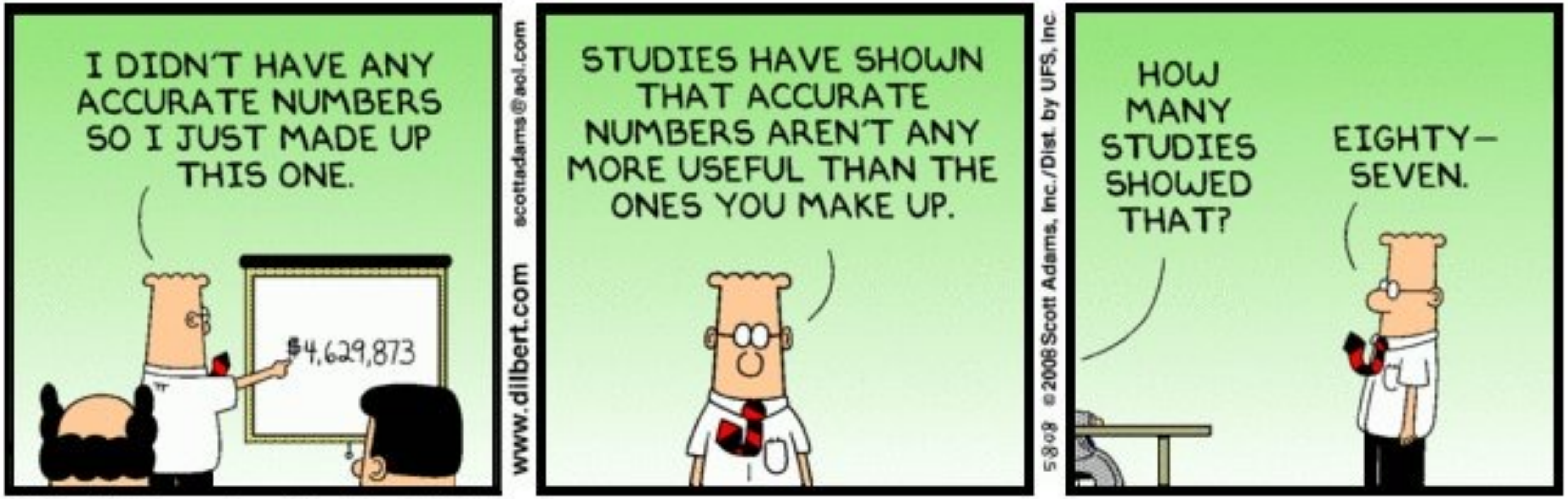


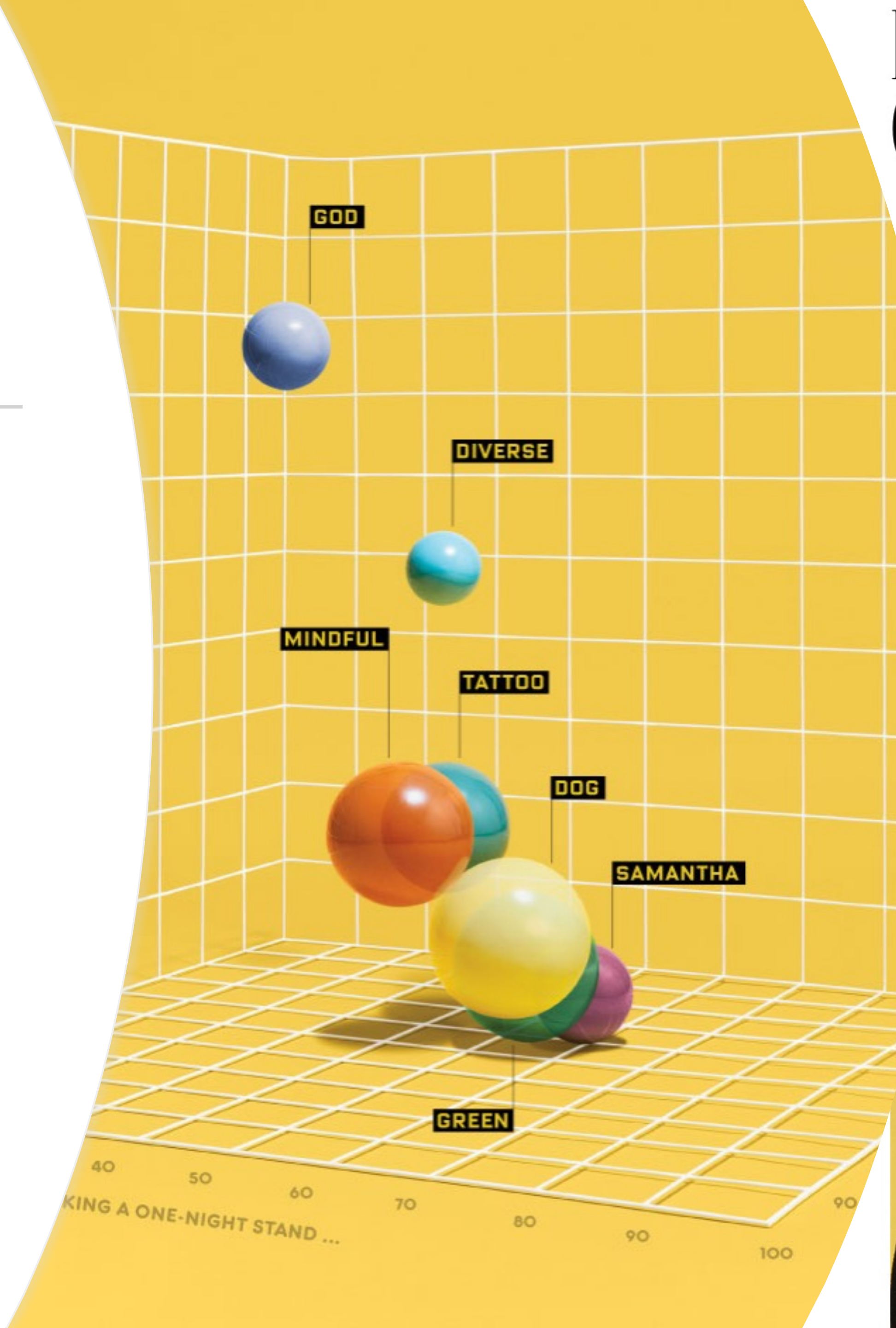
I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
scottadams@aol.com
www.dilbert.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
©2008 Scott Adams, Inc./Dist. by UFS, Inc.
5-8-08
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

HOW A MATH GENIUS HACKED OKCUPID TO FIND TRUE LOVE

- "I think that what I did is just a slightly more algorithmic, large-scale, and machine-learning-based version of what everyone does on the site," McKinlay says. Everyone tries to create an optimal profile—he just had the data to engineer one.

- "It's not like, we matched and therefore we have a great relationship," McKinlay agrees. "It was just a mechanism to put us in the same room. I was able to use OkCupid to find someone."

