Introduction to Data Mining
Methods and Tools

by Michael Hahsler
Agenda

- What is Data Mining?
- Data Mining Tasks
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues
Agenda

- What is Data Mining?
- Data Mining Tasks
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues
What is Data Mining?

One of many definitions:

"Data mining is the science of extracting useful knowledge from huge data repositories"

ACM SIGKDD, Data Mining Curriculum: A Proposal

http://www.kdd.org/curriculum
Why Data Mining? 
Commercial Viewpoint

• Businesses collect and warehouse lots of **data**.
  – Purchases at department/grocery stores
  – Bank/credit card transactions
  – Web and social media data
  – Mobile and IOT

• **Computers** are cheaper and more powerful.

• **Competition** to provide better services.
  – Mass customization and recommendation systems
  – Targeted advertising
  – Improved logistics
Why Mine Data?
Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data

- Data mining may help scientists
  - identify **patterns and relationships**
  - to **classify and segment** data
  - formulate hypotheses
Knowledge Discovery in Databases (KDD) Process

Data normalization
- Noise/outliers
- Missing data

Data/dim. reduction
- Features engineering
- Feature selection

Understand domain

Decide on task & algorithm Performance?
CRISP-DM Reference Model

- **Cross Industry Standard Process for Data Mining**

- De facto standard for conducting data mining and knowledge discovery projects.

- Defines tasks and outputs.

- Now developed by IBM as the Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM).

- SAS has SEMMA and most consulting companies use their own process.

Tasks in the CRISP-DM Model

**Business Understanding**
- Determine Business Objectives
  - Background
  - Business Objectives
  - Business Success Criteria
- Assess Situation
  - Inventory of Resources
  - Requirements, Assumptions, and Constraints
  - Risks and Contingencies
  - Terminology
  - Costs and Benefits
- Determine Data Mining Goals
  - Data Mining Goals
  - Data Mining Success Criteria
- Produce Project Plan
  - Project Plan
  - Initial Assessment of Tools and Techniques

**Data Understanding**
- Collect Initial Data
  - Initial Data Collection Report
- Describe Data
  - Data Description Report
- Explore Data
  - Data Exploration Report
- Verify Data Quality
  - Data Quality Report

**Data Preparation**
- Select Data
  - Rationale for Inclusion/Exclusion
- Clean Data
  - Data Cleaning Report
- Construct Data
  - Derived Attributes
  - Generated Records
- Integrate Data
  - Merged Data
- Format Data
  - Reformatted Data
  - Dataset
  - Dataset Description

**Modeling**
- Select Modeling Techniques
  - Modeling Technique
  - Modeling Assumptions
- Generate Test Design
  - Test Design
- Build Model
  - Parameter Settings
  - Models
  - Model Descriptions
- Assess Model
  - Model Assessment
  - Revised Parameter Settings

**Evaluation**
- Evaluate Results
  - Assessment of Data Mining Results w.r.t. Business Success Criteria
  - Approved Models
- Review Process
  - Review of Process
- Determine Next Steps
  - List of Possible Actions
  - Decision

**Deployment**
- Plan Deployment
  - Deployment Plan
- Plan Monitoring and Maintenance
  - Monitoring and Maintenance Plan
- Produce Final Report
  - Final Report
  - Final Presentation
- Review Project
  - Experience
  - Documentation

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model
Agenda

- What is Data Mining?
- **Data Mining Tasks**
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues
Data Mining Tasks

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

- **Predictive Methods**
  - Use some features (variables) to predict and unknown or future value of other variable.
Data Mining Tasks

Classification

Regression

Cluster Analysis

Predictive Modeling

Association Analysis

Anomaly Detection

Data

<table>
<thead>
<tr>
<th>ID</th>
<th>Home Owner</th>
<th>Marital Status</th>
<th>Annual Income</th>
<th>Defaulted Borrower</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>80K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison Wesley, 2006
Data Mining Tasks

- Classification
- Regression
- Cluster Analysis
- Predictive Modeling
- Association Analysis
- Anomaly Detection
Clustering

Group points such that
- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

Ideal grouping is not known → **Unsupervised Learning**

Euclidean distance based clustering in 3-D space.
Clustering Market Segmentation

- **Goal**: subdivide a market into distinct subsets of customers. Use a different marketing mix for each segment.

- **Approach**:
  - Collect different attributes of customers based on their geographical and lifestyle related information and observed buying patterns.
  - Find clusters of similar customers.
Clustering Documents

- **Goal**: Find groups of documents that are similar to each.

