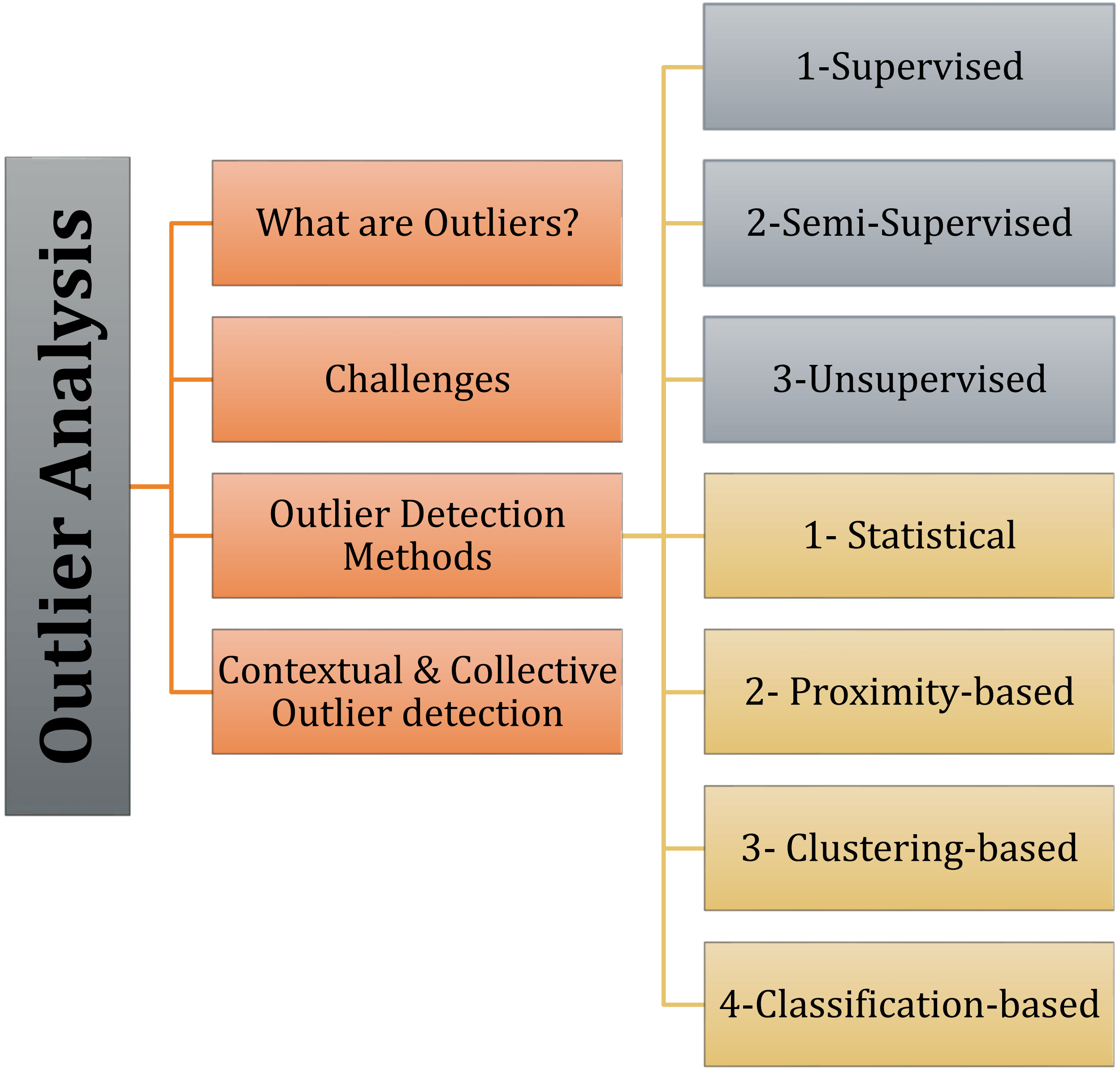
IS422P - DATA MINING OUTLIER ANALYSIS

AMIRA REZK

INFORMATION SYSTEM DEPARTMENT


Outlier Analysis
What are Outliers?
Challenges
Outlier Detection Methods
Contextual & Collective Outlier detection
1-Supervised
2-Semi-Supervised
3-Unsupervised
1- Statistical
2- Proximity-based
3- Clustering-based
4-Classification-based