- https://www.wired.com/2014/01/how-to-hack-okcupid/



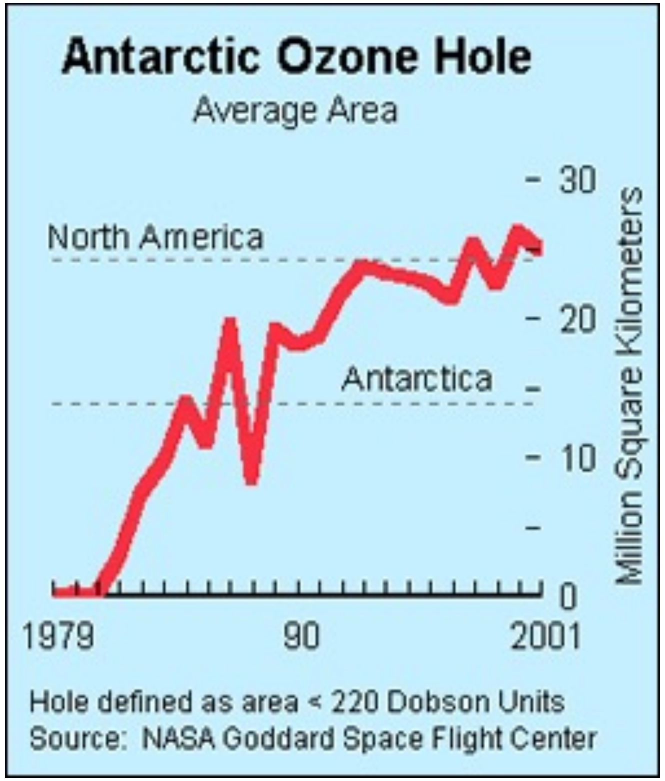
Importance of Anomaly Detection

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

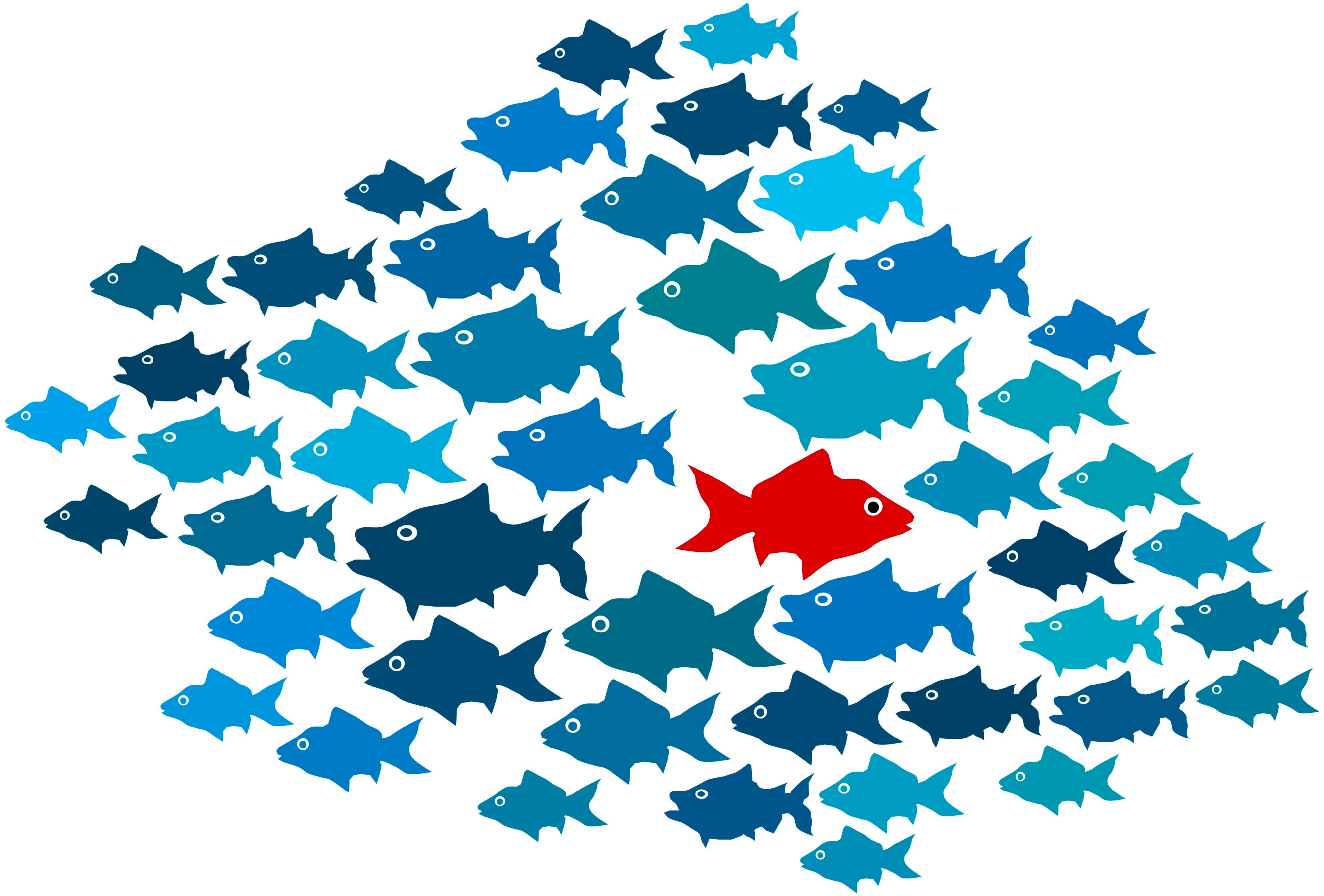
- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?

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



Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

Anomaly Detection



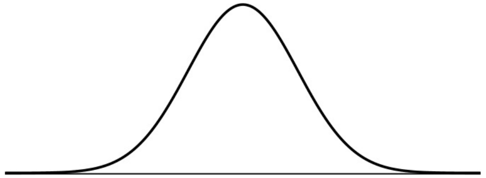
Anomalies

- Data from different classes

  “An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism.”

  –Douglas Hawkins, statistician

- Natural Variation



- Errors

Issues to Consider

- How many attributes define an anomaly?
- Global vs local perspective
- Degree to which a point is an anomaly
- Identifying one at a time vs many at once
    - Masking & Swamping

Supervised vs Unsupervised

- Supervised: Requires a labeled training set with both normal and anomalous objects. Note that there may be more than one normal or anomalous class.
- Unsupervised: Detect which objects are anomalous based on the data alone. Note that the presence of many anomalies that are similar to each other can cause them all to be labeled normal.
- Semi-supervised: Sometimes training data contains labeled normal data, but has no anomalous objects. Model the normal data. Anything not adhering to that model is anomalous. (Also called Novelty Detection.)

Statistical Approaches

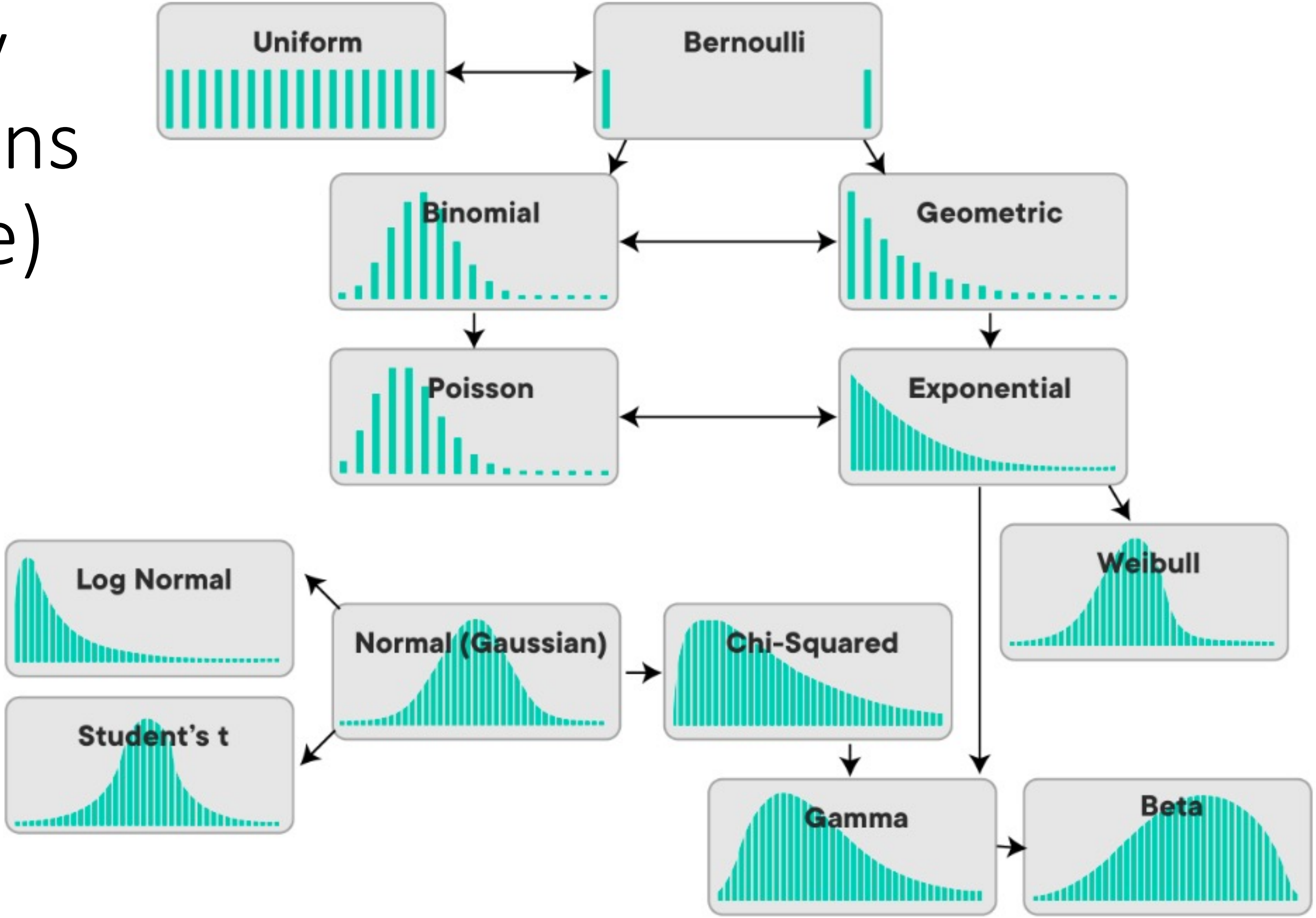
- Model-based: create a model for the data, evaluate objects based on their probability under the model (parametric or non-parametric)

- Probabilistic Definition of an Outlier (discordant observation):

