An Introduction to Web Retrieval

Ricardo Baeza-Yates

Yahoo! Labs

Barcelona, Spain

ESSIR 2011, Koblenz, Germany

Contents

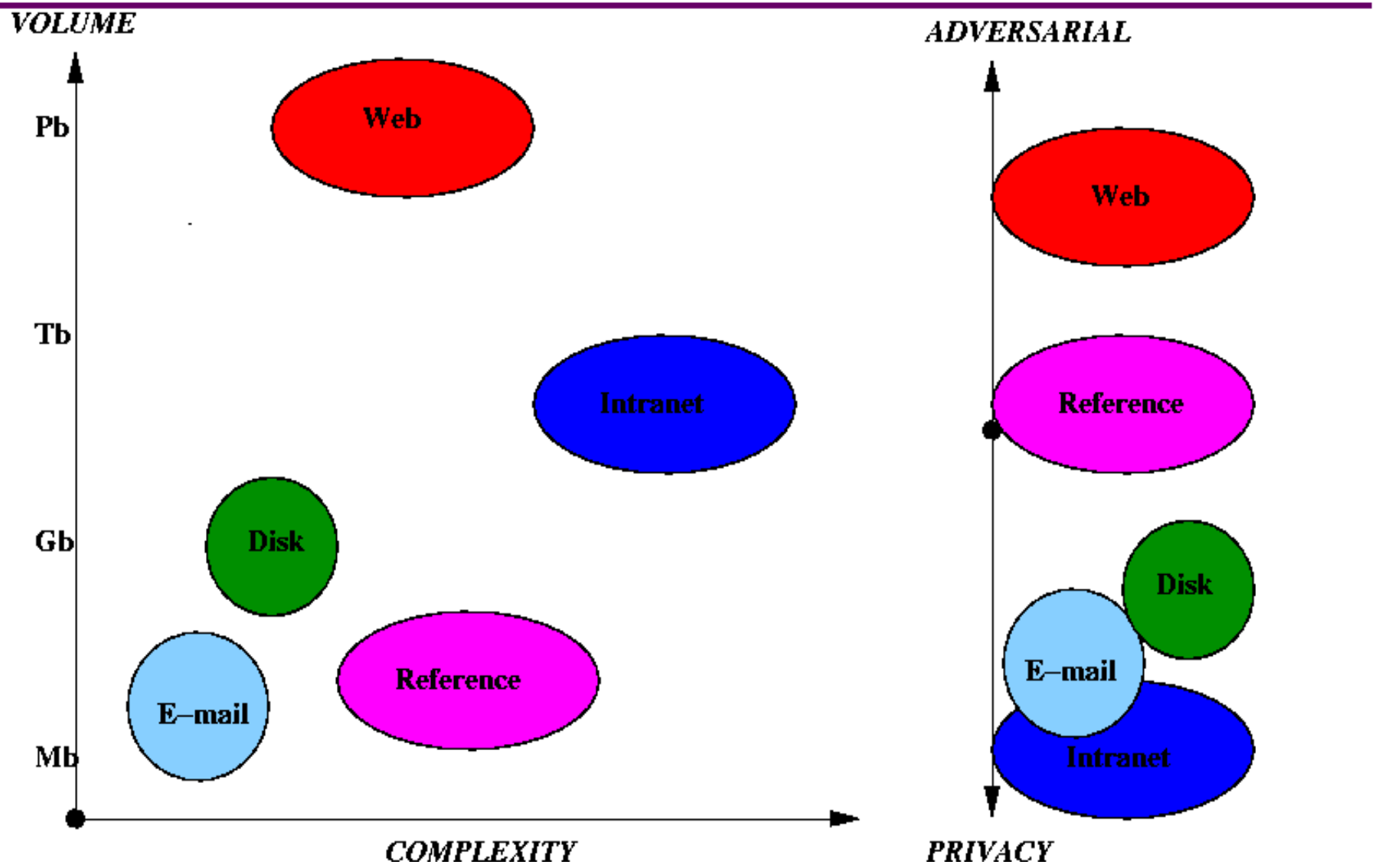
1. The Web
   - Basic concepts, wisdom of crowds, the long tail, advertising, Web content, Web spam
2. Web Search
   – Basic architecture, scalability, Web queries
3. Ranking
   – Link analysis, mixing features
4. Crawling
   – Goals, algorithms, evaluation

(1) The Web

Internet and the Web Today

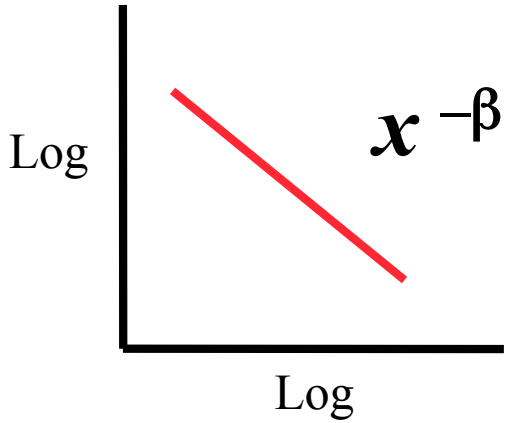
- **Between 1 and 2.5 billion people connected**
  - 5 billion estimated for 2015
- **More than 2 billion mobile phones today**
  - At least 500 million with Internet
- **Internet traffic has increased 20 times in the last 5 years**
- **Today there are more than 400 million Web servers**
- **The Web is in practice unbounded**
  - Dynamic pages are unbounded
  - Static pages over 20 billion?

Different Views on Data



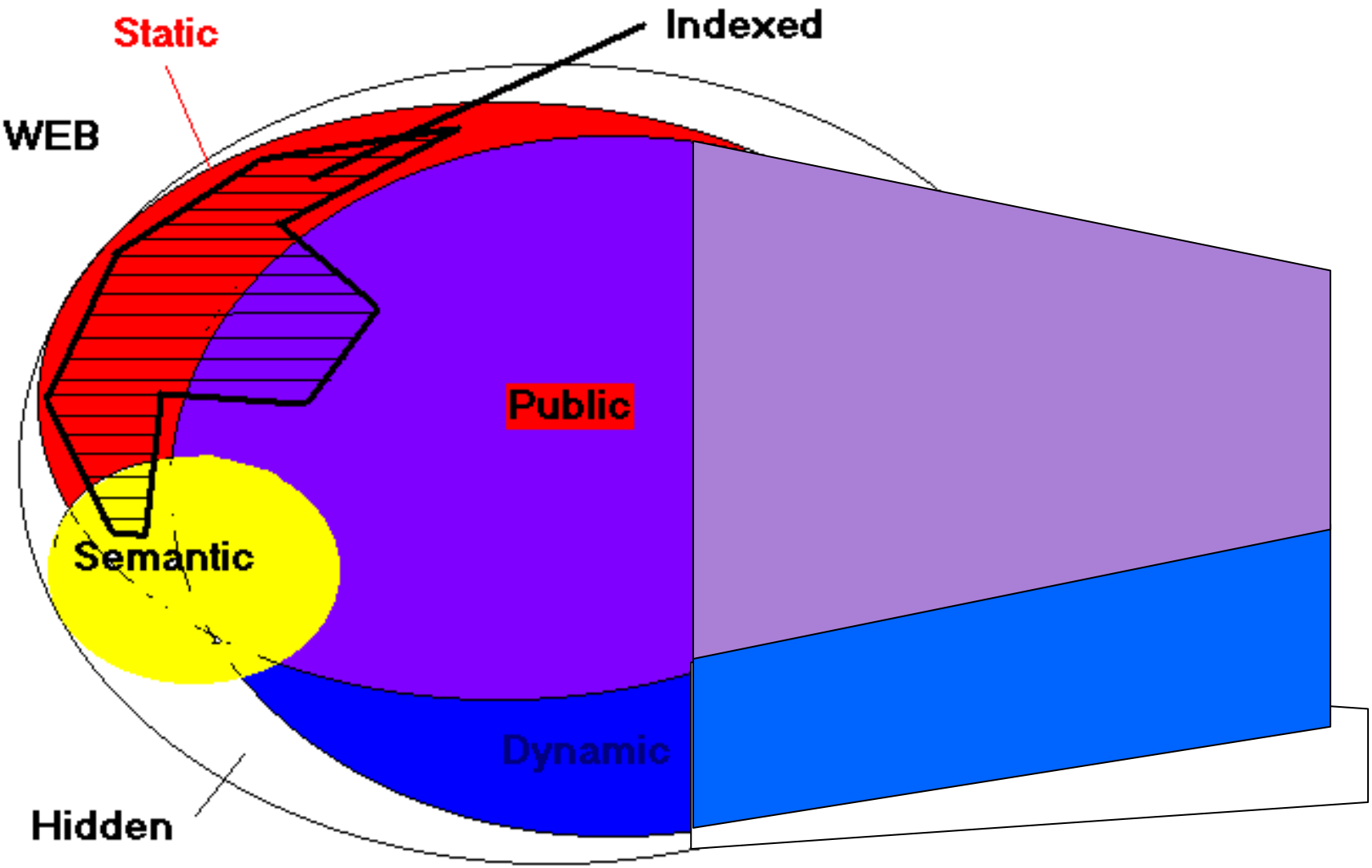
The Web

- **Largest public repository of data**
- **Today, there are more than 463 million Web servers (Aug 2011) and more than 820 million hosts (Jan 2011)**
- **Well connected graph with out-link and in-link power law distributions**

