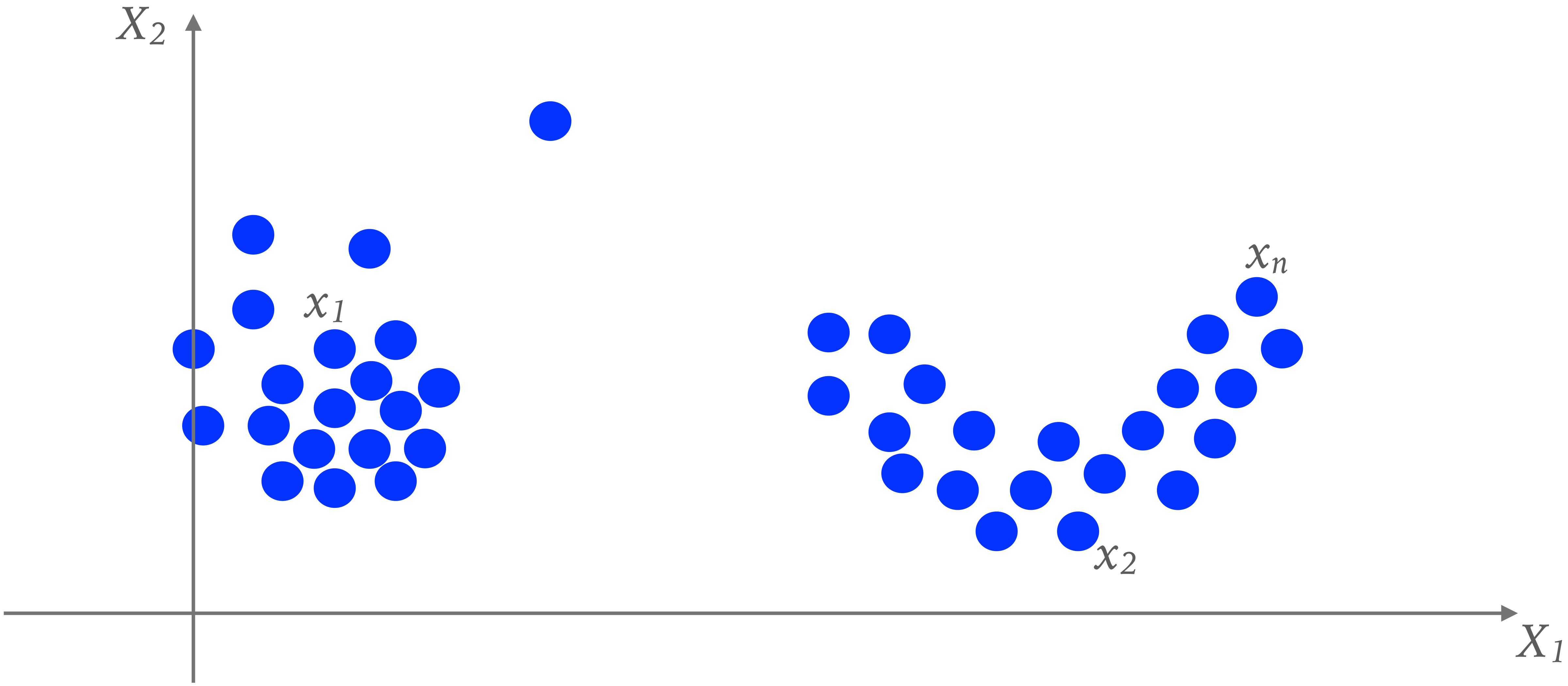
CSCI 347: Data Mining

Clustering Introduction

WHAT ARE CLUSTERS IN A DATA SET?

Data Matrix

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



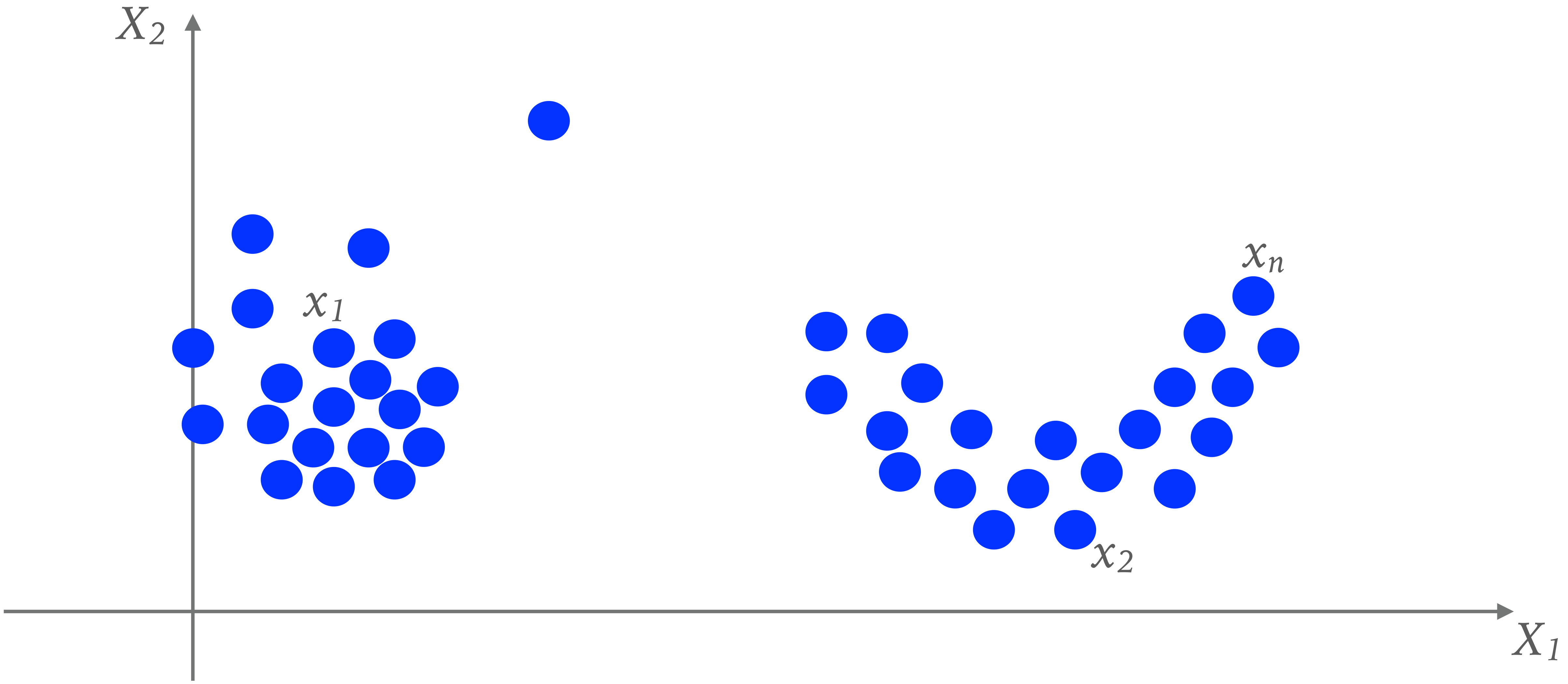
WHAT ARE SOME APPLICATIONS OF CLUSTERING?

