Data Mining & Analytics
Data Mining Reference Model
Data Warehouse
Legal and Ethical Issues

Slides by Michael Hahsler
Data Mining & Analytics

- Analytics is the discovery and communication of meaningful patterns in data.
- Analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.
- Analytics often favors data visualization to communicate insight.
Analytics and Visualization

- Infoviz is a field by its own.
- Napoleon's Army in Russia by Charles Minard (around 1850)
Data Mining & Analytics

- Stochastic Optimization: How can we achieve the best outcome including the effects of variability?
- Optimization: How can we achieve the best outcome?
- Predictive modeling: What will happen next if?
- Forecasting: What if these trends continue?
- Simulation: What could happen....?
- Alerts: What actions are needed?
- Query/drill down: What exactly is the problem?
- Ad hoc reporting: How many, how often, where?
- Standard Reporting: What happened?

Competitive Advantage

Degree of Complexity

Data Mining / Stats
- Statistics
- Machine Learning

OR

DB / CS

Based on: Competing on Analytics, Davenport and Harris, 2007
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
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*
Problem: Mining Point of Sale (POS) Data

<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</strong></td>
<td><strong>Select Modeling Techniques</strong></td>
<td><strong>Evaluate Results</strong></td>
<td><strong>Plan Deployment</strong></td>
</tr>
<tr>
<td>Business Background</td>
<td>Rationale for Inclusion/Exclusion</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></td>
<td>Modeling Assumptions</td>
<td>Approved Models</td>
<td></td>
</tr>
<tr>
<td>Business Success Criteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Assess Situation</strong></td>
<td><strong>Clean Data</strong></td>
<td><strong>Generate Test Design</strong></td>
<td><strong>Review Process</strong></td>
<td><strong>Plan Monitoring and Maintenance</strong></td>
</tr>
<tr>
<td>Inventory of Resources</td>
<td>Data Cleaning Report</td>
<td>Test Design</td>
<td>Review of Process</td>
<td>Monitoring and Maintenance Plan</td>
</tr>
<tr>
<td>Requirements, Assumptions, and Constraints</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risks and Contingencies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terminology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Costs and Benefits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Determine Data Mining Goals</strong></td>
<td><strong>Construct Data</strong></td>
<td><strong>Build Model</strong></td>
<td><strong>Determine Next Steps</strong></td>
<td><strong>Produce Final Report</strong></td>
</tr>
<tr>
<td>Data Mining Goals</td>
<td>Derived Attributes</td>
<td>Parameter Settings</td>
<td>List of Possible Actions</td>
<td>Final Report</td>
</tr>
<tr>
<td>Data Mining Success Criteria</td>
<td>Generated Records</td>
<td>Models</td>
<td>Decision</td>
<td>Final Presentation</td>
</tr>
<tr>
<td><strong>Produce Project Plan</strong></td>
<td><strong>Integrate Data</strong></td>
<td><strong>Assess Model</strong></td>
<td></td>
<td><strong>Review Project</strong></td>
</tr>
<tr>
<td>Project Plan</td>
<td>Merged Data</td>
<td>Model Assessment</td>
<td></td>
<td>Experience</td>
</tr>
<tr>
<td>Initial Assessment of Tools and Techniques</td>
<td></td>
<td>Revised Parameter Settings</td>
<td></td>
<td>Documentation</td>
</tr>
</tbody>
</table>

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model
Problem: How is POS data stored?

- Relational data base?
- How do the tables look like?
- Has every store/region its own data base?

- What if I want to know how many units of product A were sold in the last three month in Texas?

- This must be easier!
Data Warehouse
ELT: Extract, Transform and Load

- **Extracting** data from outside sources
- **Transforming** it to fit analytical needs. E.g.,
  - Clean (missing data, wrong data)
  - Translate (1 → "female")
  - Join (from several sources)
  - Calculate and aggregate data
- **Loading** it into the end target (data warehouse)
Data Warehouse

- **Subject Oriented:** Data warehouses are designed to help you analyze data in a certain area (e.g., sales).

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

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

- **Time Variant:** they maintain both historical and (nearly) current data.
OLAP: OnLine Analytical Processing

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

For fast operation OLAP requires a special database structure (Snow-flake scheme)
<table>
<thead>
<tr>
<th></th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>users</strong></td>
<td>clerk, IT professional</td>
<td>knowledge worker</td>
</tr>
<tr>
<td><strong>function</strong></td>
<td>day to day operations</td>
<td>decision support</td>
</tr>
<tr>
<td><strong>DB design</strong></td>
<td>application-oriented</td>
<td>subject-oriented</td>
</tr>
<tr>
<td><strong>data</strong></td>
<td>current, up-to-date</td>
<td>historical, summarized, multidimensional</td>
</tr>
<tr>
<td></td>
<td>detailed, flat relational</td>
<td>integrated, consolidated</td>
</tr>
<tr>
<td></td>
<td>isolated</td>
<td></td>
</tr>
<tr>
<td><strong>usage</strong></td>
<td>repetitive</td>
<td>ad-hoc</td>
</tr>
<tr>
<td><strong>access</strong></td>
<td>read/write</td>
<td>lots of scans</td>
</tr>
<tr>
<td></td>
<td>index/hash on prim. key</td>
<td></td>
</tr>
<tr>
<td><strong>unit of work</strong></td>
<td>short, simple transaction</td>
<td>complex query</td>
</tr>
<tr>
<td><strong># records accessed</strong></td>
<td>tens</td>
<td>millions</td>
</tr>
<tr>
<td><strong># users</strong></td>
<td>thousands</td>
<td>hundreds</td>
</tr>
<tr>
<td><strong>DB size</strong></td>
<td>100MB-GB</td>
<td>100GB-TB</td>
</tr>
<tr>
<td><strong>metric</strong></td>
<td>transaction throughput</td>
<td>query throughput, response</td>
</tr>
</tbody>
</table>
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....