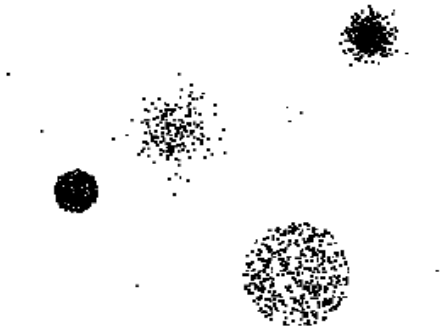
Anomaly Detection

Tozammel Hossain

Data Science & Analytics
University of Missouri

Anomaly 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**
  - 200 pound, 2 year old
  - Unusually high blood pressure



Types of Anomaly Detection

- **Outlier Detection**
  - Observations that is far from others (inliers)
  - Focus on regions where the training data is the most concentrated
- **Novelty Detection**
  - Training data is not polluted by outliers
  - Detecting whether a **new** observation is an outlier
- **Aka deviation detection, exception mining**

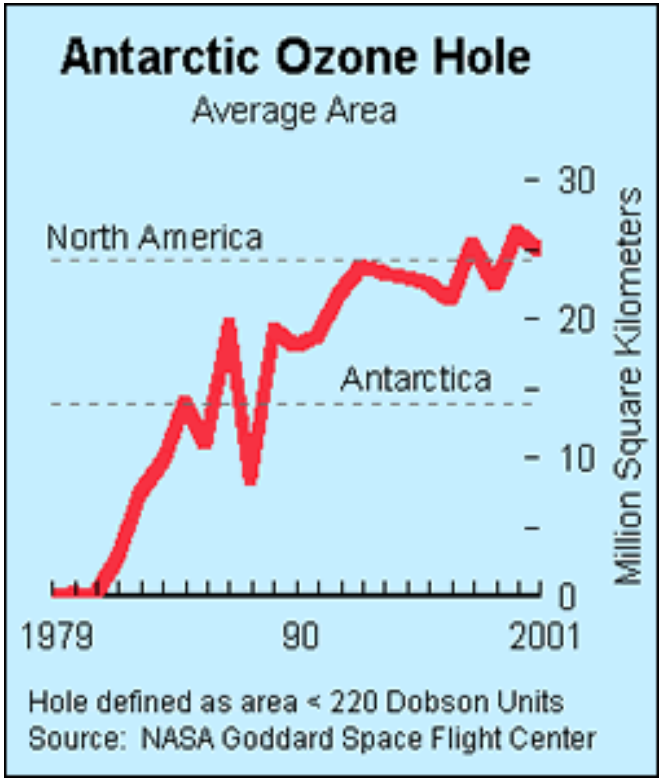
Importance of Anomaly Detection

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!



Applications

- **Fraud Detection**
- **Intrusion Detection**
- **Public Health**
- **Medicine**
- **Ecosystem**

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

Noise vs Anomalies

- **Noise doesn't necessarily produce unusual values or objects**
- **Noise is not interesting**
- **Noise and anomalies are related but distinct concepts**

General Issues: Anomaly Scoring

- **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

Type of Anomaly Detection Problems

- 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
- 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

Model-Based Anomaly Detection

- **Unsupervised (Outlier Detection)**
  - Anomalies are those points that don't fit well
  - Anomalies are those points that distort the model
- **Supervised (Rare Class Detection)**
  - Anomalies are regarded as a rare class
  - Need to have training data
- **Semi-supervised (Novelty Detection)**
  - Normal data is given for training

