Pokročilé metody analýzy dat v neurovědách

RNDr. Eva Koriťáková, Ph.D.
doc. RNDr. Ladislav Dušek, Dr.

Přínos kurzu

- orientace v principech vícerozměrné analýzy dat s důrazem na zpracování medicínských dat, a to především z neurovědního výzkumu
- schopnost zvolit a aplikovat adekvátní metodu analýzy a klasifikace dat k dosažení požadovaných výsledků
- schopnost správné interpretace dosažených výsledků včetně vyhodnocení úspěšnosti klasifikace
- schopnost praktické analýzy dat v software MATLAB, STATISTICA, SPSS či R

Osnova kurzu

1. Úvod do pokročilé vícerozměrné analýzy dat:
   - význam, cíle a příklady využití vícerozměrné analýzy dat
   - vícerozměrná data a jejich tabulkové a grafické zpracování
2. Vícerozměrné statistické testy a rozložení:
   - vícerozměrný průměr, kovarianční matice, matice korelačních koeficientů
   - vícerozměrný t-test, vícerozměrná analýza rozptylu
   - transformace a jiné úpravy vícerozměrných dat
3. Podobnosti a vzdálenosti ve vícerozměrném prostoru:
   - metriky pro určení vzdálenosti
   - metriky pro určení podobnosti a asociační matice
4. Shluková analýza:
   - shluková analýza hierarchická – hierarchické aglomerativní shlukování, hierarchické divizivní shlukování
   - shluková analýza nehierarchická
   - identifikace optimálního počtu shluků

Osnova kurzu – pokračování

5. Ordinační analýzy I:
   - principy redukce dimenzionality dat; selekce a extrakce proměnných
   - analýza hlavních komponent (PCA), faktorová analýza (FA)
6. Ordinační analýzy II:
   - analýza nezávislých komponent (ICA), korespondenční analýza (CA), vícerozměrné škálování (MDS), redundanční analýza (RDA), kanonická korelační analýza (CCorA)
7. Klasifikace I:
   - principy a cíle klasifikace
   - diskriminační analýza pomocí diskriminačních fcí, minimální vzdálenosti a pomocí hranic – Fisherova LDA
8. Klasifikace II:
   - metoda podpůrných vektorů (SVM), přehled dalších klasifikačních metod
   - hodnocení úspěšnosti klasifikace

Požadavky ke kolokviu

- Předmět je ukončen kolokviem sestávajícím se z teoretických otázek a analýzy praktických příkladů na počítači.

- Je nutné porozumět probíraným tématům a umět aplikovat vícerozměrné statistické metody při analýze reálných datových souborů.

Doporučená literatura

- Koriťáková, E. et al.: online výukové materiály Vícerozměrné metody pro analýzu a klasifikaci dat http://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--vicerozmerne-metody-pro-analyzu-dat

- DUDA, R.O. et al. *Pattern Classification*. New York: Wiley-Interscience,, 2000, 680 pp.
- BISHOP, C. *Pattern Recognition and Machine Learning*. New York: Springer, 2006, 738 pp.
- FLACH, P.A. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press, 2012, 396 pp.
- CHUNG, M.K. *Statistical and computational methods in brain image analysis*. Boca Raton: CRC Press, 2014, 400 s.
- KUNCHEVA, L.I. *Combining Pattern Classifiers: Methods and Algorithms*. New Jersey: Wiley-Interscience,, 2004, 376 pp.
- JOHNSON, R. et al. *Applied multivariate statistical analysis*. 6th ed. Upper Saddle River, N.J.: Prentice Hall, 2007, 773 pp.

- MELOUN, M. et al. *Statistická analýza vícerozměrných dat v příkladech*. Praha: Academia, 2012, 750 s.
- EVERITT, B. et al. *An introduction to applied multivariate analysis with R*. New York: Springer, 2011, 273 pp.
- JAMES, G. et al. *An introduction to statistical learning: with applications in R*. New York: Springer, 2013, 426 pp.
- THEODORIDIS, S. et al. *Introduction to pattern recognition: a MATLAB approach*. Amsterdam: Academic Press, 2010, 219 pp.

Blok 1

Úvod do pokročilé vícerozměrné analýzy dat

Osnova

1. Význam, cíle a příklady využití vícerozměrné analýzy dat
2. Vícerozměrná data, jejich popis a vizualizace
3. Předzpracování dat

Význam, cíle a příklady využití vícerozměrné analýzy dat

Význam a cíle vícerozměrné analýzy dat

- většina dat pořízených při výzkumu jsou data vícerozměrná – chceme zjistit celou řadu vlastností daných subjektů či objektů

PROMĚNNÉ (VLASTNOSTI)

SUBJEKTY	ID	Pohlaví	Věk	Váha	MMSE skóre	Objem hipokampu	…
	1	muž	84	85,5	29	7030	
	2	žena	25	62,0	28	6984	
	…						

- zpravidla nestačí analyzovat každou proměnnou zvlášť – pro úplné pochopení vztahů většinou potřeba analyzovat proměnné současně

→ použití VÍCEROZMĚRNÝCH METOD

Význam a cíle vícerozměrné analýzy dat II

- vícerozměrné metody umožňují:
  - znázornit a popsat vícerozměrná data
  - zjišťovat vztahy mezi jednotlivými proměnnými a mezi subjekty (resp. objekty)

- mnoho způsobů dělení vícerozměrných metod do skupin – např. dělení podle cíle, kterého chceme vícerozměrnou analýzou dosáhnout:

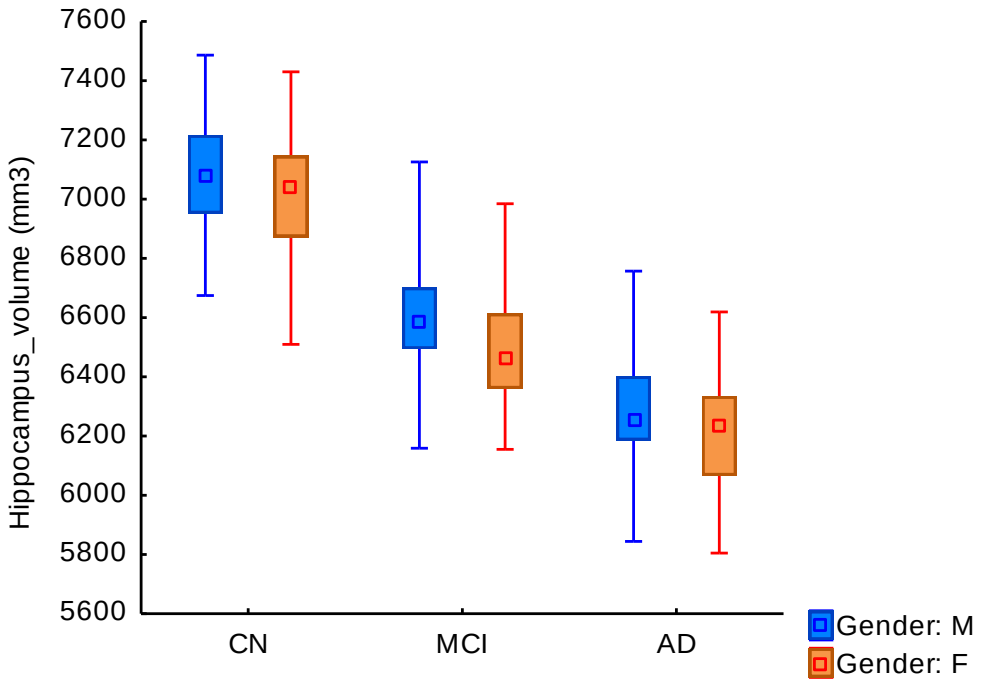
1. Testování hypotéz o vícerozměrných datech
2. Vytvoření shluků subjektů, objektů nebo proměnných
3. Redukce vícerozměrných dat
4. Klasifikace subjektů či objektů

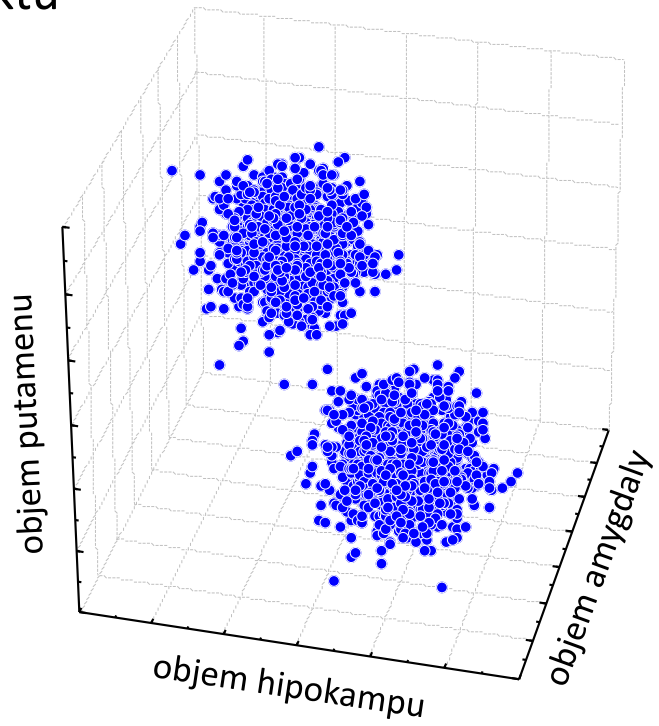
Cíle vícerozměrné analýzy dat

1. Testování hypotéz o vícerozměrných datech

Příklady:

- výzkum vztahu pohlaví a typu onemocnění na objem hipokampu
- zjištění, zda je rozdílná spotřeba elektrické energie ve městech a na vesnicích během týdne a o víkendu
- ověření, zda objem hipokampu, amygdaly a putamenu dokáže odlišit pacienty se schizofrenií od zdravých subjektů



