Inteligentná analýza údajov

Róbert Móro, Jakub Ševcech

20.9.2018

https://tinyurl.com/iau2018-19

Úvod do inteligentnej analýzy údajov

Úvod do dátovej vedy

Dátový vedec: Najsexy povolanie 21. storočia*

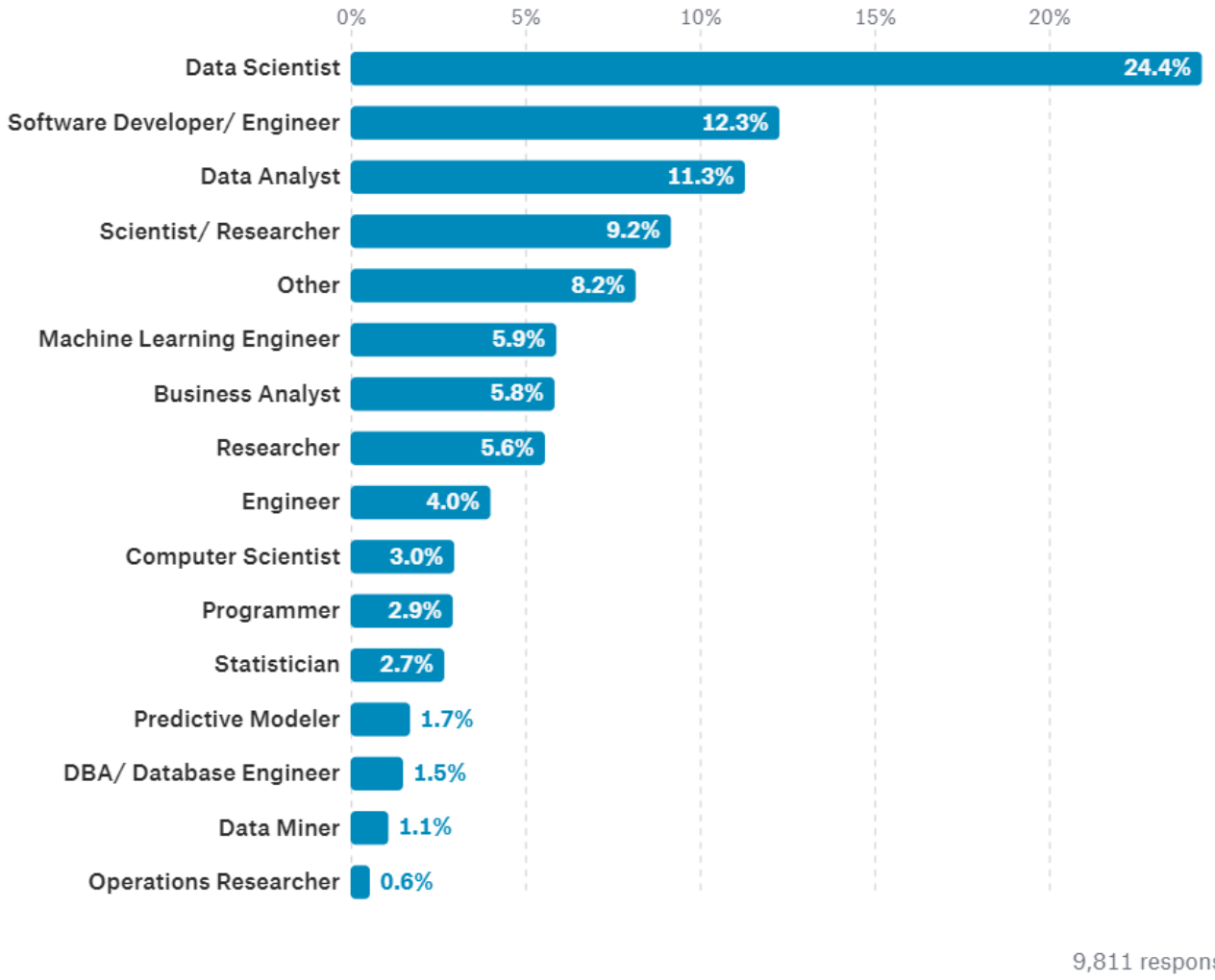
* Thomas H. Davenport, D.J. Patil: *Data Scientist: The Sexiest Job of the 21st Century*, Harvard Business Review, 2012

Neexistuje presná definícia, čo je to dátová veda. Dátový vedec je...

...pokus o rebranding; niekto, kto ovláda štatistiku lepšie ako ktorýkoľvek softvérový inžinier a je lepší v softvérovom inžinierstve ako ktorýkoľvek štatistik

