Úvod do analýzy dat

Matematické metody pro ITS (11MAMY)

Ondřej Přibyl (Jan Přikryl)

Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní

Obsah prezentace

- Měřené veličiny
- Chyby měření
- Vizualizace dat
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

Diskuze

- Jaký je rozdíl mezi:
  - DATY,
  - INFORMACÍ a
  - ZNALOSTMI?

- Uveďte na příkladech.

Data, informace a znalosti

Data

Jakékoli vyjádření (reprezentace) skutečnosti, schopné přenosu, interpretace či zpracování. Účelem dat je přenášet a dále zpracovávat odraz skutečnosti. Jsou to jakékoli zaznamenané poznatky či fakta.

Informace

Data, která mají smysl (význam). Jsou to sdělitelné (komunikovatelné) znalosti. Je to údaj, ke kterému si člověk přiřadí význam.

Znalost

To, co jednotlivec ví po osvojení dat a informací a po jejich začlenění do souvislostí.
Účelem znalostí je porozumění modelům.

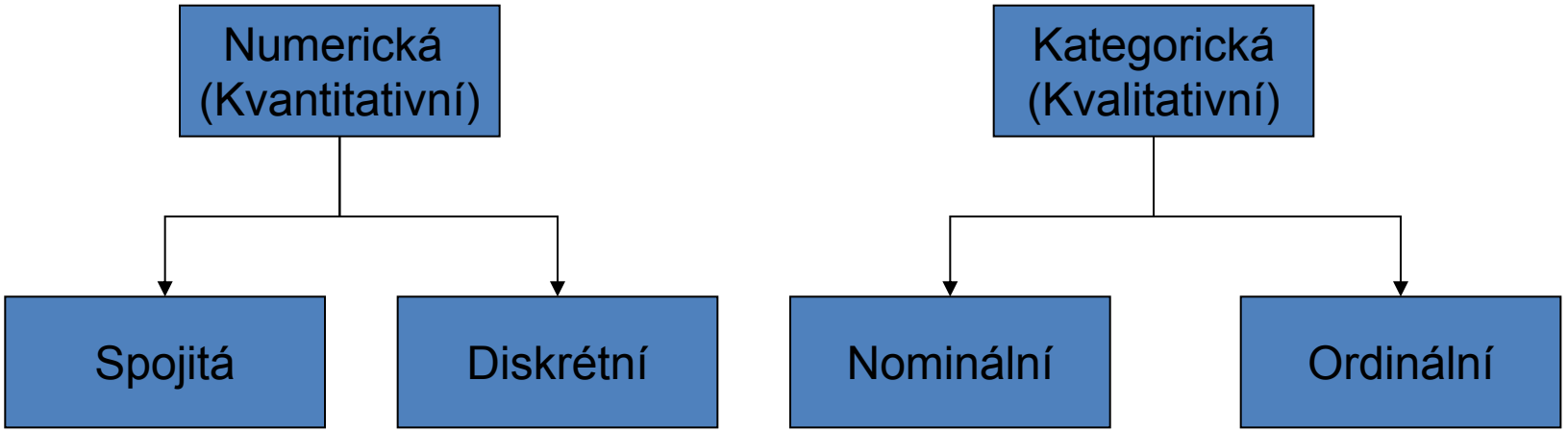
Moudrost

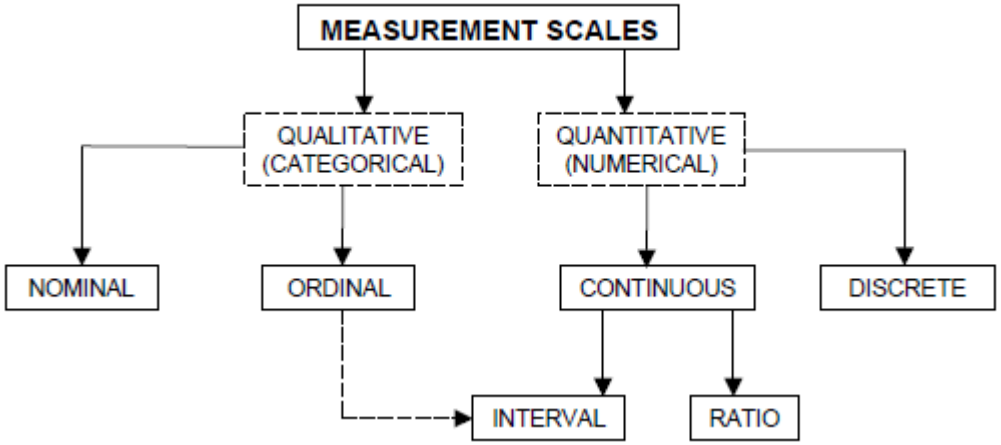
Porozumění principům.

Diskuze

- Jaké znáte typy měřených dat?
- Co mají společného, v čem se liší?

Přehled kategorií měřených dat





Numerická data

Spojitá data

- stat. znak, který může nabývat všech reálných hodnot v rámci konečného nebo nekonečného intervalu
- Příklady:
  - MTBF - doba do poruchy zařízení
  - Doba jízdy
  - Hmotnost vozidla

Diskrétní / Nespojitá

- stat. znak který může nabývat v daném intervalu pouze izolovaných číselných hodnot
- zpravidla se jedná o přirozená čísla + 0, tedy {0, 1, 2, 3, …, n}
- Příklady:
  - Počet cest automobilem za týden
  - Počet dopravních nehod

Kategorická data

Nominální

- Nabývají konečného a nízkého počtu diskrétních hodnot
- nelze nad nimi vytvořit uspořádání.
- Příklady:
  - Druhy dopravních prostředků
  - Barvy vozidel

Ordinální

- Od nominálních proměnných se liší v tom, že nad nimi lze vytvořit uspořádání.
- Příklady:
  - malý, střední, veliký
  - nikdy<občas<často<vždy
- **Binární** (speciální případ)
  - Nabývají hodnot 0 a 1

Marital status

1.	Never married	☐	4.	Married/Cohabiting	☐
2.	Divorced	☐	5.	Separated	☐
3.	Widowed	☐			

Employee's performance

1.	Excellent	☐	4.	Poor	☐
2.	Good	☐	5.	Very poor	☐
3.	Average	☐			

Příklady z dopravy (zatřídění a veličiny)

- Uveďte jednotky dané veličiny a klasifikujte ji dle typu
  - Intenzita dopravy
  - Obsazenost detektoru
  - Stupeň dopravy
  - Počet vozidel v domácnosti
  - Doba jízdy
  - Třídy vozidel
  - Hustota

Obsah prezentace

- Měřené veličiny
- **Chyby měření**
- Základní charakteristiky dat
- Vizualizace dat
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

Kde vznikají chyby při měření dopravních dat?

Chyby…

Chyby zásadně ovlivňují měření, dělí se na **náhodné** (dynamická charakteristika) a **systematické** (statická charakteristika)

- Chyba detektoru
  - Chyba měřicího zařízení
    - způsobena nedokonalostí měřicích přístrojů
  - Chyba pozorovatele (chyby způsobené lidským faktorem)
    - nesprávná volba metody měření,
    - chybné zapojení přístrojů do obvodu,
    - nevhodná volba měřicího rozsahu,
    - chybné čtení údajů, atp.

Kde vznikají chyby při měření dopravních dat?

Chyby…

Chyby zásadně ovlivňují měření, dělí se na náhodné (dynamická charakteristika) a systematické (statická charakteristika)

- Chyba přenosu
  - Chyba způsobená výpadkem v přenosové cestě
- Chyba metody
  - jejich příčinou jsou různá zjednodušení vztahů pro výpočet měřené veličiny, zjednodušení zapojení, vliv spotřeby měřicího přístroje na jeho údaj, atd.
  - Tyto chyby je obvykle možno vypočítat a výsledek měření podle nich korigovat.

Úvod do problematiky

- Každé měření v sobě obsahuje **nejistotu správnosti** výsledku.

- **Systematické chyby**
  - jsou statického rázu, zkreslují výsledek stejným, kontrolovatelným způsobem bez ohledu na počet provedených měření.
  - zdroji těchto chyb je omezená přesnost přístrojů, použitá metoda měření a osobní chyby.
  - do chyb způsobených omezenou přesností spadají např. aditivní a multiplikativní chyby.
- Náhodné chyby
- Hrubé chyby

Úvod do problematiky

- Každé měření v sobě obsahuje **nejistotu správnosti** výsledku.

- Systematické chyby
- **Náhodné chyby**
  - vyskytují se zcela nepravidelně, jejich výskyt je náhodný (ale: pravděpodobnostní distribuce chyb)
  - jsou způsobeny nekontrolovatelnými vlivy
  - nelze je odstranit
  - zjistit je můžeme až při opakovaném měření
  - neplést si s náhodnými vlivy na řízený systém (viz přednáška 2)
- Hrubé chyby

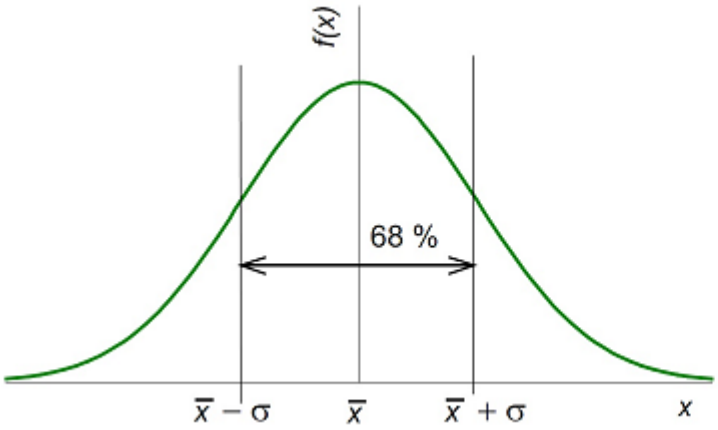
Úvod do problematiky

- Každé měření v sobě obsahuje **nejistotu správnosti** výsledku.

- Systematické chyby
- Náhodné chyby
- **Hrubé chyby**
  - někdo je považuje za první dvě kategorie chyb
  - vychýlené hodnoty (bias) … systematická
  - odlehlá měření (outliers) … náhodná
  - důvod: selhání měřicí aparatury, nesprávný záznam výsledku

Náhodné rozdělení chyb

- Normální (Gaussovo) rozdělení, střední hodnota odpovídá nejpravděpodobnější hodnotě opakovaného měření.

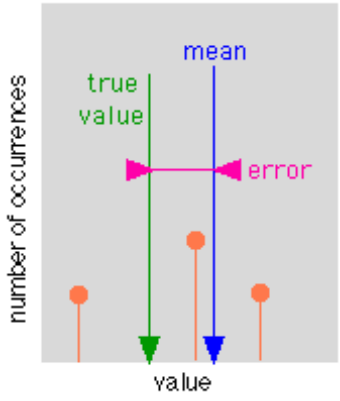


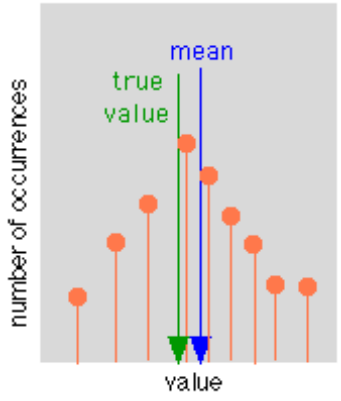
hustota pravděpodobnosti

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\bar{x})^2}{2\sigma^2}}$$

- Výsledky platí pro velké množství měření (n →∞).

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}.$$





„Přesnost“ versus „správnost“

Přesnost (precision)

- rozmezí statistické nejistoty výsledků
- souvisí s náhodnými chybami
- odpovídá reprodukovatelnosti měření
- vyjadřuje se jako rozptyl naměřených výsledků kolem průměru z *n* naměřených hodnot.
- lze odhadnout statisticky

Správnost (accuracy)

- udává průměrnou odlehlost (vzdálenost) výsledků měření od skutečné hodnoty
- souvisí se systematickými chybami
- odpovídá odchýlení měření od teoretické hodnoty.
- nelze ji odhadnout, je nutno ji stanovit s využitím standardů nebo měřením na více přístrojích

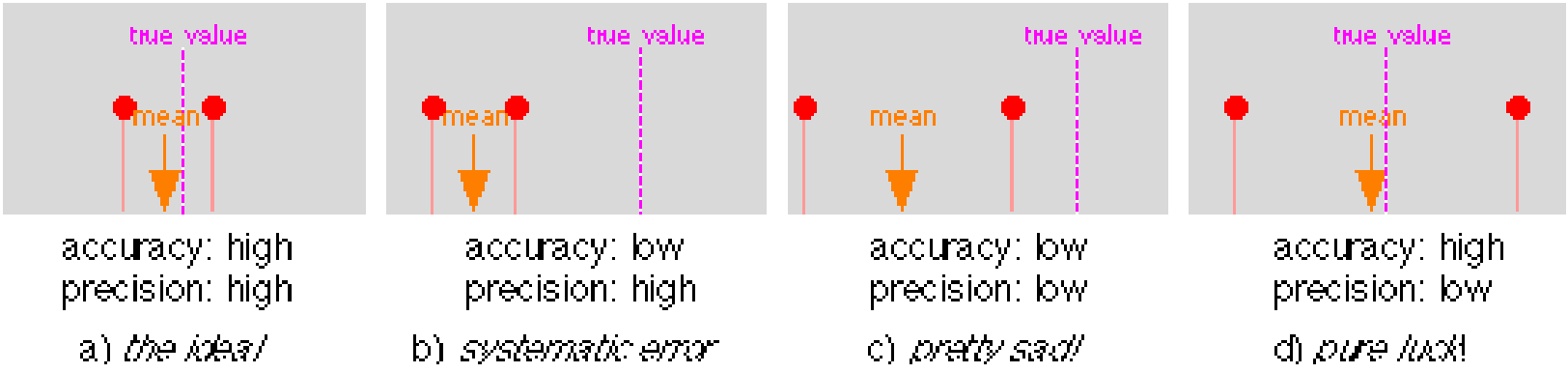
„Přesnost“ versus „správnost“

Přesnost (precision)

- rozmezí statistické nejistoty výsledků
- přesnost přístroje lze odhadnout na základě statistické analýzy

Správnost (accuracy)

- udává průměrnou odlehlost výsledků měření od skutečné hodnoty
- nelze ji odhadnout, je nutno ji stanovit s využitím standardů nebo měřením na více přístrojích

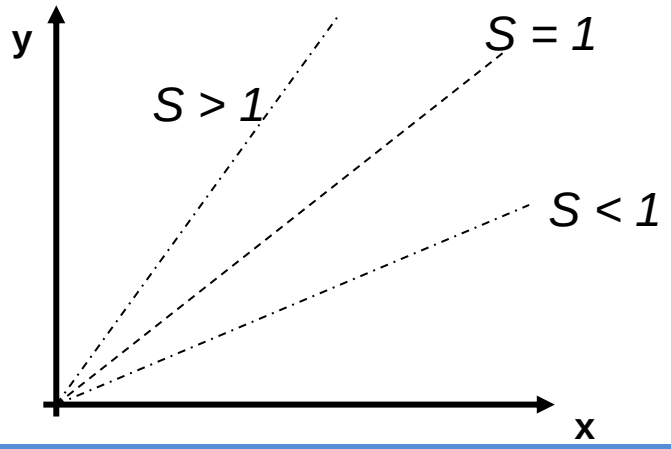


Citlivost (sensitivity) měřicího přístroje

Schopnost reagovat za stanovených podmínek na požadovanou změnu hodnoty měřené vstupní veličiny.

- podíl změny přístrojového údaje (výstupní veličiny) k požadované změně měřené (vstupní) veličiny, která změnu údaje vyvolává.
- *na přístrojích s ručkovým ukazatelem* je to velikost dílku stupnice, který odpovídá velikosti změny měřené veličiny,
- *u digitálních přístrojů* je to počet desetinných míst, s jakým je hodnota měřené veličiny udávána.

$$S = \Delta y / \Delta x$$



Diskuze

- „Při měření intenzity dopravy byla naměřena chyba 5 vozidel„
  - Je to hodně nebo málo?

Chyby měření

1. **Absolutní chyba měření**
   y_N ... naměřená hodnota
   y_S ... správná hodnota

$$\Delta_y = y_N - y_S$$

2. **Relativní chyba měření**

$$\delta_y = \frac{|\Delta_y|}{y_S}$$

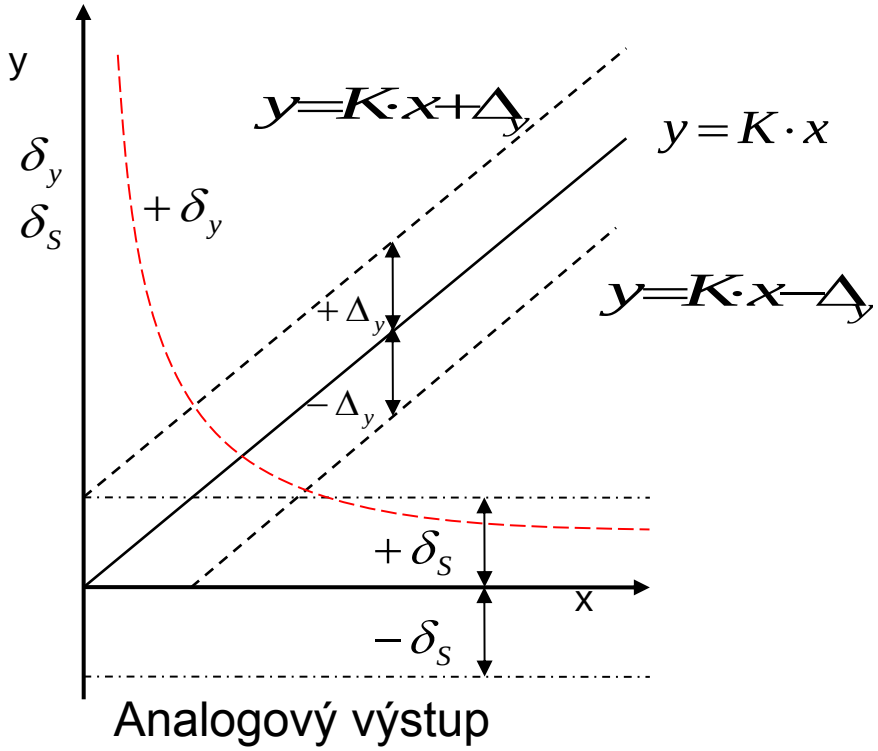
3. **Relativní chyba senzoru**
   Poměr maximální absolutní chyby měření vůči rozsahu hodnot měřené veličiny

$$\delta_s = \frac{\max|\Delta_y|}{y_{max} - y_{min}}$$

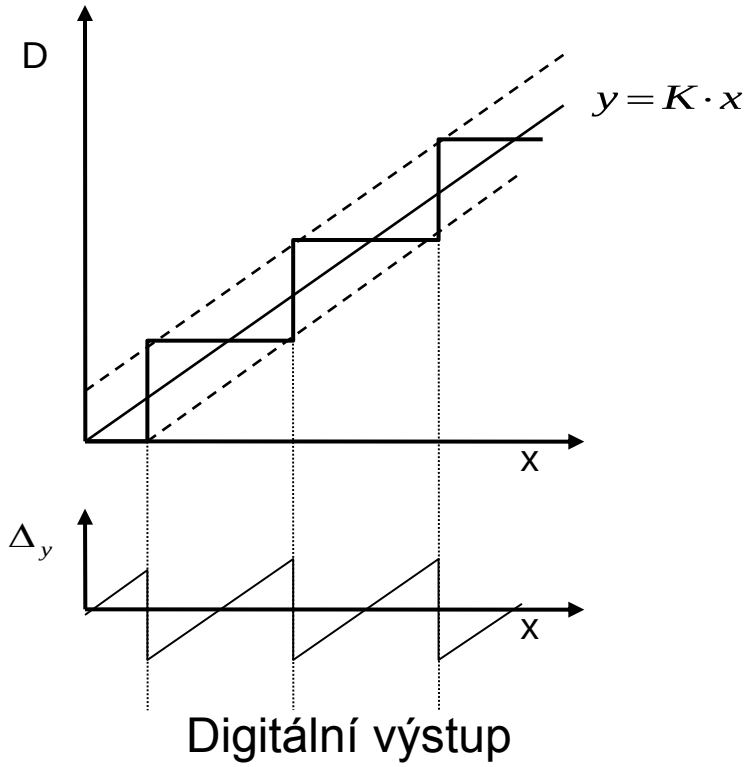
Chyby měření (pokrač.)

4. Aditivní chyba měření

- způsobena posunem jmenovité lineární charakteristiky
- Δ_y je konstantní
- δ_y je nepřímo úměrná měřené hodnotě



Analogový výstup

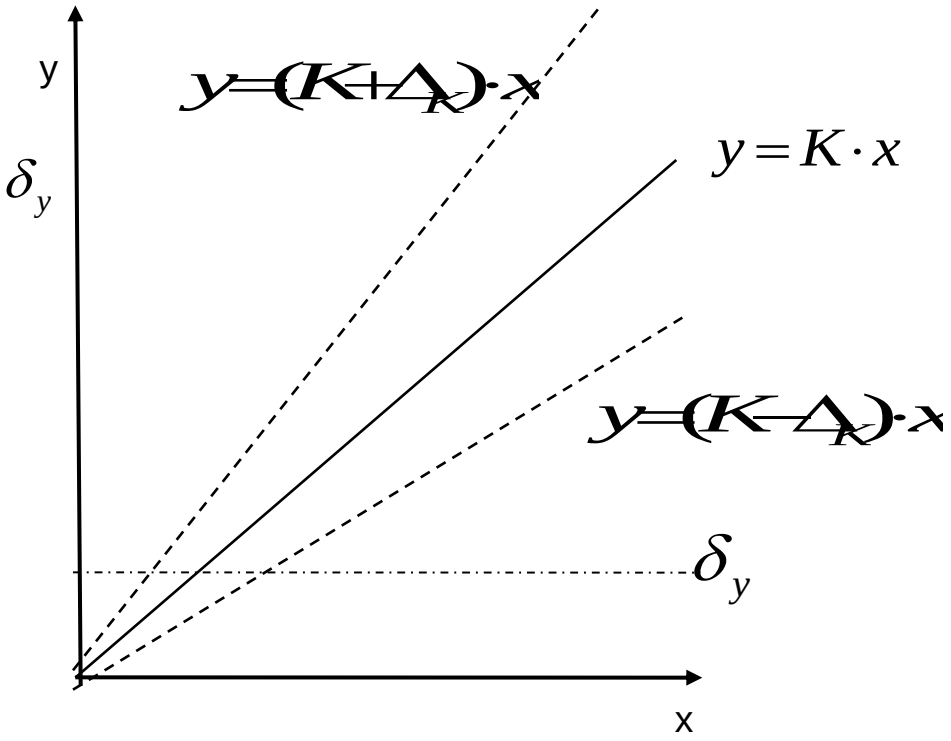


Digitální výstup

Chyby měření

5. Multiplikativní chyba měření

- ekvivalentní změně citlivosti senzoru podle x
- Δ_y závislá na hodnotě měřené veličiny
- δ_y je konstantní



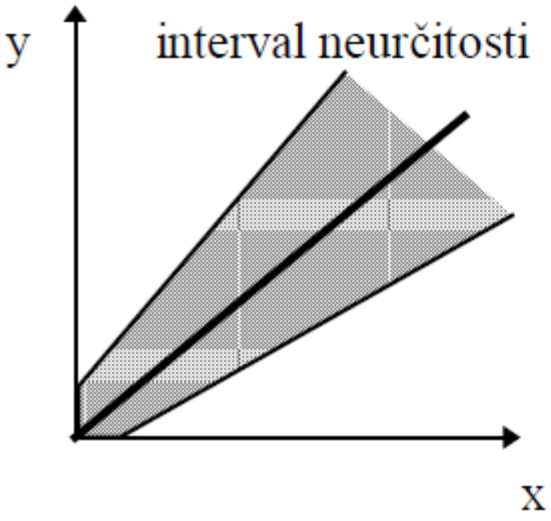
$$\Delta_y = \Delta_K \cdot x$$

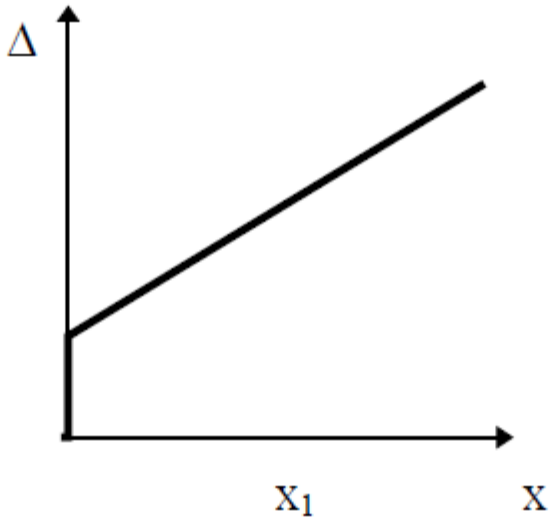
$$\delta_y = \frac{\Delta_y}{y} = \delta_K = kon\check{s}t.$$

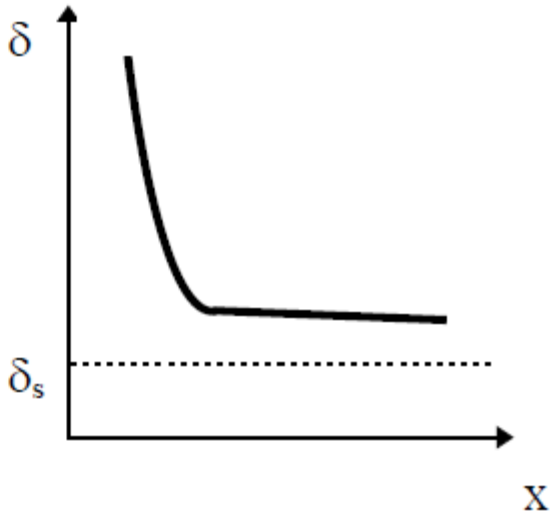
δ_K Chyba měření

Chyby měření

- Kombinovaná chyba měření
- Kombinace aditivní a multiplikativní chyby







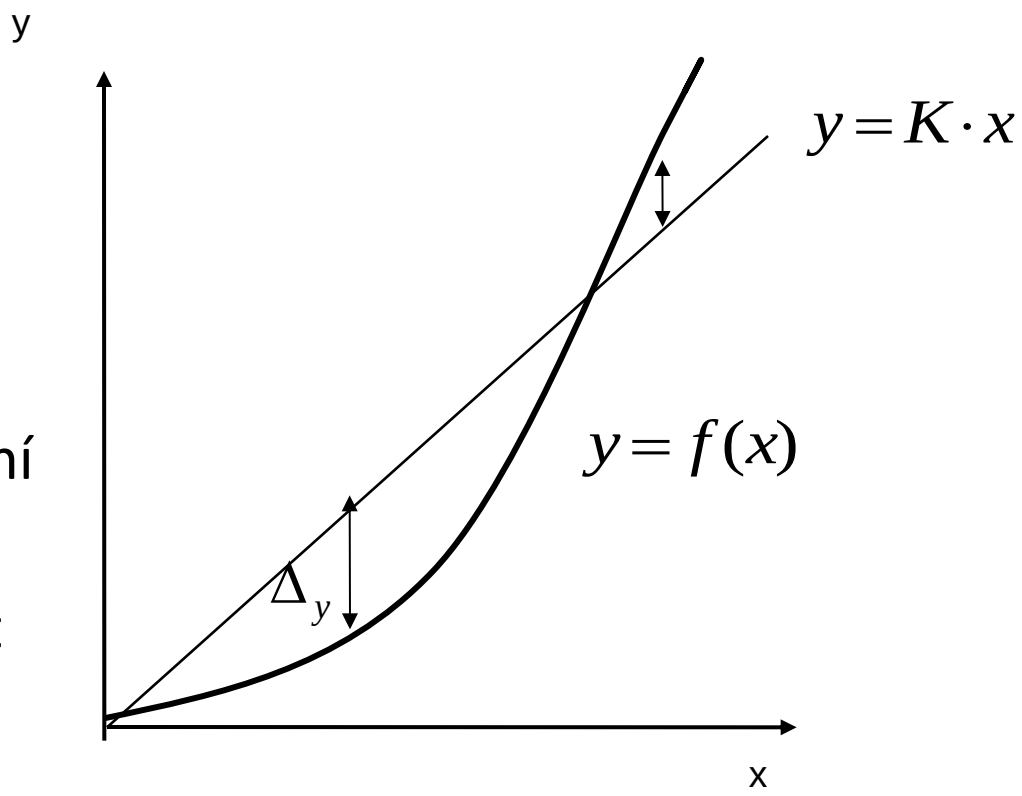
Chyby měření

Chyba linearity

- Dána odchylkou od ideální lineární charakteristiky
- je udávána vztahem:

$$\delta_L = \left(\frac{y_n - y_L}{y_{\max} - y_{\min}} \right)_{\max}$$

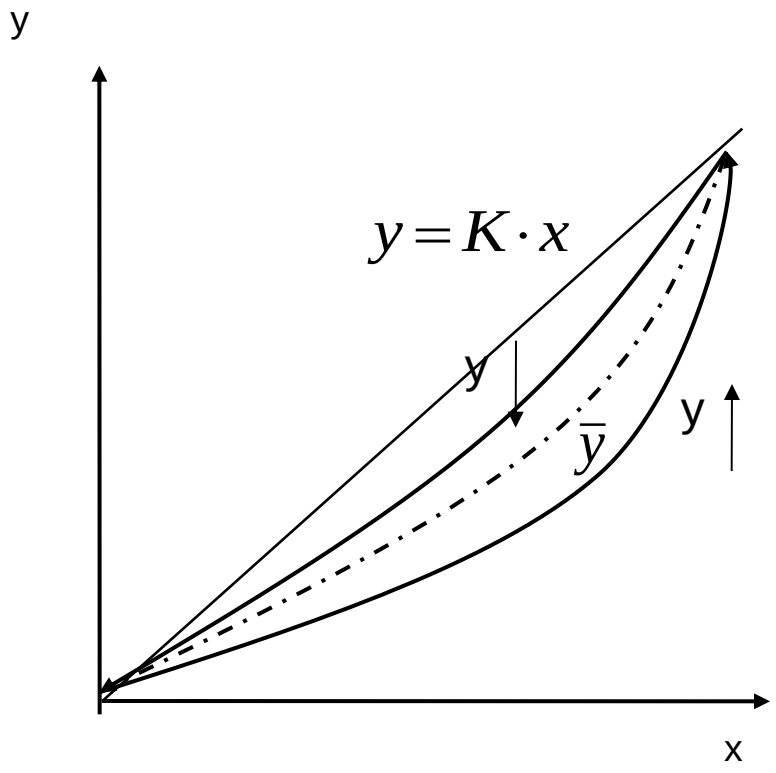
- kde y_L je definována ideální funkcí $y = y_0 + Kx$,
- parametr K lze odhadnout pomocí lineární regrese.



Chyby měření

Chyba hystereze

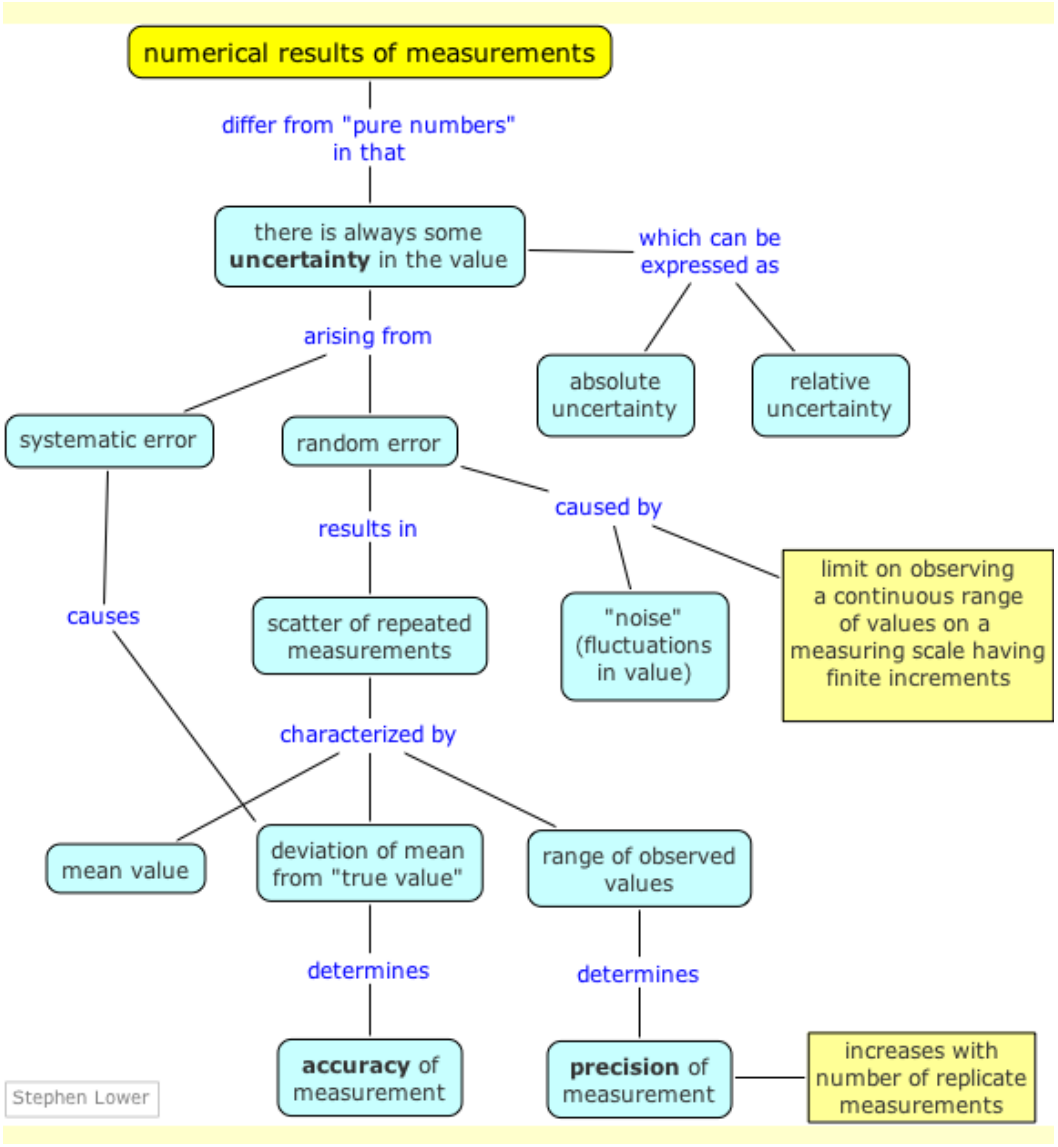
- Vyjadřuje závislost měření na předchozích stavech měřené veličiny (paměťový efekt)



$$\delta = \left(\frac{y - \bar{y}}{y_{max}} \right)$$

kde $\bar{y}$ je střední hodnota vzestupné a klesající křivky závislosti $y = f(x)$

Přehled


numerical results of measurements
differ from "pure numbers" in that
there is always some **uncertainty** in the value
which can be expressed as
absolute uncertainty
relative uncertainty
arising from
systematic error
random error
caused by
results in
causes
scatter of repeated measurements
"noise" (fluctuations in value)
limit on observing a continuous range of values on a measuring scale having finite increments
characterized by
mean value
deviation of mean from "true value"
range of observed values
determines
determines
accuracy of measurement
precision of measurement
increases with number of replicate measurements
Stephen Lower

Obsah prezentace

- Měřené veličiny
- Chyby měření
- **Základní charakteristiky dat**
- Vizualizace dat
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

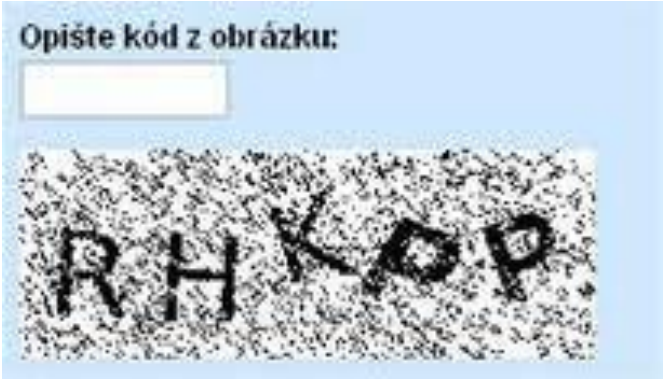
Co je to průzkumová analýza dat?