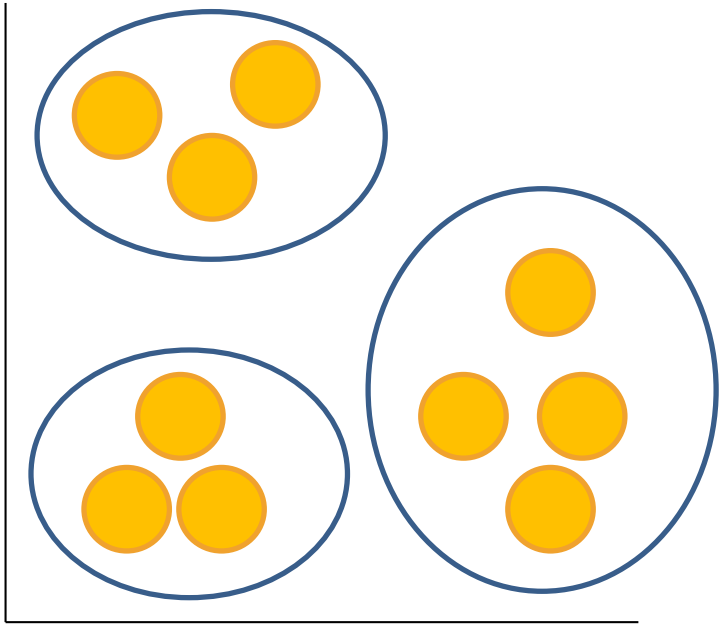


Cíle vícerozměrné analýzy dat

2. Vytvoření shluků subjektů, objektů nebo proměnných

Příklady:

- vytvoření skupin diagnóz onemocnění s podobnými léčebnými náklady
- vytvoření skupin lokalit podle výskytu určitých druhů rostlin a živočichů
- vytvoření skupin genů a subjektů na základě dat genové exprese
- vytvoření skupin subjektů se schizofrenií podle kognitivních skóre a neurologických parametrů



Cíle vícerozměrné analýzy dat

3. Redukce vícerozměrných dat

Příklady:

- vytvoření souhrnného skóre odpovědi pacientů na radioterapii z původních několika proměnných
- vytvoření menšího počtu nových proměnných z původních dat, které nám umožní znázornit vícerozměrná data ve 2-D či 3-D grafech
- výběr oblastí mozku, které nejvíce odlišují pacienty s neuropsychiatrickým onemocněním od zdravých subjektů

	A	B	C	D	E	F	G	H	I	J	K
1	ID	Group	Gender	Age	MMSE	Hippocampus_	Amygdala_	Thalamus_	Pallidum_	Putamen_	Nucl_caud_
2	101	1	M	84	28	6996	2725	12800	3914	11227	3528
3	102	1	F	76	29	7187	2916	12277	3606	11236	3773
4	103	1	M	79	30	7030	2835	12906	3638	11430	4294
5	104	1	F	89	30	7263	2919	12432	3678	11018	3585
6	105	1	F	71	30	6867	2887	12383	3689	11304	3723
7	106	1	F	70	30	7331	3081	12415	3553	11372	3969
8	107	1	F	88	30	6705	2823	12575	4150	11303	2886
9	108	1	F	86	28	6586	2860	12454	3945	11328	3741
10	109	1	F	84	29	7036	3017	12361	3827	11382	3737

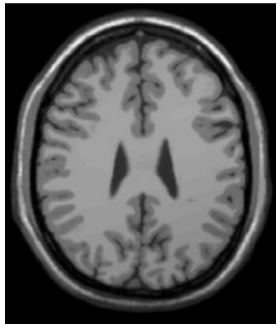
Cíle vícerozměrné analýzy dat

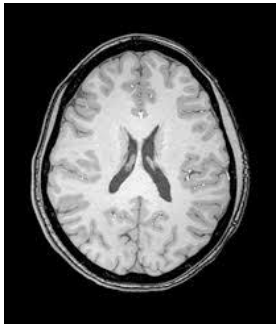
4. Klasifikace subjektů či objektů

Příklady:

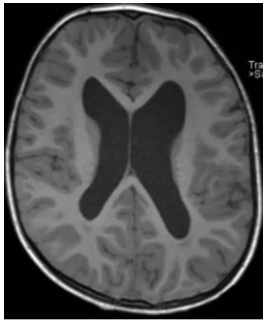
- zjištění (diagnostika) schizofrenie na základě kognitivních testů
- rozhodnutí, zda banka poskytne či neposkytne hypotéku danému subjektu na základě jeho příjmů, rodinné situace atd.
- diagnostika demence (tzn. zařazení nového subjektu do skupiny pacientů či kontrol) podle obrázku mozku

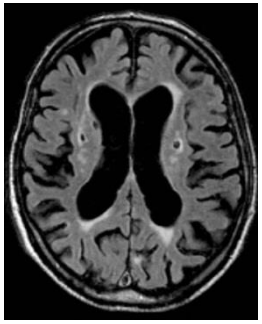
Zdravé subjekty



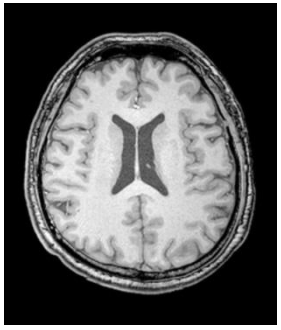


Pacienti





Nový subjekt



Pacient? x Zdravý?

Vícerozměrná data, jejich popis a vizualizace

Vícerozměrná data

PROMĚNNÉ

OBJEKTY (SUBJEKTY)

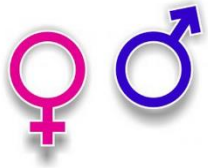
ID	Pohlaví	Věk	Váha	MMSE skóre	Objem hipokampu	…
1	muž	84	85,5	29	7030	
2	žena	25	62,0	28	6984	
3						
4						
…						

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

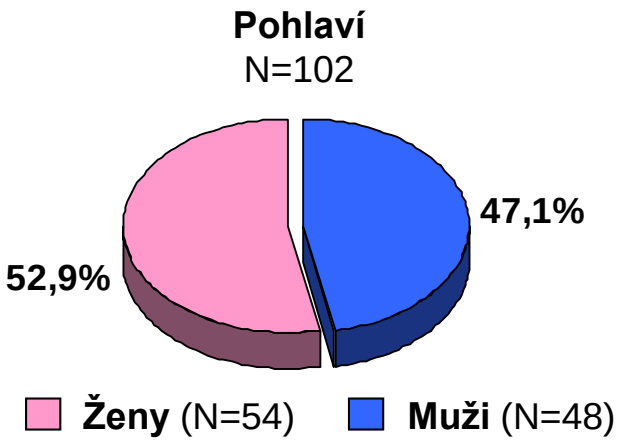
  - Ordinální data

- **Kvantitativní data:**
  - Intervalová data

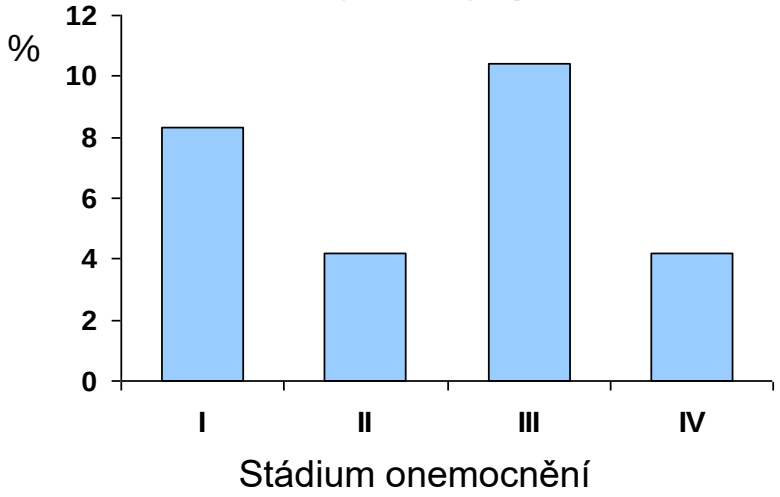
  - Poměrová data

Vizualizace jednorozměrných dat - opakování

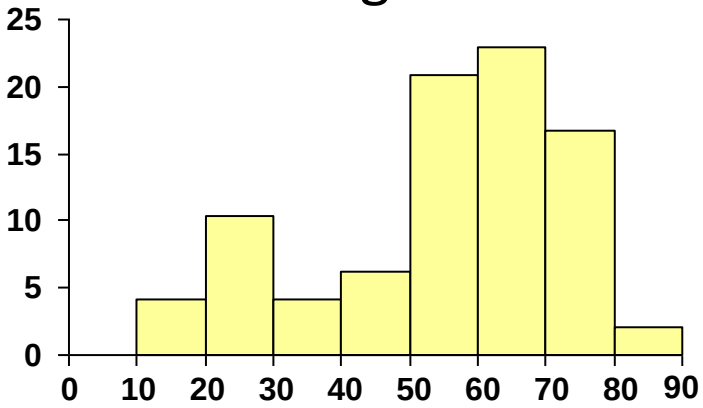
Koláčový graf



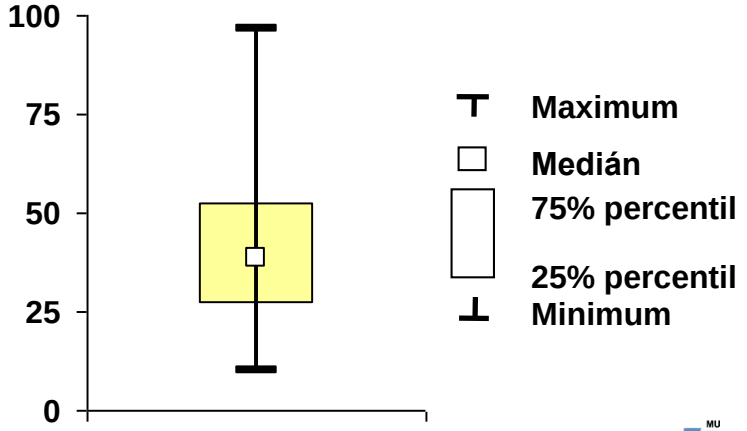
Sloupkový graf



Histogram



Krabicový graf (Box Plot)



K čemu nám může pomoci vizualizace dat?