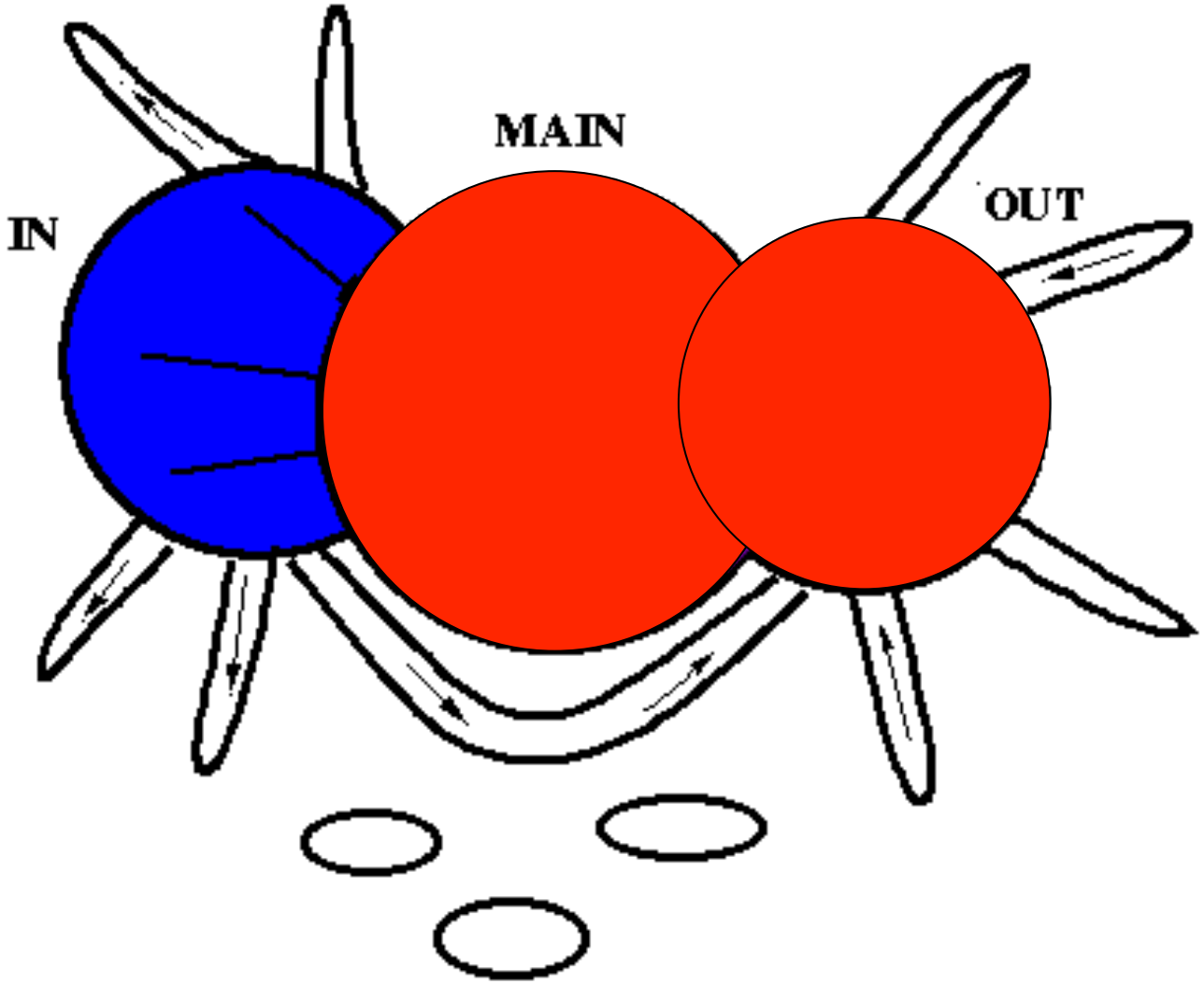


Self-similar &
Self-organizing

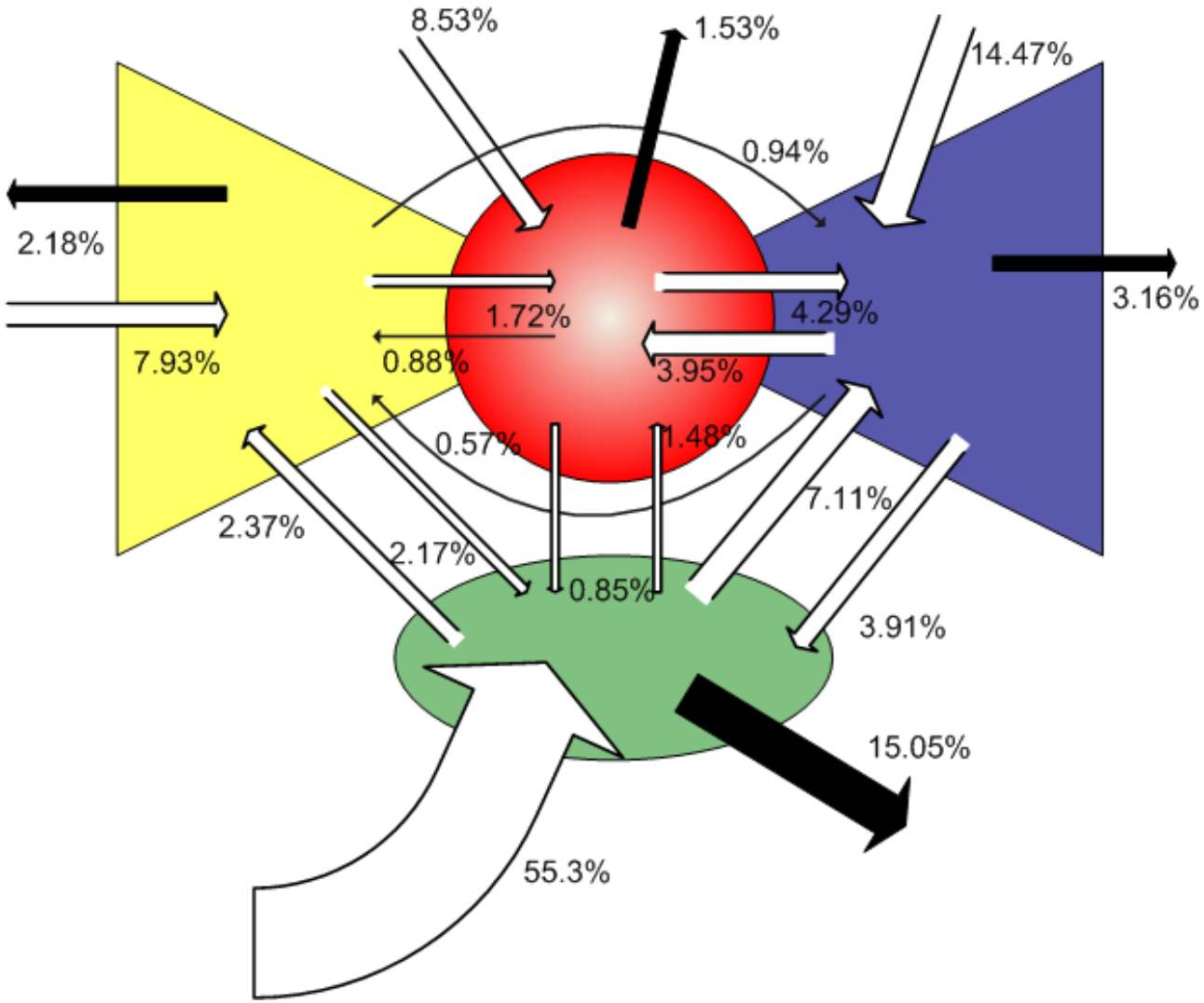
The Different Facets of the Web



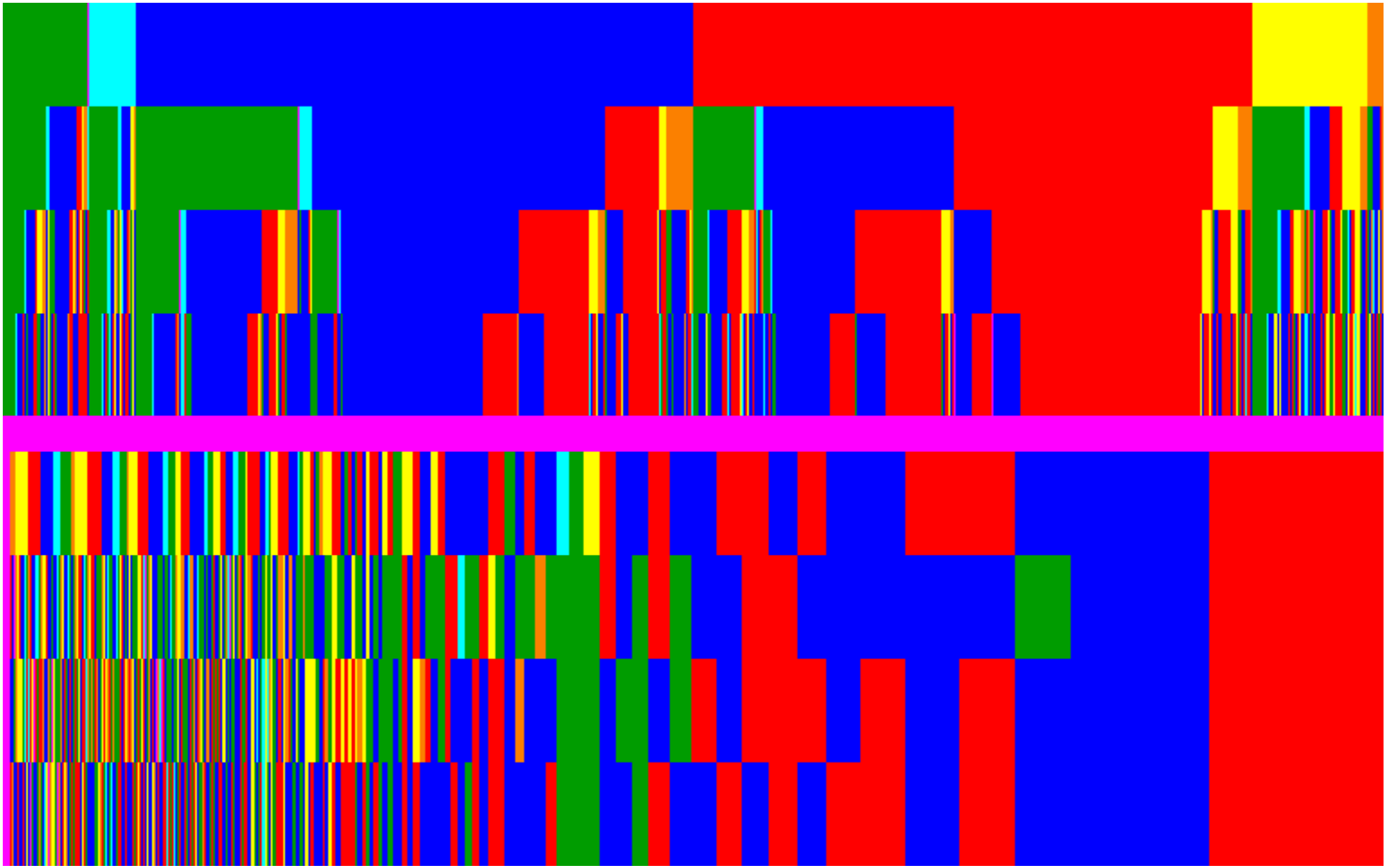
The Structure of the Web



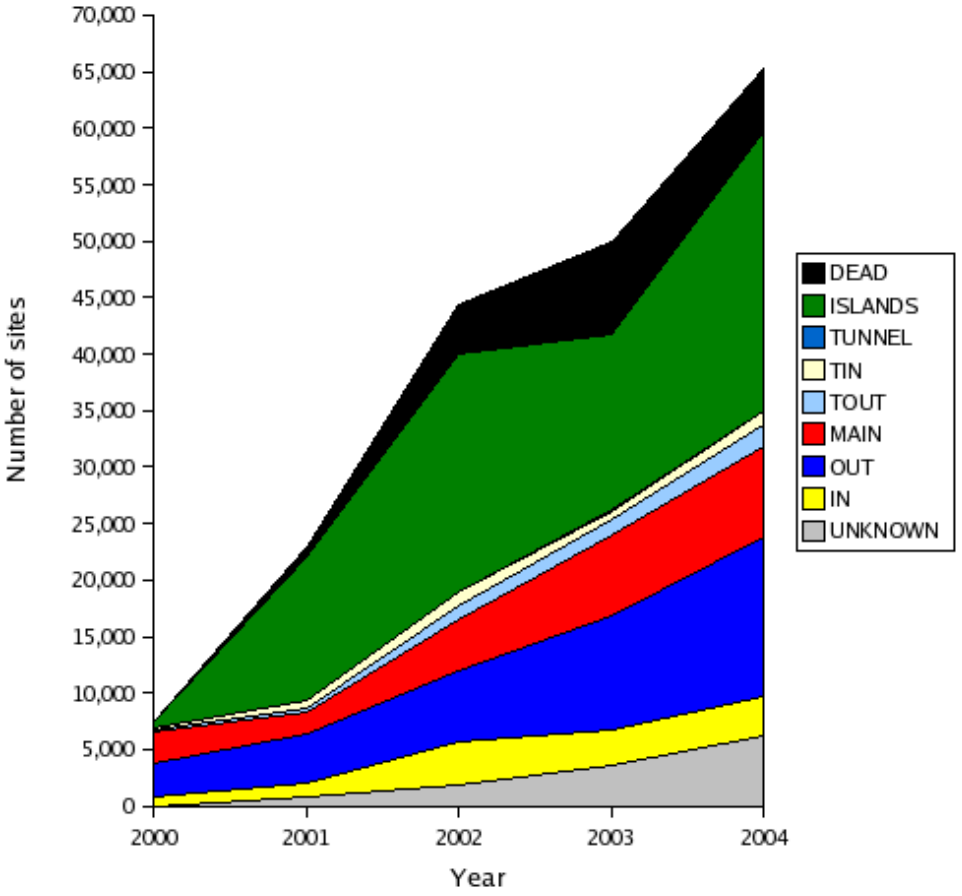
Structure Macro Dynamics

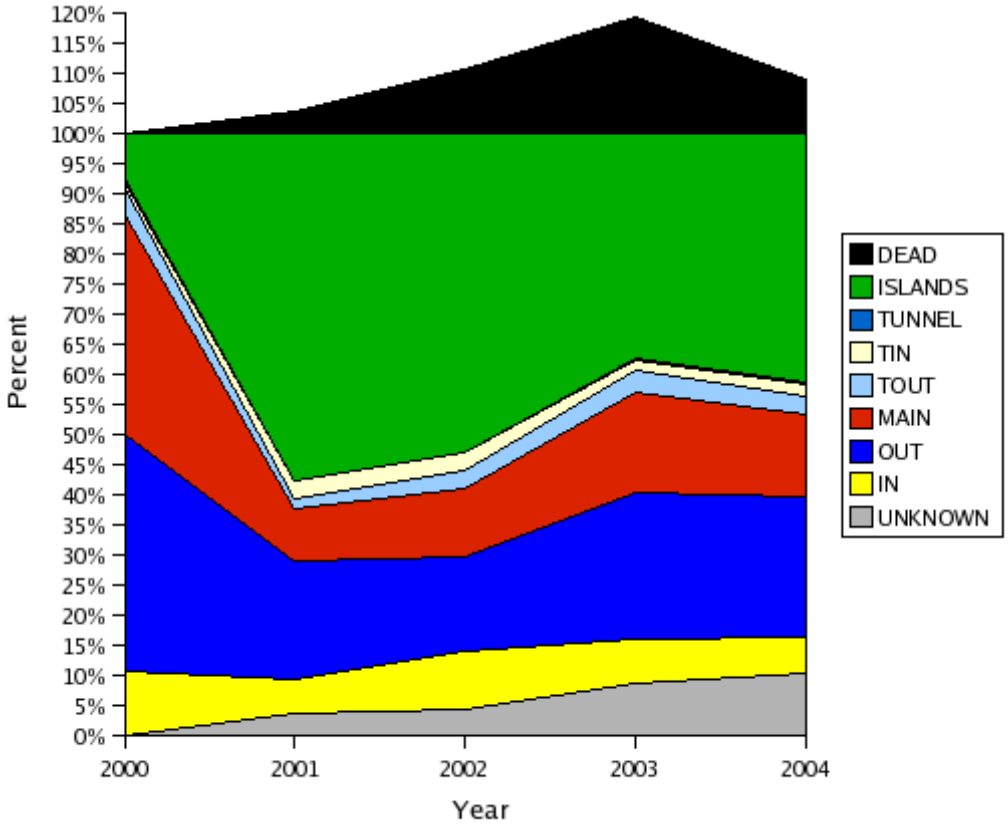


Structure Micro Dynamics

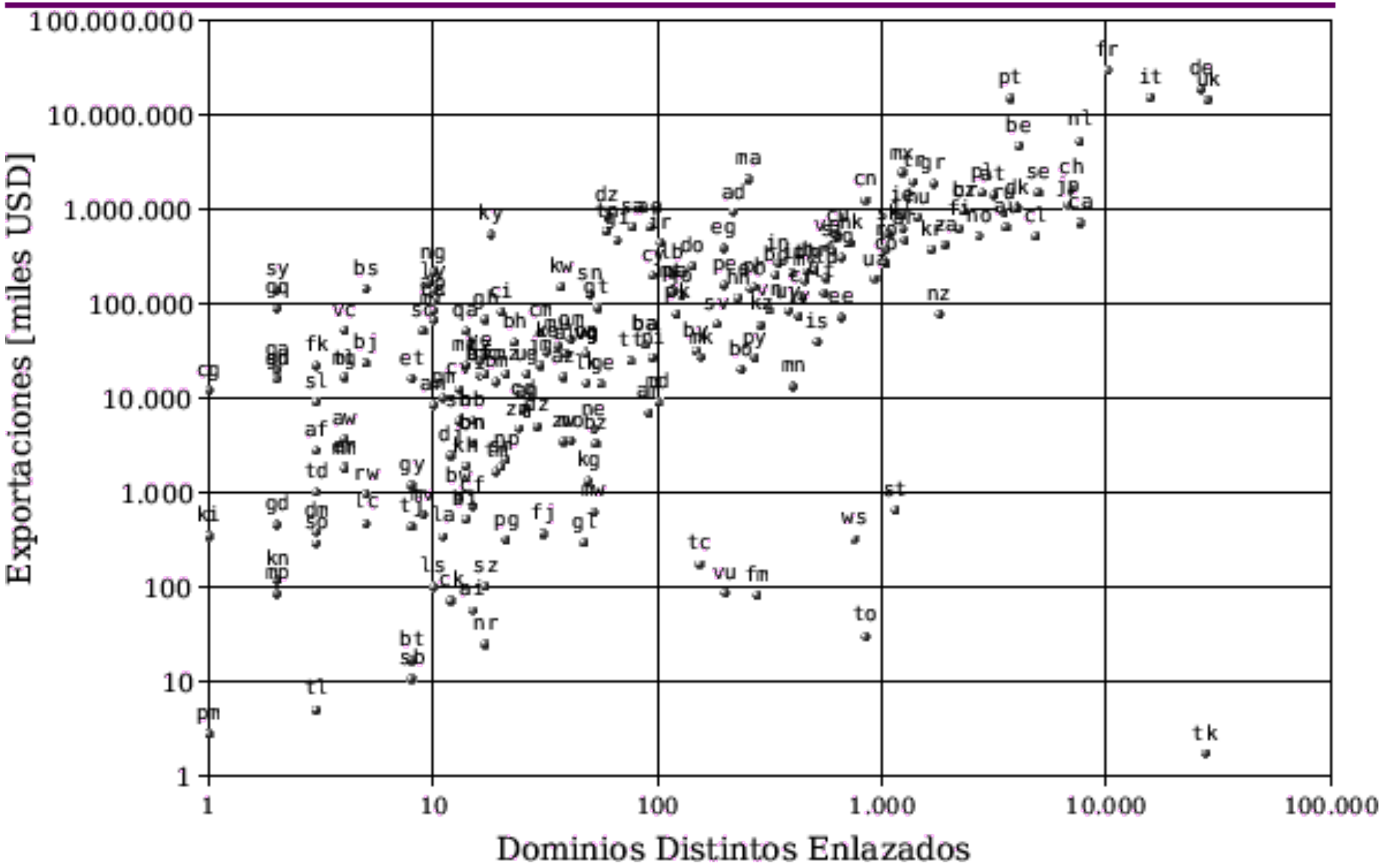


Size Evolution

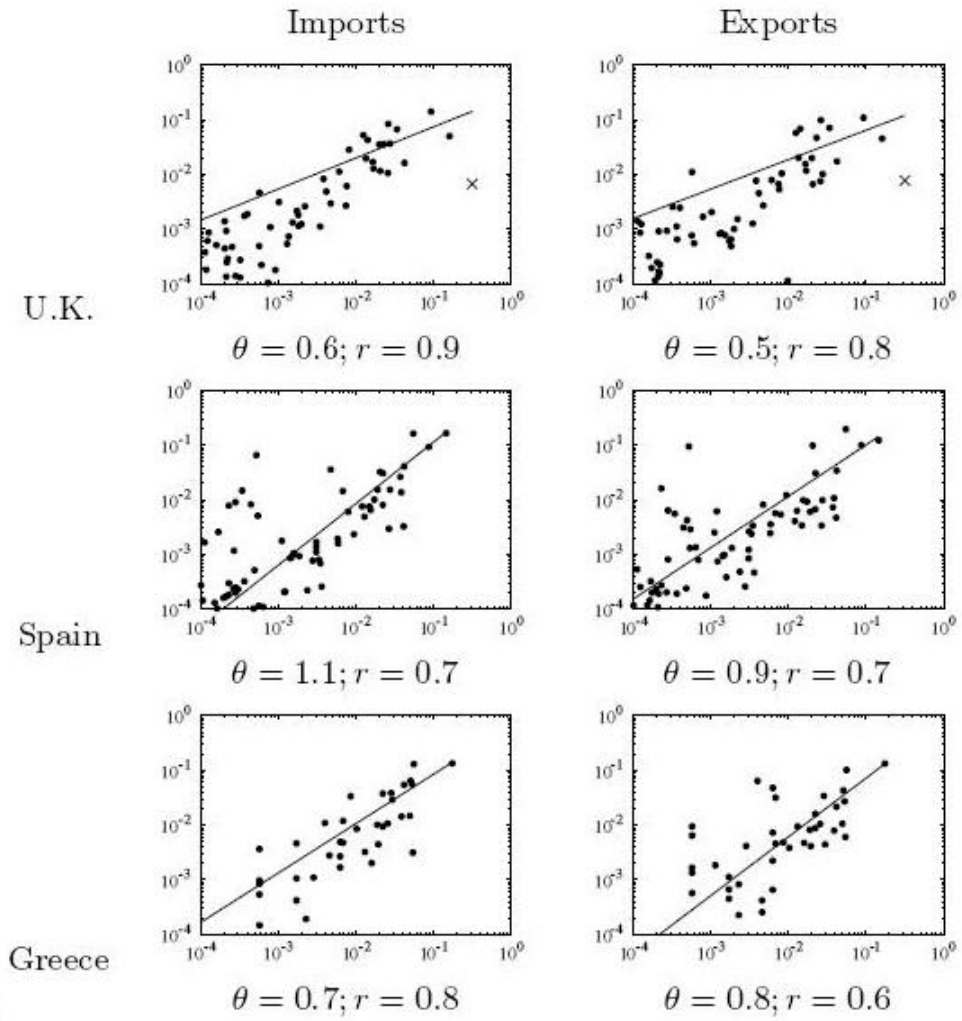


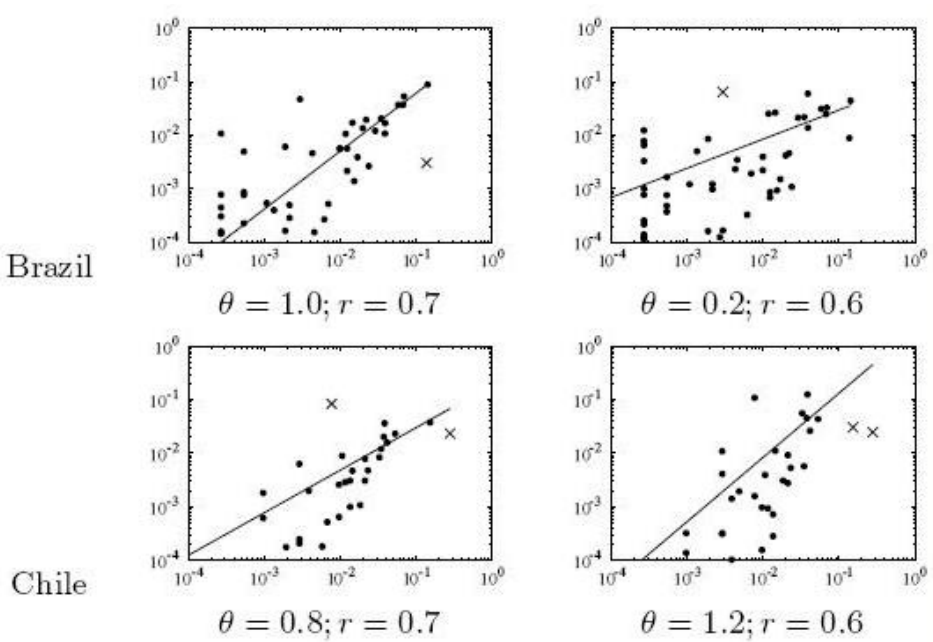


Mirror of the Society



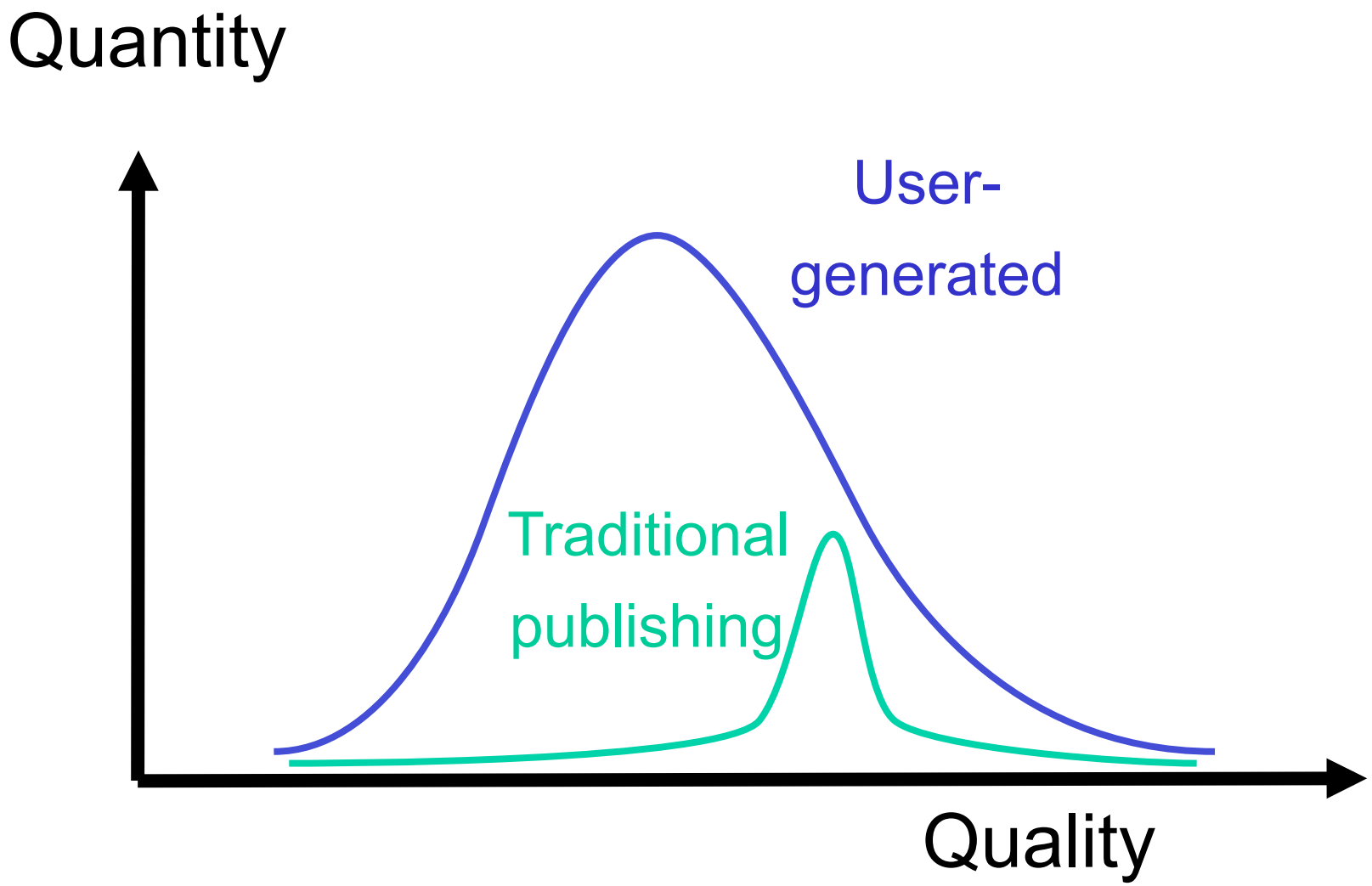
Exports/Imports vs. Domain Links





Baeza-Yates & Castillo, WWW2006

What is in the Web?



Content and Metadata trends

Content type	Amount of content produced per day
Published content	3-4 GB
Professional web content	$\sim$ 2 GB
User generated content	8-10 GB
Private text content	$\sim$ 3 TB (300x more)
Upper bound on typed content	$\sim$700 TB ($\sim$200x more)

Metadata type	Amount of metadata produced per day
Anchortext	100 MB
Tags	40 MB
Pageviews	180 GB
Reviews	Around 10 MB

[Ramakrishnan and Tomkins 2007]

The Wisdom of Crowds

- James Surowiecki, a *New Yorker* columnist, published this book in 2004
  - "Under the right circumstances, groups are remarkably intelligent"
- Importance of diversity, independence and decentralization

Aggregating data

"large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future".

Tags / jaguar / clusters

jaguar SEARCH

(Or, try an advanced search.)





car, **cars**, **auto**, etype, automobile, classic, vintage, autoshow, red, show

See more in this cluster...





zoo, **animal**, **cat**, animals, bigcat, seattle, woodlandparkzoo, sleep, edinburgh, caged

See more in this cluster...





guitar, **fender**

See more in this cluster...



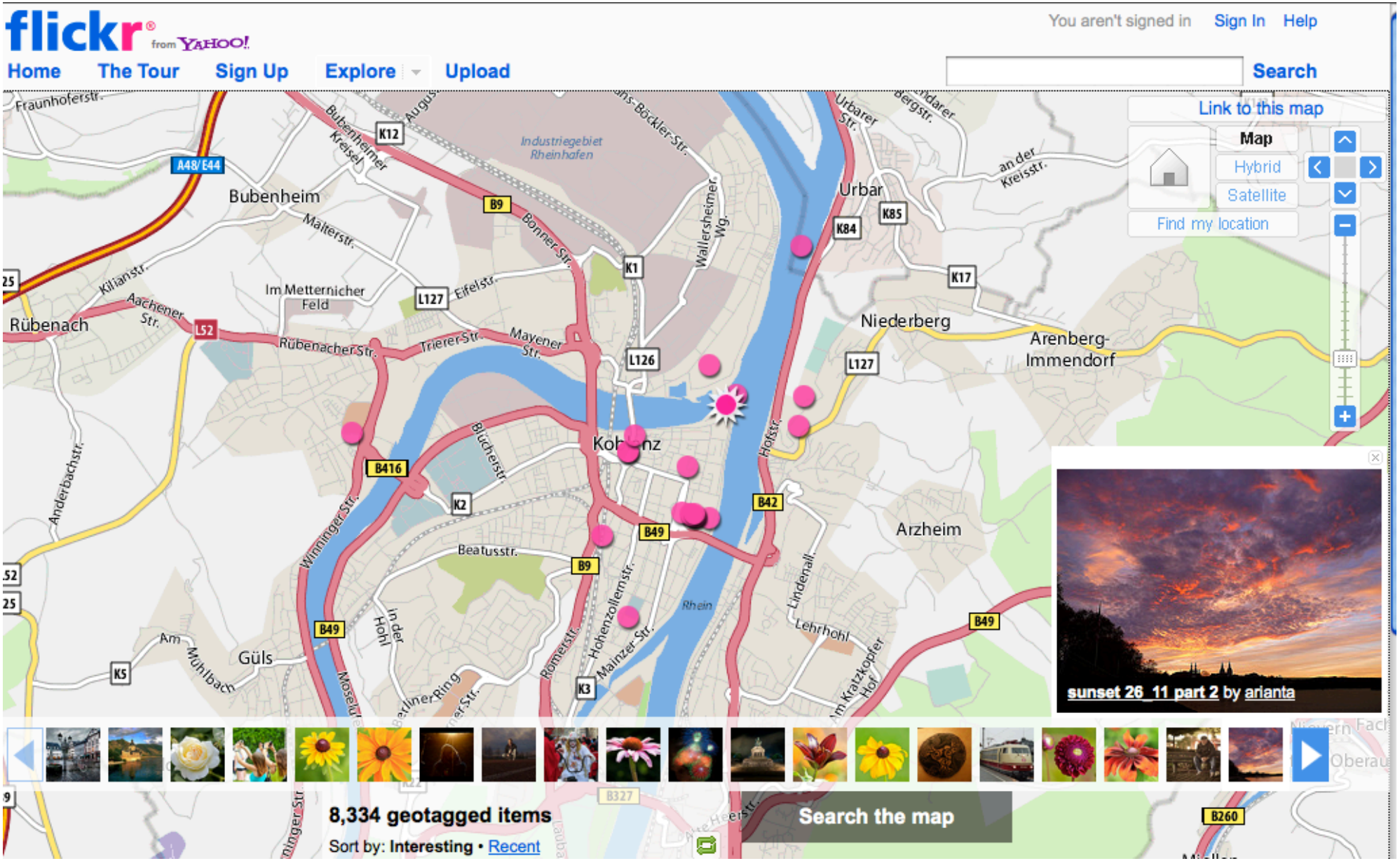


aircraft, **raf**

See more in this cluster...

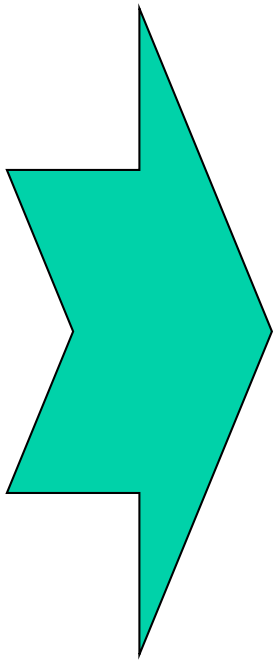
These are the *most recent* photos tagged with **jaguar**. See more...

Flickr: Geo-tagged pictures



The Wisdom of Crowds

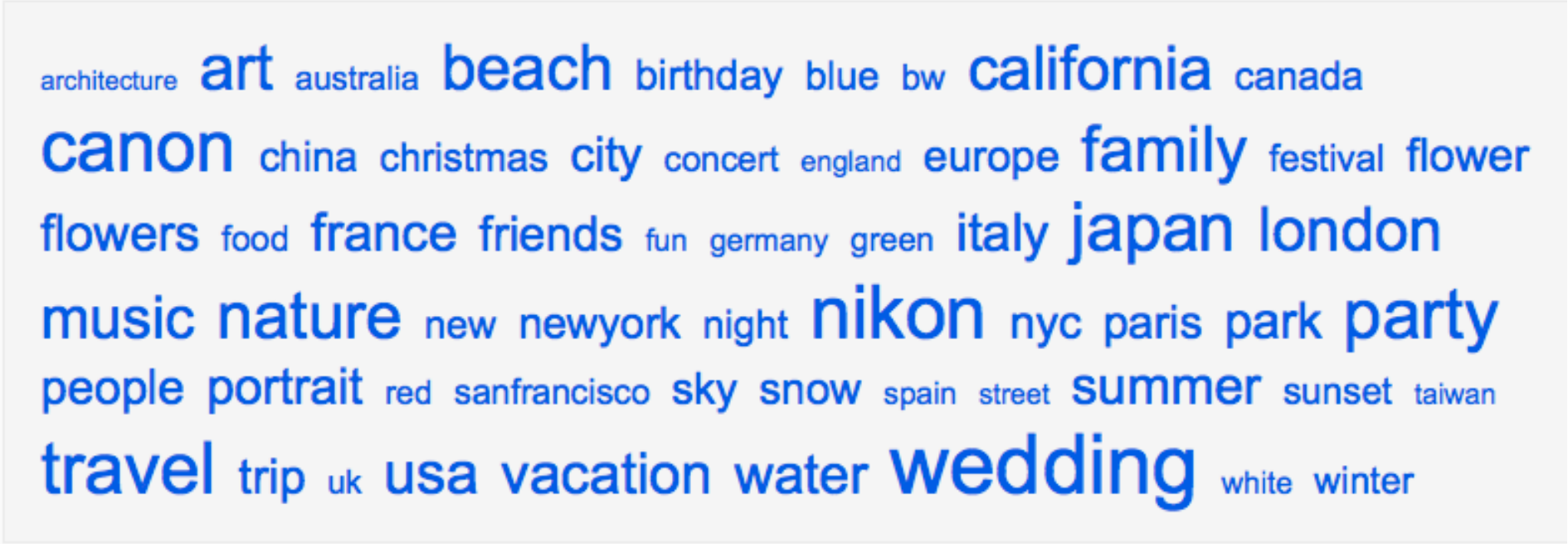
- Popularity
- Diversity
- Quality
- Coverage



Long Tail

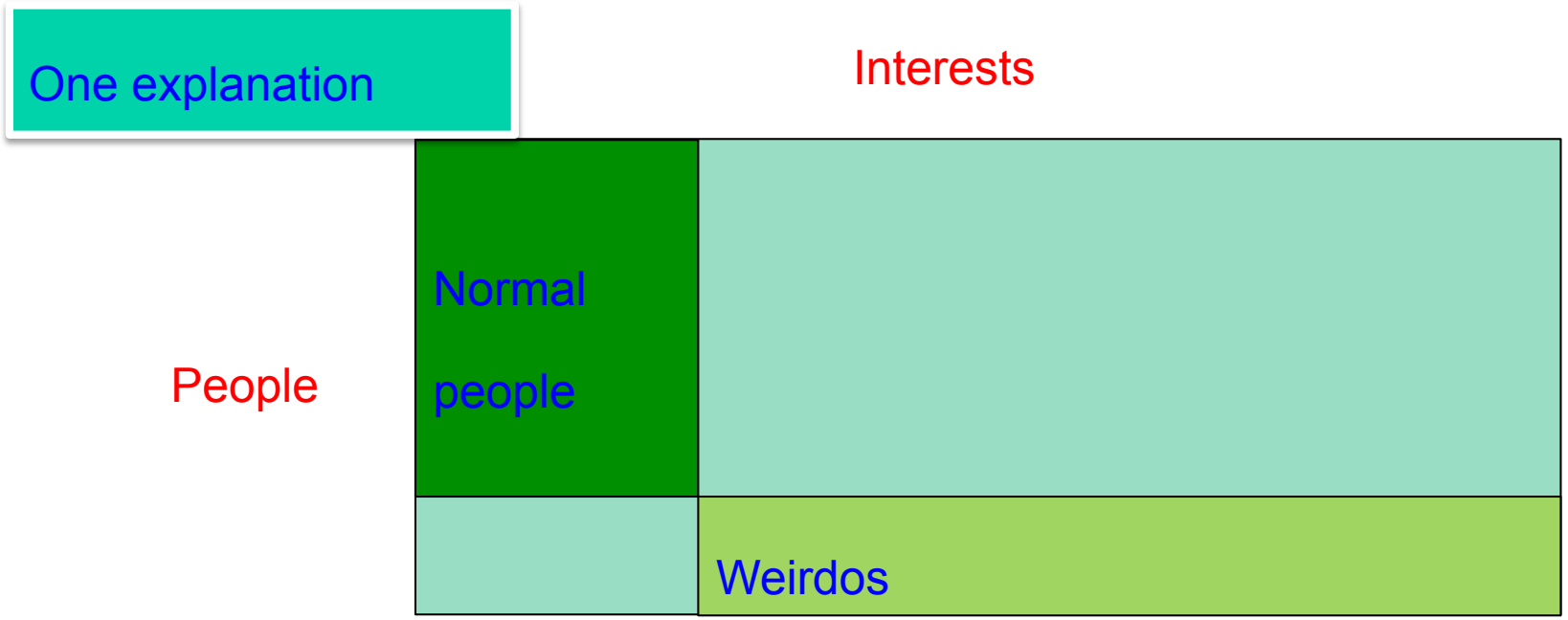
The Long Tail

Explore Flickr through tags



Heavy tail of user interests

- **Many queries, each asked very few times, make up a large fraction of all queries**
  - Movies watched, blogs read, words used …



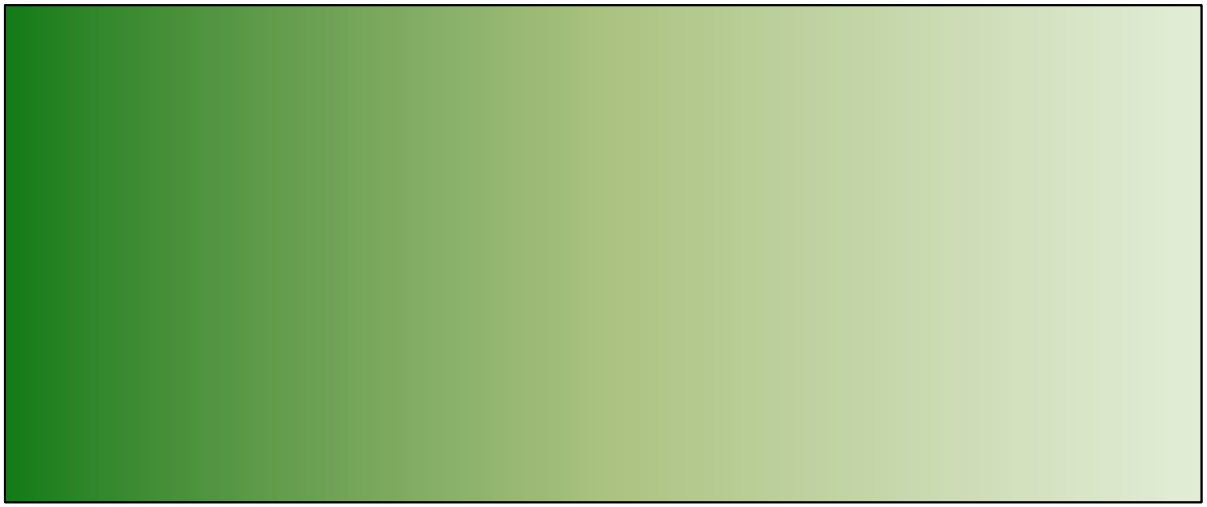
Heavy tail of user interests

- **Many queries, each asked very few times, make up a large fraction of all queries**
- **Applies to word usage, web page access …**
- **We are all partially eclectic**

The reality

Interests

People

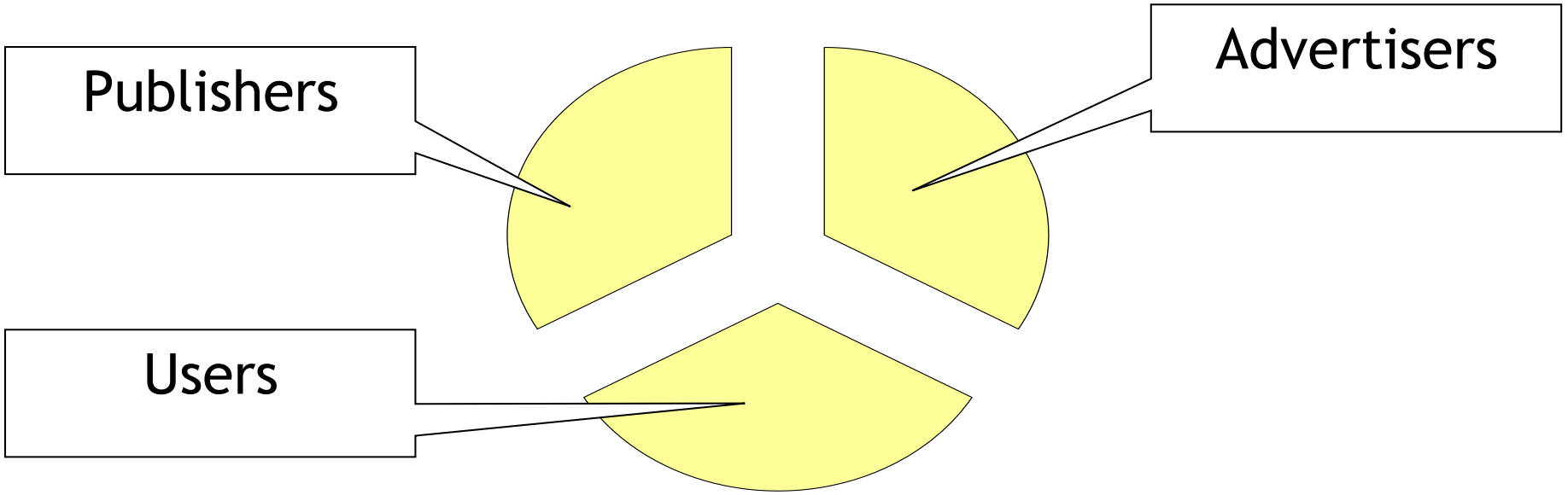


Broder, Gabrilovich, Goel, Pang; WSDM 2009

Why the heavy tail matters

- Not because the worst-sellers make a lot of money
- But because they matter to a lot of people

Advertising



and Spammers! Grrr...

Two Main Types

- Display advertising
  - Contextual match

What is in the Web?

- Information
- Interaction
- Adult content

- Web Spam: On-line casinos + Free movies + Cheap software + Buy a MBA diploma + Prescription - free drugs + V!-4-gra + Get rich now now now!!!

What is in the Web?

