AI Akademie

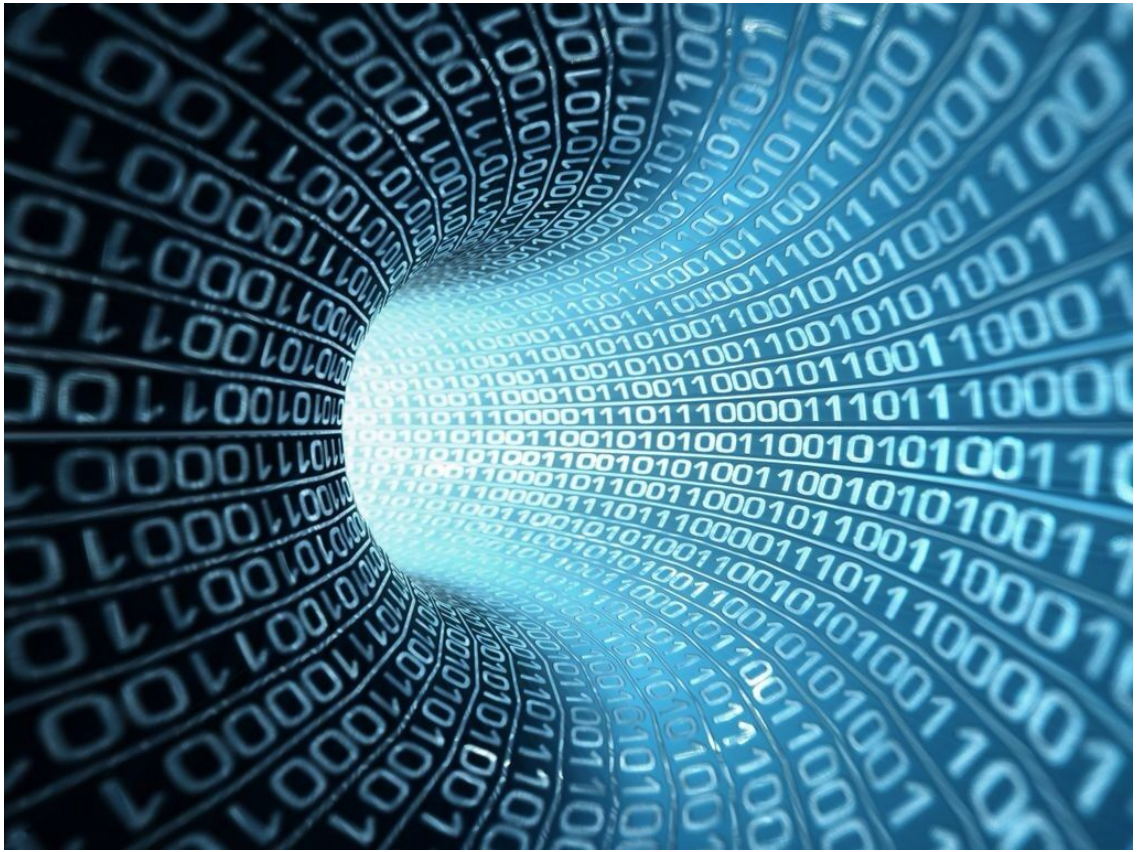
Kapitola 2: Data a informace

Machine Learning College

prg.ai

PRAHA
PRAGUE
PRAGA
PRAG

Data a informace



Data jsou soubory údajů, popisující objekty nebo události.

Může se jednat o čísla, text, obrázky a podobně.

Informace jsou data doplněná o význam a kontext.

Machine Learning College

Data a informace - příklady

- *1603*, *1492*, *1324* jsou číselná **data**. Pokud doplníme, že se jedná o nadmořské výšky Sněžky, Pradědu a Lysé hory v metrech, dostaneme **informaci**.
- *(255, 245, 135)*, *(255, 55, 67)*, *(98, 52, 64)*, … posloupnost trojic čísel je další příklad **dat**. Jestliže se dozvíme, že každá trojice reprezentuje jeden pixel v obrázku slona, dostaneme **informaci**.
- *6d e1 6d 61 20 6d 65 6c 65 20 6d 61 73 6f* posloupnost bajtů v šestnáctkové soustavě je další příklad dat. Pokud doplníme, že je máme interpretovat jako písmena, dostaneme text "*máma mele maso*", což je druh **informace**.

Machine Learning College

Způsoby vzniku dat

- **měření a snímání** (audiovizuální záznamy, srdeční puls měřený chytrými hodinkami apod.)
- **ruční tvorba** (ruční přepis skenů dokumentů na text, ruční vyznačení objektů na obrázku apod.)
- **průzkumy** (volební preference, uživatelská hodnocení produktů)
- **generování** (texty, obrázky, videa nebo zvukové stopy vzniklé pomocí tzv. *deep fake* technik)
- **a další**

Machine Learning College

Otevřená data

Otevřená data jsou úplná, snadno dostupná a strojově čitelná data, zveřejněná na internetu. Pravidla jejich použití jsou co nejvolnější a jasně definovaná.

Hlavními poskytovateli otevřených dat jsou veřejné instituce.

Přehled dostupných otevřených dat státní správy lze nalézt zde: https://data.gov.cz/

Machine Learning College

Strukturovaná a nestrukturovaná data

Strukturovaná

Počítačem snadno zpracovatelná data, která mají jasnou strukturu.

Například tabulky v databázi, data v buňkách Excelu, strukturované formáty souborů jako CSV apod.

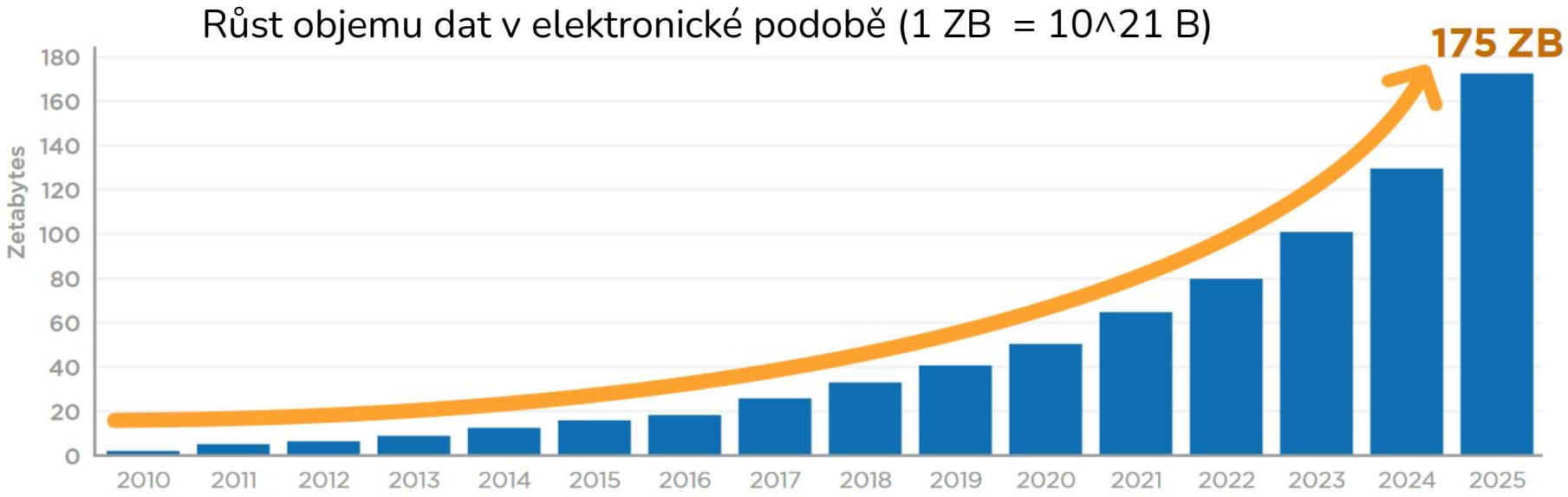
Nestrukturovaná

Data bez struktury. Pro člověka typicky srozumitelná, ale pro počítač obtížně zpracovatelná.

