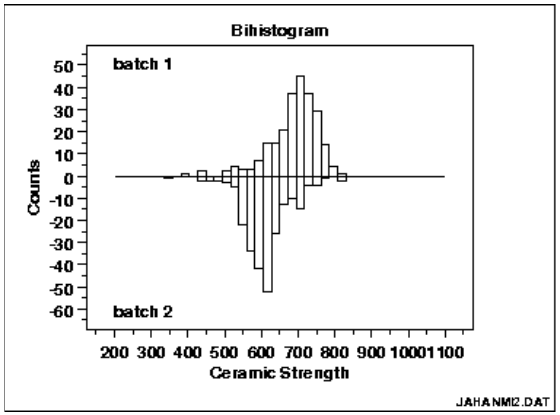
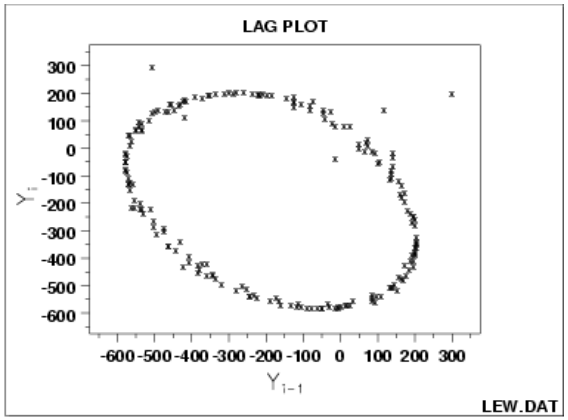
Exploratory Data Analysis





Lektor Dr.techn. Alexander K. Seewald
Österreichisches Forschungsinstitut
für Artificial Intelligence

alex@seewald.at / alex.seewald.at

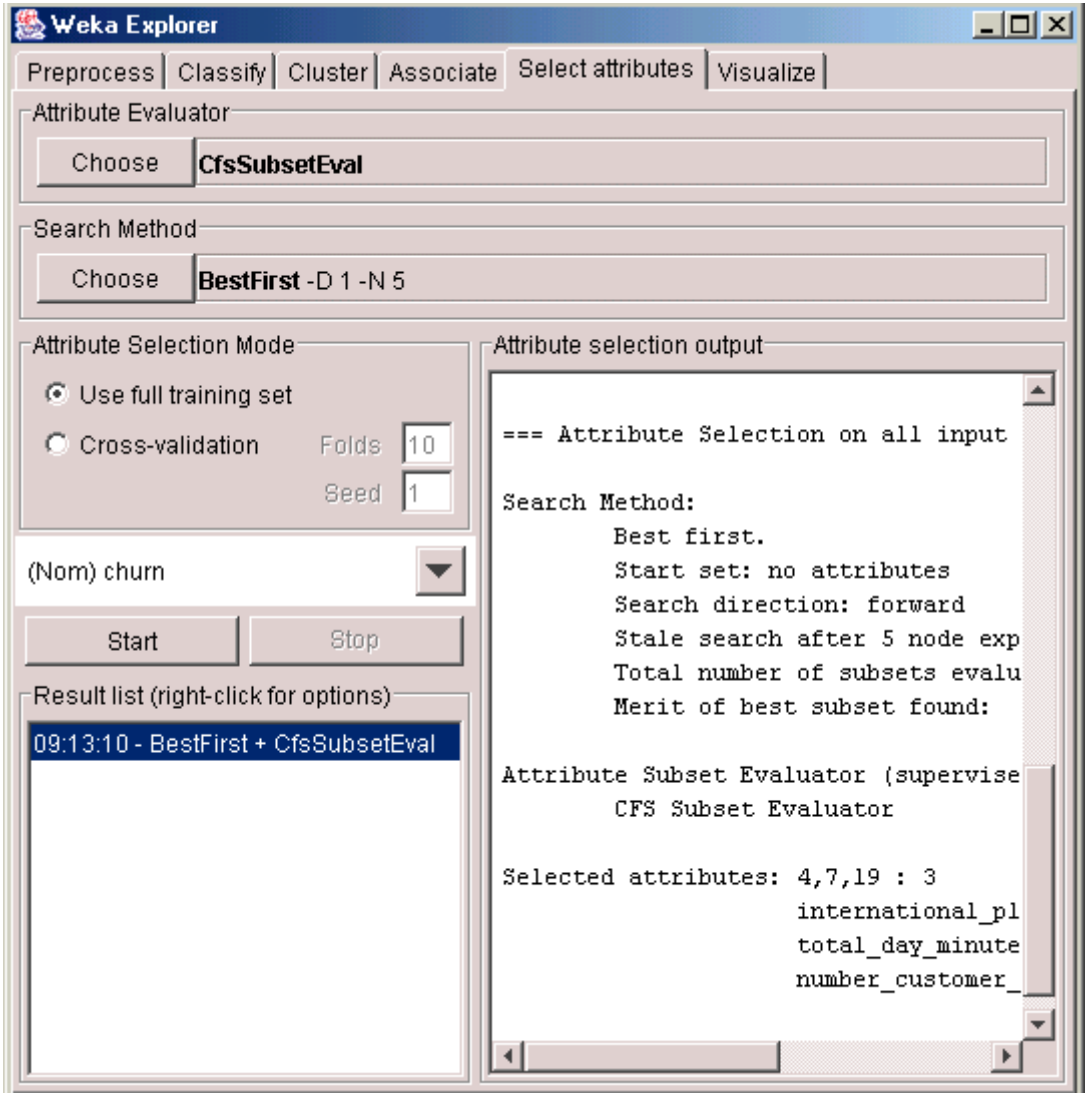
Feature Selection (from last week)

