Data mining in practice

T-61.3050

27.11.2007
Xtract / Juha Vesanto

XTRACT

Xtract Ltd
Hitsaajankatu 22
00810 Helsinki
FINLAND

T +358 9 222 4122
F +358 9 222 4155
contact@xtract.com
www.xtract.com

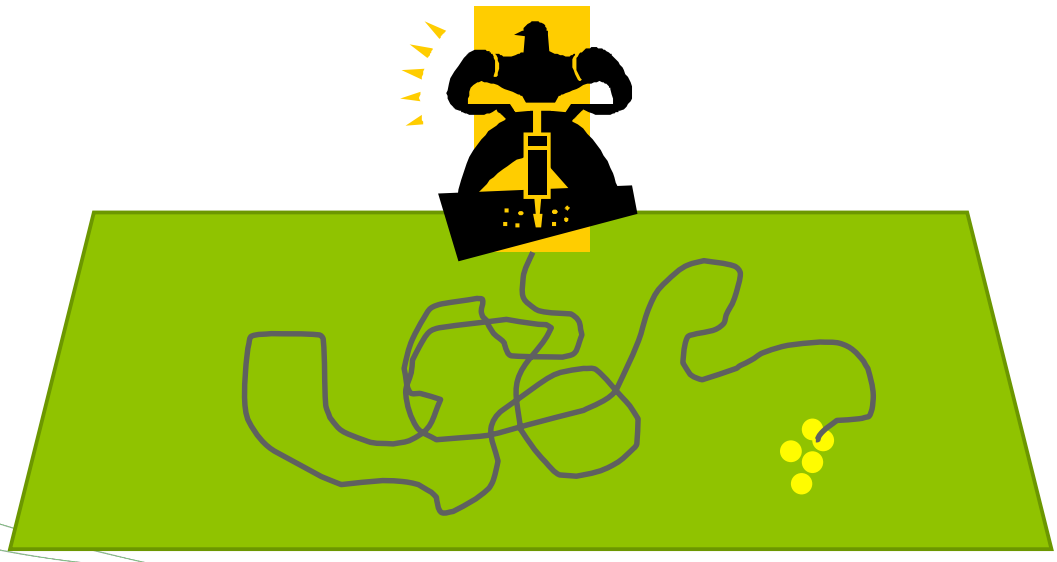
Intro

- My history
  - Juha Vesanto
  - M.Sc. in Engineering Physics 1997
  - Dr. Tech. in Information Science 2002
    - IDE research group
    - Dissertation: "Data mining using the Self-Organising Map"
- Xtract history
  - Founded in 2001
  - Main areas of operation:
    - analytics and business consulting on data-based analytics
    - software and integration services
    - data
  - Analytics specialities
    - customer analytics
    - segmentation, targeting
    - social network analytics
  - Personnel: 40-50 in Helsinki, London, and sales representatives elsewhere
  - This year forecasted revenue: >3.5 M€
  - Customers: Nokia, SanomaMagazines, Lehtipiste, Tradeka, Luottokunta, Vodafone, ...

