Data kolem nás

Život v kyberprostoru

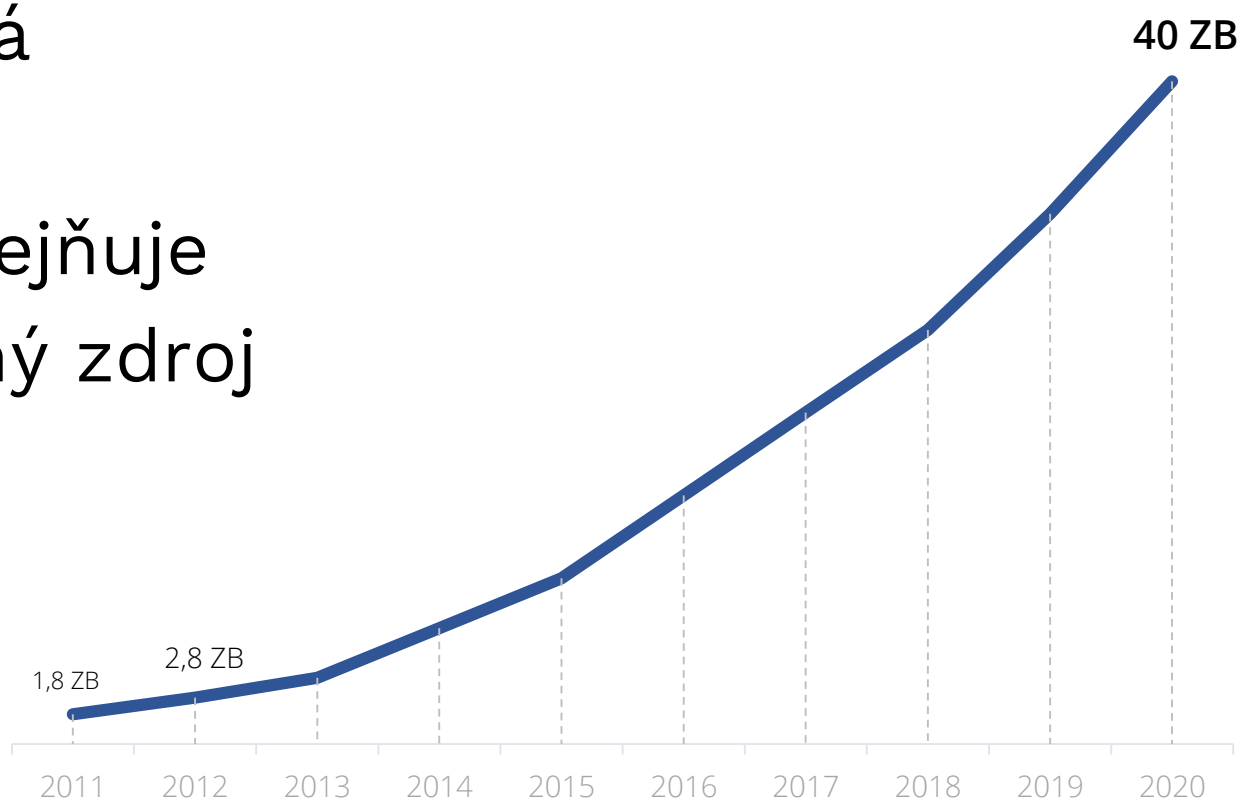
4. 12. 2019

Data

- výraz pro údaje, používané pro popis nějakého jevu
- popis vlastnosti pozorovaného objektu
- získávají se zápisem nebo měřením

Data kolem nás

- data jsou nová ropa
- všechno se měří a ukládá
- mnoho dat je k ničemu
- mnoho se jich dnes zveřejňuje
- některá představují cenný zdroj



A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

294bn

billion emails are sent

Radicati Group

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

3.9bn

people use emails

4TB

of data produced by a connected car

Intel

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

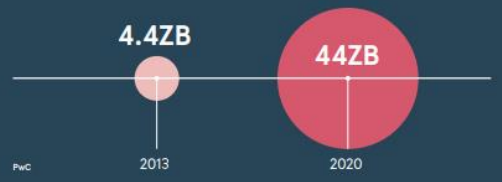
*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn

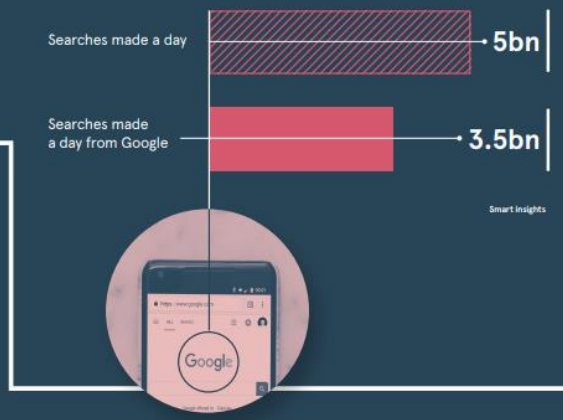
messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

ACCUMULATED DIGITAL UNIVERSE OF DATA



PwC



Smart Insights

463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business

28PB

to be generated from wearable devices by 2020

Statista

RACONTEUR

K čemu jsou data dobrá?

