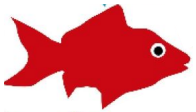
# Anomáliakeresés Pythonban

Gáspár Csaba
senior data scientist
dmlab
gaspar.csaba@dmlab.hu /
www.linkedin.com/in/gasparcsabahu/
+36208234154

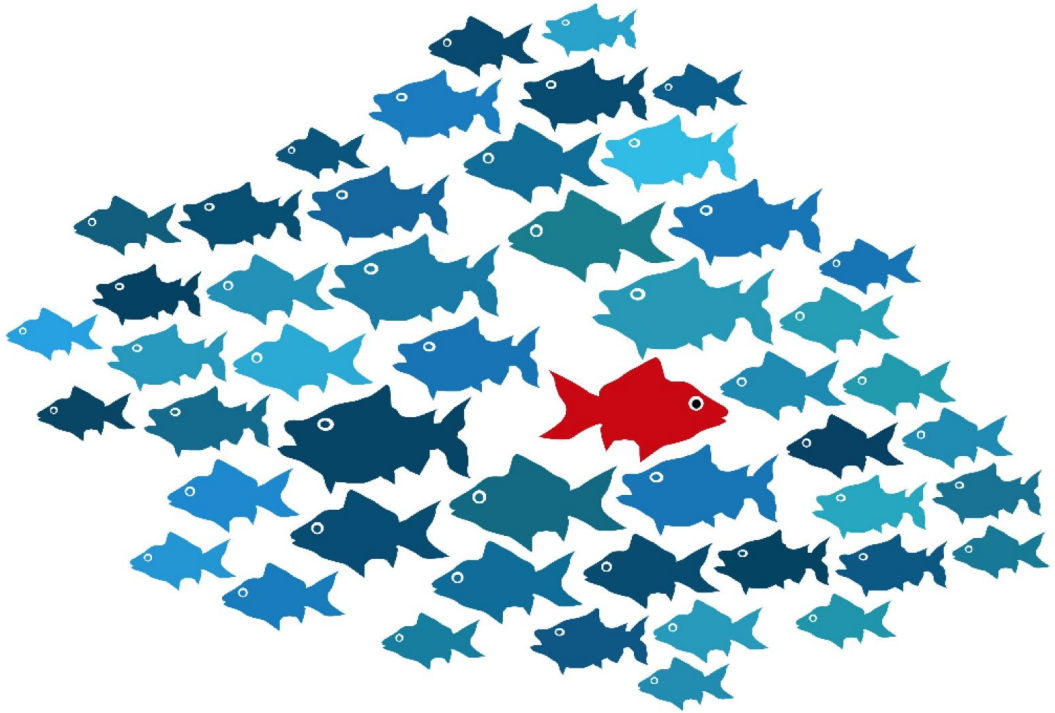
# Mi az anomália?

- Mintázat, ami nem illeszkedik az elvárt viselkedéshez.
- Szabályok helyett sok-sok viselkedési adat.
- Kapcsolat az adatokban.

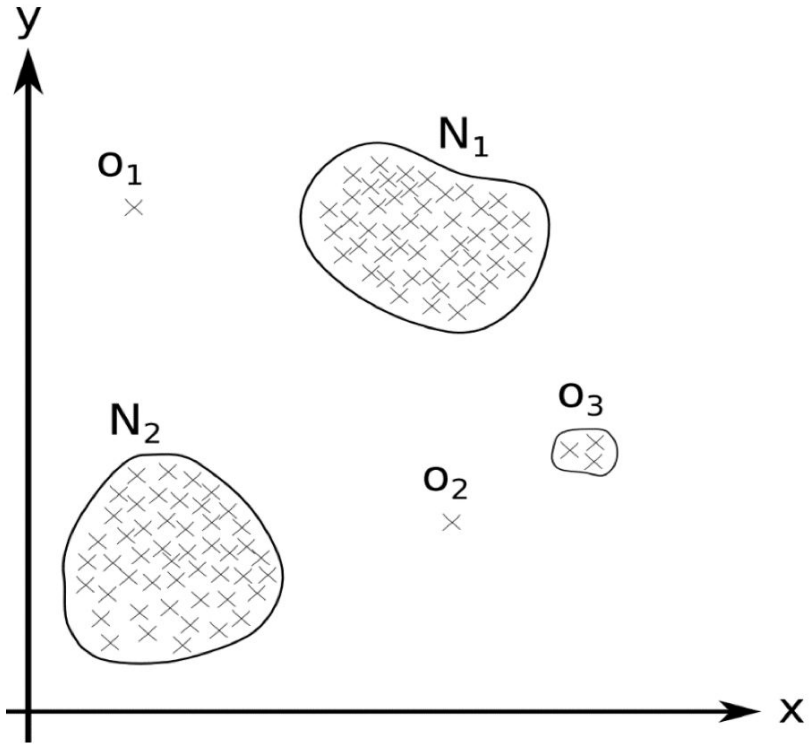


# Mi az anomália?

- Mintázat, ami nem illeszkedik az elvárt viselkedéshez.
- Szabályok helyett sok-sok viselkedési adat.
- Kapcsolat az adatokban.

