Úvod do strojového učenia
Machine learning

Zuzana Rošťáková

23. september 2021

Seminár ÚM

Doplnková literatúra I

- **James G., Witten D., Hastie T., Tibshirani R. (2021)**
  An Introduction to Statistical Learning: with applications in R.
  Springer. Second Edition. ISBN: 978-1-0716-1417-4.
  https://doi.org/10.1007/978-1-0716-1418-1

  - detailný popis vybraných metód strojového učenia
  - príklady na reálnych dátach v štatistickom softvéri R

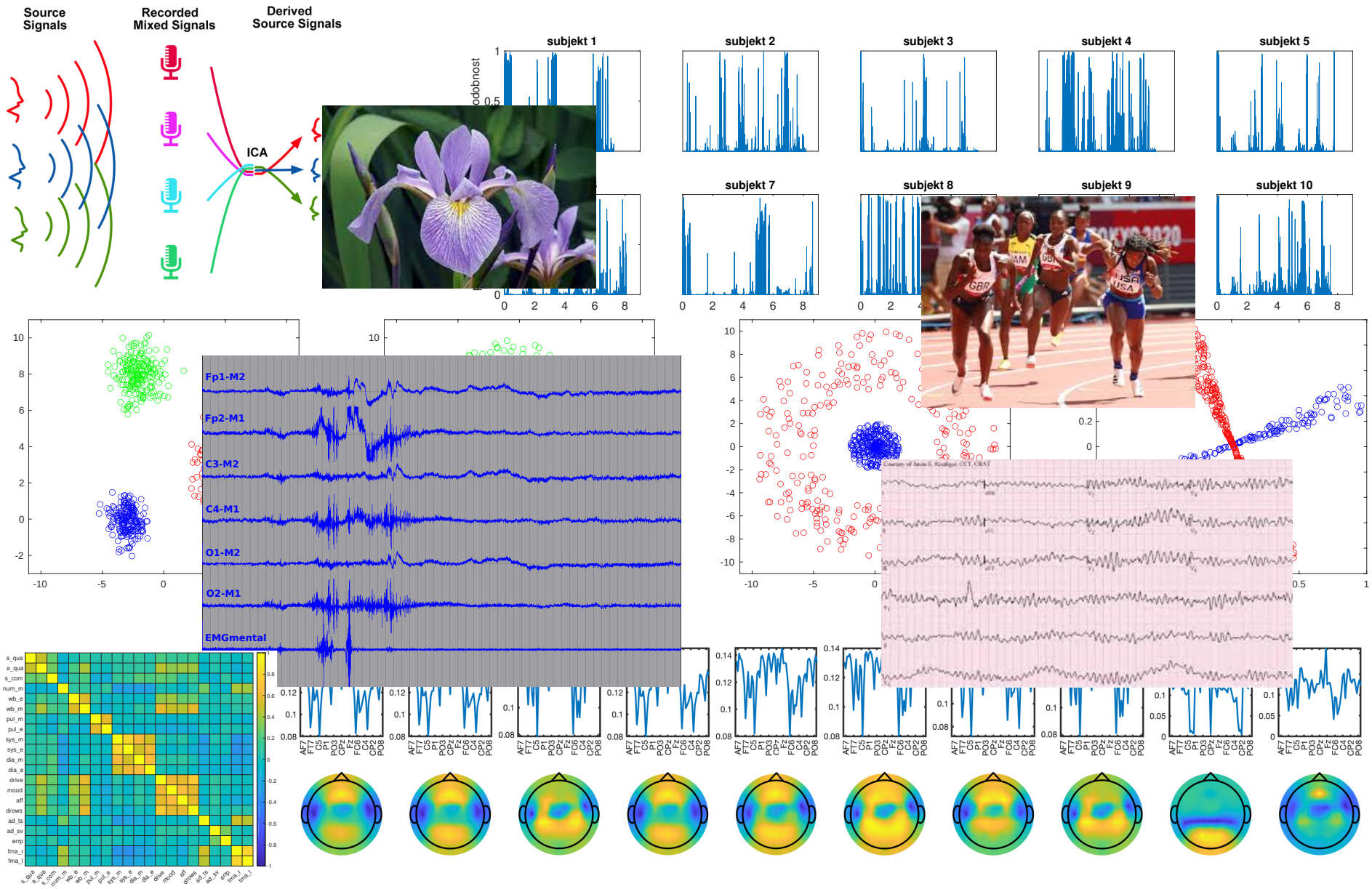
  https://link.springer.com/content/pdf/10.1007%2F978-1-0716-1418-1.pdf

- **Everitt B. S.(2005)**
  An R and S-plus Companion to Multivariate Analysis
  Springer. ISBN 978-1-85233-882-4.
  https://doi.org/10.1007/b138954

