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.

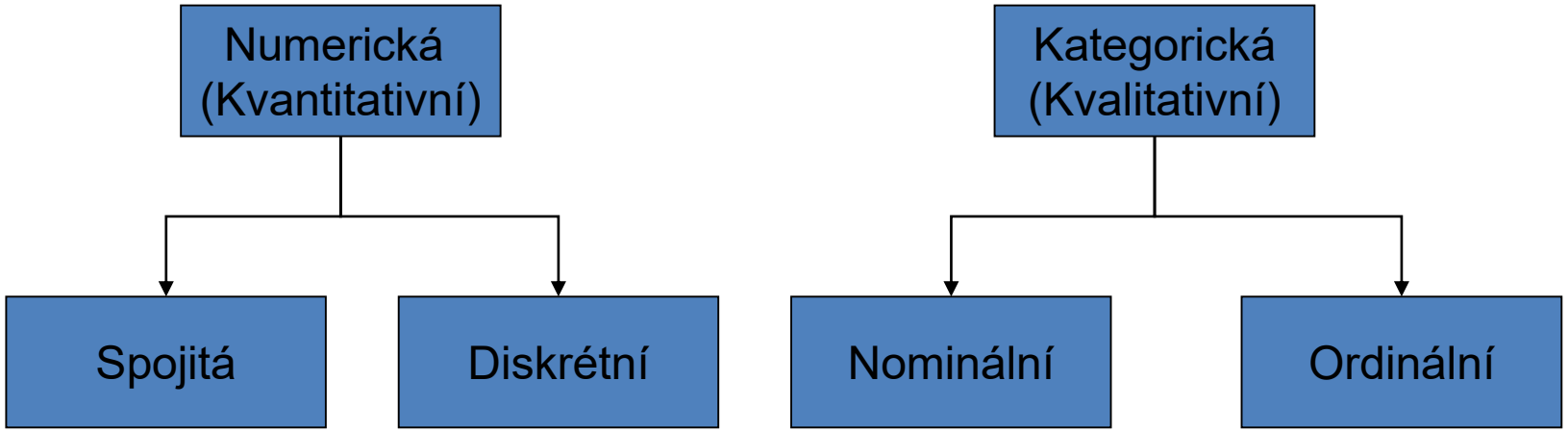
Moudrost

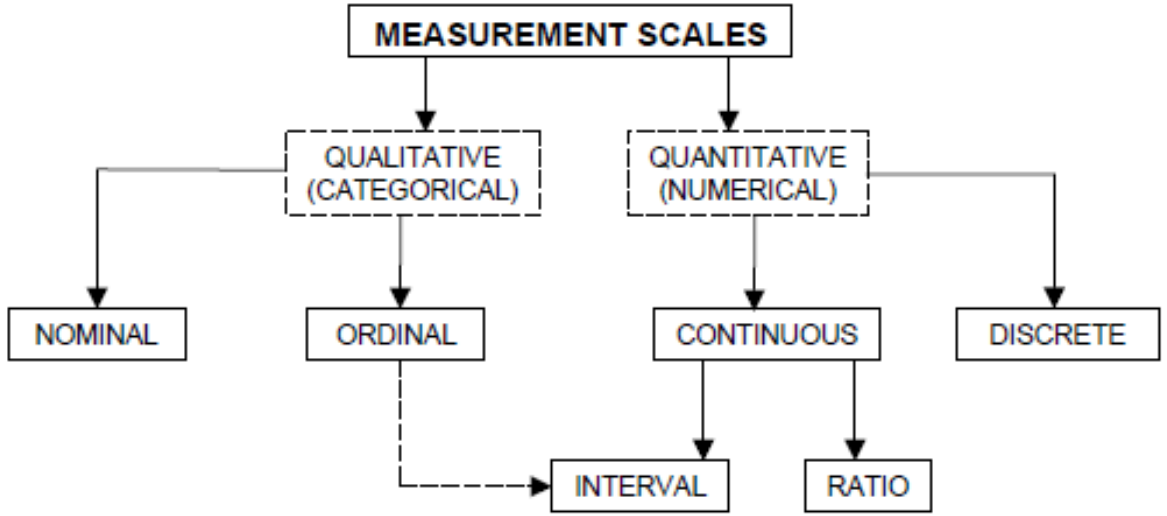
Porozumění principům.

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

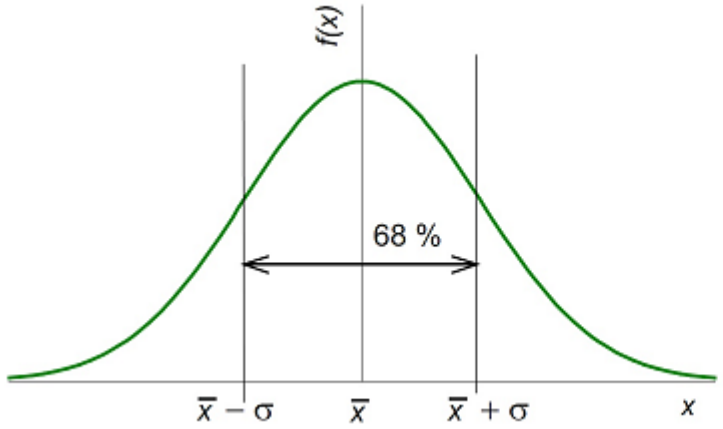
Úvod do problematiky

- 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.

