

# Přednáška 7

## Práce s daty

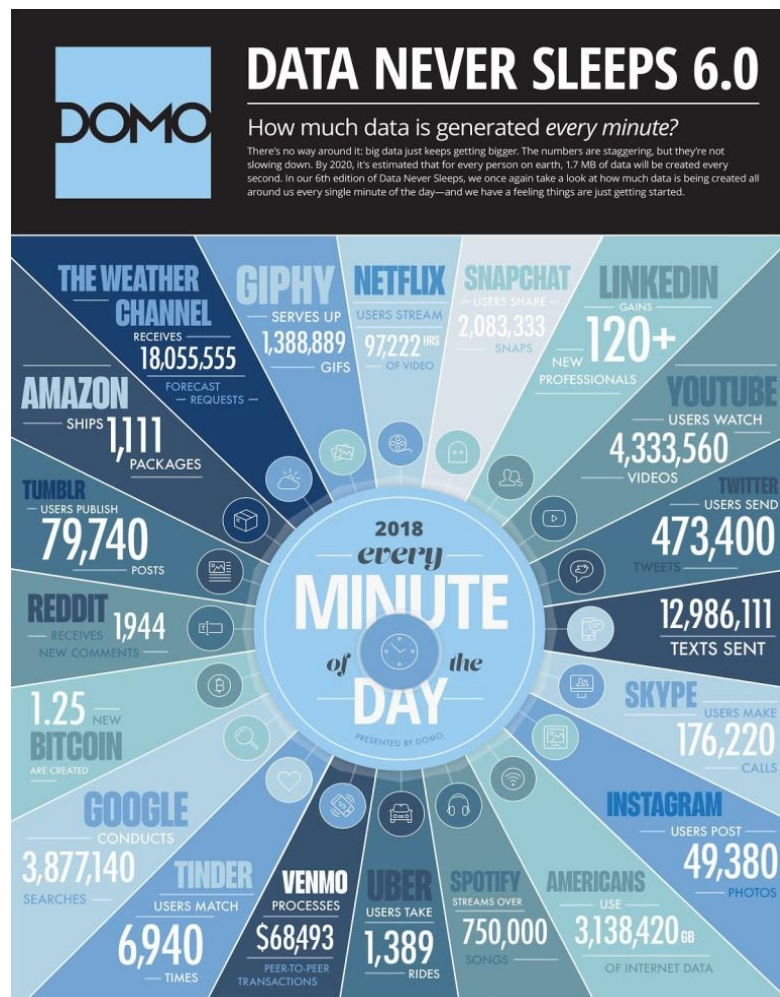
CORE013 Vývoj softwarových systémů: od myšlenky k funkčnímu řešení

# 7. Práce s daty

- Využití dat
- Velká data
- Databáze, cloud computing
- Správa dat, data science
- Digitální stopa uživatele, etické kontroverze