Například obrázky, videa, texty apod.

Machine Learning College

Big data



Velké objemy dat, které se typicky nevejdou na jeden počítač a není možné je tak zpracovávat tradičními způsoby.

Machine Learning College

Data a výpočty versus intuice

Intuice a pocity nás velmi často klamou. Proto je výhodnější dělat důležitá rozhodnutí na základě faktů a dat.

Příkladem “datově řízených” firem jsou úspěšné společnosti jako Google, Facebook nebo Amazon. Rozhodování na základě dat je však důležité i mimo business ve většině oblastí lidského bytí.

Machine Learning College

Příklady selhání intuice

Představme si, že natáhneme provaz kolem rovníku celé Země. Pokud budeme považovat Zemi za přesnou kouli s poloměrem r = 6 378 km, bude délka provazu přibližně 40 074 km.

Bez dlouhého přemýšlení a počítání zkuste odhadnout, jak se délka lana prodlouží, pokud budeme chtít lano táhnout všude 1 m nad povrchem Země.

a) 10 m
b) 10 km
c) 1000 km

Machine Learning College

Příklady selhání intuice

Představme si, že natáhneme provaz kolem rovníku celé Země. Pokud budeme považovat Zemi za přesnou kouli s poloměrem r = 6 378 km, bude délka provazu přibližně 40 074 km.

Bez dlouhého přemýšlení a počítání zkuste odhadnout, jak se délka lana prodlouží, pokud budeme chtít lano táhnout všude 1 m nad povrchem Země.

$$p = 2\pi(r + 0.001) - 2\pi r = 2\pi \cdot 0.001 \approx 0.0063$$

Velikost prodloužení (označené p) bude 6.3 m a vůbec nezávisí na poloměru Země!

Machine Learning College

Příklady selhání intuice

Ve třídě sedí 30 žáků. Jaká je šance, že alespoň 2 žáci ze třídy mají narozeniny ve stejný den?

a) 20 %
b) 50 %
c) 70 %

Machine Learning College

Příklady selhání intuice

Ve třídě sedí 30 žáků. Jaká je pravděpodobnost, že alespoň 2 žáci ze třídy mají narozeniny ve stejný den?

Problém si pro jednoduchost otočíme a bude počítat pravděpodobnost situace, kdy ve třídě nejsou žádní dva žáci se stejným datem narození:

$$p = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{336}{365} \approx 0.294$$

Původně požadovaná pravděpodobnost je tedy
1 - 0.294 = **0.706**

Machine Learning College

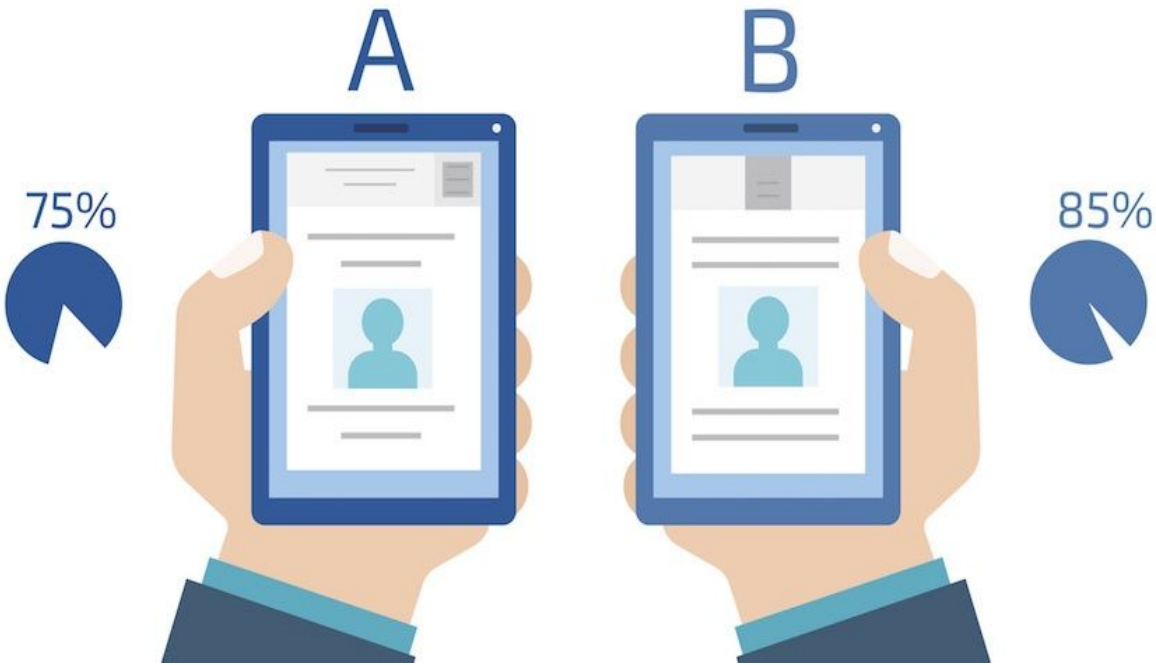
Další příklady selhání intuice

- Začátkem roku 2020 málokdo připouštěl možnost devastujícího dopadu pandemie covid-19 na celý svět, přestože data o šíření nemoci to již jasně ukazovala. Důvodem byla absence podobné zkušenosti většiny z nás a popírání něčeho do té doby tak nepředstavitelného.
- Sociální bubliny, ve kterých žijeme, způsobují velmi zkreslené vidění světa. Jeden příklad za všechny - *“Nechápu, jak mohl být zvolen politik XYZ. Neznám jediného člověka, který ho volil.”*
- Jedním ze zdrojů rasismu či jiné skupinové diskriminace může být zobecnění negativní zkušenosti na celou skupinu lidí s podobnými rysy, které však se zdrojem negativní zkušenosti nijak nesouvisí.

