

CS 345A
Data Mining
Lecture 1

Introduction to Web Mining

What is Web Mining?

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

Web Mining v. 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
-

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
 - Until last year, Google claimed 8 billion(?), Yahoo claimed 20 billion
 - Google recently announced that their index contains 1 trillion pages
 - How to explain the discrepancy?
-

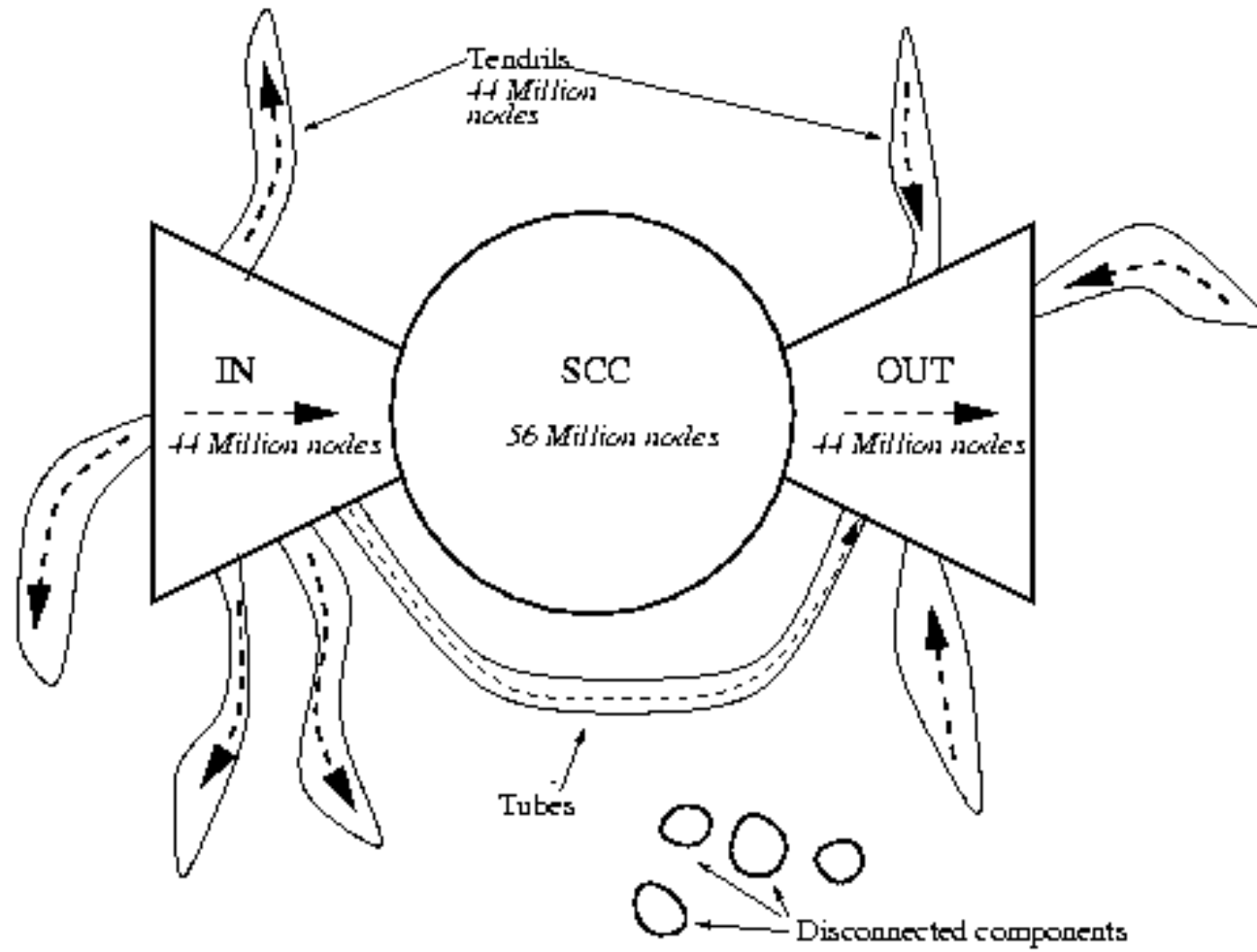
The web as a graph

- Pages = nodes, hyperlinks = edges
 - Ignore content
 - Directed graph
 - High linkage
 - 10-20 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

