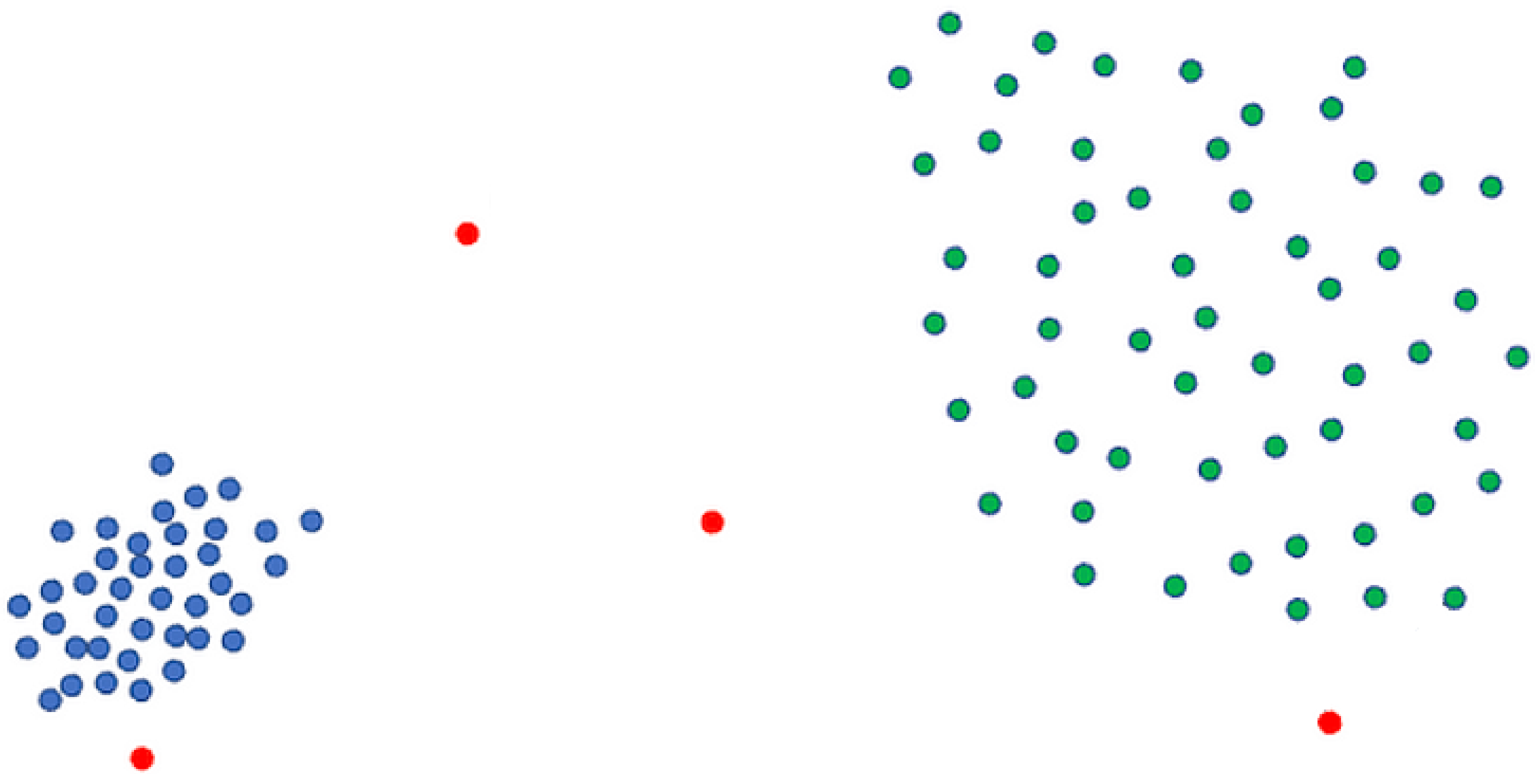
# Outlier detection & Isolation Forest

12. 11. 24

Glötzl Alexander
**FAKULTÄT FÜR PHYSIK**

Universität Regensburg

# What is an outlier?

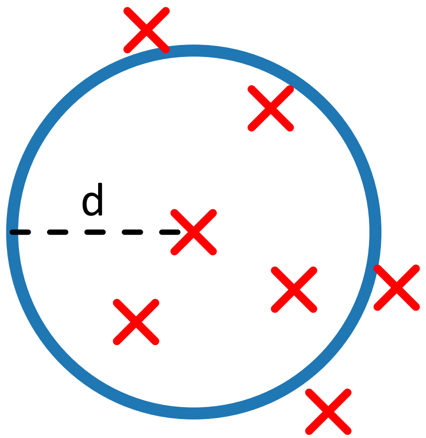


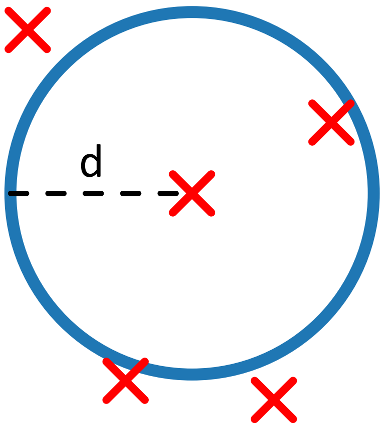
→ 5 different definitions

Universität Regensburg

# 1. Distance-based methods

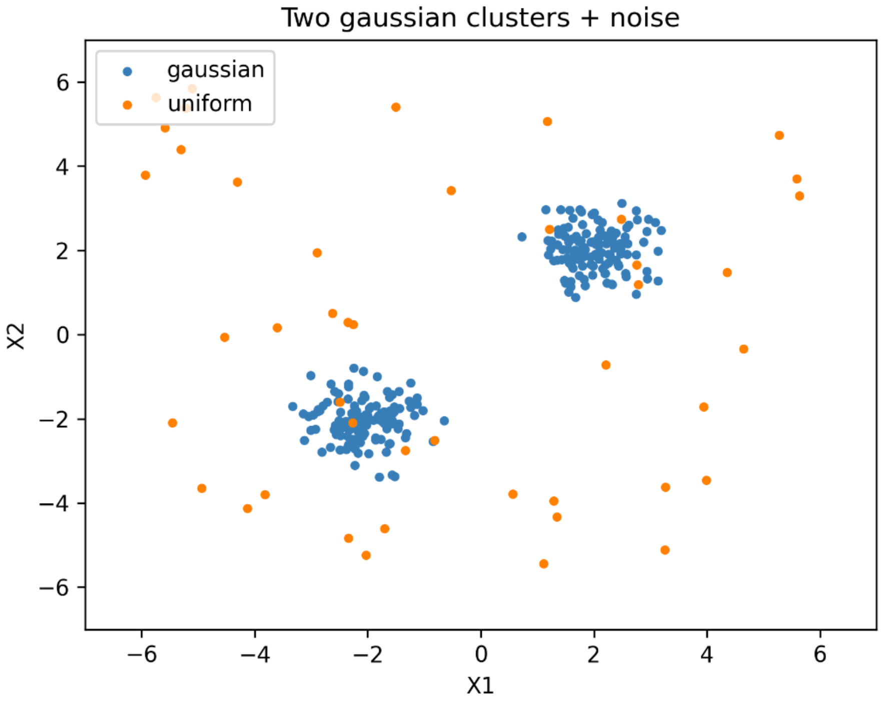
- **Definition 1:**
  Outliers are the examples for which there are fewer than **p** other examples within distance **d**

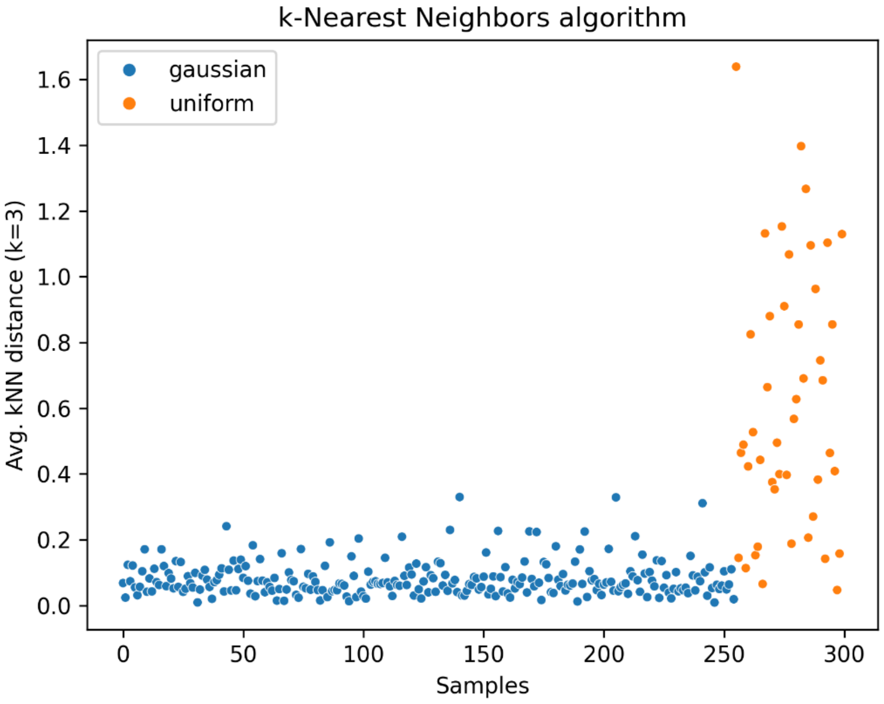


Universität Regensburg

- **Definition 2:**
  Outliers are the top **n** examples whose average distance to the **k** nearest neighbors is greatest

