OntarioTech Science

Kourosh Davoudi
kourosh@uoit.ca

Anomaly Detection

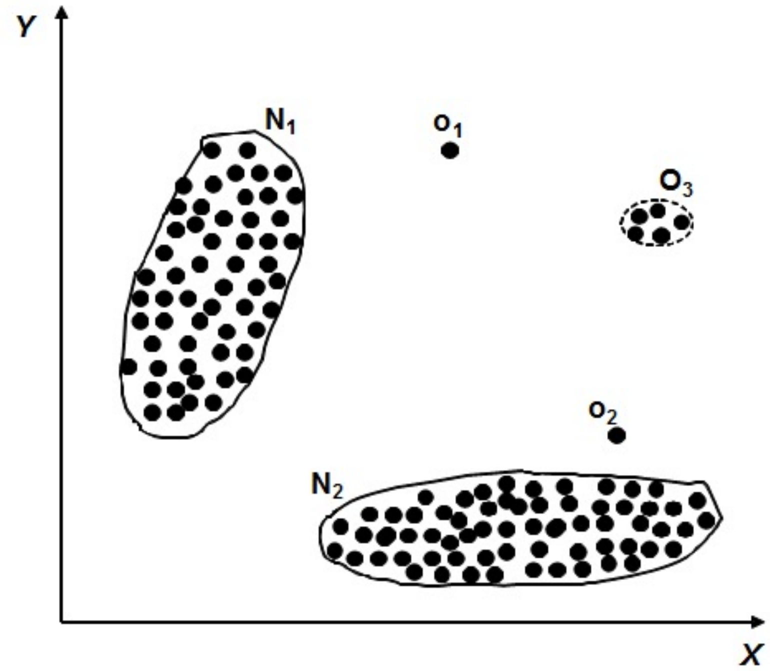
CSCI 4150U: Data Mining

Learning Outcomes

- Understand the anomaly detection tasks:
  - Basic definitions
  - Importance, Challenges, Causes, Applications
- Explain different anomaly detection approaches:
  - Visual-based
  - Statistical
  - Distance-based
  - Density-based
  - Cluster-based

Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Natural implication is that anomalies are relatively rare
  - One in a thousand occurs often if you have lots of data
  - Context is important, e.g., freezing temps in July
- Can be important or a nuisance
  - 10 foot tall 2 year old
  - Unusually high blood pressure



Motivation

Anomalous events usually occur relatively infrequently

However, when they do occur, their consequences can be quite dramatic and quite often in a negative sense

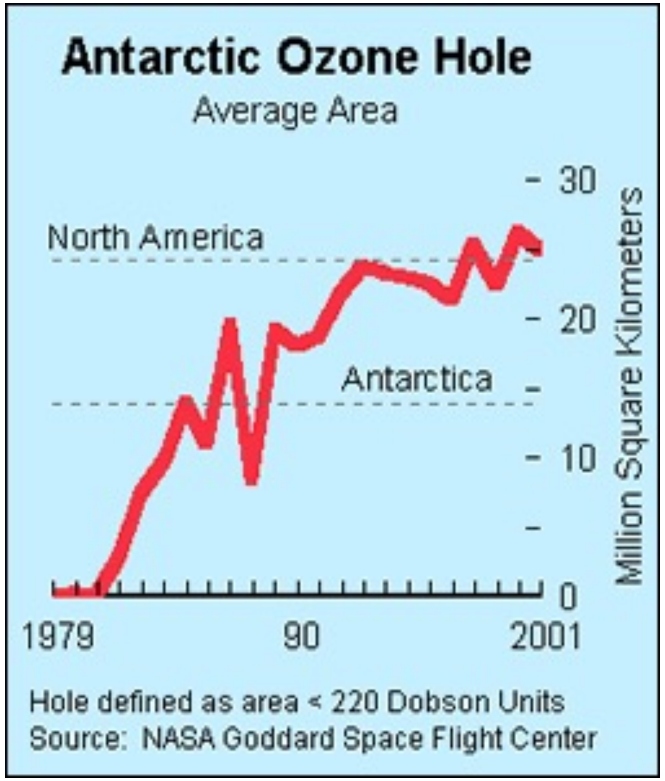
Importance of Anomaly Detection

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels

- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?

- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



http://www.epa.gov/ozone/science/hole/size.html

Causes of Anomalies

- Data from different classes
  - Measuring the weights of oranges, but a few grapefruit are mixed in
- Natural variation
  - Unusually tall people
- Data errors
  - 200 pound 2 year old

Is noise always an outlier?

A. Yes
B. No

Distinction Between Noise and Anomalies

- Noise is erroneous, perhaps random, values or contaminating objects
  - Weight recorded incorrectly
  - Grapefruit mixed in with the oranges

- Noise doesn't necessarily produce unusual values or objects
- Noise is not interesting
- Anomalies may be interesting if they are not a result of noise!
- Noise and anomalies are related but distinct concepts

Anomaly Detection Challenges

- Defining a representative normal region is challenging
- Method is often unsupervised, so validation can be quite challenging
- How many outliers are there in the data?
- The boundary between normal and outlying behavior is often not precise
- The exact notion of an outlier is different for different application domains
- Availability of labeled data for training/validation
- Normal behavior keeps evolving

"Mining needle in a haystack. So much hay and so little time"

Working assumption:

- There are considerably more "normal" observations than "abnormal" observations(outliers/anomalies) in data

Applications of Anomaly Detection

- Insurance / Credit card fraud detection (e.g. abnormally high purchase made on a credit card, etc.)
- Network intrusion detection
- Telecommunication fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining

Fraud Detection

Fraud detection refers to detection of criminal activities occurring in commercial organizations

- Malicious users might be the actual customers of the organization or might be posing as a customer (also known as identity theft).

Types of fraud

- Credit card fraud
- Insurance claim fraud

Challenges:

- Fast and accurate real-time detection
- Misclassification cost is very high

Money laundry detection is an example outlier detection task.