Anomaly Detection Techniques

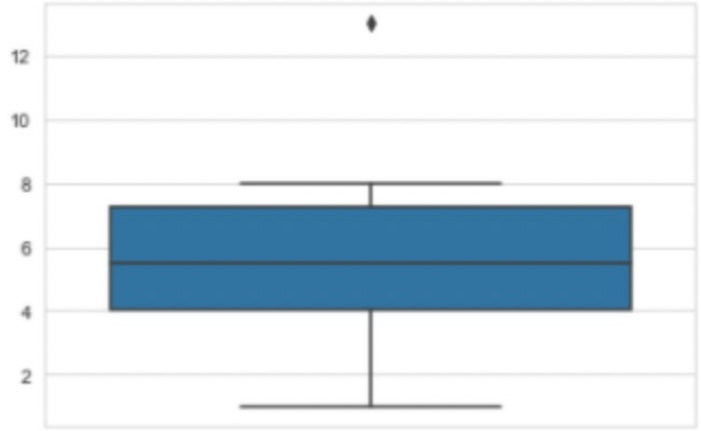
- **Statistical Approaches**
- **Proximity-based**
  - Anomalies are points far away from other points
- **Clustering-based**
  - Points far away from cluster centers are outliers
- **Reconstruction Based**

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)

- **Apply a statistical test that depends on**
  - Data distribution
  - Parameters of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)

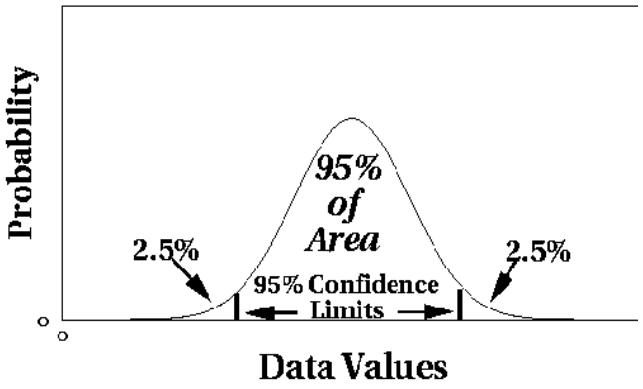
Box Plot

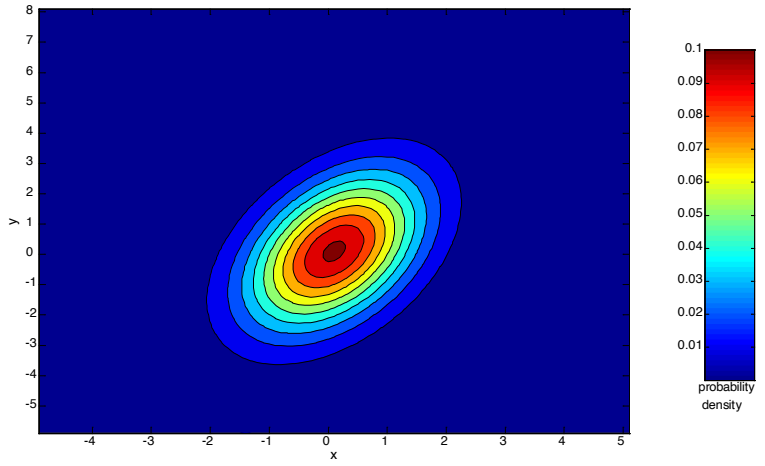

12
10
8
6
4
2


Min = Q1 - 1.5*IQR
Max = Q3 + 1.5*IQR
IQR
1st quartile
3rd quartile

Elliptic Envelop

- **Assumption: Data came from a normal/Gaussian distribution**
- **Fit a Gaussian model and try to define a shape of the data**