- **Approach**: Identify frequently occurring terms in each document. Define a similarity measure based on term co-occurrences. Use it to cluster.

- **Gain**: Can be used to organize documents or to create recommendations.
Clustering Data Reduction

- **Goal**: Reduce the data size for predictive models.

- **Approach**: Group data given a subset of the available information and then use the group label instead of the original data as input for predictive models.
Data Mining Tasks

- Classification
- Regression
- Cluster Analysis
- Predictive Modeling
- Association Analysis
- Anomaly Detection
Association Rule Discovery

- Given is a set of transactions. Each contains a number of items.
- Produce dependency rules of the form \( \text{LHS} \rightarrow \text{RHS} \)
  which indicate that if the set of items in the LHS are in a transaction, then the transaction likely will also contain the RHS item.

<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>

Transaction data

<table>
<thead>
<tr>
<th></th>
<th>Discovered Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Milk} → {Coke}</td>
<td></td>
</tr>
<tr>
<td>{Diaper, Milk} → {Beer}</td>
<td></td>
</tr>
</tbody>
</table>
Let the rule discovered be

\{\text{Potato Chips, ...} \} \rightarrow \{\text{Soft drink}\}

- **Soft drink as RHS:** What should be done to boost sales? Discount Potato Chips?
- **Potato Chips in LHS:** Shows which products would be affected if the store discontinues selling Potato Chips.
- **Potato Chips in LHS and Soft drink in RHS:** What products should be sold with Potato Chips to promote sales of Soft drinks!
Association Rule Discovery
Supermarket shelf management

- **Goal**: To identify items that are bought together by sufficiently many customers.

- **Approach**:
  - Process the point-of-sale data to find dependencies among items.
  - Place dependent items
    - close to each other (convenience).
    - far from each other to expose the customer to the maximum number of products in the store.
Goal: Anticipate the nature of repairs to keep the service vehicles equipped with right parts to speed up repair time.

Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover co-occurrence patterns.
Data Mining Tasks

Classification

Regression

Cluster Analysis

Predictive Modeling

Association Analysis

Anomaly Detection
Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Studied in statistics and econometrics.

Applications:
- Predicting sales amounts of new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of stock market indices (autoregressive models).
Data Mining Tasks

- Classification
- Regression

Cluster Analysis

Association Analysis

Anomaly Detection

Predictive Modeling

Data

<table>
<thead>
<tr>
<th>ID</th>
<th>Home Owner</th>
<th>Marital Status</th>
<th>Annual Income</th>
<th>Defaulted Borrower</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>80K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Classification

Find a **model** for the class attribute as a function of the values of other attributes/features.

Class information is available → **Supervised Learning**
Classification

Find a **model** for the class attribute as a function of the values of other attributes/features.

**Goal:** assign new records to a class as accurately as possible.

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Single</td>
<td>75K</td>
<td>?</td>
</tr>
<tr>
<td>Yes</td>
<td>Married</td>
<td>50K</td>
<td>?</td>
</tr>
<tr>
<td>No</td>
<td>Married</td>
<td>150K</td>
<td>?</td>
</tr>
<tr>
<td>Yes</td>
<td>Divorced</td>
<td>90K</td>
<td>?</td>
</tr>
<tr>
<td>No</td>
<td>Single</td>
<td>40K</td>
<td>?</td>
</tr>
<tr>
<td>No</td>
<td>Married</td>
<td>80K</td>
<td>?</td>
</tr>
</tbody>
</table>
Classification
Direct Marketing

**Goal:** Reduce cost of mailing by targeting a set of consumers likely to buy a **new** product.

**Approach:**

- Use the **data** for a similar product introduced before or from a focus group. We have customer information (e.g., demographics, lifestyle, previous purchases) and know which customers decided to buy and which decided otherwise. This **buy/don’t buy** decision forms the **class attribute**.
- Use this information as input attributes to **learn a classifier model**.
- Apply the model to new customers to **predict** if they will buy the product.
Goal: To predict whether a customer is likely to be lost to a competitor.

Approach:
- Use detailed record of transactions with each of the past and present customers, to find attributes (frequency, recency, complaints, demographics, etc.).
- Label the customers as loyal or disloyal.
- Find a model for disloyalty.
- Rank each customer on a loyal/disloyal scale (e.g., churn probability).
Classification
Sky Survey Cataloging

- **Goal:** To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).

- **Approach:**
  - Segment the image to **identify objects**.
  - **Derive features** per object (40).
  - Use known objects to **model the class** based on these features.