# 21. STOLETÍ – DOBA DATOVÁ

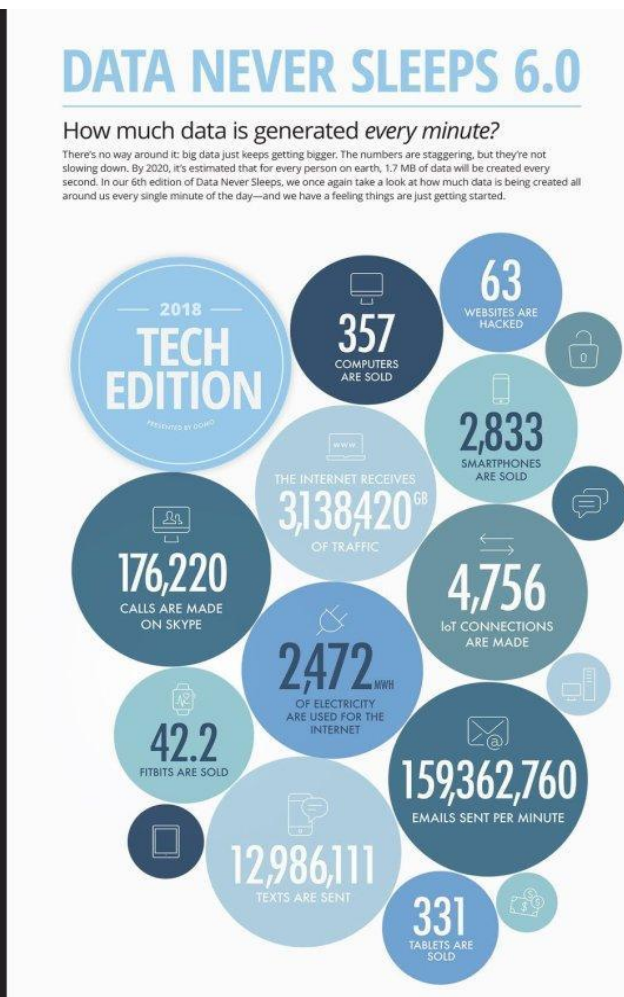
# Motivace



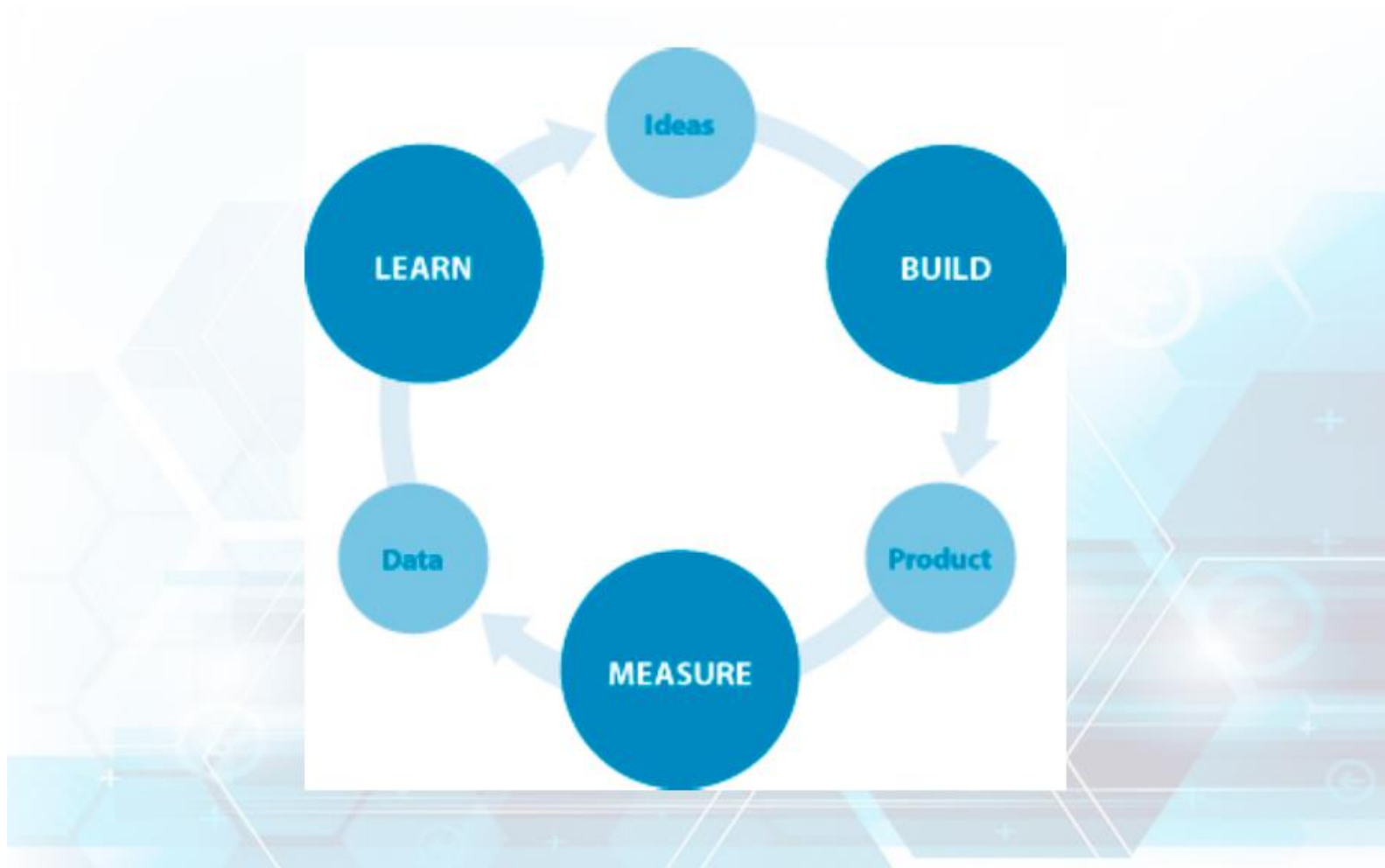
*For*  
**TECH**

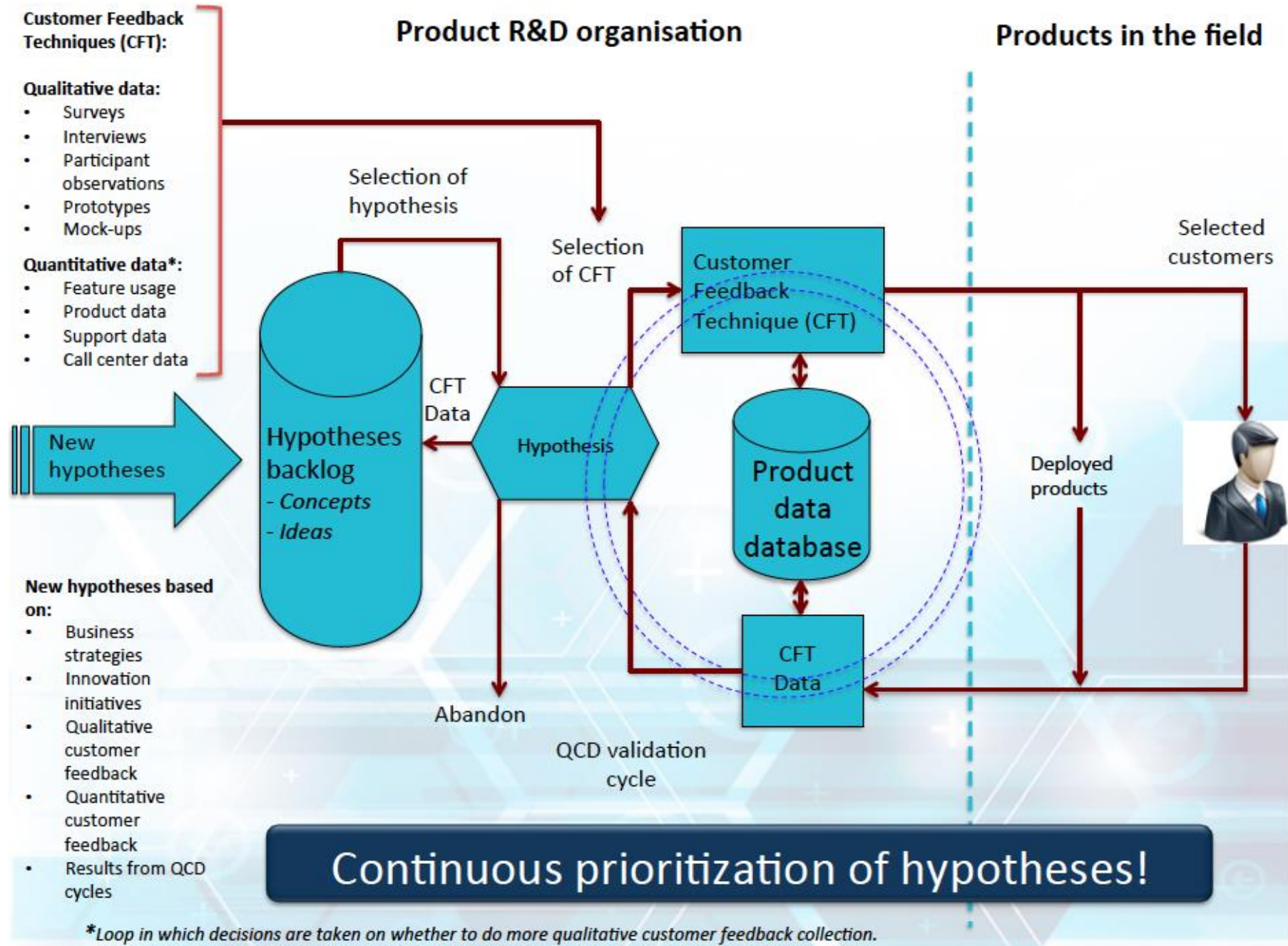
The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.

Learn more at [domo.com](http://domo.com)



# Koloběh inovací

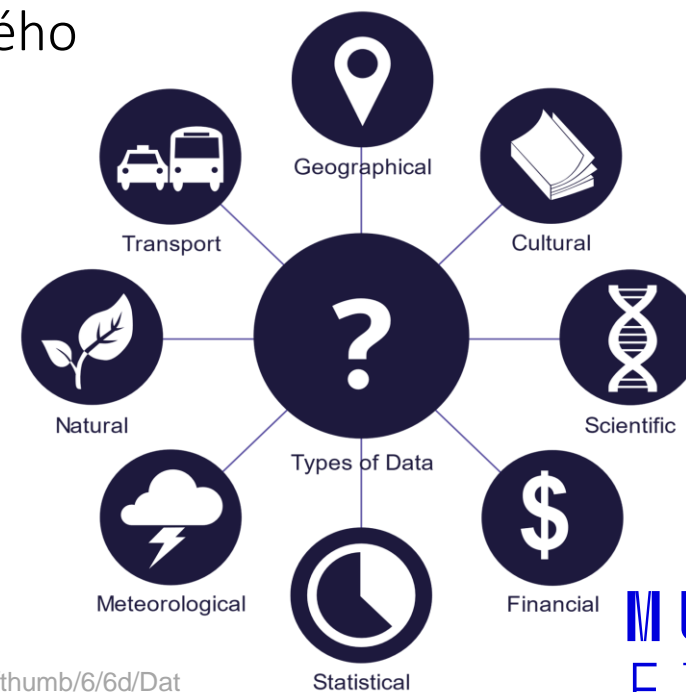




# (VELKÁ) DATA

# Data

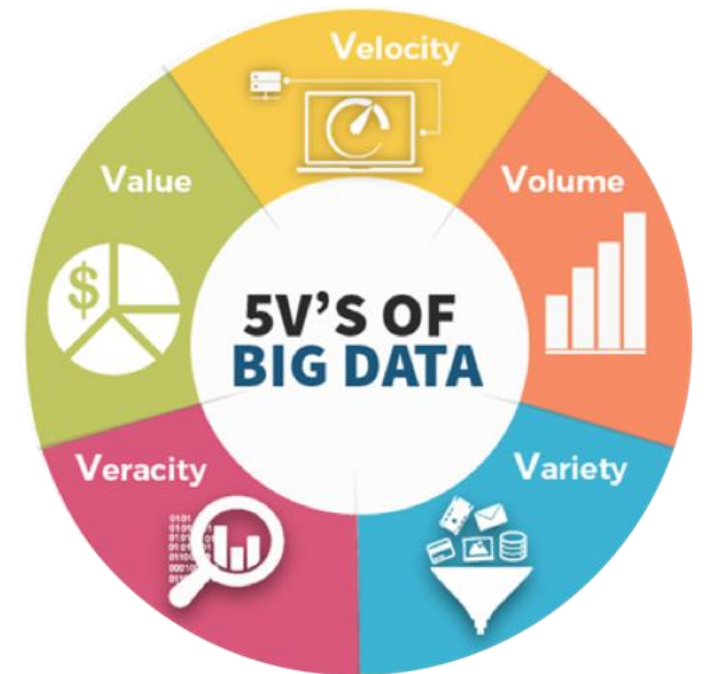
- Informace převedené do binární digitální podoby
  - informace převedené do formy, která je efektivní pro přenos a zpracování.
- Lze je vytvářet, zpracovávat, ukládat a uchovávat v digitální podobě.
  - To umožňuje přenášet data z jednoho počítače do druhého
- Digitální informace (tj. data) ve srovnání s analogovými informacemi se v průběhu času nezhoršuje a po použití neztrácí kvalitu





