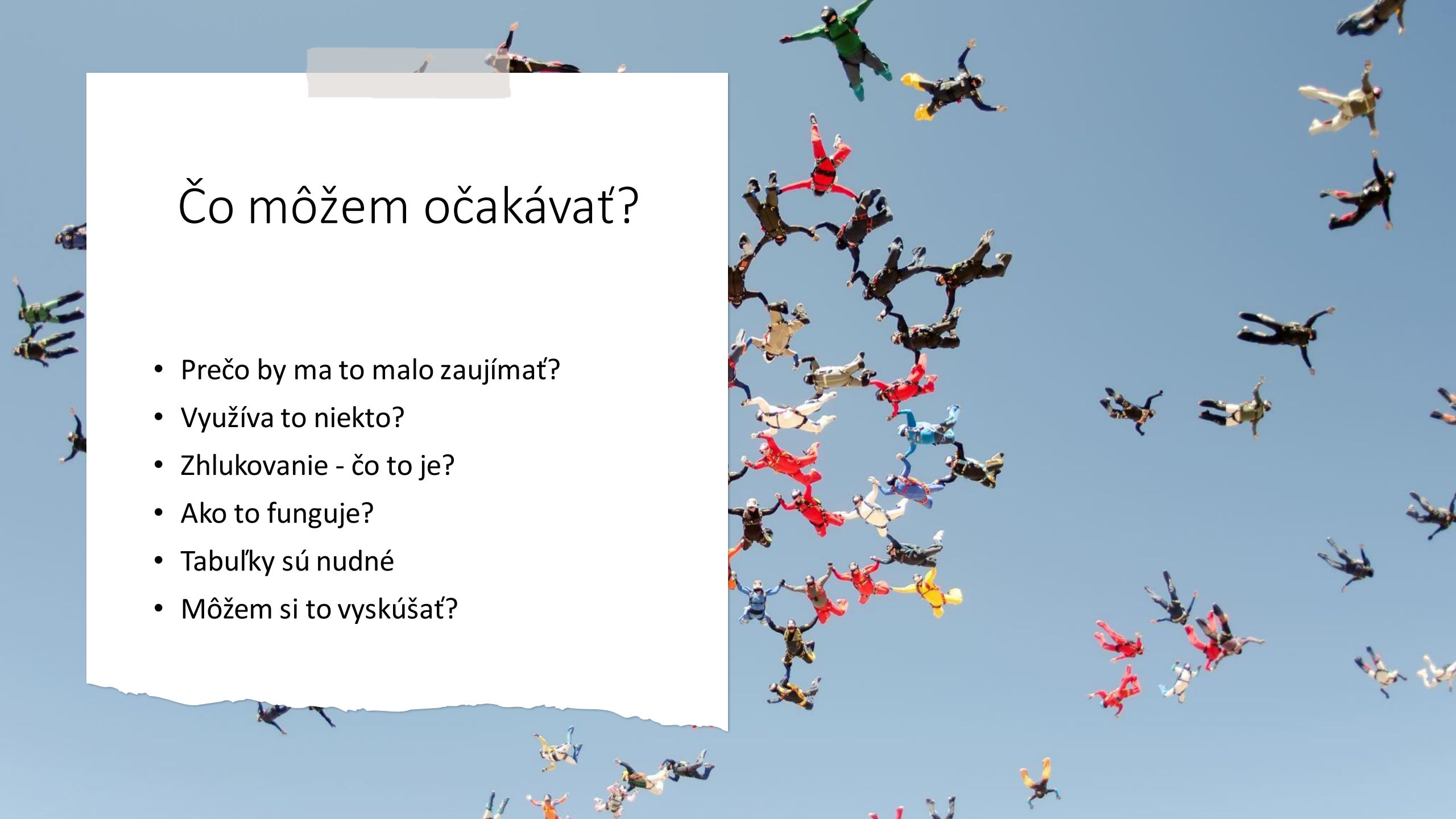
Učenie bez učiteľa

zhlukovanie
a
vizualizácia dát

Operačný program
Efektívna verejná správa

Európska únia
Európsky sociálny fond

MINISTERSTVO
VNÚTRA
SLOVENSKEJ REPUBLIKY



Čo môžem očakávať?

- Prečo by ma to malo zaujímať?
- Využíva to niekto?
- Zhlukovanie - čo to je?
- Ako to funguje?
- Tabuľky sú nudné
- Môžem si to vyskúšať?