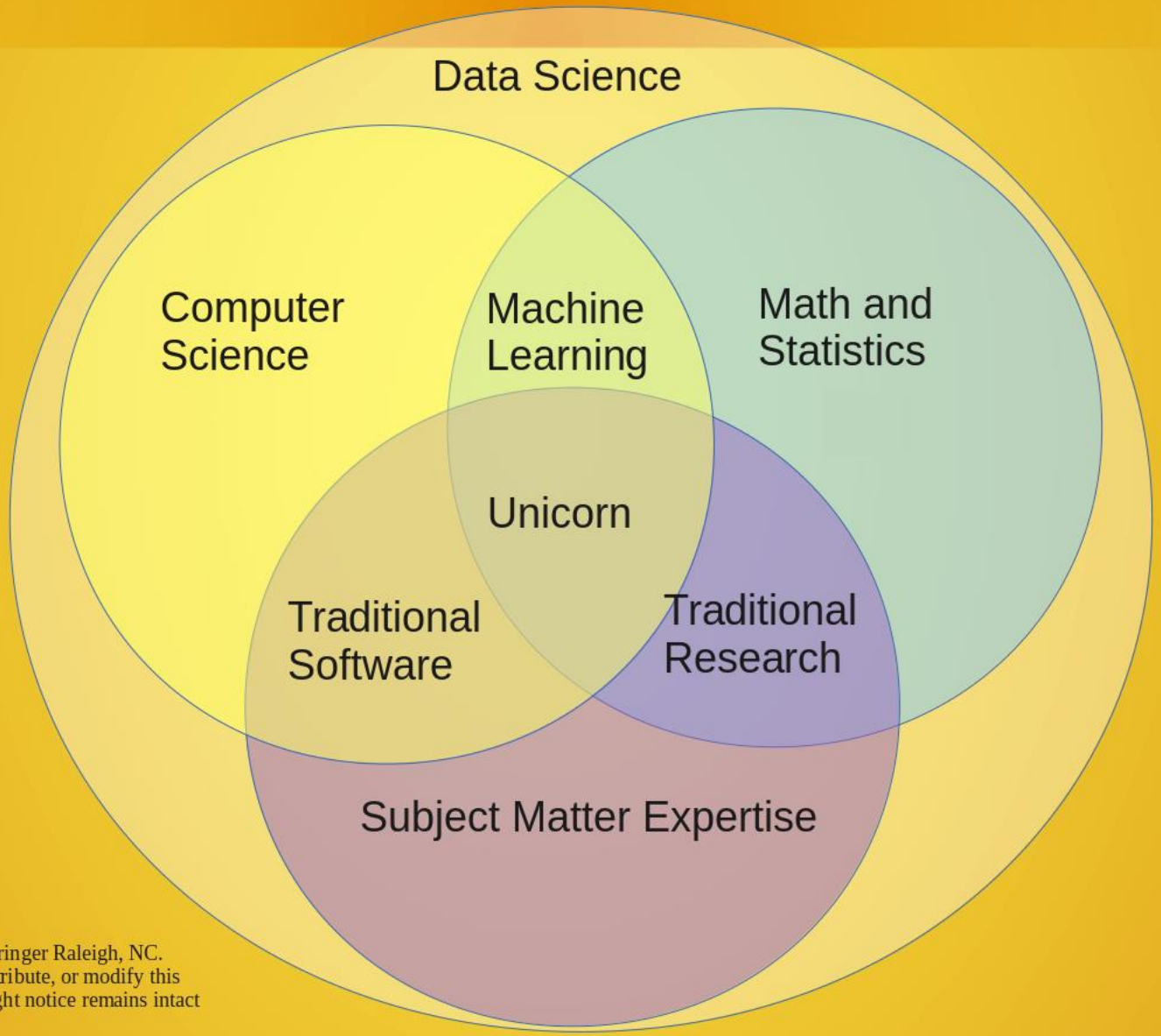
https://www.quora.com/What-is-data-science/answer/Rahul-Agarwal-10

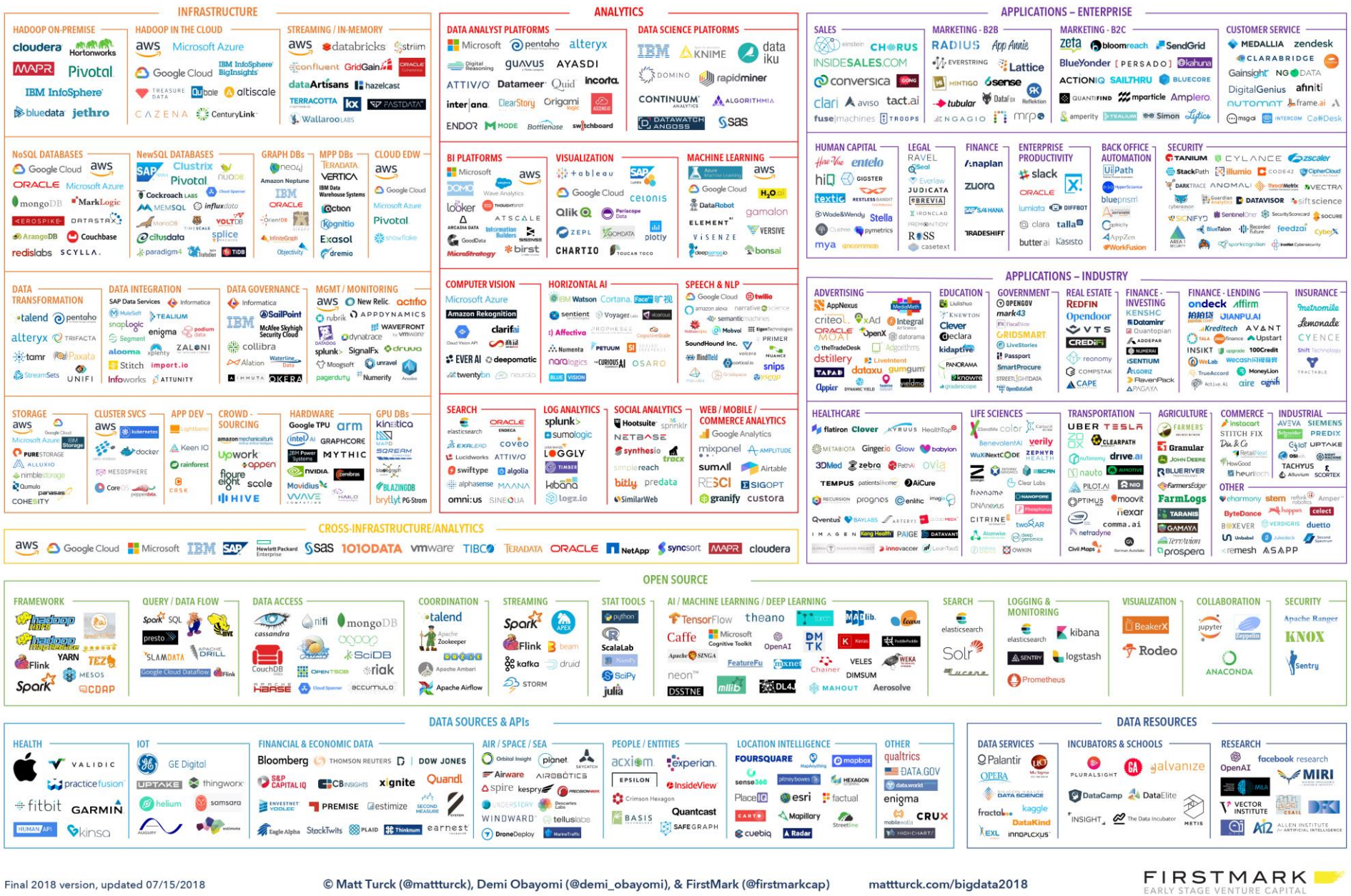
Ani pracovná pozícia sa tak vždy nemusí volať