A. True
B. False

Healthcare Informatics

- Detect anomalous patient records
  - Indicate disease outbreaks, instrumentation errors, etc.
- Key Challenges
  - Only normal labels available
  - Misclassification cost is very high
  - Data can be complex: spatio-temporal

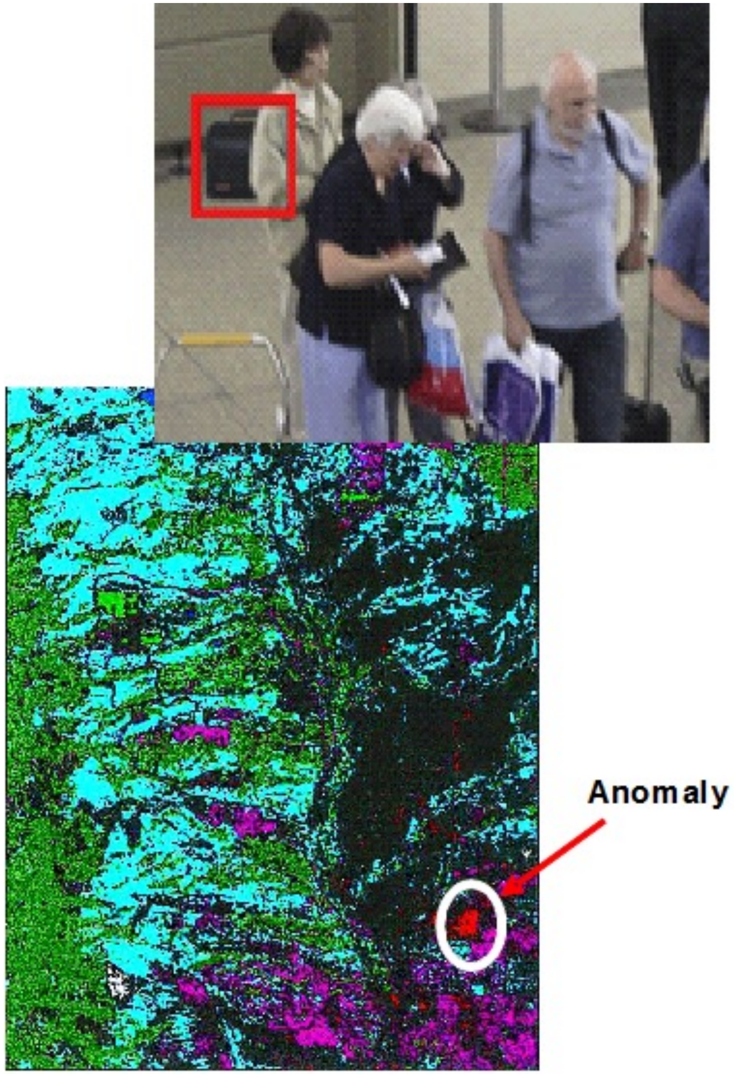
Industrial Damage Detection

- Industrial damage detection refers to detection of different faults and failures in complex industrial systems

- Example: Aircraft Safety
  - Anomalies in engine combustion data
- Key Challenges
  - Data is extremely huge, noisy and unlabelled
  - Most of applications exhibit temporal behavior
  - Detecting anomalous events typically require immediate intervention

Image Processing/Video Surveillance

- Detecting outliers in images monitored over time
- Detecting anomalous regions within an image
- Used in
  - mammography image analysis
  - video surveillance
  - satellite image analysis
- Key Challenges
  - Detecting collective anomalies
  - Data sets are often very large



Anomaly: Number of Attributes

- Many anomalies are defined in terms of a single attribute
  - Height
  - Shape
  - Color

- Can be hard to find an anomaly using all attributes
  - Noisy or irrelevant attributes
  - Object is only anomalous with respect to some attributes

- However, an object **may not be** anomalous in any one attribute

Anomaly: Scoring/Binary Categorization

- Many anomaly detection techniques provide only a binary categorization
  - An object is an anomaly or it isn’t
  - This is especially true of classification-based approaches

- Other approaches assign a score to all points
  - This score measures the degree to which an object is an anomaly
  - This allows objects to be ranked

- In the end, you often need a binary decision
  - Should this credit card transaction be flagged?
  - Still useful to have a score

Variants of Anomaly Detection Problems

- Given a data set D, find all data points $x \in D$ with anomaly scores greater than some threshold t

- Given a data set D, find all data points $x \in D$ having the top-n largest anomaly scores

- Given a data set D, containing mostly normal (but unlabeled) data points, and a test point x, compute the anomaly score of x with respect to D

Model-Based Anomaly Detection

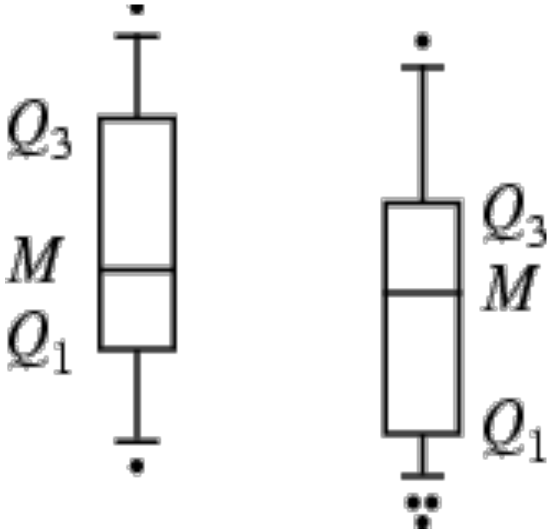
- Build a model for the data
  - Unsupervised
    - Anomalies are those points that don’t fit well
    - Anomalies are those points that distort the model
    - Examples:
      - Statistical distribution
      - Clusters
      - Graph
  - Supervised
    - Anomalies are regarded as a rare class
    - Need to have training data