Data Matrix →

Movies

People

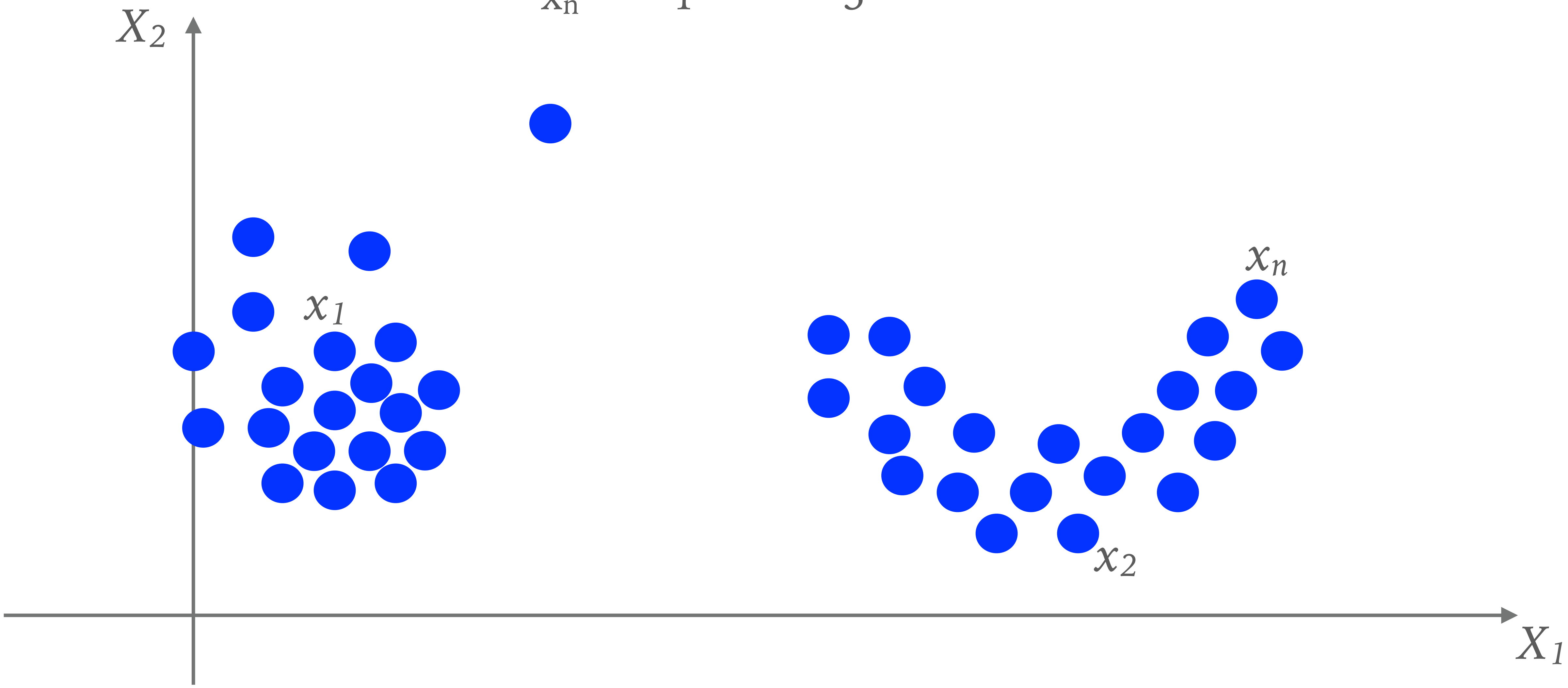
	X_1	X_2
x_1	2	3
x_2	5	4
⋮		
x_n	1	3



WHAT ARE SOME APPLICATIONS OF CLUSTERING?

Data Matrix →

		Genes	
		X_1	X_2
Samples	x_1	2	3
	x_2	5	4
	⋮		
	x_n	1	3



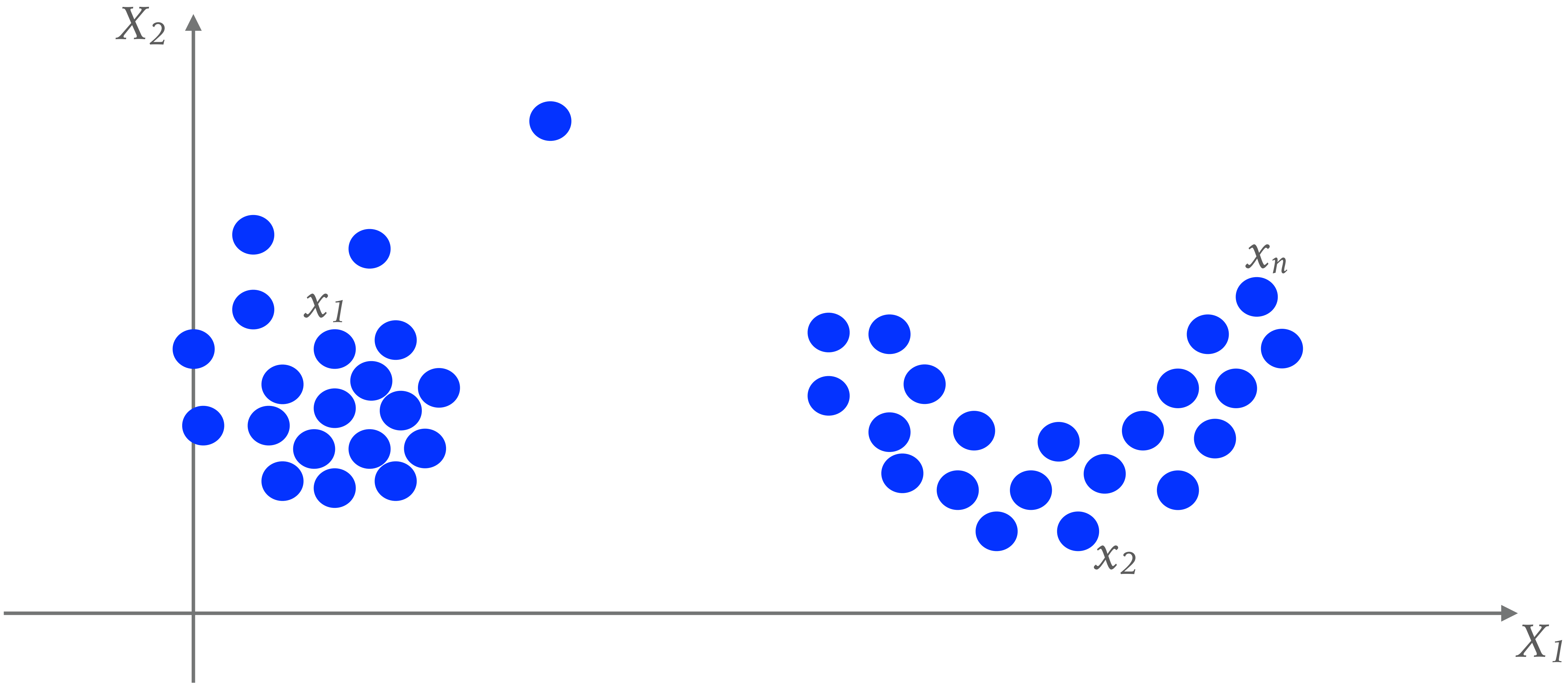
WHAT ARE SOME APPLICATIONS OF CLUSTERING?

