Introduction to Regression

(linear, logistic, multivariate, nonlinear)

and a message at the end

Martin Sebera

Magdeburg, January 2024

Content

- I. Linear regression
- II. Logistic regression
- III. Multivariate Regression
- IV. Nonlinear regression - Neural net

What it is regression?

- statistical method that helps us **understand the relationship between variables**.
- exploring how one variable affects another.
- In a sports setting, we can use linear regression to predict outcomes or analyze performances. Let's demonstrate this with three examples:

History of regression

- The term regression comes from the works of anthropologist and meteorologist Francis Galton, which he presented to the public between 1877 and 1885.
- The question of heredity and specifically the relationship between the height of fathers and their first-born sons.
- The "return tendency" of the next generation towards the mean was called regression by Galton (he originally called this phenomenon reversion, which he later changed to regression = a step back).
- Although the current concept of regression analysis has little in common with Galton's original intention, the idea of accessing empirical data has remained, and the term regression has become so accepted that it is still used today

Correlation

- Correlation - the mutual relationship between two variables.
- If there is a correlation between two variables, it is likely that they depend on each other, but this does not mean that one of them must be the cause and the other the effect. **The correlation alone does not allow to decide.**

Procedure

- Model design, where we choose the appropriate **shape of the regression function**. If the theoretical model is not known, we analyze the point diagram and the graph of conditional averages.
- **Estimation** of regression parameters and tests of their significance.
- **Regression diagnostics**, when we perform residual analysis and identification of influential points.
- Assessment of **model quality**. The result is either the acceptance of the proposed model or the design of another model.

Regression procedure

- Working with regression models is actually much more difficult.
- It is necessary to test many assumptions (normality, homogeneity of variances, multicollinearity), choose an appropriate method (method of least squares, maximum likelihood), test residuals, analyze the quality of the model (residual variance, index of determination, Akaike information criterion, ROC curve, Gain graph), etc. .
- The following examples are more emotive, which are intended to show the possibilities of regression.

I. Linear regression

- Linear regression analysis is used **to predict** the value of a variable based on the value of another variable. The variable you want to predict is called the ***dependent variable***. The variable you are using to predict the other variable's value is called the ***independent variable***.
- It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation.

The most frequently used functions

- regression line $Y = \beta_0 + \beta_1\, x$
- regression parabola $Y = \beta_0 + \beta_1\, x + \beta_2\, x^2$
- logarithmic regression $Y = \beta_0 + \beta_1 \ln x$
- regression hyperbola $Y = \beta_0 + \beta_1 \frac{1}{x}$

Coefficient of determination R² - how well the data fit the regression model (how well the model explains the data).

- **A value of 0** - the model does not explain any variability in the dependent variable.
- **A value of 1** - the model perfectly explains all the variability in the dependent variable.
- A higher R² value means that the model better explains the variability of the dependent variable. However, it is important to keep in mind that a higher R² does not necessarily mean that the model is appropriate or accurate.