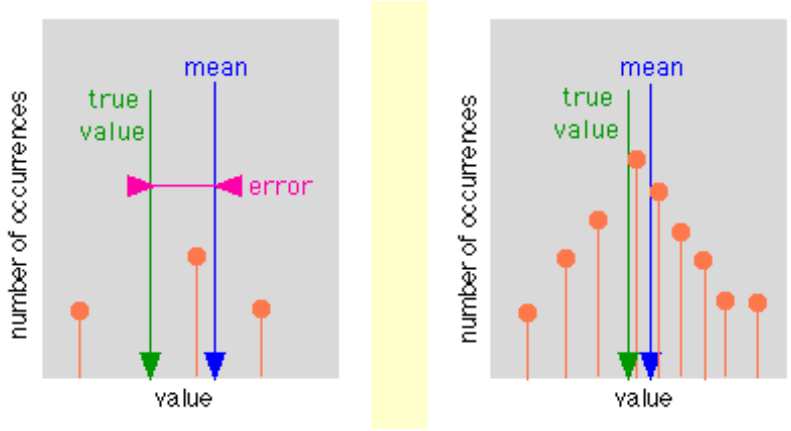


hustota pravděpodobnosti

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

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

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



„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

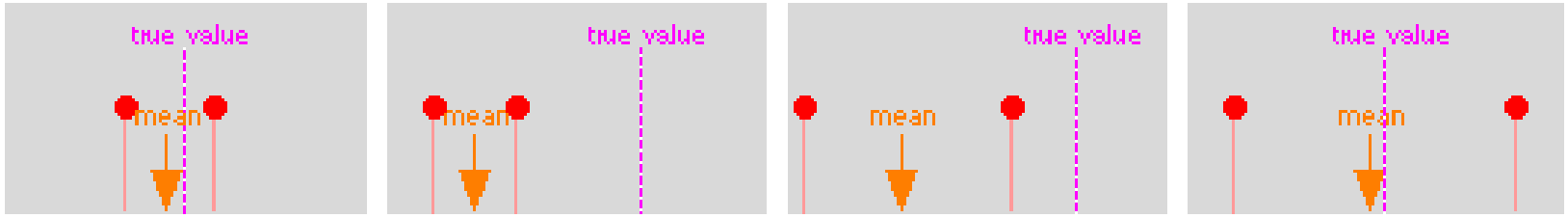
„Přesnost“ versus „správnost“

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

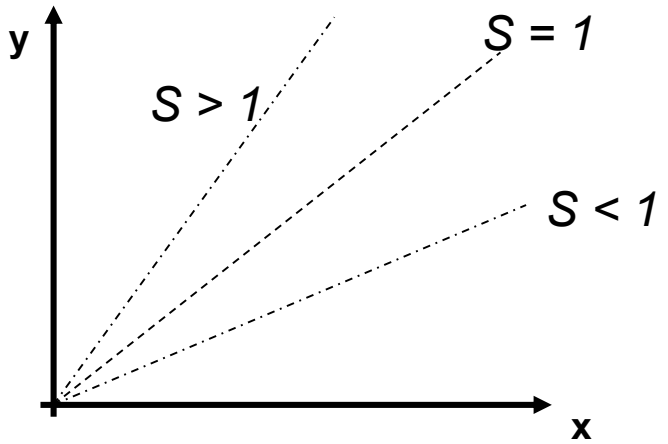


Citlivost (sensitivity) měřicího přístroje

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**

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

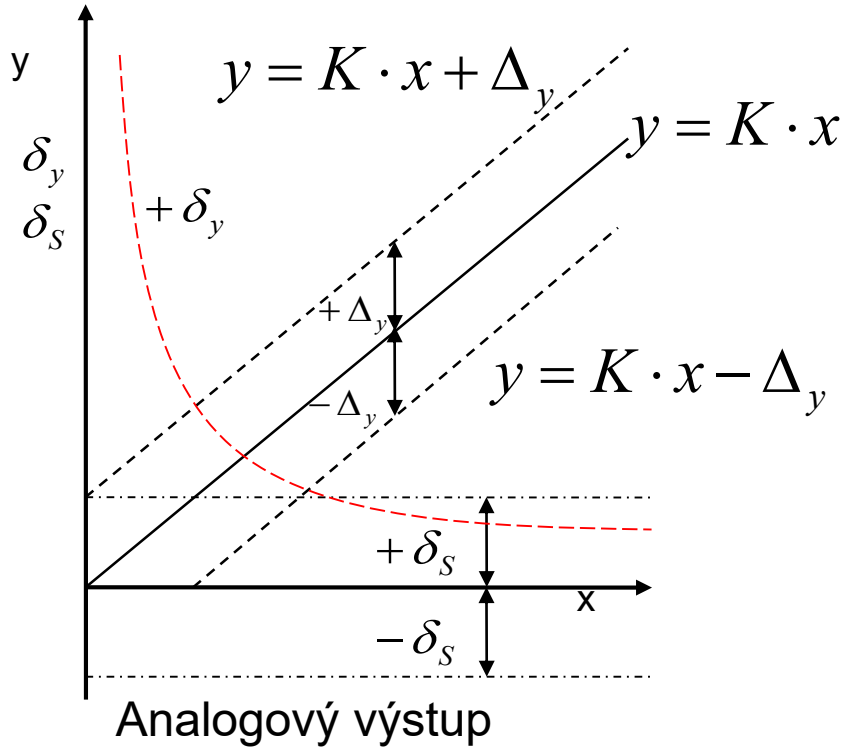
3. **Relativní chyba senzoru**
   Poměr maximální absolutní chyby měření vůči rozsahu hodnot měřené veličiny

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

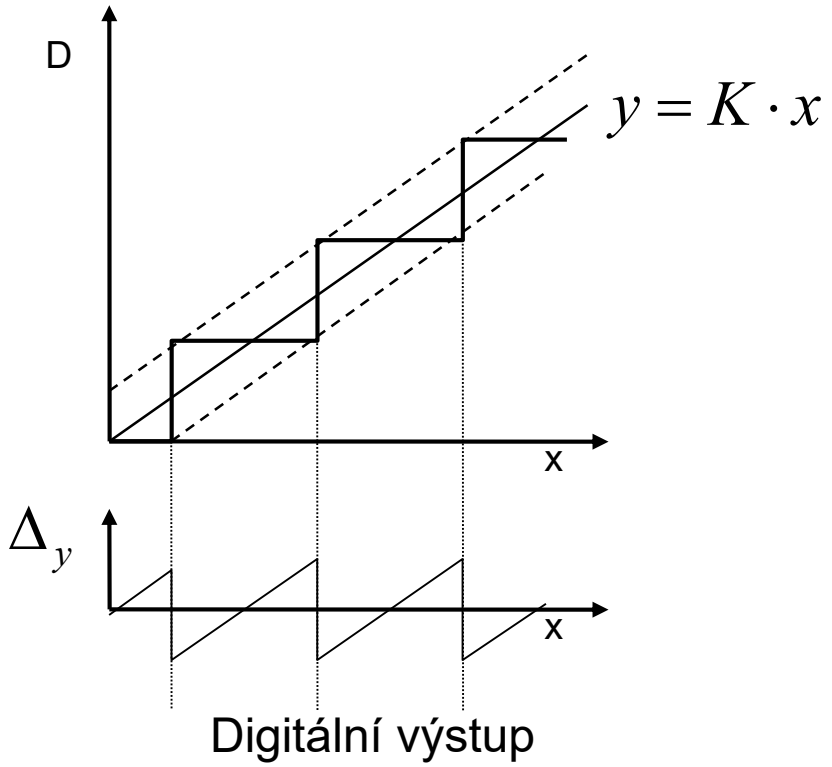
Chyby měření (pokrač.)

