Data Foundations

- Data Attributes and Features
- Data Pre-processing
- Data Storage
- Data Analysis

Data Attributes

- Describing data content and characteristics
- Representing data dimensions
- Set of all attributes: attribute vector or attribute array.
Attribute Types

**Nominal Data**

<table>
<thead>
<tr>
<th>Point</th>
<th>airport</th>
<th>town</th>
<th>mine</th>
<th>capital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line</td>
<td>river</td>
<td>road</td>
<td>boundary</td>
<td>pipeline</td>
</tr>
<tr>
<td>Area</td>
<td>orchard</td>
<td>desert</td>
<td>forest</td>
<td>water</td>
</tr>
</tbody>
</table>

**Sortable Data**

**Numerical Attributes**

Discrete vs. Continuous
Statistical Features of Data

Mean:
\[ \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}. \]

Median:
\[ Q_{\frac{1}{2}}(x) = \begin{cases} x_{\frac{n+1}{2}}, & \text{if } n \text{ is odd.} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & \text{if } n \text{ is even.} \end{cases} \]

Standard Deviation:
\[ \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 = \left( \frac{1}{N} \sum_{i=1}^{N} x_i^2 \right) - \bar{x}^2, \]

Similarity Matrix

Distance for Categorical Data:
- \( p \): total # of types;
- \( m \): # of same types

\[ d(i, j) = \frac{p - m}{p} \]
Distance Measures (Numerical)

- **Euclidean Distance:** $d_{EC} = \sqrt{\sum_{i=1}^{d} (P_i - Q_i)^2}$
- **Manhattan Distance:** $d_{CB} = \sum_{i=1}^{d} |P_i - Q_i|$  
- **Minkowski Distance:** $d_{Mt} = \left(\sum_{i=1}^{d} |P_i - Q_i|^p\right)^{\frac{1}{p}}$
- **Cosine Distance:** $s(X,Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|}$

Data Uncertainty

Source of uncertainty:
- Attribute error
- Missing attributes
- Data integration error
- Resolution conversion
- Application uncertainty
Data Preprocessing

1. Data Cleaning
2. Data Integration

✓ Data Quality
  – Accuracy
  – Completeness
  – Consistency
  – Timeliness
  – Believability
  – Interpretability
Data Visualization Quality

Data-Ink Ratio:

Data Error Types

✓ Missing data
  – Replace with constants
  – Replace with average value of attributes
  – Regression
  – Manual filling

✓ Noisy data
  – Regression
  – Outlier analysis
Visual Data Cleaning

✓ Using visualization techniques for data cleaning

Data Integration

✓ Combining data from multiple sources
  – Structural conflicts
  – Schema differences
  – Data Conflicts
  – Repeated data
✓ Providing a uniform visualization
Data Integration Example

Form 1:

<table>
<thead>
<tr>
<th>No</th>
<th>LName</th>
<th>FName</th>
<th>Gender</th>
<th>Address</th>
<th>Phone/Fax</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>Smith</td>
<td>Christoph</td>
<td>M</td>
<td>23 Harley St, Chicago, IL, 60633-2294</td>
<td>333-222-6542 / 333-222-6599</td>
</tr>
<tr>
<td>493</td>
<td>Smith</td>
<td>Kris L.</td>
<td>F</td>
<td>2 Harley Place, South Fork, MN, 48503, 5998</td>
<td>444-555-6666</td>
</tr>
</tbody>
</table>

Form 2:

<table>
<thead>
<tr>
<th>CID</th>
<th>Name</th>
<th>Street</th>
<th>City</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Kristen Smith</td>
<td>2 Harley Pl</td>
<td>South Fork, MN</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>Christian Smith</td>
<td>Hurley St 2</td>
<td>S Fork MN</td>
<td>1</td>
</tr>
</tbody>
</table>

Integrated Form:

<table>
<thead>
<tr>
<th>No</th>
<th>LName</th>
<th>FName</th>
<th>Gender</th>
<th>Street</th>
<th>City</th>
<th>ZIP</th>
<th>Phone</th>
<th>Fax</th>
<th>CID</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Smith</td>
<td>Kristen L.</td>
<td>F</td>
<td>2 Harley Place</td>
<td>South Fork, MN</td>
<td>48503-5998</td>
<td>444-555-6666</td>
<td>11</td>
<td>493</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Smith</td>
<td>Christian</td>
<td>M</td>
<td>2 Harley Place</td>
<td>South Fork, MN</td>
<td>48503-5998</td>
<td>444-555-6666</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Smith</td>
<td>Christoph</td>
<td>M</td>
<td>23 Harley Street</td>
<td>Chicago, IL</td>
<td>60633-2294</td>
<td>333-222-6542</td>
<td>333-222-6599</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>