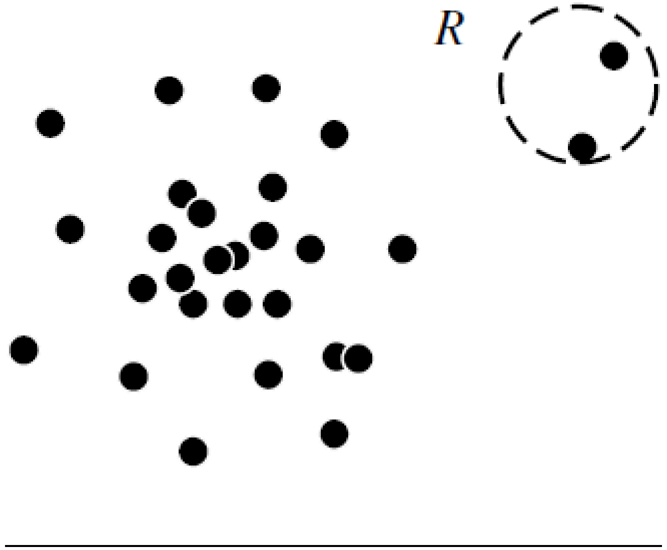
WHAT ARE OUTLIERS?

- Outlier → a data object that deviates significantly from the normal objects
  - Ex: a student with exceptionally high grades
  - Normal versus anomalous data objects → how to define normalcy?
- Outliers are different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection
- Outlier detection vs. novelty detection:
  - early stage, outlier; but later merged into the model
- Applications:
  - Credit card fraud detection, Telecom fraud detection, Customer segmentation

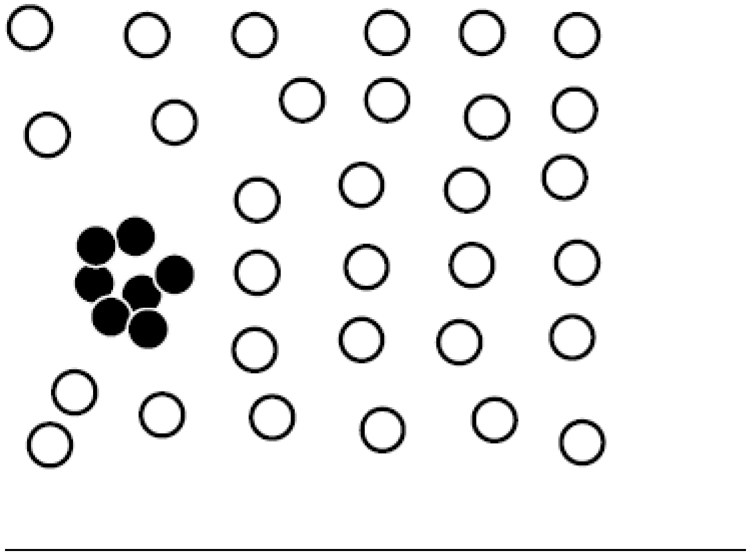
TYPES OF OUTLIERS

- Global
  - Deviate significantly from the rest of the dataset,   Ex: Intrusion Detection
  - How to measure deviation?
- Contextual
  - Deviate with respect to context (time, location),   Ex: Temperature values
  - Contextual attributes used to evaluate context
  - Behavioral attributes used to evaluate outlier behavior
- Collective
  - A subset of objects collectively deviates significantly from the dataset
  - Ex: Multiple order delays, DoS attacks

