Bad for IR, Worse for Recommenders

Missing Data and the External Validity of Offline Evaluations

MICHAEL D. EKSTRAND

BOISE STATE UNIVERSITY

People and Information Research Team

TL;DR

Missing data seriously compromises the external validity of standard IR evaluation paradigms

IR has techniques that try to address these issues

They don’t work for recommender systems

And the problems are worse

What can editorials in 1960 sci-fi magazines tell us about evaluating recommender systems?

Agenda

- Rehearse Evaluation
- Missing Data Problems
- IR Solutions
- Why They Don’t Work For RecSys
- Why We Don’t Have a Solution
- Promising Directions

Evaluation Strategies

Online, by measuring live user response

Offline, by using existing data sets

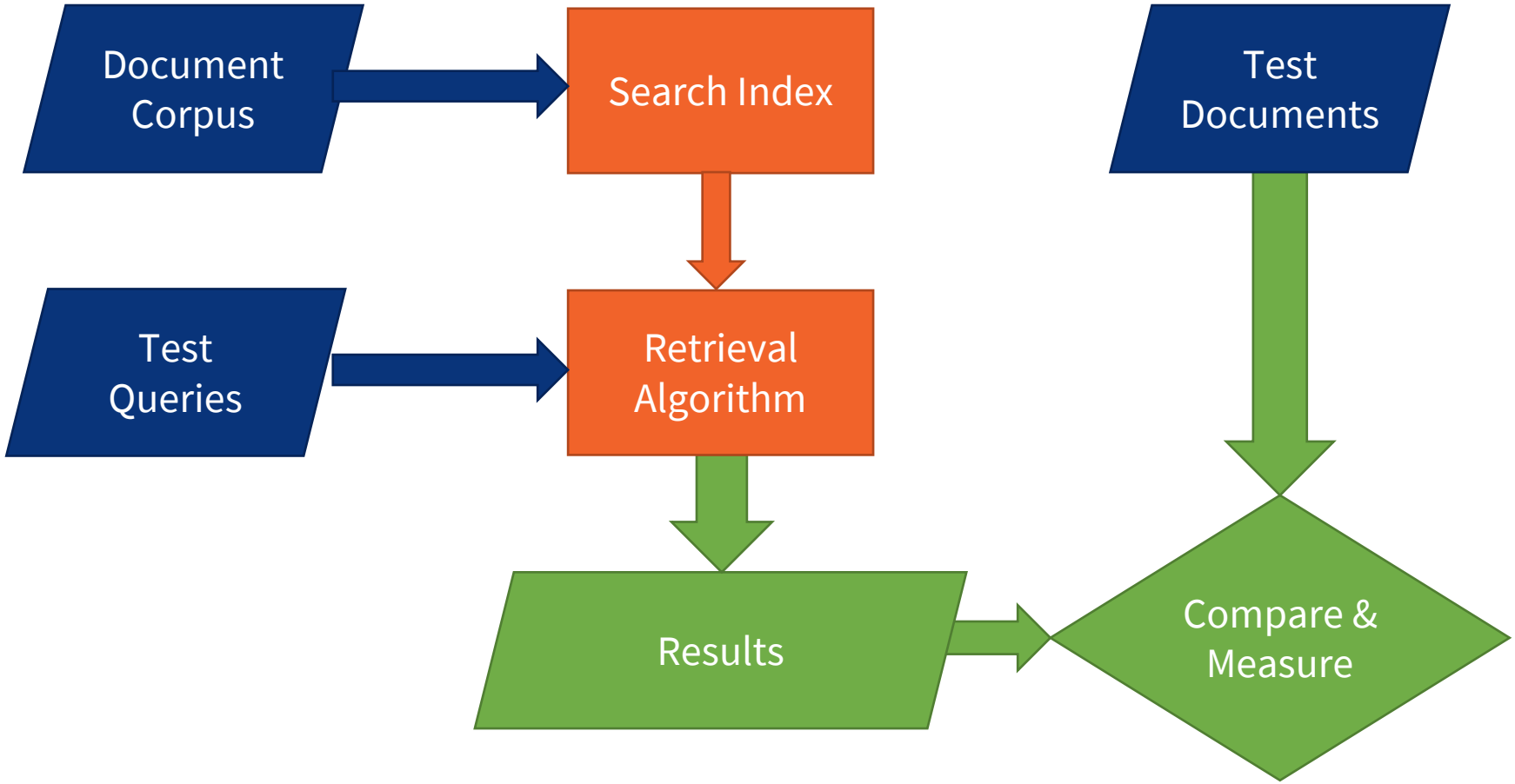
- **Prediction accuracy** with rating data (RMSE)
- **Retrieval (Top-*N*) accuracy** with ratings, purchases, clicks relevance, etc. (MAP, MRR, P/R, AUC, nDCG)