Feature Construction **may yield a large number of features.** ***Feature Selection*****, i.e. reducing the number of features, can improve classification accuracy as well as speeding up the learning process. It is also essential to achieve simpler, more comprehensible models.**

- Feature selection is well supported in WEKA under weka.attributeSelection (e.g. CfsSubsetEval, ChiSquaredAttributeEval and ReliefFAttributeEval).
- Feature construction can in simple cases be done via weka.filters.unsupervised.attribute.AddExpression
- More complex feature construction can be done in Java, or in external programs which output the ARFF file format.

alex@seewald.at / alex.seewald.at

Feature Selection in WEKA



AttributeEvaluator: determines the merit of a specific feature subset

Many can be used, including accuracy of specific classifiers (WrapperSubsetEval)

SearchMethod: How to search in the set of all feature subsets. For n features there are 2^n subsets - only very few can be considered, so search is essential.

Output: A set of most relevant attributes (4,7,19 in this case)

alex@seewald.at / alex.seewald.at

Issues in Feature Selection

Two reasons why attribute/feature is found relevant by FS

- It is really relevant to determine the true classification
- It correlates with the class purely due to chance.

Some common approaches to distinguish these two cases

- Use a CV instead of training data. If a feature is chosen in all folds, it is more likely to be relevant.
- Use multiple feature selection methods. A feature chosen by more than one method is more likely to be relevant.
- Common sense: Are the chosen features plausible?

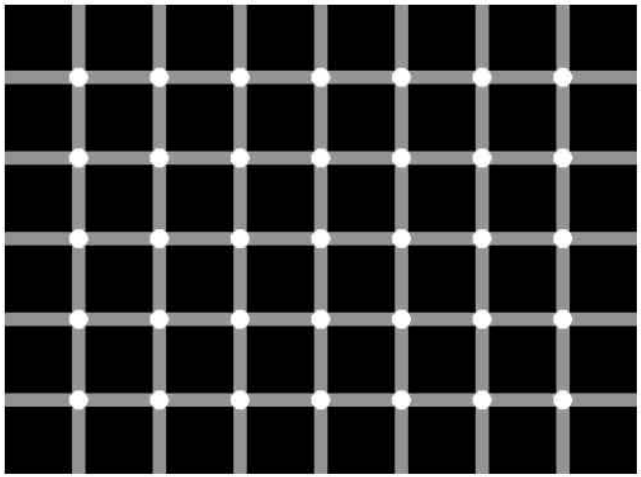
Feature Selection on full training data followed by CV should be avoided. FS feeds back information on attribute distributions and class correlations into the training set and increases potential for overfitting.

alex@seewald.at / alex.seewald.at

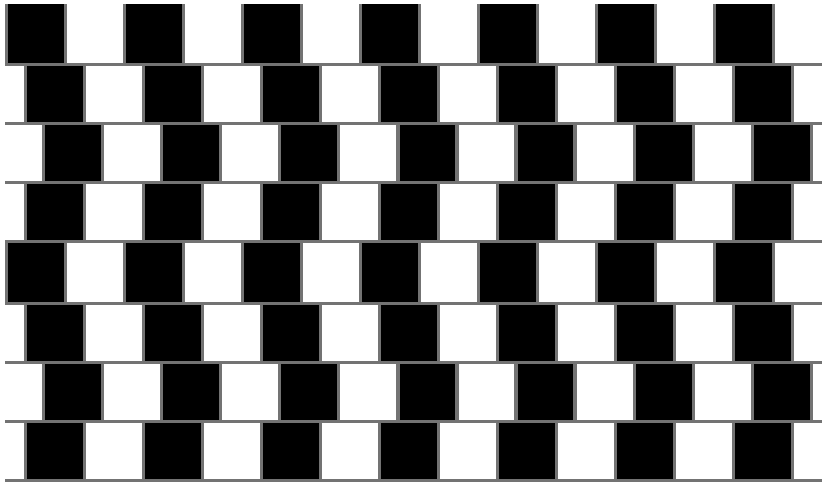
Exploratory Data Analysis

- Part of Data Analysis that is conceptually most similar to Artificial Intelligent Data Analysis.
- Instead of algorithms, EDA focusses on utilizing human visual sense for analysis as well as statistical methods.
- Complementary to AI DA. Not without issues, see below.

