Analýza a klasifikace dat – přednáška 1

RNDr. Eva Janoušová

Podzim 2014

Přínos předmětu

- orientace v principech rozpoznávání a klasifikace dat s důrazem na zpracování medicínských a biologických dat
- schopnost zvolit a aplikovat adekvátní metodu analýzy a klasifikace dat k dosažení požadovaných výsledků
- schopnost správné interpretace dosažených výsledků včetně vyhodnocení úspěšnosti klasifikace

Požadavky ke zkoušce

- předmět je ukončen ústní zkouškou – je nutné porozumět probíraným tématům, zvolit správnou metodu na daný analytický problém a umět získané výsledky interpretovat

- přednášky jsou kombinovány s cvičením – maximálně 2 absence

Doporučená literatura

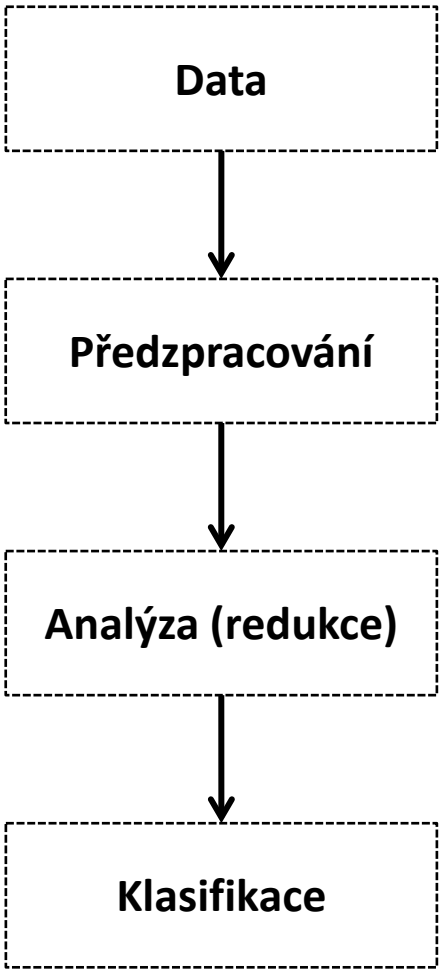
- Holčík, J.: Analýza a klasifikace dat. Brno, CERM 2012, 112s.
  http://www.iba.muni.cz/res/file/ucebnice/holcik-analyza-klasifikace-dat.pdf
  http://www.iba.muni.cz/index.php?pg=vyuka--ucebnice

- Janoušová, E. et al.: online výukové materiály vícerozměrným metodám a analýze a klasifikaci dat
  http://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--vicerozmerne-statisticke-metody

- DUDA R. O., HART P. E., STORK D. G., 2000: Pattern Classification. Wiley-Interscience, New York, 680 pp.
- BISHOP C., 2006: Pattern Recognition and Machine Learning. Springer, New York, 738 pp.
- KUNCHEVA L. I., 2004: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, New Jersey, 376 pp.

Vícerozměrná data, jejich popis a vizualizace

Schéma analýzy a klasifikace dat



	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M		90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

?

nebo

?

Vícerozměrná data

PROMĚNNÉ

OBJEKTY (SUBJEKTY)

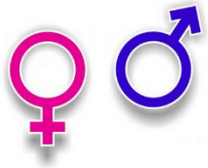
ID	Pohlaví	Věk	Váha	…
1	muž	84	85,5	
2	žena	25	62,0	
3				
4				
…				

Poznámka: proměnné označovány i jako znaky, pozorování, diskriminátory, příznakové proměnné či příznaky

Anglicky označení pouze jedním termínem: feature

Typy dat - opakování

- **Kvalitativní (kategoriální) data:**
  - Binární data



  - Nominální data

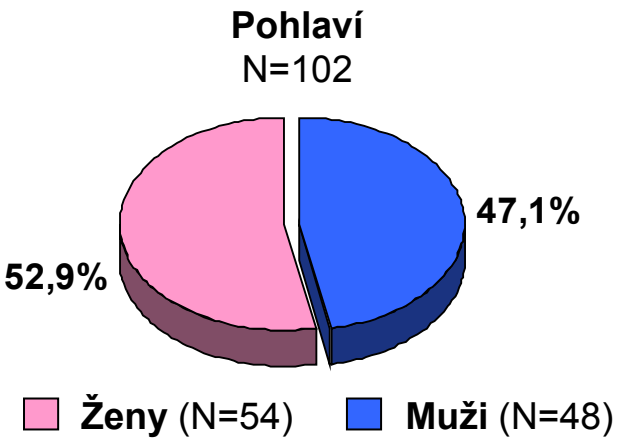
  - Ordinální data

- **Kvantitativní data:**
  - Intervalová data

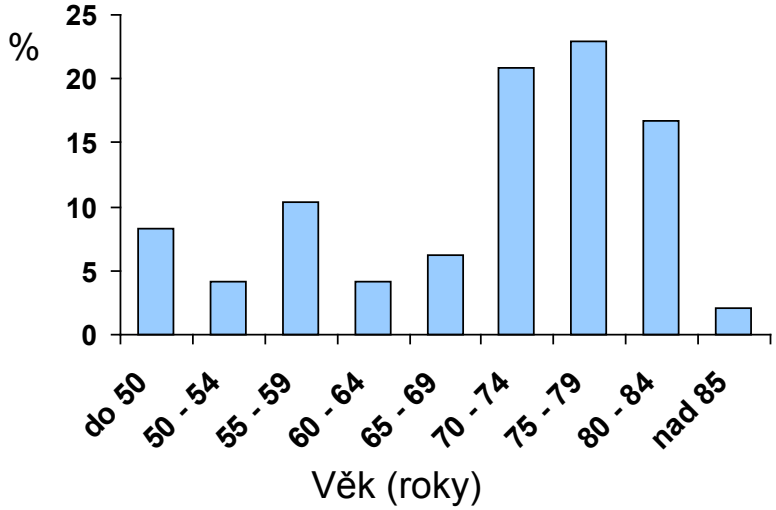
  - Poměrová data

Vizualizace jednorozměrných dat - opakování

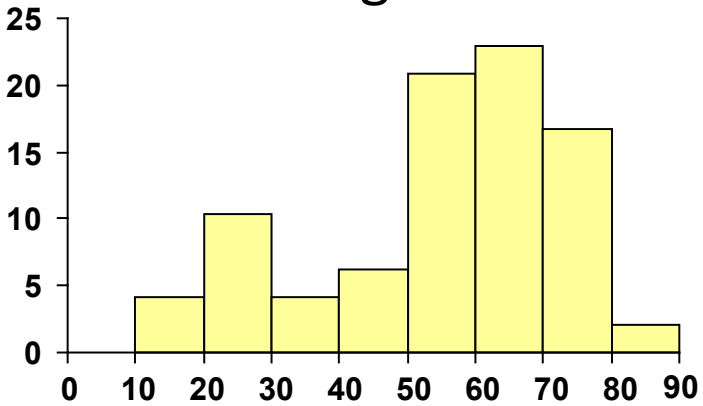
Koláčový graf



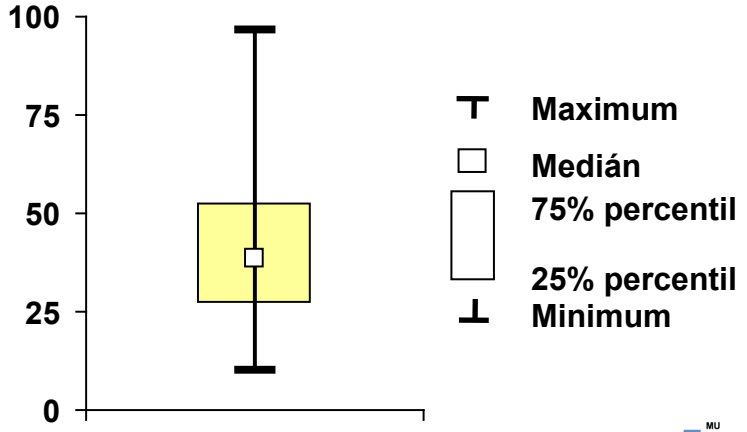
Sloupkový graf



Histogram



Krabicový graf (Box Plot)

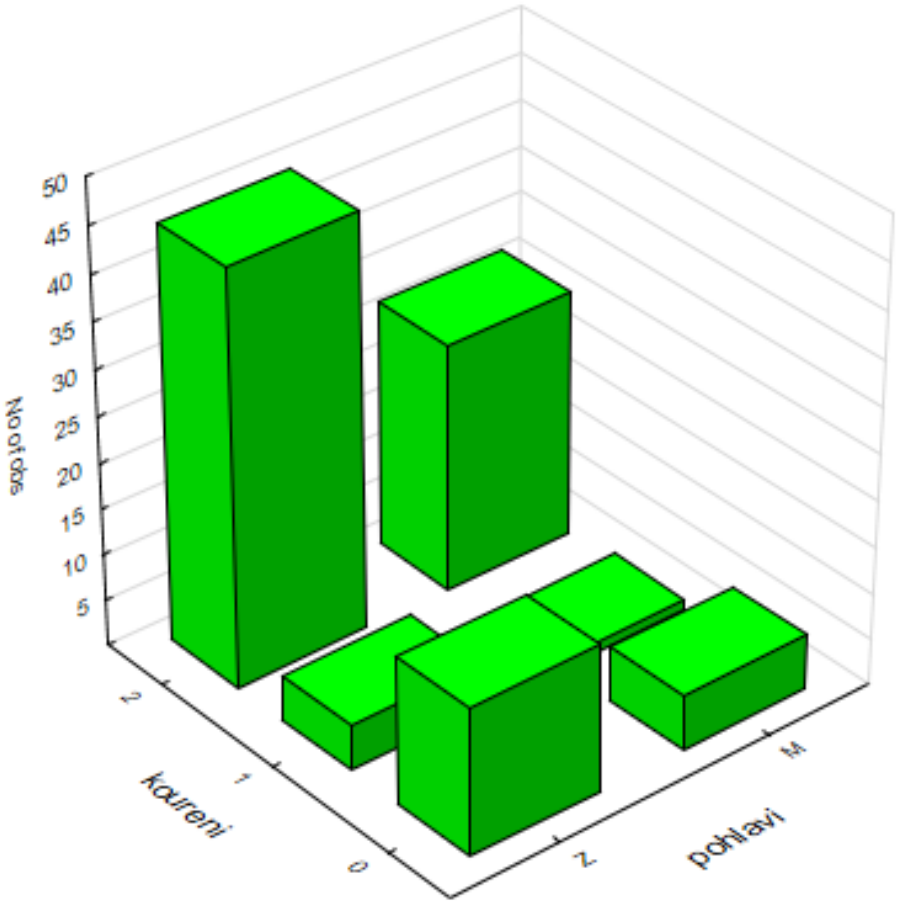


Vizualizace vícerozměrných dat

- 3D sloupkové grafy
- dvourozměrný histogram
- maticové grafy
- krabicové grafy pro více proměnných
- ikonové (symbolové) grafy:
  - profilové sloupce
  - profily
  - paprskové (hvězdicové) grafy
  - polygony
  - pavučinové grafy
  - Chernoffovy tváře

