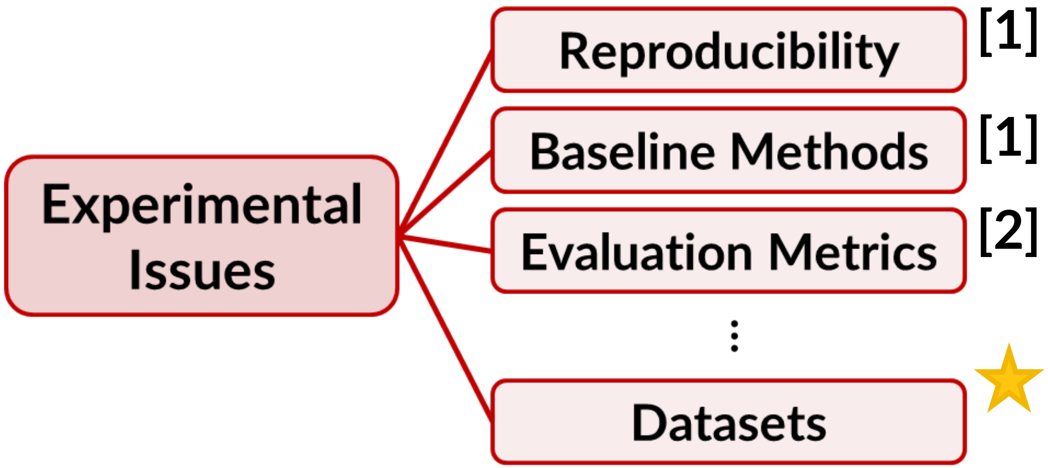
The Datasets Dilemma:
How Much Do We Really Know About Recommendation Datasets?

Jin Yao CHIN, Yile CHEN, Gao CONG

School of Computer Science and Engineering
Nanyang Technological University
Singapore

WSDM 2022

Experimental Issues



- ▷ **Reproducibility ☹**
  - Missing/incomplete datasets, source code, etc.
- ▷ **Baseline Methods ☹**
  - Weak or poorly tuned baselines
- ▷ **Evaluation Metrics ☹☹**
  - “Sampled metrics”, i.e. ranking based on a *randomly sampled subset* of candidate items

[1] A Worrying Analysis of Recent Neural Recommendation Approaches, RecSys 2019 (Best Paper)
[2] On Sampled Metrics for Item Recommendation, KDD 2020 (Best Paper)

The 'Datasets Dilemma'

- For other research fields, there are certain *"benchmark datasets"*
  - E.g., ImageNet, etc. for ***Computer Vision***
  - E.g., Stanford Question Answering Dataset (**SQuAD**) for ***Question & Answering***

- As for ***Recommendation Systems***…
  - **No** benchmark datasets
  - The choice of datasets used for empirical evaluation seems to be a **fundamental** but **often neglected** aspect

The 'Datasets Dilemma'

"How much do we really know about recommendation datasets?"

1. <u>**How**</u> are different datasets being utilised in recent papers?
   - Are there any patterns?

2. <u>**What**</u> are the **similarities** as well as **differences** between various datasets?
   - Can we define them using objective measures?

3. <u>**If**</u> the choice of datasets used could **influence** the **observations** and/or **conclusions obtained**?
   - Empirical study using a variety of item recommendation algorithms