Count the black dots

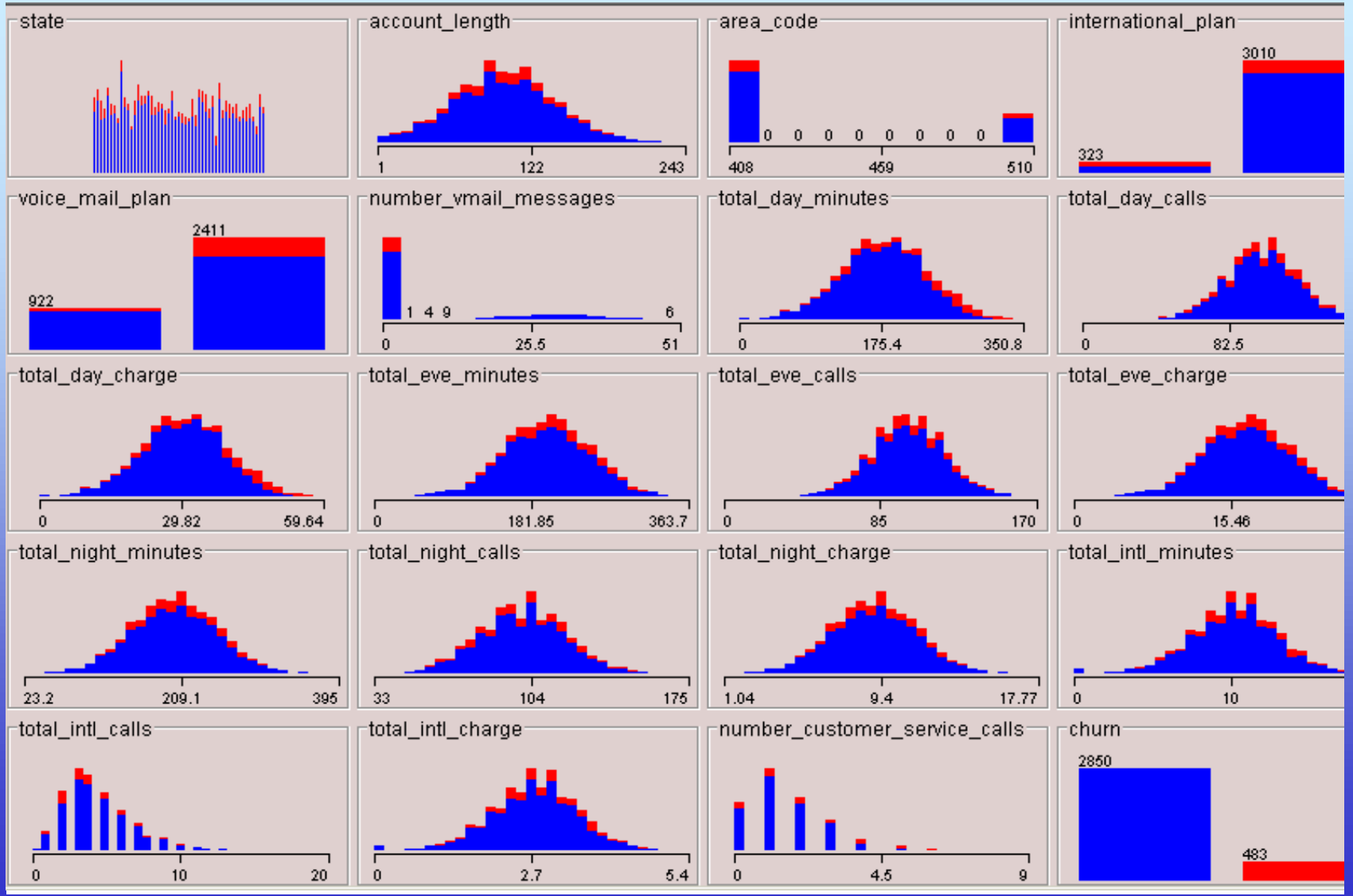


Straight horizontal lines



alex@seewald.at / alex.seewald.at

Histograms (WEKA: Visualize All)