Why Offline Still Matters

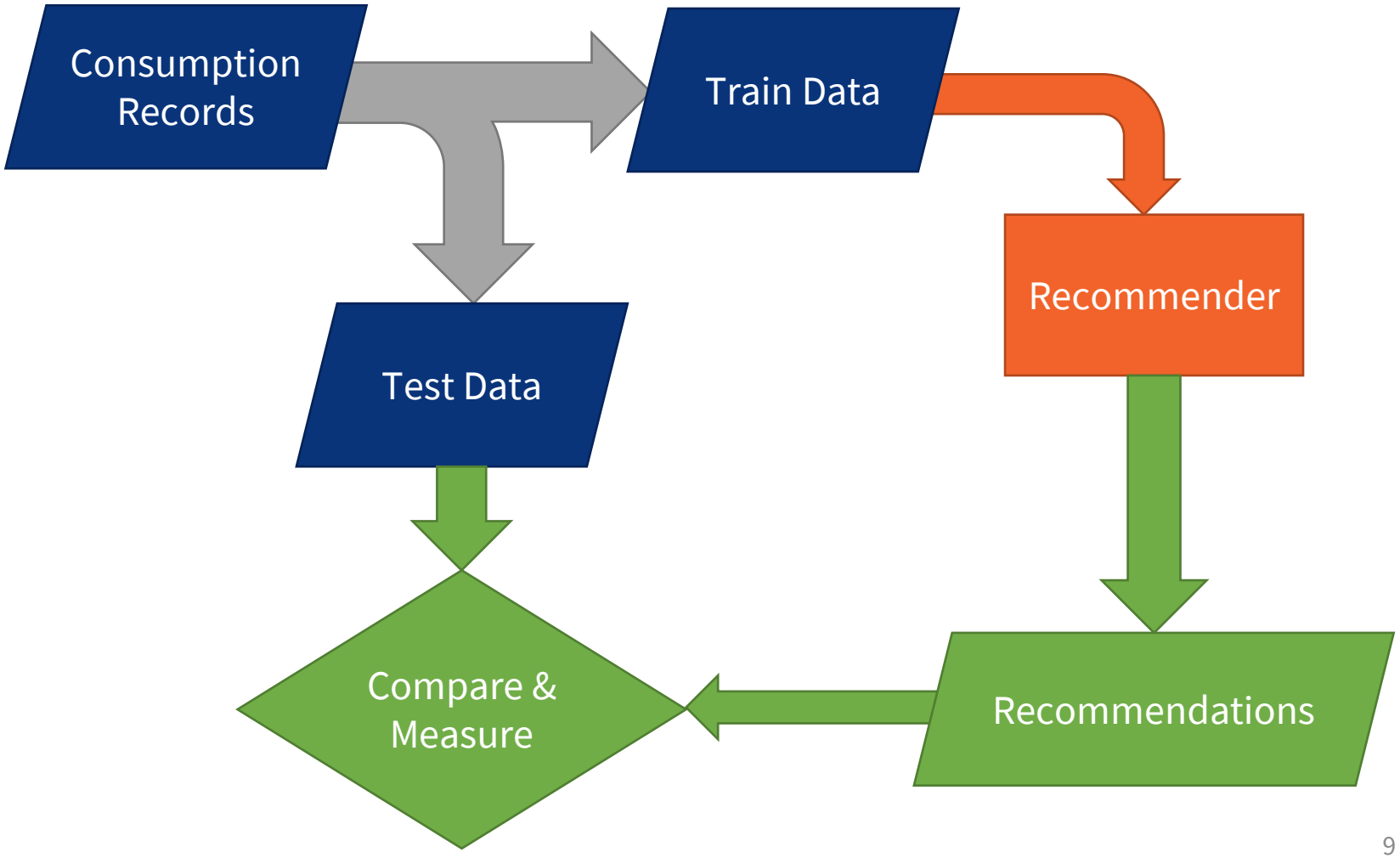
Pre-select Algorithms

Optimize Algorithms

Offline Evaluation (IR)



