QSPR II

Pokročilá chemoinformatika

PARAMETRIZACE

Parametrizace neboli učení modelu

- lineární, logistická, zobecněná regrese

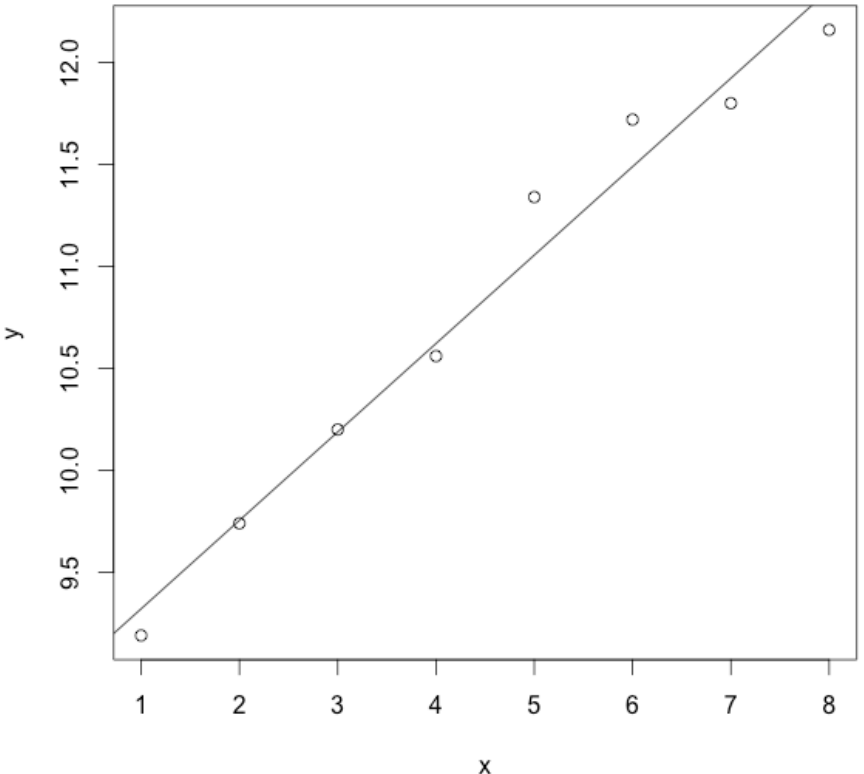
- MLR (Multiple Linear Regression)
- KNN (K-Nearest Neighbors)
- Decision Tree
- ASNN (ASsociative Neural Networks)
- Naive Bayes

Lineární regrese

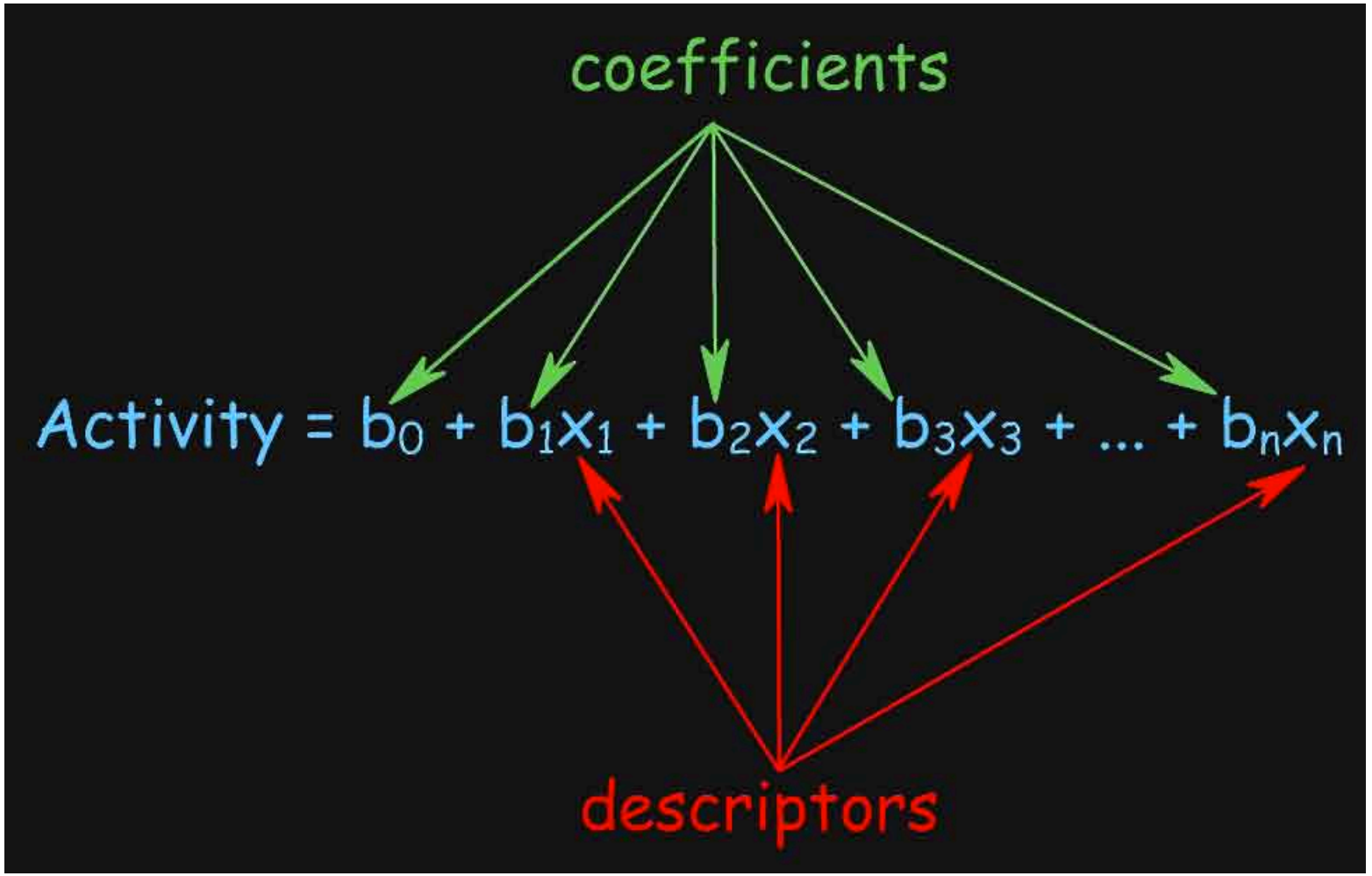
- proměnné **x** a **y**, hledáme takové *a* a *b*, které nejlépe popiše vzájemný lineární vztah

$$y = ax + b$$

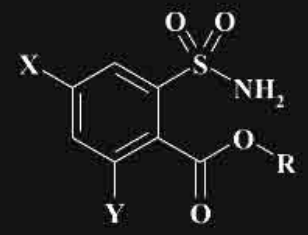
- Používáme metodu nejmenších čtverců pro minimalizaci výsledné sumy vzdálenosti bodů od přímky



MLR - princip multilineární regrese



MLR - důvod multilineární regrese



bad model

good model

model	r
log 1/C = 0.009 Es + 3.411	0.03
log 1/C = -0.626 σ + 3.314	0.27
log 1/C = -0.078 logP + 3.432	0.38
log 1/C = -0.210 logP - 2.214 σ + 3.154	0.80
log 1/C = 0.21 Es - 0.238 logP - 3.81 σ + 3.046	0.95

