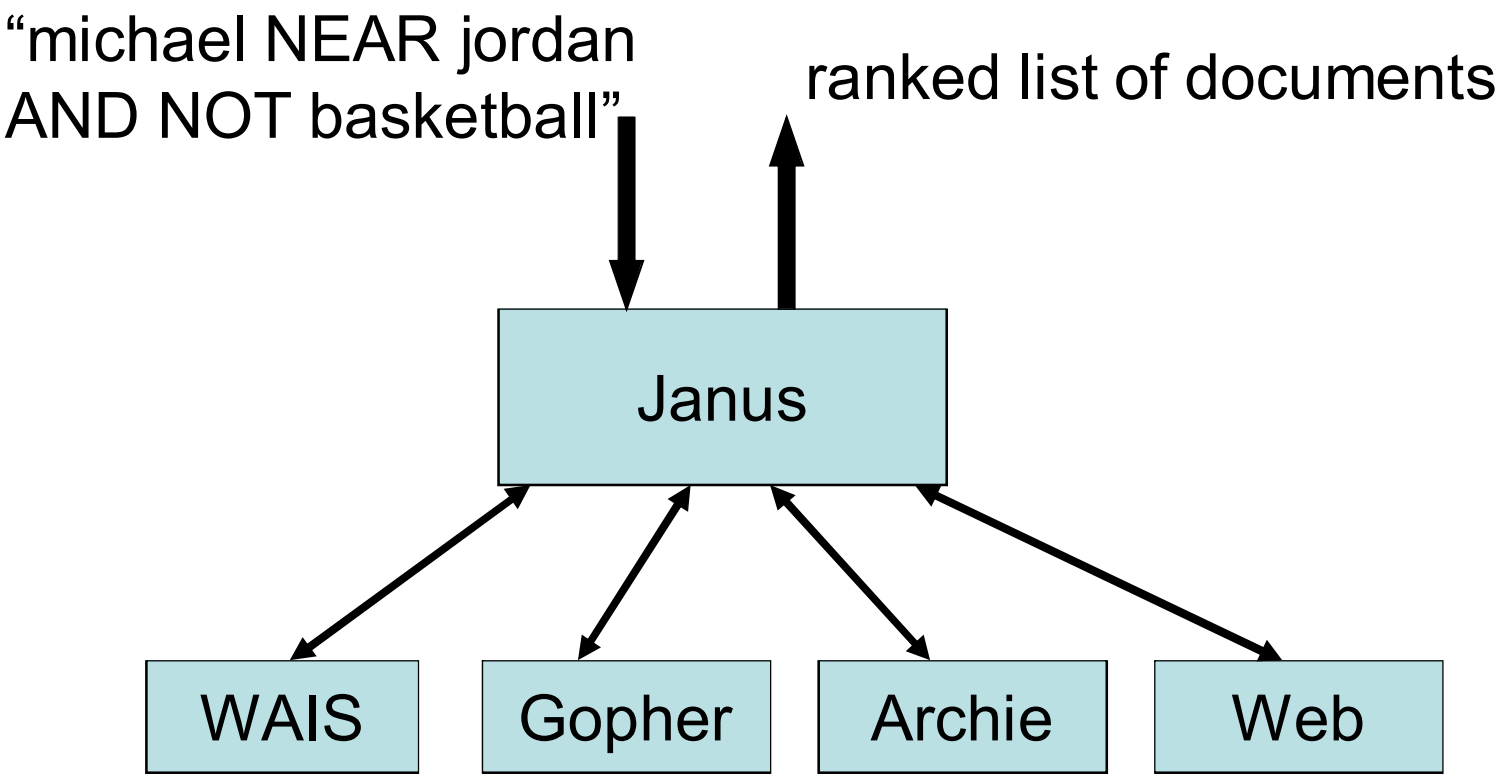
Harnessing the Deep Web

Anand Rajaraman

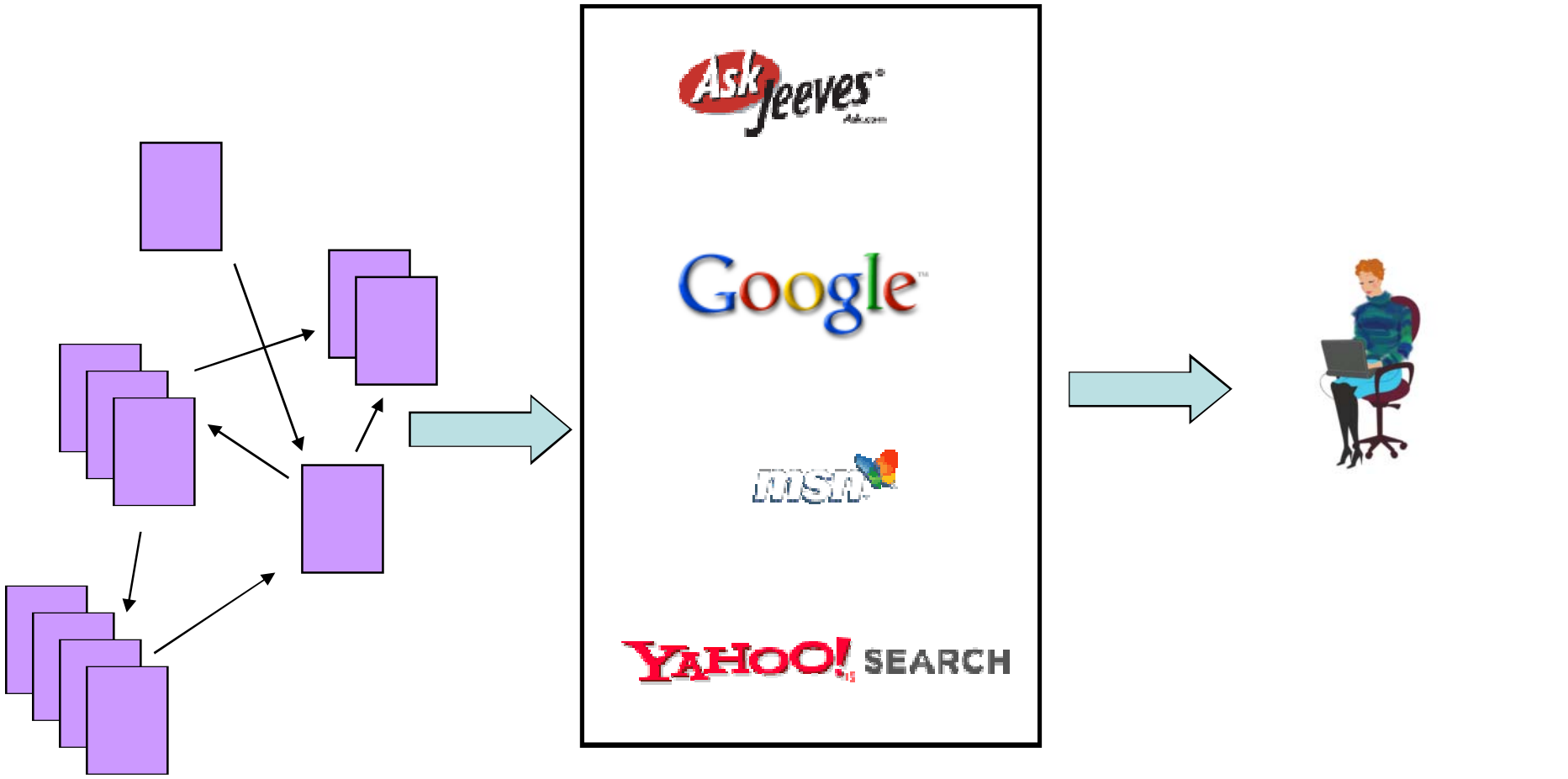
Kosmix

Flashback: PARC, Summer ‘94