Prečo by ma to malo zaujímať?
Učenie bez učiteľa

- Veľké množstvo dát bez jasnej štruktúry
- Minimum ľudského dohľadu
- Neexistujúce označenia

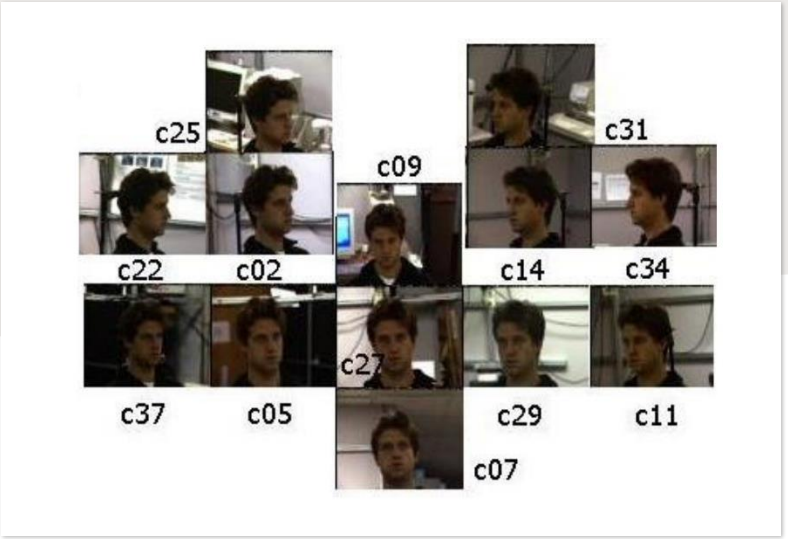
Prečo by ma to malo zaujímať?

- Nájdenie neznámych štruktúr
- Žiadne predsudky
- Zvýraznenie skrytých štruktúr
- Škálovanie na veľké dátové štruktúry
- Aplikovateľné na variabilné dátové štruktúry

Využíva to niekto?

- Detekcia podvodov
  - Poistovne
- Spam filter
- Identifikácia falošných správ

Zhlukovanie v oblasti spracovania obrazu



- Vstup do systému 30-120 fps
- Nájdenie reprezentatívnej tváre

Zhlukovanie - čo to je?

How many clusters?

How many clusters?