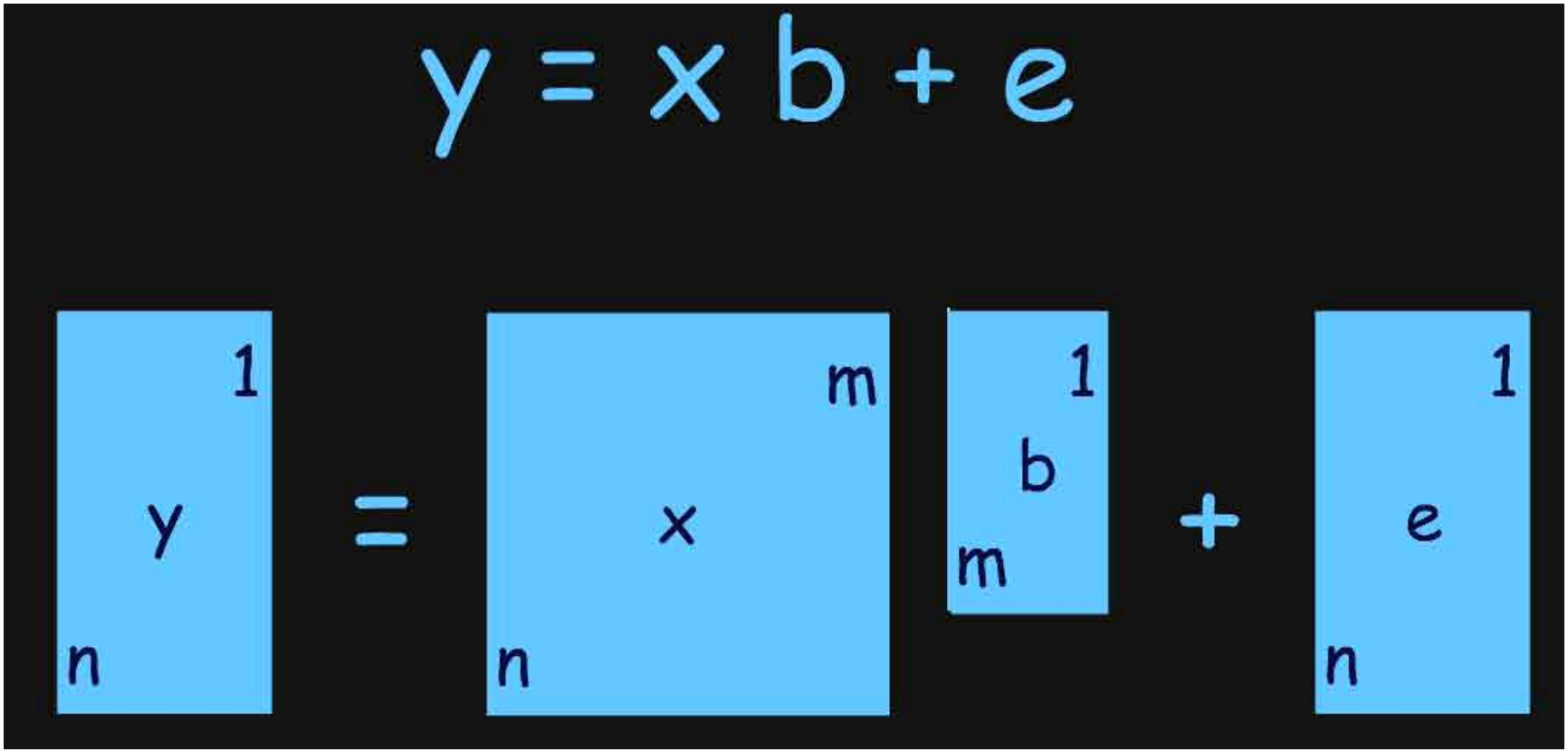
MLR – maticový zápis multilineární regrese