- *Exploratory data analysis*: První krok při analýze nových dat
- Kombinace grafických, semigrafických a číselných tabulkový postupů, které podají základní informace o vlastnostech souboru

Cíle

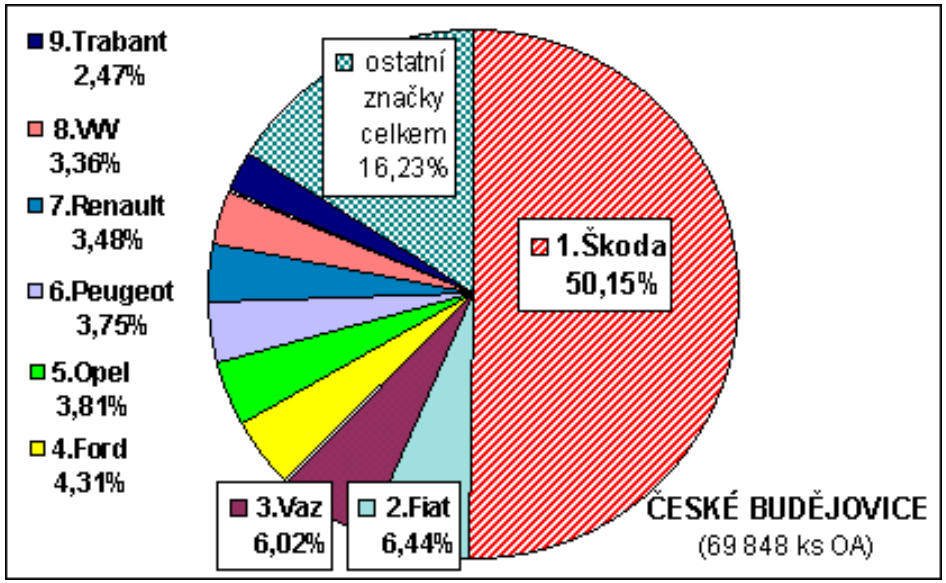
- získat přehled o datech, jejich kvalitě a vlastnostech
- vybrat vhodný nástroj pro předzpracování dat
- využít lidských schopností dříve, než je vybrán automatický nástroj

Lidé jsou schopni rozpoznat charakteristiky dat, které nemohou být rozpoznány (nebo jen velmi obtížně) automatickými systémy:



Frekvence atributu a rozsah hodnot

- **Frekvence atributu**
  - Procentuální vyjádření četnosti výskytu dané hodnoty v datech
  - Na příklad v ČR je frekvence výskytu vozidel Škoda 50,15%
- **Rozsah hodnot**
  - Rozdíl mezi maximální a minimální hodnotou daného atributu



Modus, medián a aritmetický průměr atributu

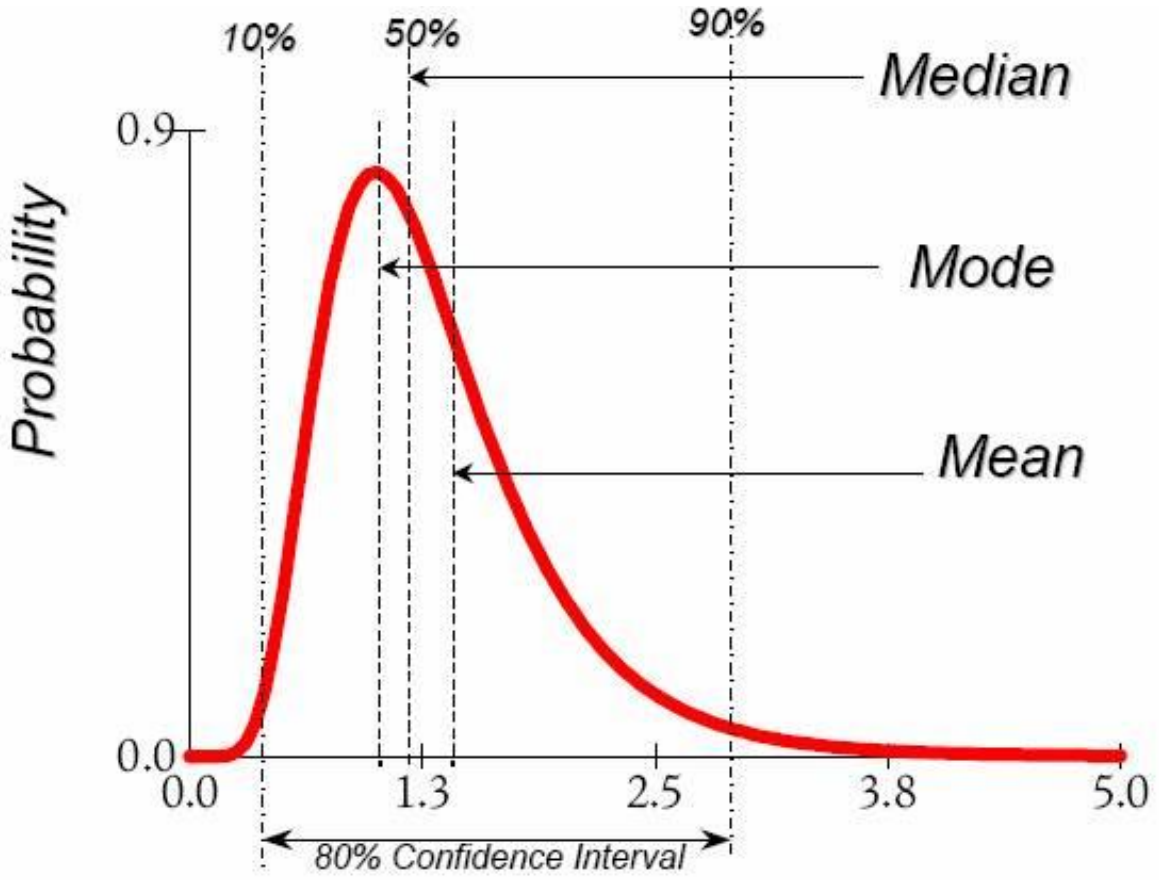
- *Modus atributu (mode)*
  - **nejčastější** hodnota v daném statistickém souboru
  - hodnota znaku s největší relativní četností
  - určení předpokládá roztřídění souboru podle **obměn** znaku
- *Medián (median)*
  - hodnota, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny
  - Je-li rozsah statistického souboru sudé číslo, pak je medián určen jako aritmetický průměr dvou prostředních hodnot
  - Platí, že 50 % hodnot je menších nebo rovných a 50 % hodnot je větších nebo rovných mediánu

Modus, medián a aritmetický průměr atributu

- *Aritmetický průměr (mean)*
  - statistická veličina, která v vyjadřuje typickou hodnotu
  - součet všech hodnot vydělený jejich počtem

Modus, aedián a aritmetický průměr atributu

- Aritmetický průměr (mean)
- Modus atributu (mode)
- Medián (median)



Příklad

- Nalezni
  - Modus,
  - Medián
  - Aritmetický průměr (mean)
  - Rozsah hodnot a
  - Frekvenci výskytu hodnoty 13

- pro následující hodnoty: 13, 18, 13, 14, 13, 16, 14, 21, 13

Příklad

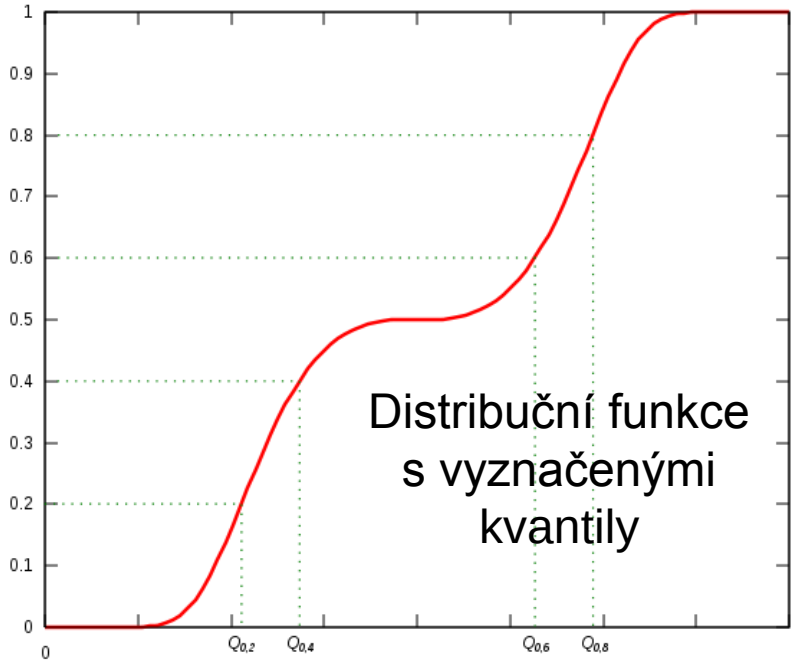
Řešení:

- Aritmetický průměr:
  - (13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) ÷ 9 = 15
- Medián (uspořádej vzestupně): 13, 13, 13, 13, 14, 14, 16, 18, 21
  - Z devíti čísel je střední (9 + 1) ÷ 2 = 5té číslo : 14
- Modus: nejčastěji se opakující číslo: 13
- Rozsah hodnot: 21 – 13 = 8
- Frekvence výskytu hodnoty 13: 4 / 9 *100 = 44,4%

Kvantily

Dělí soubor seřazených hodnot na několik stejně velkých částí:

- **Medián** – kvantil $Q_{0,5}$
  - rozděluje statistický soubor na dvě stejně početné poloviny
- **Kvartil** – kvantil $Q_{0,25}$ a $Q_{0,75}$ rozděluje na horní/dolní čtvrtinu
  - 25 % prvků má hodnoty menší, než dolní kvartil $Q_{0,25}$ a
  - 75 % prvků hodnoty menší, než horní kvartil $Q_{0,75}$
- **Percentil** – kvantil Q_k
  - dělí statistický soubor na setiny, jako k-tý percentil označujeme Q_k / 100.



Obsah prezentace

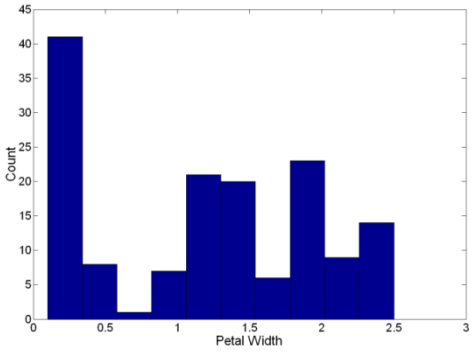
- Měřené veličiny
- Chyby měření
- Základní charakteristiky dat
- **Vizualizace dat**
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

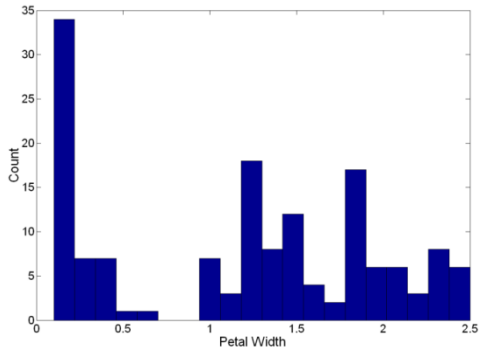
Vizualizace

- převedení dat do vizuální či tabulkové podoby pro potřeby analýzy dat
- velmi silným nástrojem pro **průzkumovou analýzu** dat.
  - Lidé mají velkou schopnost analyzovat velké množství dat prezentované vizuálně
  - Je možné identifikovat obecné trendy a struktury
  - Je možné identifikovat obecné outliery
- Techniky:
  - Histogram
  - Box plot
  - Korelační diagram

Vizualizační techniky: Histogram

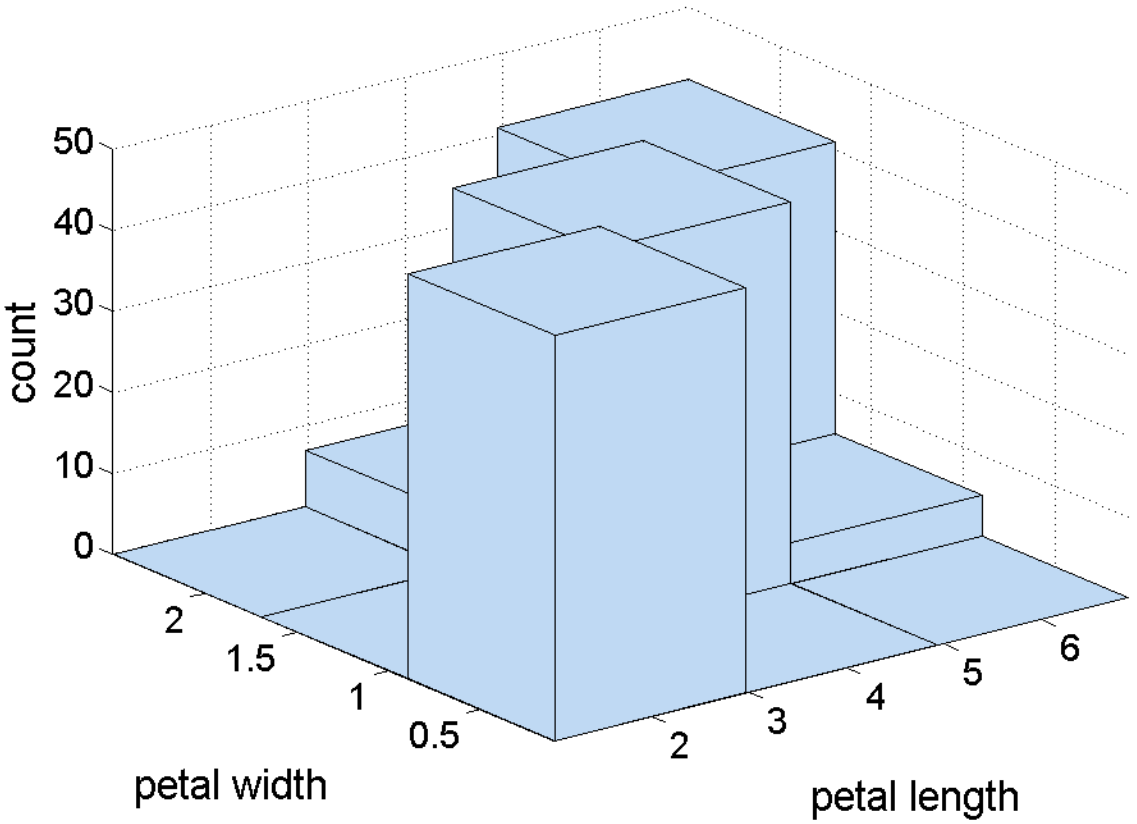
- Histogram
  - Rozdělí hodnoty do intervalů a zobrazí jejich četnosti
  - Výška sloupce udává počet objektů v daném intervalu
  - říká zda je soubor homogenní, nebo zda se rozpadá do dílčích menších podsouborů
    - jen jedna nejčetnější hodnota (homogenní soubor)
    - více hodnot s většími četnostmi
  - někdy lze zjistit přítomnost extrémních výchylek v datech





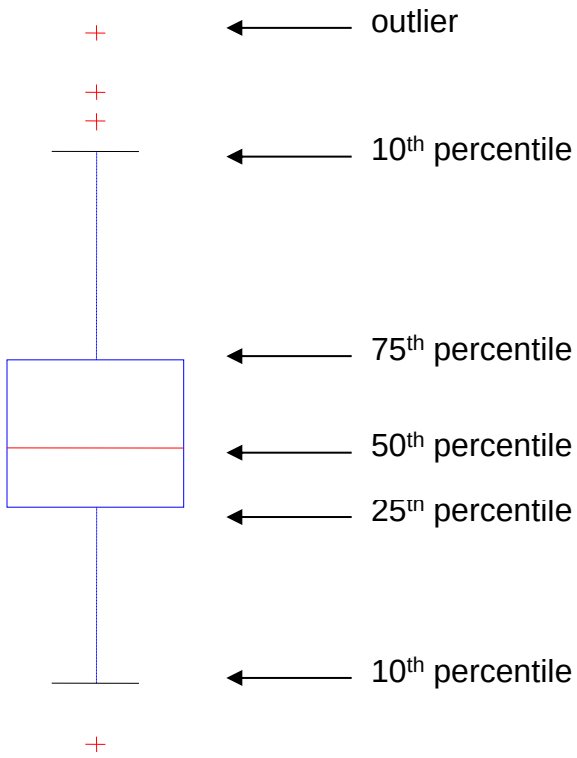
Dvoudimensionální Histogram

- Zobrazuje společné rozdělení dvou atributů



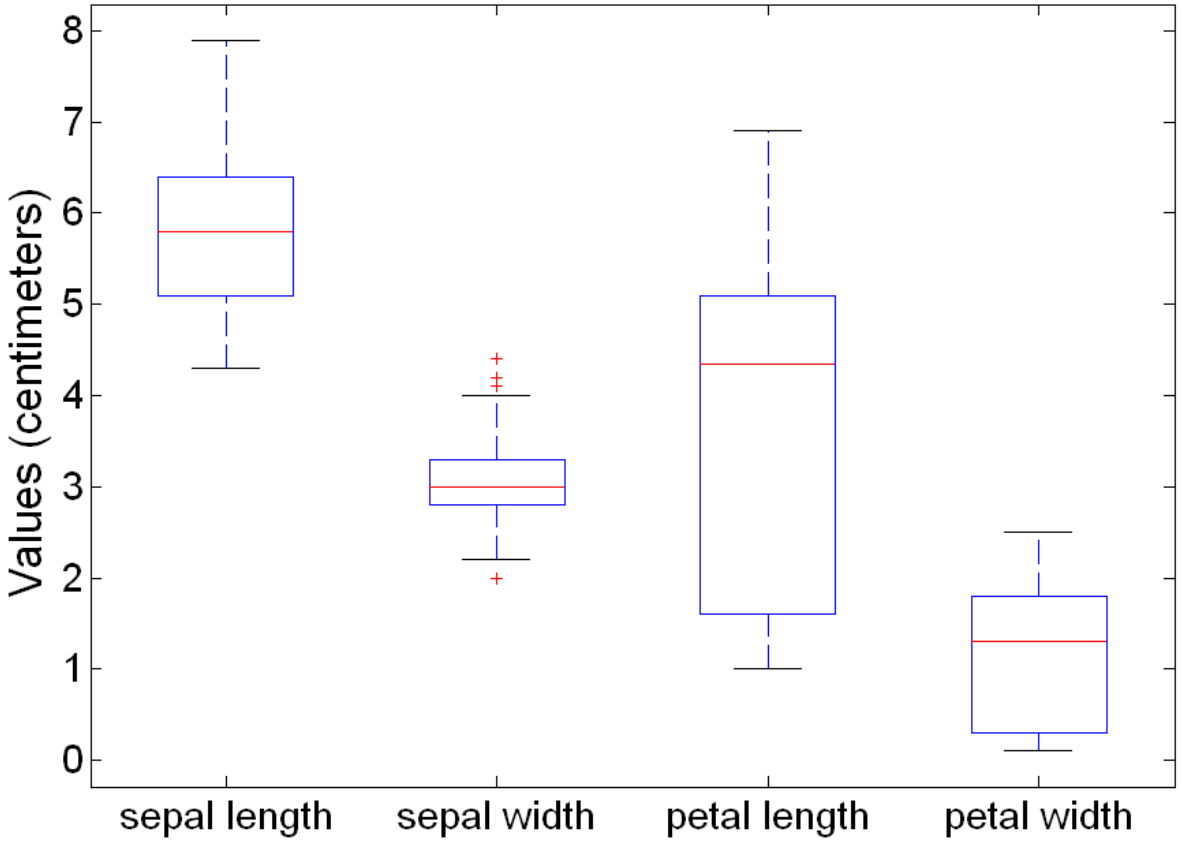
Vizualizační techniky: Box Plots

- Box Plots (J. Tukey)
  - grafické zobrazení tzv. 5-číselného souhrnu



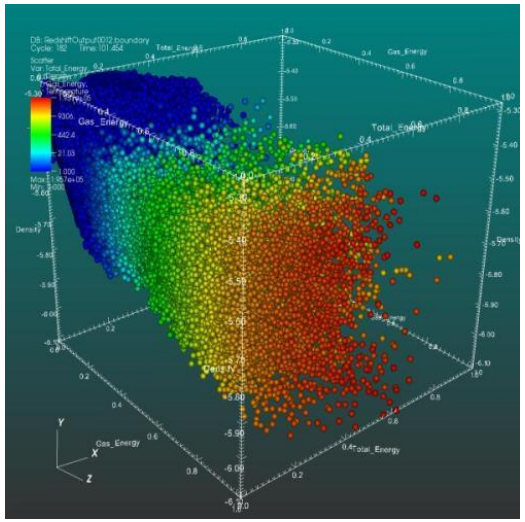
Příklad Box Plots

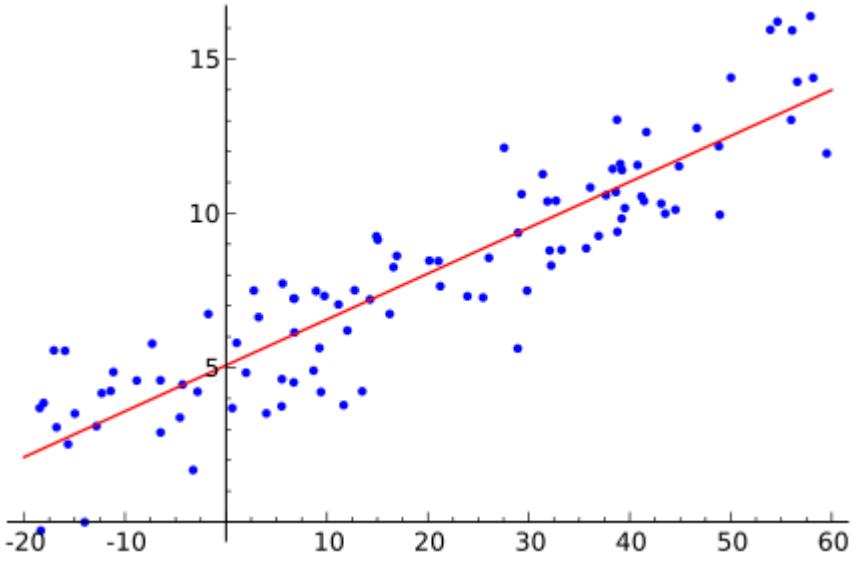
- Box plots se využívají k porovnání atributů



Vizualizační techniky: Korelační diagram

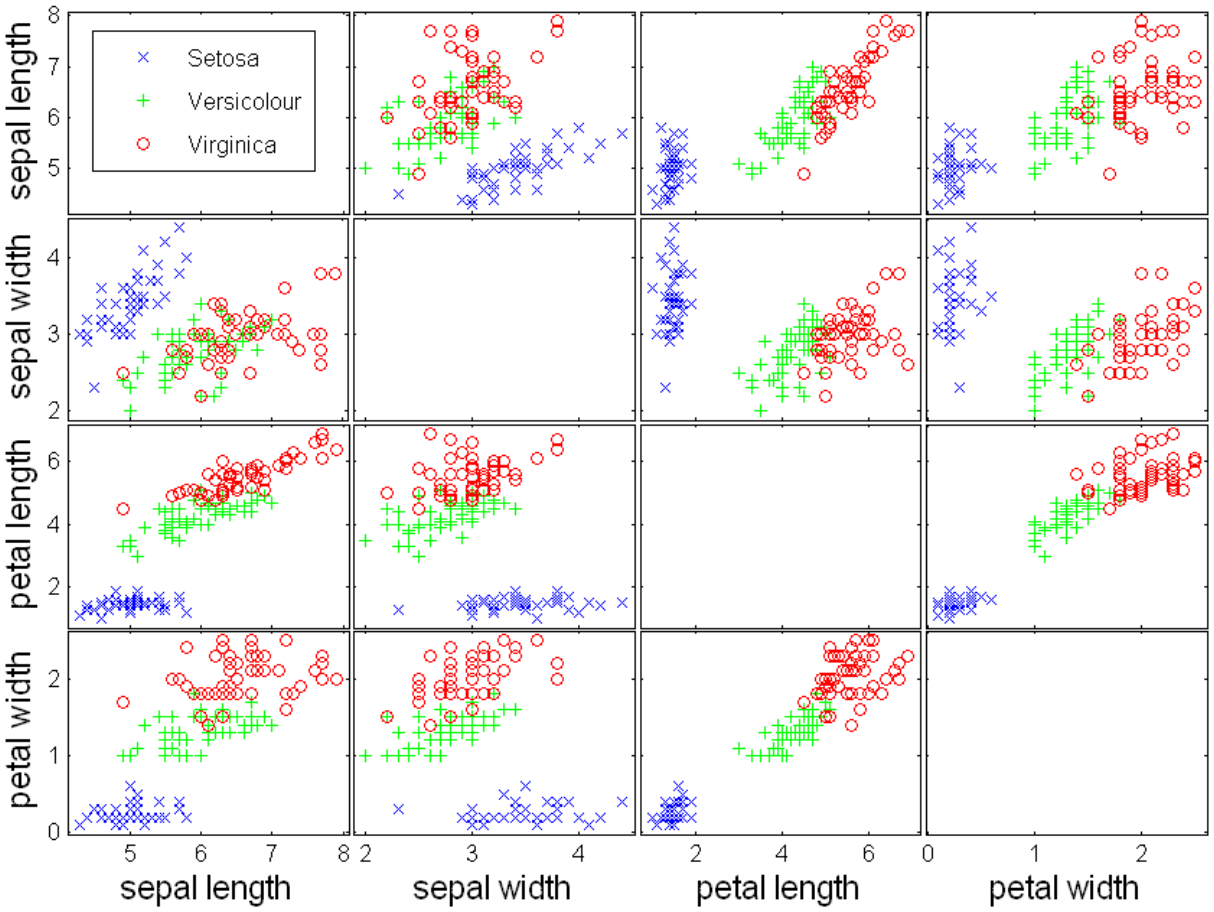
- též bodový graf (angl. *scatter plot*)
- matematické schéma užívající kartézských souřadnic pro zobrazení souboru dat o dvou (či tří) proměnných (na osy).
- je možné jednoduše zjistit vzájemný vztah (korelaci) mezi oběma proměnnými





Pole korelačních diagramů

- Vícerozměrné zobrazení je nepřehledné
- Zobrazuje vzájemné vztahy více proměnných



Obsah prezentace

- Měřené veličiny
- Chyby měření
- Základní charakteristiky dat
- Vizualizace dat
- **Další aspekty analýzy dat**
- Hlavní kroky při analýze dat

Co jsou data?

- Kolekce datových objektů a jejich atributů

- Atribut
  - vlastnost či charakteristika objektu
  - Příklad: barva vozidla, objem motoru, a další
- Datový objekt (záznam (DB), instance, vzorek, entita, ...) je popsán kolekcí atributů

Atributy

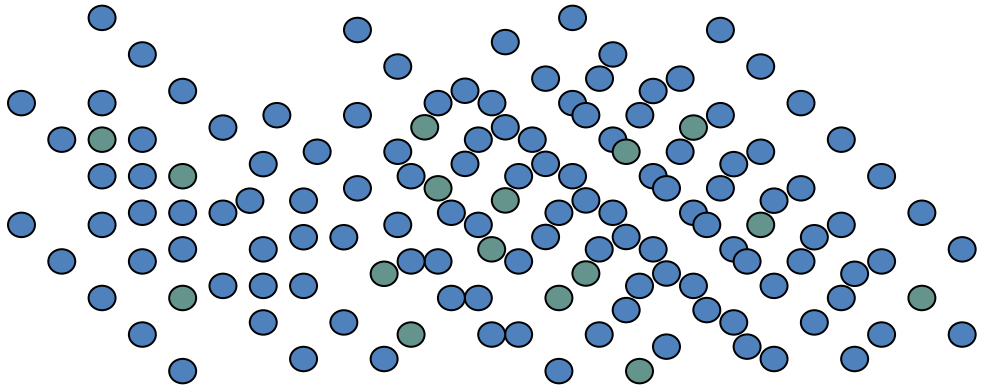
Objekty

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Populace versus náhodný výběr

- *Základní soubor (populace):* všechny jednotky
  - např. všichni řidiči v ČR
  - označujeme se písmeny řecké abecedy (μ, σ, ...)
- *Výběrový soubor:* vybrané jednotky, náhodný výběr
  - např. všichni řidiči, kteří v konkrétním dni jeli autem a stali se účastníky dotazníku
  - označujeme písmeny latinské abecedy ($\bar{x}$, s, ...)

Zdroj http://www.nedarc.org/



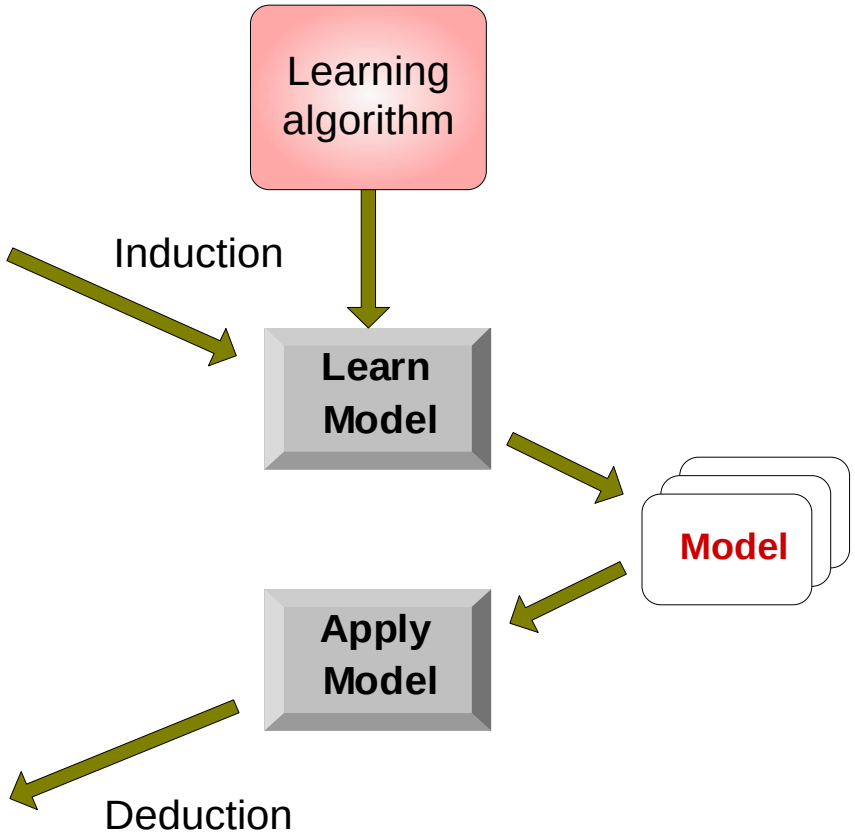
Aplikace modelu

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

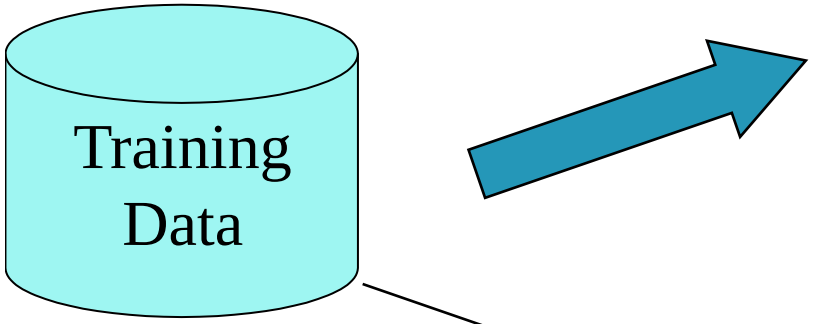
Training Set

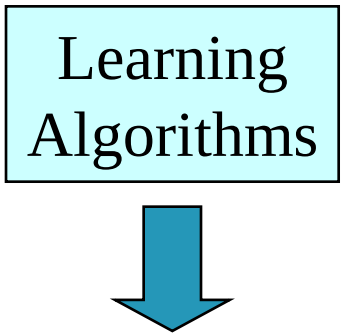
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

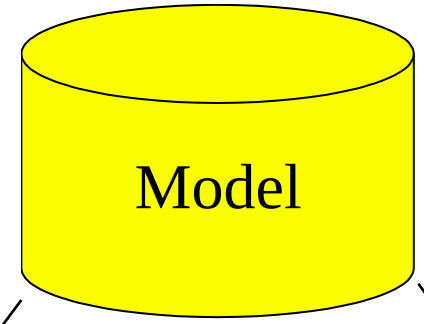
Test Set



Indukce



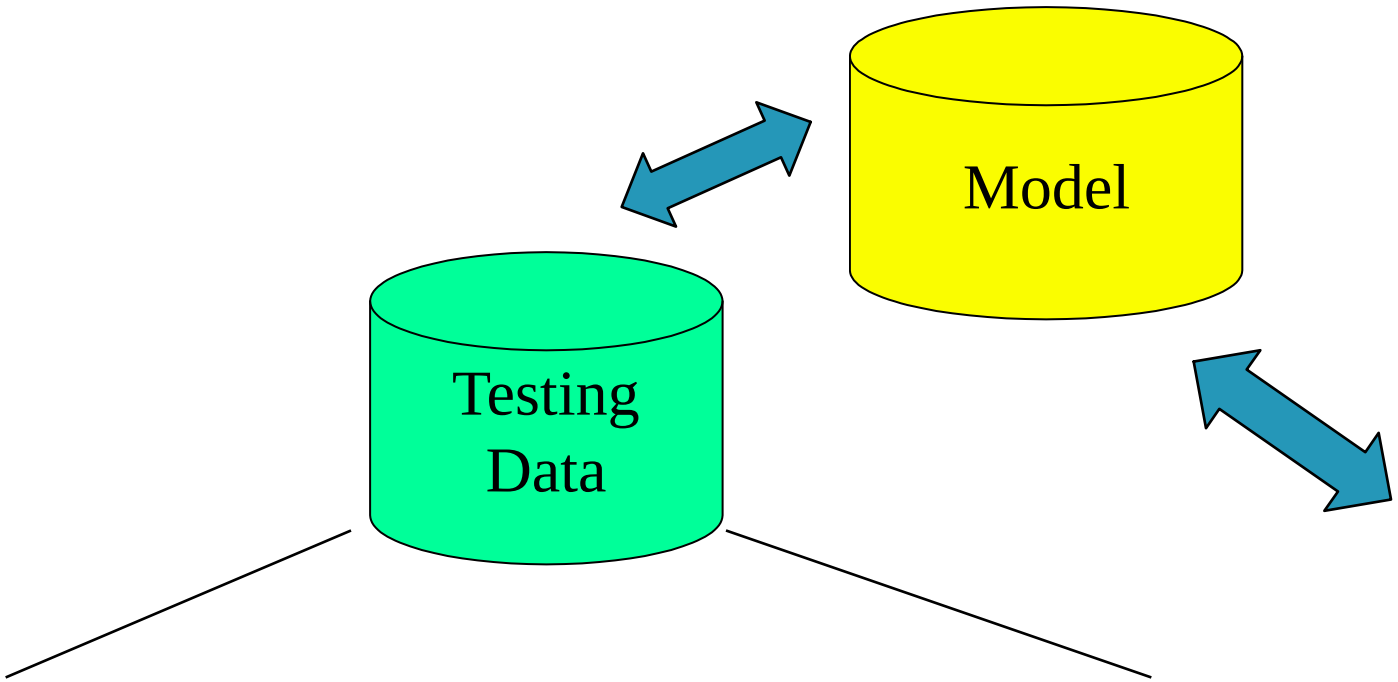




NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

Dedukce



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Tenured?

Yes

Jak vyhodnotit výsledný model?

