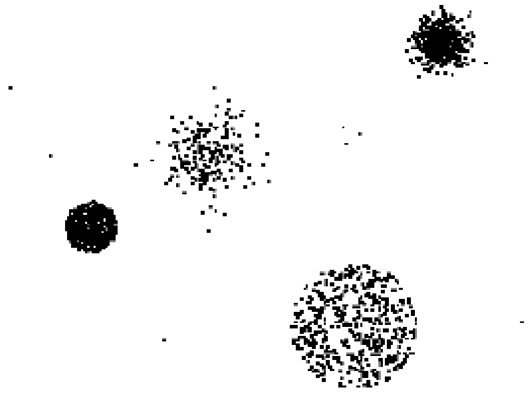
CS5344: Anomaly Detection

Anthony Tung
Department of Computer Science
anthony@comp.nus.edu.sg

Acknowledgement: Modified from Introduction to Data Mining , 2nd Edition By Tan, Steinbach, Kumar

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

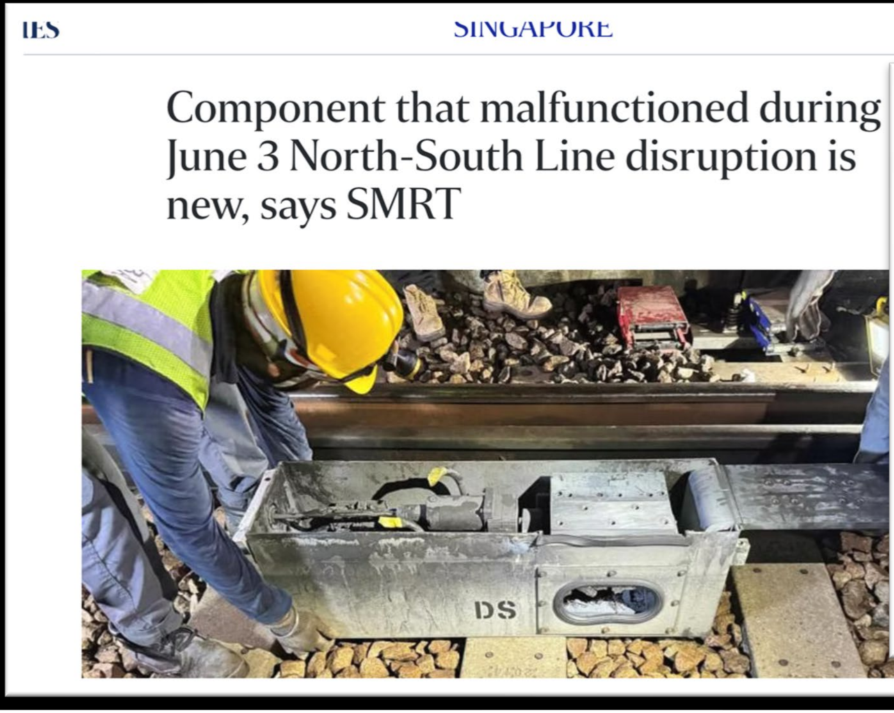
One in a thousand occurs often if you have lots of data

Context is important, e.g., freezing temps in July

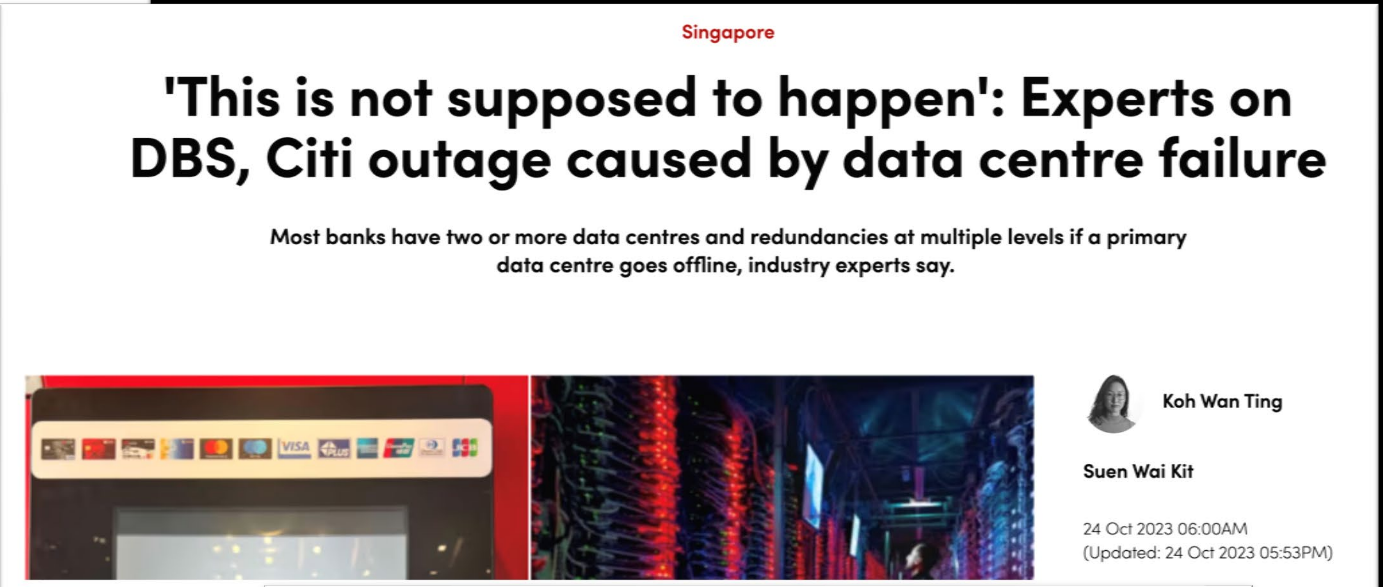
Can be important or a nuisance

Unusually high blood pressure

200 pound, 2 year old



Component that malfunctioned during June 3 North-South Line disruption is new, says SMRT


Singapore

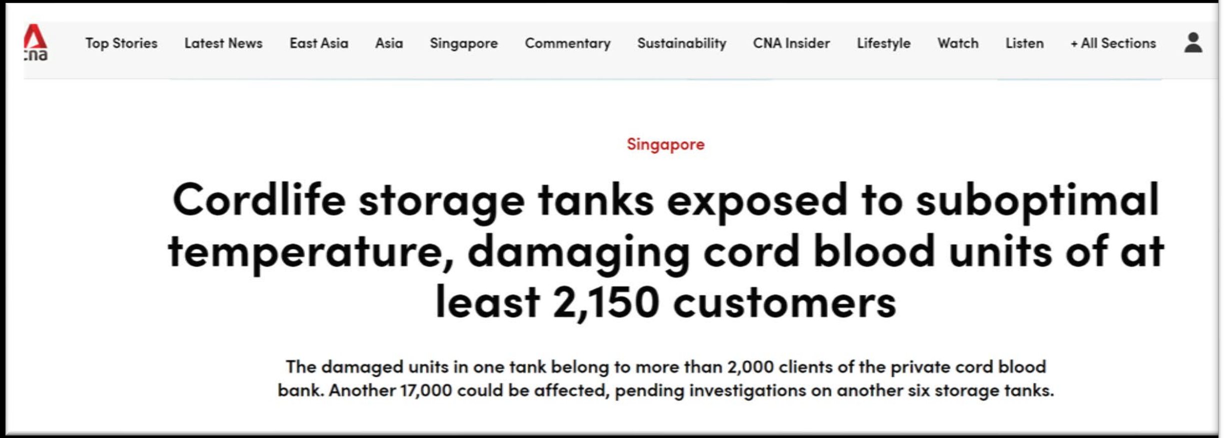
'This is not supposed to happen': Experts on DBS, Citi outage caused by data centre failure

Most banks have two or more data centres and redundancies at multiple levels if a primary data centre goes offline, industry experts say.

