EPL 451

Data Mining on the Web

Τα Διαδικαστικά

- Το μάθημα απευθύνεται σε προπτυχιακούς φοιτητές Πληροφορικής
- Ό,τι πληροφορία χρειαστείτε για το μάθημα (συμβόλαιο, πρόγραμμα μαθημάτων, παραπομπές βιβλιογραφίας, επικοινωνία και ανακοινώσεις) θα την βρείτε στην Ιστοσελίδα:
  - http://www.cs.ucy.ac.cy/courses/EPL451
- Ενημέρωση και απορίες για τις διαλέξεις/εργασίες μέσω του piazza
- Ώρες γραφείου διδάσκοντα

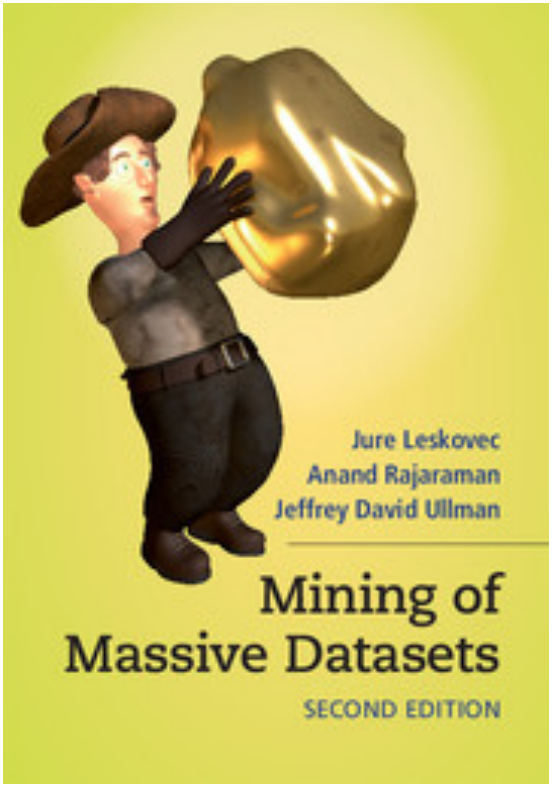
Διδασκαλία

- Δύο διαλέξεις την εβδομάδα (Δευ/Πεμ)
  - 16:30-18:00
- Φροντιστήριο (Πέμπτη)
  - 18:00 - 19:00
- Εργαστήριο (Τετάρτη)
  - Hadoop, Weka…
  - 18:00-20:00

Βιβλιογραφία

- Βιβλίο μαθήματος:
  - Mining Massive Datasets, by Jure Leskovec, Anand Rajaraman and Jeff Ullman, Cambridge University Press, 2014
  - http://mmds.org/

Για περισσότερες πληροφορίες, ανατρέξετε στην ιστοσελίδα *"Resources"* του ιστιακού τόπου του μαθήματος.

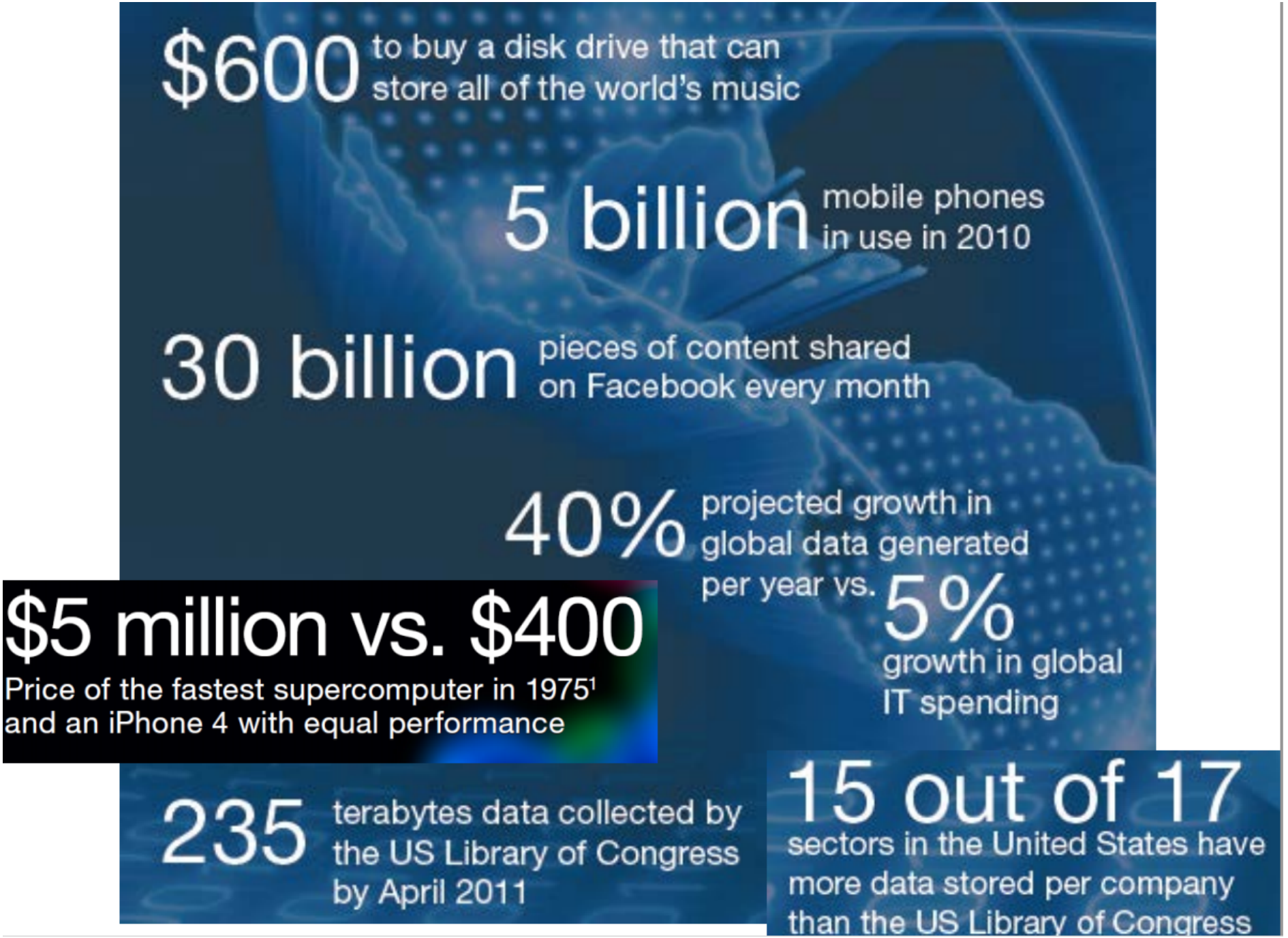


Αξιολόγηση

- 2 γραπτές εξετάσεις:
  - Ενδιάμεση (25%)
  - Τελική (40%)
- Ασκήσεις:
  - 2 Προγραμματιστικές (10%)
  - Εργασία εξαμήνου (25%)
- Quiz της ημέρας:
  - 2 απροειδοποίητα διαγνωστικά τεστ πάνω στην ύλη του μαθήματος που διδάχτηκε τη συγκεκριμένη μέρα. Η βαθμολογία τους προσμετρείται αν ο φοιτητής/ρια επιτυγχάνει πάνω από 60% (5%).

Above all

- The goal of the course is to learn and enjoy
- The basic principle is to ask questions when you don't understand
- Say when things are unclear; not everything can be clear from the beginning
- Participate in the class as much as possible


$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated per year vs. 5% growth in global IT spending
$5 million vs. $400
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance
235 terabytes data collected by the US Library of Congress by April 2011
15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

Data contains value and knowledge

Big Data Technologies are the catalyst, innovation happens in the Data Market

- **Big Data is..**

A new generation of technologies designed to economically extract value from very large volumes of data, by enabling their high-velocity capture, discovery and/or analysis.

- **The Data Market is...**

The market where digital products or services derived from data are exchanged, generating innovation, improving efficiency and effectiveness of production processes, and better understanding of customers.

- **The Data Economy is...**

The economy affected by the exploitation of data, including the direct impacts of the data market, the indirect impacts on the user industries, and the overall induced impacts on consumption and growth

IDC
Analyze the Future

Data Mining

- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - And **ANALYZED** ← this class

