

CS 345  
Data Mining  
Lecture 1

---

## Introduction to Web Mining

# What is Web Mining?

---

- Discovering useful information from the World-Wide Web and its usage patterns
  - Applications
    - Web search e.g., Google, Yahoo,...
    - Vertical Search e.g., FatLens, Become,...
    - Recommendations e.g., Amazon.com
    - Advertising e.g., Google, Yahoo
    - Web site design e.g., landing page optimization
-

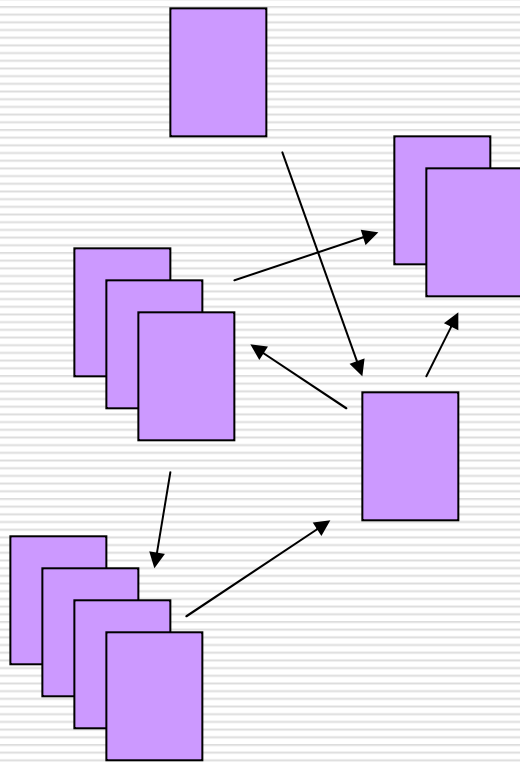
# How does it differ from “classical” Data Mining?

---

- The web is not a relation
    - Textual information and linkage structure
  - Usage data is huge and growing rapidly
    - Google’s usage logs are bigger than their web crawl
    - Data generated per day is comparable to largest conventional data warehouses
  - Ability to react in real-time to usage patterns
    - No human in the loop
-

# The World-Wide Web

---



The Web

- ❑ Huge
  - ❑ Distributed content creation, linking (no coordination)
  - ❑ Structured databases, unstructured text, semistructured
  - ❑ Content includes truth, lies, obsolete information, contradictions, ...
  - ❑ Our modern-day Library of Alexandria
-

# Size of the Web

---

## Number of pages

### ■ Technically, infinite

- Because of dynamically generated content

- Lots of duplication (30-40%)

