

# Distributed Query-Sampling: A Quality-Conscious Approach

James Caverlee      Georgia Tech

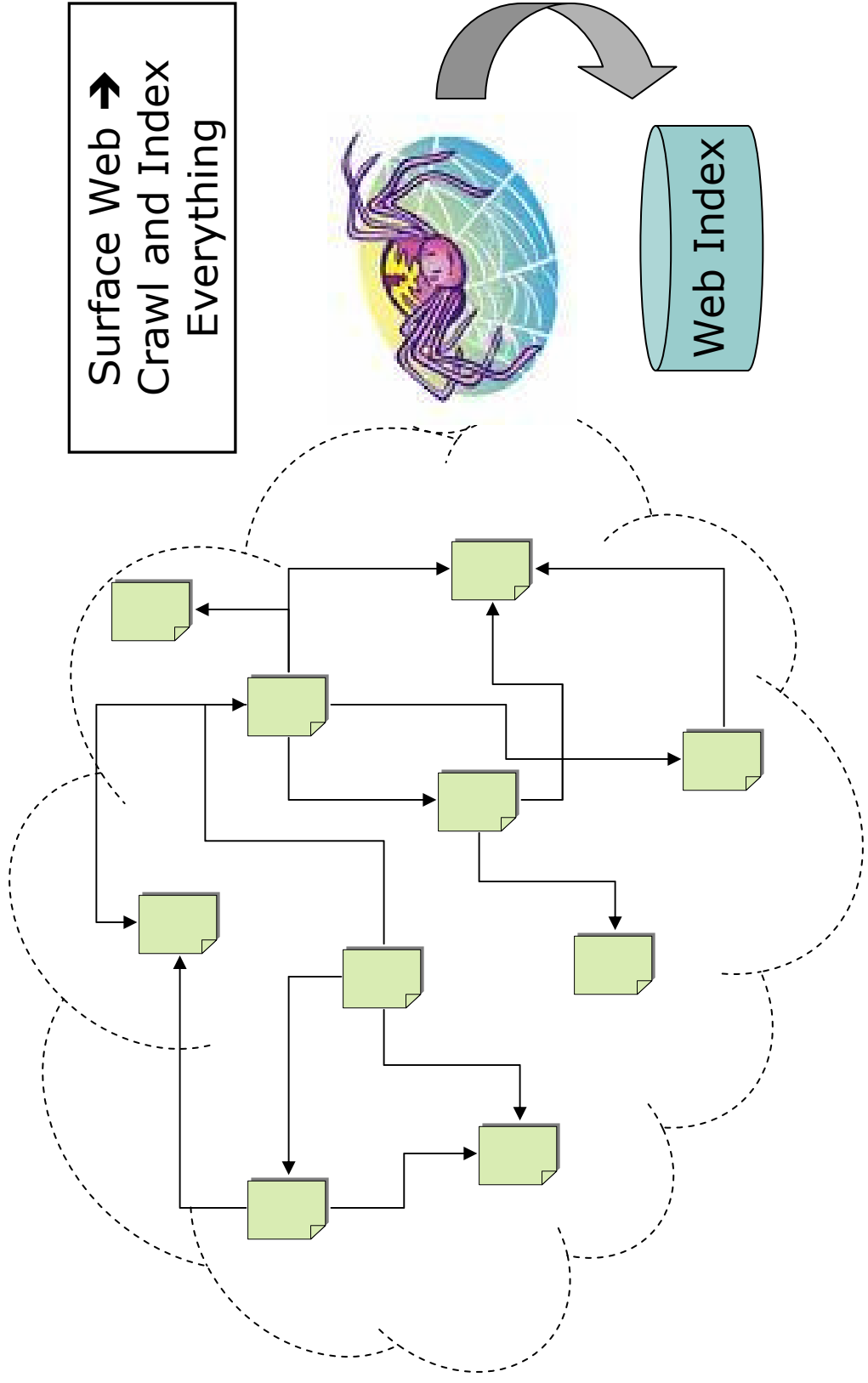
Ling Liu

Georgia Tech

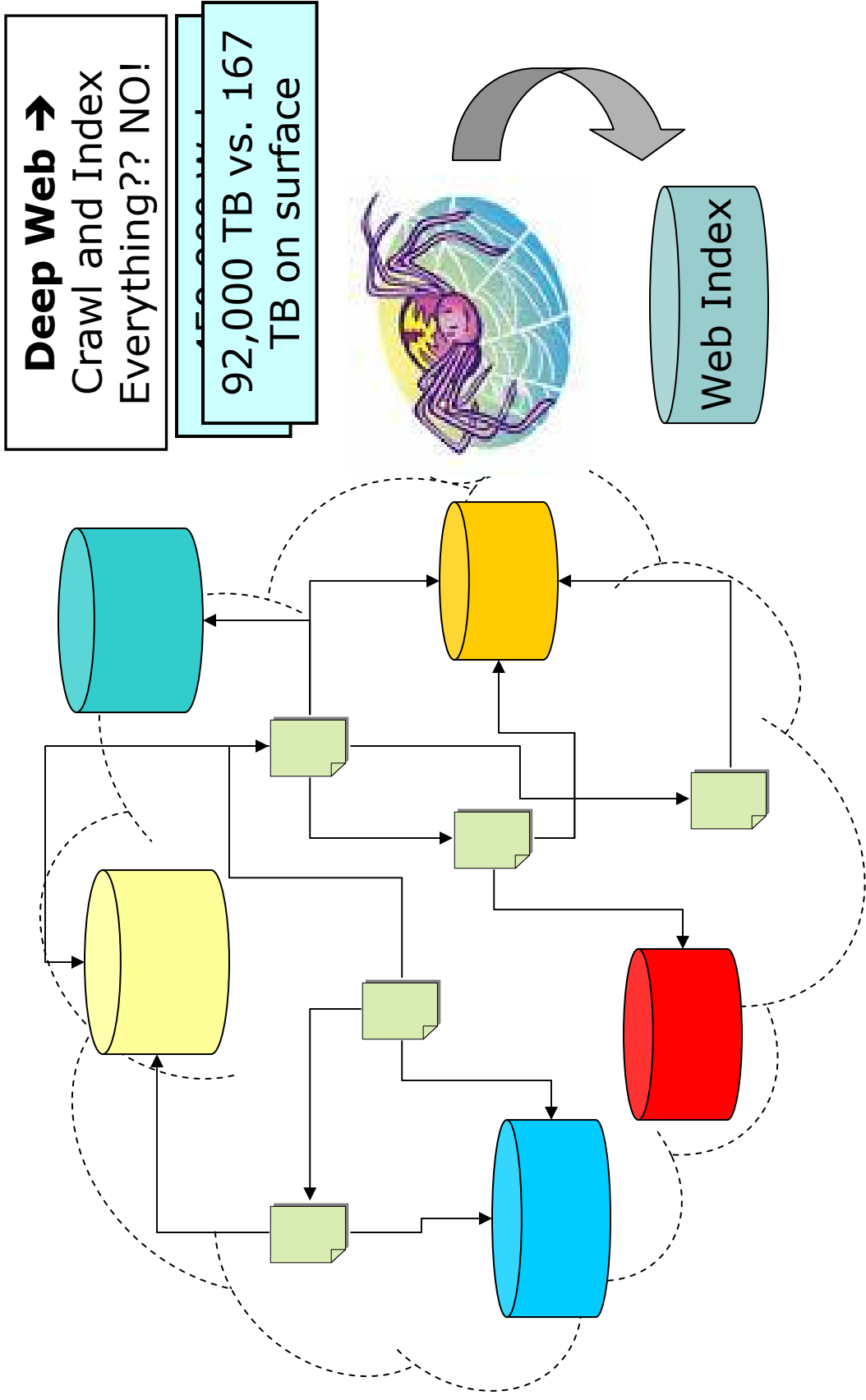
Joonsoo Bae

Chonbuk Natl U.

