

Machine Learning Crash Course Part II: Clustering

Tamara Broderick
UC Berkeley
August 21, 2012



Outline

Outline

0. What is clustering?

Outline

- 0. What is clustering?
- 1. K means algorithm

Outline

- 0. What is clustering?
- 1. K means algorithm
- 2. Clustering evaluation

Outline

0. What is clustering?
1. K means algorithm
2. Clustering evaluation
3. Clustering trouble-shooting

Outline

0. What is clustering?
1. K means algorithm
2. Clustering evaluation
3. Clustering trouble-shooting
4. Example

Outline

0. What is clustering?

1. K means algorithm
2. Clustering evaluation
3. Clustering trouble-shooting
4. Example

Clustering

Clustering

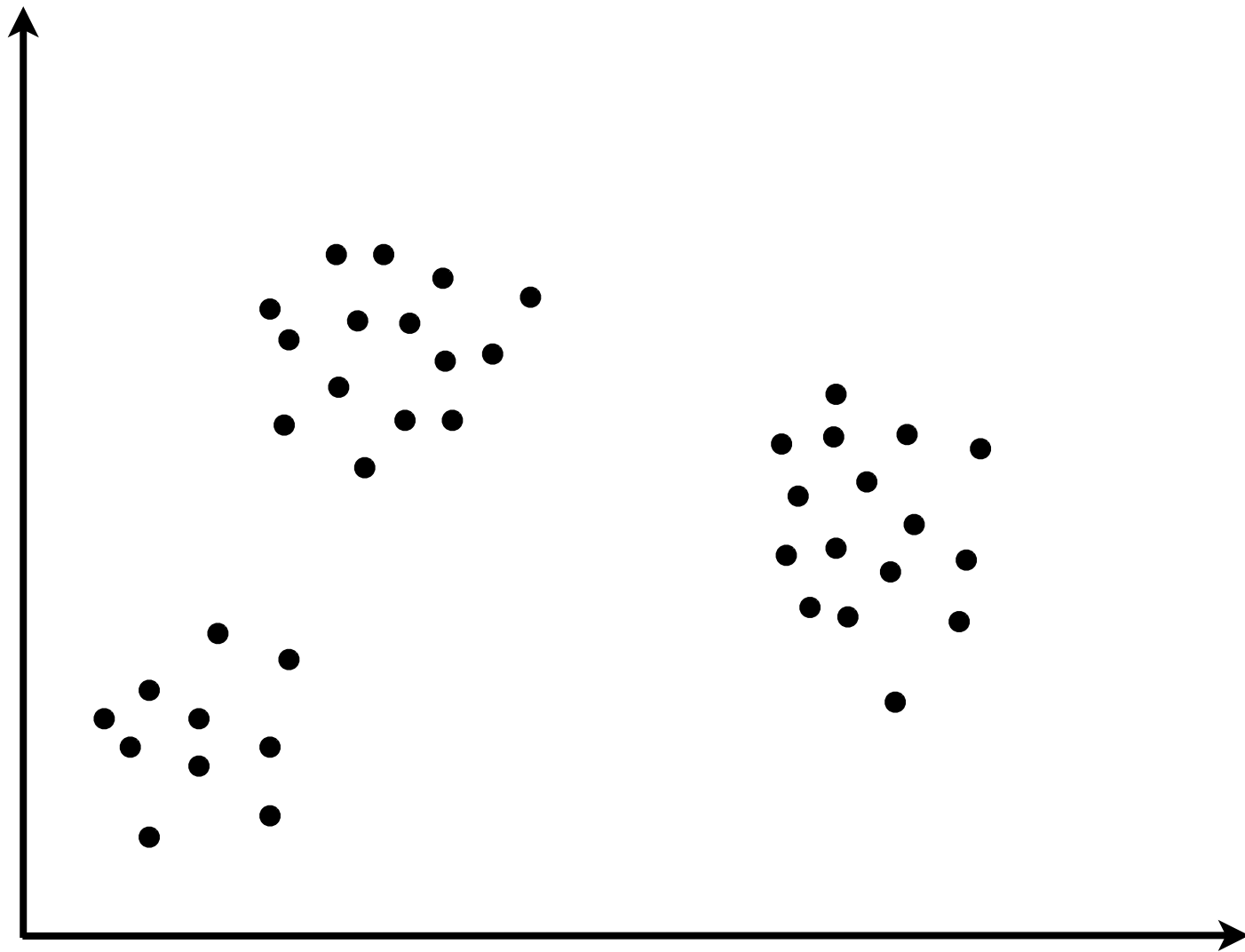
Grouping data according to similarity.

Clustering

Grouping data according to similarity.

Clustering

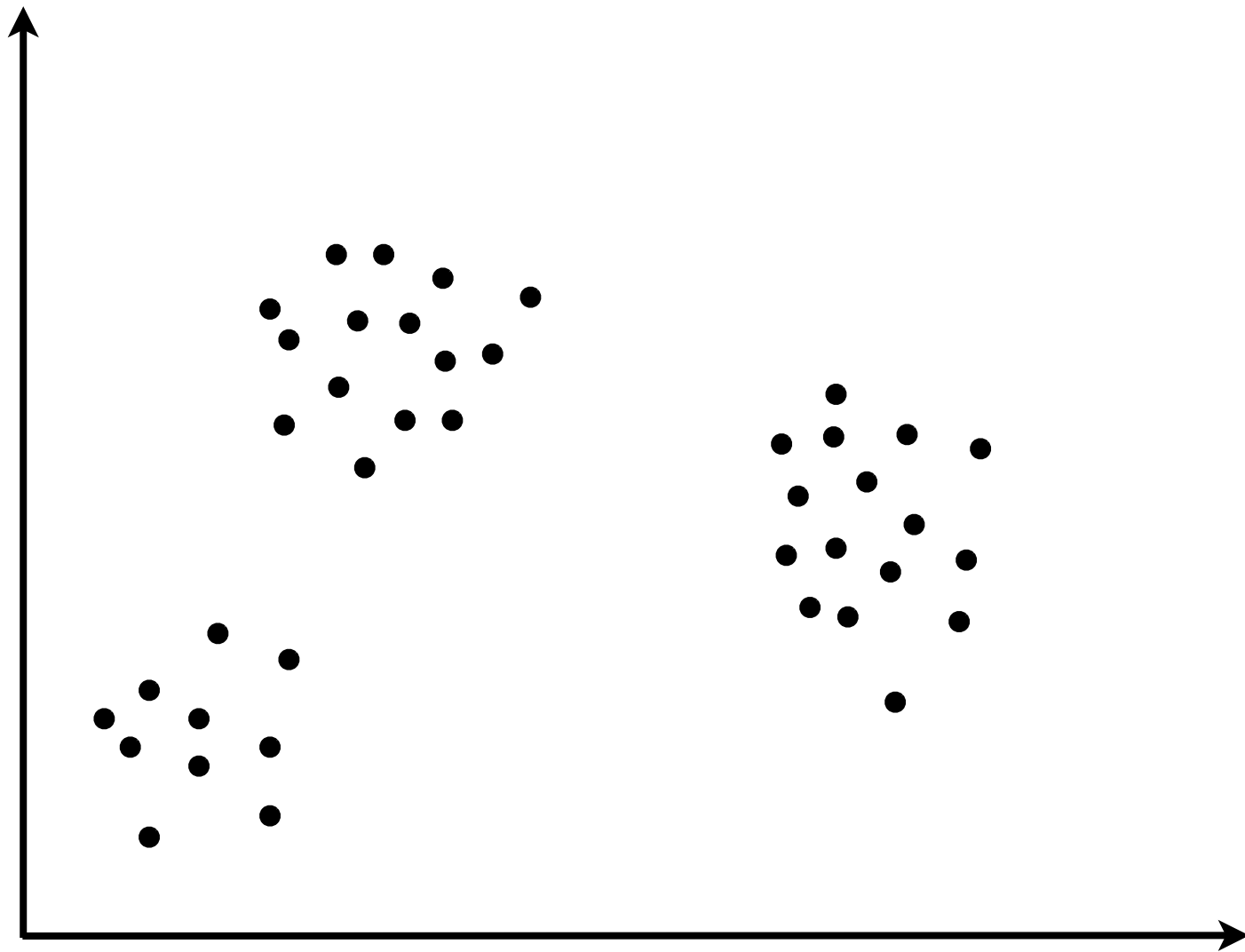
Grouping data according to similarity.



Clustering

Grouping data according to similarity.

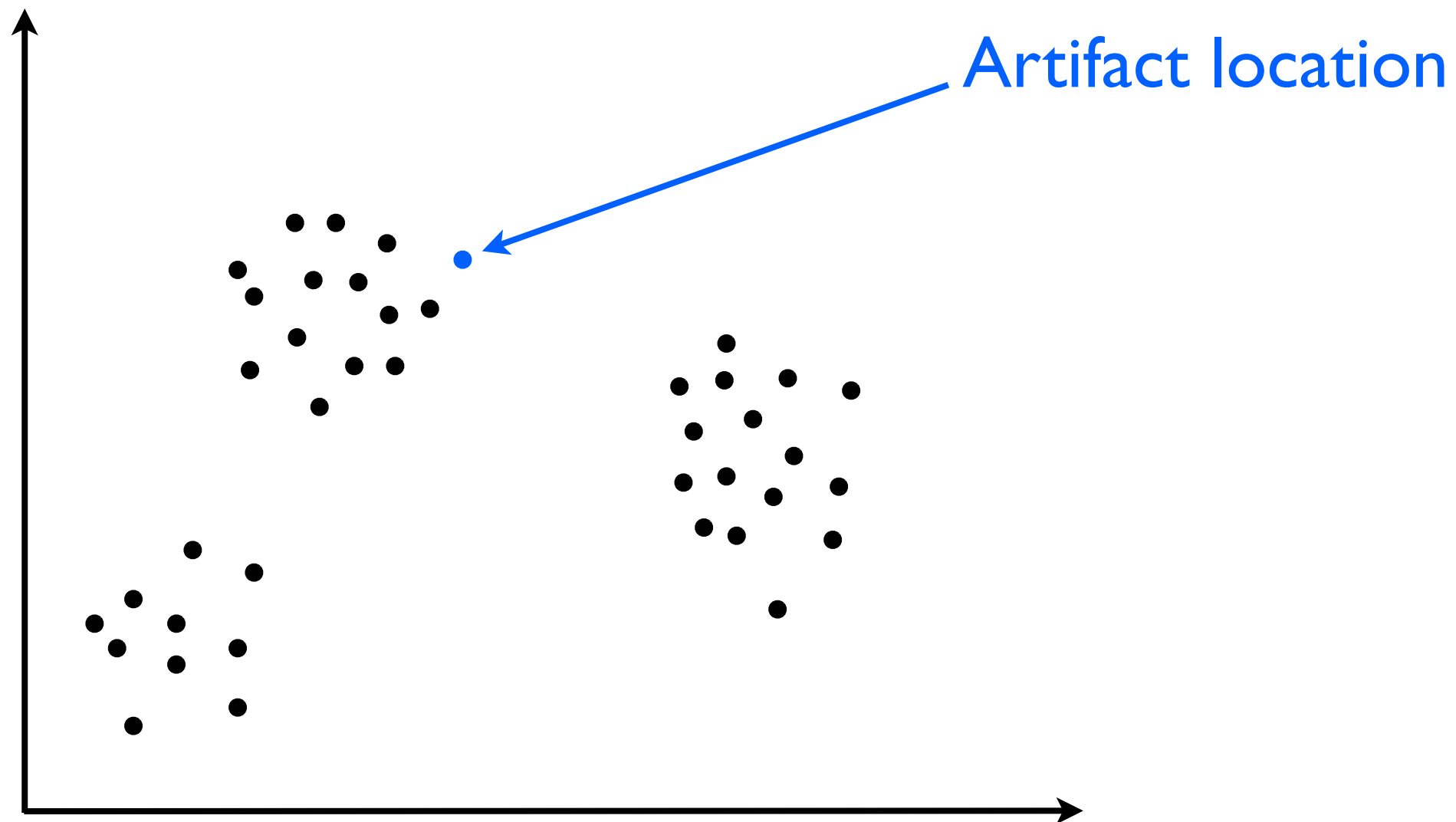
E.g. archaeological dig



Clustering

Grouping data according to similarity.

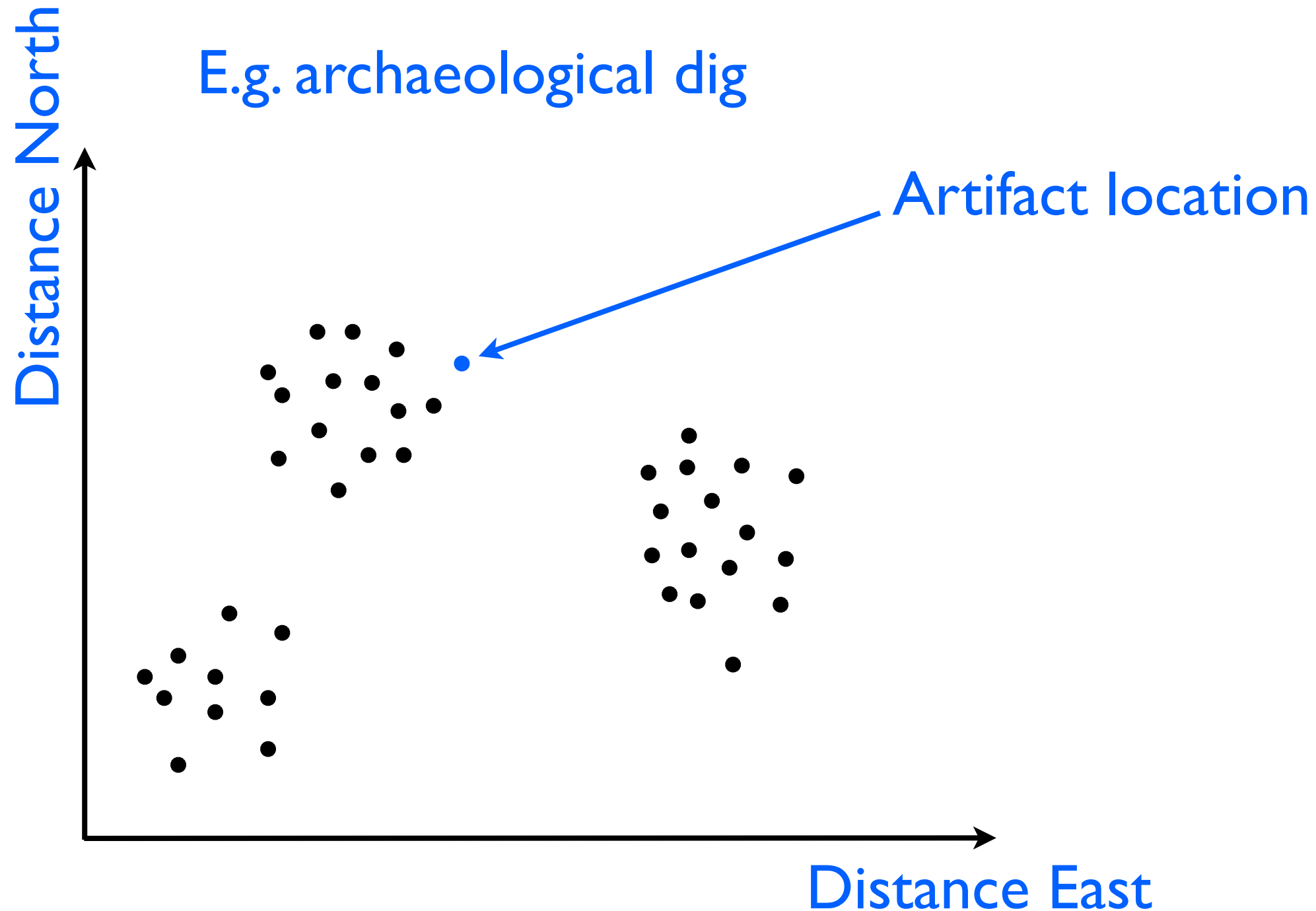
E.g. archaeological dig



Clustering

Grouping data according to similarity.

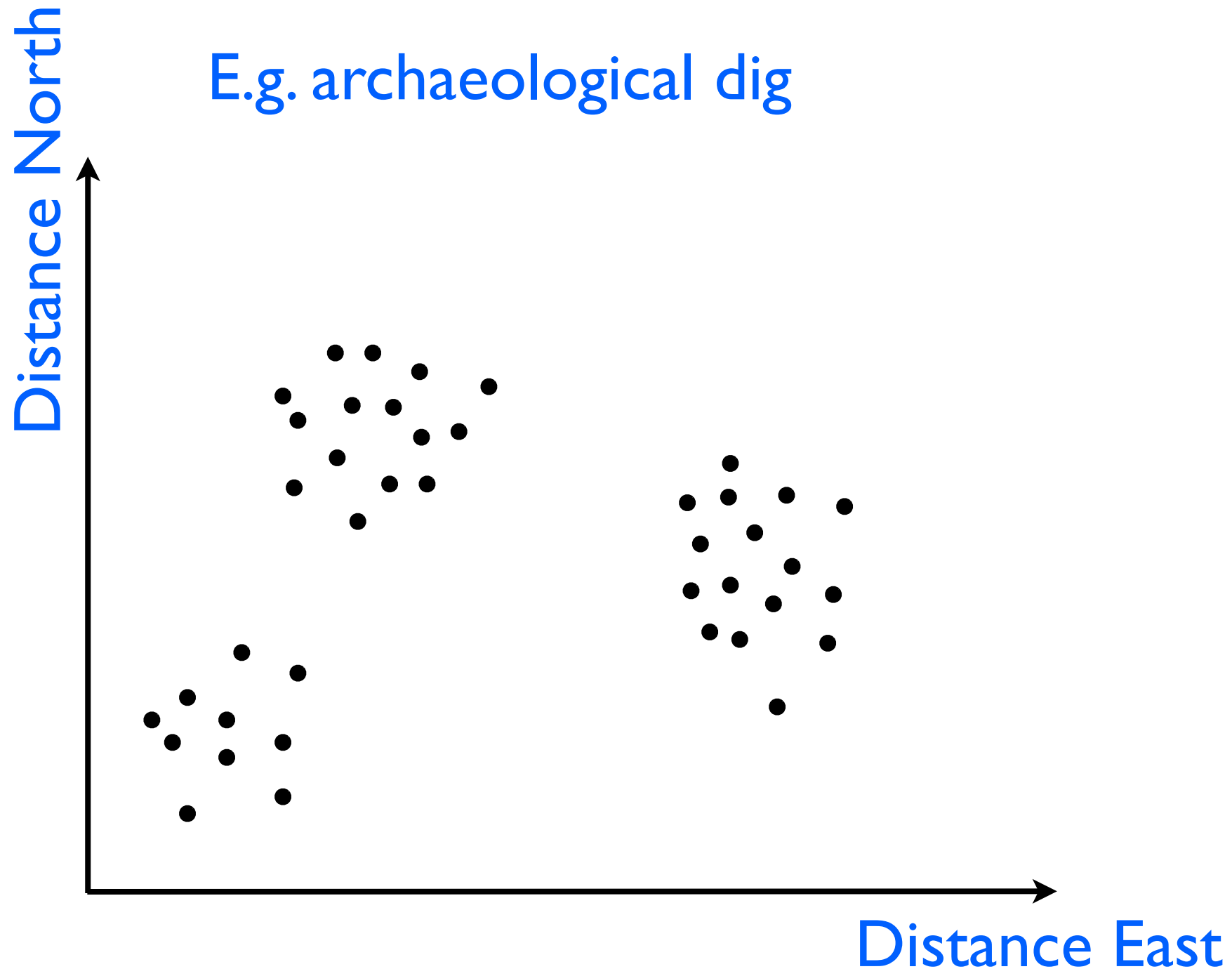
E.g. archaeological dig



Clustering

Grouping data according to similarity.

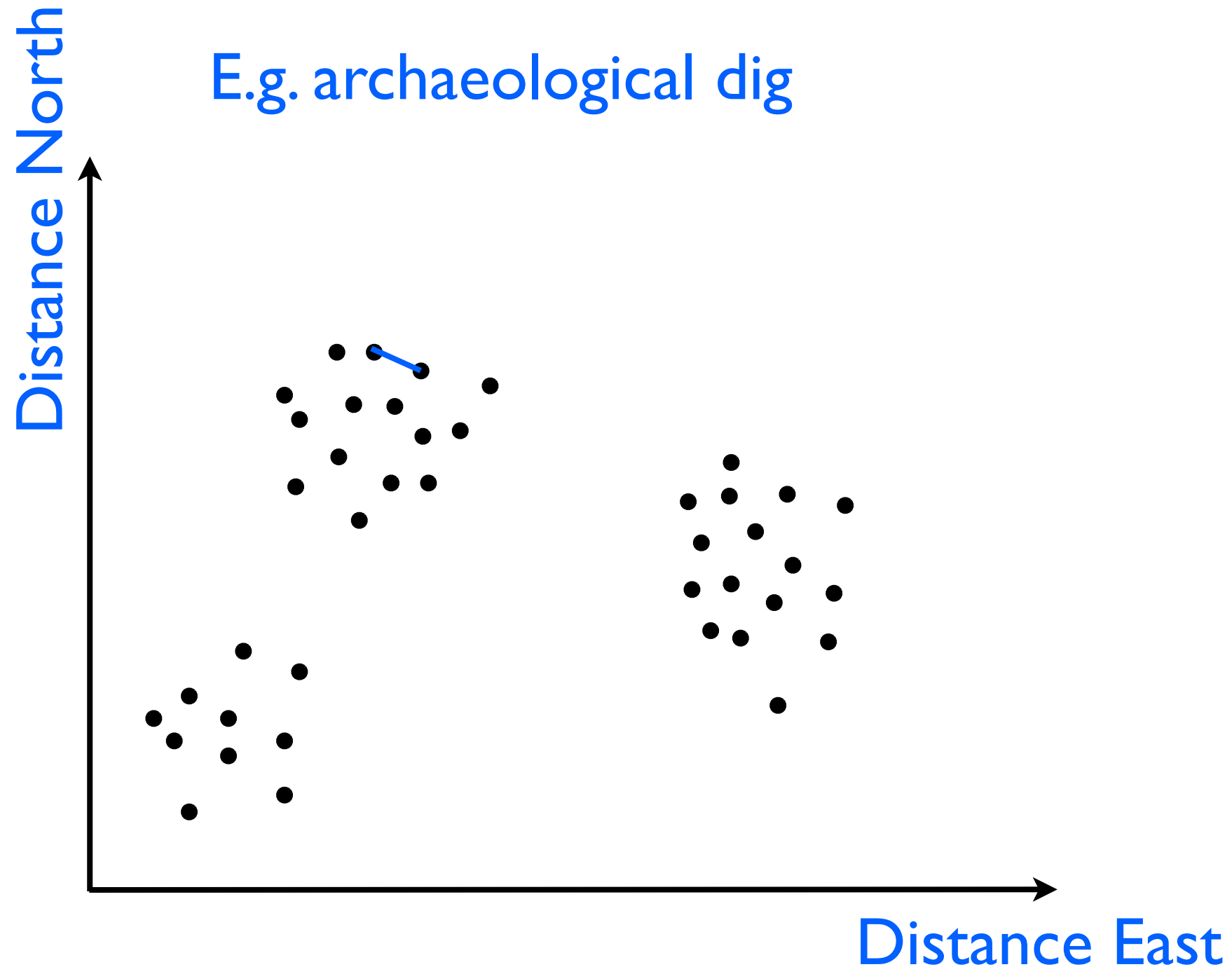
E.g. archaeological dig



Clustering

Grouping data according to similarity.

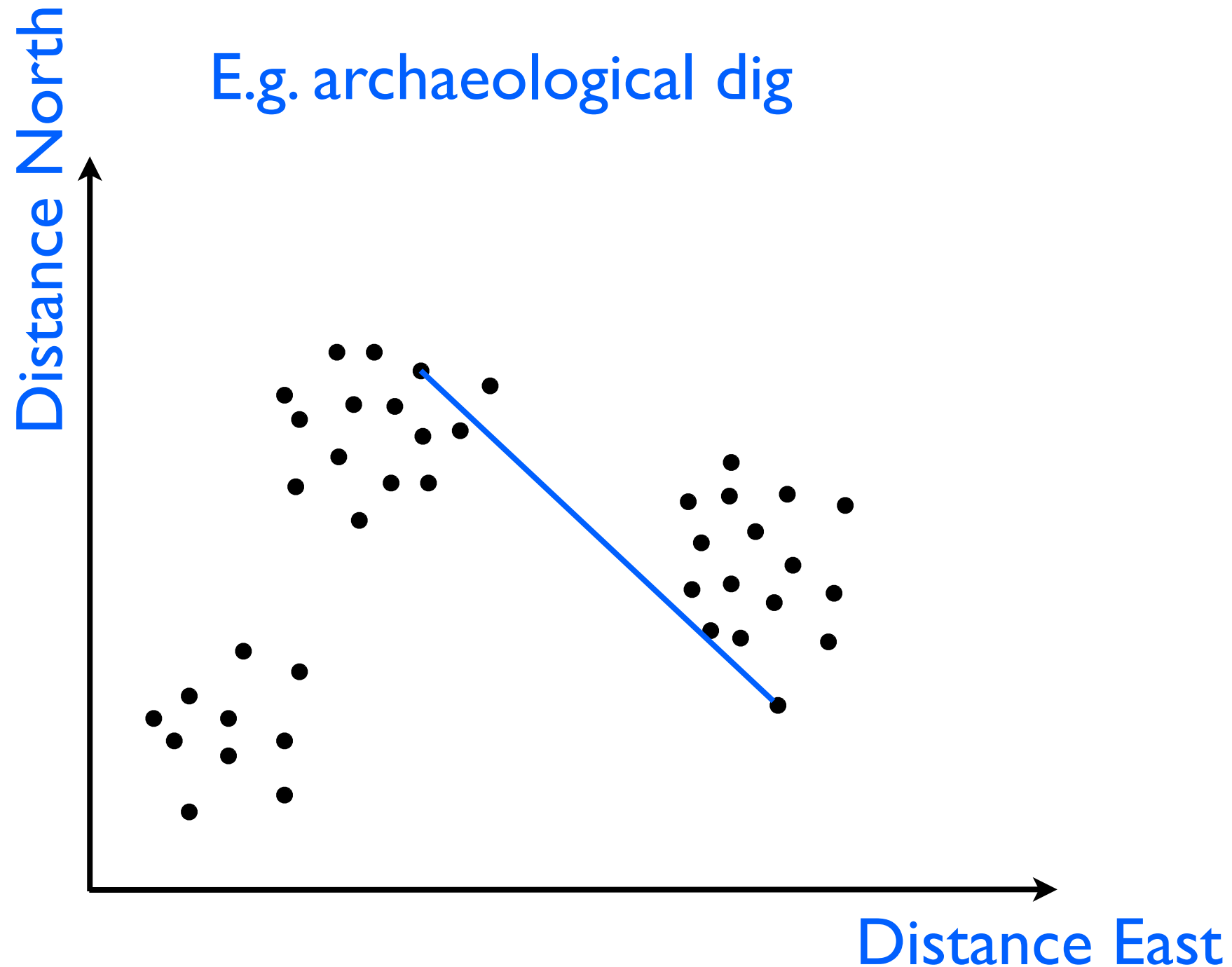
E.g. archaeological dig



Clustering

Grouping data according to similarity.

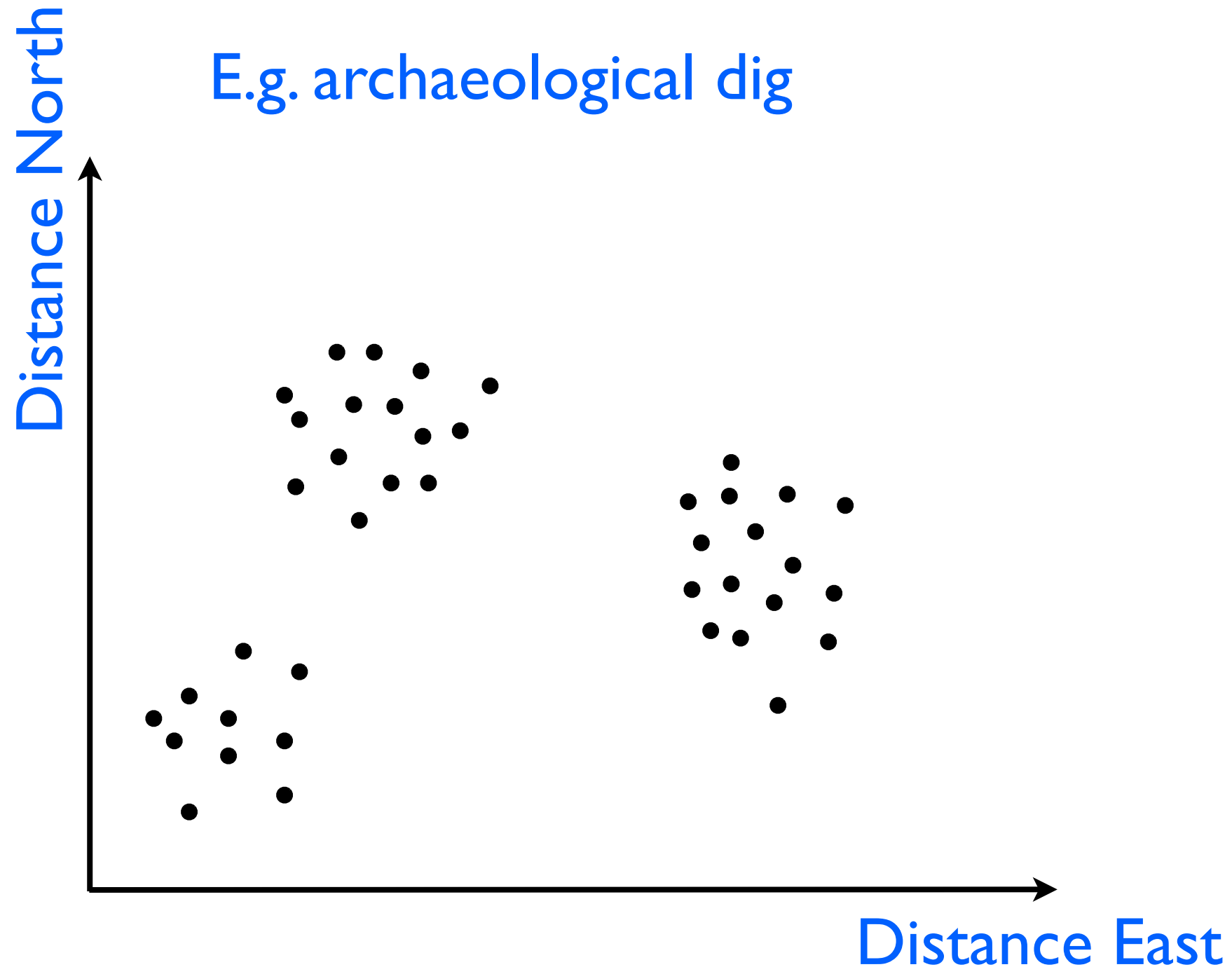
E.g. archaeological dig



Clustering

Grouping data according to similarity.

