# DCML-CPS - Module 7

# Unsupervised ML

Tommaso Zoppi
University of Florence - Florence (Italy)
tommaso.zoppi@unifi.it

# Course Map

**Monitoring**
- 1. Basics of Metrology
- 2. Monitoring

**Testing**
- 3. Fault Injection
- 4. Robustness Testing

**Anomaly Detection**
- 5. Data Analysis
- 6. Supervised ML
- **7. Unsupervised ML**
- 8. Meta-Learning
- 9. Error/Intrusion Detection

# Unsupervised ML

# Unsupervised Algorithms

- Unsupervised Algorithms
  - They learn a normal behaviour
  - Without assuming any knowledge of anomalous events

**No labels needed for training!**

# Supervised vs Unsupervised

- Detection capabilities of unsupervised are equal when detecting both known (that appear in the training set) and unknown events
- Instead, supervised algorithms are very good in detecting known issues, but have essentially no means to detect unknowns

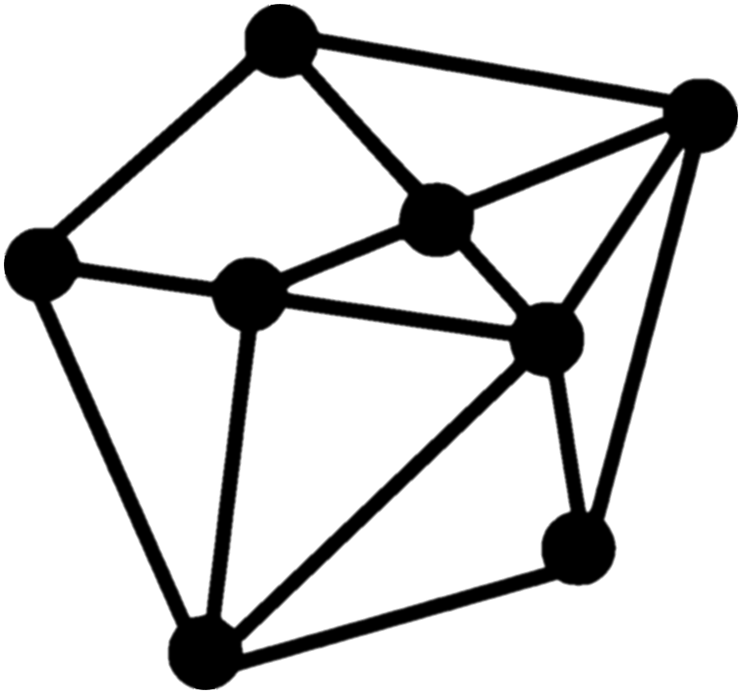
| | Known Issue | Unknown Issue |
|---|---|---|
| Supervised | Very Good! | Very Bad |
| Unsupervised | Average Good | |

# Unsupervised Families

- Families of Unsupervised Algorithms
  - Clustering
  - Density-Based
  - Angle-Based
  - Classification
  - Statistical
  - Neural Network

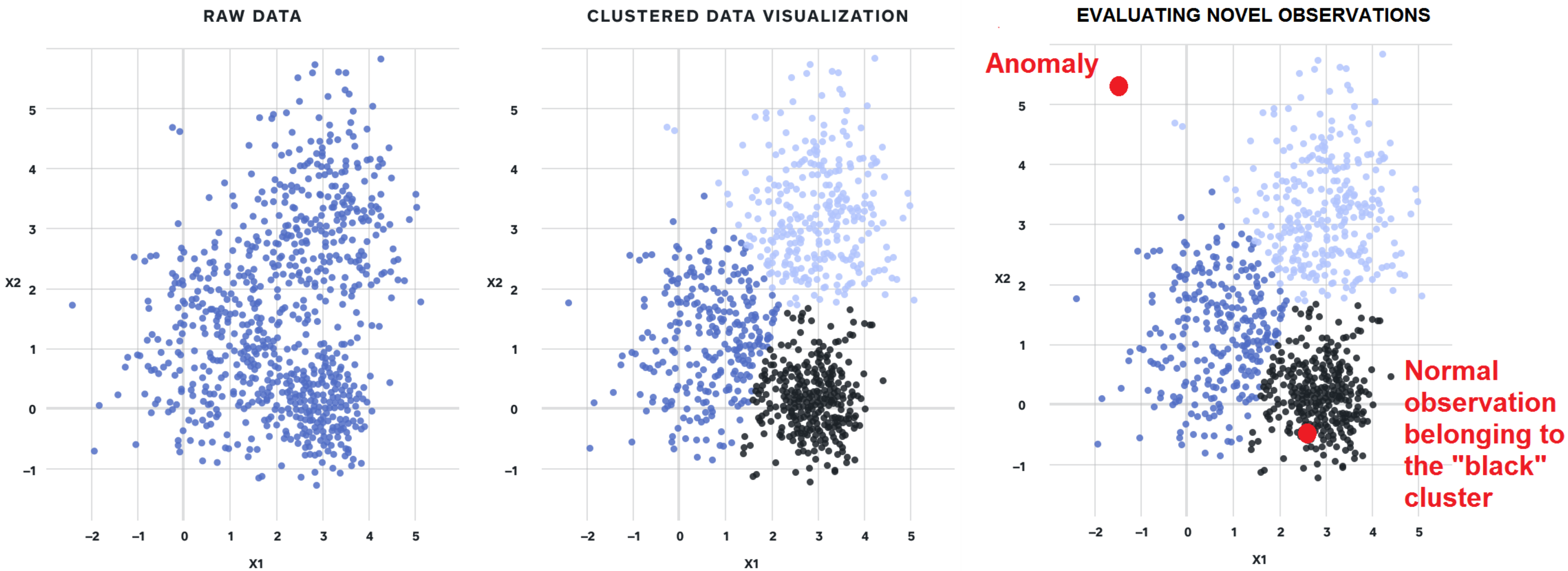
# Unsupervised Families

- Families of Unsupervised Algorithms
  - Clustering
  - Density-Based
  - Angle-Based
  - Classification
  - Statistical
  - Neural Network



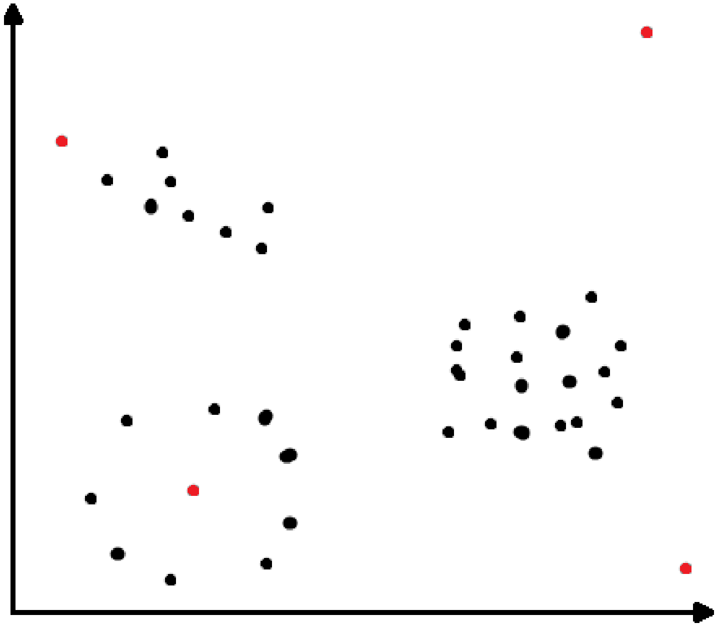
# Clustering

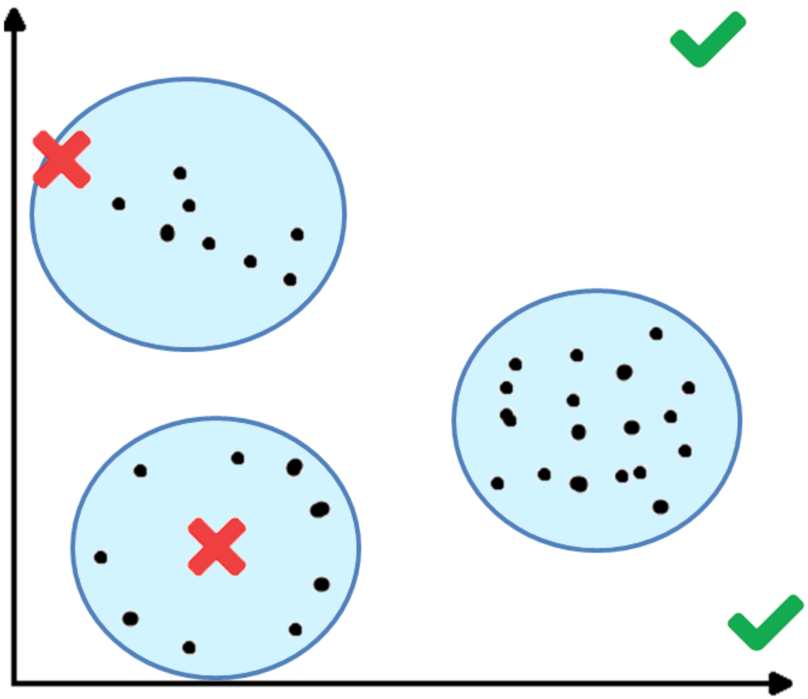
- Clustering algorithms
  - analyse the data flow, to
  - derive clusters which identify groups of similar data points
  - Data points that are far (do not "belong") from all the clusters are labeled as anomalies