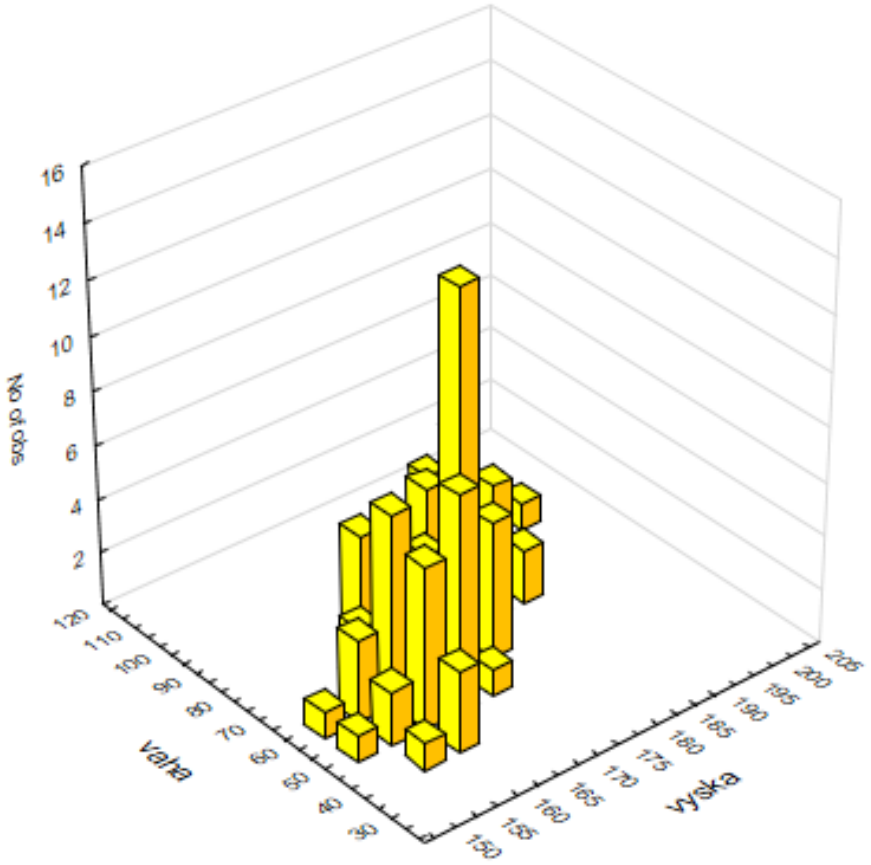
3D sloupkové grafy

- vzájemný výskyt kategorií dvou kategoriálních proměnných
- v softwaru Statistica: Graphs – 3D Sequential Graphs – Bivariate Histograms...



Dvourozměrný histogram

- pro vykreslení vztahu dvou spojitých proměnných
- v softwaru Statistica: Graphs – 3D Sequential Graphs – Bivariate Histograms...

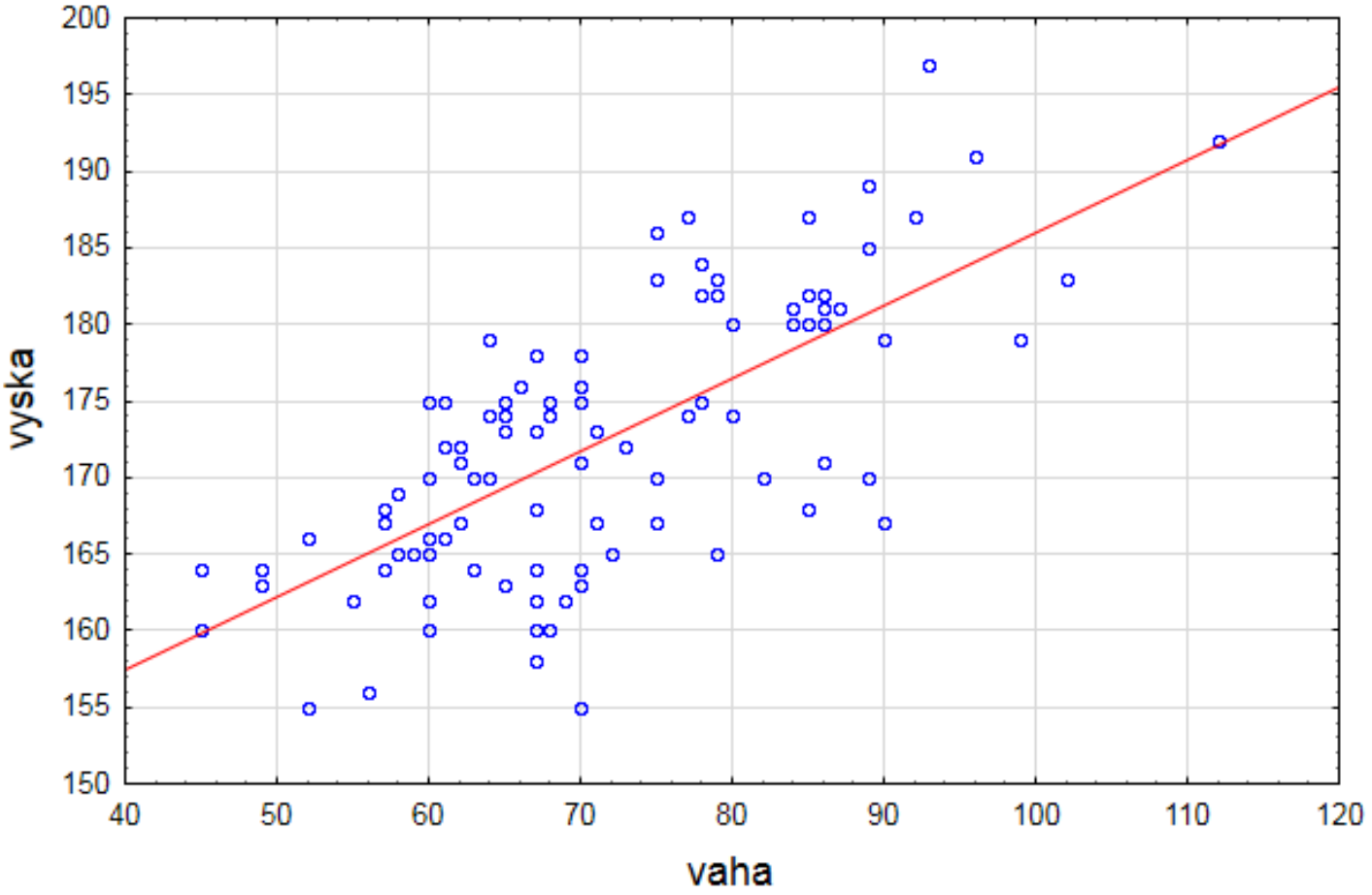


Úkol 1

- vykreslete dvourozměrný histogram pro věk a systolický tlak
- změňte barvu pozadí grafu na transparentní
- změňte barvu sloupečků (např. na červenou)
- zvětšete velikost písma u popisků os (u hodnot i názvů proměnných)

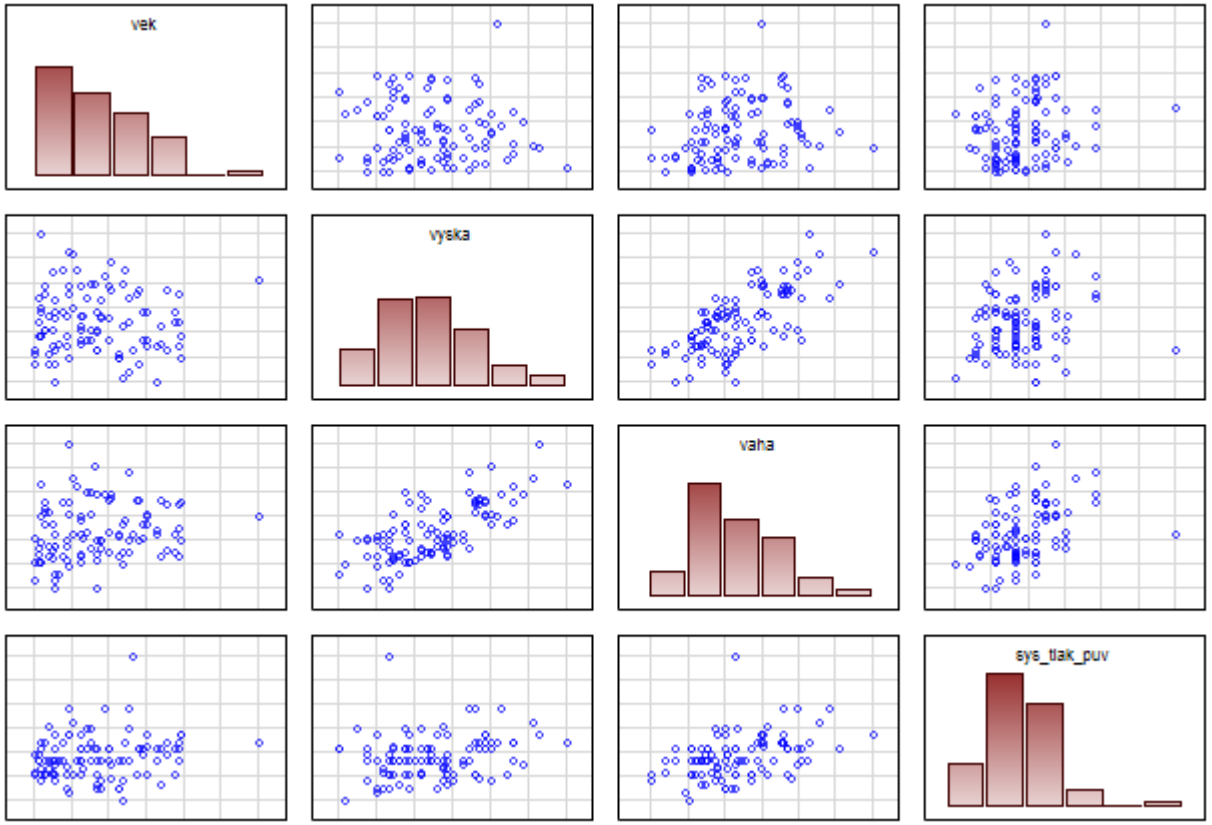
Tečkový graf

- rovněž pro vykreslení vztahu dvou spojitých proměnných
- v softwaru Statistica: Graphs – Scatterplots...



