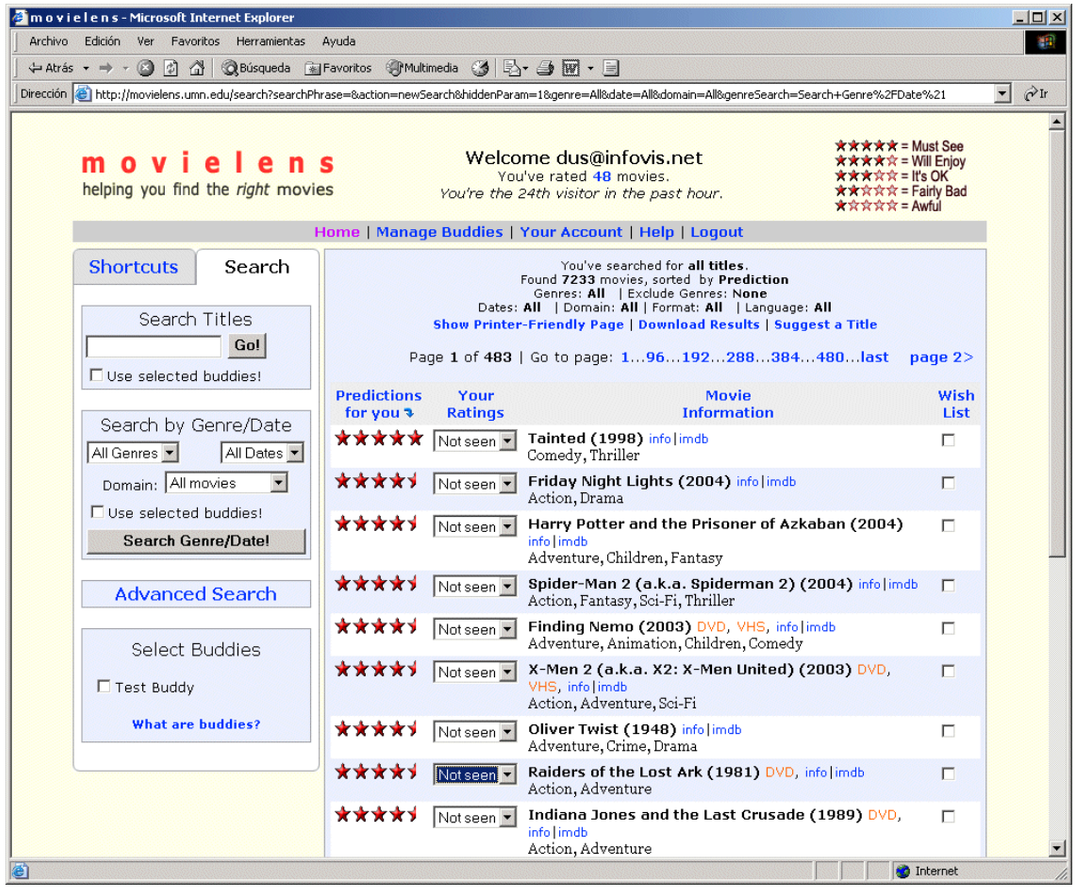
Evaluating Collaborative Filtering Recommender Systems

Jonathan L. Herlocker, Joseph A. Konstan, Loren g. Terveen, and John T. Riedel

Human Computer Interaction Group
École Polytechnique Fédérale de Lausanne (EPFL)
Rong Hu

Challenges



vs.