TYPES OF OUTLIERS EXAMPLES



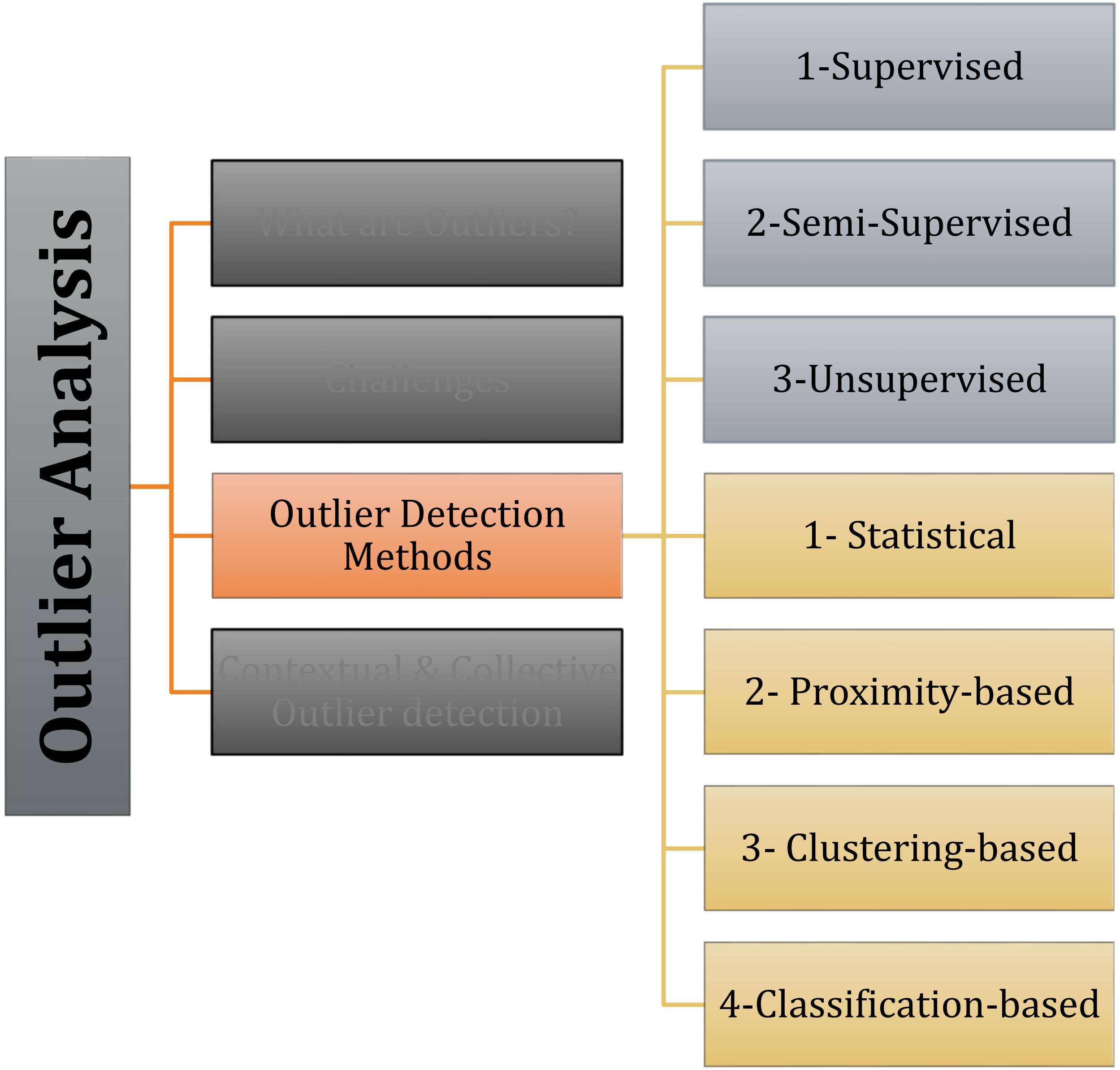
The objects in region *R* are outliers.



The black objects form a collective outlier.

CHALLENGES FOR OUTLIER DETECTION

- Modeling normal objects & outliers
  - Normal models are challenging to build
  - Distinction between normalcy and anomaly is ambiguous
- Application-specific outlier detection
  - How much deviation is considered an anomaly/outlier?
- Handling noise in outlier detection
  - How to detect outliers in the presence of noise?
  - Noise sometimes "hides" outliers!
- Understandability
  - Understand why these are outliers
  - Specify the degree of an outlier



Outlier Analysis

What are Outliers?

Challenges

Outlier Detection Methods

Contextual & Collective Outlier detection

1-Supervised

2-Semi-Supervised

3-Unsupervised

1- Statistical

2- Proximity-based

3- Clustering-based

4-Classification-based

OUTLIERS DETECTION METHODS

- Two ways to categorize outlier detection methods:
  - Based on whether user-labeled examples of outliers can be obtained:
    - Supervised, semi-supervised vs. unsupervised methods
  - Based on assumptions about normal data and outliers:
    - Statistical, proximity-based, and clustering-based methods

OUTLIERS DETECTION METHODS
SUPERVISED METHODS

- Modeling outlier detection as a classification problem
  - Samples examined by domain experts used for training & testing
- Methods for Learning a classifier effectively:
  - Model normal objects & report those not matching the model as outliers, or
  - Model outliers and treat those not matching the model as normal
- Challenges
  - Imbalanced classes, i.e., outliers are rare: Boost the outlier class and make up some artificial outliers
  - Catch as many outliers as possible, i.e., recall is more important than accuracy (i.e., not mislabeling normal objects as outliers)

