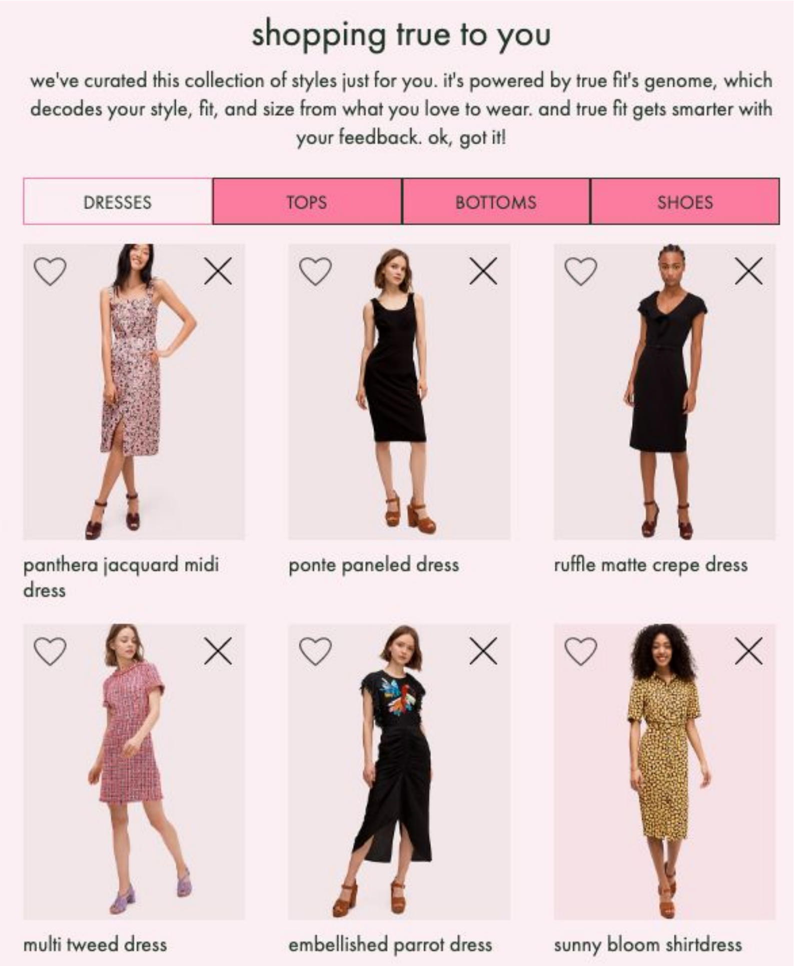
Assessing Fashion Recommendations:
A Multifaceted Offline Evaluation Approach

Jake Sherman, Chinmay Shukla, Rhonda Textor, Su Zhang, Amy A. Winecoff

About True Fit

- We provide footwear and apparel size and style recommendations
- Our clients range from large, multi-brand retailers (e.g., Macy’s), to smaller, single-brand retailers (e.g., Kate Spade)
- Over 100M people have received a recommendation from True Fit



Challenges of the fashion domain

- Different recommendations for different users (i.e., personalization) is a goal
- Accuracy alone is insufficient to measure offline performance
- Acute cold-start problem due to volume of “new” users
- Exceptional data sparsity

Objective

Developing a holistic offline evaluation approach that:

- Includes metrics to measure whether or not different users are getting different recommendations
- Performs evaluations for multiple user slices based on user interaction histories (i.e., new versus existing users) to measure cold-start performance

Measuring distinctness

Start by measuring the distinctness of a pair of users' top-k recommendations:

$$AD_{k,i,j} = |L_{k,i} \triangle L_{k,j}| = |(L_{k,i} - L_{k,j}) \cup (L_{k,j} - L_{k,i})|$$

This is the symmetric difference between the two sets of recommendations

Then, take the average $AD_{k,i,j}$ across all possible pairs of users:

$$AD_k = \frac{1}{\frac{1}{2}(U^2 - U)} \cdot \sum_{i=1}^{U} \sum_{j=i+1}^{U} AD_{k,i,j}$$

Distinctness example for two users

A pair of users’ top-5 recommendations:

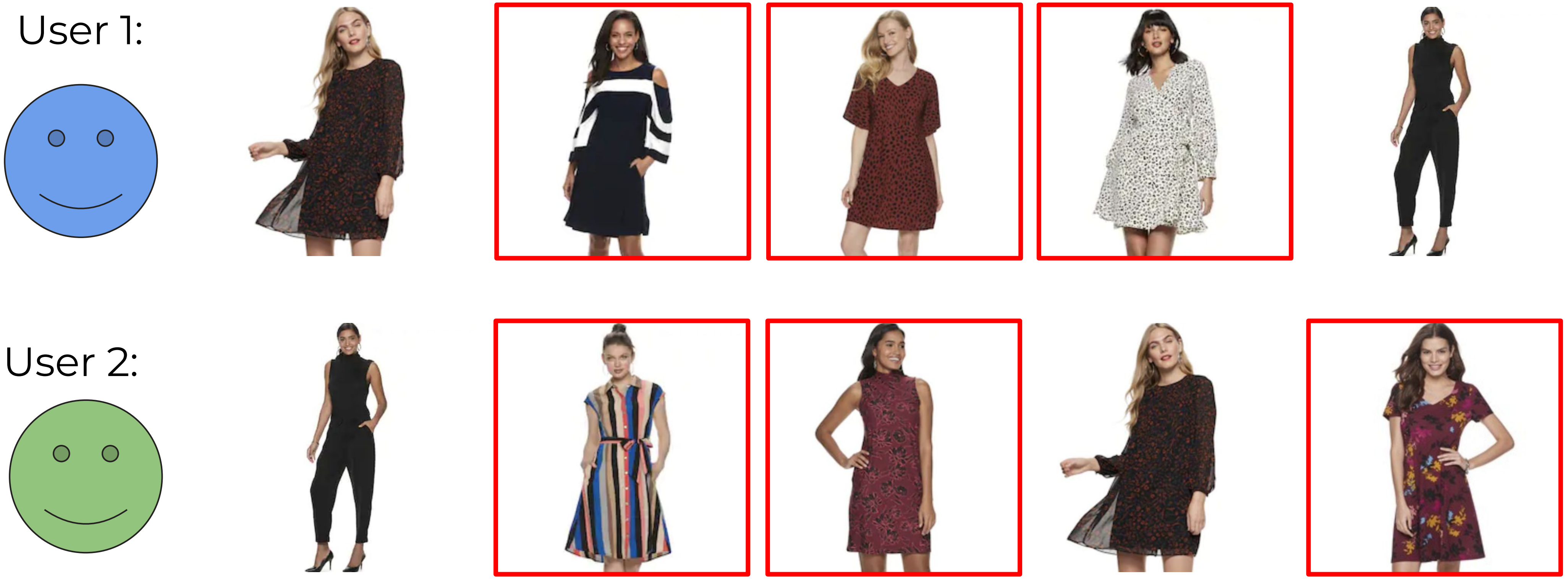
User 1:

User 2:

Distinctness example for two users

A pair of users' top-5 recommendations:

6 recommendations out of the 10 are distinct



Measuring popularity

Start by measuring the relative popularity of a user's top-k recommendations:

$$RP_{k,u} = \frac{\sum_{i=1}^{k} Q_{u,i}}{\sum_{i=1}^{k} Q_i}$$

Quantity sold of user's top-k recommendations

Quantity sold of the k most popular items across all users

Then, take the average of $RP_{k,u}$ across all users:

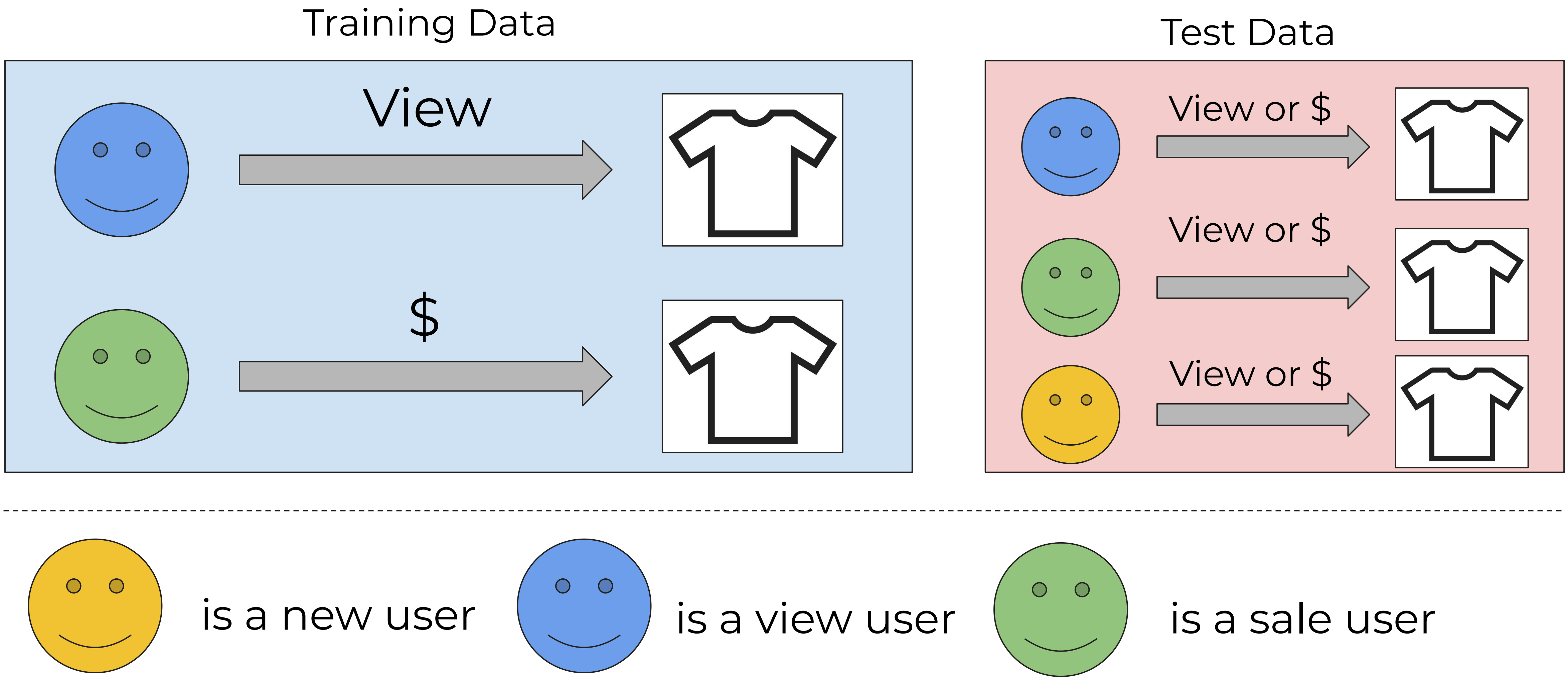
$$RP_k = \frac{1}{U} \cdot \sum_{u=1}^{U} RP_{k,u}$$