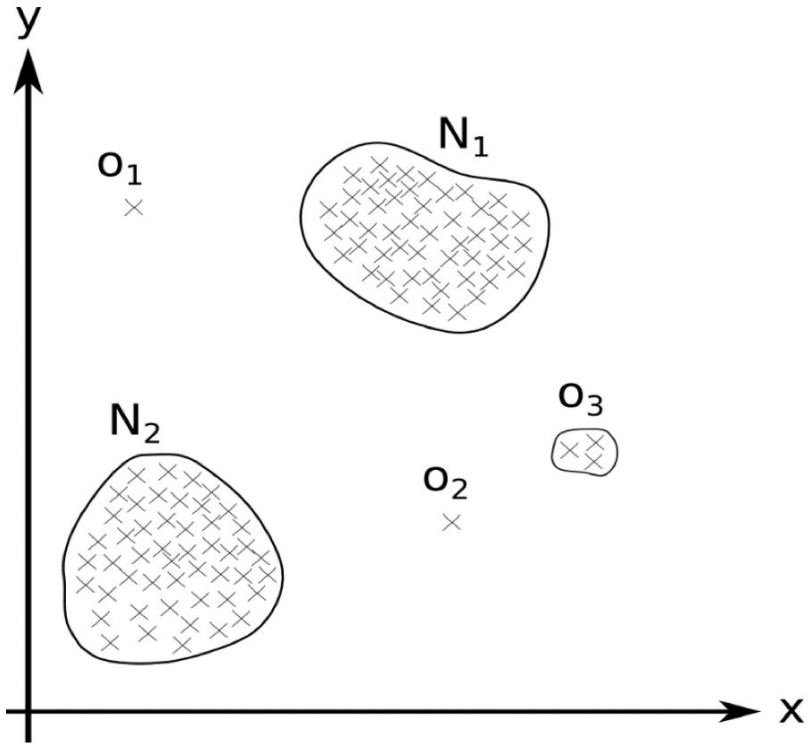


# Anomália szemléltetve




- Nominális pontok tartománya
- Outlier-ek fogalma
- Kevés dimenzió esetén könnyen észrevehetőek

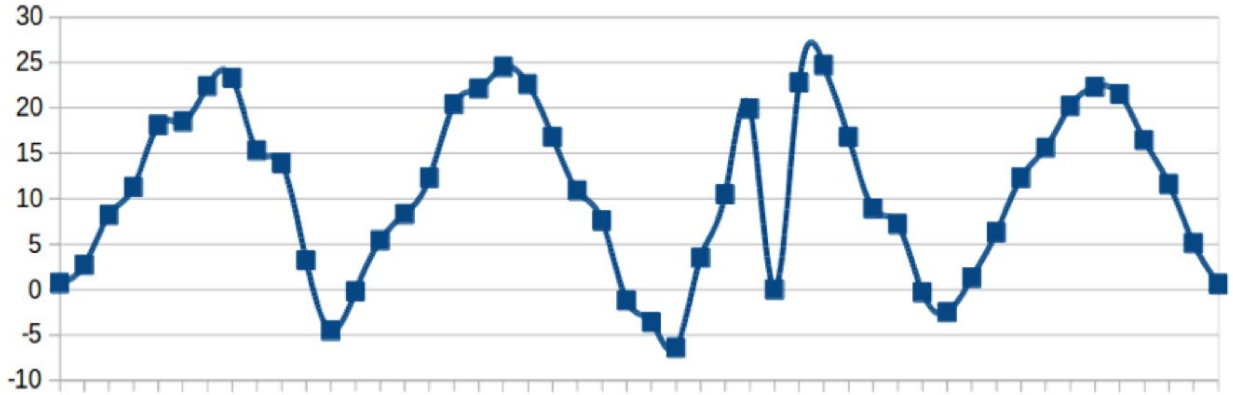
# Anomália fajtái




- Pontszerű anomália
- Nem csak egy pontból állhat, ez kihívást jelenthet bizonyos esetekben

# Anomália fajtái

- Kontextuális anomália
  - Környezetéhez képest kiugró
  - Kontextus: Idő, helyszín, kapcsolati háló