OUTLIERS DETECTION METHODS
UNSUPERVISED METHODS

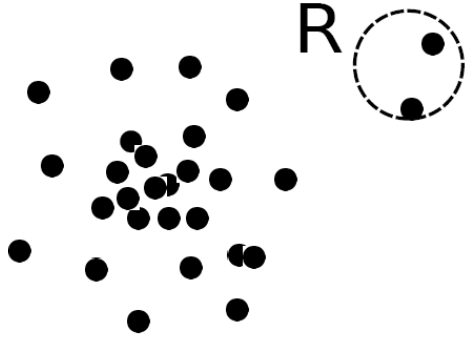
- Normal objects are somewhat ``clustered" into multiple groups, each having some distinct features
  - An outlier is expected to be far away from any groups of normal objects
- Weakness: Cannot detect collective outlier effectively
  - Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area
- Ex. In some intrusion or virus detection, normal activities are diverse
  - Unsupervised methods may have a high false positive rate but still miss many real outliers.
  - Supervised methods can be more effective, e.g., identify attacking some key resources

OUTLIERS DETECTION METHODS
SEMI-SUPERVISED METHODS

- In many applications, the number of labeled data is often small:
  - Labels could be on outliers, normal objects, or both
- If some labeled normal objects are available
  - Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
    - Those not fitting the model of normal objects are detected as outliers
- If only some labeled outliers are available
  - a small number of labeled outliers many not cover the possible outliers well
  - To improve the quality of detection → models for normal objects learned from unsupervised methods

OUTLIERS DETECTION METHODS
STATISTICAL METHODS

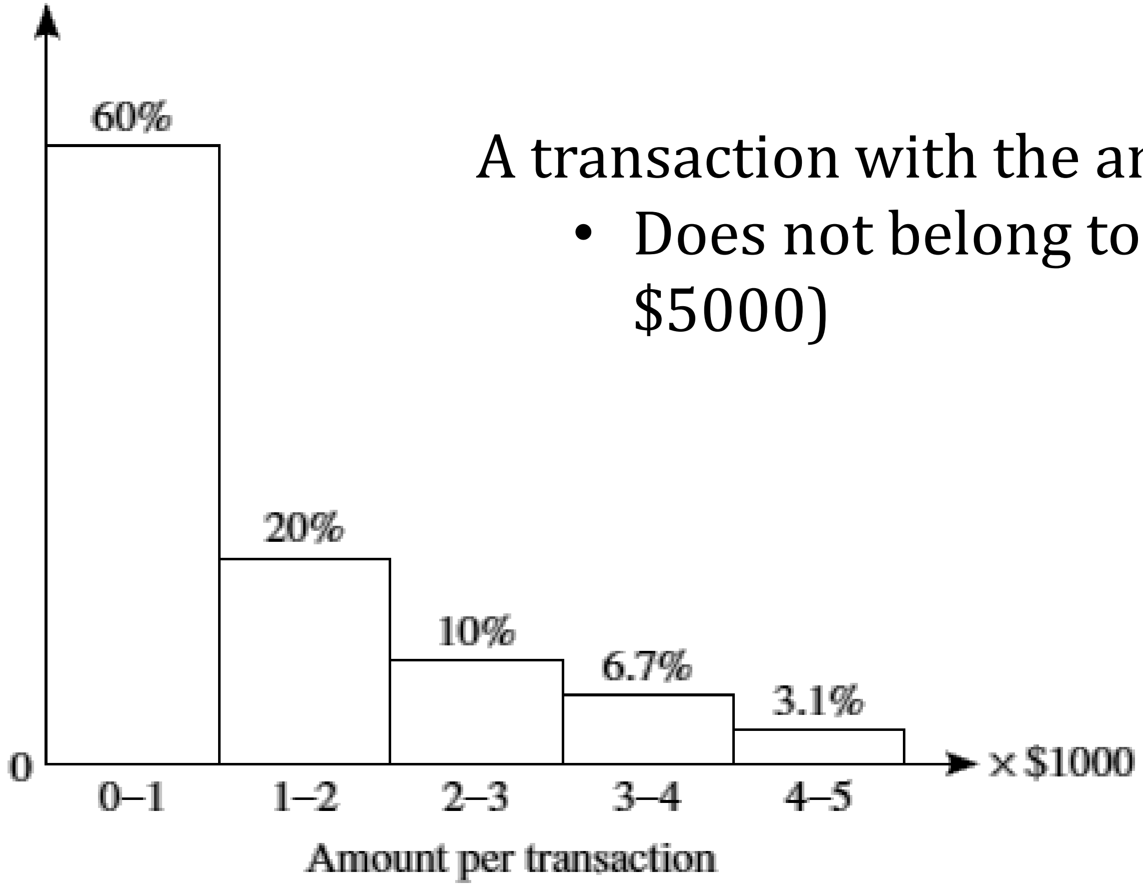
- known as model-based methods: the normal data follow some statistical model (a stochastic model)
  - The data not following the model are outliers.
- Effectiveness: highly depends on whether the assumption of statistical model holds in the real data
- There are rich alternatives to use various statistical models
  - E.g., parametric vs. non-parametric



