure Proportional to Relevance

Items	$\hat{h}(x)$
A_1	0.82
A_2	0.81
A_3	0.80
B_1	0.79
B_2	0.78
B_3	0.77

$\times$

Exposure@k
e_1
e_2
e_3
e_4
e_5
e_6



(a) DCG=3.8193, DTR=1.7483

(b) DCG=3.8044, DTR=1.0000

Without Fairness Constraint

$\mathbb{P}_{\text{fair}}$: Proportional Exposure

Solution: Ranking Policy

What if these relevance predictions are biased?

How to incorporate these constraints into a learning to rank framework?

[Singh & Joachims, KDD 2018]

Learning-to-Rank with fairness constraints

For a query x, rank a candidate set $\mathcal{S}_x = \{d_1, d_2, d_3, \dots\}$ of items

- d_i represented by features $\psi(d_i|x)$, and
- d_i has a merit score (e.g., relevance—whether a user would click it or not).

Ranking Policy π maps $\mathcal{S}_x$ to a ranking.

[Singh & Joachims, NeurIPS 2019]

Learning-to-Rank with fairness constraints

For a query x, rank a candidate set $\mathcal{S}_x = \{d_1, d_2, d_3, \dots\}$ of items

- d_i represented by features $\psi(d_i|x)$, and
- d_i has a merit score (e.g., relevance—whether a user would click it or not).

Ranking Policy π maps $\mathcal{S}_x$ to a ranking.

Learning objective: Find policy π that maximizes expected utility U with small disparity D

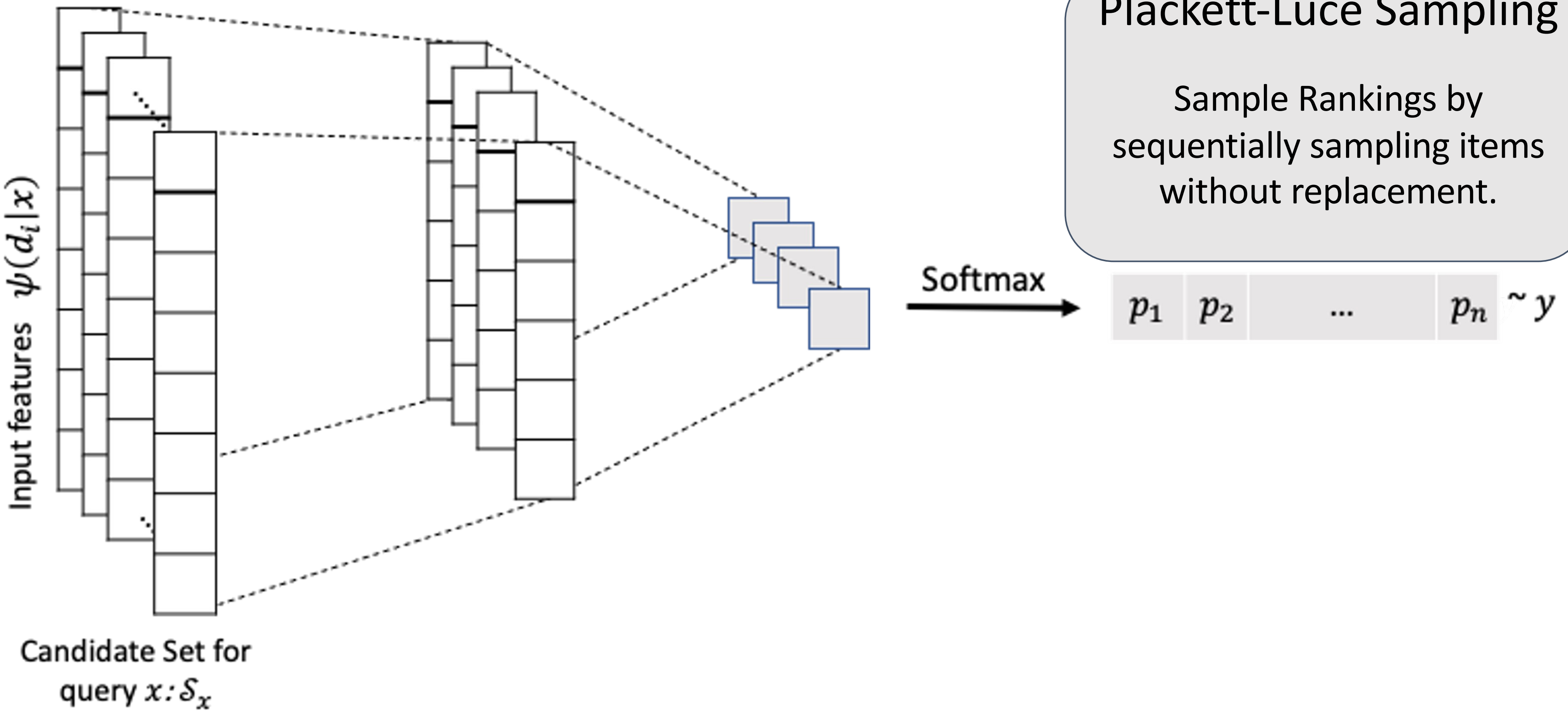
$$\pi^* = \text{argmax}_\pi\ \mathbb{E}_x[U(\pi|x)] \text{ s.t. } \mathbb{E}_x[D(\pi|x)] \leq \delta.$$

