Introduction to Clustering

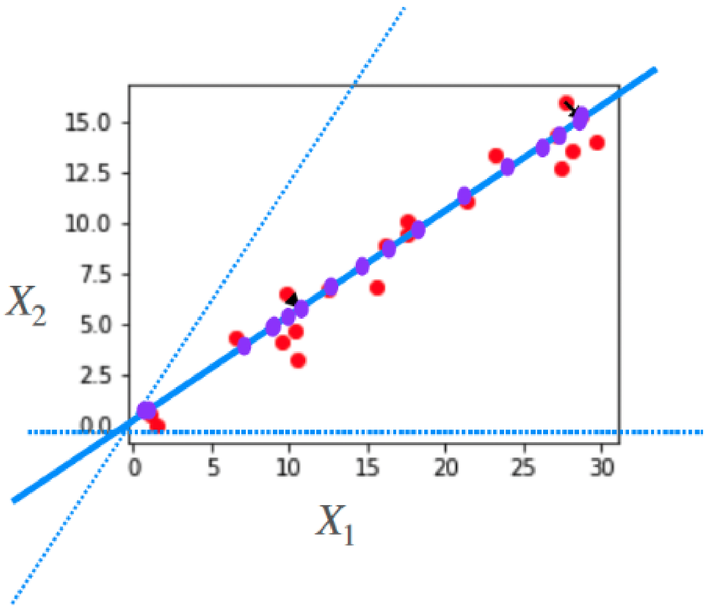
CSCI 347

Adiesha Liyana Ralalage

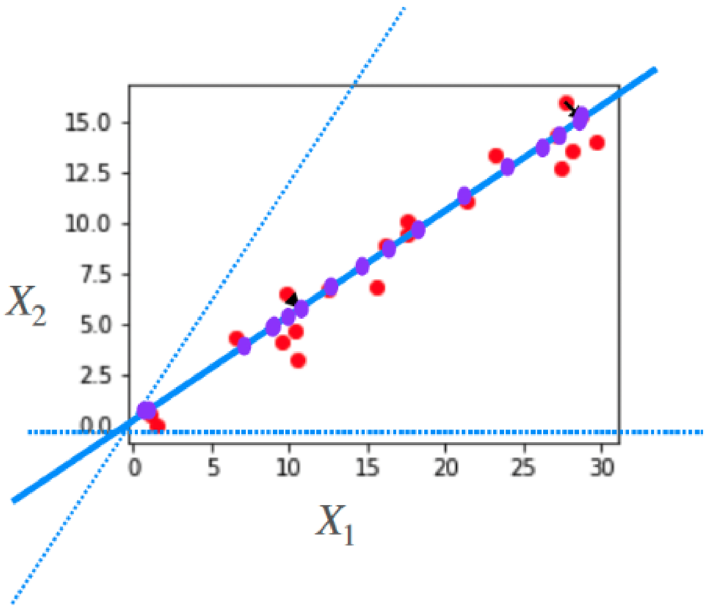
PCA summary

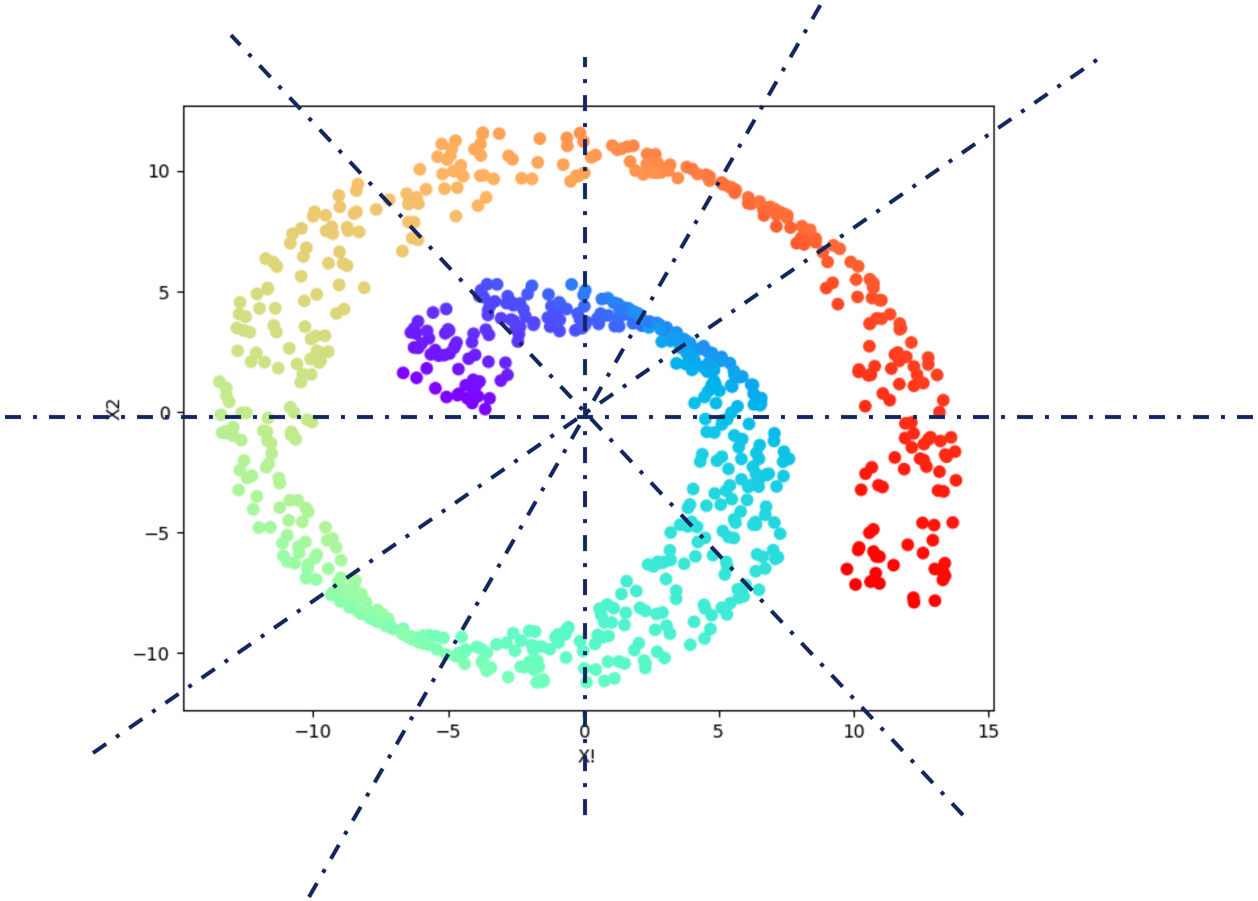
- A good “go-to” method for dimensionality reduction.
  - Advantages
    - Preserves a large fraction of the total variance.
    - Creates independent (uncorrelated) new attributes
    - Fast (very useful on large datasets)
  - Disadvantages
    - Not Interpretable (what does $0.8X_1 + 2X_2$ mean?)
    - Sensitive to scaling (need to do standardization or normalization)
    - Fails to capture nonlinear relationships.