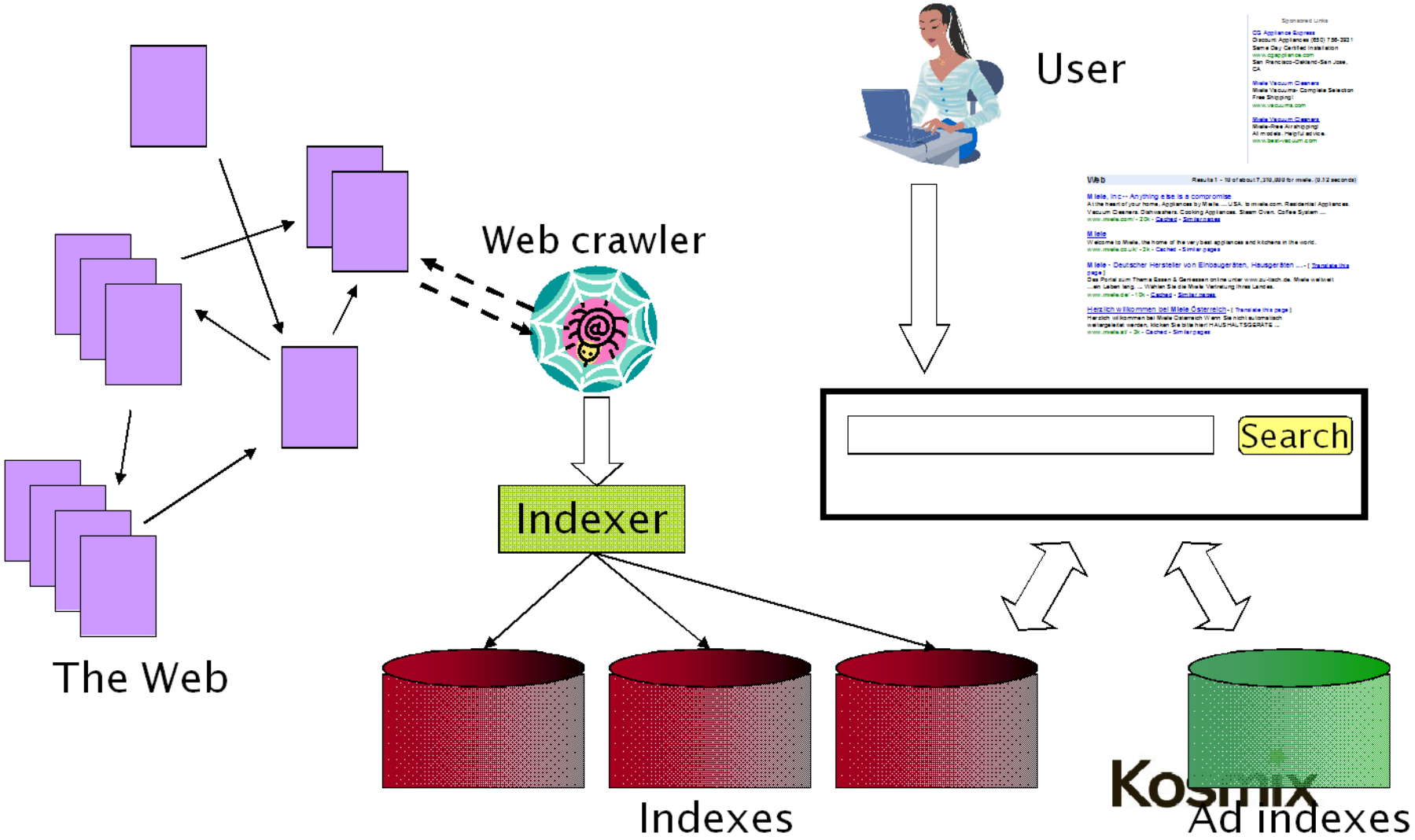
Kosmix

Searching the Web



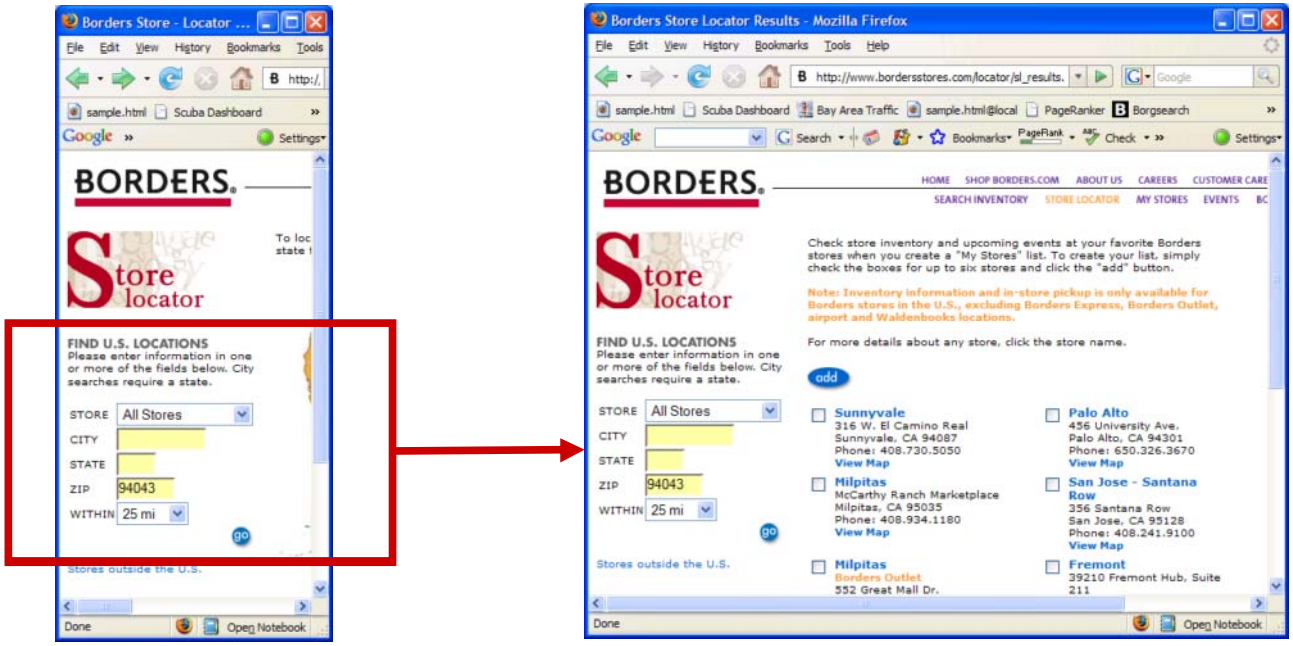
Kosmix

Search Engine 101



What is the Deep Web?