Fight Spam

- Adversarial Web Retrieval
- Text Spam (e.g. Cloaking)
- Link Spam (e.g. Link Farms)
- Metadata spam
- Ad spam (e.g. Clicks, Bids)

(2) Web Search

The evolution of commercial web search engines

- **First generation -- use only "on page", text data**
  - Word frequency, language

1994-1997 AV, Excite, Lycos, etc

- **Second generation -- use off-page, web-specific data**
  - Link (or connectivity) analysis
    - Sophisticated mathematical methods
  - Click-through data (What results people click on)
  - Anchor-text (How people refer to this page)

From 1998. Made popular by Google but everyone now

- **Third generation -- answer "the need behind the query"**

  -- Focus on user need, rather than on query

Still evolving

  - Semantic analysis -- what is this about?
  - Integrates multiple sources of data
  - Context determination
    - spatial (user location/target location), query stream (previous queries), personal (user profile), etc
  - Help the user
    - UI, spell checking, query refinement, query suggestion, syntax driven feedback, context help, context transfer, etc
  - Integration of search and text analysis

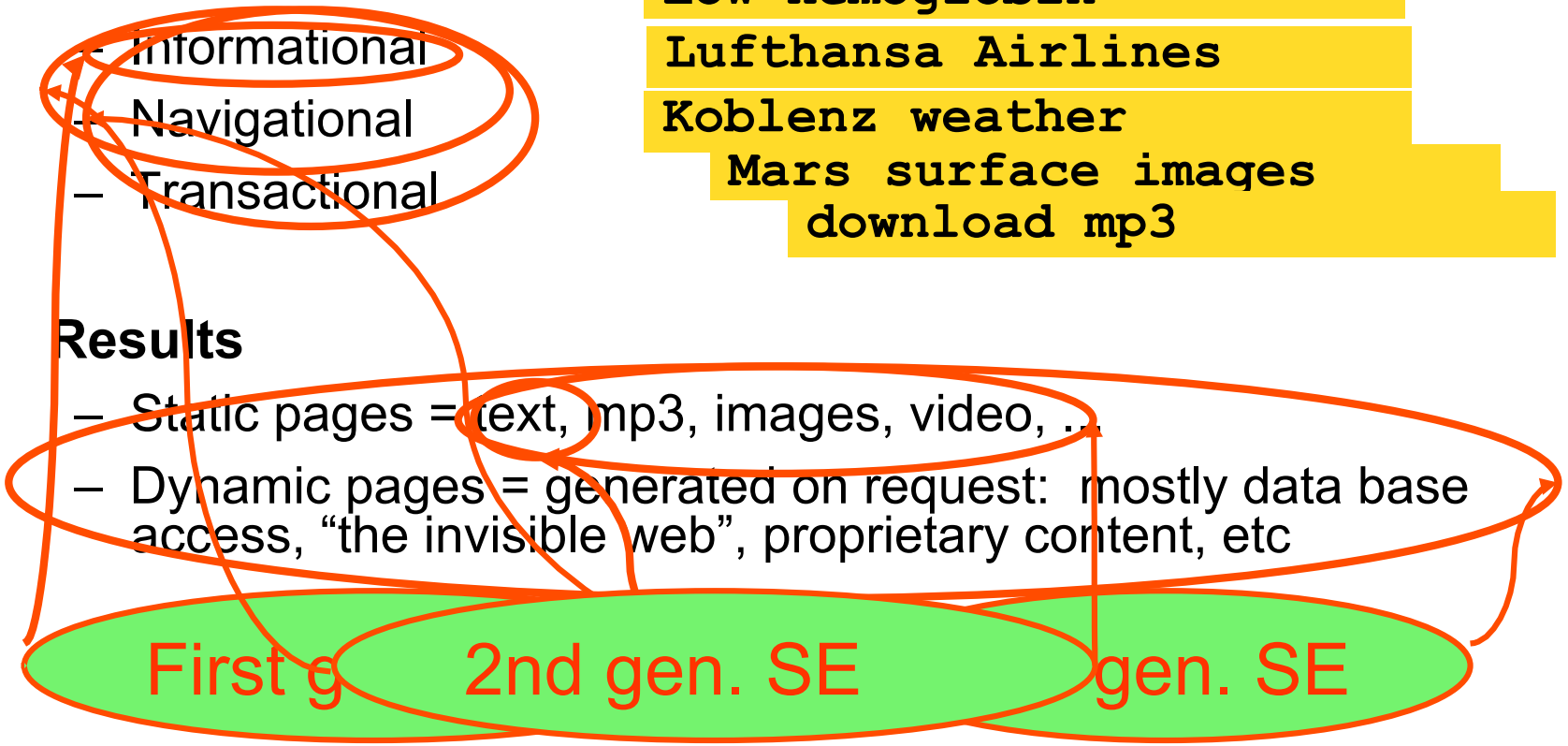
Web Search today

- **Corpus: The publicly accessible Web**
- **Goal: Retrieve high quality results that are relevant to user's need**

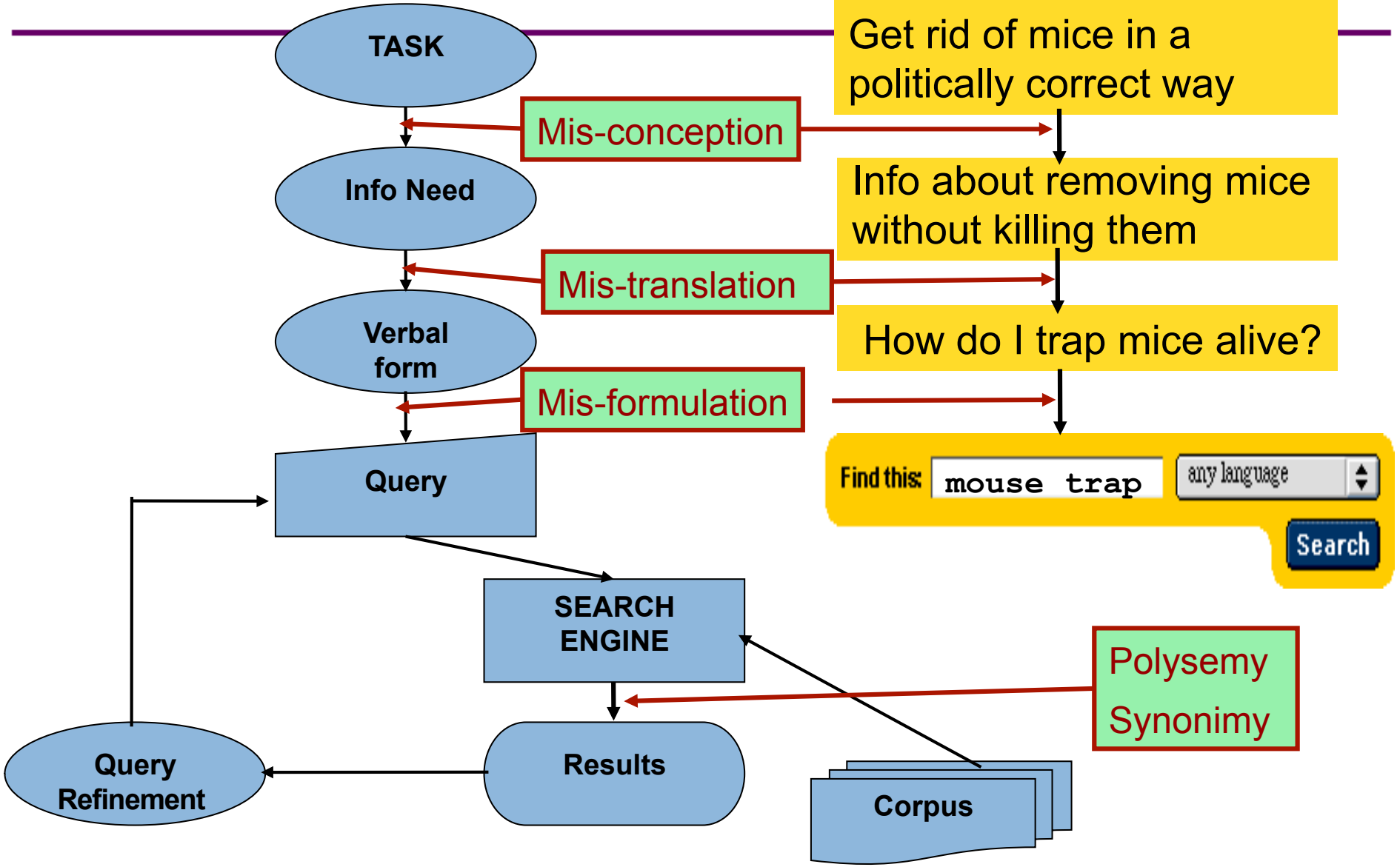
- **Need**
  - Informational
  - Navigational
  - Transactional

Low hemoglobin
Lufthansa Airlines
Koblenz weather
Mars surface images
download mp3

- **Results**
  - Static pages = text, mp3, images, video, …
  - Dynamic pages = generated on request: mostly data base access, "the invisible web", proprietary content, etc

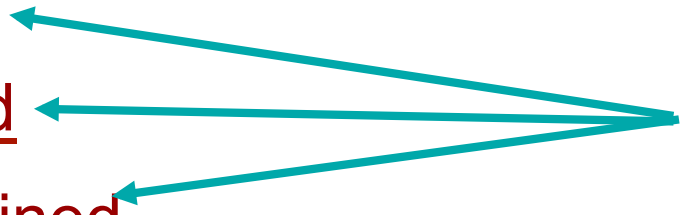


The classic search model

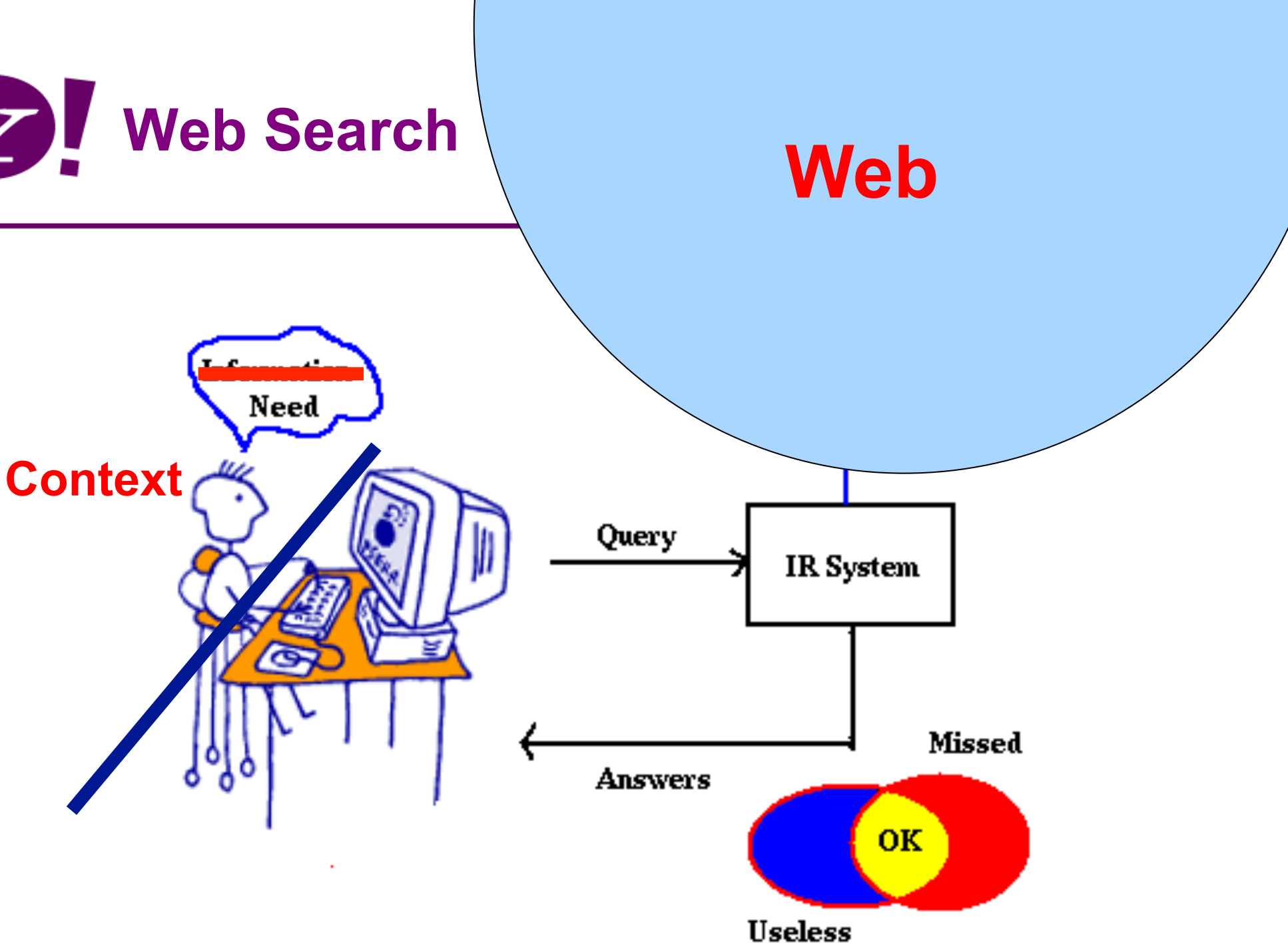


Classic IR Goal

- **Classic relevance**
  - For each query Q and stored document D in a given corpus assume there exists relevance Score(Q, D)
    - Score is the average over users U and contexts C
  - Optimize Score(Q, D) as opposed to Score(Q, D, U, C)
  - That is, usually:
    - Context ignored
    - Individuals ignored
    - Corpus predetermined



Web Search



Web Search

- **This is one of the most complex data engineering challenges today:**
  - Distributed in nature
  - Large volume of data
  - Highly concurrent service
  - Users expect very good & fast answers
- **Solution: Replicated centralized system**

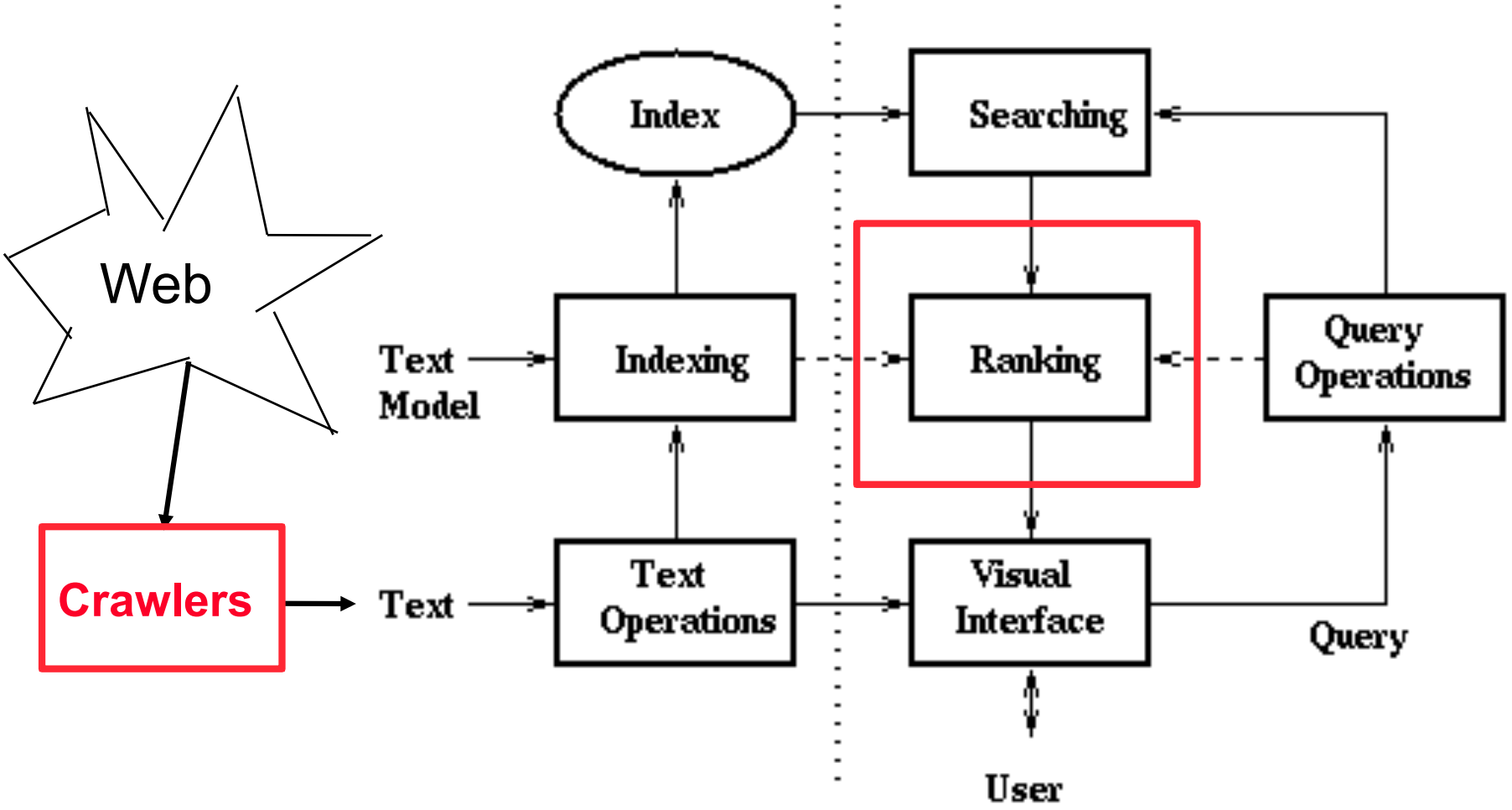
Main techniques

- Centralized Software Architecture
- Hypertext Structure
  - Allows to include link ranking
- On-line Quality Evaluation
- Distributed Data
  - Crawling
- Locally Distributed Index
  - Parallel Indexing
  - Parallel Query Processing
- Business Model based in Advertising
  - E.g. Word based and pay-per-click

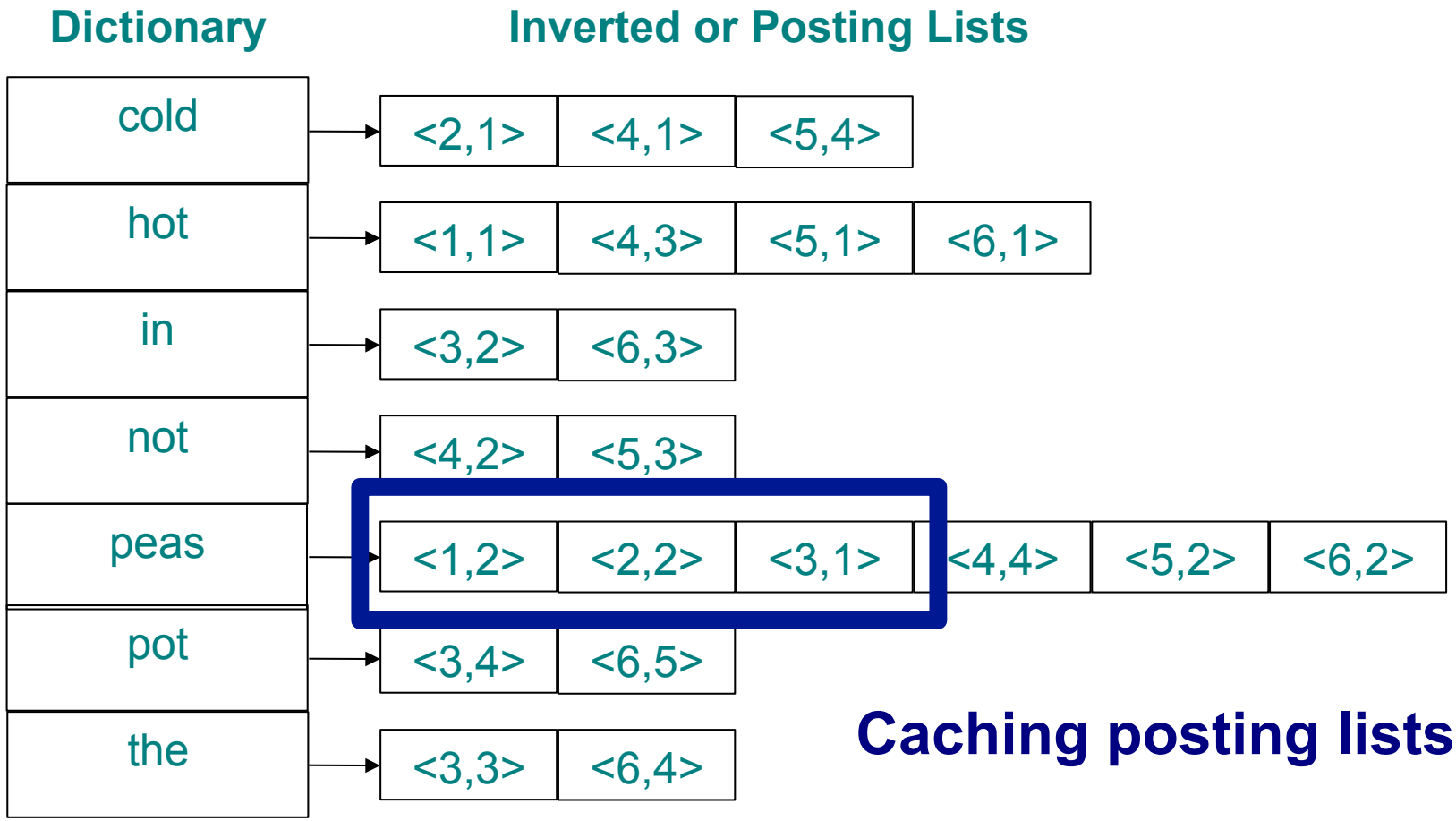
Main challenges

- Problems:
  - volume
  - fast rate of change and growth
  - dynamic content
  - redundancy
  - organization and data quality
  - diversity
  - …..
- Deal with data overload

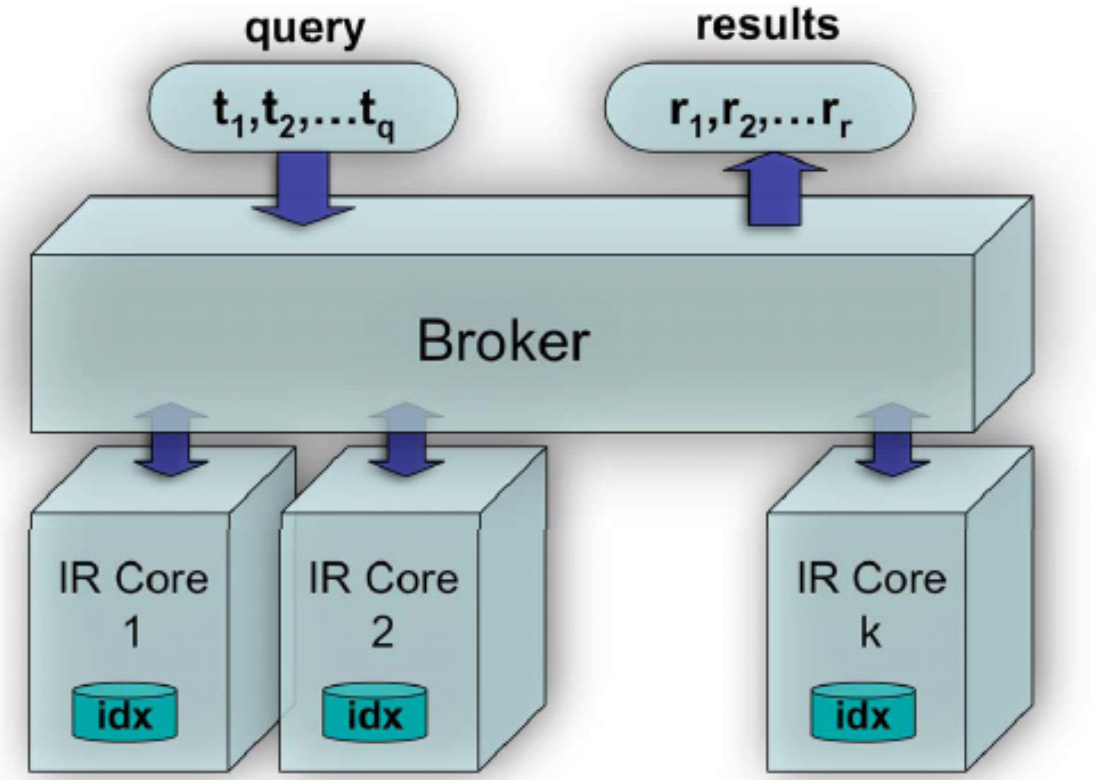
WR Logical Architecture



Inverted Index

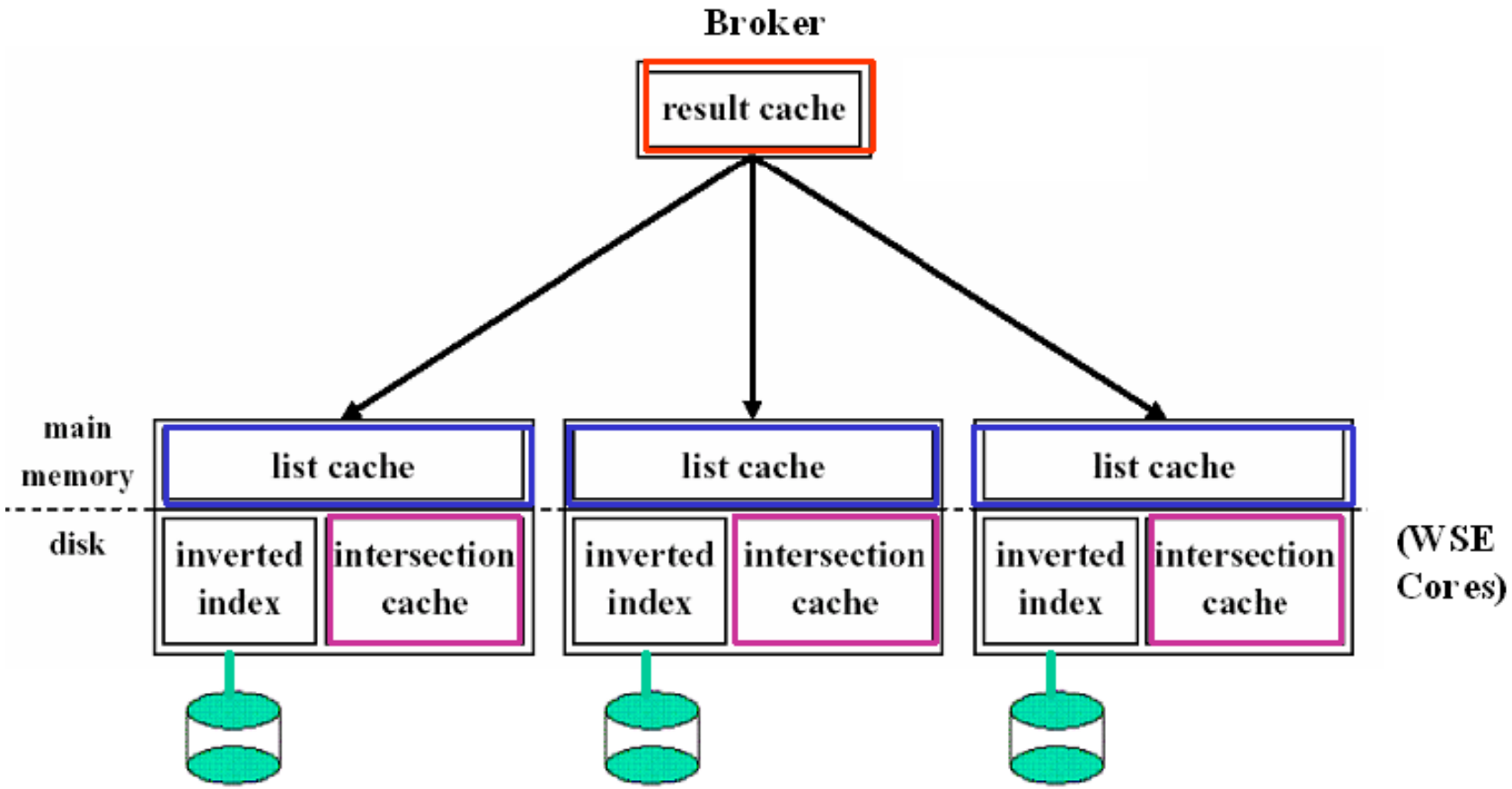


A key goal is optimizing throughput but making sure that query response time is below a given upper bound