→ odhalení problémů v datech

id	vek	pohlavi	cholesterol	vyska	vaha	obvod_pasu	obvod_boku	BMI	sys_tlak	dia_tlak
1	38	Z	4.6	164	45	60	87	16.7	120	80
2	36	Z	4.35	167	90	97	112	32.3	130	80
3	26	Z		178	70	72	94	22.1	127	80
4	25	Z	4.2	165	59	65	92	21.7	130	80
5	47	M	5.65	158		92	96	26.8	155	90
6	21	Z	6.35	172	61	69	98	20.6	135	80
7	23	Z	3.45	170	82	92	113	28.4	130	80
8	35	M	7.99	179	90	101	110	28.1	140	88
9	33	Z	4.88	167	57	70	92	20.4	140	85
10	48	Z	9.56	164	70	93	107	26.0	250	97
11	25	M	3.1	186	75	81	102	21.7	120	70
12	41	Z	10	167	62	71	101	22.2	140	90
13	29	ZZ	4.2	165	58	66	98	21.3	120	80
14	24	M	5.62	174	80	92	107	26.4	156	90
15	58	Z	7.9	164	63	73	100	23.4	135	90

Chybné hodnoty

Chybějící hodnoty

Odlehlé hodnoty

→ k vytvoření představy, jaké výsledky analýzy máme asi očekávat

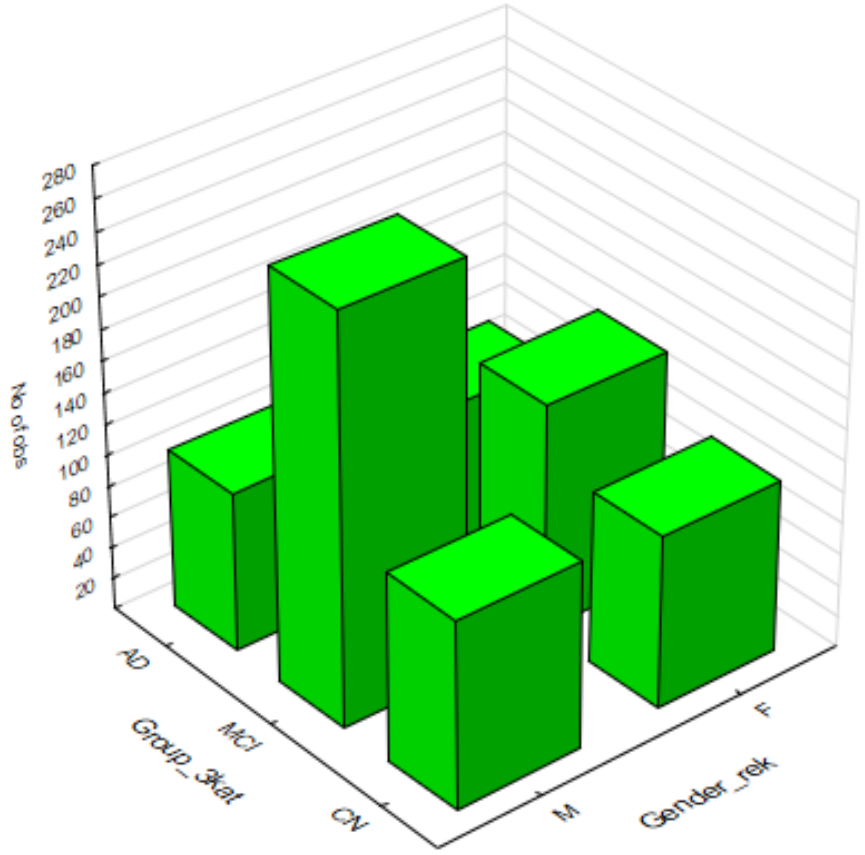
→ ke zjištění vztahu mezi proměnnými, ...

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

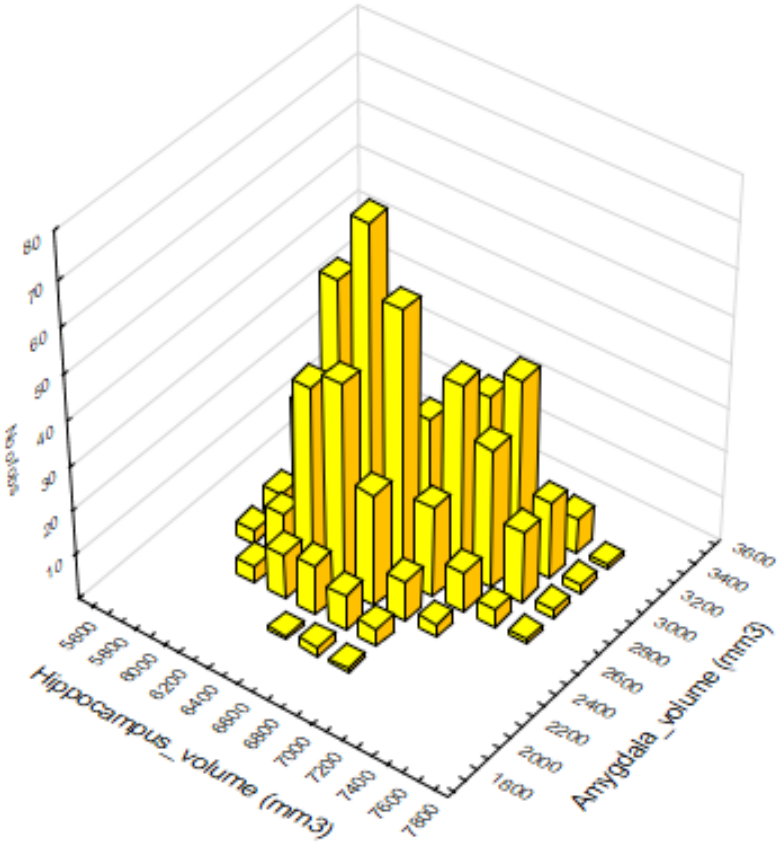
3D sloupkové grafy

- 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...