Podle prediktivní schopnosti, ne podle rychlost algoritmu:

- *Matice záměn (Confusion Matrix)*
  - TP je počet správných predikcí, že daná instance je negativní
  - FN je počet nesprávných predikcí, že daná instance je negativní
  - FP je počet nesprávných predikcí, že daná instance je pozitivní
  - TN je počet správných predikcí, že daná instance je pozitivní
- (ROC křivka)

	PREDIKOVANÁ TŘÍDA		
SKUTEČNÁ TŘÍDA		Třída=Ano	Třída=Ne
	Třída=Ano	TP	FN
	Třída=Ne	FP	TN

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metriky pro porovnání metod

- Přenost (accuracy)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

	PREDIKOVANÁ TŘÍDA		
SKUTEČNÁ TŘÍDA		Třída=Ano	Třída=Ne
	Třída=Ano	a	b
	Třída=Ne	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Problémy s touto metrikou (přesnost)

Máme problém s dvěma třídami

- Počet vzorků pro třídu 0 = 9990
- Počet vzorků pro třídu 1 = 10

Pokud model predikuje vše do třídy 0, přesnost je

$$\frac{9990}{9990 + 10} = 0.999 \equiv 99.9\ \%$$

Cost Matrix (cena)

	PREDIKOVANÁ TŘÍDA		
SKUTEČNÁ TŘÍDA	$C(i\|j)$	Třída=Ano	Třída=Ne
	Třída=Ano	C(Ano|Ano)	C(Ne|Ano)
	Třída=Ne	C(Ano|Ne)	C(Ne|Ne)

$C(i|j)$: Cena za špatnou klasifikaci vzorku z třídy j do třídy i

Výpočet ceny klasifikace

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

Model M_2	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	250	45
	-	5	200

Accuracy = 80%

*(150+250)/(150+250+60+40)*100*

Cost = 5890

150(-1)+40*100+60*1+250*0*

Accuracy = 90%

Cost = 4255

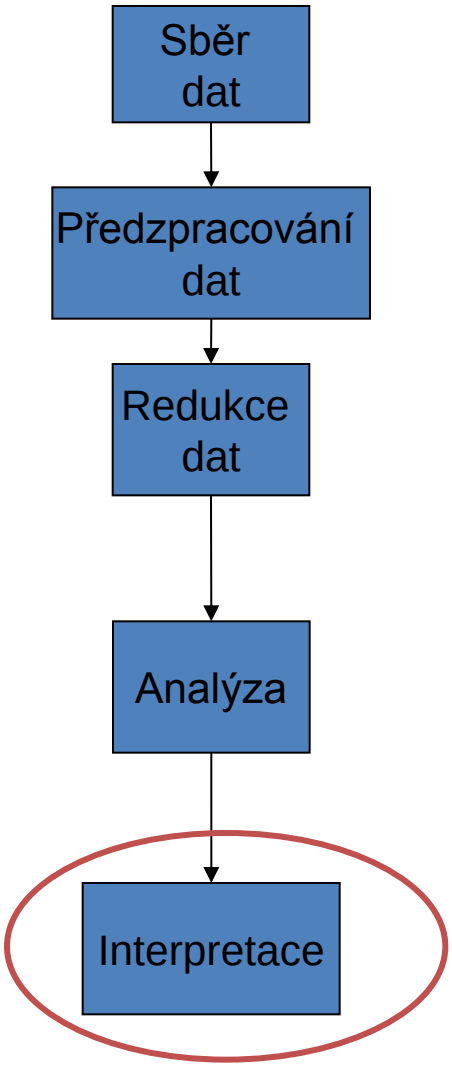
Obsah prezentace

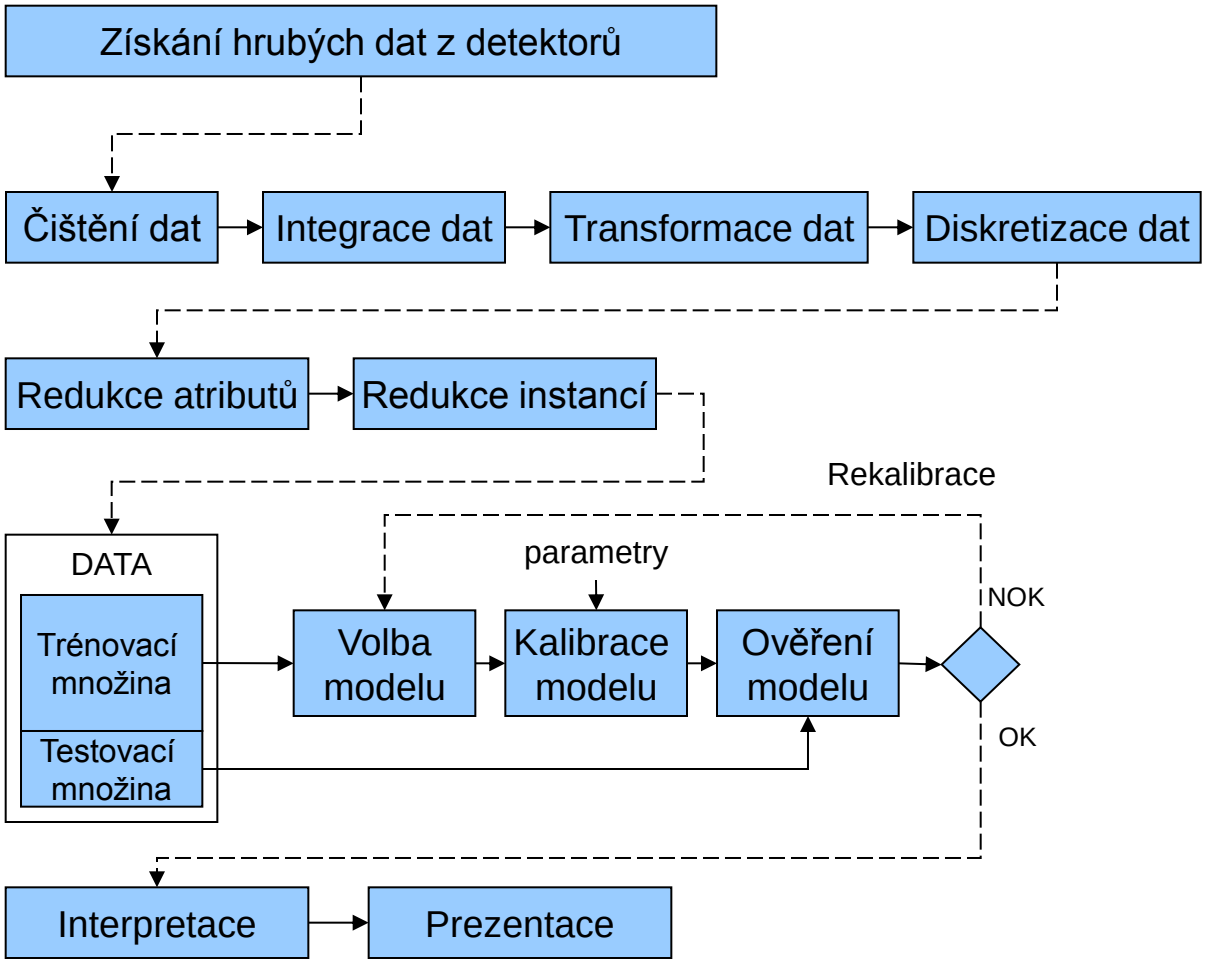
- Měřené veličiny
- Chyby měření
- Vizualizace dat
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

Diskuze

- Co je třeba udělat s daty před aplikací vlastního matematického modelu?
  - Uveďte na příkladech

Hlavní kroky





Interpretace výsledků

Zjistit komu je určen výsledek naší analýzy!

- Forma:
  - Osobní – prezentace
  - Dokument
- Obsah:
  - Technikům: Detailní technický popis
    - Čísla, porovnání, technické zdůvodnění, …
  - Obecné veřejnosti
    - Grafy, tabulky, obrázky
  - Vedoucí manažer, starosta, …
    - Jeden přehledný obrázek
    - Je třeba výsledky prodat (většinou jde o peníze)
    - Na této úrovni je často forma prezentace stejně důležitá jako technický obsah

Co by v prezentaci výsledků nemělo chybět

- Název, datum, kontakty
- Stručný úvod (o co vlastně jde)
- Představení autorů, poděkování sponzorům
- Cíle
- Popis použité metody
  - Proč byla použita tato metoda a ne jiná
  - Základy
- Všechny předpoklady použité při zpracování
  - Je třeba zajistit reprodukovatelnost výsledků
- Výsledky
  - Graficky, přehledně (Je třeba je správně „prodat“)
- Závěr
  - Co se podařilo, ale i co se nepodařilo
  - Další kroky

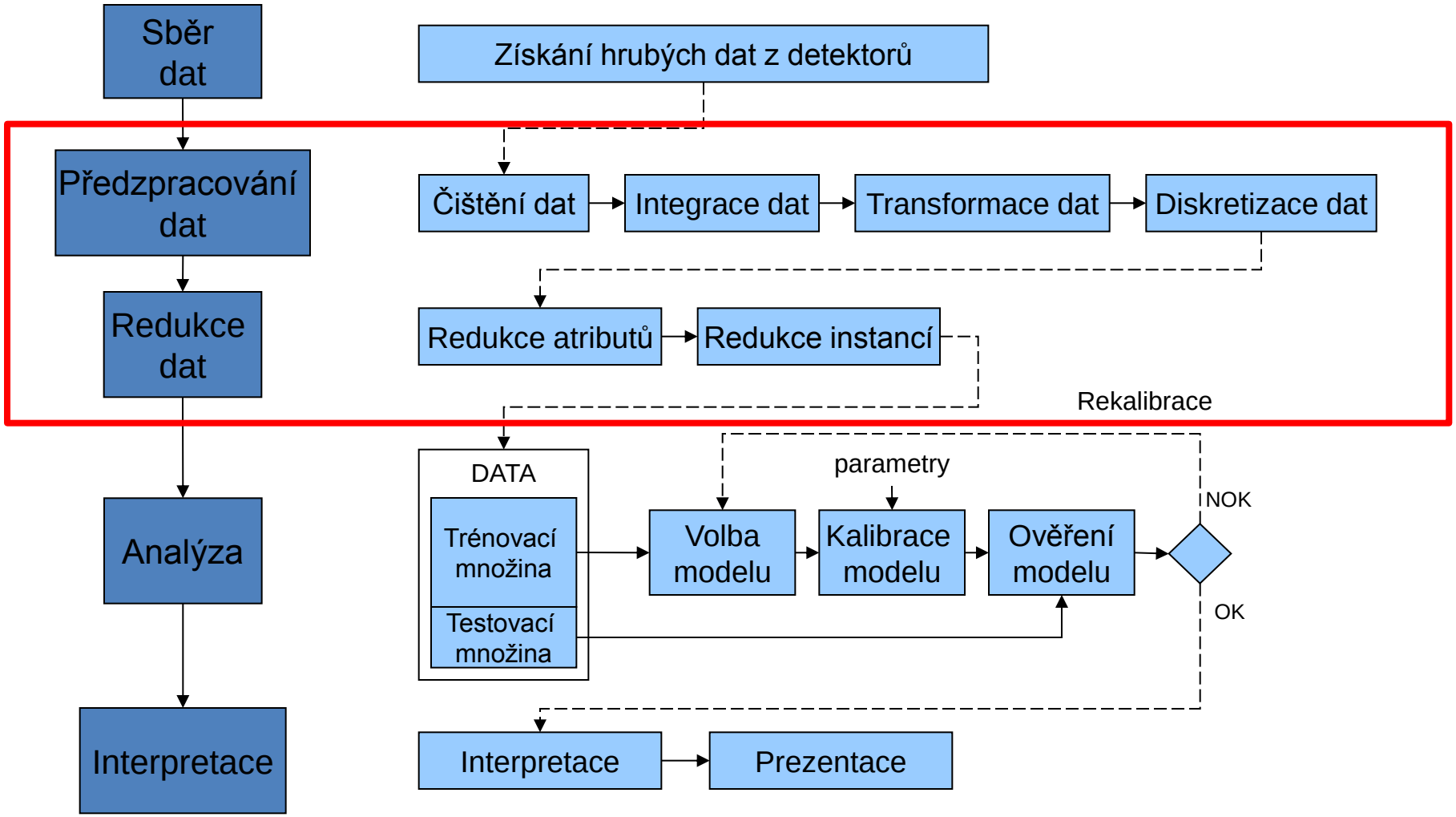
Předzpracování dat

Matematické metody pro ITS (11MAMY)

Ondřej Přibyl (Jan Přikryl)

Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní

Hlavní kroky – obsah prezentace



Diskuse

- Co znamená - předzpracování dat?
- Proč je předzpracování dat nutné?

Proč je třeba předzpracovávat data?

Data v reálném světě jsou „špinavá“

- Nekompletní
- Obsahují šum
- Nekonzistentní
  - Data obsahují protichůdné informace
- Chybná
  - obsahují špatné údaje vzniklé chybami měřicích přístrojů i lidské obsluhy
- Nejednoznačná
  - popsána pomocí příliš mnoha atributů – není zřejmé které jsou relevantní
- Složitá
  - Data mají formu složitého relačního schématu a
  - ne jednoduché tabulky nutné pro matematické algoritmy

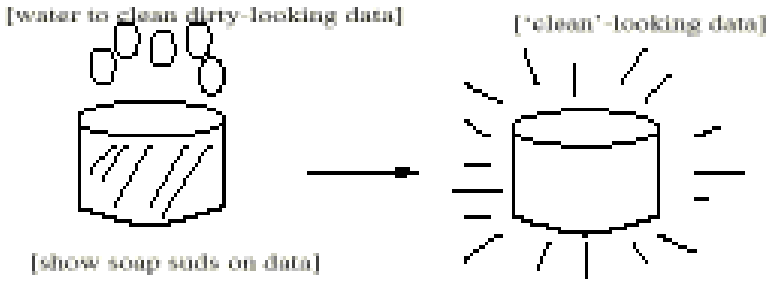
Motto: Pokud nejsou kvalitní data, nebudou kvalitní ani výsledky analýzy!

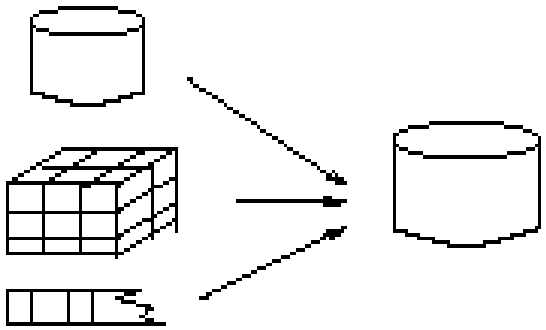
Jak určit kvalitu dat? (Metriky kvality dat)

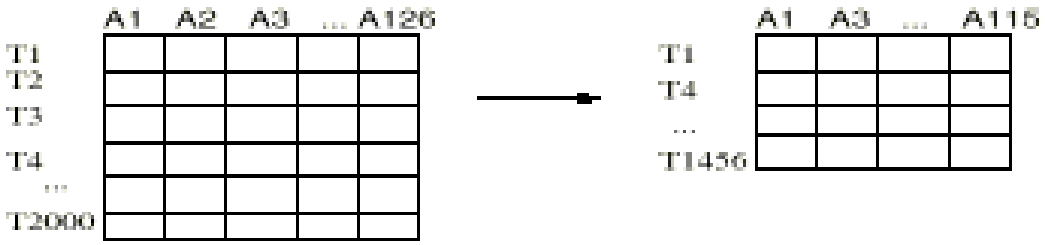
- **Přesnost** (accuracy)
  - měřeno obvykle statistickými charakteristikami pro chybu, např. směrodatná odchylka
- **Úplnost** (completness)
  - zda statistické charakteristiky dat nejsou ovlivněny výběrovými efekty.
- **Konzistence** (consistency)
- **Včasnost** (timeliness)
  - za jakou dobu lze data aktualizovat
- **Důvěryhodnost** (believability)
- **Přidaná hodnota** (added value),
- **Interpretabilita** (interpretability),
- **Dostupnost** (accessibility)
  - Technologické, legislativní a procesní bariéry

Hlavní oblasti předzpracování dat

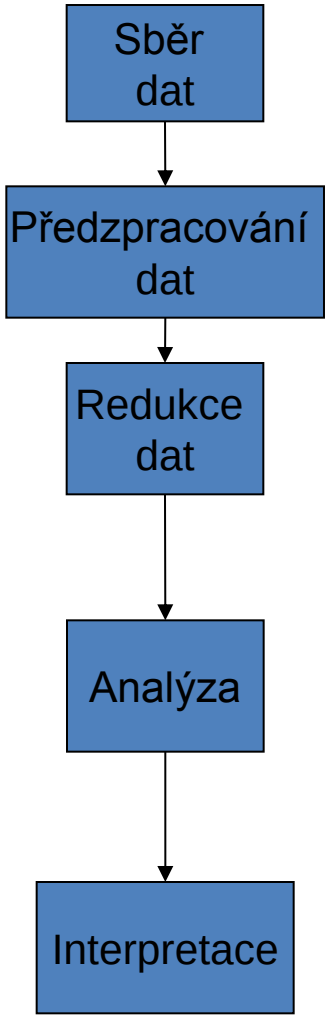
- Čištění dat
- Integrace dat z více zdrojů
- Transformace dat
- Redukce dat
- Diskretizace dat

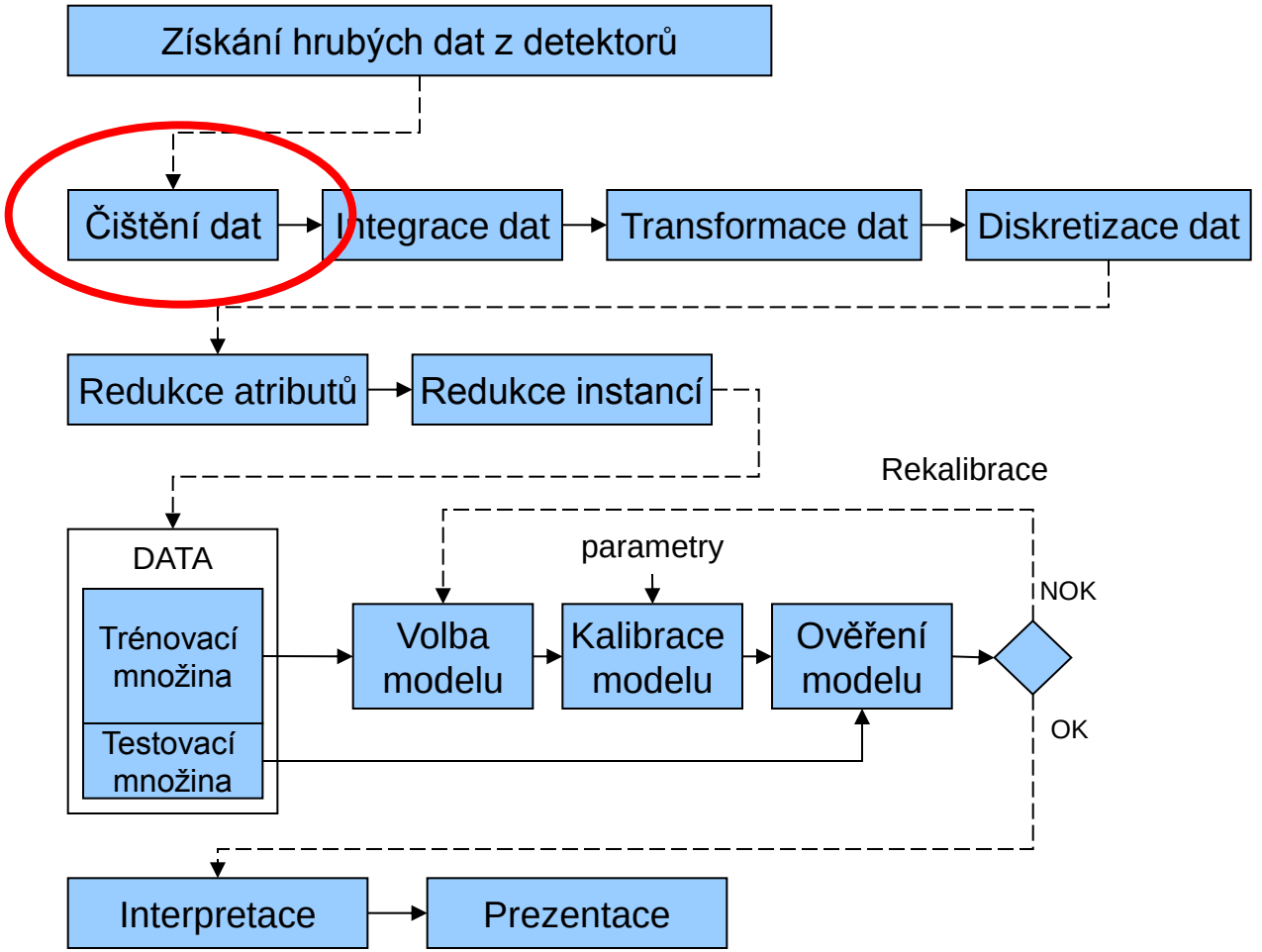






Hlavní kroky





Cíle čištění dat

- Zajistit, že v datech nejsou chybějící či jinak nepřípustné hodnoty

- Identifikuj a nahraď chybějící hodnoty
- Identifikuj extrémní výchylky
- Oprav nekonzistentní data

- Vyhlaď zašuměná data

Rozdělení metod čištění dat

- Dle způsobu měření dat

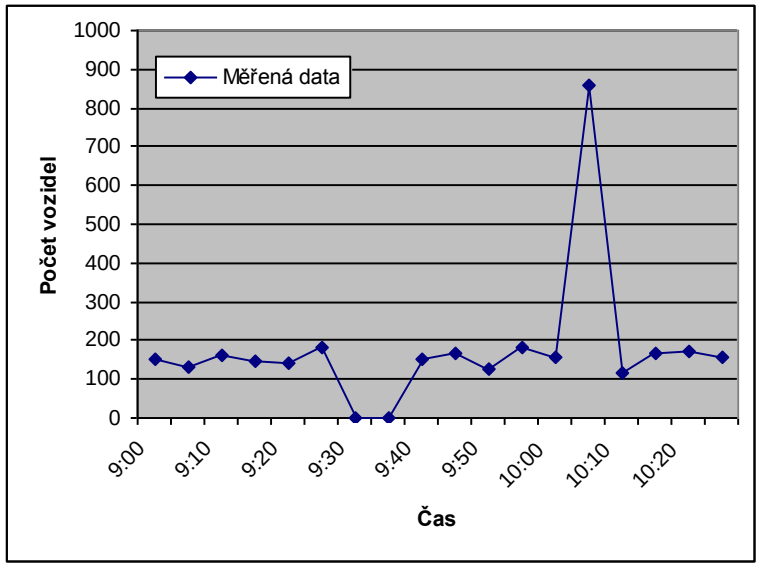
- A. Data v tabulární formě

Atributy

Objekty

Jméno	Zařazení	Roky	Počet publikací
Petr	asistent	3	5
Jan	docent	8	15
Michaela	profesor	12	54
Jiří	asistent	5	?
David	docent	7	?
Jana	asistent	1	1
Martina	asistent	3	12
Petr	docent	5	51
Michal	profesor	10	35
Martin	profesor	8	45

- B. Časové řady



Chybějící data

Jak dojde ke ztrátě dat?

- nefunkčnost senzorů či problém při přenosu a zápisu dat.
- Nevyplněný dotazník
- **Jak opravit následující tabulku?**

Jméno	Zařazení	Roky	Počet publikací
Petr	asistent	3	5
Jan	docent	8	15
Michaela	profesor	12	54
Jiří	asistent	5	?
David	docent	7	?
Jana	asistent	1	1
Martina	asistent	3	12
Petr	docent	5	51
Michal	profesor	10	35
Martin	profesor	8	45

Jak opravit chybějící data - A. Tabulární forma?

- Ignorování vzorku
  - obvykle se použije pokud chybí označení třídy (při klasifikaci)
- Manuální vyplnění
  - náročné, nemožné (vytvoření stejných podmínek, stejné subjekty)?
- Vyplnění globální konstantou
  - např. „Chybějící“
- Vyplnění střední hodnotou všech vzorků
- Střední hodnota pro dané atributy
- Pravděpodobnostní modely
  - regrese, rozhodovací stromy, …

Jak opravit chybějící data - A. Tabulární forma?

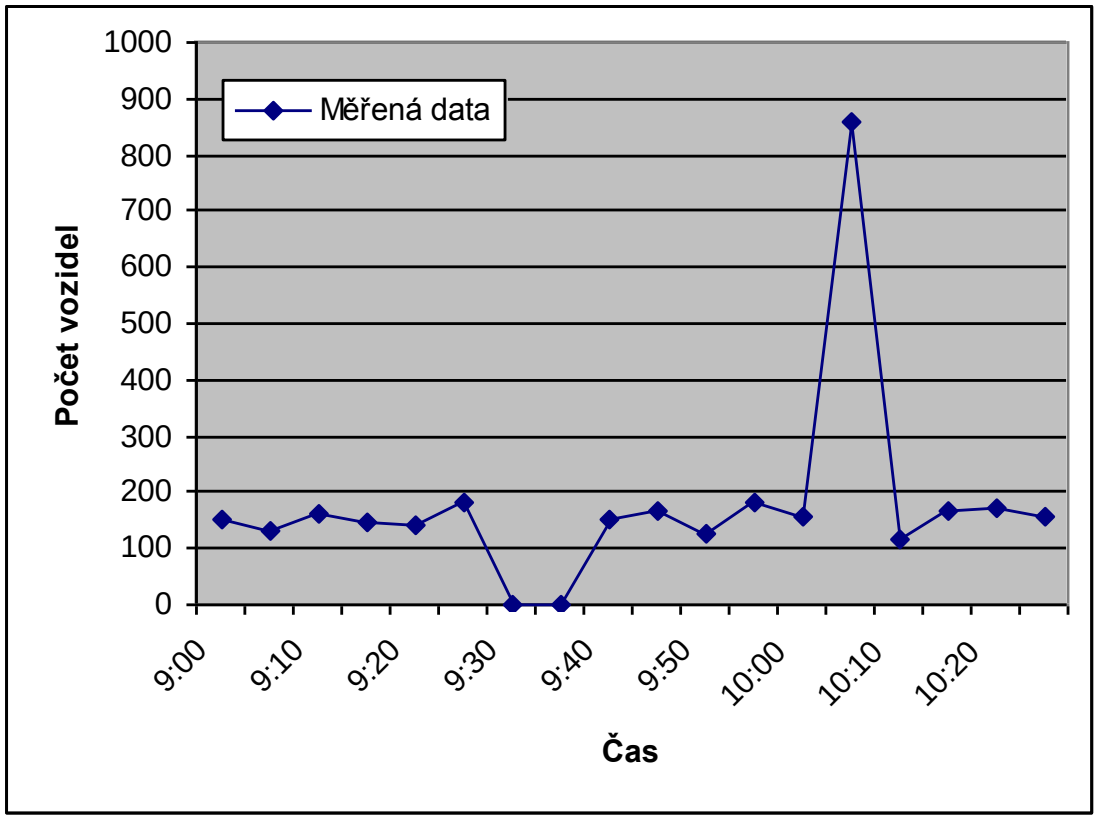
- Vyplnění střední hodnotou všech vzorků
- Střední hodnota pro dané atributy
- Pravděpodobnostní modely (regrese, rozhodovací stromy, …)

Průměr	30
Průměr asistent	6
Průměr docent	33
Průměr profesor	45
Průměrný roční počet publikací	4

Jméno	Zařazení	Roky	Počet publikací
Petr	asistent	3	5
Jan	docent	8	15
Michaela	profesor	12	54
Jiří	asistent	5	?
David	docent	7	?
Jana	asistent	1	1
Martina	asistent	3	12
Petr	docent	5	51
Michal	profesor	10	35
Martin	profesor	8	45

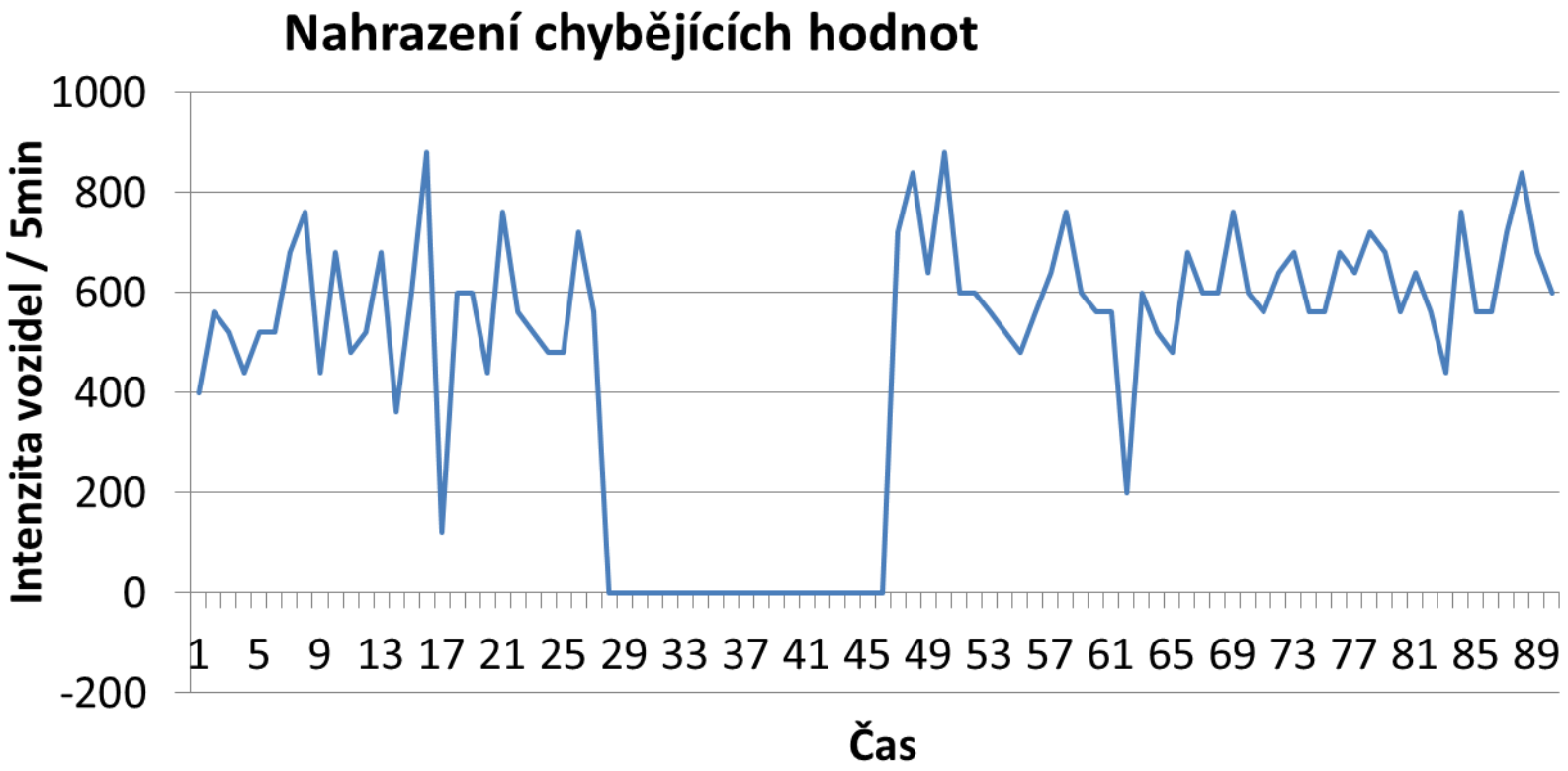
Jak opravit chybějící data - B. Časová řada?

- V prvním kroku je třeba identifikovat nečistá data !
- **Diskuse** - Jak identifikovat chybějící hodnoty v případě časové řady?
- Jak zjistit zda se jedná o naměřenou hodnotu či o chybu?
  - Rozsah hodnot,
  - Statistika,
  - Kontext
  - …



Jak opravit chybějící data - B. Časová řada?

- **Diskuse**
  - Jak nahradit chybějící hodnoty v případě časové řady?



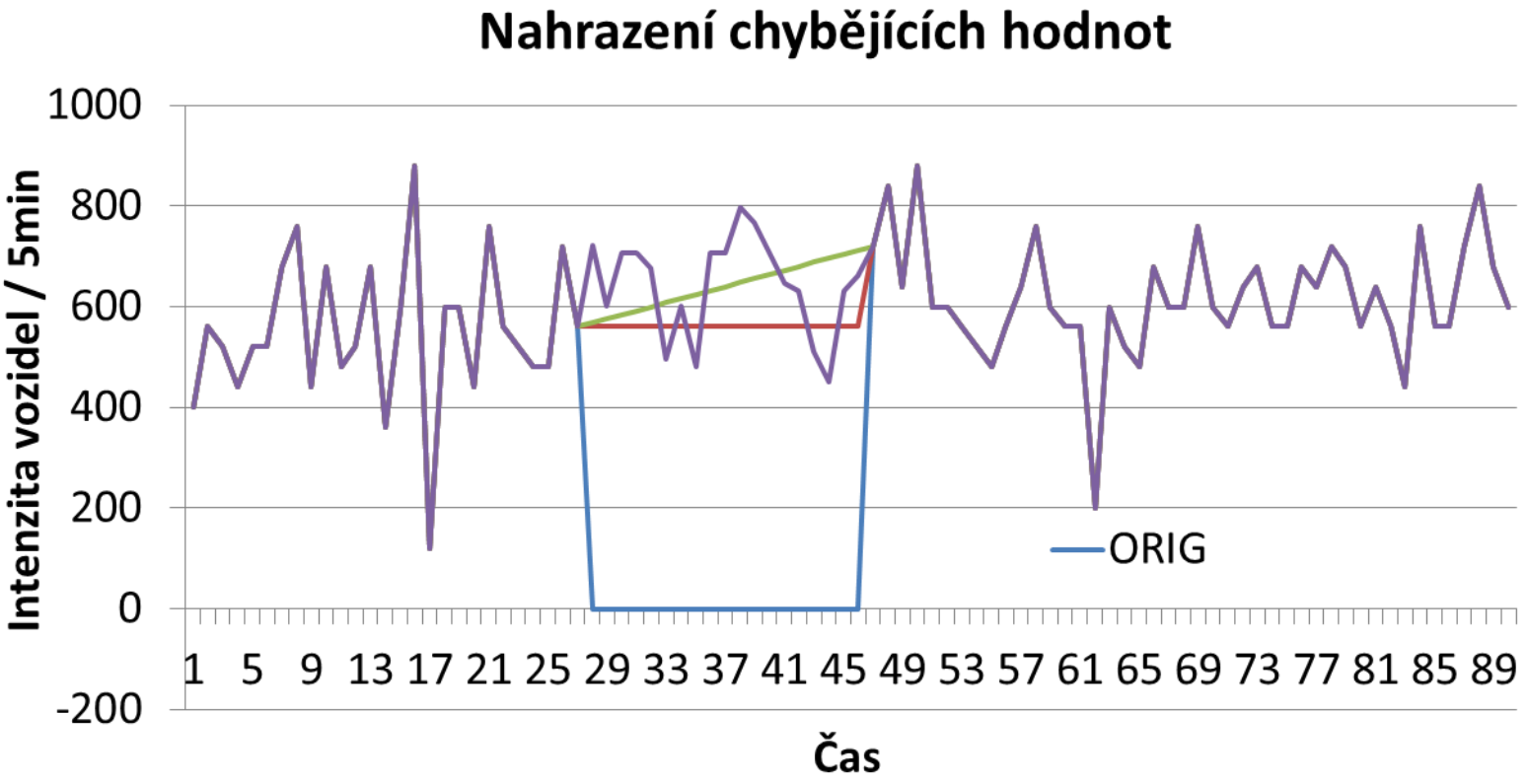
Chybějící hodnoty – časová řada

- Náhrada poslední hodnotou
  - nahradí se poslední správně naměřenou hodnotou.
- Průměr platných hodnot
  - místo chybné hodnoty se použije průměr poslední platné hodnoty před výpadkem a první po výpadku.

- Lineární spojnice platných hodnot
  - Další strana
- Nahrazení dle statického modelu
  - Další strana

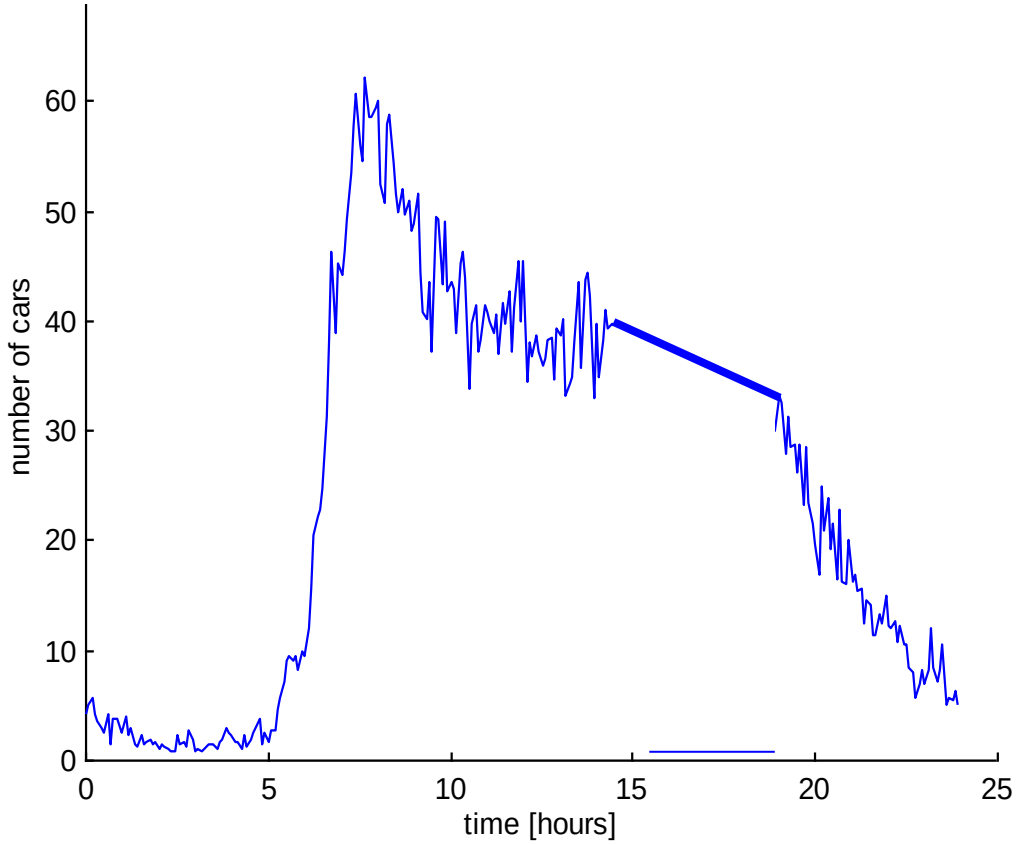
Jak opravit chybějící data - Řešení

- Diskuse
  - Jak identifikovat a nahradit chybějící hodnoty v případě časové řady?



