Federated Search

Jaime Arguello
INLS 509: Information Retrieval
jarguell@email.unc.edu

November 21, 2016

Up to this point...

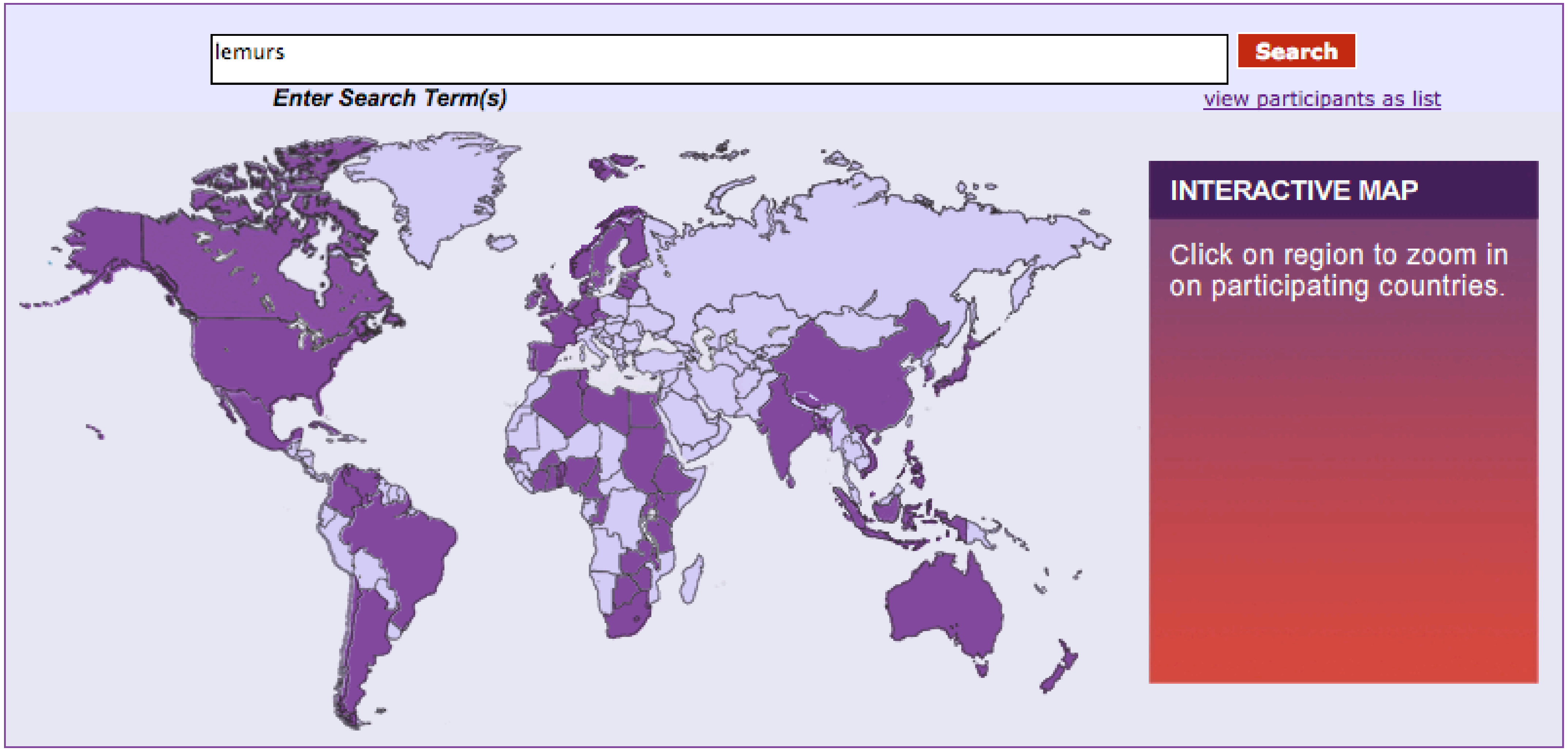
- Classic information retrieval
  - search from a single centralized index
  - all queries processed the same way
- Federated search
  - search across <u>multiple distributed collections</u>
  - a.k.a: resources, search engines, search services, etc.

Motivation

- Some content cannot be crawled and centrally indexed (exposed only via a search interface)
    - also referred to as “the hidden web”
- Even if crawl-able, we may prefer searchable access to this content via the third-party search engine. why?
    - content updated locally
    - unique document representation (e.g., metadata)
    - customized retrieval

Federated Search Examples
(World Wide Science)

- Exhaustive search (across all collections)



Federated Search Examples

(World Wide Science)

Refine Search | lemurs | New Search | Advanced Search | by Deep Web Technologies

Search: **Full Record: lemurs**
301 ranked results of 1,625 available

Create an alert from this search

Summary of All Results

62 of **62** sources complete

Results **1** – **10** of 301 Sort by: Rank Limit to: All Sources 1 2 3 4 5

1 **Lemurs - Ambassadors for Madagascar**
Thalmann, Urs
Madagascar Conservation & Development 2006-01-01
PDF
Directory of Open Access Journals (Sweden)

2 **The dental comb of lemurs**
Roberts, D.
UK PubMed Central
Full Text Available
Digital Repository Infrastructure Vision for European Research (DRIVER)

3 **The Placentation of Lemurs**
Turner
UK PubMed Central
Full Text Available
Digital Repository Infrastructure Vision for European Research (DRIVER)

4 **Object permanence in lemurs.**
Deppe, Anja M.
Vascoda (Germany)

Federated Search Examples
(World Wide Science)

Refine Search | lemurs | New Search | Advanced Search

Search: **Full Record: lemurs**
301 ranked results of 1,625 available

Create an alert from this search

Summary of All Results

Results **1** – **10** of 301 Sort by: Rank Limit to: All Sources 1 2 3

1 **Lemurs - Ambassadors for Madagascar**
Thalmann, Urs
Madagascar Conservation & Development 2006-01-01
Directory of Open Access Journals (Sweden)

2 **The dental comb of lemurs**
Roberts, D.
UK PubMed Central
Digital Repository Infrastructure Vision for European Research (DRIVER)

3 **The Placentation of Lemurs**
Turner
UK PubMed Central
Digital Repository Infrastructure Vision for European Research (DRIVER)

4 **Object permanence in lemurs.**
Deppe, Anja M.
Vascoda (Germany)

Summary of All Results for this Search

National Library of Latvia	✓	3
National Library of the Czech Republic Manuscriptorium	✓	0
Nepal Journals Online (Nepal)	✓	0
Norwegian Open Research Archives (NORA)	✓	0
OpenSIGLE	✓	9
Philippines Journals Online (Philippines)	✗	0
Science.gov (United States)	✓	100
Scientific Electronic Library Online (Argentina)	✓	0
Scientific Electronic Library Online (Brazil)	✓	0
Scientific Electronic Library Online (Chile)	✓	0
Scientific Electronic Library Online (Colombia)	✓	0
Scientific Electronic Library Online (Cuba)	✓	0
Scientific Electronic Library Online (Mexico)	✓	0
Scientific Electronic Library Online (Portugal)	✓	0
Scientific Electronic Library Online (Spain)	✓	0
Scientific Electronic Library Online (Venezuela)	✓	0

Federated Search Examples
(World Wide Science)

Federated Search Examples
(Vertical Aggregation in Web Search)

pittsburgh | Search

maps

[Pittsburgh, PA](#) maps.google.com

web

[City of **Pittsburgh**, Pennsylvania - Pghgov.com](#)
Official city site including information on economic development, resident information, links, tourism and contact information.
www.city.**pittsburgh**.pa.us/ - Cached - Similar

images

[Images for **pittsburgh**](#) - Report images

web

[**Pittsburgh** - Wikipedia, the free encyclopedia](#)
Pittsburgh is the second-largest city in the U.S. Commonwealth of Pennsylvania and the county seat of Allegheny County. Regionally, it anchors the largest ...
History of Pittsburgh - Neighborhoods - List of people from the Pittsburgh ... - 1936
en.wikipedia.org/wiki/**Pittsburgh** - Cached - Similar

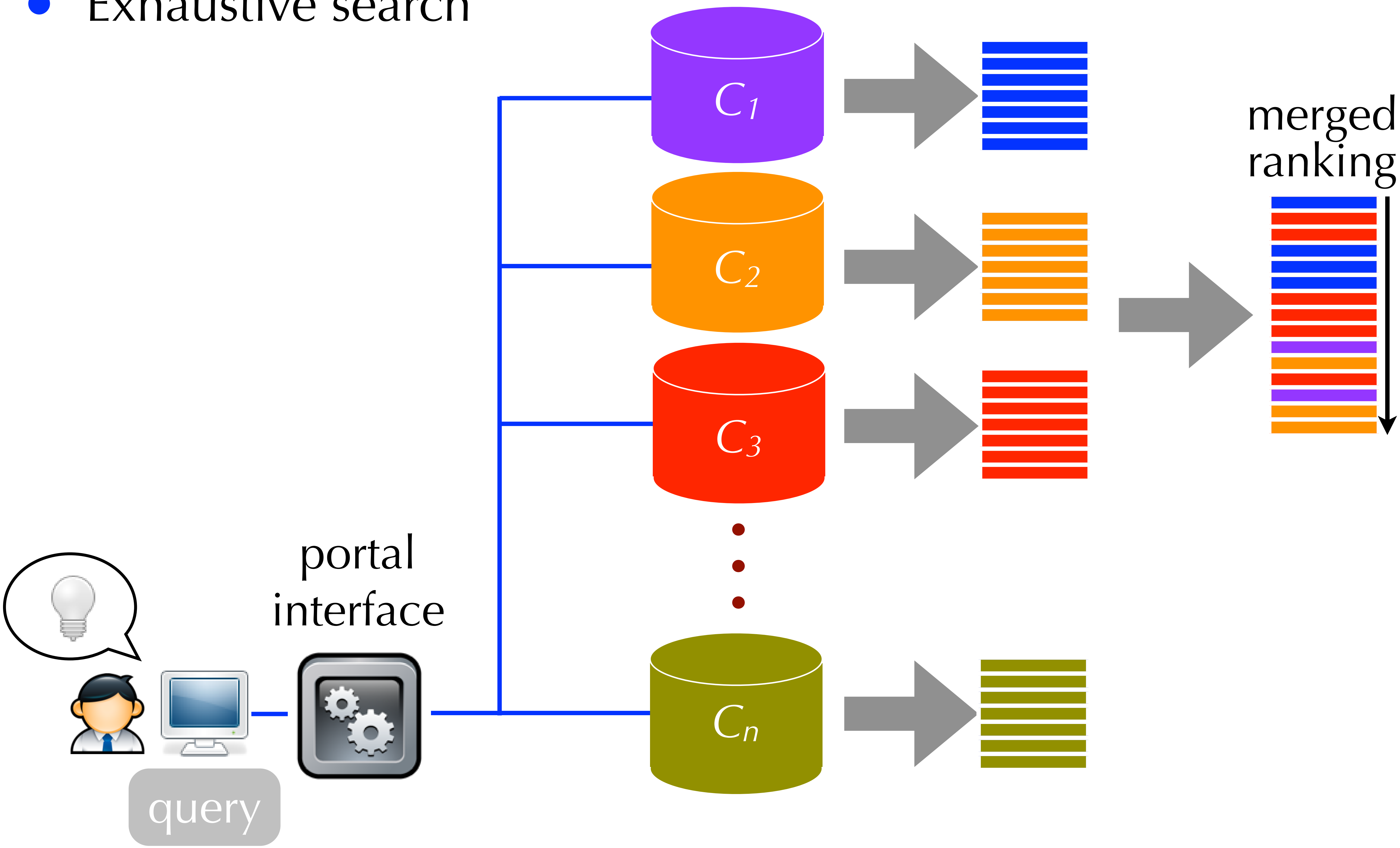
books

[Books for **pittsburgh**](#)

[**Pittsburgh**: a sketch of its early social life](#) - Charles William Dahlinger - 1916 - 216 pages
[**Pittsburgh**:: 1758-2008](#) - Pittsburgh Post-Gazette, Carnegie Library of Pittsburgh - 2008 - 128 pages
Pittsburgh: 17582008 surveys the citys evolution from strategic fort in the wilderness ...
books.google.com

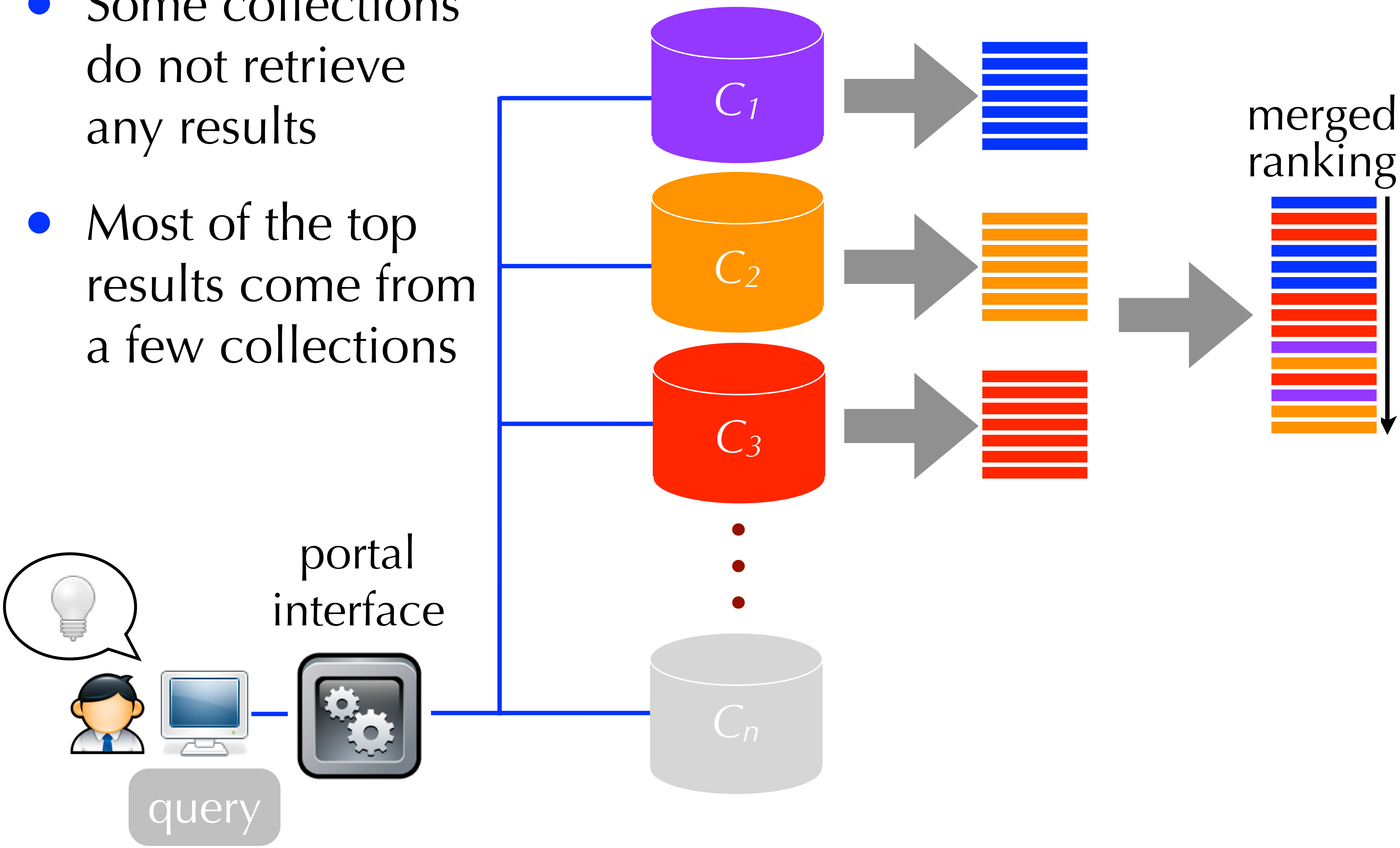
Federated Search

- Exhaustive search



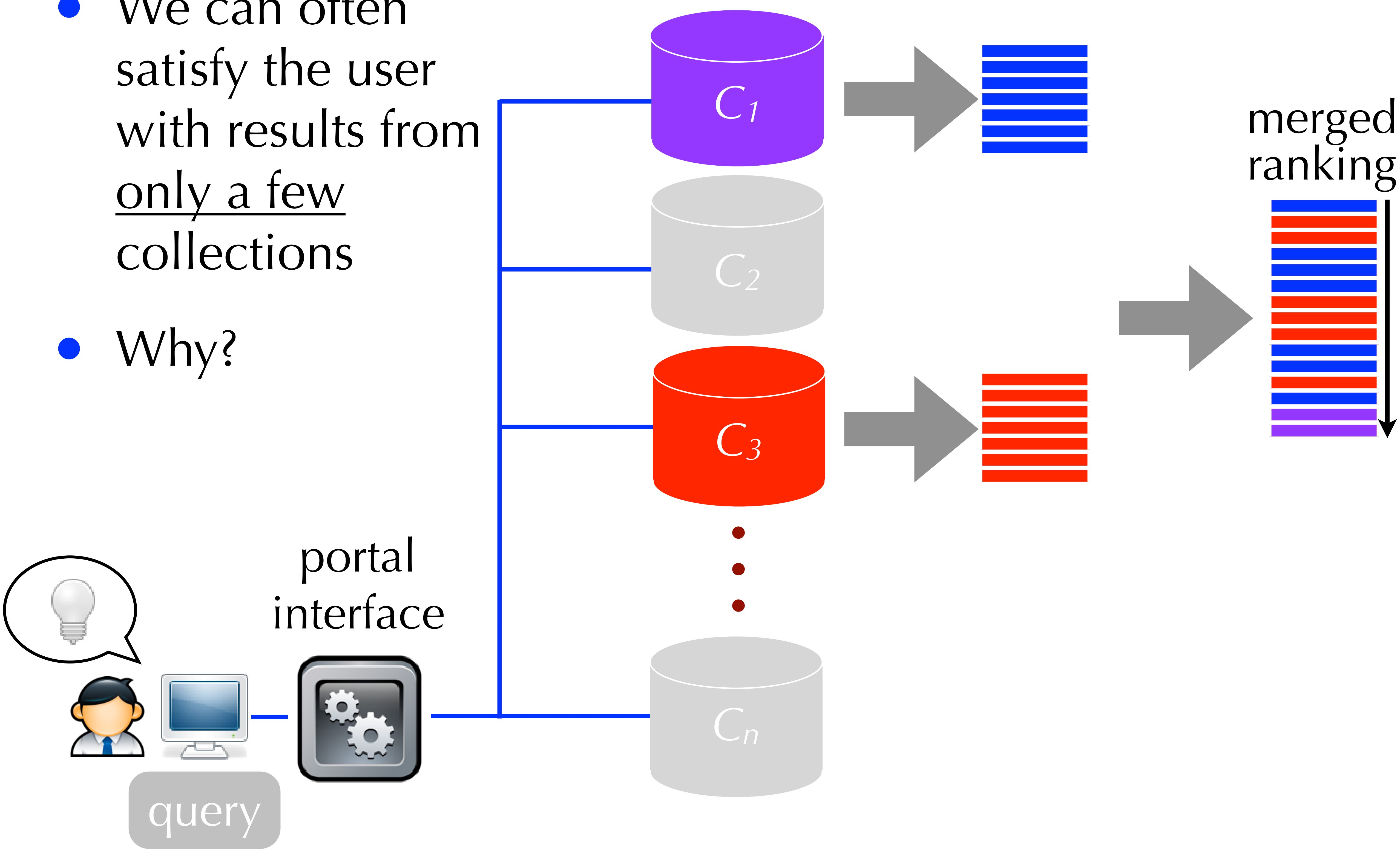
Federated Search

- Some collections do not retrieve any results
- Most of the top results come from a few collections



Federated Search

- We can often satisfy the user with results from <u>only a few</u> collections
- Why?