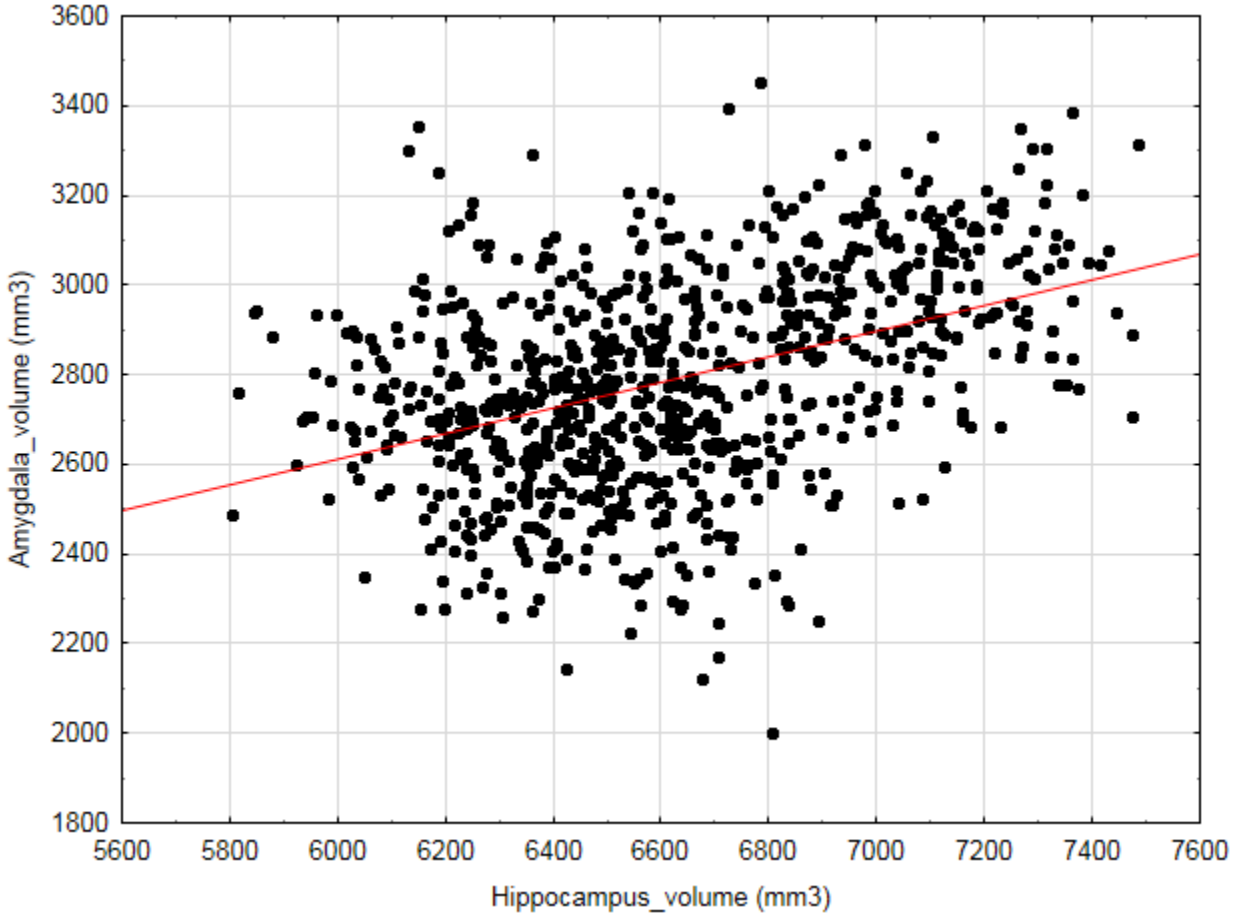


Úkol 1

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

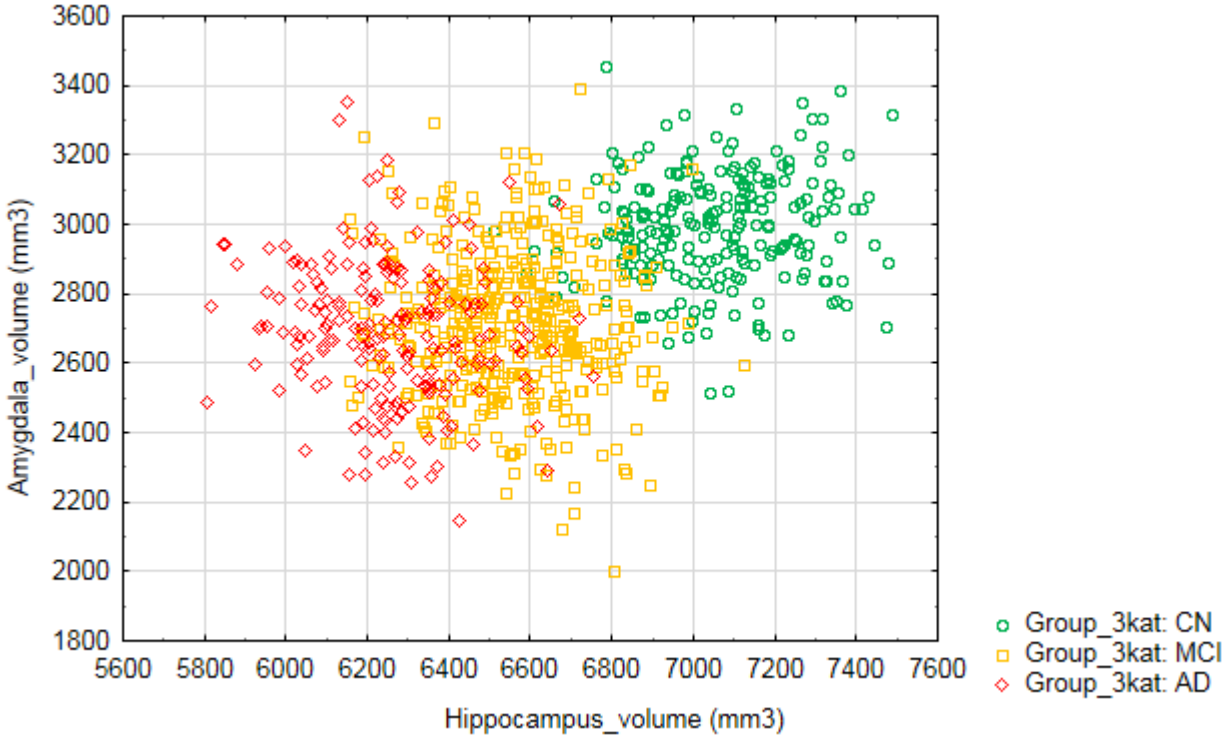
Tečkový graf

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



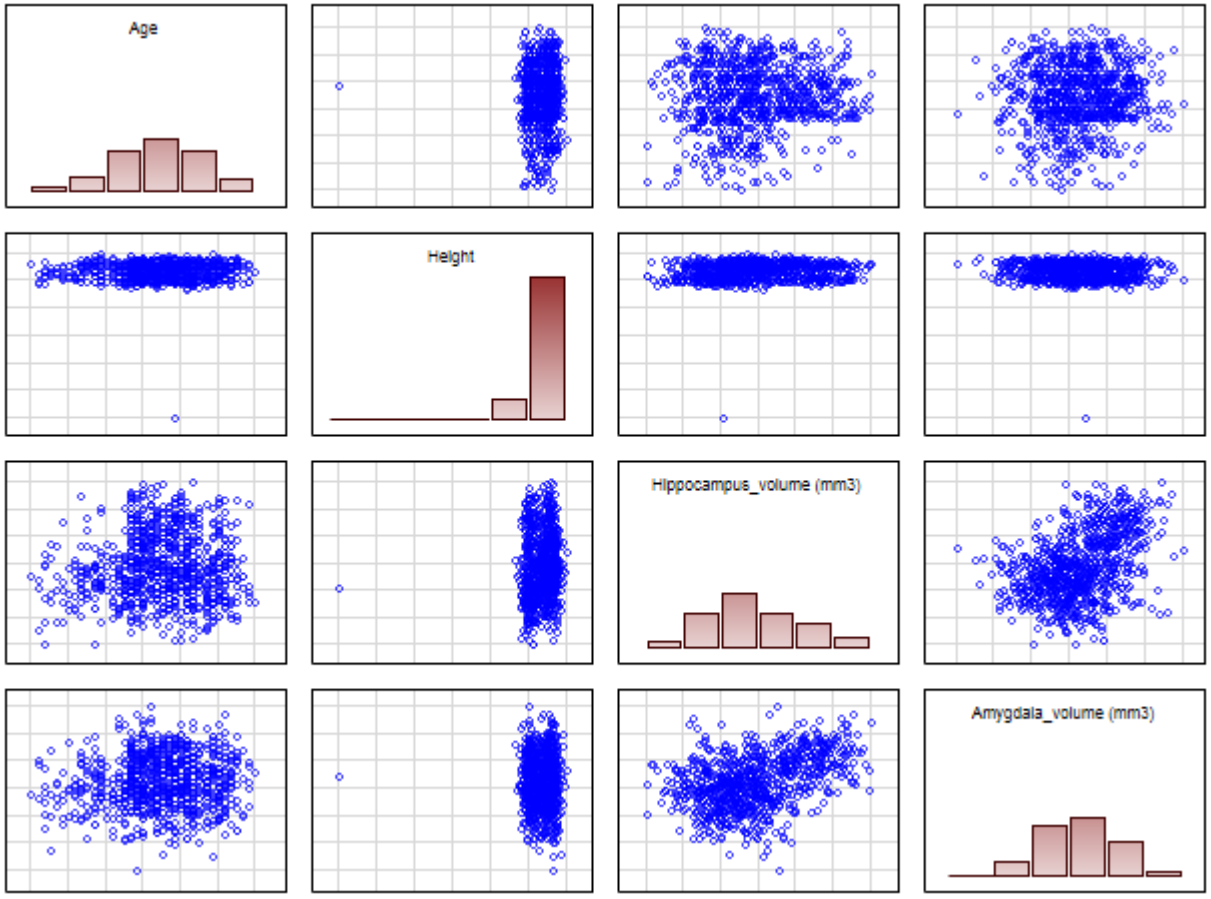
Tečkový graf – přidání kategoriální proměnné

- zahrnutí kategoriální proměnné do grafu použitím různých symbolů či barev pro jednotlivé skupiny určené danou kategoriální proměnnou
- v softwaru Statistica: Graphs – Scatterplots – na záložce Categorized zahrnout On u X-Categorized, vybrat kategoriální proměnnou pomocí Change Variable a změnit Layout na Overlaid