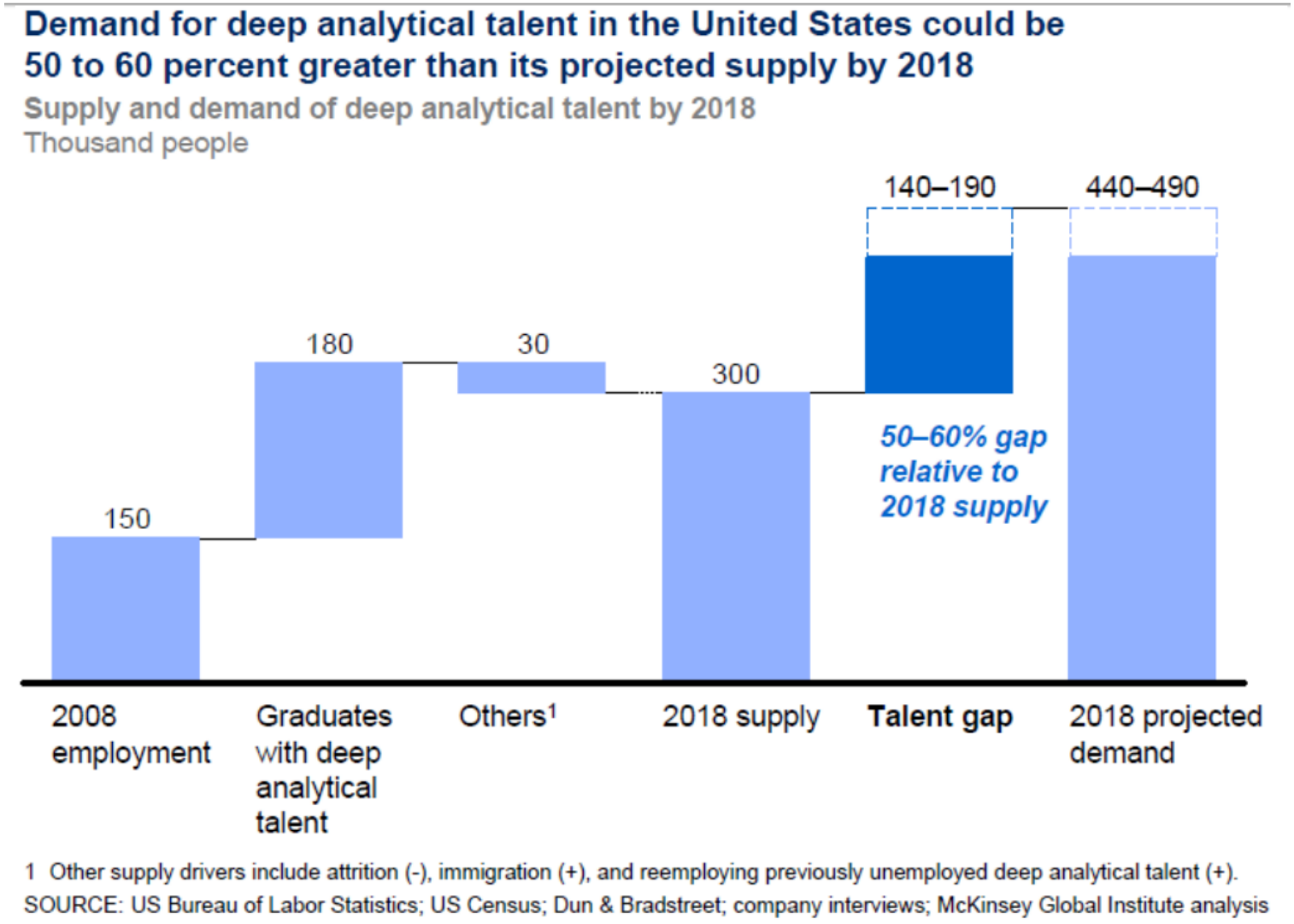
Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

Good News: Demand for Data Mining

- Recent studies by the EU (IDC and McKinsey) estimate that, by 2020, big and open data can improve the European GDP by **1.9%**, an equivalent of one full year of economic growth in the EU.

- The increased adoption of big data can have positive impact on employment, and is expected to result in millions of jobs in the EU by 2020.

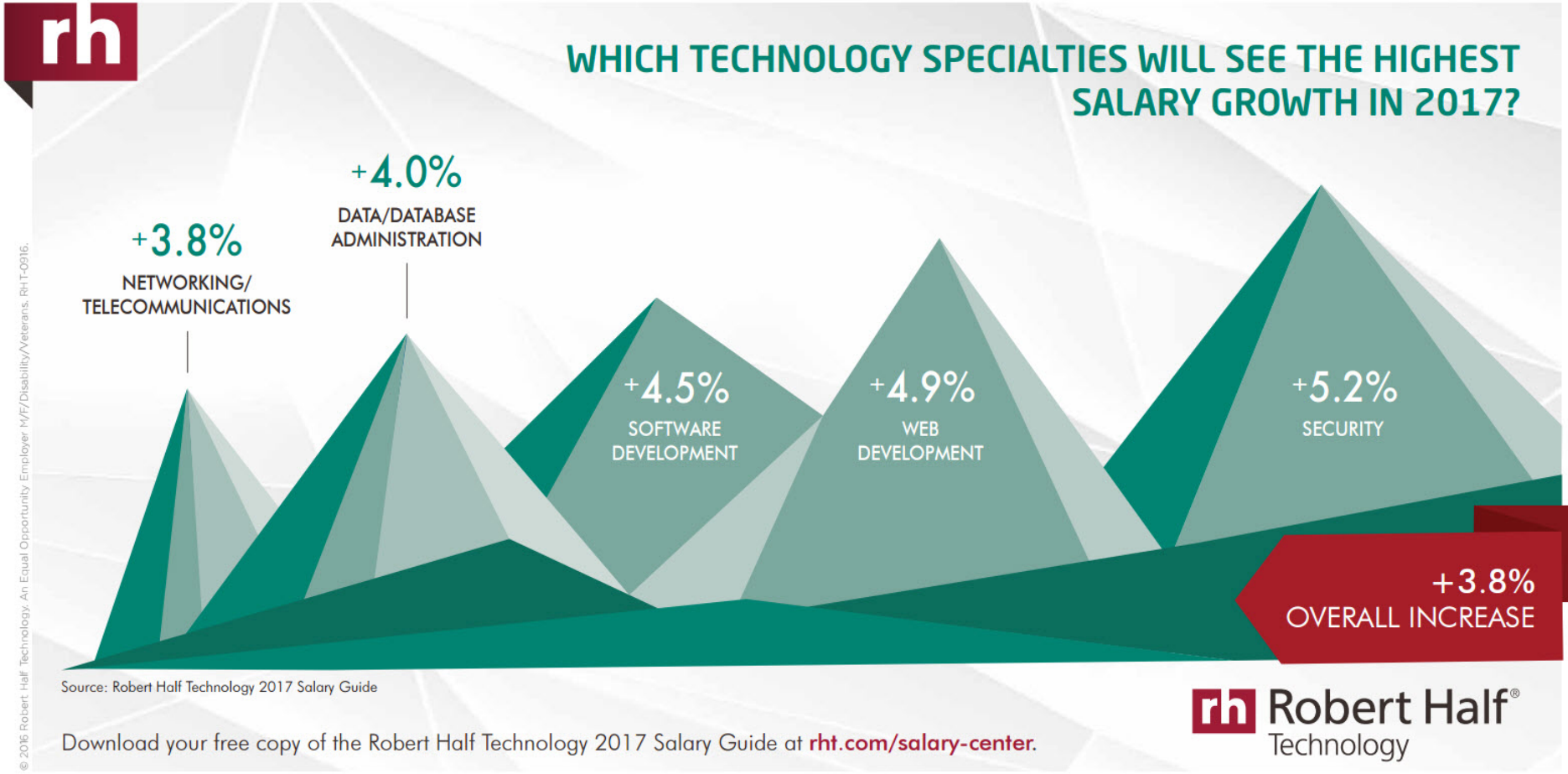
Good News: Demand for Data Mining



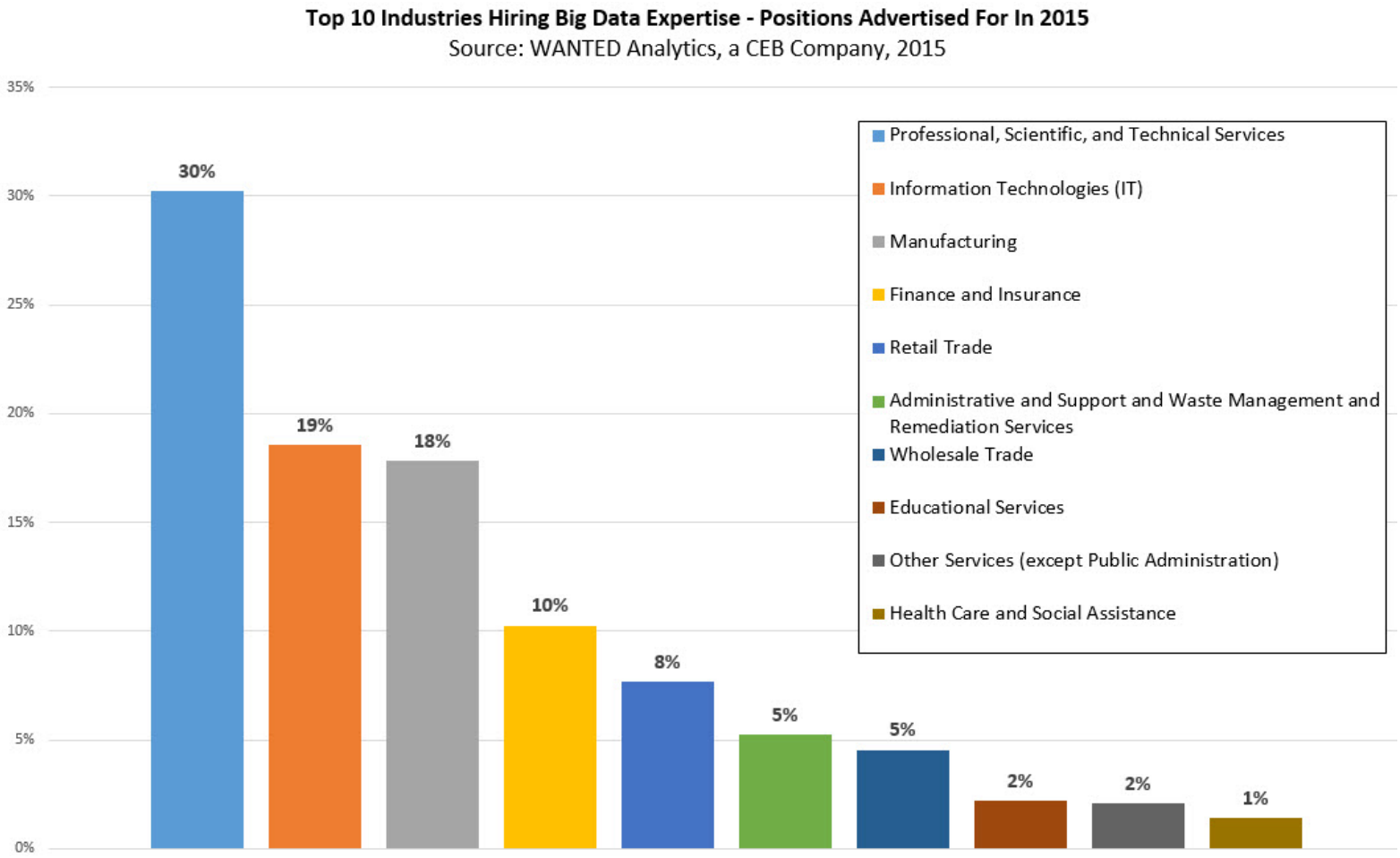
Market Salary

The median advertised salary for professionals with big data expertise is $124,000 a year. Sample jobs in this category include Software Engineer, Big Data Platform Engineer, Information Systems Developer, Platform Software Engineer, Data Quality Director, and many others. The distribution of median salaries across all industries shown below:

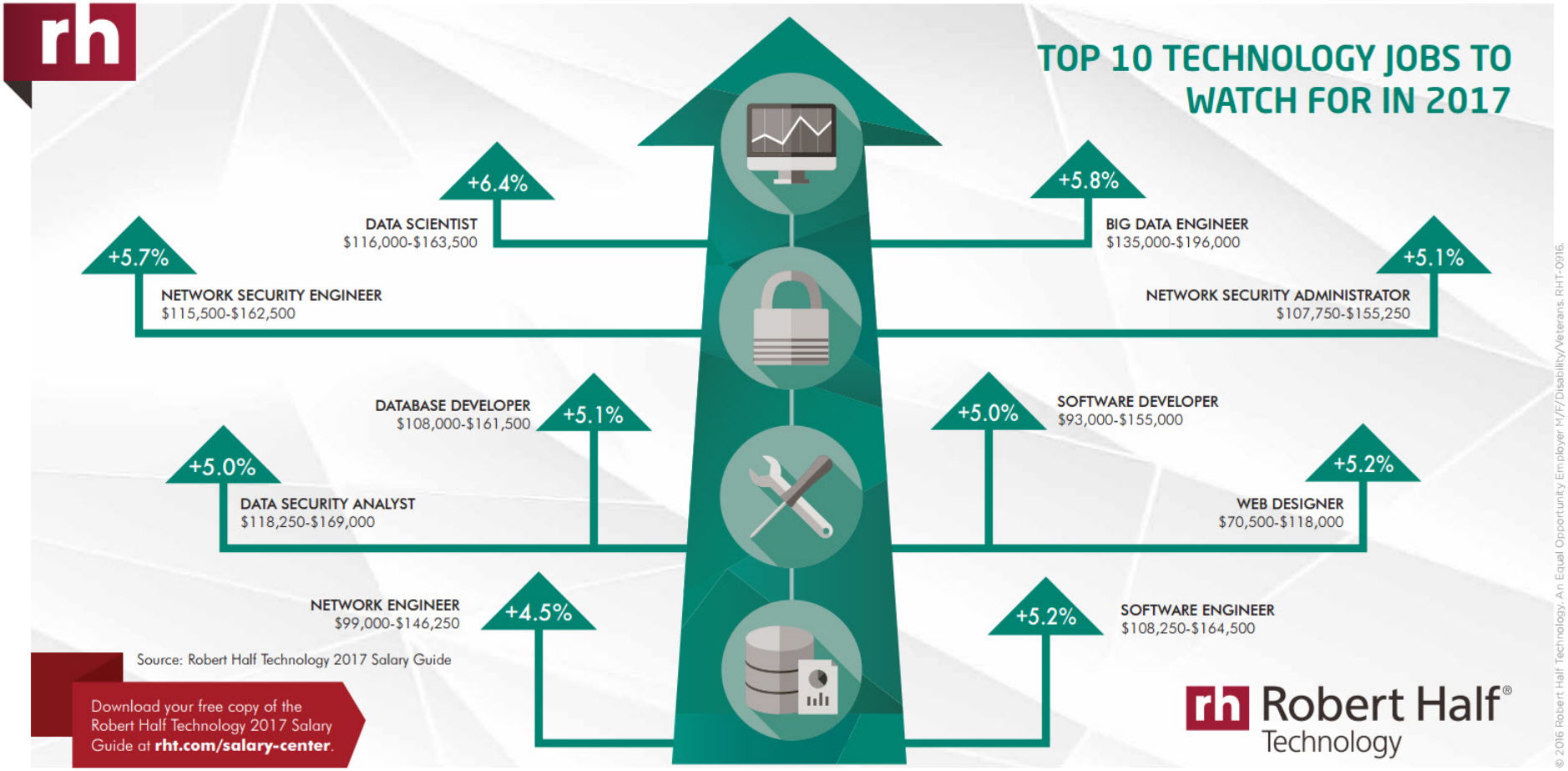
Which technology Specialties will see the highest salary growth in 2017?



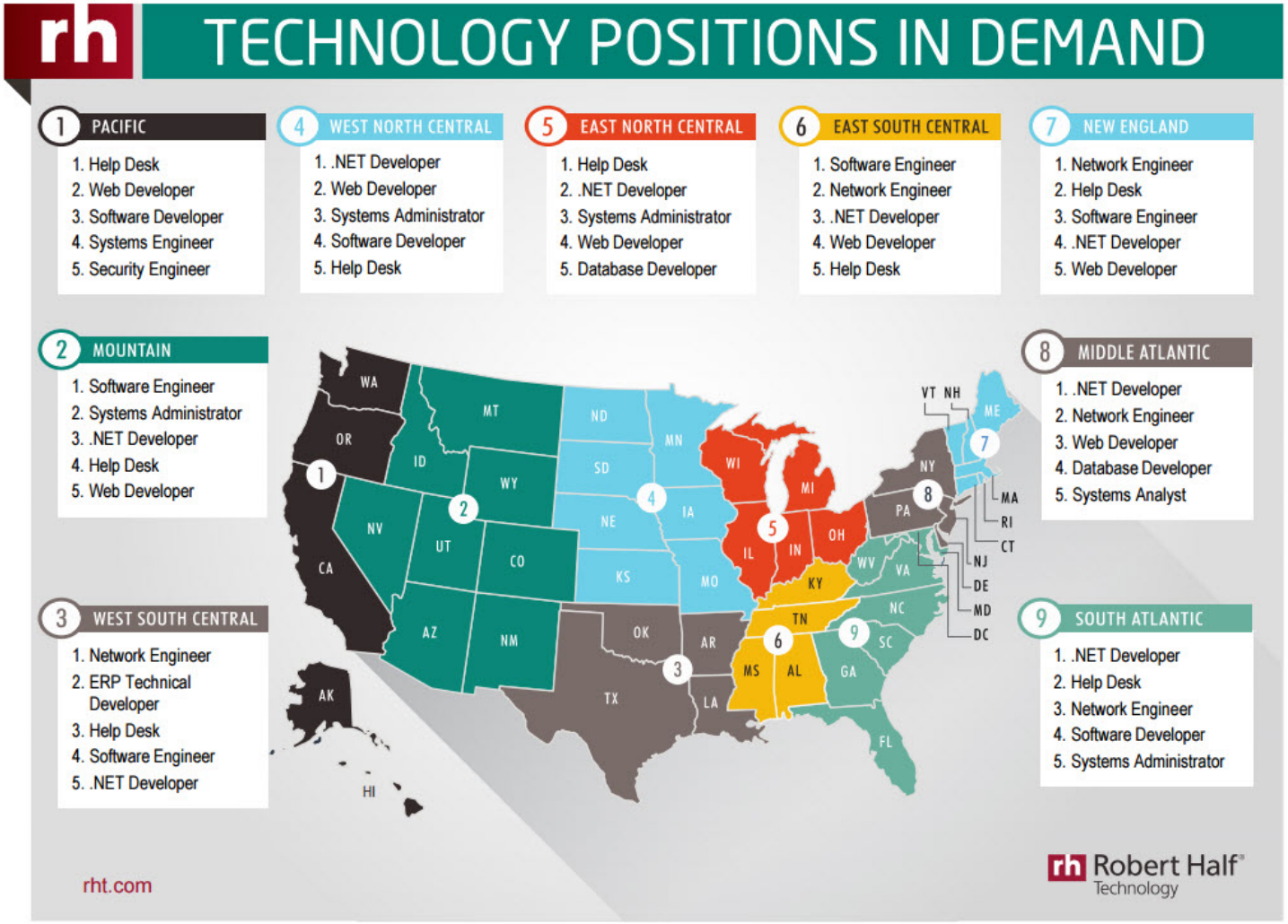
Top 10 Industries Hiring Big Data Expertise