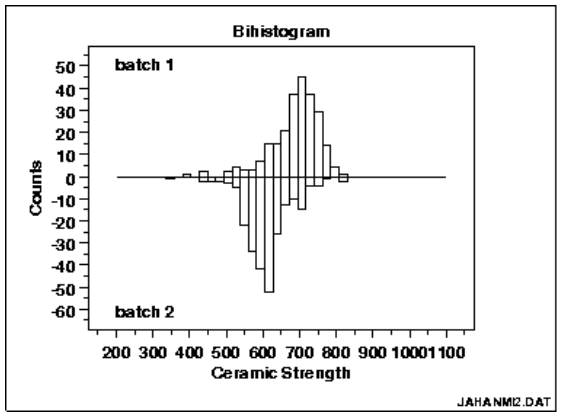


alex@seewald.at / alex.seewald.at

Histograms (2)

A **Histogram** shows the number of examples for which an attribute has a specific nominal value, or a value within a numeric interval. Numeric attributes need to be discretized before a histogram can be computed. In some cases, the class is also shown (as before). If there were a single attribute which correlates well with the class, it would be immediately apparent ⇒ graphical feature selection.

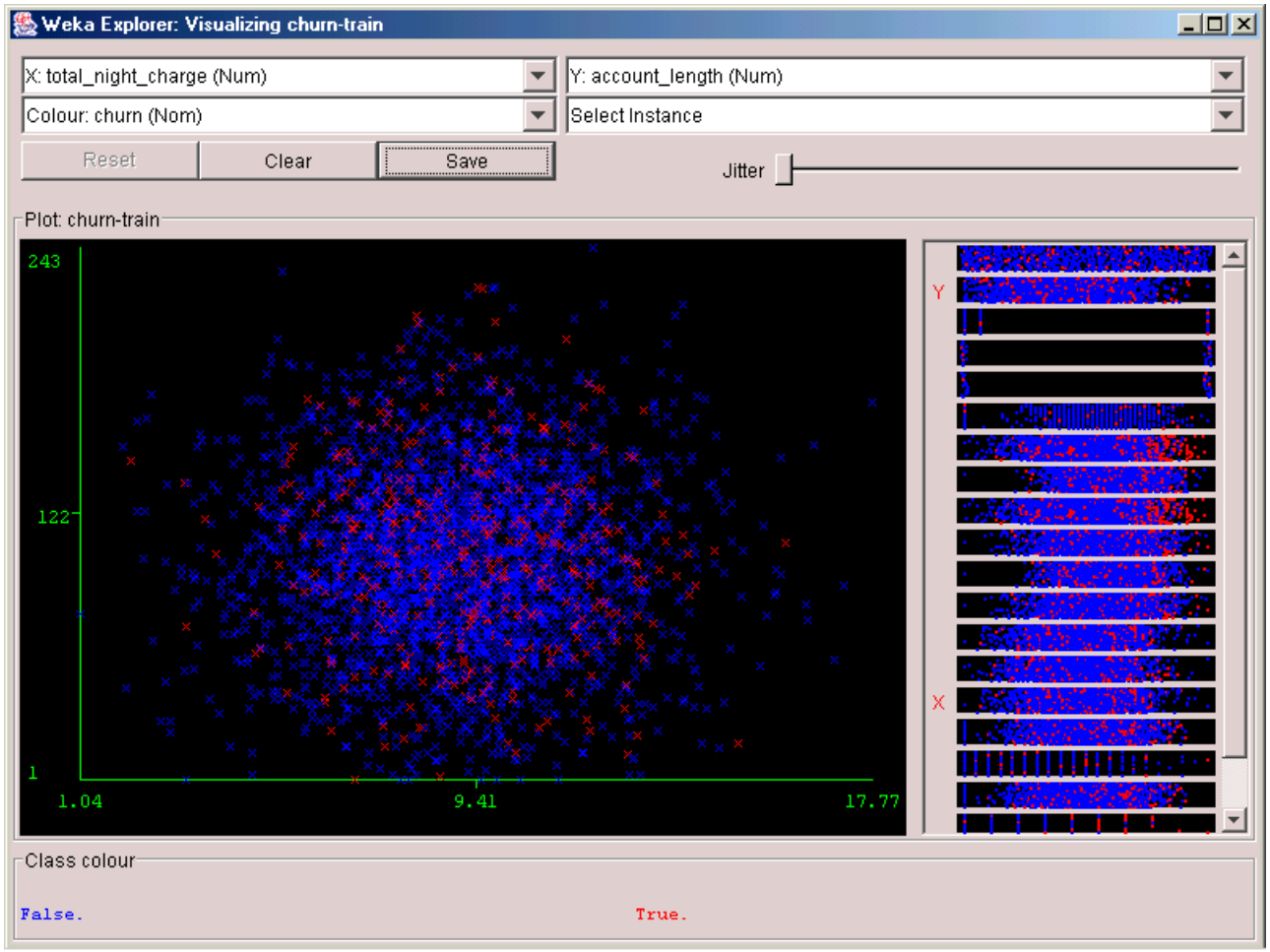
- **Bi-Histogram:** two histograms along the same axis, above and below; for comparing difference sets of examples (e.g. from different production batches, different locations etc.)



alex@seewald.at / alex.seewald.at

Scatterplots

Plot one feature (as X) vs. another feature (as Y)



No correlation between X and Y in this case (spherical appearance)

Classes are shown in **red** (churn=true) and **blue**.