Data Matrix →

Words

Documents

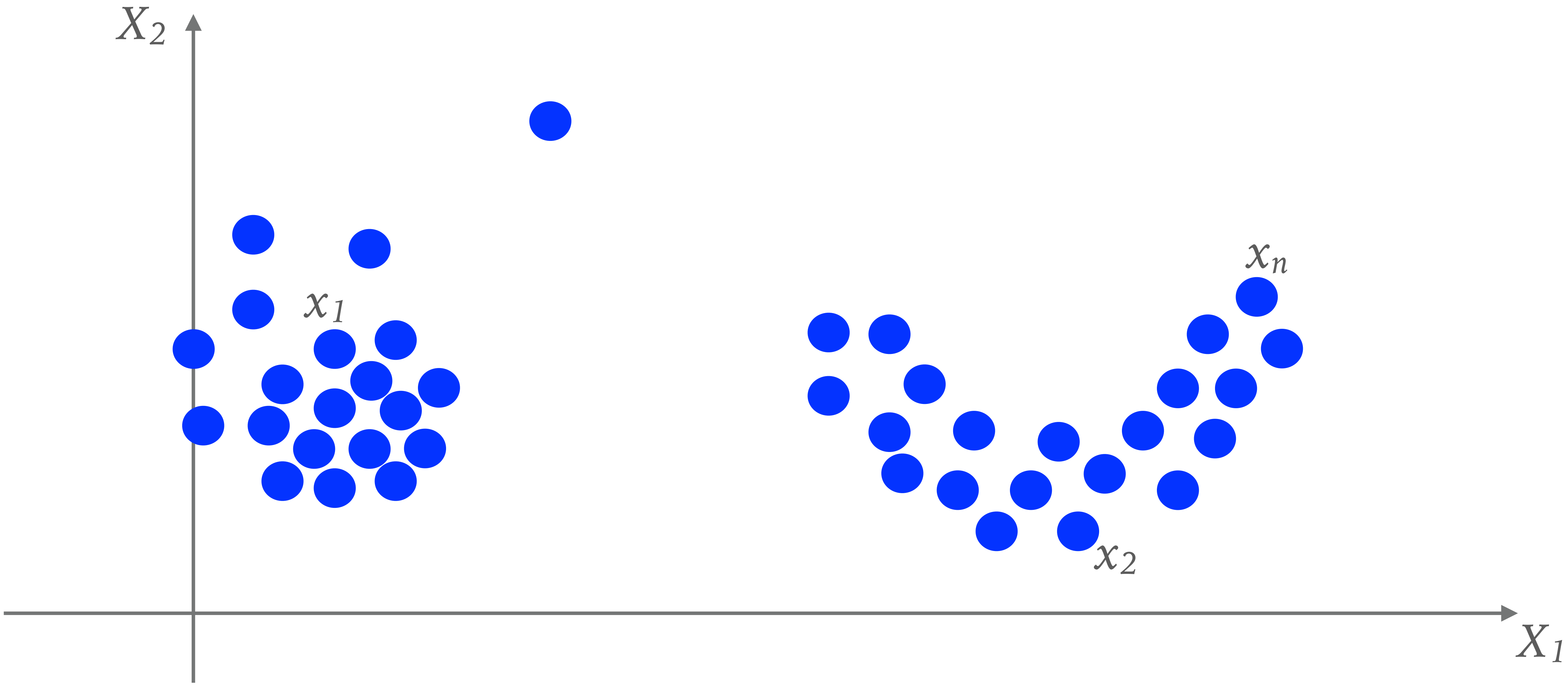
	X_1	X_2
x_1	10	3
x_2	0	42
⋮		
x_n	1	6



HOW DO WE FIND CLUSTERS IN A DATA SET?

