# Basic Understanding of Quantitative Modelling

Vadim Strijov

Computing Center of the Russian Academy of Sciences

September 12th, 2012
RWTH Aachen, School of Business and Economics

## Quantitative modelling

with examples in Marketing/Financial/Environmental Engineering

The main goal is to show
**how a quantitative model could be recognized**
among daily routines.

## Quantitative modelling, definition

*The model is a mathematical representation of our knowledge about some investigated phenomenon.*

**The quantitative model**

is based mainly on measured data and may concern our knowledge about processes underlying the phenomenon.

VERSUS

**The mathematical model**

is based on our knowledge about processes underlying the phenomenon and may concern measured data.

## Notation

**The model is the parametric family of functions**

The model

$$f : \mathbf{x} \mapsto \hat{y}$$

maps the object description to its corresponding class label.

**Classification, Recognition, Regression, Decision making** are types of modelling problems.

## Classification, RecognitioN, Regression, Decision making

The model

$$f : \mathbf{x} \mapsto \hat{y}.$$

1. $\mathbf{x}$ – patient $\mapsto$ y – treatment result
2. $\mathbf{x}$ – bank client $\mapsto$ y – grant/reject a credit
3. $\mathbf{x}$ – stock share price $\mapsto$ y – buy/sell
4. $\mathbf{x}$ – telecom subscriber $\mapsto$ y – will go to another provider
5. $\mathbf{x}$ – photograph $\mapsto$ y – identity of the person
6. $\mathbf{x}$ – protein fragment $\mapsto$ y – type of secondary structure
7. $\mathbf{x}$ – text message $\mapsto$ y – is spam or not
8. $\mathbf{x}$ – chemical combination structure $\mapsto$ y – its property
9. $\mathbf{x}$ – technological process parameters $\mapsto$ y – product quality
10. $\mathbf{x}$ – history of sales $\mapsto$ y – customer demand forecast
11. $\mathbf{x}$ – description of apartment $\mapsto$ y – price to sell
12. $\mathbf{x}$ – pair (client, commodity) $\mapsto$ y – rating of commodity

{C,N,R,D}.

## Classification, Recognition, Regression, Decision making

The model

$$f : \mathbf{x} \mapsto \hat{y}.$$

1. [C] $\mathbf{x}$ – patient $\mapsto$ y – healthy or not
2. [C] $\mathbf{x}$ – bank client $\mapsto$ y – grant/reject a credit
3. [C] $\mathbf{x}$ – stock share price $\mapsto$ y – buy/sell
4. [C] $\mathbf{x}$ – telecom subscriber $\mapsto$ y – will go to another provider
5. [N] $\mathbf{x}$ – photograph $\mapsto$ y – identity of the person
6. [C] $\mathbf{x}$ – protein fragment $\mapsto$ y – type of secondary structure
7. [C] $\mathbf{x}$ – text message $\mapsto$ y – is spam or not
8. [R] $\mathbf{x}$ – chemical combination structure $\mapsto$ y – its property
9. [R] $\mathbf{x}$ – technological process parameters $\mapsto$ y–product quality
10. [R] $\mathbf{x}$ – history of sales $\mapsto$ y – customer demand forecast
11. [R] $\mathbf{x}$ – description of apartment $\mapsto$ y – price to sell
12. [D] $\mathbf{x}$ – pair (client, commodity) $\mapsto$ y – rating of commodity

$$\{C,N,R,D\}.$$

## Learning of the model

The problem above was [R], the recognition problem, where
**x** is a text string (*in fact, your knowledge about it*),
$y$ is a label in the set {C,N,R,D},
the model is

$$f : \mathbf{x} \mapsto \hat{y}$$

and the error (or loss) function is

$$S = \frac{1}{12} \sum_{i=1}^{12} [y_i \neq \hat{y}_i],$$

where $[\cdot]$ means

$$[y_i \neq \hat{y}] = \begin{cases} 0, & \text{if } y = \hat{y}; \\ 1, & \text{if } y \neq \hat{y}. \end{cases}$$

## Historical data

There given the sample set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, *D* stands for "Data"

### the object description (or object) $\mathbf{x}$

- is a vector, $\mathbf{x}_i = [x_{i1}, \ldots, x_{ij}, \ldots, x_{in}]$ of $n$ components, which are called **features**;
- or more complex structure;

### the label $y$

is a scalar and could be of

- binary set, $y \in \{0, 1\}$;
- countable finite set, $y \in \{1, \ldots, z\}$;
- set of real numbers, $y \in \mathbb{R}$;
- etc.

The pair $(\mathbf{x}, y)$ is called **a precedent** or a historical sample.

# Data scales

| Scale | Mathematical operations |
|---|---|
| Nominal | No linear operations allowed |
| Ordinal | Only comparison allowed |
| Linear (Real) | Linear operations: "$+$", "$\times$" |
| Interval | Linear operations with restrictions |
| Binary | Linear or Boolean operations (and, or, not) |

## Nominal, Ordinal, Linear, Interval, Binary, etc. (Unspecified)

1. Education degrees
2. Wind force
3. Moody's bank ratings
4. BBC News titles
5. Google search results
6. Protein amino acids
7. Spectrum colors
8. Time in physics
9. Facebook connections
10. Diesel engine combustion pressure
11. The distance to home
12. Traffic light's lights

{N,O,L,I,B,U}.

## Nominal, Ordinal, Linear, Interval, Binary, etc. (Unspecified)

1. [O] Education degrees
2. [I] Wind force
3. [O] Moody's bank ratings
4. [N] BBC News titles
5. [O] Google search results
6. [N] Protein amino acids
7. [I] Spectrum colors
8. [L] Time in physics
9. [B] Facebook connections
10. [I] Diesel engine combustion pressure
11. [L] The distance to home
12. [U] Traffic light's lights

{N,O,L,I,B,U}.

# Data type conversion: binarization

## Applicant's industry, nominal scale

| Nominal | Tourism | Banking | Telecom |
|---|---|---|---|
| John | 1 | 0 | 0 |
| Thomas | 0 | 1 | 0 |
| Sara | 0 | 0 | 1 |

## Applicant's education, ordinal scale

| Ordinal | Primary | Secondary | Higher |
|---|---|---|---|
| John | 1 | 0 | 0 |
| Thomas | 1 | 1 | 0 |
| Sara | 1 | 1 | 1 |

# Parametric model

Introduce model parameters, the vector $\mathbf{w}$ and call $f$ the parametric model:
The element-wise mapping

$$f : (\mathbf{w}, \mathbf{x}) \mapsto \hat{y},$$

the vector mapping

$$\mathbf{f} : (\mathbf{w}, X) \mapsto \hat{\mathbf{y}},$$

or

$$\mathbf{f} : \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_m \end{bmatrix}, \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \ddots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \mapsto \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_m \end{bmatrix}.$$

## Classification of patients with Cardio-Vascular Disease

There given two groups of patients: $y \in \{A1, A3\}$; each patient is described by a set of markers $\mathbf{x}$.

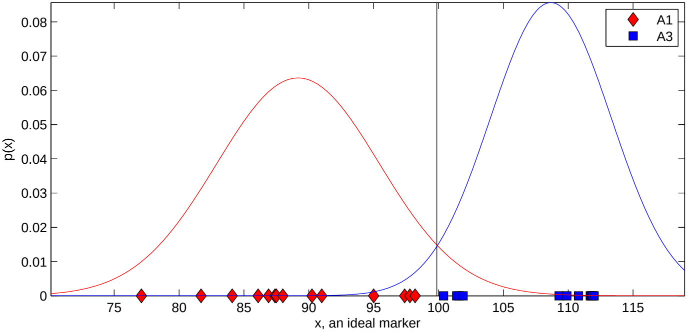
| | | | |
|---|---|---|---|
| Classes | $\longrightarrow$ | Groups of patients | The patients have classification labels "A1" and"A3". |
| Objects | $\longrightarrow$ | Patients | We have measured data for 14 patients in the group "A1" and 17 patients in the group "A3". |
| Features | $\longrightarrow$ | Markers | We have 20 markers: K, L, K/M, L/M, K/N, K/O, L/O, K/P, L/P, K/Q, K/R, L/R, L/R/SA, L/T/SA, L/T/SO, U/V, U/W, U/X, U/Y, U/Z |

# Object–Feature (Patient–Marker) table, an extract

| Class | Patient name | K | L | K/M | L/M | |
|---|---|---|---|---|---|---|
| A1 | C001 | 58.3 | 16.7 | 0.52 | 0.00 | |
| A1 | C004 | 40.2 | 6.0 | NaN | NaN | |
| A1 | C005 | 54.3 | 13.1 | NaN | NaN | |
| A1 | C008 | 48.7 | 9.8 | 0.05 | 0.02 | etc. |
| A3 | 023 | 46.6 | 21.2 | 0.40 | 0.08 | |
| A3 | 026 | 50.7 | 26.2 | 0.12 | 0.00 | |
| A3 | 027 | 45.3 | 24.5 | 0.05 | 0.02 | |
| A3 | D037 | 46.3 | 13.1 | 1.23 | 0.13 | |
| | | | | etc. | | |

Can we show that the groups are significantly different?

