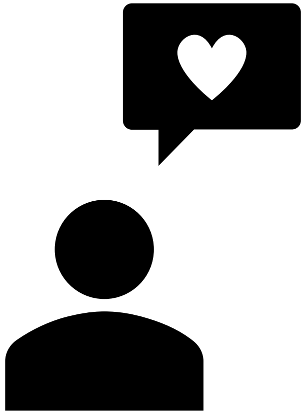
Debiasing Item-to-Item Recommendations With Small Annotated Datasets

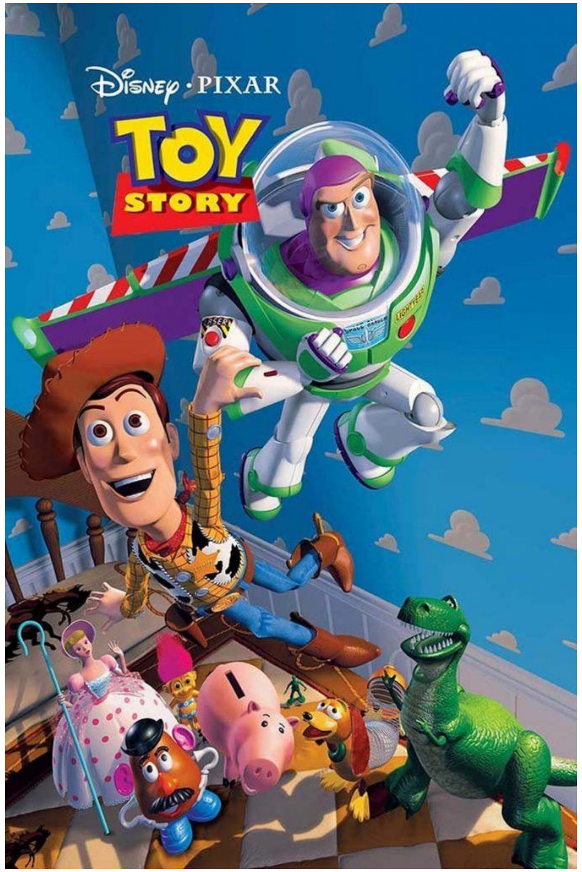
RecSys 2020

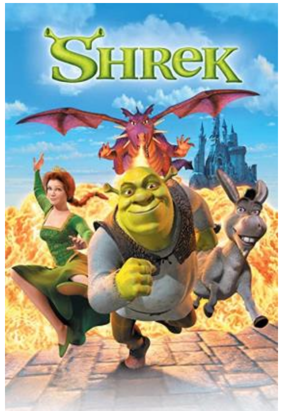
TOBIAS SCHNABEL, PAUL BENNETT

MICROSOFT

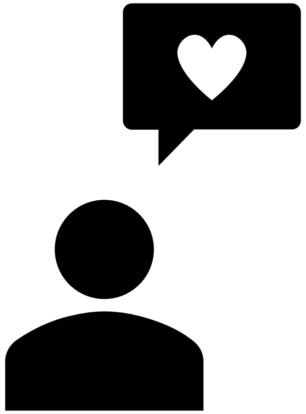
What movie would you recommend?

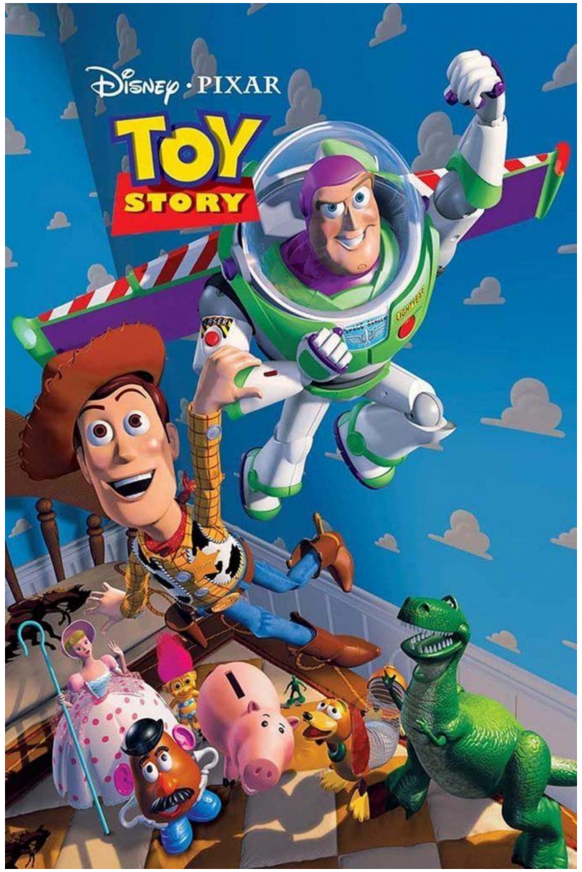






What movie would you recommend?