Offline Evaluation (RecSys)



Evaluation Data

- Expert-judged relevance
- User-judged relevance/utility (rating)
- User response (click)

Measuring Accuracy

Metric	Rank-Aware?	Judgement Used
Precision, Recall, F	No	Binary Rel.
AUC	Partially	Binary Rel.
MRR	Yes	Binary Rel.
MAP	Yes	Binary Rel.
NDPM	Partially	Relative Rel.
NDCG	Yes	Utility

Computing Metrics

☒ *Zootopia*

☑ *The Iron Giant*

☑ *Frozen*

☒ *Seven*

☒ *Tangled*

RR = 0.5

AP = 0.417

nDCG = 0.815

Missing Data

- ☐ *Zootopia*
- ☑ *The Iron Giant*
- ☑ *Frozen*
- ☒ *Seven*
- ☐ *Tangled*

RR = 0.5

AP = 0.417

nDCG = 0.815

IR: can't get complete relevance judgements

RecSys: users haven't rated or bought all items

Only Positives

- ☐ *Zootopia*
- ☑ *The Iron Giant*
- ☑ *Frozen*
- ☐ *Seven*
- ☐ *Tangled*

RR = 0.5

AP = 0.417

nDCG = 0.815

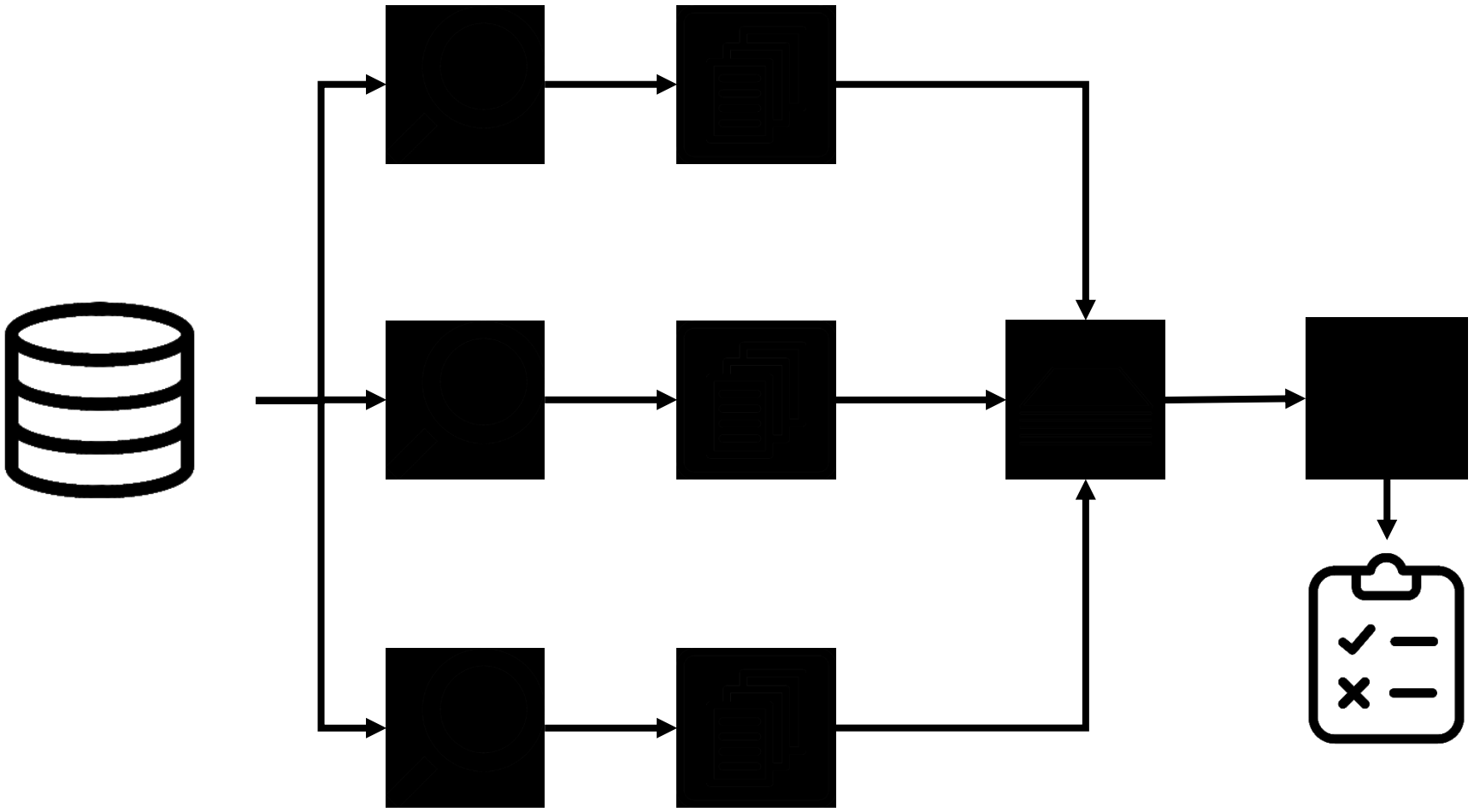
Assume:

unknown ⇒ irrelevant

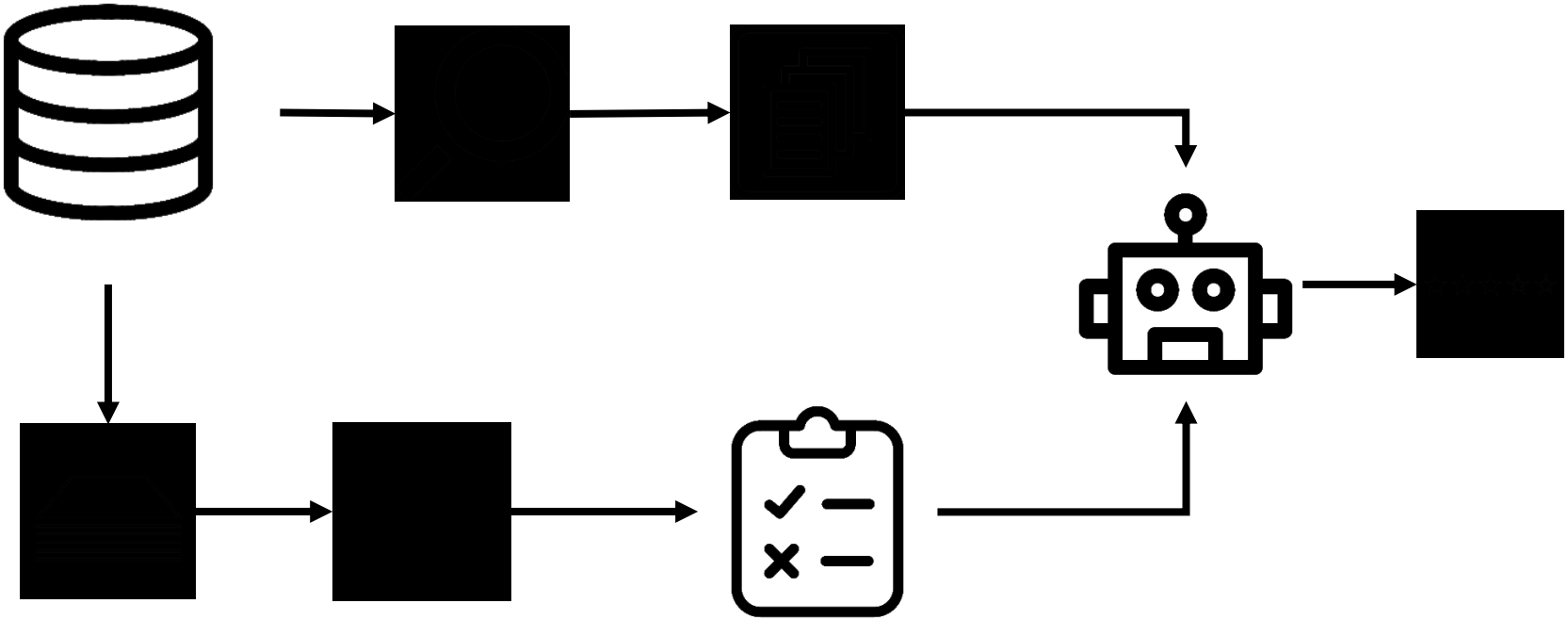
Not true, so how to fix?