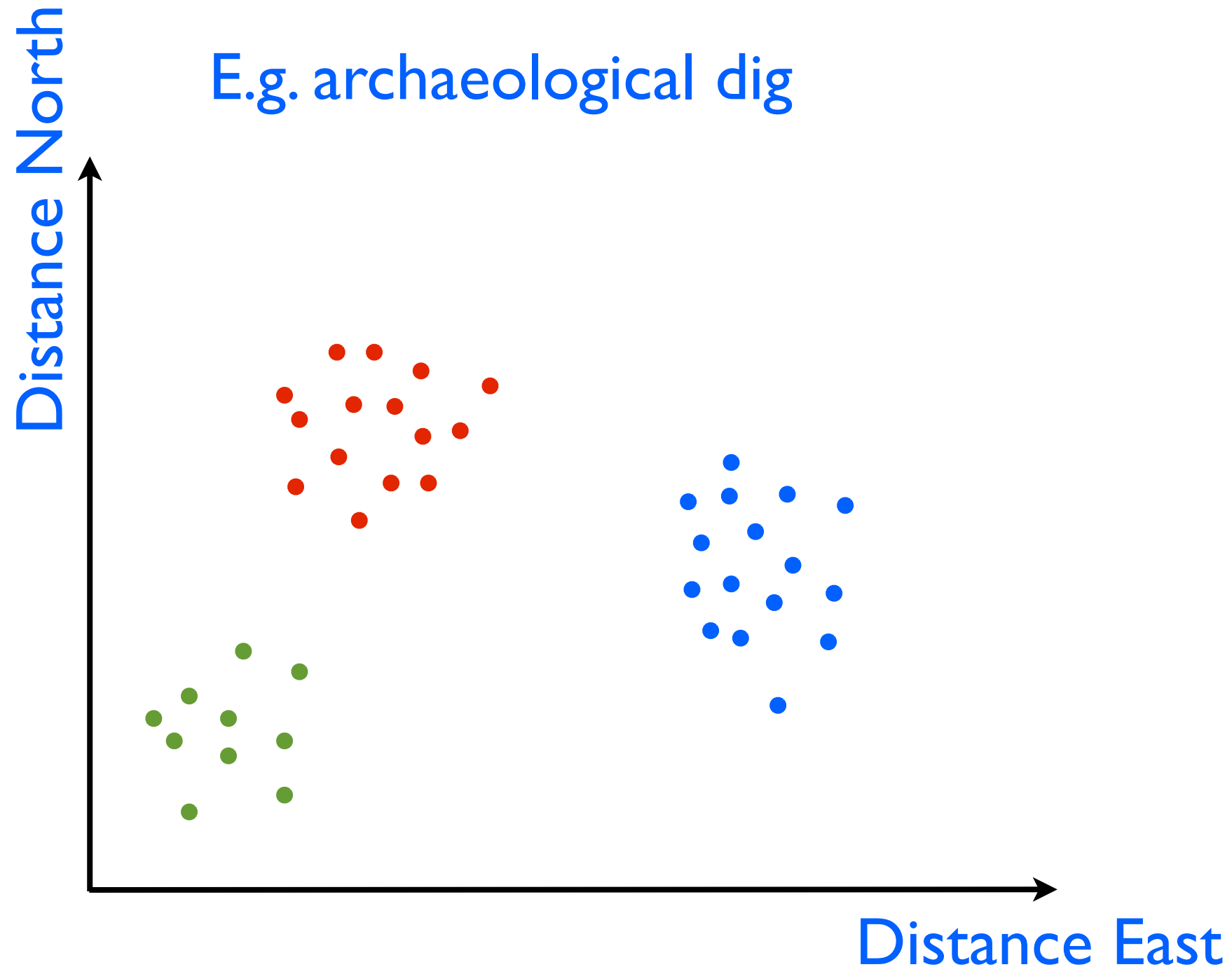
E.g. archaeological dig



Clustering

Grouping data according to similarity.

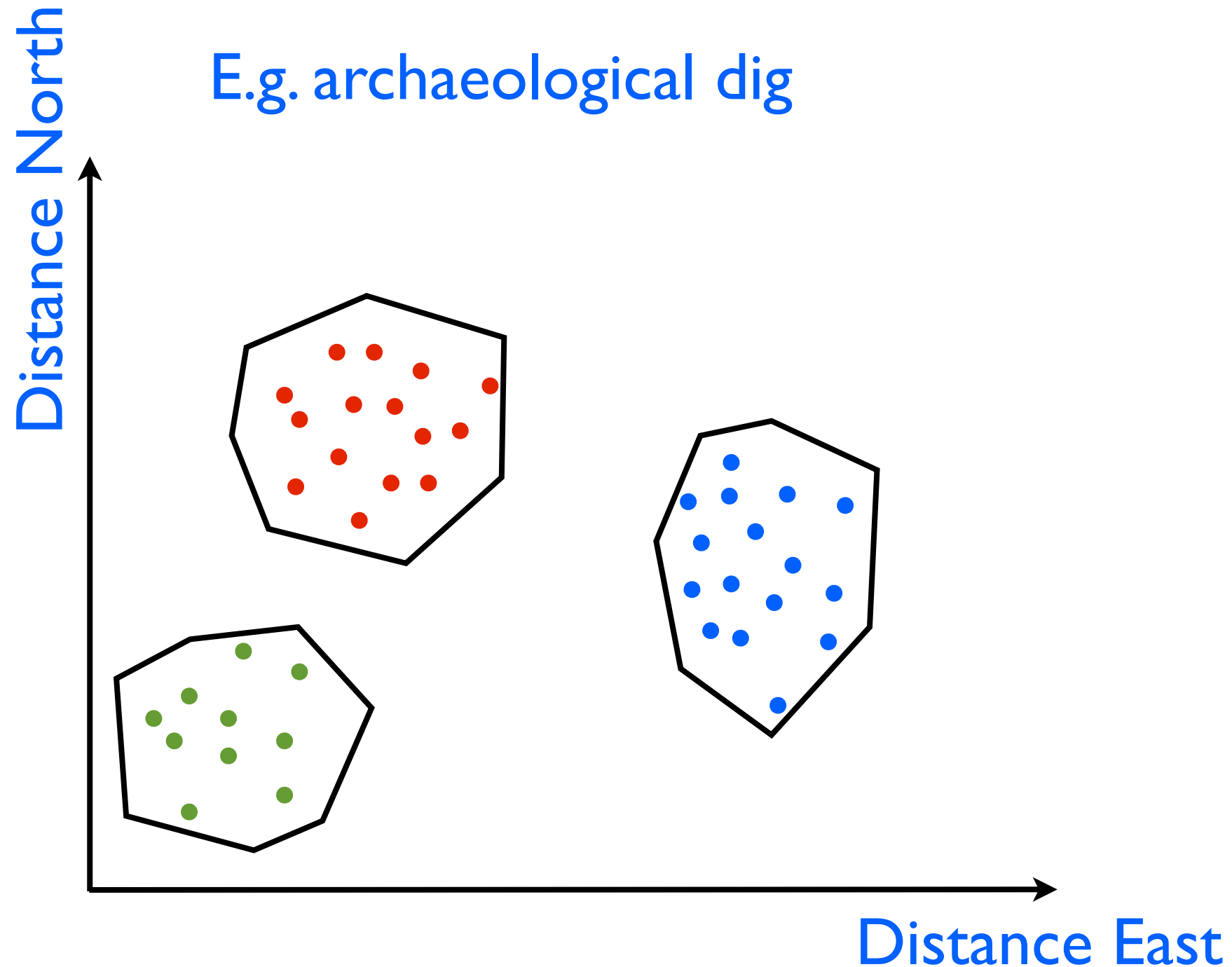
E.g. archaeological dig



Clustering

Grouping data according to similarity.

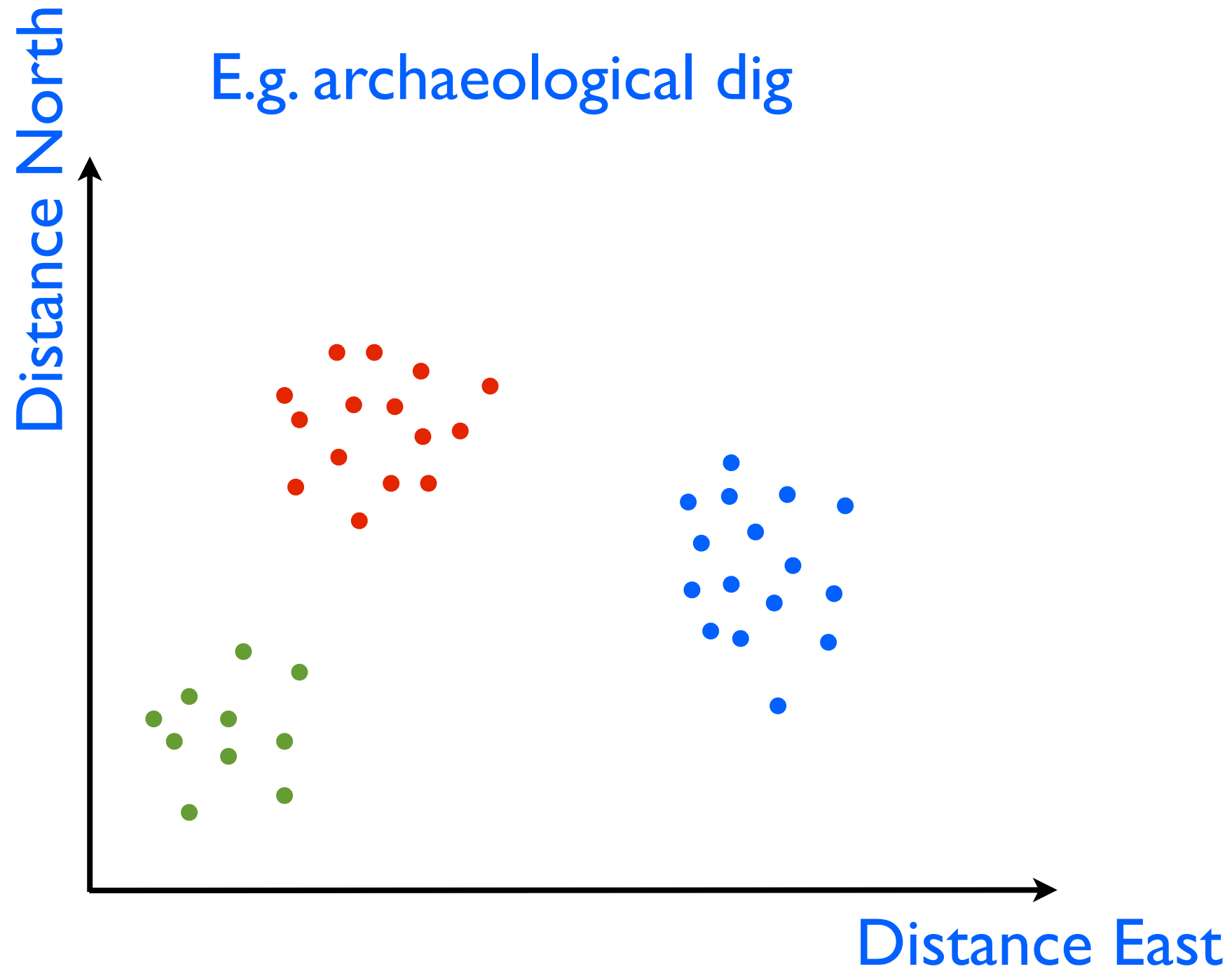
E.g. archaeological dig



Clustering

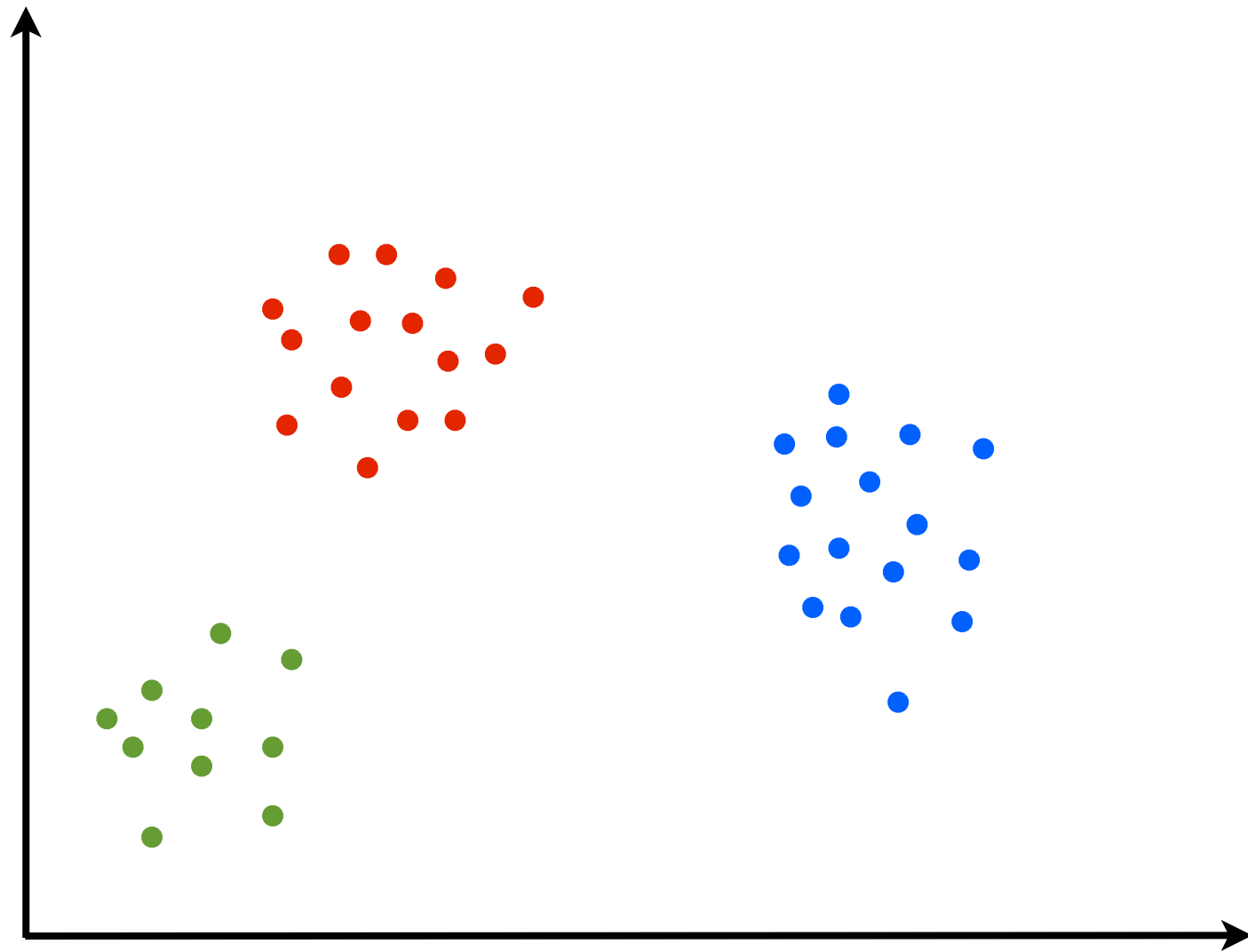
Grouping data according to similarity.

E.g. archaeological dig



Clustering vs. Classification

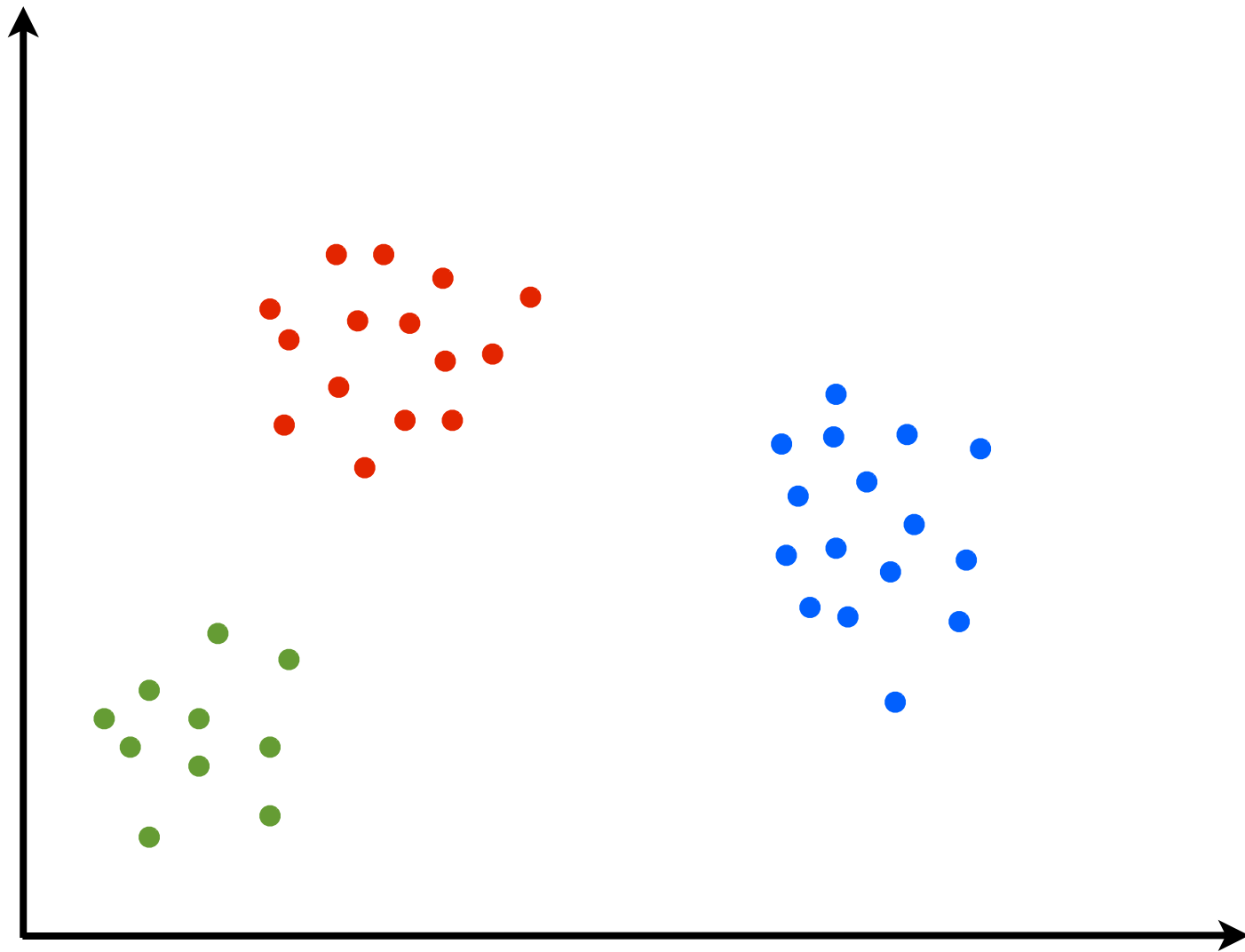
Grouping data according to similarity.



Clustering vs. Classification

Grouping data according to similarity.

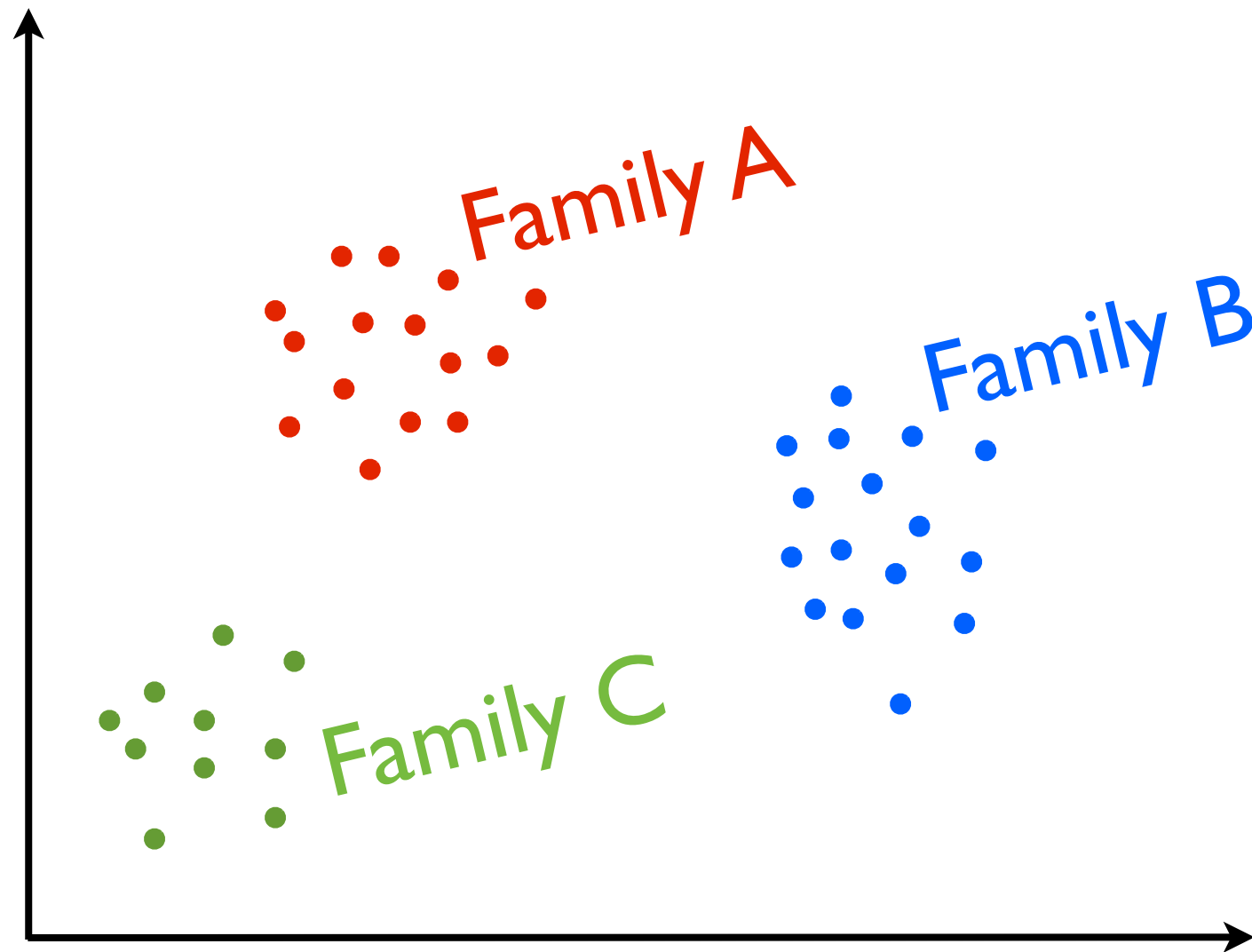
Predicting new labels from old labels.



Clustering vs. Classification

Grouping data according to similarity.

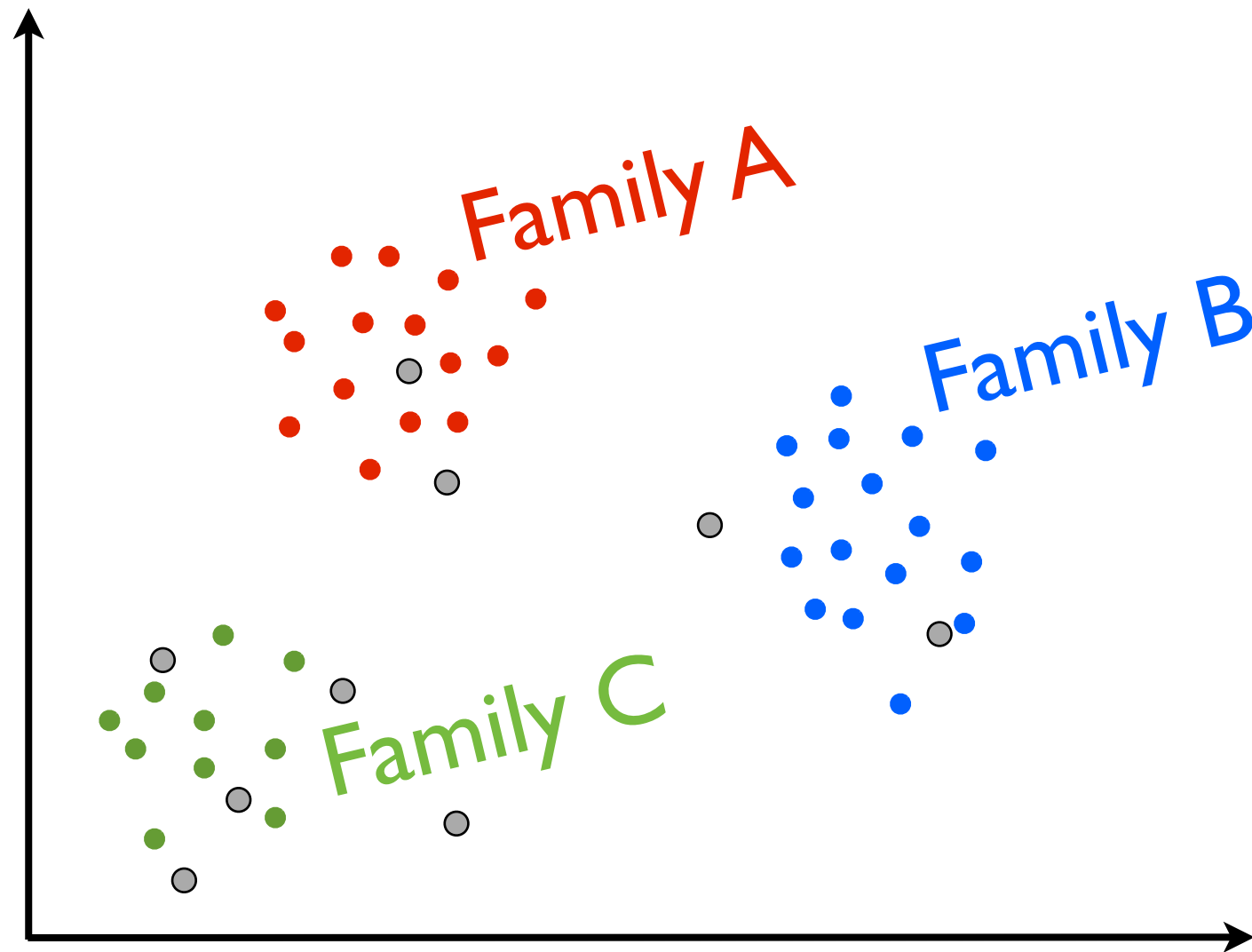
Predicting new labels from old labels.



Clustering vs. Classification

Grouping data according to similarity.

Predicting new labels from old labels.



Why use clustering...

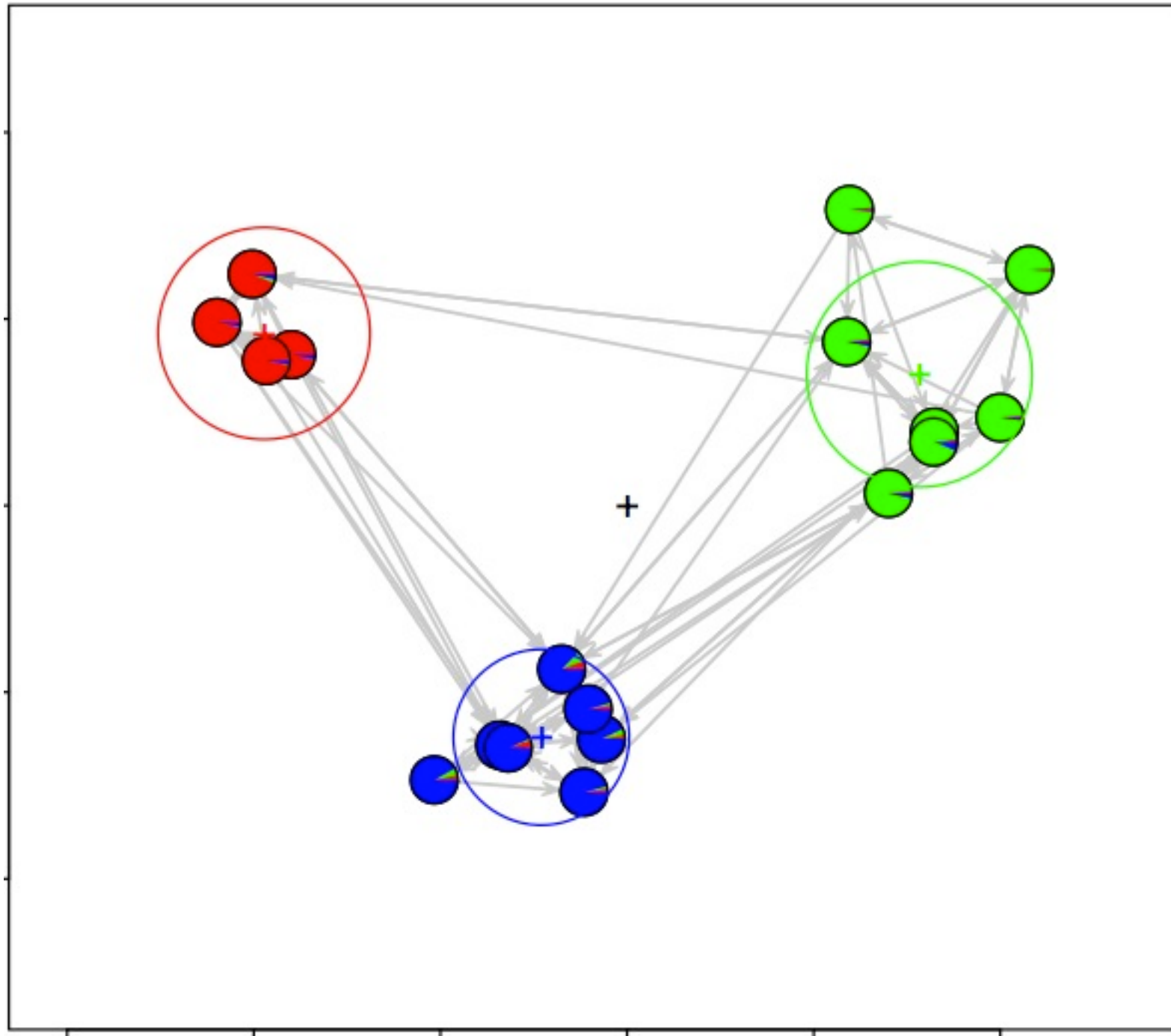
... instead of classification?

- Exploratory data analysis

Why use clustering...

... instead of classification?