Idea: Use co-occurrence counts (MLE)

$$\frac{\text{counts}(i, \text{"}Toy\ Story\text{"})}{\text{counts}(\text{"}Toy\ Story\text{"})}$$

A data-driven answer

- Common approach: Find movies most likely to co-occur with „Toy Story“

rank	title	$\hat{P}^{MLE}(S[i] \mid S[j])$
1.	Forrest Gump (1994)	0.634
2.	Star Wars: Episode IV - A New Hope (1977)	0.610
3.	Shawshank Redemption, The (1994)	0.593
4.	Pulp Fiction (1994)	0.578
5.	Silence of the Lambs, The (1991)	0.554
6.	Matrix, The (1999)	0.554
7.	Jurassic Park (1993)	0.537
8.	Star Wars: Episode VI - Return of the Jedi (1983)	0.520
9.	Star Wars: Episode V - The Empire Strikes Back (1980)	0.506
10.	Back to the Future (1985)	0.500

A data-driven answer

confounding

- Common approach: Find movies most likely to co-occur with „Toy Story“

rank	title	$\hat{p}^{MLE}(S[i] \mid S[j])$	popularity rank
1.	Forrest Gump (1994)	0.634	2
2.	Star Wars: Episode IV - A New Hope (1977)	0.610	6
3.	Shawshank Redemption, The (1994)	0.593	1
4.	Pulp Fiction (1994)	0.578	3
5.	Silence of the Lambs, The (1991)	0.554	4
6.	Matrix, The (1999)	0.554	5
7.	Jurassic Park (1993)	0.537	8
8.	Star Wars: Episode VI - Return of the Jedi (1983)	0.520	16
9.	Star Wars: Episode V - The Empire Strikes Back (1980)	0.506	11
10.	Back to the Future (1985)	0.500	23

Why naive estimation fails

- **Model:**
  - Get a number of sessions S_k
  - $S_k[i] = 1$ means item i occurs in session k
- **Task:** Want estimator for $P(S[i] = 1 \ \mid S[j = \text{"Toy Story"}] = 1)$
- **Problem:** Don‘t observe S_k, but only partial S_k^{obs} (m.n.a.r.)

S_k	1	0	0	1	0	1	1	0	0
S_k^{obs}	0	0	0	1	0	1	0	0	0

$\Longrightarrow$ Naive estimator (MLE) not consistent

IPS to the rescue

- **Solution:** Use Inverse Propensity Scoring (IPS) to get consistent estimator

$$\hat{P}^{IPS}(S[i] = 1 \mid S[j] = 1) = \frac{p_i^{-1} p_j^{-1} \text{counts}(i, j)}{p_j^{-1} \text{counts}(j)}$$

- **Problem:** Where to get propensities?
  - Randomization often infeasible
  - Fitting propensity model on observational data relies on strong assumptions

Estimating propensities

- **Question:** What if we had some labeled data?
  - relevance("Up" | "Toy Story") > relevance("Star Wars IV" | "Toy Story")