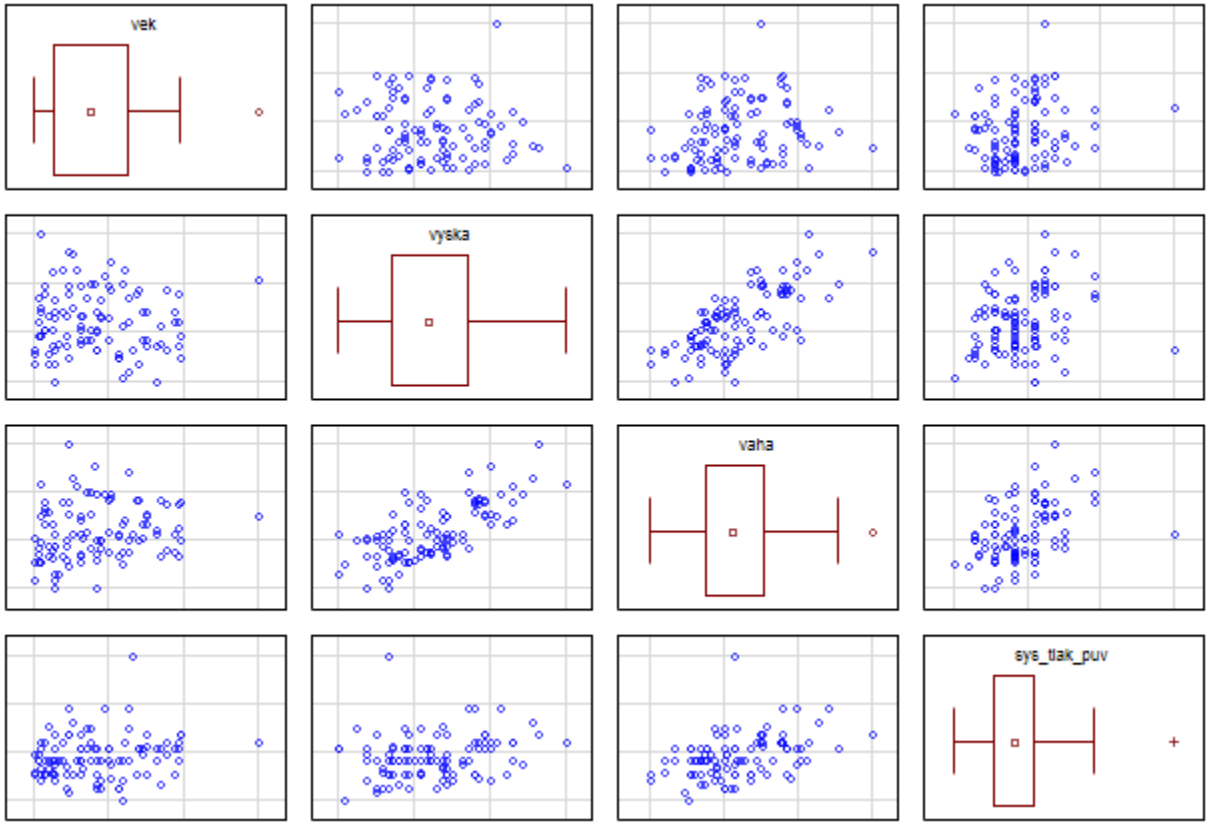
Maticový graf

- vykreslení vztahu více spojitých proměnných
- v softwaru Statistica: Graphs – Matrix Plots...
- upozornění: nastavení, jak se vypořádat s chybějícími hodnotami



Maticový graf – na diagonále krabicové grafy

- v softwaru Statistica: Graphs – Matrix Plots...; na záložce Advanced zatrhnout Display: Box plot

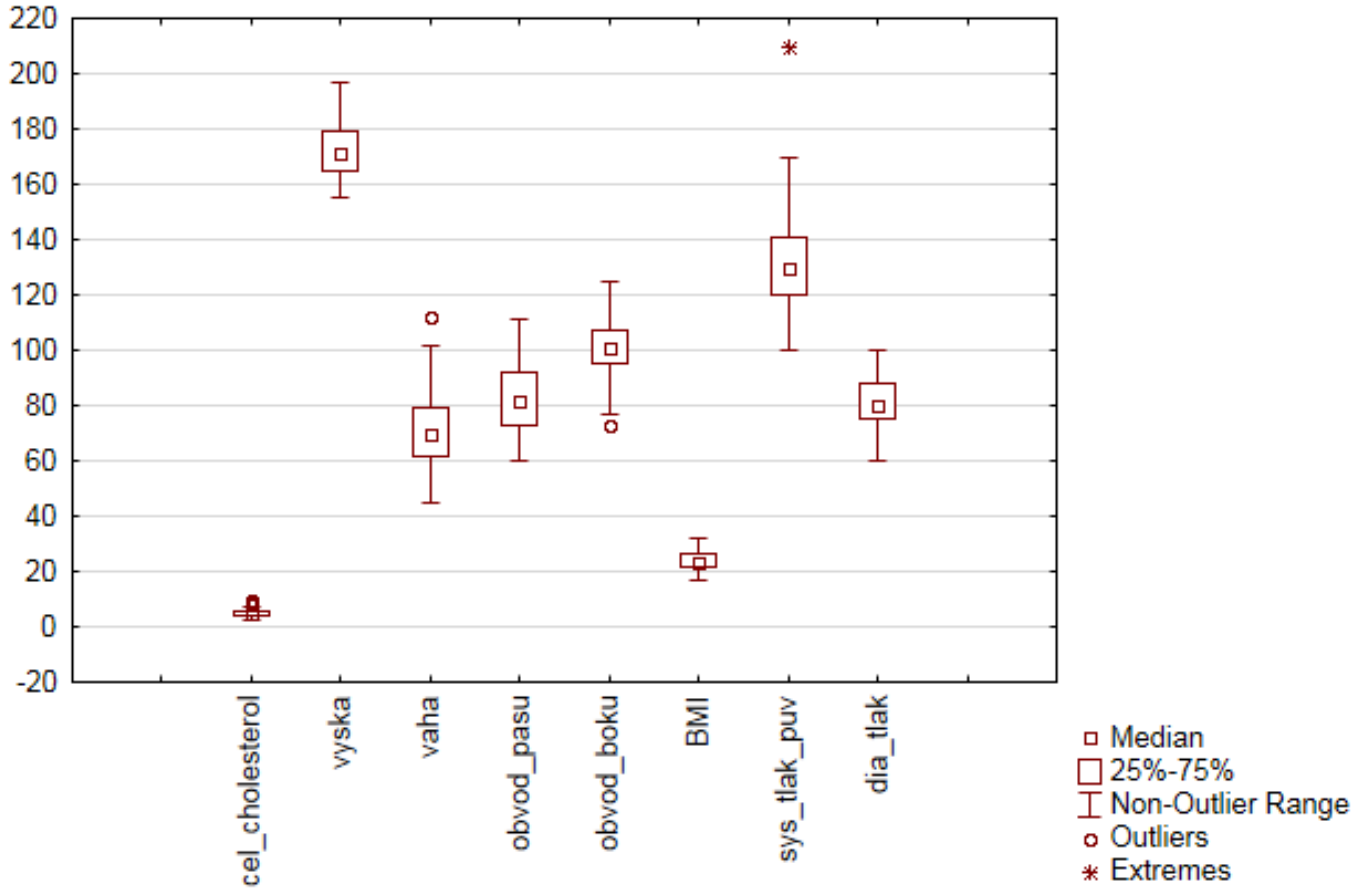


Úkol 2

- vykreslete maticový graf pro proměnné: věk, LDL, HDL i celkový cholesterol, systolický a diastolický tlak, přičemž na diagonále budou krabicové grafy
- změňte barvu krabicového grafu na černou (můžete nastavit i výplň)
- změňte barvu tečkových grafů
- zrušte čáry mřížky u tečkových grafů (gridlines)

Krabicové grafy pro více proměnných

- ukáží nám, zda mají proměnné podobný rozsah hodnot
- v softwaru Statistica: označit příslušné sloupečky v datech – Graphs – Graphs of Block Data – Box Plot: Block columns

