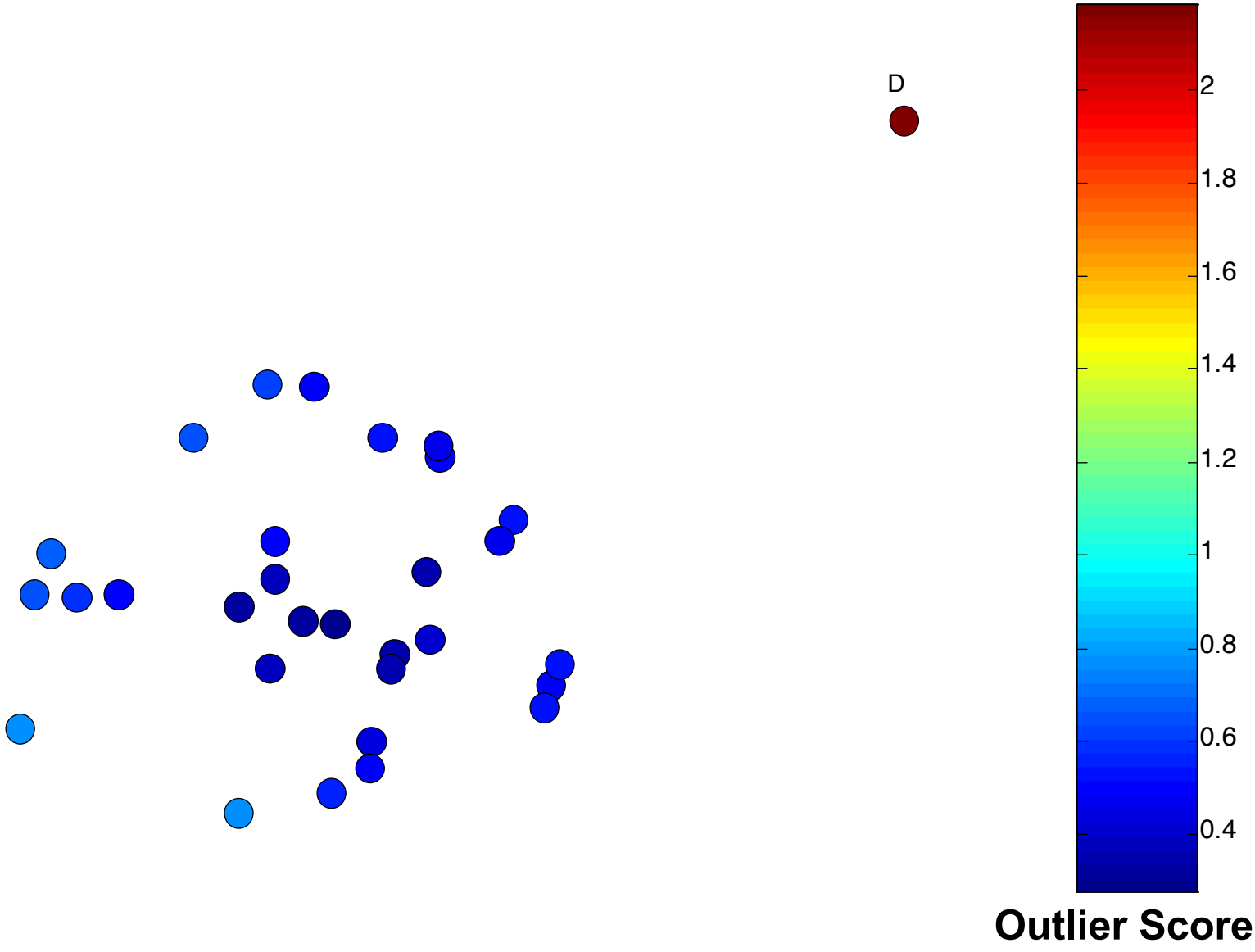
Techniques for unsupervised anomaly detection

- Proximity-based
- Density-based
- Clustering-based
- Isolation forest
- Auto Encoder
- Deep One-Class Classification
- Deep isolation forest

