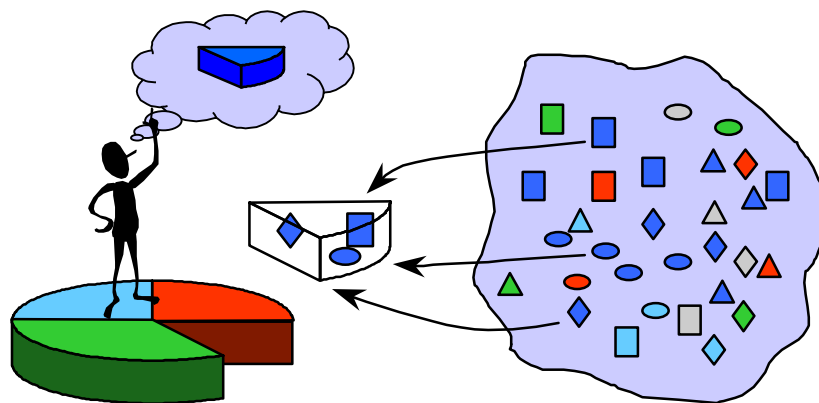


Personalized Information Environments



James C. French, Andrew S. Grimshaw
Department of Computer Science
University of Virginia

Charles L. Viles
School of Information and Library Science
University of North Carolina

Outline

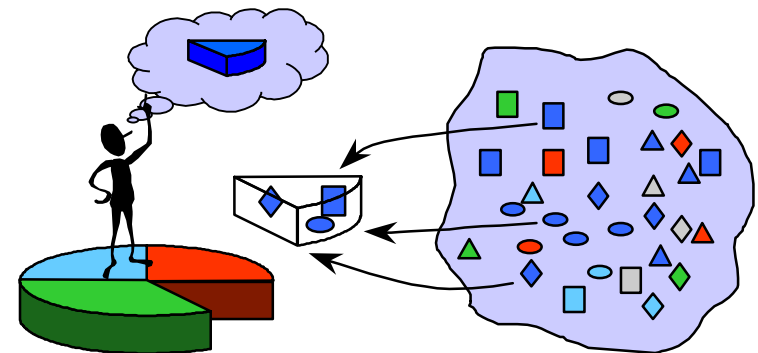


- **Personalized Information Env. (PIE)**
- **PIE Architecture**
- **Implementation Status**
- **Research Activities**
 - **Distributed search**
 - **Information Exploration**
 - **Metadata Management**

What is the problem?



- **vast sources of information**
 - organization is *provider*-centric
- **want *user*-centric strategy to:**
 - select databases for search
 - conduct searches
 - merge results
 - awareness services (SDI)



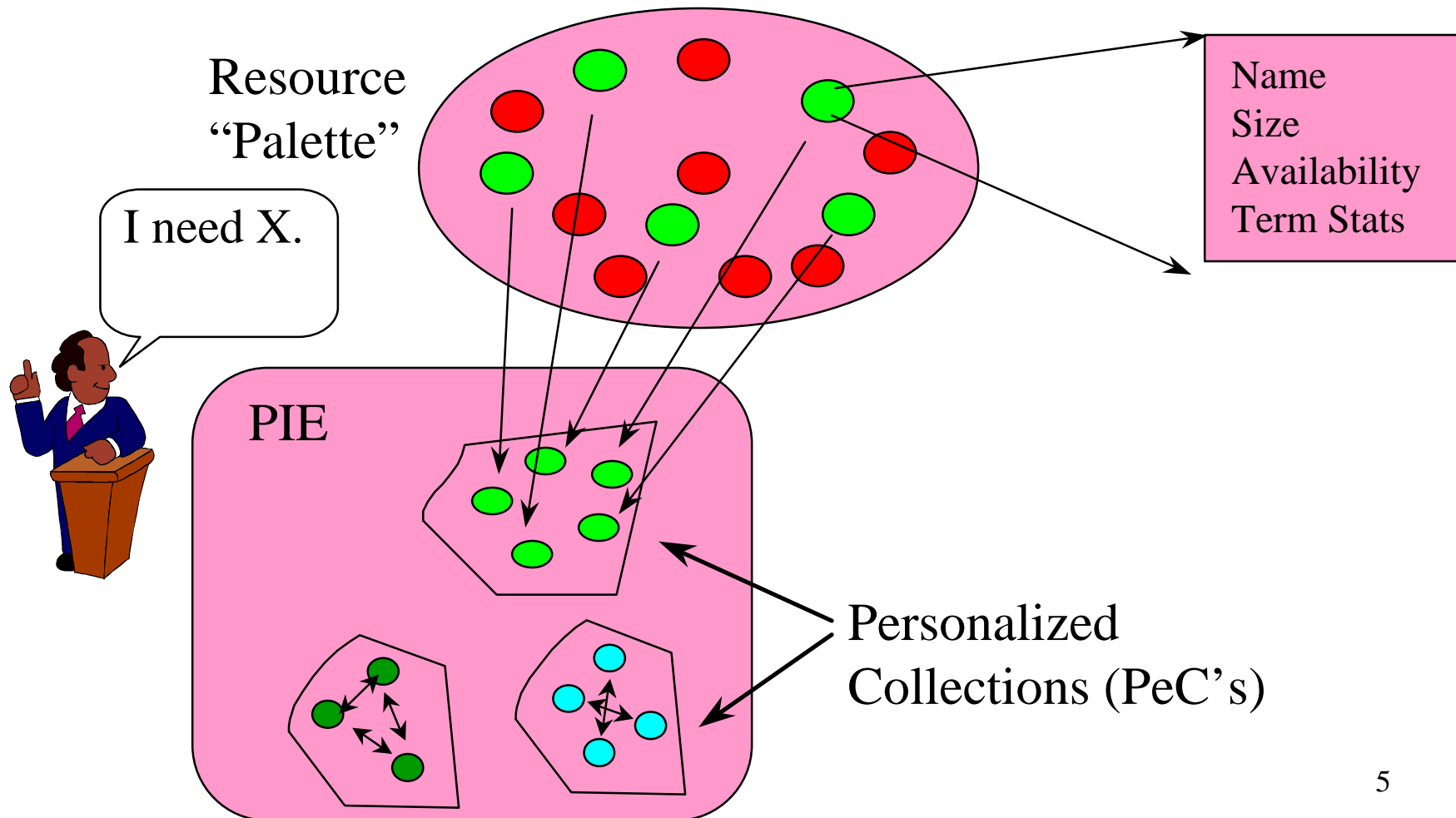
Our solution



Create personalized information environments (PIE) that:

- **are persistent**
- **are user-customizable**
- **are private or sharable**
- **provide effective search efficiently**
- **are secure**

