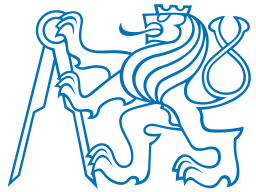
Úvod do analýzy dat

Matematické metody pro ITS (11MAMY)

Ondřej Přibyl (Jan Přikryl)

Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní

verze 2022-03-23 19:37: Opraveny drobné chyby v prezentaci, sloučeno.



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.

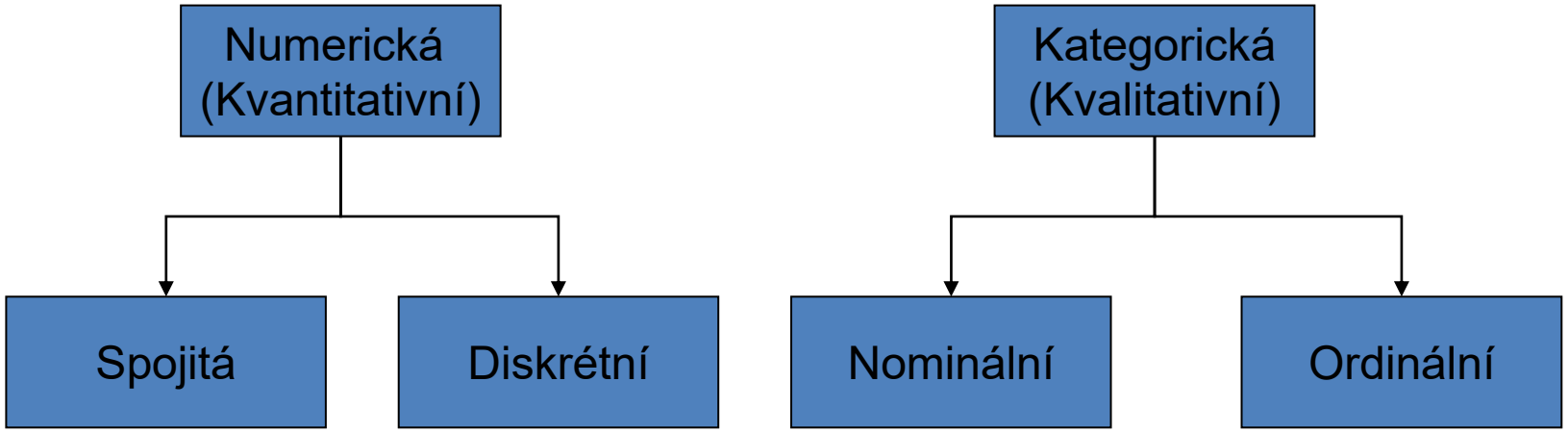
Moudrost

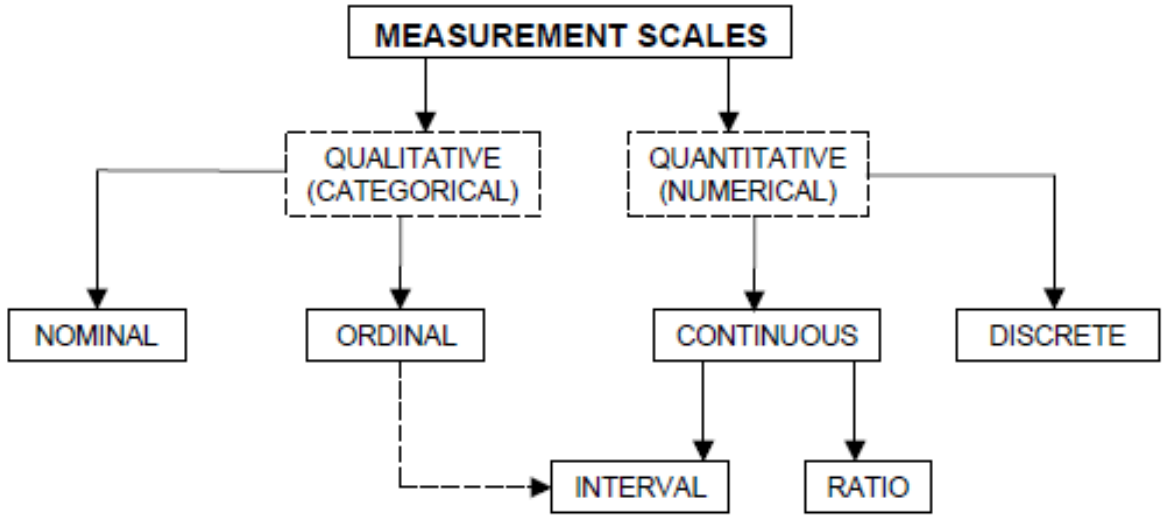
Porozumění principům.

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.
Dělíme je na **náhodné** (dynamická charakteristika) a **systematické** (statická charakteristika)

- **Chyba detektoru**
  - Chyba měřicího zařízení (způsobená 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.
Dělíme je na **náhodné** (dynamická charakteristika) a **systematické** (statická charakteristika)

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

Viz také Eulerova metoda pro numerické řešení ODR …

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

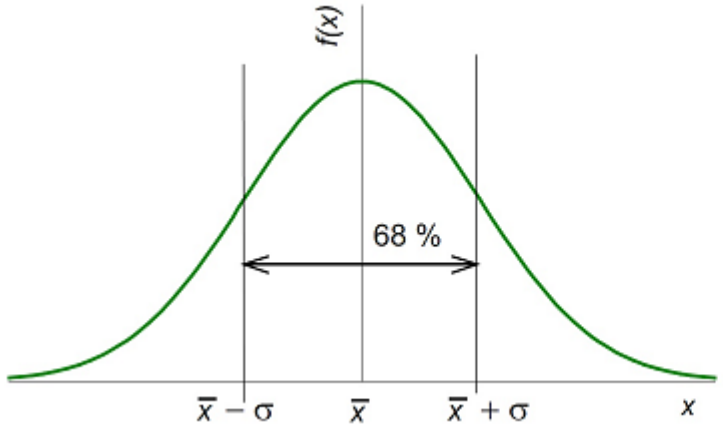
Úvod do problematiky

- 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.

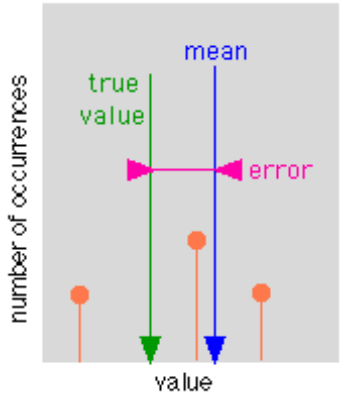


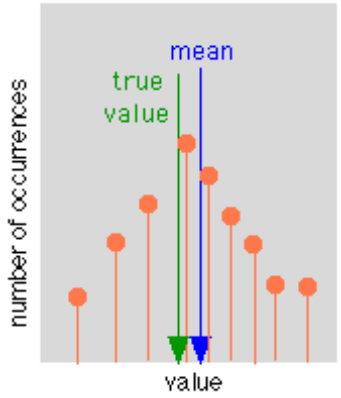
hustota pravděpodobnosti

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

- 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

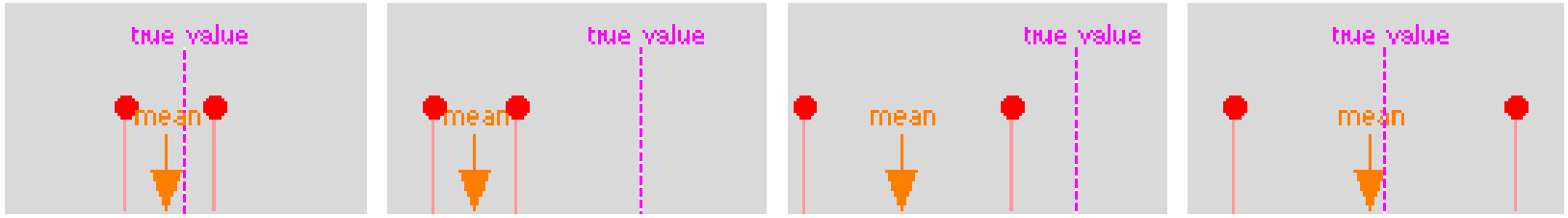
„Přesnost“ versus „správnost“

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

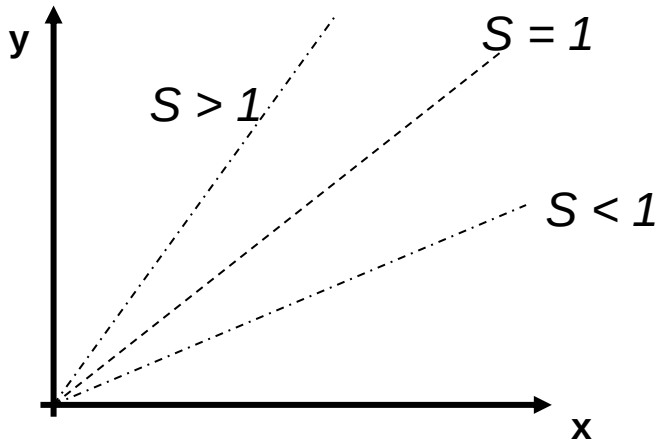


Citlivost (sensitivity) měřicího přístroje

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**

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

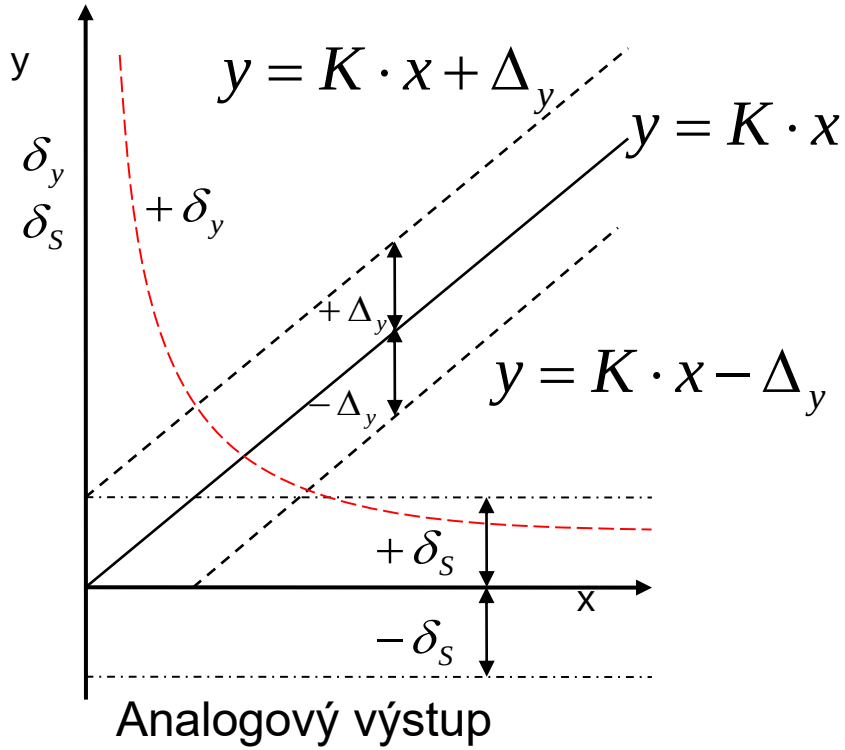
3. **Relativní chyba senzoru**
   Poměr maximální absolutní chyby měření vůči rozsahu hodnot měřené veličiny

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

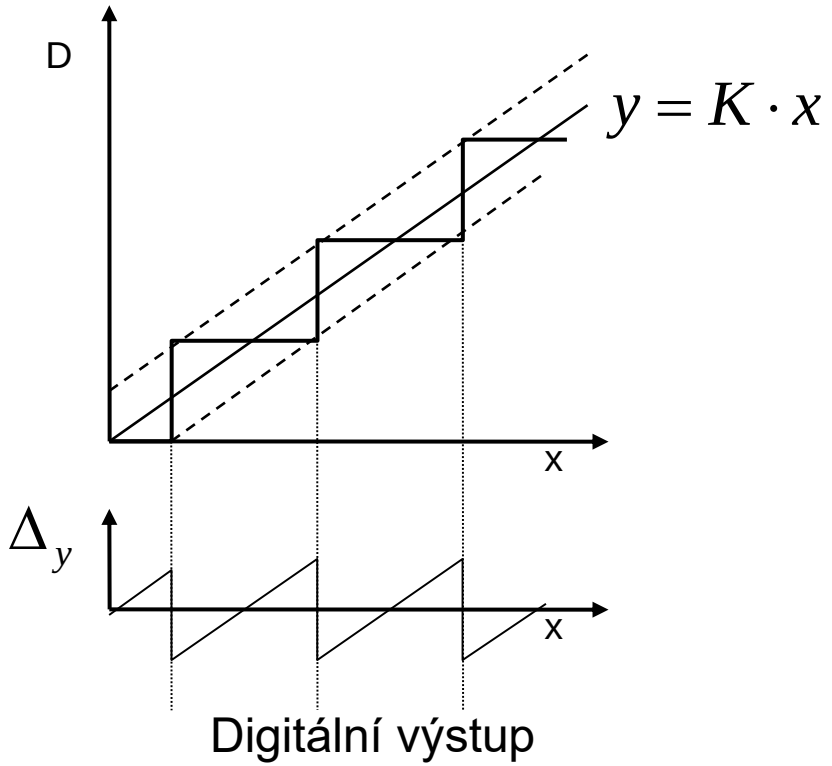
Chyby měření (pokrač.)

4. Aditivní chyba měření

- Způsobena posunem jmenovité lineární charakteristiky
- Absolutní chyba měření x je konstantní
- Relativní chyba měření x závisí hyperbolicky na x



Analogový výstup

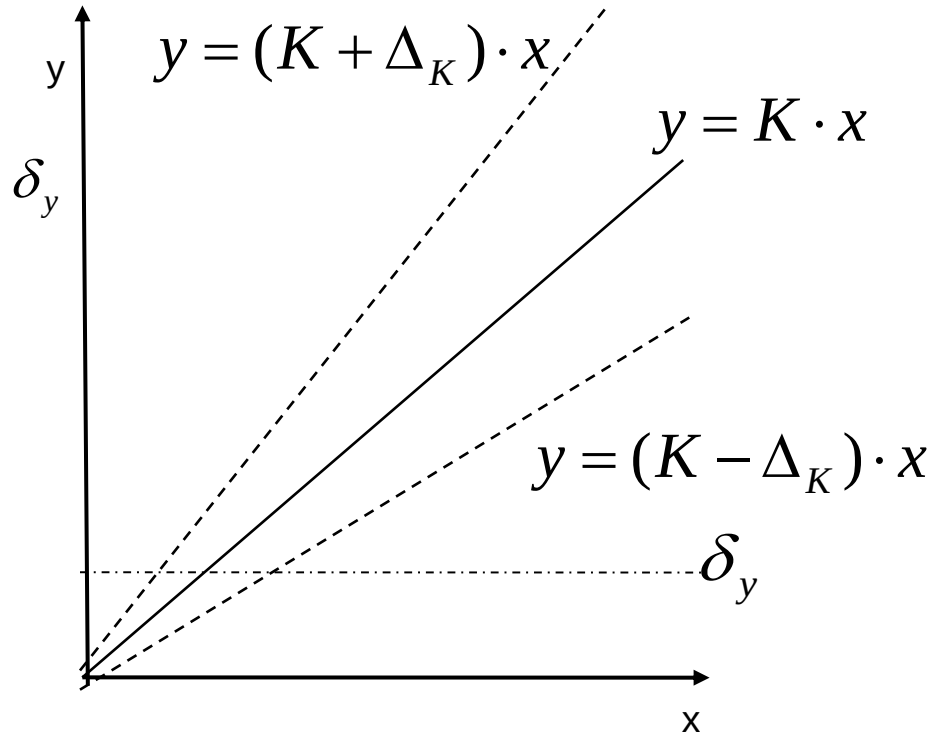


Digitální výstup

Chyby měření

Multiplikativní chyba měření

- Je ekvivalentní změně citlivosti senzoru
- Absolutní chyba je závislá na hodnotě měřené veličiny
- Relativní chyba měření je konstantní



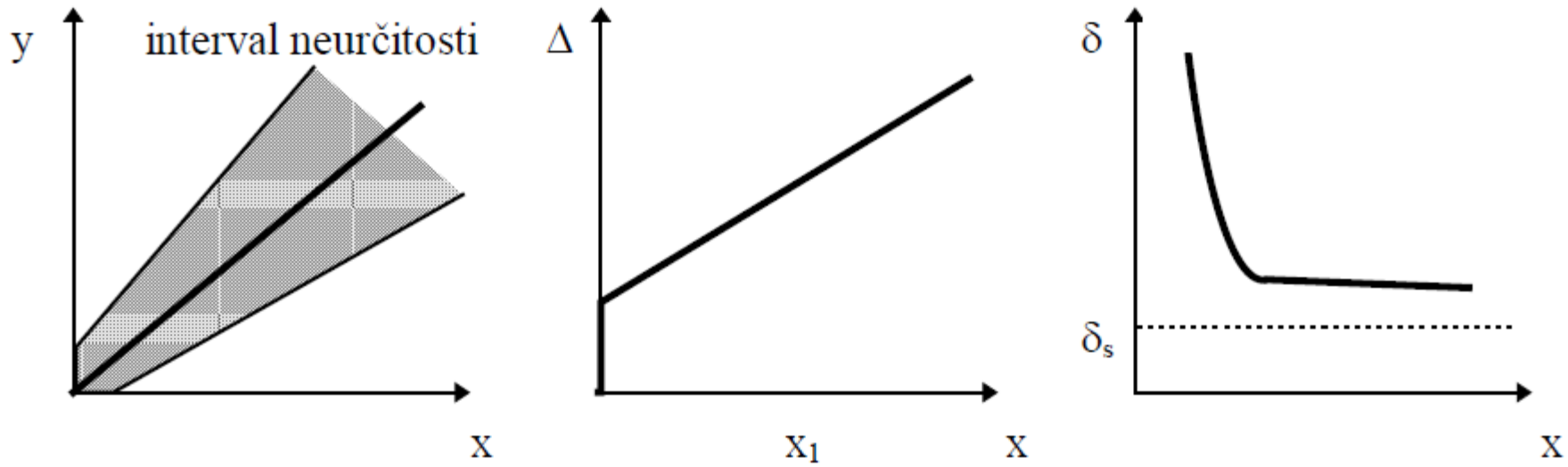
$$\Delta_y = \Delta_K \cdot x$$

$$\delta_y = \frac{\Delta_y}{y} = \delta_K = konst.$$

δ_K Chyba měření

Chyby měření

- Kombinovaná chyba měření
- Kombinace aditivní a multiplikativní chyby

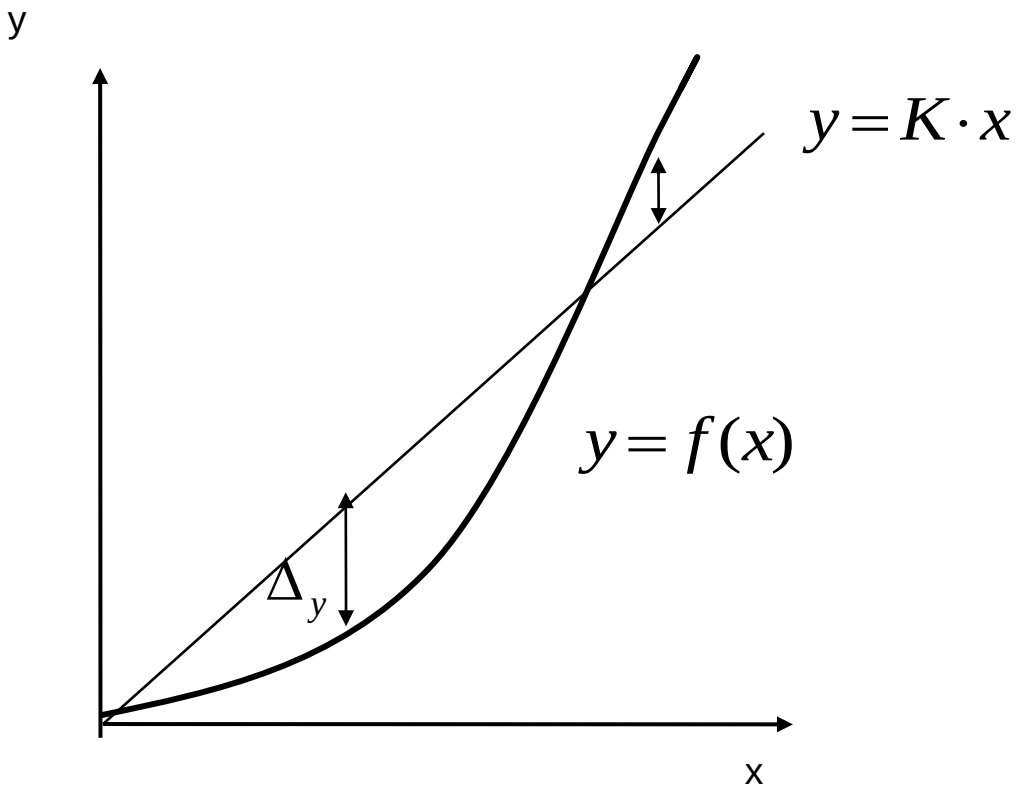


Chyby měření

Chyba linearity

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

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

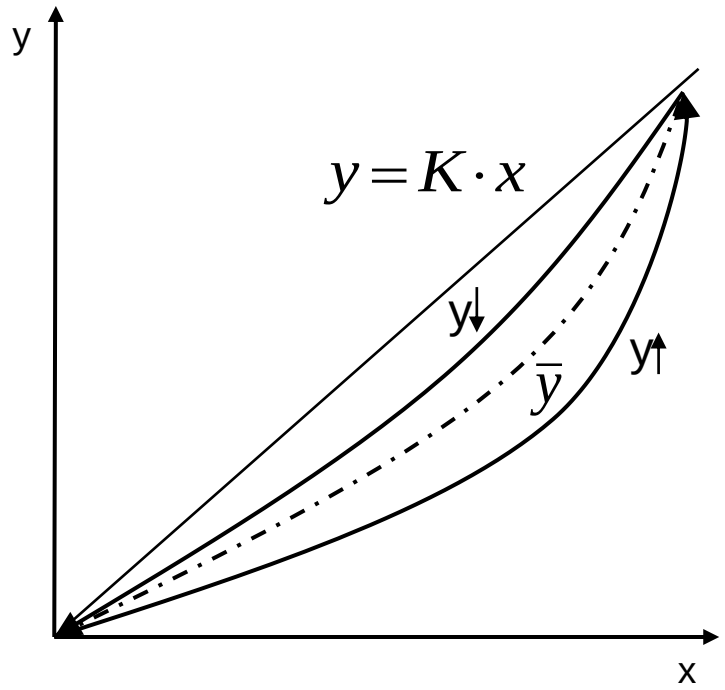


- kde y_L je definována ideální funkcí $y = y_0 + Kx$,
- parametr K lze odhadnout pomocí lineární regrese.