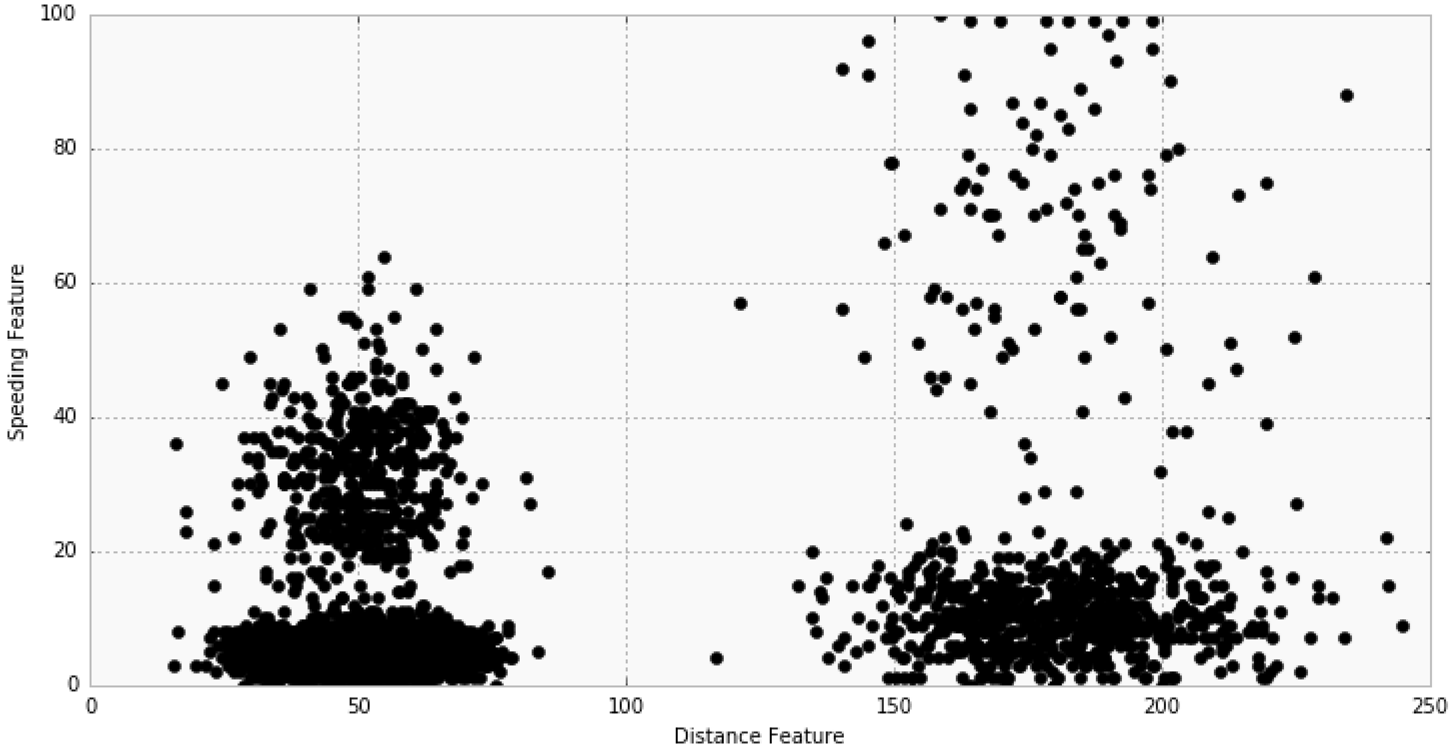
# Clustering: At a Glance

- On the left there is the dataset, red dots are anomalies
- On the right a graphical view of clustering
  - Crosses mean False negatives, ticks mean true positives





# Clustering: K-Means

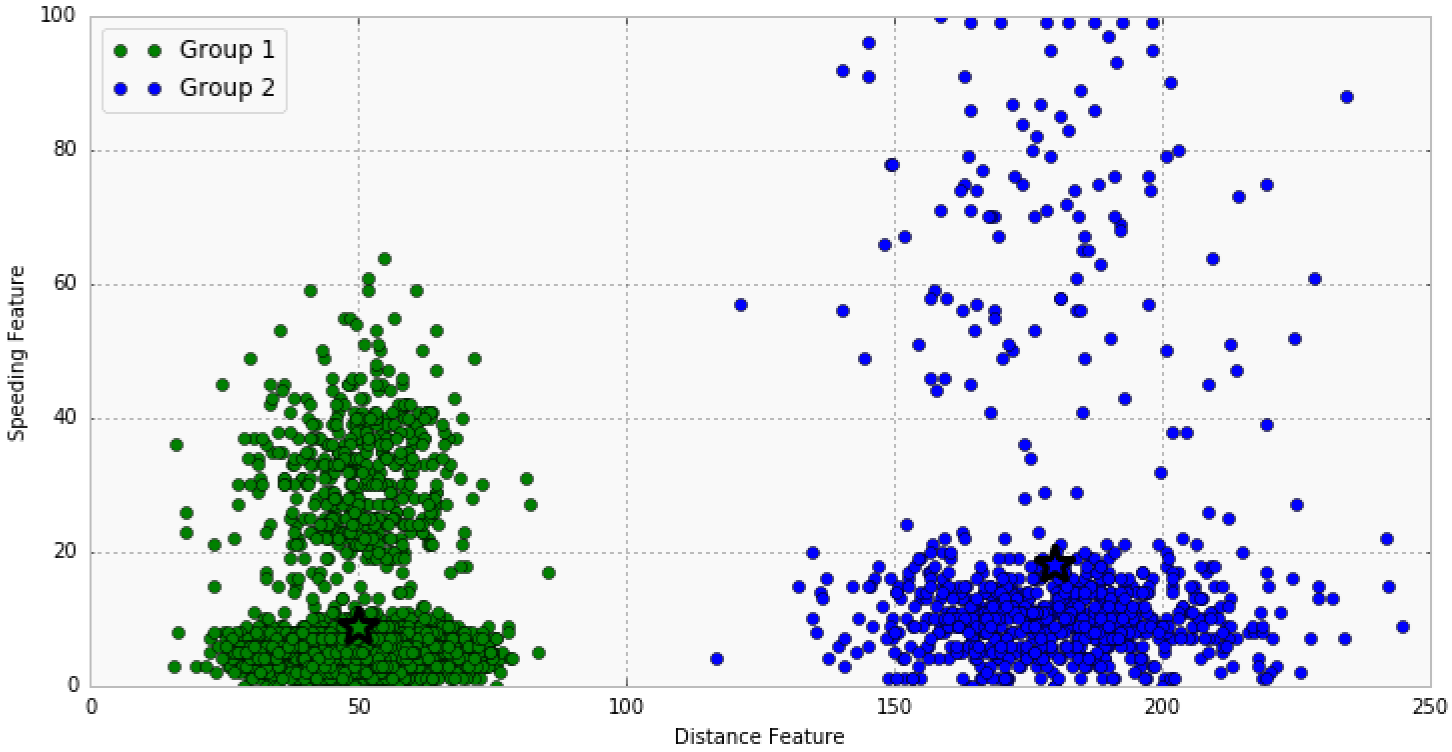



**How important is the number of clusters?**

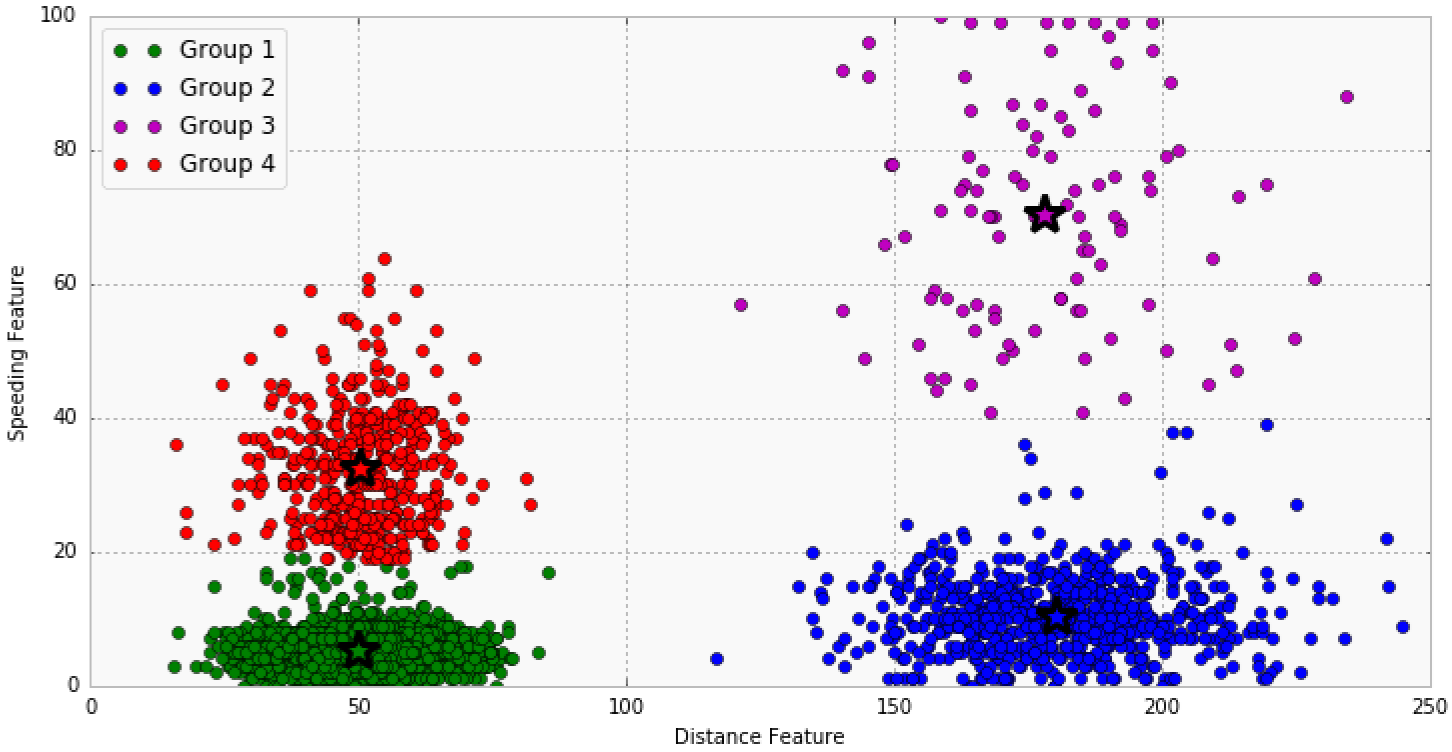
G-Means: Automatic tuning of K in K-Means

Hamerly, Greg, and Charles Elkan. "Learning the k in k-means." Advances in neural information processing systems. 2004.

