What is Data Mining?

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

Knowledge Discovery in Databases (KDD)
## What is (not) Data Mining?

<table>
<thead>
<tr>
<th>What is not Data Mining?</th>
<th>What is Data Mining?</th>
</tr>
</thead>
<tbody>
<tr>
<td>– Look up phone number in phone directory</td>
<td>– Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)</td>
</tr>
<tr>
<td>– Query a Web search engine for information about “Amazon”</td>
<td>– Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)</td>
</tr>
</tbody>
</table>
Data Explosion

“We are drowning in data, but starving for knowledge”

“The amount of data stored in various media has doubled in three years, from 1999 to 2002. The amount of data put into storage in 2002, five exabytes (one quintillion bytes), was equal to the contents of a half a million new libraries, each containing a digitised version of the print collection of the entire US Library of Congress”

(Lyman and Varian, UC Berkeley, 2003)
Scale of Data

<table>
<thead>
<tr>
<th>Organization</th>
<th>Scale of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walmart</td>
<td>~ 20 million transactions/day</td>
</tr>
<tr>
<td>Google</td>
<td>&gt; 4.2 billion Web pages</td>
</tr>
<tr>
<td>Yahoo</td>
<td>~10 GB Web data/hr</td>
</tr>
<tr>
<td>NASA satellites</td>
<td>~ 1.2 TB/day</td>
</tr>
<tr>
<td>NCBI GenBank</td>
<td>~ 22 million genetic sequences</td>
</tr>
<tr>
<td>France Telecom</td>
<td>29.2 TB</td>
</tr>
<tr>
<td>UK Land Registry</td>
<td>18.3 TB</td>
</tr>
<tr>
<td>AT&amp;T Corp</td>
<td>26.2 TB</td>
</tr>
</tbody>
</table>

“The great strength of computers is that they can reliably manipulate vast amounts of data very quickly. Their great weakness is that they don’t have a clue as to what any of that data actually means”

Why Mine Data?

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

From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”
Why Mine Data?

“More often, data mining yields unexpected nuggets of information that open the company’s eyes to new markets, new ways of reaching customers and new ways of doing business.”

[M.Betts, ComputerWorld, April 2003]

“The concept of data mining is one of those things that applies across the spectrum, from business looking at financial data to scientists looking at scientific data... Homeland Security Department will mine data for information from biological sensors, for example... Once we do get a dense enough sensor network out there, we are going to be inundated with data and a lot of the data mining techniques that have been used in industry ... particularly the financial one, will be applied to those data sets.”

[D.Bolka, Director of HSARPA, 2004]
## Data Mining Applications

<table>
<thead>
<tr>
<th>Application</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Intelligence</td>
<td>Customer purchase history, credit card information</td>
<td>What products are frequently bought together by customers</td>
</tr>
<tr>
<td>Collaborative Filtering</td>
<td>User-provided ratings for movies, or other products</td>
<td>Recommended movies or other products</td>
</tr>
<tr>
<td>Network Intrusion Detection</td>
<td>TCPdump trace or Cisco NetFlow logs</td>
<td>Anomaly score assigned to each network connection</td>
</tr>
<tr>
<td>Web search</td>
<td>Query provided by user</td>
<td>Documents ranked based on their relevance to user input</td>
</tr>
<tr>
<td>Medical Diagnosis</td>
<td>Patient history, physiological, and demographic data</td>
<td>Diagnosis of patient as sick or healthy</td>
</tr>
<tr>
<td>Climate Research</td>
<td>Measurements from sensors aboard NASA Earth observing satellites</td>
<td>Relationships among Earth Science events, trends in time series, etc</td>
</tr>
<tr>
<td>Process Mining</td>
<td>Event-based data from workflow logs</td>
<td>Discrepancies between prescribed models and actual process executions</td>
</tr>
</tbody>
</table>
Origins 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 Tasks

- Predictive Methods
  - Use some variables to predict unknown or values of other variables.

- Descriptive Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996
Data Mining Tasks...