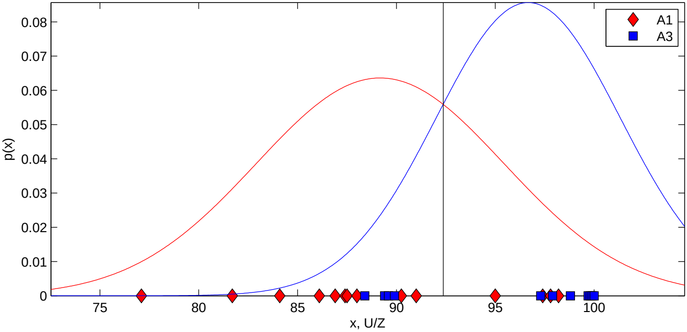
# One-dimensional analysis, ideal example




**Separate two groups using statistical hypothesis;**

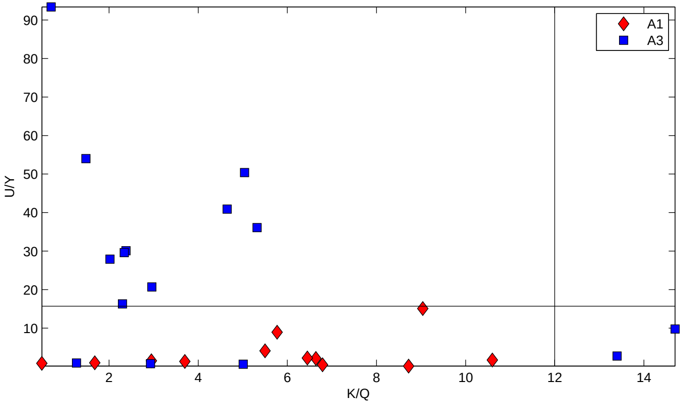
try the null-hypothesis in one of the following tests: Student's t-test, Welch's t-test or Mann-Whitney's U-test.

# One-dimensional analysis, real data



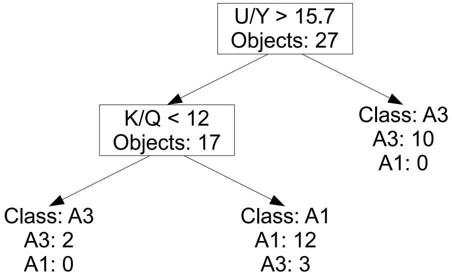

- ✓ It is very simple to visualize one-dimensional data.
- ✓ One-dimensional statistics is well-developed and recognized.
- × And give poor results if one deals with a complex problem.

# Classification rules and decision trees




if U/Y > 15.7 then A3 else ( if K/Q < 12 then A1 else A3 )

## Decision tree




if U/Y $>$ 15.7 then A3 else ( if K/Q $<$ 12 then A1 else A3 )

- ✓ Different subsets of markers produce trees of different quality.
- ✓ One can use several trees to make a voting algorithm.

# Decision forest and voting algorithms

1. If U/Y $<$ 15.7 then A1 else A3
2. If U/Z $<$ 88.2 then A1 else (if U/V $<$ 51.9 then A1 else A3)
3. If U/Z $<$ 88.2 then A1 else (if K/N $<$ 31.9 then A3 else A1)

| Class | Patient | Rule 1 | Rule 2 | Rule 3 | Vote |
|---|---|---|---|---|---|
| A1 | C014 | A1 | ~~A3~~ | A1 | A1 |
| | C015 | NaN | A1 | A1 | A1 |
| | D034 | A1 | A1 | ~~A3~~ | A1 |
| | L107 | A1 | NaN | ~~A3~~ | NaN |
| | etc. | ... | ... | ... | ... |
| A3 | 023 | A3 | A3 | A3 | A3 |
| | 026 | ~~A1~~ | A3 | A3 | A3 |
| | 027 | A3 | NaN | NaN | NaN |
| | 009 | ~~A1~~ | A3 | A3 | A3 |
| | etc. | ... | ... | ... | ... |

## Linear classifier

The equation

$$\mathbf{w}^\mathsf{T}\mathbf{x} = b$$

describes a separating hyperplane in the feature space.
Let $\mathbf{x}_i$ be the patient's markers and $\mathbf{w}$ be the parameters. Then

$$\hat{y}_i = f(\mathbf{w}, \mathbf{x}_i) = \text{sign}\left(\sum_j w_j x_{ij} - b\right) = \text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x}_i - b)$$

is the class of the $i$-th patient.

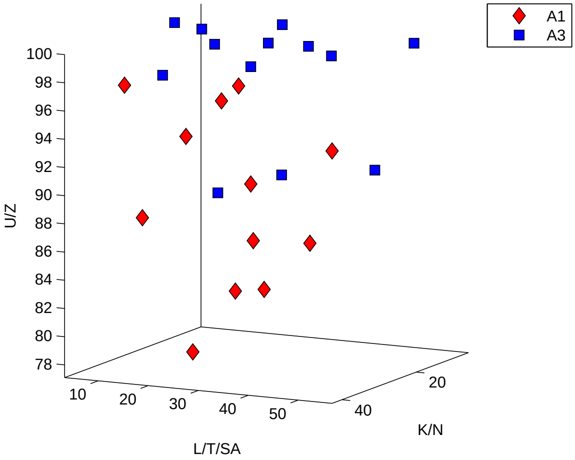
### Decision tree

✓ assumes the markers do not depend on each other.

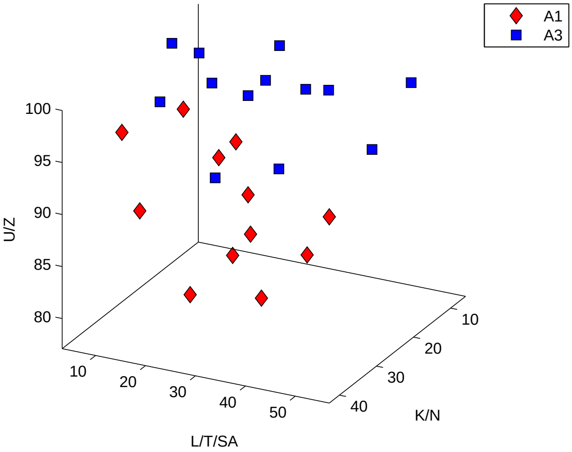
### Linear classifier

✓ assumes the markers depend on each other.

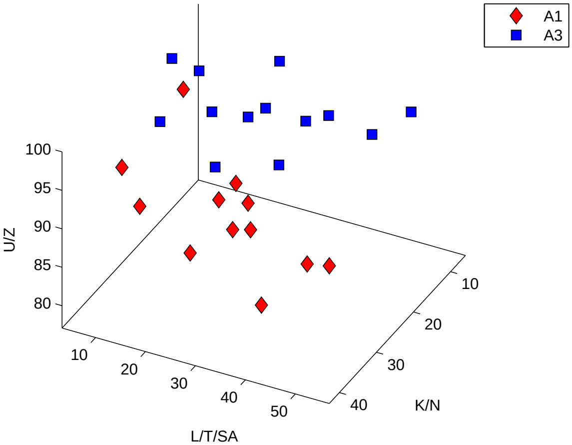
# Linear classifier



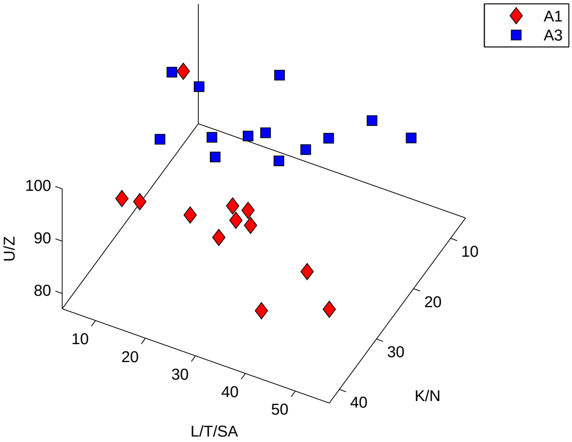
# Linear classifier



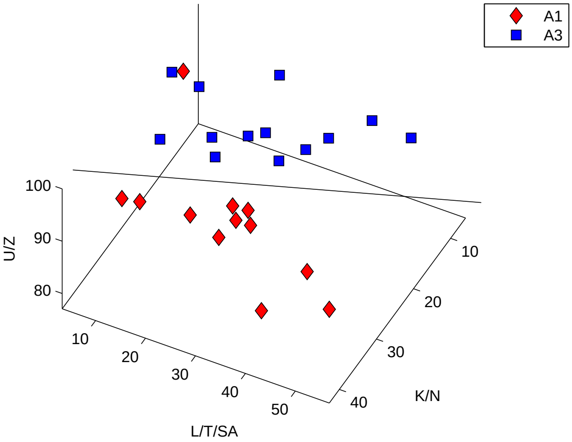
# Linear classifier



# Linear classifier



# Linear classifier



## Classification results

After classification a pair (A1 vs. A3) we obtain the following information:

1. classified patients

$$(\text{C014}, \text{D034}, \text{L107}, \ldots, \sout{\text{C008}}, 023, 026, \ldots, \text{C015}),$$