Machine Learning College

Příklady využití dat k rozhodování v praxi

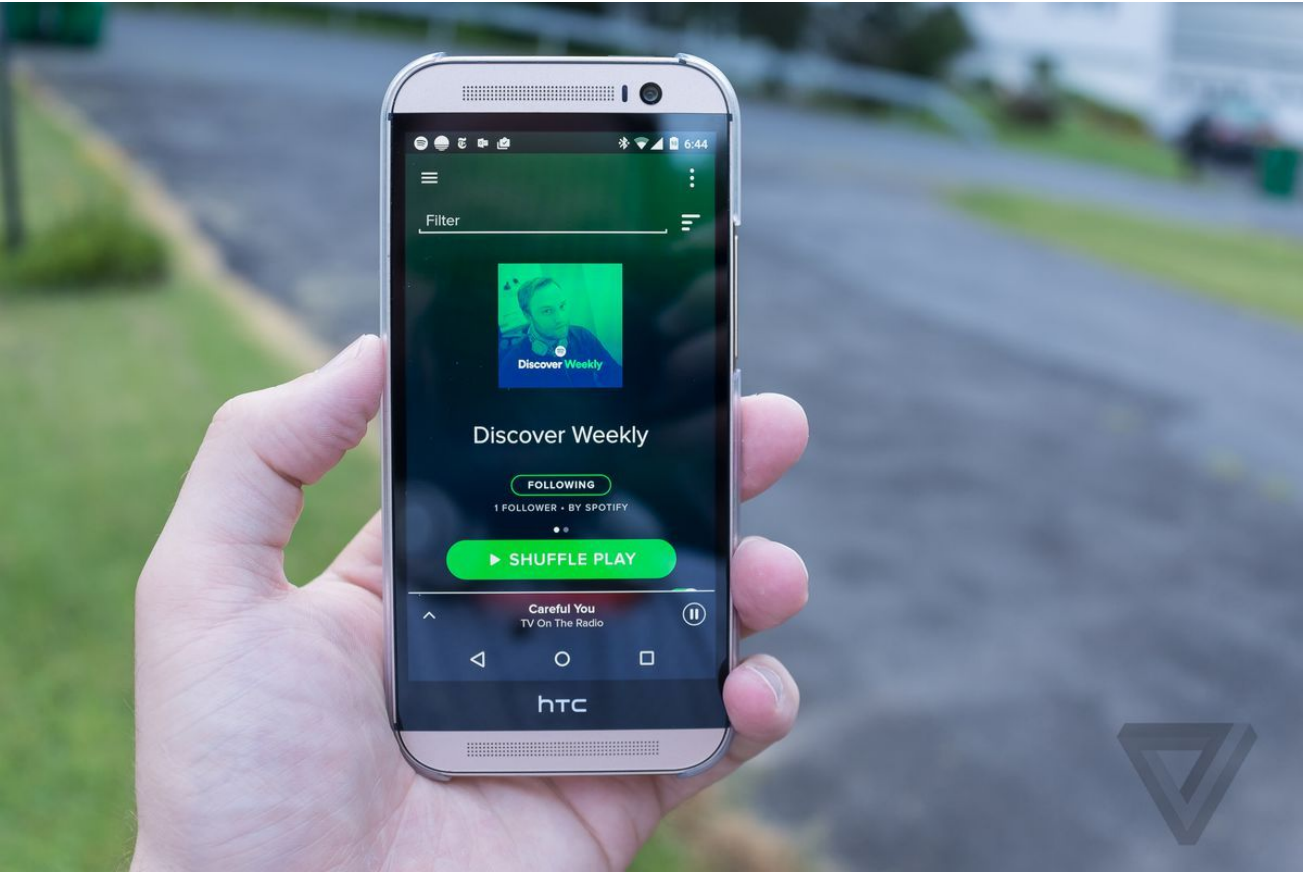
A/B testování verzí produktu



Po určitou dobu se uživatelům náhodně zobrazují dvě nebo více verzí produktu (např. designu e-shopu) a sbírají se data o chování uživatelů. Poté se vybere nejúspěšnější varianta (např. na základě prodejů), která se použije pro všechny uživatele.

Machine Learning College

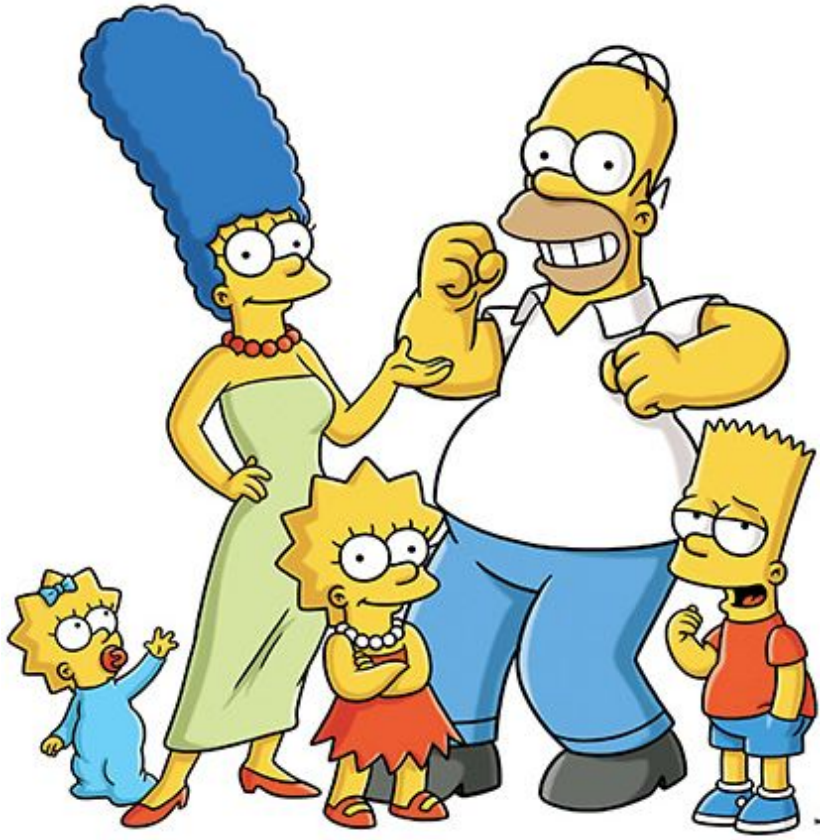
Příklady využití dat k rozhodování v praxi



Personalizované doporučování obsahu ve streamovacích aplikacích (Spotify, Youtube, Netflix, apod.)

Machine Learning College

Příklady využití dat k rozhodování v praxi



Natáčení nových dílů televizních seriálů na základě dat o sledovanosti a prodeji.

První díl seriálu Simpsonovi byl vytvořen v roce 1987 a točí se dodnes.

Machine Learning College

Popisná statistika dat

Data jsou typicky příliš velká na to, aby je člověk dokázal interpretovat přímo. Pro základní porozumění datům slouží *popisná statistika*.

Mezi nejčastěji používané statistické charakteristiky patří:

- průměr (střední hodnota)
- medián a další kvantily
- rozptyl
- směrodatná odchylka

Machine Learning College

Popisná statistika - průměr (střední hodnota)

Vzorek dat (měsíční mzda v ČR náhodné skupiny lidí v tis. Kč): [21, 38, 31, 34, 180, 18, 41, 39, 32, 29]

Výpočet aritmetického průměru datového vzorku

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{1}{10}(21 + 38 + \cdots + 29) = 46.3$$

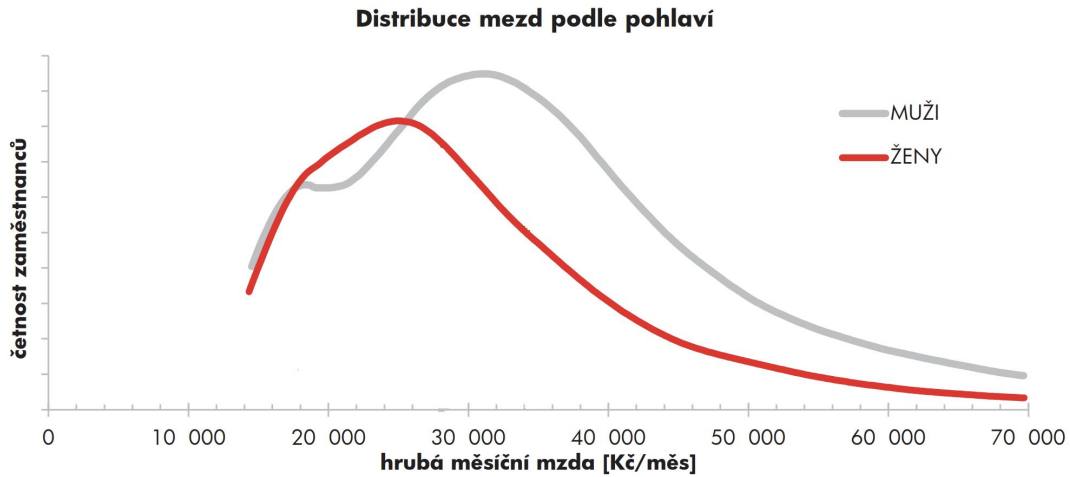
Machine Learning College

Popisná statistika - průměr (střední hodnota)