- Content hidden behind HTML forms

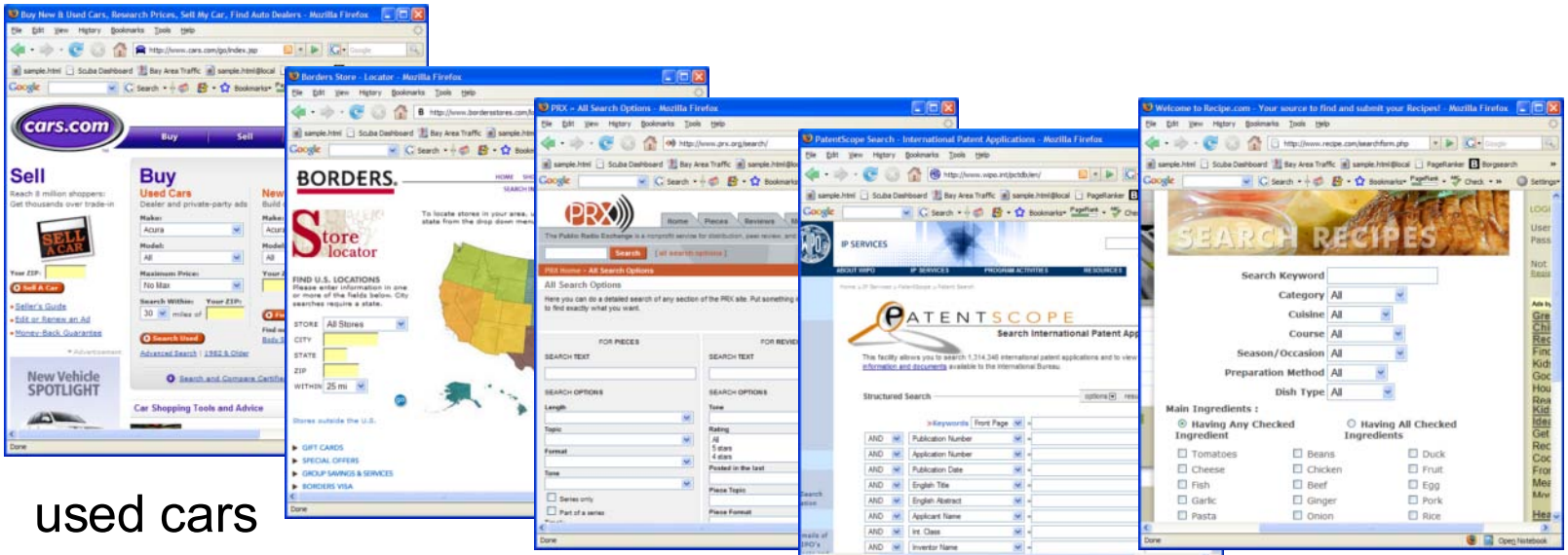


Deep = not accessible through search engine crawlers

Kosmix

The Deep Web

- Major gap in search engine coverage



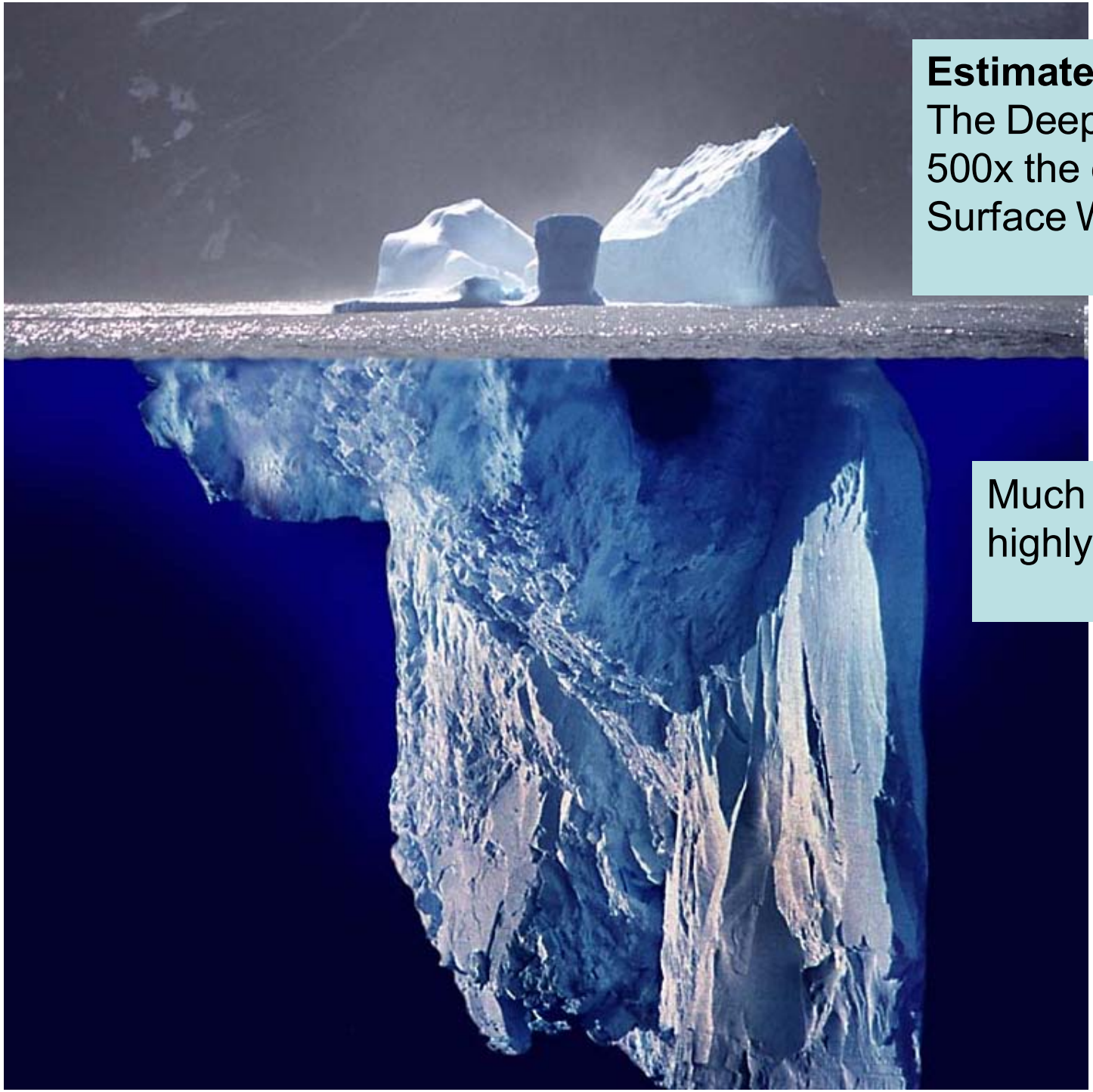
used cars

store locations

radio stations

patents

recipes



Estimate:
The Deep Web contains 500x the data found in the Surface Web

Much of this data is highly structured

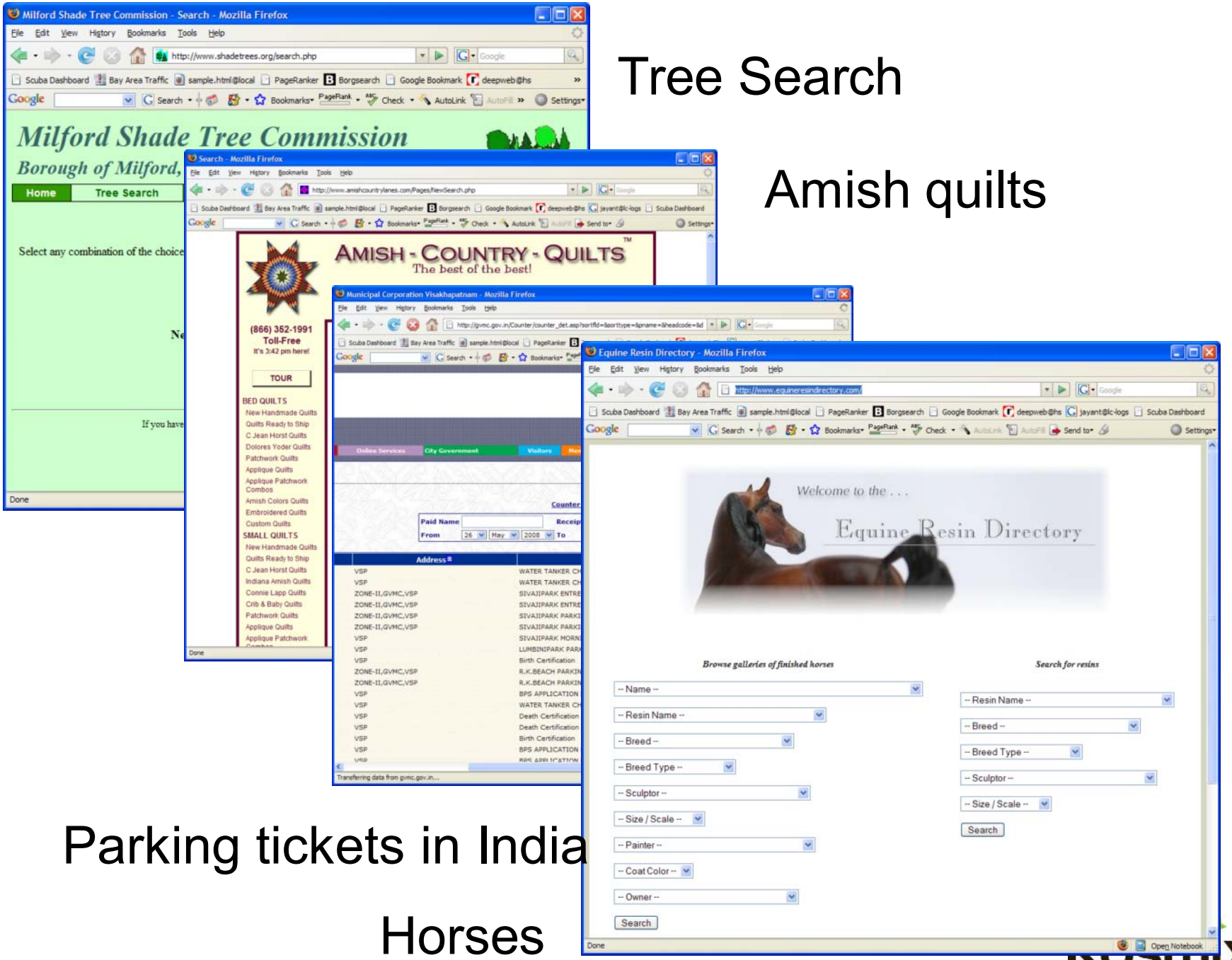
Kosmix

Harnessing the Deep Web

- Two different approaches to deep web:
  - Surfacing content: Google
  - Federated Approach: Kosmix