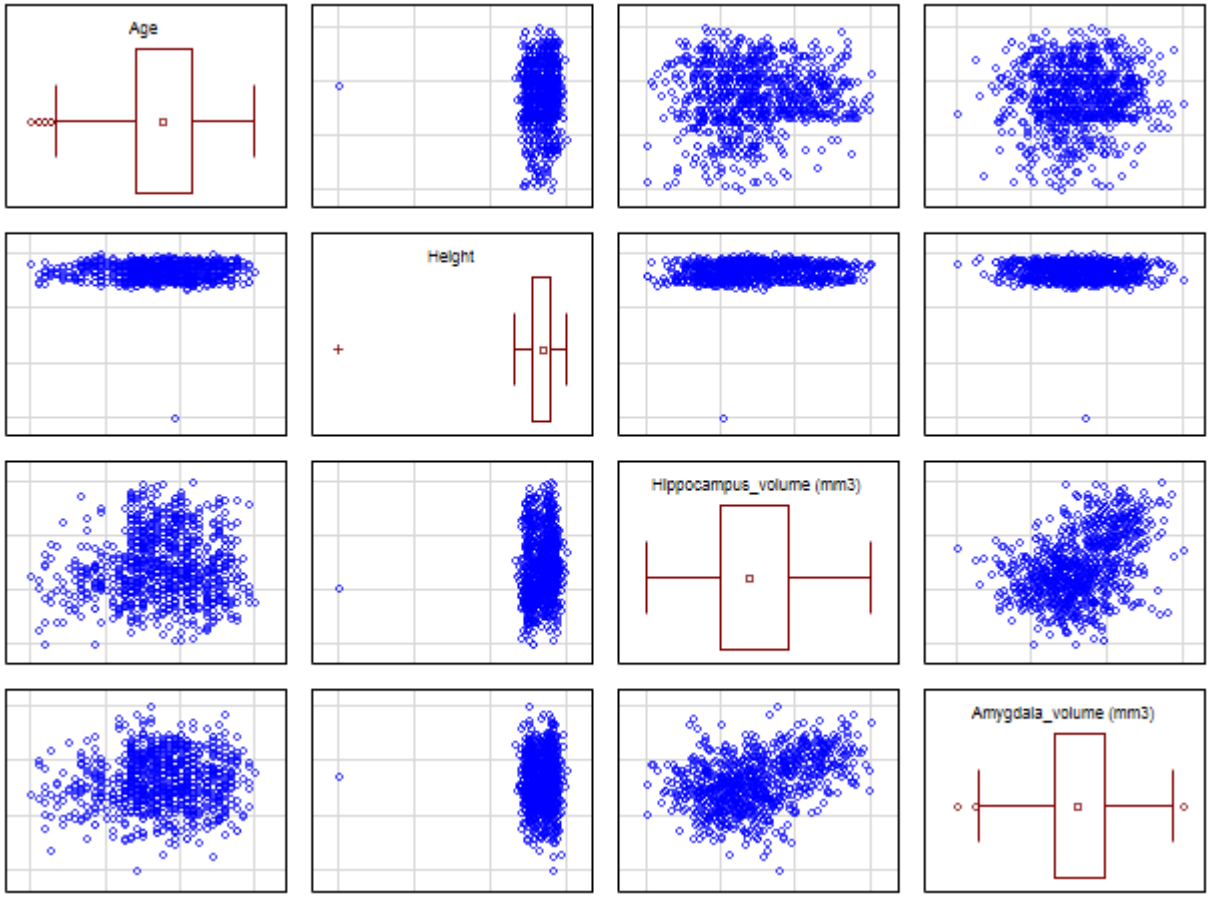
Maticový graf

- 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

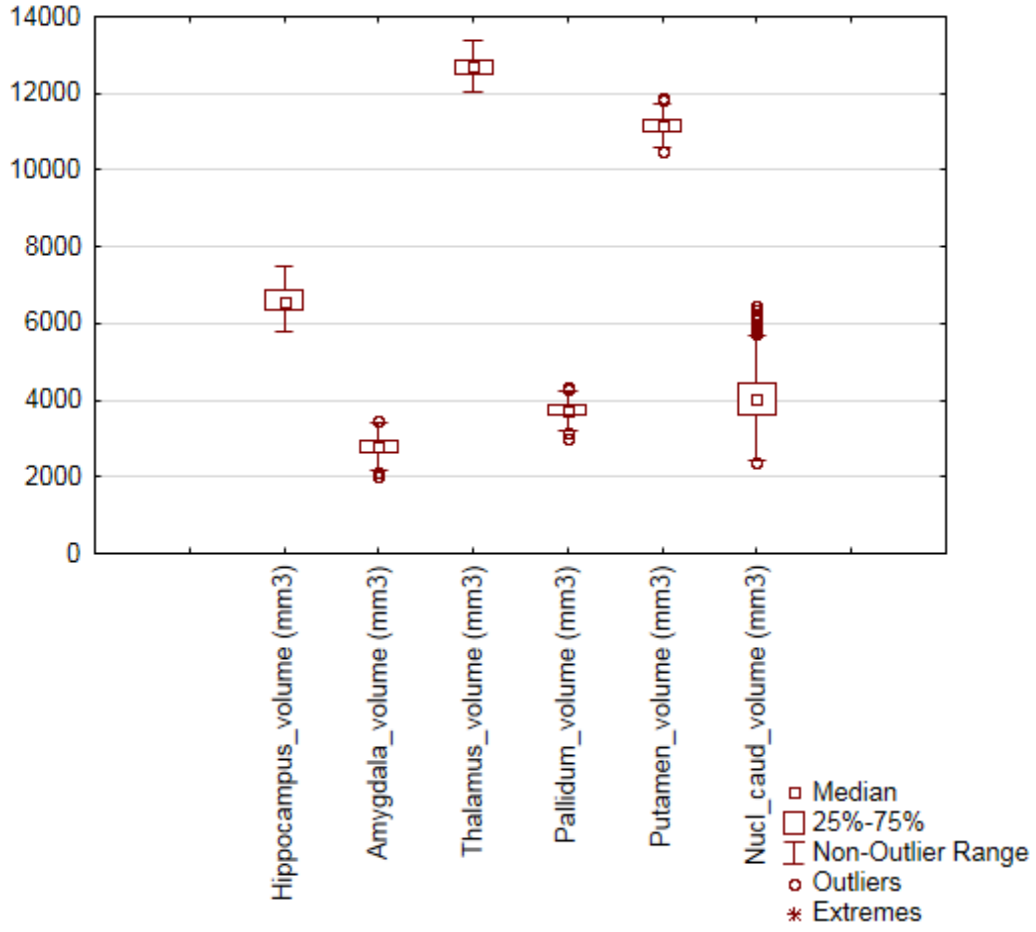
- vykreslete maticový graf pro proměnné: objem hipokampu, amygdaly, thalamu, pallida a putamenu, přičemž na diagonále budou krabicové grafy
- změňte barvu krabicového grafu na černou (můžete nastavit i výplň)
- změňte barvu tečkových grafů
- zrušte čáry mřížky u tečkových grafů (gridlines)

Krabicové grafy pro více proměnných

- ukáží nám, zda mají proměnné podobný rozsah hodnot

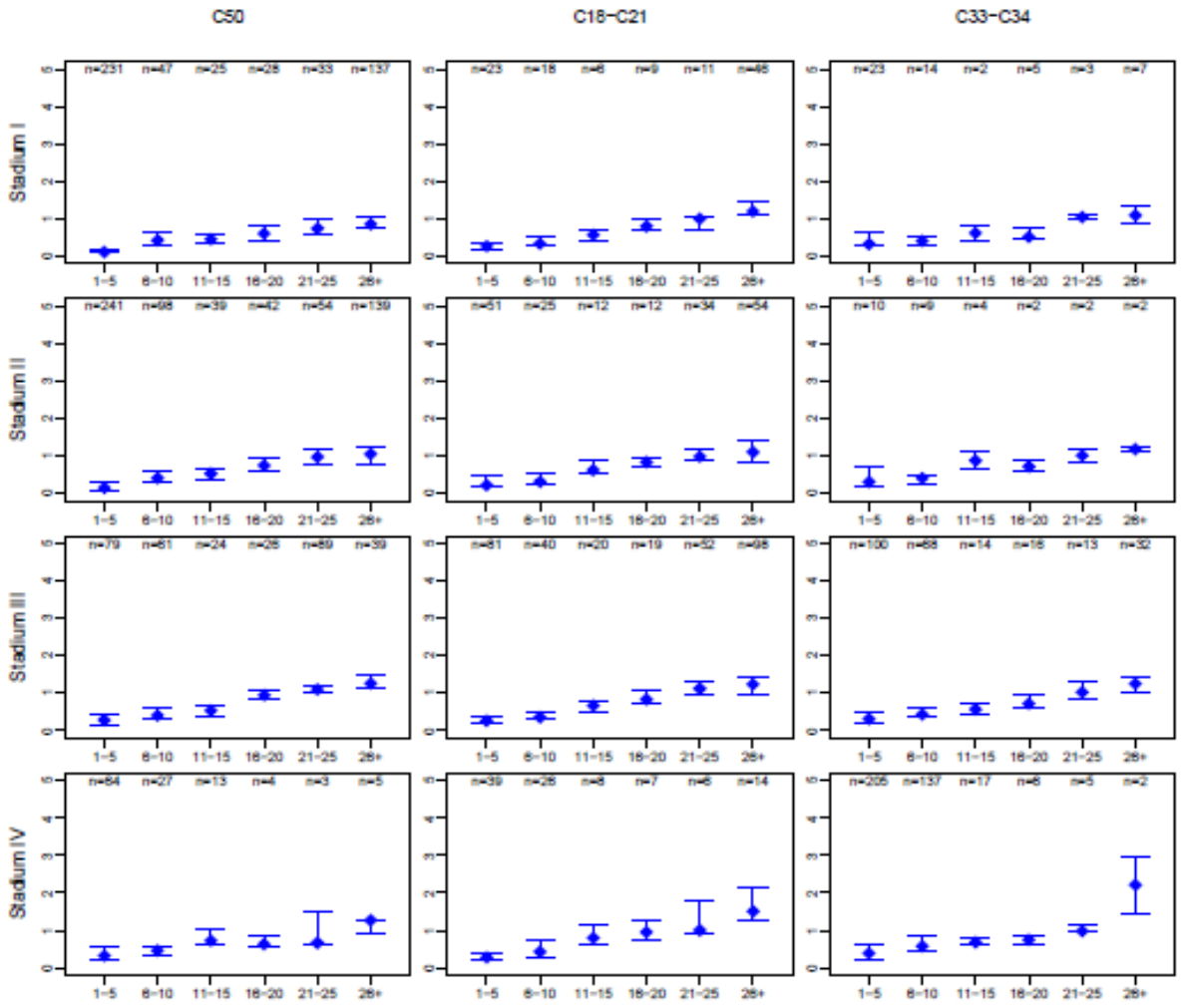
V softwaru STATISTICA – 2 způsoby:

1. označit příslušné sloupečky v datech – Graphs – Graphs of Block Data – Box Plot: Block columns
2. Statistics – Basic Statistics/ Tables – Descriptive statistics – Box & whisker plot for all variables (na záložce Options lze zvolit, že krabicové grafy mají být typu Median/Quartiles/Range nebo po kliknutí do grafu lze v Plot - Box/Whisker měnit Middle point, Box value, Whisker value a po kliknutí na More i zapnutí Outliers)



Vícenásobné krabicové grafy

- umožňují znázornění vztahu několika kvalitativních proměnných a jedné kvantitativní proměnné