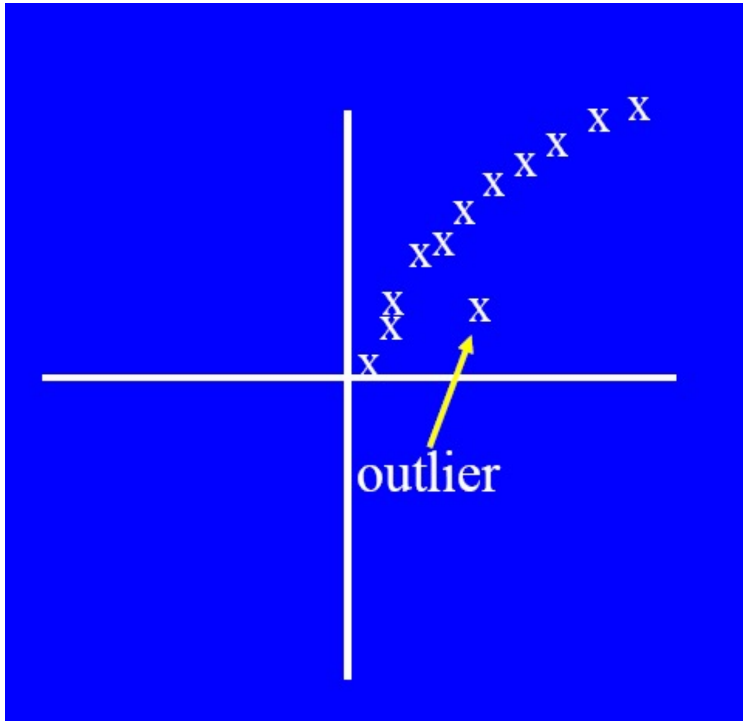
Additional Anomaly Detection Techniques

- Proximity-based
  - Anomalies are points far away from other points
  - Can detect this graphically in some cases
- Density-based
  - Low density points are outliers
- Pattern matching
  - Create profiles or templates of typical but important events or objects
  - Algorithms to detect these patterns are usually simple and efficient

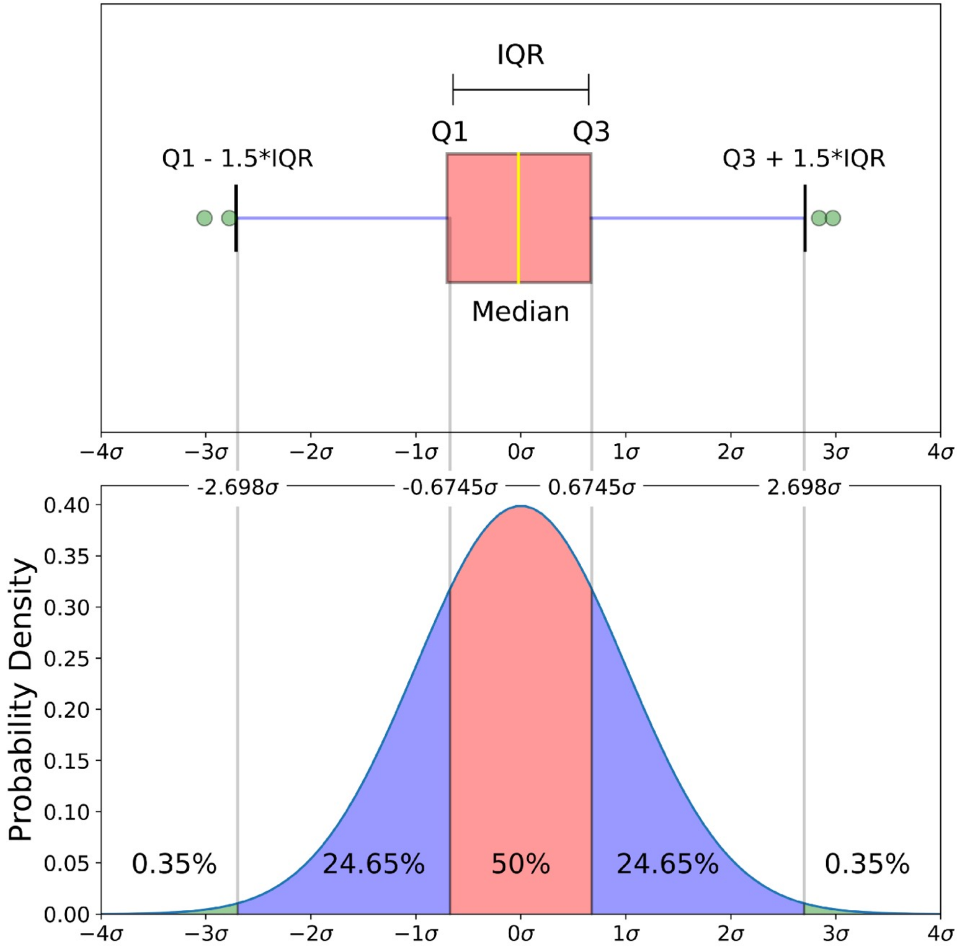
Visual Approaches

- Boxplots or scatter plots
- Limitations
  - Not automatic
  - Subjective





Boxplots (Detail)



- 25 % of data is less than Q1
- 75 % of data is less than Q2
- IQR = Q3 – Q2

Example:

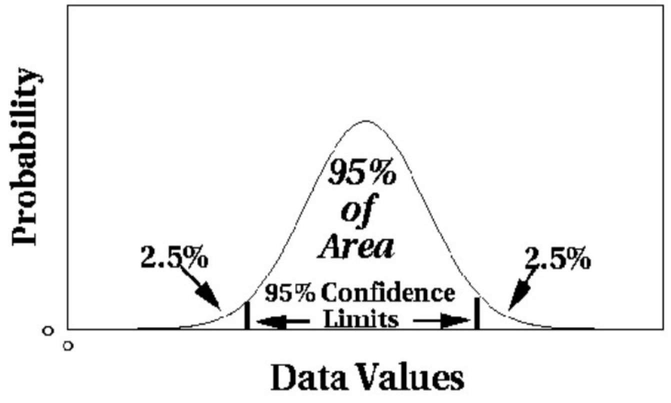
IQR = Q3 - Q1

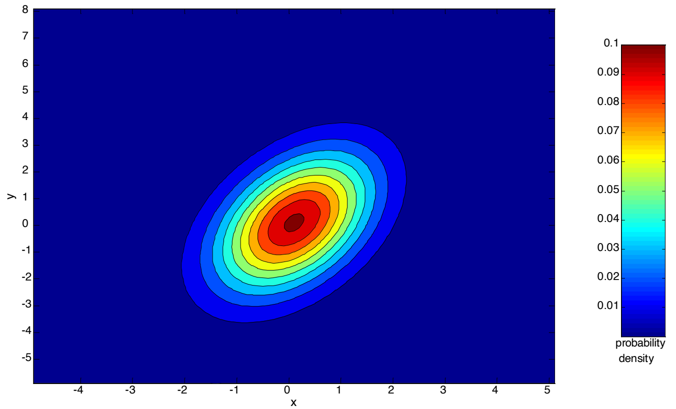
P is an Outlier if P > Q3 + 1.5 IQR
P is an Outlier if P < Q1 - 1.5 IQR
P is an Extreme Outlier if P > Q3 + 3 IQR
P is an Extreme Outlier if P < Q1 - 3 IQR

https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Statistical Approaches

- Probabilistic definition of an outlier:
  - An outlier is an object that has a **low probability** with respect to a probability distribution model of the data.
- Usually assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Approaches:
  - (1) Statistical test
    - Example: Grubbs' Test
      - Detect outliers in **univariate** data
      - Assume data comes from **normal** distribution
  - (2) Likelihood Approach
- Issues
  - Identifying the distribution of a data set
    - Heavy tailed distribution
  - Number of attributes
  - Is the data a mixture of distributions?