Chyby měření

Chyba hystereze

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

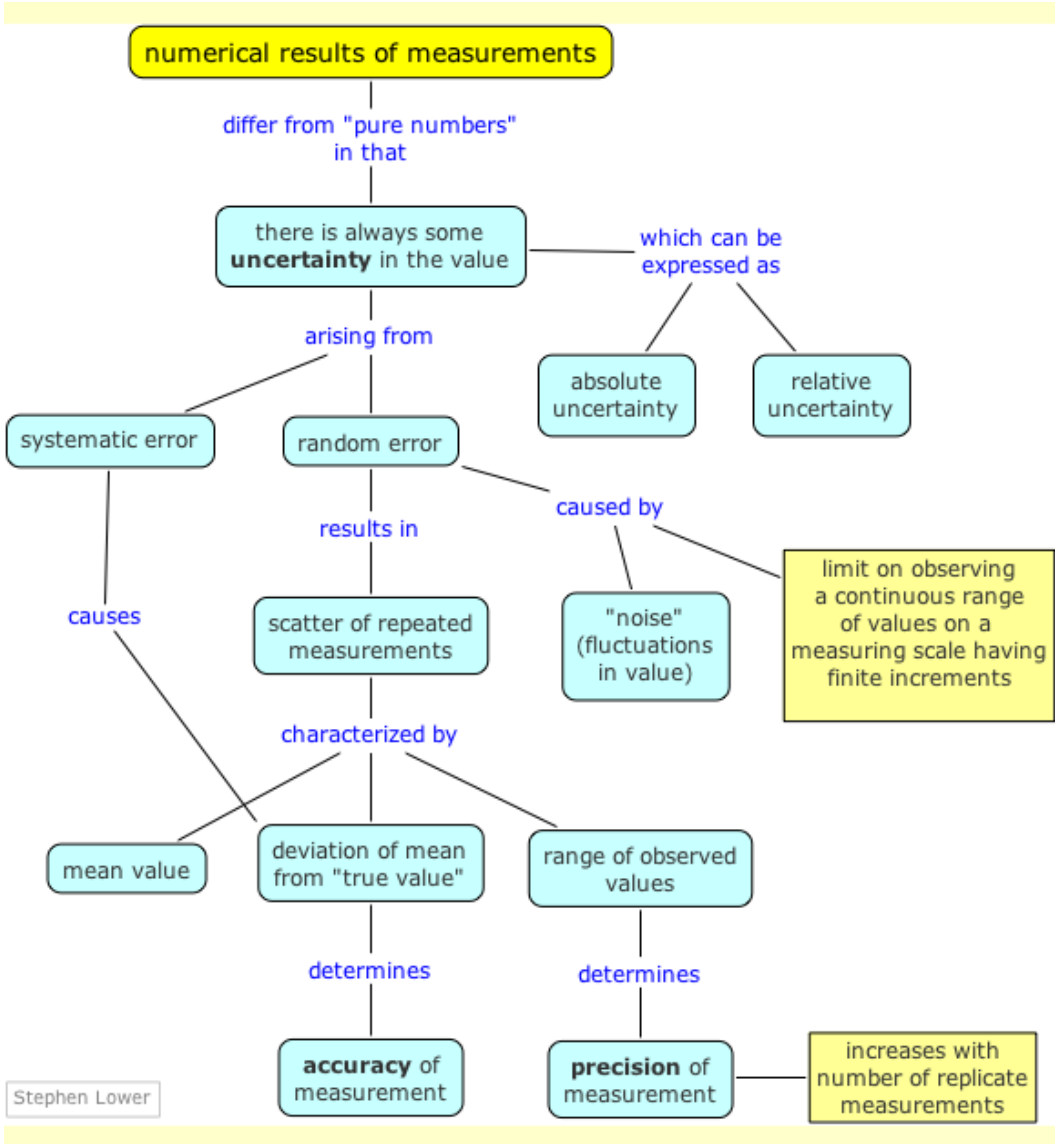


$$\delta_S = \left(\frac{y\downarrow - y\uparrow}{y_{\max}} \right)_{\max} = \left(\frac{\Delta_{yH}}{y_{\max}} \right)_{\max}$$

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

kde $\bar{y}$ je střední hodnota vzestupné a klesající závislosti y.

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

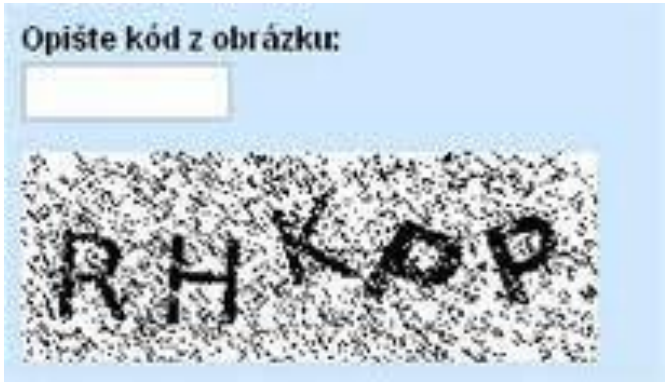
- Měřené veličiny
- Chyby měření
- **Základní charakteristiky dat**
- Vizualizace dat
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

Co je to průzkumová analýza dat?

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

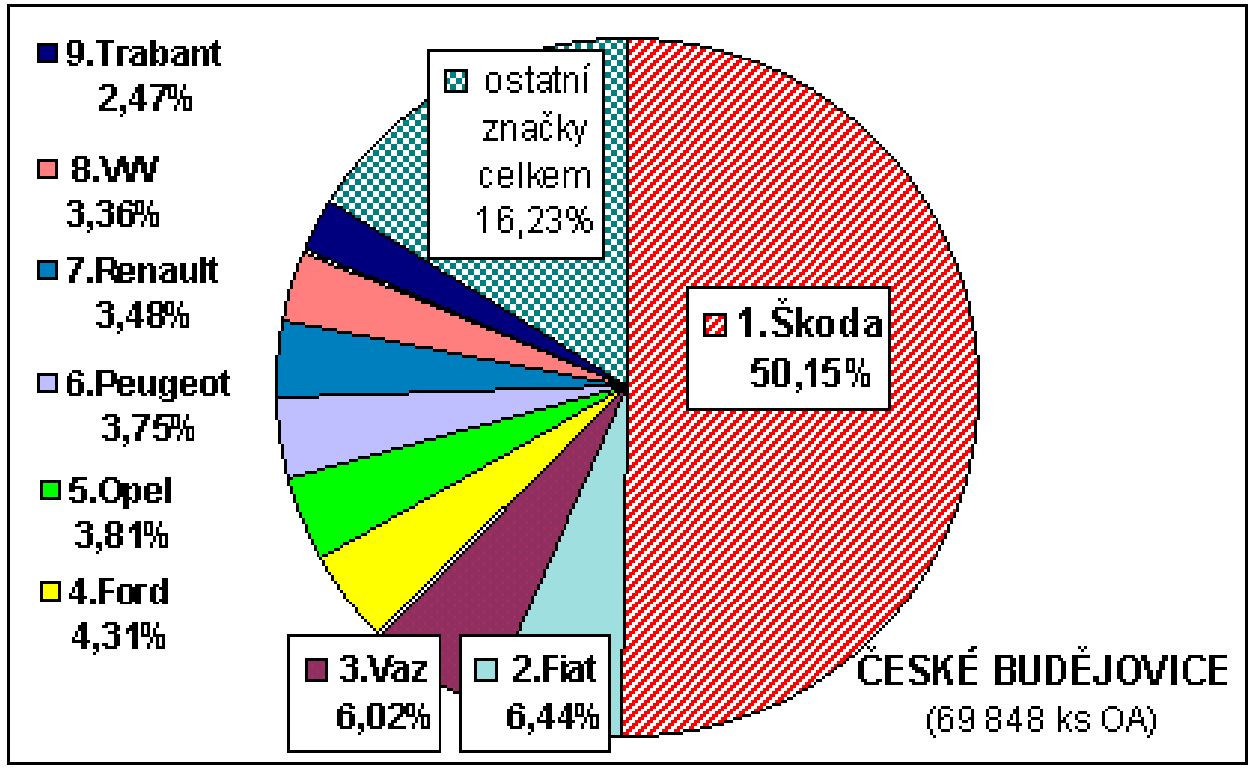
Cíle

- 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

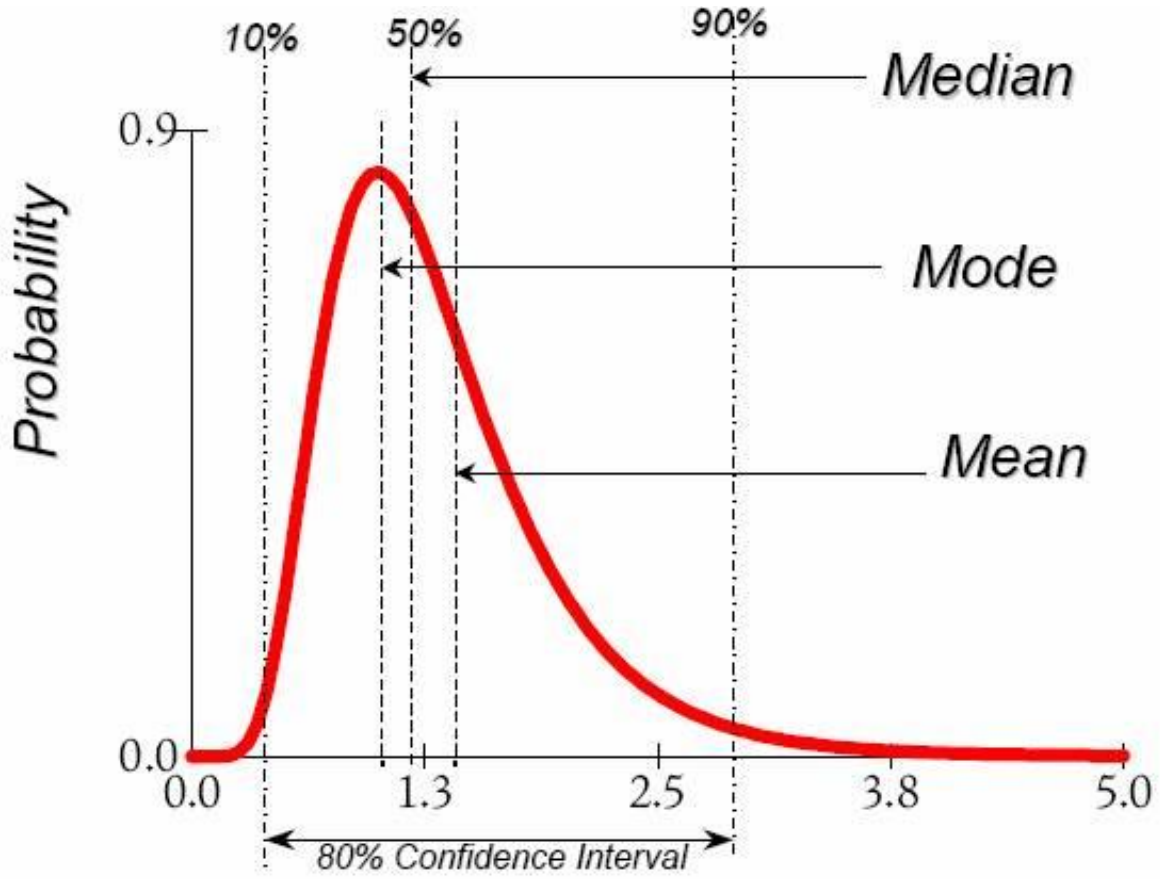


Modus, Medián a Aritmetický průměr atributu

- **Aritmetický průměr** (mean)
  - Statistická veličina, která v vyjadřuje typickou (očekávanou) hodnotu
  - Součet všech hodnot vydělený jejich počtem
- **Modus** atributu (mode)
  - *Nejčastější* hodnota v daném statistickém souboru
  - Hodnota znaku s největší relativní četností
  - Určení předpokládá roztřídění souboru podle *obměn* znaku.
- **Medián** (median)
  - Dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny
  - Pro sudý počet prvků je to aritmetický průměr dvou prostředních hodnot
  - 50% hodnot je menších nebo rovných a 50% je větších nebo rovných mediánu

Modus, Medián a Aritmetický průměr atributu

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



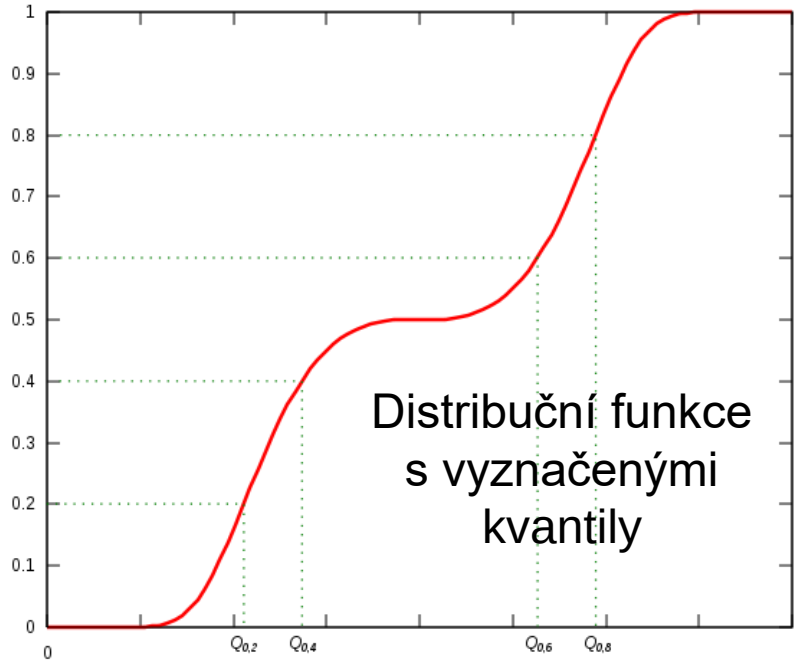
Příklad

- 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

Kvantily

- Dělí soubor seřazených hodnot na několik stejně velkých částí.
- **Medián** - kvantil $Q_{0,5}$.
  - Kvantil rozdělující statistický soubor na dvě stejně početné
- **Kvartil (**rozdělují statistický soubor na čtvrtiny.)
  - 25 % prvků má hodnoty menší než dolní kvartil $Q_{0,25}$ a
  - 75 % prvků hodnoty menší než horní kvartil $Q_{0,75}$
- **Percentil**
  - Percentil 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

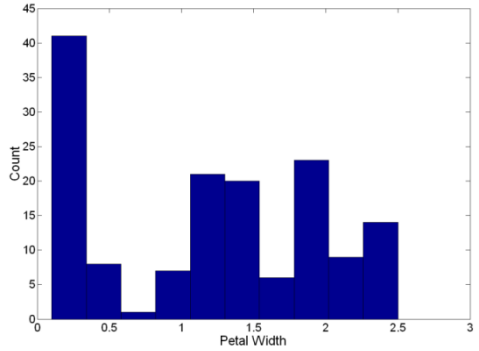
Vizualizace

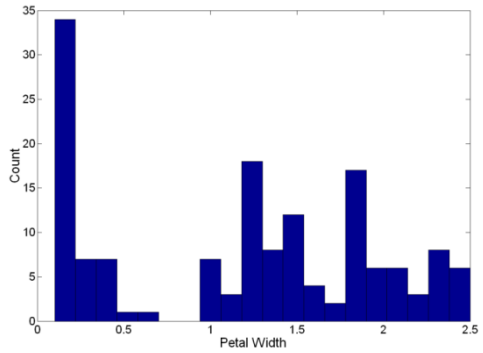
- Převedení dat do vizuální či tabulkové podoby pro potřeby analýzy dat
- Velmi silným nástrojem pro **průzkumovou analýzu** dat (angl. *exploratory analysis*)
  - Lidé mají velkou schopnost analyzovat velké množství dat prezentované vizuálně
  - Je možné identifikovat obecné trendy a struktury
  - Je možné identifikovat obecné vychýlené hodnoty (outliery)

- Techniky:
  - Histogram
  - Box plot
  - Korelační diagram

Vizualizační techniky: Histogram

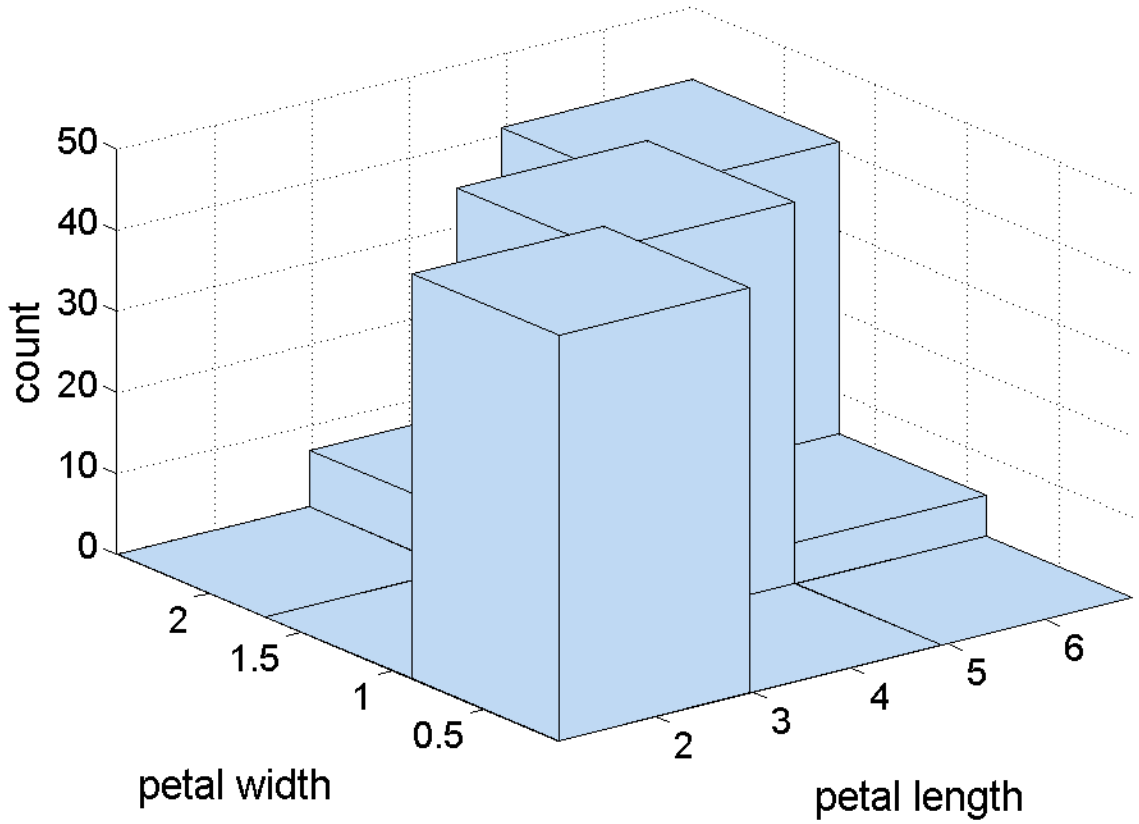
- Histogram
  - Rozdělí hodnoty do intervalů a zobrazí jejich četnosti
  - Výška sloupce udává počet objektů v daném intervalu
  - Říká, zda je soubor homogenní, nebo zda se rozpadá do dílčích menších podsouborů
    - jen jedna nejčetnější hodnota (homogenní soubor)
    - více hodnot s většími četnostmi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





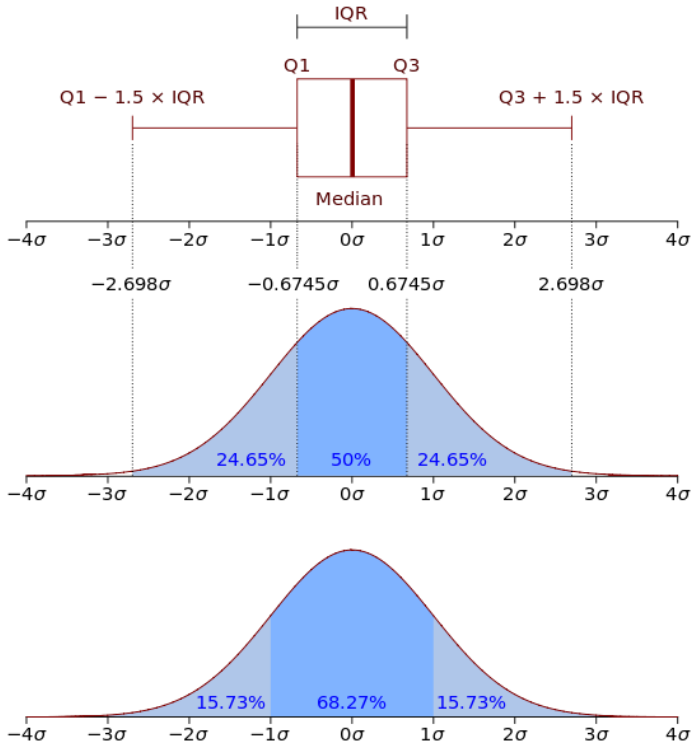
Dvoudimenzionální Histogram

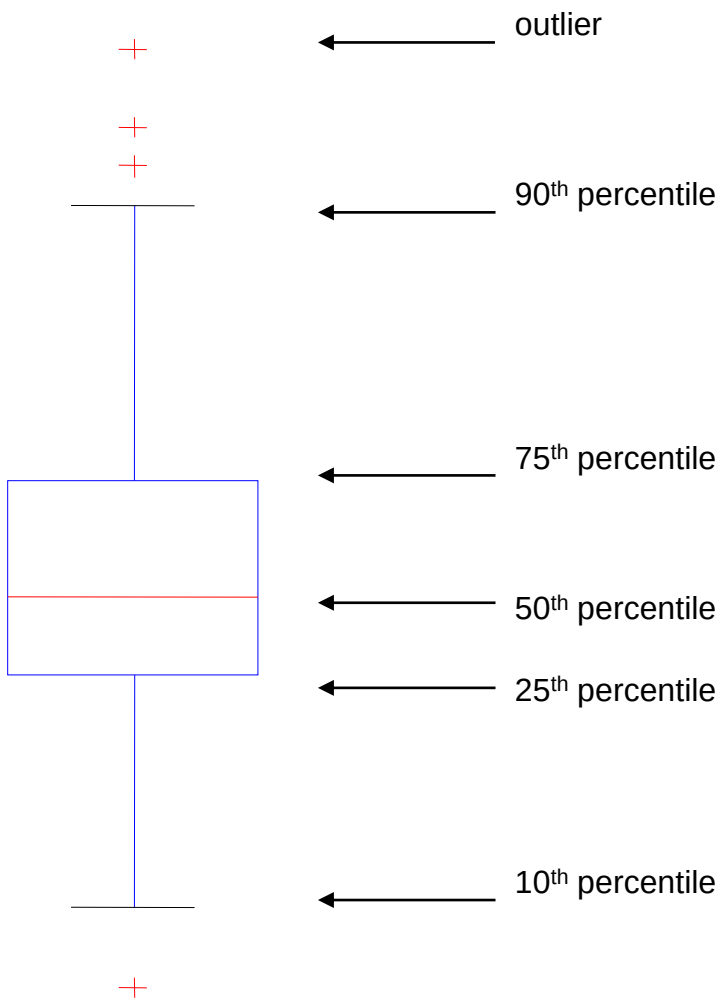
- Zobrazuje společné rozdělení dvou atributů