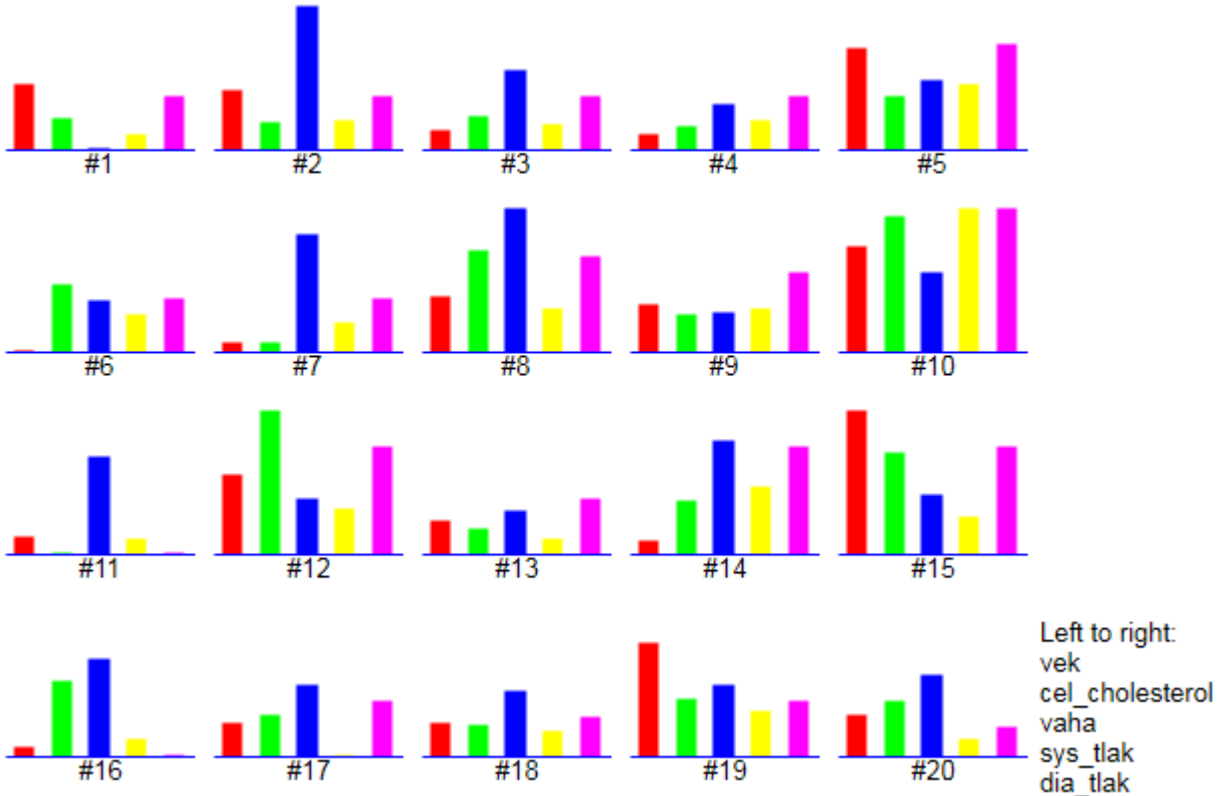


Ikonové (symbolové) grafy

- hodnoty znaků znázorněny jako geometrické útvary či symboly
- každému objektu (subjektu) odpovídá jeden obrazec složený z těchto geometrických útvarů či symbolů
- umožní vizuálně porovnat, které objekty (subjekty) jsou si podobné
- mnoho druhů, v softwaru Statistica např.:
  1. Profilové sloupce
  2. Profily
  3. Paprskové (hvězdicové) grafy
  4. Polygony
  5. Pavučinové grafy
  6. Chernoffovy tváře

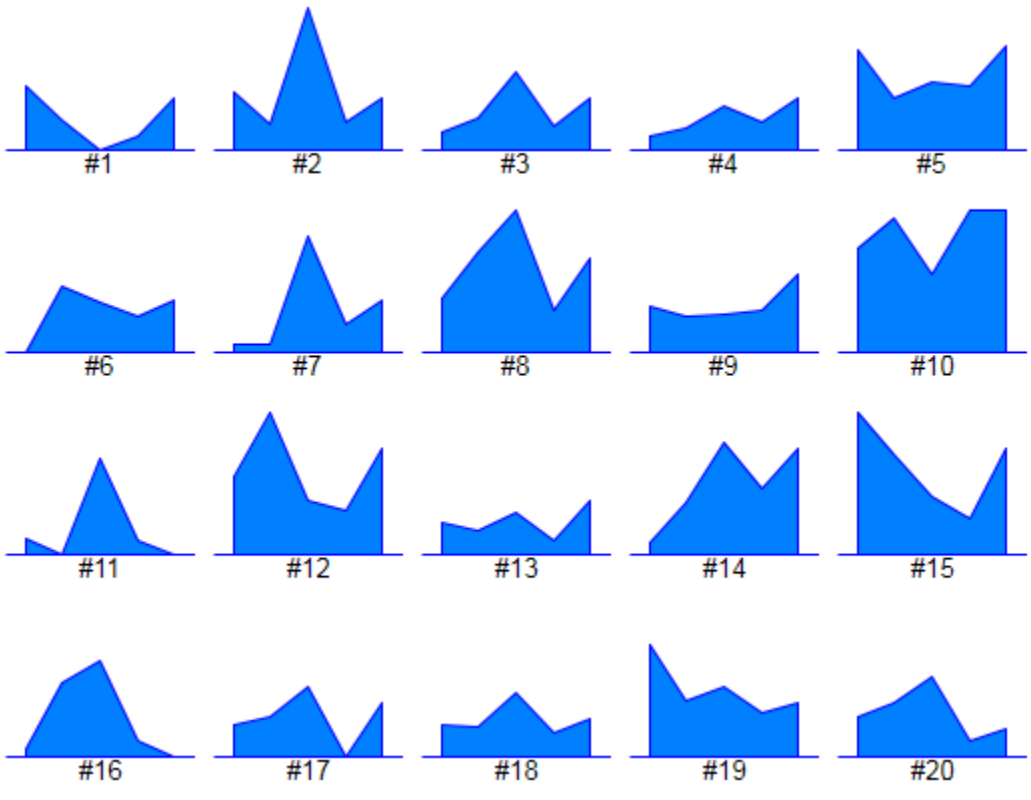
Ikonové grafy – profilové sloupce

- výšky sloupců odpovídají relativním hodnotám proměnných (relativní hodnota je podíl původní hodnoty a maxima z absolutních hodnot dané proměnné)
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Columns** – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



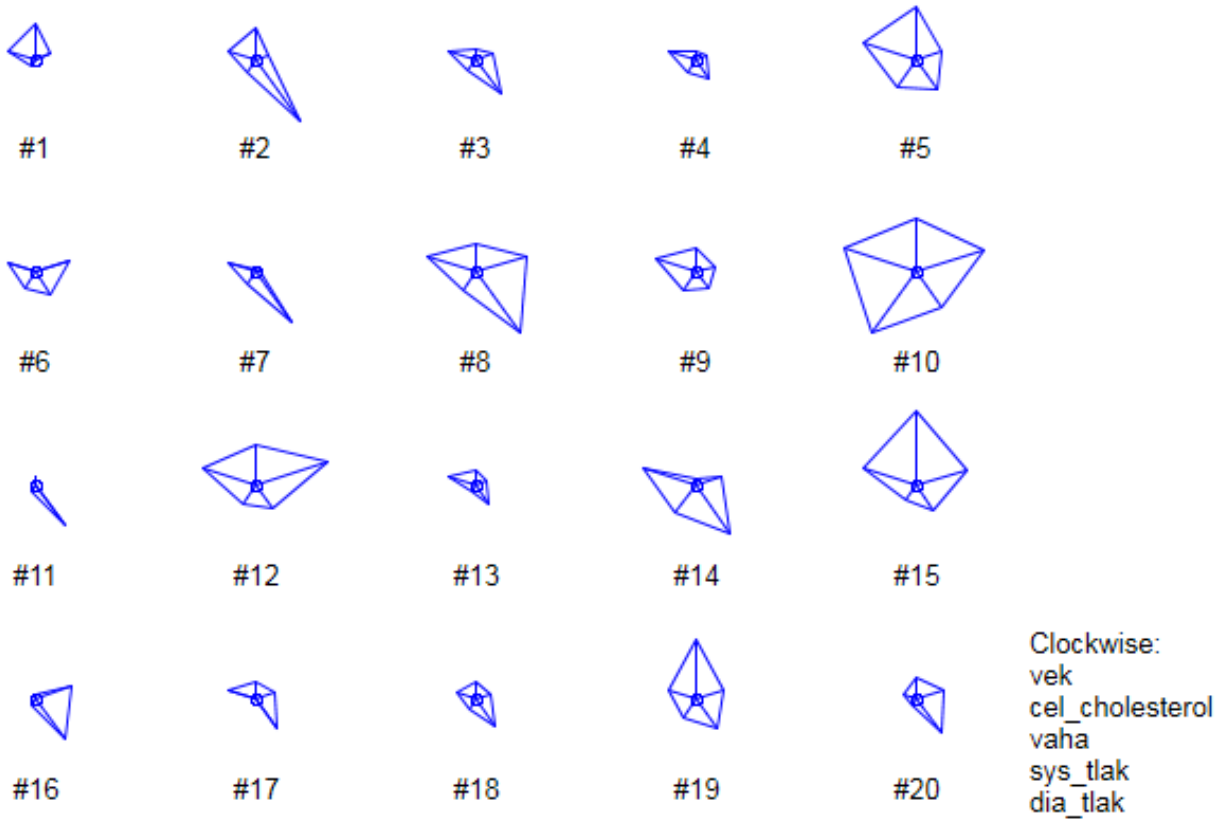
Ikonové grafy – profily

- obdoba profilových sloupců, jen se středy horních hran profilových sloupců spojí úsečkami
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Profiles** – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



