Introduction to Data Mining

a.j.m.m. (ton) weijters

(slides are partially based on an introduction of Gregory Piatetsky-Shapiro)

Overview

- Why data mining (data cascade)
- Application examples
- Data Mining & Knowledge Discovering
- Data Mining versus Process Mining

Why Data Mining

- Cascade of data
  - Different growth rates, but about 30% each year is a low growth rate estimation
- The possibility to use computers to analyze data
  - 1975 computer for the whole university (main frame) with 1MB working memory, now a PC with 512 MB working memory

Cascade of data

- Business and government systems (transactions system, ERP systems, Workflow systems, ...)
- Scientific data: astronomy, biology, etc
- Web, text, and e-commerce (new regularities, about data storage to prevent attempts)
- Hospitals, internal revenue service
- ...

Examples large data bases

- AT&T handles billions of calls per day
  - so much data, it cannot be all stored -- analysis has to be done “on the fly”
- Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session
- Google

First conclusion

- Very little data will ever be looked at by a human
- Data Mining algorithms and computers are **NEEDED** to make sense and use of data.

Overview

- Why data mining (data cascade)
- Application examples
- Data Mining & Knowledge Discovering
- Data Mining versus Process Mining

Application examples I

- Customer Relationship Management (CRM)
  - Based on a data base with client information and behavior try to select other potential consumers of a product.
  - Euro miles.
- Profiling tax cheaters
  - Based on the profile of the tax payer and some figures from the tax (electronic) form try to product tax cheating.

Application examples II

- Health care
  - Given the patient profile and the diagnoses try to predict the number of hospital days. Information is used in planning system.
- Industry
  - Job shop planning. Based on already accepted jobs, try to product the delivery time of a new offered job.

Type of applications

- **Classification** (supervised)
  - Credit risk: result of data mining are rules that can be used to classify new clients as: high, normal, low
- **Estimation** (supervised)
  - Credit risk: output is not a classification but a number between -1 and 1 to indicate risk (-1.0 very low, 0.0 normal, +1.0 very high)
- **Clustering** (unsupervised)
- **Associations:** e.g. Bier & Chips & Peanuts occur frequently in a shopping list of one person
- **Visualization:** to facilitate human discovery

Supervised verses unsupervised

- Supervised (Credit risk)
  - Starting point is a historical data base with client information and his/her financial data including credit history (classification). This data base is used to induce credit risk rules.
- Unsupervised (Clustering)
  - Try to cluster customers into similar groups (how many groups, in which sense similar)

E-commerce – Case Study

- A person buys a book (product) at Amazon.com.
- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought “Advances in Knowledge Discovery and Data Mining”, also bought “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”
- Recommendation program is quite successful

Hands-on-project I

- Historical consumer data
  - Age, education, sex, relationship, etc.
  - Income
- Model to predict income above 50K
- Use the model to select consumers for direct mailing

Problems Suitable for Data-Mining

- have sub-optimal current methods
- have accessible, sufficient, and relevant data
- provides high payoff for the right decisions!
- (have a changing environment)

Overview

- Why data mining (data cascade)
- Application examples
- Data Mining & Knowledge Discovering
- Data Mining versus Process Mining