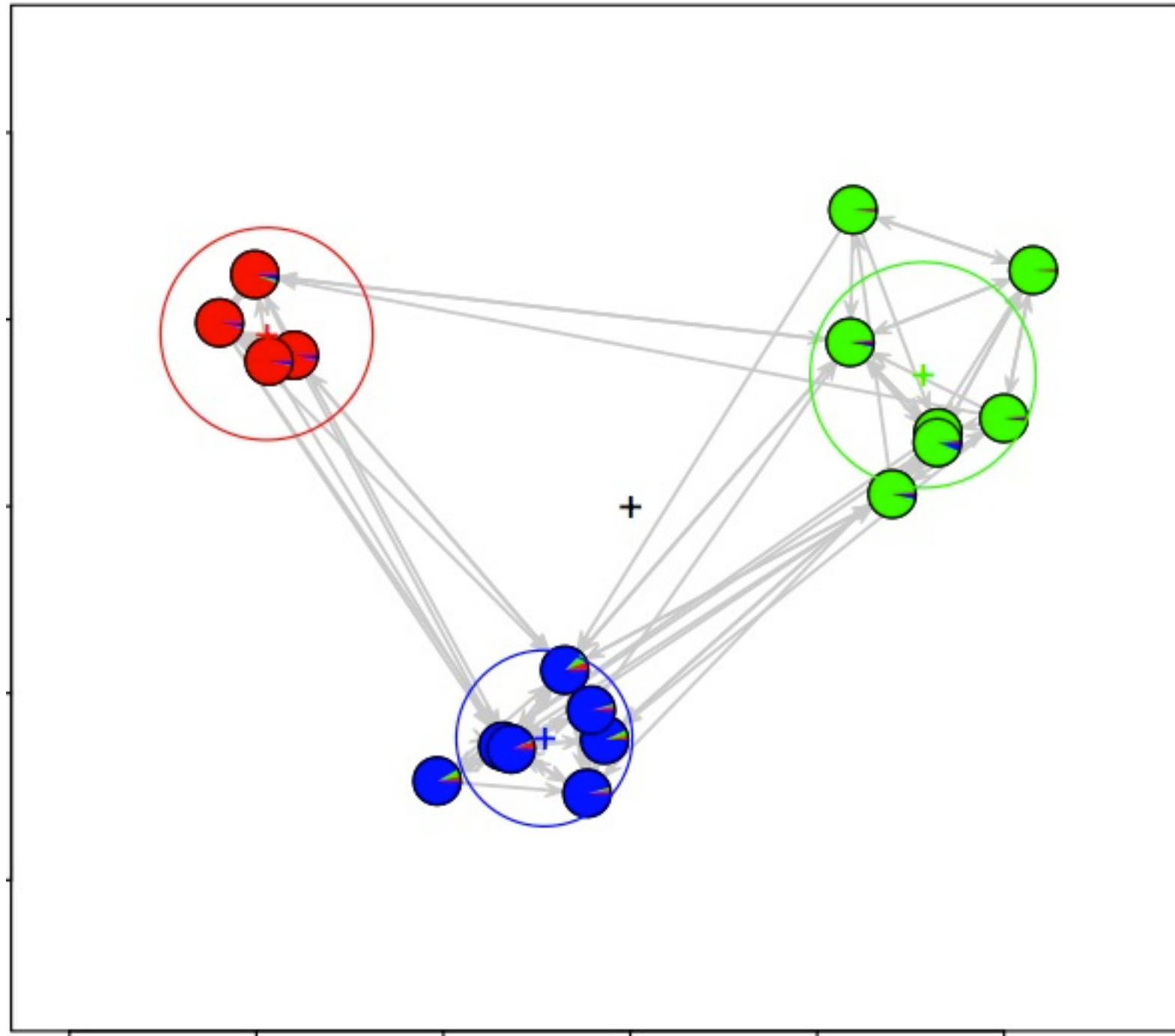
- Exploratory data analysis



Why use clustering...

... instead of classification?

- Exploratory data analysis



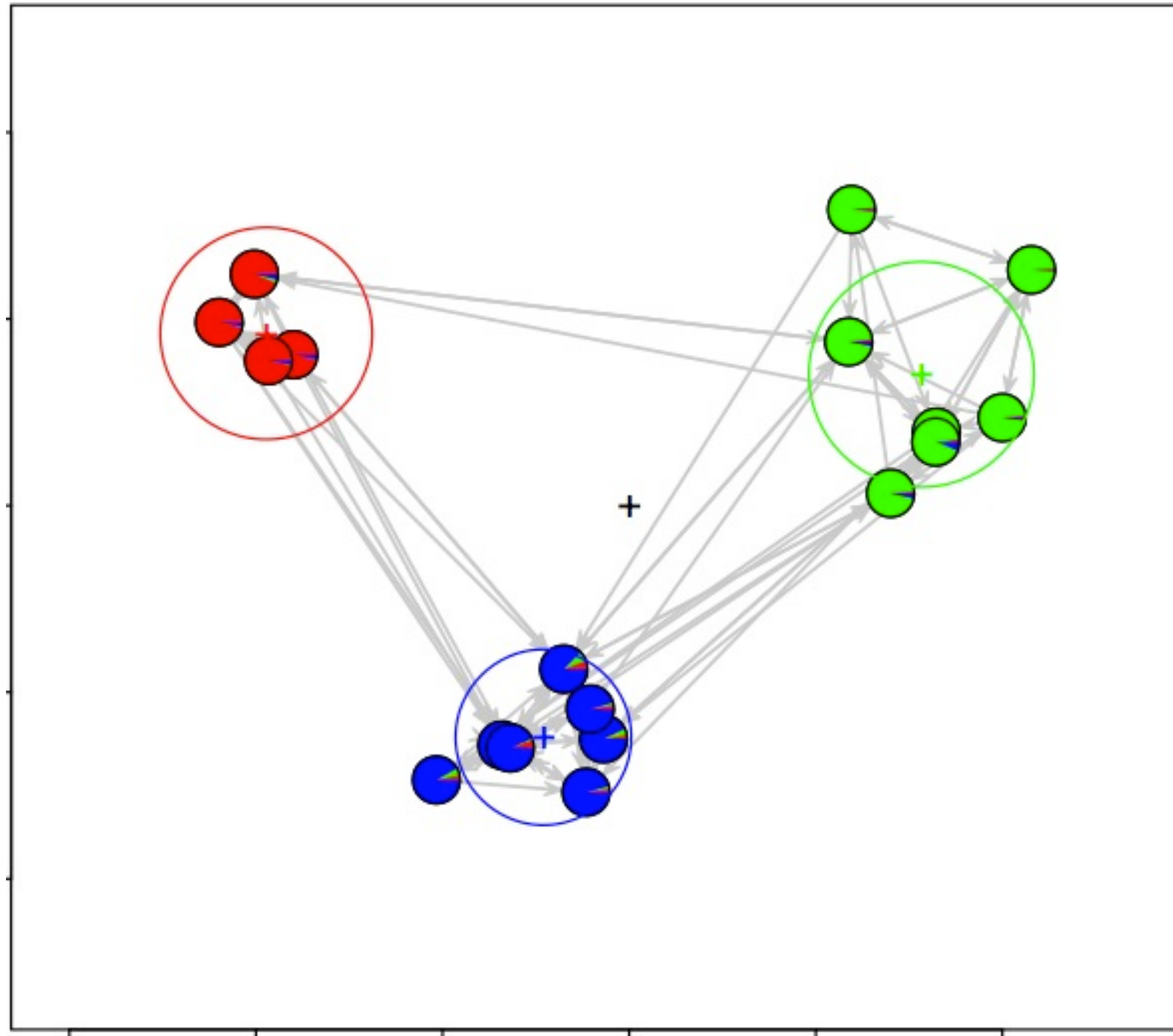
Datum: person

Similarity: the number of common interests of two people

Why use clustering...

... instead of classification?

- Exploratory data analysis



Datum: a binary vector specifying whether a person has each interest

Similarity: the number of common interests of two people

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Topic Analysis

NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Philharmonic and Juilliard School. "Our board felt that we had a mark on the future of the performing arts with these grants an act our traditional areas of support in health, medical research, education Hearst Foundation President Randolph A. Hearst said Monday in Lincoln Center's share will be \$200,000 for its new building, which and provide new public facilities. The Metropolitan Opera Co. and will receive \$400,000 each. The Juilliard School, where music and

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Topic Analysis

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Topic Analysis

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Datum: word

Similarity: how many documents exist where two words co-occur

the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Topic Analysis

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Datum: binary vector indicating document occurrence

Similarity: how many documents exist where two words co-occur

the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Document clustering

The screenshot shows the Carrot2 search engine interface. At the top, there is a navigation bar with links: About | More demos | Download | Carrot2 @ sf.net | Carrot Search. Below this is a search bar with the text 'tiger' and a 'Search' button. To the left of the search bar are logos for Yahoo!, Google, MSN, PUT, Wikipedia, ODP, and Jobs. The main content area displays search results for 'tiger'. On the left, there is a sidebar with a folder tree under 'All results (100)'. The folders listed are: Mac OS (9), Tiger Woods (5), Tiger Cubs (4), Computer (4), Onitsuka Tiger by Asics (4), Information on the Tiger (6), Security Tool (3), Technology Tiger Attack Helicopter (3), Sign (3), Siberian Tiger (3), and Geographic (2). The 'Tiger Woods (5)' folder is selected. The main results area shows three search results:

- 5: Official Website for Tiger Woods
Official site for pro golfer Tiger Woods, complete with video interviews, photos, stats, and features.
<http://www.tigerwoods.com/>
- 34: tiger -- Encyclopædia Britannica
tiger ... Woods, Tiger ... tiger beetle ...
<http://www.britannica.com/eb/article-9072439/tiger>
- 66: Abilene Reporter News: Tiger Woods
Tiger Woods Haunted by Tears, Failure. Bulk of Masters Field Set by Final Rank ... Tiger Finishes the Season in Style. Els Wins South African Open by 3 Strokes ...
http://www.reporternews.com/abil/sp_tiger_woods/0,1874,ABIL_

At the bottom of the interface, it says: Query: tiger -- input: Yahoo! (100 results) -- Clusterer: Lingo

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Document clustering

Query: tiger -- input: Yahoo! (100 results) -- Clusterer: Lingo

Datum: document

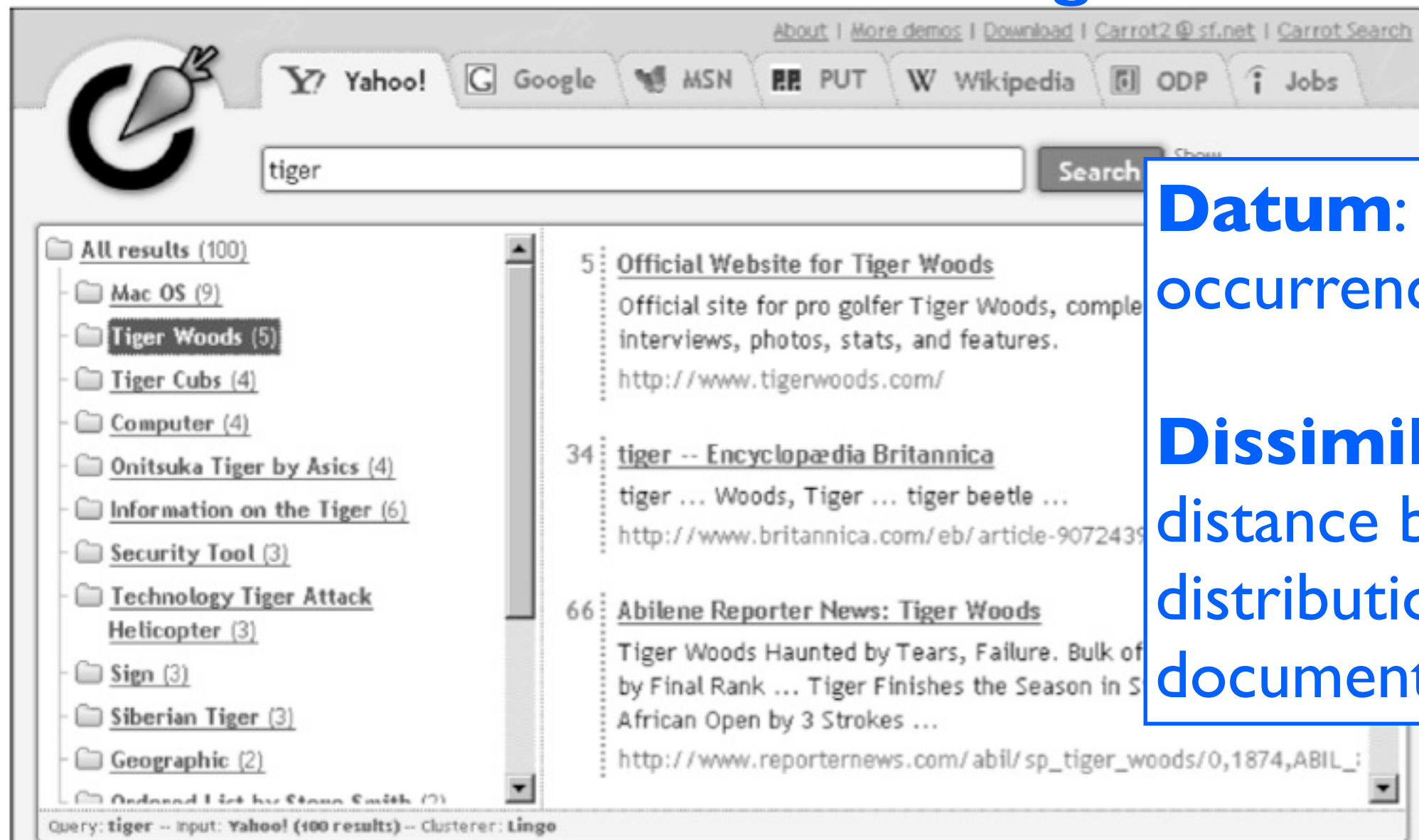
Dissimilarity:
distance between topic
distributions of two
documents

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Document clustering



Datum: vector of topic occurrences

Dissimilarity:
distance between topic distributions of two documents

Why use clustering...

... instead of classification?

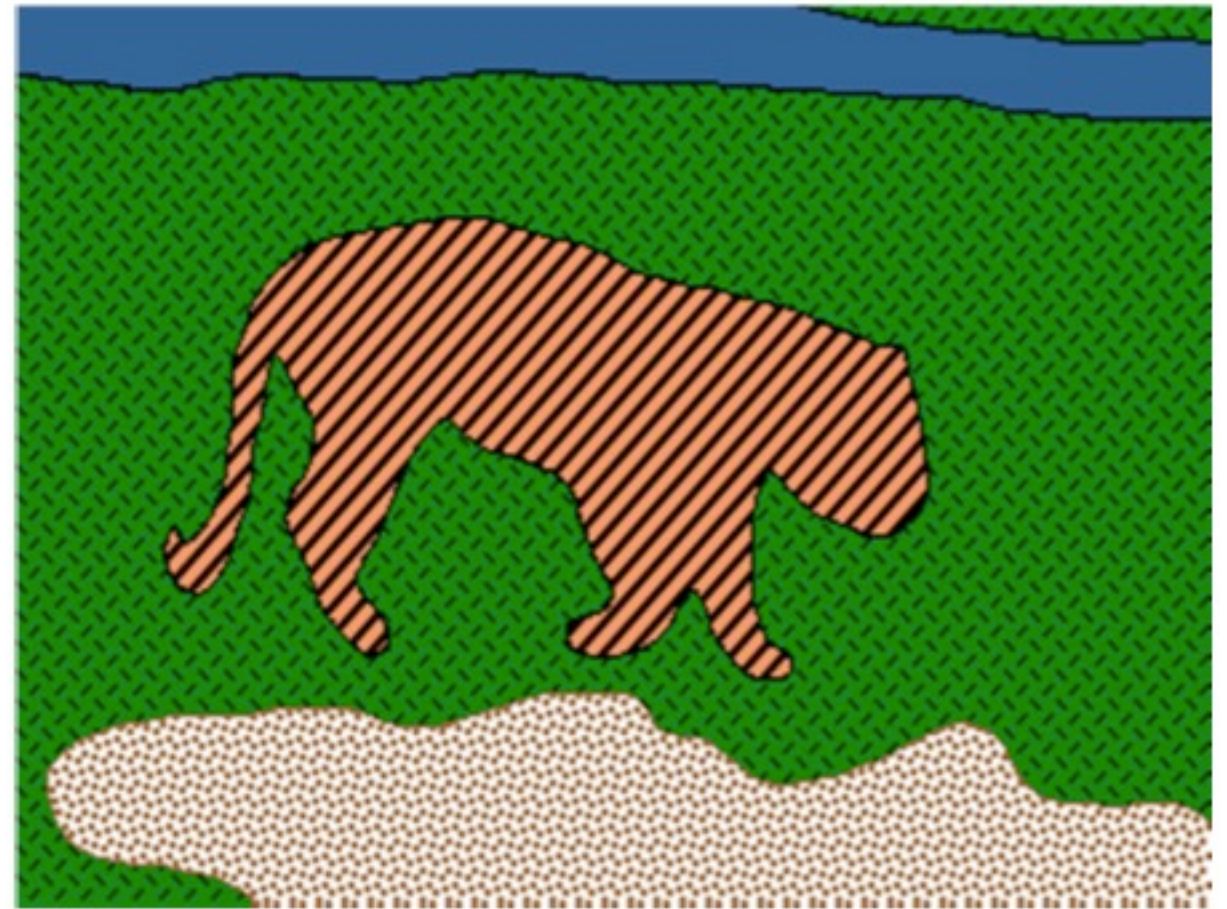
- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Image segmentation

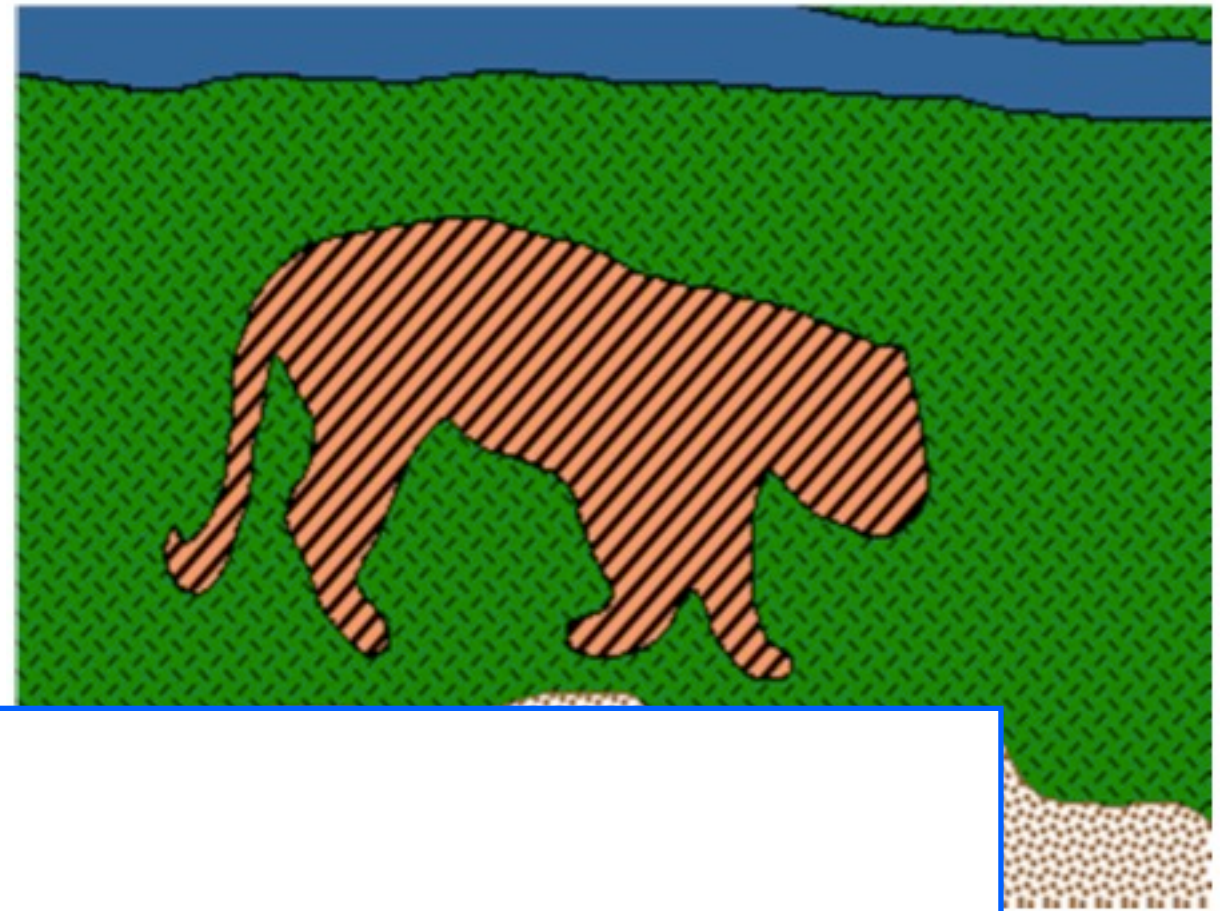


Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Image segmentation



Datum: pixel

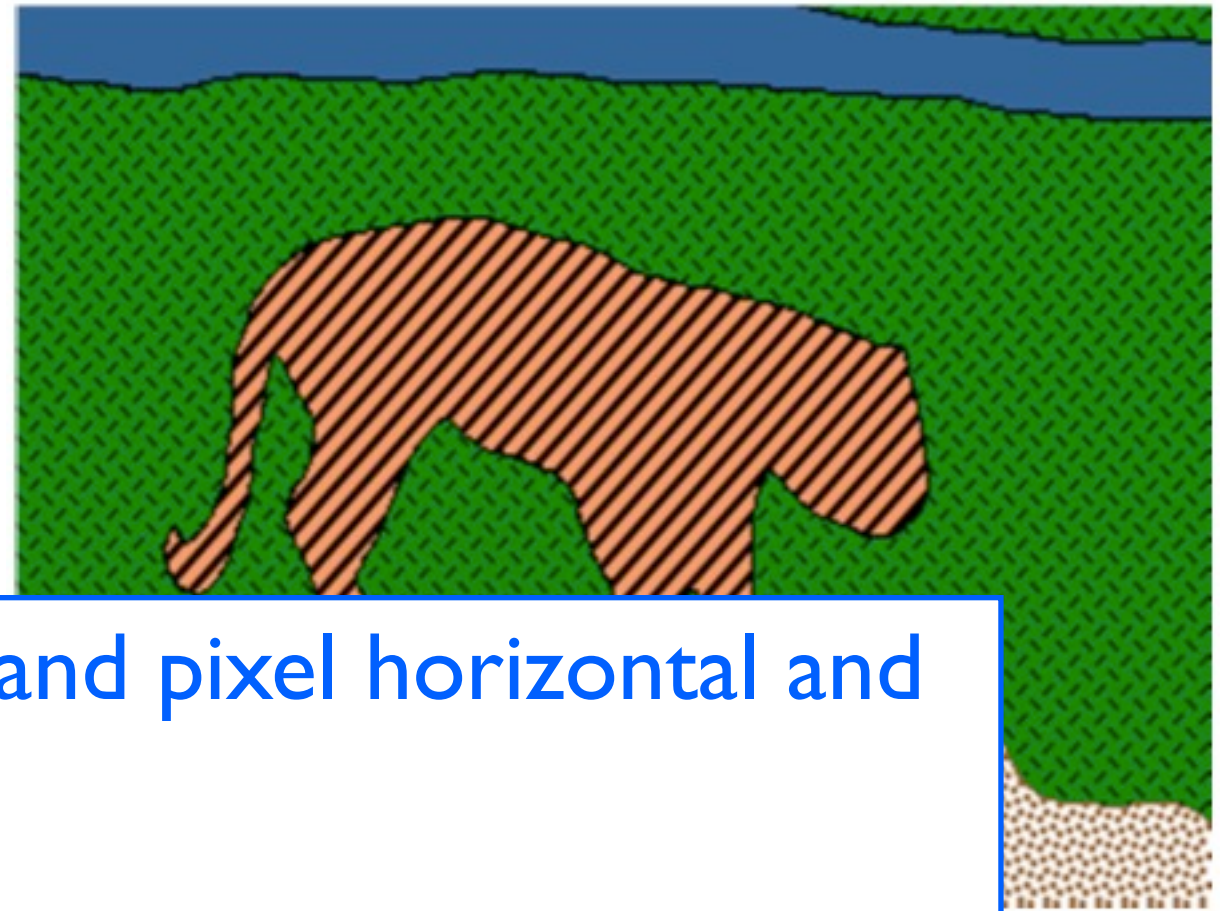
Dissimilarity: difference in color + difference in location

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

Image segmentation



Datum: pixel RGB values and pixel horizontal and vertical locations

Dissimilarity: difference in color + difference in location

Why use clustering...

... instead of classification?

- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

... when the cartoon looks so easy?

- High-dimensional data
- Big data
- Data not numerical

Outline

0. What is clustering?

1. K means algorithm
2. Clustering evaluation
3. Clustering trouble-shooting
4. Example

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
3. Clustering trouble-shooting
4. Example

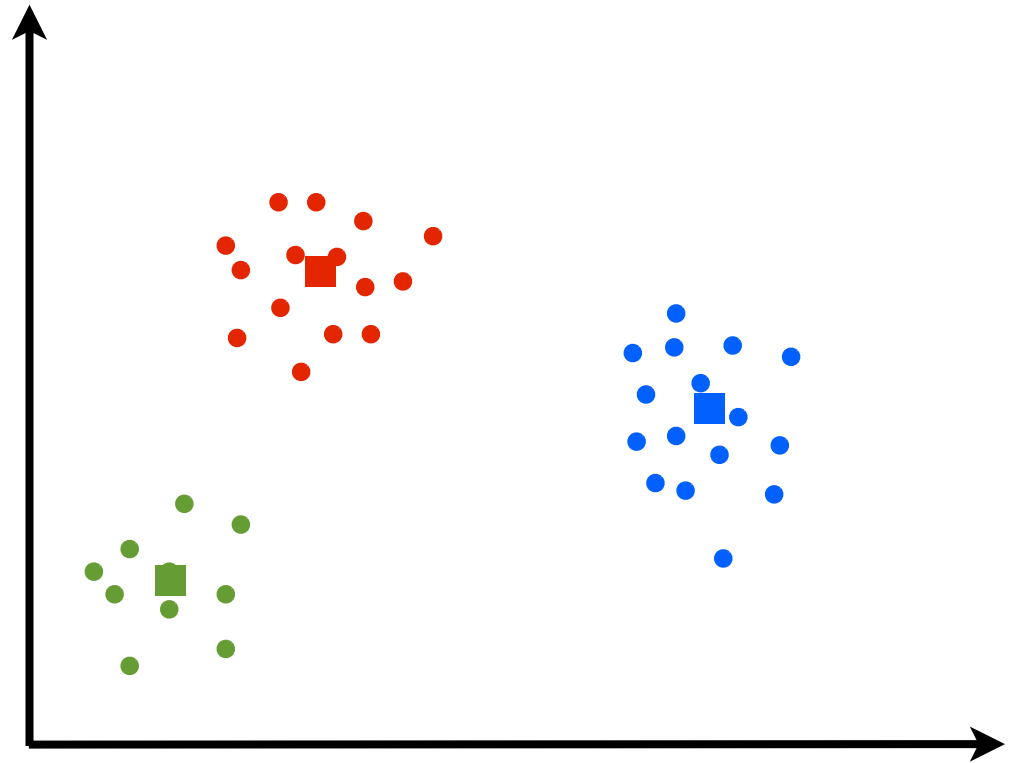
Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
3. Clustering trouble-shooting
4. Example