Statistical-based – Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)

$$D = (1 - \lambda)\, M + \lambda\, A$$

- General Approach:
  - Initially, assume all the data points belong to M
  - Let $LL_t(D)$ be the log likelihood of D at time t *(LL is log likelihood)*
  - For each point x_t that belongs to M, move it to A
    - Let $LL_{t+1}(D)$ be the new log likelihood.
    - Compute the difference, $\Delta = LL_t(D) - LL_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistical-based – Likelihood Approach

- Data distribution, $D = (1 - \lambda)\, M + \lambda\, A$
- M is a probability distribution estimated from data
  - Can be based on any modeling method
- A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

Strengths/Weaknesses of Statistical Approaches

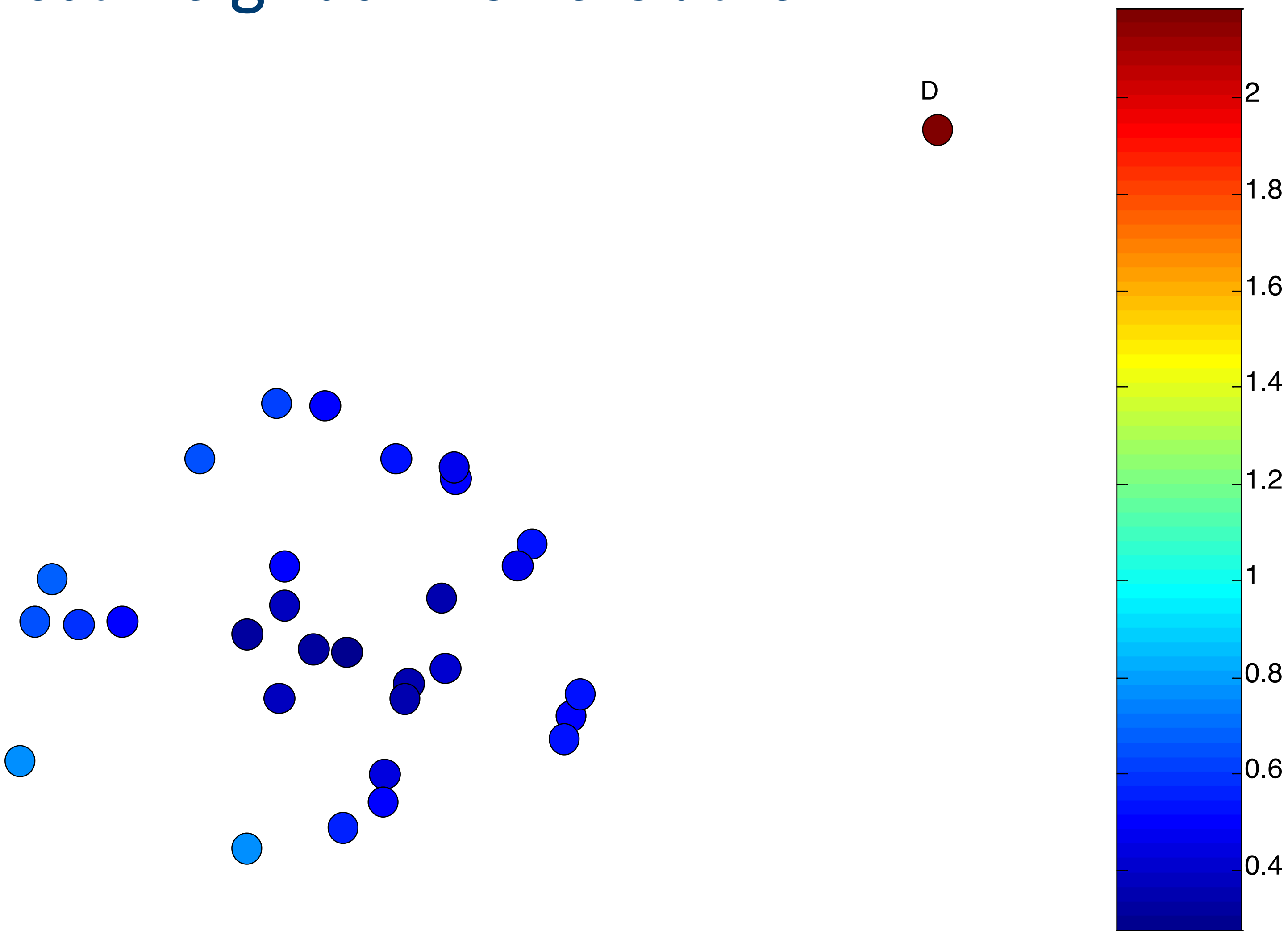
- Firm mathematical foundation
- Can be very efficient
- Good results if distribution is known
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution
- Anomalies can distort the parameters of the distribution

Proximity/Distance-Based Approaches

- Approach 1: A point x is an outlier if at least fraction p of the points in the dataset lies greater than distance d from the object x
  - You need to specify d and p