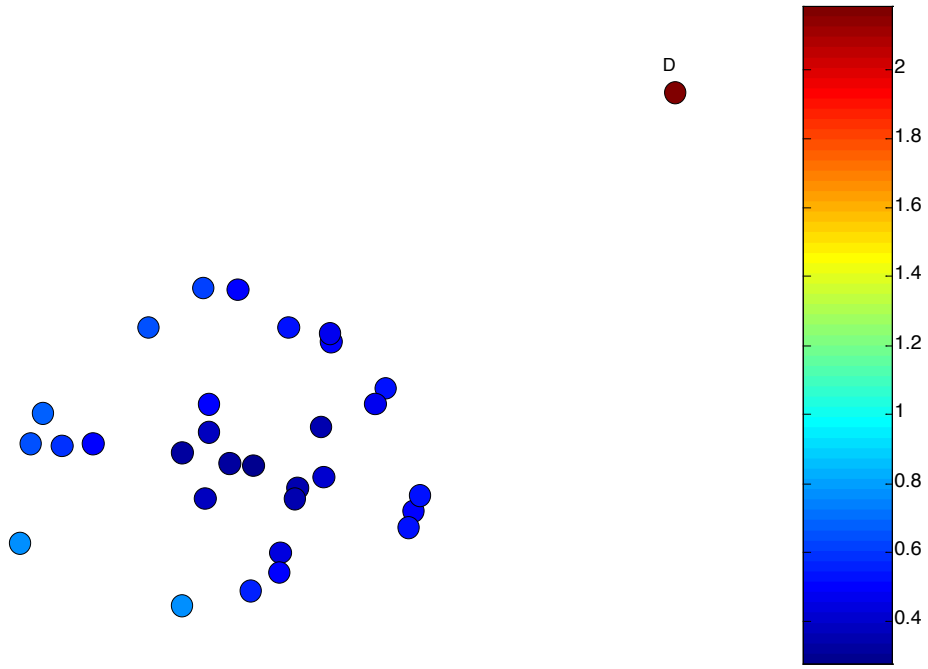


Strengths/Weaknesses of Statistical Approaches

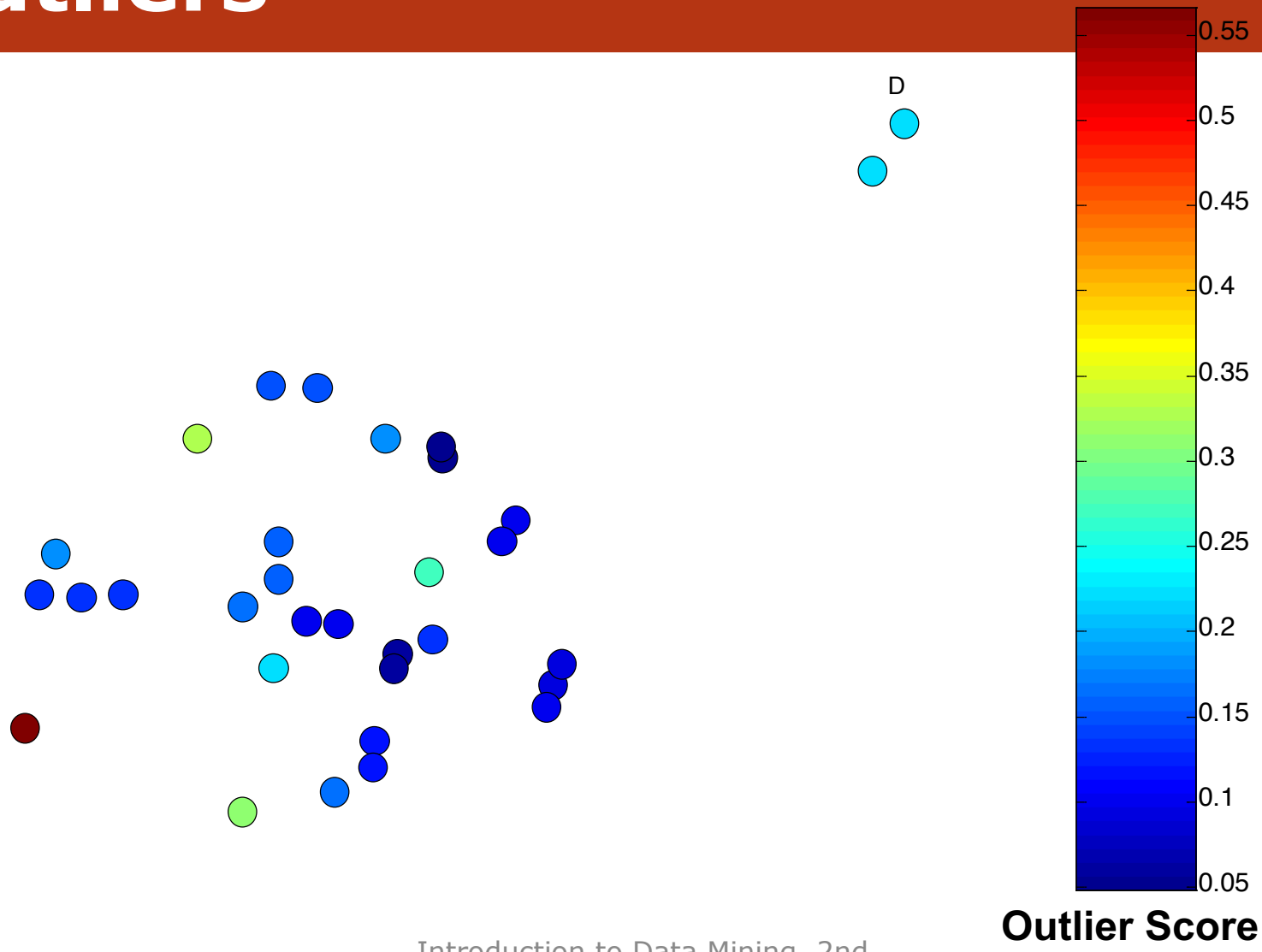
- Firm mathematical foundation
- Can be very efficient
- Good results if distribution is known
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution
- Anomalies can distort the parameters of the distribution