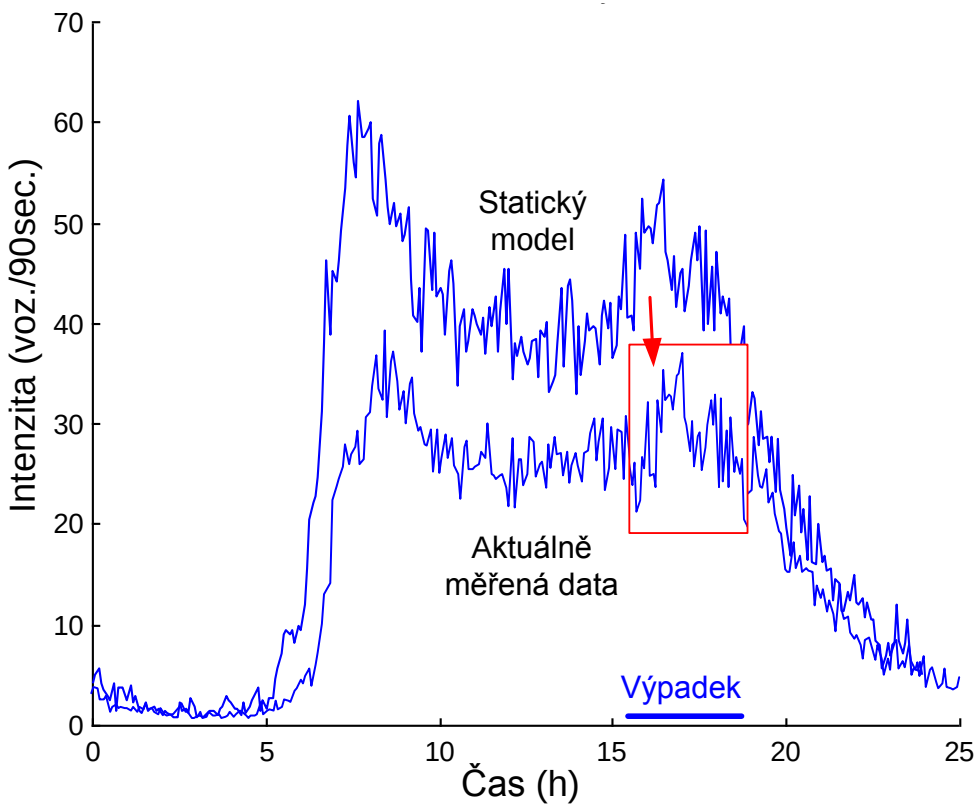
Chybějící hodnoty – časová řada

- Lineární spojnice platných hodnot
  - místo chybné hodnoty se lineální spojnice poslední platné hodnoty před výpadkem a první po výpadku.



Chybějící hodnoty – časová řada

- Nahrazení dle statického modelu
  - „typické“ chování (takzvaný statický model) je možné použít pro odhad chybějících hodnot.
  - úprava (koeficientem) pro přizpůsobení aktuálnímu rozsahu dat

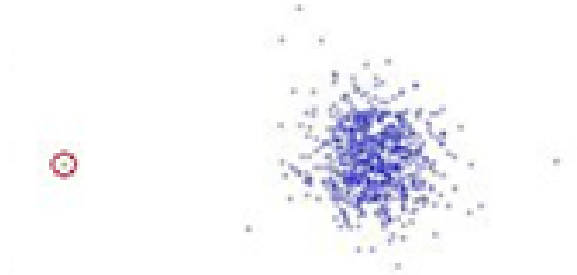


Diskuze

- Co je to outlier?
- Jak identifikovat outliery ?
- Jak odstranit outliery?

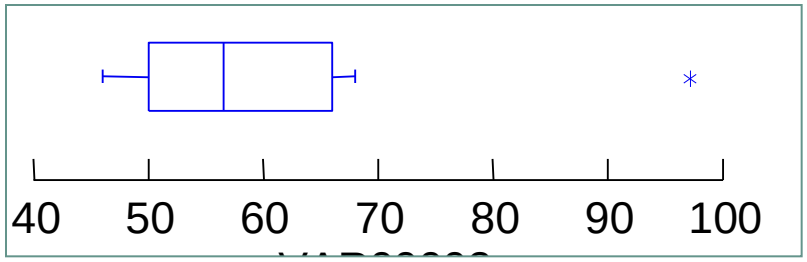
Detekce extrémních hodnot

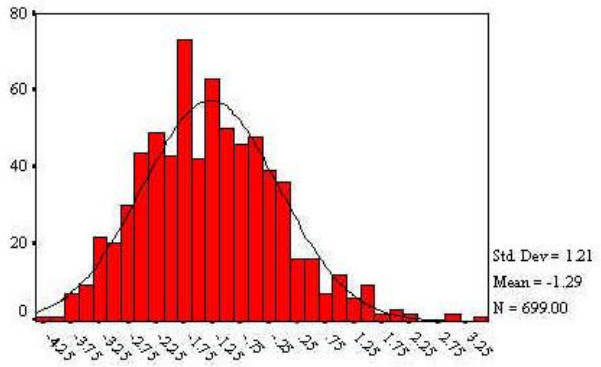
- Outliers je třeba Identifikovat, popřípadě odstranit



- Metody – **jednorozměrné**
  - Gausovské rozložení - Z score
  - Histogram
  - Boxplot

$$z_i = \frac{\left(x_i - \bar{x}\right)}{s}$$

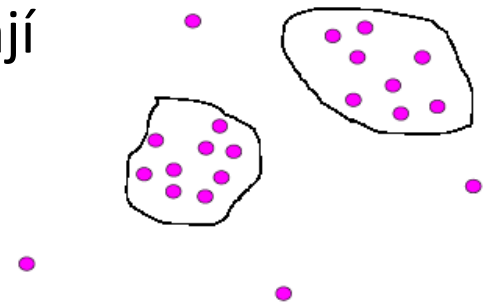




Detekce extrémních hodnot

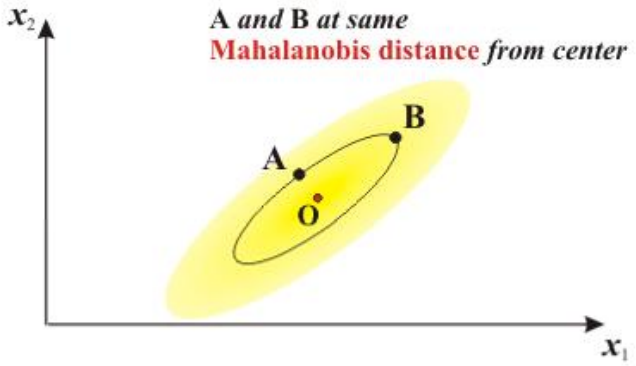
Metody – **vícerozměrné**

- Shlukování
  - Najdi „podobná“ data a odstraň ta která se vymykají

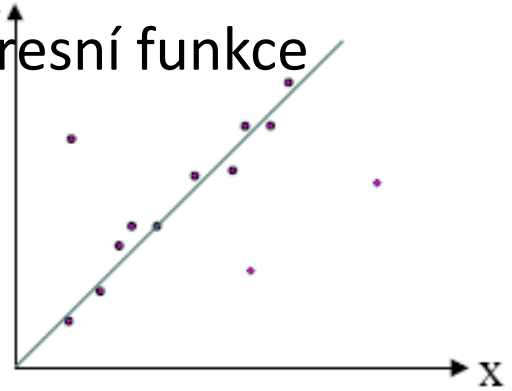


- Binning
  - Seřaď hodnoty atributů a sluč je do skupin (bins)
  - Vyhlaď je podle středních hodnot,

- Mahalanobisova vzdálenost
  - vzdálenost od centroidu



- Regrese
  - Najdi data ležící daleko od regresní funkce



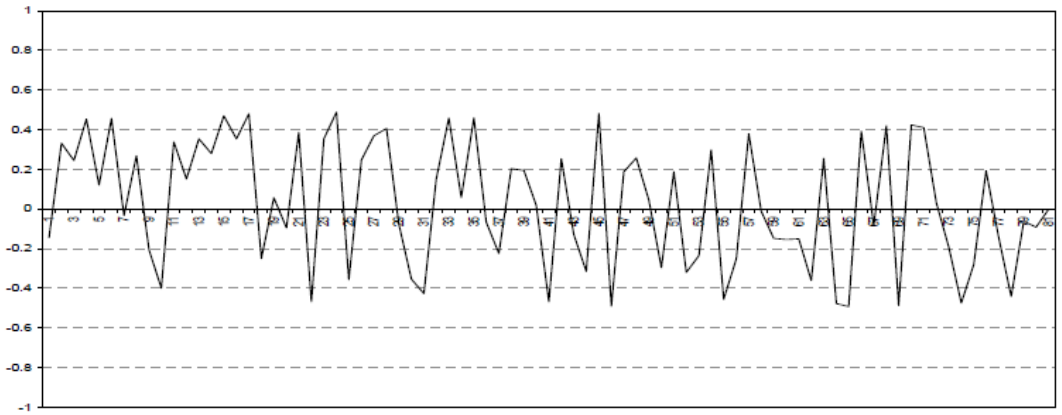
FILTRACE DAT

Diskuze

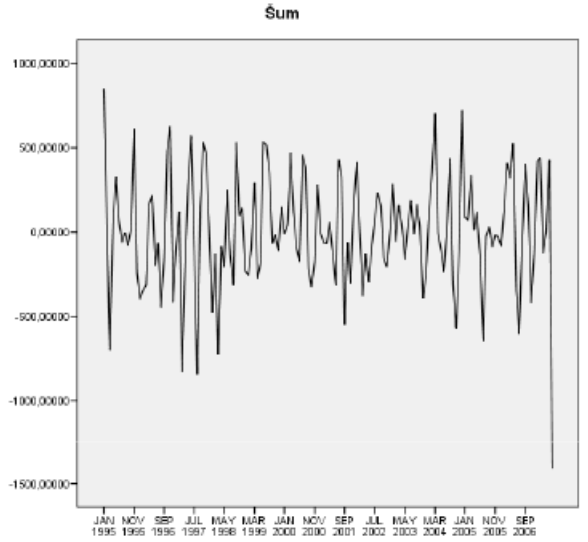
- Co je cílem filtrace?
- Co je to bílý šum (noise)?

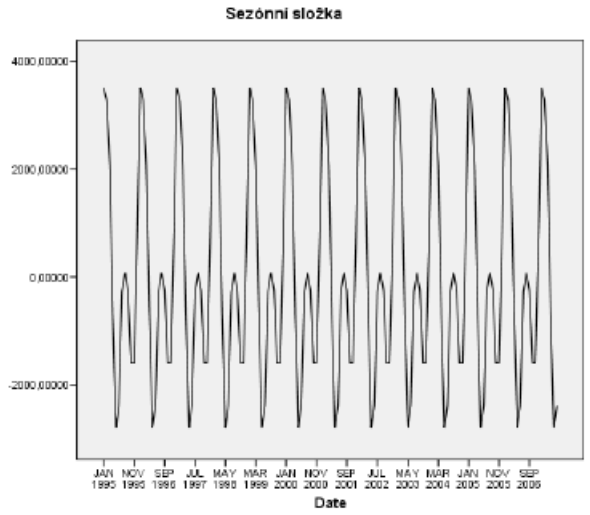
Šum analytického signálu

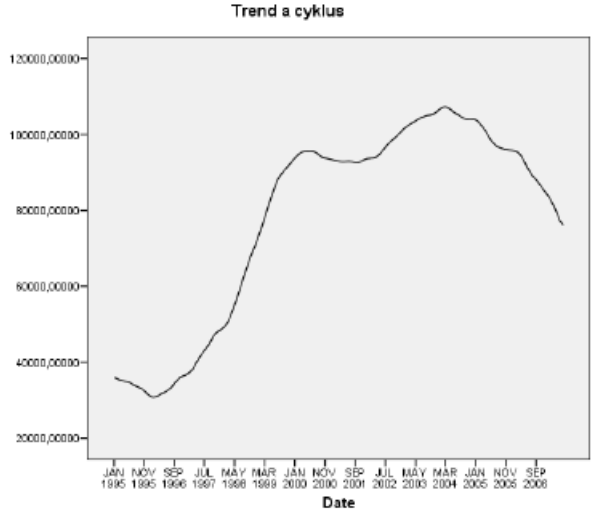
- náhodné zvýšení nebo snížení měřeného signálu
- Šum, jehož suma je **nulová** v časovém intervalu pozorování, se označuje jako bílý šum
- Šum, jehož suma je **nenulová** v časovém intervalu pozorování, se označuje jako náhodný šum
- Šum je významný jen z hlediska intervalu pozorování
- Intenzivní a náhlé změny signálu (spiky) nelze doslovně považovat za šum - jejich původ bývá v okamžitém porušení funkce měřícího zařízení



Dekompozice časové řady







source: http://www.spss.cz/files/ruzne/vsb/casove_rady.pdf

Filtrování dat

- **Důvody pro využití filtrování dat**
  - metoda pro čištění dat
  - odstranění náhodné složky z dat – odstranění šumu (náhodné složky)

A) Filtrace v **časové oblasti**

  - Obvykle se jedná o okno definované velikosti ve kterém se spočítá střední hodnota a ta se použije pro odstranění šumu.

B) Filtrace ve **frekvenční oblasti**

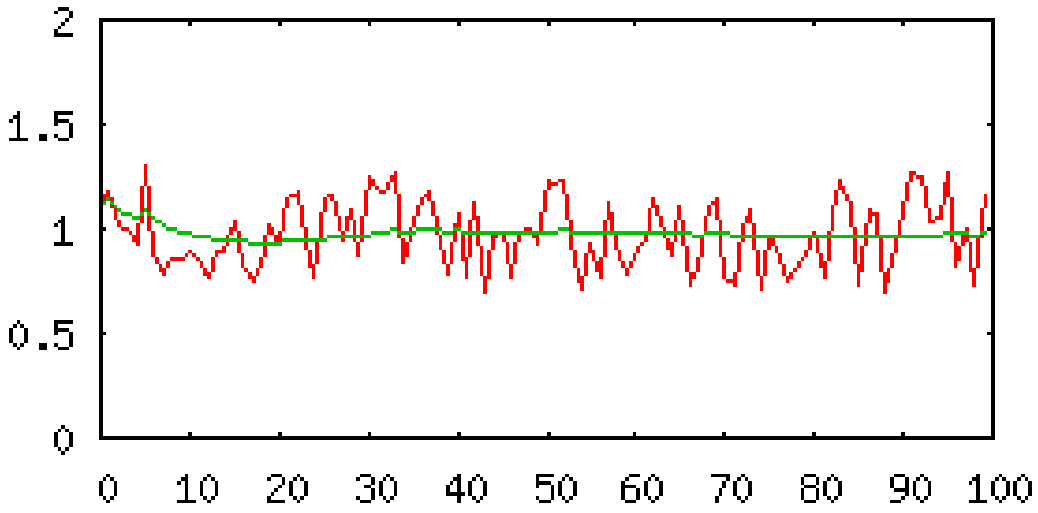
  - Data jsou transformována z časové oblasti do oblasti frekvenční.
  - Odebráním vysokofrekvenčních složek dojde k odfiltrování šumu a náhodných složek.

Filtrace dat v časové oblasti

- Veličiny s konstantní hodnotou
  - jednoduchým průměrováním
- Veličiny s pomalu se měnící hodnotou
  - Plovoucí průměr (moving average)
  - Vážený plovoucí průměr (weighted moving average)
  - Exponenciální vyhlazování (exponential smoothing)

Filtrace dat v časové oblasti

- **Jednoduché průměrování**

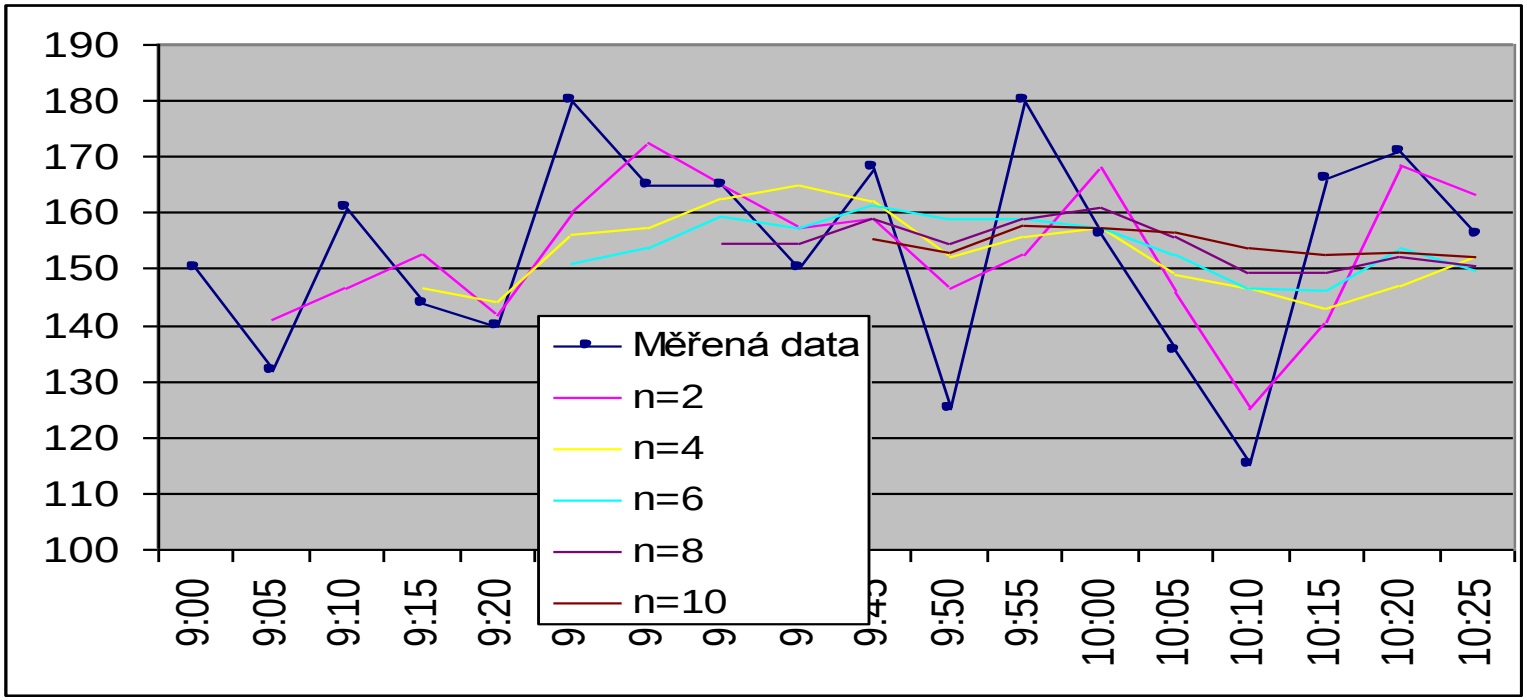


$$\hat{x}_n = \frac{\hat{x}_{n-1} \cdot (n-1) + x_n}{n}$$

Filtrace dat v časové oblasti

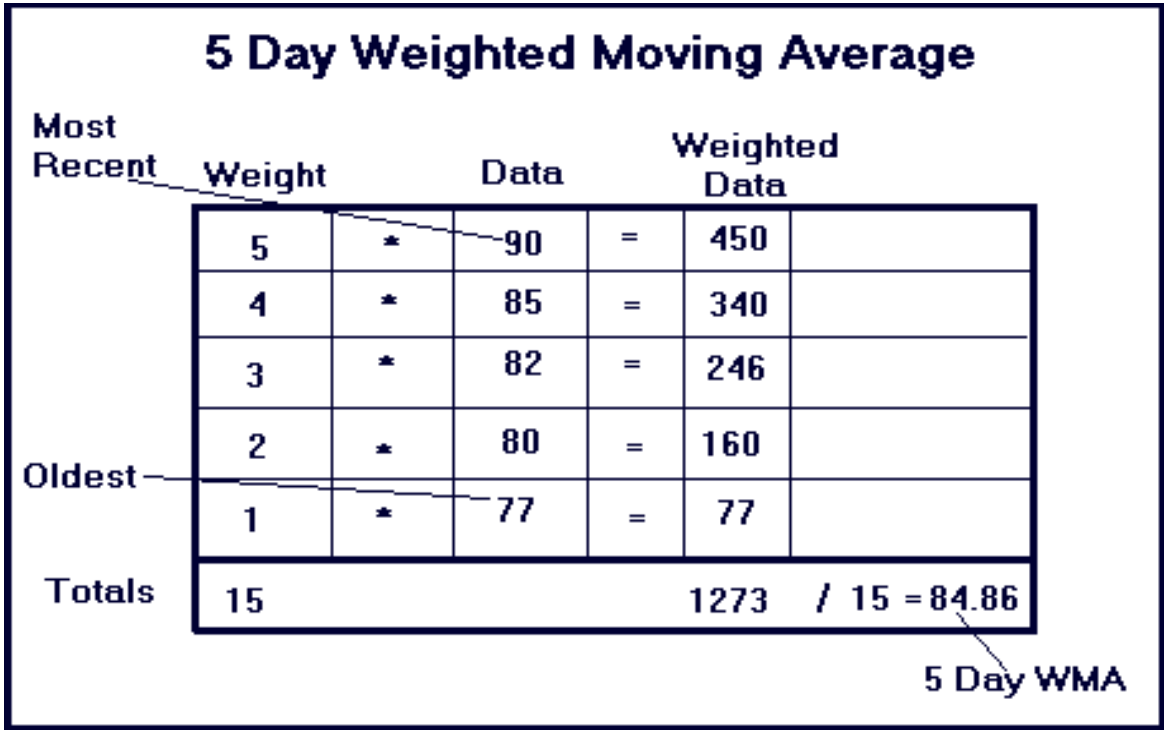
- **Plovoucí průměr** (moving average)
  - Pokud neexistuje periodický cyklus
  - Všechny hodnoty mají stejnou váhu
  - k … velikost ‚paměti‘

$$\hat{x}_n = \frac{1}{k}\sum_{i=1}^{k} x_{(n-i)+1}$$



Filtrace dat v časové oblasti

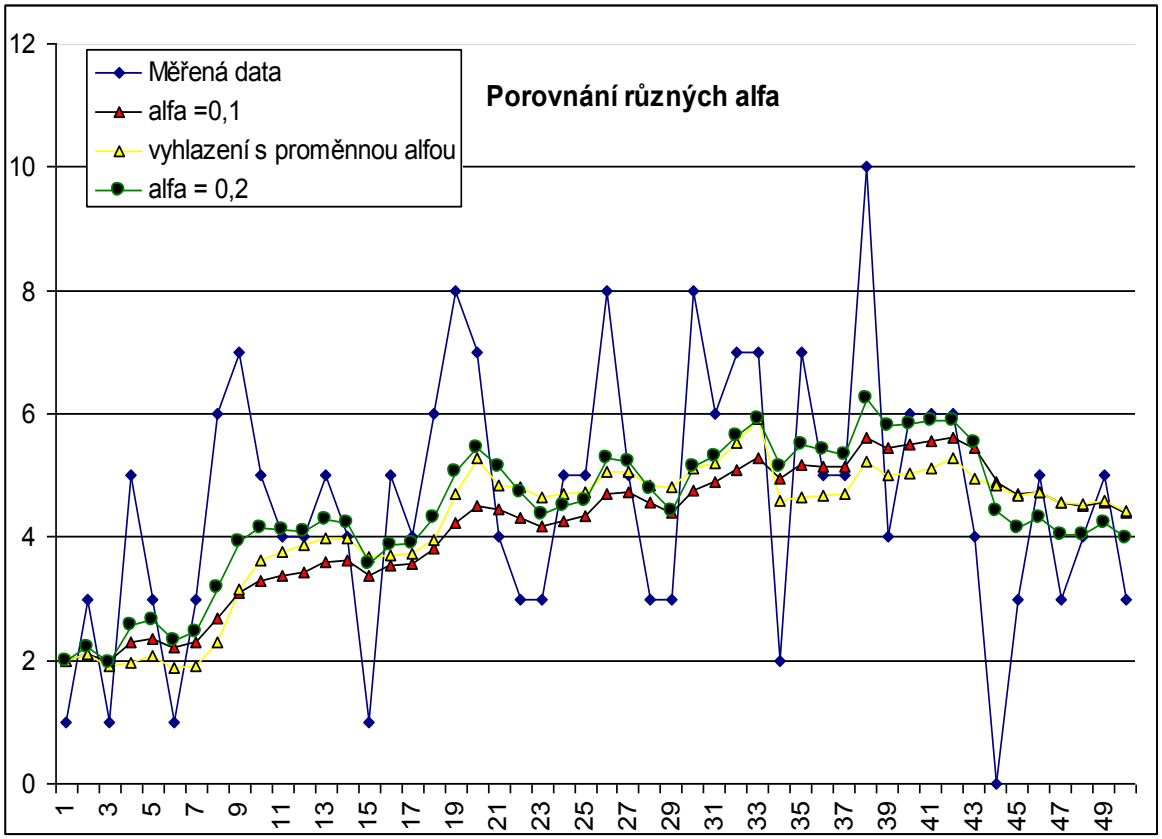
- **Vážený plovoucí průměr** (weighted moving average)
  - Rozlišuje vliv jednotlivých měření



$$\hat{x}_n = \frac{\sum_{i=1}^{k} w_i \cdot x_{(n-i)+1}}{\sum_{i=1}^{k} w_i}$$

Filtrace dat v časové oblasti

- Exponenciální vyhlazování (exponential smoothing)
  - S pevnou hodnotou alfa
  - S proměnnou hodnotou alfa

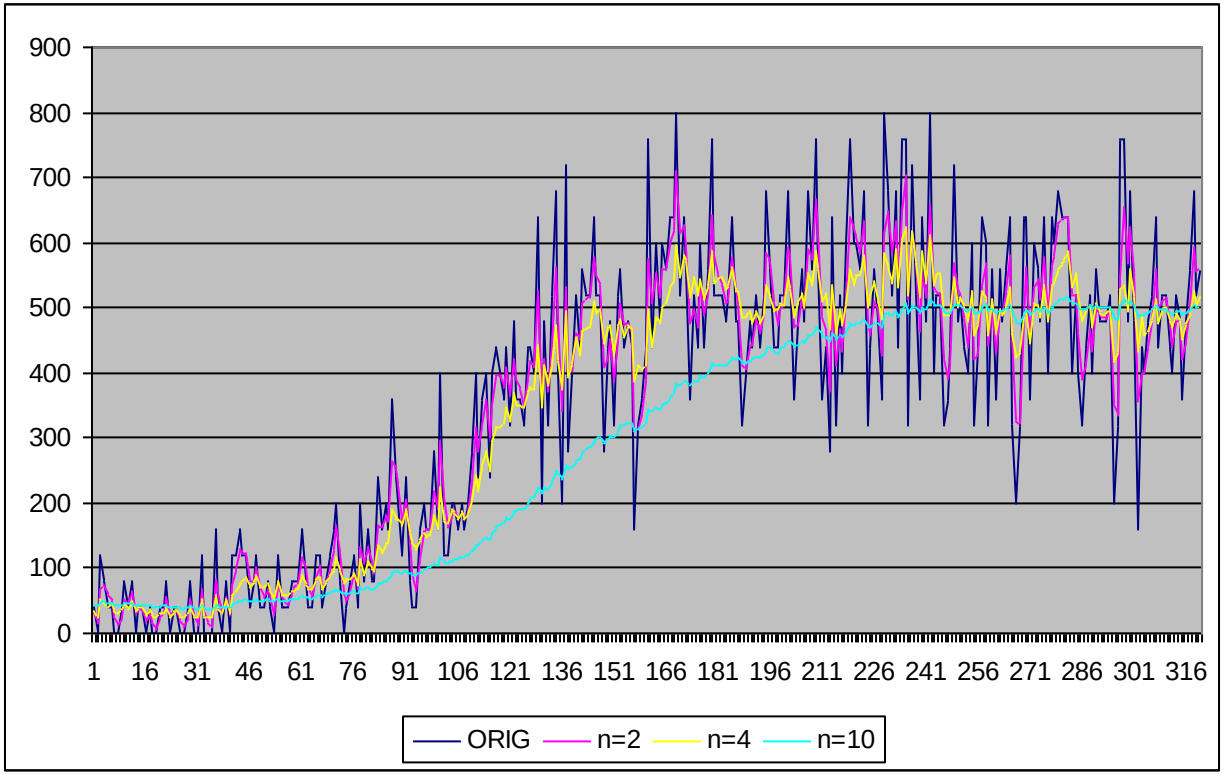


$$\bar{v}_1 = v_1$$

$$\bar{v}_t = \bar{v}_{t-1} + \alpha(v_t - \bar{v}_{t-1})$$

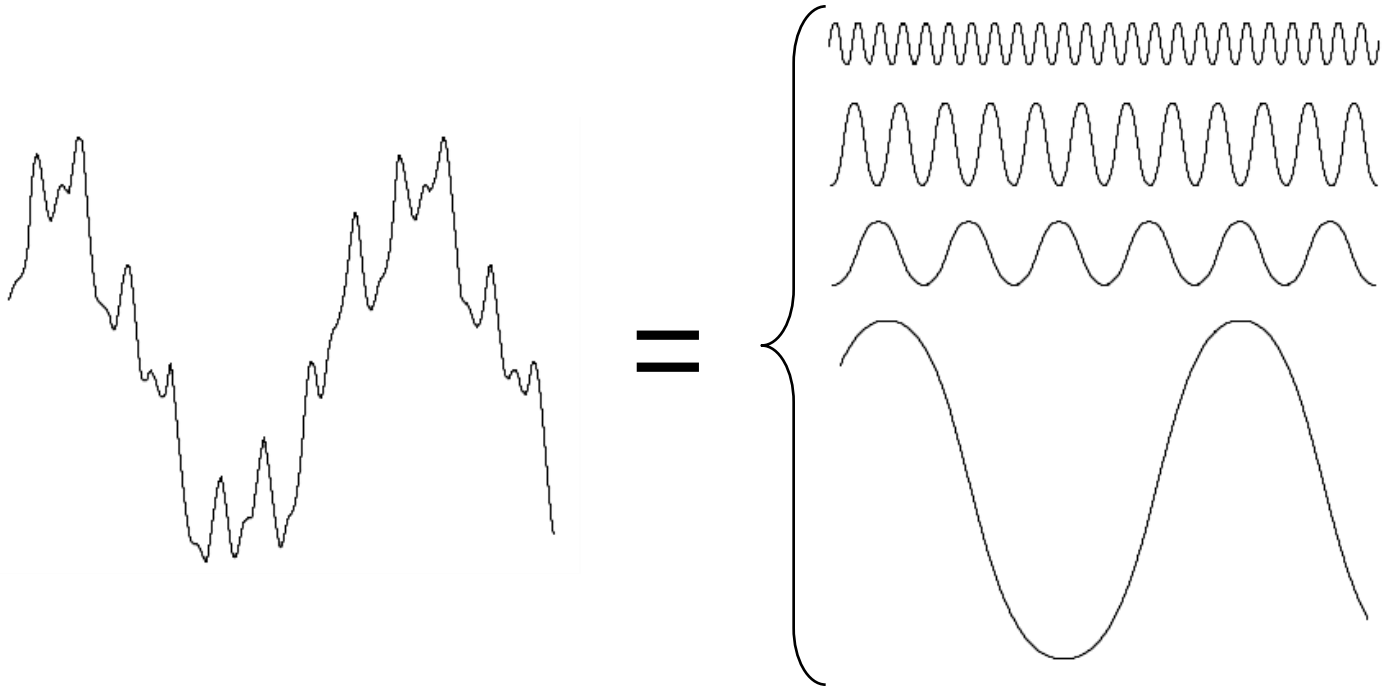
Diskuze

- Jak volit optimální velikost okna pro plovoucí průměr? Je lepší n=5, n=10 nebo n = 20?



Filtrování dat ve frekvenční oblasti

- Fourierova transformace.
  - Libovolnou periodickou časovou řadu lze nahradit superpozicí sinusových a cosinusových funkcí



Základní kroky pro filtrování ve frekvenční oblasti

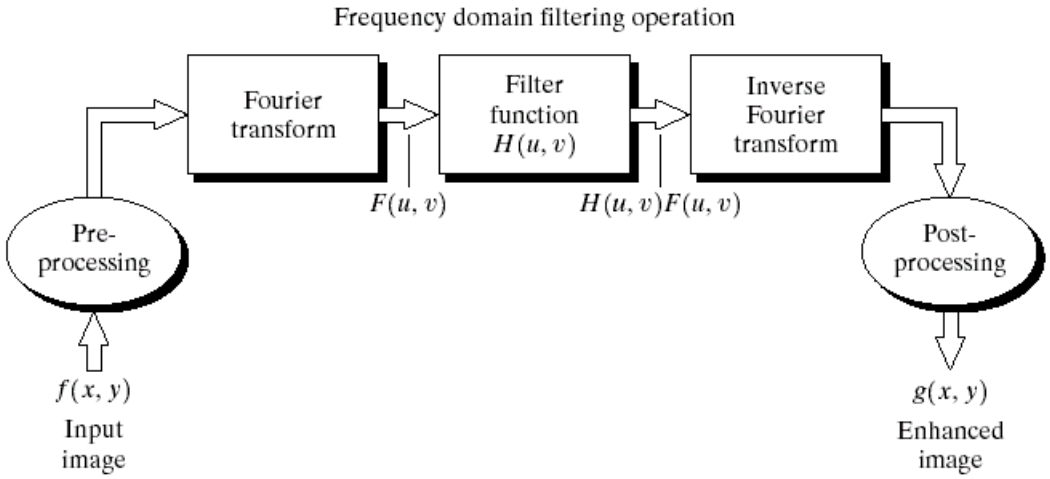
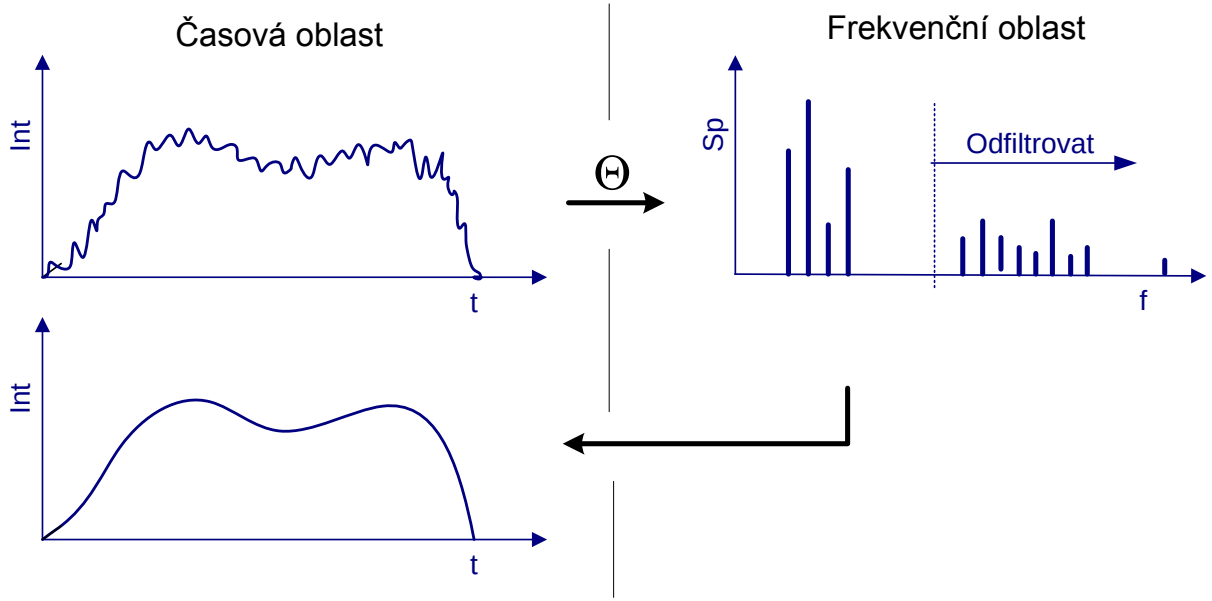


FIGURE 4.5 Basic steps for filtering in the frequency domain.