- Együttes anomália

# Anomália detekció Use-Case-ek

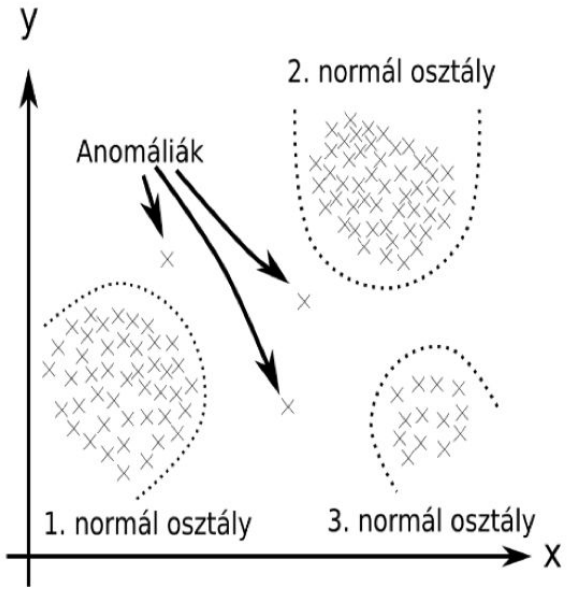
- Online csalás detektálás (vagy adócsalás)
  - Különböző adatokat gyűjtünk a felhasználóinkról, a furcsa viselkedéseket akarjuk kiszűrni
  - Story: Seon - (https://forbes.hu/uzlet/magyar-startup-nagy-uzlet-befektetes-seon/)
- Gyártás
  - Repülőgép hajtóművek: mérjük a vibrációt, hőgenerálódást és meg akarjuk mondani, hogy adott hajtómű rendben van-e.
- Adatközpontok (vagy távközlési hálózatok)
  - Mérjük az egyes számítógépeknek a memória felhasználását, a CPU leterheltségét, a hálózati forgalmat. Ezek alapján melyik gép fog lerobbanni, hol kell karbantartást végezni?

# Algoritmikus megközelítések

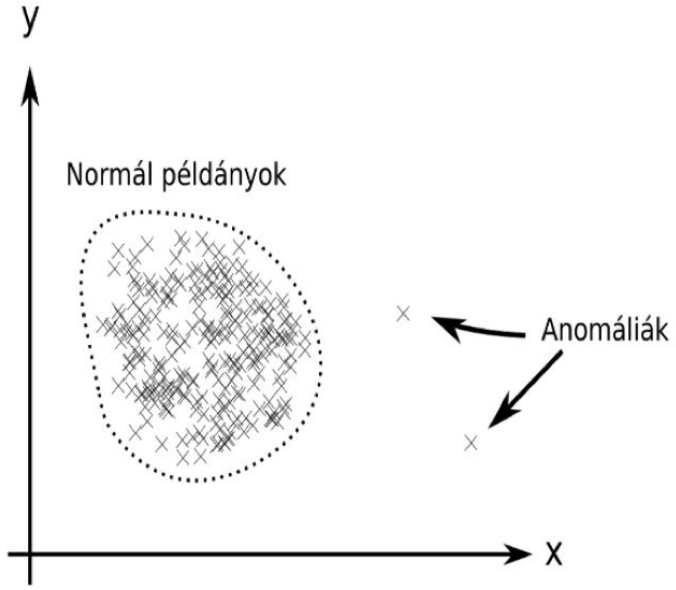
- Supervised learning: a tanító adathalmazban az egyes adatok fel vannak címkézve: Osztályozás
  - Probléma: sok nominális pont, kevés anomália, nehéz jó modellt építeni. Mi van akkor ha olyan anomália keletkezik, amiről nincsen adat?

- Unsupervised learning: Nincsenek címkék, a tanuló adat “tiszta”, teszt adatban vannak anomáliák. Vagy a tanító adatban is vannak anomáliák, amiket meg akarunk találni.
  - Probléma: Outlier anomaly vs. Inlier anomaly