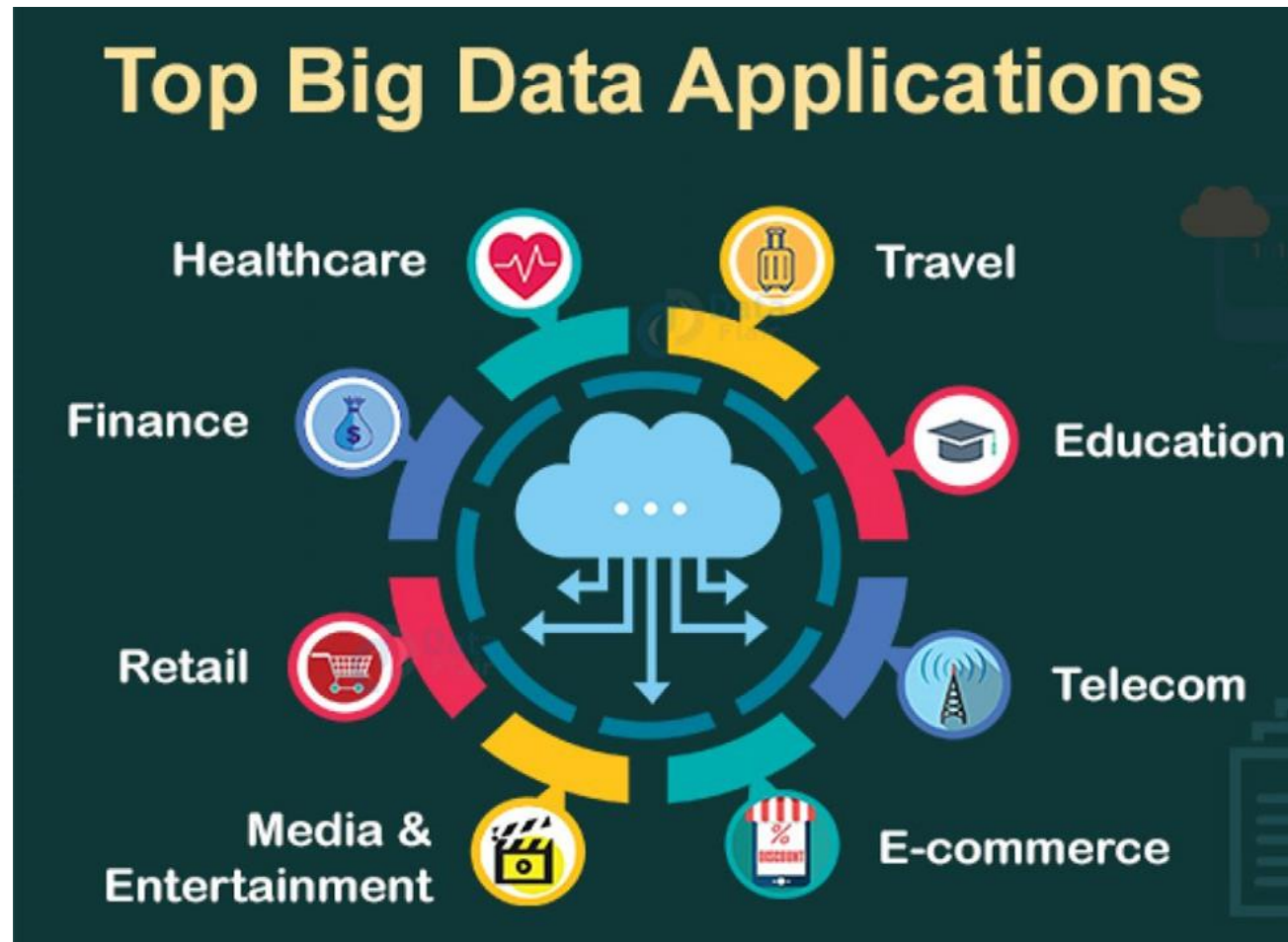
# Velká data

- **Soubor dat**, který je obrovský nebo exponenciálně roste s časem.
- Zpracování pomocí tradičních databázových a softwarových nástrojů je obtížné nebo nemožné.
- Charakteristika - 5 V's:
  - **Volume** – objem dat je obrovský
  - **Variety** – rozmanitost, různé zdroje a formát dat
  - **Velocity** – data jsou generována velmi rychle
  - **Veracity** – data jsou ověřena a validována
  - **Value** – data mají vysokou hodnotu



[https://img.techentice.com/media/2019/04/5\\_V\\_Big\\_data-1.png](https://img.techentice.com/media/2019/04/5_V_Big_data-1.png)

# Velká data

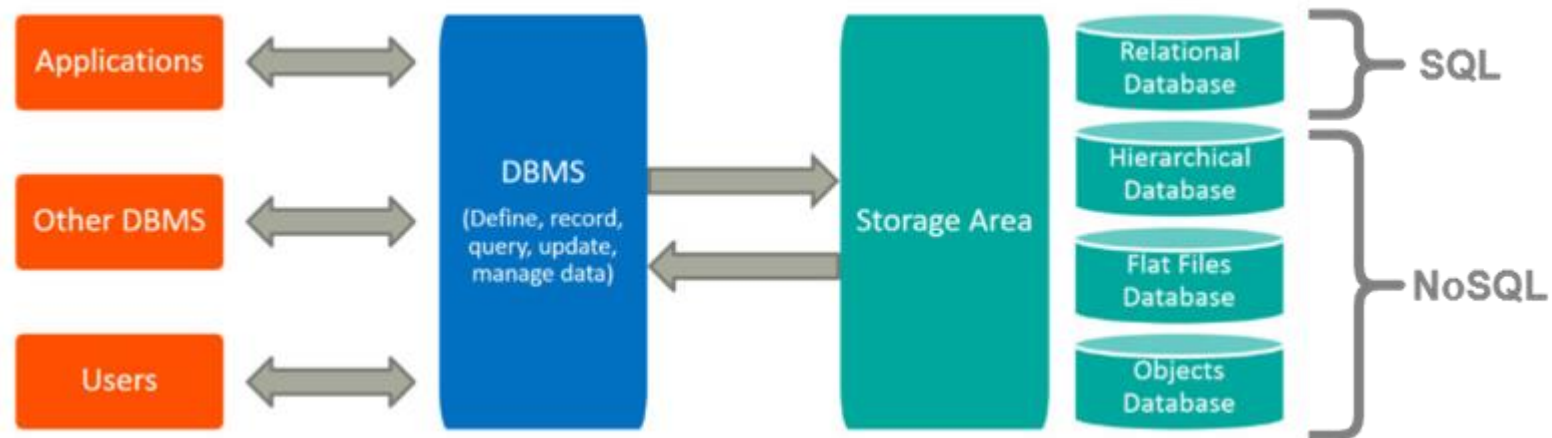


<https://data-flair.training/blogs/wp-content/uploads/sites/2/2019/12/top-big-data-applications-2-1280x720.jpg>

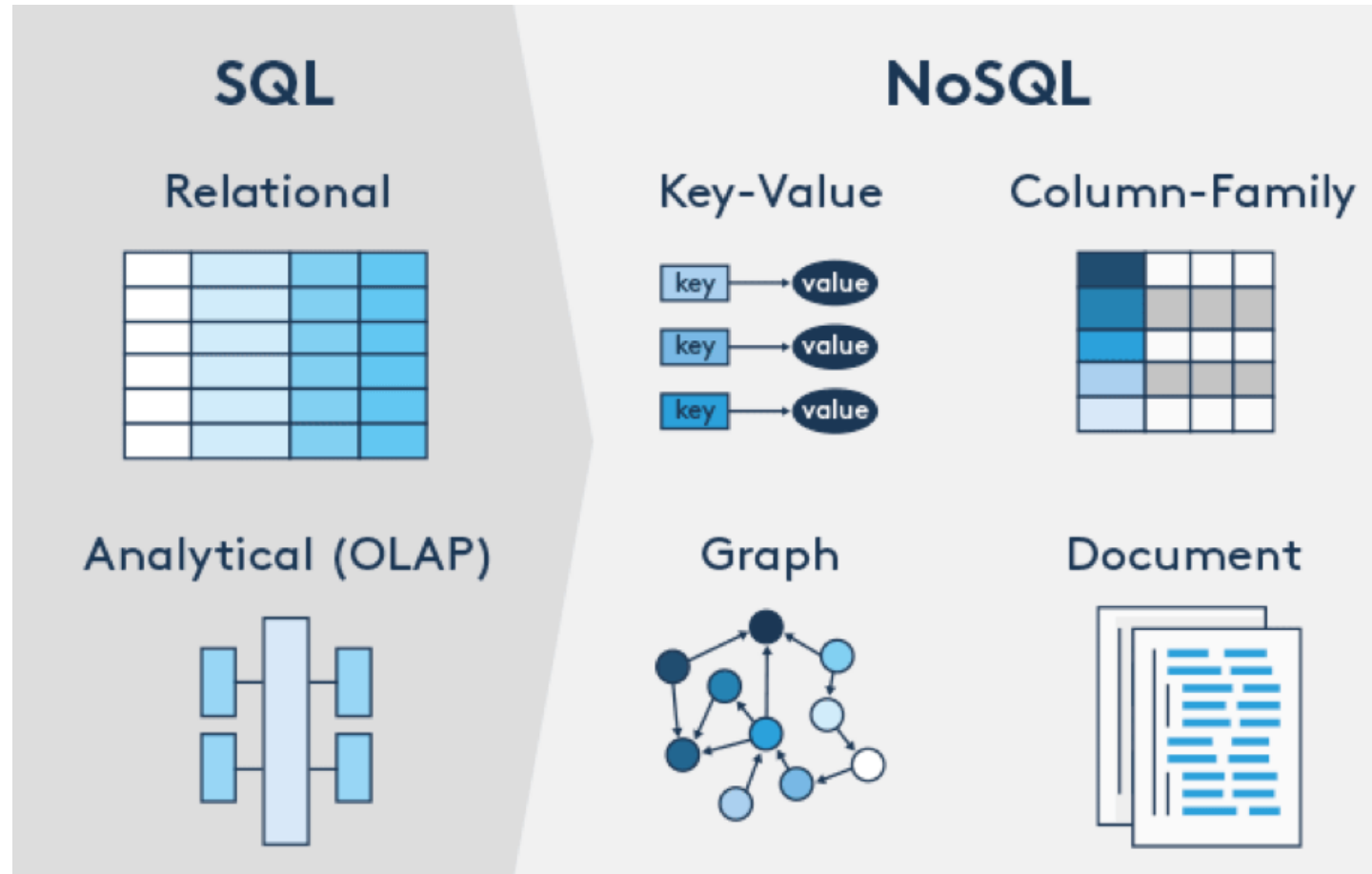
# DATABÁZE

# Databáze

- Uspořádané uložení
- Podporuje přístup k datům, jejich ukládání a manipulaci s nimi.
- Obvykle jako řádky a sloupce v tabulce.
- Nejpoužívanějším jazykem je SQL (Structured Query Language).
- Řízeno systémem pro správu databází (DBMS)