Objective

Developing a holistic offline evaluation approach that:

- ✅ **Includes metrics to measure whether or not different users are getting different recommendations**
- Performs evaluations for multiple user slices based on user interaction histories (i.e., new versus existing users) to measure cold-start performance

Defining user slices based on user interactions in the training data



Objective

Developing a holistic offline evaluation approach that:

- ✓ **Includes metrics to measure whether or not different users are getting different recommendations**
- ✓ **Performs evaluations for multiple user slices based on user interaction histories (i.e., new versus existing users) to measure cold-start performance**

Demonstrating the value of our approach

In order to demonstrate the effectiveness of our proposed offline evaluation approach, we will:

- Create recommendations using 3 different recommendation strategies, for 3 different retailers
- Use our evaluation approach to reveal the strengths and weaknesses of each recommendation strategy

Our data is extremely sparse and faces major cold-start challenges

The majority of our users are new (no view or sale in the training data)

Table 1: Descriptive Statistics for Training Data

	Users	Products	Sales (%)	Views (%)	Unobserved (%)
Retailer 1	39,307	376	7,461(0.05%)	103,829(0.7%)	14,668,142(99.2%)
Retailer 2	42,490	865	8,276(0.02%)	143,781(0.4%)	36,601,793(99.6%)
Retailer 3	60,333	386	21,904(0.1%)	141,320(0.6%)	23,125,314(99.3%)

Table 2: Descriptive Statistics for Test Data

	New Users (%)	View Users (%)	Sale Users (%)	Products	Sales (%)	Views (%)	Unobserved (%)
Retailer 1	2,477(73.8%)	667(19.9%)	213(6.3%)	319	1,727(0.2%)	8,850(0.8%)	1,060,306(99.0%)
Retailer 2	6,048(69.0%)	1,997(22.8%)	720(8.2%)	676	5,171(0.1%)	50,443(0.9%)	5,869,526(99.1%)
Retailer 3	5,164(71.2%)	1,513(20.9%)	578(7.9%)	314	2,753(0.1%)	19,578(0.9%)	2,255,739(99.0%)