OUTLIERS DETECTION METHODS
STATISTICAL METHODS

- **Parametric** → assume normal data is generated by a distribution with parameter *Θ*
  - *PDF* of distribution *f*(*x*, *Θ*) yields probability that *x* is generated by distribution → ***smaller means outlier***
  - For **univariate outliers** → *boxplots* → parameters are *mean* and *IQR*
  - For multivariate outliers → χ^2*-statistic* → parameter is *mean*
- **Nonparametric** → learn normal model from input data
  - *histograms*

OUTLIERS DETECTION METHODS
STATISTICAL METHODS

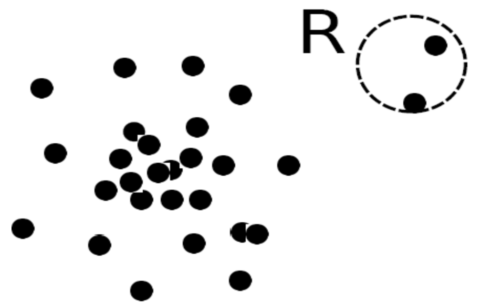


A transaction with the amount of $7500 is considered an outlier

- Does not belong to any of the bins (0.2% of transactions > $5000)

OUTLIERS DETECTION METHODS
PROXIMITY-BASED METHODS

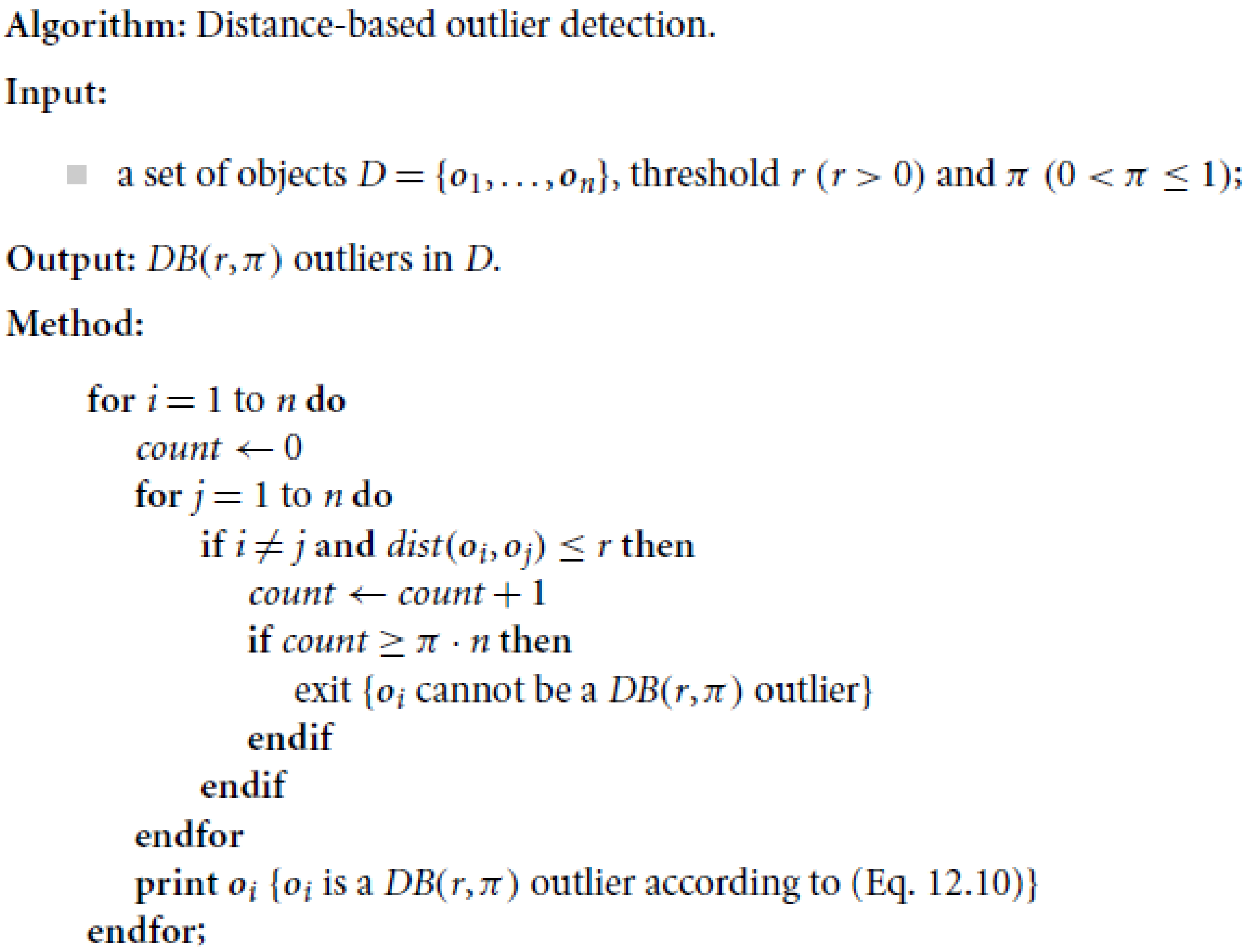
- An object is an outlier if the nearest neighbors of the object are far away, i.e., the proximity of the object is significantly deviates from the proximity of most of the other objects in the same data set
- Effectiveness: highly relies on the proximity measure.
- Challenges:
  - proximity or distance measures cannot be obtained easily.
  - finding a group of outliers which stay close to each other
- Two major types of proximity-based outlier detection
  - Distance-based vs. density-based