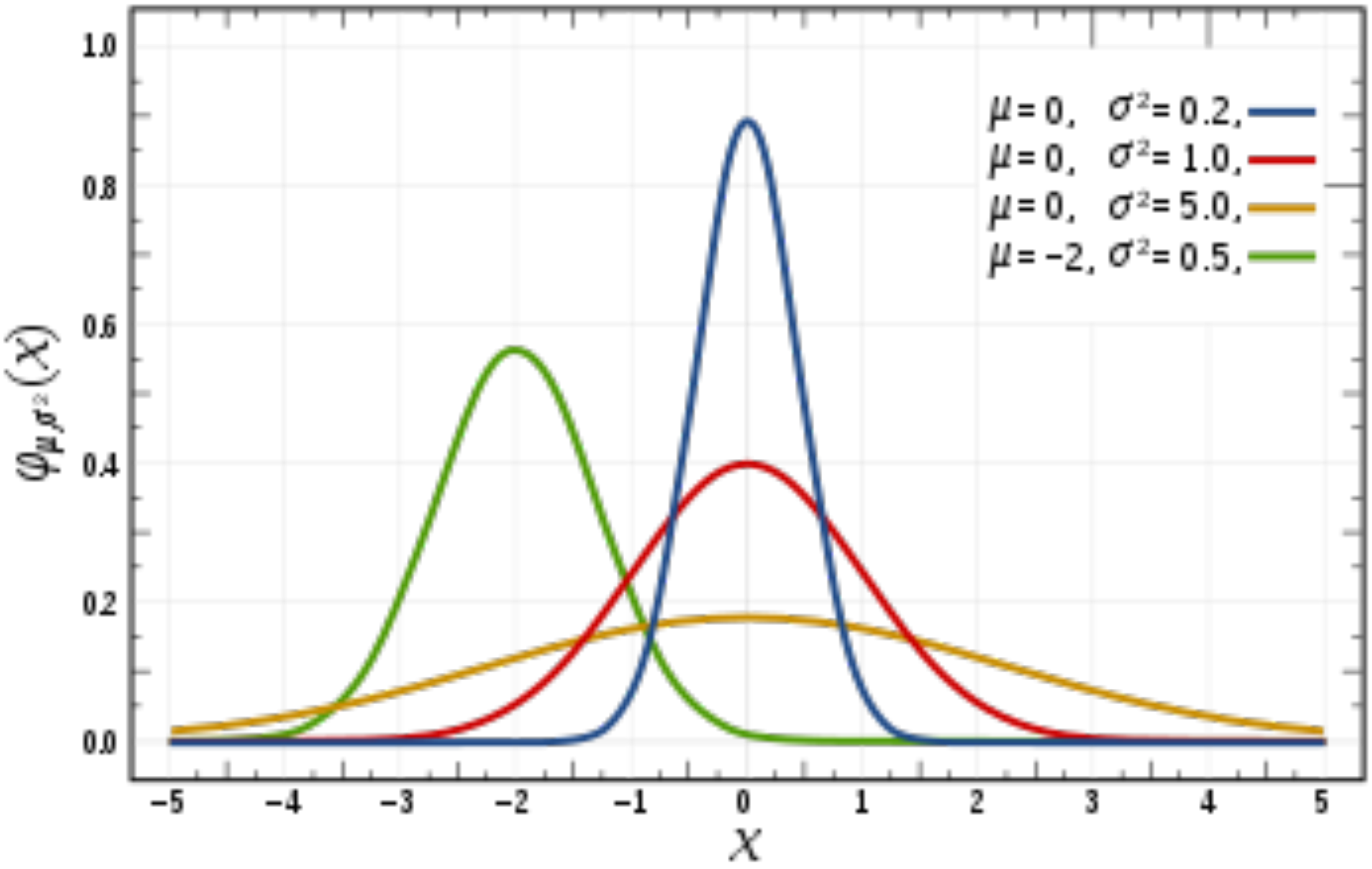
  An outlier is an object that has a low probability with respect to a probability distribution model of the data

- Most statistical outlier detection techniques are only univariate or bivariate

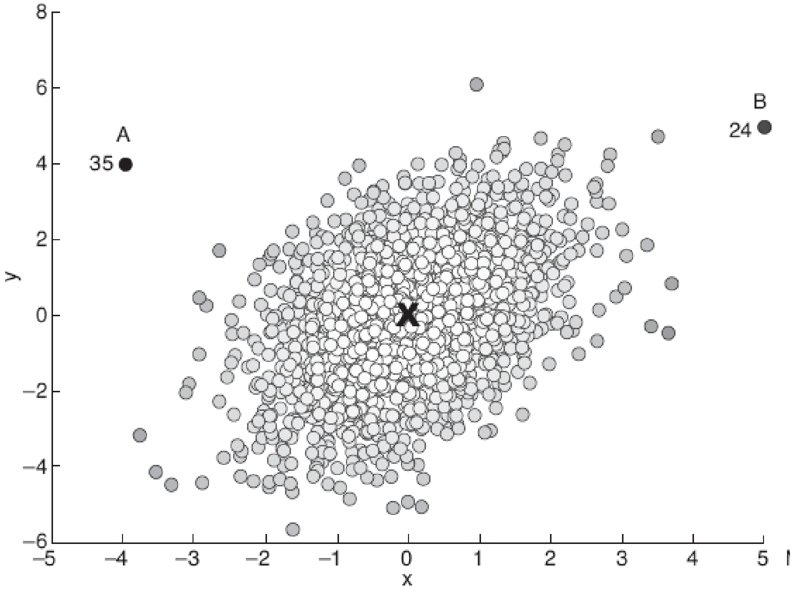
Probability Distributions (Univariate)



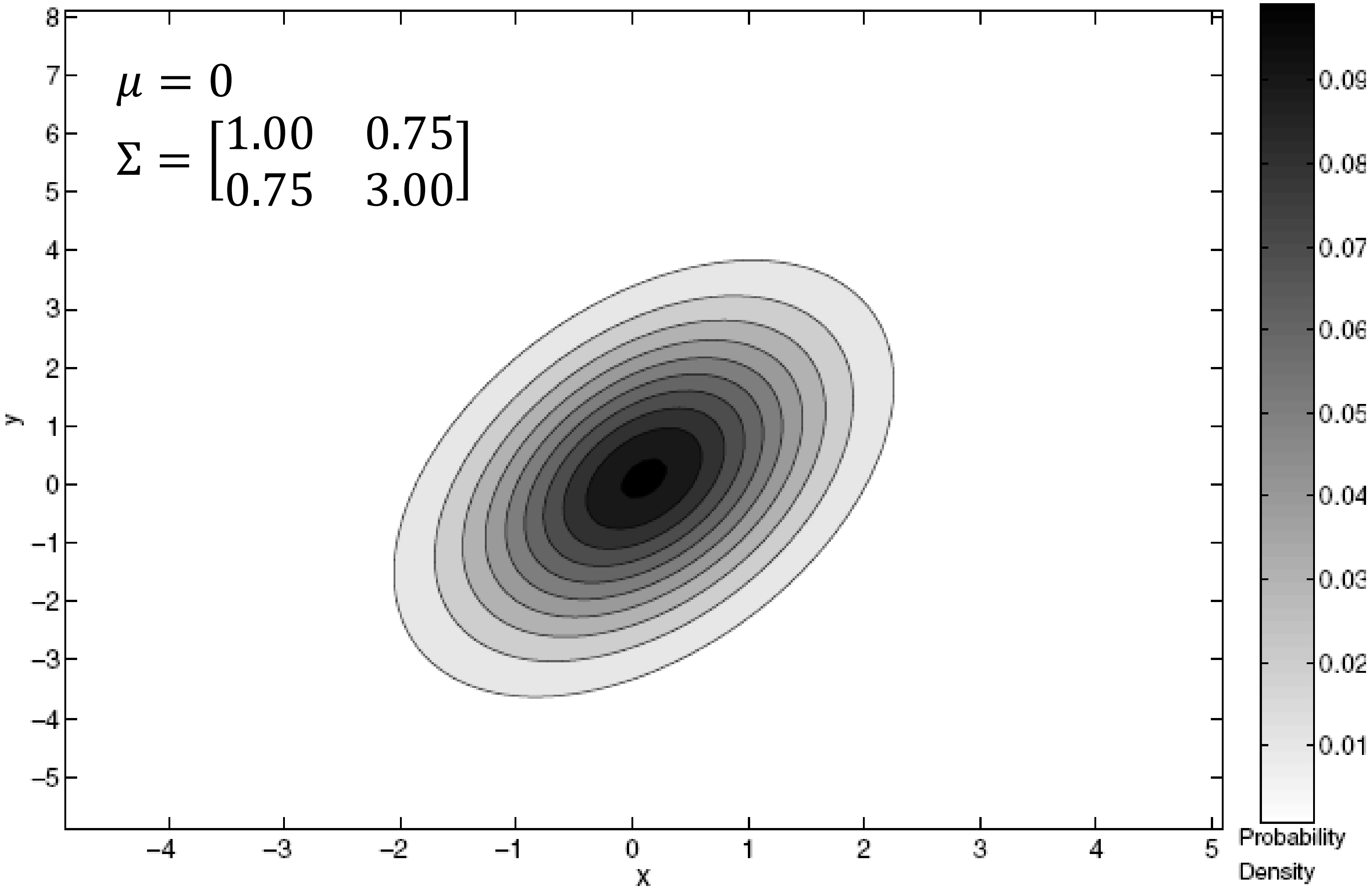
Normal Distribution



Multivariate Gaussian Distribution



Original data



$$\mu = 0$$

$$\Sigma = \begin{bmatrix} 1.00 & 0.75 \\ 0.75 & 3.00 \end{bmatrix}$$

Probability distribution

Mixture Models

- Assumes the data is a mixture of two distributions: one for normal data points, and one for outliers
- Find the parameters of both distributions that maximize the likelihood of the data
- Define a distribution for all the data, then iteratively move data points from the normal set to the outlier set. Update the two distributions at each step.