Our data is extremely sparse and faces major cold-start challenges

The overwhelming majority of the user-item matrix is empty

Table 1: Descriptive Statistics for Training Data

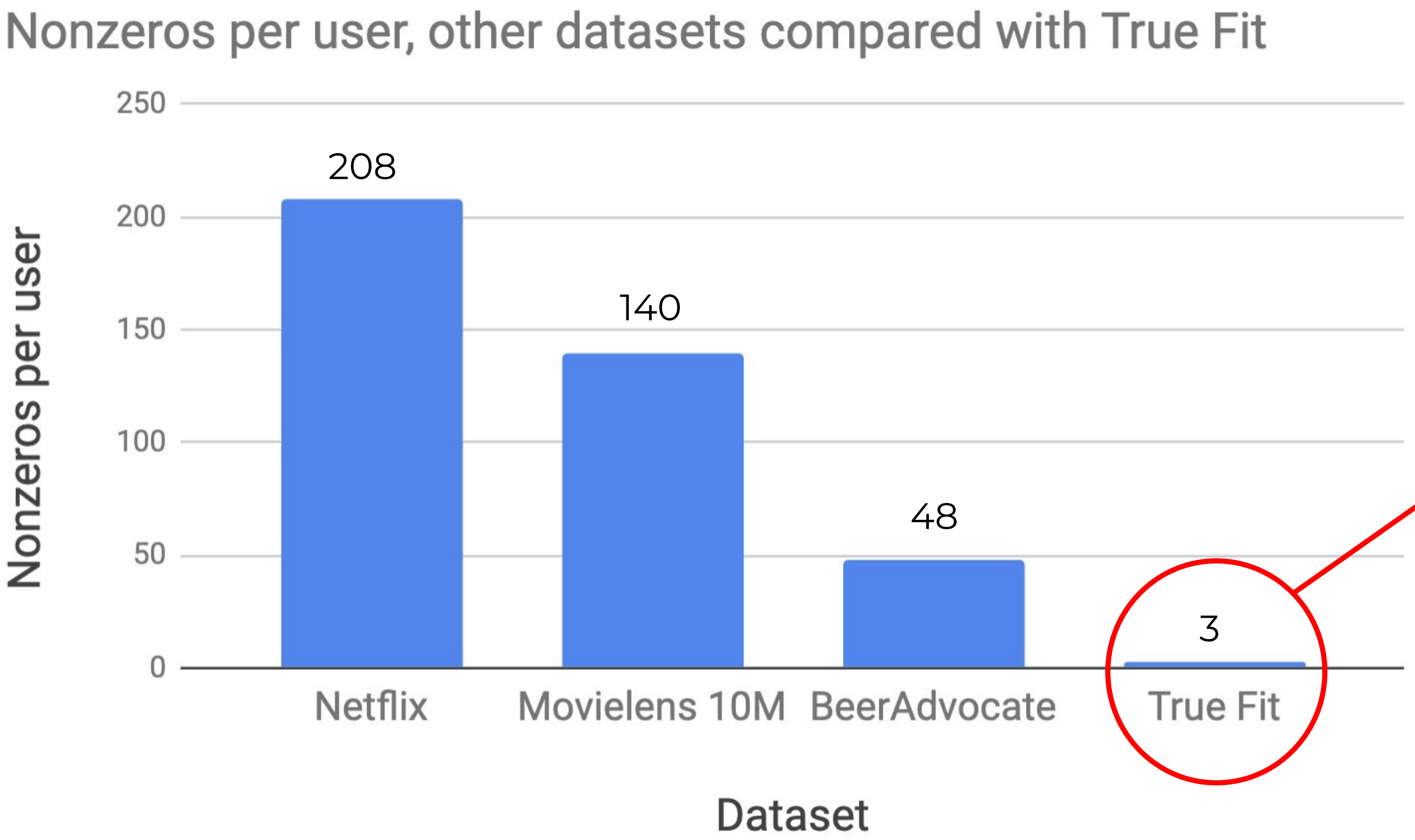
	Users	Products	Sales (%)	Views (%)	Unobserved (%)
Retailer 1	39,307	376	7,461(0.05%)	103,829(0.7%)	14,668,142(99.2%)
Retailer 2	42,490	865	8,276(0.02%)	143,781(0.4%)	36,601,793(99.6%)
Retailer 3	60,333	386	21,904(0.1%)	141,320(0.6%)	23,125,314(99.3%)

The majority of our users are new (no view or sale in the training data)

Table 2: Descriptive Statistics for Test Data

	New Users (%)	View Users (%)	Sale Users (%)	Products	Sales (%)	Views (%)	Unobserved (%)
Retailer 1	2,477(73.8%)	667(19.9%)	213(6.3%)	319	1,727(0.2%)	8,850(0.8%)	1,060,306(99.0%)
Retailer 2	6,048(69.0%)	1,997(22.8%)	720(8.2%)	676	5,171(0.1%)	50,443(0.9%)	5,869,526(99.1%)
Retailer 3	5,164(71.2%)	1,513(20.9%)	578(7.9%)	314	2,753(0.1%)	19,578(0.9%)	2,255,739(99.0%)

Fashion data is exceptionally sparse



Many fewer nonzeros per user for True Fit compared with other datasets

How we setup our experiment

Recommendation strategies:

1. Most popular items (MP)
2. Collaborative filtering (CF)
3. Content-based modeling (CB)

Evaluation metrics:

- Standard metrics: *normalized discounted cumulative gain at k* ($NDCG_k$)
- Our metrics: *average distinctness at k* (AD_k), *relative popularity at k* (RP_k)

Recommending popular items maximizes accuracy...