4. Aditivní chyba měření

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



Analogový výstup

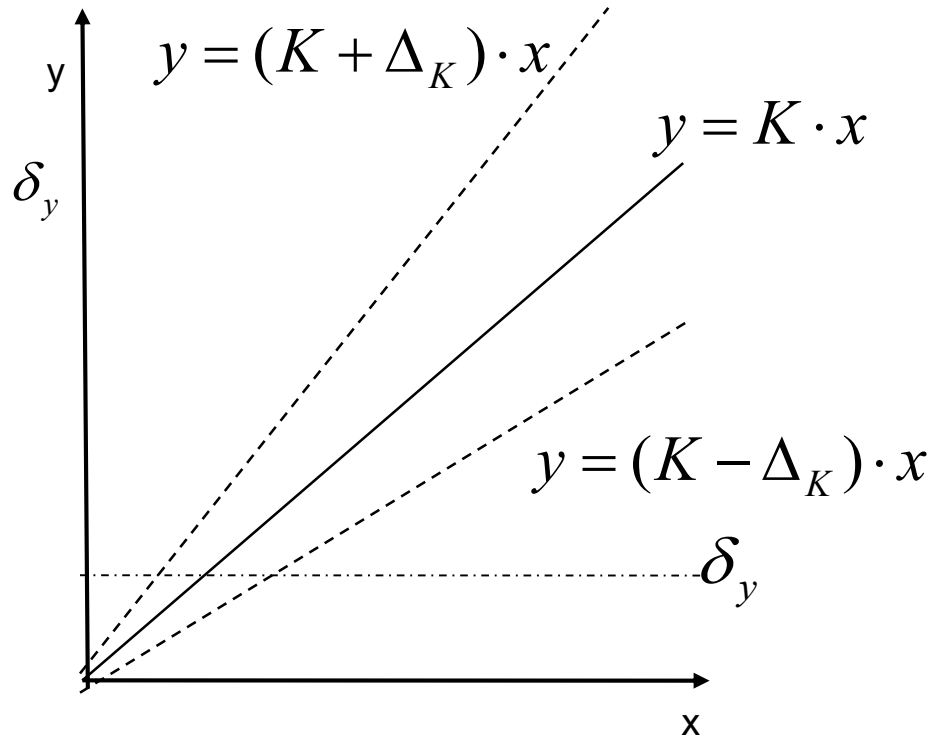


Digitální výstup

Chyby měření

Multiplikativní chyba měření

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



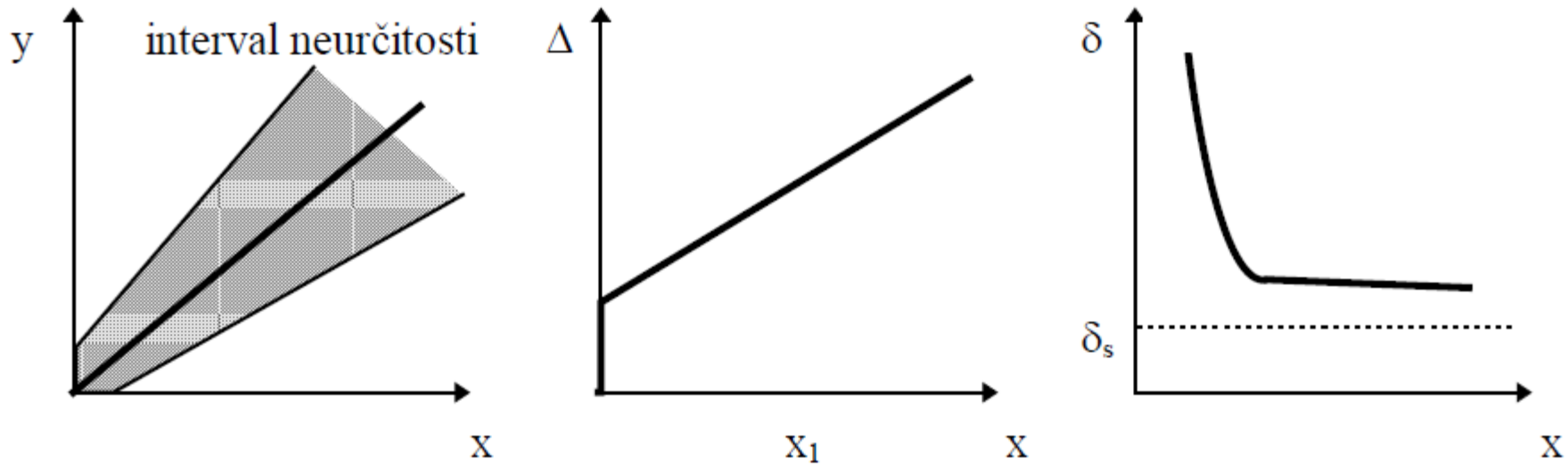
$$\Delta_y = \Delta_K \cdot x$$

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

δ_K Chyba měření

Chyby měření

- Kombinovaná chyba měření
- Kombinace aditivní a multiplikativní chyby

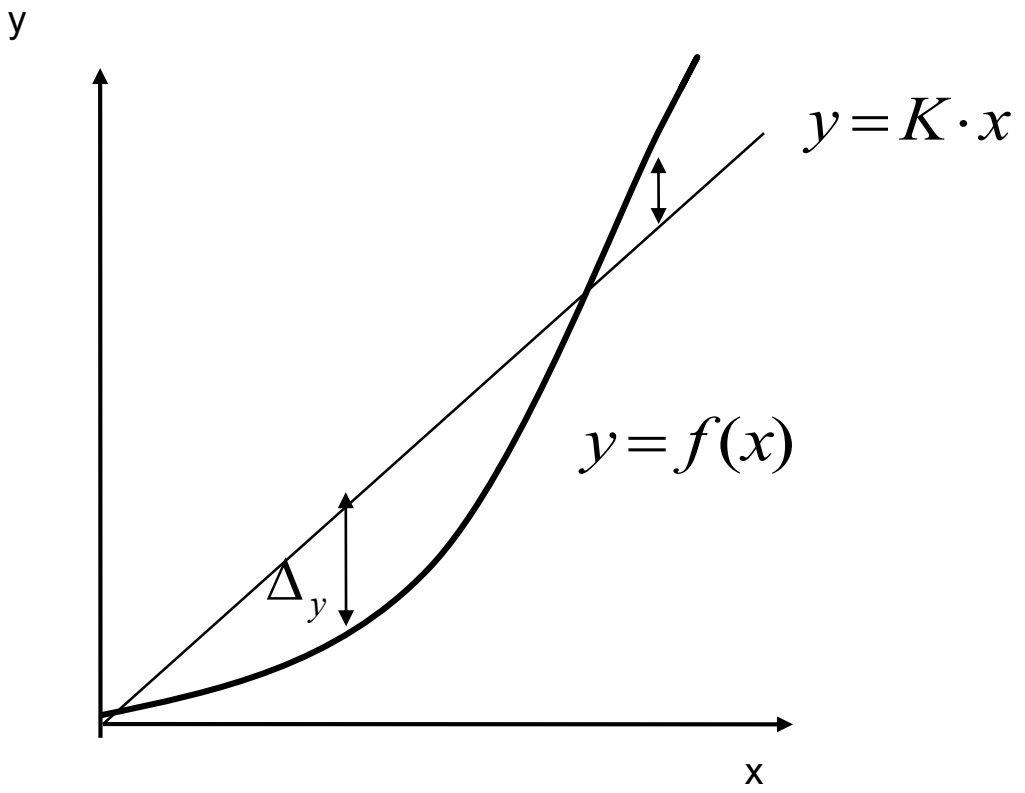


Chyby měření

Chyba linearity

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

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

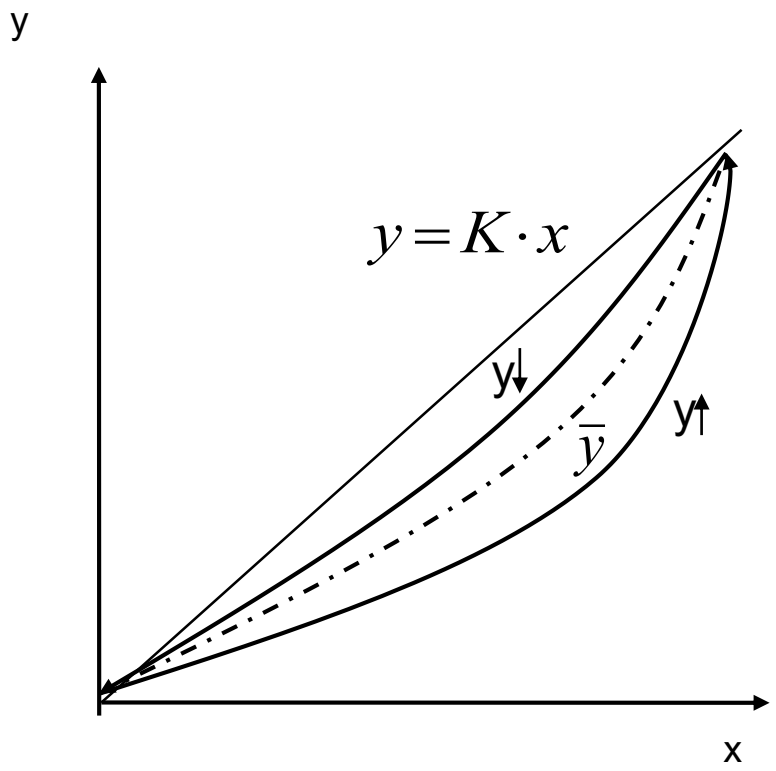


- kde y_L je definována ideální funkcí $y = y_0 + Kx$,
- parametr K lze odhadnout pomocí lineární regrese.

Chyby měření

Chyba hystereze

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

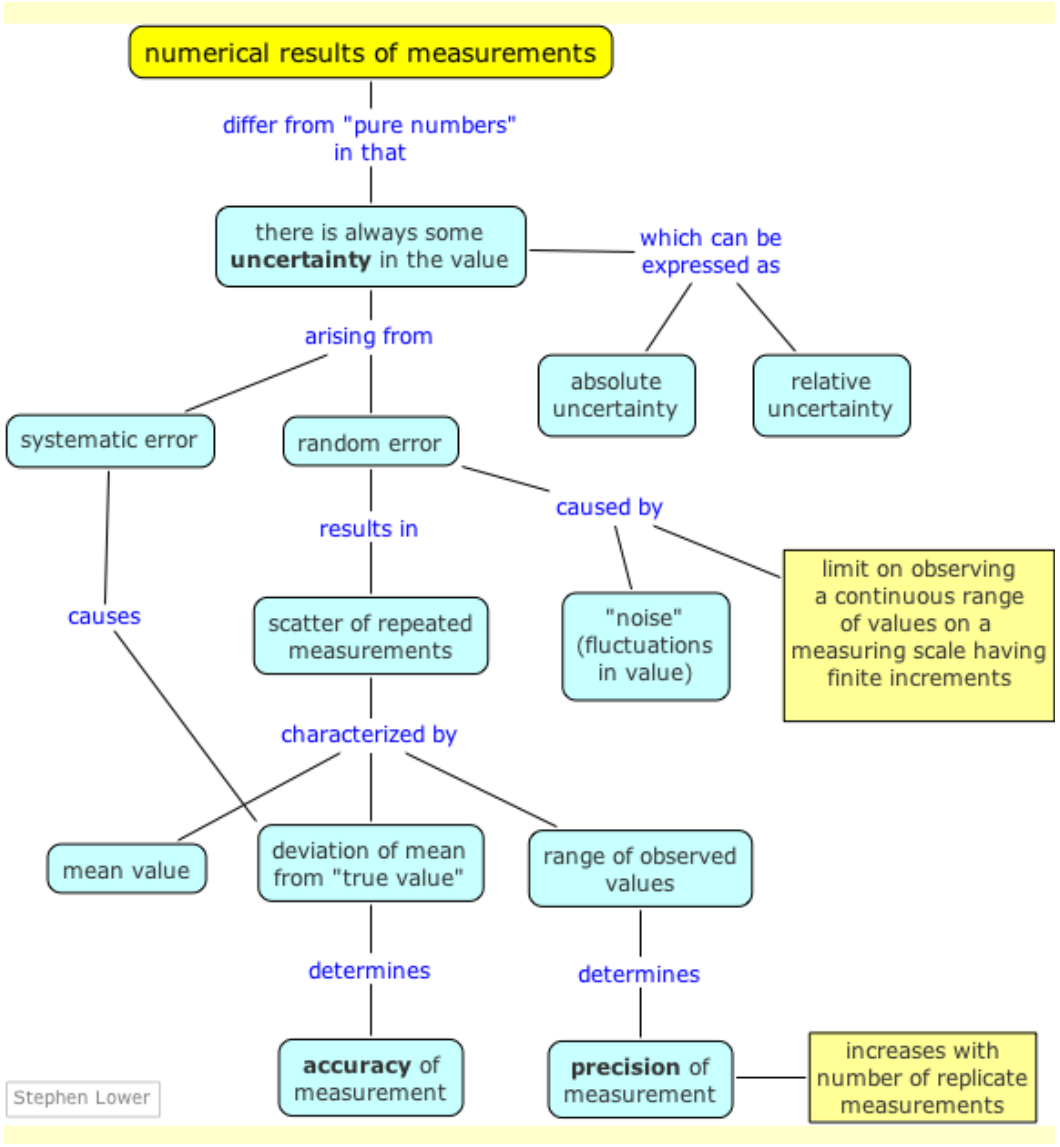


$$\delta_S = \left(\frac{y\downarrow - y\uparrow}{y_{\max}} \right)_{\max} = \left(\frac{\Delta_{yH}}{y_{\max}} \right)_{\max}$$