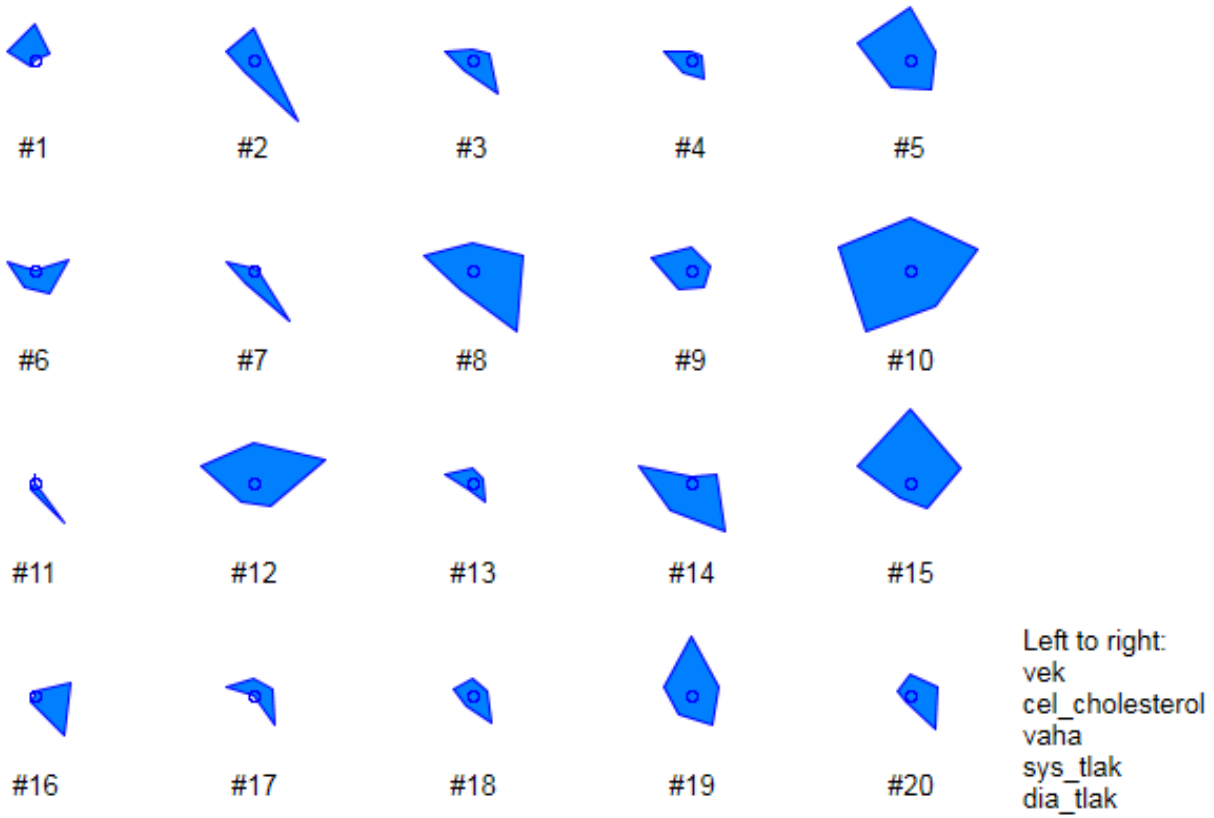
Ikonové grafy – paprskové (hvězdicové) grafy

- vzdálenosti od středu odpovídají relativním hodnotám proměnných
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Stars** – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



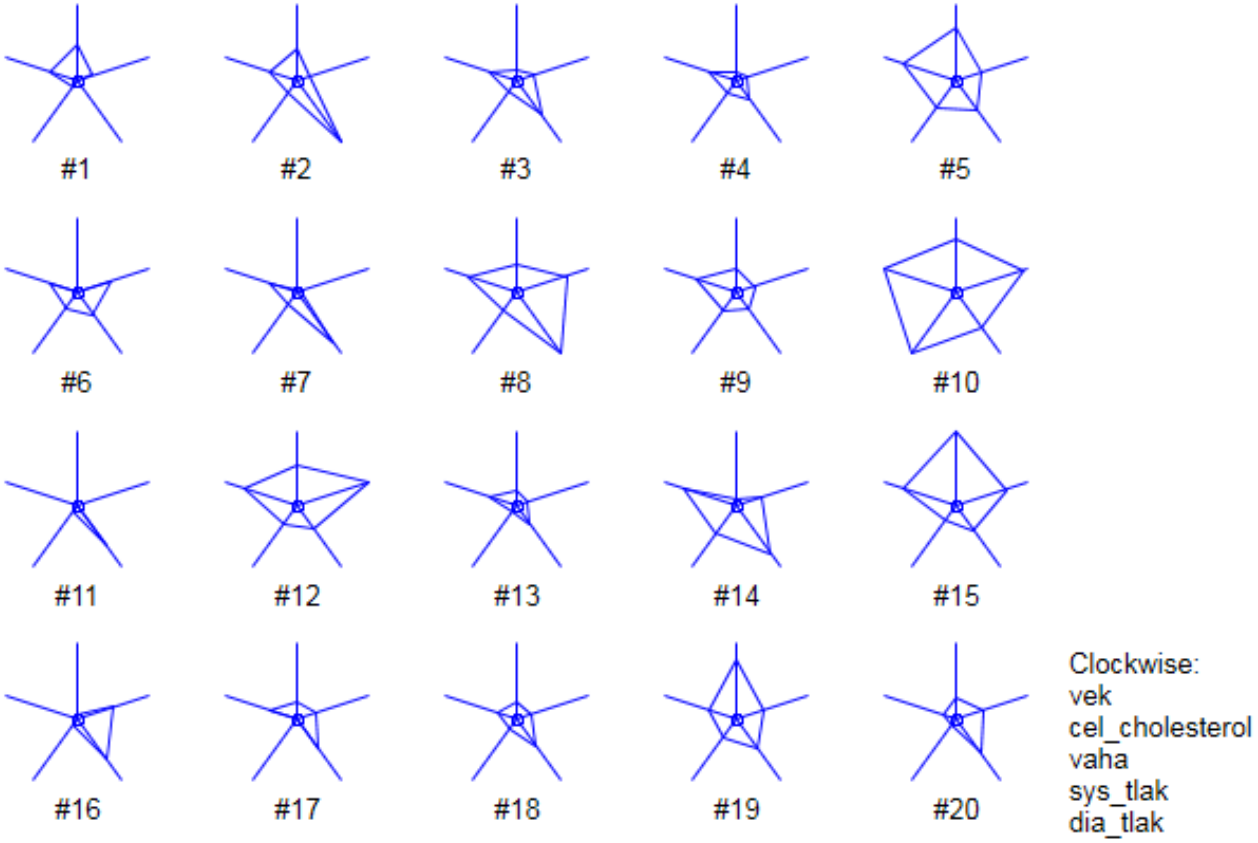
Ikonové grafy – polygony

- obdoba paprskových grafů, jen jsou vyplněné
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Polygons** – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



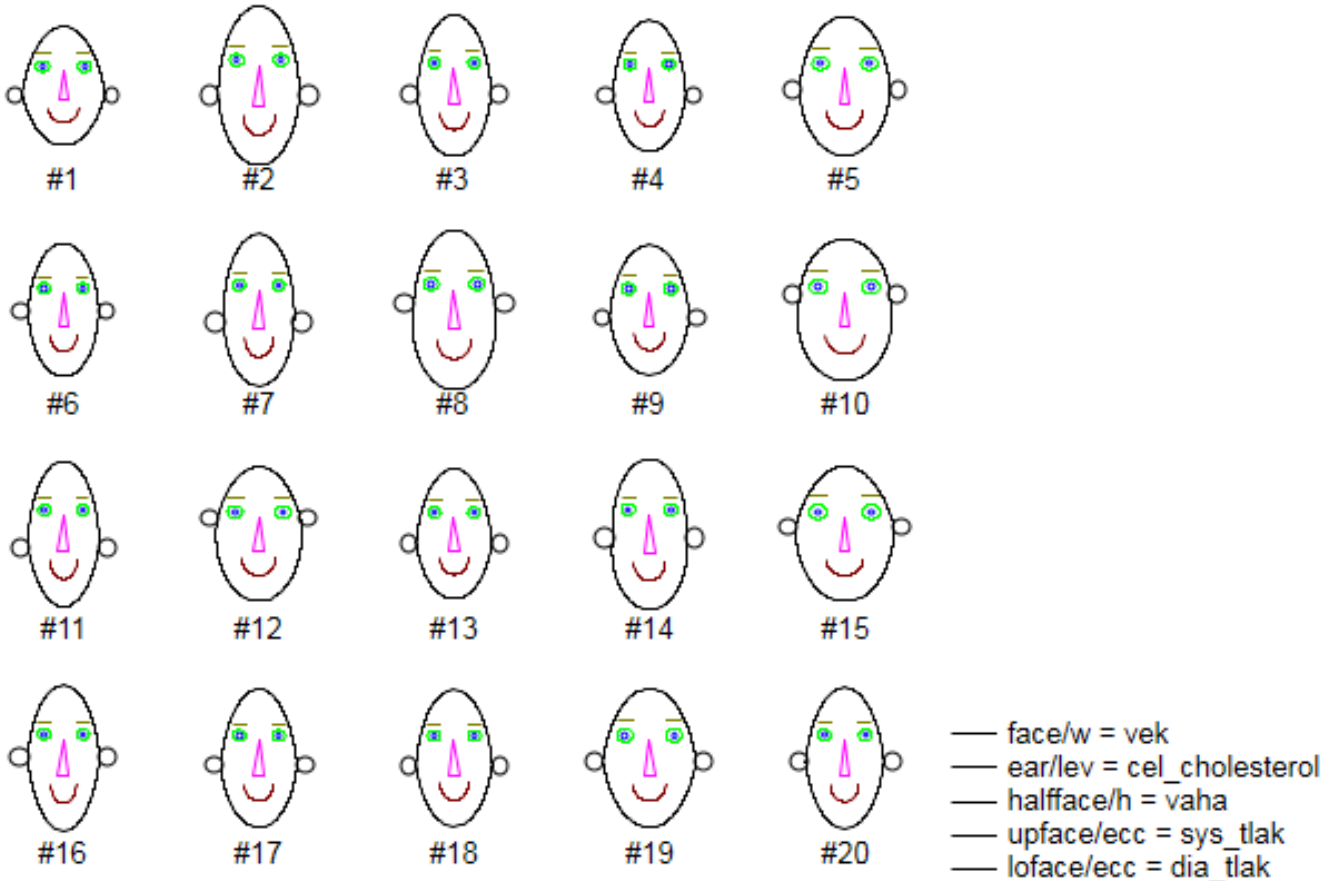
Ikonové grafy – pavučinové grafy

- obdoba paprskových grafů, přidáno znázornění maxima absolutních hodnot
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Sun Rays** – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



Ikonové grafy – Chernoffovy tváře

- proměnné znázorněny jako části obličeje
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Chernoff Faces** – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“

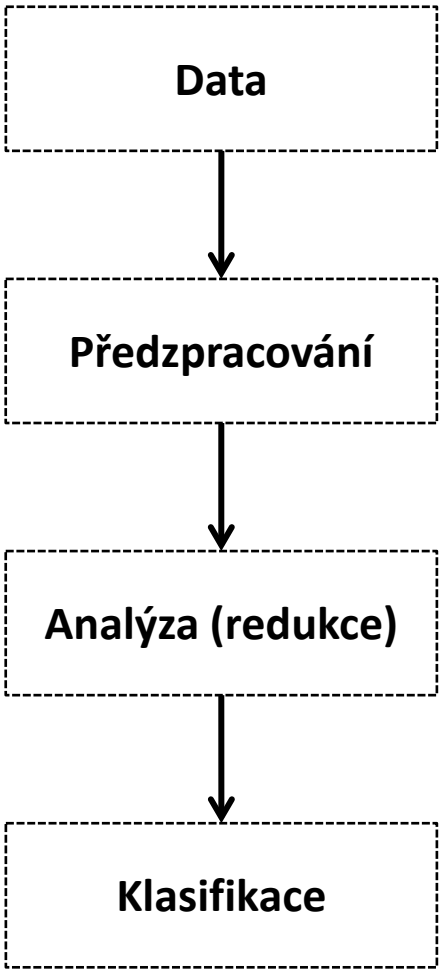


Úkol 3

- zvolte si typ ikonových grafů, které se Vám zdají nejpřehlednější, a vykreslete graf pro subjekty 80 až 100 s využitím proměnných věk, výška, váha, obvod pasu a boků a BMI

Analýza a klasifikace dat

Schéma analýzy a klasifikace dat



	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M		90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

?

