



Data Science and Big Data

Oracle Junior Program - IT technológiák és architektúrák nagyvállalati környezetben

Szakai Ádám

Miről lesz ma szó

- ▶ Fogalmak tisztázása
 - ▶ DataScience
 - ▶ Mesterséges intelligencia
 - ▶ Machine learning
 - ▶ Deep learning
 - ▶ Big Data
- ▶ Példák

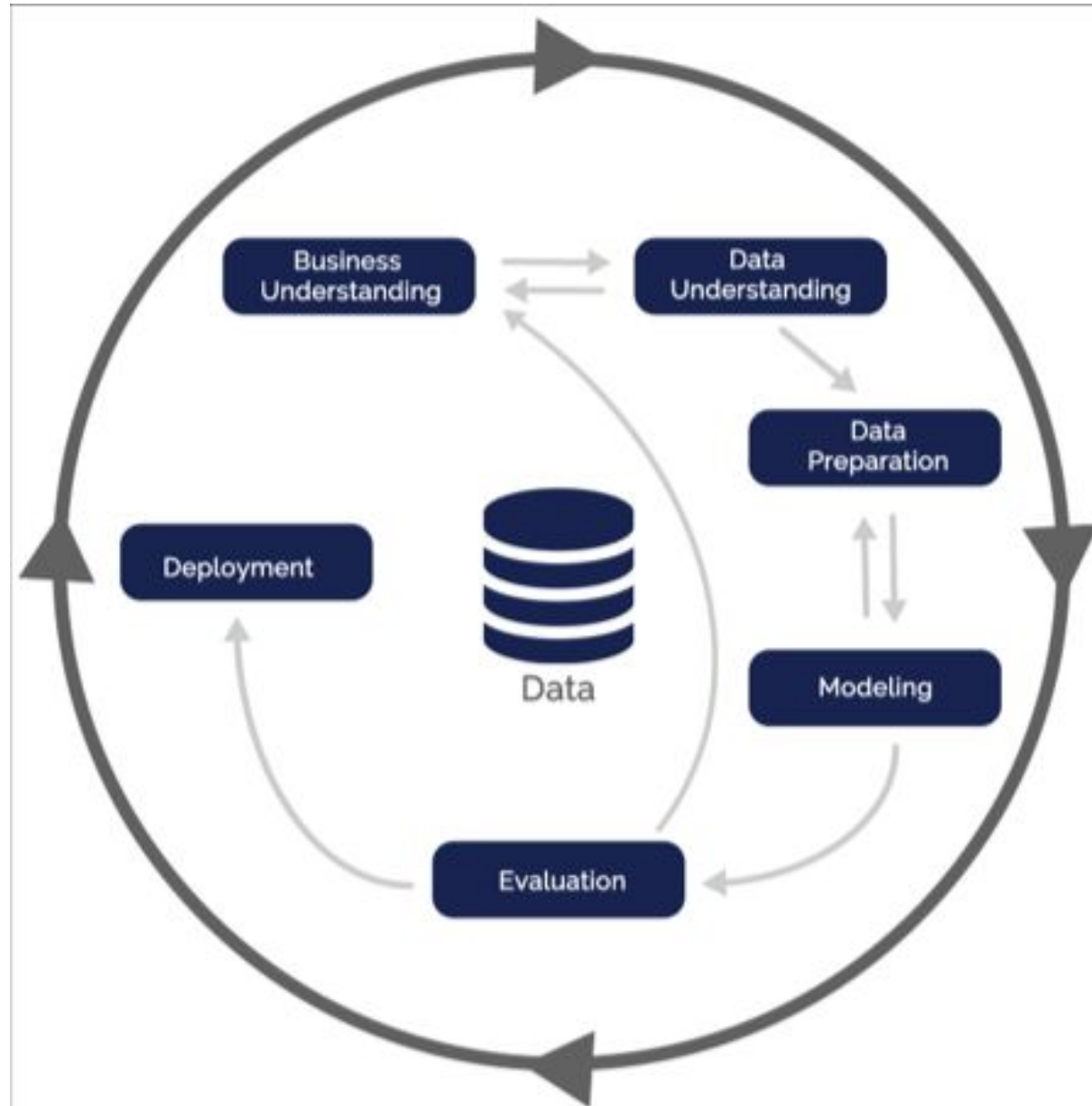
Definíciók

- ▶ “Data science provides meaningful information based on large amounts of complex data or big data. Data science, or data-driven science, combines different fields of work in statistics and computation to interpret data for decision-making purposes.” - Investopedia
- ▶ “The use of scientific methods to obtain useful information from computer data, especially large amounts of data” - Cambridge dictionary
- ▶ Fancy name for statistics?

Data Scientist

- ▶ 2020-ban is a “legszexibb” foglalkozás
- ▶ Statisztika/Matematika
- ▶ Programozás
- ▶ Domain tudás
- ▶ Machine learning algoritmusok ismerete
- ▶ Gyors tanulási képesség
- ▶ Egyetemeken info-s képzés szakirány

Mit csinál egy Data Scientist



Nehézségek

- ▶ Gyorsan fejlődő tudományág -> Folyamatos tanulást igényel
- ▶ Az adattisztítás időigényes és nem túl hálás feladat
- ▶ Az eredmények prezentálásához jó kommunikációs, tárgyaló és tanító készségek szükségesek
- ▶ Leleményesség és önálló munkavégzés
- ▶ Nem mindig állít elő kézzelfogható terméket