Vizualizační techniky: Krabicový graf

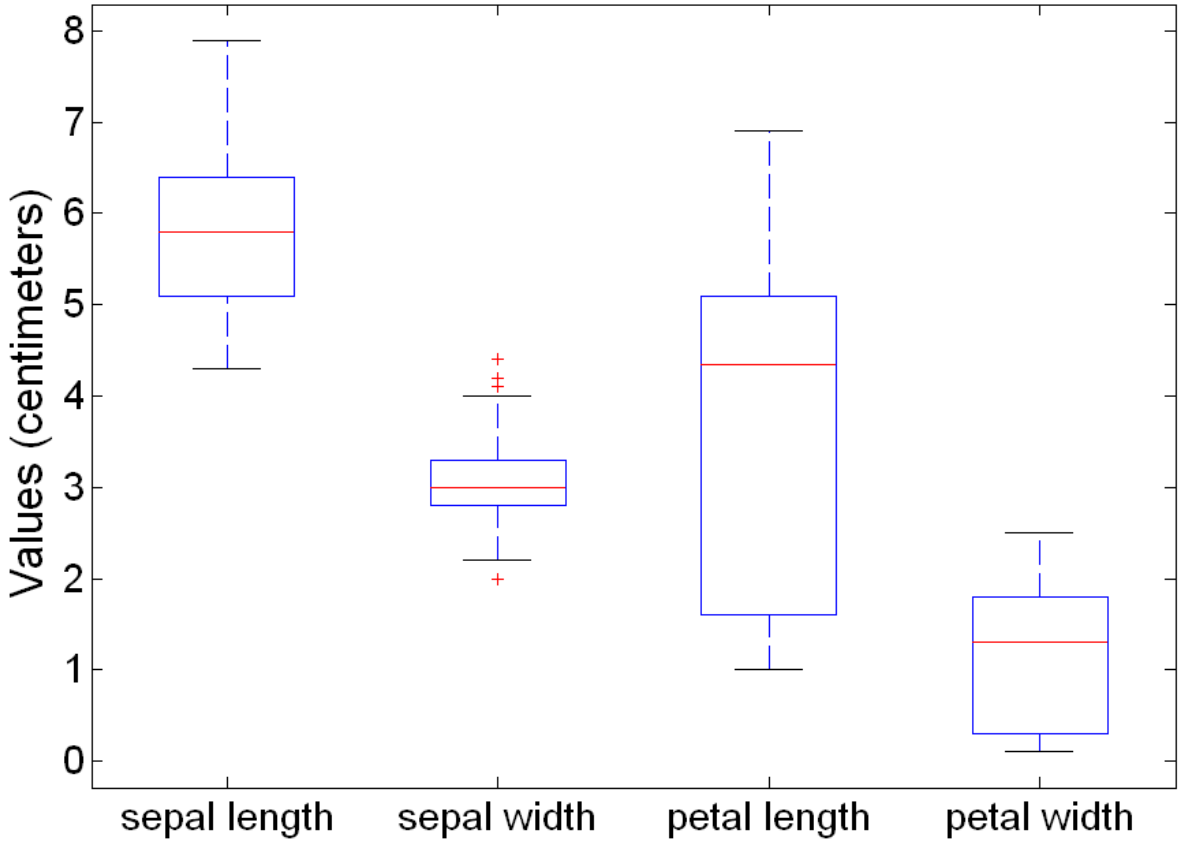
- Krabicový graf (krabicový diagram, box plot)
  - grafické zobrazení tzv. 5číselného souhrnu
  - autorem je statistik J. Tukey





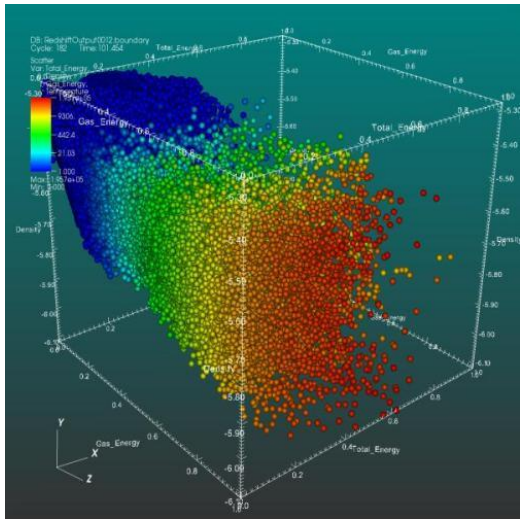
Příklad: Krabicový graf

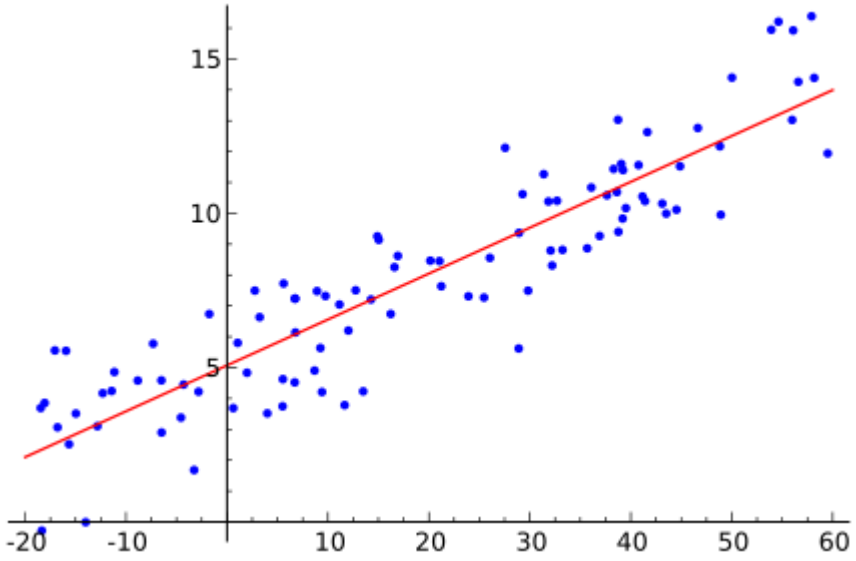
- Krabicové grafy se využívají k porovnání atributů



Vizualizační techniky: Korelační diagram

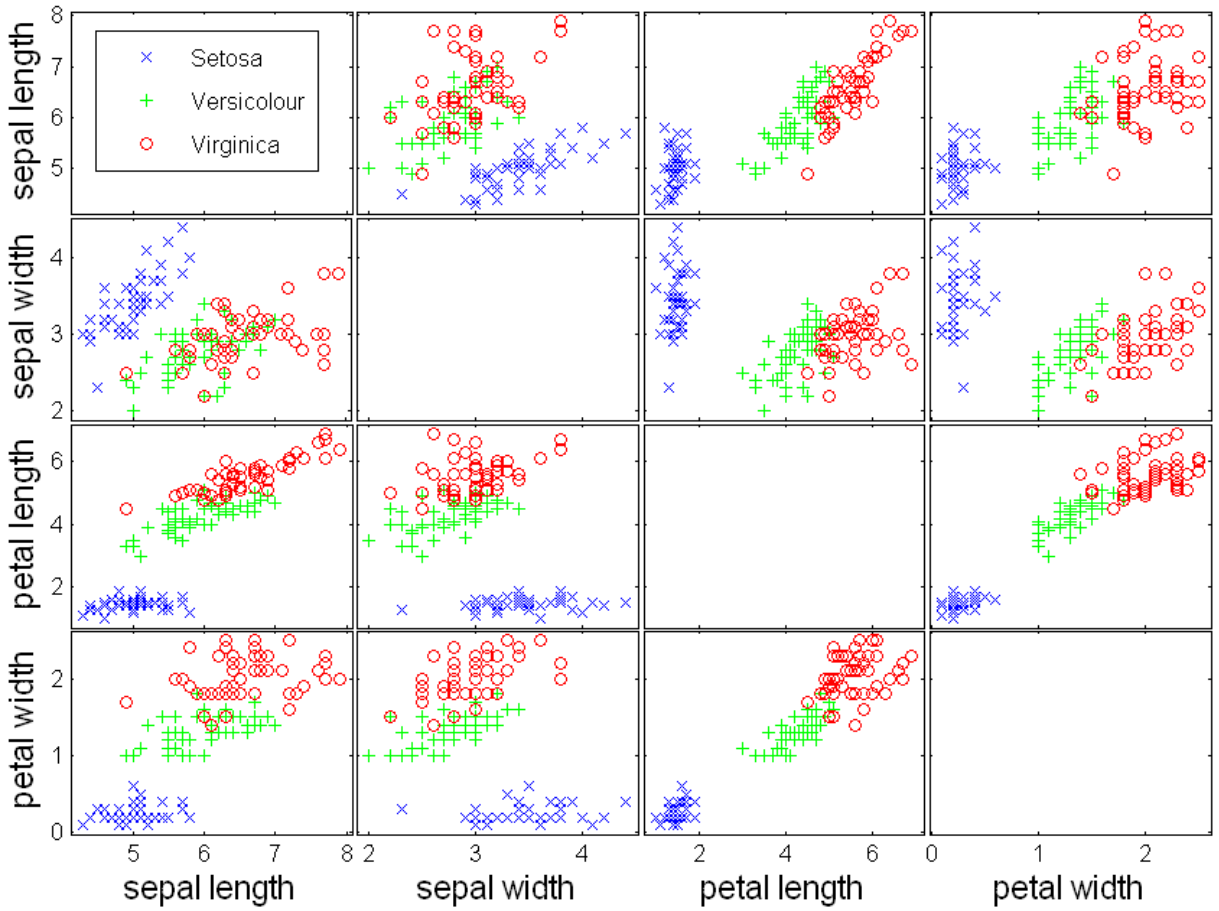
- též bodový graf (angl. *scatter plot*)
- matematické schéma užívající kartézských souřadnic pro zobrazení souboru dat o dvou (či tří) proměnných (na osy).
- je z něj možné jednoduše zjistit vzájemný vztah (korelaci) mezi oběma proměnnými





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

- **Datový objekt**
  - záznam v databázi, instance, vzorek, entita, ...
  - příklad: vozidlo, pacient, respondent
  - popsán kolekcí atributů

- **Atribut** (někdy také **příznak**)
  - vlastnost či charakteristika objektu
  - příklad: barva vozidla, objem motoru, a další

Atributy

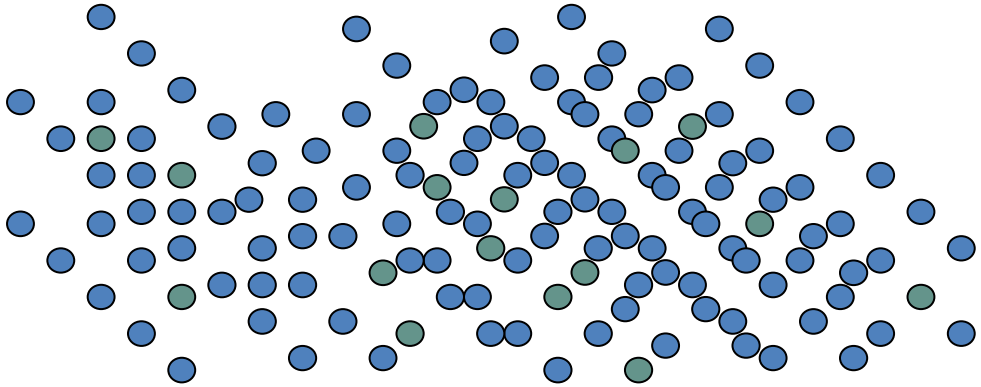
Objekty

ID	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

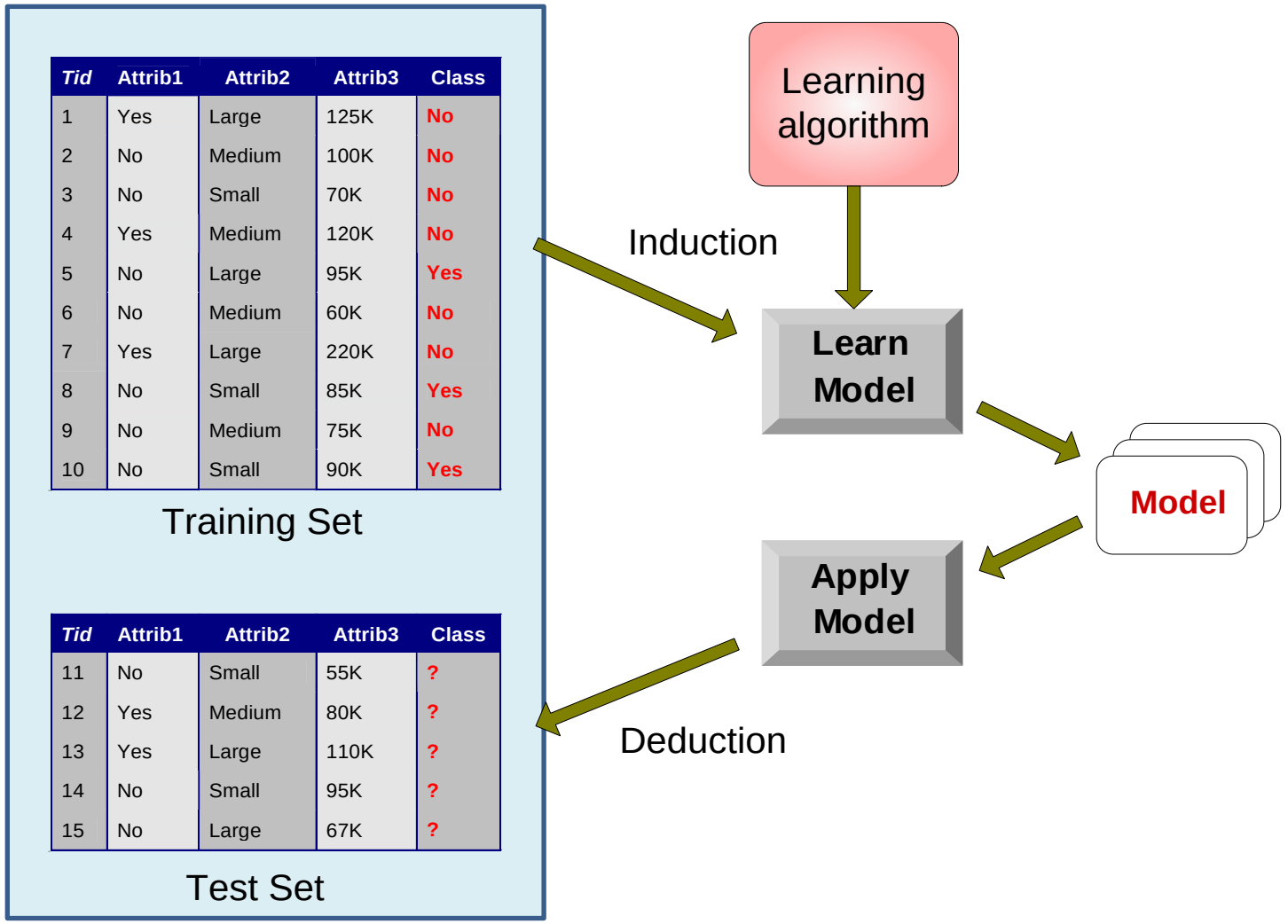
- **Základní soubor** (populace) - všechny jednotky
  - příklad: všichni řidiči v ČR
  - parametry označujeme písmeny řecké abecedy ($\mu, \sigma, \epsilon, \ldots$)
- **Výběrový soubor** – vybrané jednotky, náhodný výběr
  - např. všichni řidiči, kteří v konkrétním dni jeli autem a stali se účastníky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

Zdroj http://www.nedarc.org/

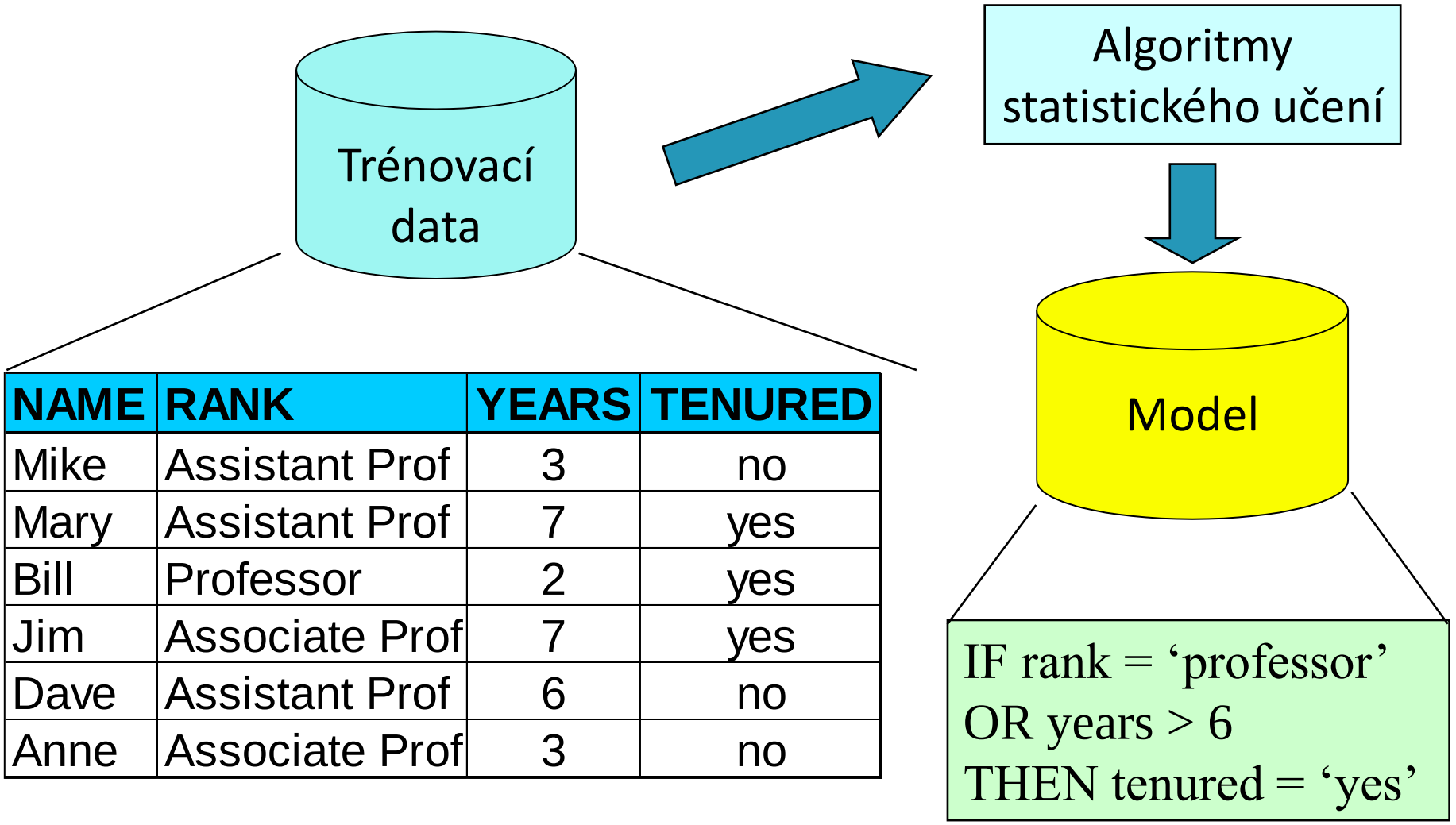


Aplikace statistického modelu

Data můžeme dělit na trénovací a testovací sadu, lepší je ale **křížová validace**:

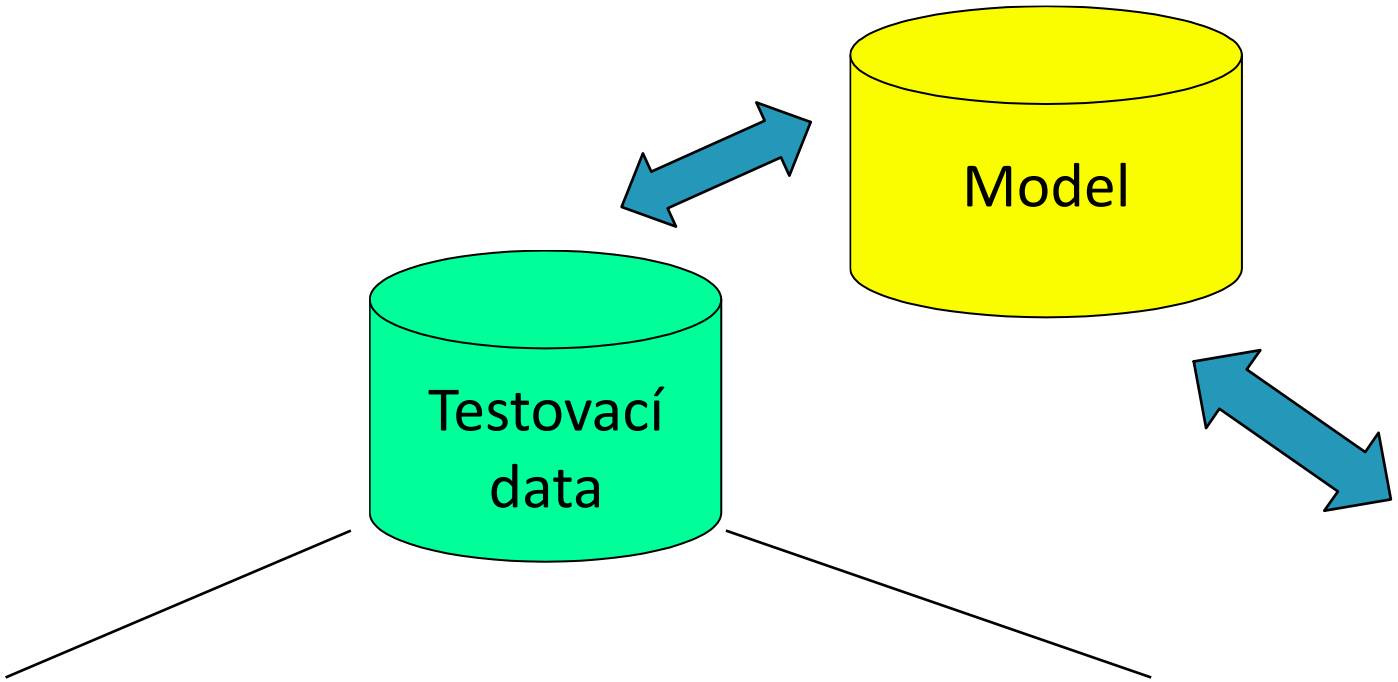


Indukce



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

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 rychlosti algoritmu!)
- **Matice záměn** (confusion matrix):
  - a je počet správných predikcí, že daná instance je pozitivní (*true positive*, TP)
  - b je počet nesprávných predikcí, že daná instance je negativní (*false negative*, FN)
  - c je počet nesprávných predikcí, že daná instance je pozitivní (*false positive*, FP)
  - d je počet správných predikcí, že daná instance je pozitivní (*true negative*, TN)

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

Metriky pro porovnání metod: Správnost

- Správnost (*accuracy*, A) – míra správné klasifikace

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

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

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metriky pro porovnání metod: Přesnost

- Přesnost (*precision*, P) – míra exaktnosti modelu

$$\text{Precision} = \frac{a}{a+c} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

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

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metriky pro porovnání metod: Přesnost

- Úplnost (*recall*, R) – míra kompletnosti modelu

$$\text{Recall} = \frac{a}{a+b} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

		PREDIKOVANÁ TŘÍDA	
		Třída=Ano	Třída=Ne
SKUTEČNÁ TŘÍDA	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 metrikou A (správnost)

- Máme problém s dvěma nevyváženými třídami
  - Počet vzorků pro třídu 0 = 9990
  - Počet vzorků pro třídu 1 = 10

- Pokud model predikuje vše do třídy 0, přesnost je
  - 9990/10000 = 99,9 %

Cost Matrix (cena)

		PREDIKOVANÁ TŘÍDA	
		Třída=Ano	Třída=Ne
SKUTEČNÁ TŘÍDA	Třída=Ano	$C(\text{Ano}\|\text{Ano})$	$C(\text{Ne}\|\text{Ano})$
	Třída=Ne	$C(\text{Ano}\|\text{Ne})$	$C(\text{Ne}\|\text{Ne})$

- $C(y|x)$: Cena za klasifikaci vzorku z třídy x do třídy y
- Lze využít v případě, kdy model poskytuje pravděpodobnost Ano/Ne k modifikaci rozhodnutí
- Lze využít k odstranění vlivu nevyvážených vstupních dat

Výpočet ceny klasifikace

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

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	60	40
	-	80	320

A = 76%

*(60+320)/(60+320+60+40)*100*

Cost = 4080

*60*0+40*100+80*1+320*0*

Model M_2	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	55	45
	-	5	395

A = 90%

(55+395)/(55+395+5+45)*100

Cost = 4255

(55*0+45*100+5*1+395*0)

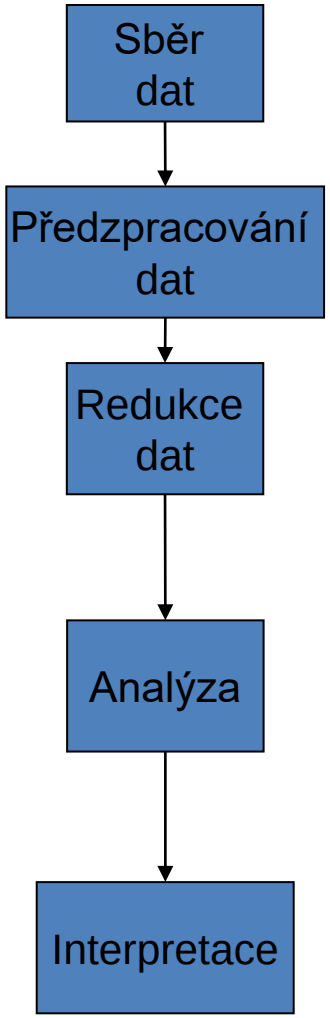
Obsah prezentace

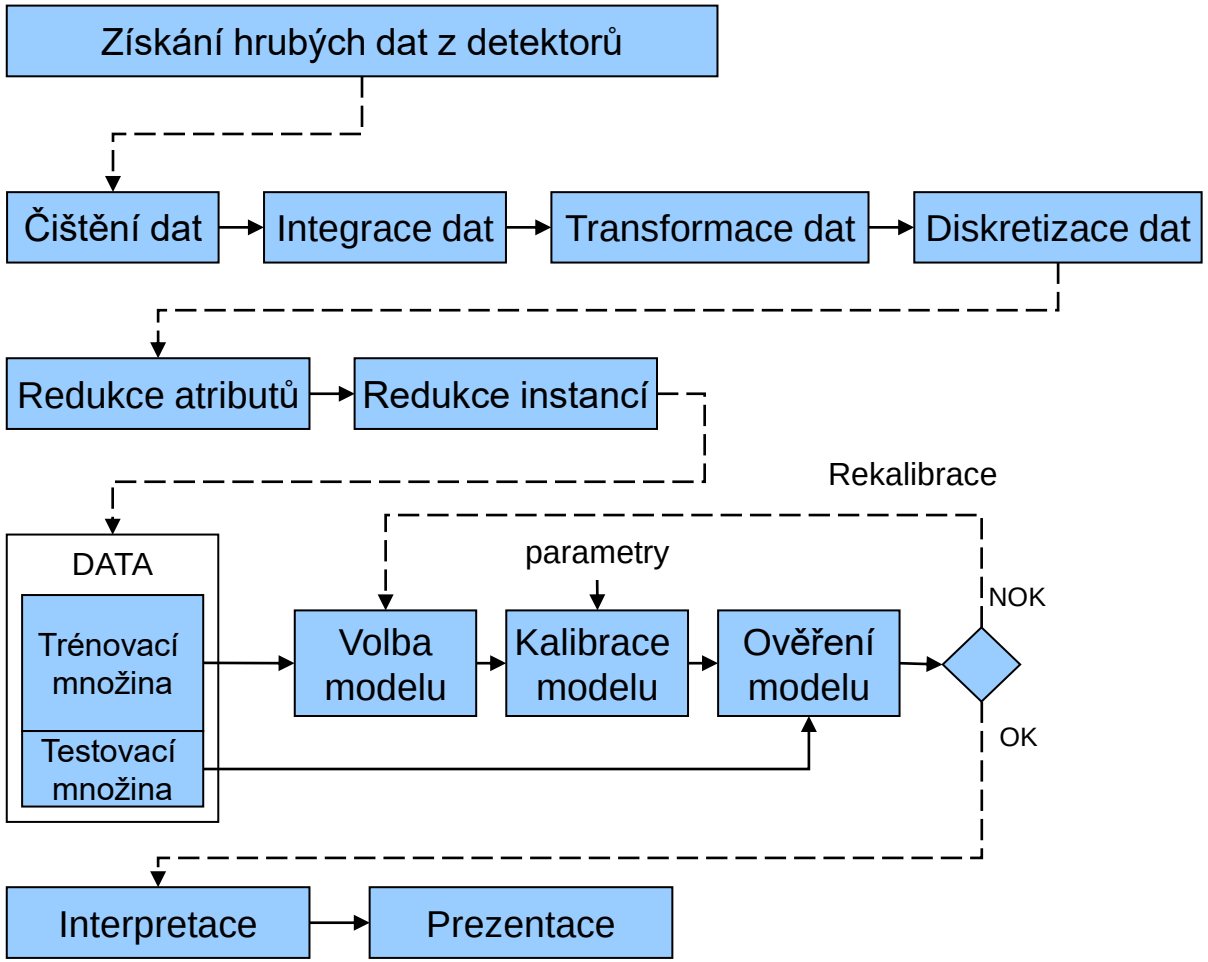
- Měřené veličiny
- Chyby měření
- Vizualizace dat
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

Diskuze

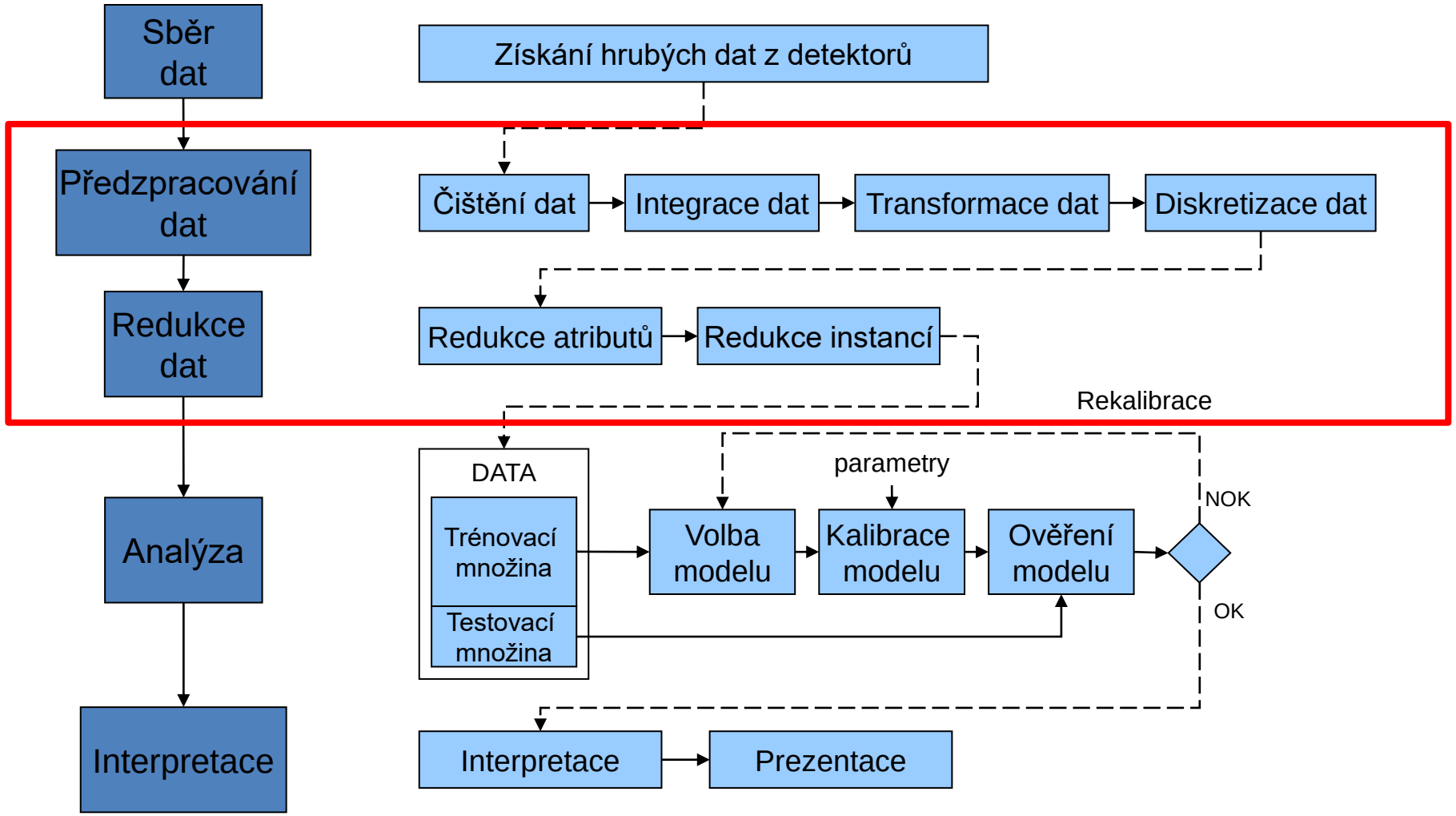
- Co je třeba udělat s daty před aplikací vlastního matematického modelu?
  - Uveďte na příkladech

Hlavní kroky





Hlavní kroky



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

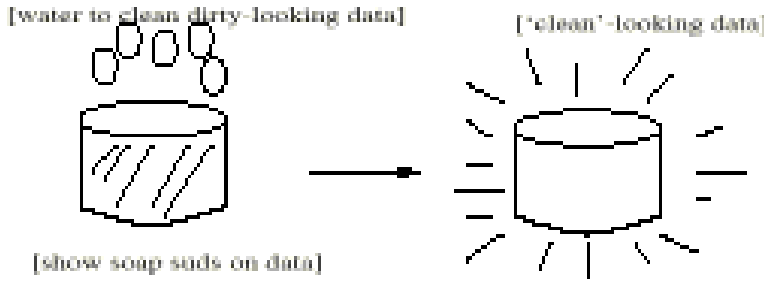
Motto: Pokud nejsou kvalitní data, nebudou kvalitní ani výsledky analýzy!

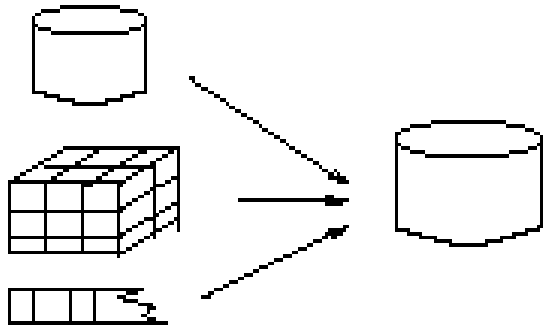
Jak určit kvalitu dat? (Metriky kvality dat)

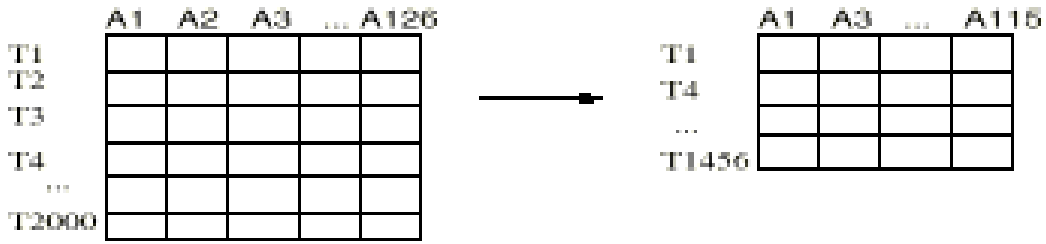
- **Přesnost** (accuracy)
  - měřeno obvykle statistickými charakteristikami pro chybu, např. směrodatná odchylka
- **Úplnost** (completness)
  - zda statistické charakteristiky dat nejsou ovlivněny výběrovými efekty.
- **Konzistence** (consistency)
- **Včasnost** (timeliness)
  - za jakou dobu lze data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

Hlavní oblasti předzpracování dat

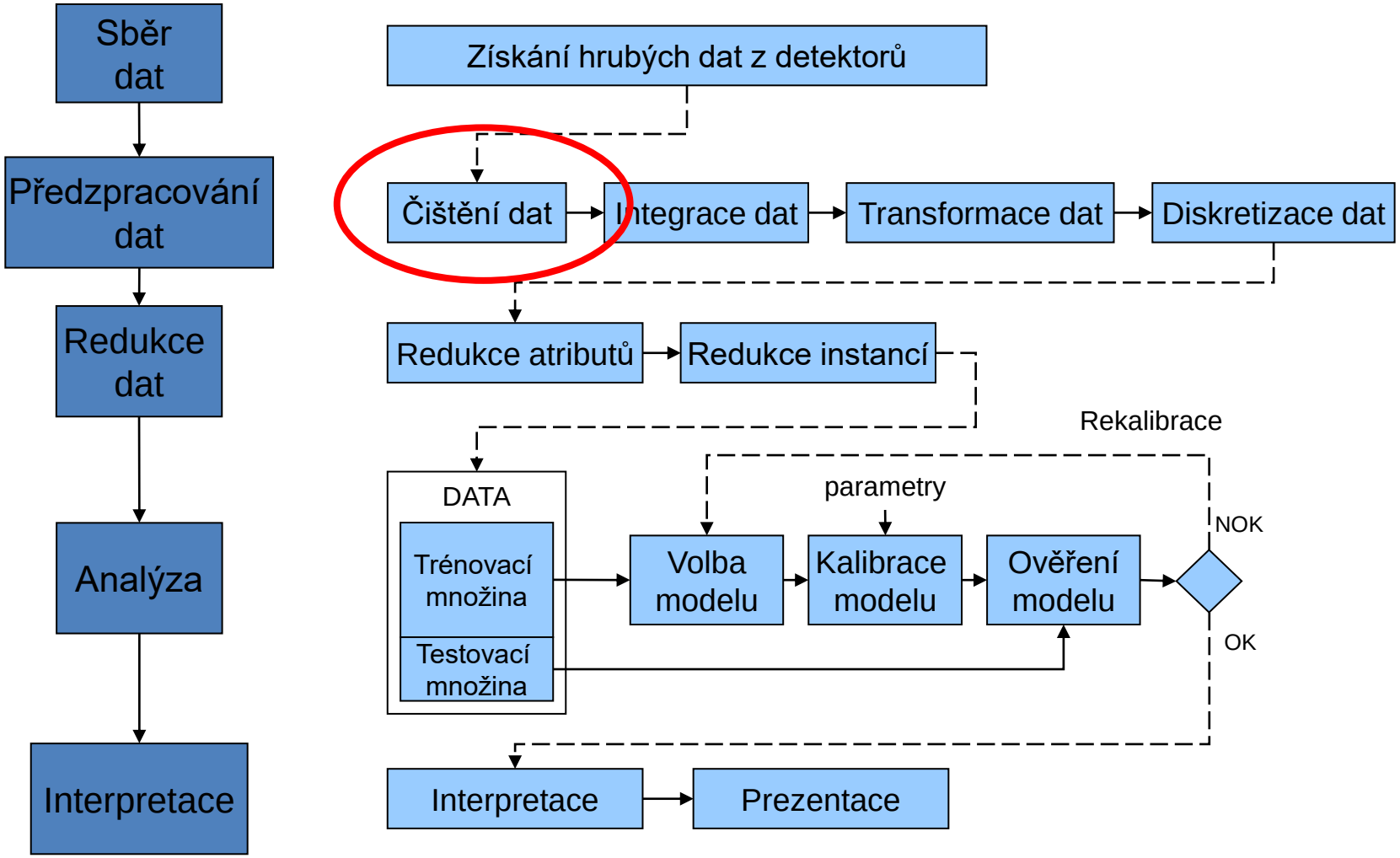
- Čištění dat
- Integrace dat z více zdrojů
- Transformace dat
- Redukce dat
- Diskretizace dat







Hlavní kroky



Cíle čištění dat

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

Činnosti:

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

- Vyhlaď zašuměná data

Rozdělení metod čištění dat

- Dle způsobu měření dat

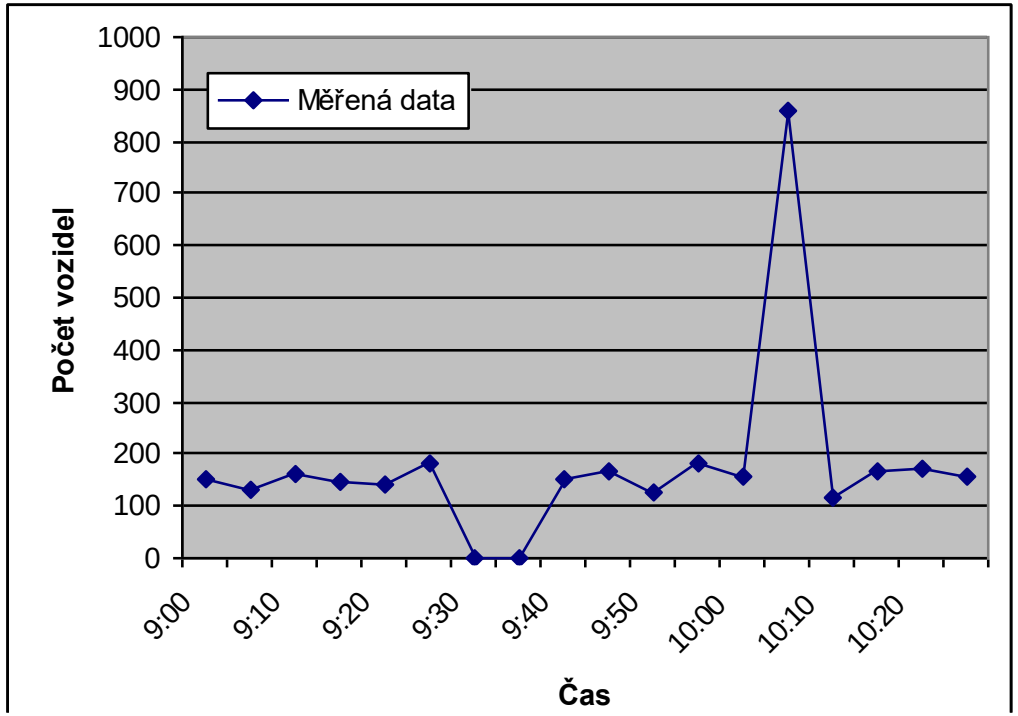
- A. Data v tabulární formě

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 v tabulární formě?

- 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 v tabulární formě?