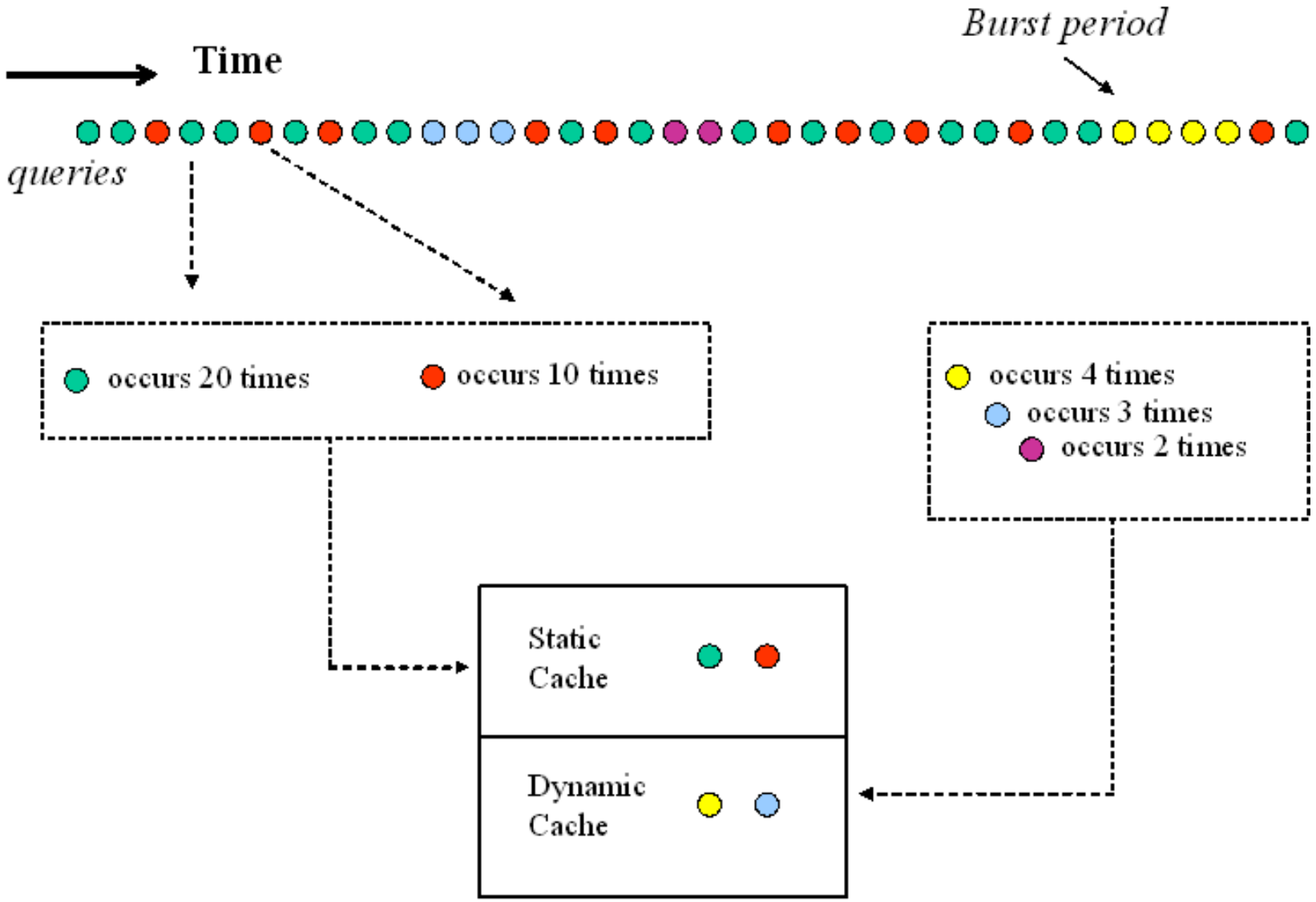


Throughput → number of queries per unit of time

Cache hierarchies are used to improve throughput

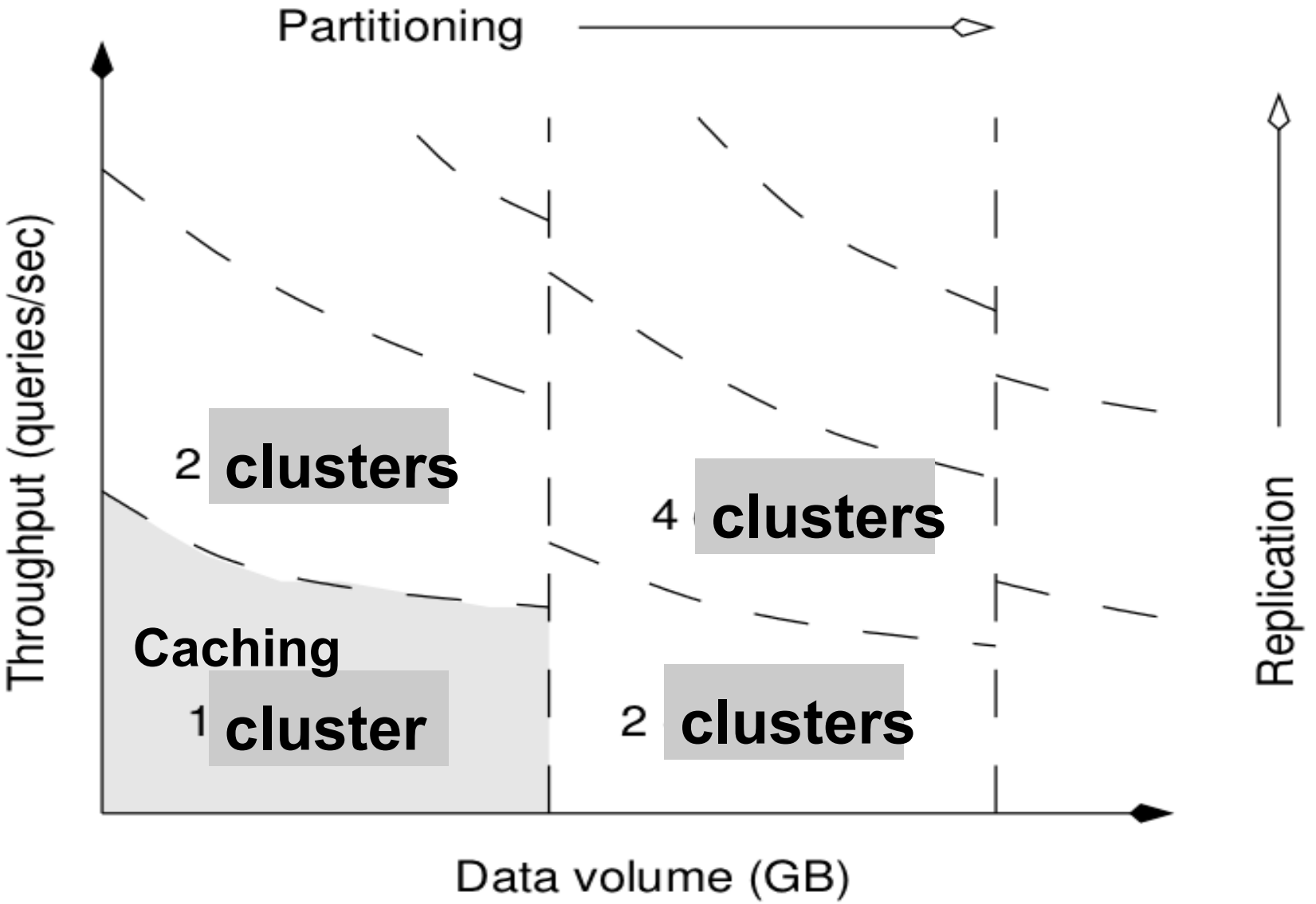


Cache policies are tailored to user behavior



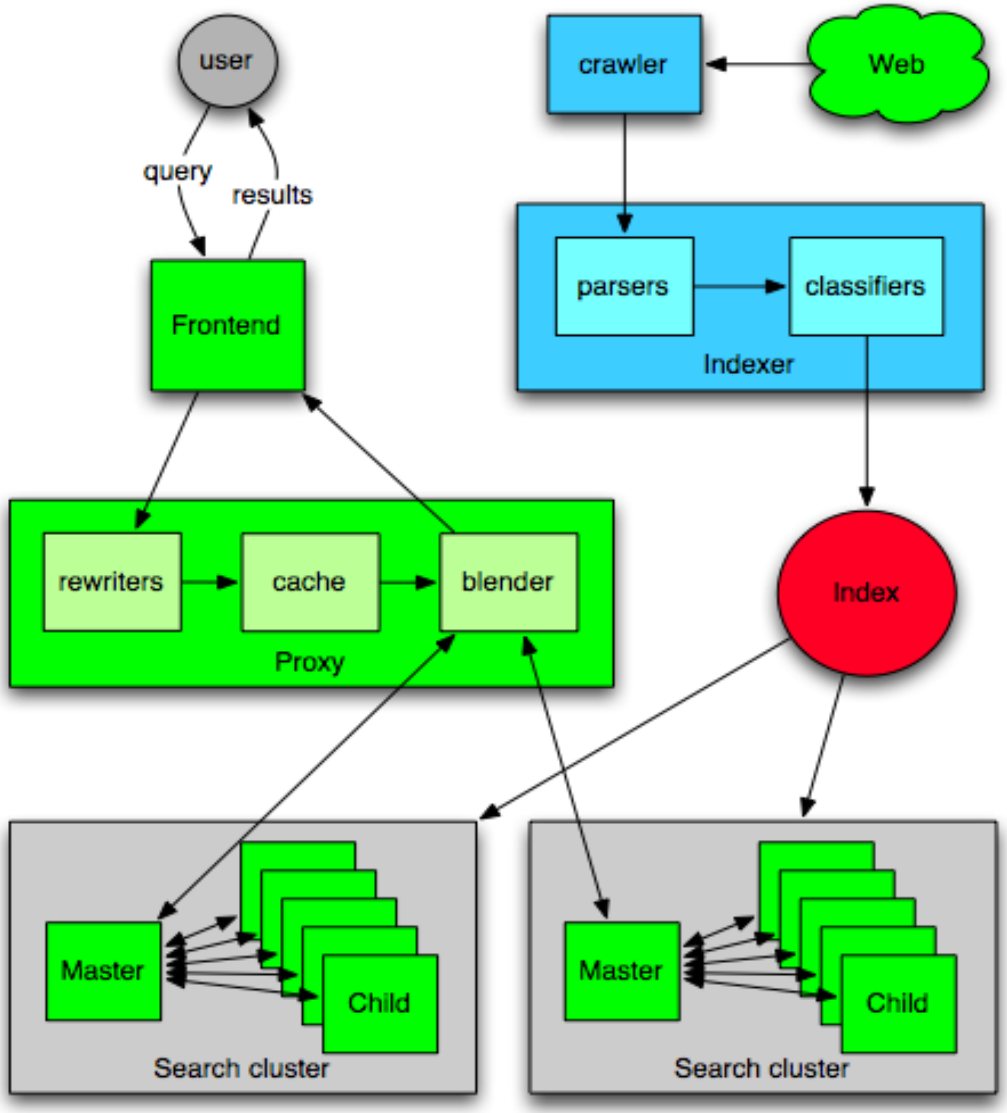
Scaling Up

Adapted from Moffat and Zobel, 2004.

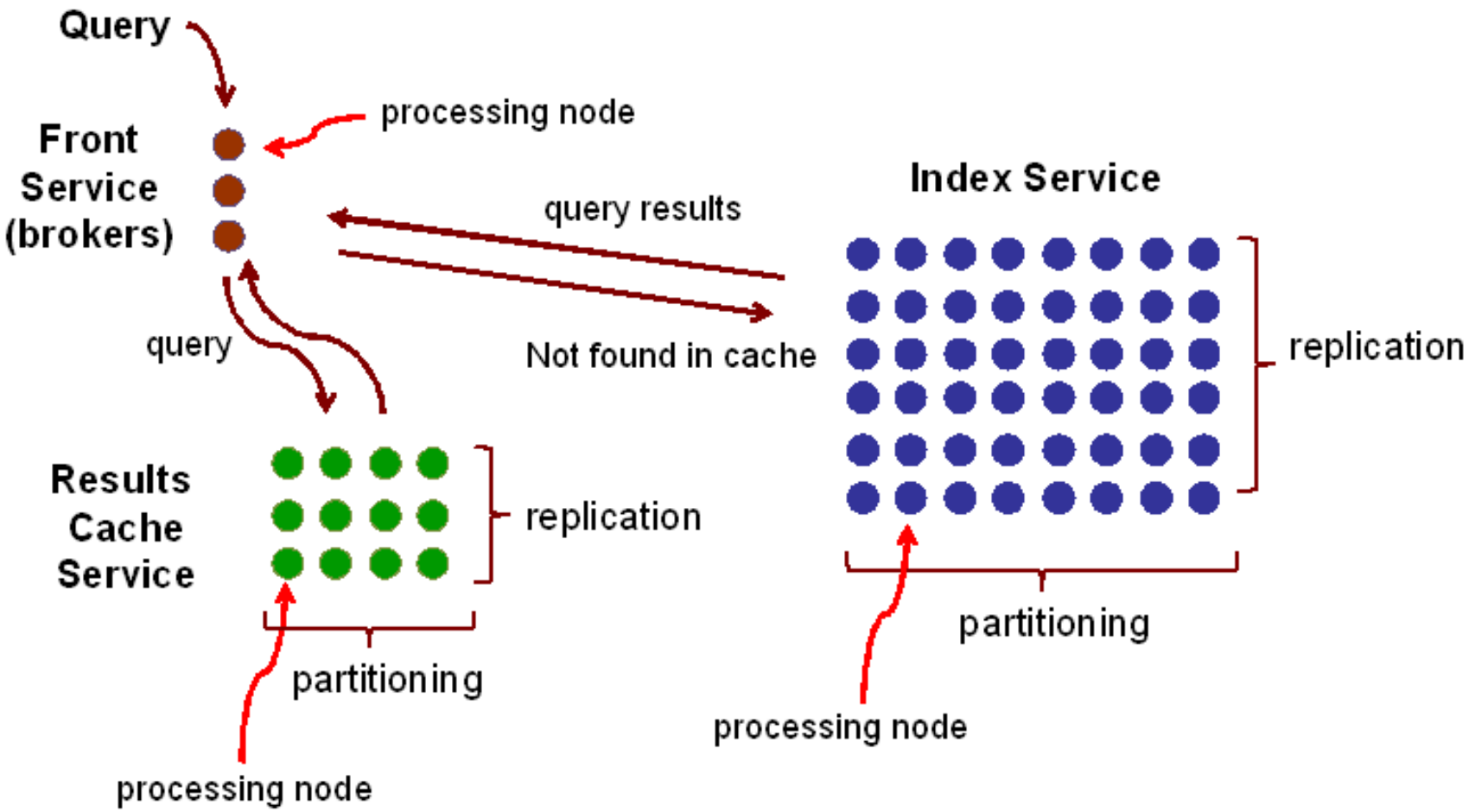


A Typical Web Search Engine

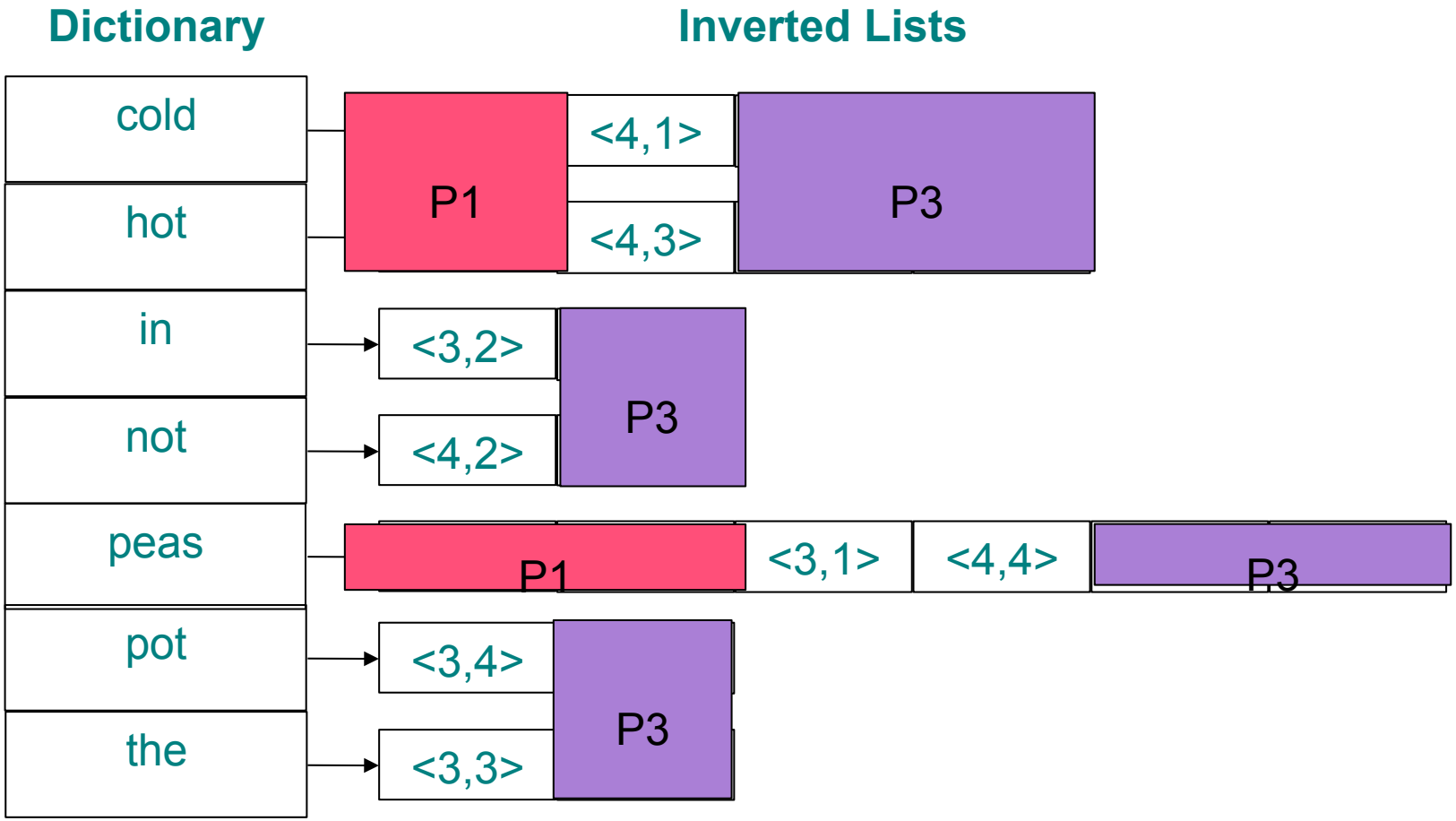
- Caching
  - result cache
  - posting list cache
  - document cache
- Replication
  - multiple clusters
  - improve throughput
- Parallel query processing
  - partitioned index
    - document-based
    - term-based
  - Online query processing



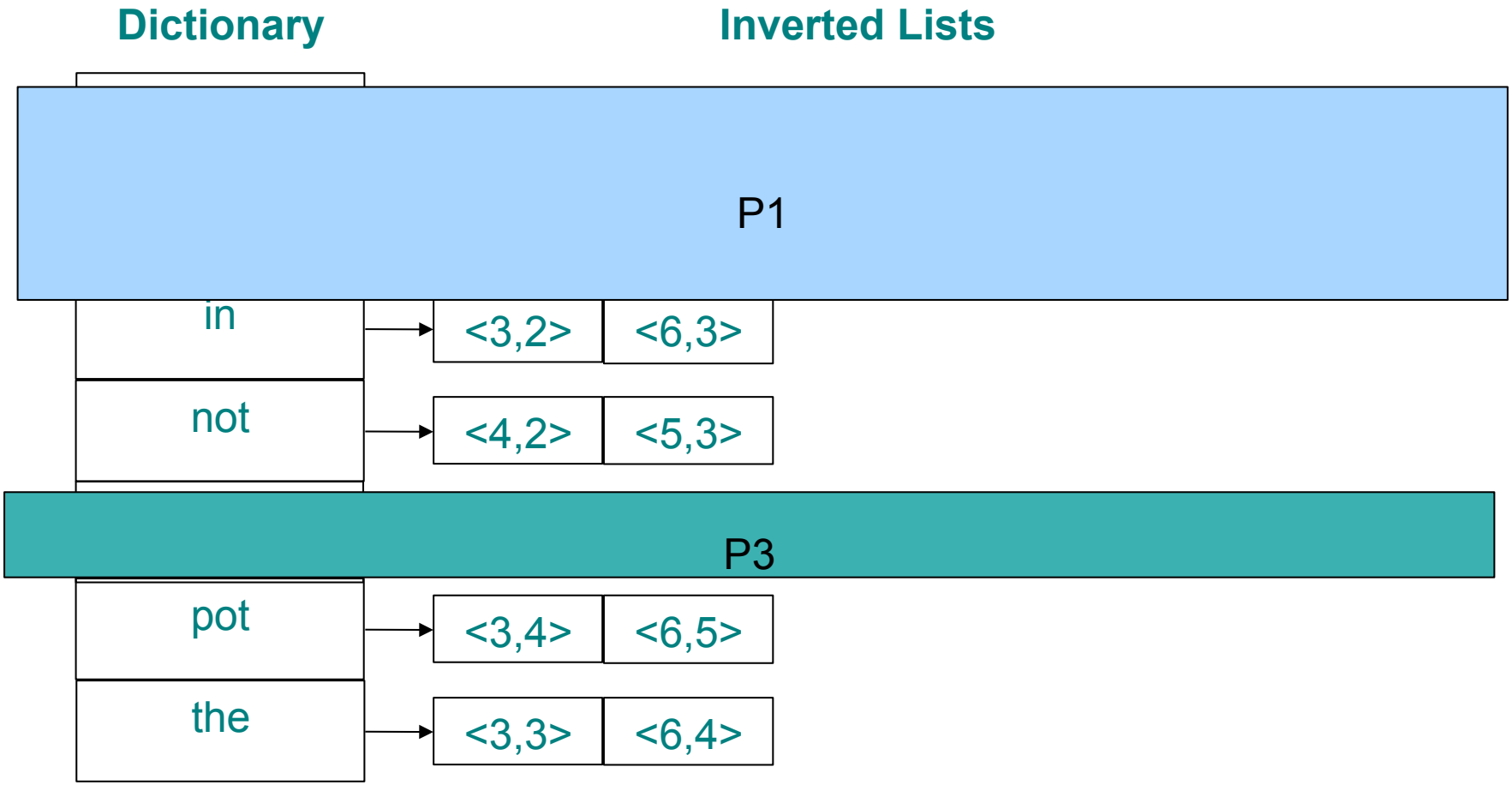
The engine is divided into a collection of services Each service is deployed in a set of processing nodes



Document Partitioning



Term Partitioning

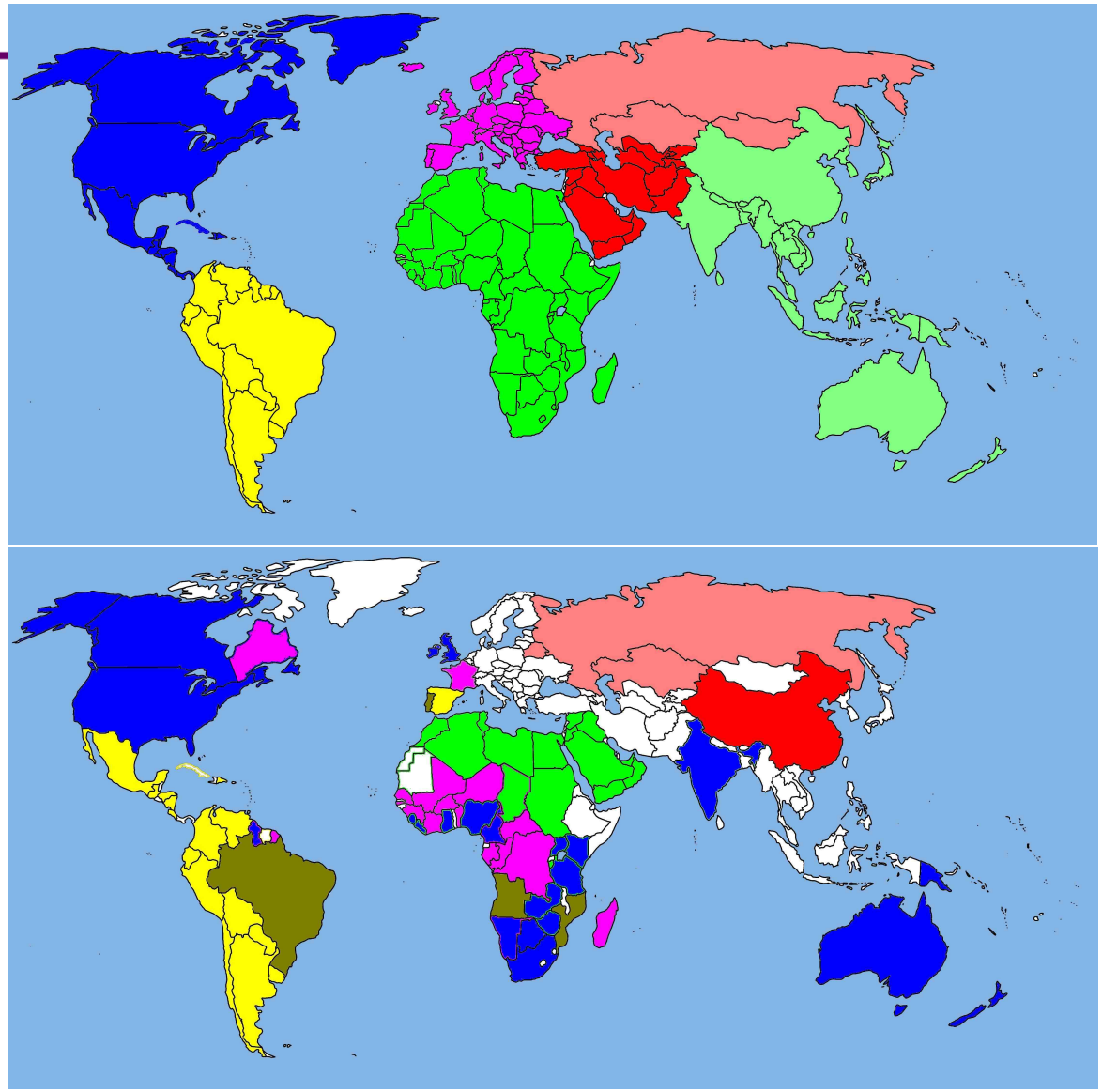


Index Partitioning: Comparison

• By documents	By terms
• Easy to partition	Random partition
• *Easier to build*	Hard to build
• No concurrency	Concurrent
• Perfect balance	Less balanced
• Less variance	Higher variance
• *Easier to maintain*	Harder to maintain

Index Partitioning: Practice

- **Within a cluster**
  - **term-based**
    - **performance**
  - **document-based**
    - **fault tolerance**
    - **load balance**
- **Across data centers**
  - **geographical**
  - **language-based**



Algorithmic Challenges

- Crawling:
  - Quantity
  - Freshness
  - Quality
  - Politeness vs. Usage of Resources

Conflict

Adversarial IR

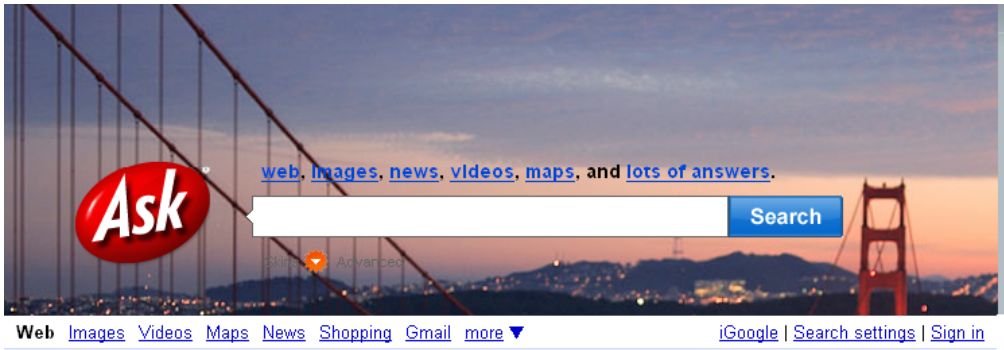
- Ranking
  - Words, links, usage logs, … , metadata
  - Spamming of all kinds of data
  - Good precision, unknown recall

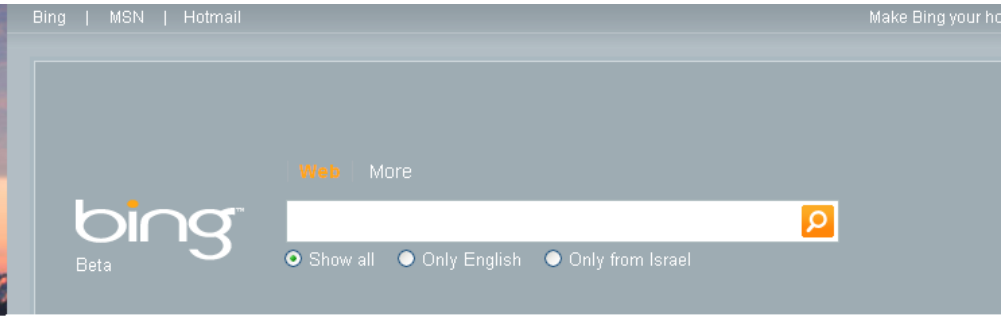
Search rectangle

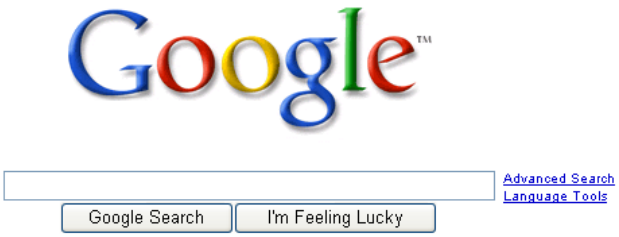
Very little differences between major search engines
A rectangle – text box for your queries

Other forms of rectangles?