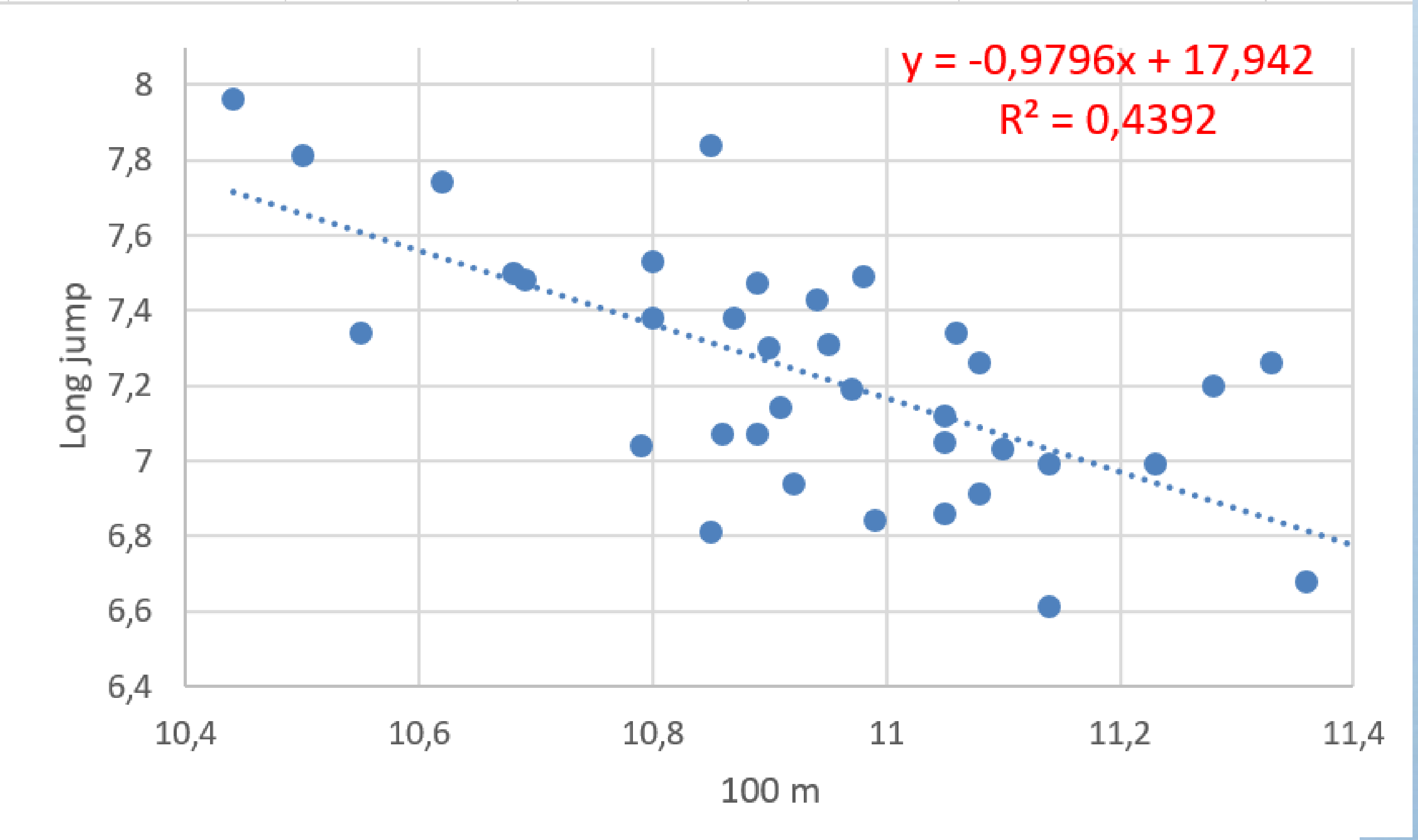
Example 1

- dependence of performance in the long jump on performance in the 100 m run

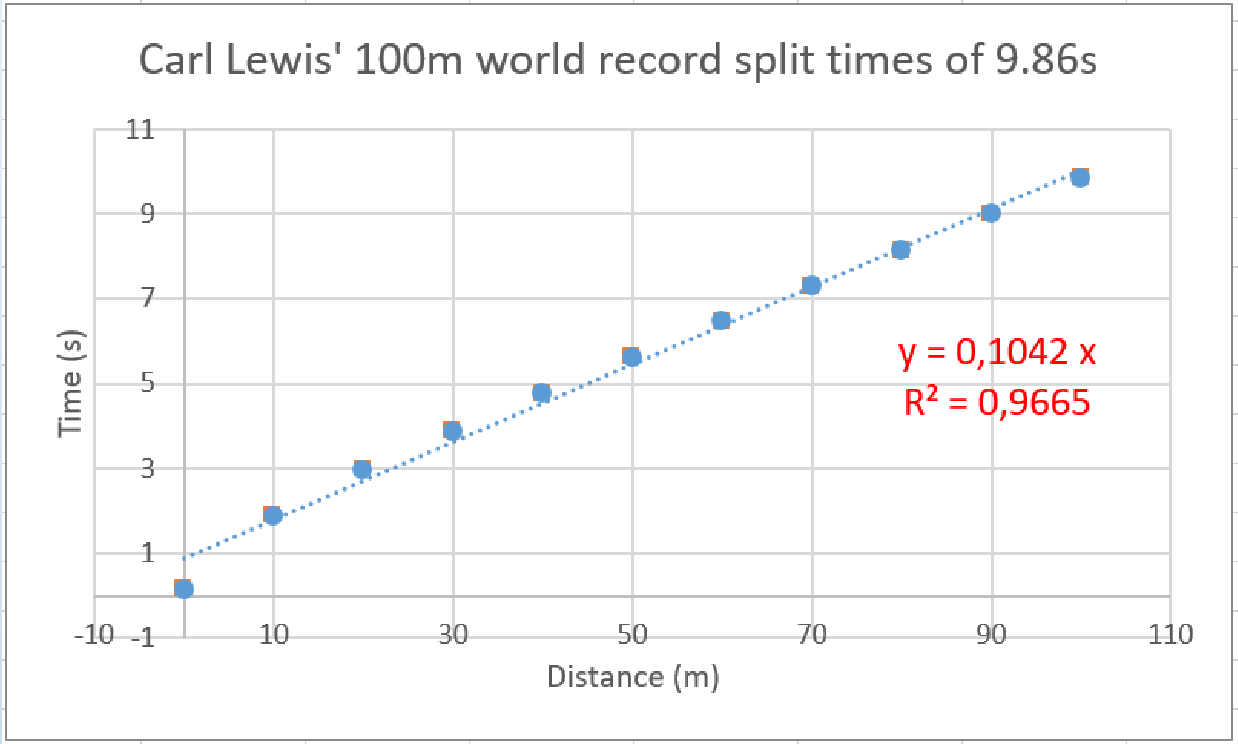
- **LJ** = -0,98 * **100m** + 17,94

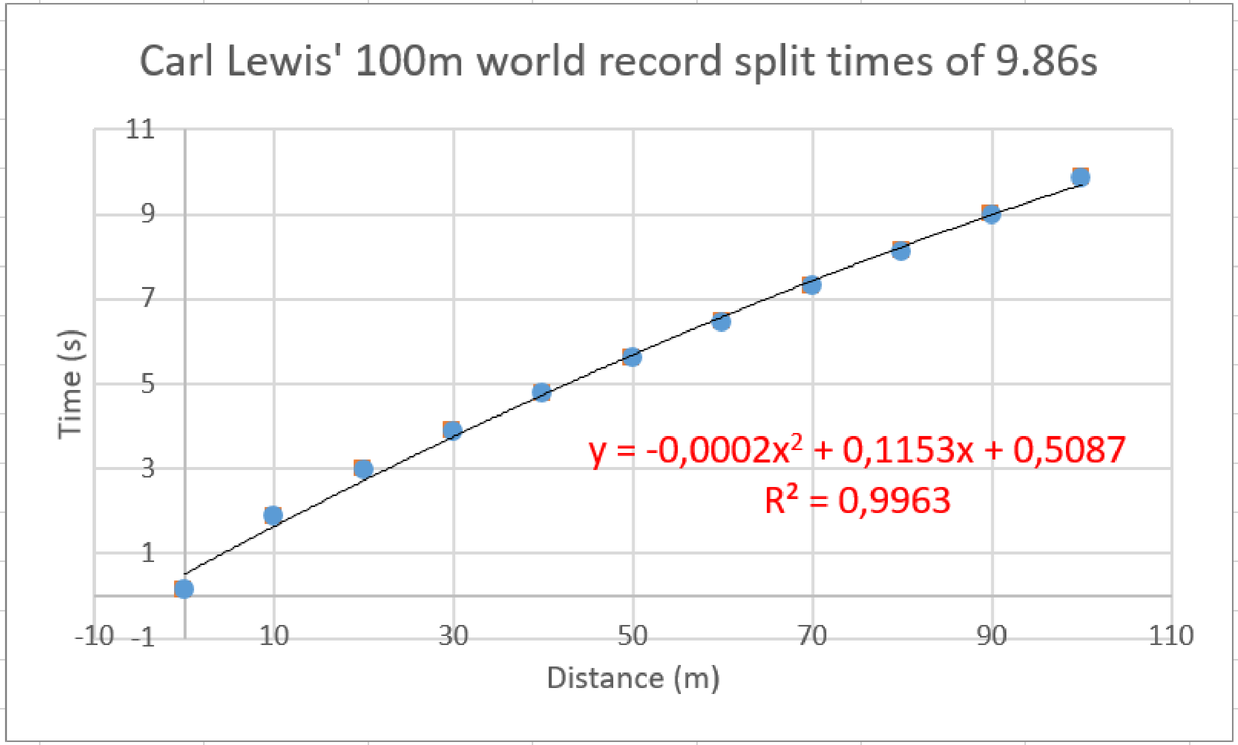
- Geometrically speaking, the coefficient of the independent variable is the tangent of the angle the line makes with the x-axis.

- arctg(-0,9796) = -45°



Example 2



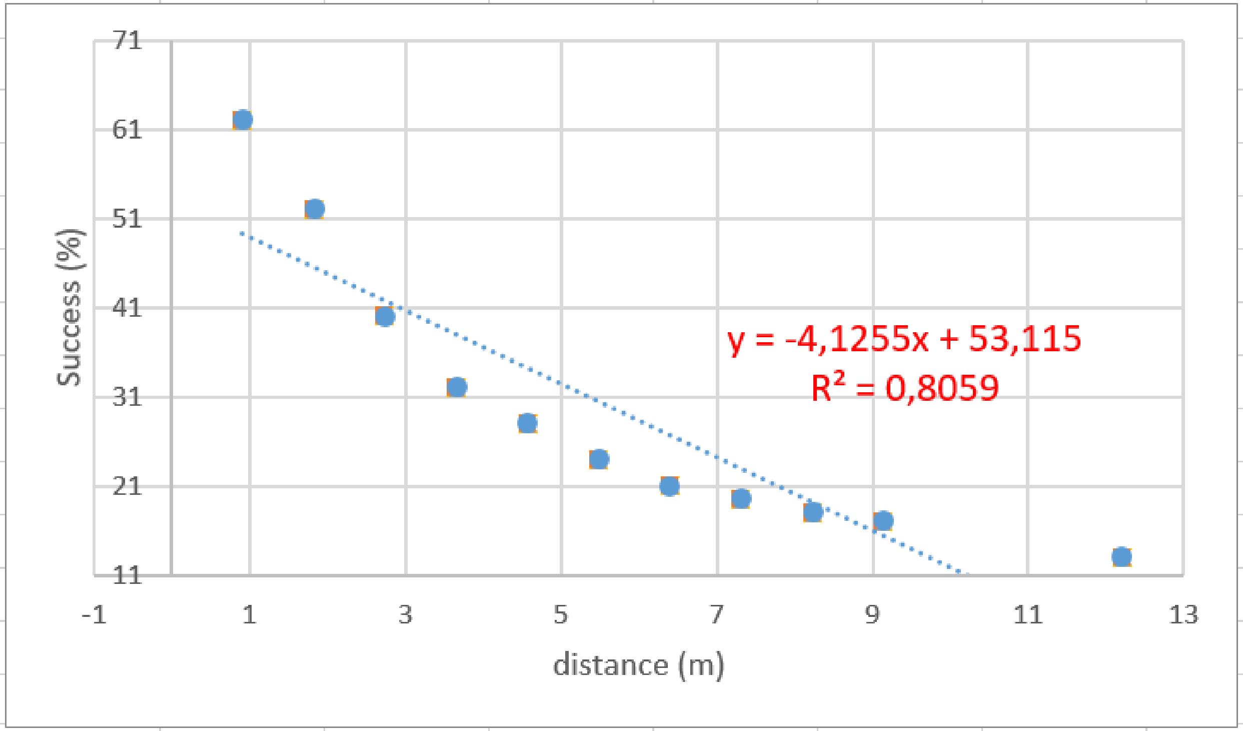


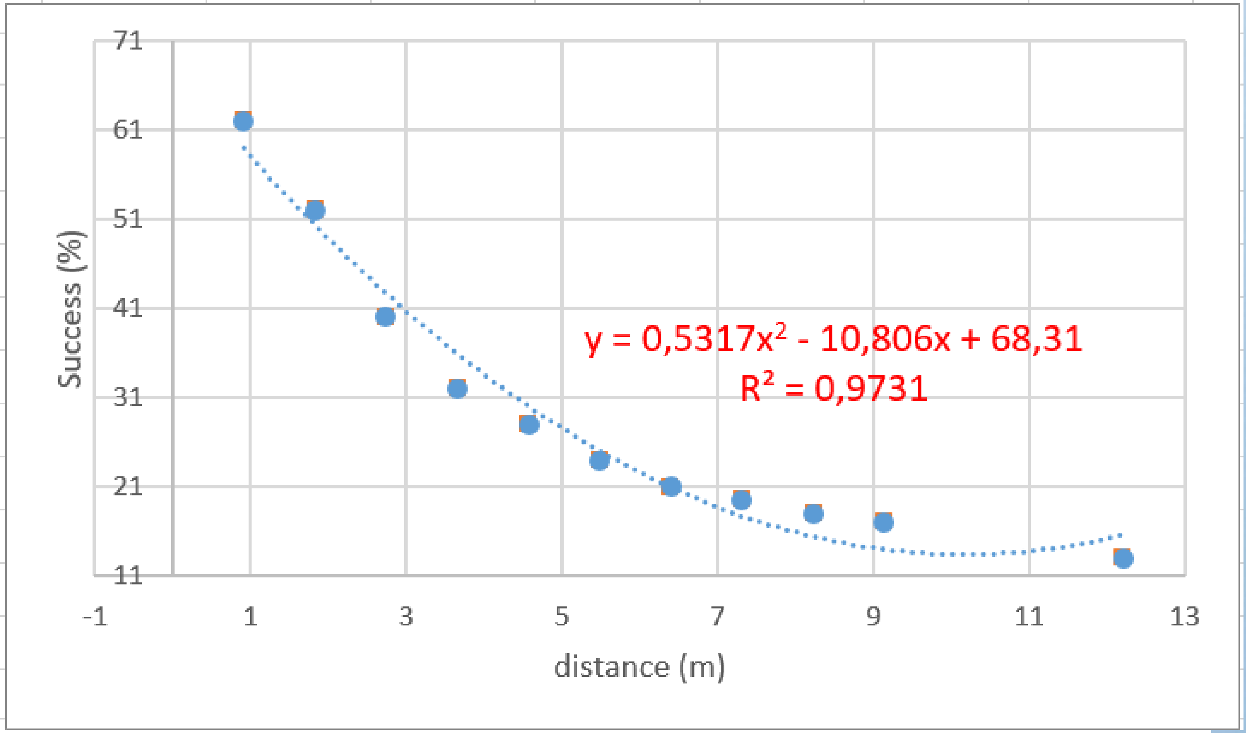
- Linear and quadratic linear regression models.
- Quadratic regression has a slightly higher quality (both models are very accurate, because $R^2 \to 1$). The quadratic model takes into account "fatigue" - the decrease in speed during the sprint

Example 3 - Shooting success in basketball

- If the horizontal distance of the basketball player from the basket increases, the percentage of shooting success decreases with this distance

foot	m	%
3	0,9	62
6	1,8	52
9	2,7	40
12	3,7	32
15	4,6	28
18	5,5	24
21	6,4	21
24	7,3	20
27	8,2	18
30	9,1	17
40	12	13