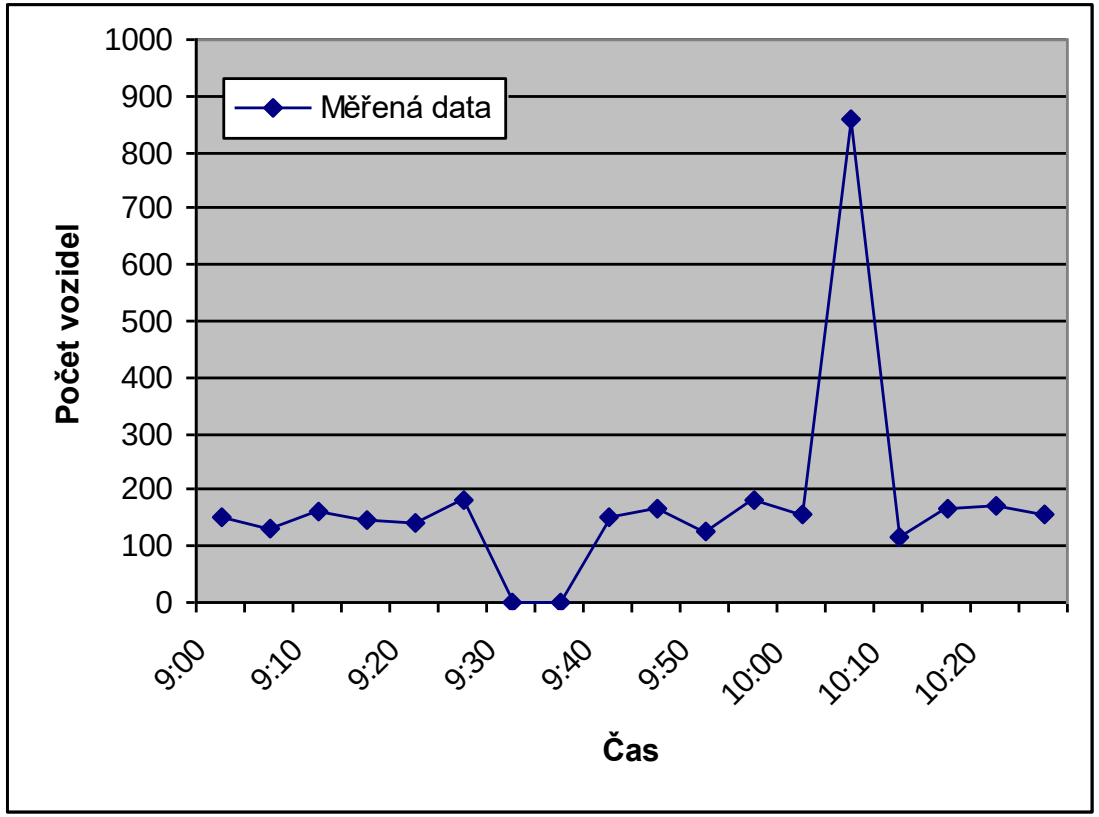
- Vyplnění střední hodnotou všech vzorků
- Střední hodnota pro dané atributy
- Pravděpodobnostní modely (regrese, rozhodovací stromy, …)

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

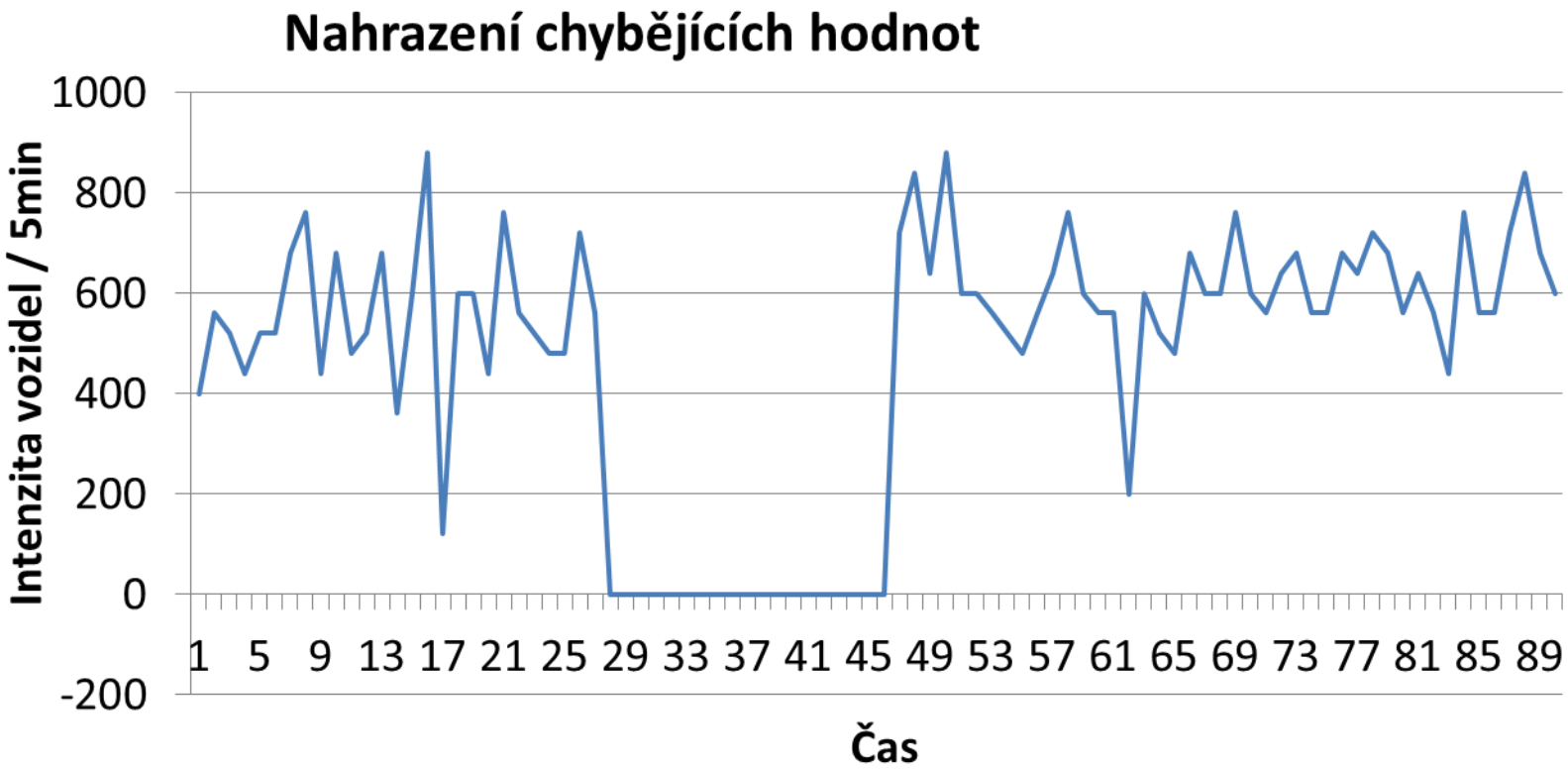
Jak opravit chybějící data - B. Časová řada?

- 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?



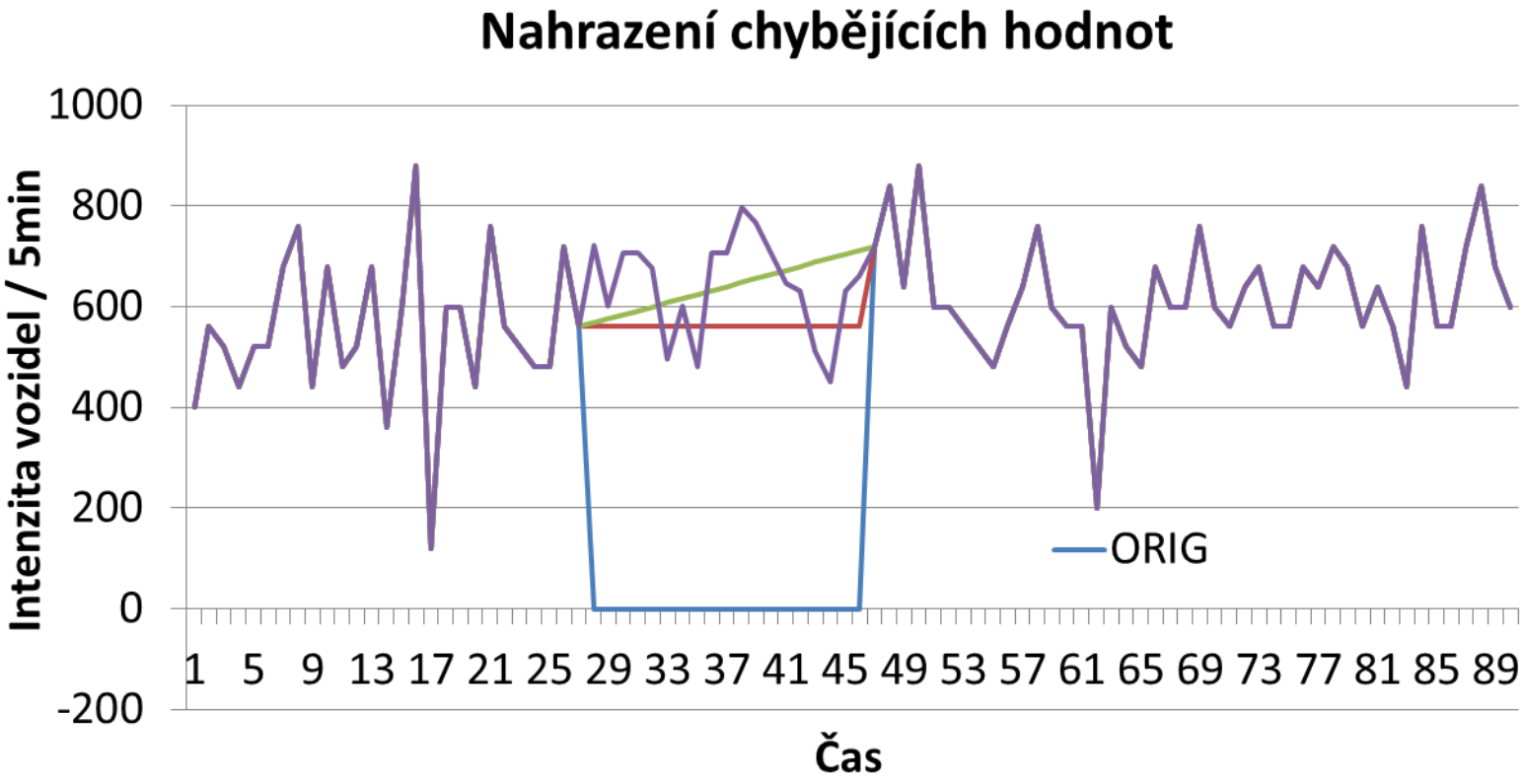
Chybějící hodnoty – časová řada

- 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

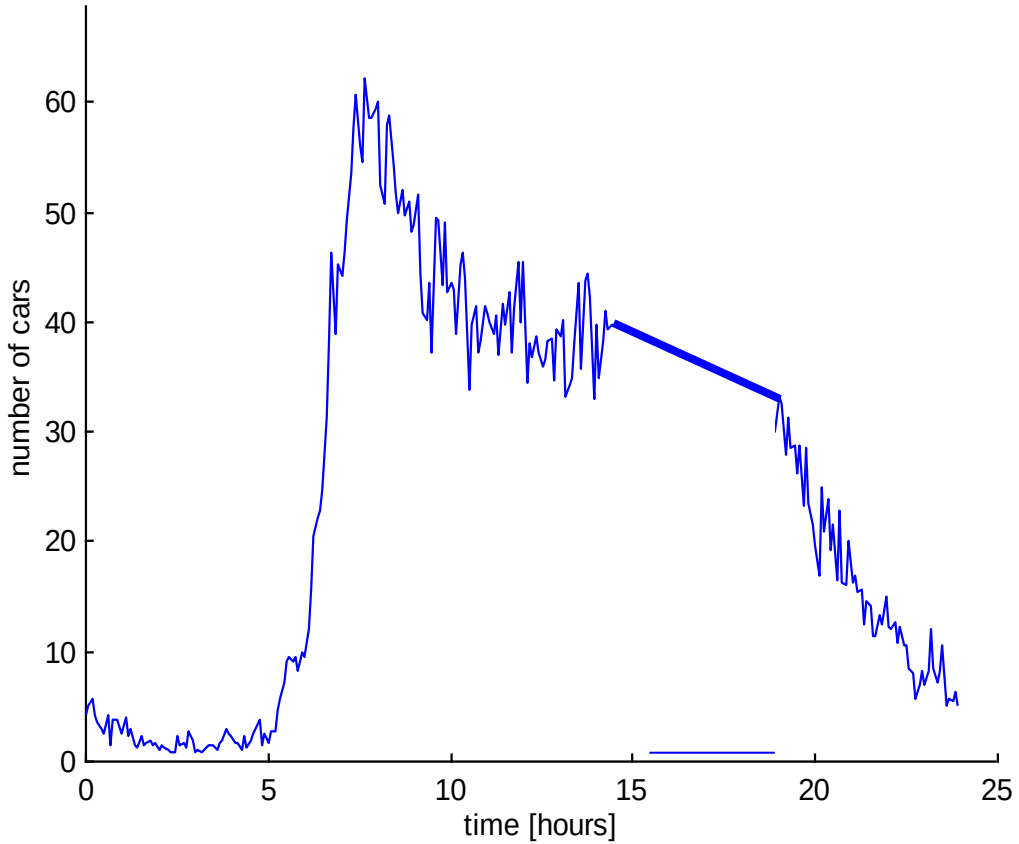
Jak opravit chybějící data - Řešení

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



Chybějící hodnoty – časová řada

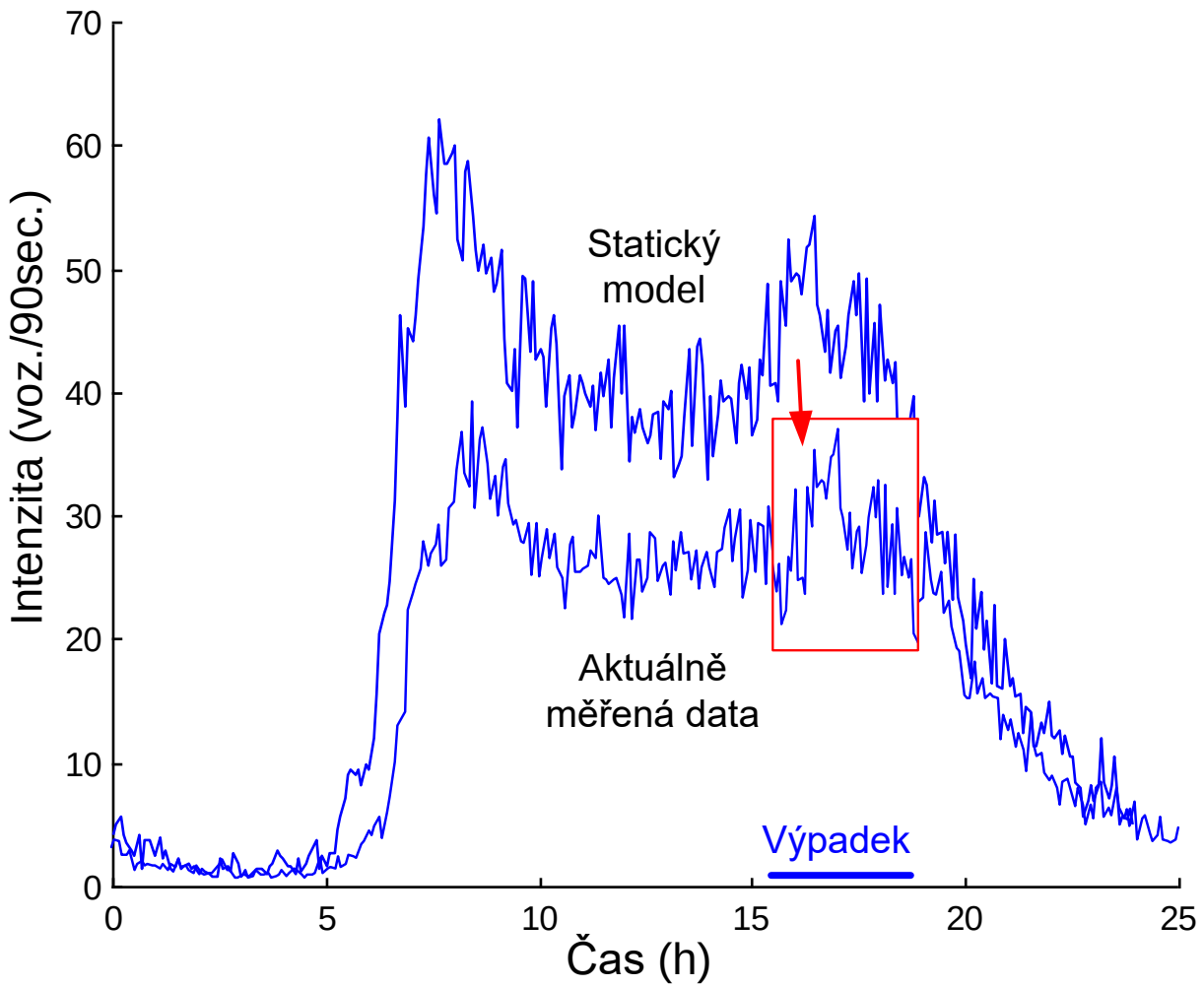
- Lineární spojnice platných hodnot
  - místo chybné hodnoty se lineální spojnice poslední platné hodnoty před výpadkem a první po výpadku.



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

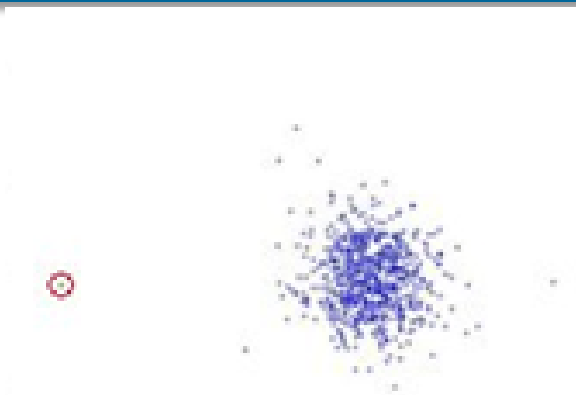


Diskuze

- Co je to vychýlená hodnota (outlier)?
- Jak spolehlivě vychýlené hodnoty identifikovat?
- Jak je odstranit?

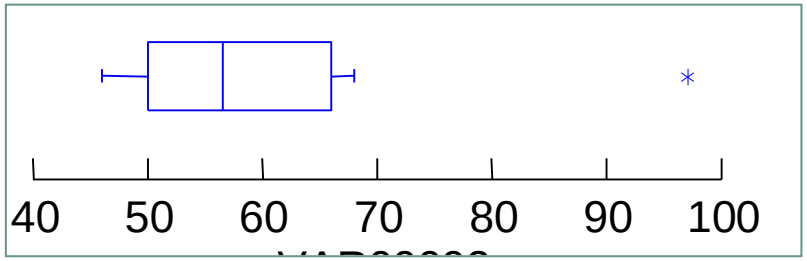
Detekce vychýlených hodnot

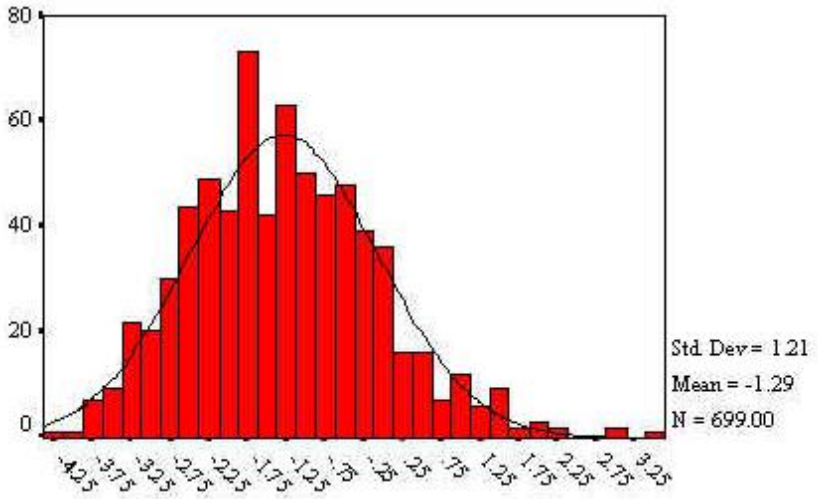
- Outliery je třeba identifikovat, popřípadě odstranit



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

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

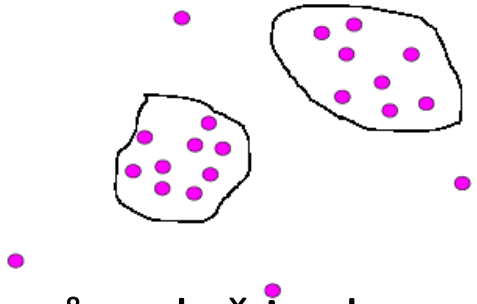




Detekce extrémních hodnot

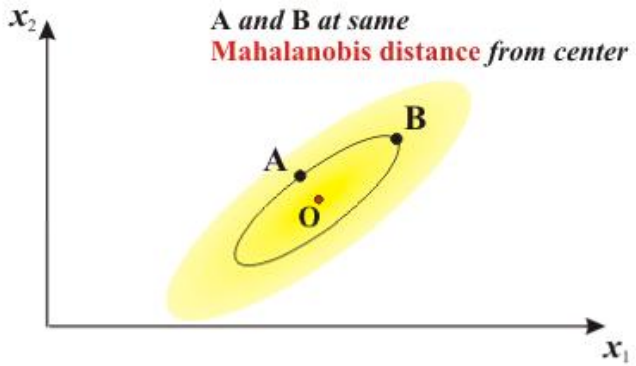
Metody – **vícerozměrné**

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

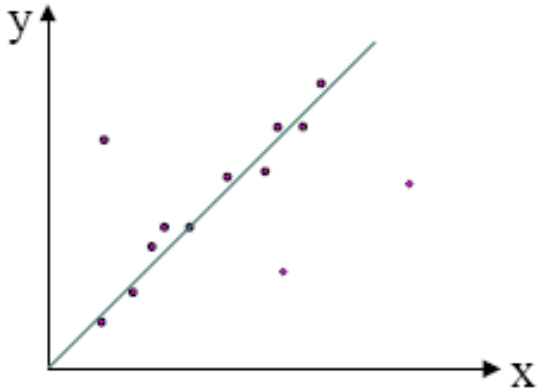


- Binning
  - Seřaď hodnoty atributů a sluč je do skupin (bins)
  - Vyhlaď je podle středních hodnot,

- Mahalanobisova vzdálenost
  - vzdálenost od centroidu



- Regrese
  - Najdi data ležící daleko od regresní funkce



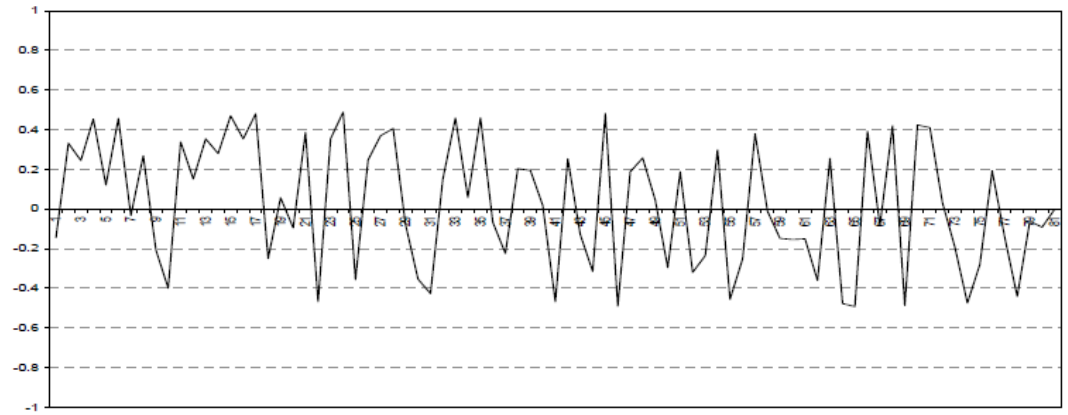
FILTRACE DAT

Diskuze

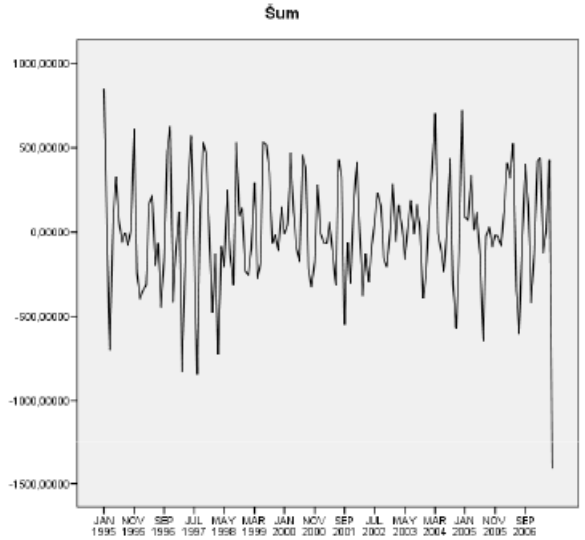
- Co je cílem filtrace?
- Co je to bílý šum (noise)?