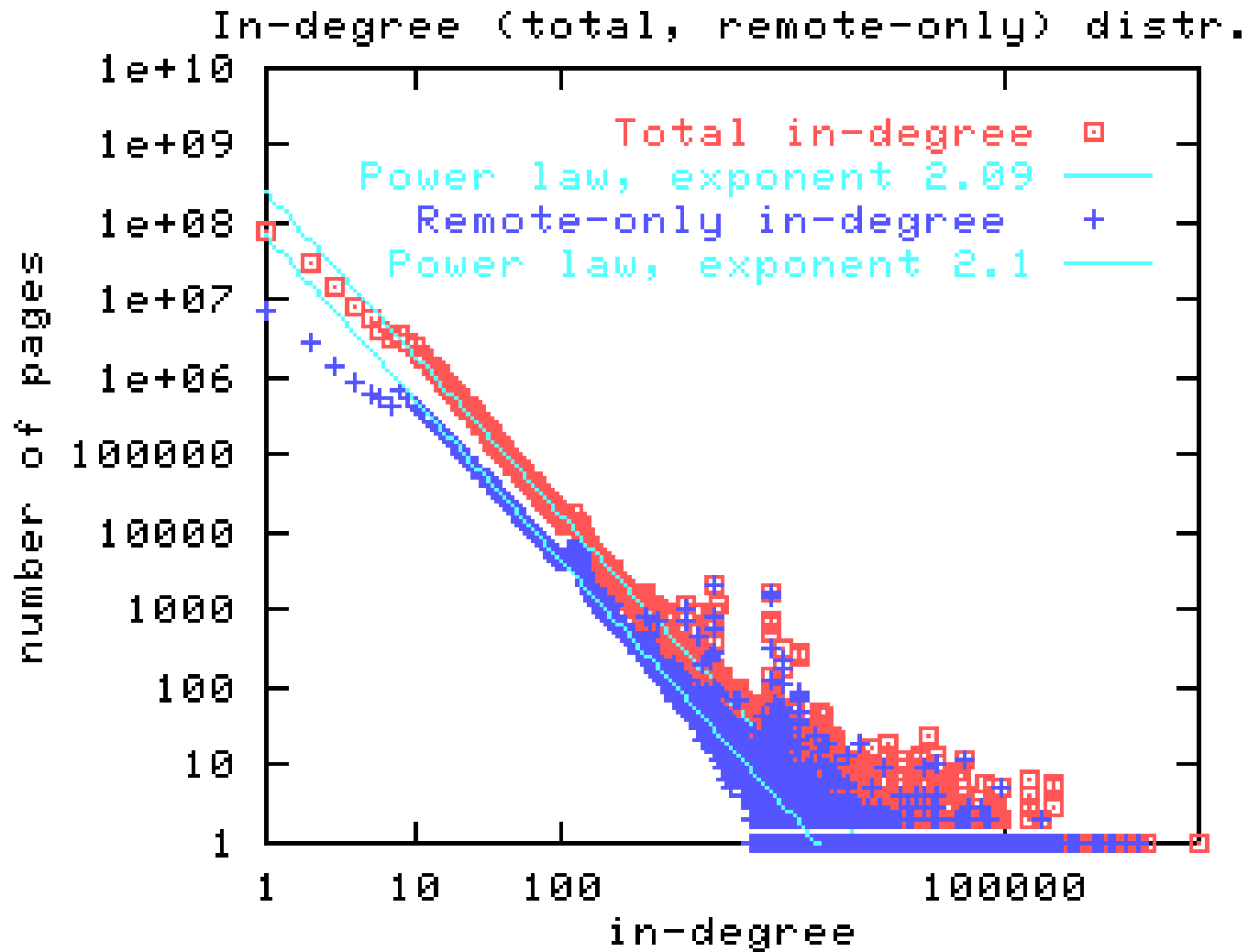
What can the graph tell us?