Vzorek dat (měsíční mzda v ČR náhodné skupiny lidí v tis. Kč):

[21, 38, 31, 34, 180, 18, 41, 39, 32, 29]

Aritmetický průměr **není** vhodným ukazatelem výše mezd ve společnosti. Většina lidí má relativně nízké mzdy a velmi málo lidí má mzdy velmi vysoké. To způsobuje nadhodnocení průměru a iluzi bohatší společnosti.



Distribuce mezd v ČR podle pohlaví za rok 2020. Zdroj: https://ispv.cz/

Machine Learning College

Popisná statistika - medián

Vzorek dat (měsíční mzda v ČR náhodné skupiny lidí v tis. Kč): [21, 38, 31, 34, 180, 18, 41, 39, 32, 29]

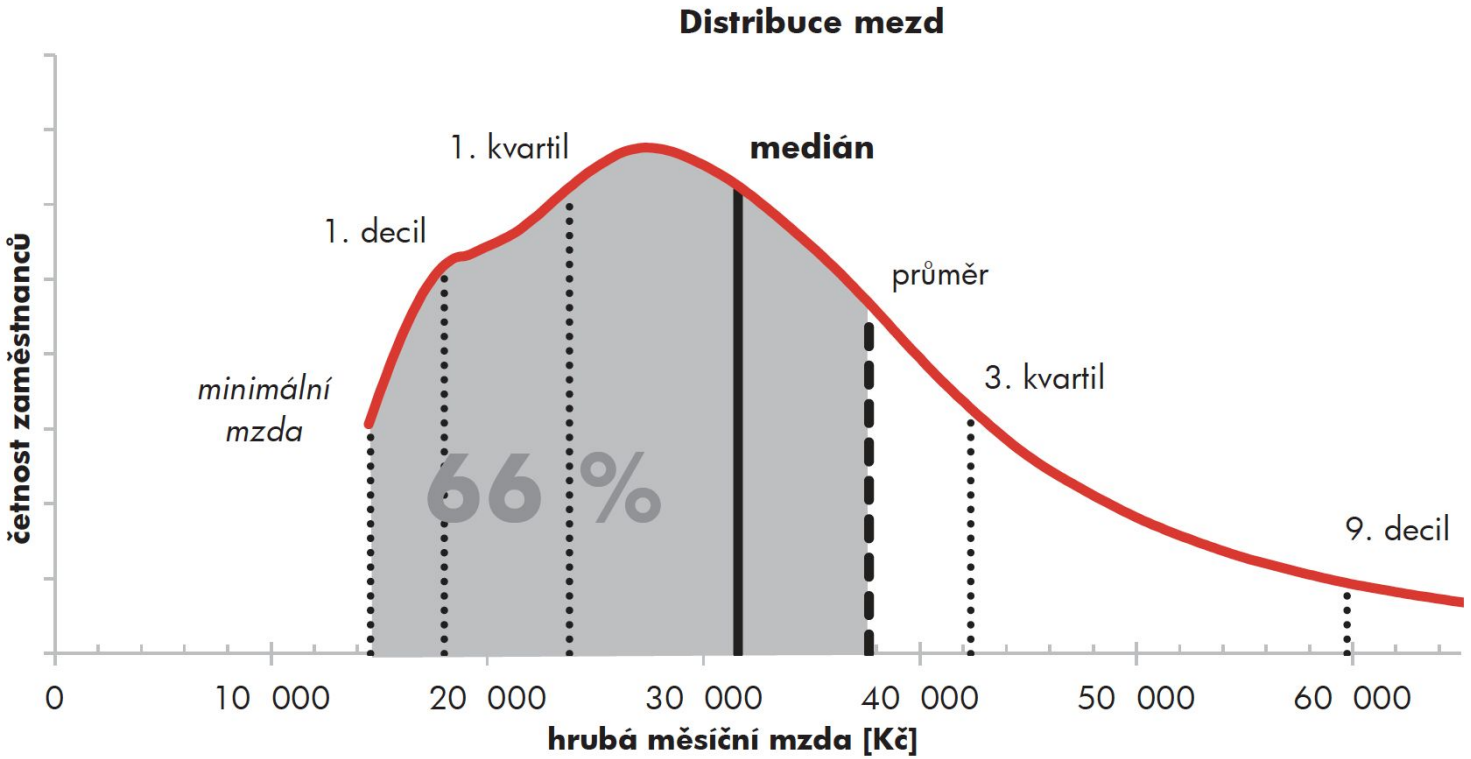
Medián je hodnota, jež dělí řadu vzestupně seřazených výsledků na dvě stejně početné poloviny. V případě sudého počtu prvků je medián průměrem dvou prostředních. Je to mnohem vhodnější ukazatel mezd ve společnosti než aritmetický průměr.

Pokud bychom řadu rozdělili jinde než v polovině, dostaneme další statistické ukazatele, nazývané obecně **kvantily**.

Medián našeho vzorku je 33.

Machine Learning College

Popisná statistika - distribuce mezd v ČR za 2020



Zdroj: https://ispv.cz/

Machine Learning College

Popisná statistika - rozptyl

Vzorek dat (měsíční mzda v ČR náhodné skupiny lidí v tis. Kč): [21, 38, 31, 34, 180, 18, 41, 39, 32, 29]

Při zkoumání vzorku dat je pro nás důležitou informací jeho variabilita. Průměrné odchylky od střední hodnoty (aritmetického průměru) zkoumají statistické ukazatele **rozptyl** a **směrodatná odchylka**.

Rozptyl (σ^2) je definovaný jako průměr kvadrátů odchylek od průměru.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{10}\sum_{i=1}^{n}(x_i - 46.3)^2 = 2035.61$$

Machine Learning College

Popisná statistika - směrodatná odchylka

Vzorek dat (měsíční mzda v ČR náhodné skupiny lidí v tis. Kč): [21, 38, 31, 34, 180, 18, 41, 39, 32, 29]

Při zkoumání vzorku dat je pro nás důležitou informací jeho variabilita. Průměrné odchylky od střední hodnoty (aritmetického průměru) zkoumají statistické ukazatele **rozptyl** a **směrodatná odchylka**.