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_mx_m + e$$

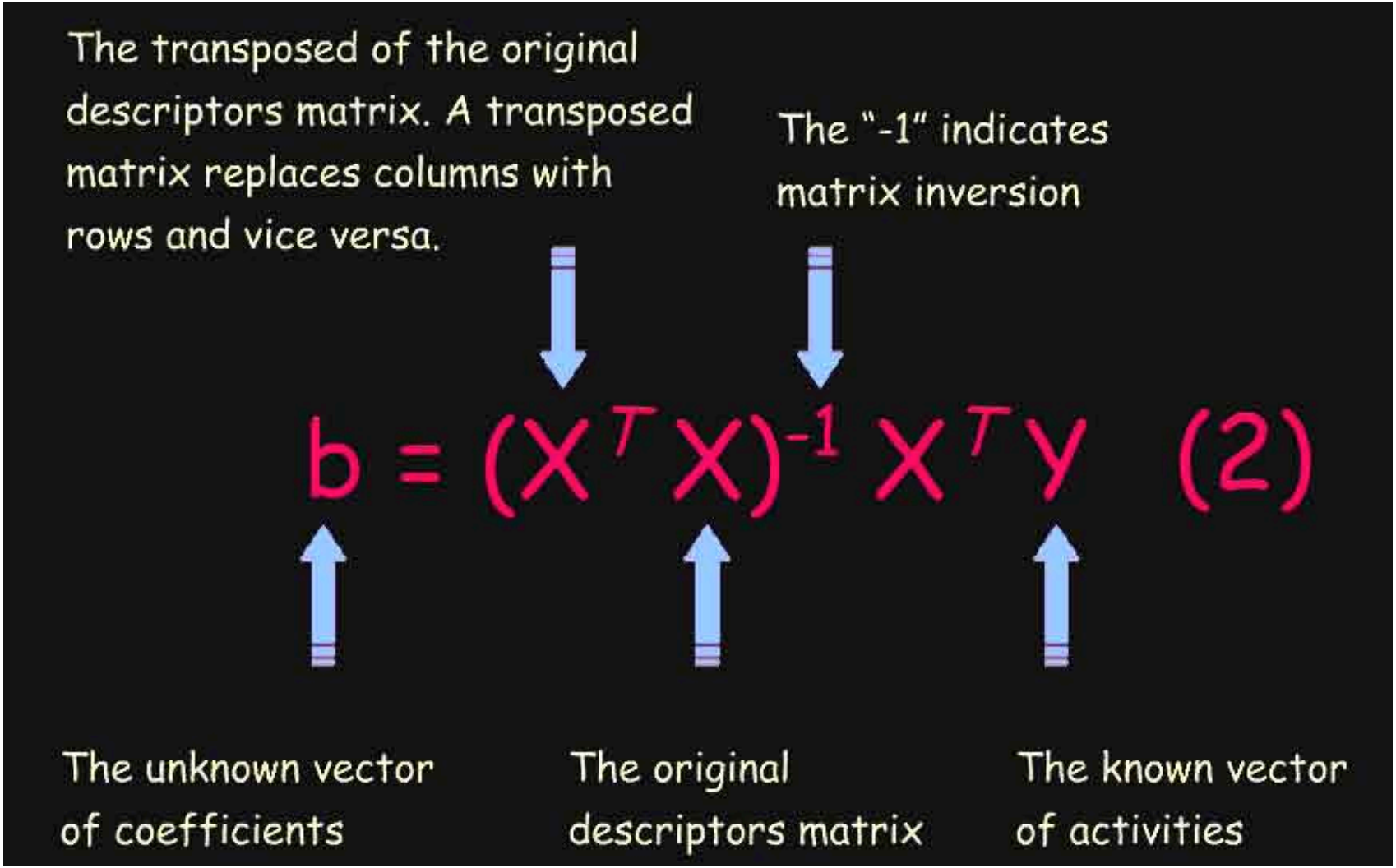
$$y = \sum_{j=1}^{m} b_jx_j + e$$

Matrix notation: $y = x^T b + e$

MLR – hledáme b I
multilineární regrese

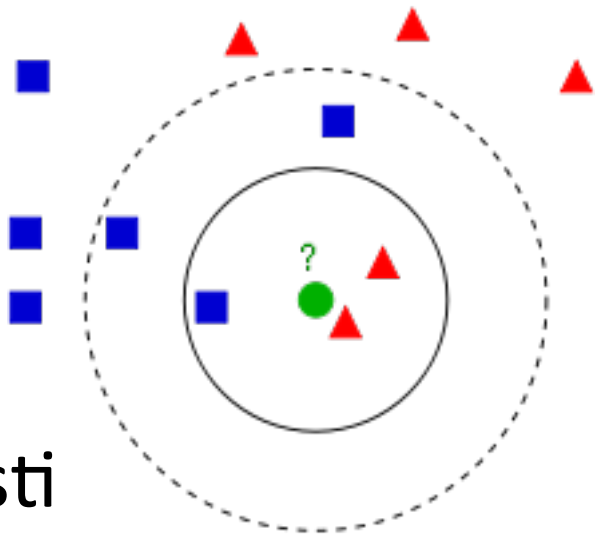


MLR – hledáme b II
multilineární regrese



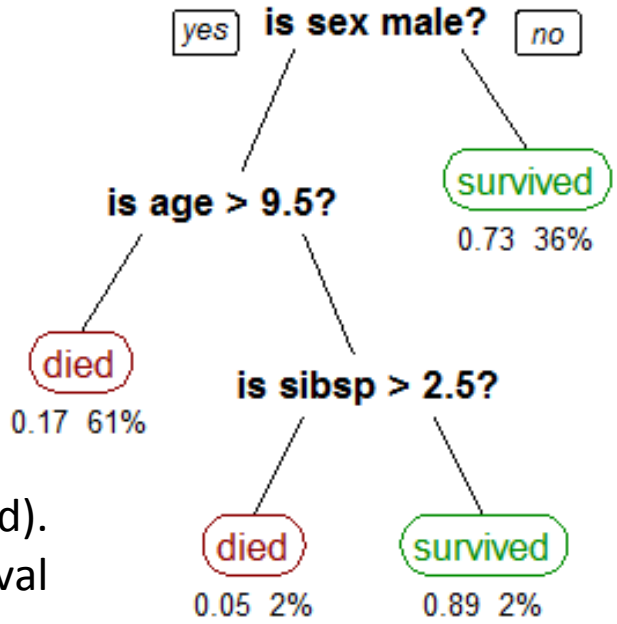
KNN (K-Nearest Neighbors)
Algoritmus k-nejbližších sousedů

- patří mezi nejjednodušší "machine lerning" algoritmy
- může sloužit ke klasifikaci nebo predikci
- pokud budeme klasifikovat zelené kolečko podle 3-KK (bereme v úvahu 3 sousedy) bude patřit do trojúhelníků, v 5-KK do čtverců
- v případě regrese se hodnota bude počítat na základě nejbližších sousedů a vzdálenosti



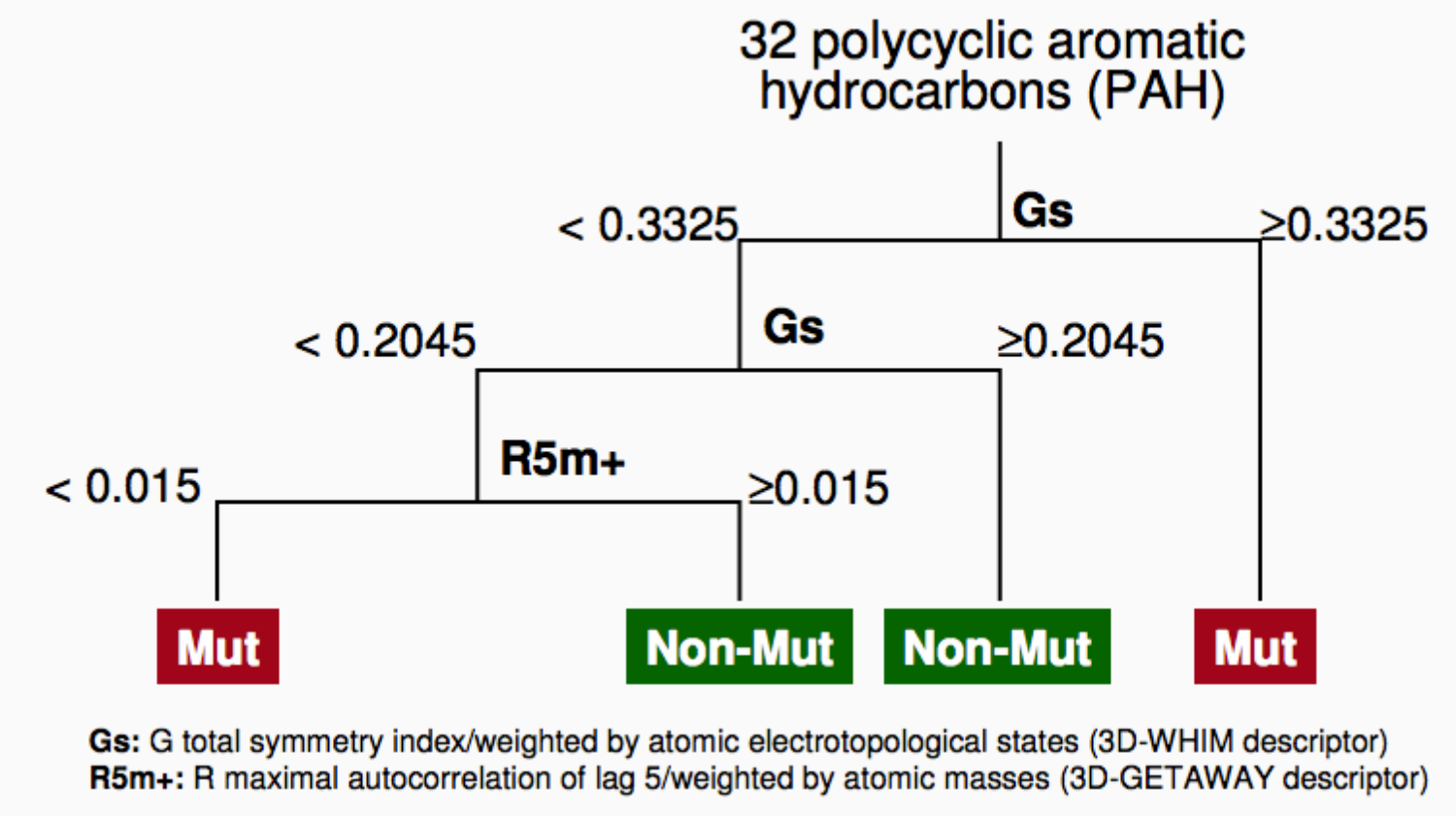
Decision Tree
Rozhodovací stromy

- může sloužit ke klasifikaci (classification trees) nebo predikci (regression trees)
- V těchto stromových strukturách představují listy (leafs) třídy a větve představují spojky mezi třídami.