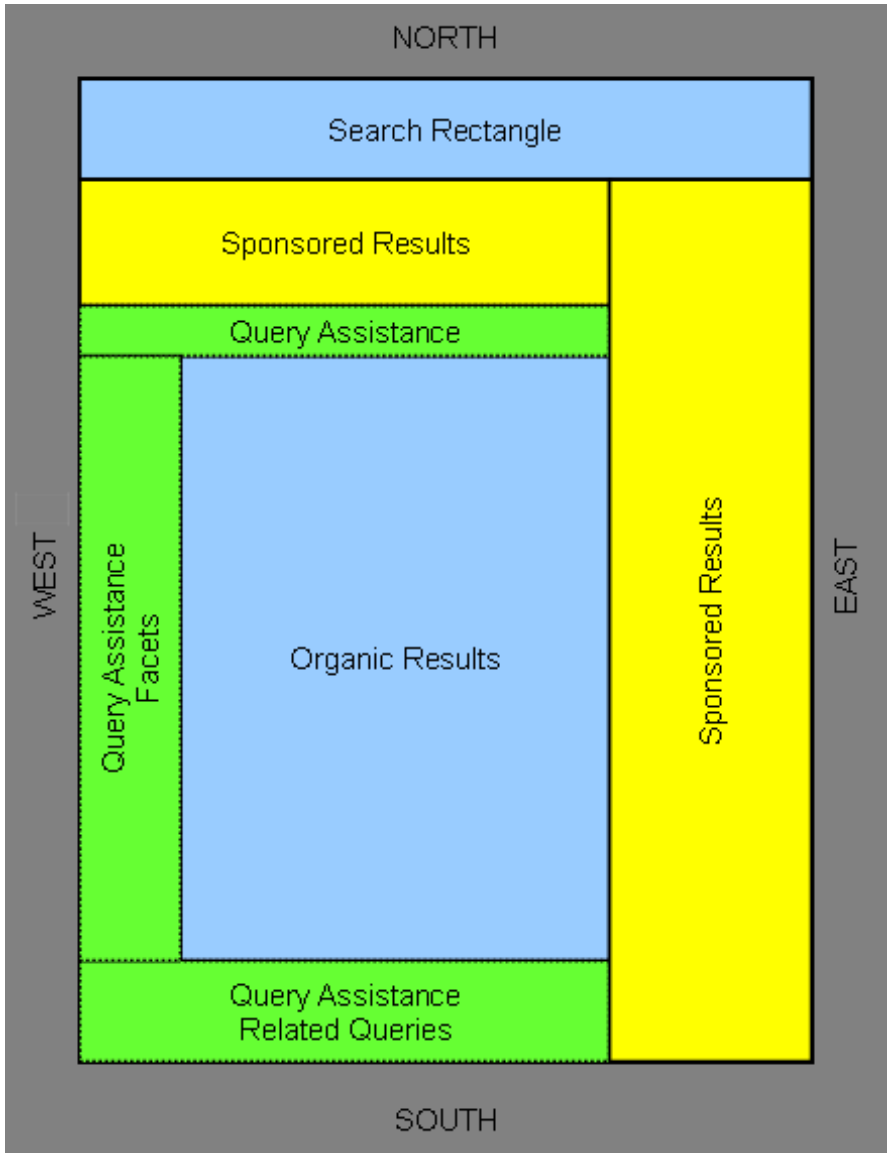
- **Embedded in a portal**
- **Always here in a toolbar**
- **Ultimate rectangle: omnibox**







SERP – basic layout



Web Search Queries

- **Cultural and educational diversity**
- **Short queries & impatient interaction**
  - few queries posed & few answers seen
- **Smaller & different vocabulary**
- **Different user goals [Broder, 2000]:**
  - Information need
  - Navigational need
  - Transactional need
- **Refined by Rose & Levinson, WWW 2004**

User Needs

- **Basic Taxonomy (Broder 2002)**
  - **Informational** – want to learn about something (~40% / 65%)

    Low hemoglobin
  - **Navigational** – want to go to that page (~25% / 15%)

    Lufthansa
  - **Transactional** – want to do something (web-mediated) (~35% / 20%)
    - Access a service

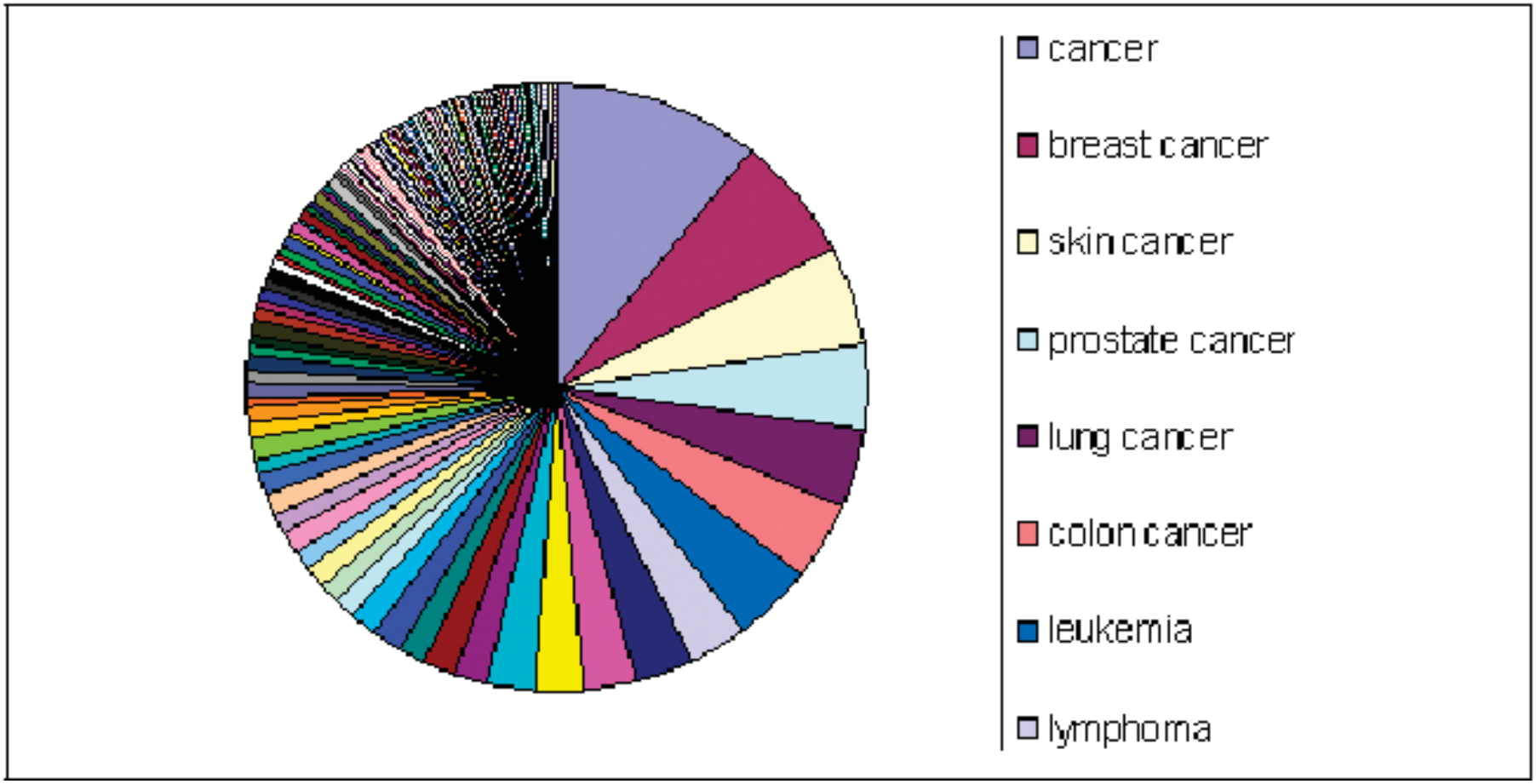
      Koblenz weather
    - Downloads

      Mars surface images
    - Shop

      Digital camera
  - Gray areas
    - Find a good hub

      Car rental Frankfurt
    - Exploratory search "see what's there"

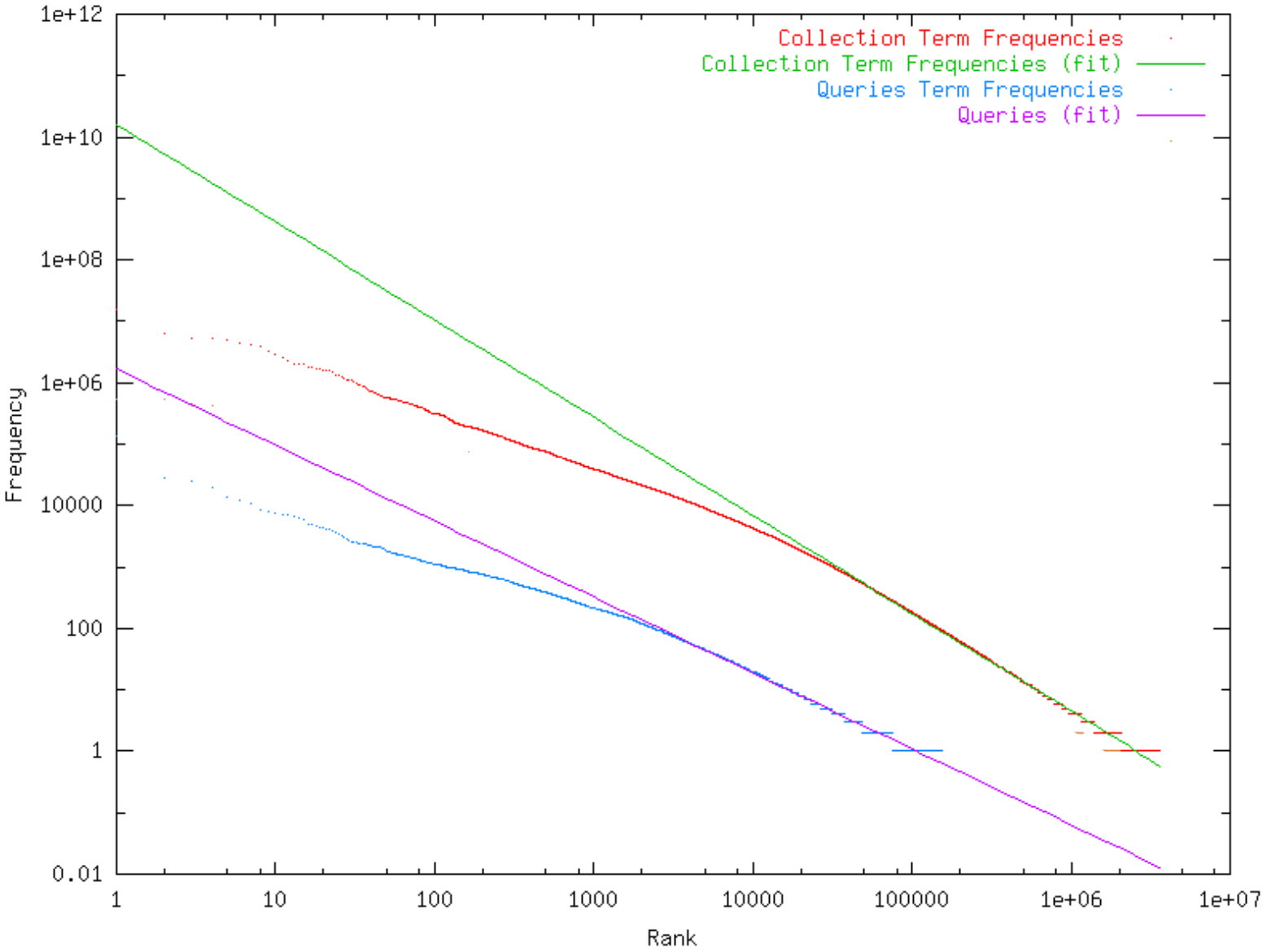
Query Distribution



Power law: few popular broad queries, many rare specific queries

Queries and text

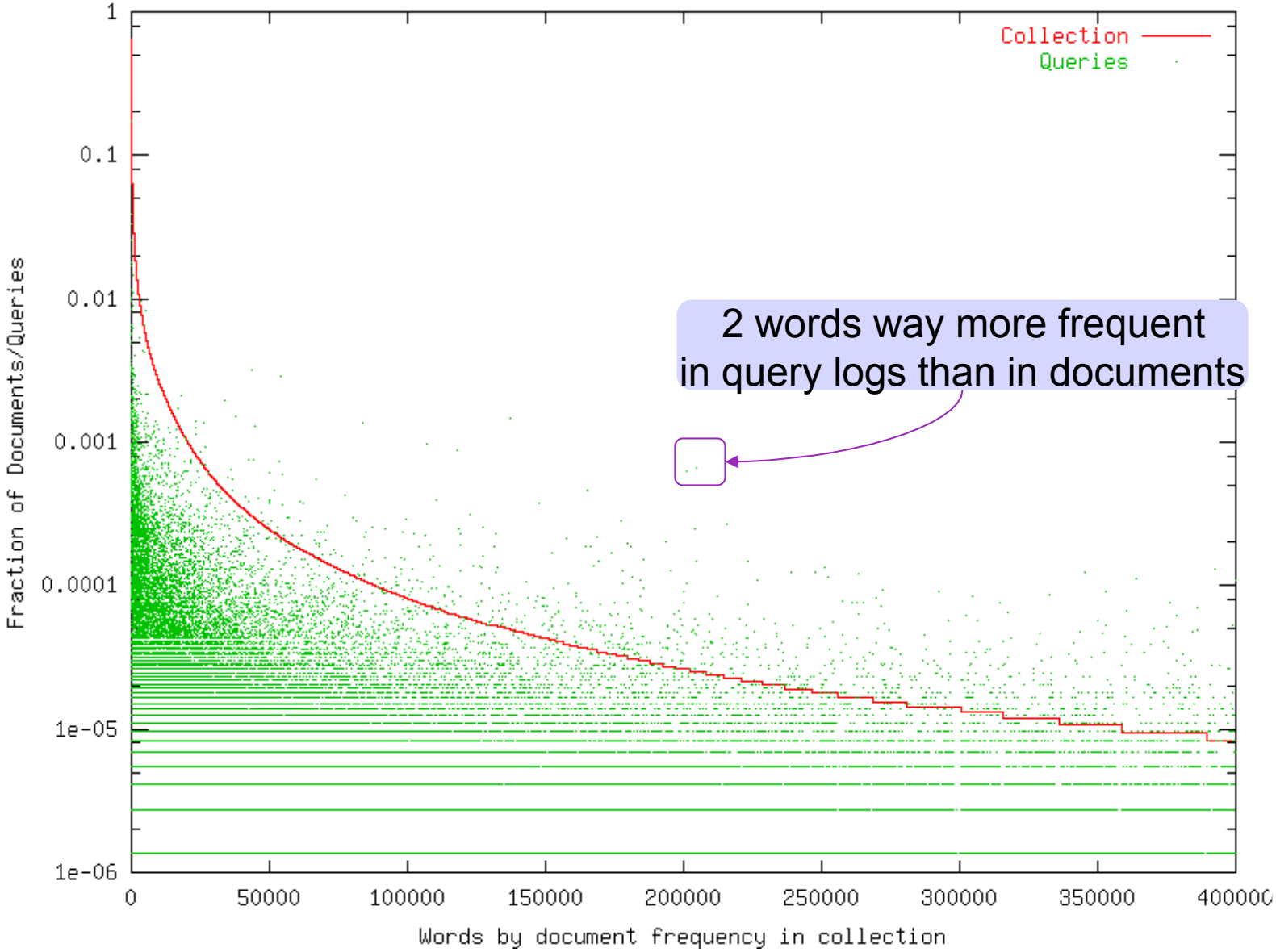
Word distribution in queries and in documents are different



[Baeza-Yates & Saint-Jean, SPIRE'2003] *"A Three Level Search Engine Index based in Query Log Distribution"*

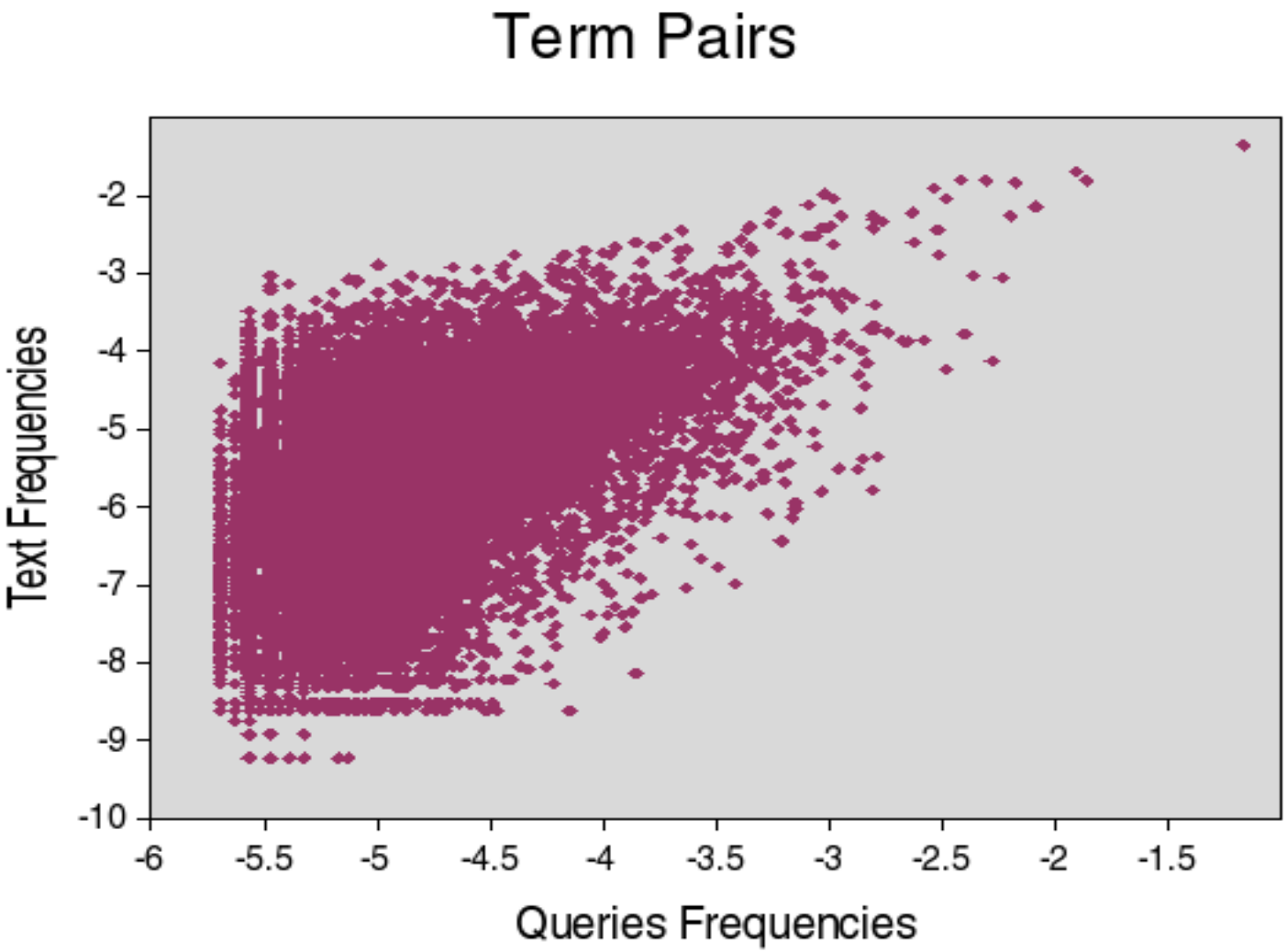
Queries and text

Word distribution in queries in documents now sorted such that same word on x

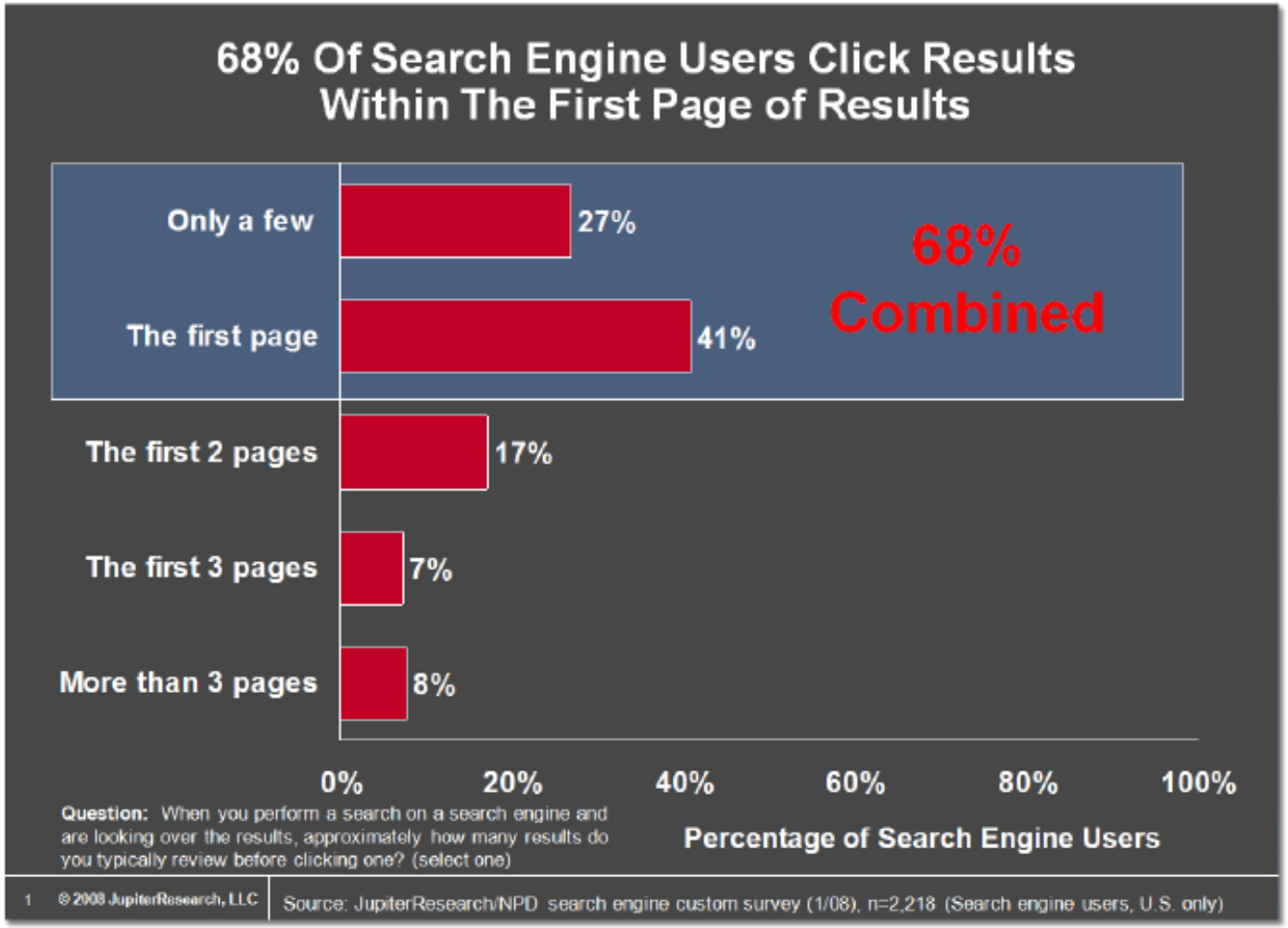


Queries and text

Log-log plot – weak correlation bw query & document vocabularies



Other implicit signals
How far do people look for results?