Data often lies close to one dimensional linear space



Data often lies close to one dimensional linear space





MONTANA STATE UNIVERSITY

What can we use for nonlinear dimensionality reduction?

- Kernel PCA
- Laplacian eigenmaps
- Locally linear embedding (*)
- T-SNE (t-distributed Stochastic Neighbor Embedding)
    - Looks at local structure
    - It converts the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities.
- UMAP (Uniform Manifold Approximation and Projection)
    - based on the idea that high-dimensional data often lies on or near a lower-dimensional manifold.
- ISOMAP (Isometric Mapping)
    - Aims to preserve the geodesic distances between data points.
    - Geodesic Distance: measures the shortest distance along the curved manifold.
    - ISOMAP assumes that the data lies on or near a lower-dimensional manifold embedded in a high-dimensional space. It tries to "unfold" this manifold to reveal its true structure.

Locally linear embedding

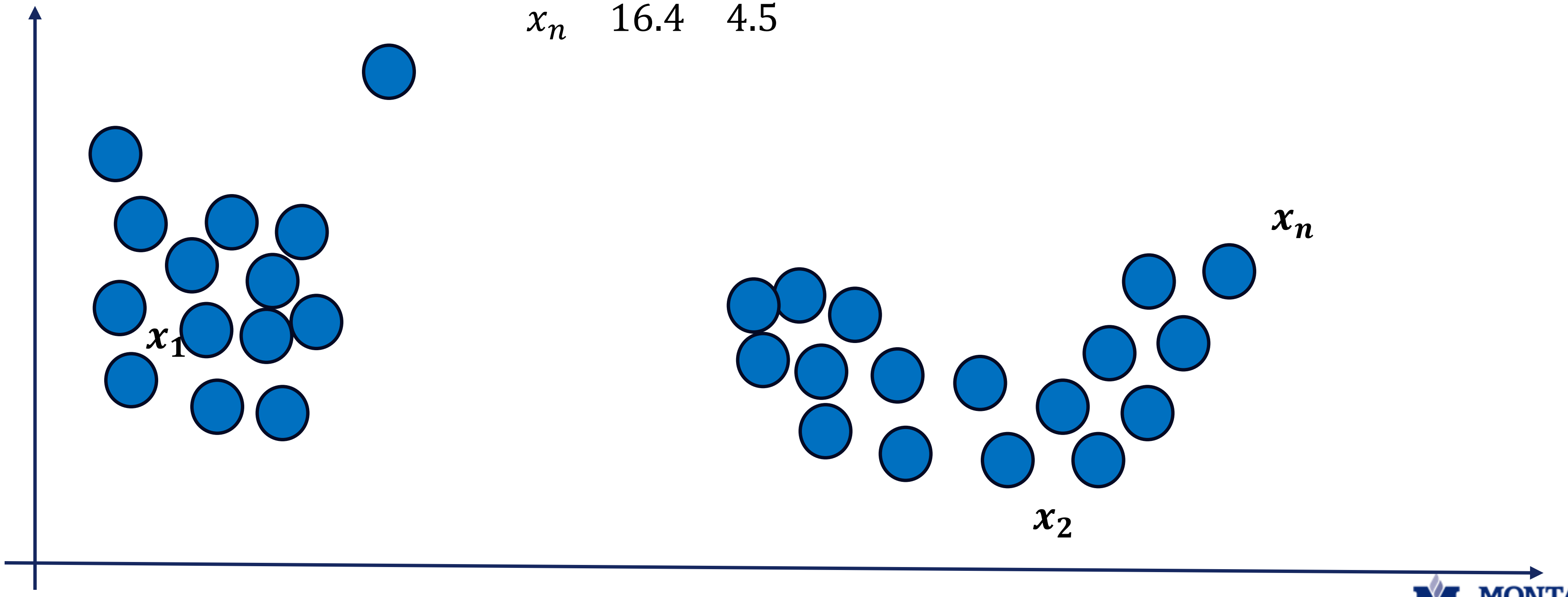
Key Concepts:

- Local Linearity: LLE assumes that the data is locally linear, meaning that each data point can be approximated as a linear combination of its neighbors.
- Neighborhood Preservation: The primary goal of LLE is to preserve the local neighborhood relationships in the low-dimensional embedding.
- Points that are close together in the high-dimensional space should remain close in the low-dimensional space.
- Weight Matrix: LLE constructs a weight matrix that captures the contribution of each neighbor in reconstructing a data point from its neighbors.
- Let’s look at an example for this.

What are clusters in a dataset

Data Matrix

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5



Clustering

- Clustering is the task of partitioning the points into natural groups.
- These natural groups are called clusters.
- Points within these clusters are very similar, whereas points across clusters are dissimilar as possible.