MovieLens

Netflix

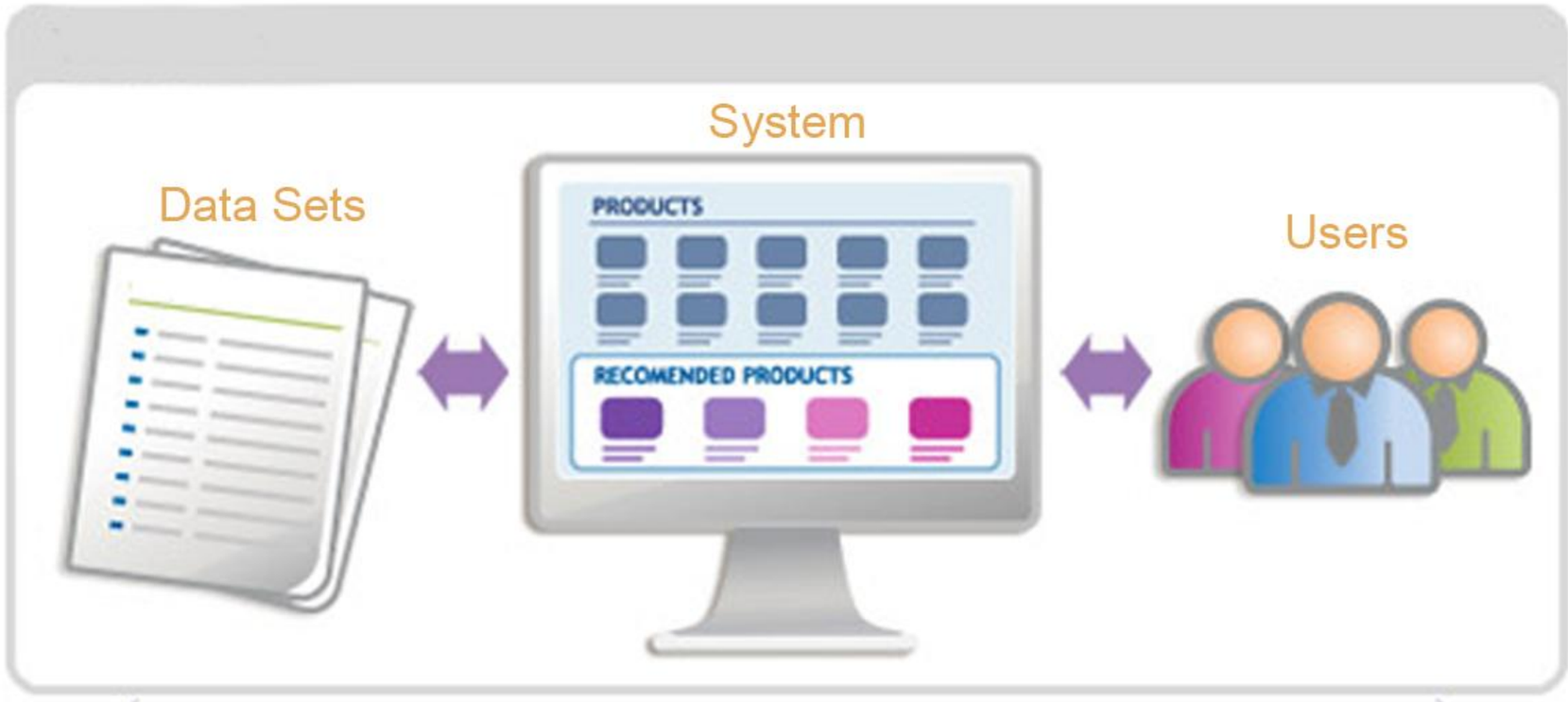
Challenges

- Evaluating recommender systems and their algorithms is **inherently difficult**
  - Different algorithms may be better or worse on different **data sets**
  - **Goals** for which an evaluation is performed may differ
  - Deciding what **combination** of measures to use in comparative evaluation is a significant challenge

Recommendation Algorithm Evaluation

Evaluation Considerations

- User Tasks
- Evaluation Methods
- Datasets
- Evaluation Metrics



User Tasks

- Annotation in Context
  - Filter through structure database to annotate messages in context using predictions
  - Examples:
    - Tapestry [Goldberg et al. 1992], GroupLens [Resnick et al. 1994]
- Find Good Items
  - Suggest specific items to users, and provide users with a ranked list of the recommended items
  - Examples:
    - Ringo [Shardanand and Maes 1995], Bellcore Video Recommender [Hill et al. 1995]

Suggested User Tasks

- Find All Good Items
  - Coverage becomes important in this task
- Recommend Sequence
  - The whole list should be pleasing as whole
- Just Browsing
  - Interface, ease of use, level and nature of information
- Find Credible Recommender
  - Useful recommendations that the user is sure to enjoy.