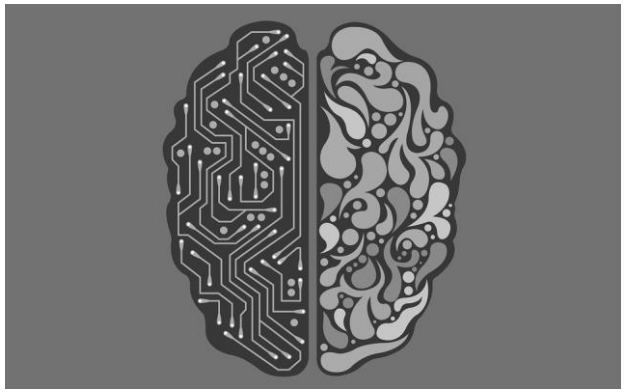
Zdroj: The State of Data Science & Machine Learning. 2017. Kaggle. https://www.kaggle.com/surveys/2017

Data Science Venn Diagram v2.0

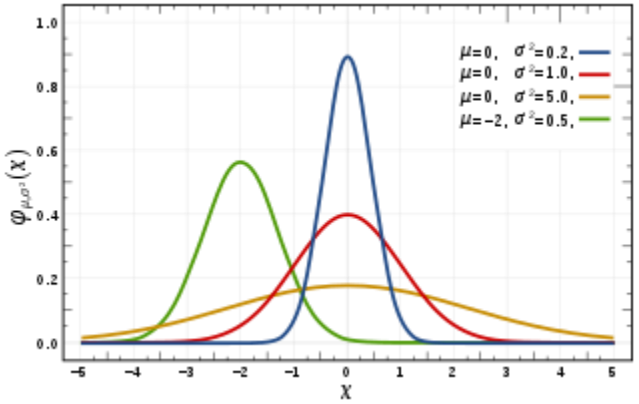




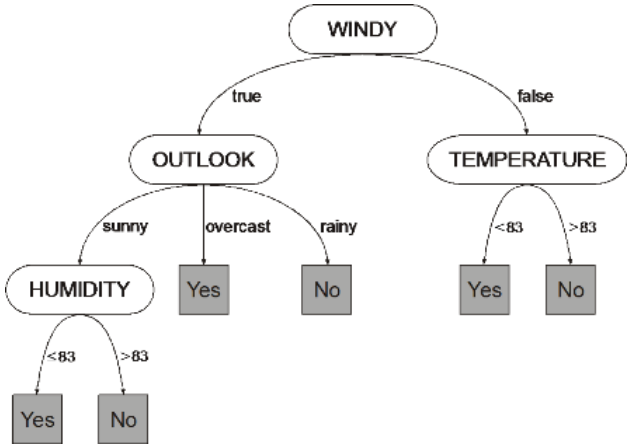
http://mattturck.com/bigdata2018/



Umelá inteligencia

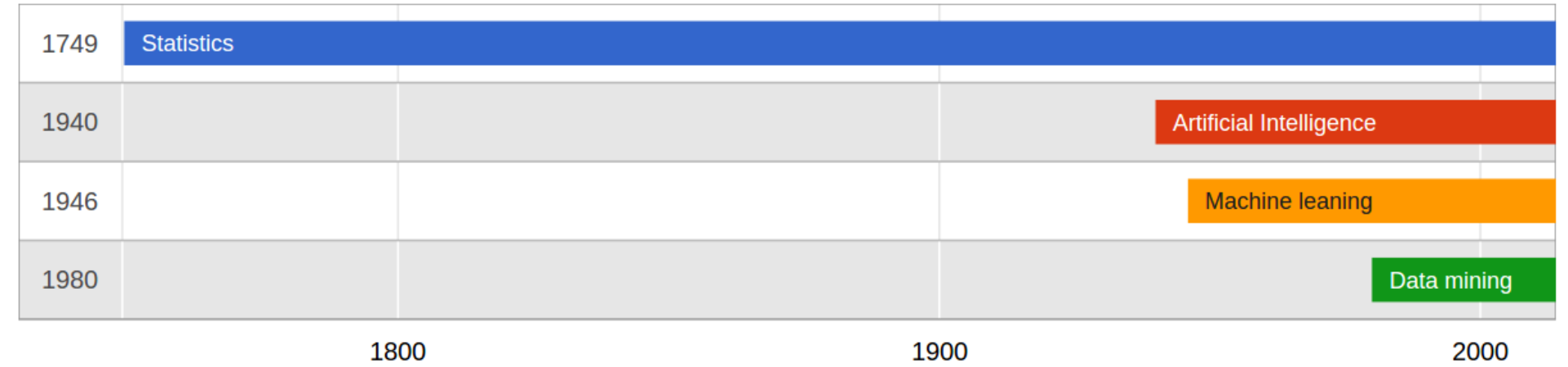


Štatistika



Strojové učenie

Dolovanie v dátach



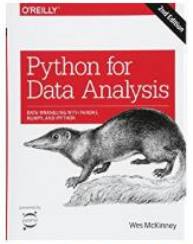
- Ak vyvíjate algoritmy napodobňujúce ľudské správanie (uvažovanie, reprezentácia znalostí, plánovanie a pod.), robíte v oblasti *umelej inteligencie*
- Ak opisujete dáta (deskriptívna štatistika) alebo odvádzate závery o populácii na základe jej vzorky (inferenčná štatistika), tak ste zrejme *štatistik*
- Ak využívate inferenčnú štatistiku na tvorbu algoritmov, ktoré sú schopné samé sa učiť, tak zrejme robíte *strojové učenie*
- Ak využívate strojové učenie na riešenie konkrétneho problému a deskriptívnu štatistiku na opísanie dát a výsledkov, tak zrejme *dolujete v dátach*

Zdroj: http://shakthydoss.com/technical/what-is-the-difference-between-artificial-intelligence-machine-learning-statistics-and-data-mining/

Príklad: Chceme zlepšiť odporúčač na stránke

Customers who bought this item also bought

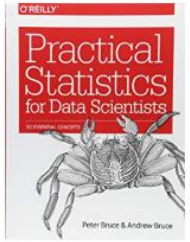
Page 1 of 15



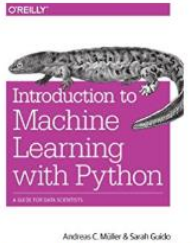
Python for Data Analysis: Data Wrangling with Pandas, NumPy, and…
› Wes McKinney
47
Paperback
$41.45 prime

Hands-On Machine Learning with Scikit-Learn and TensorFlow:…
› Aurélien Géron
227
#1 Best Seller in Data Processing
Paperback
$38.20 prime

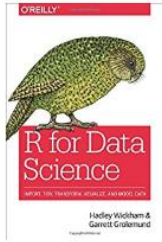
Data Science from Scratch: First Principles with Python
› Joel Grus
108
Paperback
$27.30 prime