Směrodatná odchylka (σ) je definovaná jako odmocnina rozptylu.

$$\sigma = \sqrt{\sigma^2} = \sqrt{2035.61} \approx 45.11$$

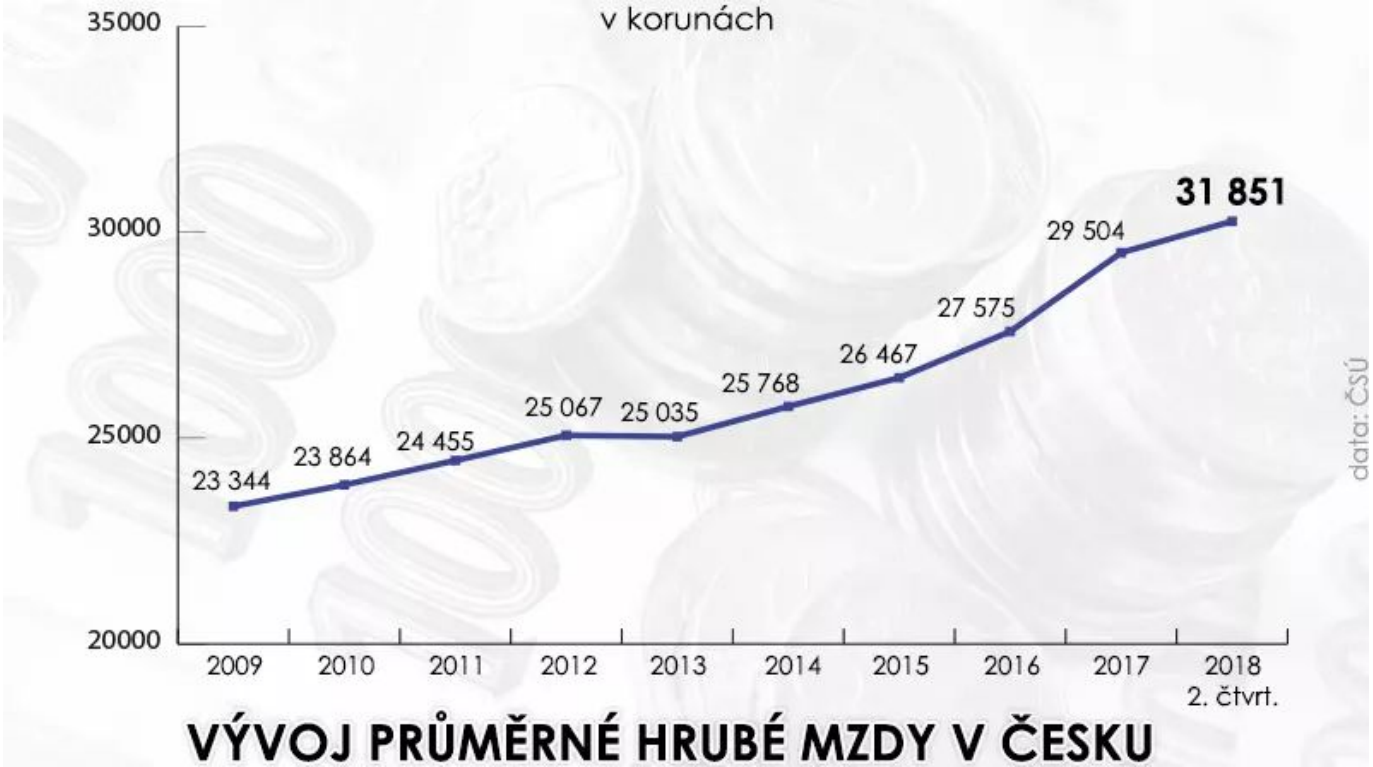
Machine Learning College

Vizualizace dat

- Liniový (spojnicový) graf
- Mapa
- Koláčový graf
- Sloupcový graf
- Krabicový graf