- Comparisons and future challenges

Kosmix



Tree Search

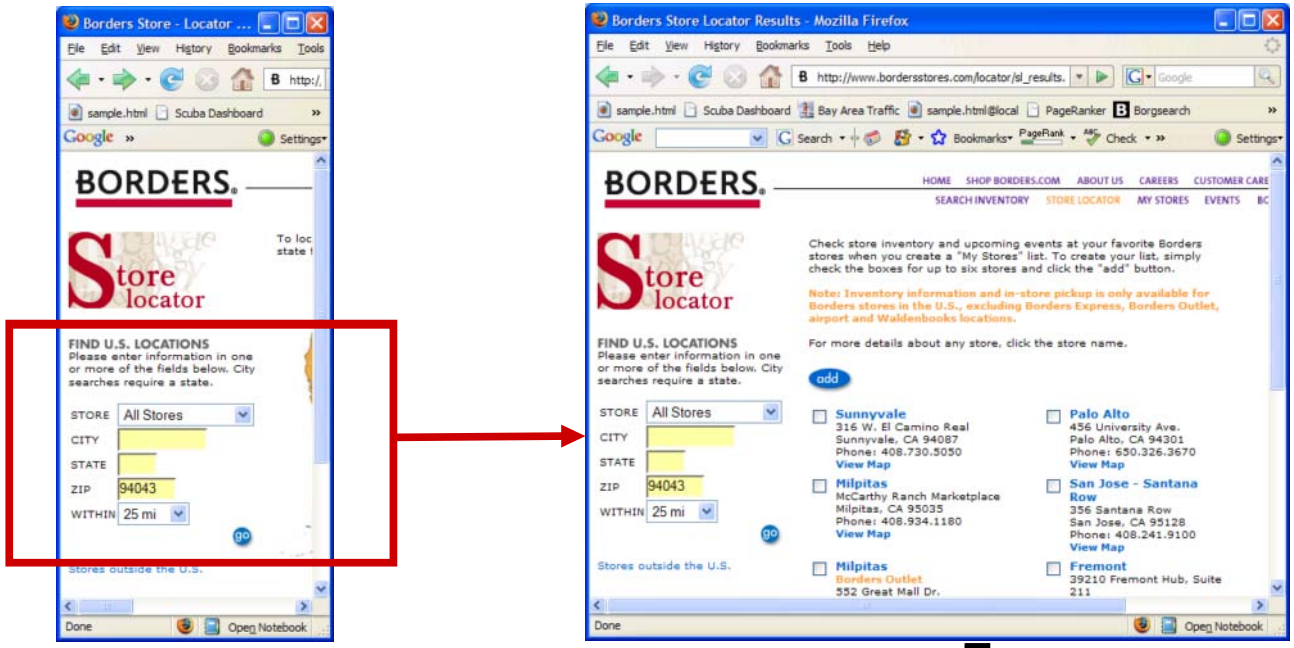
Amish quilts

Parking tickets in India

Horses

Kosmix

Surfacing the Deep Web



Web crawler

Indexer

Index

Kosmix

Surfacing

- At Crawl time
  - Precompute “interesting” form submissions
  - Insert resultimg pages into Google index
- At Query time
  - Nothing!
  - Pages are already in the index

Kosmix

Surfacing Challenges

[Madhavan et al., VLDB 2008]

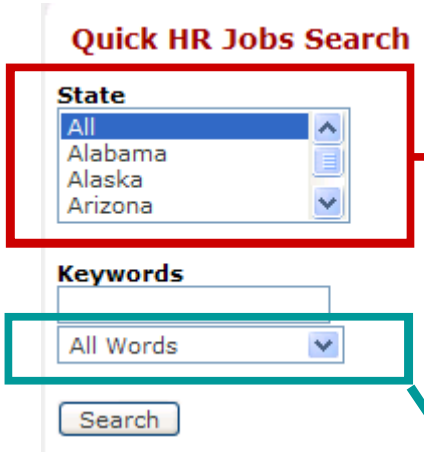
1. Predicting the correct input combinations
   - Generating all possible URLs is wasteful and unnecessary
   - Cars.com has ~500K listings, but 250M possible queries

2. Predicting the appropriate values for text inputs
   - Valid input values are required for retrieving data
   - Ingredients in recipes.com and zipcodes in borderstores.com

3. Don't do anything bad!

Kosmix

Informative Query Templates



Result pages different ➔ informative

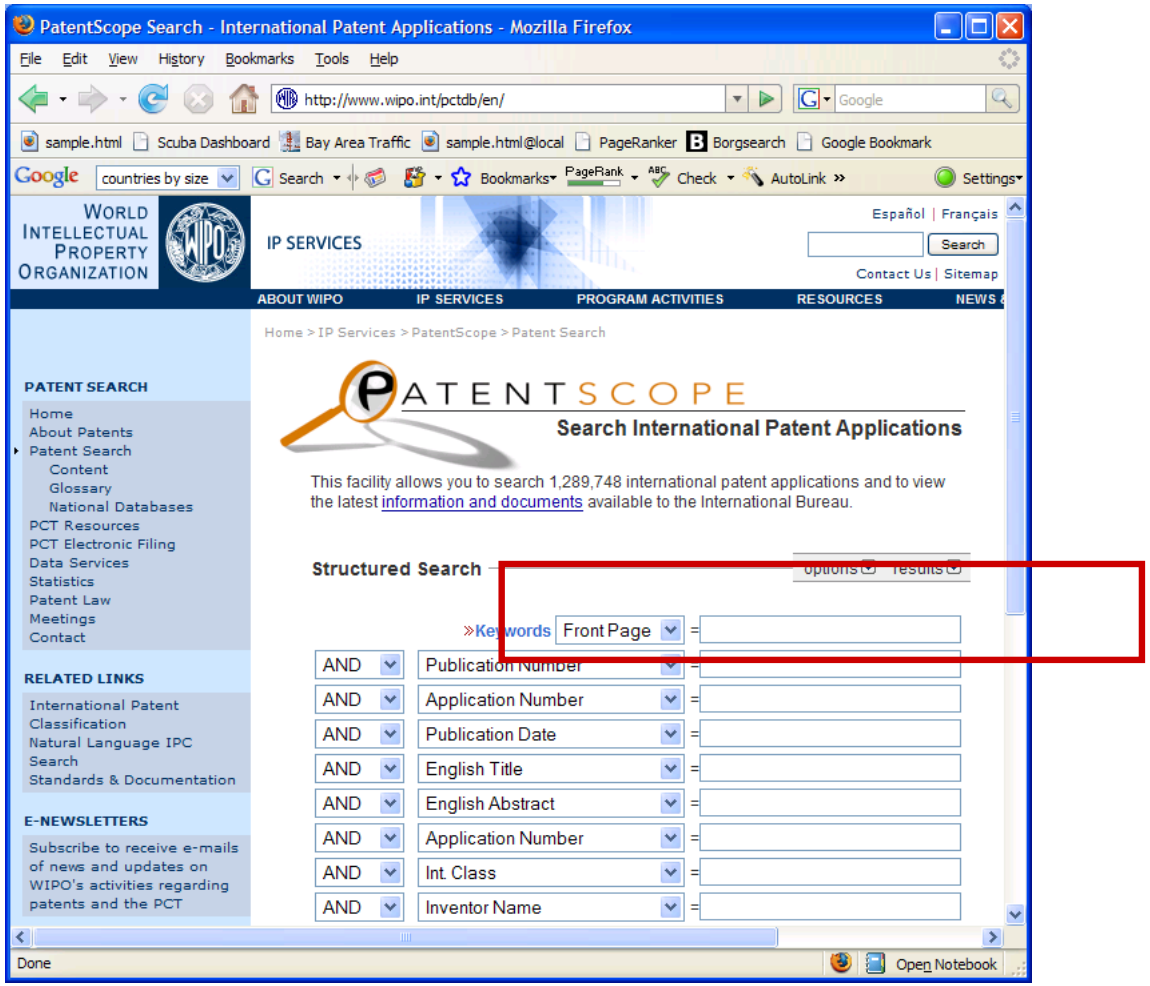
http://jobs.shrm.org/search?**state=All**&kw=&type=All
http://jobs.shrm.org/search?**state=AL**&kw=&type=All
http://jobs.shrm.org/search?**state=AK**&kw=&type=All
…
http://jobs.shrm.org/search?**state=WV**&kw=&type=All

Result pages similar ➔ un-informative

http://jobs.shrm.org/search?state=All&kw=&**type=ALL**
http://jobs.shrm.org/search?state=All&kw=&**type=ANY**
http://jobs.shrm.org/search?state=All&kw=&**type=EXACT**

