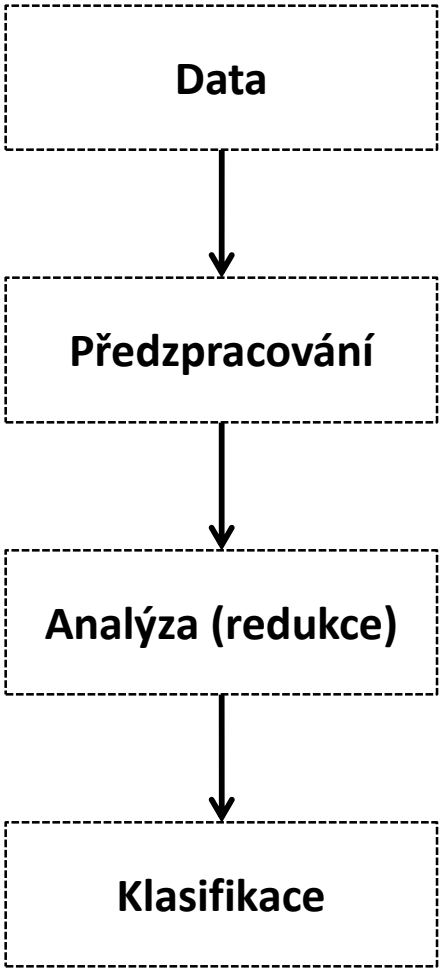
Analýza a klasifikace dat – přednáška 1

RNDr. Eva Koriťáková, Ph.D.

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

?

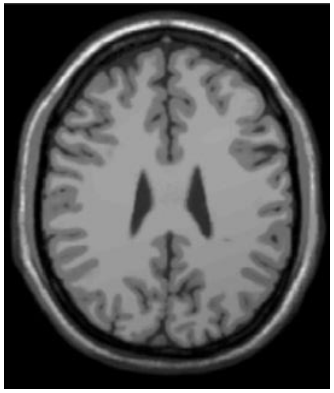
nebo

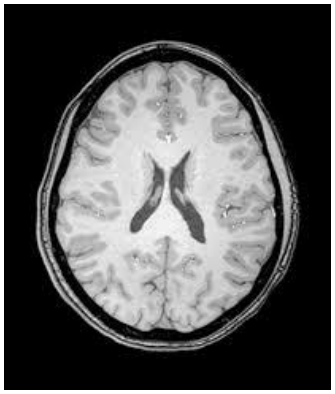
?

Proč používat klasifikaci dat?

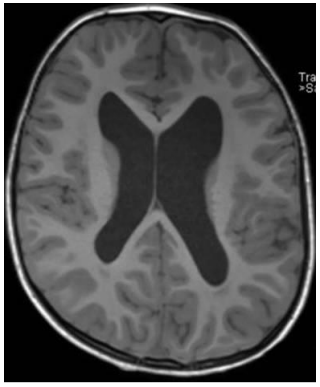
1. Podpora diagnostiky onemocnění mozku (Alzheimerova choroba, schizofrenie atd.):

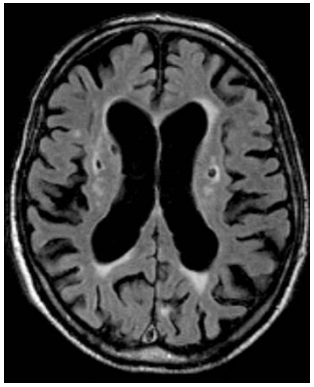
Zdravé subjekty



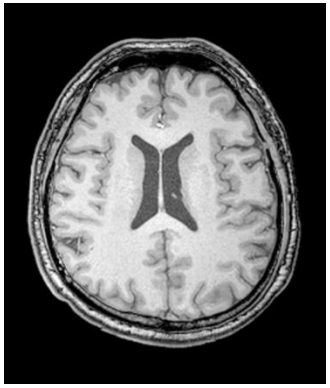


Pacienti





Nový subjekt

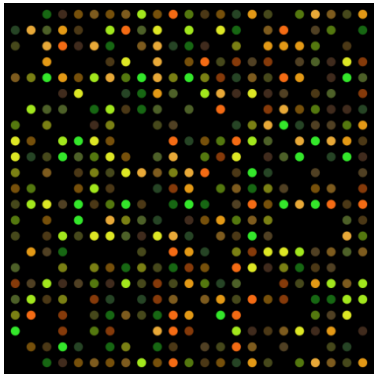


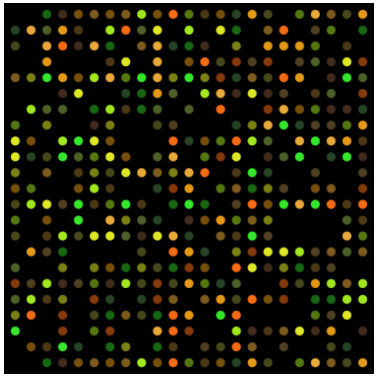
Pacient? x Zdravý?

Proč používat klasifikaci dat?