The Cluster Hypothesis
(van Rijsbergen, 1979)

- Similar documents are relevant to similar information needs
  - used in cluster-based retrieval
  - document score normalization
  - pseudo-relevance feedback
  - federated search

Federated Search

- Objective: given a query, predict which few collections have relevant documents and combine their results into a single document ranking

Federated Search

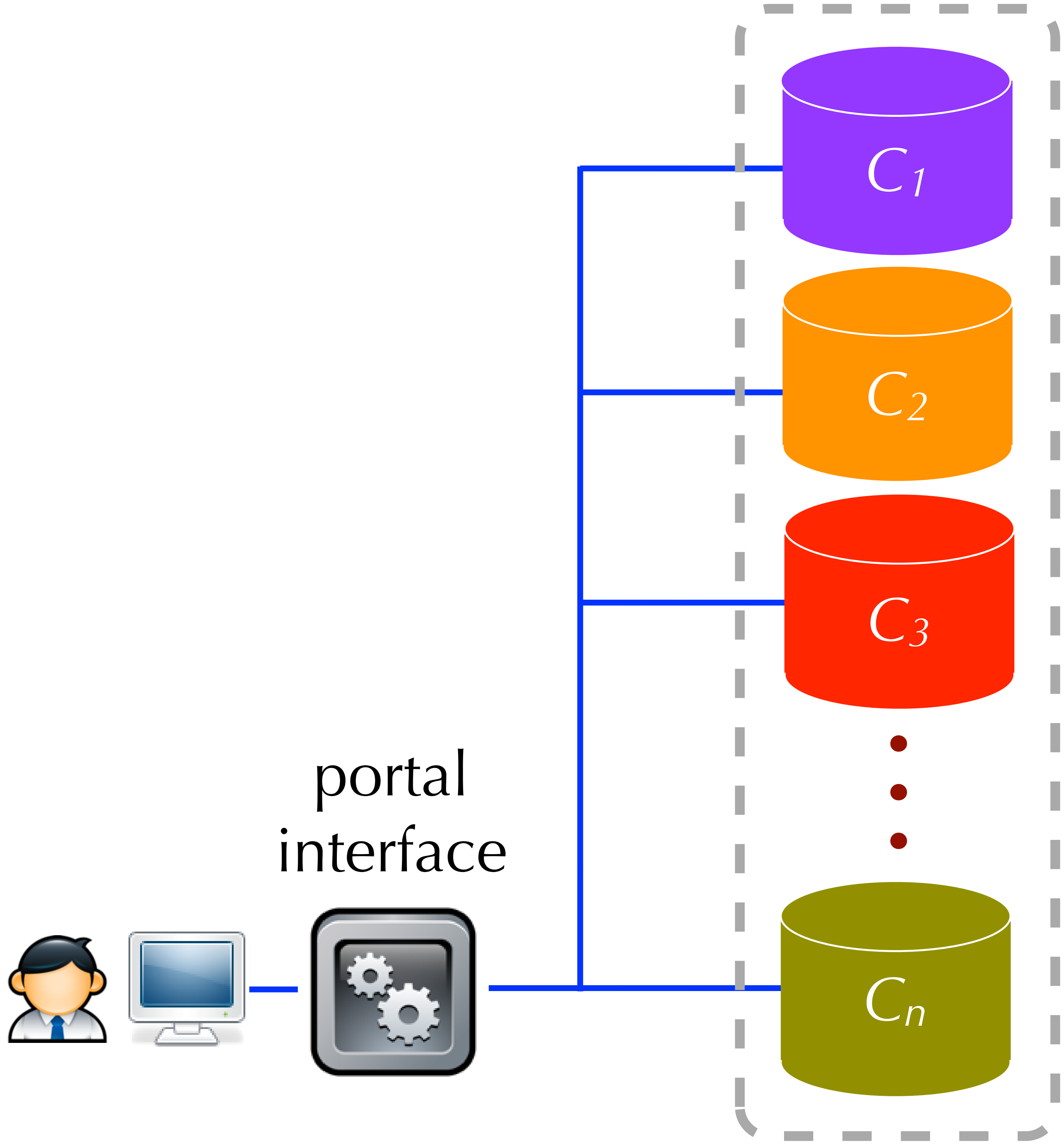
Resource representation

Resource selection

Results merging

Federated Search

Resource Representation

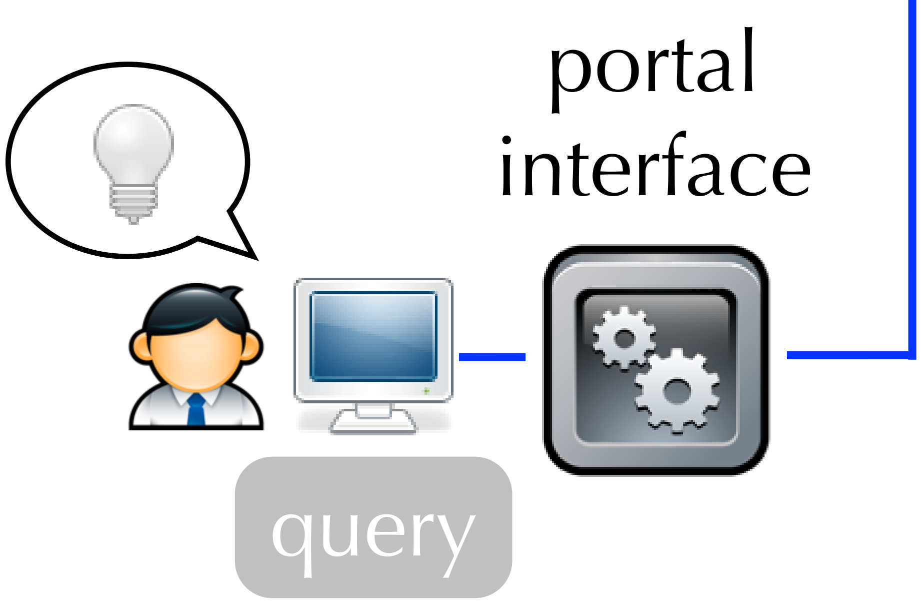


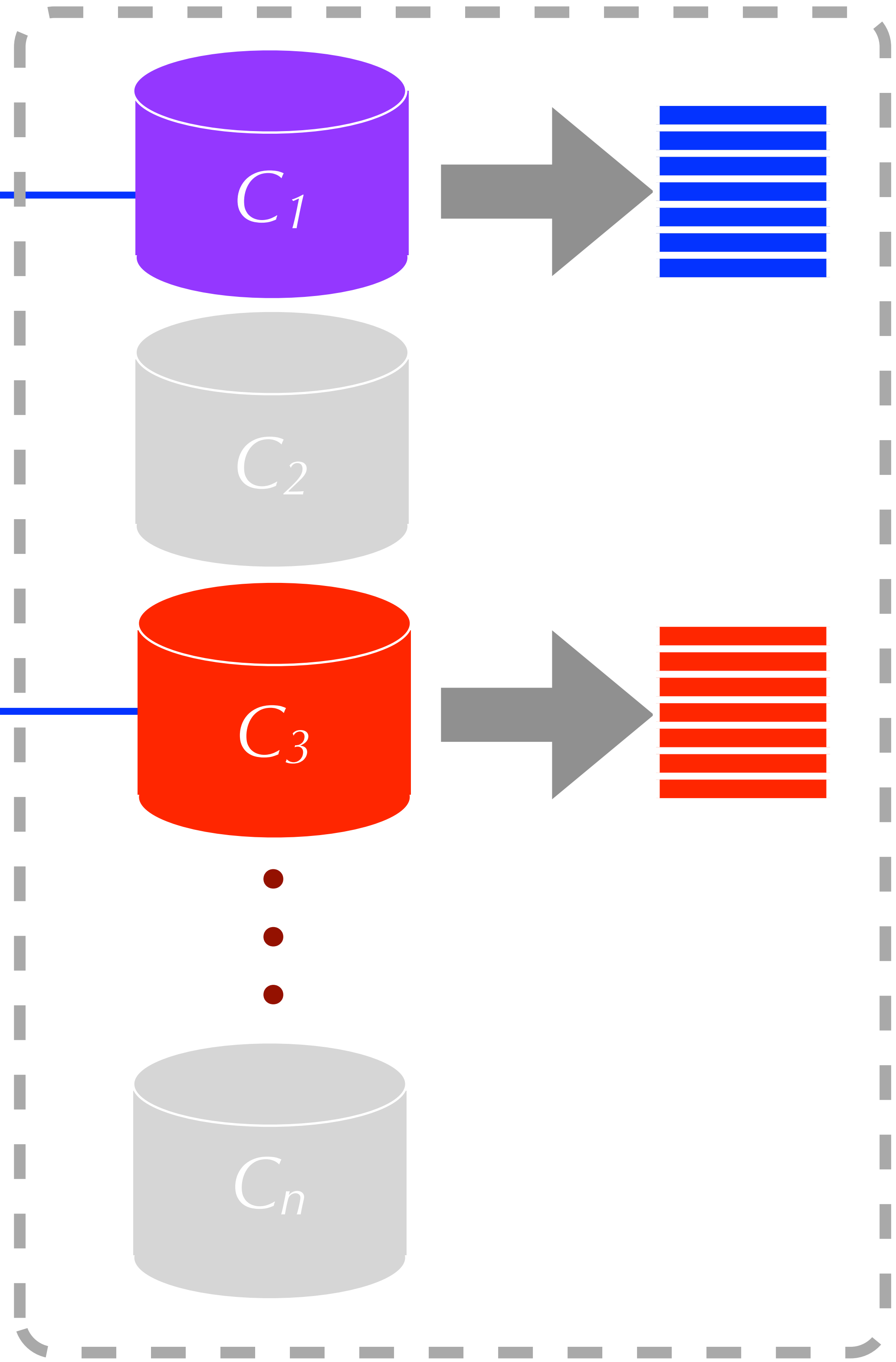
- Gathering information about what each resource contains
- What types of information needs does each resource satisfy?

Federated Search

Resource Selection

- Deciding which few resources to search given a user's query

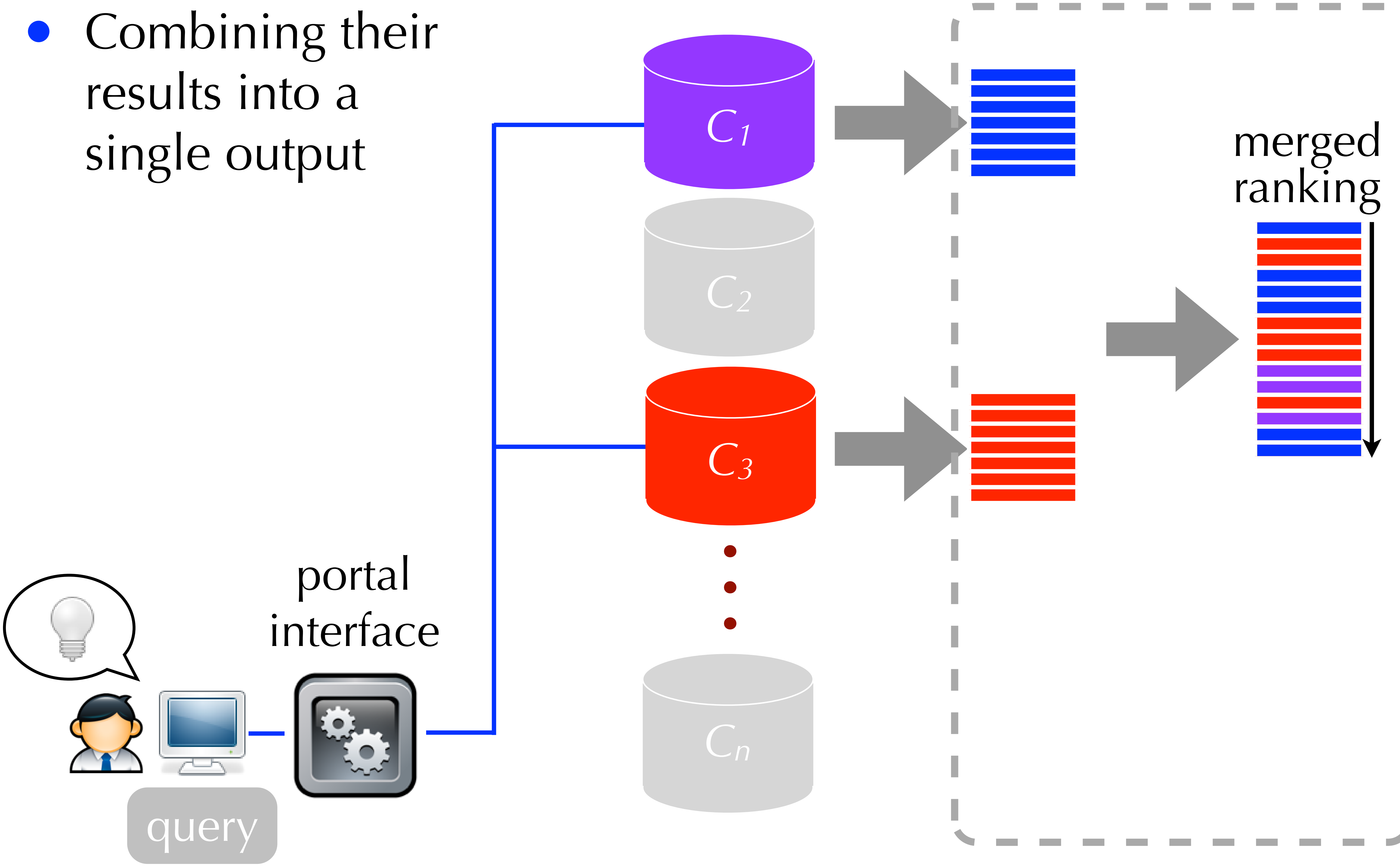




Federated Search

Results Merging

- Combining their results into a single output



Federated Search

off-line

resource representation

resource selection

results merging

at query-time

Cooperative vs. Uncooperative

- Cooperative environment
  - assumption: resources provide accurate and complete information to facilitate selection and merging
  - centrally designed protocols and APIs
- Uncooperative environment
  - assumption: resources provide no special support for federated search
  - only a search interface
- Different environments require different solutions

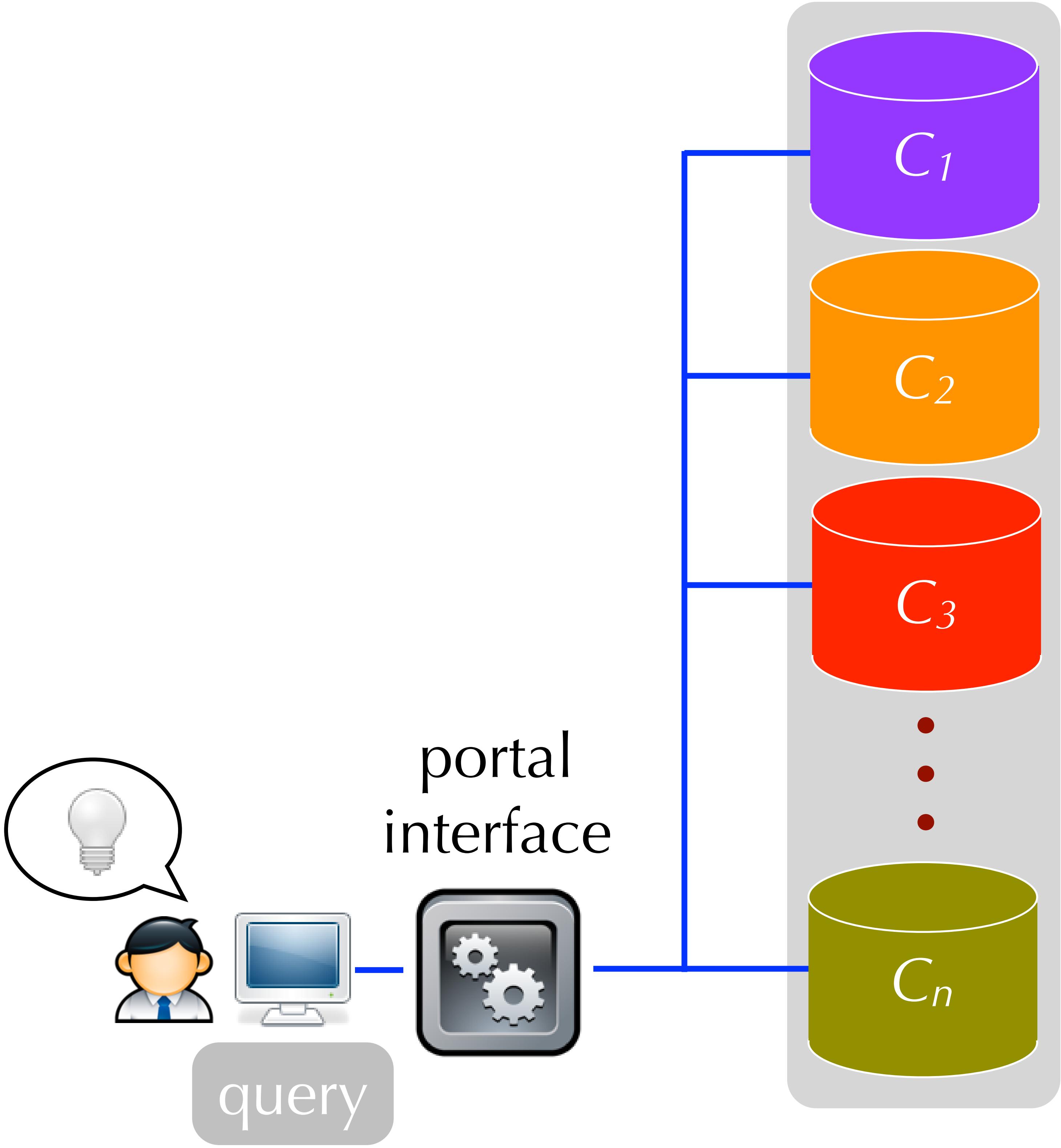
Resource Representation

Resource Representation

- Objective: to gather information about what each resource contains
  - but, ultimately to inform resource selection
- Discussion: what sources of evidence could we use to do this?

Resource Representation

using content

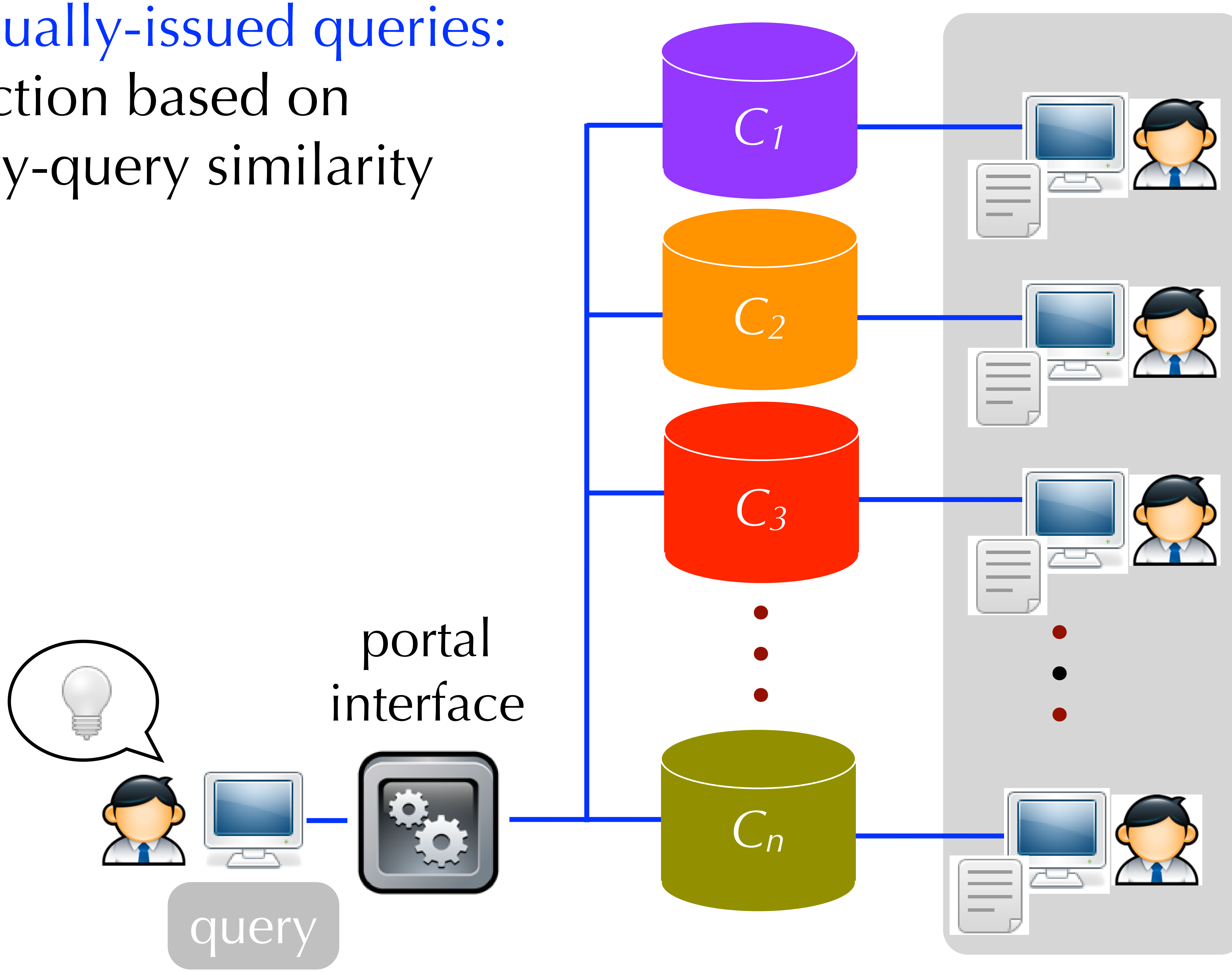


- **Term frequencies:** selection based on the query-collection similarity
- **A set of "typical" docs:** selection based on the predicted relevance of sampled documents