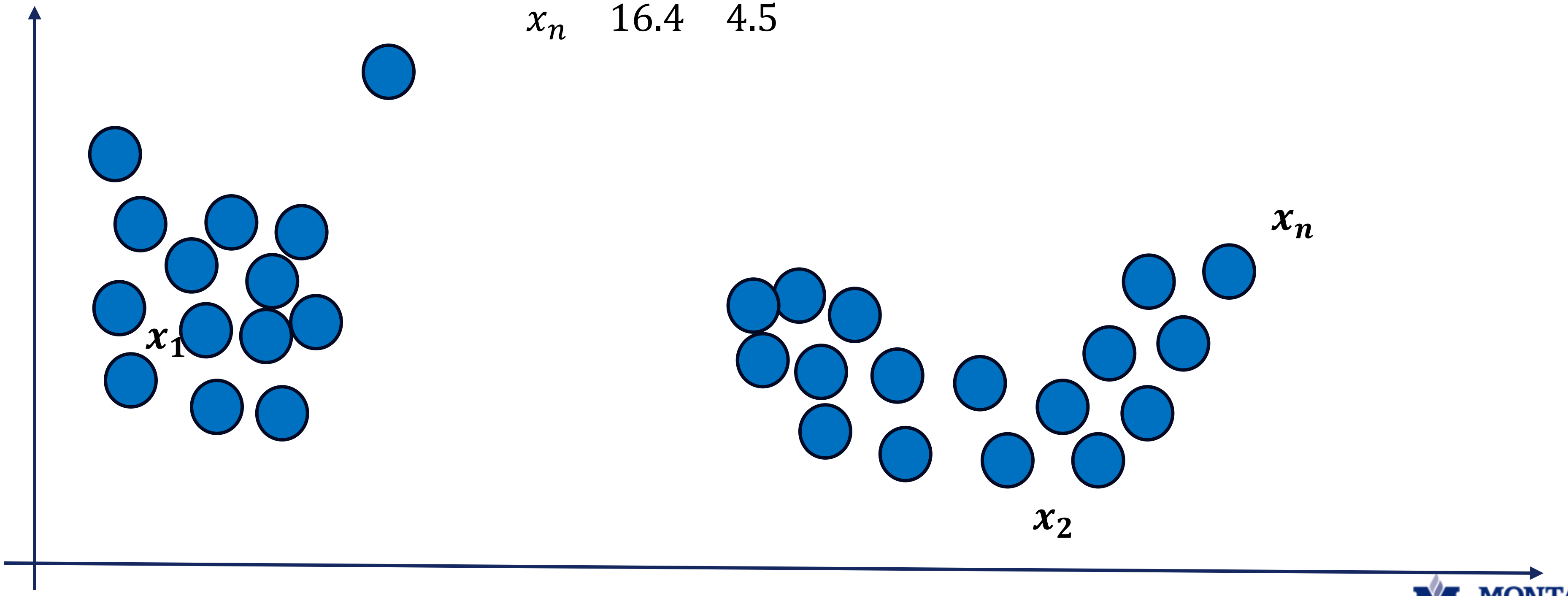
Applications of clustering

Movies

Data Matrix

People

$$\begin{array}{ccc} & X_1 & X_2 \\ x_1 & 2 & 3.2 \\ x_2 & 13.1 & 0.9 \\ \vdots & \vdots & \vdots \\ x_n & 16.4 & 4.5 \end{array}$$