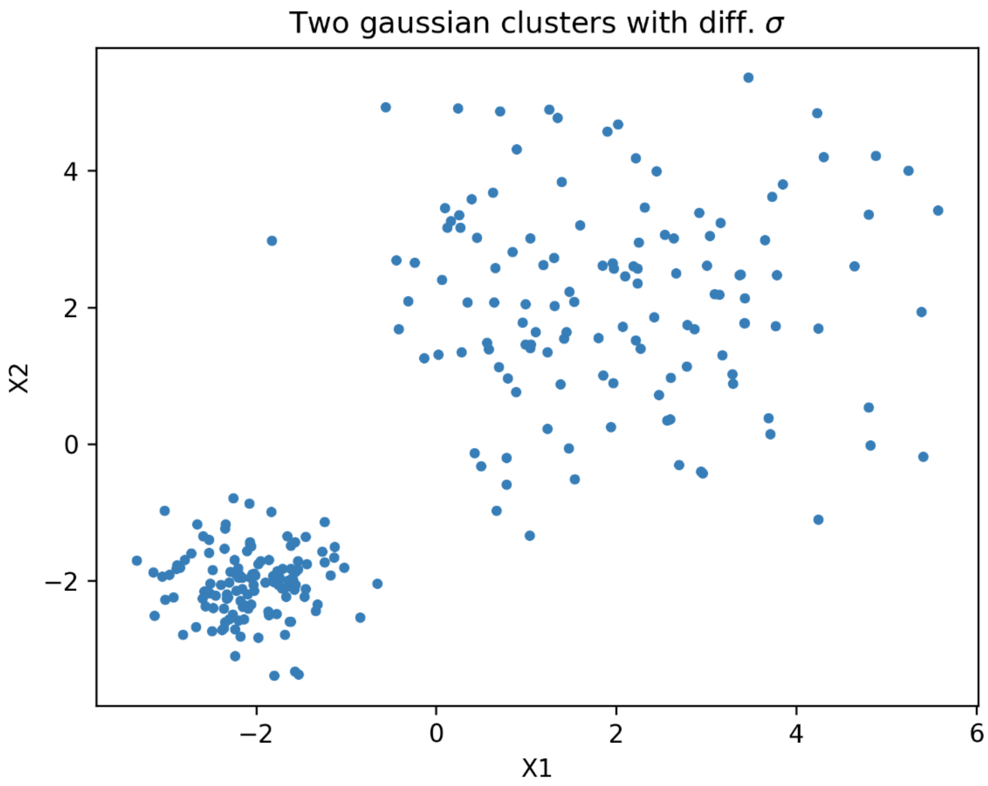


Universität Regensburg

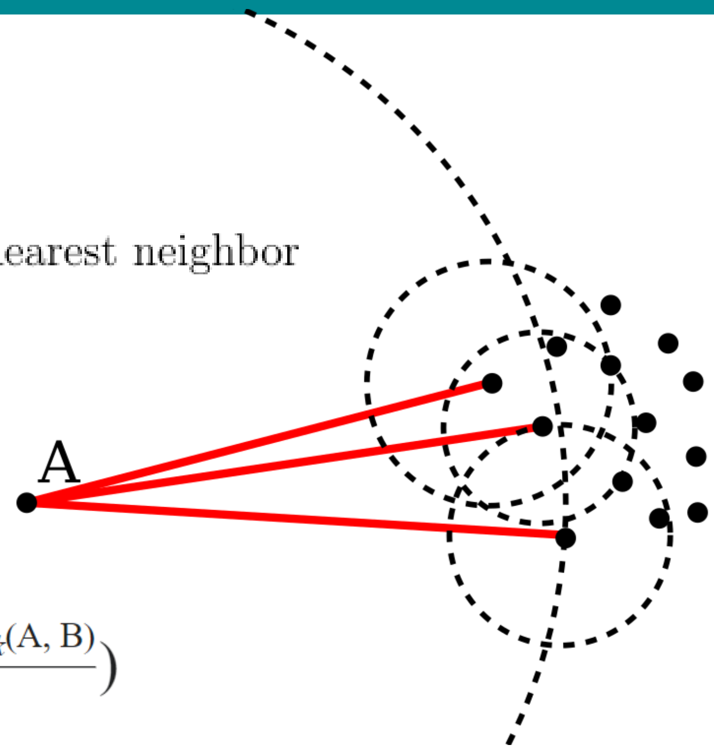
# 2. Non-Parametric Density-Based Methods

- **Example 1:** Local Outlier Factor (LOF)



$k\text{–distance}(A) :=$ distance between object A to its k-th nearest neighbor

$N_k(A) :=$ set of k nearest neighbors




$$\text{reachability–distance } k(\text{A,B}) = \max\{k\text{–distance}(\text{B}), \text{d}(\text{A,B})\}$$

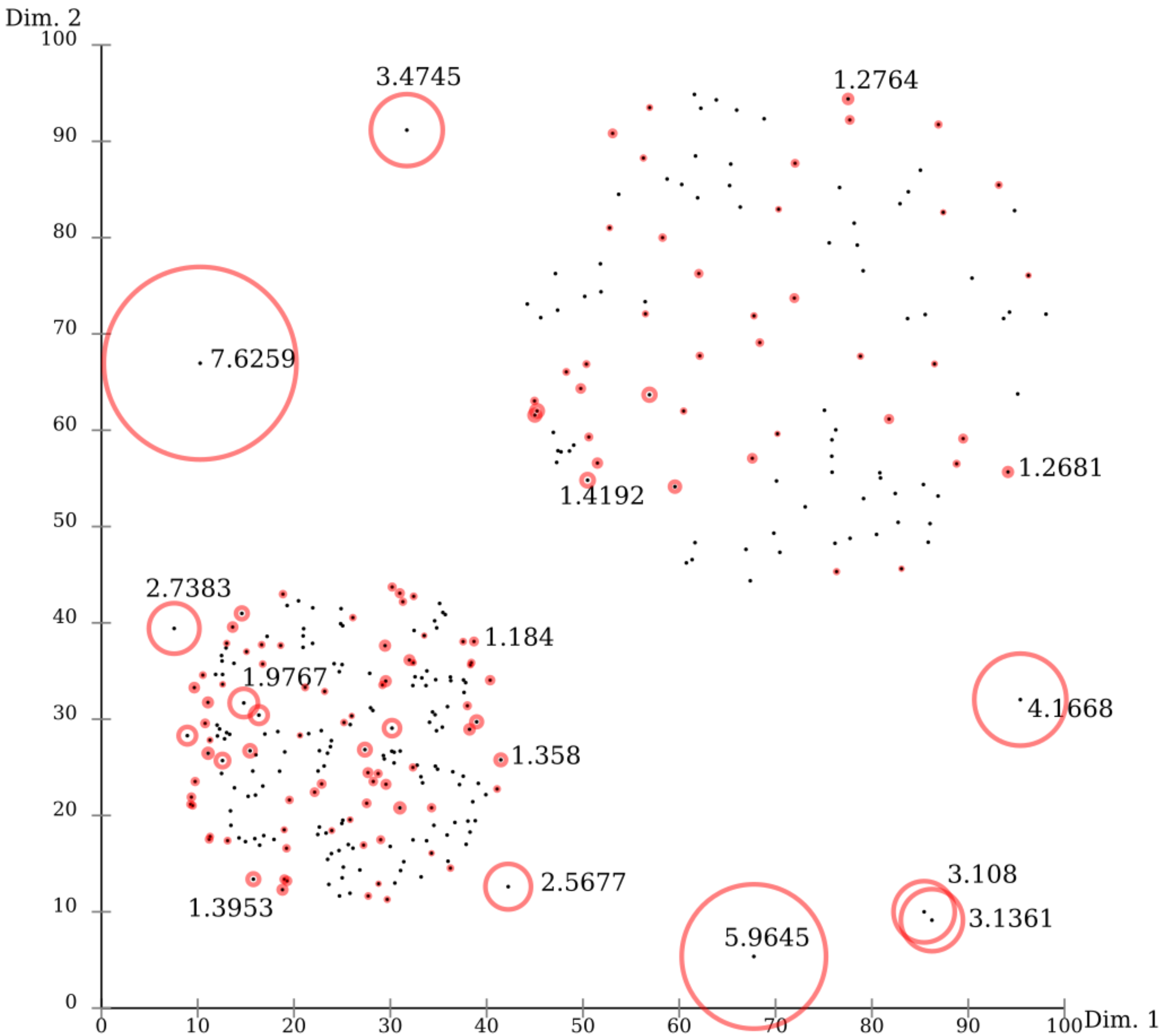
$$\text{local reachability density}_k(\text{A}) := 1/\left(\frac{\sum_{\text{B} \in \text{N}_k(\text{A})} \text{reachability-distance}_k(\text{A, B})}{|\text{N}_k(\text{A})|}\right)$$

$$\text{LOF}_k(\text{A}) := \frac{\sum_{\text{B} \in \text{N}_k(\text{A})} \frac{\text{lrd}_k(\text{B})}{\text{lrd}_k(\text{A})}}{|\text{N}_k(\text{A})|}$$

$\text{LOF(k)} \sim 1$ means **Similar density as neighbors,**

$\text{LOF(k)} < 1$ means **Higher density than neighbors (Inlier),**

$\text{LOF(k)} > 1$ means **Lower density than neighbors (Outlier)**




**Advantage**:
- identifies outliers under consideration of *local* neighborhood

**Disadvantage**:
- there is no clear boundary for LOF-score (LOF=2.0 might be outlier for one dataset but not the other)

# 2. Non-Parametric Density-Based Methods

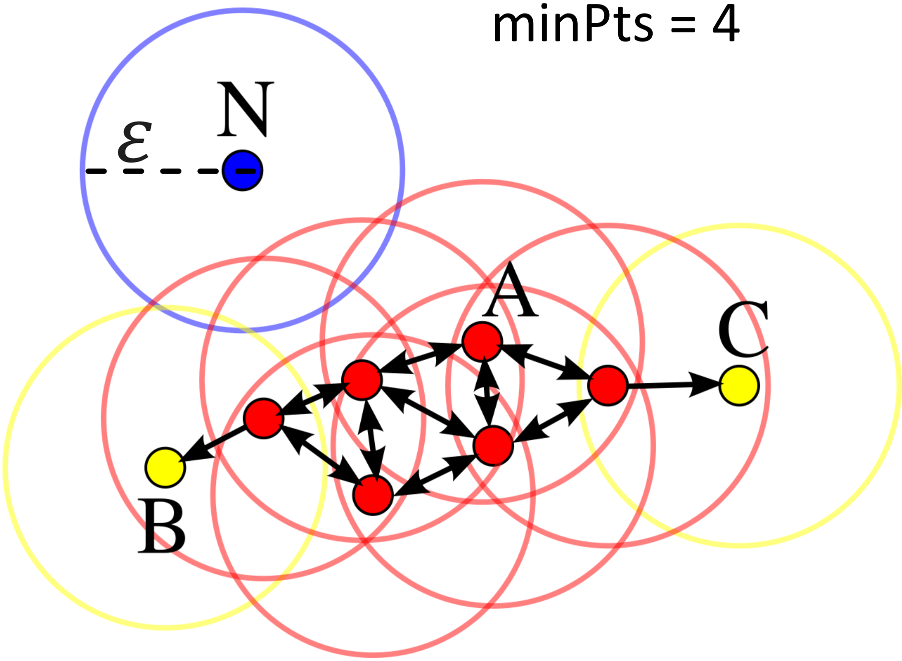
- **Example 2:** Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