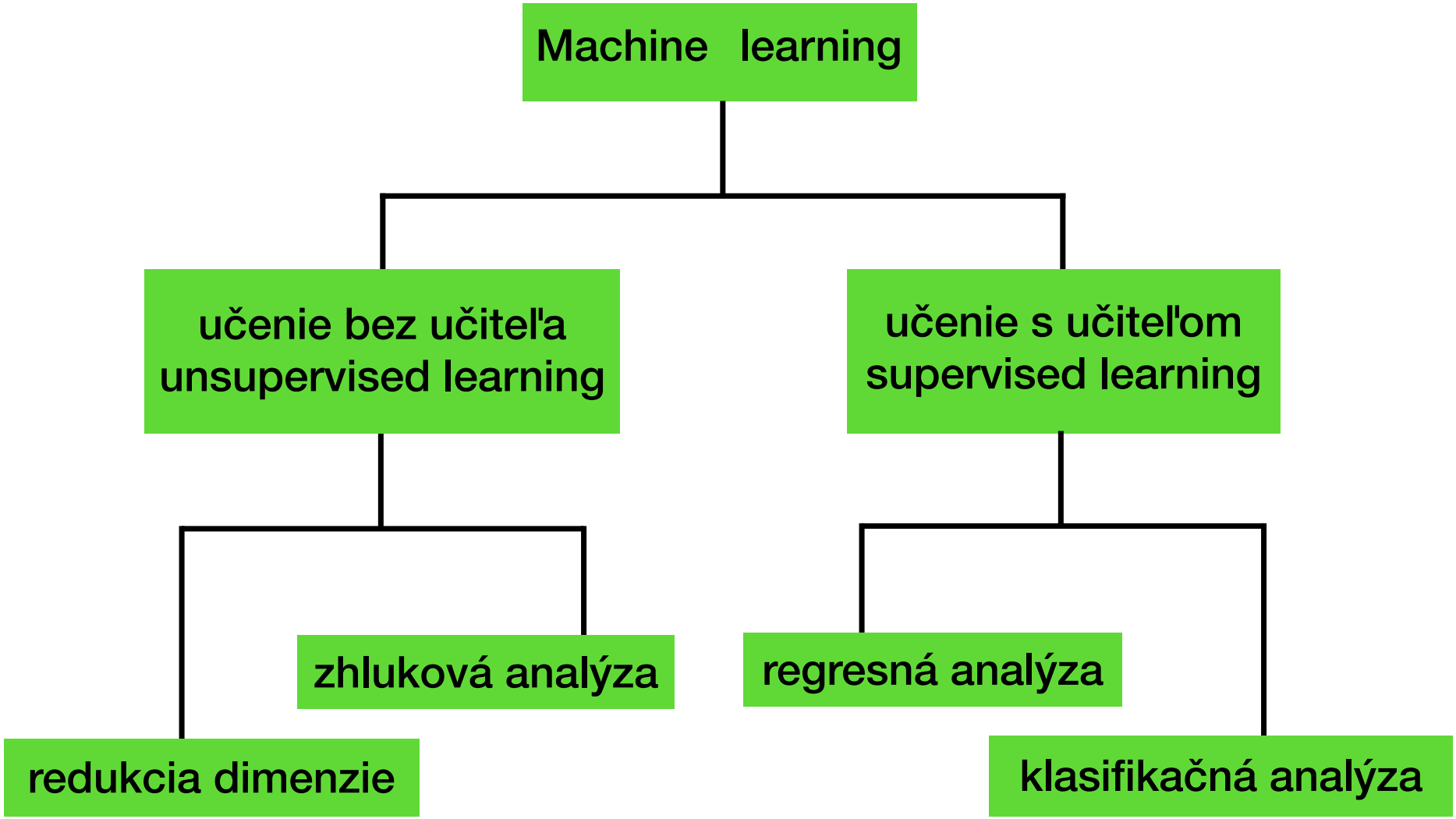
Doplnková literatúra II

- elektronické zdroje k predmetu ***Analýza zhlukov a klasifikácia dát*** od **doc. Mgr. Radoslava Harmana, PhD.**
  - http://www.iam.fmph.uniba.sk/ospm/Harman/VSAp.pdf
- **Lamoš, F., Potocký, R. (1998)**
  Pravdepodobnosť a matematická štatistika (štatistické analýzy)
  Univerzite Komenského, Bratislava. ISBN: 80-223-1262-2
- **Mohammed, M., Khan, M. B., Bashier, E. B. M. (2016).**
  Machine learning: algorithms and applications.
  Crc Press. ISBN: 978-1-31537-165-8
  https://doi.org/10.1201/9781315371658
  - príklady implementácie jednotlivých metód v MATLAB-e

Machine learning - strojové učenie

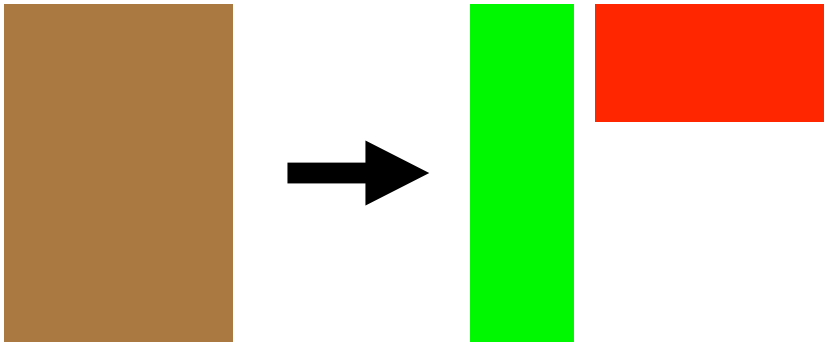


Machine learning - strojové učenie

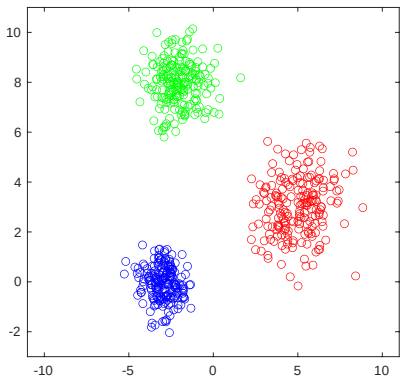


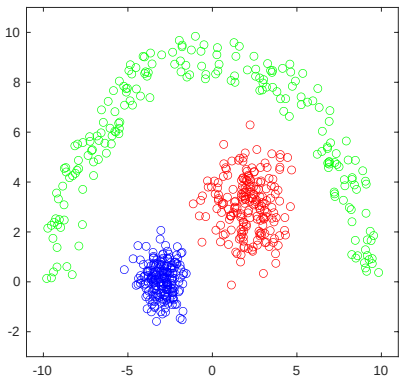
Učenie bez učiteľa - unsupervised learning

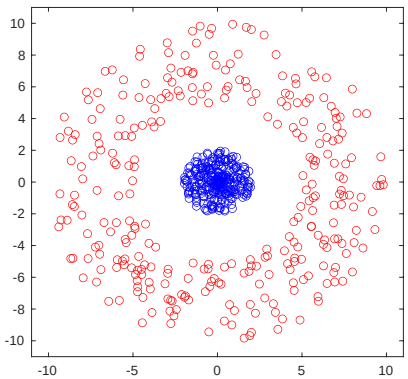
- **cieľ:**
  - **redukcia dimenzie:** nájsť menší počet skrytých (latentných) premenných

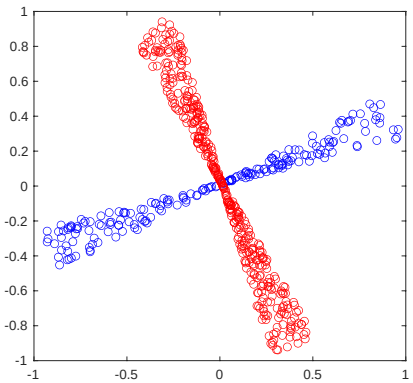


  - **zhluková analýza:** nájsť skryté podskupiny podobných pozorovaní



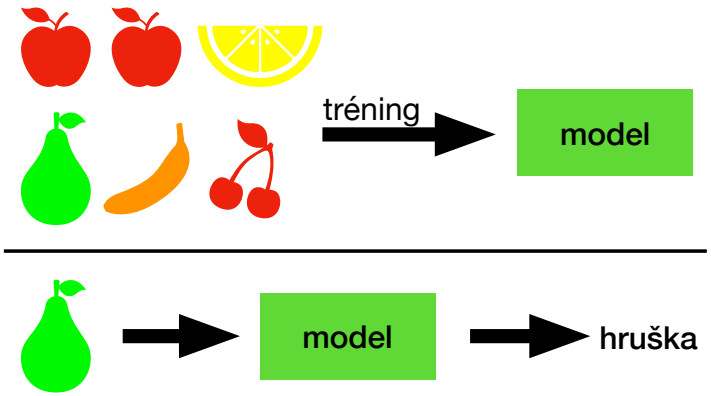






Učenie s učiteľom - supervised learning

- cieľ:
  - **klasifikačná analýza:** vytvoriť model, ktorý by na základe vstupných údajov zaradil pozorovania do skupín