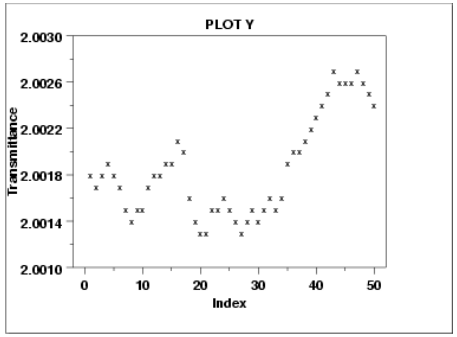
Also shows distribution of all attribute's values on the right.

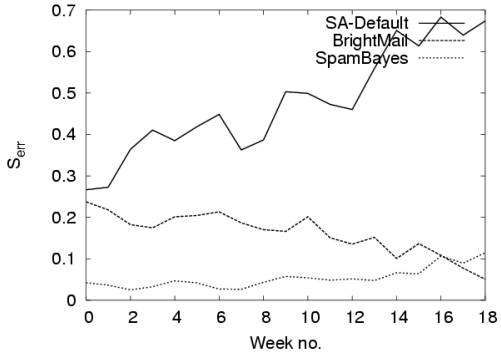
alex@seewald.at / alex.seewald.at

Scatterplots (2)

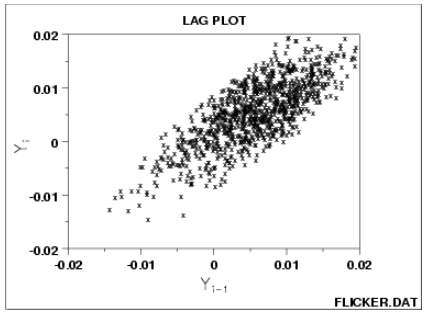
Many variants...

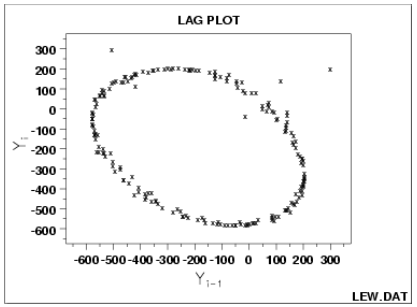
- **Run Sequence Plot:** X = example no. / time, Y = response variable; used to check for time dependencies and order effects.

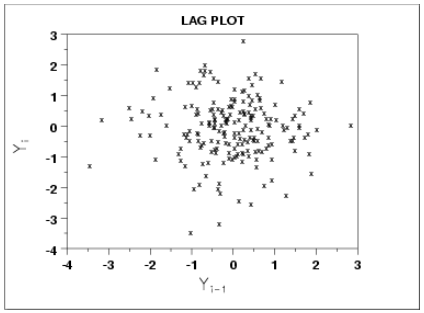




- **Lag plot:** Y = response variables Y_i for all i; X = Y_{i-LAG} for all i with arbitrary LAG (usually 1); for autocorrelation