Personalized Information Environments



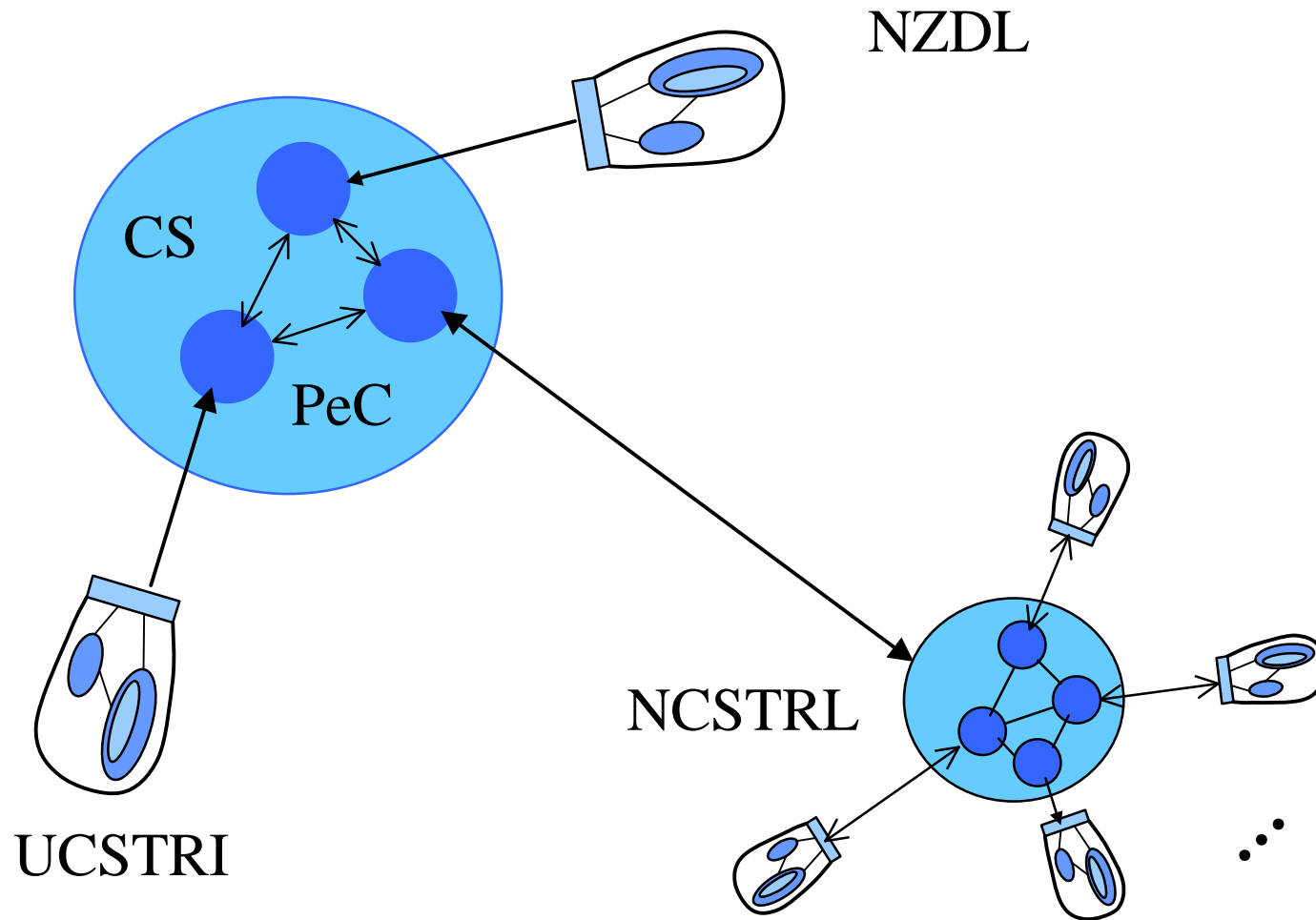
PIE Architecture



Personalized Information Environment components:

- **Personalized Collections (PeC)**
 - user specified information resources
- **Virtual Repositories (VIRP)**
 - uniform encapsulation of information sources

CS Technical Reports PeC



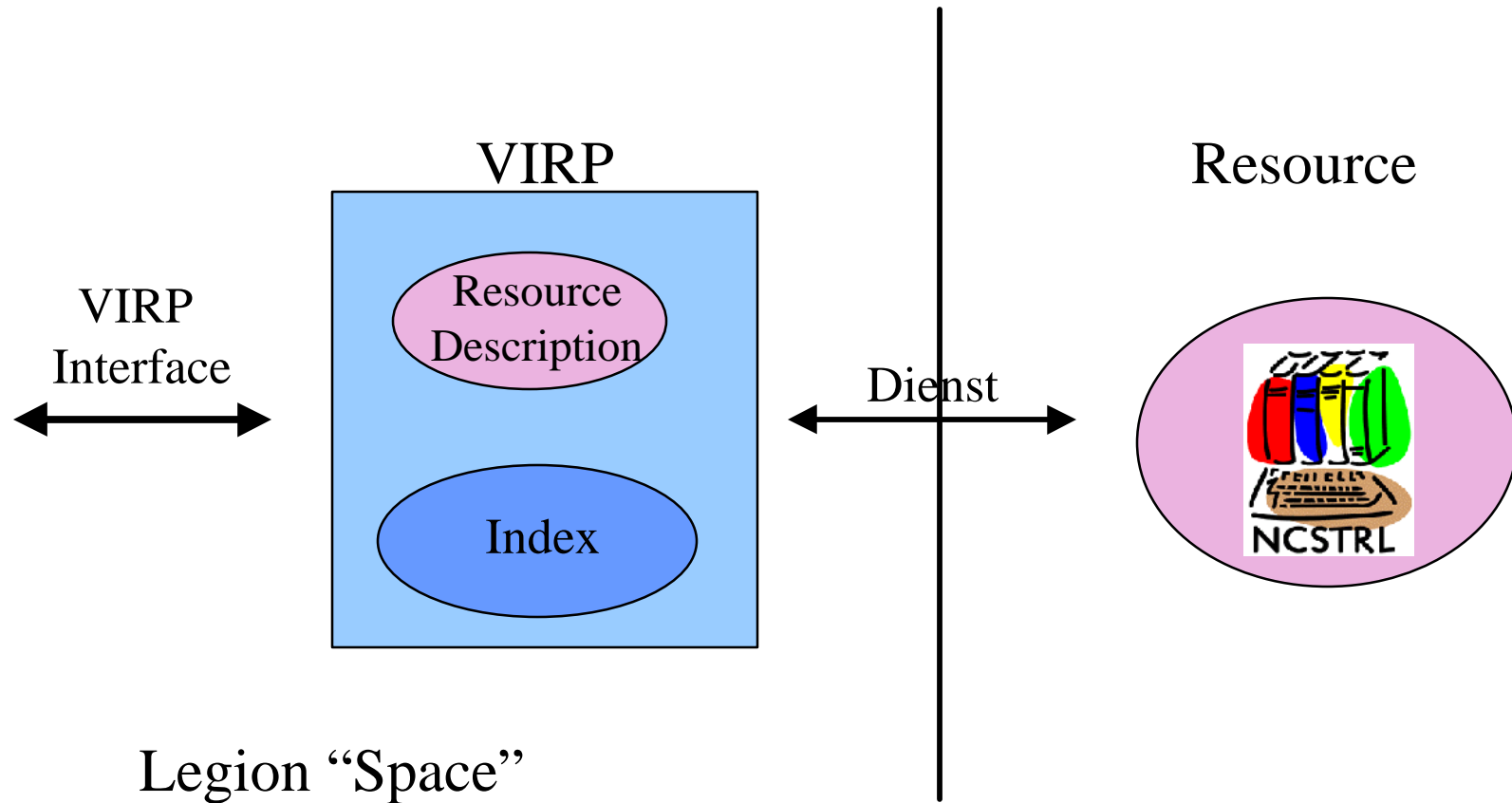
Implementation Status



Initial prototype

- deployed over NCSTRL**
- built on Legion metasystem**

The NCSTRL VIRP



The “Resource Description”