Selecting Data Sets

- Key Decisions
  - Live User Experiment vs. Offline Analyses
  - Synthesized vs. Natural Data Sets
  - Properties of Data Sets

Live User Experiments vs. Offline Analyses

- Offline Analyses
  - Algorithm Evaluation
  - Predict certain withheld values from a dataset
- Advantage
  - **Quick and economical** to conduct large evaluations
- Weaknesses
  - **Natural sparsity of ratings data sets** limits the set of items that can be evaluated
  - They are limited to objective evaluation of prediction results

Live User Experiments vs. Offline Analyses

- Live User Experiments
  - Controlled experiments
  - Field studies
- Advantage
  - Evaluate user performance, satisfaction, participation, et
- Disadvantage
  - Expensive
  - The bias of users' responses

Synthesized vs. Natural Data Sets

- Natural data sets
  - May imperfectly match the properties of the target domain and task
- Synthesized data sets
  - Specifically to match those properties
  - Appropriate for new domain
  - Unfair to other algorithms
    - It fits their approach too well
    - Drawing comparative conclusions is risky

Properties of Data Sets

- What properties should the dataset have in order to best model the tasks for which the recommender is being evaluated?
- Data Set Properties
  - Domain Features
    - e.g., user task, content domain
  - Inherent Features
    - e.g., implicit/explicit ratings, rating scales, item content attributes, demographic information
  - Sample Features
    - e.g., density, size, distribution

Accuracy Metrics

- Predictive Accuracy Metrics
- Classification Accuracy Metrics
- Rank Accuracy Metrics

Predictive Accuracy Metrics

- Measure how close the recommender system's predicted ratings are to the true user ratings

- Particular important for evaluating tasks where the predicting rating will be displayed to the user, e.g., *Annotation in Context*

Predictive Accuracy Metrics

- Mean Absolute Error (MAE)

$$|\overline{E}| = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N}$$

- Advantages
  - Simple and easy to understand
  - Has well studied statistical properties
- Weakness
  - Less appropriate for the tasks returning a rank list, e.g., *Find Good Items*
  - Less appropriate when the granularity of true preference is small

Predictive Accuracy Metrics

- Mean Squared Error
  - Emphasis on large errors

- Root Mean Squared Error
  - Emphasis on large errors

- Normalized mean absolute error
  - Allow comparison between prediction runs on different datasets.

Classification Accuracy Metrics

- Measure the frequency with which a recommender system makes correct or incorrect decisions about whether an item is good
- Appropriate for tasks when users have true binary preferences, such *Find Good Items*

Classification Accuracy Metrics

- Precision and Recall

	Selected	Not Selected	Total
Relevant	N_{rs}	N_{rn}	N_r
Irrelevant	N_{is}	N_{in}	N_i
Total	N_s	N_n	N

