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
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
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

Decide on task & algorithm
- Performance?

Understand domain

CRISP-DM Reference Model

- Cross Industry Standard Process for Data Mining
- Open standard process model
- Industry, tool and application neutral
- 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 similar process.

## Tasks in the CRISP-DM Model

<table>
<thead>
<tr>
<th>Business Understanding</th>
<th>Data Preparation</th>
<th>Modeling</th>
<th>Evaluation</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Determine Business Objectives</strong></td>
<td><strong>Select Data Rationale for Inclusion/Exclusion</strong></td>
<td><strong>Select Modeling Techniques</strong></td>
<td><strong>Evaluate Results</strong></td>
<td><strong>Plan Deployment</strong></td>
</tr>
<tr>
<td>Background</td>
<td>Clean Data</td>
<td>Modeling Technique</td>
<td>Assessment of Data Mining Results w.r.t. Business Success Criteria</td>
<td>Deployment Plan</td>
</tr>
<tr>
<td>Business Objectives</td>
<td>Data Cleaning Report</td>
<td>Modeling Assumptions</td>
<td>Approved Models</td>
<td>Plan Monitoring and Maintenance</td>
</tr>
<tr>
<td>Business Success Criteria</td>
<td>Explore Data</td>
<td>Generate Test Design</td>
<td>Review Process</td>
<td>Monitoring and Maintenance Plan</td>
</tr>
<tr>
<td><strong>Assess Situation</strong></td>
<td>Construct Data</td>
<td>Test Design</td>
<td>Review of Process</td>
<td>Produce Final Report</td>
</tr>
<tr>
<td>Inventory of Resources Requirements, Assumptions, and Constraints</td>
<td>Derived Attributes</td>
<td>Build Model</td>
<td>Final Report</td>
<td>Final Presentation</td>
</tr>
<tr>
<td>Risks and Contingencies</td>
<td>Generated Records</td>
<td>Parameter Settings Models</td>
<td>Review Project</td>
<td>Experience Documentation</td>
</tr>
<tr>
<td>Terminology</td>
<td>Integrate Data</td>
<td>Model Descriptions</td>
<td>List of Possible Actions Decision</td>
<td></td>
</tr>
<tr>
<td>Costs and Benefits</td>
<td>Merged Data</td>
<td>Assess Model Model Assessment Revised Parameter Settings</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data Quality Report</td>
<td>Data Quality Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data Description Report</td>
<td>Data Collection Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Initial Data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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 unknown or future value of other variable.
Data Mining Tasks

- Classification
- Regression

### Data

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

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>
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:**
1. Collect different attributes of customers based on their geographical and lifestyle related information and observed buying patterns.
2. 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 $$LHS \rightarrow RHS$$
- 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>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
<td></td>
</tr>
</tbody>
</table>

Discovered Rules

- $$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$$
- $$\{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}$$
Association Rule Discovery
Marketing and Sales Promotion

- 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.
Association
Rule
Discovery
Inventory
Management

- **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
- Predictive Modeling
- Cluster Analysis
- 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

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**

<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>

**Training Set** → **Learn Classifier** → **Model**
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>

**Refund Marital Status Taxable Income Cheat**

<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
- Data quality
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

Data Science?
Analytics?
Big Data?
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.
**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

Based on: Competing on Analytics, Davenport and Harris, 2007
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.
Data Science

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

Good luck finding this person!
Probably a team effort!
Agenda

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

Tools: Popularity

Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll

- Python
- RapidMiner
- R Language
- Excel
- Anaconda
- SQL Language
- Tensorflow
- Keras
- scikit-learn
- Tableau
- Apache Spark

https://www.kdnuggets.com/polls/
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**
  - Numpy, scikit-learn, pandas
  - Jupyter notebook

→ Both have similar capabilities. Slightly different focus:
  - R: statistical computing and visualization
  - Python: Scripting, big data
  - Interoperability via rpy2 and rediculate
R

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

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 Tasks
- Relationship to Statistics, Optimization, Machine Learning and AI
- Tools
- Data
- Legal, Privacy and Security Issues
### Data