# Why is Sampling Useful?



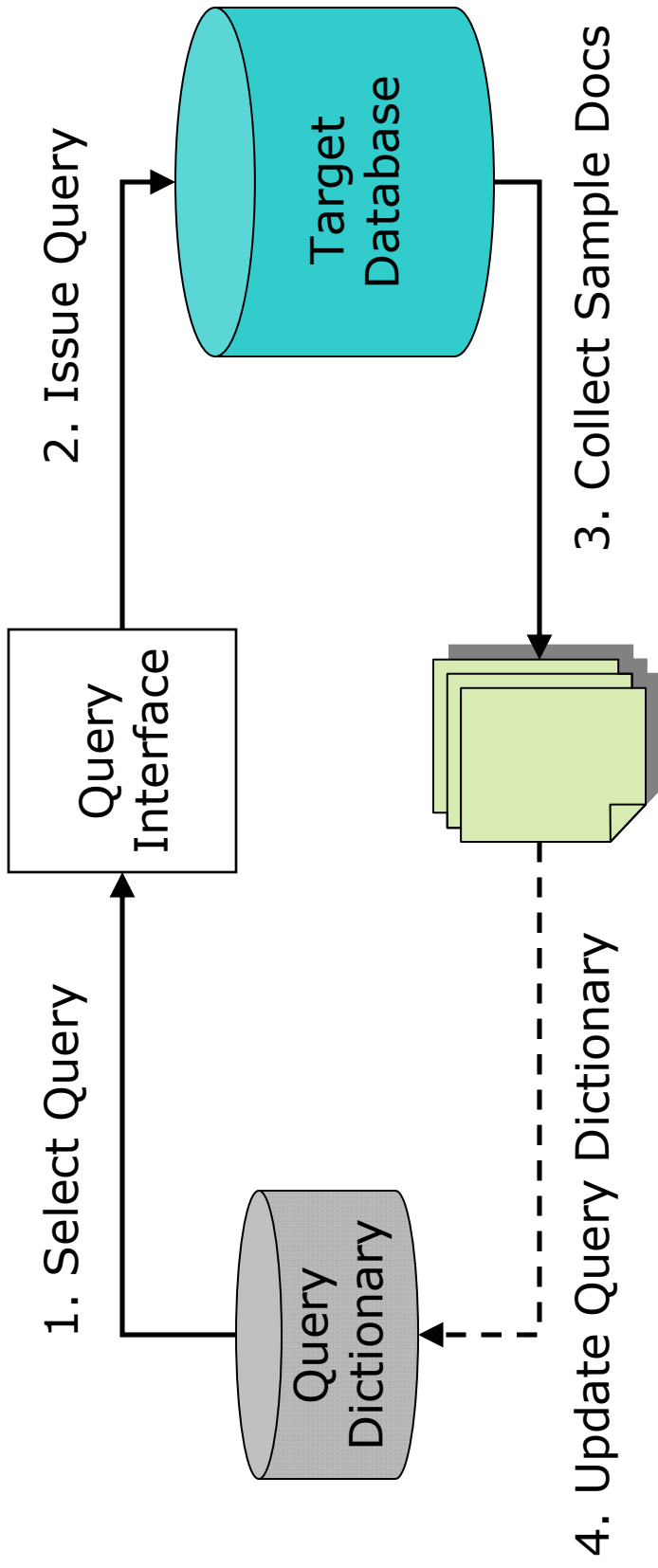
# Why is Sampling Useful?



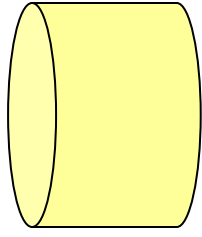
# How Do We Sample?

- Query-Based Sampling

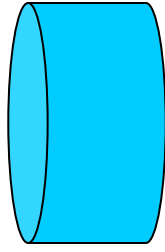
- [Callan et al. TOIS '01, SIGMOD '99]



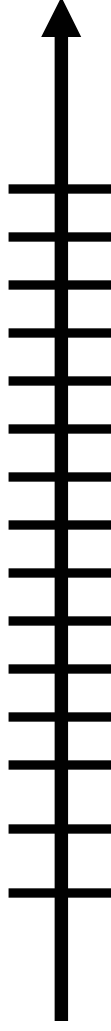
# Key Challenges for



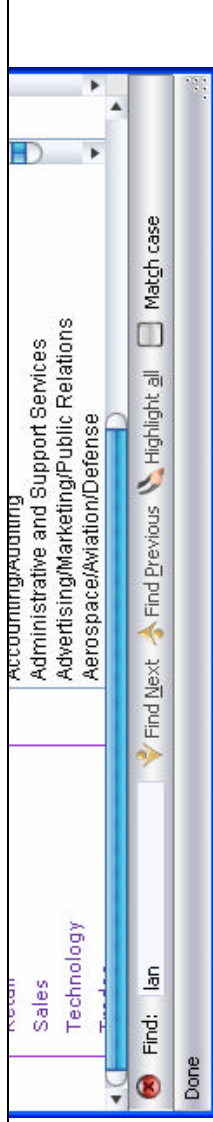
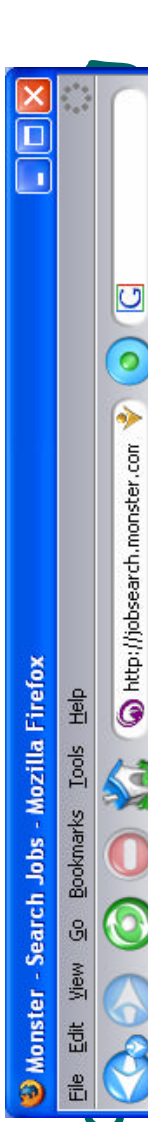
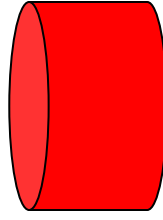
1.



2.