Distance-Based Approaches

- **The outlier score of an object is the distance to its kth nearest neighbor**



One Nearest Neighbor - Two Outliers



Strengths/Weaknesses of Distance-Based Approaches

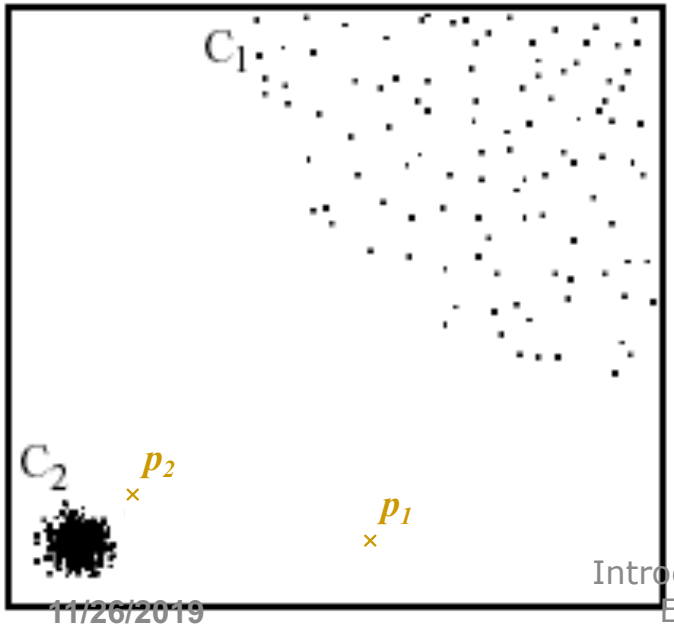
- **Simple**
- **Expensive – $O(n^2)$**
- **Sensitive to parameters**
- **Sensitive to variations in density**
- **Distance becomes less meaningful in high-dimensional space**

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 kth neighbor
  - Another definition: Inverse of the average distance to k neighbors
  - DBSCAN, LocalOutlierFactor (LOF)
- **If there are regions of different density, this approach can have problems**

Relative Density-based: LOF approach

- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors
- Outliers are points with largest LOF value



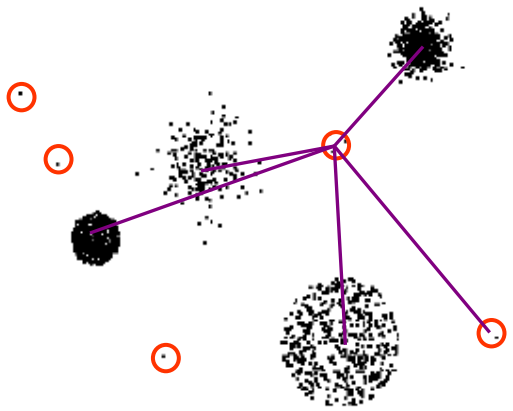
In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Strengths/Weaknesses of Density-Based Approaches