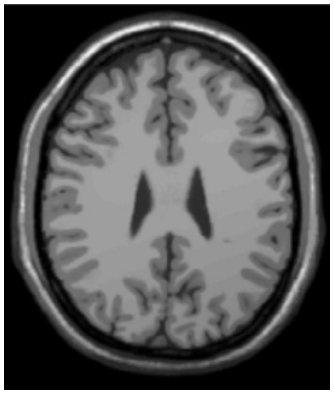
nebo

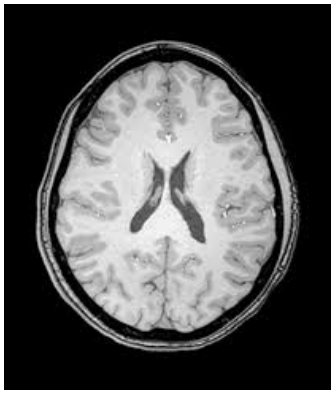
?

Proč používat klasifikaci dat?

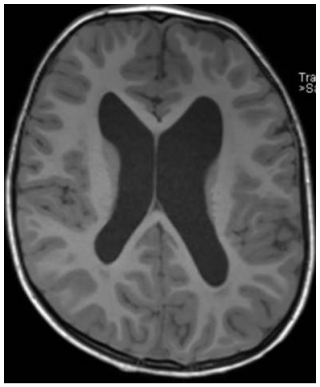
1. Podpora diagnostiky onemocnění mozku (Alzheimerova choroba, schizofrenie atd.):

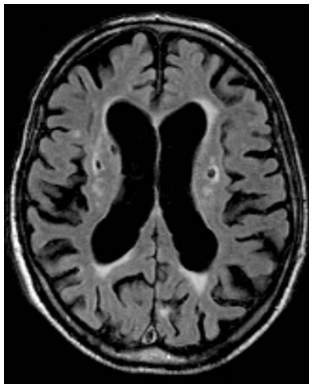
Zdravé subjekty



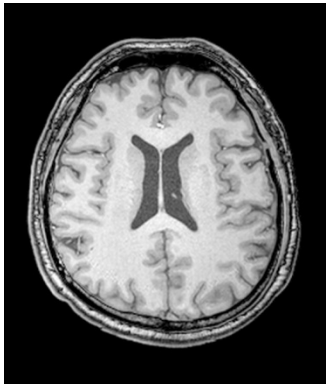


Pacienti





Nový subjekt

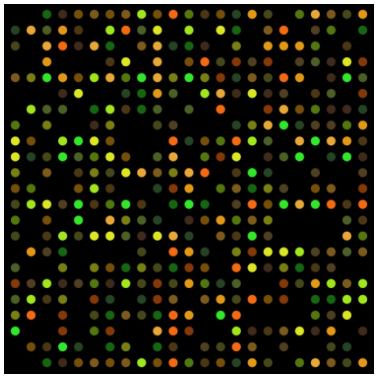


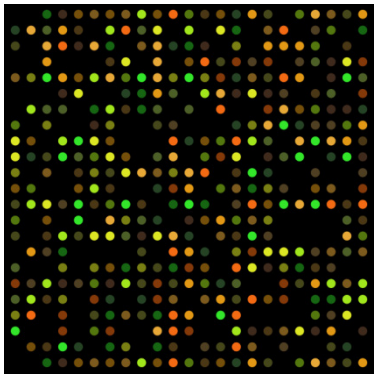
Pacient? x Zdravý?

Proč používat klasifikaci dat?

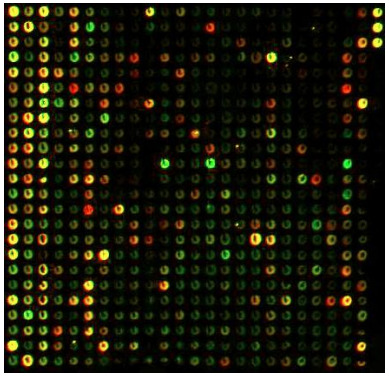
2. Odhalení genetického onemocnění na základě dat s microarray experimentů:

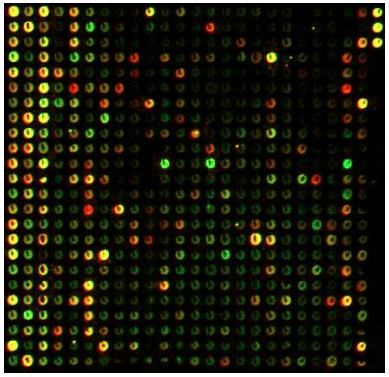
Zdravé subjekty



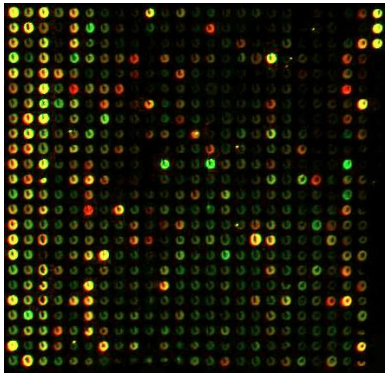


Pacienti





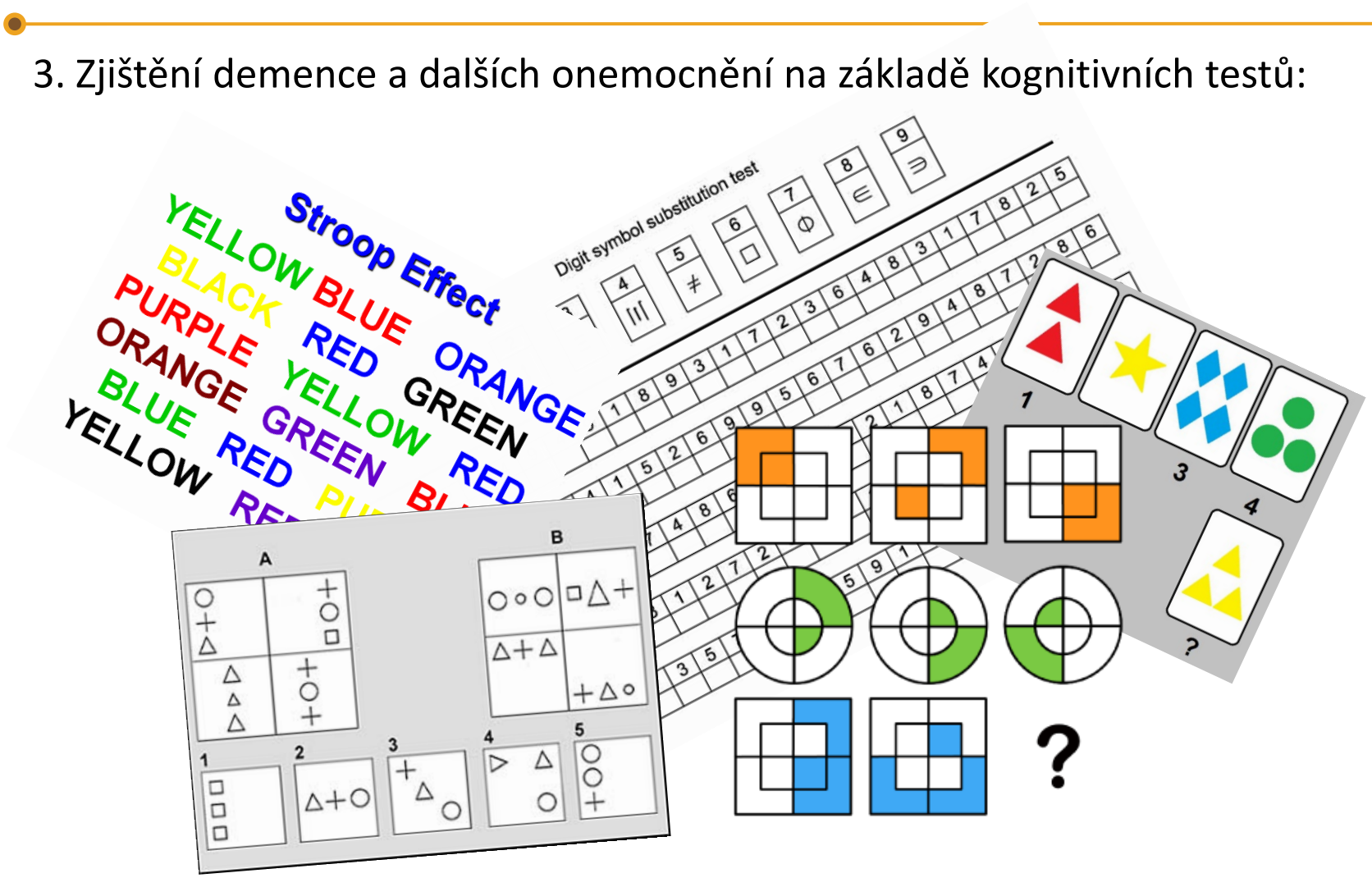
Nový subjekt



Pacient? x Zdravý?

Proč používat klasifikaci dat?

3. Zjištění demence a dalších onemocnění na základě kognitivních testů:



Demence ano? x Demence ne?