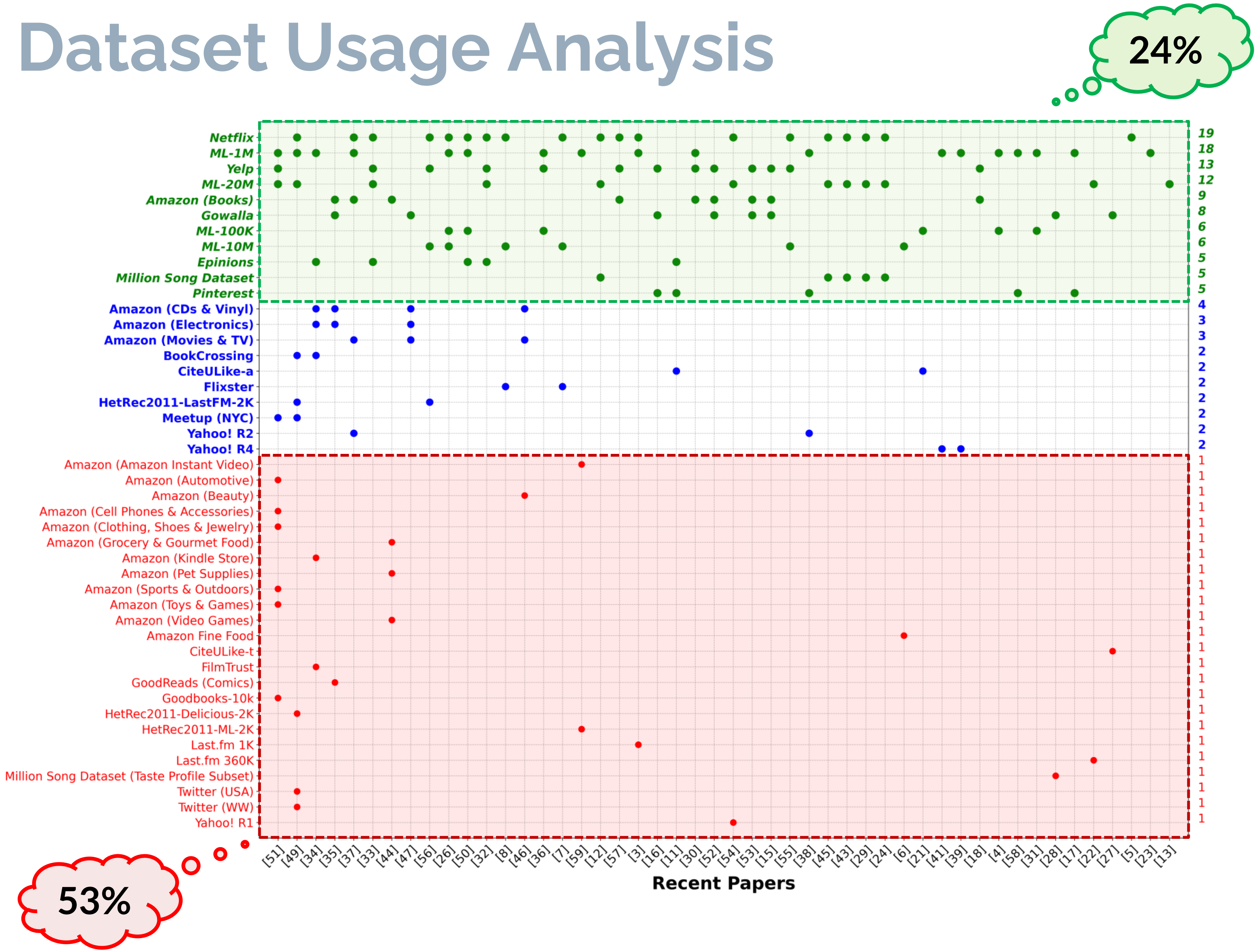
Paper and Dataset Collection

- **Conferences**: KDD, SIGIR, TheWebConf, WSDM, and RecSys
- **Years**: 2016 to 2020
- ***Keyword search*** based on title → **~400 full papers**
- ***Manual filtering***
  1. Implicit feedback-based top-K recommendation
  2. Evaluated using classification and/or ranking metrics
  3. Utilizes at least 1 publicly available dataset
- Obtained a total of **48** full papers

Useful property for analyzing usage patterns

"A dataset used in any single one of these papers can be used in every other paper as well."

Dataset Usage Analysis



Frequent Combinations

Same Author

Datasets	Papers
Epinions, ML-20M, Netflix, Yelp	[32, 33]
ML-20M, Million Song Dataset, Netflix	[12, 24, 29, 43, 45]
Amazon (Books), Gowalla, Yelp	[15, 52, 53]
Amazon (CDs & Vinyl; Electronics), Gowalla	[35, 47]
Flixster, ML-10M, Netflix	[7, 8]
ML-100K, ML-1M, Netflix	[26, 50]
ML-10M, Netflix, Yelp	[55, 56]
ML-1M, ML-20M, Meetup (NYC)	[49, 51]

▷ We use the Apriori algorithm to determine the **combinations of datasets** which have been used together in *2 or more papers*

▷ *Most frequent pairing would be {ML-20M, Netflix}*

- Evaluated at the same time in 9 separate papers

The 'Datasets Dilemma'

"How much do we really know about recommendation datasets?"