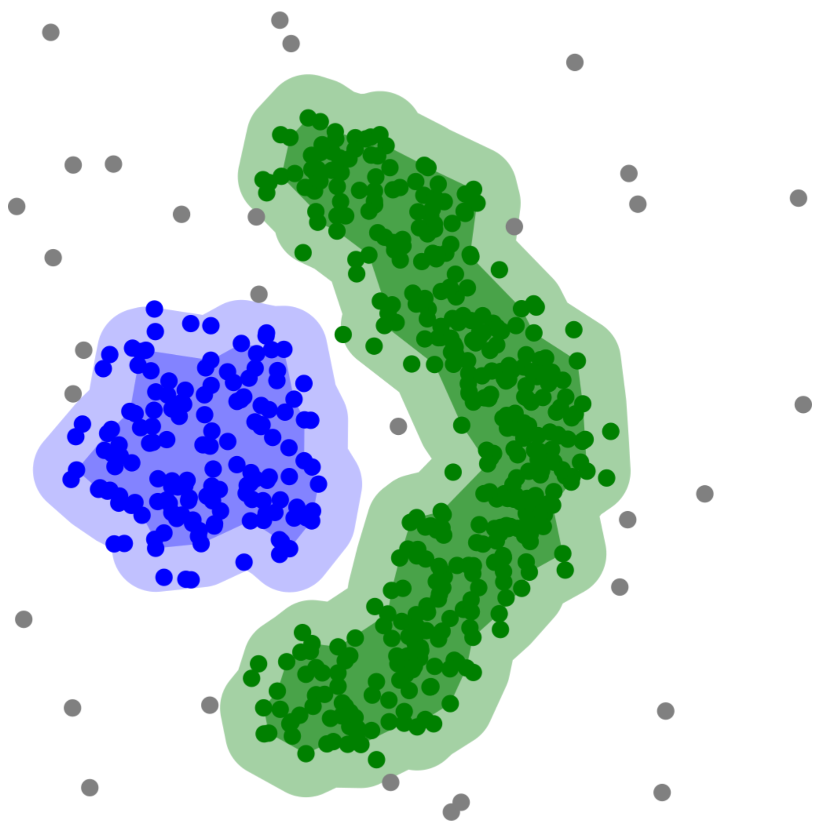
**Parameters:** minPts, $\varepsilon$

**Core Point** A**:**
- if at least minPts are within distance $\varepsilon$

**Outlier**:
- all points that are not reachable

minPts = 4





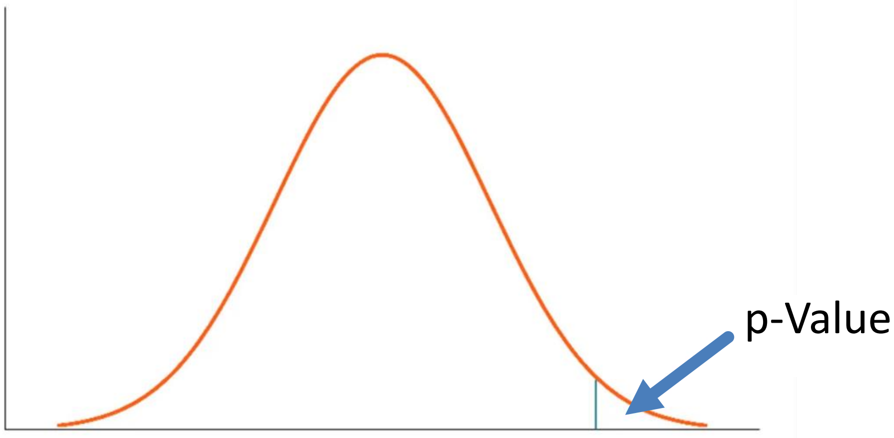
**Advantage**:
- outliers are "byproduct"
- finds clusters without specifying number of clusters a priori

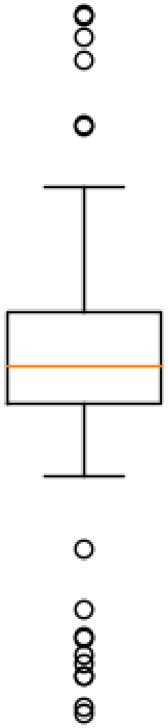
**Disadvantage**:
- does not handle different densities well
- depends on distance measure
  → in high-dimensional data: "curse of dimensionality"

# 3. Parametric Density-Based Methods

$$z = \frac{x - \mu}{\sigma}$$

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$$



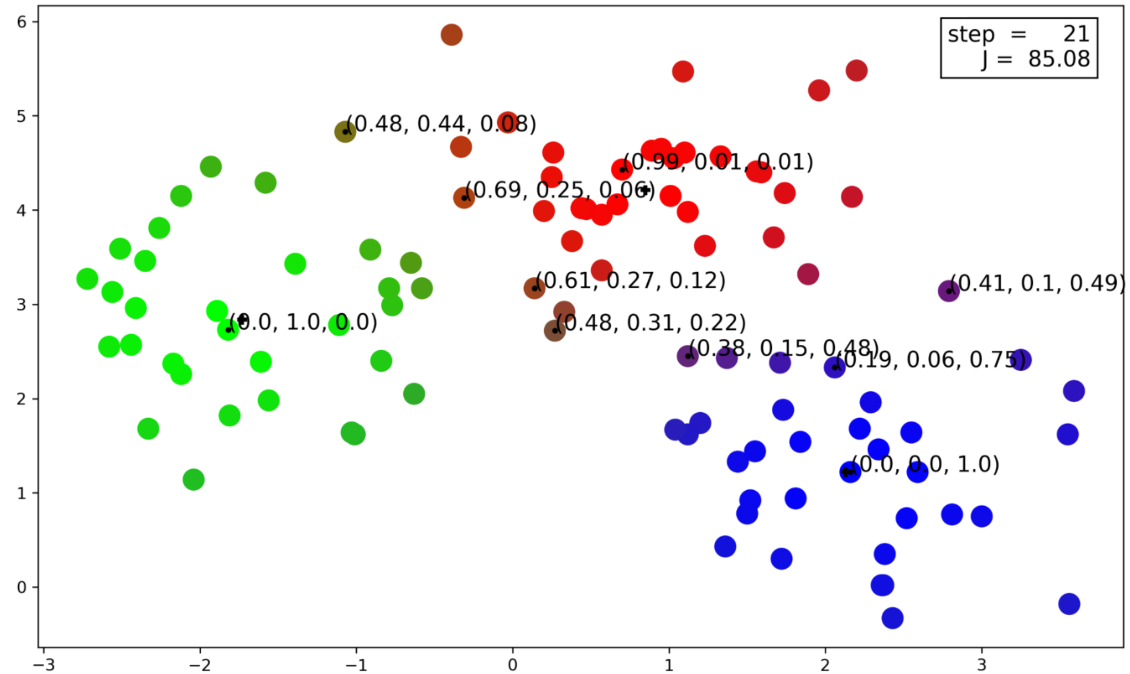



# 4. Cluster-based methods

Anomaly scores based on:

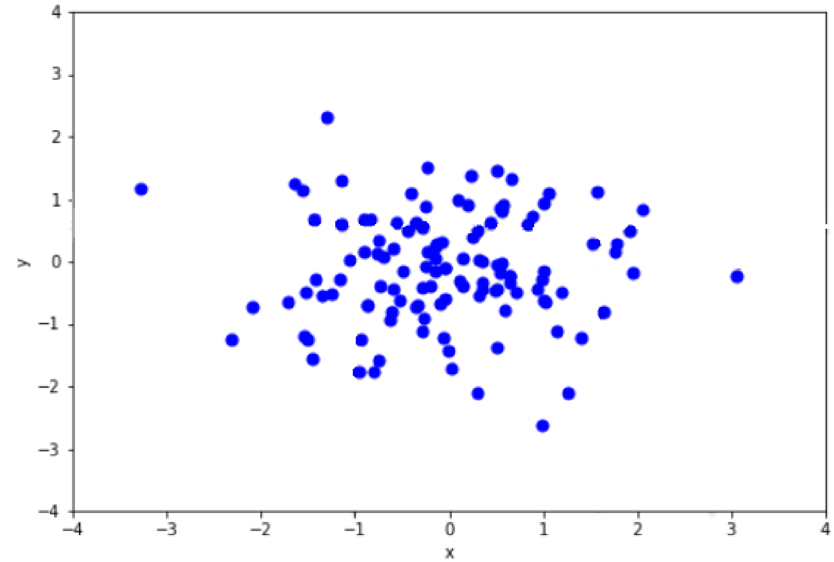
- cluster membership
- distance from other clusters
- size of the closest cluster

