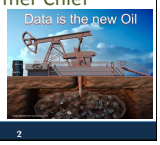
## Data Mining


Machine Learning, Data Mining, and Knowledge Discovery: An Introduction



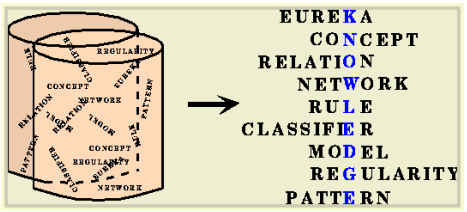
## Some recent happenings …

- CodeSprint2 (Jan 2012)
  - http://cs2.interviewstreet.com/
- Career of the future: Data Scientist
  - http://mashable.com/2012/01/13/career-of-the-future-data-scientist-infographic/
  - "Data is the new oil," Andreas Weigend, Head of the Social Data Lab at Stanford and the former Chief Scientist at Amazon






## My early interpretation of Data Mining






## Outline

- **Introduction: Data Flood**
- Data Mining Application Examples
- Data Mining & Knowledge Discovery
- Data Mining Tasks



## Data Flood

- More data is generated:
  - Bank, telecom, other business transactions ...
  - Scientific data: astronomy, biology, etc
  - Web, text, and e-commerce



## Large databases

- Commercial databases:
  - Winter Corp. 2003 Survey: France Telecom has largest decision-support DB, ~30TB; AT&T ~ 26 TB
- Web
  - Google searches 4+ Billion pages, many hundreds TB
  - Internet Archive (http://www.archive.org), 4.5 petabyte (2009)
    - The "Way back" machine
    - Alexa internet archive: 7 years of data, 500 TB

## Large Databases

- Astronomy
  - Sloan Digital Sky Survey (SDSS) – DR8 (http://www.sdss.org/)
  - National Virtual Observatory (NVO) (http://www.us-vo.org/)
- Biology/Medecine
  - National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/)
  - PubMed has over 21.47 million records (Jan 2012)



## From terabytes to exabytes to …

- UC Berkeley 2003 estimate: 5 exabytes (5 million terabytes) of new data was created in 2002.
  - www.sims.berkeley.edu/research/projects/how-much-info-2003/
- US produces ~40% of new stored data worldwide
- 2006 estimate: 161 exabytes (IDC study)
  - www.usatoday.com/tech/news/2007-03-05-data_N.htm
- 2010 estimate: 988 exabytes



## Largest Databases in 2005

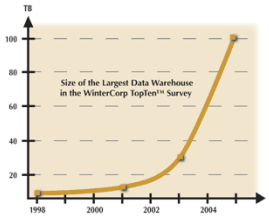
Winter Corp. 2005 Commercial Database Survey:

1. Max Planck Inst. for Meteorology , 222 TB
2. Yahoo ~ 100 TB (Largest Data Warehouse)
3. AT&T ~ 94 TB

www.wintercorp.com/VLDB/2005_TopTen_Survey/TopTenWinners_2005.asp



## Data Growth




From 2003 to 2005, the size of the largest database TRIPLED!

Conservative estimation: ~30% growth rate
Very little data will ever be looked at by a human



## Data Growth

- Surveys sees major expansion of world's data centers ( NYTimes, Sept 2011)
  - Construction of new data centers to grow 19%

Knowledge Discovery is **NEEDED** to make sense and use of data.



## Outline

- Introduction: Data Flood
- **Data Mining Application Examples**
- Data Mining & Knowledge Discovery
- Data Mining Tasks

## Machine Learning / Data Mining Application areas

- Science
  - astronomy, bioinformatics, drug discovery, …
- Business
  - CRM (Customer Relationship management), fraud detection, e-commerce, manufacturing, sports/entertainment, telecom, targeted marketing, health care, …
- Web:
  - search engines, advertising, web and text mining, …
- Government
  - Surveillance, crime detection, profiling tax cheaters, …



## Application Areas

What do you think are some of the most important and widespread business applications of Data Mining?



## Data Mining for Customer Modeling

- Customer Tasks:
  - attrition prediction
  - targeted marketing:
    - cross-sell, customer acquisition
  - credit-risk
  - fraud detection
- Industries
  - banking, telecom, retail sales, …



## Customer Attrition: Case Study

- Situation: Attrition rate at for mobile phone customers is around 25-30% a year!
- With this in mind, what is our task?
  - Assume we have customer information for the past N months.



## Customer Attrition: Case Study

- Task:
  - Predict who is likely to attrite next month.
  - Estimate customer value and what is the cost-effective offer to be made to this customer.



## Customer Attrition Results

- Verizon Wireless built a customer data warehouse
- Identified potential attriters
- Developed multiple regional models
- Targeted customers with high propensity to accept the offer
- Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers) (Reported in 2003)

## Assessing Credit Risk: Case Study

- Situation: Person applies for a loan
- Task: Should a bank approve the loan?
- Note: People who have the best credit don't need the loans, and people with worst credit are not likely to repay. Bank's best customers are in the middle.



## Credit Risk - Results

- Banks develop credit models using variety of machine learning methods.
- Mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan
- Widely deployed in many countries



## e-commerce

- A person buys a book (product) at Amazon.com

What is the task?



## Successful e-commerce – Case Study

- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**"
- Recommendation program is quite successful



## Genomic Microarrays – Case Study

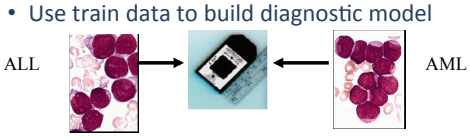
Given microarray data for a number of samples (patients), can we

- Accurately diagnose the disease?
- Predict outcome for given treatment?
- Recommend best treatment?



