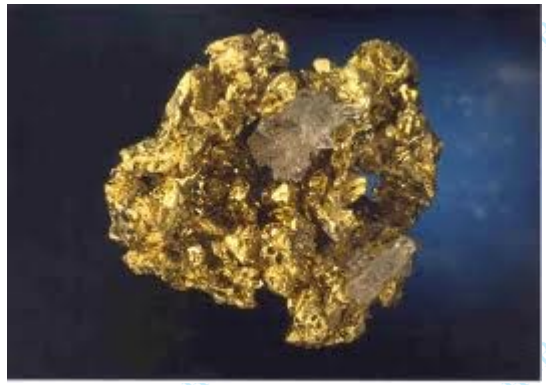
دانشگاه آزاد اسلامی واحد تبریز

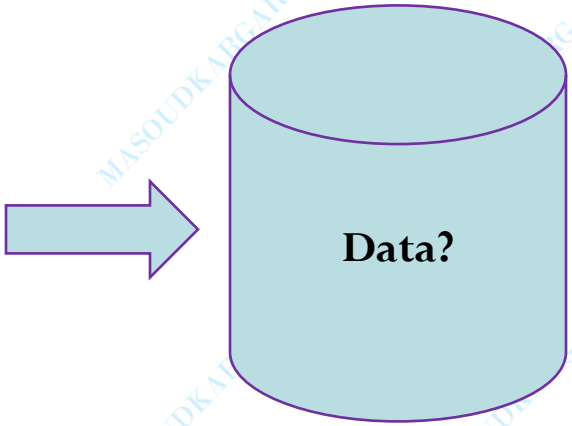
نام درس: داده کاوی

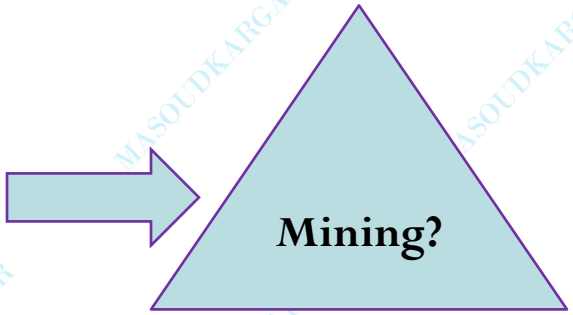
بخش: مقدمه ای بر داده کاوی

نام استاد: دکتر مسعود کارگر

Why You are Here?







The Social Layer in an Instrumented Interconnected World

12+ TBs of tweet data every day

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

? TBs of data every day

100s of millions of GPS enabled devices sold annually

25+ TBs of log data every day

76 million smart meters in 2009… 200M by 2014

2+ billion people on the Web by end 2011

Bigger and Bigger Volumes of Data

- Retailers collect click-stream data from Web site interactions and loyalty card data
  - This traditional POS information is used by retailer for shopping basket analysis, inventory replenishment, +++
  - But data is being provided to suppliers for customer buying analysis
- Healthcare has traditionally been dominated by paper-based systems, but this information is getting digitized
- Science is increasingly dominated by big science initiatives
  - Large-scale experiments generate over 15 PB of data a year and can't be stored within the data center; sent to laboratories
- Financial services are seeing large and large volumes through smaller trading sizes, increased market volatility, and technological improvements in automated and algorithmic trading
- Improved instrument and sensory technology
  - Large Synoptic Survey Telescope's GPixel camera generates 6PB+ of image data per year or consider Oil and Gas industry

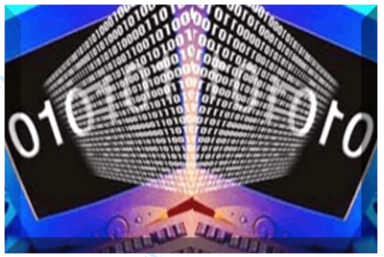
Applications for Big Data Analytics

Smarter Healthcare

Multi-channel

Finance

Log Analysis



Homeland Security

Traffic Control

Telecom

Search Quality

Manufacturing

Trading Analytics

Fraud and Risk

Retail: Churn, NBO

Most Requested Uses of Big Data

- Log Analytics & Storage
- Smart Grid / Smarter Utilities
- RFID Tracking & Analytics
- Fraud / Risk Management & Modeling
- 360° View of the Customer
- Warehouse Extension
- Email / Call Center Transcript Analysis
- Call Detail Record Analysis
- +++

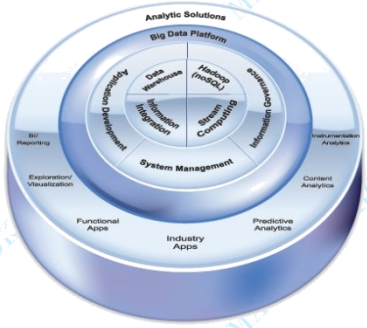
The IBM Big Data Platform

InfoSphere BigInsights
Hadoop-based low latency analytics for variety and volume

Hadoop

Information Integration

InfoSphere Information Server
High volume data integration and transformation



Stream Computing

InfoSphere Streams
Low Latency Analytics for streaming data

MPP Data Warehouse

IBM InfoSphere Warehouse
Large volume structured data analytics

IBM Netezza High Capacity Appliance
Queryable Archive Structured Data

IBM Netezza 1000
BI+Ad Hoc Analytics on Structured Data

IBM Smart Analytics System
Operational Analytics on Structured Data

IBM Informix Timeseries
Time-structured analytics

Big Data Values

Big data can generate significant financial value across sectors

US health care
- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration
- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data
- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail
- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

Manufacturing
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

SOURCE: McKinsey Global Institute analysis

What This course can do for You?