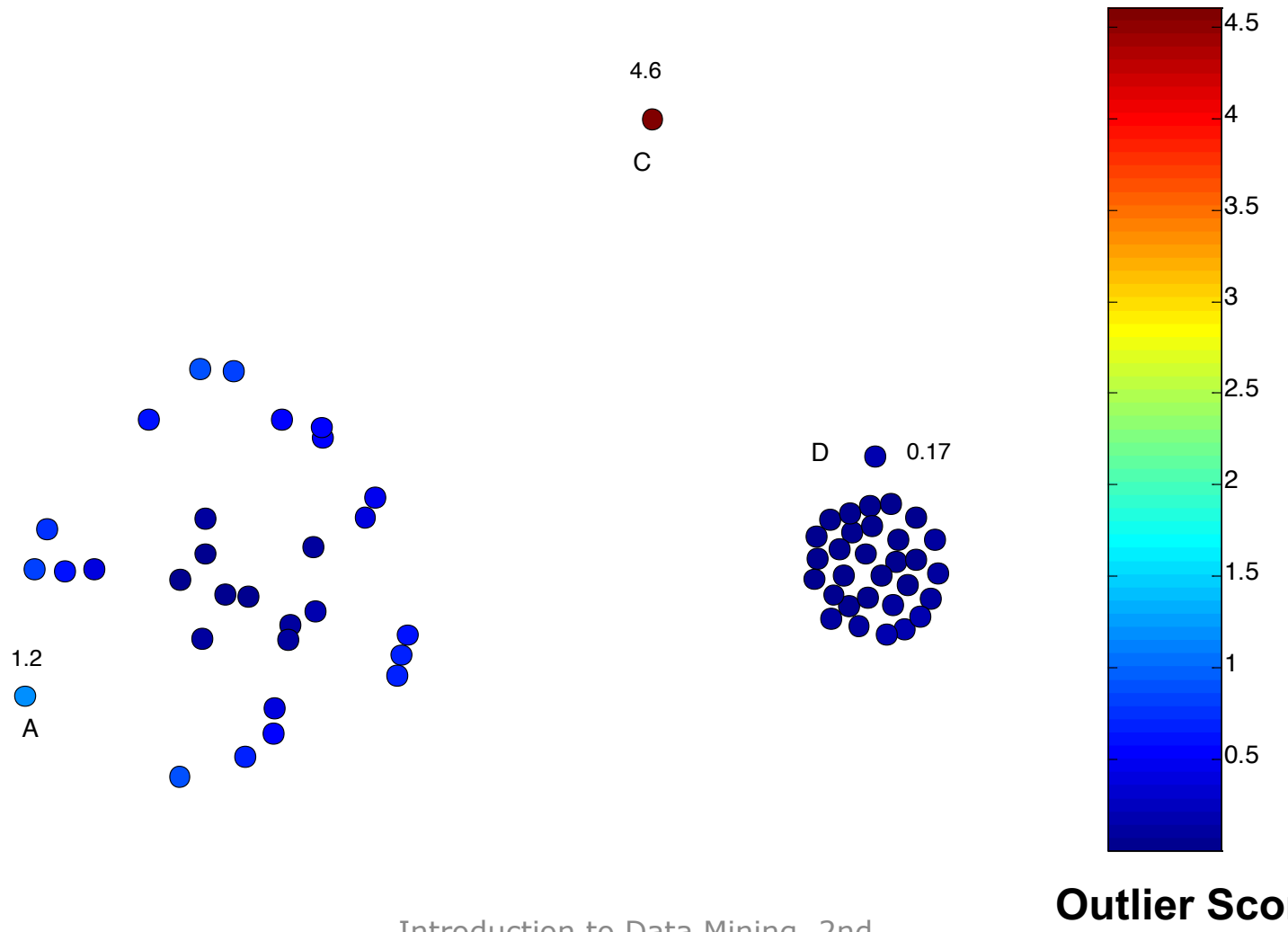
- **Simple**
- **Expensive – $O(n^2)$**
- **Sensitive to parameters**
- **Density becomes less meaningful in high-dimensional space**

