

Supporting Ranking and Clustering as Generalized Order-By and Group-By

Chengkai Li (UIUC)

joint work with

Min Wang

Lipyeow Lim

Haixun Wang (IBM)

Kevin Chang (UIUC)



DAIS
The Database and Information Systems Laboratory
at The University of Illinois at Urbana-Champaign
Large Scale Information Management



Boolean database queries

PARTS		SUPPLIERS		PRICES		
PARTNO	NAME	SUPPNO	NAME	PARTNO	SUPPNO	PRICE
P107	Bolt	S51	Acme	P107	S51	.59
P113	Nut	S57	Ajax	P107	S57	.65
P125	Screw	S63	Amco	P113	S51	.25
P132	Gear			P113	S63	.21
				P125	S63	.15
				P132	S57	5.25
				P132	S63	10.00

Fig. 1(b). A Relational Database.

Example 1:

```
SELECT SUPPNO, PRICE
FROM QUOTES
WHERE PARTNO = '010002'
AND MINQ< = 1000 AND MAXQ> = 1000;
```

(Relational Algebra and System R papers)

Data Retrieval



A screenshot of a Flickr search result. The search query is "21st Century Document Manag". The top navigation bar includes "Home", "Learn More", "Sign Up!", and "Explore". The search results show a large image of a pile of shredded paper. The Flickr logo and navigation are visible at the top of the page.



A screenshot of a Flickr search result. The search query is "SPAM". The top navigation bar includes "Home", "Learn More", "Sign Up!", and "Explore". The search results show a large image of several cans of SPAM, including "Classic" and "Crazy Tasty" varieties. The Flickr logo and navigation are visible at the top of the page.

Example: What Boolean queries provide

```
SELECT      *
FROM        Houses H
WHERE       200K<price<400K AND #bedroom = 4
```

	query semantics	<i>results organization</i>
Boolean query	hard constraints (<i>True or False</i>)	<input type="checkbox"/> <i>a flat table</i> <input type="checkbox"/> <i>too many (few) answers</i>

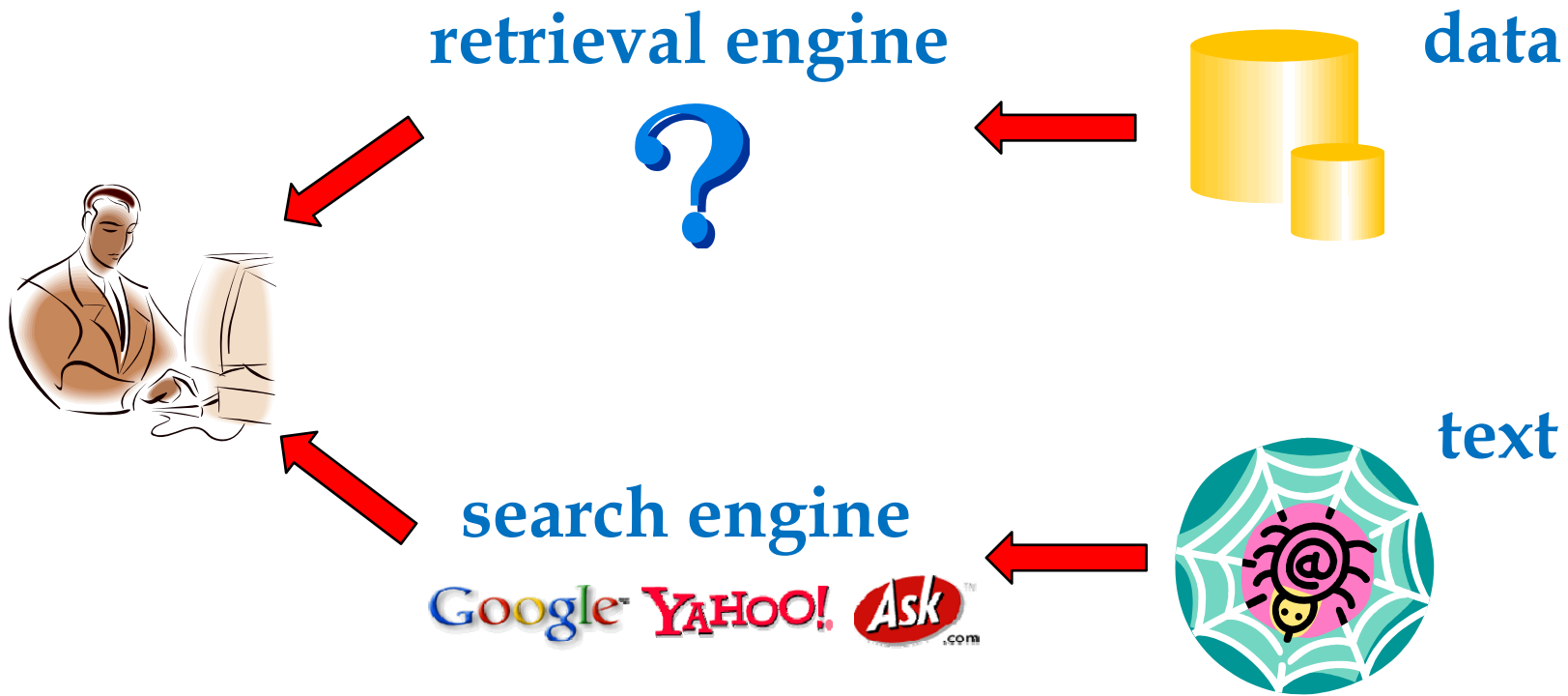
Example: What may be desirable

- < 500K is more acceptable
- but willing to pay more for big house
- close to the lake is a plus
- avoid locations near airport
- ...