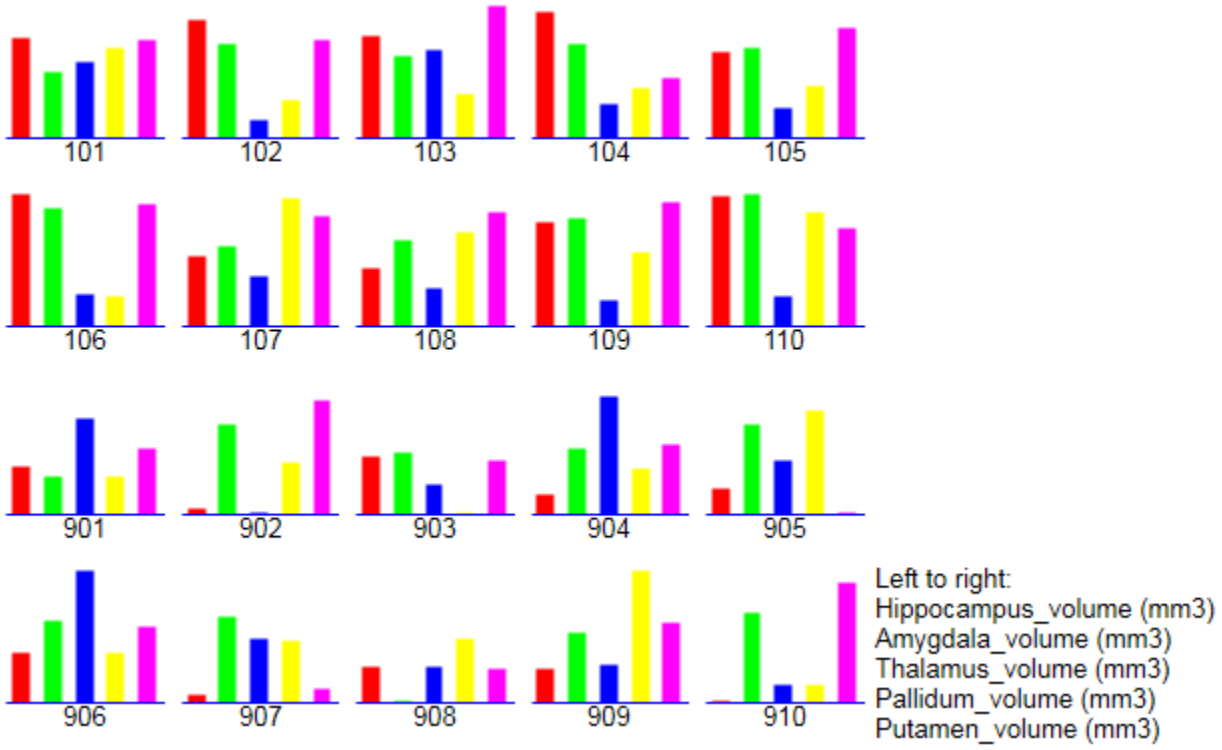


Ikonové (symbolové) grafy

- 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

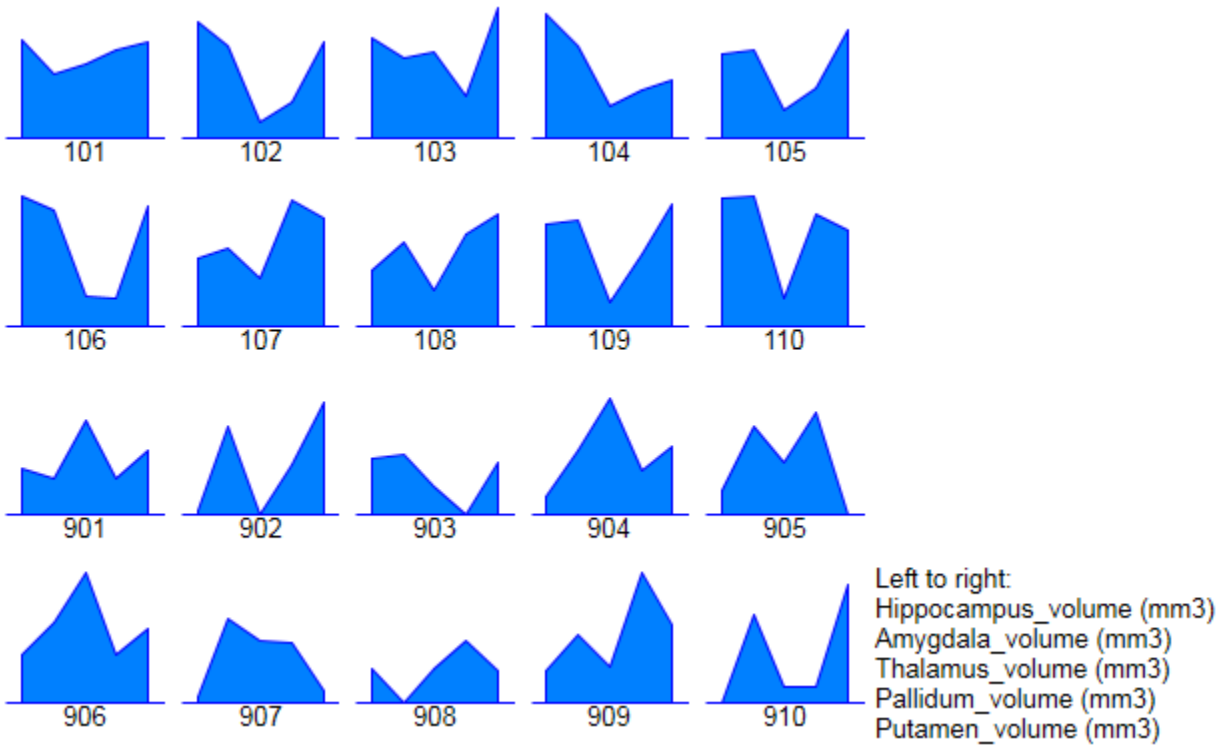
Ikonové grafy – profilové sloupce

- 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“



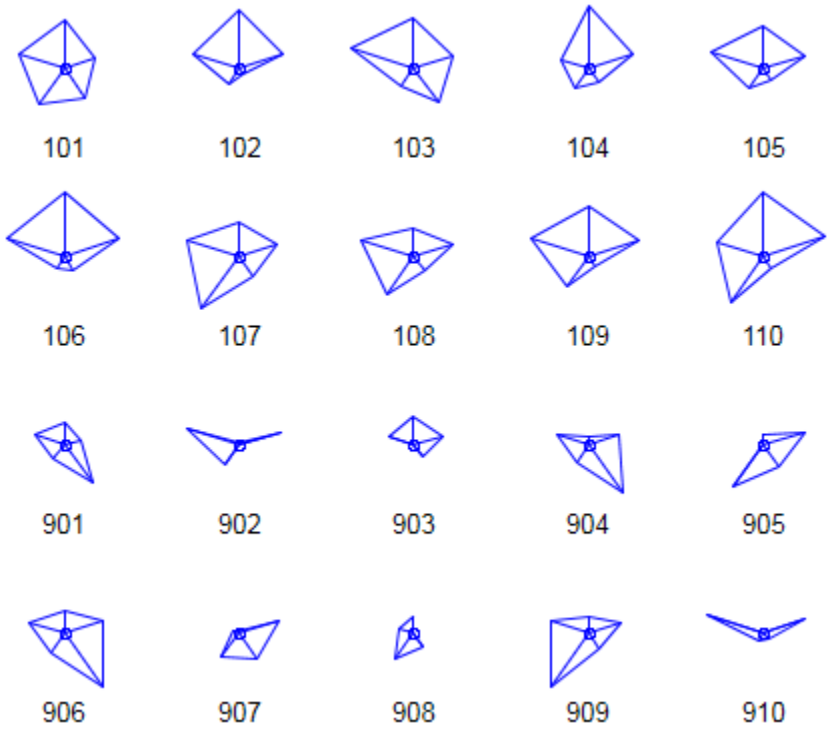
Ikonové grafy – profily

- 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“



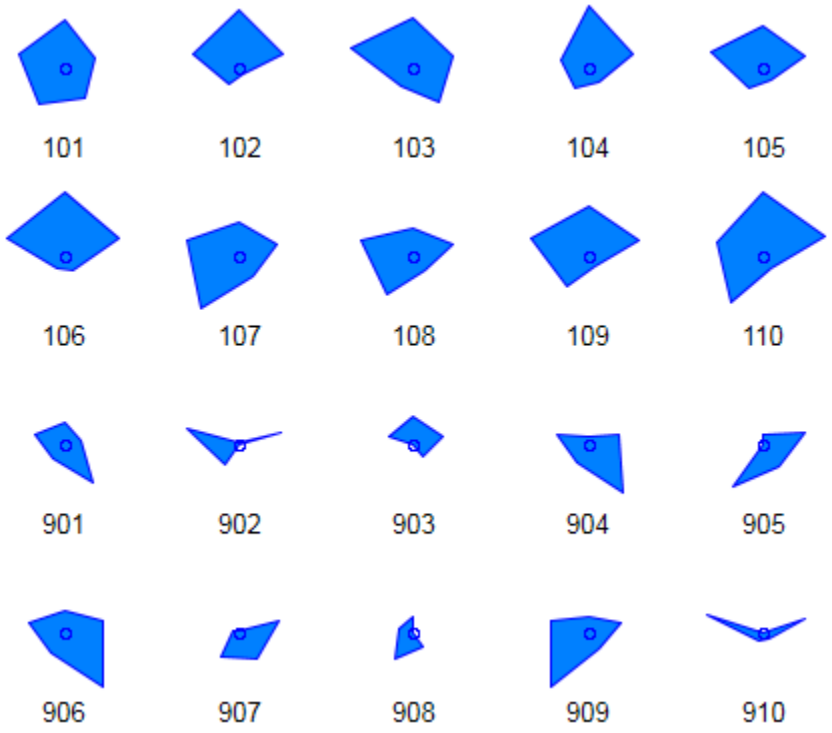
Ikonové grafy – paprskové (hvězdicové) grafy

