Data Science
Topic: Data Mining
Intro
&
Case Studies

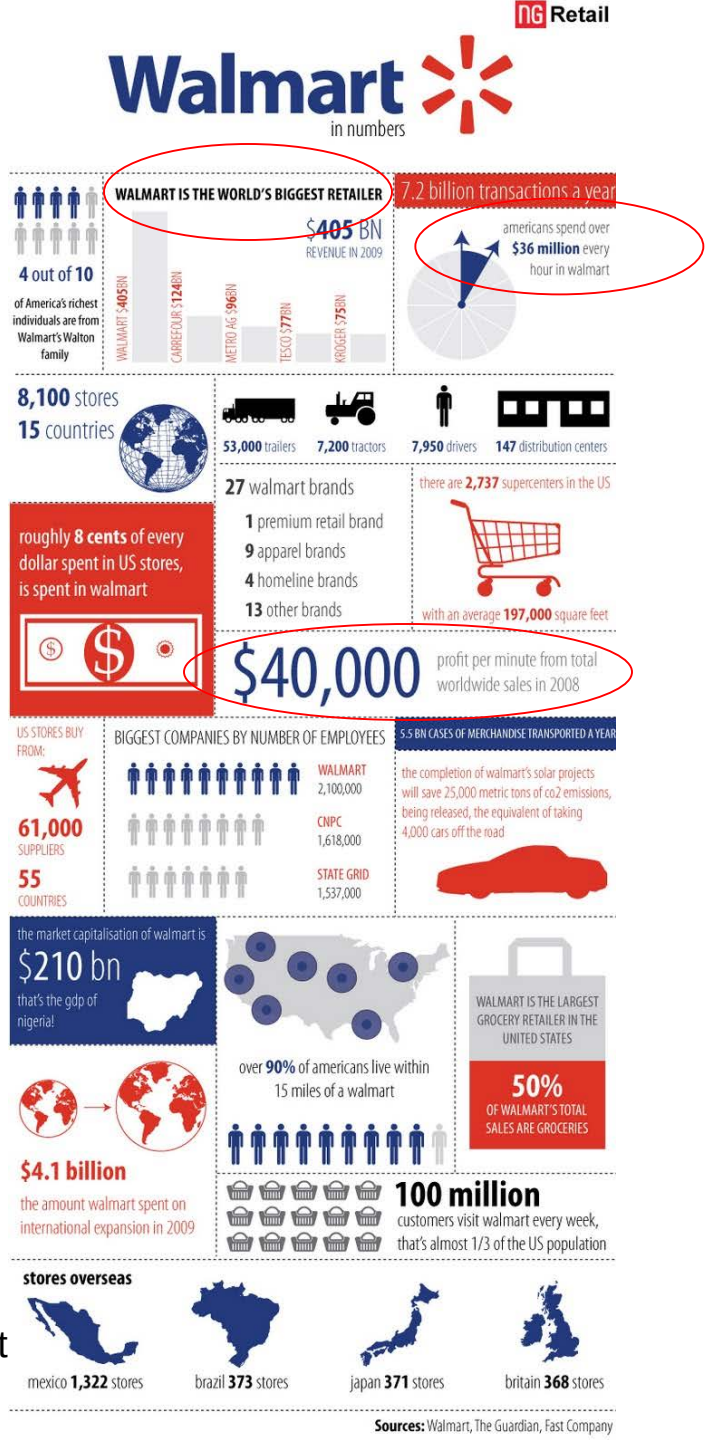
Mannes Poel

Why Data Mining

- There is a gap between data creation and our understanding of (the knowledge hidden in) the data.
- Classical examples (outdated figures):
  - YouTube: more than 5 hour of videos uploaded every second.
  - WalMart: more than 7.2 billion transactions a year.
  - Webpages: at least 7.44 billion pages.
  - Visa CC transaction volume: more than 30.7 billion.

Walmart

Want to know more? Search on “Walmart in numbers”



How to increase revenues?

Dat

Exploring new domains: Health

- E-Health: Interaction between clients and care givers shift from the physical domain to the virtual domain.
  New challenges for client profiling and recommender systems.
- For instance:
  - Can one give a patient (semi-)automatic referral advice based on a questionnaire?
  - Can one predict if a client drops out of eHealth program, for instance a stop smoking program? Similar to churn management in the telecom industry.
  - Safe and early dispatch of patients.

Exploring new domains

- Sentiment analysis in the social or health domain. Does someone feel depressed?
- Data driven recommender systems in the health domain.
- Sports mining
- Automatic image description (first step to explanation-based classification of images.

Applications in the medical domain

- Automatic Atrial Fibrillation detection
- Scheduling of OR: prediction of surgery duration.
- Prognostication of Postanoxic Coma Patients
- Smart watches: lots medical and physical data of humans under natural circumstances. Automatic detection of an outbreak of an epidemic of influenza.

Relationships

AI can predict whether your relationship will last based on how you speak to your partner

September 29, 2017 10.25am BST

I'm TALKING. Roman Samborskyi/Shutterstock

website

Other challenges?

Who knows other data mining challenges for or possible fruitful applications of Data Mining?

How AI is changing medicine

Regina Barzilay teaches computers how to learn.
A professor at the Massachusetts Institute of Technology, her work focused on natural language processing – training computers to understand human speech – until a breast cancer diagnosis three years ago.
“Going through it, I realized that today we have more sophisticated technology to select your shoes on Amazon than to adjust treatments for cancer patients,”

Summarizing

- Knowledge Generation and Knowledge based systems are more and more data and data mining driven.