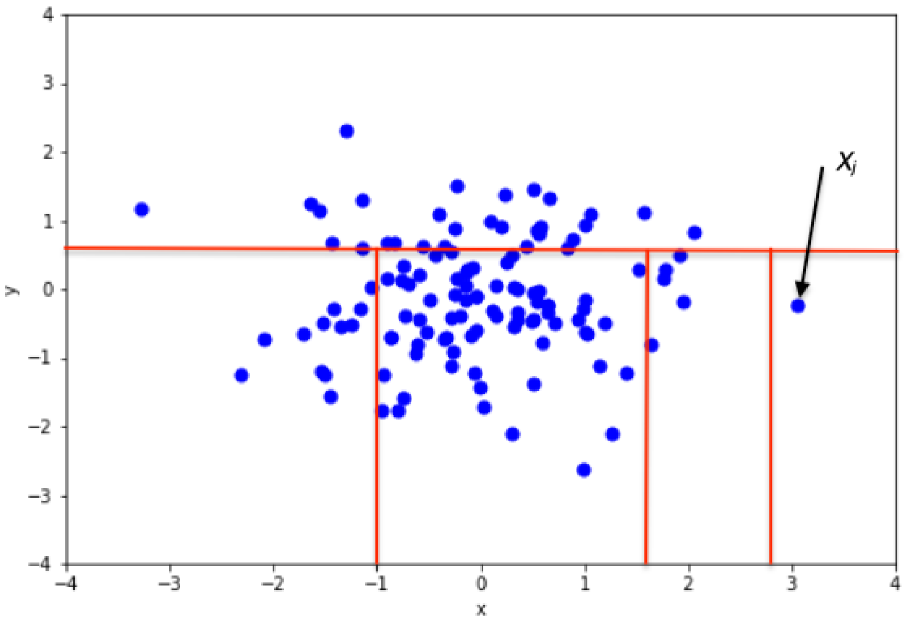
fuzzy C-means Clustering

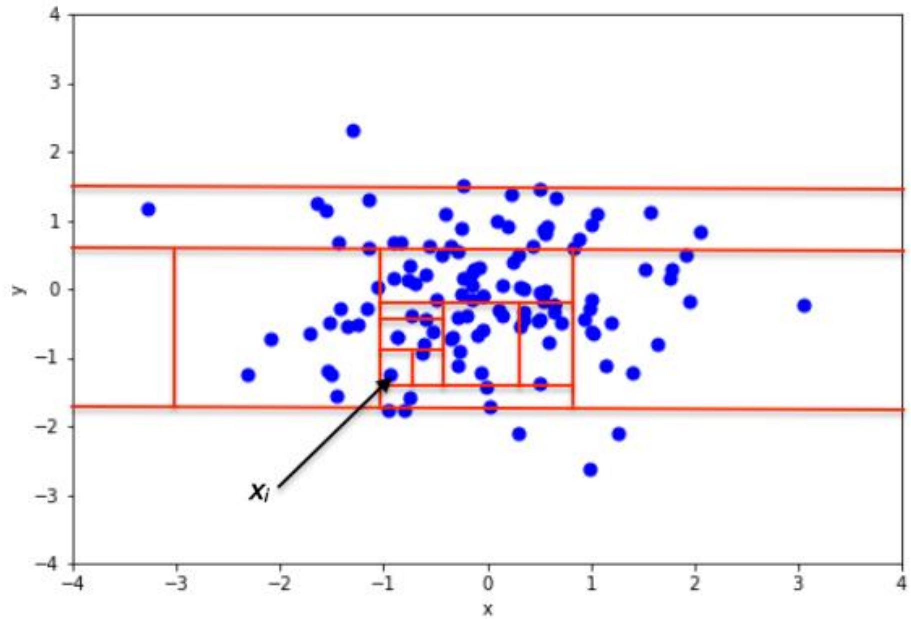


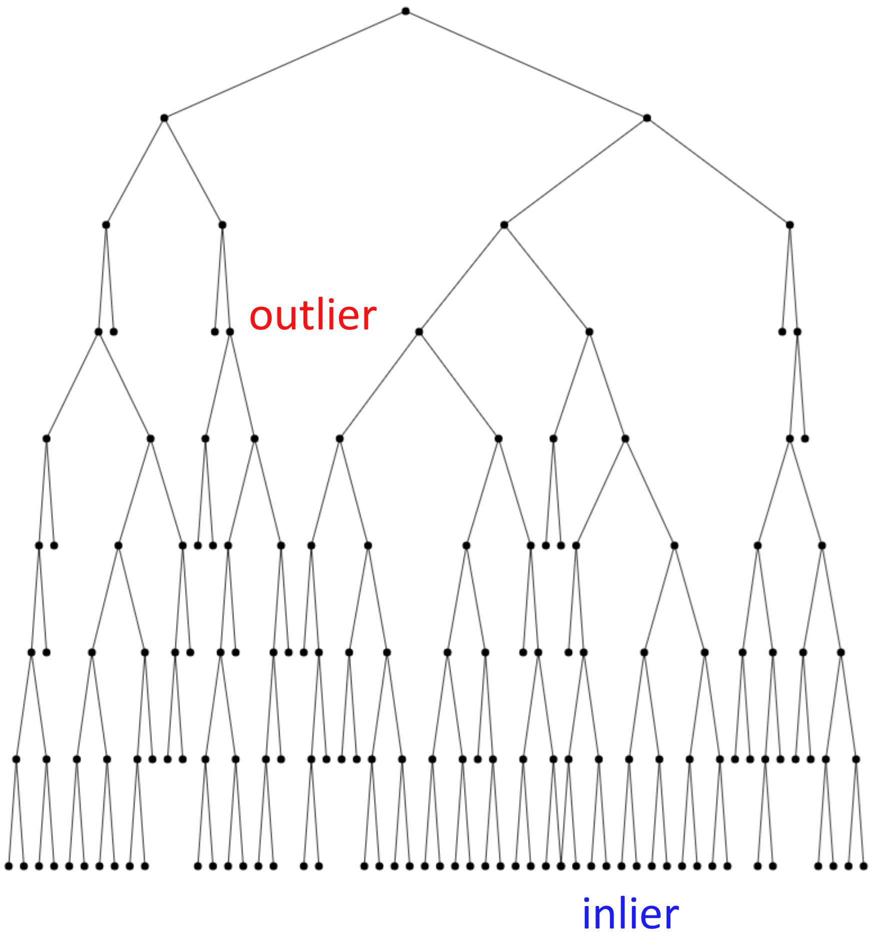
# 5. Depth-based method

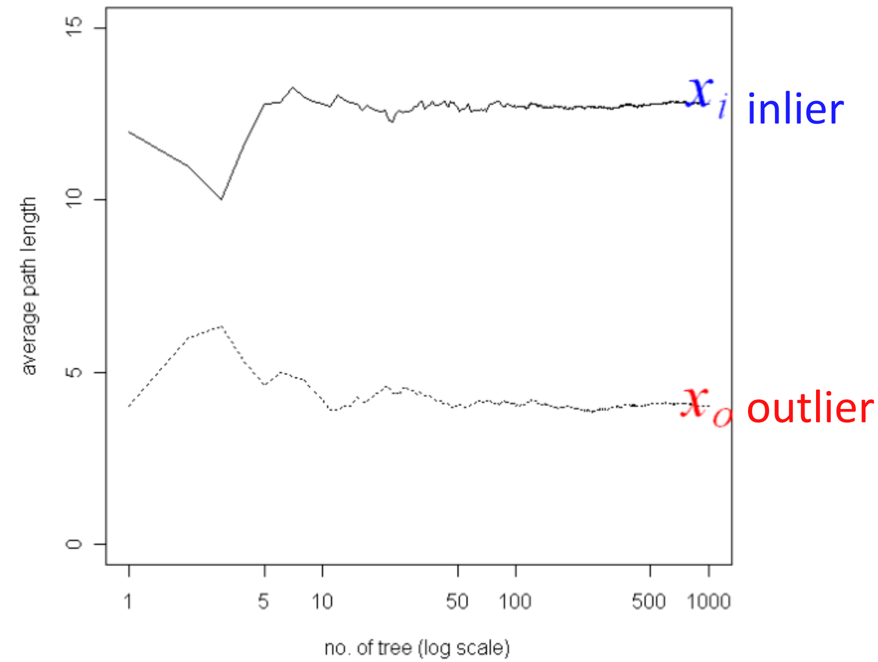
- **Example:** Isolation Forest

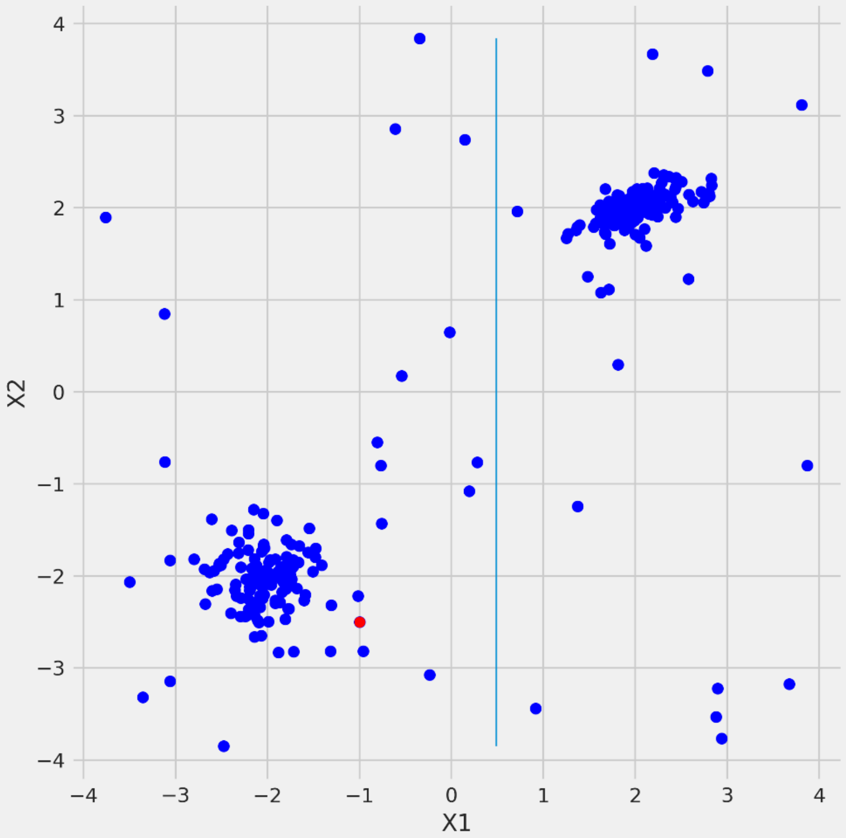