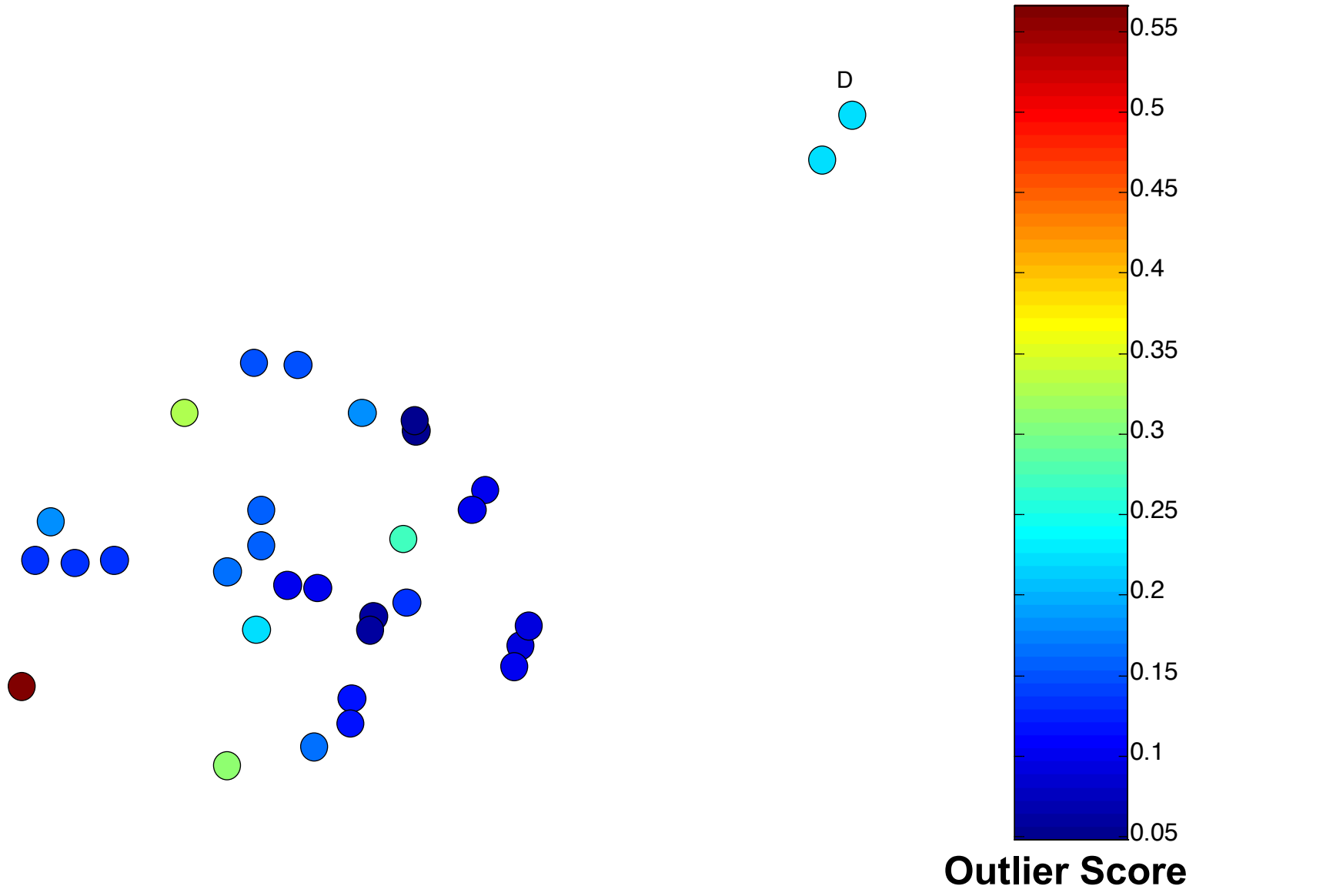
Distance-Based Approaches

- An object is an outlier if a specified fraction of the objects is more than a specified distance away (Knorr, Ng 1998)
  - Some statistical definitions are special cases of this
- The outlier score of an object is the distance to its kth nearest neighbor