Zdroj: wiki | A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

Decission tree – příklad predikce mutagenity

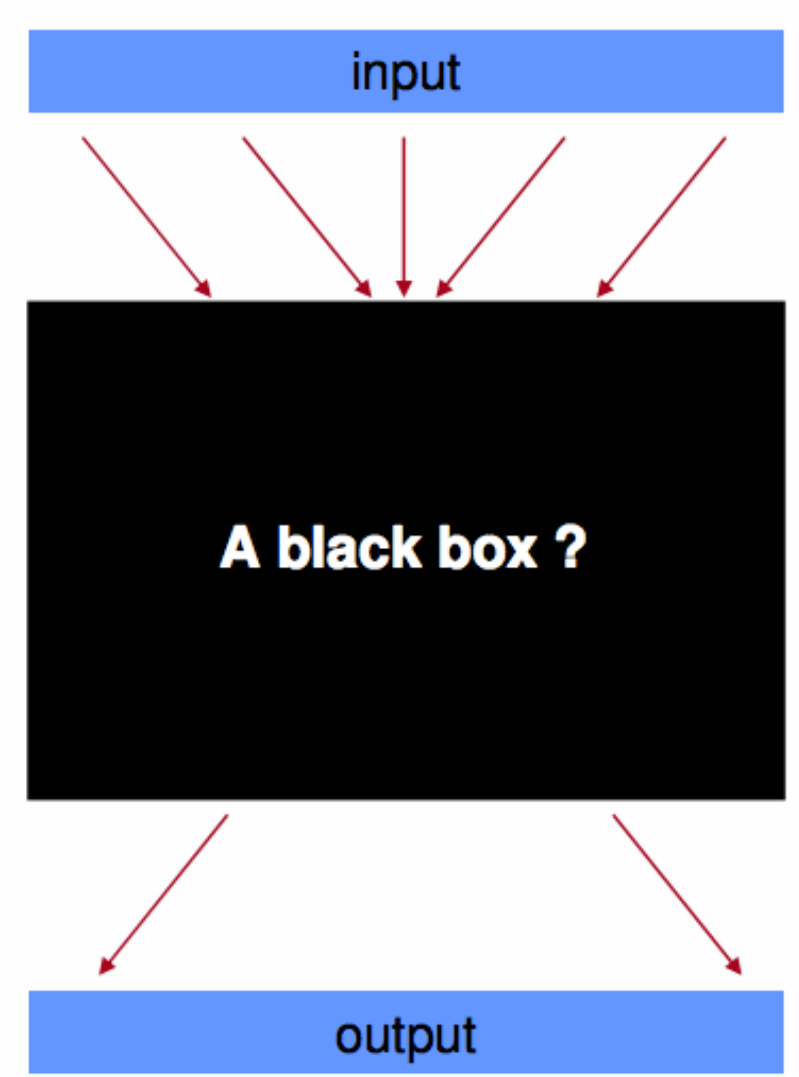


Zdroj: P. Gramatica, E. Papa, A. Marrocchi, L. Minuti, A. Taticchi, Ecotoxicology and Environmental Safety 2007, 66(3), 353-361.

ASNN (ASsociative Neural Networks)

- Statistické metody používají informace a “učení”.
- Mozek ale nepotřebuje žádné statistické metody pro učení.
- Neuronové sítě simulují nervový systém za použití algoritmů a matematických modelů.

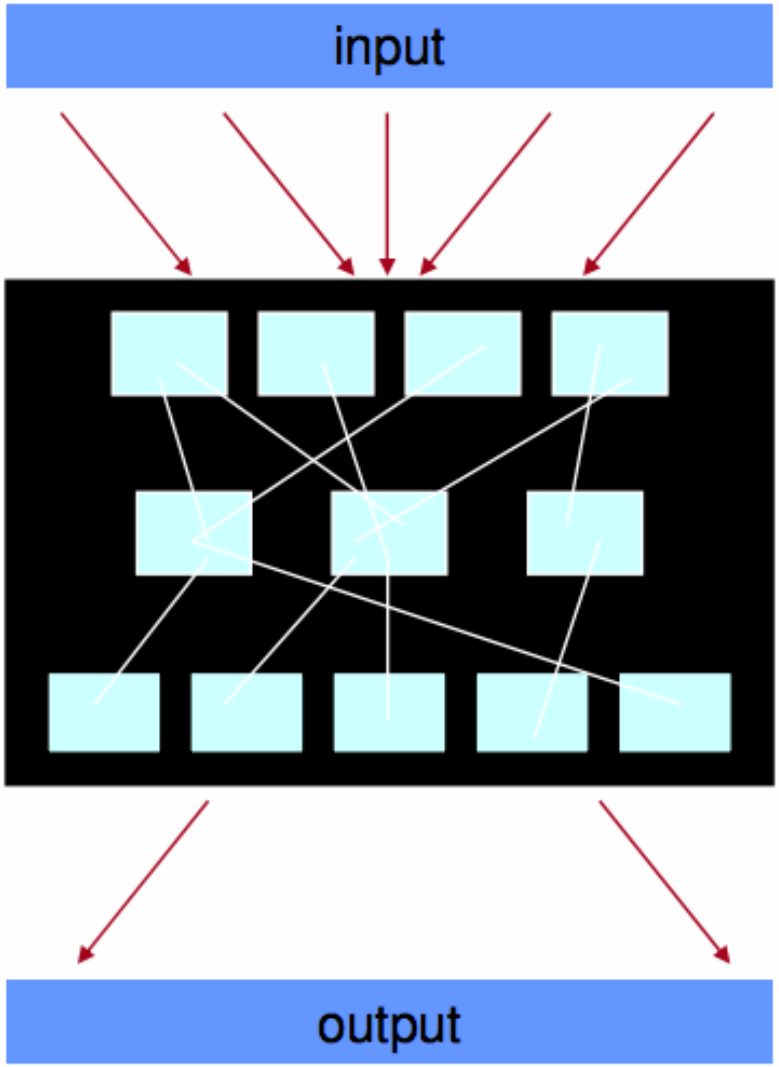
NN – black box?