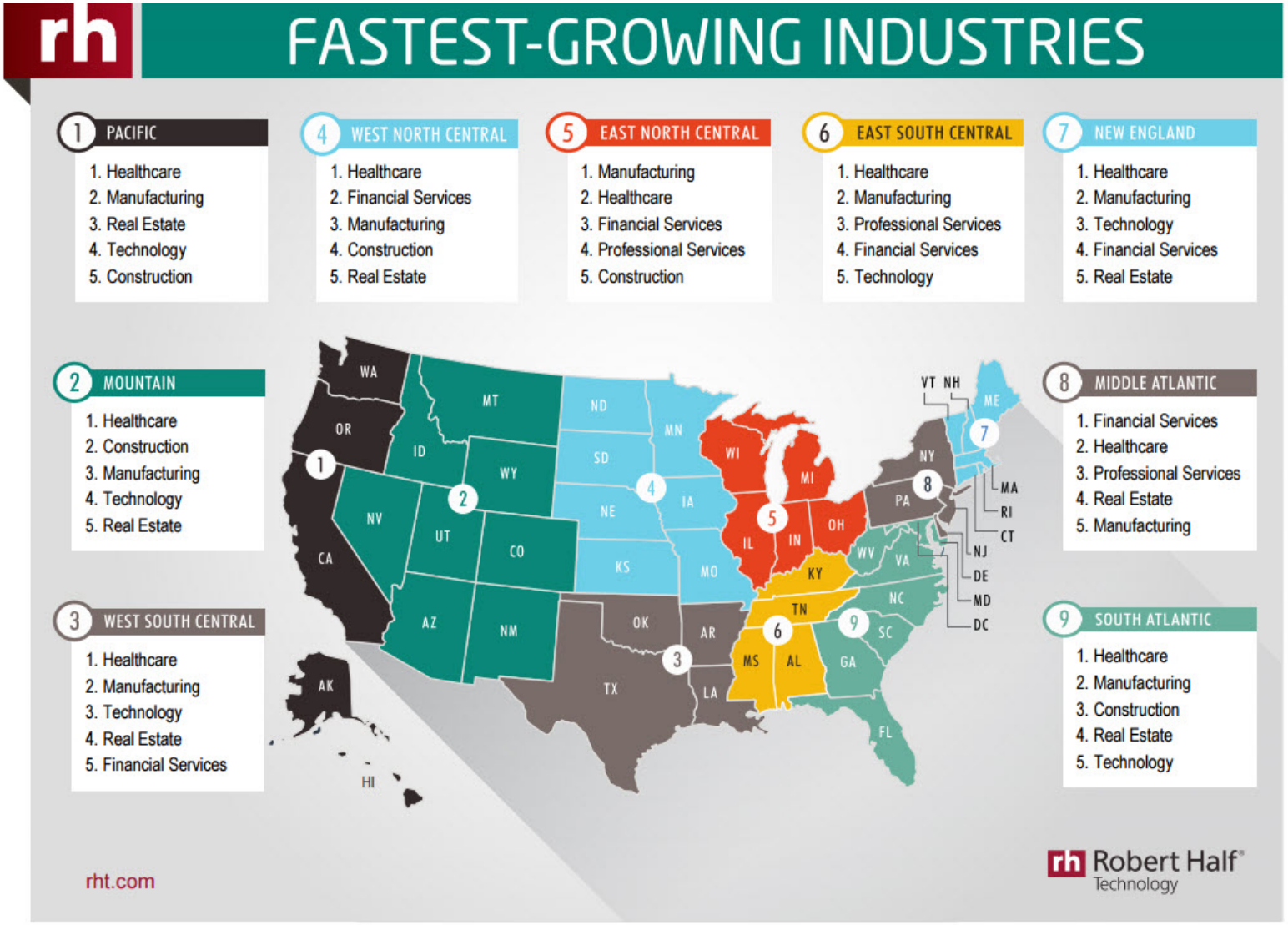
Top 10 Technology Jobs to Watch for in 2017



Top Technology Positions in Demand at USA



Fastest Growing Industries at USA



Big Data Jobs Skills

Skill	# of Big Data Jobs Mentioning This Skill Set (multiple responses allowed)	% Growth in Demand For This Skill Set Over The Previous Year
Big Data	206,967	96.29%
Java	59,886	79.63%
Python	55,907	197.19%
Hadoop	55,063	97.77%
Structured query language	47,706	65.81%
Data warehousing	42,781	209.13%
Open source technology	41,579	333.66%
Linux	41,218	110.17%
VMware	37,679	798.19%
NoSQL	35,802	169.86%

What is Data Mining?

- Given lots of data discover patterns and models that are:
  - valid: hold on new data with some certainty
  - novel: non-obvious to the system
  - useful: should be possible to act on the item
  - understandable: humans should be able to interpret the pattern

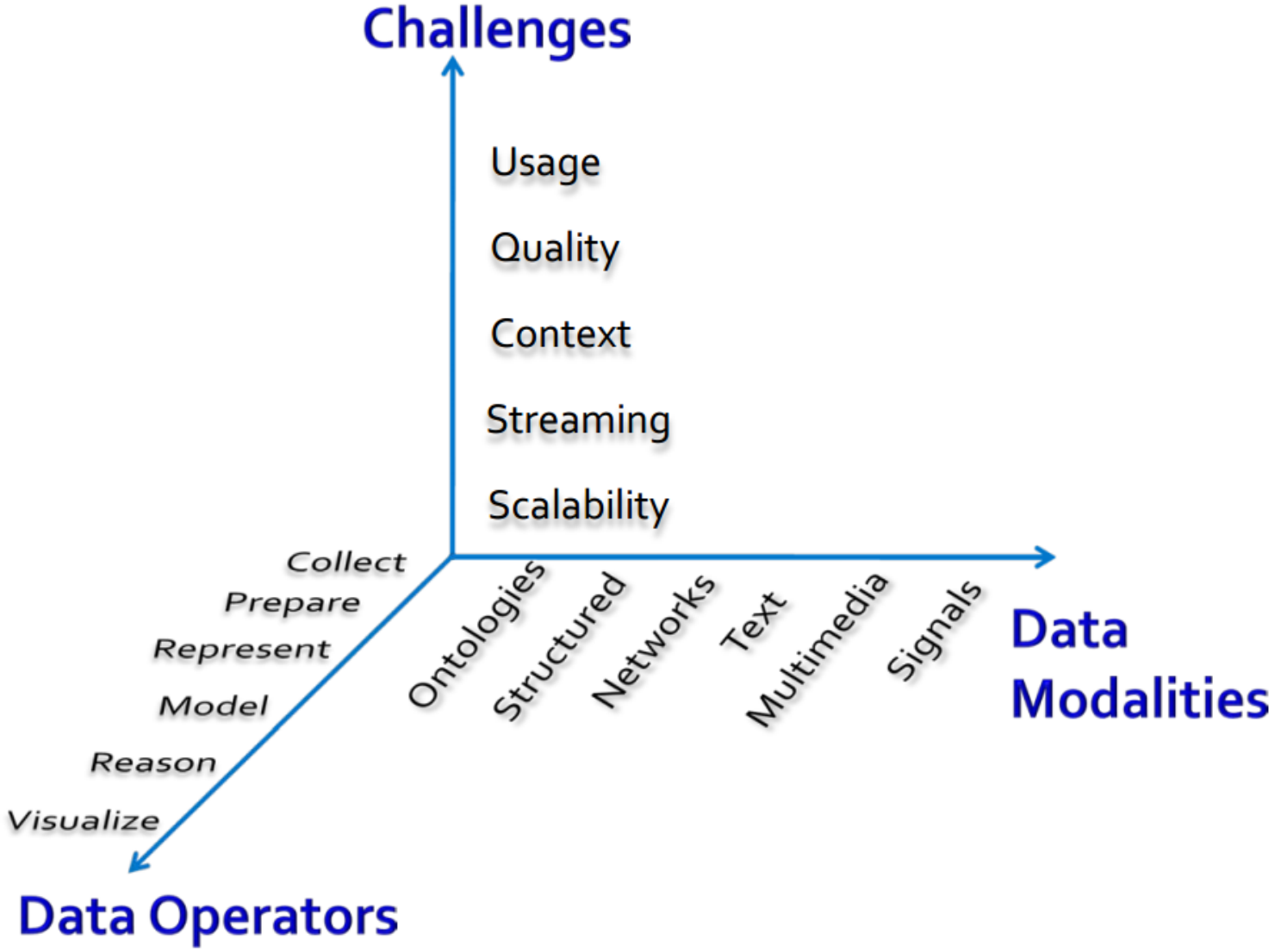
What is Data Mining?

- Subsidiary issues:
  - Data cleansing: detection of bogus data
    - E.g., age = 150
    - Entity resolution
  - Visualization: something better than megabyte files of output
  - Warehousing of data (for retrieval)

Data Mining Tasks

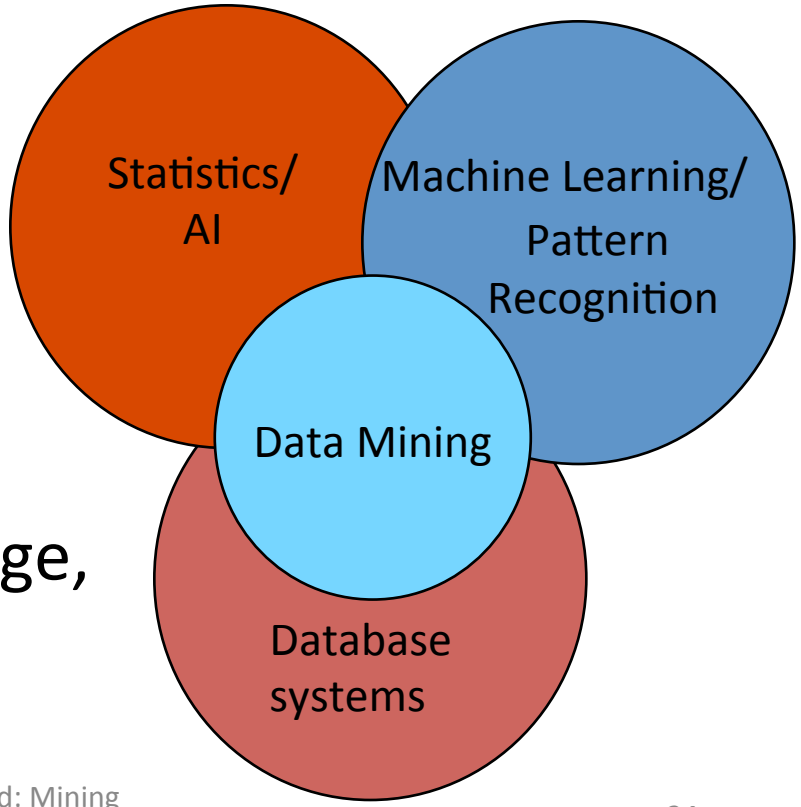
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

What matters when dealing with Data?