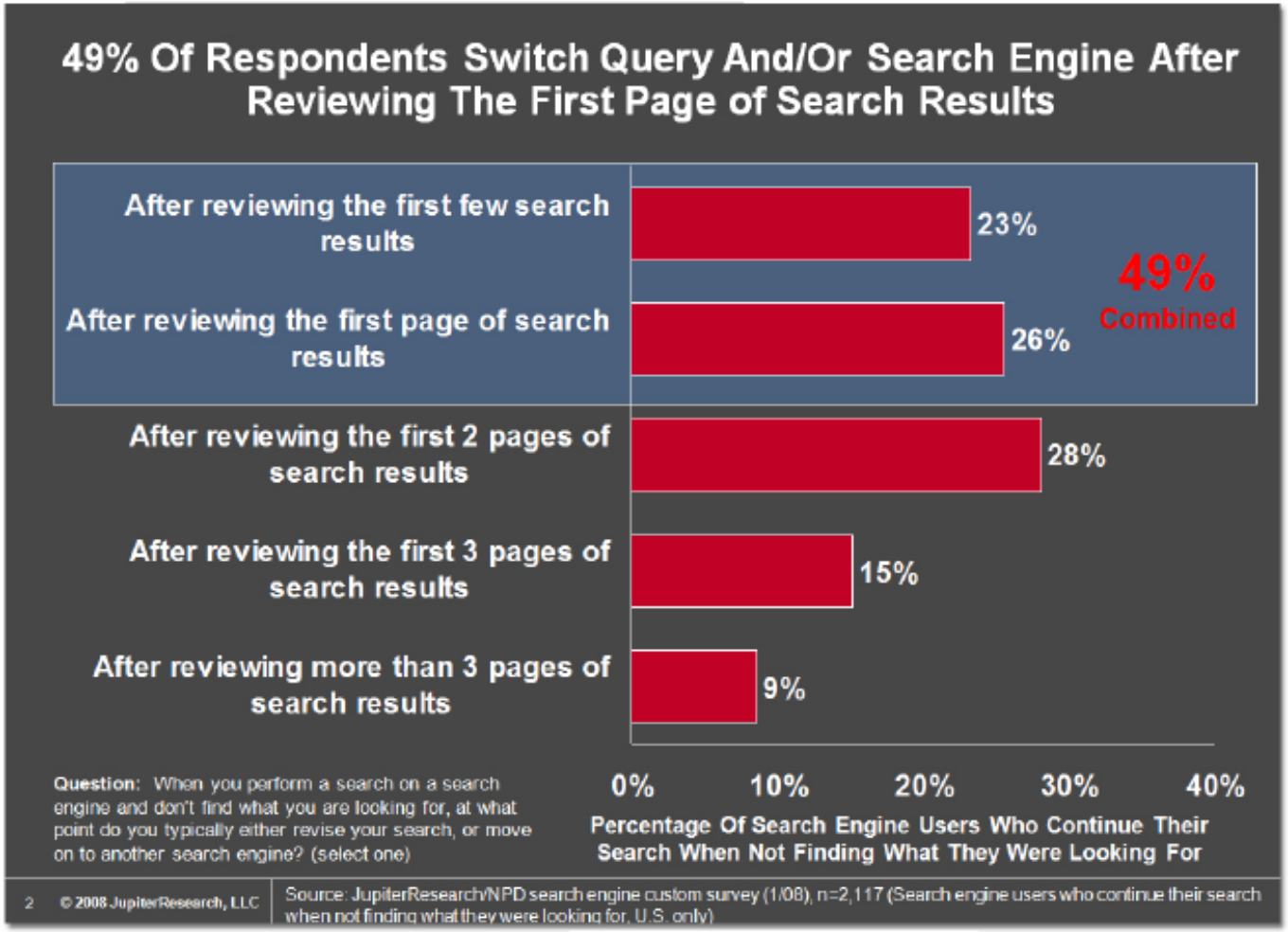
Evolution of behavior

“The first three pages of search results now appear to be the *last frontier*”

	2008	2006	2004	2002
Only a few	27%	23%	24%	16%
The first page	41%	39%	36%	32%
The first 2 pages	17%	19%	20%	23%
The first 3 pages	7%	9%	8%	10%
More than 3 pages	8%	10%	12%	19%

Source: iprospect.com *iProspect Blended Search Results Study – April 2008*

“When you […] don’t find what you are looking for, [when] do you […] revise your search, or move on to another search engine?”



Source: iprospect.com *iProspect Blended Search Results Study – April 2008*

Same question over time

	2008	2006	2004	2002
After reviewing the first few search results	23%	16%	23%	14%
After reviewing the first page of search results	26%	25%	19%	14%
After reviewing the first 2 pages of search results	28%	27%	26%	28%
After reviewing the first 3 pages of search results	15%	20%	15%	22%
After reviewing more than 3 pages of search results	9%	12%	17%	22%

Source: iprospect.com *iProspect Blended Search Results Study – April 2008*

Typical Session

- Two queries of
- .. two words, looking at…
- .. two answer pages, doing
- .. two clicks per page

- What is the goal?

MP3

games

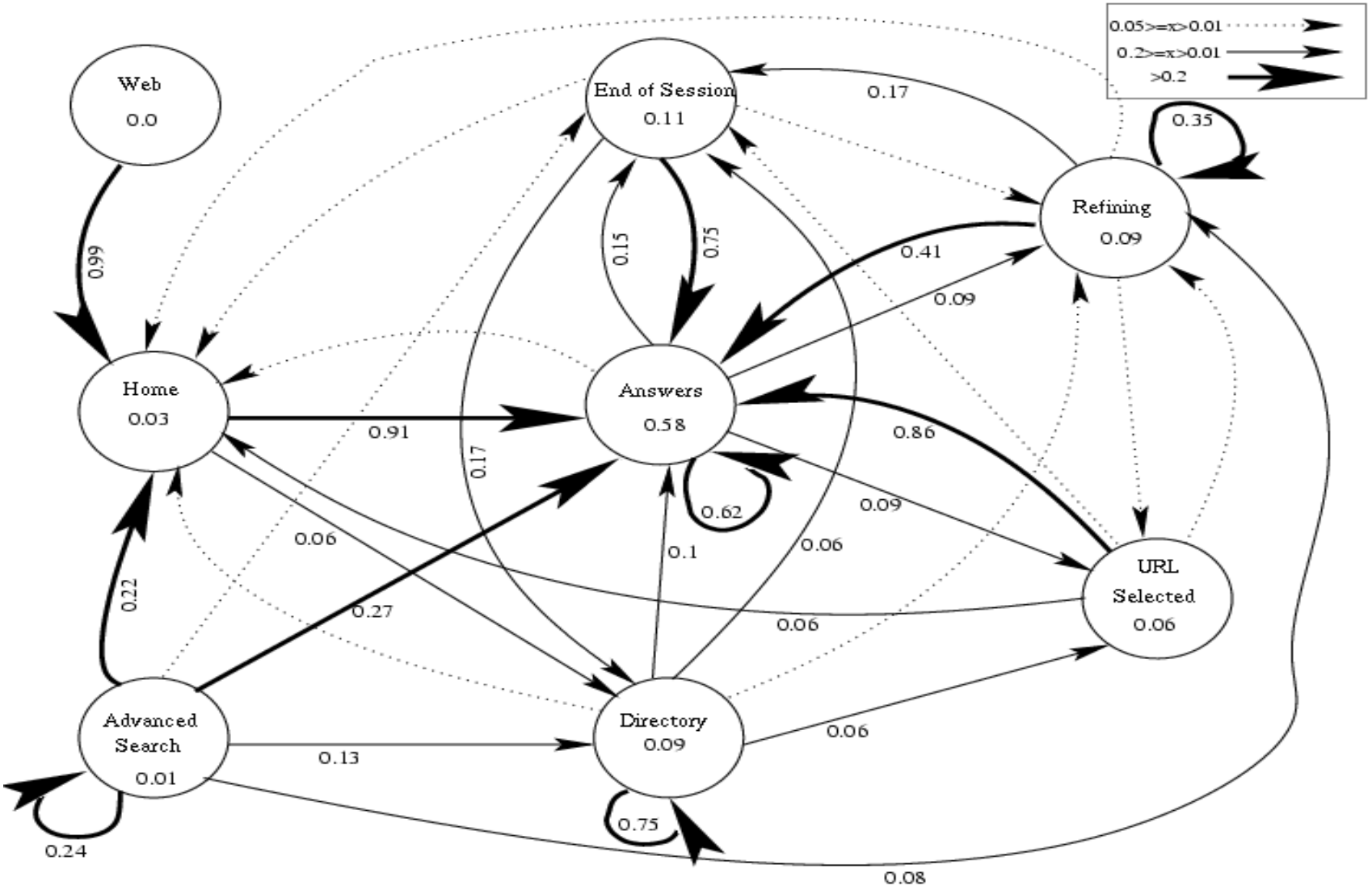
cars

lady gaga

pictures

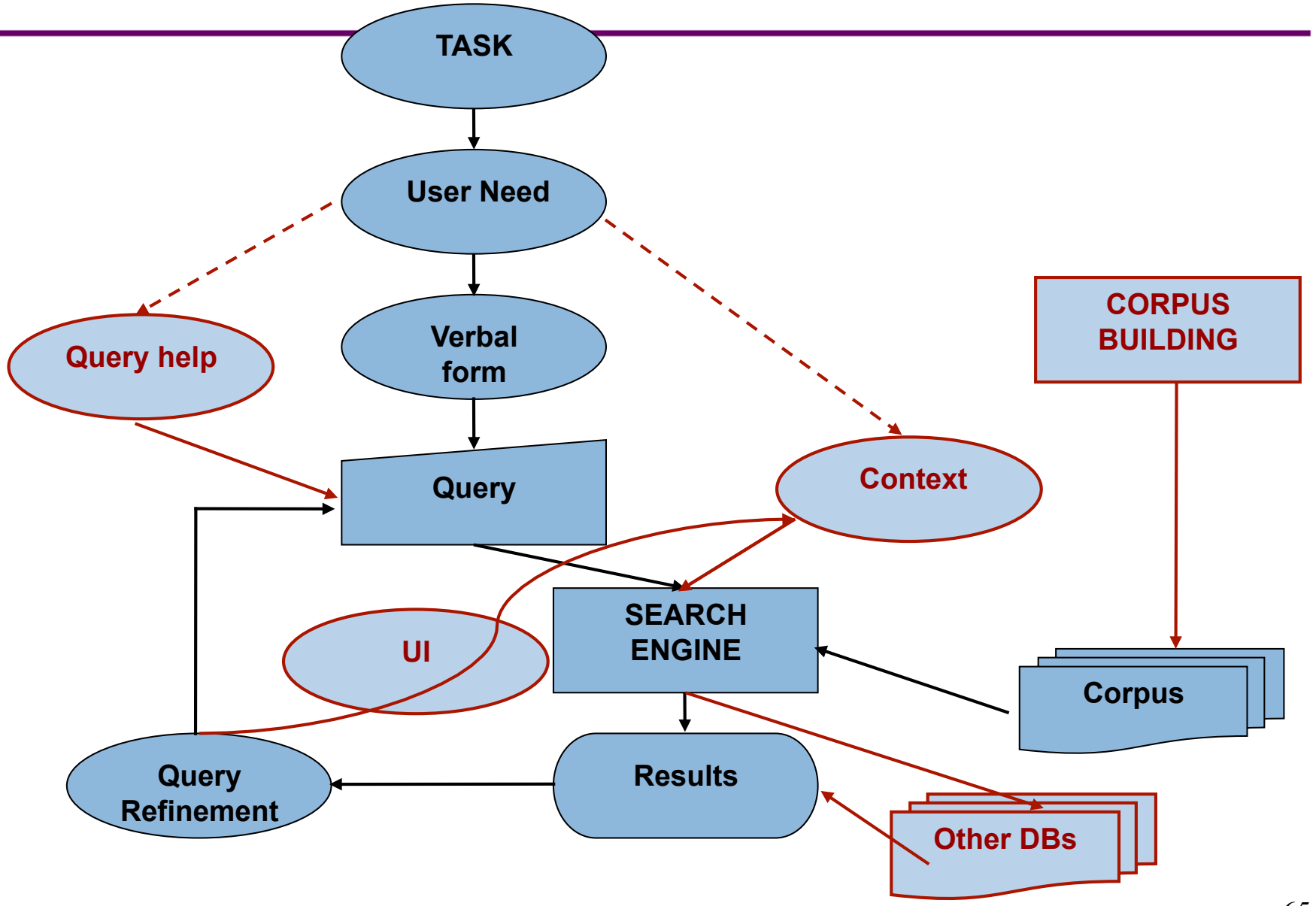
ski

User Modeling



Web IR Model

Search engine ➔ Universal Information finding environment



(3) Ranking

Link analysis

- Infer properties of Web entities based on their connectivity / link structure of graph structures they belong to
- Such properties can be importance of nodes or similarity between nodes
- Mostly focused on Web pages, but ideas apply to many domains: social networks, query logs, etc.

- Prestige, centrality, co-citation, PageRank, HITS

Social sciences and bibliometry

"...we are involved in an 'infinite regress': [an actor's status] is a function of the status of those who choose him; and their [status] is a function of those who choose them, and so ad infinitum"

[Seeley, 1949]

Citation Analysis

- **Citation frequency**
- **Co-citation coupling frequency**
  - Co-citations with a given author measures "impact"
  - Co-citation analysis [Mcca90]
    - Convert frequencies to correlation coefficients, do multivariate analysis/clustering, validate conclusions
    - E.g., co-citation in the "Geography and GIS" web shows communities [Lars96]
- **Bibliographic coupling frequency**
  - Articles that co-cite the same articles are related
- **Citation indexing**
  - Who is a given author cited by? (Garfield [Garf72])
    - E.g., Science Citation Index (*http://www.isinet.com/*)
    - CiteSeerX (*http://citeseerx.ist.psu.edu*) [Lawr99a]

Prestige

- Consider a graph $G=(V,E)$
- $E[u,v] = 1$ if there is a link from u to v
- $E[u,v] = 0$ otherwise
- prestige vector: $p[u]$ the prestige score of node u

$$p' = E^T p$$

because

$$p[u] = \sum_v E[v,u]\, p[u] = \sum_v E^T[u,v]\, p[u]$$

- After each iteration normalize by setting $||p|| = 1$
- p converges to the principal eigenvector of E^T

Centrality

- Importance notion based on centrality
- Used by epidemiology, social-network analysis, etc.: removing a central node disconnects the graph to a big extent

- *d(u,v)* the shortest-path distance between *u* and *v*
- $r(u) = \max_v d(u,v)$ *radius* of node *u*
- $\arg\min_u r(u)$ *center* of the graph

- Various other notions of centrality in the literature

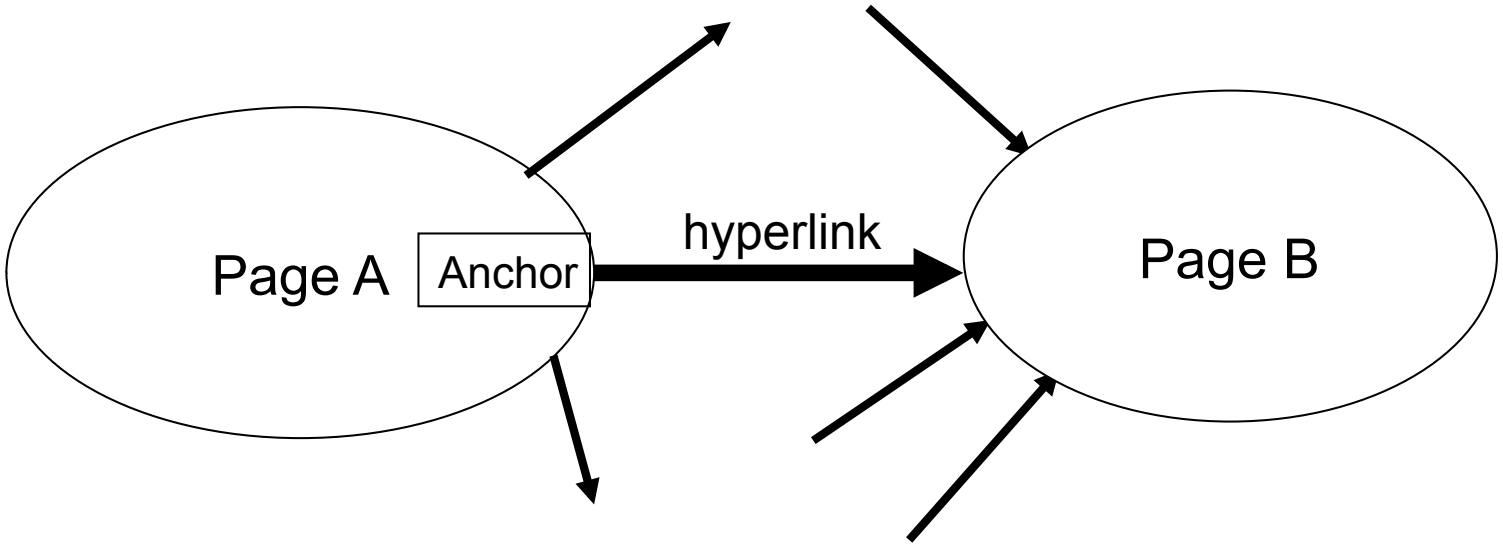
Co-citation

- Measure of similarity between nodes
- If nodes *v* and *w* are both linked by node *u*, then they are co-cited
- If *E* is the adjacency matrix of the graph, the number of nodes that co-cite both *v* and *w* is

$$p[u] = \sum_u E[u,v]\, E[u,w] = \sum_u E^T[v,u]\, E[u,w] = (E^T E)[v,w]$$

- Thus similarity is captured in the entries of the matrix $E^T E$

The Web as a Directed Graph

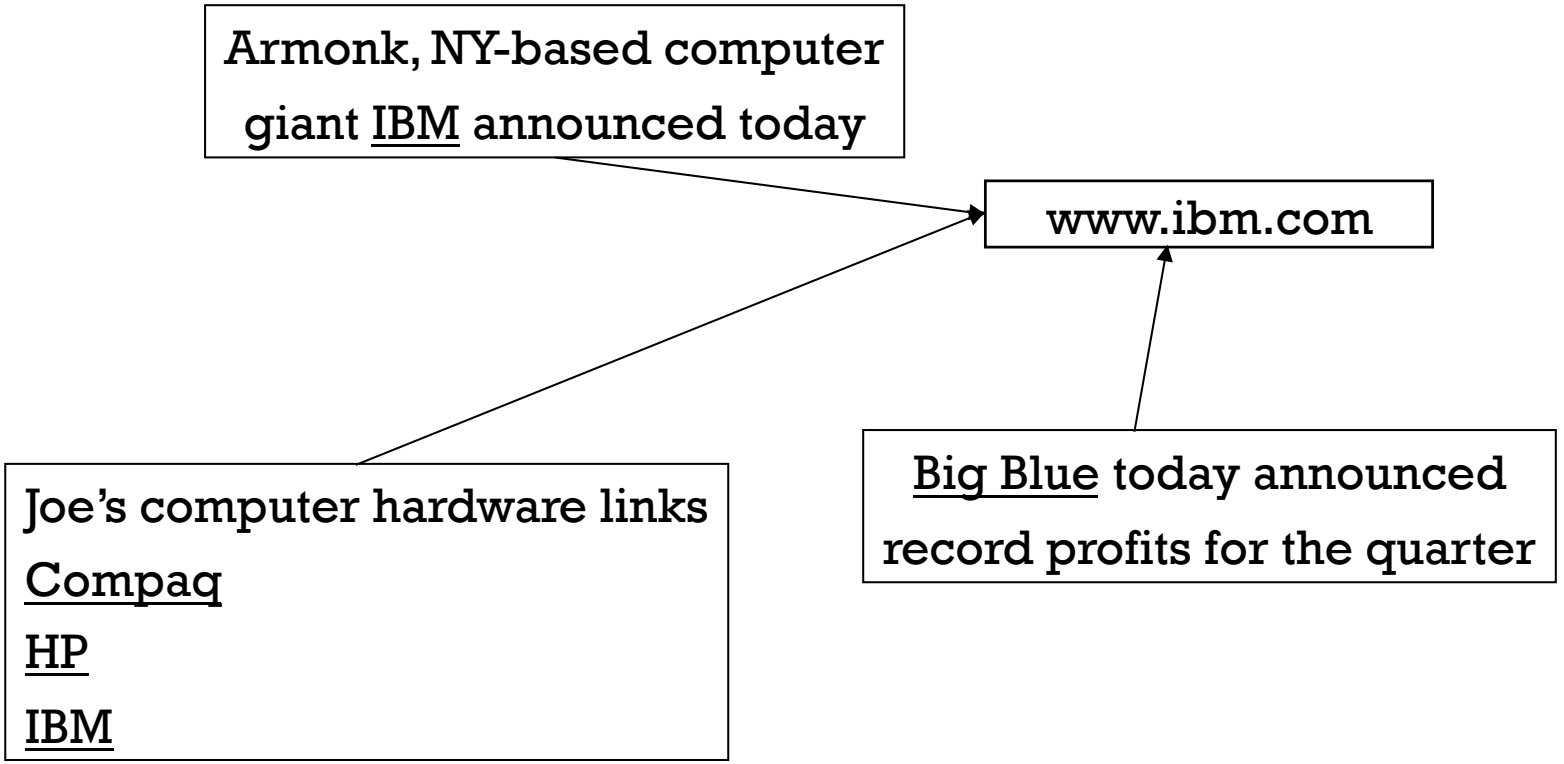


Assumption 1: A hyperlink between pages denotes author perceived relevance (quality signal)

Assumption 2: The anchor of the hyperlink describes the target page (textual context)

Indexing anchor text

- **When indexing a document *D*, include anchor text from links pointing to *D*.**



Link Ranking

- **Incoming links count & variations**
  - Li / Marchiori / Carriere *et al.,* 1997; Joo & Myaeng, 1998
- **HITS (Kleinberg, 1998)**
  - Authorities: good pages
  - Hubs: good links
- **PageRank (Page & Brin, 1998)**
  - Random walk + random jumps if "bored"
- **Many variations of these ideas**
- **Good to find communities, spam, etc.**
- **Application to other problems (e.g. ranking DB relations)**
- **Today: just a component of a search engine ranking**

PageRank

- *good pages point to good pages* **[Brin and Page, 1998]**
- **Algorithm for ranking results in web search**
- **An authority score is assigned to each Web page**
- **Authority scores independent of the query**

- **Authority scores corresponds to the stationary distribution of a random walk on the graph:**
  - With probability α follow a link in the graph
  - With probability $1-\alpha$ go to a node chosen uniformly at random (teleportation)

- **Random walk also known as random surfer model**

PageRank

- Let *E* be the adjacency matrix of the graph, and *L* the row-stochastic version of *E*
- Each row of *E* is normalized so that it sums to *1*
- Authority score defined by

$$p_{(i+1)} = L^T p_{(i)}$$

- problematic if the graph is not strongly connected, So:

$$p_{(i+1)} = \alpha\, L^T p_{(i)} + (1-\alpha) 1/n\, 1$$

- where *1* is the matrix with all entries equal to *1*
- and $\alpha \in [0,1]$, common value $\alpha = 0.85$

PageRank variants and enchancements

- **Dealing with sink nodes**
- **Personalized PageRank**
  - Teleportation to a set of pages defining the preferences of a particular user
- **Topic-sensitive PageRank [Haveliwala 02]**
  - Teleportation to a set of pages defining a particular topic
- **TrustRank [Gyöngyi 04]**
  - Teleportation to “trustworthy” pages

- **Many papers on analyzing PageRank and numerical methods for efficient computation**

HITS

- **[Kleinberg 1998]**
- **Exploit the intuition that there are:**
  - pages that contain high-quality information (authorities)
  - pages with good navigational properties (hubs)

Good hubs point to good authorities and good authorities are pointed by good hubs

HITS algorithm

- Given a query *q*
- Use a standard wen IR system to find a set of pages *R* relevant to *q* (*root set*)
- Expand to the set of pages connected to *R* (*expanded set*) and form the graph *G=(V,E)*
- authority vector: *a[u]* the authority score of node *u*
- *h* hub vector: *h[u]* the hub score of node *u*

$$a = E^T h \qquad h = E a$$

- *a* converges to the principal eigenvector of $E^T E$
- *h* converges to the principal eigenvector of $E E^T$

HITS

- **HITS is related to SVD on the graph matrix E**
- **non-principal eigenvectors provide different topics**
- **HITS sensitive to local-topology**
- **PageRank is more stable – due to trandom jump step**
- **Researchers attempted to make HITS more stable**
  - SALSA stochastic algorithm for link analysis [Lempel and Moran, 01]:
  - A random surfer model in which the surfer follows alternatively random inlinks and outlinks
  - [Ng et al. 01] introduce a random jump step in the HITS model

Discussion

- HITS introduces the notion of hub, which does not exist in PageRank
- HITS is query sensitive
- PageRank does not depend on the query; thus the authority scores can be pre-computed
- Hence HITS is harder to compute on-line

Ranking in Practice

- Ranking function
  - Many features, hard to optimize
- Learning to rank
  - Machine learning finds the (linear) function
- Learn the ranking function
  - Use genetic algorithms to find the function

Main Considerations

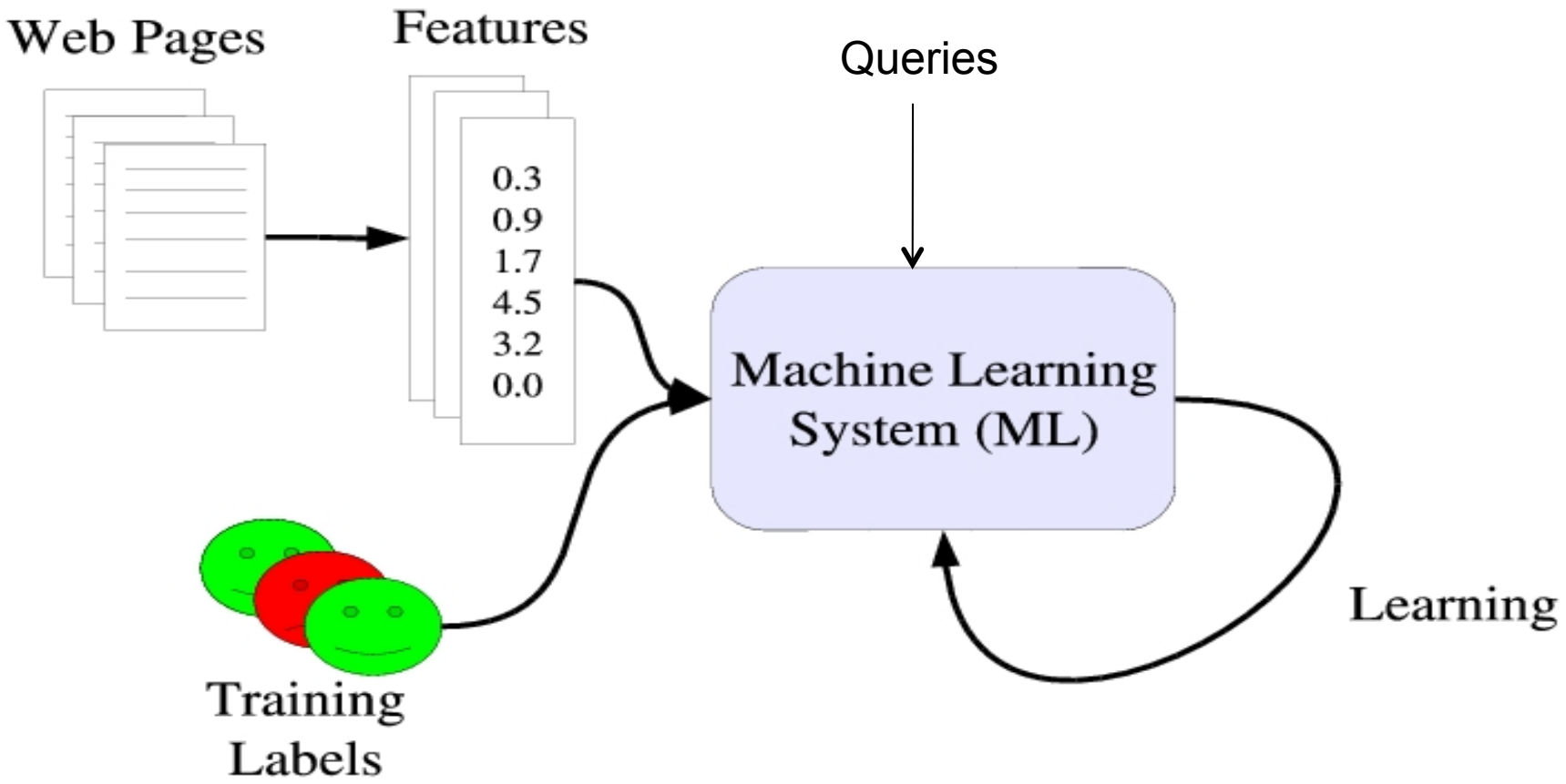
- Ranking must be integrated with the index
- Partial evaluation must be used
- Based on many features:
  - text content, link analysis, usage information, metadata, etc.
- Robust against spamming
- Evaluation: manual and automatic (click streams)

Features: Wisdom of Crowds

- Crucial for Search Ranking
- Text: Web Writers & Editors
  - not only for the Web!
- Links: Web Publishers
- Tags: Web Taggers
- Queries: All Web Users!
  - Queries and actions (or no action!)