Machine Learning College

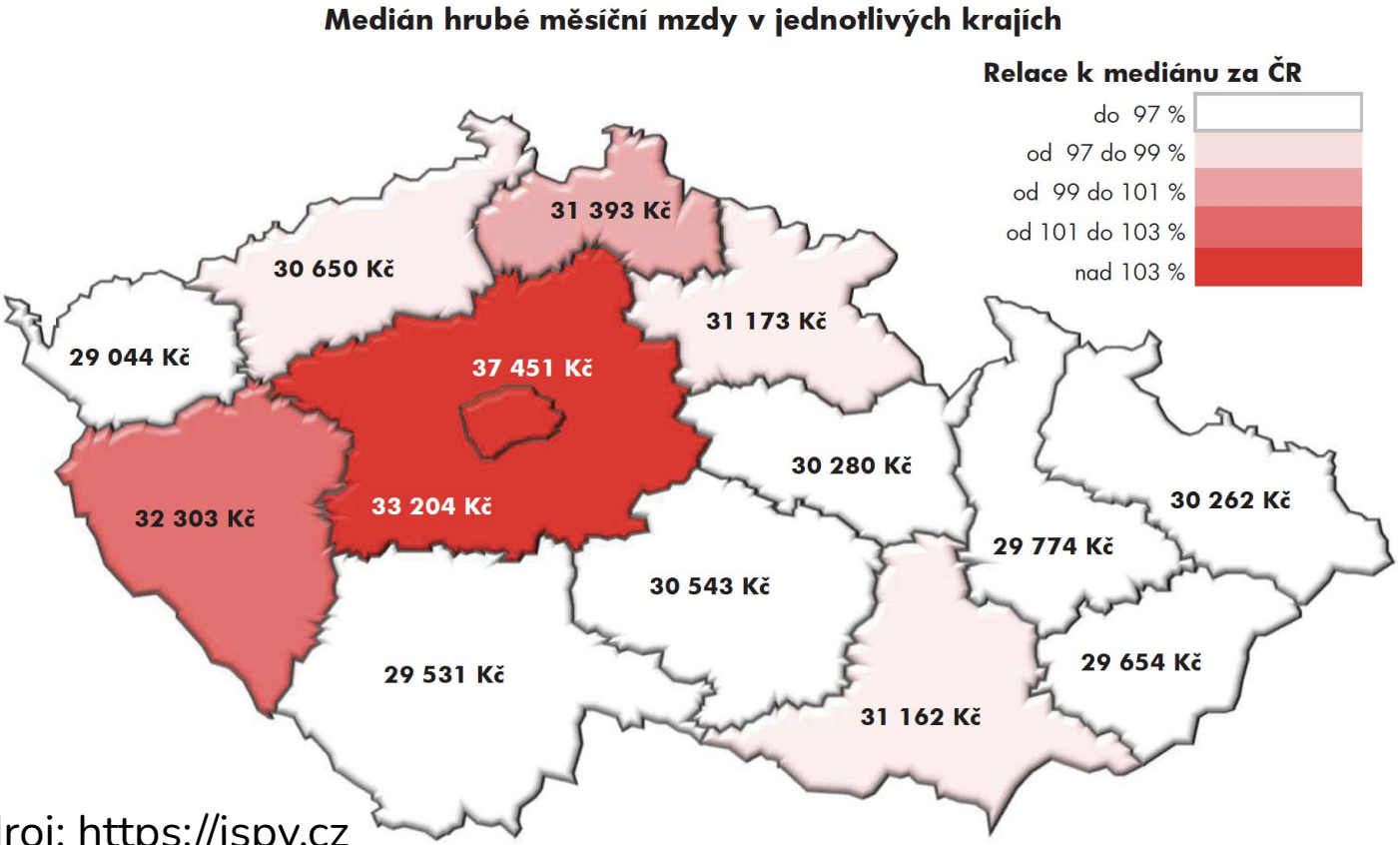
Liniový (spojnicový) graf



Zdroj: Český statistický úřad

Machine Learning College

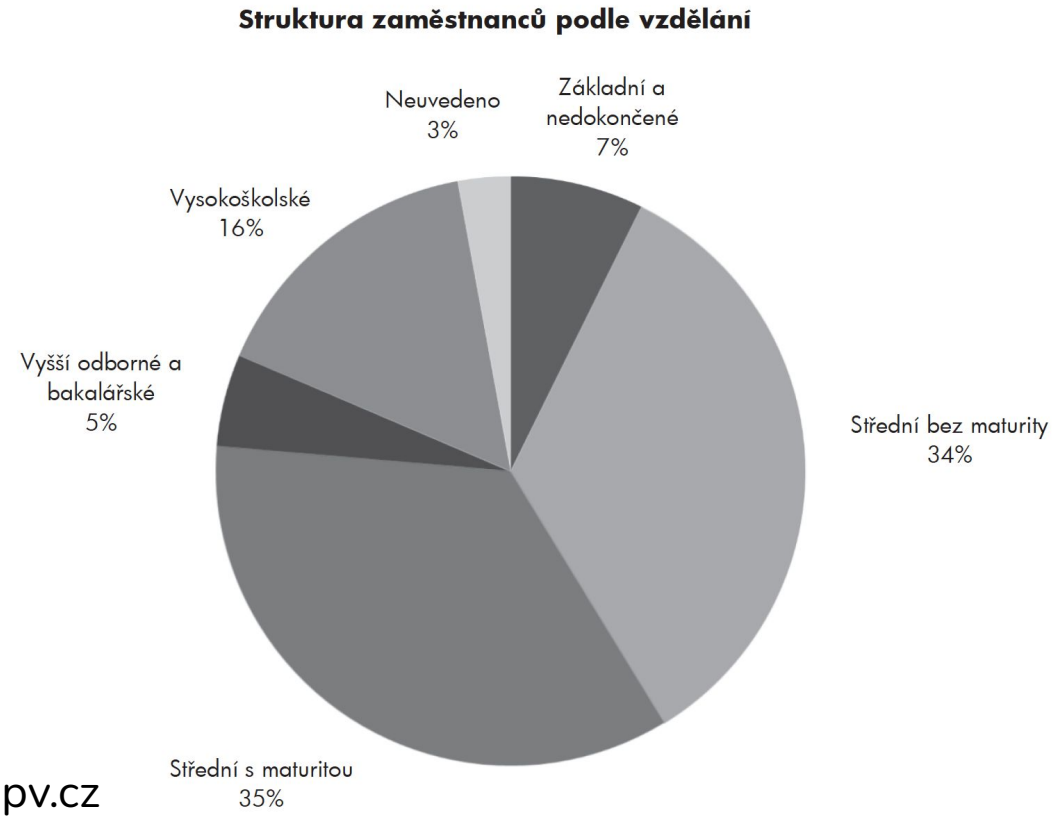
Vizualizace dat - mapa



Zdroj: https://ispv.cz

Machine Learning College

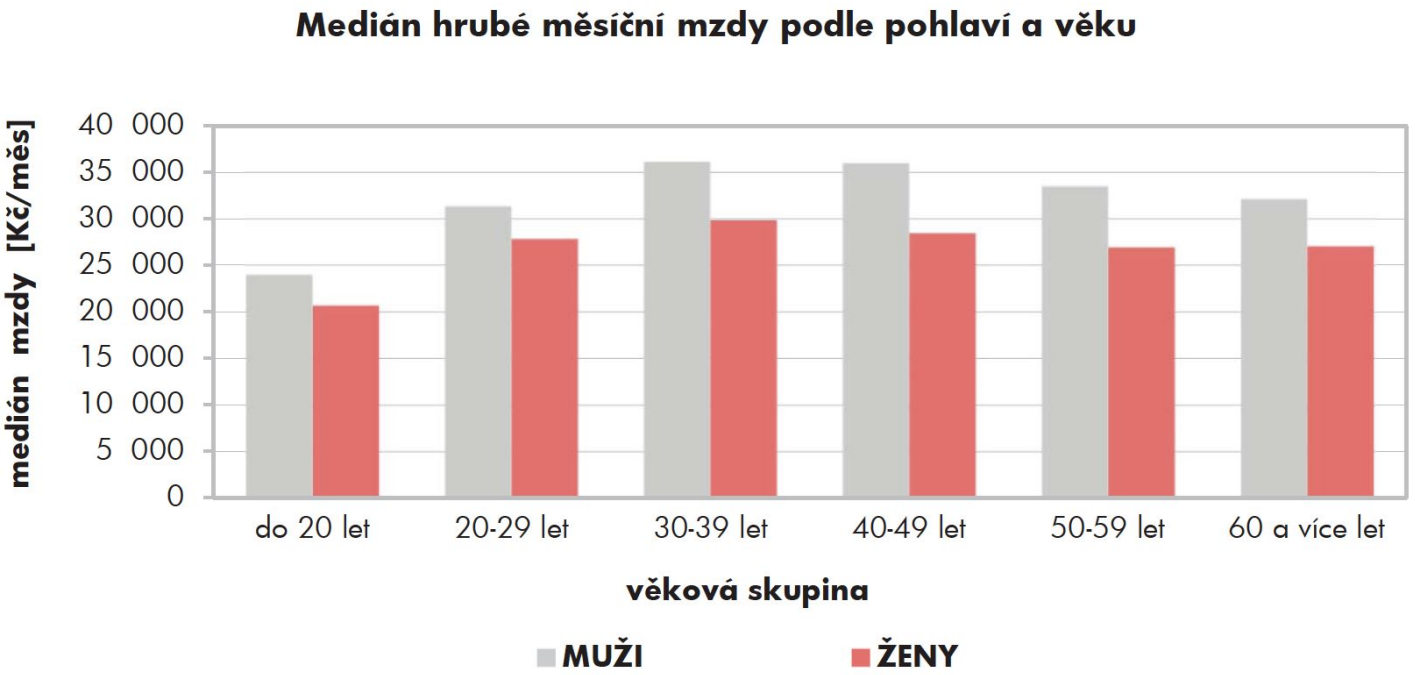
Vizualizace dat - koláčový graf



Zdroj: https://ispv.cz

Machine Learning College

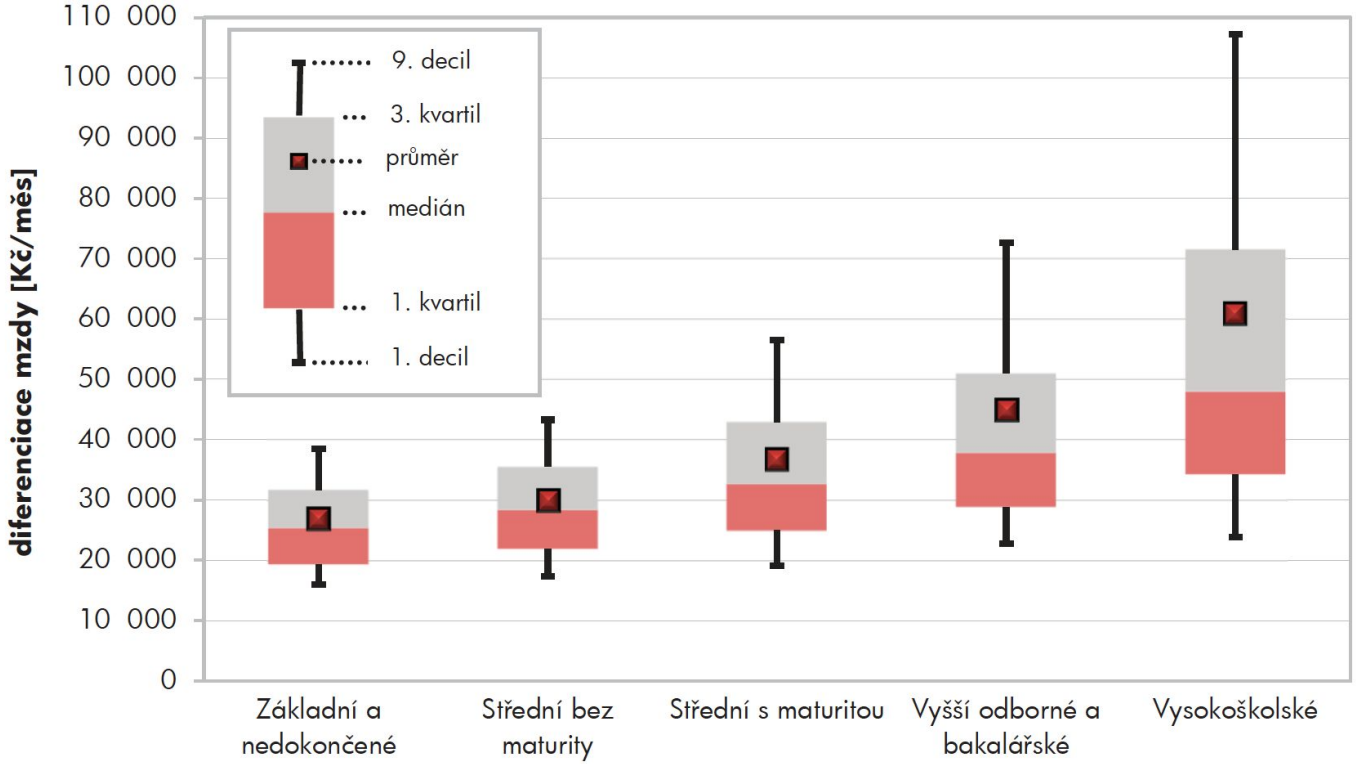
Vizualizace dat - sloupcový graf



Zdroj: https://ispv.cz

Machine Learning College

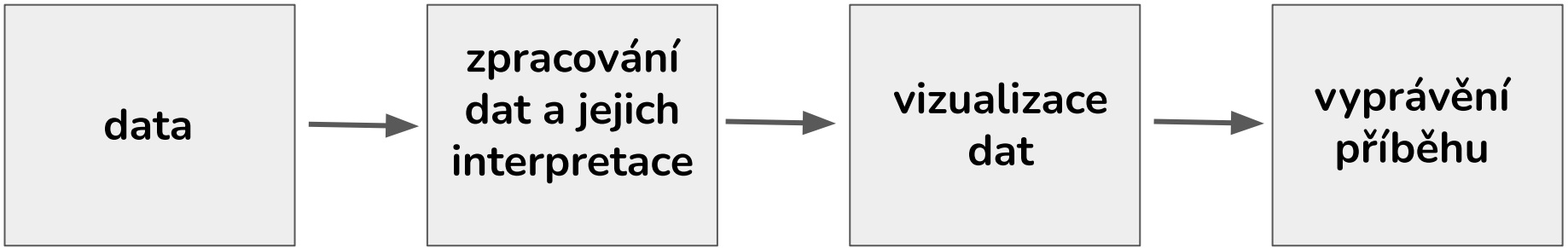
Vizualizace dat - krabicový graf



Zdroj: https://ispv.cz

Machine Learning College

Datová žurnalistika



Ukázky kvalitní datové žurnalistiky například zde:

https://www.irozhlas.cz/zpravy-tag/datova-zurnalistika

Machine Learning College

Chybná interpretace dat

- Reprezentativita dat
