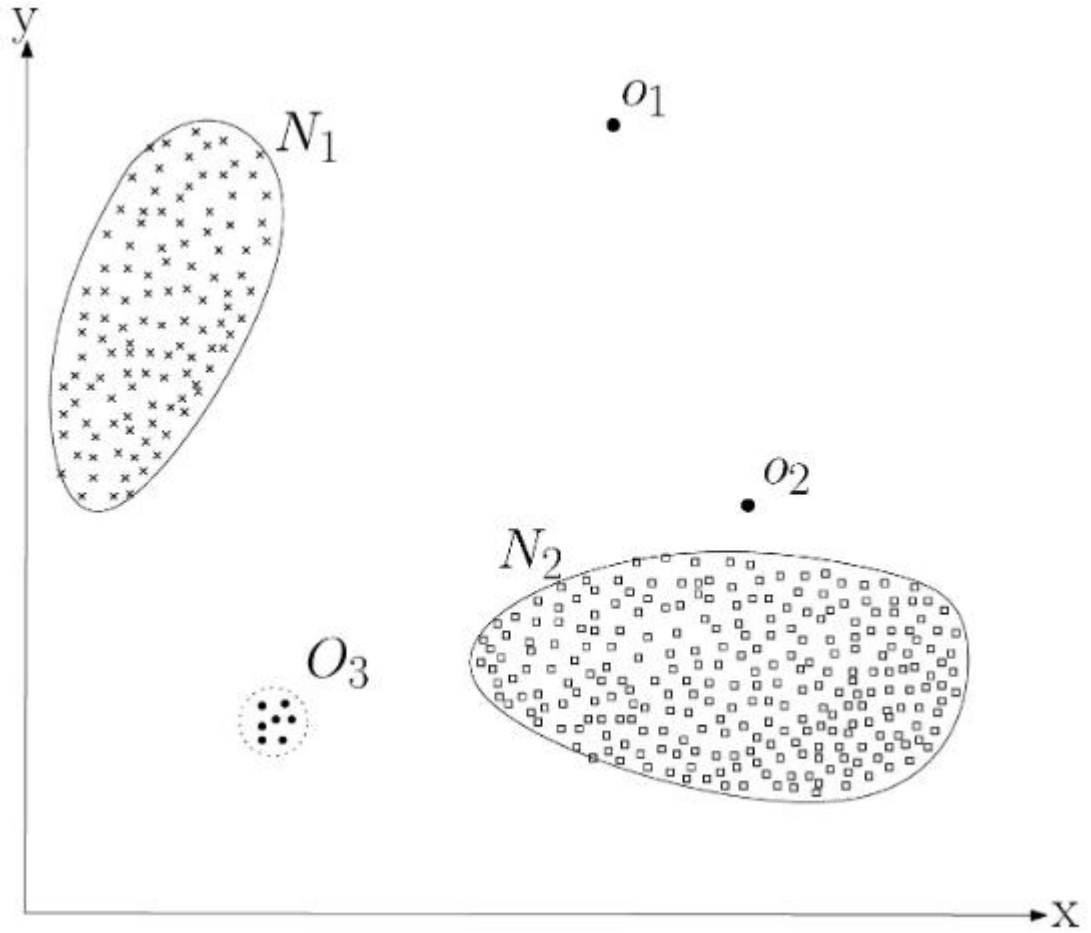
Introduction to Anomaly detection

CSCI 347 – Adiesha Liyanage

What is an Anomaly?

“Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are often referred to as anomalies [or] outliers...in different application domains”

-V. Chandola et al. “Anomaly Detection: A Survey.” 2009



MONTANA STATE UNIVERSITY

What are some applications of anomaly detection?

- Fraud detection
  - Credit card transactions
- Intrusion detection
  - Network traffic
- Fault Detection
  - Safety-critical systems
- Cancer Screening
  - MRI scans

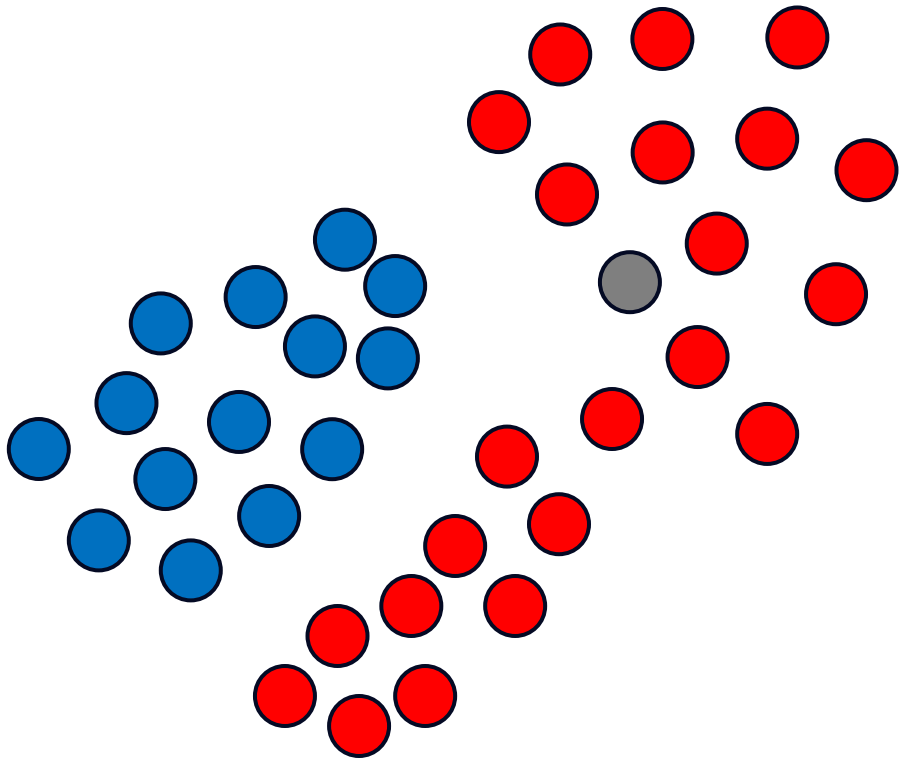
Approaches to Anomaly Detection

- KNN
- Local Outlier Factor
- Isolation Forest
  - Tries to isolate anomalies
  - Tree based algorithm
- One-class SVM
  - It tries to learn the boundary around the "normal" data points and classifies anything outside that boundary as an anomaly
- Cluster-based methods
  - You can use existing clustering methods.
- Low-pass filter
  - Used in time series data to smooth out short-term fluctuations (noise) and highlight long-term trends.

K-NEAREST NEIGHBOR CLASSIFICATION

Given: Data with class labels (blue and red points)

We want to find: Class label of new data instance (grey point)



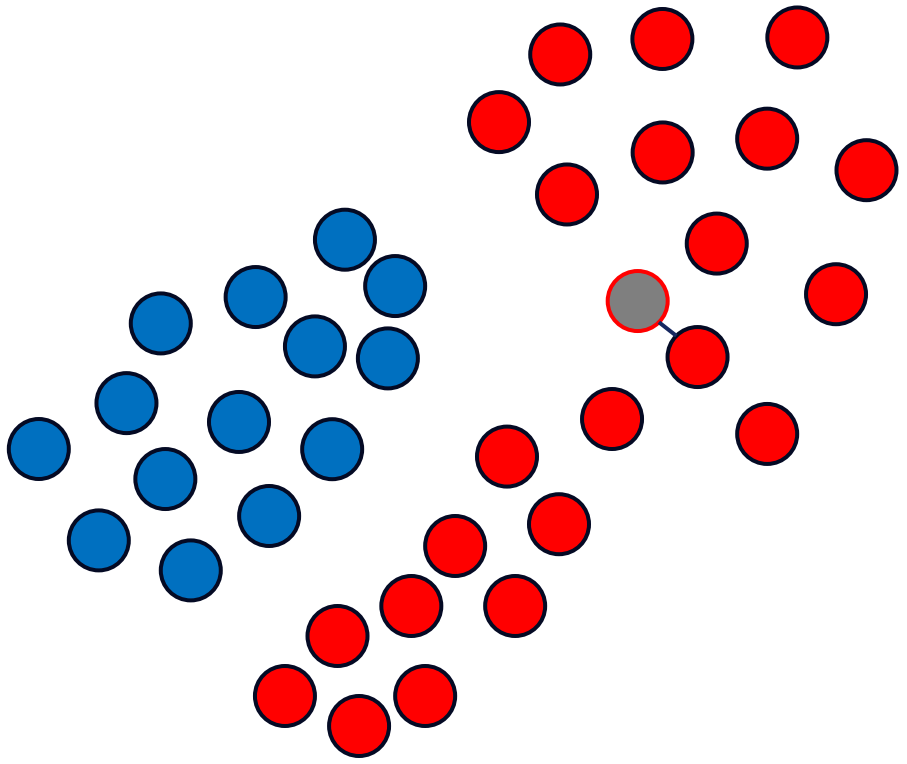
Find k nearest neighbors of the new data instance, and assign it the majority class label