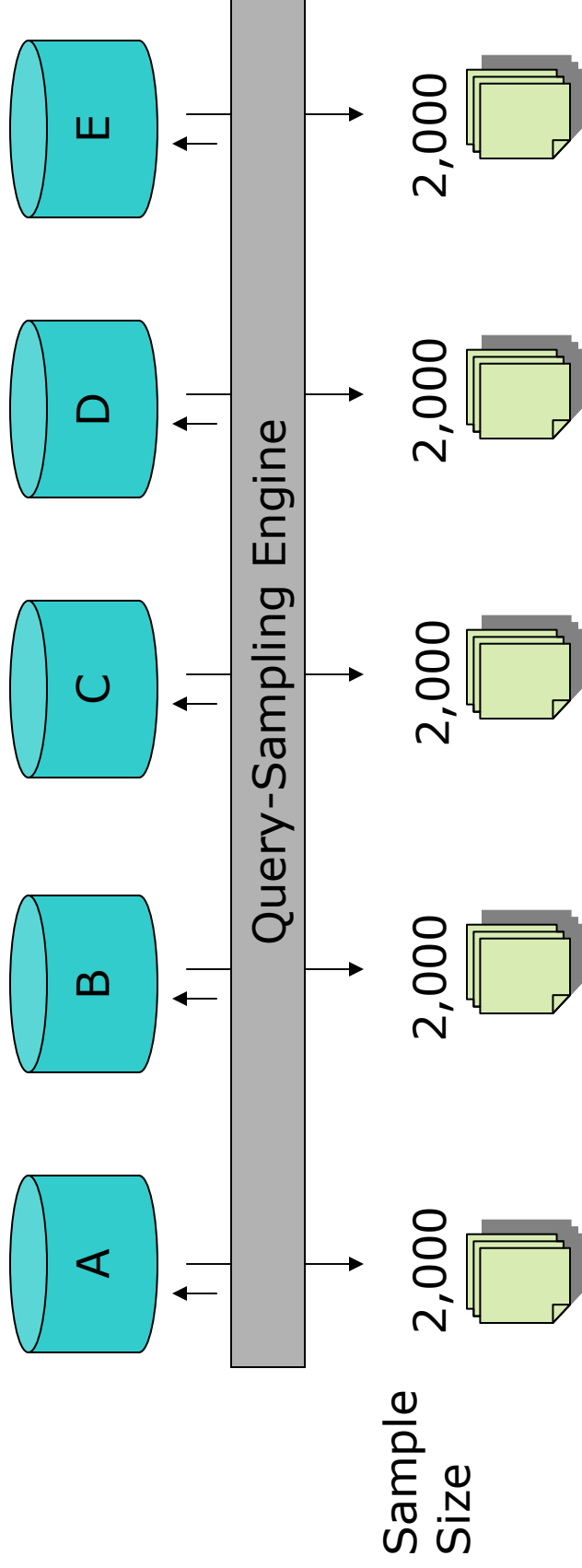
3.



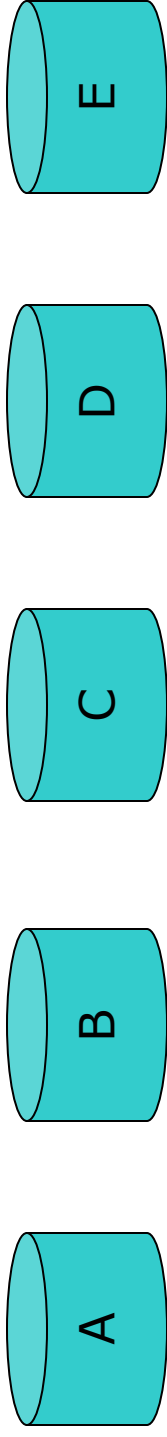
# Popular Solution - Uniform Sampling

Total # of sample docs = 10,000

Total # of databases = 5



# Uniform Sampling Problems



- Size differences
  - A has 1,000 docs; B has 1 million
- Diversity of information
  - B has basketball info; C has basketball, football, baseball, ...
- Duplicate (or near duplicate) databases
  - D and E are mirrors

○ ...

## Key Ideas:

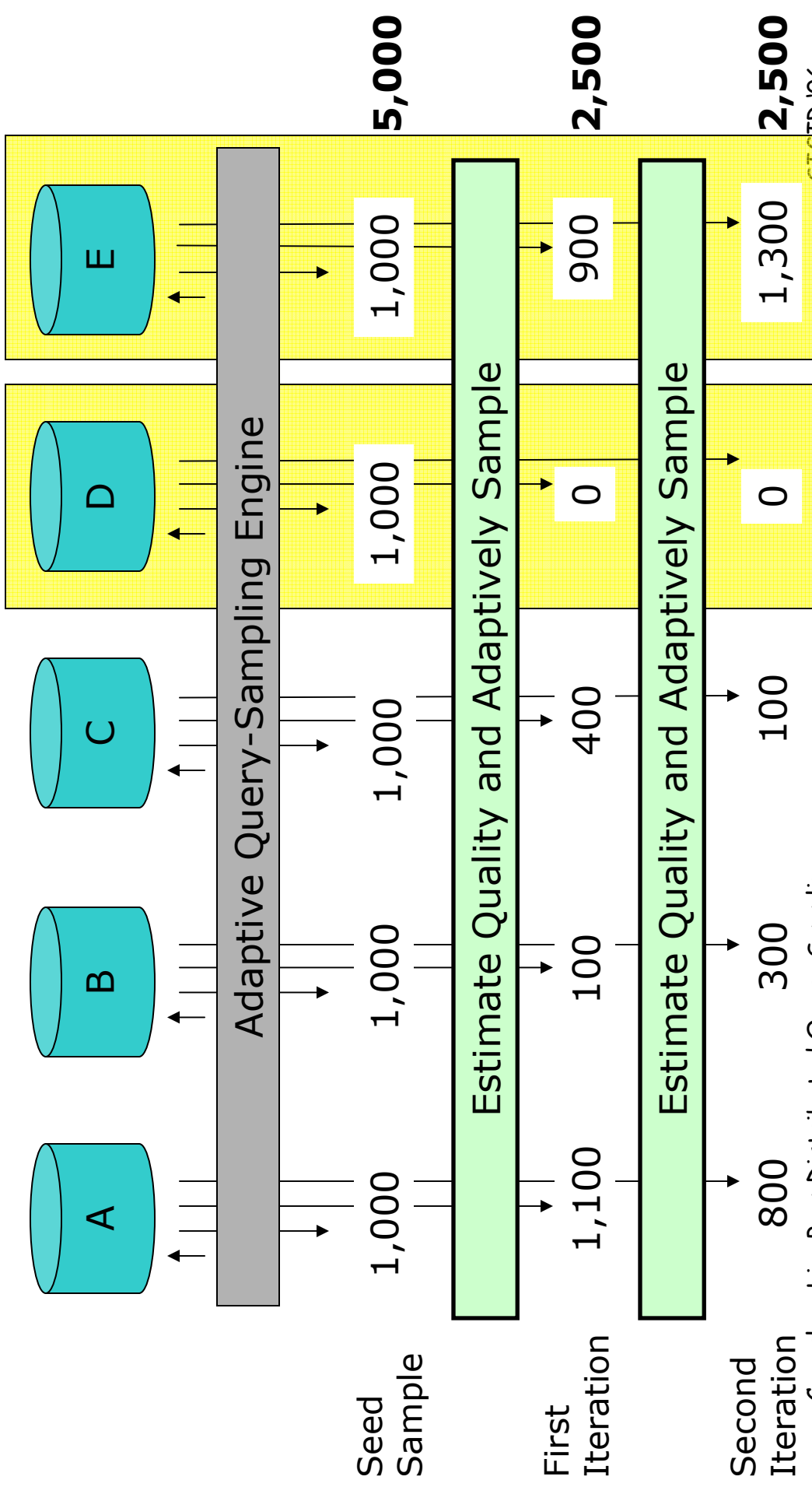
- Databases vary in quality
- Adaptively allocate resources to reflect this quality