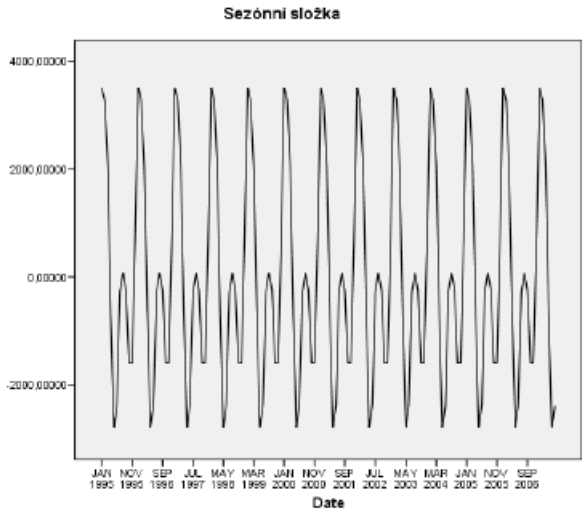
Šum analytického signálu

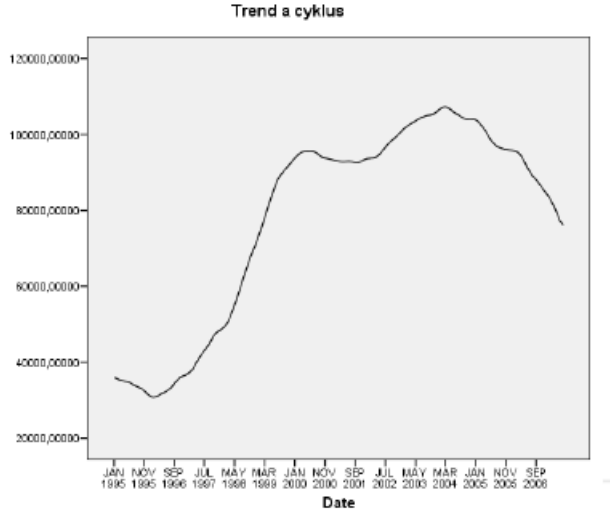
- náhodné zvýšení nebo snížení měřeného signálu
- šum, jehož suma je **nulová** v časovém intervalu pozorování, se označuje jako **bílý šum**
- šum, jehož suma je **nenulová** v časovém intervalu pozorování, se označuje jako náhodný šum
- šum je významný jen z hlediska intervalu pozorování
- intenzivní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

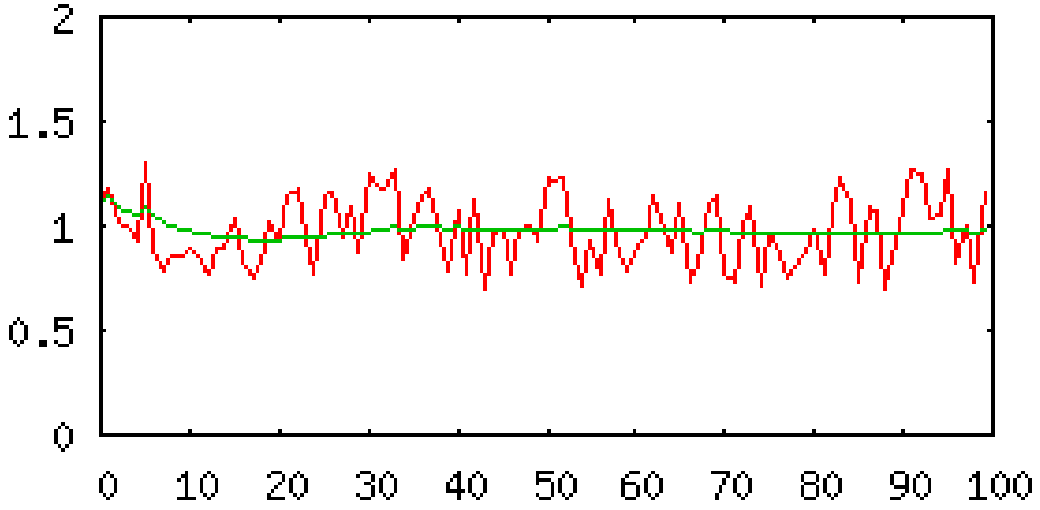
Filtrace dat v časové oblasti

- 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**

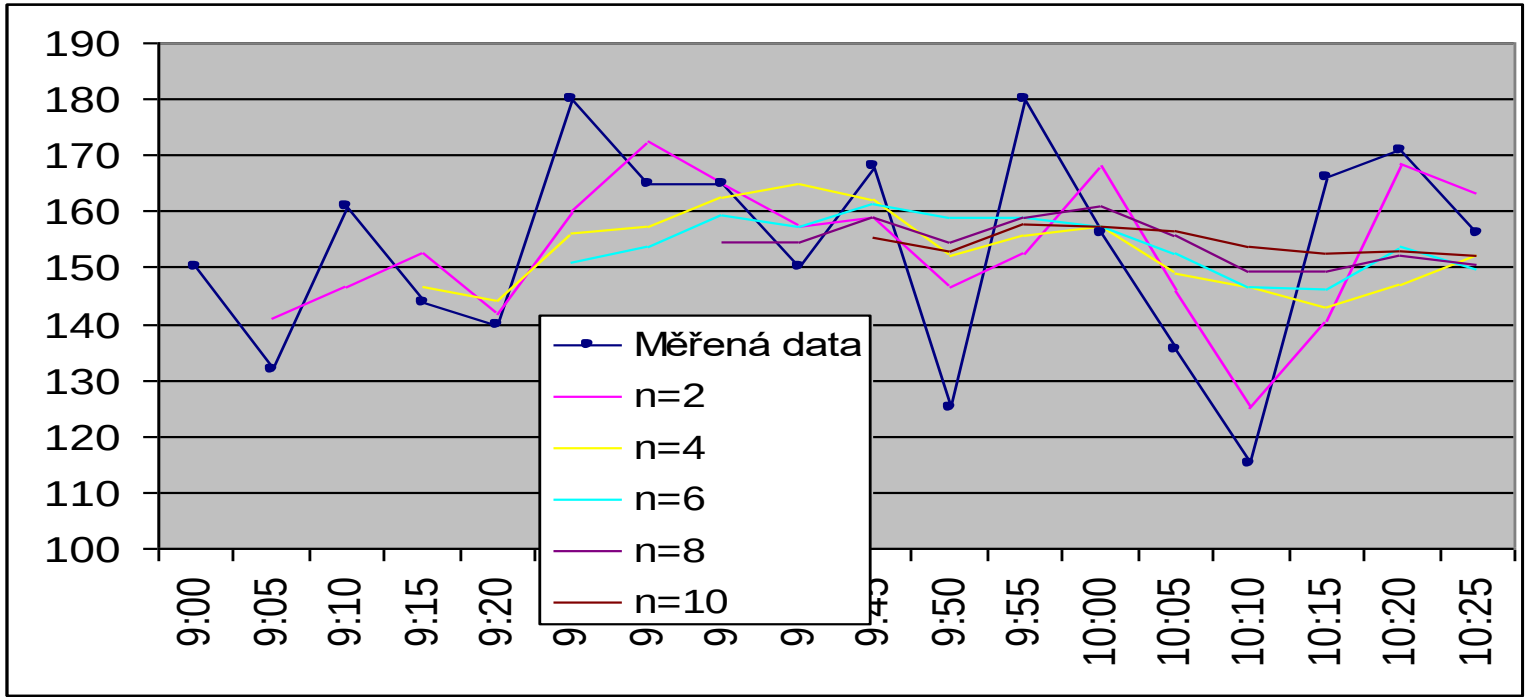


$$\hat{x}_n = \frac{\hat{x}_{n-1} \cdot (n-1) + x_n}{n}$$

Filtrace dat v časové oblasti

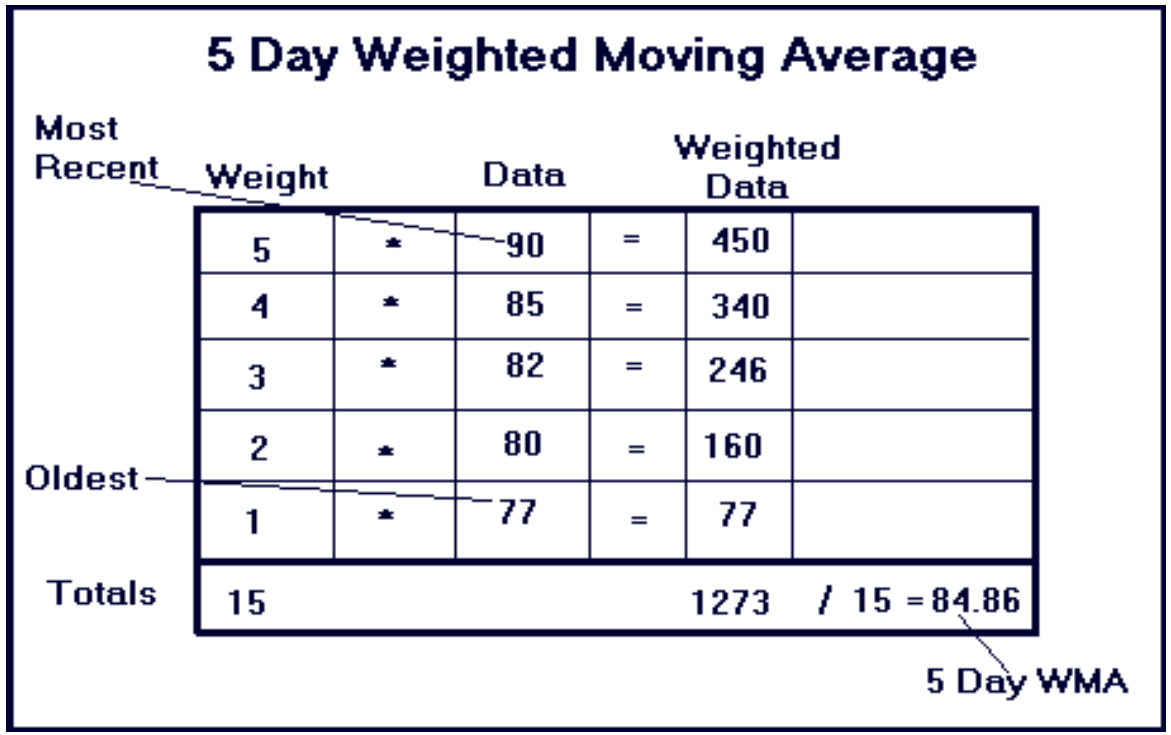
- **Plovoucí průměr** (moving average)
  - Pokud neexistuje periodický cyklus
  - Všechny hodnoty mají stejnou váhu
  - k … velikost ‚paměti‘

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



Filtrace dat v časové oblasti

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

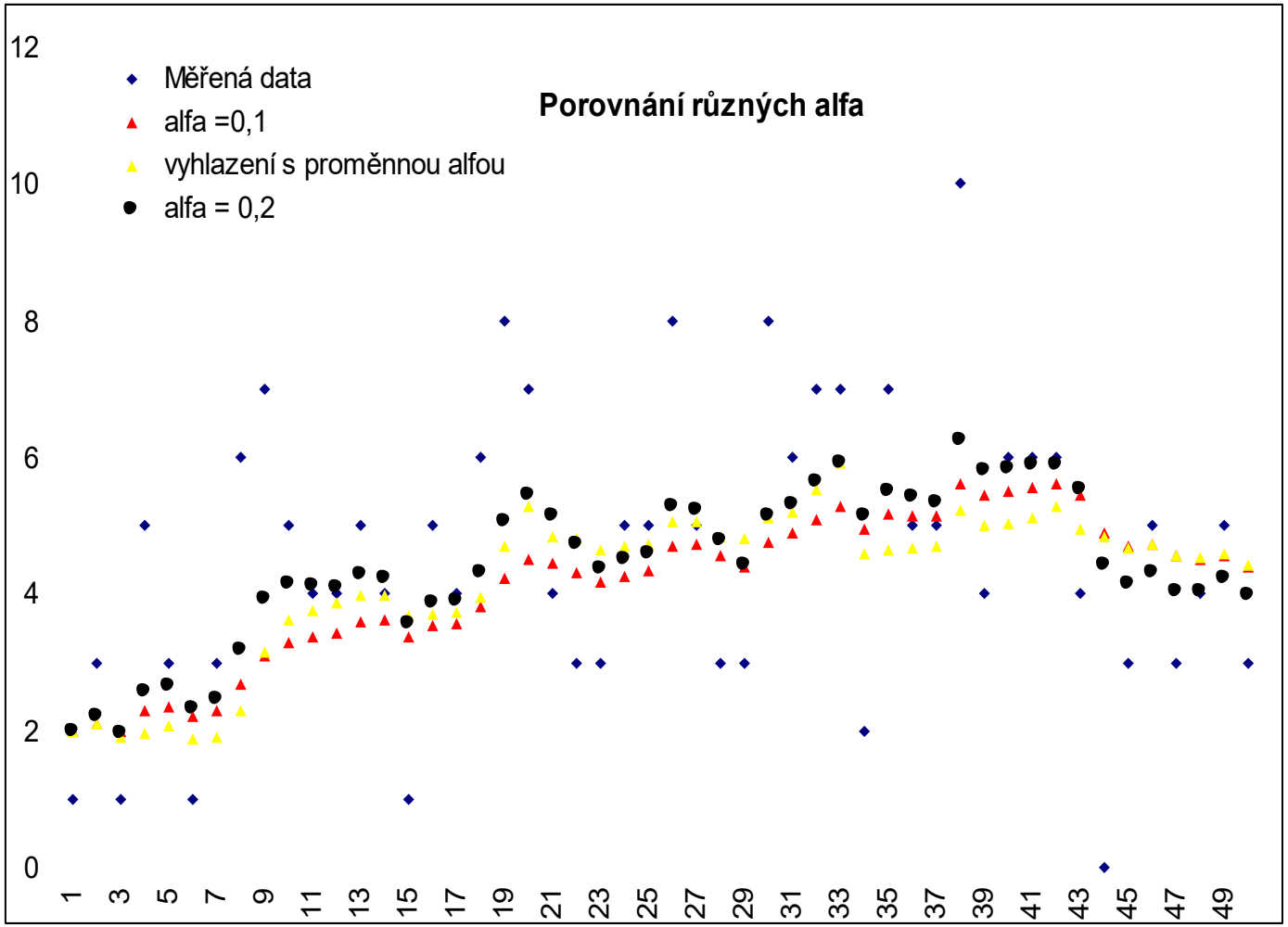


$$\hat{x}_n = \frac{\sum_{i=1}^{k} w_i \cdot x_{(n-i)+1}}{\sum_{i=1}^{k} w_i}$$

Filtrace dat v časové oblasti

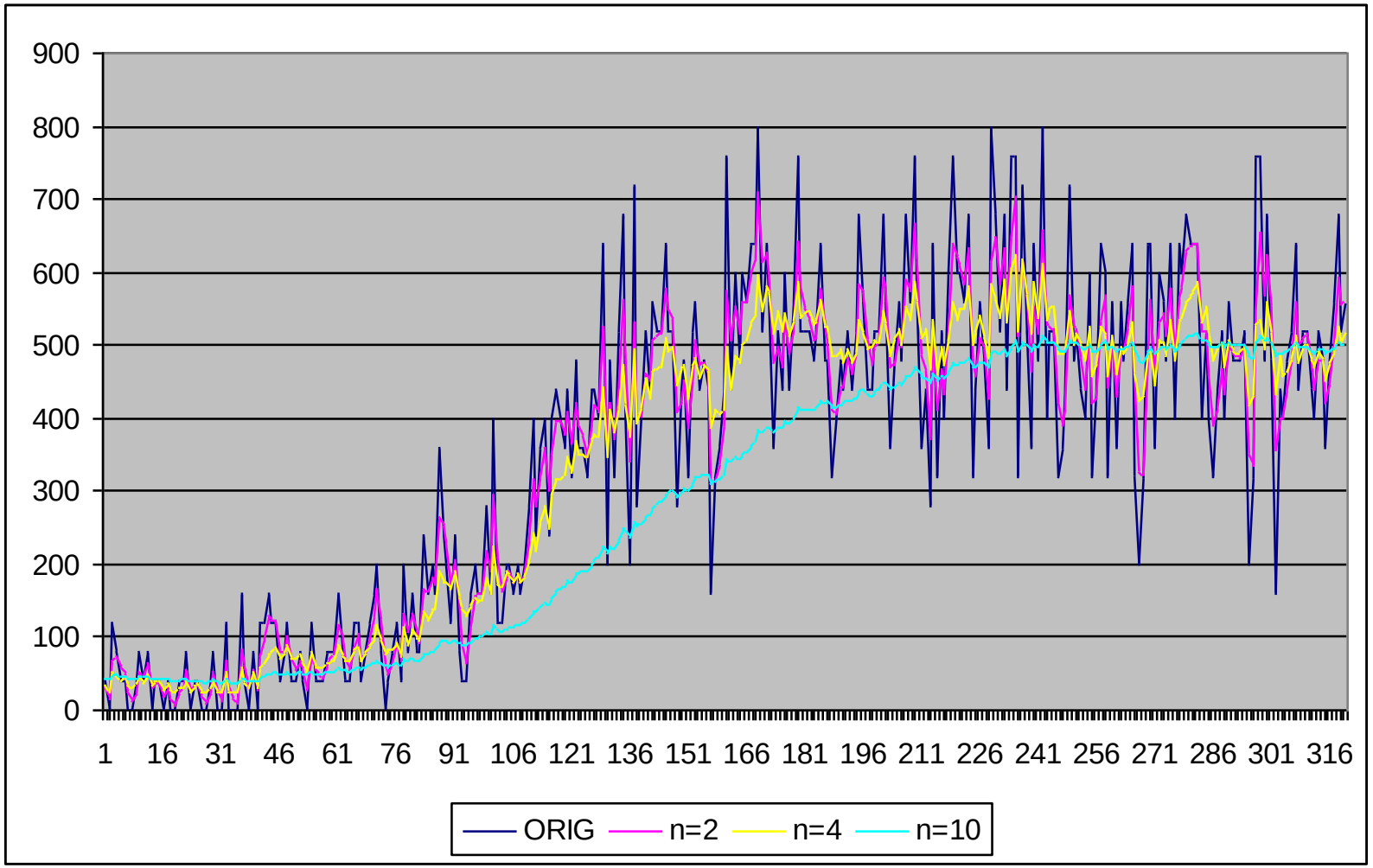
- Exponenciální vyhlazování (exponential smoothing)
  - s pevnou hodnotou α
  - s proměnnou hodnotou α