- They expect you are a DM insider!
- How do they know if you are a proficient Data Miner?
  - You know what they are talking about:
    - Glossary: cross-validation, boosting, missing values, sensitivity etc.
  - You know what algorithm solutions exist for their projects
  - You know what software tools/packages are available
  - You can quickly prototype a system using existing or your own code
  - You know how to evaluate the tools
  - You know how to tune or customize the data/tools/algorithms for better performance
  - You know the DM literature/progress in your area (new fancy data mining?

What Data Mining can Do for you

- Commercial:
  - Business intelligence
  - Customer targeting
- Scientific Research
  - Extract hidden patterns from enormous amount of data
  - Material science, Text mining, disease gene discovery

Voice from the Real-world

- Job Ads of Nexttag, San Jose, CA
  - Solid knowledge and hands-on experience in statistics, data clustering, predictive modeling, and/or text classification. Familiar with various modeling techniques such as regression, neural networks, decision trees, SVM, etc. Strong programming skills, especially in C++, Java, and/or Perl.
- **Corporate Analytics & Modeling: eBay Inc.**
  - Reduce losses by analyzing and correlating **fraud patterns** across all companies and suggesting new technologies, techniques and models

    Explore the use of statistical techniques like **machine learning/neural networks, clustering**, link analysis, graph theory and network theory to gain new insights on cross-company data, which in turn result in actionable ways to reduce fraud and risk without compromising business growth
  - Analysis will generally be project based and will often be complex in nature, whereby **large volumes of data** are extracted and synthesized into complex **models** and actionable **recommendations**. Analyses may involve segmentation, profiling, data mining, clustering and predictive modeling.

Voice from the real-world DM

- Washington Mutual Funds: Senior Data Mining Analyst - Customer Behavior Analytics
  - The role will require ability to extract data from various sources & to design/construct complex analysis and communicate that to client as actionable intelligence.
  - He/she will routinely engages in quantitative analysis on many non-standard and unique business problems and uses computer-intensive data mining techniques (decision trees, neural networks, etc.) to deliver actionable output.
  - Ad hoc prototyping skills using multiple techniques to solve a myriad of business scenarios.

Case Study 1: Beat Google!

- Google's Adsense System

Advertiser select Keywords for Ads : cell phone watch

Google Adsense System

Publisher/webmaster's webpages/blog/forum ……. ……phone…. Watch….

?

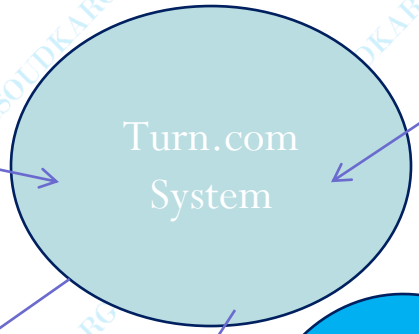
K1, K2, K3, K4….	W1, W2, W3, ….W100	→2/100
T1, T2, K3, T4…	W1, W3, ….W10	→4/100
Show which ads?	W1, W3, ….W10	→Max Profit

Case Study 1: Beat Google! 18M$

- Startup Company Turn.com 's New idea

•Select Keywords is not easy!
•Advertiser DO nothing!
• 100% automatic!
•Advertiser's URL/website as input

Publisher/webmaster's webpages/blog/forum
……
……phone….
Watch….

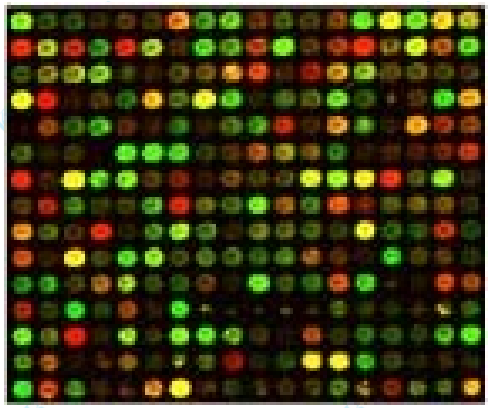


?