Kosmix

Example: www.wipo.int



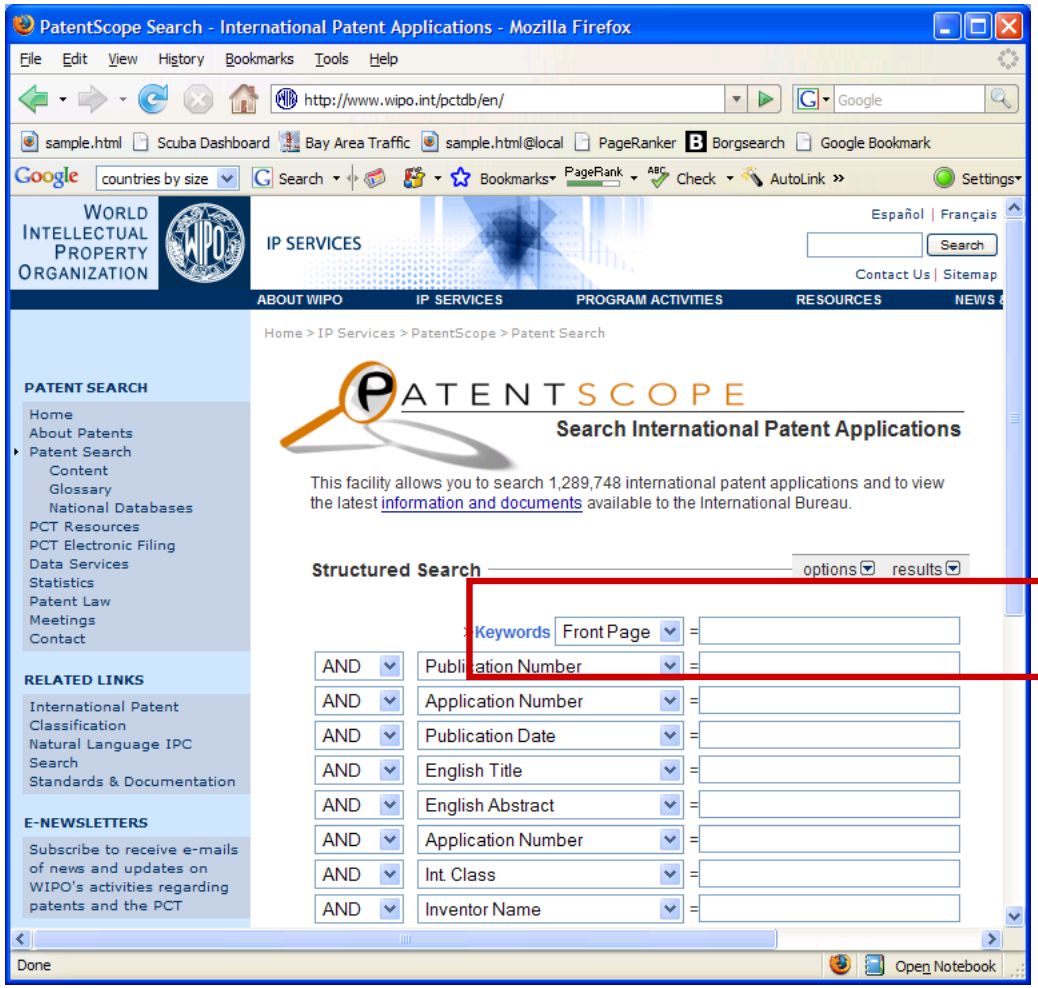
Kosmix

Input values for Generic Search

- Iterative Probing for search boxes
  - Select an initial list of candidate keywords
  - Download pages based on current set of keywords
  - Extract more candidate keywords from result pages
  - Refine the current set of keywords
  - Repeat until no more new candidate keywords
  - Prune list of candidate keywords
- Related Work:
  - Classifying Deep-Web sources [Ipeirotis+]
  - Extracting text documents [Ntoulas+, Barbosa+]

Kosmix

Example: www.wipo.int



Metalworking
Protein
Antibody
Pyrazole
Immobilizer
Vasoconstriction
Phosphinates
Nosepiece
Sandbridge
Viscosity
Carboxydiphenylsulphide
Ozonizer
…

Kosmix

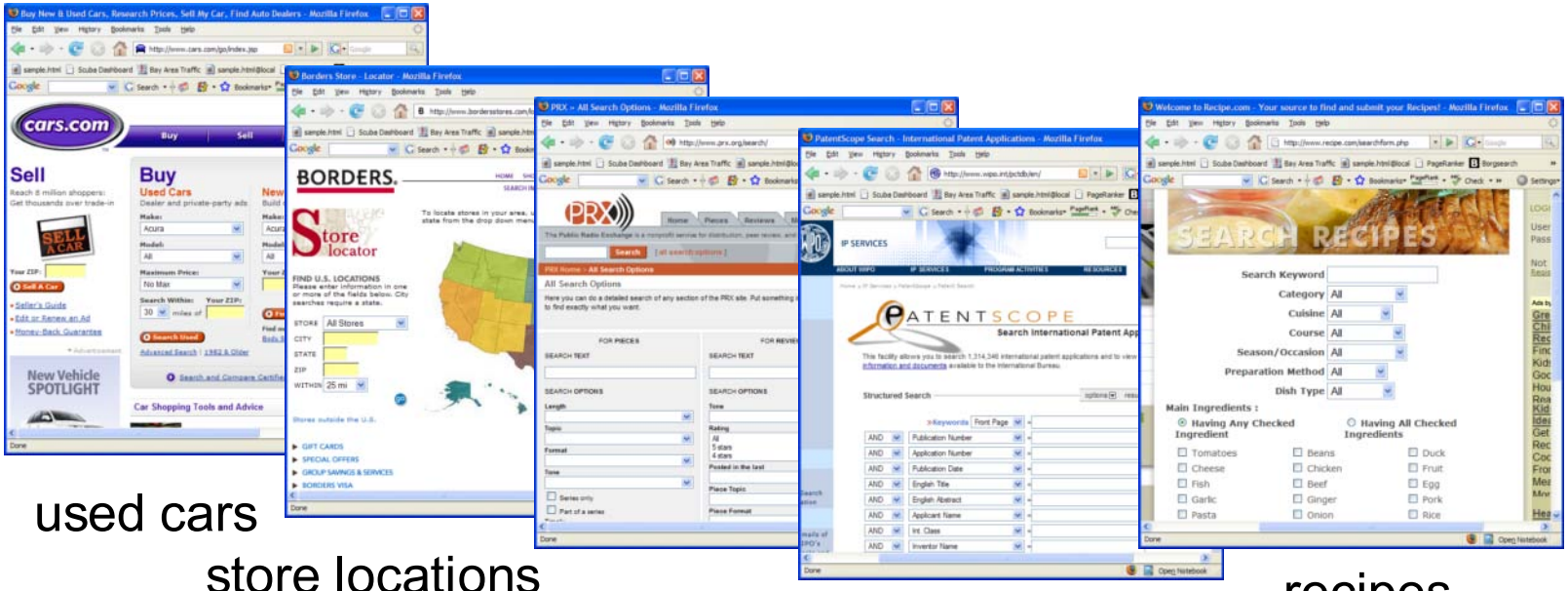
Impact on Query Stream

[Madhavan et al, VLDB 2008]

- Crawled ~3M sites
- 50 languages, hundreds of domains
- 1000 queries per-second get results from the deep web!
- *Impact mostly on the long and heavy tail of queries*

Kosmix

A Closer Look



used cars

store locations

radio stations

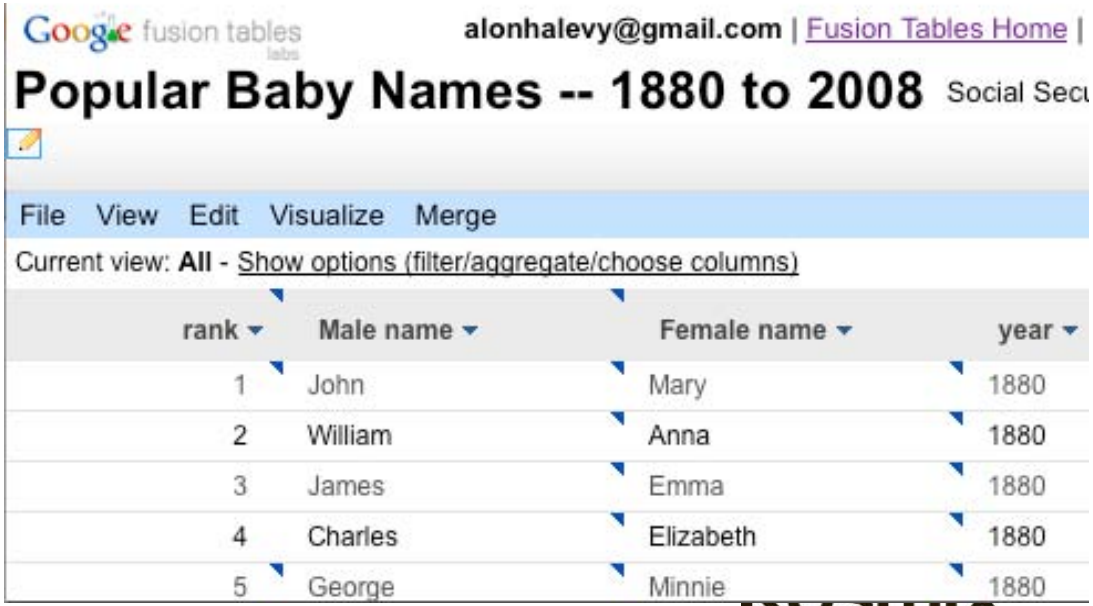
patents

recipes

Kosmix

Structured Data!