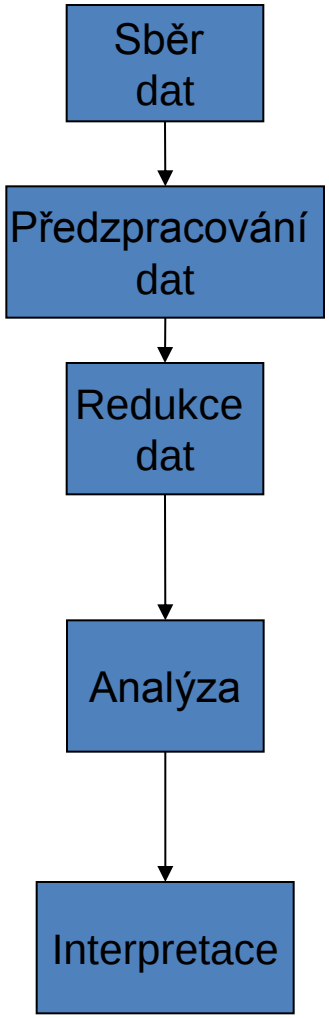


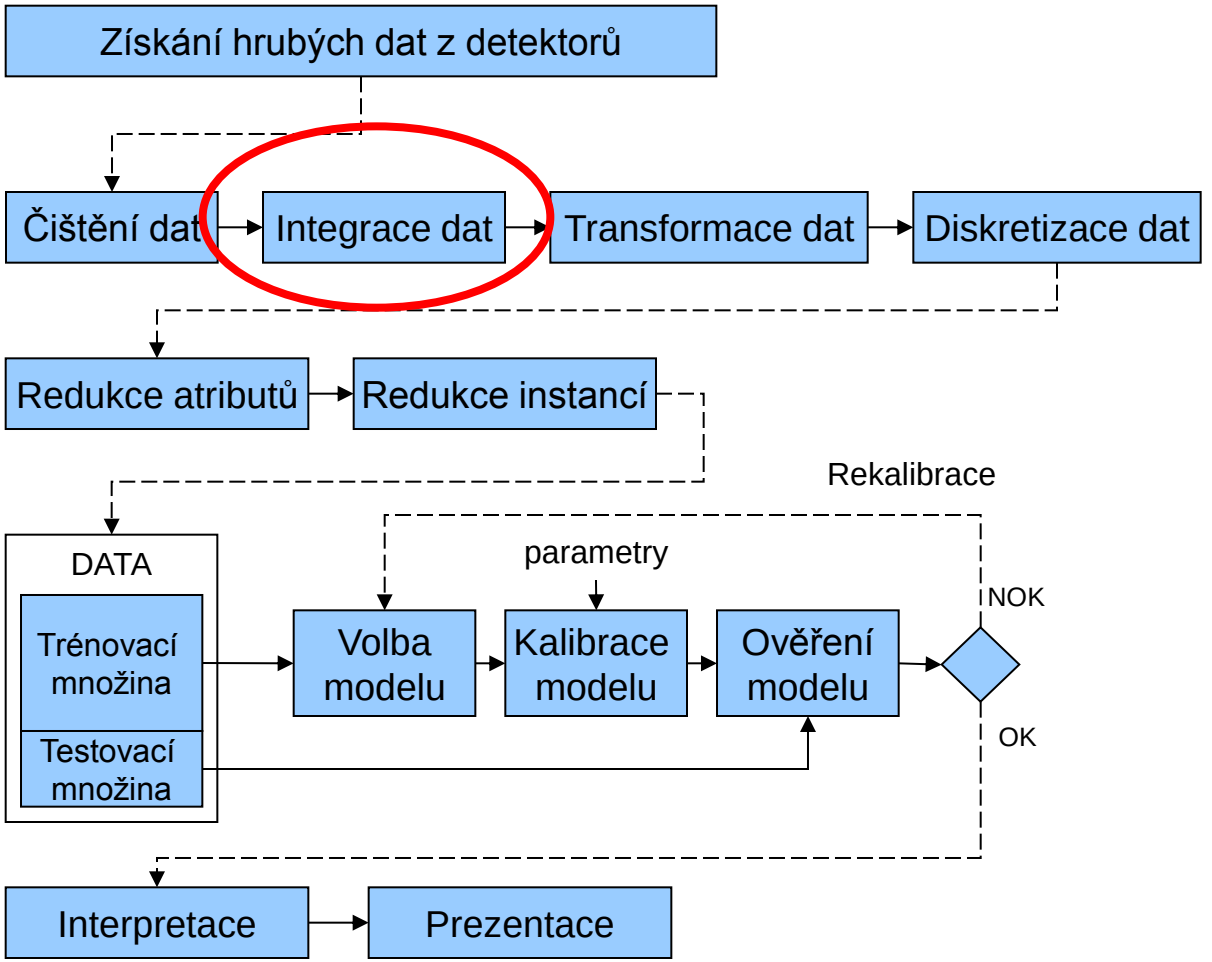
Filtrace

- Filtrace vždy **znamená ztrátu nějaké informace**
  - Cílem je odfiltrování bílého šumu (tedy informace s nulovou střední hodnotou)
- Přílišná filtrace může poškodit data
- Metoda a nastavení filtrace je třeba volit
  - dle cíle a
  - dle následné analytické metody

- V důsledku filtrace ke **zpoždění informace** (pokud potřebujeme rychle reagovat na změnu trendu dat (např. identifikace nehod))
- Bayesovské metody se o filtraci postarají, při lineární regresi je určité filtrování spíše výhodou

Hlavní kroky





Integrace dat

- Integrace dat:
  - Kombinování dat z více zdrojů
- Důvod pro integraci dat?
  - Je třeba **odstranit duplicity a konflikty** v datech, aby nedošlo k nekonzistentním stavům
- Důvody nekonzistence?
    - I / int / Int1
    - Věk / Datum narození
    - tabA_Int / tabB_Int
    - Cena (EUR) / Cena (Kc)
    - Rychlost (kmh) / Rychlost (mph)

Diskuse

- Jak odhalím redundantní data?

Jak zjistit redundantní data při integraci?

- Redundantní data (časové řady) mohou být zjištěna pomocí korelační analýzy
  - popisuje lineární vztahy mezi veličinami (míra závislosti)
- Párový (Pearsonův) **korelační koeficient**
  - míra vyjádření "těsnosti **lineární** vazby". (od –1 do +1)

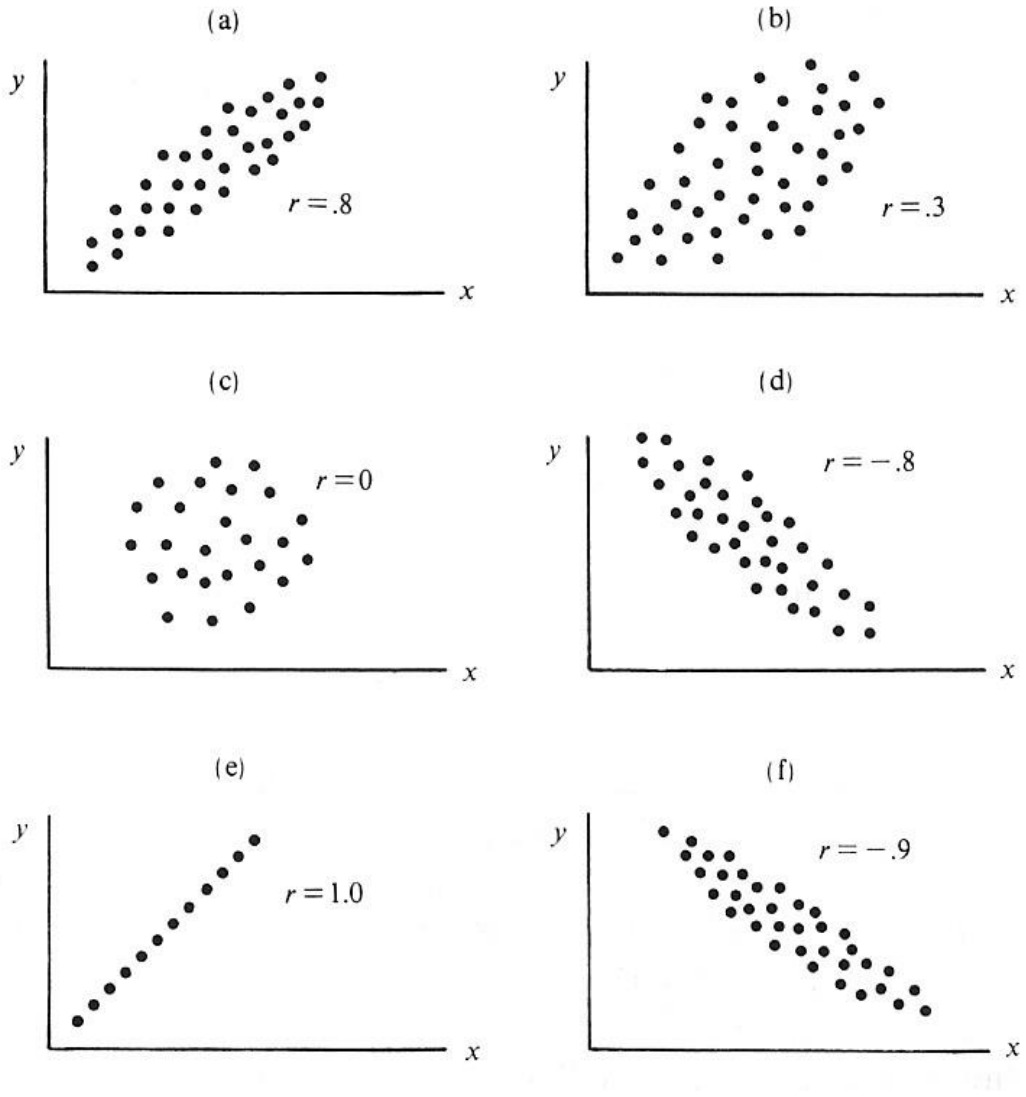
$$R = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1)\, s_x\, s_y} \qquad R = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Vypočteme **aritmetické průměry** souborů X a Y, vynásobíme **sumy čtverců odchylek** od těchto průměrů obou souborů. Tím jsme spočetli tzv. **kovarianci**, což je však absolutní veličina, pro výpočet relativní veličiny pak kovarianci dělíme odmocninou násobku rozptylu souboru X a souboru Y.

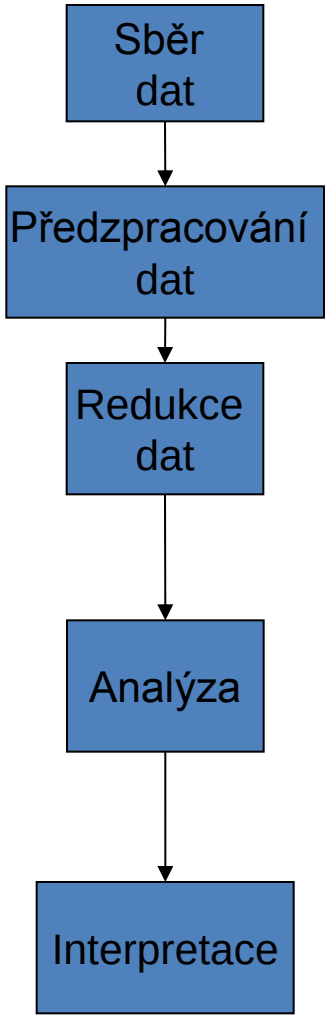
Diskuse

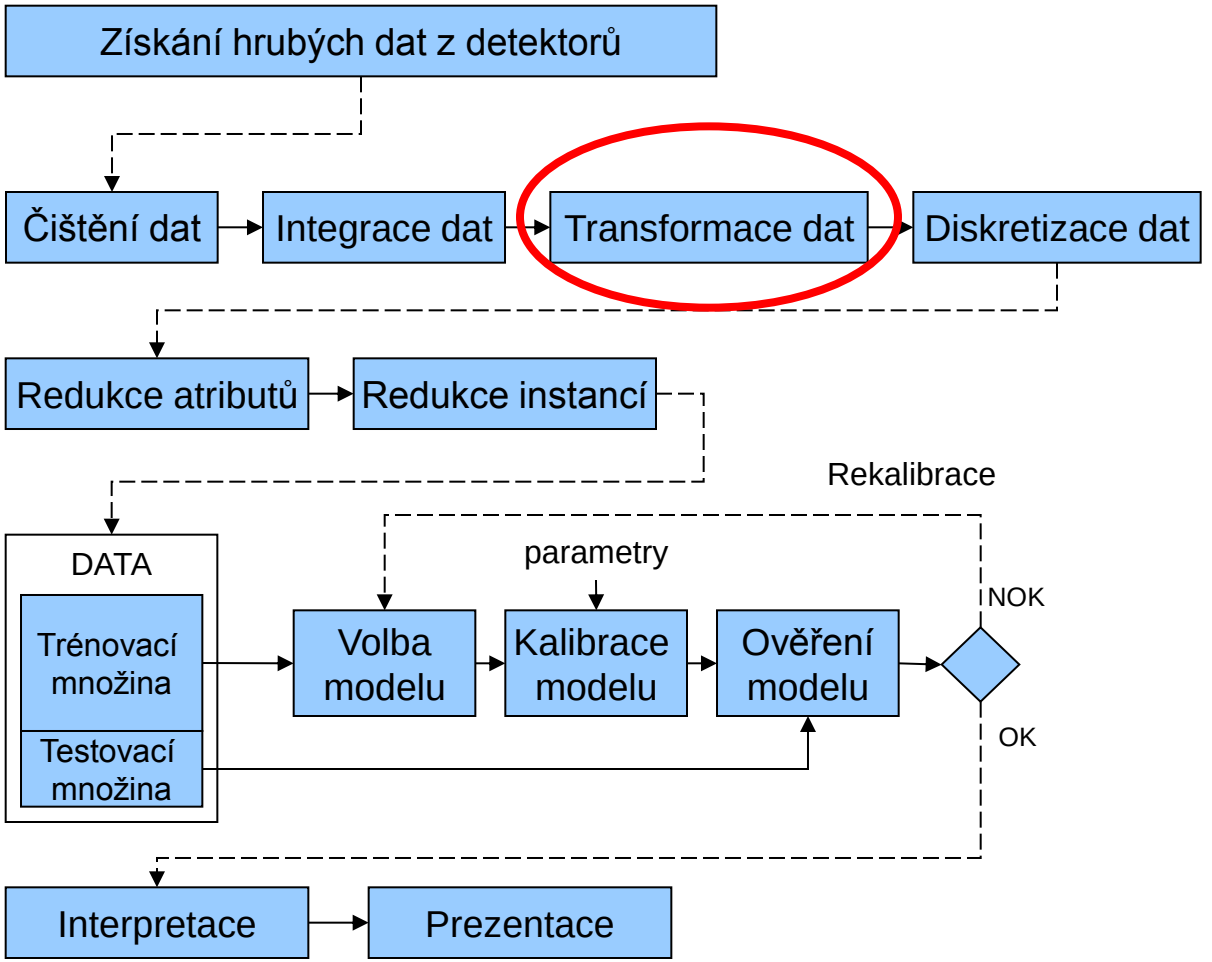
- Co znamenají hodnoty korelačního koeficientu rovné
  - 0?
  - 1?
  - -1?
  - 0,5?

Ukázka korelačních koeficientů



Hlavní kroky

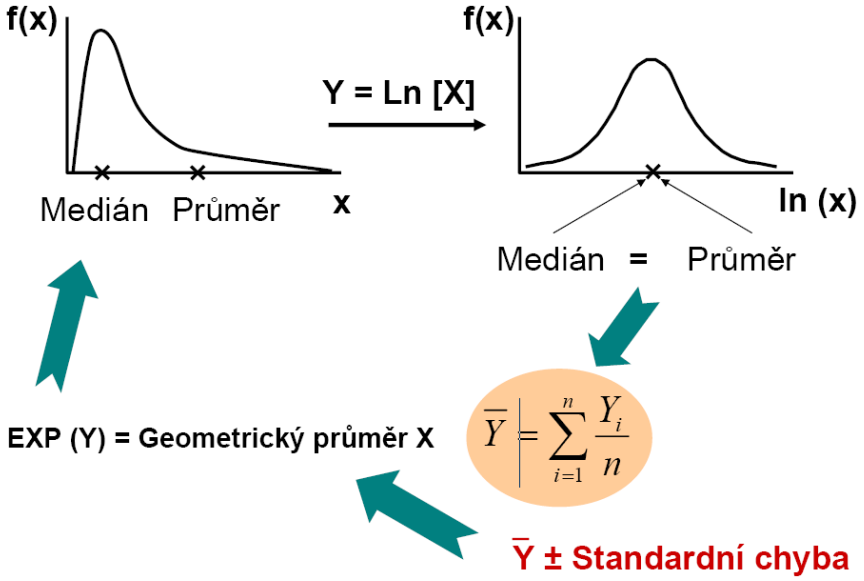




Transformace dat

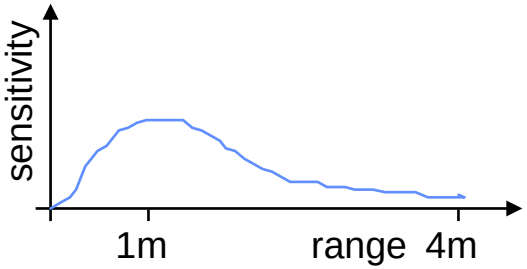
Odstranění závislosti atributů na jednotkách měření

- Transformace
  - **Logaritmická**, odmocninová, …
- Agregace
  - souhrny, vytváření globálních pravidel
- Generalizace
  - koncept hierarchického rozvrstvení
- **Normalizace**
  - Změna měřítka tak, aby data náležela do určitého intervalu
  - Vhodné pro porovnávání různých dat

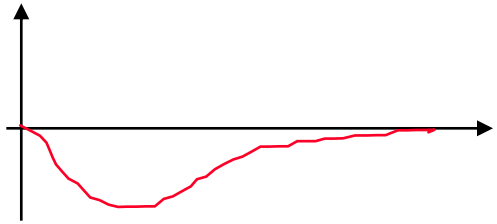


Příklad transformace -
nelinearita

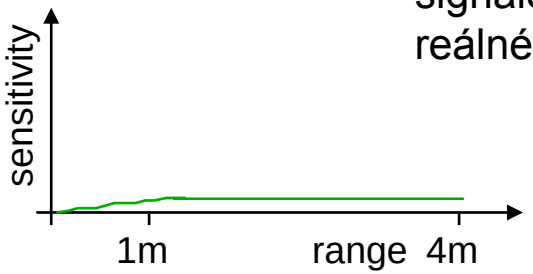
- **Korekce mlhy**



Skenery do venkovního prostředí- Outdoor- vykazují při opticky nečistém prostředí zvýšenou citlivost okolo 1,3m



Podle křivky je citlivost upravována opozitním signálem, který je přičítán k reálnému signálu

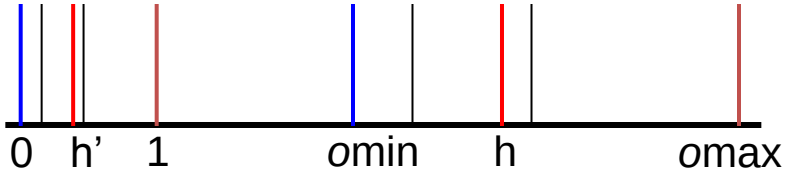


Výsledná křivka citlivosti je téměř plochá. To zaručuje konstantní citlivost pro celou zaručenou vzdálenost detekce.

Transformace dat - normalizace

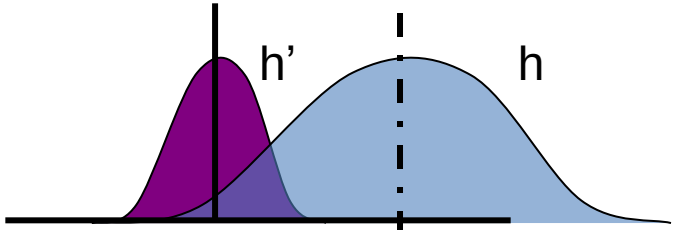
- Lineární normalizace

$$h' = \frac{(h - o\min)(new\max - new\min)}{o\max - o\min} + new\min$$

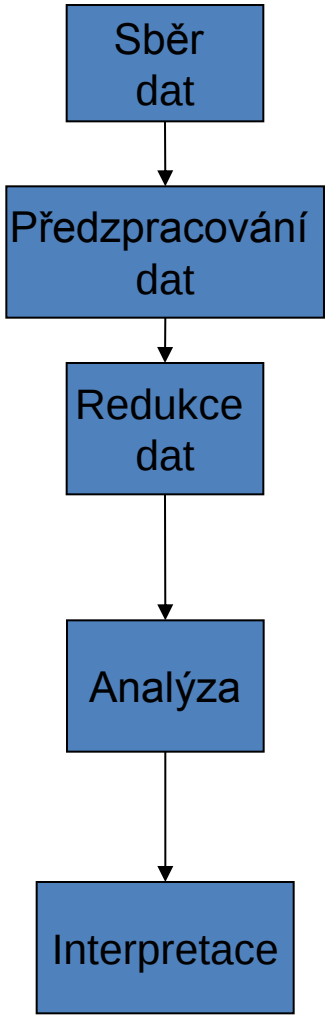


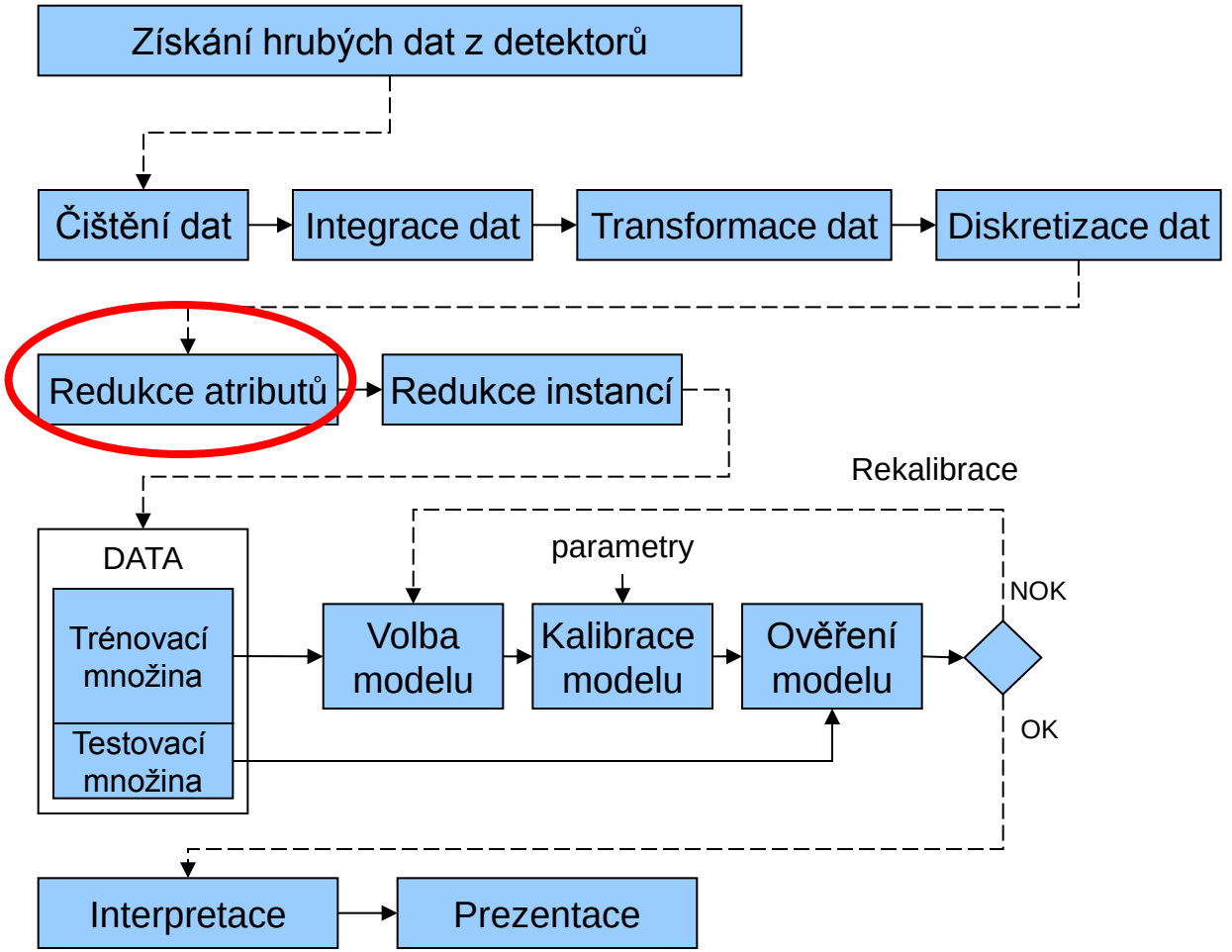
- Z-score normalizace

$$h' = \frac{h - mean_A}{std_A}$$



Hlavní kroky



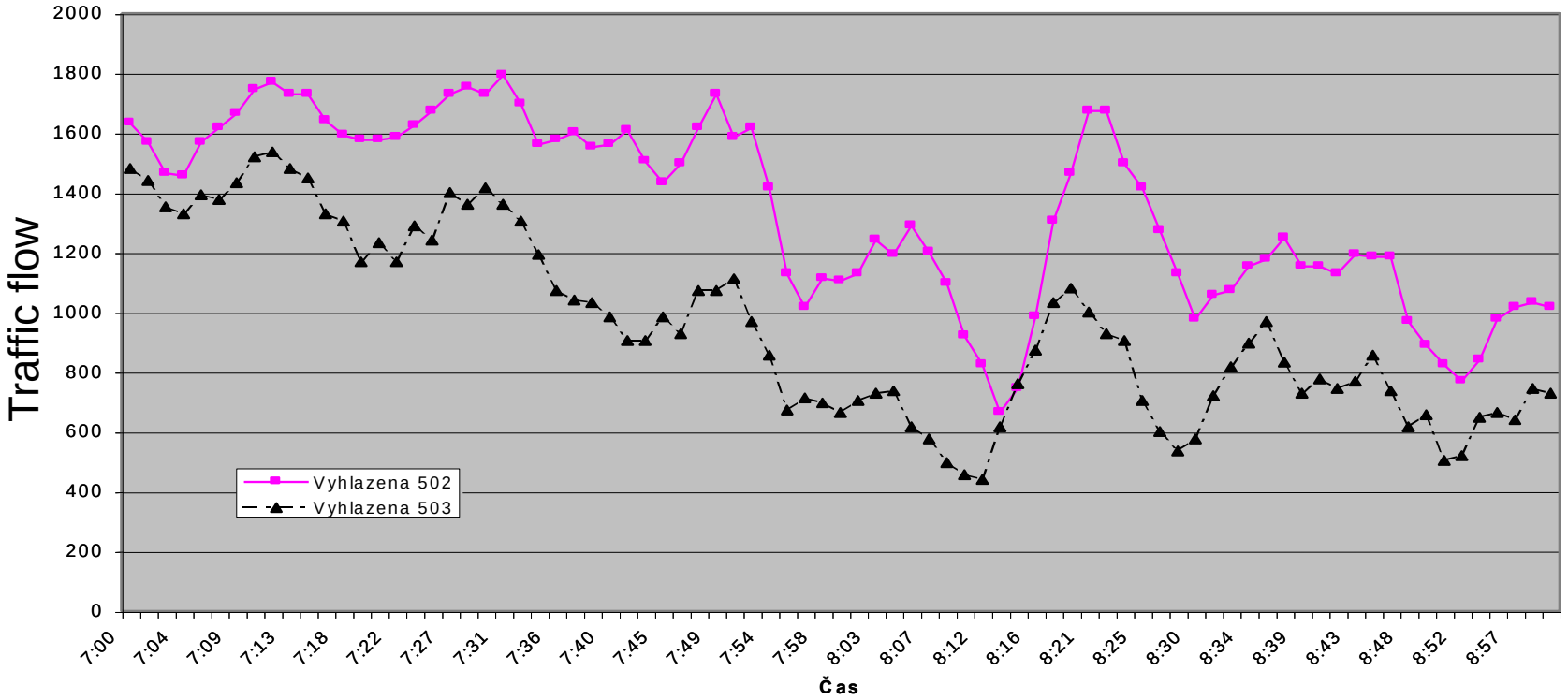


Proč je třeba redukce dat?

- V praxi se sbírá **obrovské množství dat**
  - Na většině řízených křižovatek se nachází v každém jízdním pruhu alespoň jedna indukční smyčka
  - Z každé indukční smyčky se v pravidelných intervalech 90 sekund sbírají informace o intenzitě dopravy a obsazenosti detektoru.
- Toto obrovské množství dat se přenáší na hlavní dopravní řídící ústřednu v Praze
- **Co s tím?**
  - Klasické algoritmy jsou zahlceny
  - Není možné najít „důležitá data“ – informaci o dopravní nehodě, a pod.

Redukce dat - Motivace

- Sbíraná data často nesou podobnou, či zcela stejnou informaci
  - Viz. detektory umístěné na jedné linii za sebou



Redukce dat

- Strategie redukce dat
  - Redukce dimenzionality
  - Komprese dat
  - Redukce vzorků
  - Diskretizace

Redukce dat - cíl

Získat redukovanou množinu dat, která je mnohem menší objemově, ale produkuje (téměř) stejné analytické výsledky.

- **Výběr podmnožiny atributů**
  - Vyber minimální podmnožinu všech atributů, které dostatečně reprezentují původní rozdělení dat

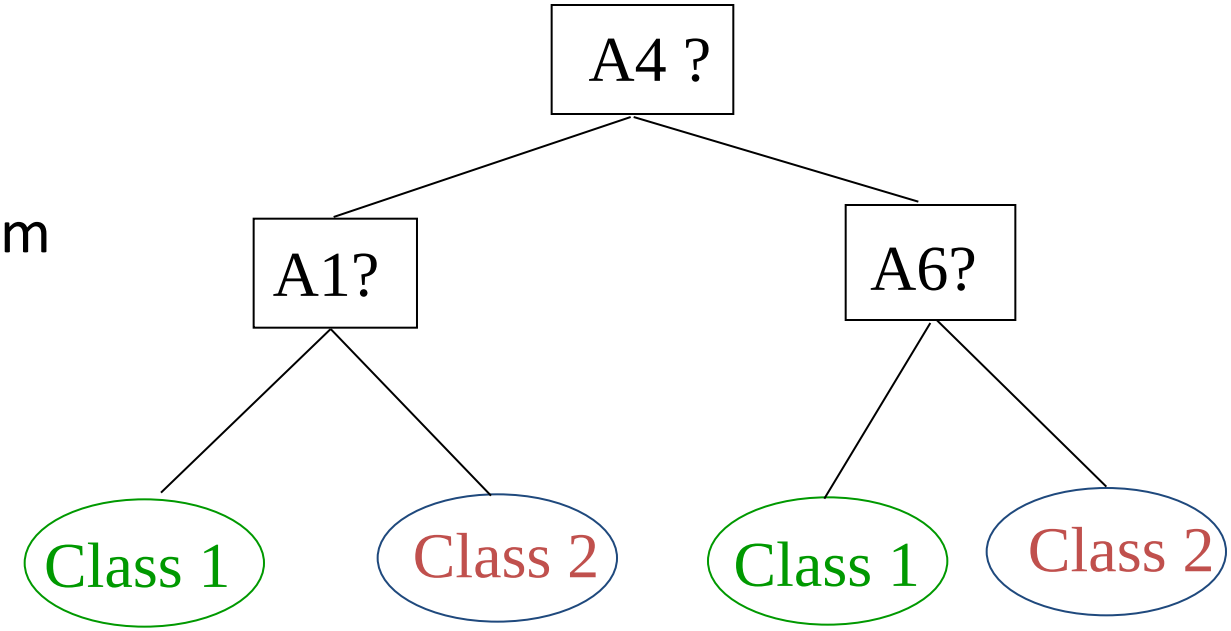
Redukce dimenzionality - Jak vybrat atributy?

Heuristické metody (exponenciální počet možností) jsou **greedy**

- **Vynecháním**
  - Konstantních atributů
  - Řídce obsazených atributů
  - Atributů s duplicitní informací (věk x datum narození)
    - Korelační analýza, ANOVA
- **Sloučením** atributů
- **Analyticky**
  - Rozhodovací stromy
  - Fourierova transformace, Wavelet transformace
  - Analýza hlavních komponent (PCA, *Principal component analysis*)
  - Shlukování

Příklad rozhodovacího stromu

- Původní množina atributů **{A1, A2, A3, A4, A5, A6}**
- Výsledný strom

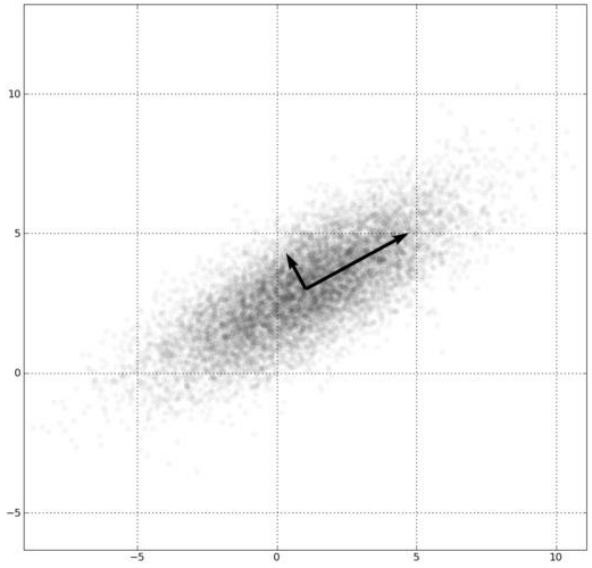


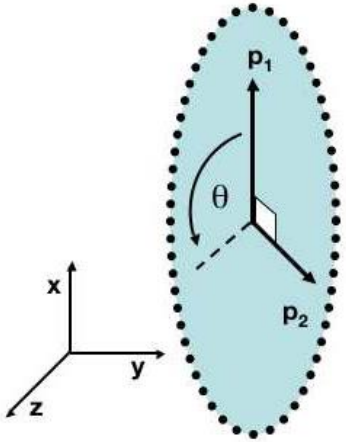
- Redukovaná množina atributů: **{A1, A4, A6}**

Metoda PCA (Principal Component Analysis)

- Analýza hlavních komponent
  - Snížení dimenze dat s co nejmenší ztrátou informace

- Nové atributy/dimenze:
  - Lineární kombinace původních
  - Nekorelované
  - Ortogonální
  - Zachycují co nejvíce původní variance v datech





Komprese dat

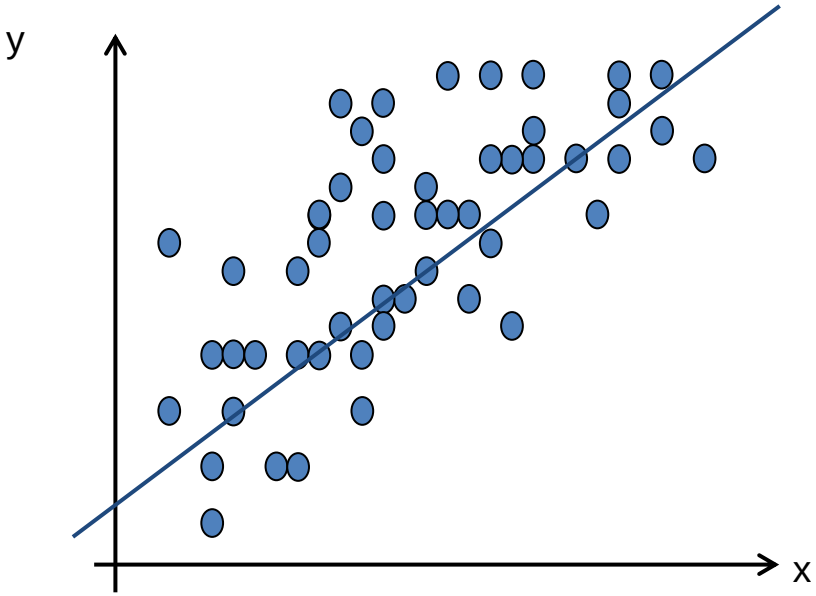
- Komprese řetězců
  - Mnohé existující algoritmy (LZW, LZMA, MPEG2, MPEG4, JPEG, …)
  - Obvykle bezeztrátové
- Využívá se pro kompresi audio/video dat
- Příklad - Kódování RLE (Run-Length Encoding)
  - posloupnost opakujících se bytů se nahradí jednou hodnotou s uvedením počtu opakování

"AAAhoooooj"

"<3A>h<4o>j"

Redukce počtu vzorků – parametrické metody

- Parametrické metody
  - Pokud původní data odpovídají určitému modelu, je možné je nahradit tímto modelem



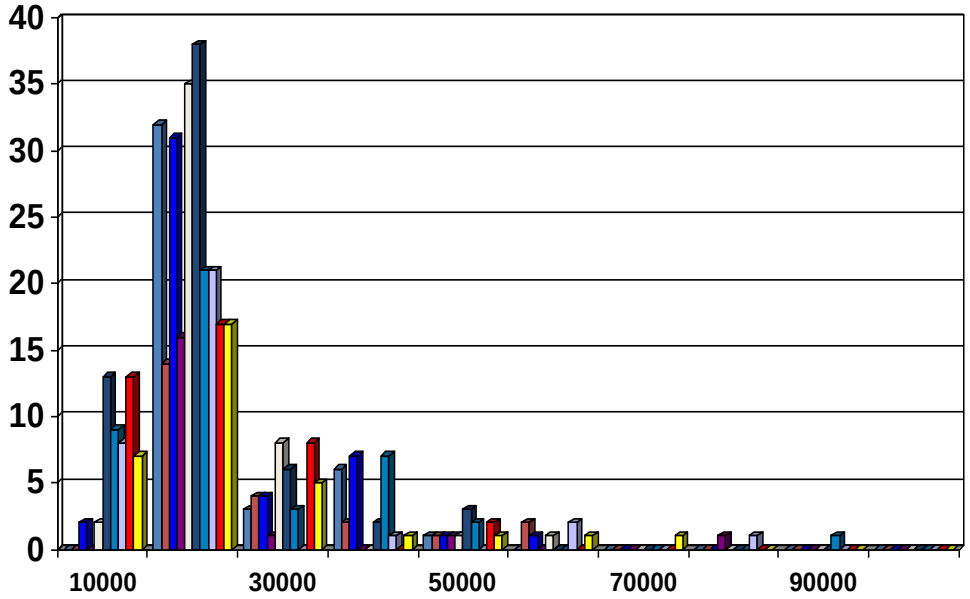
$$y = \alpha + \beta \cdot x$$

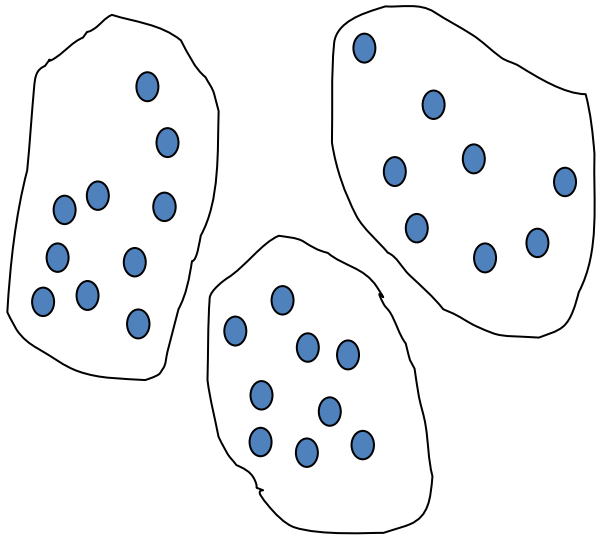
$$y = 5 + 10 \cdot x$$

$$\{5;10\}$$

Redukce počtu vzorků – neparametrické metody

- Neparametrické metody
  - Data není možné nahradit modelem
  - Hlavní skupiny: Histogram, shlukování, a další