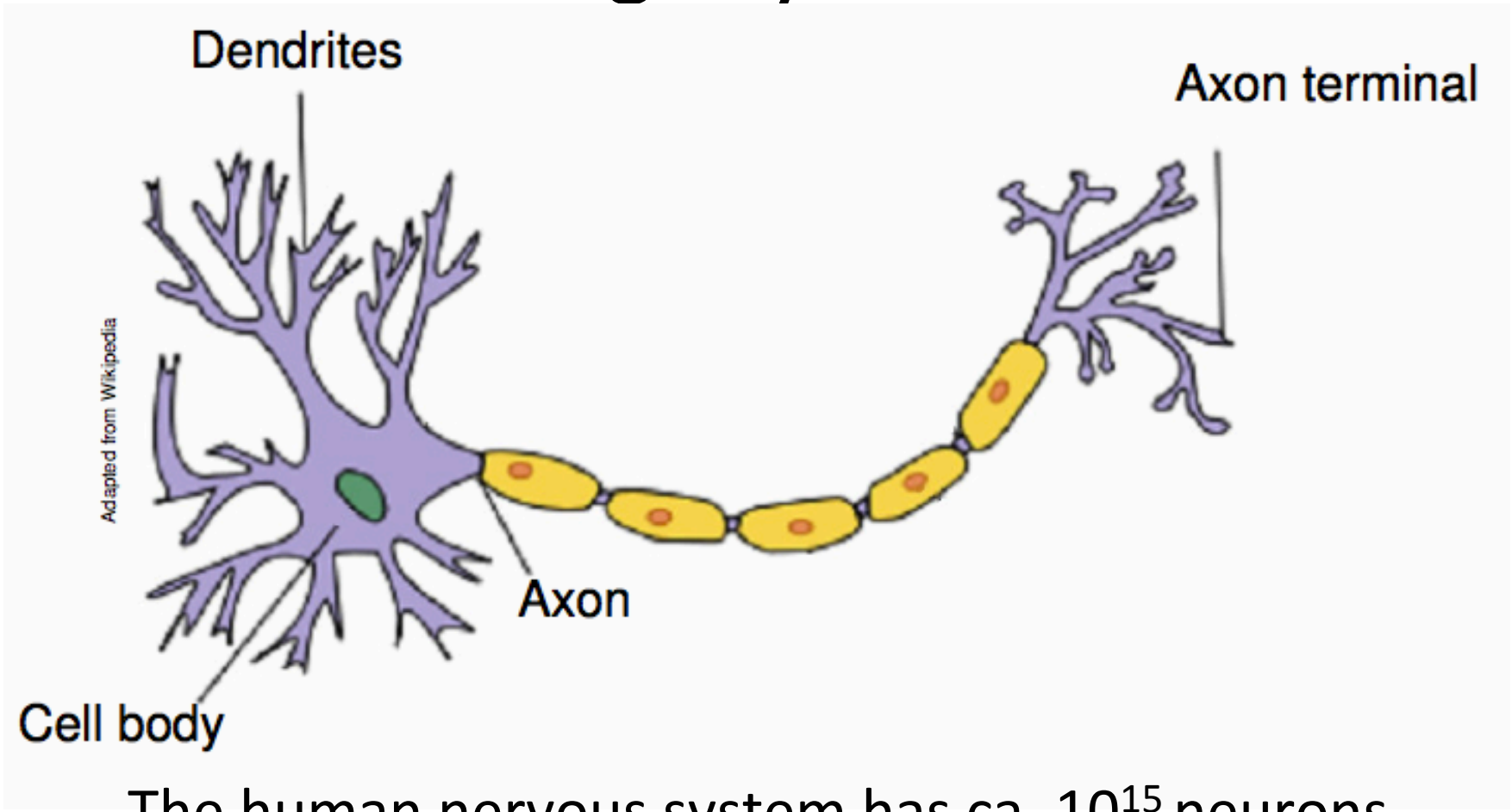
NN – black box? NE!

spojené
funkční
jednotky

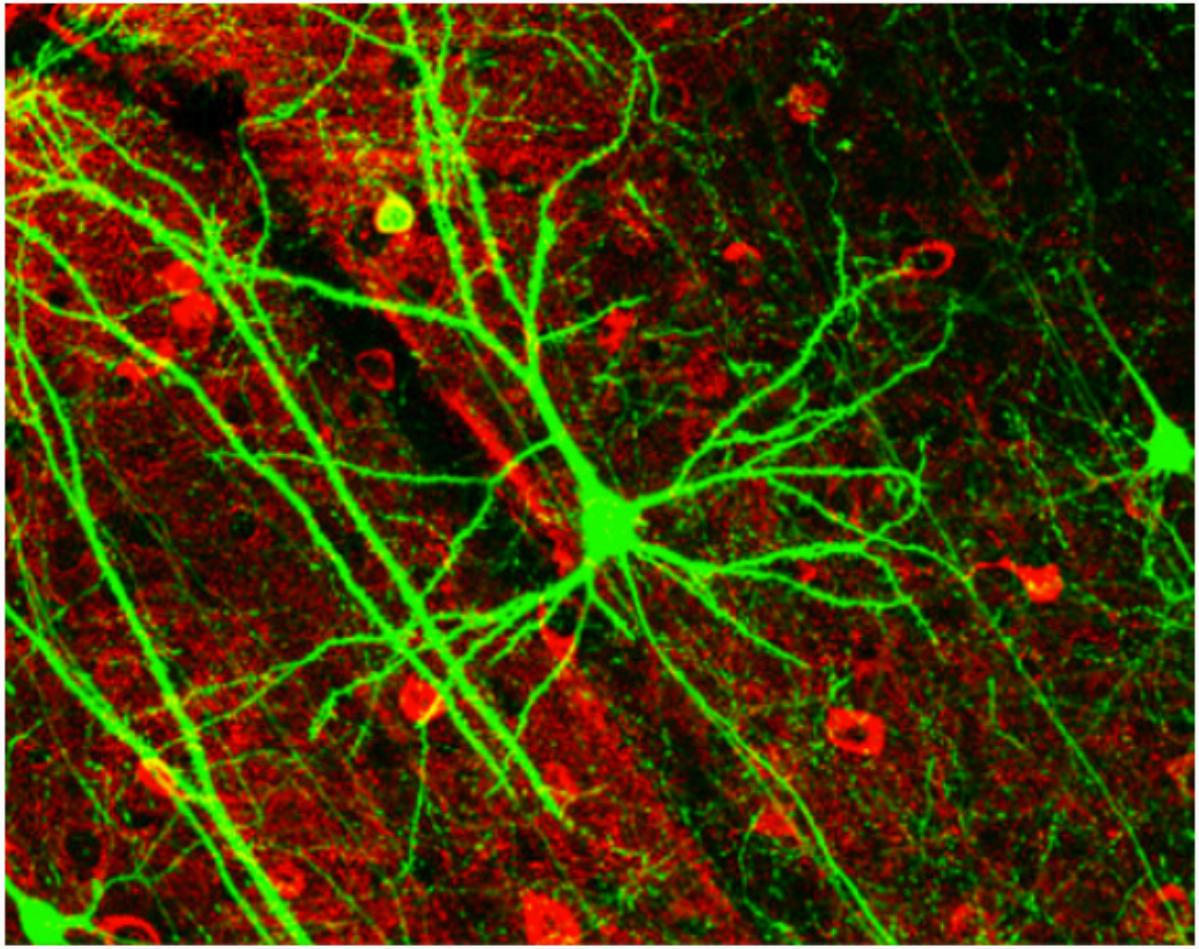
NEURONY



Biologický neuron



The human nervous system has ca. 10^{15} neurons. Transmission of an electric signal between dendrites and axons occurs through the transport of ions.



Biologický neuron

Neurons in the superficial layers of the visual cortex in the brain of a mice.