Origins of Data Mining

- Overlaps with machine learning, statistics, artificial intelligence, databases, visualization but more stress on
  - scalability of number of features and instances
  - stress on algorithms and architectures
  - automation for handling large, heterogeneous data



Focus: Web Mining

- The course will be devoted to ways to data mining on the Web:
  - Mining to discover things about the Web
    - E.g., PageRank, finding spam sites
  - Mining data from the Web itself
    - E.g., analysis of click streams, similar products at Amazon, making recommendations.

Focus 2: MapReduce

- Much of the course will be devoted to **Large scale computing** for **data mining**
- Challenges:
  - How to distribute computation?
  - Distributed/parallel programming is hard
- **Map-reduce** addresses all of the above
  - Google's computational/data manipulation model
  - Elegant way to work with big data

Course Outline (1)

- Association rules, frequent itemsets
- PageRank and related measures of importance on the Web (link analysis)
  - Spam detection
  - Topic-specific search Recommendation systems
  - E.g., what should Amazon suggest you buy?

Course Outline (2)

- Finding similar Web pages
- Clustering data
- Extracting structured data (relations) from the Web
- Managing Web advertisements
- Mining data streams
- Social Network Mining
- Sentiment Analysis

Prerequisites

- Algorithms (EPL 231)
  - Dynamic programming, basic data structures
- Programming (EPL 131, EPL 132, EPL 233):
  - Your choice, but C++/Java will be very useful
- Databases (EPL 345)

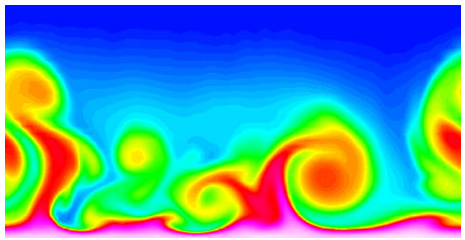
Why Mine Data? Industry

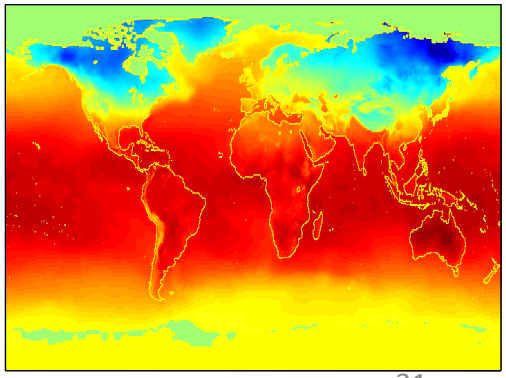
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

- Computers are cheap and powerful
- Goal:
  - Provide better, customized services (e.g. in Customer Relationship Management)

Why Mine Data? Science

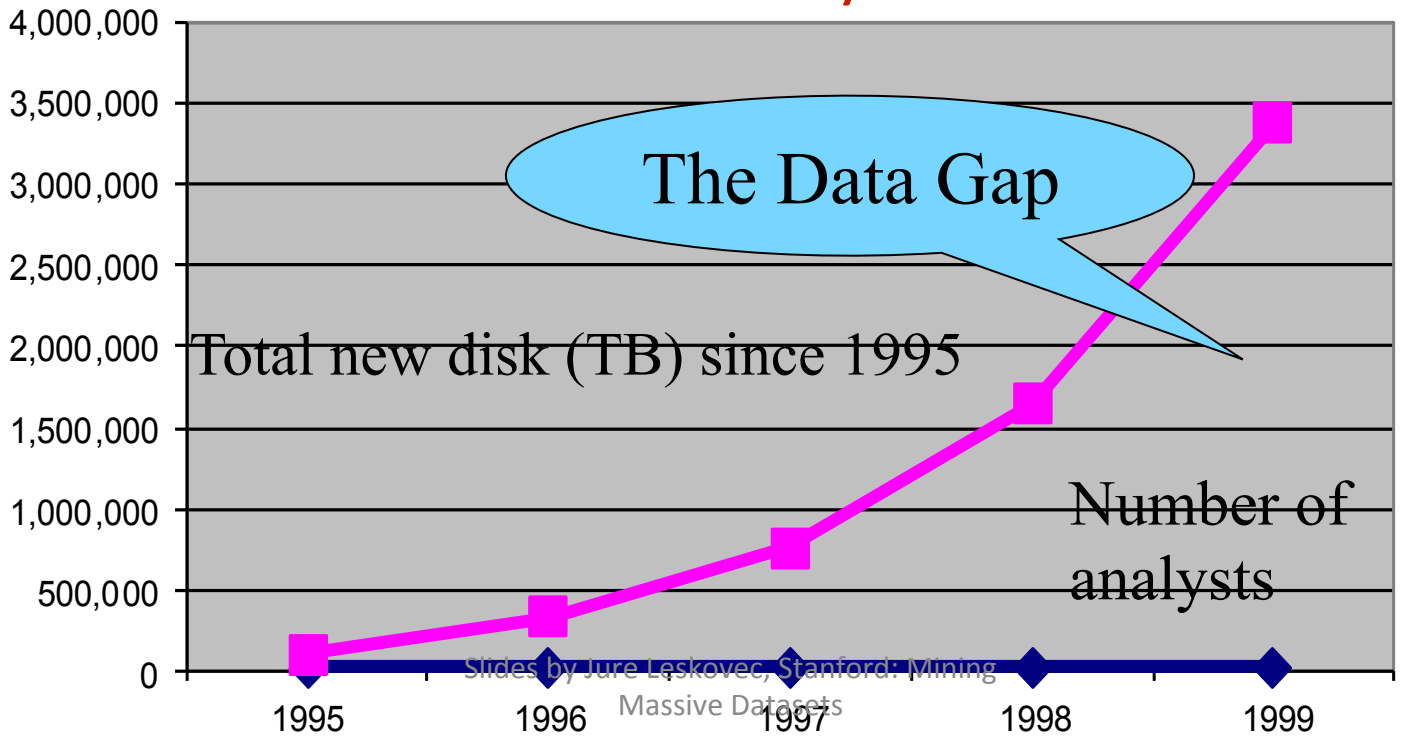
- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining helps scientists
  - in classifying and segmenting data
  - in Hypothesis Formation





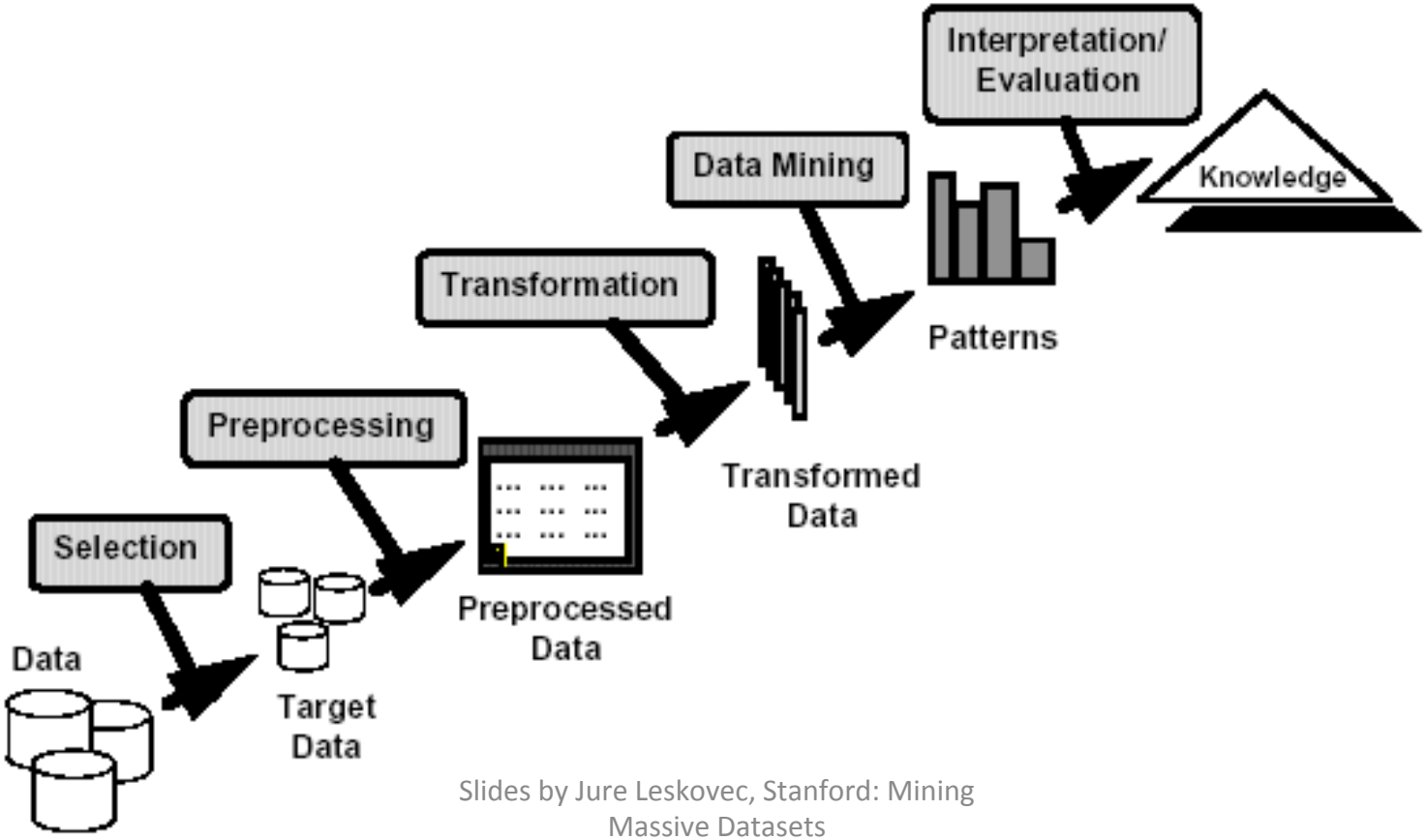
Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts take weeks to discover useful information
- Much of the data is never analyzed at all



What is Data Mining?