Characteristics of Statistical Approaches

- When there is sufficient knowledge of the data, and type of distribution that should be applied, these tests have a firm statistical foundations and are very accurate and effective
- There are a wide variety of statistical outlier tests for single attributes
- Fewer options are available for multivariate tests
- Statistical tests don’t really work in high dimensions

Proximity Based Approaches

- An object is an outlier if it is distant from most other points

- The top n data points whose distance to the kth nearest neighbor is greatest
- OR: The top n data points whose average distance to the k nearest neighbors is greatest

Sensitive to the value of k

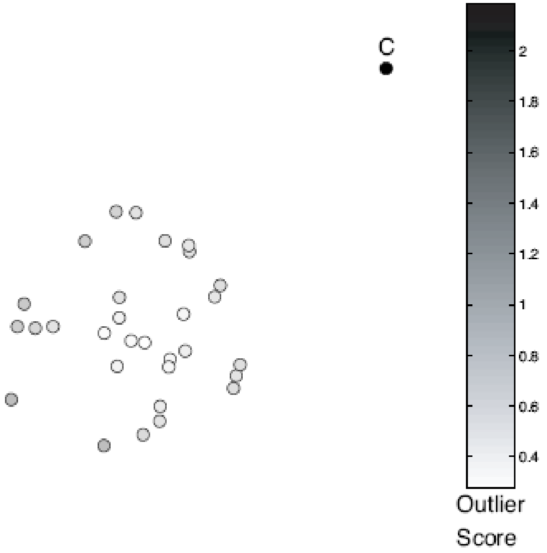


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

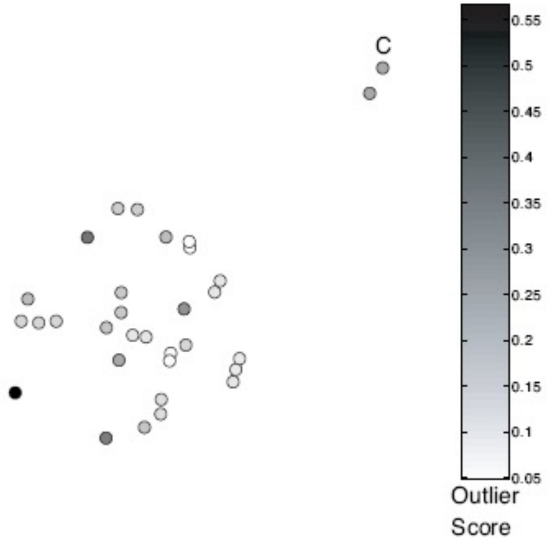


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

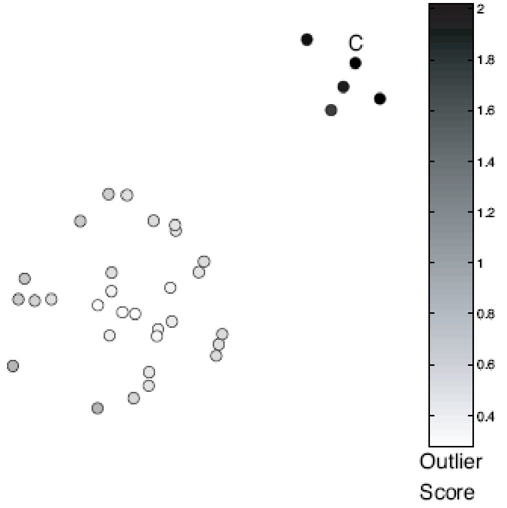


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

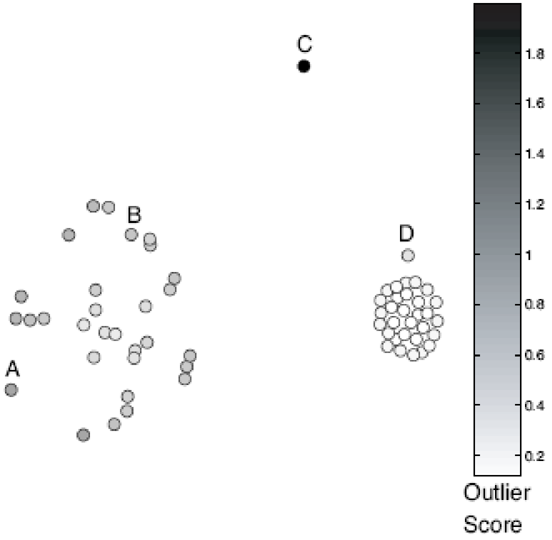
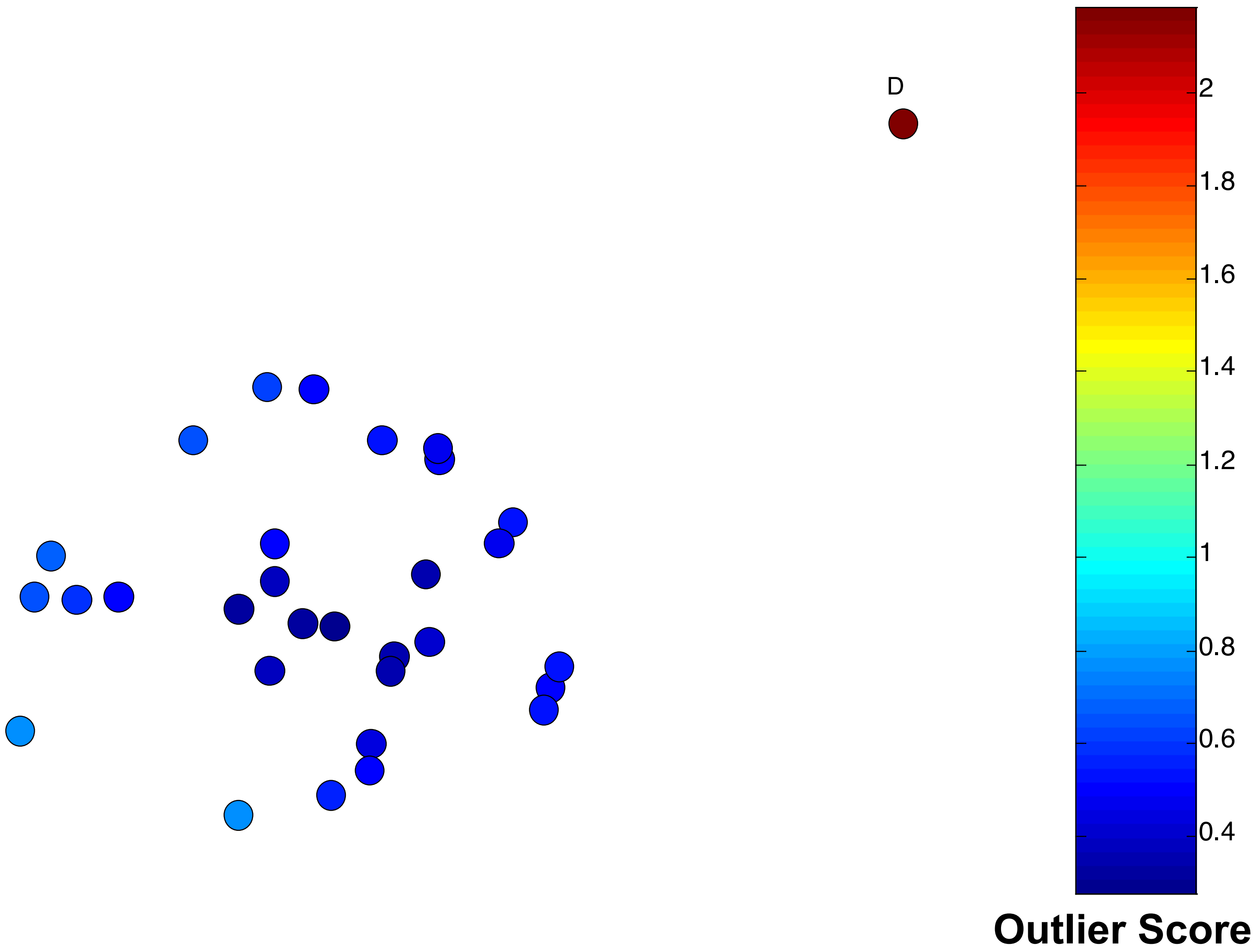
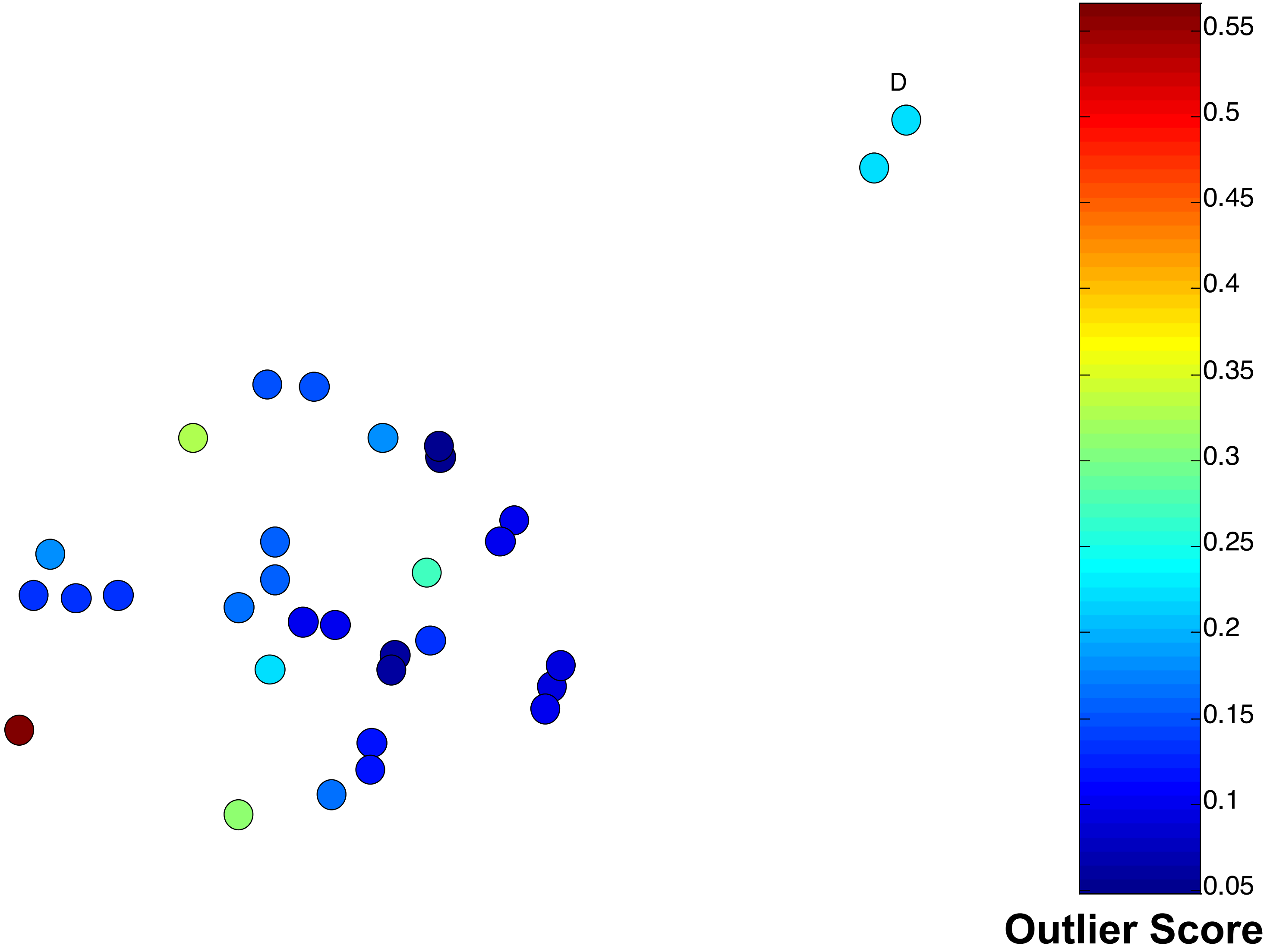