# Our Solution: Quality-Conscious Adaptive Sampling

- 1. Use fraction of total resources to collect seed samples from each database
  - 2. Use collected samples to estimate quality of each database
  - 3. Generate a new sampling allocation based on relative quality
  - 4. Collect additional quality-conscious samples
  - 5. Goto 2 until resources exhausted
- Iterate

10,000 sample docs total; 5 databases

# Adaptive Sampling: Simple Example



# Challenges for Adaptive Sampling

- 1. How do we estimate database quality?
  - We describe three schemes in the paper
- 2. How do we tell if one scheme is better than another?
- 3. How should we divide the total resources between the seed and adaptive phases?
- 4. For how many iterations should we adaptively sample?

# 1. How Do We Estimate Database Quality?

- Proportional Document Ratio [PD]
- Proportional Vocabulary Ratio [PV]
- Vocabulary Growth [VG]

# Sampling Scheme 1: Proportional Document Ratio

- Idea: Collect the same document proportion from each database

How do we estimate  $|D|$  ??

→ sample-resample  
[Si & Callan, SIGIR 2003]

This estimates **ideal sample size** for each database – need to scale to account for seed sampling [see paper]

Uniform: 2,000    2,000    2,000    2,000

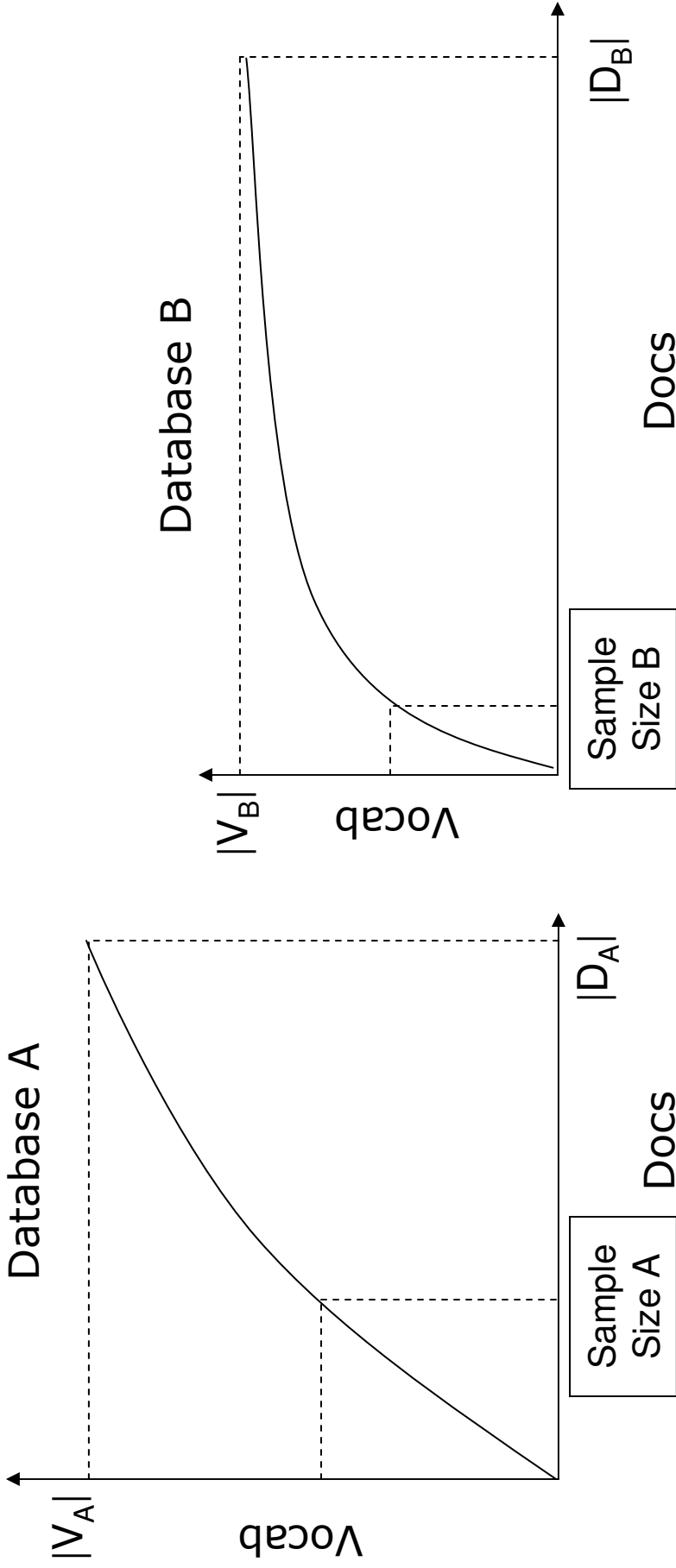
PD: 200    1,000    2,000    5,000    1,800

# Sampling Scheme 2: Proportional Vocabulary Ratio

- Idea: Collect the same vocabulary proportion from each database
  - $\text{ratio}_{PV}$
  - E.g., 10% of the vocabulary terms at  $D_1, D_2, \dots, D_n$
- Favors databases with large vocabularies
- Challenge:
  - Each database will have different size vocabularies and different growth rates of vocabulary
  - Need to estimate vocabulary size and growth rate

# Sampling Scheme 2: Proportional Vocabulary Ratio

How big should our sample be to collect  $\text{ratio}_{PV} = 50\%$  of vocabulary terms?



# Challenge: Estimate Vocabulary Size

$$|V| = K^* (\text{text\_size})^\beta$$

$K^*$  (avg\_doc\_size)<sup>\*</sup> |D|<sup>β</sup>

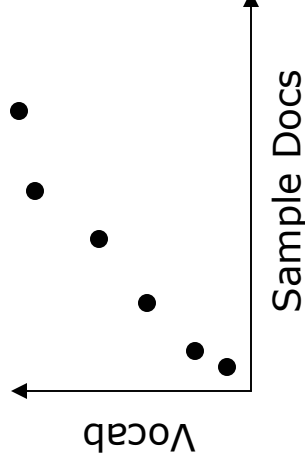
But these parameters are unknown!

