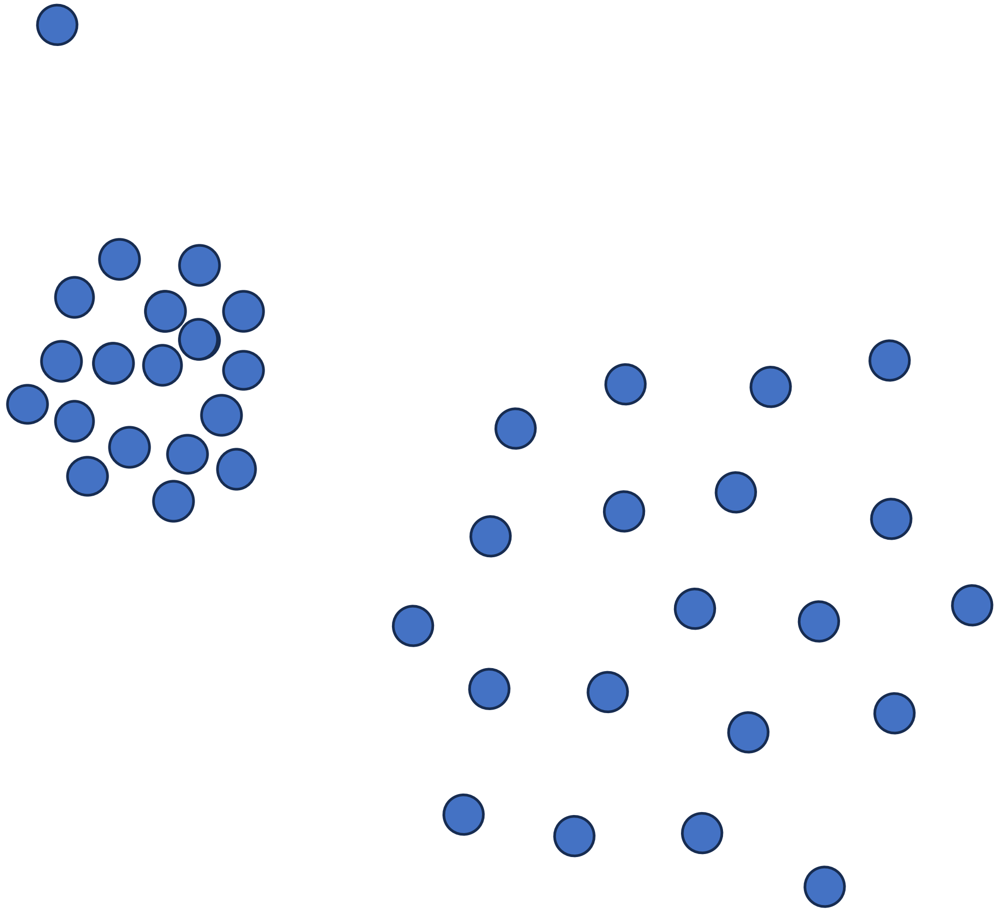
Data Science & Machine Learning Lab

Anomaly Detection

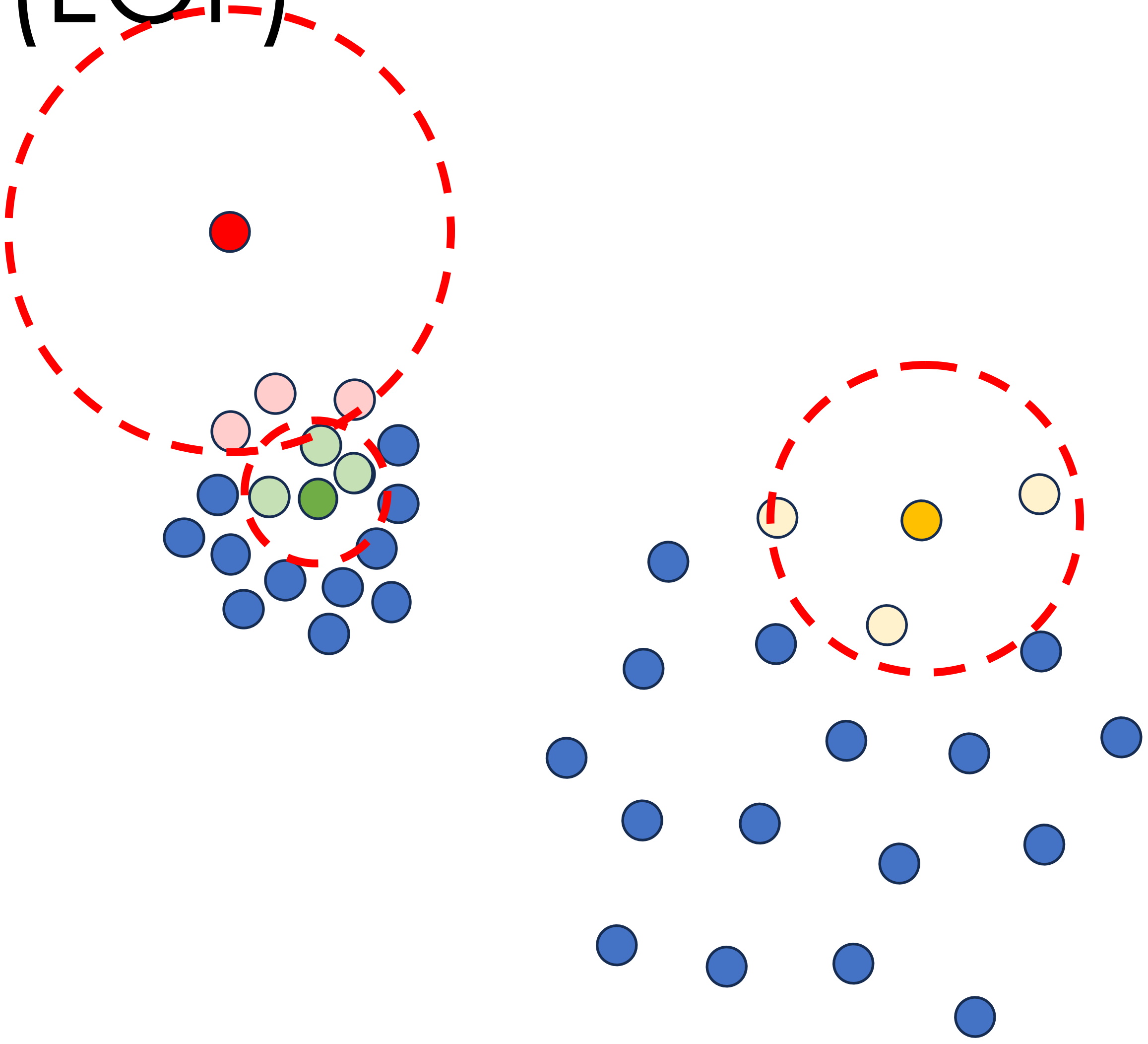
Flavio Giobergia

Local Outlier Factor (LOF)

- LOF considers points to be anomalous if they are found in non-dense regions of space.
- Since different regions of space can have different densities, LOF estimates the ”local” density of each region of space
- If a point has a smaller density than its (k) nearest neighbors, it is considered an anomaly.



Local Outlier Factor (LOF)