- **Result:** Found 16 new high red-shift quasars.

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

- Classification
- Regression
- Cluster Analysis
- Predictive Modeling
- Association Analysis
- Anomaly Detection
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.
Other Data Mining Tasks

- **Text mining** – document clustering, topic models
- **Graph mining** – social networks
- **Data stream mining**/real time data mining
- **Mining spatiotemporal data** (e.g., moving objects)
- **Visual data mining**
- **Distributed data mining**
Challenges of Data Mining

- Scalability
- Dimensionality
- Complexity and heterogeneous data
- Data quality
- Data ownership and privacy
Agenda

- What is Data Mining?
- Data Mining Tasks
- **Relationship to Statistics, Optimization, Machine Learning and AI**
- Tools
- Data
- Legal, Privacy and Security Issues
Origins of Data Mining

Draws ideas from AI, machine learning, pattern recognition, statistics, and database systems.

There are differences in terms of
- used data and
- the goals.

https://rayli.net/blog/data/history-of-data-mining/
Relationship to other Fields

Method
- Artificial Intelligence
- Optimization
- Statistics

+ Application Areas

Learning Strategy
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Online Learning

Analytics? Data Science? Big Data?

Machine Learning

Statistical Learning

Data Mining

Math
Artificial Intelligence: Create an autonomous agent that perceives its environment and takes actions that maximize its chance of reaching some goal.

Areas: reasoning, knowledge representation, planning, learning, natural language processing, and vision.
**Relationship to other Fields**

**Optimization**: Selection of a best alternative from some set of available alternatives with regard to some criterion.  
**Techniques**: Linear programming, integer programming, nonlinear programming, stochastic and robust optimization, heuristics, etc.
**Statistics:** Study of the collection, analysis, interpretation, presentation, and organization of data.

**Techniques:** Descriptive statistics, statistical inference (estimation, testing), design of experiments.
Learning Strategy: From what data do we learn?

- Is a training set with correct answers available? → Supervised learning
- Long-term structure of rewards? → Reinforcement learning
- No answer and no reward structure? → Unsupervised learning
- Do we have to update the model regularly? → Online learning
**Statistical learning:** deals with the problem of finding a **predictive function** based on data.

**Tools:** (Linear) classifiers, regression and regularization.
Machine Learning involves the study of algorithms that can extract information automatically, i.e., without on-line human guidance.

Techniques: Focus on supervised learning.
Data Mining: Manually analyze a given dataset to gain insights and predict potential outcomes.

Techniques: Any applicable technique from databases, statistics, machine/statistical learning. New methods were developed by the Data Mining community.
Data Mining & Analytics

<table>
<thead>
<tr>
<th>Competitive Advantage</th>
<th>Degree of Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stochastic Optimization</td>
<td>How can we achieve the best outcome including the effects of variability?</td>
</tr>
<tr>
<td>Optimization</td>
<td>How can we achieve the best outcome?</td>
</tr>
<tr>
<td>Predictive modeling</td>
<td>What will happen next if?</td>
</tr>
<tr>
<td>Forecasting</td>
<td>What if these trends continue?</td>
</tr>
<tr>
<td>Simulation</td>
<td>What could happen….?</td>
</tr>
<tr>
<td>Alerts</td>
<td>What actions are needed?</td>
</tr>
<tr>
<td>Query/drill down</td>
<td>What exactly is the problem?</td>
</tr>
<tr>
<td>Ad hoc reporting</td>
<td>How many, how often, where?</td>
</tr>
<tr>
<td>Standard Reporting</td>
<td>What happened?</td>
</tr>
</tbody>
</table>

Based on: Competing on Analytics, Davenport and Harris, 2007

OR

Data Mining / Stats

Statistics

OR

Machine Learning

DB / CS
Prescriptive Analytics

What decisions should we make now to achieve the best future outcome?

Issues:
- What are the decision variables? Causality?
- Relationship can be non-linear. Convex?
- Uncertainty about quality and reliability of the predictive model.
Good luck finding this person! Probably a team effort!

