QUERY RECOMMENDATIONS FOR INTERACTIVE DATABASE EXPLORATION

Gloria Chatzopoulou*, UC Riverside
Magdalini Eirinaki, San Jose State Univ
Neoklis Polyzotis, UC Santa Cruz

*work performed while author affiliated with UC Santa Cruz

SSDBM ’09 New Orleans, Louisiana

Motivation

- Scientific disciplines use relational DBMS for storage and retrieval of information
  - Biologists (e.g. UCSC Genome, BMRB)
  - Astronomers (e.g. Skyserver)
  - Chemists (e.g. PubChem)
- DBs are accessible online by users with diverse information needs
- Typical users do interactive exploration

Motivation (cont'd)

- Typical users are not SQL experts
- Scientific datasets increase in size
- Users may miss interesting information
  - They do not write the "right" query
  - They are not aware of all parts of the database

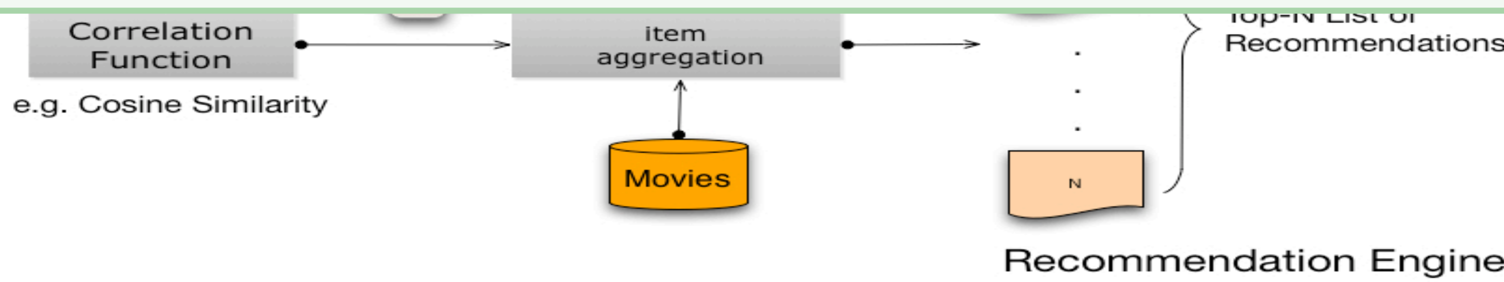
Our goal: Assist users in finding useful information