1. How are different datasets being utilised in recent papers?
   - The choice of datasets is often ***determined arbitrarily***
   - Difficult to compare results between different papers
2. What are the **similarities** as well as **differences** between various datasets?
   - Can we define them using objective measures?
3. If the choice of datasets used could **influence** the **observations** and/or **conclusions obtained**?
   - Empirical study using a variety of item recommendation algorithms

Dataset Characteristics

Two different ***types*** of dataset characteristics [1]

1. Structural
2. Distributional

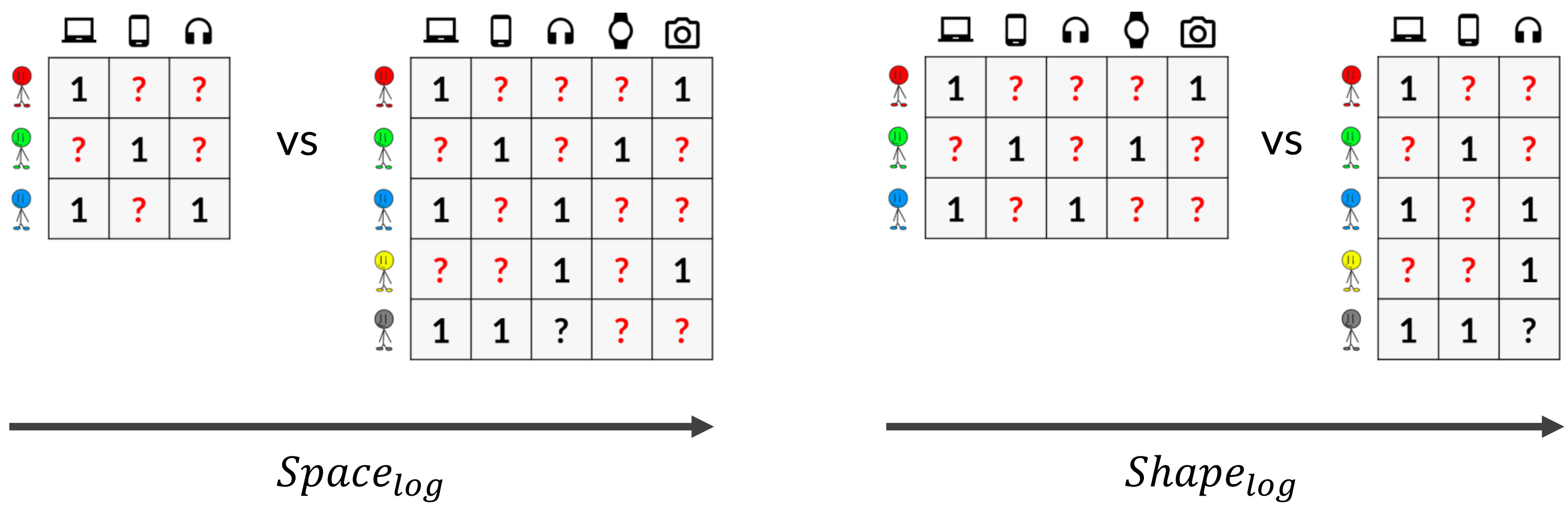
Easily derived
from the
user-item
interaction matrix

[1] Impact of Data Characteristics on Recommender Systems Performance. ACM TMIS, 3(1), April 2012

Structural Characteristics

- ▷ $Space_{log} = log_{10}\left(\frac{|U| \text{ x } |I|}{1000}\right)$
- ▷ $Shape_{log} = log_{10}\left(\frac{|U|}{|I|}\right)$
- ▷ $Density_{log} = log_{10}\left(\frac{|K|}{|U| \text{ x } |I|}\right)$

- $|U|$ = # of Users
- $|I|$ = # of Items
- $|K|$ = # of Ratings



Distributional Characteristics

▷ $Gini_{user} = 1 - 2\sum_{u=1}^{|U|}\left(\frac{|U|+1-u}{|U|+1}\right) \text{x} \left(\frac{|K_u|}{|K|}\right)$