# Osztályozás

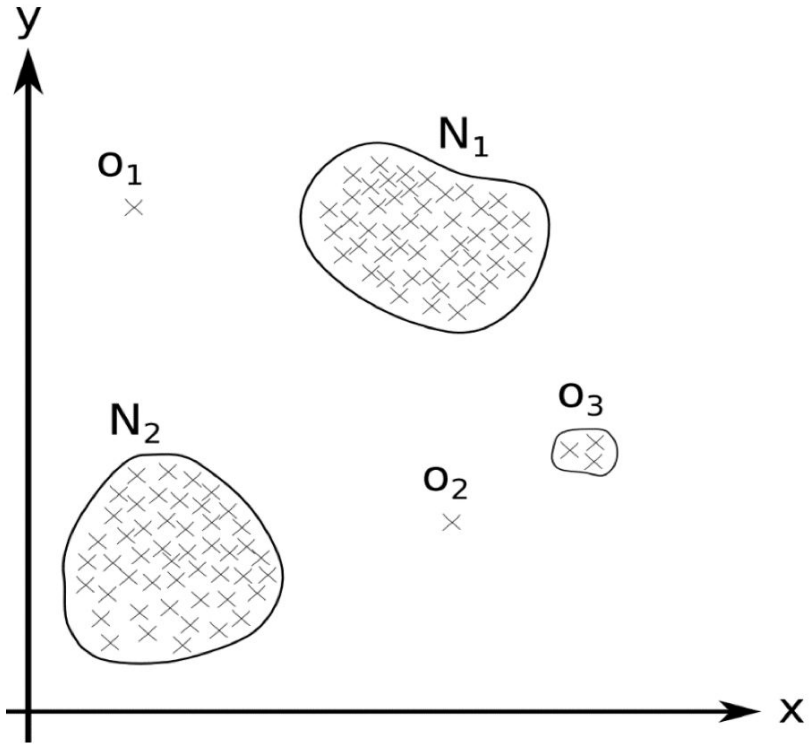



(a) Több osztályú anomália detekció



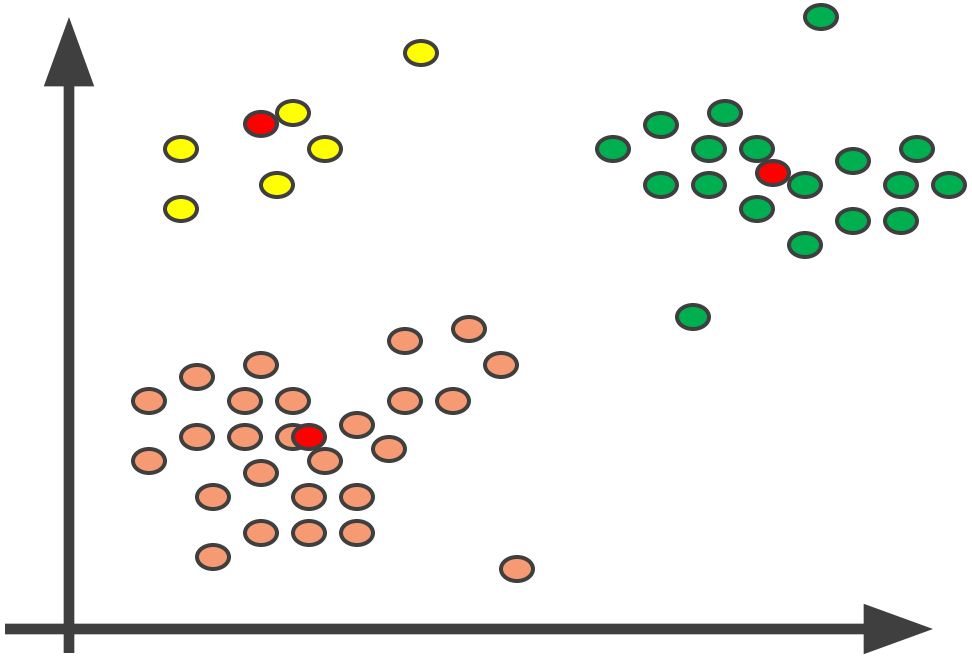

(b) Egy osztályú anomália detekció

# K. legközelebbi szomszéd módszer




- 1. Módszer
  - Keressük ki a legközelebbi szomszédot – milyen messze van?
  - Küszöbérték felett
- 2. Módszer
  - Keressük ki a k. legközelebbi szomszédot – milyen messze van?
  - Küszöbérték felett