Koh Wan Ting

Suen Wai Kit

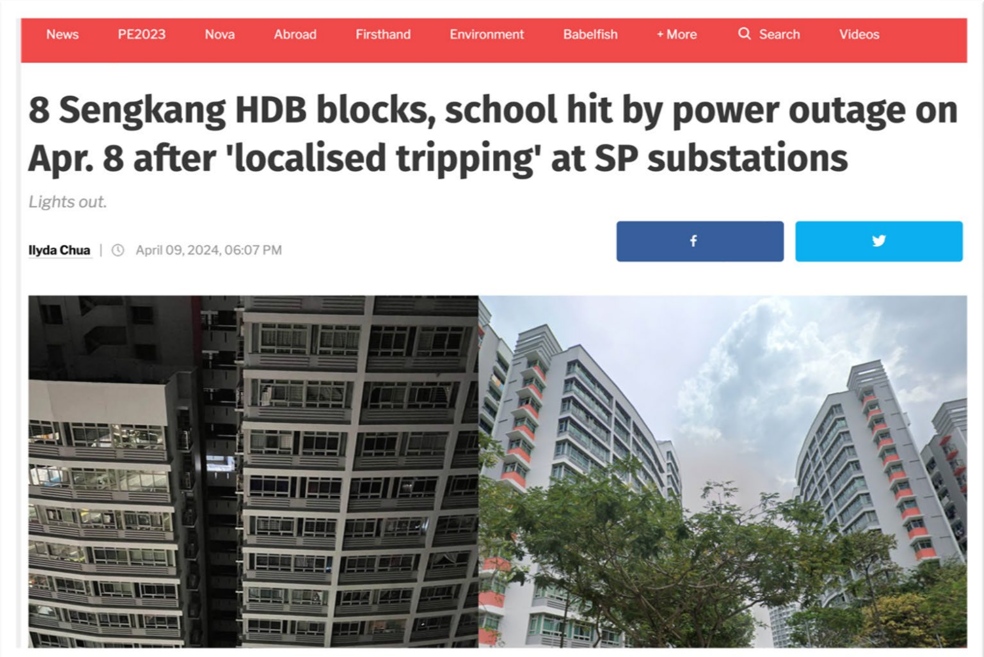
24 Oct 2023 06:00AM
(Updated: 24 Oct 2023 05:53PM)



Singapore

Cordlife storage tanks exposed to suboptimal temperature, damaging cord blood units of at least 2,150 customers

The damaged units in one tank belong to more than 2,000 clients of the private cord blood bank. Another 17,000 could be affected, pending investigations on another six storage tanks.



8 Sengkang HDB blocks, school hit by power outage on Apr. 8 after 'localised tripping' at SP substations

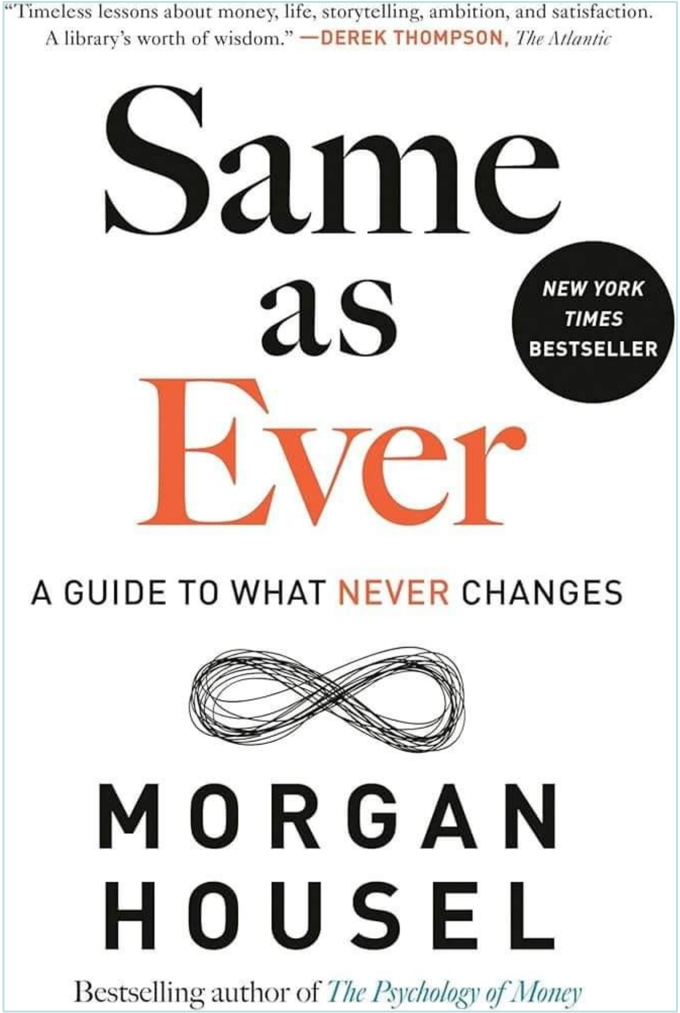
Lights out.

Ilyda Chua | April 09, 2024, 06:07 PM

Surveillance of AI Systems

As AI systems get deployed in manufacturing plants and other domains, errors from these systems will get more costly

Deploying white-box systems to monitor and audit these systems is much cheaper than trying to make these systems "perfect"



"Timeless lessons about money, life, storytelling, ambition, and satisfaction. A library's worth of wisdom." —DEREK THOMPSON, *The Atlantic*

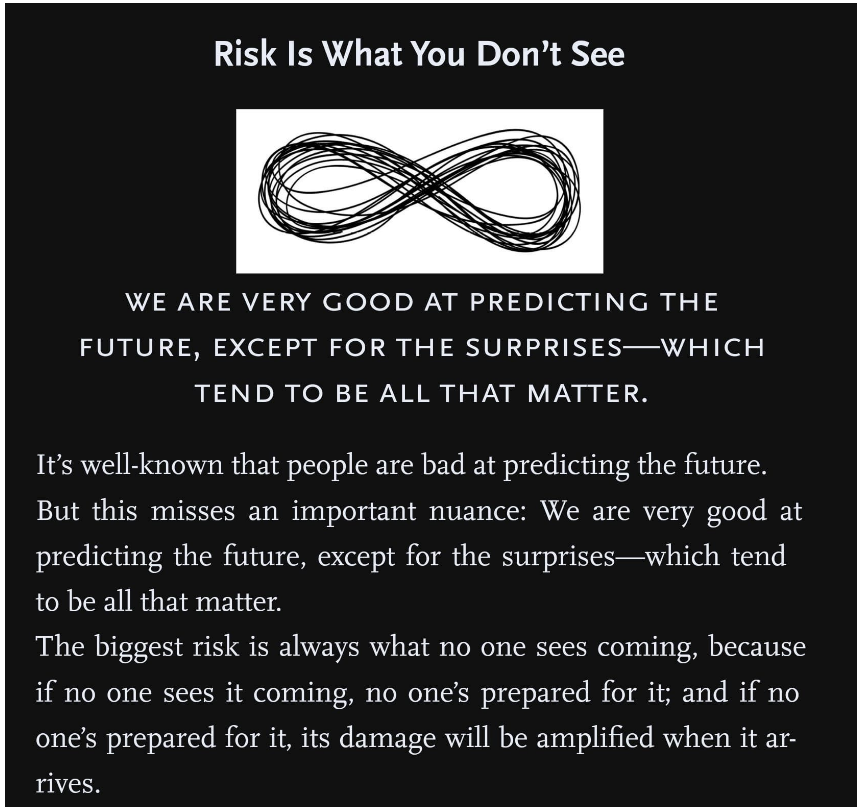
Same as Ever

NEW YORK TIMES BESTSELLER

A GUIDE TO WHAT NEVER CHANGES

MORGAN HOUSEL

Bestselling author of *The Psychology of Money*



Risk Is What You Don't See

WE ARE VERY GOOD AT PREDICTING THE FUTURE, EXCEPT FOR THE SURPRISES—WHICH TEND TO BE ALL THAT MATTER.