Web Collaborative Filtering

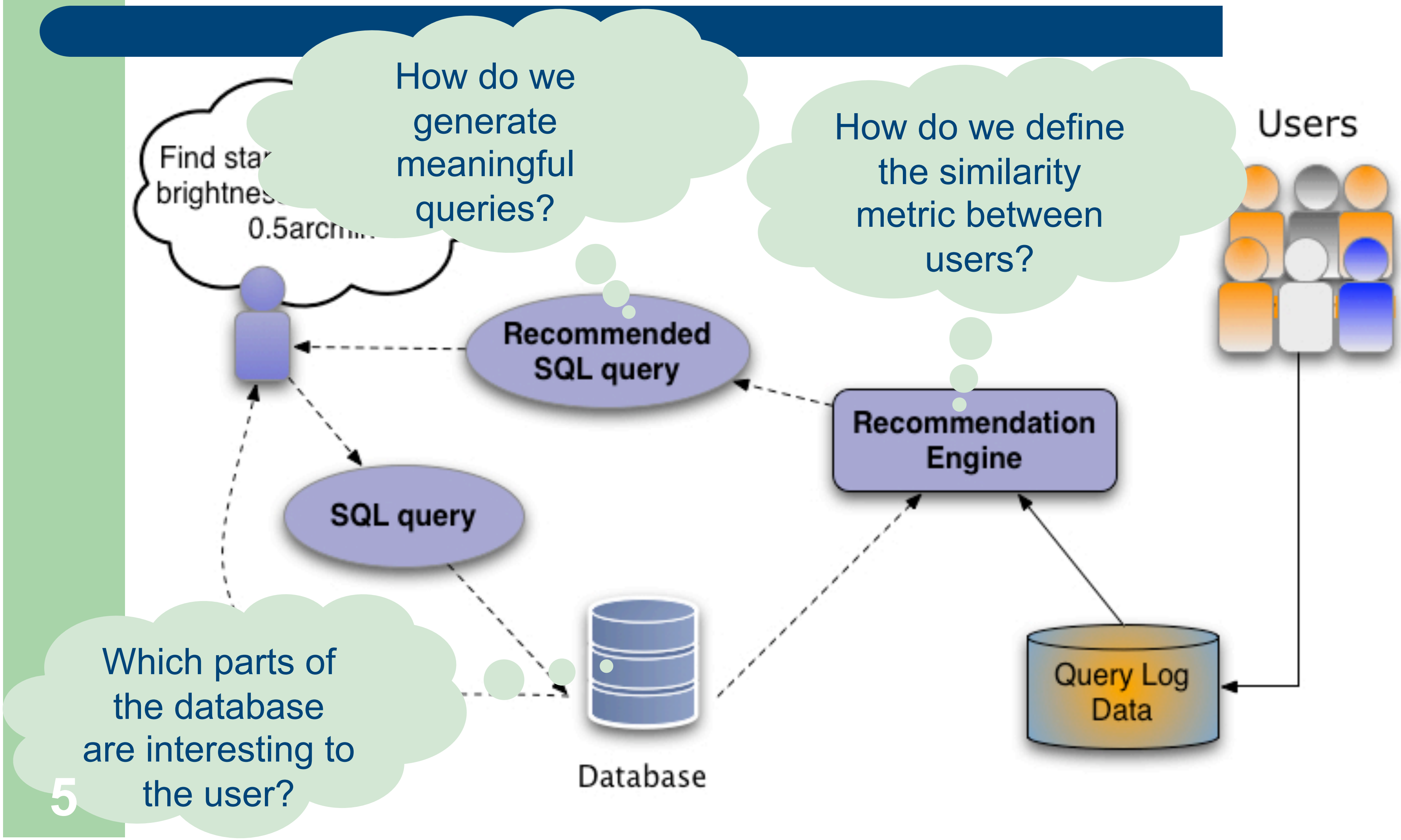
<u>Example:</u> Movie Recommendations

If Alice and Bob **both** like movie X and Alice likes movie Y
then
Bob is likely to be interested in seeing movie Y

If Alice and Bob **both** query data X and Alice queries data Y
then
Bob is likely to be interested in querying data Y