# K közép (K means) algoritmus



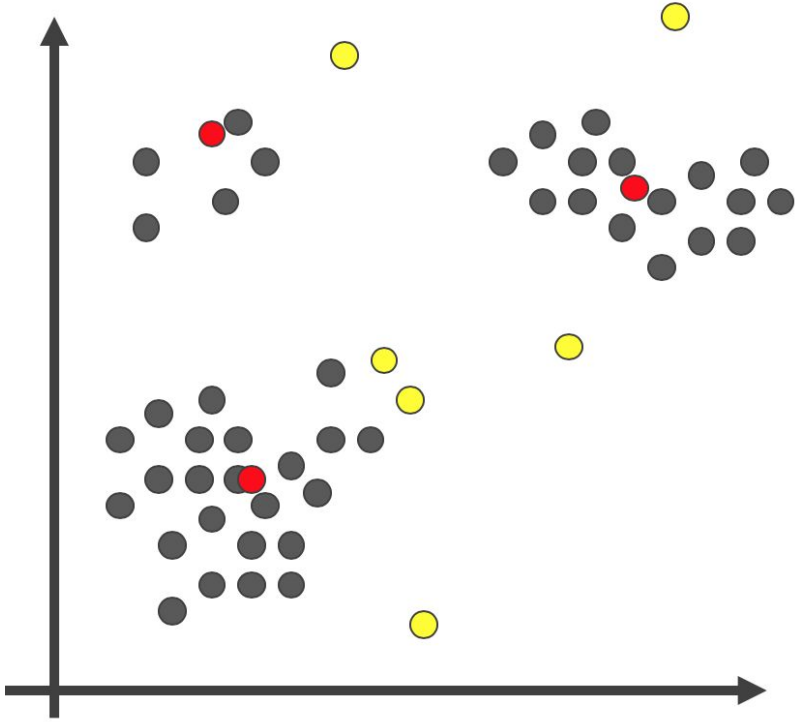
K-means algoritmus

- Véletlenszerű pontok kiindulási centroid

1. Legközelebbi centroidhoz rendelt pontok
2. Centroid újraszámítás

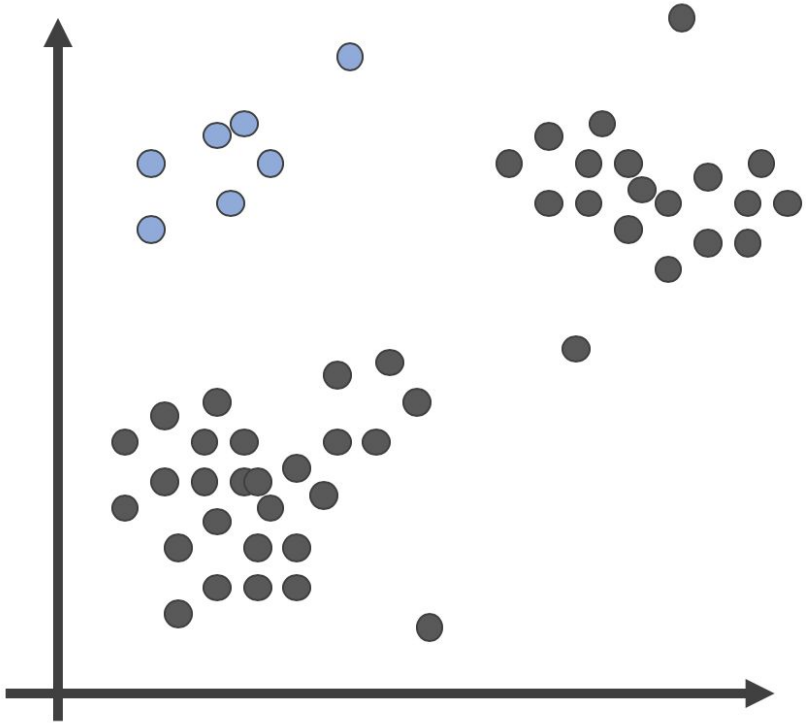
- Iteráció, míg nem változik

# K-Means klaszterezés



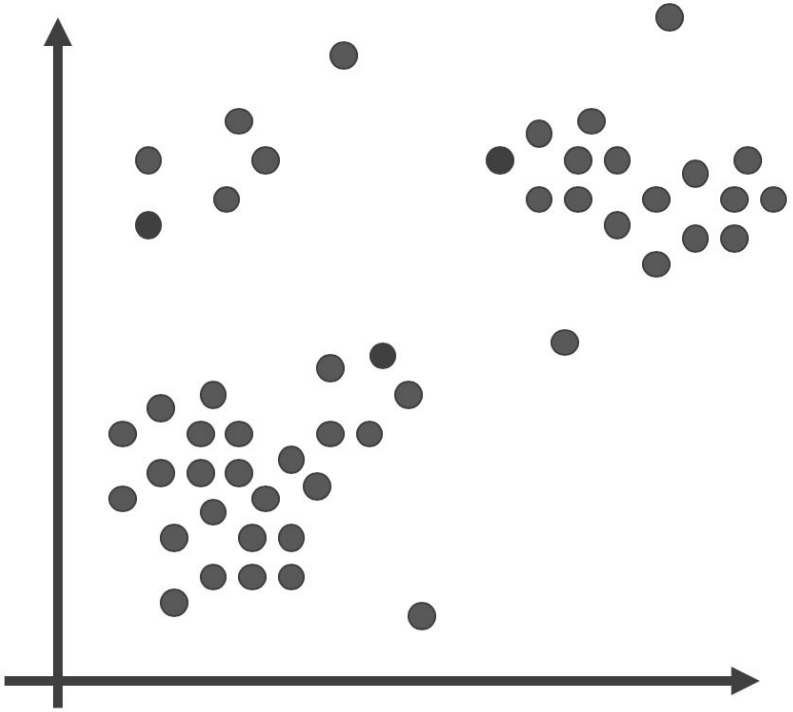
- Első megközelítés
  - Anomália az a pont, amely messze esik a centroidjától.