K means algorithm

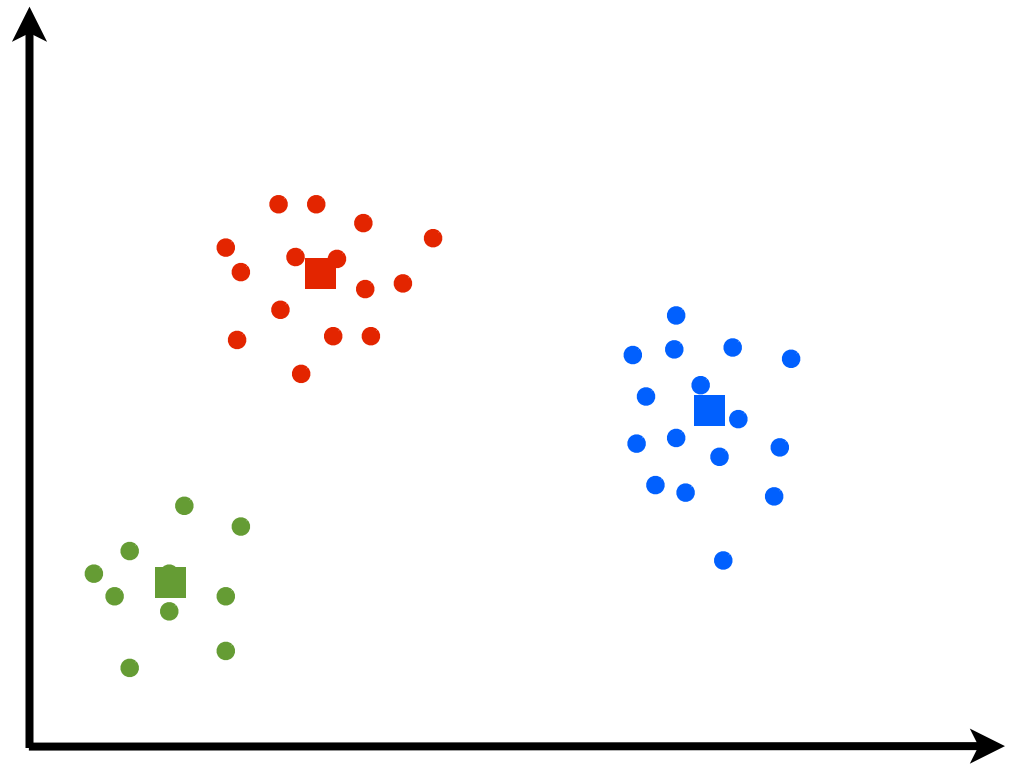
Benefits



K means algorithm

Benefits

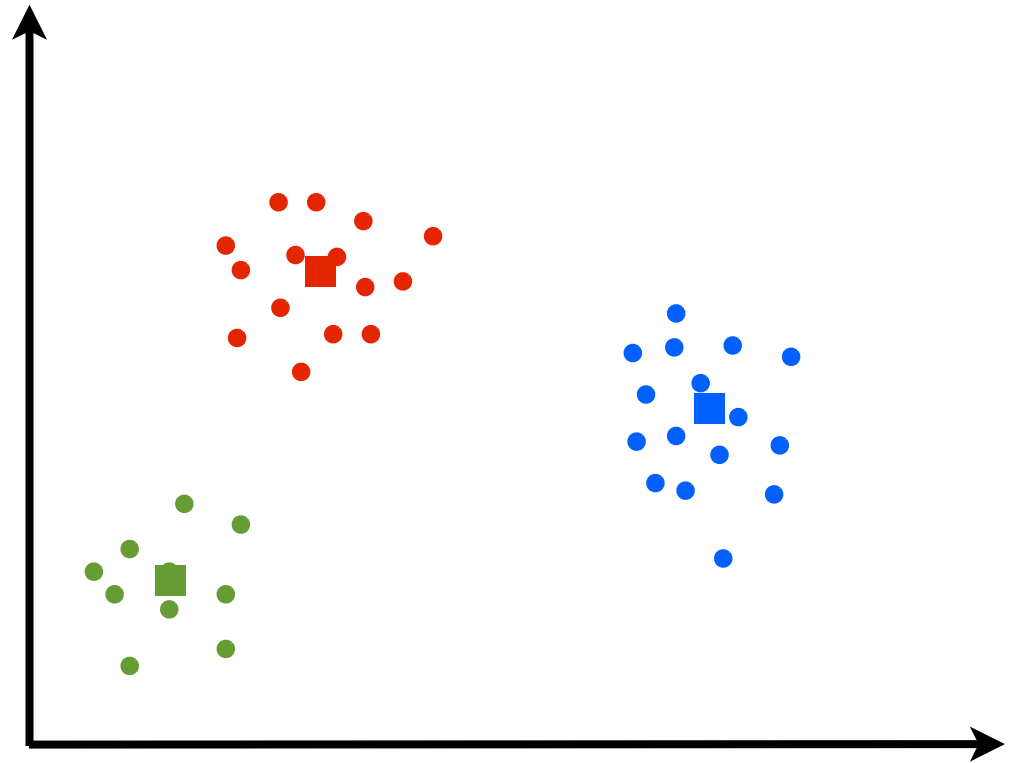
- Fast



K means algorithm

Benefits

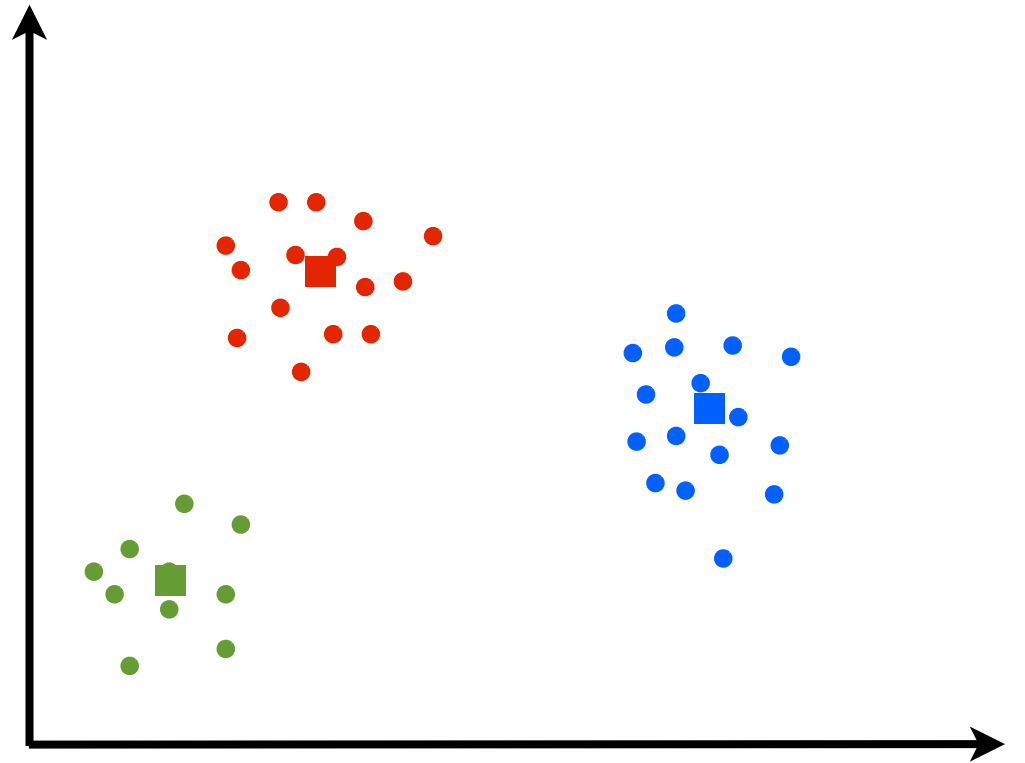
- Fast
- Fast



K means algorithm

Benefits

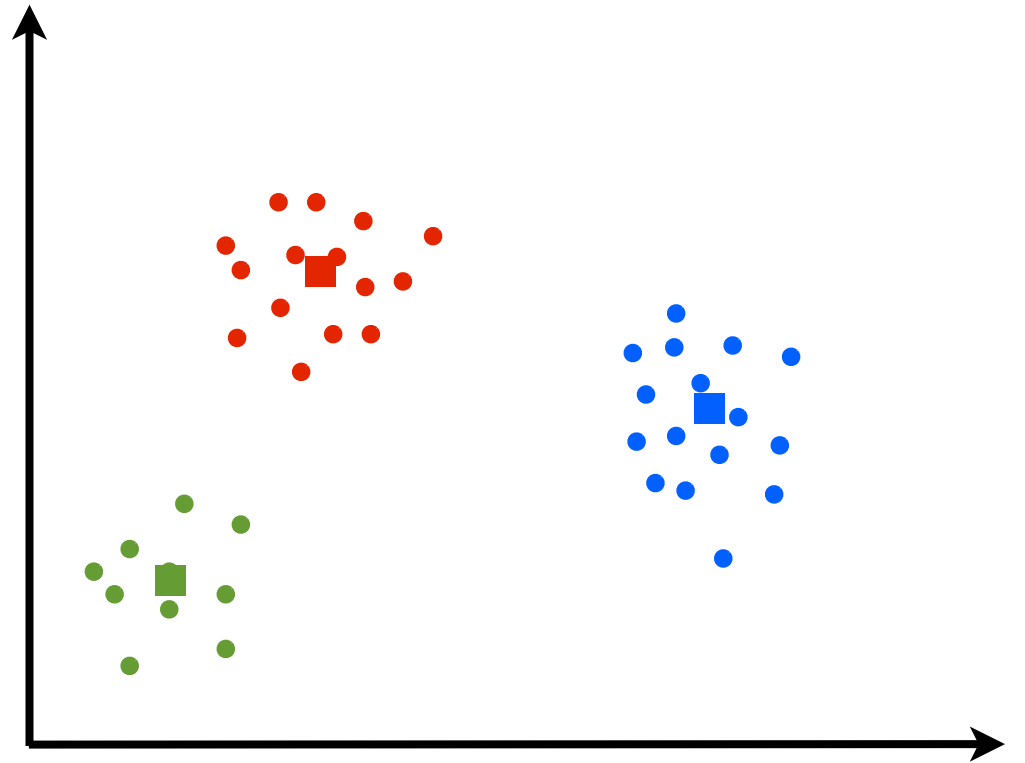
- Fast
- Fast
- Fast



K means algorithm

Benefits

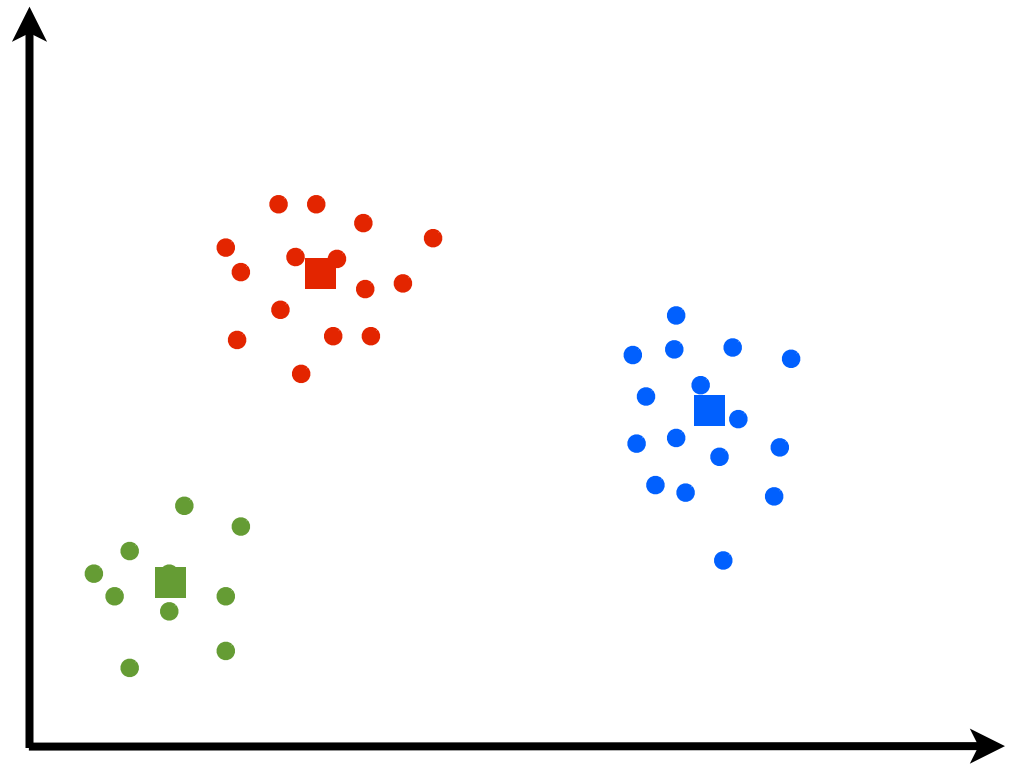
- Fast
- Conceptually straightforward



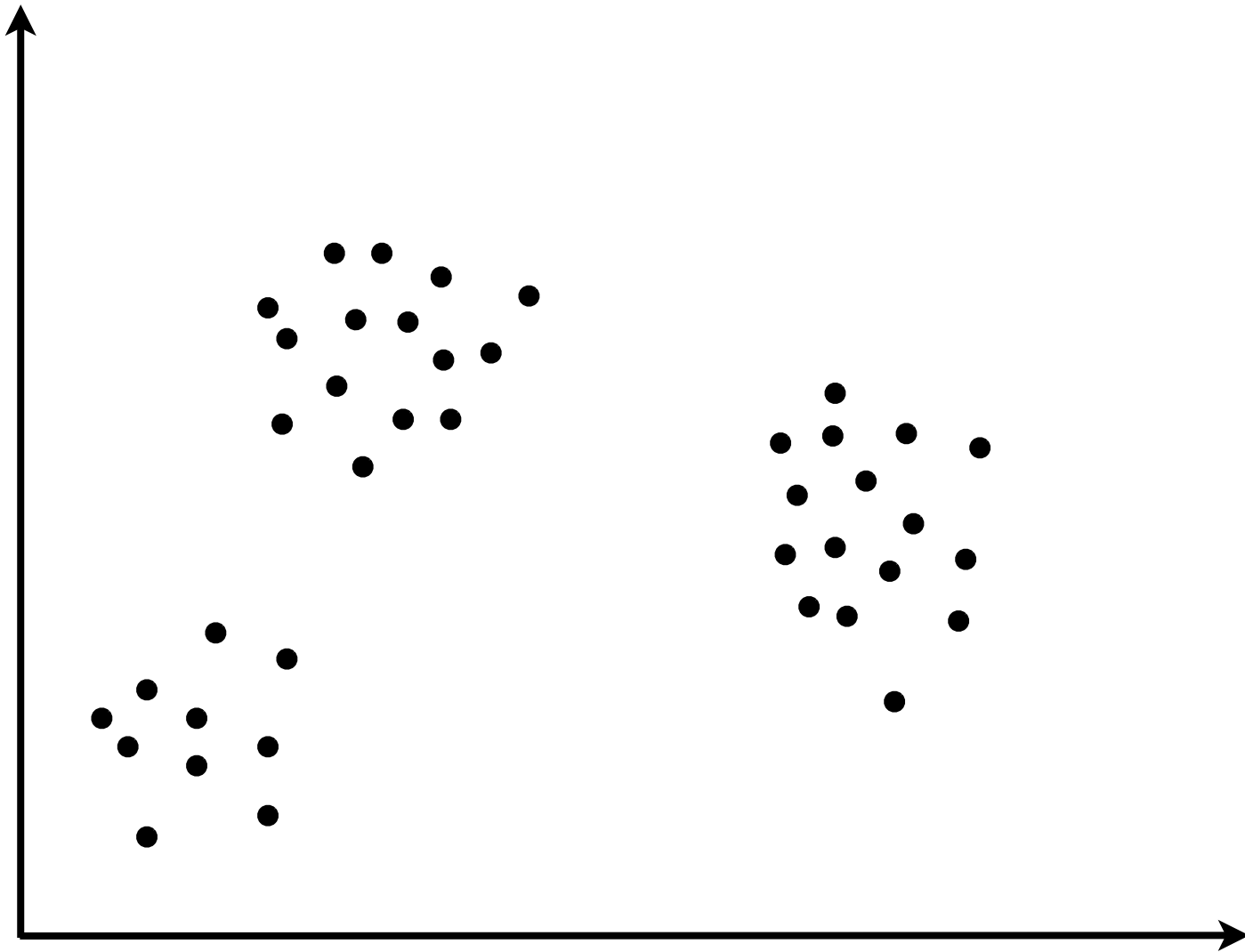
K means algorithm

Benefits

- Fast
- Conceptually straightforward
- Popular

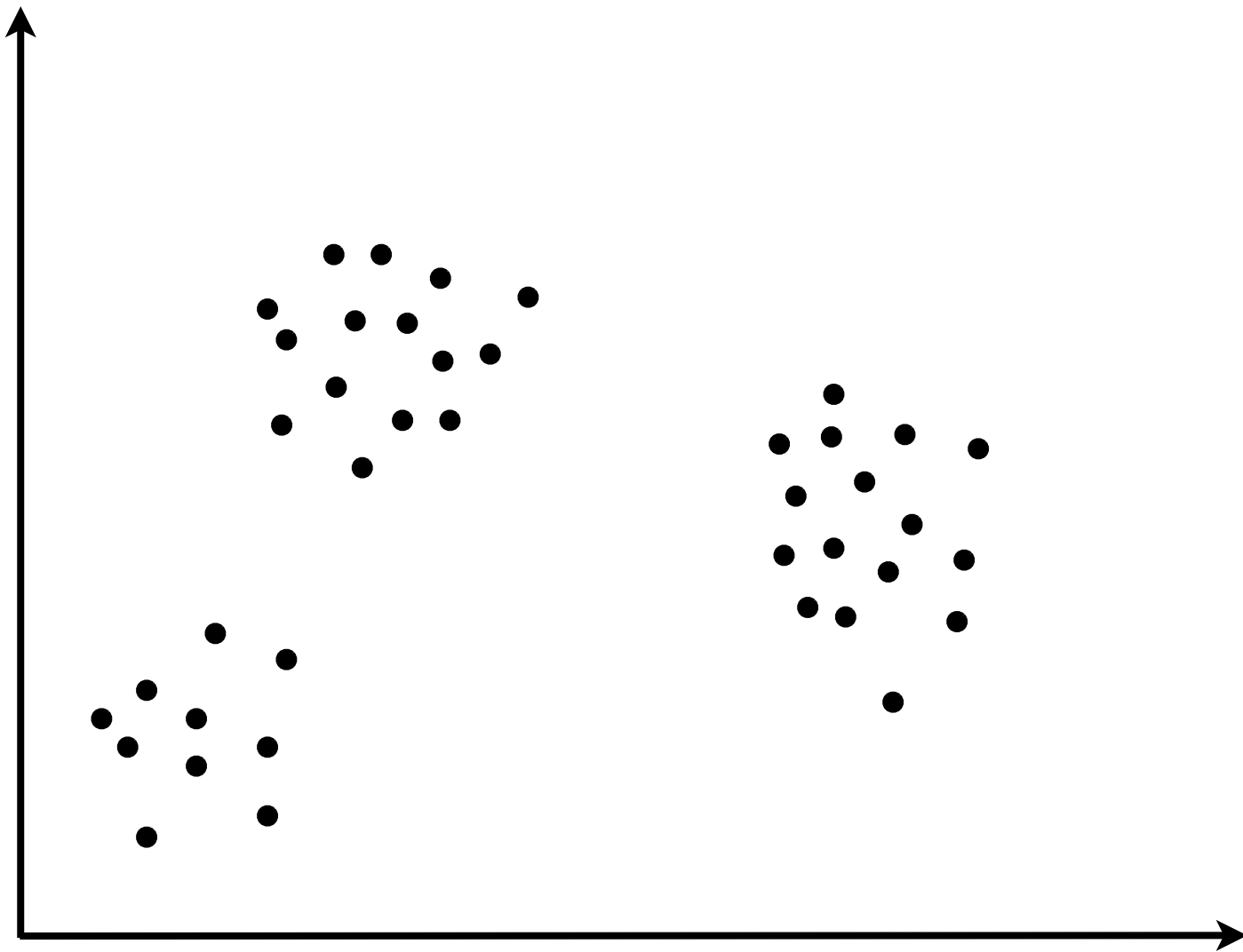


K means: preliminaries



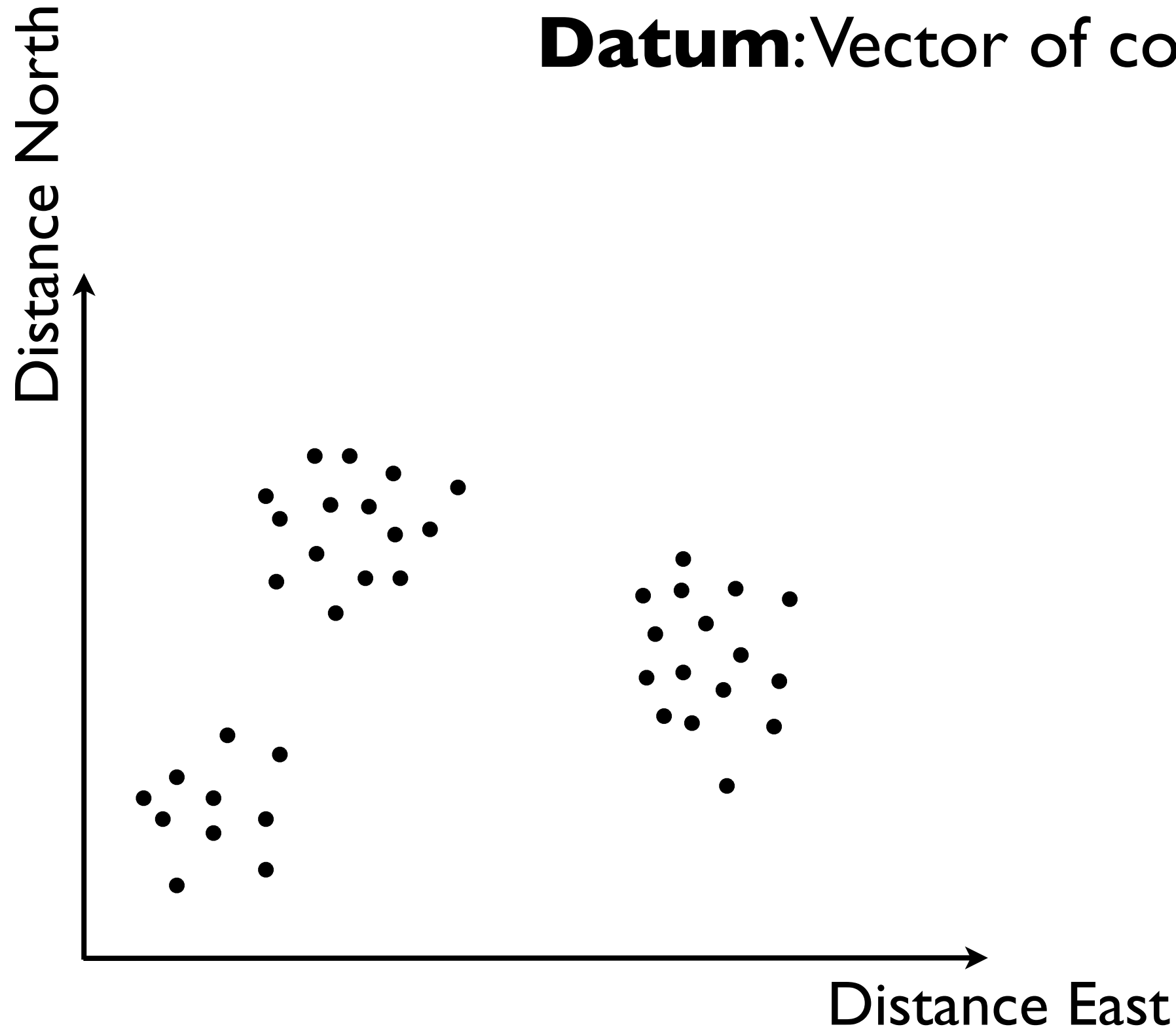
K means: preliminaries

Datum: Vector of continuous values



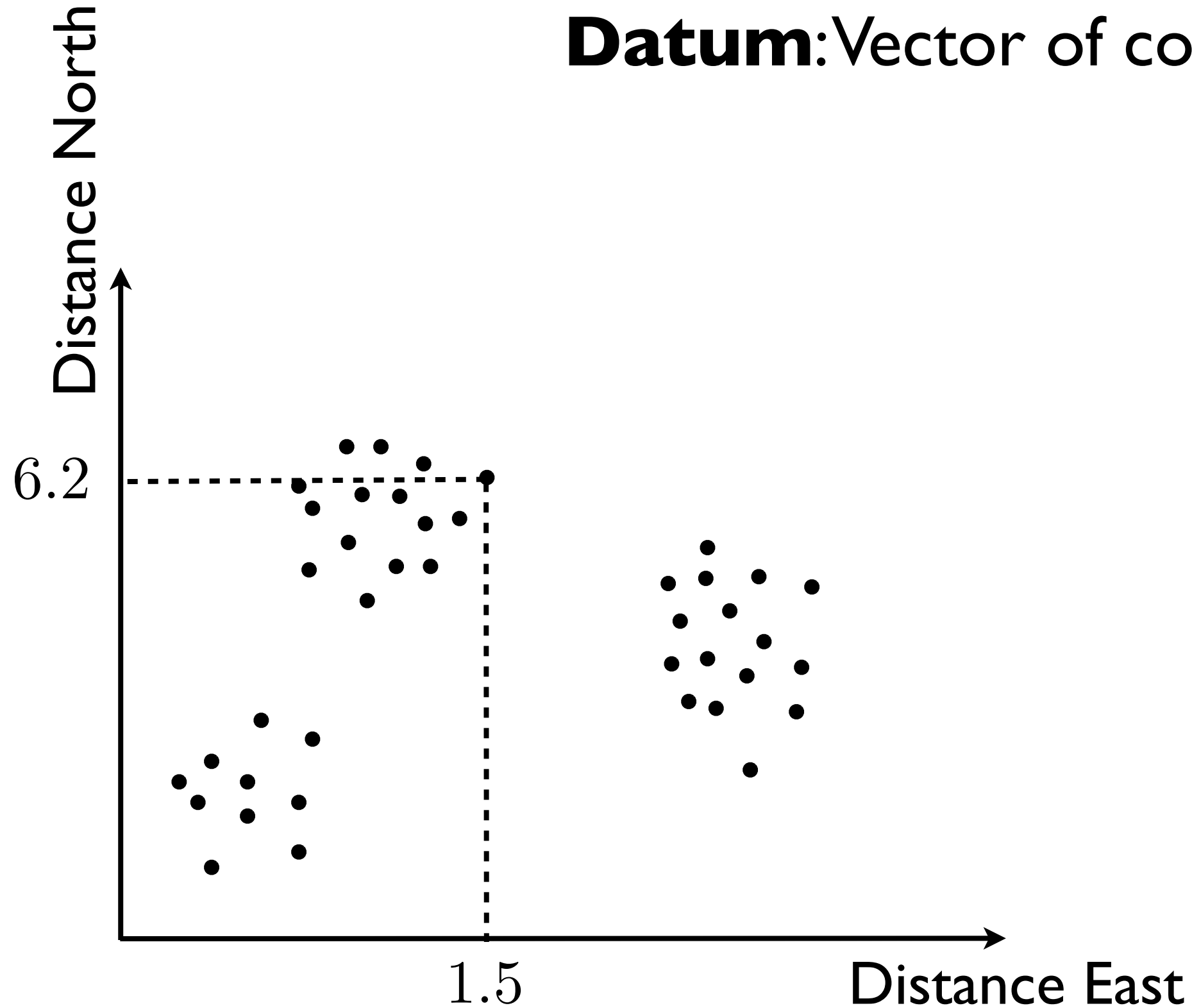
K means: preliminaries

Datum: Vector of continuous values



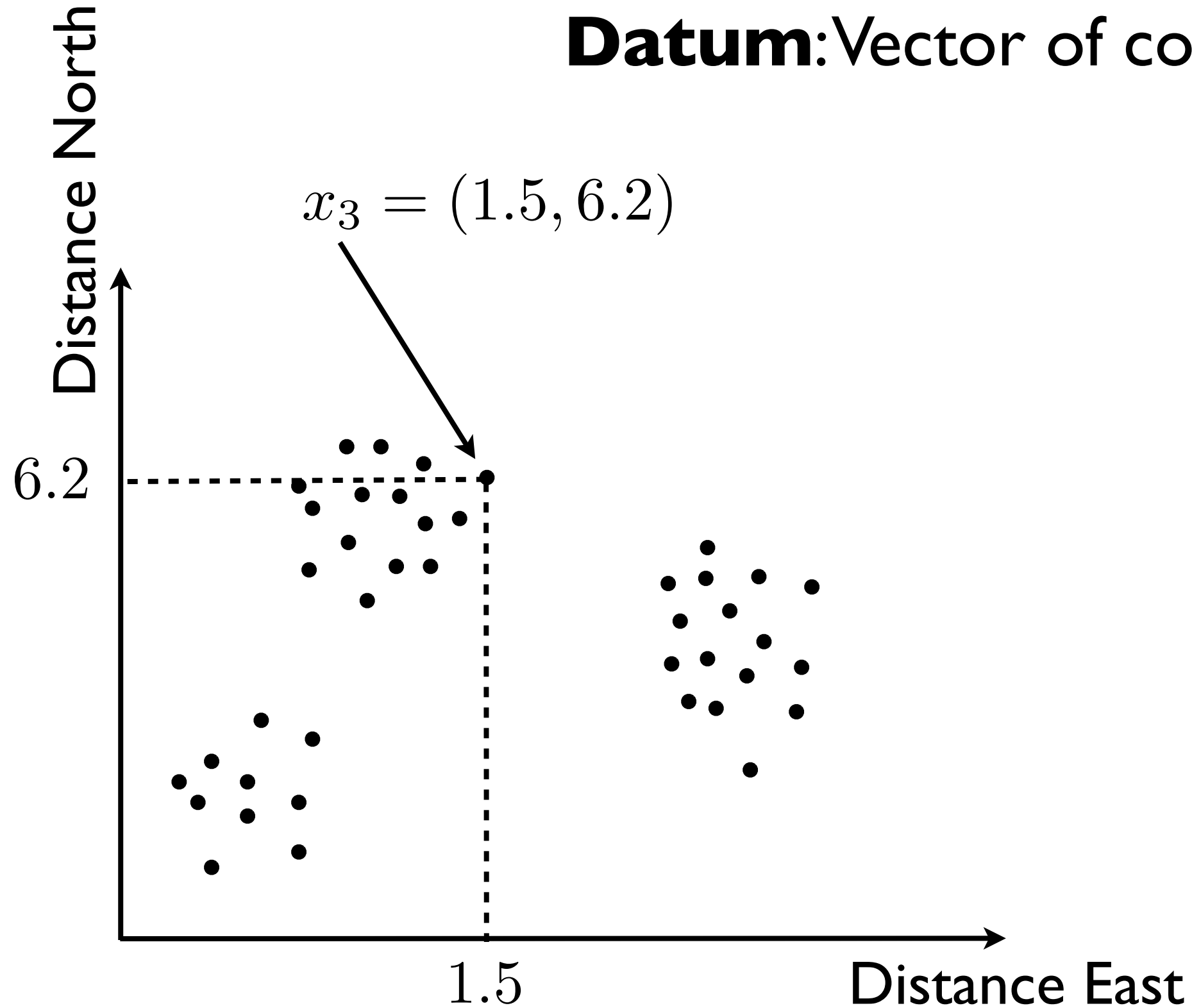
K means: preliminaries