  - **regresná analýza:** nájsť vzťah medzi vstupnými premennými a výstupnou premennou (premennými)

$$\mathbf{Y} = \mathbf{X} \beta + \mathbf{E}$$

Vizualizácia mnohorozmerných dát

Vizualizácia mnohorozmerných dát

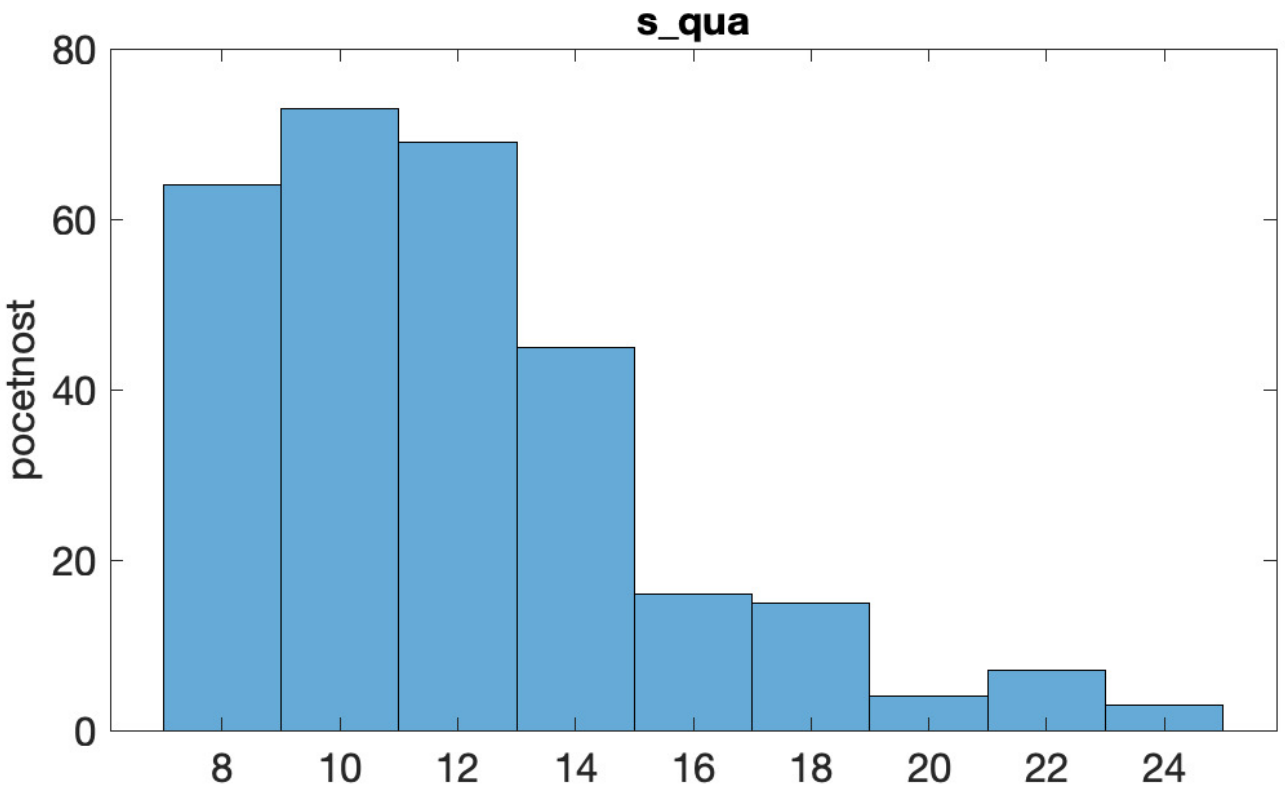
- pozorujeme **p** premenných na **n** objektoch
- **dáta:** matica X o rozmere $n \times p$

$$X = \begin{bmatrix} — & \mathbf{x_1^T} & — \\ — & \mathbf{x_2^T} & — \\ & \vdots & \\ — & \mathbf{x_n^T} & — \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{x_1^\star} & \mathbf{x_2^\star} & \dots & \mathbf{x_p^\star} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times p}$$

  - $\mathbf{x_i} = (x_{i1}, x_{i2}, \dots, x_{ip})^T$
    $\rightarrow$ vektor hodnôt **p** premenných pre i-ty objekt, $i = 1, \dots, n$
  - $\mathbf{x_l^\star} = (x_{1l}, x_{2l}, \dots, x_{nl})^T$
    $\rightarrow$ vektor hodnôt l-tej premennej, $l = 1, \dots, p$ pre **n** objektov
- pred samotnou analýzou je vhodné získať určitý **obraz** o dátach
- rôzne prístupy k vizualizácií p-rozmerných dát

Histogram

- sĺpcový diagram
- základňa každého obdĺznika diagramu má dĺžku zvoleného intervalu
- výška každého obdĺžnika je rovná počtu pozorovaní, ktoré patria do zvoleného intervalu
- MATLAB: *histogram*(x, *vektor_hranic_intervalov*)



Histogram - Príklad: Dotazníky

- 148 ľudí, 2 noci v spánkovom laboratóriu $\rightarrow$ 296 pozorovaní [Rosipal et al., 2013]

- večer a ráno vypĺňali dotazníky ohľadom ich subjektívneho stavu

$\rightarrow$ **9 premenných**

  - Self-rating questionnaire for sleep quality, awakening quality and somatic complaints (s_qua, a_qua,s_com)

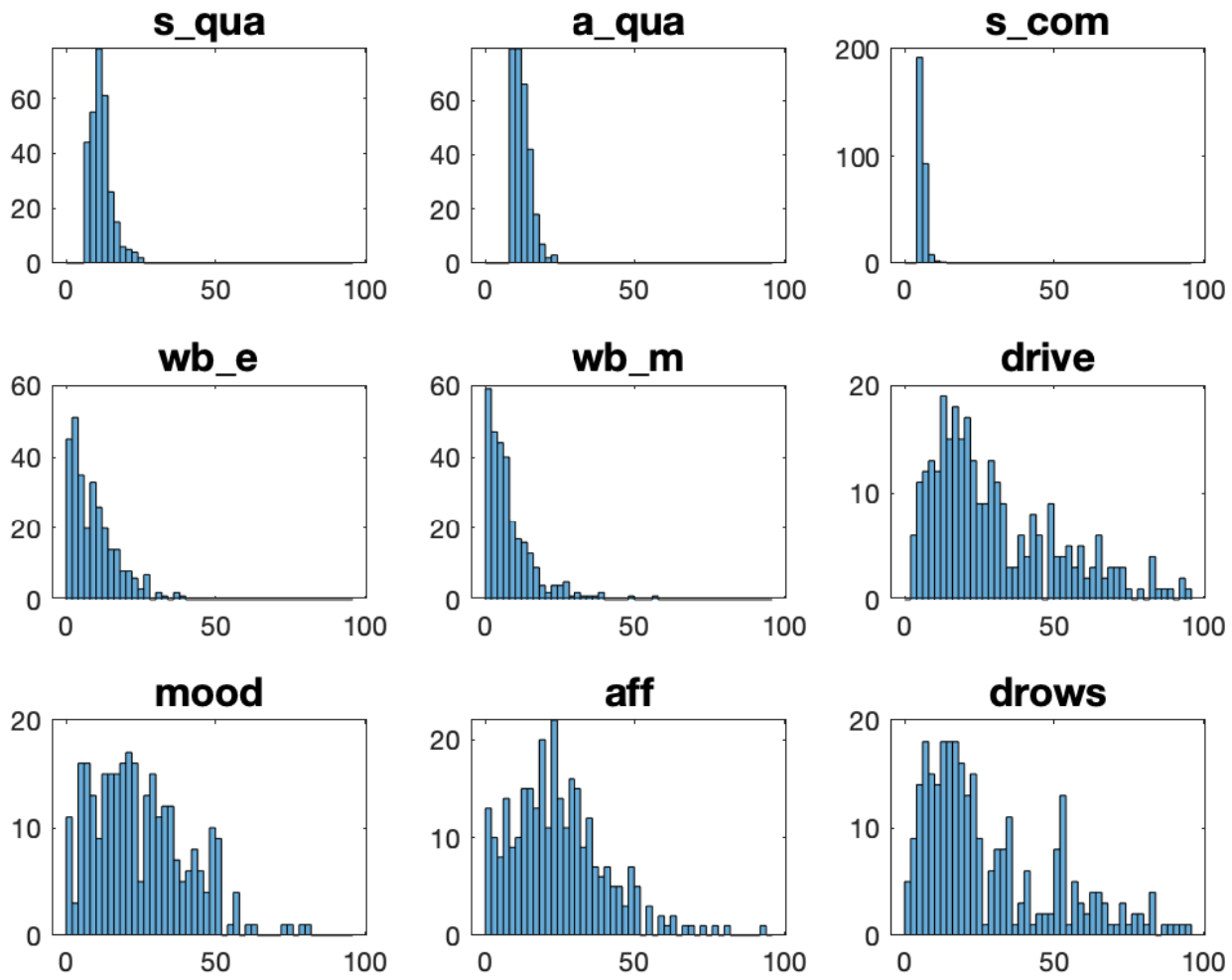
  - Well-being self-assessment scale evening/morning (wb_e, wb_m)