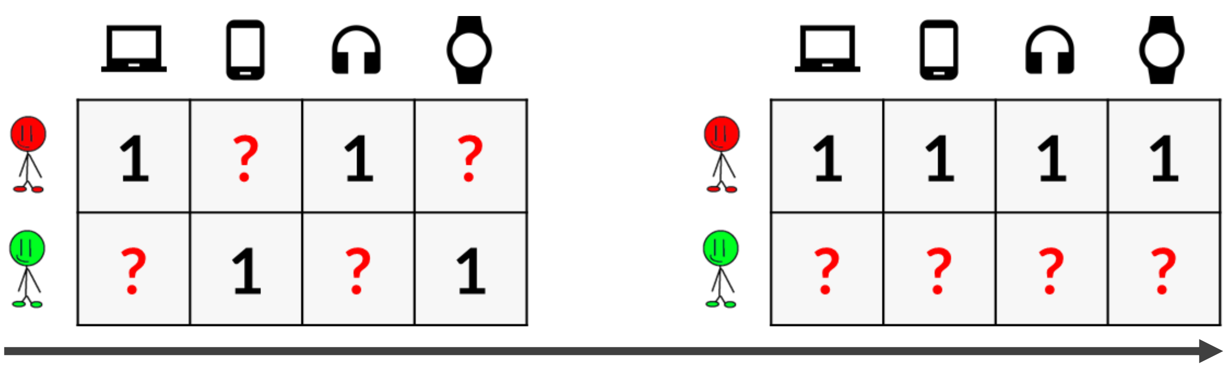
Distribution of Interactions over Users

- $|U|$: Number of Users
- $|K|$: Number of Interactions
- $|K_u|$: Number of Interactions for User u

▷ $Gini_{item} = 1 - 2\sum_{i=1}^{|I|}\left(\frac{|I|+1-i}{|I|+1}\right) \text{x} \left(\frac{|K_i|}{|K|}\right)$

Distribution of Interactions over Items

- $|I|$: Number of Items
- $|K|$: Number of Interactions
- $|K_i|$: Number of Interactions for Item i