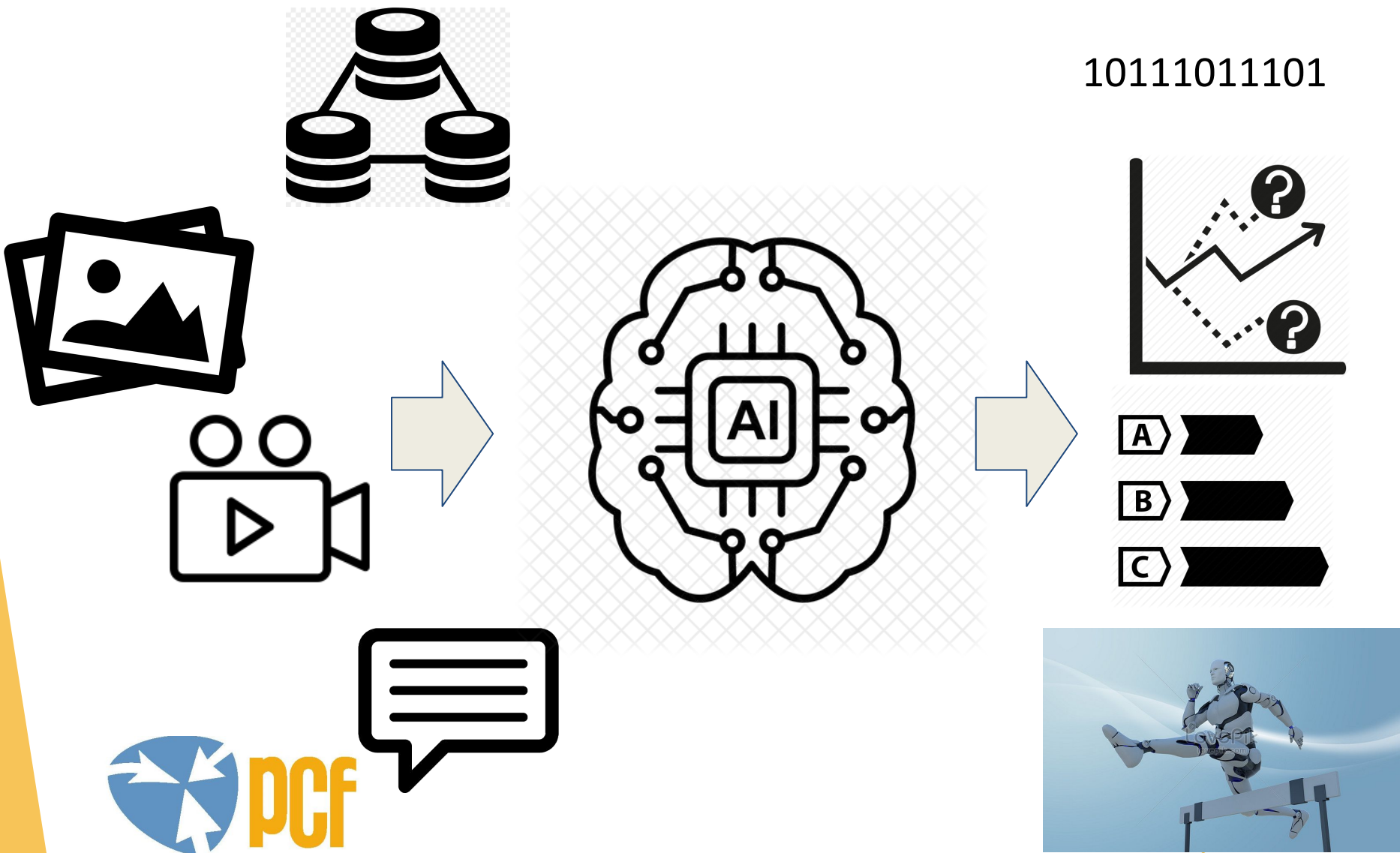


<https://towardsdatascience.com/4-reasons-why-you-shouldnt-be-a-data-scientist-e3cc6c1d50e>

Mesterséges intelligencia(AI)