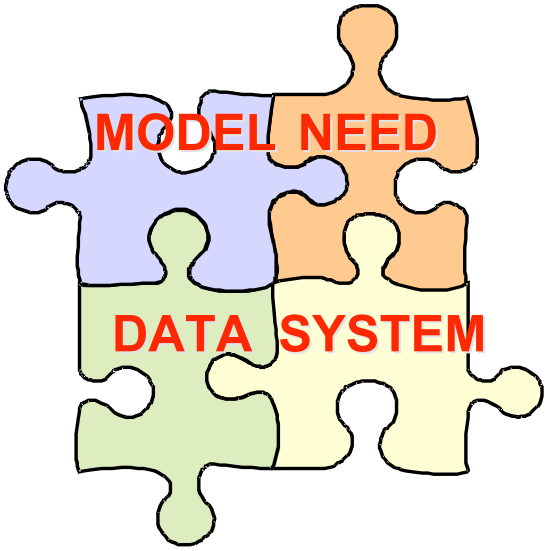
Data mining in practice

BUSINESS DATA MINING

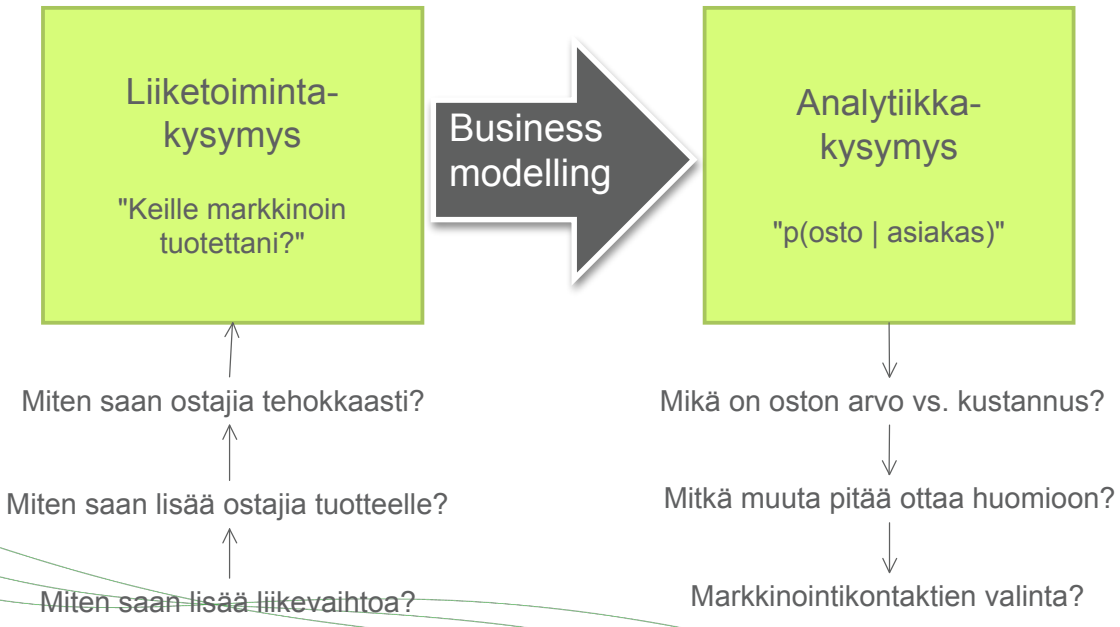
Data mining in practice - not



Business data mining



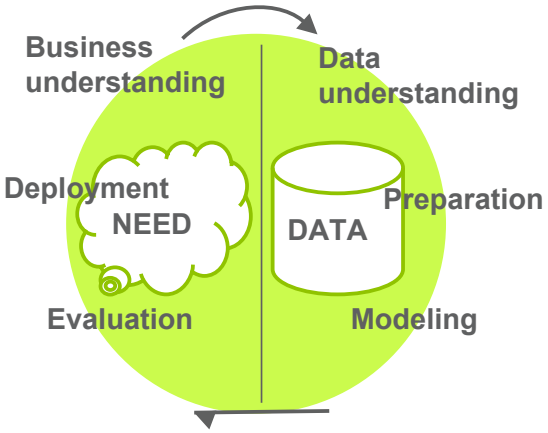
Business modelling



Business and analytics viewpoints

Business viewpoint

- modelling answers business needs
- aims at results deployment



Analytics viewpoint

- data mining is about finding something interesting from data
- data mining starts with and revolves around data

Data mining in practice

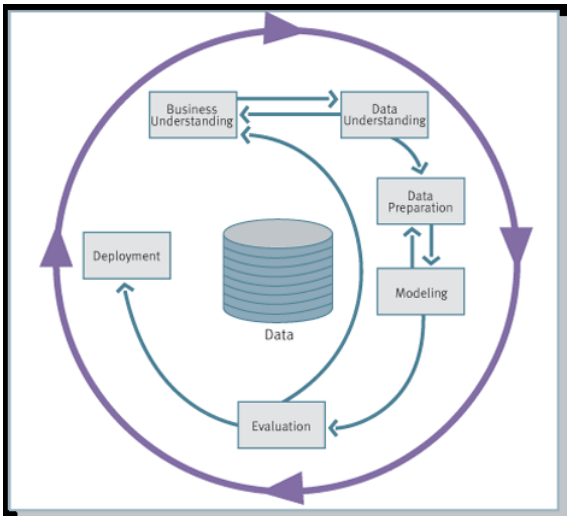
DATA MINING PROCESS

CRISP-DM

CRoss-Industry Standard Process for Data mining

www.crisp-dm.org

partners: Teradata, SPSS, DaimlerChrysler, OHRA
+ special interest group



"51% of data miners use CRISP-DM methodology"

http://www.kdnuggets.com/polls/2002/methodology.htm

CRISP-DM Phases

1. Business understanding
- business need
- data mining target
- project planning

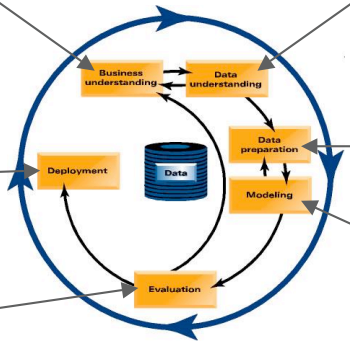
2. Data understanding
- data collection
- data review