# K-Means klaszterezés



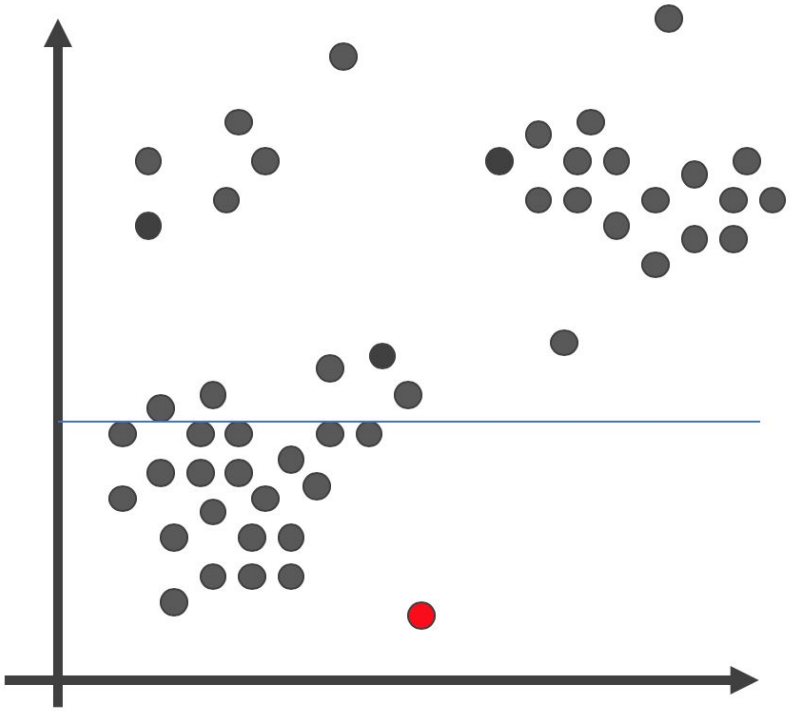
- Első megközelítés
  - Anomália az a pont, amely messze esik a centroidjától.
- Második megközelítés
  - Anomália = kicsit és ritka klaszterben van

# Isolation Forest



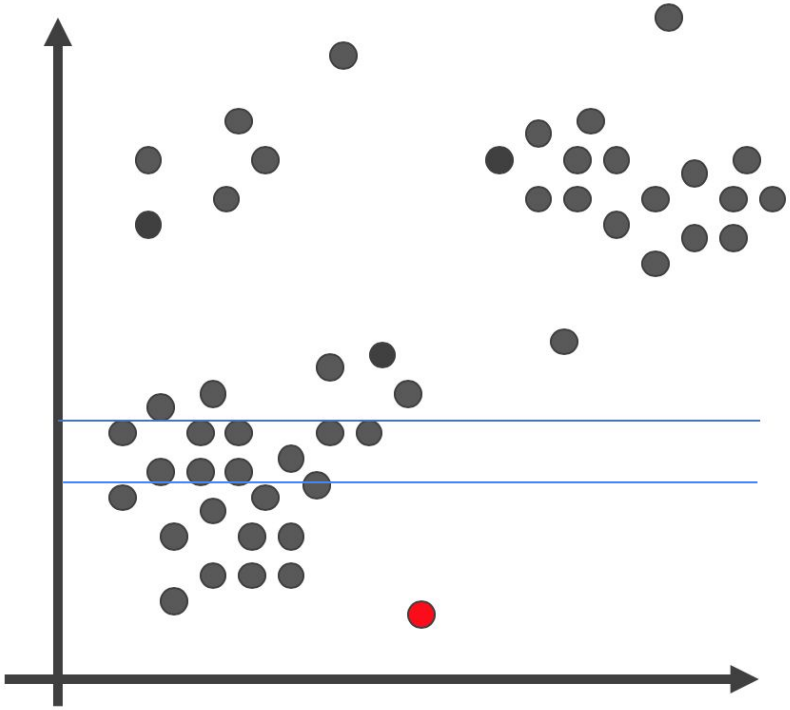
- Isolation Forest algoritmus
  - Válasszunk ki egy pontot, amit izolálni szeretnénk
  - Válasszunk egy random változót
  - Csináljunk egy random vágást ennek a változónak a segítségével
  - Ismételjük meg az előző 2 lépést, amíg teljesen el nincs izolálva a pont
  - Isolation Number: hány lépésre volt szükség?