- 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“



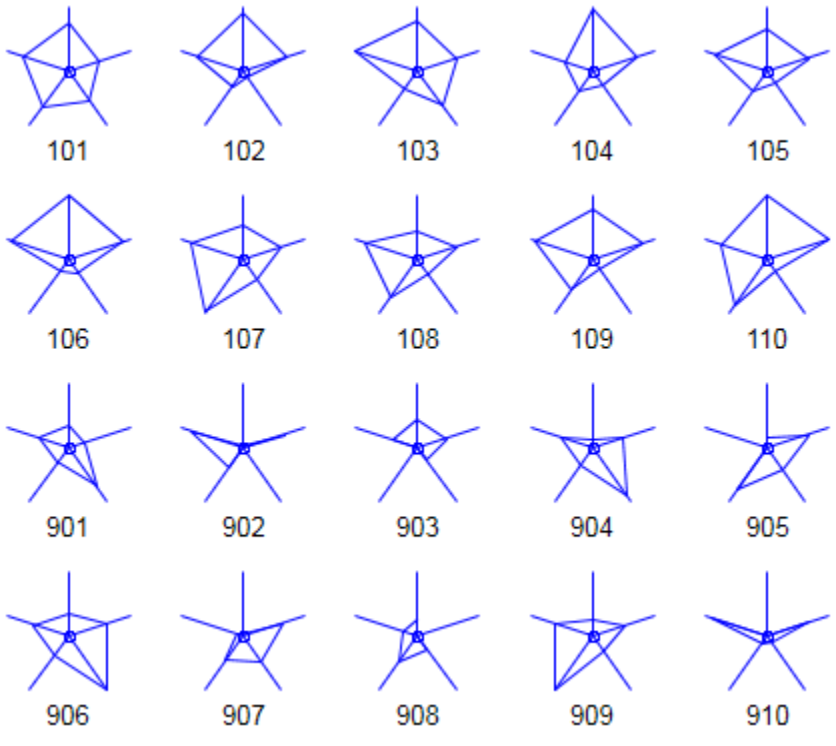
Ikonové grafy – polygony

- 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“



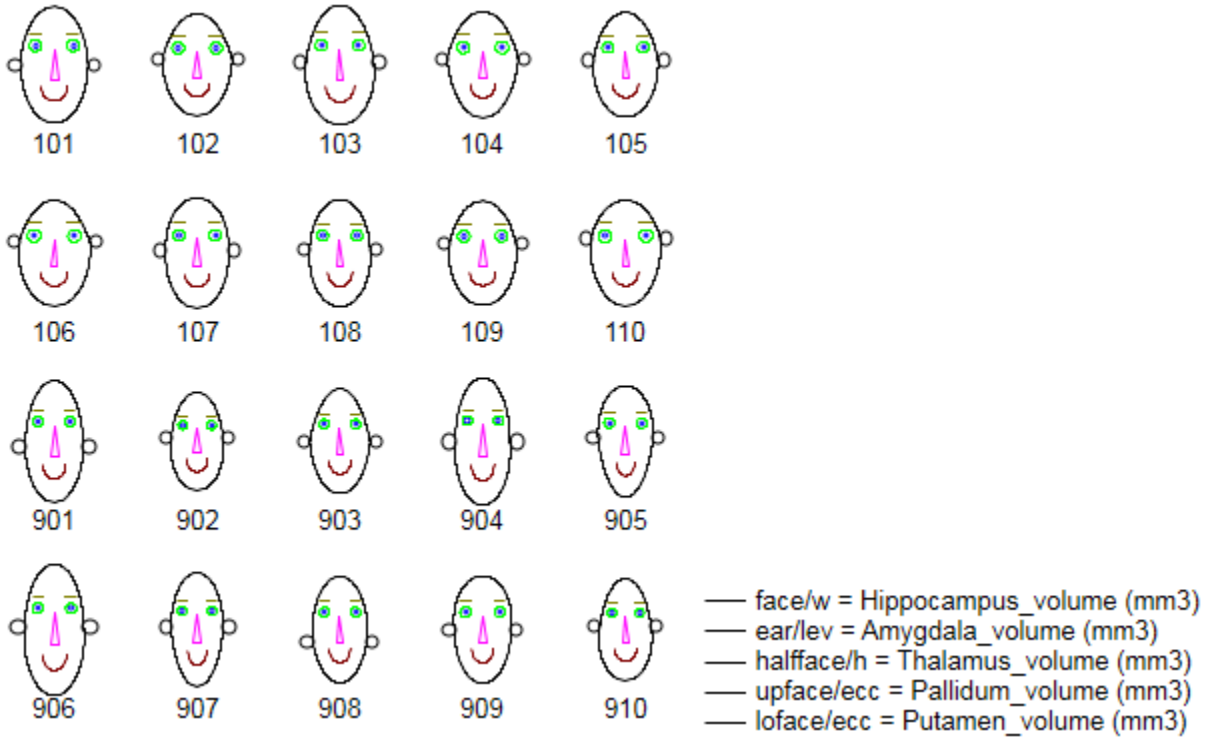
Ikonové grafy – pavučinové grafy

- 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“

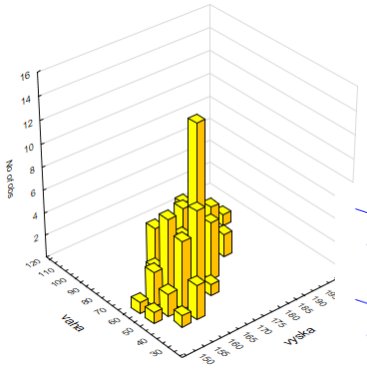


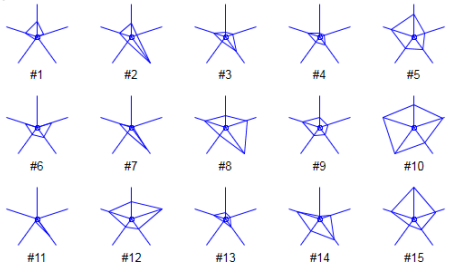
Úkol 3

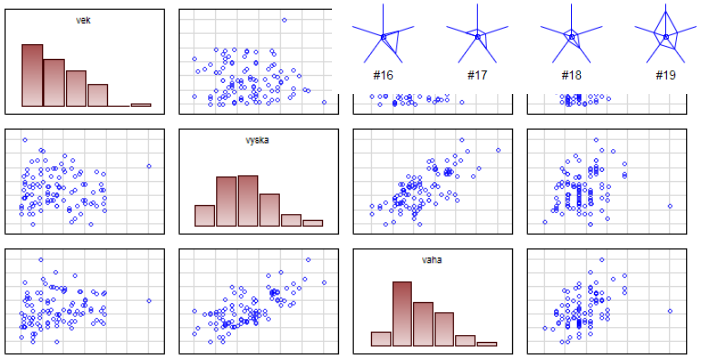
- zvolte si typ ikonových grafů, které se Vám zdají nejpřehlednější, a vykreslete graf pro subjekty 201 až 230 s využitím proměnných věk, MMSE, objem hipokampu a objem nucleus caudatus

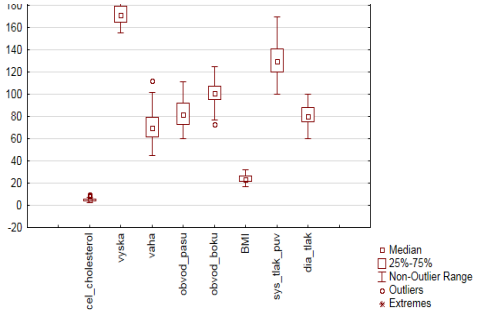
Vizualizace vícerozměrných dat - shrnutí

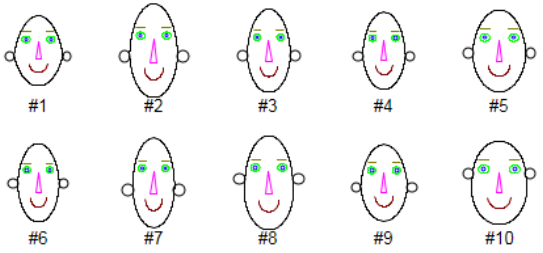
- 3D sloupkové grafy
- dvourozměrný histogram
- maticové grafy
- krabicové grafy pro více proměnných
- ikonové (symbolové) grafy:
  - profilové sloupce
  - profily
  - paprskové (hvězdicové) grafy
  - polygony
  - pavučinové grafy
  - Chernoffovy tváře

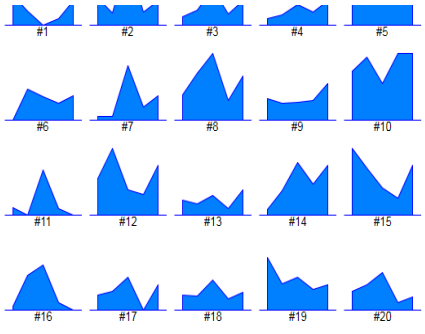






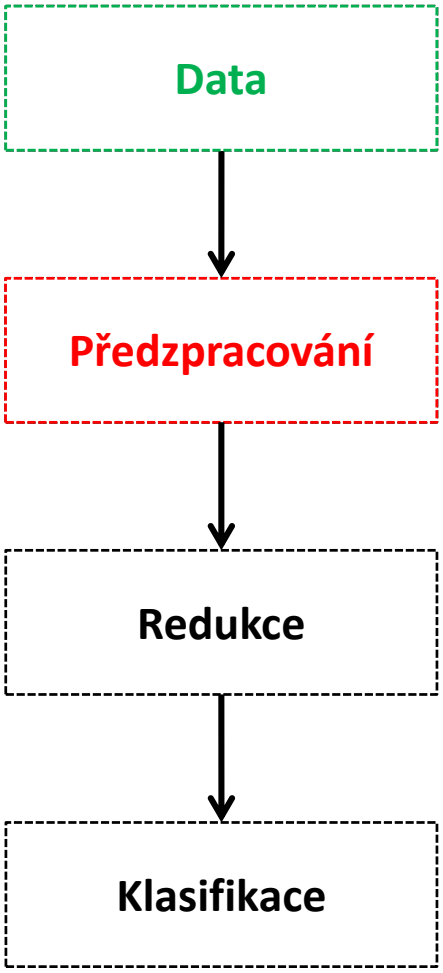






Předzpracování dat

Schéma analýzy a klasifikace dat



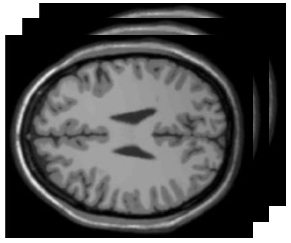
Ukázka - kognitivní data apod.

