Oregon State University

School of Electrical Engineering and Computer Science

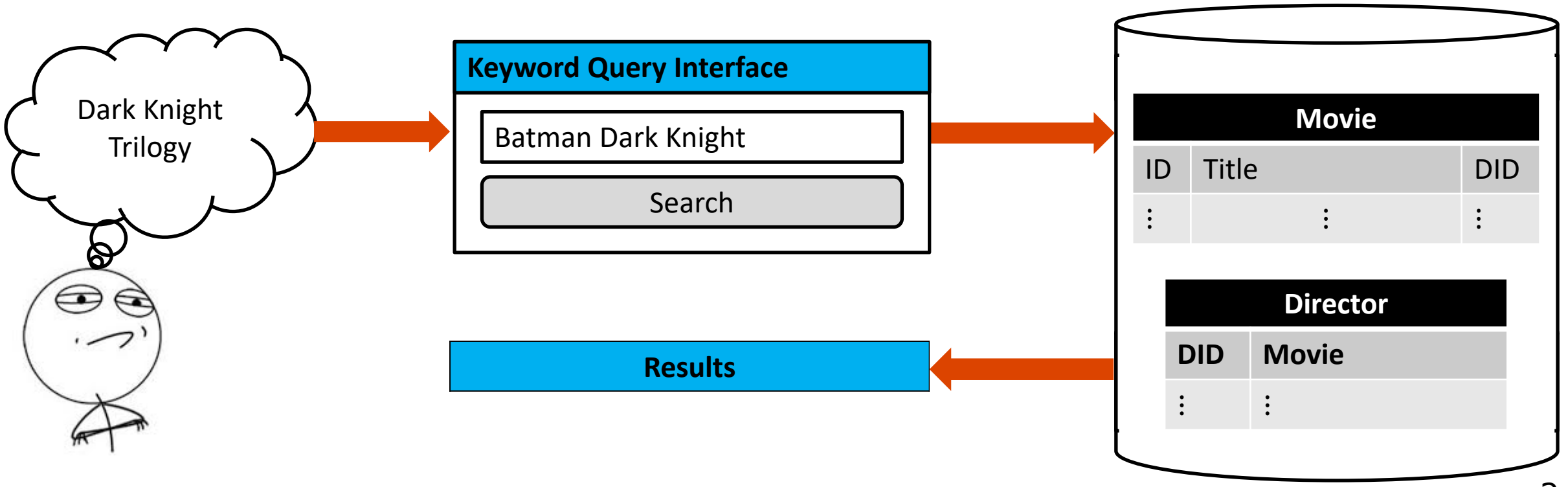
There is no dichotomy between effectiveness and efficiency in keyword search over databases

Vahid Ghadakchi, Arash Termehchy

IDEA Lab

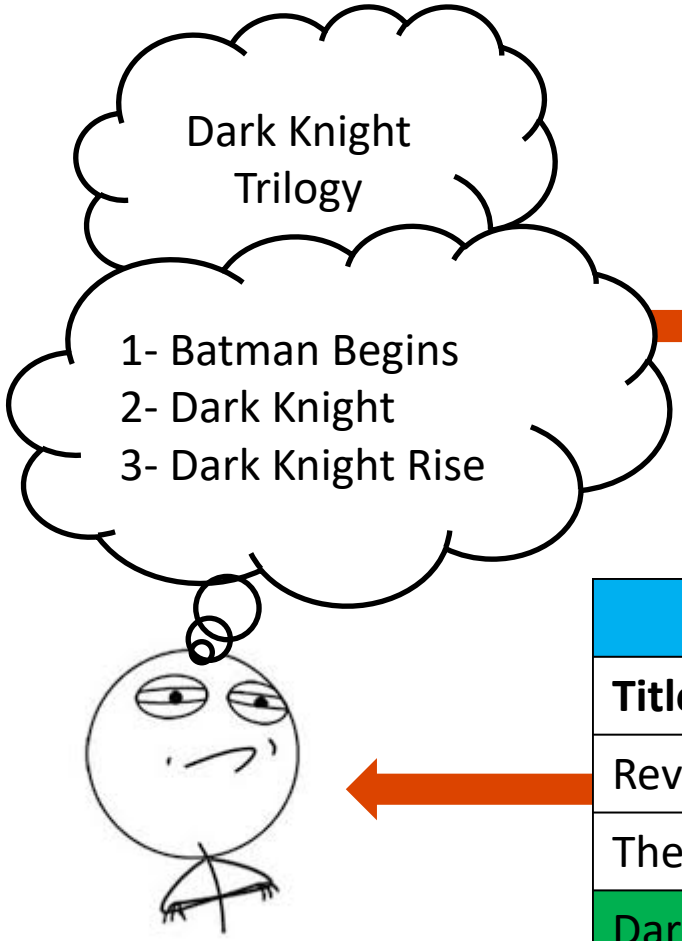
Most users can not express their intent over databases