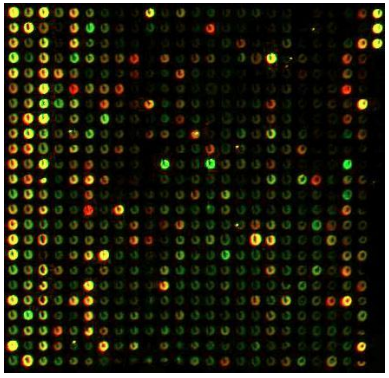
2. Odhalení genetického onemocnění na základě dat z microarray experimentů:

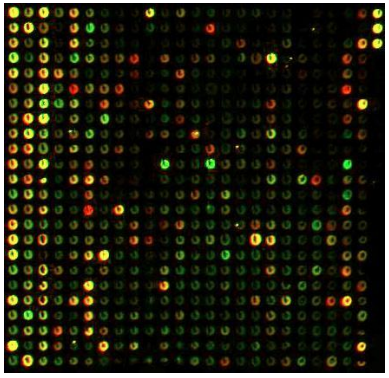
Zdravé subjekty



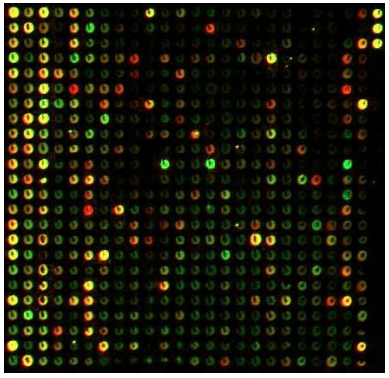


Pacienti





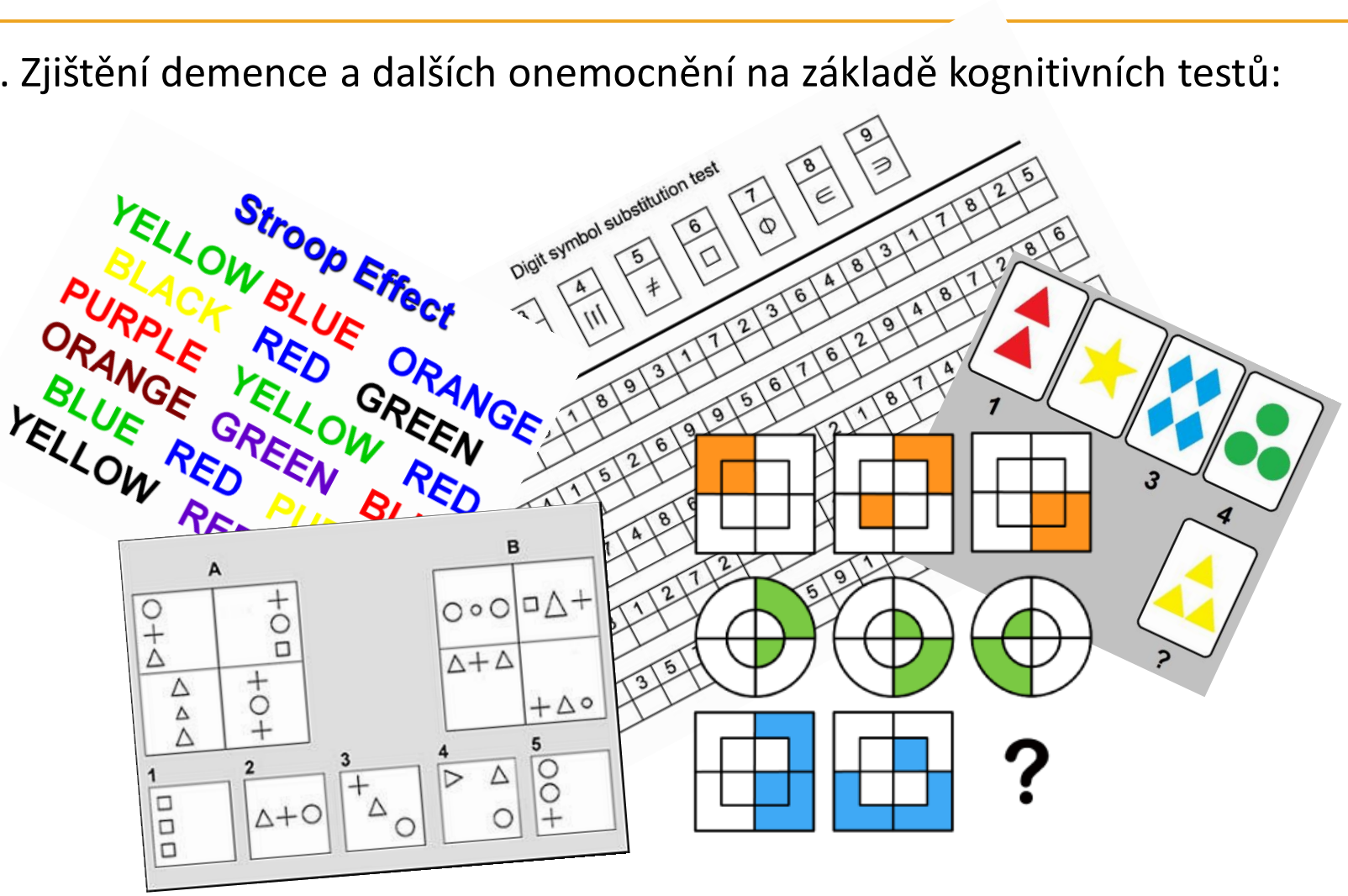
Nový subjekt



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

?