- Non-trivial extraction of implicit, previously unknown and useful information from data



Meaningfulness of Answers

- A big data-mining risk is that you will "discover" patterns that are meaningless.

- Bonferroni's principle: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap.

Rhine Paradox – (1)

- A parapsychologist in the 1950's hypothesized that some people had Extra-Sensory Perception
- He devised an experiment where subjects were asked to guess 10 hidden cards – red or blue
- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right

Rhine Paradox – (2)

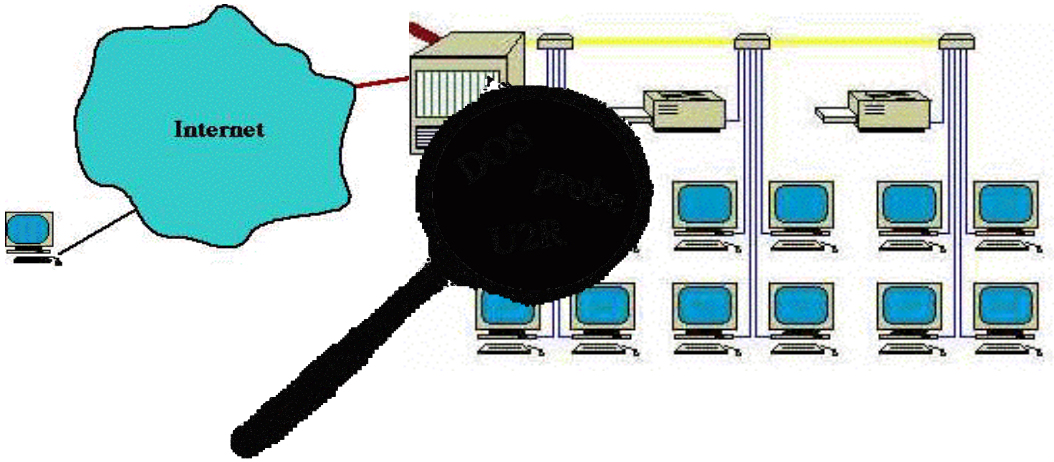
- He told these people they had ESP and called them in for another test of the same type
- Alas, he discovered that almost all of them had lost their ESP
- What did he conclude?
- He concluded that you shouldn't tell people they have ESP; it causes them to lose it ☺

Collaborative Filtering

- Given database of user preferences, predict preference of new user
- Example:
  - Predict what new movies you will like based on
    - your past preferences
    - others with similar past preferences
    - their preferences for the new movies
- Example:
  - Predict what books/CDs a person may want to buy
  - (and suggest it, or give discounts to tempt customer)

Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection



Association Rule Discovery

- Supermarket shelf management:
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule:
    - If a customer buys diaper and milk, then he is likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Some Success Stories (1)

- Network intrusion detection using a combination of sequential rule discovery and classification tree on 4 GB DARPA data
  - Won over (manual) knowledge engineering approach
  - http://www.cs.columbia.edu/~sal/JAM/PROJECT/ provides good detailed description of the entire process

Some Success Stories (2)

- Major US bank: Customer attrition prediction
  - Segment customers based on financial behavior: 3 segments
  - Build attrition models for each of the 3 segments
  - 40-50% of attritions were predicted == factor of 18 increase

Some Success Stories (3)

- Targeted credit marketing: major US banks
  - find customer segments based on 13 months credit balances
  - build another response model based on surveys
  - increased response 4 times – 2%

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Applications

- Banking: loan/credit card approval:
  - predict good customers based on old customers
- Customer relationship management:
  - identify those who are likely to leave for a competitor
- Targeted marketing:
  - identify likely responders to promotions
- Fraud detection: telecommunications, finance
  - from an online stream of event identify fraudulent events
- Manufacturing and production:
  - automatically adjust knobs when process parameter changes

Applications (continued)

- Medicine: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relationship between diseases
- Molecular/Pharmaceutical:
  - identify new drugs
- Scientific data analysis:
  - identify new galaxies by searching for sub clusters
- Web site/store design and promotion:
  - find affinity of visitor to pages and modify layout

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

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?