- Weakness: knowledge is most of the time rather implicit coded in the system so hard to check. This resulted in branch of research called Explainable AI

Conclusion

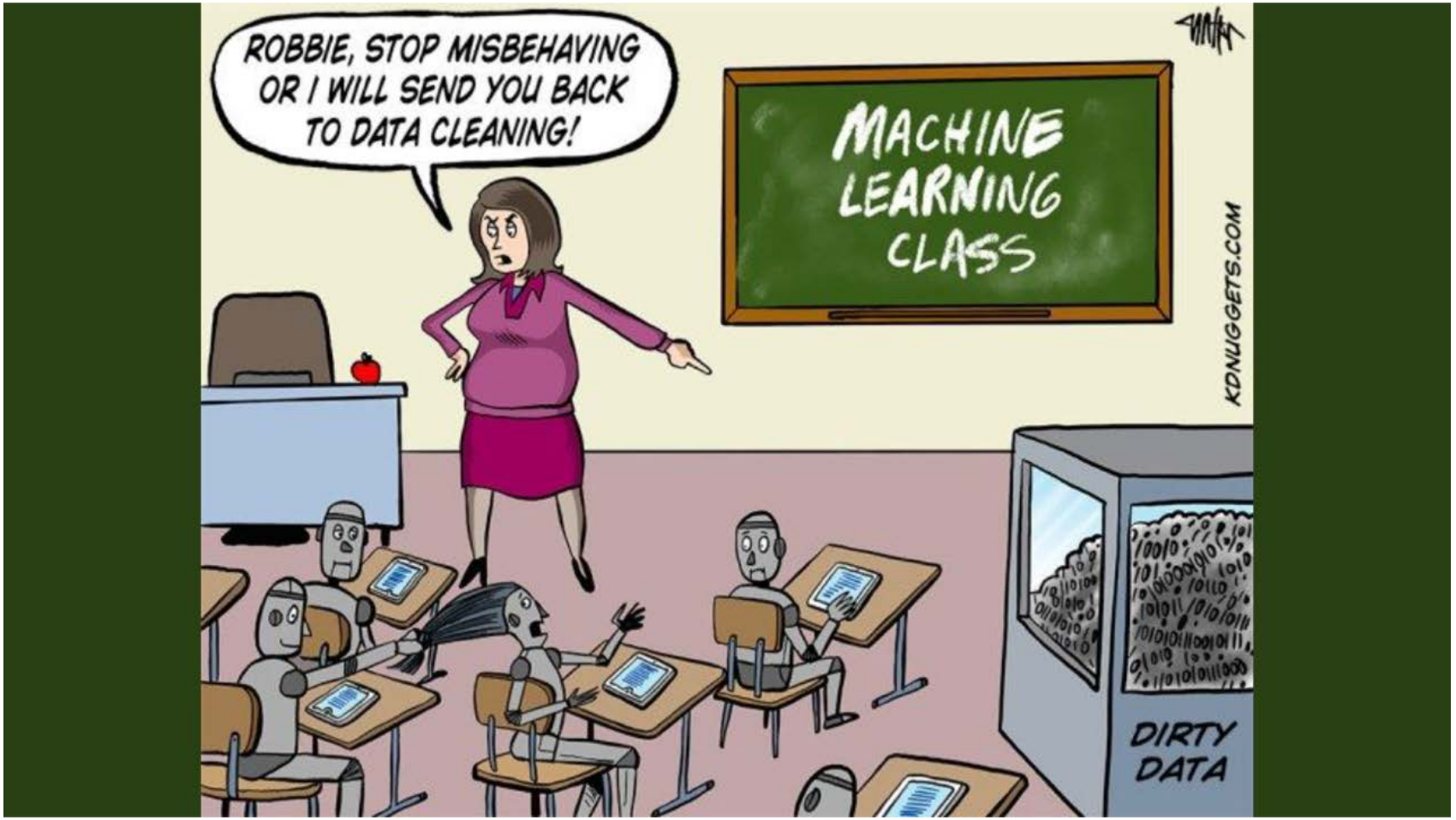
- Need for techniques which can **automatically** extract **knowledge** from data.
- We need tools to automatic:
  - Data cleaning and preparation. Extract/construct relevant (high level) features from data. Transforming the object of interest to feature space.
  - Analyze data in feature space (visualization, outlier detection, data quality)
  - Find useful (relevant) knowledge (information) in feature space.

This is all part of Data Mining

Data Cleaning/Preparation

- Load data in suitable DB.
- Detecting and removing outliers.
- Dealing with missing values. Removing data or imputing missing values.
- Dealing with class imbalance
- Enrich the data with data from external sources.
- Transforming feature values. Some DM models can only deal with numerical values.

Data Cleaning/Preparation



How is this knowledge represented?

- Pattern:
  - Classification algorithm/model (a relation between features and categories) which classifies new data into categories.
    Knowledge is most of the time implicit (black box).
  - A pattern in the data that repeats.
  - Relations in data of behavior that are not obvious, for instance association rules in basket data analysis.
    Explicit knowledge in terms of rules.
  - A model to predict.
    Knowledge is most of the time implicit.

How to Extract Useful Patterns from Data?

- By hand, but most of the time not feasible!
- Let the Computing Machine do it, this is called Data Mining/Machine Learning/ Patter recognition/....
  - But feature engineering (transforming object to feature space) is still a challenge in some domains. Especially in domains for which there is not that much data.

Assumptions

- The Past is a Good Representation of the Future.
- The Data is Available.
- The Data Contains the Knowledge, Information, Patterns one Wants to Extract.

Garbage in ➔Garbage out!

Case study 1: Classifying Iris flowers

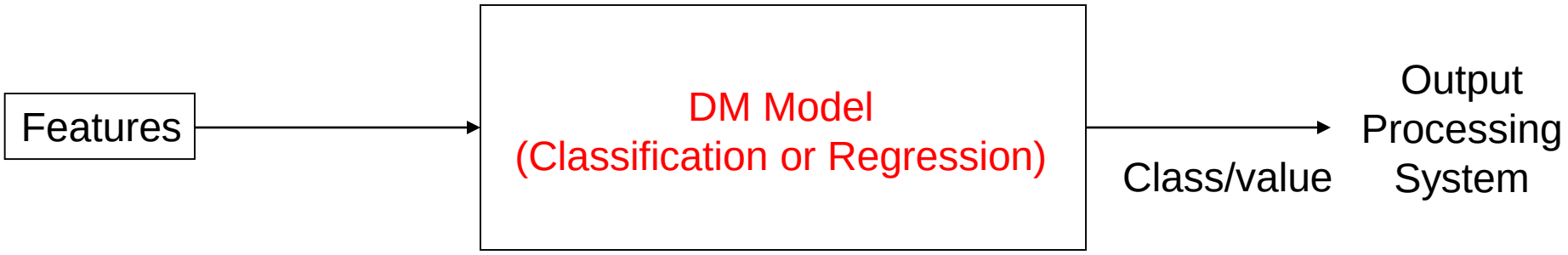
Setosa

Versicolor

Virginica

DM problem:
Can one build an application (model) which automatically classifies Iris flowers

On a conceptual level



Based on the given features, automatically classify the Iris flowers

1. In almost all Data Mining applications features are given or constructed from available raw data or data bought from external sources.
2. Historical examples (so the available data) are stored in a database.