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



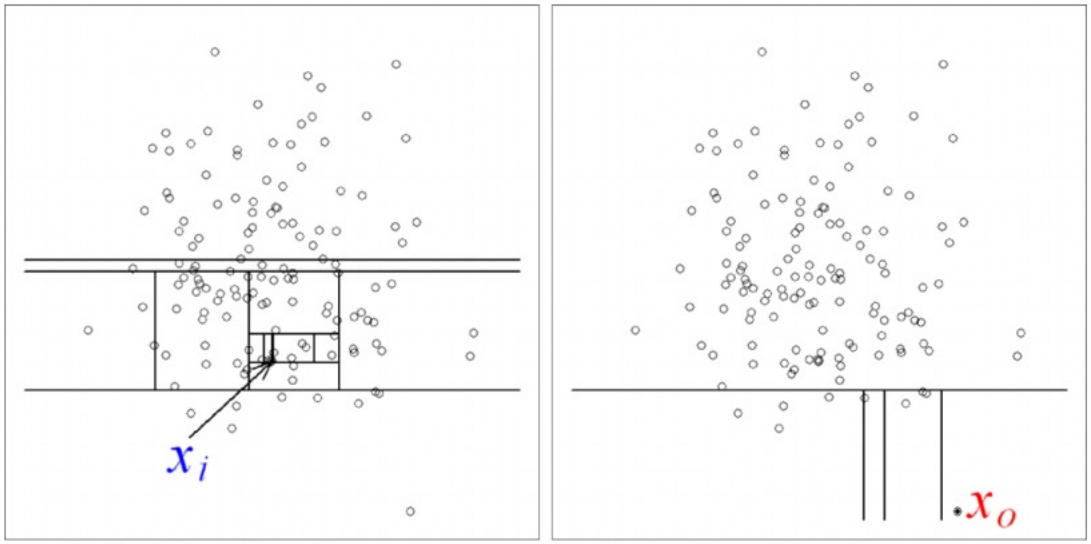
Strengths/Weaknesses of Clustering-Based Approaches

- **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**

Isolation Forest

- **Based on Random Forest**

Liu Tink Zhou icdm2008



(a) Isolating x_i (b) Isolating x_o

Reconstruction-Based Approaches

- **Based on assumptions there are patterns in the distribution of the normal class that can be captured using lower-dimensional representations**
- **Reduce data to lower dimensional data**
  - Can use Principal Components Analysis (PCA) or other dimensionality reduction techniques
  - Can also use neural networks
- **Measure the reconstruction error for each object**
  - The difference between original and reduced dimensionality version

