

Data Mining and Information Retrieval

Introduction to Web Mining

What is Web Mining?

Discovering useful information from the World-Wide Web and its usage patterns.

Web Mining vs. Data Mining

- **Structure (or lack of it)**

- Textual information and linkage structure

- **Scale**

- Data generated per day is comparable to largest conventional “data warehouses”

- **Speed**

- Often need to react to evolving usage patterns in real-time (e.g., merchandising)

Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

Size of the Web

■ Number of pages

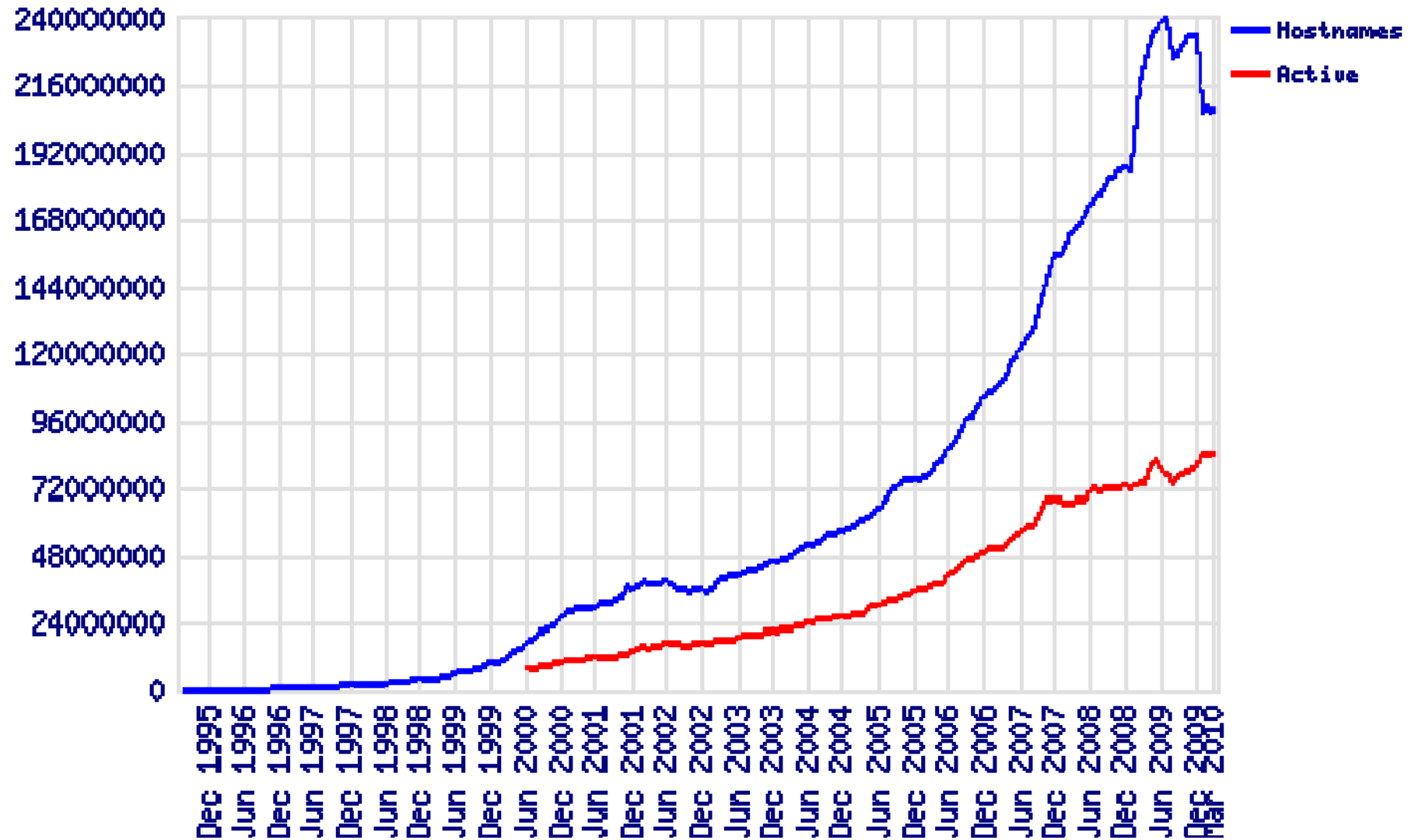
- Technically, infinite
- Much duplication (30-40%)
- Best estimate of “unique” static HTML pages comes from search engine claims
 - Google = 8 billion(?), Yahoo = 20 billion