# Clustering: K-Means

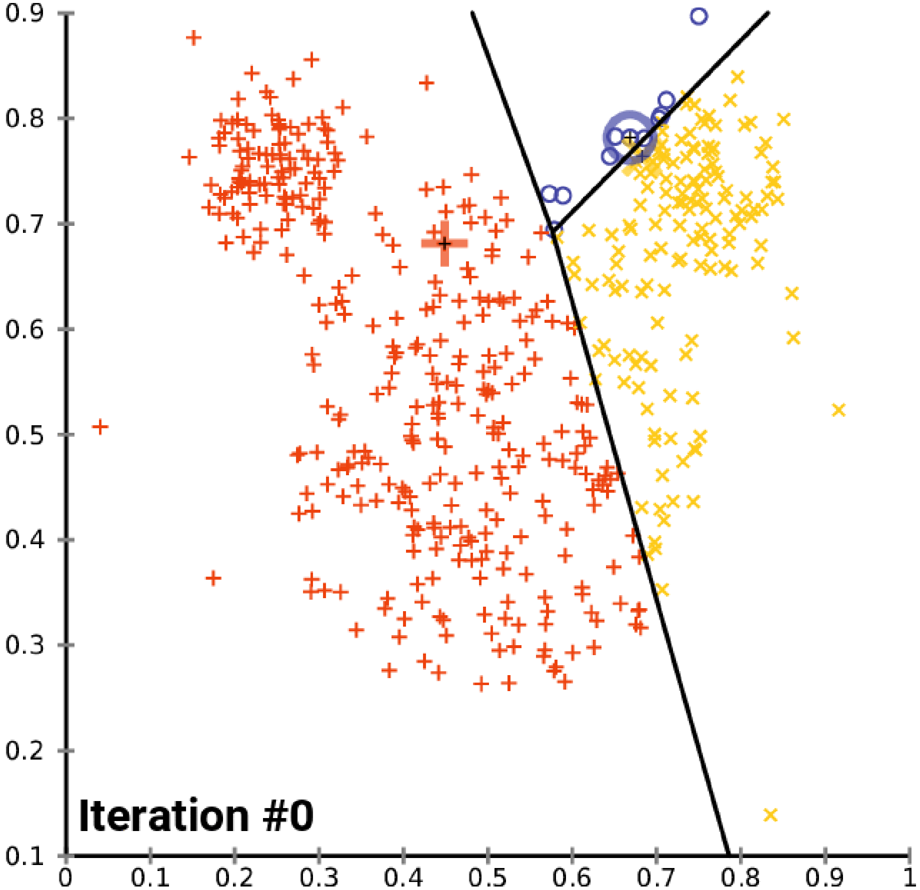


# Clustering: K-Means



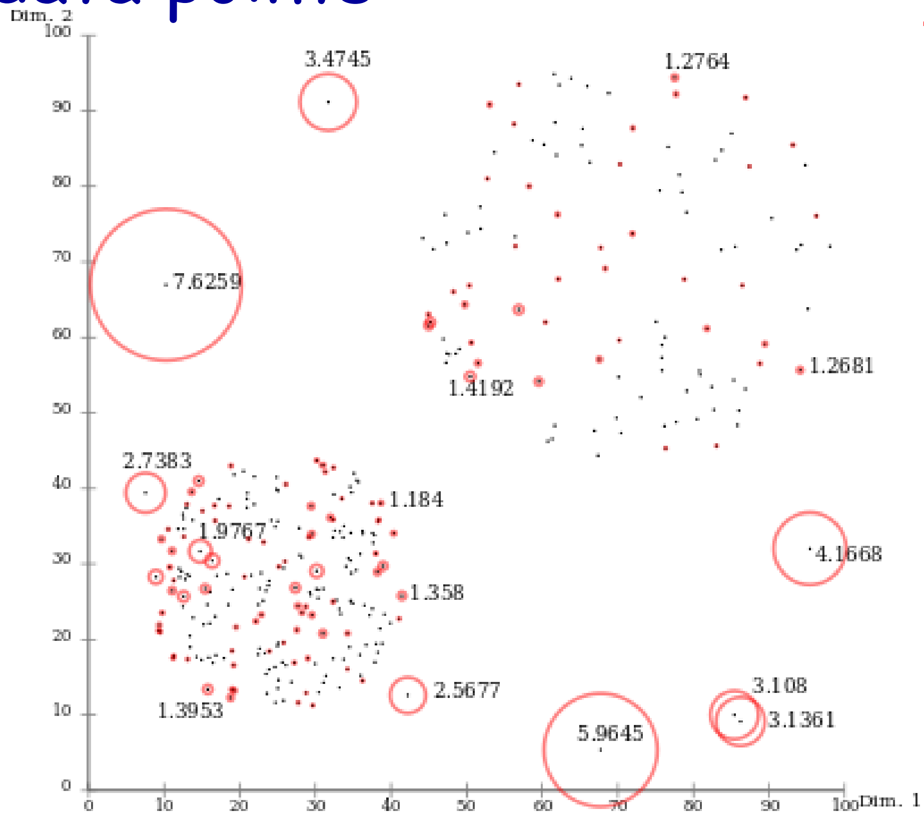
# Example: Training K-Means

- Example of convergence of K-Means (11 iterations)
  - From https://commons.wikimedia.org/wiki/File:K-means_convergence.gif



# Density-Based

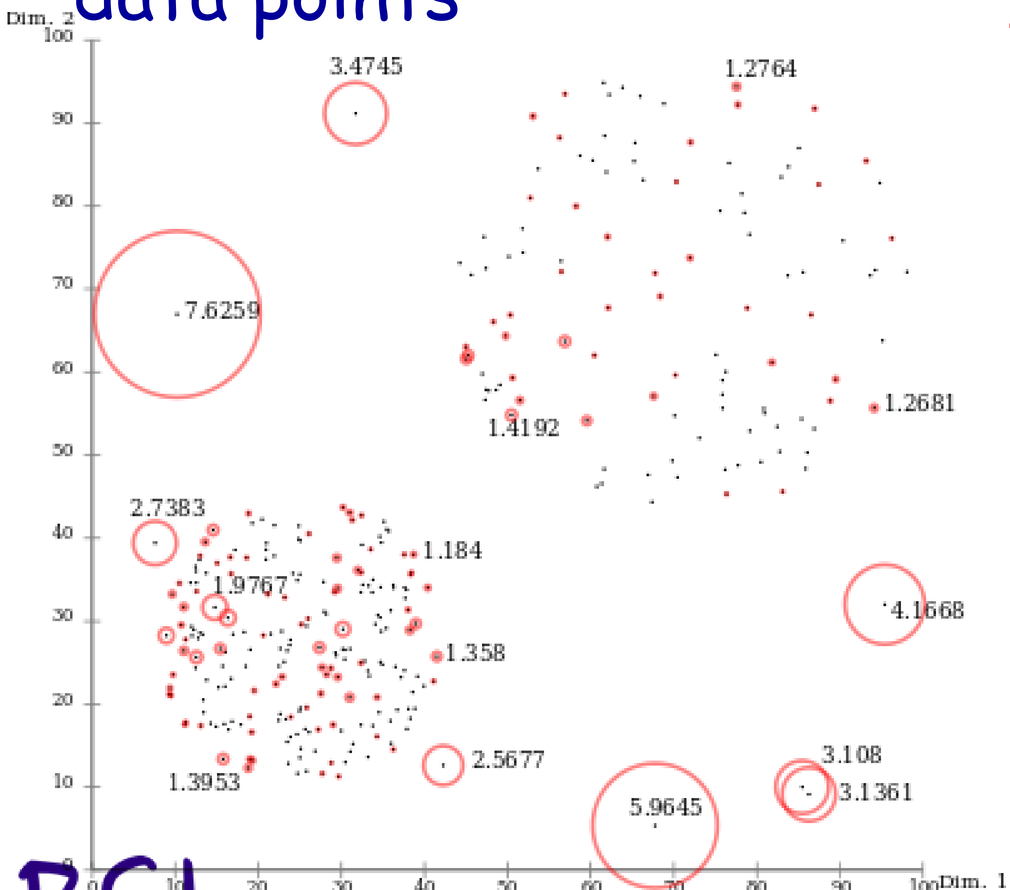
- Density-based algorithms label data points as anomalies if they are far from most of the other data points



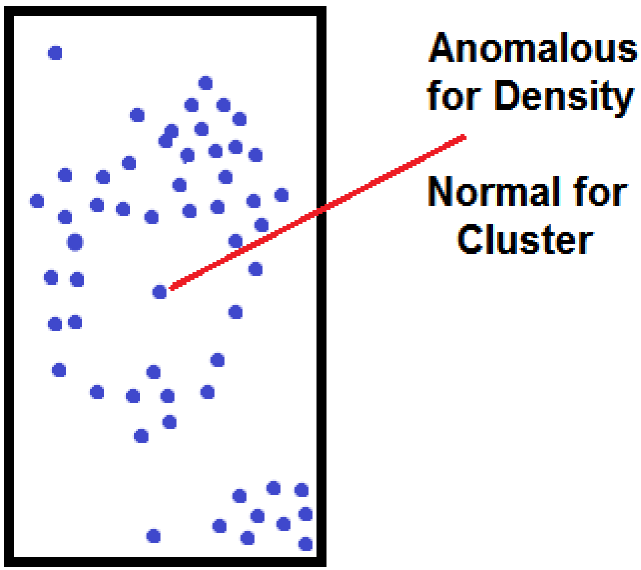

**–Why are they different from clustering algorithms?**

# Density-Based

- Density-based algorithms label data points as anomalies if they are far from most of the other data points