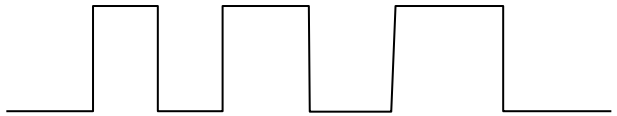


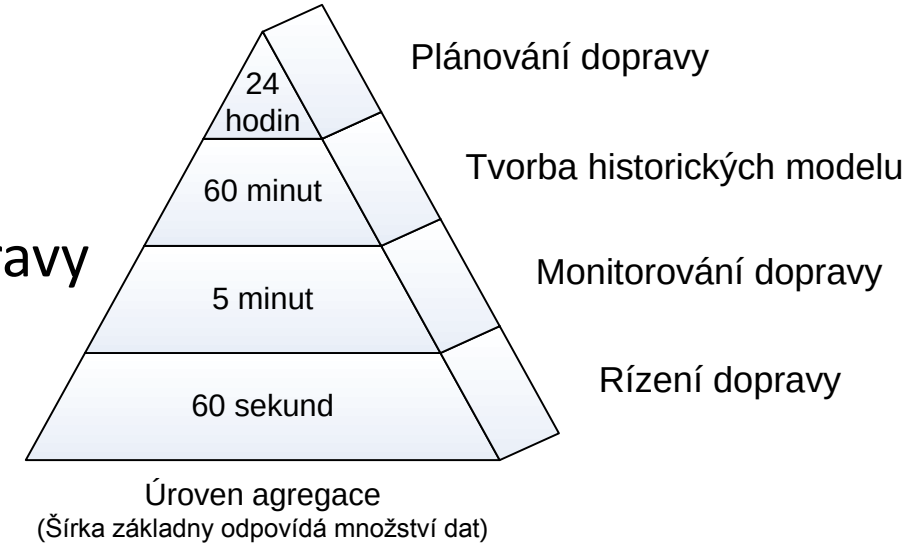
Hierarchická redukce - agregace

- Různé dopravní problémy mají různé požadavky na data

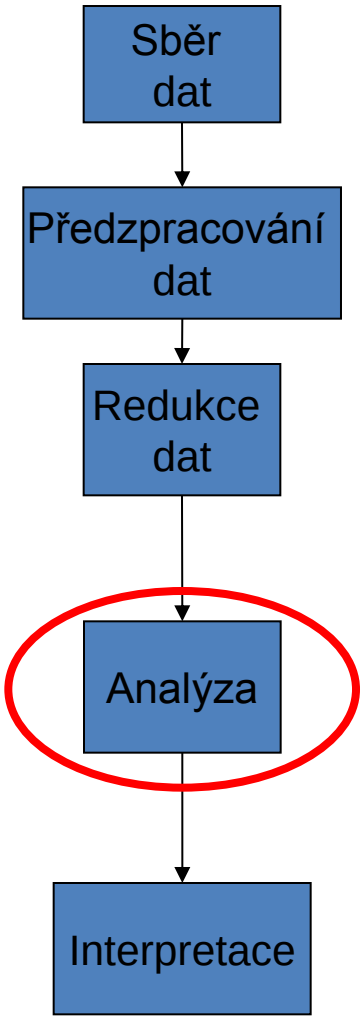
Agregace

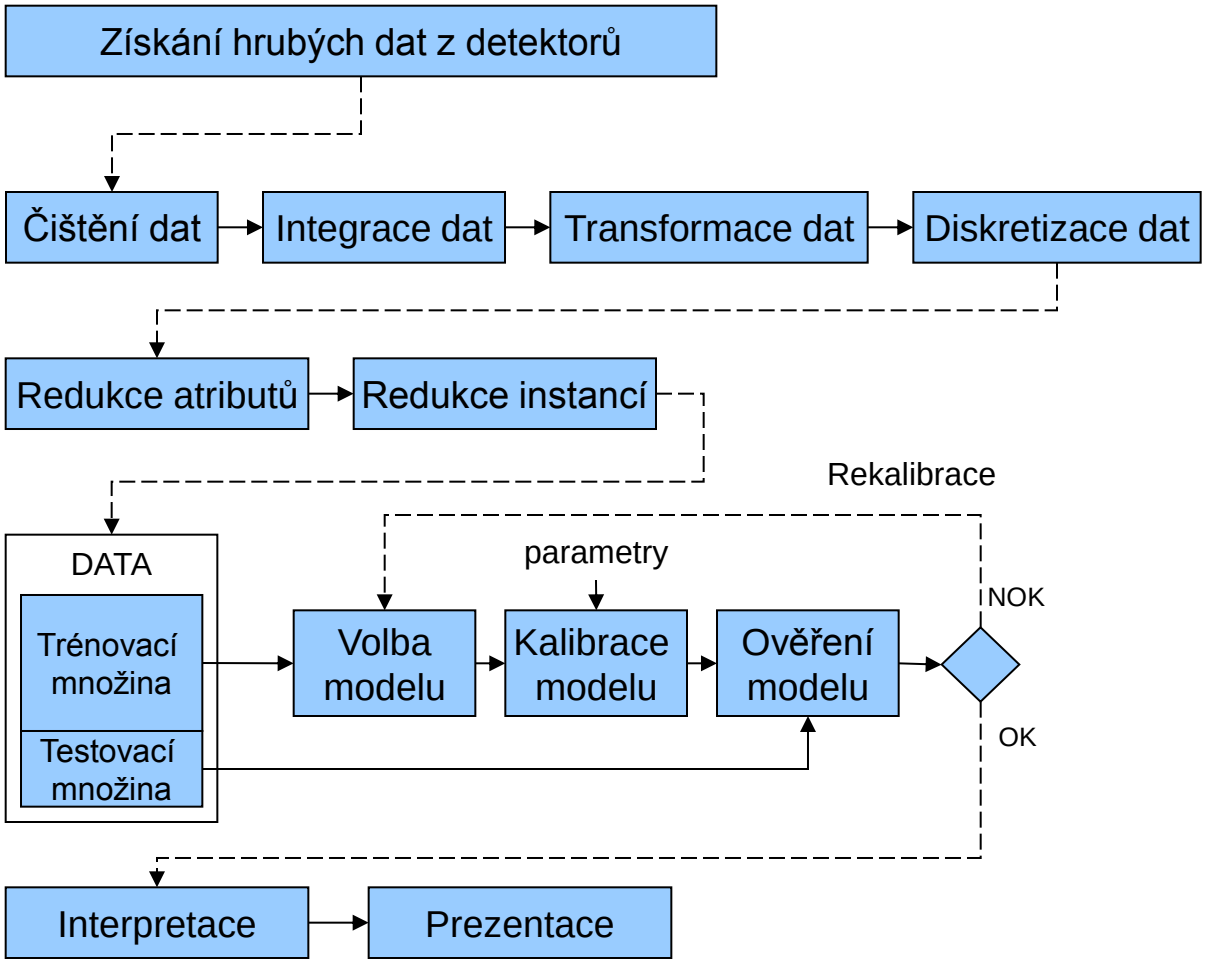
- Příklad – Měření intenzity dopravy
  - Původně měřená data
    individuální vozidla
  - Potřeby řízení uzlu
    [počet vozidel/60 sec]
  - Potřeby monitorování dopravy
    [počet vozidel/5min]





Hlavní kroky





Základní analytické metody

Matematické metody pro ITS (11MAMY)

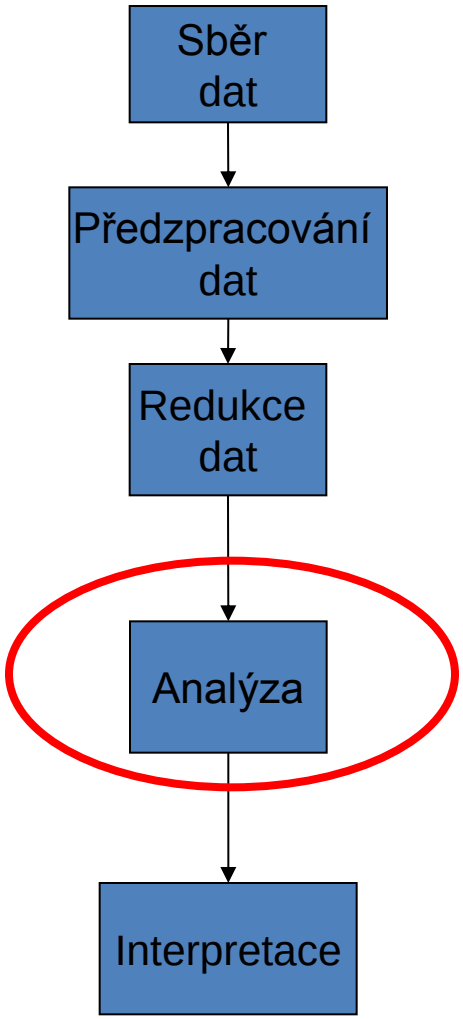
Ondřej Přibyl (Jan Přikryl)

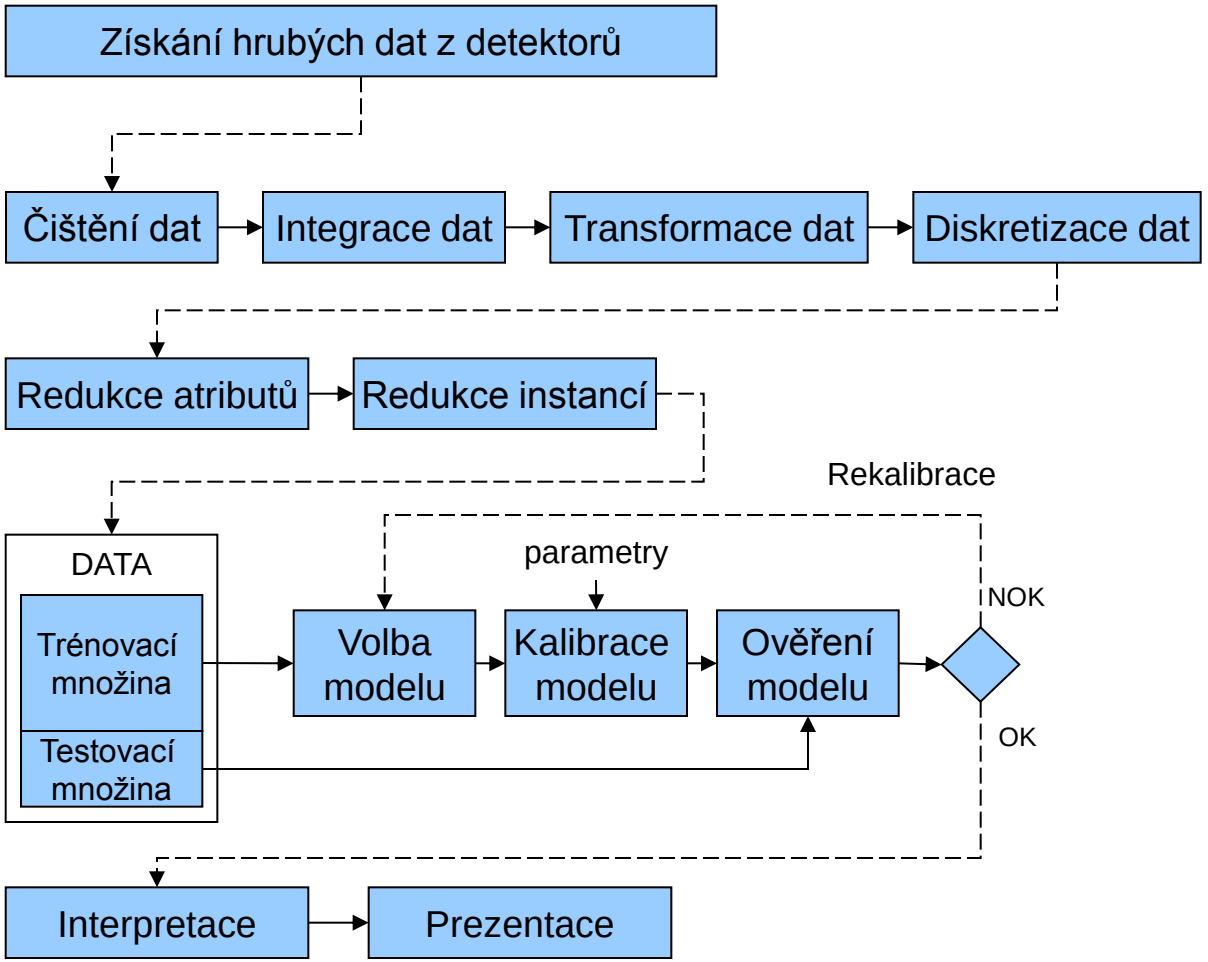
Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní

Obsah prezentace

- Úvod do analýzy dat
- Shlukování
- Rozhodovací stromy
- Lineární regrese

Hlavní kroky





Dělení analytických nástrojů

dle toho jak nastavují či modifikují své parametry (fáze učení či kalibrace):

- **Učení s učitelem** (i v české literatuře se často používá anglický termín „supervised learning“) kdy **existuje informace** o správné kategorii či o chybě predikce pro každý vzor

- **Učení bez učitele** (unsupervised learning, reinforcement learning) kdy **neznáme** správnou kategorii ani informaci o výsledné chybě predikce a systém sám vytváří přirozené shluky na základě podobnosti mezi objekty

Jak složitý model použít?

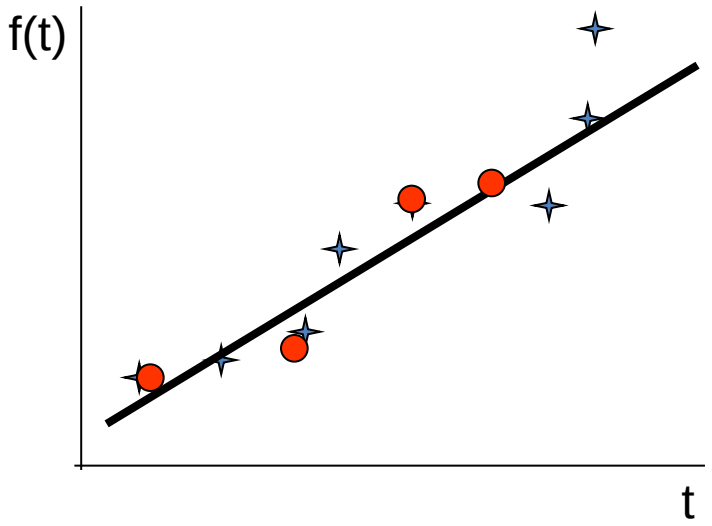
- **příliš jednoduchý** model
  - není schopen popsat složitější procesy v datech.

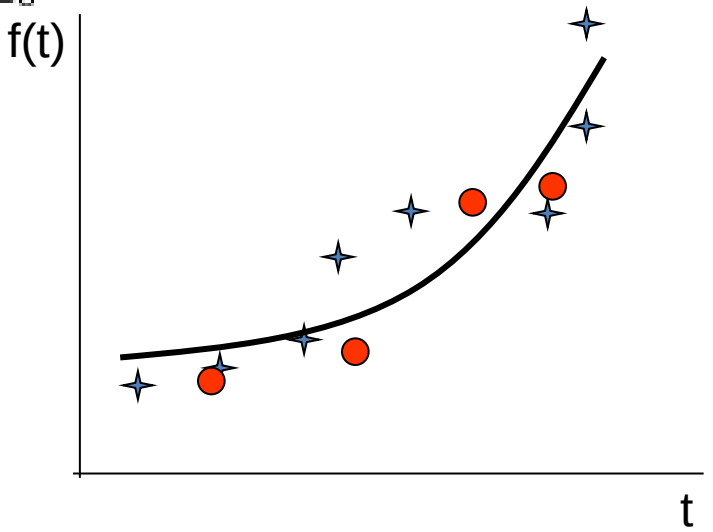
- **příliš komplexní** model
  - dojde k jeho nakalibrování nejen s ohledem na obecné trendy v trénovací množině, ale i všech **náhodných procesů** a složek zachycených v těchto datech.
  - Toto je ekvivalentní procesu učení adaptivních systémů, například neuronových sítí, proto se používá výraz přetrénování.

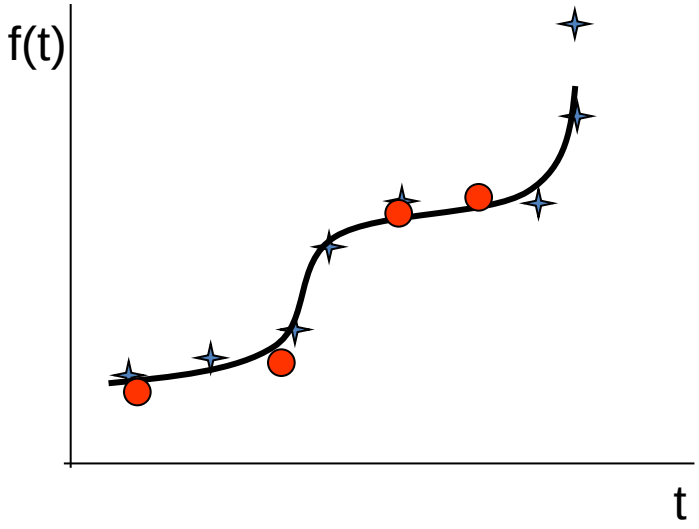
Přetrénování (Overfitting)

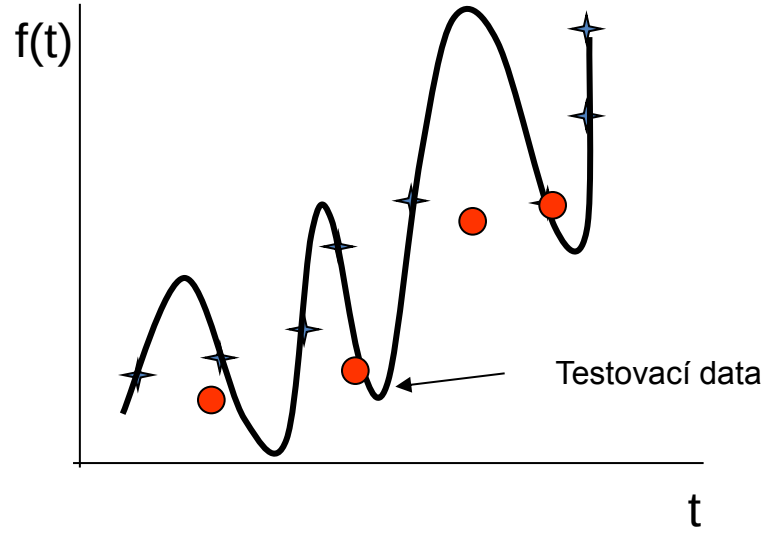
- Polynom n-tého řádu
- Neuronové sítě

$$p(x) = \sum_{i=0}^{n} a_i x^i = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

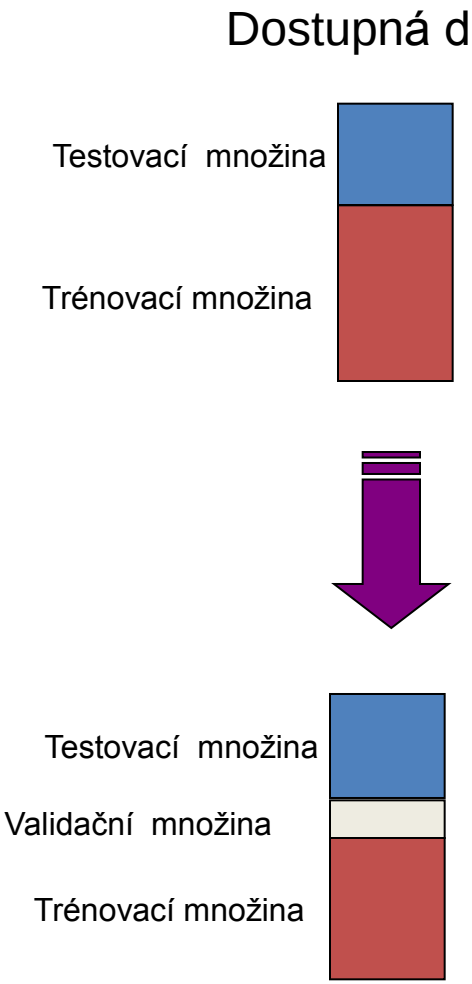


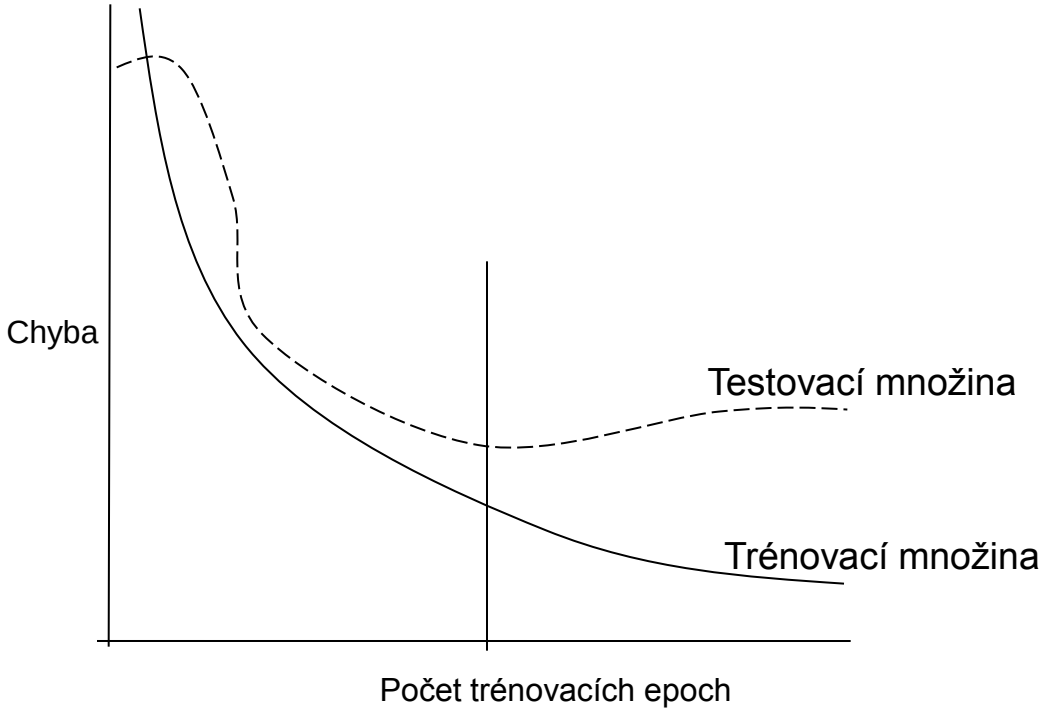






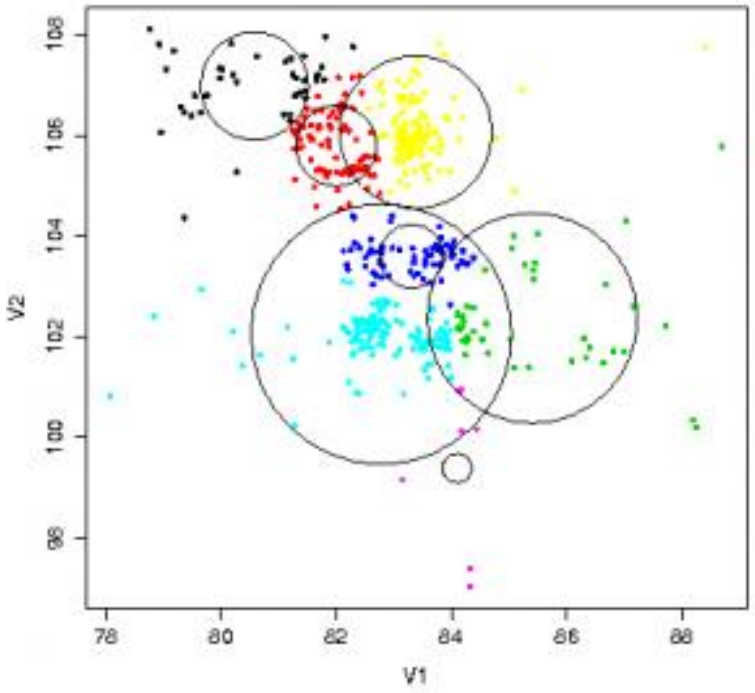
Jak odstranit problém přetrénování?





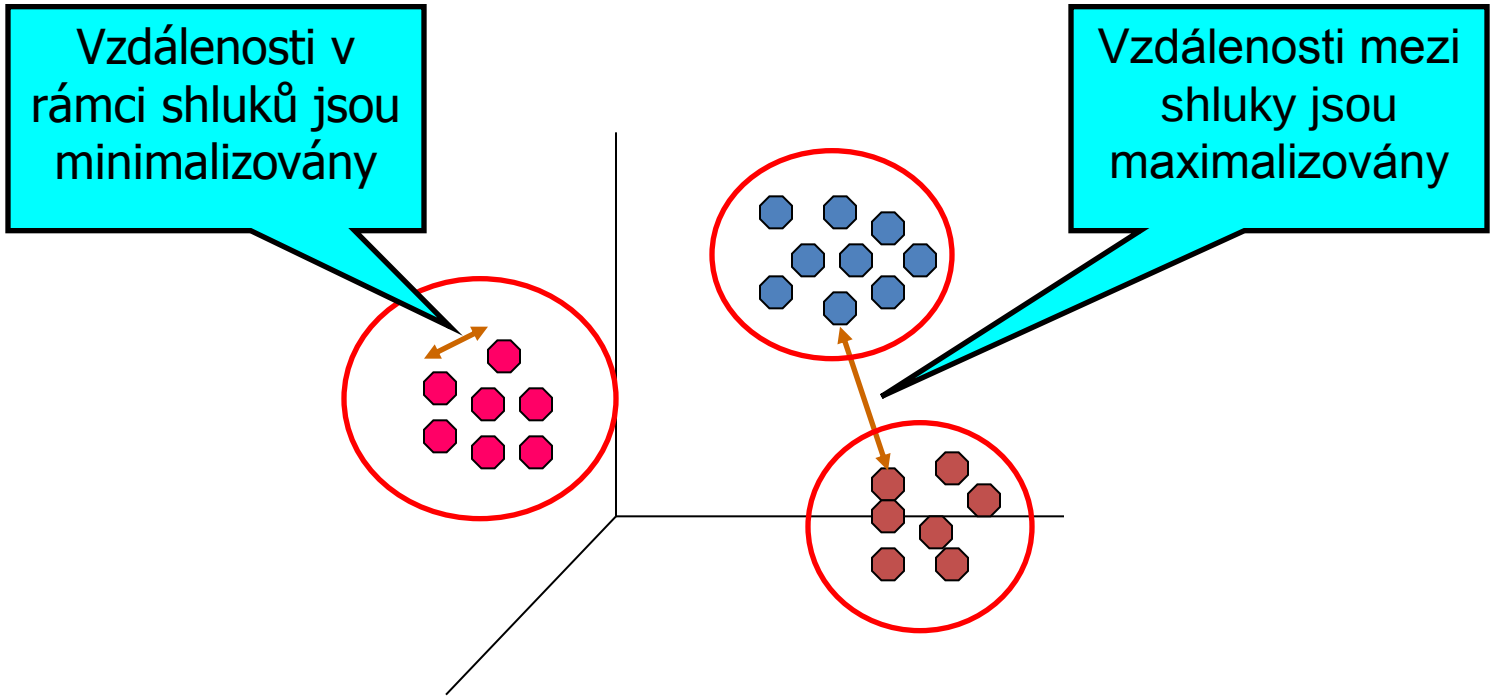
Diskuse

- Co je shlukování?
  - Jedná se o učení s učitelem či bez učitele?



Shlukování - Cíle

- Nalézt přirozené shluky dat.
- Každý objekt je reprezentován vektorem l atributů.
- Shlukování se uskutečňuje na **základě podobnosti** mezi jednotlivými objekty, tak že objekty v každém shluku jsou si navzájem více podobné než objektům v ostatních shlucích.



Diskuze

- Co znamená míra podobnosti?
- Jak by jste vyjádřili míru podobnosti pro:
  - Numerická data

    A (3,10) ;
    B (6, 6)

  - Binární data

    A (00101101) ;
    B (11011111)

  - Speciální data
    - např. intenzita dopravy (časová řada)

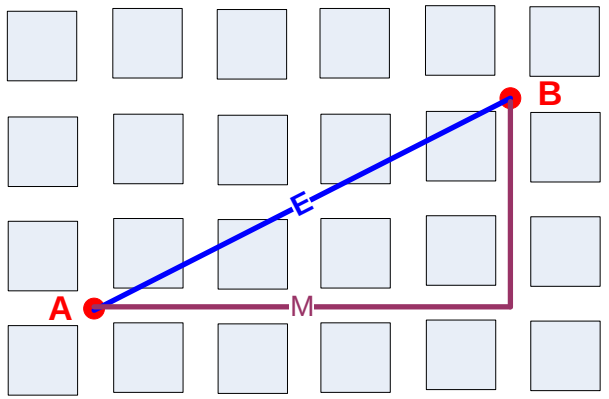
Míra podobnosti objektů

Numerická data

- Minkowského metrika

$$\left(\sum_{k=1}^{p} \left(x_k(i) - x_k(j) \right)^{\lambda} \right)^{\frac{1}{\lambda}}$$

  - E - Euklidovská vzdálenost (Lambda = 2)
  - M - Manhattan vzdálenost - městský blok (Lambda = 1)



Binární data

- Například počtem souhlasných bitů
  - distance(a,b) = {Počet ai ≠ bi } / i

Dělení metod shlukové analýzy

- Hierarchické metody
- Optimizační metody

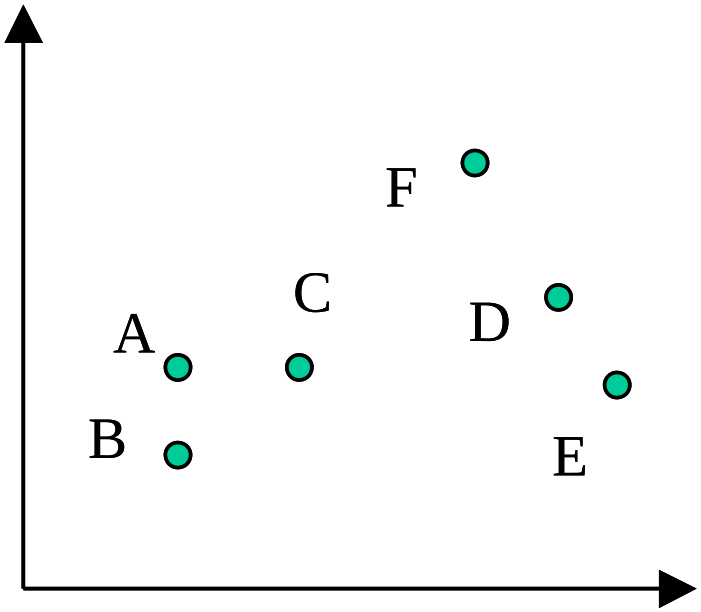
Dělení metod shlukové analýzy

Hierarchické metody

- Pracují **iterativně** pro všechny hodnoty počtu shluků K.
- Existují dva základní principy:
  - **aglomerační** (ze zdola nahoru) - začínají tak, že každý objekt je zároveň samostatným shlukem. V každém kroku pak dojde ke spojení dvou nejpodobnějších shluků a zároveň je vypočítána poloha centra tohoto shluku
  - **rozdělovací** (ze shora dolů) - začínají tak, že všechny objekty patří do jediného shluku. V každém kroku potom dojde k rozdělení tohoto shluku v nejvhodnějším místě.
- Výhody
  - vytvoří shluky pro všechny hodnoty K.

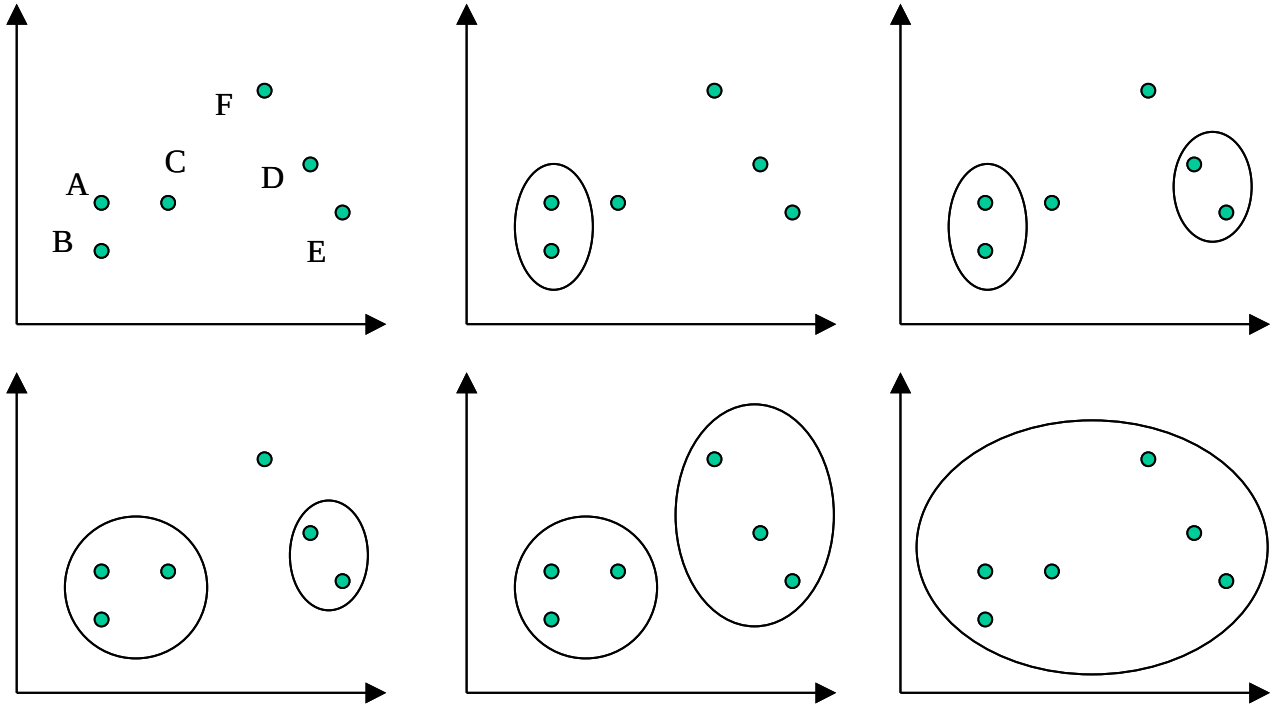
Příklad hierarchického shlukování

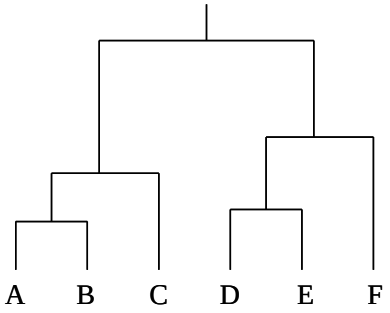
- Zadání
  - Úkolem je najít shluky v příkladě šesti 2D objektů, {A, B, C, D, E, F}



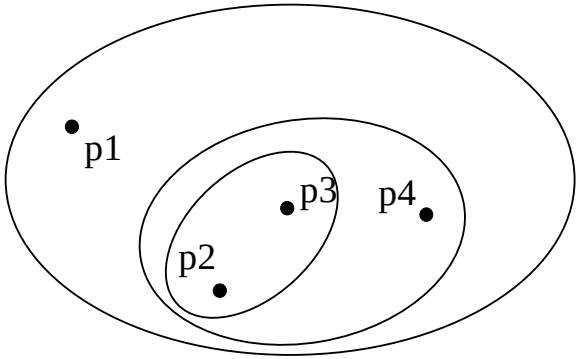
Příklad hierarchického shlukování

- **Dendrogram** - graf na jehož ypsilonové ose jsou vzdálenosti mezi jednotlivými shluky.