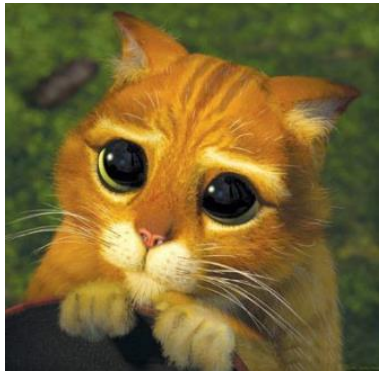
Matička bez vady nebo
s vnitřní prasklinou?

Proč používat klasifikaci dat?

6. Rozpoznání tváře při vstupu do zabezpečené budovy:

Nemá přístup do budovy

Má přístup do budovy

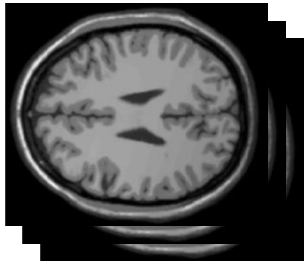


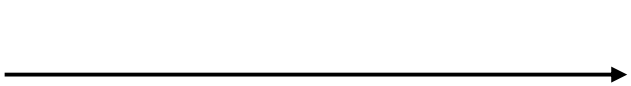
?

Dostane se do budovy: ano? x ne?

Proč používat redukci dat?

Obrazová data

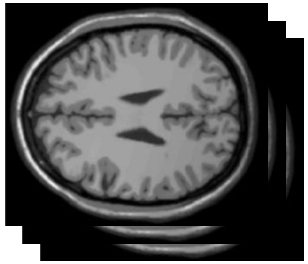




Klasifikace

Proč používat redukci dat?

Obrazová data



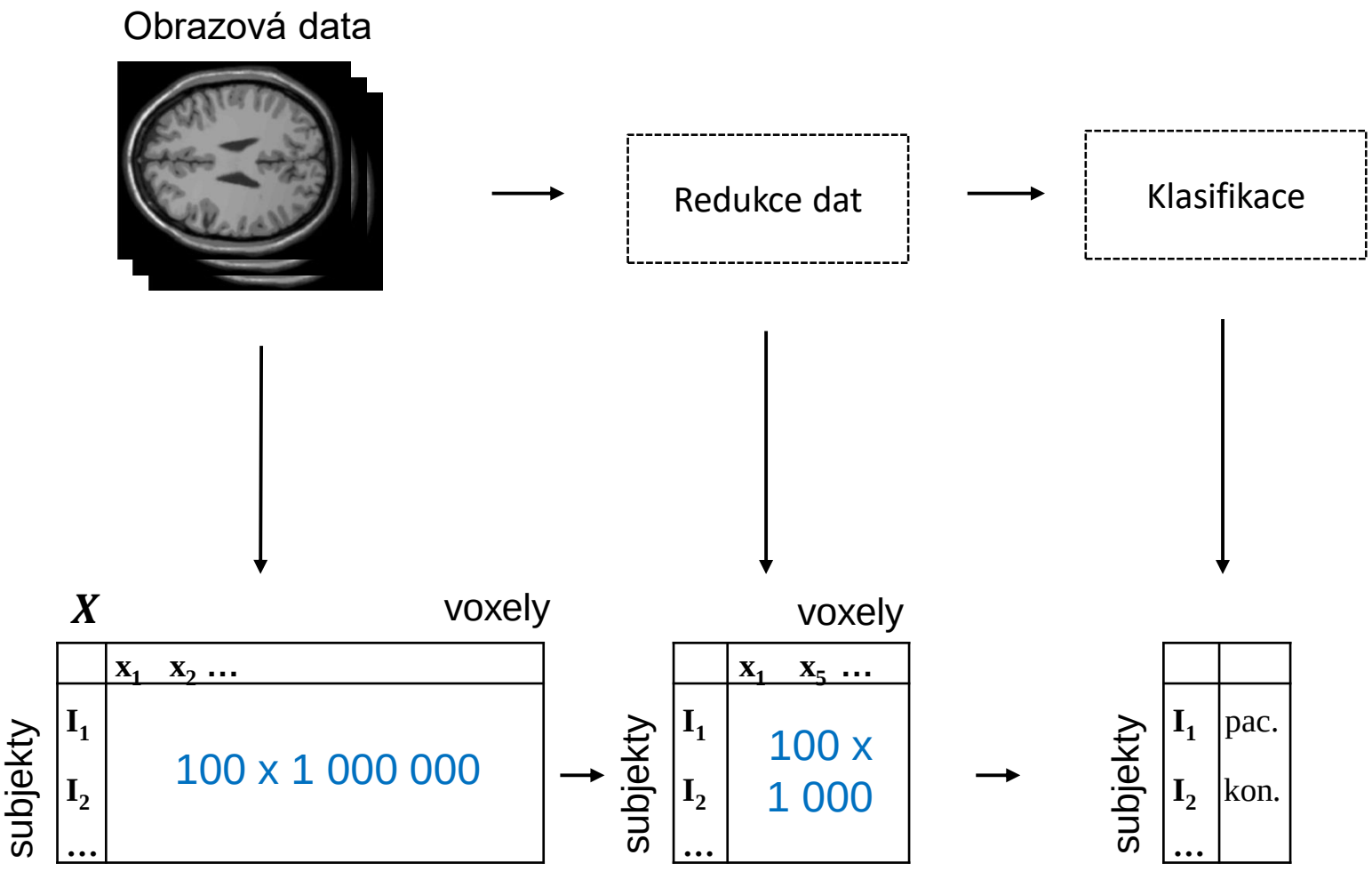
Klasifikace

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?