Rely on seed sample to estimate:

avg\_doc\_size → straightforward

|D| → use sample-resample [Si & Callan, SIGIR 2003]

K and Beta? curve fitting to estimate

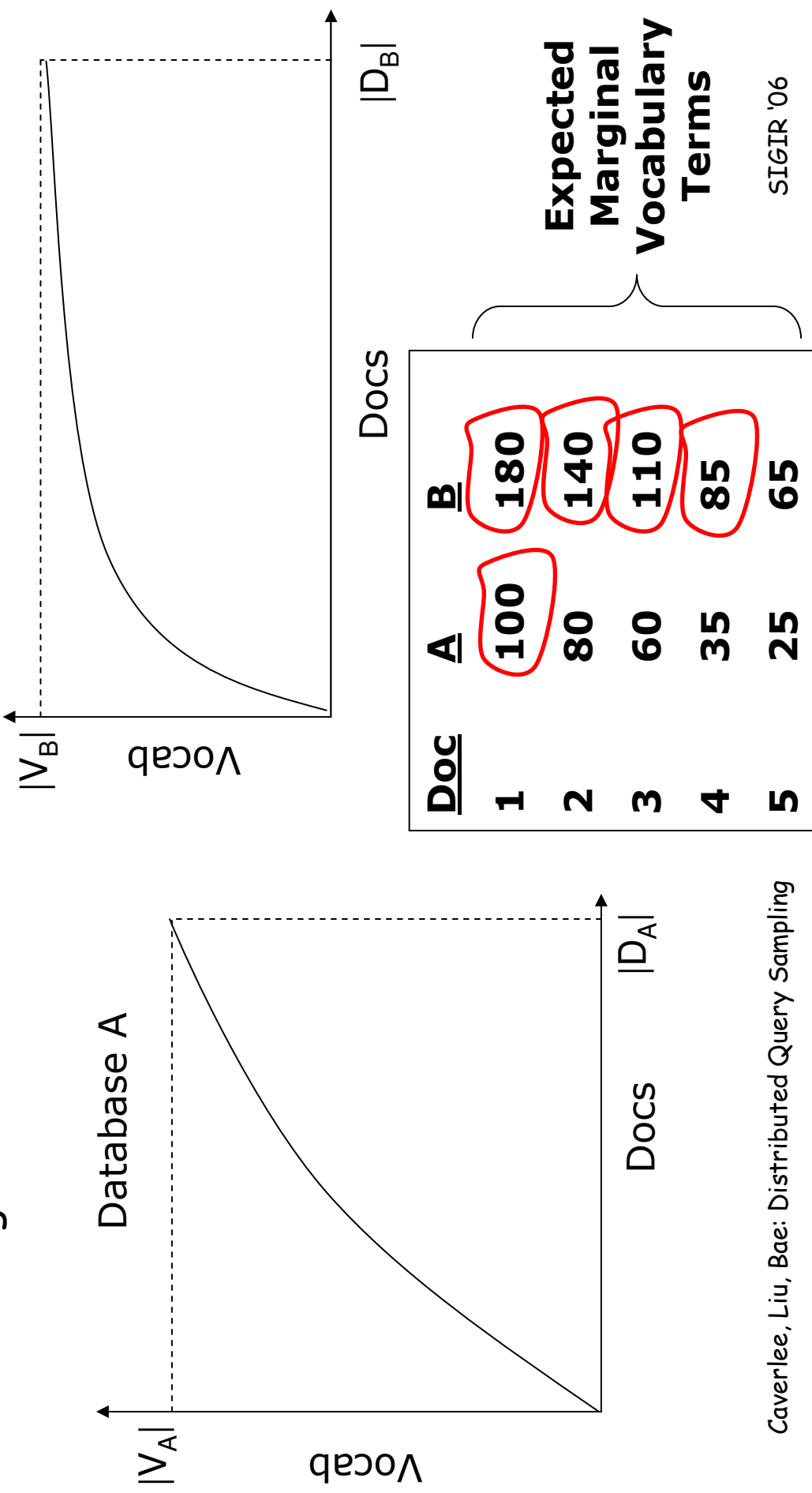


# Challenge: Estimate Sample Size

- $n$  equations like:
  - $\text{ratioPV} * |V| = K * (\text{avg\_doc\_size} * \text{sample\_size})^B$
- Solve for *sample\_size*
  - $\text{sample\_size} = e^{\wedge}(\dots \text{ratioPV} \dots)$
  - But **ratioPV** is still unknown
- Vary **ratioPV** until we have:
  - $N = \text{sample\_size}(A) + \text{sample\_size}(B) + \dots$

# Sampling Scheme 3: Vocabulary Growth

- Favors databases with a fast growing vocabulary, regardless of total size



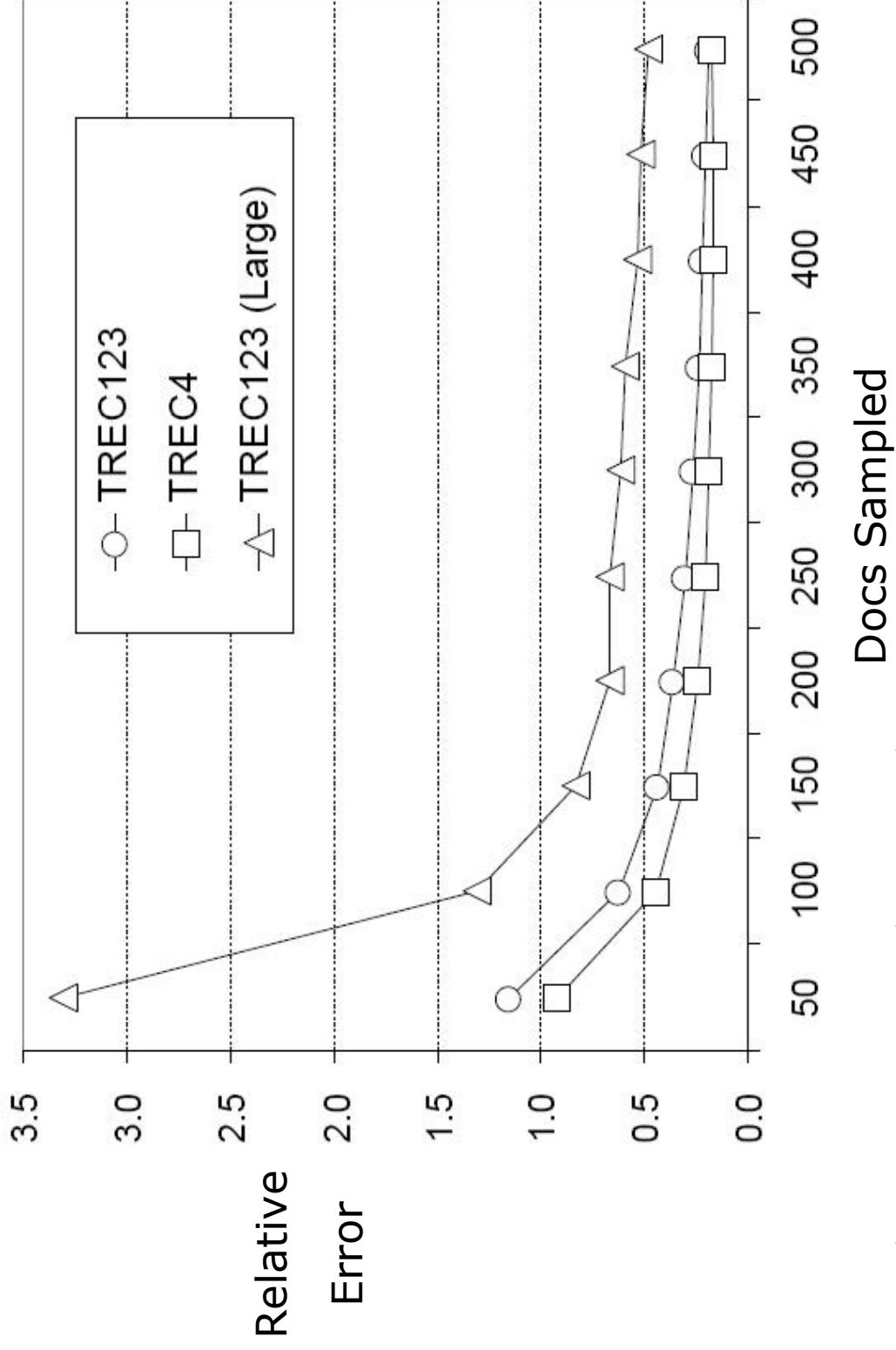
## 2. How Do We Evaluate Sampling Schemes?