- Predictive Modeling [Predictive]
  - Classification
  - Regression

- Clustering [Descriptive]

- Pattern Discovery [Descriptive]
  - Association Rule Mining
  - Sequential Pattern Discovery
  - Tree/Subgraph Mining

- Anomaly Detection [Predictive]
Classification: Definition

- **Given:**
  - A collection of records (training set)
    - Each record contains a set of attributes
    - One of the *nominal* attributes is designated as the *class attribute*

- **Task:**
  - Find a model for the class attribute as a function of other attributes
  - Use the model to predict the class for previously unseen records

- **Goal:**
  - Model should accurately predict the class for previously unseen records
    - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.
## Illustrating Classification

![Table and Diagram]

<table>
<thead>
<tr>
<th>Tid</th>
<th>Attrib1</th>
<th>Attrib2</th>
<th>Attrib3</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Large</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Medium</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Small</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Medium</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Large</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Medium</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Large</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Small</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Medium</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Small</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Test Set

<table>
<thead>
<tr>
<th>Tid</th>
<th>Attrib1</th>
<th>Attrib2</th>
<th>Attrib3</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>No</td>
<td>Small</td>
<td>55K</td>
<td>?</td>
</tr>
<tr>
<td>12</td>
<td>Yes</td>
<td>Medium</td>
<td>80K</td>
<td>?</td>
</tr>
<tr>
<td>13</td>
<td>Yes</td>
<td>Large</td>
<td>110K</td>
<td>?</td>
</tr>
<tr>
<td>14</td>
<td>No</td>
<td>Small</td>
<td>95K</td>
<td>?</td>
</tr>
<tr>
<td>15</td>
<td>No</td>
<td>Large</td>
<td>67K</td>
<td>?</td>
</tr>
</tbody>
</table>

### Training Set

- Learn Classifier
- Model
Classification: Applications

- **Direct marketing**
  - Predict consumers who will most likely buy a new product based on their demographic, lifestyle, and previous buying behavior

- **Spam detection**
  - Categorize email messages as spam or non-spam based on message header and content

- **Functional classification of proteins**
  - Assign sequences of unknown proteins to their respective functional classes

- **Galaxy classification**
  - Classify galaxies based on their image features

- **Automated target recognition**
  - Identify target objects (enemy tanks, trucks, etc) based on signals gathered from sensor arrays
Classifying Galaxies

**Early**

Class:
- Stages of Formation

**Intermediate**

**Late**

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Courtesy: http://aps.umn.edu
Regression: Definition

- **Given:**
  - A collection of records (training set)
    - Each record contains a set of attributes
    - One of the continuous-valued attributes is designated as the target variable

- **Task:**
  - Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency

- Greatly studied in statistics, neural network fields
Regression: Applications

- **Marketing**
  - Predicting sales amounts of new product based on advertising expenditure

- **Earth Science**
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc

- **Finance**
  - Time series prediction of stock market indices

- **Agriculture**
  - Predicting crop yield based on soil fertility and weather information

- **Socio-economy**
  - Predicting electricity consumption in single family homes based on outdoor temperatures
Regression Analysis for Climate Data

SOI: a climate index related to the El-Nino phenomenon

Using SOI to predict precipitation in Australia
Clustering: Definition

- **Given:**
  - A set of data points
  - Each data point has a set of attributes
  - A distance/similarity measure between data points
    - E.g., Euclidean distance, cosine similarity, and edit distance
- **Task**
  - Partition the data points into separate groups (clusters)
- **Goal:**
  - Data points that belong to the same cluster are very similar to one another
  - Data points that belong to different clusters are less similar to one another
Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intra-cluster distances are minimized

Inter-cluster distances are maximized
Clustering: Applications

- Market Segmentation
  - Subdivide customers based on their geographical and lifestyle related information

- Document clustering
  - Find groups of documents that are similar to each other based on the important terms appearing in them

- Time series clustering
  - Find groups of similar time series (e.g., stock prices, ECG, seismic waves) based on their shapes

- Sequence clustering
  - Find groups of sequences (e.g., Web or protein sequences) with similar features
Association Rule Mining: Definition