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

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

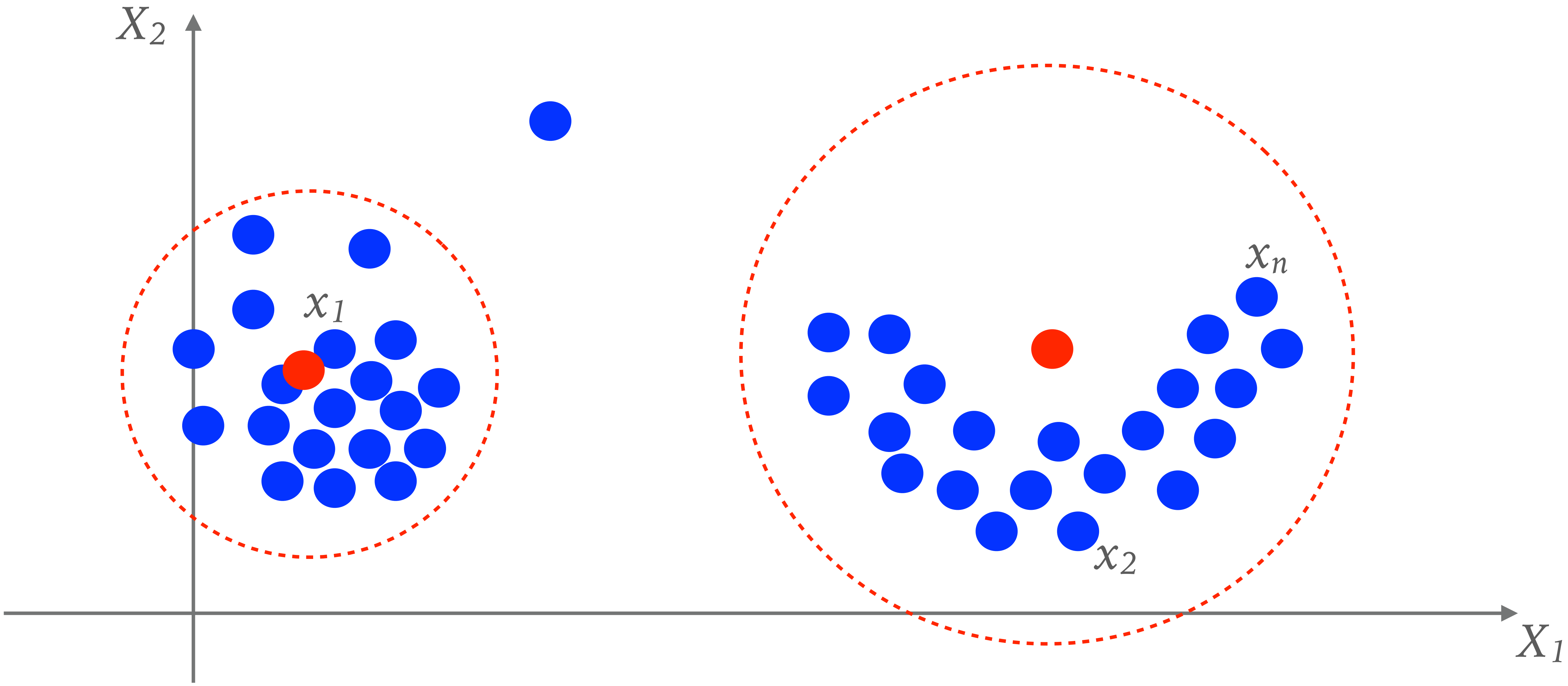


HOW DO WE FIND CLUSTERS IN A DATA SET?

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
⋮		
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

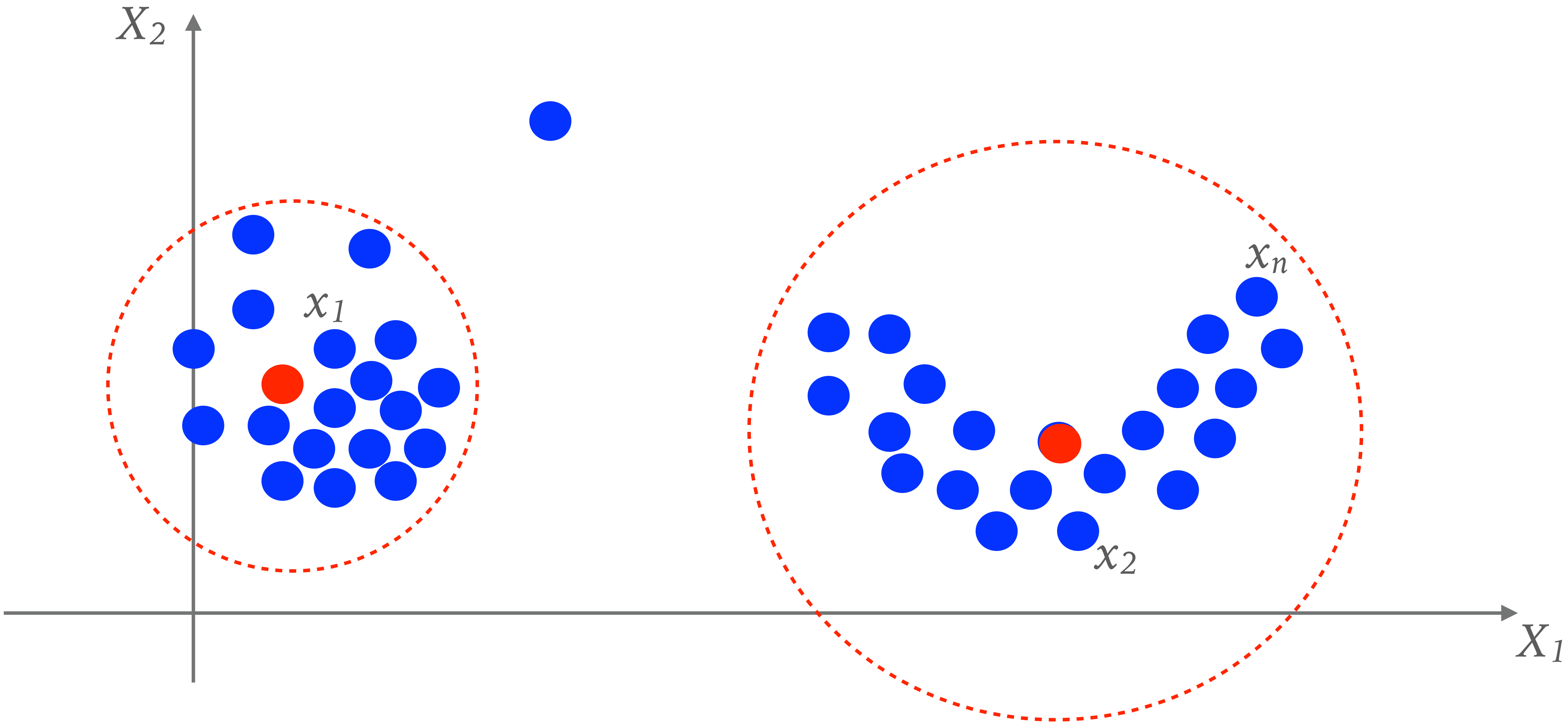


HOW DO WE FIND CLUSTERS IN A DATA SET?

	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
⋮		
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