Machine learning - Gépi tanulás

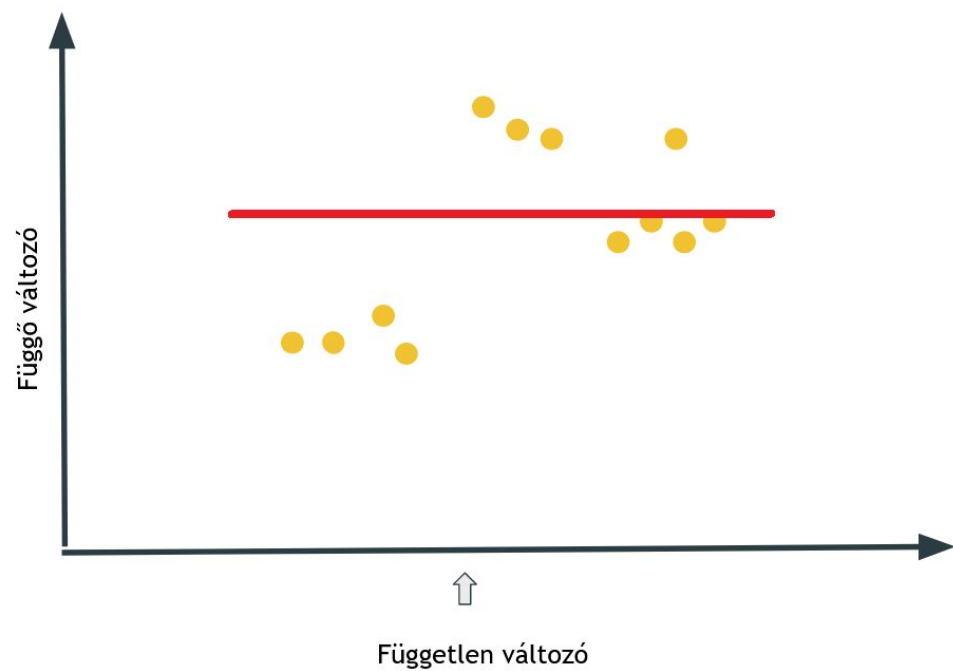


Machine learning

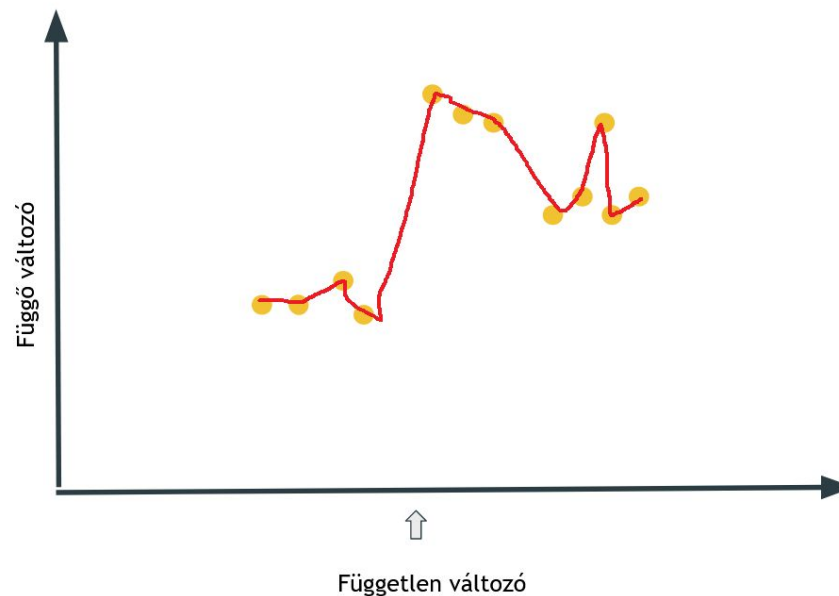


Machine learning - Bias and Variance

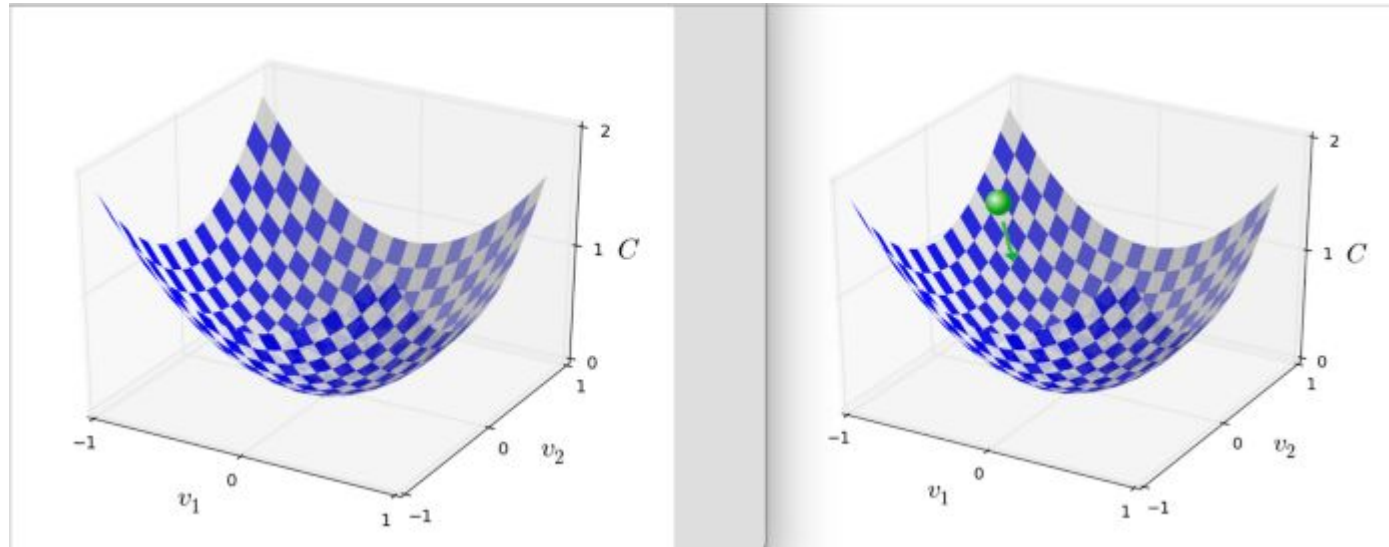
High bias