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

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

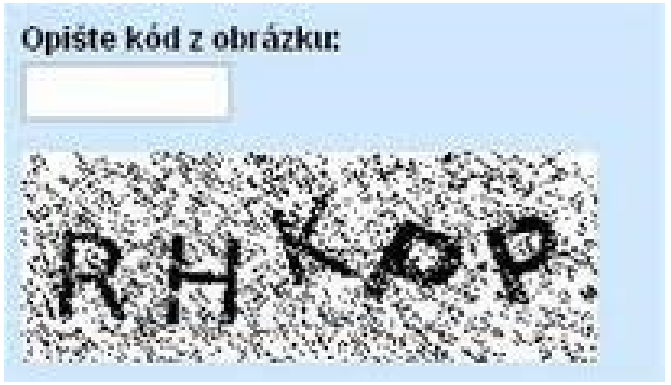
- Měřené veličiny
- Chyby měření
- **Základní charakteristiky dat**
- Vizualizace dat
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

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

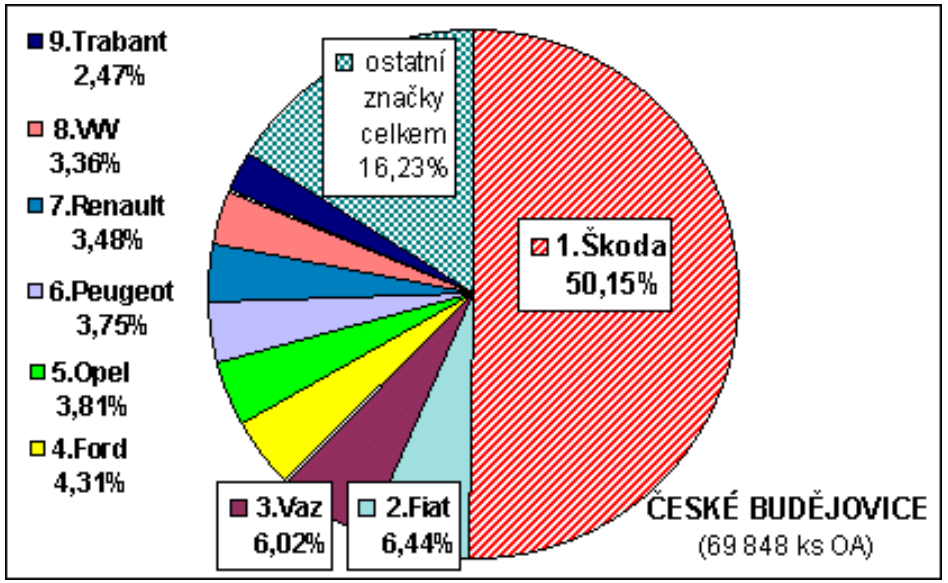
Cíle

- 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

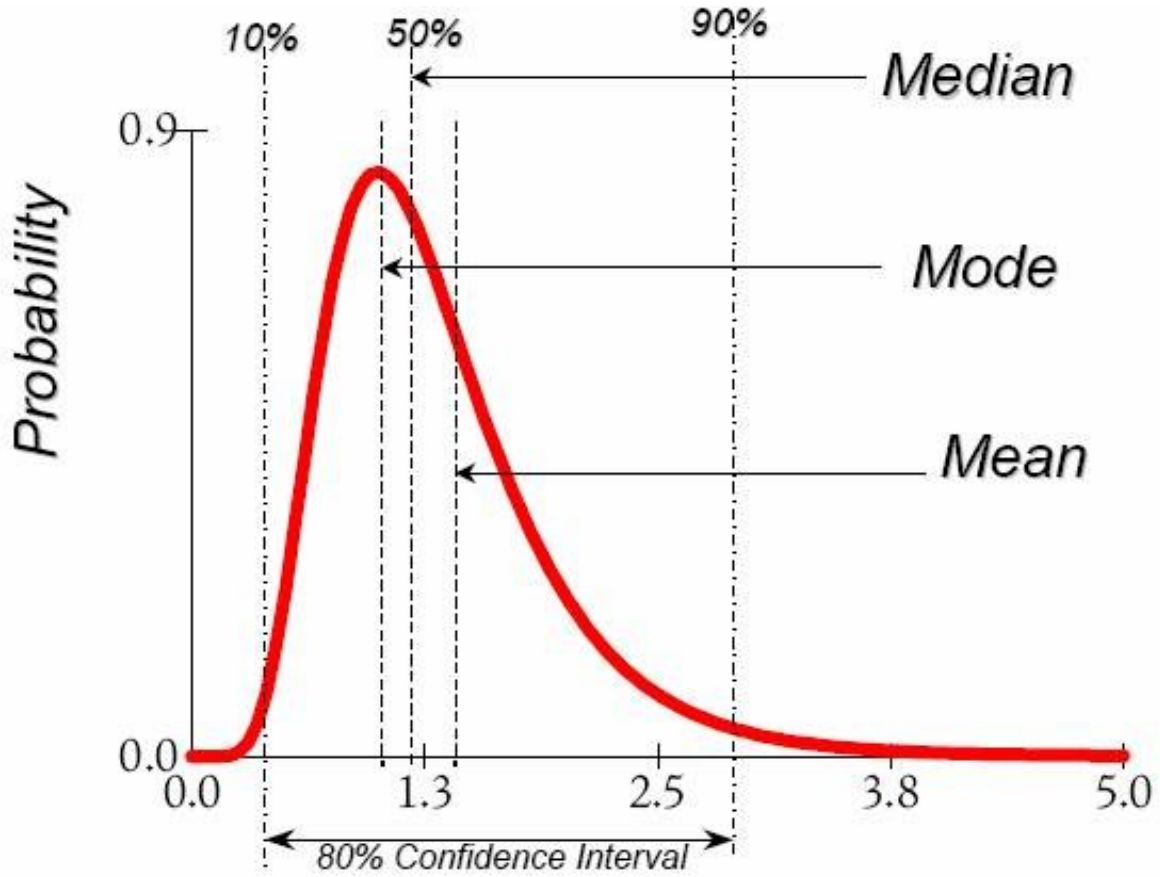


Modus, Medián a Aritmetický průměr atributu

- Aritmetický průměr (mean)
  - statistická veličina, která v vyjadřuje typickou hodnotu
  - součet všech hodnot vydělený jejich počtem.
- Modus atributu (mode)
  - **nejčastější** hodnota v daném statistickém souboru
  - je to 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)
- Modus atributu (mode)
- Medián (median)



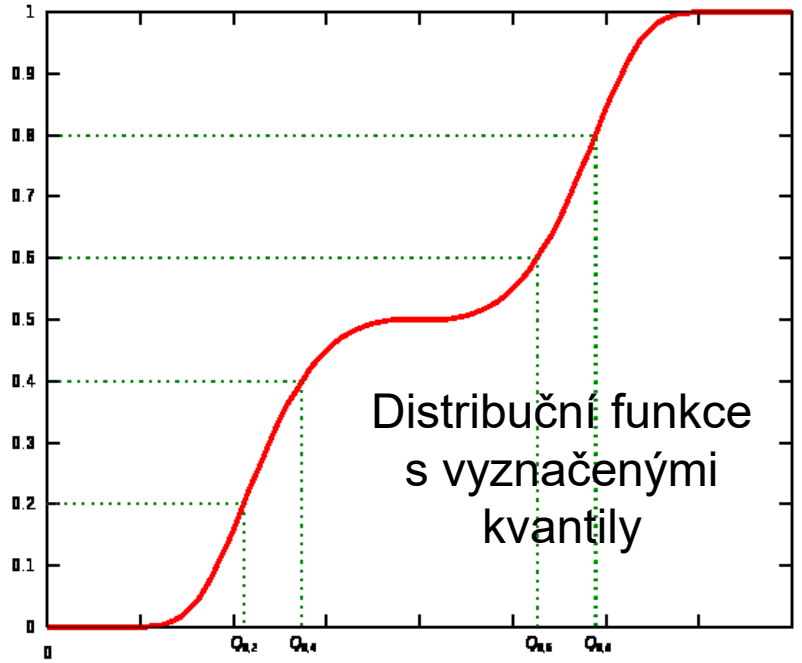
Příklad

- 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

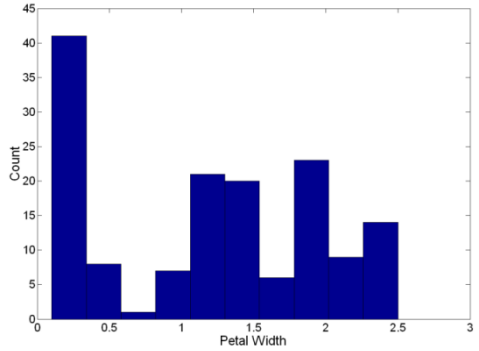
- Měřené veličiny
- Chyby měření
- Základní charakteristiky dat
- **Vizualizace dat**
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