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

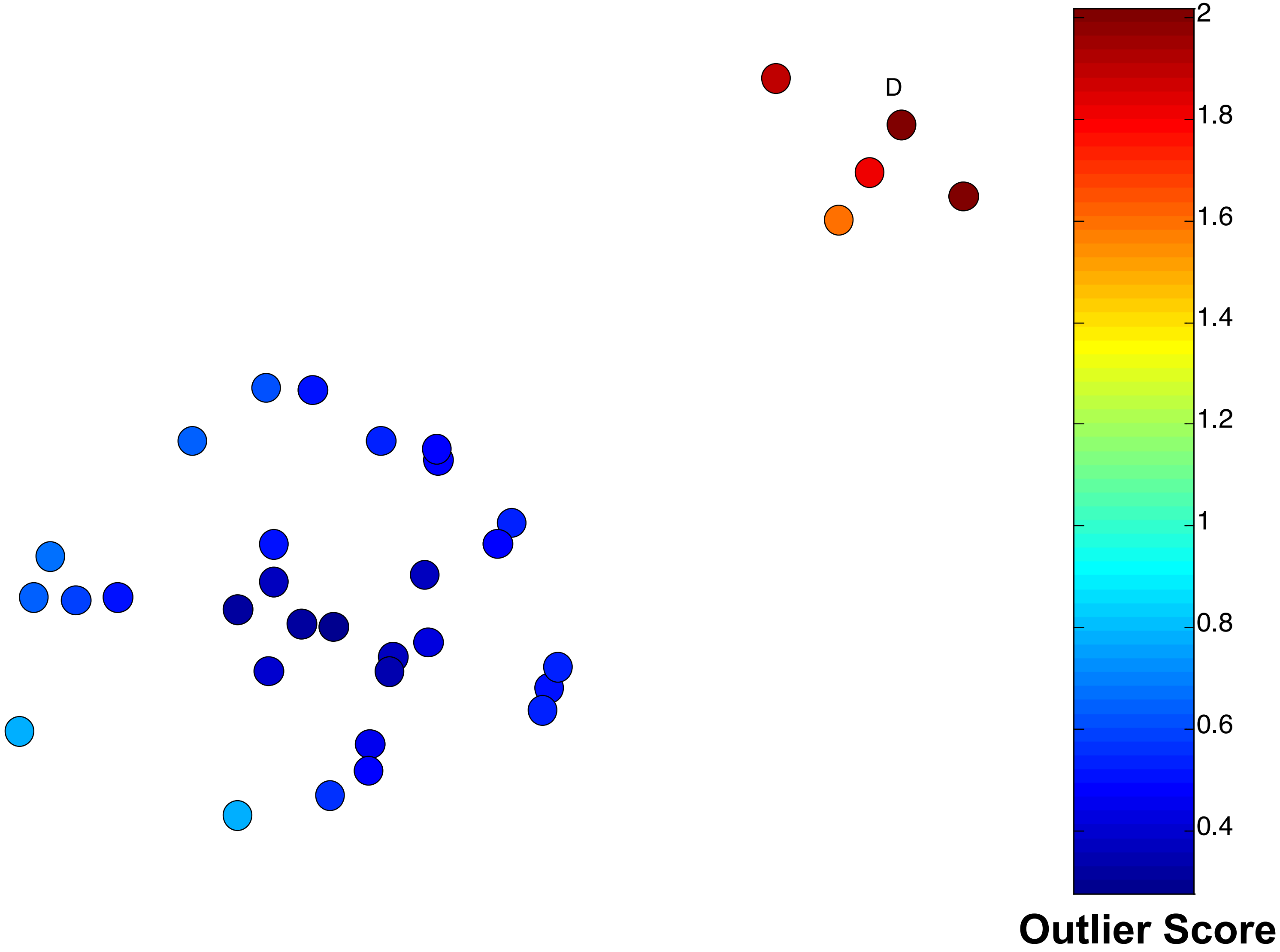
One Nearest Neighbor - One Outlier



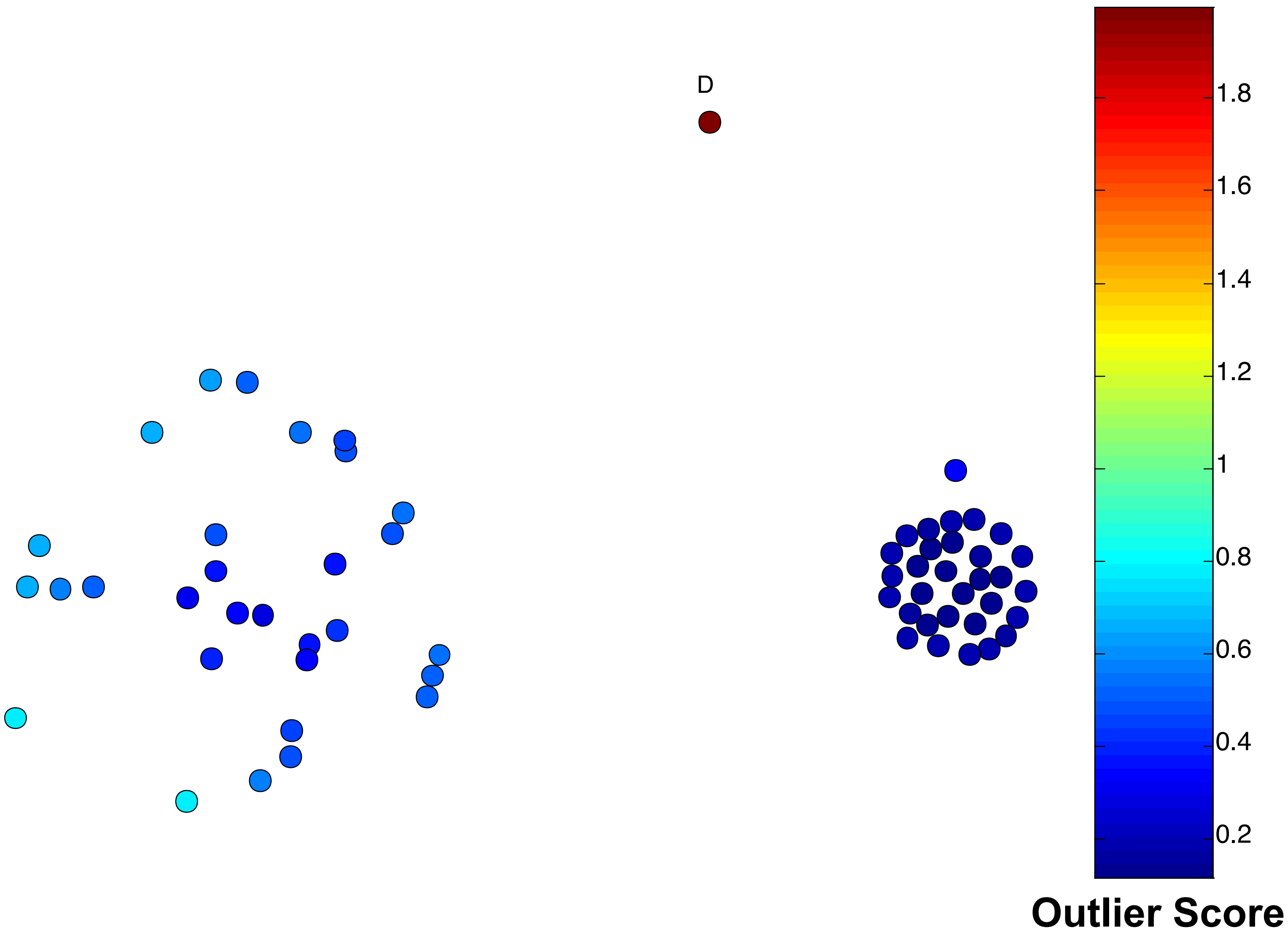
One Nearest Neighbor - Two Outliers



Five Nearest Neighbors - Small Cluster



Five Nearest Neighbors - Differing Density



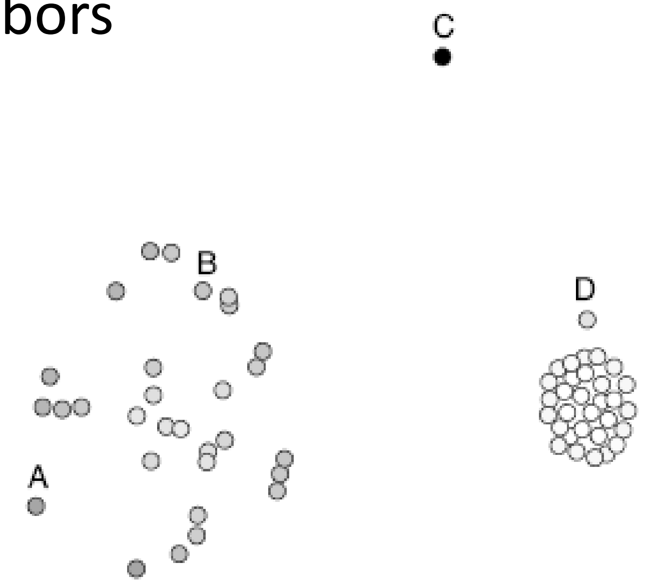
Density-Based Approaches

- Outliers are objects in regions of low density

- Defining density:
  - Count of points within a radius (DBSCAN)
  - Inverse of the average distance to the k nearest neighbors

- Struggles when data contains regions of

differing densities



Relative Density

- Relative density is the ratio of the density of point x to the average density of its k nearest neighbors
- Relative density is also often called Local Outlier Factor (LOF)

$$relative\ density\ (x) = \frac{density(x)}{\sum_{i=1}^{k} density(x_i)/k}$$

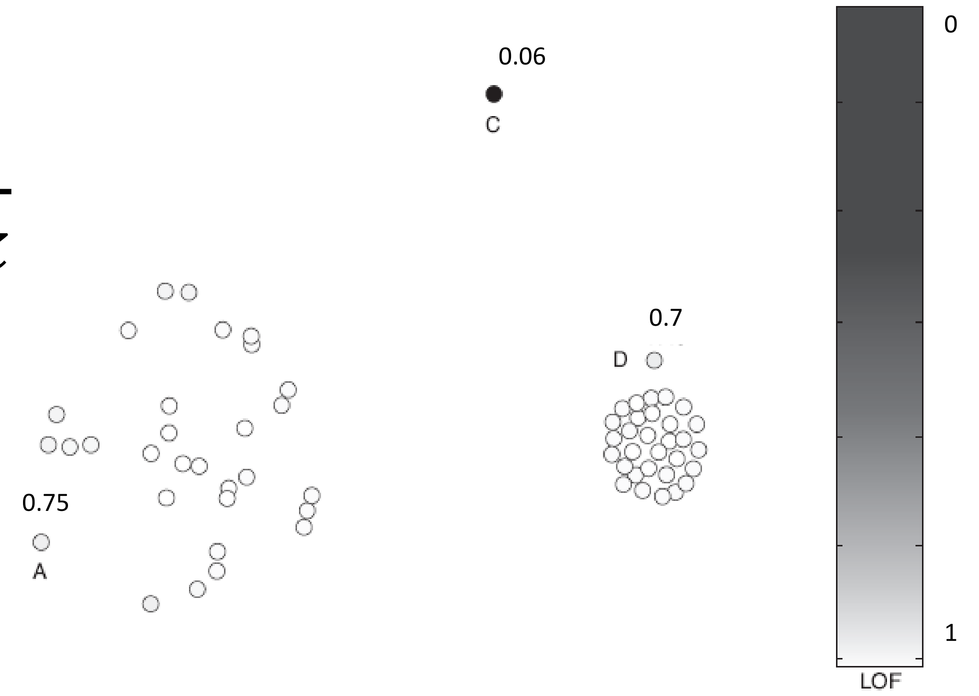


Figure 10.8. Relative density (LOF) outlier scores for two-dimensional points of Figure 10.7.

Characteristics of Proximity and Density Based Approaches

- More easily applied that statistical-based approaches. Easier to define a proximity/density measure than a probability distribution.
- Both proximity and density/relative density give a measure of the degree to which an object is an outlier.
- Relative density can work even if there are regions of differing densities