- Most users are not familiar with SQL, schema and exact content



Keyword queries are inherently vague

Oregon State University OSU



Keyword Query Interface

Batman Dark Knight

Search

Movie

ID	Title	DID
⋮	⋮	⋮
4	Dark Knight Rises	40
⋮	⋮	⋮
10	Batman Begins	40

Results

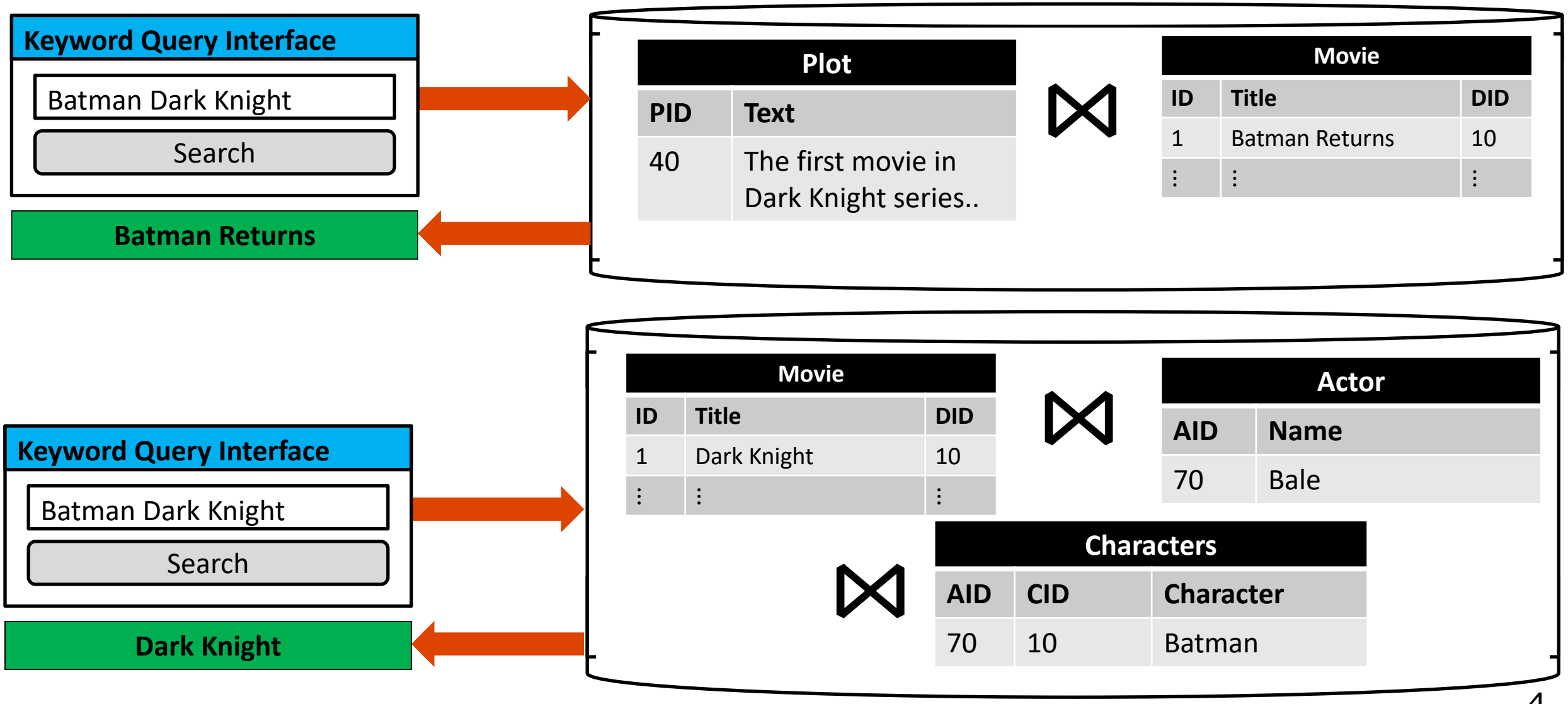
Title	Director
Reviews: Batman Dark Knight	Antwiller
The Dark Knight Movie Review	Rodriguez
Dark Knight	Nolan
Dark Knight Parody	Bane
Dark Knight Aurora	Lopez