databaseFootball.com



A Different Approach

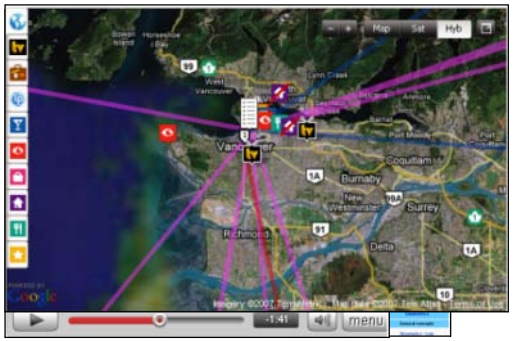
Then

- Explosion of documents

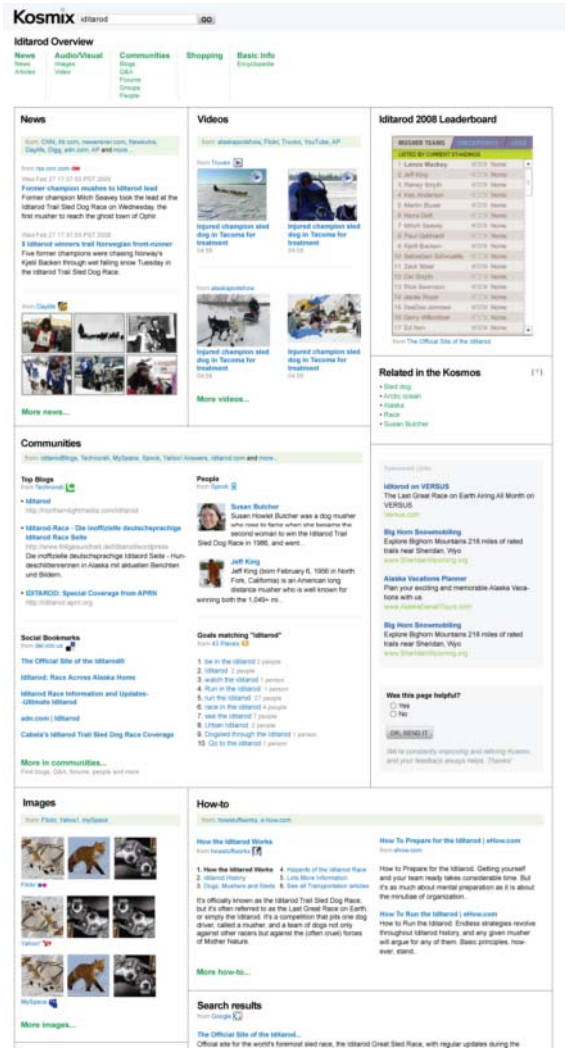
- But the documents were primarily of the same information type
- Search engines surfaced the top 10 relevant documents for the query with a snippet for each document

Now

- Explosion of structured information



Solution
Kosmix Topic Pages



Kosmix: Introducing Topic Pages

Search Results = Find needle in haystack

Caltrain schedule
Bi Rite Creamery San Francisco

Topic Page = Get 360-degree overview of topic

Kate Winslet
Citizen Kane
Cholesterol

Kosmix

Kosmix Demo

Kosmix

Consumer Validation -- RightHealth

- Second most visited health site on the Web per Hitwise
- 6.4 million monthly unique users per comScore

Most Popular Websites ranked by Visits

Month of September, 2009

Health and Medical - Information

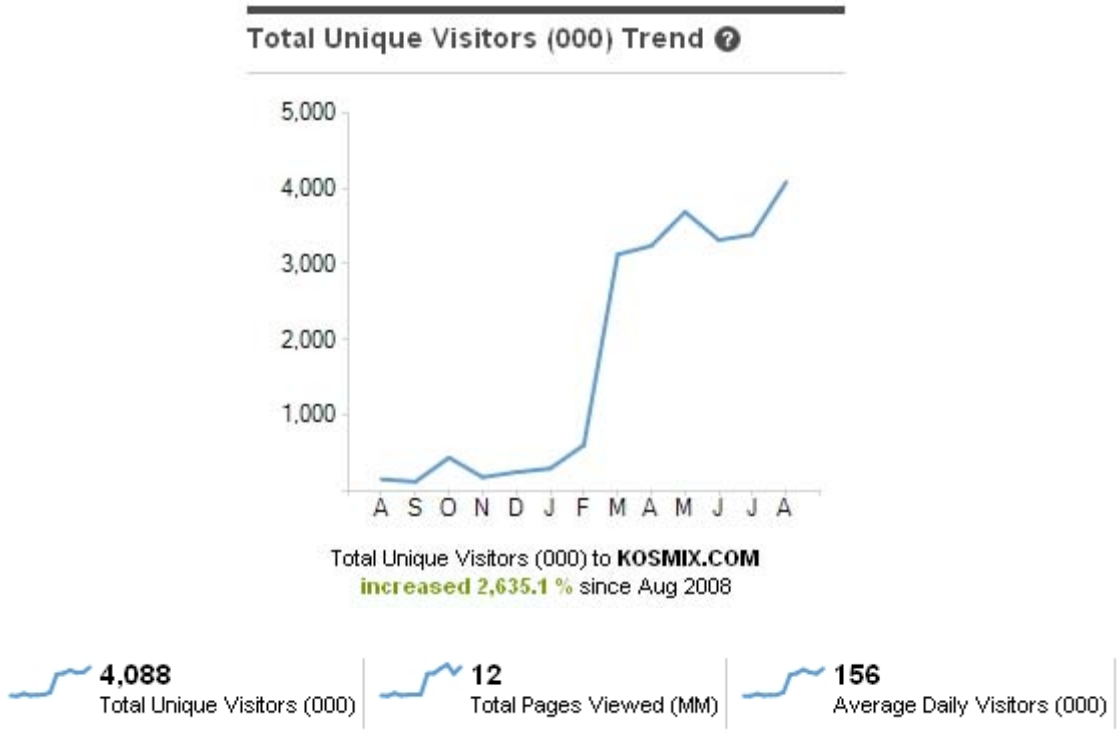
	Websites (2,737 returned)	Visits ▼
1	WebMD	10.20%
2	Righthealth.com	5.67%
▲ 3	MedicineNet	3.48%
▼ 4	MayoClinic.com	3.48%
▲ 5	Drugs.com	3.33%
▼ 6	MSN Health	3.23%
▲ 7	AOL Health	2.58%
▼ 8	Yahoo! Health	2.05%
▲ 9	QualityHealth.com	1.66%
▼ 10	MedlinePlus	1.66%

hitwise
An Experian company

Kosmix

Consumer Validation -- Kosmix

Kosmix.com was launched in Dec 2008, has already grown to 4.1M uniques*



Kosmix

* Source: comScore

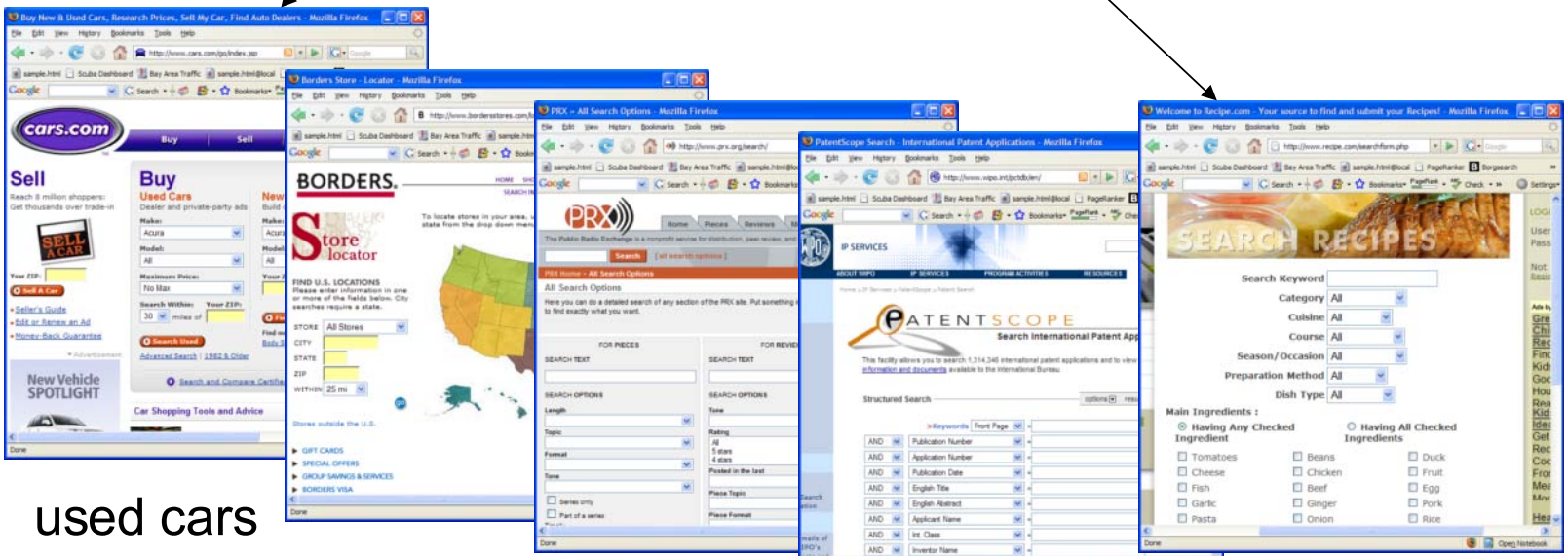
Kosmix RightHealth

- 10 million unique visitors/month
- 35 million searches/month
- Rapid growth over the past few months

Federated Approach

toyota prius

pumpkin pie



used cars

store locations

radio stations

patents

recipes

Kosmix

Federated approach to the Deep Web

- It is hard to separate structured data from its applications
  - 15 years of data integration research & experience
- Let domain specialists create and maintain structured data
  - And expose it as web services
- Route queries to specialist services