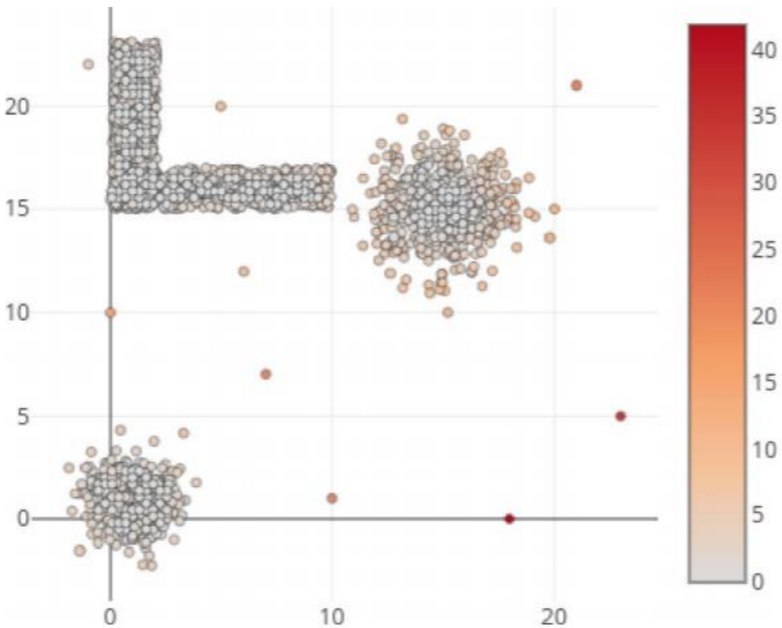


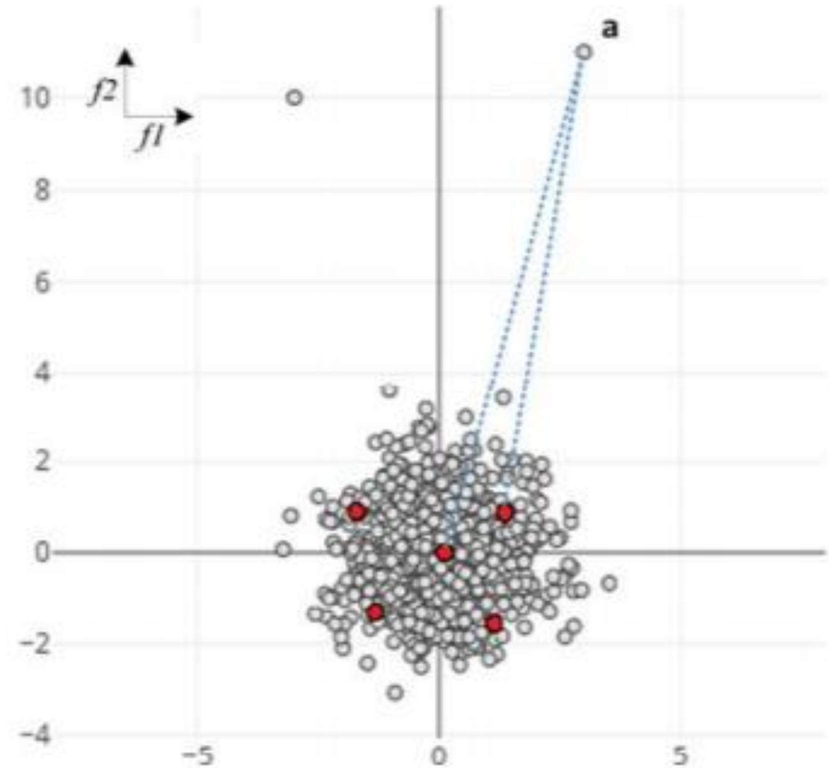

–**Why are they different from clustering algorithms?**



# Density: Sparse Data Observers

- Chooses some data points in training set as observers.
  - If a data point is far from observers, it is an anomaly



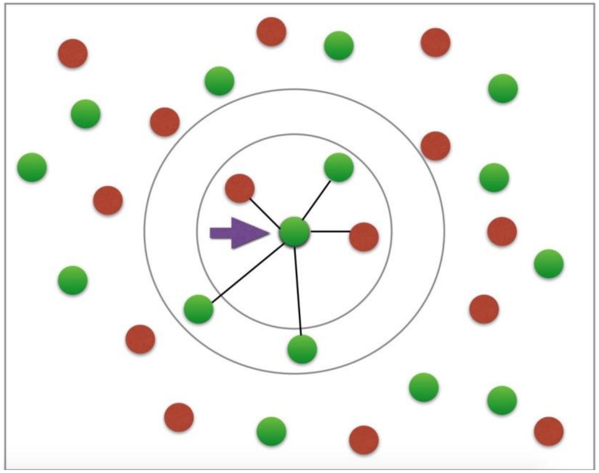

Vázquez, F. I., Zseby, T., & Zimek, A. (2018, November). Outlier detection based on low density models. In *2018 IEEE international conference on data mining workshops (ICDMW)* (pp. 970-979). IEEE.

# Density-Neighbour-Based

- These algorithms are based on the idea of neighbourhood
  - Which is naturally used in Supervised ML
  - However, there are algorithms which are based on the concept of neighbourhood to estimate density
    - Those go unsupervised (i.e., they do not need labeled training data)

# Density-Neighbour-Based: ODIN (I)

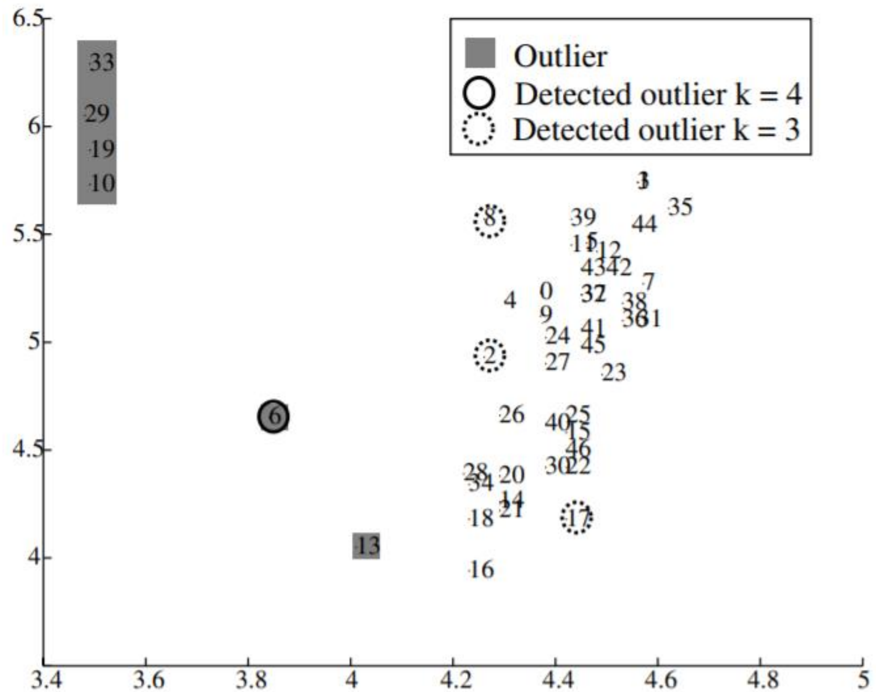
- Uses the kNN Graph
- A bi-directional graph where
  - nodes are data points, and
  - An edge exists from node A to node B if
    - B is one of the k-NN of A, or
    - A is one of the k-NN of B



Hautamaki, V., Karkkainen, I., & Franti, P. (2004, August). Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* (Vol. 3, pp. 430-433). IEEE.

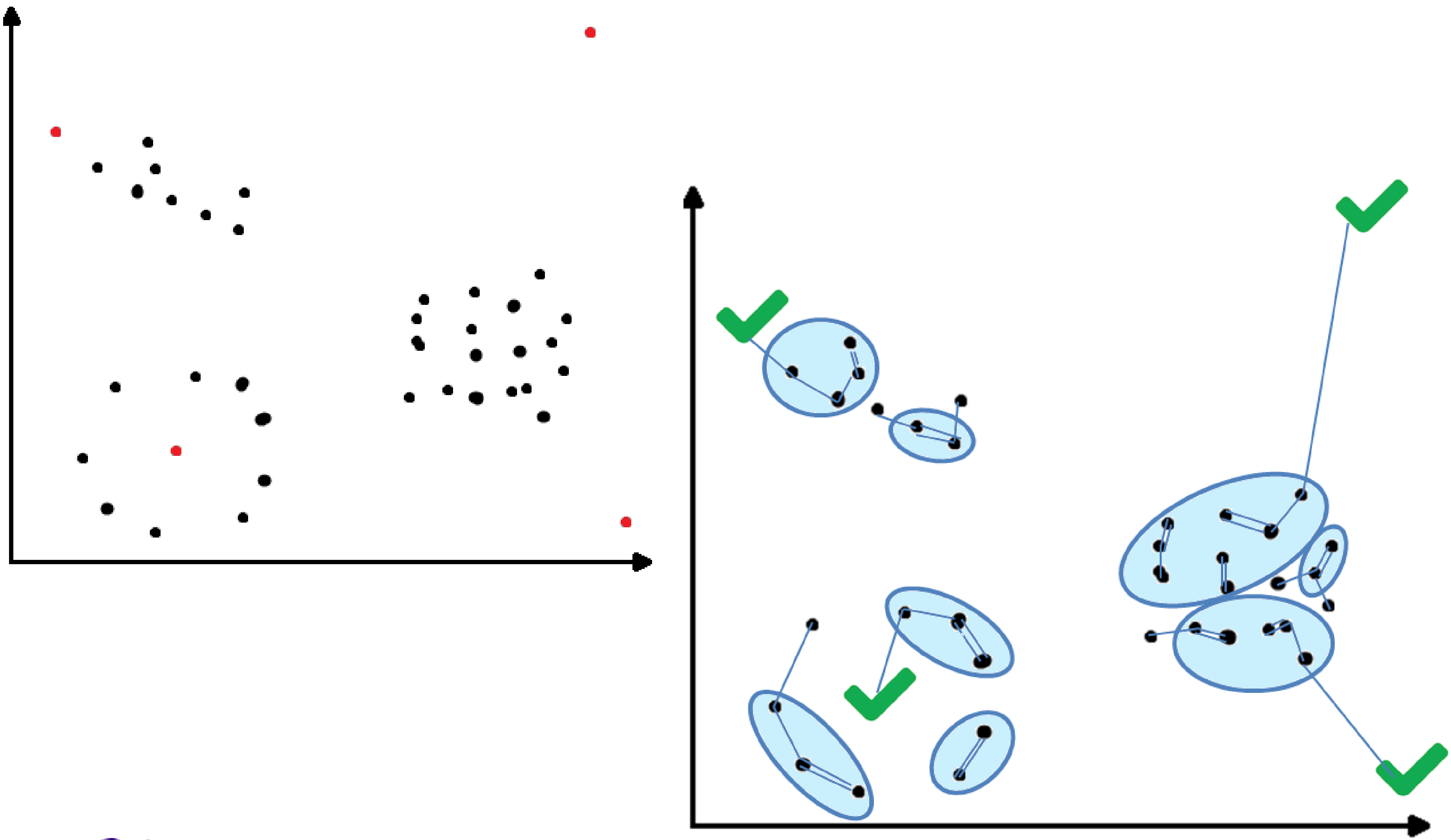
# Density-Neighbour-Based: ODIN (II)