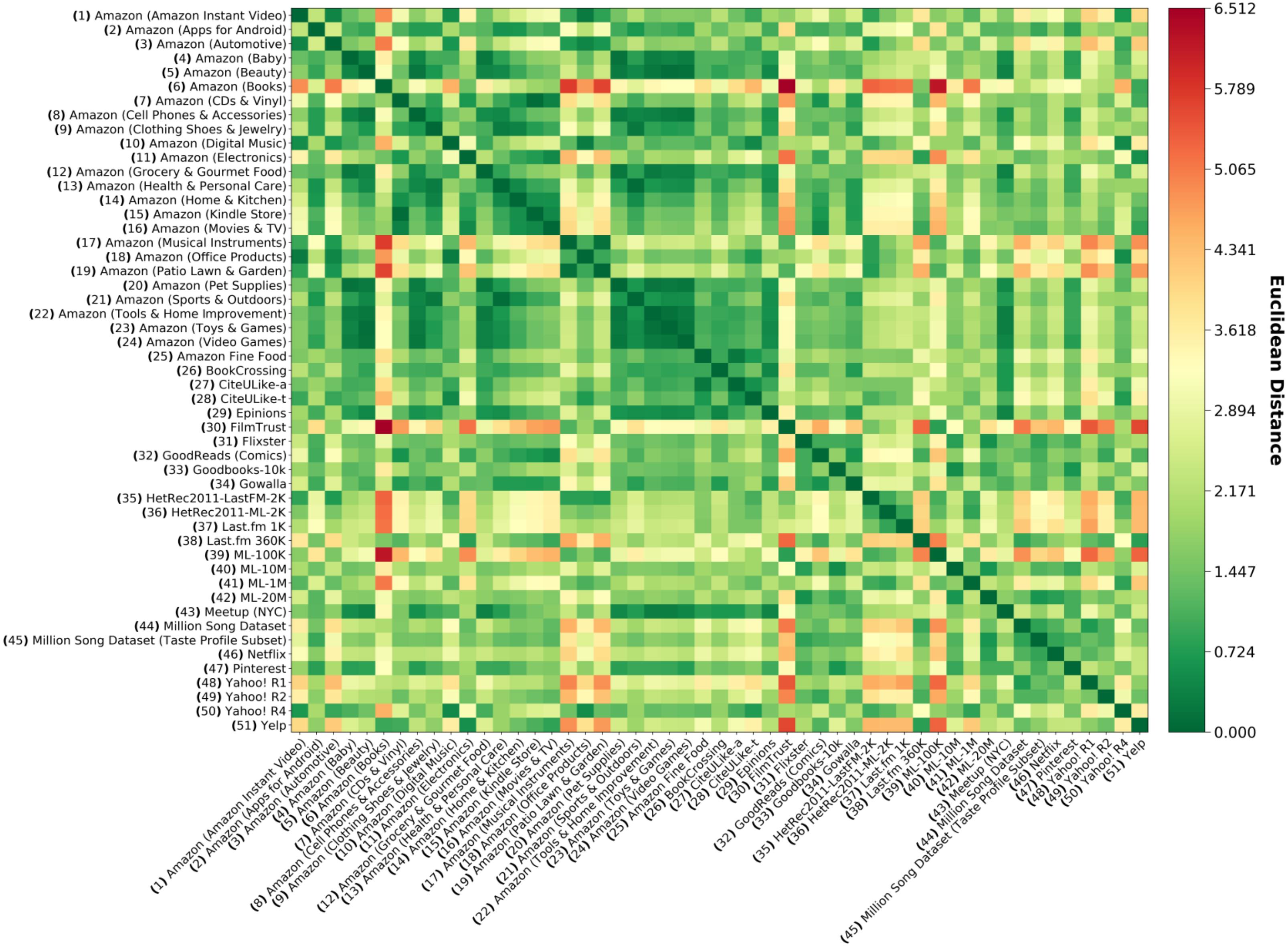


$Gini_{user} = \mathbf{0}$ $Gini_{user} = \mathbf{1}$

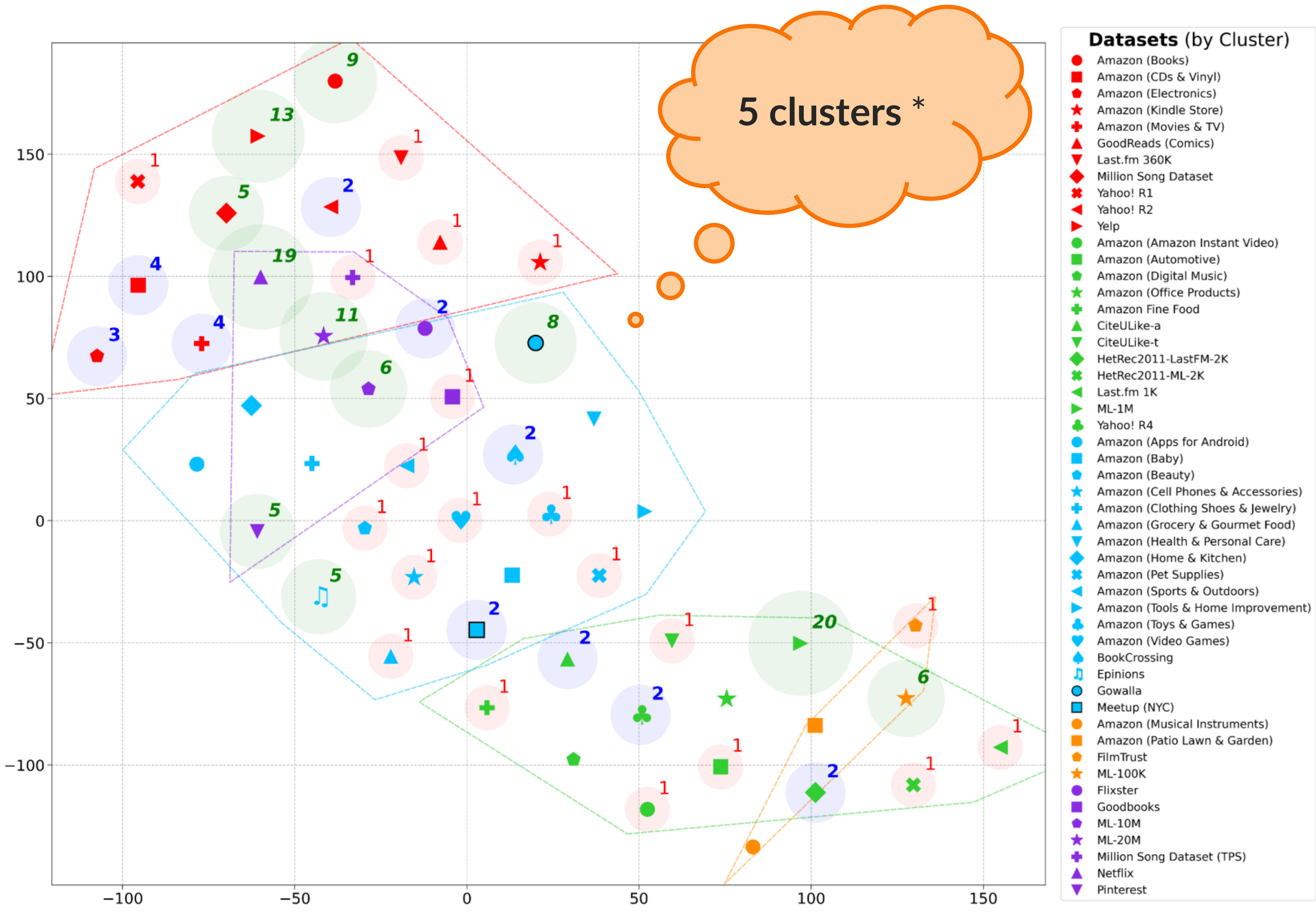
Datasets Used for Analysis & Experiments