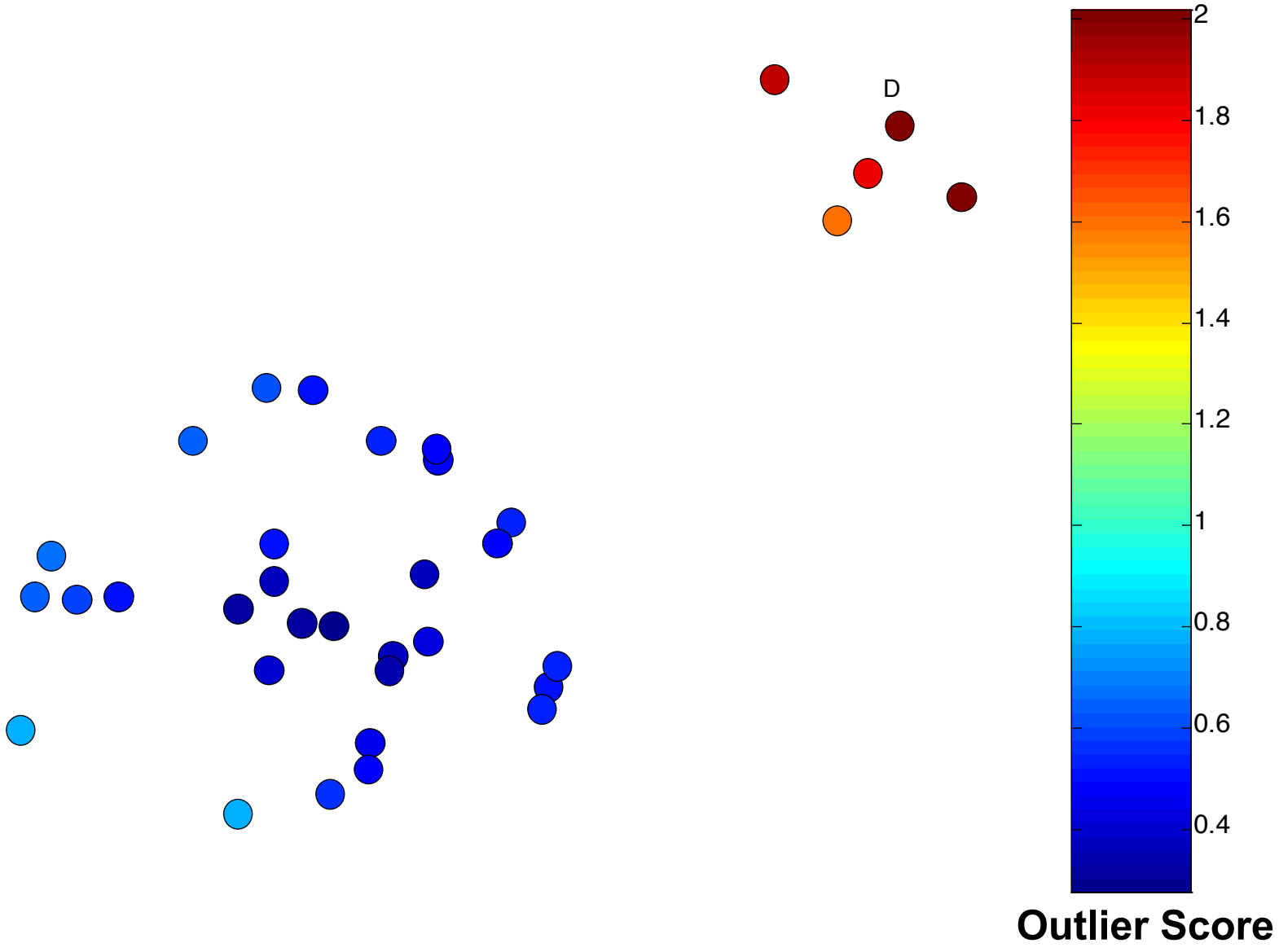
Nearest Neighbor – K=5



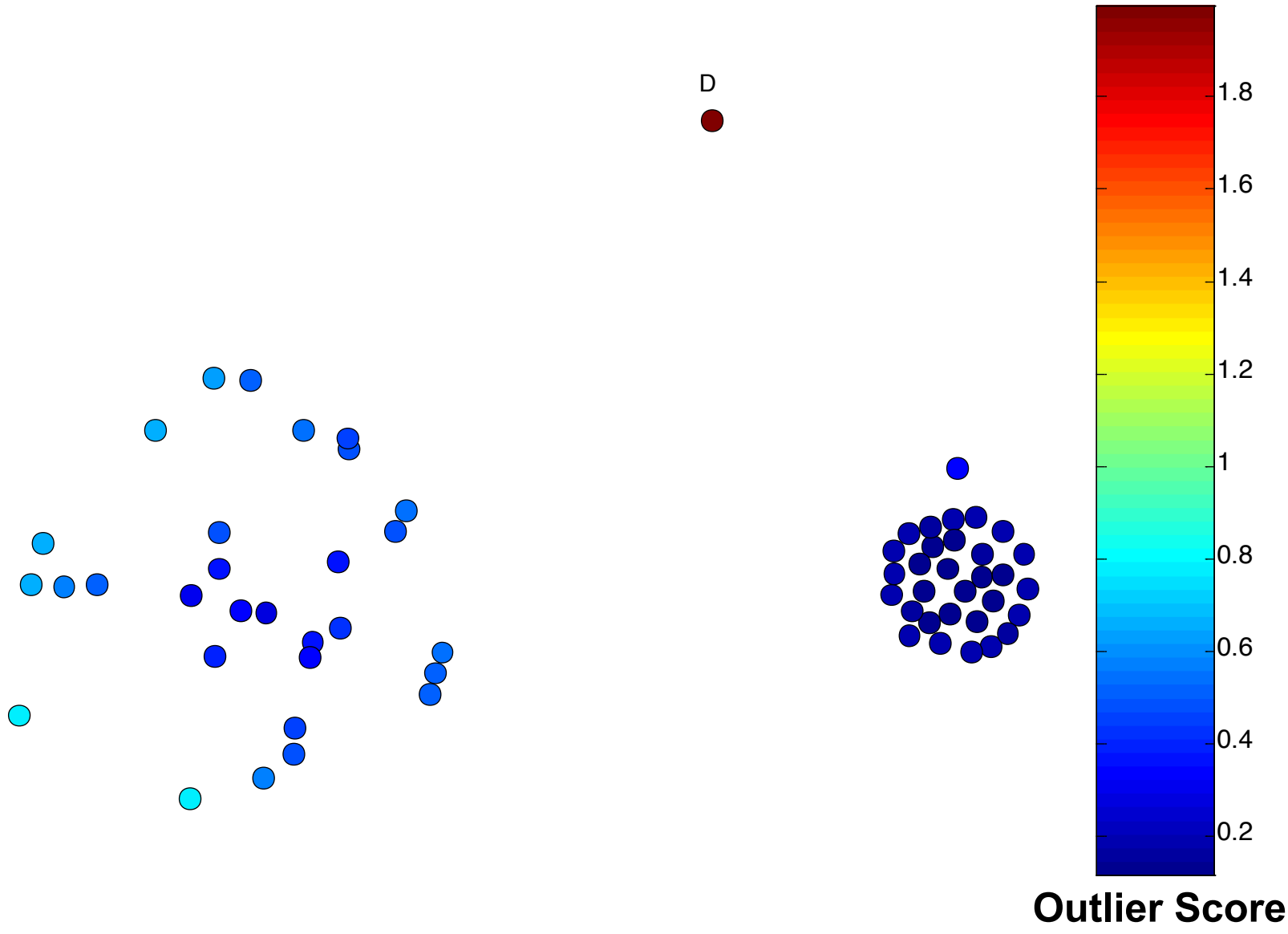
Nearest Neighbor – K=1



Five Nearest Neighbors - Small Cluster



Five Nearest Neighbors - Differing Density



Strengths/Weaknesses of Distance-Based Approaches

- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Sensitive to variations in density
- Distance becomes less meaningful in high-dimensional space

Density-Based Approaches

- **Density-based Outlier:** The outlier score of an object is the inverse of the density around the object.
  - Can be defined in terms of the k nearest neighbors
  - One definition: Inverse of distance to the k^{th} neighbor
  - Another definition: Inverse of the average distance to k neighbors
  - DBSCAN definition

- If there are regions of different density, this approach can have problems

Relative density outlier score (Local Outlier Factor, LOF)