Proč data předzpracovávat?

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

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. „listwise“= „casewise” 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 např. tečkové, maticové či krabicové grafy
- 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

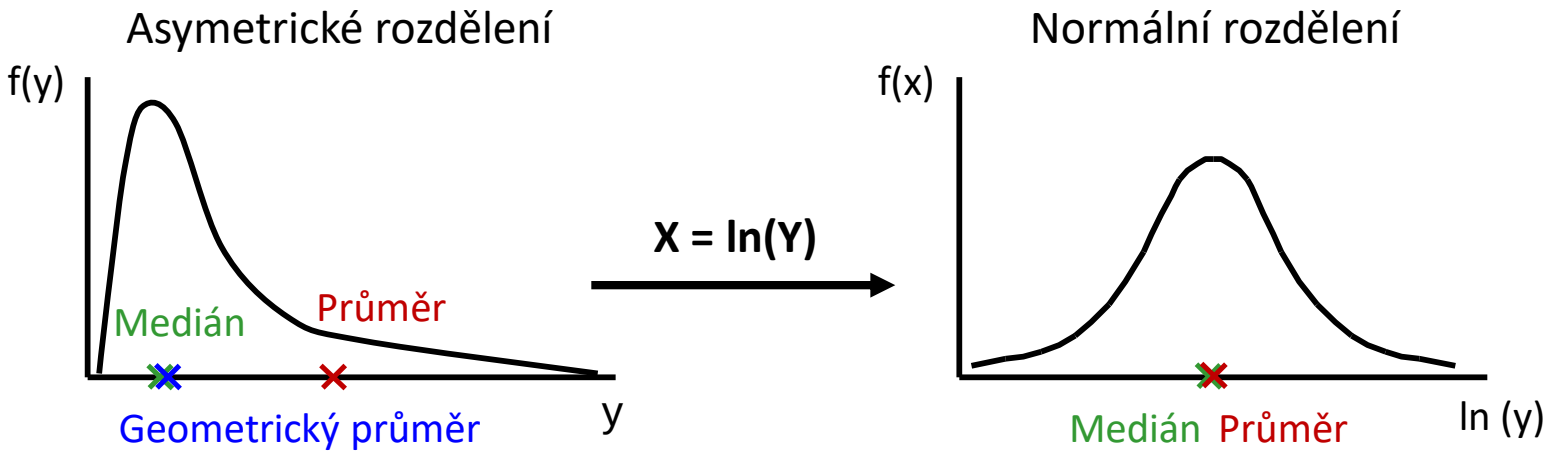
2. **odlehlé hodnoty, které jsou fyziologické** (tzn. jsou to reálné hodnoty) → je vhodné tyto hodnoty v datech ponechat, pokud je to možné a nezkreslí to analýzu a použít neparametrické metody analýzy dat
   - příklad, kdy je vhodné odlehlou hodnotu v souboru ponechat: pacienti Alzheimerovou chorobou 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

Předzpracování dat – transformace

- normalizace dat (= převod na normální rozdělení)
- standardizace dat
- min-max normalizace
- centrování dat
- odstranění vlivu kovariát

Normalizace dat

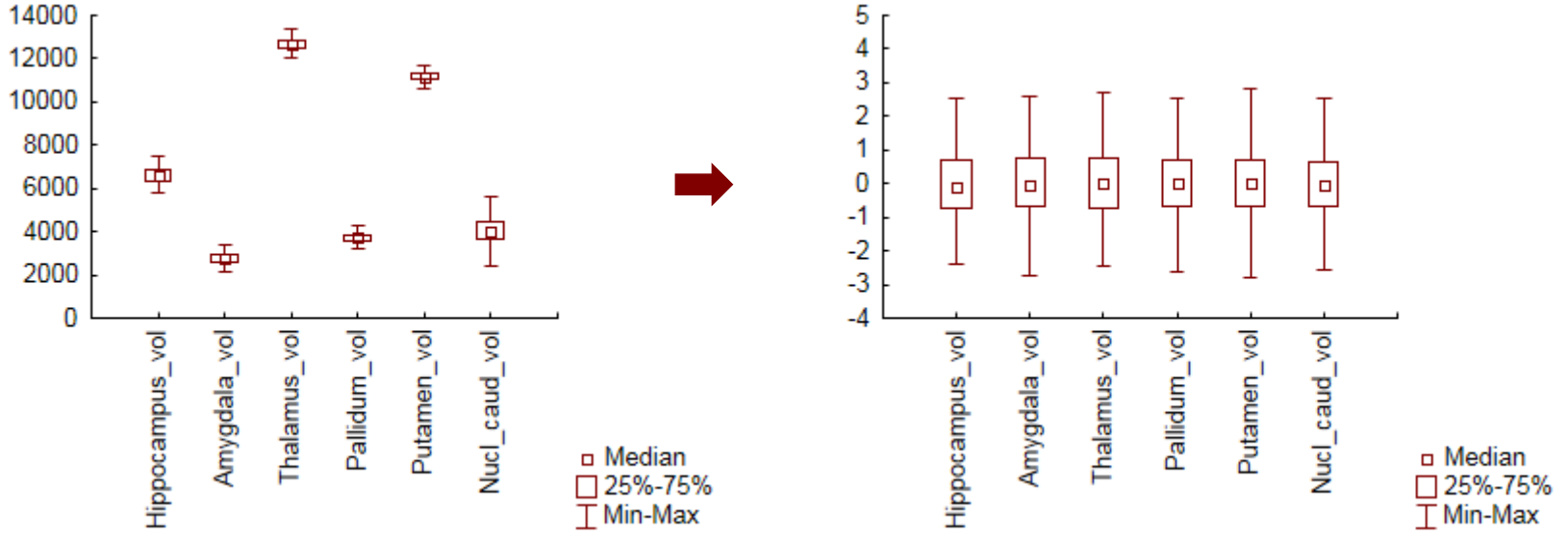
- převod na normální rozdělení (normalita je předpokladem řady statistických testů).
- např. **logaritmická transformace**: X = ln(Y) nebo X = ln(Y+1), pokud data obsahují hodnotu 0