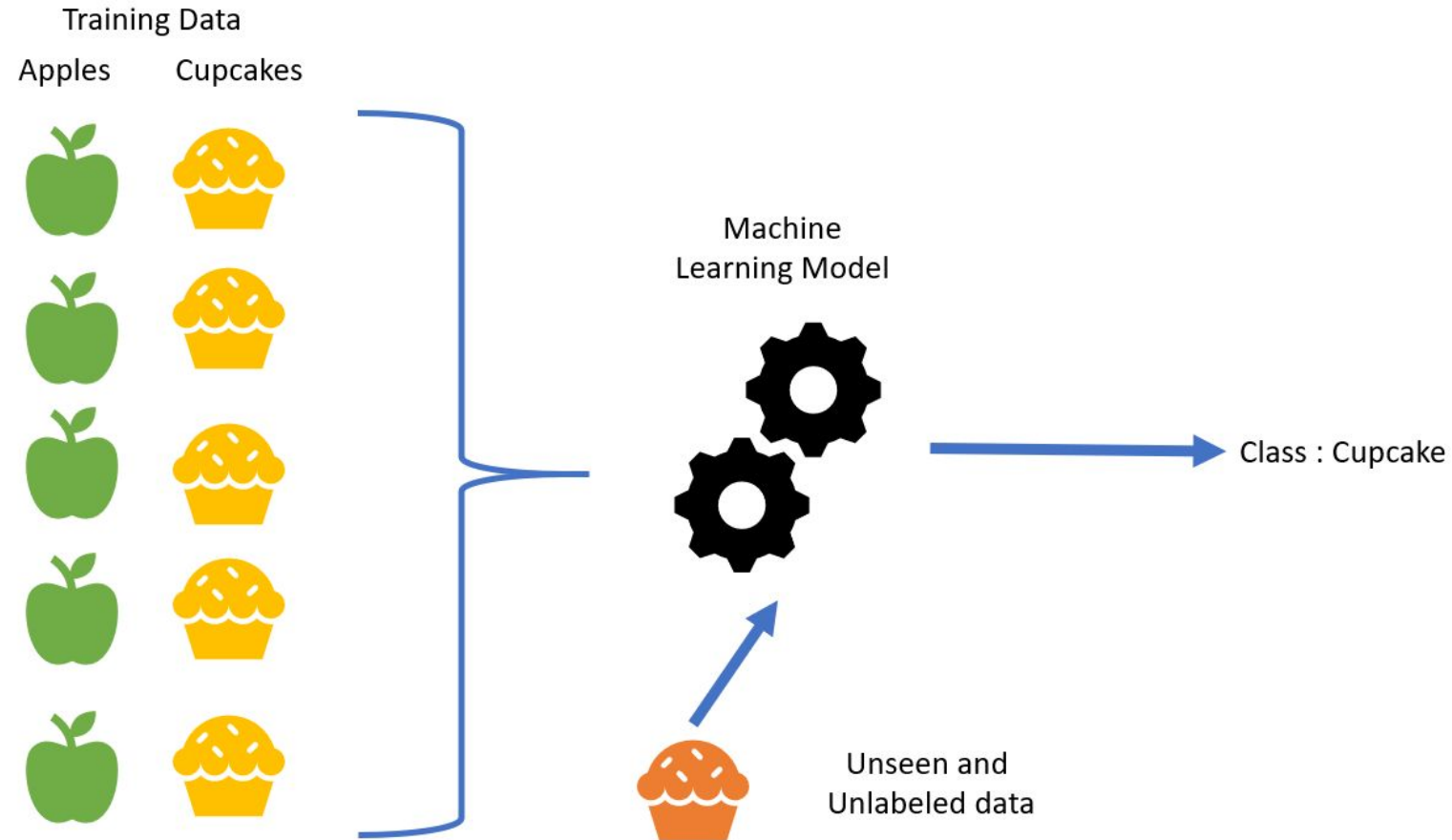
High Variance



Machine learning - Minimize error

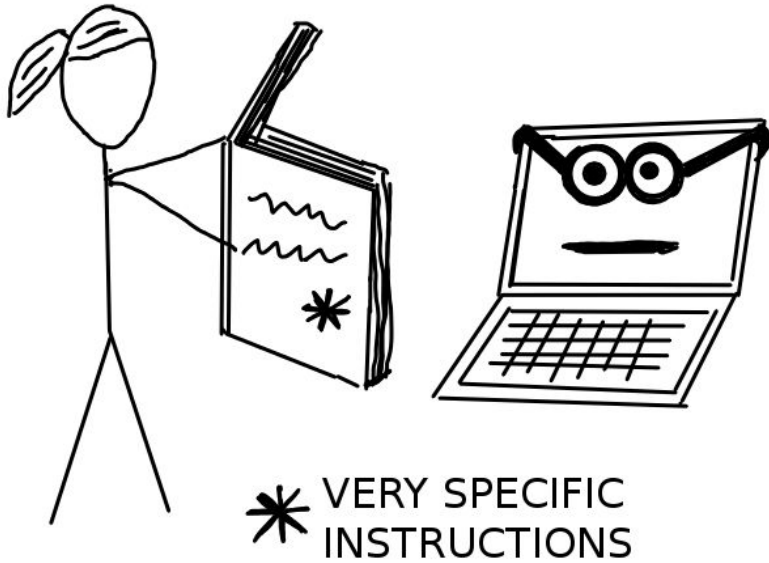


Machine learning

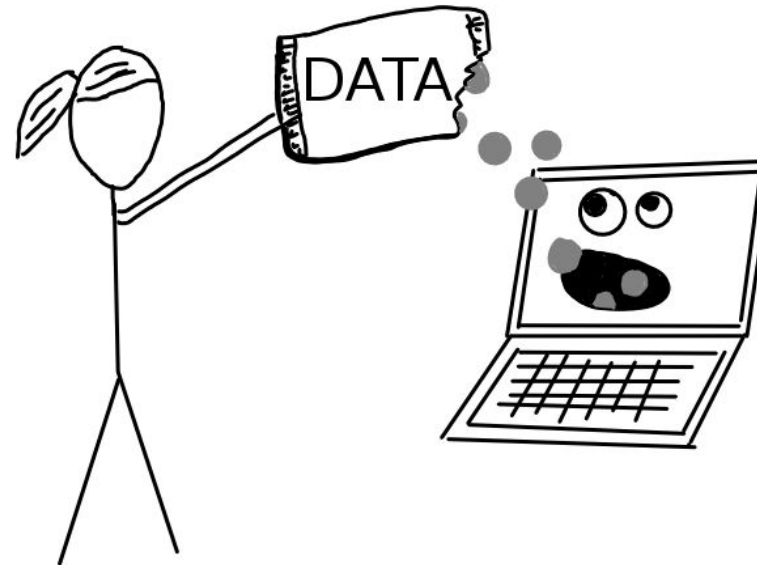


Machine learning

Without Machine Learning