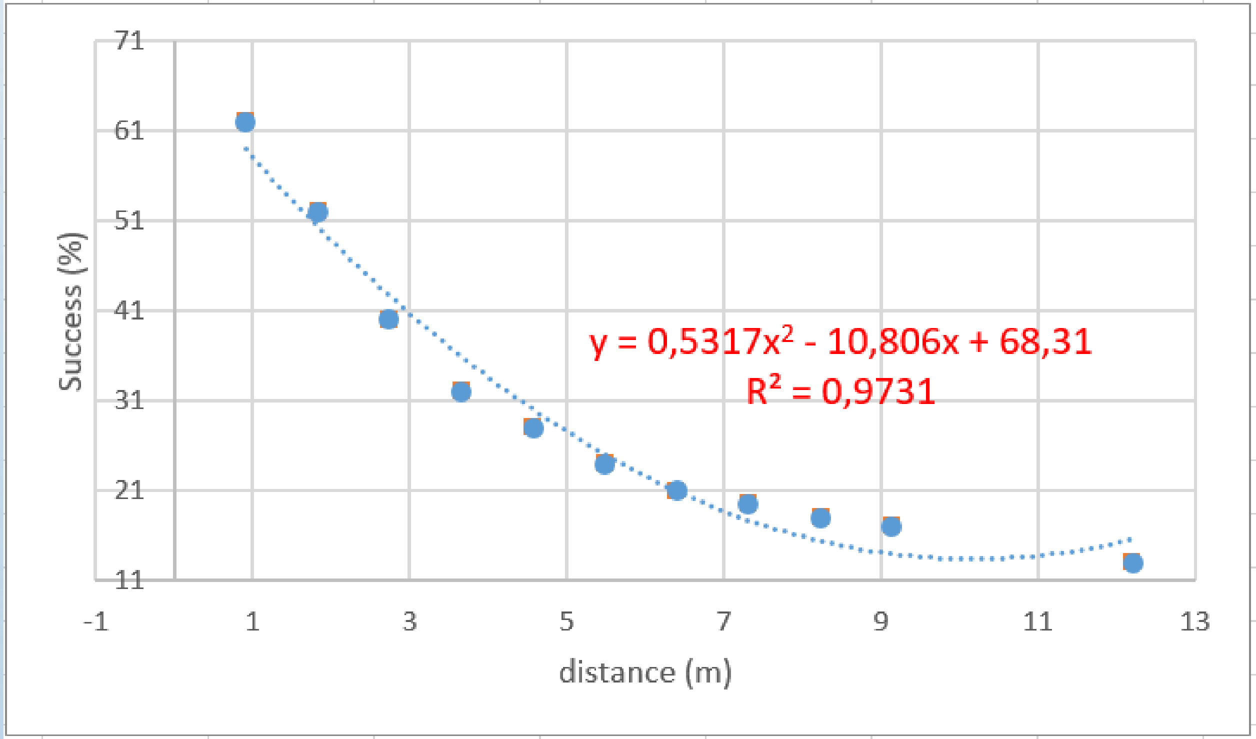


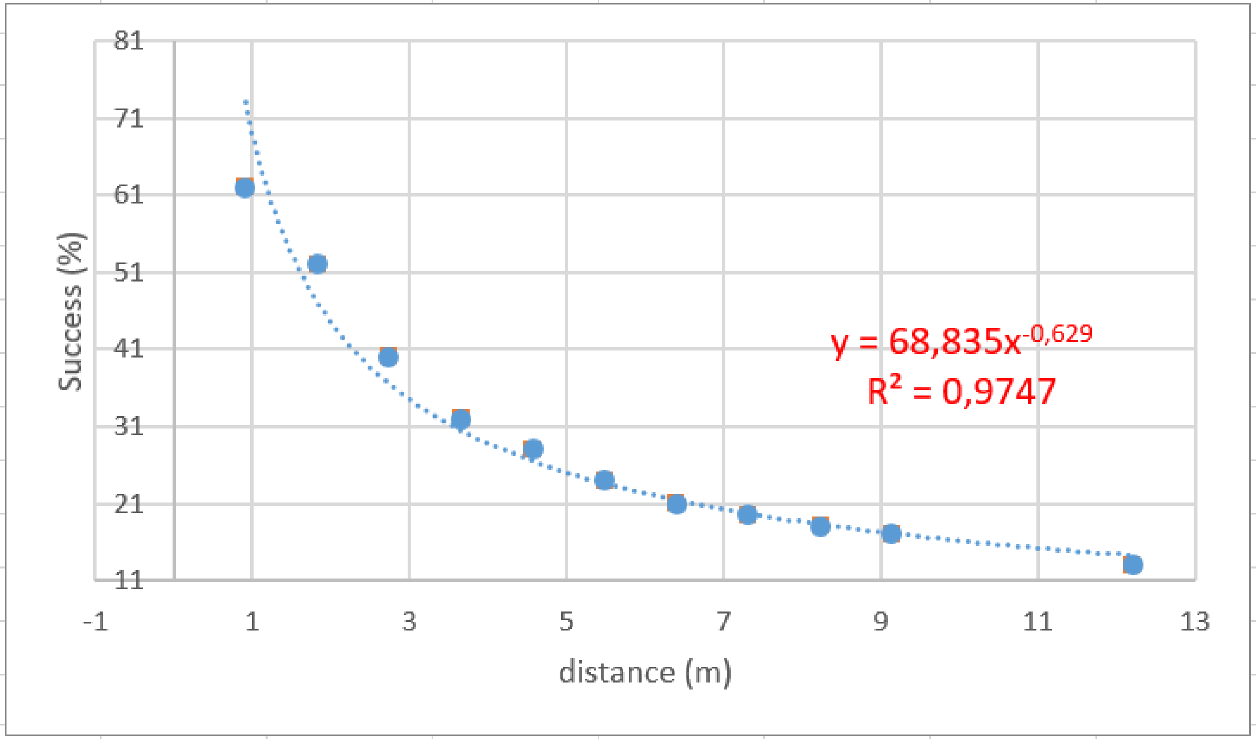


Example 3 - Shooting success in basketball

- However, the best regression model will be a power model. Why? See how the curve would behave with further distance...

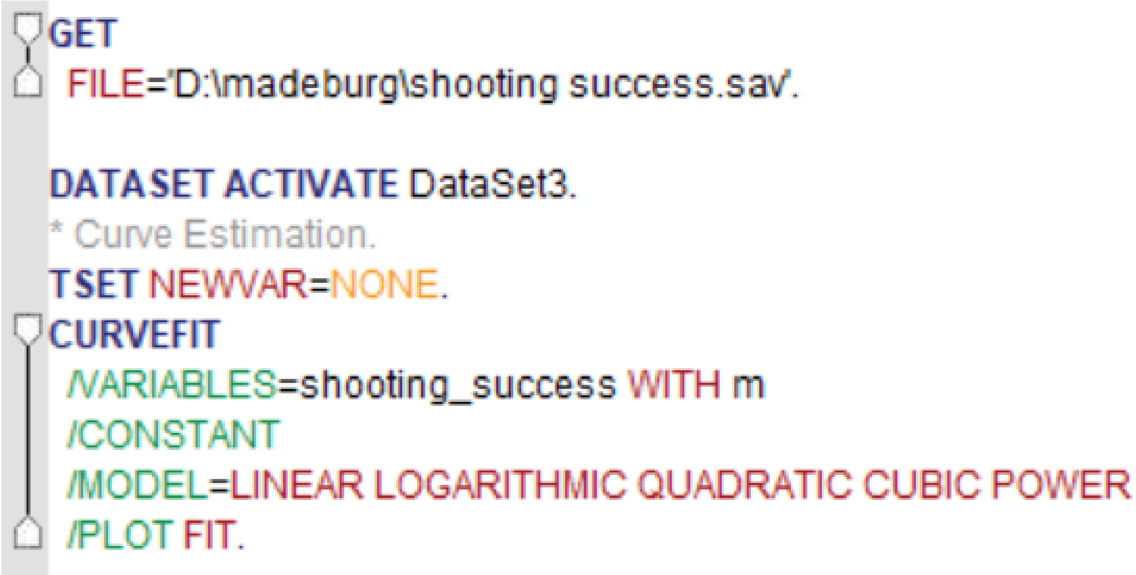
foot	m	%
3	0,9	62
6	1,8	52
9	2,7	40
12	3,7	32
15	4,6	28
18	5,5	24
21	6,4	21
24	7,3	20
27	8,2	18
30	9,1	17
40	12	13





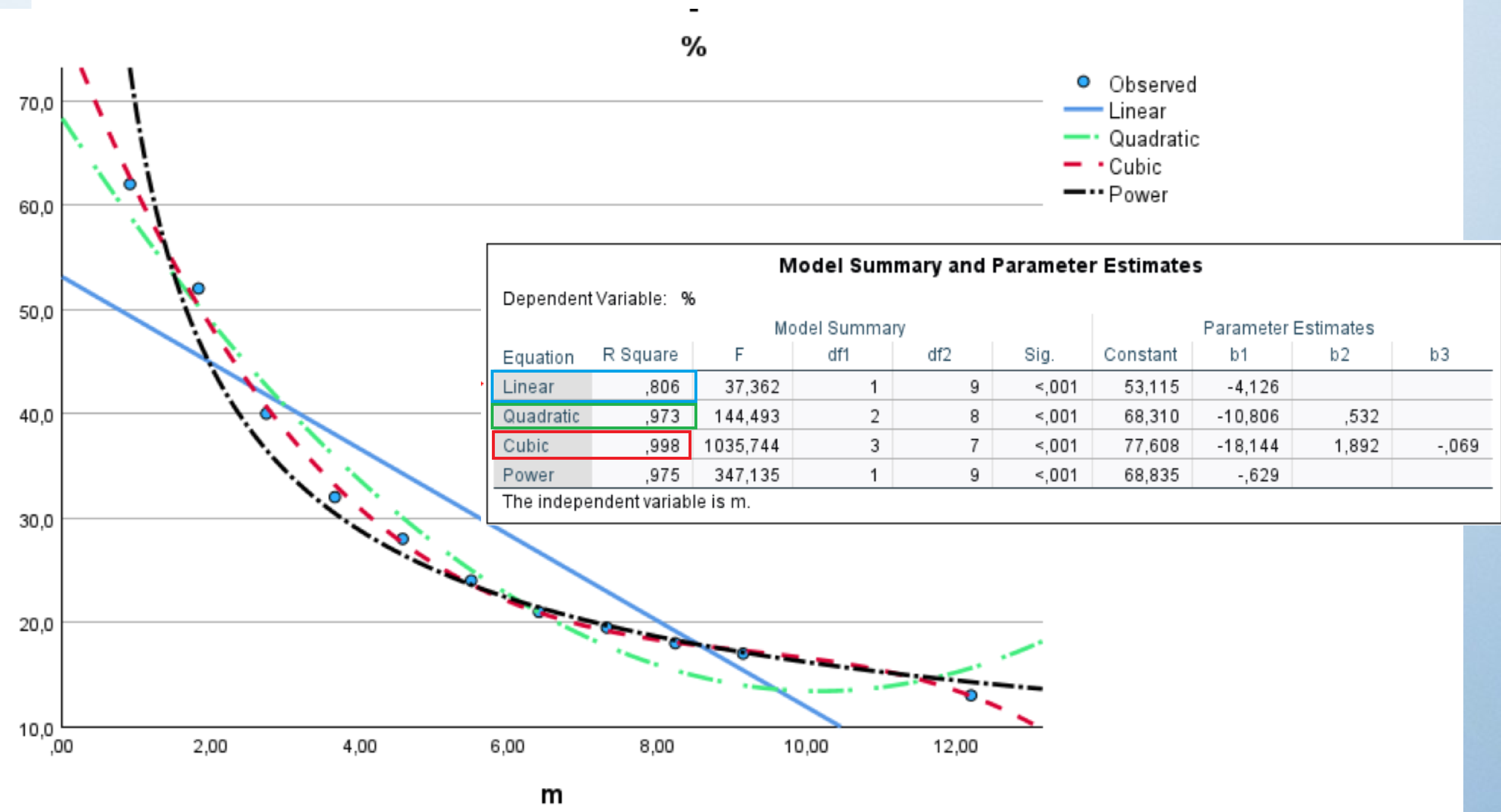
Example 3 - Shooting success in basketball in SPSS

- The chosen type of regression function must first of all respect the logical and objective connections of the phenomena and their regularities
- SPSS → Analyze → Regression → Curve estimation



```
GET
  FILE='D:\madeburg\shooting success.sav'.

DATASET ACTIVATE DataSet3.
* Curve Estimation.
TSET NEWVAR=NONE.
CURVEFIT
  /VARIABLES=shooting_success WITH m
  /CONSTANT
  /MODEL=LINEAR LOGARITHMIC QUADRATIC CUBIC POWER
  /PLOT FIT.
```



Model Summary and Parameter Estimates

Dependent Variable: %

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	,806	37,362	1	9	<,001	53,115	-4,126		
Quadratic	,973	144,493	2	8	<,001	68,310	-10,806	,532	
Cubic	,998	1035,744	3	7	<,001	77,608	-18,144	1,892	-,069
Power	,975	347,135	1	9	<,001	68,835	-,629		

The independent variable is m.

Example 4 – height, weight

- *We take the measures of people in the class and try to estimate the shape of the regression curve*

II. Logistic regression

- Logistic regression is a statistical method used to analyze data where the dependent variable is categorical, usually binary (ie has two possible values, such as yes/no, success/failure, 0/1).
- The main goal of logistic regression is **to model the probability** that a given input sample belongs to one of two categories.
- Example:
  - Pollard, R., & Reep, C. (1997). Measuring the Effectiveness of Playing Strategies at Soccer. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *46*(4), 541–550.