2. classification error

$$\frac{|a1|}{|A1|} + \frac{|a3|}{|A3|},$$

3. most important markers

$$(\text{L}, \text{K/N}, \text{U/Z}),$$

4. parameters of the algorithm

$$\hat{y}_i = \text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x}_i - b) = \text{sign}([0.35, 0.72, 0.29]^\mathsf{T}\mathbf{x}_i - 34.16).$$

## Database from R. A. Fisher: Iris data, 1936

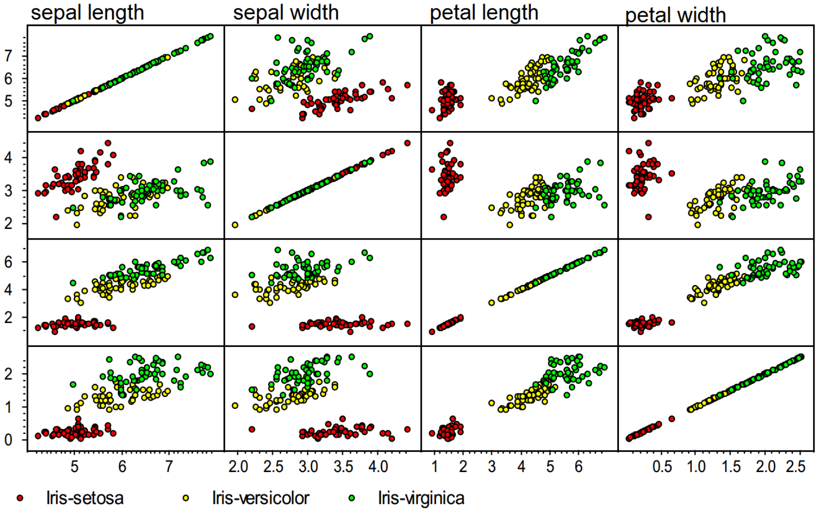
Object description, $\mathbf{x}$:

- sepal length in cm,
- sepal width in cm,
- petal length in cm,
- petal width in cm.

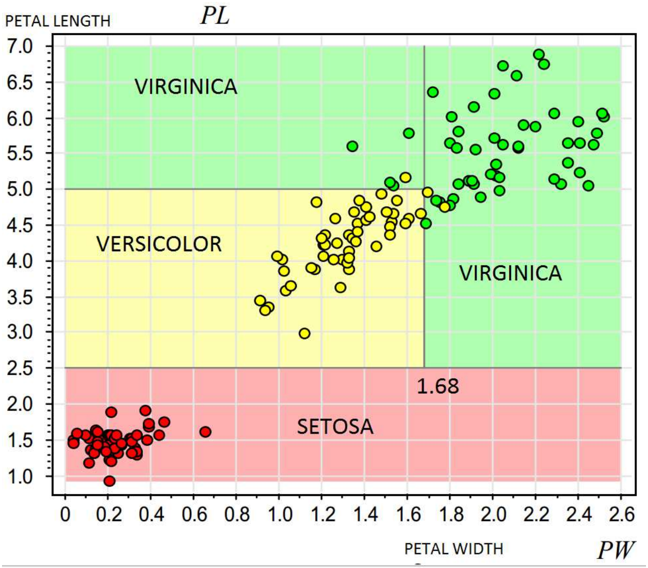
Class label, $y$:

- Iris Setosa,
- Iris Versicolour,
- Iris Virginica.

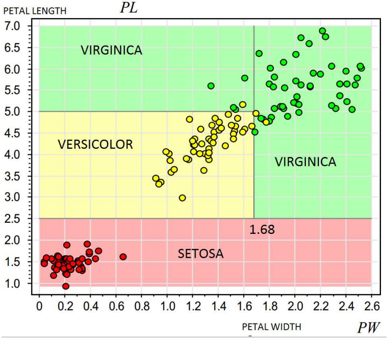
# Scatterplot for Iris data

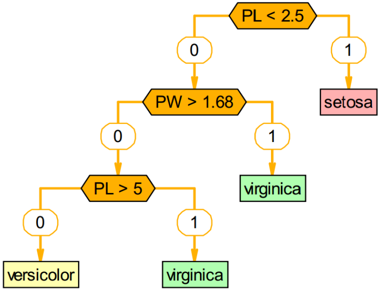


## Classify the Iris



# Classify the Iris




| | |
|---|---|
| setosa | $r_1(x) = [PL \leqslant 2.5]$ |
| virginica | $r_2(x) = [PL > 2.5] \wedge [PW > 1.68]$ |
| virginica | $r_3(x) = [PL > 5] \wedge [PW \leqslant 1.68]$ |
| versicolor | $r_4(x) = [PL > 2.5] \wedge [PL \leqslant 5] \wedge [PW < 1.68]$ |

## The workflow of the bank credit scoring

Client's application & history

↓

Client's score: probability of fraud / default

↓

Accept (refuse) the application

↓

Make the agreement, start client's history

### Type of detection

Fraud: deliquency 90+ on $3^{rd}$

$$0 \longrightarrow \underbrace{\overbrace{30+ \longrightarrow 60+ \longrightarrow 90}+ \longrightarrow 120+} \longrightarrow 150+$$

Default: deliquency 90+ on any, but $1^{st}$

## Statistics od banking data

- Loans of 90+ delinquency, default cases, applications
- The fraud cases are rejected
- Overall number of cases
  - $\sim 10^4$ for long-term credits
  - $\sim 10^6$ point-of-sale credits
  - $\sim 10^7$ for churn analysis
- Default rate $\sim$ 8–16%
- Period of observing: no less 91 days after approval
- Number of source variables $\sim$ 30–50
- Number records with missing data $> 0$, usually very small
- Number of cases with outliers $> 0$, $3\sigma^2$-cutoff

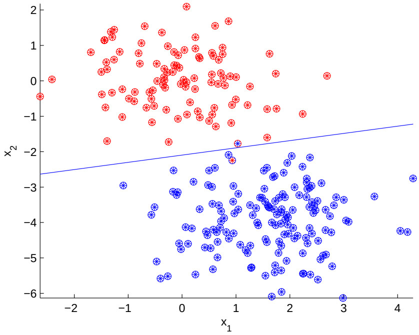
# List of features (fields in questionary)

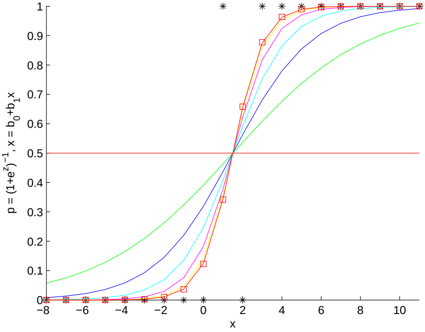
| Variable | Type | Categories |
|---|---|---|
| Loan currency | Nominal | 3 |
| Applied amount | Linear | |
| Monthly payment | Linear | |
| Term of contract | Linear | |
| Region of the office | Nominal | 7 |
| Day of week of scoring | Linear | |
| Hour of scoring | Linear | |
| Age | Linear | |
| Gender | Nominal | 2 |
| Marital status | Nominal | 4 |
| Education | Ordinal | 5 |
| Number of children | Linear | |
| Industrial sector | Nominal | 27 |
| Salary | Linear | |
| ... | ... | ... |

# Scoring problem statement

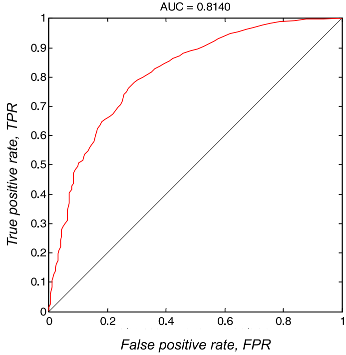
1. The data set: $\mathbf{x} \in \mathbb{R}^n$, $y \in \{0, 1\}$,
$$D = \{(\mathbf{x}_i, y_i)\}, \quad i \in \{1, \dots, m\};$$
2. the design matrix $X \in \mathbb{R}^{m \times n}$,
$$X = [\mathbf{x}_1^\mathsf{T}, \dots, \mathbf{x}_m^\mathsf{T}]^\mathsf{T};$$
3. class labels $\mathbf{y} \sim$ Bernoulli;
$$\mathbf{y} = [y_1, \dots, y_m]^\mathsf{T},$$
4. the model
$$\mathbf{f}(\mathbf{w}, X) = \frac{1}{1 + \exp(-X\mathbf{w})}.$$

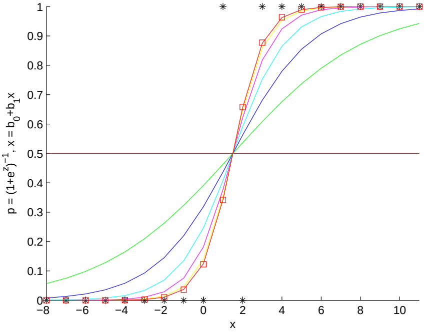
# Separating surface



# ROC-curve (receiver operating characteristic) as quality criterion




| $\hat{y} \backslash y$ | $P$ | $N$ |
|---|---|---|
| $\hat{P}$ | $TP$ | $FP$ |
| $\hat{N}$ | $FN$ | $TN$ |

$TPR = TP/P$ True Positive Rate

$FPR = FP/N$ False Positive Rate

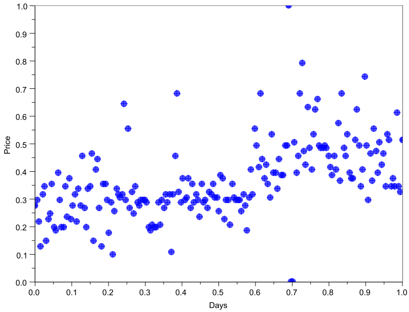
## Linear regression

The data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m = (X, \mathbf{y})$,
where $\mathbf{y} = [y_1, \ldots, y_m]^T$ is the target vector and $X$ is the design matrix

$$X = \begin{bmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^m \end{bmatrix} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \, .$$

# White bread price forecasting

White bread prices: $m = 195$ pairs $(x_i, y_i)$, the data are mapped to the segment $(0, 1)$.