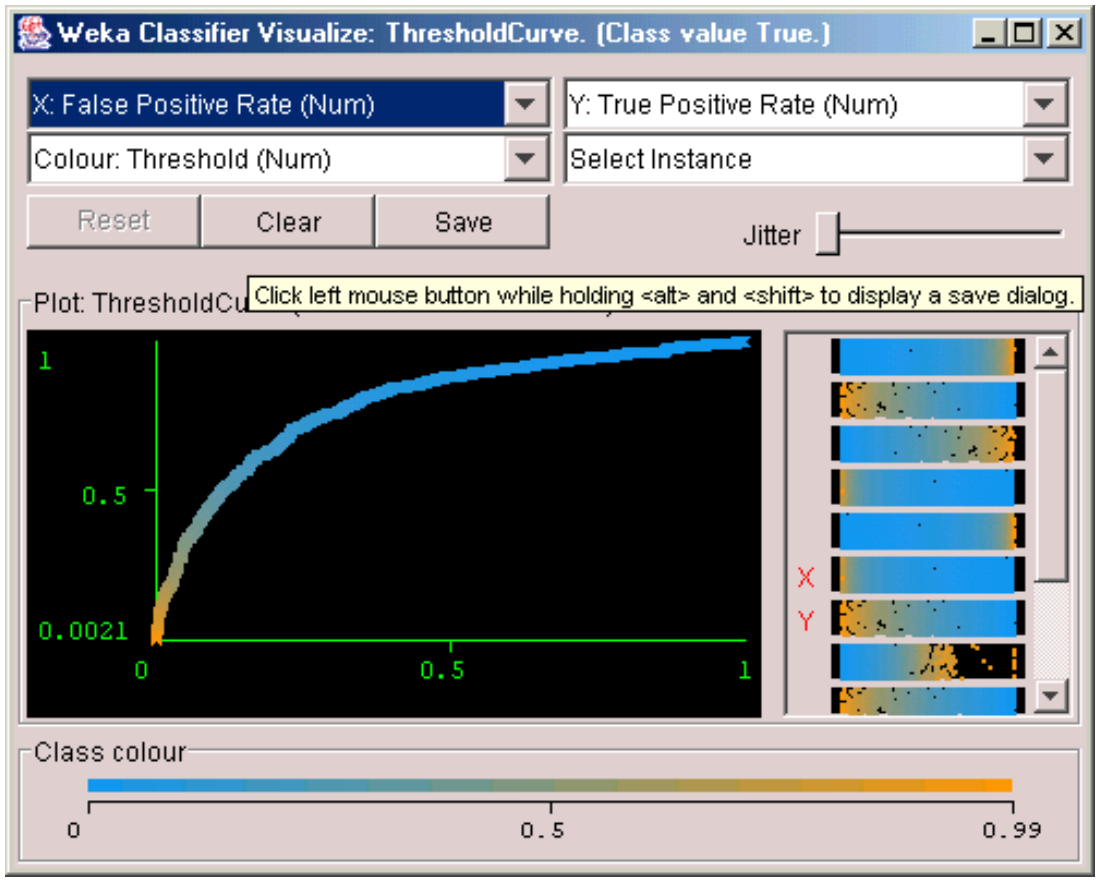




alex@seewald.at / alex.seewald.at

Scatterplots (3)

- **ROC curve:** X=False Positive (FP) rate, Y=True Positive (TP) rate; shows classifier performance in more detail



Available in WEKA by right-clicking on model in result list after training, and choosing *View Threshold Curve* for a specific class.

If the ROC curve of learning algorithm A is always above the ROC curve of learning algorithm B, it is unambiguously better. This seldom happens in practice.

alex@seewald.at / alex.seewald.at

Scatterplots (4)

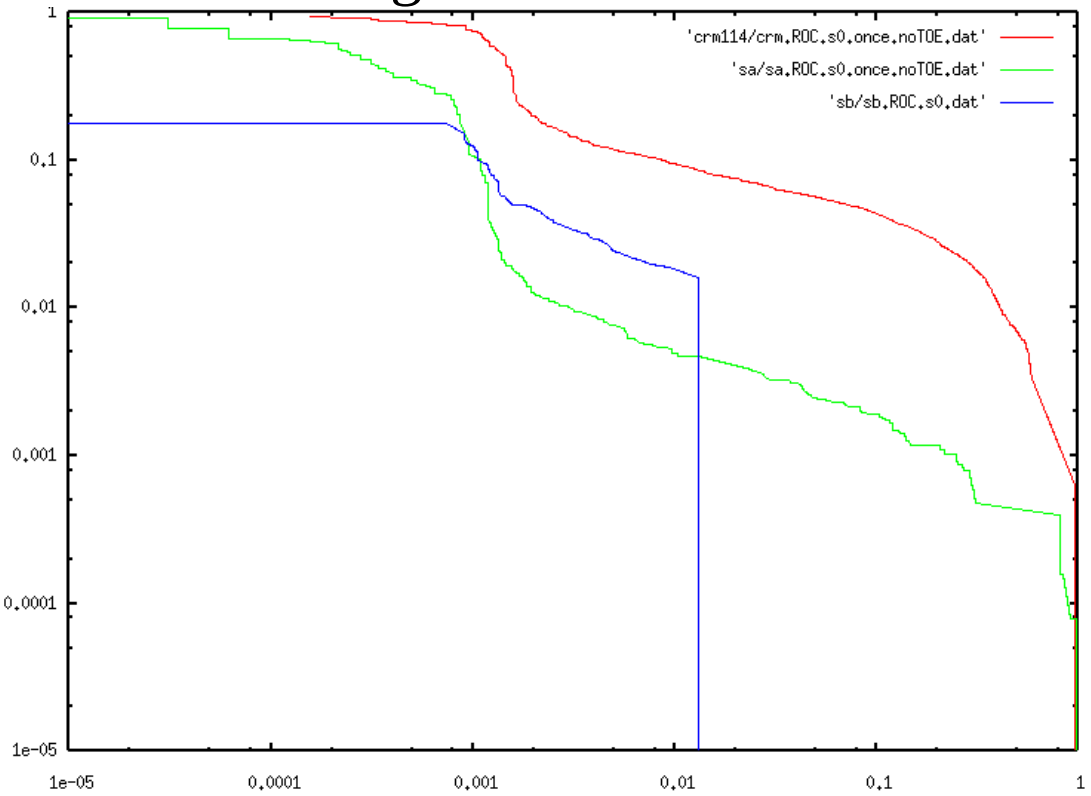
- **Ham / Spam Error Curve:** X = ham error, Y = spam error; for comparing spam filtering systems. E.g. at an acceptable ham error of 0.1%, the spam error of current systems is around 5%. Notice logarithmic scale on X & Y!