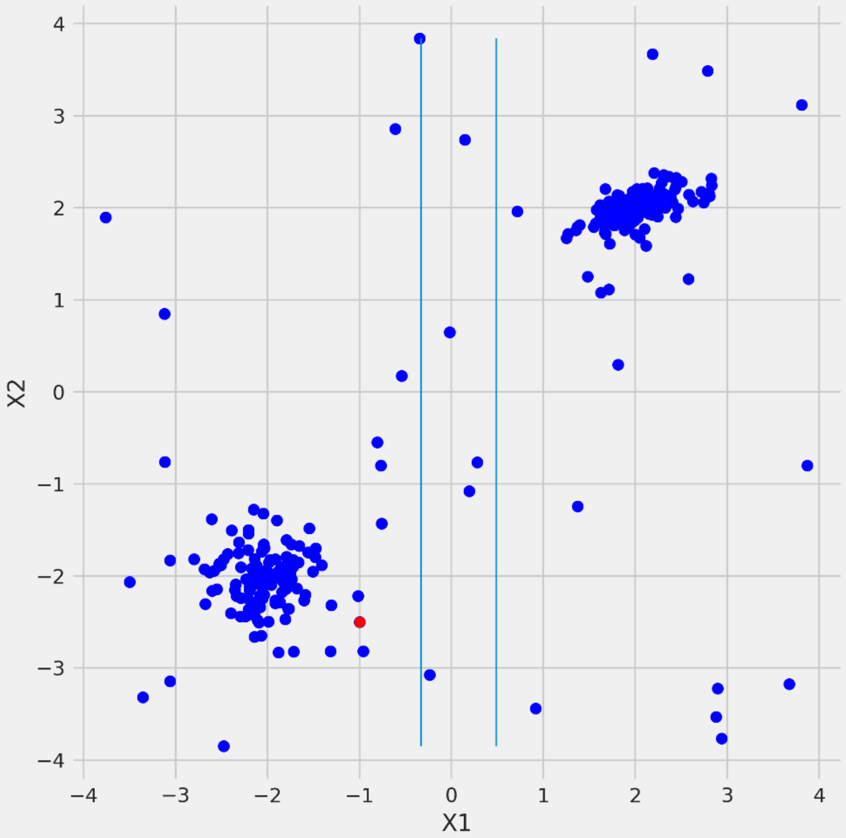



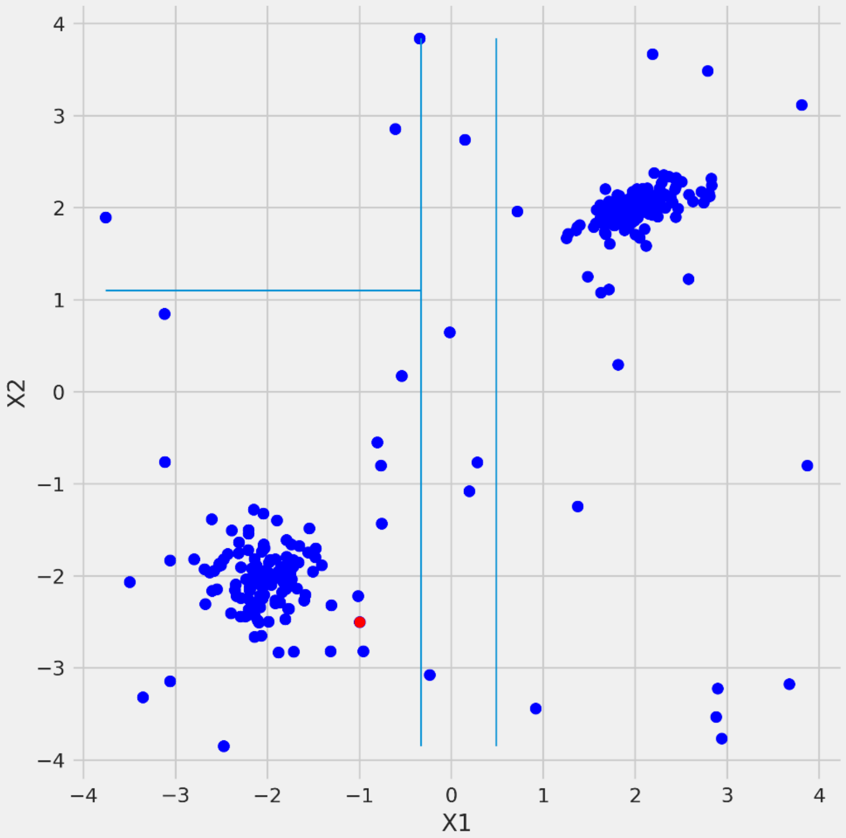

outlier
inlier

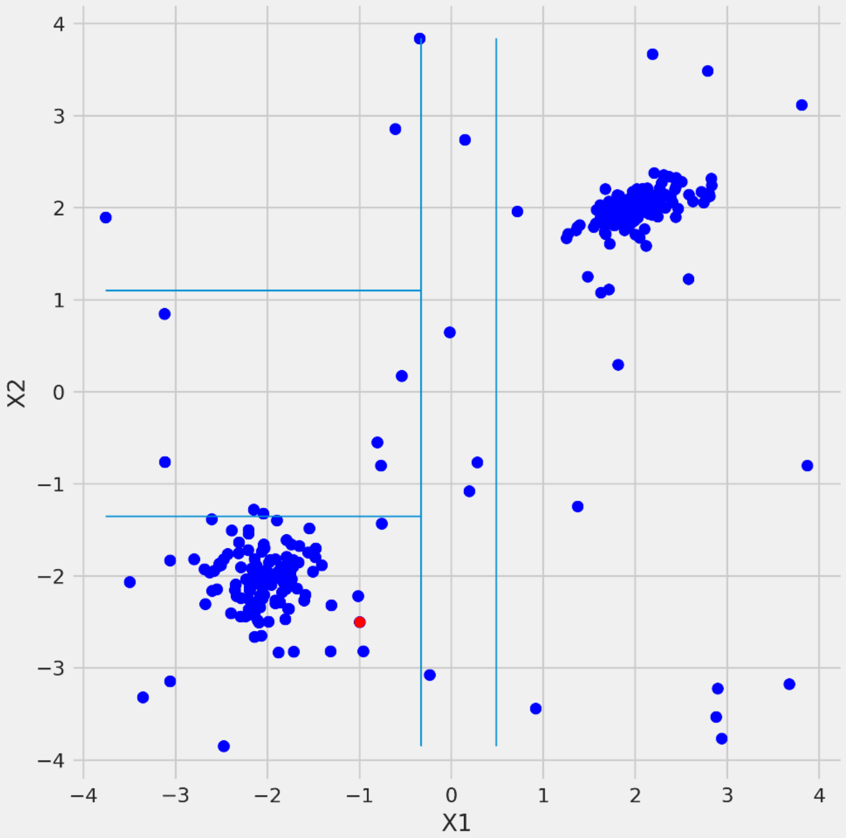

15
10
5
0
average path length
1
5
10
50
100
500
1000
no. of tree (log scale)
$x_i$ inlier
$x_o$ outlier