OUTLIERS DETECTION METHODS
PROXIMITY-BASED METHODS

- Distance-based → for an object o, examine the number of other objects in its r-neighborhood
  - r is a distance threshold
  - π is a fraction threshold → min # objects needed in neighborhood
- Density-based → for an object o, examine its density relative to the density of its local neighbors
  - A local outlier factor (LOF) is computed in terms of the K-NN of an object in comparison to its neighbors

OUTLIERS DETECTION METHODS PROXIMITY-BASED METHODS



Algorithm: Distance-based outlier detection.

Input:

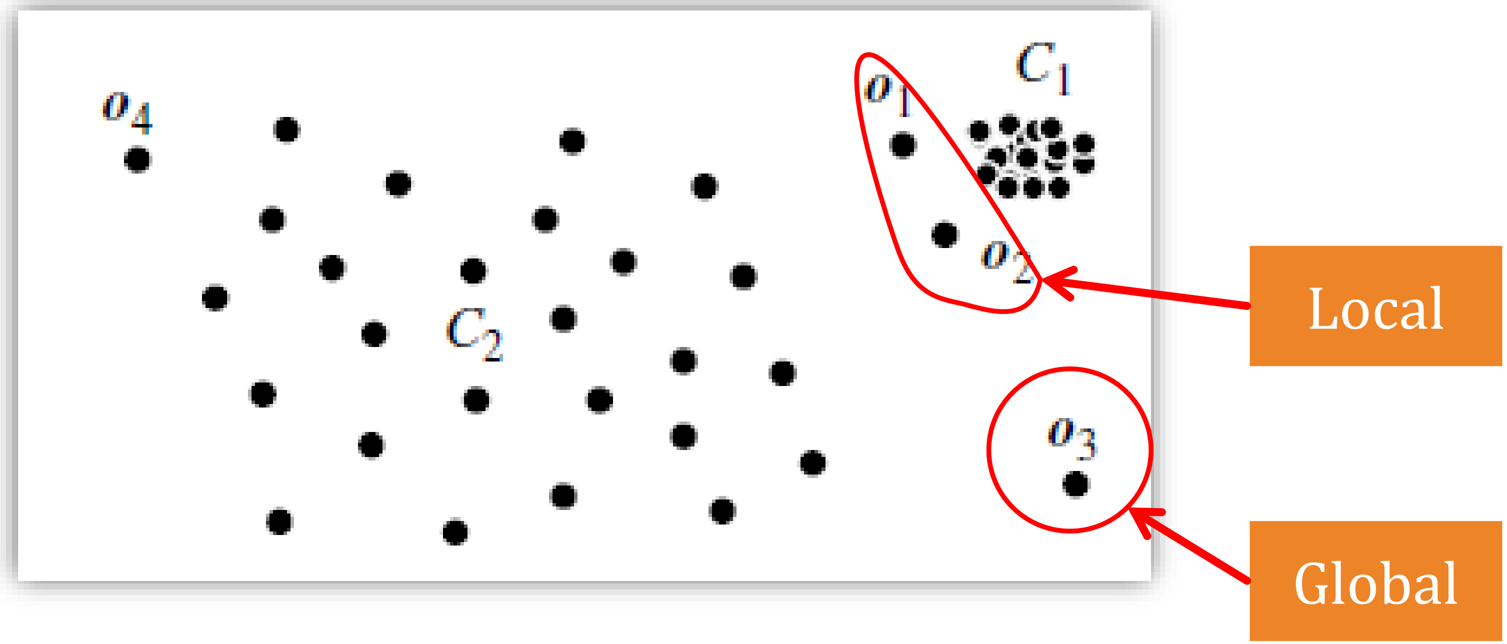
- a set of objects $D = \{o_1, \ldots, o_n\}$, threshold r ($r > 0$) and π ($0 < \pi \leq 1$);