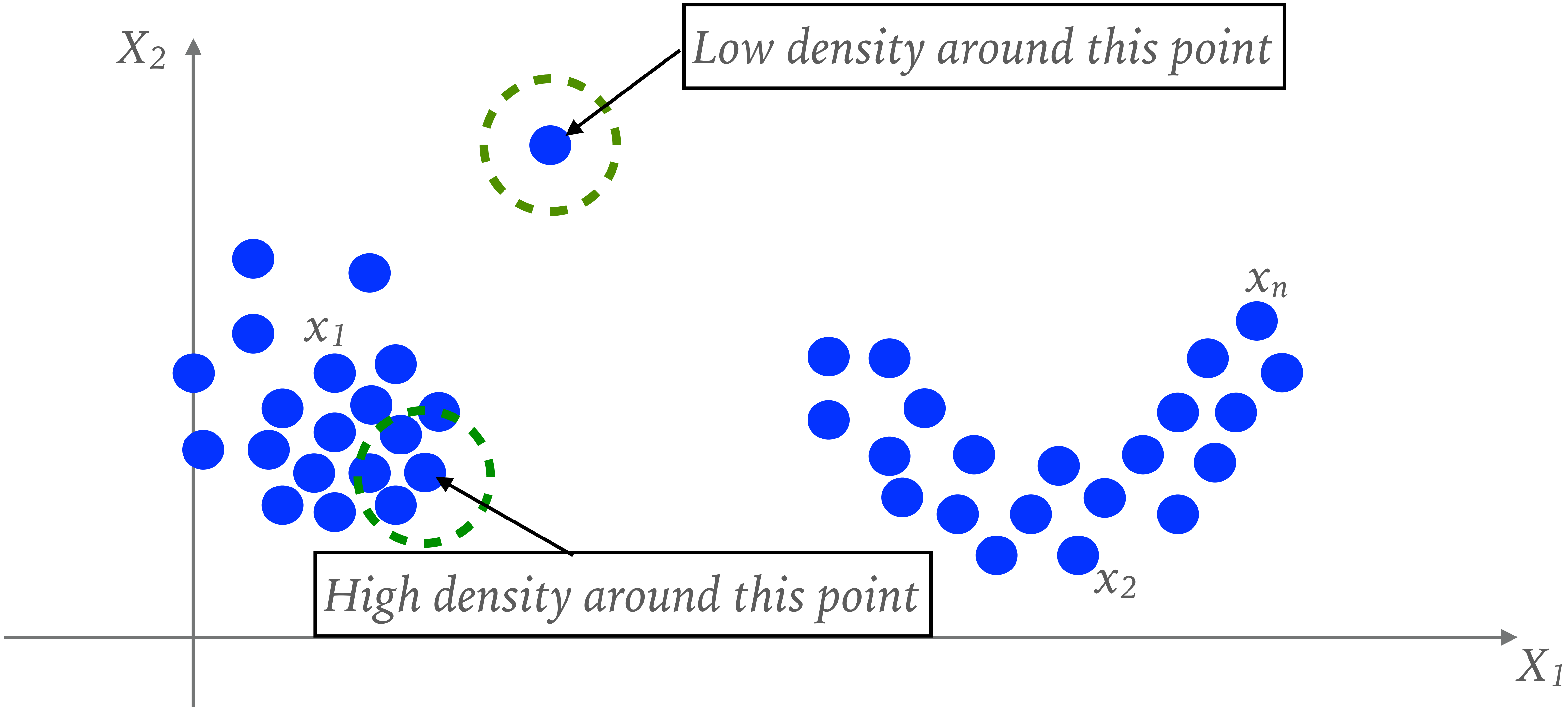


HOW DO WE FIND CLUSTERS IN A DATA SET?

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

Density-based methods:

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



Low density around this point

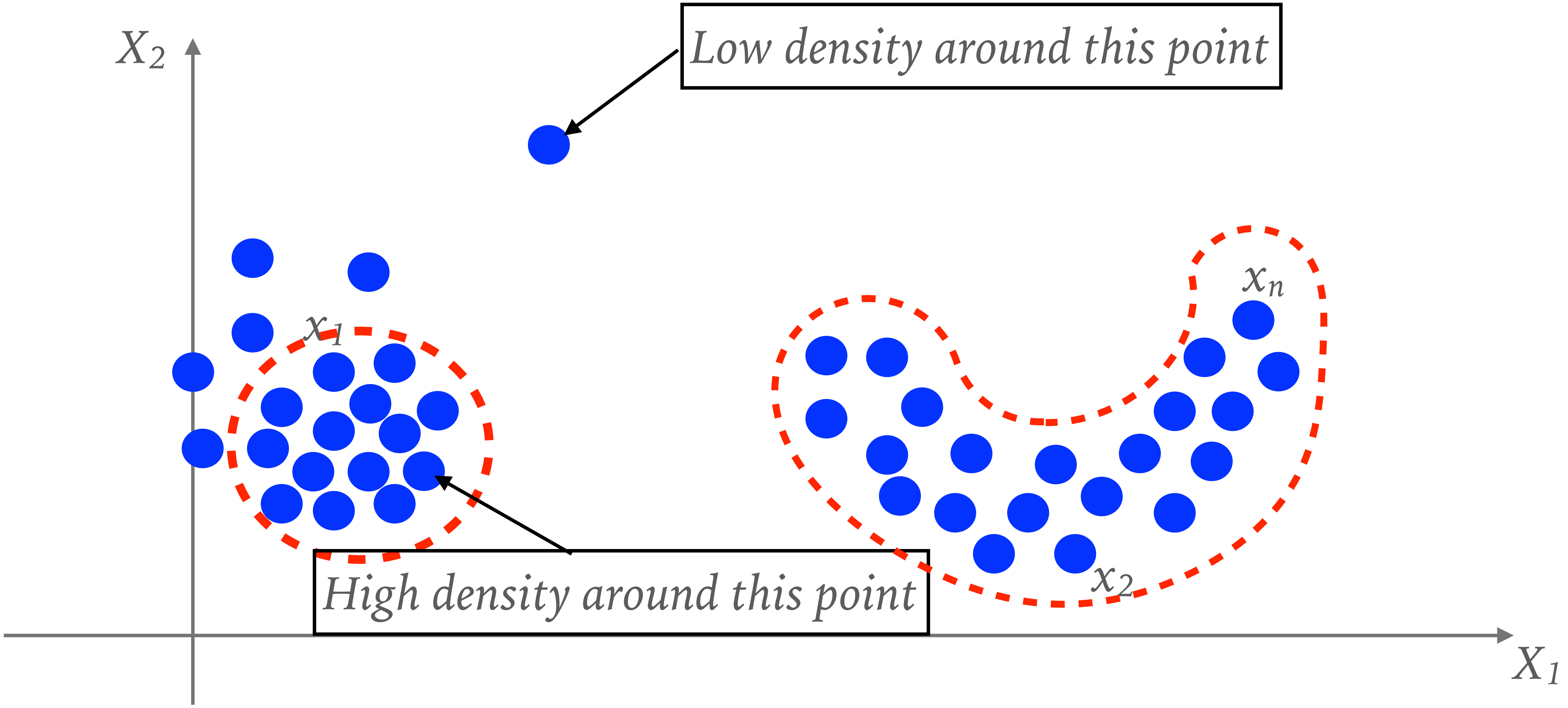
High density around this point

HOW DO WE FIND CLUSTERS IN A DATA SET?

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

Density-based methods:

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



Low density around this point

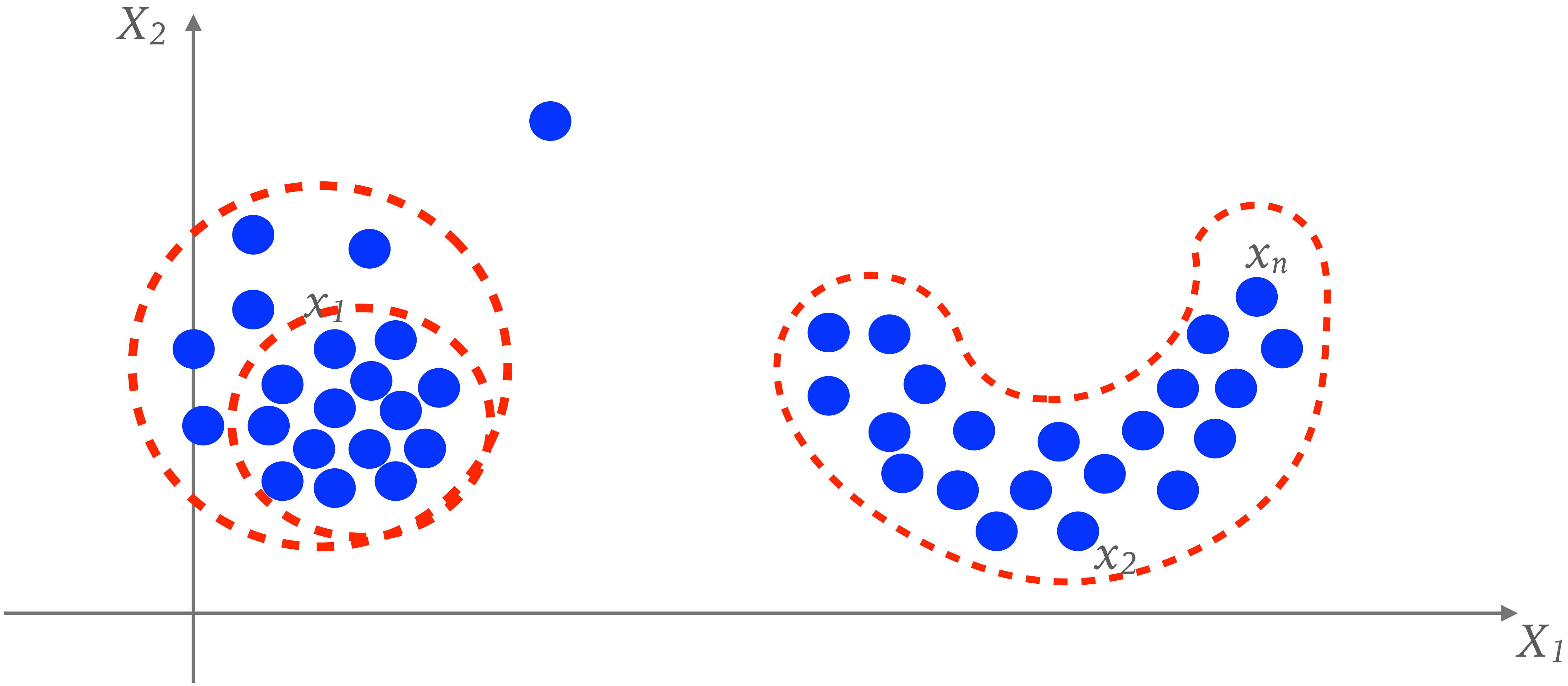
High density around this point

HOW DO WE FIND CLUSTERS IN A DATA SET?

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

Hierarchical methods:

Clusters within clusters



HOW DO WE FIND CLUSTERS IN A DATA SET?

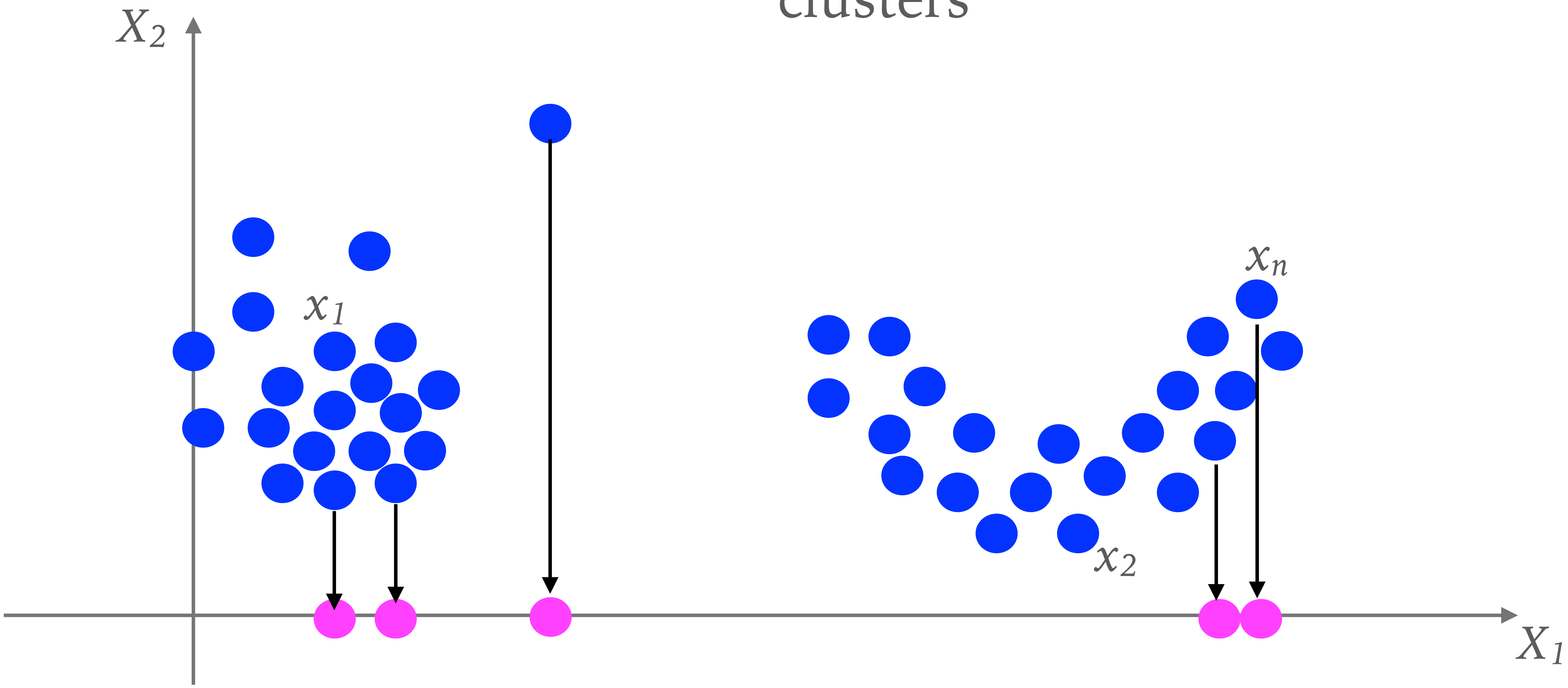
	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
⋮		
x_n	16.4	4.5

→

	X_1
x_1	2
x_2	13.1
⋮	
x_n	16.4

Spectral and subspace methods:

Find a lower dimensional space that better represents the clusters



HOW DO WE FIND CLUSTERS IN A DATA SET?

