Outlier/Anomaly Detection

(for your reference only)

What is an outlier?

- An observation (or measurement) that is usually different (large or small) relative to the other values in a data set.
- Outliers typically are attributable to one of the following causes:
  - Error: the measurement or event is observed, recorded, or entered into the computer incorrectly.
  - Contamination: the measurement or event comes from a different population.
  - Inherent variability: the measurement or event is correct, but represents a rare event.

Many Names for Outlier Detection

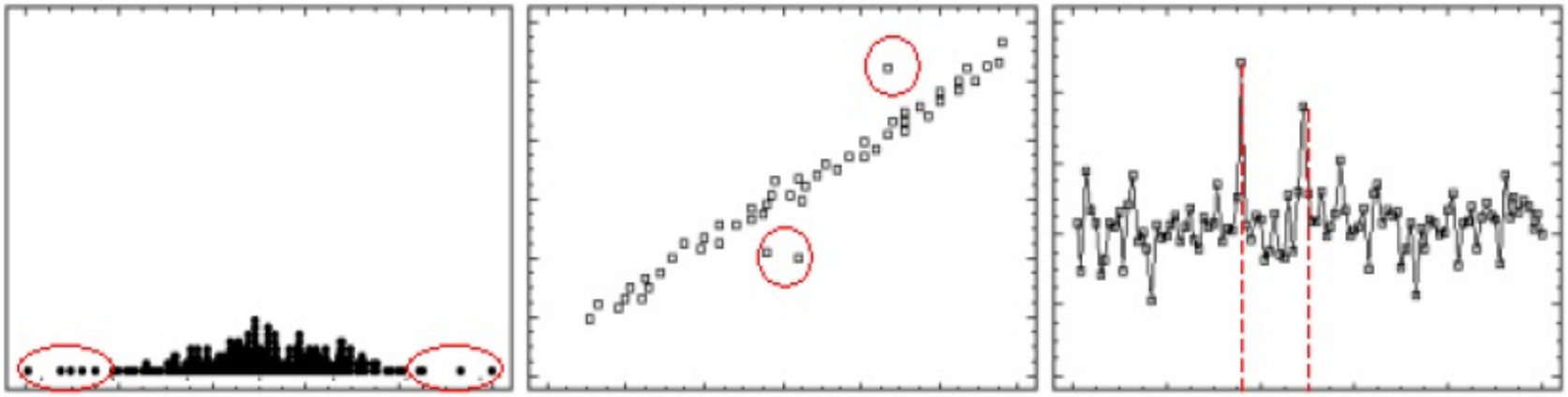
- Outlier analysis
- Anomaly detection
- Intrusion detection
- Misuse detection
- Surprise discovery
- Rarity detection
- Unusual event detection

Applications of Anomaly Detection

- Network intrusion detection
- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining
- …