- A total of **51** datasets
  - Excluded datasets which are too small after preprocessing
  - Included some missing Amazon datasets
- Preprocessing
  - Removed users/items with <5 interactions
  - For datasets with explicit feedback (i.e., ratings), convert all the observed entries into positive interactions
- Some publicly available datasets are in a *pre-processed form*
  - E.g., MovieLens datasets do not include users with <20 interactions

Similarities and Differences



Dataset Clusters



* Number of clusters chosen based on internal validation measures

Dataset Clusters – Centroids

Cluster 1: Gigantic but sparse
- Most number of users/items

Amazon (Books)
Million Song Dataset
Yelp

Cluster	$Space_{log}$	$Shape_{log}$	$Density_{log}$	$Gini_{user}$	$Gini_{item}$
1	7.274 (1)	0.497 (2)	-3.412 (5)	0.477 (2)	0.657 (2)
2	4.340 (4)	-0.134 (5)	-2.162 (3)	0.441 (3)	0.517 (4)
3	5.619 (3)	0.272 (3)	-3.106 (4)	0.337 (4)	0.504 (5)
4	3.167 (5)	0.116 (4)	-1.670 (1)	0.289 (5)	0.557 (3)
5	6.307 (2)	0.878 (1)	-2.120 (2)	0.502 (1)	0.767 (1)

Cluster 4: Tiny but dense

MovieLens-100K

Cluster 5: |Users| >> |Items|
- Highly concentrated

MovieLens-10M
MovieLens-20M
Netflix
Pinterest

The ‘Datasets Dilemma’

“How much do we really know about recommendation datasets?”

1. **How** are different datasets being utilised in recent papers?
   - The choice of datasets is often ***determined arbitrarily***
   - Difficult to compare results between different papers
2. **What** are the **similarities** as well as **differences** between various datasets?
   - Sparse vs Dense, Ratio of Users to Items, …
   - Datasets can be ***distinctively different*** from one another
3. **If** the choice of datasets used could **influence** the **observations** and/or **conclusions obtained**?
   - Empirical study using a variety of item recommendation algorithms

Experimental Setup

▷ **Sampling**

- Impractical to evaluate on all 51 datasets
- For each cluster, select the 3 datasets which are **closest** to the cluster centroid