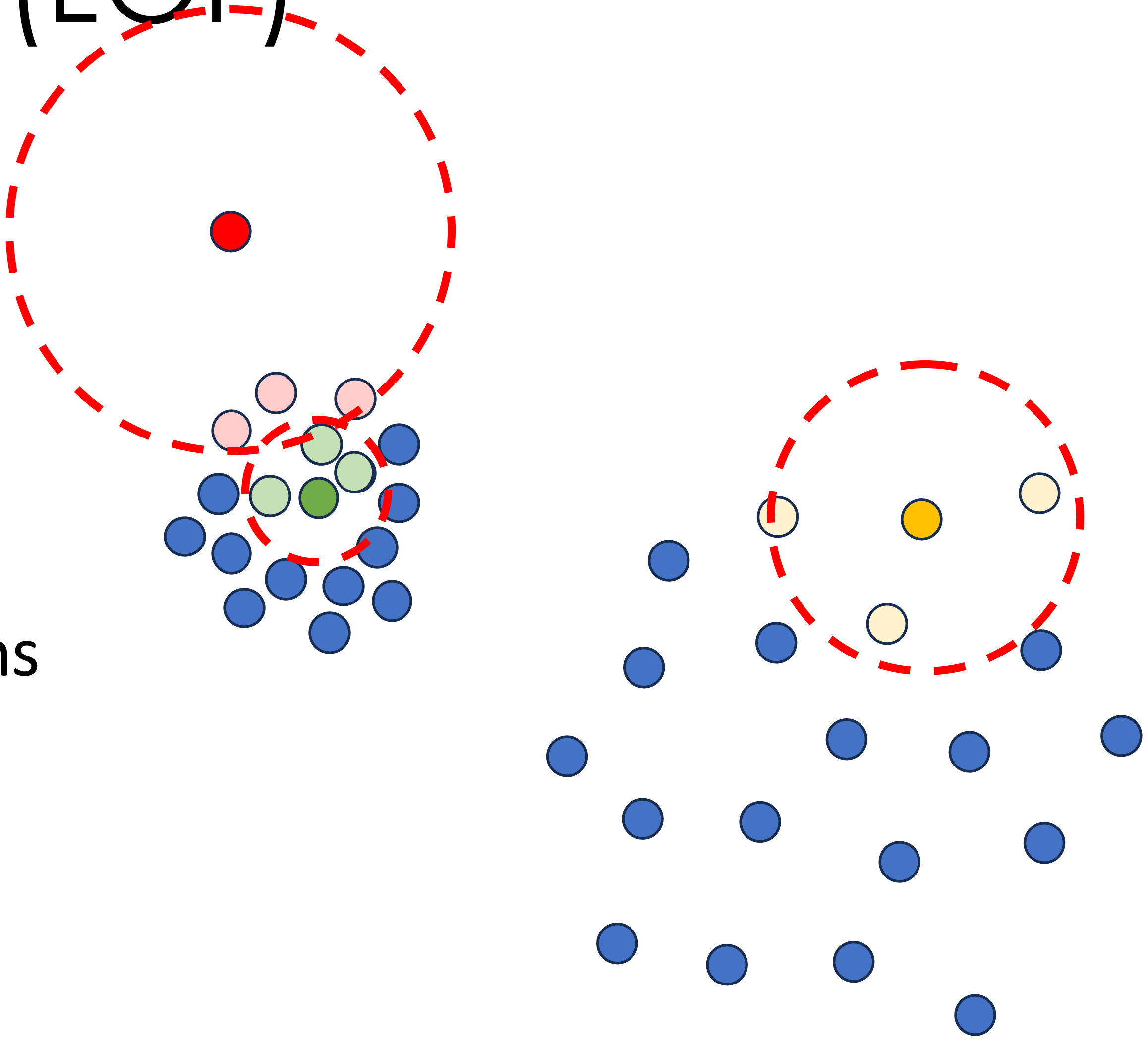


- For a given point, we compute its “local density”, which is inversely proportional to the distance to its k-th nearest neighbor*
- Large distance: neighbors are far away
  ➔ low-density region
- Small distance: neighbors are close by
  ➔ high-density region

* **Note**: this is a slight simplification w.r.t. actual LOF – where the “**Local Reachability Distance**” (LRD) is computed as an estimate of the local density

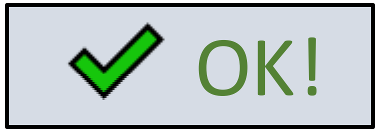
Local Outlier Factor (LOF)