Logistic regression - example

Pollard and Reep (1997) used logistic regression to investigate the effectiveness of different strategies in soccer and their effect on the probability of scoring a goal. They wanted to discover how certain characteristics of the situation affect the probability of a goal being scored. As the basic event (dependent variable), they chose a situation that ended with a shot on goal. There were 489 of these events. They also identified the characteristics of the situations that could affect the outcome of the event. They chose as predictors:

- **distance from the goal** in meters (DIST);
- **the angle** (ANGLE) to the nearest goal post;
- measure of **how many touches** the player had with the ball before shooting: one (TOUCH = 0), more than one (TOUCH = 1);
- distance measure of the **closest opponent**: less than one meter (TIGHTNESS = 0), more than one meter (TIGHTNESS = 1);
- origin of ball **acquisition**: from play (**GAIN** = 0), free kick or throw from the sideline (GAIN = 1).

Logistic regression - example

The available information made it possible to complete these variables for all 489 events. Head shots and kick shots were analyzed in particular. For 410 kick shots, the regression equation was found:

Ln(goal chance) = 1.245 - 0.219 DIST - 1.578 ANGLE + 0.947 TIGHTNESS - 1.069 GAIN

This formula allows you to calculate the probability of scoring a goal in different situations.

For example, let's assume a kick from 15 meters (DIST=15) directly in front of the goal (ANGLE=0) with an opponent less than one meter away (TIGHTNESS=0) when the player got to the ball after a free kick (GAIN=0). The value of Ln(goal chance) = y = 3.109. The probability is calculated according to the formula

$$p^{(x)} = \frac{e^{y}}{1+e^{y})} = \frac{e^{3,109}}{1+e^{3,109})} = 0{,}9572; \qquad \ln(0{,}9572) = \mathbf{0{,}043}$$

The probability of scoring a goal in this situation is 4,3 %

Logistic regression - example

The chosen model allows for further interpretations. For example

- from the coefficient of the variable **DIST** we get by transforming $e^{0,219}$ = 1.24, which is a value that says that with each meter to the goal, the probability of scoring increases by 24 %.
- Similarly, the value for **TIGHTNESS** $e^{0,947}$ = 2.58 means that a player who can be more than one meter away from an opponent doubles the probability of scoring/a goal.

The equation for head shooting has the form:

Ln(goal chance) = 1.520 - 0.237 OIST - 3.117 ANGLE - 1.784 GAIN

and allows for the same interpretations as calculated probabilities for kicking.

Pollard, R., & Reep, C. (1997). Measuring the Effectiveness of Playing Strategies at Soccer. *Journal of the Royal Statistical Society.*, 46(4), 541–550.

III. Multivariate Regression

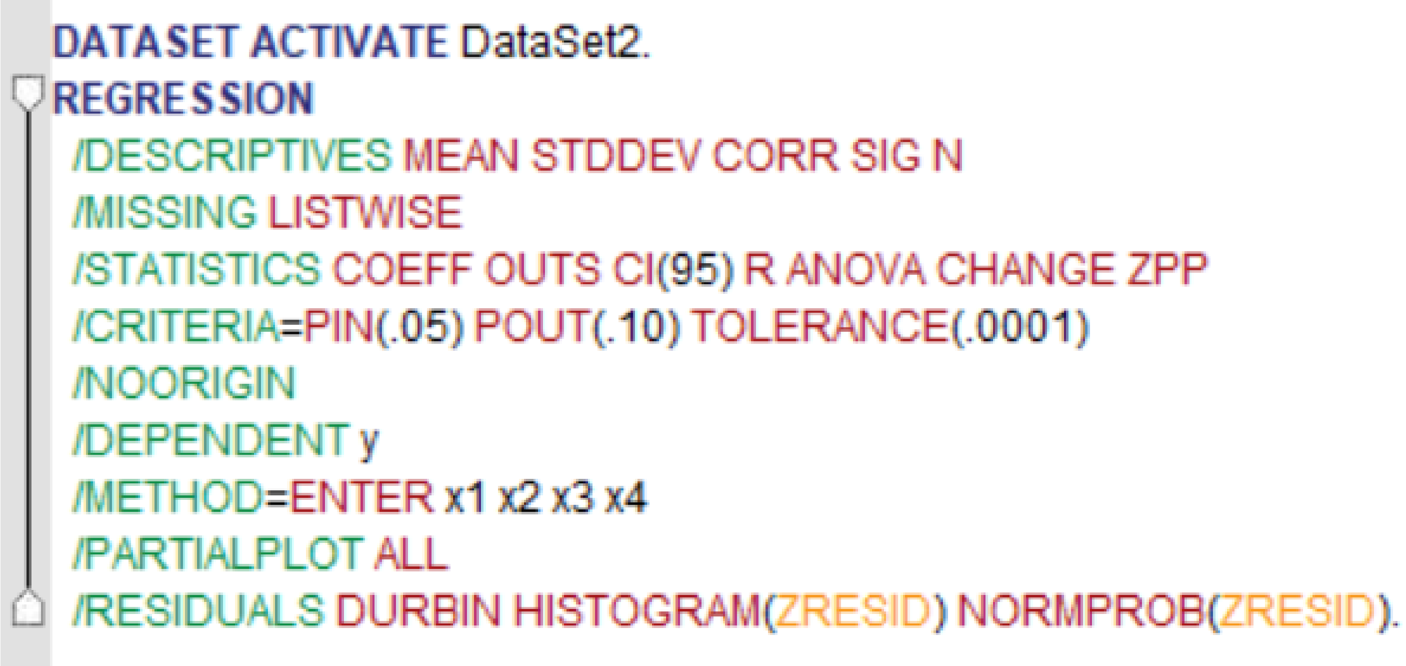
- Multivariate Regression is a method used to measure the degree at which more than one independent variable (predictors) and more than one dependent variable (responses), are linearly related.

Multivariate Regression - example

- For twenty selected households, data on **quarterly expenditure** on food and beverages (**y**), quarterly **household income (x1)**, **number of children (x2)**, **average age** of earning household members (**x3**) and **number of household members (x4)** were obtained.
- Decide which variables contribute significantly to explaining the variability in quarterly spending values.
- Try to guess which independent variables what will they be?