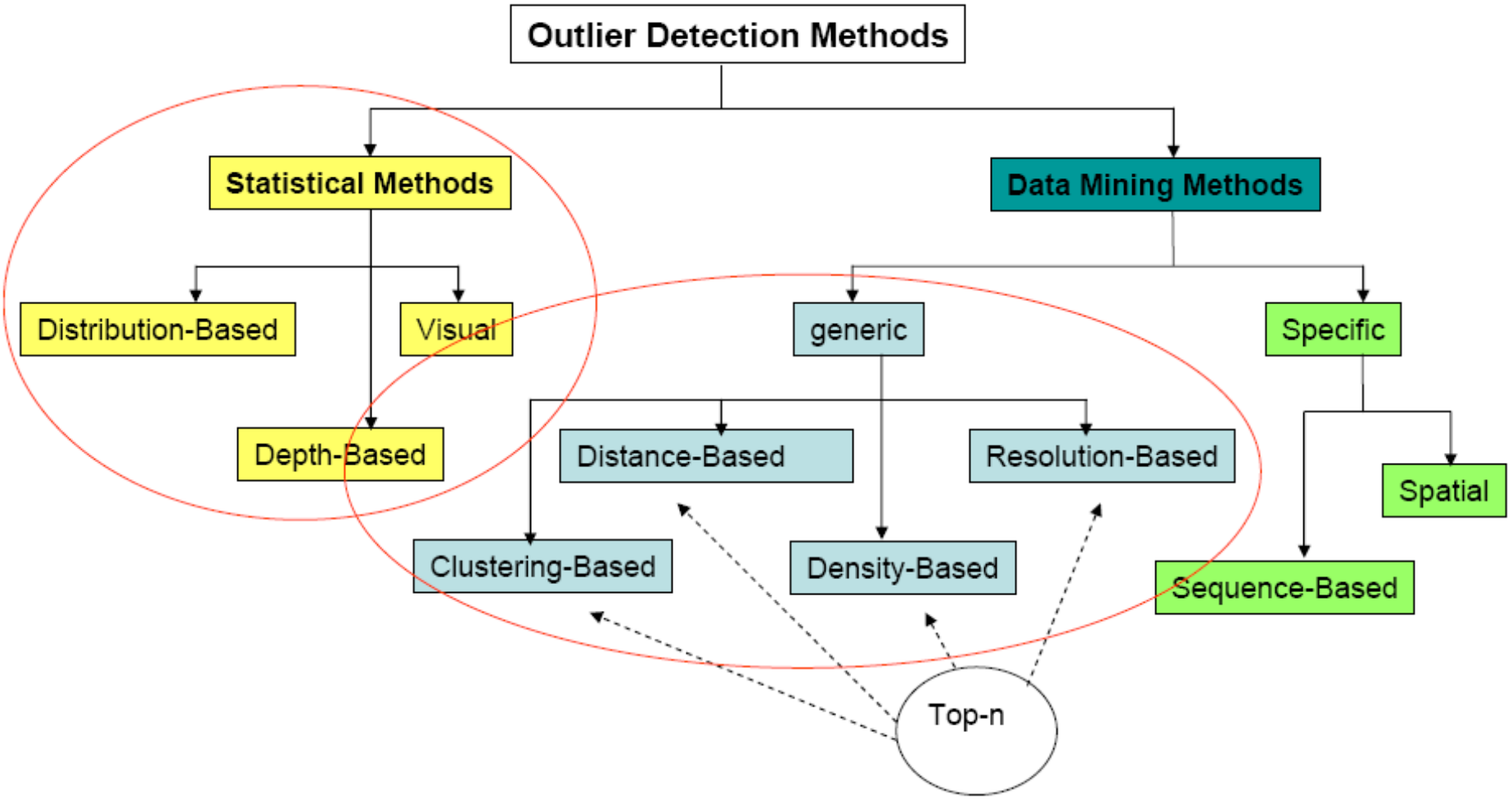
Relativity of an Outlier

- The notion of outlier is subjective and highly application-domain-dependent.
- There is always an ambiguity in defining an outlier.



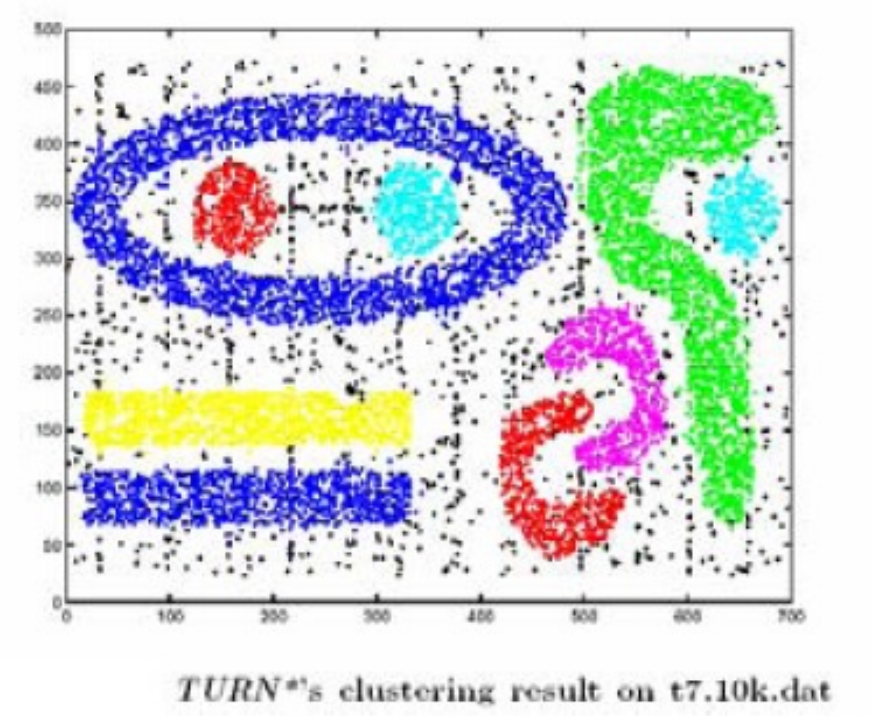
(a) Outliers w.r.t. a distribution (b) Outliers w.r.t. a pattern (c) time series outliers

Topology for Outlier Detection

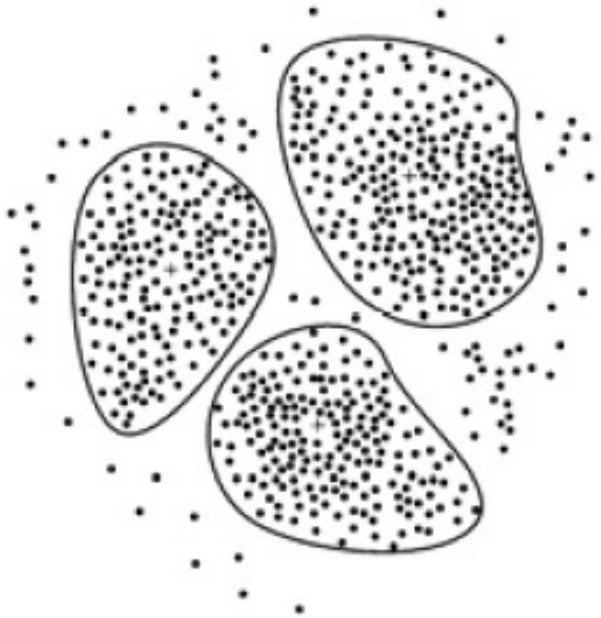


Clustering-Based Outlier Mining

- Some clustering techniques distinguish between isolated points and clustered points – non-sensitive to noise, e.g. DBSCAN
- Identified small clusters and singletons are labeled outliers.