Source: T. Stadelmann, et al., Applied Data Science in Europe
Agenda

- What is Data Mining?
- Data mining techniques
- Relationship to Statistics, Optimization, Machine Learning and AI

Tools

- Data
- Legal, Privacy and Security Issues
Gartner 2016 Magic Quadrant for Advanced Analytics Platforms (changes from 2015)
Tools
Popularity

Rexer Analytics

2015

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2895 voters]

Legend: red: Free/Open Source tools
green: Commercial tools
Fuchsia: Hadoop/Big Data tools

<table>
<thead>
<tr>
<th>Software</th>
<th>% users in 2016</th>
<th>% users in 2015</th>
<th>% users in 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>R (1419)</td>
<td>49.0%</td>
<td>46.9%</td>
<td></td>
</tr>
<tr>
<td>Python (1325)</td>
<td>45.8%</td>
<td>30.3%</td>
<td></td>
</tr>
<tr>
<td>SQL (1029)</td>
<td>35.5%</td>
<td>30.9%</td>
<td></td>
</tr>
<tr>
<td>Excel (972)</td>
<td>22.9%</td>
<td>25.8%</td>
<td></td>
</tr>
<tr>
<td>RapidMiner (944), 11.7 % alone</td>
<td>32.6%</td>
<td>31.5%</td>
<td>44.2%</td>
</tr>
<tr>
<td>Hadoop (641)</td>
<td>22.1%</td>
<td>18.4%</td>
<td>12.7%</td>
</tr>
<tr>
<td>Spark (624)</td>
<td>21.6%</td>
<td></td>
<td>11.3%</td>
</tr>
</tbody>
</table>

n = 1,220 analytic professionals


Tools
Types

- Simple graphical user interface
- Process oriented
- Programming oriented
Tools

Simple GUI

- **Weka**: Waikato Environment for Knowledge Analysis (Java API)
- **Rattle**: GUI for Data Mining using R
Tools
Process oriented

- SAS Enterprise Miner
- IBM SPSS Modeler
- RapidMiner
- Knime
- Orange
Tools
Programming oriented

- **R**
  - Rattle for beginners
  - RStudio IDE, markdown, shiny
  - Microsoft Open R

- **Python**
  - Scikit-learn, pandas
  - IPython, notebooks

→ Both have similar capabilities. Slightly different focus:
  - R: statistical computing and visualization
  - Python: Machine learning and big data
R

```r
library(GGally)
ggpairs(nba[, c("ast", "fg", "trb")])
```

Python

```python
import seaborn as sns
import matplotlib.pyplot as plt
sns.pairplot(nba[['ast', 'fg', 'trb']])
plt.show()
```
Agenda

- What is Data Mining?
- Data mining techniques
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues
Data

Web     Customer ID,Email,Country,Telephone,First name,Surname,Company,..
456,AQA.com,United Kingdom,123,A,A,A,..
457,BQB.com,United States,124,B,B,B,..
458,CQC.com,Aran Emerates,125,C,C,C,..
459,DQD.com,New Zealand,126,D,D,D,..
460,EQE.com,United Kindom,127,E,E,E,..
461,EQP.com,United States,128,F,F,F,..
462,GAG.com,Aran Emerates,129,G,G,G,..
463,HGH.com,New Zealand,130,H,H,H,..
464,IHI.com,United Kindom,131,I,I,I,..
465,JQQ.com,United States,132,J,J,J,..
466,KQQ.com,Aran Emerates,133,K,K,K,..
467,LLL.com,New Zealand,134,L,L,L,..
468,MMI.com,United Kindom,135,M,M,M,..
469,United States,136,N,N,N,..
470,QQS.com,Aran Emerates,137,O,O,O,..
471,POP.com,New Zealand,138,P,P,P,..
472,United Kindom,139,O,O,O,..
473,RRS.com,United States,140,R,R,R,..
474,SOS.com,Aran Emerates,141,S,S,S,..
475,TQQ.com,New Zealand,142,T,T,T,..
476,XXX.com,United Kindom,143,U,U,U,..
477,VVV.com,United States,144,V,V,V,..
478,WWW.com,Aran Emerates,145,W,W,W,..
479,XXX.com,New Zealand,146,X,X,X,..
480,YYY.com,United Kindom,147,Y,Y,Y,..
481,ZQQ.com,United States,148,Z,Z,Z,..