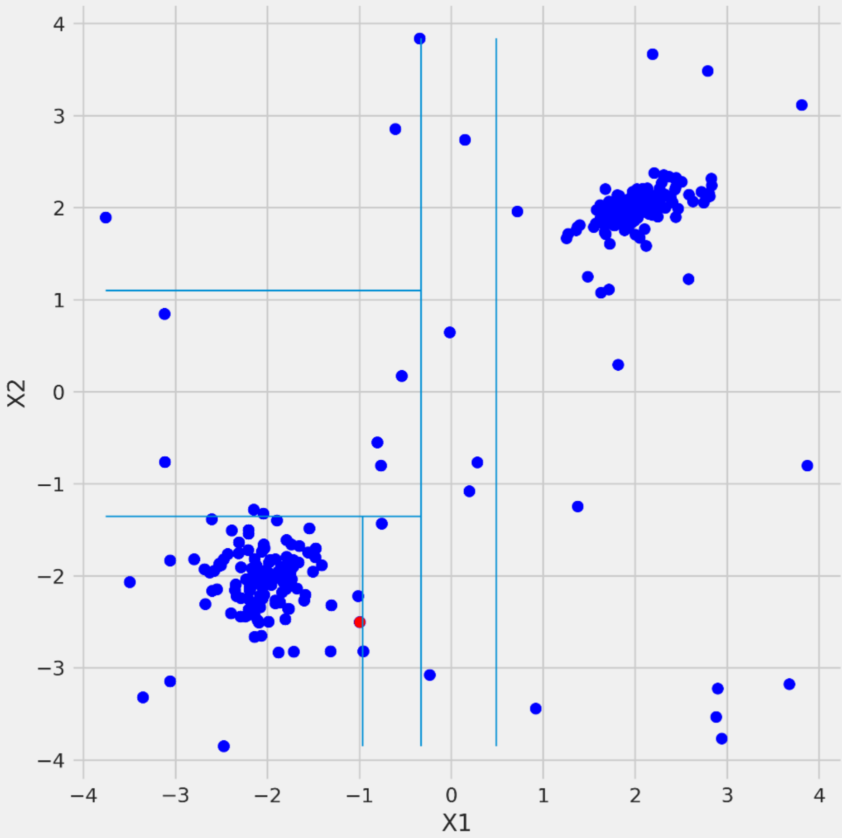
X1
X2
−4
−3
−2
−1
0
1
2
3
4

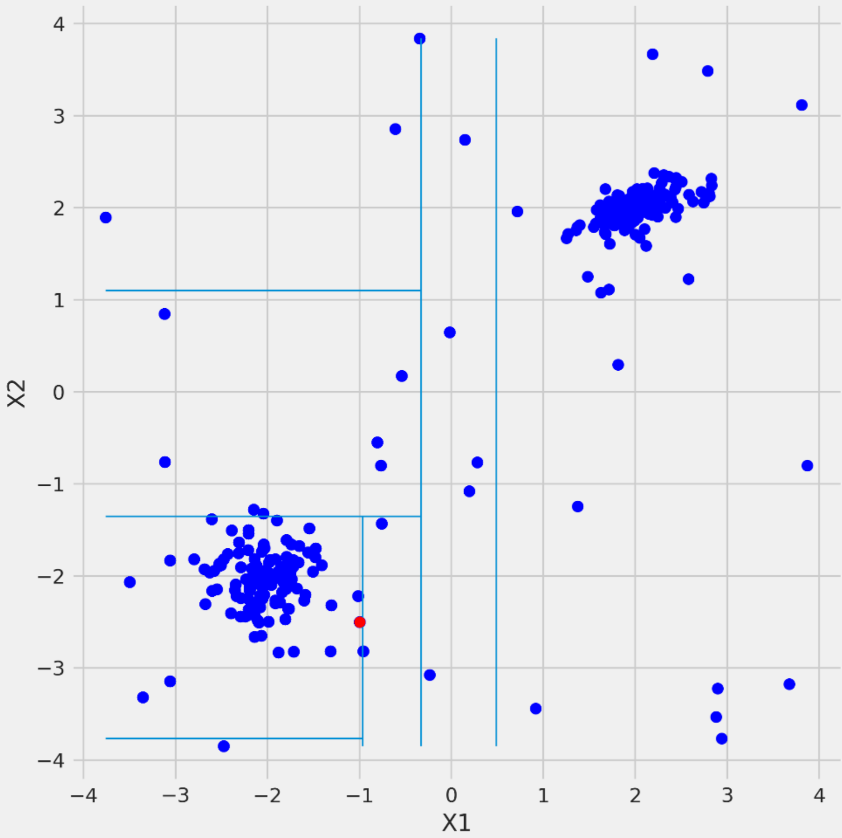

X1
X2
−4
−3
−2
−1
0
1
2
3
4

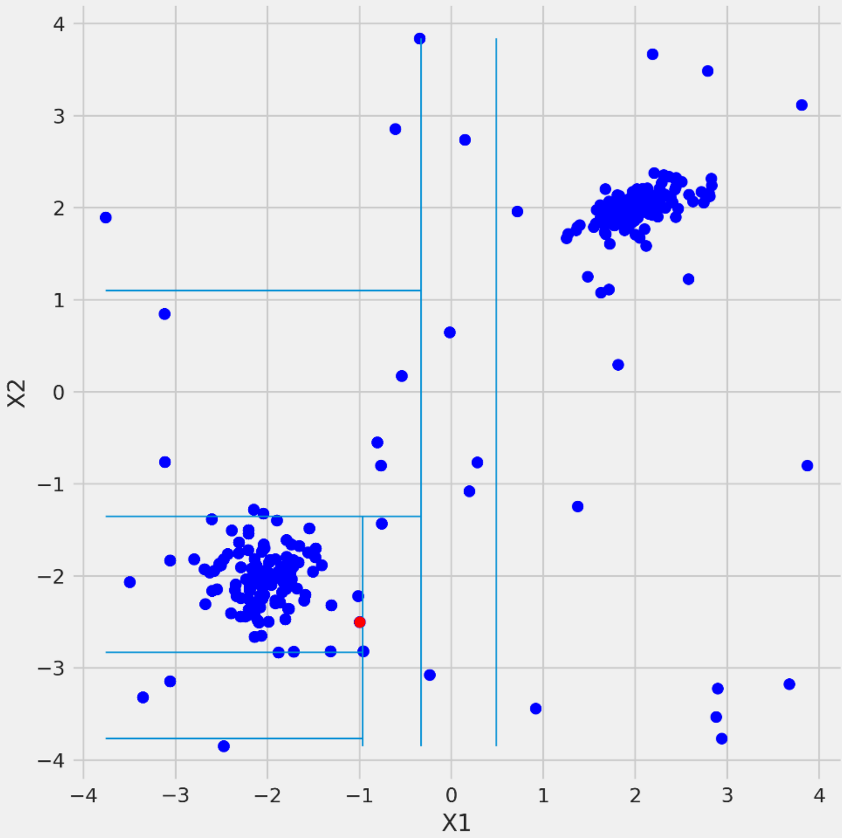

X2
X1
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4

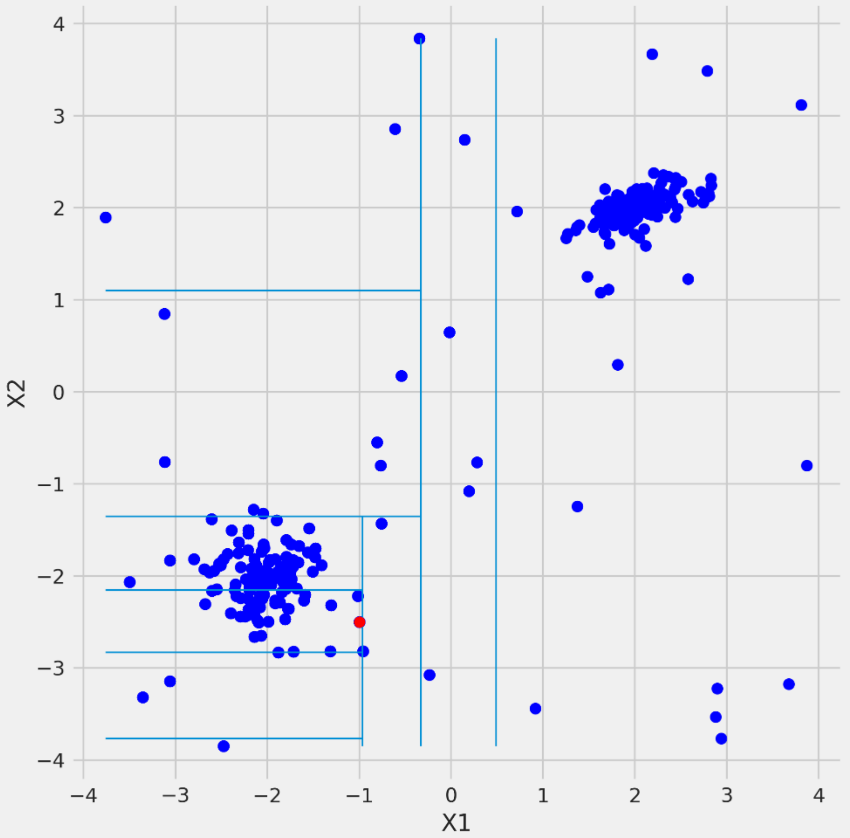


4
3
2
1
0
−1
−2
−3
−4
X2
−4
−3
−2
−1
0
1
2
3
4
X1

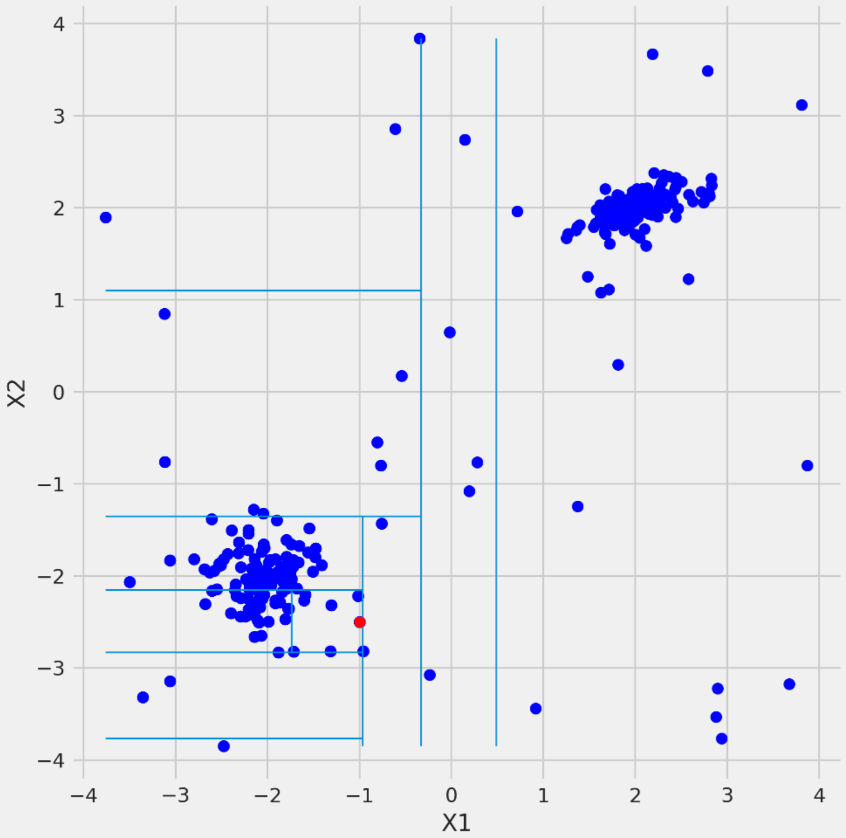

X1
X2
−4
−3
−2
−1
0
1
2
3
4

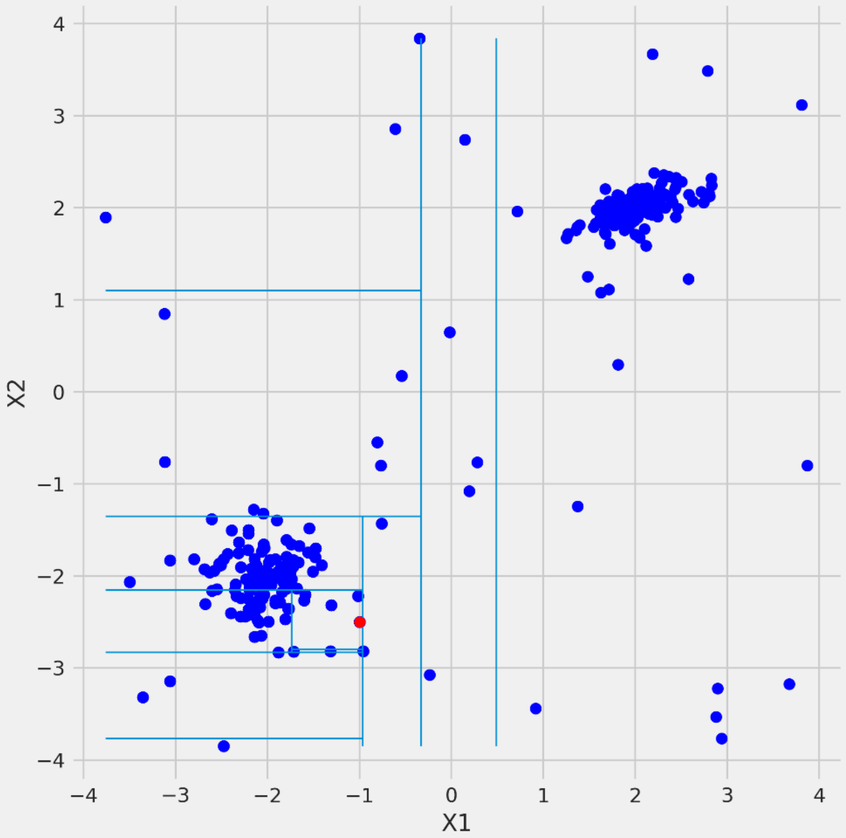

X1
X2

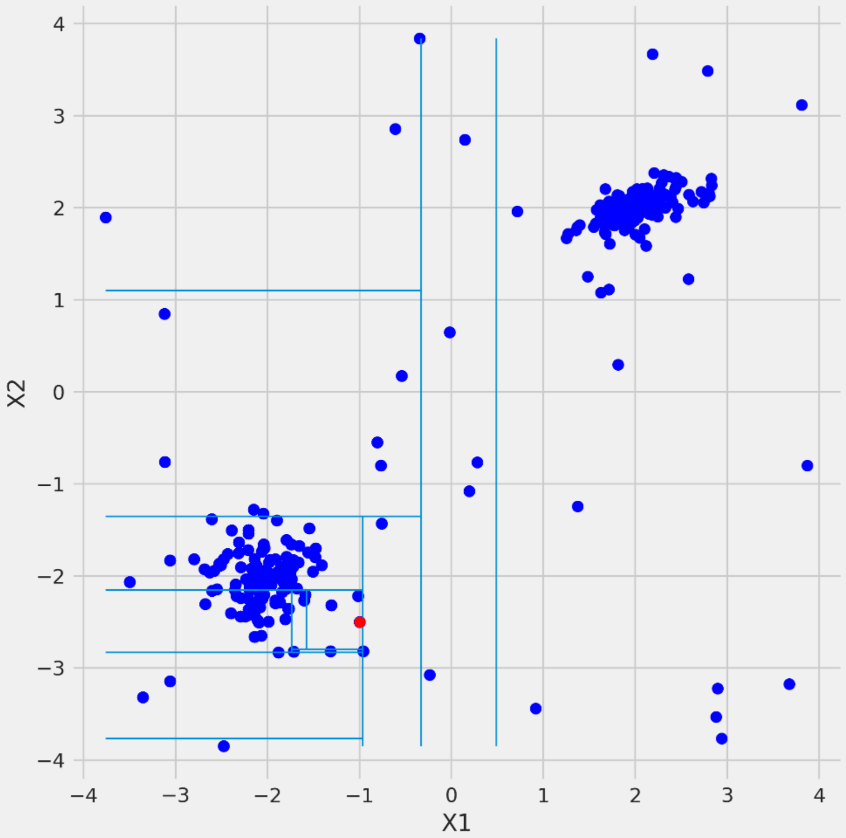

4
3
2
1
0
−1
−2
−3
−4
X2
−4
−3
−2
−1
0
1
2
3
4
X1

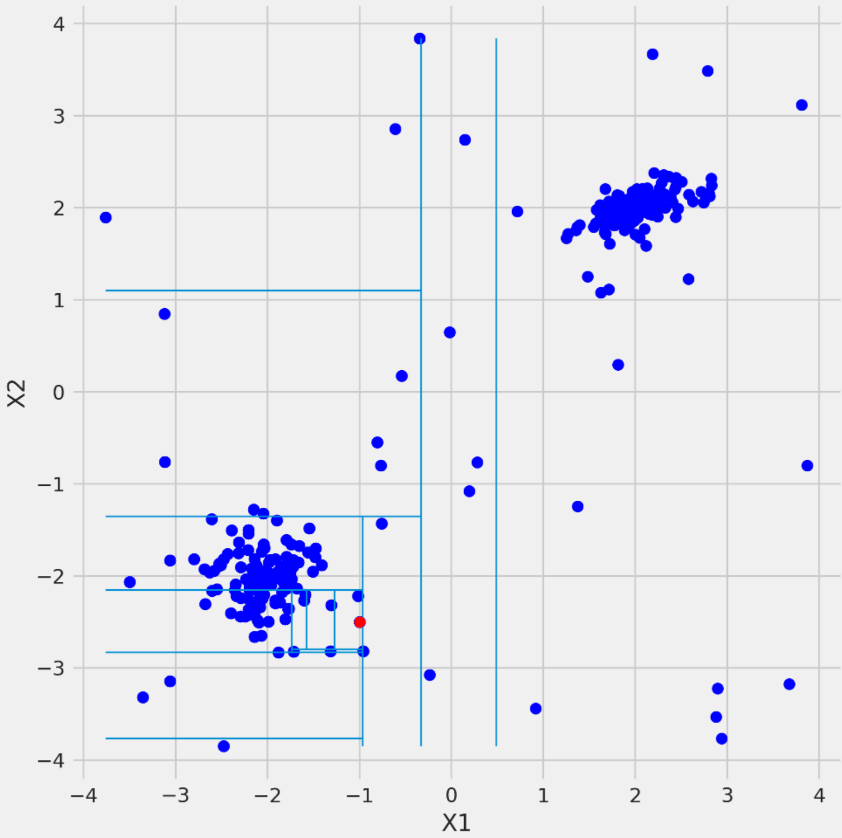

X1
X2
−4
−3
−2
−1
0
1
2
3
4

Universität Regensburg

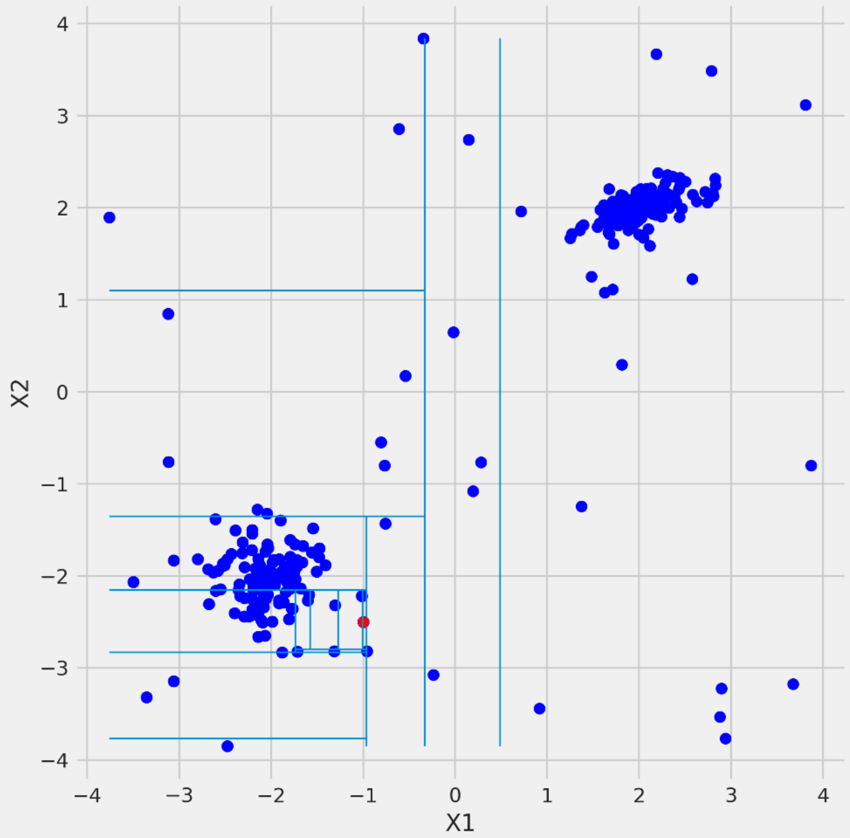



X1
X2
−4
−3
−2
−1
0
1
2
3
4


X1
X2
−4
−3
−2
−1
0
1
2
3
4



4
3
2
1
0
−1
−2
−3
−4
X2
−4
−3
−2
−1
0
1
2
3
4
X1

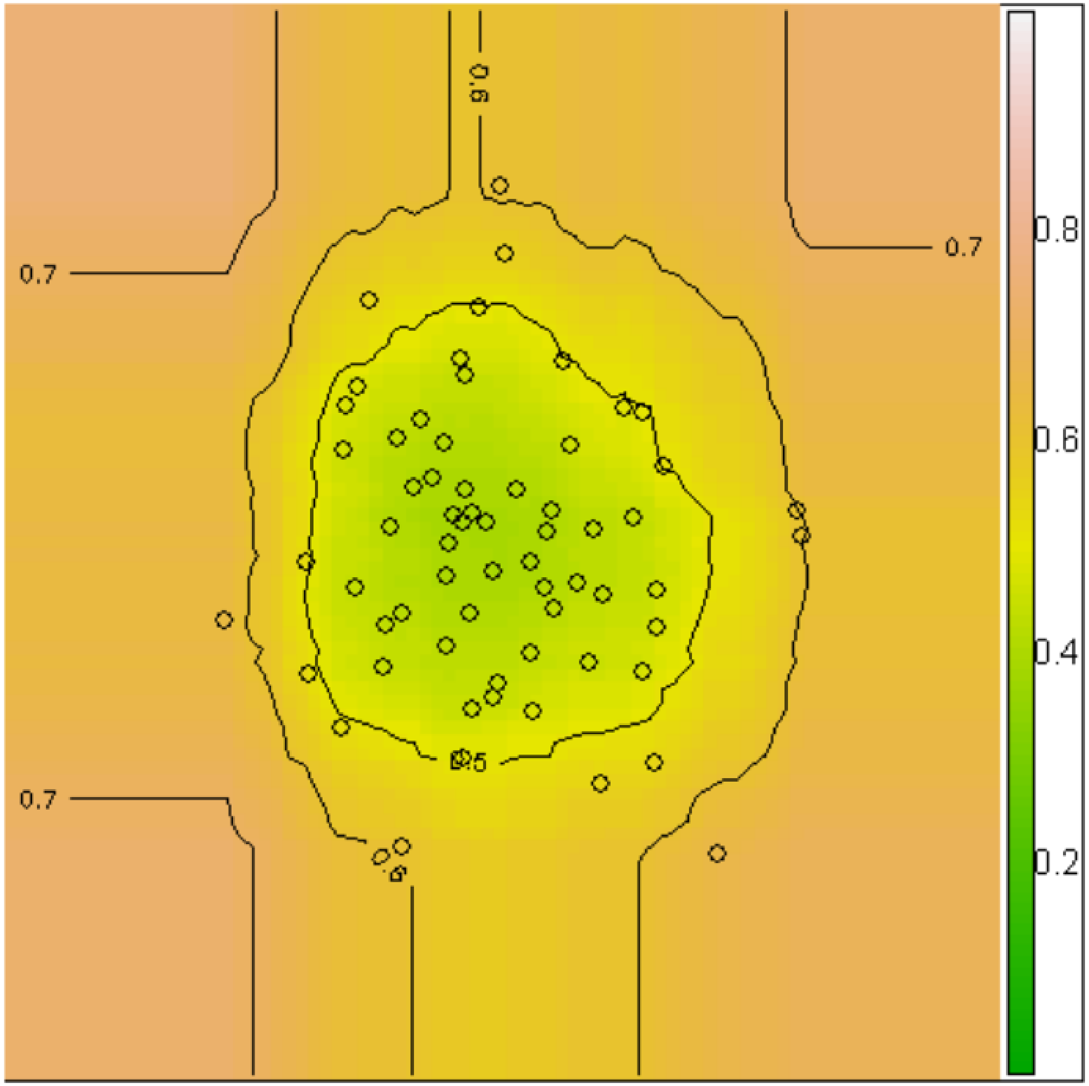

X1
X2
−4
−3
−2
−1
0
1
2
3
4

anomalie score s:

$$c(n) = 2H(n-1) - (2(n-1)/n)$$