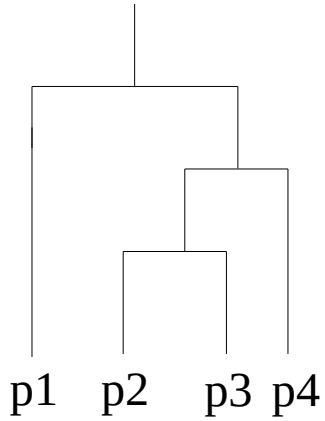




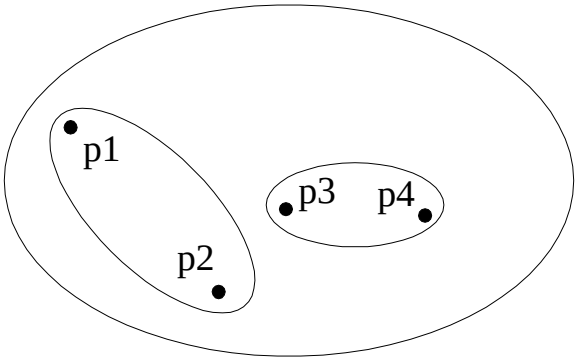
Hierarchické shlukování



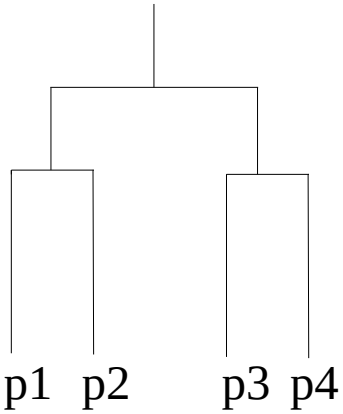
Tradiční shlukování



Tradiční Dendrogram



Alternativní shlukování



Alternativní Dendrogram

Diskuze

- Jaká jsou omezení hierarchických metod?

Problémy hierarchického shlukování

- **neoptimalizují žádnou účelovou** funkci.
- **hledají v každém kroku nejlepší možné řešení**, ovšem to nezaručuje dosažení globálního optima.
- neumožňují následně měnit shluky. Jakmile jednou dojde ke spojení (či rozdělení) dvou shluků, jedná se o spojení konečné.
- jsou také poměrně **citlivé na odchylky**, šum a okrajová data.
- V neposlední řadě jsou tyto algoritmy pro větší objemy dat poměrně **náročné na výpočetní čas**.

Optimalizační metody shlukování

- Na rozdíl od hierarchických metod, **vytváří požadovaný počet shluků** najednou.

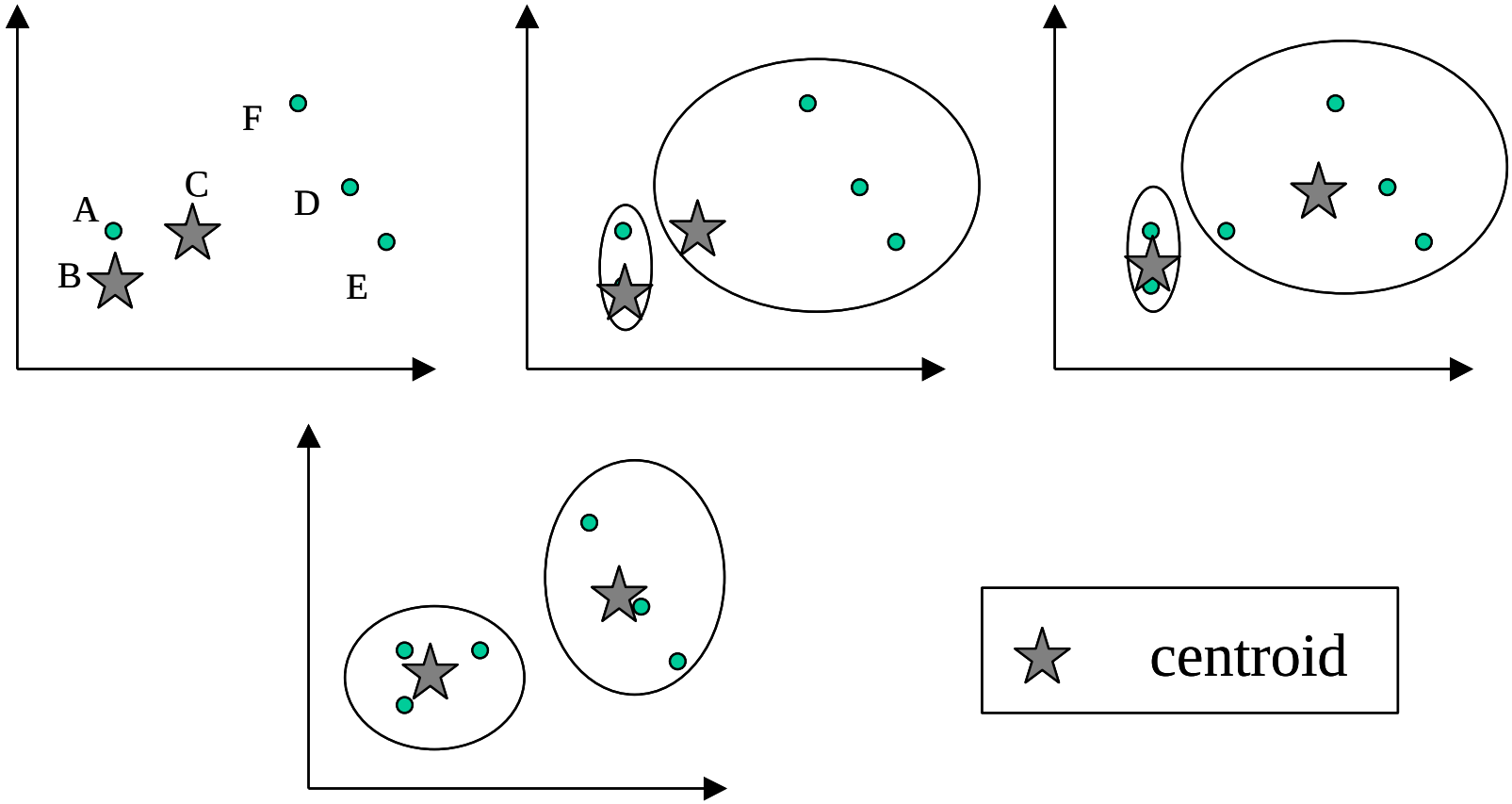
- Z toho je zřejmé, že **počet shluků musí být znám** před analýzou a je jedním ze vstupních parametrů těchto metod.

- Nejznámější je metoda zvaná k-means

Metoda *k*-means

- Iterativní metoda
- Každý shluk reprezentován geometrickým středem - centroid
- Začínáme s náhodně generovanými shluky.
  Shluky následně posouvány tak, aby se zvýšila podobnost objektů v rámci jednotlivých shluků a zvýšila odlišnost objektů náležejícím do různých shluků.
- Celý proces je popsán následujícími kroky:
  1. Vybrat náhodně k objektů (centroidů) které reprezentují jednotlivé shluky
  2. Přiřadit každý objekt v databázi k nejbližšímu centroidu
  3. Spočítat nové centroidy shluků (*geometrický průměr*)
  4. Krok 2 a 3 se opakuje dokud dochází ke změně ve shlucích

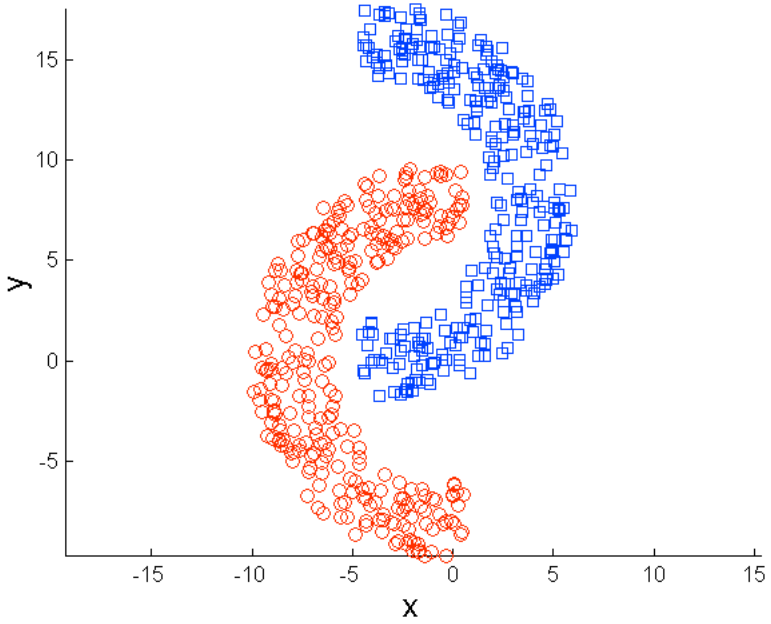
Příklad k-means

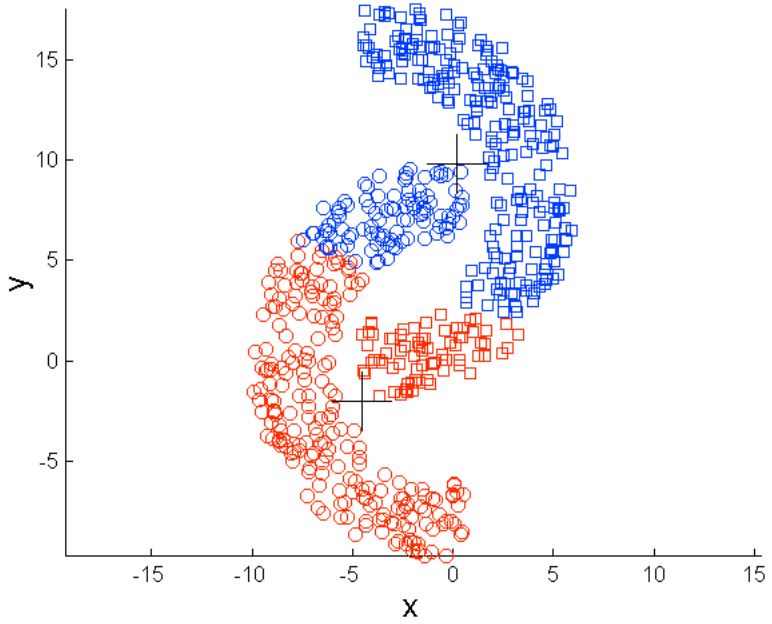


Diskuze

- Jaká jsou omezení k-means?
- Jak je překonat?

Omezení k-means: Nekulovité tvary v datech

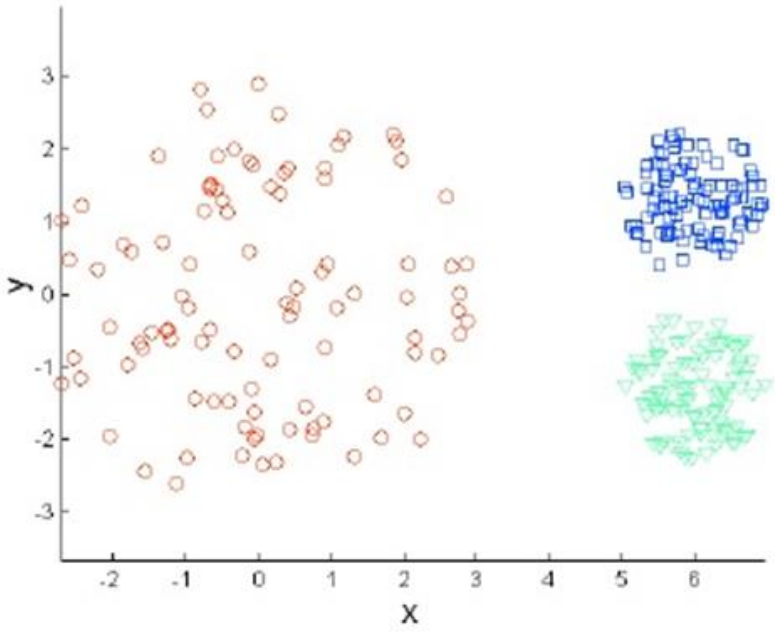




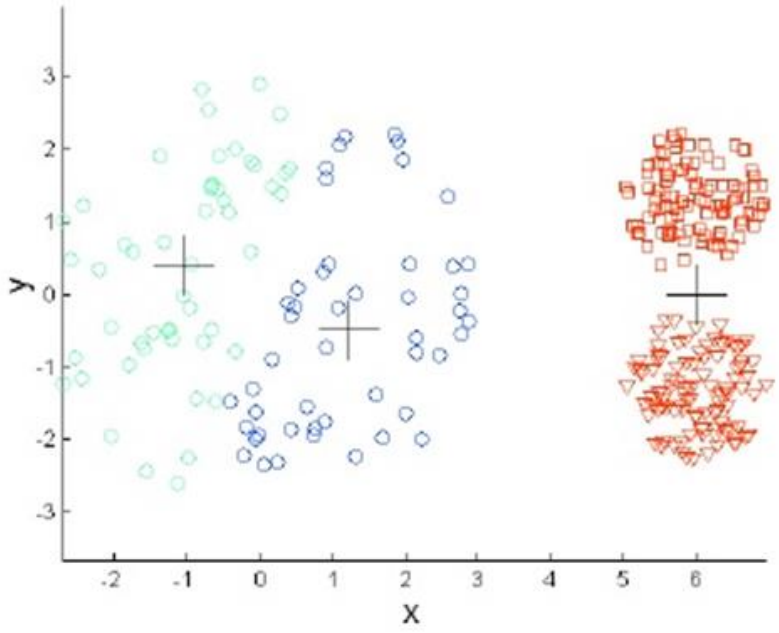
Data

Zdroj: http://www.scribd.com/doc/6646055/16/Limitations-of-K-means-Non-globular-Shapes

Omezení k-means: Různá hustota dat



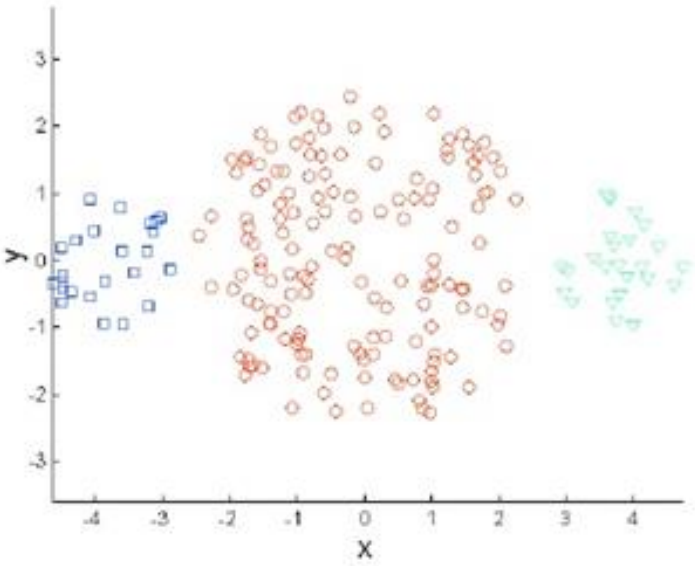
Data



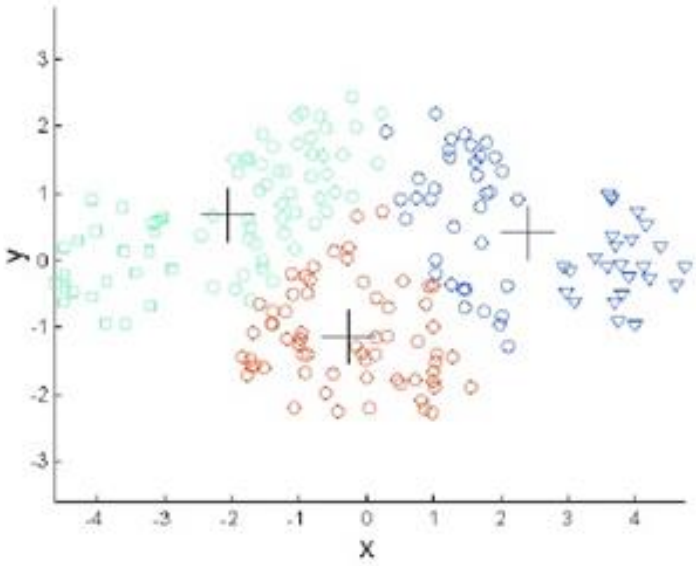
K-means (3 Shluky)

Zdroj: http://www.scribd.com/doc/6646055/16/Limitations-of-K-means-Non-globular-Shapes

Omezení k-means: Různá velikost shluků



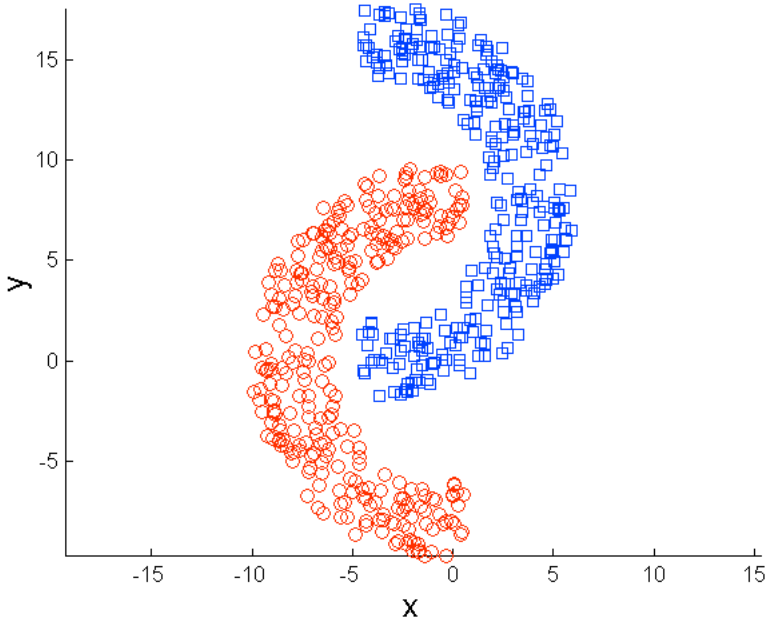
Data



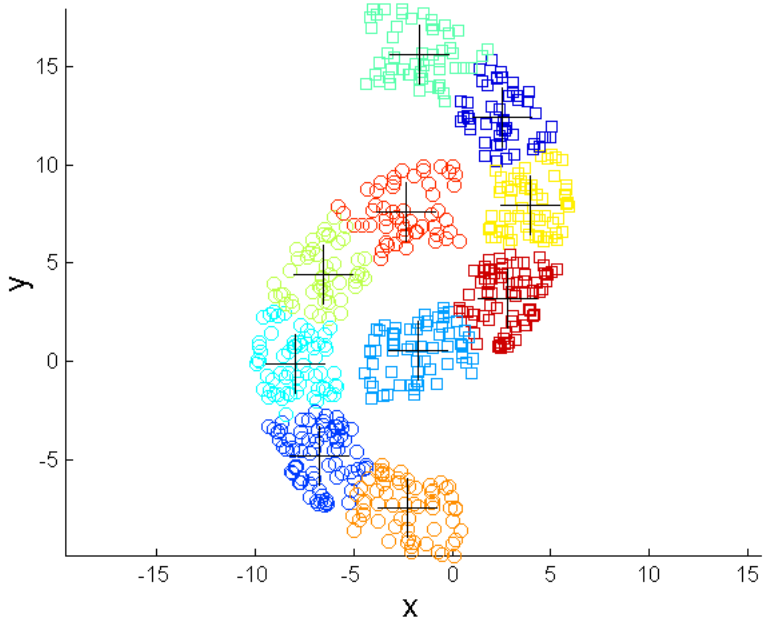
K-means (3 Shluky)

Zdroj: http://www.scribd.com/doc/6646055/16/Limitations-of-K-means-Non-globular-Shapes

Overcoming K-means Limitations



Original Points



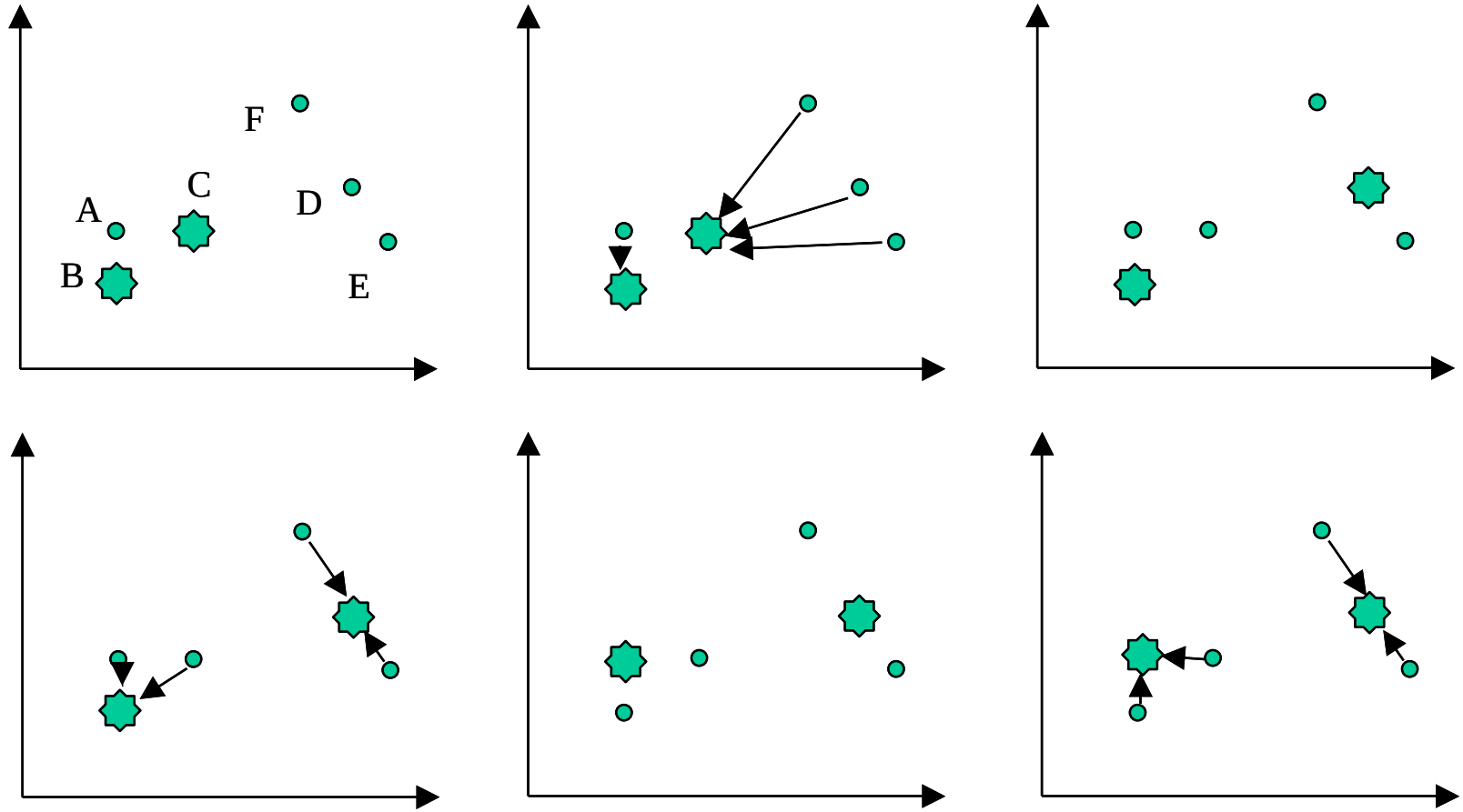
K-means Clusters

Zdroj: http://www.scribd.com/doc/6646055/16/Limitations-of-K-means-Non-globular-Shapes

Problémy k-means algoritmu

- velmi jednoduchý algoritmus, vždy najde řešení
- velmi citlivý k volbě prvních centroidů a v závislosti na nich může najít pouze lokální optimum.
- omezení pro
  - různé hustoty dat
  - různé velikosti shluků
  - nekulovité tvary dat
- omezen na data v Eukleidovském prostoru, memůže být použit pro shlukování kategorických dat
  - k-medoids: Již se nepočítá geometrické centrum každého shluku, ale algoritmus nalezne objekt, který je mu nejblíže, a který potom tento shluk reprezentuje

Příklad k-medoids algoritmu



Jak určit optimální počet shluků?

- Počet shluků K musí být zadán před zahájením shlukování.
- Jak jej určit?
- Obvyklým řešením je, že se shlukování provede pro celou řadu počtu shluků. Následně se provede analýza a vybere nejvhodnější řešení.
- Analytické řešení
  - Na základě vzdáleností uvnitř a mezi shluky
  - Silhouette koeficienty

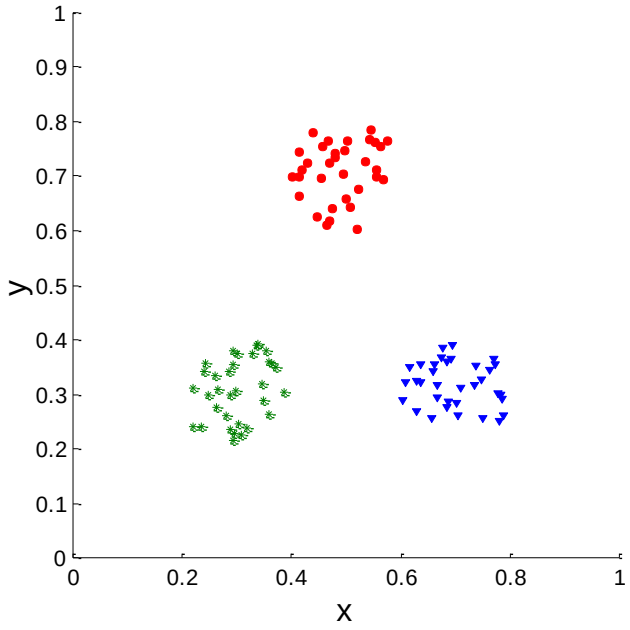
Silhouette koeficienty

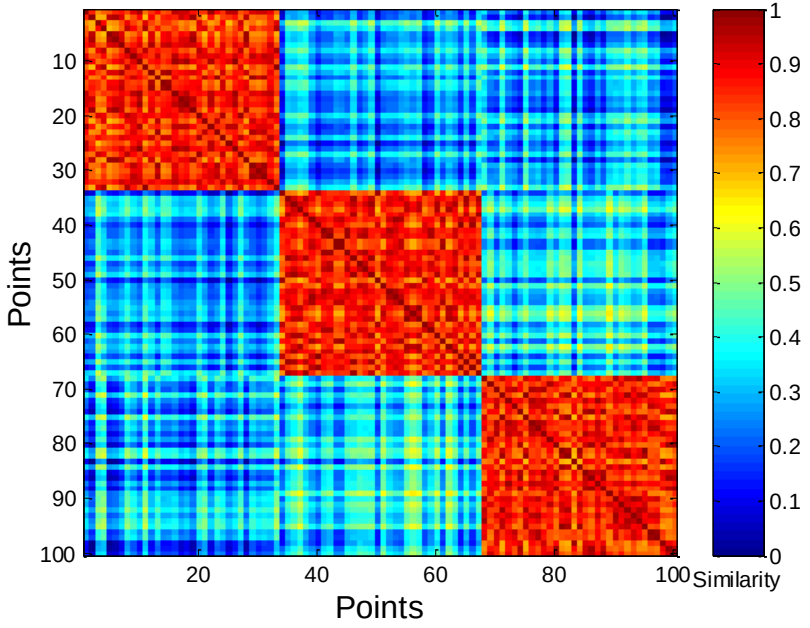
- Pro každý objekt i bude podle následující rovnice spočítána jeho takzvaná silueta, která v podstatě porovnává spojitost každého shluku vůči vzdálenosti mezi shluky
- kde a_i udává průměrnou vzdálenost objektu i vůči ostatním objektům v tomtéž shluku, a b_i udává průměrnou vzdálenost objektu i vůči všem objektům v dalším nejbližším shluku (další nejlepší kandidát pro umístění objektu i).
- Je zřejmé, že je definováno na intervalu (-1,1).

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

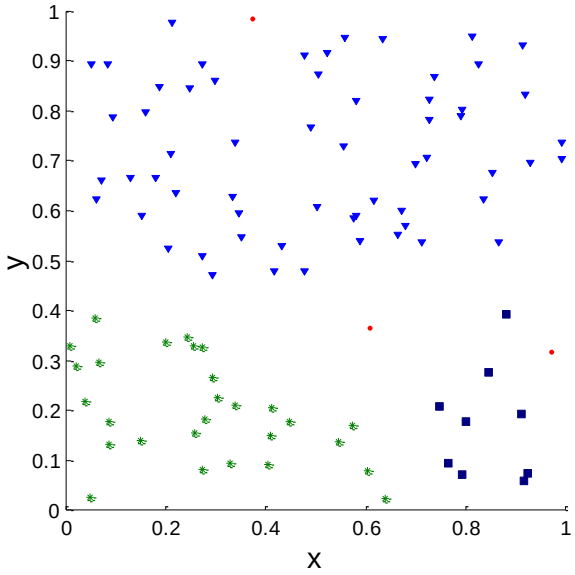
SC	Doporučený význam
0.71 – 1.00	Byla nalezena silná struktura
0.51 – 0.70	Byla nalezena dostatečná struktura
0.26 – 0.50	Nalezená struktura je slabá a doporučuje se vyzkoušet jiný přístup
< 0.25	Žádná podstatná struktura nalezena nebyla

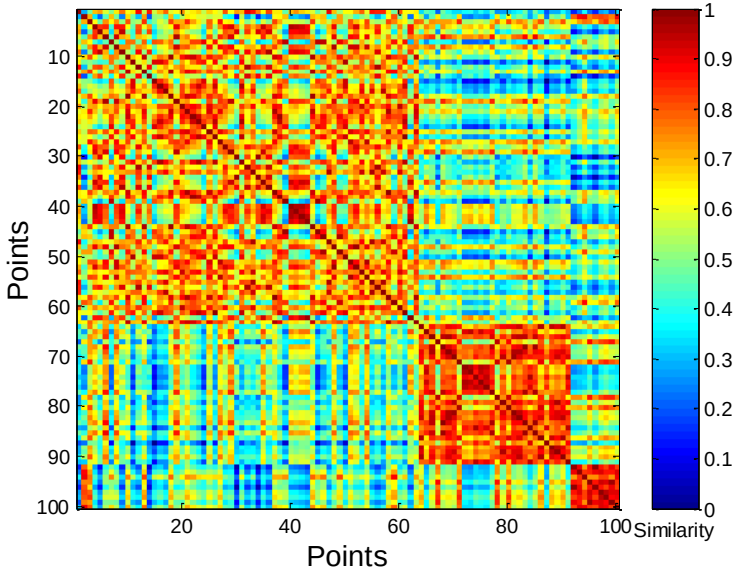
matice podobnosti pro ověření kvality shluků





matice podobnosti pro ověření kvality shluků





ROZHODOVACÍ STROMY

Co jsou rozhodovací stromy?

- Analytický nástroj
  - Popisuje vztah mezi nezávislými a závislými proměnnými
- Podporuje rozhodování
- Používají se často v marketingu – segmentace trhu
  - Jaká cílová skupina kupuje naše zboží?
- Výsledný rozhodovací strom je snadno čitelný pro člověka
  - Rozhodovací pravidla (IF-THEN rules)

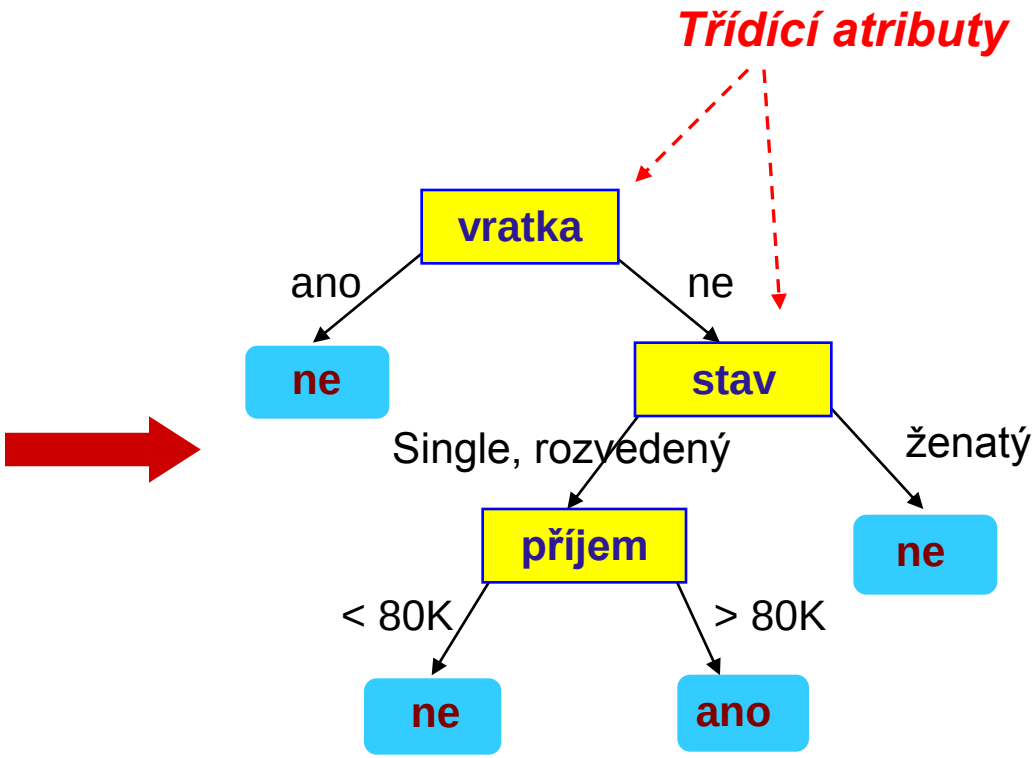
Základní dělení

- Binární × Multinomiální
- Spojité × diskrétní × kombinované hodnoty proměnných

Příklad

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

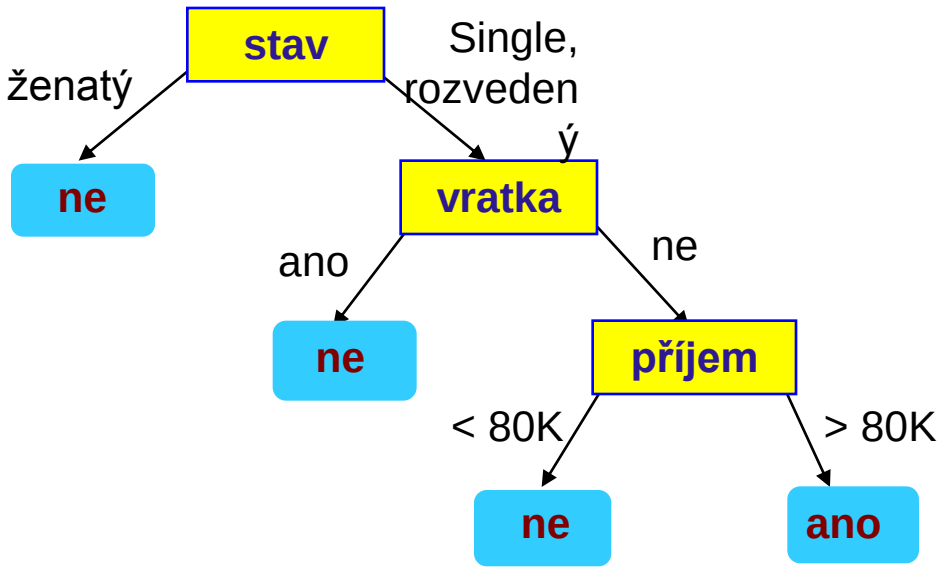
Trénovací množina



Model: Rozhodovací strom

Jiný model

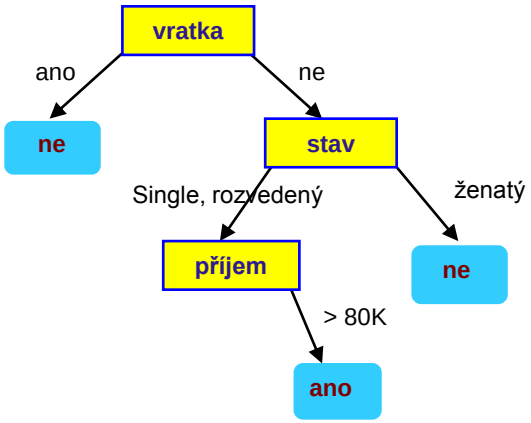
	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

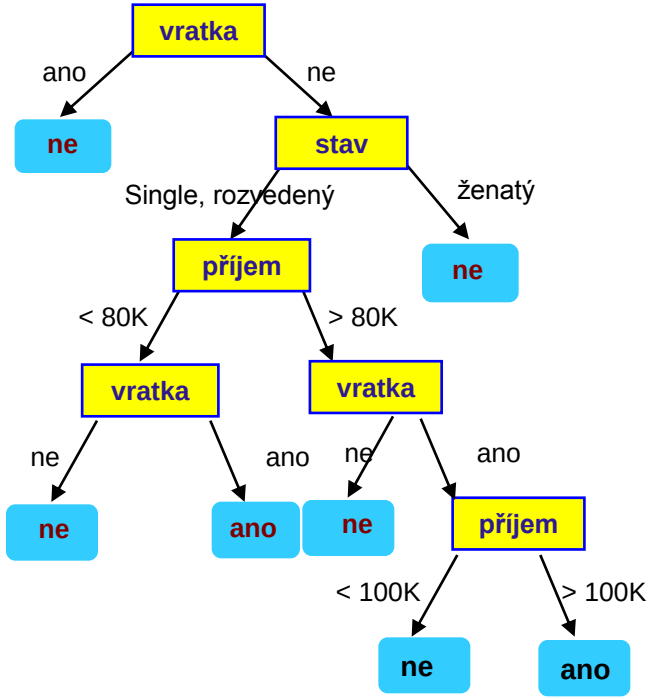


There could be more than one tree that fits the same data!