With Machine Learning

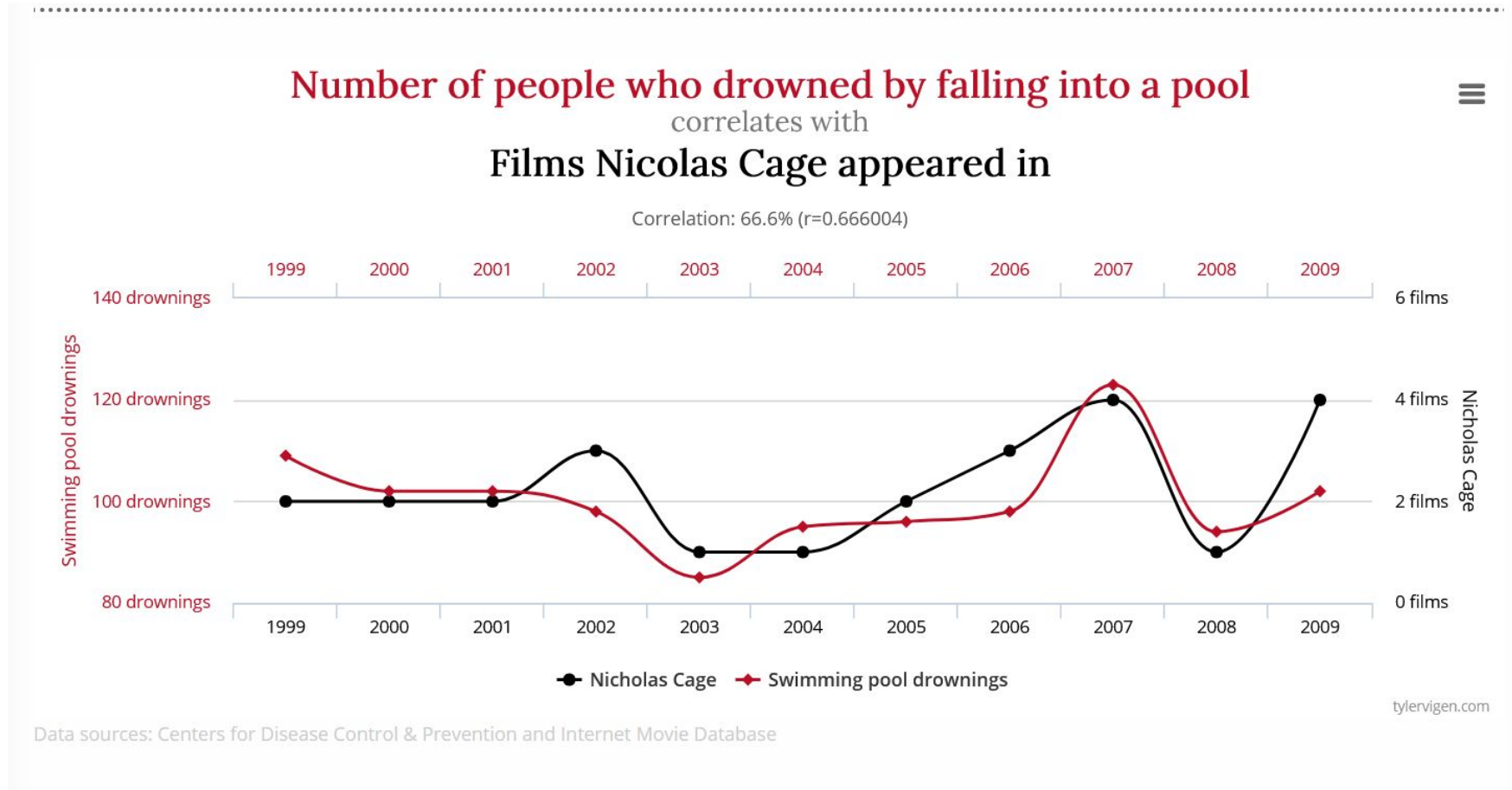


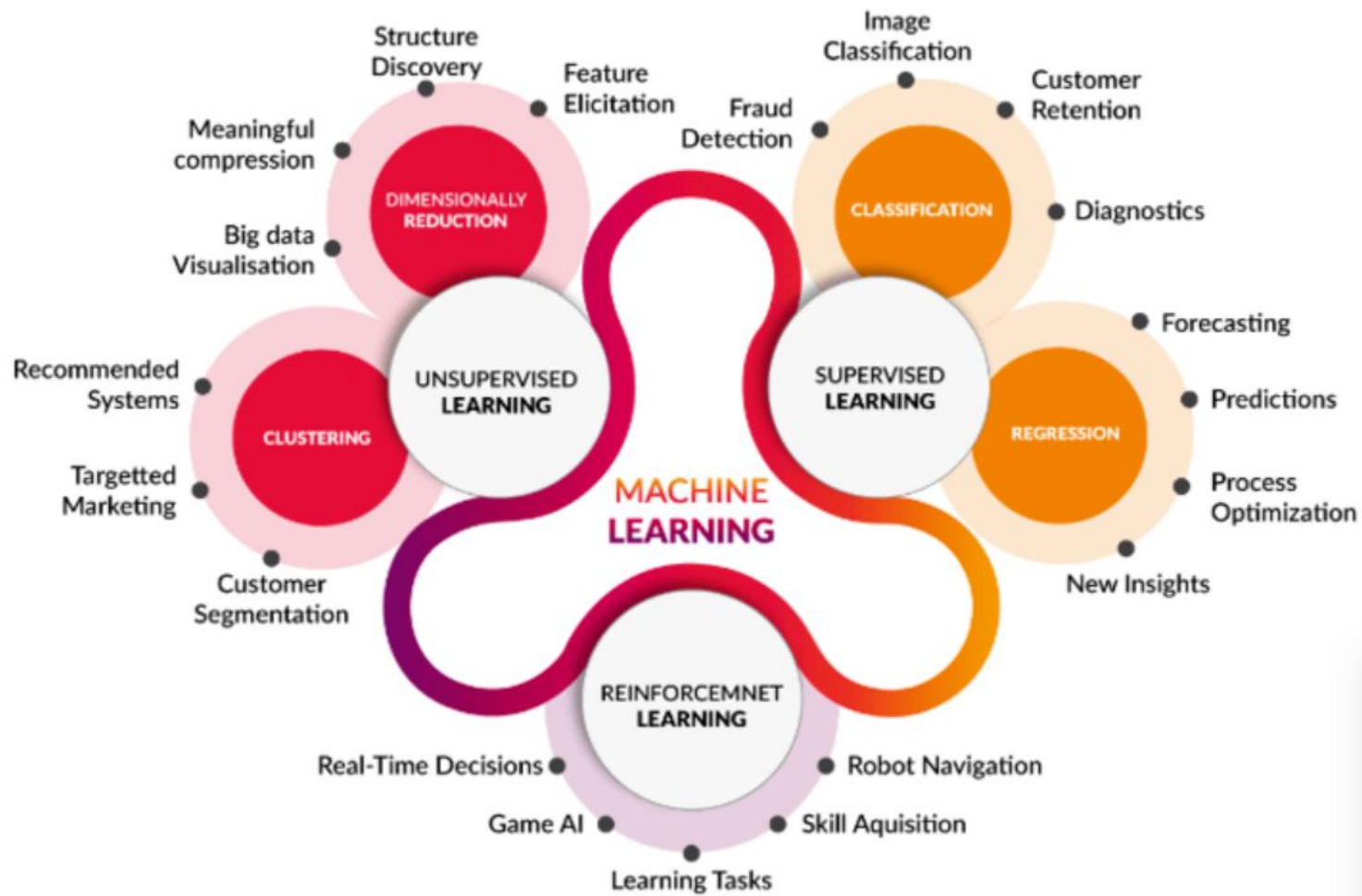
Data Preparation

- ▶ Outlier filtering
- ▶ Handling missing values
- ▶ Dependency - correlation
- ▶ Multiple correlation
- ▶ Transform data