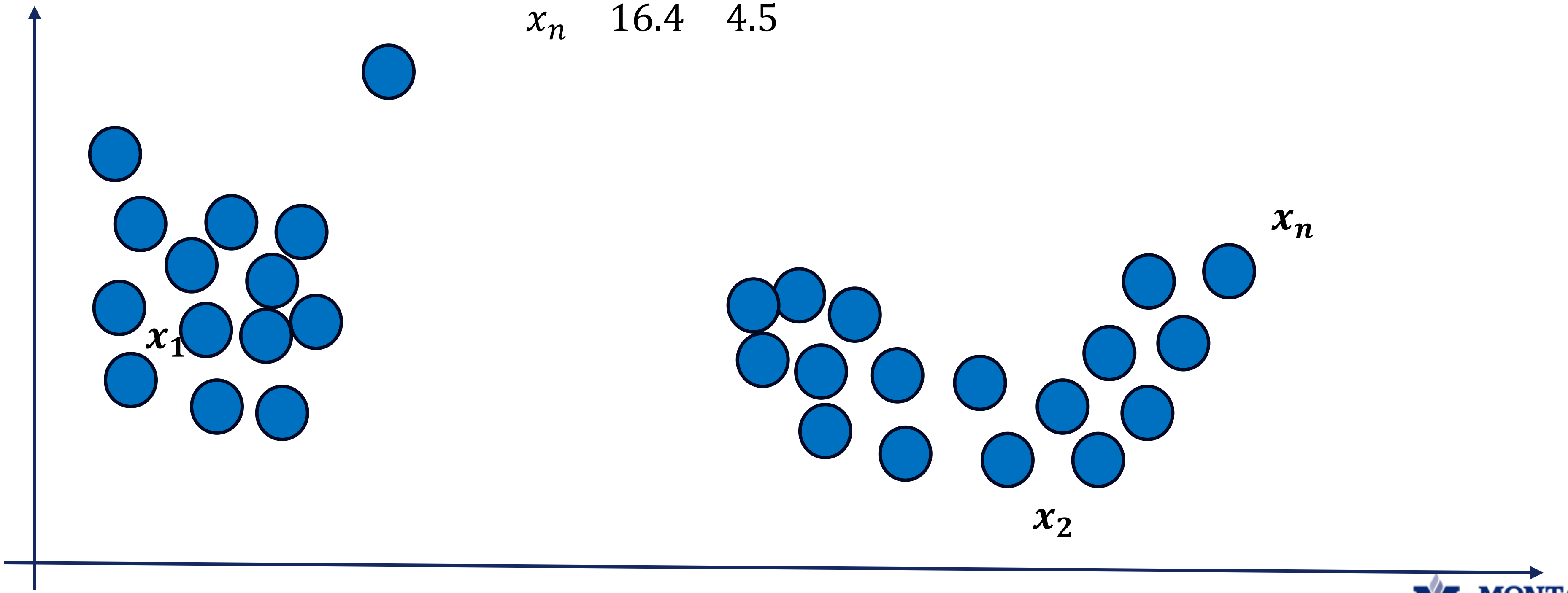
MONTANA STATE UNIVERSITY

Applications of clustering

Samples

Data Matrix

Genes	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5



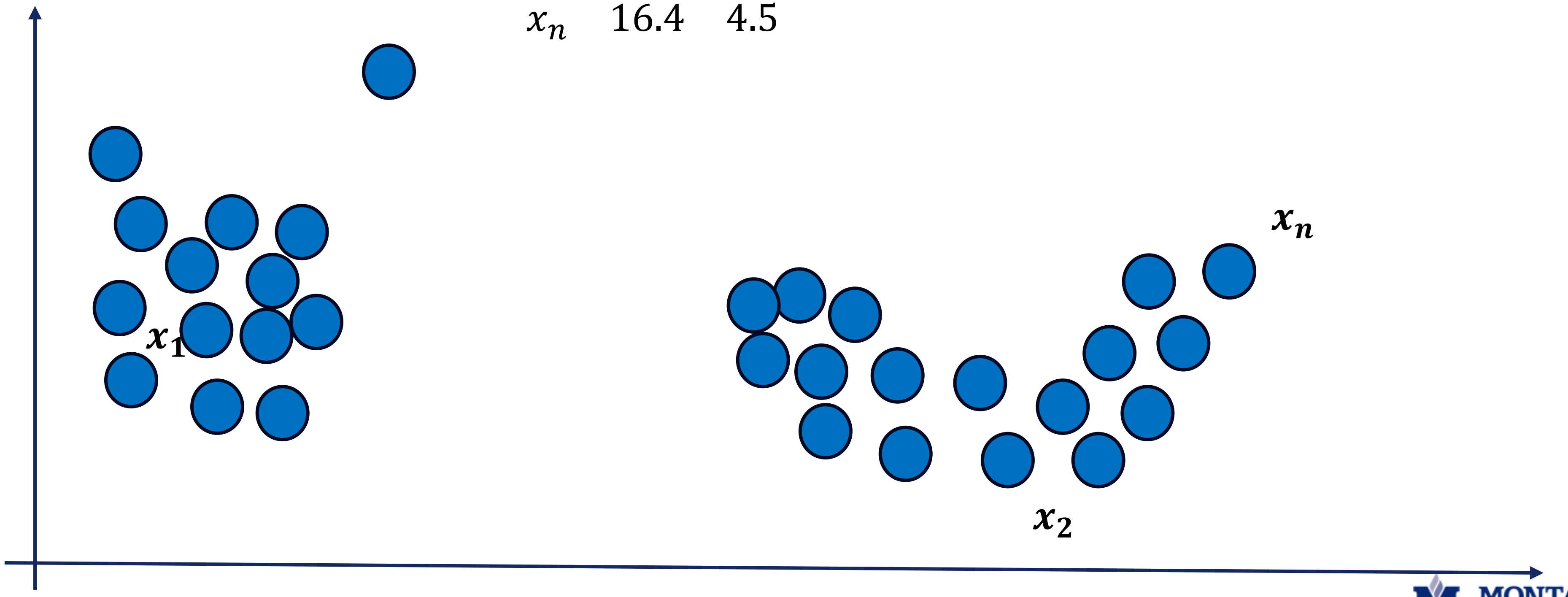
MONTANA STATE UNIVERSITY

Applications of clustering

Data Matrix

Words

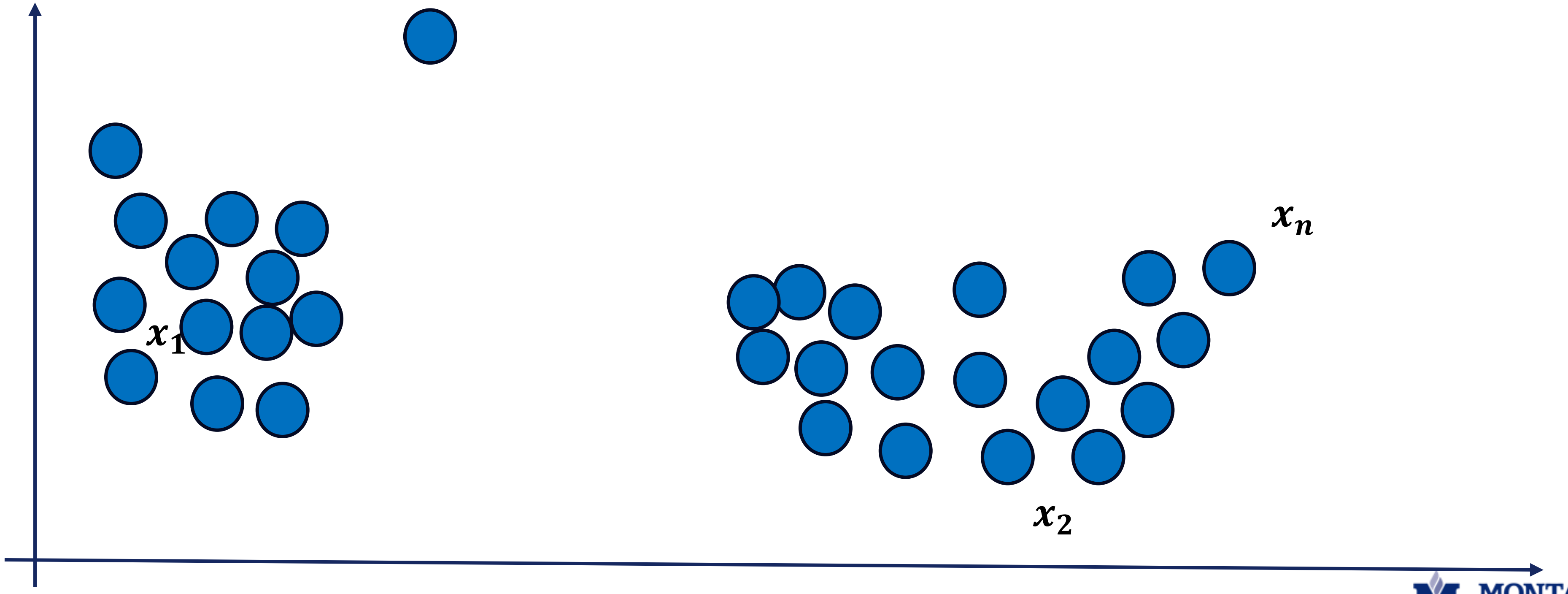
Documents	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5



How do we find clusters in a dataset?

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5

Our goal is to gather data instances into groups with high within-group similarity

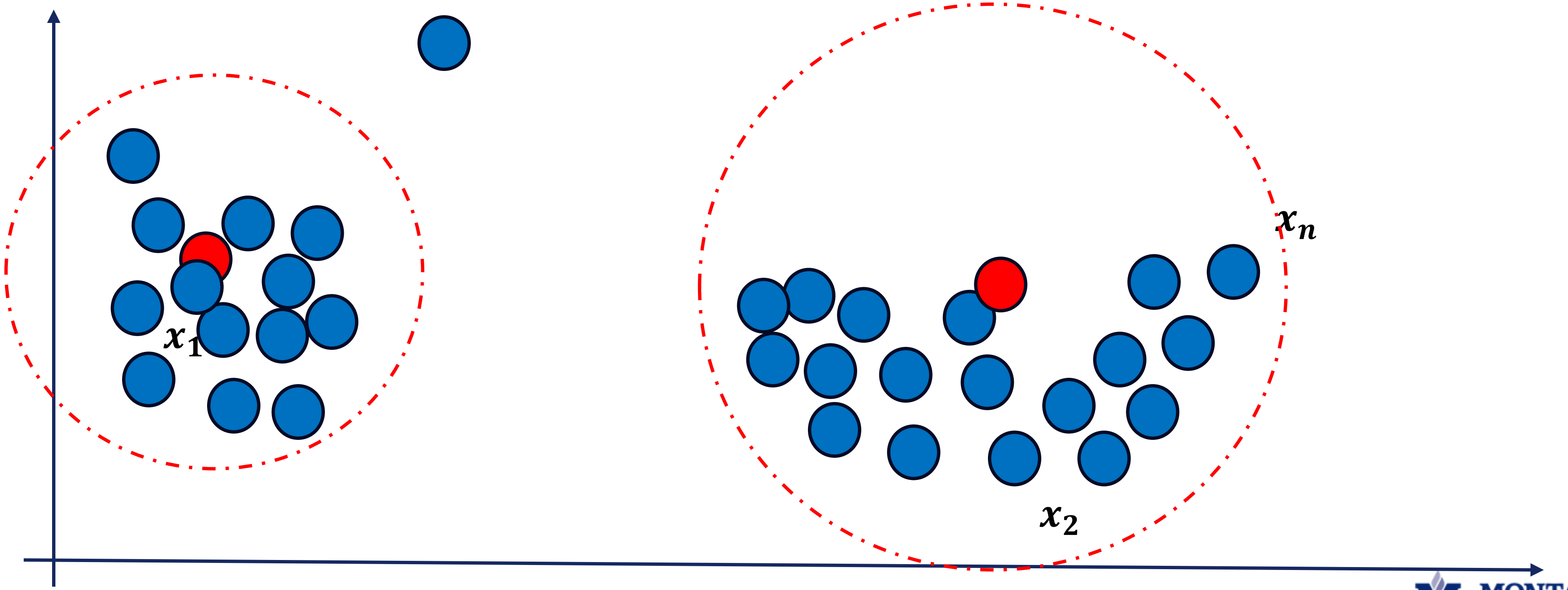


How do we find clusters in a dataset?