Iris data

Sepal length	Sepal width	Petal length	Petal width	Class
5.1	3.5	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.3	3.3	6.0	2.5	virginica
…..	…..	…..	…..	…..
…..	…..	…..	…..	…..
…..	…..	…..	…..	…..
…..	…..	…..	…..	…..

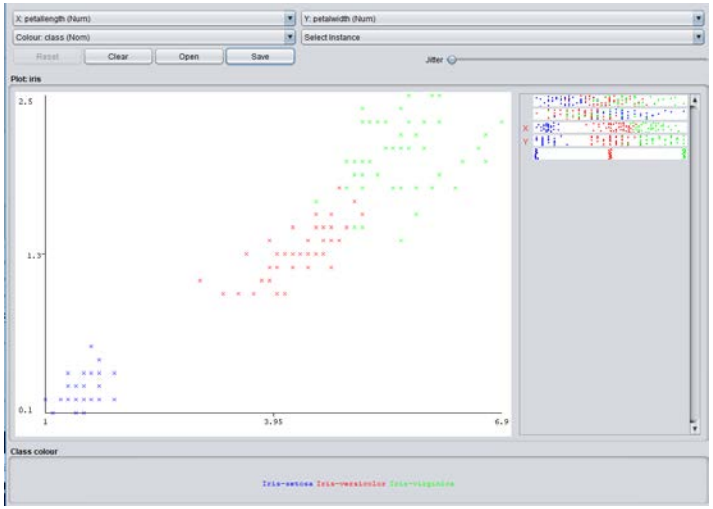
First step: Visual analysis of the data (if possible)

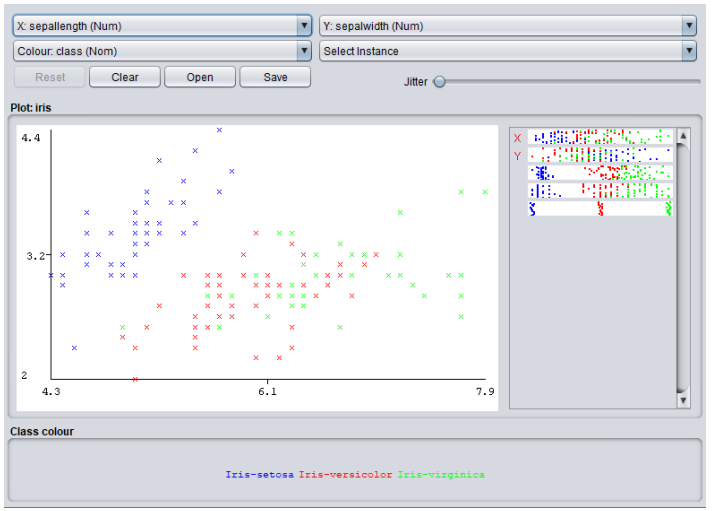
- Why:
  - Inspect data quality
  - Class distributions
- How:
  - Use of data mining tools (R, Matlab, Octave, Weka, SPSS, etc.)

For this lecture I will use WEKA to study and mine the data.

Visualizing the data

- Visualize the projection on two-dimensional attribute spaces (scatterplot).





- Which tuple of attributes is most predictive?

Build model to classify the data

- Naïve Bayes
- Decision Tree (J4.8)
- Which model is best?

Performance Measures

- Accuracy
- Precision
- Recall
- Sensitivity
- Specificity
- Or a combination these.
- All these measures can be computed from the confusion matrix!!!

Confusion matrix

Predicted class

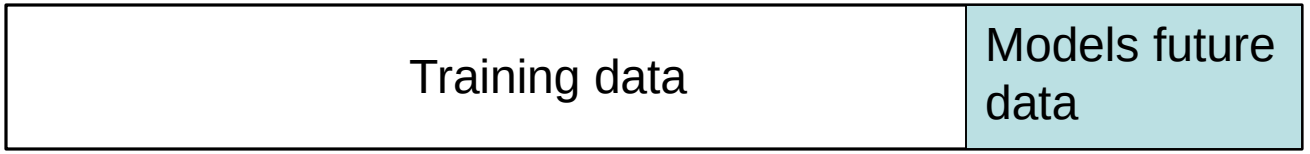
Act /Pred	Spam	Ham
Spam	97 (TP)	51 (FN)
Ham	11 (FP)	825 (TN)

- Accuracy?
- Recall for Spam?
- Precision for Ham?

- Answers: 0.937, 0.898 and 0.987

How to measure the performance?

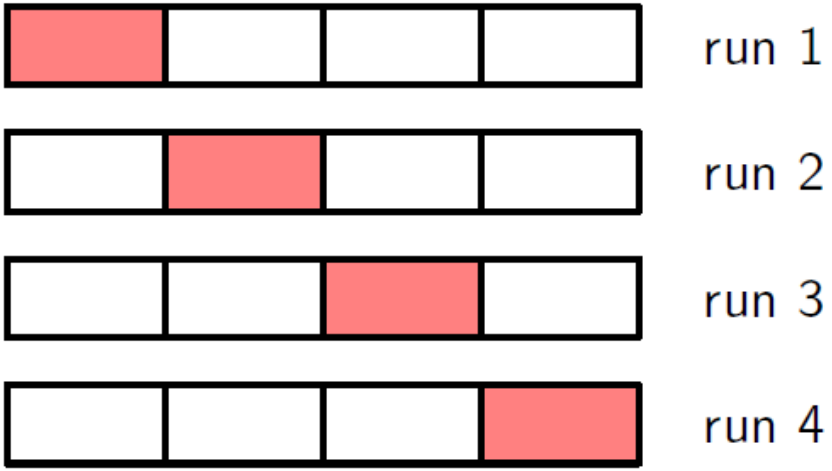
- Indication of performance on new data, future data.
- So needs to model data that will be generated in the future.
- Use part of your available data.



- Train model on training data and asses performance on model of future data.

K-fold Cross Validation

- Split available data in K equal pieces.
- Train model on K-1 pieces and test on 1 piece.
- Repeat K times picking each time a different test piece.
- Each data point is used K-1 times for training and 1 time for testing.



Predicting Bankruptcy

Based on annual reports.

Challenge

The goal DM challenge is to design and validate a data mining approach for classifying companies. This classification must be done on basis of a certain amount of financial features determined from the annual reports of the companies. The companies must be classified in two categories:

1. not bankrupt within 5 years of the publishing date of the annual report and
2. bankrupt within 5 years after the publishing date of the annual report.

Feature Engineering

The financial features extracted from the annual reports are determined by financial experts. But the relation between these financial features and the two classes, bankrupt or not bankrupt, are not clear, not even to experts.