Soln. 1: Pooling

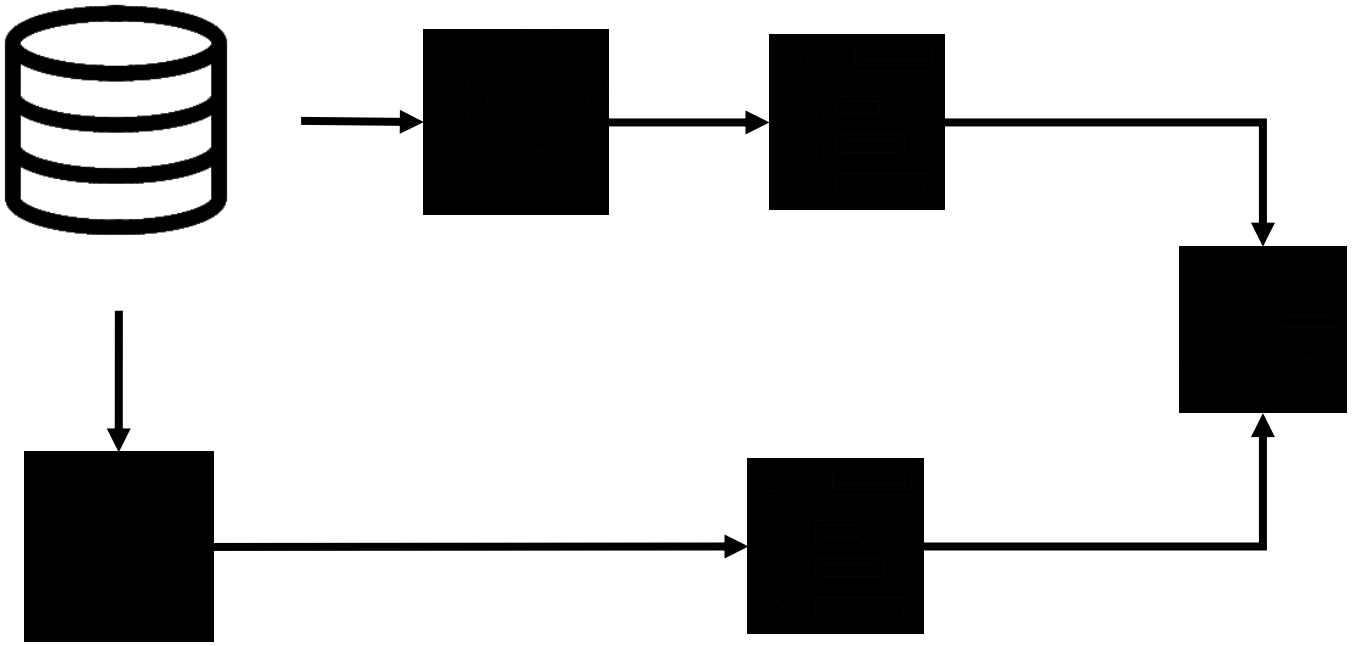
TREC [Buckley and Voorhees, 2004]



Soln. 2: Relevance Inference
[Aslam and Yilmaz, 2007]



Soln. 3: Rank Effectiveness
[Aslam and Yilmaz, 2006]



Soln. 4: Graded Relevance
[Jiang et al., 2015]

- Collect expert judgements of user satisfaction
- Train model to predict from user activity logs
- Estimate satisfaction for unjudged sessions

Requires expert judges, only works for online system logs

Where We're At

Several solutions seem to improve IR situation (for objective settings).

Let's talk about recommendation…

It Gets Worse

Recommendation relevance is subjective.

- Only expert is the user
- Do they find it interesting?

Popularity has a strong effect

Ungraded items may be the b

Offline metrics known to be weak predictors of online performance [Rossetti et al., 2016]

Solution Suitability

- ~~Pooling~~
  - Requires expert judgements
- ~~Relevance Inference~~
  - Begs the question
- Rank Effectiveness
  - Works if we have relative feedback
- ~~Graded Relevance~~
  - Insufficient data in offline data sets

Popularity Bias

Offline evaluations favor recommenders that recommend popular items

[Bellogin et al., 2011; Bellogin 2012]

Misclassified Decoys

☐ *Zootopia*

☑ *The Iron Giant*

☑ *Frozen*

☒ *Seven*

☐ *Tangled*

3 possibilities for *Zootopia*:

- I don't like it
- I do but data doesn't know
- I do but **I don't know yet**

RR = 0.5

AP = 0.417

Misclassified Decoys

If I would like *Zootopia*

But have not yet seen it

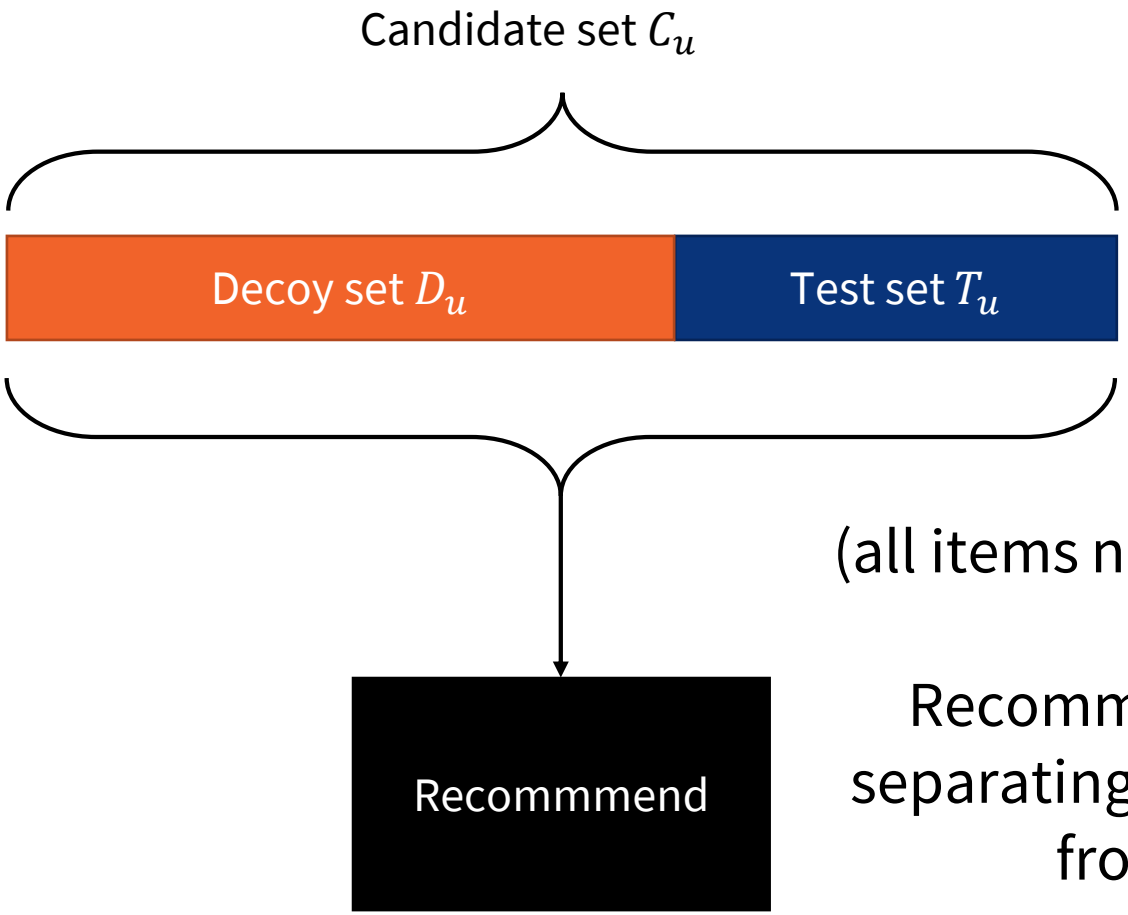
Then it is likely a **very good** recommendation

But the recommender is penalized

Solution Space

- Change **test rating selection**
  - Can reduce popularity bias [Bellogin, 2012]
- Change **candidate set**
  - Can mitigate misclassified decoys
  - Enables rank effectiveness
- ???

The Candidate Set



Often: $C_u = I \setminus R_u$
(all items not rated in training)

Recommender is a *classifier* separating relevant items (T_u) from decoy items (D_u)

Sturgeon's Law

Ninety percent of everything is crud.