Resource Representation

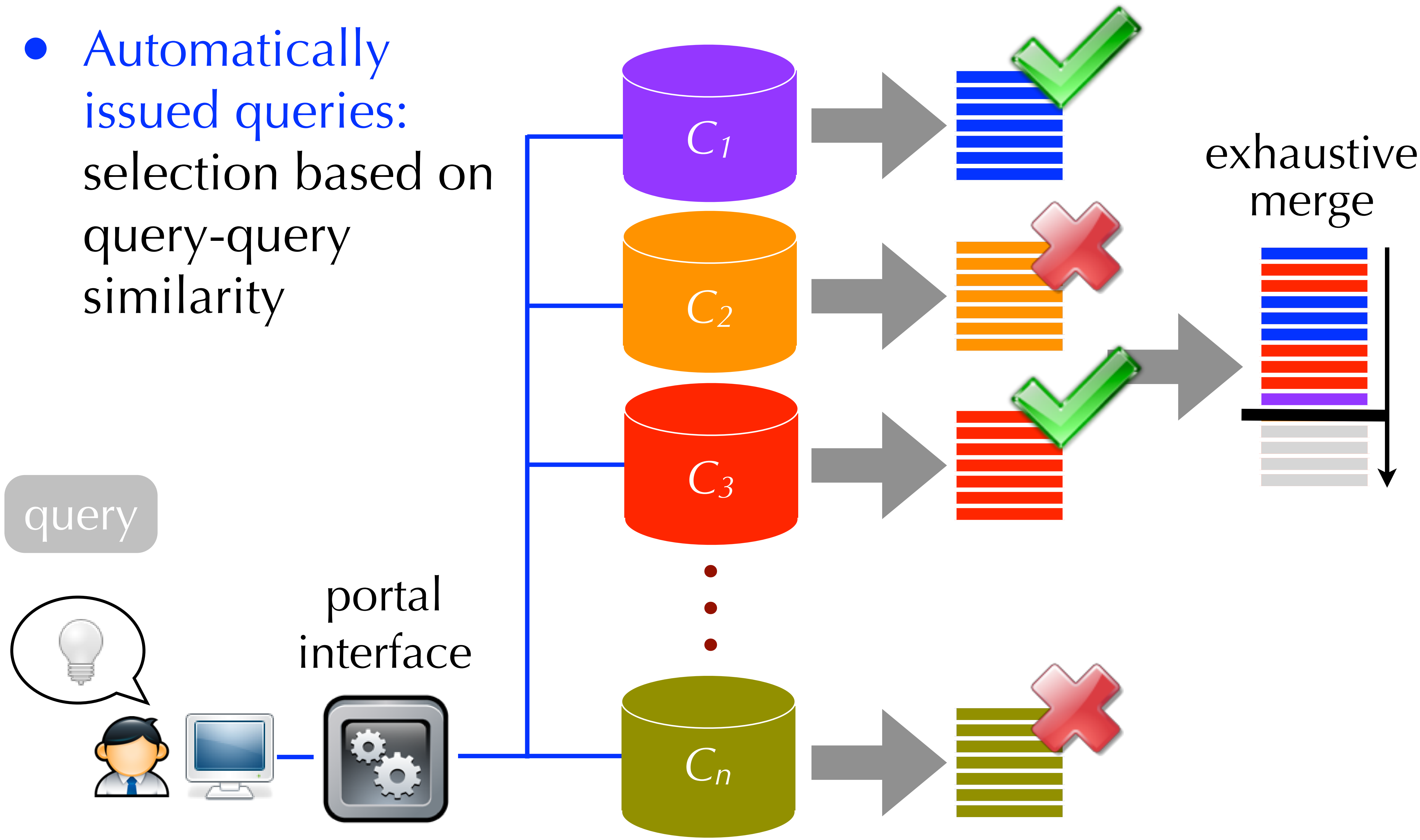
using manually-issued queries

- Manually-issued queries: selection based on query-query similarity



Resource Representation
using previous retrievals

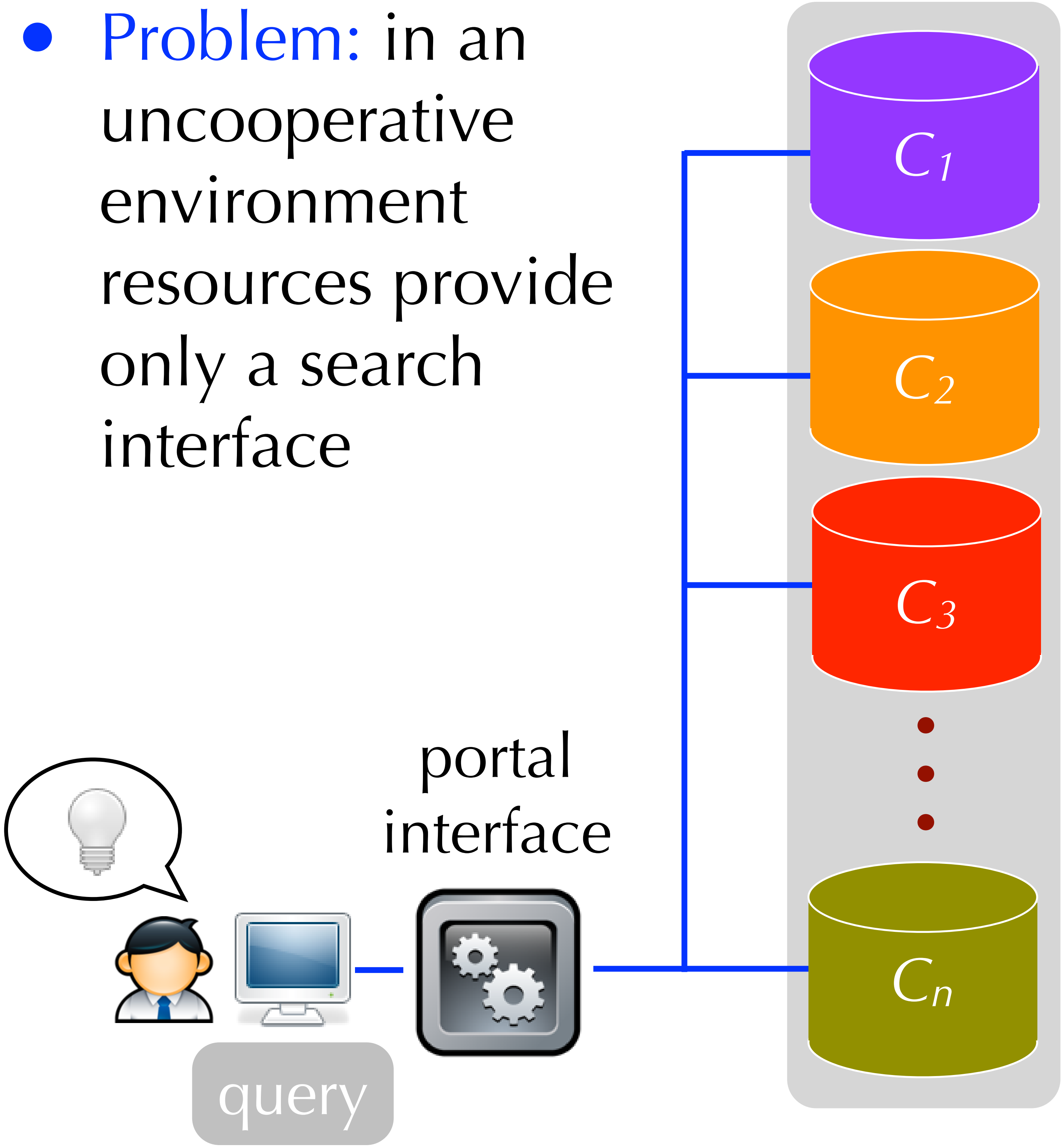
- Automatically issued queries: selection based on query-query similarity



Resource Representation

using content

- **Problem:** in an uncooperative environment resources provide only a search interface



- **Term frequencies:** selection based on the query-collection similarity

- **A set of 'typical' docs:** selection based on the predicted relevance of sampled documents

Query-based Sampling
(Callan and Connell, 2001)

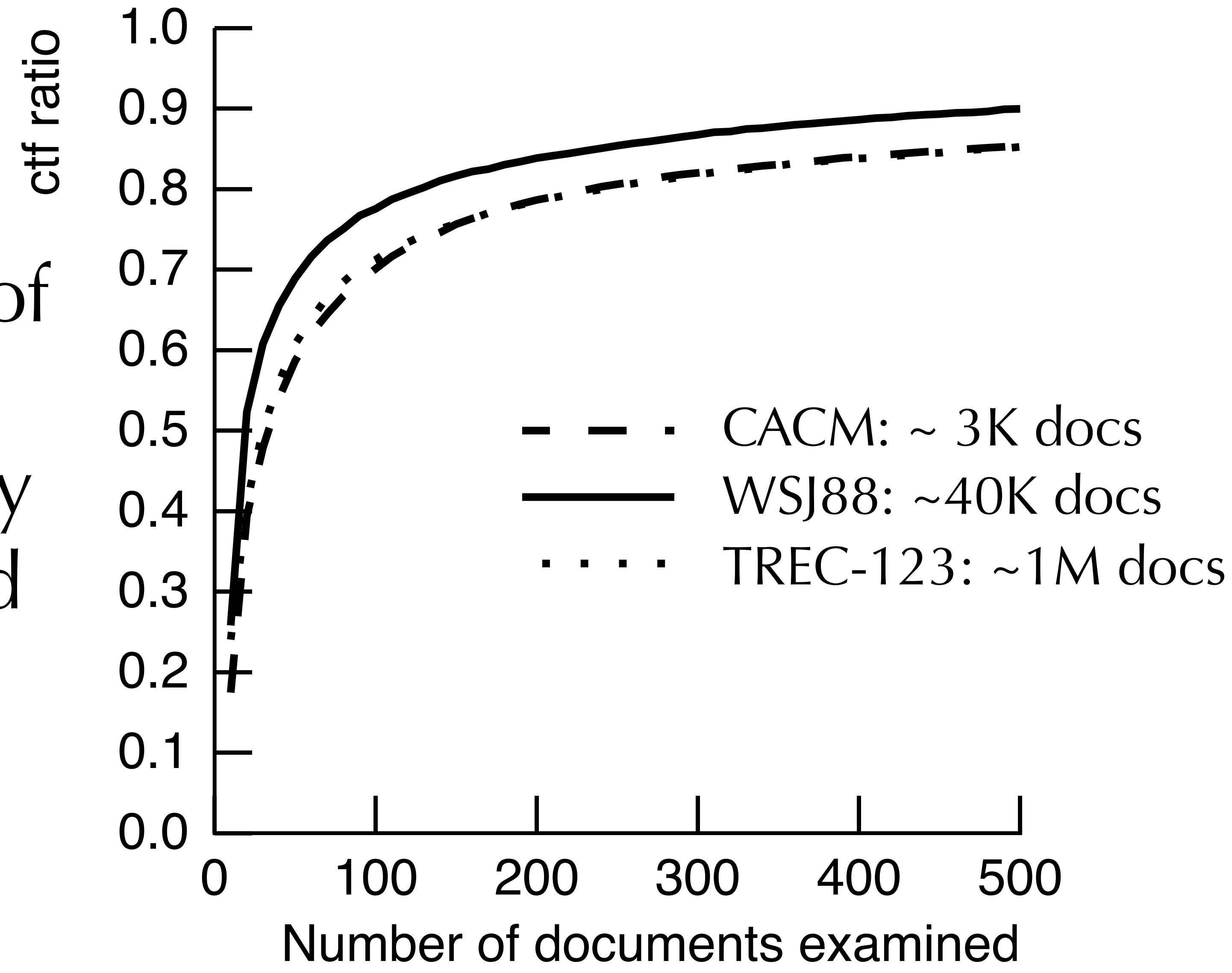
- Repeat *N* times (e.g., *N*=100),
  1. submit a query to the search engine
  2. download a few results (e.g., 4)
  3. update the collection representation (e.g., term frequencies)
  4. select a new query for sampling (e.g., from the emerging representation)

Query-based Sampling

- Discussion: suppose we want to represent resources using term frequency information, how many samples do we need?
- Hint: zipf's law states that the number of <u>new</u> terms seen in each additional document decreases exponentially

Query-based Sampling
(Callan and Connell, 2001)

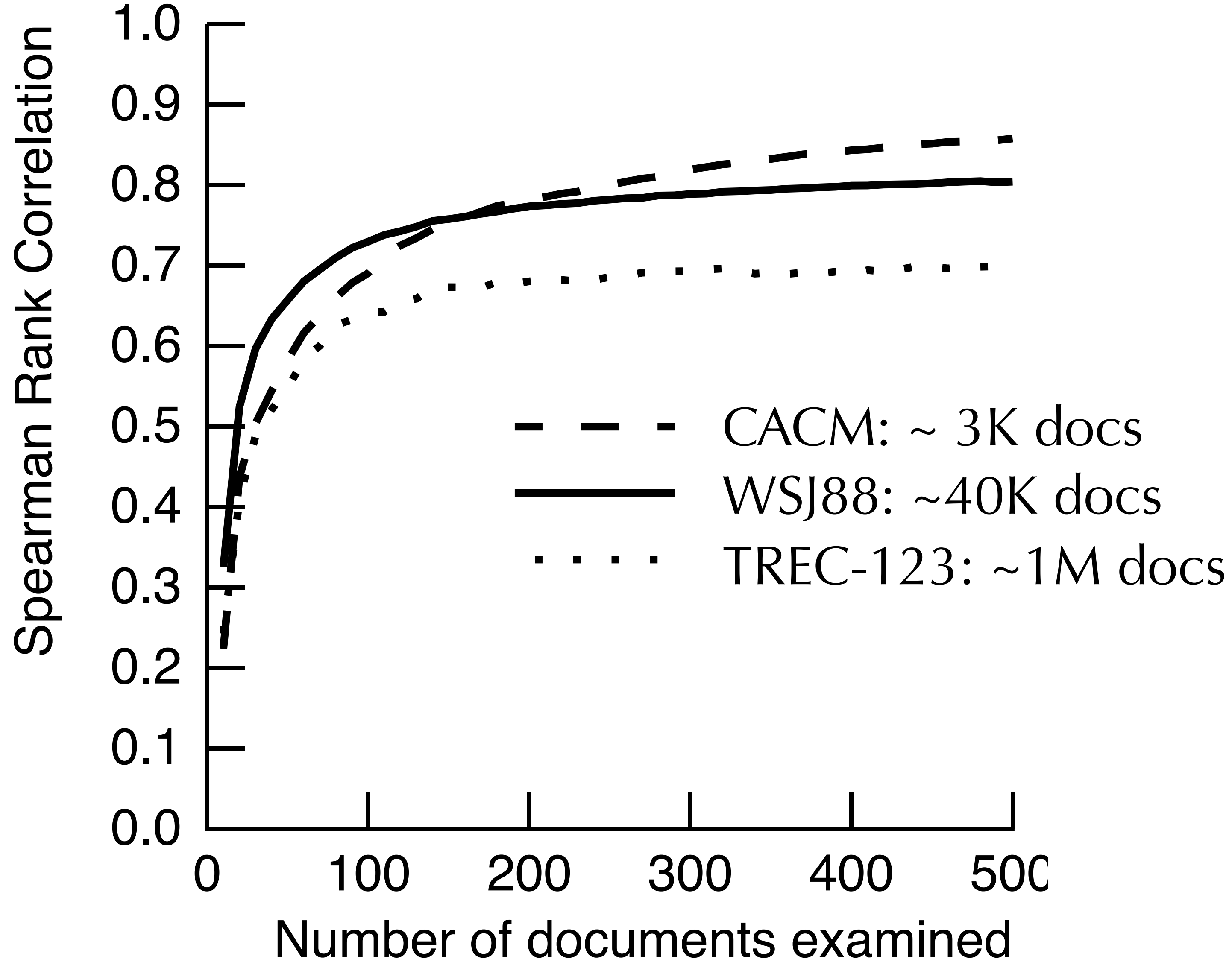
ctf ratio: % of collection "covered" by the observed terms



(a)

- After 500 docs we've seen enough vocabulary to account for about 80-90% all term occurrences

Query-based Sampling
(Callan and Connell, 2001)



(b)

- The ordering of terms (by frequency) based on sample set statistics approximates the actual one

Query-based Sampling
Extensions

- Adaptive sampling: sample until rate of unseen terms decreases below threshold (Shokouhi *et al.*, 2006)
    - slight improvement
- Sampling using (popular) query-log queries
    - web query-log (Shokouhi *et al.*, 2007), resource-specific query-log (Arguello *et al.*, 2009)
- Re-sampling to avoid stale representations
    - re-sampling according to collection size is a good heuristic (Shokouhi *et al.*, 2007b)

Resource Selection

Resource Selection

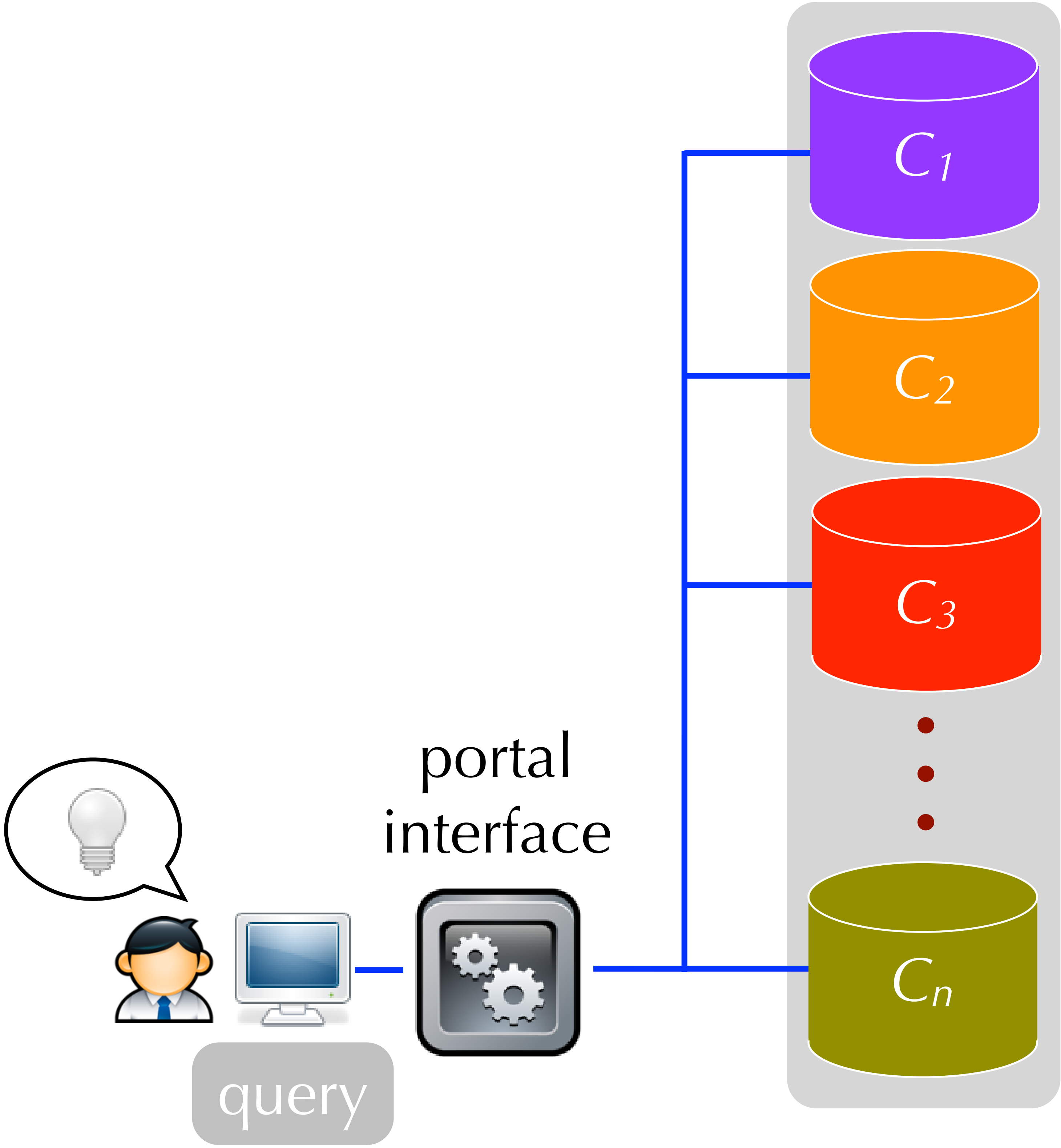
- Objective: deciding which resources to search given a user's query
- Most prior work casts the problem as resource ranking
  - given a query, select the $k \ll n$ collections that produce good merged results
  - k is given (an interesting research problem)

Resource Selection

- Content-based methods: score resources based on the similarity between the query and content from the resource
  - large vs. small document models
- Query-similarity methods: score resources based on the effectiveness of previously issued queries that are similar to the query (will be covered at high level)

Resource Representation

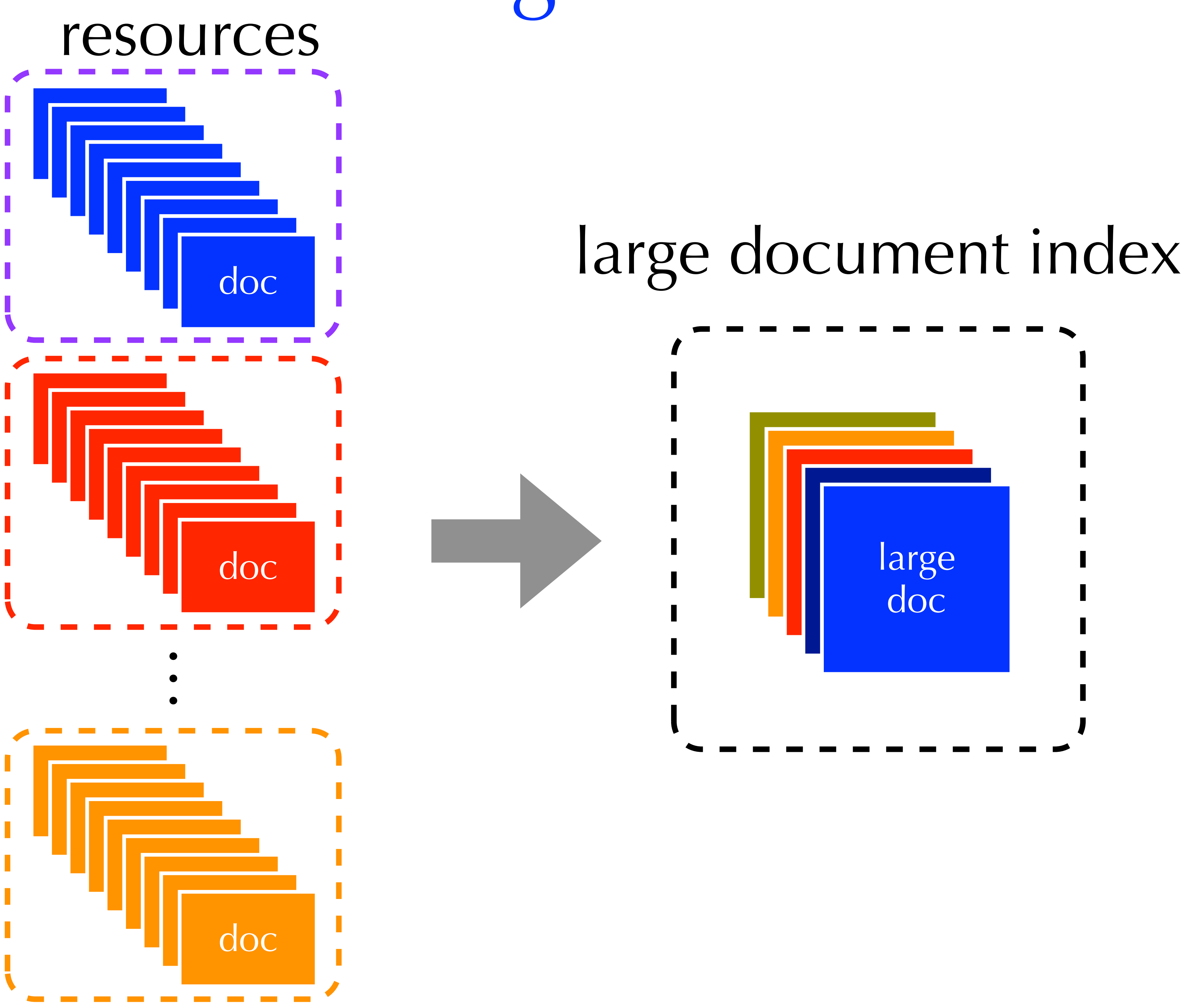
using content



- Term frequencies: selection based on the query-collection similarity

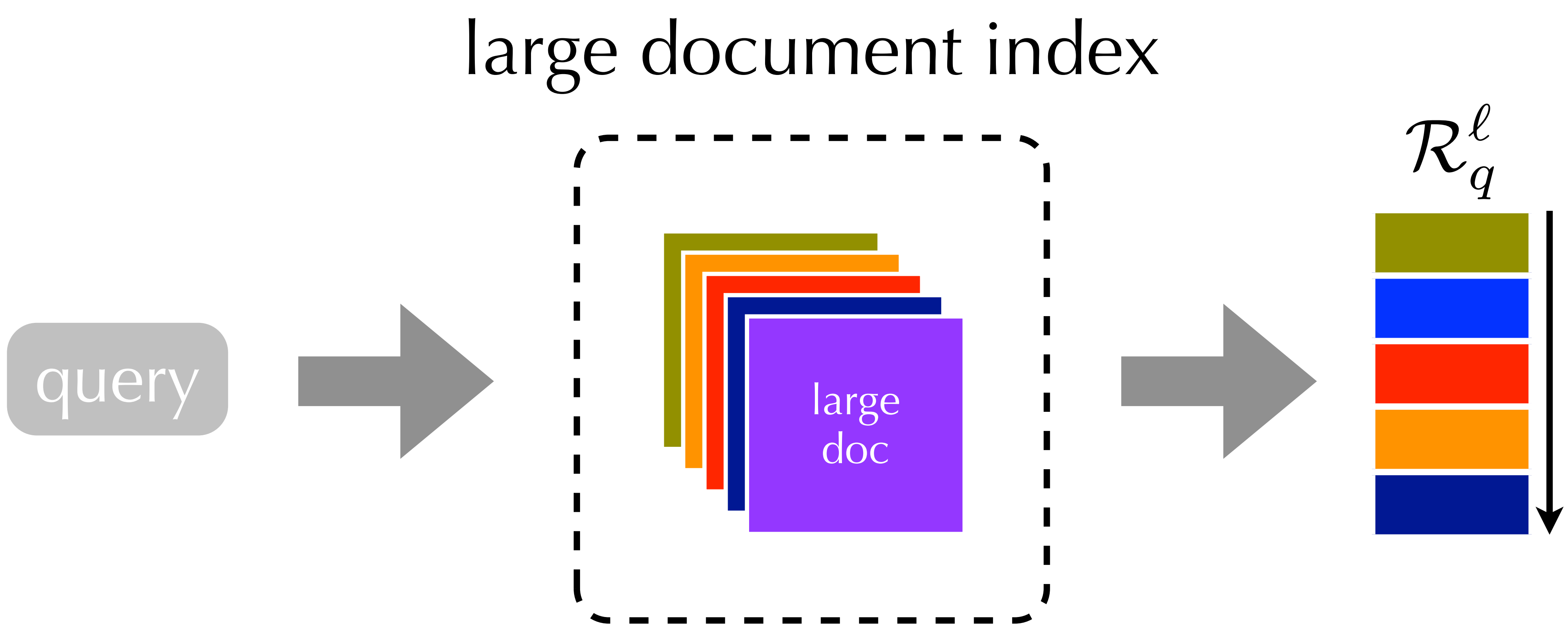
- A set of 'typical' docs: selection based on the predicted relevance of sampled documents