Features

- Profitability ratios:
  - Column 1: net profit / equity
  - Column 2: net profit / total assets
  - Column 3: earnings before interest and taxes / total assets
  - Column 4: earnings before deprecation, interest and taxes / total assets

Features

- Liquidity ratios:
  - Column 5: current assets / current liabilities
  - Column 6: working capital / total assets
  - Column 7: working capital / sales

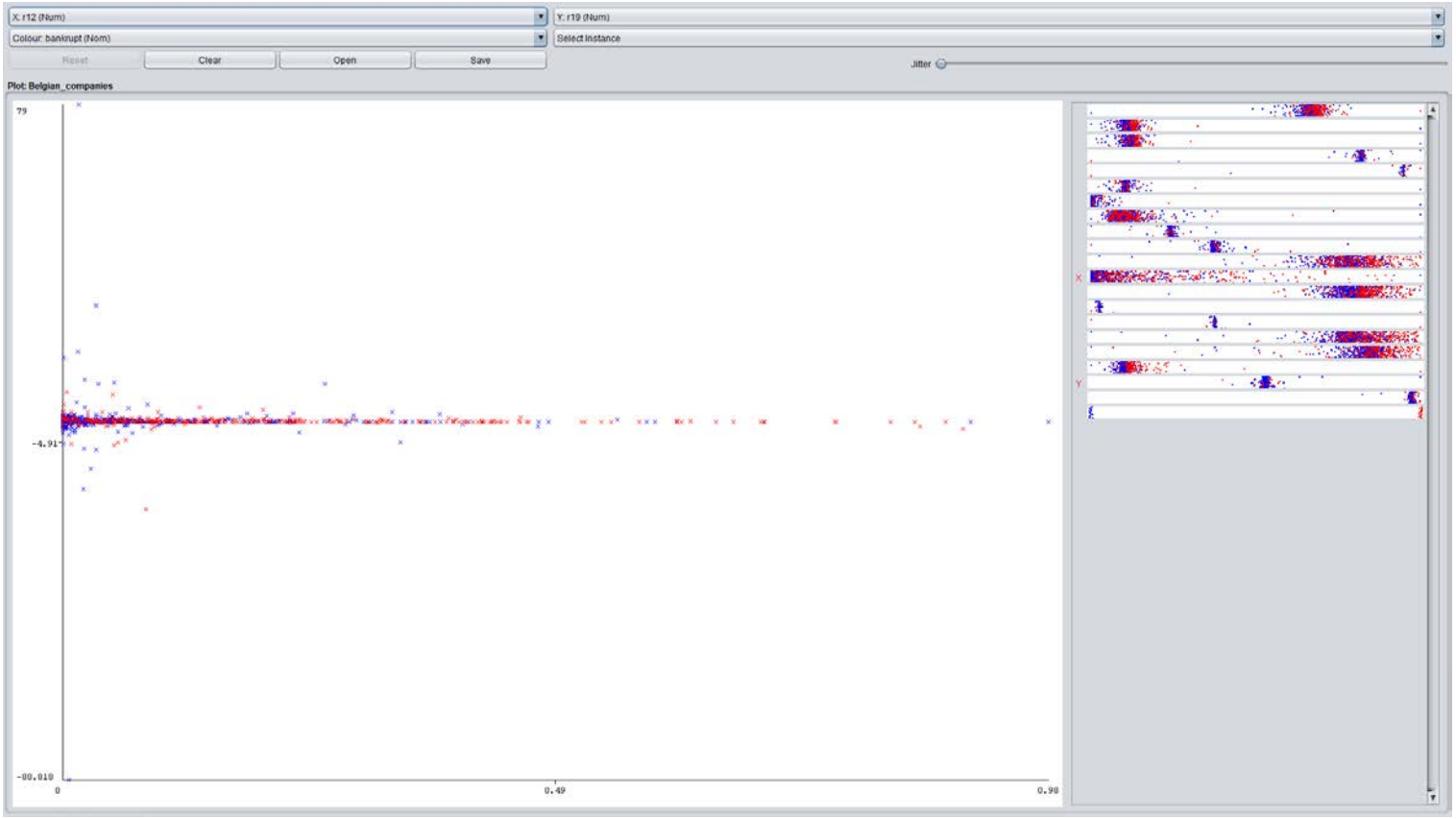
Features

- Activity ratios:
  - Column 8: accounts receivable / sales
  - Column 9: accounts payable / sales
  - Column 10: inventory / sales
  - Column 11: net profit / total assets
  - Column 12: net profit / sales
  - Column 13: earnings before deprecation, interest and taxes / sales
  - Column 14: earnings before interest and taxes / sales
  - Column 15: added value / total assets
  - Column 16: added value / fixed assets

Features

- Coverage ratios:
  - Column 17: equity / total assets
  - Column 18: cash flow / total debt
  - Column 19: retained earnings / total assets
  - Column 20: earnings before interest and taxes / interest

Visualization (scatterplots)