Datum: Vector of continuous values



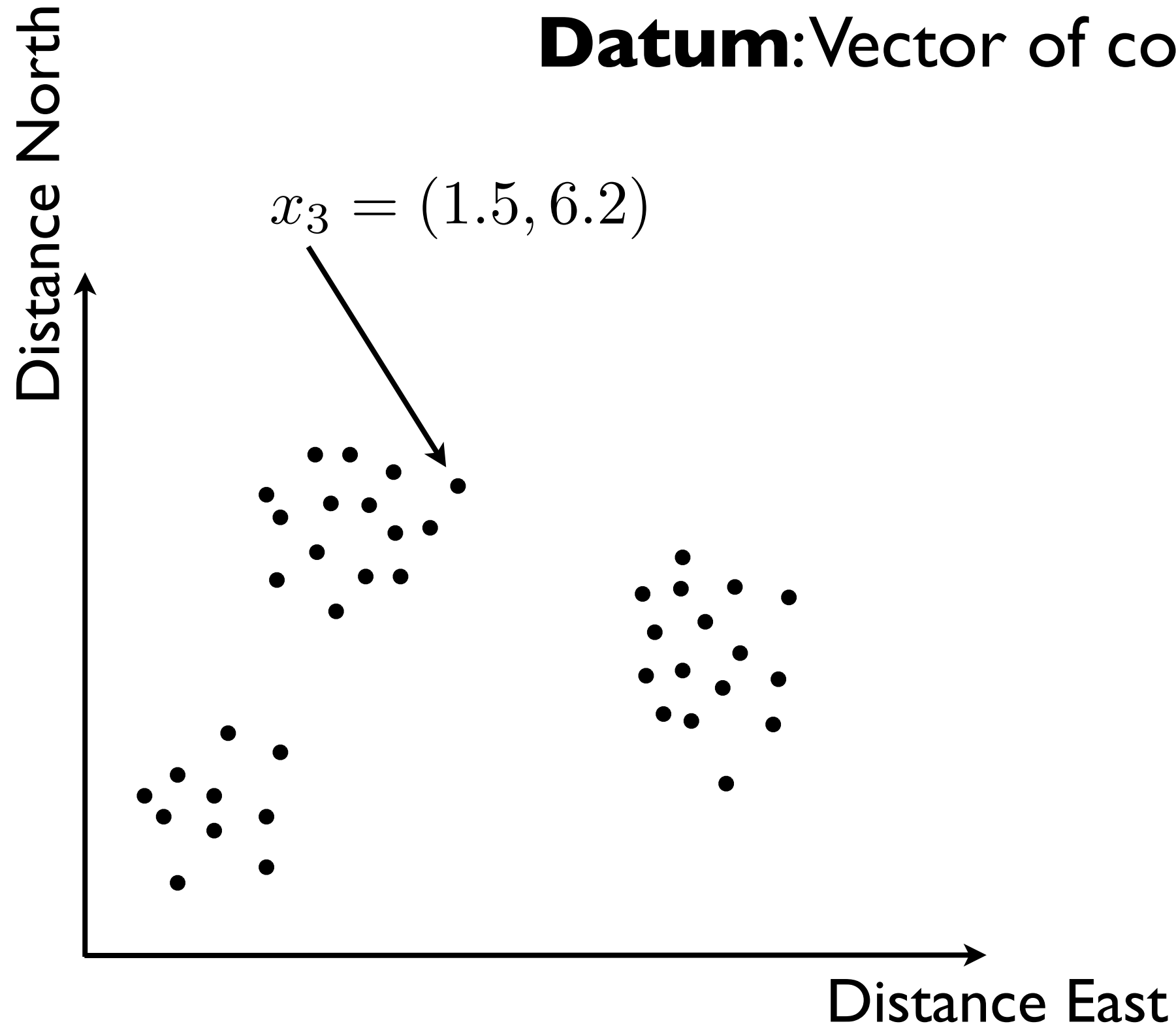
K means: preliminaries

Datum: Vector of continuous values



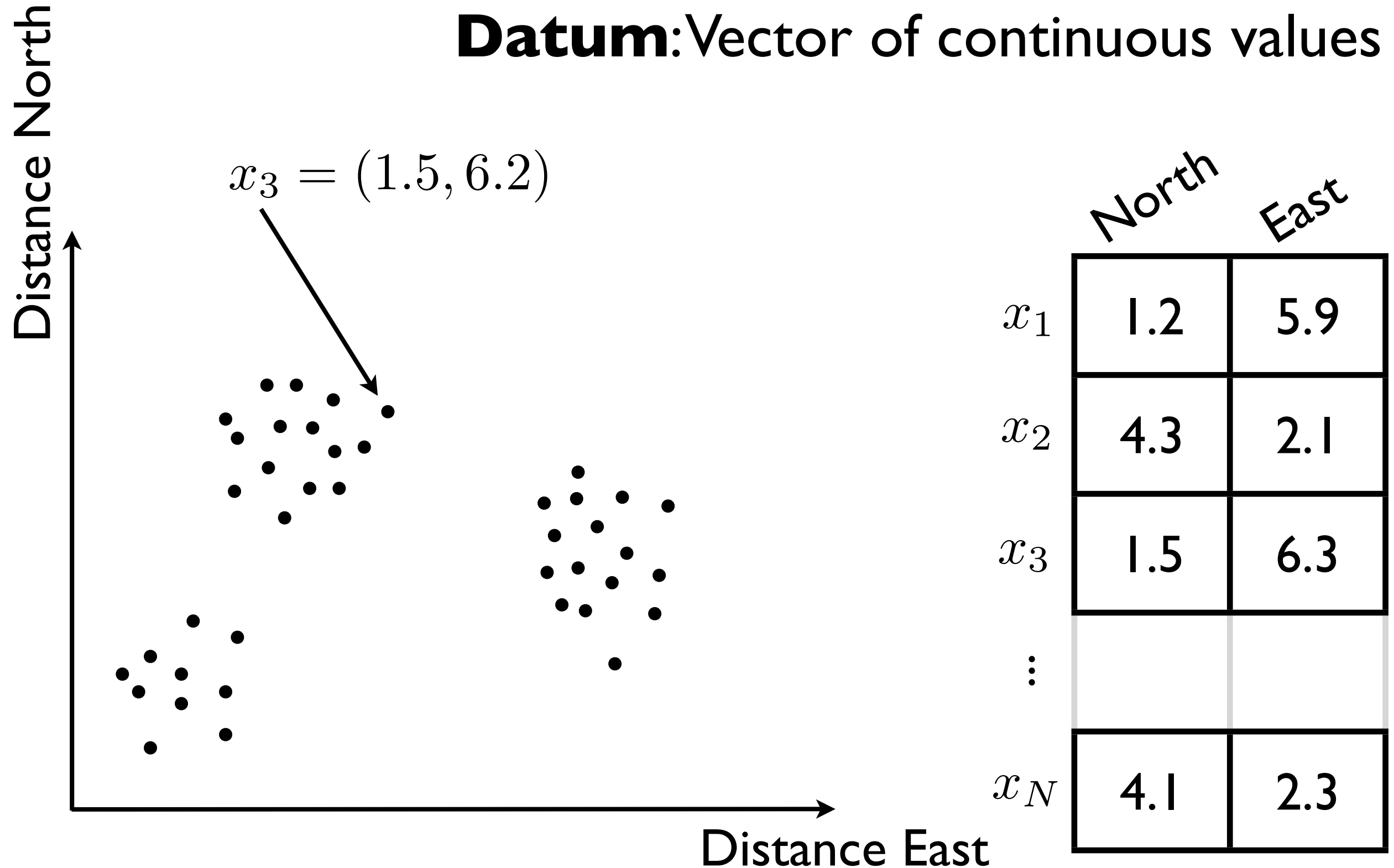
K means: preliminaries

Datum: Vector of continuous values



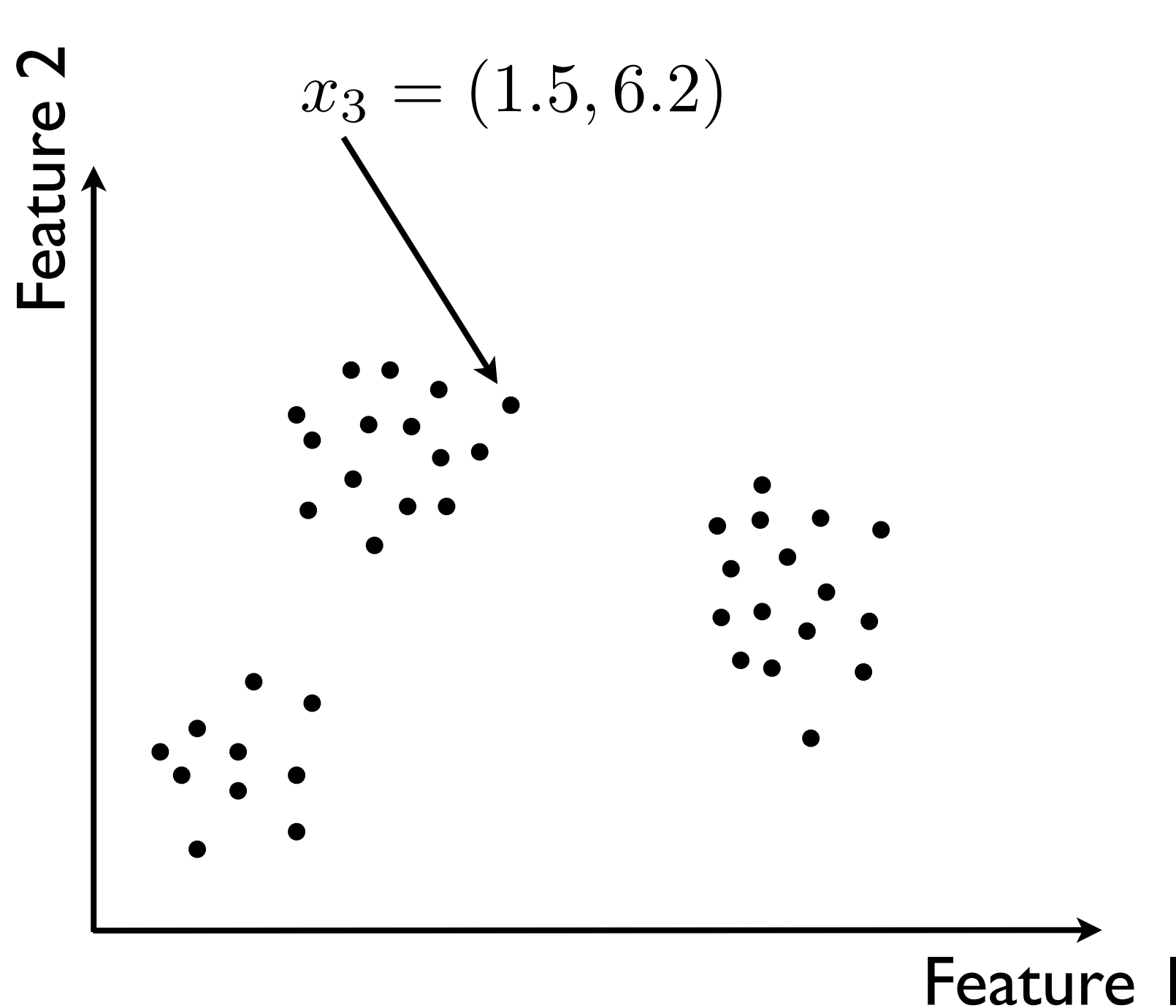
K means: preliminaries

Datum: Vector of continuous values



K means: preliminaries

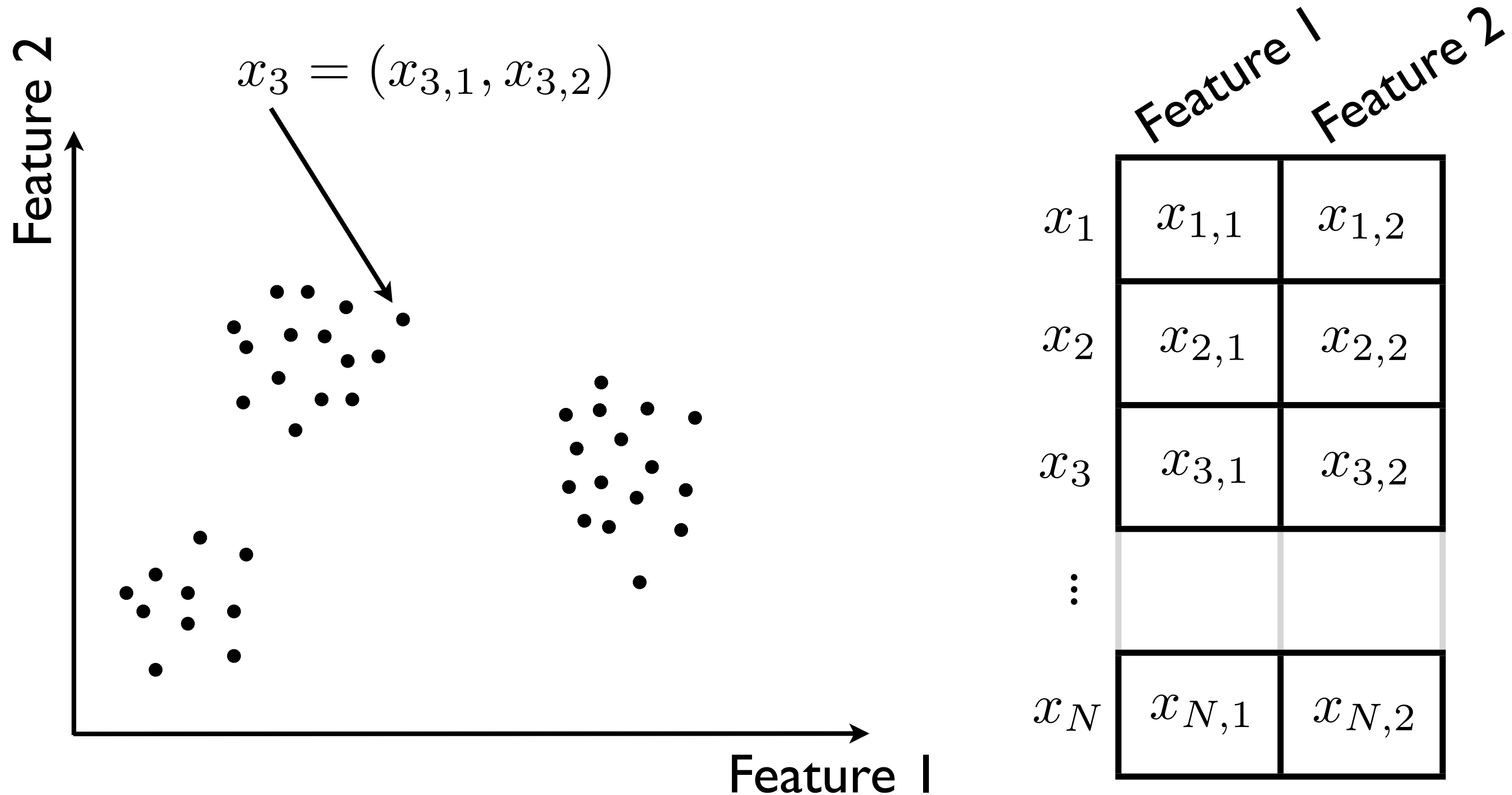
Datum: Vector of continuous values



	Feature 1	Feature 2
x_1	1.2	5.9
x_2	4.3	2.1
x_3	1.5	6.3
$\vdots$		
x_N	4.1	2.3

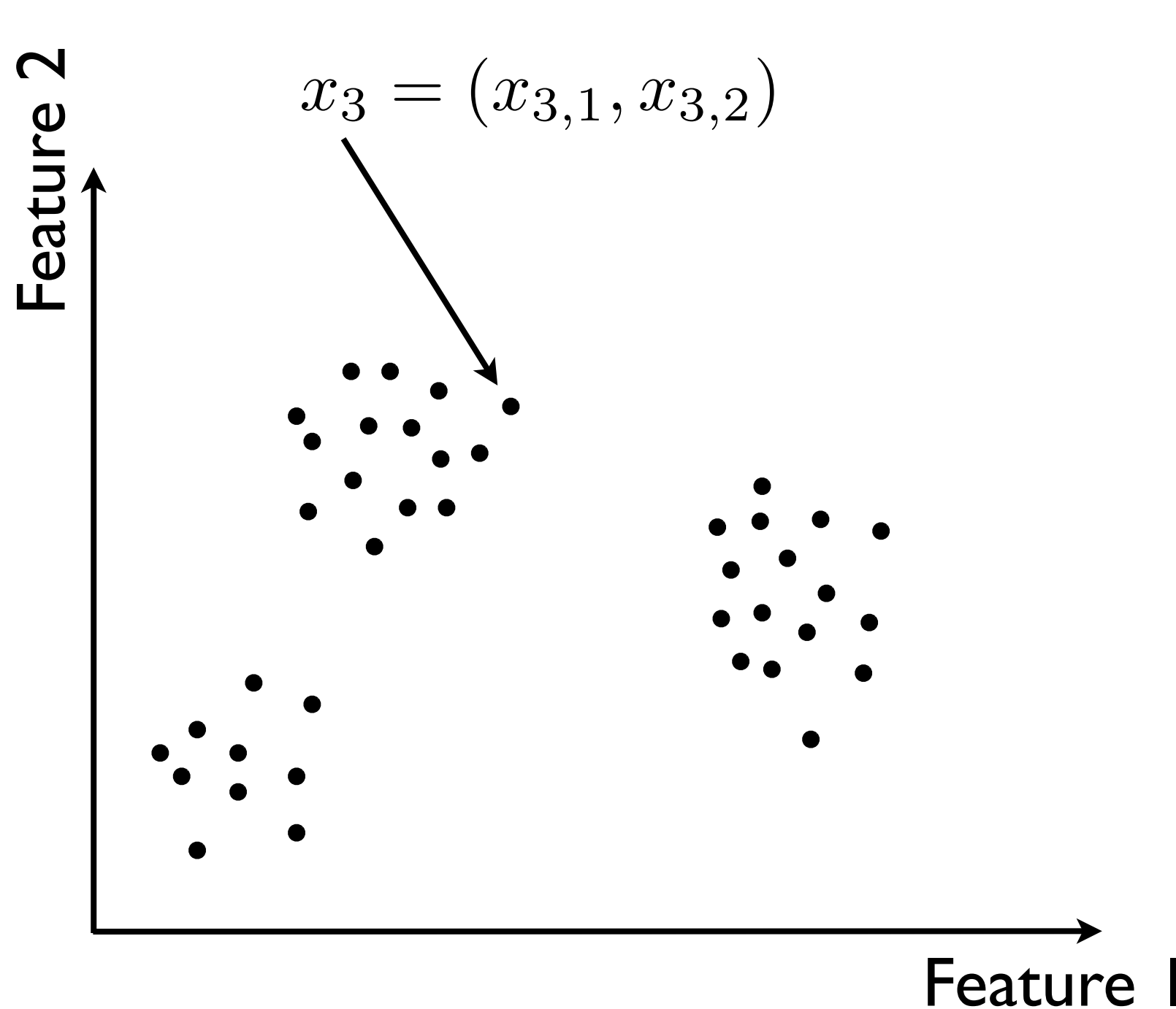
K means: preliminaries

Datum: Vector of continuous values



K means: preliminaries

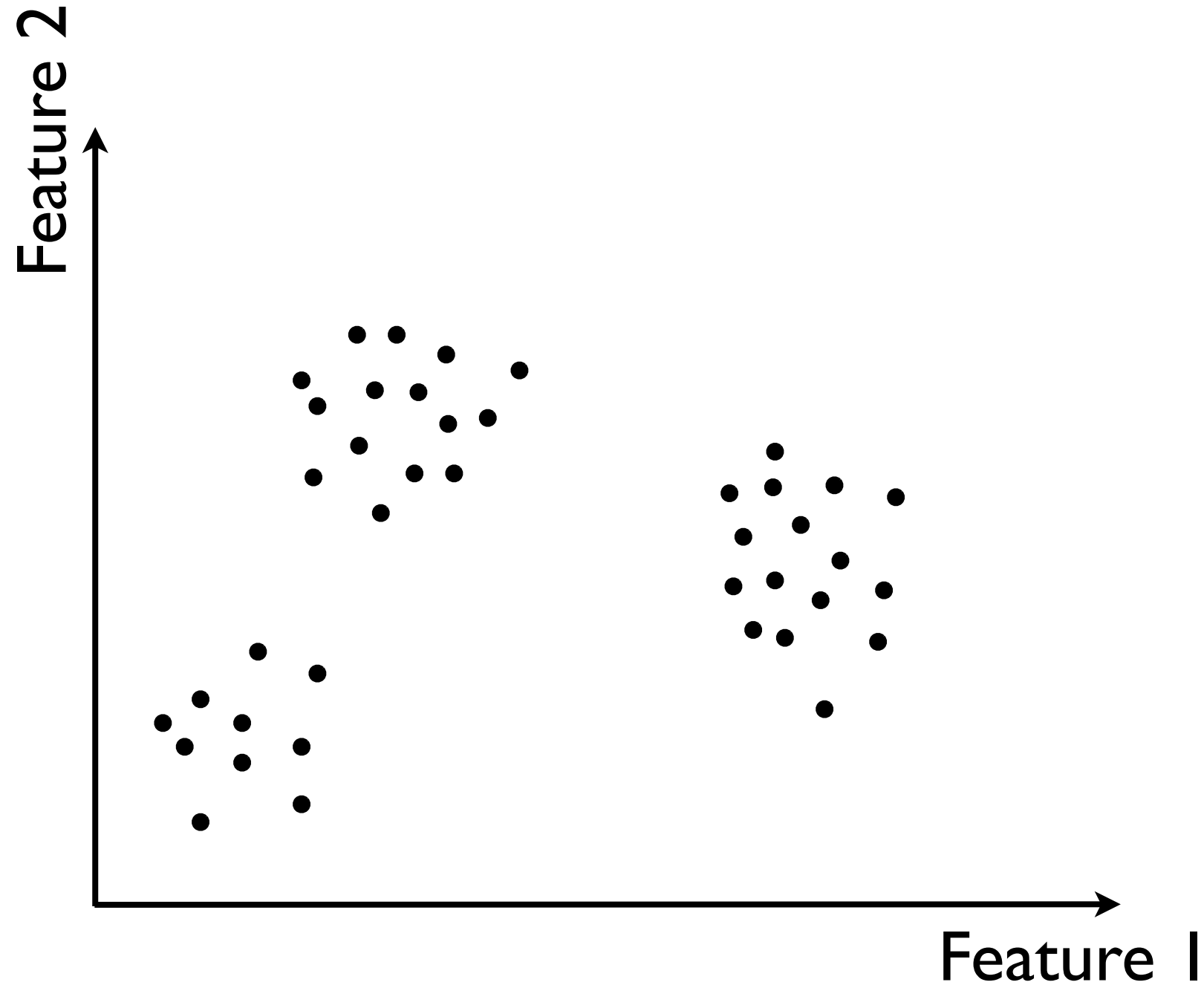
Datum: Vector of **D** continuous values



	Feature 1	Feature 2	...
x_1	$x_{1,1}$	$x_{1,2}$	
x_2	$x_{2,1}$	$x_{2,2}$	
x_3	$x_{3,1}$	$x_{3,2}$	
$\vdots$			
x_N	$x_{N,1}$	$x_{N,2}$	