  - Visual analog scale test for drive, mood, affectivity and drowsiness (drive, mood, aff, drows)

Histogram - Príklad: Dotazníky

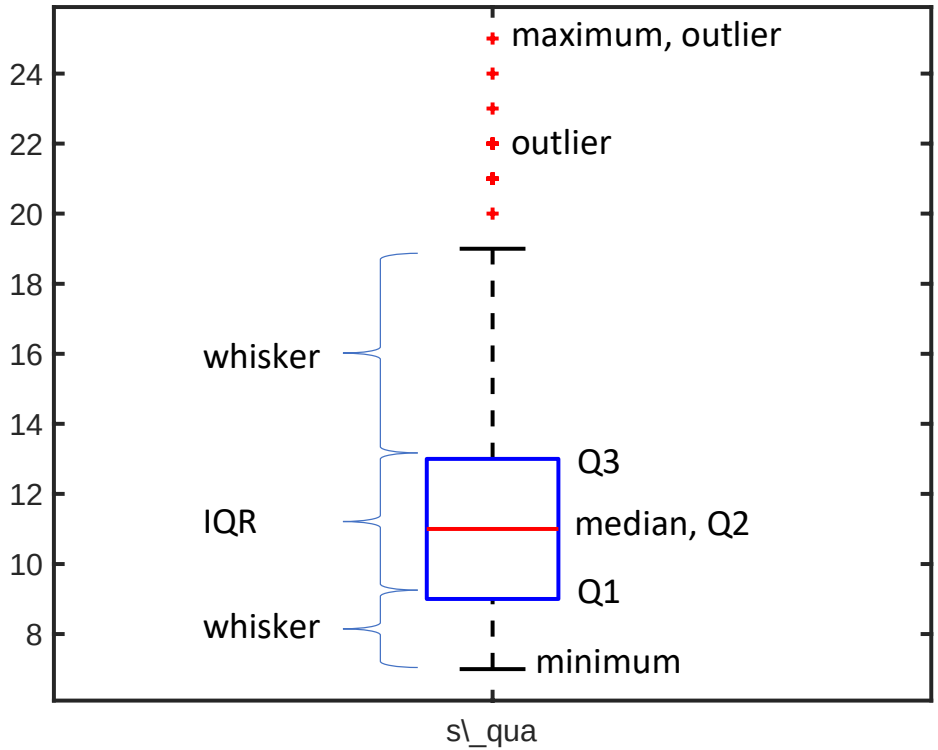
- každú premennú osobitne znázorníme pomocou histogramu

$$histogram(data, 0:2:100)$$

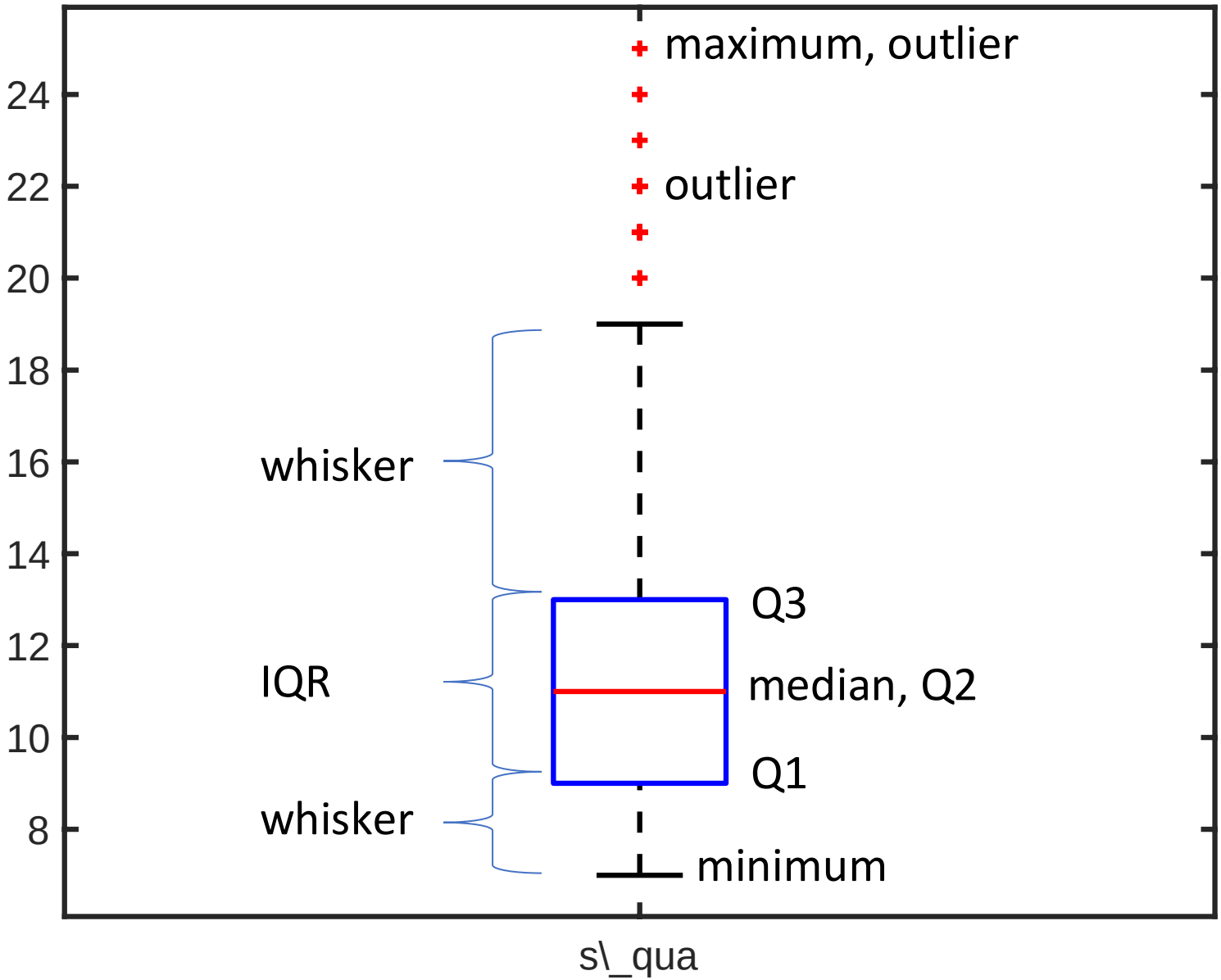


Boxplot - krabicový graf

- autor: John Tukey (navrhnuté 1970, publikované 1977)
- medián, minimum, maximum
- prvý (Q_1, 25%-ný kvantil) a tretí kvartil (Q_3, 75%-ný kvantil)
- $IQR = Q_3 - Q_1$, medikvartilové rozpätie
- whiskers (fúzy) = 1.5 x IQR (alebo vlastná definícia)