### ■ Best estimate of “unique” static HTML pages comes from search engine claims

- Google = 8 billion, Yahoo = 20 billion

- Lots of marketing hype

## Number of unique web sites

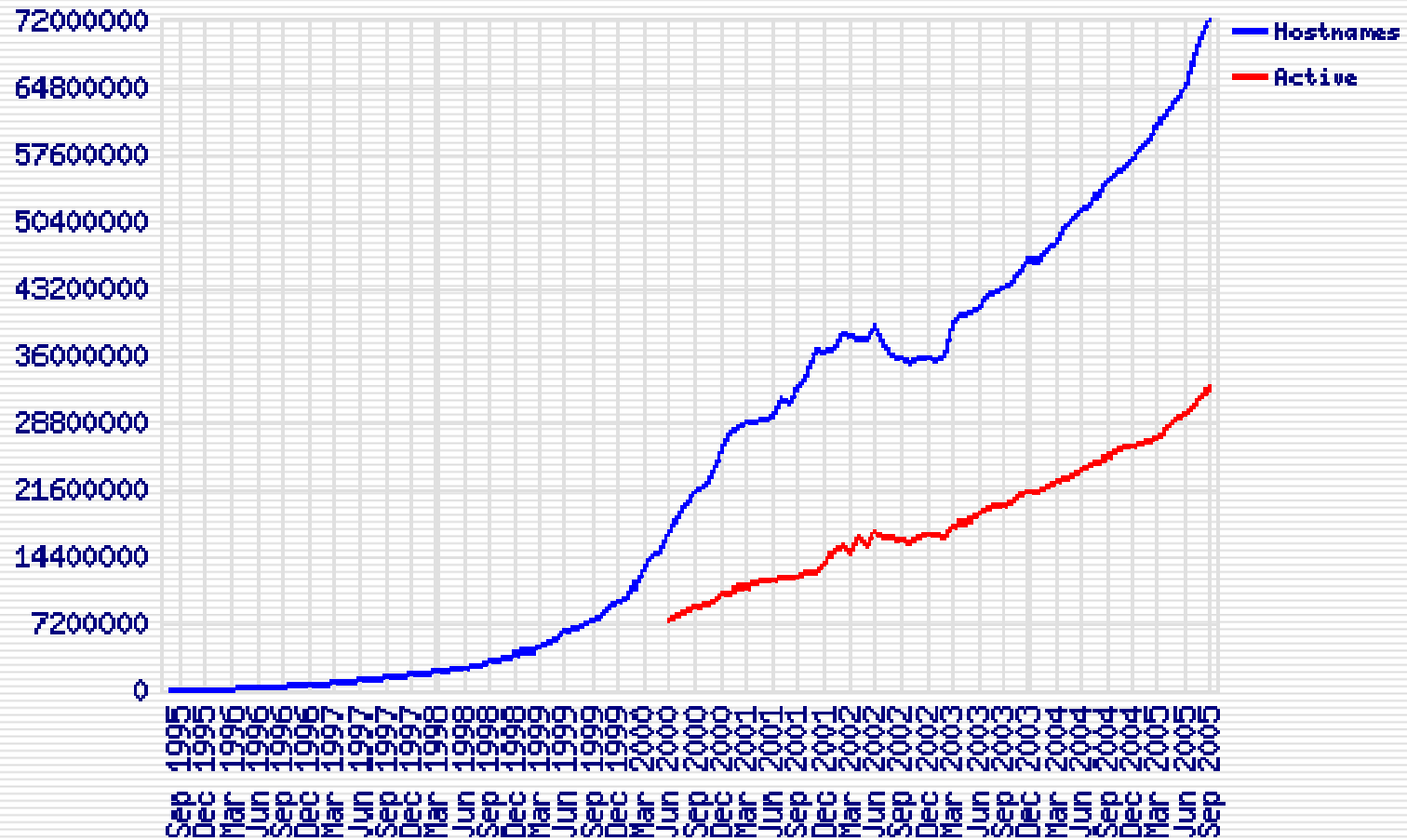
### ■ Netcraft survey says 72 million sites

([http://news.netcraft.com/archives/web\\_server\\_survey.html](http://news.netcraft.com/archives/web_server_survey.html))