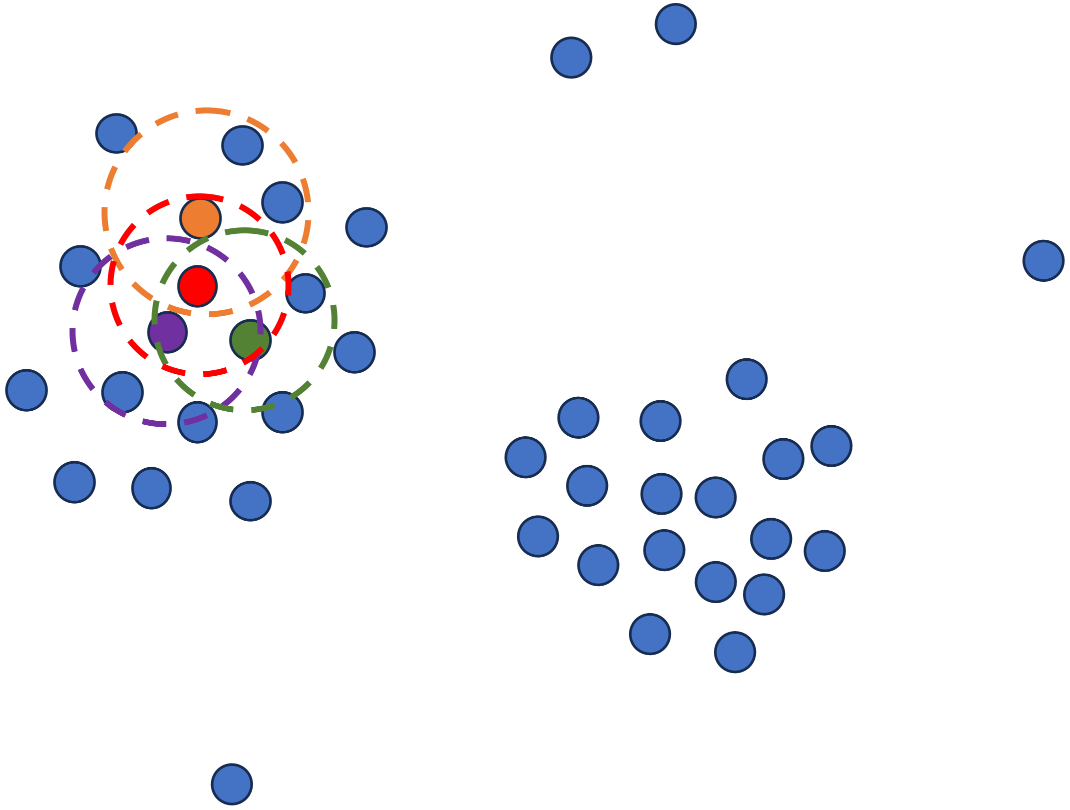
- If all high-density regions of space had the same density, we could simply define a threshold distance τ
- All points having k-th NN more distant than τ are anomalies
- However, the space may have regions of space of varying distance
- So, we need to "estimate" a τ separately for each point



Local Outlier Factor (LOF)