Proč používat klasifikaci dat?

4. Rozpoznání hmyzu:

Nejedovaté housenky

Jedovaté housenky

?

Jedovatá nebo nejedovatá housenka?

Proč používat klasifikaci dat?

5. Rozpoznání vadných výrobků:

Matičky bez vady

Matičky s vnitřní prasklinou

?

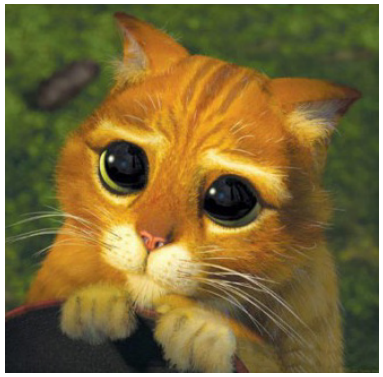
Matička bez vady nebo s vnitřní prasklinou?

Proč používat klasifikaci dat?

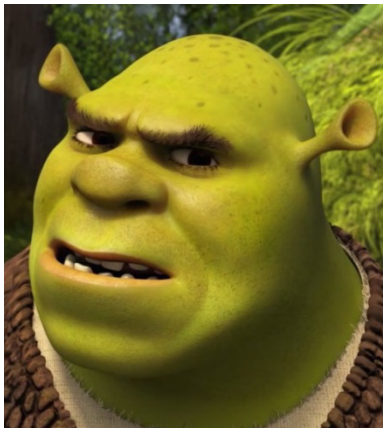
6. Rozpoznání tváře při vstupu do zabezpečené budovy:

Nemá přístup do budovy

Má přístup do budovy



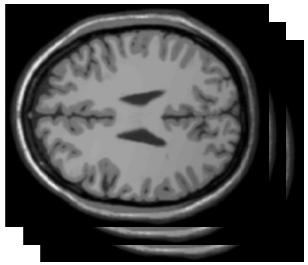
?



Dostane se do budovy: ano? x ne?

Proč používat redukci dat?

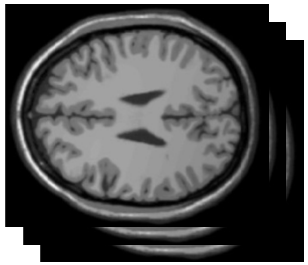
Obrazová data



Klasifikace

Proč používat redukci dat?

Obrazová data



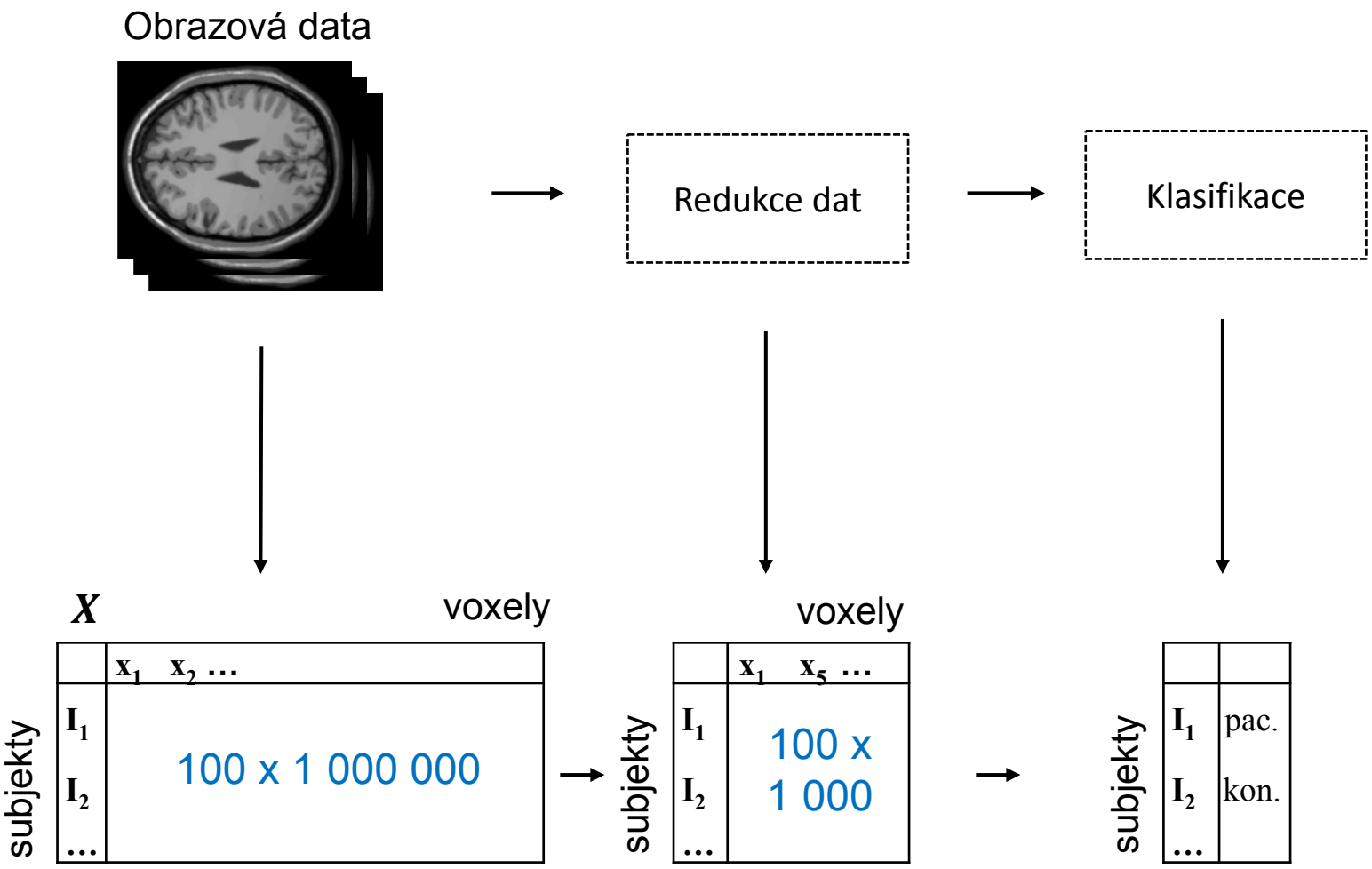
Klasifikace

X voxely

subjekty	x_1 x_2 …
I_1	100 x 1 000 000
I_2	
…	

subjekty	
I_1	pac.
I_2	kon.
…	

Proč používat redukci dat?

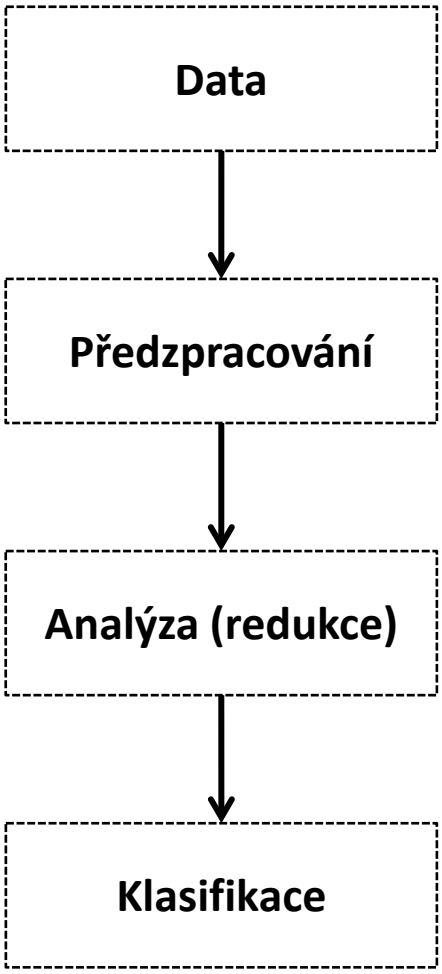


Cíle analýzy a klasifikace dat - shrnutí

- **rozhodnutí o typu či charakteru objektu** – např. že daná rostlina je pomněnka lesní (*Myosotis sylvatica*), zvíře že je medvěd hnědý (*Ursus arctos*), nebo že daná budova je vystavěna v renesančním slohu – ***klasifikační,*** resp. ***rozpoznávací úloha***;
- **posouzení kvality stavu analyzovaného objektu** – např. zda je pacient v pořádku, nebo má infarkt myokardu, cirhózu jater, apod. – opět ***klasifikační,*** resp. ***rozpoznávací úloha***;
- **rozhodnutí o budoucnosti objektu** – např. zda lze pacienta léčit a vyléčit, zda les po 20 letech odumře, jaké bude sociální složení obyvatelstva na daném území a v daném čase – ***klasifikační*****,** *resp.* ***predikční úloha***

- poznámka: v některých oblastech se pojem predikce a klasifikace rozlišuje:
  - pojem ***klasifikace*** je používán, použije-li se klasifikačního algoritmu pro známá data; pokud jsou data nová, pro která apriori neznáme klasifikační třídu, pak hovoříme o ***predikci*** klasifikační třídy
  - pojem ***klasifikace*** používáme, pokud vybíráme identifikátor klasifikační třídy z určitého diskrétního konečného počtu možných identifikátorů; pokud určujeme (predikujeme) spojitou hodnotu, např. pomocí regrese, pak hovoříme o ***predikci***, i když tento pojem nemá časovou dimenzi

Schéma analýzy a klasifikace dat



	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M		90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

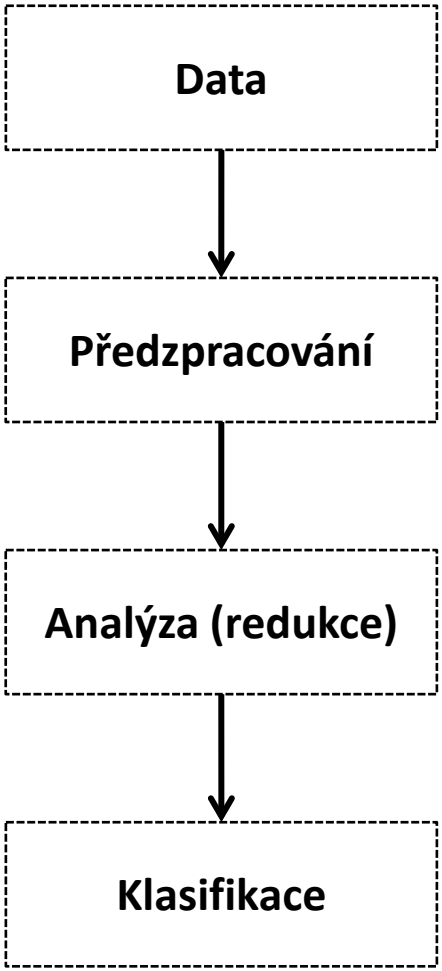
	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

?

nebo

?