- **Assume:** $P(S[\text{"Up"}] = 1 \mid S[\text{"Toy Story"}] = 1)$

  $> P(S[\text{"Star Wars IV"}] = 1 \mid S[\text{"Toy Story"}] = 1)$

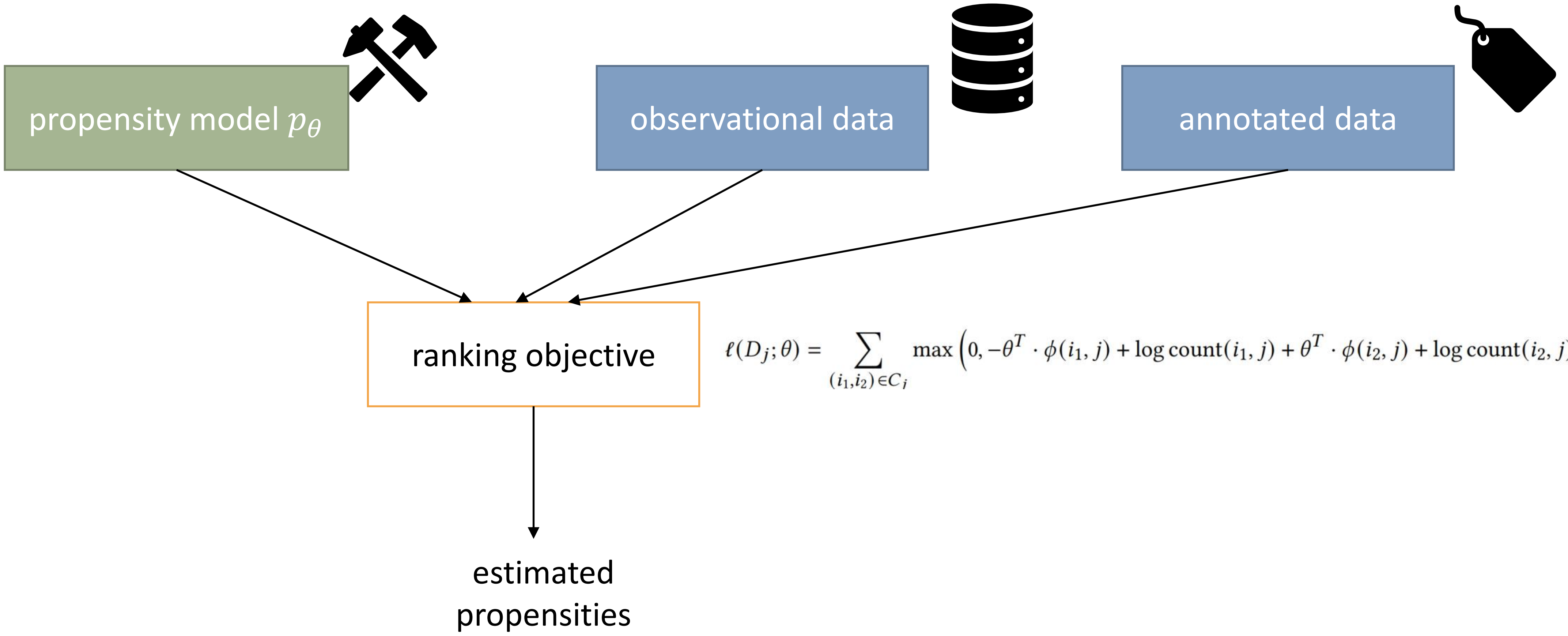
Estimating propensities

- **Question:** What if we had some labeled data?
  - relevance("Up" | "Toy Story") > relevance("Star Wars IV" | "Toy Story")

- **Assume:** $\hat{P}^{IPS}(S[\text{"Up"}] = 1 \mid S[\text{"Toy Story"}] = 1)$
  $> \hat{P}^{IPS}(S[\text{"Star Wars IV"}] = 1 \mid S[\text{"Toy Story"}] = 1)$

Find propensity model which satisfies constraints

General framework



$$\ell(D_j;\theta) = \sum_{(i_1,i_2)\in C_j} \max\Big(0, -\theta^T \cdot \phi(i_1,j) + \log \mathrm{count}(i_1,j) + \theta^T \cdot \phi(i_2,j) + \log \mathrm{count}(i_2,j)\Big).$$

Experiment setup

- **Observational data:** MovieLens 25M dataset (binarized)
- **Annotated data** [Yao & Harper, 2018]:
  - 67 rankings – each for one seed movie
  - ~ 10 relevant candidates per seed ranking
  - Sample for negatives to create relevance pairs

$$\text{relevance}(\textit{rated movie} \mid j) > \text{relevance}(\textit{random sample} \mid \text{j})$$

- **Propensity model:**
  - Uses release date, popularity, ratio of ratings w.r.t. to seed movie

Experiment setup (c'ed)

- **Metrics:**
  - Recall@k – robust to missing movies
  - Mean ranks of relevant items

- **Baselines:**
  - POPULARITY, RANDOM
  - SUPERVISED: Learn relevance label directly
  - MF-based: PURESVD, WRMF, BPR, SLIM