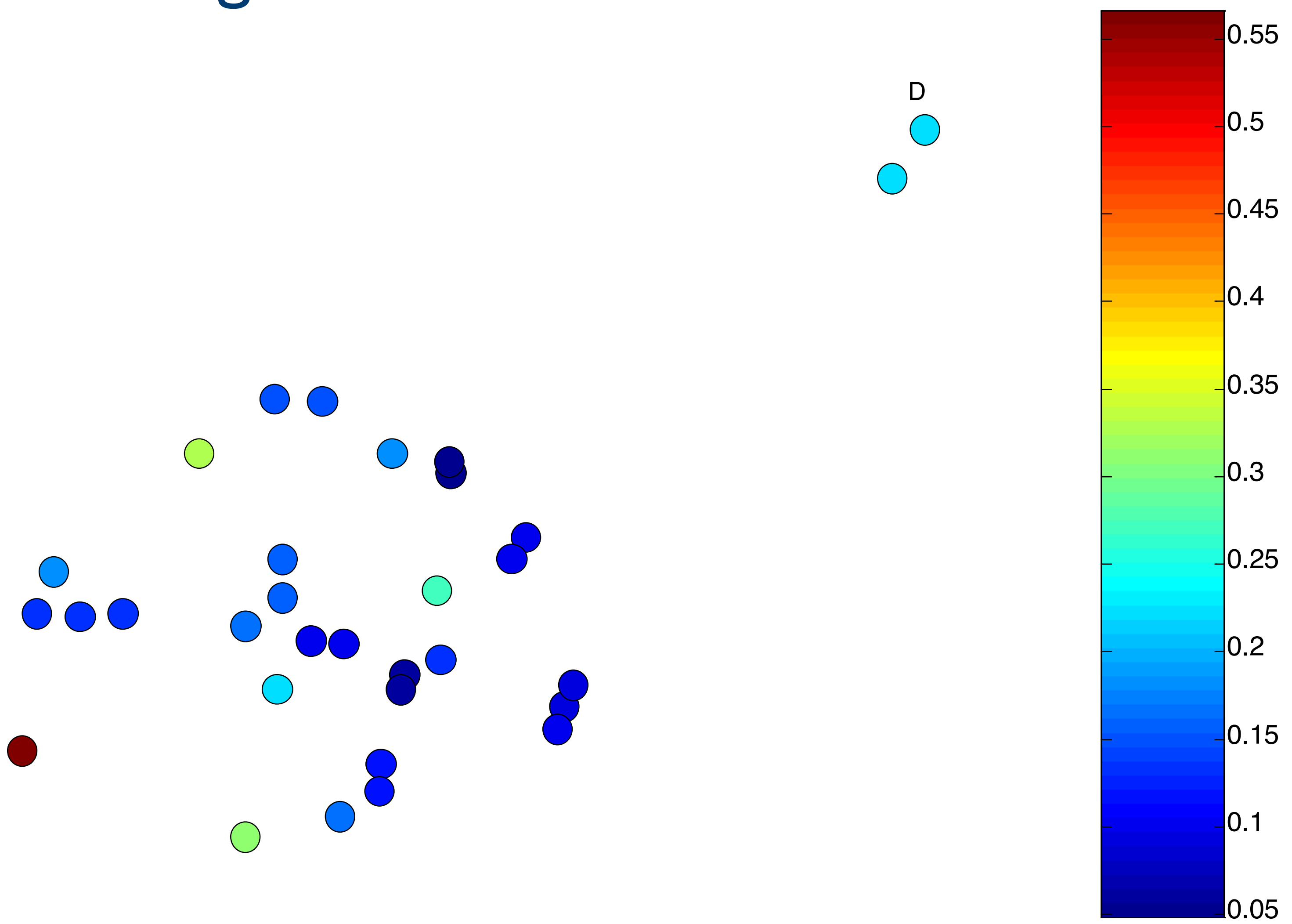
- Approach 2: The outlier score of an object is the distance to its k'th nearest neighbor
  - You need to specify k