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5

Representative-based methods

Find a representative that best represents each cluster, and group points based on their closest representative.



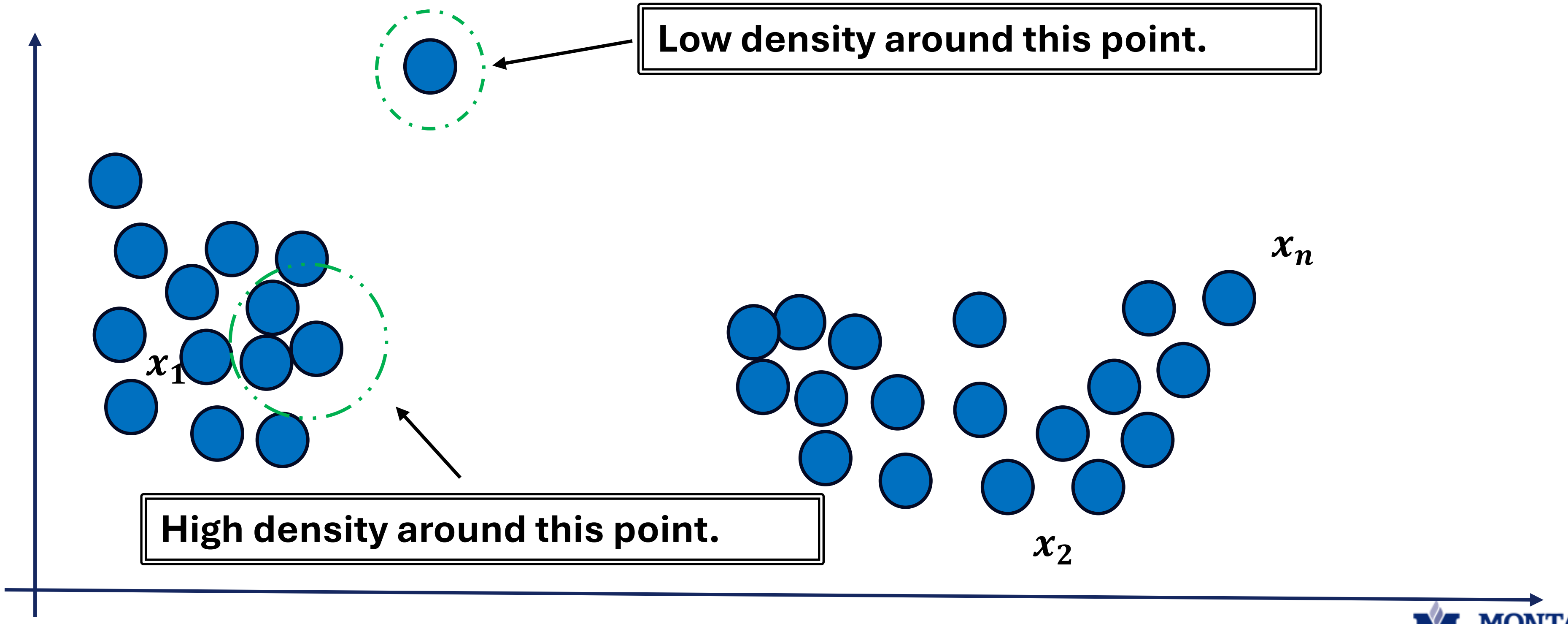
MONTANA STATE UNIVERSITY

How do we find clusters in a dataset?

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5

Density-based methods:

Find regions of high density (# points / some small volume)



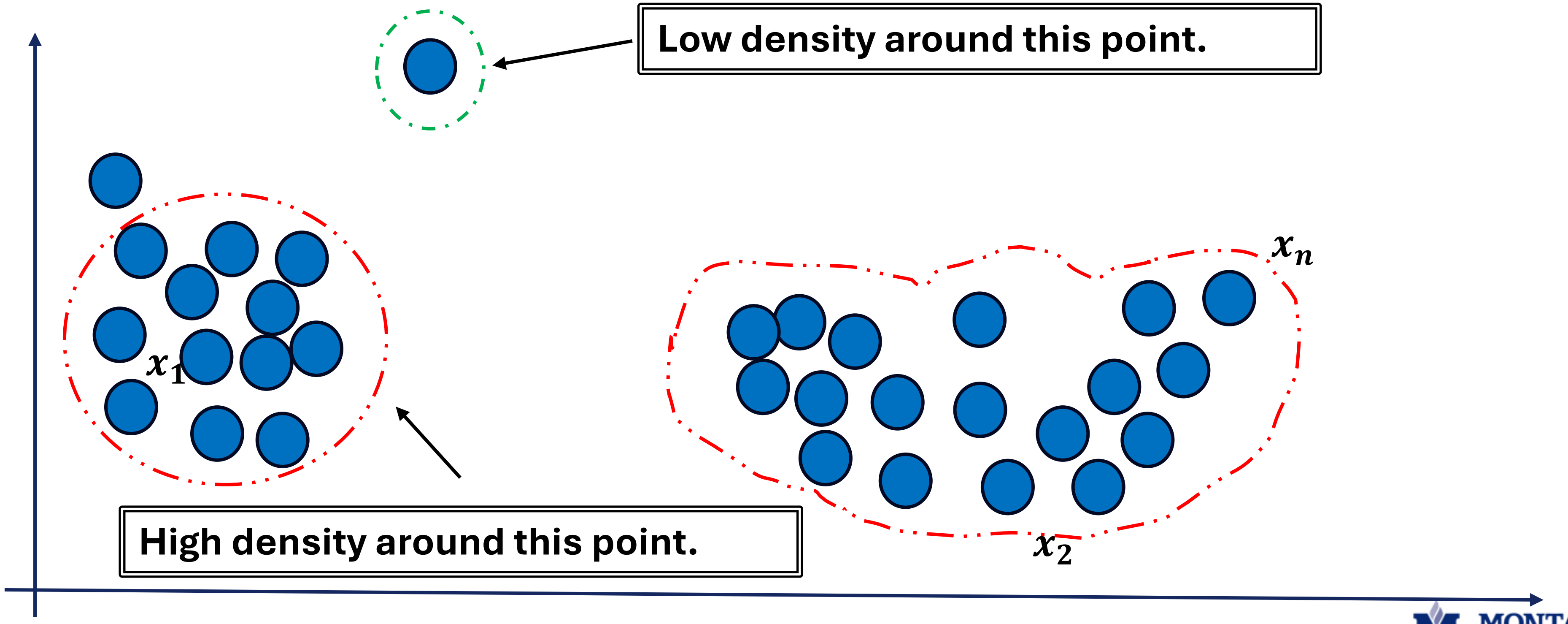
MONTANA STATE UNIVERSITY

How do we find clusters in a dataset?