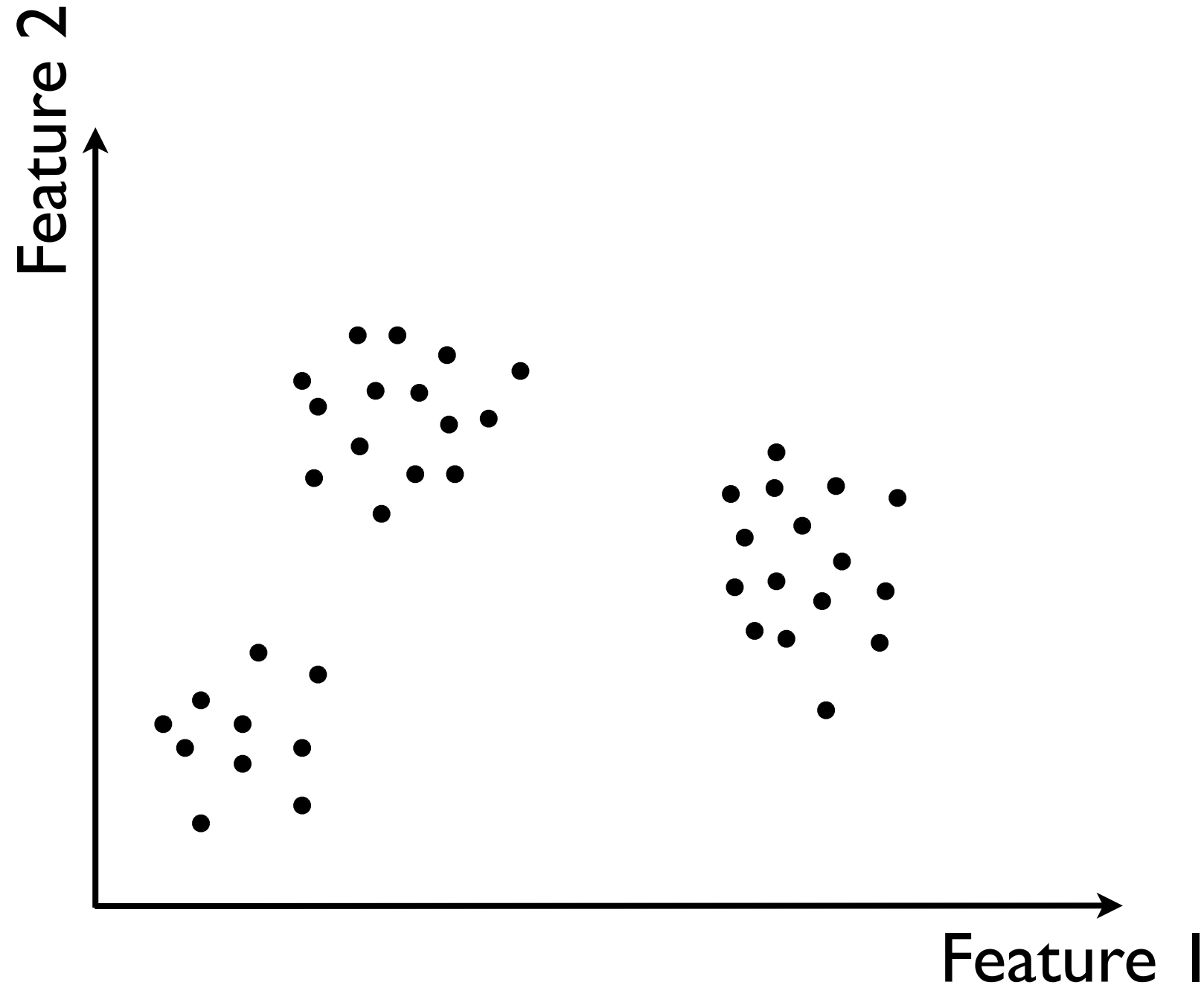
K means: preliminaries

Datum: Vector of D continuous values



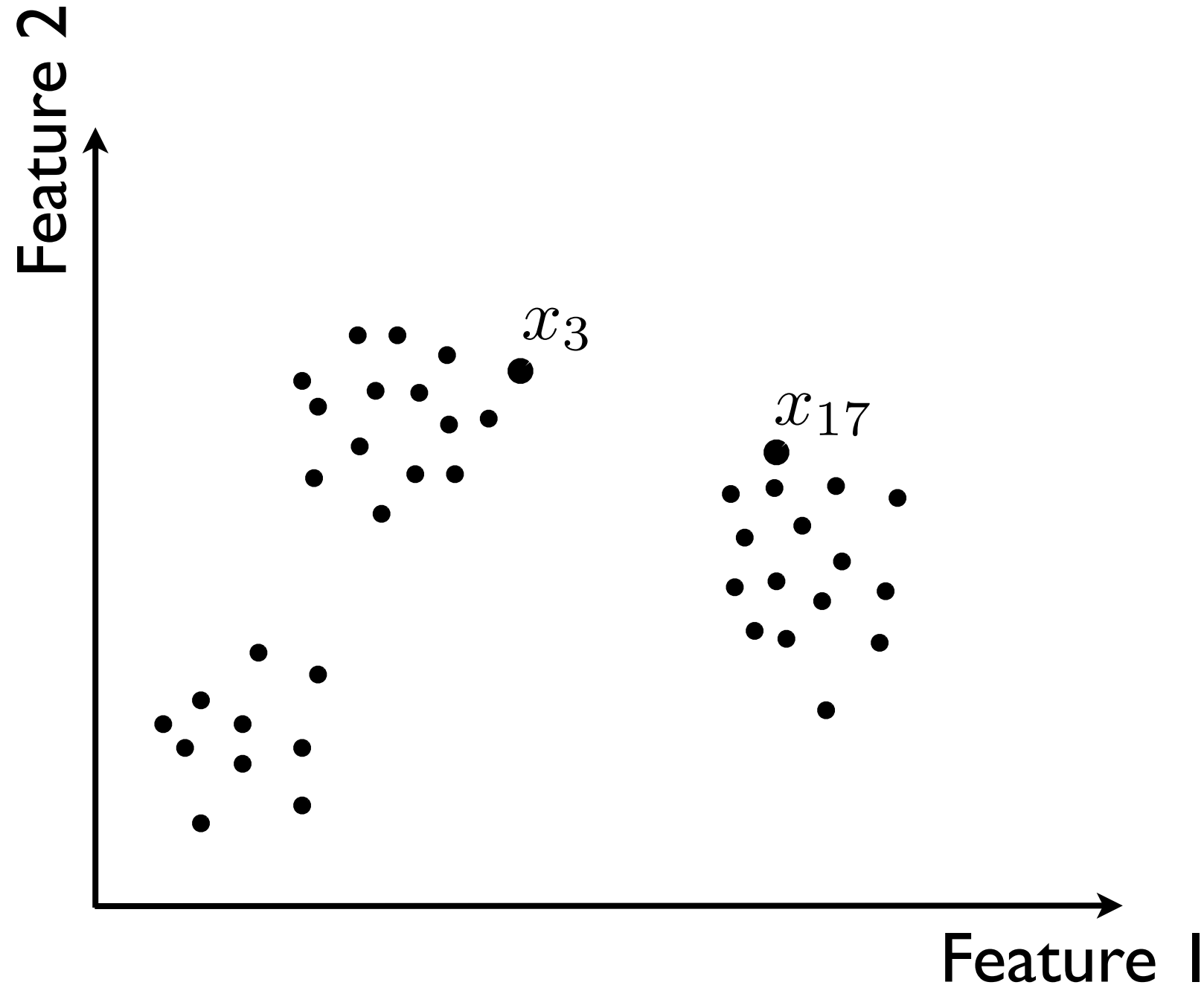
K means: preliminaries

Dissimilarity: Distance as the crow flies



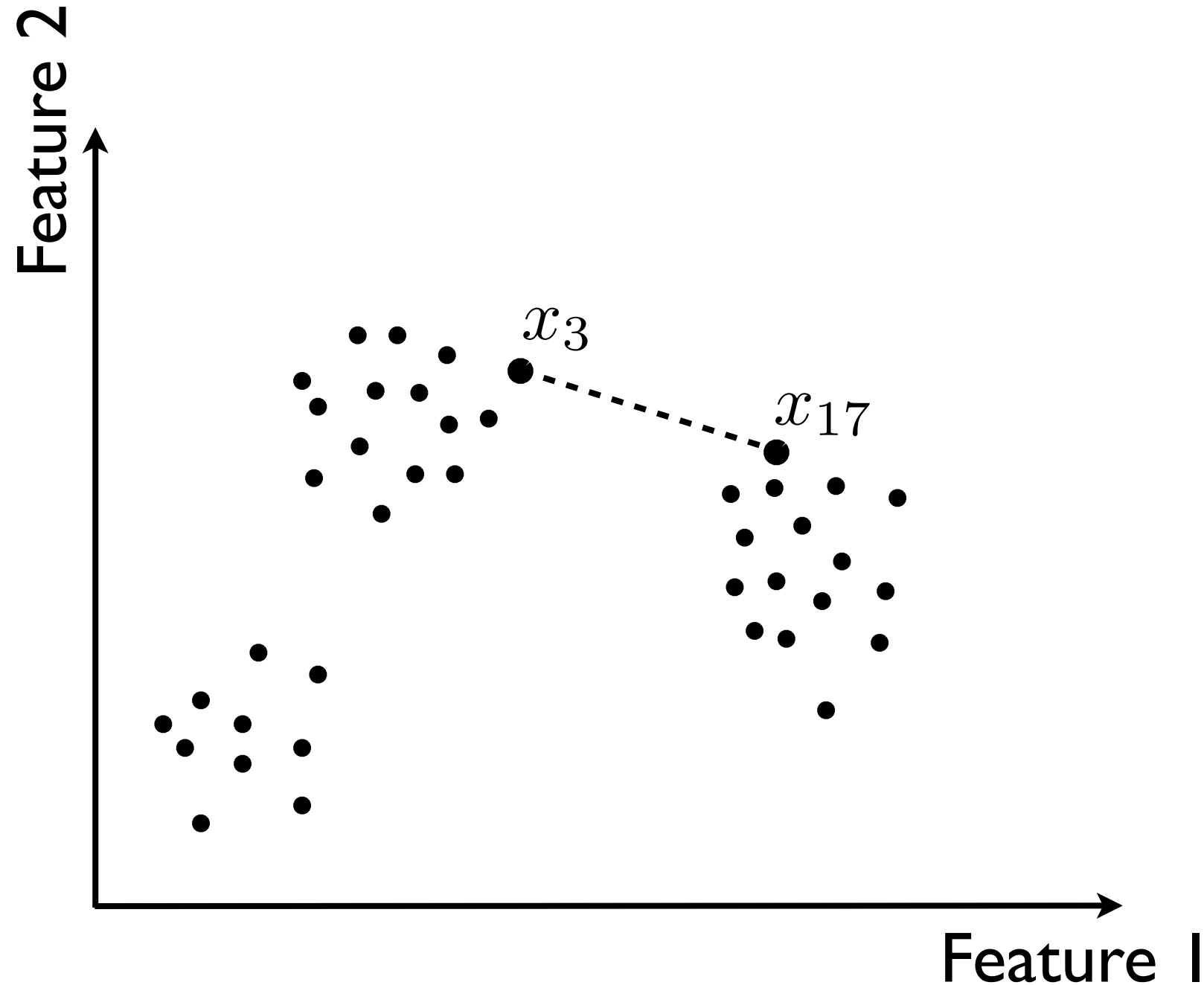
K means: preliminaries

Dissimilarity: Distance as the crow flies



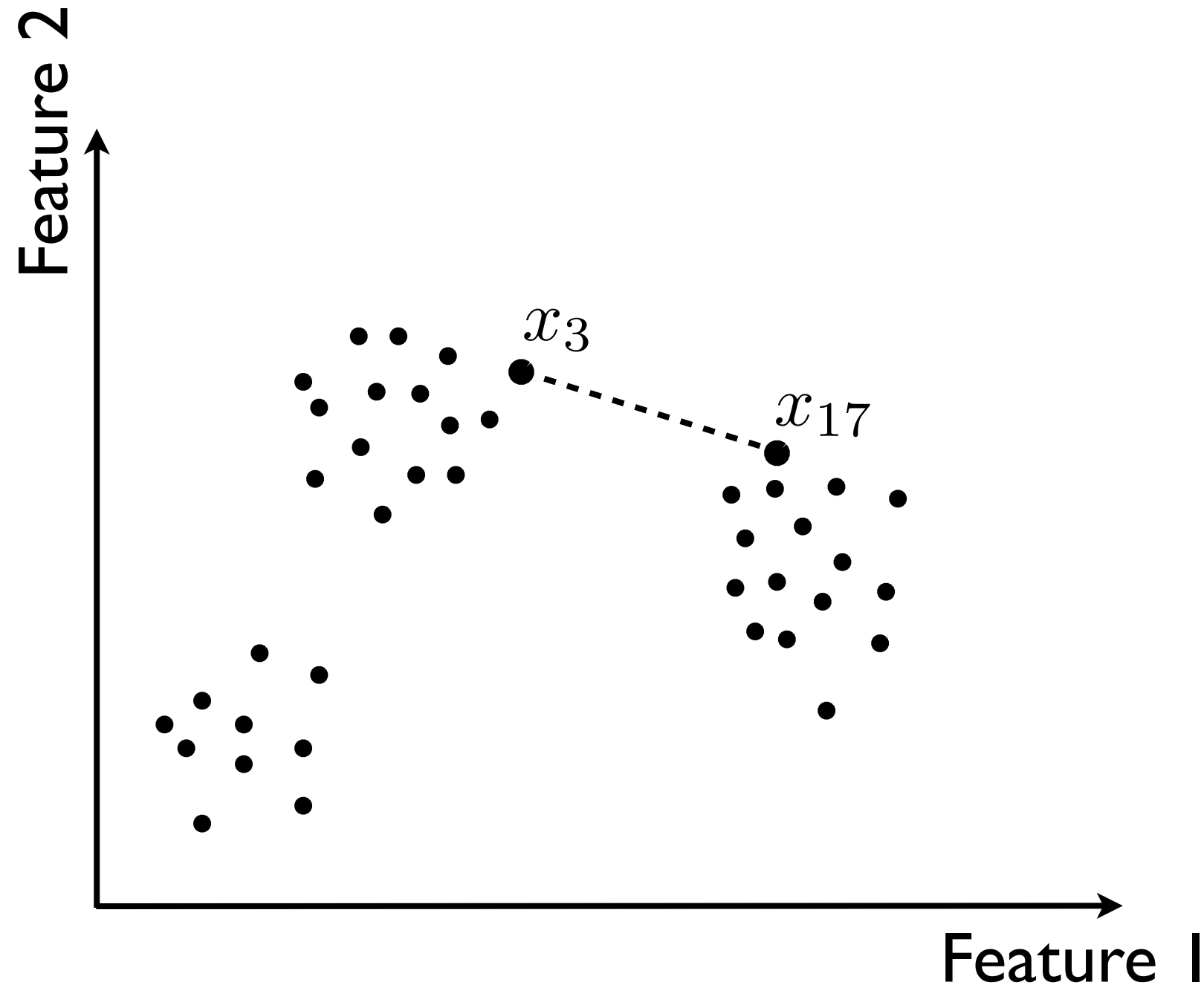
K means: preliminaries

Dissimilarity: Distance as the crow flies



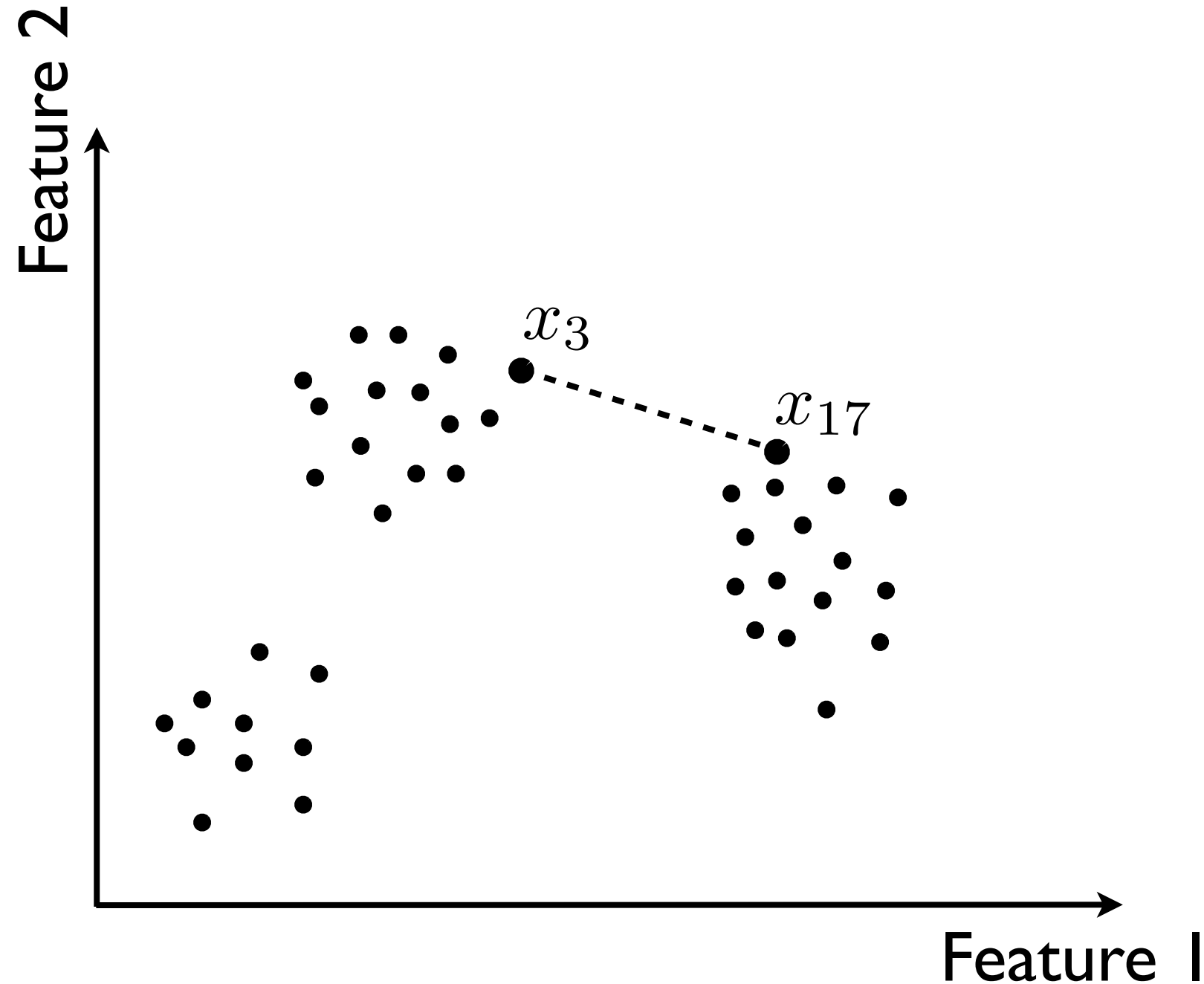
K means: preliminaries

Dissimilarity: Euclidean distance



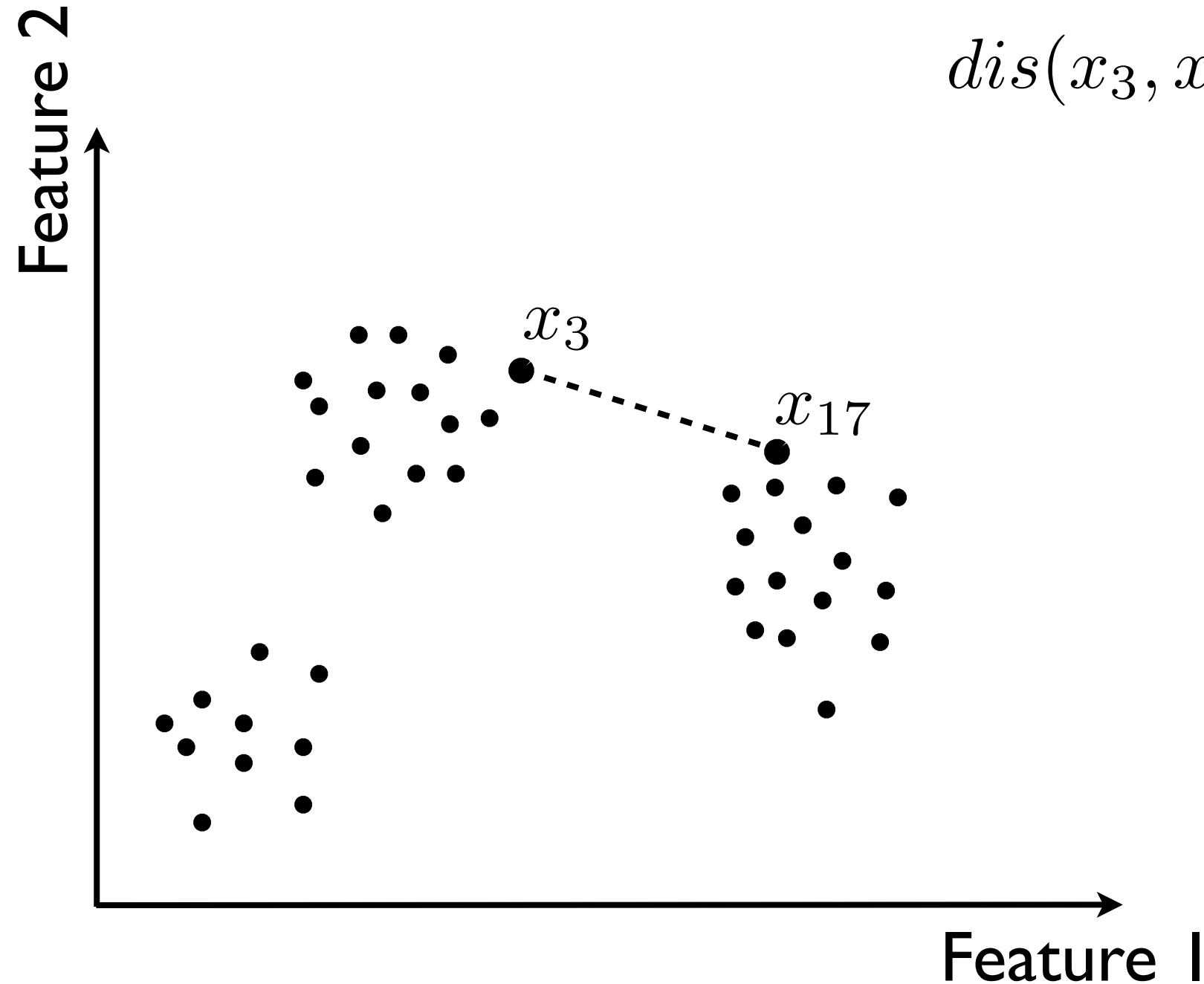
K means: preliminaries

Dissimilarity: Squared Euclidean distance



K means: preliminaries

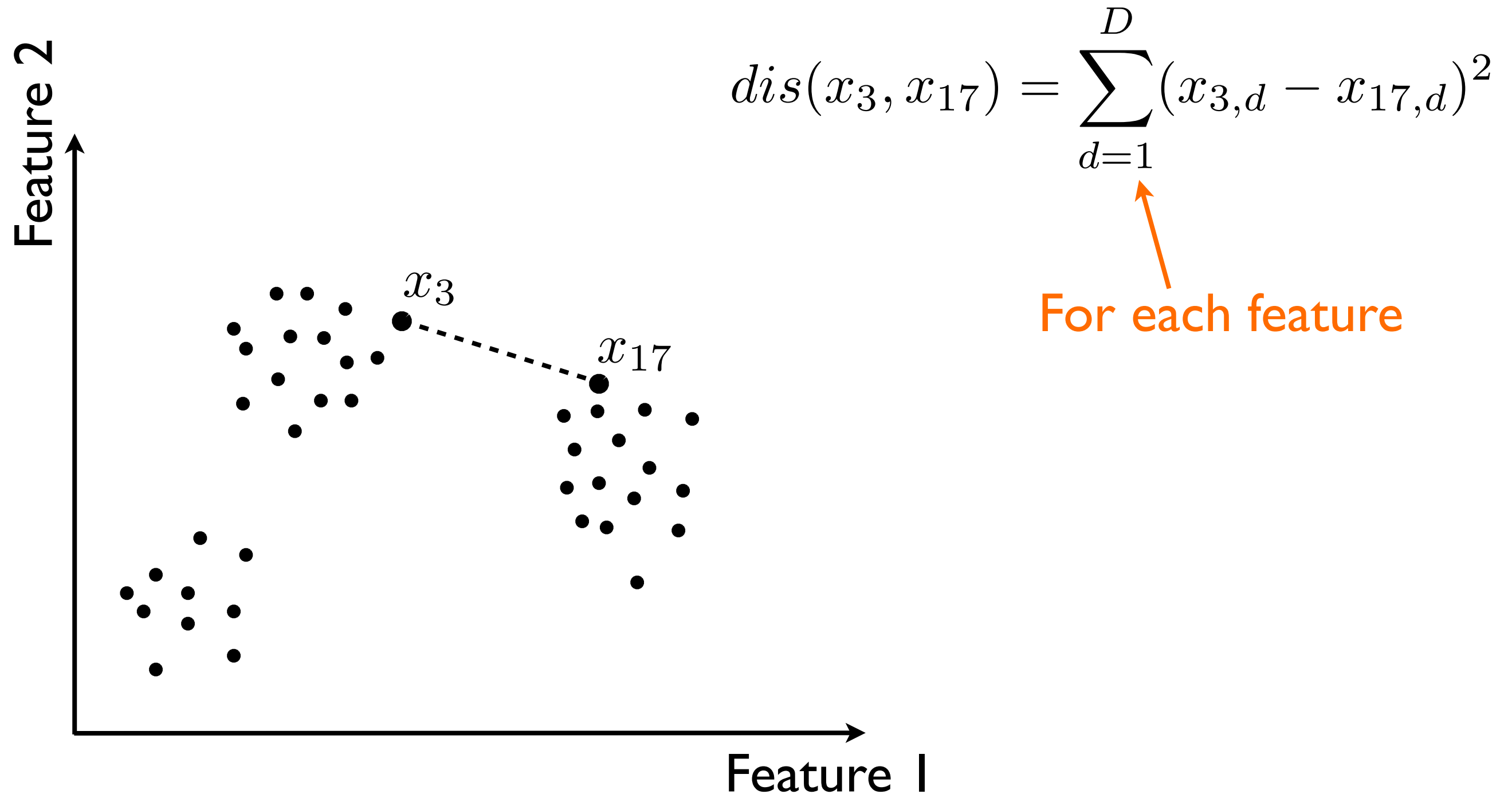
Dissimilarity: Squared Euclidean distance



$$\begin{aligned} dis(x_3, x_{17}) = & (x_{3,1} - x_{17,1})^2 \\ & + (x_{3,2} - x_{17,2})^2 \end{aligned}$$