One Nearest Neighbor - One Outlier



Outlier Score

One Nearest Neighbor - Two Outliers



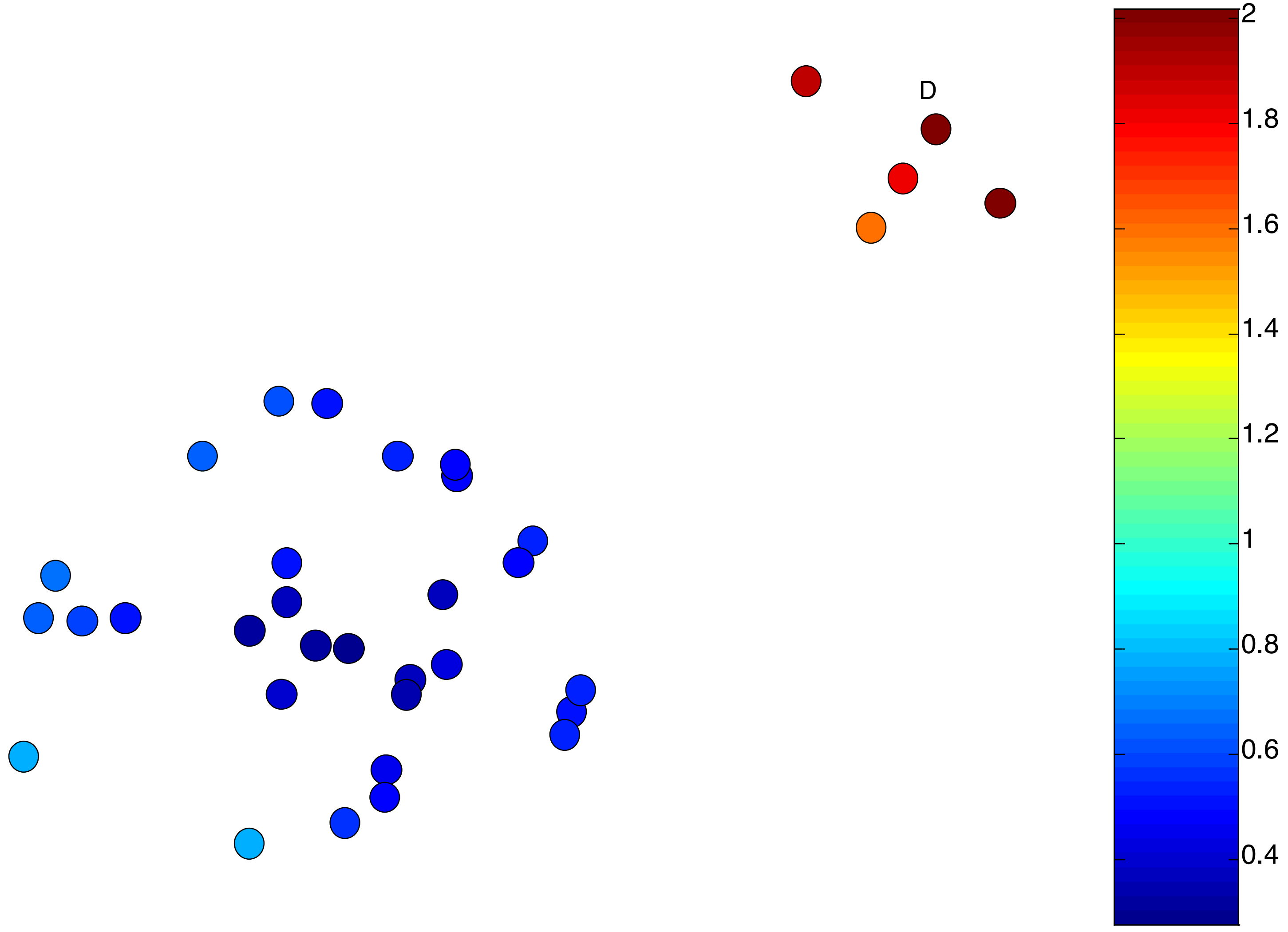
Outlier Score

In the The outlier **score** of an object is the distance to its k'th NN. Clusters with less than or equal k points can have high score.

A. True

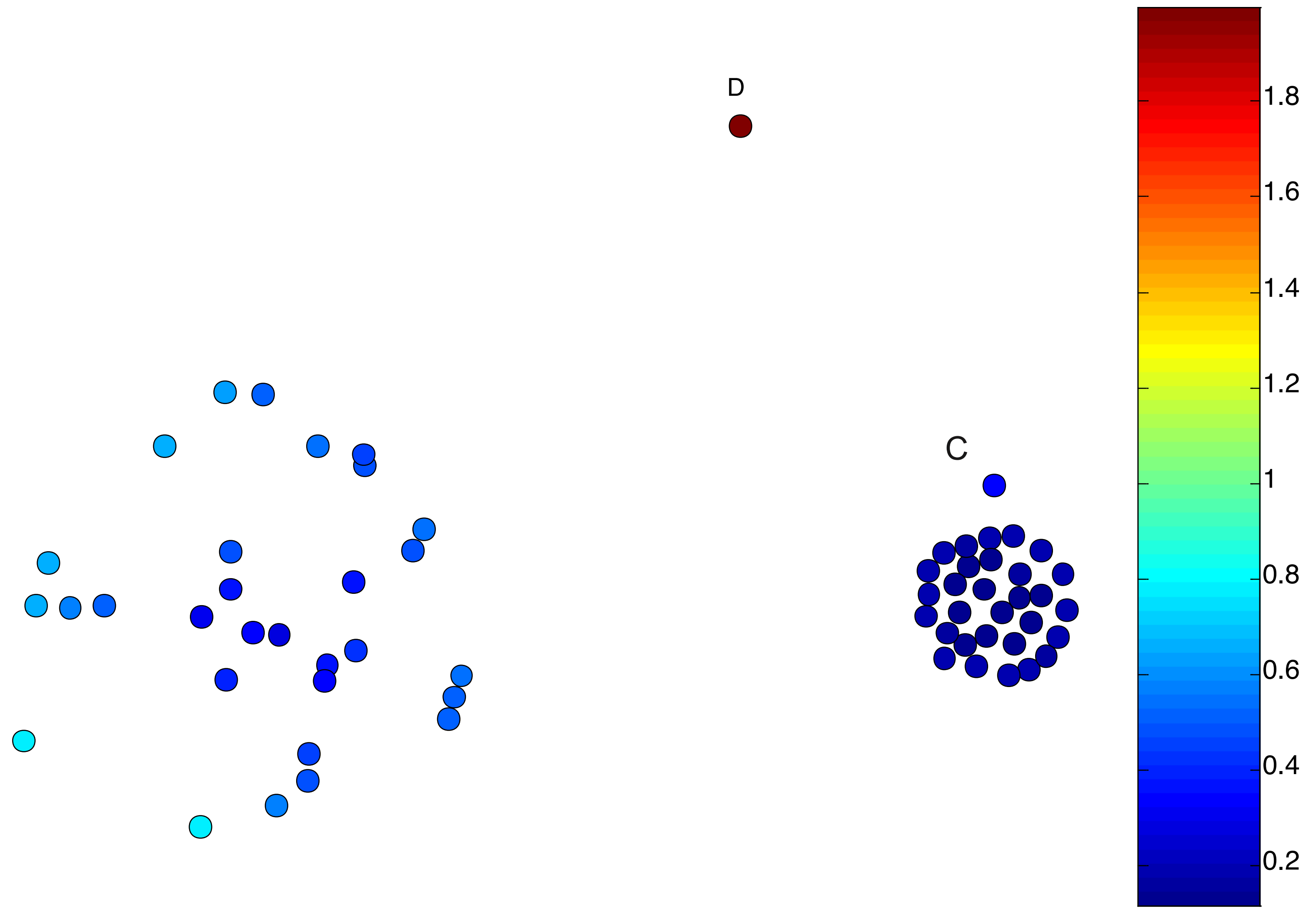
B. False

Five Nearest Neighbors - Small Cluster



Outlier Score

Five Nearest Neighbors - Differing Density



Outlier Score

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.
- Density definitions:
  - One definition: Inverse of distance to 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