	query semantics	results organization
Boolean query	hard constraints (<i>True or False</i>)	<input type="checkbox"/> <i>a flat table</i> <input type="checkbox"/> <i>too many (few) answers</i>
fuzzy retrieval	“soft” constraints (preference, similarity, relevance,...)	<input type="checkbox"/> <i>a ranked list</i> <input type="checkbox"/> <i>a grouping of results</i> <input type="checkbox"/> <i>etc.</i>

“Retrieval” of DATA: From Boolean query to fuzzy retrieval



Retrieval mechanisms: Learning from Web search

The image shows a screenshot of the Vivisimo search engine interface. The search bar contains the text "database" and "the Web". The search results are displayed in a "Clustering" view, showing a "Navigation Map" of related websites. A "Facets" panel is visible on the left, listing categories like "Database Journal", "SQL", and "Oracle". A "Categorization" panel is also present, showing terms like "popular", "term", "industry", and "oracle".

Ranking

Clustering

Navigation_Map

Facets

Categorization

Generalizing SQL constructs for data retrieval



NEW search [auctions](#) at [Clusty.com](#)

Top 211 results of at least 196,730,000 retrieved for the query **database**

Clustered Results

- database (215)
 - Software (38)
 - Images (20)
 - Downloads (16)
 - Management (17)
 - SQL (14)
 - Database management (5)
 - SQL database engine (3)
 - Sql92 (2)
 - Other Topics (4)
 - Oracle (14)
 - Statistics, Data (11)
 - Research (9)

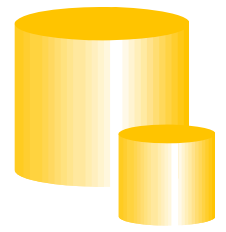
Database [new window]
Create Your **database** in less than 3 min. Get started now.
[www.WebOffice.com](#) - Sponsored Listings 1

Vista & Office 2007 Books [new window]
Essential guides from Wiley The right resource just for you
[www.bookpool.com](#) - Sponsored Listings 2

- The Internet Movie Database (IMDb)** [new window] [frame] [cache]
Features plot summaries, reviews, cast lists, and theatre schedules
[www.imdb.com](#) - Wisenut 1, MSN 3, Open Directory 4, Wisenut 9
- MySQL AB :: The world's most popular open source database** [new window] [frame] [review] [clusters]
The software's official homepage with news, downloads and documentation
world's most popular open source **database** ...
[www.mysql.com](#) - Wisenut 2, Ask 2, Open Directory 21
- The Internet Movie Database (IMDb)** [new window] [frame] [pre]
IMDb: The biggest, best, most award-winning movie site on the planet

Order-By

Group-By



From *crispy ordering* to *fuzzy ranking*

■ Crispy ordering

Order By

Houses.size, Houses.price

- by attribute values
- equality of values
- order \neq desirability

■ Fuzzy ranking

Order By

*Houses.size - 4*Houses.price*

Limit

5

- by ranking function
- combine matching criteria
- order \Rightarrow desirability : top-k

From crispy grouping to fuzzy clustering

■ Crispy grouping

Group By

Houses.size, Houses.price

- by attribute values
- equality partition
- no limit on output size

■ Fuzzy clustering

Group By

k-means(H.size, H.price)

Into

5

- by distance function
- proximity of values
- number of clusters

Need for combining ranking with clustering

- **Clustering-only**

- **A group can be big**

- “too many answers” problem persists

- **How to compare things within each group?**

- **Ranking-only**

- **Lack of global view**

- top-k results may come from same underlying group
(e.g., cheap and big houses come from a less nice area.)

- **Different groups may not be comparable**

Contributions

- Concepts

- generalize Group-By to fuzzy clustering, parallel to the generalization from Order-By to ranking

- integrate ranking with clustering in database queries

- Efficient processing framework

- summary-based approach

Related works

■ Clustering

- ❑ not on dynamic query results
- ❑ use summary (grid with buckets)

(e.g., STING [WangYM97] WaveCluster [SheikholeslamiCZ98])

■ Ranking (top-k) in DB: many instances (e.g., [LiCIS05])