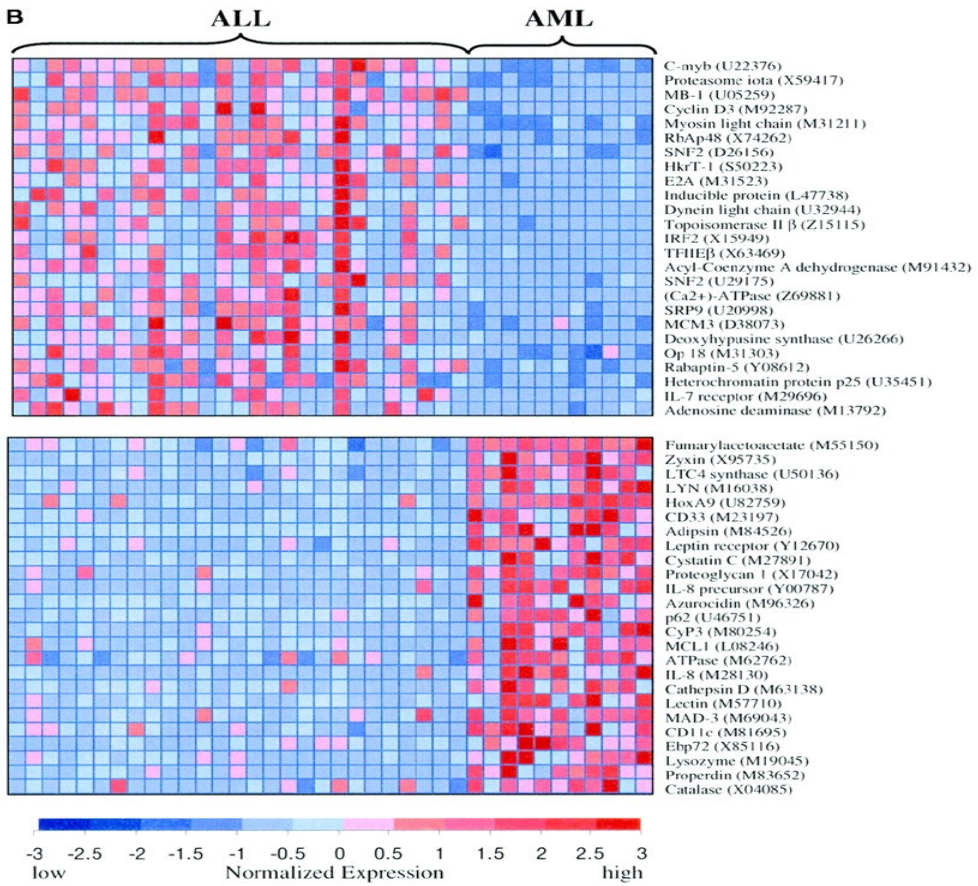
K1, K2, K3, K4….	W1, W2, W3, ….W100+60 variables	→2/100
T1, T2, K3, T4…	W1, W3, ….W10+60 variables	→4/100
Show which ads?	W1, W3, ….W10+60 variables	→Max Profit

DM Case Study 2: Molecular Classification of Cancer

- Acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML)