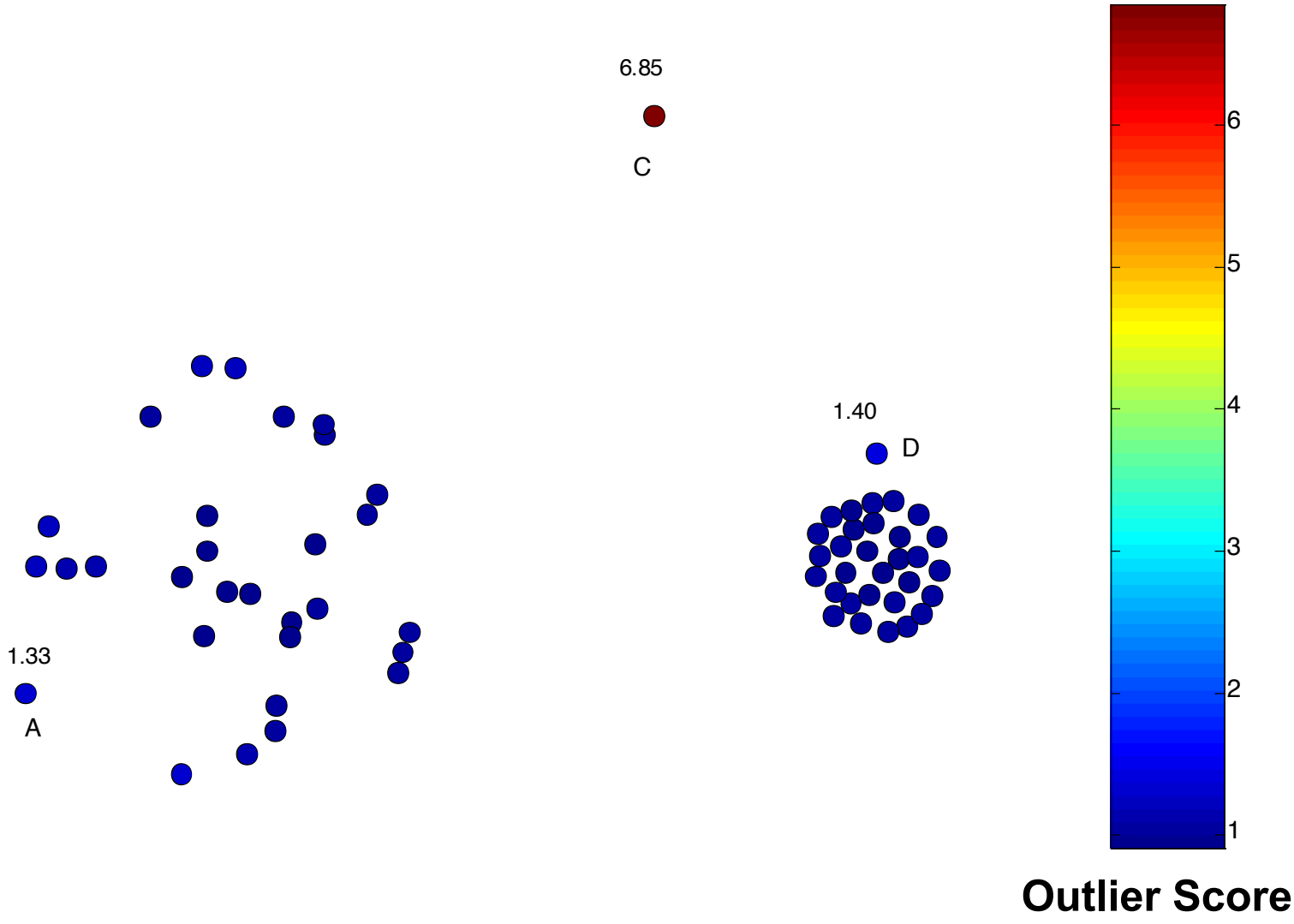
- Consider the density of a point relative to that of its k nearest neighbors

$$average\ relative\ density(\mathbf{x}, k) = \frac{density(\mathbf{x}, k)}{\sum_{\mathbf{y} \in N(\mathbf{x},k)} density(\mathbf{y}, k)/|N(\mathbf{x}, k)|}. \quad (10.7)$$

Algorithm 10.2 Relative density outlier score algorithm.

1: $\{k$ is the number of nearest neighbors$\}$
2: **for all** objects $\mathbf{x}$ **do**
3: Determine $N(\mathbf{x}, k)$, the k-nearest neighbors of $\mathbf{x}$.
4: Determine $density(\mathbf{x}, k)$, the density of $\mathbf{x}$, using its nearest neighbors, i.e., the objects in $N(\mathbf{x}, k)$.
5: **end for**
6: **for all** objects $\mathbf{x}$ **do**
7: Set the $outlier\ score(\mathbf{x}, k) = average\ relative\ density(\mathbf{x}, k)$ from Equation 10.7.
8: **end for**

Relative Density Outlier Scores



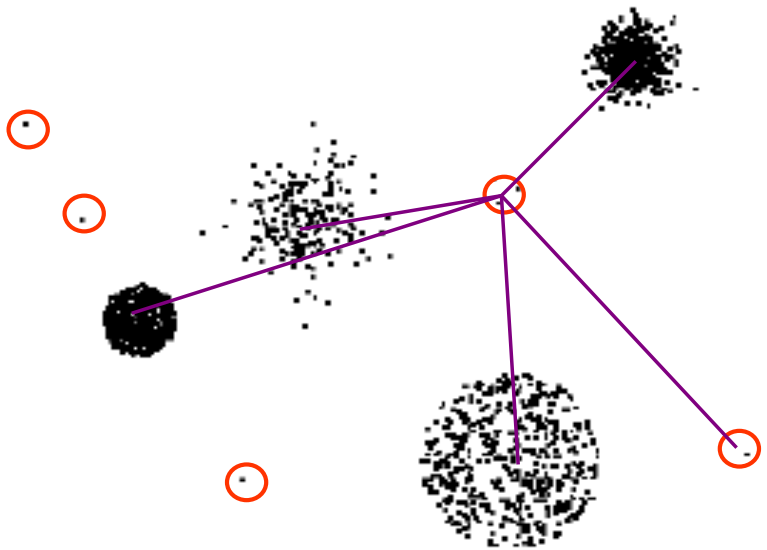
Density-Based Approaches

- Pros
  - Quantitative measure of degree to which object is an outlier.
  - Can work well even if data has variable density.