- data pro aktivní občanství?
- data v pracovním životě?
- data pro řešení každodenních problémů?
- **datový mindset**

Datový mindset

- Jakou rozlohu mají v ČR evropsky významné lokality?
- https://data.gov.cz/

Otevřená data jsou informace a data bezplatně a volně dostupná na internetu ve strukturované a strojově čitelné podobě a jsou zpřístupněna způsobem, který jejich využití neklade zbytečné technické či jiné překážky.

U nás je tato definice velmi důležitá...

https://opendata.gov.cz/_media/standardy_publikace_a_katalogizace_otevrenych_dat_vs_cr.pdf

Open data | +/-

Jaké výhody má otevírání dat?

Open data | +/-

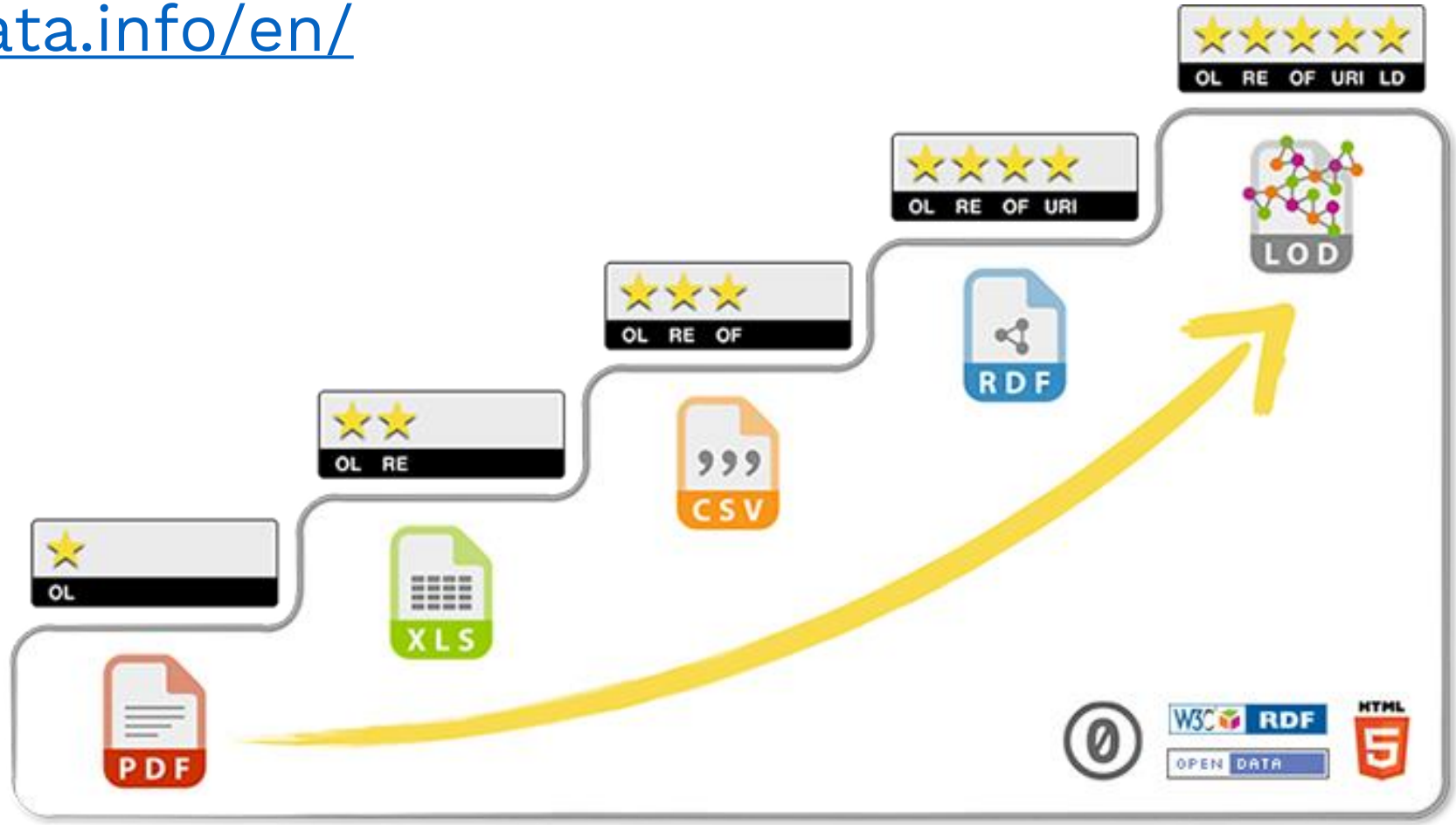
- zvýšení efektivity – sdílení a analýza
- podpora ekonomiky – zdroj inovací, surovina
- transparentnost a kontrola
- zapojení občanů do rozhodování
- datová žurnalistika