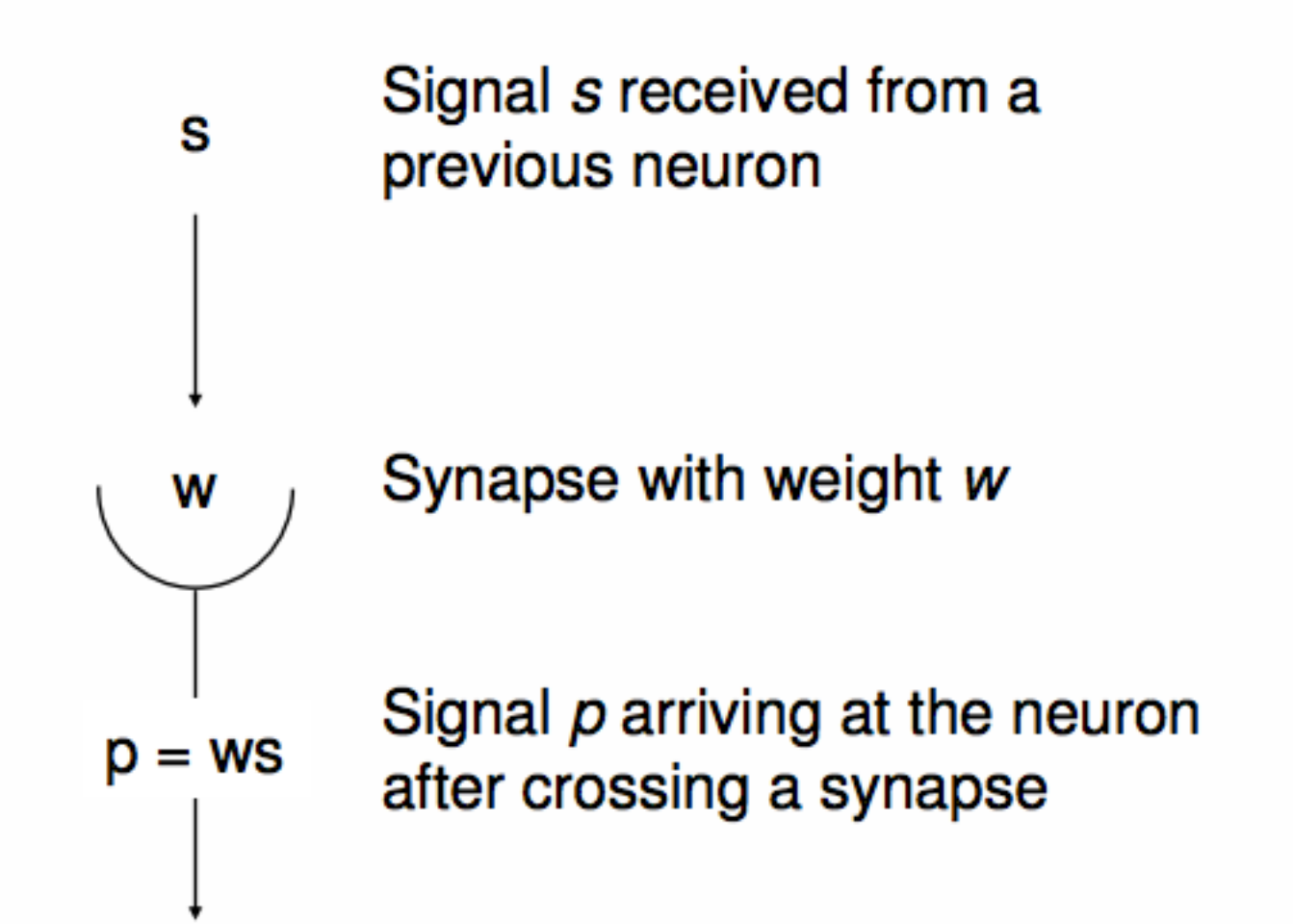
PLoS Biology Vol. 4, No. 2, e29 DOI: 10.1371/journal.pbio.0040029

Co je důležité pro neurony?

SÍŤ (NETWORK)

Co je důležité pro neurony?

Přenos signálu

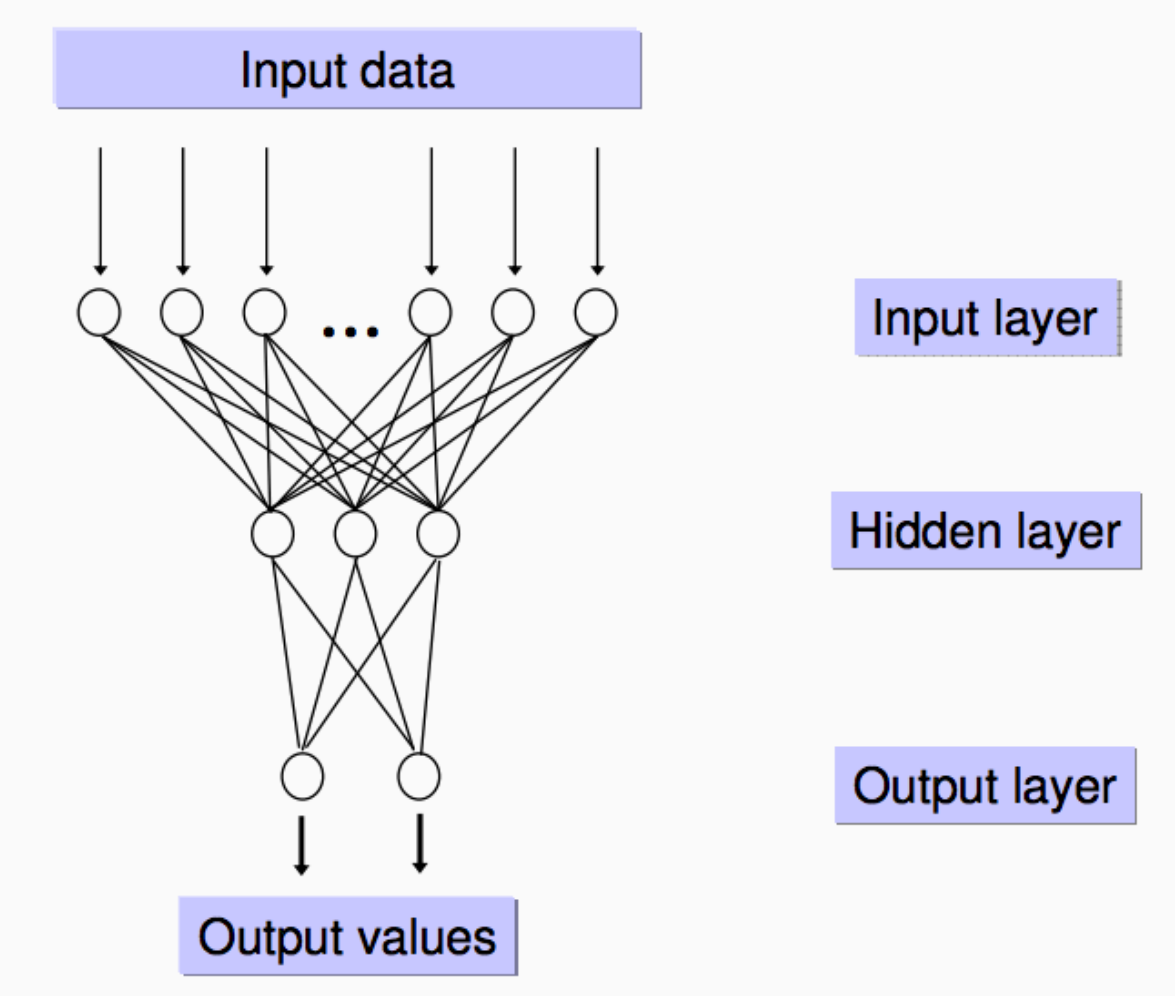


In artificial neurons, the synaptic strength is called **weight**.

Synapse a učení

- **Učení a pamět jsou považovány za výsledek dlouhodobých změn synaptické síly.**
- **V umělých neuronových sítích dochází k učení opravou váhy.**

Neuronové sítě



KVALITA MODELŮ

Kvalita QSPR modelů I

- kvalitu modelu můžeme posuzovat podle dvou kritérií

① kvalitu modelu na tréninkové sadě dat

- **reprodukce**
- data byla použita pro naučení modelu
- jak moc dobré modely jsme připravili?