- **Exportable**
- **Given on demand by the VIRP**
- **Contains summary information about the Resource**
- **Can be “Individual” or “Composite”**



Handle to VIRP

Document Count

Author List

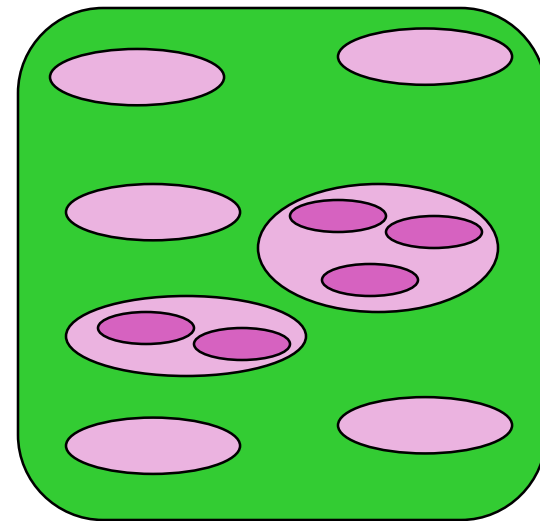
Site Name

Type of Resource

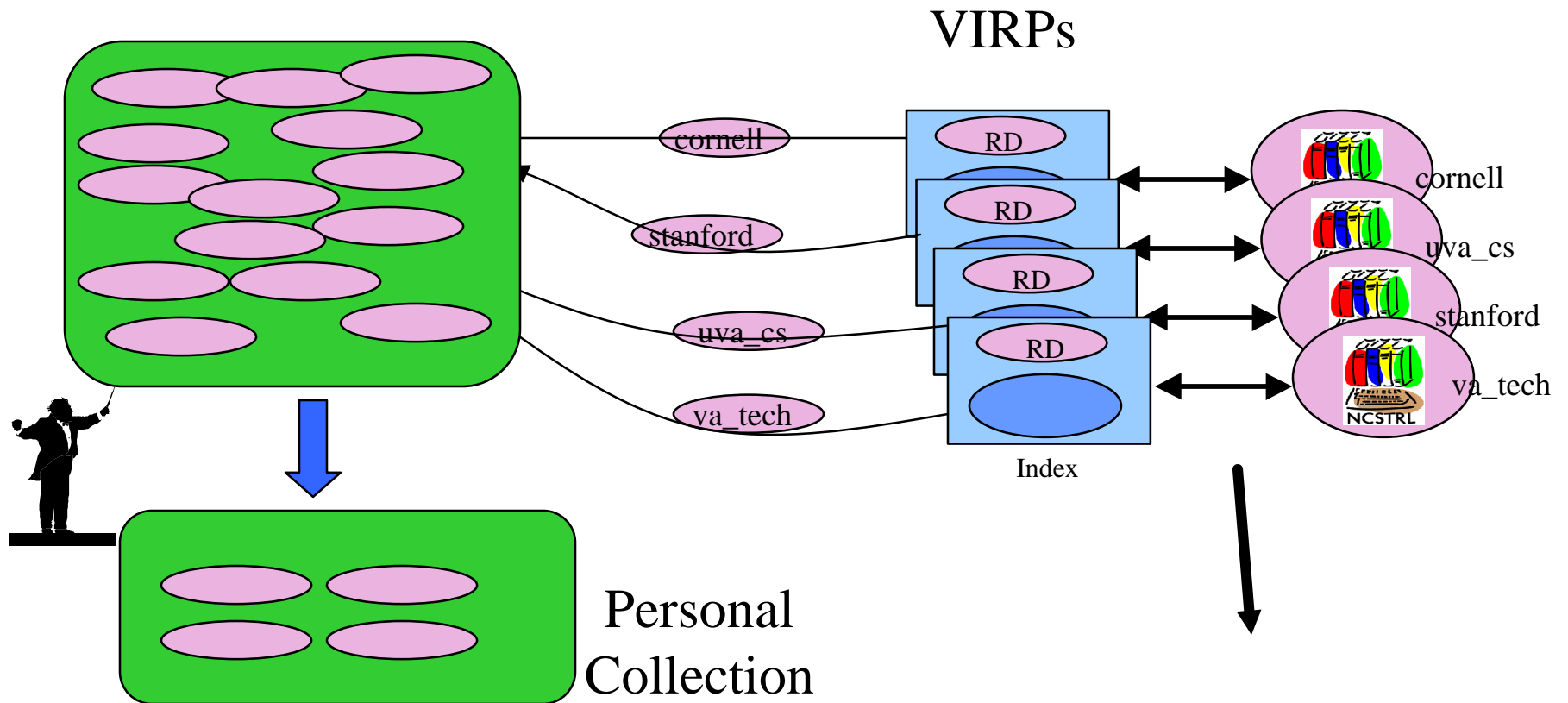
Personal Collection



- **Users manipulate Resource Descriptions**
- **PeC creation is interactive**
- **Persistence and sharing are important**



Current PIE



Implementation: Next Steps



- **Fully functional search**
- **Document Delivery**
- **Persistence (Legion)**
- **Multiple Hosts (Legion)**
- **Sharing of Personalized Collections (Legion)**

Research Activities



- **Distributed search**
 - query routing
 - database selection
 - evaluation
 - testbed specification
 - metrics
- **Information Exploration**
- **Metadata Management**