Not that much info

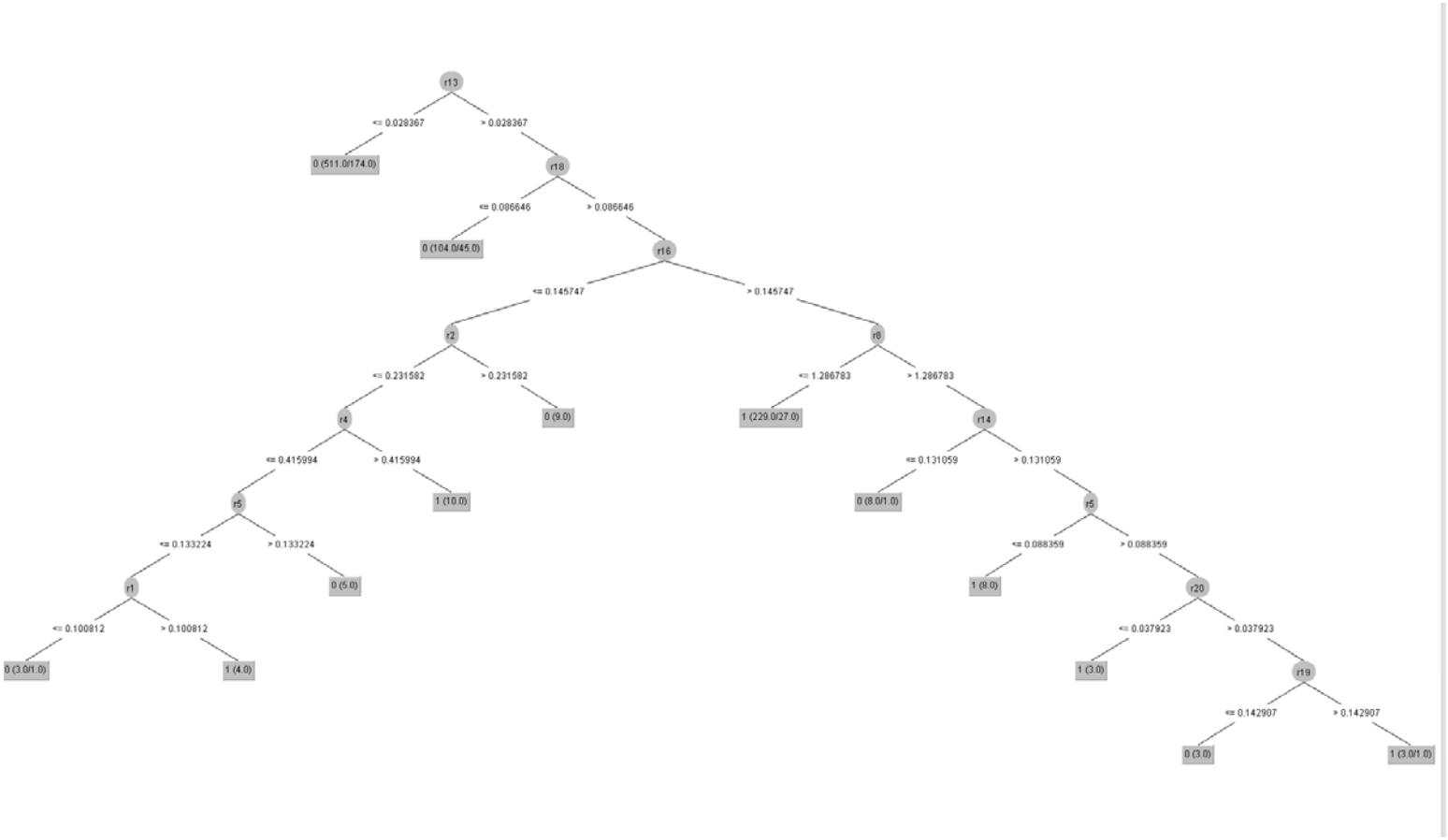
Classification (Logistic regression)

```
Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         629               69.8889 %
Incorrectly Classified Instances       271               30.1111 %
Kappa statistic                          0.3978
Mean absolute error                      0.4131
Root mean squared error                  0.4592
Relative absolute error                 82.6128 %
Root relative squared error             91.8466 %
Total Number of Instances              900

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,778     0,380     0,672       0,778   0,721      0,403    0,738     0,670     0
               0,620     0,222     0,736       0,620   0,673      0,403    0,738     0,740     1
Weighted Avg.  0,699     0,301     0,704       0,699   0,697      0,403    0,738     0,705

=== Confusion Matrix ===

   a   b   <-- classified as
 350 100 |   a = 0
 171 279 |   b = 1
```

Classification
Decision Trees

```
Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         607               67.4444 %
Incorrectly Classified Instances       293               32.5556 %
Kappa statistic                          0.3489
Mean absolute error                      0.3975
Root mean squared error                  0.4826
Relative absolute error                 79.4994 %
Root relative squared error             96.5269 %
Total Number of Instances              900

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,762    0,413    0,648      0,762    0,701      0,354    0,684     0,628     0
               0,587    0,238    0,712      0,587    0,643      0,354    0,684     0,660     1
Weighted Avg.  0,674    0,326    0,680      0,674    0,672      0,354    0,684     0,644

=== Confusion Matrix ===

   a   b   <-- classified as
 343 107 |   a = 0
 186 264 |   b = 1
```

Visualization of the DT



Classification
Random Forest

```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.35 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         634               70.4444 %
Incorrectly Classified Instances       266               29.5556 %
Kappa statistic                          0.4089
Mean absolute error                      0.3897
Root mean squared error                  0.4475
Relative absolute error                 77.9378 %
Root relative squared error             89.4967 %
Total Number of Instances              900

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,773    0,364    0,680      0,773    0,723      0,413    0,751     0,704     0
               0,636    0,227    0,737      0,636    0,683      0,413    0,751     0,762     1
Weighted Avg.  0,704    0,296    0,708      0,704    0,703      0,413    0,751     0,733

=== Confusion Matrix ===

   a   b   <-- classified as
 348 102 |   a = 0
 164 286 |   b = 1
```

Medical fraud detection

Example of unsupervised data mining

Problem description

- Medicaid & Medicare (U.S. national social insurance programs)
- Over 70M enrollees, high increasingly growing expenditures $432B
- Total expenditure of 17.6% of the Gross Domestic Product (GDP) in 2010
- Health care fraud estimates ranging from 3 to 10%
- Fraud detection: audits, market signals and automatic fraud detection