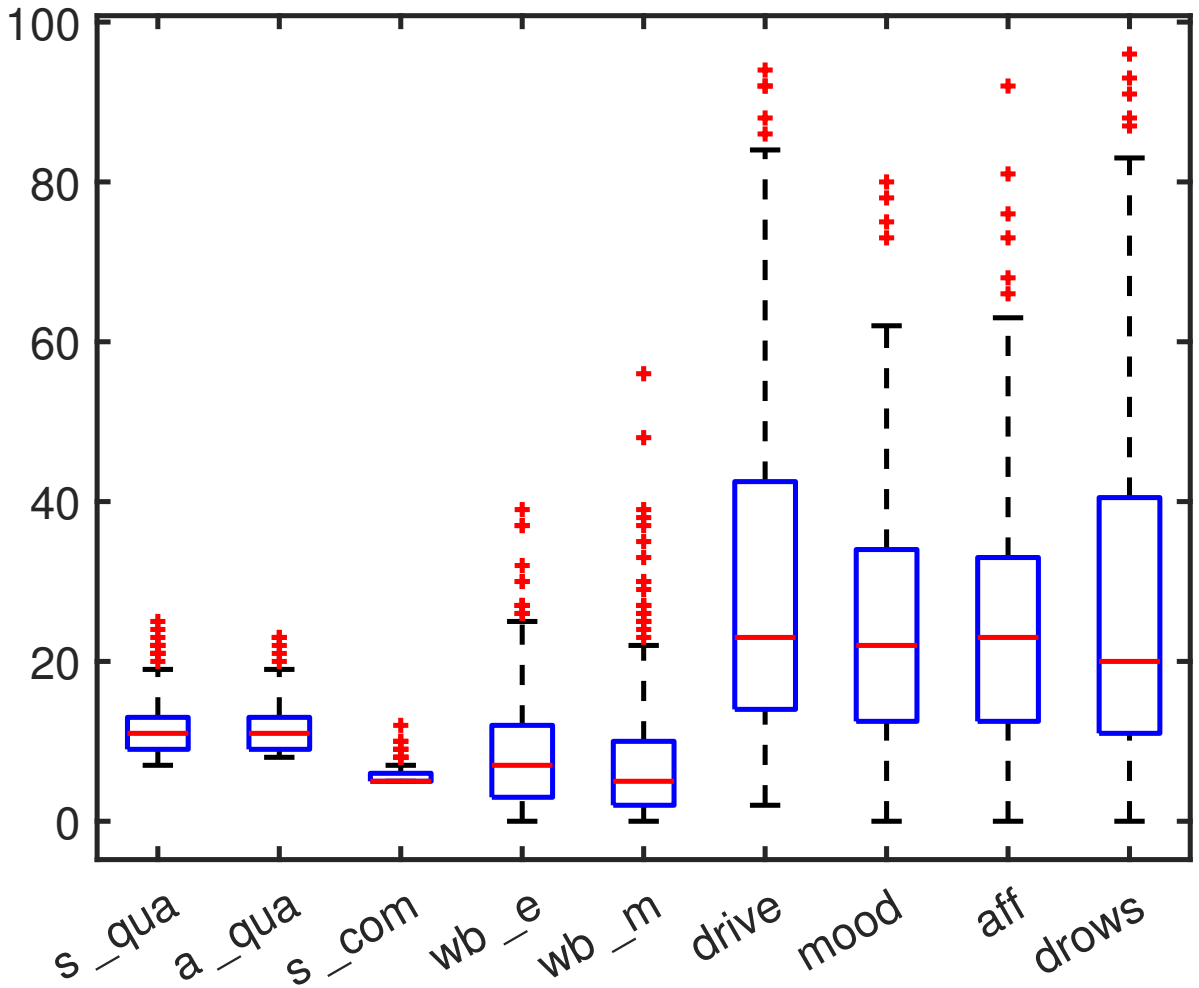


Boxplot - krabicový graf



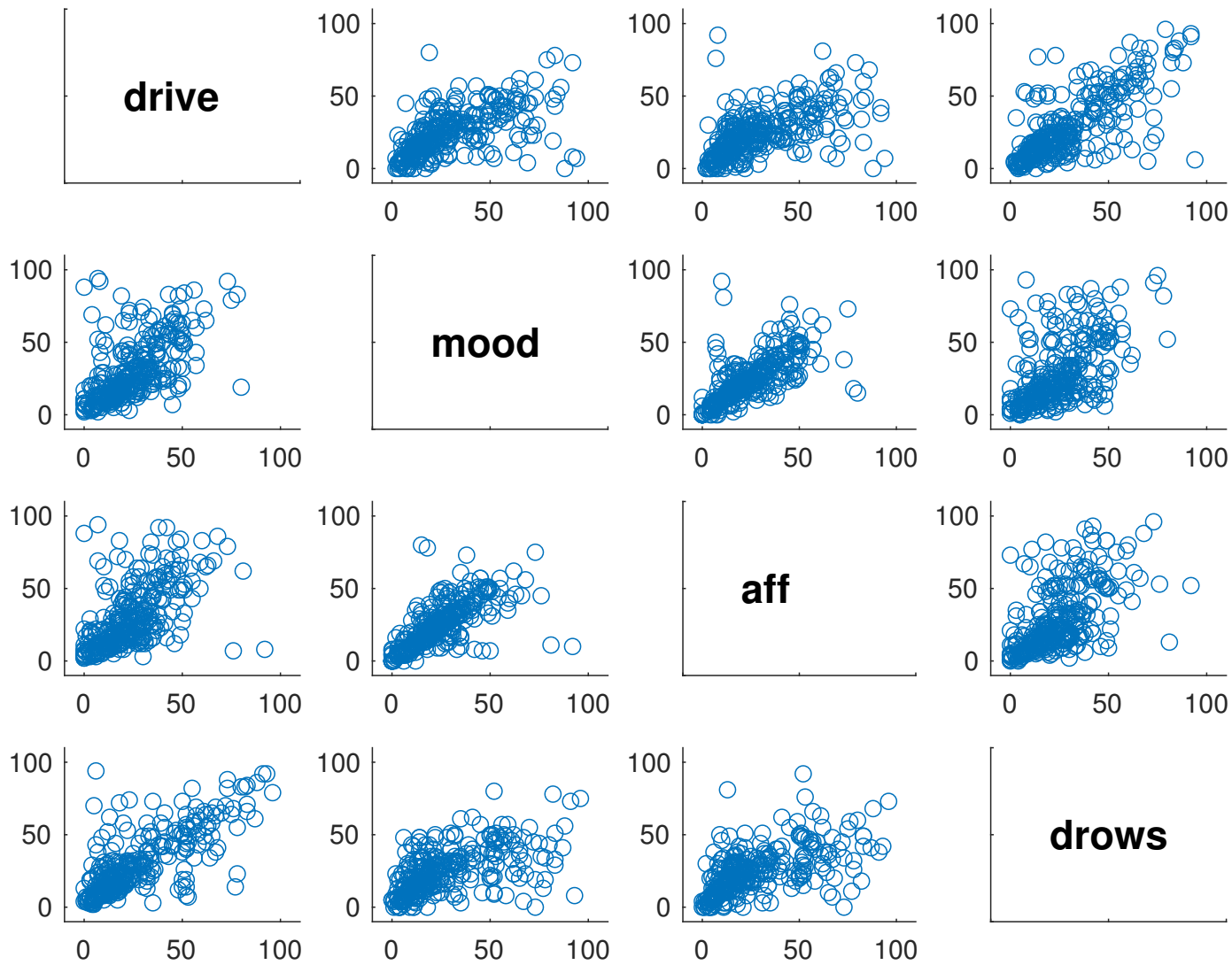
Boxplot - Príklad: Dotazníky

- *boxplot*(*data*, '*labels*', *variable_names*)



Scatterplot

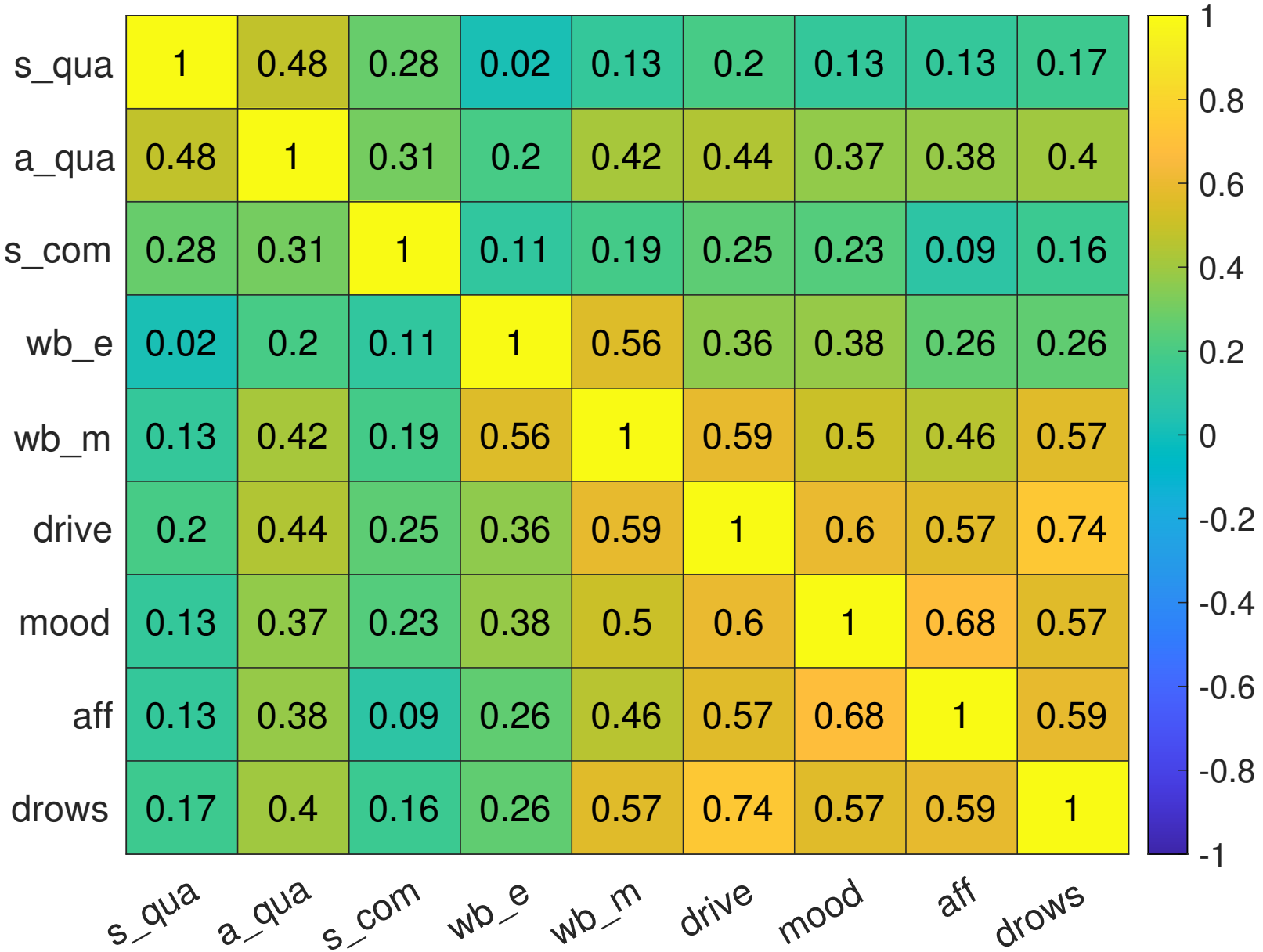
$scatter(x, y)$ alebo $plot(x, y, 'o')$



Heatmap

- vizualizácia korelačnej matice dát
- $R = corr(data)$
  - $p \times p$ matica
  - R_{ij} = korelácia medzi i-tou a j-tou premennou
- $h = heatmap(xval, yval, R)$
  - $xval, yval$ = označenie premenných na osi x, y
  - $R = p \times p$ korelačná matica