<u>Ham error</u> (X):

Probability that a good ham mail will get lost (i.e. classified as spam)

<u>Spam error</u> (Y):

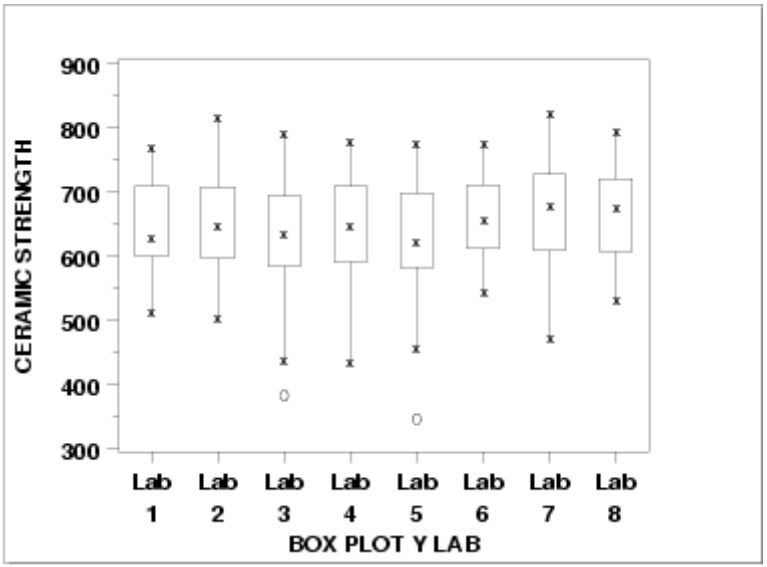
Probability that a bad spam mail will get through (i.e. classified as ham)

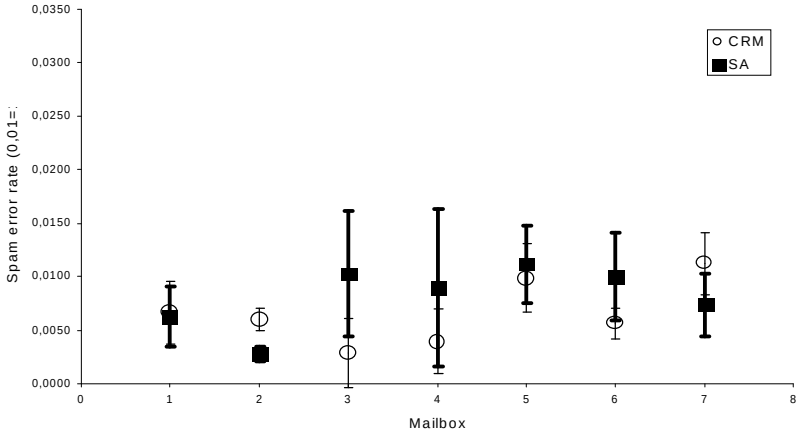


alex@seewald.at / alex.seewald.at

Boxplots

- Boxplots show mean, standard deviation, position of 1st and 3rd quartile, and outliers (see below, left)
- A simplified version which only shows mean and standard deviation is also often used (*error bars*; see below, right). No overlap between error bars indicates a significant difference at around 95% significance level (α=0.05)