- Cons
  - $O(n^2)$ complexity
  - Must choose parameters
    - k for nearest neighbor
    - d for distance threshold

Clustering-Based Approaches

- **Clustering-based Outlier:** An object is a cluster-based outlier if it does not strongly belong to any cluster
  - For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
  - For density-based clusters, an object is an outlier if its density is too low
  - For graph-based clusters, an object is an outlier if it is not well connected
- Other issues include the impact of outliers on the clusters and the number of clusters



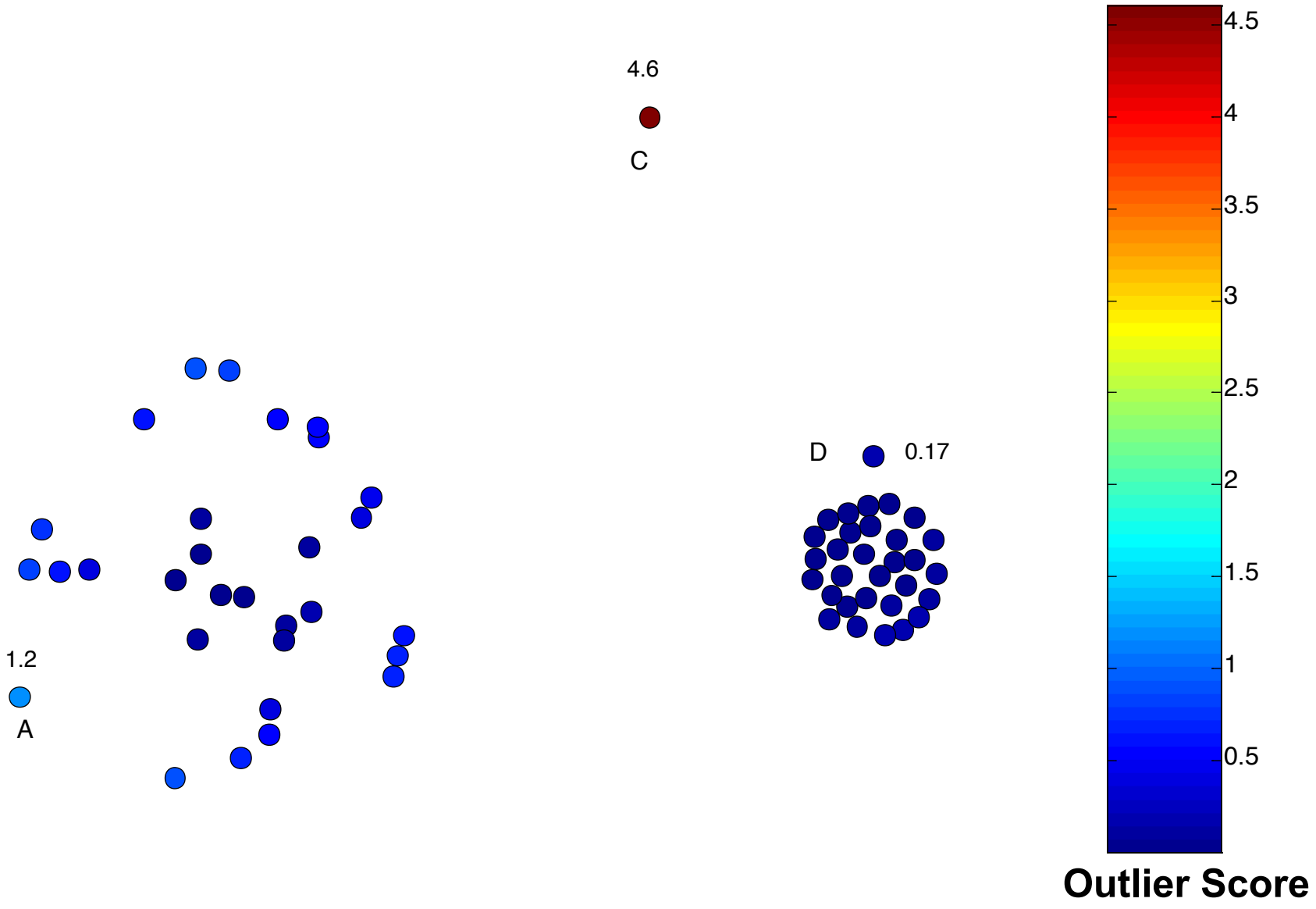
Clustering-Based Approaches

Assess degree to which object belongs to any cluster.