- 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

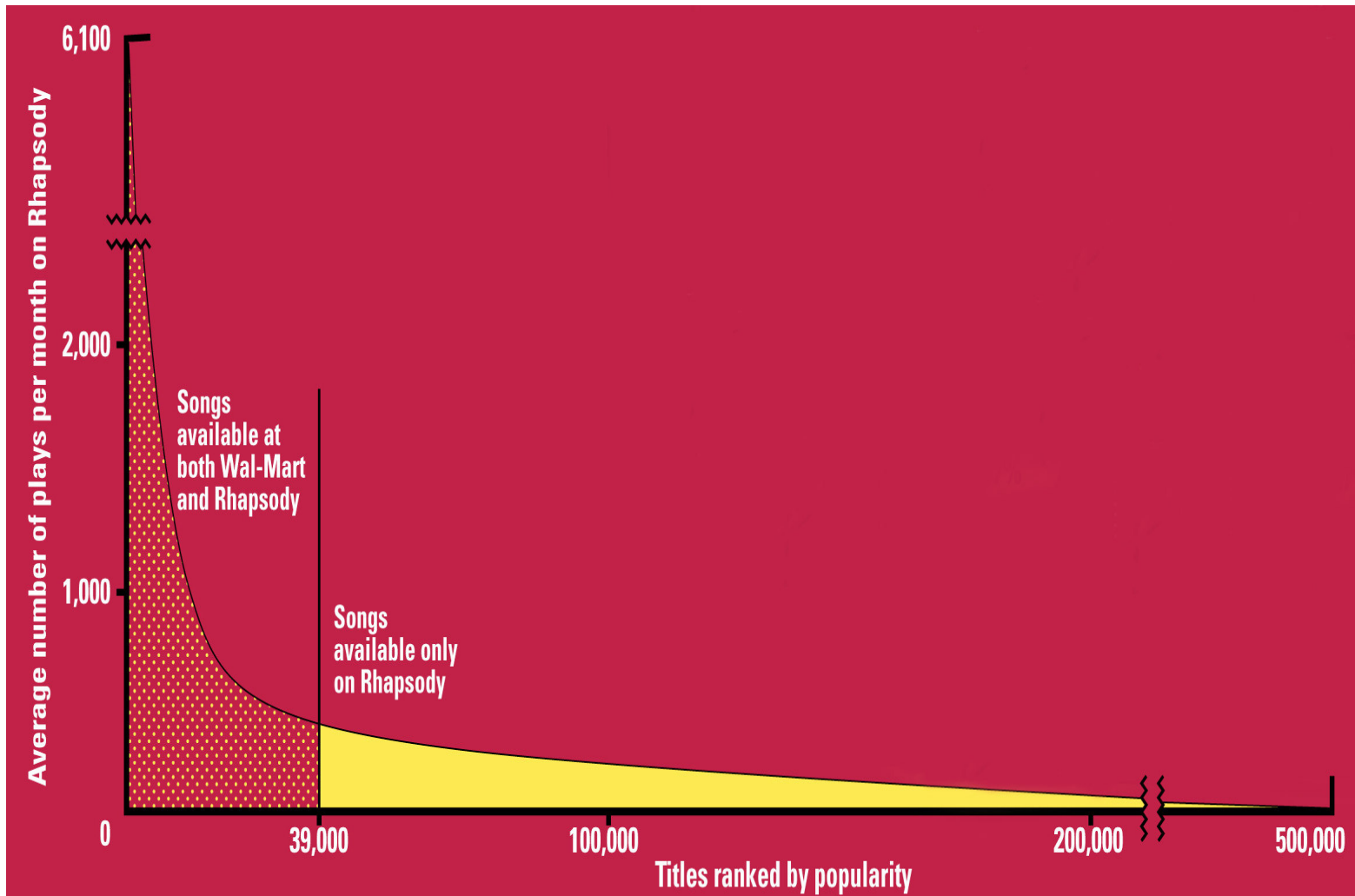
□ Structure

- In-degrees
- Out-degrees
- Number of pages per site

□ Usage patterns

- Number of visitors
 - Popularity e.g., products, movies, music
-

The Long Tail



Source: Chris Anderson (2004)

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

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 choice necessitates better filters
 - Recommendation engines (e.g., Amazon)
 - How **Into Thin Air** made **Touching the Void** a bestseller
-

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 left is the logo 'simplyhired'. To its right are navigation links: 'search', 'browse', and 'suggestions'. Below these are two input fields: 'software engineer' (labeled 'keywords') and 'Mountain View, CA' (labeled 'location'). A 'search' button is to the right of the location field, and a link for 'advanced search' is below it. A grey bar below the search area contains the text '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: 'Kaidara Software (www.kaidara.com) provides software solutions that enable firms to effectively harness the experience and know-how within an organization to reduce the cost of delivering superior customer service. We are looking for a Software Implementation Consultant / Engineer to add to our...'. 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 states: '... server-side data updates and various data manipulation tools. You'll participate in the design and development of Internet/Intranet application software to deliver the next generation of our products line that allows our customers to engage in business-to-business, e-commerce and global...'. 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.