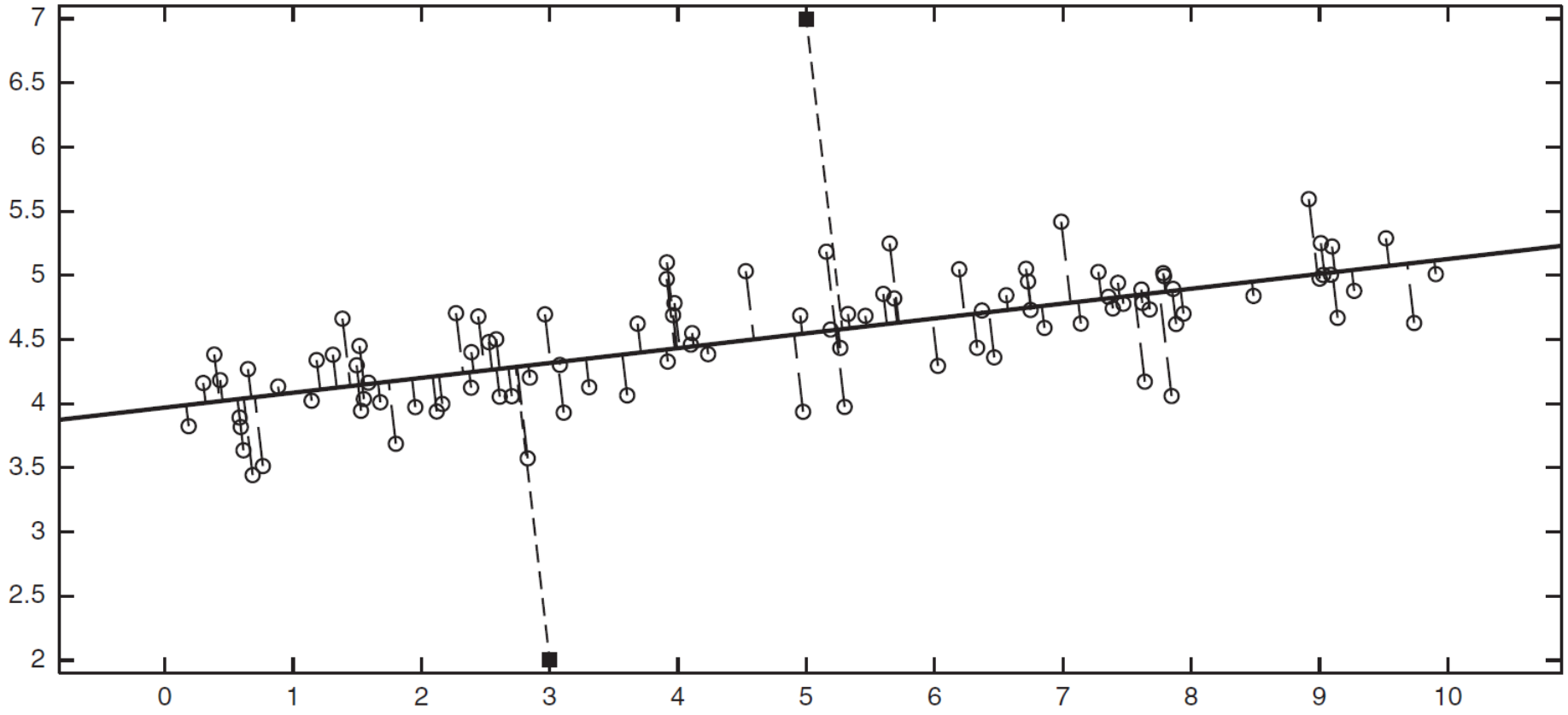
Reconstruction Error

- Let $\mathbf{x}$ be the original data object
- Find the representation of the object in a lower dimensional space
- Project the object back to the original space
- Call this object $\hat{\mathbf{x}}$

$$\text{Reconstruction Error}(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|$$