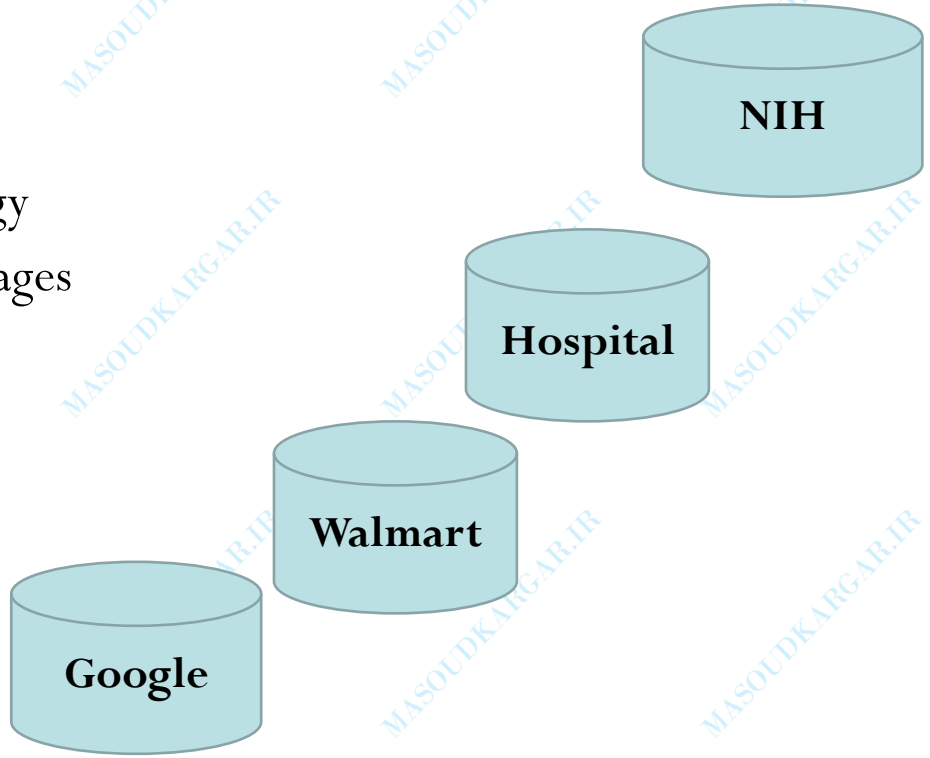


294.934	336.441	280.649	339.055	386.737	299.84
254.397	251.721	233.85	291.528	285.749	255.097
374.489	393.215	397.667	410.991	453.056	317.424
743.2	852.32	814.38	864.983	916.213	724.985
691.036	661.555	610.911	752.899	740.483	666.639
433.061	417.579	369.553	476.694	491.008	423.264
2062.46	2148.22	1975.06	2324.45	2306.12	1969.33
5044.33	4840.89	4881.37	5611.05	5305.85	5107.22
7347.44	6396.52	6819.01	8258.17	7028.87	7091.95
57.4888	64.9644	51.1723	62.0184	58.9424	62.8922
70.0174	63.6578	61.8539	68.3837	67.6043	71.3727
73.943	61.8324	62.1907	67.415	66.0033	60.0997
14.552	16.5048	13.3464	11.5526	15.6981	14.8824
59.4447	68.4437	64.8264	70.4665	63.1142	68.0021
88.6177	85.4299	88.6463	81.0773	90.681	90.4933
33.8383	30.5347	32.9086	31.3637	28.7242	30.8204
37.2913	44.7138	49.0442	31.0727	42.467	37.6109
102.903	102.687	96.2443	109.515	94.7417	98.8921
79.3472	71.5034	68.2442	82.266	75.2751	87.5428
33.9511	30.7839	33.6008	28.6972	28.9501	30.3945
54.4191	52.9862	52.254	56.0972	54.6459	60.6052
86.7367	95.5013	94.9733	78.3674	84.3088	89.1522
50.65	40.3147	40.8639	42.772	43.4843	37.5306



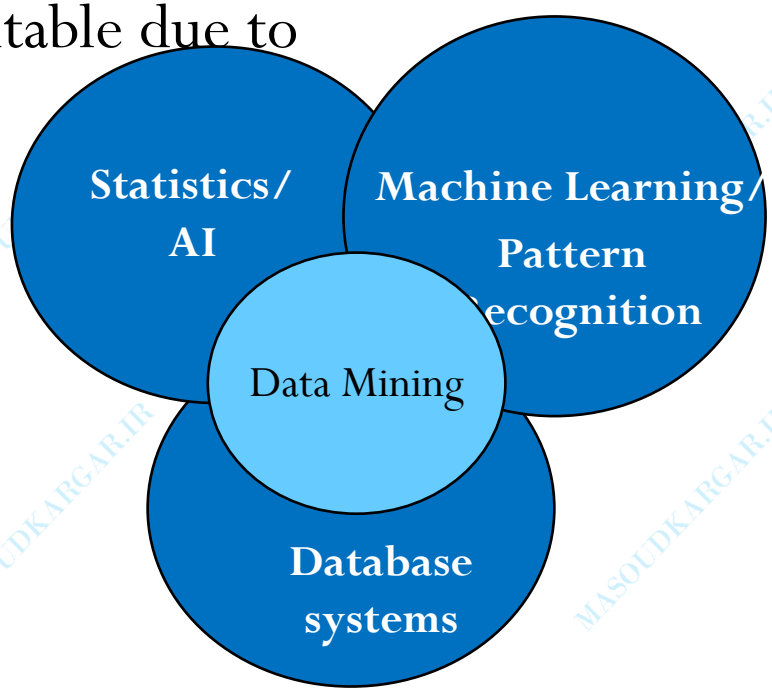
Data mining: What, Who, Why, How?

- What is data mining?
  - Use historical (large-scale) data to uncover regularities and improve future decisions
  - Everybody has some data:
    - Science: physics, chemistry, biology
    - Health care: patients, diseases, images
    - Business: sales, marketing
    - Internet: web
  - What data can you get?



Data mining: what?

- What is the original of Data mining?
- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Data mining: What, Who, Why, How?

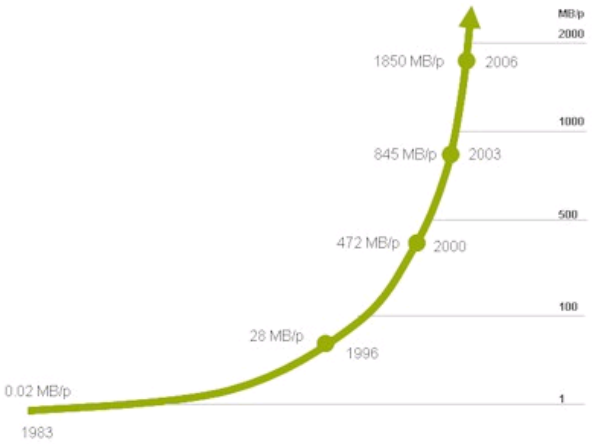
- WHO is doing data mining?

retail, financial, communication, and marketing organizations

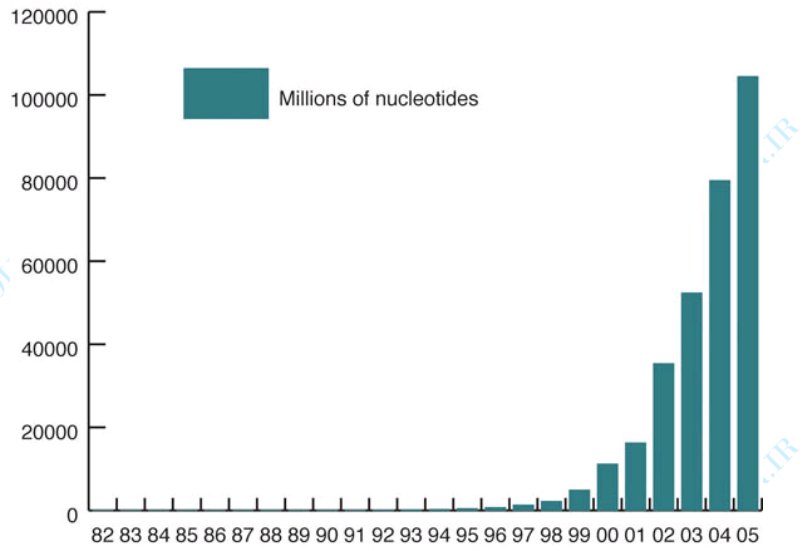
Data mining: What, Who, Why, How?