HeatMap - Príklad: Dotazníky



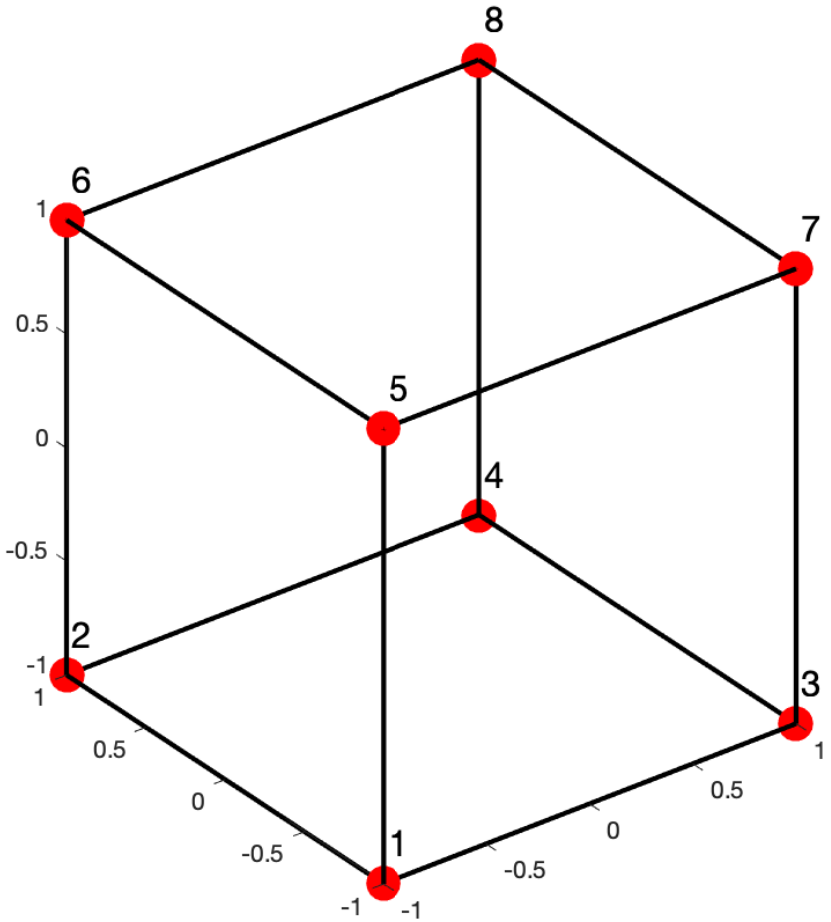
Mnohorozmerné škálovanie

- transformácia p-rozmerných dát do "2D" pre lepšiu vizualizáciu
- v MATLAB-e:
  1. matica vzájomných vzdialeností pomocou funkcie *pdist*
     - $D = pdist(X, \text{'}metric\text{'}) \rightarrow n \times n$ matica
     - 'metric' = euclidean, squaredeuclidean, cityblock, correlation, ...
  2. mnohorozmerné škálovanie pomocou funkcie *cmdscale* (*classical multidimensional scaling*))
     - $Xd = cmdscale(D, pocet\_dimenzii)$

       $\rightarrow$ ak chceme dáta len v "2D", tak $pocet\_dimenzii = 2$
  3. vykreslenie dát v "2D"
     - scatter(Xd(:,1),Xd(:,2))

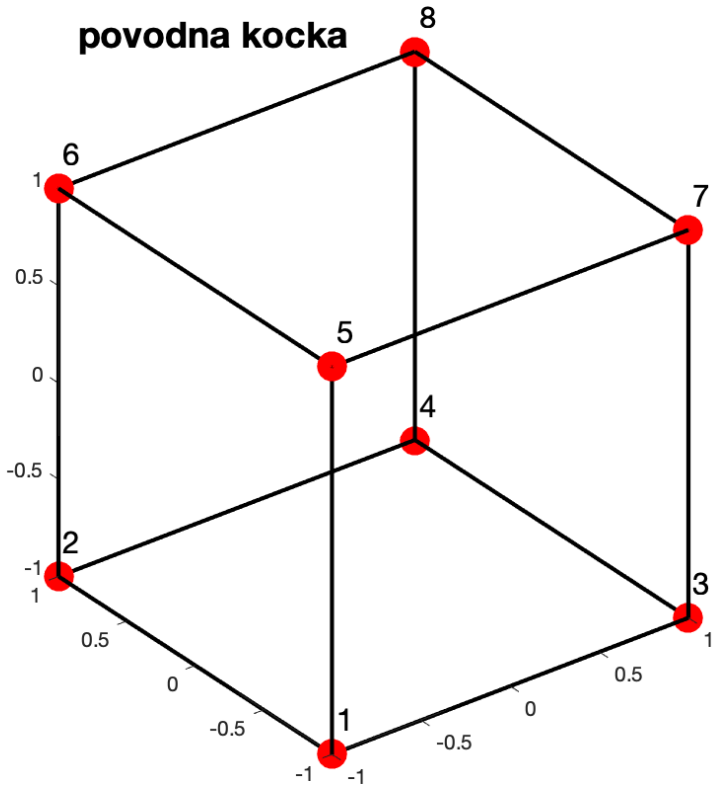
Mnohorozmerné škálovanie - Príklad: Kocka

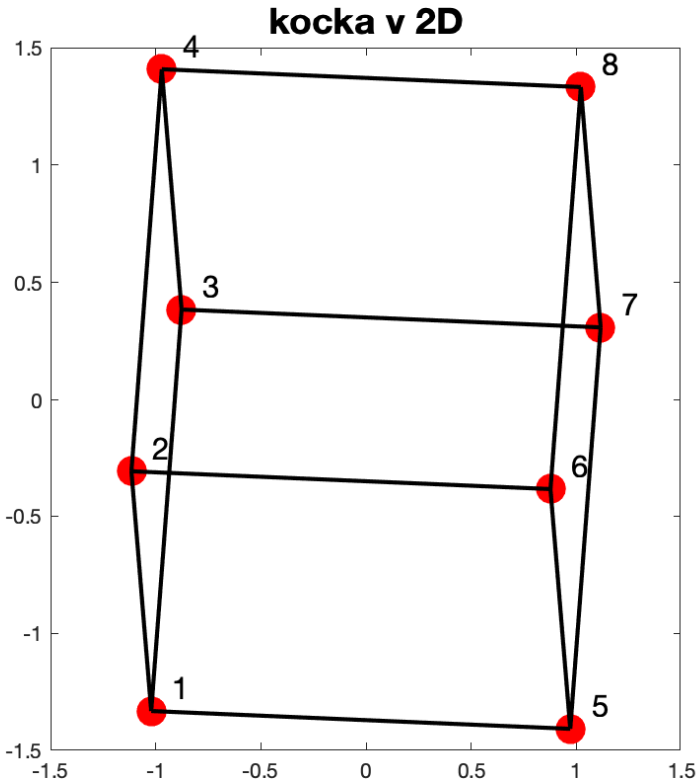
- $X \in \mathbb{R}^{8\times 3} \rightarrow$ kocka so stredom v [0,0,0]
- vykreslene pomocou funkcie $plot3(x, y, z)$