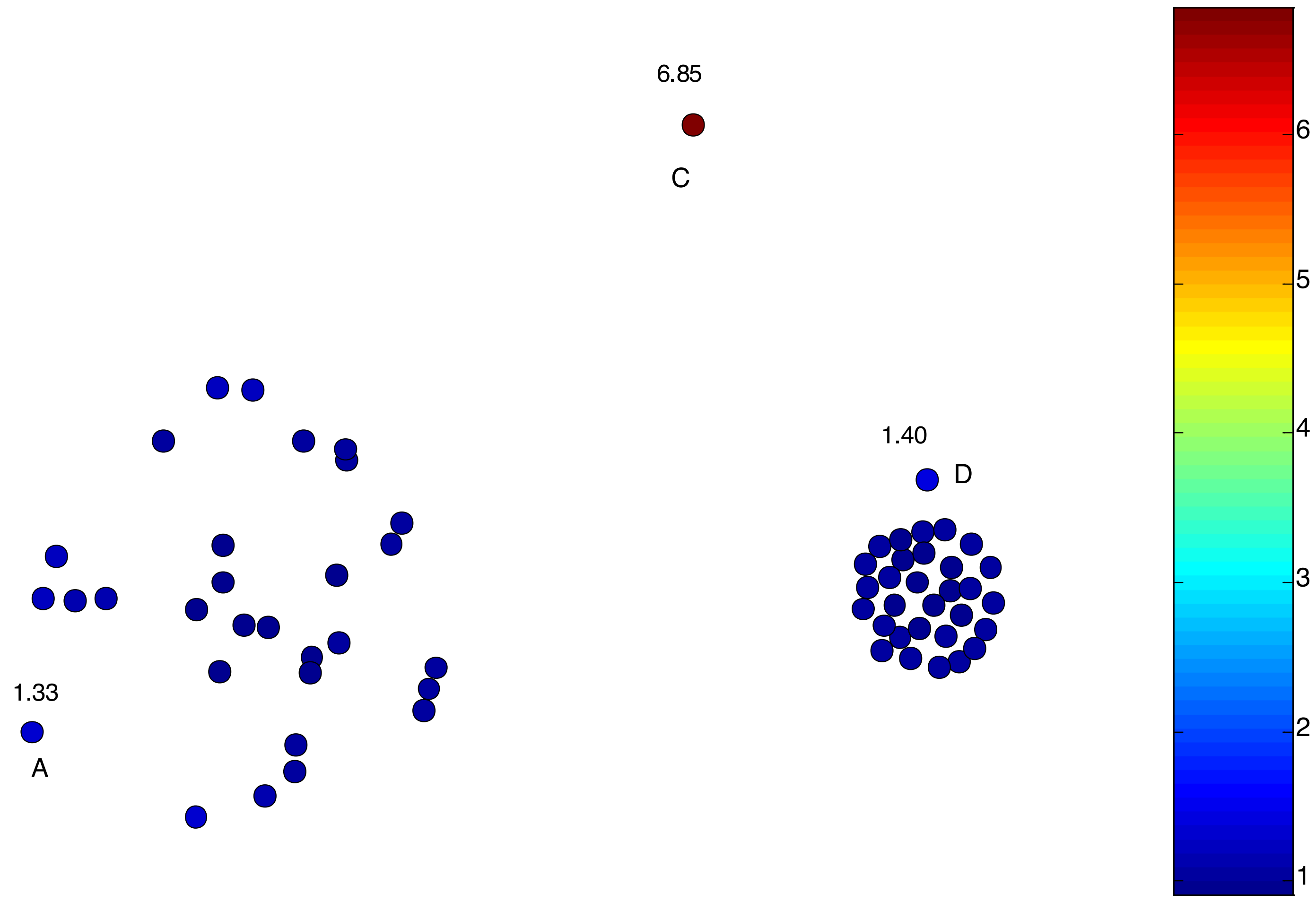
- Consider the density of a point relative to that of its k nearest neighbors

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

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 $\textit{density}(\mathbf{x}, k)$, 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 $\textit{outlier score}(\mathbf{x}, k) = \textit{average relative density}(\mathbf{x}, k)$ from Equation 10.7.
8: **end for**