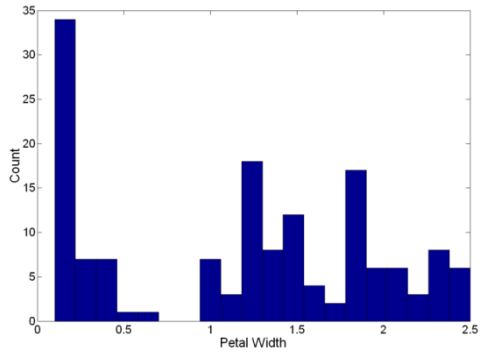
Vizualizace

- převedení dat do vizuální či tabulkové podoby pro potřeby analýzy dat
- velmi silným nástrojem pro **průzkumovou analýzu** dat.
  - 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

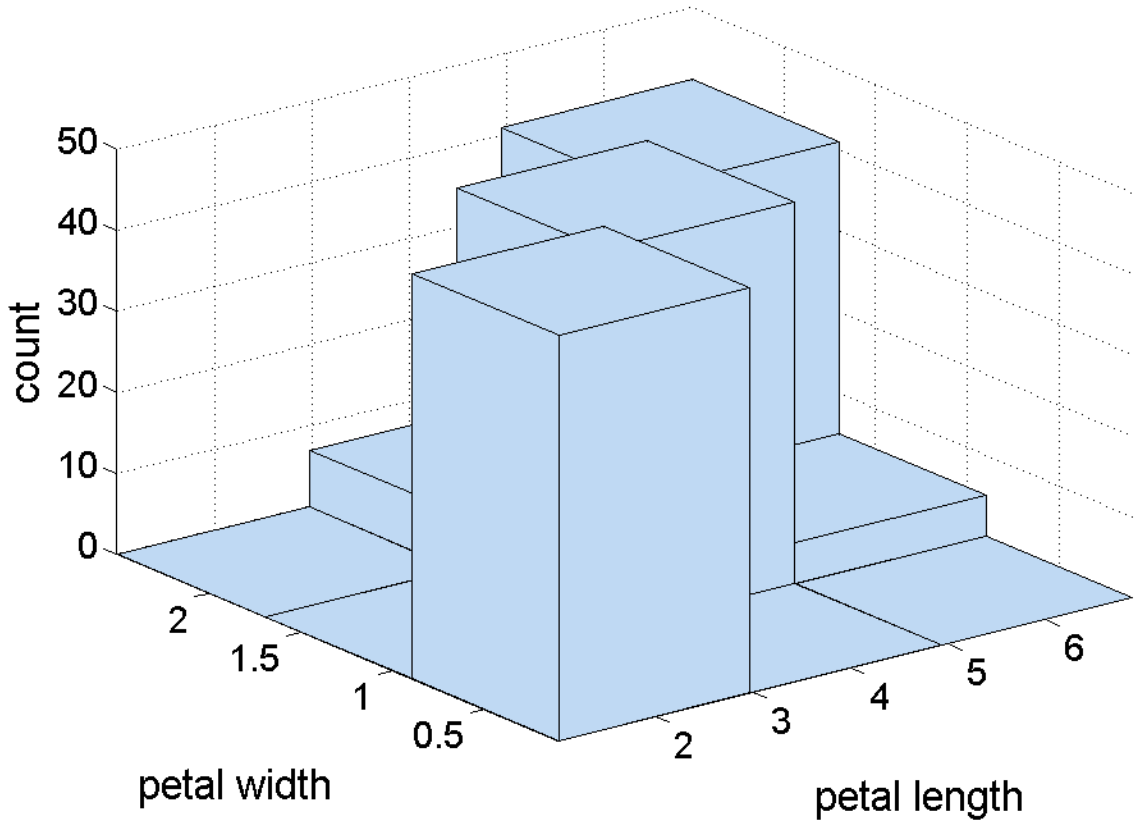
- Histogram
  - Rozdělí hodnoty do intervalů a zobrazí jejich četnosti
  - Výška sloupce udává počet objektů v daném intervalu
  - říká zda je soubor homogenní, nebo zda se rozpadá do dílčích menších podsouborů
    - jen jedna nejčetnější hodnota (homogenní soubor)
    - více hodnot s většími četnostmi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