Large Document Models



- Represent each resource (or its samples) as a single "large document"

Large Document Models



1. Given the query, rank "large documents" using functions adapted from document retrieval
2. Select the top k

Large Document Models

- CORI (Callan, 1995)

$$\mathrm{CORI}_w(C_i) = b + (1-b) \times \frac{df_{w,i}}{df_{w,i} + 50 + 150 \times \frac{col\_len}{avg\_col\_len}} \times \frac{\log\left(\frac{|\mathcal{C}|+0.5}{cf_w}\right)}{\log(|\mathcal{C}| + 1.0)}$$

- adapted from BM25

$$P(w|d) = b + (1-b) \times \frac{tf}{tf + 0.5 + 1.5 \times \frac{doc\_len}{avg\_doc\_len}} \times \frac{\log\left(\frac{N+0.5}{df}\right)}{\log(N + 1.0)}$$

Large Document Models

- KL-Divergence (Xu and Croft 1999)

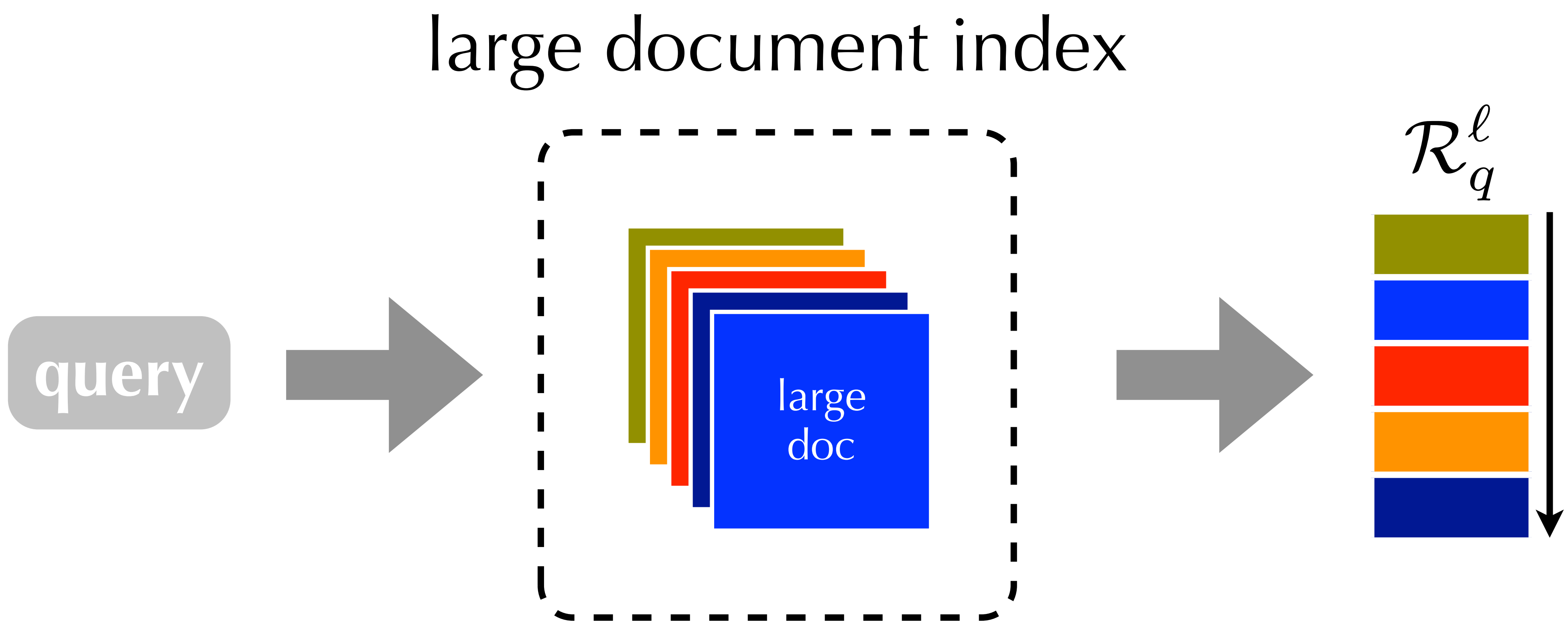
$$KL_q(C_i) = \sum_{w \in q} P(w|q) \log \left(\frac{P(w|q)}{P(w|C_i)} \right)$$

- Query Likelihood (Si *et al.*, 2002)

$$P(q|C_i) = \prod_{w \in q} \lambda P(w|C_i) + (1 - \lambda) P(w|G)$$

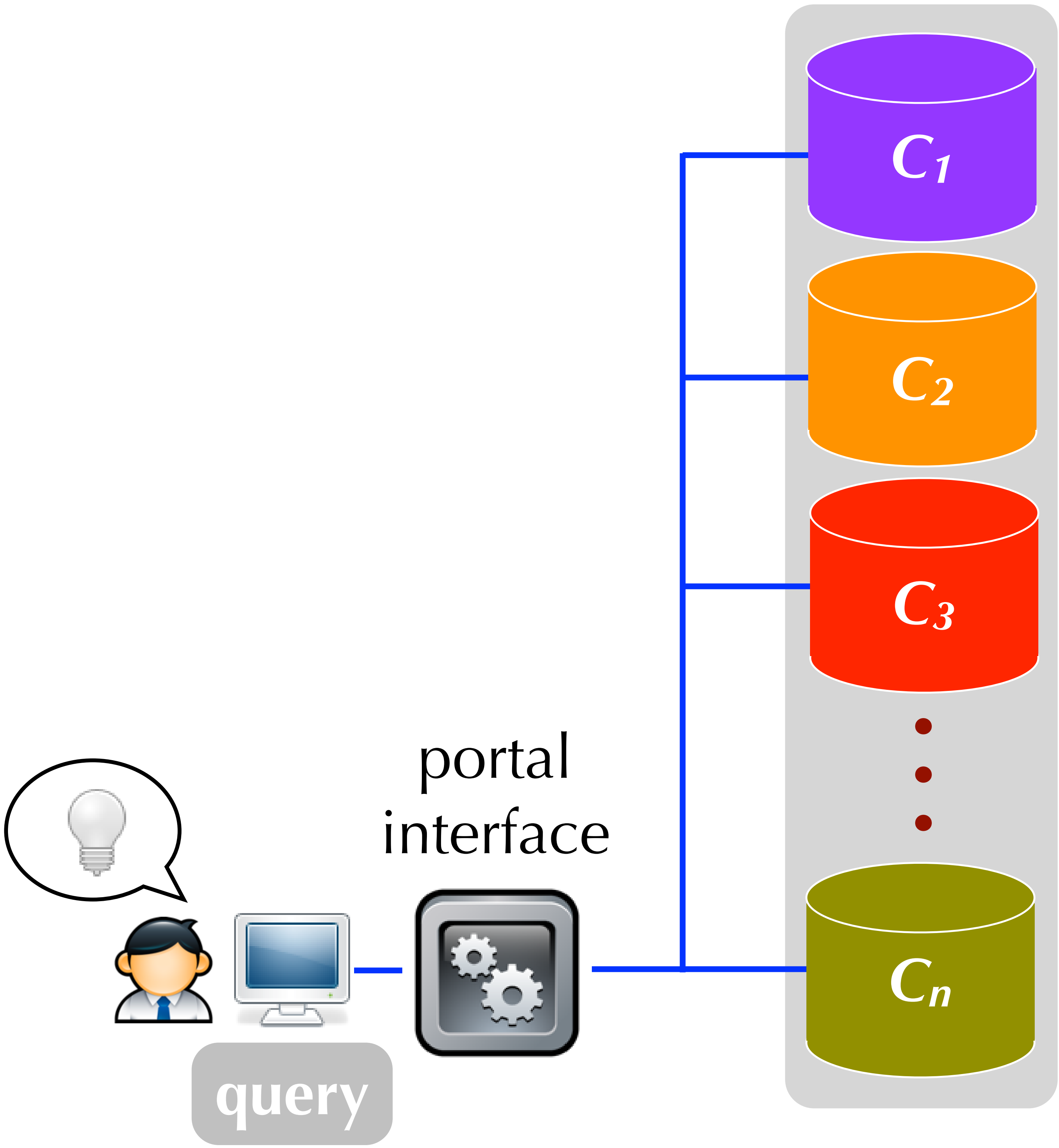
Large Document Models

- Discussion: potential limitations?



Resource Representation

using content

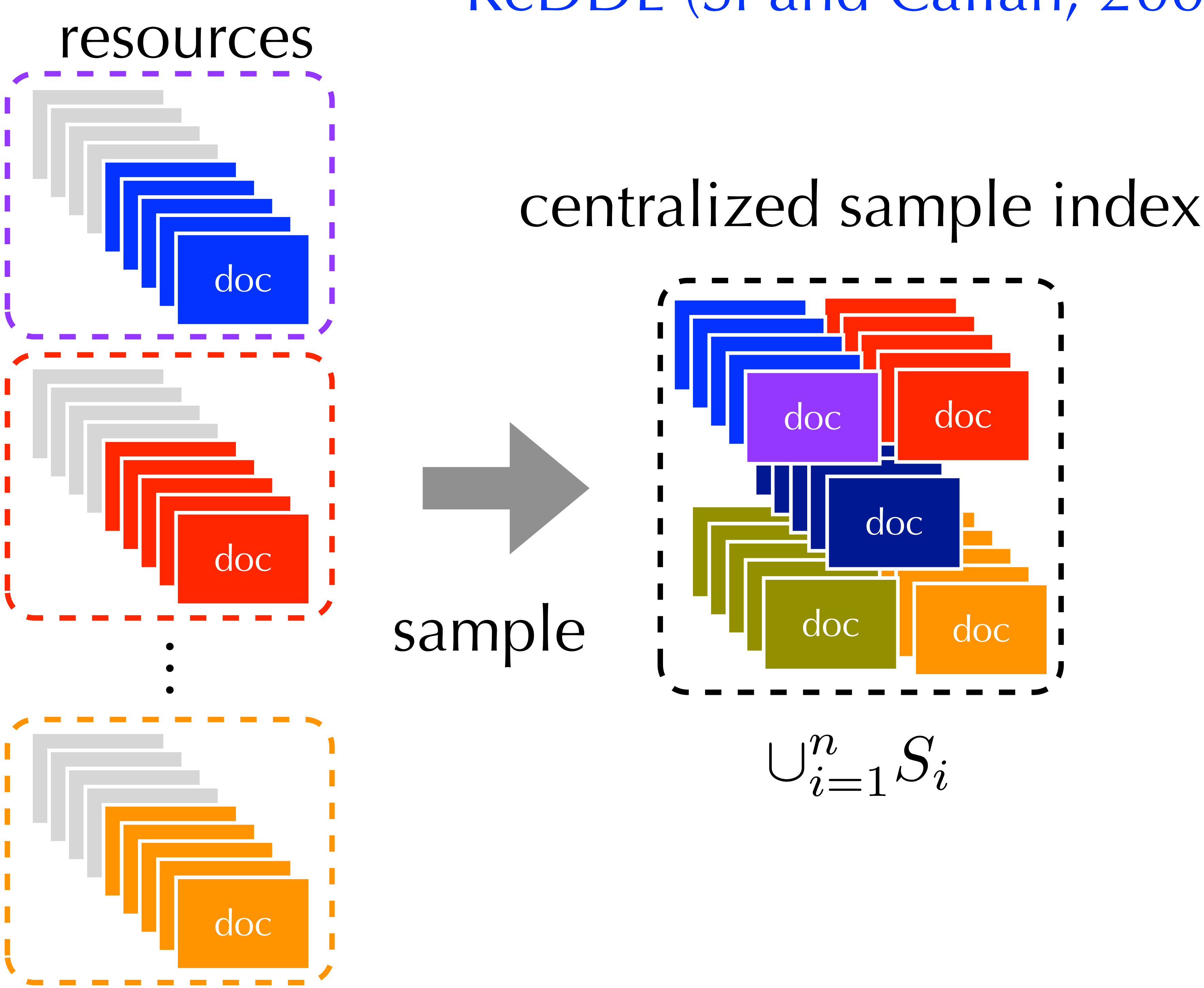


- Term frequencies: selection based on the query-collection similarity

- A set of 'typical' docs: selection based on the predicted relevance of sampled documents

Small Document Models

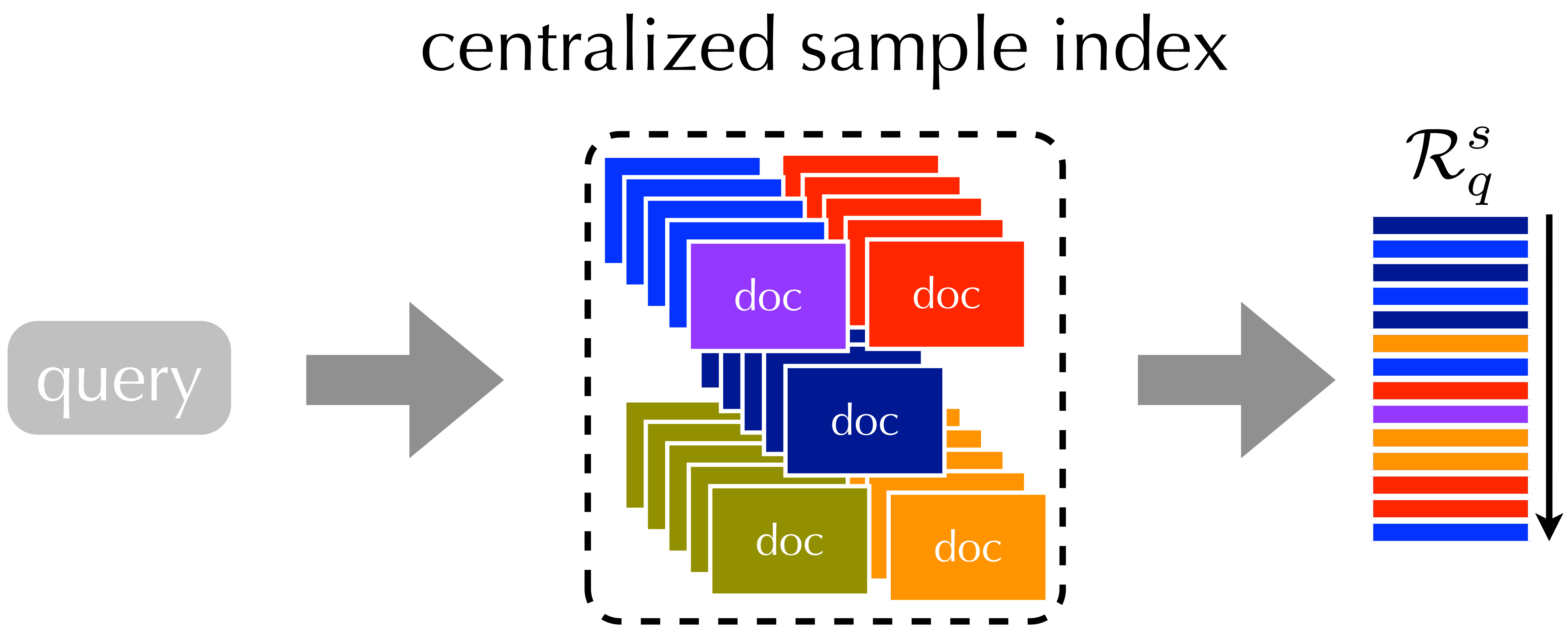
ReDDE (Si and Callan, 2003)



- Combine samples in a centralized index, keeping track of which collection each sample came from

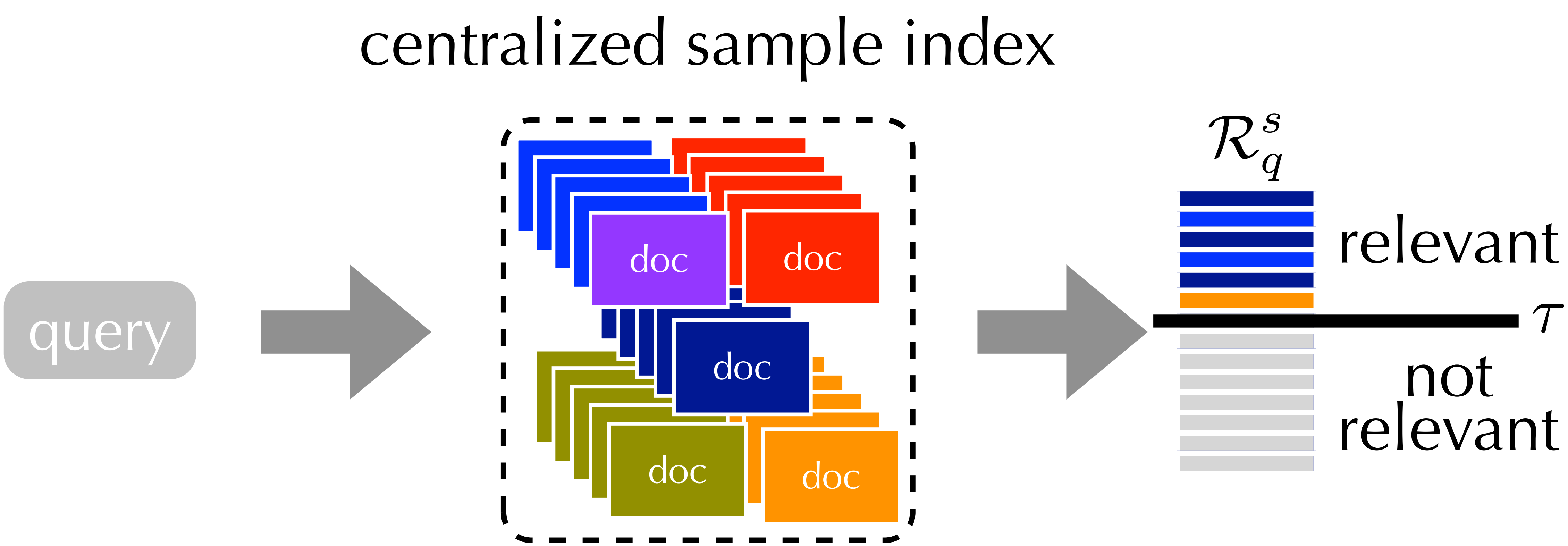
Small Document Models

ReDDE (Si and Callan, 2003)



- Given a query, conduct a retrieval from the centralized sample index

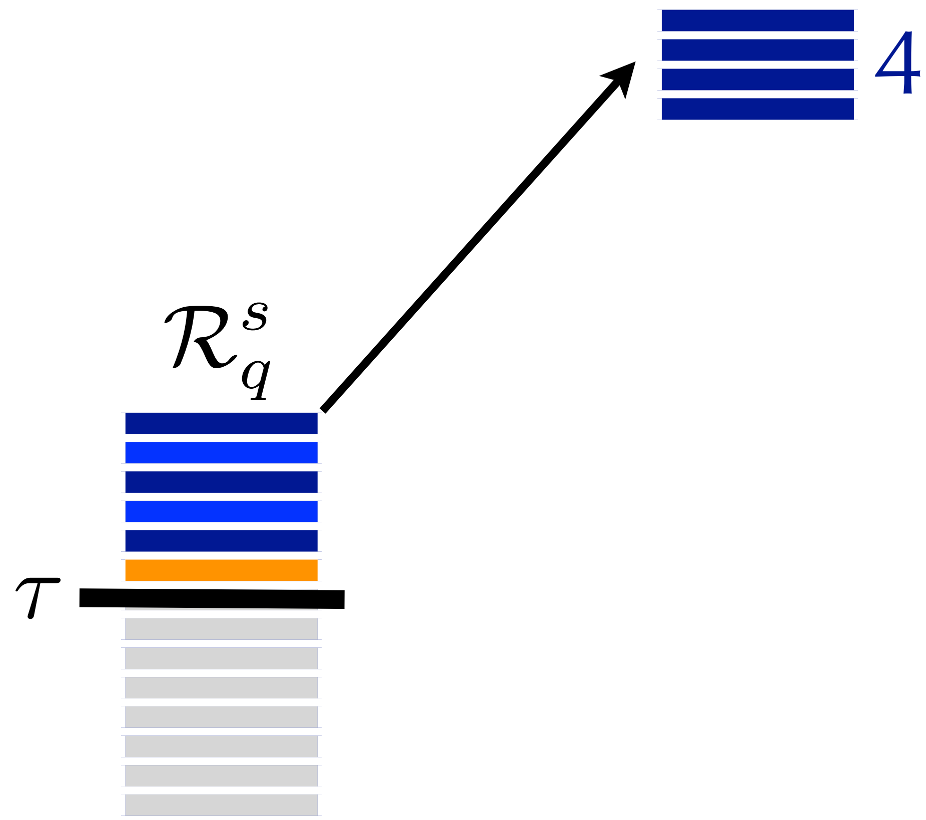
Small Document Models
ReDDE (Si and Callan, 2003)

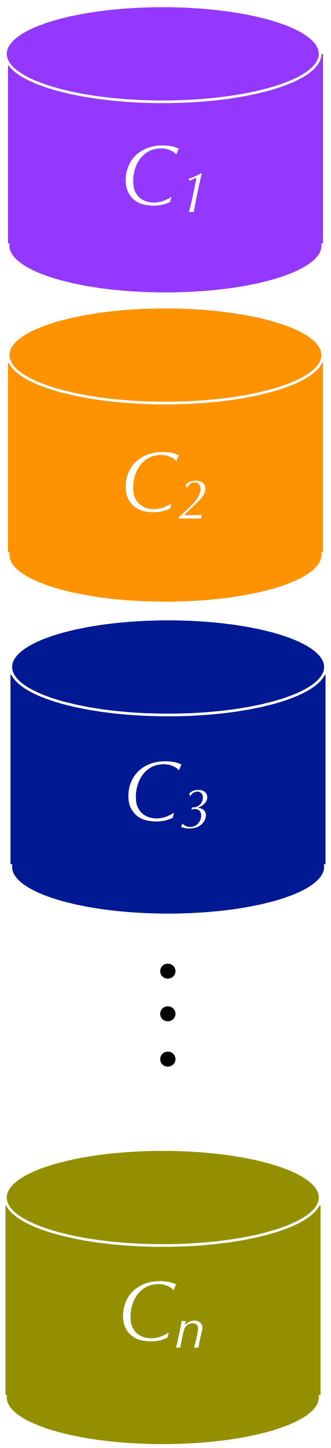


- Use a rank-based threshold to predict a set of relevant samples

Small Document Models

ReDDE (Si and Callan, 2003)