MP results in more accurate (higher $NDCG_{10}$) recommendations than CF or CB

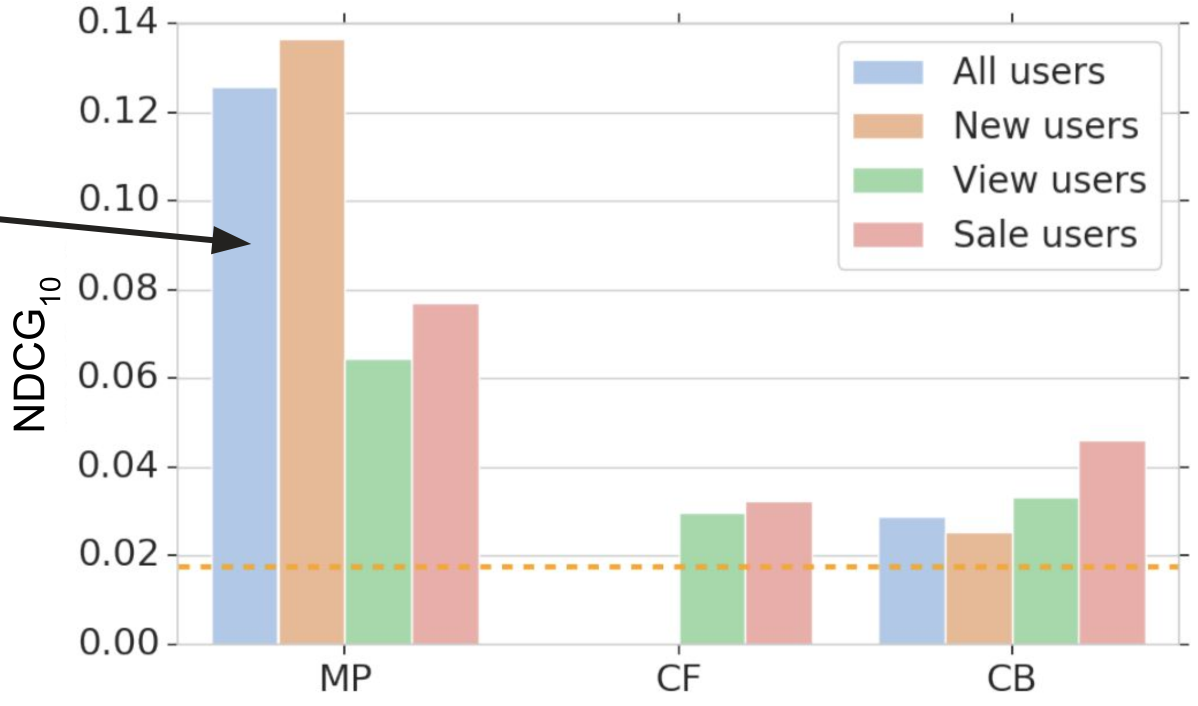


Figure 1: $NDCG_{10}$ for Retailer 1. The yellow dotted line corresponds to the $NDCG_{10}$ value for Retailer 1 that would result from a random ranking of the items.

Recommending popular items maximizes accuracy...

MP results in more accurate (higher $NDCG_{10}$) recommendations than CF or CB

CF suffers from the cold-start problem, **cannot make recommendations for 70%+ of users**