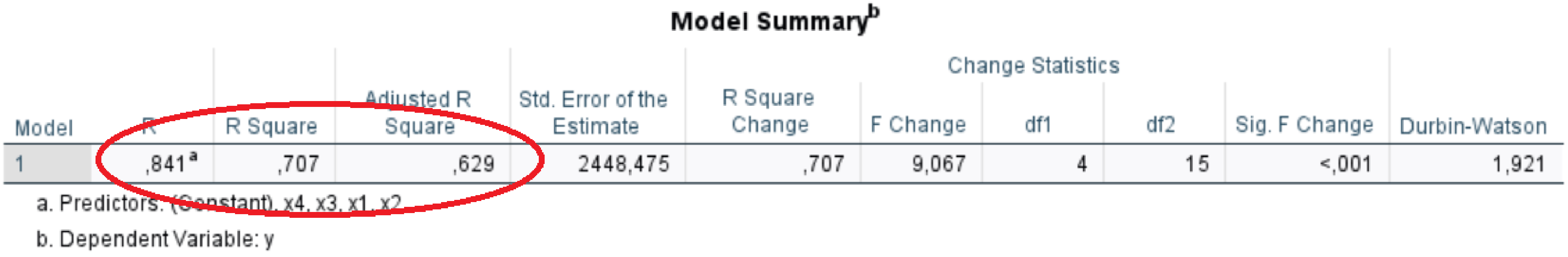
Multivariate Regression - example

x_1	x_2	x_3	x_4	y
11172	0	55	1	3464
8868	0	21	1	1982
17414	0	49	1	3228
10730	0	22	1	3034
24110	0	62,5	2	10146
38530	0	57	2	8202
22902	0	54,5	2	9332
25448	0	57,5	2	7096
20326	0	28	2	6248
39186	1	38,5	3	13816
28758	1	45,5	3	10328
33658	1	28,5	3	4786
24272	1	36	3	9710
30386	2	35	4	10778
31750	2	30,5	4	10568
39456	2	32,5	4	14260
48458	2	38	4	10934
37990	2	37	4	6388
24920	2	33,5	4	8584
40064	3	47	5	16950



```
DATASET ACTIVATE DataSet2.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10) TOLERANCE(.0001)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER x1 x2 x3 x4
  /PARTIALPLOT ALL
  /RESIDUALS DURBIN HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

Multivariate Regression – example

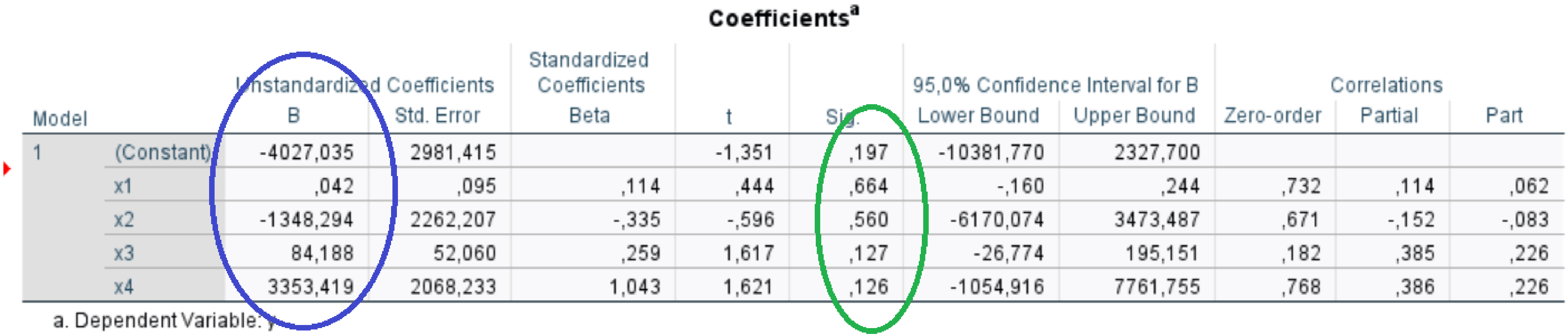


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,841[a]	,707	,629	2448,475	,707	9,067	4	15	<,001	1,921

a. Predictors: (Constant), x4, x3, x1, x2

b. Dependent Variable: y



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations		
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-4027,035	2981,415		-1,351	,197	-10381,770	2327,700			
	x1	,042	,095	,114	,444	,664	-,160	,244	,732	,114	,062
	x2	-1348,294	2262,207	-,335	-,596	,560	-6170,074	3473,487	,671	-,152	-,083
	x3	84,188	52,060	,259	1,617	,127	-26,774	195,151	,182	,385	,226
	x4	3353,419	2068,233	1,043	1,621	,126	-1054,916	7761,755	,768	,386	,226

a. Dependent Variable: y

$y = -4027 + 0.042063\ x1 - 1348.3\ x2 + 84.188\ x3 + 3353.4\ x4$

with the adjusted coefficient of determination, which takes into account the number of independent variables, $R^2 = 0.629$ and the residual standard deviation se = 2448.5.

IV. Nonlinear regression - Neural net

- assumptions: normality of data, homogenity of variances
  (→ parametric vs. nonparametric methods)
  nominal, ordinal, categorical variables cannot be combined in model

- often these conditions are not met

- Many inputs generate an output that is a nonlinear function of the weighted sum of these inputs.

- The weights assigned to each of the inputs are obtained on the basis of a learning process, where the generated outputs are compared with the so-called target outputs.

- The obtained deviations between the known values and the obtained outputs serve as feedback for the adjustment of the weights.

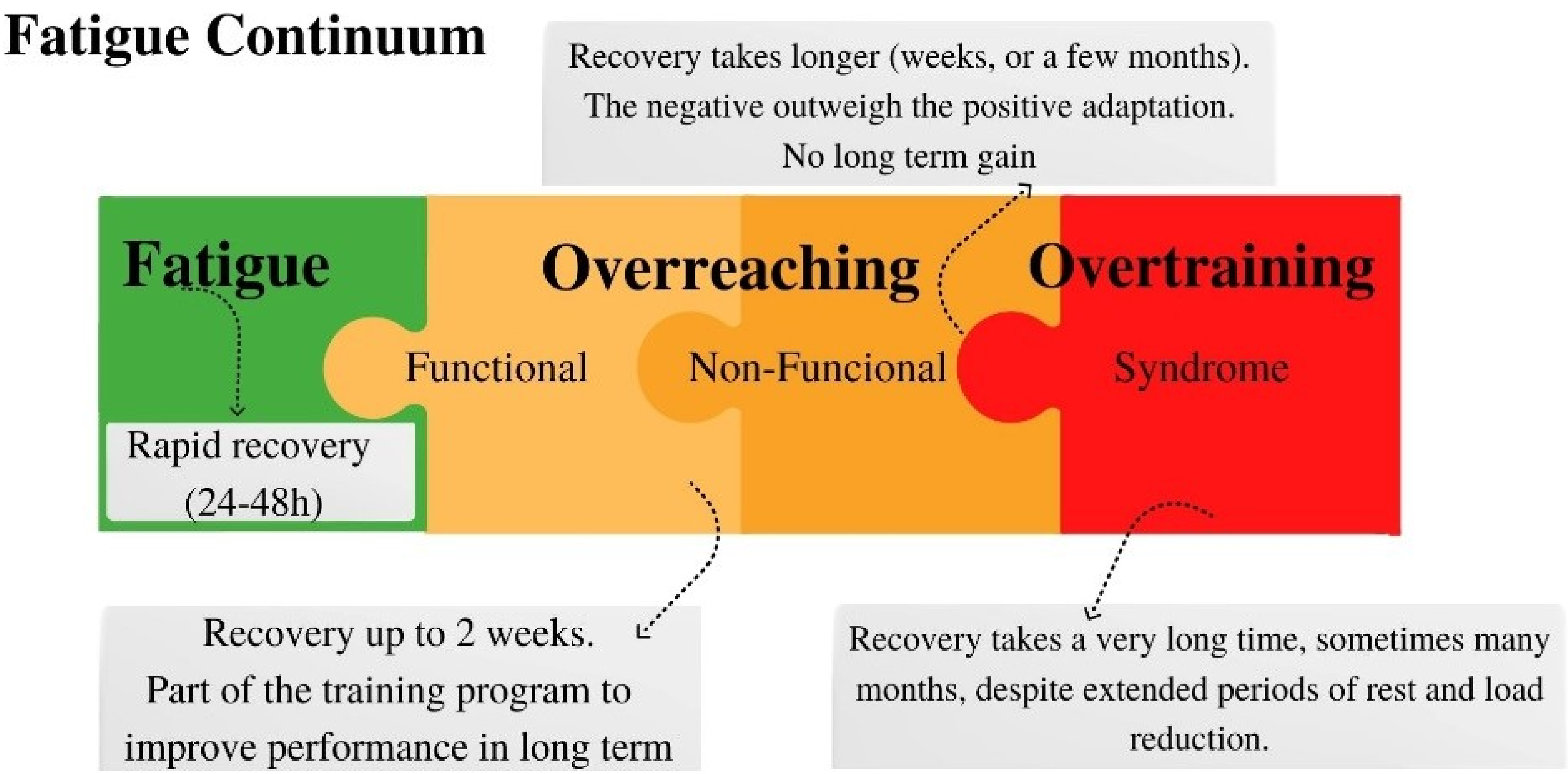
Nonlinear regression - Neural net

- NN is a method in artificial intelligence
- NN teaches computers to process data in a way that is inspired by the human brain.
- It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.
- In other words, it is a very complex regression, where I have one dependent and many independent

Nonlinear regression - Neural net

- Multilayer Perceptron (MLP): class of feed-forward neural networks