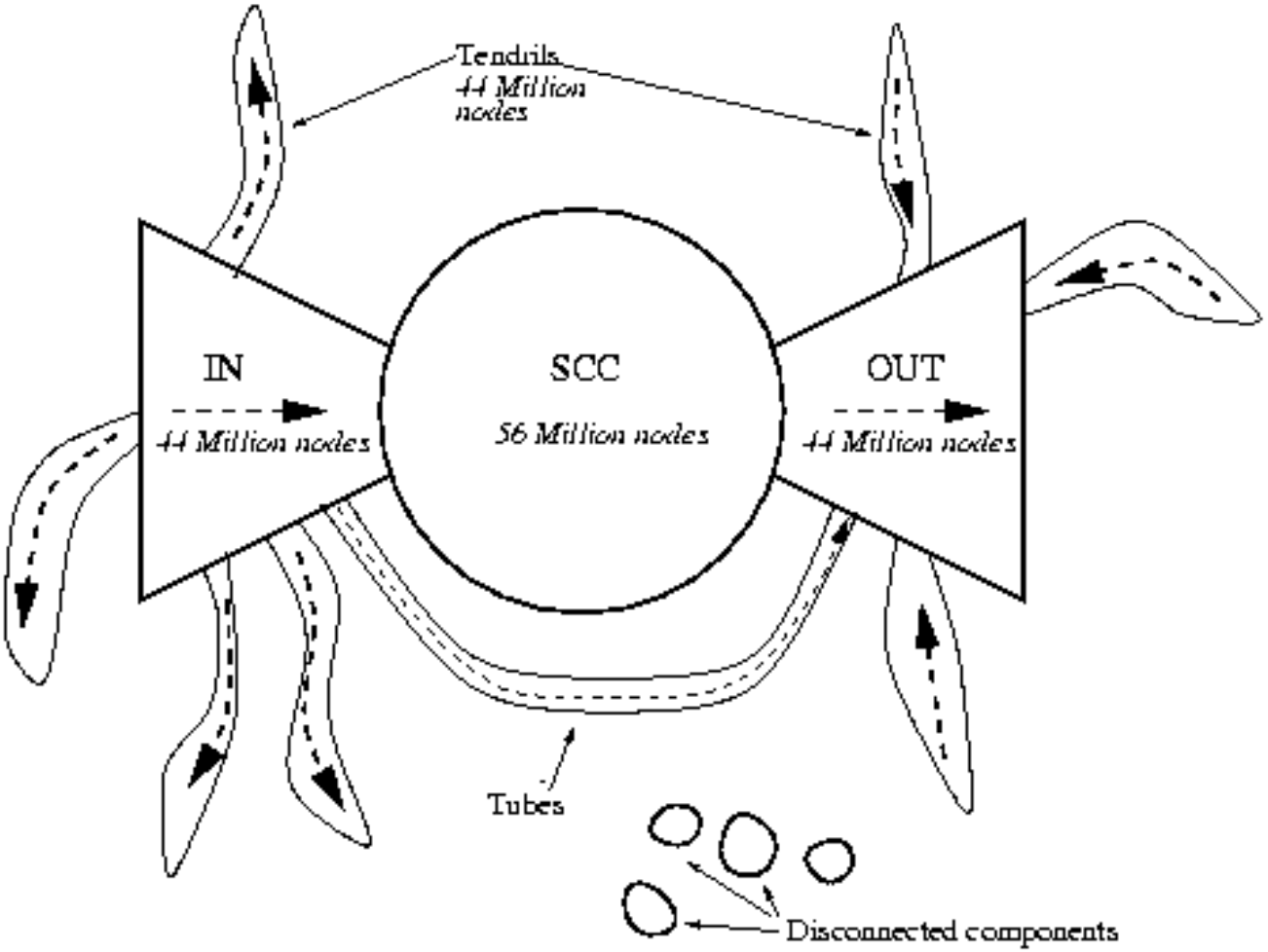
The web as a graph

- 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

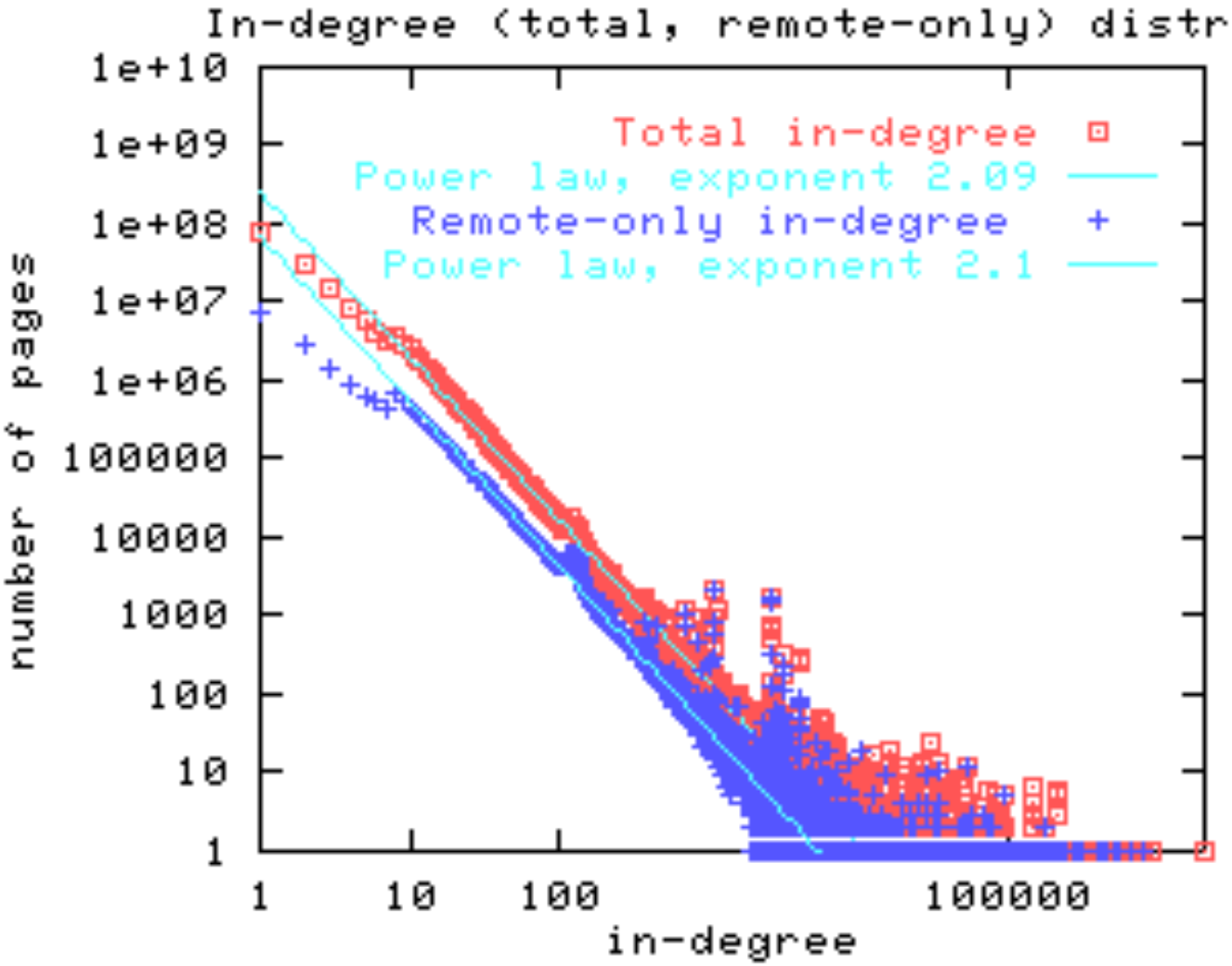
What can the graph tell us?

- 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



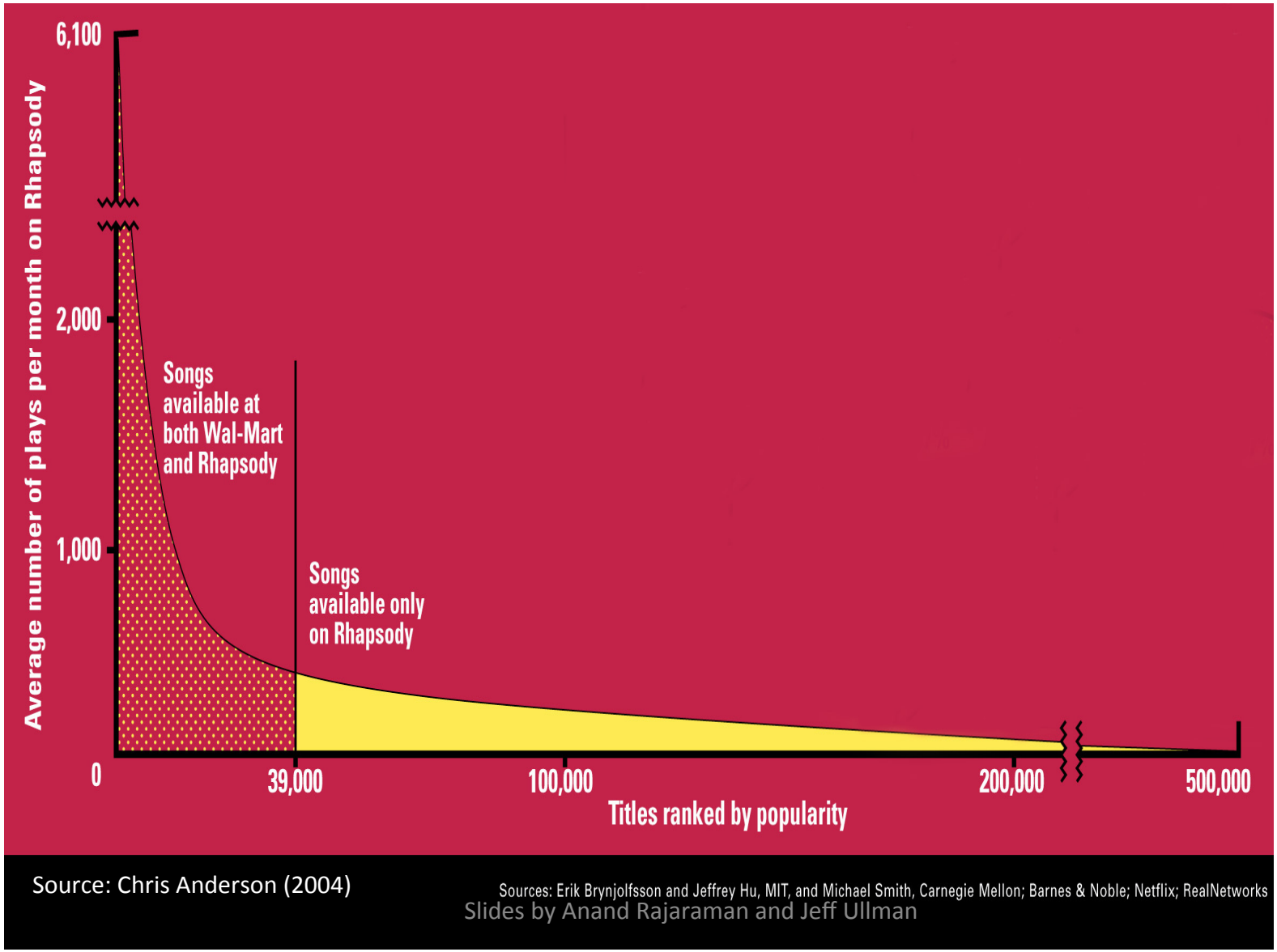
Source: Broder et al, 2000

Slides by Jure Leskovec, Stanford: Mining Massive Datasets

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)