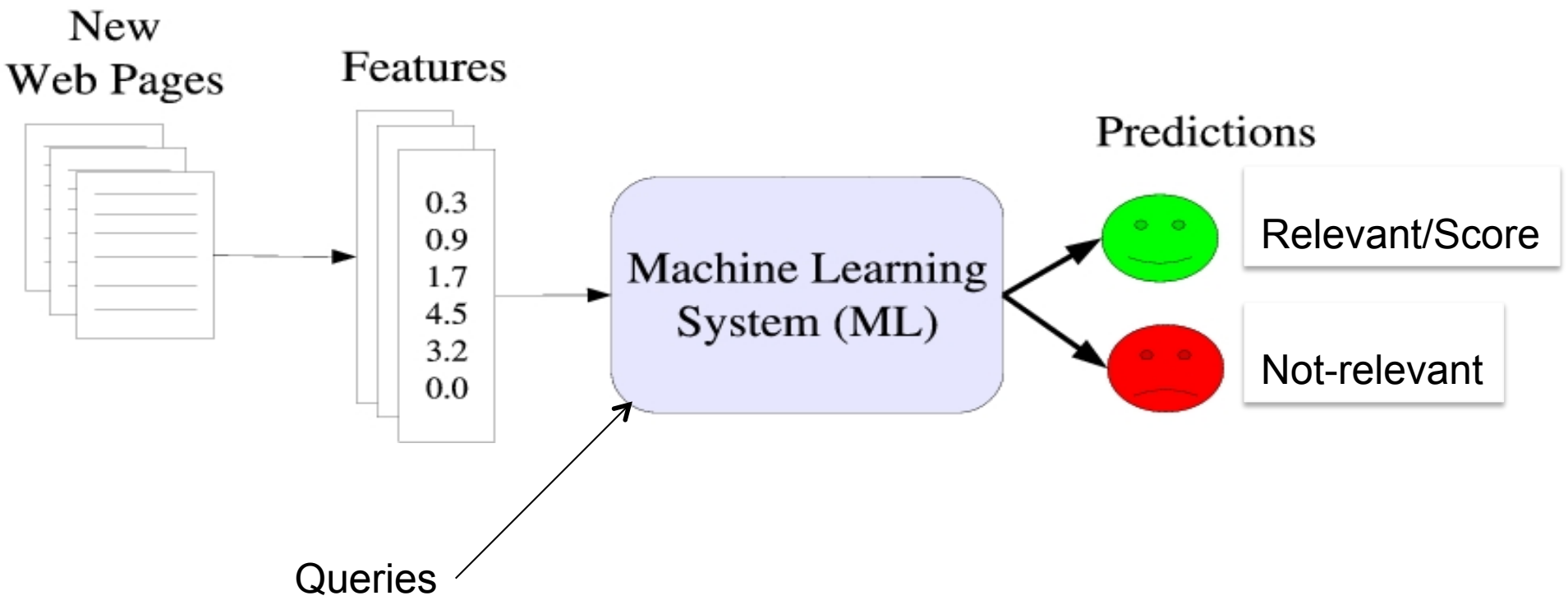
Learning to Rank: Content Features

- **Machine-learning approach --- training**



Learning to Rank: content features

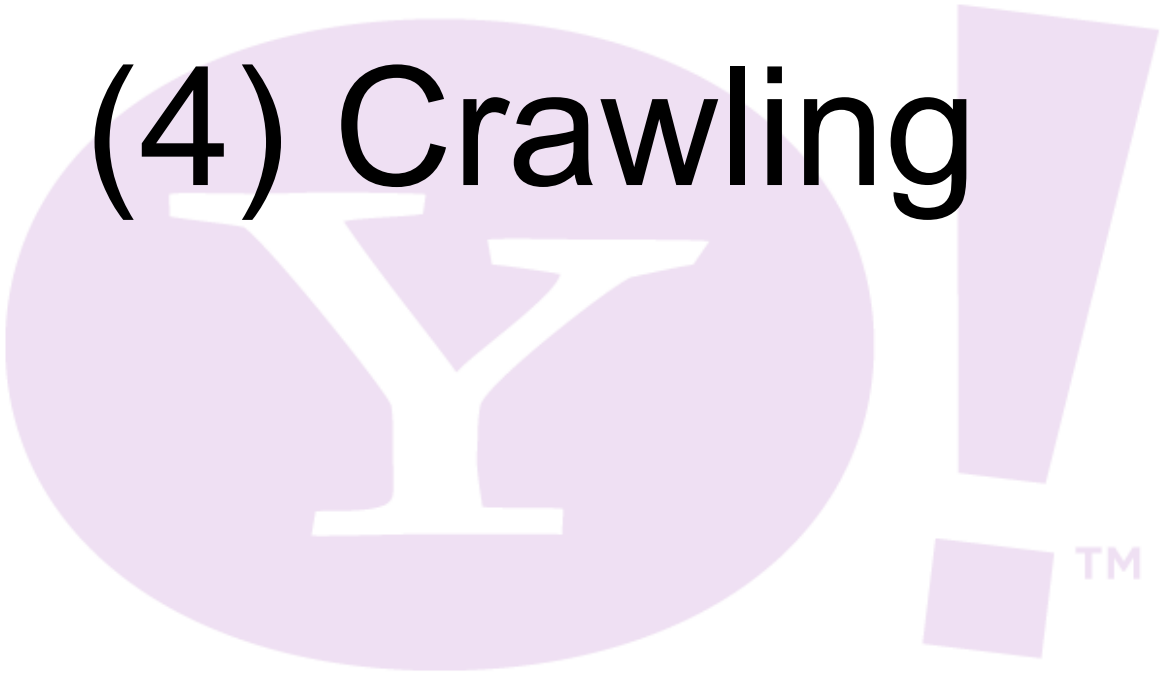
- **Machine-learning approach --- prediction**



Quality Evaluation

- **Manual: user quality evaluations**
- **Automatic: click-through as positive relevance feedback**
  - Monitor average click rank
  - Important to unbias the distribution
    - ranking bias
    - interface bias

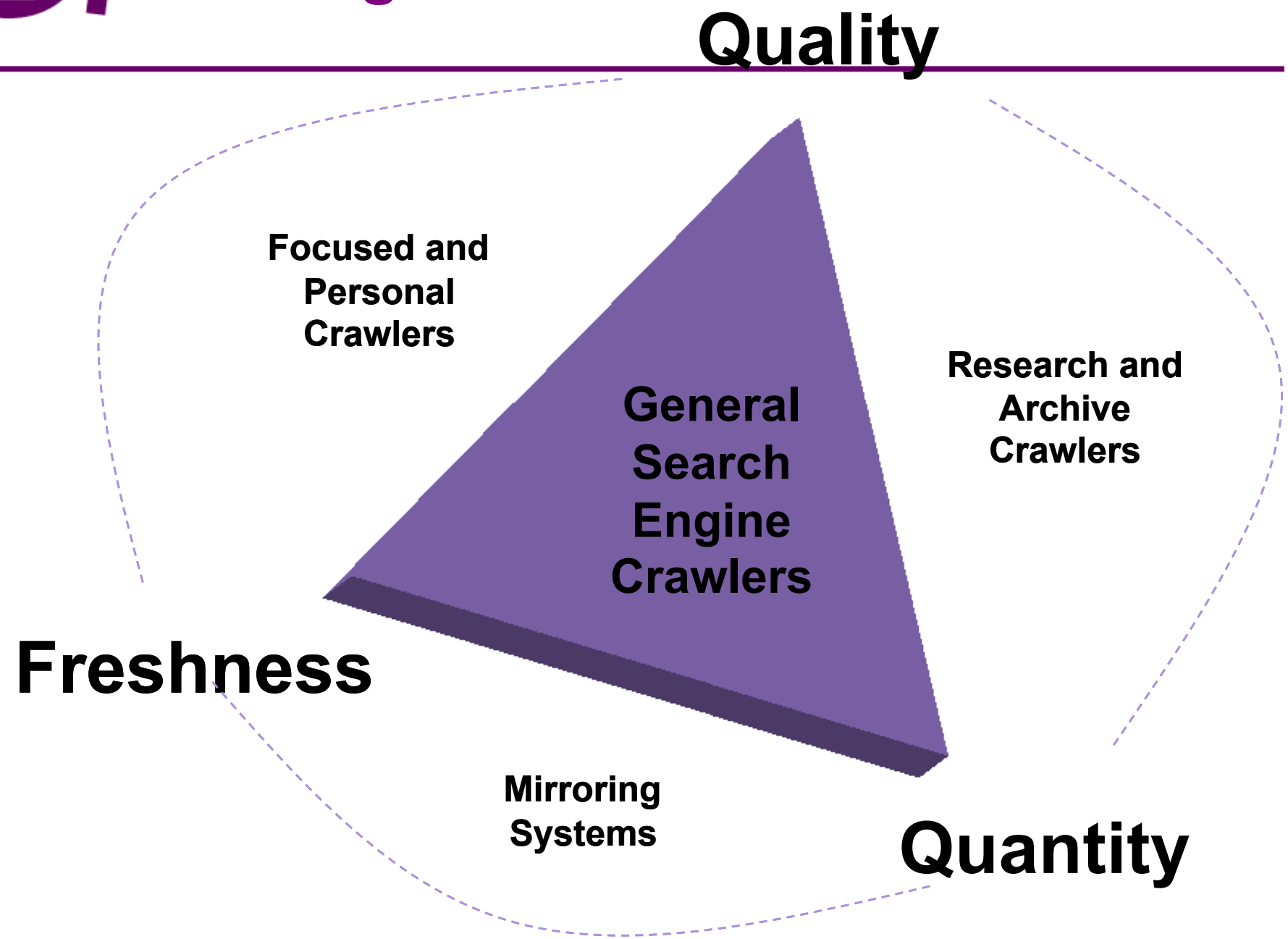
(4) Crawling

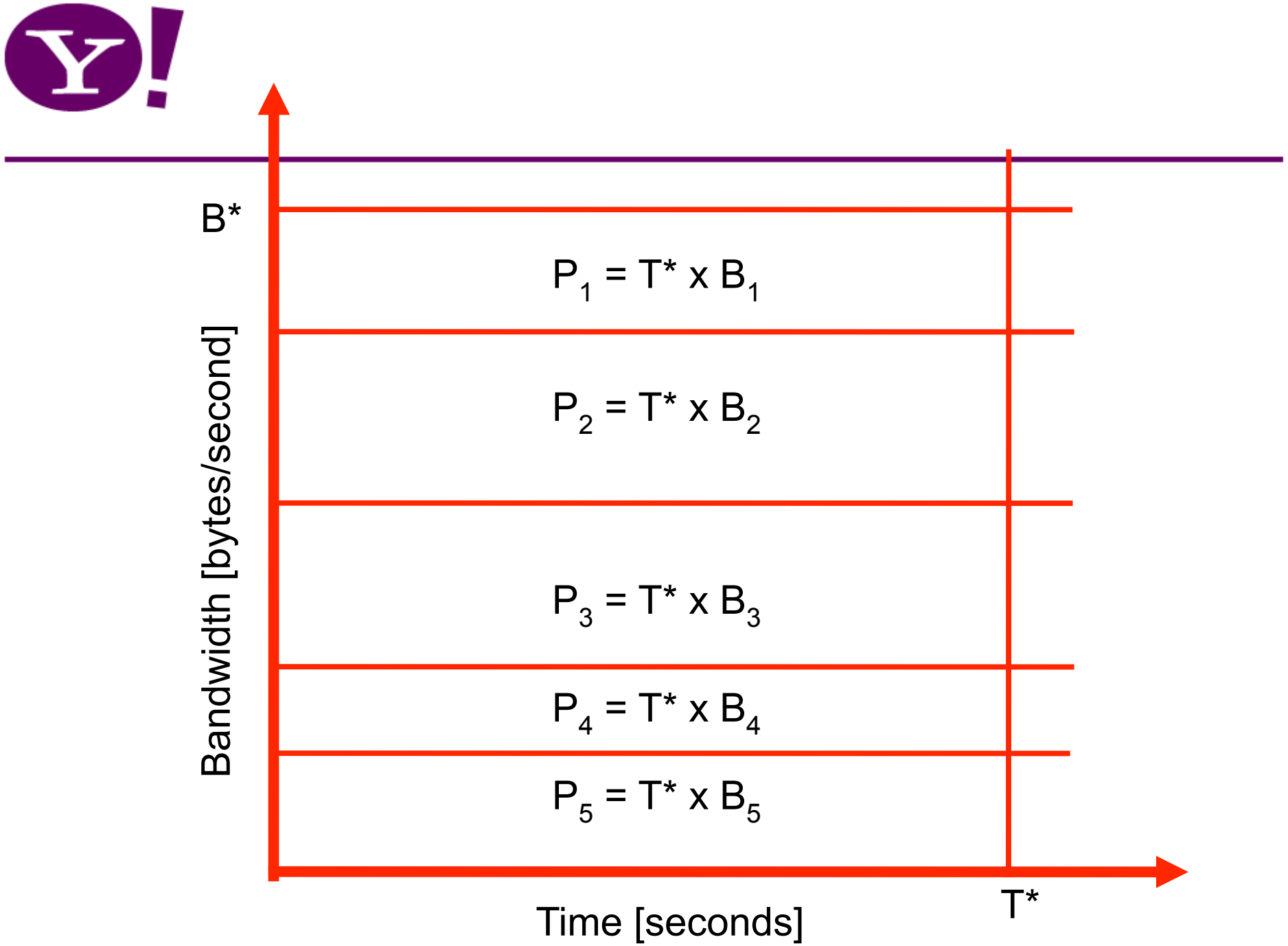


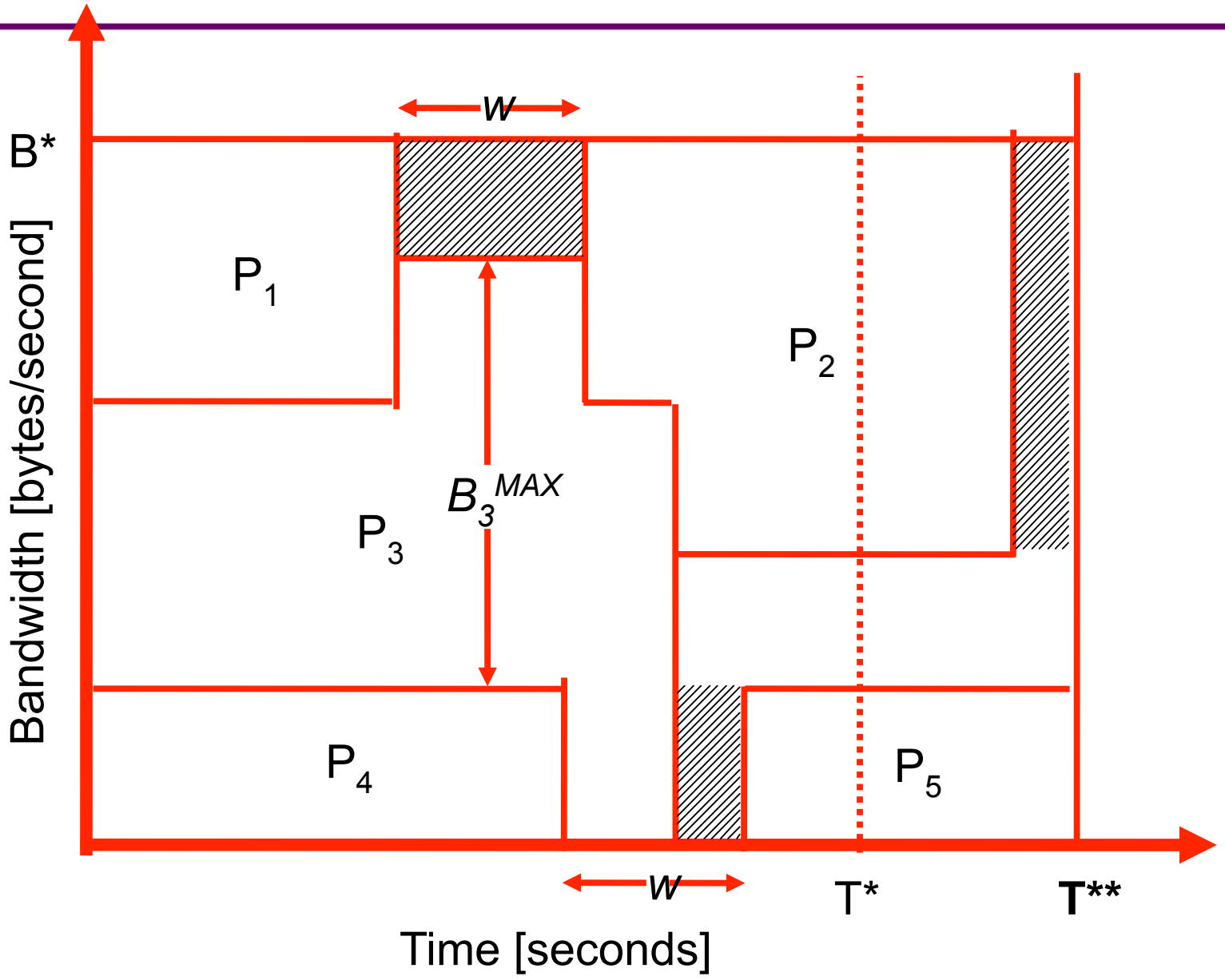
Crawling

- NP-Hard Scheduling Problem
- Different goals
- Many Restrictions
- Difficult to define optimality
- No standard benchmark

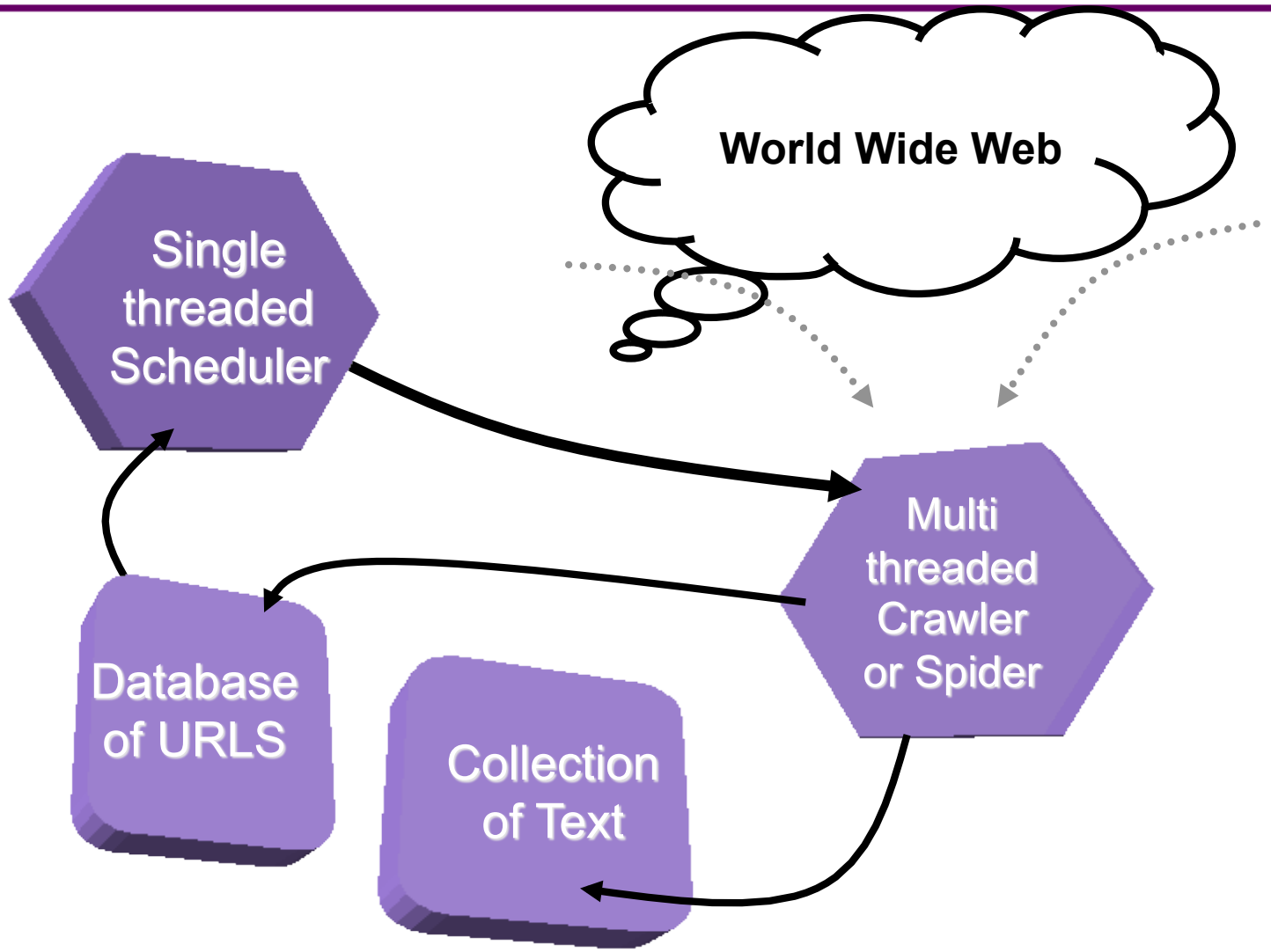
Crawling Goals

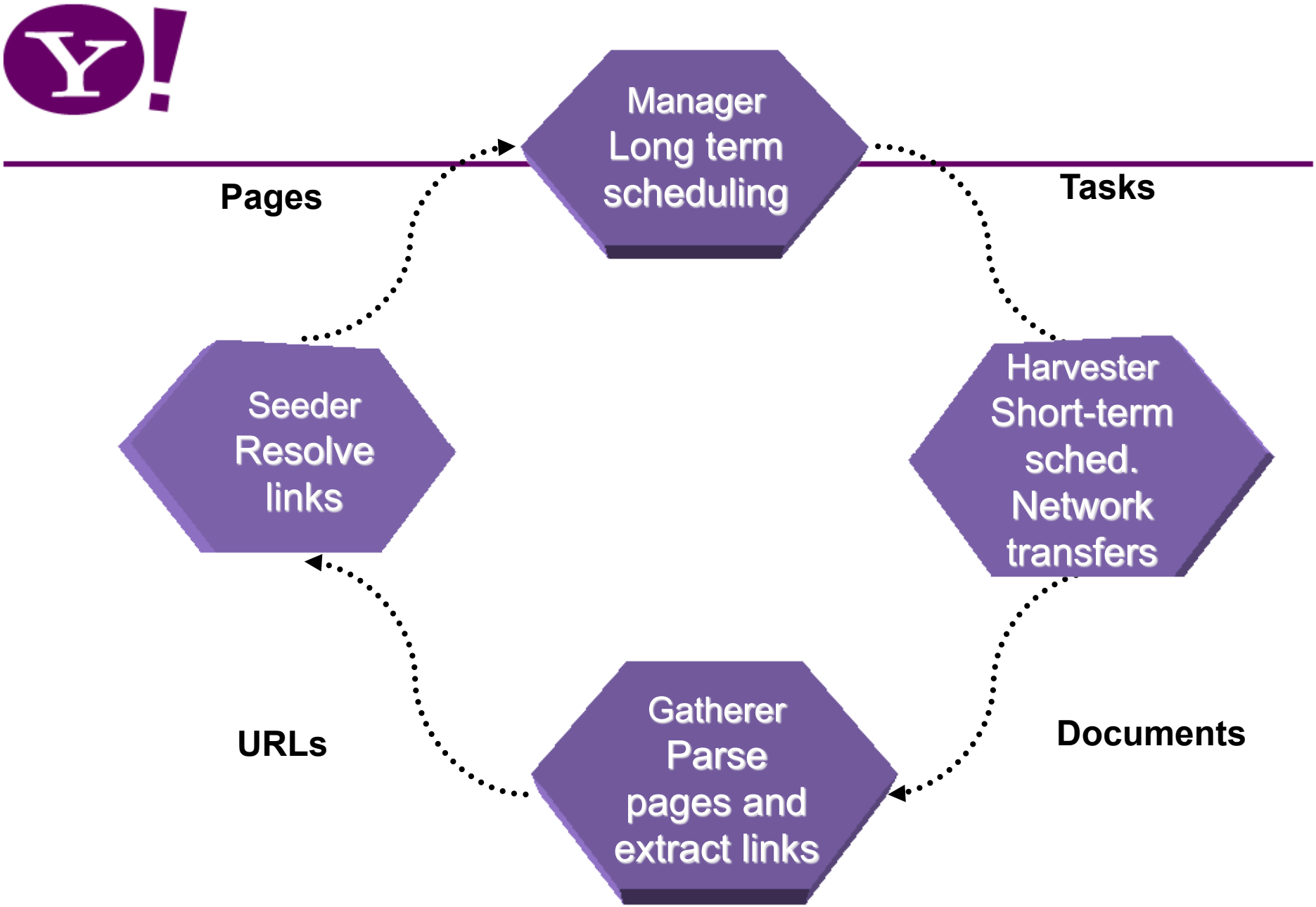



B*
Bandwidth [bytes/second]
P1 = T* x B1
P2 = T* x B2
P3 = T* x B3
P4 = T* x B4
P5 = T* x B5
T*
Time [seconds]

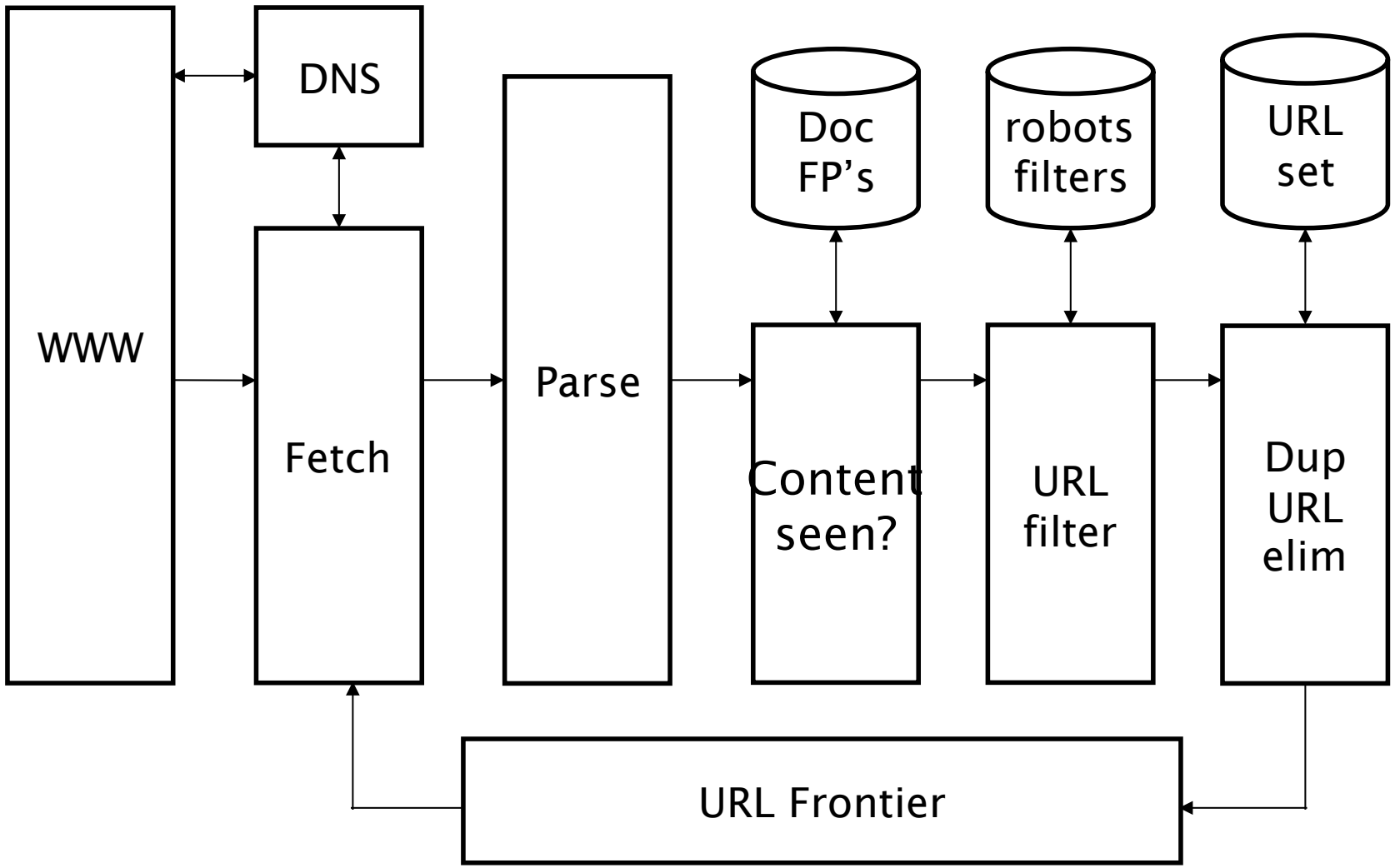

w
B*
Bandwidth [bytes/second]
P1
P2
P3
B3 MAX
P4
P5
w
T*
T**
Time [seconds]