$$\bar{v}_1 = v_1$$
$$\bar{v}_t = \bar{v}_{t-1} + \alpha(v_t - \bar{v}_{t-1})$$



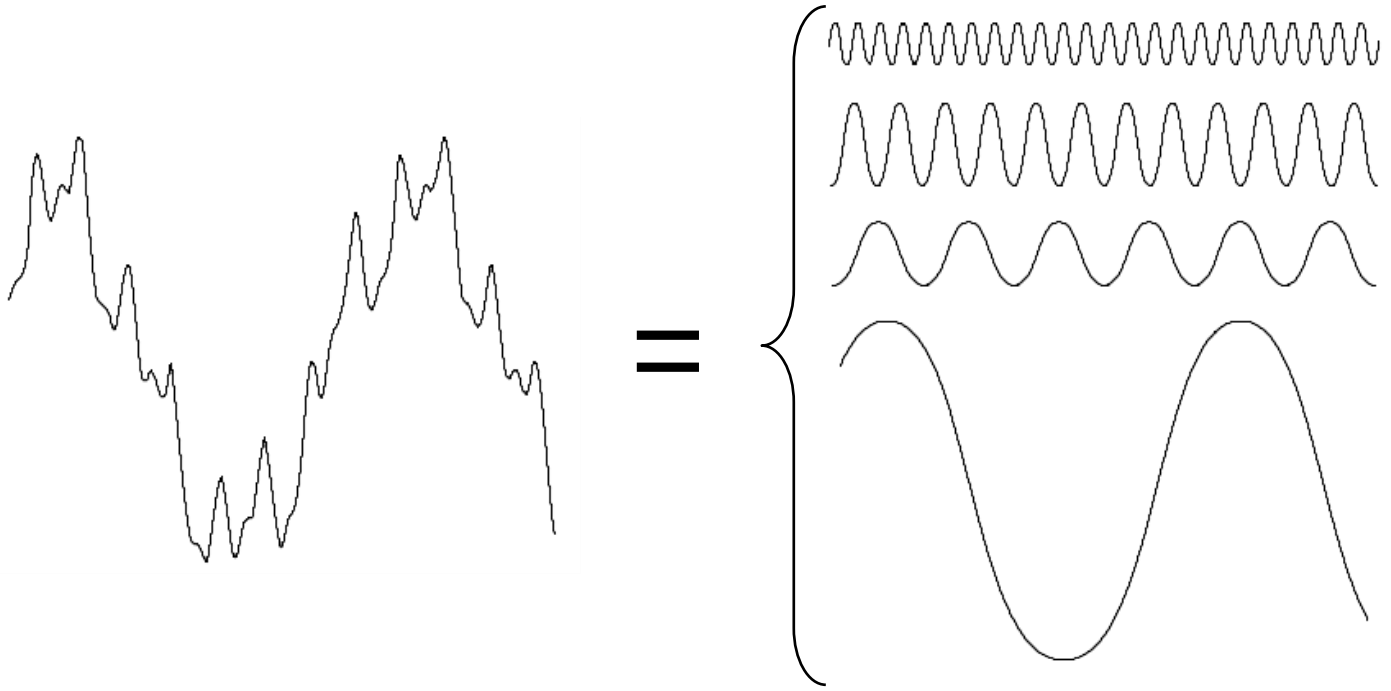
Diskuze

- 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

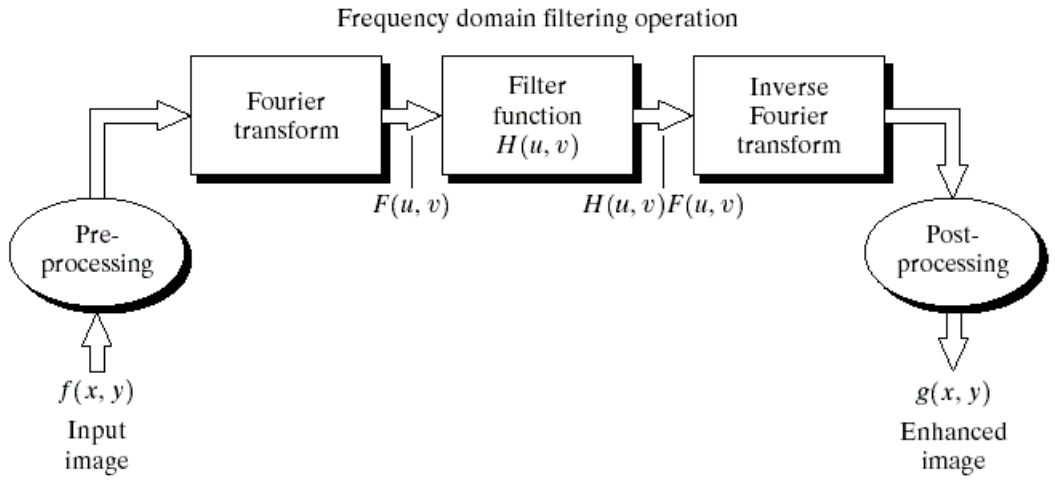
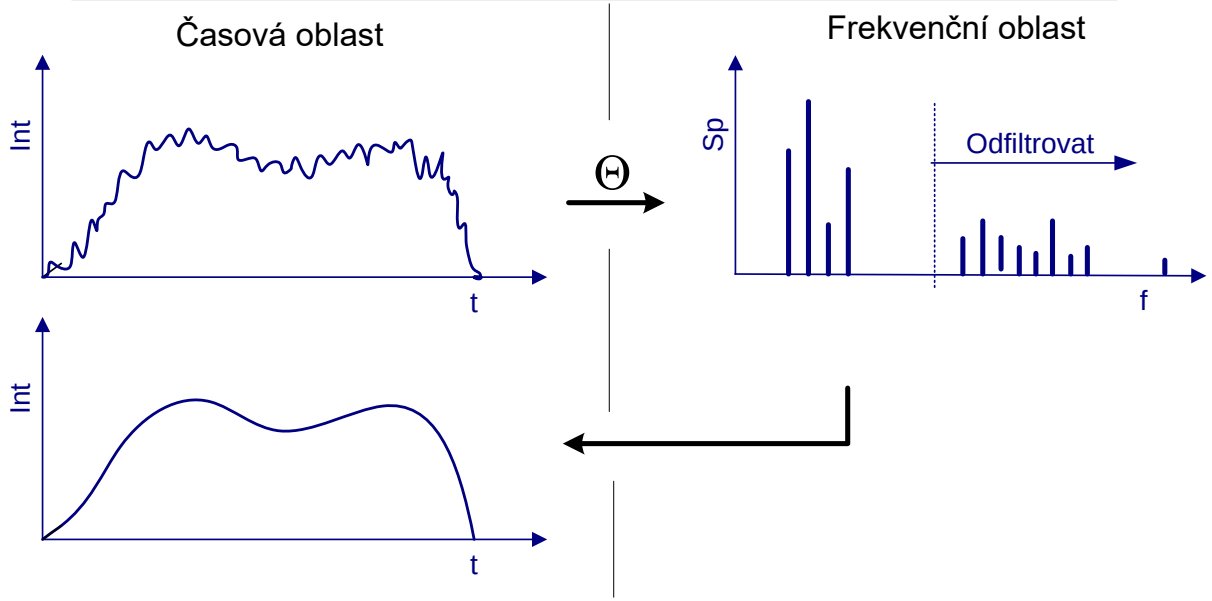


FIGURE 4.5 Basic steps for filtering in the frequency domain.

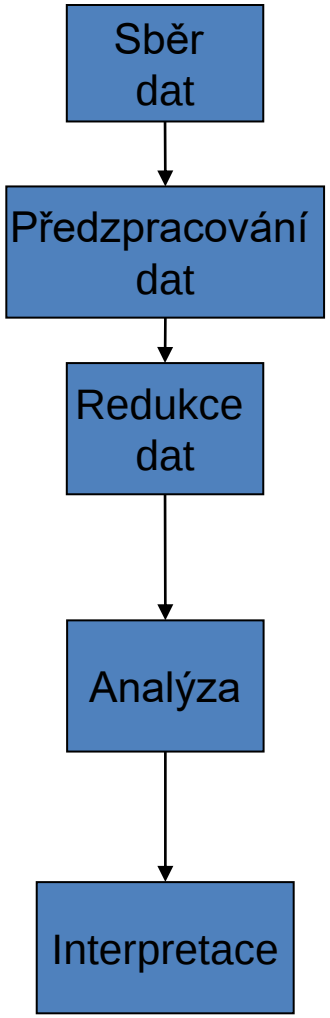


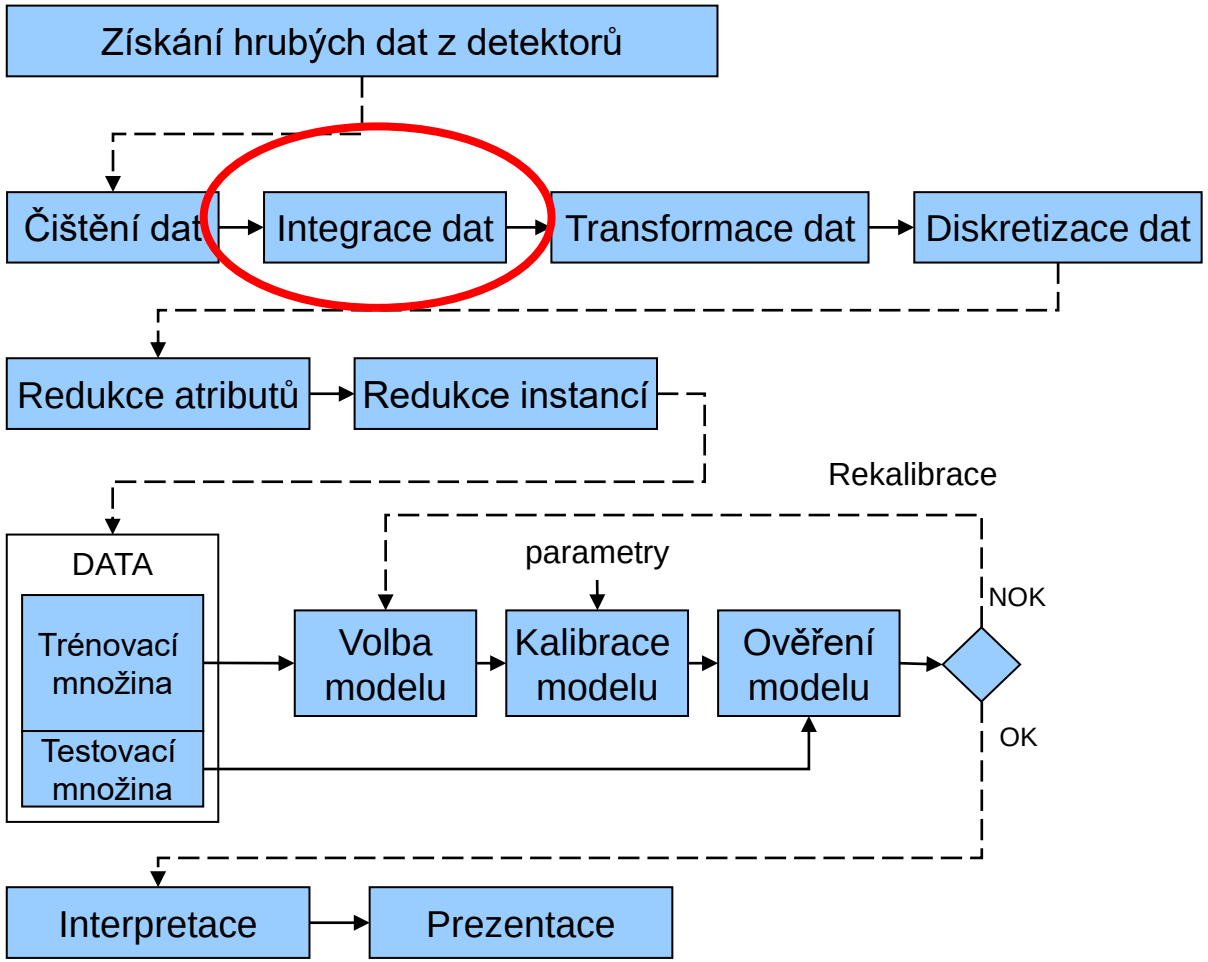
Filtrace

- 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 dochází ke **zpoždění informace** – pozor, pokud potřebujeme rychle reagovat na změnu trendu dat (např. při identifikaci 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 (Kč)
    - Rychlost (km/h) / 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)

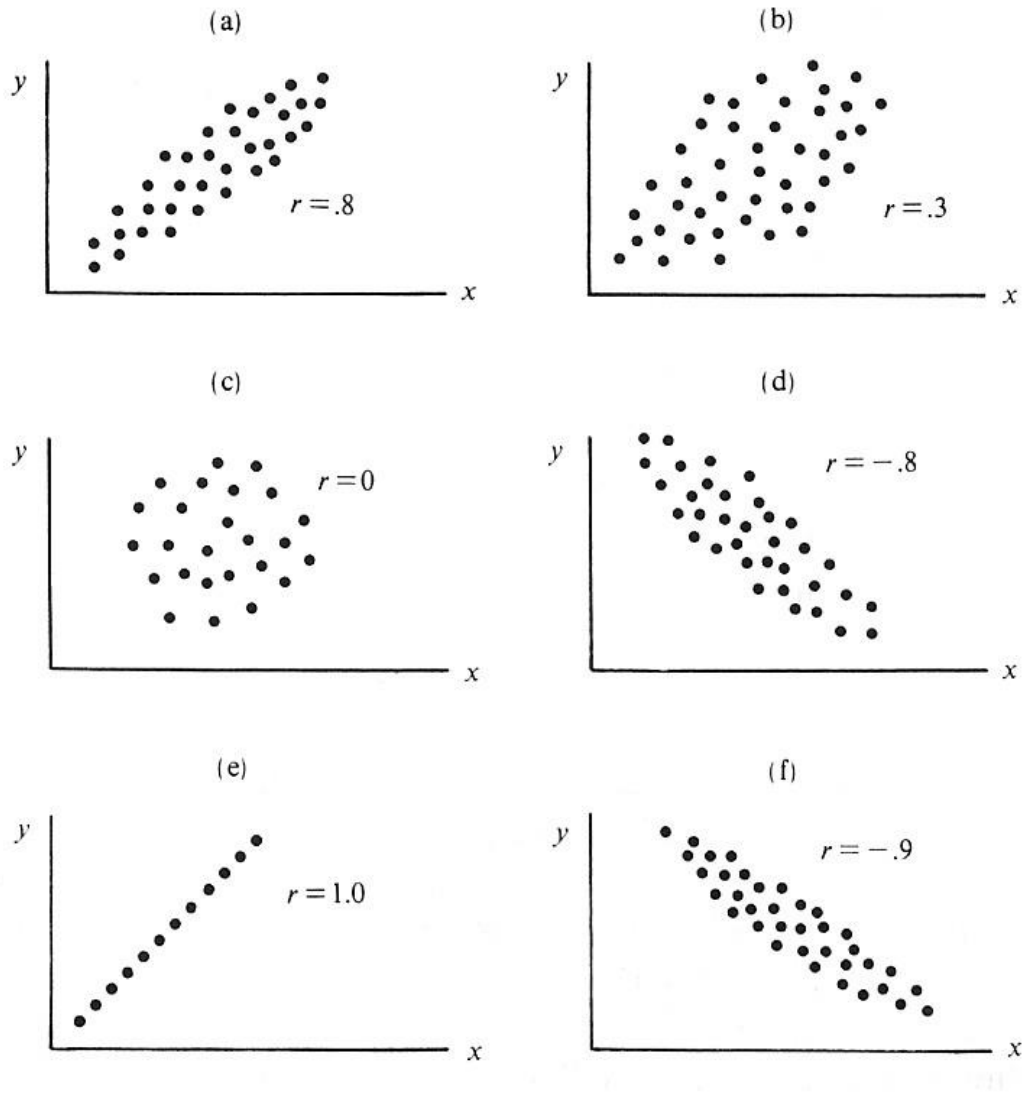
$$R = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1)\, s_x\, s_y} \qquad R = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

  - "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. "

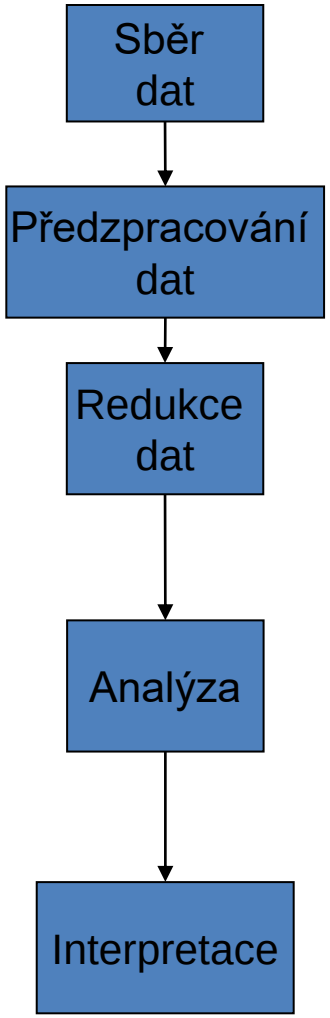
Diskuse

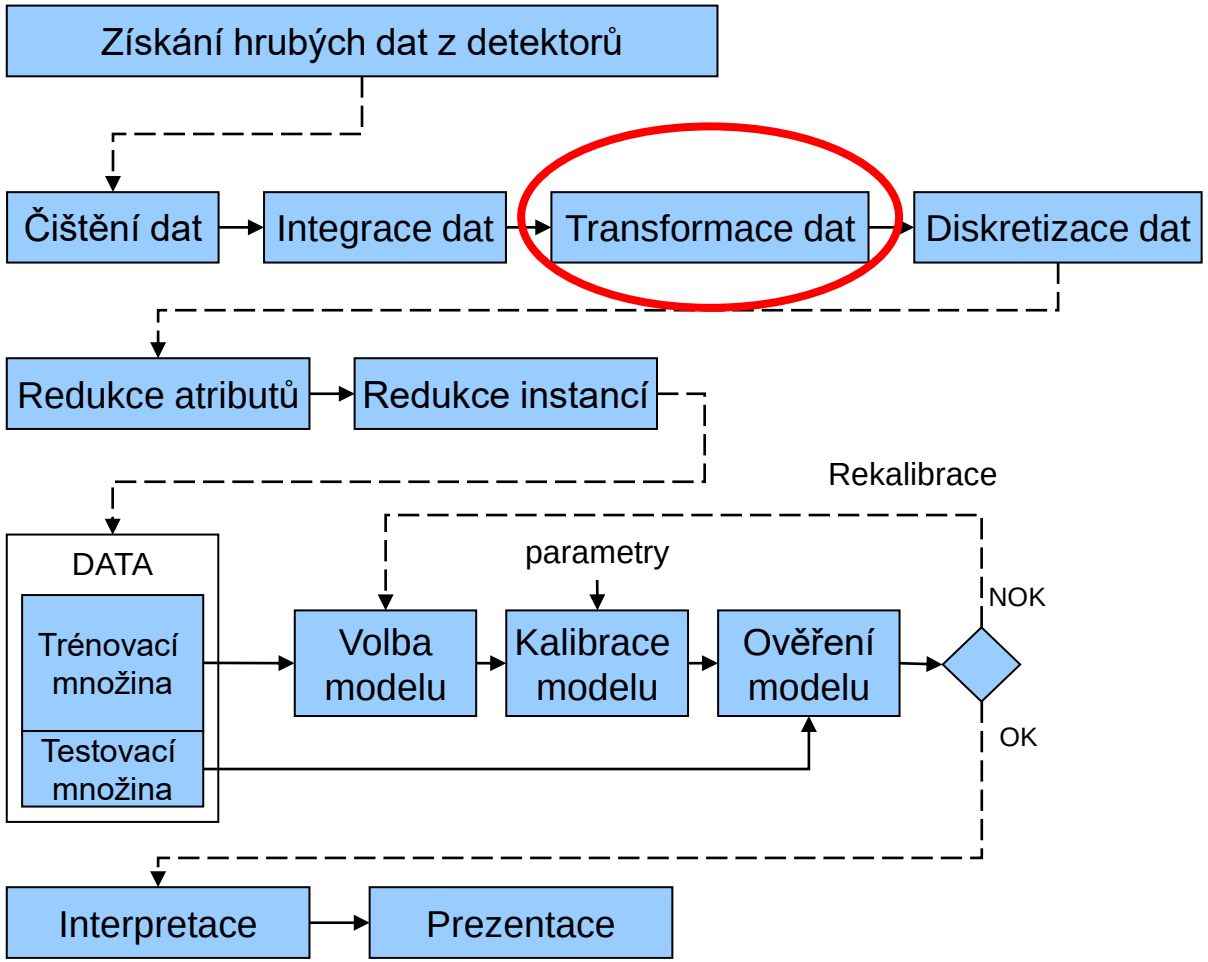
- Co znamenají hodnoty korelačního koeficientu rovné
  - 0?
  - 1?
  - -1?
  - 0,5?

Ukázka korelačních koeficientů



Hlavní kroky

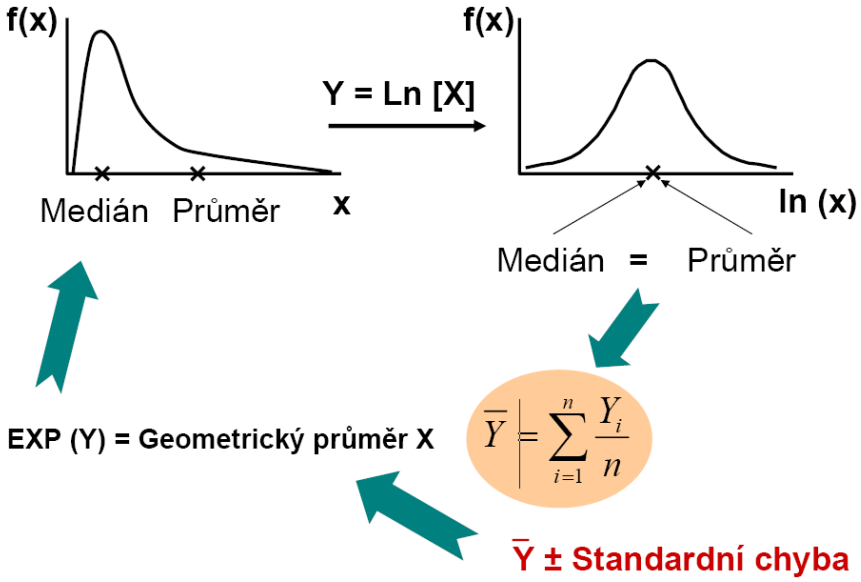




Transformace dat

Odstranění závislosti atributů na jednotkách měření

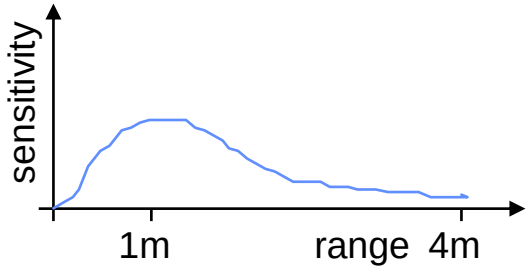
- Transformace
  - **Logaritmická**, odmocninová, …
- Agregace
  - souhrny, vytváření globálních pravidel
- Generalizace
  - koncept hierarchického rozvrstvení
- **Normalizace**
  - Změna měřítka tak, aby data náležela do určitého intervalu
  - Vhodné pro porovnávání různých dat



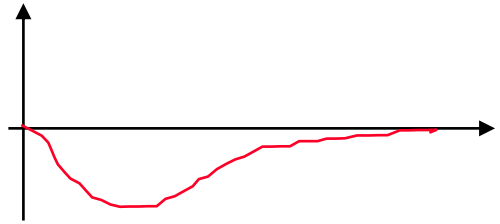
Příklad transformace -

nelinearita

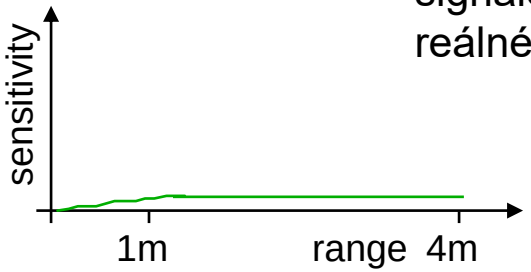
- **Korekce mlhy**



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

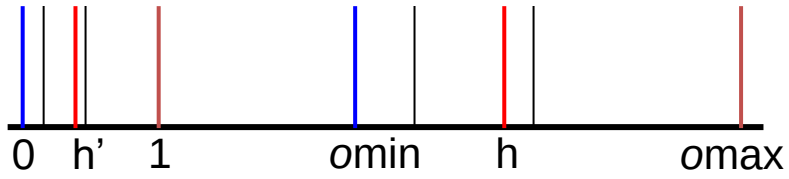


Výsledná křivka citlivosti je téměř plochá. To zaručuje konstantní citlivost pro celou zaručenou vzdálenost detekce.