It's well-known that people are bad at predicting the future. But this misses an important nuance: We are very good at predicting the future, except for the surprises—which tend to be all that matter.

The biggest risk is always what no one sees coming, because if no one sees it coming, no one's prepared for it; and if no one's prepared for it, its damage will be amplified when it arrives.

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

Distinction Between Noise and Anomalies

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

Model-based vs Model-free

- **Model-based Approaches**
  - Model can be parametric or non-parametric
  - Anomalies are those points that don’t fit well
  - Anomalies are those points that distort the model
- **Model-free Approaches**
  - Anomalies are identified directly from the data without building a model
- **Often the underlying assumption is that the most of the points in the data are normal**

General Issues: Label vs Score

- Some anomaly detection techniques provide only a binary categorization

- Other approaches measure the degree to which an object is an anomaly
  - This allows objects to be ranked
  - Scores can also have associated meaning (e.g., statistical significance)

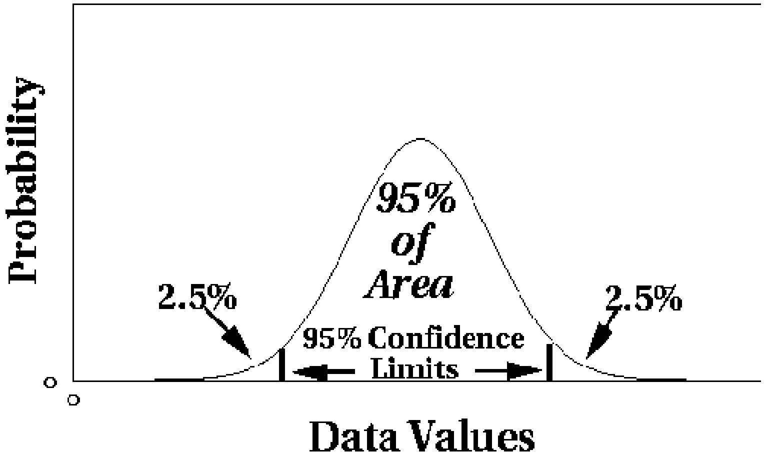
Anomaly Detection Techniques

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

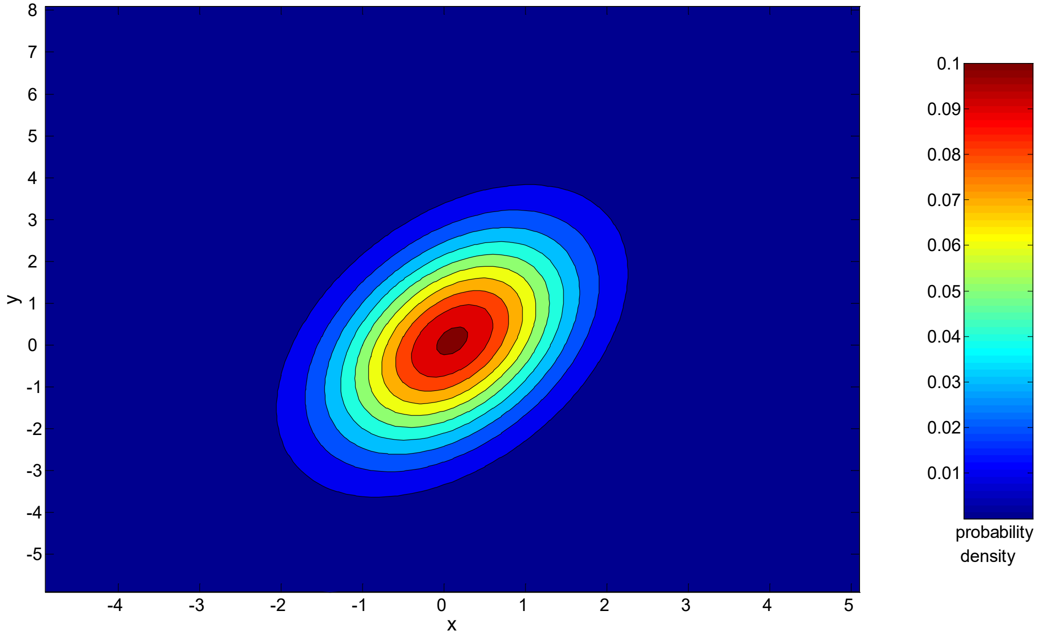
Statistical Approaches

- **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)
- **Issues**
  - Identifying the distribution of a data set
    - Heavy tailed distribution
  - Number of attributes
  - Is the data a mixture of distributions?

Normal Distributions



One-dimensional Gaussian



Two-dimensional Gaussian

Grubbs' Test

Detect outliers in univariate data

Assume data comes from normal distribution

Detects one outlier at a time, remove the outlier, and repeat

- H_0: There is no outlier in data
- H_A: There is at least one outlier

Grubbs' test statistic:

$$G = \frac{\max\left|X - \overline{X}\right|}{s}$$

Reject H_0 if:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N, N-2)}}{N - 2 + t^2_{(\alpha/N, N-2)}}}$$

Statistically-based – Likelihood Approach

- **Assume the data set D contains samples from a mixture of two probability distributions:**
  - M (majority distribution)
  - A (anomalous distribution)
- **General Approach:**
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time t
  - For each point x_t that belongs to M, move it to A
    - Let L_{t+1} (D) be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}$ (D)
    - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistically-based – Likelihood Approach

Data distribution, D = (1 – λ) M + λ A

M is a probability distribution estimated from data

- Can be based on any modeling method (naïve Bayes, maximum entropy, etc.)