Mnohorozmerné škálovanie - Príklad: Kocka

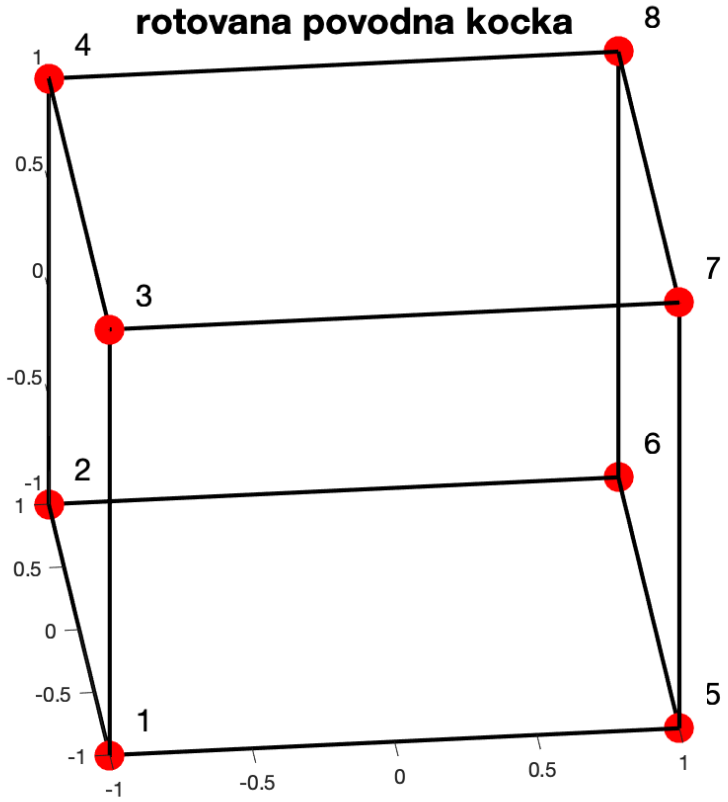
- kocka v 2D pomocou funkcie *cmdscale*

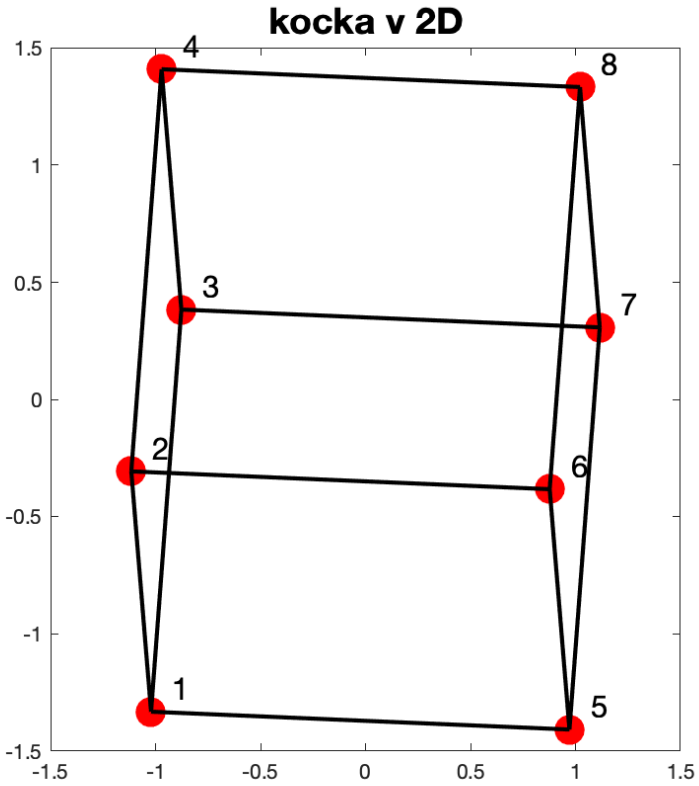




Mnohorozmerné škálovanie - Príklad: Kocka

- kocka v 2D pomocou funkcie *cmdscale*

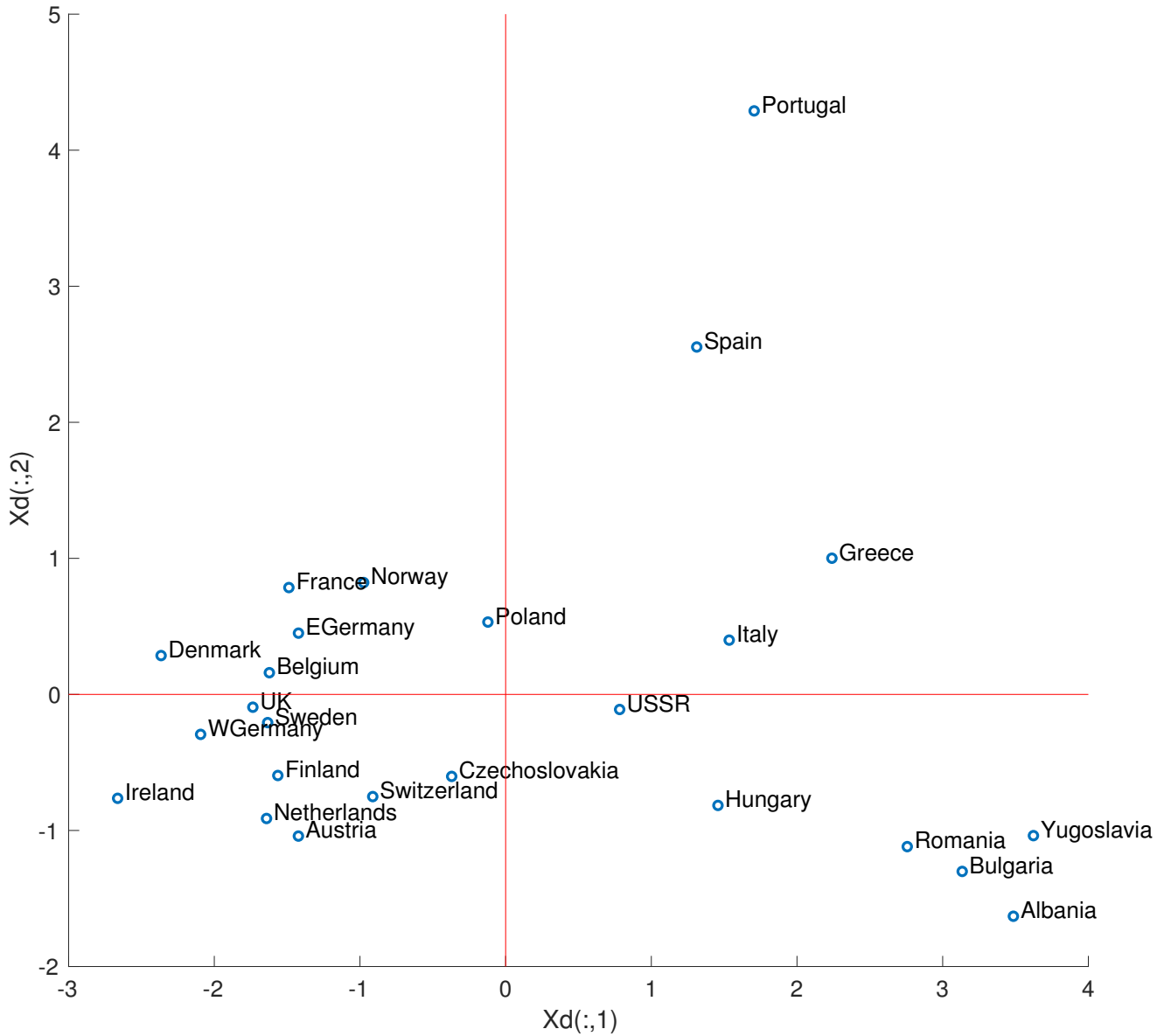




Mnohorozmerné škálovanie - Príklad: Potraviny

- 25 krajín Európy
- 9 premených (v mil. ton):
  - červené mäso - hovädzie, bravčové, ...
  - biele mäso - hydina, ...
  - vajíčka
  - mlieko
  - ryby
  - obilniny
  - potraviny obsahujúce škrob - zemiaky, ...
  - orechy
  - ovocie, zelenina
- staršie dáta (Juhoslávia, Československo, ...)
- zdroj:
  http://www.iam.fmph.uniba.sk/ospm/Harman/data/nutrition.txt

Mnohorozmerné škálovanie - Príklad: Potraviny



Otázky ?

Literatúra

Rosipal, R., Lewandowski, A., and Dorffner, G. (2013).
In search of objective components for sleep quality indexing in normal sleep.
Biological Psychology, 94(1):210–220.