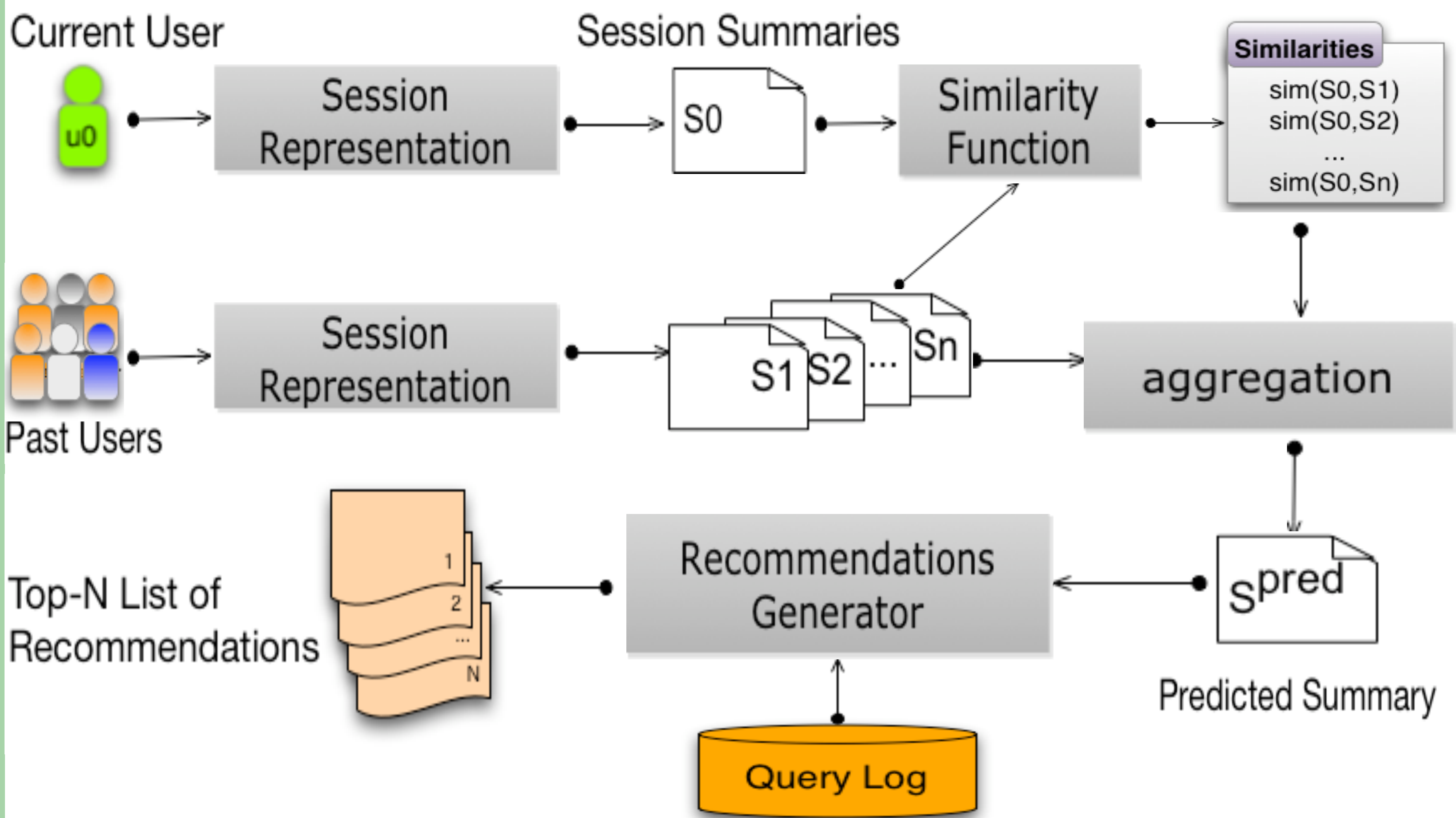
System Architecture



Roadmap

- Introduction
- QueRIE Recommendation Framework
- Experiments
- Conclusions

Conceptual Framework



Session Summaries

R	a	b
	y	3
	s	4
	w	3
	r	2

L	a	c
	y	9
	s	3
	s	5
	t	8

q1: $R \bowtie_{R.a=L.a} L$

q2: $\sigma_{R.b=4} (F \bowtie_{R.a=L.a} L)$

Binary Weighting Scheme
q1 = <1,1,0,0,1,1,1,0>
q2 = <0,1,0,0,0,1,1,0>
s0 = <1,2,0,0,1,2,2,0>2

Result Weighting Scheme
q1 = <0.33,0.33,0,0,0.33,0.33,0.33,0>
q2 = <0,0.50,0,0,0,0.50,0.50,0>
s0 = <0.33,0.83,0,0,0.33,0.83,0.83,0>