Correlation is not causation

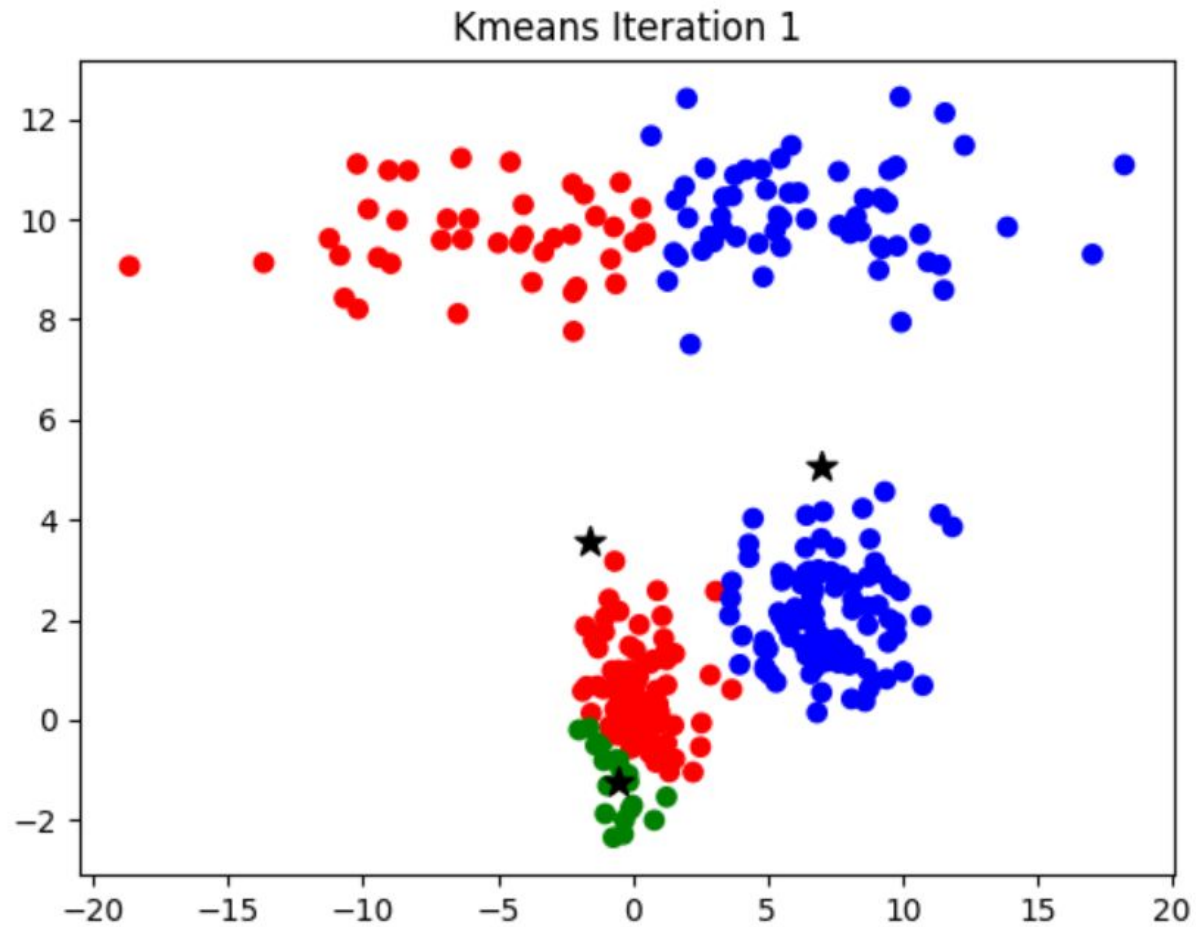




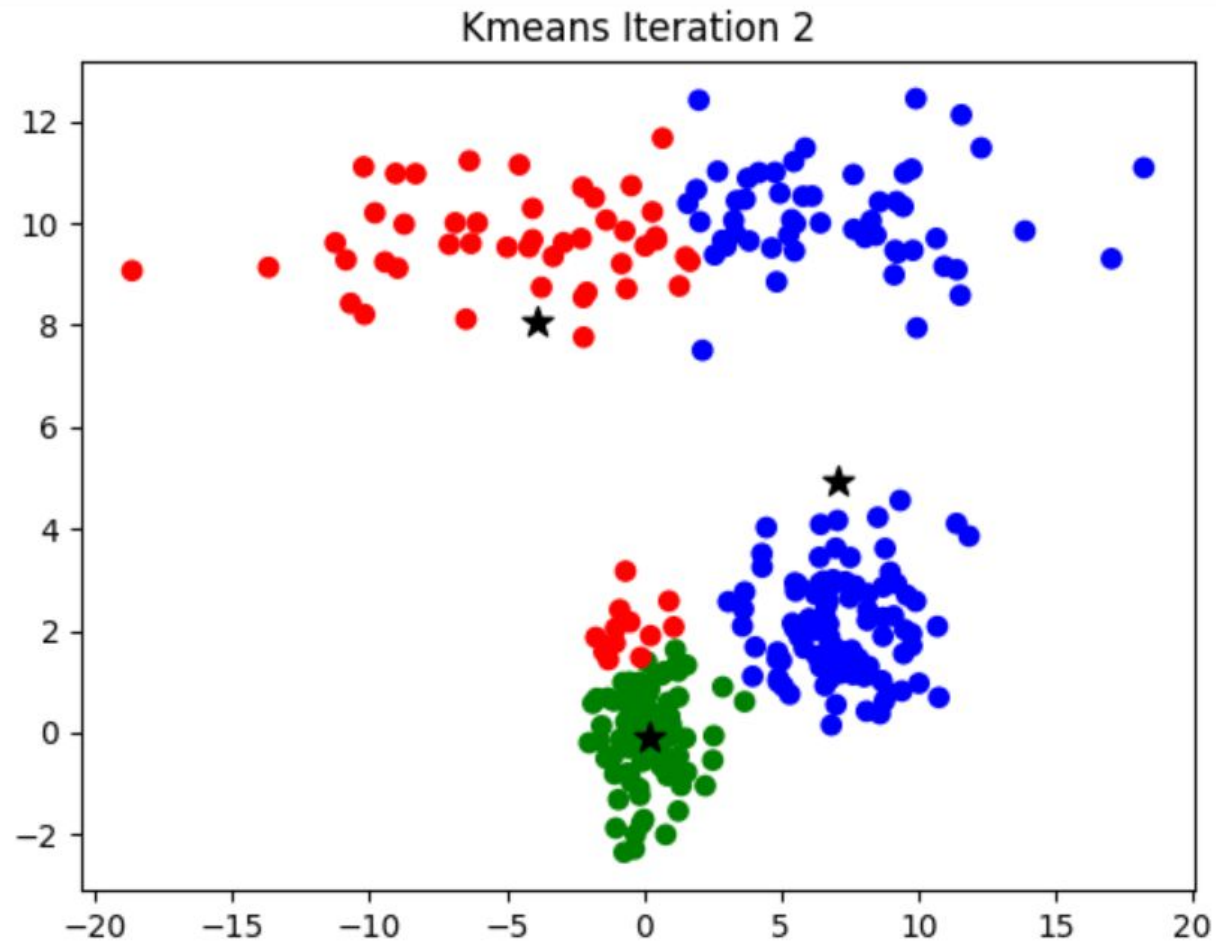
Reinforcement learning



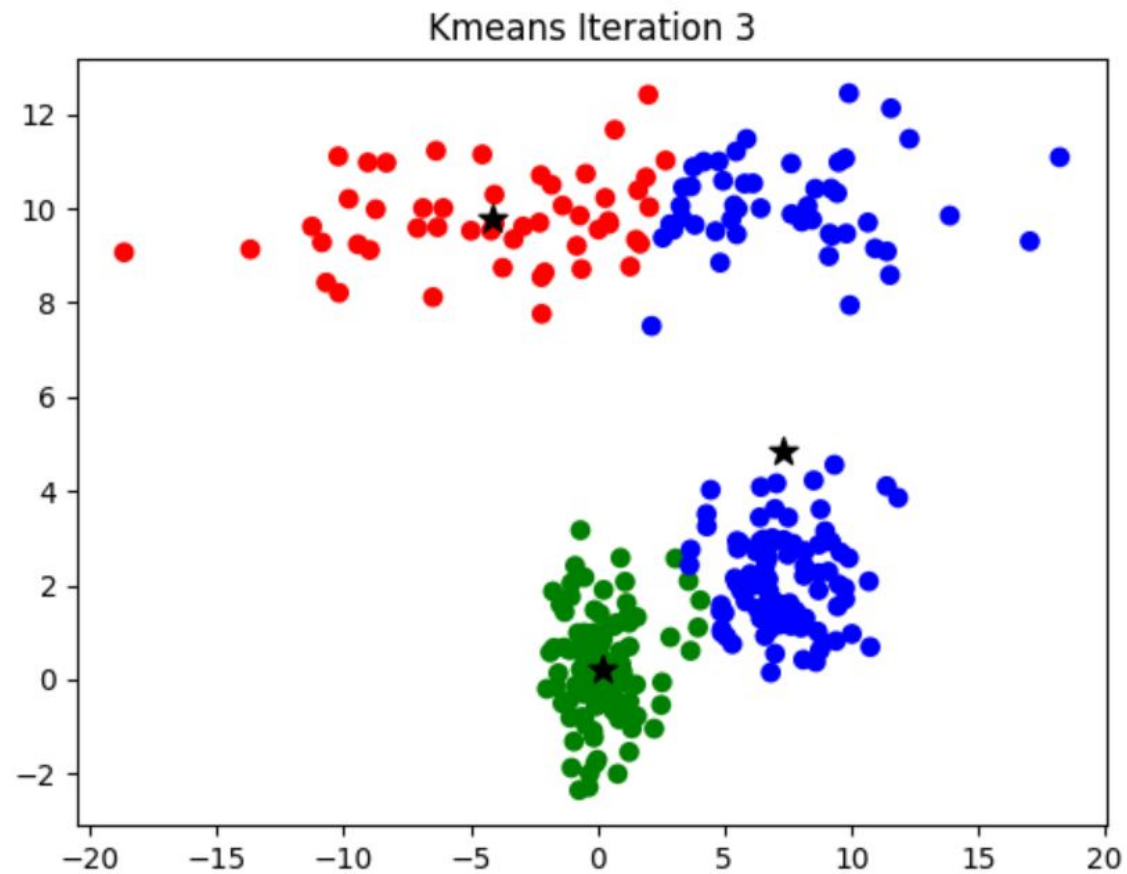
Clustering - Kmeans



Clustering Kmeans



Clustering Kmeans



Clustering Kmeans

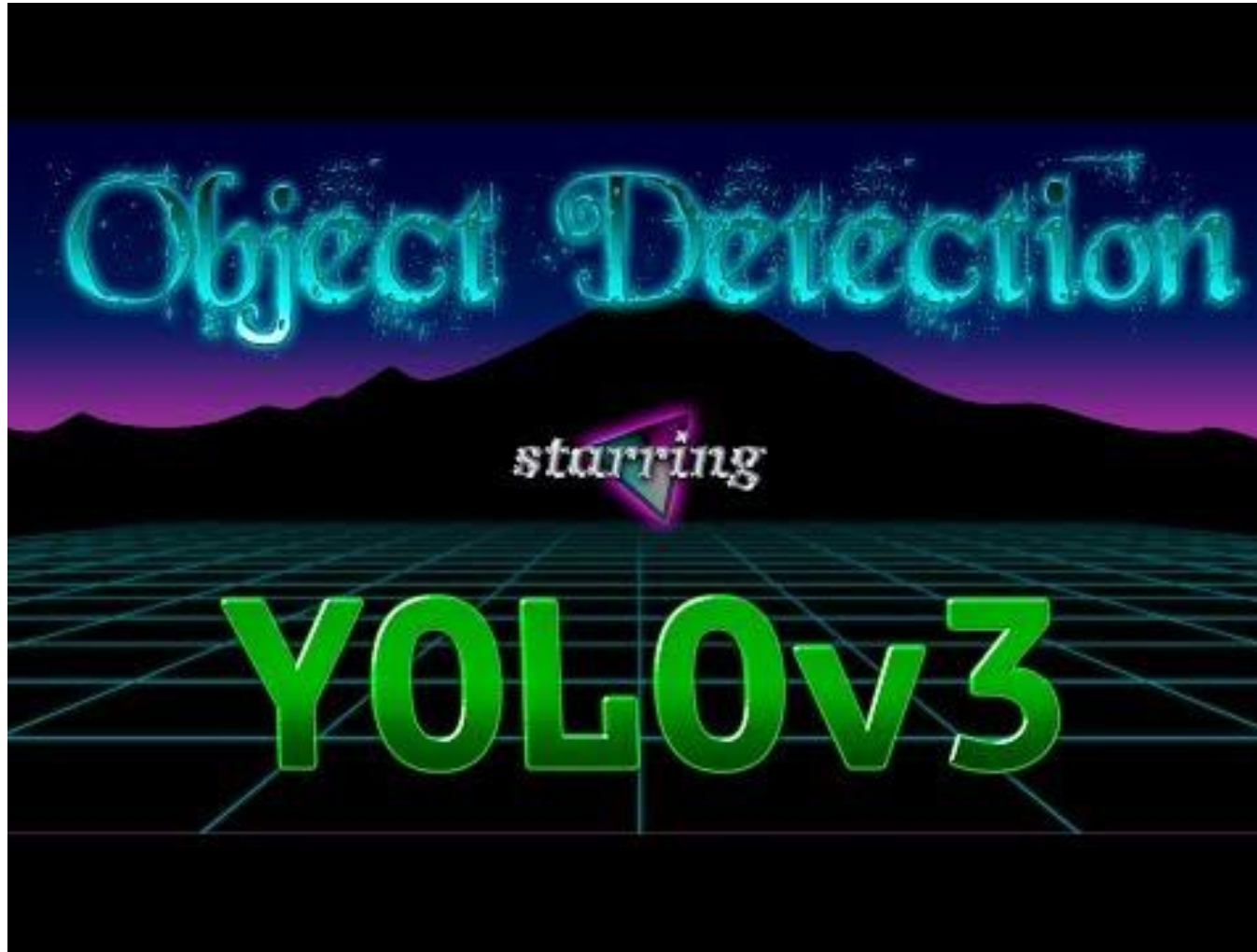
Hogyan definiáljuk a távolságot?

Hány klasztert válasszunk?

Mi a klaszterek jelentése?

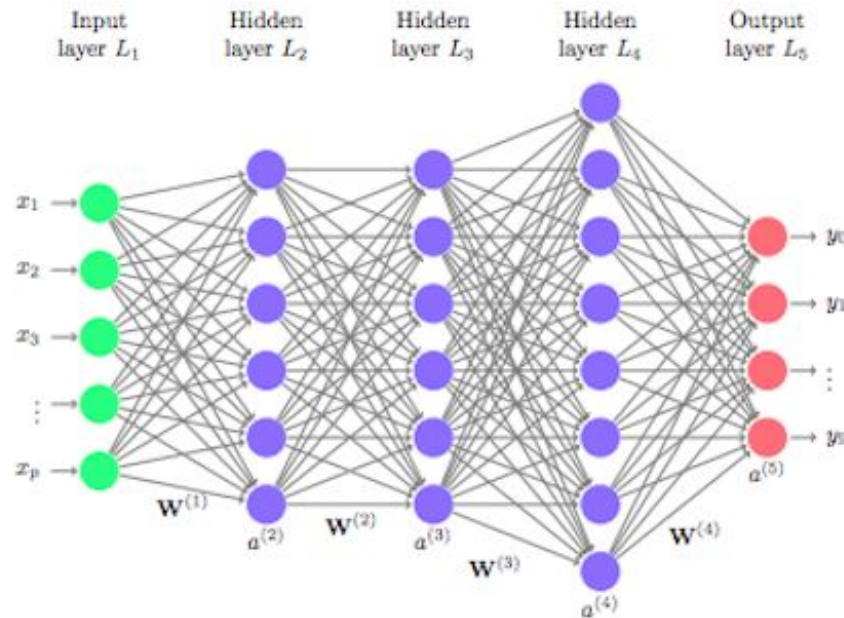


Classification



Deep learning

- ▶ Fancy név a többszintű neurális hálóra
- ▶ Layerek lehetnek különböző típusúak
- ▶ Teljesítmény / bemenő adatok összefüggés jobban skálázódik



Python

- ▶ Python is an open source script language
- ▶ High-level, general purpose
- ▶ Dynamic type system, automatic memory management
- ▶ Interpreted
- ▶ Fields of application
 - ▶ Numeric computation
 - ▶ Simulation
 - ▶ Data analysis
- ▶ It's popular because:
 - ▶ It's easy to learn and use
 - ▶ It's fast (like really fast)
 - ▶ It's resource-friendly
 - ▶ It's almost limitless (there's a Python library for almost everything)



További lehetőségek

- ▶ R
- ▶ SAS
- ▶ Matlab
- ▶ Húzókatós programok
- ▶ PL/SQL

Big data

Egyszerűen megfogalmazva

- „Olyan adatmennyiség, amitől az Excel már crash-el.”

Oxford dictionary

- „data sets that are too large and complex to manipulate or interrogate with standard methods or tools”

Dictionary.com

- „data sets, typically consisting of billions or trillions of records, that are so vast and complex that they require new and powerful computational resources to process

O'Reilly Media (M. Loukides)

- „As storage capacity continues to expand, today's “big” is certainly tomorrow's “medium” and next week's “small.” The most meaningful definition I've heard: “big data” is when the size of the data itself becomes part of the problem.”