- 3 types of layers - the input layer, output layer and hidden layer
- Activation Functions: **defines how the weighted sum of the input is transformed into an output from a node or nodes in a layer of the network**

Example Overtraining



Fatigue Continuum

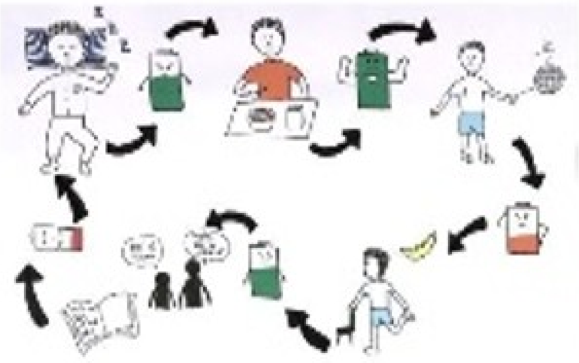
Recovery takes longer (weeks, or a few months).
The negative outweigh the positive adaptation.
No long term gain

Rapid recovery (24-48h)

Recovery up to 2 weeks.
Part of the training program to improve performance in long term

Recovery takes a very long time, sometimes many months, despite extended periods of rest and load reduction.

Bernacikova, M., Kumstat, M., Buresova, I., Kapounkova, K., Struhar, I., ☺ Sebera, M., & Paludo, A. C. (2022). **Preventing chronic fatigue in Czech young athletes: The features description of the “SmartTraining” mobile application**. *FRONTIERS IN PHYSIOLOGY*, *13*, 919982. https://doi.org/10.3389/fphys.2022.919982

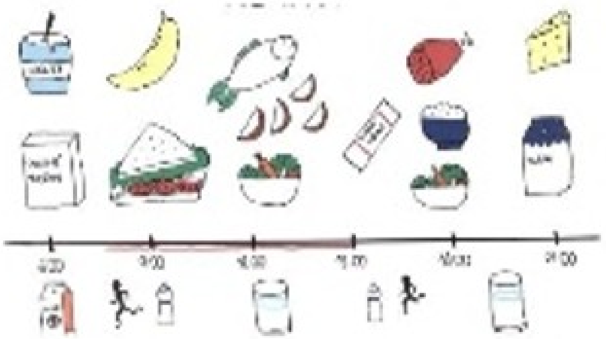


INTRODUCTION

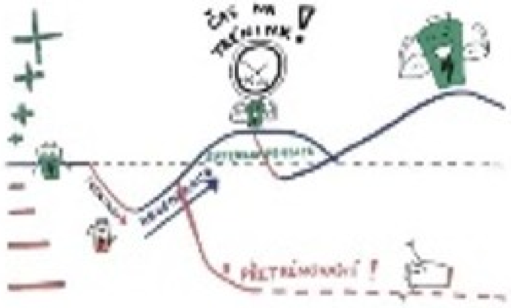
SPÁNKOVÝ RITUÁL

SLEEP

NUTRITION

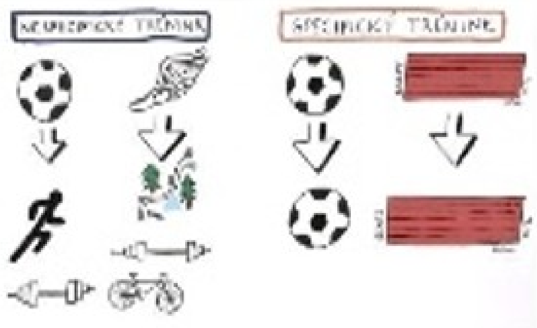


NUTRITION AND TRAINING

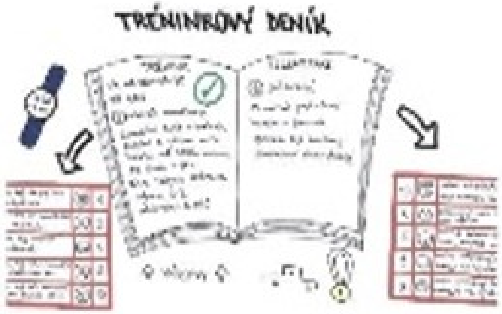


REGENERATION

COMMUNICATION



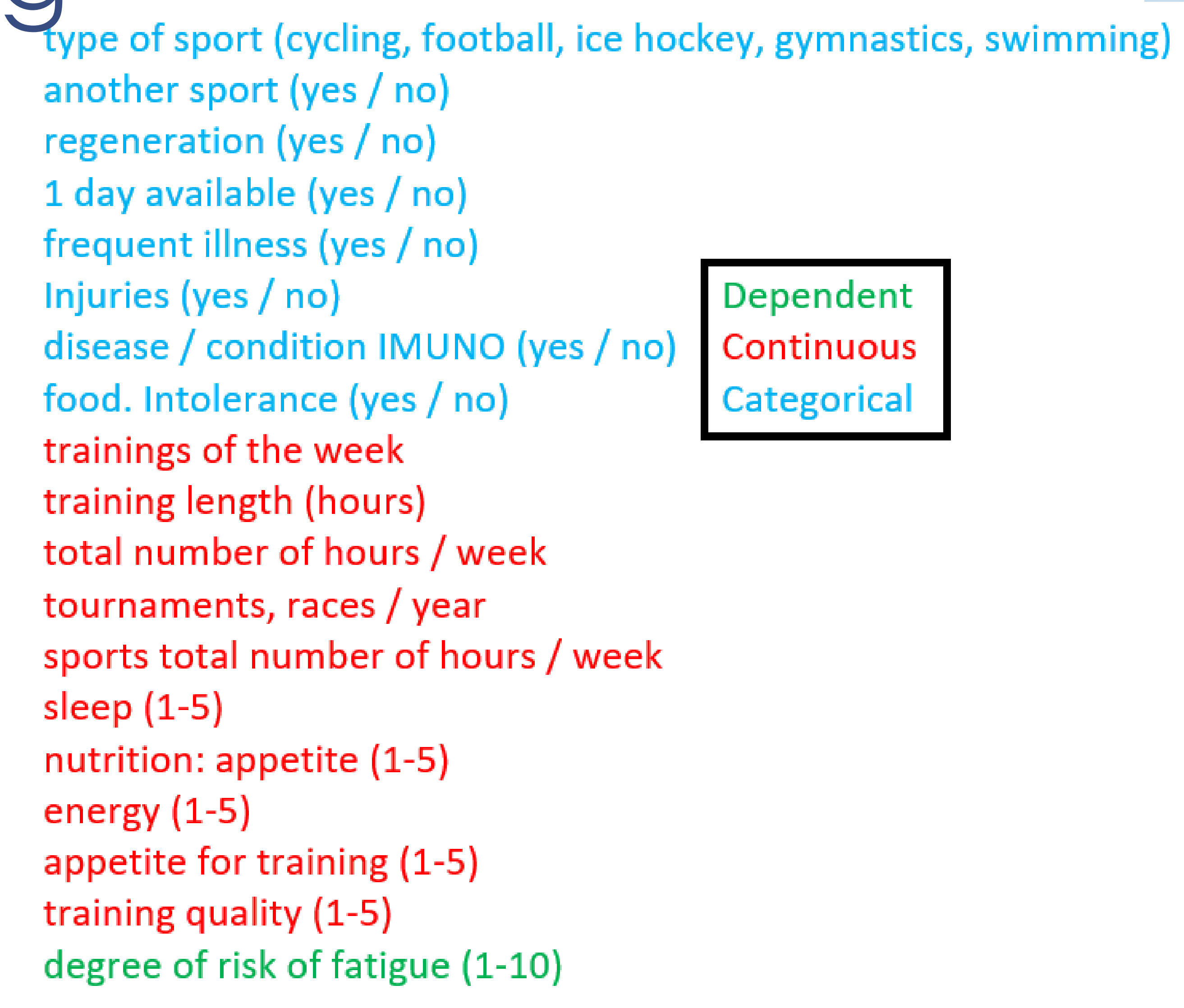
TRAINING PARAMETERS



MONITORING PROCESS

Example Overtraining

- How to get variables that are numerical, categorical, nominal ordinal into one regression model?
- The assumptions of data normality, homogeneity of variances, etc. are not met.



type of sport (cycling, football, ice hockey, gymnastics, swimming)
another sport (yes / no)
regeneration (yes / no)
1 day available (yes / no)
frequent illness (yes / no)
Injuries (yes / no)
disease / condition IMUNO (yes / no)
food. Intolerance (yes / no)
trainings of the week
training length (hours)
total number of hours / week
tournaments, races / year
sports total number of hours / week
sleep (1-5)
nutrition: appetite (1-5)
energy (1-5)
appetite for training (1-5)
training quality (1-5)
degree of risk of fatigue (1-10)

Dependent
Continuous
Categorical

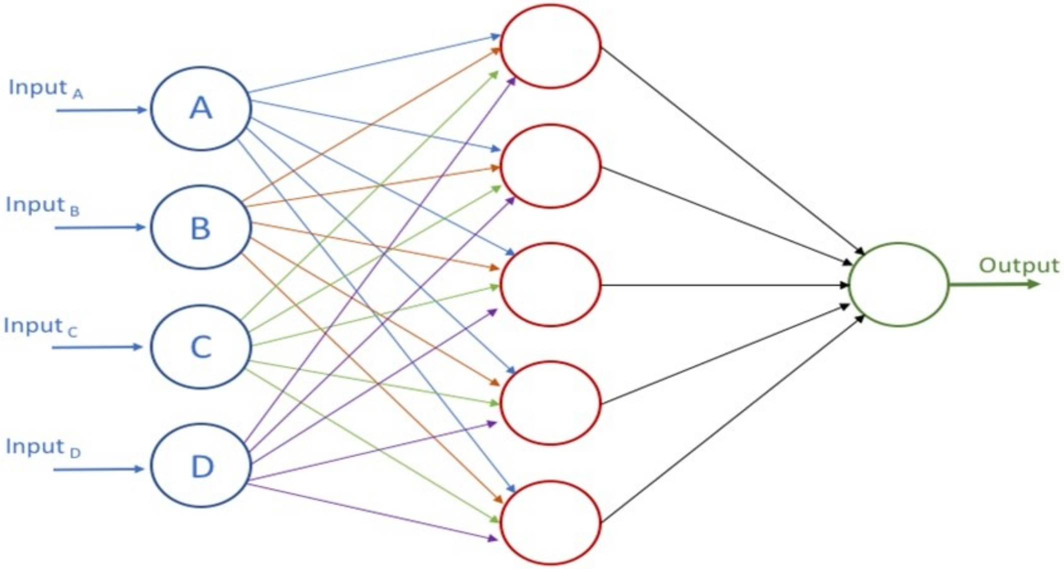
Example Overtraining

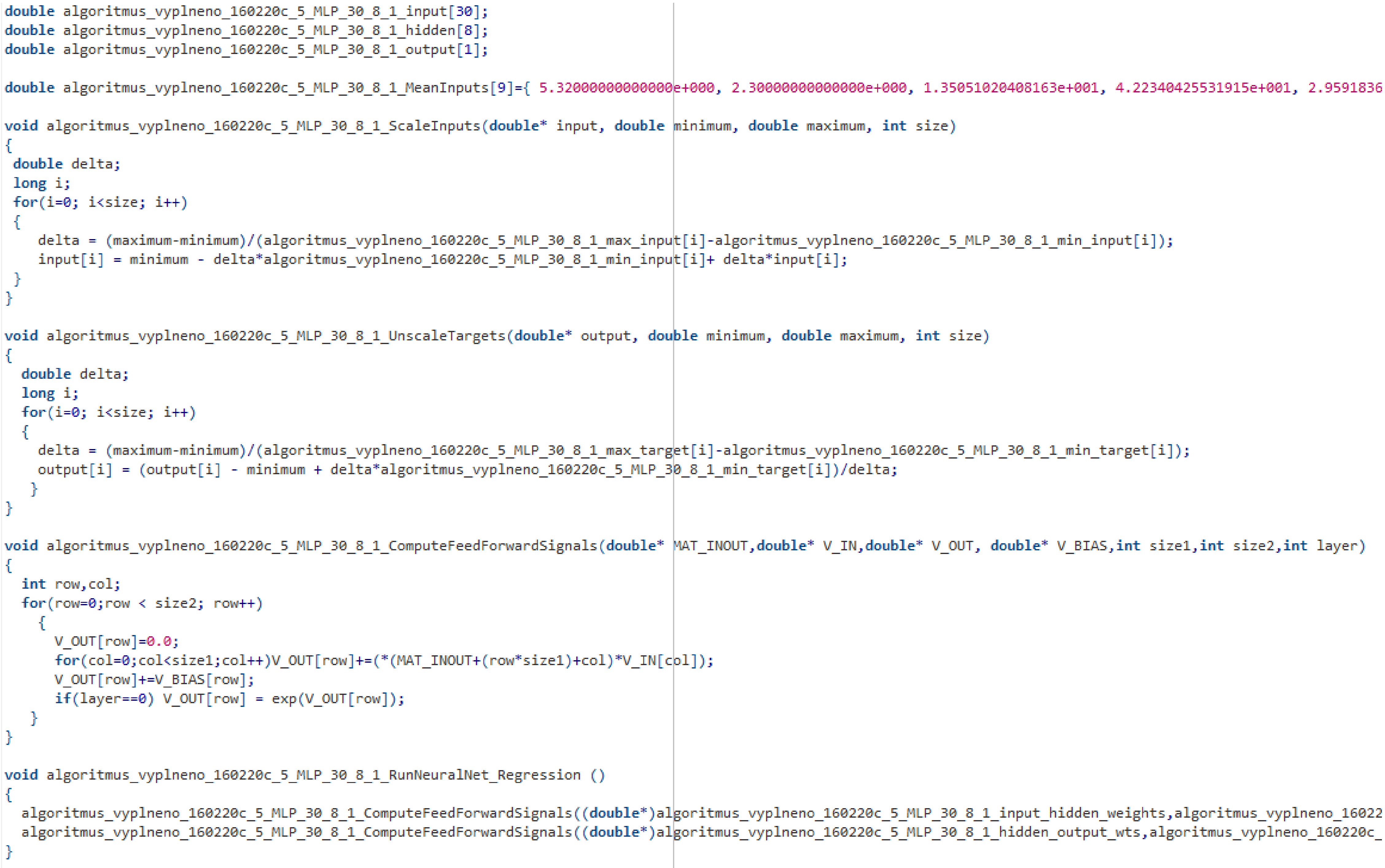
A neural network with 30 inputs, one hidden layer with 11 hidden neurons and 1 output neuron.

type of sport (cycling, football, ice hockey, gymnastics, swimming)
another sport (yes / no)
regeneration (yes / no)
1 day available (yes / no)
frequent illness (yes / no)
Injuries (yes / no)
disease / condition IMUNO (yes / no)
food. Intolerance (yes / no)
trainings of the week
training length (hours)
total number of hours / week
tournaments, races / year
sports total number of hours / week
sleep (1-5)
nutrition: appetite (1-5)
energy (1-5)
appetite for training (1-5)
training quality (1-5)
degree of risk of fatigue (1-10)

Dependent
Continuous
Categorical

MLP 30-11-1