■ Number of web sites

- Netcraft survey says **206,675,938** sites (March 2010)

(http://news.netcraft.com/archives/web_server_survey.html)

Netcraft Survey



http://news.netcraft.com/archives/web_server_survey.html

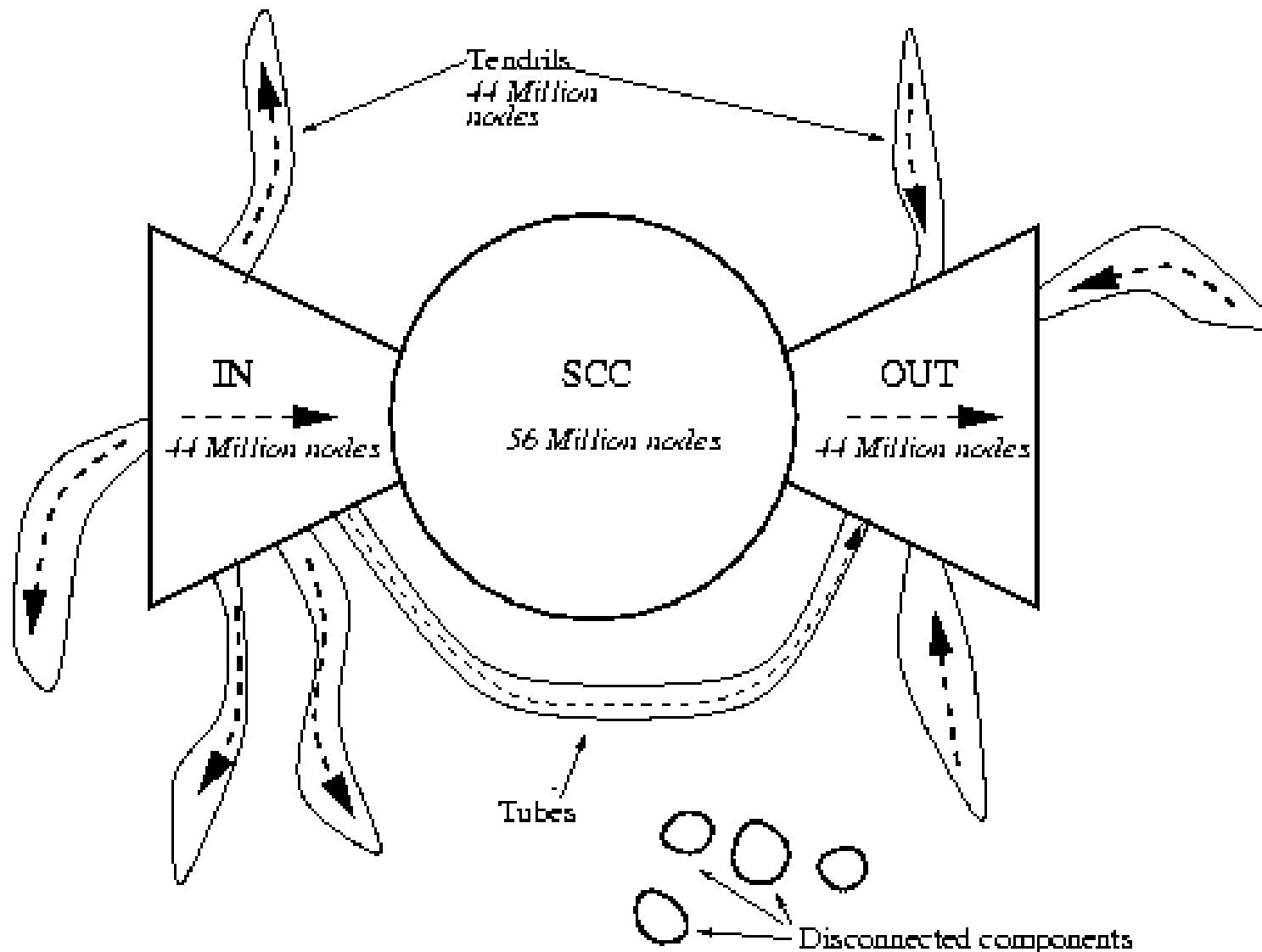
The Web as a Graph

- Pages = nodes, hyperlinks = edges
 - Ignore content
 - Directed graph
- High linkage
 - 8-10 links/page on average
 - Power-law degree distribution

Structure of Web Graph

- Let's take a closer look at structure
 - Broder et al (2000) studied a crawl of 200M pages and other smaller crawls
 - Bow-tie structure
 - Not a "small world"