# Databáze



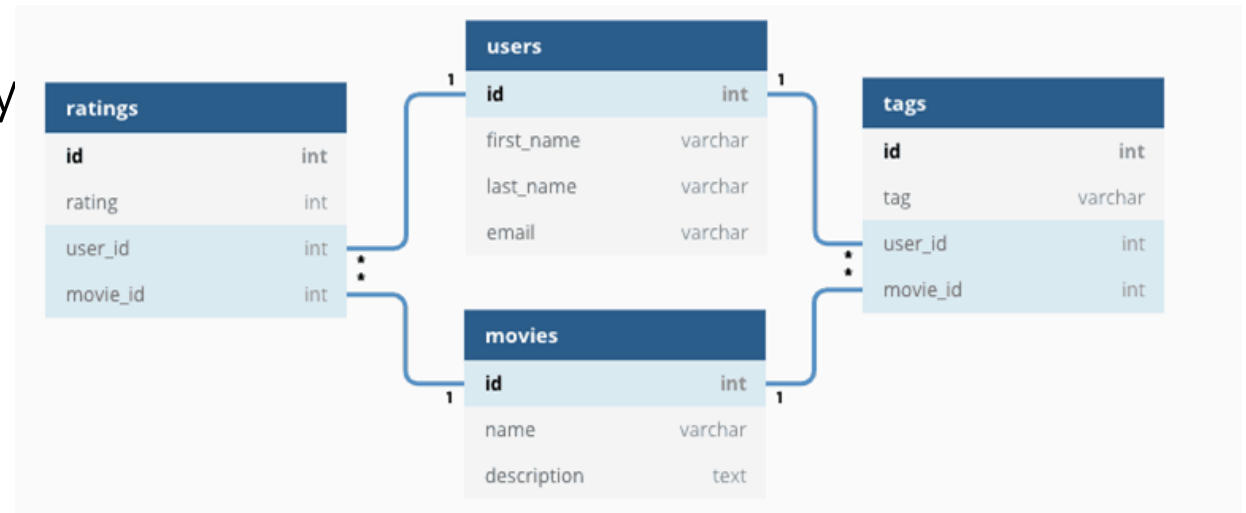
<https://asesoftware.com/site/wp-content/uploads/2019/06/asesoftware-sql-nosql.png>

# Relační databáze

- Ukládá data jako řadu dvourozměrných tabulek s řádky a sloupci s předem definovanými vztahy mezi daty.
- Vztah mezi tabulkami a typy polí se nazývá schéma. Schéma musí být jasně definováno před přidáním jakýchkoli informací.
- Každá tabulka má své vlastní sloupce a každý řádek v tabulce má stejnou sadu sloupců a jedinečné ID nazývané klíč.
- Pro dotazování se používá strukturovaný dotazovací jazyk (SQL).

# Relační databáze

- Používá se v:
  - Transakčně orientované systémy
  - Účetní software
  - Nástroje pro řízení
- Příklady zahrnují:
  - PostgreSQL
  - MySQL



[https://assets-global.website-files.com/5debb9b4f88fbc3f702d579e/5e3c1a71724a38245aa43b02\\_99bf70d46cc247be878de9d3a88f0c44.png](https://assets-global.website-files.com/5debb9b4f88fbc3f702d579e/5e3c1a71724a38245aa43b02_99bf70d46cc247be878de9d3a88f0c44.png)

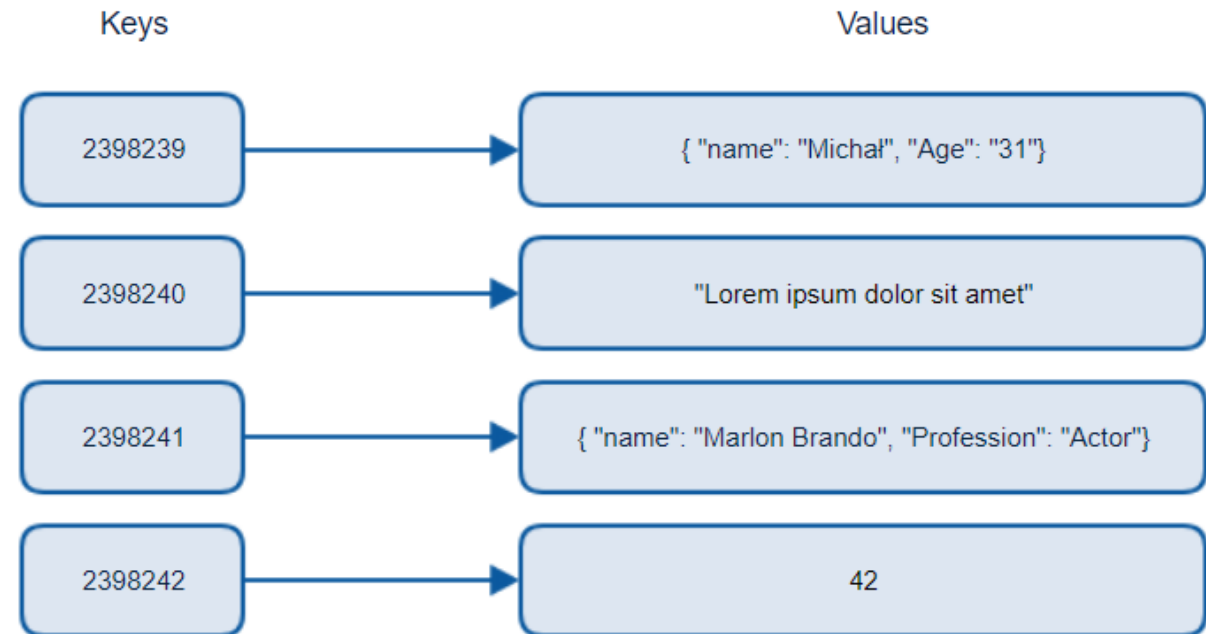
# Databáze klíč-hodnota

- NoSQL databáze.
- Ukládá data jako kolekci dvojic klíč-hodnota, v níž klíč slouží jako jedinečný identifikátor.
- Datová struktura, která je dnes známá spíše jako slovník.
- Nemá dotazovací jazyk; poskytuje jednoduchý způsob ukládání, dotazování a aktualizace dat pomocí příkazů get, put a delete - není optimalizován pro dotazování podle hodnoty.



# Databáze klíč-hodnota

- Používá se v:
  - Nákupní košík v e-shopech
  - Ukládání do mezipaměti
  - Hry pro více hráčů
- Příklady zahrnují:
  - Redis
  - Apache Cassandra
  - Amazon Dynamo DB
  - Microsoft Azure Cosmos DB



<https://www.michalbialecki.com/wp-content/uploads/2018/03/cosmos-db-key-value-schema.png>

# Dokumentové databáze

- NoSQL databáze, která uchovává soubor pojmenovaných polí a dat (tzv. dokumentů).
- Uložená data mohou být zakódována v různých formátech, např. XML, YAML, JSON, prostý text.
- Nevyžaduje, aby všechny dokumenty měly stejnou strukturu – poskytuje flexibilitu pro ukládání různých dat
- Umožňuje dotazování a filtrování dokumentů podle hodnoty jednoho nebo více polí a úpravu hodnot bez přepisování celého dokumentu.

# Dokumentové databáze

- Používá se v:
  - Profily uživatelů
  - Velká data v reálném čase
  - Správa obsahu
- Příklady zahrnují
  - MongoDB
  - Úložiště Google Cloud Firestore
  - Microsoft Azure Cosmos DB

## Document 1

```
{  
  "id": "1",  
  "name": "John Smith",  
  "isActive": true,  
  "dob": "1964-30-08"  
}
```

<https://lennilobel.files.wordpress.com/2015/07/i4.png>

## Document 2

```
{  
  "id": "2",  
  "fullName": "Sarah Jones",  
  "isActive": false,  
  "dob": "2002-02-18"  
}
```

## Document 3

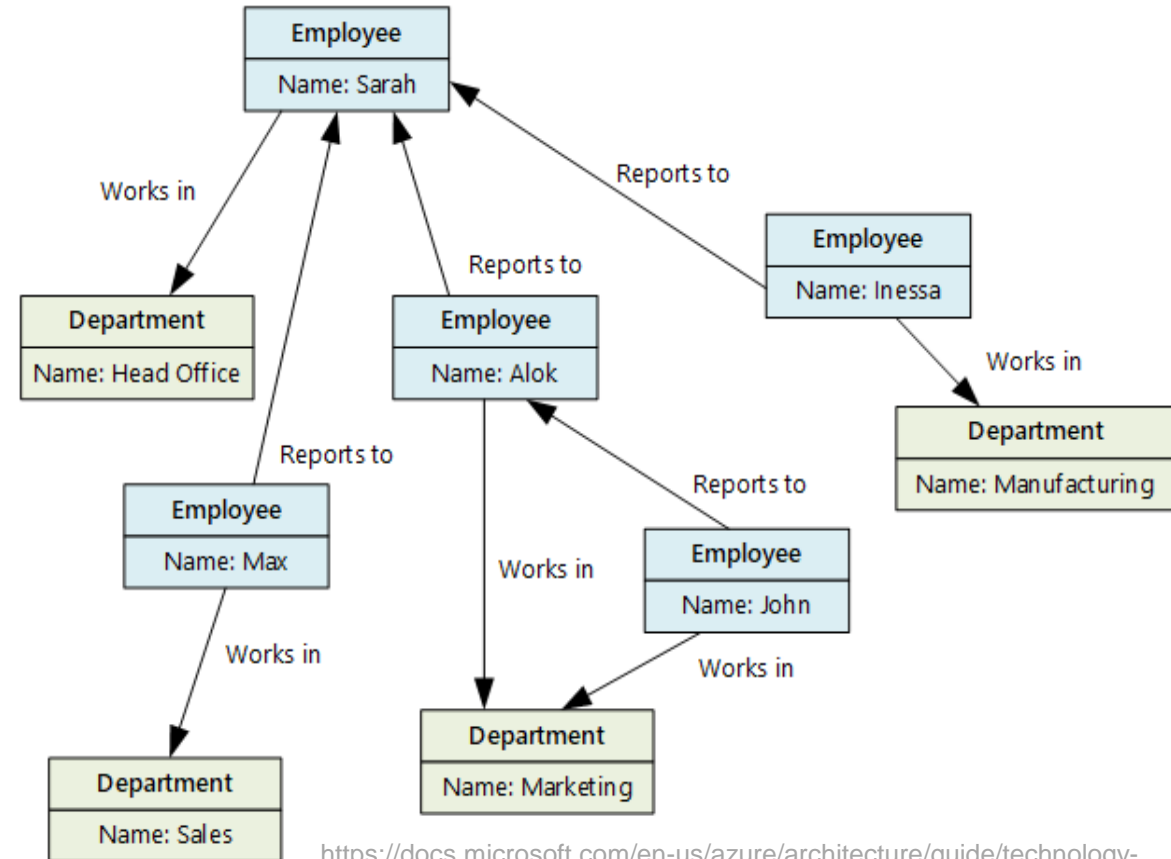
```
{  
  "id": "3",  
  "fullName":  
  {  
    "first": "Adam",  
    "last": "Stark"  
  },  
  "isActive": true,  
  "dob": "2015-04-19"  
}
```