A is initially assumed to be uniform distribution

Likelihood at time t:

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

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

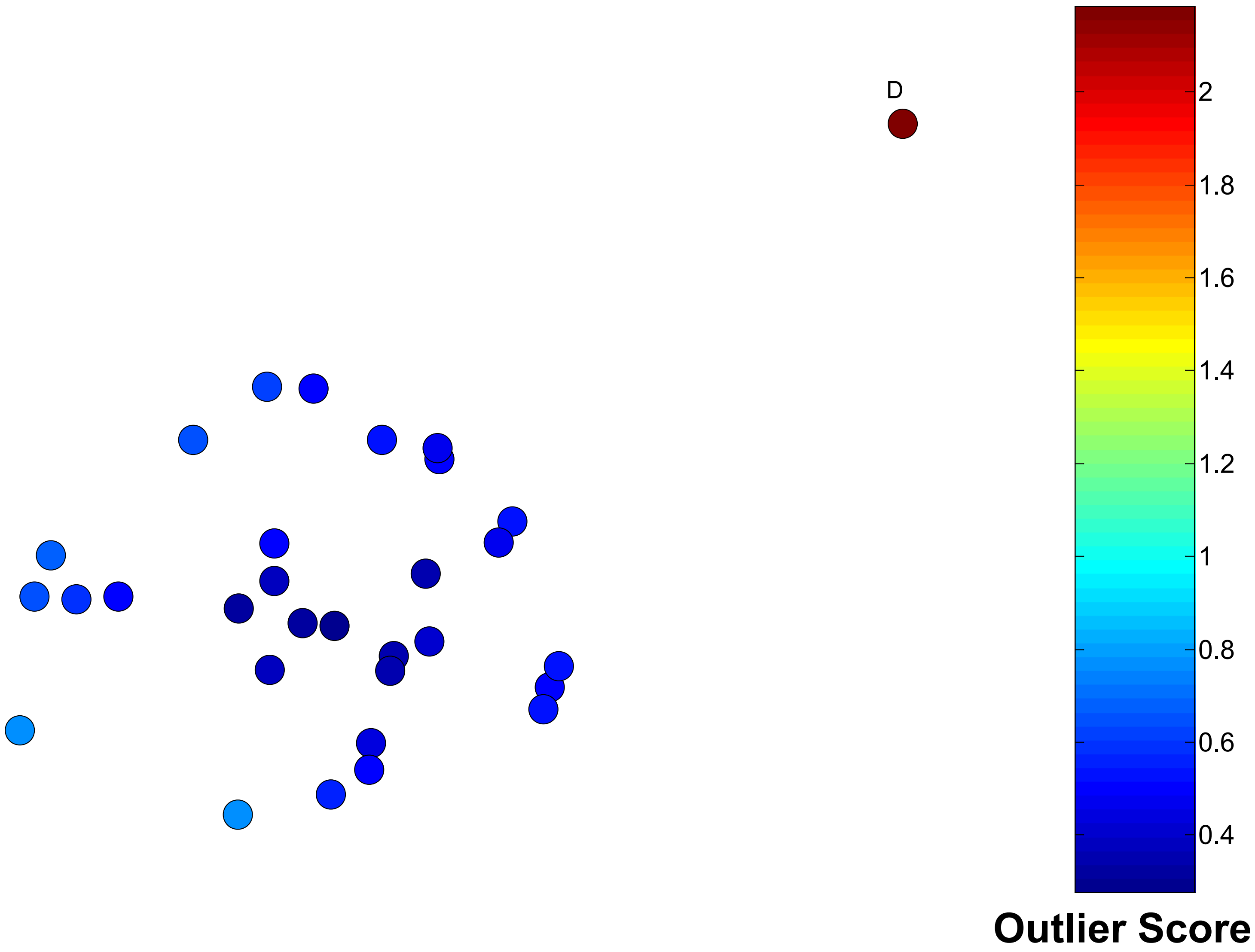
Strengths/Weaknesses of Statistical Approaches