Relative Density Outlier Scores



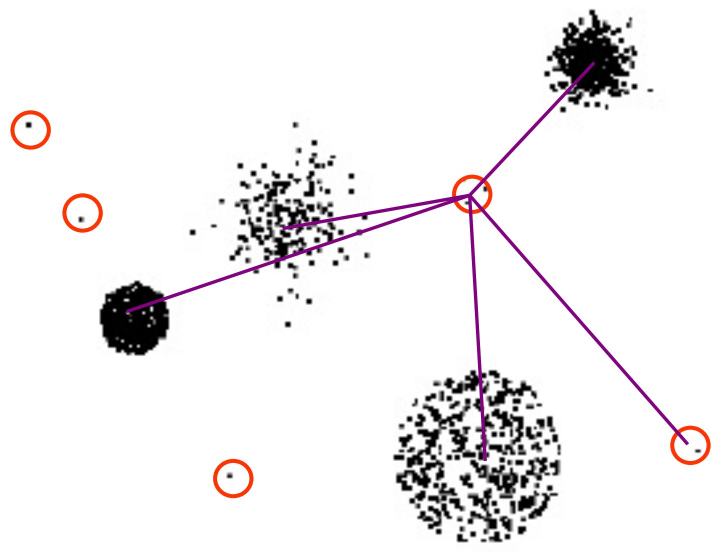
Outlier Score

Strengths/Weaknesses of Density-Based Approaches

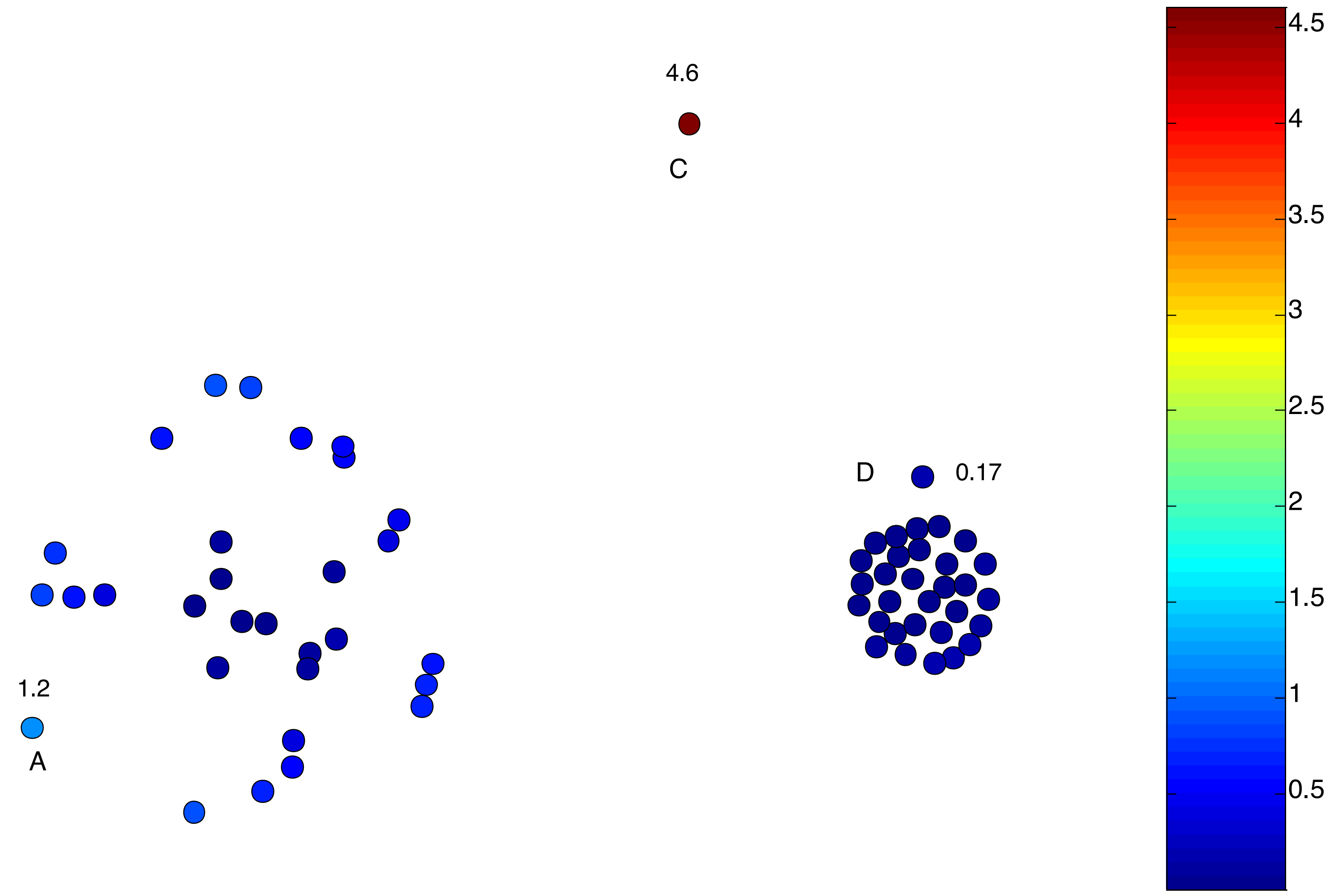
- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Density becomes less meaningful in high-dimensional space

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



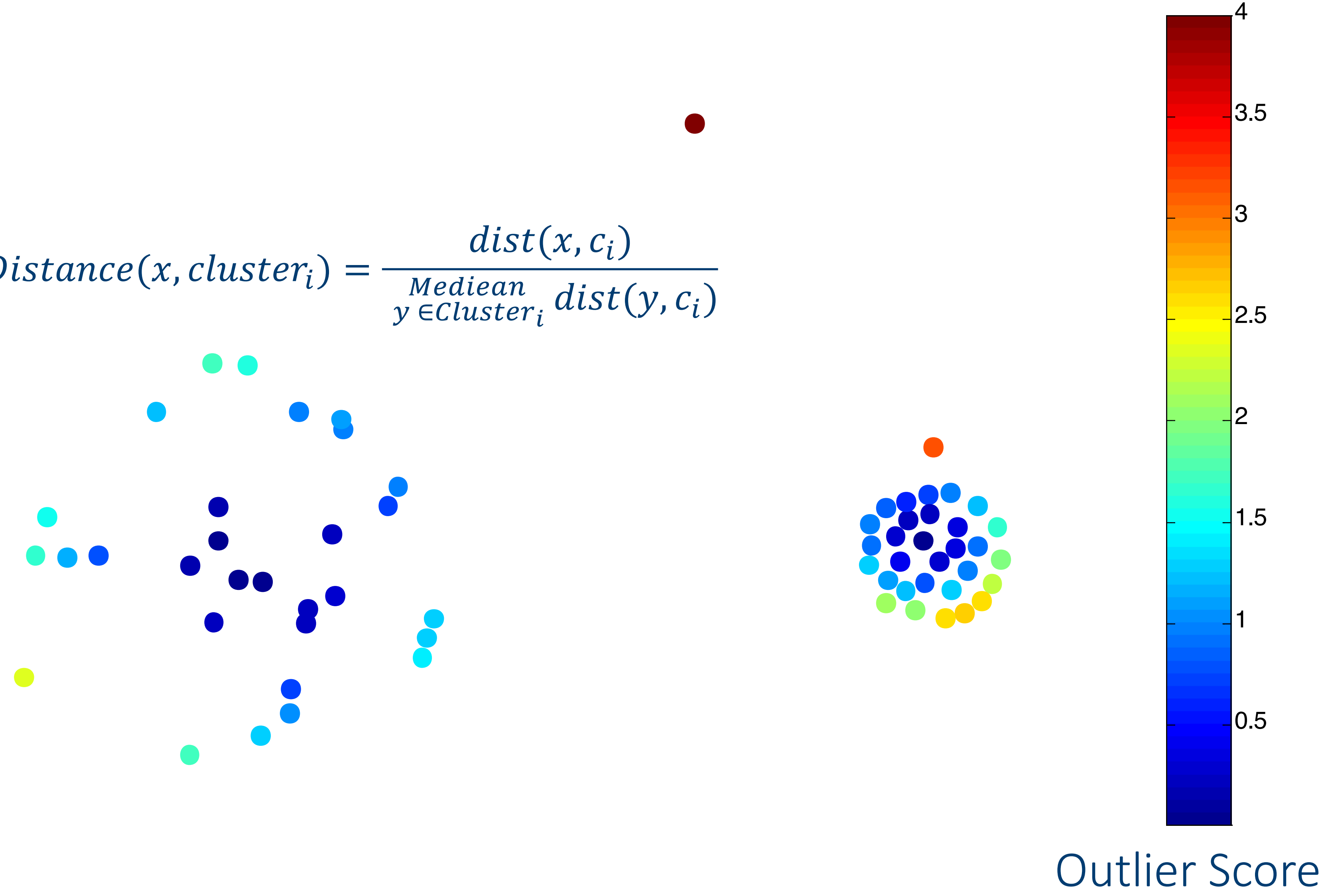
Distance of Points from Closest Centroids



Outlier Score

Relative Distance of Points from Closest Centroid

$$Relative\ Distance(x, cluster_i) = \frac{dist(x, c_i)}{\underset{y \in Cluster_i}{Mediean}\ dist(y, c_i)}$$



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

Participant Leaders

Points	Participant	Points	Participant