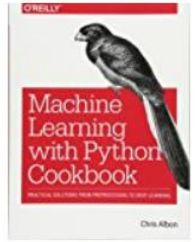
Practical Statistics for Data Scientists: 50 Essential Concepts
› Peter Bruce
42
#1 Best Seller in Data Warehousing
Paperback
$33.99 prime



Introduction to Machine Learning with Python: A Guide for Data Scientists
Andreas C. Müller
37
Paperback
$24.18 prime



R for Data Science: Import, Tidy, Transform, Visualize, and Model Data
› Hadley Wickham
96
#1 Best Seller in Mathematical & Statistical…
Paperback
$18.17 prime



Machine Learning with Python Cookbook: Practical Solutions from…
› Chris Albon
6
Paperback
$46.72 prime

Dátový vedec...

1. Formuluje otázky a hypotézy
   a. Čo to znamená, že aktuálny odporúčač nefunguje (dobre)?
   b. Aká metrika ma zaujíma?
   c. Aké dáta mám k dispozícii? Čo sa zaznamenáva?
   d. Ktoré signály sú dobré na predikciu správania (nákupu/záujmu) používateľa?
2. Definuje ideálnu dátovú sadu
   a. Získava dáta
   b. Čistí ich
   c. Skúma ich vlastnosti (prieskumná analýza)
3. Identifikuje a realizuje vhodný typ analýzy
   a. Návrh a implementácia nového odporúčača
   b. Overenie (porovnanie) s aktuálnym
4. Interpretuje a komunikuje výsledky (dátový produkt)
5. Automatizuje kroky 2-4 pre predikcie na nových dátach
   a. Nasadenie odporúčača v produkčnom prostredí; škálovateľnosť

Práca dátového vedca začína otázkou

- **Deskriptívne**
  - Nerobíme závery, nezovšeobecňujeme
  - *Koľko ľudí volilo konzervatívcov?*
- **Exploratívne**
  - Pozeráme na dáta a hľadáme niečoho nové, čo sme nevedeli, nesnažíme sa to potvrdiť
  - *Ukazujú sa nejaké vzťahy medzi vekom a volebnou účasťou, resp. voľbou konzervatívcov/liberálov?*
- **Inferenčné**
  - Zoberieme malú vzorku dát (pozorovaní) a snažíme sa extrapolovať a zovšeobecniť to pre väčšiu populáciu/vzorku
  - *Vplyv veku na volebnú účasť a voľbu*