Data Storage

- File Systems
- Databases and DBMS
- Data Warehouse
Structured Files

- **XML files**: eXtensible Markup Language

  ```xml
  <note>
    <to>Tove</to>
    <from>Jani</from>
    <heading>Reminder</heading>
    <body>Don't forget me this weekend!</body>
  </note>
  ``

- **XML Extensions**: IVOA VOtable (Space programs), KML (Web maps), etc.

- **Special formats**: e.g. HDF (Hierarchical Data Format) for scientific data
Databases

✓ A database is an organized collection of data. The data is typically organized to model aspects of reality in a way that supports processes requiring information. For example, modelling the availability of rooms in hotels in a way that supports finding a hotel with vacancies.

Relational Databases

✓ DBMS: Database Management System
✓ Data definition and query languages (SQL)
Visualization of a database

**Challenge:** Interactivity!

**Example:** National Science Foundation Database:

Data Warehouse

✓ A data warehouse is a system used for reporting and data analysis. It is a central repository of integrated data from one or more disparate sources.
Database vs Data Warehouse

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Database</th>
<th>Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>Data operation</td>
<td>Information in the data</td>
</tr>
<tr>
<td>Users</td>
<td>Employee, DB Administrator</td>
<td>Analyst, manager, executive</td>
</tr>
<tr>
<td>Functionalities</td>
<td>Daily operations</td>
<td>decision making support</td>
</tr>
<tr>
<td>Data</td>
<td>Current</td>
<td>Historical, time-variant</td>
</tr>
<tr>
<td>Access</td>
<td>Read, write, mean, etc</td>
<td>Read</td>
</tr>
<tr>
<td>Focus</td>
<td>Input, query</td>
<td>information output</td>
</tr>
<tr>
<td>Size</td>
<td>1 GB ~ &lt; 1 TB</td>
<td>TBs</td>
</tr>
</tbody>
</table>

Data Analysis

- Statistical Analysis
- Exploratory Analysis
- Data Mining
Statistical Analysis

- Statistical description: properties, parameters, distribution, correlations.
- Statistical prediction: using probability methods and sampling theory to predict statistical properties (distribution, correlation, parameters, forecasting, etc.)

Data Exploration Using Visualization

- Raw data drawing
- Statistical drawing
- Multi-view
**Data Trajectory**

**MOBILE RENTAL SCORES**  
*JULY 13 – JULY 17, 2011*

<table>
<thead>
<tr>
<th></th>
<th>Netflix</th>
<th>Hulu</th>
<th>Redbox</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W</td>
<td>Th</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>Th</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>Su</td>
<td>Su</td>
<td>Su</td>
</tr>
</tbody>
</table>

**Data Comparison**

**Questionnaire: 2012**  
*Survey*  
*Sample Size:*  
*Purpose:*  
*Date:*  
*Venue:*  
*Methodology:*  
*Participants:*  
*Analysis:*  
*Conclusion:*
Trend and Patterns

Relations
Line Chart

Peak Break-Up Times
According to Facebook status updates

Spring Break
“spring clean”
Valentine’s Day
April Fool’s Day
Mondays
Summer holiday

2 weeks before
winter holidays

Christmas
“too cruel”

Line Chart: Sunspots

α = 1.0

Sunspot Number

0 50 100 150

1750 1800 1850 1900

Year
Bar Chart Labeling

Bar Chart: Scale
Bar Chart: Variant

Stacked Bar Chart
Stacked Bar Chart

Stacked Chart
Pie Chart

http://www.youtube.com/watch?feature=player_detailpage&v=DjKtAYUlfms#t=438s
Pie Charts of Van Gogh Paintings

Histogram
Contour Map

Scatter Plot Matrix
Reference Lines in Scatter Plot

Heatmap
Box Plot

- Maximum
- Upper Quartile
- Median
- Lower Quartile
- Minimum

Box Plot Variations

- Different box plot variations with various notations and outlier representations.
Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive repositories or data streams” – J. Han and M Kamber, “Data Mining: Concepts and Techniques”
Data Mining Tasks

**Description:**
- Data
- Algorithm
- Features

**Prediction**
1. Model
2. Trained Model

**Descriptive Tasks:**
- Concept Description
- Association Mining
- Clustering
- Outlier Analysis

**Predictive Tasks:**
- Classification
- Evolution Analysis
Data Mining Methods

- **Statistical Method:** Regression, Parameter Estimation
- **Machine Learning:** Decision tree, Neural Networks
- **Algorithmic Method:** K-mean, Graph operations
- **Statistical Learning:** Probability model, Bayesian networks

Data Mining New Applications

- **Text Mining** – summarizes, navigates, and clusters documents contained in a database.
- **Web Mining** – integrates data and text mining within a Web site; enhances the Web site with intelligent behavior, such as suggesting related links or recommending new products to the consumer
Data Visualization Pipeline

Looping Model
Visual Data Mining and Visual Analytics

✓ Users are involved in the data mining process through visualization and user interactions.
✓ Certain tasks are difficult to be automated
  – Validation of clustering results
  – Checking data focal points and noise
  – Expert knowledge input, etc.

Visual Analytics Paradigm
Visual Analytics

Visual Data Mining Examples: Visualizing data correlations
Visual Data Mining Example:
Biomarker detection

Visual Data Mining Example:
Facial feature detection for medical diagnosis
Visual Data Mining Example:
Concept detection in text data

Visual Data Mining Example:
Visualizing decision trees
IEEE Conference on Visual Analytics Science and Technology

BaobabView: Interactive Construction and Analysis of Decision Trees