```c
double algoritmus_vyplneno_160220c_5_MLP_30_8_1_input[30];
double algoritmus_vyplneno_160220c_5_MLP_30_8_1_hidden[8];
double algoritmus_vyplneno_160220c_5_MLP_30_8_1_output[1];

double algoritmus_vyplneno_160220c_5_MLP_30_8_1_MeanInputs[9]={ 5.32000000000000e+000, 2.30000000000000e+000, 1.35051020408163e+001, 4.22340425531915e+001, 2.9591836

void algoritmus_vyplneno_160220c_5_MLP_30_8_1_ScaleInputs(double* input, double minimum, double maximum, int size)
{
 double delta;
 long i;
 for(i=0; i<size; i++)
 {
    delta = (maximum-minimum)/(algoritmus_vyplneno_160220c_5_MLP_30_8_1_max_input[i]-algoritmus_vyplneno_160220c_5_MLP_30_8_1_min_input[i]);
    input[i] = minimum - delta*algoritmus_vyplneno_160220c_5_MLP_30_8_1_min_input[i]+ delta*input[i];
 }
}

void algoritmus_vyplneno_160220c_5_MLP_30_8_1_UnscaleTargets(double* output, double minimum, double maximum, int size)
{
  double delta;
  long i;
  for(i=0; i<size; i++)
  {
    delta = (maximum-minimum)/(algoritmus_vyplneno_160220c_5_MLP_30_8_1_max_target[i]-algoritmus_vyplneno_160220c_5_MLP_30_8_1_min_target[i]);
    output[i] = (output[i] - minimum + delta*algoritmus_vyplneno_160220c_5_MLP_30_8_1_min_target[i])/delta;
   }
}