Transformace dat - normalizace

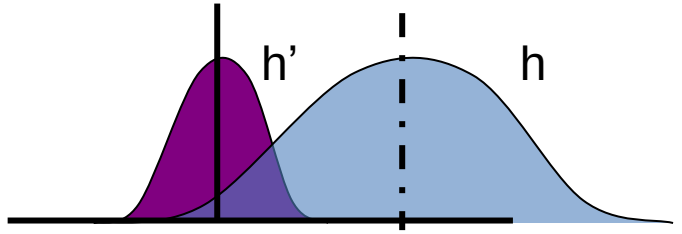
- Lineární normalizace

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

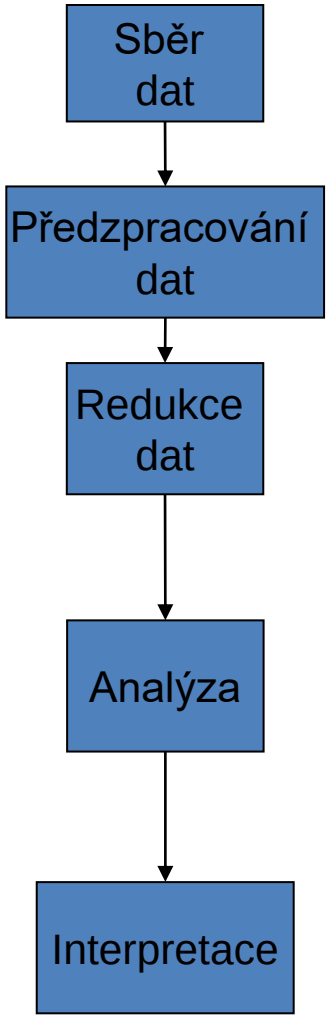


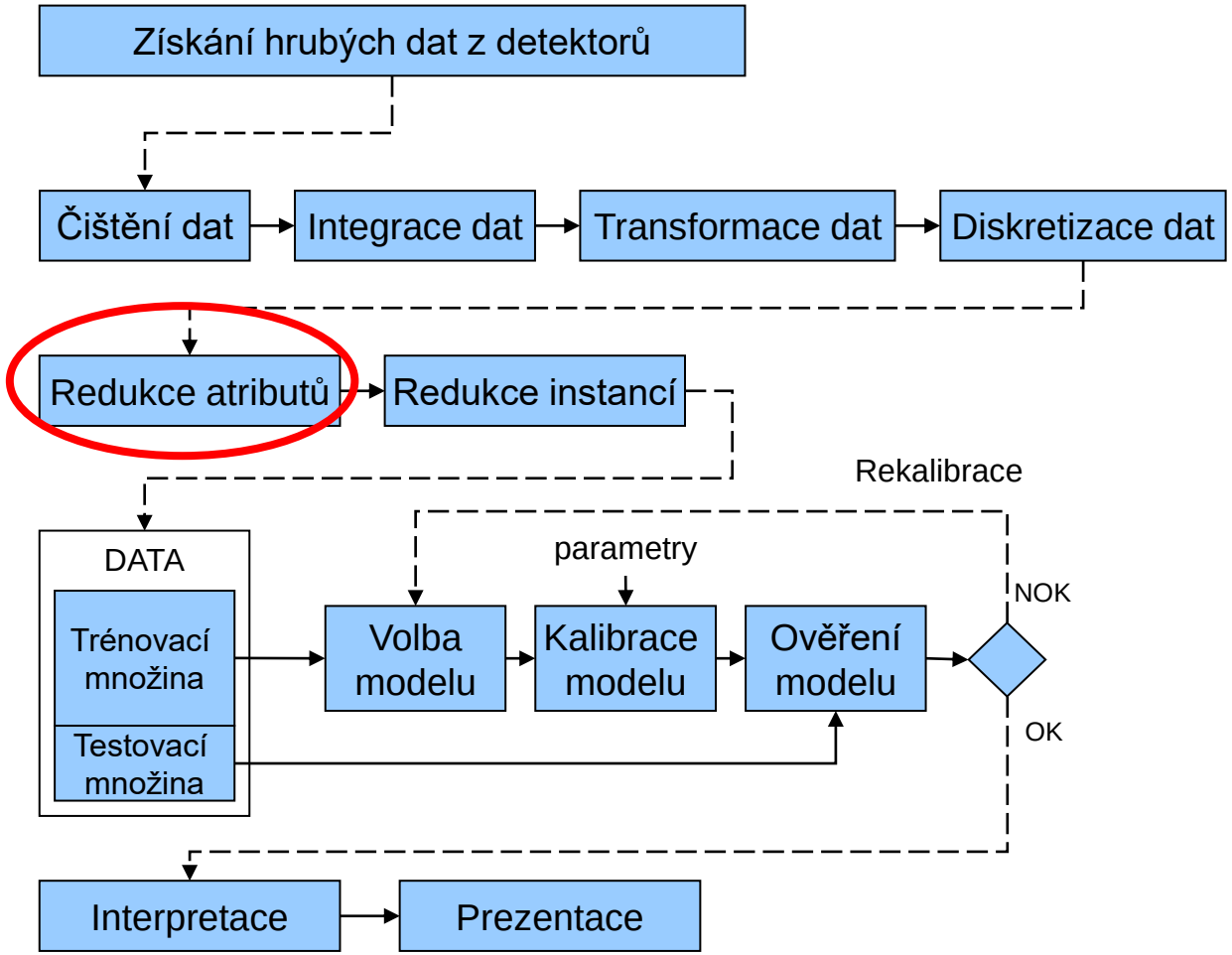
- Z-score normalizace

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



Hlavní kroky



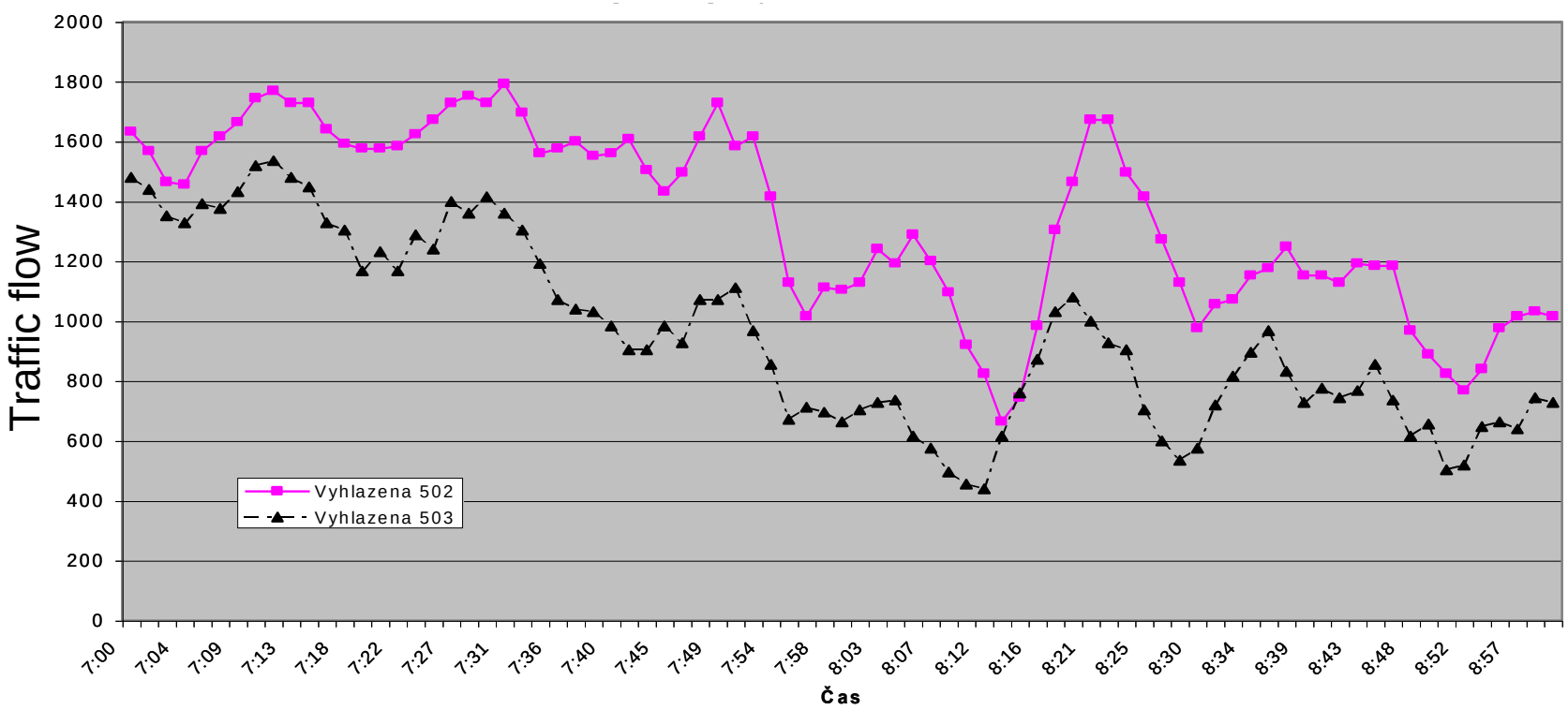


Proč je třeba redukce dat?

- 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

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

Redukce dat – cíl

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

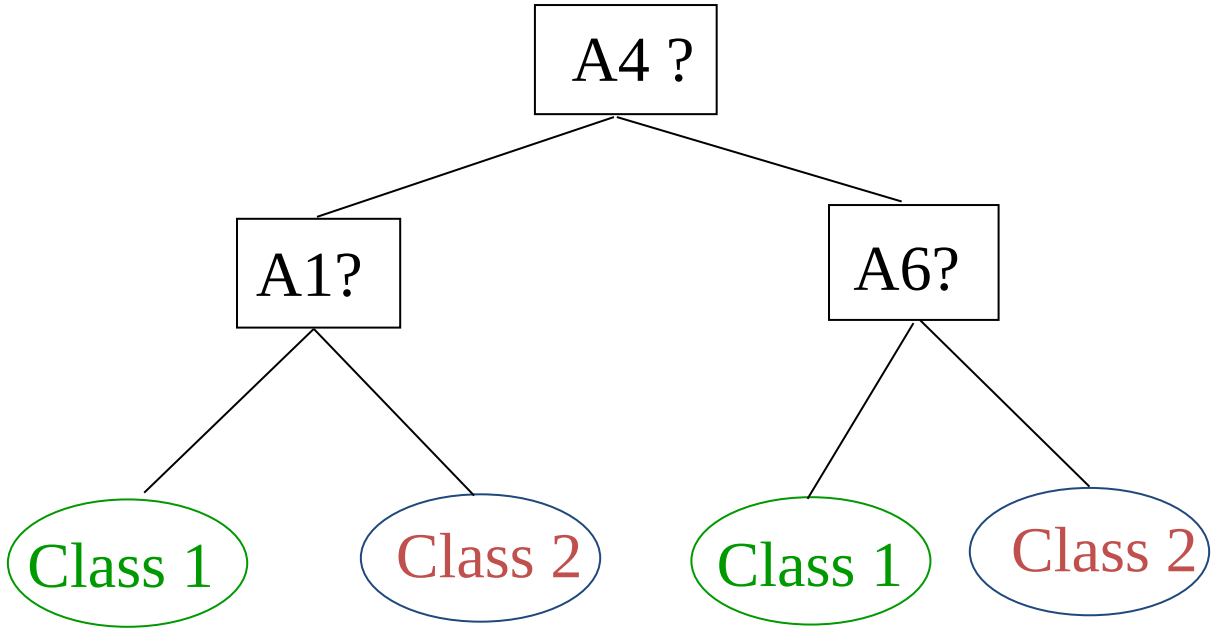
Redukce dimenzionality - Jak vybrat atributy?

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

- **Vynecháním**
  - Konstantních atributů
  - Řídce obsazených atributů
  - Atributů s duplicitní informací (věk × datum narození)
    - Korelační analýza, ANOVA
- **Sloučením** atributů
- **Analyticky**
  - Rozhodovací stromy
  - Fourierova transformace, vlnková (wavelet) transformace
  - Analýza hlavních komponent (angl. *principal component analysis*, PCA)
  - Shlukování

Příklad rozhodovacího stromu

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

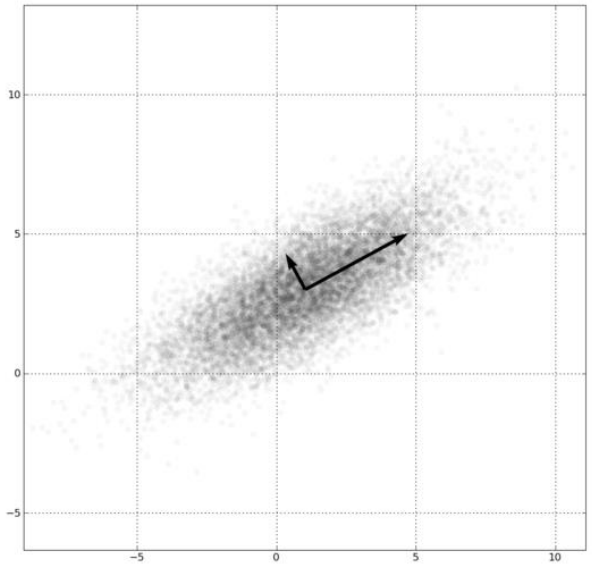


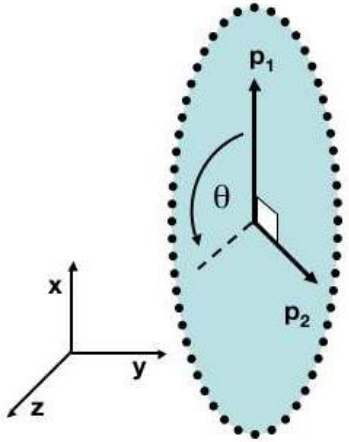
- Redukovaná množina atributů: **{A1, A4, A6}**

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

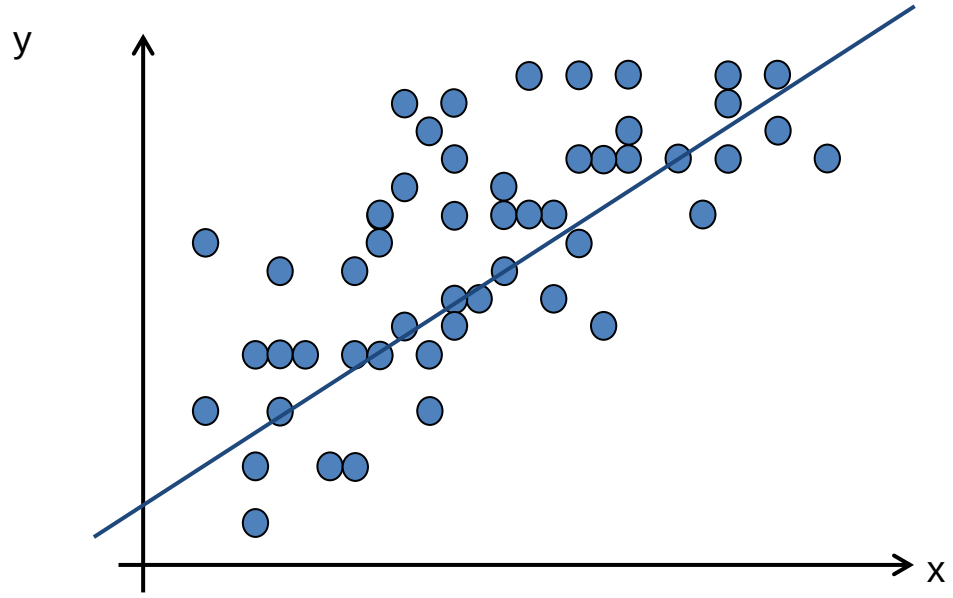
- Komprese řetězců
  - Mnohé existující algoritmy (mpeg4, jpg, …)
  - Obvykle bezeztrátové
- Využívá se pro kompresi audio/video dat
- Příklad - Kódování RLE (Run-Length Encoding)
  - posloupnost opakujících se bytů se nahradí jednou hodnotou s uvedením počtu opakování

"AAAhoooooj"

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

Redukce počtu vzorků – parametrické metody

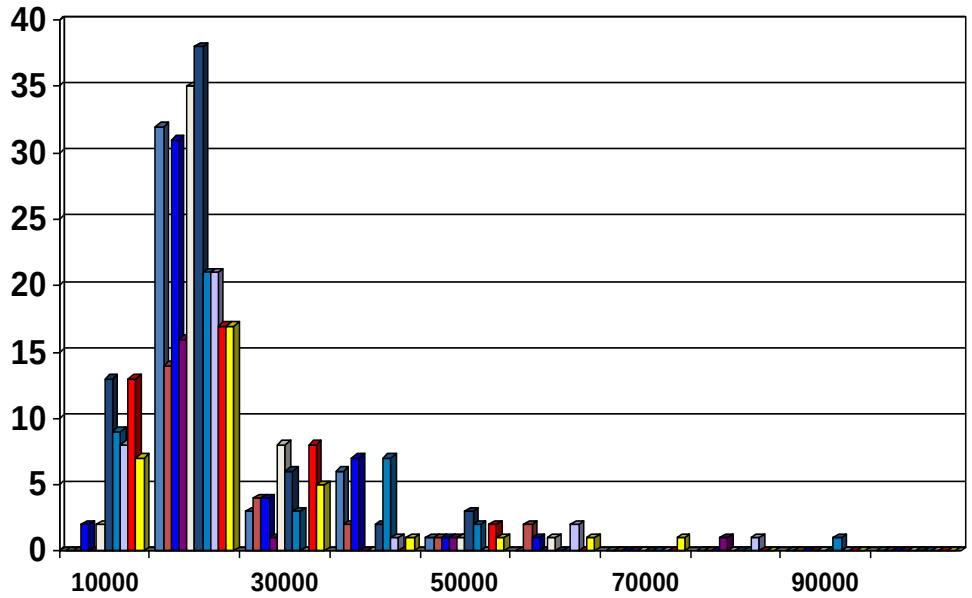
- Parametrické metody
  - Pokud původní data odpovídají určitému modelu, je možné je nahradit tímto modelem

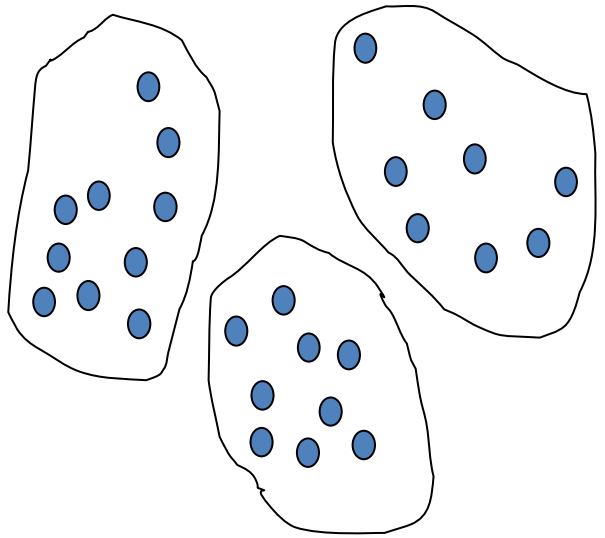


$$y = \beta_0 + \beta_1 \cdot x$$
$$y = 5 + 10 \cdot x$$
$$\rightarrow \{5; 10\}$$

Redukce počtu vzorků – neparametrické metody

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



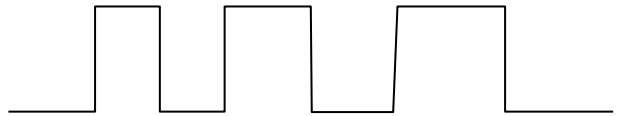


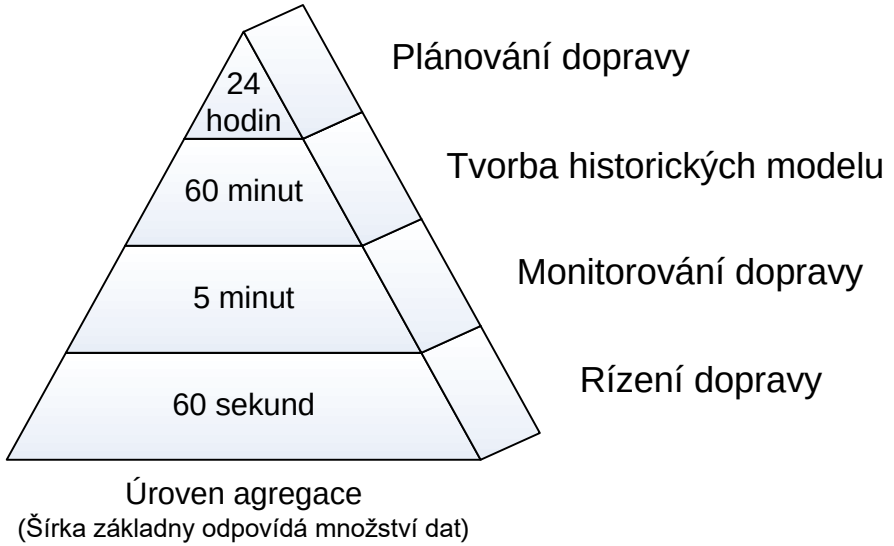
Hierarchická redukce – agregace

- 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