Software Architecture

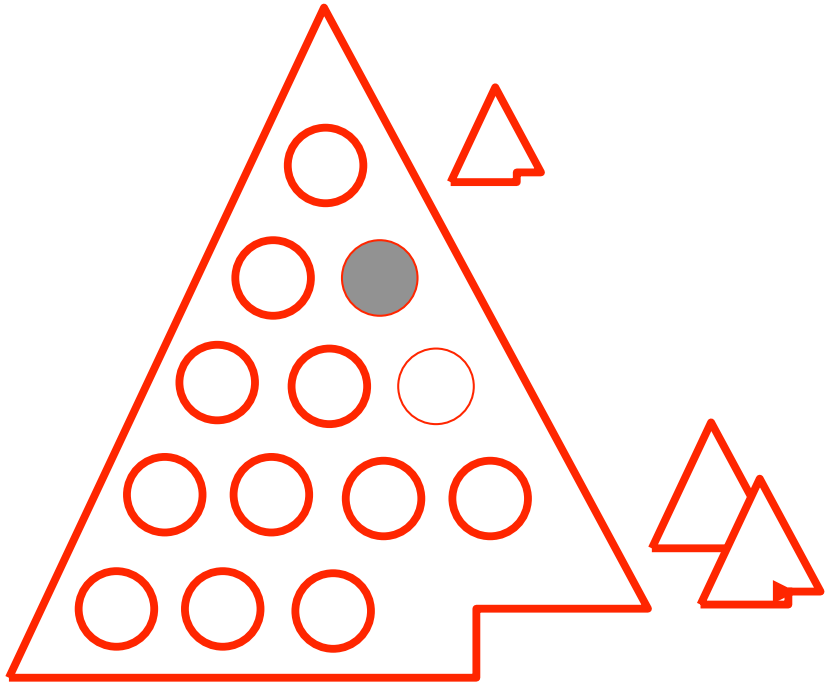



Manager
Long term
scheduling
Pages
Tasks
Seeder
Resolve
links
Harvester
Short-term
sched.
Network
transfers
URLs
Gatherer
Parse
pages and
extract links
Documents

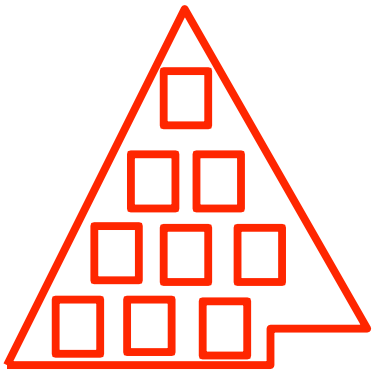
Basic Crawl Architecture



Priority Queues



Queue of Web sites
(long-term scheduling)



Queue of Web pages
for each site
(short-term scheduling)

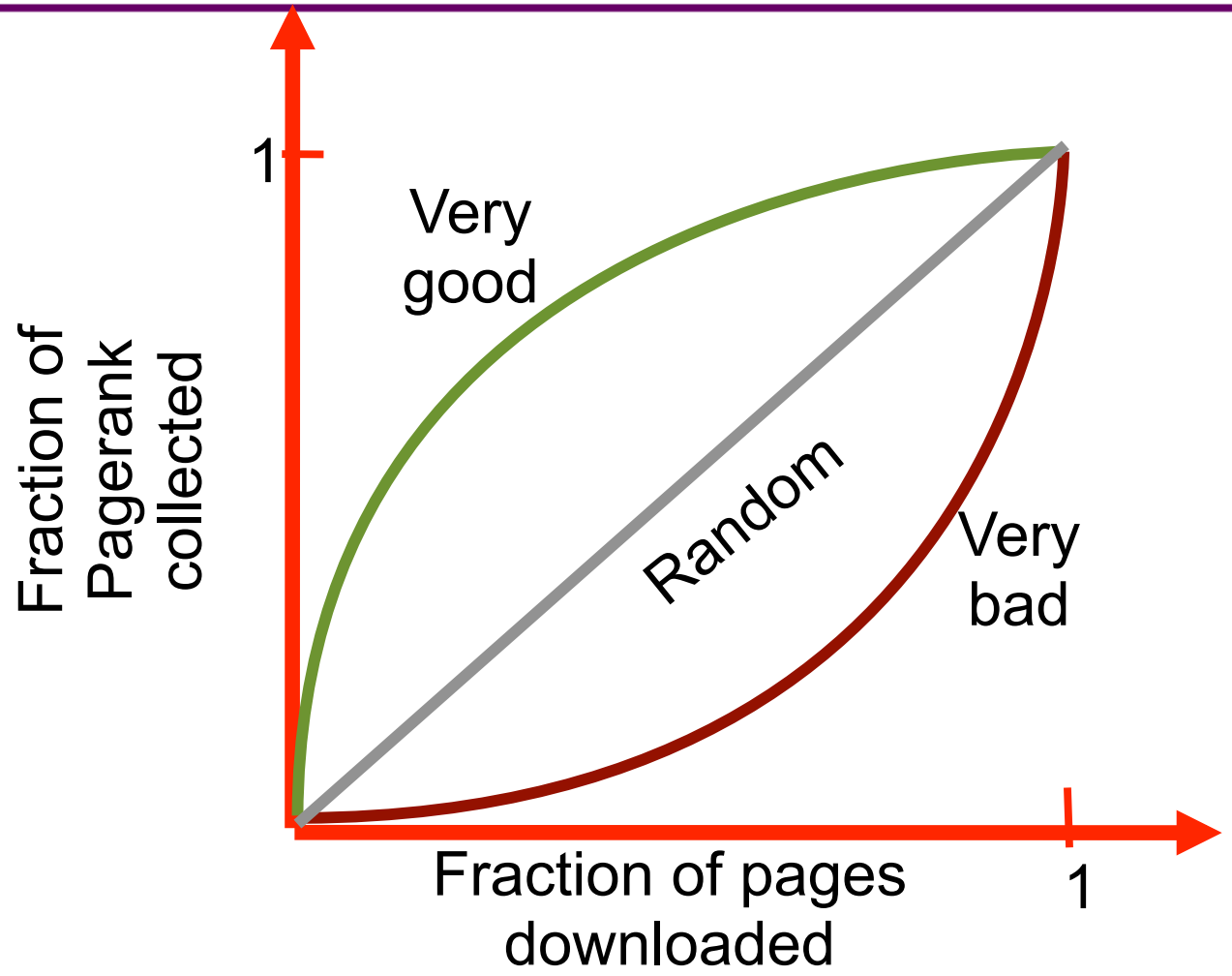
Formal Problem

- **Find a sequence of page requests *(p,t)* that:**
  - Optimizes a function of the volume, quality and freshness of the pages
  - Has a bounded crawling time
  - Fulfils politeness
  - Maximizes the use of local bandwidth
- **Must be on-line: how much knowledge?**

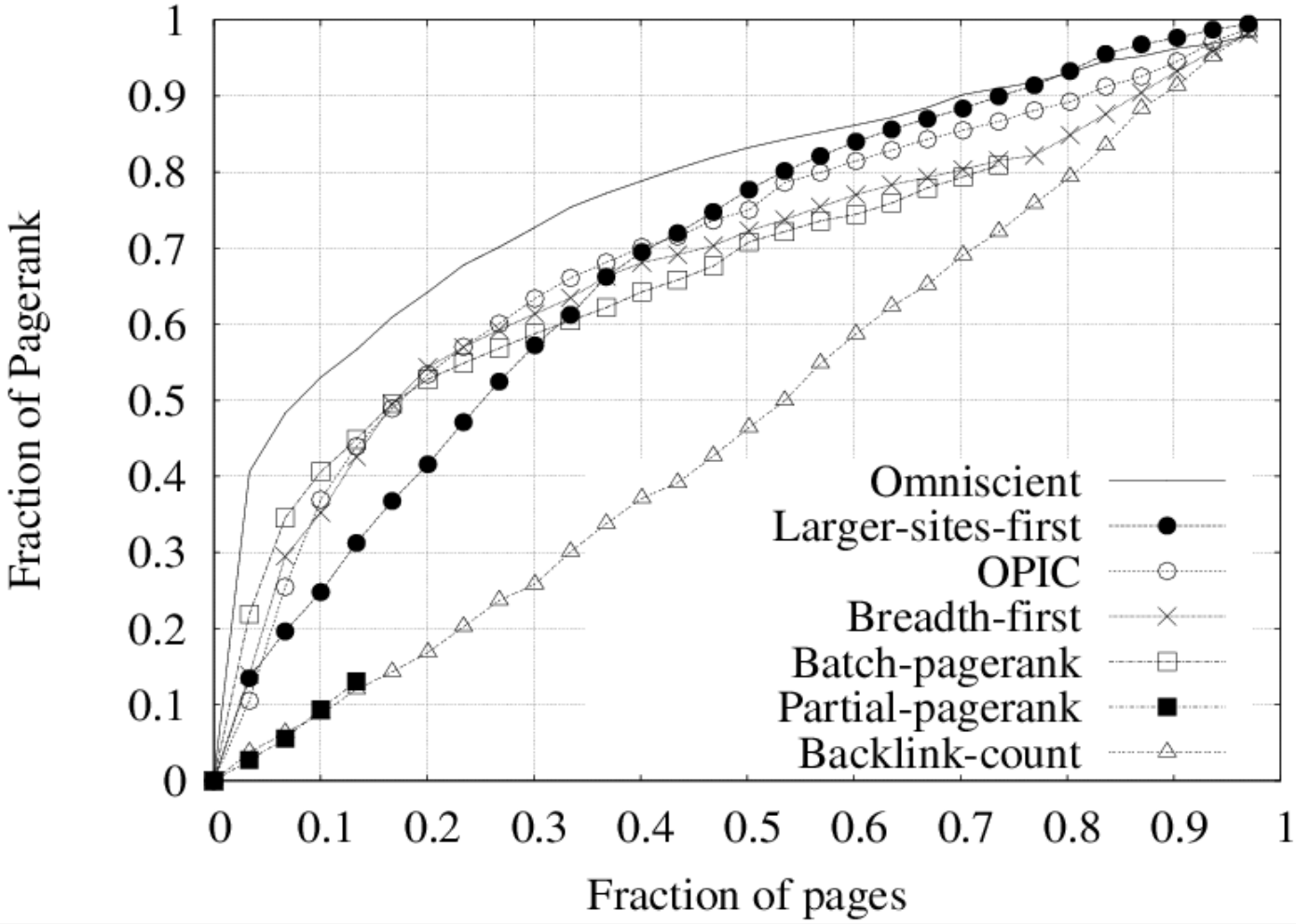
Crawling Heuristics

- **Breadth-first**
- **Ranking-ordering**
  - PageRank
- **Largest Site-first**
- **Use of:**
  - Partial information
  - Historical information
- **No Benchmark for Evaluation**
  - Use simulation

Comparing crawling algorithms

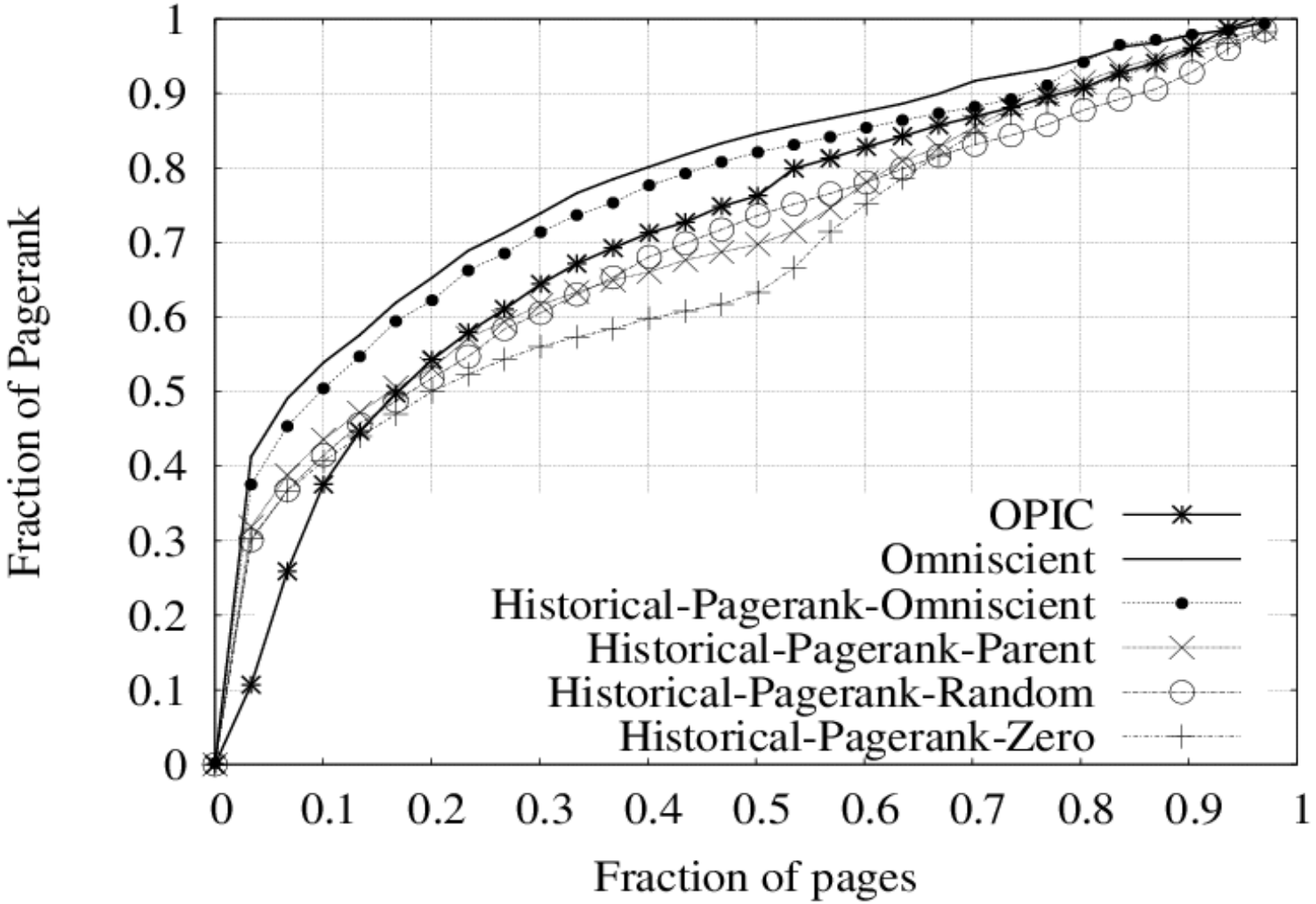


No Historical Information

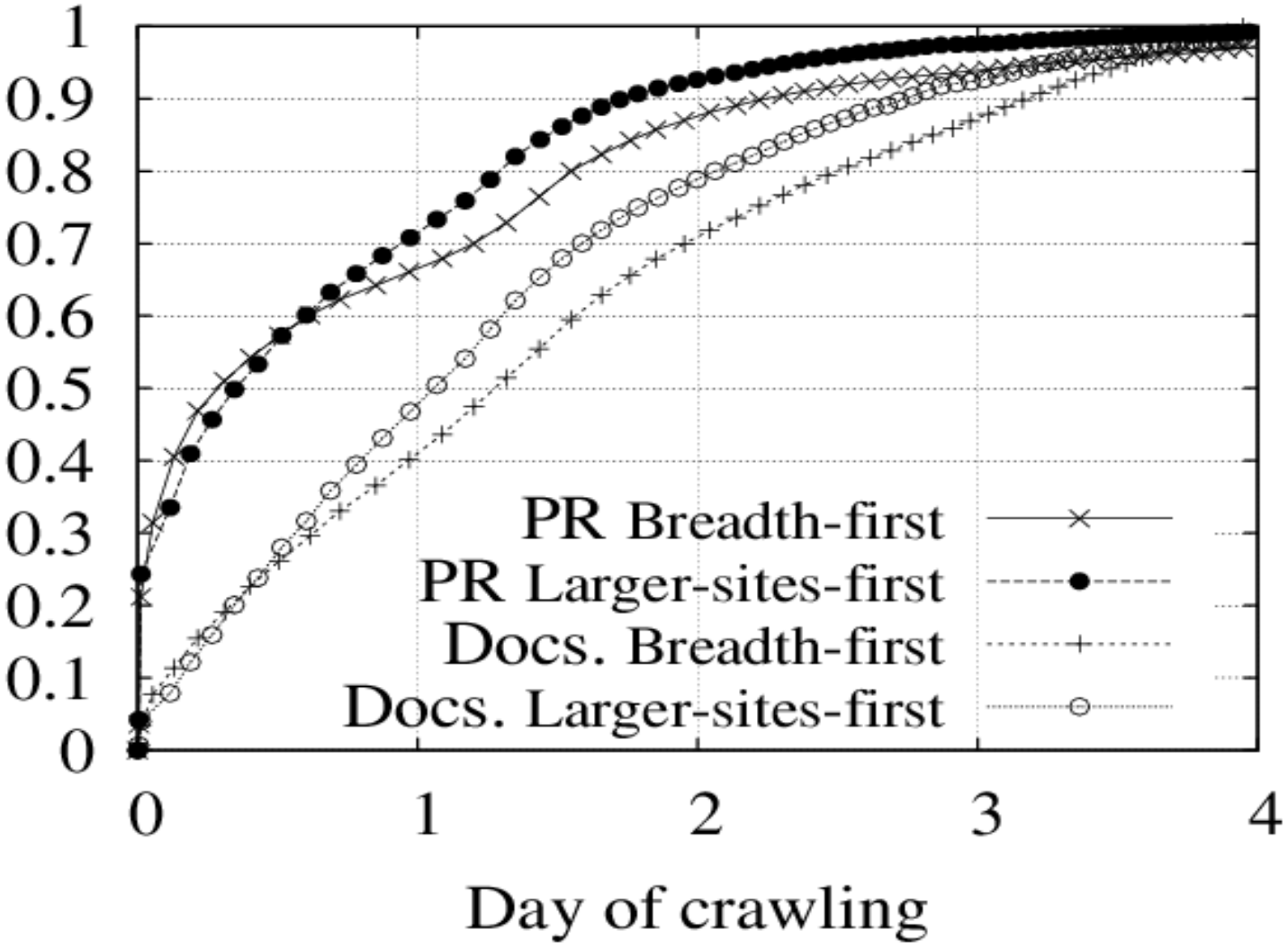


Baeza-Yates, Castillo, Marin & Rodriguez, WWW2005

Historical Information



Validation in the Greek domain

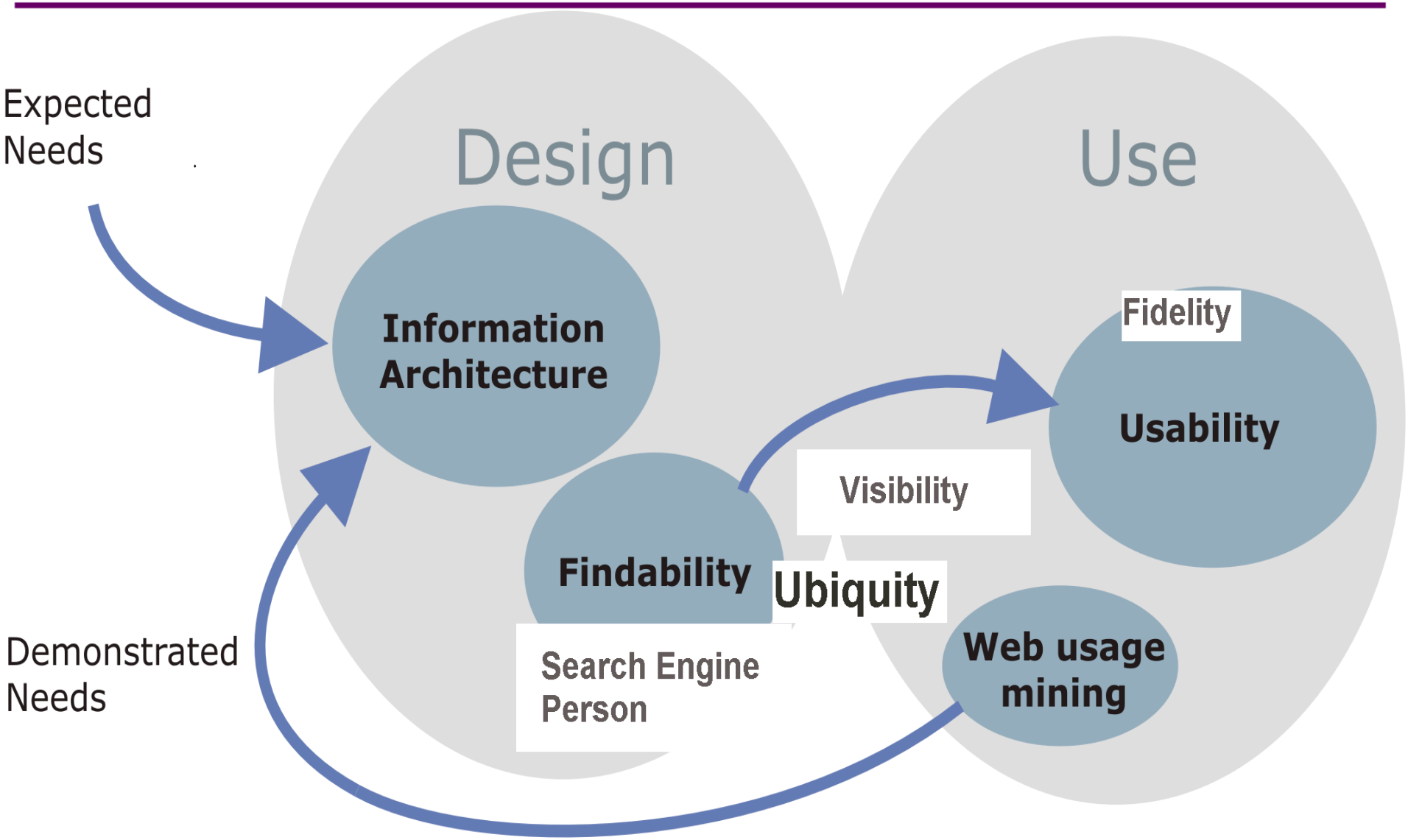


(5) Final Remarks

Young research field

- The Web is scientifically young
- The Web is intellectually diverse
- The technology mirrors the economic, legal and sociological reality
- Search is evolving to “task completion” and implicit search
- Plenty of open problems

Web Design and Search



Main Open Problems

Large scale usage data is key BUT

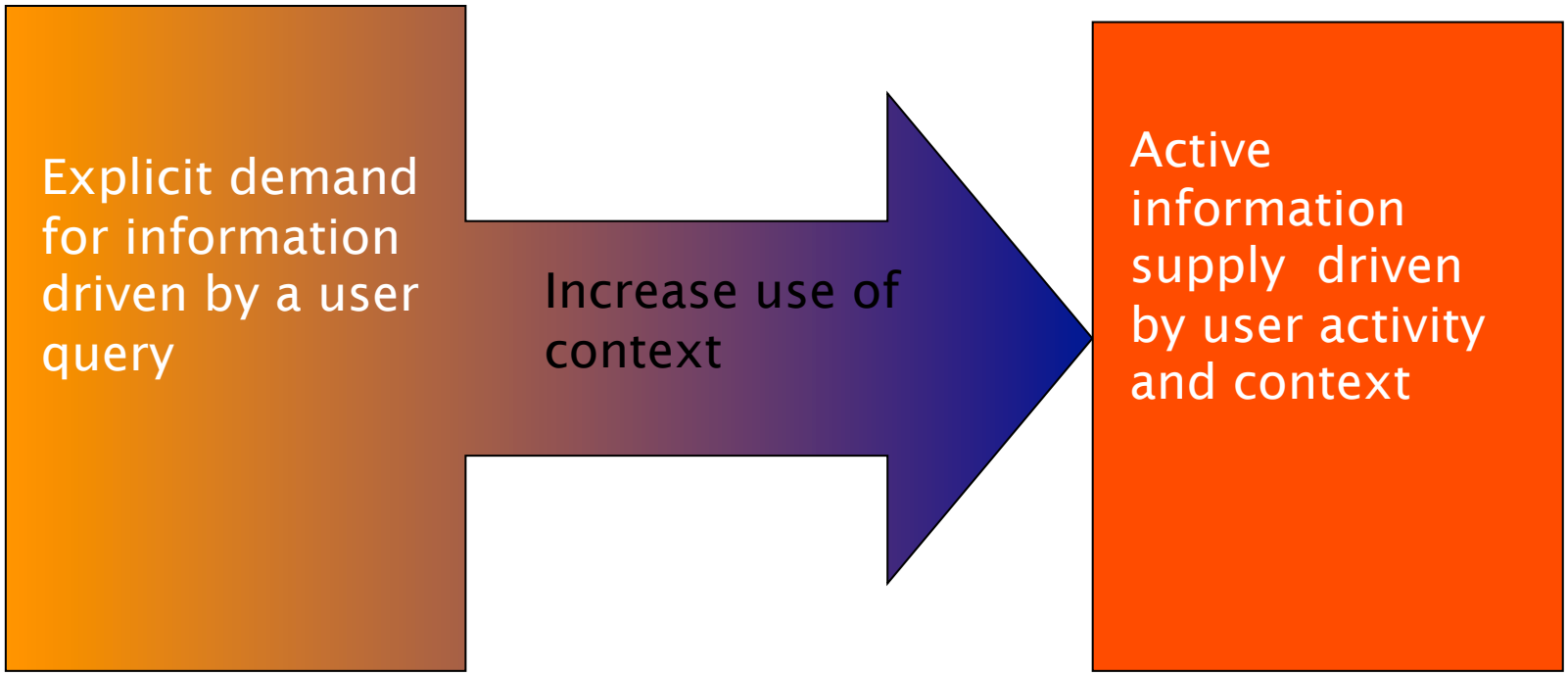
Usage data at a very large scale	
over larger and larger populations over longer and longer periods of time	
Personalization	**Privacy**
More data via larger communities, makes data less personalized *wisdom of crowds does not work well on small corpora*	Over personalization endangers privacy Long-term logs endanger privacy

The New frontiers

- **Front-end and user experience**
  - The most probable reason for users to switch between quasi-equivalent engines is a better user experience

- **Depart from the rectangle/ranked list paradigm**
  - Get rid of queries? Implicit search
    - Content delivery is one flavor
    - But in general, why should we even have to formulate a query?

What's next? Fourth generation: **From Information Retrieval to Information Supply**



Bibliography – General

- **Modern Information Retrieval**

  by R. Baeza-Yates & B. Ribeiro-Neto, Addison-Wesley, 1999. Second edition in 2011.

- **Introduction to Information Retrieval**

  by C. Manning, P. Raghavan & H. Schuetze,

  Cambridge University Press, 2008.

- **Web Search: The Role of the Users.**

  by R. Baeza-Yates & Y. Maarek, 2011.

  Tutorial at SIGIR 2011 and other conferences.

- **Websites:**
  - http://www.searchenginewatch.com/
  - http://www.searchengineshowdown.com/

- **Main upcoming conferences:**

  SPIRE 2011, October, Pisa, Italy
  WSDM 2012, February, Seattle, USA
  ECIR 2012, April, Barcelona, Spain
  ACM SIGIR 2012, July, Portland, USA