- Works in multi-dimensions, but does suffer from the curse of dimensionality (all data points are far from each other in high dimensional space)
- Expensive: $O(m^2)$ time complexity
- Sensitive to parameter values; Parameter selection can be difficult (k or Eps & MinPts)

Clustering Based Approaches

- Clustering finds groups of strongly related objects
- Anomalies are objects that are not strongly related to other objects
- Cluster the data, then assess the degree to which an object belongs to any cluster

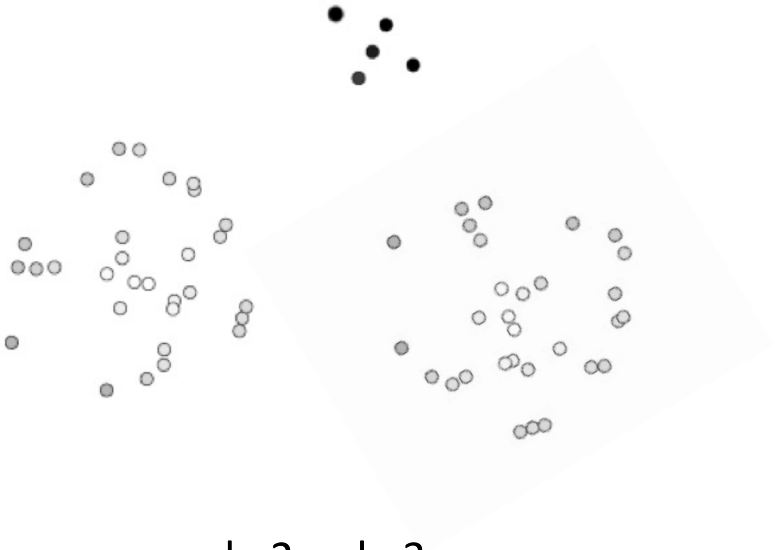
Outliers impact the clustering

- Approach 1:
  - Cluster the data
  - Identify and remove outliers
  - Cluster the data again, without outliers

- Approach 2:
  - Have a group for objects that don’t fit well into any cluster
  - At each step of k-means, the clusters change. Objects that no longer fit well into a cluster are moved to the outlier set. Objects currently in the outlier set are tested again to see if they now belong to a cluster.
  - Objects in the outlier set when the clustering stops, are outliers

Characteristics of Clustering Based Approaches

- Since the definition of a cluster is complementary to that of an outlier, you can find clusters and outliers at the same time
- The set of outliers found is heavily dependent on the number of clusters used



k=2 vs k=3

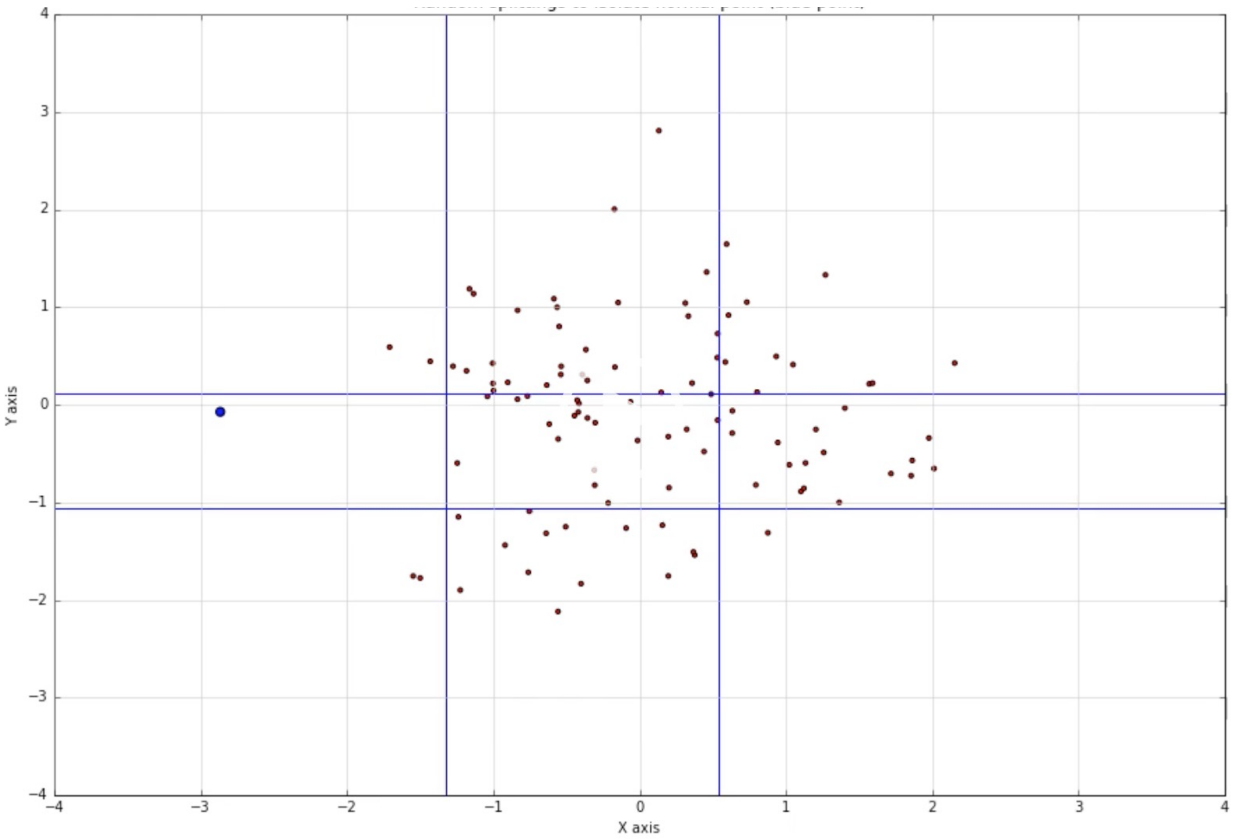
Isolation Based Approach (Isolation Forests)