Diskuse

Jaký z následujících modelů zvolit?





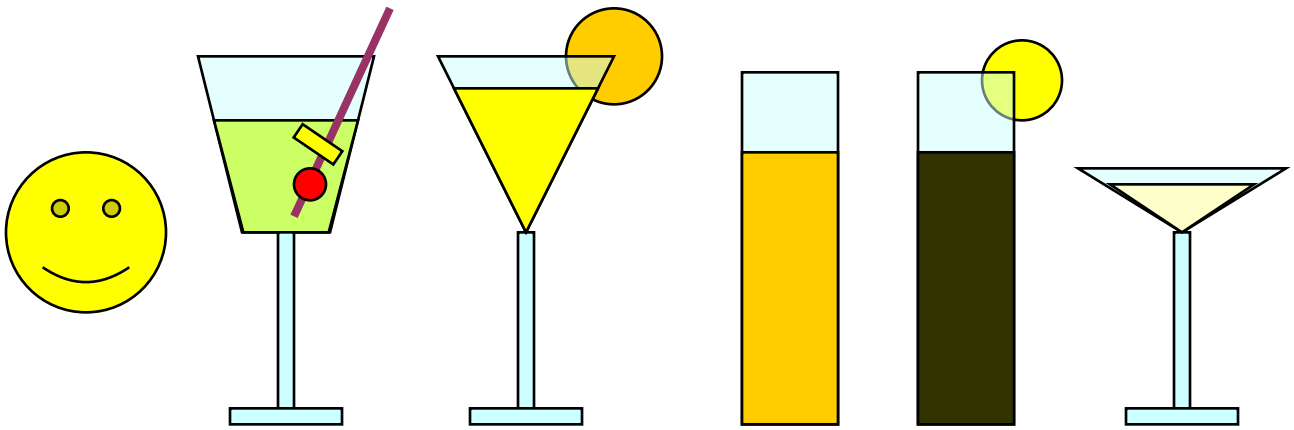
Okamova břitva:

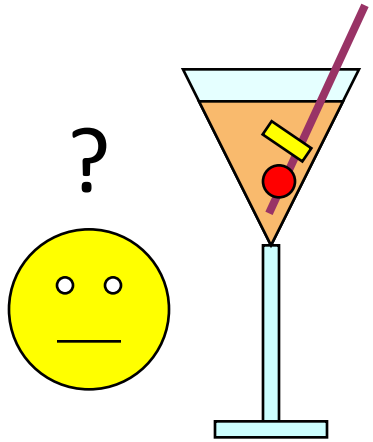
Pokud pro nějaký jev existuje vícero vysvětlení, je lépe ***upřednostňovat to nejméně komplikované****.*

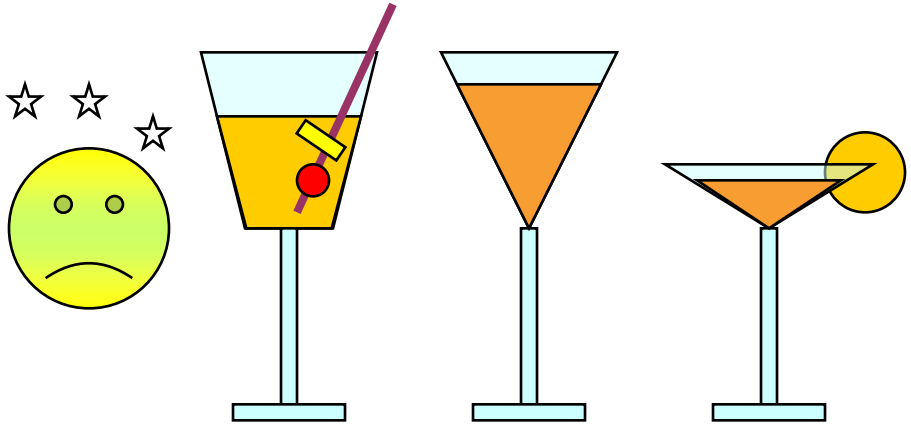
Indukce rozhodovacích stromů

- Mnoho různých algoritmů
  - CART
  - C4.5
  - CHAID
- Jedná se o „**greedy**“ metody
  - Vždy zvolí optimalizaci v každém kroku,
    Může vést k nalezení lokálních extrémů
- **Otázky**
  - Jak rozdělit záznamy?
    - Jaké zvolit atributy?
    - Jaké zvolit hranice?
  - Kdy skončit dělení?

Příklad



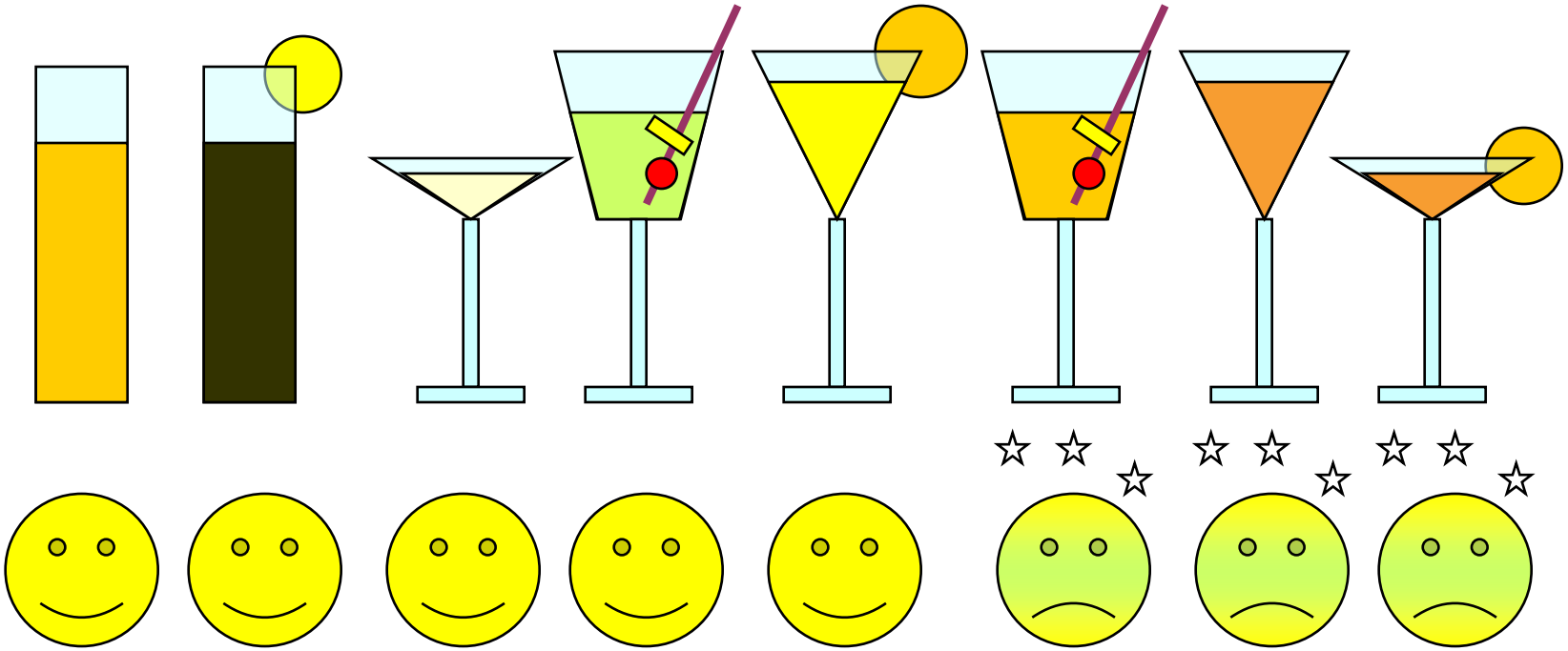




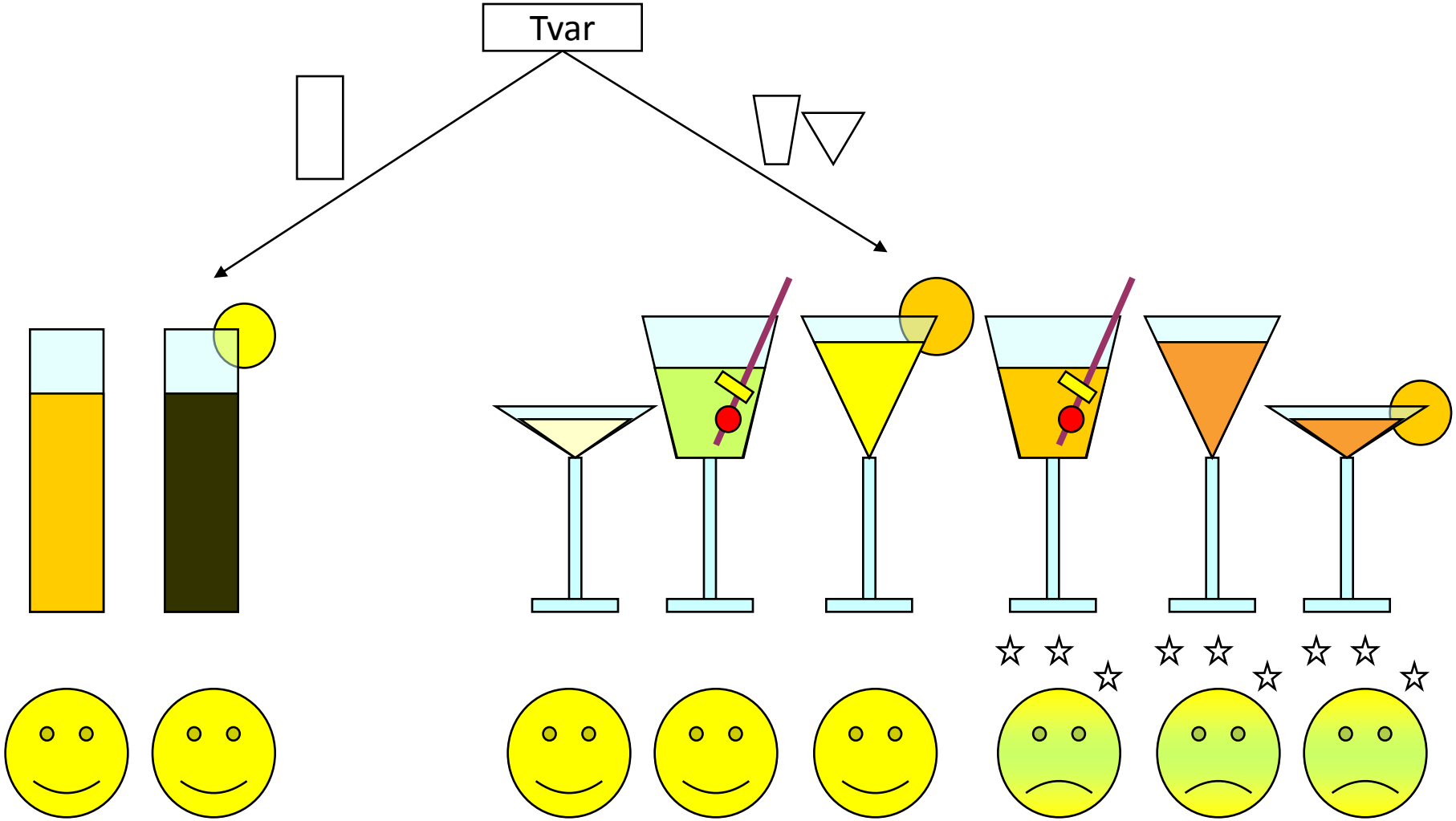
Zkušenosti

Otázka: Bude mi po tomto nápoji špatně či nikoli?

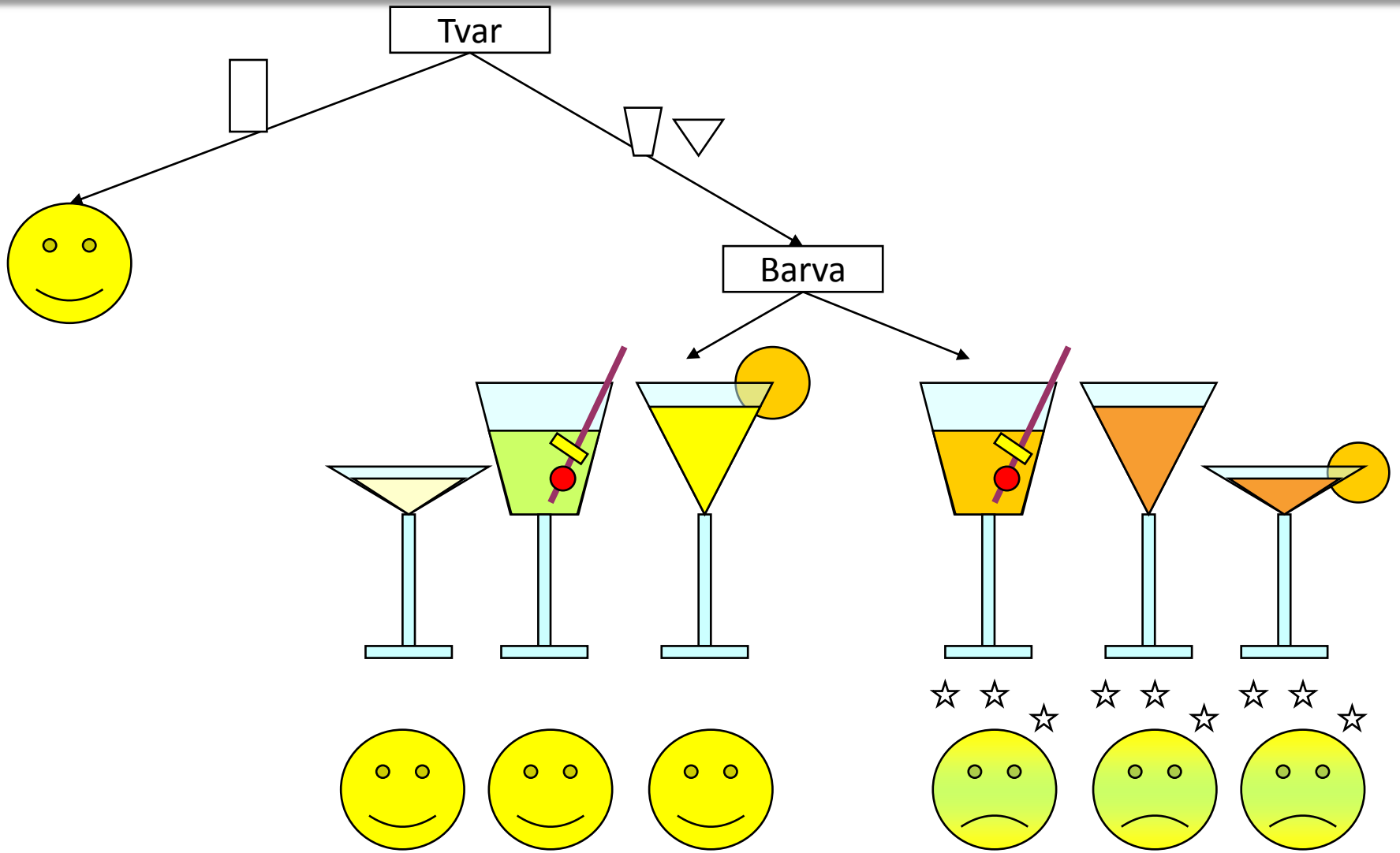
Množina: 8 klasifikovaných prvků



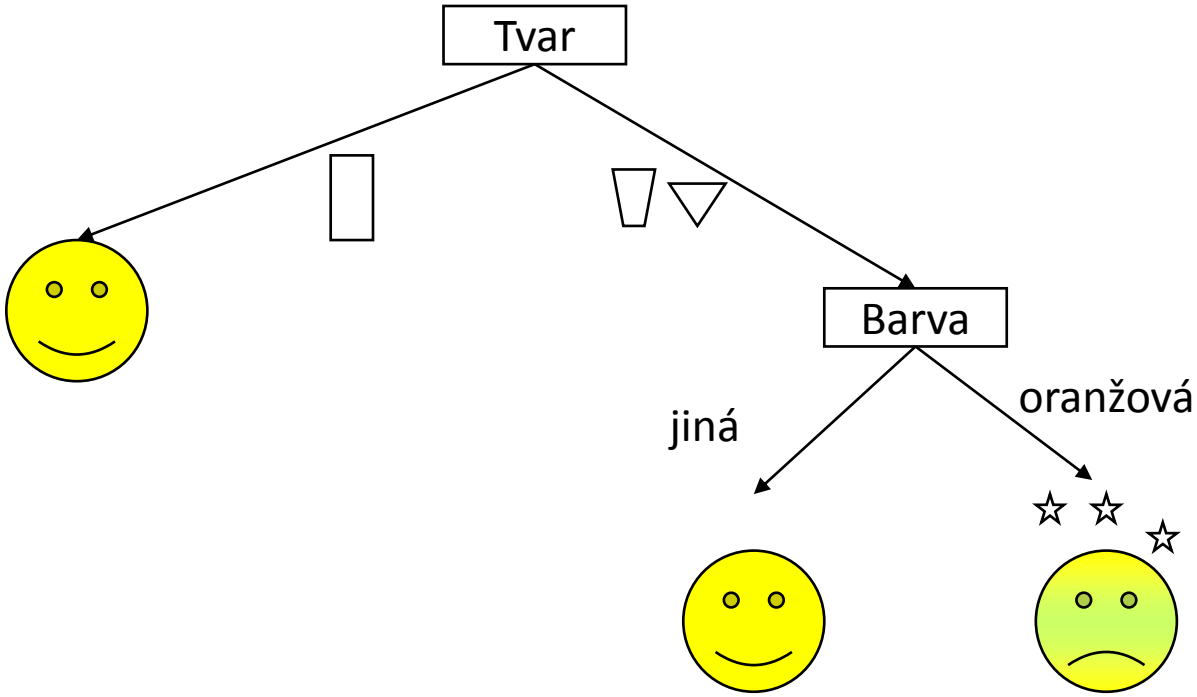
Pozorování 1: TVAR je důležitý

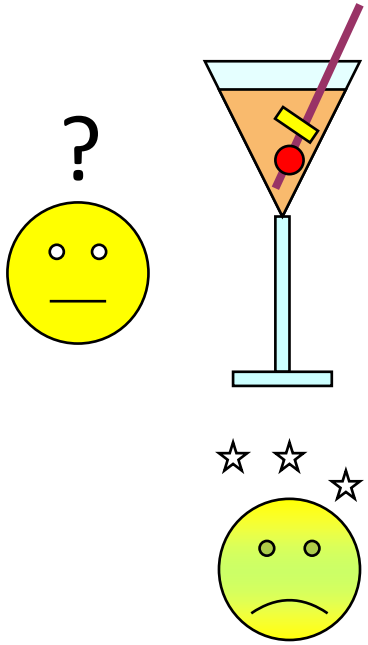


Pozorování 2: Pro některé tvary je důležitá BARVA



Dedukce





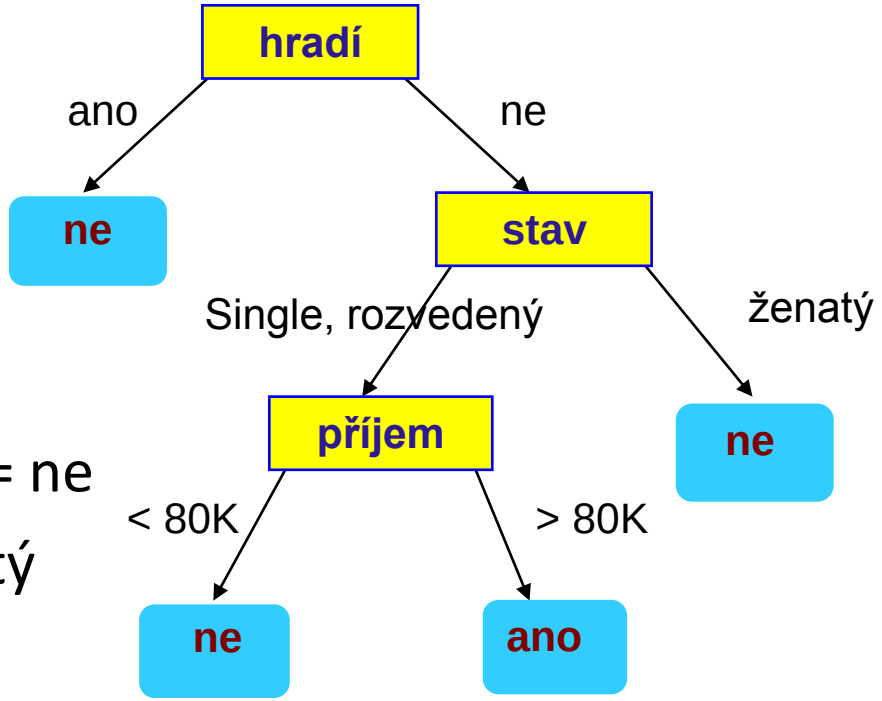
Indukce rozhodovacích stromů - princip

Indukce:

- Vyber atribut, **který nejvíce popisuje** výstupní atribut (který od sebe nejlépe odliší příklady z různých tříd)
- Rozděl tento atribut tak aby bylo maximalizováno určité kritérium
  - **GINI index** (“Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset”)
  - **Entropie** (udává homogenitu atributu)
  - **Počet chybně klasifikovaných vzorků**
  - …

Výstupy

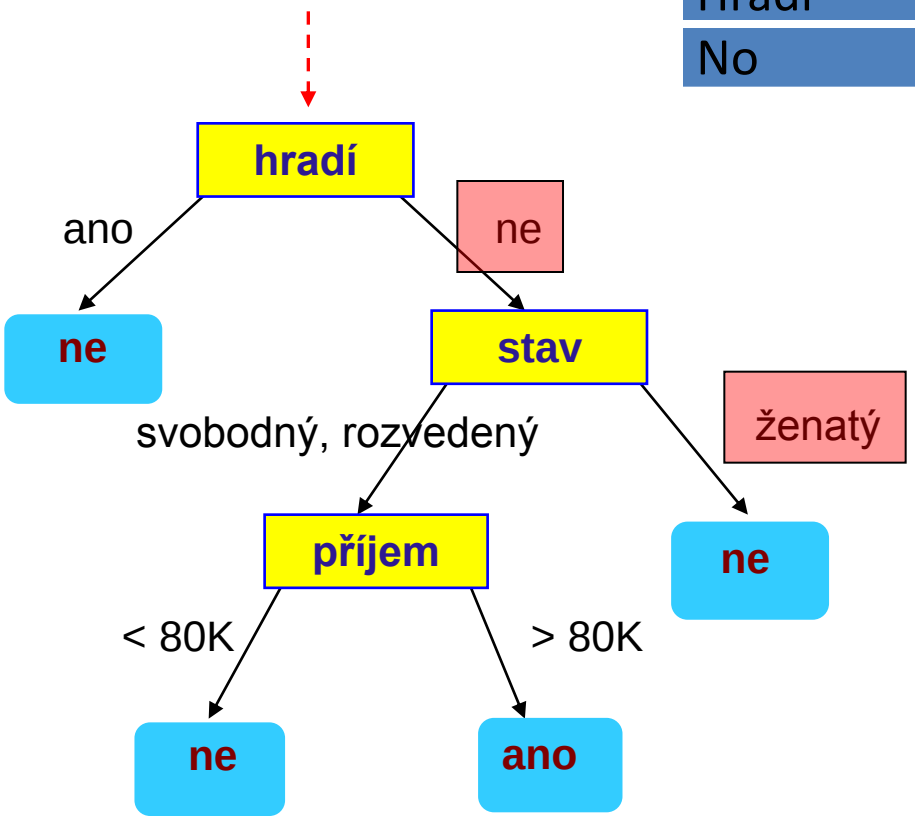
- **Rozhodovací strom**
  - Podvádí či ne?



- **Rozhodovací pravidla**
  - IF hradí = ano THEN podvádí = ne
  - IF hradí = ne AND stav = ženatý THEN podvádí = ne
  - IF hradí = ne AND stav = Single, rozvedený AND příjem < 80K THEN podvádí = ne
  - ELSE podvádí = ano

Aplikování modelu

Začít od kořene stromu



Testovací data

Hradí	Stav	příjem	podvádí
No	Ženatý	80K	?

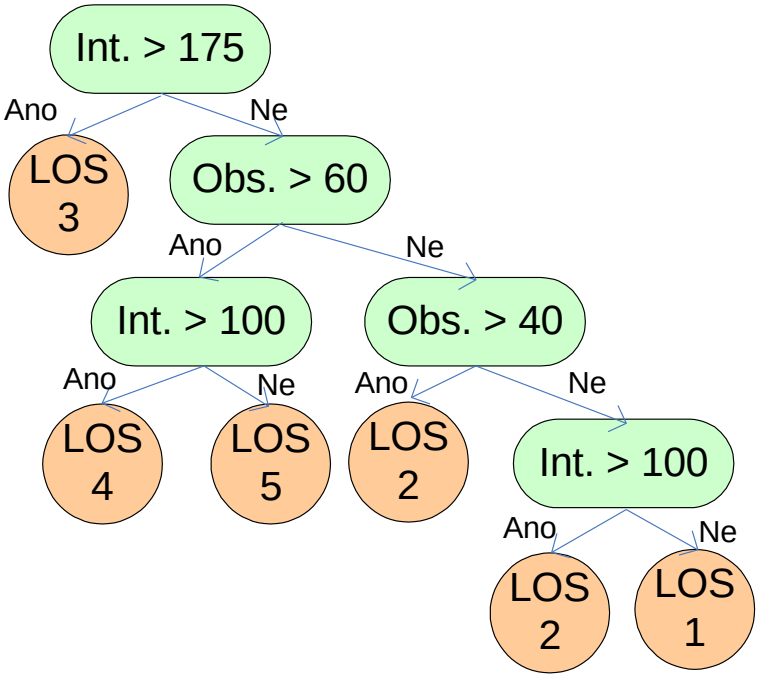
Podvádí? = ne

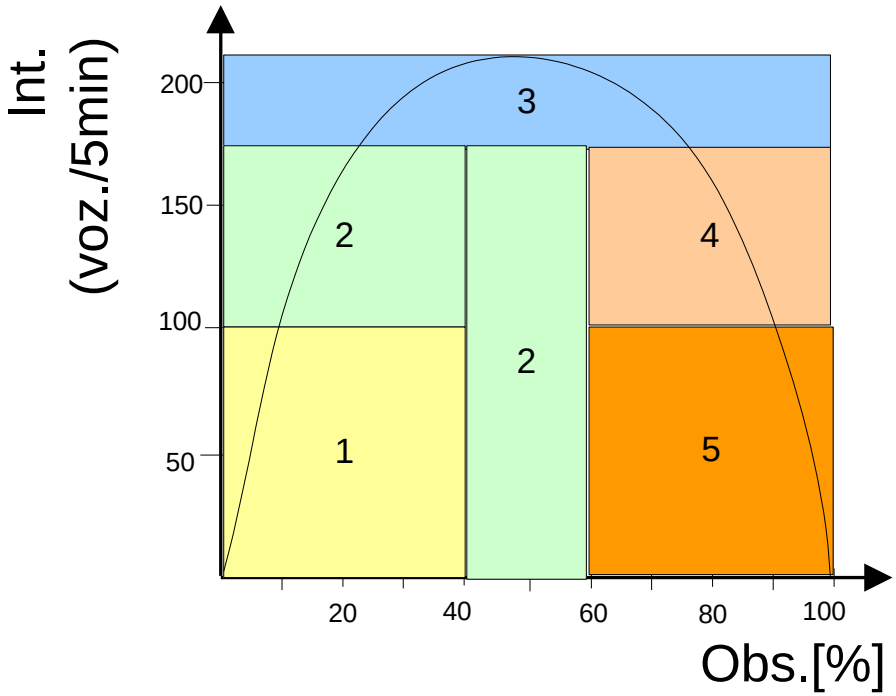
Diskuze

- Jmenujte některé aplikace z oblasti dopravy / ITS.
- Vlastnosti rozhodovacích stromů? (výhody, nevýhody)

Příklad využití v dopravě

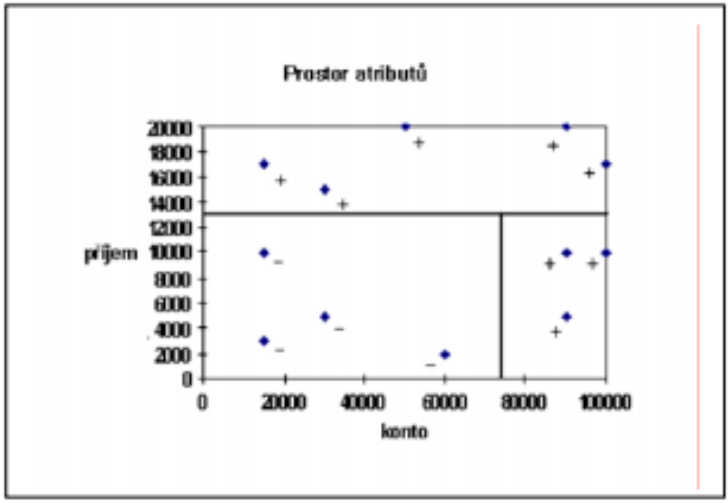
- Automatický klasifikátor stupně dopravy
- Pracuje v grafu závislosti intenzita / obsazenost

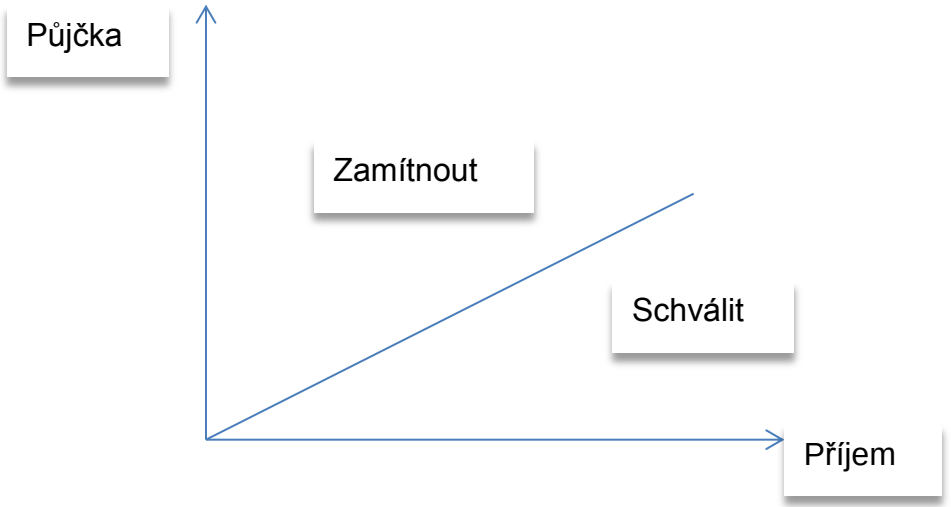




Rozhodovací stromy - vlastnosti

- Výhody
  - Akceptují chybějící hodnoty
  - Akceptují spojité i diskrétní hodnoty
  - Výsledek ve formě snadno pochopitelných pravidel
  - Nalezne důležité proměnné
- Nevýhody
  - Vhodné pro oblasti s obdélníkovými oblastmi
  - Mohou být příliš velké pro některé aplikace
    - Doplňující prořezávání

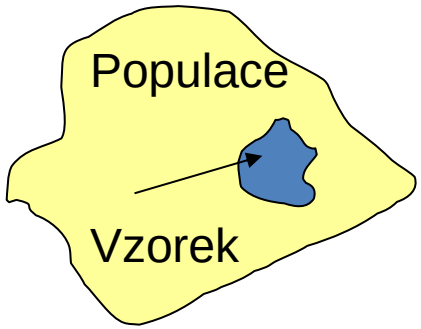


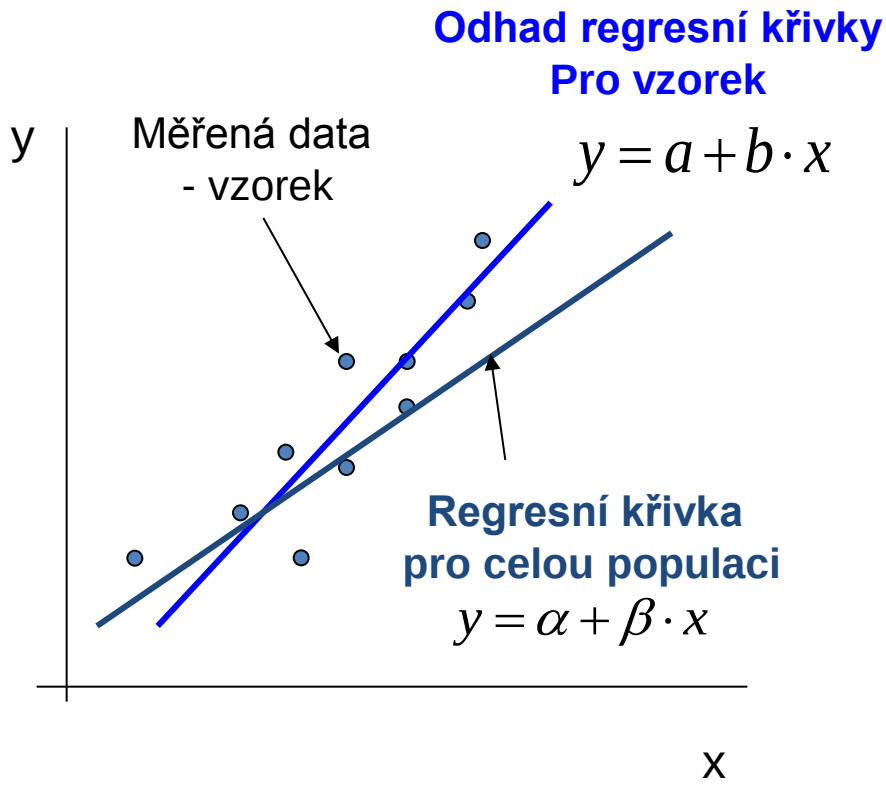


Princip metody lineární regrese

- Předpoklad
  - x…nezávislé proměnné
  - y…závislá proměnná
  - koeficienty

- Hledáme odhad těchto koeficientů
  - a,b α, β $y = \alpha + \beta \cdot x$
- Metody odhadu
  - Nejmenší čtverce





Metoda nejmenších čtverců

Gauss (ale i Legendre, podobné postupy u Laplace, Cotese)

- Metoda pro řešení přeurčené soustavy lineárních rovnic

$$\boldsymbol{X}\boldsymbol{z} = \boldsymbol{y}$$

- Soustava nemá řešení, reziduum $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{z}$ má ale *minimum*
- Hodnotu

$$\hat{\boldsymbol{z}} = \arg\min_{\boldsymbol{z}} |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{z}|^2$$

  lze určit jako $\boldsymbol{z} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$

- Pro lineární regresi s $y_i = \alpha x_i + \beta$ je ta soustava

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Lineární regrese - vlastnosti

Lineární regresní model je vztah, který *sumarizuje soubor dat*

- Hledáme nejlepší odhad neznámých parametrů tohoto modelu (zde Gauss vs. Legendre)
- Snahou je nalézt odhady b parametrů β, které jsou
  - nejlepšími (Best),
  - nestrannými (Unbiased) a
  - lineárními (Linear)
  - odhady

  } (NNLO/BLUE)

- Odhady mají asymptoticky normální rozdělení

Diskuse

- Jak je možné dosáhnout nalezení NNLO (BLUE) odhadů a mají asymptoticky normální rozdělení?

- Za splnění základních předpokladů:

Předpoklady metody nejmenších čtverců

1. Regresní parametry β mohou nabývat libovolných hodnot.
2. Regresní model je lineární v parametrech.
3. Matice nenáhodných, nastavovaných hodnot vysvětlujících proměnných X má hodnost rovnou právě m. To znamená, že žádné její dva sloupce x_j , x_k nejsou kolineární, tj. rovnoběžné
4. Náhodné chyby g_i mají nulovou střední hodnotu $E(g_i) = 0$, konstantní a konečný rozptyl $E(var(g_i)) = \sigma^2$ a celkově mají normální rozdělení $N(0, \sigma)$.
5. Také podmíněný rozptyl $D(y/x) = \sigma^2$ je konstantní.
6. Náhodné chyby gi jsou vzájemně nekorelované a platí Cov(gi,gj) = E(gi gj) = 0.
7. Pokud mají chyby normální rozdělení, jsou nezávislé. Tento požadavek odpovídá požadavku nezávislosti měřených veličin y.

Omezení metody nejmenších čtverců

1. Předpokládá, že x_i jsou přesná
2. Gramián $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ může být špatně podmíněný, odhad „zesiluje“ chybu měření
3. Minimalizuje se $|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{z}}|$, někdy bychom raději $|\boldsymbol{z} - \hat{\boldsymbol{z}}|$

Vlastnosti lineární regrese (zjednodušeně)

- Výhody
  - Velmi kvalitně zpracované metody
  - Poměrně jednoduché
  - Výpočetně nenáročné
- Nevýhody
  - Téměř nikdy nemohou být splněny všechny předpoklady
    - Existuje řada opatření a jiných statistických modelů, která se to snaží vyřešit (Logit, Probit, …)
- Alternativní řešení
  - Použití jiných metod
    - Rozhodovací stromy
    - Soft-computing (FS, ANN, GA, …)