Bow-tie Structure



Source: Broder et al, 2000

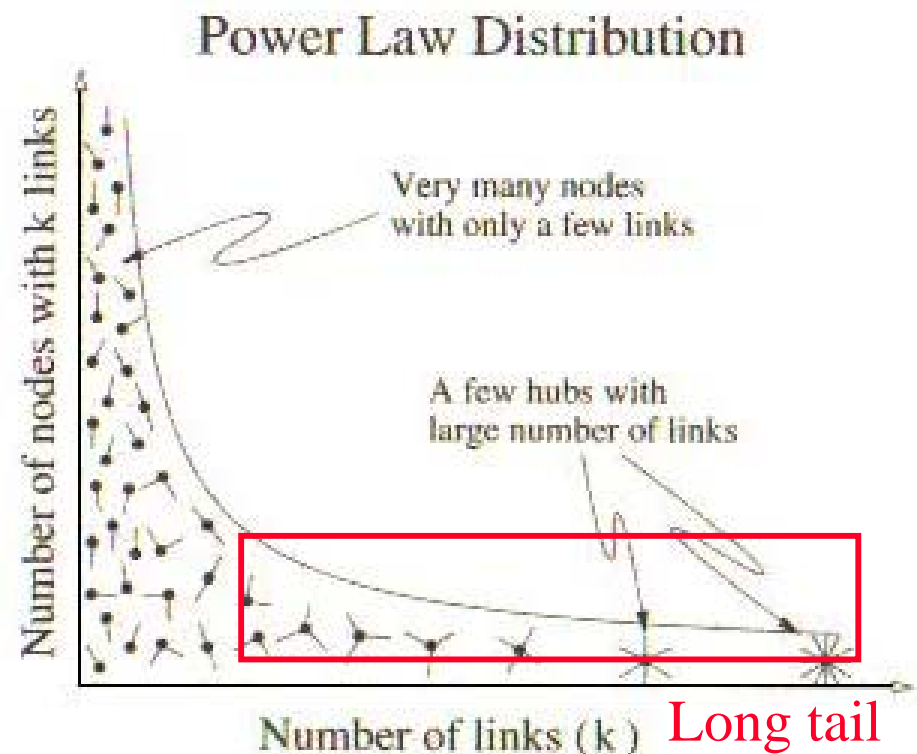
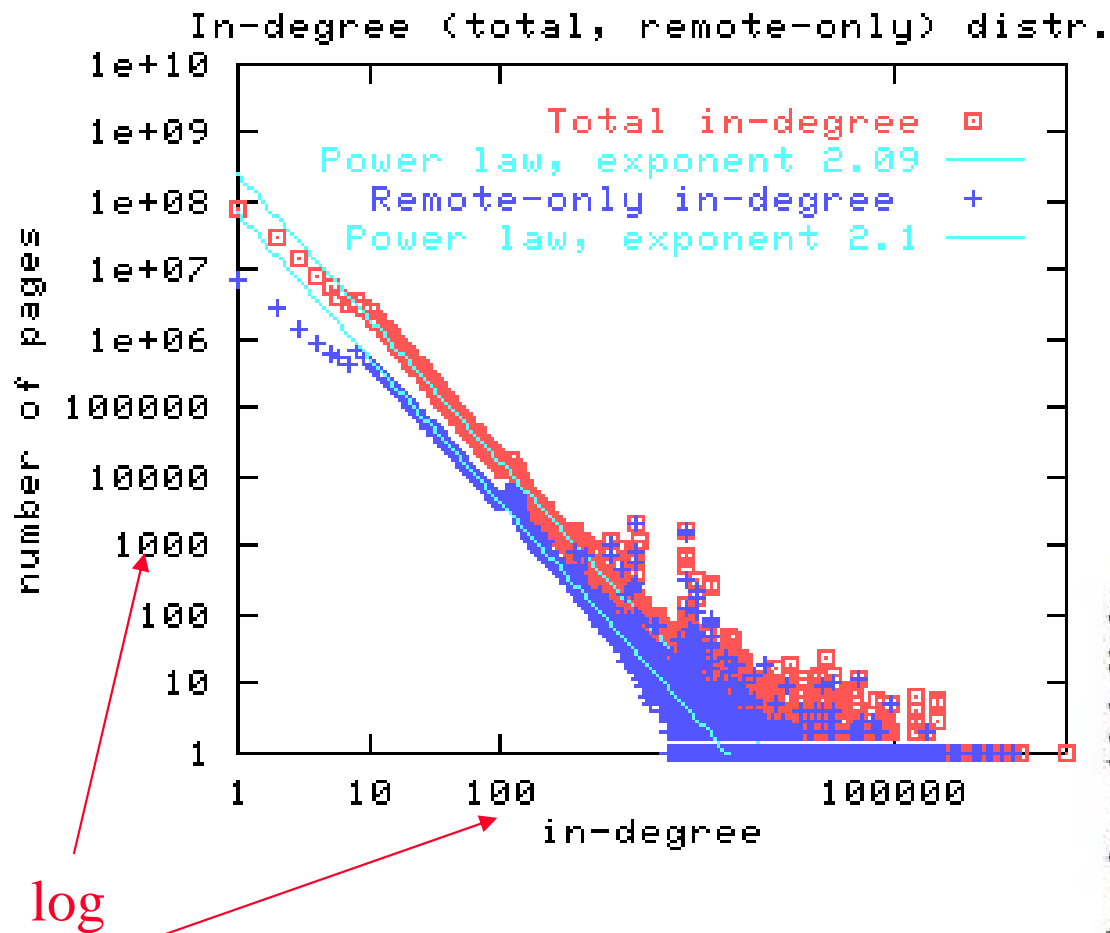
What can the graph tell us?

- Distinguish “important” pages from unimportant ones
 - Page rank
- Discover communities of related pages
 - Hubs and Authorities
- Detect web spam
 - Trust rank

Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

Power-law degree distribution



Source: Broder et al, 2000

Power-laws galore

- Structure
 - In-degrees
 - Out-degrees
 - Number of pages per site
- Usage patterns
 - Number of visitors
 - Popularity
- And much more...

Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- **Structured data extraction**
- Web advertising
- Systems Issues

Extracting Structured Data

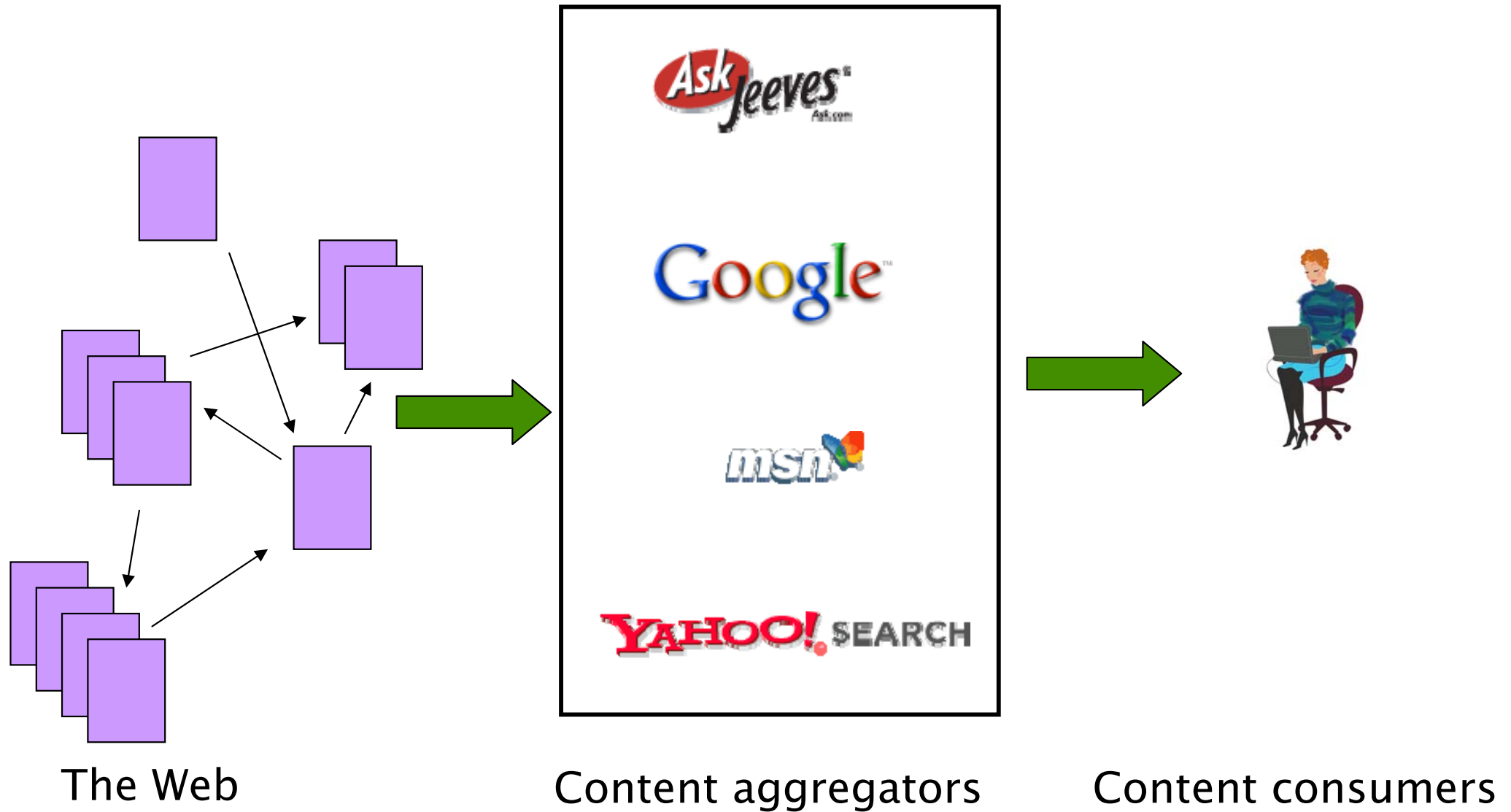
The screenshot shows the SimplyHired website interface. At the top, there are navigation links for 'search', 'browse', and 'suggestions'. The main search area includes the 'simply|hired' logo, a 'keywords' input field containing 'software engineer', a 'location' input field containing 'Mountain View, CA', and a 'search' button. Below the search bar, there is a 'sorted by:' section with options for 'best match first' and 'newest job first'. The first search result is for a 'Software Implementation Consultant / Engineer' position at 'Kaidara Software (Los Altos, CA)'. The job description states that Kaidara Software provides software solutions to reduce the cost of delivering superior customer service and is looking for a Software Implementation Consultant / Engineer. The job was posted '2 days and 3 hours ago' from 'Monster'. Below the job description are four buttons: 'who do i know?™', 'research salary', 'send-to-friend', and 'apply now'. The second search result is for a 'Software Engineer' position at 'ESP Enviromental Software (Mountain View, CA)'. The job description mentions server-side data updates and various data manipulation tools. The job was posted '2 days and 19 hours ago' from 'Dice'.