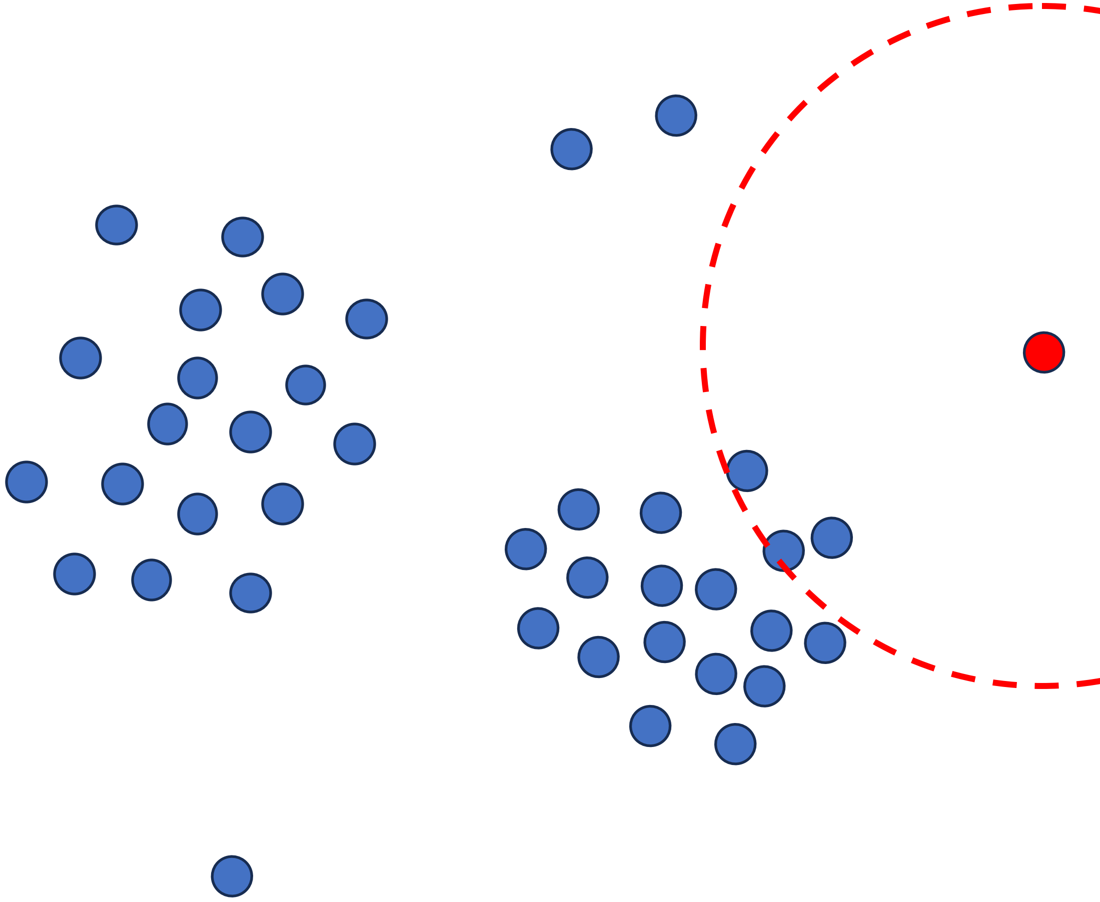
- We estimate the “local density” of a sample with the average distance of the k-th NN of the neighbors of



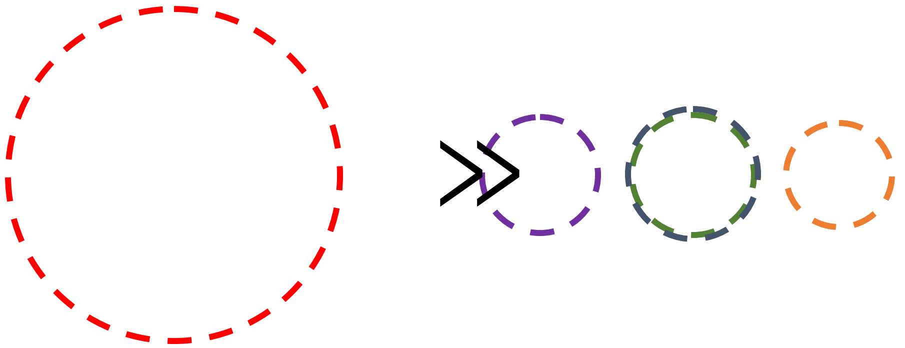


Local Outlier Factor (LOF)