- **Precision**: the probability that a selected item is relevant

$$P = \frac{N_{rs}}{N_s}$$

- **Recall**: the probability that a relevant items will be selected

$$R = \frac{N_{rs}}{N_r}$$

Classification Accuracy Metrics

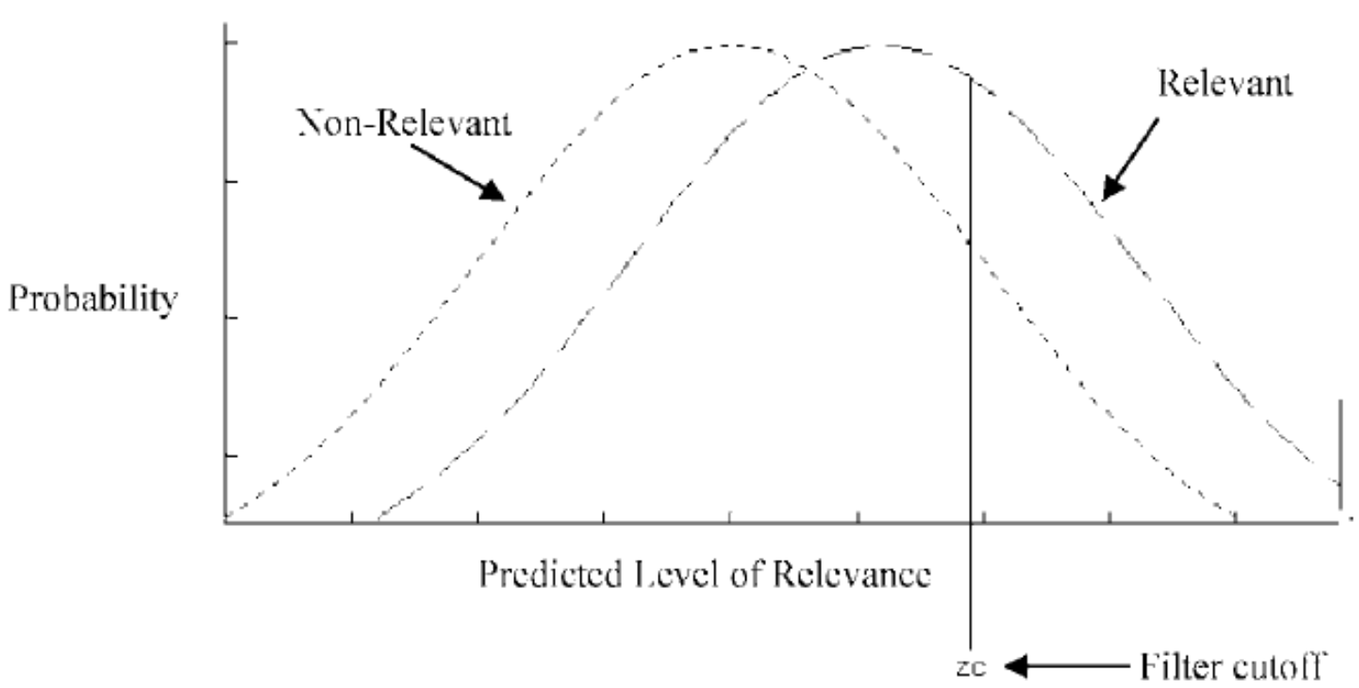
- Precision and Recall
  - Advantages:
    - More comprehensive
  - Disadvantages:
    - Need to convert multi-scale ratings to binary scale
    - Relevance is more inherently subjective in RS
    - Recall and Precision should be considered together

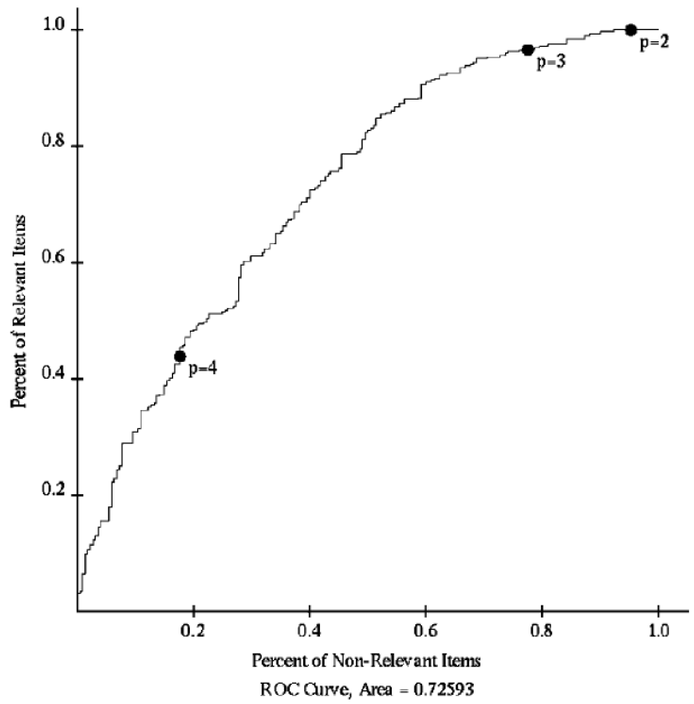
- **F1** metric combines precision and recall