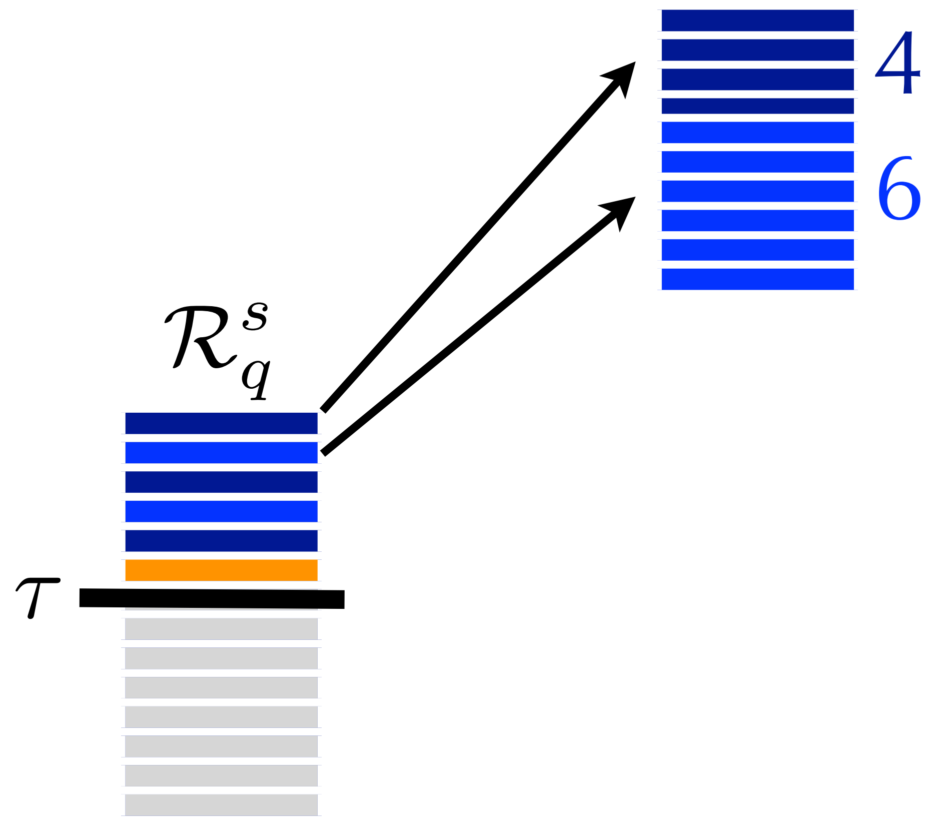


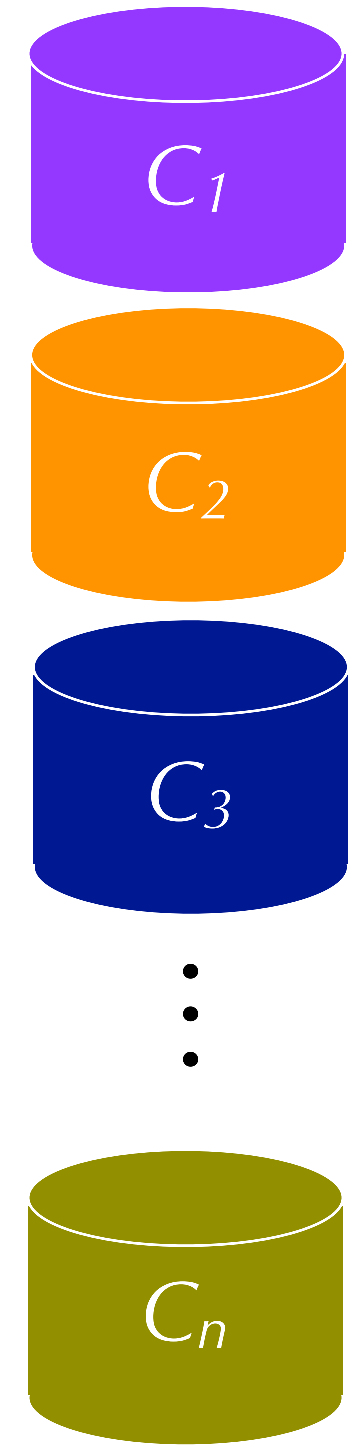
$$\text{scale factor}(C_i) = \frac{|C_i|}{|S_i|}$$

- Assume that each relevant sample represents some number of relevant documents in its original collection

Small Document Models

ReDDE (Si and Callan, 2003)



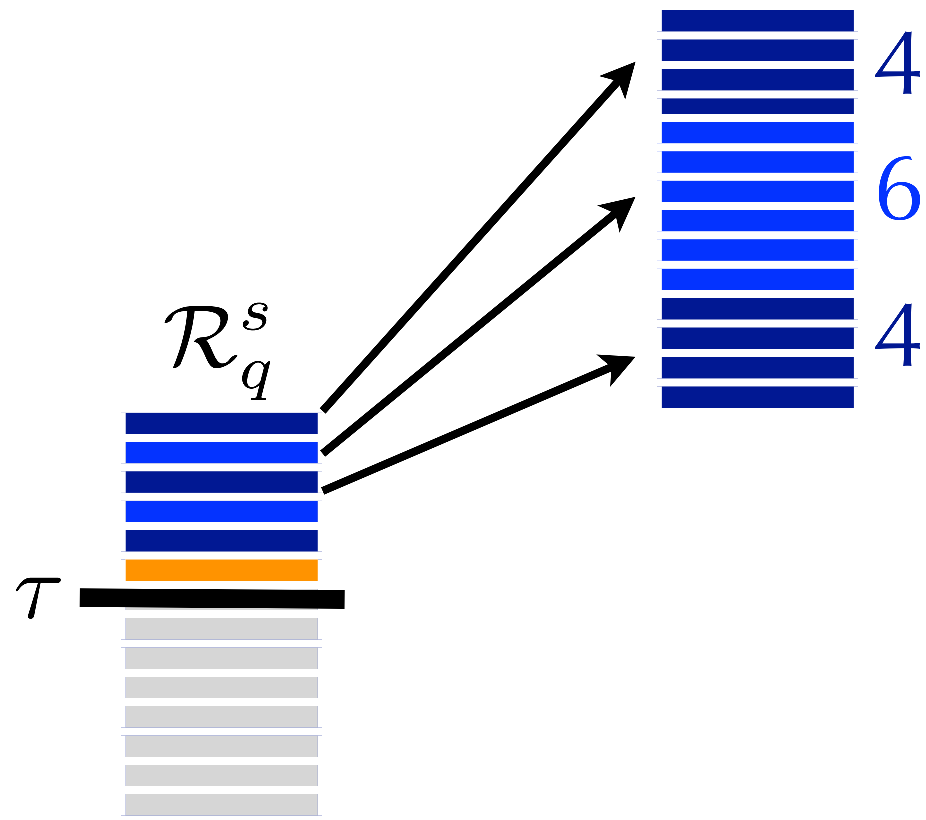


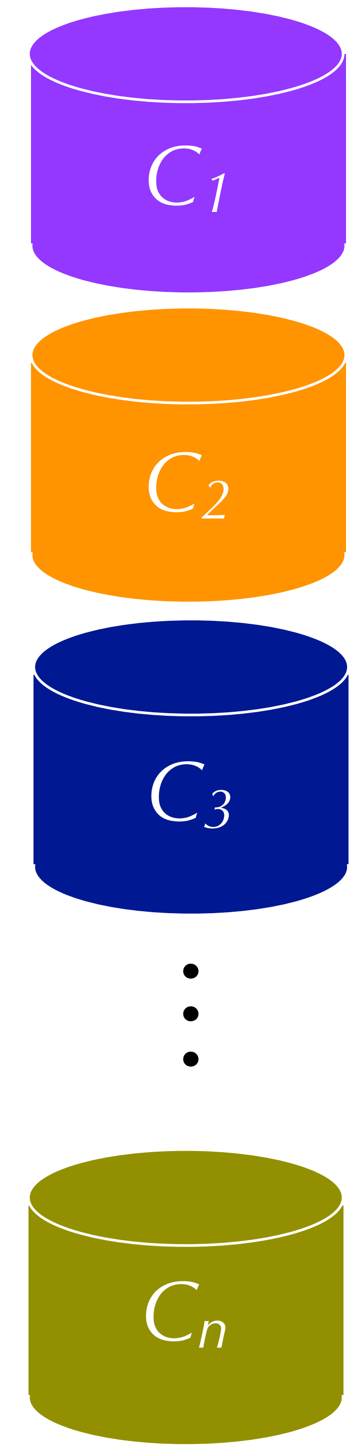
$$\text{scale factor}(C_i) = \frac{|C_i|}{|S_i|}$$

- "Scale-up" sample retrieval

Small Document Models

ReDDE (Si and Callan, 2003)



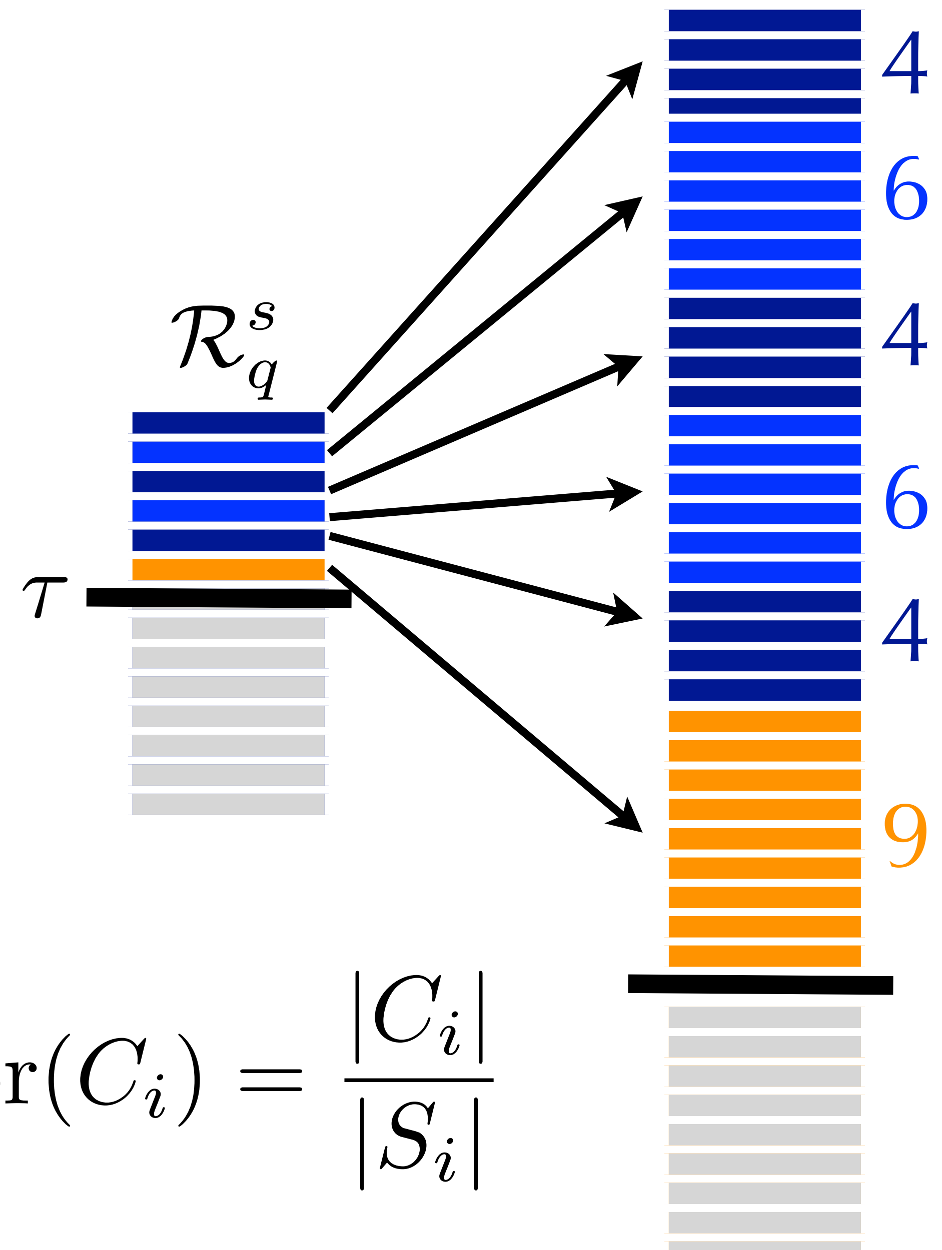


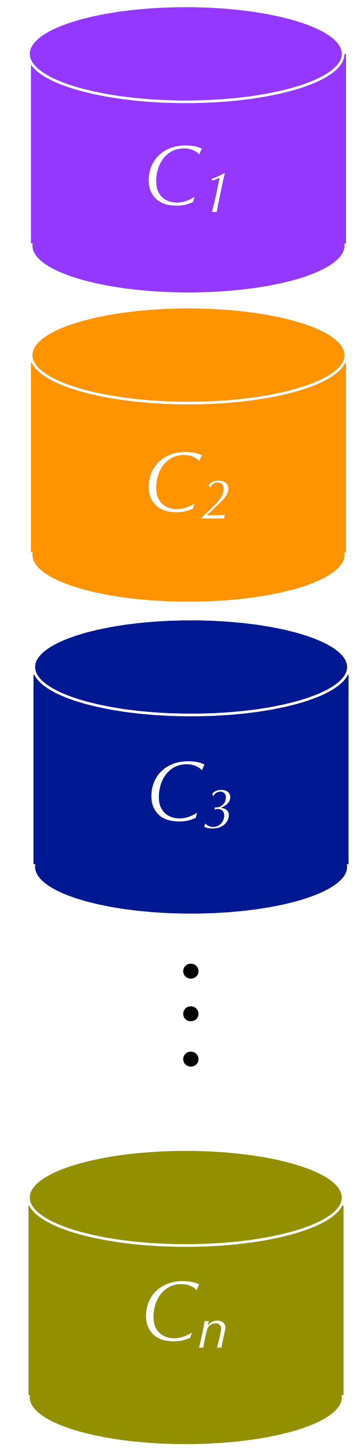
$$\text{scale factor}(C_i) = \frac{|C_i|}{|S_i|}$$

- "Scale-up" sample retrieval

Small Document Models

ReDDE (Si and Callan, 2003)





$$\text{scale factor}(C_i) = \frac{|C_i|}{|S_i|}$$

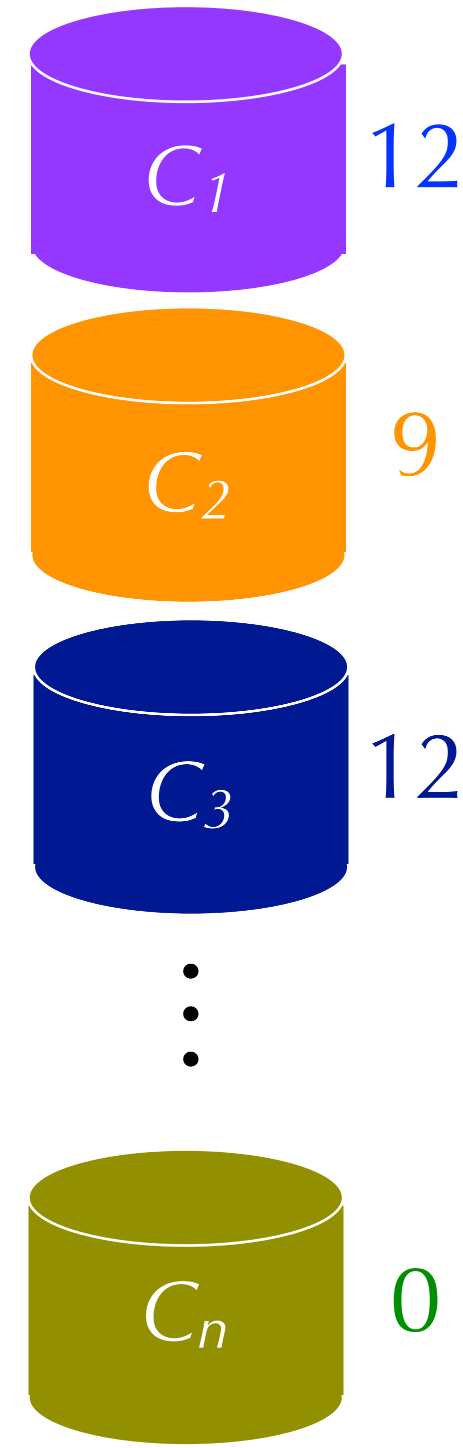
- "Scale-up" sample retrieval

Small Document Models

ReDDE (Si and Callan, 2003)

1. Score collections by their <u>estimated</u> number of relevant documents

2. Select the top *k*



$$\text{scale factor}(C_i) = \frac{|C_i|}{|S_i|}$$

Small Document Models
ReDDE Variants

- ReDDE can be viewed as a voting method: each (predicted) relevant sample is a vote for its collection
- Discussion: potential limitations?

Small Document Models
ReDDE Variants

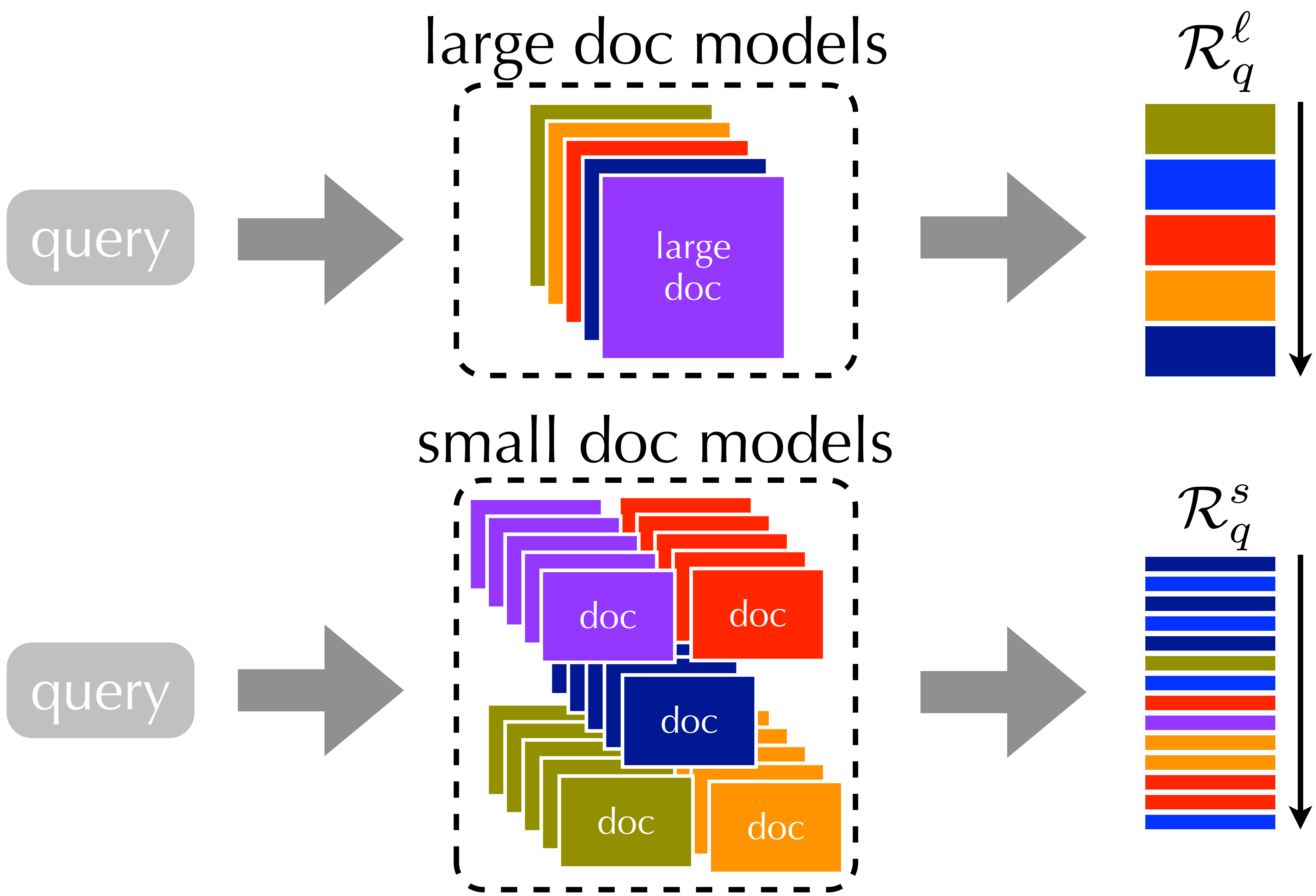
- ReDDE can be viewed as a voting method: each (predicted) relevant sample is a vote for its collection
- Discussion: potential limitations?
  - sensitivity to threshold parameter: samples that are more relevant (i.e., ranked higher) should get more votes (Shokouhi, 2007; Thomas, 2009)
  - a resource may not retrieve its relevant documents: samples from resources predicted to be more reliable should get more votes (Si and Callan, 2004)
- No ReDDE variant outperforms another across all experimental testbeds

Resource Selection
ReDDE vs. CORI

- ReDDE wins: it never does worse and often does better
- ReDDE outperforms CORI when the collection size distribution is skewed
  - CORI is biased towards small, topically-focused collections
  - favors collections that are proportionately relevant
  - misses large collections with many relevant documents