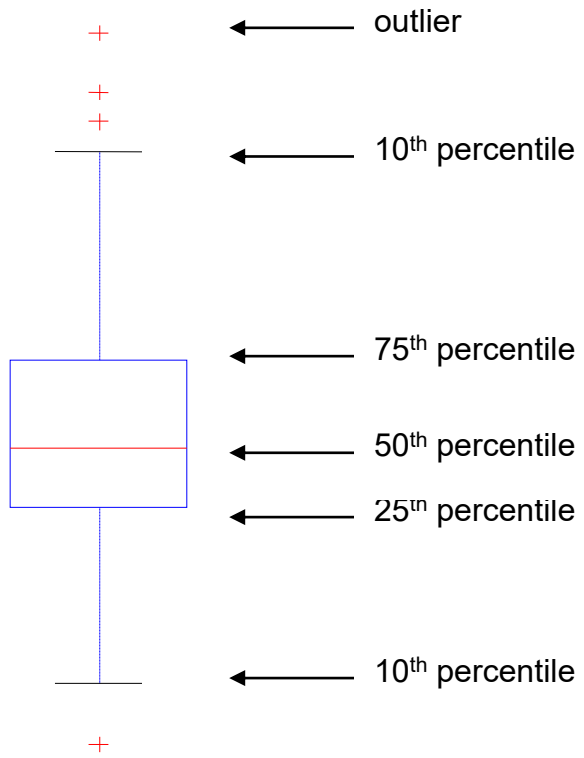
Dvoudimensionální Histogram

- Zobrazuje společné rozdělení dvou atributů



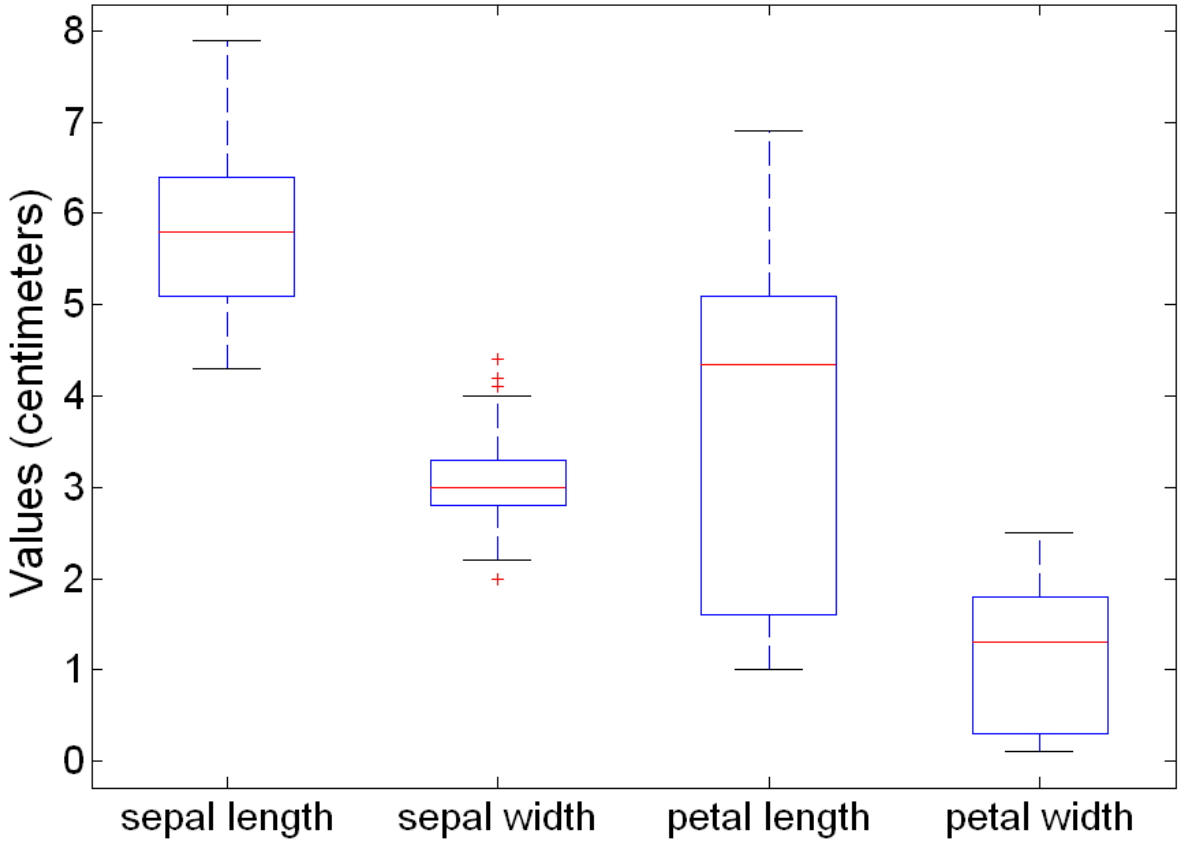
Vizualizační techniky: Box Plots

- Box Plots
  - grafické zobrazení tzv. 5ti číselného souhrnu
  - Autor J. Tukey



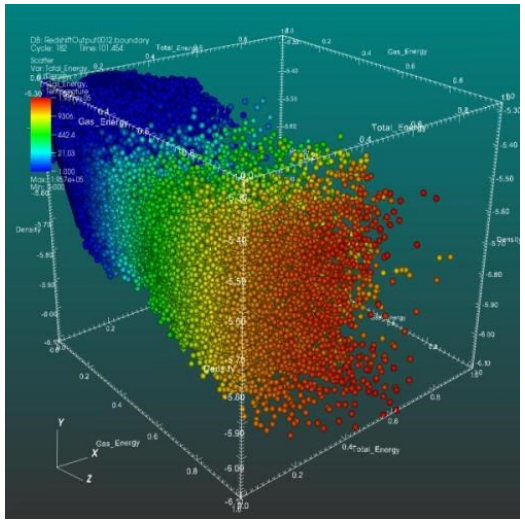
Příklad Box Plots

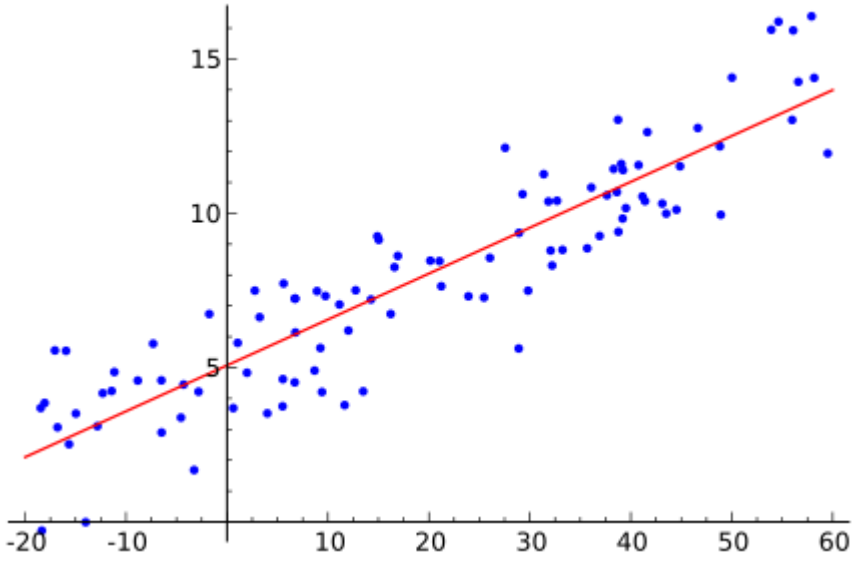
- Box plots se využívají k porovnání atributů



Vizualizační techniky: Korelační diagram

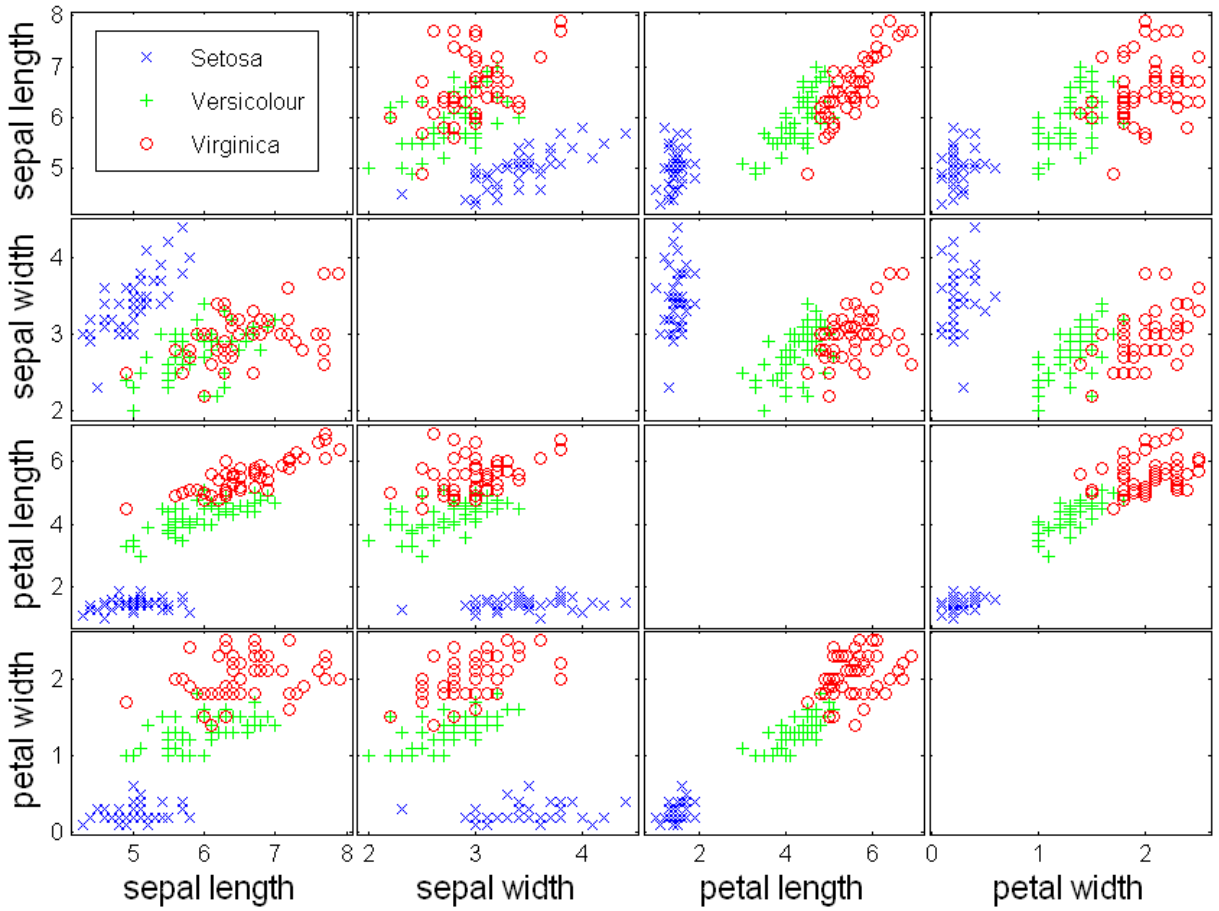
- též bodový graf (anglicky Scatter plot)
- matematické schéma užívající kartézských souřadnic pro zobrazení souboru dat o dvou (či tří) proměnných (na osy).
- Takto je možné jednoduše zjistit vzájemný vztah (korelaci) mezi oběma proměnnými





Pole korelačních diagramů

- Vícerozměrné zobrazení je nepřehledné
- Zobrazuje vzájemné vztahy více proměnných



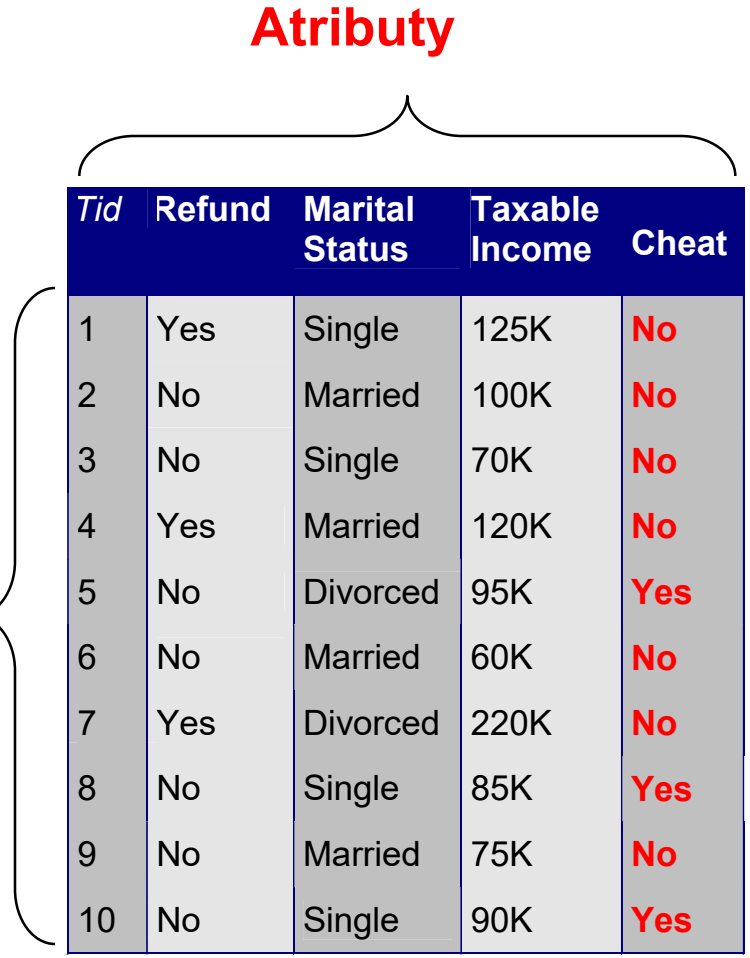
Obsah prezentace

- 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 (rekord (DB), instance, vzorek, entita, ...) je popsán kolekcí atributů