- Šum a chyby v datech
- Bias v datech
- Změny podmínek při sběru dat
- Korelace a kauzalita

Machine Learning College

Reprezentativita dat

Jestliže máme pro popis zkoumaného jevu málo dat a data daný jev nereprezentují dostatečně, říkáme, že data nejsou *reprezentativní*.

Příkladem nedostatečně reprezentativních dat může být náš vzorek mezd 10 lidí (je jich málo).

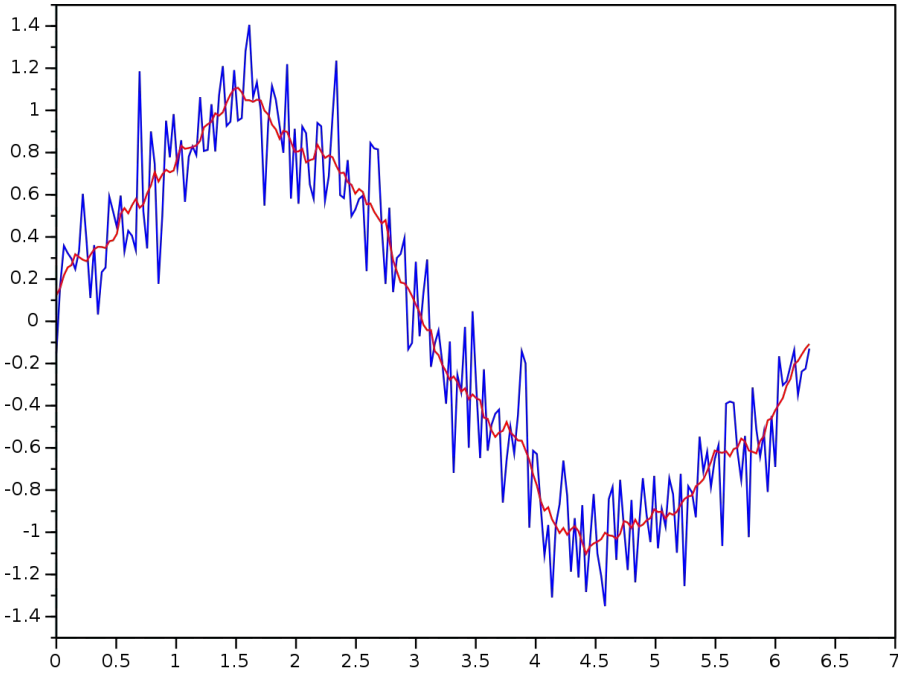
[21, 38, 31, 34, 180, 18, 41, 39, 32, 29]

Aby byla data reprezentativní, je třeba mít vzorek co největší a správně vybraný.

Machine Learning College

Šum a chyby v datech

Šum - náhodná chyba, která ovlivňuje data. Pokud šum není příliš velký, tak nevadí.



Machine Learning College

Bias v datech

Bias je systematická chyba, která ovlivňuje data. Tento druh chyby nám vadí, protože může výrazně ovlivňovat globální statistiky jako průměr, medián apod.