K-NEAREST NEIGHBOR CLASSIFICATION

Given: Data with class labels (blue and red points)

We want to find: Class label of new data instance (grey point)

$k = 1$: Red Label



Find k nearest neighbors of the new data instance, and assign it the majority class label

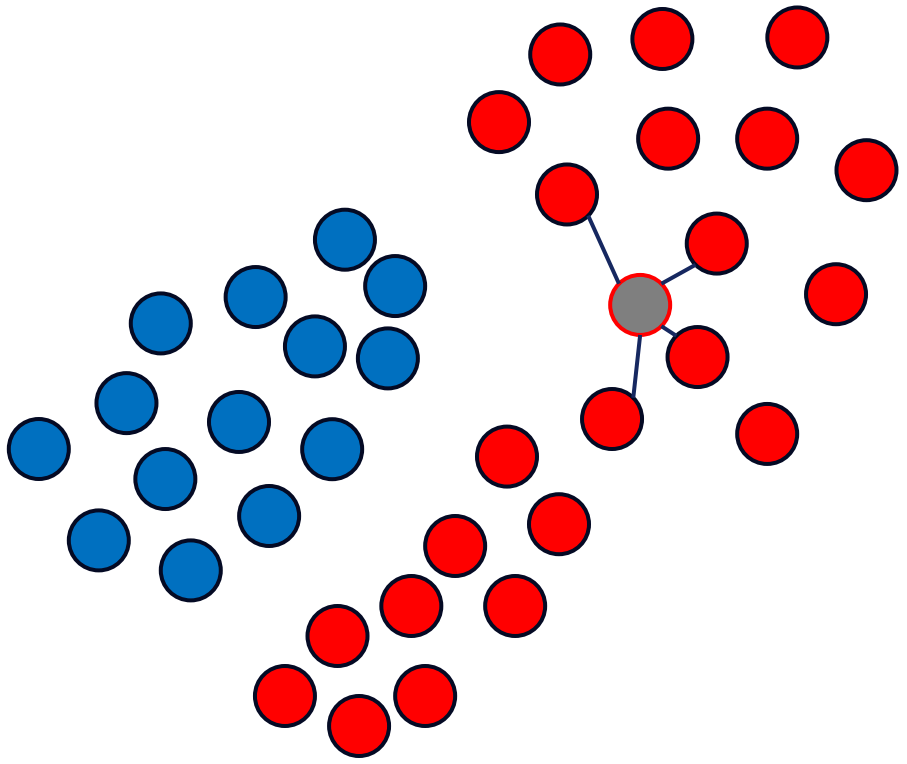
MONTANA STATE UNIVERSITY

K-NEAREST NEIGHBOR CLASSIFICATION

Given: Data with class labels (blue and red points)

We want to find: Class label of new data instance (grey point)

$k = 4$: Red Label



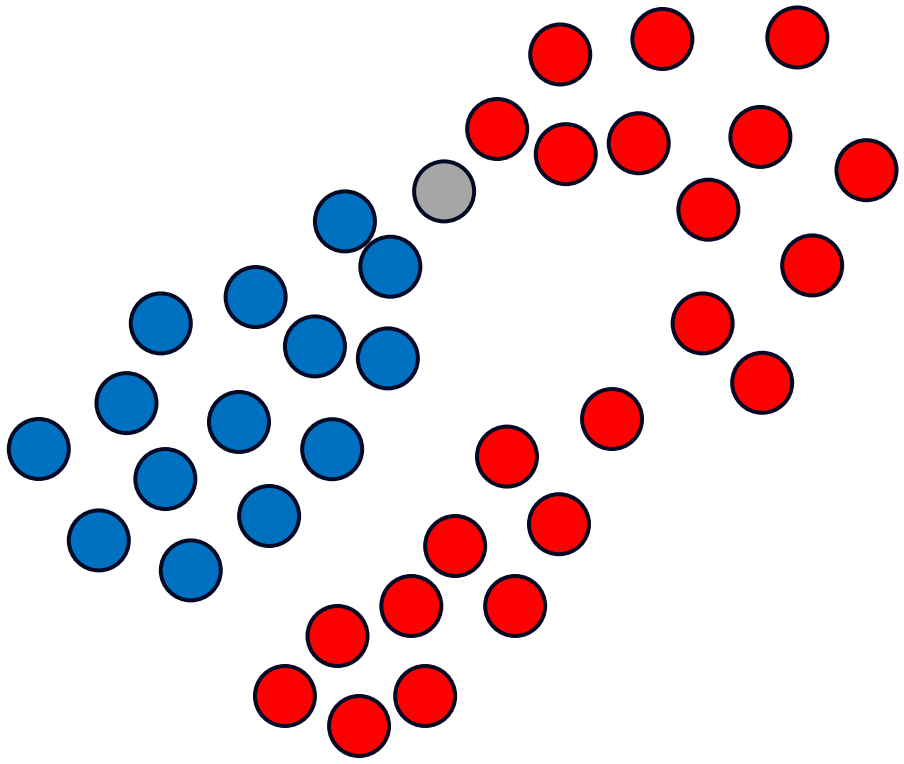
Find k nearest neighbors of the new data instance, and assign it the majority class label

MONTANA STATE UNIVERSITY

K-NEAREST NEIGHBOR CLASSIFICATION

Given: Data with class labels (blue and red points)

We want to find: Class label of new data instance (grey point)



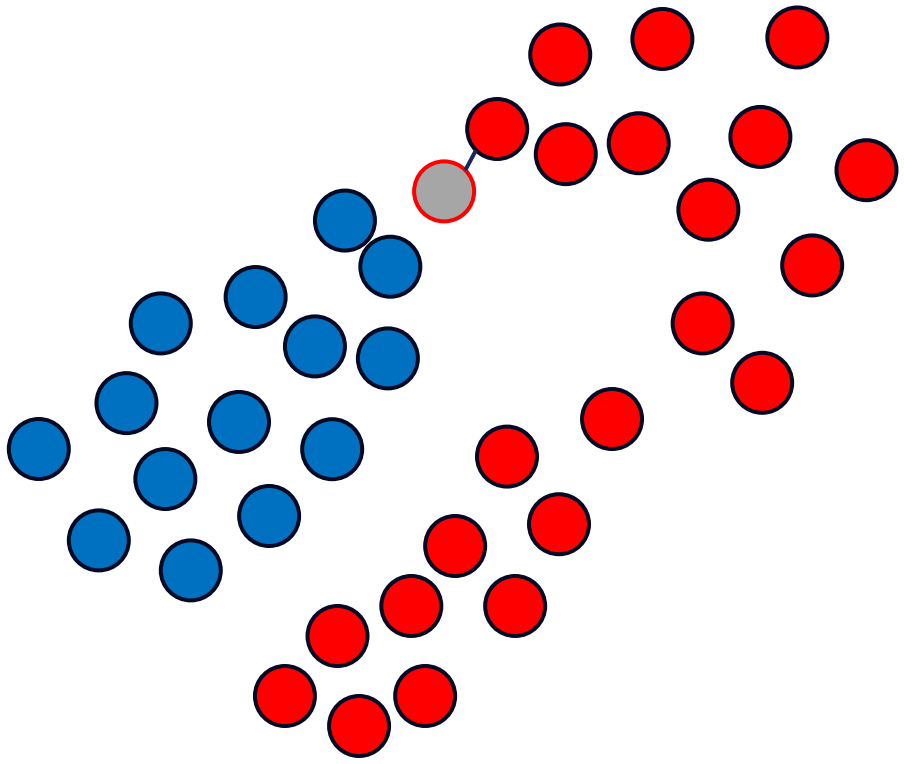
Find k nearest neighbors of the new data instance, and assign it the majority class label

K-NEAREST NEIGHBOR CLASSIFICATION

Given: Data with class labels (blue and red points)

We want to find: Class label of new data instance (grey point)

$k = 1$: Red Label



Find k nearest neighbors of the new data instance, and assign it the majority class label