Resource Selection

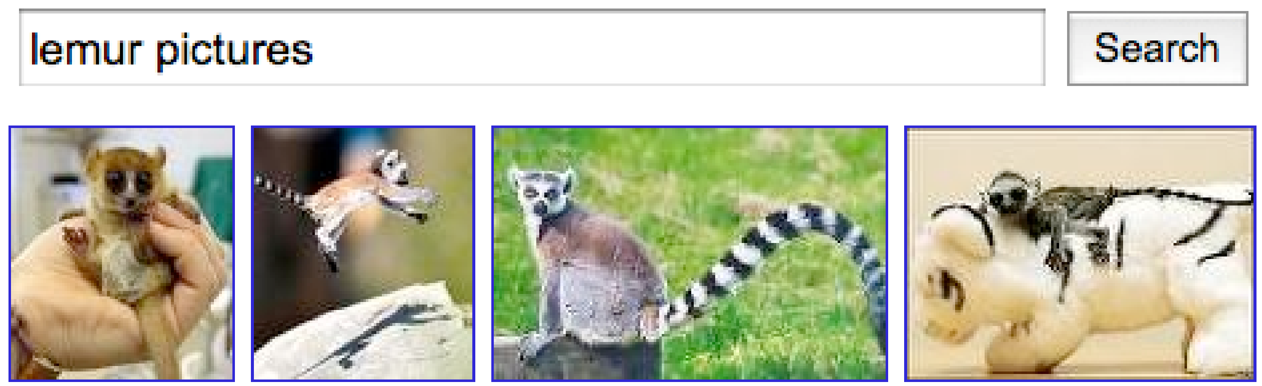
content-based methods

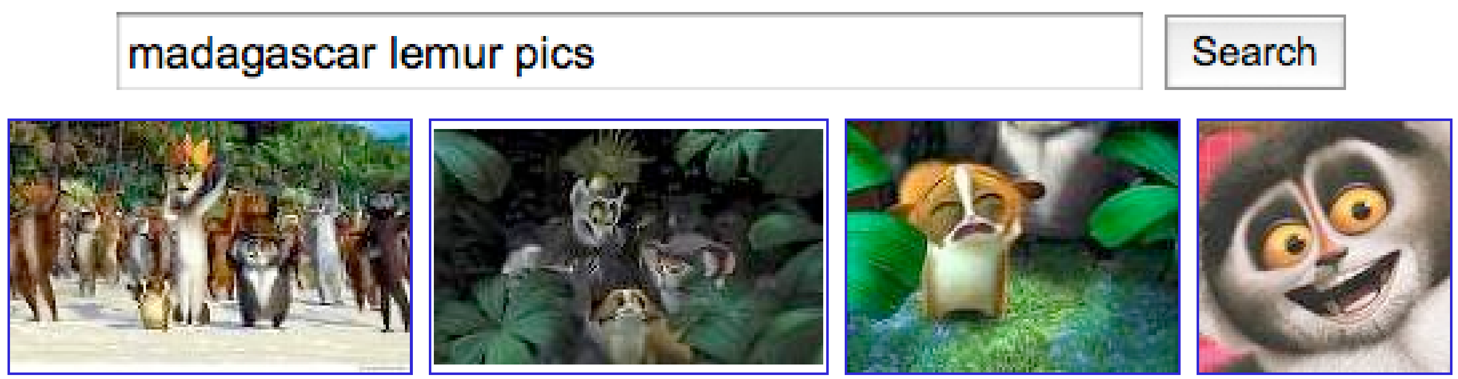


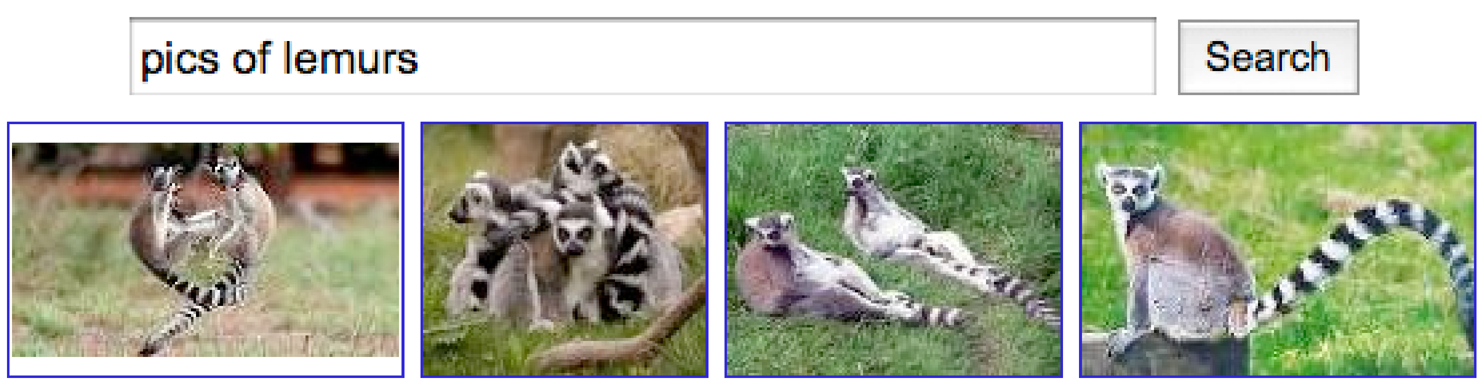
- <u>Resource relevance</u> as a function of <u>content relevance</u>

Resource Selection
query-similarity methods

- Key assumption: similar queries retrieve similar results







Resource Selection

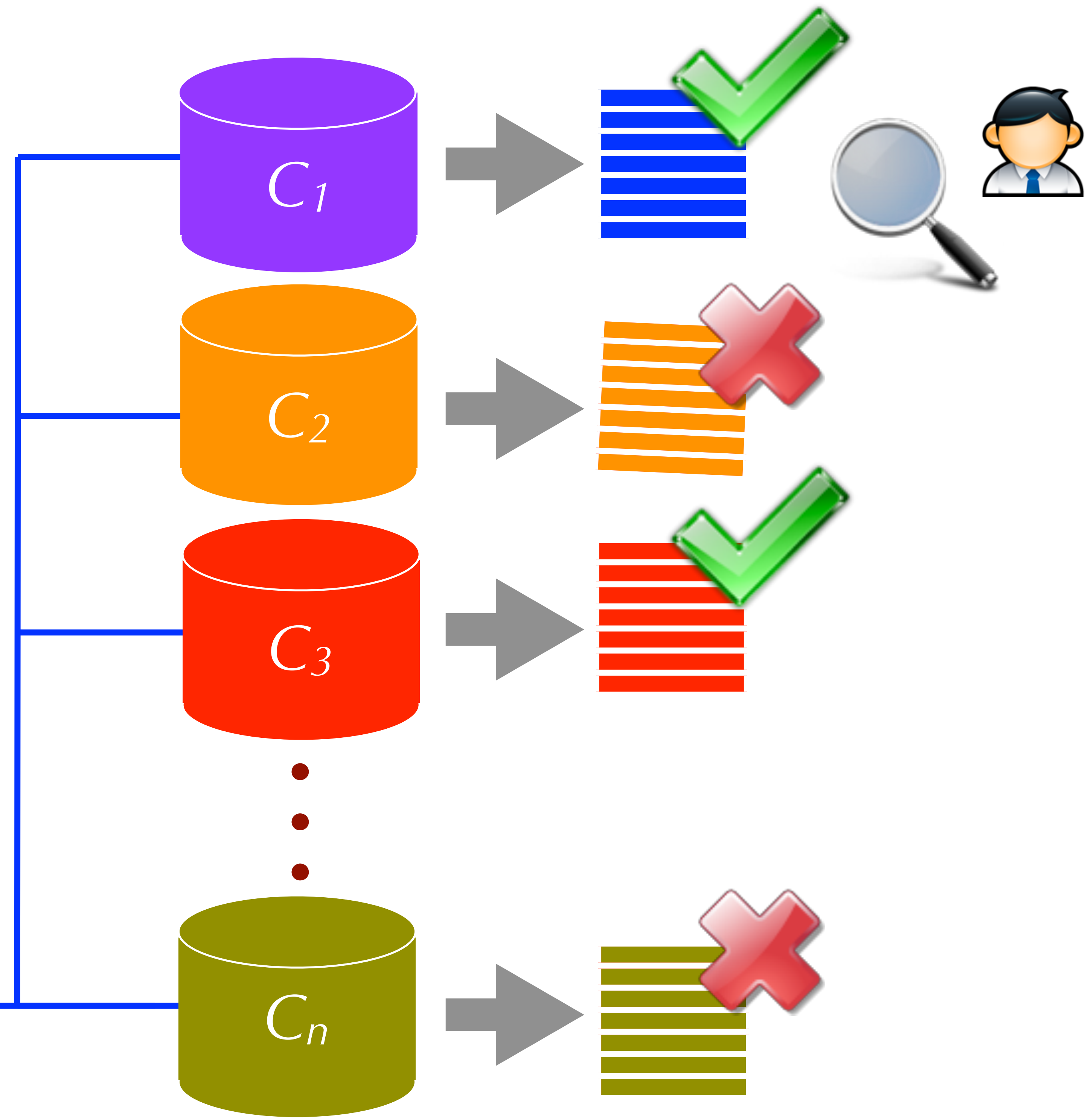
query-similarity methods

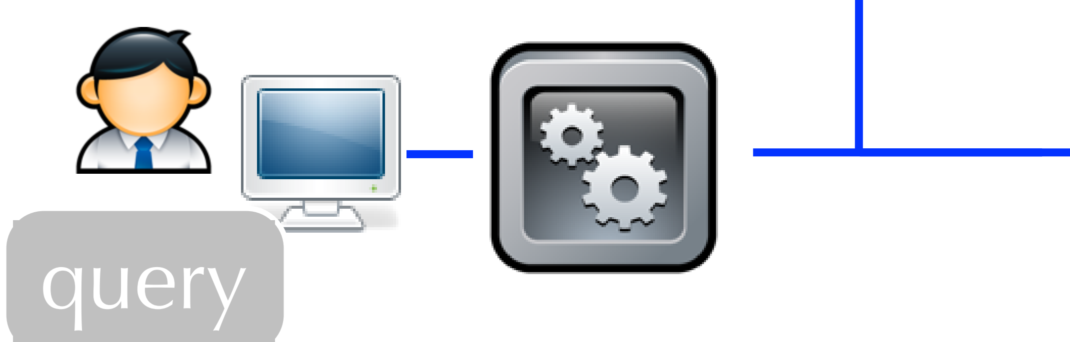
- Select resources based on their <u>expected retrieval effectiveness</u> for the given query
- Requires two components:
  1. retrieval effectiveness: a way to determine that a previously seen query produced an effective retrieval from the resource
  2. query-similarity: a way to predict that a new (unseen) query will retrieve similar results from the resource

Query-Similarity Methods
(Voorhees *et al.*, 1995)

- Training phase: did the resource <u>retrieve</u> relevant documents?

- e.g., use human relevance judgements



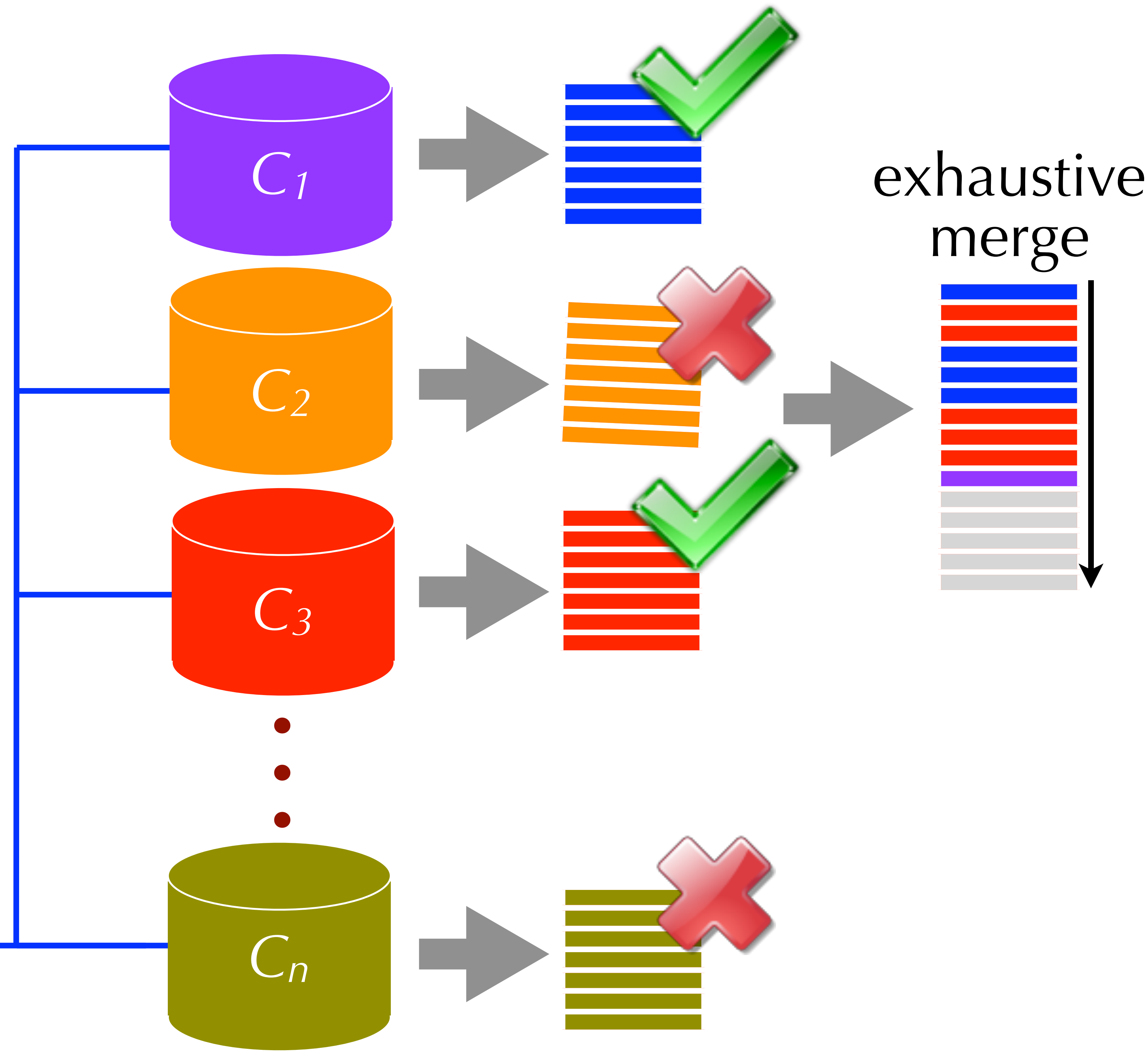


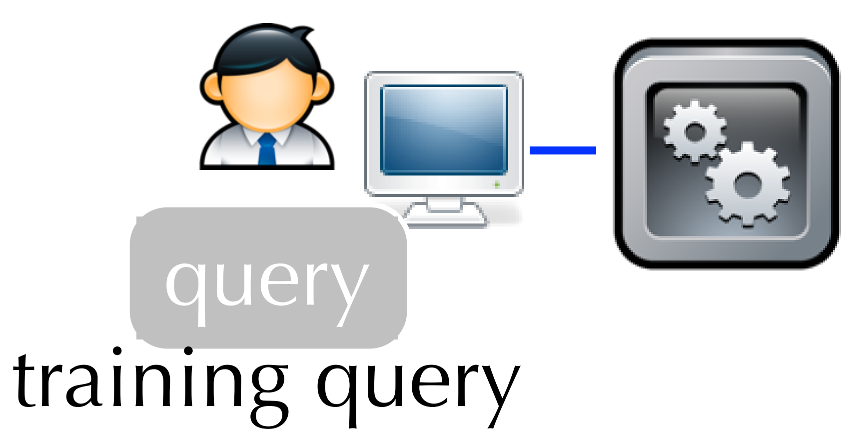
training query

Query-Similarity Methods
(Arguello *et al.*, 2008)

- Training phase: did the resource <u>retrieve</u> relevant documents?

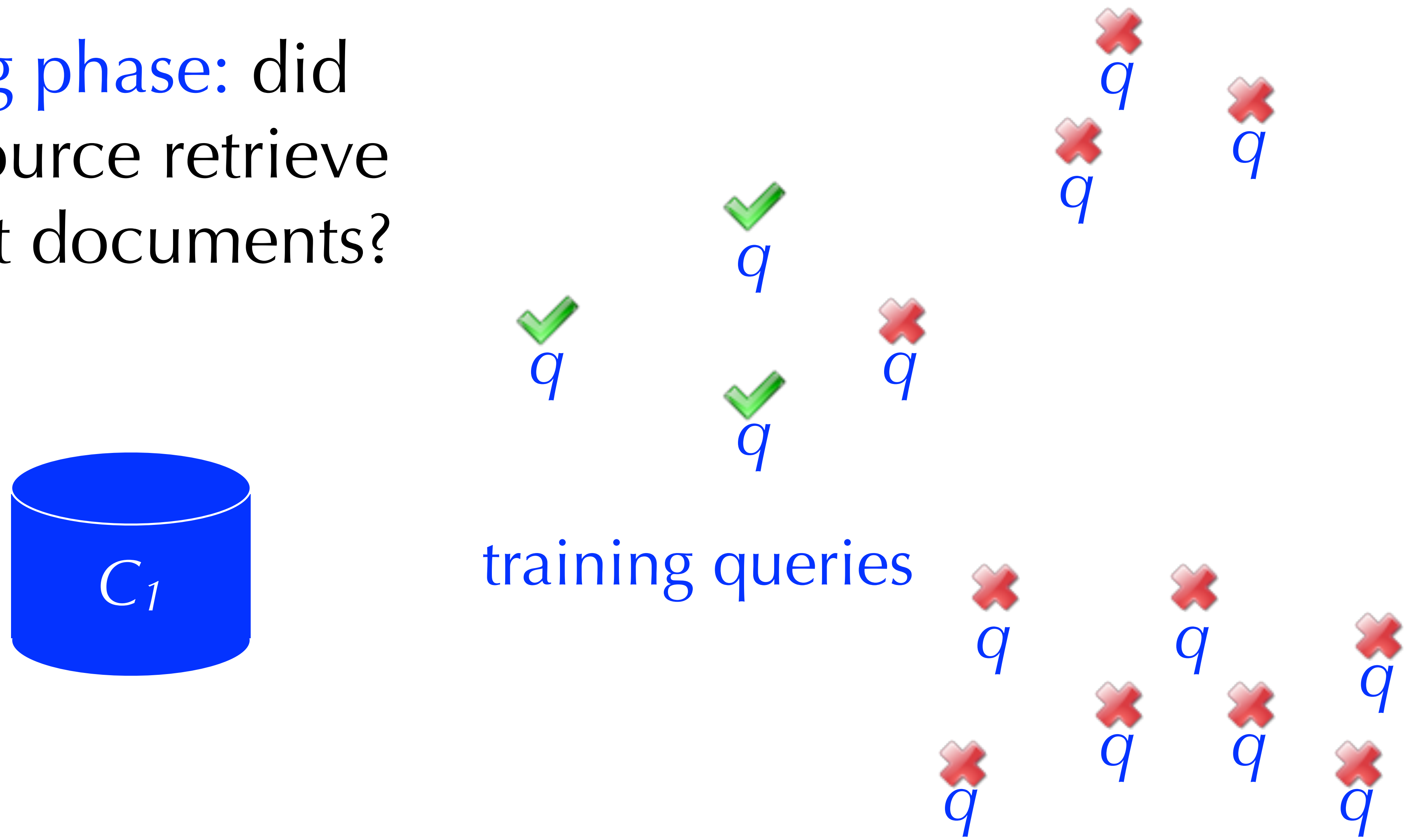
- e.g., use retrievals that merge content from <u>every</u> resource





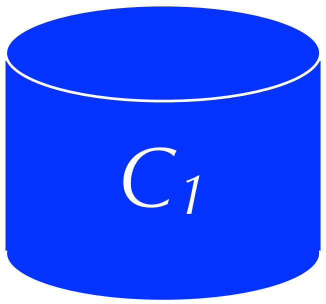
Query-Similarity Methods

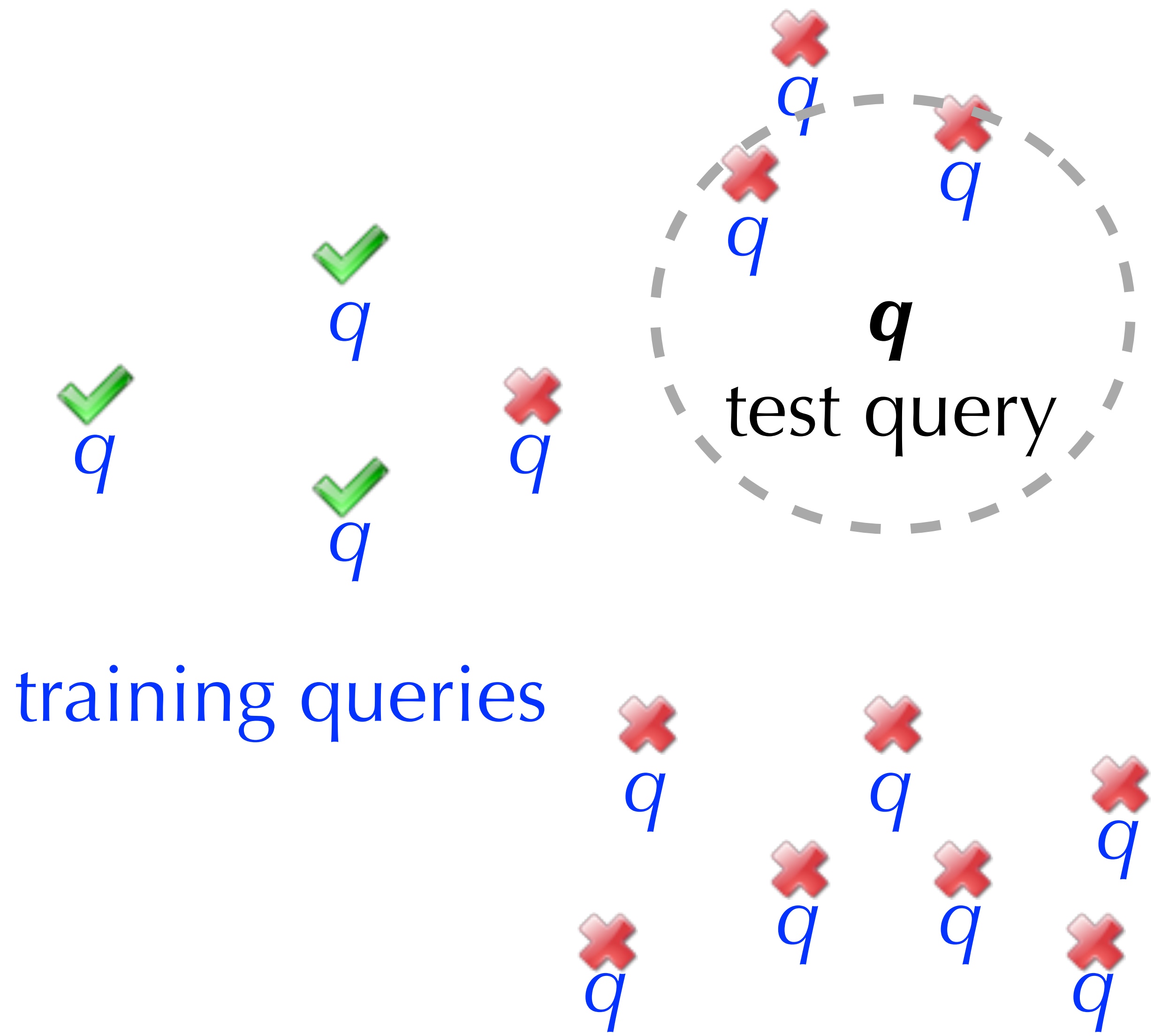
- Training phase: did the resource retrieve relevant documents?



