Data Mining OGSA Data Use Case

Susan Malaika

Data Mining

Data Mining is the automated discovery of hidden patterns and information in large volumes of data using techniques from statistics and pattern recognition.

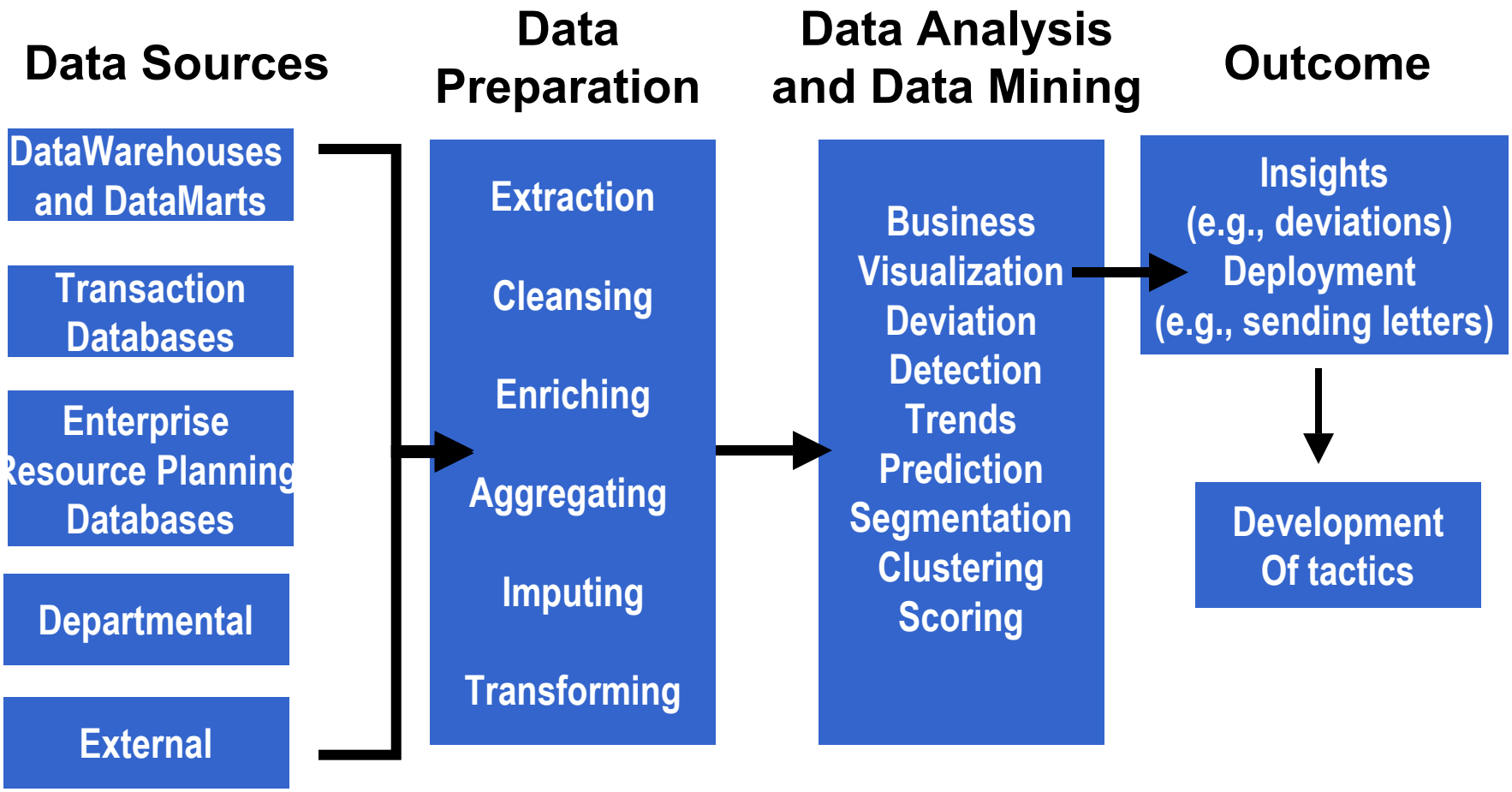
An example of a Data Mining Application:

Database marketing is a technique where you use all information available within your company's database as well as any external information to improve your marketing efforts.

Data Mining in Distribution

- Retailers often ask
  - How do I predict (forecast) demand – so as to reduce stock-outs?
  - Which new products will appeal to my established customers?
  - To whom should I target promotions for specific products or offerings?
  - Which products are likely cross-sell candidates for a given product?
  - Which products are potential (and profitable) substitutes for a given product?
  - How can we detect shrinkage reliably enough to warrant investigation?
- In recent years, data mining techniques have been developed that augment statistical analysis to address these questions

A Data Mining Process



Mining identifies key indicators

Data Mining Methods
Discovering

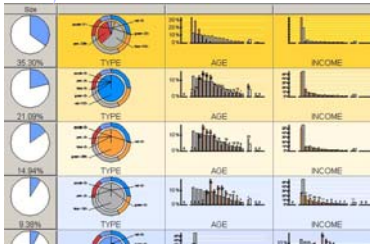
- Discovering unknown Patterns
  - Clustering
    - Grouping data records into subsets where items in subsets are more similar to one another – Demographics techniques
      - e.g., Most males get married between the ages of 30 to 35
  - Association Discovery
    - Complementary behavior where one action leads to another action at the same time.
      - e.g., shoppers who purchase milk will also purchase bread or cheese
  - Sequential Pattern Discovery
    - One action will lead to another action within a time frame.
      - e.g, applicants of Citibank VISA card will apply for ReadyCredit within 30 days of applications.