Approximate the prices using the linear model

$$y_i = f(\mathbf{w}, x_i) + \varepsilon_i = w_2 x_i + w_1, \quad \mathbf{w} = [w_1, w_2]^T.$$

## Univariate linear regression

Use mapping $g_1 = (\cdot)^0, \ g_2 = (\cdot)^1$ obtain the matrix

$$X = \begin{bmatrix} g_1(x_1) & g_2(x_1) \\ g_1(x_2) & g_2(x_2) \\ \dots & \dots \\ g_1(x_m) & g_2(x_m) \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_m \end{bmatrix}.$$

According to the Least Squares,

$$\hat{\mathbf{w}} = (X^T X)^{-1} X^T \mathbf{y}, \quad \text{where } y = [y_1, \dots, y_m]^T.$$

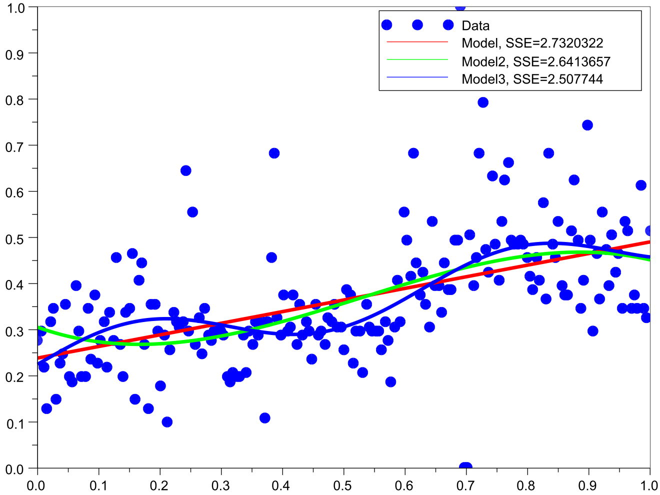
In the matrix notation, the linear model is

$$\mathbf{y} = X\mathbf{w} + \varepsilon,$$

and the regression function $\hat{\mathbf{y}} = X\hat{\mathbf{w}}$.
The error function is the Sum of Squared Errors,

$$SSE = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}).$$

## Univariate linear regression

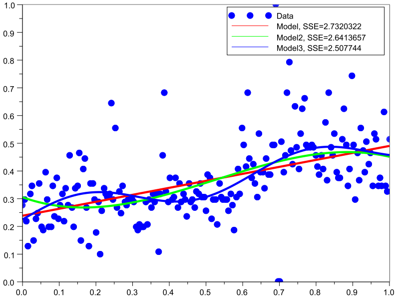


## Primitive functions

Introduce the set of the primitives:
$G = \{g_1, \dots, g_5\} = \{x^0, x^1, x^2, x^3, sin(10x)\}$ and the linear model

$$\hat{y}_i = \sum_{j \in \mathcal{A}} w_j g_j(x_i), \quad \text{for short } \hat{\mathbf{y}} = X_{\mathcal{A}}\mathbf{w}, \quad \text{where } \mathcal{A} \subseteq \{1, \dots, 5\};$$

and obtain the regression functions.



## Sales planning

The set of retailers problems:

- custom inventory,
- calculation of optimal insurance stocks,
- consumer demand forecasting.

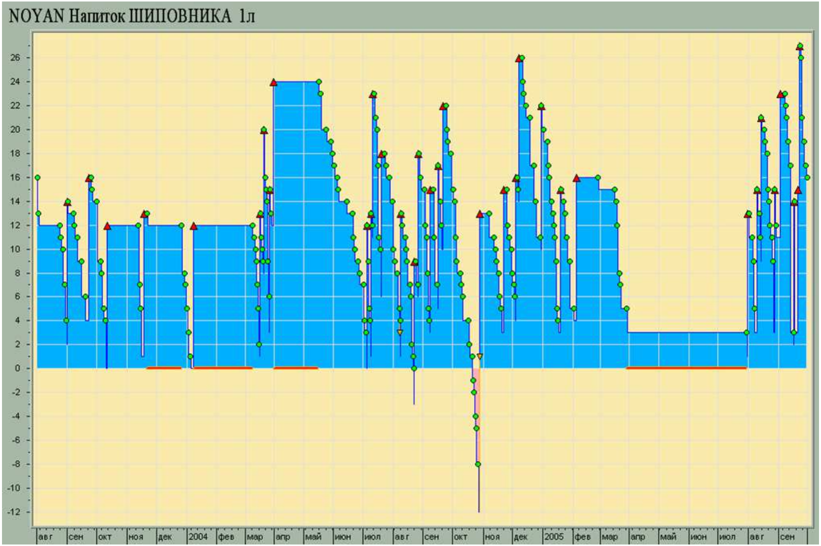
There given:

- time-scale,
- historical time series,
- additional time series;

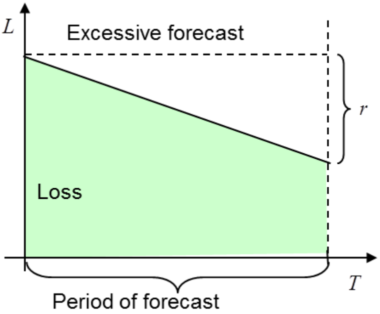
**We must forecast the time series.**

The quality of forecasting is the minimum loss of money.

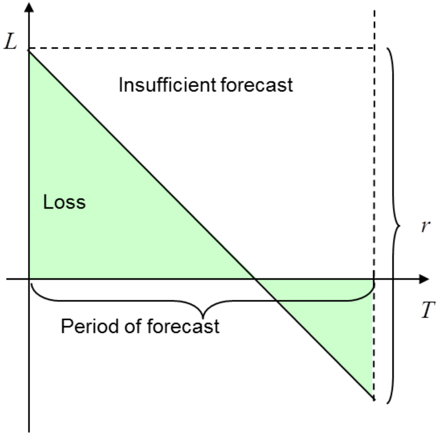
## Custom inventory



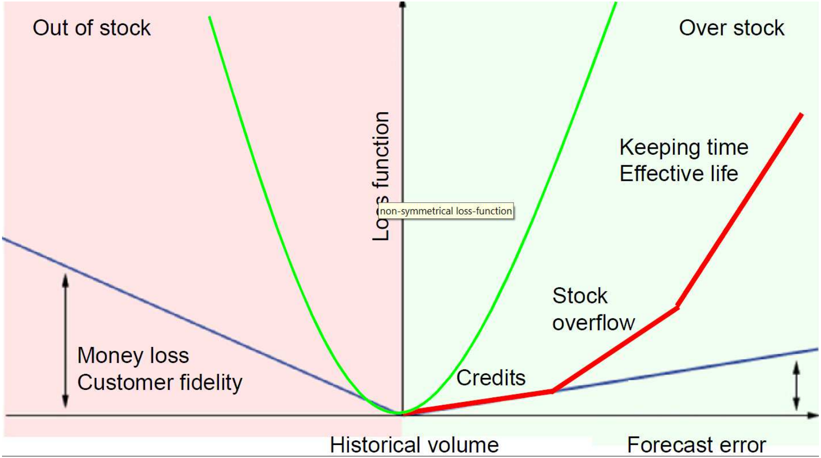
# Excessive forecast



# Insufficient forecast



# Error (loss) function

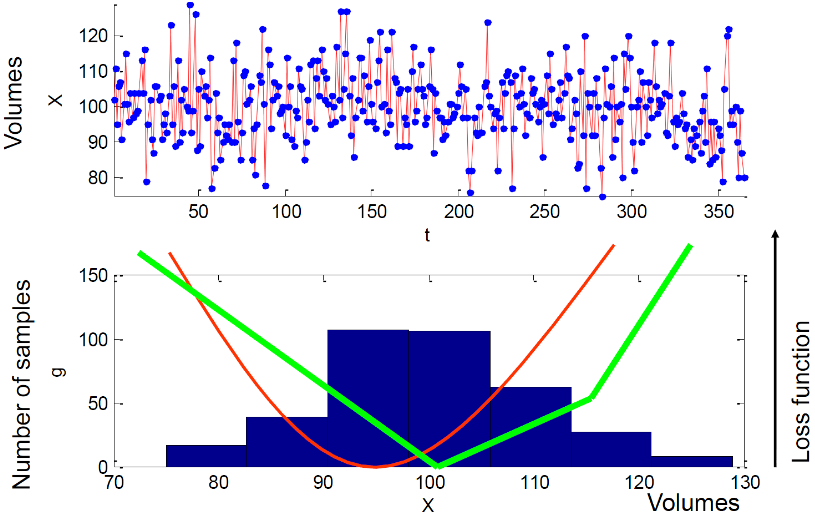



Quadratic function, Linear function, Asymmetric function

# Noisy time series forecasting

- There is a historical time series of the volume off-takes (i.e. foodstuff).
- Let the time series be homoscedastic (its variance is the time-constant).
- Using the loss function one must forecast the next sample.