---

# Netcraft survey

---



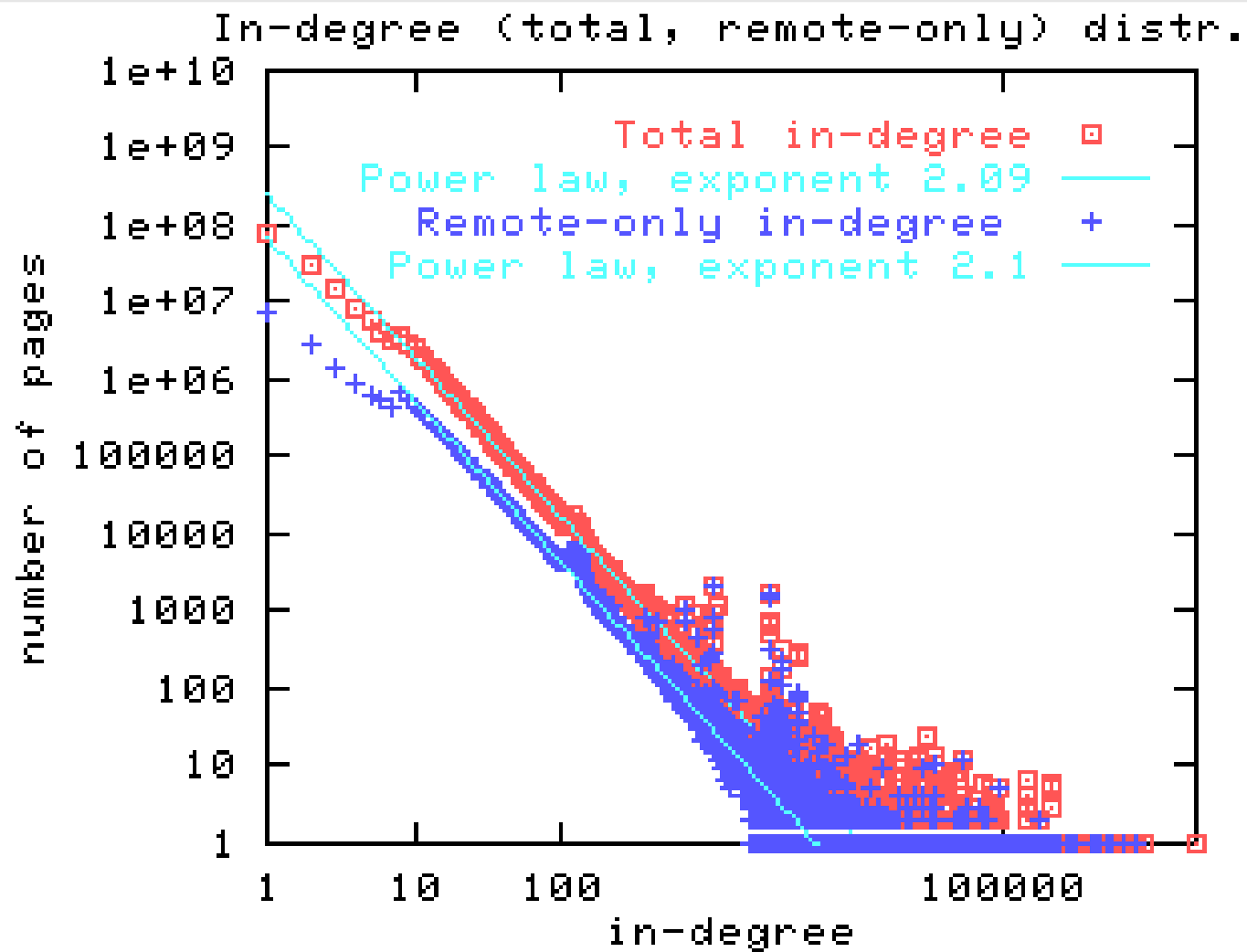
[http://news.netcraft.com/archives/web\\_server\\_survey.html](http://news.netcraft.com/archives/web_server_survey.html)