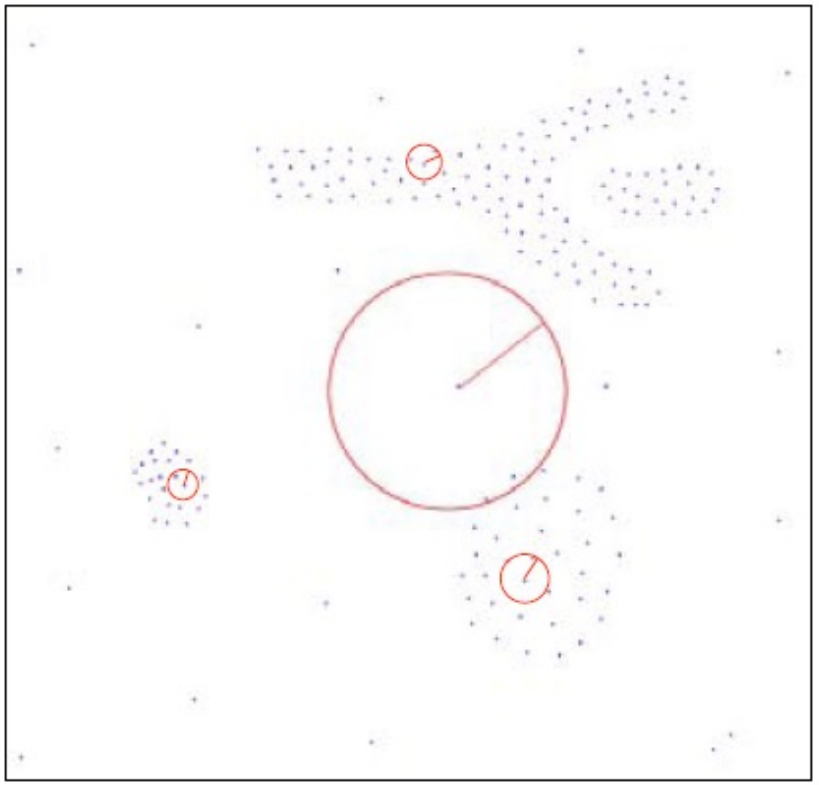


*TURN**'s clustering result on t7.10k.dat



k-Nearest Neighbor Approach

- Given k, for each point calculate the average distance to its k nearest neighbours. The larger the average distance the higher the likelihood the point is an outlier.
- Could sort in descending order the points based on their average distance to their k-NN.

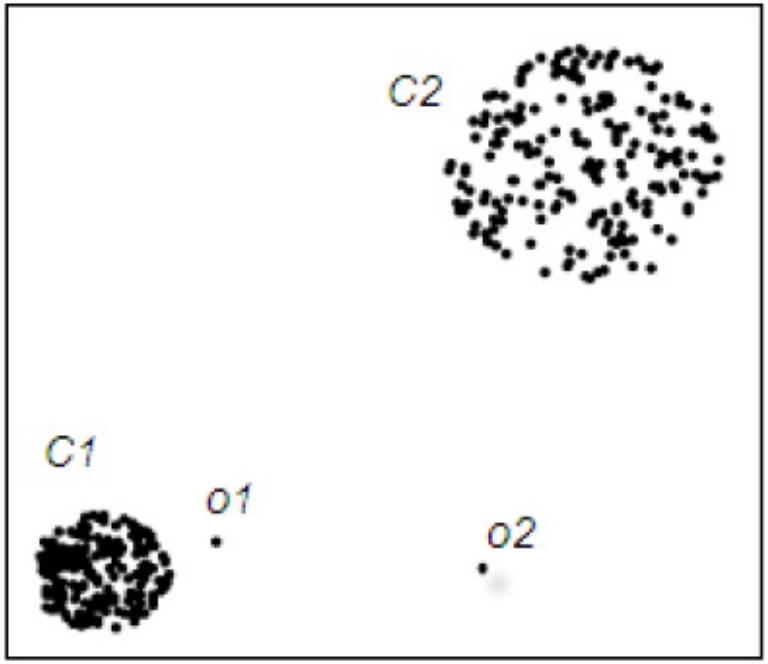


Distance-Based Outlier Analysis

- A typical distance-based outlier notion – DB(p,d)
- Distance-based outlier definition: A $DB(p,D)$-outlier is an object o in a dataset T such that at least a fraction p of the objects in T lies at a distance greater than D from o
- Can effectively identify outliers which deviate from the majority

Distance-Based Issues

- Tend to find outliers global to the whole dataset.
- Not good for dataset consisting of clusters of diverse density.
- In this example, C_1 is significantly denser than C_2 and o_1 is not found as outlier since C_1 is too dense relative to C_2.



To conclude…

- There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise.
- Again, a domain specific requirement is usually required.