Query-Similarity Methods

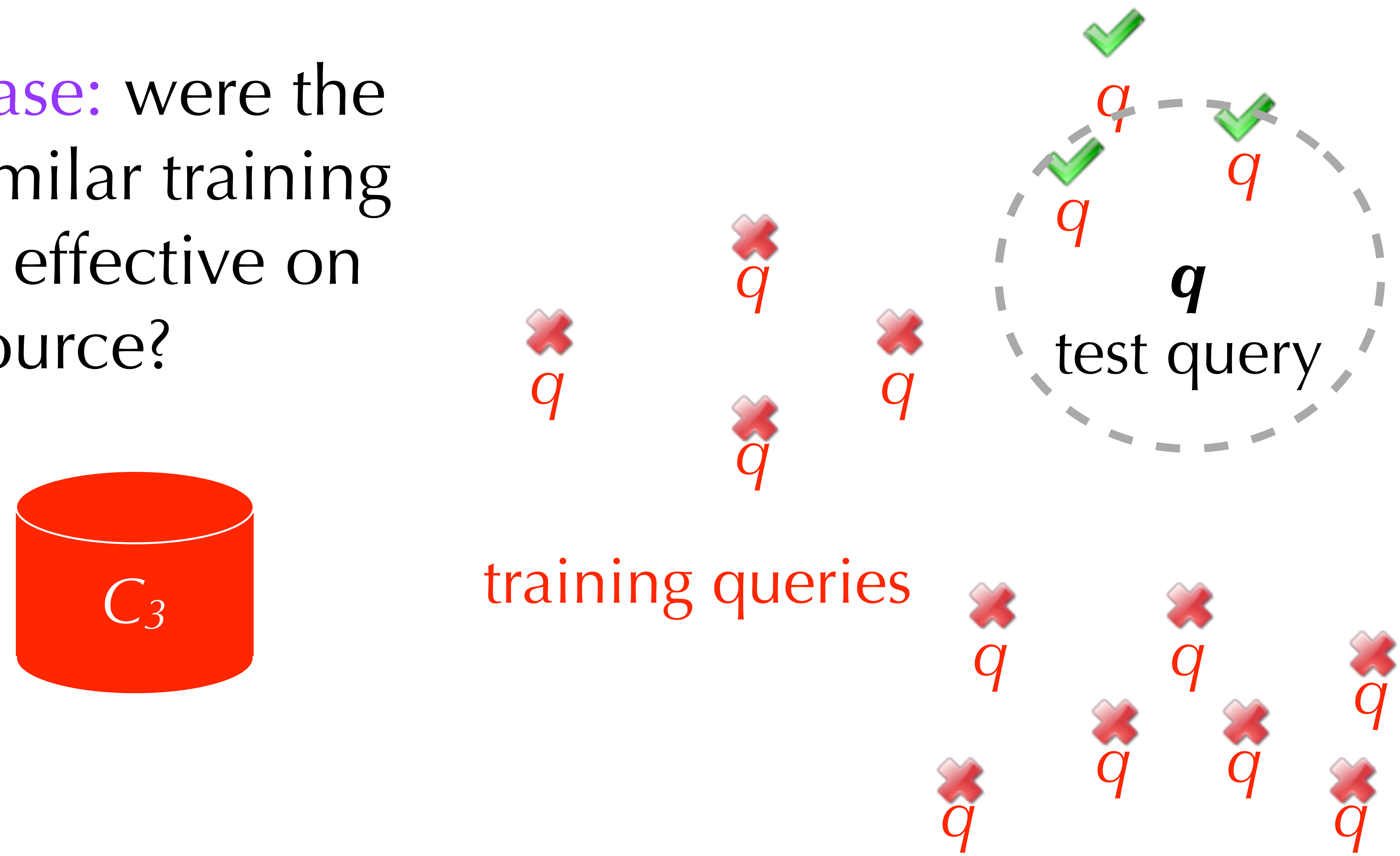
- Test phase: were the most similar training queries effective on the resource?





Query-Similarity Methods

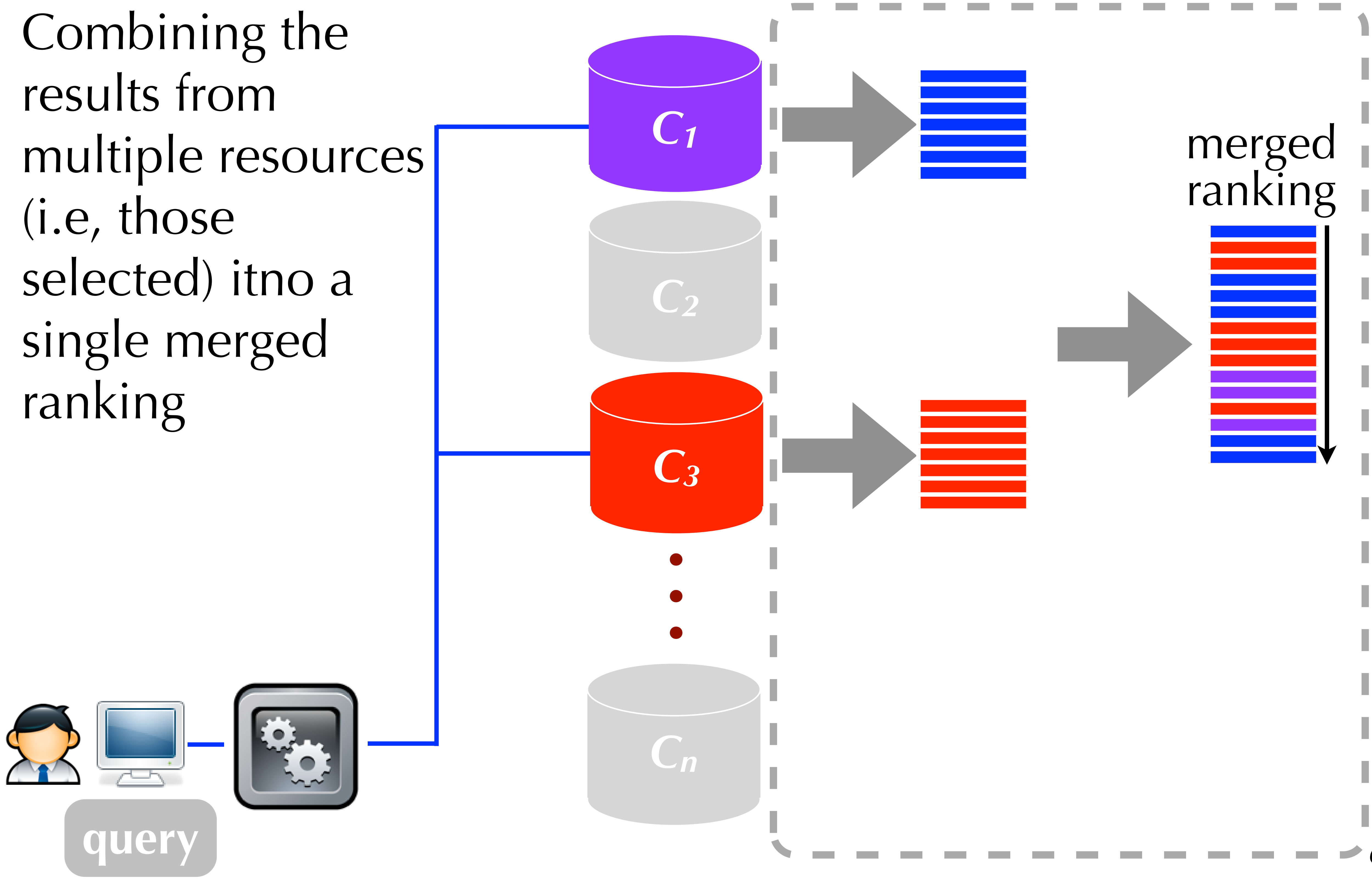
- Test phase: were the most similar training queries effective on the resource?



Results Merging

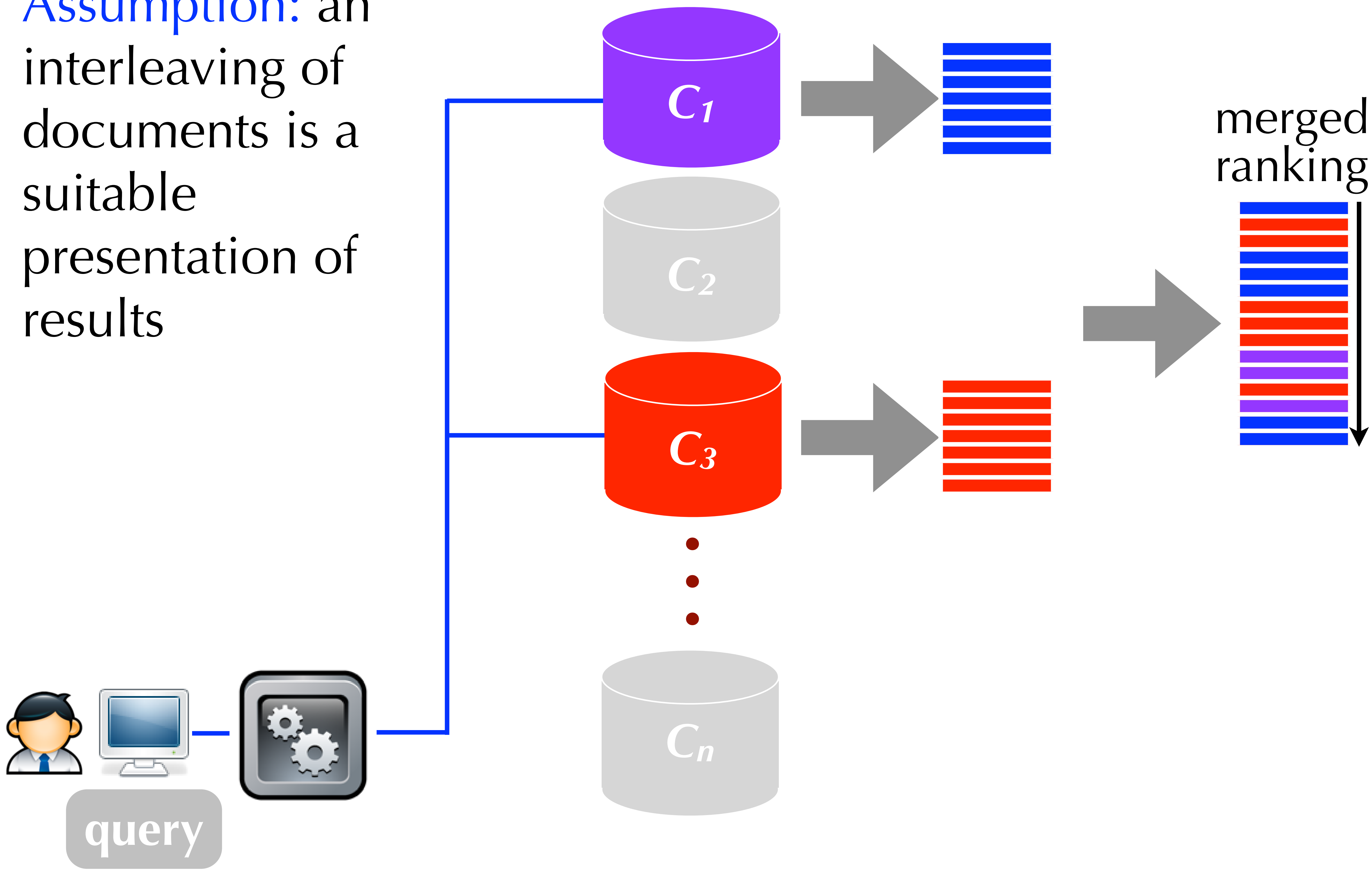
Results Merging

- Combining the results from multiple resources (i.e, those selected) itno a single merged ranking



Results Merging

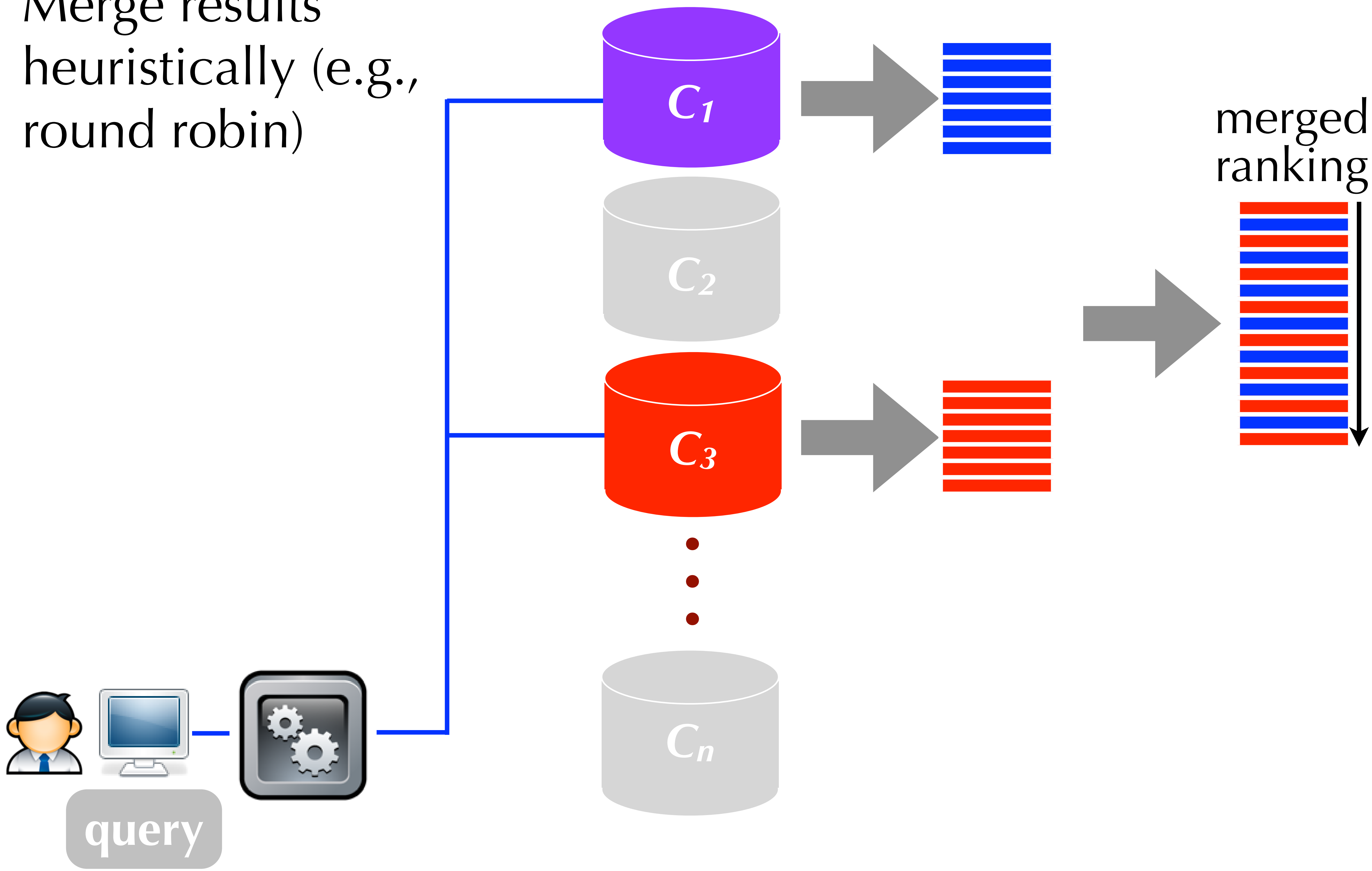
- Assumption: an interleaving of documents is a suitable presentation of results



Results Merging

Naive Interleaving

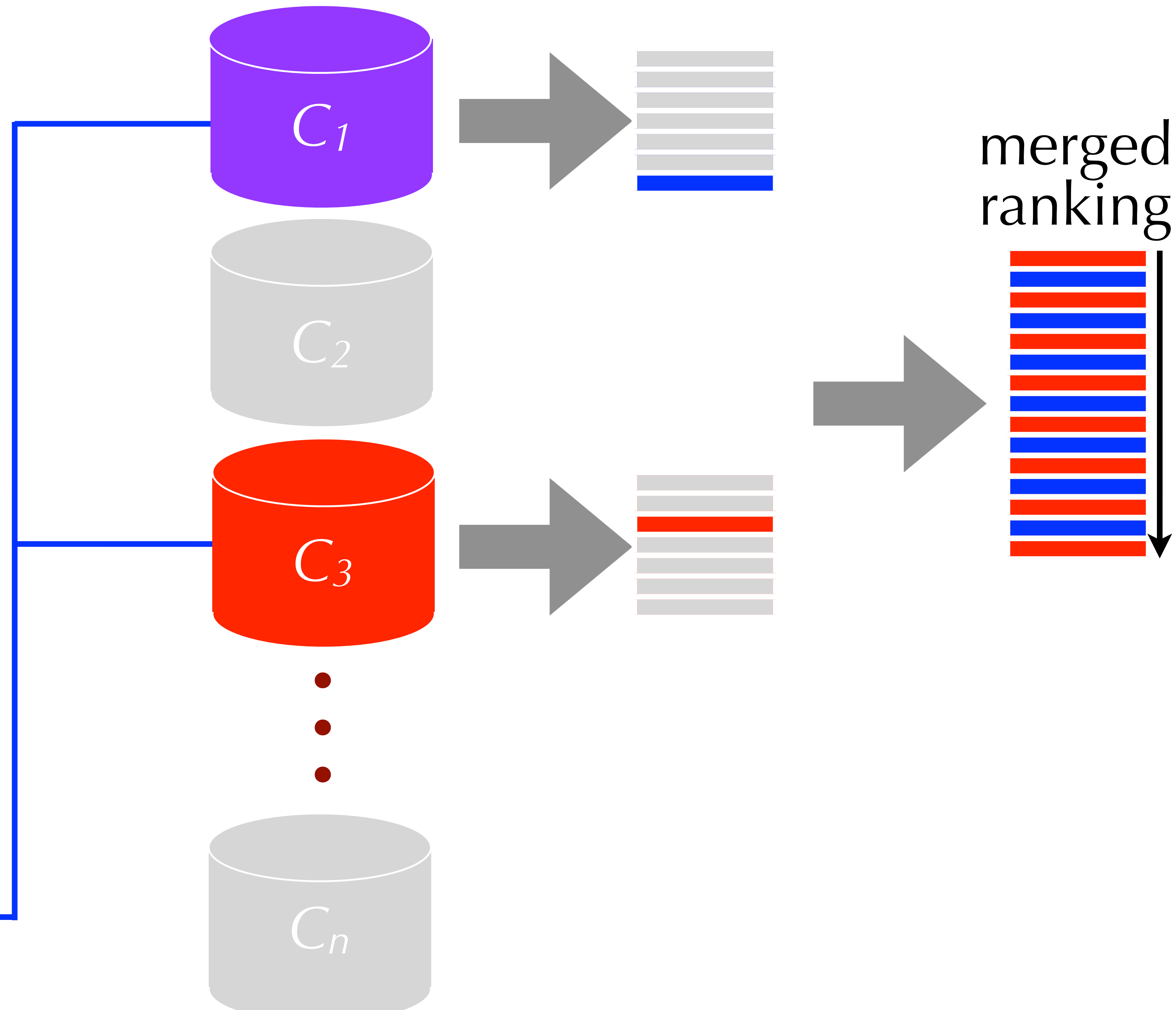
- Merge results heuristically (e.g., round robin)

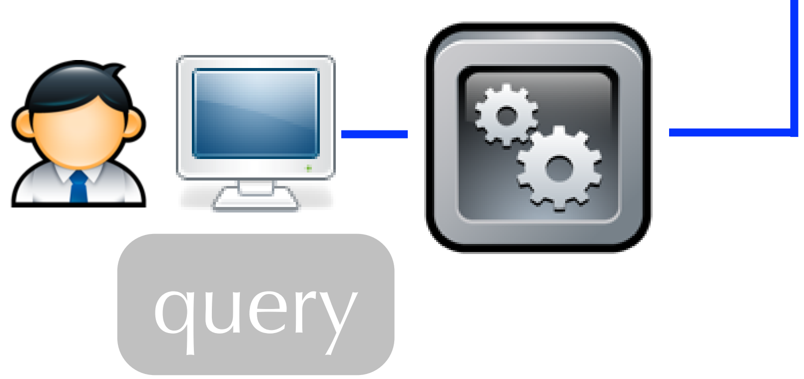


Results Merging

Naive Interleaving

- Problem:
  rank 7 from C_1
  may be more
  relevant than
  rank 3 from C_3.
  why?
- what other option do we have?