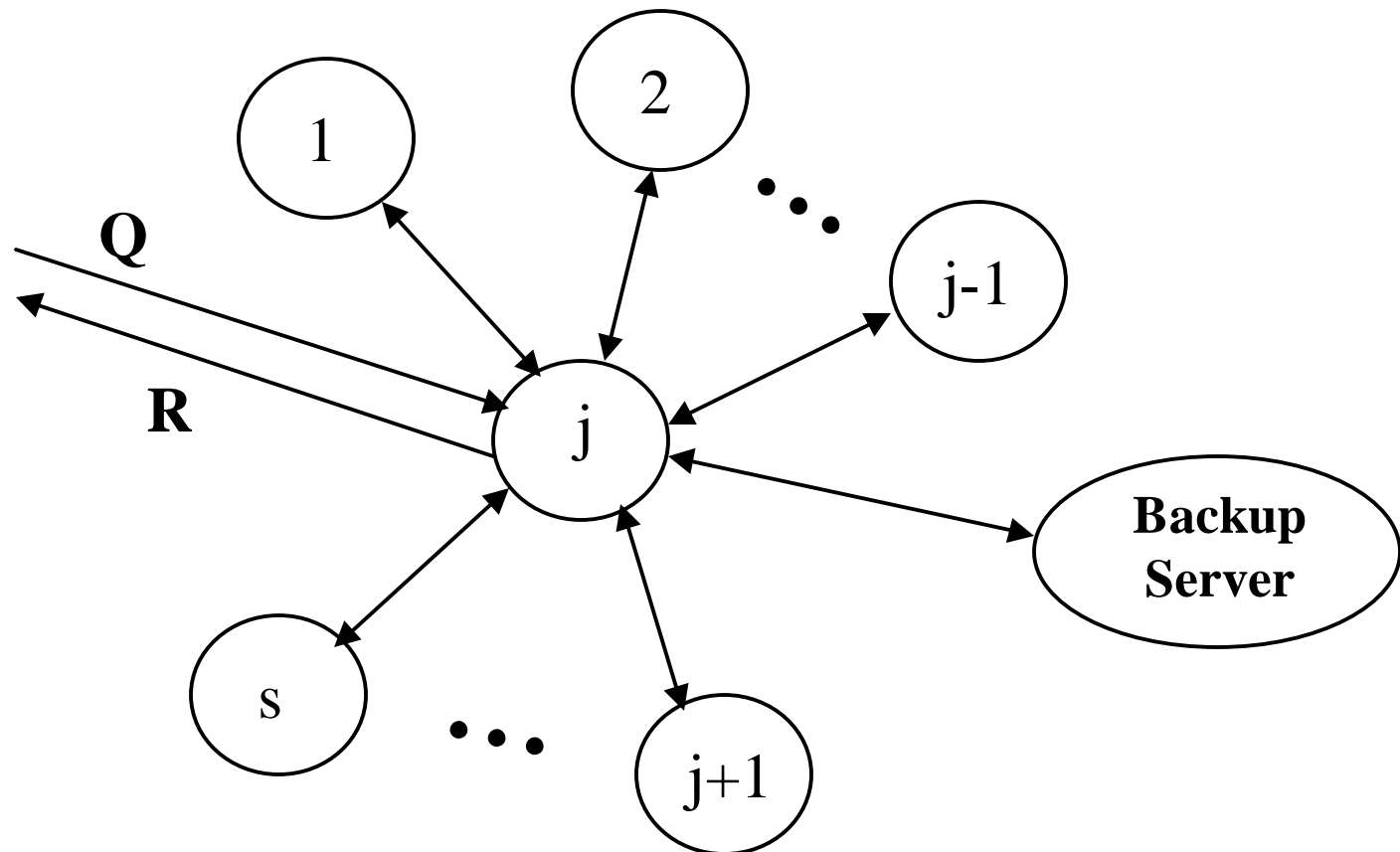
Distributed query processing



Performance issues along two dimensions

- **efficiency - how quickly is search performed?**
- **effectiveness - what is the quality of the results returned?**

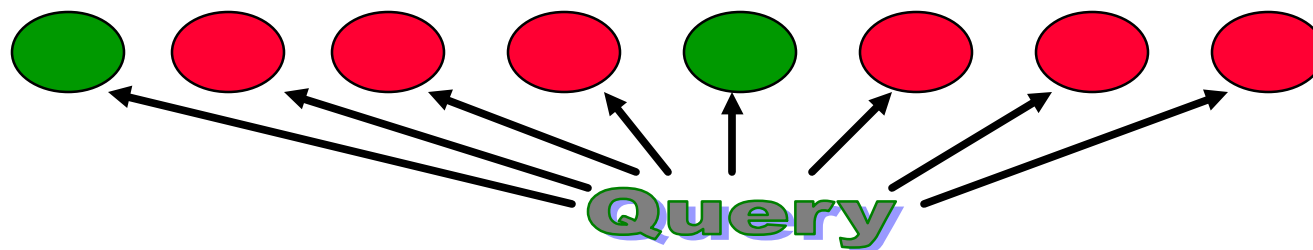
Distributed query processing



Database Selection



- Consider a distributed collection with queries frequently sent to all sites.
- For many queries, few sites contain matching documents.



Selection Criteria



- **Author last name**
- **Term in document title**

smith: Duke, Cornell, ...

jones: Univ. Hamburg, UVA, MIT

digital: Stanford, MIT, UC Berkeley, ...

library: UT Knoxville, Stanford, UC Irvine, ...

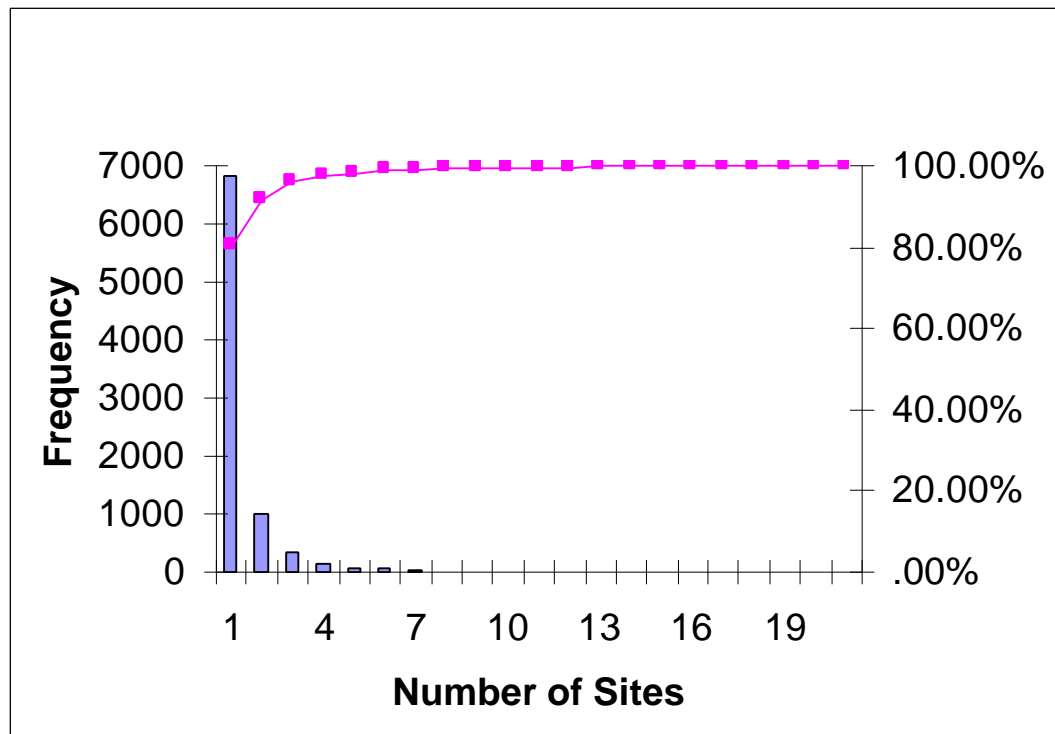
NCSTRL Example



Author Last Name	Standard only (39 sites)	Standard plus Lite (82 sites)
Creighton	0	0
French	1	3
Powell	2	3
Jones	5	12
Lee	15	25
Smith	15	24

Number of sites where query is processed

Distribution of Author Names



Sites	Names	Cumulative %
1	6835	80.07%
2	996	91.74%
3	354	95.89%
4	127	97.38%
5	72	98.22%
6	59	98.91%
7	22	99.17%
8	17	99.37%
9	12	99.51%
10	8	99.60%
11	6	99.67%
12	5	99.73%
13	6	99.80%
14	2	99.82%
15	2	99.85%
16	4	99.89%
17	0	99.89%
18	2	99.92%
19	2	99.94%
20	1	99.95%
More	4	100.00%

Results



- **Significant reduction in search space.**
- **Amount of information small enough to replicate at all sites.**
- **No change in search effectiveness.**
- **Transparent to the user.**

DL'98

Evaluating Selection Algorithms



Testbed requirements

- data collection
- set of queries (with relevance judged?)
- one or more baselines for comparison
 - optimal (U. Mass.), RBR (U. Va.)
 - ideal (Stanford)
 - SBR (U.Va.)
- metrics