Schéma analýzy a klasifikace dat



- rekonstrukce a doplnění chybějících údajů;
- filtrace rušivých x zvýraznění užitečných složek dat;
- konverze typu dat (A/Č převod);

- určení a výběr hodnot příznaků (reprezentativních parametrů) – pro příznakové klasifikátory;
- nalezení primitiv (charakteristických tvarových segmentů) – strukturální klasifikátory;
- redukce dat

- zatřídění do skupin (tříd, kategorií)

Analýza dat

- **Analýza** (z řečtiny – *rozbor, rozčlenění*) je vědecká metoda založená na dekompozici celku na elementární části. Cílem analýzy je identifikovat podstatné a nutné vlastnosti elementárních částí celku, poznat jejich podstatu a zákonitosti.
- opakem je **syntéza** – označení pro proces spojení dvou nebo více částí do jednoho celku (následuje po analýze – spojíme znalosti dohromady; např. lékaři dělají syntézu výsledků, které jim pošlou statistici)
- ze statistického pohledu: analýza = celý proces zpracování dat

- především ve smyslu redukce dat:
  - výběr proměnných z předem zvolené množiny proměnných
  - vyjádření původních proměnných pomocí menšího počtu skrytých (tzv. latentních) nezávislých proměnných

- poznámka: není ale jisté, že celková množina proměnných bude obsahovat všechny podstatné proměnné – musíme se spolehnout na experty

Klasifikace versus diskriminační analýza

- **klasifikace –** rozdělení (konkrétní či teoretické) dané skupiny (množiny) objektů na konečný počet dílčích skupin (podmnožin), v nichž všechny objekty mají dostatečně podobné společné vlastnosti. Předměty (jevy), které mají podobné uvažované vlastnosti tvoří třídu (skupinu).
- **diskriminační analýza –** hledá vztah mezi kategoriální proměnnou a množinou vzájemně vázaných proměnných; je to podskupina klasifikačních metod

- poznámka: analýza a klasifikace dat občas nazývána souhrnně jako „rozpoznávání obrazů“ (*pattern recognition*), „dolování z dat“ (*data mining*) či „strojové učení“ (*machine learning*)

Typy klasifikátorů – podle reprezentace vstupních dat

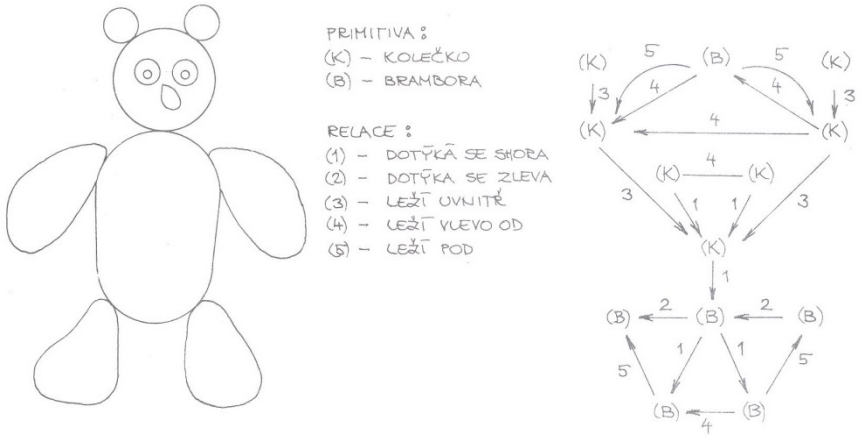
1. **Podle reprezentace vstupních dat:**
   - **příznakové klasifikátory:** paralelní x sekvenční
   - **strukturální (syntaktické) klasifikátory**
   - **kombinované klasifikátory**
2. **Podle jednoznačnosti zařazení do skupin:**
   - **deterministické klasifikátory**
   - **pravděpodobnostní klasifikátory**
3. **Podle typů klasifikačních a učících algoritmů:**
   - **parametrické klasifikátory**
   - **neparametrické klasifikátory**
4. **Podle způsobu učení:**
   - **učení s učitelem:** dokonalým x nedokonalým
   - **učení bez učitele**
5. **Podle podle principu klasifikace:**
   - **klasifikace pomocí diskriminačních funkcí**
   - **klasifikace pomocí vzdálenosti od etalonů klasif. tříd**
   - **klasifikace pomocí hranic v obrazovém prostoru**

Typy klasifikátorů – podle reprezentace vstupních dat

- **příznakové** – vstupní data vyjádřena vektorem hodnot jednotlivých proměnných (příznaků):
  - paralelní – zpracování vektoru jako celku (např. Bayesův klasifikátor)
  - sekvenční – zpracování (občas i měření) proměnných postupně (např. klasifikační stromy)

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

- **strukturální (syntaktické)** – vstupní data popsána relačními strukturami



- **kombinované** – jednotlivá primitiva doplněna příznakovým popisem

Typy klasifikátorů – dle jednoznačnosti zařazení do skupin

- **deterministické klasifikátory:**
  - každý objekt musí patřit do nějaké třídy a nemůže být současně ve více třídách
  - pozn. použití termínu „**deterministický klasifikátor**“ v případě, že klasifikátor daná data zpracuje vždy se stejným výsledkem (např. Bayesův klasifikátor) x „**nedeterministický klasifikátor**“, který může při opakovaném zpracování daných dat klasifikovat různě (např. neuronové sítě – záleží na tom, jaká bude inicializace)

- **pravděpodobnostní klasifikátory:**
  - stanoví pravděpodobnost zařazení obrazů do daných klasifikačních tříd
  - např. člověk má s pravděpodobností 0,6 infarkt, s pstí 0,3 má atrofii srdeční komory a s pstí 0,1 je zdravý

Typy klasifikátorů – dle typů klasifikačních a učících algoritmů

- **parametrické klasifikátory:**
  - potřeba nastavit či určit parametry
  - např. prahová klasifikace (potřeba stanovit práh), metoda podpůrných vektorů (potřeba stanovit parametr „C“) atd.

- **neparametrické klasifikátory:**
  - není potřeba nastavovat žádné parametry
  - např. klasifikace podle vzdáleností od reprezentativního objektu (tzv. „etalonu“) skupin

- pozn. z tohoto pohledu jsou klasifikační stromy parametrické klasifikátory, pokud to však hodnotíme ze statistického pohledu, jsou to neparametrické metody, protože nemají předpoklad normálního rozdělení

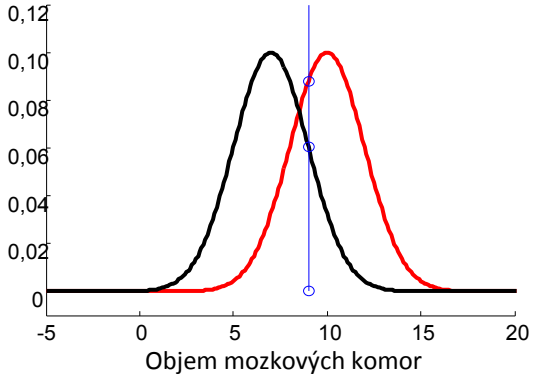
Typy klasifikátorů – podle způsobu učení

- **učení s učitelem** – k dispozici trénovací množina, u níž známe zařazení každého objektu do jednotlivých klasifikačních tříd
  - učení s dokonalým učitelem – učitel se nemůže splést (tzn. předpokládáme, že všechny trénovací objekty jsou správně označené, že patří do dané třídy)
  - učení s nedokonalým učitelem – připouštíme, že v trénovací množině mohou být nesprávně označené subjekty (např. u některých duševních onemocnění se lékař může splést a označit pacienta za schizofrenika, i když trpí bipolární poruchou, což se však prokáže až za několik let, takže v naší trénovací množině je takto špatně zařazený subjekt)

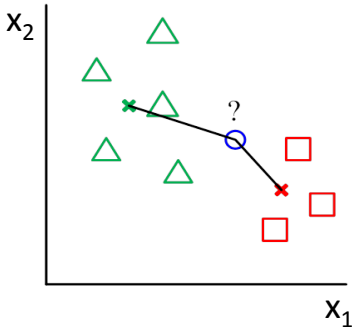
- **učení bez učitele:**
  - trénovací množina není k dispozici a často ani předem neznáme, jaké třídy (skupiny) se v datech budou vyskytovat
  - typickým příkladem je shlukování

Typy klasifikátorů – podle principu klasifikace

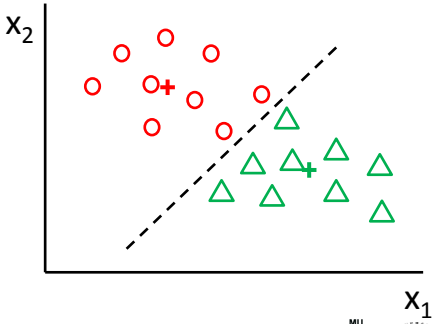
- **klasifikace pomocí diskriminačních funkcí:**
  - diskriminační funkce určují míru příslušnosti k dané klasifikační třídě
  - pro danou třídu má daná diskriminační funkce nejvyšší hodnotu



- **klasifikace pomocí vzdálenosti od etalonů klasif. tříd:**
  - etalon = reprezentativní objekt(y) klasifikační třídy
  - počet etalonů klasif. třídy různý – od jednoho vzorku (např. centroidu) po úplný výčet všech objektů dané třídy (např. u klasif. pomocí metody průměrné vazby)



- **klasifikace pomocí hranic v obrazovém prostoru:**
  - stanovení hranic (hraničních ploch) oddělujících klasifikační třídy



Příprava nových učebních materiálů pro obor Matematická biologie

je podporována projektem OPVK

č. CZ.1.07/2.2.00/28.0043

„Interdisciplinární rozvoj studijního oboru Matematická biologie“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