Distributed IR

IS4200/CS6200
Information Retrieval

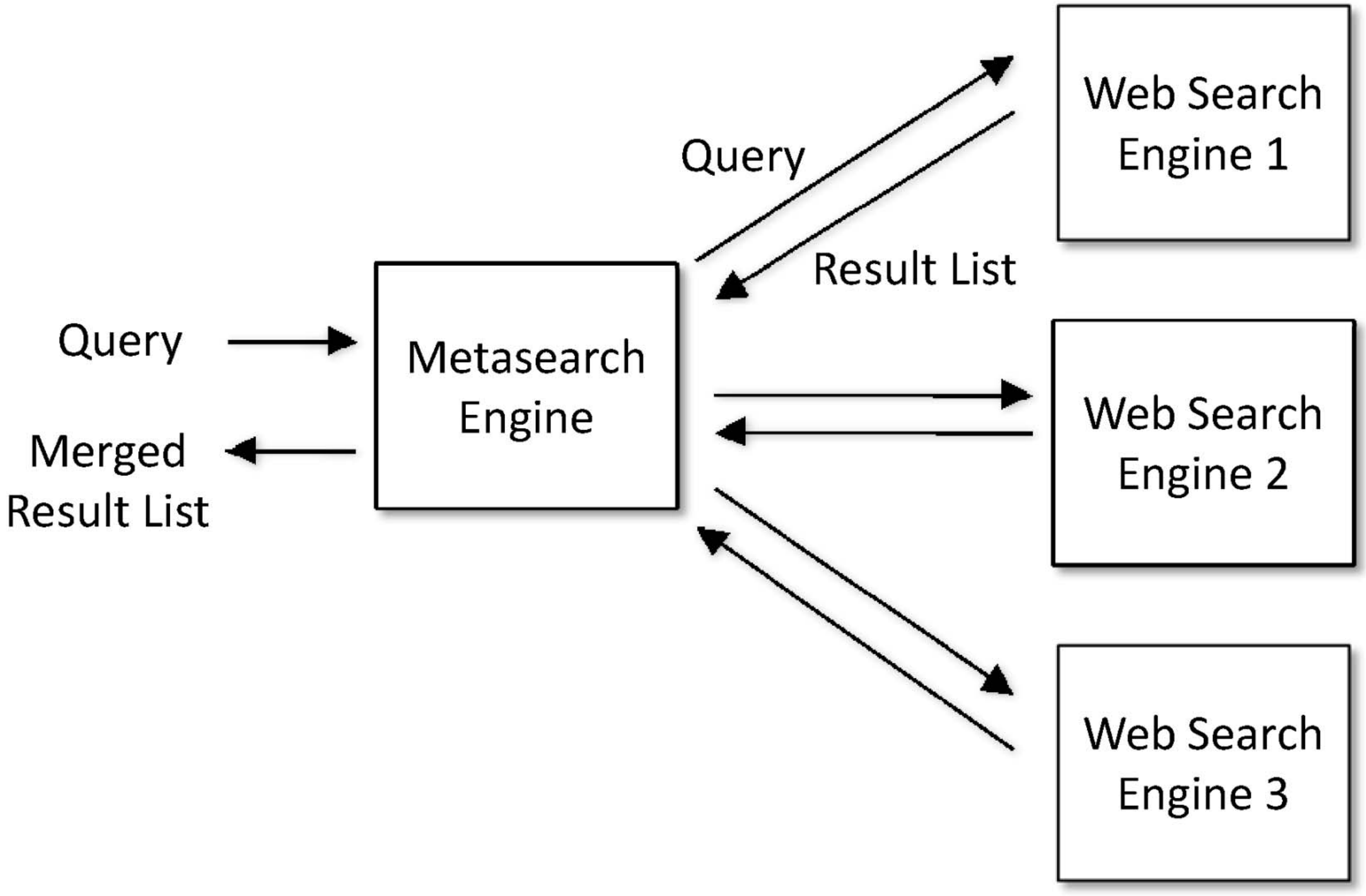
Distributed IR

- IR is usually views as searching a single collection of documents
- What is a collection?
  - A single source, e.g.,Wall Street Journal? (What time period?)
  - A single location, e.g., Snell Library?
  - A set of libraries, e.g., all Northeastern Libraries?

Distributed IR

- What is distributed search?
  - Searching over networks or communities of nodes
  - Each node contains some searchable data
- Distributed search applications
  - Locally distributed search (on LAN, for efficiency)
  - Metasearch (WAN, Internet, federated architecture)
    - Node: search engines
    - Data: index
  - Peer−to−peer (P2P)
    - Node: user machines
    - Data: index, files, etc.

Metasearch Architecture



Distributed IR

- Resource representation
  - How is a node represented?
- Resource selection
  - Which nodes should be searched for the given information need?
- Result merging
  - How do we combine the results obtained from all of the nodes?

Distributed IR

- Partition large collections across processors
  - To increase speed
  - Because of political or administrative requirements
- Networks, with hundreds or thousands of collections
  - Consider number of collections indexed on the Web
- Heterogeneous environments, many IR systems
- Economic costs of searching everything at a site
- Economic costs of searching everything on a network

Issues

- Resource representation
  - Contents, search engine, services, etc
- Resource selection
  - Deciding which collection(s) to search
  - Ranking collections for a query
  - Selecting the best subset from a ranked list
- Searching
  - Interoperability, cooperativeness
- Result merging: Merging a set of document rankings
  - Different underlying corpus statistics
  - Different search engines with different output information
- Metrics:
  - Generality, effectiveness, efficiency, consistency of results, amount of manual effort, etc.

Collection Selection

- Single Site / LAN / Few Sites
  - Select everything
  - Group manually (and select manually)
  - Rule-based selection
  - Relevant document distribution (RDD)
  - Query Clustering
  - Query Probing
- Many Sites / WAN / Internet
  - Content-based collection ranking (and selection)

Selection: Exhaustive

- Found in LANs, e.g. where a large collection is partitioned
- Works well with the unranked Boolean model
  - Result set is the union of all search results
- Can work with statistical models
  - Merge-sort all search results, to obtain merged ranked list
  - But, scores from different databases aren't comparable,due to different corpus statistics, e.g., idf, avg_doclen
  - Scores can be made comparable by imposing one set of corpus statistics on all databases, e.g., global statistics, first database
- Ignores costs of searching collections
  - e.g., time, money
- Does not scale to WAN / Internet
- Some parallelism by distributing search

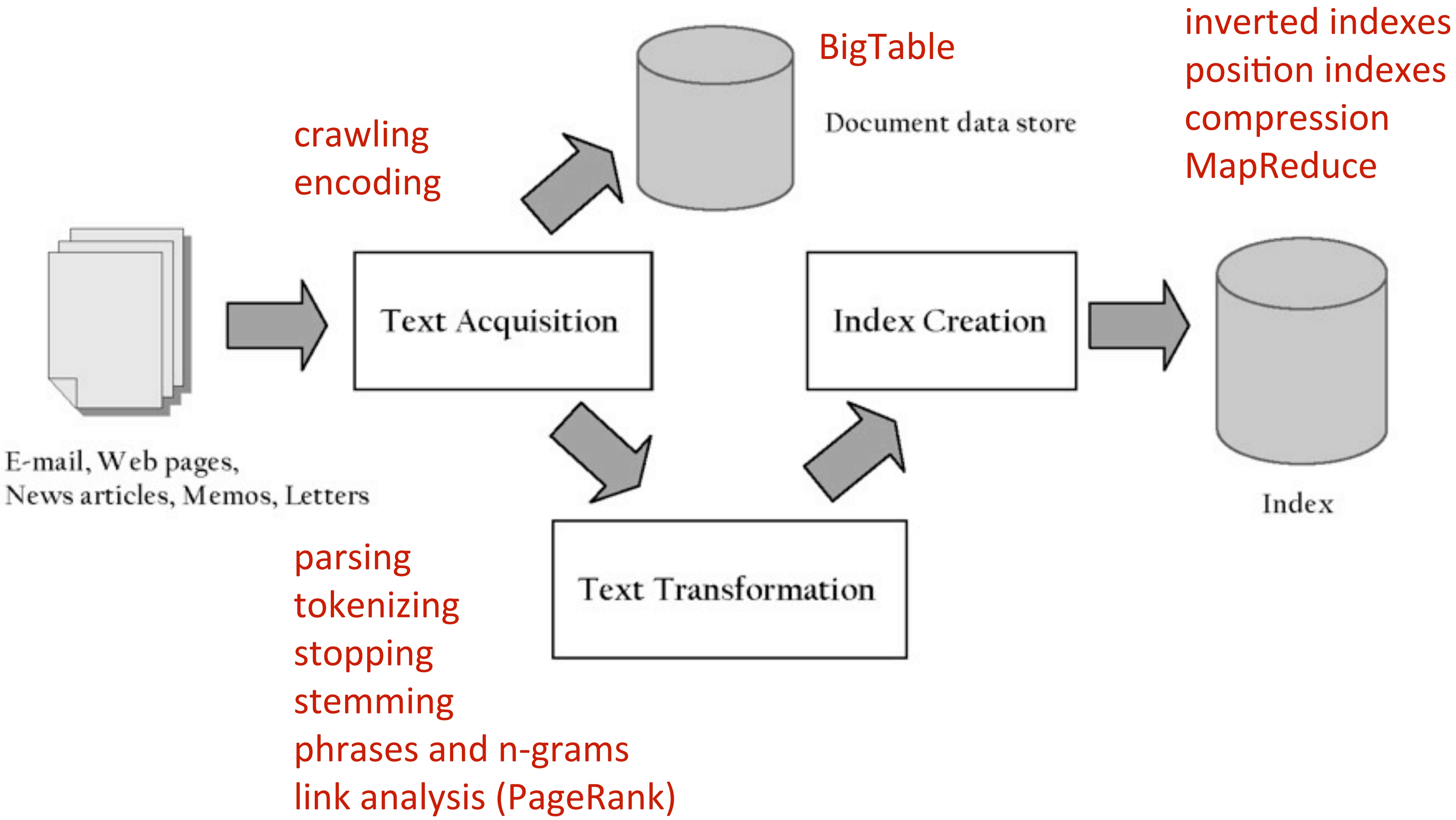
Selection: Manual

- Collections are organized into groups with a common theme
  - e.g., finance, technology, appellate court decisions
- User selects which group to search
- Found in commercial service providers
  - e.g., Dialog, WestLaw
- Groupings determined manually
  - time consuming, inconsistent groupings, coarse groupings, not good for unusual information needs
- Groupings determined automatically
  - Broker agents maintain a centralized cluster index by periodically querying collections on each subject
  - Automatic creation, better consistency, coarse groupings
  - Not good for unusual information needs

Selection: Rule-based

- The contents of each collection are described in a knowledge-base
  - ✤ few details provided by authors of such systems
- A rule-based system selects the collections for a query
  - ✤ few details provided by authors of how this works
- CONIT, a research system, never deployed widely
  - ✤ tested on static and homogeneous collections
  - ✤ time consuming to create
  - ✤ inconsistent selection if rules change
  - ✤ coarse groupings so not good for unusual information needs

Indexing Process



E-mail, Web pages,
News articles, Memos, Letters

crawling
encoding

Text Acquisition

BigTable

Document data store

parsing
tokenizing
stopping
stemming
phrases and n-grams
link analysis (PageRank)

Text Transformation

Index Creation

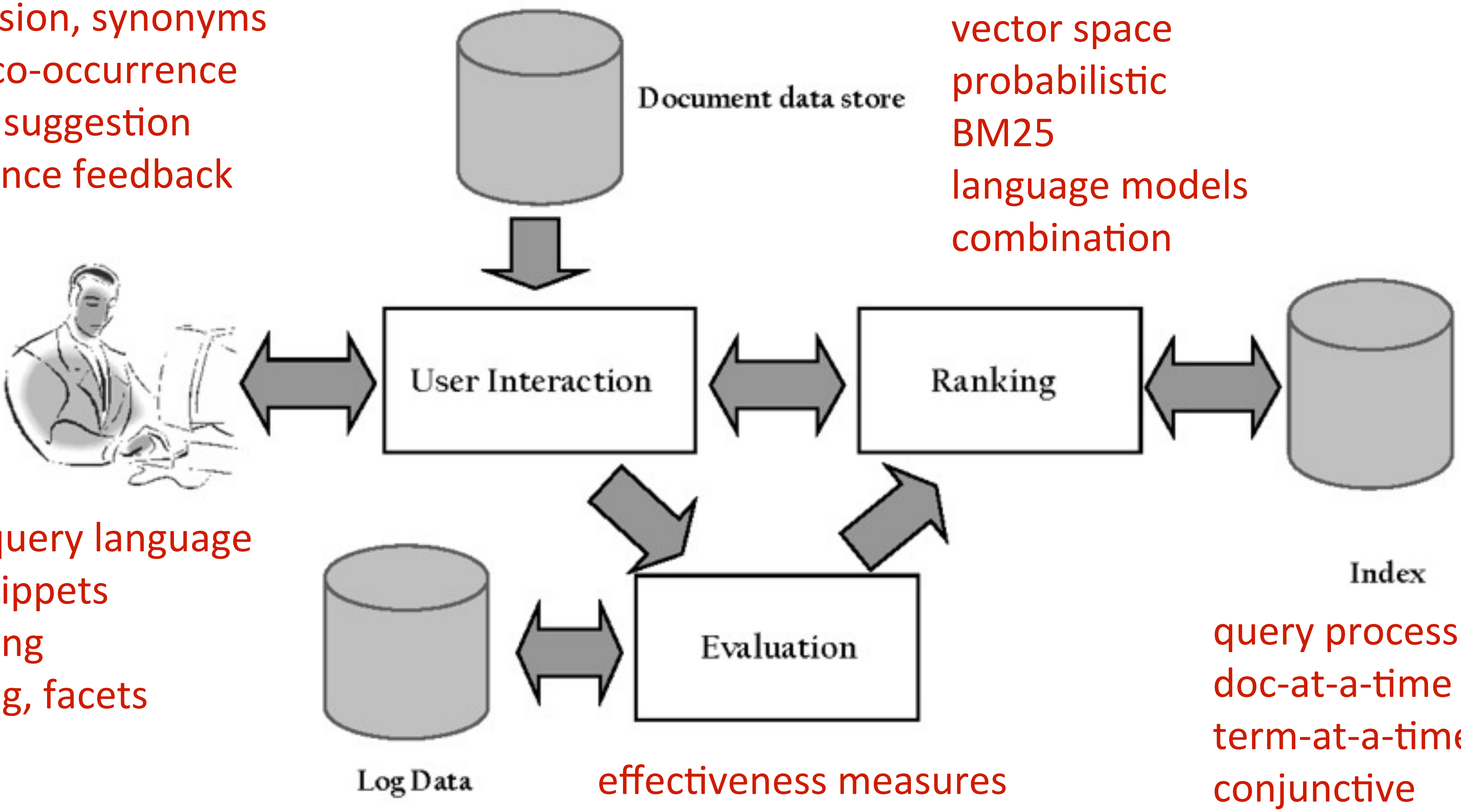
inverted indexes
position indexes
compression
MapReduce

Index

Query Process

Information needs
query transformation
spelling correction
expansion, synonyms
term co-occurrence
query suggestion
relevance feedback

retrieval models
Boolean
vector space
probabilistic
BM25
language models
combination



Galago query language
result snippets
advertising
clustering, facets

query logs

effectiveness measures
recall, precision, MAP
NDCG, significance

query processing
doc-at-a-time
term-at-a-time
conjunctive
optimization

Current Research Issues

- Understanding queries
  - NLP and queries, question answering, “semantic search”, query reformulation representations, query sessions, diversity, mapping queries to structure, rare queries, query similarity, query suggestion, genre classification
- Retrieval models
  - Learning to rank, Markov Random Field model, variations of language models, filtering models

Current Research Issues

- Evaluation
  - New metrics for new tasks (e.g., diversity, sessions), crowdsourcing, simulation, games
- New applications
  - Entity search, social search, personal search, multimedia search, aggregated search, opinion retrieval
- New architectures
  - Real-time search, mobile search, MapReduce

Careers and Study in IR

- Here: ML course, NLP course, independent study
- Graduate degrees: many possibilities, here, Umass Amherst, CMU, UIUC
- Careers:
  - Industry: Google, Microsoft, Yahoo, Amazon, Ebay, LinkedIn, Facebook, Twitter, etc. (all levels from B.S. to Ph.D.)
  - Academic: More in ML, "information" schools, Europe