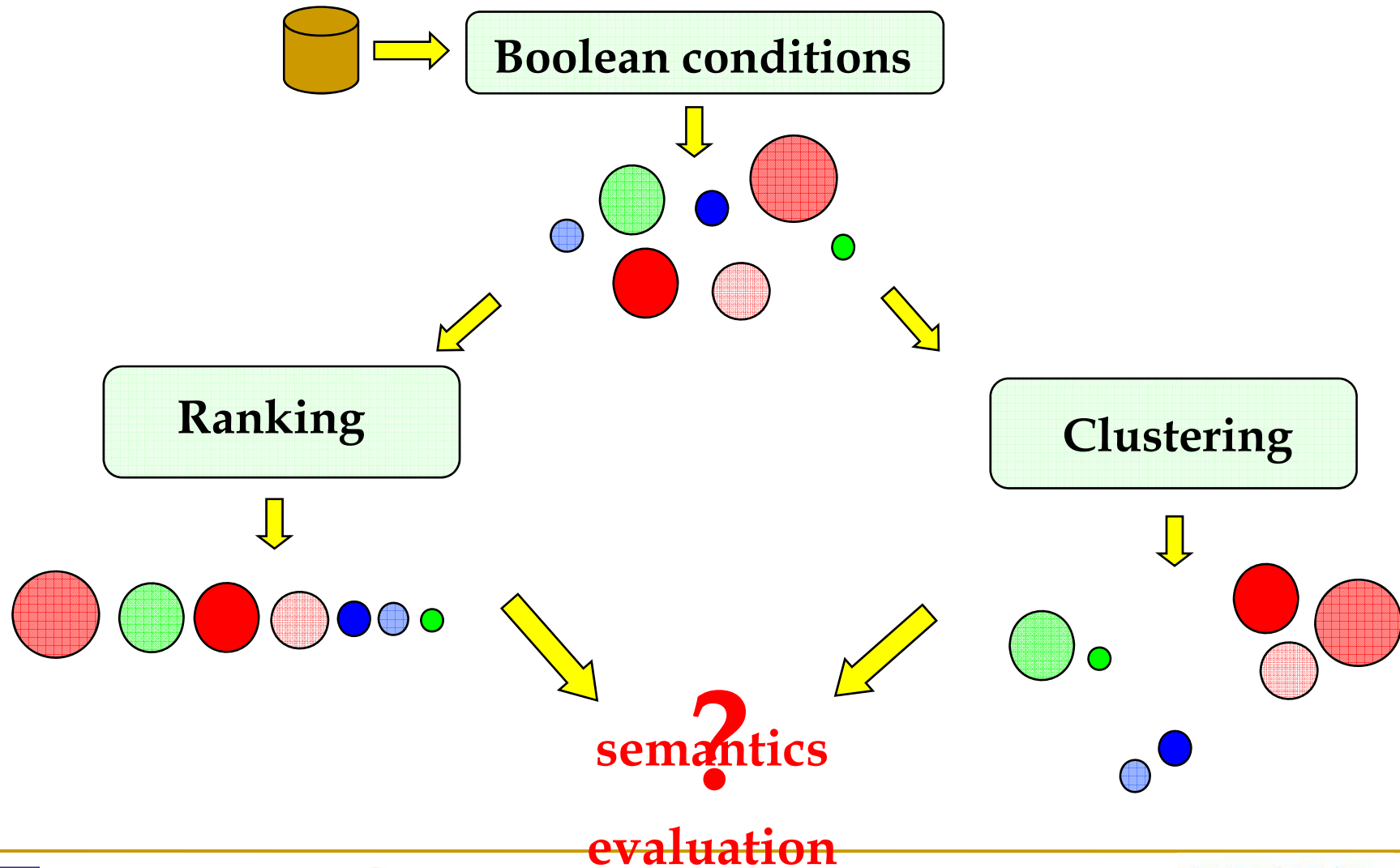
- ❑ use summary (histogram) in top-k to range query translation [ChaudhuriG99]

■ Categorization of query results [ChakrabartiCH04]

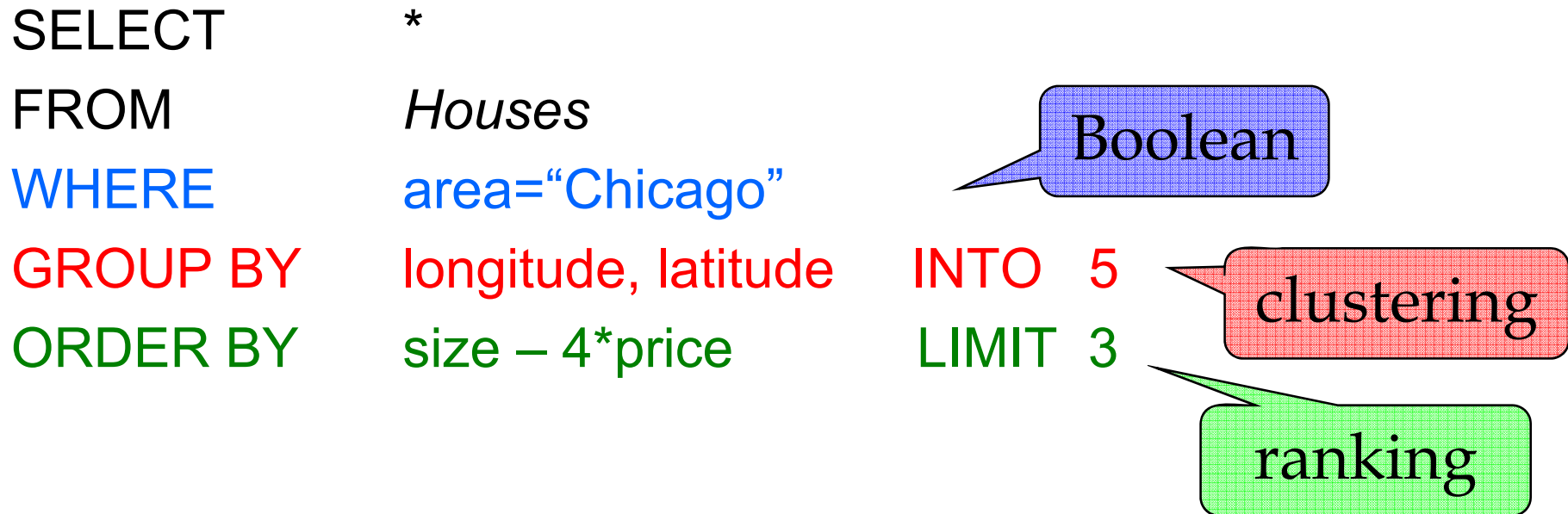
- ❑ different from clustering
- ❑ no integration with ranking
- ❑ focus on reducing navigation overhead, not processing

■ Web search and IR (e.g., [ZamirE99] [LeuskiA00])

Integrate the two generalizations



Query semantics: ClusterRank queries



Semantics: *order-within-groups*

Return the top k tuples within each group (cluster).