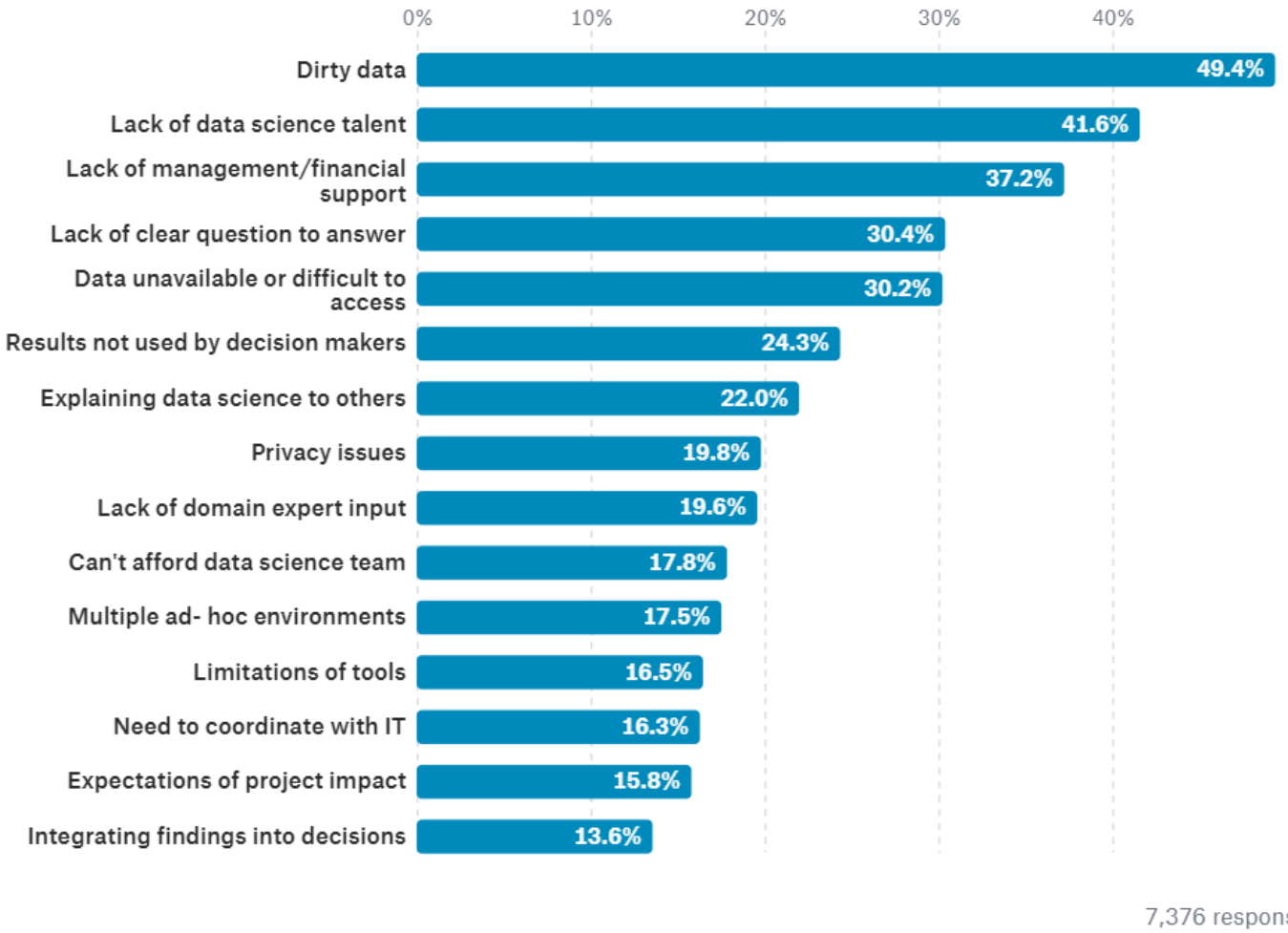
Práca dátového vedca začína otázkou

- **Prediktívne**
  - Snažíme sa použiť dáta, ktoré sme zozbierali o nejakých objektoch, na predpovedanie hodnôt pre nejaký iný objekt
  - Aj keď X predpovedá Y, neznamená to, že X spôsobuje Y
  - *Ako budú voliť prvovoliči v najbližších voľbách?*
- **Kauzálne (preskriptívne)**
  - Skúmame, čo sa stane s premennou X, ak zmeníme premennú Y
  - Na identifikovanie takejto závislosti sú potrebné randomizované štúdie
  - *Zmenila volebná kampaň názory voličov?*

Postupnosť krokov pri strojovom učení (ML Workflow)

- **Definícia problému:** (ne)formálny opis, obmedzenia, manuálne riešenie
- **Integrácia dát:** konsolidácia formátu a štruktúry dát, prepájanie entít (záznamov)
- **Prieskumná analýza:** Deskriptívna štatistika, distribúcie, korelácie, vychýlené hodnoty, ...
- **Predspracovanie dát:** Vzorkovanie, riešenie vychýlených a chýbajúcich hodnôt, ...
- **Tvorba čŕt:** extrakcia, transformácia a výber čŕt
- **Trénovanie, vyhodnocovanie a výber modelov:** zadefinovanie metodológie vyhodnotenia, optimalizácia hyperparametrov, výber optimálneho modelu
- **Prezentácia výsledkov:** vizualizácia výsledkov, obmedzenia navrhnutého riešenia, opis a zverejnenie datasetu, nasadenie a prevádzka