- další příklady:
  - **odmocninová transf.** (pro proměnné s Poissonovým rozložením nebo obecně data typu počet jedinců, buněk apod.: $X = \sqrt{Y}$ nebo $X = \sqrt{Y+1}$
  - **arcsin transfomace** (pro proměnné s binomickým rozložením)
  - **Box-Coxova tranformace**

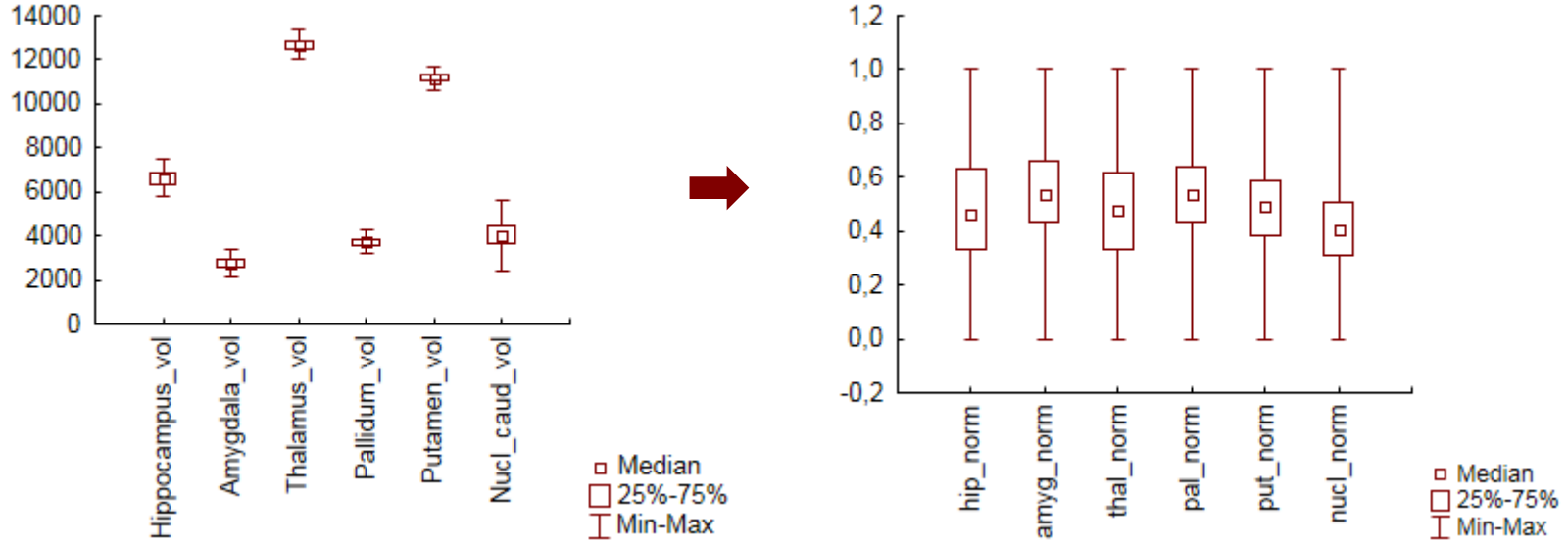
Standardizace dat

- důvod: převod proměnných na stejné měřítko
- standardizace: $z_i = \frac{x_i - \bar{x}}{s}$ (tzn. odečtení průměru od jednotlivých hodnot a podělení směrodatnou odchylkou)
- proměnné budou mít rozsah přibližně od -3 do 3
- získáme tím současně i tzv. z-skóre (které vyjadřuje, o kolik směrodatných odchylek se i-tá hodnota odchýlila od průměru)
- pozor: standardizace je nevhodná v případě, když proměnné nemají normální rozdělení a když se v datech vyskytují odlehlé hodnoty!!!