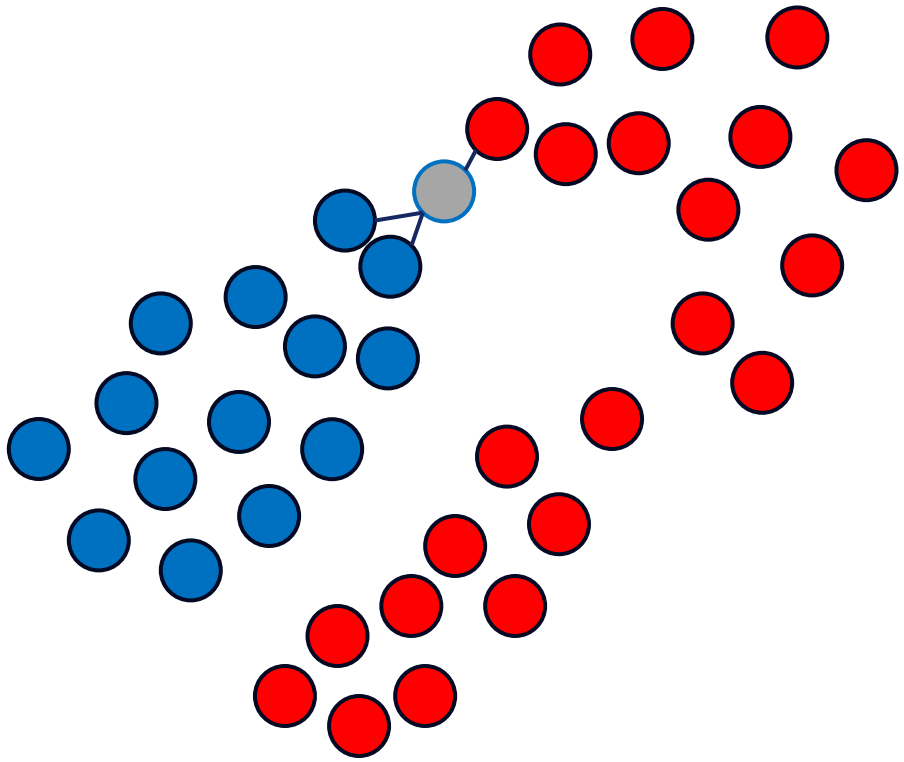
MONTANA STATE UNIVERSITY

K-NEAREST NEIGHBOR CLASSIFICATION

Given: Data with class labels (blue and red points)

We want to find: Class label of new data instance (grey point)

$k = 3$: Blue Label



Find k nearest neighbors of the new data instance, and assign it the majority class label

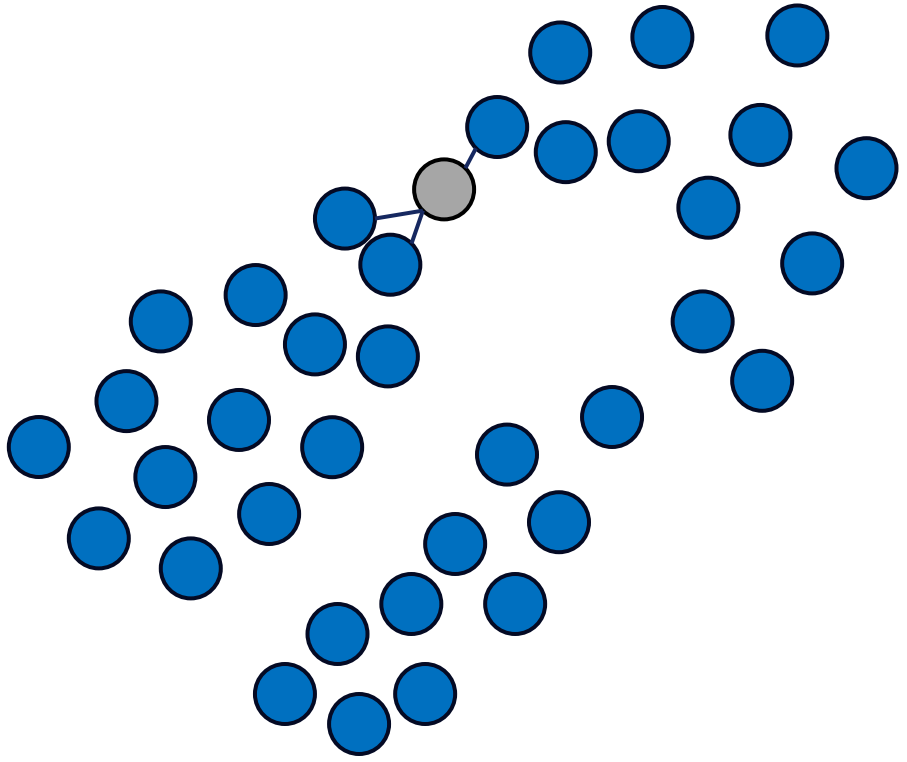
MONTANA STATE UNIVERSITY

K-NEAREST NEIGHBOR ALGORITHM FOR ANOMALY DETECTION

Given: Data Matrix

We want to find: Anomaly score for data instance (grey point)

$k = 3$:
Anomaly
score small



Find k nearest neighbors of the new data instance, and assign it an anomaly score that is its average distance to its nearest neighbors

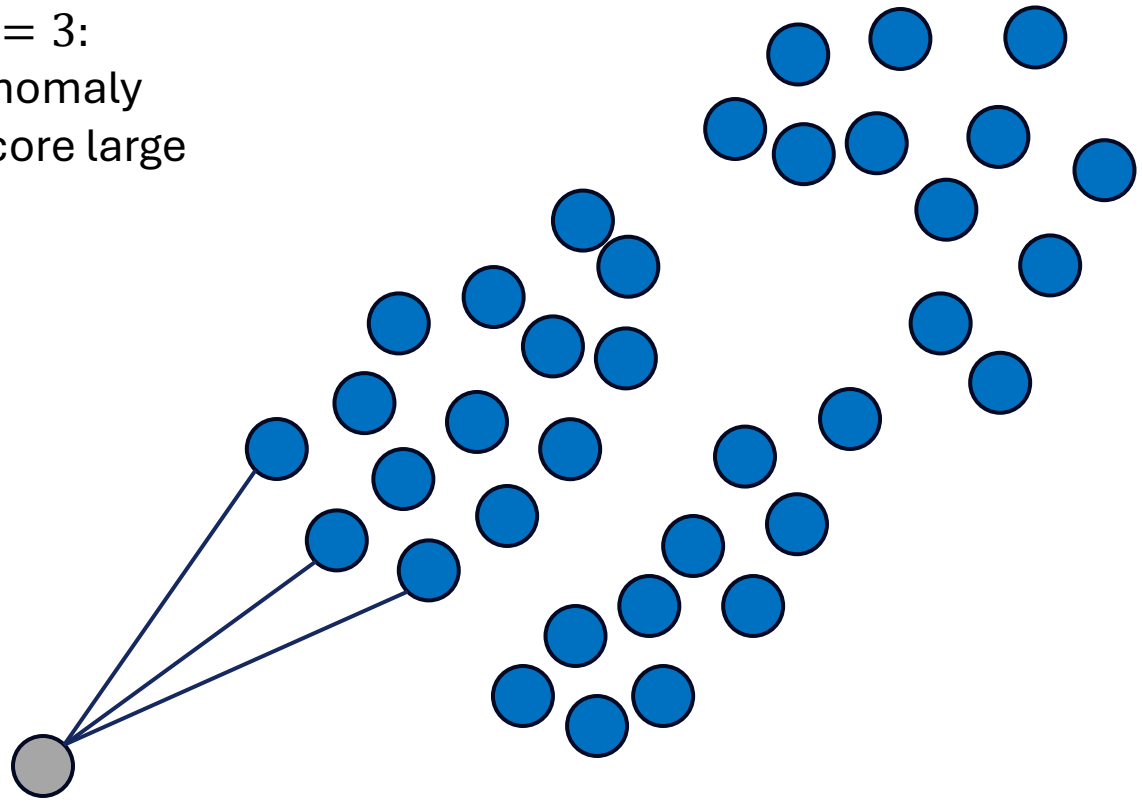
MONTANA STATE UNIVERSITY

K-NEAREST NEIGHBOR ALGORITHM FOR ANOMALY DETECTION

Given: Data Matrix

We want to find: Anomaly score for data instance (grey point)

$k = 3$:
Anomaly
score large



Find k nearest neighbors of the new data instance, and assign it an anomaly score that is its average distance to its nearest neighbors