Kosmix

Trends Favor the Federated Approach

- Social Media
- Real-Time Information
- Specialized search engines
- Innovative visualizations
- Business Model issues
- Algorithmic Content
- Availability of APIs

Deep Web Driver: Social Media

- Content volume grows rapidly
- Access controls can prevent indexing
- Opportunities for personalization

Kosmix

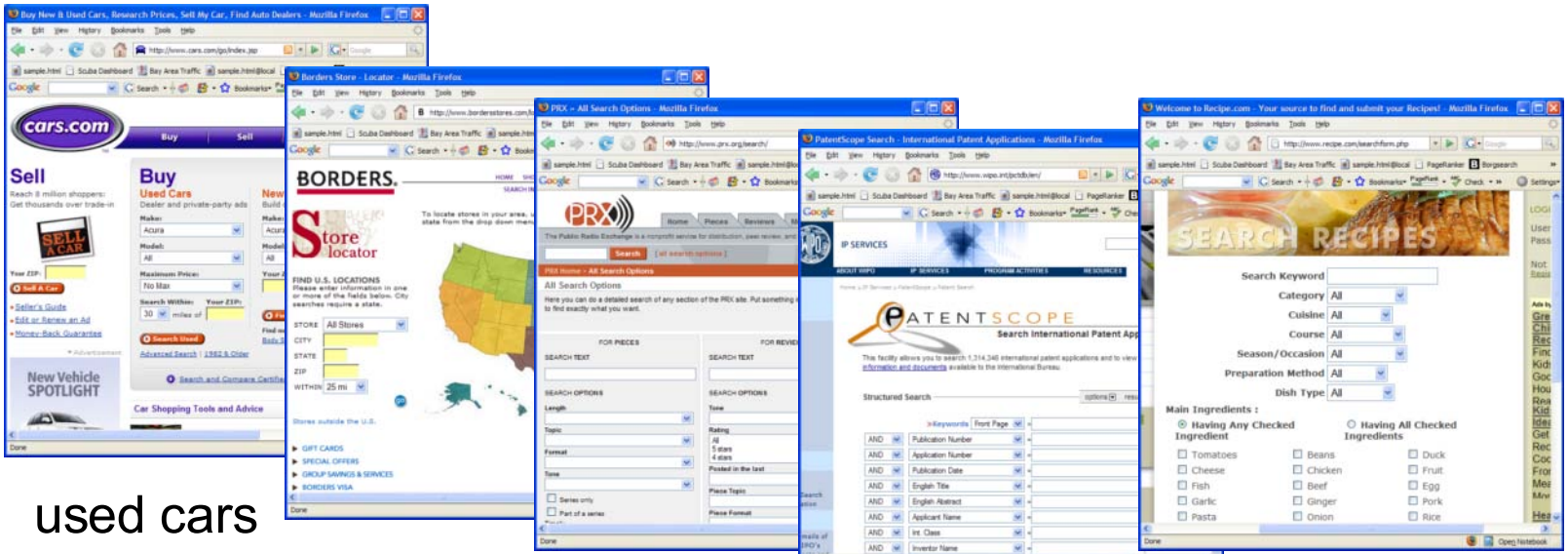
Deep Web Driver: Real-Time

- Earthquake in China (2008)
- US Airways 1549 Hudson landing (2009)
- Iran elections (2009)

Kosmix

Specialized Search Engines

It's a shame to take all this richness and compress it into 10 results links!



used cars

store locations

radio stations

patents

recipes

Kosmix

Deep Web Driver: Specialized Search Engines

- Vertical search engines exist for several domains
  - Travel, real estate, autos, …
- They use specialized ranking functions
  - Very different from standard web search
- And innovative visualizations of results
  - Going beyond 10 links per page

Kosmix

Challenges for the federated approach

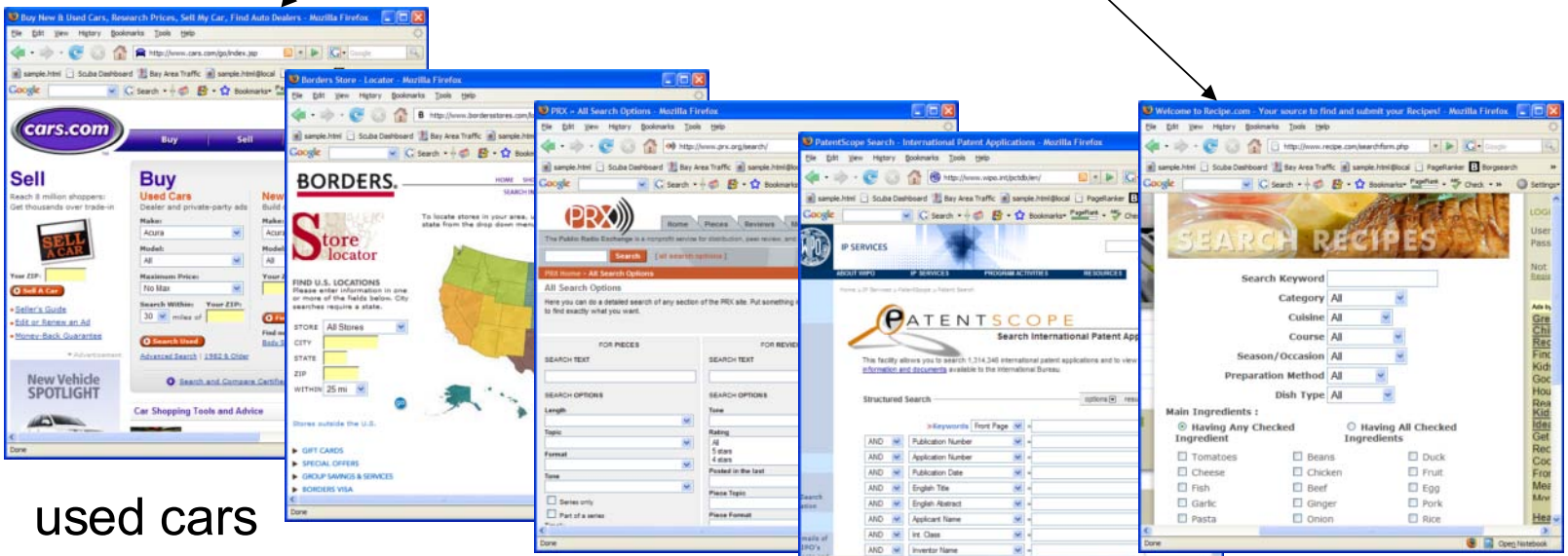
- Data source integration
- Data source selection 
- Query transformation
- Results Layout
- Performance