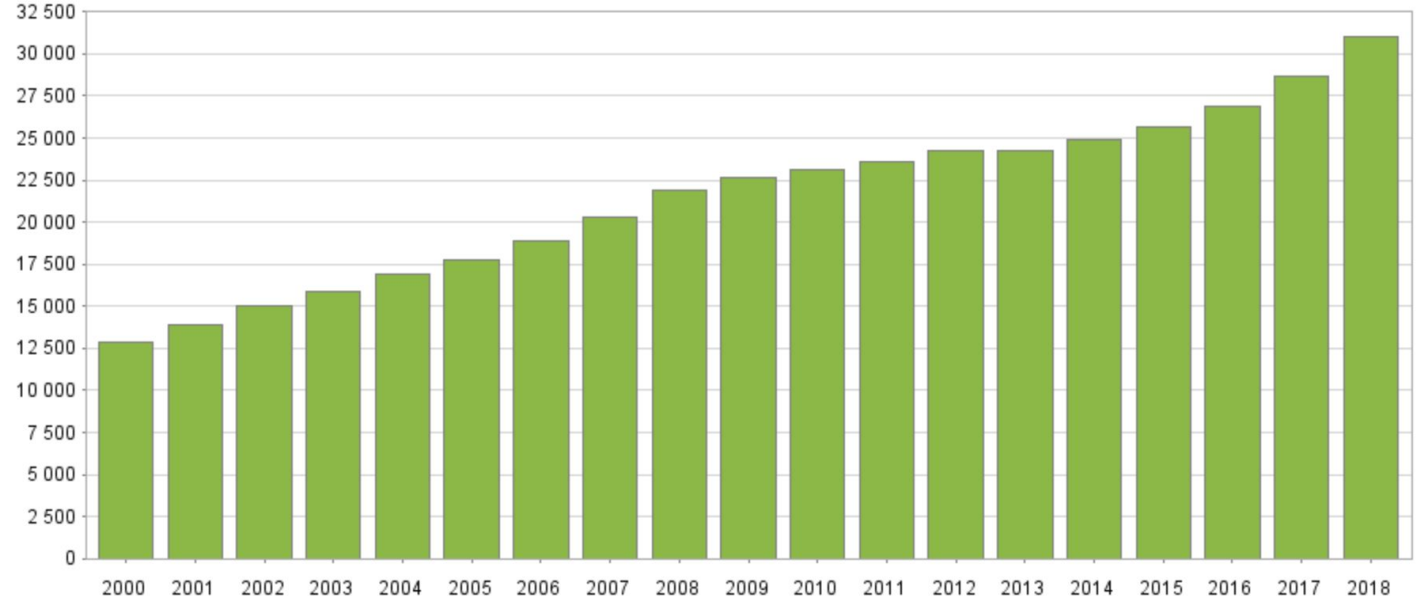
Bias může být způsobený i nevhodným výběrem vzorku dat (**výběrové zkreslení**).

Příkladem může být, pokud bychom do dat pro počítání statistik o mzdách v ČR vybrali pouze absolventy vysokých škol žijící v Praze.

Machine Learning College

Změny podmínek při sběru dat

Některé veličiny a ukazatele (například průměrná mzda) se vyvíjí v čase. Nemůžeme tedy např. sbírat data o mzdách v roce 2010 a dělat z nich závěry o mzdách v roce 2021.



Zdroj: Český statistický úřad

Machine Learning College

Korelace a kauzalita

Korelace vyjadřuje závislost mezi dvěma veličinami.

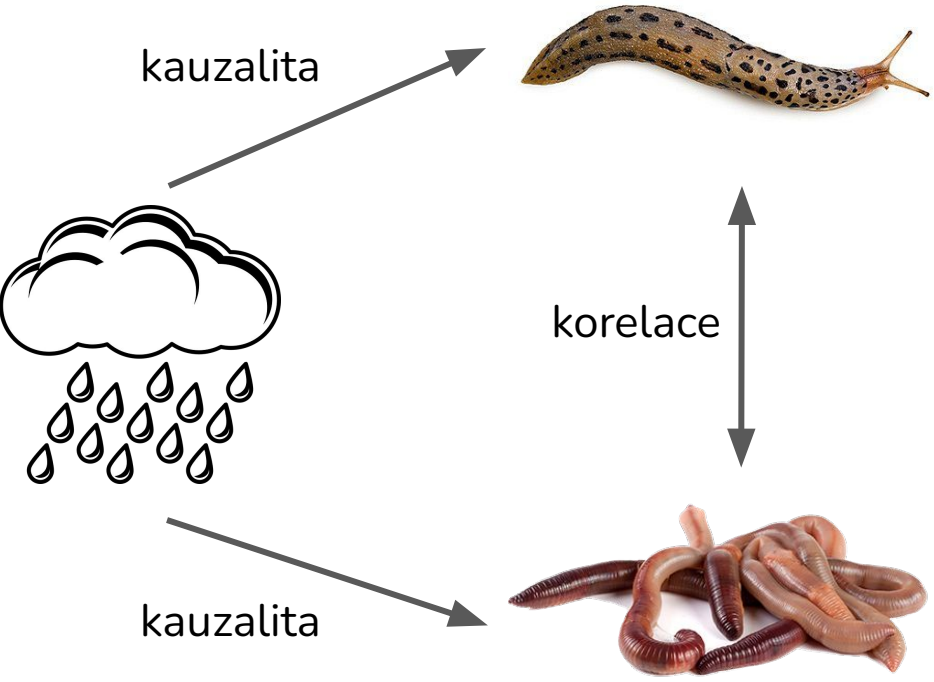
Například četnost výskytu slimáků a žížal na zahrádce spolu silně korelují. Pokud najdeme velké množství žížal, máme velkou šanci najít velké množství slimáků a naopak.

Kauzalita je vztahem dvou veličin, kde hodnota jedné přímo ovlivňuje hodnotu druhé.

Příčinou zvýšeného výskytu slimáků a žížal je většinou deštivé počasí. Zvýšený výskyt slimáků sám o sobě zvýšený výskyt žížal nezpůsobuje. Stejně tak zvýšený výskyt žížal nezpůsobuje zvýšený výskyt slimáků.

Machine Learning College

Korelace není kauzalita



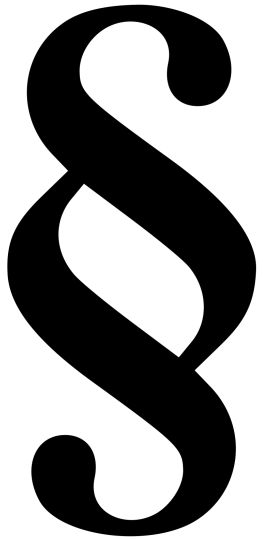
Praxi je někdy obtížné odlišit korelaci a kauzalitu. Je to však nesmírně důležité.

Dlouho například nebylo prokázáno, že kouření způsobuje rakovinu, přestože bylo jasné, že spolu korelují.

Machine Learning College

Osobní data a GDPR

GDPR (angl. General Data Protection Regulation) je regulace Evropské Unie, která zajišťuje ochranu osobních údajů v evropském prostoru. Osobní údaje je možné sbírat pouze se souhlasem daných osob a jejich uchování je možné pouze na nezbytně dlouhou dobu.



Osobní údaje jsou například jméno, adresa, datum narození, e-mail nebo IP adresa.

Zavedení GDPR v roce 2018 sice zkomplikovalo sběr a skladování dat, na druhou stranu ale zvýšilo ochranu osob před zneužitím jejich osobních údajů.

Machine Learning College

Úloha k procvičení - počty podlaží budov v Praze

- stáhněte si CSV soubor podlaznost_praha_2021.csv, obsahující informace o počtech podlaží budov v Praze. Originální zdroj pochází z otevřených dat: https://opendata.praha.eu/dataset/ipr-podlaznosti

- Data si importujte do svého oblíbeného tabulkového procesoru nebo jiného nástroje.

- Zjistěte jaký je průměr, medián a směrodatná odchylka počtu podlaží domů v Praze.

- Data vhodným způsobem vizualizujte

Machine Learning College