# The time series and the histogram



## The forecasting algorithm

**Let there be given:**

the historgam $H = \{X_i, g_i\}, i = 1, \ldots, m;$

the loss function $L = L(Z, X);$

for example, $L = |Z - X|$ or $L = (Z - X)^2$.

**The problem:**

For given $H$ and $L$, one must find the optimal forecast value $\tilde{X}$.

**Solution:**

$$\tilde{X} = \arg \min_{Z \in \{X_1, \ldots, X_m\}} \sum_{i=1}^{m} g_i L(Z, X_i).$$

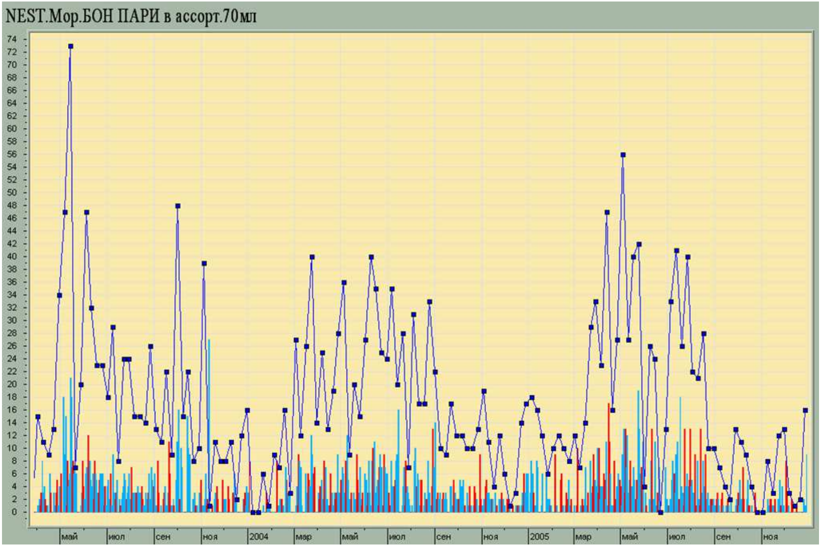
**Result:**

$\tilde{X}$ is the optimal forecast of the time series.

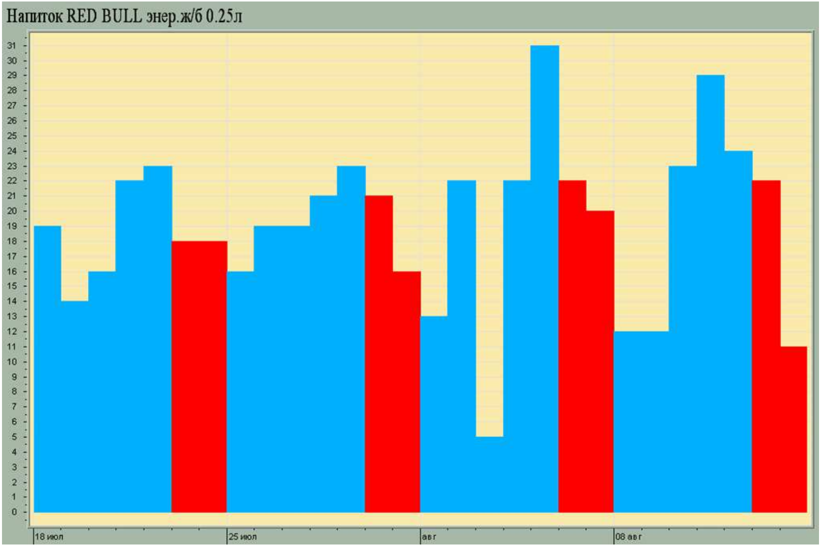
## The sales time series is non-stationary

- There is a trend — total increase or decrease in sales volume,
- periodic component — week and year cycles,
- aperiodic component — promotional actions and holidays,
- life cycle of goods — mobile phones.

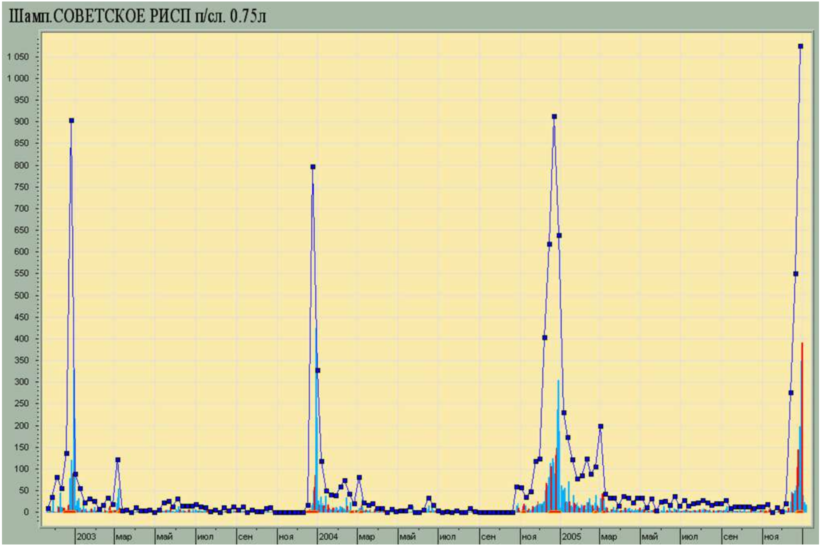
# Year seasonality



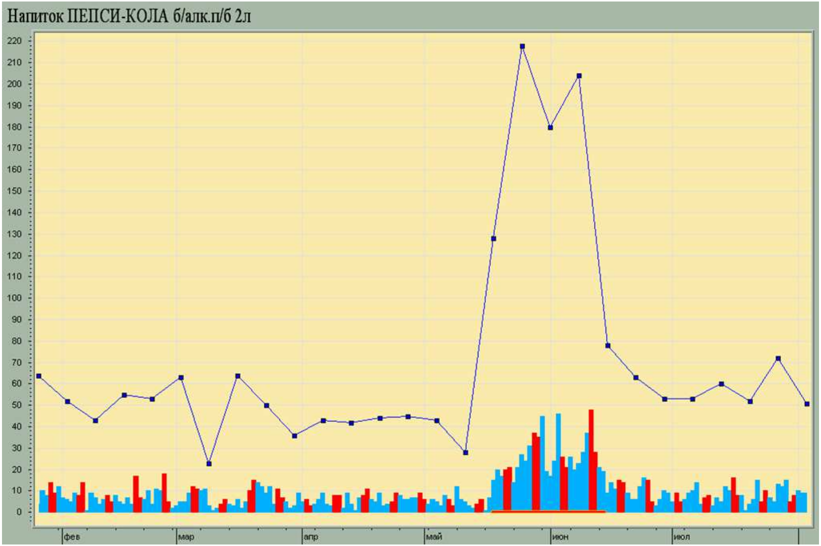
# Week seasonality



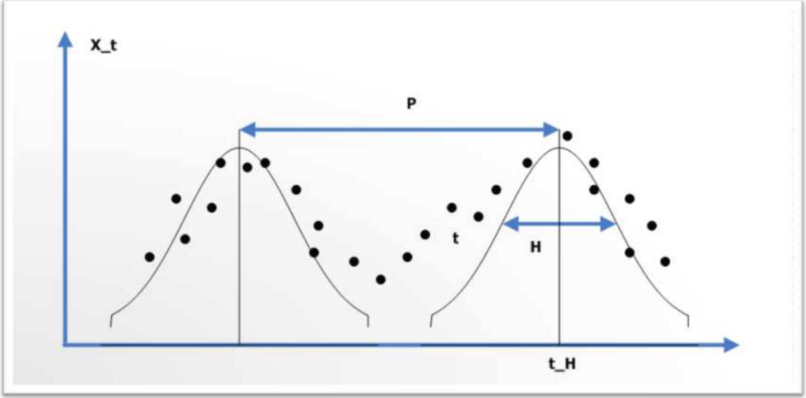
# Holidays and week-ends



# Promotional actions



## How to forecast quasi-periodical time series




Split the time series into the periods.

## Hour by Hour Energy Forecasting

Data:

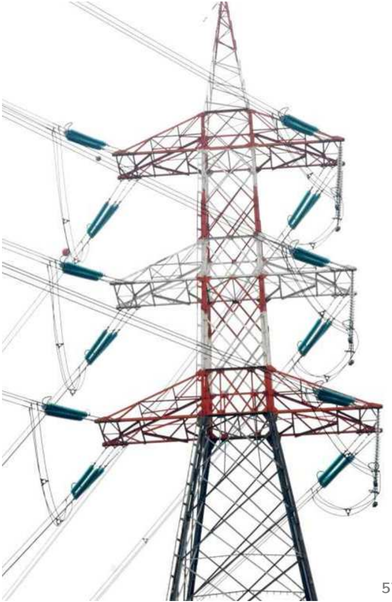
- historical consumption and prices, multivariate time series.