Six Clusters

Two Clusters

Four Clusters

Typy zhlukovania

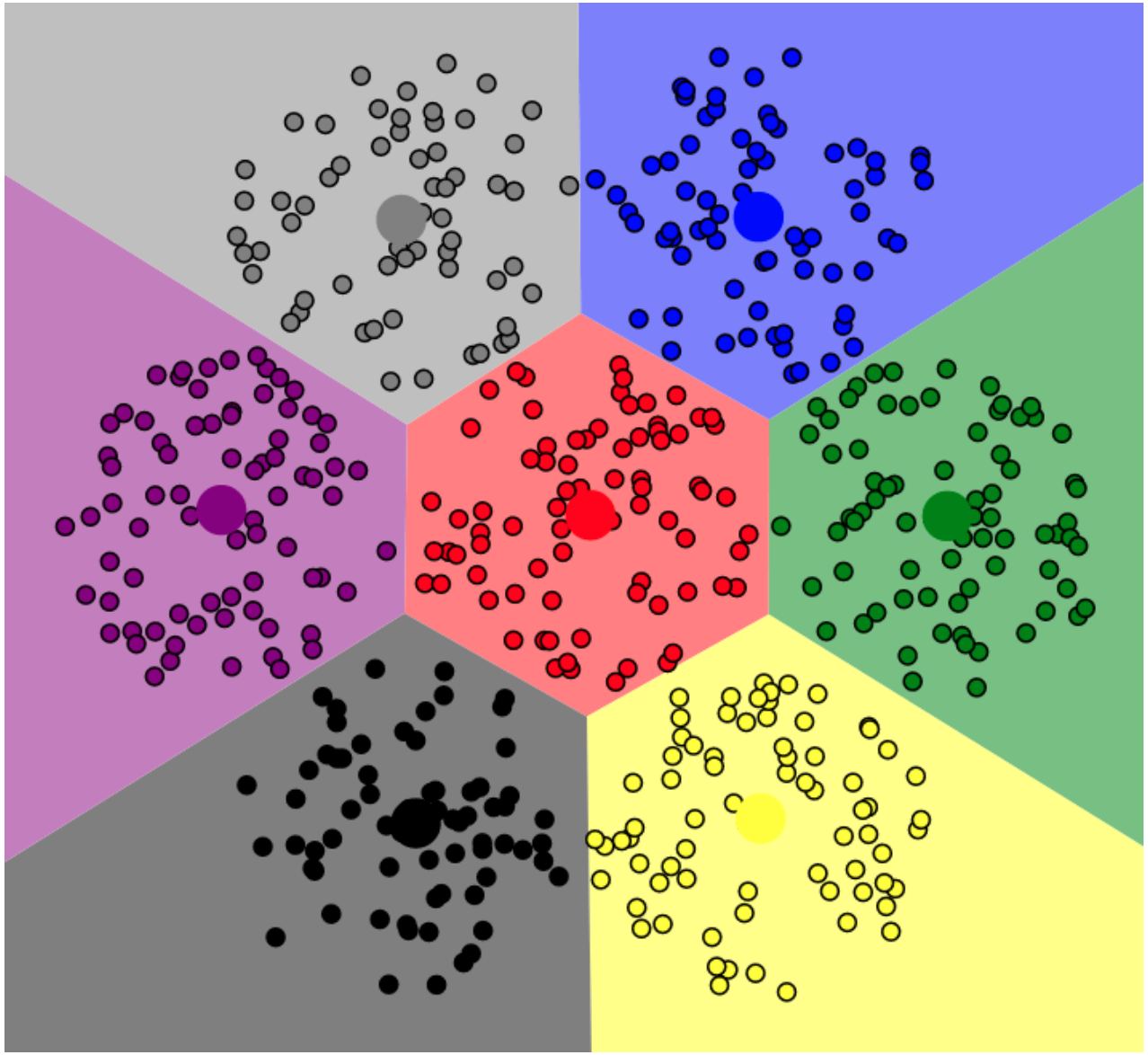
- Modely založené na centroidoch
  - **K-means**
  - Známy počet zhlukov
- Modely založené na rozdenení