Several notes

- **Non-deterministic semantics**

- clustering is non-deterministic by nature
- sacrificing the crispiness of SQL queries
- worthy for exploring the fuzziness in data retrieval?

- **Language syntax isn't our focus**

- current SQL semantics: *order-among-groups*
Select... From... Where... Group By... Order By...(RankAgg[LiCI06])
- OLAP function

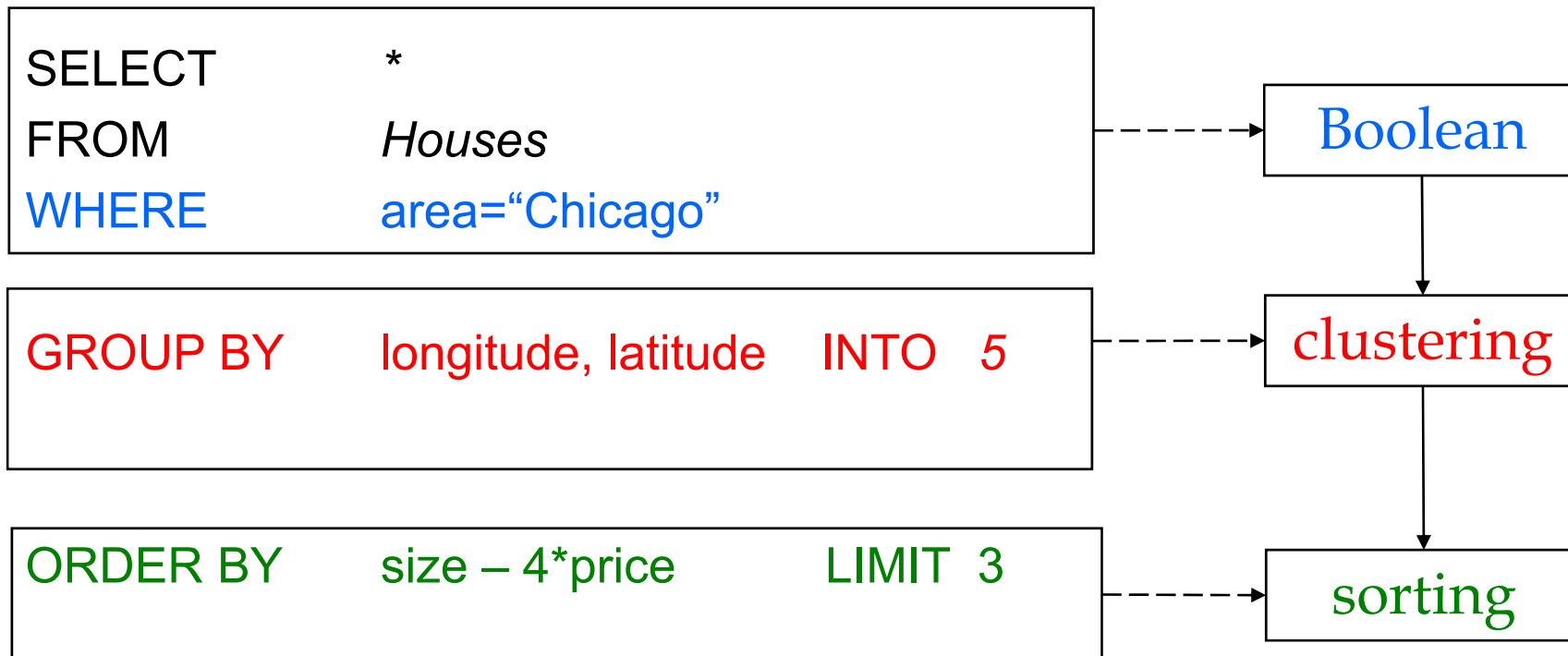
- **Clustering function**

- algorithm, distance measure hidden behind

- **Other semantics**

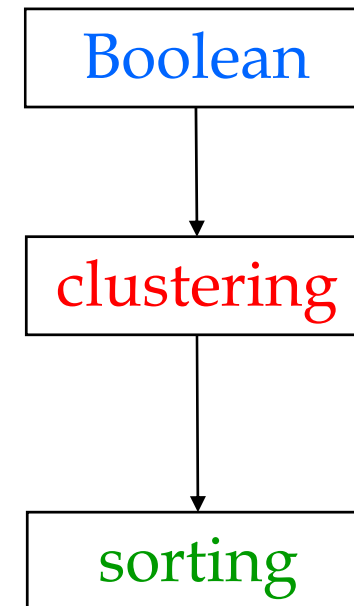
- e.g., cluster the global top k

Query evaluation: Straightforward Materialize-Cluster-Sort approach



Query evaluation: Straightforward Materialize-Cluster-Sort approach

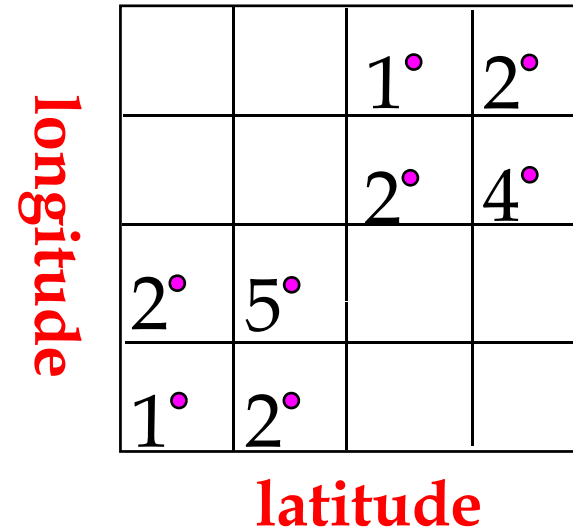
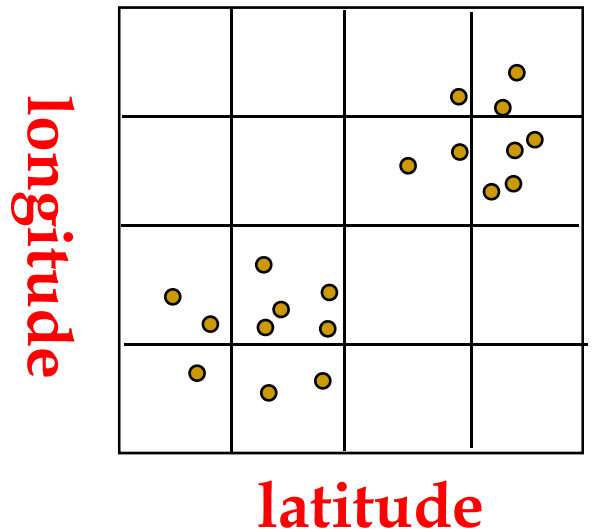
- **Overkill:**
cluster and rank all,
only top 10 in each cluster are requested
- **Inefficient:**
 - ❑ fully generate Boolean results
 - ❑ clustering large amount of results is expensive
 - ❑ sorting big group is costly





Query evaluation: Summary-driven approach

- use summary to cluster
- use summary for pruning in ranking
- use bitmap-index
 - to construct query-dependant summary
 - to bring together Boolean, clustering, and ranking