Data Mining Methods
Predicting

- Predicting the future based on past patterns
  - Classification
    - Deducing the class of groups who will behave in a certain way. Using decision tree deductions.
      - e.g., Young urban professional tends to buy expensive sports cars, while luxury sedans are purchased by the elderly wealthy person.
  - Value Prediction
    - Detect and account for interesting sequence of information in data records.
      - e.g., Sales of car increases sharply before the Chinese New Year period
  - Similar Time Sequence Discovery
    - One action cause a reaction at the same time
      - e.g.,: prices of hot chocolate increases, if the prices of milo reminds the same, its sales increases. (Price wars, discounting)

A Data Mining Process

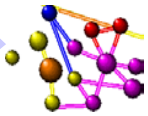
[2] Define model and run the analysis (CPU Intensive)



[3] Visualize

Models & scores used in Business Intelligence Tools

Models & scores used in Applications



Models

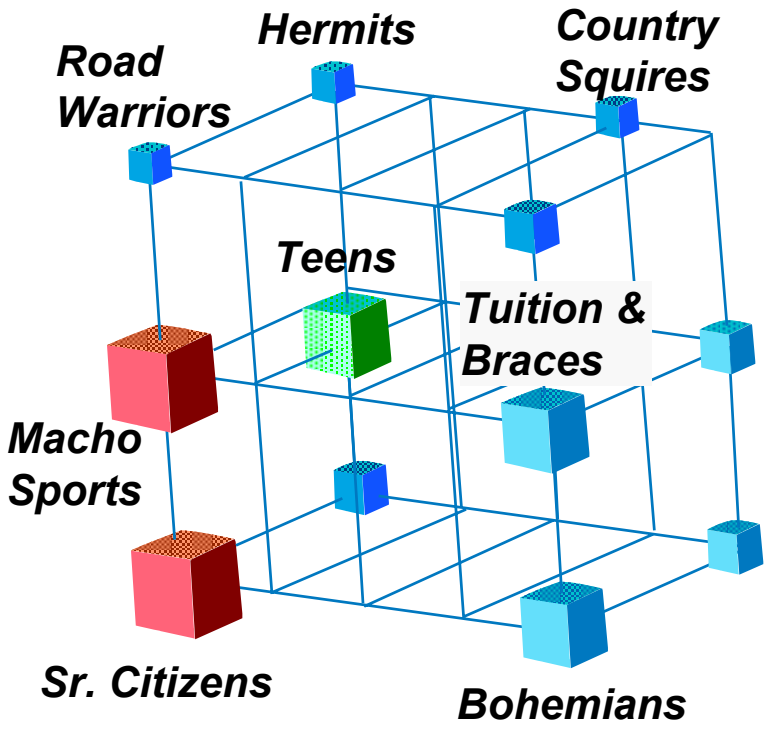
[4] Score (IO intensive)
XML (PMML)

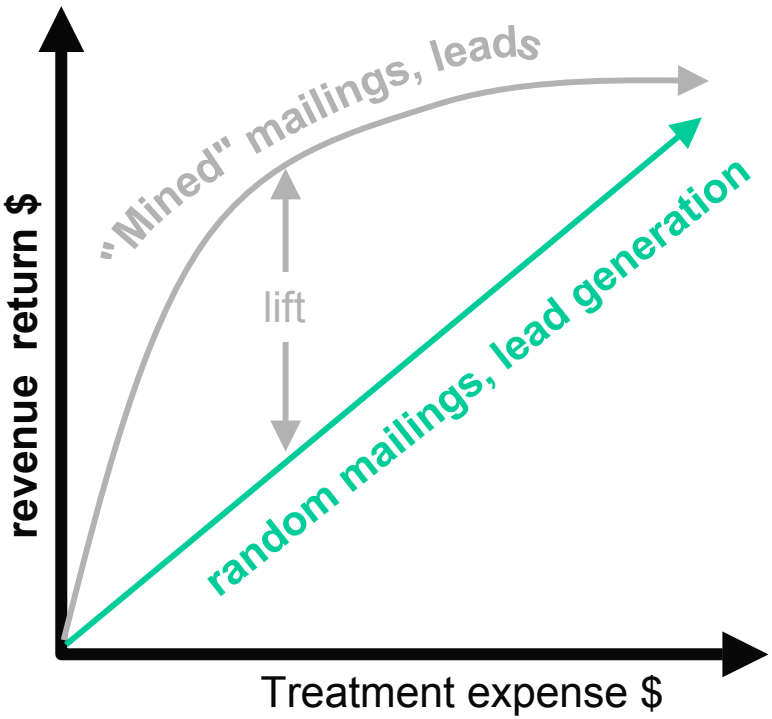
[1] Extract and prepare data

Data Warehouse

For info on PMML see http://www.dmg.org/

Typical Business Uses





Clustering & Segmentation (Group similar behavior)

Prediction (e.g., Need info about previous mailing campaigns)