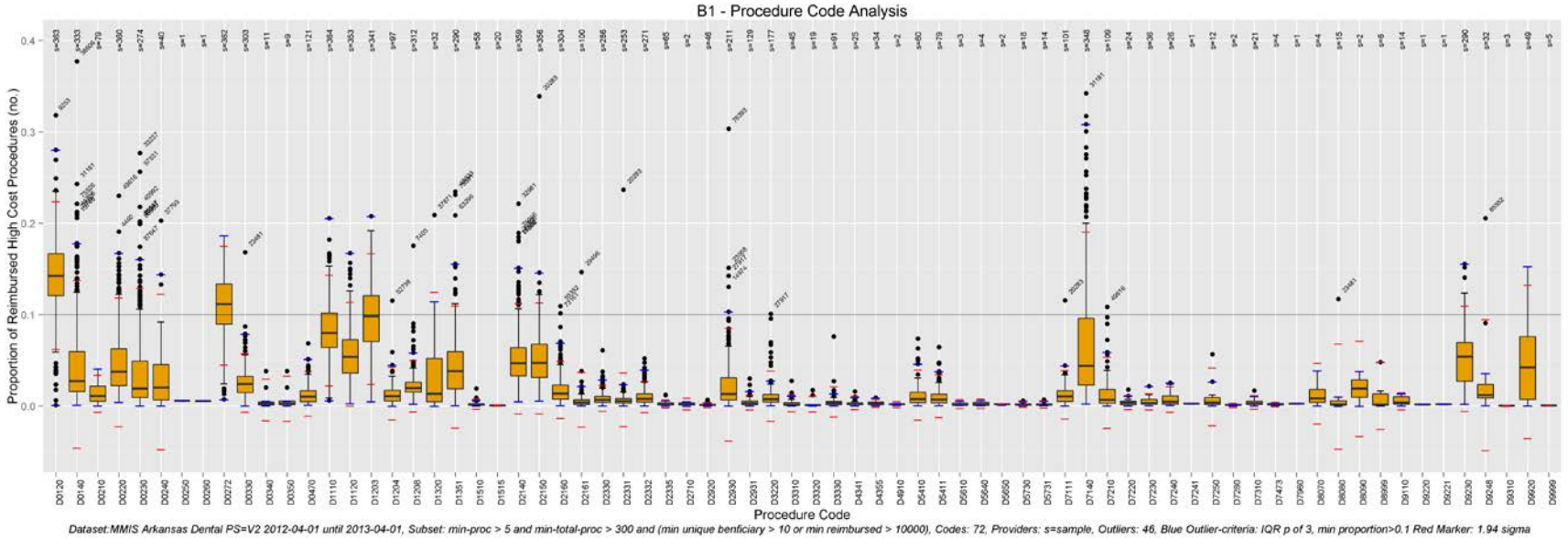
Medical Fraud detection: Goal

- Goal: Automatic warning system for medical billing fraud; fingerprint of possible fraudulent behavior.

- Mostly unlabelled data: unsupervised techniques such as clustering must be used.

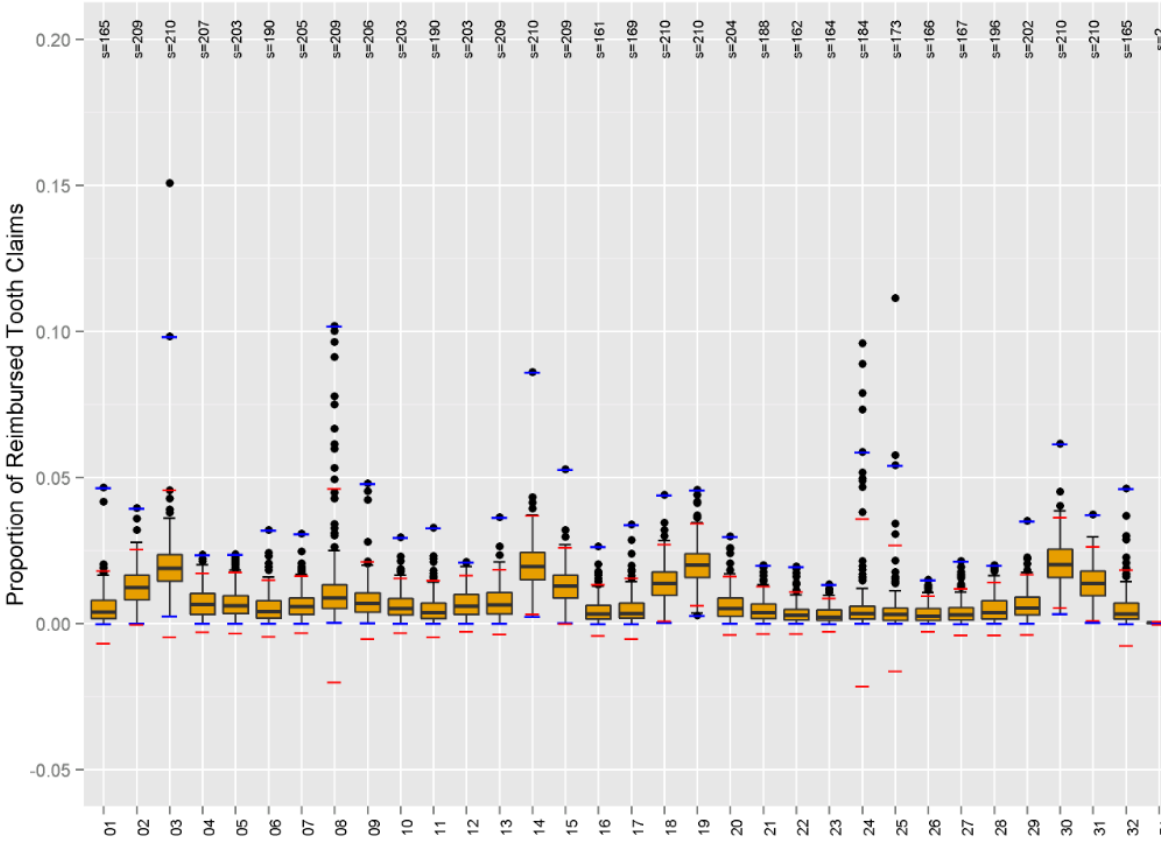
Medical Fraud detection: example of fingerprint

Declaration code analysis:



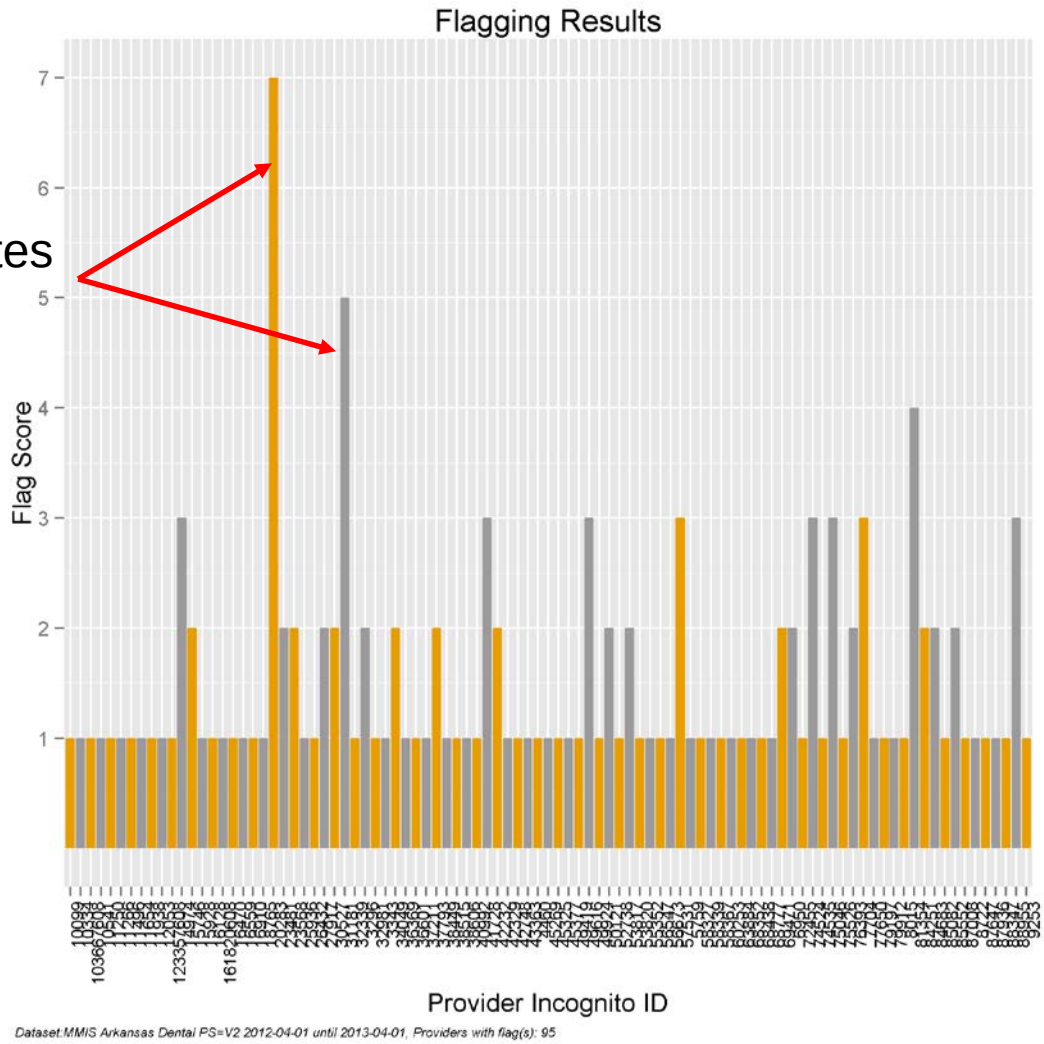
Medical Fraud detection: example of fingerprint

- Tooth code analysis



Flagging results

Possible candidates for investigation



Challenges

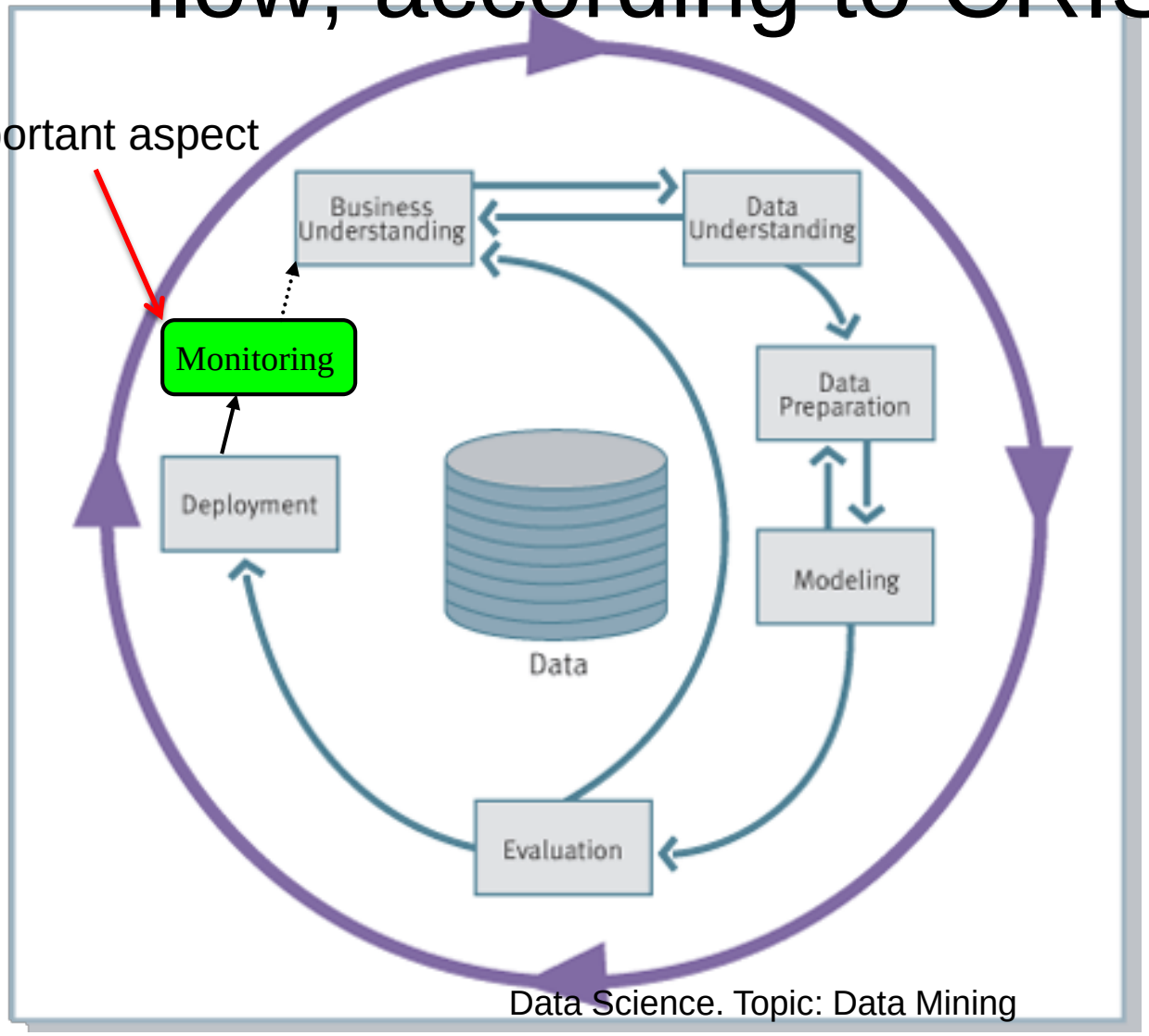
- Validation of approach; what is the precision and recall.
- Fraud Patterns will change over time.
- Migration to other health-care providers

Assumptions for Data Mining

- The Past is a Good Representation (Predictor) of the Future.
- The Data is Available.
- The Data Contains the Information (Patterns) one Wants to Extract.

One needs to test these assumptions on the data!