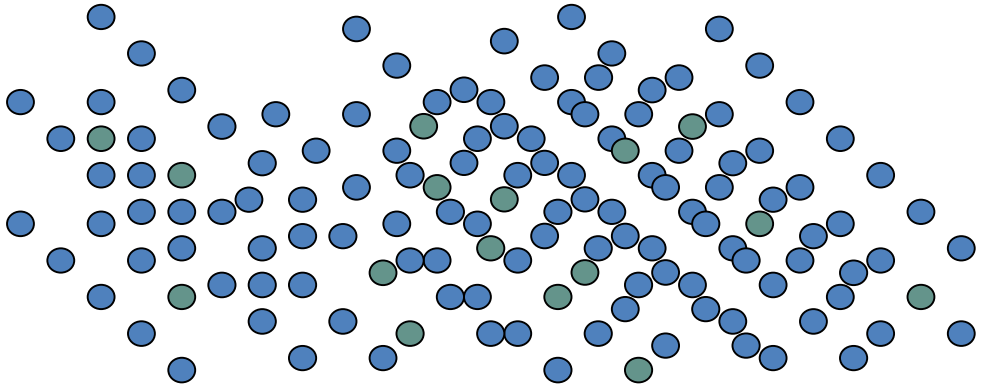


Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Populace versus náhodný výběr

- Základní soubor (populace) - Všechny jednotky
  - např. všichni řidiči v ČR
  - Označují se pásmeny řecké abecedy (μ, σ, ...)
- Výběrový soubor - vybrané jednotky, náhodný výběr
  - např. všichni řidiči, kteří v konkrétním dni jeli autem a stali se účastníky dotazníku
  - Označují se písmeny latinské abecedy (, s, ...)