Big data

Gartner definíció (D. Laney):

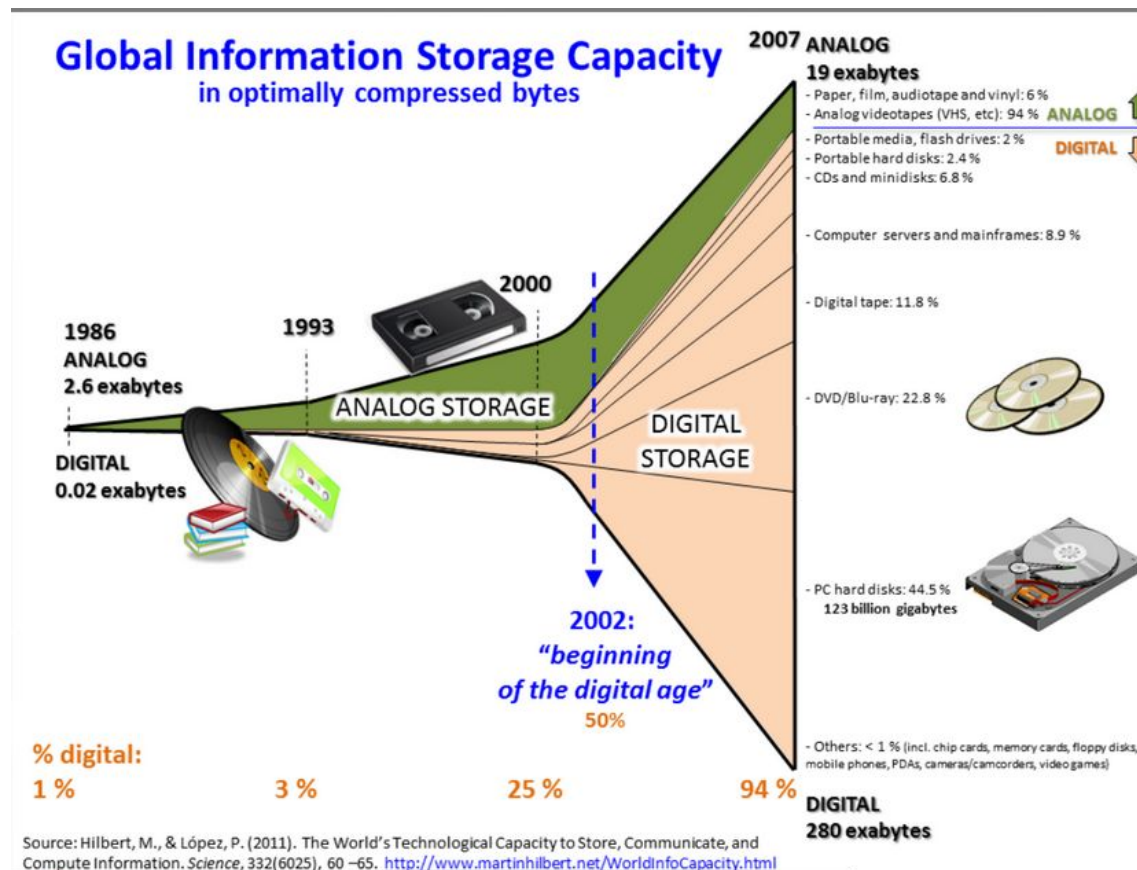
- ▶ „Big data are high **volume**, high **velocity**, and/or high **variety** information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.”

Fentit szokás 3-10V-nek is nevezni:

- ▶ Volume - méret
- ▶ Velocity - adatsebesség
- ▶ Variety - sokféle típusú adat
- ▶ ...
 - ▶ Veracity - igazságtartalom, tisztaság; Validity - helyesség; Variability - egyre rugalmasabb struktúrák; Value - nagy értékű; Visualization - vizualizálhatóság; ...

Big data - Volume

- The capacity to store information doubles roughly every 3.5 years
- Roughly 2,5 exabytes of data are generated every day



Big data - Variety

- ▶ Different data types are used: text, pictures, videos
- ▶ The capability to process important information is crucial
- ▶ Data come from different sources

Big data Velocity

- ▶ So much so that the MetLife executive stressed that: “Velocity can be more important than volume because it can give us a bigger competitive advantage. Sometimes it’s better to have limited data in real time than lots of data at a low speed.”
- ▶ Standard relational databases are not capable of handling this amount of data
- ▶ Parallel computation
- ▶ Exmaples?



Big data - Veracity

- Data veracity, in general, is how accurate or truthful a data set may be
- The actual data makes sense based on business needs
- Data source trustworthy
- The only measure that decreases over time



Map reduce



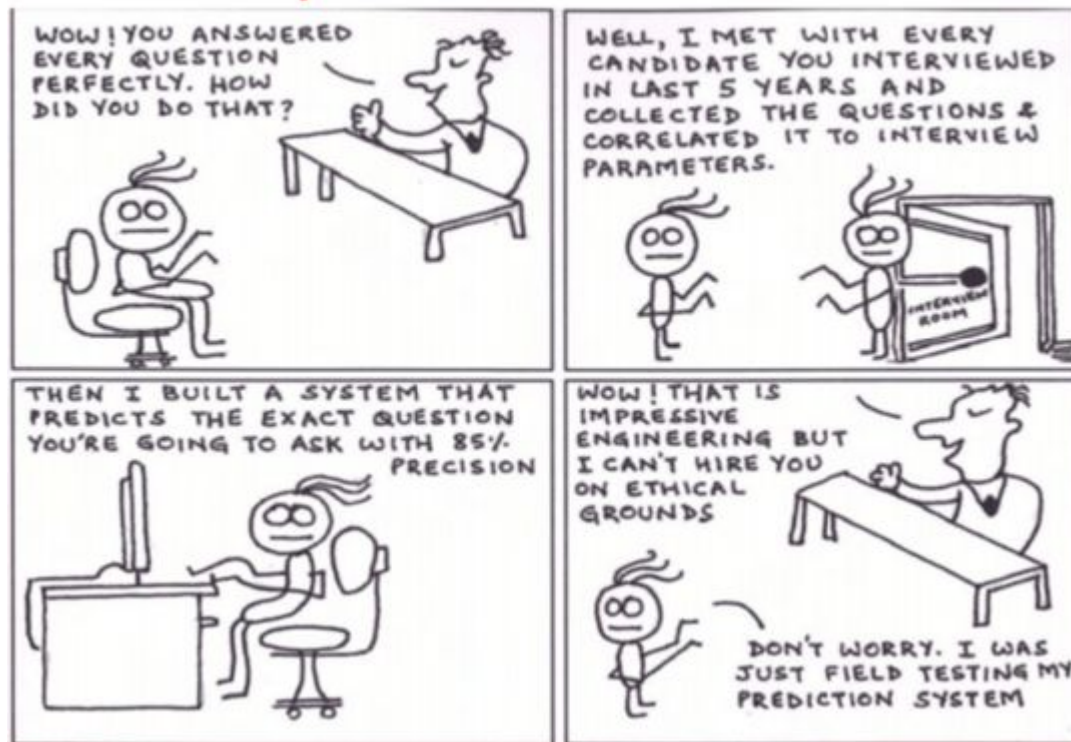
Data Science akcióban





PEOPLE COME FIRST
INFORMATIKAI SZAKÉRTŐK EGYESÜLETE

When you interview a data scientist...



Köszönöm a figyelmet!