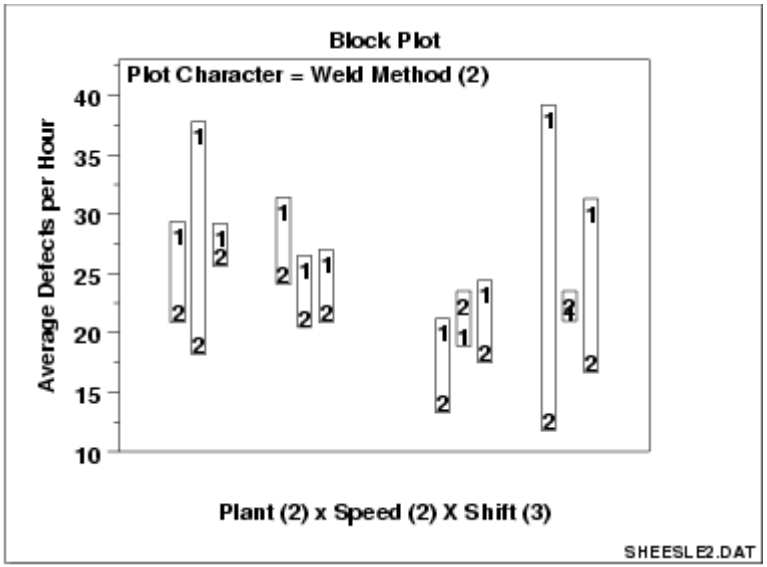


alex@seewald.at / alex.seewald.at

Blockplots

A graphical way for analysis-of-variance (ANOVA)

- Y = response variable
- X = all combinations of *nuisance* variables (i.e. all features that are supposedly not relevant vs. the response variable)
- Plot character: levels of the primary factor (i.e. the one feature whose effect we want to analyse; **1** and **2** below)

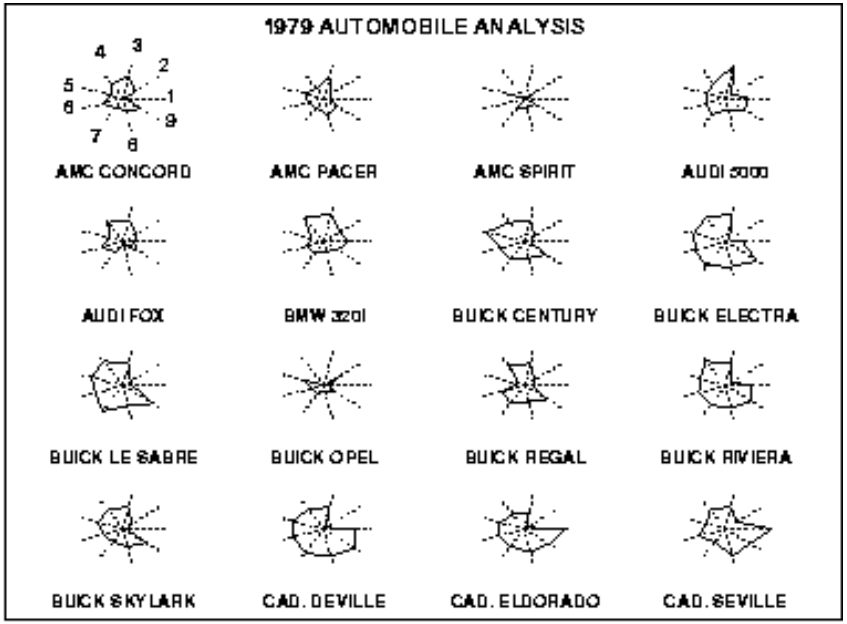


alex@seewald.at / alex.seewald.at

Starplot

Shows an arbitrary number of attributes at one glance

- Each attribute is assigned to a line in a specific direction.
- The value of the attribute defines a point on this line.
- Connecting all lines shows a single example with all values
- Related to glyph-based visualization. Only useful for small datasets.



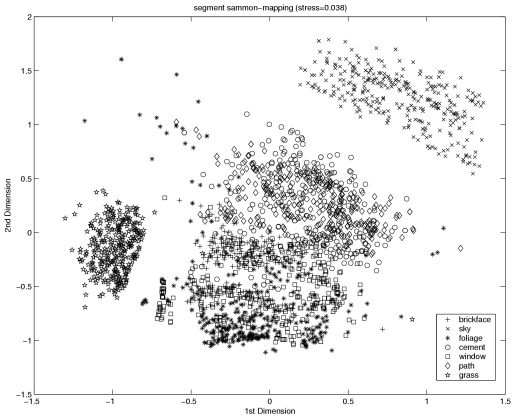
alex@seewald.at / alex.seewald.at

Dimensionality-reducing techniques

Very hard to recognize patterns in greater than 3D. Dimensionality-reducing techniques reduce the number of dimensions while preserving aspects of the data.

- **Principal Components Analysis (PCA):** computes smaller set of features that accounts for most of the variance in the data. Each new feature is a linear combination of old features.

- **Sammon Mapping:** arbitrary (non-linear) mapping which tries to preserve euclidean distances between datapoints.



alex@seewald.at / alex.seewald.at

Further Reading & Software

- Engineering Statistics Handbook (EDA methods & ref.)
  http://www.itl.nist.gov/div898/handbook/eda/eda.htm
- Dataplot Software
  http://www.itl.nist.gov/div898/software/dataplot/
- The R Project for Statistical Computing
  http://www.r-project.org/
- GnuPlot plotting software
  http://www.gnuplot.info/

alex@seewald.at / alex.seewald.at