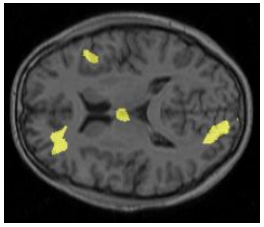
	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M		90
4	3	26	Z	178	70

	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

Ukázka - obrazová data



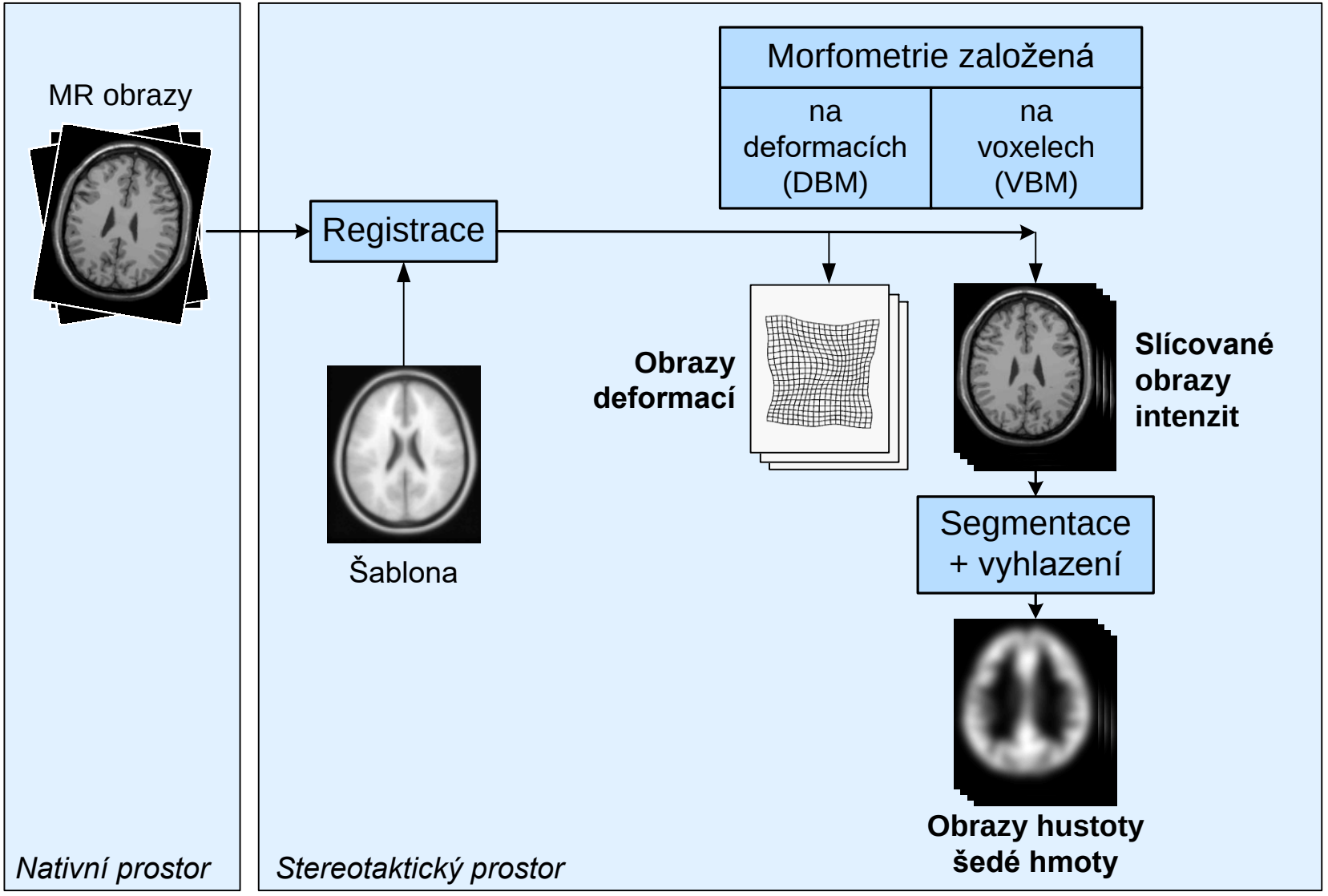


?

nebo

?

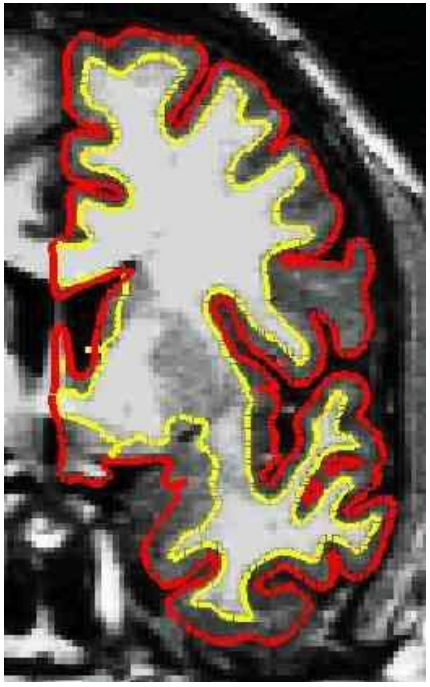
Předzpracování obrazových dat



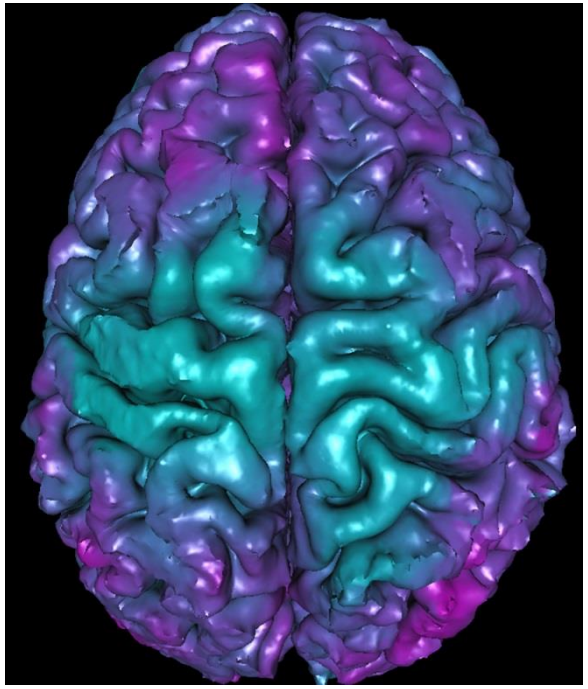
Předzpracování obrazových dat

Další typy dat, které mohou vzniknout po předzpracování obrazů:

Informace o tloušťce šedé hmoty v jednotlivých oblastech mozku



Informace o ploše jednotlivých oblastí mozku



Předzpracování dat – chybějící hodnoty

- snaha, aby v datech vůbec nenastaly
- pokud však nastanou, je silně nedoporučováno dělat každou analýzu na jinak velkém souboru (tzv. „pairwise“ odstraňování objektů) → 3 možná řešení:

1. vyloučit z analýzy všechny objekty, u nichž se vyskytla nějaká chybějící hodnota (tzv. „casewise“= „listwise“ odstranění objektů):
   - pokud chybějících hodnot mnoho, zbyde pouze málo objektů
   - pozor na systematicky chybějící hodnoty – může dojít ke zkreslení výsledků analýz
   - občas vhodné odstranit proměnné s mnoha chybějícími hodnotami místo objektů, pokud proměnné nejsou důležité pro analýzu
2. definování souboru s vyplněnými „klíčovými“ proměnnými:
   - na tomto souboru provedena většina analýz
   - další analýzy dělány na podsouboru s menším počtem subjektů
3. doplnění chybějících hodnot (tzv. imputace):
   - doplnění průměrem z hodnot, které jsou pro danou proměnnou k dispozici
   - doplnění hodnot na základě regresních modelů
   - pozor! doplnění hodnot však může zkreslit výsledky analýz

Předzpracování dat – odlehlé hodnoty

- k identifikaci odlehlých hodnot mohou pomoci tečkové, maticové či krabicové grafy
- další možné metody k identifikaci odlehlých hodnot budou probrány na příští přednášce
- je třeba rozlišovat:

1. **odlehlé hodnoty, které jsou způsobeny chybou** (měřících přístrojů apod.) - jsou to většinou nereálné hodnoty → je vhodné je smazat a dále s nimi zacházet jako s chybějícími hodnotami

2. **odlehlé hodnoty, které jsou fyziologické** (tzn. jsou to reálné hodnoty) → je vhodné tyto hodnoty v datech ponechat, pokud je to možné a nezkreslí to analýzu a použít neparametrické metody analýzy dat
   - příklad, kdy je vhodné odlehlou hodnotu v souboru ponechat: pacienti Alzheimerovou chorobou v našem souboru mají hodnotu MMSE skóre větší než 15, jeden pacient má však hodnotu skóre 7 (je to reálná hodnota, smazáním bychom uměle snížili variabilitu)
   - příklad, kdy je nevhodné odlehlou hodnotu v souboru ponechat: chceme měřit výšku 15-letých dětí – dítě trpící nanismem měřící 80 cm by průměrnou výšku velice zkreslilo, proto ho ze souboru vyřadíme

Poděkování

Příprava výukových materiálů předmětu
„DSAN02 Pokročilé metody analýzy dat v neurovědách“
byla finančně podporována prostředky projektu FRMU
č. MUNI/FR/0260/2014 „Pokročilé metody analýzy dat
v neurovědách jako nový předmět na LF MU“