# Grafové databáze

- NoSQL databáze, která uchovává dva typy informací, uzly a hrany.
- Uzly obvykle uchovávají informace o osobách, místech a věcech, zatímco hrany uchovávají informace o vztazích mezi uzly.
- Vztahy umožňují přímé propojení dat v databázi a jejich načtení pomocí jedné operace.
- Poskytuje dotazovací jazyk, který lze použít k efektivnímu procházení sítě vztahů.

# Grafové databáze

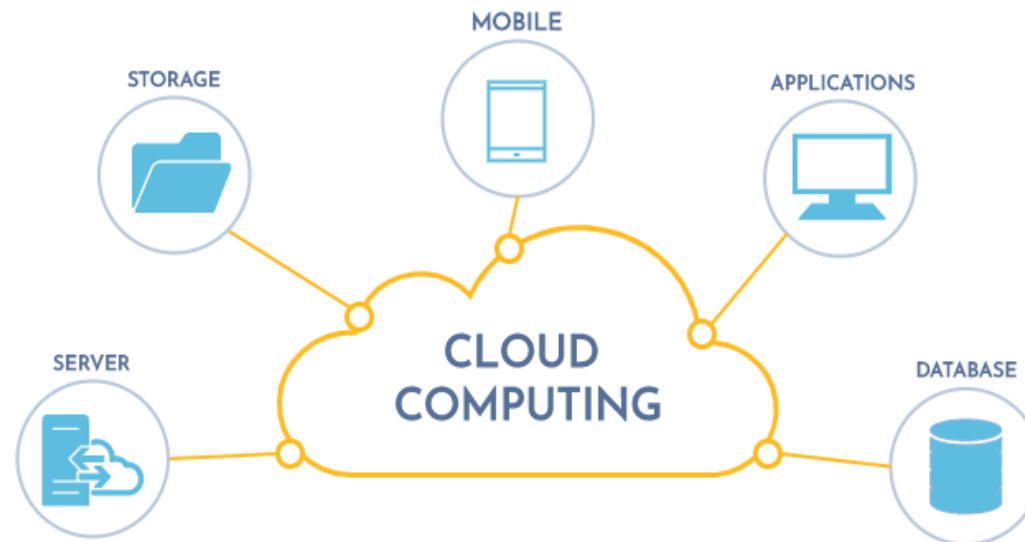
- Používá se v:
  - Sociální sítě
  - Odhalování podvodů
  - Recommendation engines
- Příklady zahrnují:
  - Neo4j
  - Amazon Neptun
  - Apache Giraph



<https://docs.microsoft.com/en-us/azure/architecture/guide/technology-choices/images/graph.png>

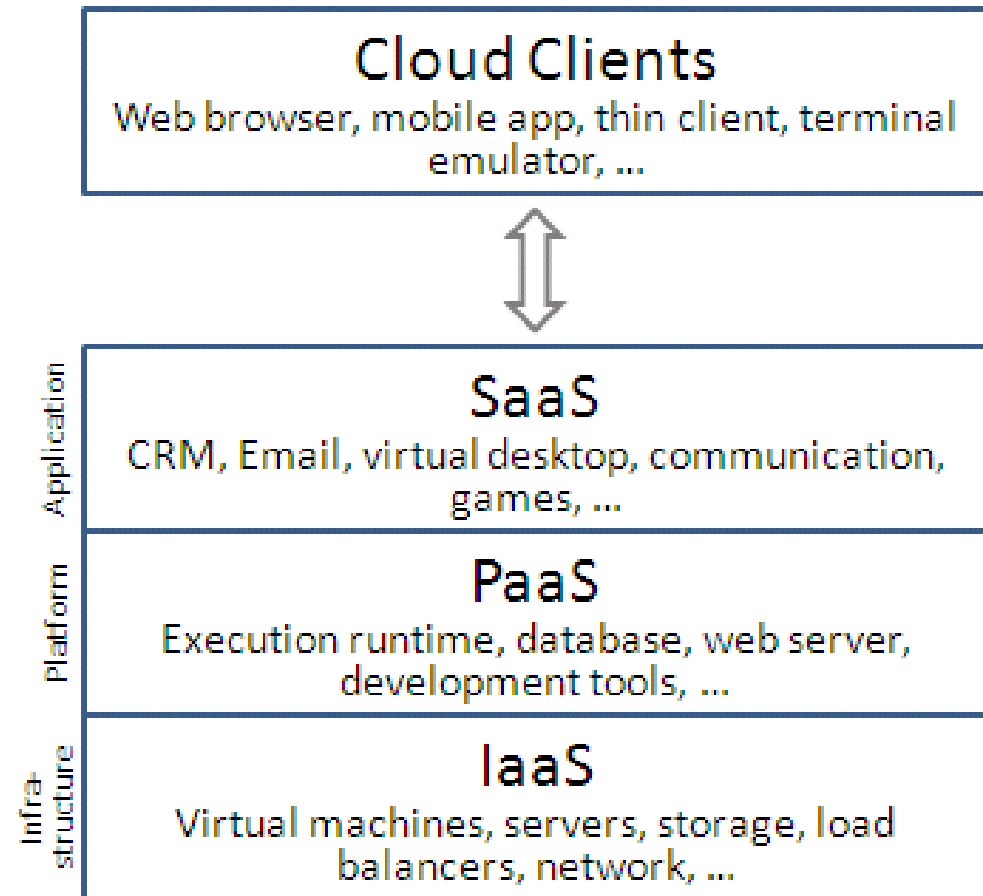
# Cloud Computing

- Cloud computing
  - Dostupnost zdrojů počítačového systému na vyžádání, zejména úložiště dat a výpočetní výkon, bez přímé aktivní správy uživatelem.
  - Včetně serverů, úložišť, databází, sítí, softwaru, analytiky a zpracování – přes internet („cloud“)
- Výhody
  - Elasticita
  - Výkonnost
  - Cena



# Cloud Computing

- Modely služeb
  - Infrastructure as a service (IaaS)
  - Platform as a service (PaaS)
  - Software as a service (SaaS)



# SPRÁVA DAT

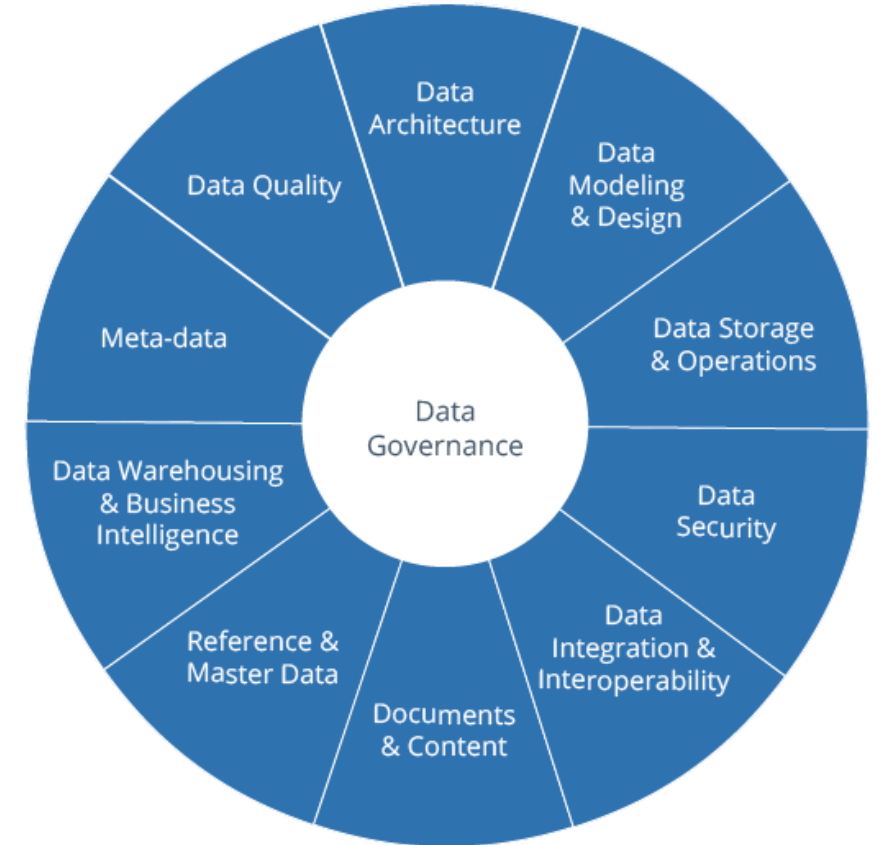


# Správa dat

- Administrativní proces, který zahrnuje **získávání, ověřování, ukládání, ochranu a zpracování** požadovaných údajů, aby byla zajištěna jejich dostupnost, spolehlivost a aktuálnost pro uživatele.
- Zahrnuje **celý životní cyklus** dat, od počátečního vytvoření dat až po jejich konečné vyřazení.
  - Zahrnuje lidi, procesy a technologie potřebné ke správě a ochraně dat společnosti s cílem zajistit obecně srozumitelná, správná, úplná, důvěryhodná, a bezpečná data.
- Některé společnosti umí dobře shromažďovat data, ale neumí je dostatečně dobře spravovat, aby z nich dokázaly vytvořit hodnotu.