<http://www.simplyhired.com>

Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- **Web advertising**
- Systems Issues

Searching the Web



Ads vs. search results

Web

Results 1 - 10 of about 2,230,000 for geico. (0.04 sec)

[GEICO Car Insurance. Get an auto insurance quote and save today ...](#)

GEICO auto insurance, online car insurance quote, motorcycle insurance quote, online insurance sales and service from a leading insurance company.

[www.geico.com/](#) - 21k - Sep 22, 2005 - [Cached](#) - [Similar pages](#)

[Auto Insurance](#) - [Buy Auto Insurance](#)

[Contact Us](#) - [Make a Payment](#)

[More results from www.geico.com »](#)

[Geico, Google Settle Trademark Dispute](#)

The case was resolved out of court, so advertisers are still left without legal guidance on use of trademarks within ads or as keywords.

[www.clickz.com/news/article.php/3547356](#) - 44k - [Cached](#) - [Similar pages](#)

[Google and GEICO settle AdWords dispute | The Register](#)

Google and car insurance firm **GEICO** have settled a trade mark dispute over ... Car insurance firm **GEICO** sued both Google and Yahoo! subsidiary Overture in ...

[www.theregister.co.uk/2005/09/09/google_geico_settlement/](#) - 21k - [Cached](#) - [Similar pages](#)

[GEICO v. Google](#)

... involving a lawsuit filed by Government Employees Insurance Company (**GEICO**). **GEICO** has filed suit against two major Internet search engine operators, ...

[www.consumeraffairs.com/news04/geico_google.html](#) - 19k - [Cached](#) - [Similar pages](#)

Sponsored Links

[Great Car Insurance Rates](#)

Simplify Buying Insurance at Safeco
See Your Rate with an Instant Quote
[www.Safeco.com](#)

[Free Insurance Quotes](#)

Fill out one simple form to get multiple quotes from local agents.
[www.HometownQuotes.com](#)

[5 Free Quotes. 1 Form.](#)

Get 5 Free Quotes In Minutes!
You Have Nothing To Lose. It's Free
[sayyessoftware.com/Insurance](#)
Missouri

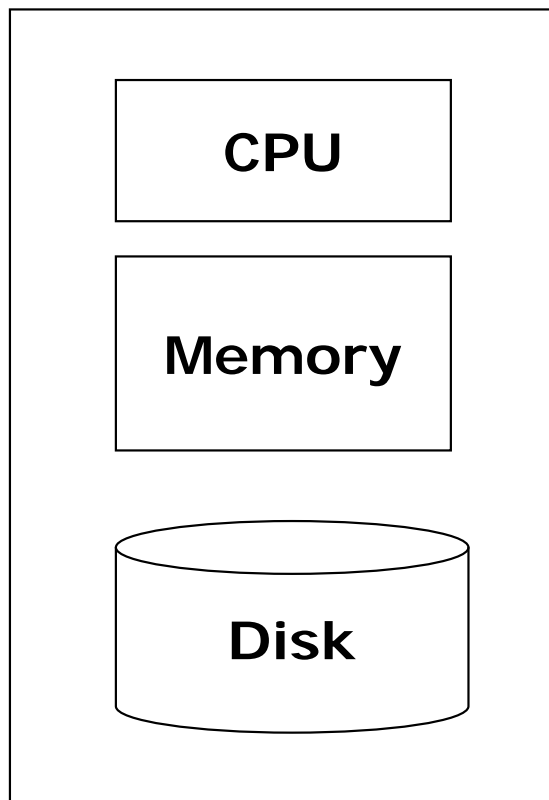
Ads vs. search results

- Search advertising is the revenue model
 - Multi-billion-dollar industry
 - Advertisers pay for clicks on their ads
- Interesting problems
 - What ads to show for a search?
 - If I'm an advertiser, which search terms should I bid on and how much to bid?

Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- **Systems Issues**

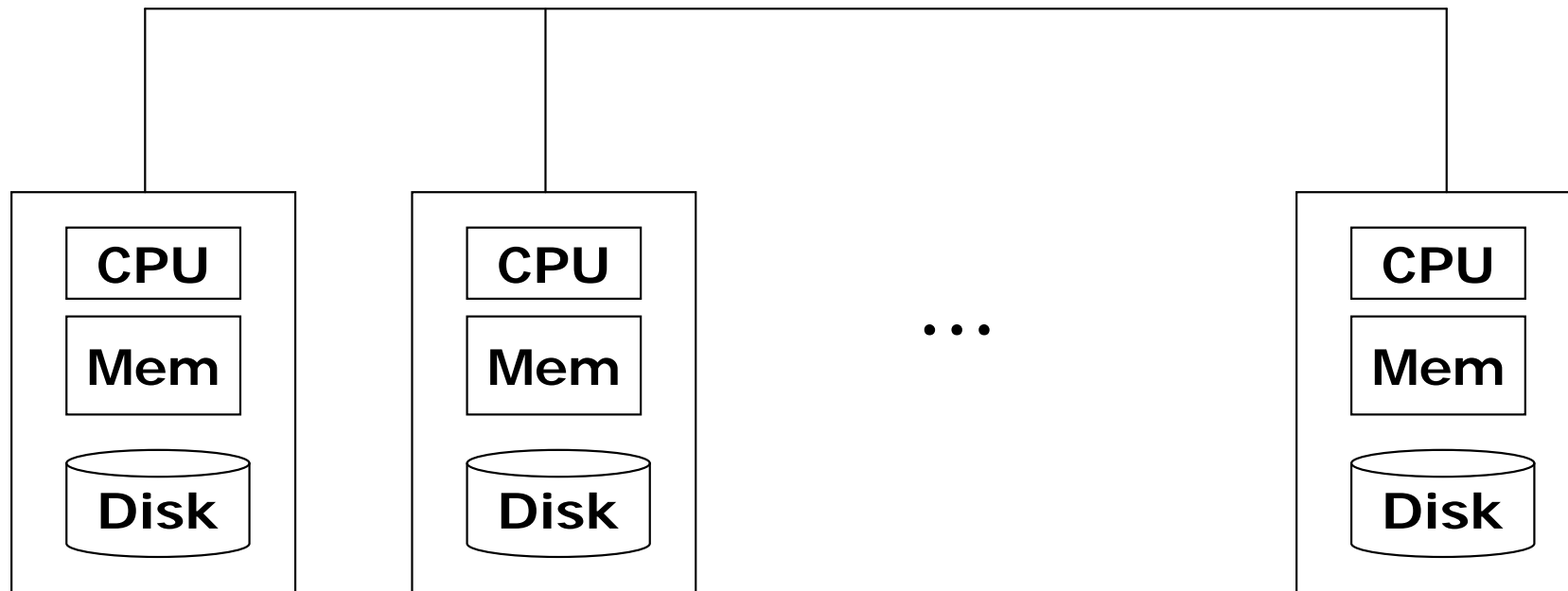
Systems architecture



Machine Learning, Statistics

"Classical" Data Mining

Very Large-Scale Data Mining



Cluster of commodity nodes

Systems Issues

- Web data sets can be very large
 - Tens to hundreds of terabytes
- Cannot mine on a single server!
 - Need large farms of servers
- How to organize hardware/software to mine multi-terabyte data sets