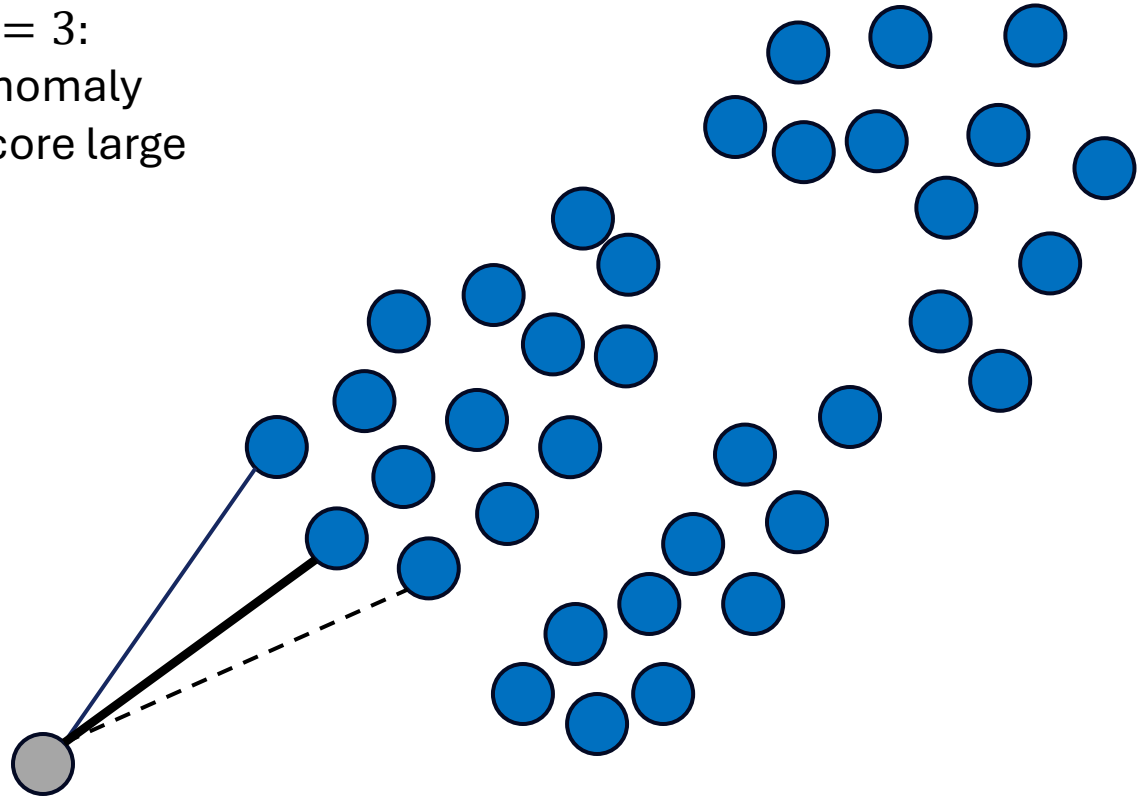
MONTANA STATE UNIVERSITY

K-NEAREST NEIGHBOR ALGORITHM FOR ANOMALY DETECTION

Given: Data Matrix

We want to find: Anomaly score for data instance (grey point)

$k = 3$:
Anomaly
score large



**Find k nearest neighbors of the new data instance,
and assign it an anomaly score that
is its average distance to its nearest neighbors**

Note: Common variants are distance to kth nearest neighbor or median distance to k nearest neighbors

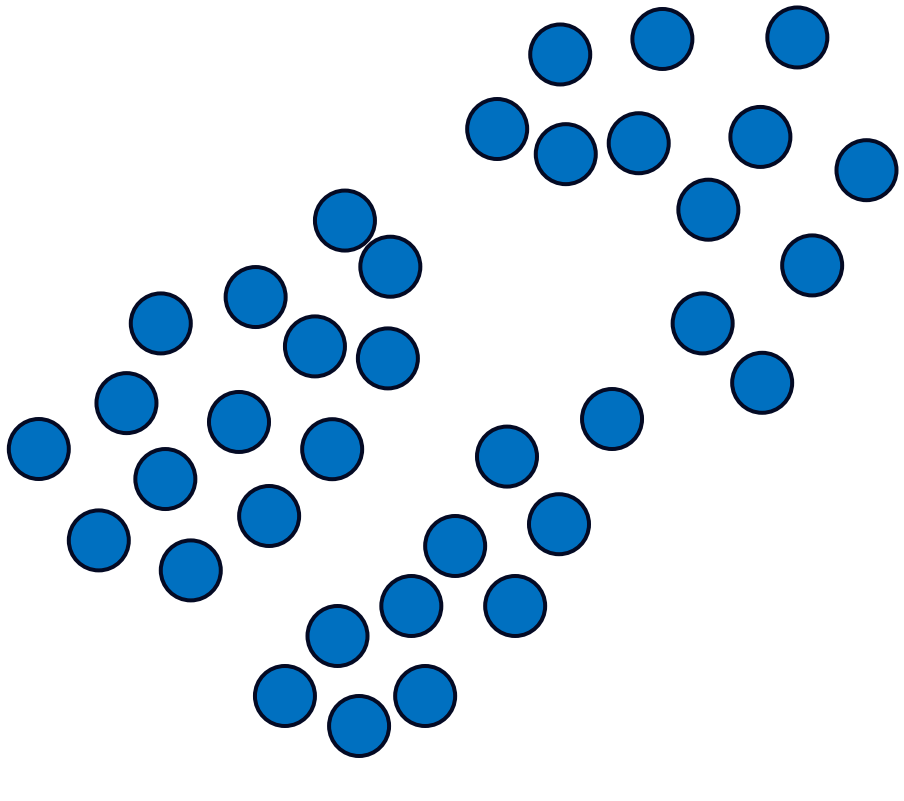
MONTANA STATE UNIVERSITY

K-NEAREST NEIGHBOR ALGORITHM FOR ANOMALY DETECTION

Given: Data Matrix

We want to find: Anomaly score for data instance (grey point)

$k = 3$:
Anomaly score large



Find k nearest neighbors of the new data instance, and assign it an anomaly score that is its average distance to its nearest neighbors

Choosing an appropriate threshold score for anomalies is not always trivial.

MONTANA STATE UNIVERSITY

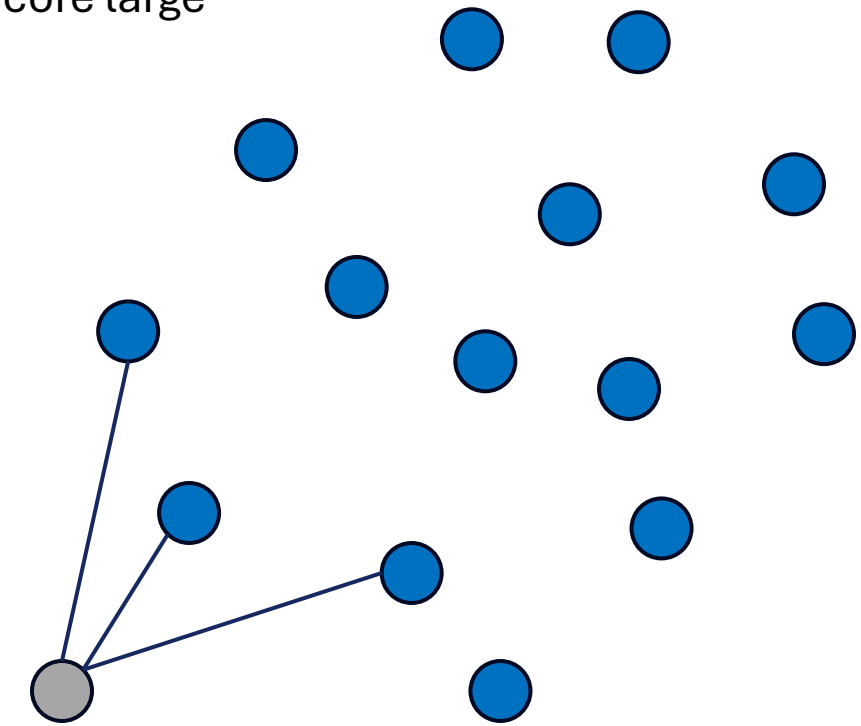
K-NEAREST NEIGHBOR ALGORITHM

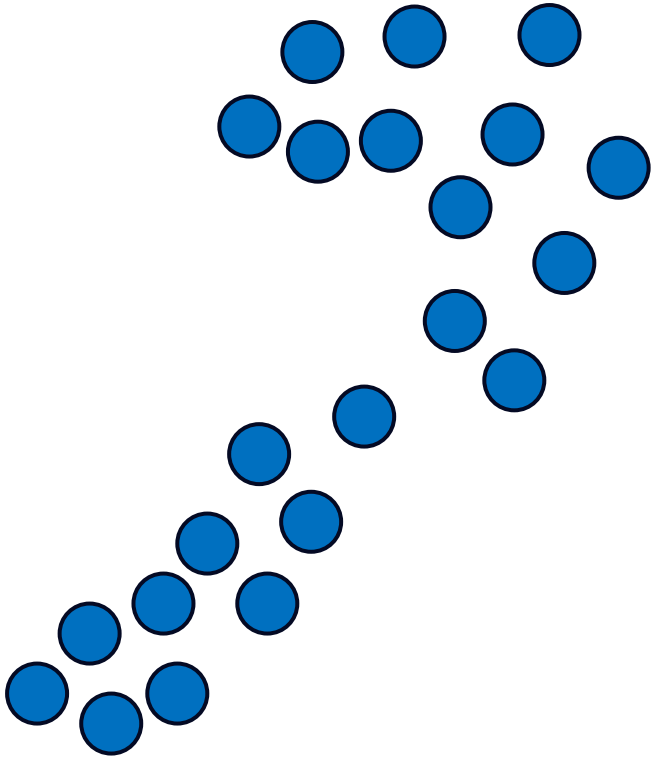
Challenge: Different densities for normal classes