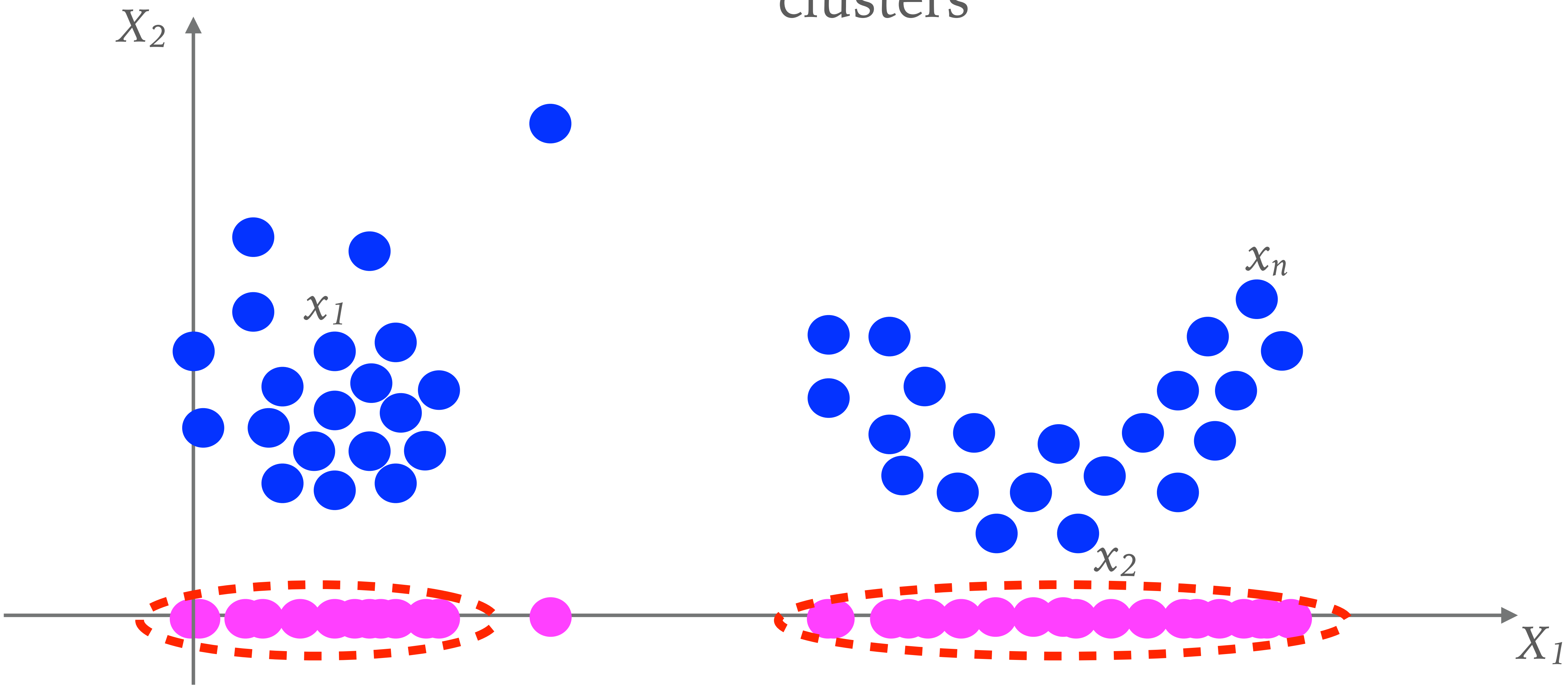
	X_1	X_2
x_1	2	3.2
x_2	13.1	0.9
⋮		
x_n	16.4	4.5

→

	X_1
x_1	2
x_2	13.1
⋮	
x_n	16.4

Spectral and subspace methods:

Find a lower dimensional space that better represents the clusters



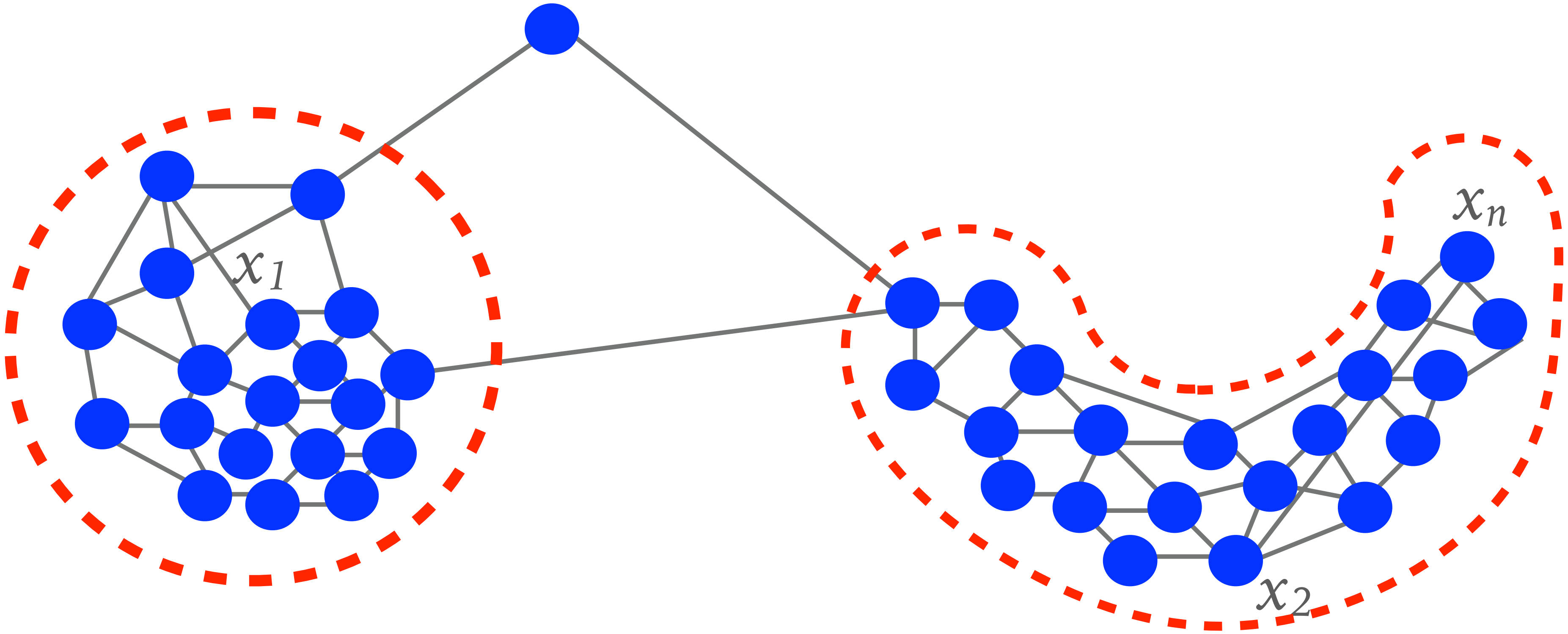
HOW DO WE FIND CLUSTERS IN A DATA SET?