3. Data preparation
- data preprocessing
- data enrichment
- feature extraction

4. Modeling
- model family selection
- model optimization
- model testing
- model review

5. Evaluation
- validation w.r.t. the need
- results review

6. Deployment
- taking results into use
- model monitoring
- updating the model



Business modelling

PRACTICE

Business & data understanding

Business

- Ymmärrä asiakkaan toiminta
  - Mikä on asiakkaan tavoite?
  - Mitä asiakas oikeasti tarvitsee?
  - Mitä toimenpiteitä asiakas on valmis / tottunut tekemään?
  - Mitä muita tekijöitä täytyy ottaa huomioon?
- Selvitä stakeholders
  - Kuka on oikeasti maksaja / tilaaja?
  - Kuka oikeasti käyttäisi tuloksia?
- Selvitä ja aseta tavoite
  - Mikä on tilaajan tavoite (lv, kate, pull, markkinaosuus)?
  - Mitä tilaaja odottaa projektin lopputuloksena?
  - Mitä tilaaja on ajatellut tekevänsä tuloksilla?

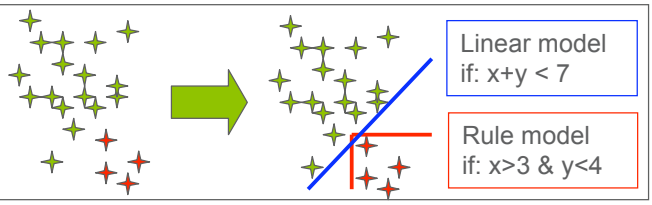
Data

- Ymmärrä asiakkaan data
  - Mitä dataa asiakkaalla on olemassa?
  - Mistä se tulee, ja milloin sitä päivitetään?
- Mallinnus
  - Miten data käännetään tuloksiksi?
  - Mallin rakenne → luotettavuus, toistettavuus, tulosten taso
- Data → Ratkaisu
  - Miten dataa voidaan käyttää ratkaisemaan asiakkaan ongelma?
  - Miten asiakas käytännössä tekee analytiikan antamilla tuloksilla?

Data preparation: compensate for imperfect nature of the data

In principle

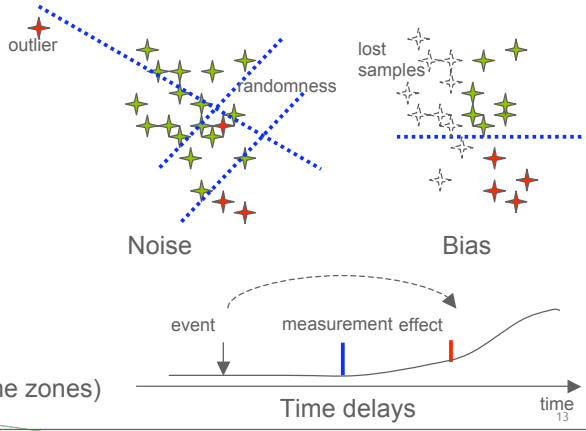
Analytical models aim at building a faithful representation of the real world



In practice

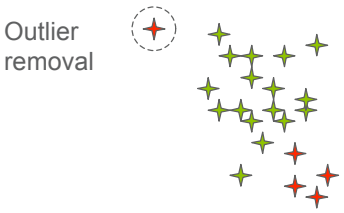
Practical difficulities arise from

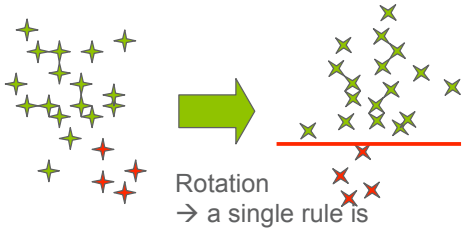
- Measurements
  - what can be measured?
  - what has been measured?
  - timing of measurements
- Data collection
  - vague concepts → misunderstanding
  - typing errors
  - differences in system settings (e.g. time zones)



Data preparation

- Read data from the data sources
- Clean the data
- Make relevant information more clearly visible
- Data enrichment
- Transform data to fit the assumptions of the modelling technique
- **Usually 80% of the work (and typically 50-90% of the end result)**





Data enrichment: CLC classes

1. Tenant suburbs of younger singles and couples

- Lower and middle income housing, occupied by students, junior administrative and service employees.
- Rental apartments in larger towns.
- High concentration of unemployment and people with low incomes.