Open data | +/-

- vznik aplikací nad otevřenými daty
- „hlídači“ státu
- https://supervizor.mfcr.cz/
- https://www.znecistovatele.cz/
- najdete bezpochyby desítky dalších

Open data | Stupně otevřenosti

- https://5stardata.info/en/



Data v různých podobách

Různé podoby dat kolem nás

- expertní systémy
- faktografické databáze
- statistické databáze
- datové repozitáře

Expertní systémy

- systémy napodobující fungování odborníka
- báze znalostí / báze dat / řídící mechanismus
- pravidla IF-THAN
- dotazování – jako skutečný expert
- postupné hledání výstupu
- podpora rozhodování

Expertní systémy

Dobrý den, mám opakované bolesti hlavy,
zvracím a trápí mě kašel.
Co by mi mohlo být?

https://symptoms.webmd.com/

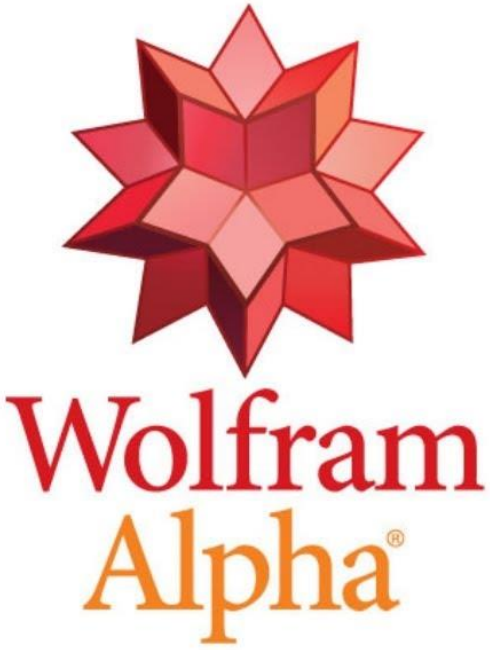
Faktografické databáze

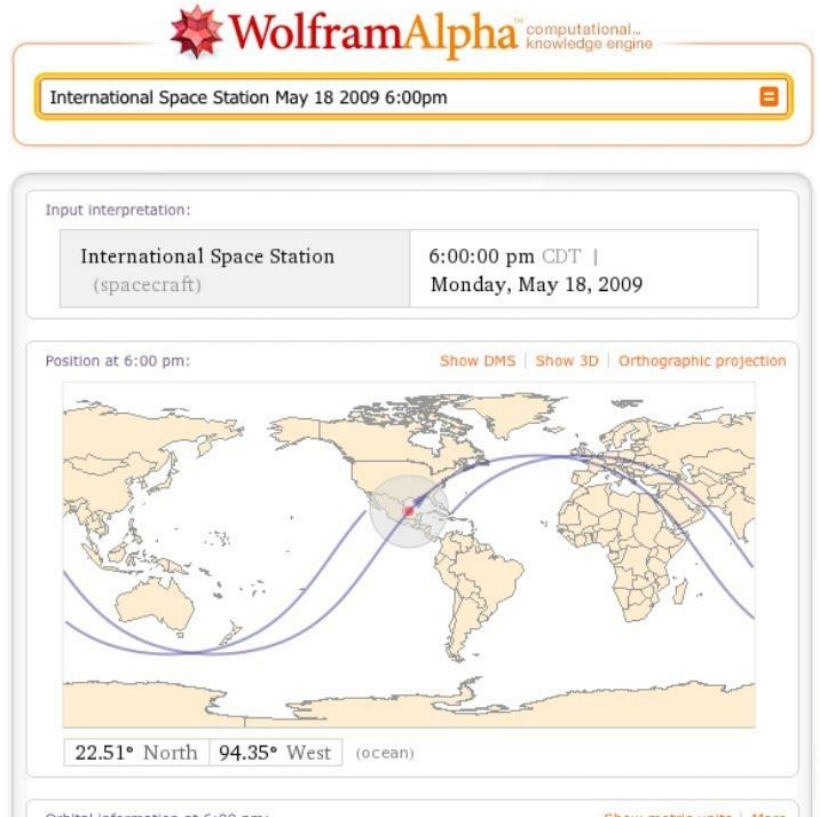
- údajovou základnu tvoří faktografické informace
- *specializované databáze (chemie, fyzika,...)*
- *statistické databáze*
- *encyklopedické databáze*