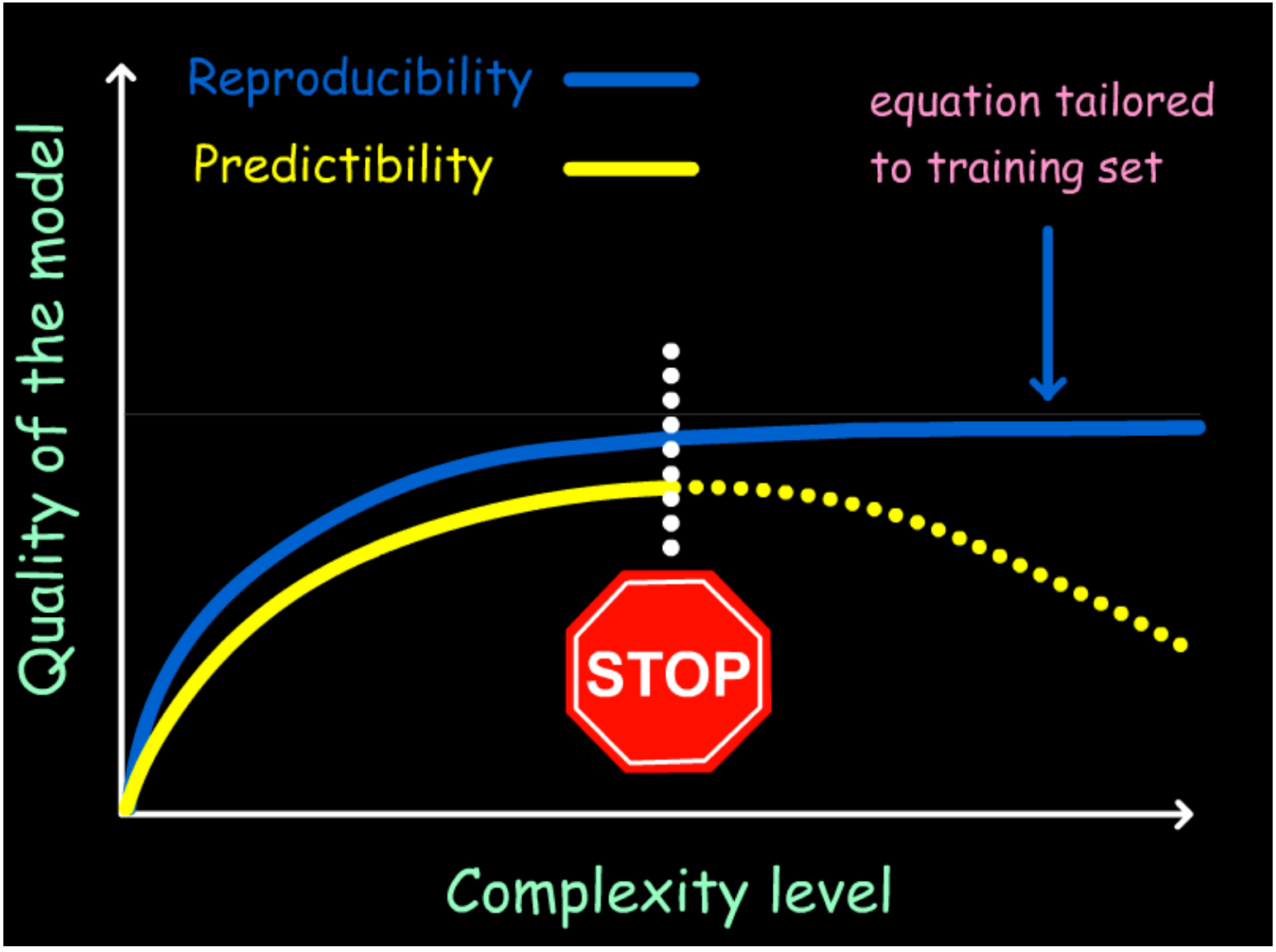
② kvalitu modelu na testovací sadě dat

- **predikce** (na nových datech)
- data nebyla použita na parametrizaci modelu
- jaká je predikční sada molekul?

Kvalita QSPR modelů II

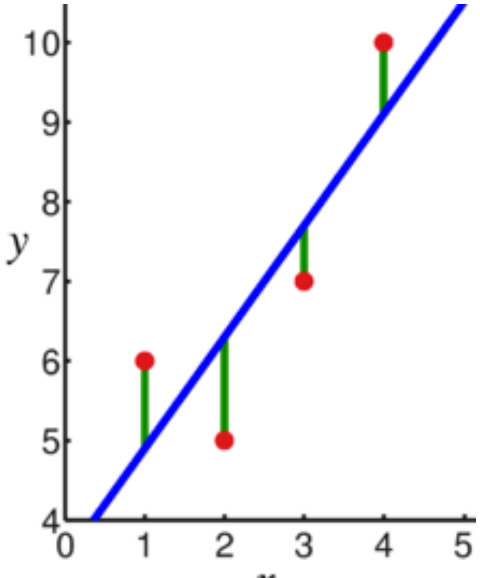
	nekvalitní model na tréninkové sadě dat	kvalitní model na tréninkové sadě dat
nekvalitní model na testovací sadě dat	–	špatně rozdělené sady, “overfiting” neboli přeučení = použito příliš moc deskriptorů
kvalitní model na testovací sadě dat	–	KVALITNÍ MODEL

Kvalita QSPR modelů III



Kvalita QSPR modelů IV

- na základě chyb modelu
- = residua, nevysvětlitelná část modelu



P^{exp}	P^{calc}	$\text{error} = P^{exp} - P^{calc}$
...	...	...
$\text{p}K_a{}^{exp}$	$\text{p}K_a{}^{calc}$	error
10.0	10.1	-0.1
...	...	...