— T. Sturgeon (1958)

Only 1% is 'really good'

— P. S. Miller (1960)

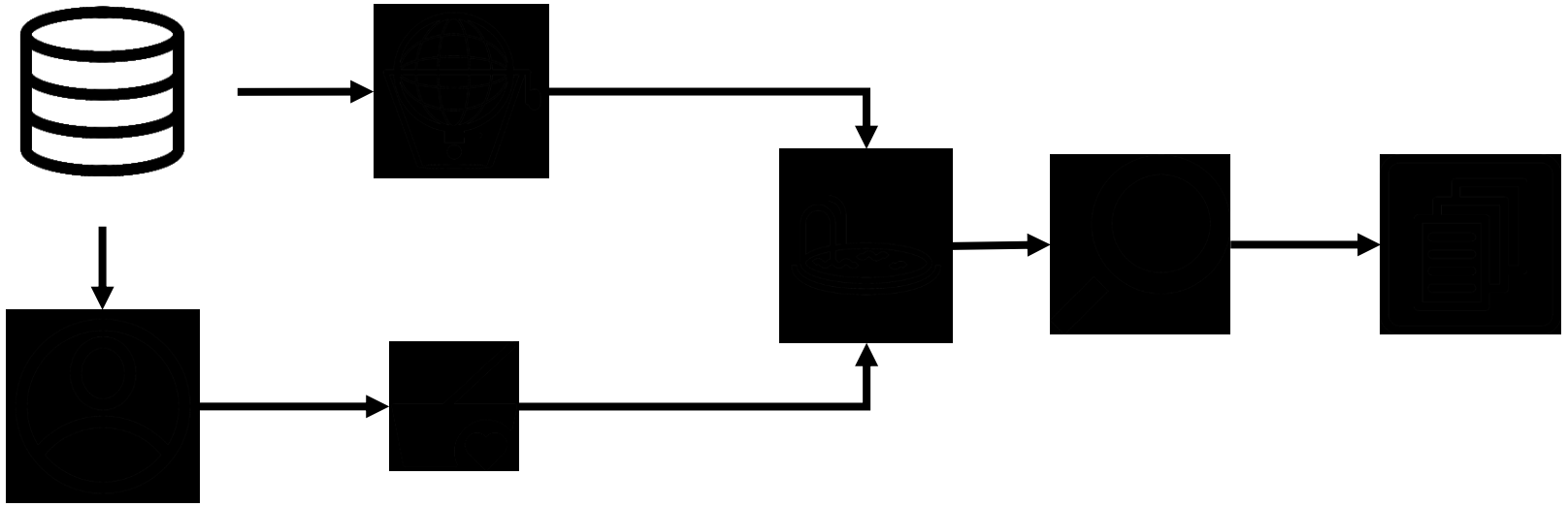
Sturgeon's Decoys

Most items are not relevant.

Corollary: a randomly-selected item is probably not relevant.

Random Decoys

One-Plus-Random protocol (Cremonesi et al. 2010)



Can generalize to test items + random items

- Koren (2008): right # of decoys is open problem

Modeling Goodness
[Ekstrand and Mahant, 2017]

Starting point:

Goodness rate $g = \Pr[i \in G_u]$, probability i is good for u

Want: $\Pr[D_u \cap G_u = \emptyset] \geq 1 - \alpha$

high likelihood of no misclassified decoys

Simplifying assumption: goodness is independent

$$\Pr[D_u \cap G_u = \emptyset] = \prod_{i \in D_u} \Pr[i \notin G_u] = (1 - g)^N$$

What's the damage?

For $\alpha = 0.05$ (95% certainty), $N = 1000$

$$1 - g = 0.95^{\frac{1}{N}}$$
$$g = 0.0001$$

Only 1 in 10,000 can be relevant!

MovieLens users like 10s to 100s of 25K films

Why so serious?

If there is even one good item in the decoy set …

… then it is the recommender's job to find that item

If no unknown items are good, why recommend?

Our evaluation strategies must account for this.

Sturgeon and Popularity

Random items are …

… less likely to be relevant

… less likely to be popular

Result: popularity is **even more** likely to separate test items from decoys

oops

Additional Problems

- Demographic bias (majority group decides popularity, dominates averages)
- Under-represented groups or topics

Some Useful Directions

- Rossetti et al.'s online/offline comparison study [2016] — relate cheap & expensive evaluations
- Counterfactual data sets and evaluation [Lefortier et al., 2016]
- Demographic-aware evaluation [Mehrotra et al., 2017]

Simulate all the things!

- Estimate reliability of test settings [Urbano 2013, 2016]
- Estimate reliability and errors in protocols?
- Estimate possible future outcomes?

A Provocation: Marginal Probability Ranking

Basic Ranking: $\Pr[i \in R_q]$

Improve: $\Pr[i \in R_q | \forall j < i . j \notin R_q]$ [Goffman, 1964]

We aren't even trying

Questions?

https://goo.gl/4u2Rof

BOISE STATE UNIVERSITY

People and Information Research Team