- ODIN score is calculated as node indegree number
  - The higher the amount of connected nodes (indegree), the more a data point is normal (close to many others)



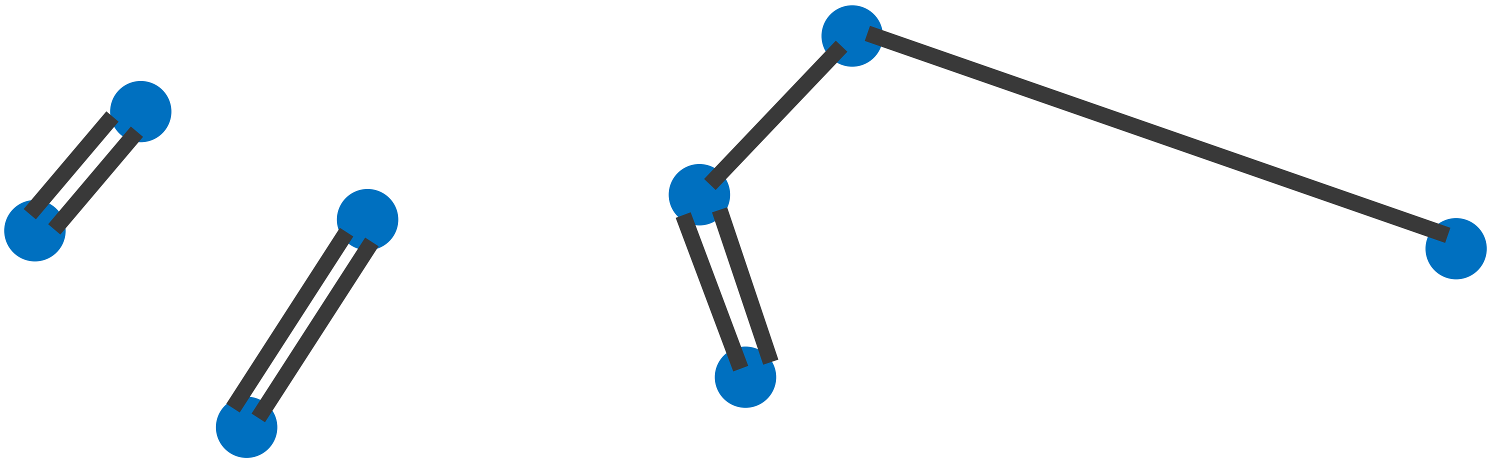

Hautamaki, V., Karkkainen, I., & Franti, P. (2004, August). Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* (Vol. 3, pp. 430-433). IEEE.

# ODIN: At a Glance



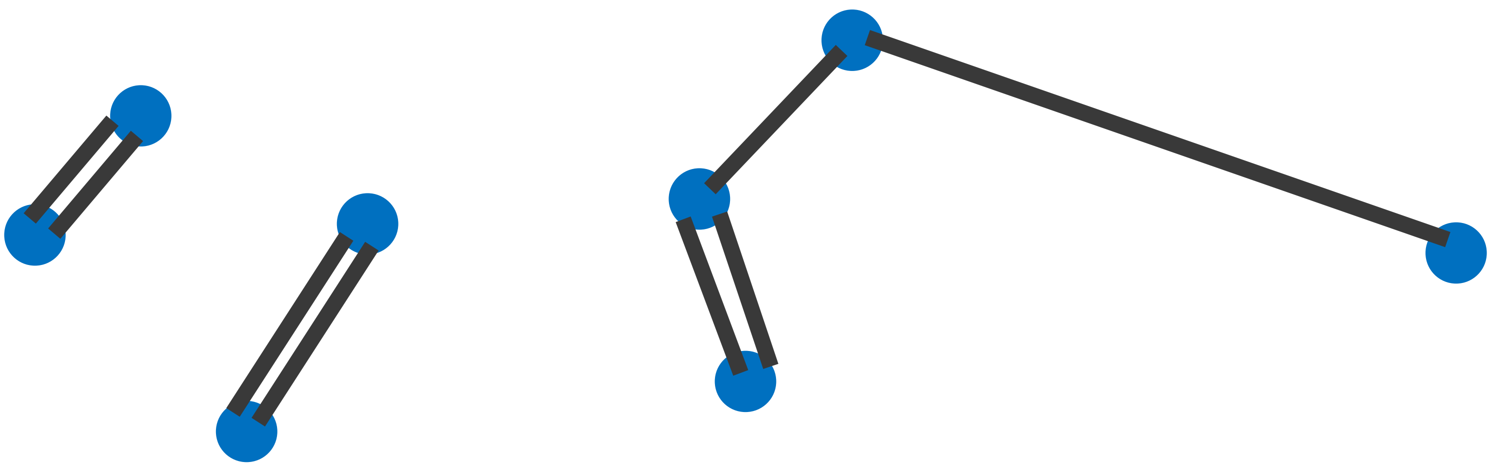
# ODIN Example

- ODIN with k=1
  - We connect each point with its closest neighbour



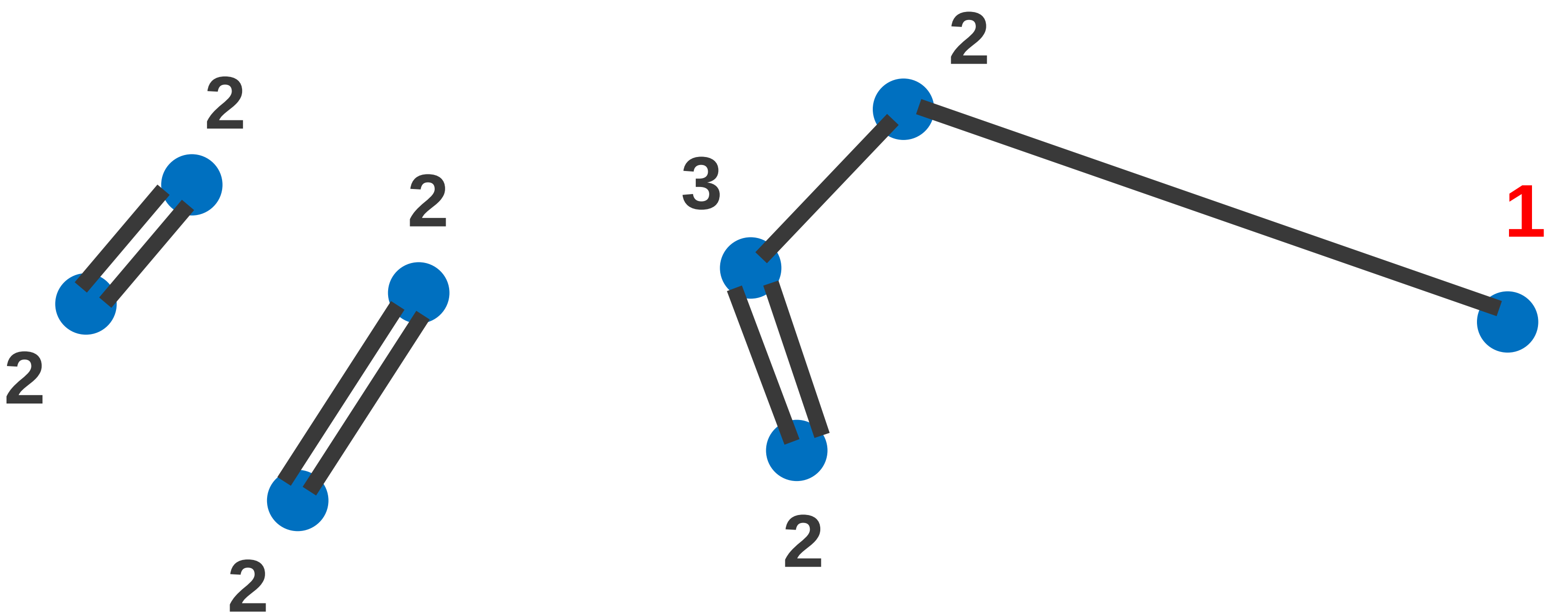
# ODIN Example

- ODIN with k=1
  - We connect each point with its closest neighbour
  - Then we calculate the indegree number
  - The lower, the more anomalous a point is