# 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
-

# Power-law degree distribution



Source: Broder et al, 2000

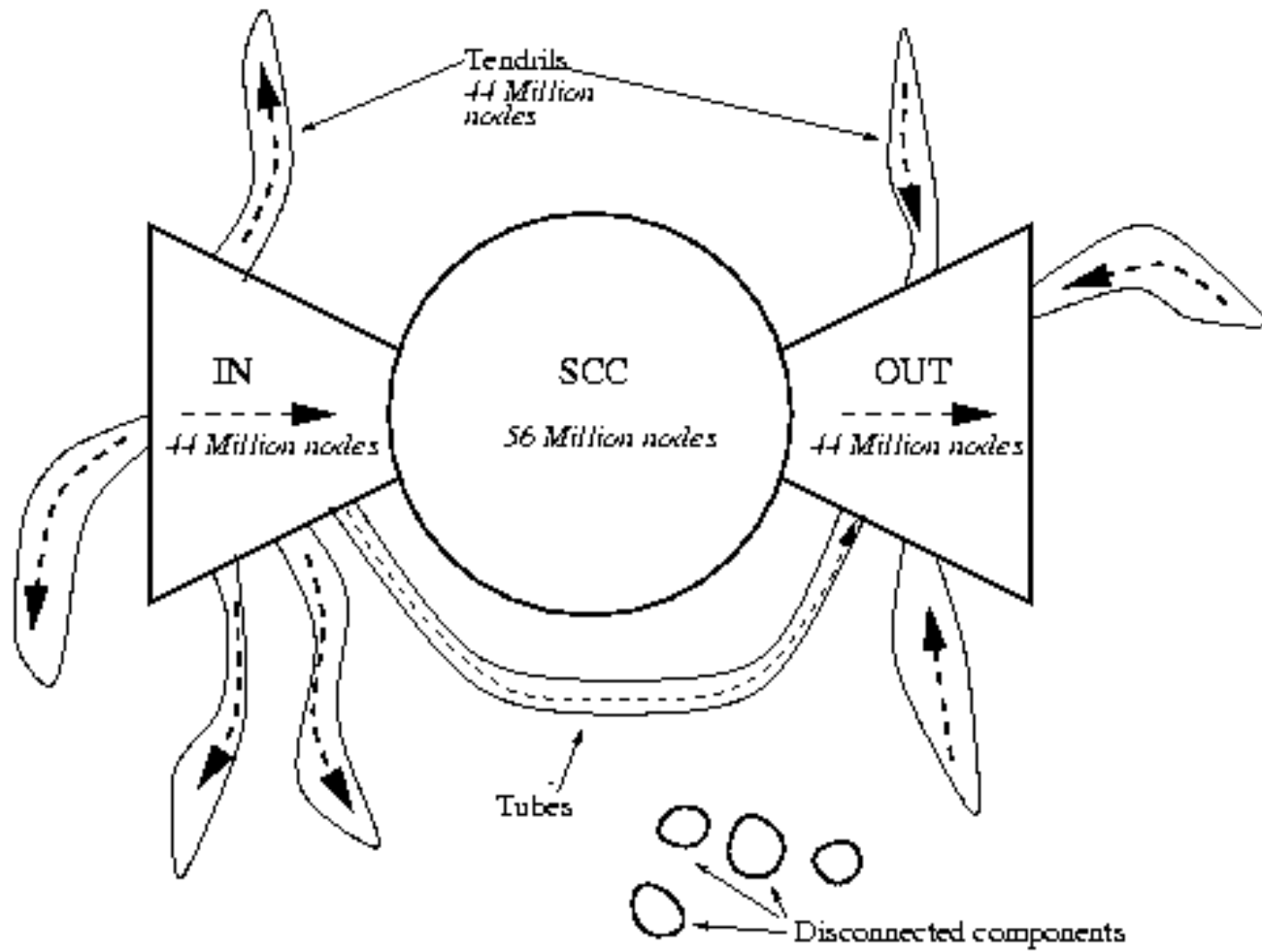


# Power-laws galore

---

- ❑ In-degrees
  - ❑ Out-degrees
  - ❑ Number of pages per site
  - ❑ Number of visitors
  - ❑ 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"
-