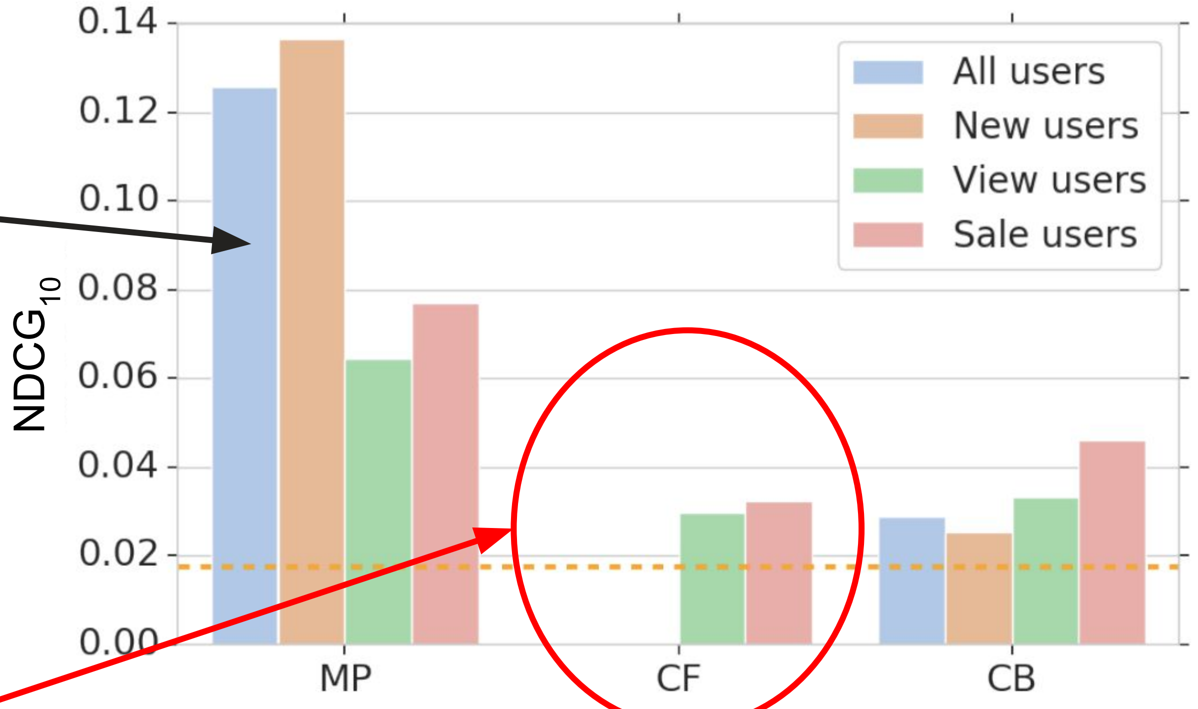


Figure 1: $NDCG_{10}$ **for Retailer 1. The yellow dotted line corresponds to the** $NDCG_{10}$ **value for Retailer 1 that would result from a random ranking of the items.**

...but results in recommendations that are not distinct...

MP recommendations are not distinct

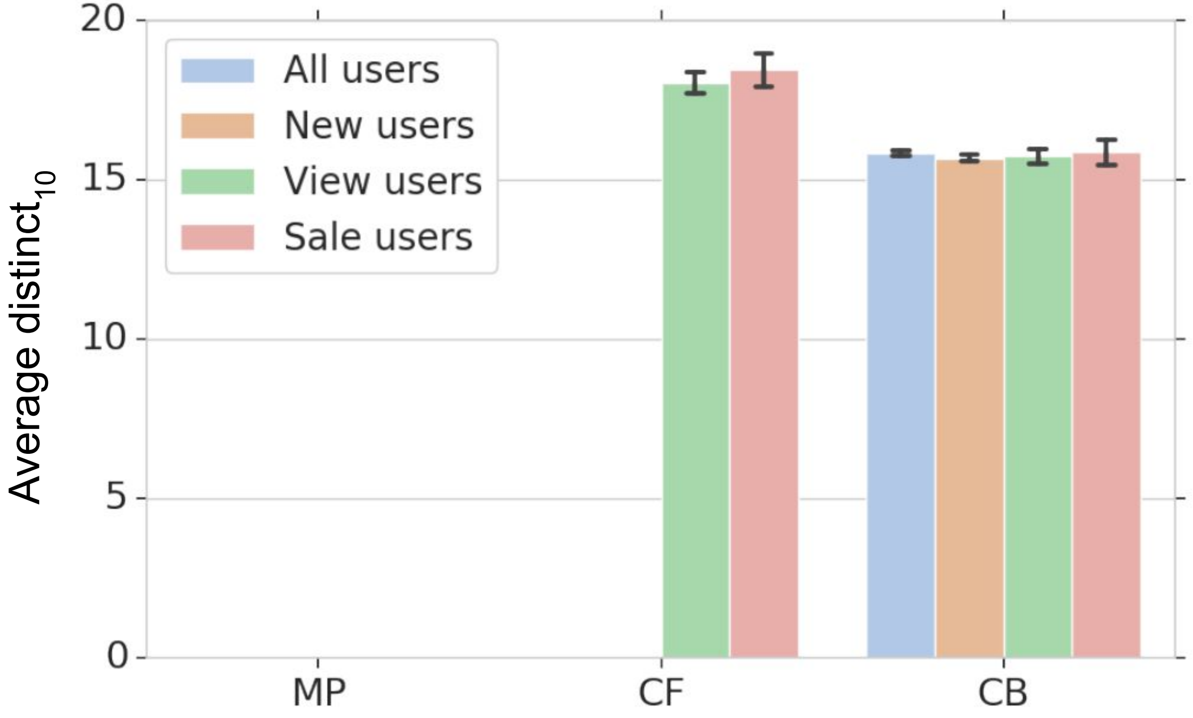


Figure 2: AD_{10} for Retailer 1. Each error bar represents the 95% confidence interval of the distribution of 1,000 bootstrap samples of $AD_{k,i,j}$ values. The MP recommendation strategy produces the same recommendations for all users, resulting in values of 0.

...but results in recommendations that are not distinct...