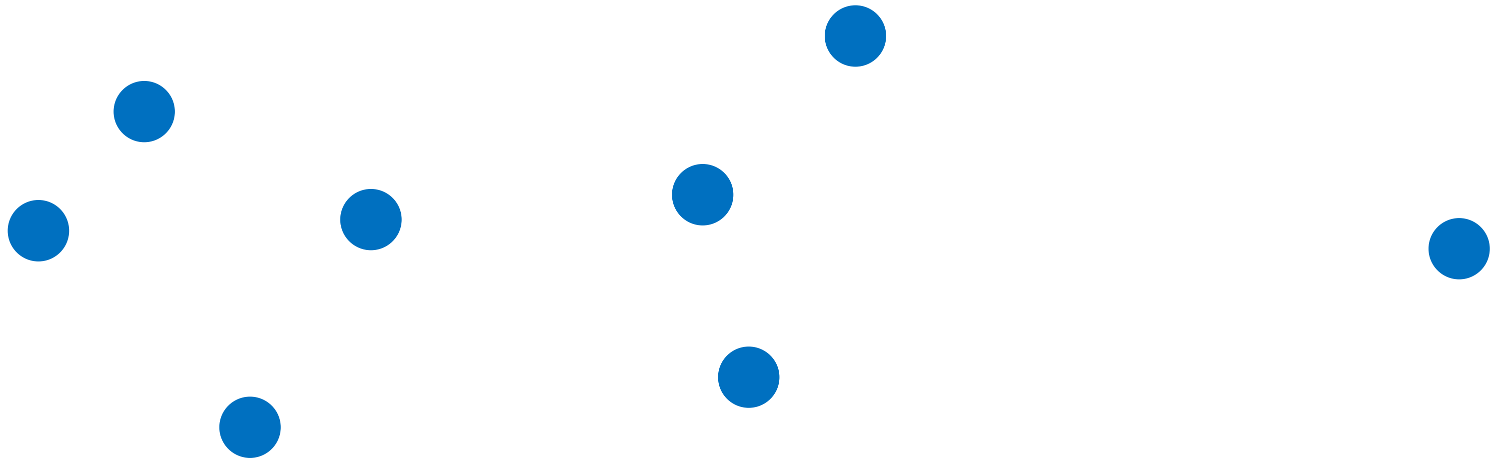
# ODIN Example

- ODIN with k=1
  - We connect each point with its closest neighbour
  - Then we calculate the indegree number
  - The lower, the more anomalous a point is



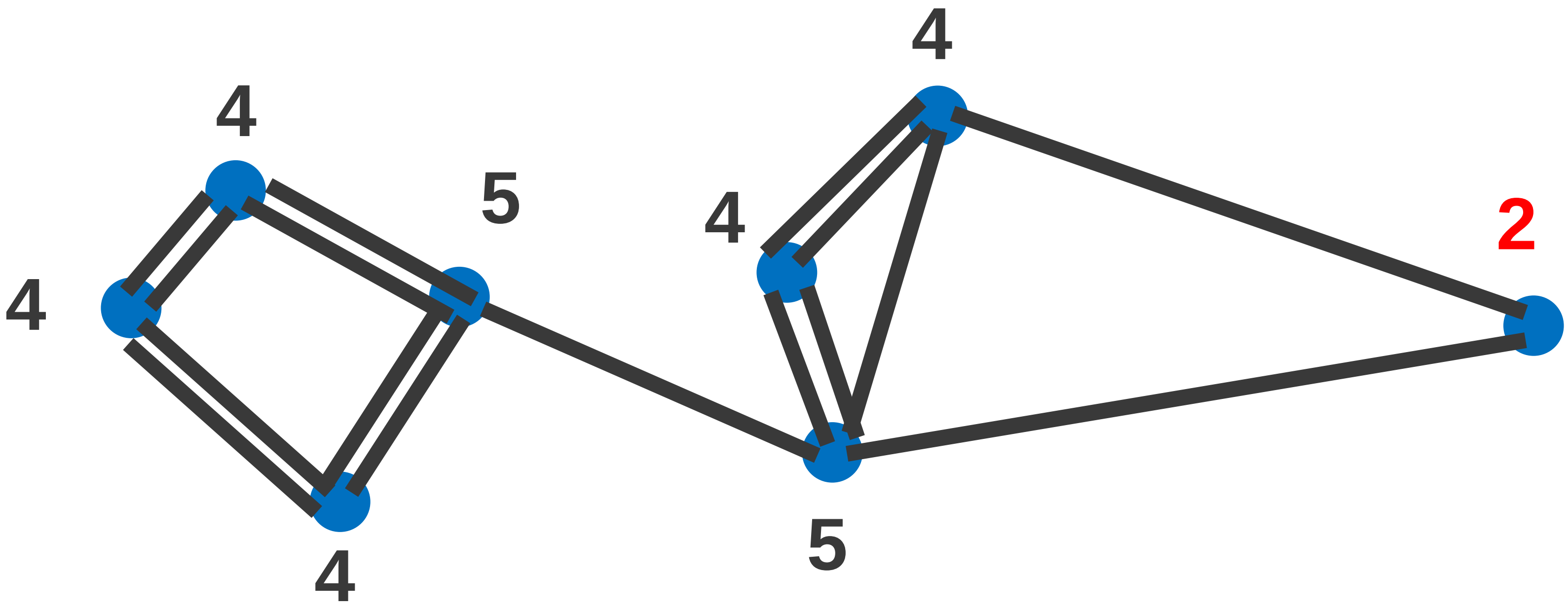
# ODIN Example

- ODIN with k=2
  - We connect each point with its two closer neighbours
  - Think about it…



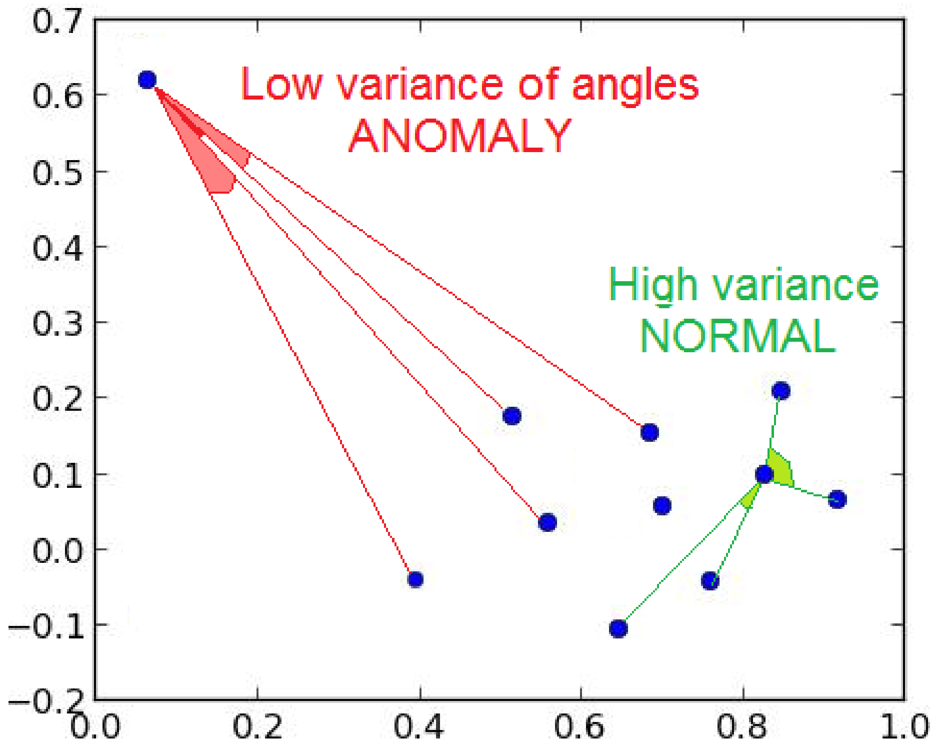
# ODIN Example

- ODIN with k=2
  - We connect each point with its two closer neighbours
  - Same result as with k=1!



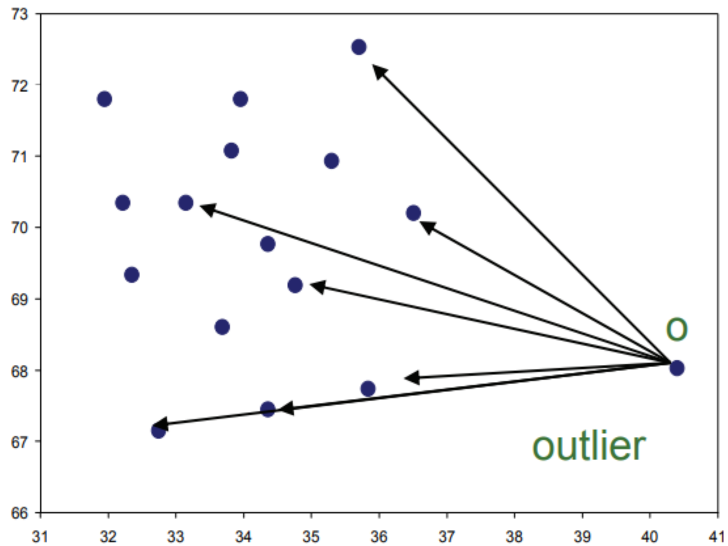
# Angle-Based

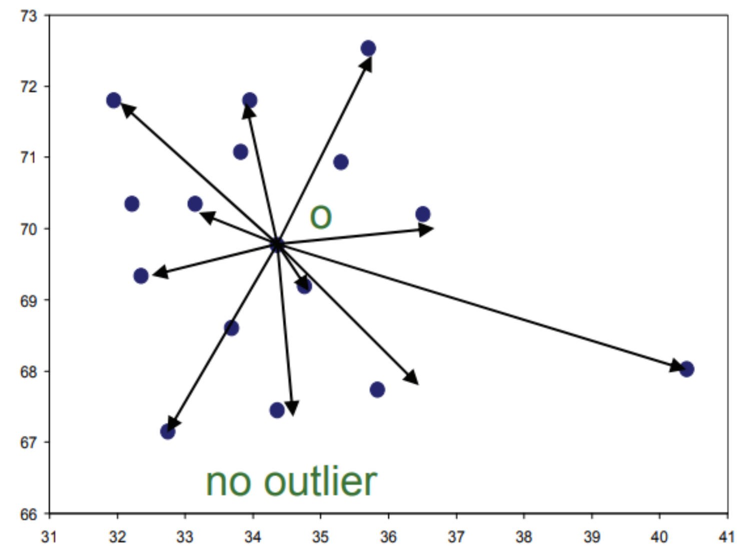
- Angle-Based algorithms identify anomalies as data points that have low variance of angles
  - Angles with respect to all the possible couples of data points