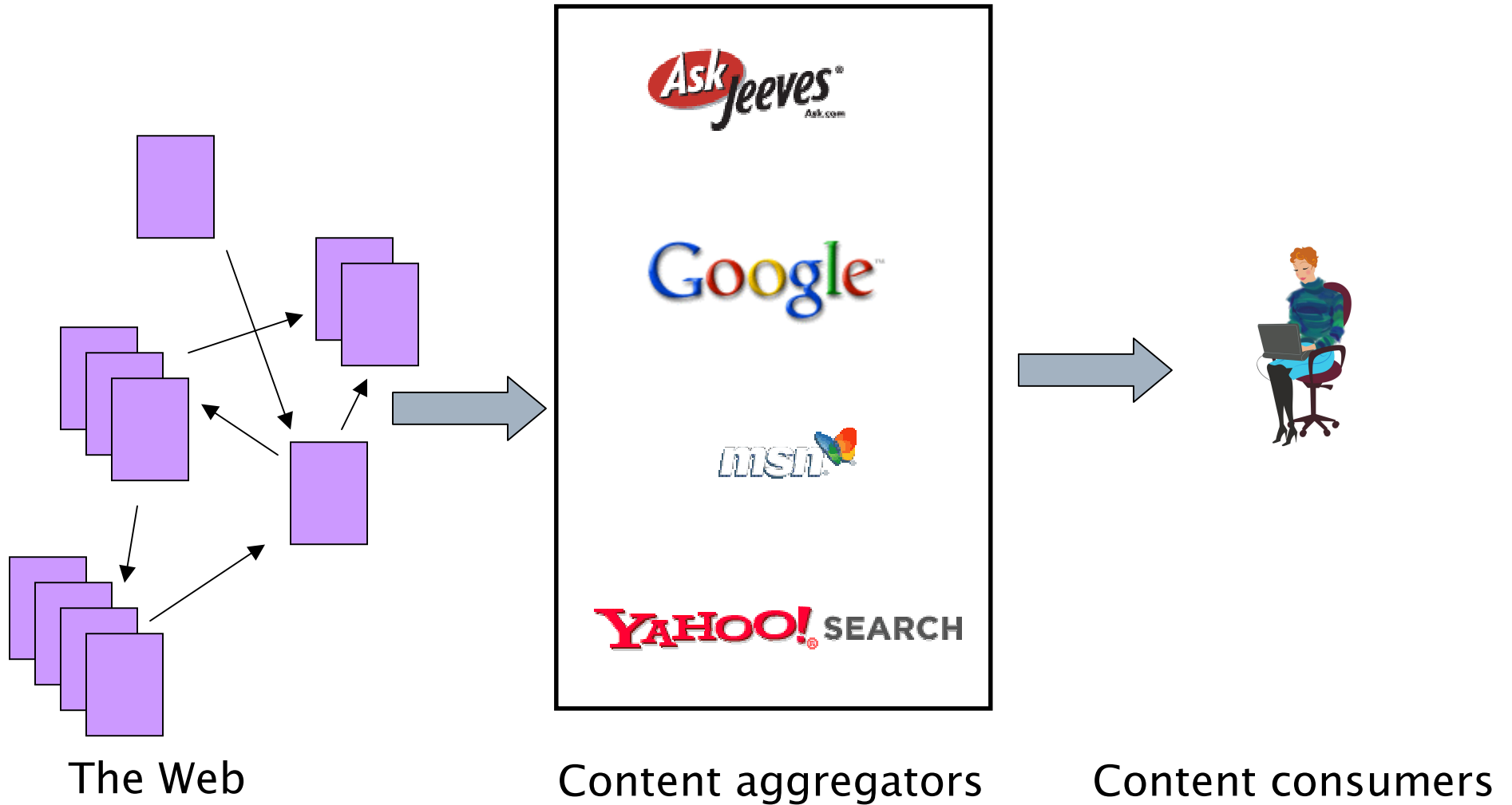
seller	section	price	
TicketLiquidator.com	42	\$184 ★	buy tix
eBay	lower	\$318 ★	buy tix
TICKET SOLUTIONS.com	108	\$155 ★	buy tix
RAZORGATOR	146	\$149 ★	buy tix
ABC Ticket Company	129	\$115 ★	buy tix
Entertainmentbroker	119	\$165 ★	buy tix