Precision = 1/5

Recall = 1/3

Keyword query interfaces has low efficiency

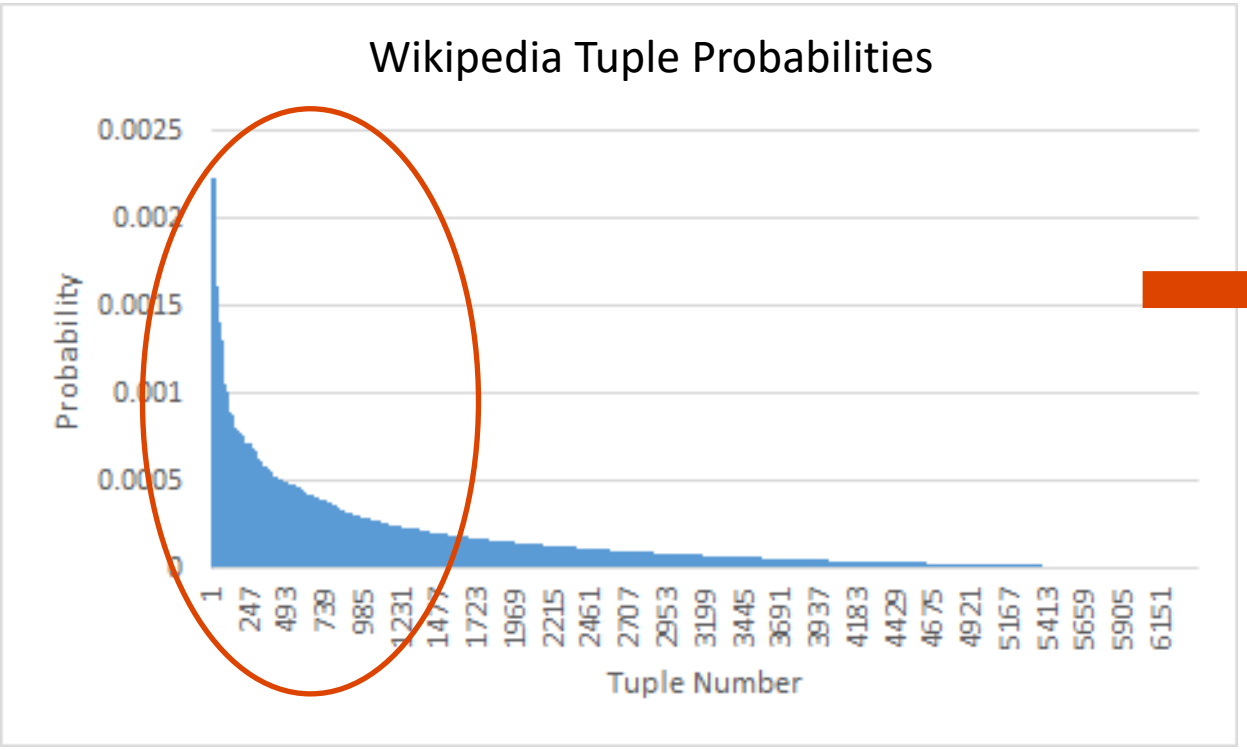
Oregon State University OSU

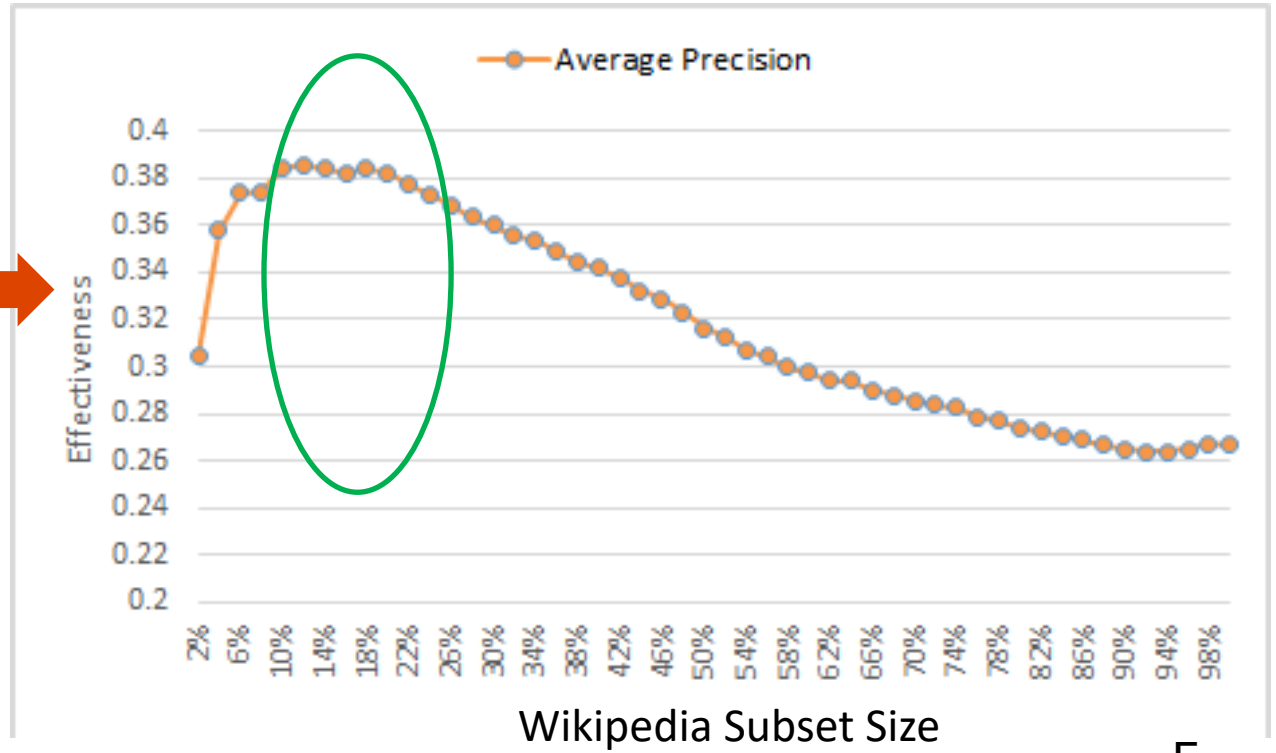


Leveraging the query distribution

Oregon State University OSU

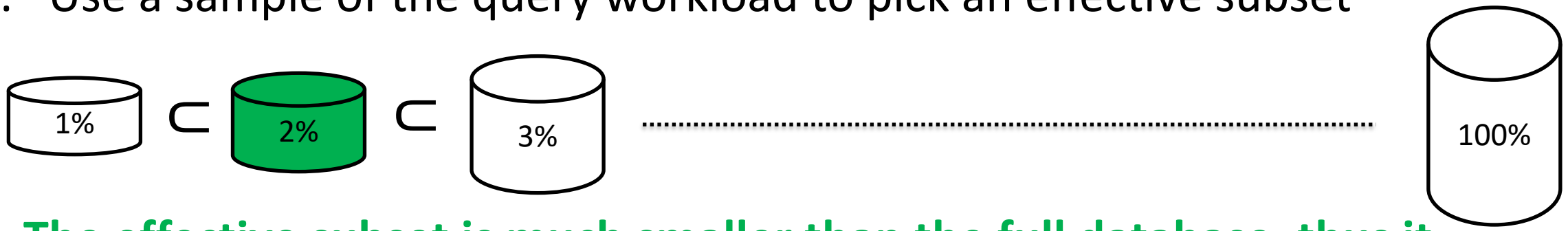
- The probability of a tuple being a relevant answer to a query follows a Zipfian distribution
- A small subset has most of the relevant answers
- **Solution: Make an effective subset using tuples with high probability**





The algorithm to pick the effective subset

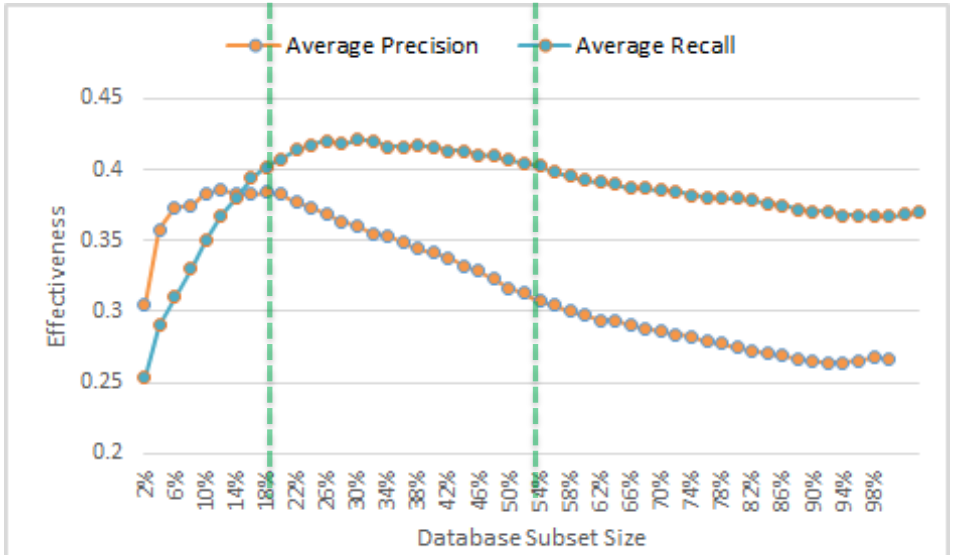
1. Compute probability of each tuple based on past interactions
2. Sort tuples based on their probability
3. Build different subsets of the database with tuples with high probability
4. Use a sample of the query workload to pick an effective subset



- **The effective subset is much smaller than the full database, thus it increases the efficiency of query answering while increasing the average precision**
- **The effective subset does not include all the tuples**
  - May decrease recall and have problem with long tail queries

How we handle recall and long-tail queries

- Recall: Effective subset can preserve recall while maintaining high precision



- Long-tail queries: Our system uses a machine learning technique to send the long-tail queries to the full database

Results on real world data and query workload

Oregon State University OSU

- Dataset: Snapshot of Wikipedia with 12 million documents
- Query Set #1: 7000 keyword queries sampled from MSN search engine
- Query Set #2: 150 keyword queries from INEX competition
- Search System: Lucene over MySQL database

	Effective Subset	Full Database
MRR of Query Set #1	0.62	0.25
MRR of Query Set #2	0.80	0.65
Average Query Time	27 (ms)	205 (ms)