Dátoví vedci najviac zápasia so špinavými dátami



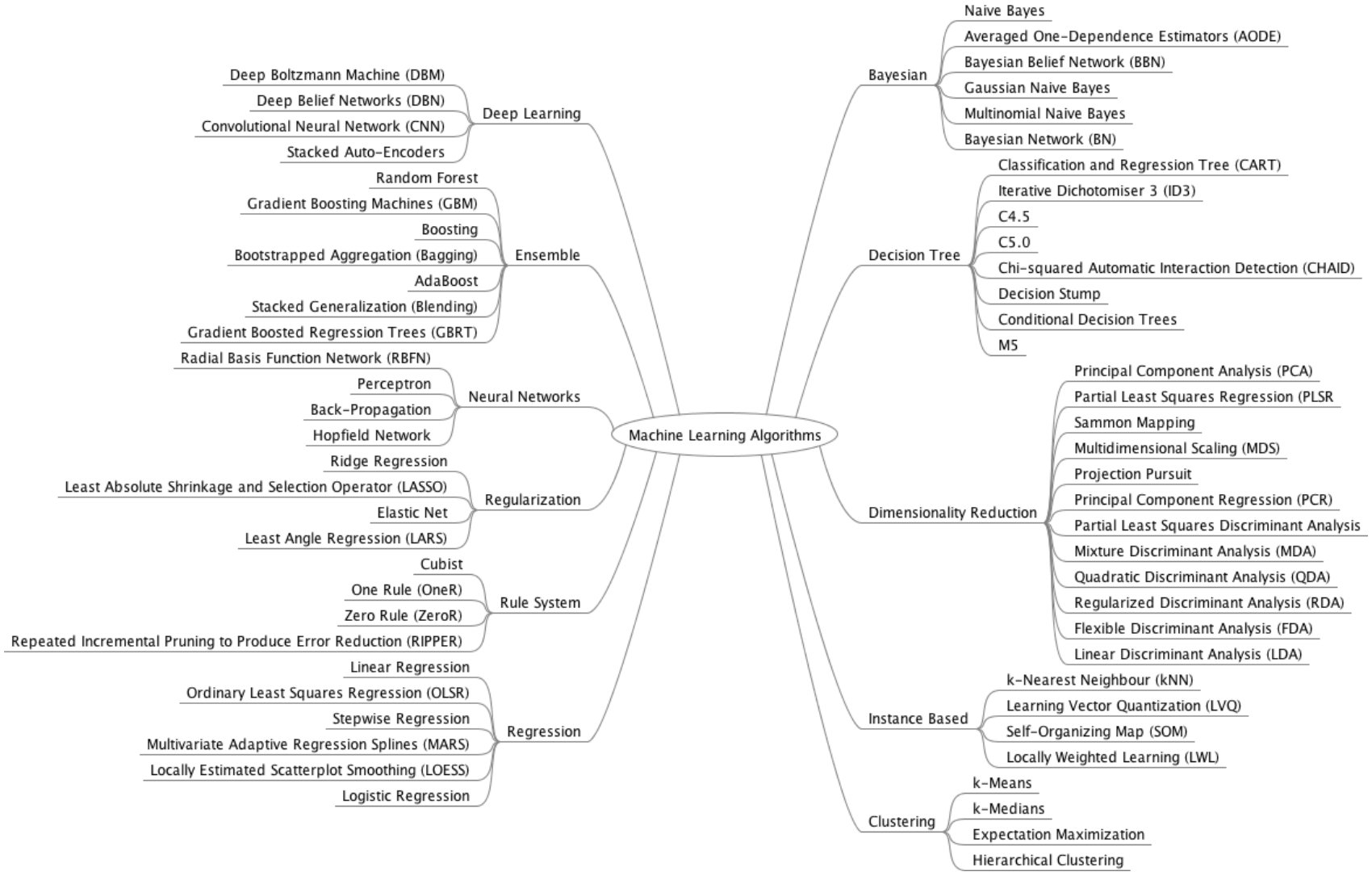
Zdroj: The State of Data Science & Machine Learning. 2017. Kaggle. https://www.kaggle.com/surveys/2017

Postupnosť krokov pri strojovom učení (ML Workflow)

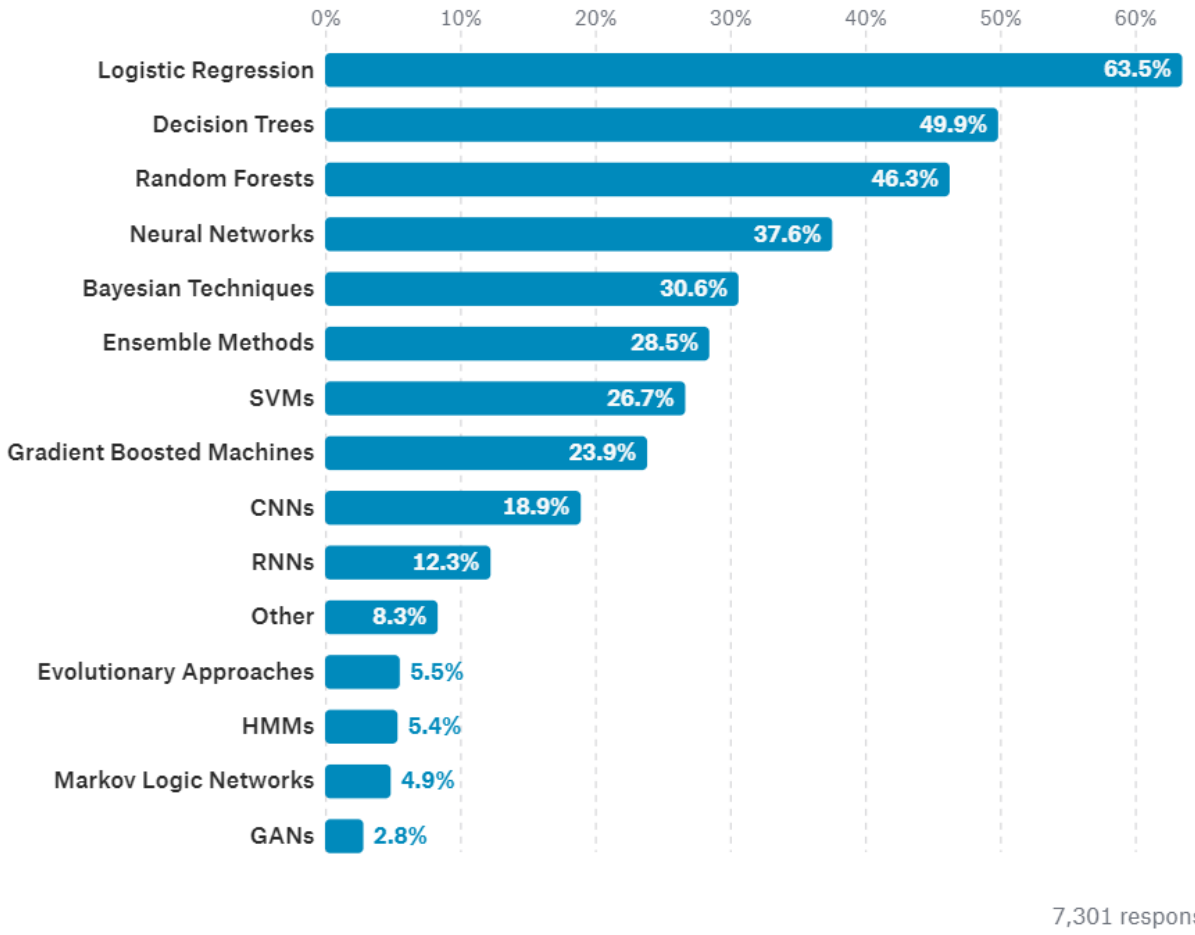
- **Definícia problému:** (ne)formálny opis, obmedzenia, manuálne riešenie
- **Integrácia dát:** konsolidácia formátu a štruktúry dát, prepájanie entít (záznamov)
- **Prie**
  vych
- **Prec**
  hodn
- **Tvorba čŕt:** extrakcia, transformácia a výber čŕt
- **Trénovanie, vyhodnocovanie a výber modelov:** zadefinovanie metodológie vyhodnotenia, optimalizácia hyperparametrov, výber optimálneho modelu
- **Prezentácia výsledkov:** vizualizácia výsledkov, obmedzenia navrhnutého riešenia, opis a zverejnenie datasetu, nasadenie a prevádzka