Adjacency matrix →

	x_1	x_2	…	x_n
x_1	0	0	…	0
x_2	0	0	…	1
⋮			⋱	
x_n	0	1	…	0

Graph-based methods:

Find subgraphs with high edge connectivity

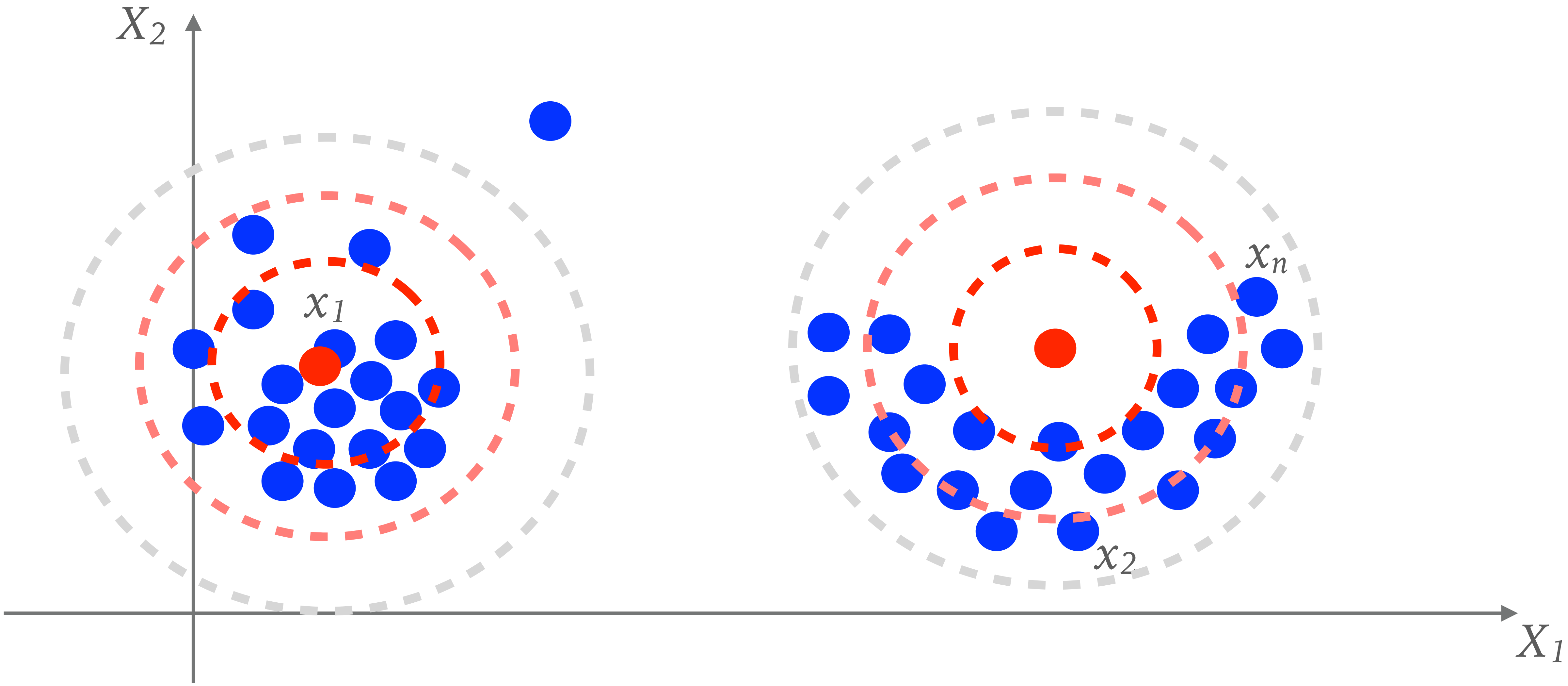


HOW DO WE FIND CLUSTERS IN A DATA SET?

	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



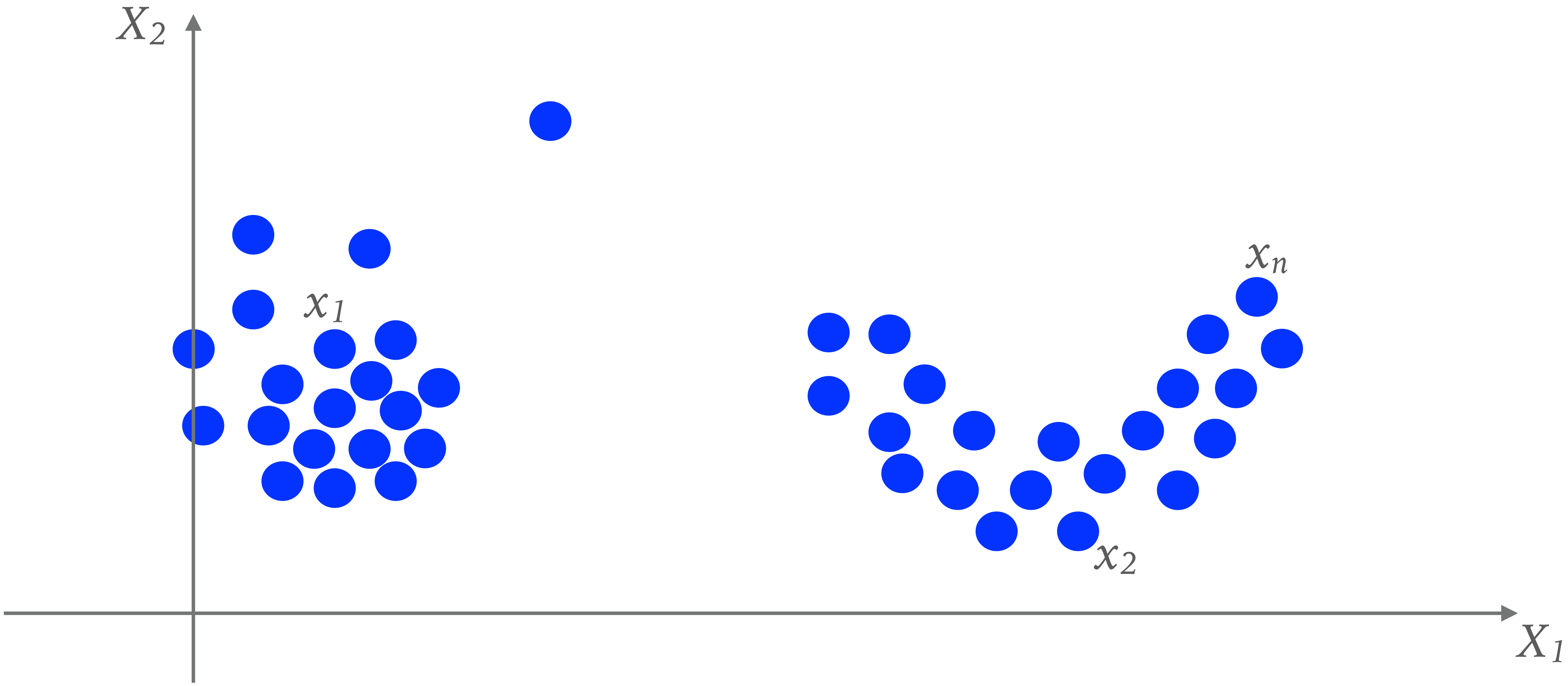
CLUSTERING TECHNIQUES

Foundations

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

Advanced topics and applications

- Parallel algorithms
- Subspace clustering
- Core sets
- Deep learning
- Document clustering
- Clustering for outlier detection


X_2
x_1
x_n
x_2
X_1