Kosmix

Data Source Selection

toyota prius

pumpkin pie



used cars

store locations

radio stations

patents

recipes

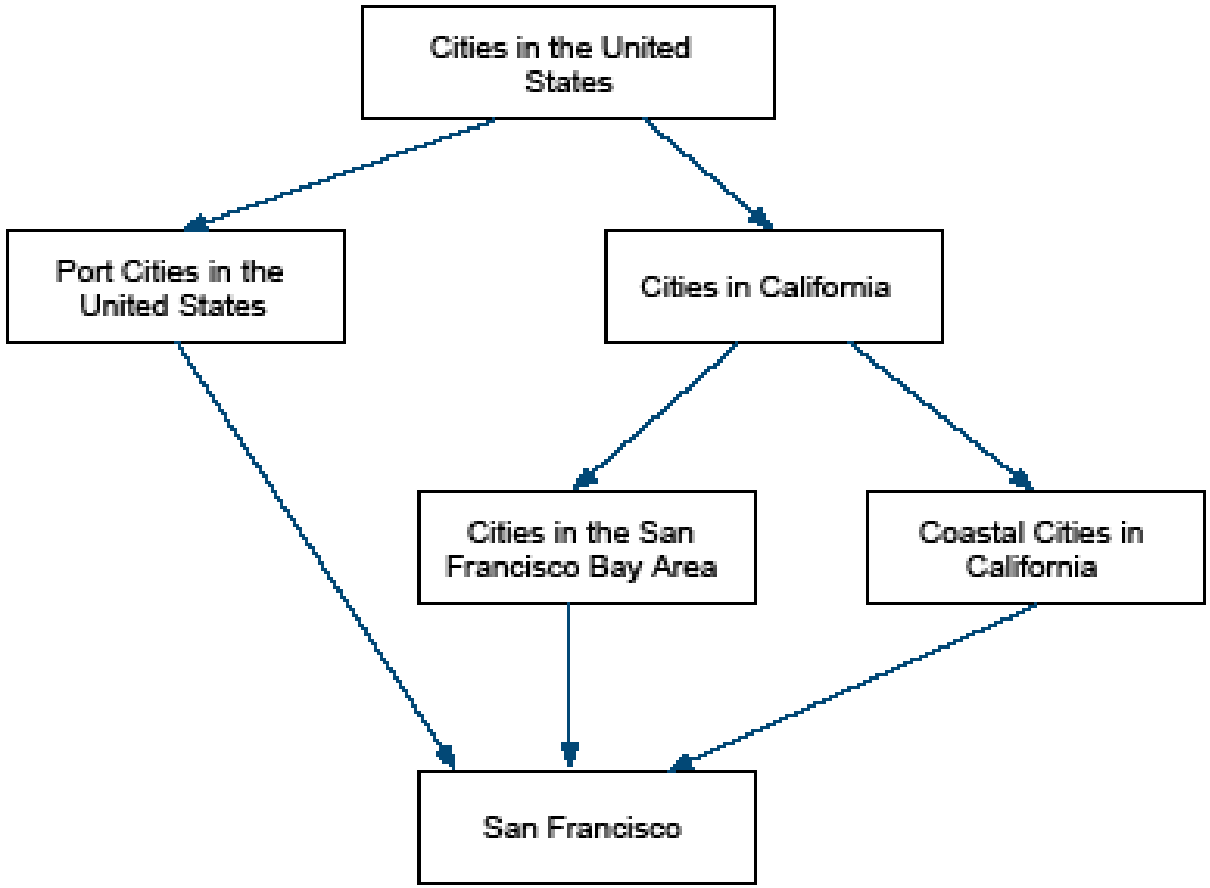
Kosmix

Data Source Selection

- The system knows about 1000s of data sources
- Cannot access each source for every query
  - Unacceptable load on service providers
- Need to select a small set of sources for each query
  - Without access to the source data

Kosmix

The Taxonomy



Kosmix

The Taxonomy

- Millions of nodes arranged in a directed acyclic graph (DAG)
  - Captures ISA relationships
  - And many other domain-specific relationships
- Created over 3 years using a combination of algorithms and human curation
  - Mined from dozens of sources, including Wikipedia, DMOZ, and deep vertical sources
  - Refreshed on a daily basis to capture new entities and relationships

Kosmix

Tagging Data Sources

- Data sources are “tagged” using taxonomy nodes
- A data source can be tagged to several nodes in taxonomy, at different levels

Last.fm → Music
Epicurious → Recipes (under Food & Wine)
MetroLyrics → Lyrics (under Music)
TripAdvisor → Hotels (under Travel)

Kosmix

Query Categorization

- Kosmix Categorization Service (KCS)
- Given a query, KCS identifies the taxonomy nodes related to the query

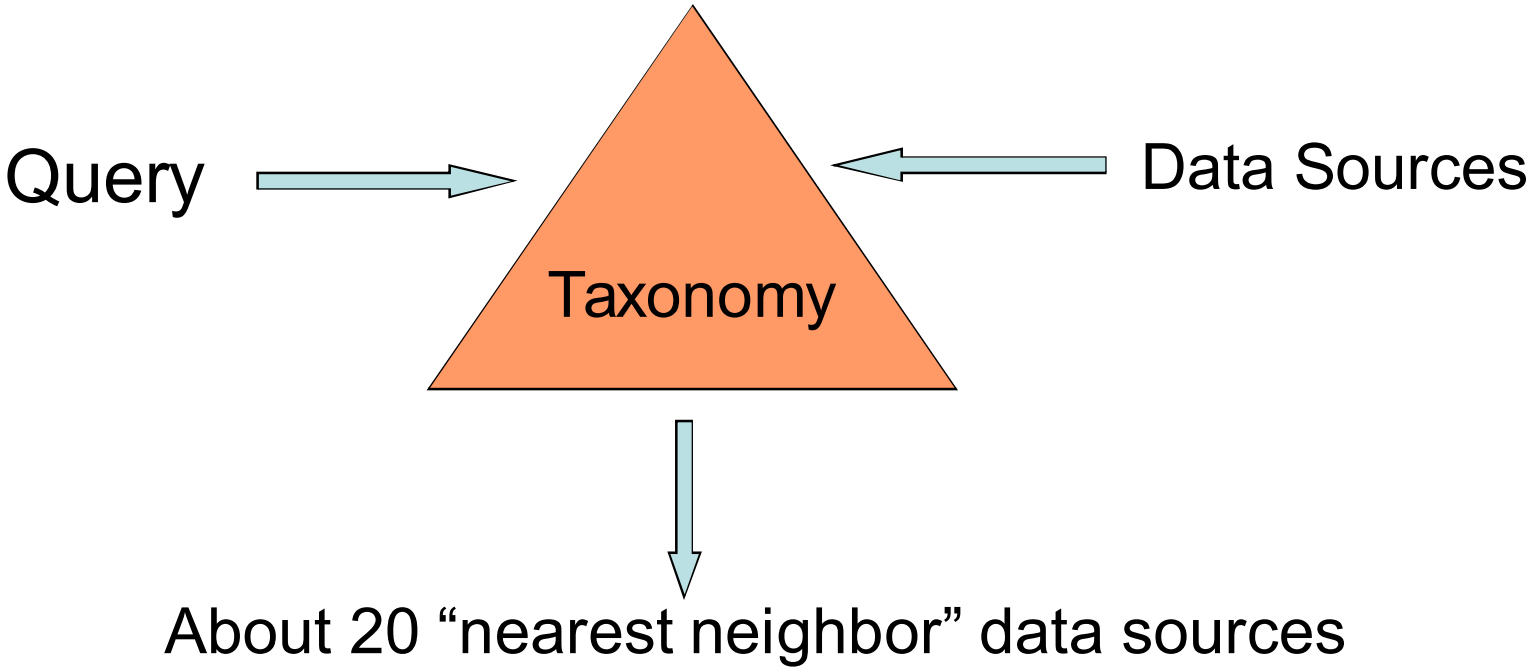
Taylor Swift → Music
Pinot Noir → Wine, Viticulture (+ many others)
Eiffel Tower → Monuments, Tourist Attractions, History of France (+ many others)

Kosmix

KCS output for Pinot Noir

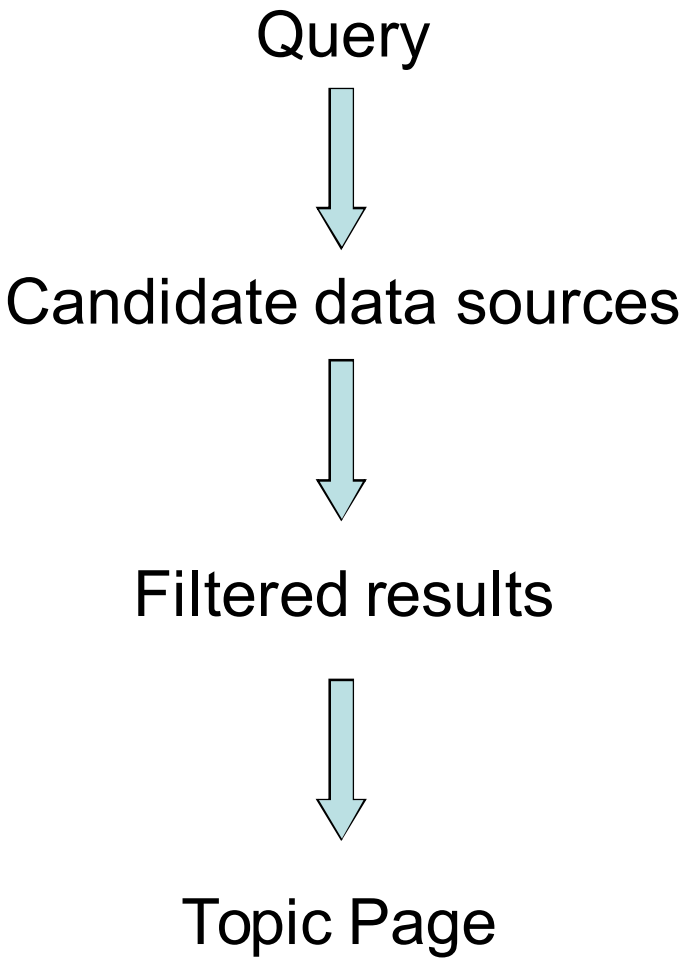
Kosmix

Data Source Selection



Kosmix

From query to topic page



Categorization

Query transformation
Data source access
Filtering

Ranking and Layout

Kosmix

Read VLDB 2009 paper for:

- Hybrid approach
  - Combines Indexing and Federation
- Taxonomy
- Categorization
- Query transformation
- Disambiguation
- Filtering
- Performance

Kosmix

Comparing the approaches

- Surfacing is great for:
  - Simplicity
  - Coverage
  - Performance
- Federated approach better for:
  - Real-time information
  - Aggregating specialized search engines
  - Social media

Kosmix

Challenges

- Can we combine the advantages of the surfacing and federated approaches?
- Is there a way to scale the federated approach rapidly?
- How can we surface data with more structure?

Kosmix

References

- Jayant Madhavan et al. *Google's Deep-Web Crawl.* VLDB 2008.

- Anand Rajaraman. *Kosmix: High Performance Topic Exploration using the Deep Web.* VLDB 2009.

Kosmix