MP recommendations are not distinct

While CF and CB both offer distinct recommendations, CF cannot make recommendations for a majority of our users

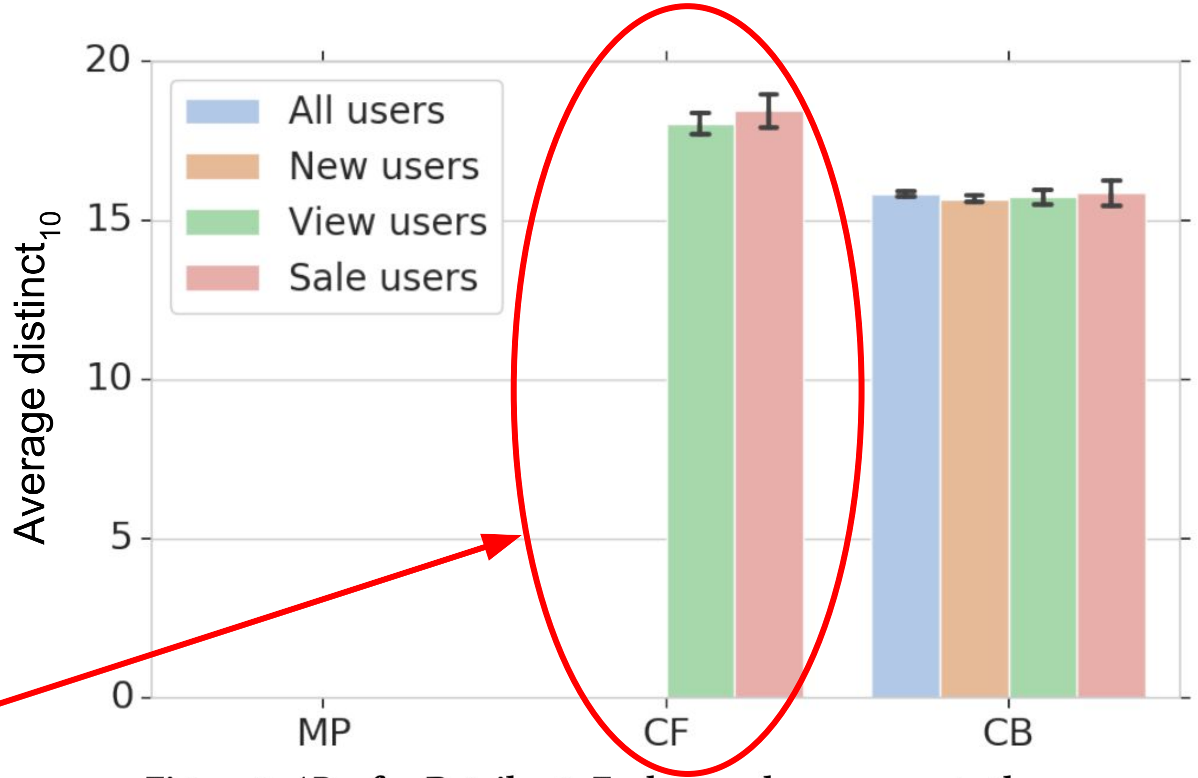


Figure 2: AD_{10} for Retailer 1. Each error bar represents the 95% confidence interval of the distribution of 1,000 bootstrap samples of $AD_{k,i,j}$ values. The MP recommendation strategy produces the same recommendations for all users, resulting in values of 0.

...and are completely popularity-biased

MP recommendations are completely popularity biased

While CF and CB each offer less popularity-based recommendations than MB, CF, once again, suffers from the cold-start problem