- 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

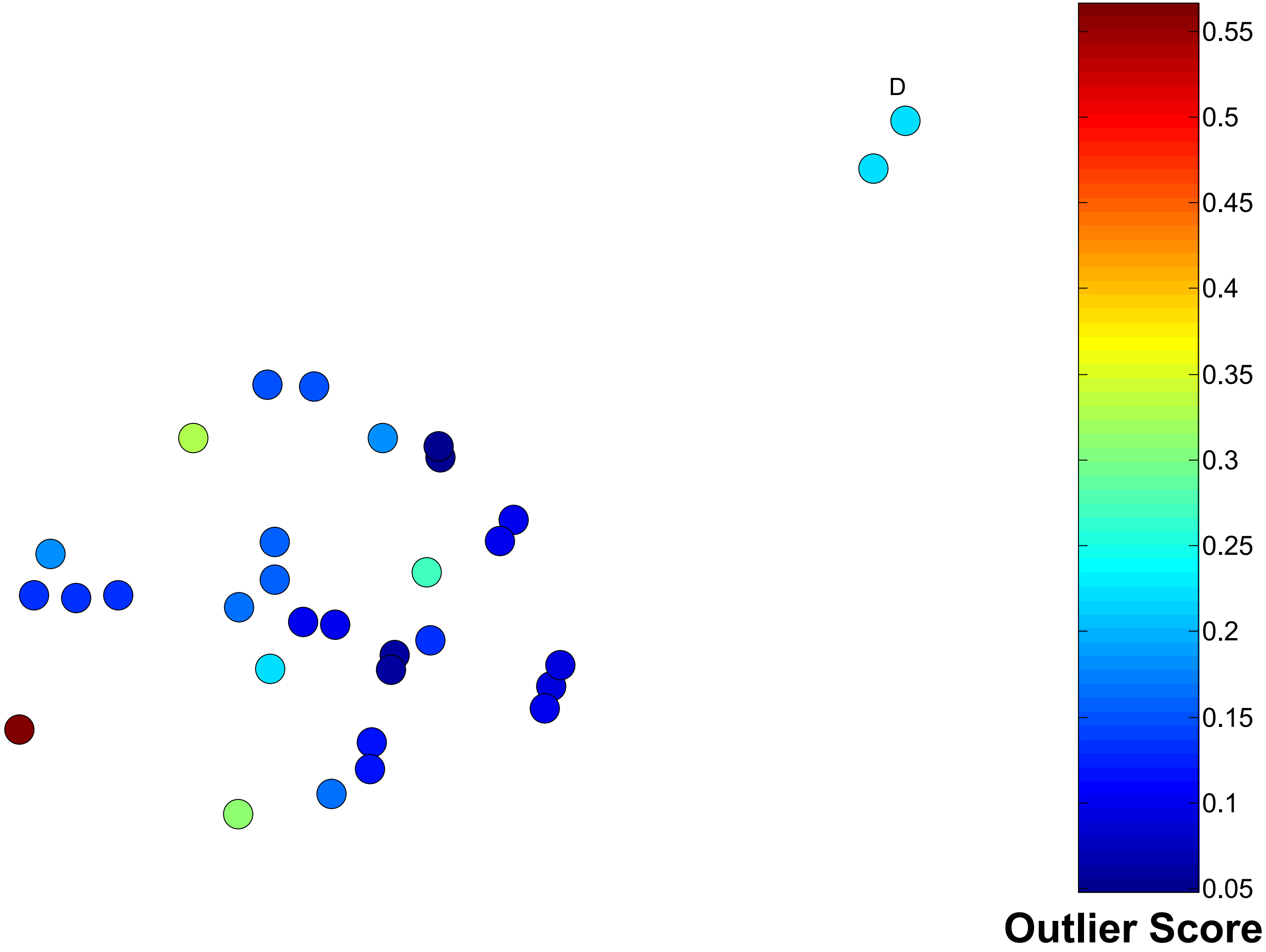
Distance-Based Approaches

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

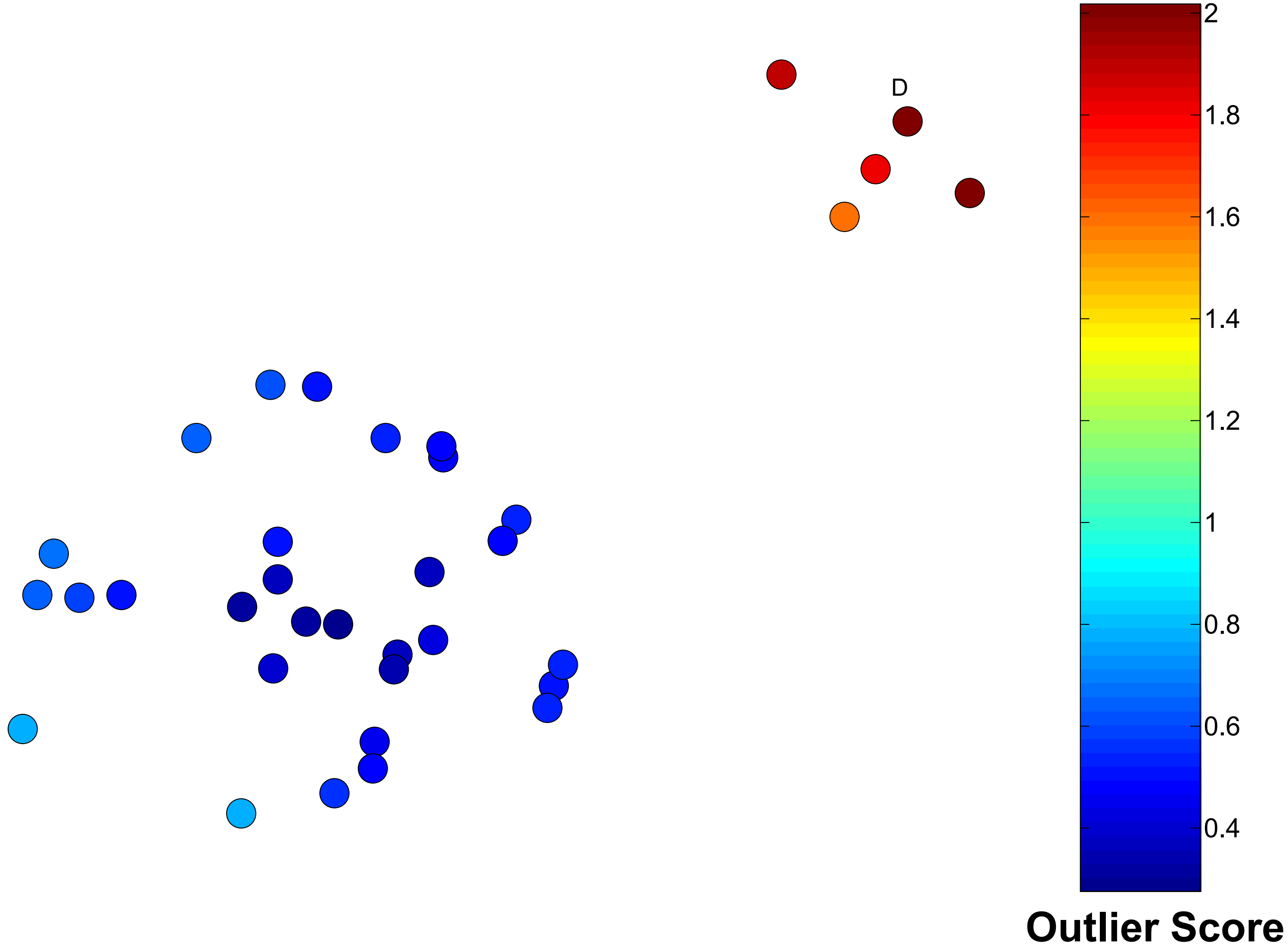
One Nearest Neighbor - One Outlier



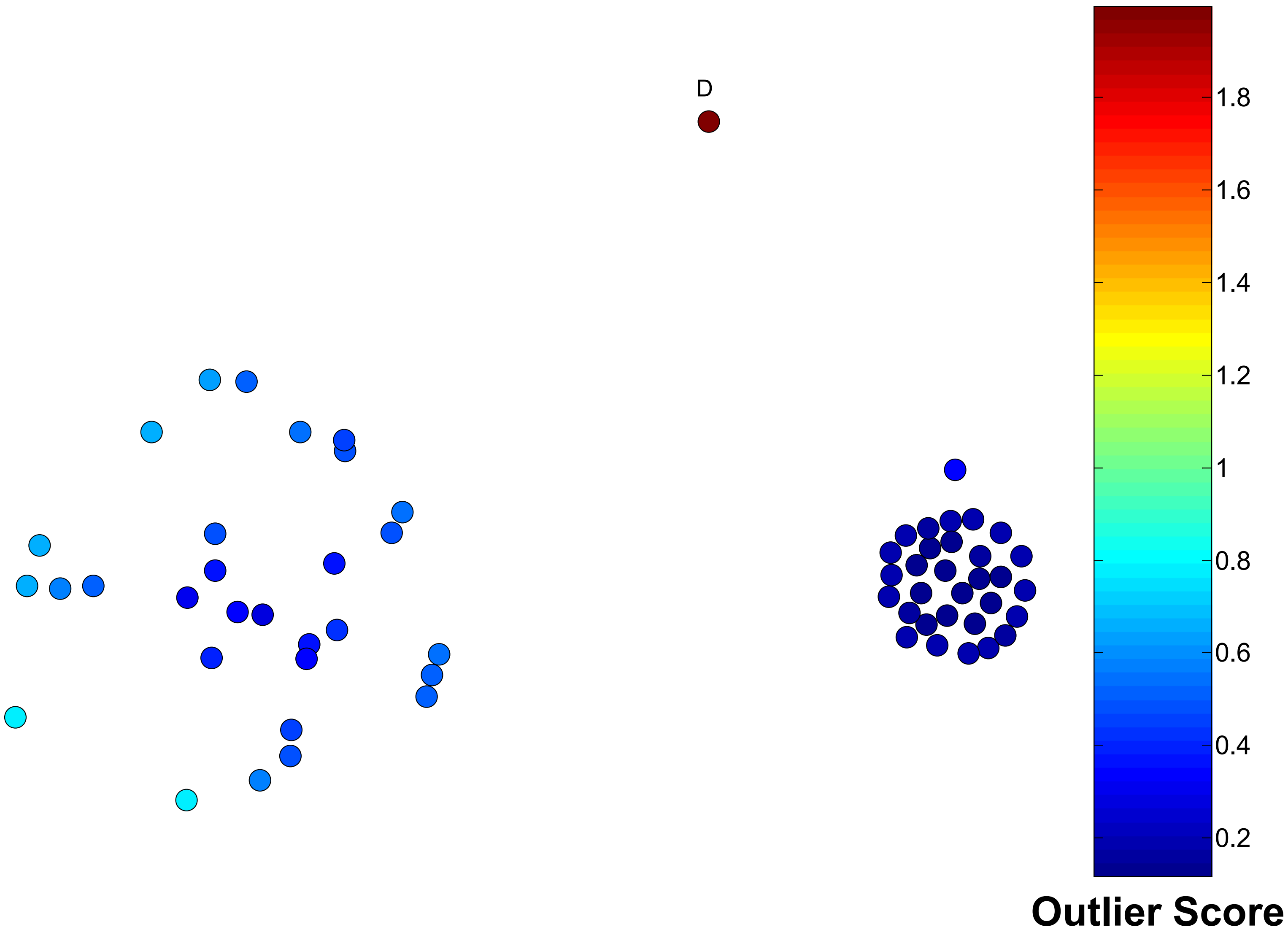
One Nearest Neighbor - Two Outliers



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 kth 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