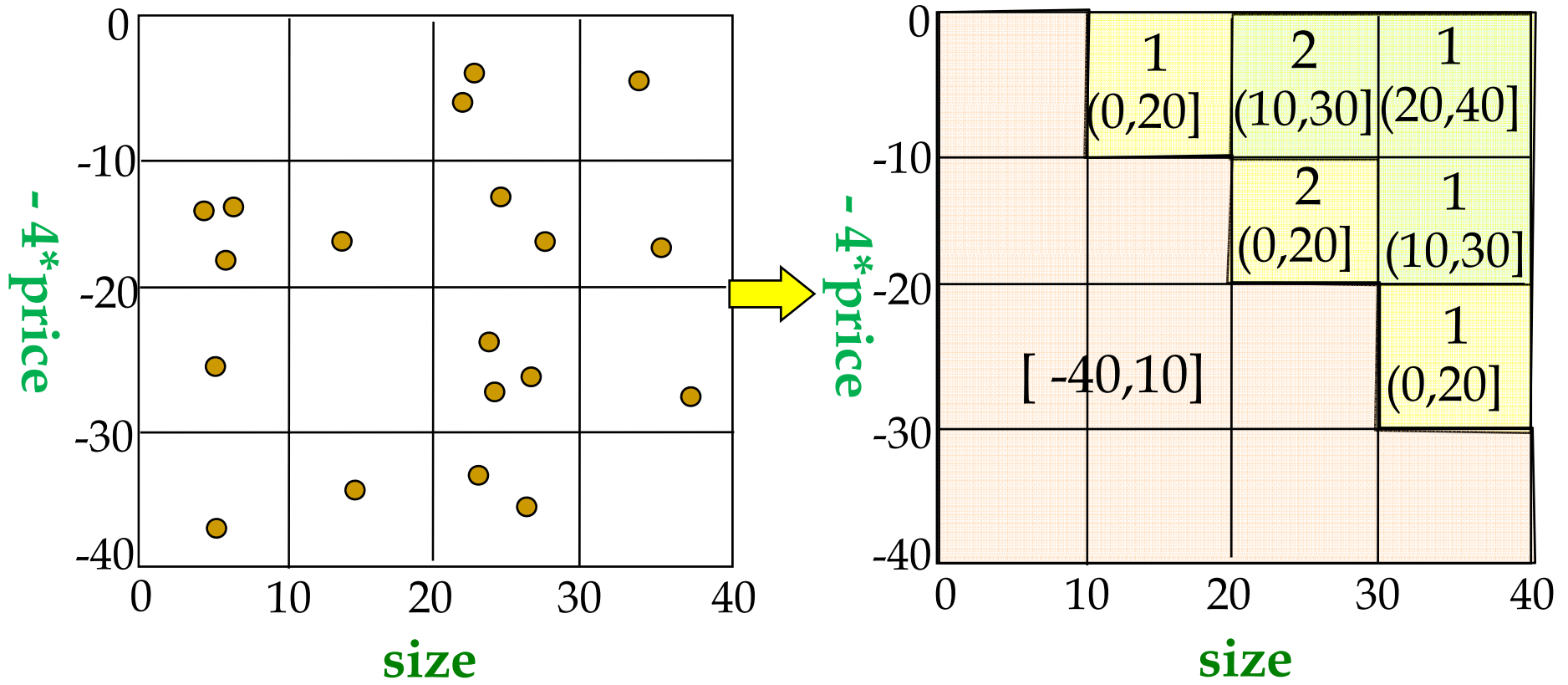
Summary for clustering



↑
K-means 
on original tuples

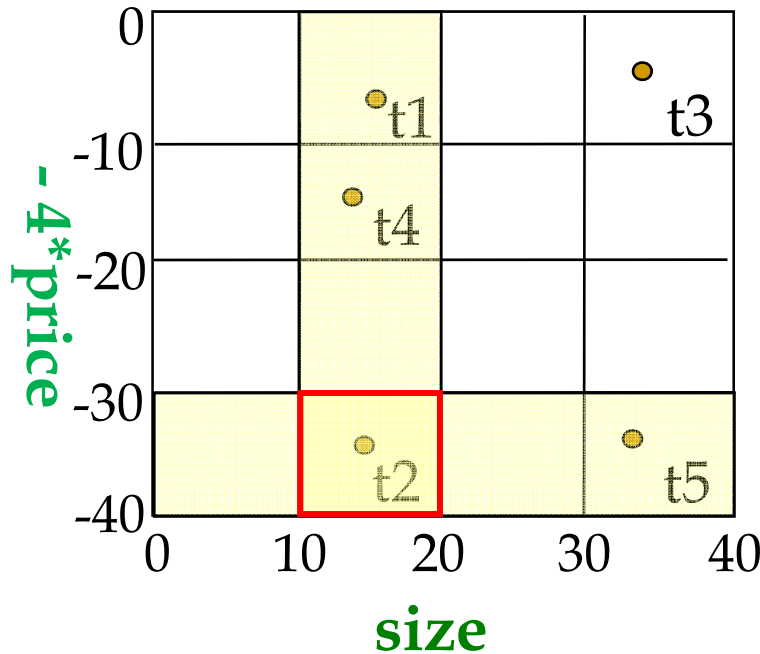
↑
weighted K-means 
on virtual tuples

Summary for ranking