# Hlavní cíle správy dat

- Minimalizace rizik
- Stanovení interních pravidel pro používání dat
- Provádění požadavků na shodu
- Zlepšení interní a externí komunikace
- Zvýšení hodnoty dat
- Usnadnění správy výše uvedené
- Snížení nákladů
- Pomáhat zajistit další existenci společnosti prostřednictvím řízení rizik a optimalizace.



<https://bi-survey.com/wp-content/uploads/2017/11/Data-Governance-topics.png>

# Životní cyklus dat

- Přehled fází úspěšné správy a uchovávání dat pro použití a opakované použití.
  - Sběr / vytváření dat
  - Ukládání dat
  - Používání dat
  - Zveřejnění dat
  - Archivace dat
  - Likvidace dat



# ZPRACOVÁNÍ DAT (DATA SCIENCE)

# Čištění dat

## DATA CLEANING CHECKLIST

### Up-to-date data



Data should be up-to-date in order to obtain maximum value from the data analysis.



### Missing values



Count missing values and analyze where in the data they are missing. Missing values can disrupt some analyses and skew the results.



### Duplicates



Duplicate IDs indicate multiple records for one person, e.g. someone holds multiple functions at the same time.



### Numerical outliers



Numerical outliers are fairly easy to detect and remove. Define minimum and maximum to spot outliers easily.



### Check IDs



Check data labels of all the fields to see whether some categorical values are mislabeled.



### Define valid output

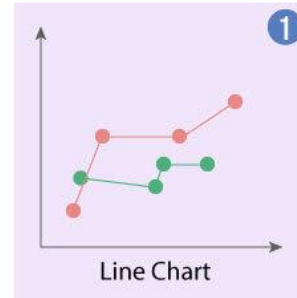


Define valid data labels for categorical data. Define data ranges for numerical variables. Non-matching data is presumably wrong.

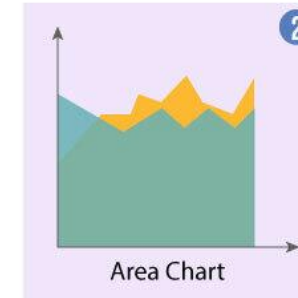


# TYPES OF DATA VISUALIZATION CHARTS

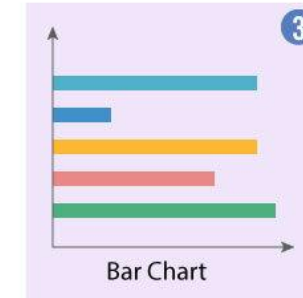
## Vizualizace dat



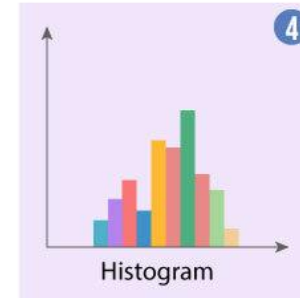
Display trends over time



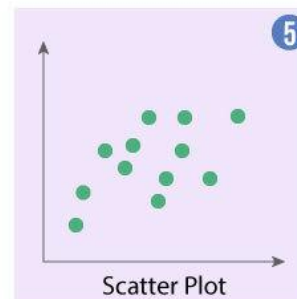
A line chart with areas below the lines filled with colors



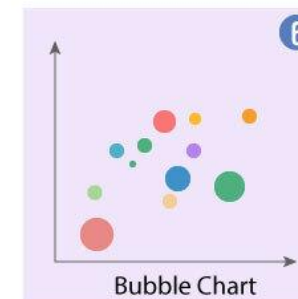
Display trends with multiple variables



Display the shape and spread of continuous dataset samples



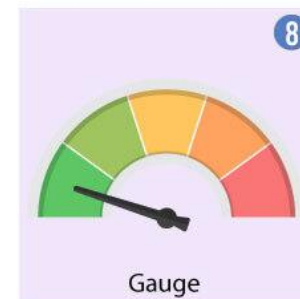
Show correlation in a dataset



Show and compare the relationship between the labelled circles



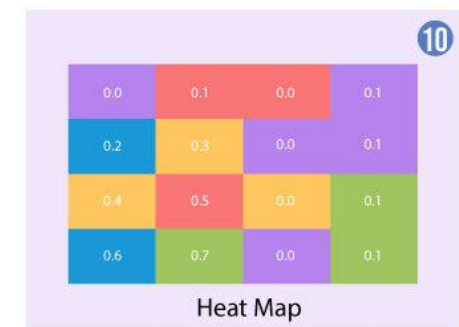
Show the contribution of data point inside a whole dataset



Visualize the distance between intervals



Show data with location as a variable



Show magnitude of a phenomenon

# Vizualization Dashboards

<https://public.tableau.com/app/profile/federal.trade.commission/viz/FraudandIDTheftMaps/AllReportsbyState>