# Angle-Based: ABOD

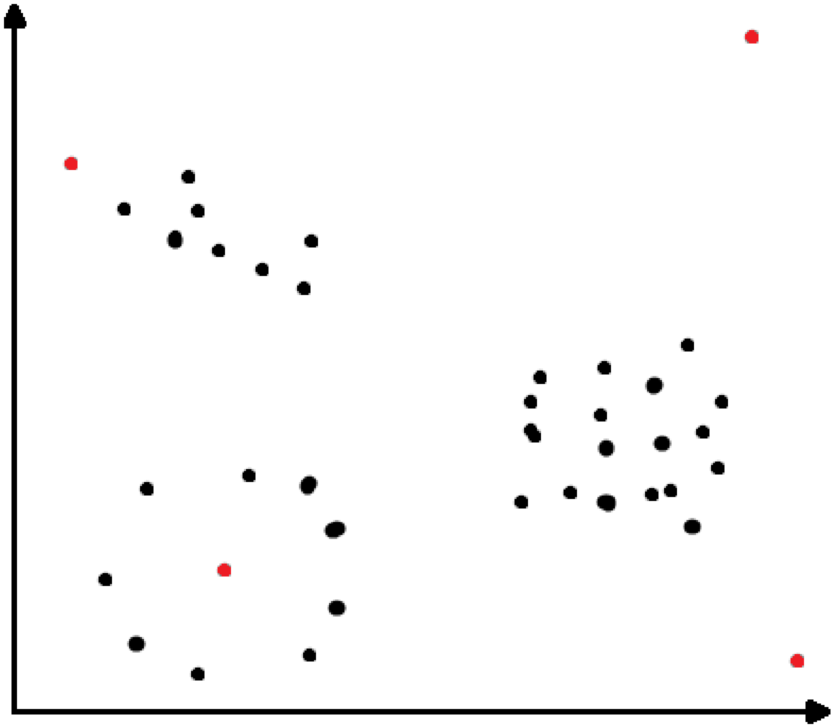
- ABOD calculates the angles of all the couples in the training set, considering the new point as vertex
  - Variance of angles is the Angle-Based Outlier Factor (ABOF)
  - Th

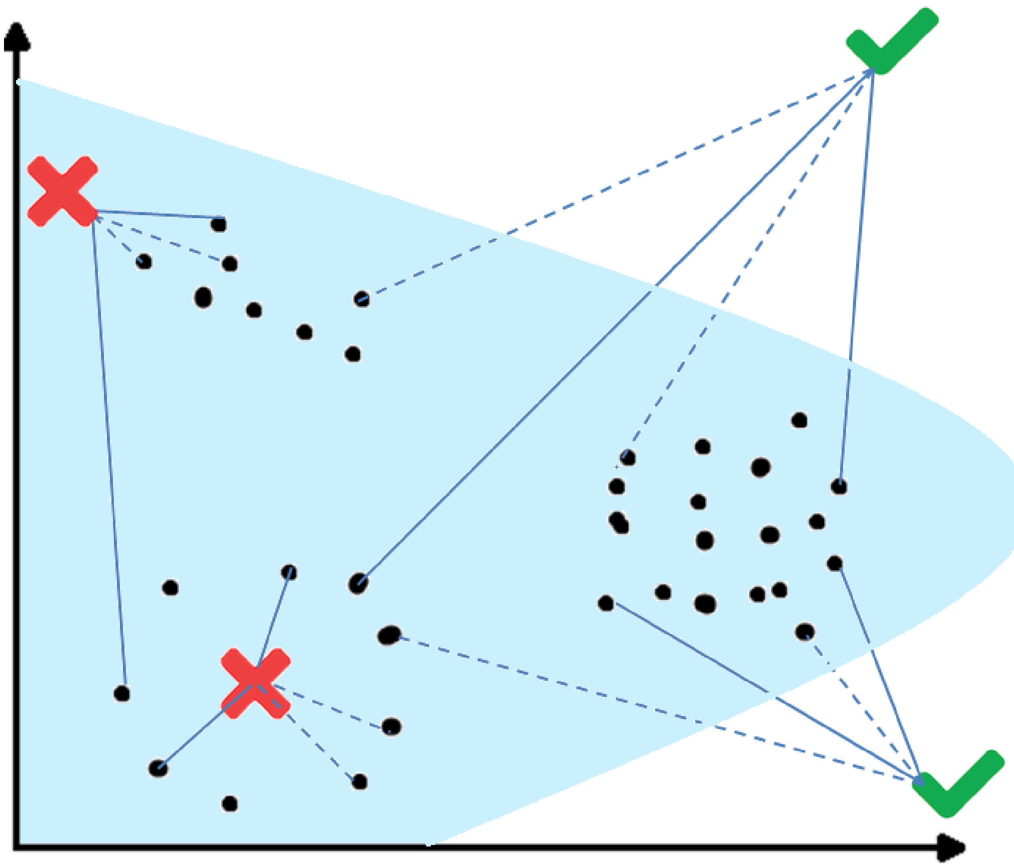



Kriegel, H. P., Schubert, M., & Zimek, A. (2008, August). Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 444-452).

# ABOD: At a Glance





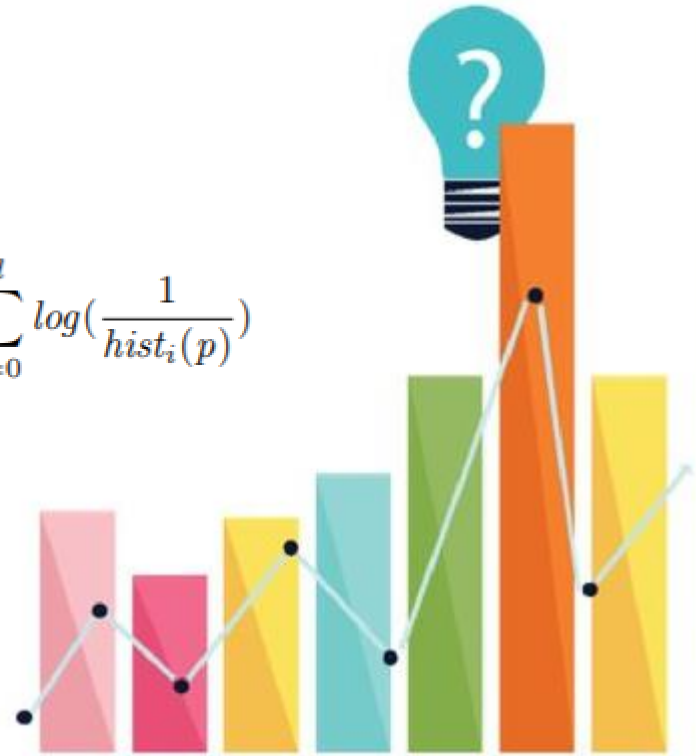
# Statistical

- They use statistical techniques to
  - extract "expected" probability distributions
  - check if data points are compliant with such distribution

# Statistical Algorithm: HBOS

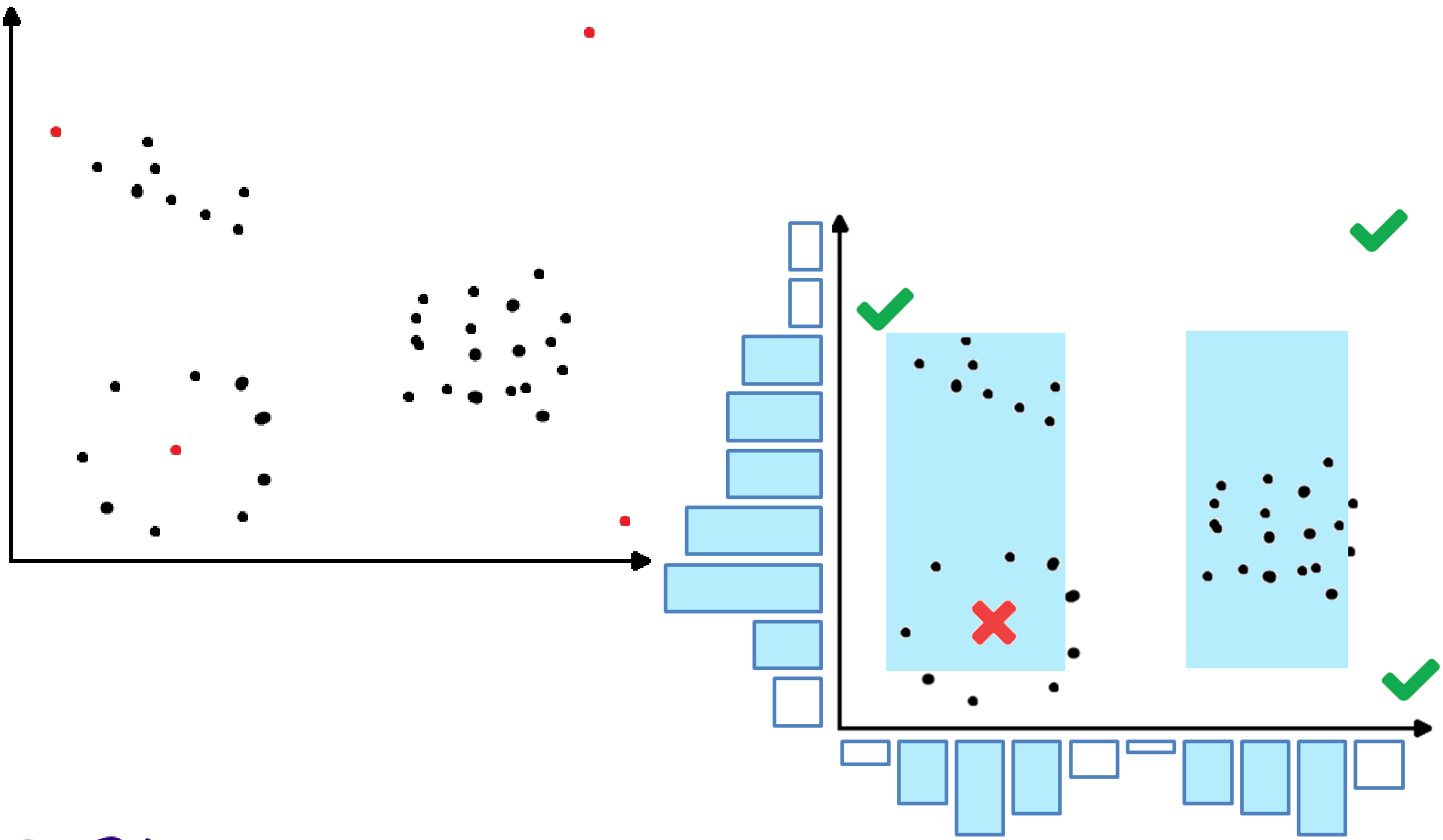
- Example: using histograms
  - For each indicator, we derive expected frequencies of values through histograms
  - Short columns = low frequencies
  - If indicators values of a data point the data point is anomalous