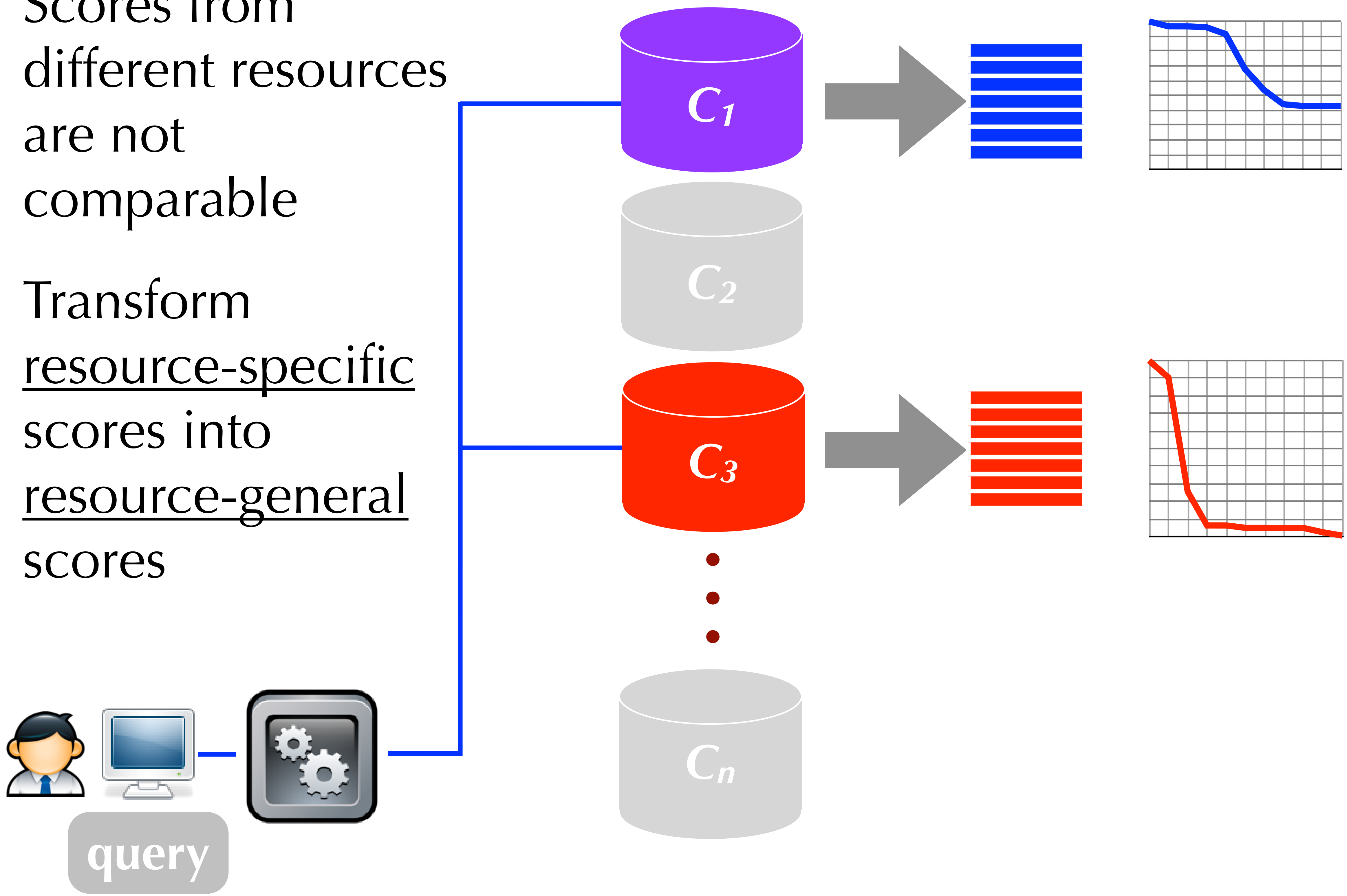




Results Merging

Score Normalization

- Scores from different resources are not comparable
- Transform <u>resource-specific</u> scores into <u>resource-general</u> scores



Results Merging

CORI-Merge (Callan *et al.*, 1995)

- Combine resource ranking and document ranking scores

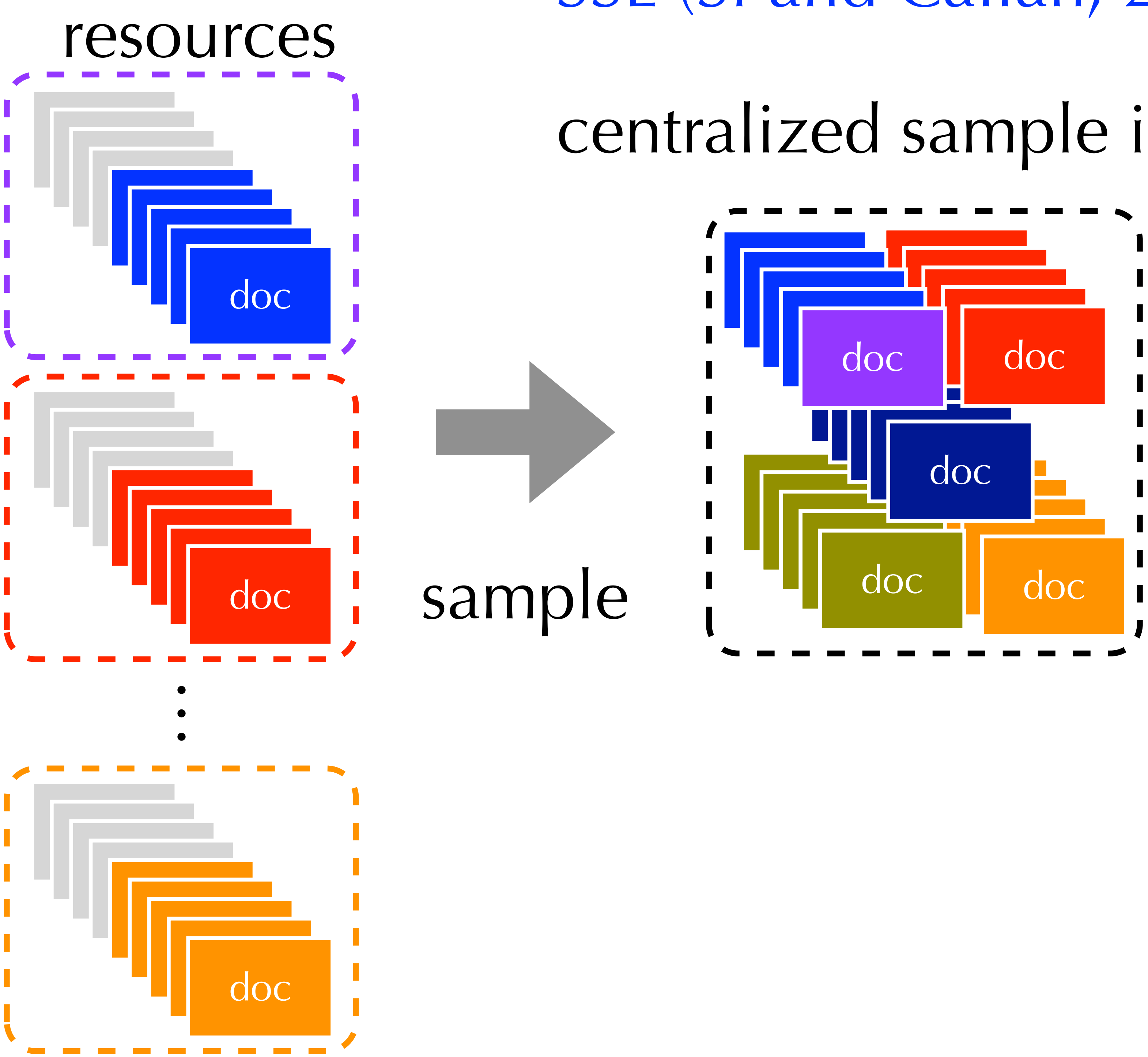
$$S_C(D) = \frac{S_i'(D) + 0.4 \times S_i'(D) \times S'(C_i)}{1.4}$$

$$S_i'(D) = \frac{S_i(D) - S_i(D_{\min})}{S_i(D_{\max}) - S_i(D_{\min})}$$

$$S'(C_i) = \frac{S(C_i) - S(C_{\min})}{S(C_{\max}) - S(C_{\min})}$$

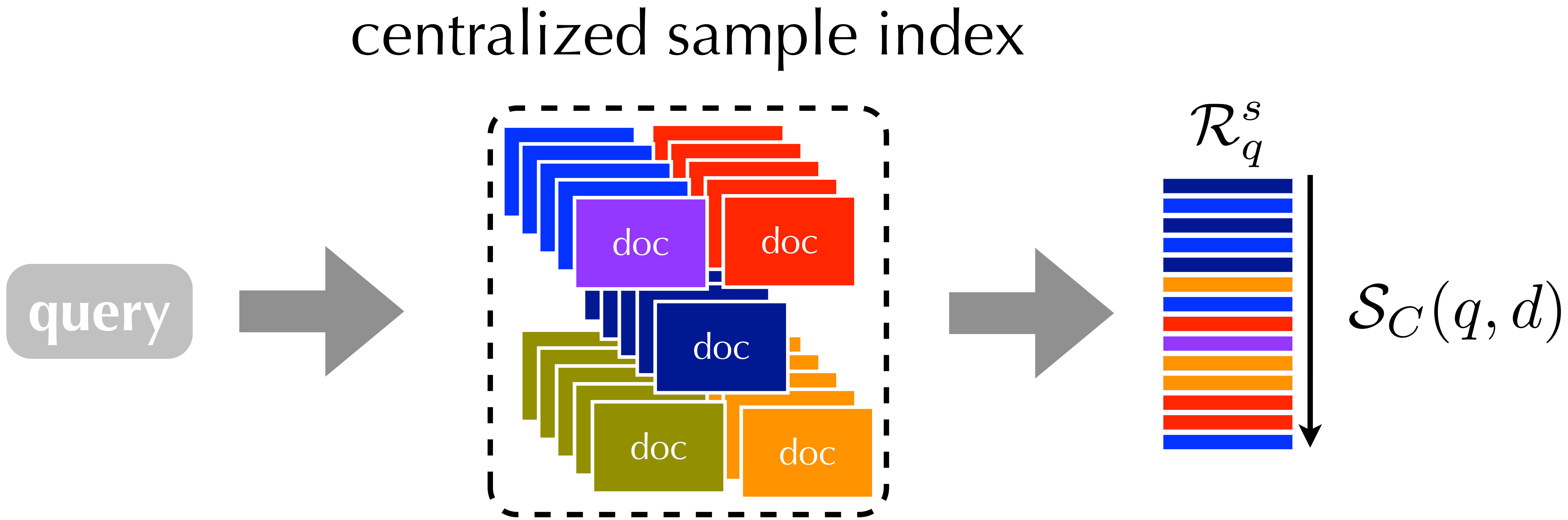
Results Merging

SSL (Si and Callan, 2003)



Results Merging

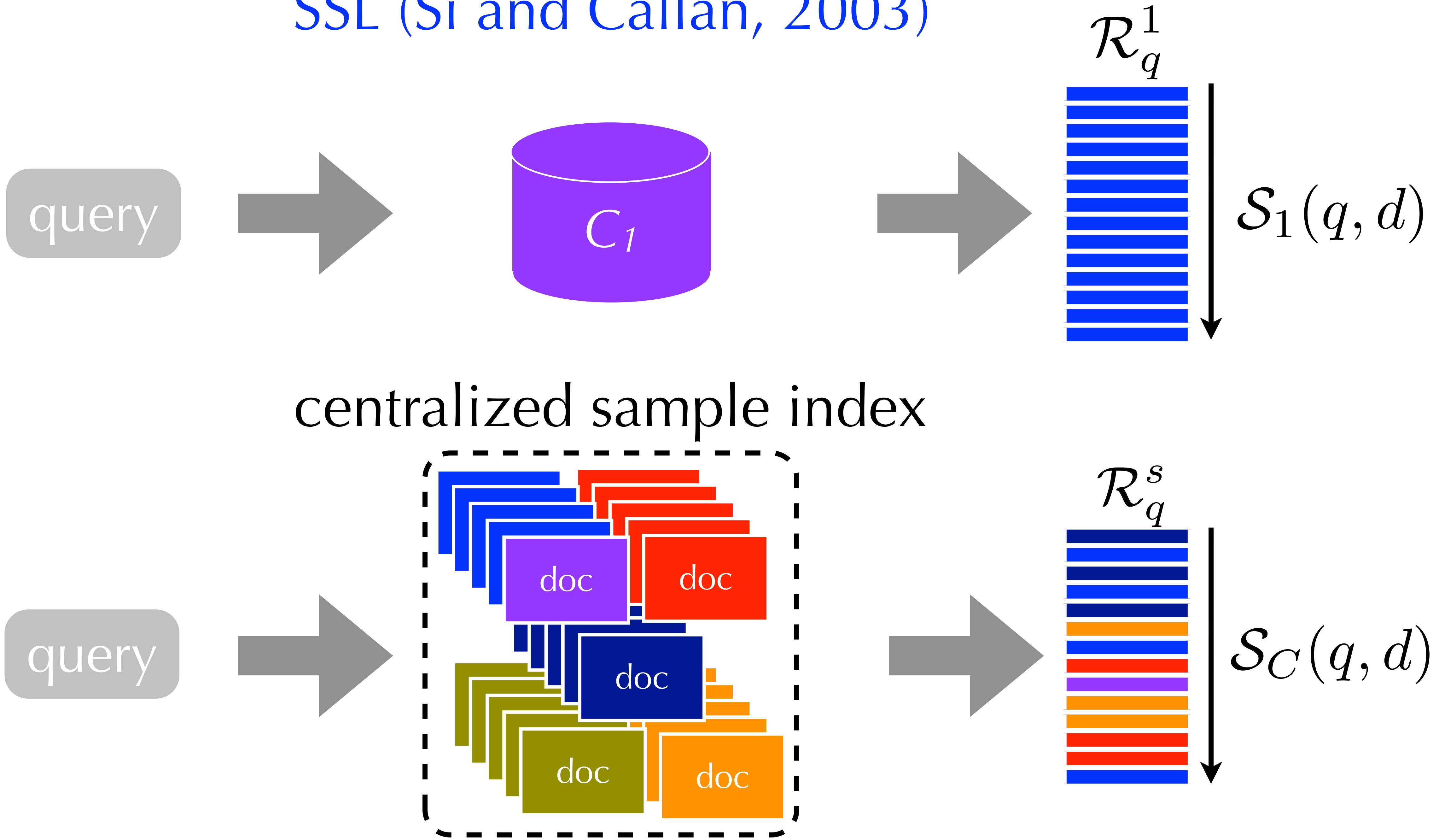
SSL (Si and Callan, 2003)



- Assumption: centralized sample index scores are directly comparable
  - same ranking/scoring algorithm
  - same IDF values
  - same document-length normalization

Results Merging

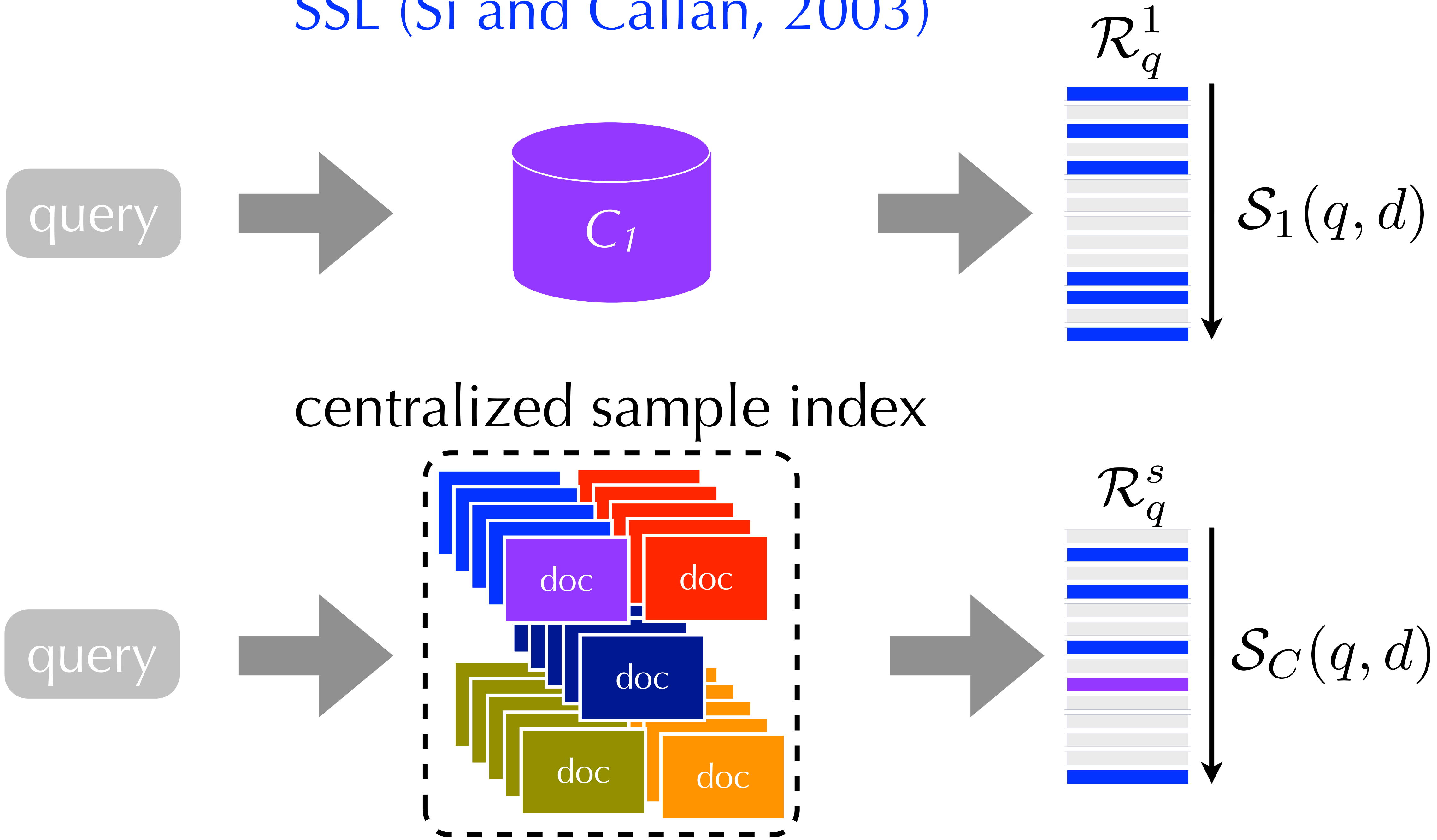
SSL (Si and Callan, 2003)



- Objective: given a query, transform C_1 scores to values that are comparable across collections

Results Merging

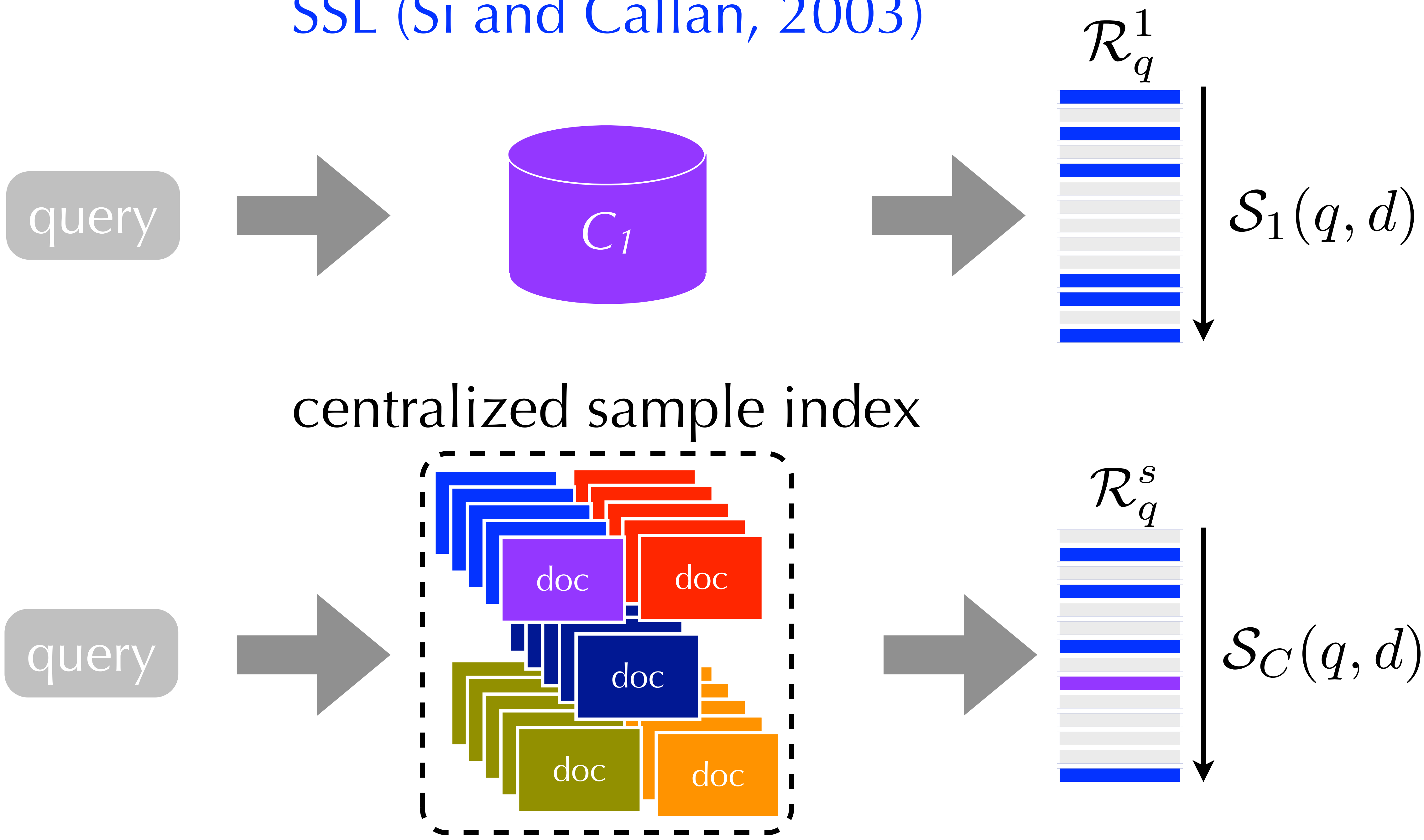
SSL (Si and Callan, 2003)



- Step 1: identify the overlap documents

Results Merging

SSL (Si and Callan, 2003)



- Step 2: use these pairs of document-scores to learn a linear transformation from C_1 scores to *CSI* scores

Results Merging

SSL (Si and Callan, 2003)

- Step 2: use these pairs of document scores to learn a linear transformation from *C_1* to *CSI* scores
- Standard linear regression (query and collection specific)

$$\mathcal{S}_C(q,d) = a \times \mathcal{S}_i(q,d) + b$$

$$\arg\min_{a,b} \sum_{d} \Big(\big(f(a, b, \mathcal{S}_i(q,d)\big) - \mathcal{S}_C(q,d) \Big)^2$$

overlap documents
(query and collection specific)

Federated Search Summary

- QBS produces effective collection representations
  - ~500 docs are enough, doesn't require cooperation
- Small document models > large document models
  - But, both assume an effective retrieval
- Query-based methods avoid this by modeling the expected retrieval using previous retrievals
  - But, require training data. or, Do they?
- Centralized sample index scores are "resource-general"
  - learn a regression model to re-score and merge

Vertical Aggregation

pittsburgh | Search

maps

[Pittsburgh, PA](#) maps.google.com

web

[City of **Pittsburgh**, Pennsylvania - Pghgov.com](#)
Official city site including information on economic development, resident information, links, tourism and contact information.
www.city.**pittsburgh**.pa.us/ - Cached - Similar

images

[Images for **pittsburgh**](#) - Report images

web

[**Pittsburgh** - Wikipedia, the free encyclopedia](#)
Pittsburgh is the second-largest city in the U.S. Commonwealth of Pennsylvania and the county seat of Allegheny County. Regionally, it anchors the largest ...
History of Pittsburgh - Neighborhoods - List of people from the Pittsburgh ... - 1936
en.wikipedia.org/wiki/**Pittsburgh** - Cached - Similar

books

[Books for **pittsburgh**](#)
[**Pittsburgh**: a sketch of its early social life](#) - Charles William Dahlinger - 1916 - 216 pages
[**Pittsburgh**:: 1758-2008](#) - Pittsburgh Post-Gazette, Carnegie Library of Pittsburgh - 2008 - 128 pages
Pittsburgh: 17582008 surveys the citys evolution from strategic fort in the wilderness ...
books.google.com

References

J. Arguello., F. Diaz, and J. Callan. (2009). Sources of evidence for vertical selection. In *SIGIR*.

J. Callan, Z. Lu, and W.B. Croft. (1995). Searching distributed collections with inference networks. In *SIGIR*.

J. Callan and M. Connell. (2001). Query-based sampling of text databases. In *TOIS*.

L. Gravano, C. Chang, H. Garcia-Molina, and A. Paepcke. (1997). STARTS. In *SIGMOD*.

L. Si and J. Callan. (2003). Relevant document distribution estimation method for resource selection. In *SIGIR*.

L. Si, R. Jin, J. Callan, and P. Ogilvie. (2002). Language modeling framework for resource selection and results merging. In *CIKM*.

M. Shokouhi. (2007). Central rank-based collection selection in uncooperative distributed information retrieval. In *ECIR*.

M. Shokouhi, M. Baillie, and L. Azzopardi. (2007). Updating collection representations for federated search. In *SIGIR*.

P. Thomas and M. Shokoui. (2009). SUSHI: Scoring scaled samples for server selection. In *SIGIR*.

J.Xu and W. B. Croft. (1999). Cluster-based language models for distributed retrieval. In *SIGIR*