Testbed Partition Statistics



Disk	Source	Number DB	Date Range	Total DB
1	WSJ (86-89)	29	12/86-11/89	67
	AP (89)	12	01/89-12/89	
	ZIFF	14	11/89-2/90	
	FR (89)	12	01/89-12/89	
	DOE			
2	WSJ (90-92)	22	04/90-03/92	54
	AP (88)	11	02/88-12/88	
	ZIFF	11	01/89-11/89	
	FR (88)	10	01/88-12/88	
3	AP (90)	12	01/90-12/90	116
	SJMN (91)	12	01/91-12/9	
	ZIFF			
	PAT	92	06/82-08/92	

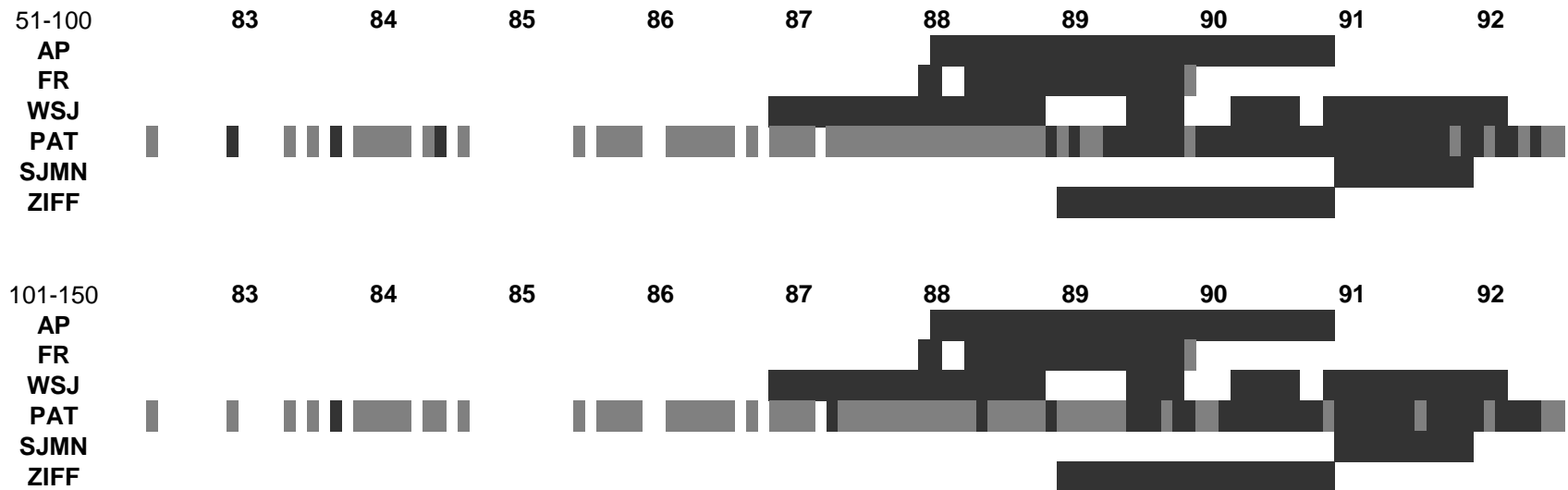
Testbed Topic Coverage



Source Disk	Topic Set					Number of Databases
	1-50	51-100	101-150	151-200	201-250	
1						67
2						54
3						116
1,2						120
2,3						170
1,2,3						236

Coverage of topics over TREC data, disks 1-3, through TREC-4. Note: Database ZIFF.89.11 is drawn from both disk 1 and 2.

Document and Query Coverage



Comparison of Ranks

Relevance Based Ranking			
Site	Rank	Mid-Rank	Merit
AP.88.12	1	1.00	21.00
WJ.91.08	2	2.00	14.00
AP.88.08	3	3.00	13.00
WJ.91.03	4	4.00	12.00
WJ.91.04	5	5.00	11.00
AP.88.10	6	6.50	10.00
WJ.91.01	7	6.50	10.00
WJ.88.06	8	8.00	9.00
WJ.88.09	9	9.50	8.00
WJ.90.08	10	9.50	8.00

Size Based Ranking			
Site	Rank	Mid-Rank	Merit
AP.88.05	1	1.00	8302
AP.88.03	2	2.00	8196
AP.88.06	3	3.00	8081
AP.88.10	4	4.50	7794
AP.88.04	5	4.50	7790
AP.89.05	6	6.50	7596
AP.89.10	7	6.50	7543
AP.89.01	8	8.50	7508
AP.88.08	9	8.50	7496
WJ.87.12	10	10.00	7384

gGloss Experiment



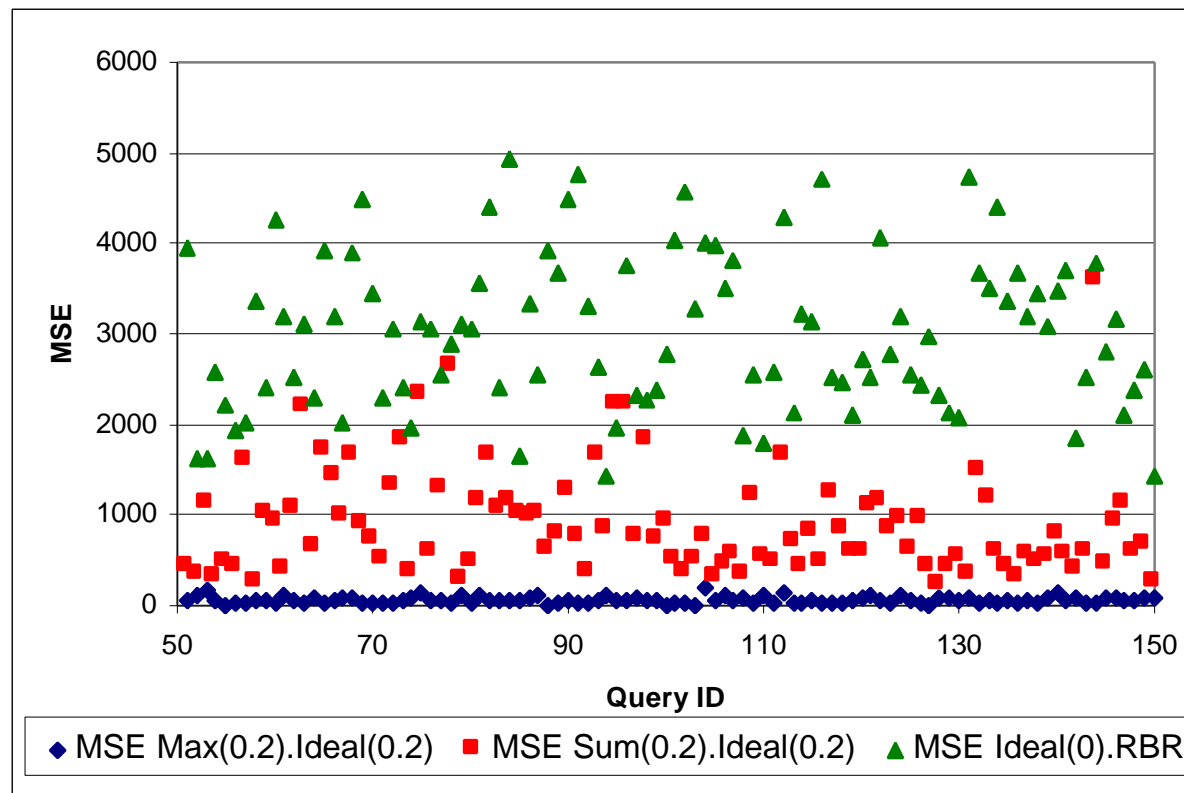
- **gGloss compared against three baselines**
 - ideal
 - **relevance based ranking (RBR)**
 - **size based ranking (SBR)**

Metrics



- **Mean squared error**
- **Recall and precision**
- **Rank correlation**

MSE



Recall and Precision Analogs



$$B_i = \text{merit}(q, db_{b_i})$$

$$E_i = \text{merit}(q, db_{e_i})$$

$$R_n = \frac{\sum_{i=1}^n E_i}{\sum_{i=1}^n B_i}$$

$$\hat{R}_n = \frac{\sum_{i=1}^n E_i}{\sum_{i=1}^{n^*} B_i}$$

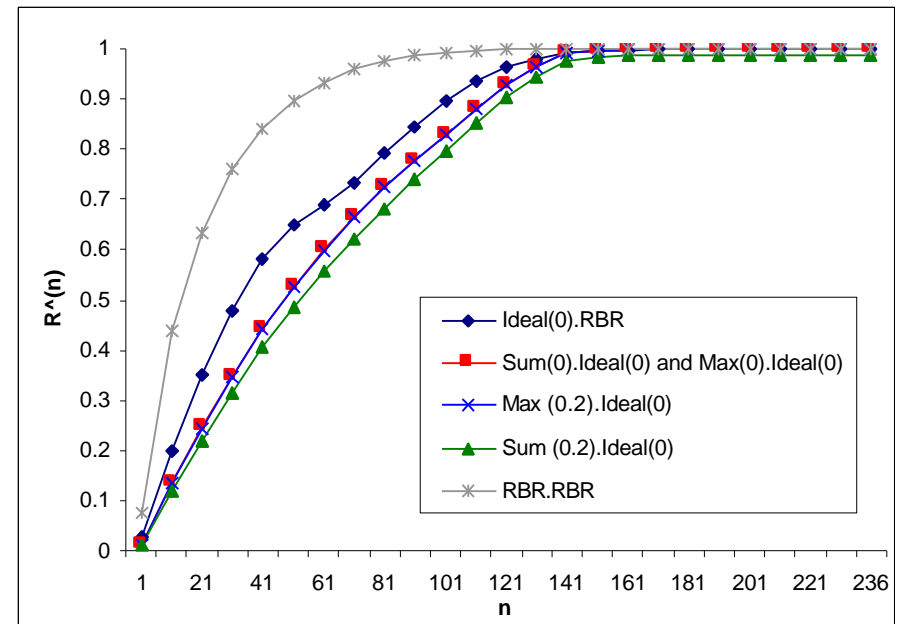
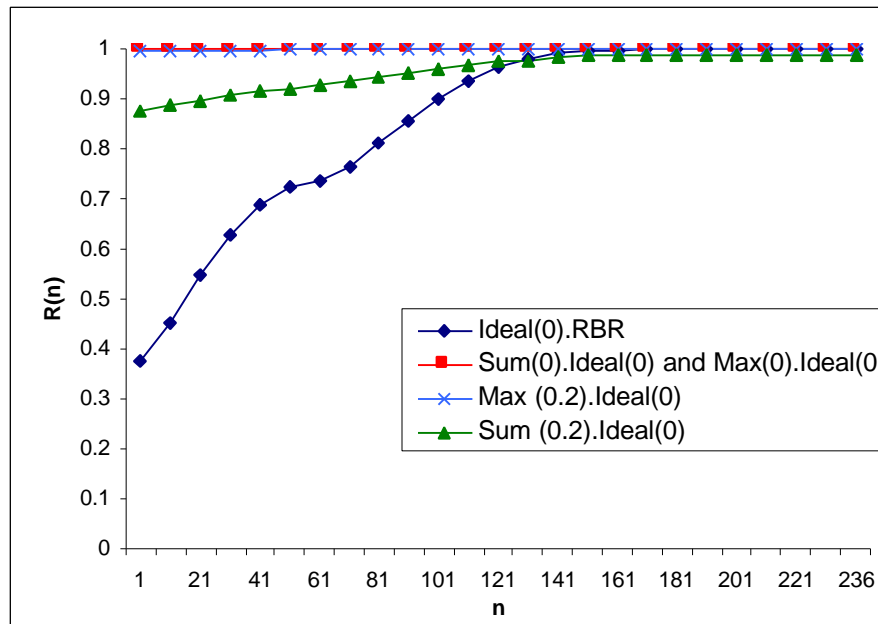
$$P_n = \frac{|\{db \in \text{Top}_n(E) \mid \text{merit}(q, db) > 0\}|}{|\text{Top}_n(E)|}$$

Example

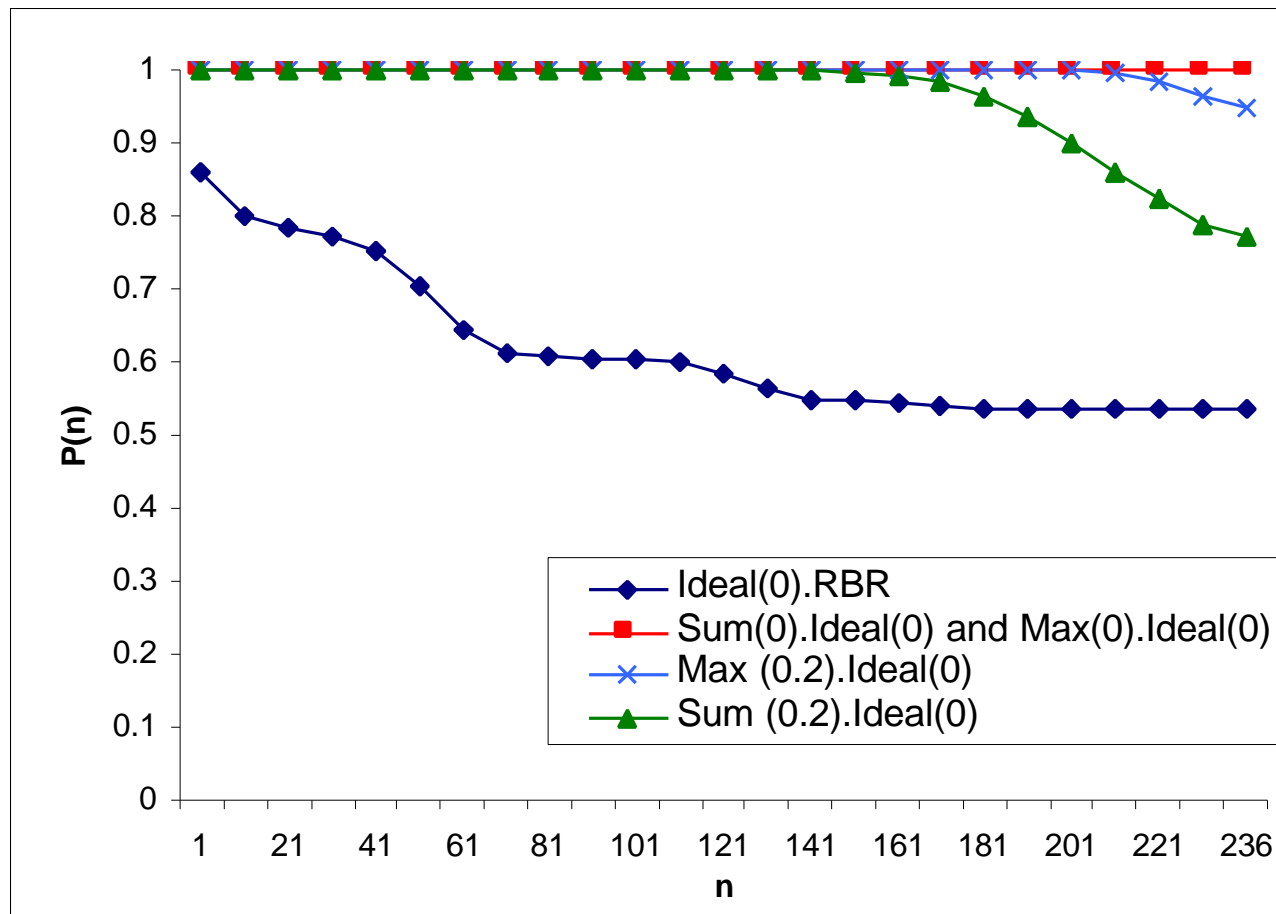


Baseline	5	3	1	0
Estimate	3	5	0	1
$R_1 = 3/5$	$\hat{R}_1 = 3/9$	$P_1 = 1/1$		
$R_2 = 8/8$	$\hat{R}_2 = 8/9$	$P_2 = 2/2$		
$R_3 = 8/9$	$\hat{R}_3 = 8/9$	$P_3 = 2/3$		
$R_4 = 9/9$	$\hat{R}_4 = 9/9$	$P_4 = 3/4$		

Recall Analogs



Precision Analog



Spearman's R

Assumes few ties.

$$R = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}$$

Where D_i is the degree of disagreement.

$$D_i = U_i - V_i$$

U and V are the sets of ranks.

Results: SBR vs. Ideal.0

<u>QID</u>	<u># Sites</u>	<u>R₁</u>	<u>R₂</u>	<u>Z₁</u>	<u>Z₂</u>
51	236	0.9678	0.9671	14.837	14.826
52	236	0.9707	0.9701	14.880	14.871
53	236	0.9482	0.9470	14.535	14.517
54	236	0.9824	0.9821	15.060	15.056
55	236	0.9811	0.9808	15.040	15.035
56	236	0.9489	0.9480	14.547	14.532
57	236	0.9763	0.9759	14.966	14.960
58	236	0.9834	0.9831	15.076	15.071
59	236	0.9865	0.9862	15.122	15.119
60	236	0.9939	0.9938	15.236	15.235

Query 1 Rankings: SBR vs. Ideal.0 vs. Ideal.2

SBR

<u>Rank</u>	<u>Site</u>	<u>Merit</u>
1	AP.88.05	8302
2	AP.88.03	8196
3	AP.88.06	8081
4	AP.88.10	7794
5	AP.88.04	7790
6	AP.89.05	7596
7	AP.89.10	7543
8	AP.89.01	7508
9	AP.88.08	7496
10	WJ.87.12	7384

Ideal.0

<u>Rank</u>	<u>Site</u>	<u>Merit</u>
1	AP.88.03	3284.63
2	AP.88.06	3230.47
3	AP.88.05	3200.08
4	AP.88.04	3153.97
5	AP.88.10	3018.42
6	AP.89.10	3015.12
7	AP.89.01	3000.82
8	AP.89.05	2995.64
9	AP.88.07	2874.42
10	AP.89.06	2871.30

Ideal.2

<u>Rank</u>	<u>Site</u>	<u>Merit</u>
1	AP.88.03	2968.65
2	AP.88.06	2914.32
3	AP.88.05	2854.79
4	AP.88.04	2844.03
5	AP.89.10	2713.71
6	AP.89.01	2707.00
7	AP.88.10	2701.68
8	AP.89.05	2693.29
9	AP.89.06	2596.15
10	AP.88.07	2585.37

Summary of results



- **gGloss estimates correlate very well with ideal**
- **gGloss does not correlate well with RBR**
- **very high correlation between gGloss and SBR**

SIGIR'98

Information Exploration



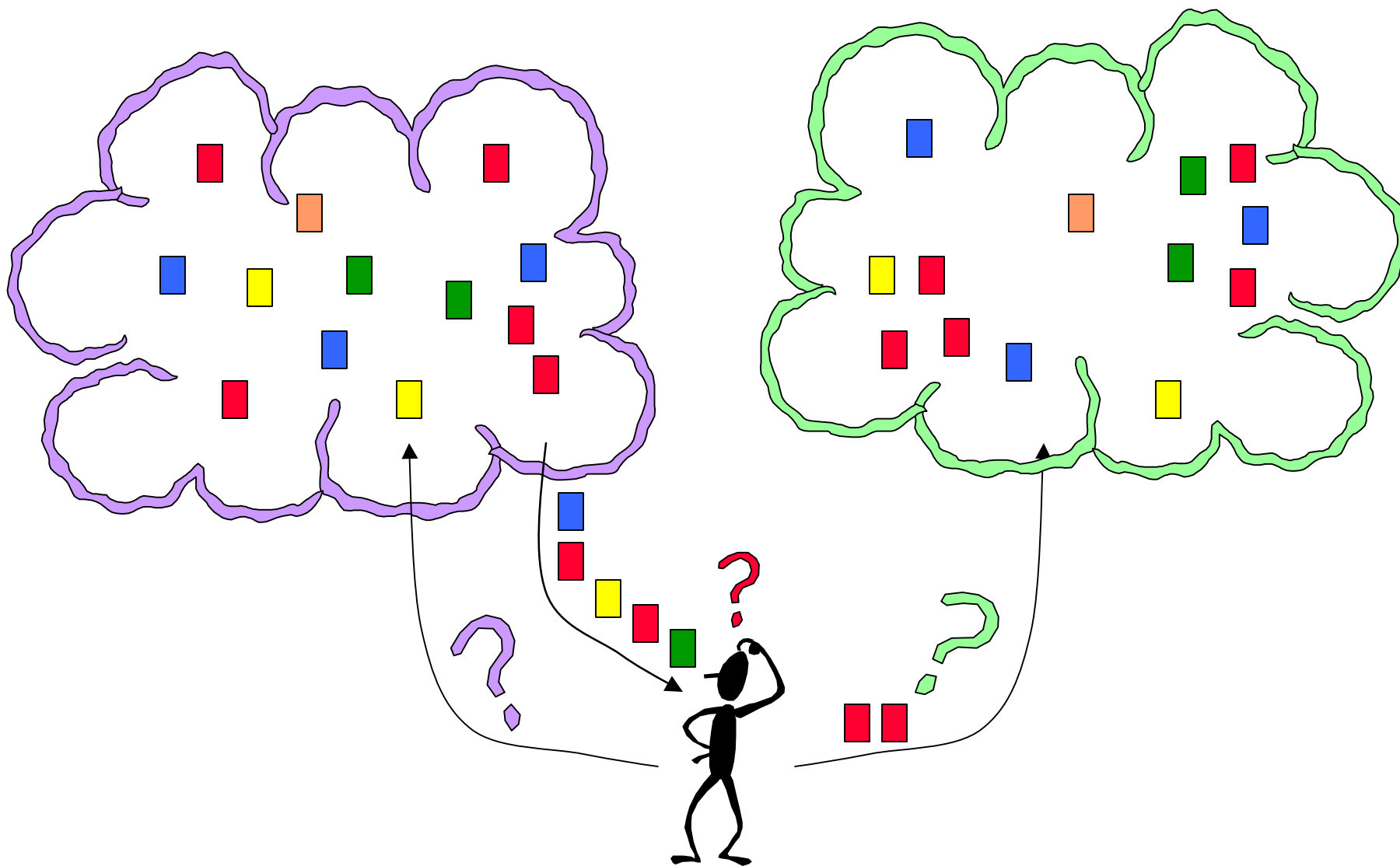
- **Traditional text-based searching sometimes frustrating**
 - Searchers may be unfamiliar with collection vocabulary
- **Utilize multiple viewpoints of a document collection**
 - term-based and topic-based
 - readily available and intuitively useful

Information Exploration (cont.)



- **Move among viewpoints to refine searches**
- **Use relevant documents located in one view to bootstrap exploration of alternate views**
- **Document relationships may be more apparent in one view than another**

User Interaction



Metadata Management

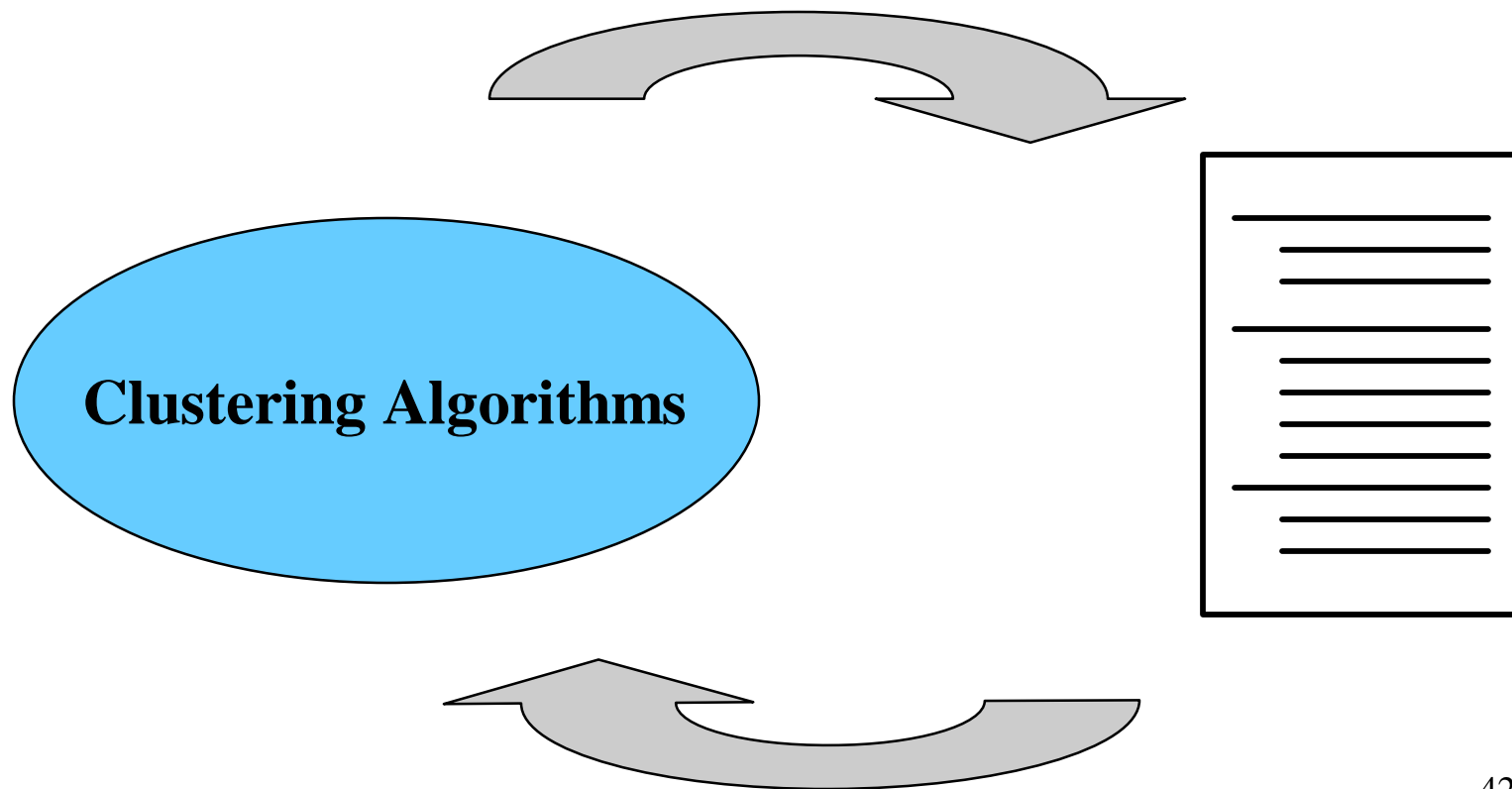


- **Authority work: detecting variant names for unique entities in the database.**
- **Authority files: files that maintain the correspondence between all allowable forms for strings in a particular bibliographic field.**

Authority file generation



- **Iterative, human-in-the-loop process**



Clustering Alternatives



- **Absolute edit distance**

$$e(u, v) \leq d$$

- **Relative edit distance**

$$e(u, v) \leq a \min(u, v)$$

- **Approximate word matching**

Hampden-Sydney College		
Hampden-Sydney College	23	
Hampden-Sydney College, Hampden-Sydney	8	
Leander McCormick Observatory, Charlottesville		
Leander McCormick Observatory, Charlottesville	30	
Leander J. McCormick Observatory, Charlottesville	2	
College of William and Mary, Williamsburg		
College of William and Mary, Williamsburg	22	
College of William and Mary, Williamsburgh	1	
Old Dominion University, Norfolk		
Old Dominion University, Norfolk	12	
Old Dominion University Research Foundation, Norfolk	1	
Science Applications, Inc., McLean		
Science Applications, Inc., McLean	9	
Science Applications International Corporation, McLean	7	
Analytical Mechanics Associates, Inc., Hampton		
Phoenix Corporation, McLean		
Phoenix Corporation, McLean	6	
BDM Corporation, McLean	3	
Mitre Corporation, McLean	1	
Computational Physics, Inc., Fairfax		
Computational Physics Inc., Fairfax	1	
Computational Physics, Inc., Fairfax	4	
Computational Physics, Inc., Annandale	2	
Beers Associates, Inc., Reston		
Science Applications International Corporation, Hampton		
Sweet Briar College, Sweet Briar		
Institute for Computer Applications in Science and Engineering, Hampton		
Institute for Computer Applications in Science and Engineering, Hampton	3	
Lockheed Engineering and Science, Hampton	1	

Examples
from the
ADS.

For more information



<http://www.cs.virginia.edu/~cyberia>

- **Distributed search**
 - SIGIR'98, DL'98, SIGMOD'99 (in prep.)
- **Information Exploration**
 - CHI'98 Information Exploration Workshop
- **Metadata Management**
 - CIKM'97, EuDL'97, ICDE'99 (submitted)

Recap



- **PIE is user-centric**
- **Research focus is on effective and efficient distributed searching**
 - retrospectively
 - SDI
- **Via simulation studies and direct observation of deployed prototypes**