## Example: ALL/AML data

- 38 training cases, 34 test, ~ 7,000 genes
- 2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)
- Use train data to build diagnostic model




Results on test data:
33/34 correct, 1 error may be mislabeled

## Other Examples

- Credit Card Fraud Detection
- Detection of Money laundering
  - FAIS (US Treasury)
- Securities Fraud
  - NASDAQ KDD system
- Phone fraud
  - AT&T, Bell Atlantic, British Telecom/ MCI
- Bio-terrorism detection at Salt Lake Olympics 2002



## Data Mining and Privacy

- in 2006, NSA (National Security Agency) was reported to be mining years of call info, to identify terrorism networks
- Social network analysis has a potential to find networks
- Invasion of privacy – do you mind if your call information is in a gov database?
- What if NSA program finds one real suspect for 1,000 false leads ? 1,000,000 false leads?



## Data Mining and Privacy

- require knowledge-based decisions
- have a changing environment
- have sub-optimal current methods
- have accessible, sufficient, and relevant data
- provides high payoff for the right decisions!

Privacy considerations important if personal data is involved



## Outline

- Introduction: Data Flood
- Data Mining Application Examples
- **Data Mining & Knowledge Discovery**
- Data Mining Tasks



## Knowledge Discovery

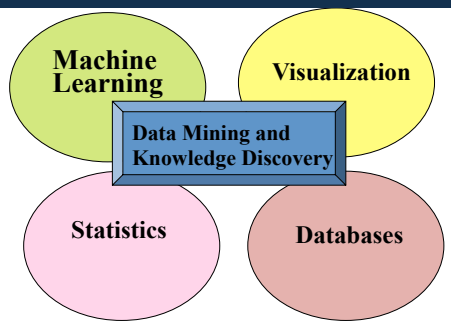
Knowledge Discovery in Data is the *non-trivial* process of identifying

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996



## Related Disciplines



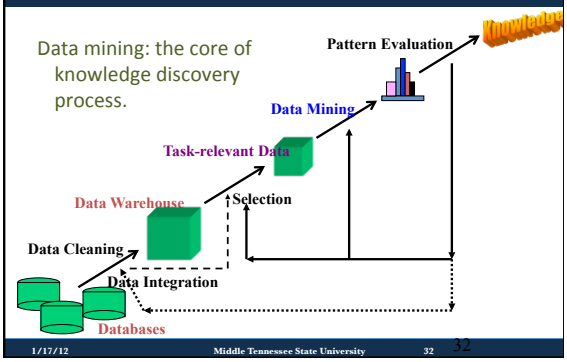
## Statistics, Machine Learning, Data Mining

- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy



## Knowledge discovery process flow

Data mining: the core of knowledge discovery process.





32

## Other Names

- Data Fishing, Data Dredging: 1960-
  - used by Statistician (as bad name)
- Data Mining :1990 --
  - used DB, business
  - in 2003 – bad image because of TIA
- Knowledge Discovery in Databases (1989-)
  - used by AI, Machine Learning Community
- also Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, ...

**Currently:**
**Data Mining and Knowledge Discovery from Databases are used interchangeably**



## Outline

- Introduction: Data Flood
- Data Mining Application Examples
- Data Mining & Knowledge Discovery
- **Data Mining Tasks**



## Data Mining Techniques

- **Classification:** predicting an item class
- **Clustering:** finding clusters in data
- **Associations:** e.g. A & B & C occur frequently
- **Visualization:** to facilitate human discovery
- **Summarization:** describing a group
- **Deviation Detection**: finding changes
- Estimation: predicting a continuous value
- Link Analysis: finding relationships
- ...



## Data Mining Functionalities (1)

- Concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Association (correlation and causality)
  - Multi-dimensional vs. single-dimensional association
  - age(X, "20..29") ^ income(X, "20..29K") à buys(X, "PC") [support = 2%, confidence = 60%]
  - contains(T, "computer") à contains(x, "software") [1%, 75%]

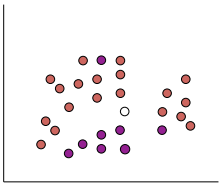
## Data Mining Functionalities (2)

- Classification and Prediction
  - Finding models (functions) that describe and distinguish classes or concepts for future prediction
  - E.g., classify countries based on climate, or classify cars based on gas mileage
  - Presentation: decision-tree, classification rule, neural network
  - Prediction: Predict some unknown or missing numerical values



## Classification

**Learn a method for predicting the instance class from pre-labeled (classified) instances**



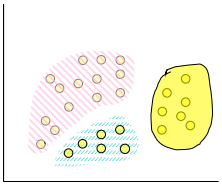
Many approaches:
Statistics,
Decision Trees,
Neural Networks,
...



## Clustering

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity

**Find "natural" grouping of instances given un-labeled data**





## Data Mining Functionalities (3)

- Outlier analysis
  - Outlier: a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses



## Are All the "Discovered" Patterns Interesting?

- A data mining system/query may generate thousands of patterns, not all of them are interesting.
  - Suggested approach: Human-centered, query-based, focused mining
- Interestingness measures: A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm



## Are All the "Discovered" Patterns Interesting?

- Objective vs. subjective interestingness measures:
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

## Summary

- Technology trends lead to data flood
  - data mining is needed to make sense of data
- Data Mining has many applications
- Knowledge Discovery Process
- Data Mining Tasks
  - classification, clustering, …



## Data mining challenges

- ACM KDD CUP (http://www.sigkdd.org/kddcup/index.php)
- Netflix Prize (http://www.netflixprize.com/)