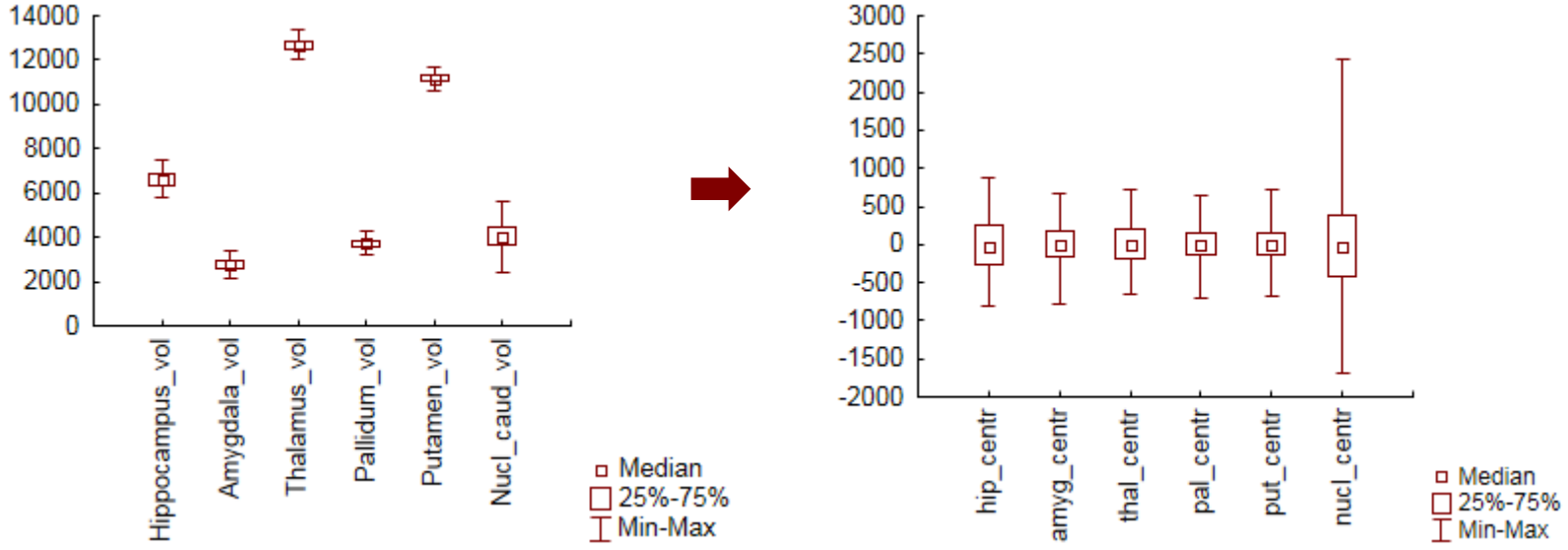
Min-max normalizace

- důvod: převod proměnných na stejné měřítko
- oproti standardizaci vhodná i na proměnné nemající normální rozdělení či obsahující odlehlé hodnoty
- min-max normalizace: $y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$
- rozsah hodnot proměnných po min-max normalizaci je od 0 do 1



Centrování dat

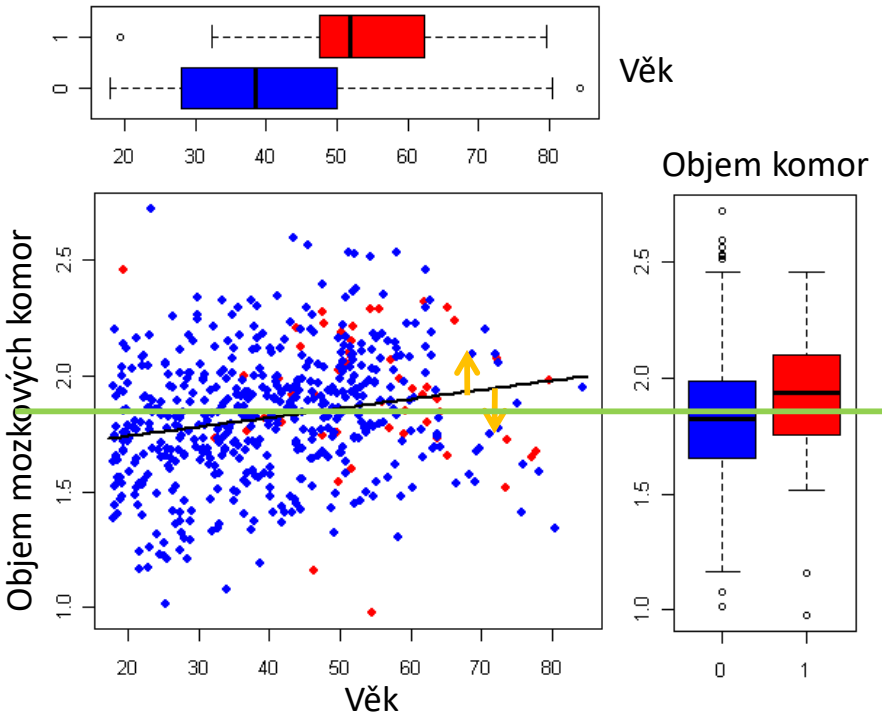
- odečtení průměru od dat – získáme novou proměnnou, která bude mít průměr roven nule
- důvod: centrování je důležitou podmínkou některých pokročilých statistických metod (např. klasifikačních)
- centrování: $z_i = x_i - \bar{x}$