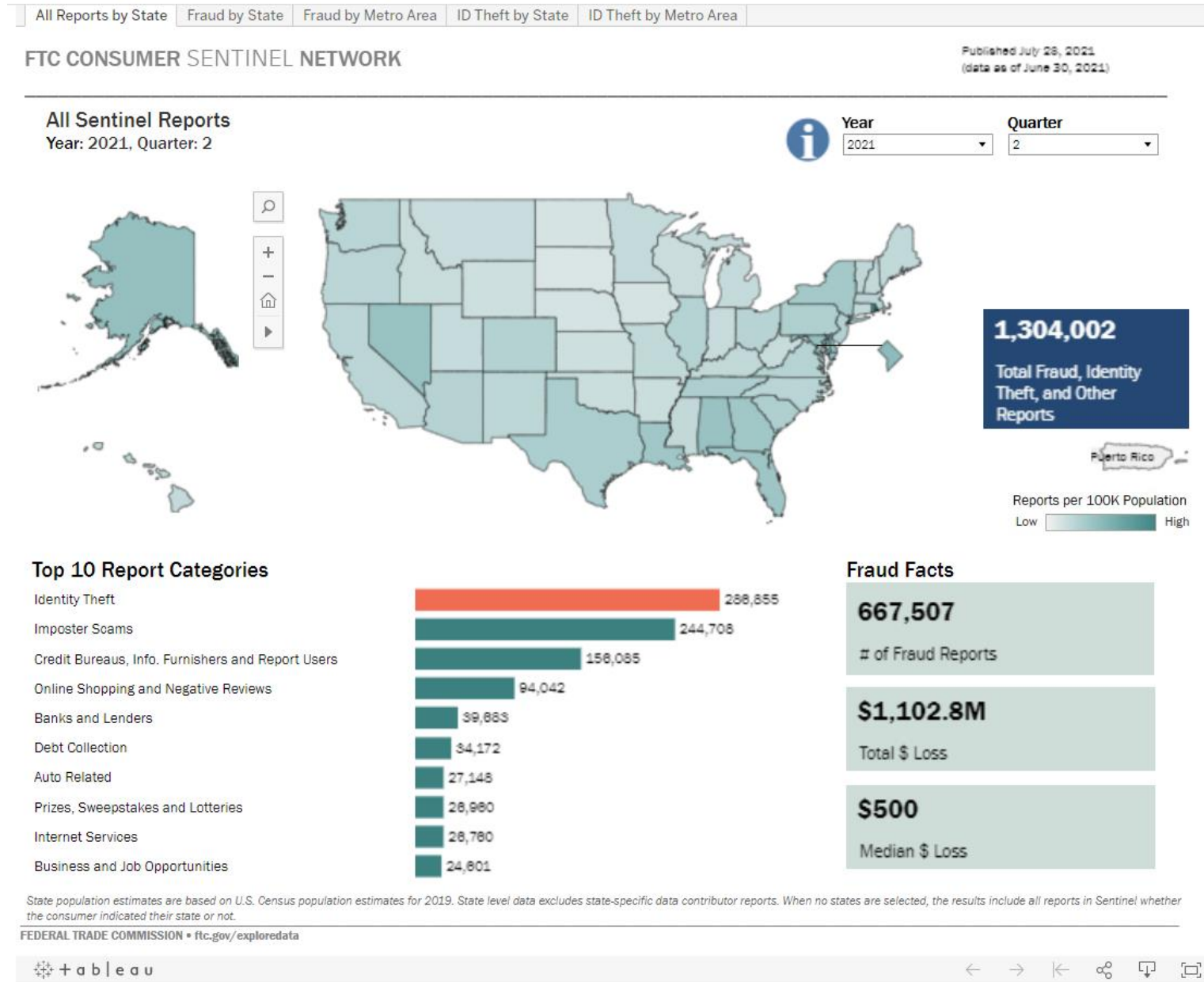
## Tableau Public

- Tableau Public access

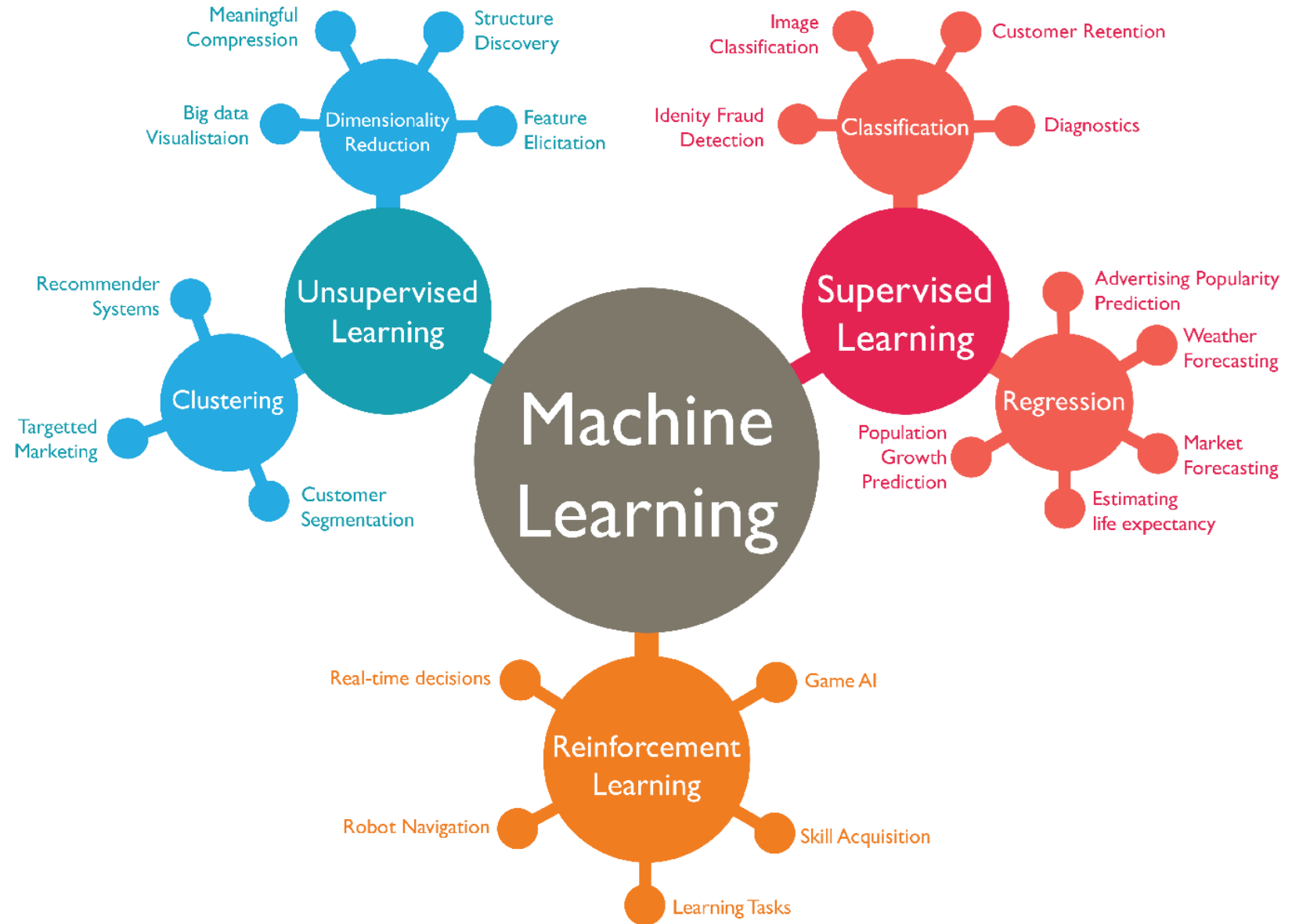
<https://public.tableau.com/en-us/s/download>