ORDER BY size - 4 * price
LIMIT 3

Construct summary by bitmap-index

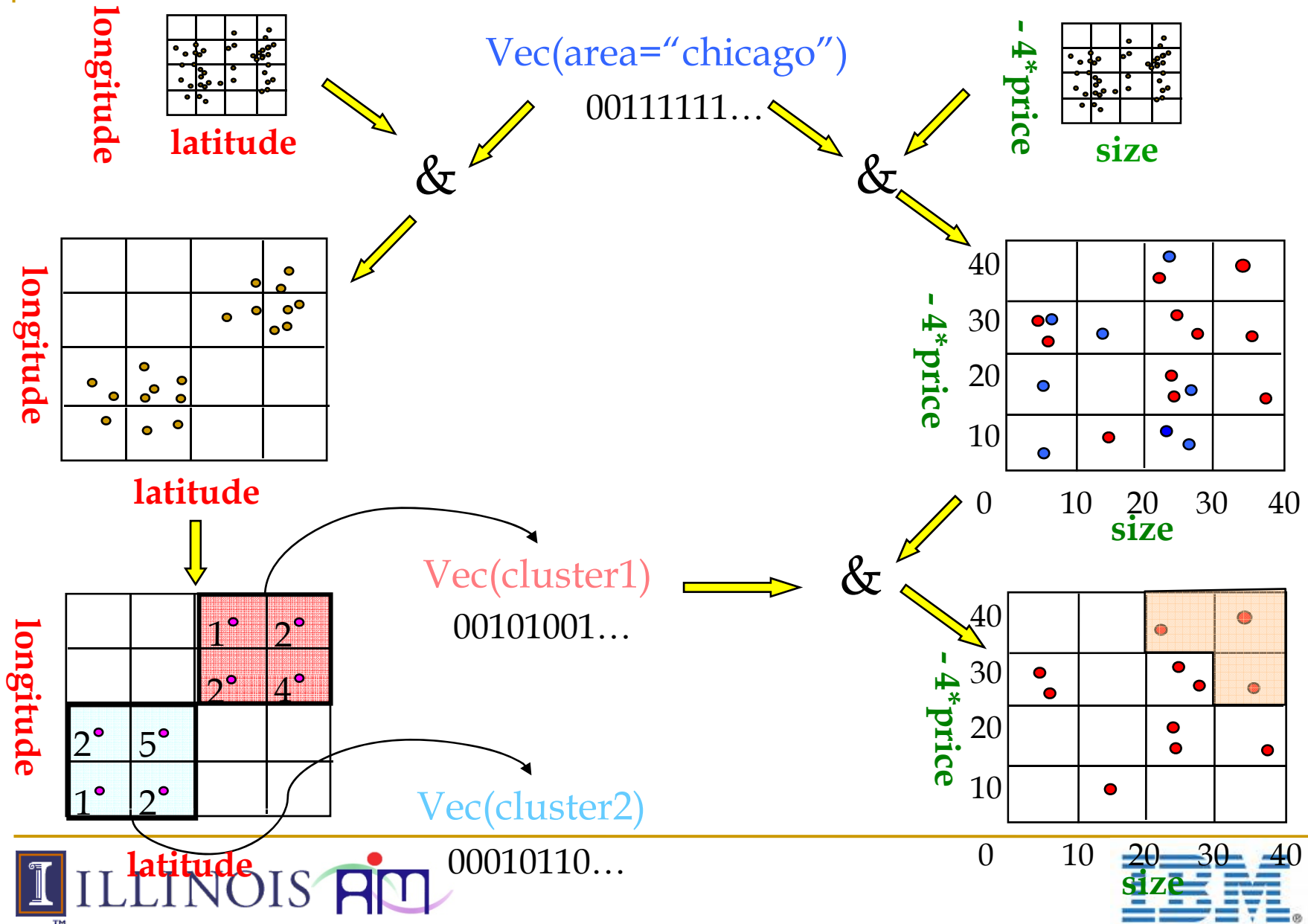


TID	size(10,20]	-4*price[-40,-30]	&
t1	0	1	0
t2	1	1	1
t3	0	0	0
t4	0	1	0
t5	1	0	0
...