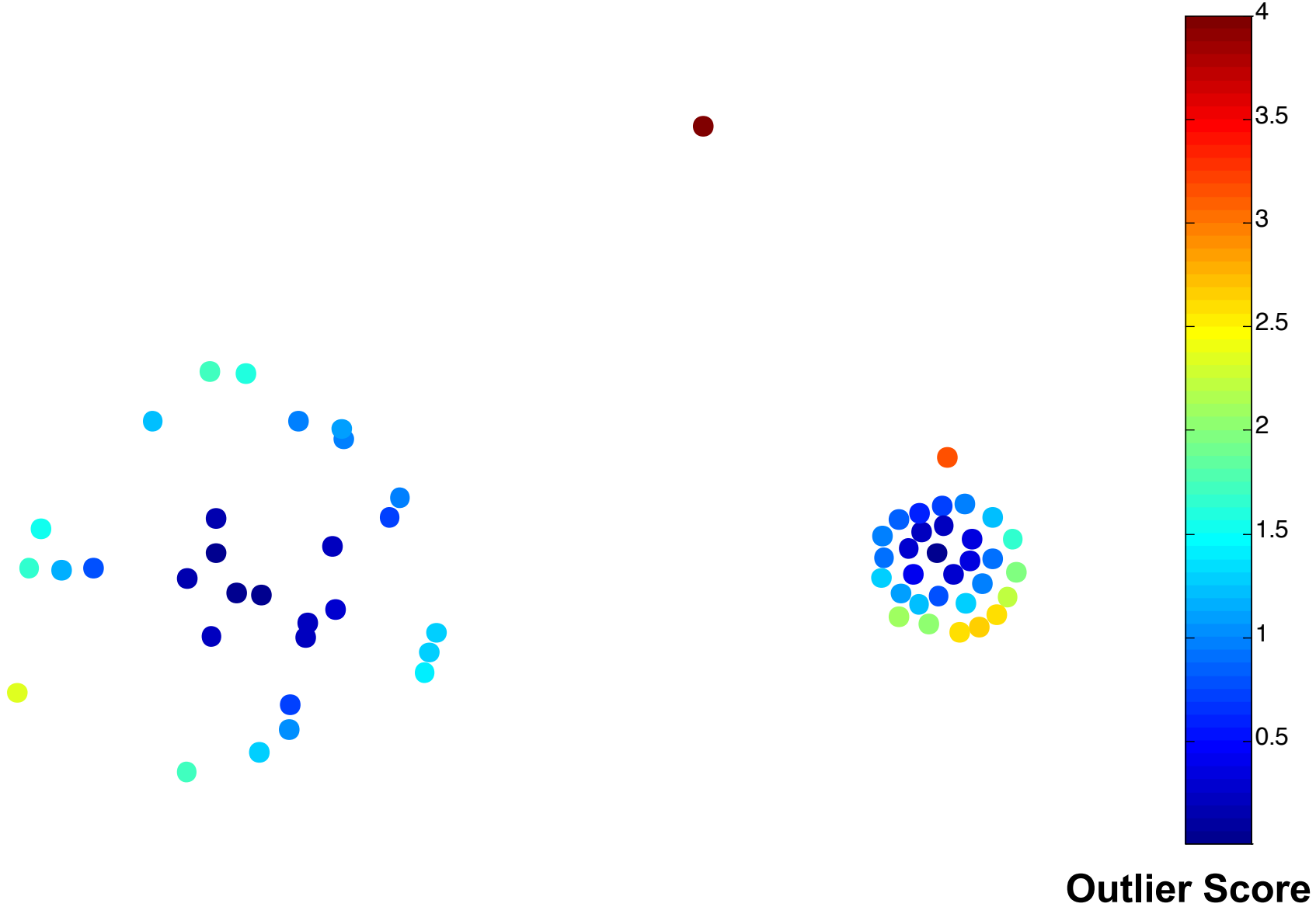
- For prototype-based clustering (e.g. k-means), use distance to cluster centers.
  - To deal with variable density clusters, use relative distance:

$$\frac{\text{distance}(\mathbf{x}, centroid_C)}{\text{median}(\{\forall_{x' \in C} \text{distance}(\mathbf{x}', centroid_C)\})}$$

Distance of Points from Closest Centroids



Relative Distance of Points from Closest Centroid



Strengths/Weaknesses of Distance-Based Approaches

- Simple
- Many clustering techniques can be used
- Can be difficult to decide on a clustering technique
- Can be difficult to decide on number of clusters
- Outliers can distort the clusters

Drawback of the existing approaches

- Many existing methods are constrained to low dimensional data and small data size because of their high computational complexity.

Isolation Forest

Isolation Forest

Fei Tony Liu, Kai Ming Ting
Gippsland School of Information Technology
Monash University, Victoria, Australia
{tony.liu},{kaiming.ting}@infotech.monash.edu.au

Zhi-Hua Zhou
National Key Laboratory
for Novel Software Technology
Nanjing University, Nanjing 210093, China
zhouzh@lamda.nju.edu.cn

Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In International Conference on Data Mining (ICDM), pp. 413-422

Isolation Forest

- This paper proposes a different type of model-based method that explicitly isolates anomalies rather than profiles normal instances.

- The proposed method takes advantage of two anomalies' quantitative properties:
  - they are the minority consisting of fewer instances
  - they have attribute-values that are very different from those of normal instances. (Anomalies are 'few and different', which make them more susceptible to isolation than normal points.)