http://www.fatlens.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
-

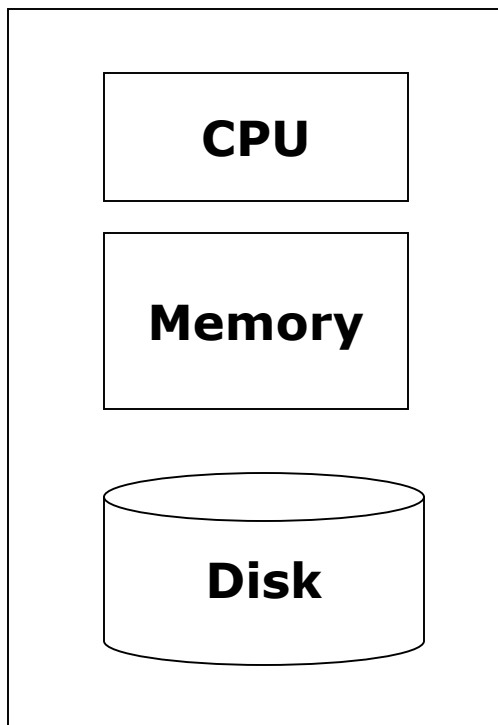
Two Approaches to Analyzing Data

- Machine Learning approach
 - Emphasizes sophisticated algorithms e.g., Support Vector Machines
 - Data sets tend to be small, fit in memory
 - Data Mining approach
 - Emphasizes big data sets (e.g., in the terabytes)
 - Data cannot even fit on a single disk!
 - Necessarily leads to simpler algorithms
-

Philosophy

- In many cases, adding more data leads to better results than improving algorithms
 - Netflix
 - Google search
 - Google ads
 - More on my blog:
Datawocky (datawocky.com)
-

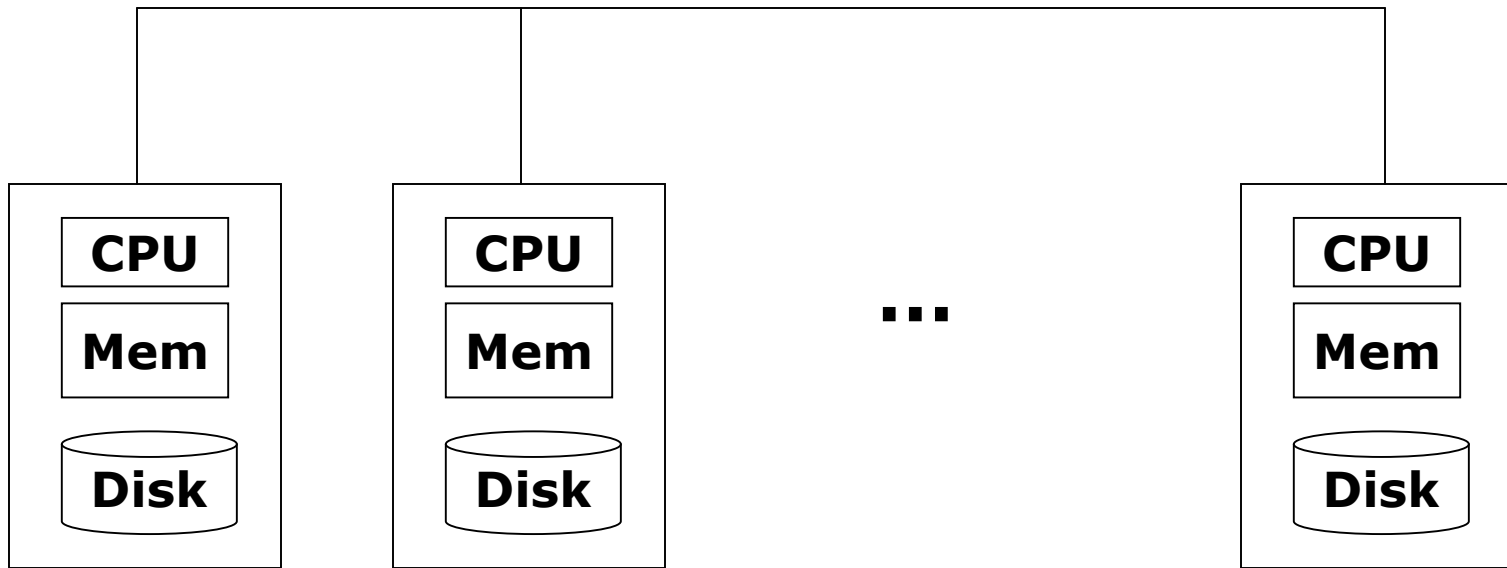
Systems architecture



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
 - Without breaking the bank!
-

Web Mining topics

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

Project

- Lots of interesting project ideas
 - If you can't think of one please come discuss with us
 - Infrastructure
 - Aster Data cluster on Amazon EC2
 - Supports both MapReduce and SQL
 - Data
 - Netflix
 - ShareThis
 - Google
 - WebBase
 - TREC
-