$$F_1 = \frac{2PR}{P + R}$$

Classification Accuracy Metrics

- ROC Curves
  - ROC: Relative Operating Characteristic / Receiver Operating Characteristic





Classification Accuracy Metrics

- ROC Curves
  - Advantages
    - Developed from solid statistical decision theory
    - A single performance measure
    - Covers different recommendation lengths.
  - Weakness
    - Ordering among relevant items has no impact.
    - A large set of potentially relevant items is needed
    - Users are only interested in one setting
    - Need a large number of data points to ensure good statistical power

Rank Accuracy Metrics

- Measure the ability of a recommendation algorithm to produce a recommended ordering of items that matches how the user would have ordered the same items

- Appropriate for ranked list and in domains where users' preferences are non-binary

Rank Accuracy Metrics

- Prediction-Rating Correlation
  - Pearson, Spearman's ρ, Kendall's Tau
  - Advantages
    - Well understood by the scientific community
    - Provide a single measurement score
  - Disadvantages
    - Interchange weakness

Rank Accuracy Metrics

- Half-life Utility Metric
  - Evaluate the utility of a ranked list to the user
  - Best for the tasks domain where there is an exponential drop in true utility as the search length increase.

$$R_a = \sum_j \frac{\max(r_{a,j} - d, 0)}{2^{(j-1)/(\alpha-1)}}$$

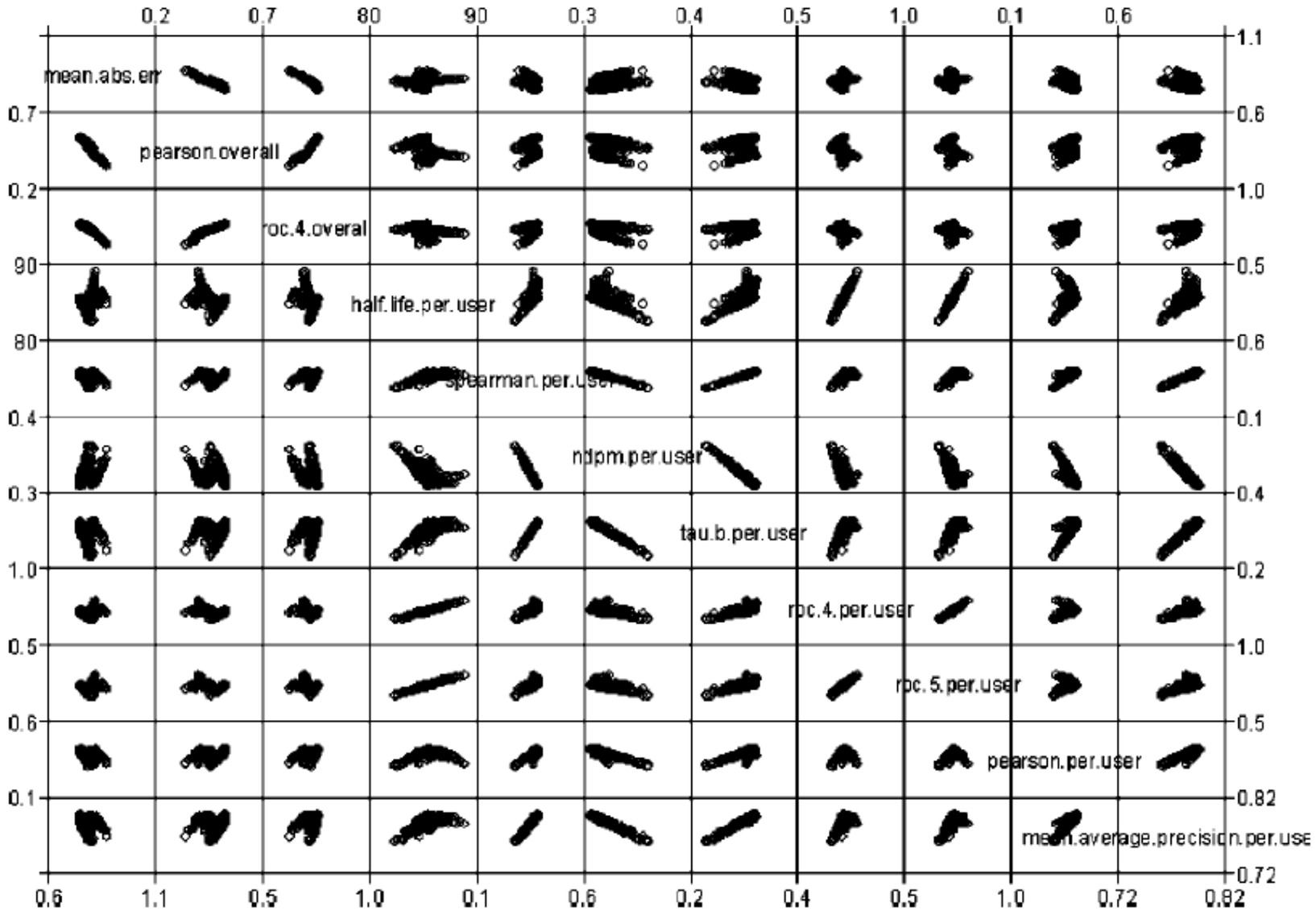
- NDPM Measure
  - Normalized Distance-based Performance Measure
  - Comparable among different datasets