  - Modely založené na štatistike
  - Pretrénovanie
- Modely založené na hustote
  - **DBSCAN**
- Neurónové siete
  - SOM (Self organizing map)

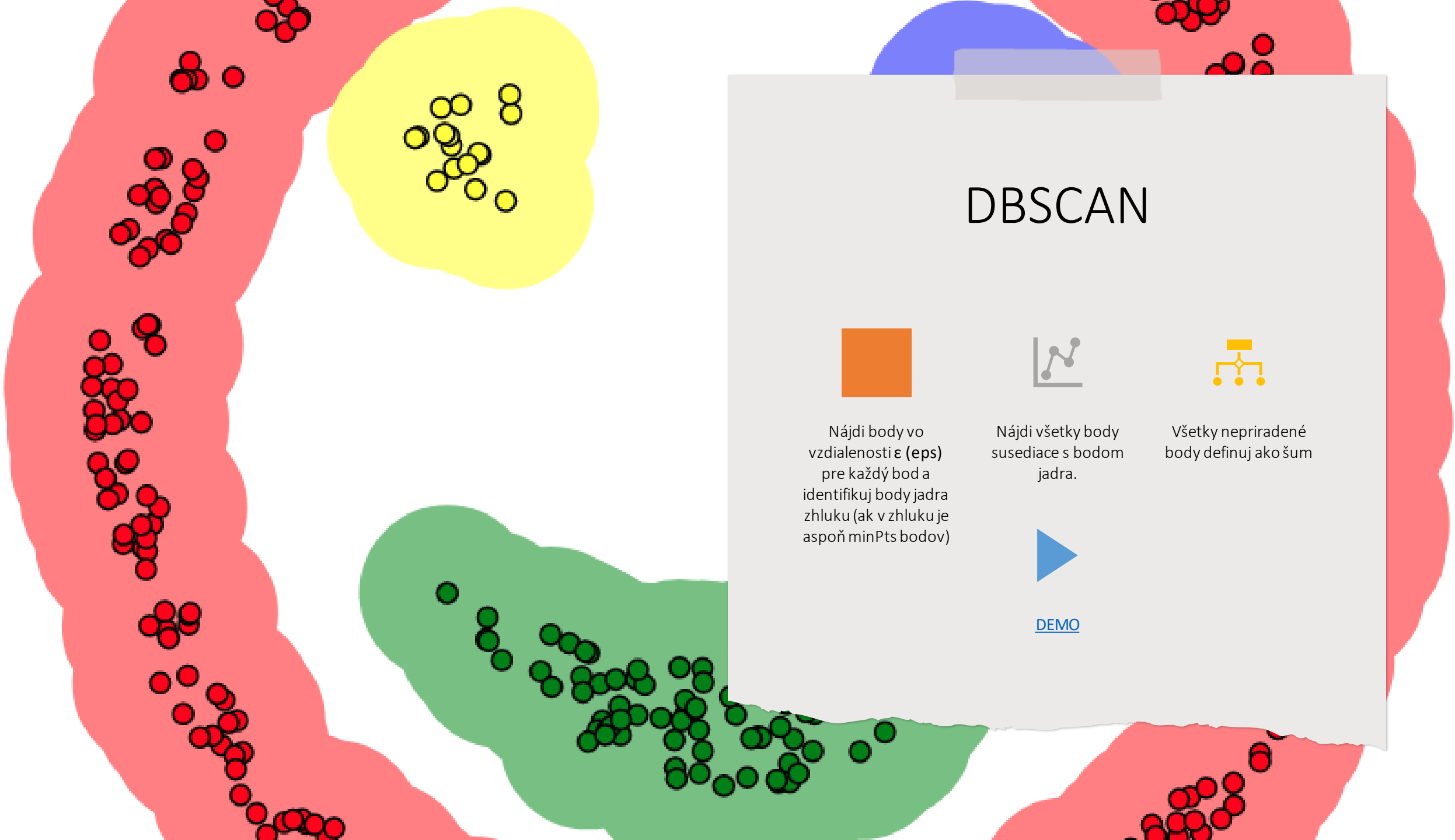
Ako to funguje?