Similarity Function

- Vector-space similarity functions can be used
  - Cosine Similarity

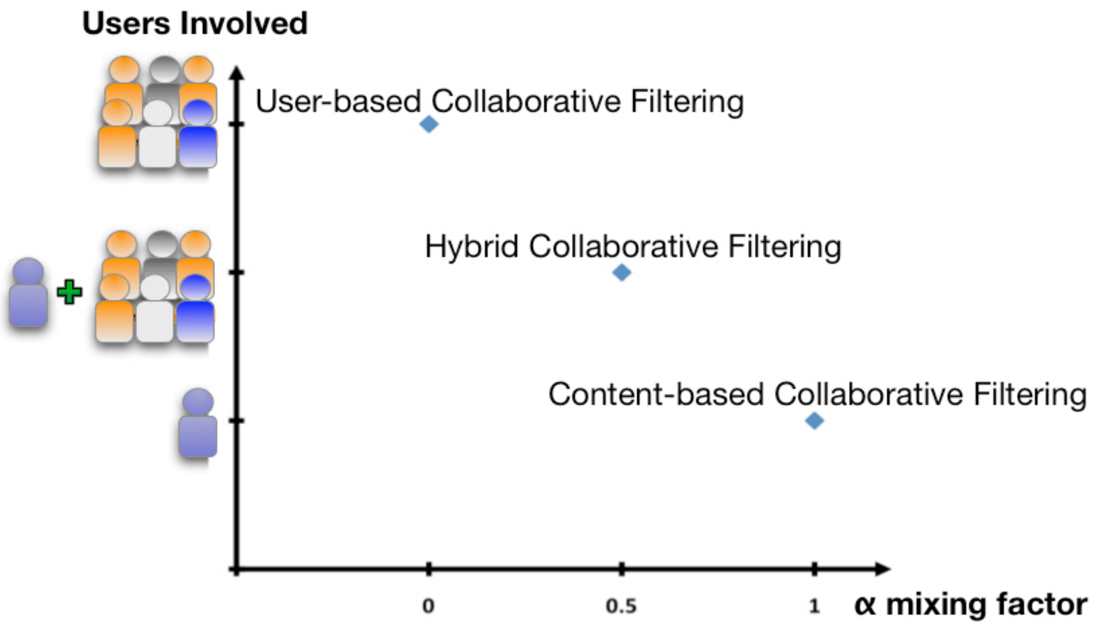
$$sim(uA, uB) = \frac{uA \bullet uB}{\|uA\| * \|uB\|}$$

- High similarity means that users are