Given: Data Matrix

We want to find: Anomaly score for data instance (grey point)

$k = 3$: Anomaly score large



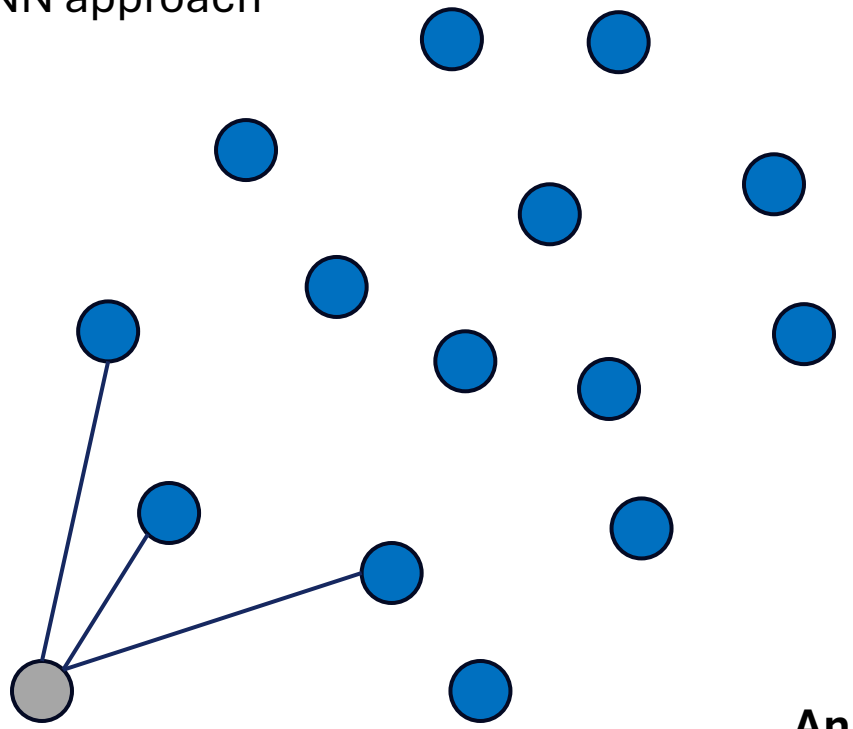


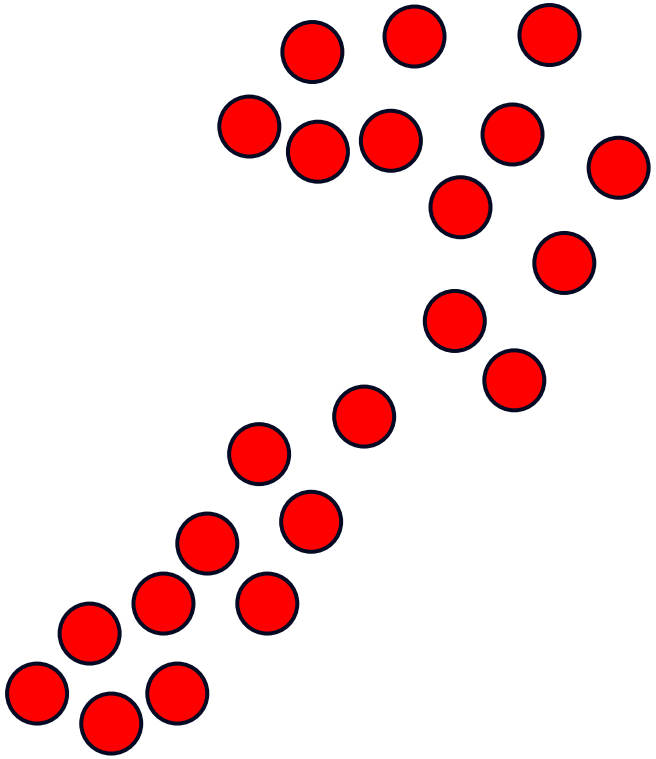
MONTANA STATE UNIVERSITY

Local outlier factor algorithm (LOF)

Instead of just looking at k nearest neighbors, consider the local density of points surrounding a data instance.

$k = 3$: large Anomaly score for KNN approach





An Anomaly has distances to neighbors that are atypical of its neighbors.

MONTANA STATE UNIVERSITY

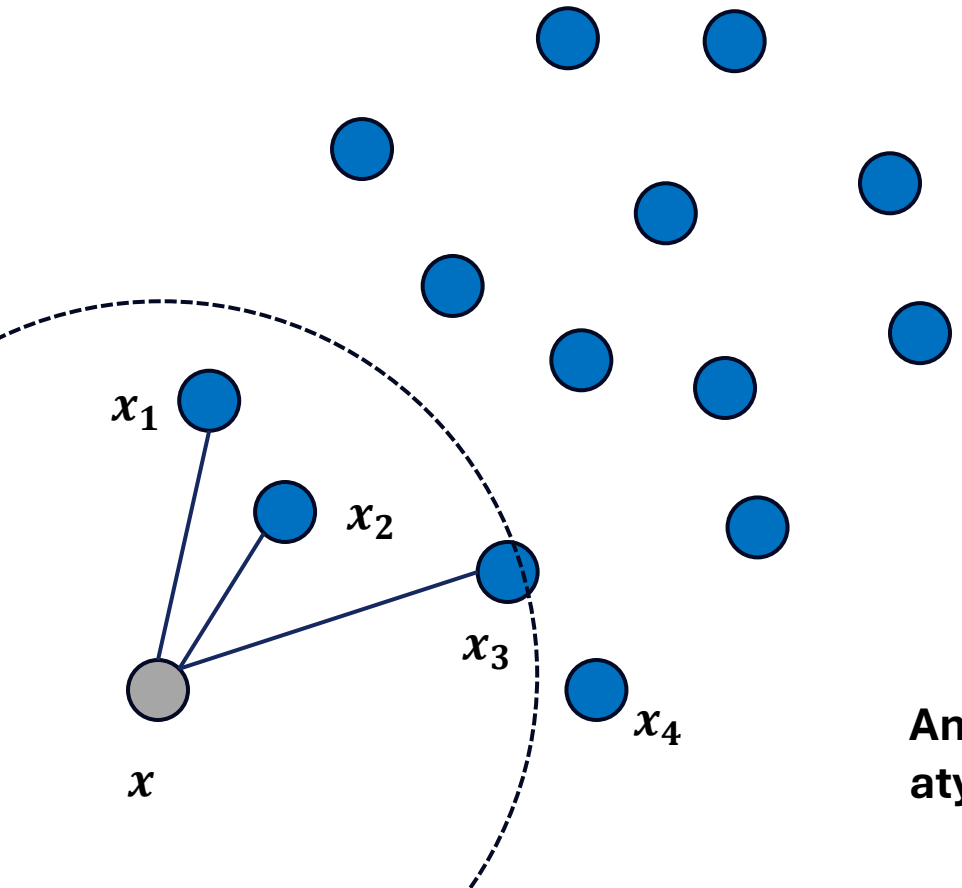
Local Outlier Factor (LOF)

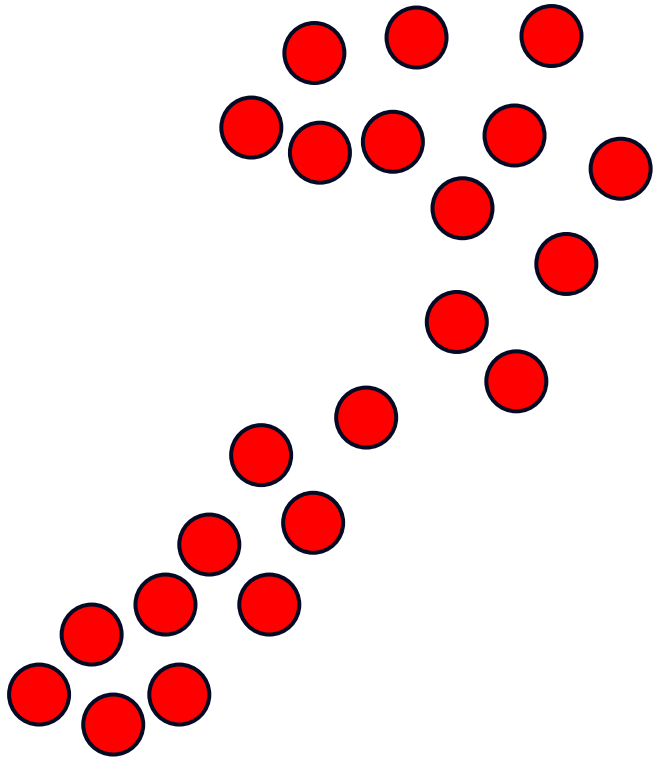
- Ratio (comparison) of the average density of the k-NN of an observation to the density of the observation itself.
  - > 1 means more likely to be an anomaly
  - <1 means less likely to be an anomaly
- What is density?
  - Inverse of the average reachability (distances) from observation to all its k-NN