To forecast:

- hour-by-hour, the next day
  - ✓ consumption and
  - ✓ price.

Solution:

- the autoregressive model generation and model selection.



## The periodic components of the multivariate time series

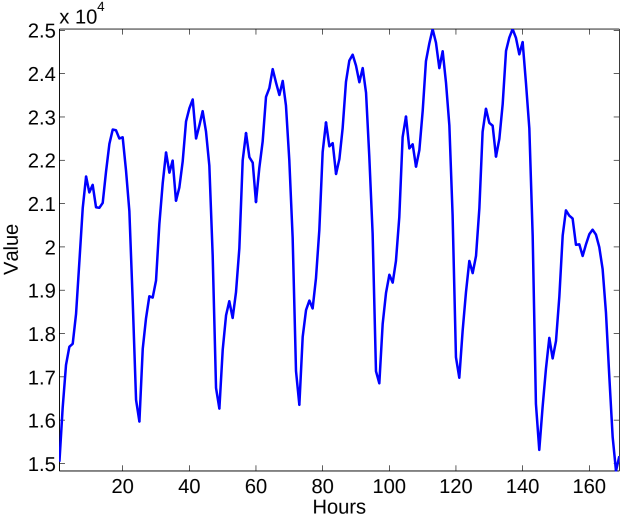
The time series:

- energy price,
- consumption,
- daytime,
- temperature,
- humidity,
- wind force,
- holiday schedule.

Periods:

- one year seasons (temperature, daytime),
- one week,
- one day (working day, week-end),
- a holiday,
- aperiodic events.

## Source time series, one week

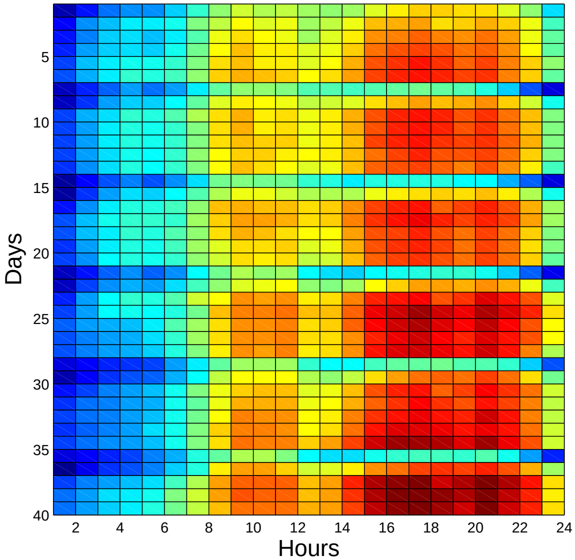


# The autoregressive matrix to forecast periodic time series

- There given the time series $\{s_1, \dots, s_\tau, \dots, s_{T-1}\}$, the length of a period is $\kappa$.
- One must to forecast the next sample $T$.
- The autoregressive matrix:
  - its $i$-th row is a period of samples,
  - its $j$-th column is a phase of the period and
  - they map into the time series sample number such that $(i-1)\kappa \mapsto \tau$; let mod $\frac{T}{\kappa} = 0$;

$$\underset{(m+1)\times(n+1)}{X^*} = \begin{pmatrix} s_T & s_{T-1} & \dots & s_{T-\kappa+1} \\ s_{(m-1)\kappa} & s_{(m-1)\kappa-1} & \dots & s_{(m-2)\kappa+1} \\ \dots & \dots & \dots & \dots \\ s_{n\kappa} & s_{n\kappa-1} & \dots & s_{n(\kappa-1)+1} \\ \dots & \dots & \dots & \dots \\ s_\kappa & s_{\kappa-1} & \dots & s_1 \end{pmatrix}.$$

# The autoregressive matrix, five week-ends



## The autoregressive matrix and the linear model

$$\underset{(m+1)\times(n+1)}{X^*} = \left(\begin{array}{c|cccc} s_T & s_{T-1} & \dots & s_{T-\kappa+1} \\ \hline \\ s_{(m-1)\kappa} & s_{(m-1)\kappa-1} & \dots & s_{(m-2)\kappa+1} \\ \dots & \dots & \dots & \dots \\ s_{n\kappa} & s_{n\kappa-1} & \dots & s_{n(\kappa-1)+1} \\ \dots & \dots & \dots & \dots \\ s_\kappa & s_{\kappa-1} & \dots & s_1 \end{array}\right).$$

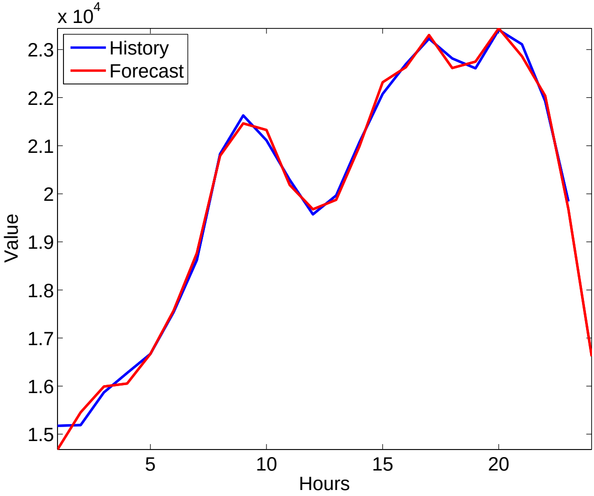
In a nutshell,

$$X^* = \left[\begin{array}{c|c} \underset{1\times 1}{s_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\ \hline \underset{m\times 1}{\mathbf{y}} & \underset{m\times n}{X} \end{array}\right].$$

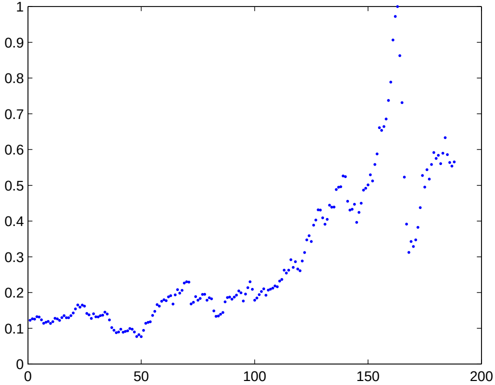
In terms of linear regression:

$$\hat{\mathbf{y}} = X\mathbf{w}, \quad \text{the forecast} \quad \hat{y}_{m+1} = \hat{s}_T = \mathbf{w}^\mathsf{T}\mathbf{x}_{m+1}^\mathsf{T}.$$

# The one-day forecast, an example

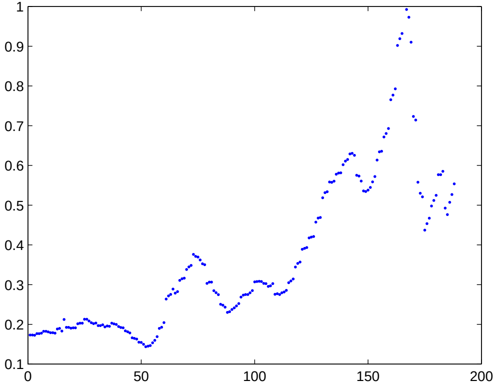


# Time series with a bubble, example 1



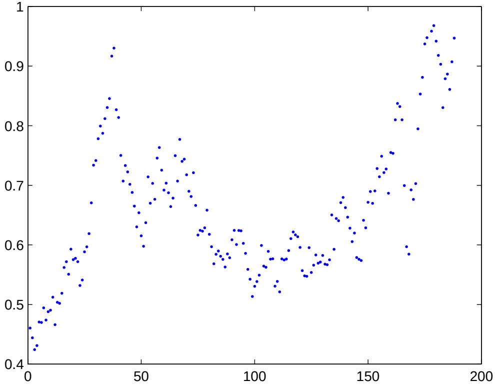

The World Bank. Global Economic Monitor 2010.
http://data.worldbank.org/data-catalog/global-economic-monitor.

## Time series with a bubble, example 2



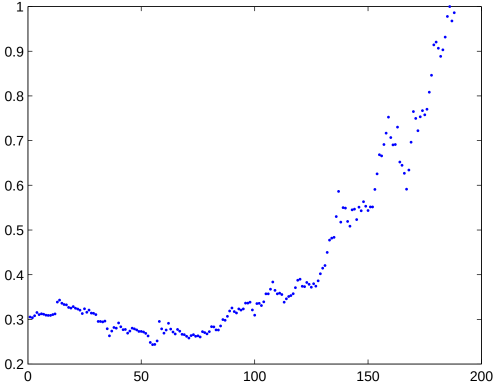

The World Bank. Global Economic Monitor 2010.
http://data.worldbank.org/data-catalog/global-economic-monitor.