▷ **Baseline methods**

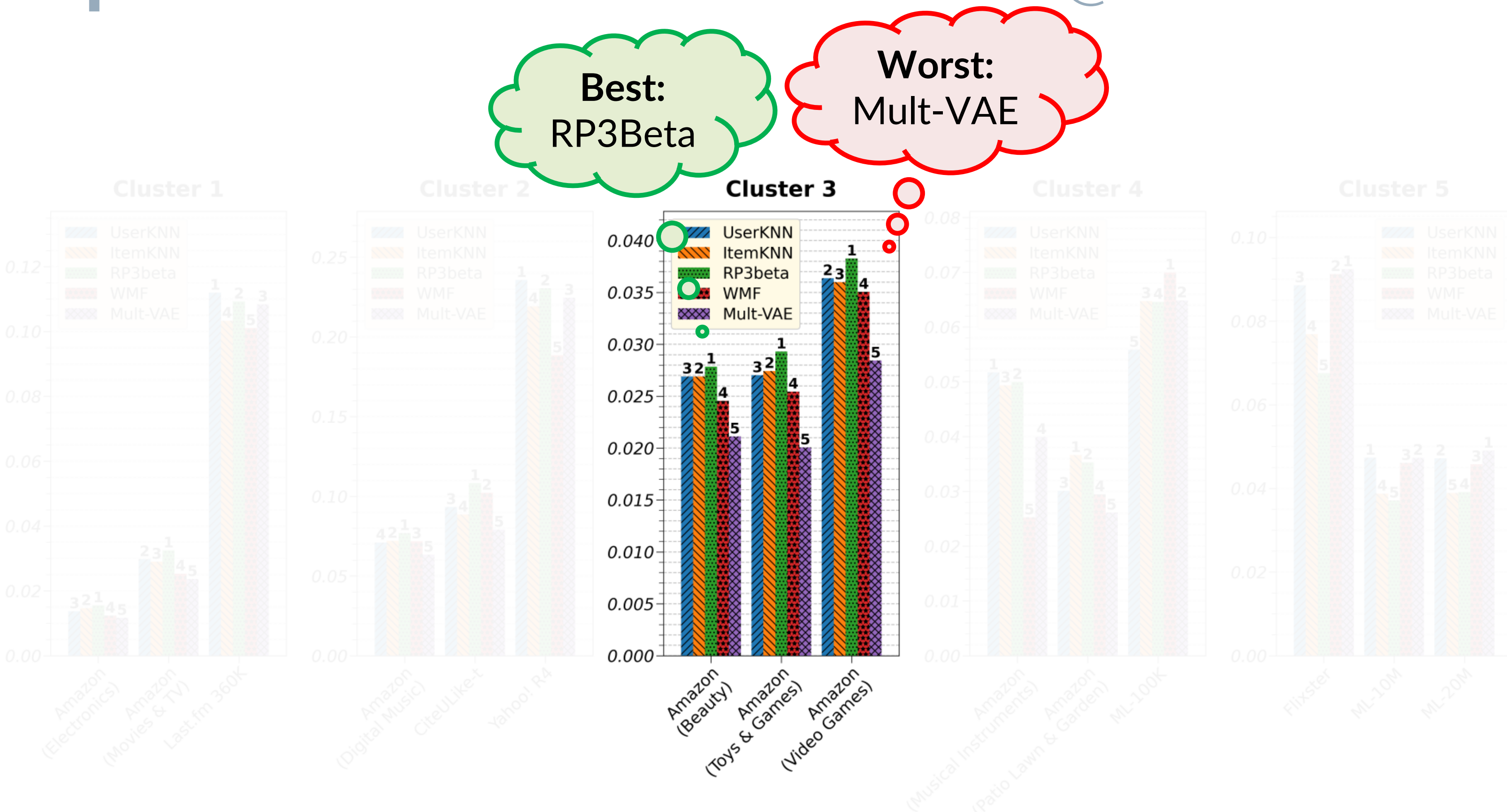
- **Neighbourhood-based**: UserKNN, ItemKNN
- **Graph-based**: RP3Beta
- **Latent Factor Model**: WMF
- **Generative Model:** Mult-VAE