#### CSV File: WebSiteUsers.csv

<table>
<thead>
<tr>
<th>Web</th>
<th>Customer ID</th>
<th>Email</th>
<th>Country</th>
<th>Telephone</th>
<th>First name</th>
<th>Surname</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>456</td>
<td>AAA.com</td>
<td>United Kingdom</td>
<td>123 A A A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>457</td>
<td><a href="mailto:B@B.com">B@B.com</a></td>
<td>United States</td>
<td>124 B B B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>458</td>
<td><a href="mailto:C@C.com">C@C.com</a></td>
<td>Aran Emerates</td>
<td>125 C C C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>459</td>
<td><a href="mailto:D@D.com">D@D.com</a></td>
<td>New Zealand</td>
<td>126 D D D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>460</td>
<td><a href="mailto:E@E.com">E@E.com</a></td>
<td>United Kingdom</td>
<td>127 E E E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>461</td>
<td><a href="mailto:F@F.com">F@F.com</a></td>
<td>United States</td>
<td>128 F F F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>462</td>
<td><a href="mailto:G@G.com">G@G.com</a></td>
<td>Aran Emerates</td>
<td>129 G G G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>463</td>
<td><a href="mailto:H@H.com">H@H.com</a></td>
<td>New Zealand</td>
<td>130 H H H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>464</td>
<td><a href="mailto:I@I.com">I@I.com</a></td>
<td>United Kingdom</td>
<td>131 I I I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>465</td>
<td><a href="mailto:J@J.com">J@J.com</a></td>
<td>United States</td>
<td>132 J J J</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>466</td>
<td><a href="mailto:K@K.com">K@K.com</a></td>
<td>Aran Emerates</td>
<td>133 K K K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>467</td>
<td><a href="mailto:L@L.com">L@L.com</a></td>
<td>New Zealand</td>
<td>134 L L L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>468</td>
<td><a href="mailto:M@M.com">M@M.com</a></td>
<td>United Kingdom</td>
<td>135 M M M</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>469</td>
<td><a href="mailto:N@N.com">N@N.com</a></td>
<td>United States</td>
<td>136 N N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>470</td>
<td><a href="mailto:O@O.com">O@O.com</a></td>
<td>Aran Emerates</td>
<td>137 O O O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>471</td>
<td><a href="mailto:P@P.com">P@P.com</a></td>
<td>New Zealand</td>
<td>138 P P P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>472</td>
<td><a href="mailto:R@R.com">R@R.com</a></td>
<td>United Kingdom</td>
<td>139 R R R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>473</td>
<td><a href="mailto:S@S.com">S@S.com</a></td>
<td>Aran Emerates</td>
<td>140 S S S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>474</td>
<td><a href="mailto:T@T.com">T@T.com</a></td>
<td>New Zealand</td>
<td>141 T T T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>475</td>
<td><a href="mailto:U@U.com">U@U.com</a></td>
<td>United Kingdom</td>
<td>143 U U U</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>476</td>
<td><a href="mailto:V@V.com">V@V.com</a></td>
<td>United States</td>
<td>144 V V V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>477</td>
<td><a href="mailto:W@W.com">W@W.com</a></td>
<td>Aran Emerates</td>
<td>145 W W W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>479</td>
<td><a href="mailto:X@X.com">X@X.com</a></td>
<td>New Zealand</td>
<td>146 X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>480</td>
<td><a href="mailto:Y@Y.com">Y@Y.com</a></td>
<td>United Kingdom</td>
<td>147 Y Y Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>481</td>
<td><a href="mailto:Z@Z.com">Z@Z.com</a></td>
<td>United States</td>
<td>148 Z Z Z</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Excel Spreadsheet: Excel training sample 3.xls

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>9/1/2005</td>
<td>$3,498.56</td>
<td>$3,200.00</td>
<td>$298.56</td>
<td>Grounds</td>
</tr>
<tr>
<td>3</td>
<td>9/1/2005</td>
<td>$1,912.11</td>
<td>$2,000.00</td>
<td>(87.89)</td>
<td>IT</td>
</tr>
<tr>
<td>4</td>
<td>9/3/2005</td>
<td>$2,121.21</td>
<td>$2,100.00</td>
<td>21.21</td>
<td>Telephones</td>
</tr>
<tr>
<td>5</td>
<td>9/8/2005</td>
<td>$1,837.27</td>
<td>$2,000.00</td>
<td>162.73</td>
<td>IT</td>
</tr>
<tr>
<td>6</td>
<td>9/10/2005</td>
<td>$323.99</td>
<td>$150.00</td>
<td>173.99</td>
<td>Grounds</td>
</tr>
<tr>
<td>7</td>
<td>9/12/2005</td>
<td>$81.61</td>
<td>$100.00</td>
<td>(18.39)</td>
<td>IT</td>
</tr>
<tr>
<td>8</td>
<td>9/14/2005</td>
<td>$2,500.00</td>
<td>$4,000.00</td>
<td>1500.00</td>
<td>Administration</td>
</tr>
<tr>
<td>9</td>
<td>9/15/2005</td>
<td>$1,000.00</td>
<td>$500.00</td>
<td>500.00</td>
<td>IT</td>
</tr>
<tr>
<td>10</td>
<td>9/15/2005</td>
<td>$31,872.22</td>
<td>$32,000.00</td>
<td>127.78</td>
<td>Administration</td>
</tr>
<tr>
<td>11</td>
<td>9/15/2005</td>
<td>$10,330.31</td>
<td>$10,000.00</td>
<td>330.31</td>
<td>Grounds</td>
</tr>
<tr>
<td>12</td>
<td>9/15/2005</td>
<td>$12,897.69</td>
<td>$12,500.00</td>
<td>397.69</td>
<td>IT</td>
</tr>
</tbody>
</table>
Data Warehouse

Data Warehouse Diagram

http://www.fulcrumlogic.com/data_warehousing.shtml
Data Warehouse

- 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
OnLine Analytical Processing (OLAP)

Operations:
- Slice
- Dice
- Drill-down
- Roll-up
- Pivot

Store data in "data cubes" for fast OLAP operations. 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
Agenda

- What is Data Mining?
- Data Mining Tasks
- 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

- Statistics
- CS (machine learning, AI, data bases)
- Optimization

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

- Data management
- Statistics
- Programming
- Communication
- + Application domain