Knowledge Discovery Process flow, according to CRISP-DM



see www.crisp-dm.org for more information

Basic Data Mining Techniques

- **Classification:** classifying a new data instance
- **Regression:** Assigning a numerical value to a new data instance.
- **Associations:** e.g. A & B in the “basket” then also C in the “basket”
- **Clustering:** finding clusters in data
- **Visualization:** to facilitate human discovery
- **Summarization:** describing a group
- **Deviation Detection**: finding strange patterns
- **Prediction**: predicting a continuous value, also called regression.

What will you learn?

- Basic skills in data visualization using R.
- Basic knowledge about the data mining pipeline.
- Basic skills in setting up a simple DM experiment using R.
- The role of the training set and test set in the DM pipeline.
- Knowledge about K-fold cross validation

Topics which should be addressed

- The design of a valid DM pipeline (including feature construction and selection) from the given data set and information/research questions (Methodology)
- How to assess the performance of the constructed DM model in a sound way. (Methodology)
- Critical reflection: strength and weakness of the methodology and results.
  Placing the results in context.

IMPORTANT IS A SOUND METHODOLOGY!

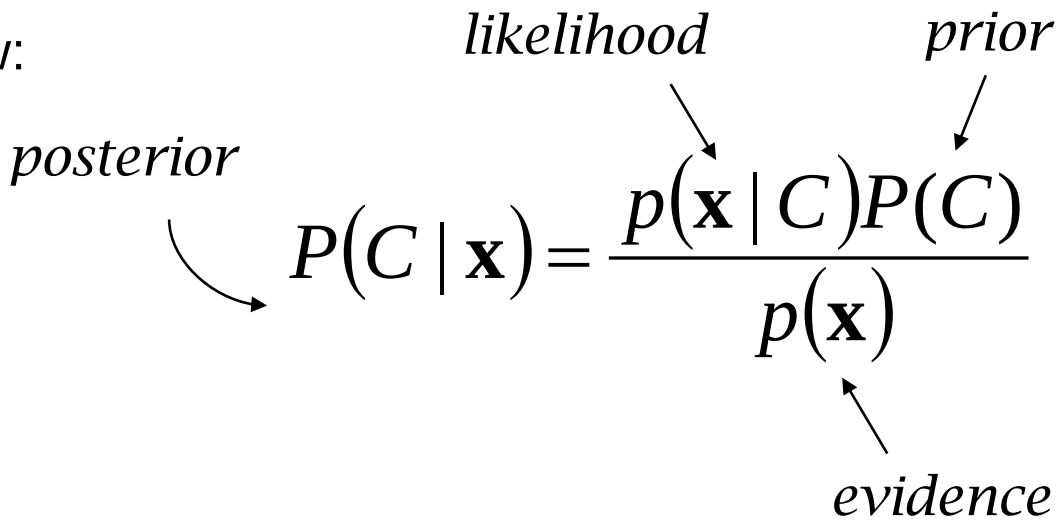
Your DM Project

- Information on the different projects can be found on Canvas.
- Read the course manual for the requirements for the final report.

PROBABILISTIC MODELS

Probabilistic Models

Bayes Law:



Classify new *x* as C_1 if $P(C_1|x) > P(C_2|x)$ else x is classified as C_2
this is equivalent to
Classify new *x* as C1 if $P(x|C_1)P(C_1) > P(x|C_2)P(C_2)$ else x is classified as C2

Problem: How to calculate $P(x|C_1)$ and $P(x|C_2)$?

Parametric Estimation for p(x|C)

- Write $p(x)$ instead of $p(x|C)$
- $X = \{ x^t \}_t$ where $x^t \sim p(x)$
- Parametric estimation:

  Assume a parametric form for p *i.e.* $p(x)=p(x \mid \theta)$ and estimate θ, its sufficient statistics, using $\mathcal{X}$

  e.g., $p=\mathcal{N}(\mu, \sigma^2)$ where $\theta = \{ \mu, \sigma^2 \}$: p is normally distributed with mean μ and variance σ^2

Multivariate Data

- Multiple measurements (sensors)
- *d* inputs/features/attributes: *d*-variate
- *N* instances/observations/examples

$$\mathbf{X} = \begin{bmatrix} X_1^1 & X_1^2 & \cdots & X_1^N \\ X_2^1 & X_2^2 & \cdots & X_2^N \\ \vdots & & & \\ X_d^1 & X_d^2 & \cdots & X_d^N \end{bmatrix}$$

Sometimes one uses another convention:
Each row is a data point!

Multivariate Parameters

$$\text{Mean}: E[\boldsymbol{x}] = \boldsymbol{\mu} = [\mu_1, ..., \mu_d]^T$$

$$\text{Covariance}: \sigma_{ij} \equiv \text{Cov}(X_i, X_j)$$

$$\text{Correlation}: \text{Corr}(X_i, X_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

$$\Sigma \equiv \text{Cov}(\boldsymbol{X}) = E[(\boldsymbol{X} - \mu)(\boldsymbol{X} - \mu)^T] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

MLE: Parameter Estimation

$$\text{Sample mean } \mathbf{m}: m_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, i = 1,..., d$$

$$\text{Covariance matrix } \mathbf{S}: s_{ij} = \frac{\sum_{t=1}^{N} \left(x_i^t - m_i\right)\left(x_j^t - m_j\right)}{N}$$

$$\text{Correlation matrix } \mathbf{R}: r_{ij} = \frac{s_{ij}}{s_i s_j}$$

This is what is called a biased estimator for the covariance:
If we divide by N-1 then we have a unbiased estimator. Hence most o the time we use 1/(N-1)

Estimator for covariance

$$\text{Covariance matrix } \mathbf{S}: s_{ij} = \frac{\sum_{t=1}^{N} \left(x_i^t - m_i\right)\left(x_j^t - m_j\right)}{N}$$

This is what is called a biased estimator for the covariance:
If we divide by N-1 then we have an unbiased estimator. Hence most of the time we use 1/(N-1)

$$\text{Covariance matrix } \mathbf{S}: s_{ij} = \frac{\sum_{t=1}^{N} \left(x_i^t - m_i\right)\left(x_j^t - m_j\right)}{N-1}$$

Independent Inputs: Naive Bayes

- If one assumes that features are independent given the class then

$$P(x_1, x_2, x_3, x_4|C) = P(x_1|C)P(x_2|C)P(x_3|C)P(x_4|C)$$

- In case of normal distribution:

$$p(\mathbf{x}|C) = \prod_{i=1}^{d} p_i(x_i|C) = \frac{1}{(2\pi)^{d/2}\prod_{i=1}^{d}\sigma_i} \exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]$$