Faktografické databáze

Wolfram Alpha

https://www.wolframalpha.com/





Wolfram Alpha | Tipy

- scrabble vs. jenga
- thorin vs. frodo
- magikarp vs. pikachu
- uncle's uncle's son's daughter's cousin
- libraries with number of books > 15 million
- 85 kg 192 cm 12 beers in 5 hours
- 42 mars bars
- F#

Faktografické databáze

Dobrý den, mám doma Citalec 20, ale ztratil jsem příbalový leták. Nedá se nějak zjistit, jestli nemůže způsobovat hypokalémii? Je to vážně hodně důležitý...

http://www.sukl.cz/modules/medication/search.php

Faktografické databáze

Dobrý den, nedávno se u nás v obci stala nehoda, kdy se srazil vlak s autem. Chtěl bych o tom vědět víc, ale našel jsem jen pár zpráv v novinách a to je všechno.

https://erail.era.europa.eu/

Statistické databáze

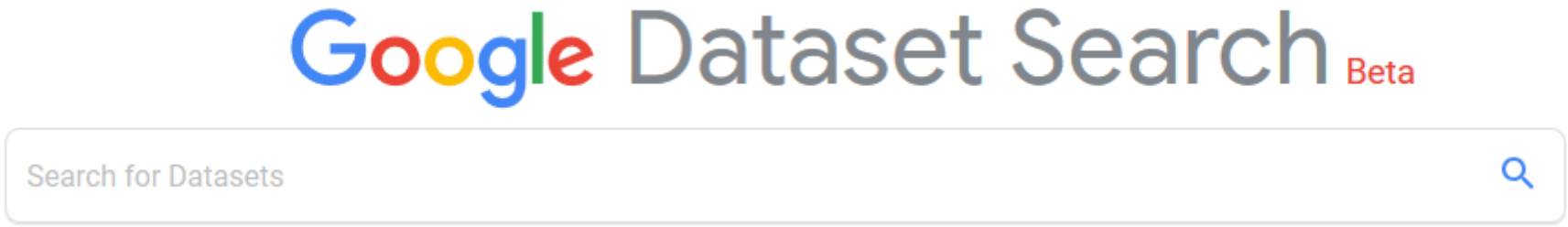
- obrovské množství databází mezinárodních organizací
- databáze od státních orgánů, úřadů
- statistické úřady

Proklikejte si:
https://data.worldbank.org/
http://apps.who.int/gho/data/node.home
https://data.unicef.org/
https://stats.oecd.org/
http://data.europa.eu/euodp/en/data/

Statistické databáze

Dobrý den, chci si otevřít hotel v Brně a potřeboval bych vědět, kolik se tady během roku ubytovává lidí a jestli to číslo klesá nebo stoupá – no prostě jestli má cenu další hotel otvírat.

https://www.czso.cz/



https://toolbox.google.com/datasetsearch

Vyhledávání dat a v datech

- složitější činnost
- kromě hledání ještě vrstva práce s daty
- vyžaduje specifické dovednosti a znalosti
- znalost formátů dat
- znalost základní práce s daty

Typy dat

- *způsob uložení dat do souboru*
- XLSX, CSV, TXT
- JSON, XML
- KML, GeoJSON – geografický rozměr

CSV

	A	B	C	D
1	ID	Gender	City	Monthly_I
2	ID000002C	Female	Delhi	20000
3	ID000004E	Male	Mumbai	35000
4	ID000007H	Male	Panchkula	22500
5	ID000008I	Male	Saharsa	35000
6	ID000009J	Male	Bengaluru	100000
7	ID000010K	Male	Bengaluru	45000
8	ID000011L	Female	Sindhudur	70000
9	ID000012M	Male	Bengaluru	20000
10	ID000013N	Male	Kochi	75000
11	ID000014C	Female	Mumbai	30000
12	ID000016C	Male	Mumbai	25000
13	ID000018S	Female	Surat	25000
14	ID000019T	Female	Pune	24000
15	ID000021V	Male	Bhubanes	27000
16	ID000022V	Female	Howrah	28000