# Time series with no bubble, example 3



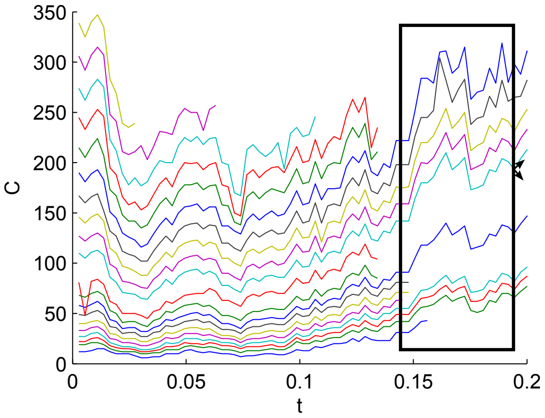

The World Bank. Global Economic Monitor 2010.
http://data.worldbank.org/data-catalog/global-economic-monitor.

# Time series with no bubble, example 4




The World Bank. Global Economic Monitor 2010.
http://data.worldbank.org/data-catalog/global-economic-monitor.

## Time series trends/events forecasting, example



## Event forecasting; One must forecast $s_{1,T+1} \in \mathbb{M} \ni \mathbb{R}$.

There are $N$ time series of length $T$ (an element $s_{n,t} \in \mathbb{M}$). Form the matrix

$$X^* = \begin{matrix} s_{1,1} & s_{1,2} & \cdots & s_{1,T-1} & s_{1,T} & s_{1,T+1} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,T-1} & s_{2,T} & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ s_{N,1} & s_{N,2} & \cdots & s_{N,T-1} & s_{N,T} & \end{matrix}$$

Denote $\Delta$ the time-lag and for the time series $[s_{1,t}]$, $t \in \{\Delta+1, \dots, T\}$ form the matrix

$$X^*_t = \begin{matrix} s_{1,t-\Delta} & s_{1,t-\Delta+1} & \cdots & s_{1,t-2} & s_{1,t-1} \\ s_{2,t-\Delta} & s_{2,t-\Delta+1} & \cdots & s_{2,t-2} & s_{2,t-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ s_{N,t-\Delta} & s_{N,t-\Delta+1} & \cdots & s_{N,t-2} & s_{N,t-1} \end{matrix}$$

and vectorizing it, obtain the sample $\mathbf{x}_t$

$$\mathbf{x}_t = [s_{1,t-\Delta}, s_{2,t-\Delta}, \dots, s_{N,t-\Delta}, s_{1,t-\Delta+1}, \dots, s_{N,t-1}]^T.$$

## Event forecasting is a classification problem

Introduce the data set $D = (X, \mathbf{y})$, where

$$X = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_{T-\Delta}^\mathsf{T} \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{T-\Delta} \end{bmatrix}.$$

Treat this classification problem as the logistic regression

$$\mathbf{f}(\mathbf{w}, X) = \frac{1}{1 + \exp(-X\mathbf{w})}.$$

## Decision support and Integral indicator construction

### The integral indicator is a measure

of object's quality. *It is a scalar, corresponded to an object.*

### The integral indicator is an aggregation

of object's features that describe various components of the term "quality". *Expert estimation of object's quality could be an integral indicator, too.*

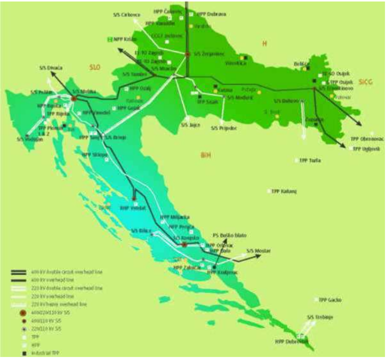
# Examples

| Index name | Objects | Features | Model |
| --- | --- | --- | --- |
| TOEFL exams | Students | Tests | Sum of scores |
| Eurovision | Singers | Televotes, Jury votes | Linear (weighted sum) |
| S&P500, NASDAQ | Time-ticks | Shares (prices, volumes) | Non-linear |
| Bank ratings | Banks | Requirements | By an expert commission |
| **Integral Indicator of Croatian PP's** | **Power Plants** | **Waste measurements** | **Linear** |

# There given a set of objects

Croatian Thermal Power Plants and Combined Heat and Power Plants

1. Plomin 1 TPP
2. Plomin 2 TPP
3. Rijeka TPP
4. Sisak TPP
5. TE-TO Zagreb CHP
6. EL-TO Zagreb CHP
7. TE-TO Osijek CHP
8. *Jetrovac TPP*



# There given a set of features

## Outcomes and Waste measurements

1. Electricity (GWh)
2. Heat (TJ)
3. Available net capacity (MW)
4. $SO_2$ (t)
5. NOX (t)
6. Particles (t)
7. $CO_2$ (kt)
8. Coal (kt)
9. Sulphur content in coal (%)
10. Liquid fuel (kt)
11. Sulphur content in liquid fuel (%)
12. Natural gas ($10^6$ $m^3$)

## How to construct an index?

**Assign a comparison criterion**

Ecological footprint of the Croatian Power Plants

**Gather a set of comparable objects**

TPP and CHP (Jetrovac TPP excluded)

**Gather features of the objects**

Waste measurements

**Make a data table: objects/features**

See 7 objects and 10 features in the table below

**Select a model**

Linear model (with most informative coefficients)

# Data table and feature optimums

| N | Power Plant | Electricity (GWh) | Heat (TJ) | Available net capacity (MW) | $SO_2$ (t) | NOx (t) | Particles (t) | $CO_2$ (kt) | Coal (kt) | Sulphur content in coal (%) | Liquid fuel (kt) | Sulphur content in liquid fuel (%) | Natural gas ($10^6$ $m^3$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Plomin 1 TPP | **452** | **0** | 98 | 1950 | 1378 | 140 | 454 | 198 | 0.54 | 0.43 | 0.2 | 0 |
| 2 | Plomin 2 TPP | **1576** | **0** | 192 | 581 | 1434 | 60 | 1458 | 637 | 0.54 | 0.37 | 0.2 | 0 |
| 3 | Rijeka TPP | **825** | **0** | 303 | 6392 | 1240 | 171 | 616 | 0 | 0 | 200 | 2.2 | 0 |
| 4 | Sisak TPP | **741** | **0** | 396 | 3592 | 1049 | 255 | 573 | 0 | 0 | 112 | 1.79 | 121 |
| 5 | TE-TO Zagreb CHP | **1374** | **481** | 337 | 2829 | 705 | 25 | 825 | 0 | 0 | 80 | 1.83 | 309 |
| 6 | EL-TO Zagreb CHP | **333** | **332** | 90 | 1259 | 900 | 19 | 355 | 0 | 0 | 39 | 2.1 | 126 |
| 7 | TE-TO Osijek CHP | **114** | **115** | 42 | 1062 | 320 | 35 | 160 | 0 | 0 | 37 | 1.1 | 24 |
| | | | | max | min | min | min | min | min | min | min | min | min |

# Notations

$X = \{x_{ij}\}$ is the $(n \times m)$ is the real matrix, the data set;
$\mathbf{y} = [y_1, \dots, y_m]^\mathsf{T}$ is the vector of integral indicators;
$\mathbf{w} = [w_1, \dots, w_n]^\mathsf{T}$ is the vector of feature importance weights;

$\mathbf{y}_0, \mathbf{w}_0$ are the expert estimations of the indicators and the weights;

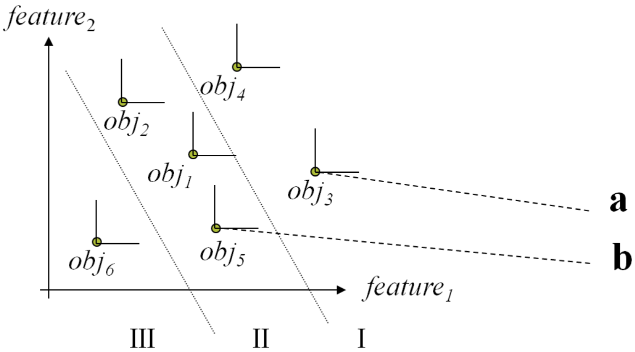
$$\begin{array}{c|c} & \mathbf{w}^\mathsf{T} \\ \hline \mathbf{y} & X \end{array} \quad = \quad \begin{array}{c|cccc} & w_1 & w_2 & \dots & w_n \\ \hline y_1 & x_{11} & x_{12} & \dots & x_{1n} \\ y_2 & x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \vdots & \dots \\ y_m & x_{m1} & x_{m2} & \dots & x_{mn} \end{array}.$$

## Usually, data prepared so that

- the minimum of each feature equals 0, while the maximum equals 1;
- the bigger value of each implies better quality of the index.

## Pareto slicing

Find the non-dominated objects at each slicing level.




### The object a is non-dominated

if there is no $\mathbf{b}_i$ such that $b_{ij} \geqslant a_j$ for all features index $j$.

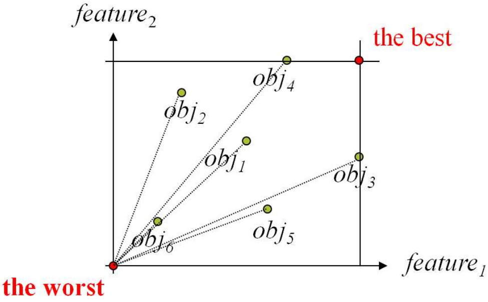
## Metric algorithm

The best (worst) object is an object that contains the (maximum) minimum values of the features.