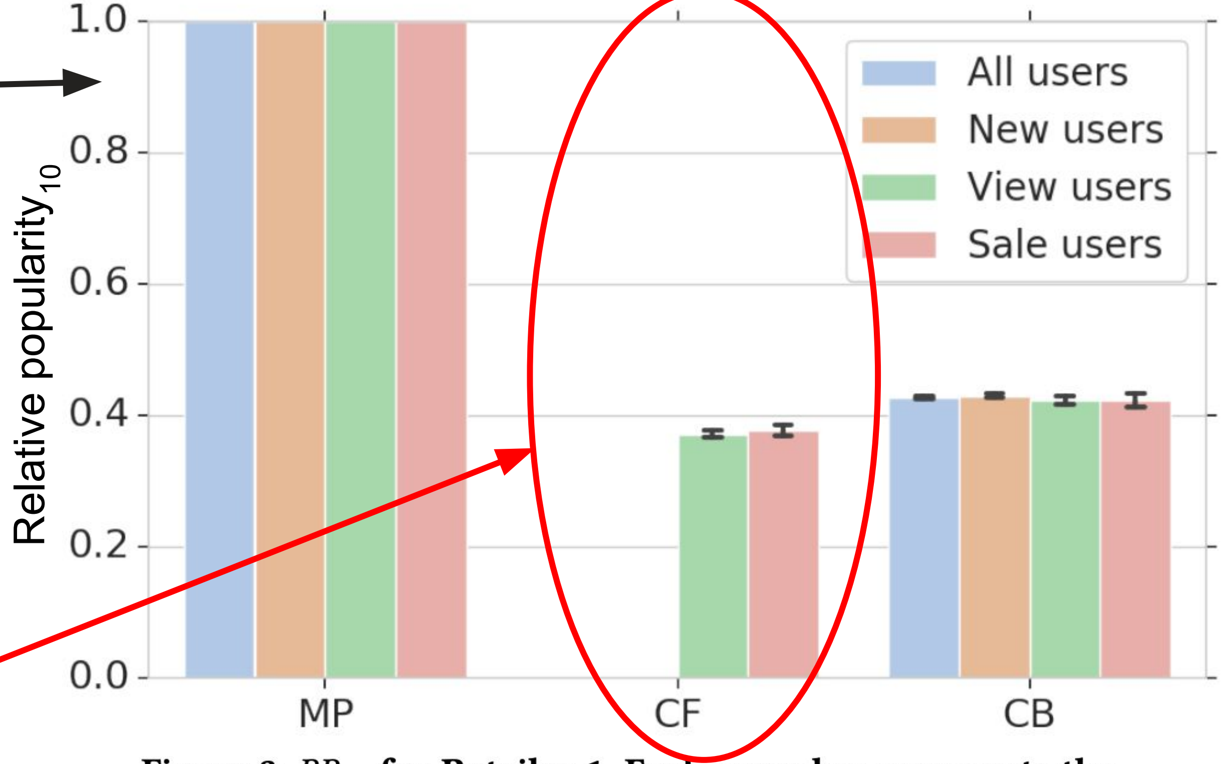


Figure 3: RP_{10} for Retailer 1. Each error bar represents the 95% confidence interval of the distribution of 1,000 bootstrap samples of $RP_{k,u}$ values. By only recommending the most popular items, the MP recommendation strategy always produces values of 1.

Conclusions

In order to perform a comprehensive offline evaluation of a fashion recommender system, one must do the following:

- Use metrics to measure whether or not different users are getting different recommendations, in addition to accuracy
- Perform evaluations for multiple user slices based on user interaction histories (new versus existing users)

Thank you

Appendix

Explaining retailer-specific results

- We suspect that differences in patterns of retailer results driven by retailer sales distributions (popularity)
- High $NDCG_{10}$ and RP_k of Retailer 2
- Retailer 2 being the exception where $NDCG_{10}$ and RP_k are higher for CF than CB

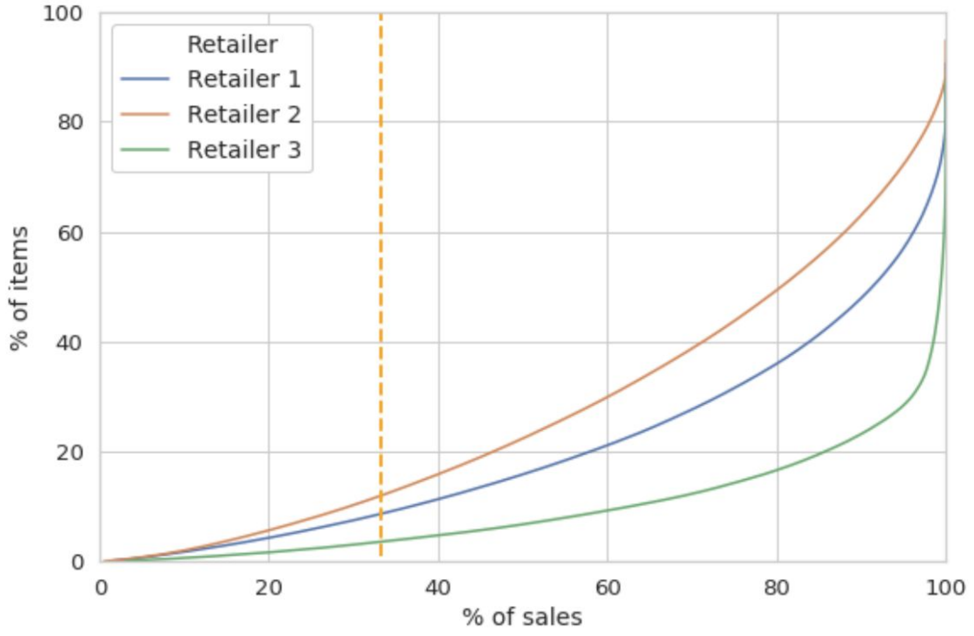


Figure 4: Sales distributions for our three retailers. Items are ordered by popularity, with the most popular items at the bottom. The set of popular items that make up a third of sales is known as the short-head, while the set of remaining items make up the long-tail [4]. The yellow dashed line provides the demarcation between the items in the short-head and long-tail.