Controversial Issues

- Data mining (or simple analysis) on people may come with a profile that would raise controversial issues of
  - Discrimination
  - Privacy
  - Security
- Examples:
  - Should males between 18 and 35 from countries that produced terrorists be singled out for search before flight?
  - Can people be denied mortgage based on age, sex, race?
  - Women live longer. Should they pay less for life insurance?

Data Mining and Discrimination

- Can discrimination be based on features like sex, age, national origin?
- In some areas (e.g. mortgages, employment), some features cannot be used for decision making
- In other areas, these features are needed to assess the risk factors
  - E.g. people of African descent are more susceptible to sickle cell anemia

Data Mining and Privacy

- Can information collected for one purpose be used for mining data for another purpose
  - In Europe, generally no, without explicit consent
  - In US, generally yes
- Companies routinely collect information about customers and use it for marketing, etc.
- People may be willing to give up some of their privacy in exchange for some benefits

Data Mining with Privacy

- Data Mining looks for patterns, not people!
- Technical solutions can limit privacy invasion
  - Replacing sensitive personal data with anon. ID
  - Give randomized outputs
    - return salary + random()
    - …

Data Mining and Security Controversy in the News

- TIA: Terrorism (formerly Total) Information Awareness Program –
  - DARPA program closed by Congress, Sep 2003
  - some functions transferred to intelligence agencies
- CAPPS II – screen all airline passengers
  - controversial
- …
- Invasion of Privacy or Defensive Shield?

Criticism of analytic approach to Threat Detection:

Data Mining will

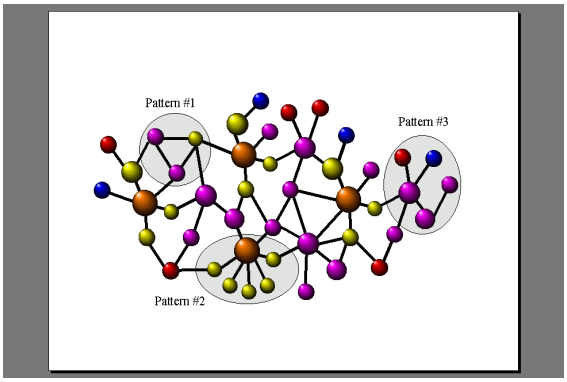
- invade privacy
- generate millions of false positives

But can it be effective?

Is criticism sound ?

- Criticism: Databases have 5% errors, so analyzing 100 million suspects will generate 5 million false positives
- Reality: Analytical models correlate many items of information to reduce false positives.
- Example: Identify one biased coin from 1,000.
  - After one throw of each coin, we cannot
  - After 30 throws, one biased coin will stand out with high probability.
  - Can identify 19 biased coins out of 100 million with sufficient number of throws

Another Approach: Link Analysis



Can Find Unusual Patterns in the Network Structure

Analytic technology can be effective

- Combining multiple models and link analysis can reduce false positives
- Today there are millions of false positives with manual analysis
- Data mining is just one additional tool to help analysts
- Analytic technology has the potential to reduce the current high rate of false positives

Data Mining and Society

- No easy answers to controversial questions
- Society and policy-makers need to make an educated choice
  - Benefits and efficiency of data mining programs vs. cost and erosion of privacy

Data Mining Future Directions

- Currently, most data mining is on flat tables
- Richer data sources
  - text, links, web, images, multimedia, knowledge bases
- Advanced methods
  - Link mining, Stream mining, …
- Applications
  - Web, Bioinformatics, Customer modeling, …

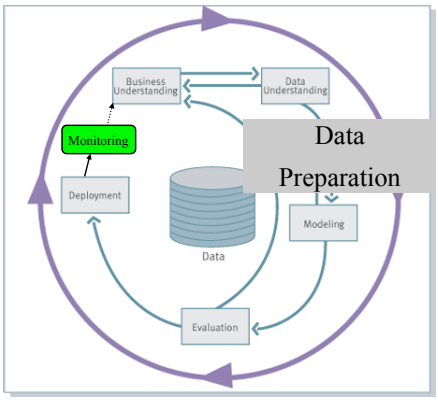
Challenges for Data Mining

- Technical
  - tera-bytes and peta-bytes
  - complex, multi-media, structured data
  - integration with domain knowledge
- Business
  - finding good application areas
- Societal
  - Privacy issues

Data Mining Central Quest

Find true patterns and avoid *overfitting* (false patterns due to randomness)

Knowledge Discovery Process



Start with Business (Problem) Understanding

Data Preparation usually takes the most effort

Knowledge Discovery is an Iterative Process

Key Ideas

- Avoid Overfitting!
- Data Preparation
  - catch false predictors
  - evaluation: train, validate, test subset
- Classification: C4.5, Bayes, …
- Evaluation: Lift, ROC, …
- Clustering, Association, Other tasks
- Knowledge Discovery is a Process