- Tutorials

<https://public.tableau.com/en-us/s/resources>



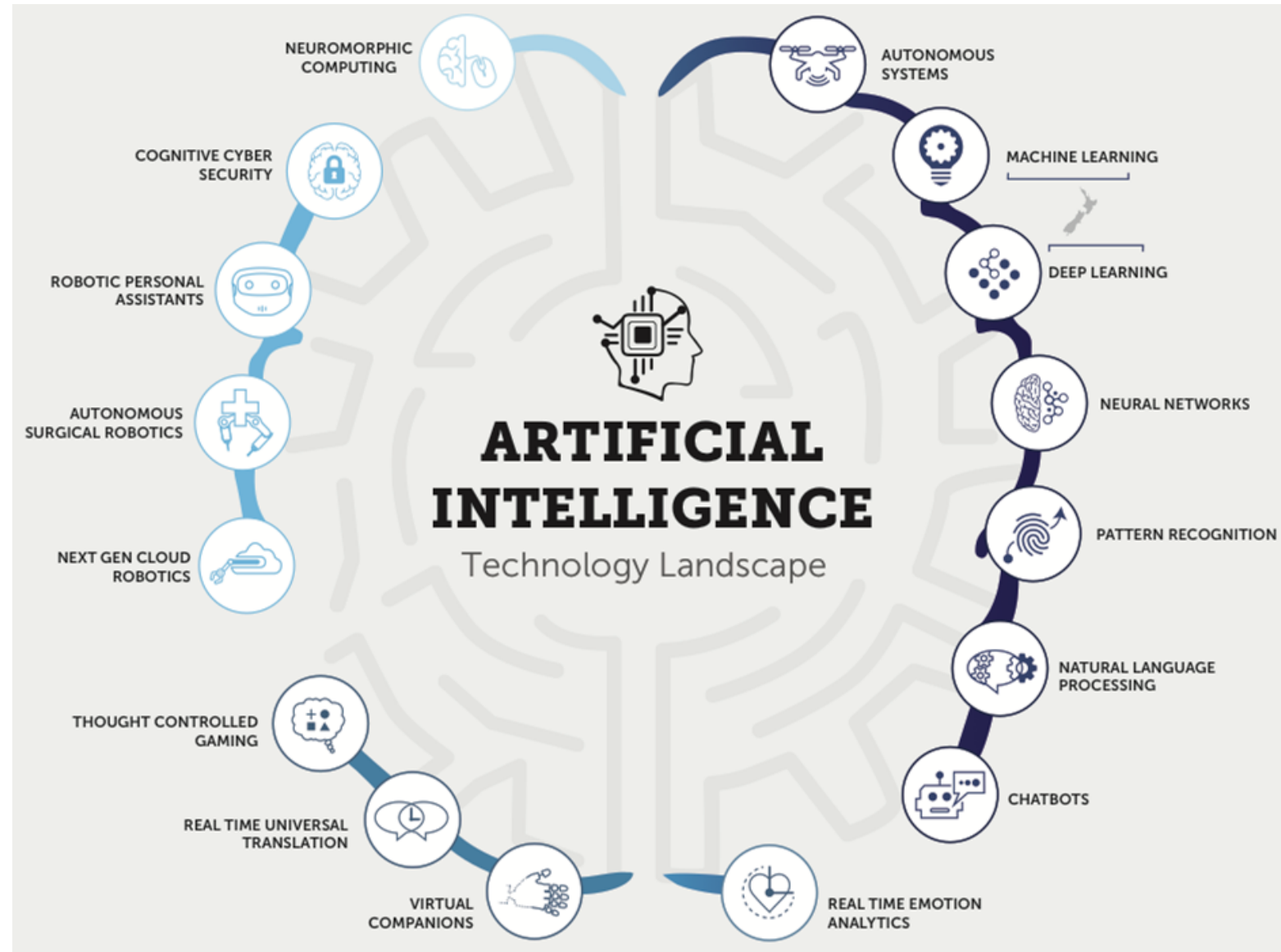
# Machine learning



<http://www.favouriteblog.com/15-algorithms-machine-learning-engineers/>



# Umělá inteligence

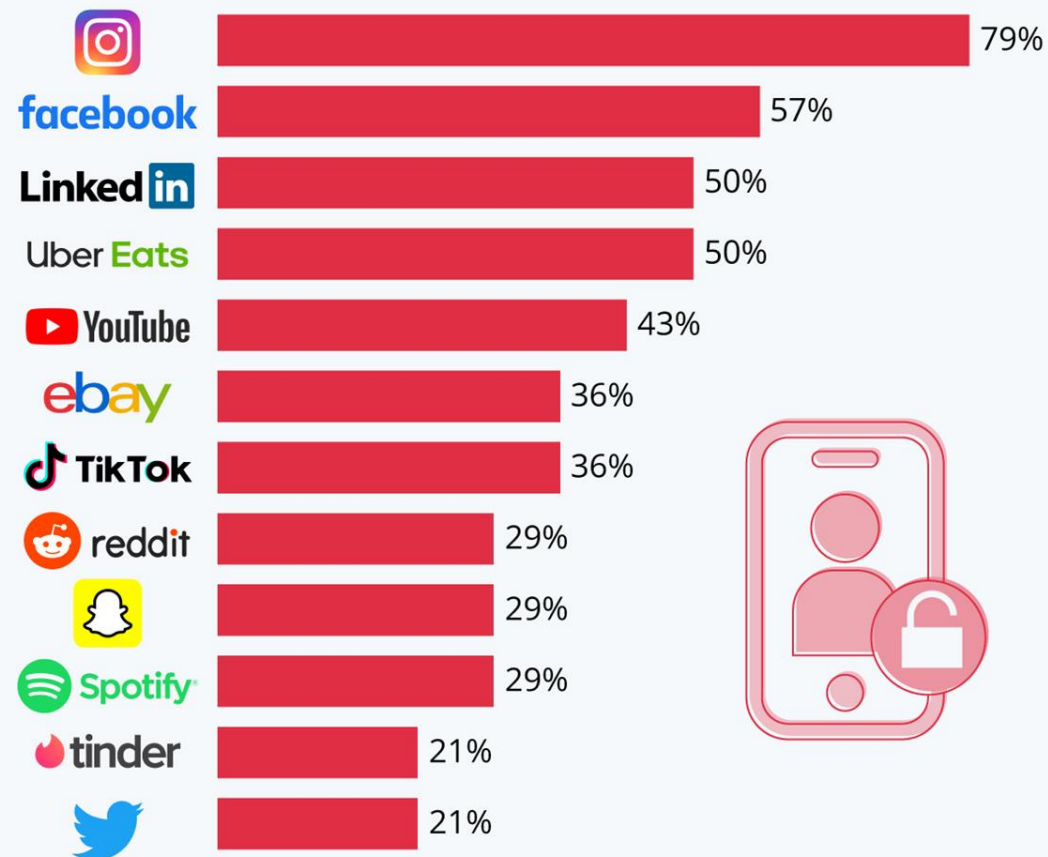


# Hodnota dat

- Kde lze využít data
  - reklama,
  - predikce,
  - pochopení zákazníka,
  - zajištění bezpečnosti anomaly detection

## Personal Data: Instagram Is a Real Tattletale

Percentage of personal data categories shared with third parties by selected iOS apps\*



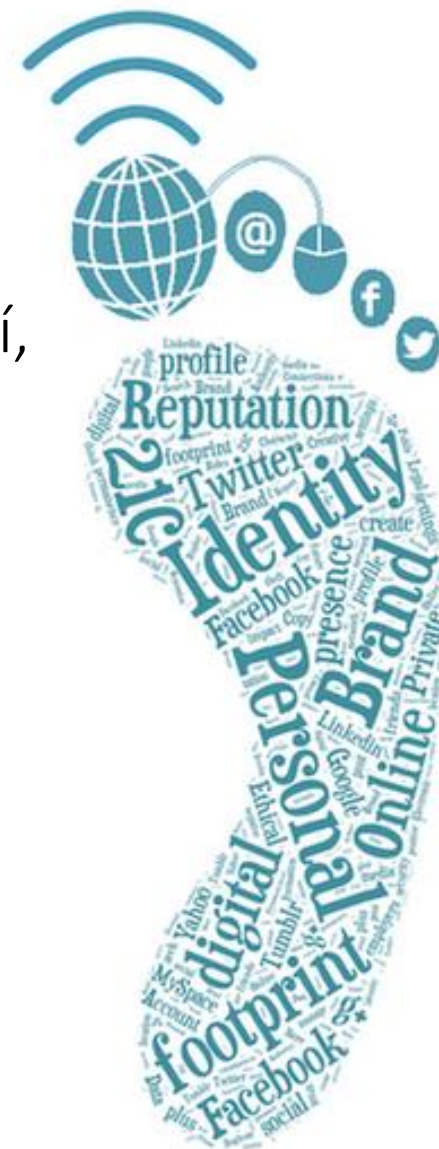
\* Based on privacy labels in the App Store, which group user data into 14 categories and inform users what data a given app collects and how it is used.

Source: pCloud



# Digitální stopa uživatele

- Digitální stopa je soubor sledovatelných digitálních aktivit, akcí, příspěvků a komunikací projevovaných na internetu nebo digitálních zařízeních.
- Může mít vliv na např.
  - Přizpůsobení obsahu a reklamy (cookies)
  - Hodnocení důvěryhodnosti
  - Screening kandidátů při pohovorech

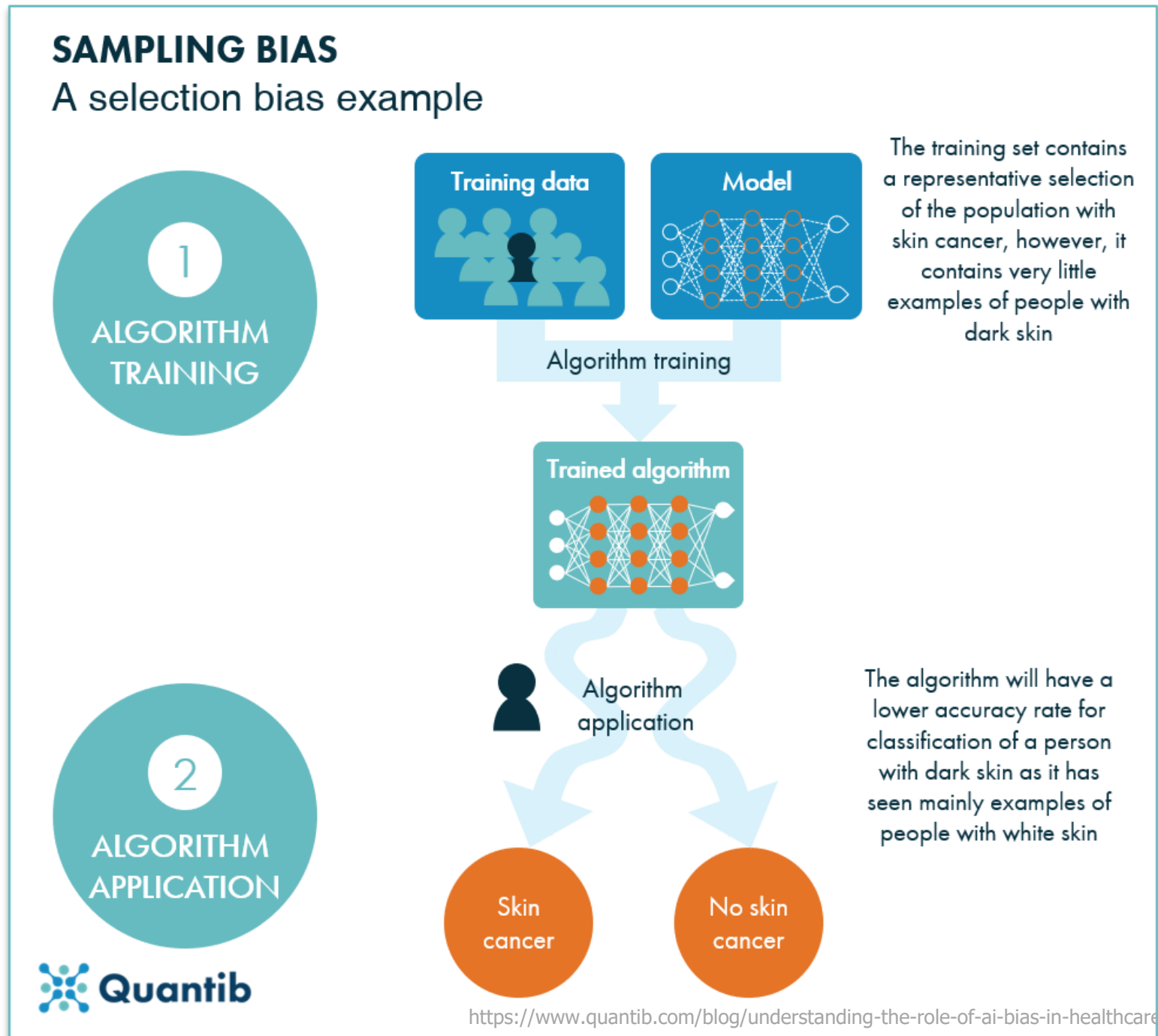


# Digitální stopa uživatele

- Zanechává obrázek o tom, kým člověk je
- Komentáře na sociálních sítích, prohlížená videa, soukromé zprávy vyměňované s přáteli, online hovory, prohlížení stránek
- Zkuste si sami sebe vložit do Googlu
- Co byste jako výsledek vyhledávání chtěli najít za 10 let?
- Riziko krádeže identity

# Kontroverze

- Etické kontroverze
  - manipulace,
  - algorithmic bias,
  - příklady ze života



# CO NÁS ČEKÁ PŘÍŠTĚ

# 8. Kvalita softwarových systémů

- Co je to kvalita, jak ji definujeme
- Jak předcházet problémům s kvalitou
- Jak kontrolovat úroveň kvality
- Testování a jeho role ve vývoji SW

## Domácí práce a příprava na příští přednášku

- Procvičení (volitelné) Tableau Public <https://public.tableau.com/en-us/s/download>  
s využitím tutoriálů <https://public.tableau.com/en-us/s/resources>
- Článek [5 of the Biggest Information Technology Failures and Scares](#)
- Článek [Do remote teams deliver lower quality software?](#)