void algoritmus_vyplneno_160220c_5_MLP_30_8_1_ComputeFeedForwardSignals(double* MAT_INOUT,double* V_IN,double* V_OUT, double* V_BIAS,int size1,int size2,int layer)
{
  int row,col;
  for(row=0;row < size2; row++)
    {
      V_OUT[row]=0.0;
      for(col=0;col<size1;col++)V_OUT[row]+=(*(MAT_INOUT+(row*size1)+col)*V_IN[col]);
      V_OUT[row]+=V_BIAS[row];
      if(layer==0) V_OUT[row] = exp(V_OUT[row]);
   }
}

void algoritmus_vyplneno_160220c_5_MLP_30_8_1_RunNeuralNet_Regression ()
{
  algoritmus_vyplneno_160220c_5_MLP_30_8_1_ComputeFeedForwardSignals((double*)algoritmus_vyplneno_160220c_5_MLP_30_8_1_input_hidden_weights,algoritmus_vyplneno_16022
  algoritmus_vyplneno_160220c_5_MLP_30_8_1_ComputeFeedForwardSignals((double*)algoritmus_vyplneno_160220c_5_MLP_30_8_1_hidden_output_wts,algoritmus_vyplneno_160220c_
}
```

The most important predictors of overtraining

- Amount of regeneration (active regeneration)
- Sleep (pasive regeneration)
- Number of tournaments/races per year
- Type of sport
- …

CONCLUSION

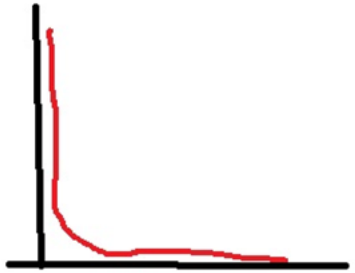
And if you've read this far, I have a final message in the form of a math quiz. Complete the text "All you need is: ....."

Draw 4 graphs

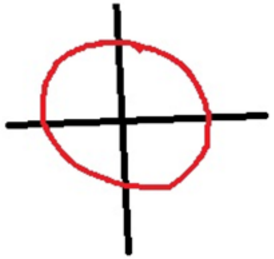
- $y = \frac{1}{x}$; x is in the interval (0; 5).
- $x^2 + y^2 = 9$
- $y = abs\,(-2x)$; x is in the interval <(-5; 5)
- $-x = 3\,abs(sin(y))$; y is in the interval $(-\pi; \pi)$

All you need is:

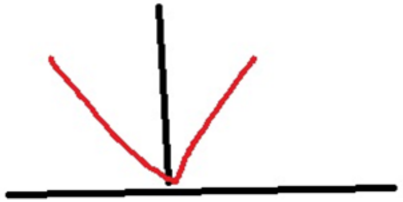
$y=\frac{1}{x}$



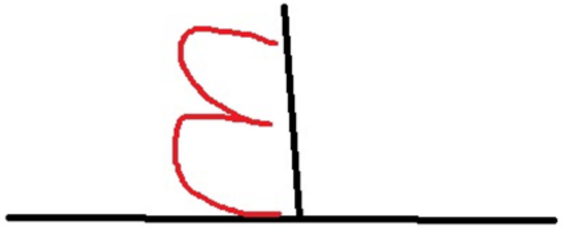
$x^2+y^2=9$



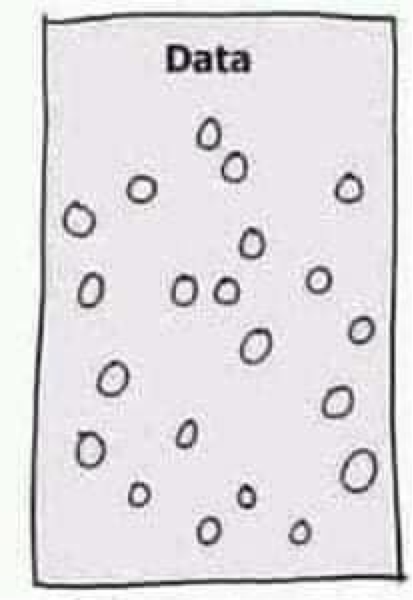
$Y=|-2x|$



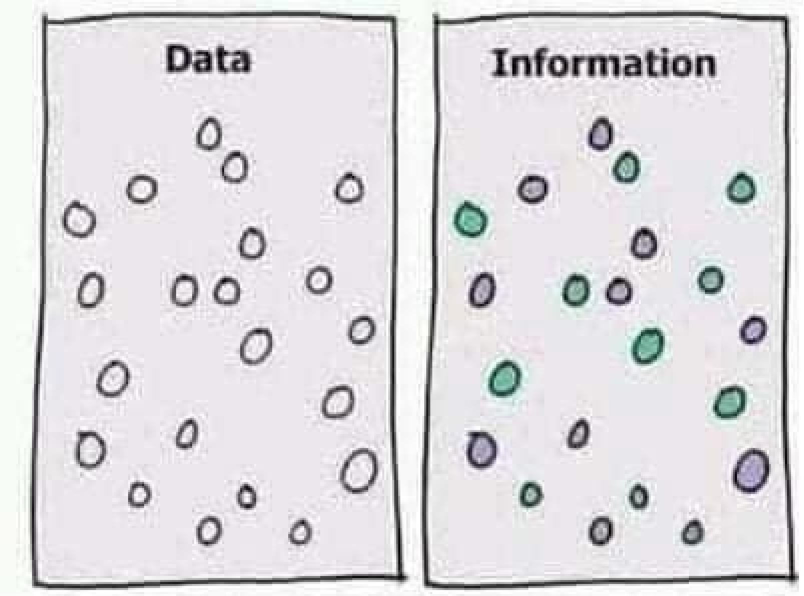
$-x=3|\sin y|$



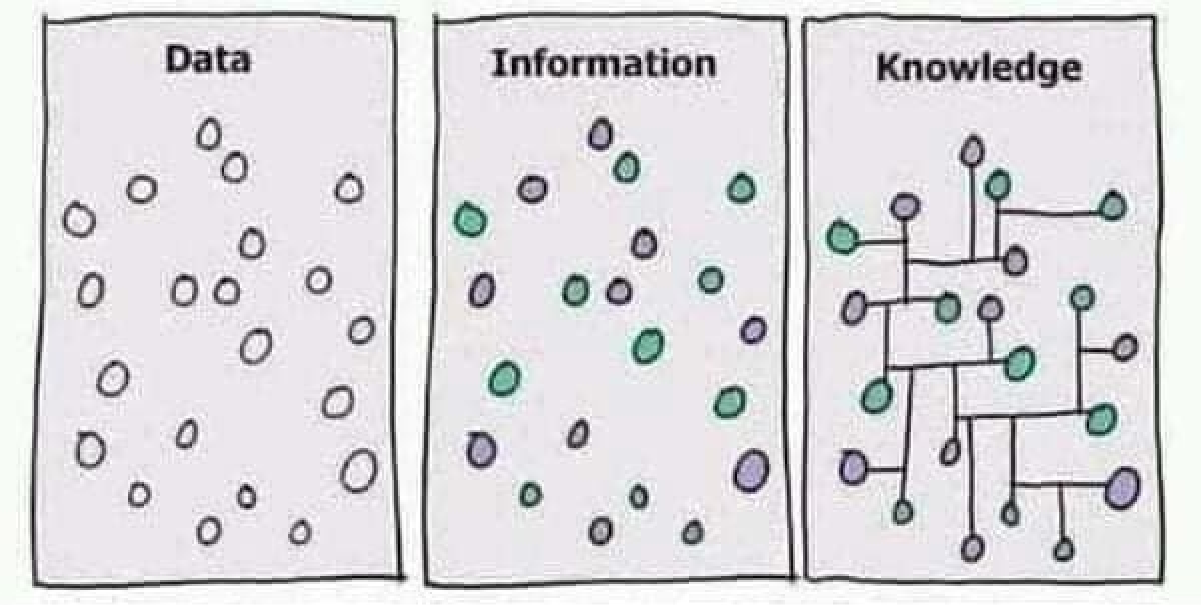
How to fight disinformation and conspiracy?



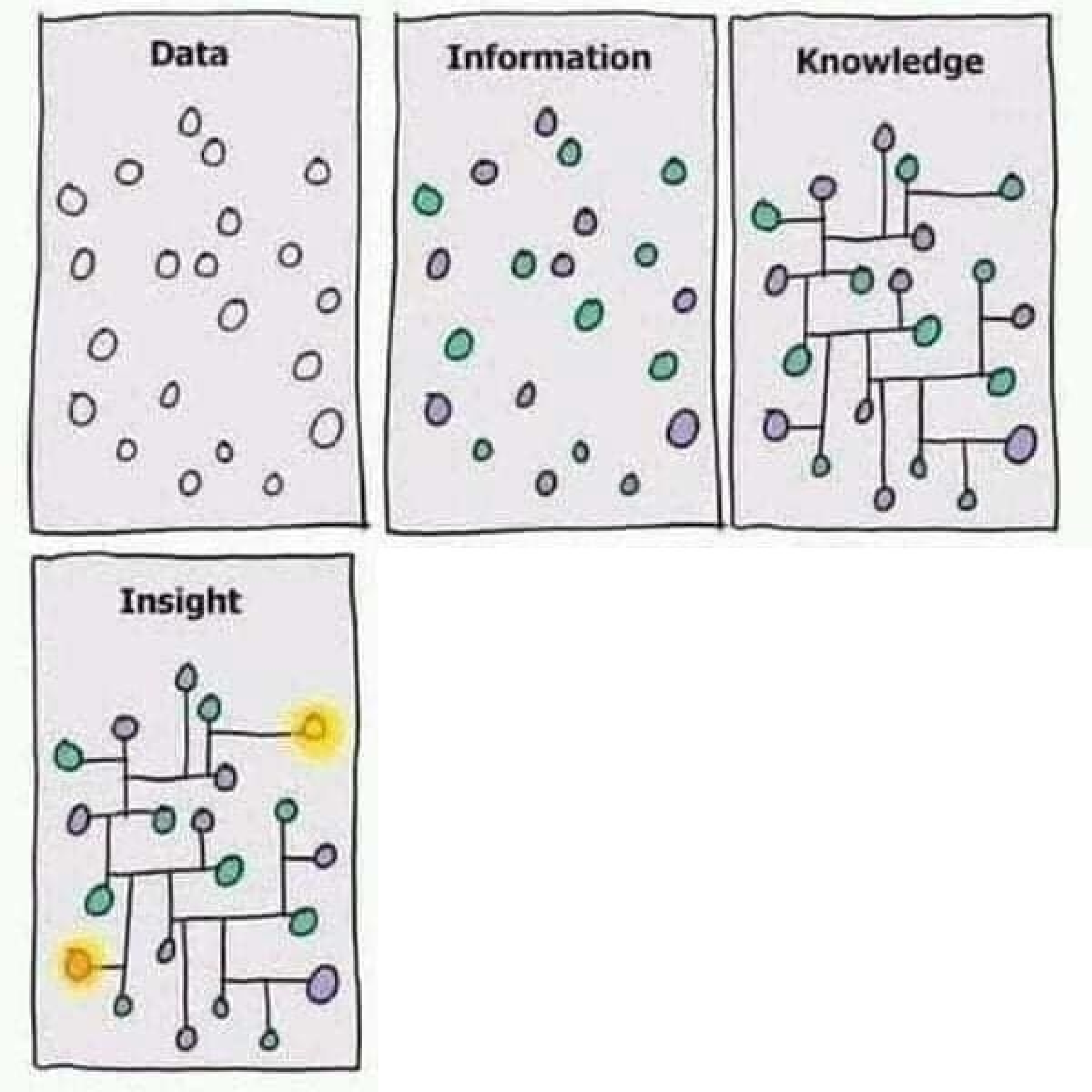
How to fight disinformation and conspiracy?



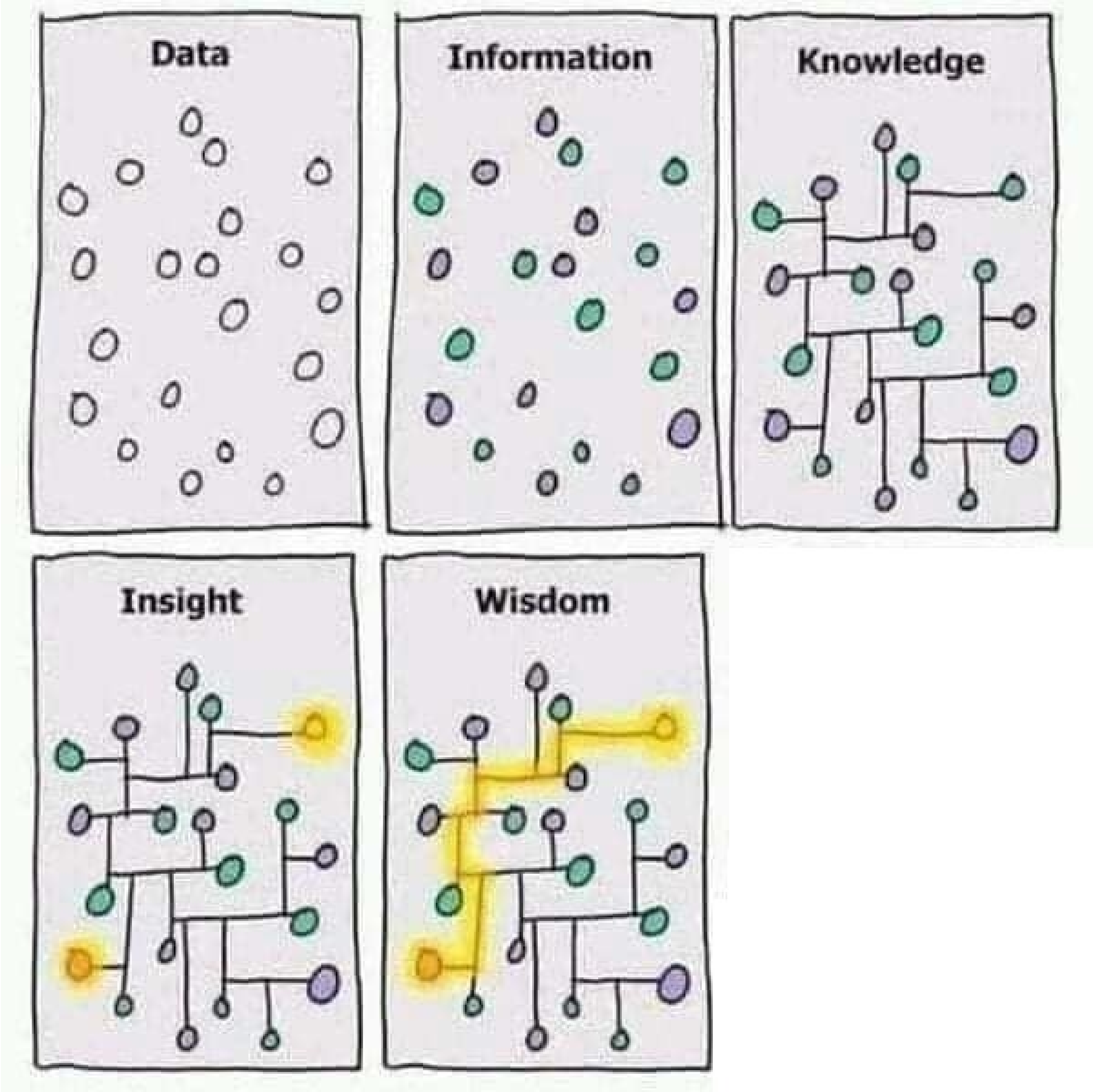
How to fight disinformation and conspiracy?



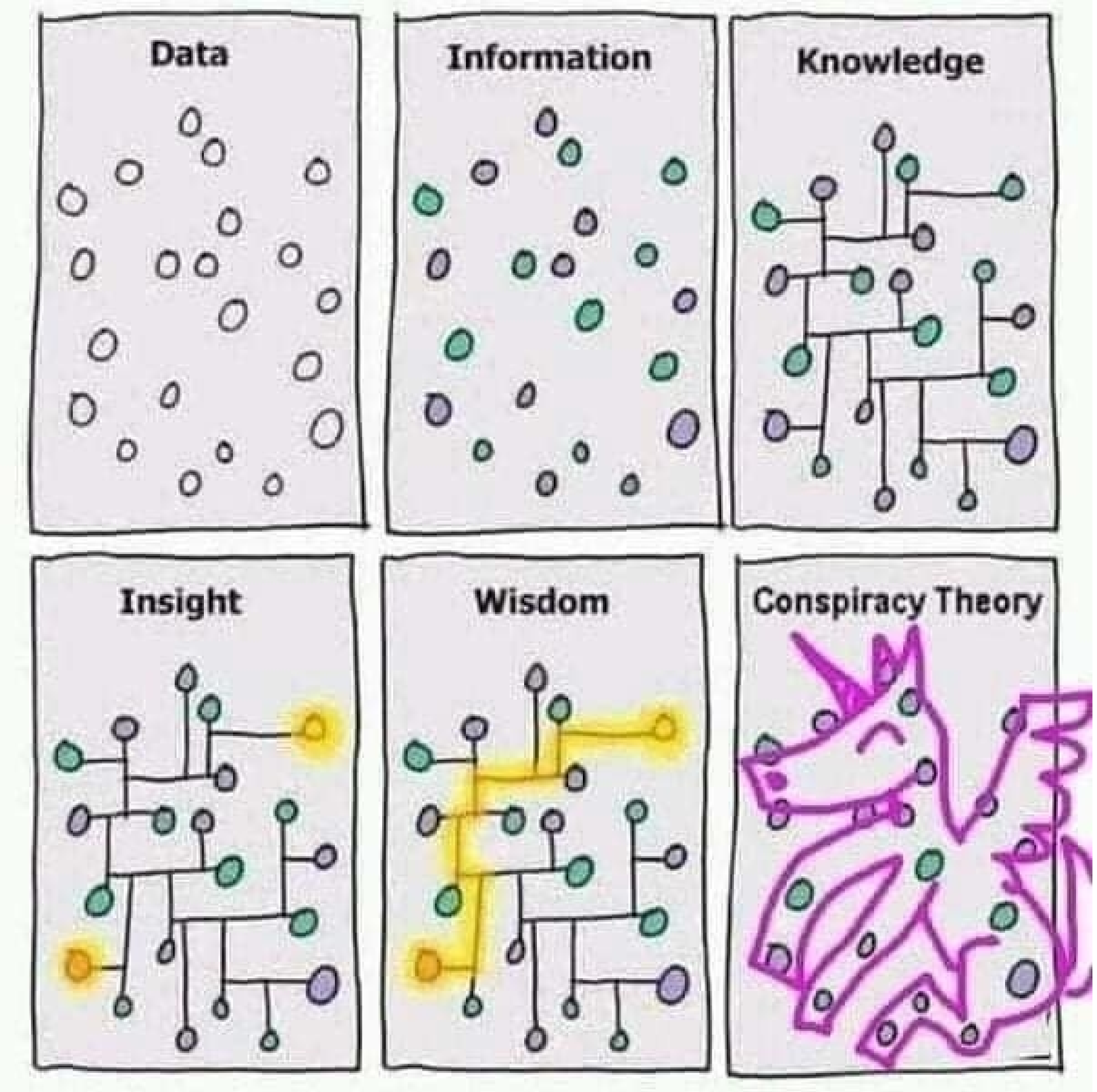
How to fight disinformation and conspiracy?



How to fight disinformation and conspiracy?



How to fight disinformation and conspiracy?



How to fight disinformation and conspiracy?

Verifying Claims and Sources:

- This involves analyzing data sets that can confirm or disprove certain claims.

Recognition and detection of data manipulation:

- Statistics offers tools for identifying unusual or improbable patterns in data, which may signal an attempt at disinformation.

Use of predictive regression models:

- Statistical modeling and machine learning can help predict the spread of misinformation and identify potential new misinformation before it spreads.

Statistics Education Initiative:

- Which teach the public how to interpret data and statistical results. This helps people better understand how data is used to support different arguments.

Promoting data transparency and openness

- so that the public can verify information and conduct independent analysis.

And that's why you need to understand statistics!

1. Dezinformace o vakcínách:
Tvrzení: "Úmrtnost na očkování je mnohem vyšší než na samotný COVID-19."
• Statistická metoda: Porovnání incidence nežádoucích účinků po očkování s incidencí a mortalitou COVID-19.
• Analýza: Použití metod k výpočtu relativního rizika, odhadování intervalů spolehlivosti a kontrola zkreslení (např. nereprezentativní výběr dat).

2. Dezinformace o imigraci:
Tvrzení: "Imigranti způsobují rapidní nárůst kriminality v ČR."
• Statistická metoda: Analýza dat z kriminálních statistik, porovnání kriminality mezi migranty a místní populací po zohlednění klíčových faktorů (věk, příjem, vzdělání).
• Analýza: Použití regresní analýzy ke kontrole vlivu různých faktorů a testování statistické významnosti.

3. Dezinformace o změně klimatu:
Tvrzení: "Globální oteplování je jen přirozený cyklus a není způsobeno lidskou činností."
• Statistická metoda: Analýza dlouhodobých teplotních trendů a jejich korelace s emisemi skleníkových plynů.
• Analýza: Použití modelování časových řad a kauzálních analýz k potvrzení vztahu mezi emisemi a teplotou.

4. Dezinformace o volebních podvodech:
Tvrzení: "Ve volbách byly miliony hlasů zmanipulovány."
• Statistická metoda: Ověření volebních výsledků pomocí forenzní analýzy (analýza anomálií v distribuci hlasů).
• Analýza: Testování pravděpodobnosti vzorů, které by naznačovaly manipulaci, oproti přirozeně očekávaným trendům.

5. Dezinformace o ekonomice:
Tvrzení: "Inflace je způsobena pouze vládními výdaji."
• Statistická metoda: Analýza vlivu jednotlivých faktorů na inflaci pomocí vícerozměrné regresní analýzy (vládní výdaje, ceny energií, globální ekonomické faktory).
• Analýza: Vyvrácení zjednodušení a propojení vícero zdrojů dat.