- Compare unigram language models for Sample Docs vs. Complete Database
  - Quality of term weights
    - Weighted Common Terms
    - Jensen-Shannon Divergence
  - Quality of term rankings
    - Spearman rank correlation coefficient
- Application scenario
  - Distributed IR - Database Selection
  - Measure Recall @ 1, 2, ..., 20

# Experimental Setup

- Data
  - TREC 123
    - 100 databases, by source and publication date
    - Avg docs ~11,000
  - TREC 4
    - 100 databases, organized by topic
    - Avg docs ~5,700
  - TREC 123 Large Databases
    - Aggregated several smaller dbs
    - Range from 45,000 docs to 242,000 docs
- Sampling
  - Stopwords removed, Porter's Stemmer
  - Indexed & Searched by Lucene
  - Max 4 docs per query
  - Initial queries selected from UNIX dictionary; subsequent queries selected randomly weighted by term frequency

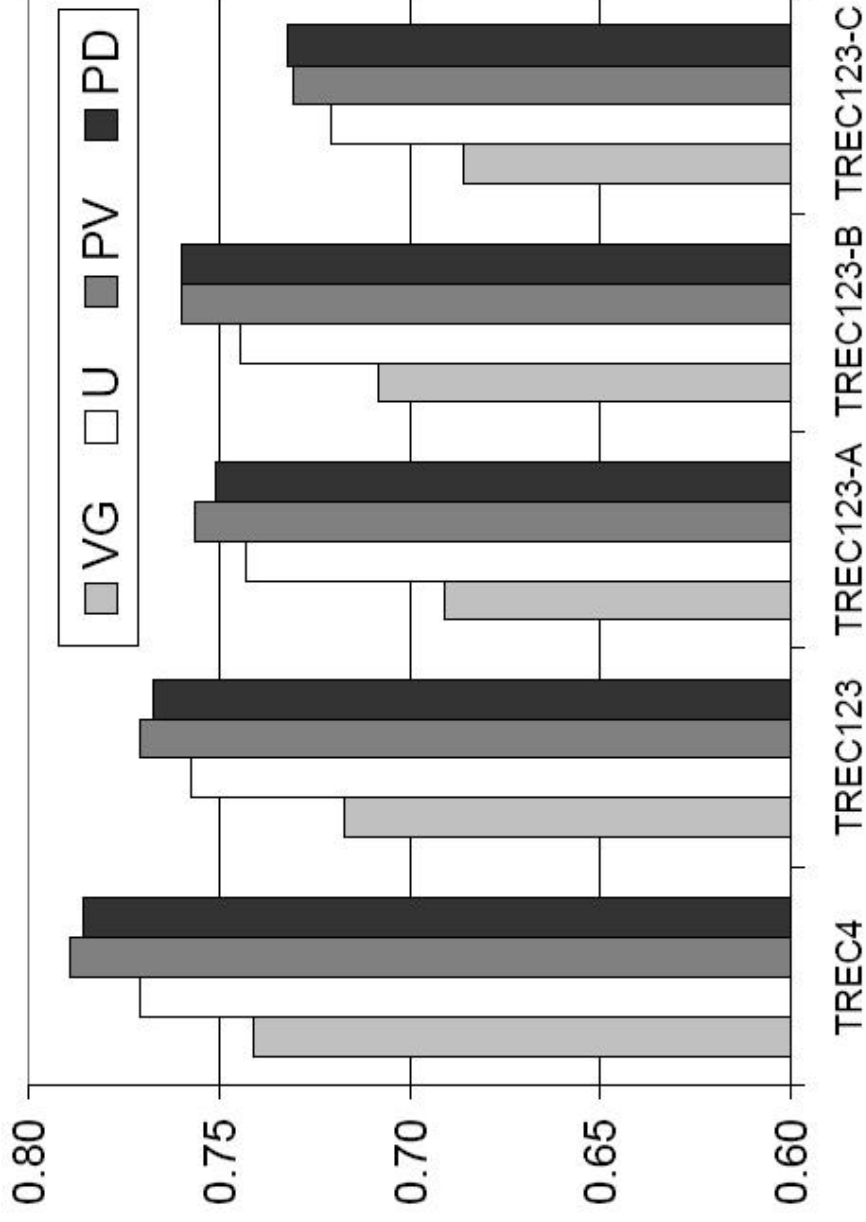
# Vocabulary Size Estimation Error



Total Sample Docs =  
300 \* # of DBs

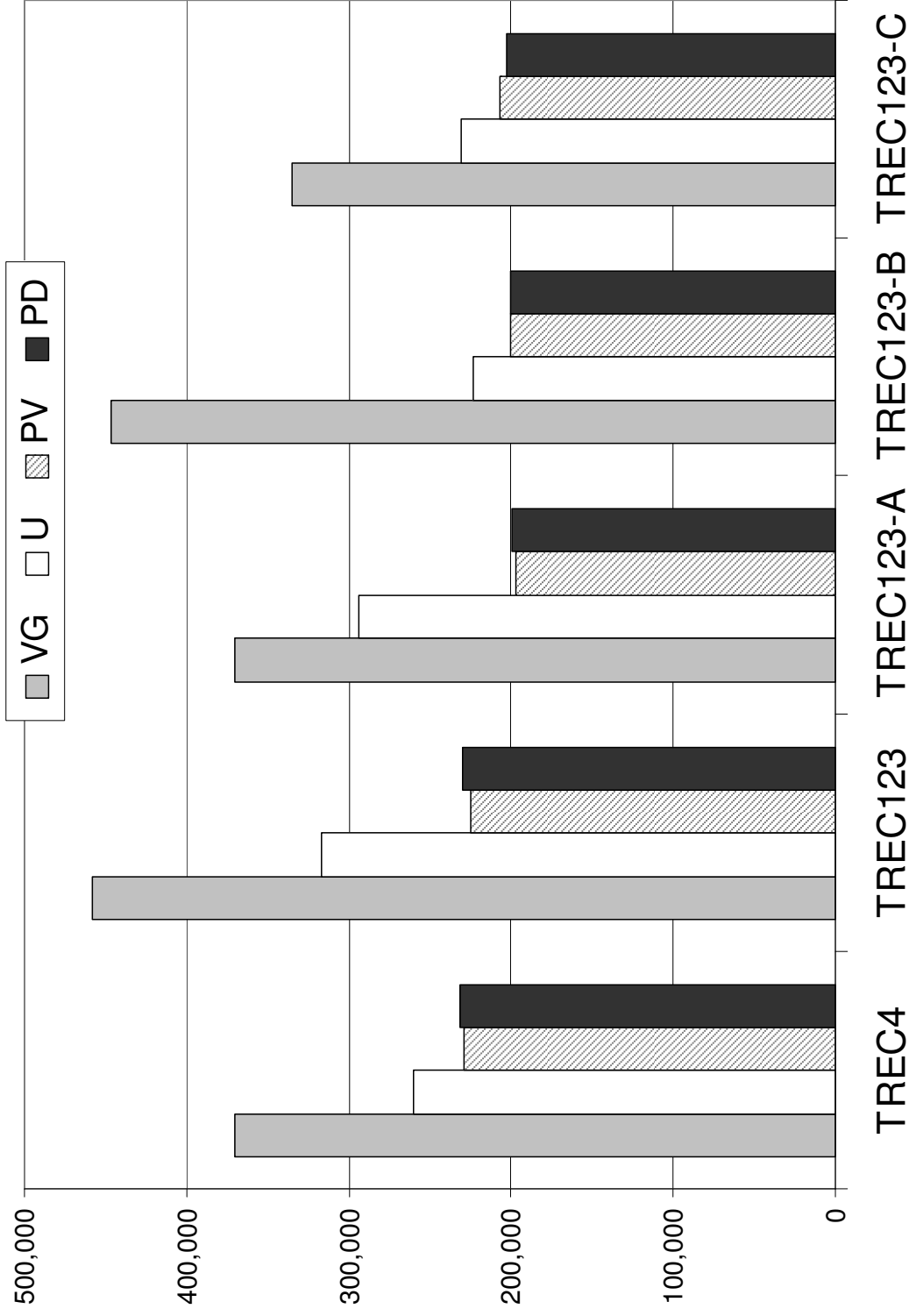
1/2 Total Sample Docs  
Used For Seed Sampling