### The index is

$$y_i = \sqrt[r]{\sum_{j=1}^{r}\left(x_{ij} - x_j^{\text{best}}\right)^r}$$

For r = 1, this algorithm coincides the weighted sum with equal weighs.



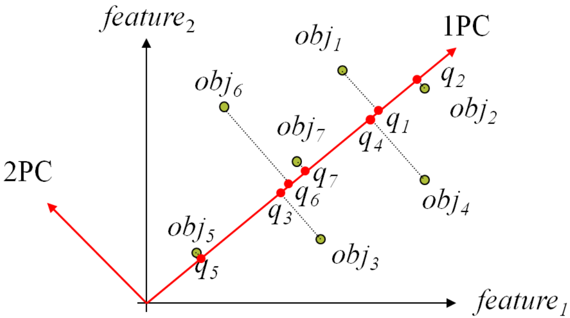
## Weighted sum

$$\mathbf{y}_1 = X\mathbf{w}_0,$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \vdots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_m \end{bmatrix}.$$

## Principal Components Analysis

$Y = XV$, where $V$ is the rotation matrix of the principal components. The indicators $\mathbf{y}_{\text{PCA}} = X\mathbf{w}_{\text{1PC}}$, where $\mathbf{w}_{\text{1PC}}$ is the $1^{\text{st}}$ column vector of the matrix $V$ in the singular values decomposition $X = ULV^{\mathsf{T}}$.

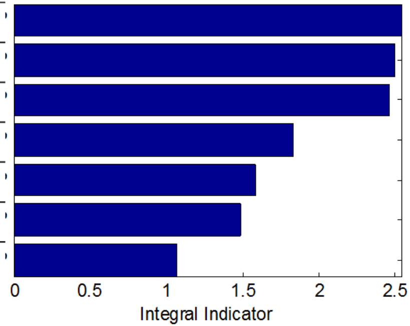



PCA gives minimum mean square error between objects and their projections.

# The Integral Indicator

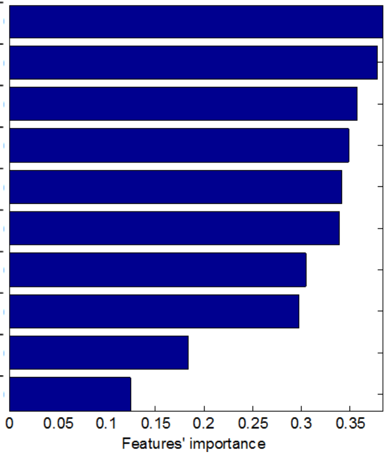
## Ecological Impact of the Croatian Power Plants

| Power Plant | Integral Indicator |
|---|---|
| TE-TO Zagreb CHP | 2.53 |
| EL-TO Zagreb CHP | 2.49 |
| TE-TO Osijek CHP | 2.46 |
| Plomin 2 TPP | 1.83 |
| Rijeka TPP | 1.57 |
| Sisak TPP | 1.48 |
| Plomin 1 TPP | 1.07 |

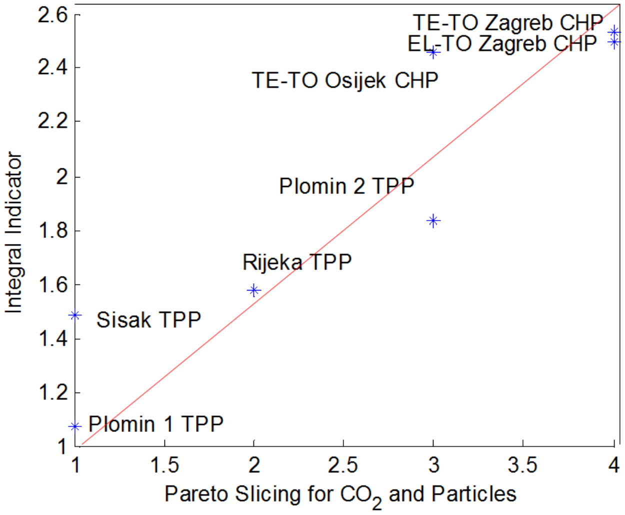


# The Importance Weights of the Features

| Feature | Weight |
|---|---|
| Coal (t) | 0.38 |
| Sulphur content in coal (%) | 0.37 |
| NOx (t) | 0.35 |
| Liquid fuel (t) | 0.34 |
| SO2 (t) | 0.34 |
| Particles (t) | 0.33 |
| Natural gas (103 m3) | 0.30 |
| CO2 (kt) | 0.29 |
| Sulphur content in l.fuel (%) | 0.18 |
| Available net capacity (MW) | 0.12 |

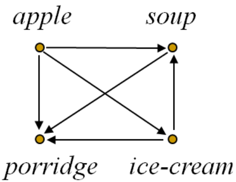


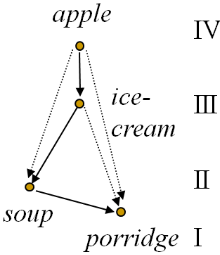
## The PCA Indicator versus Pareto Slicing



## Pair-wise comparison, toy example

| | *a* | *s* | *p* | *i-c* |
|---|---|---|---|---|
| *apple* | • | + | + | + |
| *soup* | | • | + | – |
| *porridge* | | | • | – |
| *ice-cream* | | | | • |



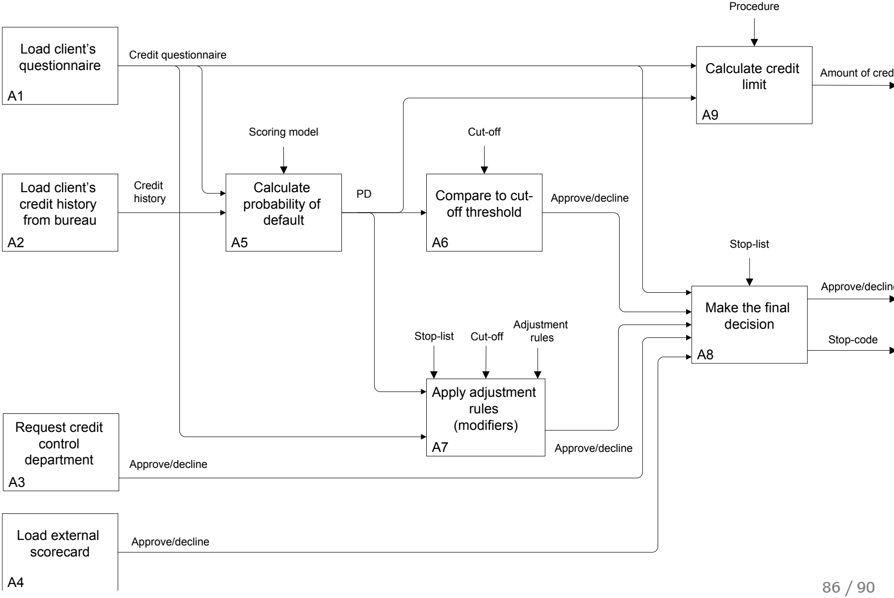




If an object in a row is better than the other one in a column then put "+", otherwise "-".

Make a graph, *row* + *column* means *row* •⟶• *column*.

Find the top and remove extra nodes.

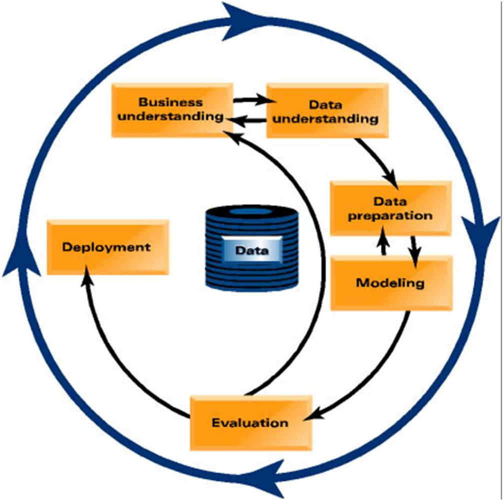
# Role of quantitative model in business process



# CRISP-DM: Cross Industry Standard Process for Data Mining

Six major phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment



## Project description

The description of the project is intended to secure the understanding of the goals, methods and results of the project.

1. **Goal**
   - The main goal, general description of the project
2. **Motivation**
   - Where and how the results of the project will be used
3. **Data**
   - Description of the sample set
4. **Quality measurement**
   - The target quality function description
5. **Requirements**
   - The conditions of successful project termination
6. **Feasibility**
   - The possible obstacles that arise during the project
7. **Methods**
   - Some recommendations on methods and algorithms which will be used

## Tools for quantitative modelling

- Matlab, Scilab, Octave
- Mathematica, Maple
- SPSS, Statistica, R-project
- SAS-Enterprize, SAS-Data Miner
- ORACLE SQL, SQL-Developer
- Ksema-XSEN
- Phyton, C-languages
- ...
- MS-Excel

## Resume

1. Data and Model
2. Classification
3. Forecasting
4. Decision making
5. Application

**The lectures and updates are on Twitter @strijov**

The link to the slides:
strijov.com/papers/Strijov2012QuantitativeModelling.pdf