2. Singles in city apartments

- Young singles or couples without children in small apartments
- Well-educated, very involved in their work.
- Prefer the vitality of the large city to the tranquility of outer suburbs.
- Low income per households (due to large share of singles).

3. Middle class in apartments

- Residential neighborhoods on the outskirts of towns and cities, mainly private housing,
- Younger singles and couples in their 30ies. The educational, income and wealth figures are raising; low unemployment

4. Well educated, high income families

- High income families in the more affluent suburbs,
- Professionals and wealthy business-people living in large and expensive owner-occupied houses.
- Two-income, two-car households.

5. Countryside

- Rural areas where agriculture and industry (where industry still remains) remain a significant source of local employment.
- Considerable variance in the levels of affluence, from the old family farm areas to the quiet small villages of only retired farmers and workers.

6. Middle class in detached houses

- (Once) less expensive areas of large detached houses in outskirts of small and medium-sized towns
- Skilled manual and white-collar workers with their families. Low rate of unemployment.
- Unpretentious areas, where sensible and self-reliant people have worked hard to achieve a comfortable and independent lifestyle.

7. Small income detached house areas

- Middle-aged households living in detached houses with small income.
- High unemployment rate, limited assets. Industry is or has been the most important employer.
- Areas located near the industrial centers of Finland.

8. Retiree areas

- Retired and soon-to-be-retired singles and couples, who typically own their houses or apartments.
- High levels of discretionary expenditure (Low household income, but low expenditure on rent, mortgages and children)

Modelling

Task	Question	Modelling
Targeting	"I want to market my product. I could send my ad to 1 million people, but I only except 2000 orders, so that's 998000 useless letters..."	Predictive scoring model • based on an earlier campaign • using available Case: publishers, banks, retailers, ...
Segmentation	"I have 1 million customers. They are a grey mass. Help?"	Segment the customers into actionable groups. Case: just about anybody, eg. operators
Pricing	"I need to set the price for my product. What is the optimal price?"	Price elasticity model log(dprice) ~ -a log(dvol) Case: just about anybody, eg. retailers
Logistics	"I have 500 retail outlets. How many products should I ship to each outlet to ensure optimal coverage?"	Seasonal variation models Case: retailers, e.g. Lehtipiste
Fraud detection	"I need to identify fraudulent credit card transactions."	Predictice scoring models Likelihood models

Analytical evaluation (& validation)

There are several ways to look at the data and the results. For the best results, it is best to check the data from all of these angles.

1. Statistics
   - compare statistics of input and output data tables (starting with N=number of samples): do they match, are the deviations as intended by the preprocessing ?
   - correlations
   - result statistics: check score histograms, segment sizes
   - model statistics
2. Cases / samples
   - pick 1-5 sample data cases, and go through the processing by hand: are the results as intended ?
3. Common sense
   - go through the results (cross-tabulations, deductions, histograms, decile profiles): do they make sense ?
4. Code review
   - what is the processing script / pipeline / program??
   - go through the code and try to find logical inconsistencies etc.

Business evaluation

Are the results practically usable?

Review by end users

Design and pilot field tests

Deployment

Lvl	Operation	Action	Benefits
1	Internal analytics	Data mining activity Distribution of the results to organization Utilization of results	Better understanding of the data for the data miner, and to the organization. Direct economic value through increased efficiency, decreased costs, or bigger revenue.
2	Repeated analytics (backoffice)	Monitoring and follow-ups	Better understanding of business & data. Identification of further opportunities. Continuing increases in economic value.
3	Scheduled analytics (batch)	Planned, scheduled updates that tie in with business processes	Further efficiency from regular usage No risk from applying outdated models
4	Integrated analytics (online)	Continuous updates to the model and scores	Reoccuring benefits from the continuously applied model Minimized operational costs & risks

Contact Details

Juha Vesanto
M +358 40 750 5515
juha.vesanto@xtract.com

Xtract Ltd
Hitsaajankatu 22
00810 Helsinki
FINLAND

T +358 9 222 4122
F +358 9 222 4155
contact@xtract.com
www.xtract.com