interested in the same parts of the database

Predicted Summary

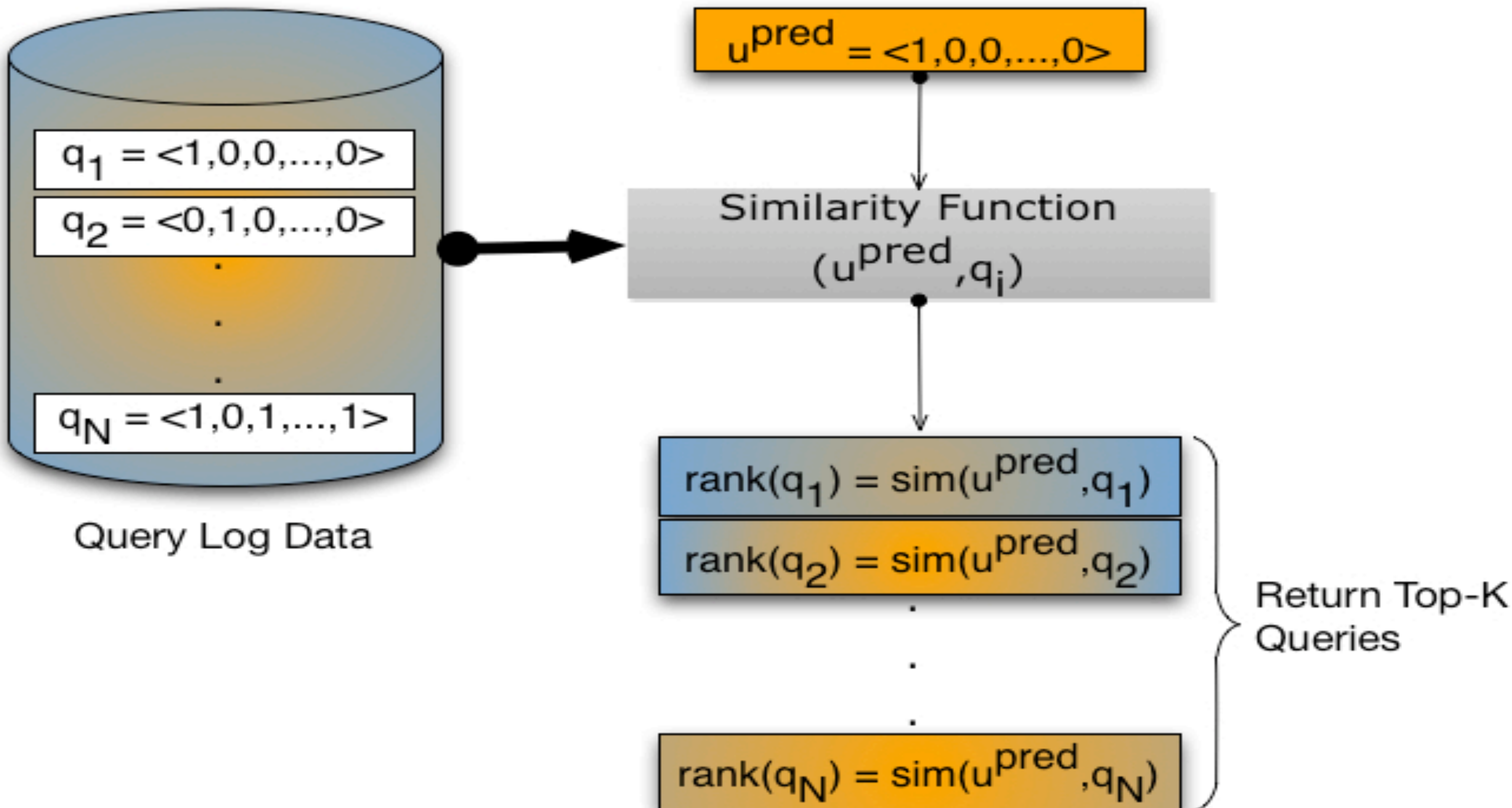
$$u^{pred} = \alpha * u + (1-\alpha) * \frac{\sum_{1 \le i \le h} sim(u,u_i) * u_i}{\sum_{1 \le i \le h} sim(u,u_i)}$$