Empirical Risk Minimization with Lagrange multiplier:

$$\pi^* = \text{argmax}_\pi \frac{1}{n} \sum_{i=1}^{n} U(\pi|x_i) - \lambda \cdot D(\pi|x_i)$$

[Singh & Joachims, NeurIPS 2019]

Stochastic Ranking Policy (π)



Plackett-Luce Sampling

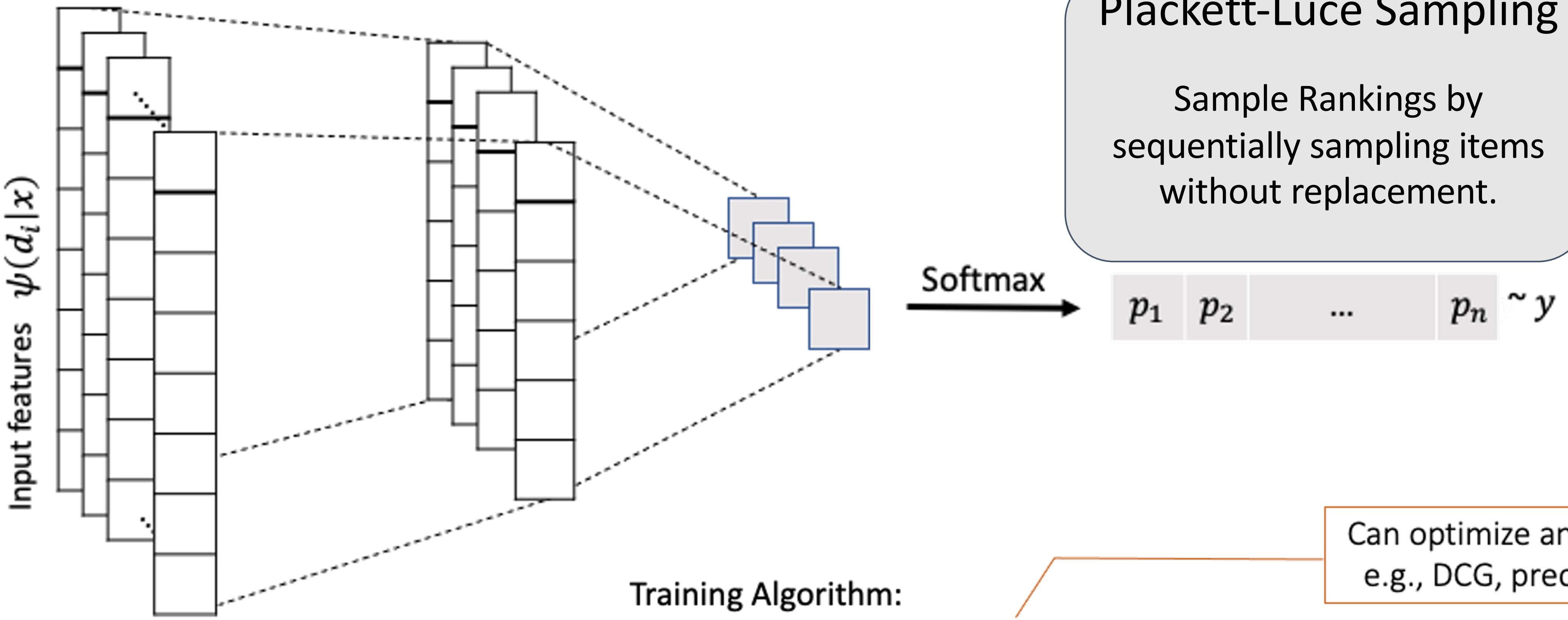
Sample Rankings by sequentially sampling items without replacement.