Knowledge Discovery Definition

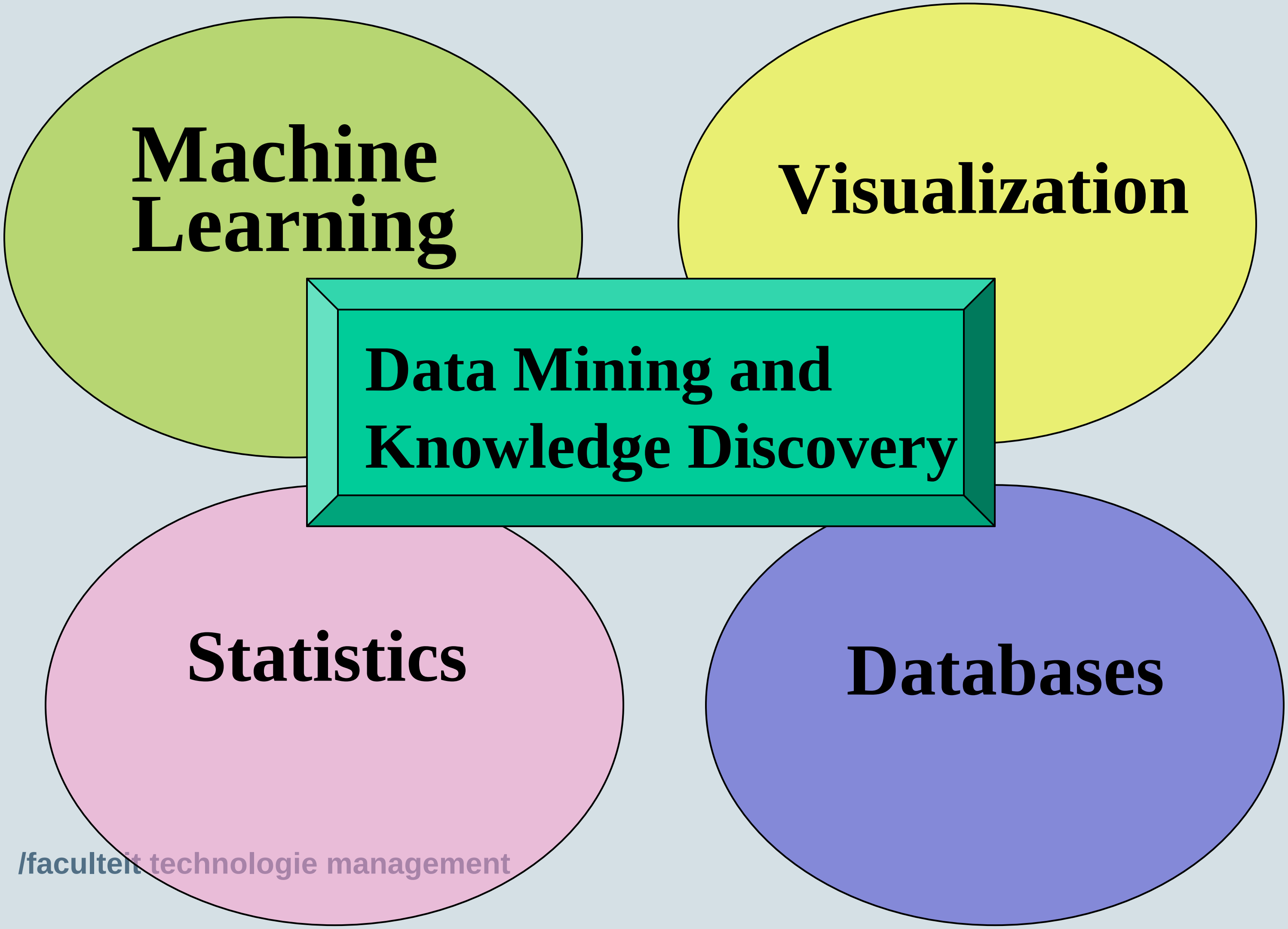
Knowledge Discovery in Data is the

non-trivial process of identifying

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

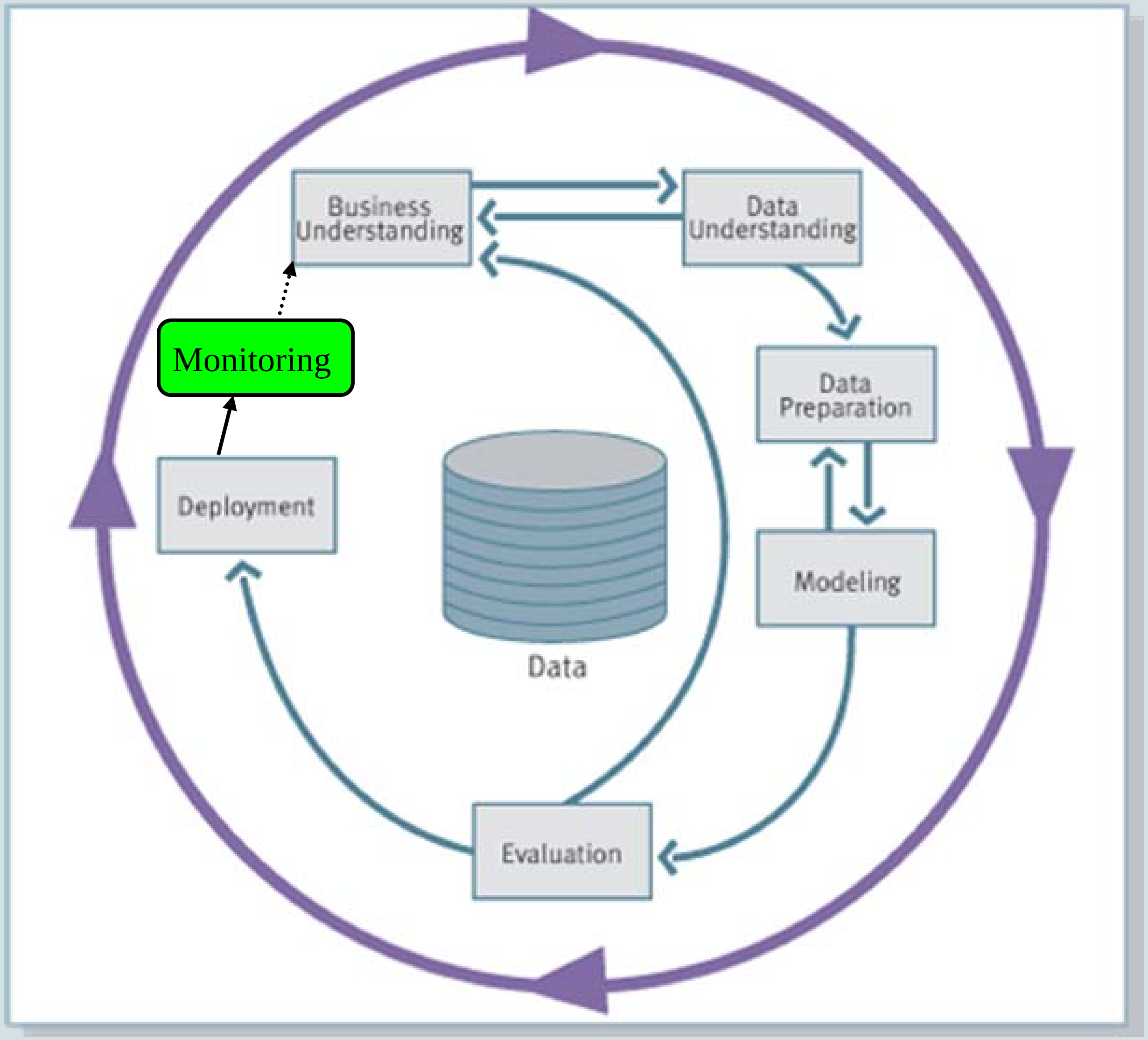
Related Fields



Statistics, Machine Learning and Data Mining

- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine Learning
  - more heuristics then theory-based
  - focused on improving performance of a learning algorithms
- Data Mining and Knowledge Discovery
  - Data Mining one step in the Knowledge Discovery process (applying the Machine Learning algorithm)
  - Knowledge Discovery, the whole process including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

Knowledge Discovery Process flow, according to CRISP-DM



Business Understanding + Data Understanding + Data Preparation 80% of the time

Modeling (applying mining algorithm) 20%

Phases and Tasks

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Asessment 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

Data Set
Data Set Description

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

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

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

Other related fields

- Data warehouse
  - A data warehouse thus not contain simply accumulated data at a central point, but the data is carefully assembled from a variety of information sources around the organization, cleaned u, quality assured, and then released (published).
- Business Intelligence (BI)
  - The use of data in the data ware house to support the managers with important information

Overview

- Why data mining (data cascade)
- Application examples
- Data Mining & Knowledge Discovering
- Data Mining versus Process Mining

Data Mining versus Process Mining

- Process Mining is data mining but with a strong business process view.
- Some of the more traditional data mining techniques can be used in the context of process mining.
- Some new techniques are developed to perform process mining (mining of process models).

Why Process Mining

- Traditional As-Is analysis of business processes strongly based on the opinion of process expert. The basic idea is to assemble an appropriate team and to organize modeling sessions in which the knowledge of the team members is used to build an adequate As-Is process model.
- The surplus values of process mining in the As-Is analysis are:
  - information based on the real performance of the process (objective)
  - more details