- Why data mining?
  - information explosion:

    Data → Knowledge/Decision/Understanding/Profit



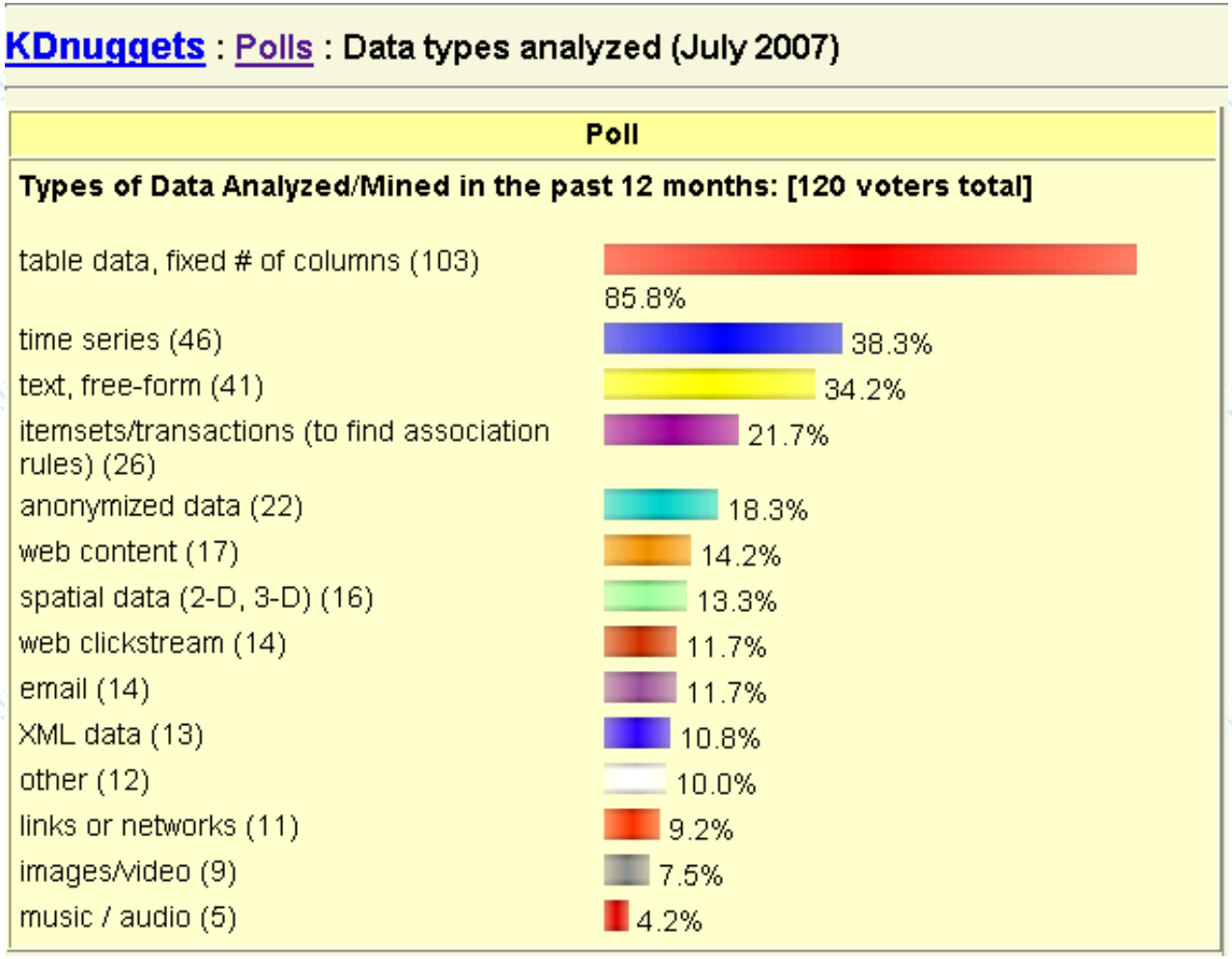
Personal Information storage



Exponential growth of the EMBL DNA sequence database

Data mining: What, Who, Why, How?

Collect Your data



Data mining: What, Who, Why, How?