Let $y_1, \ldots, y_k$ be the k nearest neighbors of x

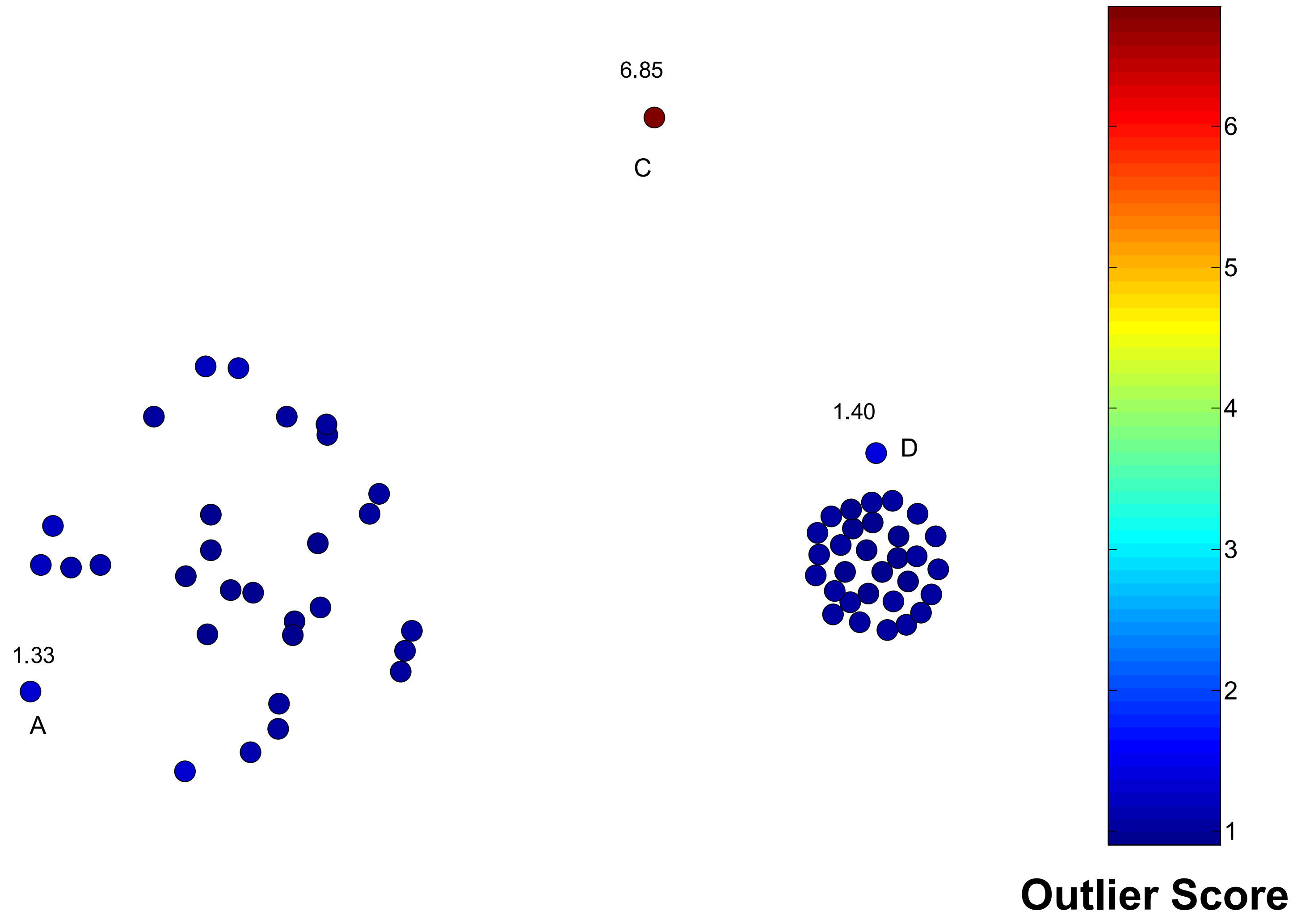
$$density(\boldsymbol{x}, k) = \frac{1}{dist(\boldsymbol{x}, k)} = \frac{1}{dist(\boldsymbol{x}, \boldsymbol{y}_k)}$$

$$relative\ density(\boldsymbol{x}, k) = \frac{\sum_{i=1}^{k} density(\boldsymbol{y}_i, k)/k}{density(\boldsymbol{x}, k)}$$

$$= \frac{dist(\boldsymbol{x}, k)}{\sum_{i=1}^{k} dist(\boldsymbol{y}_i, k)/k} = \frac{dist(\boldsymbol{x}, \boldsymbol{y})}{\sum_{i=1}^{k} dist(\boldsymbol{y}_i, k)/k}$$

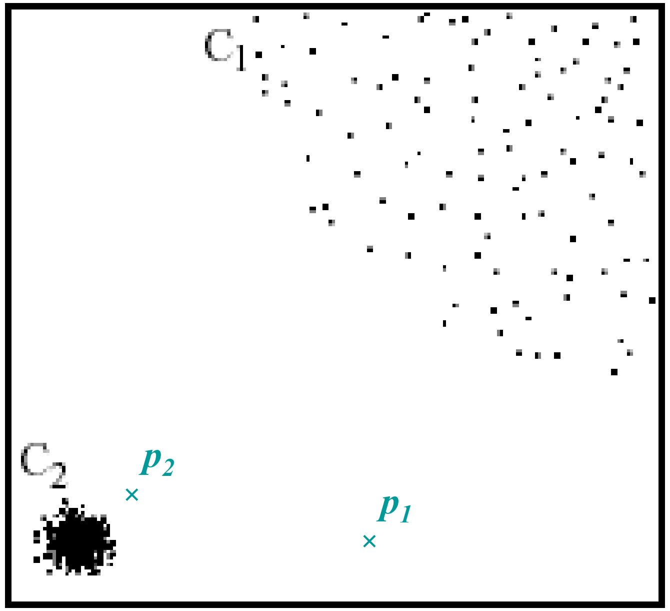
Can use average distance instead

Relative Density Outlier Scores



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

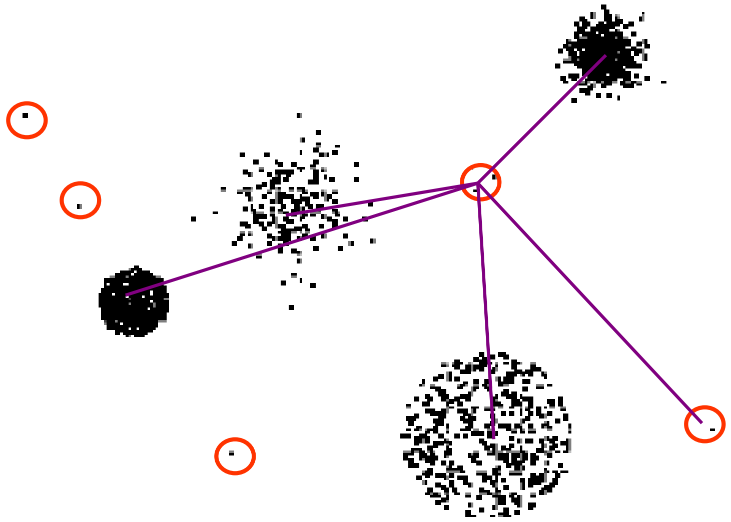
Expensive – O(n^2)