Pozor, nikdy to nie je vodopád!

Prístupov strojového učenia je množstvo, my sa dotkneme len niektorých



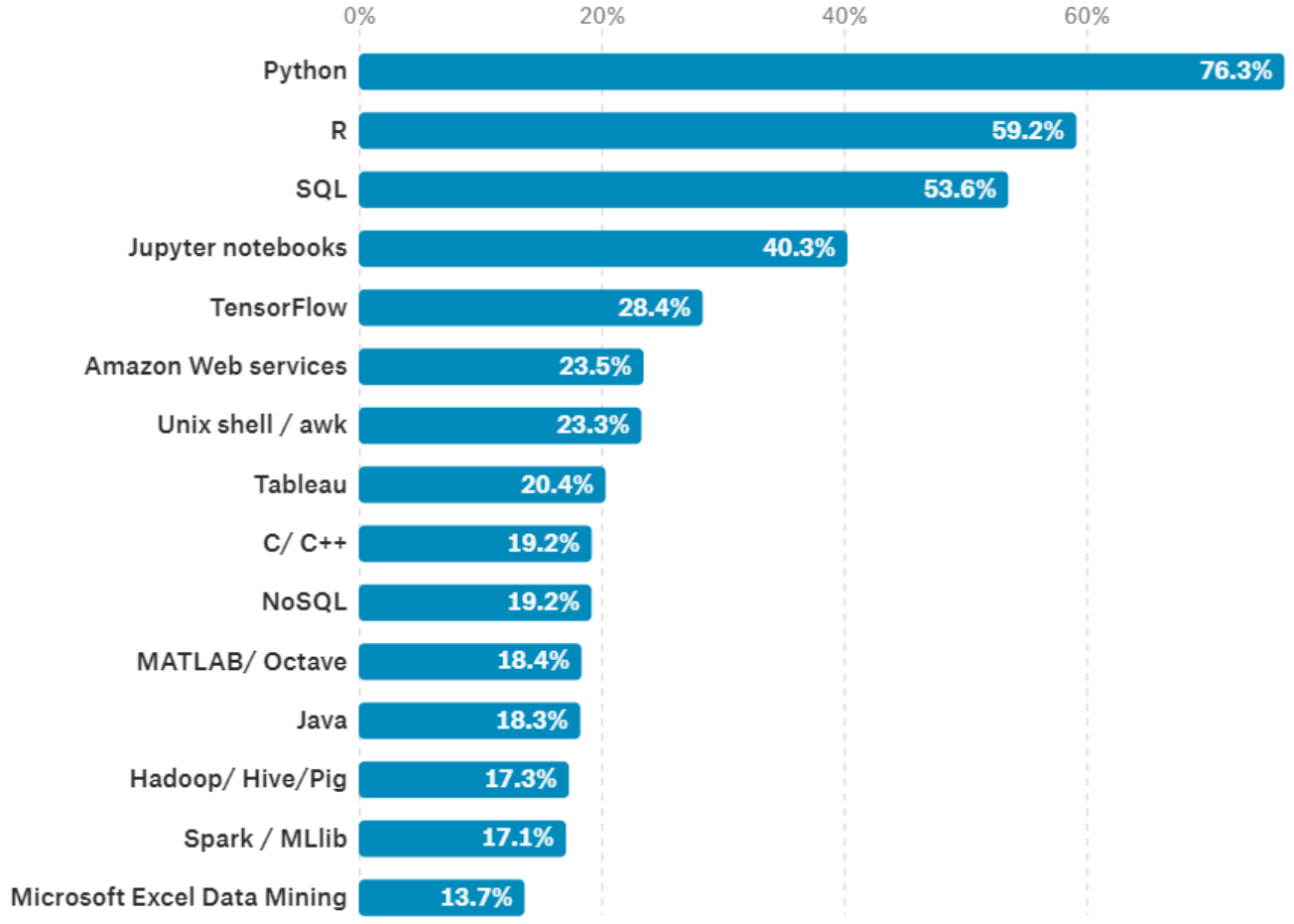
Preberieme najpoužívanejšie metódy strojového učenia – logistickú (a lineárnu) regresiu a rozhodovacie stromy



Zdroj: The State of Data Science & Machine Learning. 2017. Kaggle. https://www.kaggle.com/surveys/2017

Nástroje na analýzu dát

Aký je najpoužívanejší nástroj na analýzu dát?