The advantages of using bitmap index:

- Small
- Bit operations (&, |, ~, count) are fast
- Easily integrate Boolean, clustering, and ranking

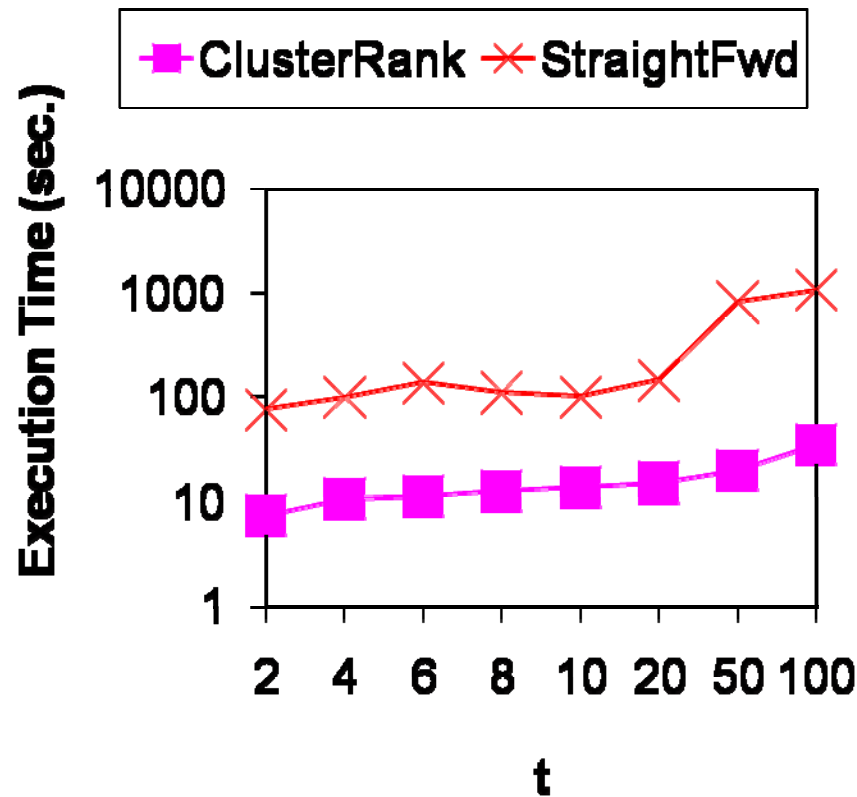
Integrating Boolean, clustering, and ranking



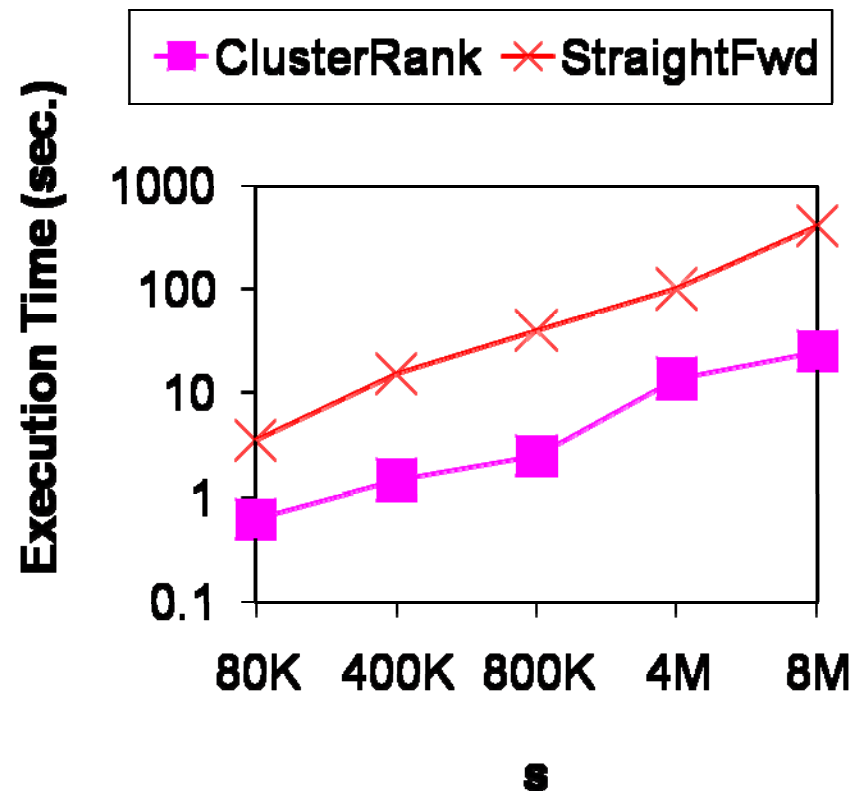
Experiments

- *ClusterRank* (summary-driven approach) vs. *StraightFwd* (materialize-cluster-rank)
 - Processing efficiency: *ClusterRank* >> *StraightFwd*
 - Clustering Quality: *ClusterRank* \approx *StraightFwd*
- synthetic data
- various configuration parameters
(#tuples, #clusters, #clustering attr, #ranking attr, #partitions per attr, k)