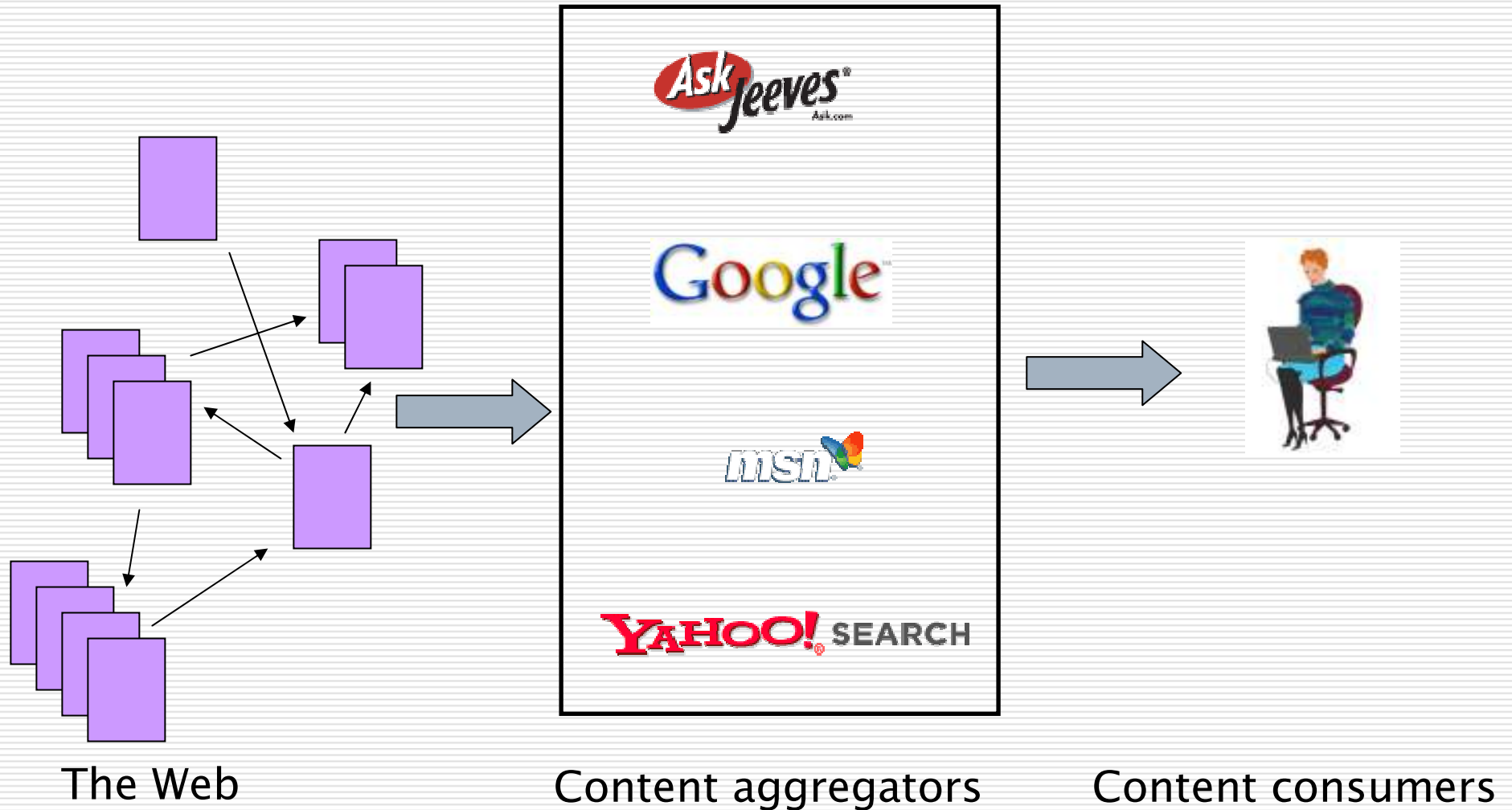
# Bow-tie Structure



Source: Broder et al, 2000

# 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
    - How to pick the top 10 results for a search from 2,230,000 matching pages?
    - What ads to show for a search?
    - If I'm an advertiser, which search terms should I bid on and how much to bid?
-

# Sidebar: What's in a name?

---

- Geico sued Google, contending that it owned the trademark "Geico"
    - Thus, ads for the keyword **geico** couldn't be sold to others
  - Court Ruling: search engines can sell keywords including trademarks
  - No court ruling yet: whether the ad itself can use the trademarked word(s)
-

# Extracting Structured Data

The screenshot shows the SimplyHired search interface. At the top left is the 'simplyhired' logo. To its right are links for 'search', 'browse', and 'suggestions'. Below these are two input fields: one for 'keywords' containing 'software engineer' and one for 'location' containing 'Mountain View, CA'. A 'search' button and a link to 'advanced search' are also present. A grey bar below the search fields indicates the sorting order: 'sorted by: best match first | newest job first'. The first job listing is for 'Software Implementation Consultant / Engineer' at 'Kaidara Software (Los Altos, CA)'. The description states that Kaidara Software provides software solutions to reduce costs and is looking for a Software Implementation Consultant / Engineer. It was posted '2 days and 3 hours ago' from 'Monster'. Below the listing are four buttons: 'who do i know?™', 'research salary', 'send-to-friend', and 'apply now'. The second job listing is for 'Software Engineer' at 'ESP Enviromental Software (Mountain View, CA)'. The description mentions server-side data updates and various data manipulation tools. It was posted '2 days and 19 hours ago' from 'Dice'.

<http://www.simplyhired.com>



# Extracting structured data

**fatlens**

"...a site the net has been waiting for." -USA TODAY

Find Tickets:

Buffalo Bills - Oakland Raiders, Network Associates Coliseum Oakland, 10-23-05

go

refine:

By Price:

All Prices

By Section:

All Sections

By Seller:

All Sellers

event tickets

Buffalo Bills - Oakland Raiders

Sunday, October 23, 2005

Network Associates Coliseum  
Oakland, CA



Click here for Seating Chart

< previous 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | next >

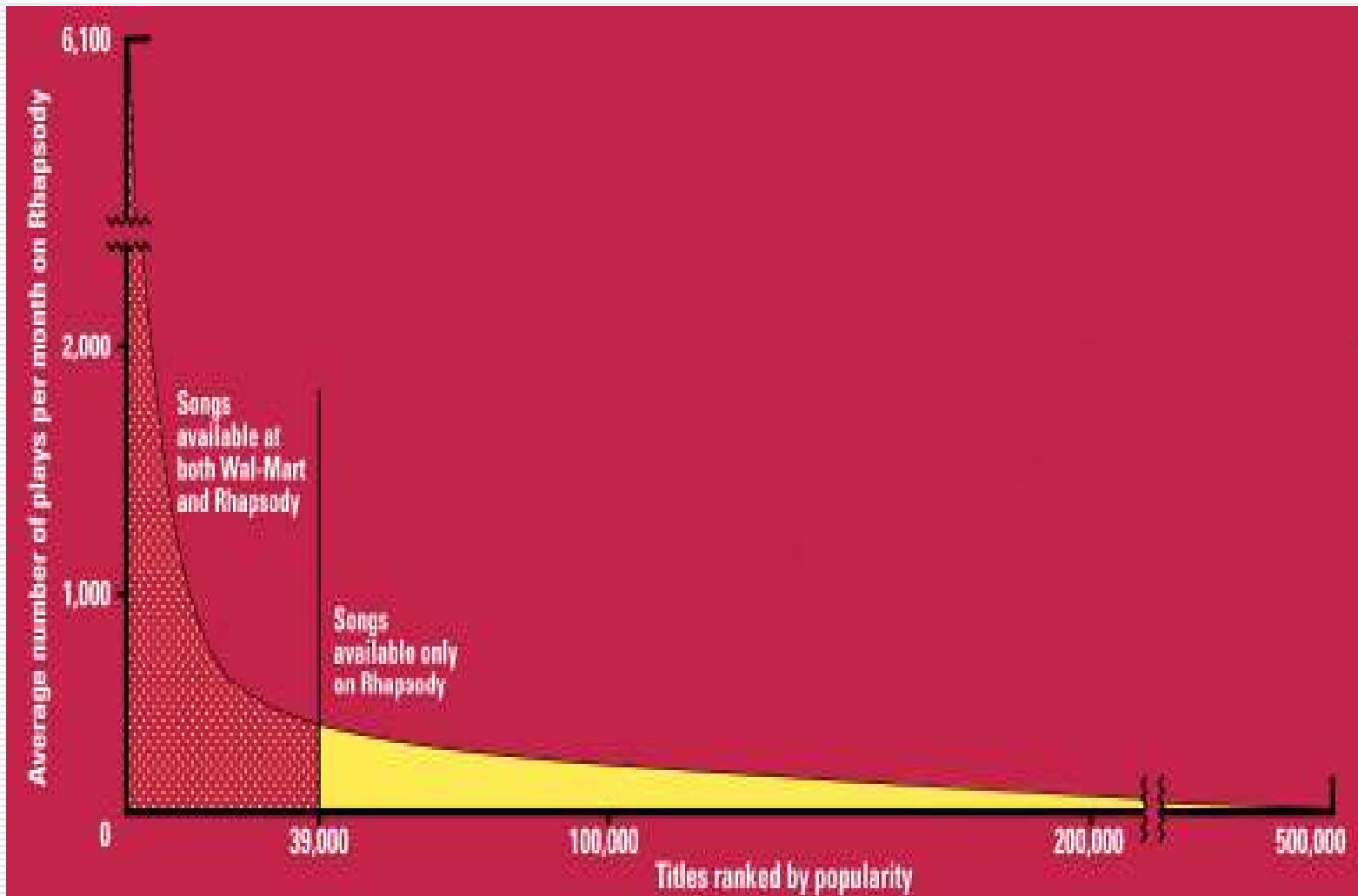
★ marks the best values in each section.

seller	section	price	
TicketLiquidator.com	42	\$184 ★	<a href="#">buy tix</a>
eBay	lower	\$318 ★	<a href="#">buy tix</a>
TICKET SOLUTIONS.com	108	\$155 ★	<a href="#">buy tix</a>
RAZORGATOR	146	\$149 ★	<a href="#">buy tix</a>
ABC Ticket Company	129	\$115 ★	<a href="#">buy tix</a>
Entertainmentbroker	119	\$165 ★	<a href="#">buy tix</a>

http://www.fatlens.com



# The Long Tail



Source: Chris Anderson (2004)

Sources: Erik Brynjolfsson and Jeffrey Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; RoadNetworks