Zdroj: The State of Data Science & Machine Learning. 2017. Kaggle. https://www.kaggle.com/surveys/2017

Poznáte už minimálne dva z nich

Excel

- Výhody
  - Možnosť vkladať štruktúrované dáta (csv, DB)
  - Filtrovanie dát
  - Vzorce -> transformácie dát, tvorba čŕt
  - Vizualizácie (scatter plot, stĺpcový diagram, histogram, krabicový diagram, ...)
  - Trendy (jednoduché regresné modely)
  - Pluginy na štatistické testovanie
- Nevýhody
  - Množstvo dát
  - Replikovateľnosť
  - Horšie možnosti automatizácie (aj keď sú makrá)

- Výhody
  - Efektívne ukladanie, indexovanie a dopytovanie nad dátami
  - Optimalizácia dopytov (query planner)
  - Funkcie na transformáciu dát (dátumy, reťazce, deskriptívna štatistika)
  - GROUP BY, window funkcie
- Nevýhody
  - Komplikovanejšia syntax zložitejších dopytov
  - Neposkytuje priamo nástroje na vizualizáciu dát alebo zložitejšie (iteratívne) algoritmy strojového učenia

SQL: Window funkcie

```
SELECT depname, empno, salary, avg(salary) OVER (PARTITION BY depname) FROM empsalary;
```

```
depname   | empno | salary |          avg
-----------+-------+--------+-----------------------
 develop   |    11 |   5200 | 5020.0000000000000000
 develop   |     7 |   4200 | 5020.0000000000000000
 develop   |     9 |   4500 | 5020.0000000000000000
 develop   |     8 |   6000 | 5020.0000000000000000
 develop   |    10 |   5200 | 5020.0000000000000000
 personnel |     5 |   3500 | 3700.0000000000000000
 personnel |     2 |   3900 | 3700.0000000000000000
 sales     |     3 |   4800 | 4866.6666666666666667
 sales     |     1 |   5000 | 4866.6666666666666667
 sales     |     4 |   4800 | 4866.6666666666666667
(10 rows)
```

Zdroj: https://www.postgresql.org/docs/9.3/static/tutorial-window.html

My budeme na predmete pracovať s Pythonom

Organizácia predmetu a podmienky absolvovania

https://github.com/sevo/IAU-2018-2019

Harmonogram prednášok

1. týždeň: Úvod do inteligentnej analýzy údajov
2. týždeň: Úvod do spracovania údajov v jazyku Python
3. týždeň: Prieskumná analýza a vizualizácia údajov
4. týždeň: Získavanie a prepájanie údajov
5. týždeň: Prieskumná analýza s využitím štatistickej analýzy
6. týždeň: Čistenie a predspracovanie údajov
7. týždeň: *Prednáška nie je (Sviatok všetkých svätých, nahrádza sa posledný týždeň semestra)*
8. týždeň: Predspracovanie textových dát
9. týždeň: Vyhodnocovanie a výber modelov
10. týždeň: Lineárna a logistická regresia
11. týždeň: Rozhodovacie stromy
12. týždeň: Numerická optimalizácia a simulácie
13. týždeň: Distribuované počítanie

Tento týždeň

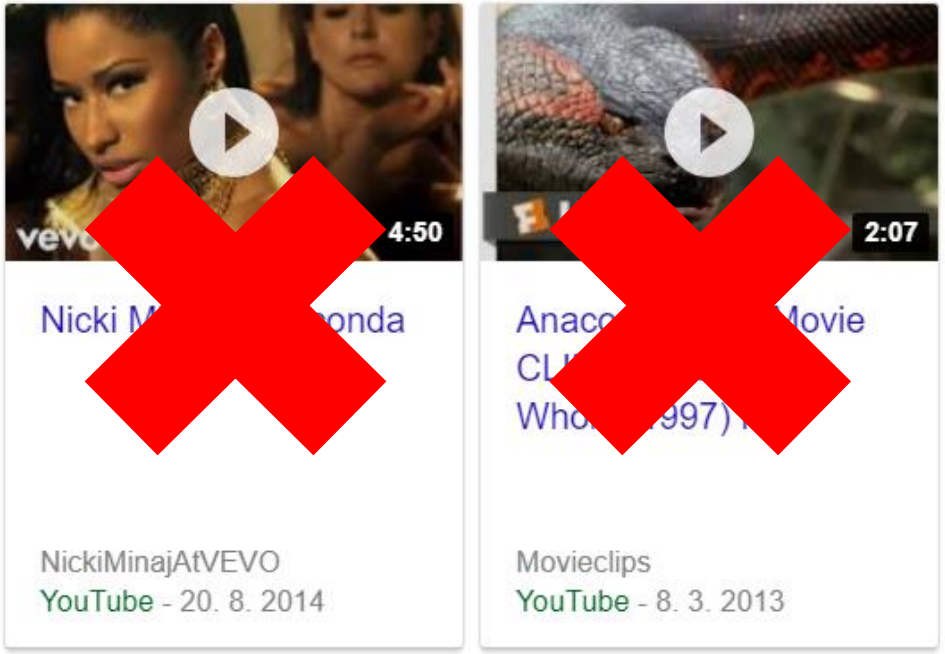
- Analýza dát pomocou tabuľkového procesora (MS Excel, LibreOffice Calc a pod.)
- Analýza dát pomocou SQL

- **Doneste si, prosím, vlastné notebooky**

Budúci týždeň

Jakub Ševcech: Úvod do spracovania údajov v jazyku Python

- Nainštalujte si Python (ideálne cez distribúciu Anaconda)



ANACONDA®

https://tinyurl.com/iau2018-19