Output: $DB(r, \pi)$ outliers in D.

Method:

```
for $i = 1$ to $n$ do
    $count \leftarrow 0$
    for $j = 1$ to $n$ do
        if $i \neq j$ and $dist(o_i, o_j) \leq r$ then
            $count \leftarrow count + 1$
            if $count \geq \pi \cdot n$ then
                exit {$o_i$ cannot be a $DB(r, \pi)$ outlier}
            endif
        endif
    endfor
    print $o_i$ {$o_i$ is a $DB(r, \pi)$ outlier according to (Eq. 12.10)}
endfor;
```

OUTLIERS DETECTION METHODS
PROXIMITY-BASED METHODS



Local

Global

OUTLIERS DETECTION METHODS
CLUSTER-BASED METHODS

- Normal data belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters
- Since there are many clustering methods, there are many clustering-based outlier detection methods as well
- Clustering is expensive:
  - straightforward adaption of a clustering method for outlier detection can be costly and does not scale up well for large data sets

OUTLIERS DETECTION METHODS
CLUSTERING-BASED METHODS

- Either an object:

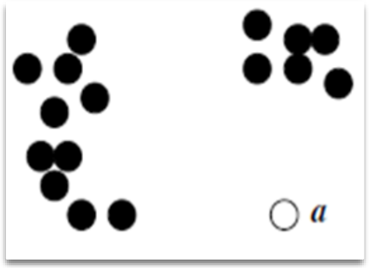
Ratio of dist (object, centroid) to average (distances between centroid and its assigned objects)

Rank clusters according to size

Assign a LOF to objects:

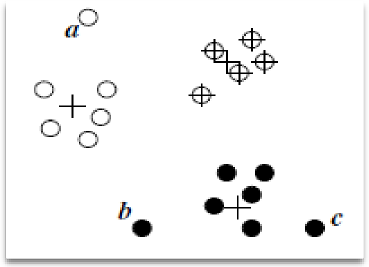
For large clusters: *size*(***C***) × *similarity* (***o***, ***C***)

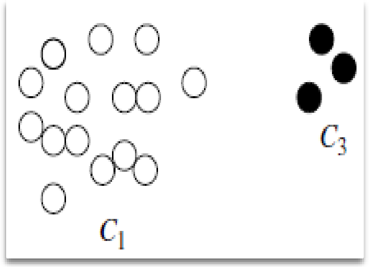
For small clusters: *size*(***C***) × *similarity*(***o***, closest large ***C***)



does not belong to any cluster

Has a large distance to the closest cluster

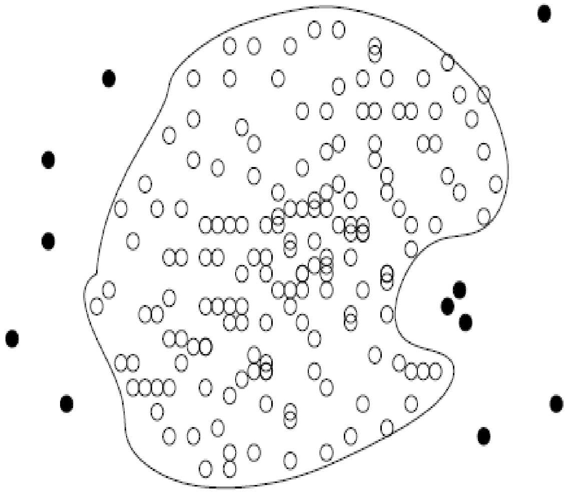




belongs to a small and remote cluster

OUTLIERS DETECTION METHODS
CLASSIFICATION-BASED METHODS

- One-class model: describe only the normal class.
  - Learn the decision boundary of the normal class using classification methods such as SVM
  - Any samples that do not belong to the normal class (not within the decision boundary) are declared as outliers
  - Advantage : detect new outliers
  - Extension: Normal objects may belong to multiple classes