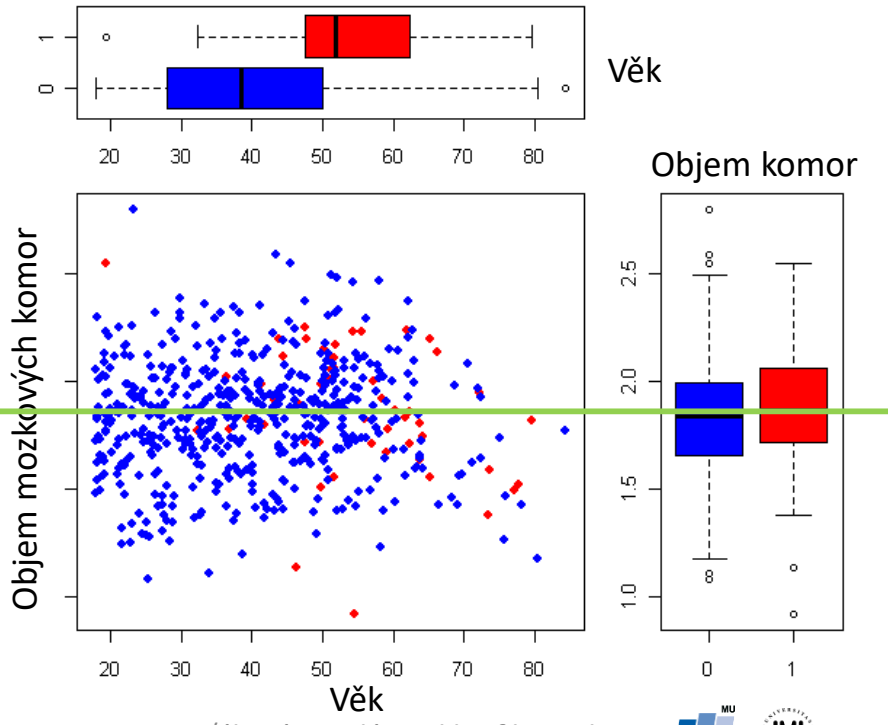
Odstranění vlivu kovariát (tzv. adjustace)

1. V prvním kroku definujeme regresní model vztahu kovariáty (např. věku) a dané proměnné
2. Pro každého pacienta je vypočteno jeho reziduum od regresní přímky ↑↓
3. Reziduum (představující hodnotu parametru po odečtení vlivu věku, jeho průměr je 0) je přičteno k průměrné hodnotě parametru —
4. Výsledná adjustovaná hodnota má odečten vliv věku, ale zároveň není změněna číselná hodnota parametru

Původní data



Adjustovaná data

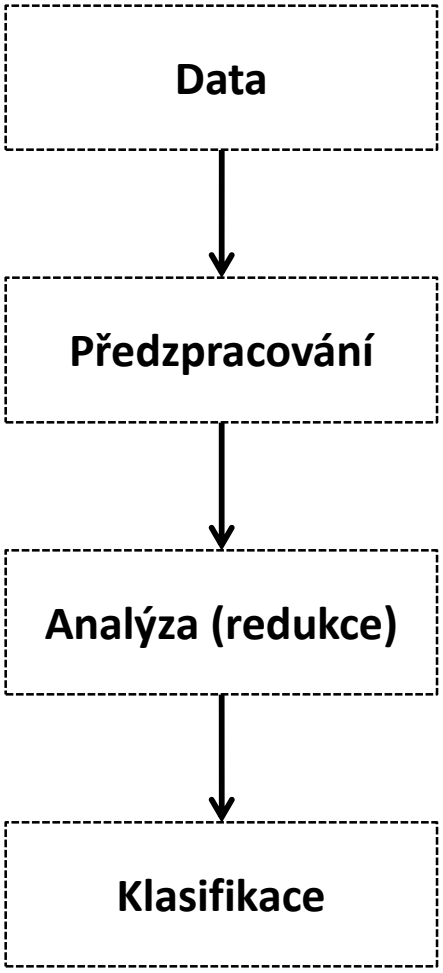


Cíle analýzy a klasifikace dat - shrnutí

- **rozhodnutí o typu či charakteru objektu** – např. že daná rostlina je pomněnka lesní (*Myosotis sylvatica*), zvíře 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 klasifikační algoritmus pro známá data; pokud jsou data nová, pro která předem neznáme klasifikační třídu, pak hovoříme o ***predikci*** klasifikační třídy
  - pojem ***klasifikace*** je používán, 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

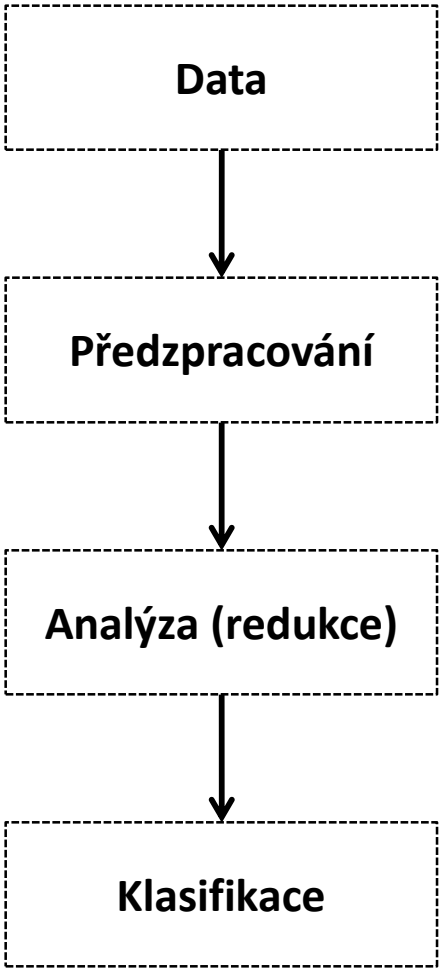
	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70


?

nebo


?