Sensitive to parameters

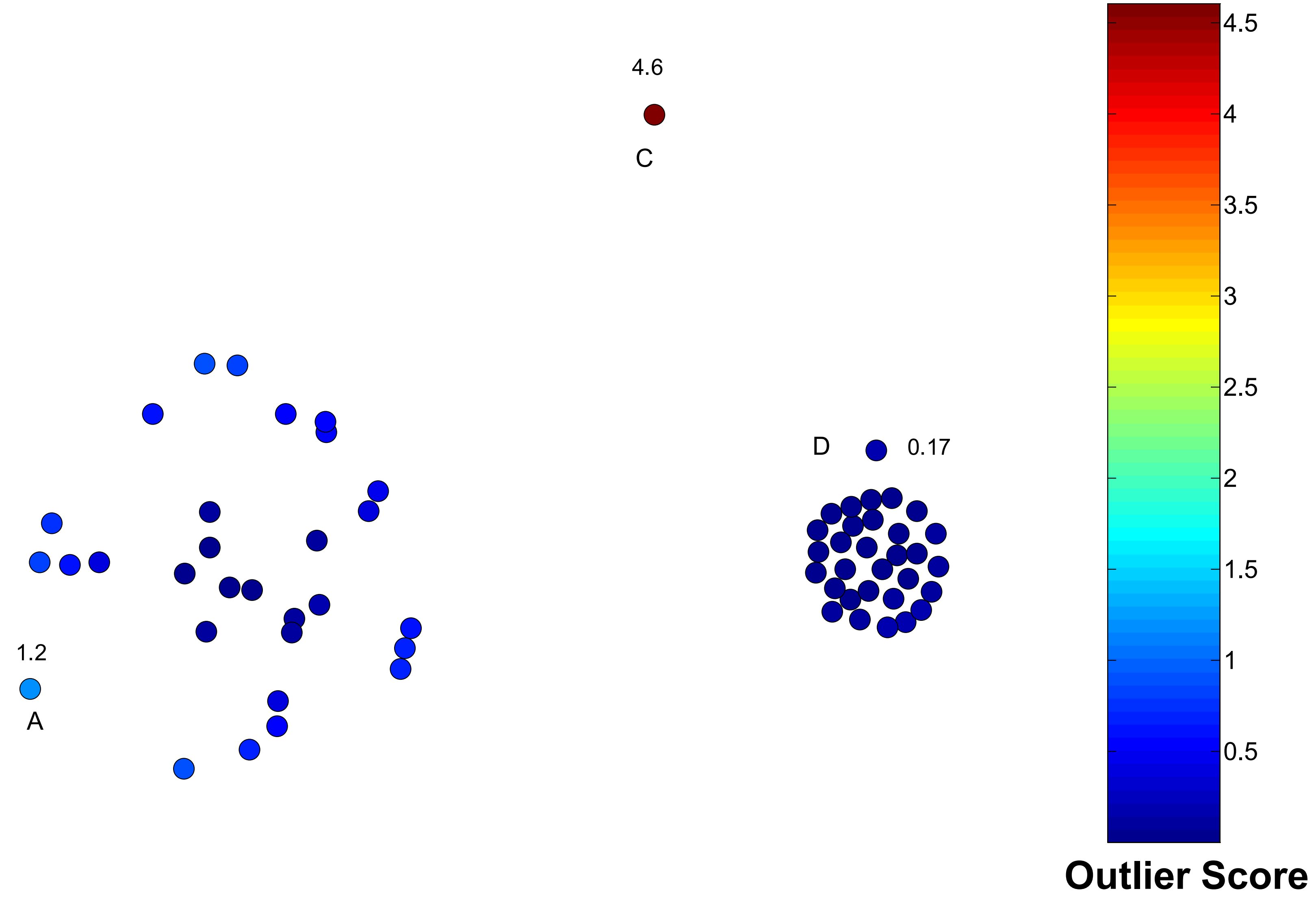
Density becomes less meaningful in high-dimensional space

Clustering-Based Approaches

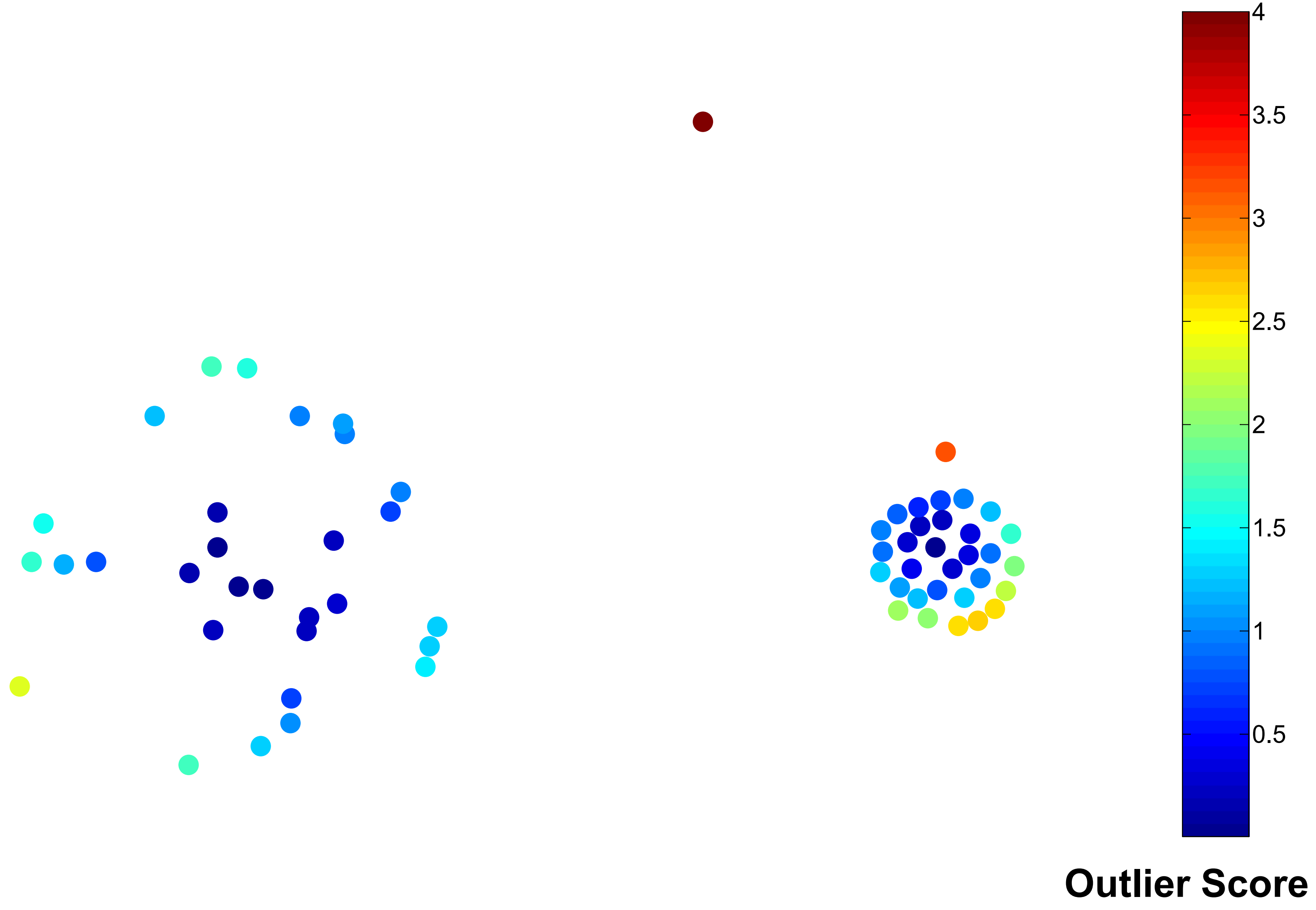
- **An object is a cluster-based outlier if it does not strongly belong to any cluster**
  - For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
    - Outliers can impact the clustering produced
  - For density-based clusters, an object is an outlier if its density is too low
    - Can’t distinguish between noise and outliers
  - For graph-based clusters, an object is an outlier if it is not well connected