$$NDPM = \frac{2C^- + C^{\cdot\cdot}}{2C^u}.$$

An Empirical Comparison of Evaluation Metrics

- Examine the extent to which the different evaluation metrics agreed or disagreed.
  - 432 different variants of the algorithm tested (Size of neighborhood, similarity, threshold, type of recommendation)
  - Examined evaluation metrics: MAE, Pearson, Spearman, ROC-4, ROC-5, half-life utility metric, mean average precision, NDPM metric
    - Per user/Overall
  - Dataset: MovieLens Dataset (100,000 movie ratings from 943 users on 1682 items)

An Empirical Comparison of Evaluation Metrics



- Coverage
  - Measures the percentage of a datasets that the recommender system is able to provide predictions
- Learning Rate
  - Measure how quickly an algorithm can produce good recommendations
- Novelty and Serendipity
  - Measure whether a recommendation is a novel possibility
  - Example: Bananas in a grocery store

- Confidence
  - Measure how sure is the recommender system that its recommendation is accurate


X]*$5@ (2000)

Personalized Prediction : ★★★★ (will enjoy it)

Your Neighbors' Ratings for this Movie

Rating	Number of Neighbors
★	1
★★	2
★★★	7
★★★★	14
★★★★★	9

([Herlocker et al. 2000])

- User Evaluation
  - How to evaluate user reaction to a recommender system

Conclusion

- Effective and meaningful evaluation of recommender systems is challenging.

- Choose appropriate evaluation **datasets**, **methods** and **metrics**, taking into account **user tasks** and **contexts**.

- Accuracy alone is NOT enough.

Thanks!

Discussion