- K-means
- DBSCAN
- SOM

K-means

- Špecifikovaný počet zhlukov *K*.
- Priebeh:
  1. Vyber K centier (rôzne spôsoby inicializácie)
  2. Priraď body ku centrám na základne ich vzájomnej vzdialenosti
  3. Získaj nové centrá
  4. Opakuj od bodu 2. kým priradenie bodov k centru nie sú rovnaké
- DEMO





DBSCAN

Nájdi body vo vzdialenosti **ε (eps)** pre každý bod a identifikuj body jadra zhluku (ak v zhluku je aspoň minPts bodov)

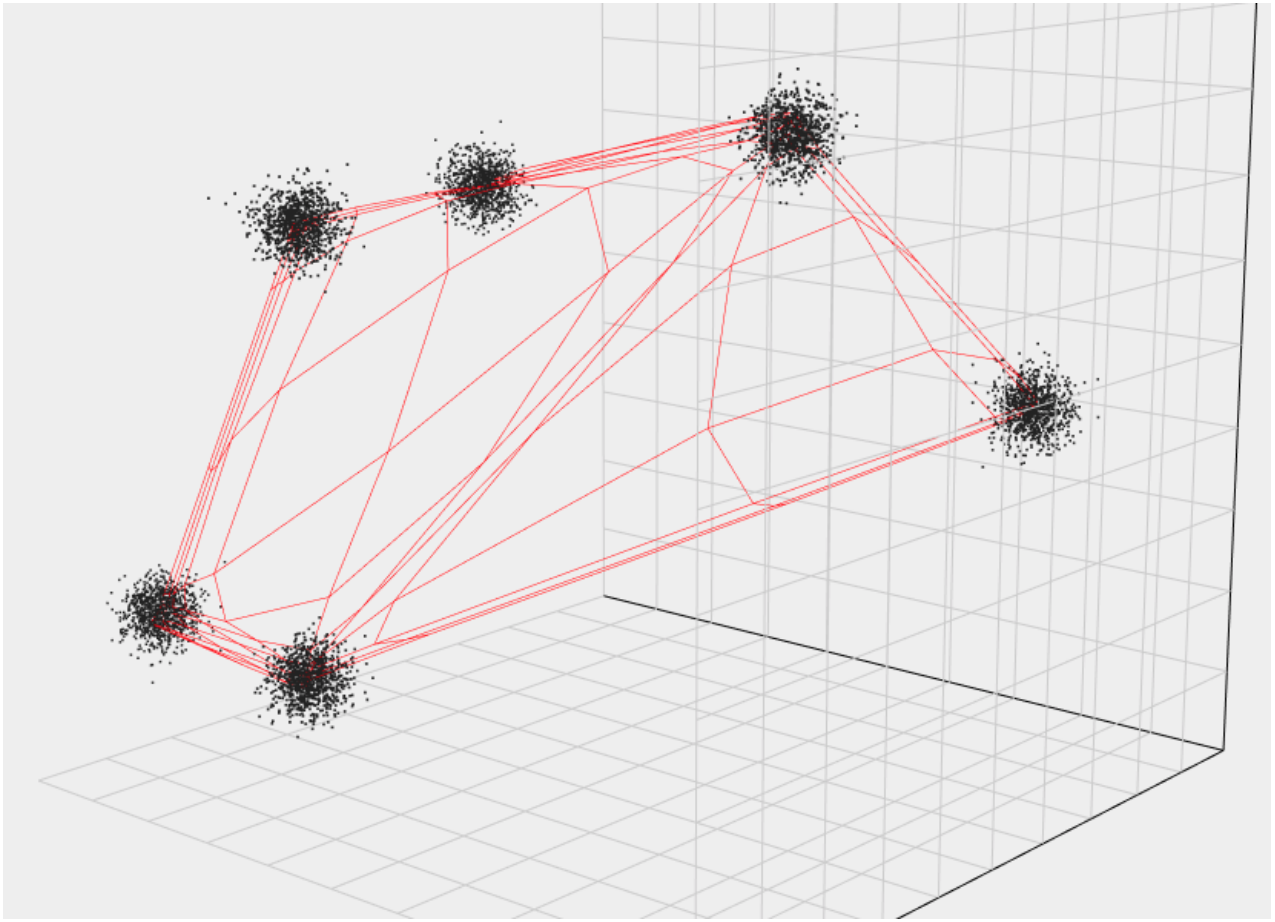
Nájdi všetky body susediace s bodom jadra.