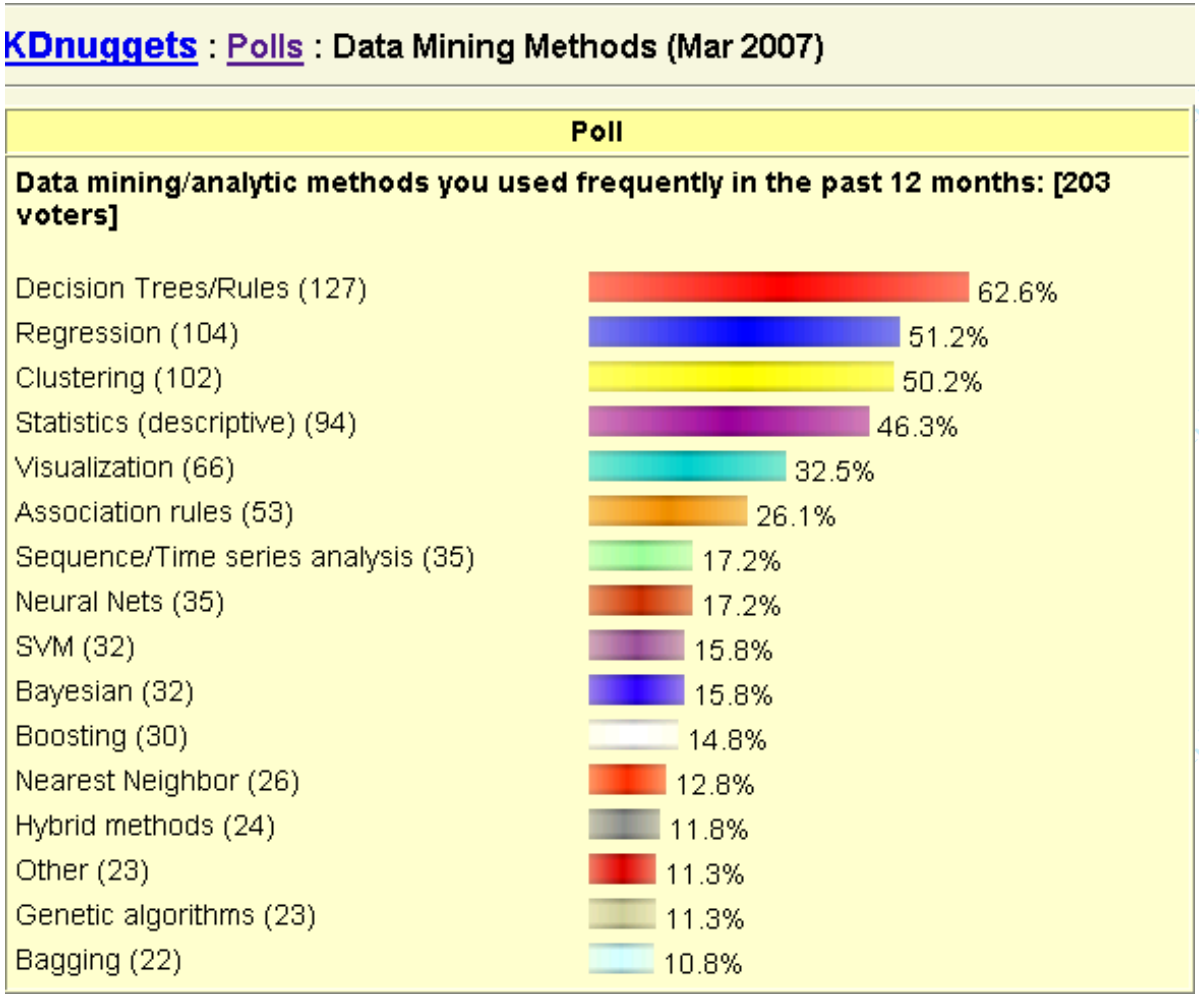
- 2. Determining the patterns you want to mine: data mining tasks
- Two main types of tasks
  - Prediction Methods
    - Use some variables to predict unknown or future values of other variables.
  - Description Methods
    - Find human-interpretable patterns/rules that describe the data.

Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Regression [Predictive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Deviation Detection [Predictive]
- Frequent Subgraph mining [Descriptive]
- …

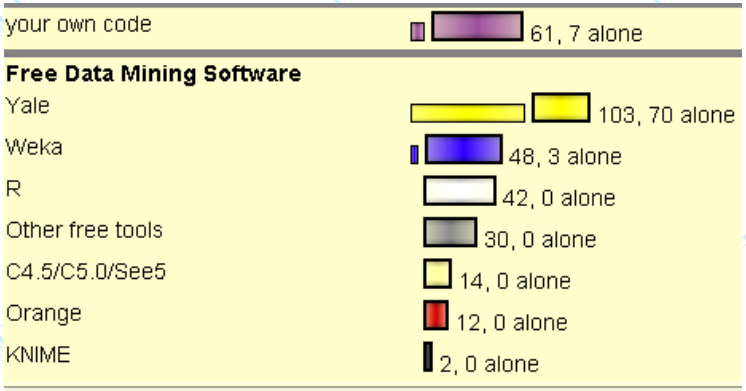
Data mining: What, Who, Why, How?