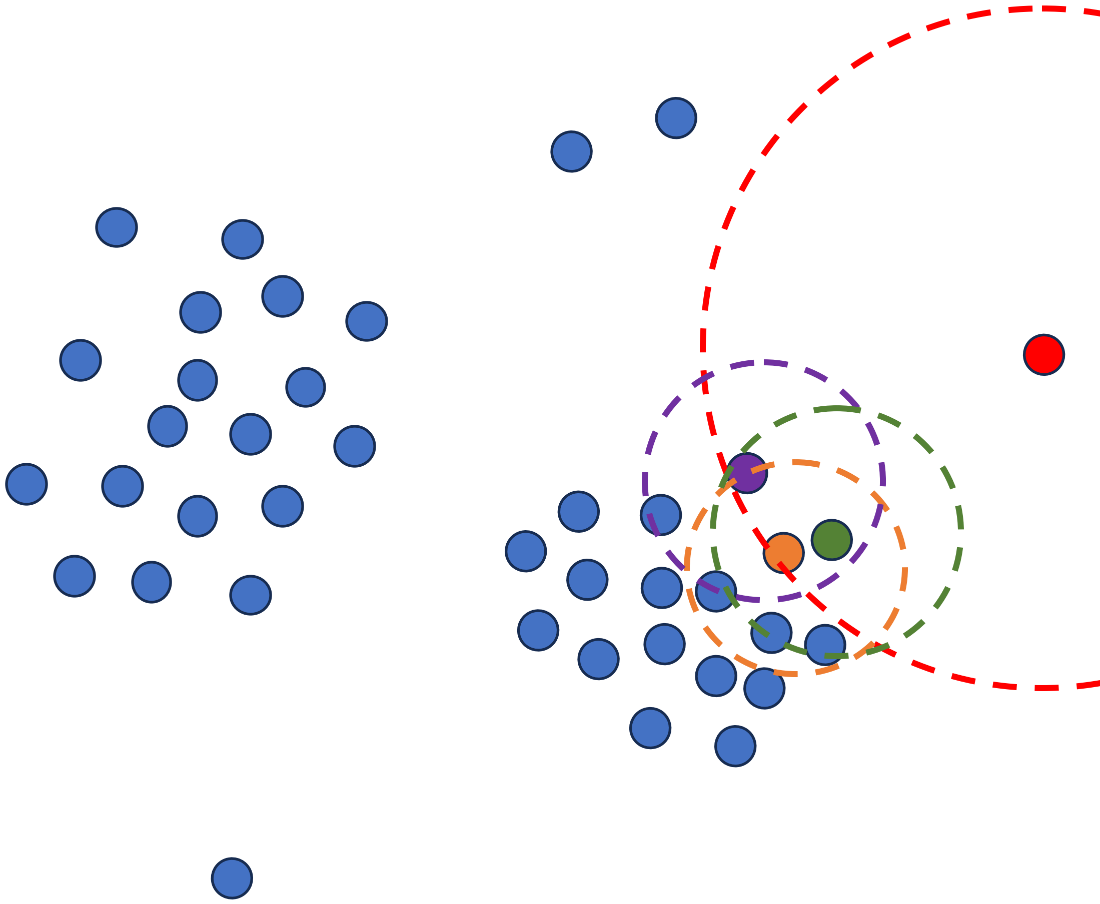
- For an anomalous point, the k-th NN will be quite far
- But, the nearest neighbors themselves (if not anomalous) will have much closer neighbors!



Local Outlier Factor (LOF)

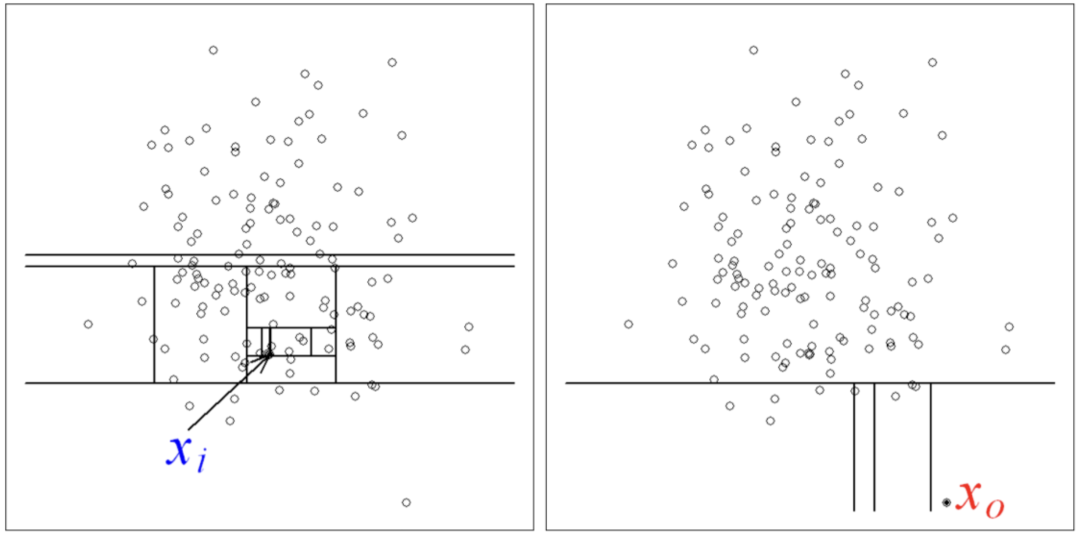


Anomalous!