[Singh & Joachims, NeurIPS 2019]

Stochastic Ranking Policy (π)

Plackett-Luce Sampling

Sample Rankings by sequentially sampling items without replacement.



Candidate Set for query x: S_x

Can optimize an arbitrary metric U, e.g., DCG, prec@k, average-rank.

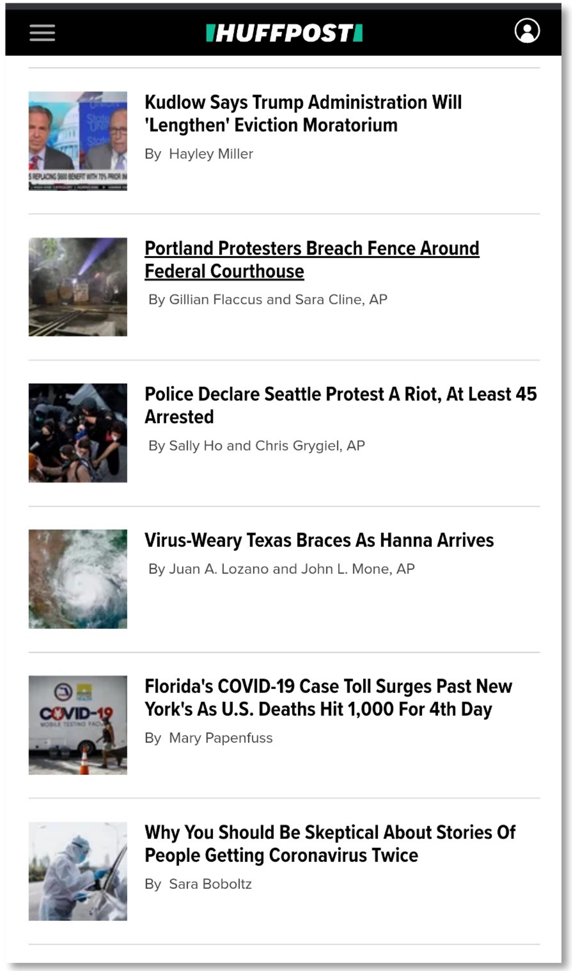
Training Algorithm:

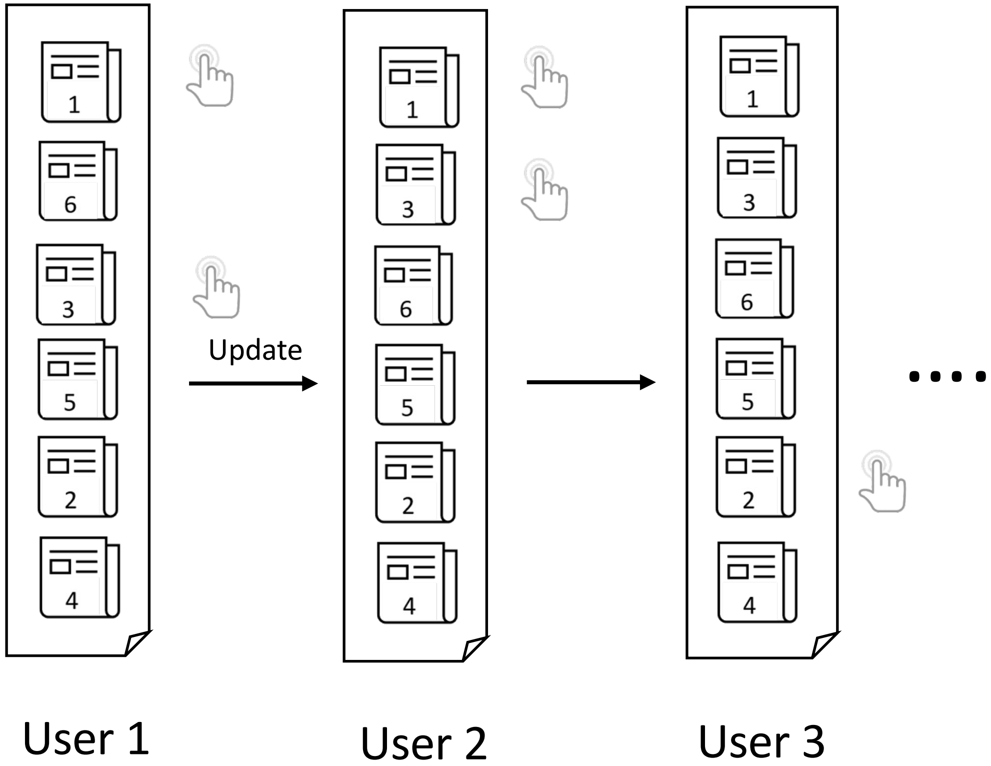
- Loss function: REINFORCE loss with the reward as $U(\pi|x_i) - \lambda \cdot D(\pi|x_i)$.
- Policy Gradient using Monte-Carlo estimates of gradient.
- Using Entropy & Variance Regularization.

[Singh & Joachims, NeurIPS 2019]

Dynamic Learning-to-Rank

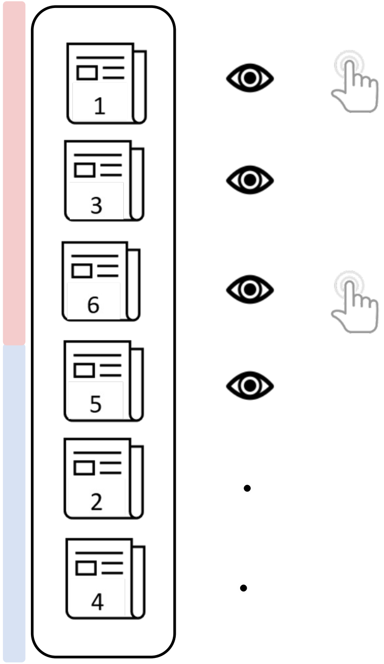
How to train a ranking policy that adapts the ranking to user interactions?





[Morik, Singh*, Hong & Joachims. SIGIR 2020]*

Dynamic Learning-to-Rank



Position Bias

Problem 1: Selection bias due to position

- Click count is not a consistent estimator of relevance.
  - Lower positions get lower attention.
  - Less attention means fewer clicks.
- Click feedback is **biased** by:
  - the deployed ranking function
  - user’s position bias

Rich-get-richer dynamic: What starts at the bottom has little opportunity to rise in the ranking.

Problem 2: Exposure disparity between groups

- Ranking solely by relevance may cause some groups to get most of the exposure on the platform.
  - For the news homepage example, this may make the platform seem biased.