# Sample Quality - Spearman

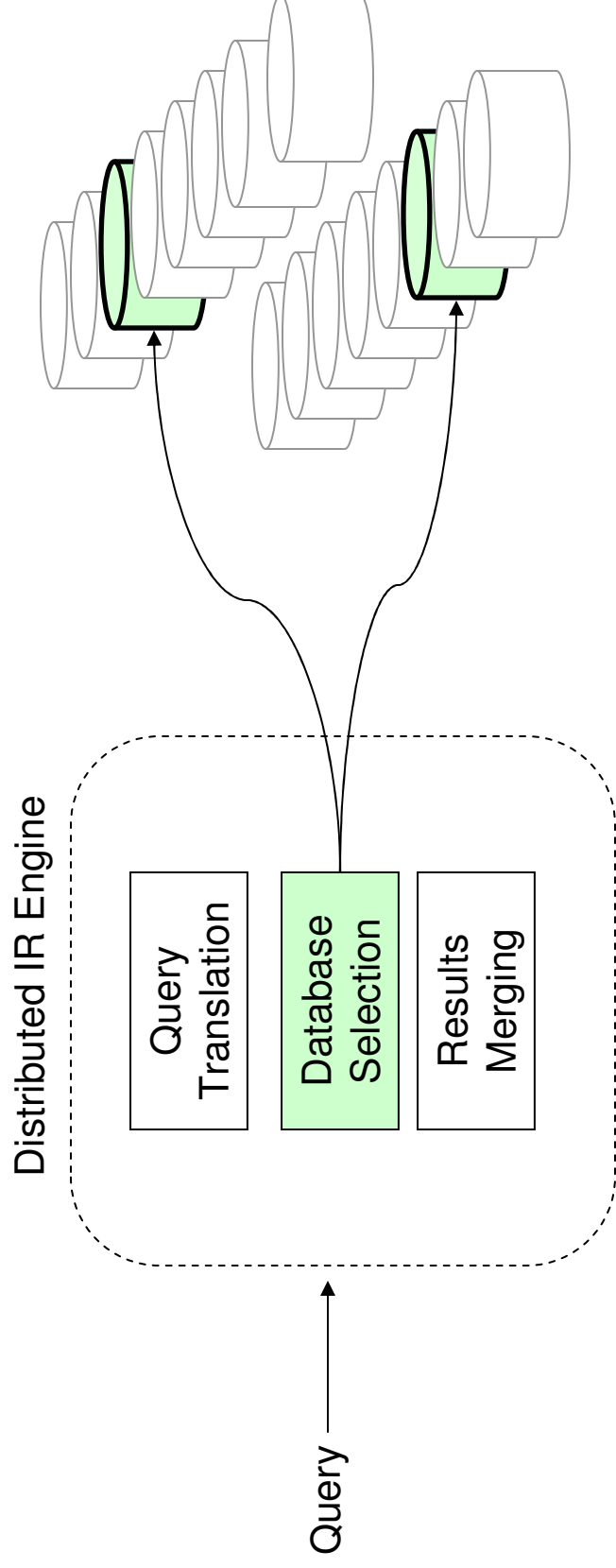


See Paper For More

# Total Collection Vocabulary Size



# Application Scenario: Database Selection



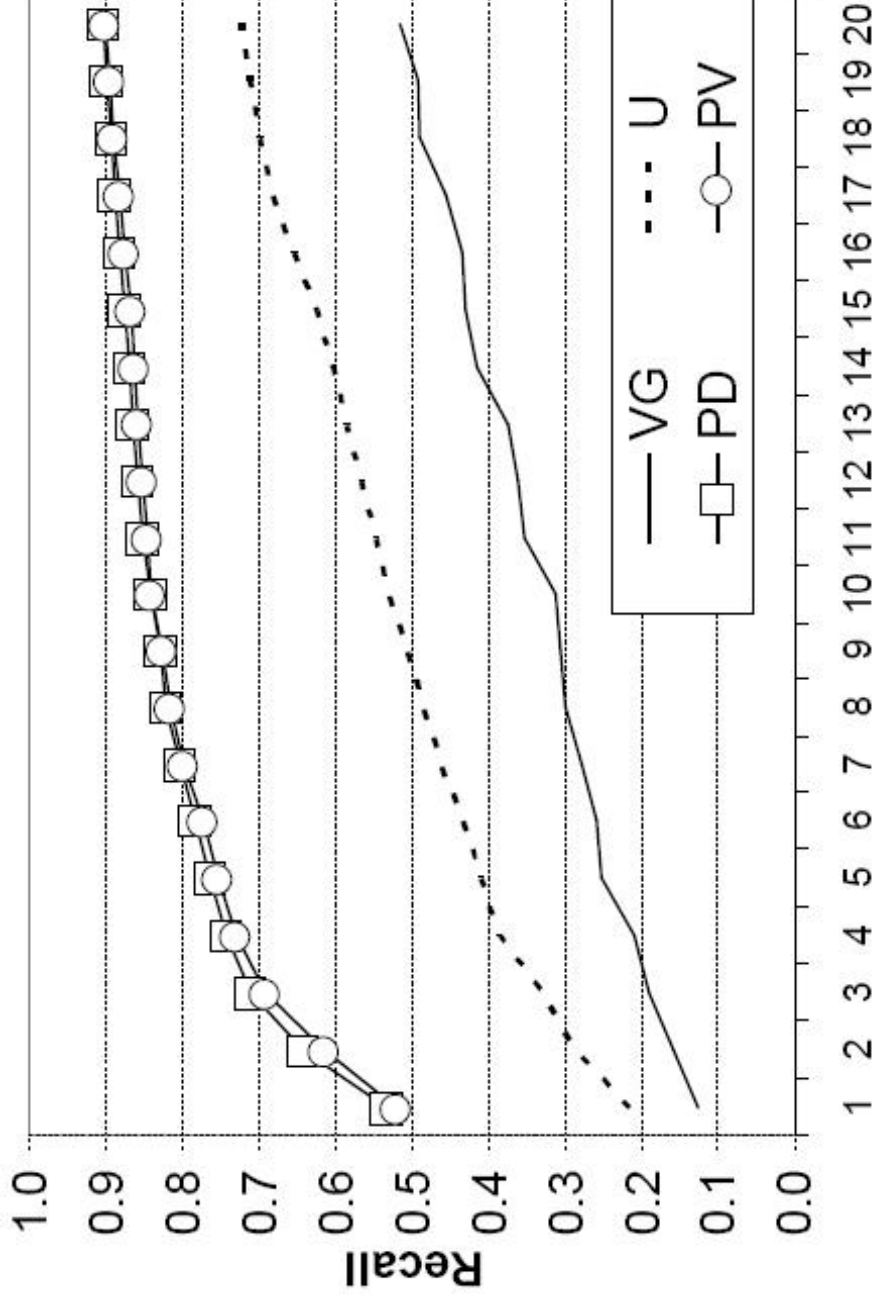
Higher-quality samples should support smarter database selection

Using  
CORI DB  
Selection

# Database Selection Recall TREC123A

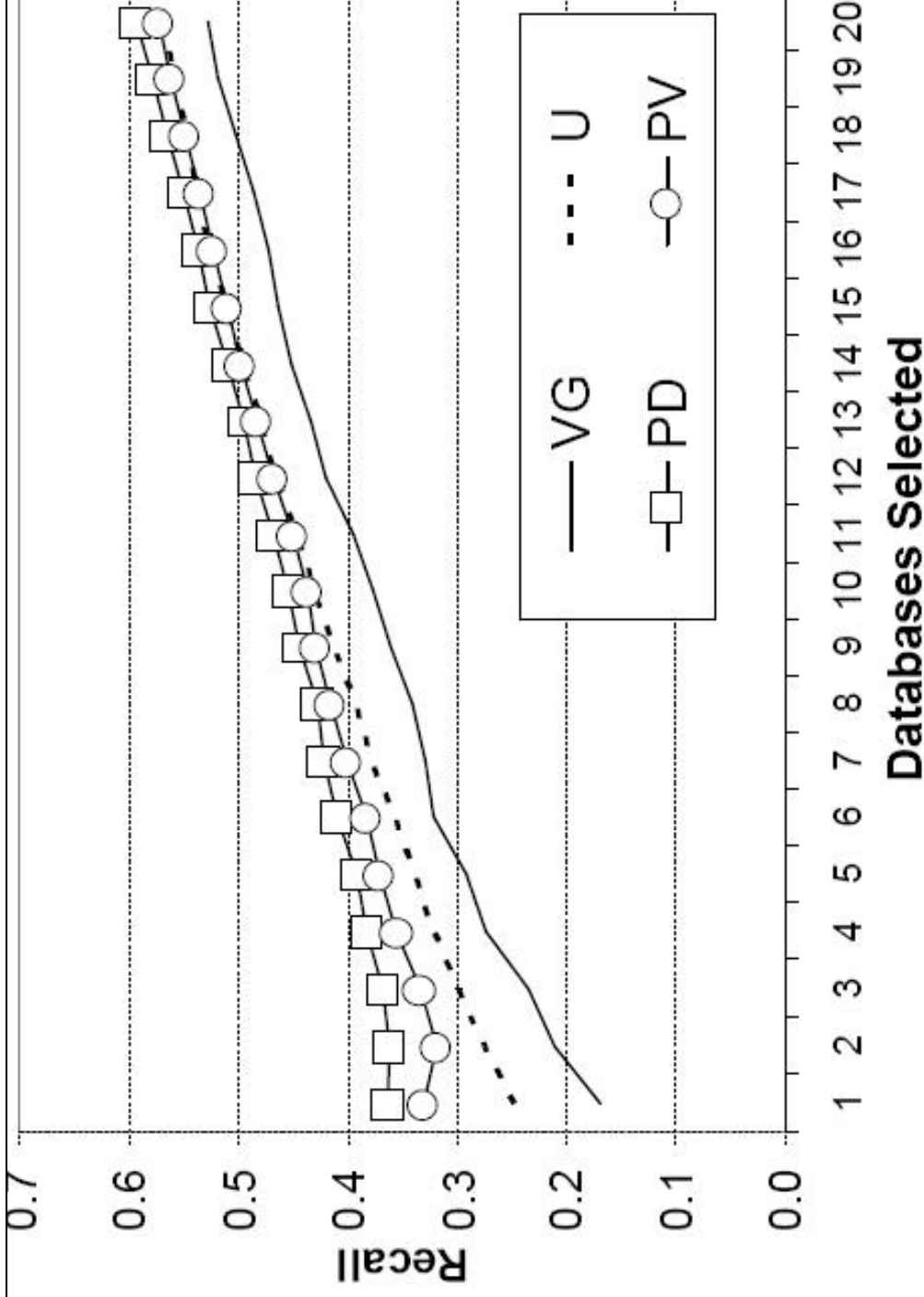
**Best case:** biggest DBs are **most relevant** to query mix

**Recall vs.  
Baseline  
Ranking  
[by # of  
relevant  
docs]**



# Database Selection Recall TREC123B

**Worst case:** biggest DBs are **least relevant** to query mix



# Contributions

- *An adaptive distributed query-sampling framework that is **quality-conscious** for extracting high-quality text database samples*
- **Relies self-configuring** ability based on the overall quality of all text databases under consideration
- **Three quality-conscious sampling schemes** for estimating database quality
- **Higher-quality** document sampling over multiple metrics compared to existing approaches

# Future Work

- Alternative types of quality
  - Novelty, Freshness, Topic-sensitivity
- Augment with other info
  - Web: link data, access patterns
  - P2P: peer longevity, reputation
- Non-uniform sampling costs

Thank You!

[caverlee@cc.gatech.edu](mailto:caverlee@cc.gatech.edu)

[www.caverlee.com](http://www.caverlee.com)