Efficiency



t : #clusters
4M tuples, 5 clustering attr,
3 ranking attr, top 5

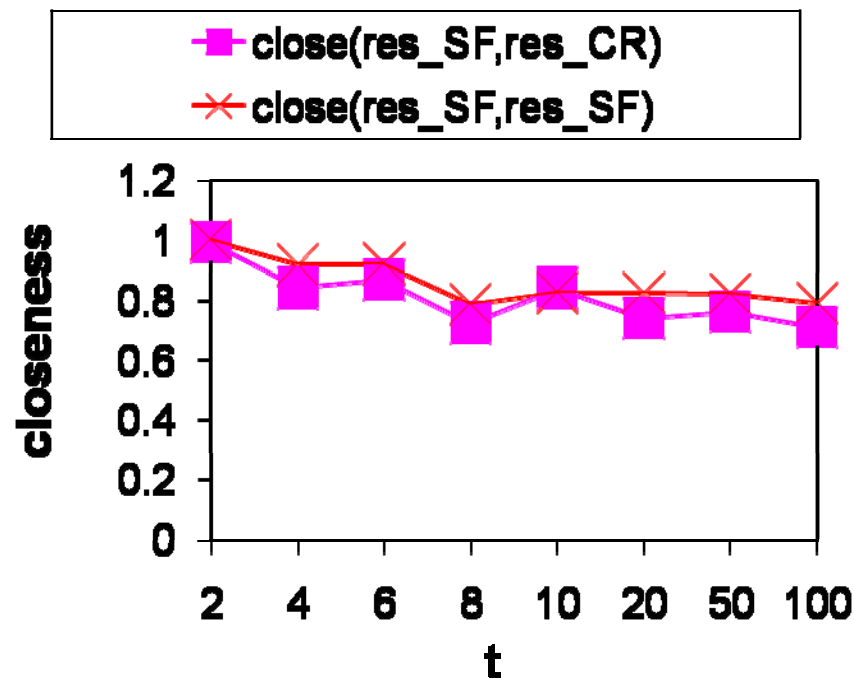


s : #tuples
10 clusters, 5 clustering attr,
3 ranking attr, top 5

Clustering quality

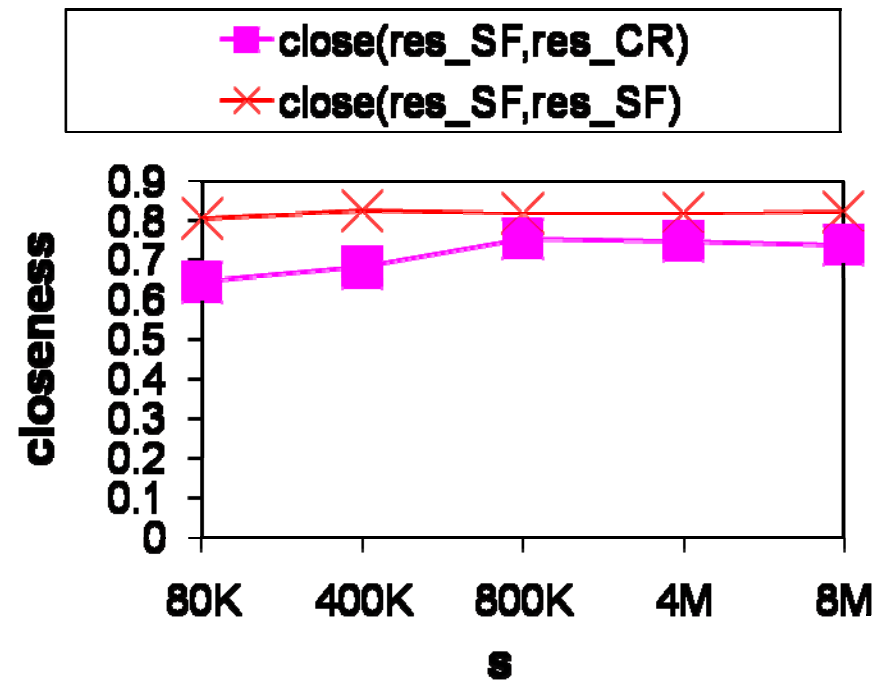
$\text{close}(\text{res_SF}, \text{res_CR})$: closeness of results from StraightFwd and ClusterRank

$\text{close}(\text{res_SF}, \text{res_SF})$: closeness of results from different runs of StraightFwd



t: #clusters

4M tuples, 8 clustering attr



s: #tuples

10 clusters, 3 clustering attr

Conclusions

- Borrow innovative mechanisms from other areas to support **data retrieval** applications
- Ranking and clustering as **generalized Order-By and Group-By**, integrated in database queries
- Query semantics: **ClusterRank** queries
- Query evaluation: **summary-driven approach** vs. materialize-cluster-sort
 - evaluation efficiency: ClusterRank >> StraightFwd
 - clustering quality: ClusterRank \approx StraightFwd

Acknowledgement

- Rishi Rakesh Sinha: source code of bitmap index
- Jiawei Han: discussions regarding presentation



Alternative semantics?

- **global clustering / local ranking (focus of this paper)**
clustering: Boolean results
ranking: local top k in each cluster
- **local clustering / global ranking**
clustering: global top k
ranking: Boolean results
- **global clustering / global ranking**
clustering: Boolean results
ranking: in each cluster, return those belonging to global top k
- **rank the clusters? (by average of local top k?)**

Join queries

- Star-schema
fact table, dimension tables, key and foreign key
- Bitmap join-index
index the fact table by the attributes in dimension tables