[Morik, Singh*, Hong & Joachims. SIGIR 2020]*

Summary so far..

- Representation learning → Embeddings for candidate retrieval
  - Bias in embeddings → bias in candidate retrieval

- Learning-to-Rank: given candidates, how do we rank them?
  - Item-side fairness: fairness for the ranked items and stakeholders
    - Fairness in learning-to-rank algorithms
    - Dynamic learning-to-rank

- Next: Practical considerations for real-world systems

Practical Recommender Systems

- ↪️ Fairness under composition
- ↪️ Two-stage recommender systems
- ↪️ Repeated Training

Practical Recommender Systems

Fairness under composition

Even if two predictors are fair, the composition of their predictions can still be unfair.
[Fairness under Composition, *Dwork and Ilvento, ITCS 2019*]

Example: $E[\text{rating}] = P(\text{click}) \times E[\text{rating}|\text{click}] = pCTR \times pRating.$

	Author demographics			
Component	non-white	non-white	white	white
$pCTR$	0.1	0.4	0.2	0.3
$pRating$	0.4	0.1	0.3	0.2
$pCTR \times pRating$	0.04	0.04	0.06	0.06

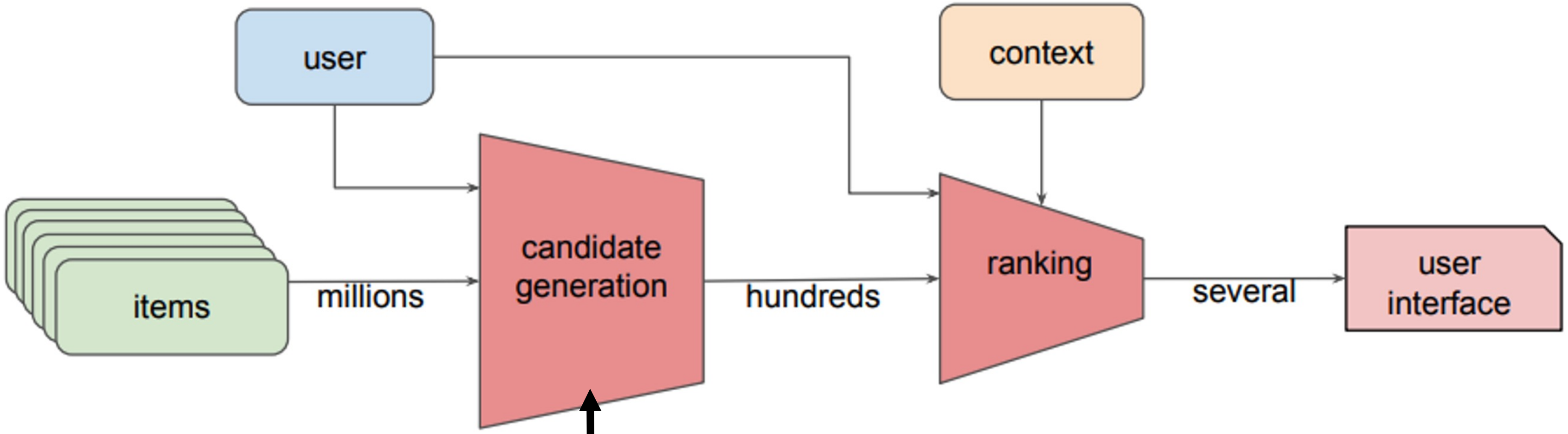
Ranking by *pCTR* or *pRating* leads to <*nw, w, w, nw*>, but ranking by their product leads to <*w, w, nw, nw*>.