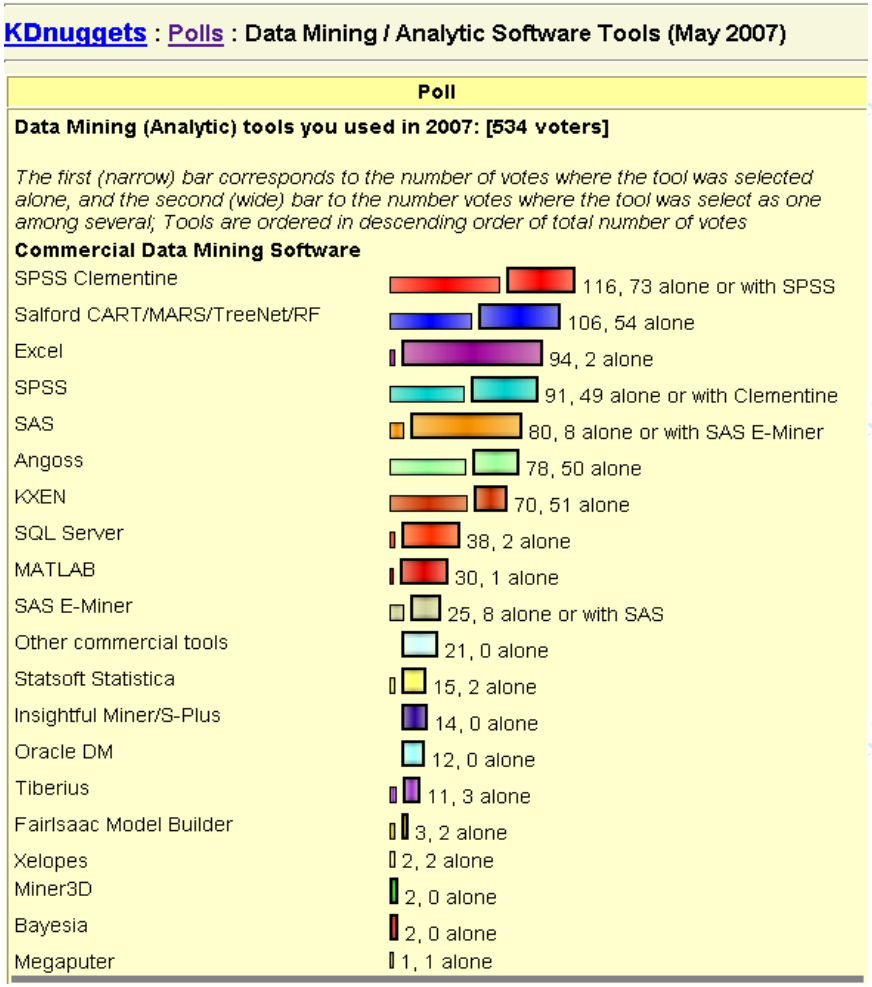
Choose the algorithm(s)



Data mining: How?

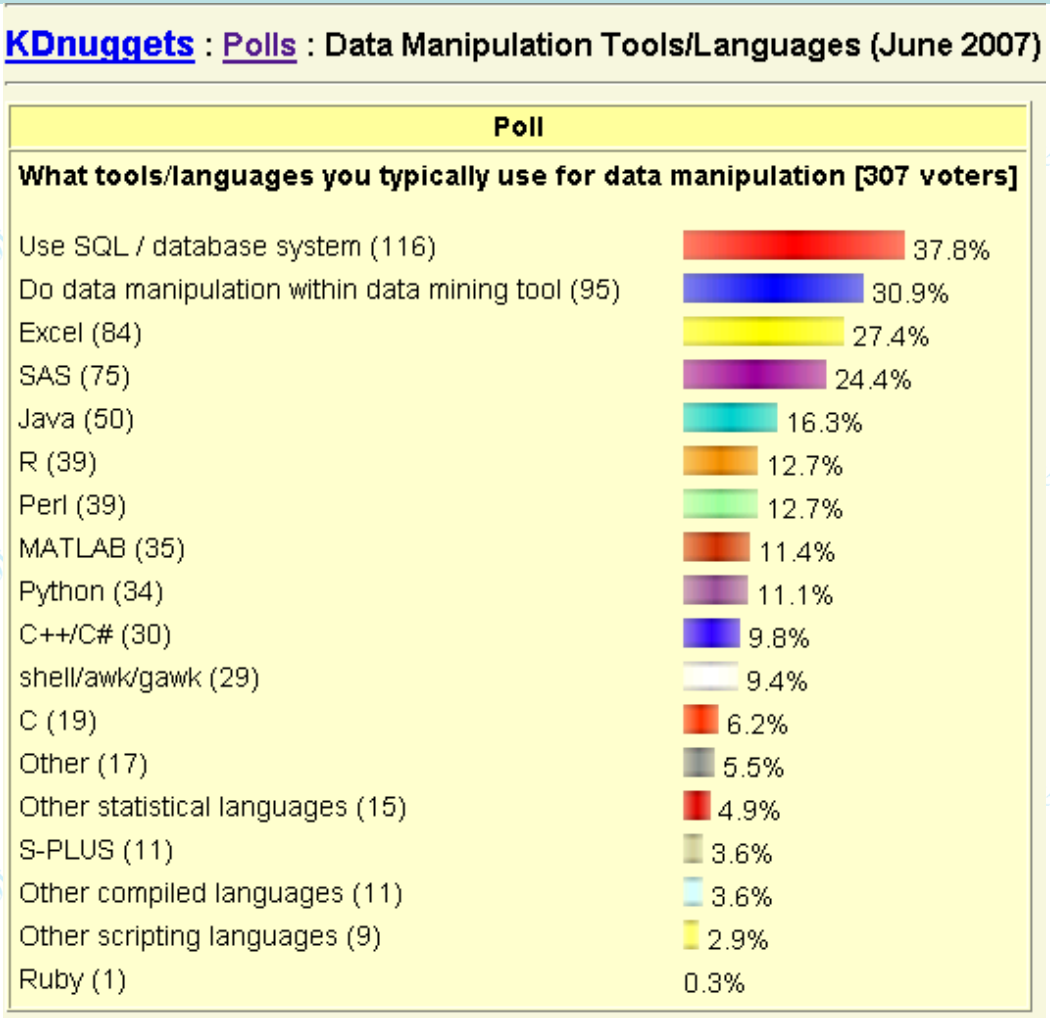
Select existing data mining software/packages