Distance of Points from Closest Centroids



Relative Distance of Points from Closest Centroid



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

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

- E.g. Use Principal Components Analysis (PCA) or Auto-encoders

Measure the reconstruction error for each object

- The difference between original and reduced dimensionality version

Reconstruction Error

Let x be the original data object

Find the representation of the object in a lower dimensional space

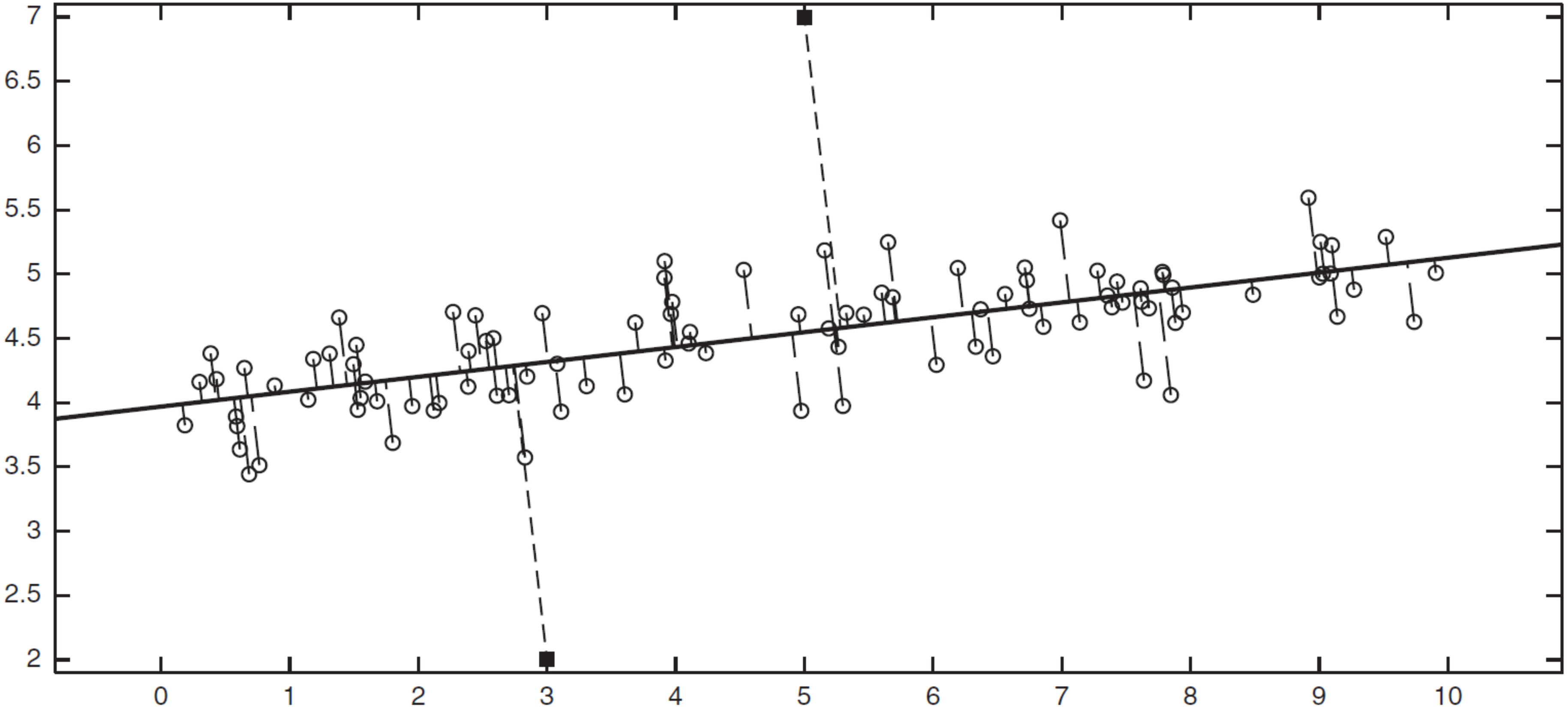
Project the object back to the original space

Call this object $\hat{\mathrm{x}}$

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

Objects with large reconstruction errors are anomalies

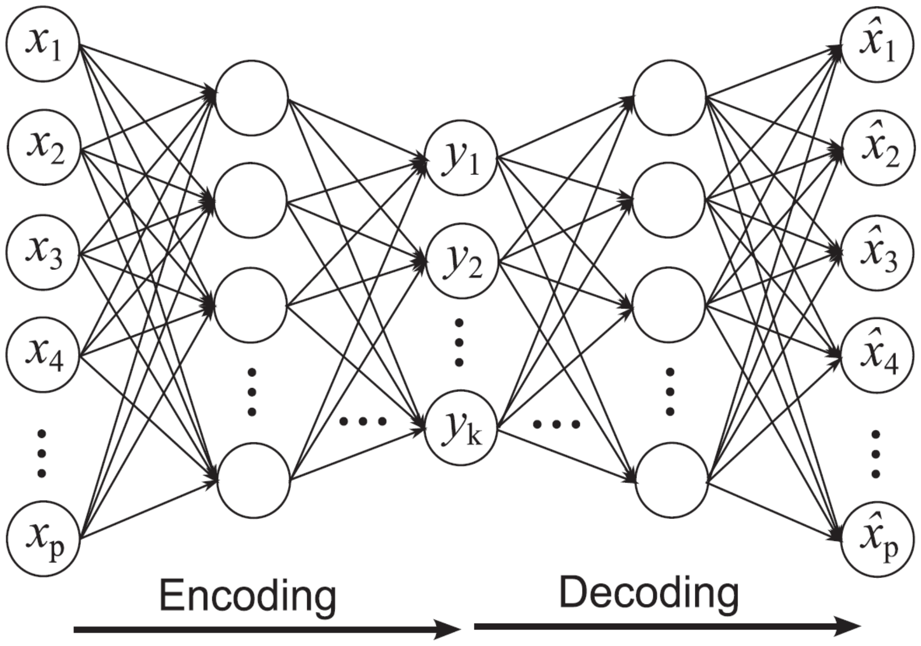
Reconstruction of two-dimensional data



Basic Architecture of an Autoencoder

An autoencoder is a multi-layer neural network

The number of input and output neurons is equal to the number of original attributes.



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

Reference

Yihao Ang, Qiang Huang, Anthony K. H. Tung, Zhiyong Huang. A Stitch in Time Saves Nine: Enabling Early Anomaly Detection with Correlation Analysis. *2023 IEEE 39th International Conference on Data Engineering* (**ICDE 2023**), pp. 1832-1845, Anaheim, CA, USA, April 3-7, 2023. [code][slides][poster][video][bibtex]
(Calling all industries! We're ***seeking collaboration for sensor time series anomaly detection***. Please drop us an email if you are interested.)