Všetky nepriradené body definuj ako šum

DEMO

Self-Organizing Map

1. Inicializuj všetky uzly "mapy". (Inicializuj váhy)
2. Náhodne vyber bod z dátovej množiny
3. Nájdi najbližší uzol mapy ku vybranému bodu (**Best Matching Unit** (BMU))
4. Uprav pozície BMU a jeho priameho okolia
5. Opakuj od 2. Kroku po dobu N iterácií



- DEMO

ata from csv:

untry	Region	Happiness.Ran
ghanistan	Southern Asia	14
bania	Central and Eastern Europe	10
geria	Middle East and Northern Africa	5
gola	Sub-Saharan Africa	14
gentina	Latin America and Caribbean	2
menia	Central and Eastern Europe	12
stralia	Australia and New Zealand	1
stria	Western Europe	1
erbaijan	Central and Eastern Europe	8
hrain	Middle East and Northern Africa	4
ngladesh	Southern Asia	11
larus	Central and Eastern Europe	6
lgium	Western Europe	1
lize	Latin America and Caribbean	5
nin	Sub-Saharan Africa	14
utan	Southern Asia	9
livia	Latin America and Caribbean	5

Tabuľky sú nudné

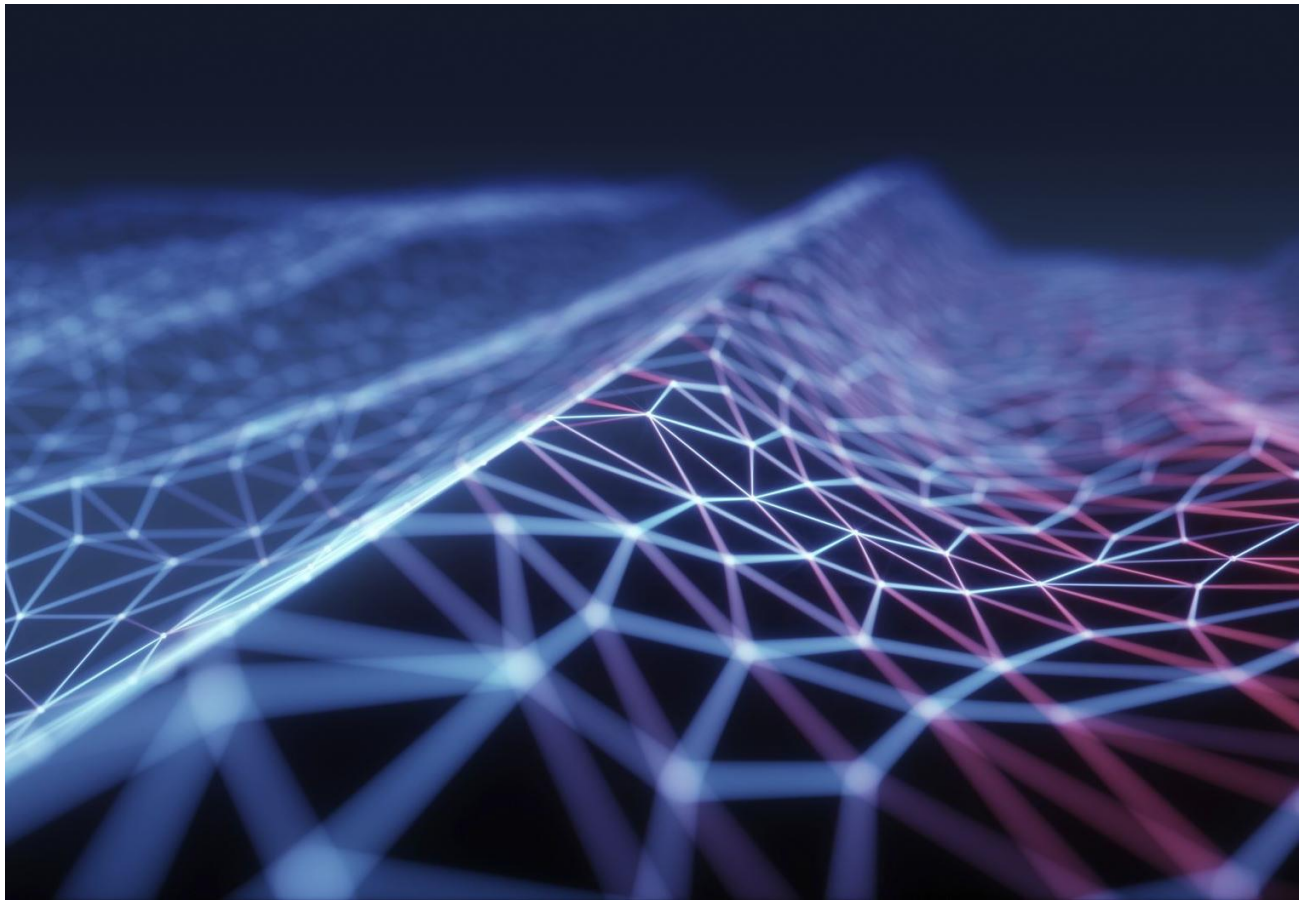
Tabuľky sú nudné, ale...

Tabuľky

- Uchovávajú mnohorozmerné dáta (veľa stĺpcov)
- Dobrý začiatočný bod pre ďalšie spracovanie
- Zložité na čítanie a pochopenie
- Takmer nemožné nájsť skryté vzťahy

Grafy

- Intuitívne chápanie
- Možnosť nájsť skryté vzťahy
- Mnoho dôležitých dát môžeme zanedbať
- Silno závyslý od vizualizačných techník



Vizualizačné techniky

ZNIŽUJÚ DIMENZIU DÁT

ZACHOVÁVAJÚ VZŤAHY MEDZI DÁTAMI

POMÁHAJÚ NÁJSŤ NOVÉ VZŤAHY

RÝCHLE SPRACOVANIE

Rôzne prístupy

- PCA (Principal component analysis)
  - Deterministický
  - Zachováva existujúce vzťahy (môžná ďalšia analýza)
  - Nerieši nekompletné dáta (ale existuje mnoho modifikácií)
  - Ponúka optimálne riešenie zo štatictického pohľadu
- T-sne
  - Iteratívny
  - Hľadá vzťahy
  - Nerieši nekompletné dáta
  - Nie vždy nájde optimálne riešenie (Vadí to?)

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Algoritmus najprv v pôvodnom priestore spočíta podobnosť všetkých párov v datasete.
- Následne sa v nižšej dimenzii snaží rozmiestniť obrazy pôvodných vzoriek tak, aby rozmiestnenie bolo čo najpodobnejšie pôvodnému.
- DEMO

PCA – Principal component analysis

- Hľadajme takú transformáciu ktorá bude optimálna z hľadiska minimalizácie strednej kvadratickej chyby rekonštruovaného a originálneho vektora dát.
- DEMO

Let's be a programmer

Colab link