- Points that are farther away from other data points are easier to isolate.

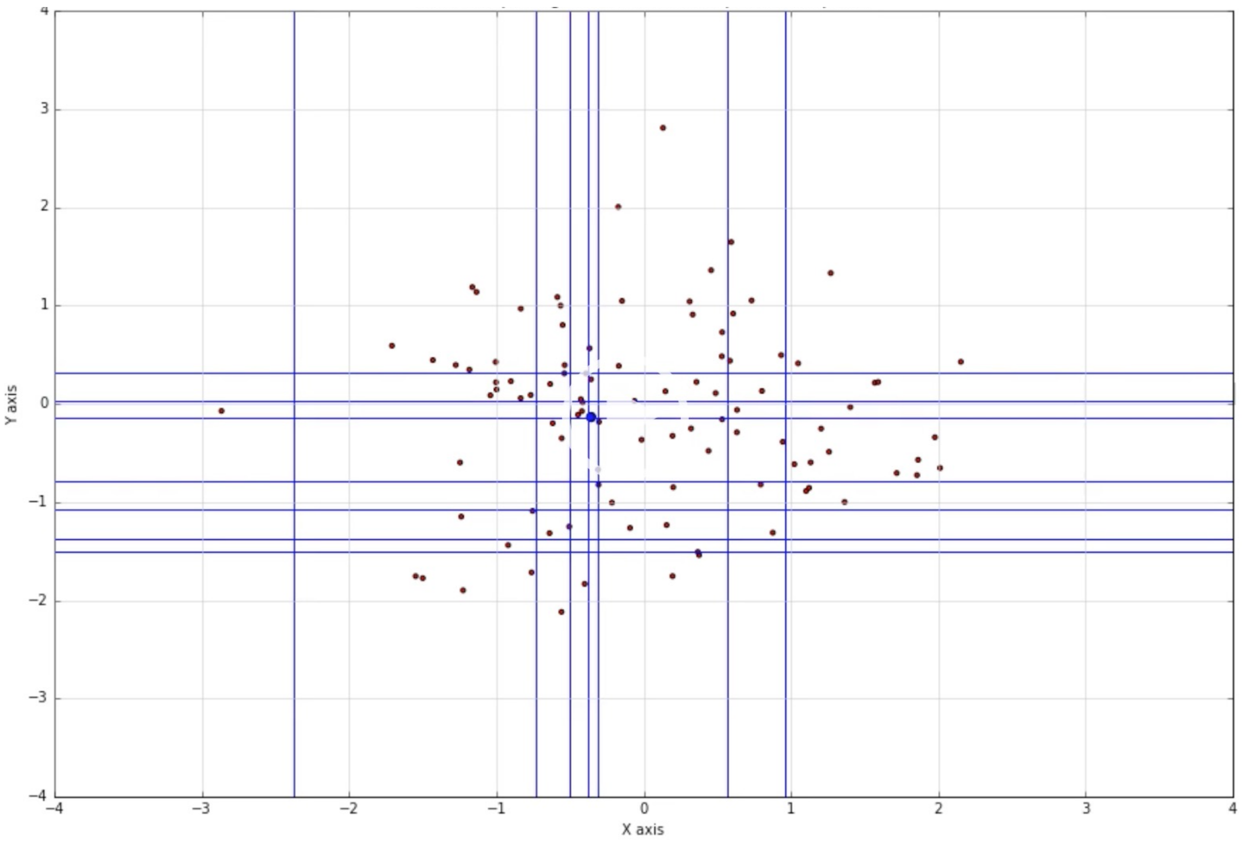
- Use trees to isolate each data point
- Randomly select a feature, then randomly select a value on which to split that feature
- Continue splitting data until all data points are isolated (or data can’t be split further)
- Keep track of the number of splits required to isolate each point

- Create a forest of these isolation trees
- Use the average number of splits required to isolate the point as an anomaly score – the smaller, the more likely to be an anomaly