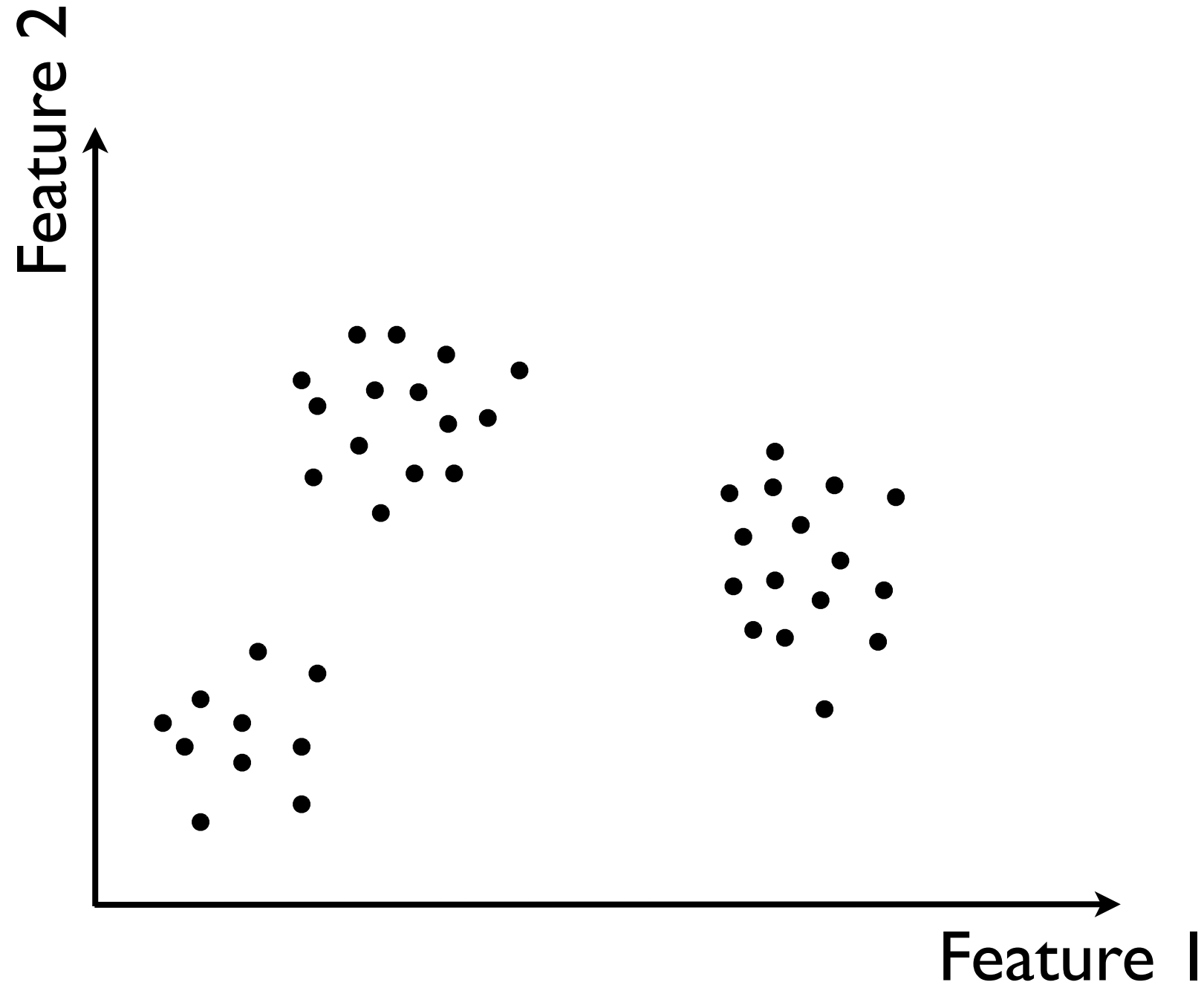
K means: preliminaries

Dissimilarity: Squared Euclidean distance



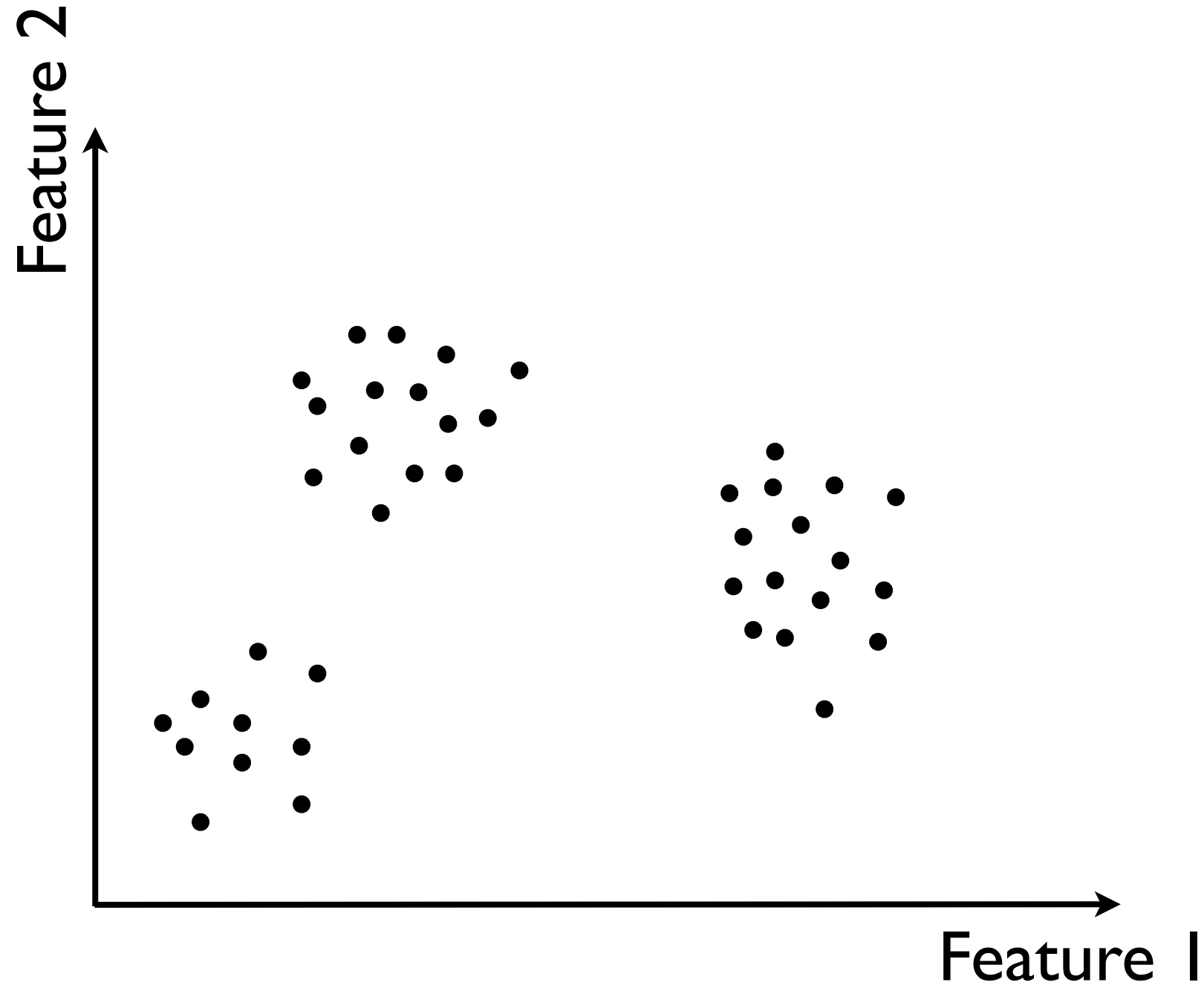
K means: preliminaries

Dissimilarity



K means: preliminaries

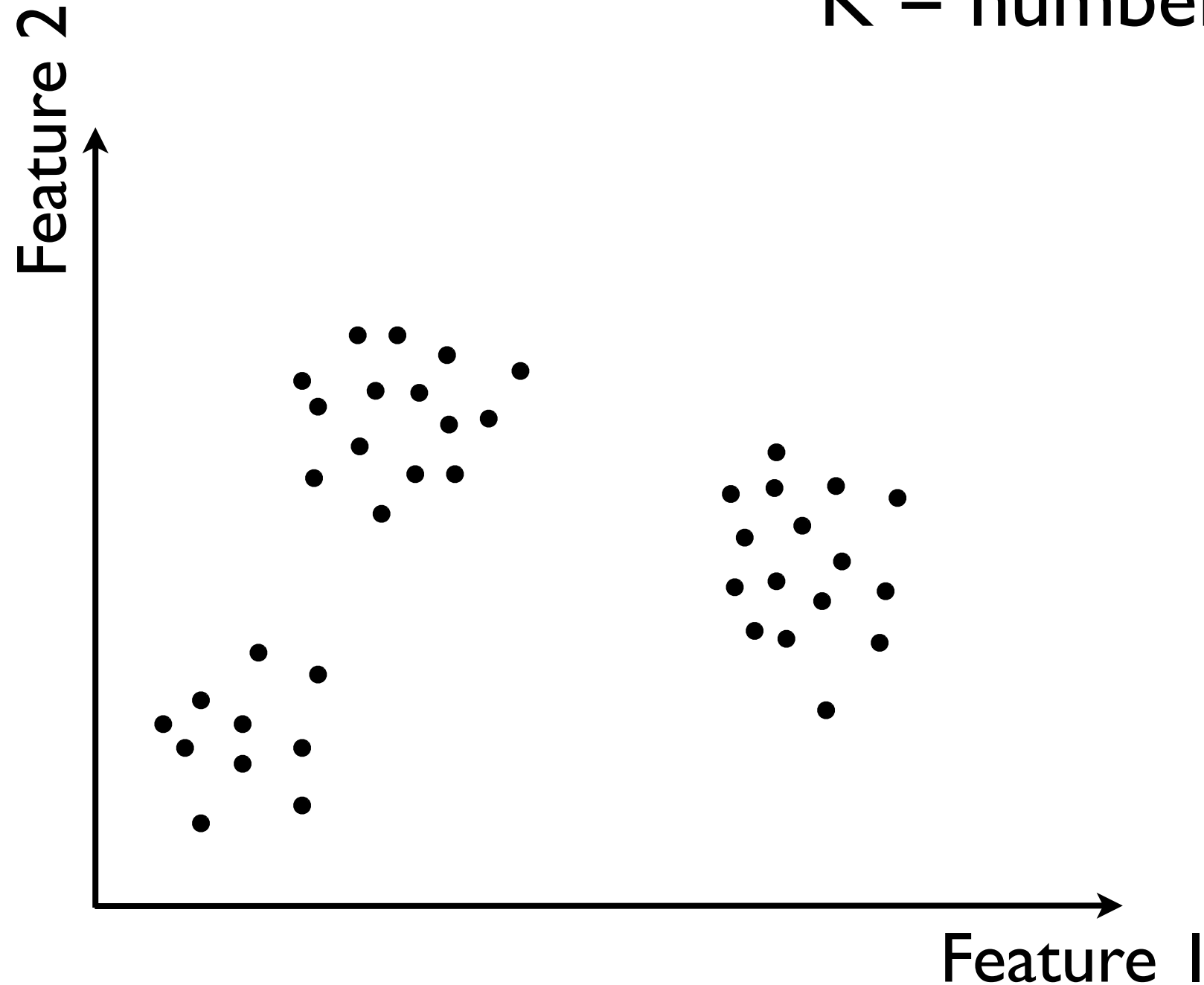
Cluster summary



K means: preliminaries

Cluster summary

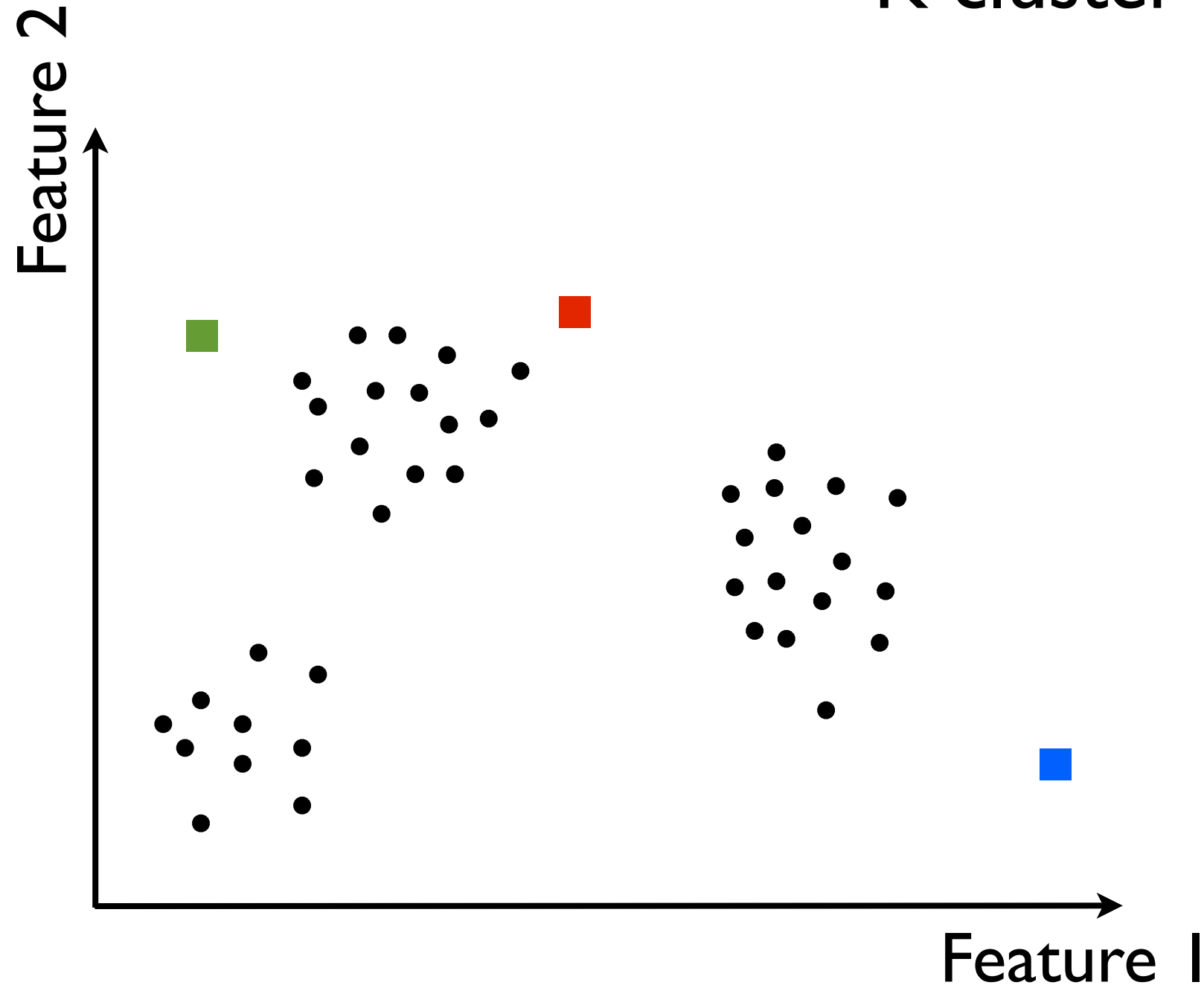
K = number of clusters



K means: preliminaries

Cluster summary

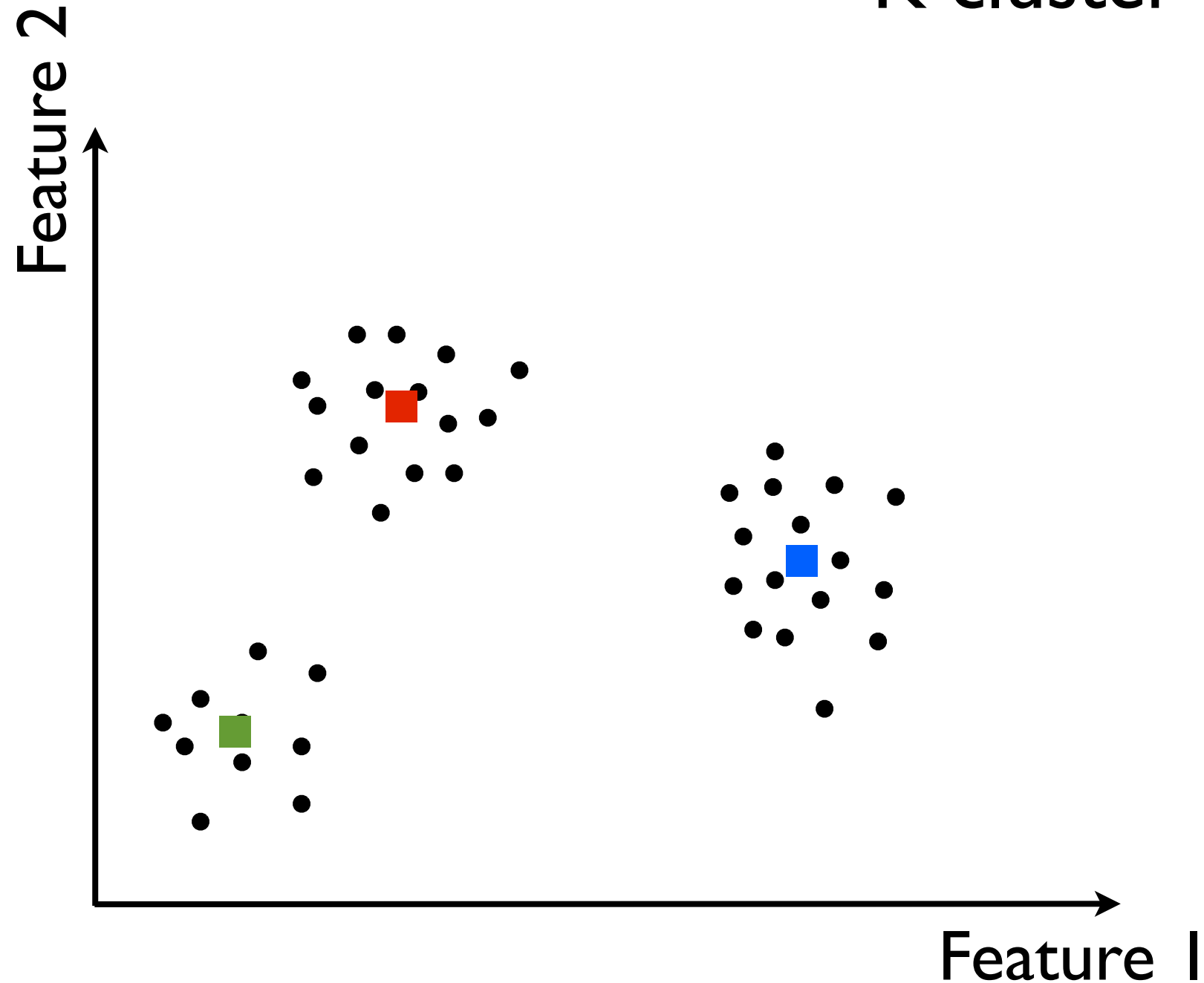
- K cluster centers



K means: preliminaries

Cluster summary

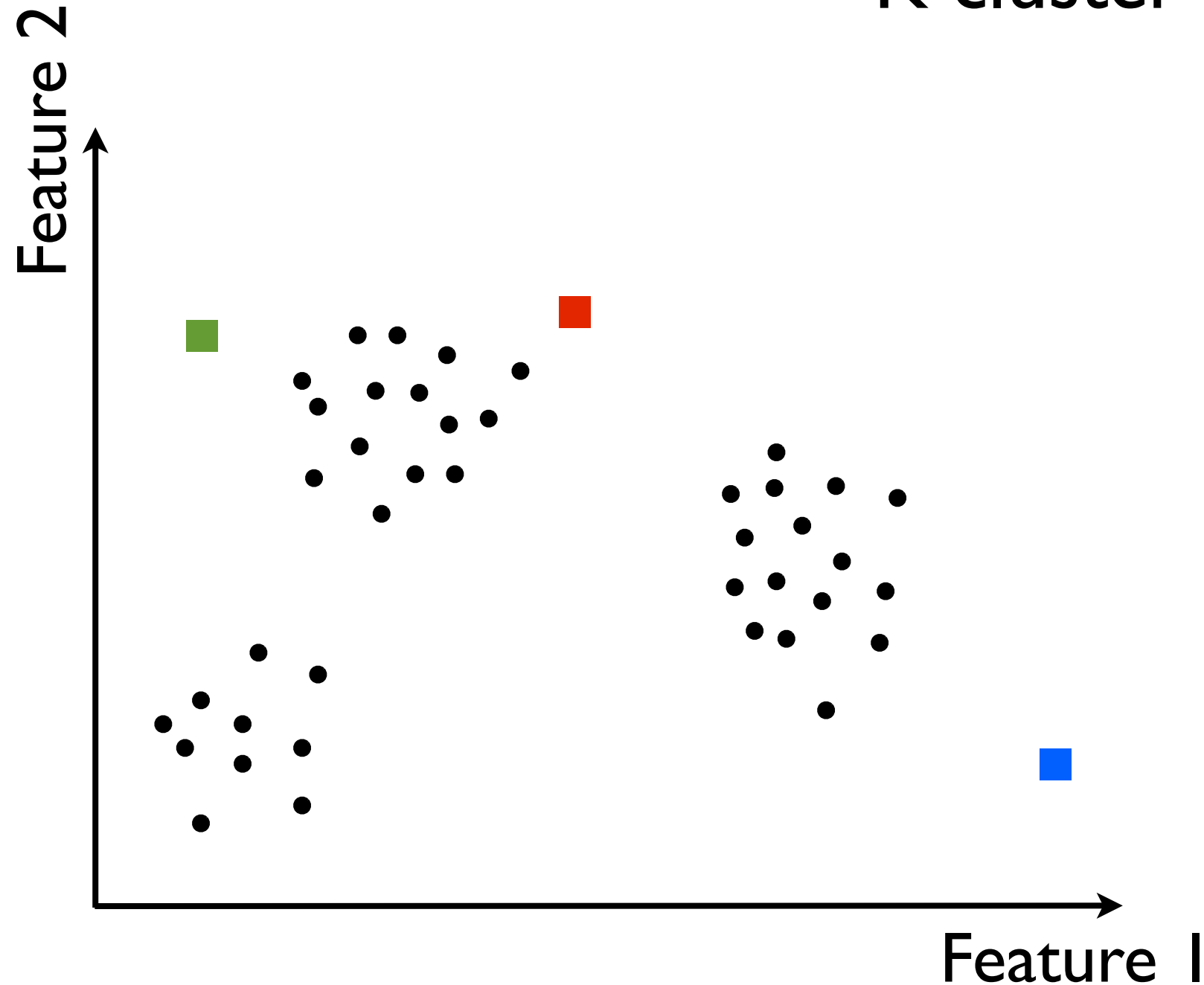
- K cluster centers



K means: preliminaries

Cluster summary

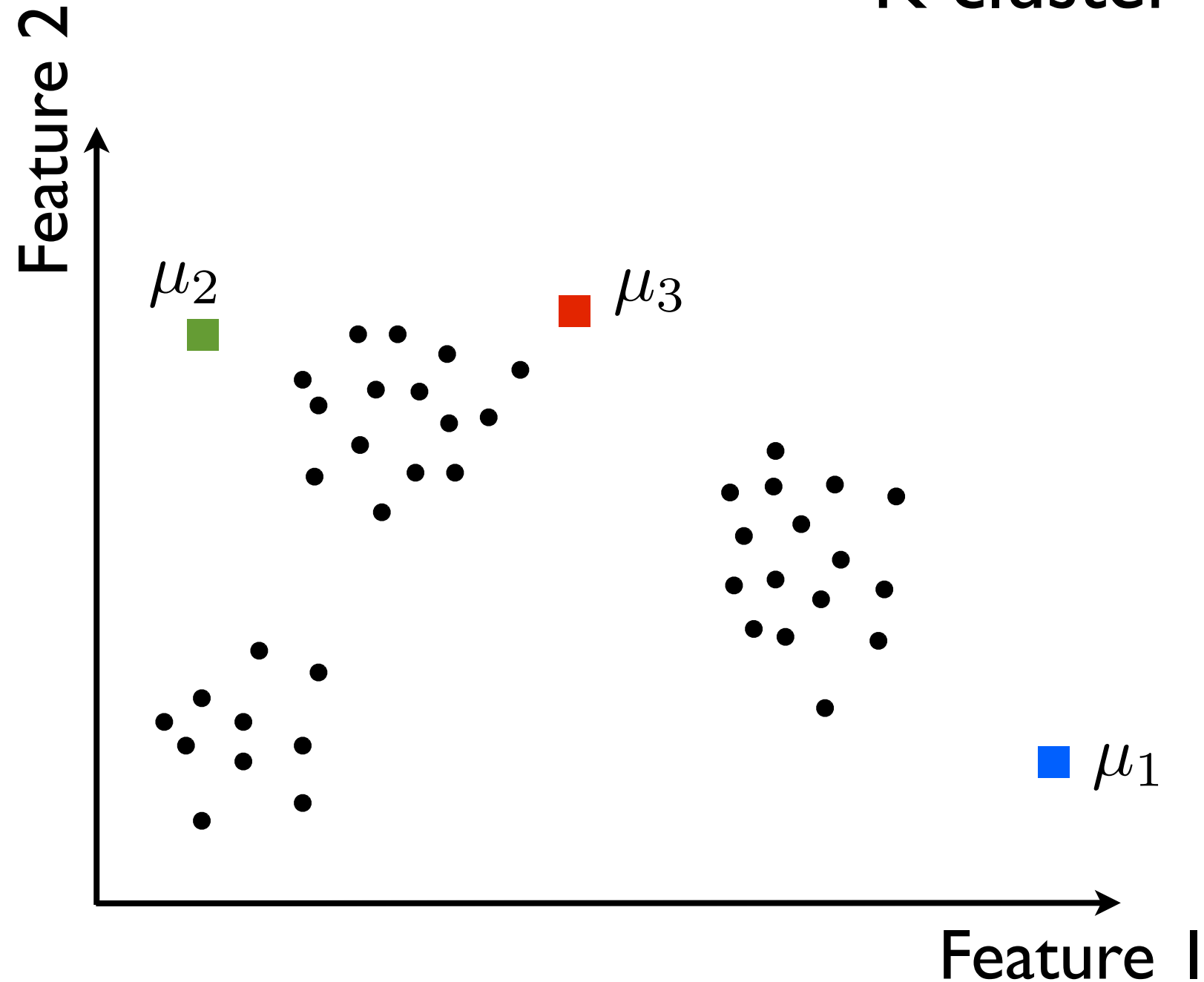
- K cluster centers



K means: preliminaries

Cluster summary

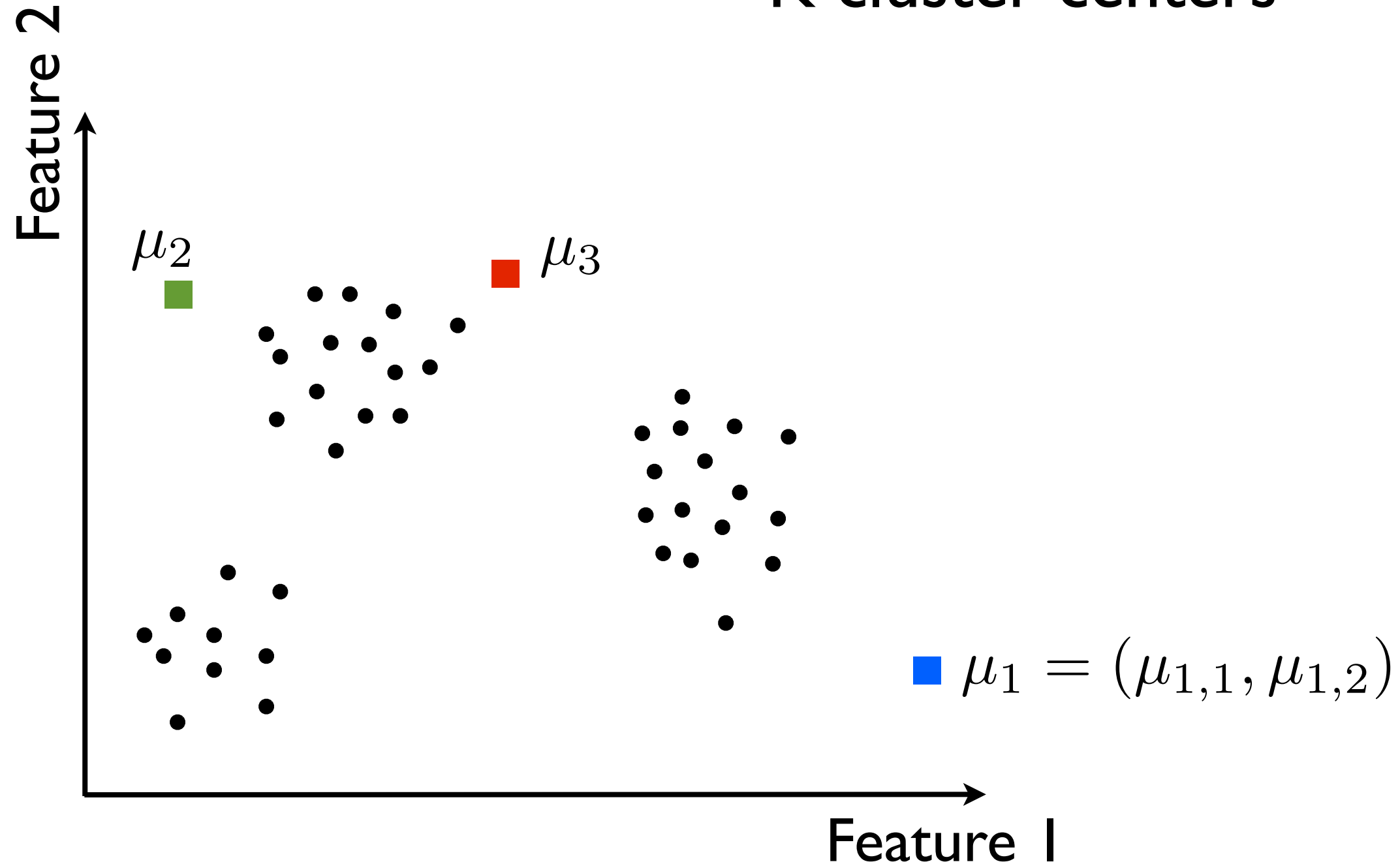
- K cluster centers



K means: preliminaries

Cluster summary

- K cluster centers

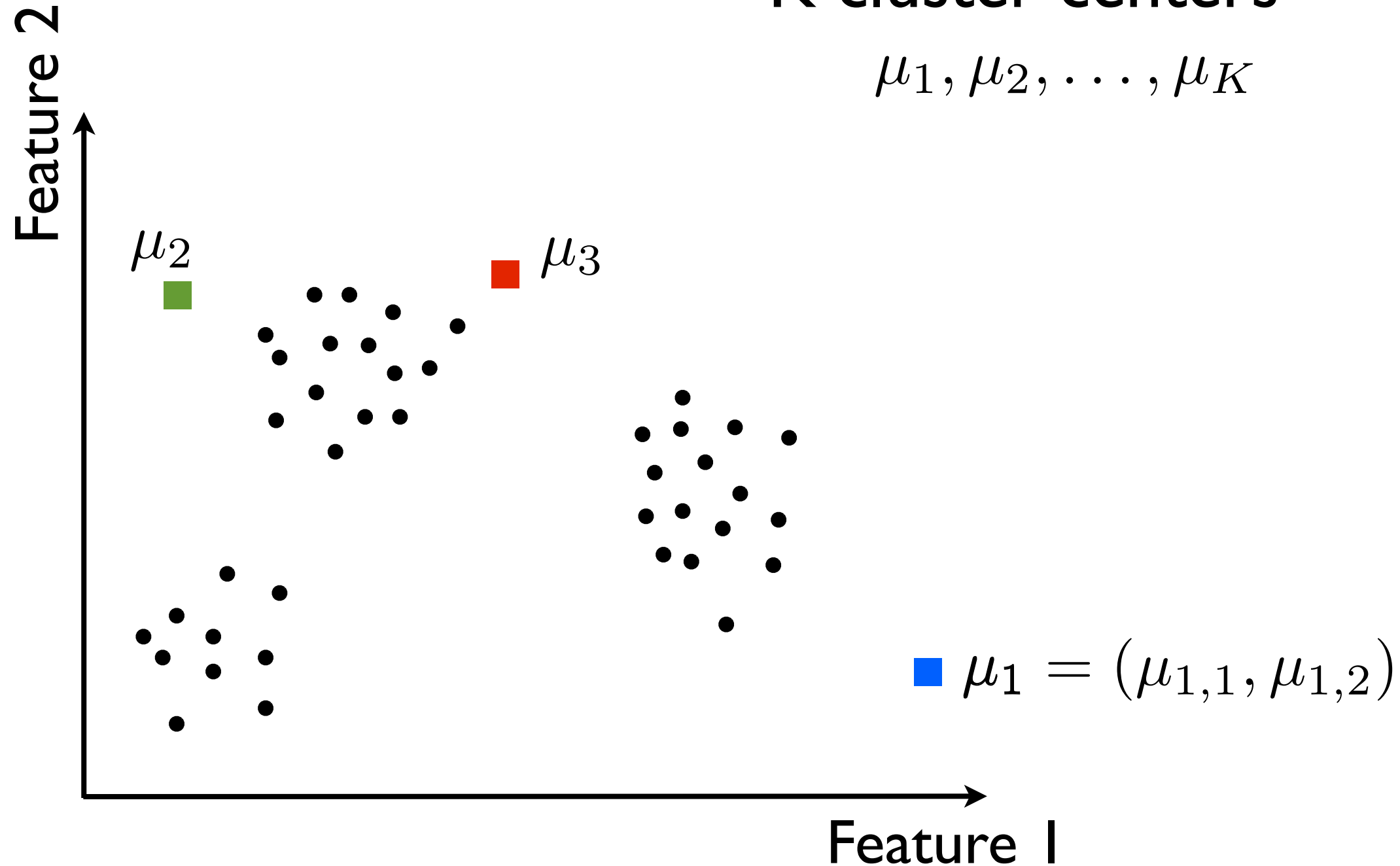


K means: preliminaries

Cluster summary

- K cluster centers

$$\mu_1, \mu_2, \dots, \mu_K$$



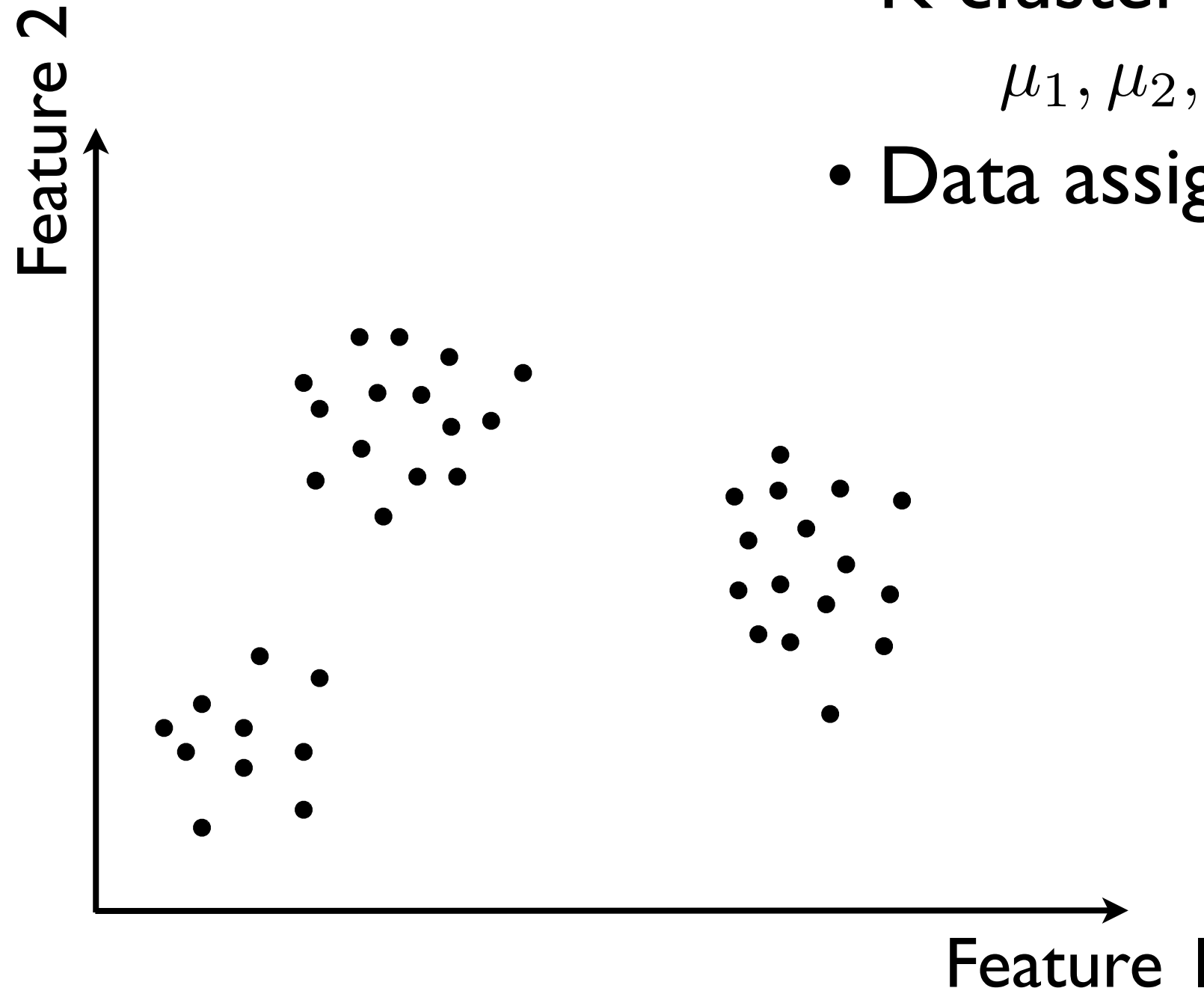
K means: preliminaries

Cluster summary

- K cluster centers

$$\mu_1, \mu_2, \dots, \mu_K$$

- Data assignments to clusters



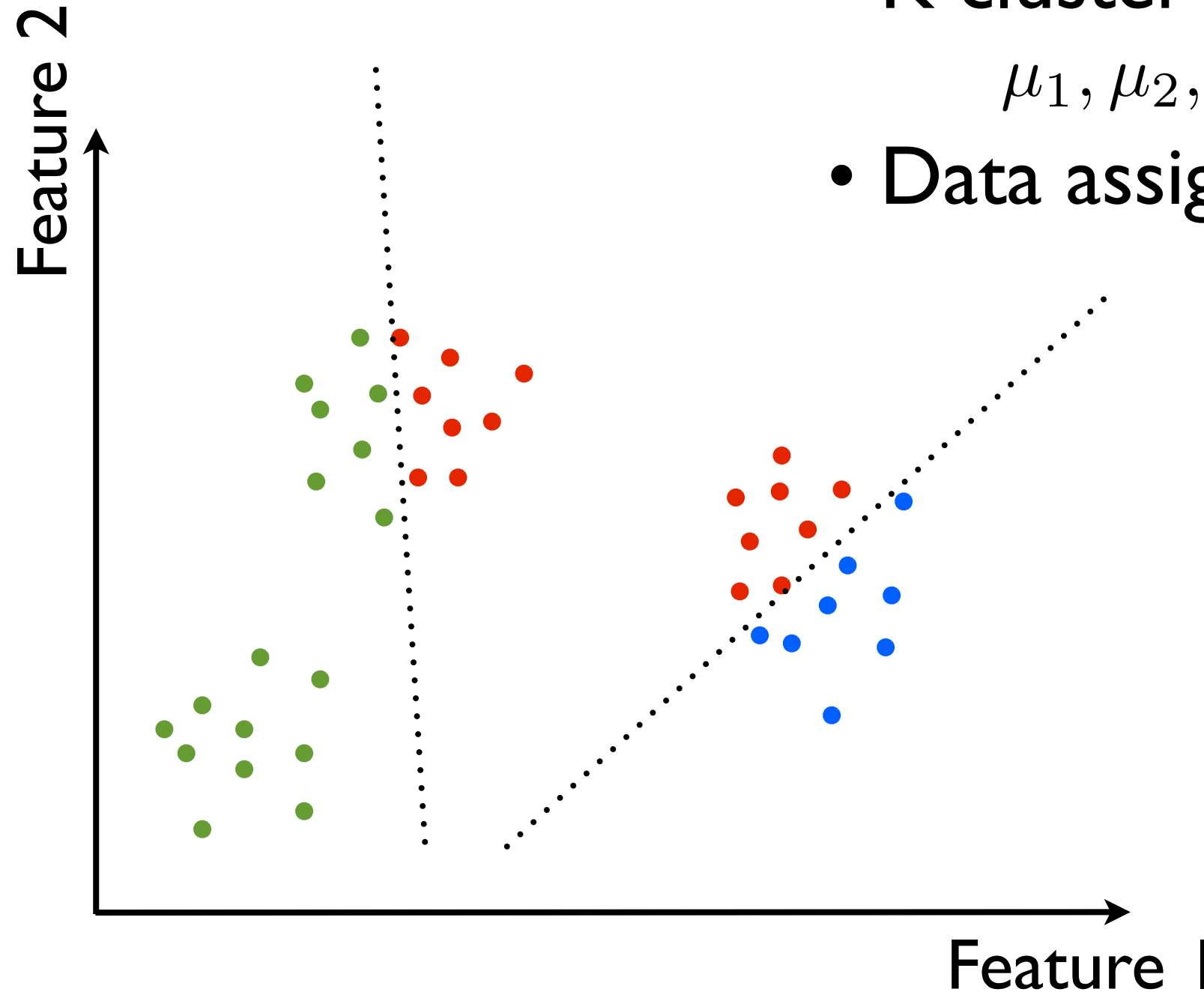
K means: preliminaries

Cluster summary

- K cluster centers

$$\mu_1, \mu_2, \dots, \mu_K$$

- Data assignments to clusters



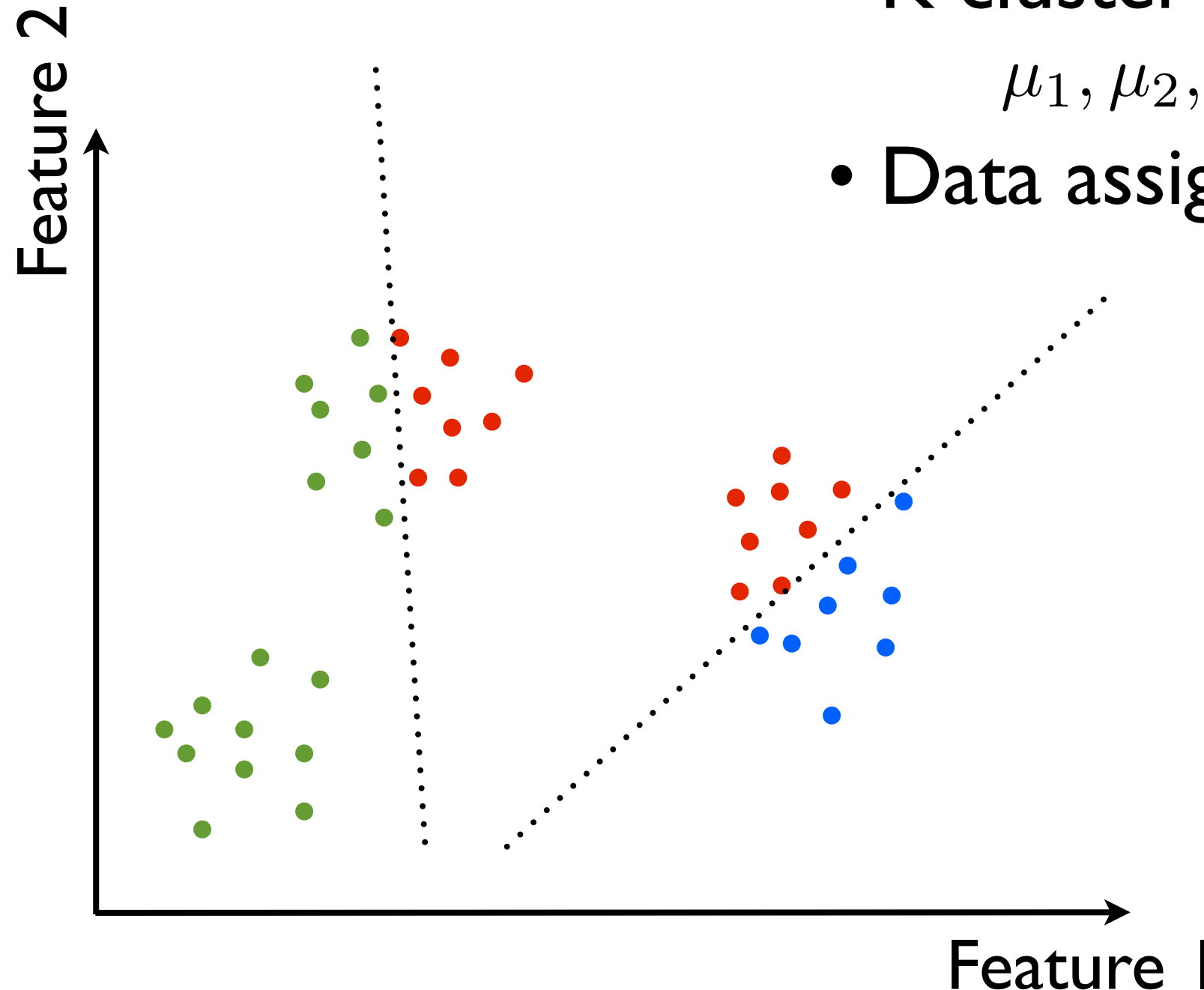
K means: preliminaries

Cluster summary

- K cluster centers