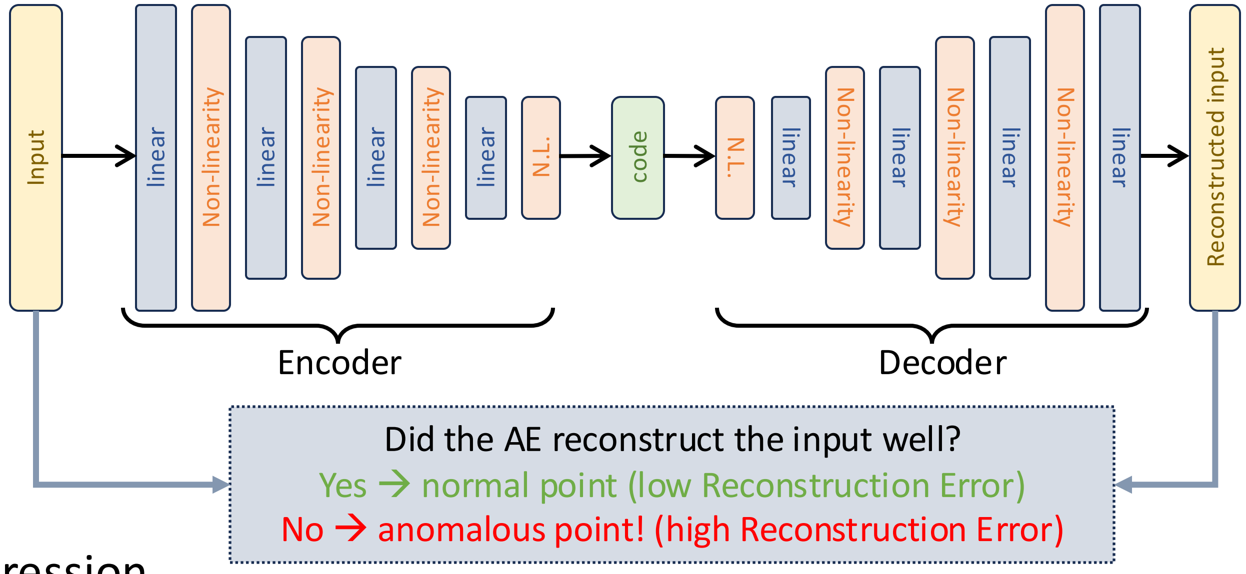


Isolation Forest

- An Isolation Forest is a collection of *randomly generated trees*
  - Randomly generated = for each node, randomly select a feature and split value, and split on “feature < value”
  - Each tree will split the input space into random partitions
- Anomalies: points isolated in fewer splits (short root-leaf path)
- Normal points: more difficult to be isolated (long root-leaf path)
- Repeat for multiple (random) trees, to get multiple paths
- $Anomaly\ score(x) = 2^{\frac{-\mathbb{E}(h(x))}{c(n)}}$
  - $\mathbb{E}(h(x))$ ➔ average path length to isolate x
  - $c(n)$ ➔ normalization factor based on dataset size, since more points = deeper trees on average = longer paths to isolate each point



Autoencoders (AE)



- Autoencoders are Neural Networks with an encoder and a decoder.
  - The encoder compresses the N-dimensional input into a D-dimensional vector (“code”)
    - Note that, if D<N (as is often the case), the compression will be lossy, if the data is full rank
  - The decoder takes the D-dimensional “compressed” vector and tries to reconstruct the original input
- During training, the AE “learns” to compress typical data and to reconstruct it properly.
- After training, if a data point is reconstructed well, it means that the AE knew how to compress/decompress the point (so, it was an ”expected” point)
- Anomalous points are those that the AE cannot reconstruct well (because the AE hasn’t seen “many” such points)