$$HBOS(p) = \sum_{i=0}^{d} log(\frac{1}{hist_i(p)})$$



Goldstein, Markus, and Andreas Dengel. "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm." *KI-2012: Poster and Demo Track* (2012): 59-63.

# HBOS: At a Glance



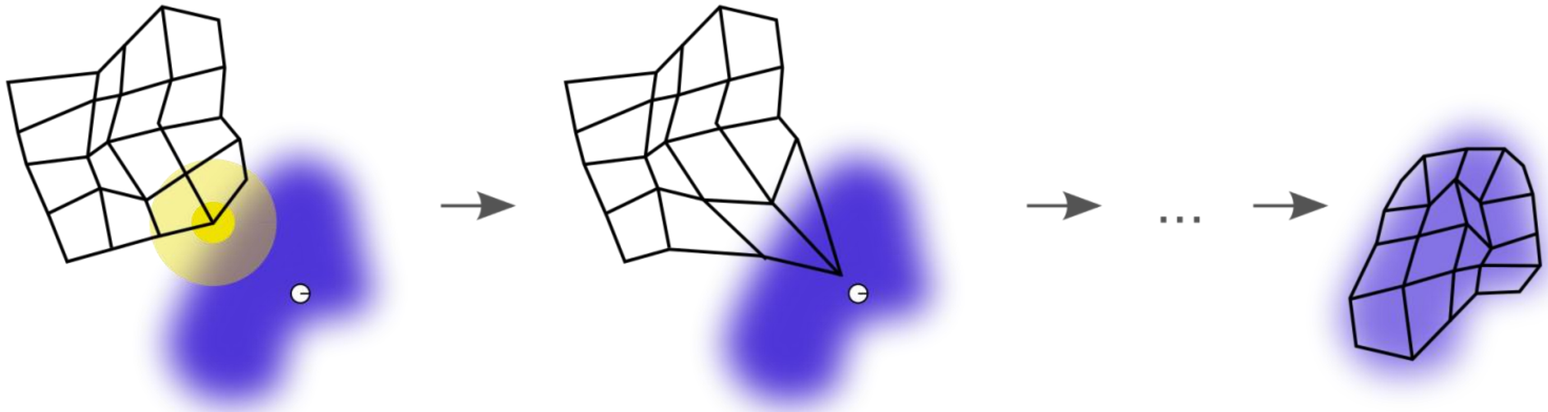
# Neural Nets (At a Glance)

- A neural network maps the human brain as a circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes.
  - The connections between neurons are modeled as weights.
  - A positive weight reflects an excitatory connection, while negative values mean inhibitory connections.
  - All inputs are modified by a weight and summed to deliver the final result.
- Mostly Supervised, but there are Unsupervised variants.

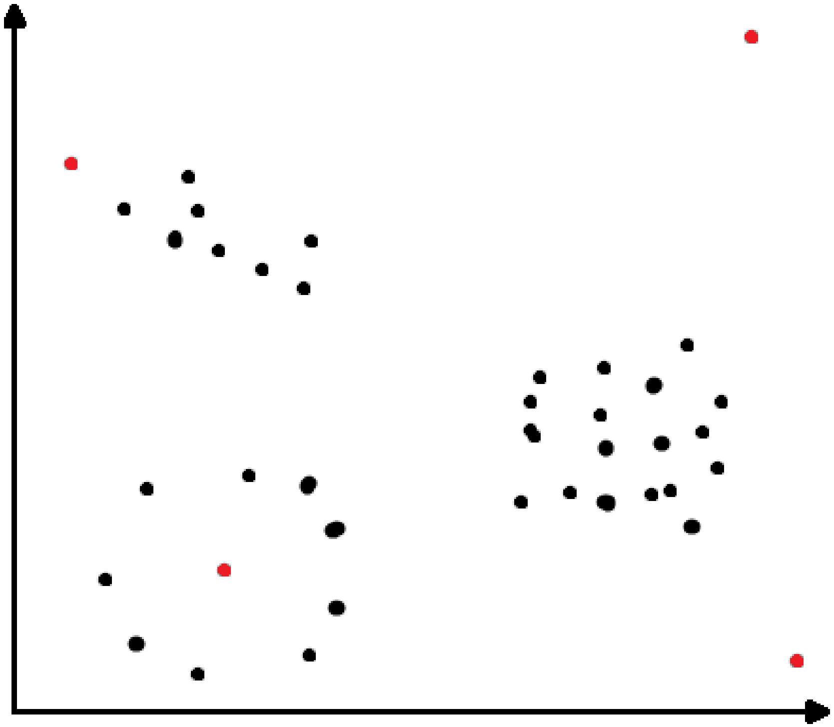
**YOU ALREADY HEARD ABOUT THAT!**

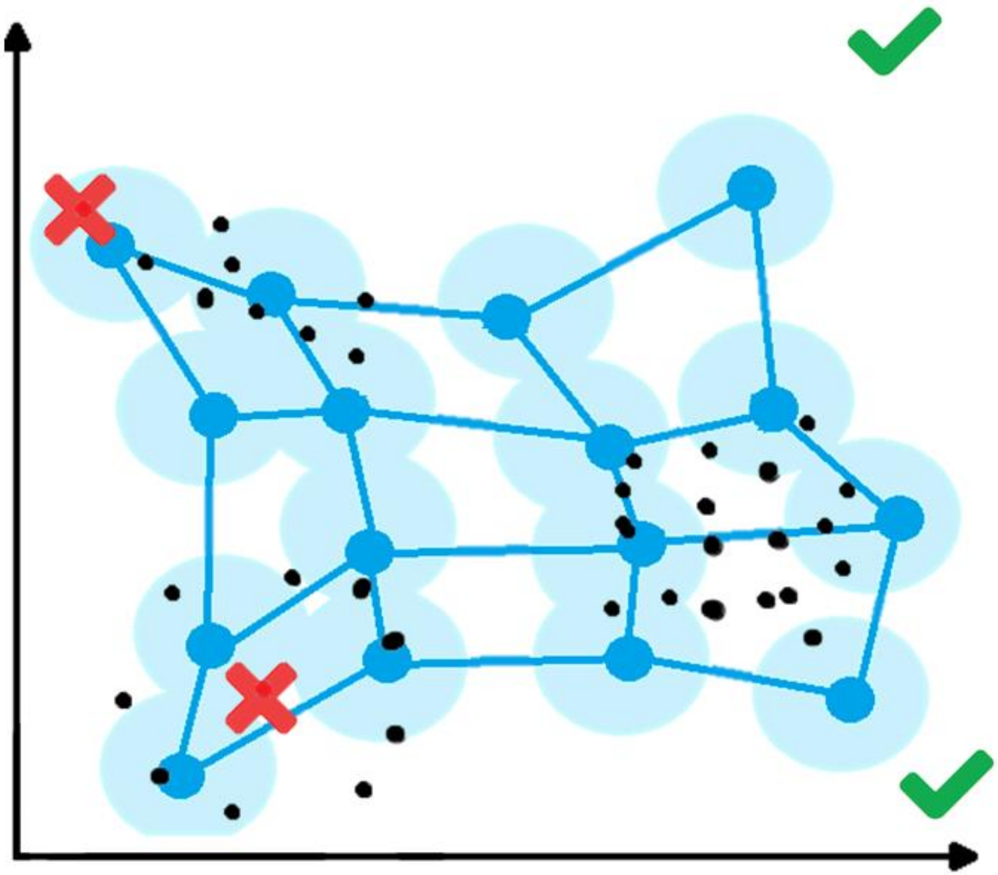
# Neural Nets: Self-Organizing Maps

- (From Wiki): Training of a Self-Organizing Map (SOM).
  - The blue blob is the distribution of the training data, and the small white disc is the current training datum drawn from that distribution.
  - At first (left) the SOM nodes are arbitrarily positioned in the data space.
  - The node (highlighted in yellow) which is nearest to the training datum is selected. It is moved towards the training datum, as (to a lesser extent) are its neighbors on the grid.
  - After many iterations the grid tends to approximate the data distribution (right).



# SOM: At a Glance





# Beware!

- The usage of deep learners for classifying tabular data was recently targeted by this study

**Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84-90**

- In which authors show that for processing tabular data, which usually does not have that many features, neural networks may not be the preferred choice
  - As it happens with unstructured data