Schéma analýzy a klasifikace dat



- rekonstrukce a doplnění chybějících údajů;
- vypořádání se s odlehlými hodnotami;
- filtrace rušivých a zvýraznění užitečných složek dat;
- konverze typu dat (A/D 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

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

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*) – obraz nejen ve smyslu obraz mozku či obraz sítnice oka, ale ve smyslu popis (tzn. „obraz“) reálného objektu
  - „dolování z dat“ (*data mining*)
  - „strojové učení“ (*machine learning*)

Typy klasifikátorů

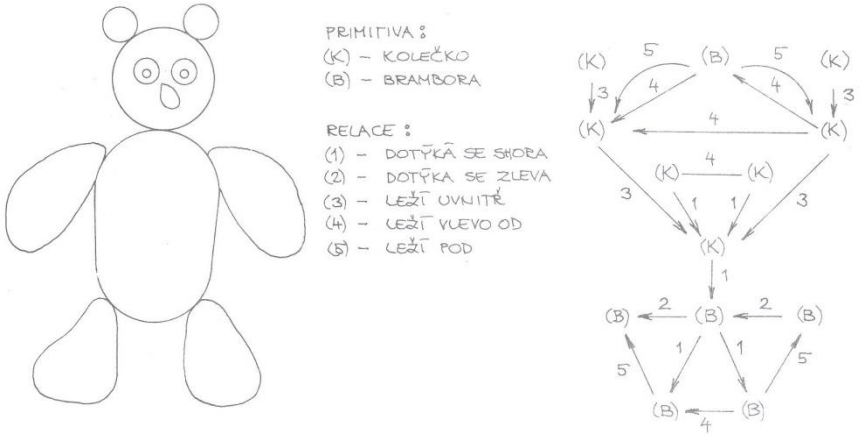
1. **Podle reprezentace vstupních dat:**
   - **příznakové klasifikátory:** paralelní x sekvenční
   - **strukturální (syntaktické) klasifikátory**
   - **kombinované klasifikátory**
2. **Podle jednoznačnosti zařazení do skupin:**
   - **deterministické klasifikátory**
   - **pravděpodobnostní klasifikátory**
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 principu klasifikace:**
   - **klasifikace pomocí diskriminačních funkcí**
   - **klasifikace pomocí vzdálenosti od etalonů klasifikačních 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

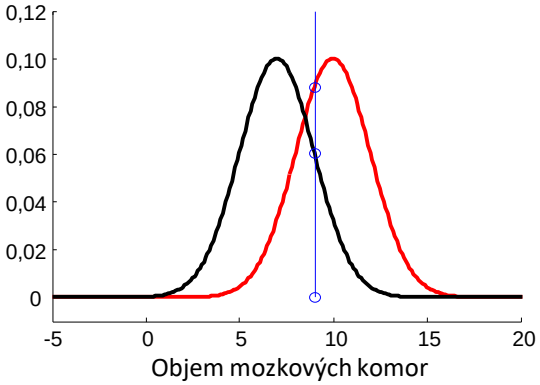
Typy klasifikátorů – podle způsobu učení

- **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)

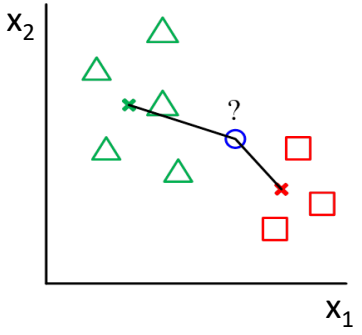
- **učení bez učitele:**
  - trénovací množina není k dispozici a často ani předem neznáme, jaké třídy (skupiny) se v datech budou vyskytovat
  - typickým příkladem je shlukování

Typy klasifikátorů – podle principu klasifikace

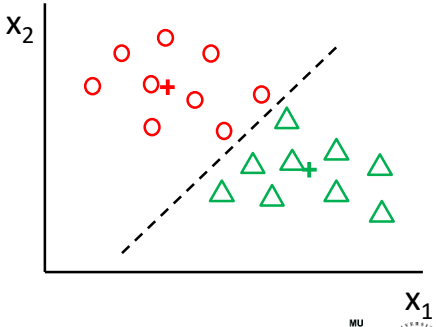
- **klasifikace pomocí diskriminačních funkcí:**
  - diskriminační funkce určují míru příslušnosti k dané klasifikační třídě
  - pro danou třídu má daná diskriminační funkce nejvyšší hodnotu



- **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



Citát na závěr:

Marriott, F. H. C. *The Interpretation of Multiple Observations*. London: Academic Press (1974):

"If the results disagree with informed opinion, do not admit a simple logical interpretation, and do not show up clearly in a graphical presentation, they are probably wrong. There is no magic about numerical methods, and many ways in which they can break down. They are a valuable aid to the interpretation of data, not sausage machines automatically transforming bodies of numbers into packets of scientific fact."

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