[Wang et al. WSDM 2021]

Practical Recommender Systems

Two stage Recommender systems:

- Candidate generation → Ranking (→ User)



Lack of diversity at candidate generation may lead to unfair results overall

[Wang & Joachims. 2022]

Practical Recommender Systems

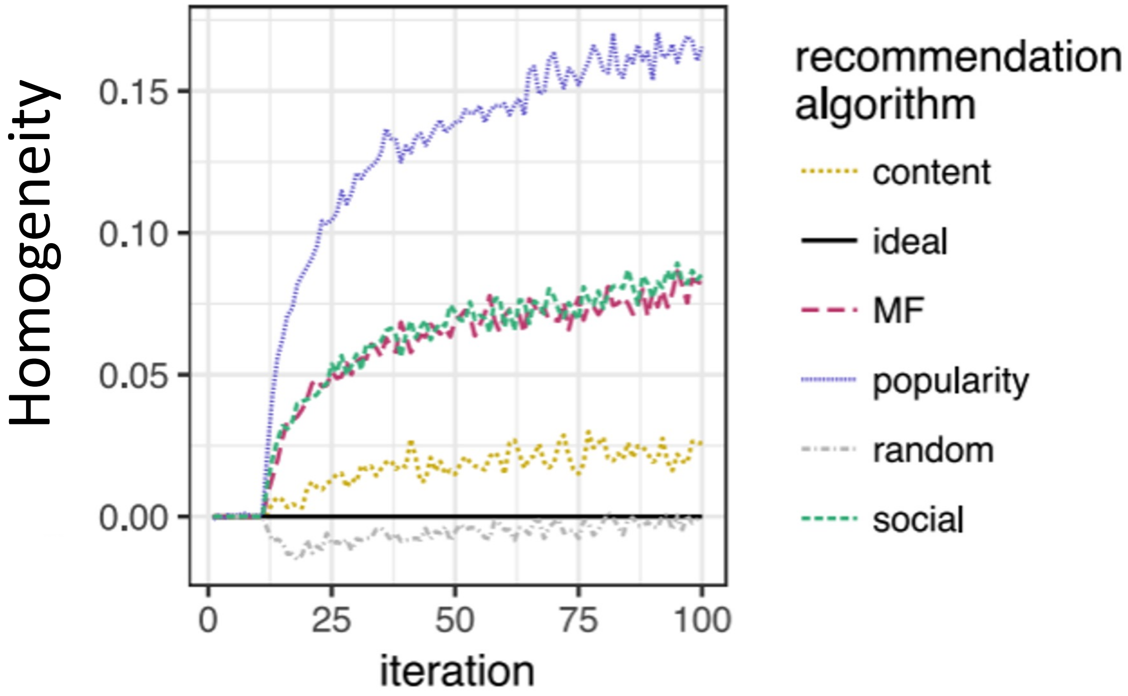
- Fairness under composition
- Two-stage recommender systems
- **Repeated Training**

Models undergo repeated training (daily, weekly, monthly).

Retraining is done using data that is confounded by algorithmic recommendations from a previously deployed system.

Consequences:

- “The recommendation feedback loop causes **homogenization of user behavior**”
- “Users experience **losses in utility** due to homogenization effects; these losses are **distributed unequally**”
- “The feedback loop **amplifies the impact of recommendation systems** on the distribution of item consumption”



Homogeneity of content recommended increases with repeated training.

[Chaney et al. RecSys 2018]

Challenges and Open Questions

- Open Questions:
  - How do users and item providers experience and perceive “unfairness”?
  - Maintaining legality:
    - How can we ensure group fairness without violating constraints around model inputs (e.g. without using protected attributes)?
    - Neutrality, monopolization, etc.
- What did we not cover but is also important?
  - Privacy
  - User safety and trust
  - Explainability and transparency

Thank you

Search and Recommender systems are the arbiters of exposure in modern two-sided online platforms.

For the long-term well-being, ranking algorithms should be able to consider utility and fairness for both users as well as creators and producers.

- Work done in collaboration with colleagues from Cornell, Google, Pinterest.
- A larger format presentation available at: https://fair-recs-tutorial.github.io/neurips-2022-tutorial/
- Feel free to reach out with questions at mail@ashudeepsingh.com