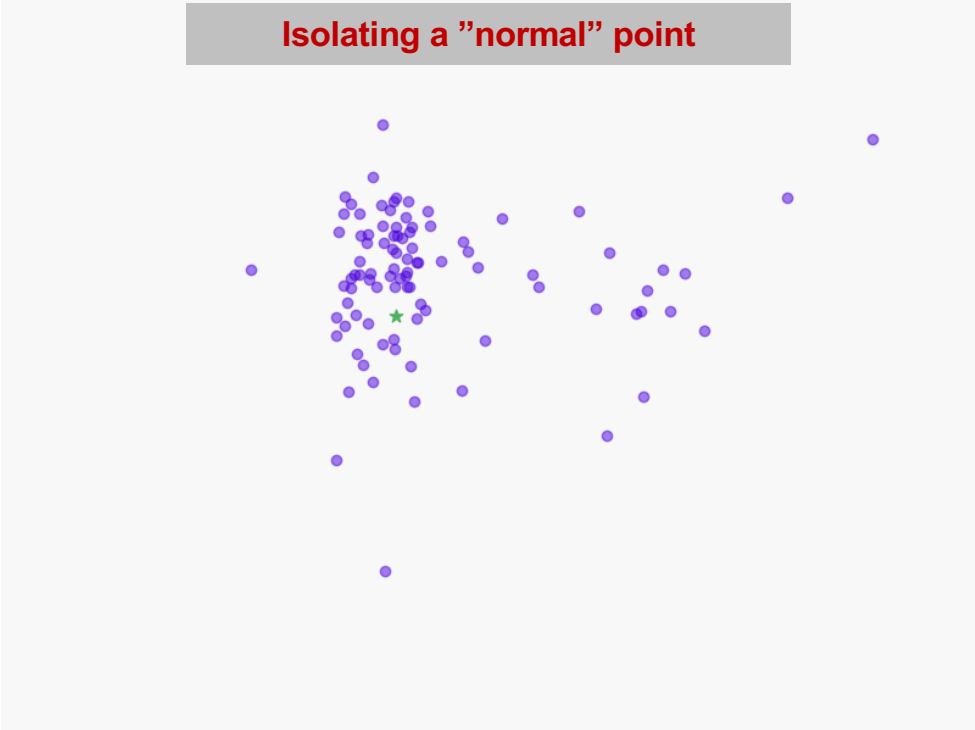
Isolation Forest

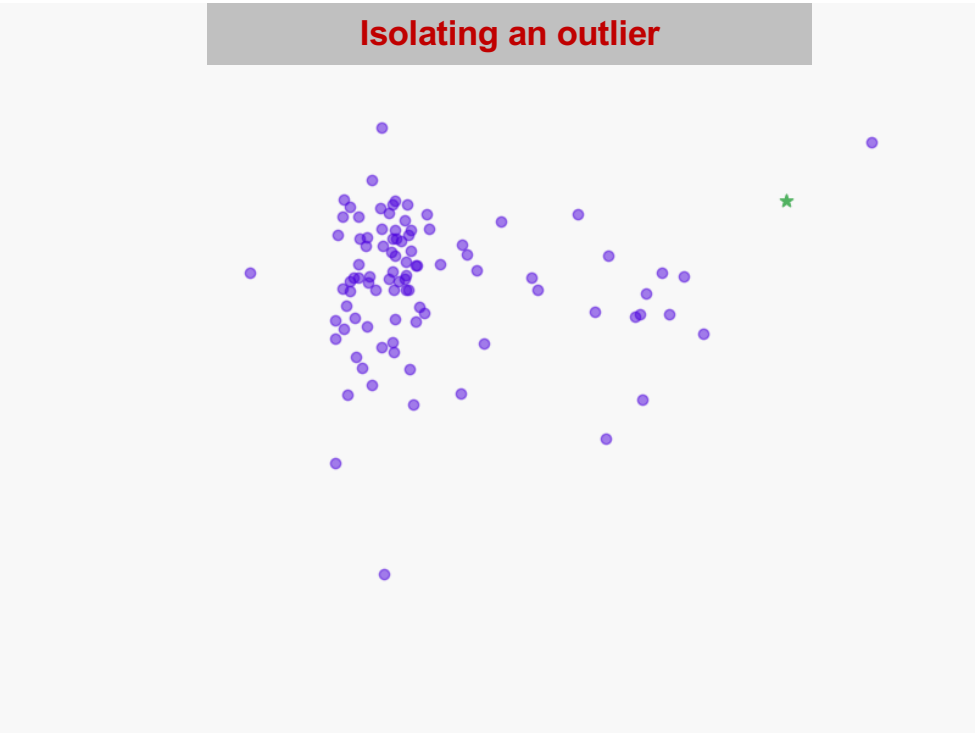
Definition : **Isolation Tree.** Let T be a node of an isolation tree. T is either an external-node with no child, or an internal-node with one test and exactly two daughter nodes (T_l,T_r). A test consists of an attribute q and a split value p such that the test $q < p$ divides data points into T_l and T_r.

Definition : **Path Length** $h(x)$ of a point x is measured by the number of edges x traverses an iTree from the root node until the traversal is terminated at an external node.

Isolation Forest

- Selects a feature, randomly selects a split between minimum and maximum values of the selected parameter
- Many splits are required in order to isolate a “normal” point
  - small number of splits: anomaly!