- Objects with large reconstruction errors are anomalies

Reconstruction of two-dimensional data



Strengths and Weaknesses

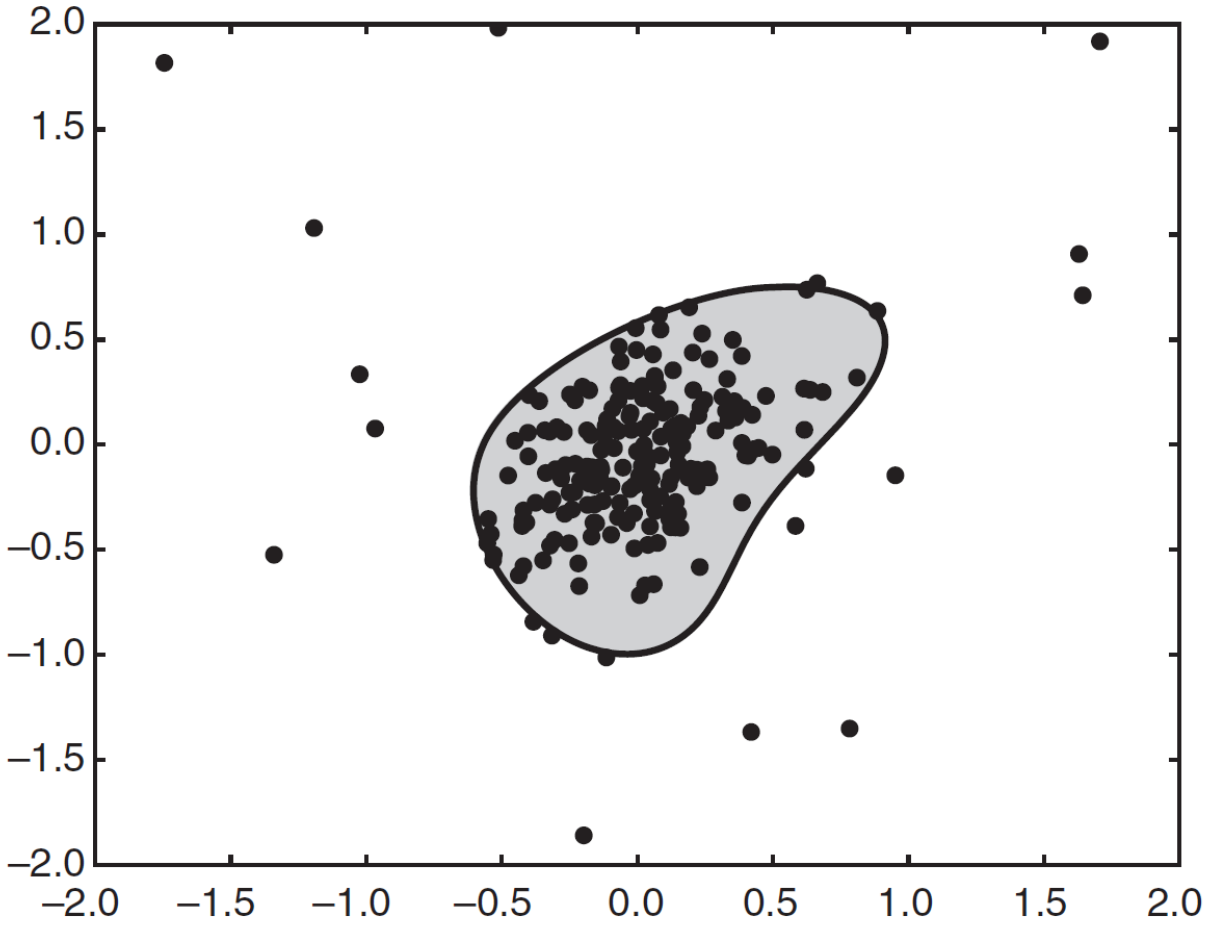
- **Does not require assumptions about distribution of normal class**
- **Can use many dimensionality reduction approaches**
- **The reconstruction error is computed in the original space**
  - This can be a problem if dimensionality is high

One Class SVM

- **Use an SVM approach to classify normal objects**
- **Uses the given data to construct such a model**
- **This data may contain outliers**
- **But the data does not contain class labels**
- **How to build a classifier given one class?**

Finding Outliers with a One-Class SVM

- De



Strengths and Weaknesses

- **Strong theoretical foundation**
- **Choice of ν is difficult**
- **Computationally expensive**

Information Theoretic Approaches

- **Key idea is to measure how much information decreases when you delete an observation**

$$Gain(x) = Info(D) - Info(D \setminus x)$$

- **Anomalies should show higher gain**

- **Normal points should have less gain**

Information Theoretic Example

-

weight	height	Frequency
low	low	20
low	medium	15
medium	medium	40
high	high	20
high	low	5

- **Eliminating last group give a gain of 2.08 – 1.89 = 0.19**

Strengths and Weaknesses

- **Solid theoretical foundation**
- **Theoretically applicable to all kinds of data**
- **Difficult and computationally expensive to implement in practice**

Evaluation of Anomaly Detection

- **If class labels are present, then use standard evaluation approaches for rare class such as precision, recall, or false positive rate**
  - FPR is also know as false alarm rate
- **For unsupervised anomaly detection use measures provided by the anomaly method**
  - Reconstruction error or gain