- **Given:**
  - A collection of transactions
  - Each transaction contains a set of items

- **Task:**
  - Discover dependency rules that will predict the presence of an item in a record based on the presence of other items

- **Goal:**
  - Rules must have high support, i.e., applicable to sufficiently large number of records
  - Rules must have high confidence, i.e., make accurate prediction
## Illustrating Association Rule Mining

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>

Rules Discovered:

- \{Milk\} \rightarrow \{Coke\}
- \{Diaper, Milk\} \rightarrow \{Beer\}
Association Rule Mining: Applications

- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management

- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period

- World-Wide Web
  - Rules are used to develop Web caching and prefetching techniques
Association Rules in Election Survey Data

Data from 2000 American National Election Studies (NEC) conducted by Center of Political Studies at U of Michigan

<table>
<thead>
<tr>
<th>CONF</th>
<th>SUPPORT</th>
<th>LIFT</th>
<th>RHAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>40.30</td>
<td>10.42</td>
<td>1.67</td>
<td>WHITE &amp; USE SURPLUS FOR TAX CUTS &amp; FOR SCHOOL VOUCHERS</td>
</tr>
<tr>
<td>32.34</td>
<td>8.36</td>
<td>1.87</td>
<td>USE SURPLUS FOR TAX CUTS &amp; FOR SCHOOL VOUCHERS &amp; COLLEGE</td>
</tr>
<tr>
<td>22.14</td>
<td>5.72</td>
<td>1.83</td>
<td>USE SURPLUS FOR TAX CUTS &amp; MIDDLE CLASS &amp; FOR SCHOOL VOUCHERS</td>
</tr>
<tr>
<td>15.92</td>
<td>4.12</td>
<td>1.79</td>
<td>SOUTH &amp; FOR SCHOOL VOUCHERS &amp; COLLEGE</td>
</tr>
<tr>
<td>15.67</td>
<td>4.05</td>
<td>1.79</td>
<td>WHITE &amp; UPPER MID CLASS &amp; FOR DEATH PENALTY</td>
</tr>
<tr>
<td>16.92</td>
<td>4.37</td>
<td>1.78</td>
<td>WHITE &amp; RURAL &amp; FOR SCHOOL VOUCHERS</td>
</tr>
<tr>
<td>22.39</td>
<td>5.79</td>
<td>1.78</td>
<td>USE SURPLUS FOR TAX CUTS &amp; MALE &amp; FOR SCHOOL VOUCHERS</td>
</tr>
<tr>
<td>36.32</td>
<td>9.39</td>
<td>1.78</td>
<td>USE SURPLUS FOR TAX CUTS &amp; FOR SCHOOL VOUCHERS &amp; FOR DEATH PENALTY</td>
</tr>
<tr>
<td>16.42</td>
<td>4.24</td>
<td>1.77</td>
<td>USE SURPLUS FOR TAX CUTS &amp; FOR SCHOOL VOUCHERS &amp; AGE 45 &amp; 64</td>
</tr>
<tr>
<td>18.91</td>
<td>4.89</td>
<td>1.76</td>
<td>WHITE &amp; SOUTH &amp; FOR SCHOOL VOUCHERS</td>
</tr>
</tbody>
</table>


The highest lift rules for Republicans tended to repeat the same items: tax cuts, school vouchers, death penalty, college-educated, middle to upper-middle class, as shown in Figure 4.

Top 10 rules for democrats, by lift, involved black, age 45-64, and urban, and support of gun control, abortion rights, environmental protection and gays in the military, as listed in Figure 5.
Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events.

Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.
Sequential Pattern Discovery: Examples

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)
    (Rectifier_Alarm) --> (Fire_Alarm)

- In point-of-sale transaction sequences,
  - Computer Bookstore:
    (Intro_To_Visual_C) (C++_Primer) -->
     (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    (Shoes) (Racket, Racketball) --> (Sports_Jacket)
Deviation/Anomaly Detection

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

Typical network traffic at University level may reach over 100 million connections per day