✓ Distinct inductive bias
✓ Simple but effective

▷ **Evaluation metrics**

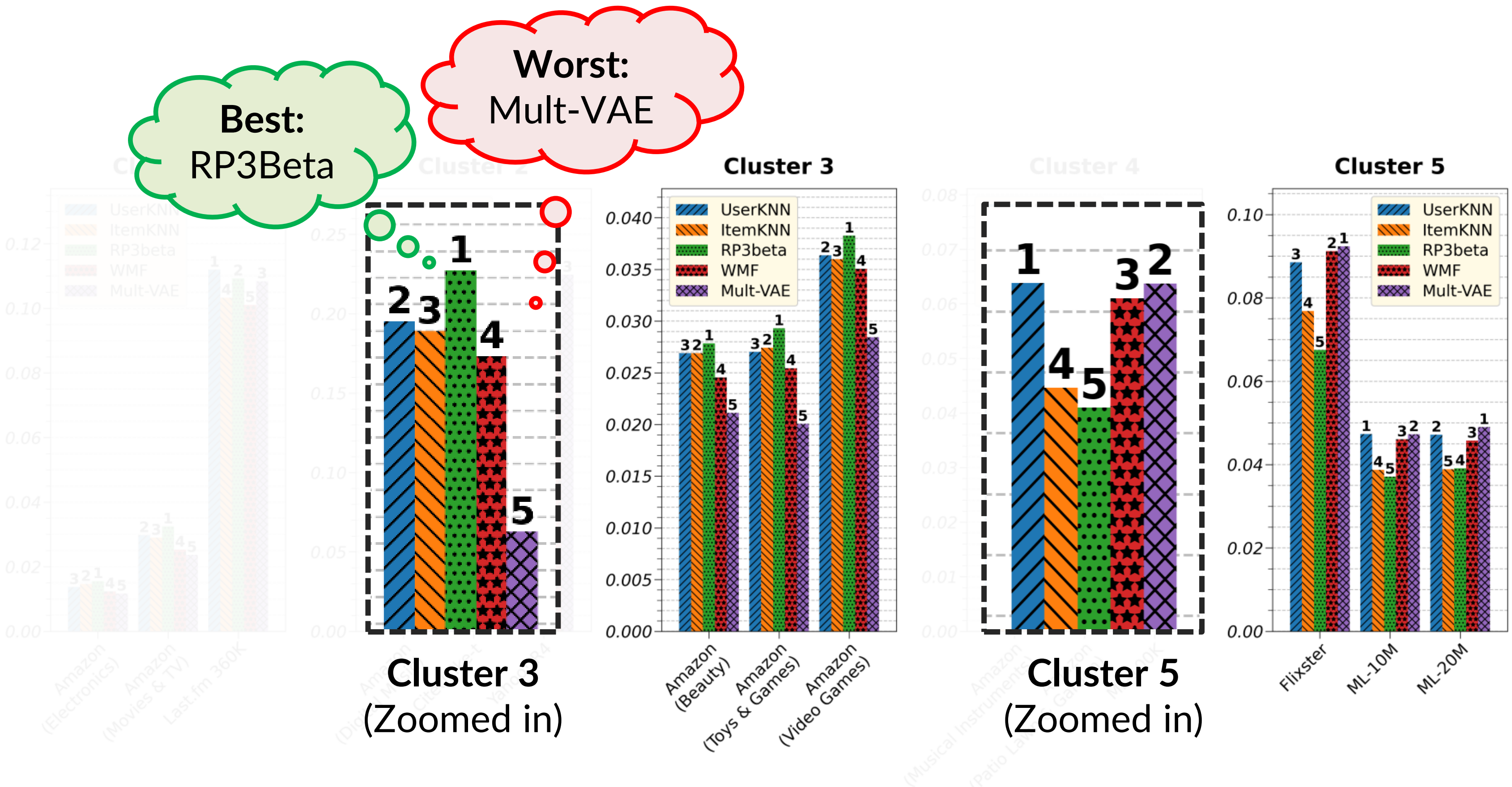
- Recall @ 10
- nDCG @ 10

Experimental Results (Recall @ 10)



- ▷ For datasets with similar characteristics, i.e. within the same cluster, some recommendation algorithm tends to perform significantly better (or worse) than the rest

Experimental Results (Recall @ 10)



▷ *'Ordering'* can *change drastically* based on dataset cluster

- **Cluster 3:** RP3Beta > UserKNN, ItemKNN > WMF > Mult-VAE
- **Cluster 5**: UserKNN, WMF, Mult-VAE >> ItemKNN, RP3Beta

The 'Datasets Dilemma'

"How much do we really know about recommendation datasets?"

1. How are different datasets being utilised in recent papers?
   - The choice of datasets is often ***determined arbitrarily***
   - Difficult to compare results between different papers
2. What are the **similarities** as well as **differences** between various datasets?
   - Sparse vs Dense, Ratio of Users to Items, …
   - Datasets can be ***distinctively different*** from one another
3. If the choice of datasets used could **influence** the **observations** and/or **conclusions obtained**?
   - Results can ***vary significantly*** based on the choice of datasets!
   - **Suggestion**: Utilising datasets with considerably different characteristics will improve robustness of evaluation procedure

Thanks!

Source Code:
https://github.com/almightyGOSU/TheDatasetsDilemma

Email:
S160005@e.ntu.edu.sg