# Isolation Forest



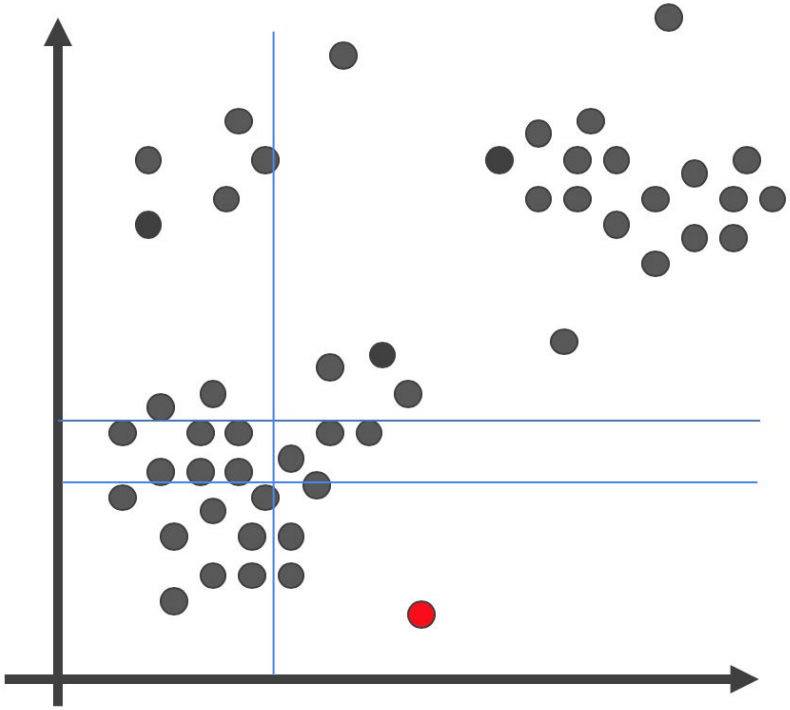
- Isolation Forest algoritmus
  - Válasszunk ki egy pontot, amit izolálni szeretnénk
  - Válasszunk egy random változót
  - Csináljunk egy random vágást ennek a változónak a segítségével
  - Ismételjük meg az előző 2 lépést, amíg teljesen el nincs izolálva a pont
  - Isolation Number: hány lépésre volt szükség?

# Isolation Forest



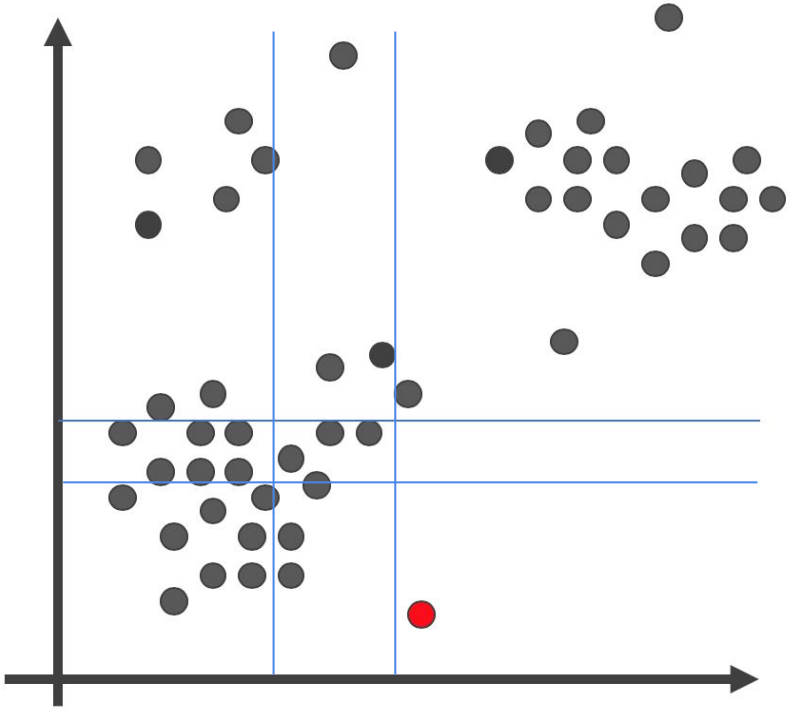
- Isolation Forest algoritmus
  - Válasszunk ki egy pontot, amit izolálni szeretnénk
  - Válasszunk egy random változót
  - Csináljunk egy random vágást ennek a változónak a segítségével
  - Ismételjük meg az előző 2 lépést, amíg teljesen el nincs izolálva a pont
  - Isolation Number: hány lépésre volt szükség?