<table>
<thead>
<tr>
<th>Date</th>
<th>Amount</th>
<th>Budgeted</th>
<th>Difference</th>
<th>Department</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005-9-1</td>
<td>$3,498.56</td>
<td>$3,200.00</td>
<td>$298.56</td>
<td>Grounds</td>
<td>Equipment</td>
</tr>
<tr>
<td>2005-9-1</td>
<td>$1,912.11</td>
<td>$2,000.00</td>
<td>$(87.89)</td>
<td>IT</td>
<td>Software</td>
</tr>
<tr>
<td>2005-9-3</td>
<td>$2,121.21</td>
<td>$2,100.00</td>
<td>$21.21</td>
<td>Telephones</td>
<td>Services</td>
</tr>
<tr>
<td>2005-9-8</td>
<td>$1,837.27</td>
<td>$2,000.00</td>
<td>$(162.73)</td>
<td>IT</td>
<td>Consulting</td>
</tr>
<tr>
<td>2005-9-10</td>
<td>$323.99</td>
<td>$150.00</td>
<td>$173.99</td>
<td>Grounds</td>
<td>Supplies</td>
</tr>
<tr>
<td>2005-9-12</td>
<td>$81.61</td>
<td>$100.00</td>
<td>$(18.39)</td>
<td>Telephones</td>
<td>Services</td>
</tr>
<tr>
<td>2005-9-14</td>
<td>$2,500.00</td>
<td>$4,000.00</td>
<td>$(1,500.00)</td>
<td>Administration</td>
<td>Supplies</td>
</tr>
<tr>
<td>2005-9-15</td>
<td>$31,872.22</td>
<td>$32,000.00</td>
<td>$(127.78)</td>
<td>Administration</td>
<td>Consulting</td>
</tr>
<tr>
<td>2005-9-15</td>
<td>$10,330.31</td>
<td>$10,000.00</td>
<td>$330.31</td>
<td>Grounds</td>
<td>Payroll</td>
</tr>
<tr>
<td>2005-9-15</td>
<td>$12,897.69</td>
<td>$12,500.00</td>
<td>$397.69</td>
<td>Grounds</td>
<td>Payroll</td>
</tr>
</tbody>
</table>
Subject Oriented: Data warehouses are designed to help you analyze data (e.g., sales data is organized by product and customer).

Integrated: Integrates data from disparate sources into a consistent format.

Nonvolatile: Data in the data warehouse are never overwritten or deleted.

Time Variant: maintains both historical and (nearly) current data.
ETL
Extract, Transform and Load

- **Extracting** data from outside sources
- **Transforming** data to fit analytical needs. E.g.,
  - Clean missing data, wrong data, etc.
  - Normalize and translate (e.g., 1 → "female")
  - Join from several sources
  - Calculate and aggregate data
- **Loading** data into the data warehouse

Source: SAS, ETL: What it is and why it matters
For fast operation OLAP requires a special database structure (Snow-flake scheme)
Big Data

"Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them." — Wikipedia

3 V's: Volume, velocity, variety, (veracity) — Gartner

---

![Diagram showing various components of big data processing](image-url)
Agenda

- What is Data Mining?
- Data mining techniques
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues
Legal, Privacy and Security Issues

DID YOU EVER THINK ABOUT SELLING OUR CONFIDENTIAL DATABASE OF CUSTOMER INFORMATION?

IT WOULD BE MASSIVELY PROFITABLE WHILE VIRTUALLY UNDETECTABLE.

BUT HIGHLY UNETHICAL.

I DON'T KNOW YOU ANYMORE.

I'M YANKING YOUR CHAIN. WHEN DO WE START?

?
Legal, Privacy and Security Issues

- Are we allowed to collect the data?
- Are we allowed to use the data?
- Is privacy preserved in the process?
- Is it ethical to use and act on the data?

Problem: Internet is global but legislation is local!
BERLIN — Angry Birds, the top-selling paid mobile app for the iPhone in the United States and Europe, has been downloaded more than a billion times by devoted game players around the world, who often spend hours slinging squawking fowl at groups of egg-stealing pigs.

When Jason Hong, an associate professor at the Human-Computer Interaction Institute at Carnegie Mellon University, surveyed 40 users, all but two were unaware that the game was storing their locations so that they could later be the targets of ads....
Pokémon Go’s constant location tracking and camera access required for gameplay, paired with its skyrocketing popularity, could provide data like no app before it.

“**Their privacy policy is vague,**” Hong said. “I’d say deliberately vague, because of the lack of clarity on the business model.”

...

The agreement says Pokémon Go collects data about its users as a “**business asset.**” This includes data used to personally identify players such as email addresses and other information pulled from Google and Facebook accounts players use to sign up for the game.

*If Niantic is ever sold, the agreement states, all that data can go to another company.*
Conclusion

Data Mining is **interdisciplinary** and overlaps significantly with many fields including

- statistics,
- CS (machine learning, AI, data bases)
- optimization.

Data Mining requires a **team effort** with members who have expertise in

- data management,
- statistics,
- programming,
- communication, and
- the application domain.