JSON

```
{
   "Employee": [
      {
         "id":"1",
         "Name": "Ankit",
         "Sal": "1000",
      },
      {
         "id":"2",
         "Name": "Faizy",
```

```
<?xml version="1.0"?>
<contact-info>
<name>Ankit</name>
<company>Anlytics Vidhya</company>
<phone>+9187654321</phone>
</contact-info>
```

Základní dovednosti

- otevřít CSV, XLSX a další běžné formáty
- převést do sloupců
- provést základní operace
- filtrovat

Pokročilejší práce s daty

Dobrý den, včera sem jel na kole a skoro sem přejel kočku.
Zajímalo by mě, jak často se něco takového v Brně děje,
že by se cyklista srazil se zvířetem...

https://data.brno.cz/

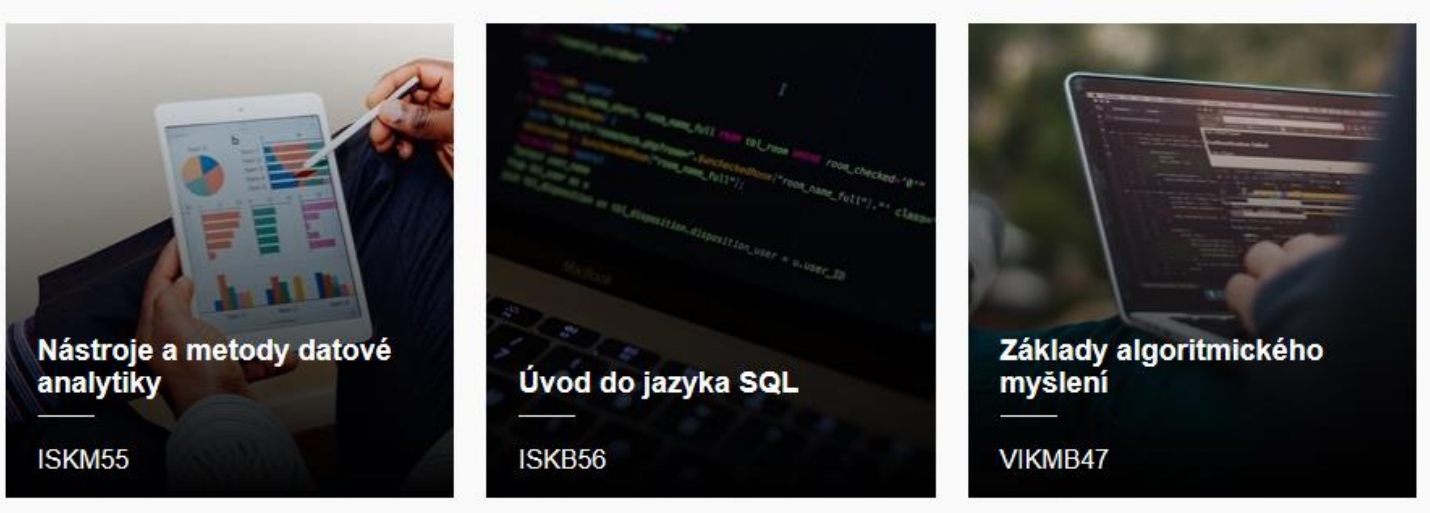
Pokročilejší práce s daty

Dobrý den, plánujeme oplotit celý náš les a potřebujeme vědět, zda v něm není biokoridor. Můžete mi sehnat data o dálkových migračních koridorech v Česku?

https://data.gov.cz/

Pokročilejší práce s daty

- dotazování DB
- parsing (příklad)
- čištění dat
- georeferencing
- …



Hodnocení kvality dat

- data nemusí být tzv. „čistá“
- data vznikají v kontextu: *kola vs. zvěř v Brně?*
- kdo je měří a publikuje? – jasný zdroj!
- jaká jsou k nim metadata?
- dokážu zjistit, co který sloupeček znamená?
- je popsaná metodologie jejich vzniku?

Evidence-based praxe

- přístupy založené na datech
- využití dat pro rozhodování
- *příklady?*

Sdílení dat ve vědě

- reprodukovatelnost výzkumu
- ověření správnosti
- možnost navázat na předchozí výzkum
- jiné využití stejných dat
- sdílení prostředí a kontextu výzkumu
- datové repozitáře

Takže...

- dat je stále více a více
- částečně i díky trendu otevřených dat
- data hrají roli ve vědě i veřejné správě
- data přicházejí v různých podobách a formátech
- představují důležitý zdroj informací
- **datový mindset**