Local outlier factor algorithm (LOF)

Reachability distance from x to x_i:

$$reachdist_k(x, x_i) = max\{dist_k(x_i), dist(x, x_i)\}$$





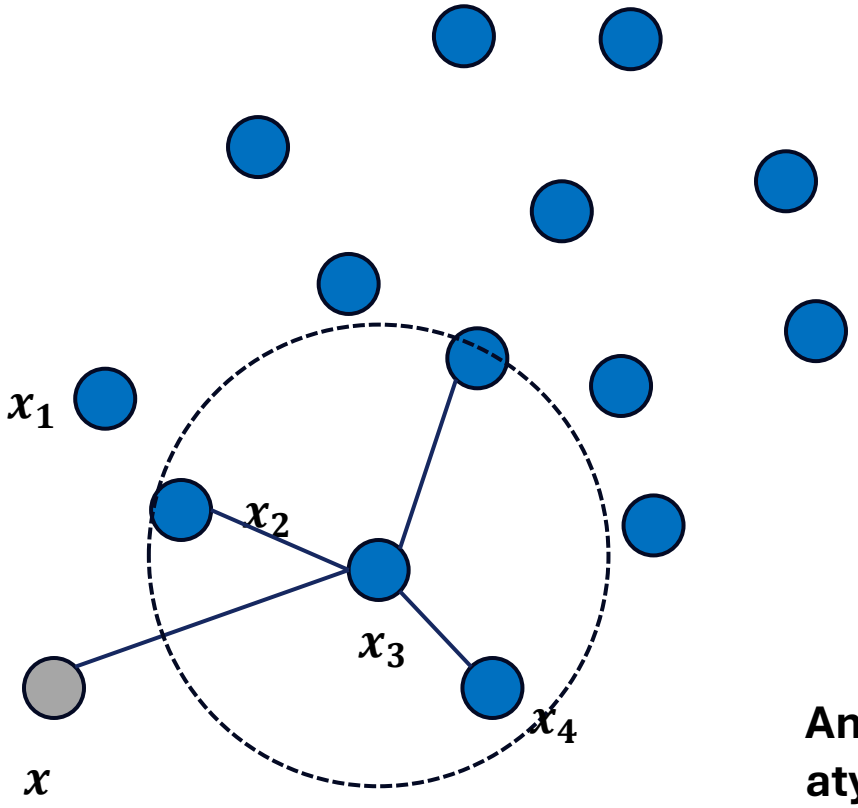
An Anomaly has distances to neighbors that are atypical of its neighbors.

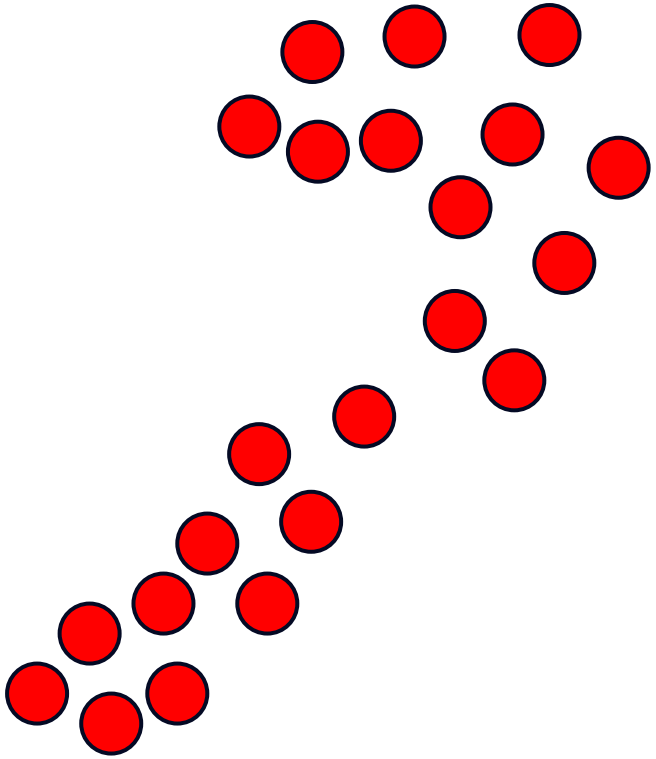
MONTANA STATE UNIVERSITY

Local outlier factor algorithm (LOF)

Reachability distance from x to x_i:

$$reachdist_k(x, x_3) = max\{dist_k(x_3), dist(x_3, x_i)\} = dist(x, x_3)$$





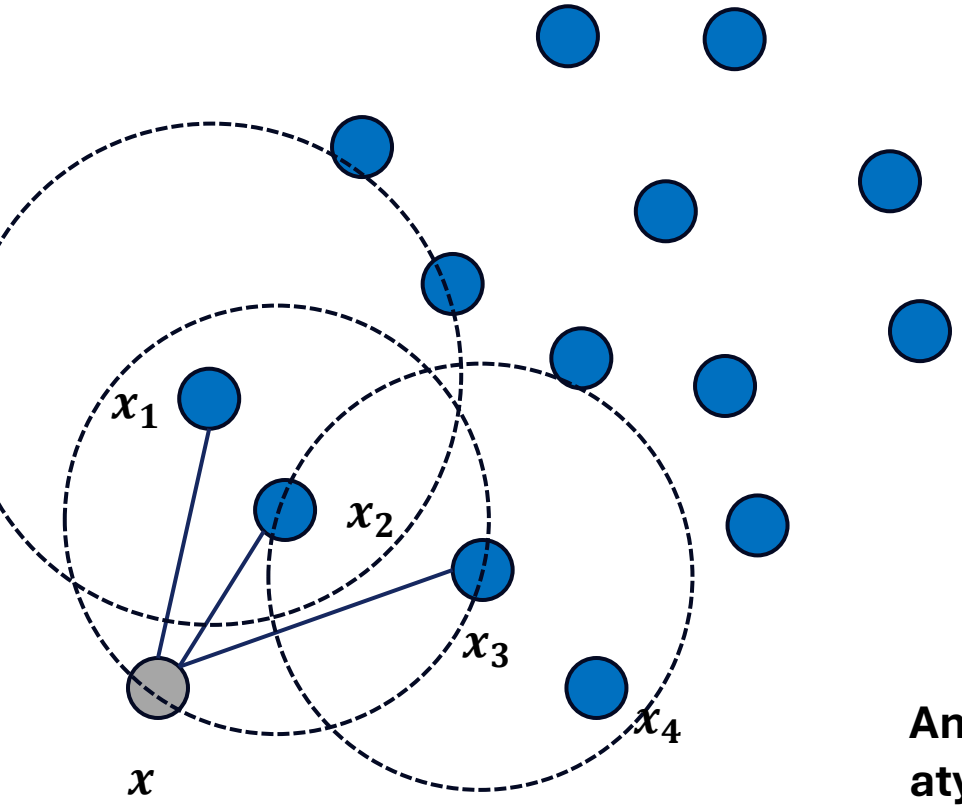
An Anomaly has distances to neighbors that are atypical of its neighbors.

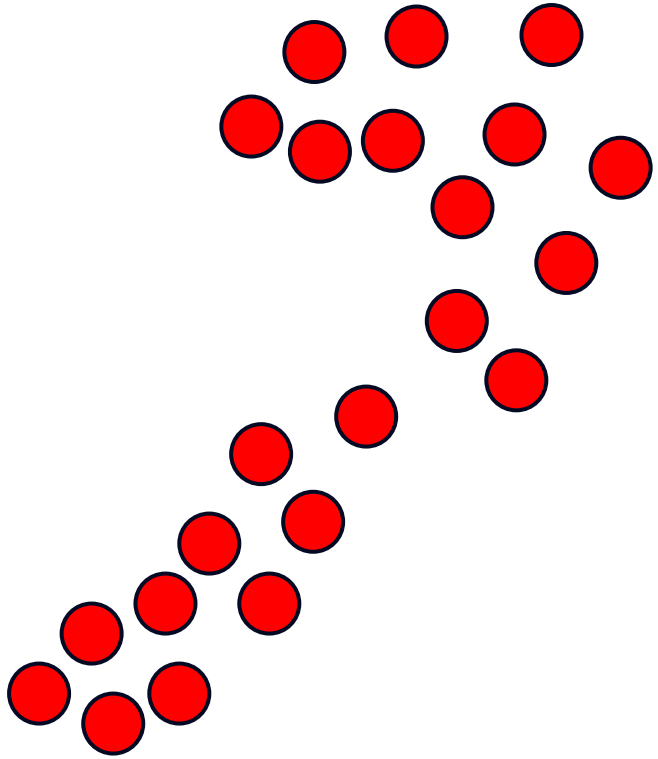
MONTANA STATE UNIVERSITY

Local outlier factor algorithm (LOF)

Local reachability density of a point x:

$$lrd_k(x) = \left(\frac{\sum_{x_i \in N_k(x)} reachdist_k(x, x_i)}{|N_k(x)|} \right)^{-1}$$





An Anomaly has distances to neighbors that are atypical of its neighbors.