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5

Density-based methods:

Find regions of high density (# points / some small volume)



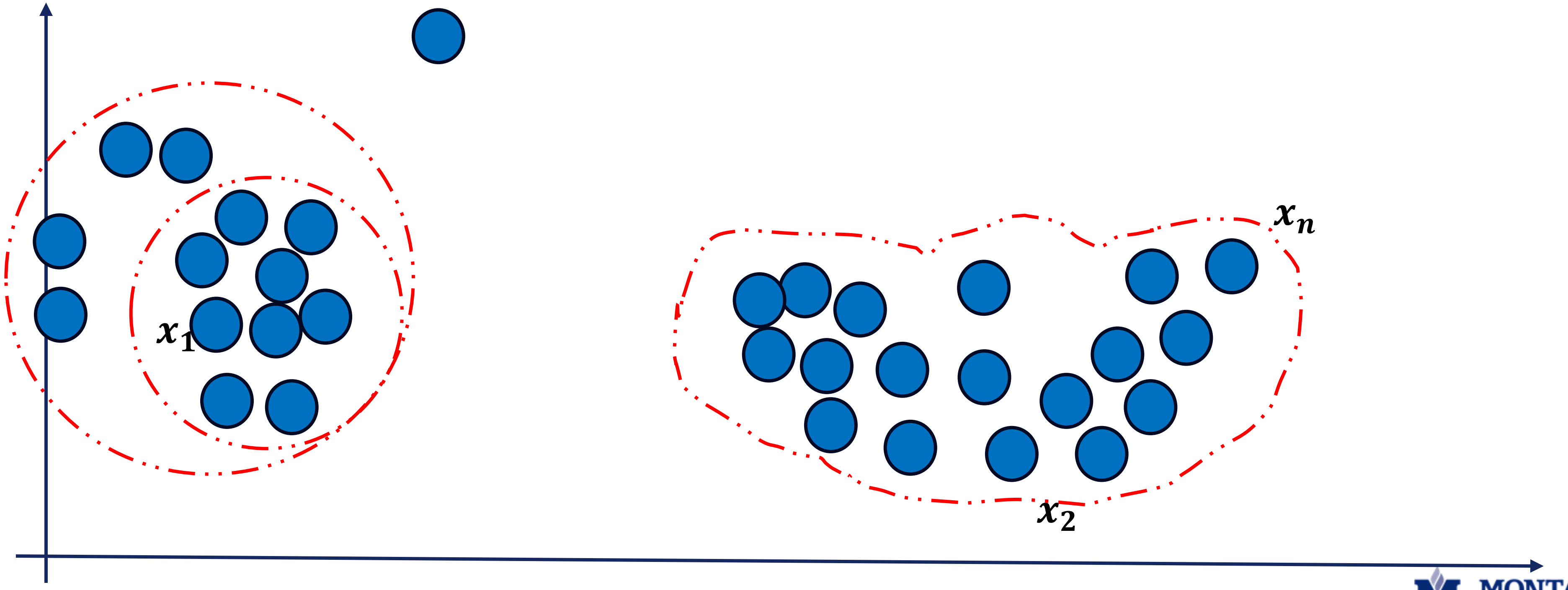
MONTANA STATE UNIVERSITY

How do we find clusters in a dataset?

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5

Hierarchical methods:

Clusters within clusters



How do we find clusters in a dataset?

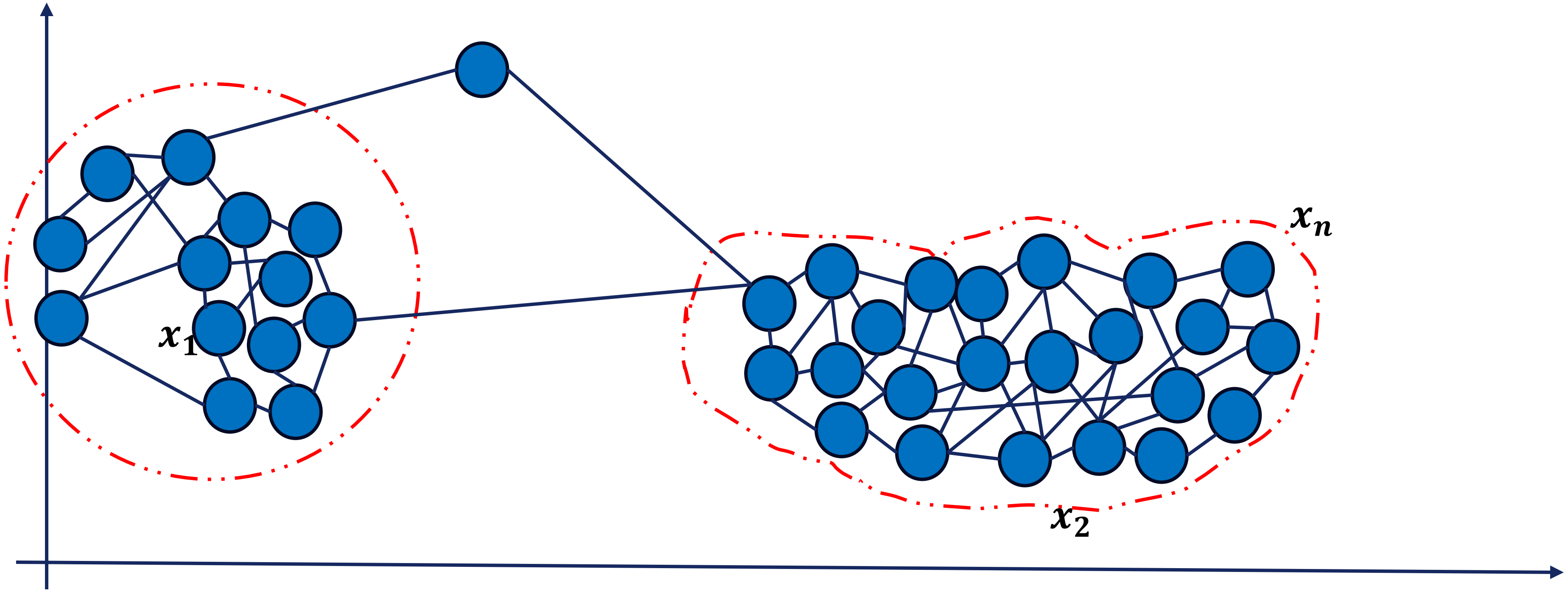
	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5

$$\longrightarrow$$

$$\begin{array}{c|cccc} & x_1 & x_2 & \cdots & x_n \\ \hline x_1 & 0 & 0 & \cdots & 0 \\ x_2 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n & 0 & 0 & \cdots & 0 \end{array}$$

Graph based methods:

Find subgraphs with high edge connectivity

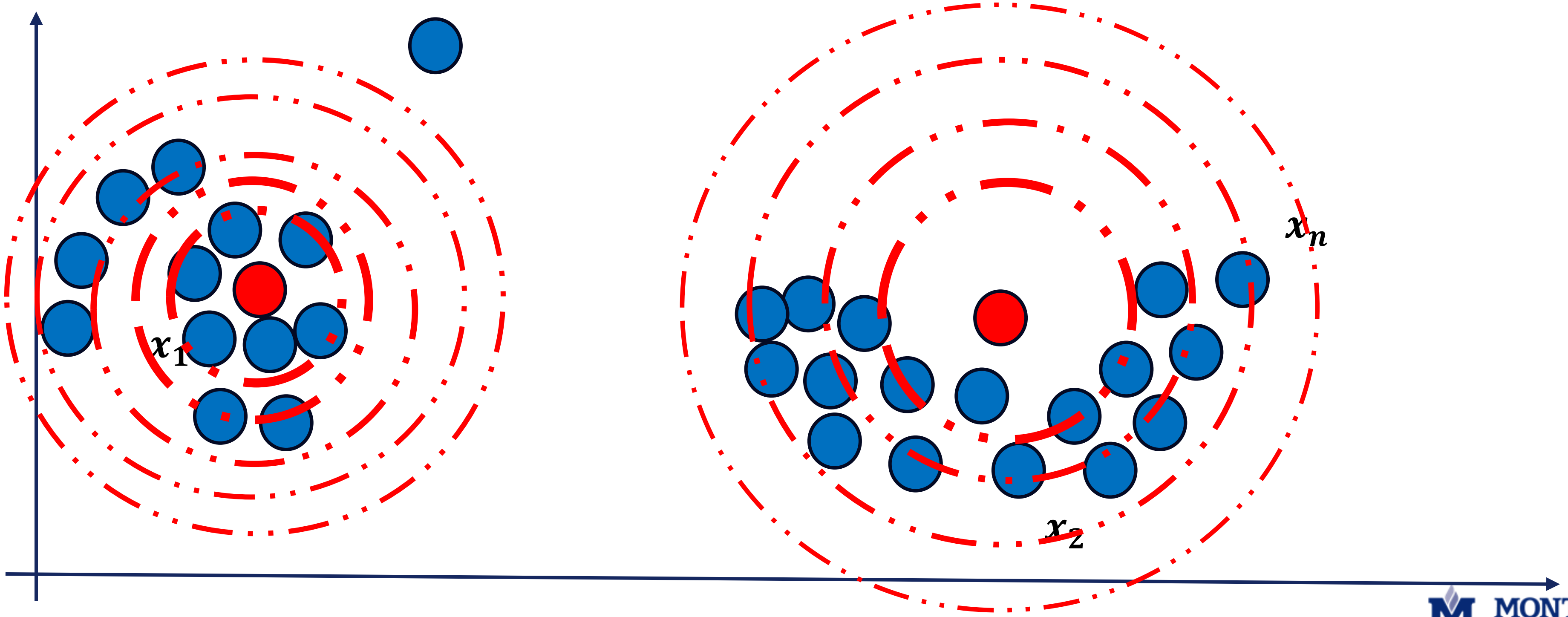


How do we find clusters in a dataset?

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
⋮	⋮	⋮
x_n	16.4	4.5

Soft clustering or probabilistic clustering

Estimate the probability distribution that the points come from



MONTANA STATE UNIVERSITY

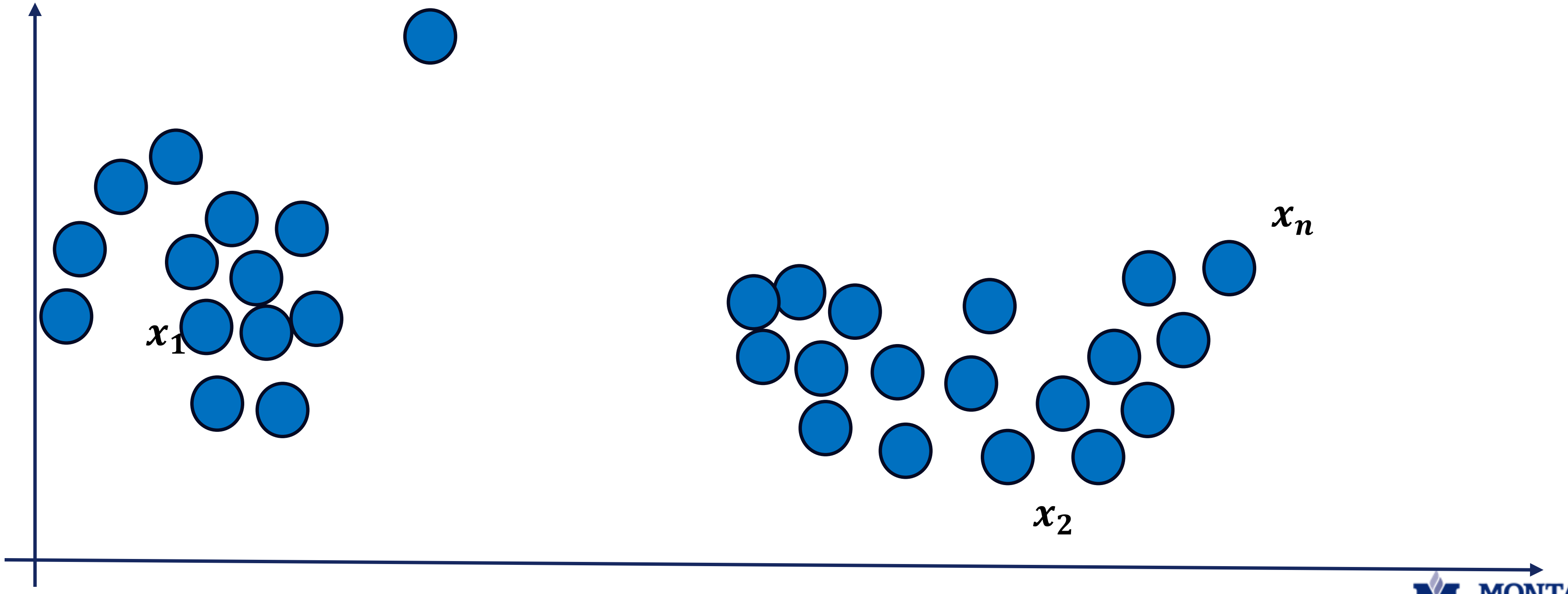
How do we find clusters in a dataset?