where α is the “mixing factor” $\alpha \in$ [0,1]



Generating Recommendations

- Use queries of past users



Roadmap

- Introduction
- QueRIE Recommendation Framework
- Experiments
- Conclusions

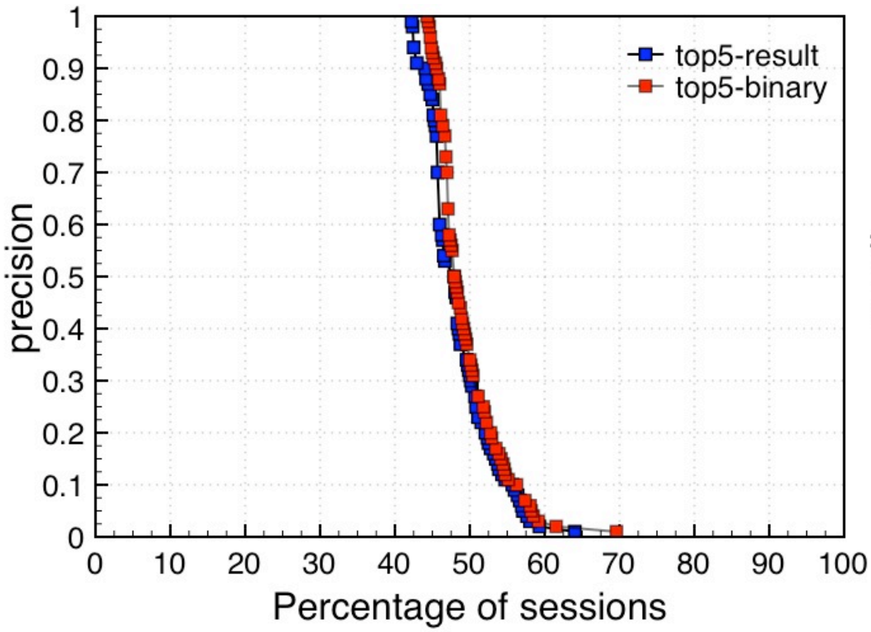
Experimental Setup

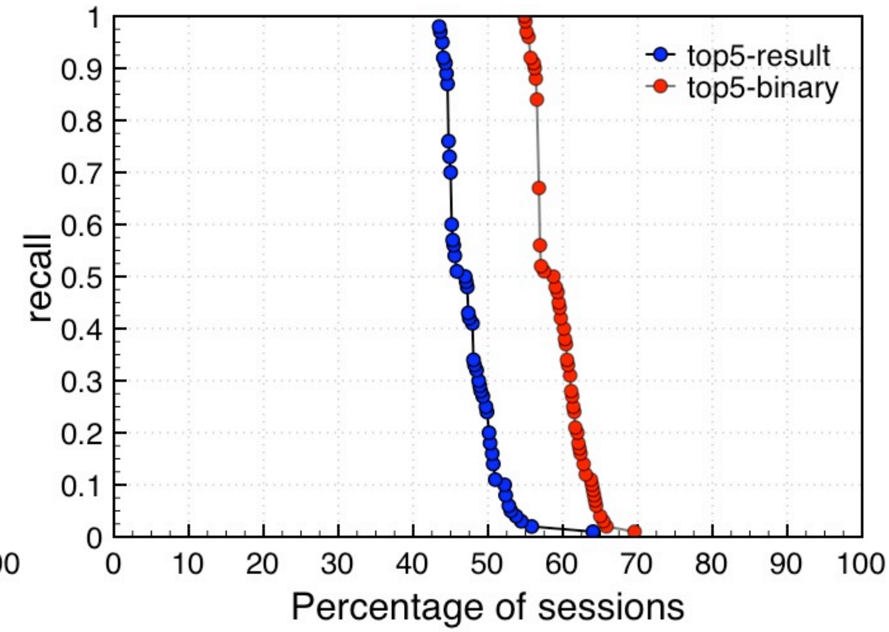
- SkyServer Dataset

Database Size	2.6TB
#Sessions	720
#Queries	6713
#Distinct Queries	4037
Avg. number of queries per session	9.3
Min. number of queries per session	3

- Evaluation Metrics: Precision and Recall
  - **High precision:** most witnesses of the recommended query are witnesses in the actual query.
  - **High Recall:** most witnesses of the actual query are witnesses in the recommended query.