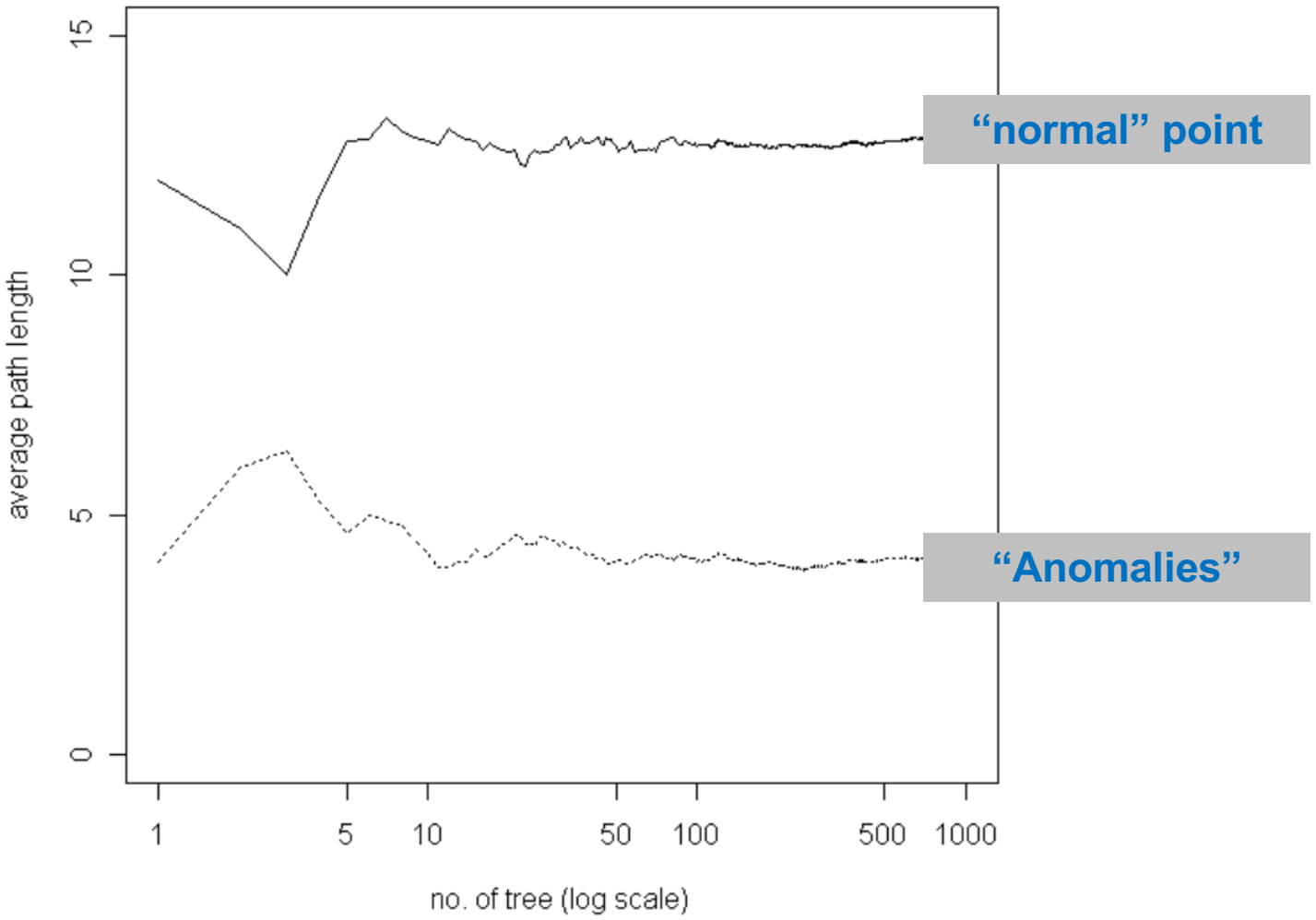




Isolation Forest

- The random partitioning produces noticeable shorter paths for anomalies since
  - the fewer instances of anomalies result in a smaller number of partitions – shorter paths in a tree structure;
  - instances with distinguishable attribute-values are more likely to be separated in early partitioning.

Isolation Forest



(c) Average path lengths converge

Isolation Forest

Algorithm 2 : $iTree(X, e, l)$

Inputs: X - input data, e - current tree height, l - height limit

Output: an iTree

1: **if** $e \geq l$ or $|X| \leq 1$ **then**
2: return *exNode*$\{Size \leftarrow |X|\}$
3: **else**
4: let Q be a list of attributes in X
5: randomly select an attribute $q \in Q$
6: randomly select a split point p from max and min values of attribute q in X
7: $X_l \leftarrow filter(X, q < p)$
8: $X_r \leftarrow filter(X, q \geq p)$
9: return *inNode*$\{Left \leftarrow iTree(X_l, e + 1, l),$
10: $Right \leftarrow iTree(X_r, e + 1, l),$
11: $SplitAtt \leftarrow q,$
12: $SplitValue \leftarrow p\}$
13: **end if**

Isolation Forest

Algorithm 1 : $iForest(X, t, \psi)$

Inputs: X - input data, t - number of trees, ψ - sub-sampling size

Output: a set of t *iTrees*

1: **Initialize** $Forest$
2: set height limit $l = ceiling(\log_2 \psi)$
3: **for** $i = 1$ to t **do**
4: $\quad X' \leftarrow sample(X, \psi)$
5: $\quad Forest \leftarrow Forest \cup iTree(X', 0, l)$
6: **end for**
7: **return** $Forest$

Isolation Forest

The anomaly score s of an instance x is defined as:

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

- c(n) is the average of h(x) given n
- E(h(x)) is the average of h(x) from a collection of isolation trees
- s is monotonic to h(x)
- when $E(h(x)) \to c(n)$, $s \to 0.5$;
- when $E(h(x)) \to 0$, $s \to 1$;
- and when $E(h(x)) \to n - 1$, $s \to 0$.
- If instances return s very close to 1, then they are definitely anomalies;
- If instances have s much smaller than 0.5, then they are quite safe to be regarded as normal instances;
- If all the instances return s ≈ 0.5, then the entire sample does not really have any distinct anomaly.

Characteristics of iForest

- iForest utilizes no distance or density measures to detect anomalies. This eliminates major computational cost of distance calculation in all distance-based methods and density-based methods.

- iForest has a linear time complexity with a low constant and a low memory requirement.

- iForest has the capacity to scale up to handle extremely large data size and high-dimensional problems with a large number of irrelevant attributes.