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5

	X_1
x_1	2
x_2	13.1
$\vdots$	$\vdots$
x_n	16.4

Spectral or subspace clustering methods:

Find a lower dimensional space that better represents the clusters



Clustering techniques

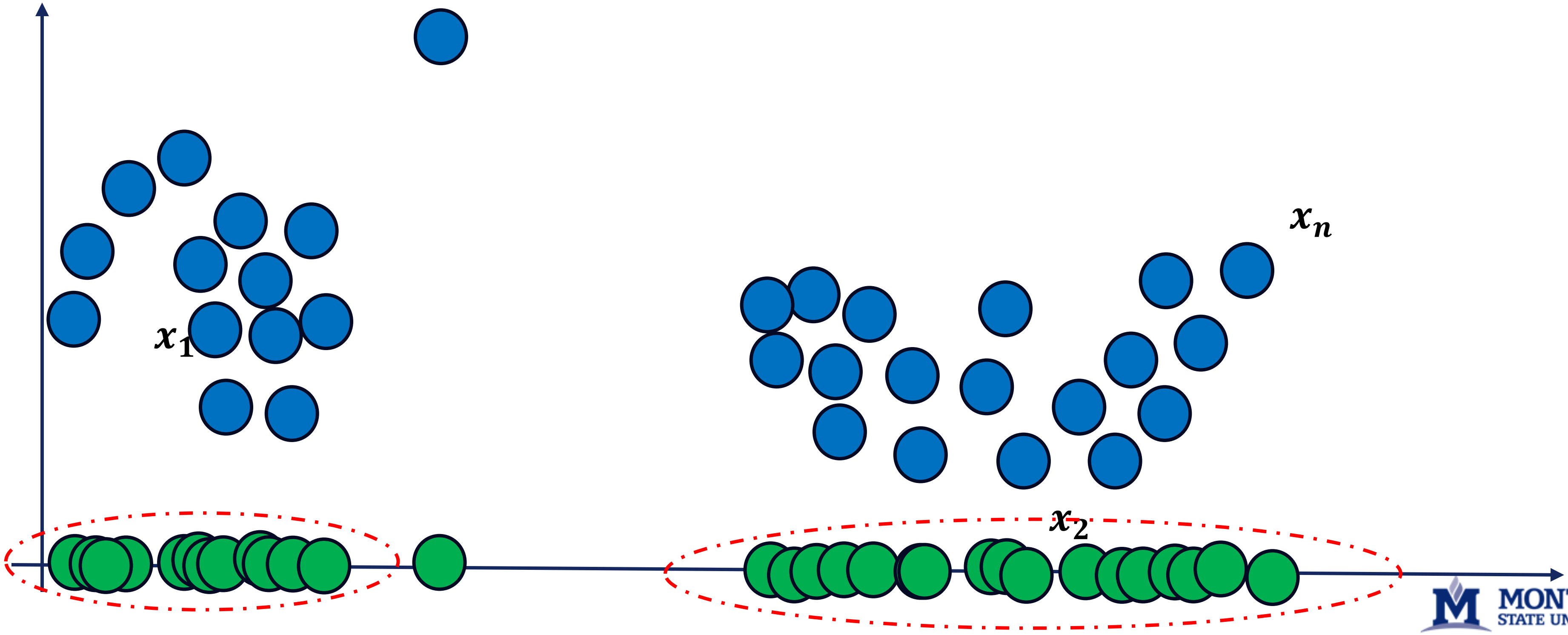
	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
$\vdots$	$\vdots$	$\vdots$
x_n	16.4	4.5

$\Rightarrow$

	X_1
x_1	2
x_2	13.1
$\vdots$	$\vdots$
x_n	16.4

Spectral or subspace clustering methods:

Find a lower dimensional space that better represents the clusters

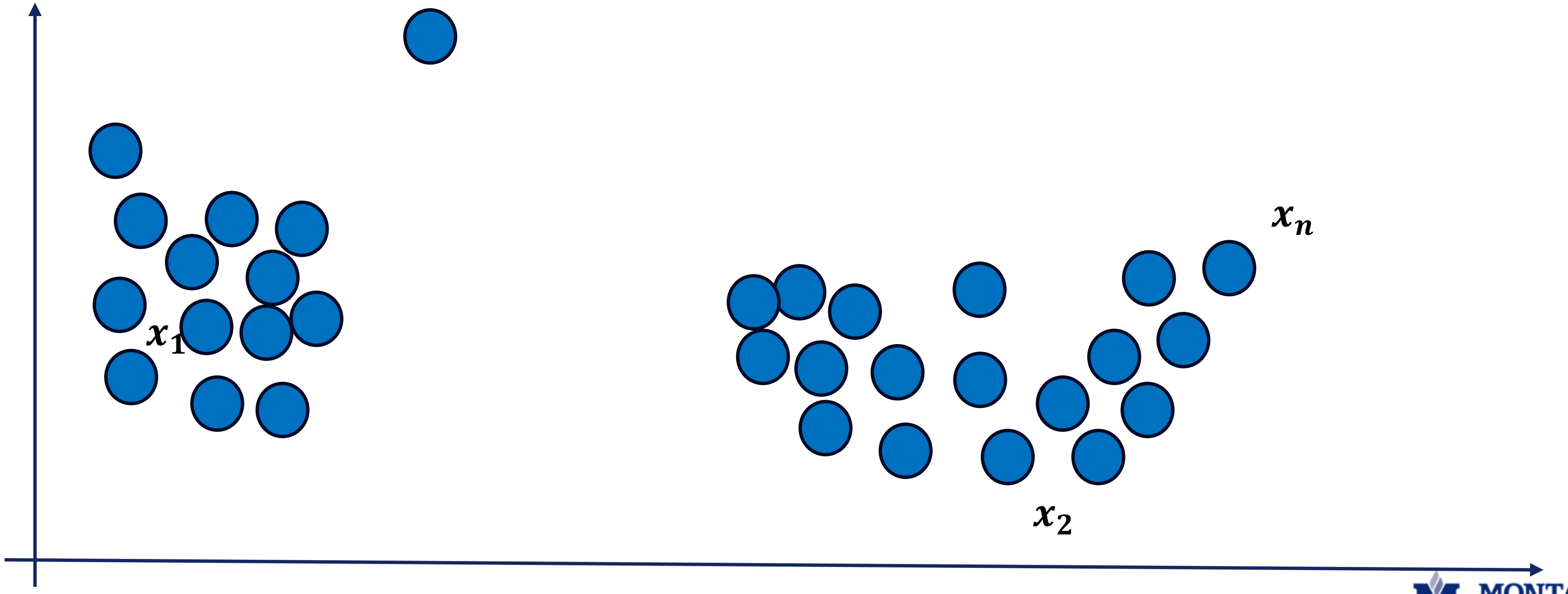


MONTANA STATE UNIVERSITY

Clustering technique

Foundations
- Representative-based method
- Density-based methods
- Hierarchical methods
- Spectral methods
- Graph-based methods

Advanced topics and applications
- Subspace clustering
- Core sets
- Deep learning
- Document clustering
- Clustering for outlier detection



Clustering

- Clustering is broadly and vaguely defined as finding groups of similar entities in a dataset.
- In this class we will learn several clustering techniques and how to validate clustering that we do.
- K-means is a representative-based algorithm that finds a specified number of k of clustering.