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

s < 0.5 → x is inlier
s ≈ 1 → x is outlier
s ≈ 0.5 for all points → all inliers


0.6
0.7
0.7
0.7
0.5
0.6
0.8
0.6
0.4
0.2

**Advantages**:
- normal instances only:
  iForest performs well, even without anomalies
- fast:
  linear time complexity and a low memory use

| | $n$ | $d$ | anomaly class |
|---|---|---|---|
| Http (KDDCUP99) | 567497 | 3 | attack (0.4%) |
| ForestCover | 286048 | 10 | class 4 (0.9%) vs. class 2 |
| Mulcross | 262144 | 4 | 2 clusters (10%) |
| Smtp (KDDCUP99) | 95156 | 3 | attack (0.03%) |
| Shuttle | 49097 | 9 | classes 2,3,5,6,7 (7%) |
| Mammography | 11183 | 6 | class 1 (2%) |
| Annthyroid | 6832 | 6 | classes 1, 2 (7%) |
| Satellite | 6435 | 36 | 3 smallest classes (32%) |
| Pima | 768 | 8 | pos (35%) |
| Breastw | 683 | 9 | malignant (35%) |
| Arrhythmia | 452 | 274 | classes 03,04,05,07, 08,09,14,15 (15%) |
| Ionosphere | 351 | 32 | bad (36%) |