Isolation Forests Example



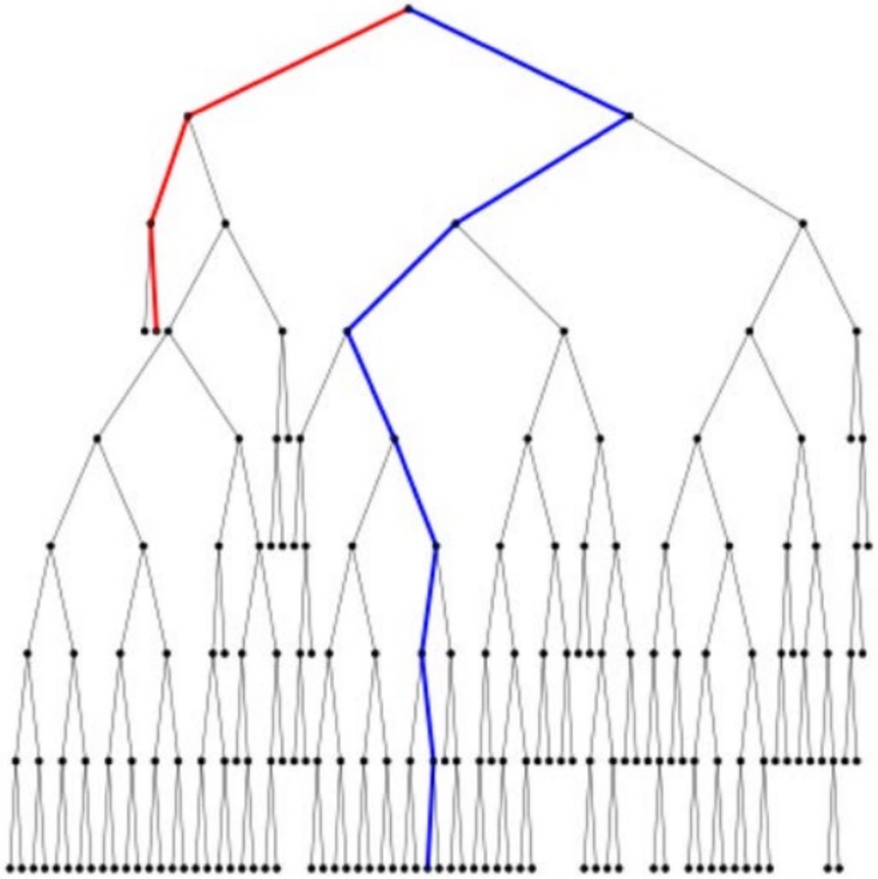
Isolating an outlier



Isolating an inlier

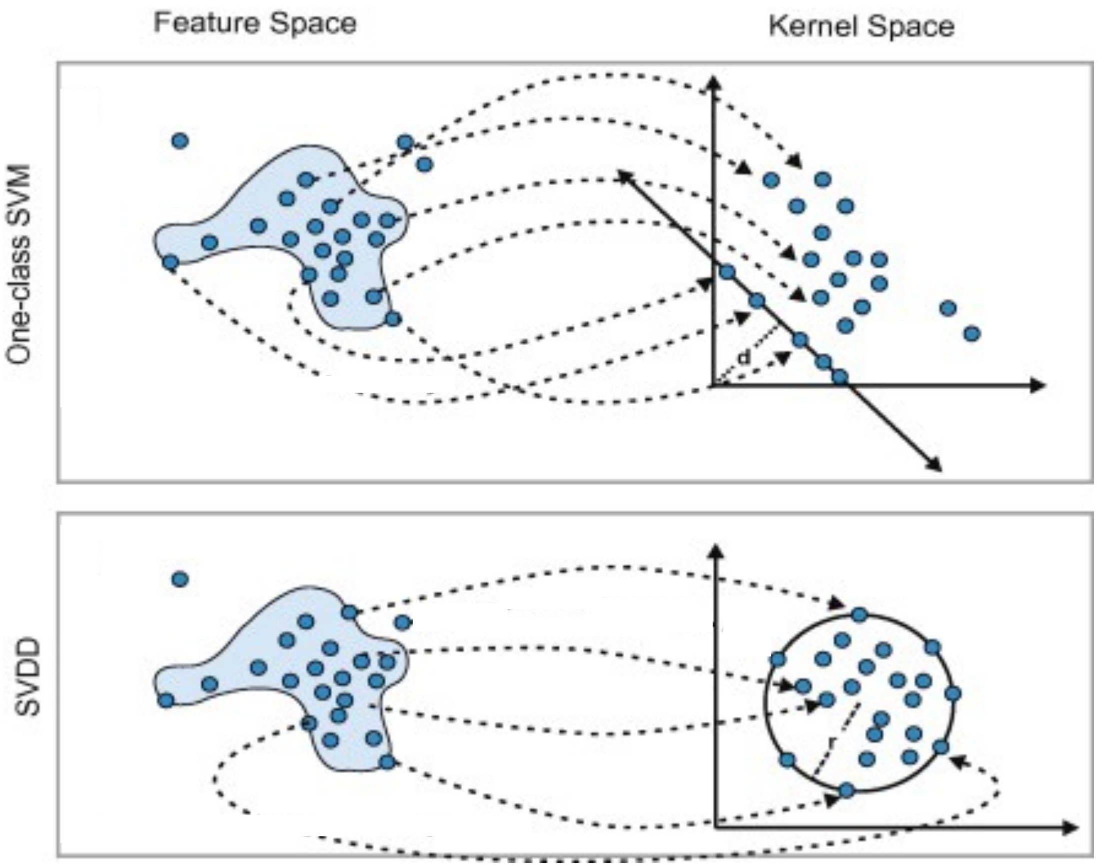
Characteristics of Isolation Forests

- No need to define a distance or density measure, nor any hyperparameters (other than number of trees)
- Not using density or distance eliminates a large computational cost
- Not susceptible to the curse of dimensionality



One-class SVMs

- Often used for **novelty detection** – data does not contain outliers, but want to detect anomalies in new data



- Use an SVM to find the hyperplane that separates all data from the origin (may need to use a kernel function)
- On which side of the hyperplane is the new data point?

- Support Vector Data Description (SVDD) – Same approach using a hypersphere instead of a hyperplane

One-Class SVM

- In a one-class SVM, the parameter ν indicates how many anomalies you’re willing to tolerate in the training data (how many training data points can be outside of the boundary)

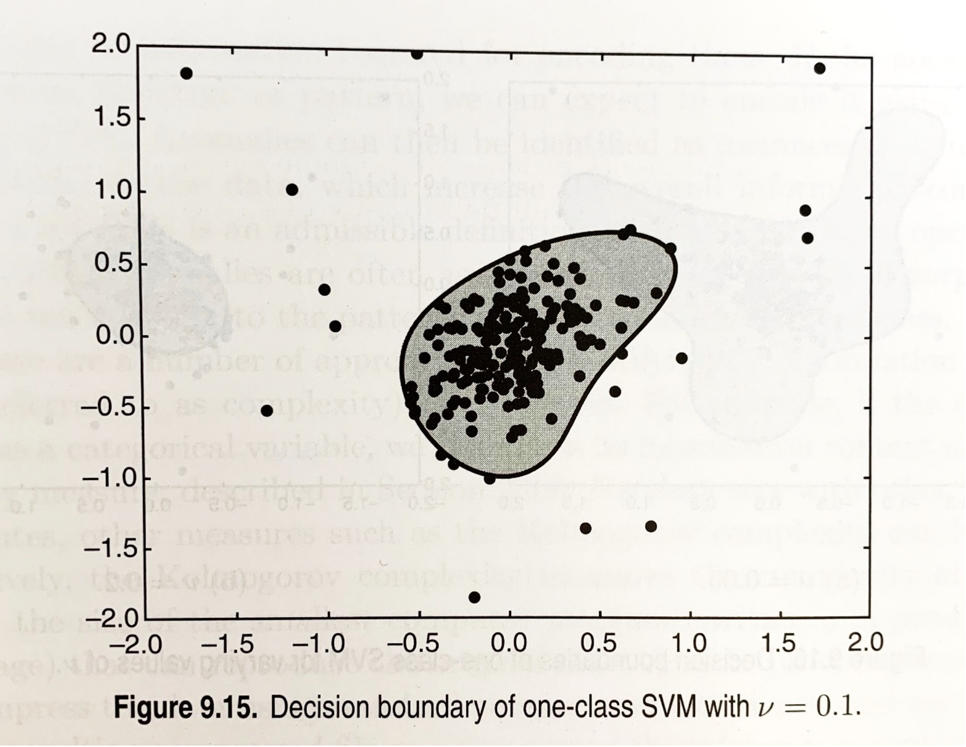
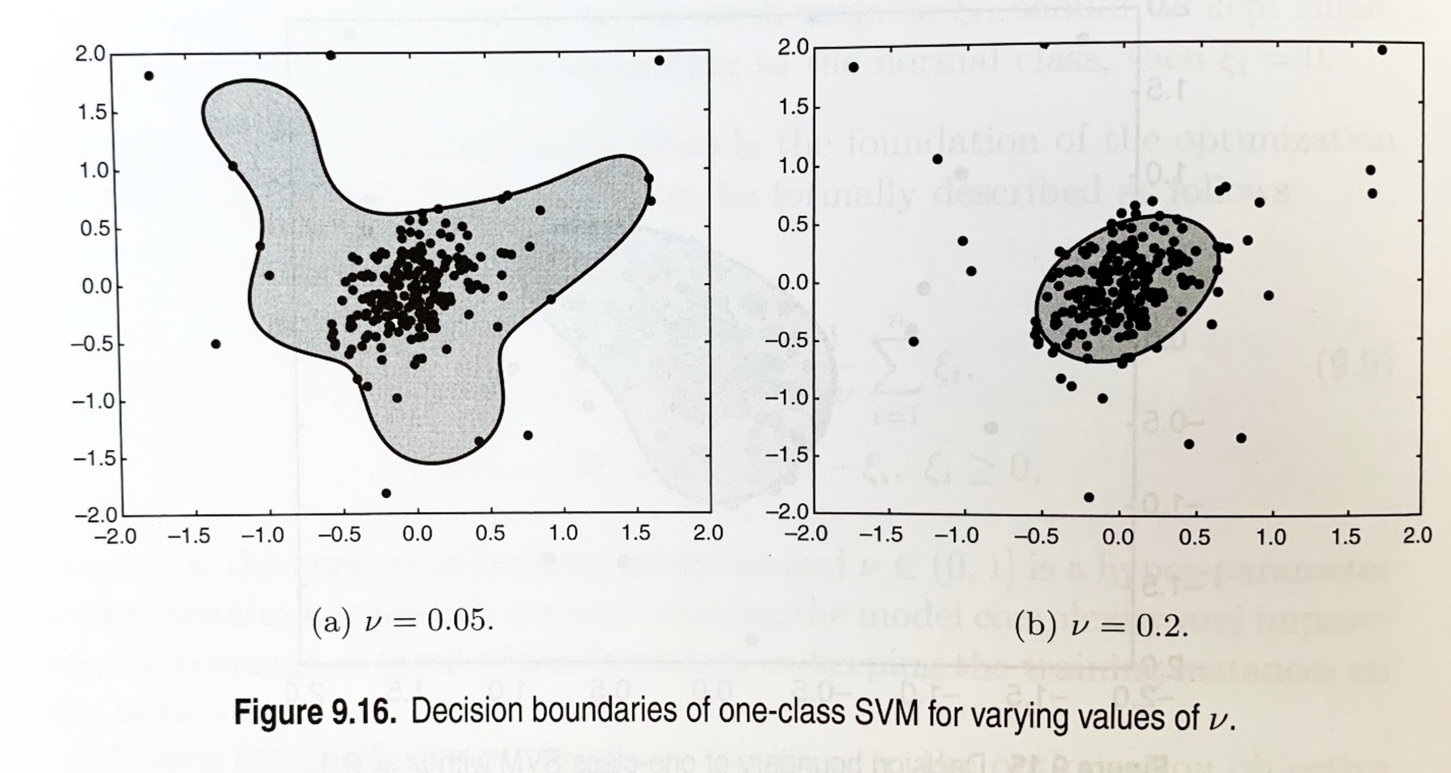


Figure 9.15. Decision boundary of one-class SVM with $\nu = 0.1$.



(a) $\nu = 0.05$.

(b) $\nu = 0.2$.

Figure 9.16. Decision boundaries of one-class SVM for varying values of ν.

Characteristics of One-class SVMs

- Often used when no anomalous examples are available – dataset contains only normal points
- Can predict if new data points are unlike the normal examples
- Requires choice of a kernel and scalar to define the hyperplane
- Not susceptible to the curse of dimensionality

Summary: Approaches to Anomaly Detection

- Model-based techniques: Build a model of the data. Anomalies are points that do not fit the model well.
  - Model can be a probability distribution, anomaly is object that is not very likely under the distribution
  - Model can be a clustering, anomaly is an object that does not strongly belong to any cluster
  - Model can be a one-class SVM, anomaly is on the other side of the hyperplane
- Proximity-based techniques: Anomalous objects are those that are distant from most other objects
- Density-based techniques: Anomalous objects are in regions of low density
- Isolation-based techniques: Anomalous objects can be easily isolated