MONTANA STATE UNIVERSITY

Local outlier factor algorithm (LOF)

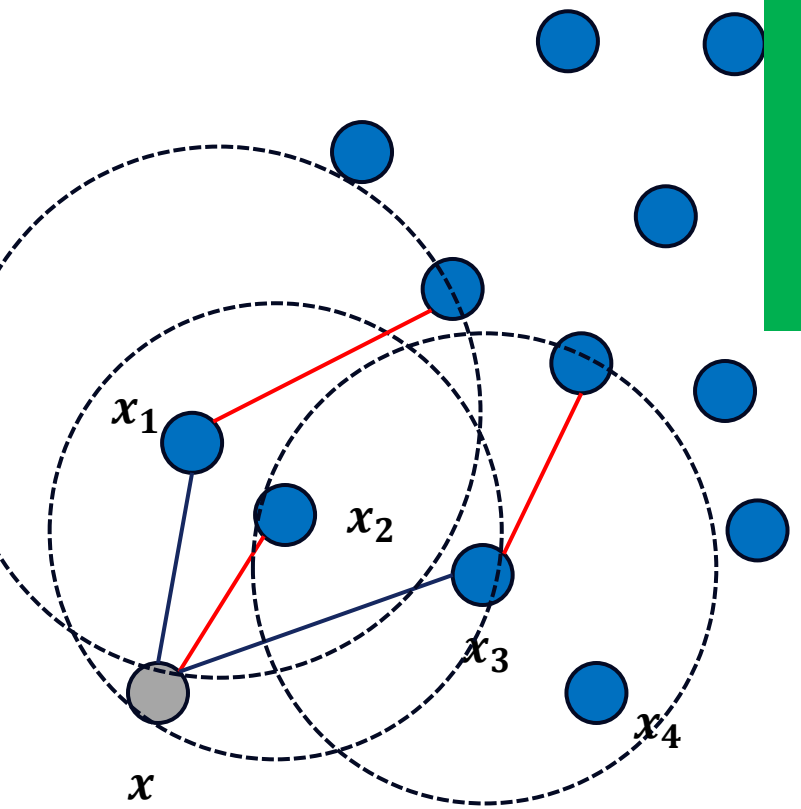
Local reachability density of a point x:

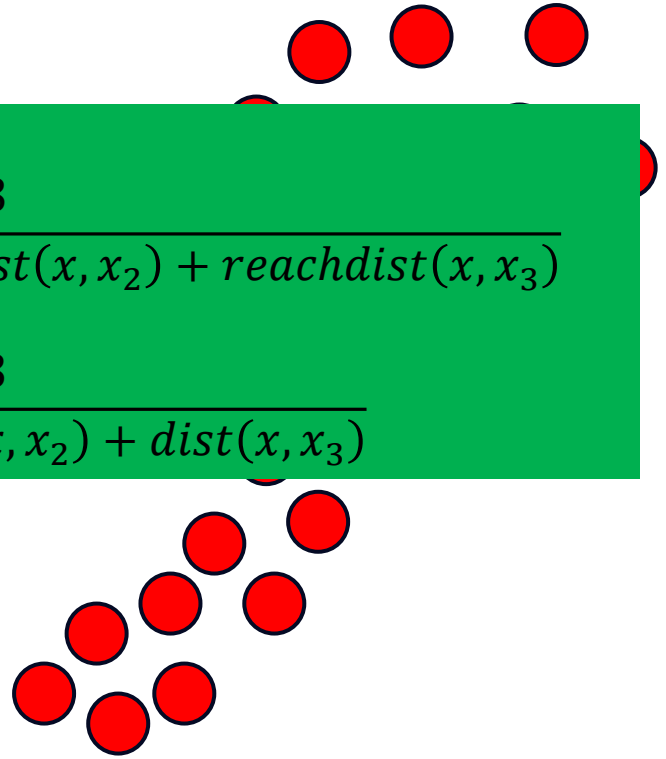
$$lrd_k(x) = \left(\frac{\sum_{x_i \in N_k(x)} reachdist_k(x, x_i)}{|N_k(x)|} \right)^{-1}$$

$x's\ lrd_3\ is:$

$$lrd_3(x) = \frac{3}{reachdist(x, x_1) + reachdist(x, x_2) + reachdist(x, x_3)}$$

$$lrd_3(x) = \frac{3}{dist_3(x_1) + dist(x, x_2) + dist(x, x_3)}$$





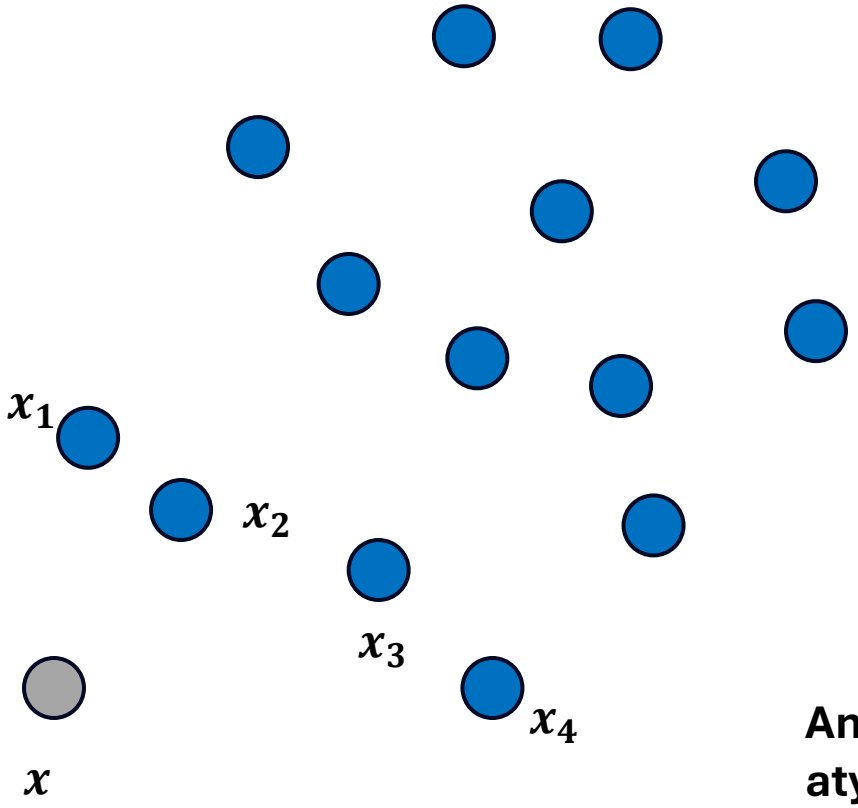
An Anomaly has distances to neighbors that are atypical of its neighbors.

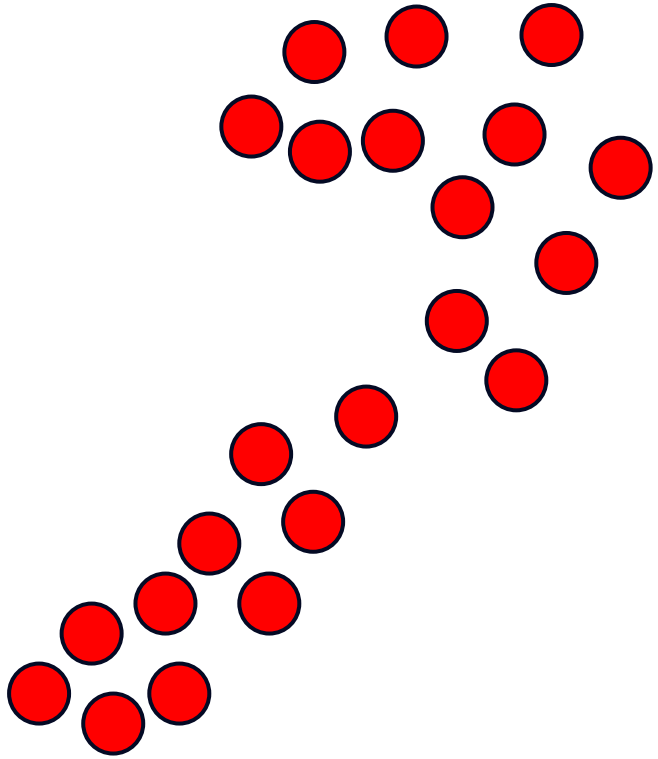
MONTANA STATE UNIVERSITY

Local outlier factor algorithm (LOF)

Local reachability density of a point x:

$$LOF_k(x) = \frac{\sum_{x_i \in N_k(x)} \frac{lrd_k(x_i)}{lrd_k(x)}}{|N_k(x)|}$$





An Anomaly has distances to neighbors that are atypical of its neighbors.