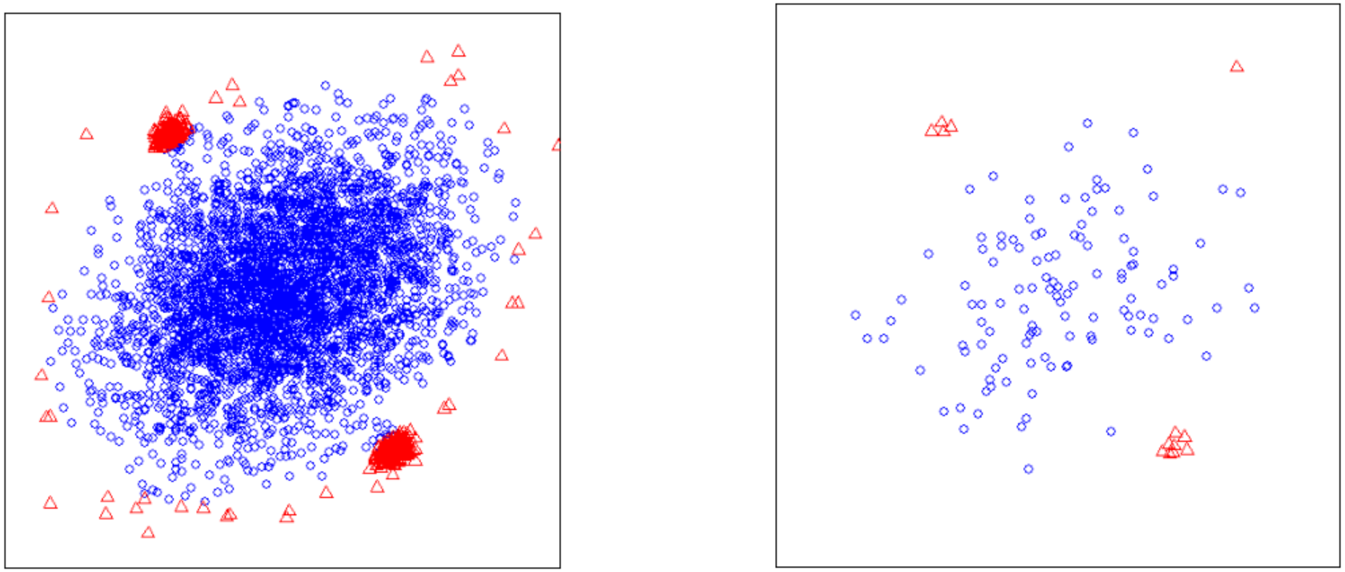
data sets

| | AUC | | | | Time (seconds) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | iForest | ORCA | LOF | RF | iForest | | | ORCA | LOF | RF |
| | | | | | Train | Eval. | Total | | | |
| Http (KDDCUP99) | **1.00** | 0.36 | NA | NA | 0.25 | 15.33 | **15.58** | 9487.47 | NA | NA |
| ForestCover | **0.88** | 0.83 | NA | NA | 0.76 | 15.57 | **16.33** | 6995.17 | NA | NA |
| Mulcross | **0.97** | 0.33 | NA | NA | 0.26 | 12.26 | **12.52** | 2512.20 | NA | NA |
| Smtp (KDDCUP99) | **0.88** | 0.80 | NA | NA | 0.14 | 2.58 | **2.72** | 267.45 | NA | NA |
| Shuttle | **1.00** | 0.60 | 0.55 | NA | 0.30 | 2.83 | **3.13** | 156.66 | 7489.74 | NA |
| Mammography | **0.86** | 0.77 | 0.67 | NA | 0.16 | 0.50 | **0.66** | 4.49 | 14647.00 | NA |
| Annthyroid | **0.82** | 0.68 | 0.72 | NA | 0.15 | 0.36 | **0.51** | 2.32 | 72.02 | NA |
| Satellite | **0.71** | 0.65 | 0.52 | NA | 0.46 | 1.17 | **1.63** | 8.51 | 217.39 | NA |
| Pima | 0.67 | **0.71** | 0.49 | 0.65 | 0.17 | 0.11 | 0.28 | **0.06** | 1.14 | 4.98 |
| Breastw | **0.99** | 0.98 | 0.37 | 0.97 | 0.17 | 0.11 | 0.28 | **0.04** | 1.77 | 3.10 |
| Arrhythmia | **0.80** | 0.78 | 0.73 | 0.60 | 2.12 | 0.86 | 2.98 | **0.49** | 6.35 | 2.32 |
| Ionosphere | 0.85 | **0.92** | 0.89 | 0.85 | 0.33 | 0.15 | 0.48 | **0.04** | 0.64 | 0.83 |

performance

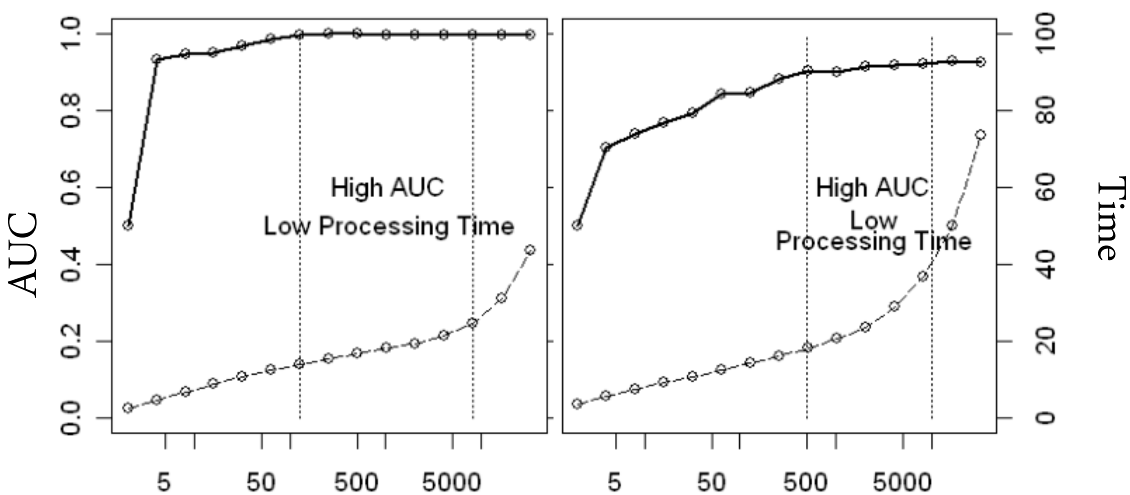
**Advantages**:
- normal instances only:
  iForest performs well, even without anomalies
- fast:
  linear time complexity and a low memory use

**Properties**:
- sub-sampling:
  iForest does not need profile of normal instances
- swamping:
  normal instances near anomalies → hard to separate
- masking:
  many anomalies close together → hard to separate



(a) Original sample
(4096 instances)

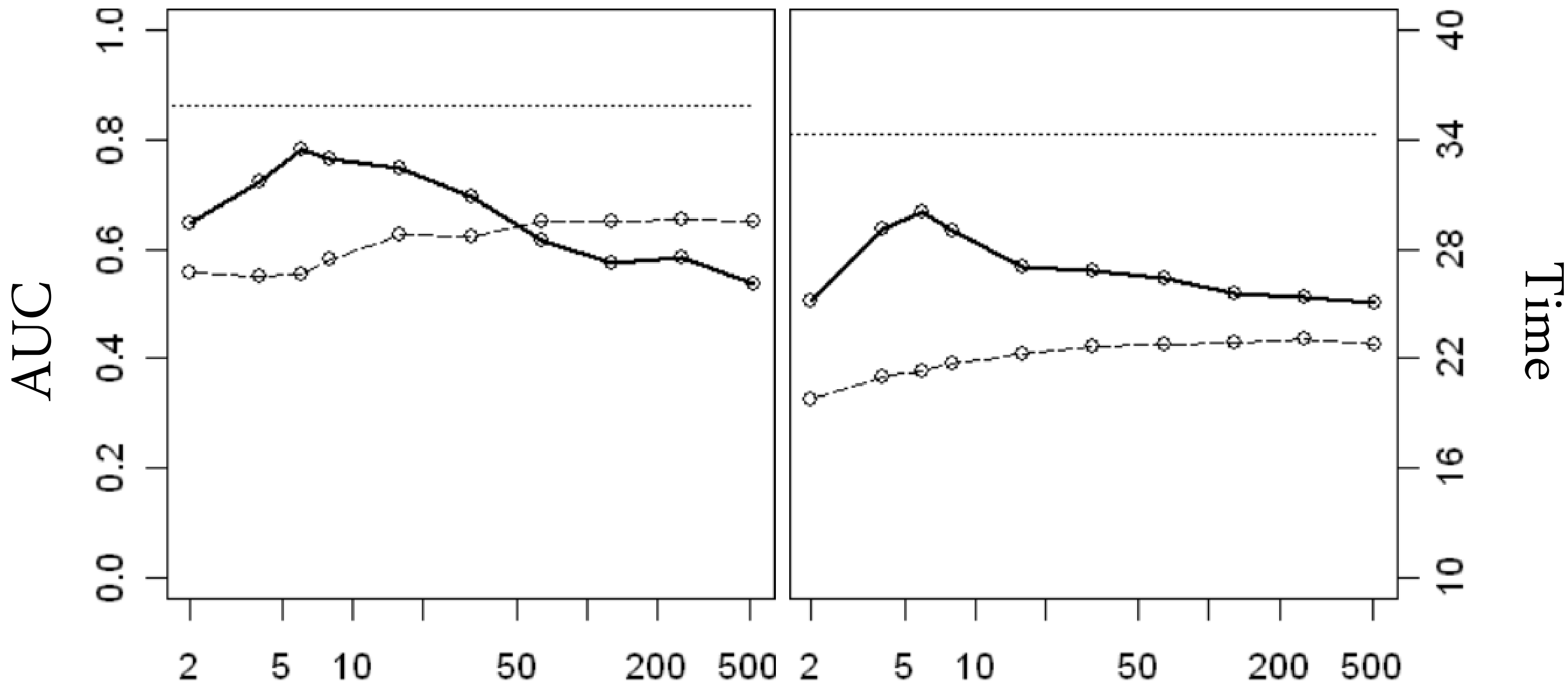
(b) Sub-sample
(128 instances)




(a) *Http* (b) *ForestCover*

**Advantages**:
- normal instances only:
  iForest performs well, even without anomalies
- fast:
  linear time complexity and a low memory use

**Properties**:
- sub-sampling:
  iForest does not need profile of normal instances
- swamping:
  normal instances near anomalies → hard to separate
- masking:
  many anomalies close together → hard to separate
- high-dim. data:
  iForest suffers from curse of dimensionality as well, but
  <u>feature selection</u> improves performance vastly




(a) *Mammography* (b) *Annthyroid*

Universität Regensburg

# Thank you