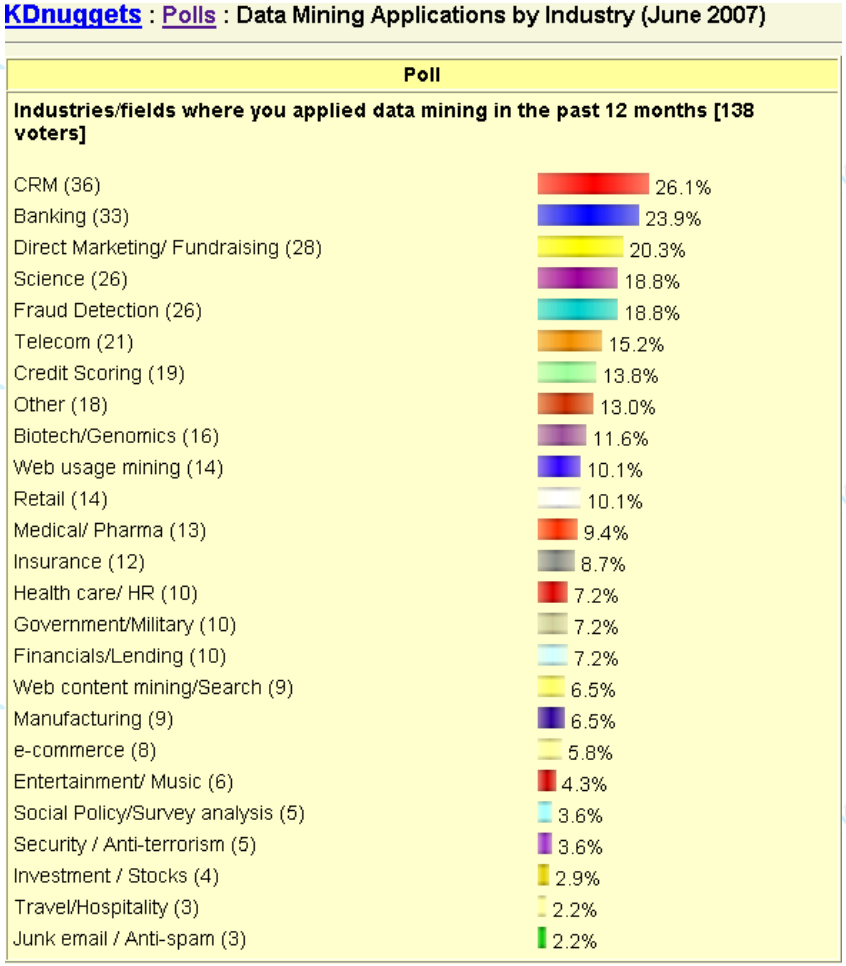


Data mining: What, Who, Why, How?

Choose the implementation platform/programming languages



Data mining: Where to work?



KDnuggets : Polls : Data Mining Applications by Industry (June 2007)

Poll

Industries/fields where you applied data mining in the past 12 months [138 voters]

CRM (36)	26.1%
Banking (33)	23.9%
Direct Marketing/ Fundraising (28)	20.3%
Science (26)	18.8%
Fraud Detection (26)	18.8%
Telecom (21)	15.2%
Credit Scoring (19)	13.8%
Other (18)	13.0%
Biotech/Genomics (16)	11.6%
Web usage mining (14)	10.1%
Retail (14)	10.1%
Medical/ Pharma (13)	9.4%
Insurance (12)	8.7%
Health care/ HR (10)	7.2%
Government/Military (10)	7.2%
Financials/Lending (10)	7.2%
Web content mining/Search (9)	6.5%
Manufacturing (9)	6.5%
e-commerce (8)	5.8%
Entertainment/ Music (6)	4.3%
Social Policy/Survey analysis (5)	3.6%
Security / Anti-terrorism (5)	3.6%
Investment / Stocks (4)	2.9%
Travel/Hospitality (3)	2.2%
Junk email / Anti-spam (3)	2.2%

قدردانی

- Dr. Jianjun Hu
  http://mleg.cse.sc.edu/edu/csce822/
- University of South Carolina
- Department of Computer Science and Engineering