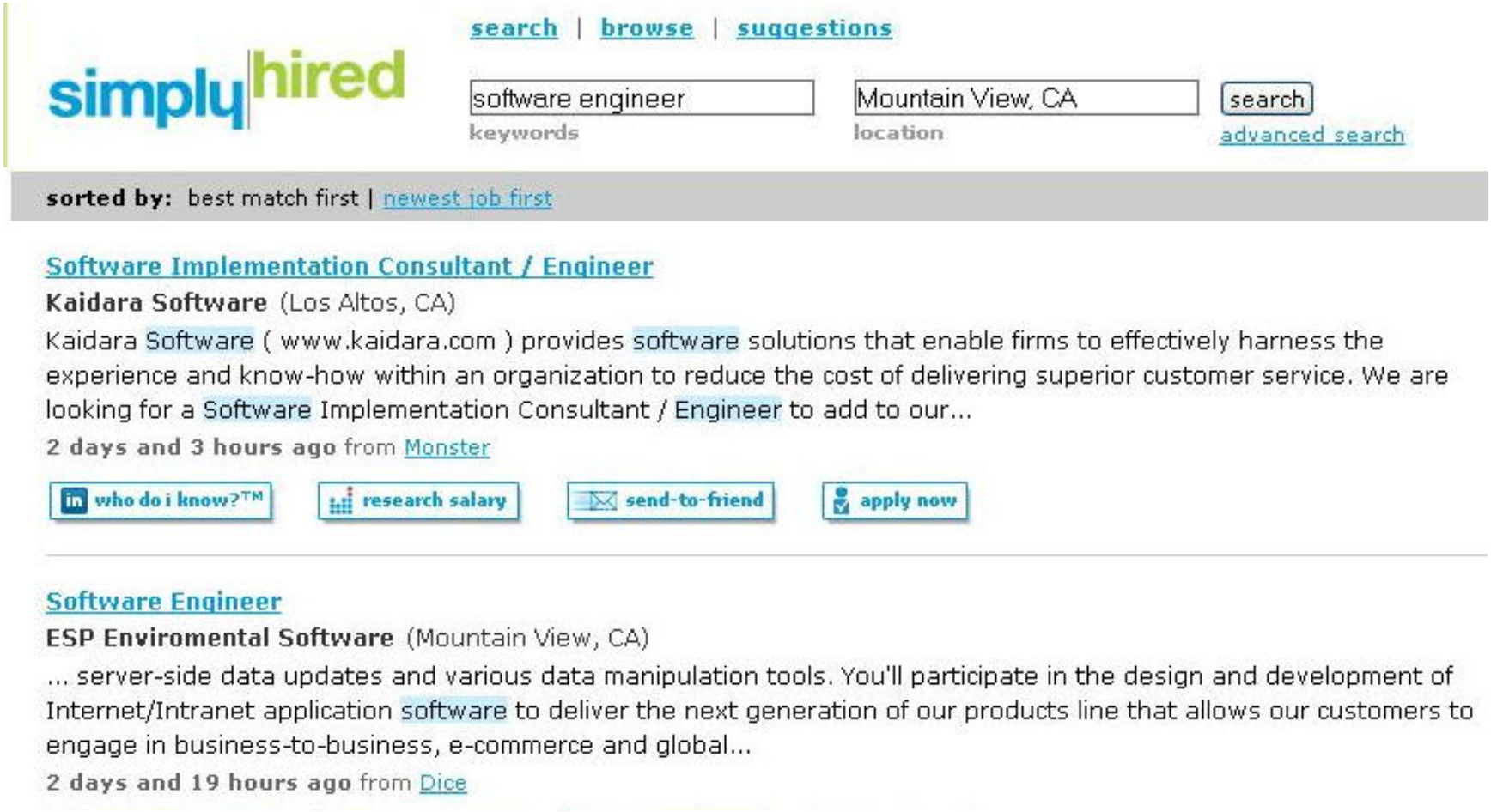
The Long Tail

- 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

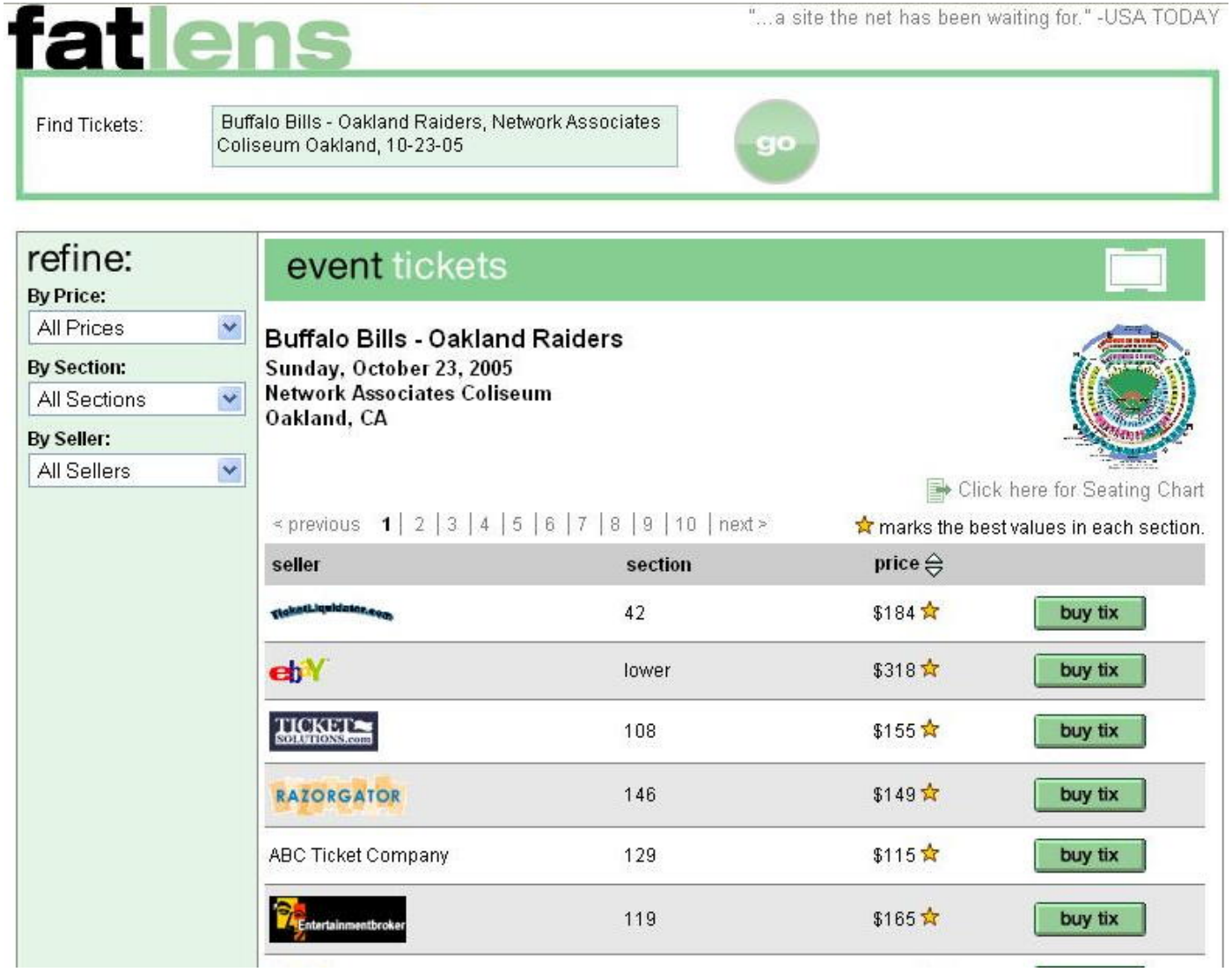
- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

Extracting Structured Data



http://www.simplyhired.com

Extracting structured data

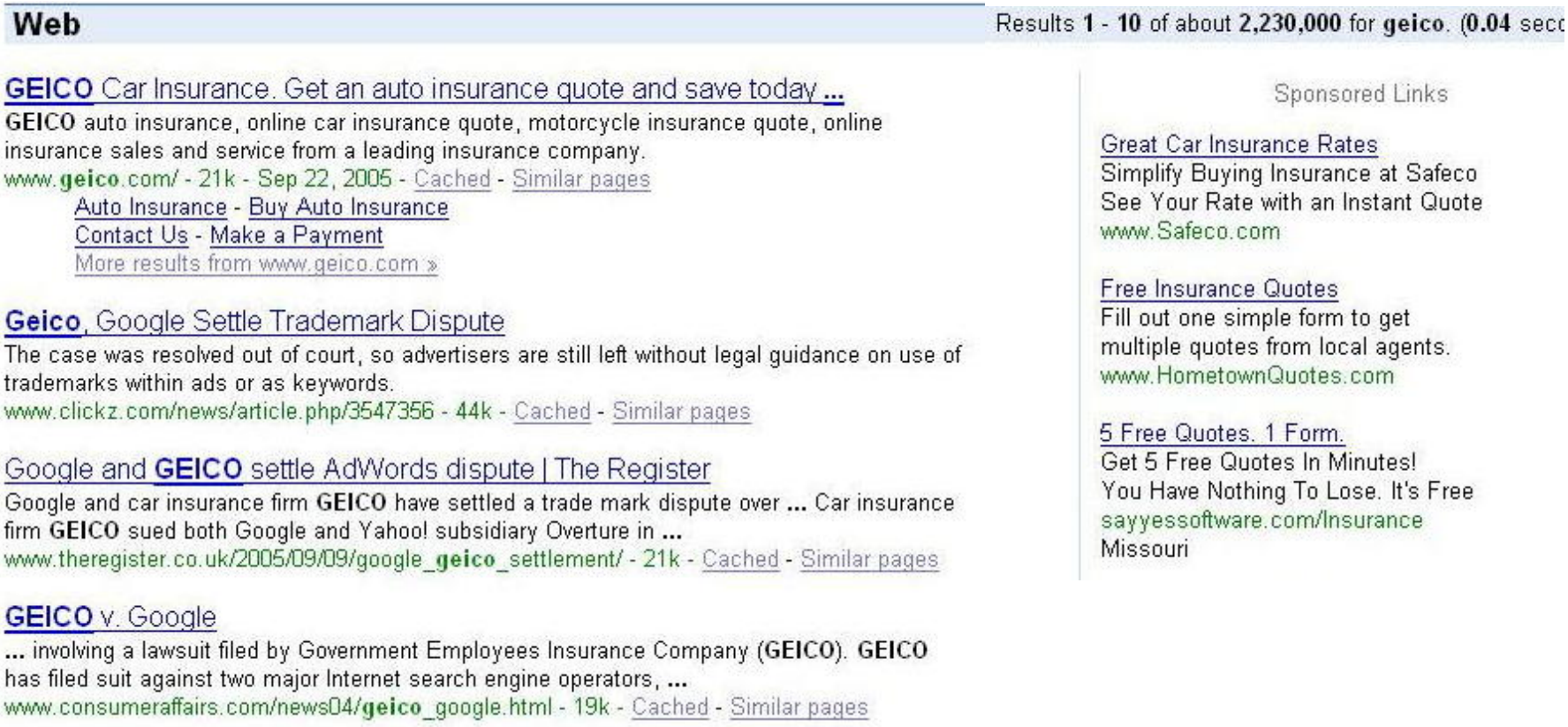


http://www.fatlens.com

Web Mining topics

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

Ads vs. search results



Web Results **1** - **10** of about **2,230,000** for **geico**. (**0.04** seco

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 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-terabye data sets
  - Without breaking the bank!