- Pick best hyperparameters for each model on val according to metric

Results

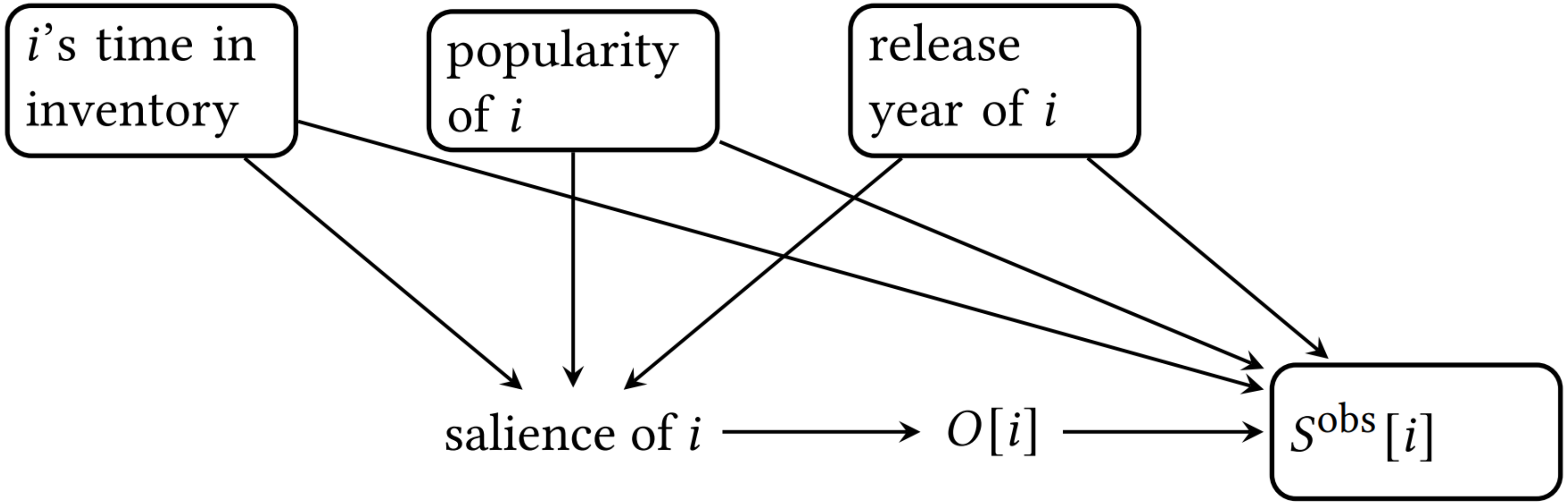
method	Recall@25	Recall@50	Recall@100	mean ranks
Random	0.000	0.000	0.000	6959.3
Pop	0.000	0.016	0.025	1850.4
Supervised	0.399	0.539	0.646	520.7
Cooccur	0.058	0.123	0.268	676.3
ItemKNN	0.436	0.529	0.594	450.9
PureSVD	0.356	0.450	0.532	673.8
WRMF	0.361	0.469	0.539	890.5
Bpr	0.365	0.455	0.515	785.8
Slim	0.487	0.639	0.657	2191.5
Ours	**0.532**	**0.652**	**0.761**	**213.5**

Results (qualitative)

rank	OURS	ITEMKNN	SLIM	WRMF
1.	Toy Story 2 (1999)	Toy Story 2 (1999)	Toy Story 2 (1999)	Toy Story 2 (1999)
2.	Toy Story 3 (2010)	Willy Wonka & t... (1971)	Toy Story 3 (2010)	Toy Story 3 (2010)
3.	Finding Nemo (2003)	● Back to the Future (1985)	Willy Wonka & t... (1971)	Muppet Treasure... (1996)
4.	Incredibles, The (2004)	Monsters, Inc. (2001)	Aladdin (1992)	James and the Gi... (1996)
5.	Monsters, Inc. (2001)	Lion King, The (1994)	● Star Wars IV (1977)	Willy Wonka & t... (1971)
6.	Shrek 2 (2004)	Bug's Life, A (1998)	Monsters, Inc. (2001)	Bug's Life, A (1998)
7.	Shrek (2001)	● Independence Day (1996)	● Independence Day (1996)	101 Dalmatians (1996)
8.	Bug's Life, A (1998)	● Star Wars IV (1977)	● Back to the Future (1985)	● Space Jam (1996)
9.	Ratatouille (2007)	Aladdin (1992)	James and the Gi... (1996)	● Star Wars IV (1977)
10.	Up (2009)	Star Wars VI (1983)	Finding Nemo (2003)	Aladdin (1992)

Picking a propensity model

- **Consider:**
  - Statistical efficiency
  - Causal validity



Conclusions

- **High-level picture:** Method that estimates causal parameters via
  - small annotated dataset
  - assumption about relationship of true causal effects and annotations
- Applied it to **item-to-item recommendation**:
  - formalized as an estimation problem from missing data
  - leverages IPS estimator to treat biases in a principled way
- **Future work:**
  - Learning guarantees / identification of parameters
  - Applying it to other scenarios where annotation is easy (e.g., search)