MONTANA STATE UNIVERSITY

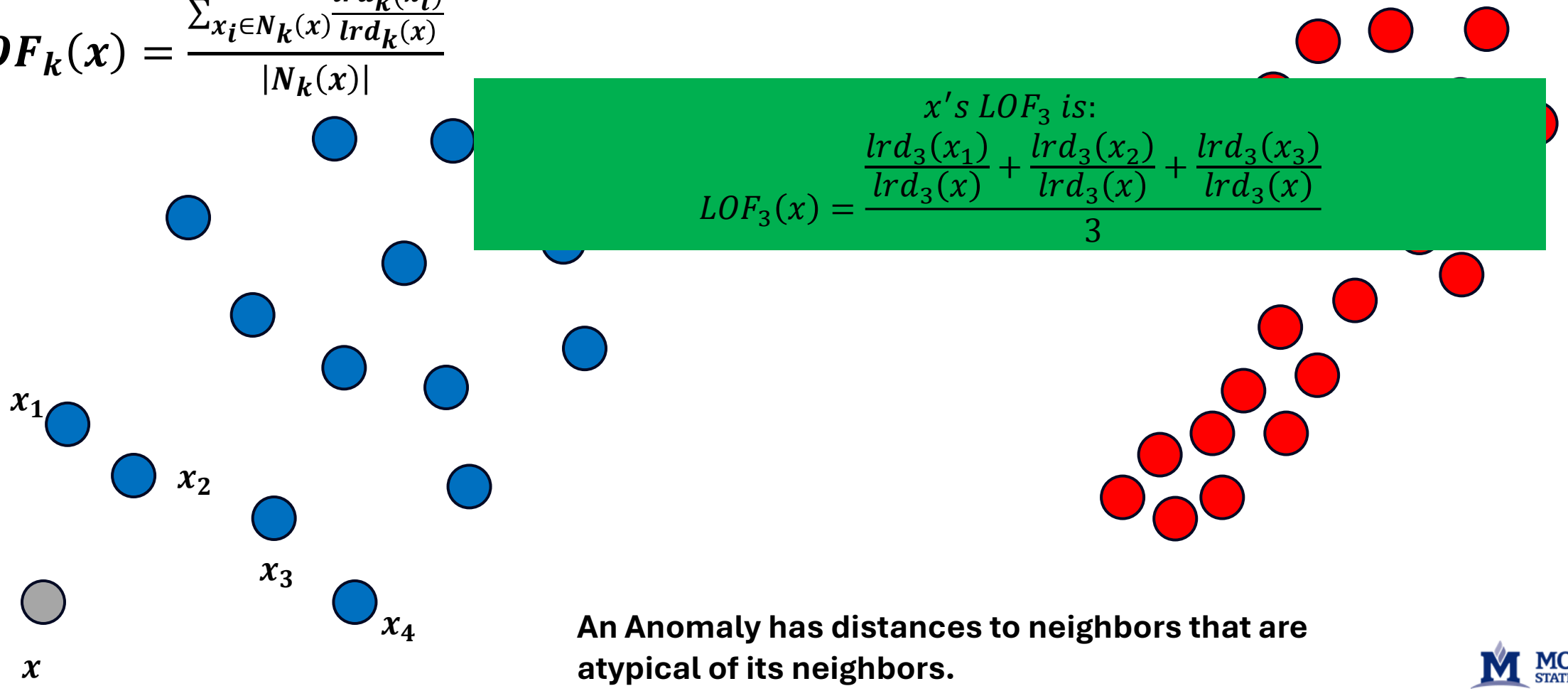
Local outlier factor algorithm (LOF)

Local reachability density of a point x:

$$LOF_k(x) = \frac{\sum_{x_i \in N_k(x)} \frac{lrd_k(x_i)}{lrd_k(x)}}{|N_k(x)|}$$

$$x's\ LOF_3\ is:$$

$$LOF_3(x) = \frac{\frac{lrd_3(x_1)}{lrd_3(x)} + \frac{lrd_3(x_2)}{lrd_3(x)} + \frac{lrd_3(x_3)}{lrd_3(x)}}{3}$$



An Anomaly has distances to neighbors that are atypical of its neighbors.

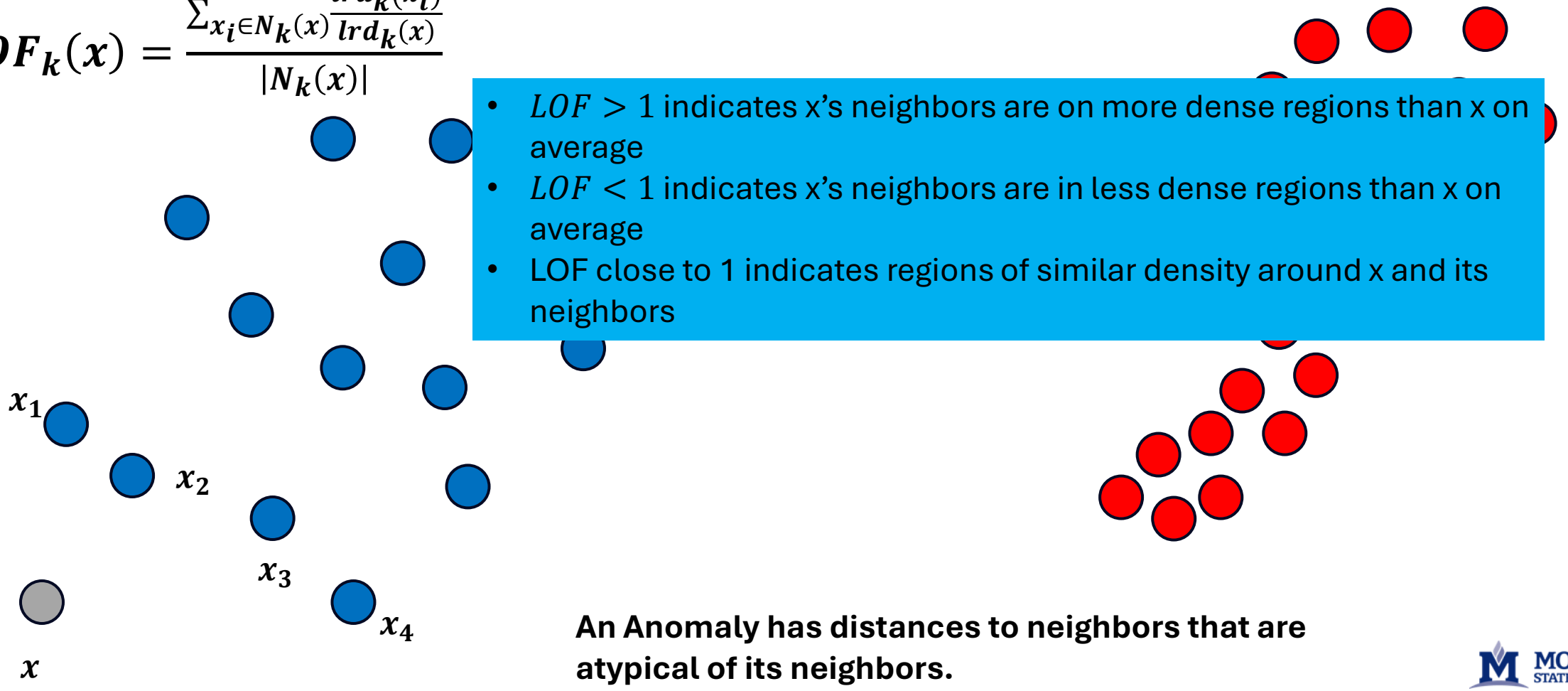
MONTANA STATE UNIVERSITY

Local outlier factor algorithm (LOF)

Local reachability density of a point x:

$$LOF_k(x) = \frac{\sum_{x_i \in N_k(x)} \frac{lrd_k(x_i)}{lrd_k(x)}}{|N_k(x)|}$$

- $LOF > 1$ indicates x's neighbors are on more dense regions than x on average
- $LOF < 1$ indicates x's neighbors are in less dense regions than x on average
- LOF close to 1 indicates regions of similar density around x and its neighbors



An Anomaly has distances to neighbors that are atypical of its neighbors.

MONTANA STATE UNIVERSITY