Zdroj http://www.nedarc.org/



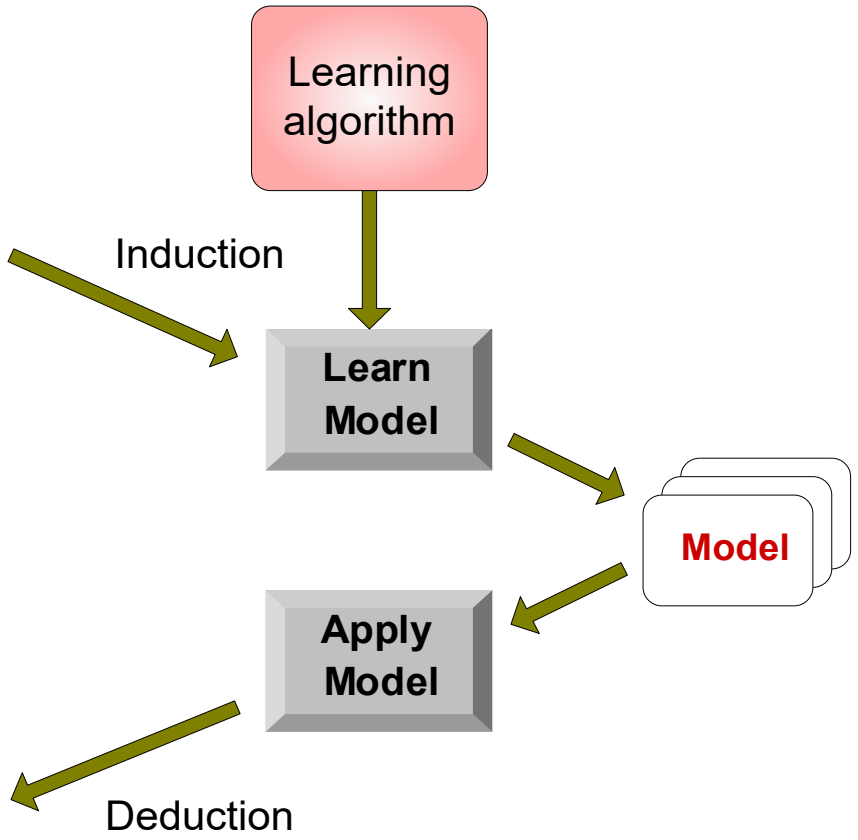
Aplikace modelu

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

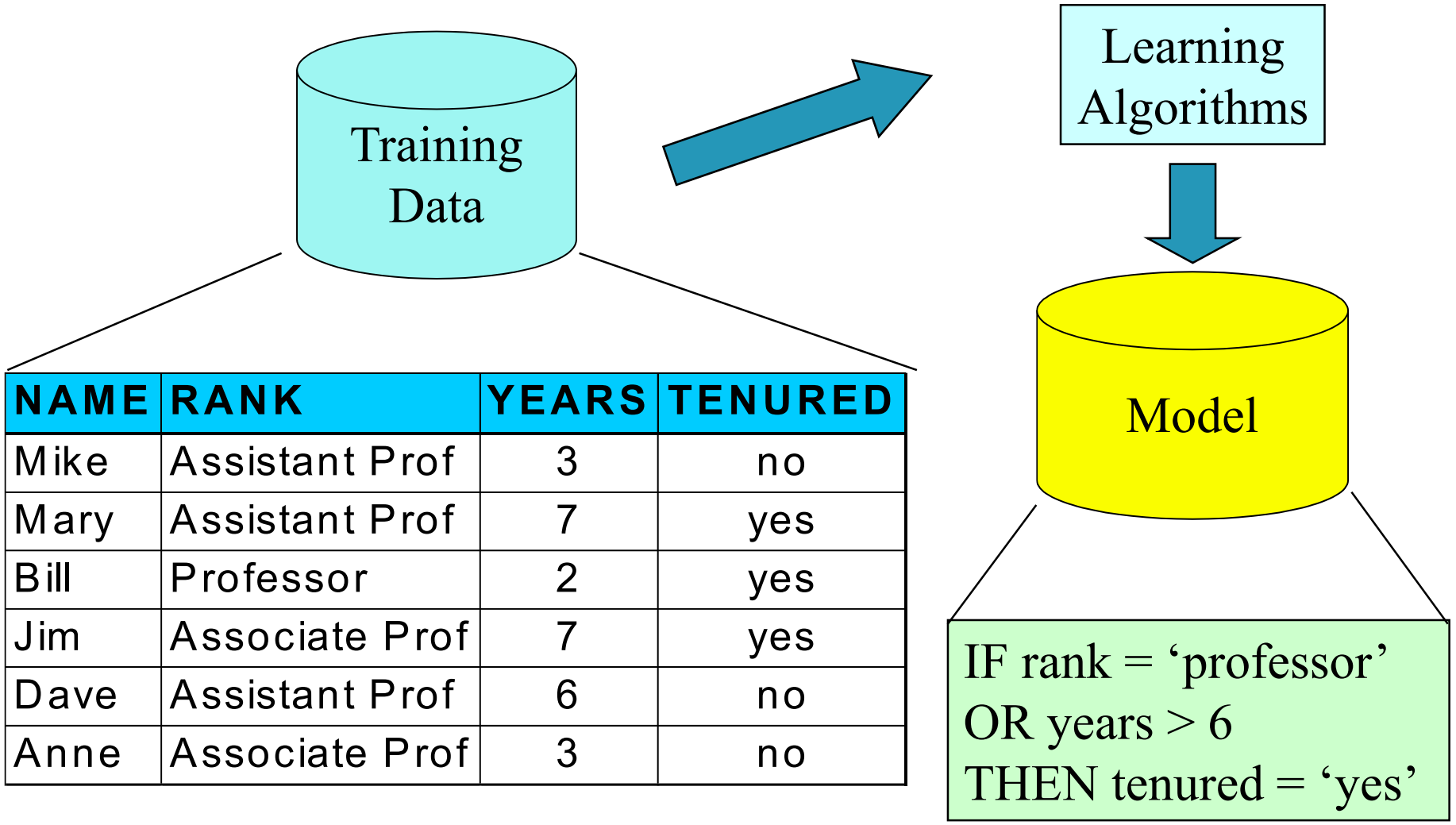
Training Set

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

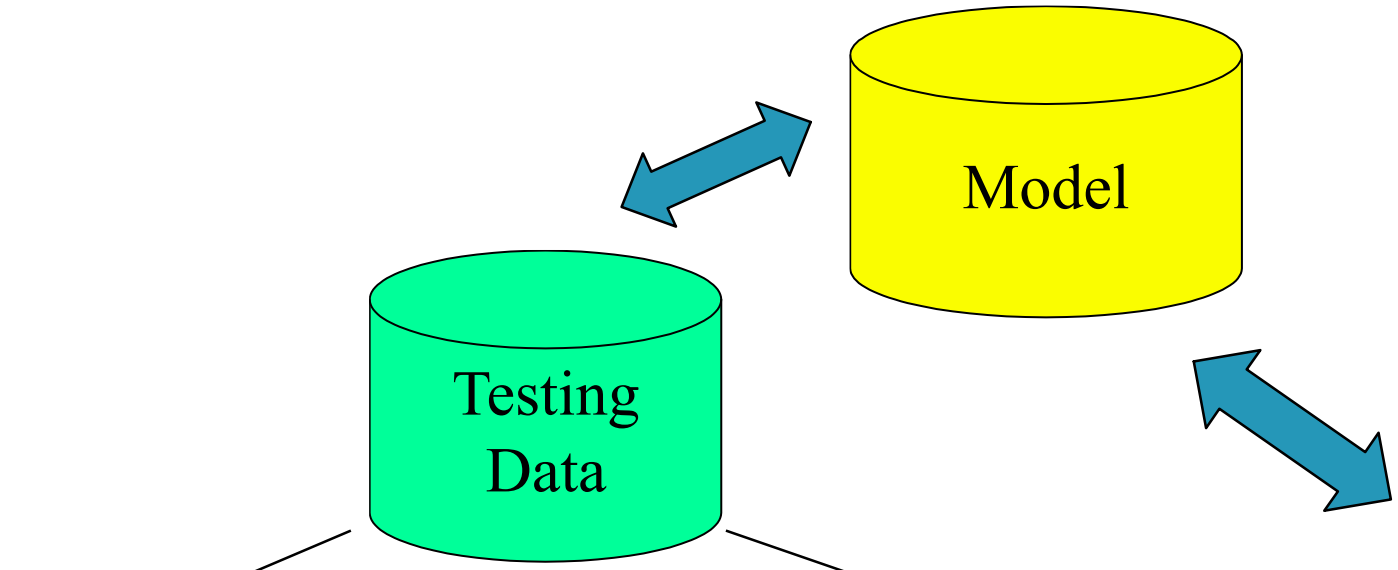


Indukce



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

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 rychlost algoritmu)
- Matice záměn (Confusion Matrix)
  - A je počet správných predikcí, že daná instance je negativní
  - B je počet nesprávných predikcí, že daná instance je negativní
  - C je počet nesprávných predikcí, že daná instance je pozitivní
  - D je počet správných predikcí, že daná instance je pozitivní

	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)

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
  - 9990/10000 = 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		
	C(i|j)	+	-
ACTUAL CLASS	+	-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

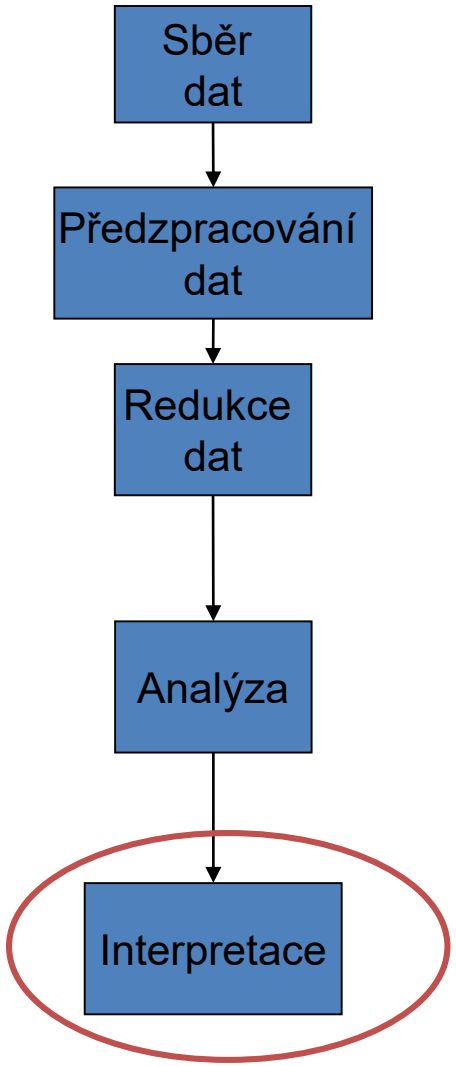
Obsah prezentace

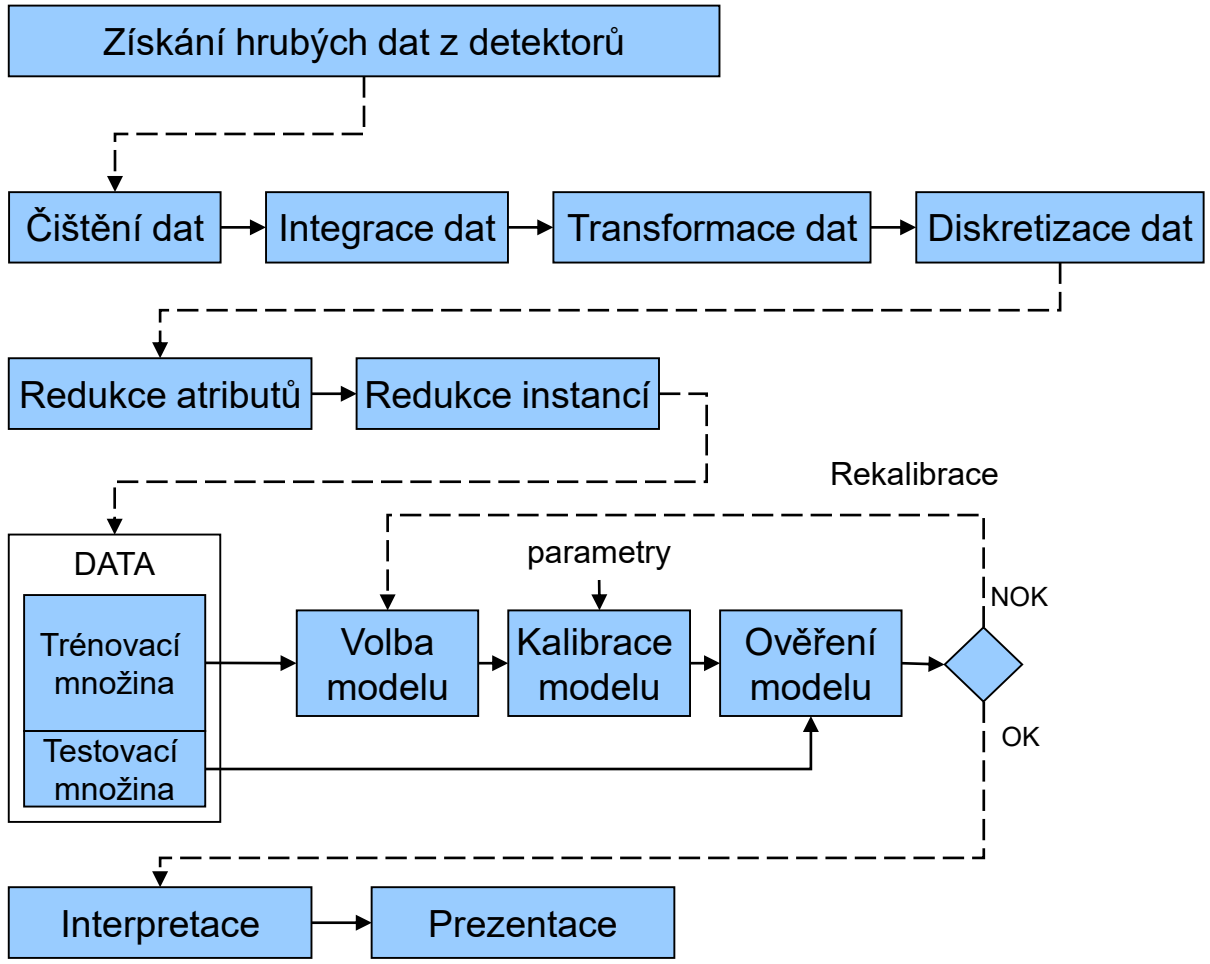
- Měřené veličiny
- Chyby měření
- Vizualizace dat
- Další aspekty analýzy dat
- Hlavní kroky při analýze dat

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

Děkuji za pozornost

Témata ke zkoušce

- Data, informace a znalosti
  - Definice a vztah
- Typy měřených veličin
  - Klasifikace s příklady
- Chyby měření
  - Klasifikace
  - Vzorce
  - Příklady a výpočet
- Populace versus náhodný výběr
- Přehled kroků analýzy dat

- Zdroje:
  - Ondřej Šlapák „Data, informace, znalosti“, ELECTRONIC JOURNAL FOR PHILOSOPHY/2003, ISSN 1211-0442
  - http://www-users.cs.umn.edu/~kumar/dmbook/index.php
  - http://www.physics.unc.edu/~deardorf/uncertainty/definitions.html
  - http://www.csse.monash.edu.au/~smarkham/resources/scaling.htm
  - http://people.math.sfu.ca/~cschwarz/Stat-301/Handouts/node5.html
  - Meloun M., Militký J.: Zpracování experimentálních dat, Plus Praha 1994