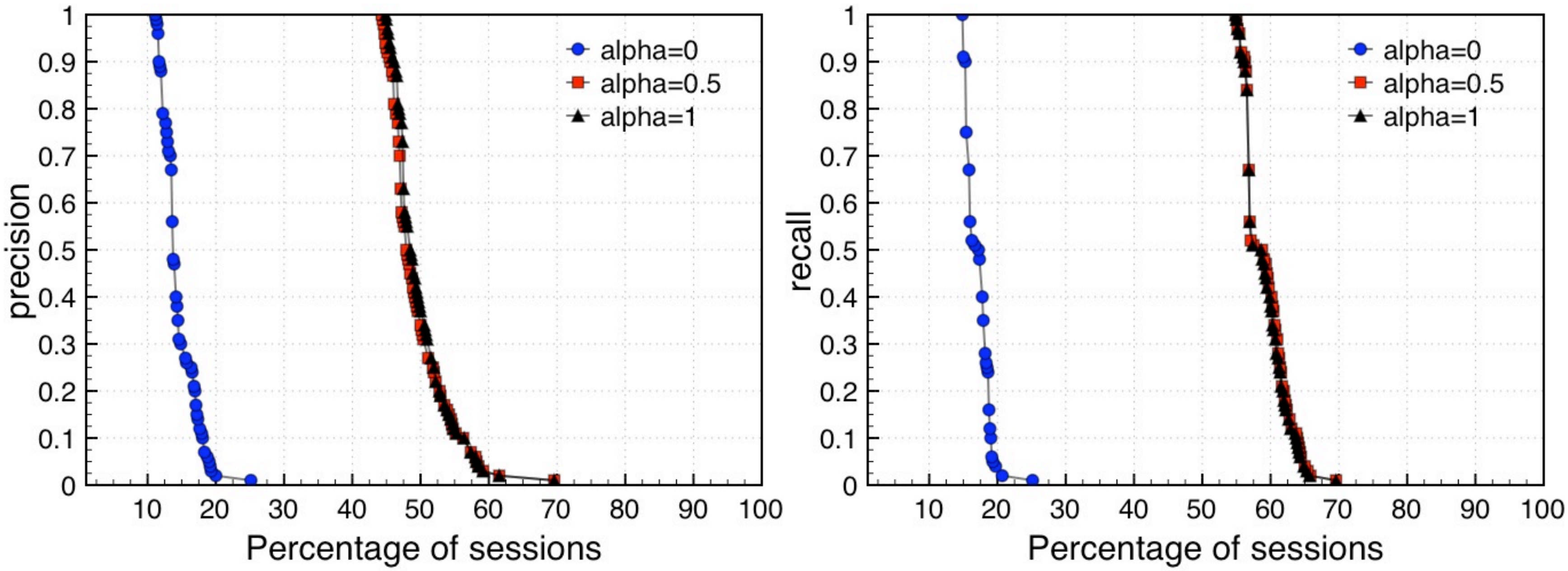
Binary vs Result Weighting Schemes





Binary outperforms Result Weighting Scheme

Effect of mixing factor α



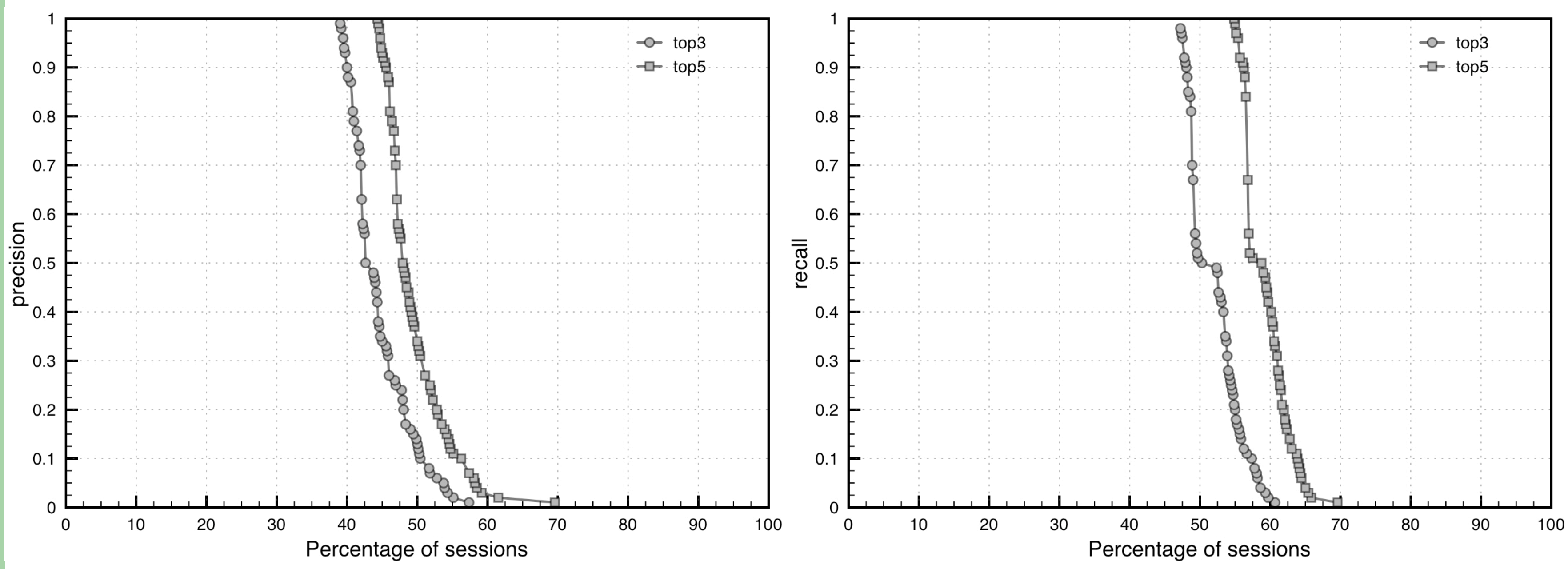
Hybrid Collaborative Filtering yields better results

Conclusions

- Scientists need help in exploring databases
- Query recommendations can be an effective tool in guiding exploration
- Collaborative filtering provides a natural method to generate recommendations
- Experiments show promising results on real-world datasets
- Ongoing Work:
  - Performance improvement
  - Use of approximation techniques

Thank you

Top-3 vs Top-5 Binary Weights



The bigger recommendation set the higher accuracy