- vyjadřujeme pomocí R^2, adjR^2, RMSE, MAE a F

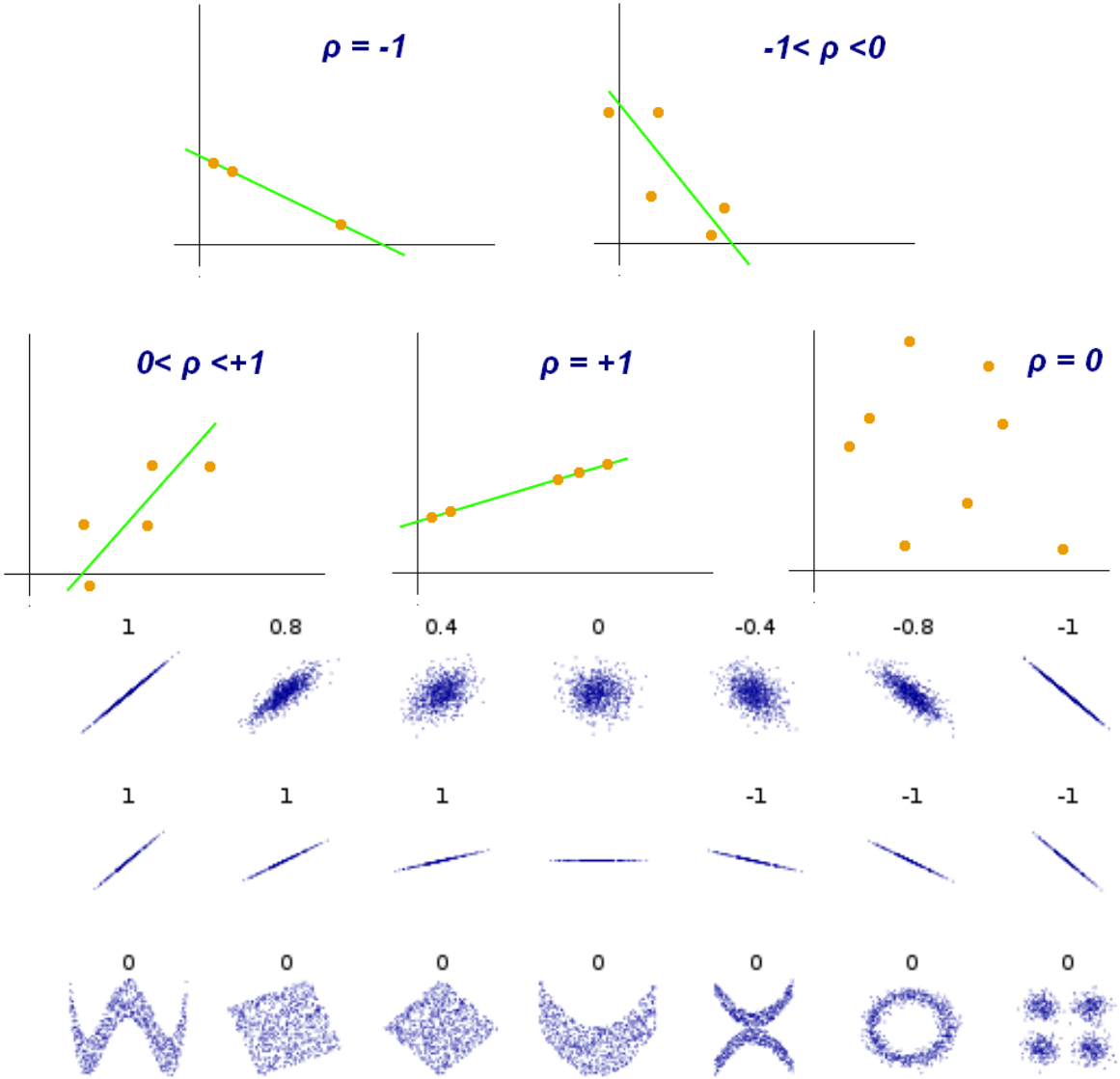
Pearsonův korelační koeficient I

$$R = \sqrt{\frac{\sum_{i=1}^{N} \left((P_i^{calc} - \bar{P}calc) \cdot (P_i^{exp} - \bar{P}exp) \right)}{\sum_{i=1}^{N} (P_i^{calc} - \bar{P}calc)^2 \cdot \sum_{i=1}^{N} (P_i^{exp} - \bar{P}exp)^2}}$$

$\bar{P}^{calc}$ průměrná vypočítaná hodnota,
$\bar{P}^{exp}$ průměrná experimentální hodnota

Nabývá hodnot od -1 do 1.

Pearsonův korelační koeficient II



Koeficient determinace R^2 I

- Leží v intervalu <0;1> a udává jaký podíl rozptylu v pozorování závislé proměnné se podařilo regresí vysvětlit (vetší hodnoty znamenají větší úspěšnost).
- Možná interpretace koeficientu R^2 je z kolika procent vysvětlují regresory (deskriptory) hodnotu závisle proměnné (predikované vlastnosti).

Koeficient determinace R^2 II

Residual sum of squares:

$$RSS = \sum_{i=1}^{N} error^2 = \sum_{i=1}^{N} (P_i^{calc} - P_i^{exp})^2$$

Total sum of squares: $TSS = \sum_{i=1}^{N} (P_i^{exp} - \bar{P}^{exp})^2$

Explained sum of squares: $ESS = (P_i^{calc} - \bar{P}^{calc})^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$R^2 = \frac{\sum_{i=1}^{N} (P_i^{calc} - \bar{P}^{calc})^2}{\sum_{i=1}^{N} (P_i^{exp} - \bar{P}^{exp})^2} = 1 - \frac{\sum_{i=1}^{N} (P_i^{calc} - P_i^{exp})^2}{\sum_{i=1}^{N} (P_i^{exp} - \bar{P}^{exp})^2}$$

Korigovaný koeficient determinace adjR^2

- pokud do modelu přidáme deskriptor, hodnota R^2 nemůže klesnout, proto se někdy používá tzv. korigovaný koeficient determinace (**adjusted coefficient of determination**), který zohledňuje počet deskriptorů

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1}$$

kde N je velikost sady, k počet deskriptorů

RMSE
root mean square error (deviation)

$$RMSE = \sqrt{mean(error^2)} = \sqrt{\frac{\sum error^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(P_i^{calc} - P_i^{epx})^2}{N}}$$

MAE
mean absolute error

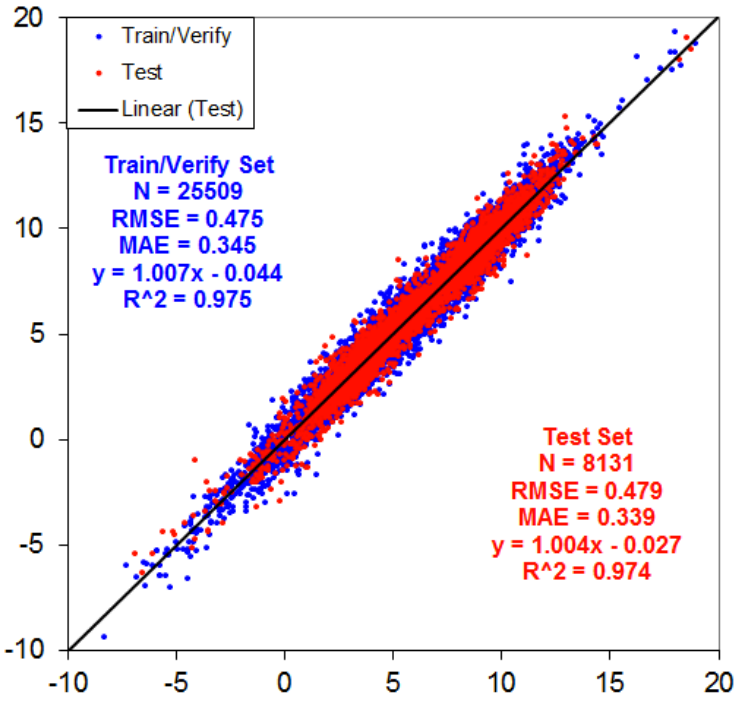
$$MAE = mean(abs(error)) = \frac{\sum |error|}{N} = \frac{\sum_{i=1}^{N} |P_i^{calc} - P_i^{epx}|}{N}$$

Test významnosti modelu F

$$F \sim \frac{N-k+1}{k} \frac{RSS-TSS}{TSS} = \frac{N-k+1}{k} \frac{R^2}{1-R^2}$$

Kvalita QSPR modelů V

- Kvalitní model by měl splňovat tato kritéria:
  - vysoké hodnoty R^2 (>0.8) a F
  - nízké hodnoty RMSE a MAE



Křížová validace
Cross validation

- v případě menší sady molekul
- nejčastěji se používá tzv. *k*-fold cross validation; příklad 5-fold: