MUNI
FI

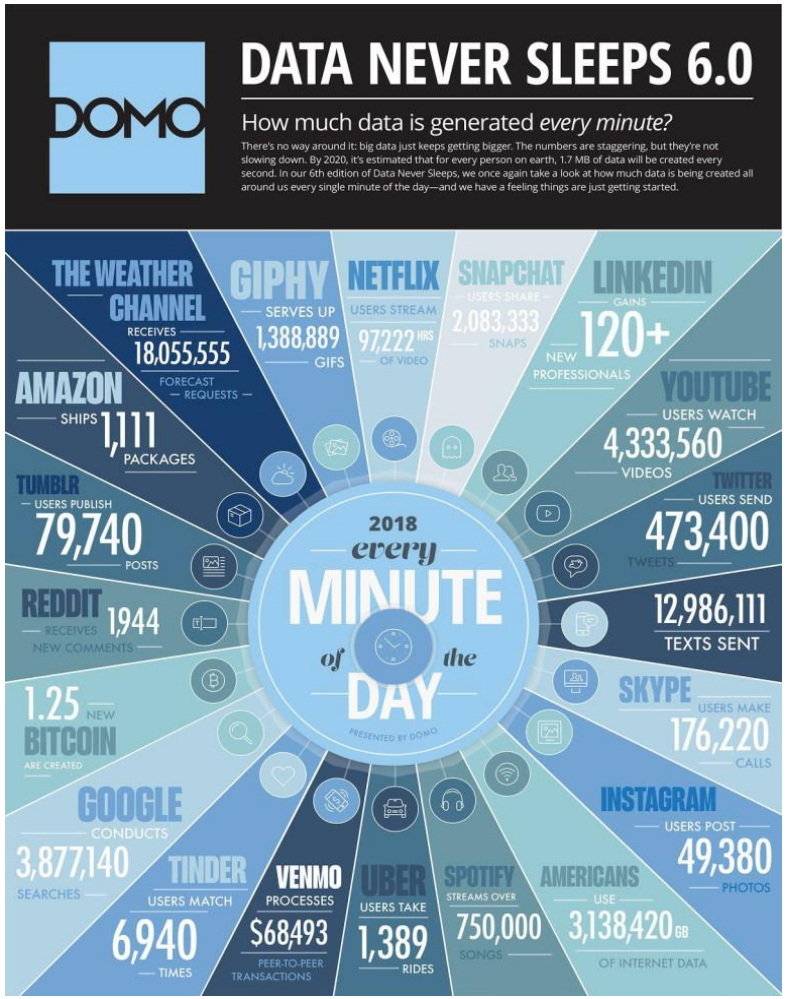
Přednáška 7
Práce s daty

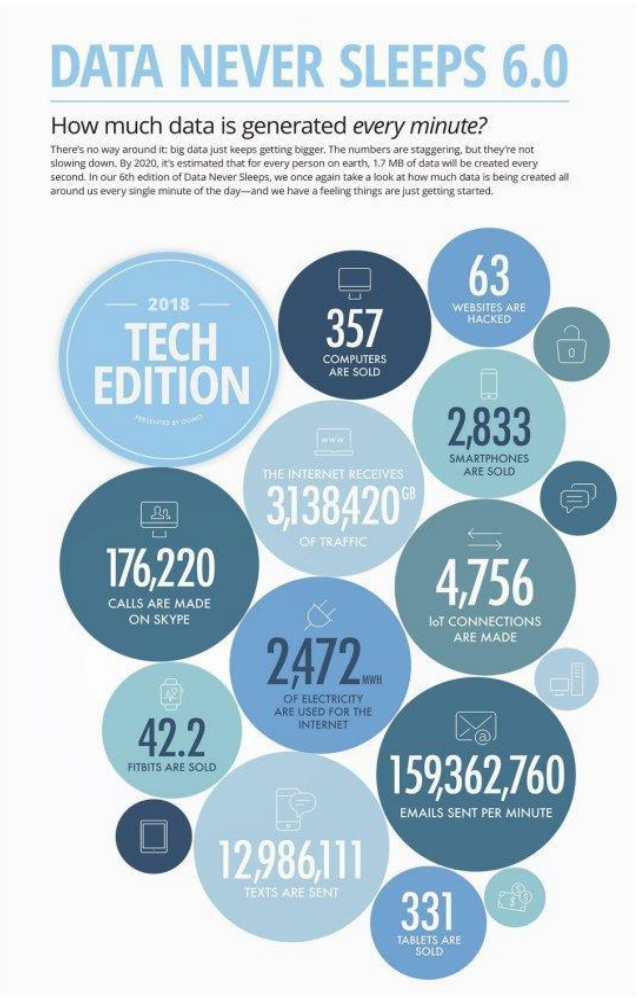
7. Práce s daty

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

21. STOLETÍ – DOBA DATOVÁ

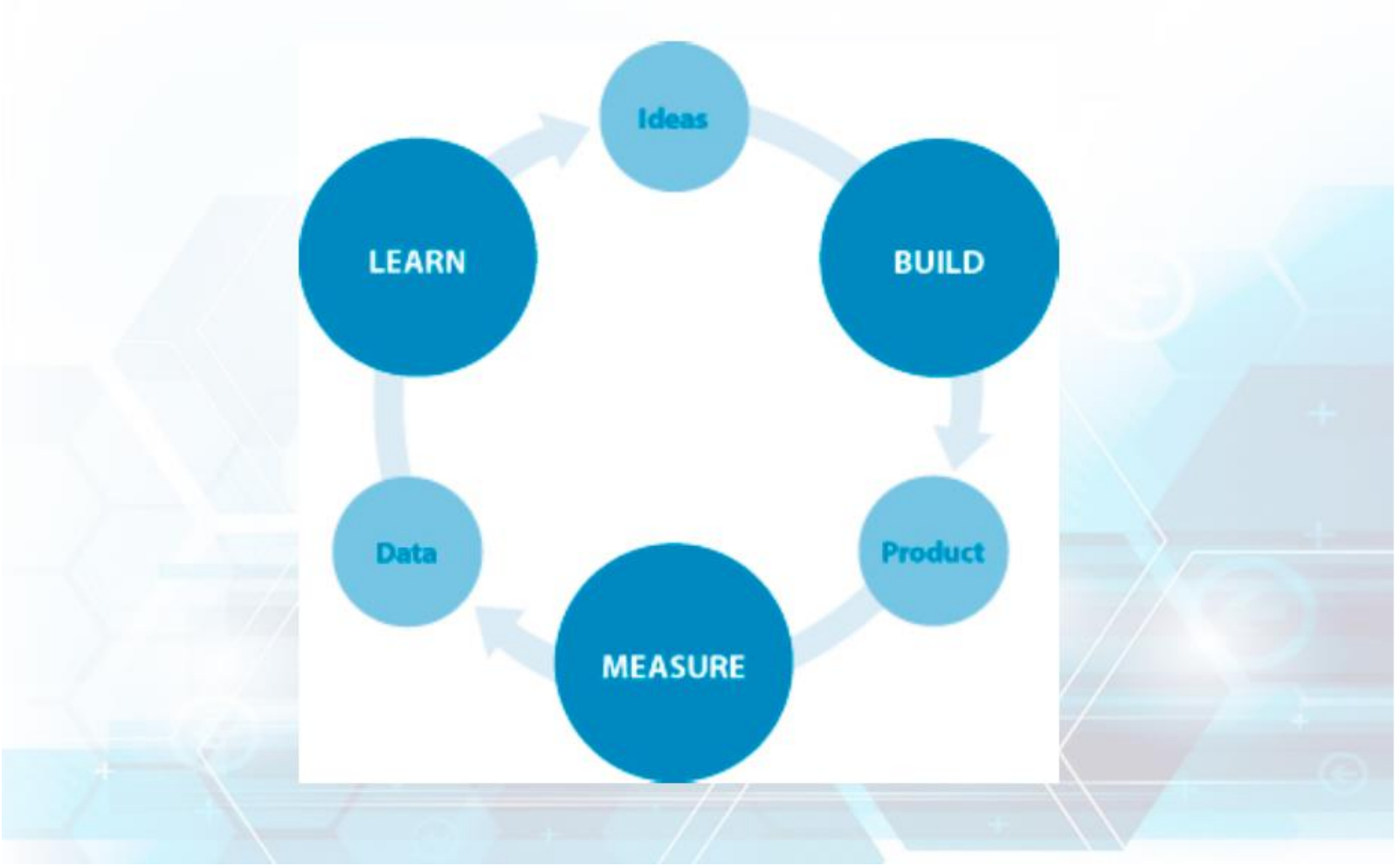
Motivace

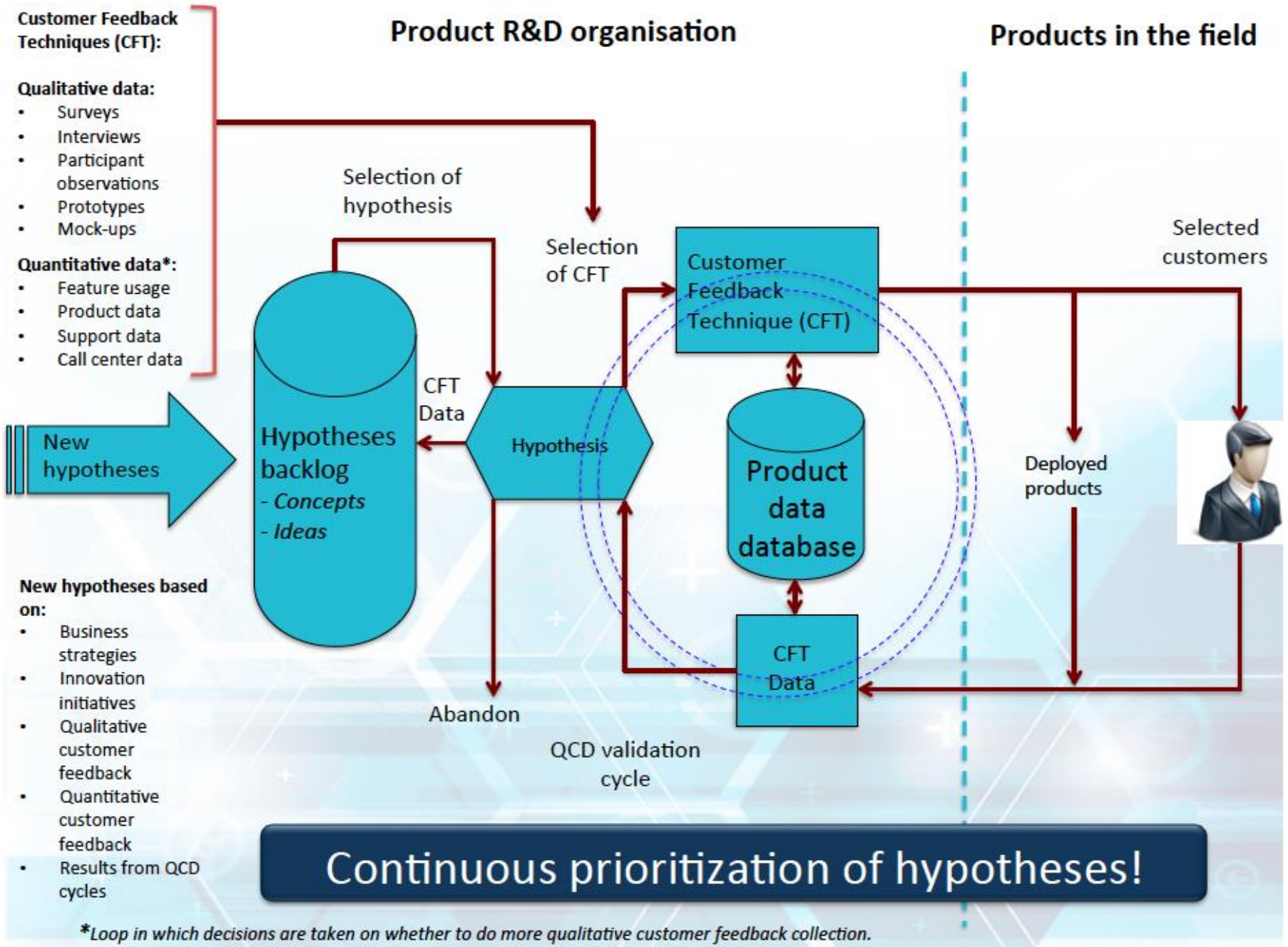




MUNI FI

Koloběh inovací

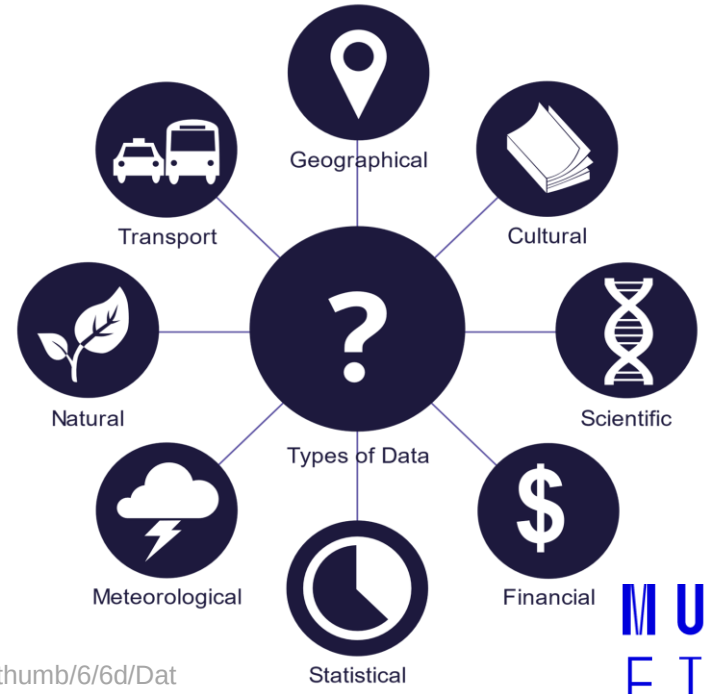



Customer Feedback Techniques (CFT):
Qualitative data:
• Surveys
• Interviews
• Participant observations
• Prototypes
• Mock-ups
Quantitative data*:
• Feature usage
• Product data
• Support data
• Call center data
Product R&D organisation
Products in the field
Selection of hypothesis
Selection of CFT
Customer Feedback Technique (CFT)
Selected customers
New hypotheses
Hypotheses backlog
- Concepts
- Ideas
CFT Data
Hypothesis
Product data database
Deployed products
CFT Data
Abandon
QCD validation cycle
New hypotheses based on:
• Business strategies
• Innovation initiatives
• Qualitative customer feedback
• Quantitative customer feedback
• Results from QCD cycles
Continuous prioritization of hypotheses!
*Loop in which decisions are taken on whether to do more qualitative customer feedback collection.

(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

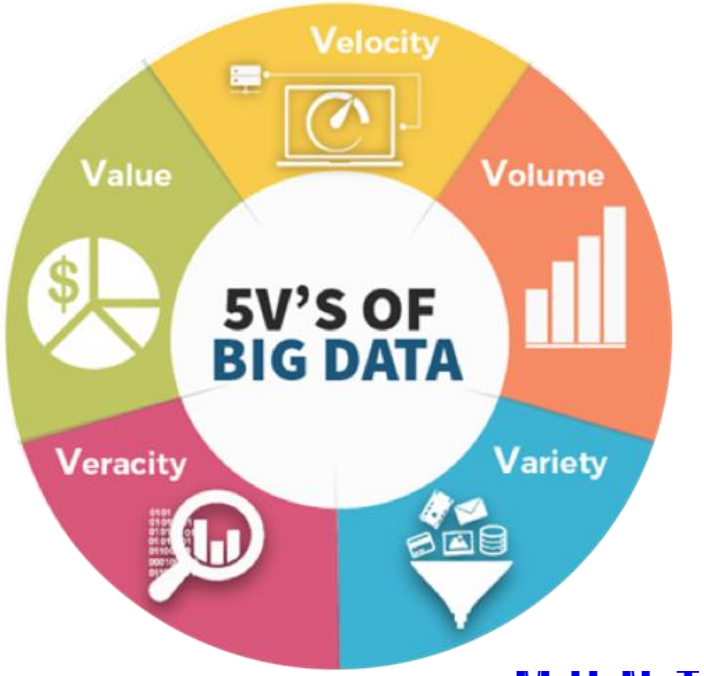


https://upload.wikimedia.org/wikipedia/commons/thumb/6/6d/Data_types_-_en.svg/300px-Data_types_-_en.svg.png

MUNI FI

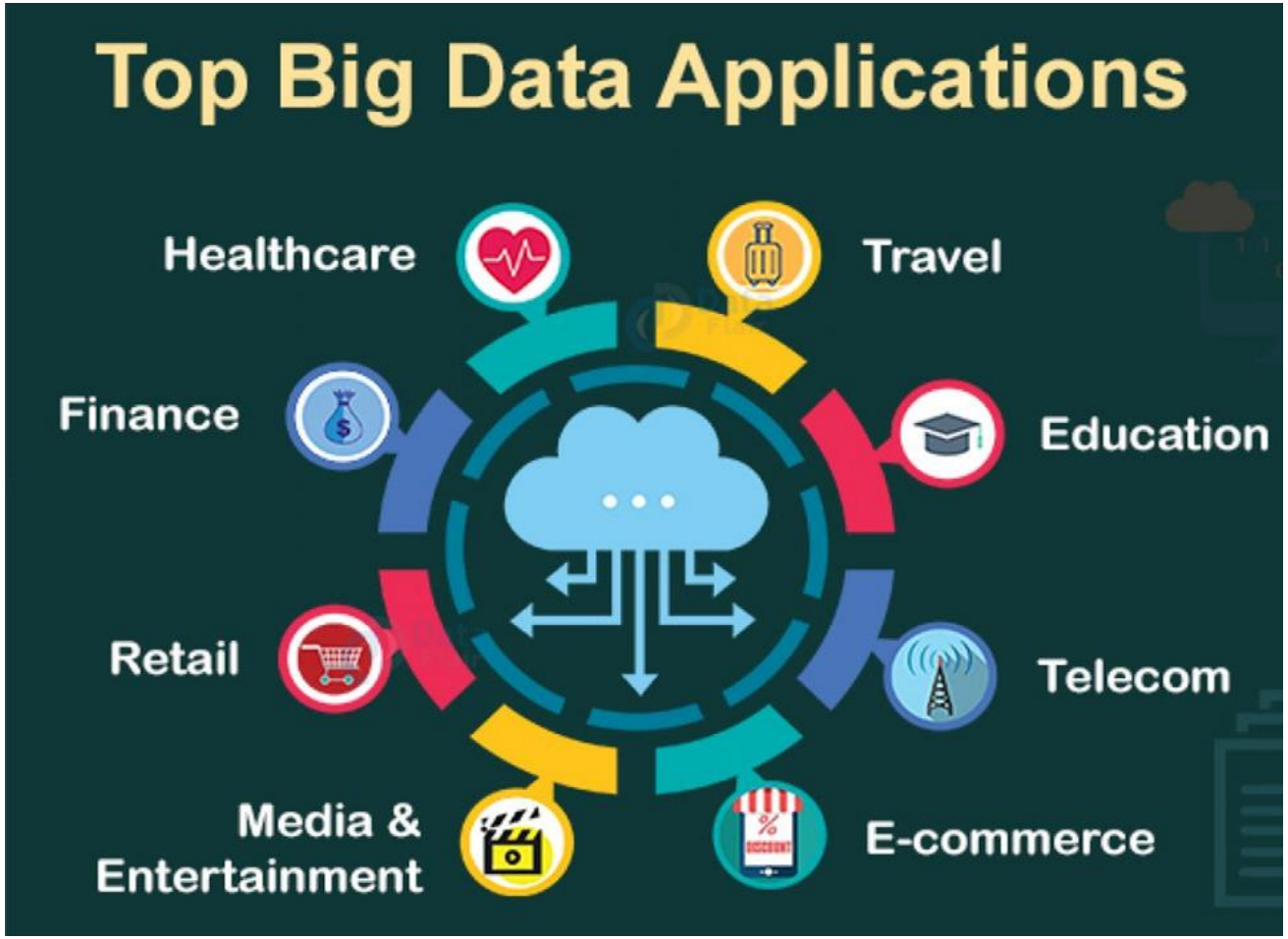
Velká data

- **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:
  - **Objem** - velikost dat je obrovská
  - **Rozmanitost** - různé zdroje a formát dat
  - **Variabilita** - údaje mohou být nekonzistentní a nepředvídatelné
  - **Rychlost** - data jsou generována velmi rychle
  - **Věrohodnost** - data jsou ověřena a validována.



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

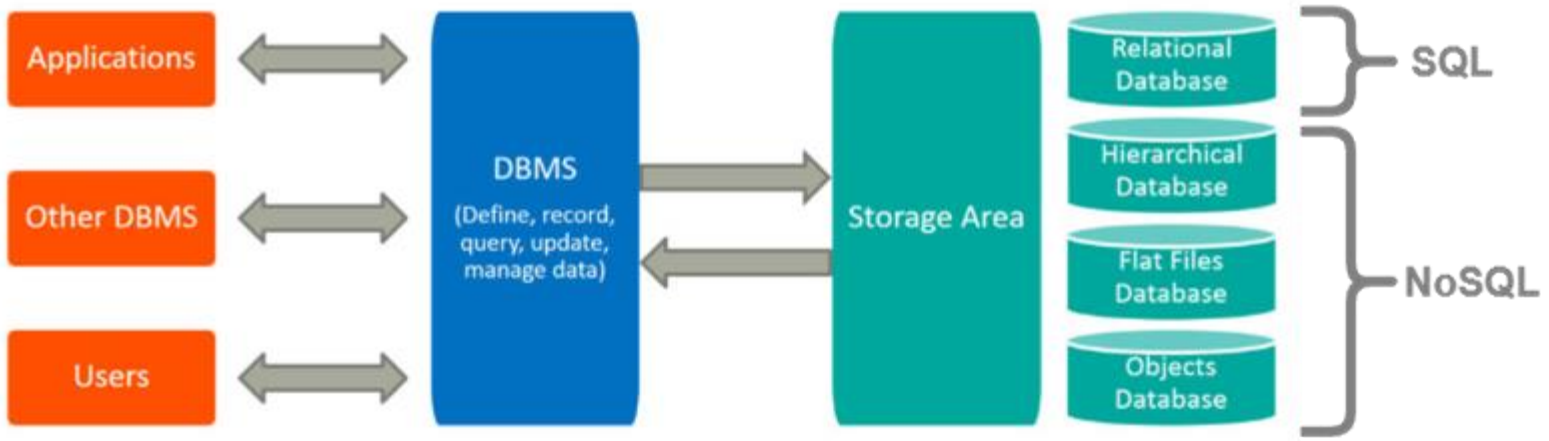
MUNI FI

DATABÁZE

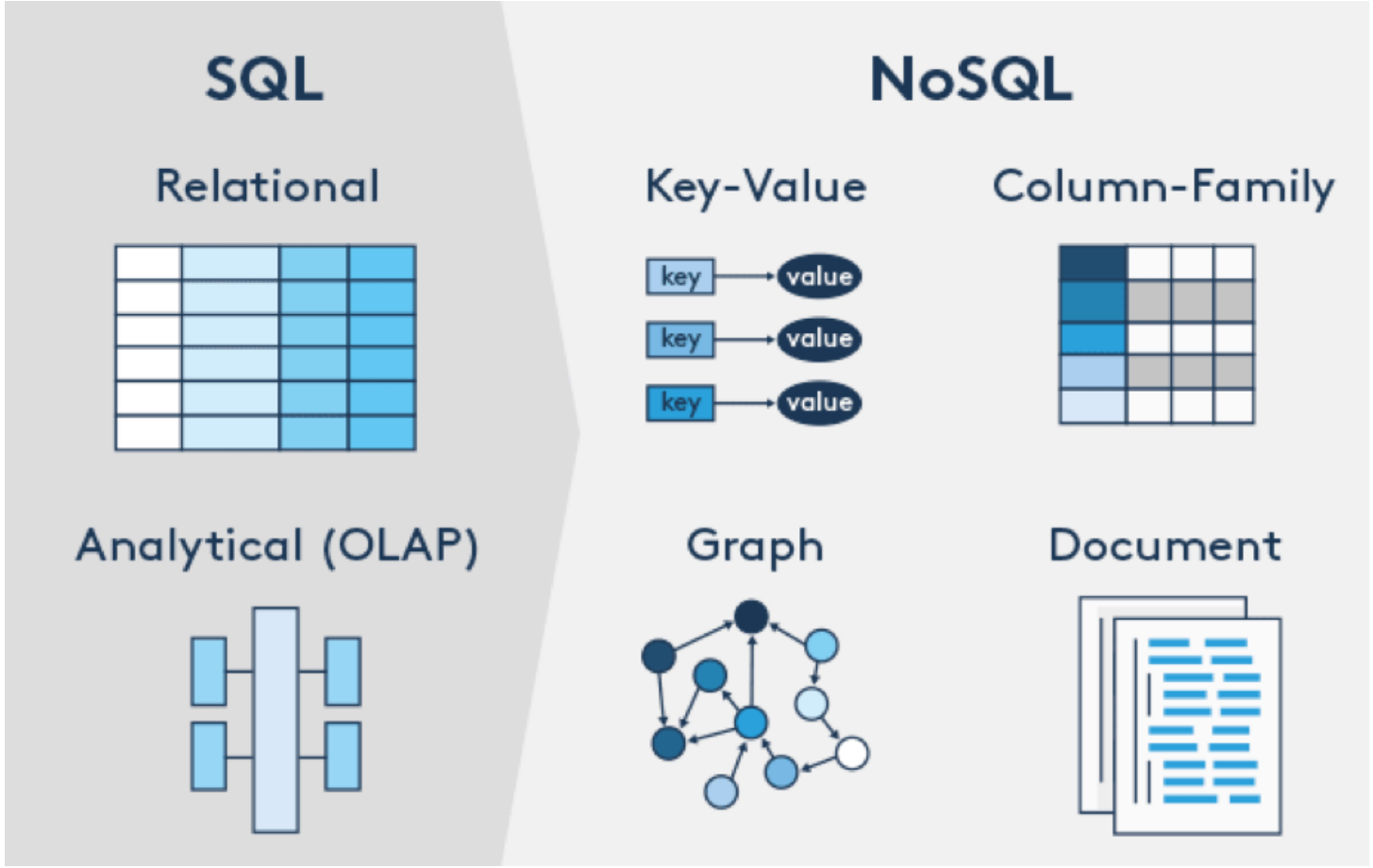
MUNI FI

Databáze

- Uspořádané uložiště
- 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

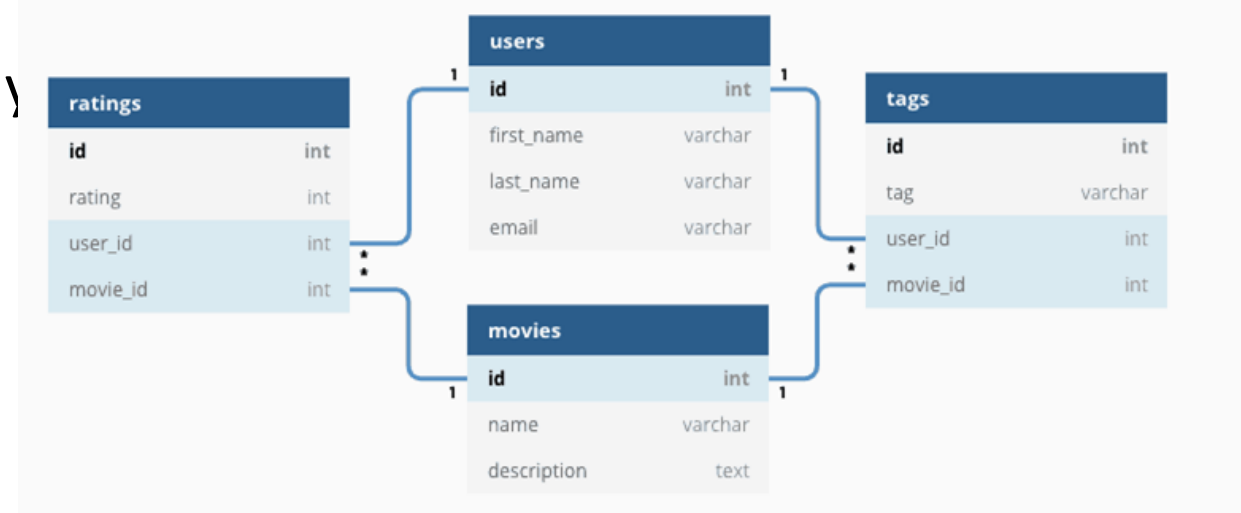
MUNI FI

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



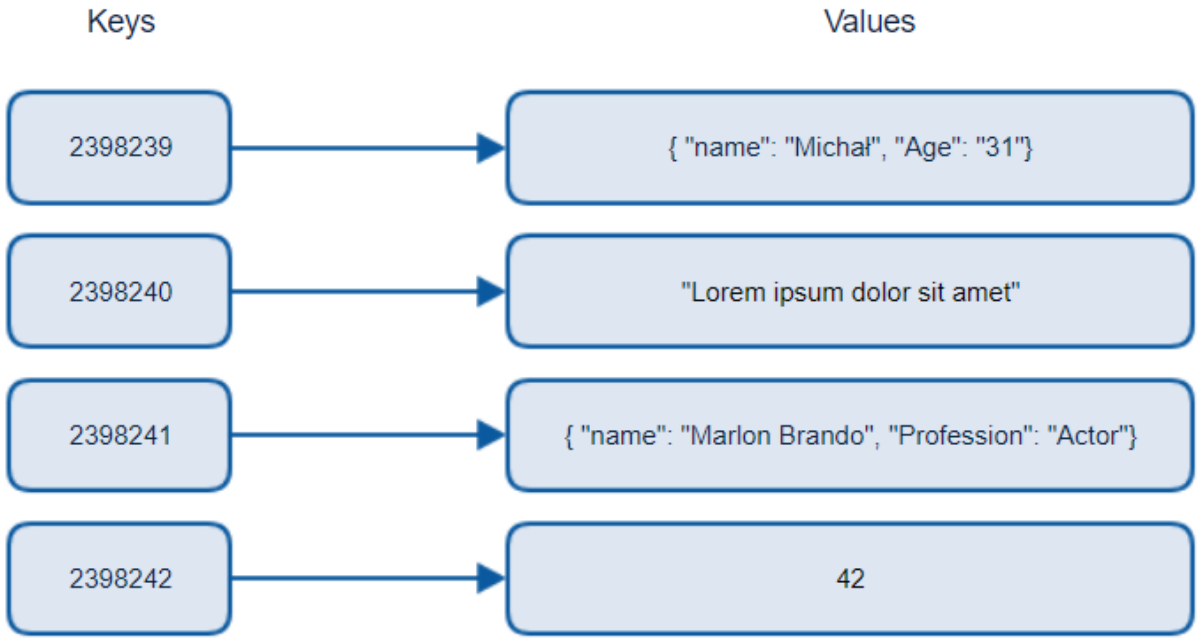
https://assets-global.website-files.com/5debb9b4f88fbc3f702d579e/5e3c1a71724a38245aa43b02_99bf70d46cc247be878de9d3a88f0c44.png

Databáze klíč-hodnota

- Nerelační 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

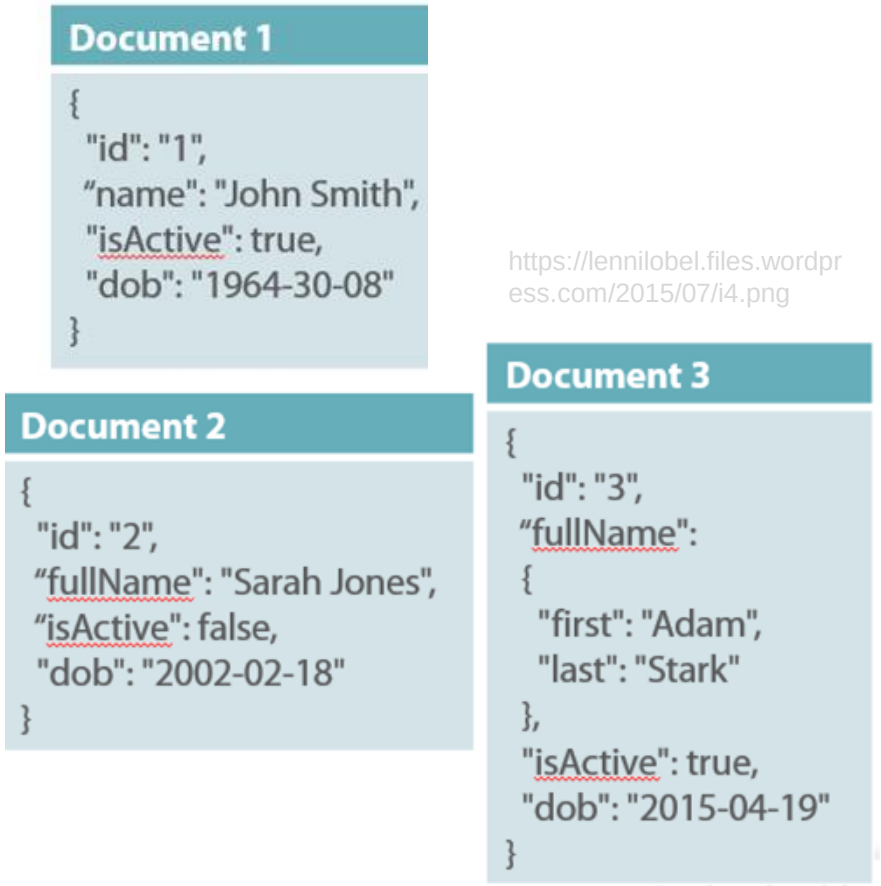
MUNI FI

Dokumentové databáze

- Nerelační 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



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

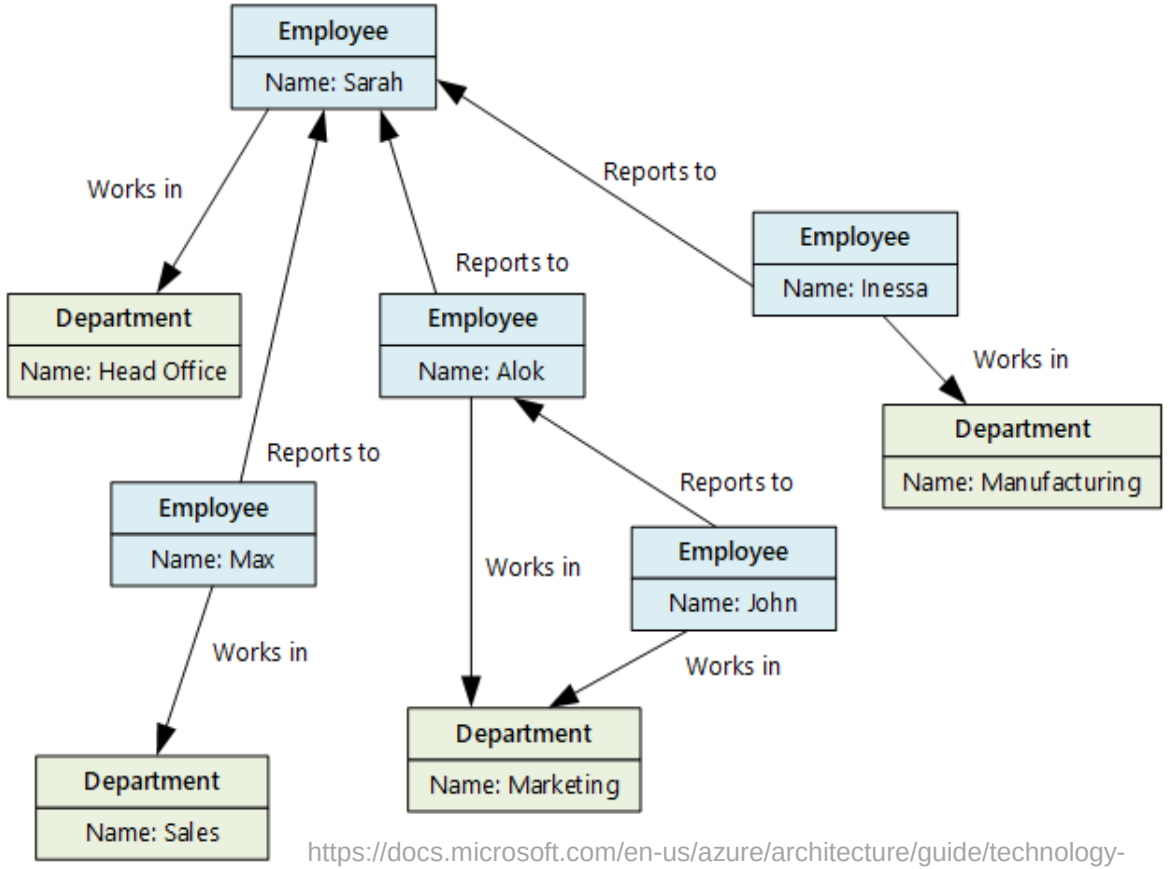
MUNI FI

Grafové databáze

- Nerelační 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

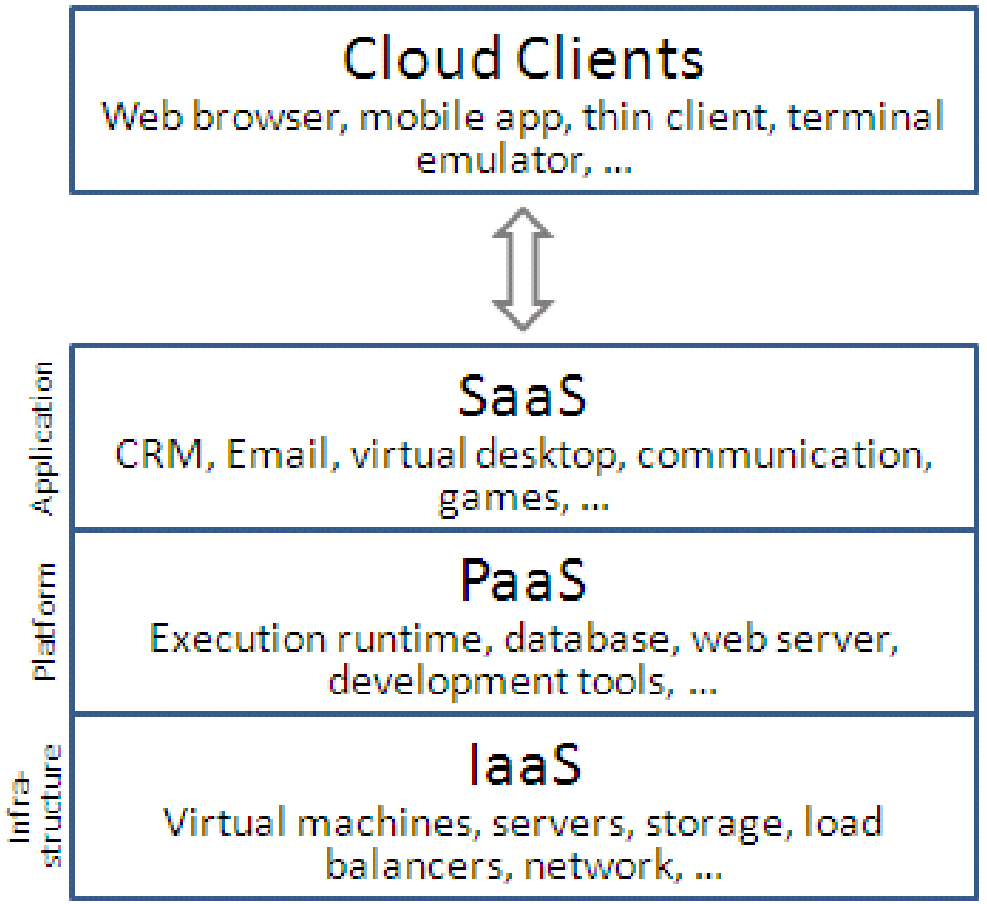
Cloud Computing

- 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

MUNI FI

Cloud Computing

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



MUNI FI

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

MUNI FI

Životní cyklus dat

- Přehled fází úspěšné správy a uchovávání dat pro použití a opakované použití.
  - **Snímání / vytváření dat**
  - **Údržba dat**
  - **Používání dat**
  - **Zveřejnění dat**
  - **Archivace dat**
  - **Čištění / ničení dat**

https://www.nexor.com/wp/wp-content/uploads/2017/03/Enabling-Secure-Information-Exchange-in-Cloud-environments-white-paper-diagram-1.png

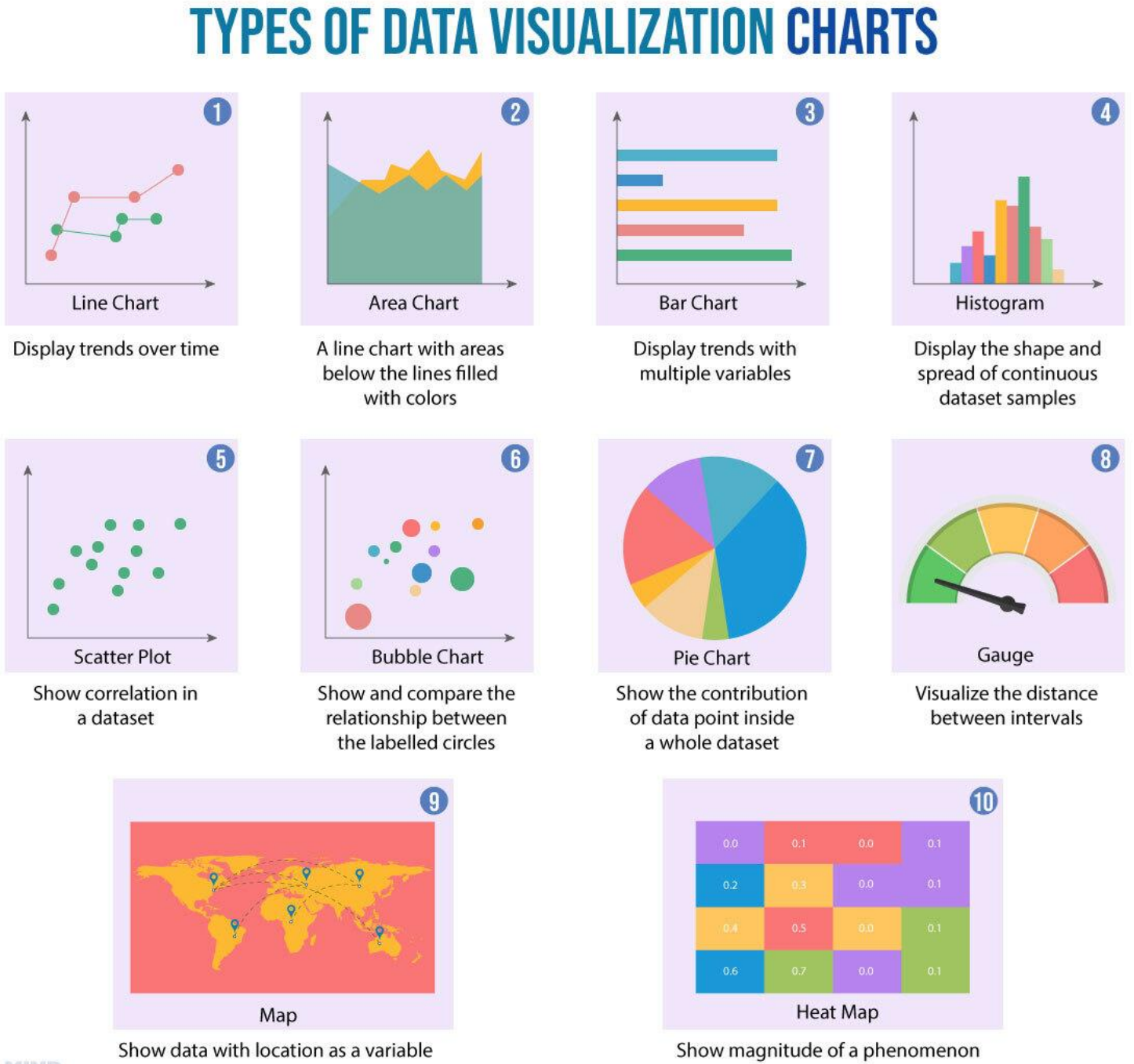
ZPRACOVÁNÍ DAT (DATA SCIENCE)

MUNI FI

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.
AIHR
BLOG & ACADEMY
https://medium.com/@nicklauswinters/data-cleaning-22b790eea834

Vizualizace dat

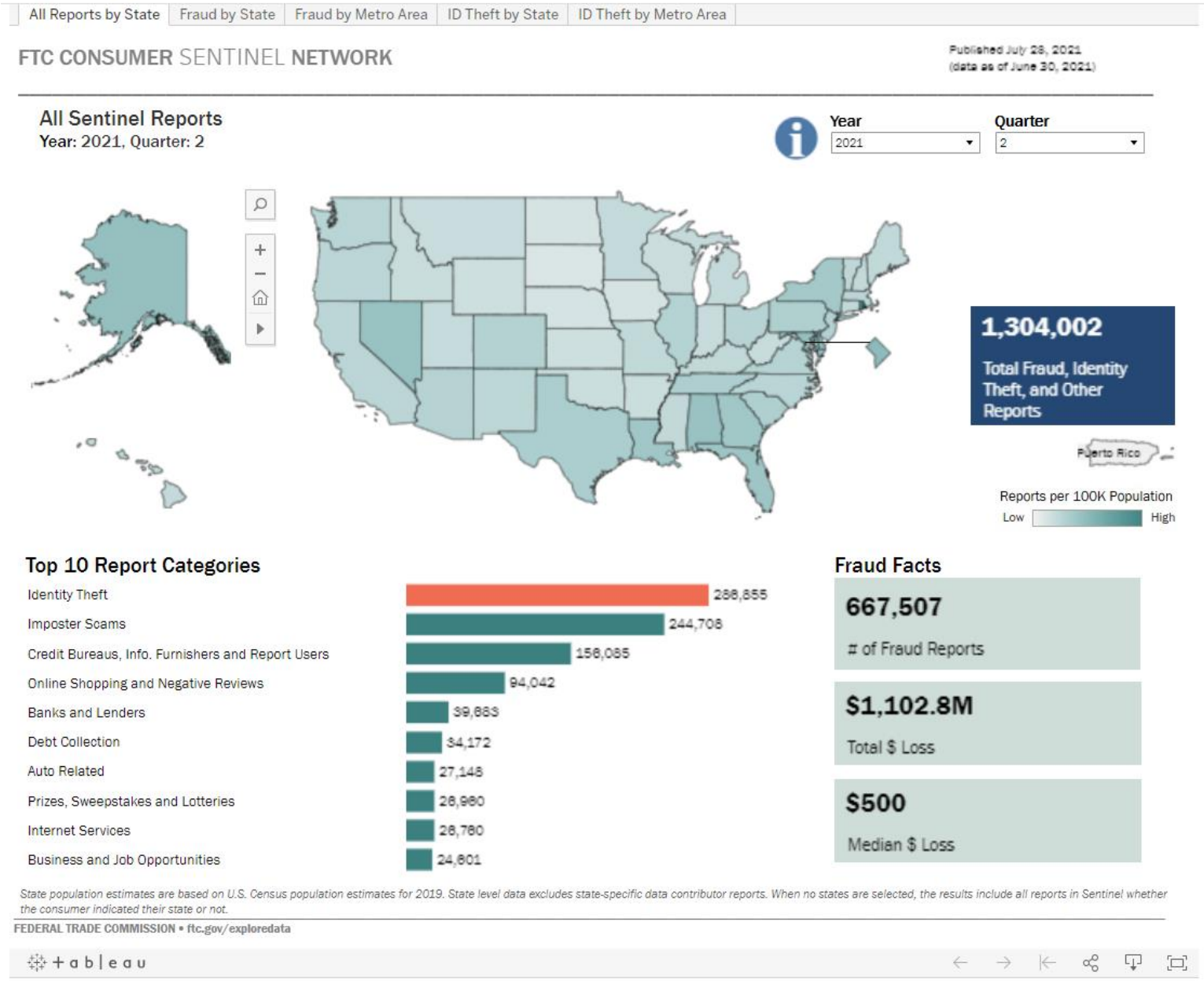


Vizualization Dashboards

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

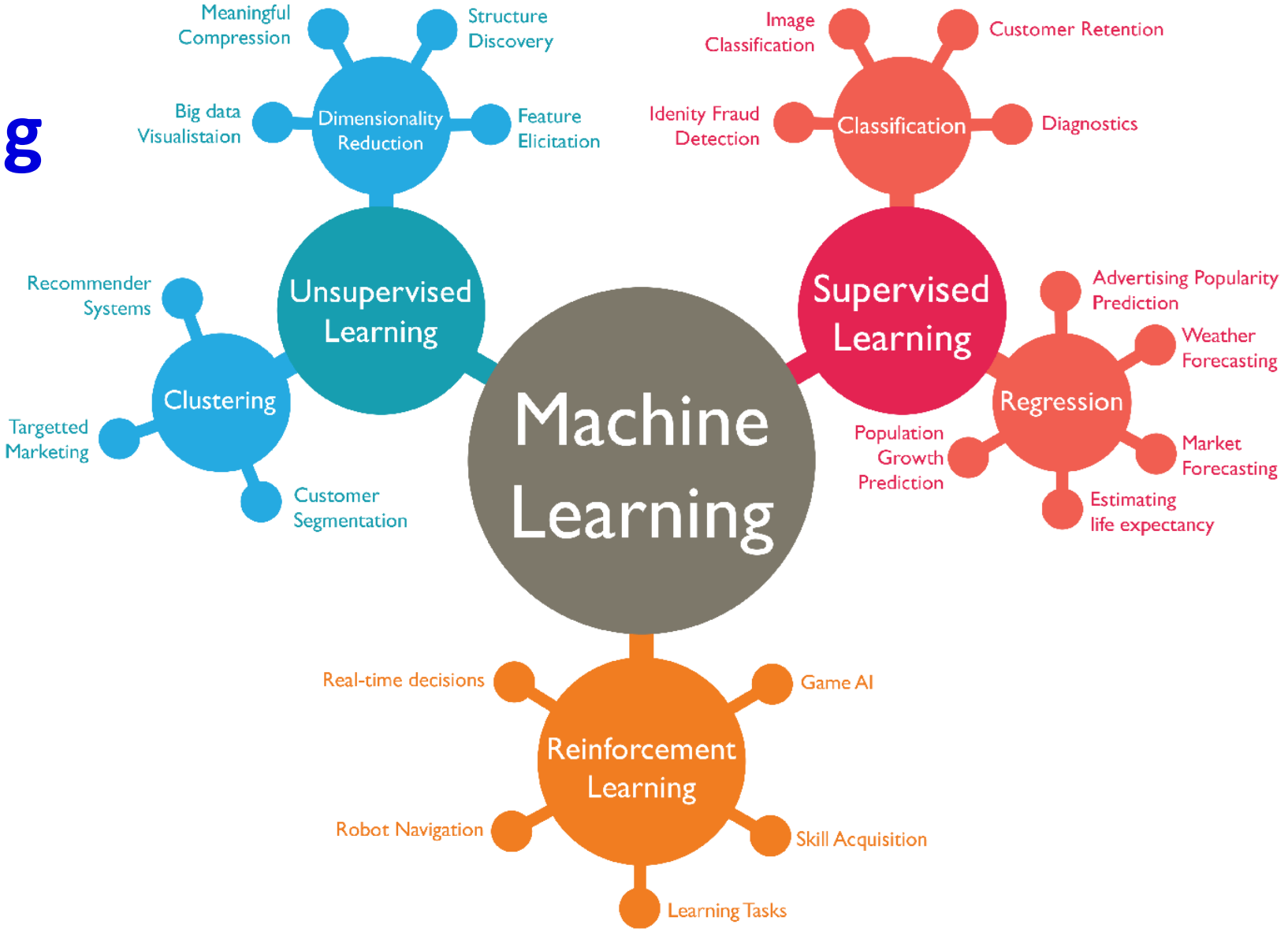
Tableau Public

- Tableau Public access
  https://public.tableau.com/en-us/s/download
- Tutorials
  https://public.tableau.com/en-us/s/resources



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

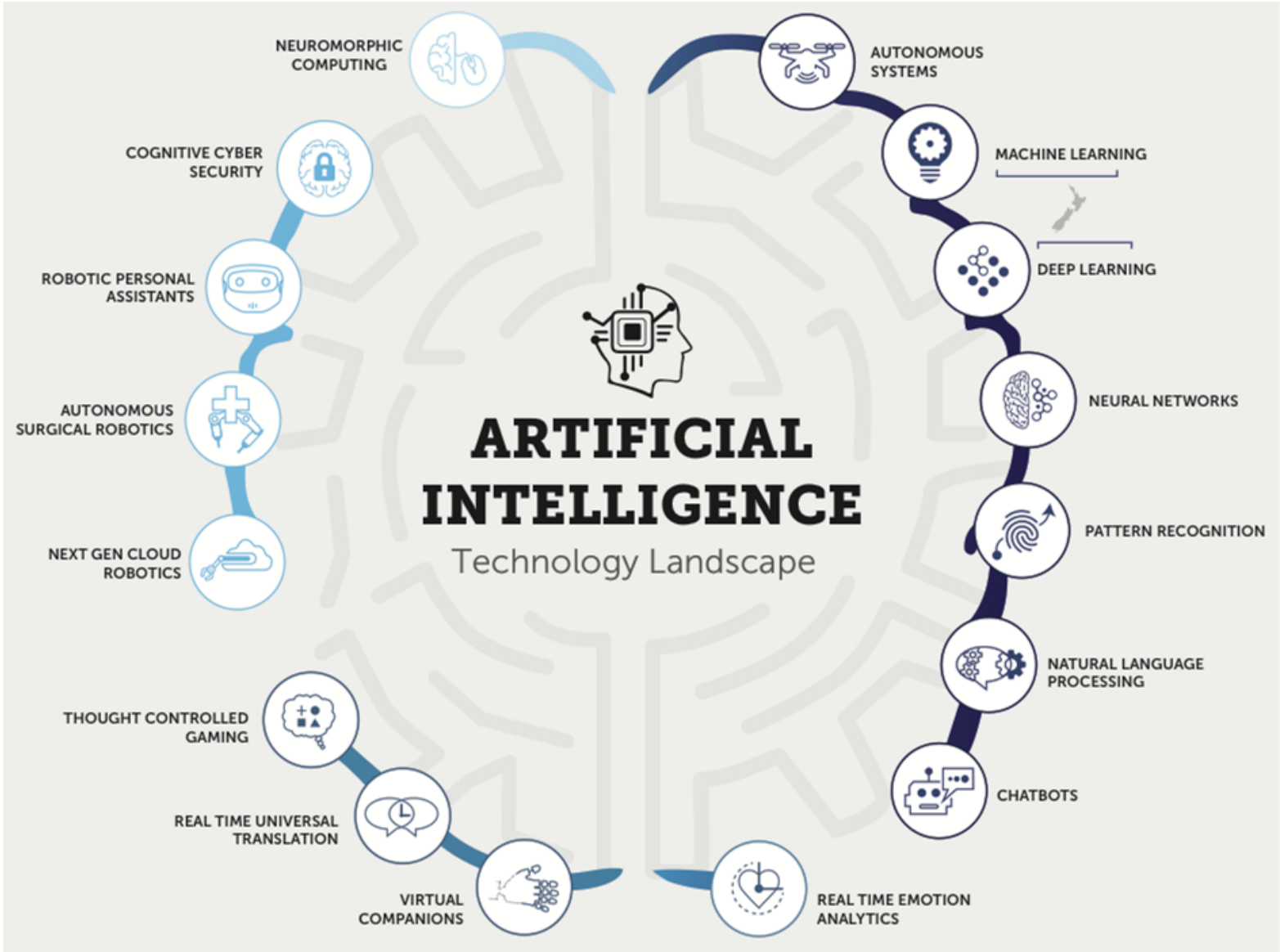
Machine learning



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

MUNI FI

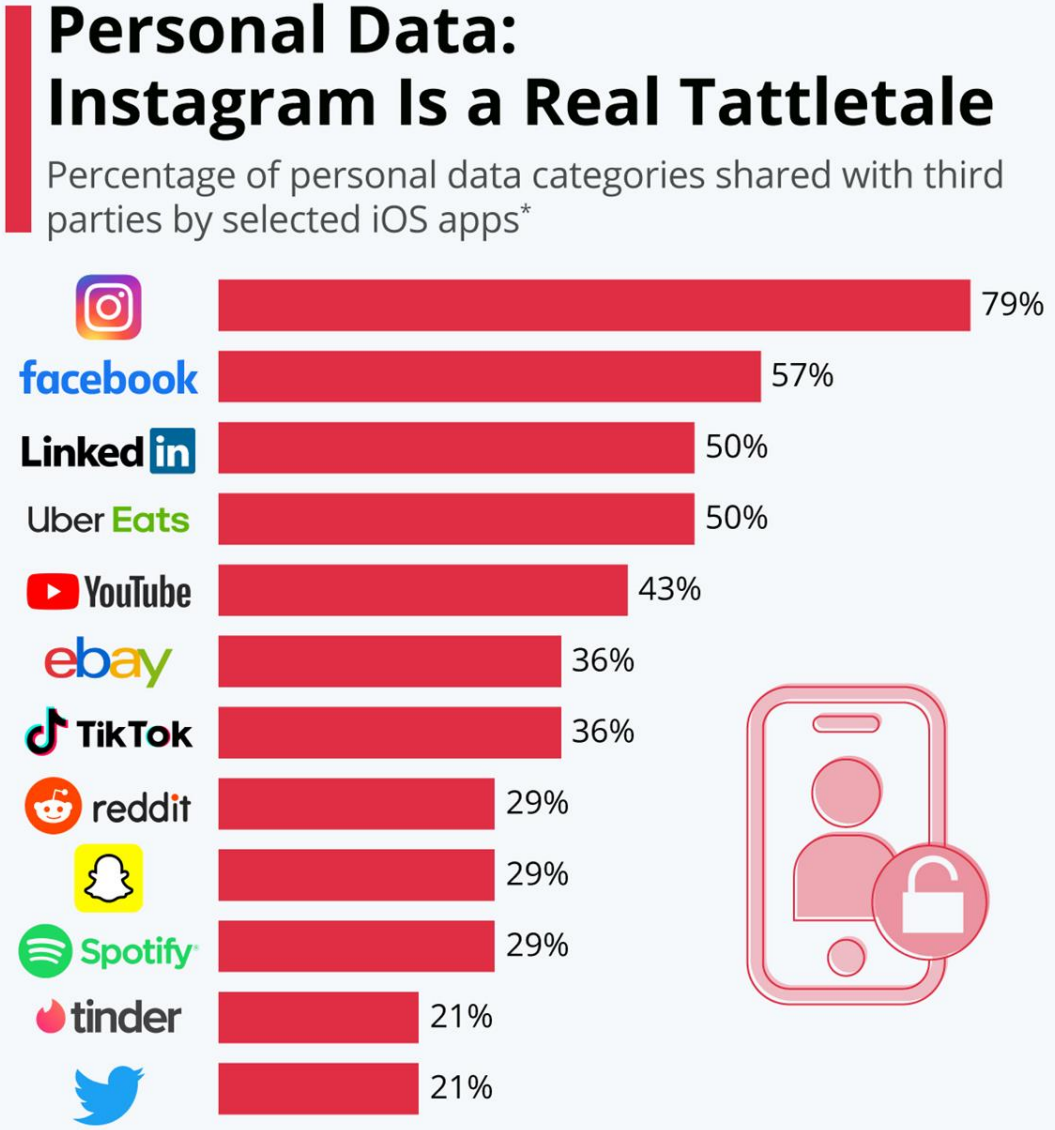
Umělá inteligence



https://miro.medium.com/max/1400/1*xiTy-b6dvVU8sDeWROcamg.png

Hodnota dat

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



https://www.statista.com/chart/24495/apps-sharing-personal-information-with-third-parties/

Digitální stopa uživatele

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

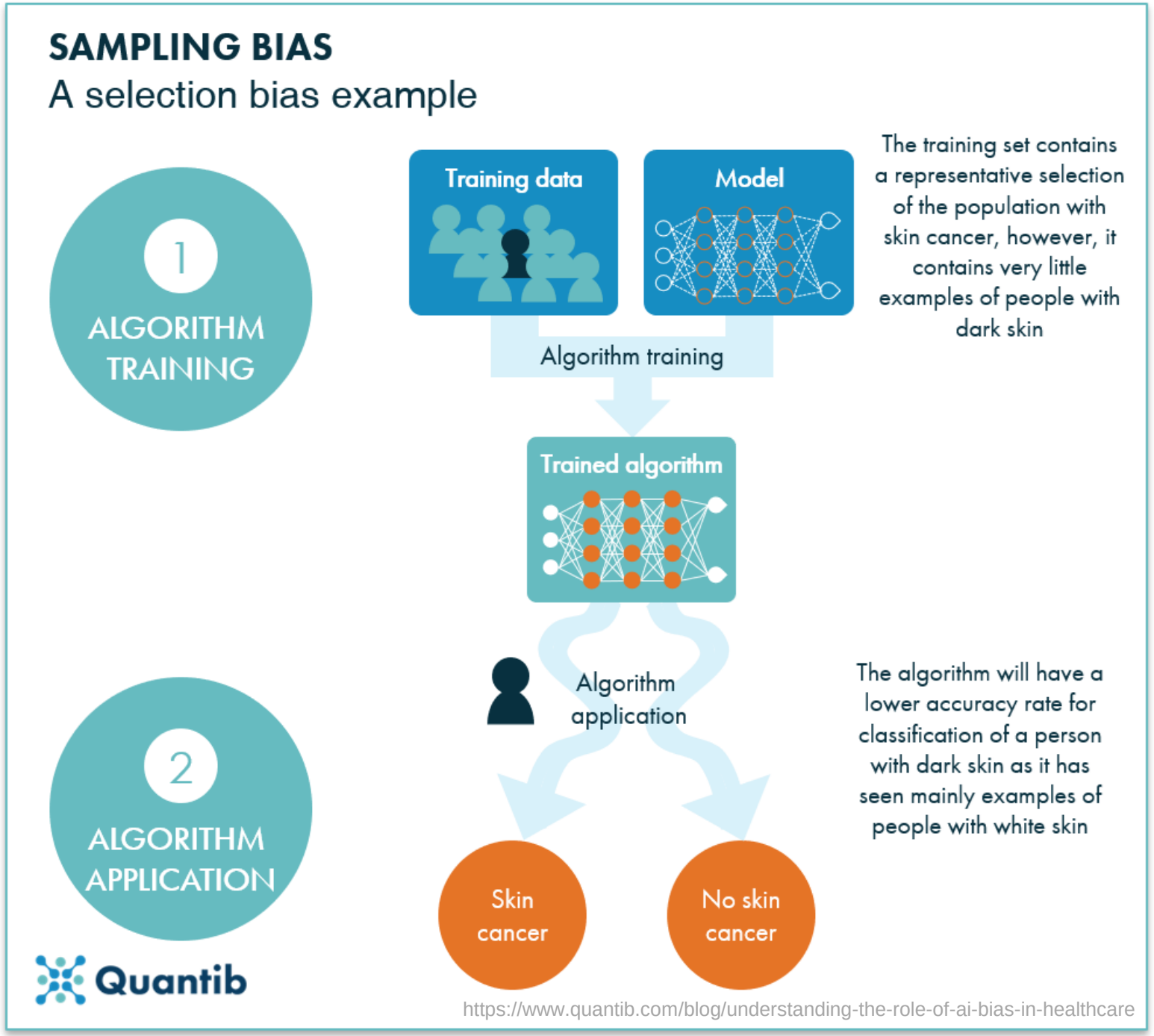
MUNI FI

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,
  - argorithmic bias,
  - příklady ze života



CO NÁS ČEKÁ PŘÍŠTĚ

MUNI FI

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?

MUNI FI