# Isolation Forest



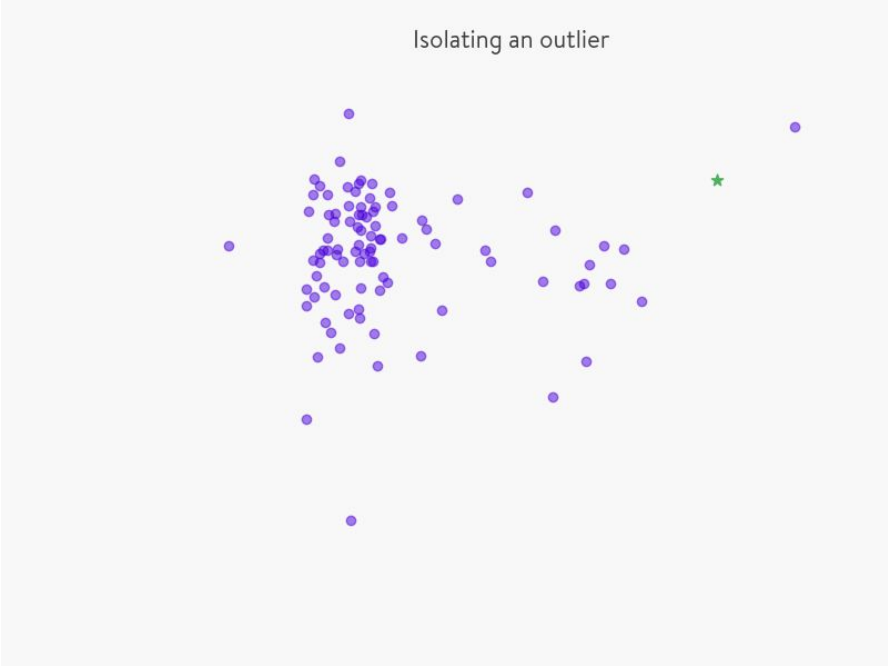
- Isolation Forest algoritmus
  - Válasszunk ki egy pontot, amit izolálni szeretnénk
  - Válasszunk egy random változót
  - Csináljunk egy random vágást ennek a változónak a segítségével
  - Ismételjük meg az előző 2 lépést, amíg teljesen el nincs izolálva a pont
  - Isolation Number: hány lépésre volt szükség?

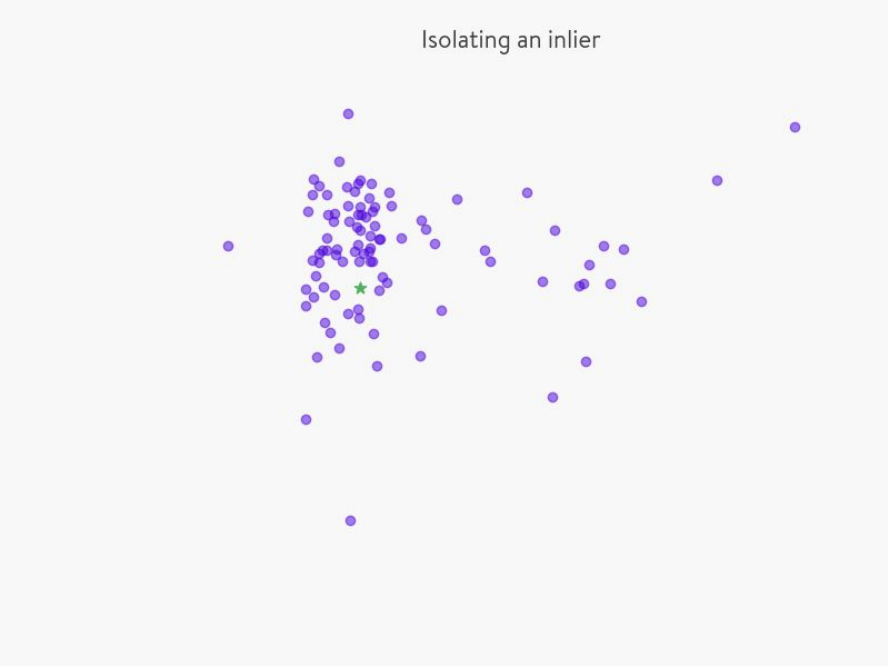
# Isolation Forest



- Isolation Forest algoritmus
  - Válasszunk ki egy pontot, amit izolálni szeretnénk
  - Válasszunk egy random változót
  - Csináljunk egy random vágást ennek a változónak a segítségével
  - Ismételjük meg az előző 2 lépést, amíg teljesen el nincs izolálva a pont
  - Isolation Number: hány lépésre volt szükség?

# Isolation Forest



# Gyakorlati munka

## Közös munkához:

- Internethozzáférés:
  - wifi: DanubiusFree
  - password: Danubius
- Adathalmaz és kiinduló kódok:

  bit.ly/2023aw

- Jupyter Notebook
  - colab.research.google.com
  - saját Notebook szerver (anaconda.com)

## Kerettörténet - 1 év munkája

- Szerződések vizsgálata (~2500 darab)
- Gyanúk alapján nyomozás - évente 111 darab
- Új data science csapat: gyanúk generálása
- Közben: tanult módszereket kipróbáljuk
- Végén ezeket kombináljuk
- Folyamatosan kiértékeljük az eredményeket

*Végső lépés - már a szimuláción kívül*

# Írd meg nekem a saját történeted!

GÁSPÁR CSABA
data scientist, CEO
gaspar.csaba@dmlab.hu
+36 (20) 823 4154
dmlab.hu/blog

dmlab