- A brute-force approach: a training set contains samples labeled as “normal” and others labeled as “outlier”
  - But, the training set is typically heavily biased: # of “normal” samples likely far exceeds # of outlier samples
  - Cannot detect unseen anomaly

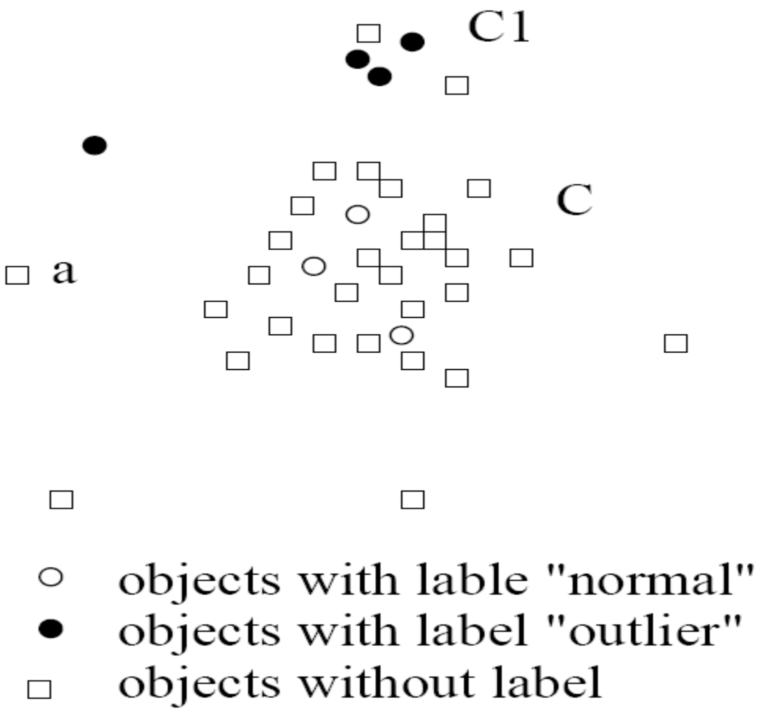
OUTLIERS DETECTION METHODS
CLASSIFICATION-BASED METHODS

- Semi-supervised learning: Combining classification-based and clustering-based methods

○ **Strength**: Outlier detection is fast

○ **Bottleneck**:

- Quality heavily depends on the availability and quality of the training set, (difficult to obtain high-quality training data)



MINING CONTEXTUAL AND COLLECTIVE OUTLIERS

- Contextual outliers:
  - Determine contextual attributes
  - Group those attributes' values
  - Determine context of object (its group)
  - Do conventional outlier detection within that group
  - Ex → customers with similar age who live within same area may have similar behavior → a customer in that age group and living in that area with a different behavior is an outlier
- Collective outliers → challenging and advanced area

EXAMPLE

- For the following nine points (2,2), (2,7), (3,1), (3,5), (4,3), (4,8), (5,2), (6,2), (6,5)
- Assume that k = 2 and initially assign (2,2) and (2,7) as the center of each cluster.
- Apply the k-means algorithm using the Manhattan distance and show the new cluster centers after the first round execution (Hint: The Manhattan distance is: d(i, j) = |xi1-xj1|+ |xi2-xj2|+ ….+ | xin-xjn|.)
- Sol:
- Cluster 1: (2,2), (3,1), (4,3), (5,2), (6,2) new mean of C1= (4,2)
- Cluster2: (2,7), (3,5), (4,8), (6,5) new mean of C2= (3.75, 6.25)

	(3,1)	(3,5)	(4,3)	(4,8)	(5,2)	(6,2)	(6,5)
(2,2)	2	4	3	8	3	4	7
(2,7)	7	3	6	3	8	9	6

EXAMPLE

- Suppose that your result in (a) is the final cluster, how can you use it to detect the outlier? What is the object which likely be an outlier?
- **Sol:** calculate Ratio of dist(object, centroid) to average (distances between centroid and its assigned objects), if object has a large distance to the closest cluster then it is local outlier.

Average of C1= 1.6

Average of C2= 2.6

(6, 5) is likely an outlier

C1	(2,2)	(3,1)	(4,3)	(5,2)	(6,2)
(4,2)	2	2	1	1	2

C2	(2,7)	(3,5)	(4,8)	(6,5)
(3.75, 6.25)	2.5	2	2	4

QUESTION?

NEXT …….