# The Long Tail

---

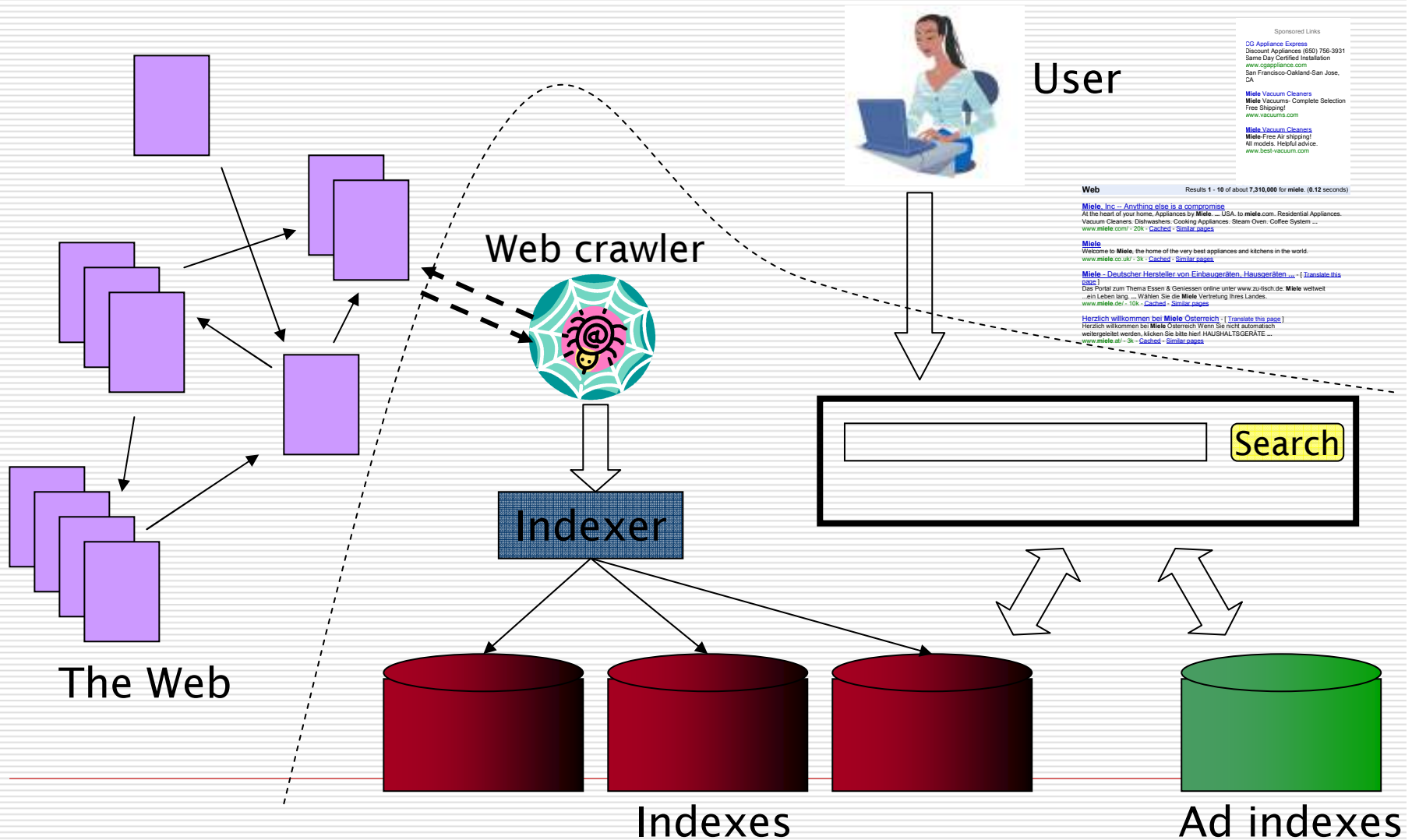
- Shelf space is a scarce commodity for traditional retailers
    - Also: TV networks, movie theaters,...
  - The web enables near-zero-cost dissemination of information about products
  - More choices necessitate better filters
    - Recommendation engines (e.g., Amazon)
    - How **Into Thin Air** made **Touching the Void** a bestseller
-

# Web Mining topics

---

- Crawling the web
  - Web graph analysis
  - Structured data extraction
  - Classification and vertical search
  - Collaborative filtering
  - Web advertising and optimization
  - Mining web logs
  - Systems Issues
-

# Web search basics



# Search engine components

---

- Spider (a.k.a. crawler/robot) – builds corpus
    - Collects web pages recursively
      - For each known URL, fetch the page, parse it, and extract new URLs
      - Repeat
    - Additional pages from direct submissions & other sources
  - The indexer – creates inverted indexes
    - Various policies wrt which words are indexed, capitalization, support for Unicode, stemming, support for phrases, etc.
  - Query processor – serves query results
    - Front end – query reformulation, word stemming, capitalization, optimization of Booleans, etc.
    - Back end – finds matching documents and ranks them
-