$$\mu_1, \mu_2, \dots, \mu_K$$

- Data assignments to clusters



S_k = set of points in
cluster k

K means: preliminaries

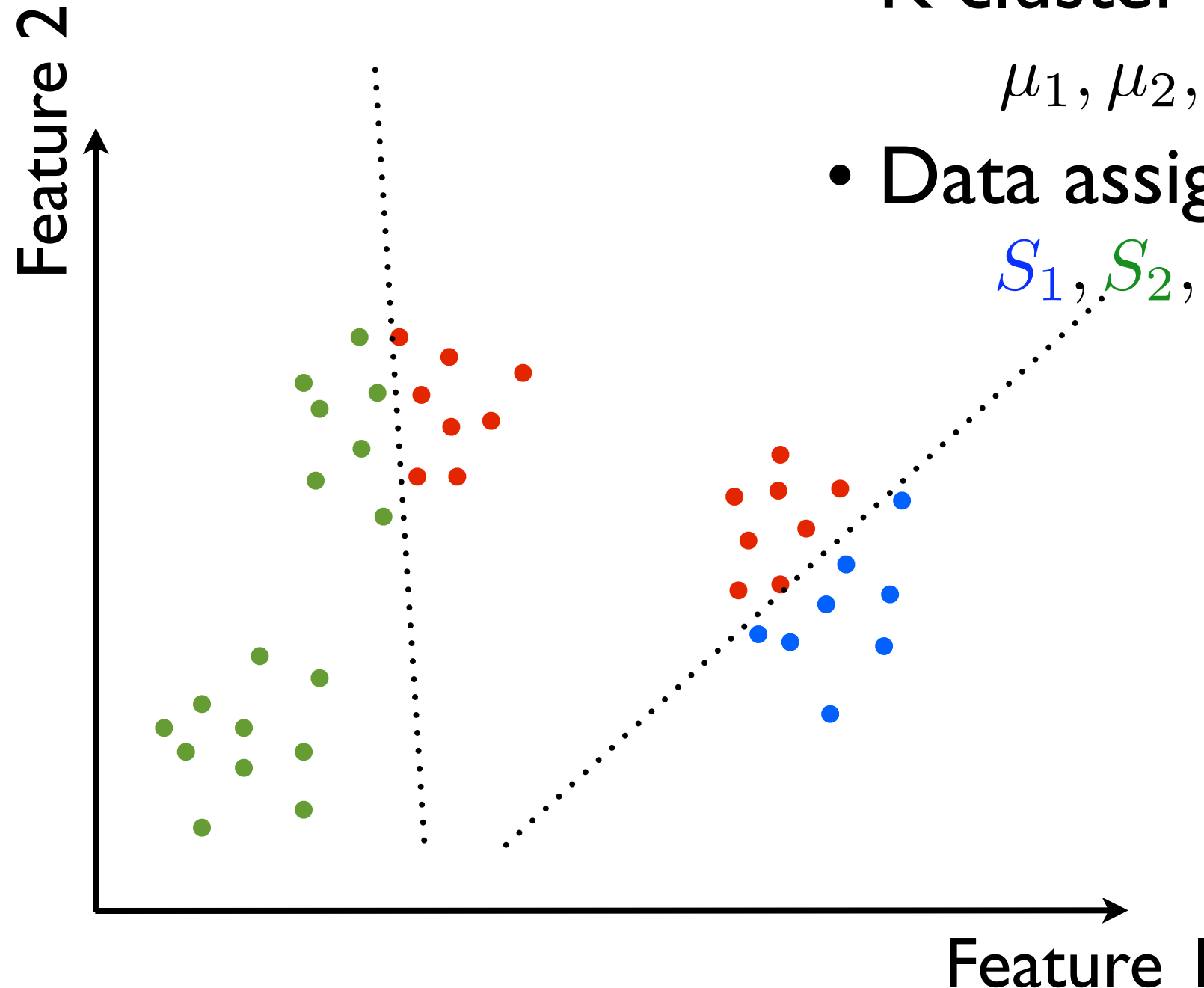
Cluster summary

- K cluster centers

$$\mu_1, \mu_2, \dots, \mu_K$$

- Data assignments to clusters

$$S_1, S_2, \dots, S_K$$



S_k = set of points in
cluster k

K means: preliminaries

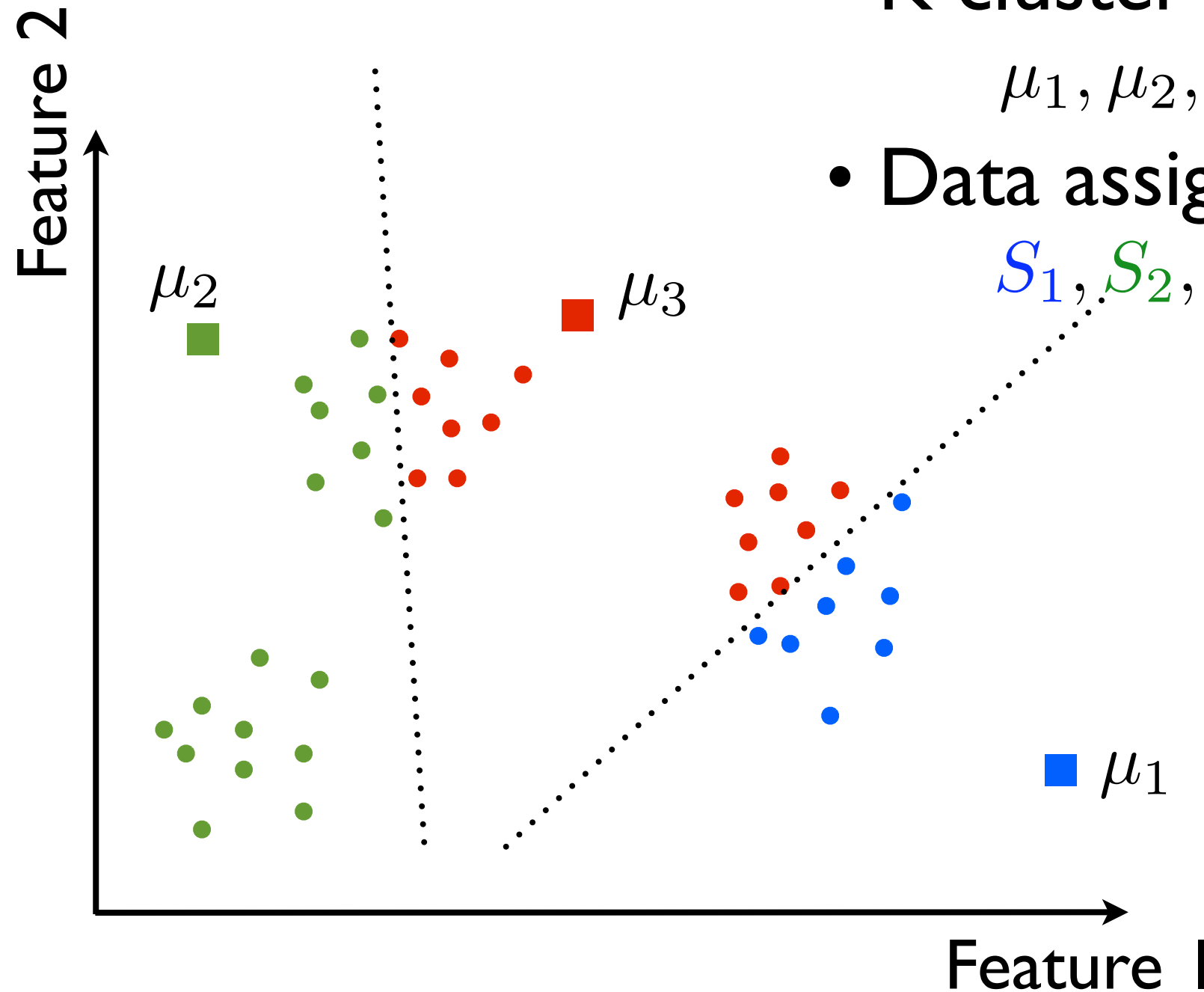
Cluster summary

- K cluster centers

$$\mu_1, \mu_2, \dots, \mu_K$$

- Data assignments to clusters

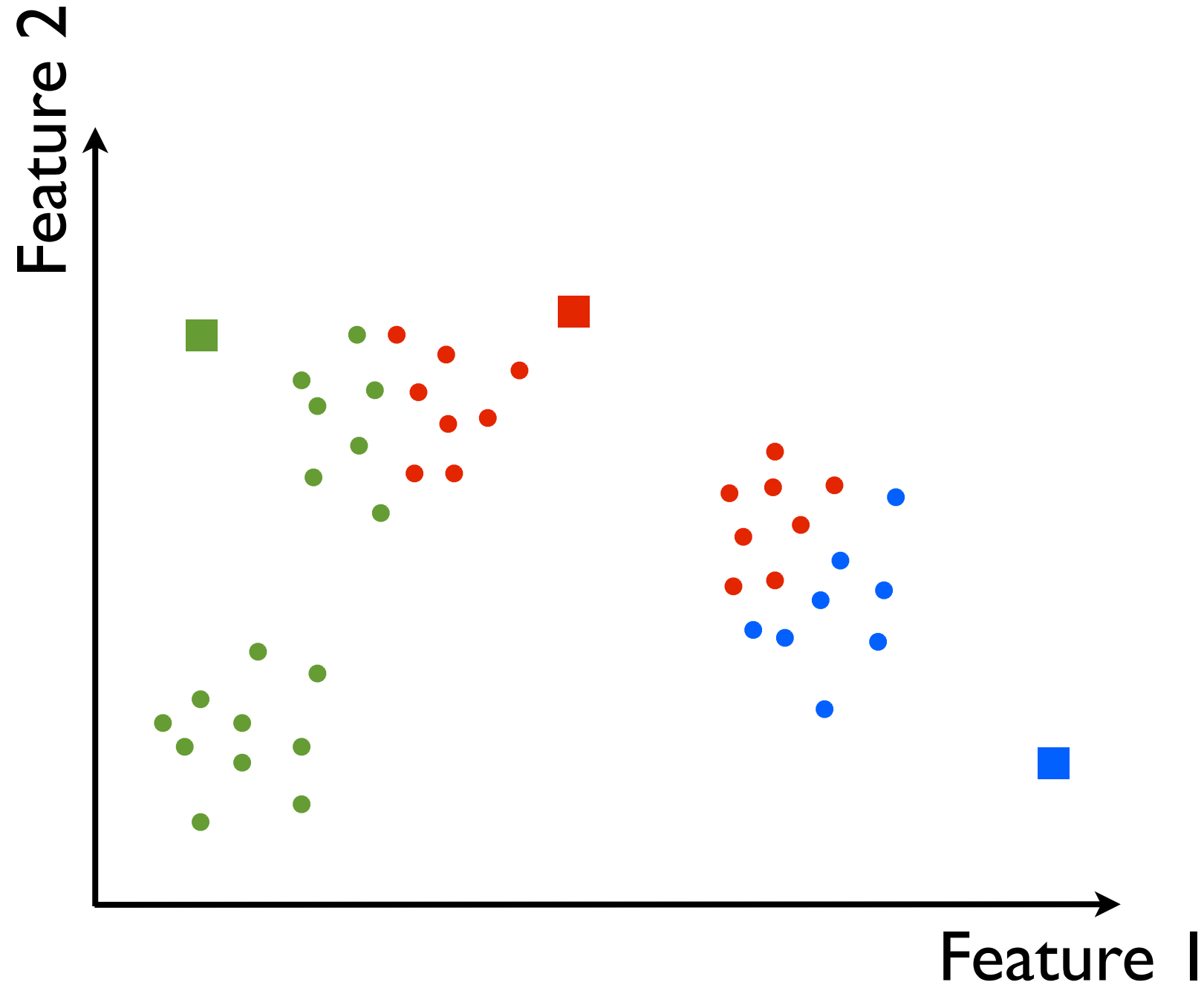
$$S_1, S_2, \dots, S_K$$



S_k = set of points in
cluster k

K means: preliminaries

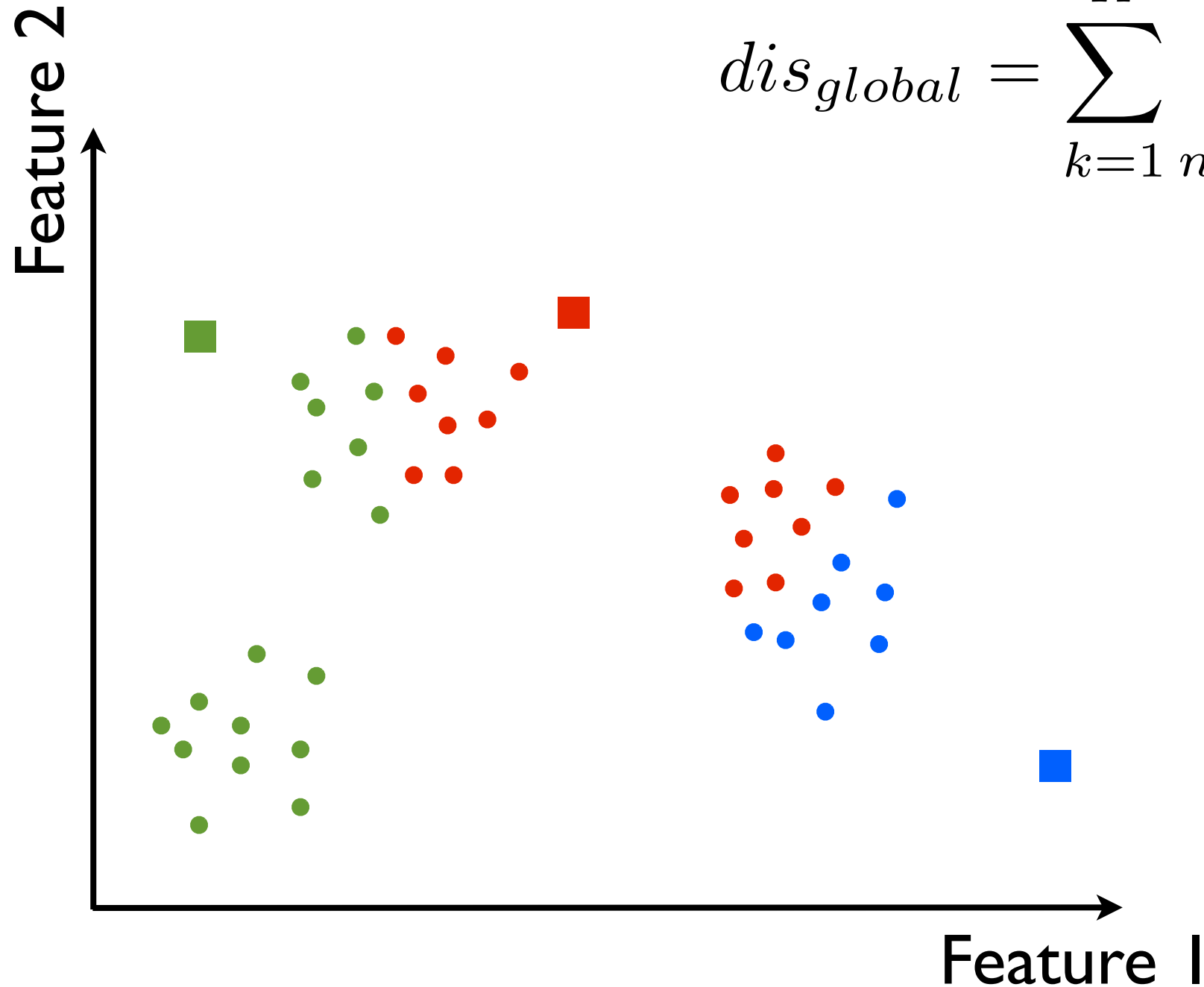
Dissimilarity



K means: preliminaries

Dissimilarity (global)

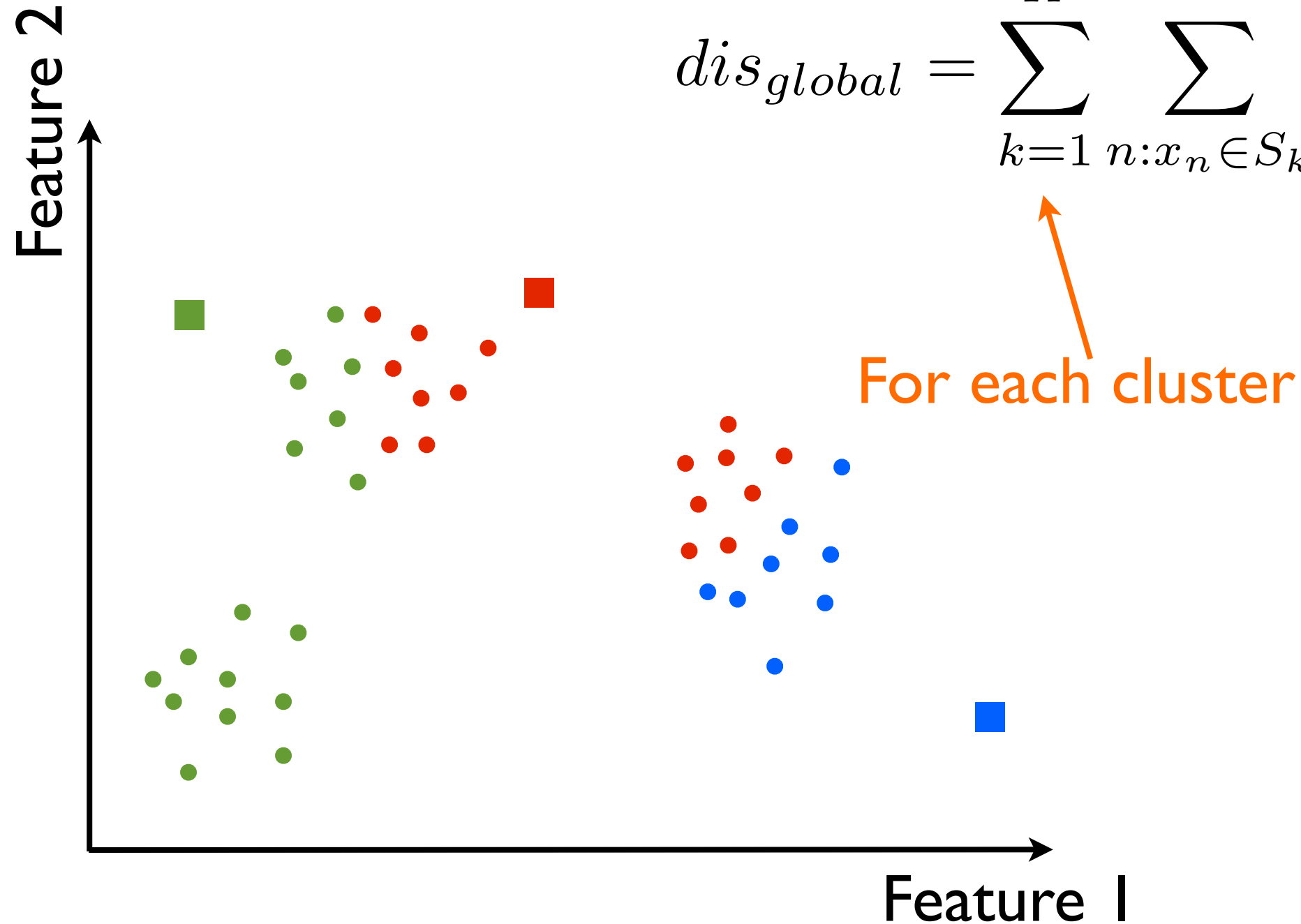
$$dis_{global} = \sum_{k=1}^K \sum_{n: x_n \in S_k} \sum_{d=1}^D (x_{n,d} - \mu_{k,d})^2$$



K means: preliminaries

Dissimilarity (global)

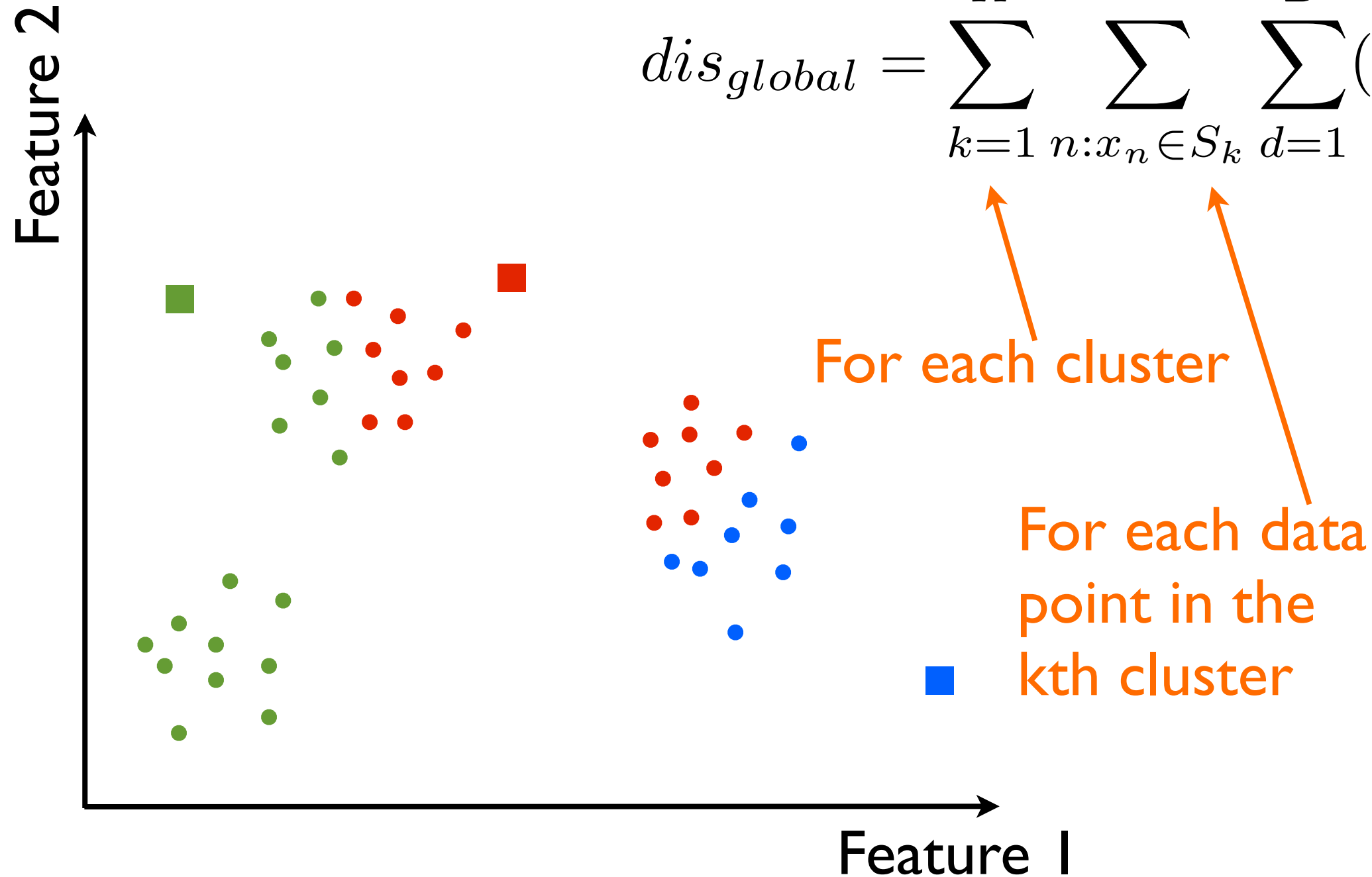
$$dis_{global} = \sum_{k=1}^K \sum_{n: x_n \in S_k} \sum_{d=1}^D (x_{n,d} - \mu_{k,d})^2$$



K means: preliminaries

Dissimilarity (global)

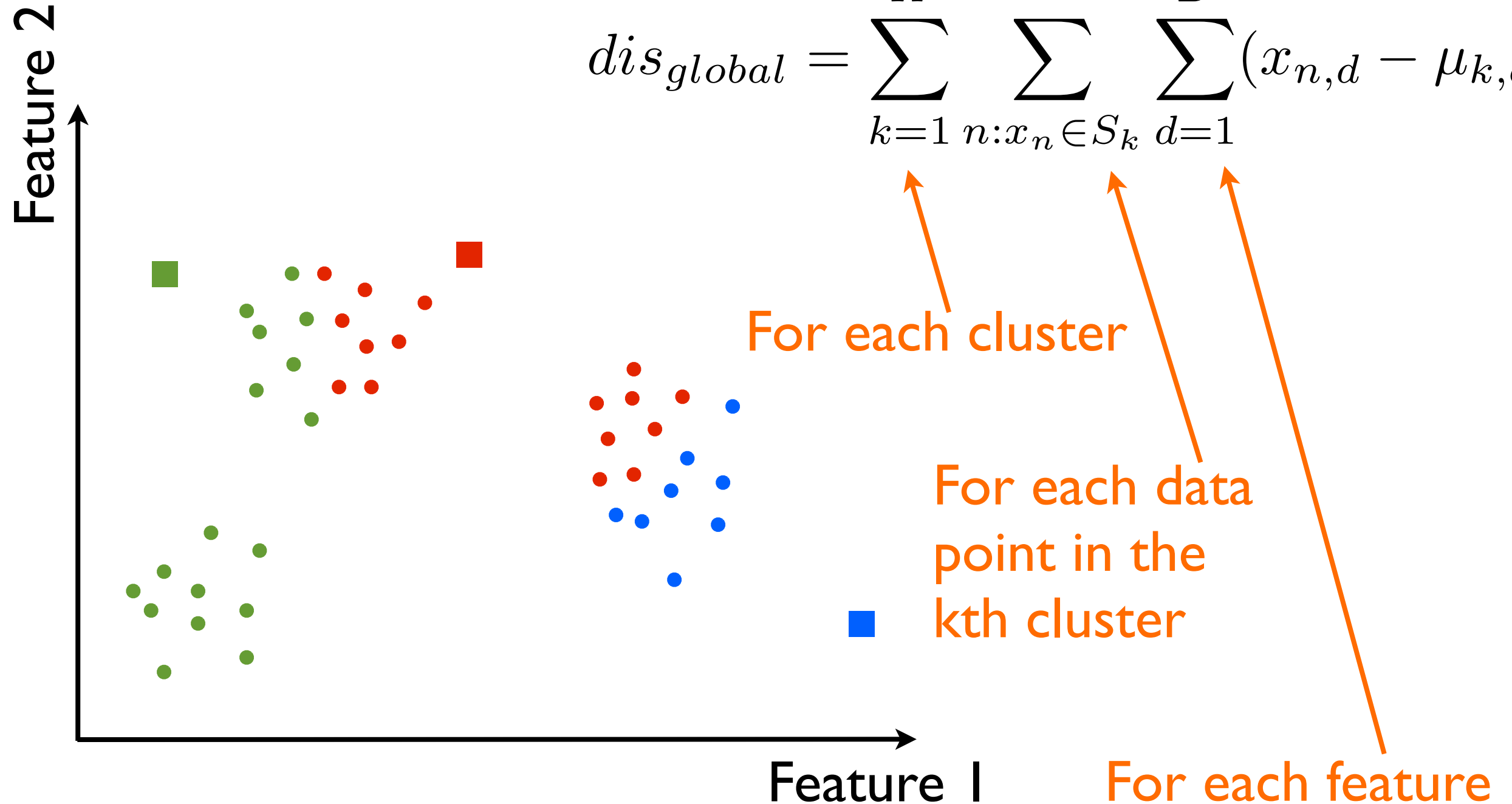
$$dis_{global} = \sum_{k=1}^K \sum_{n: x_n \in S_k} \sum_{d=1}^D (x_{n,d} - \mu_{k,d})^2$$



K means: preliminaries

Dissimilarity (global)

$$dis_{global} = \sum_{k=1}^K \sum_{n: x_n \in S_k} \sum_{d=1}^D (x_{n,d} - \mu_{k,d})^2$$

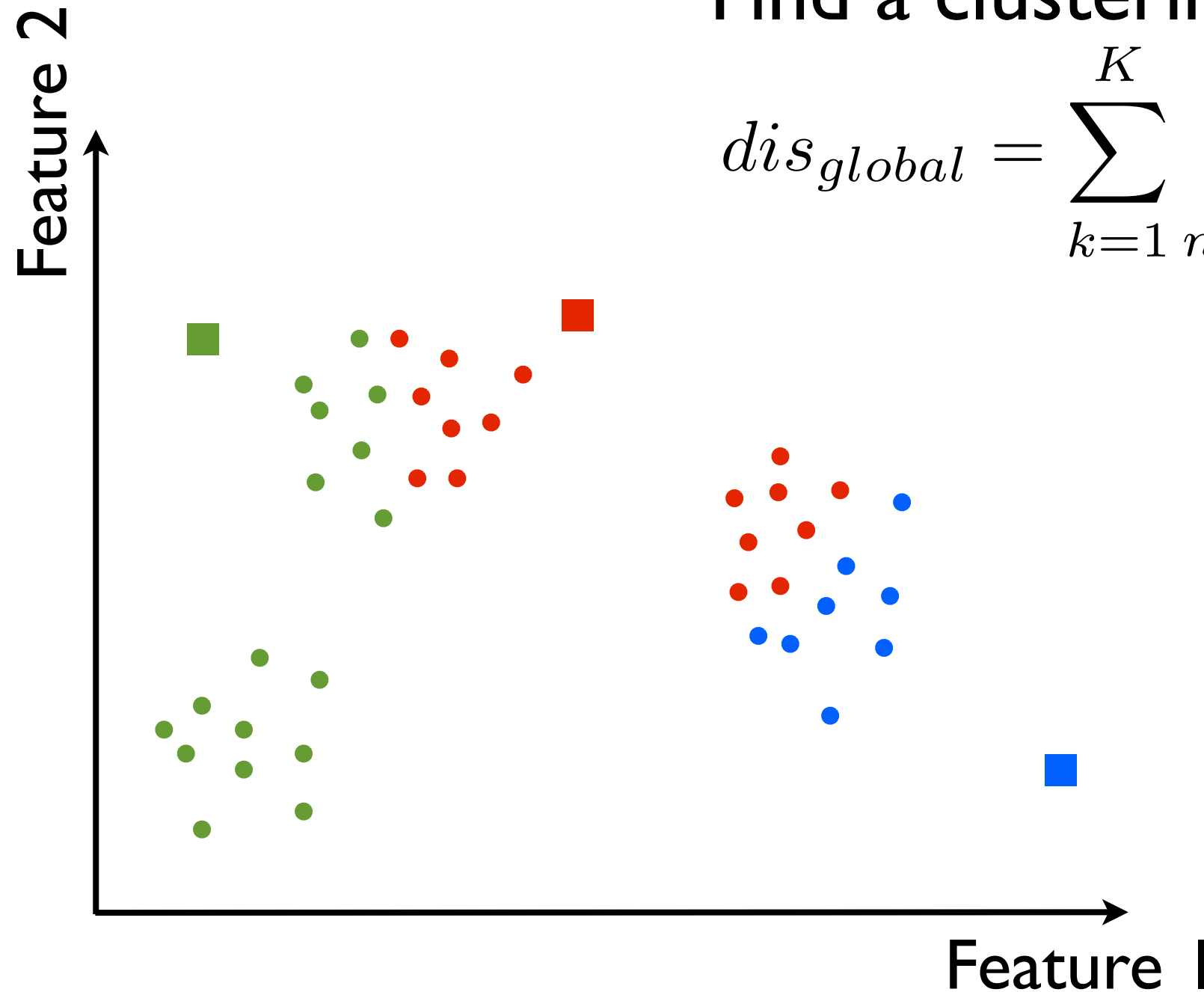


K means: preliminaries

K means objective:

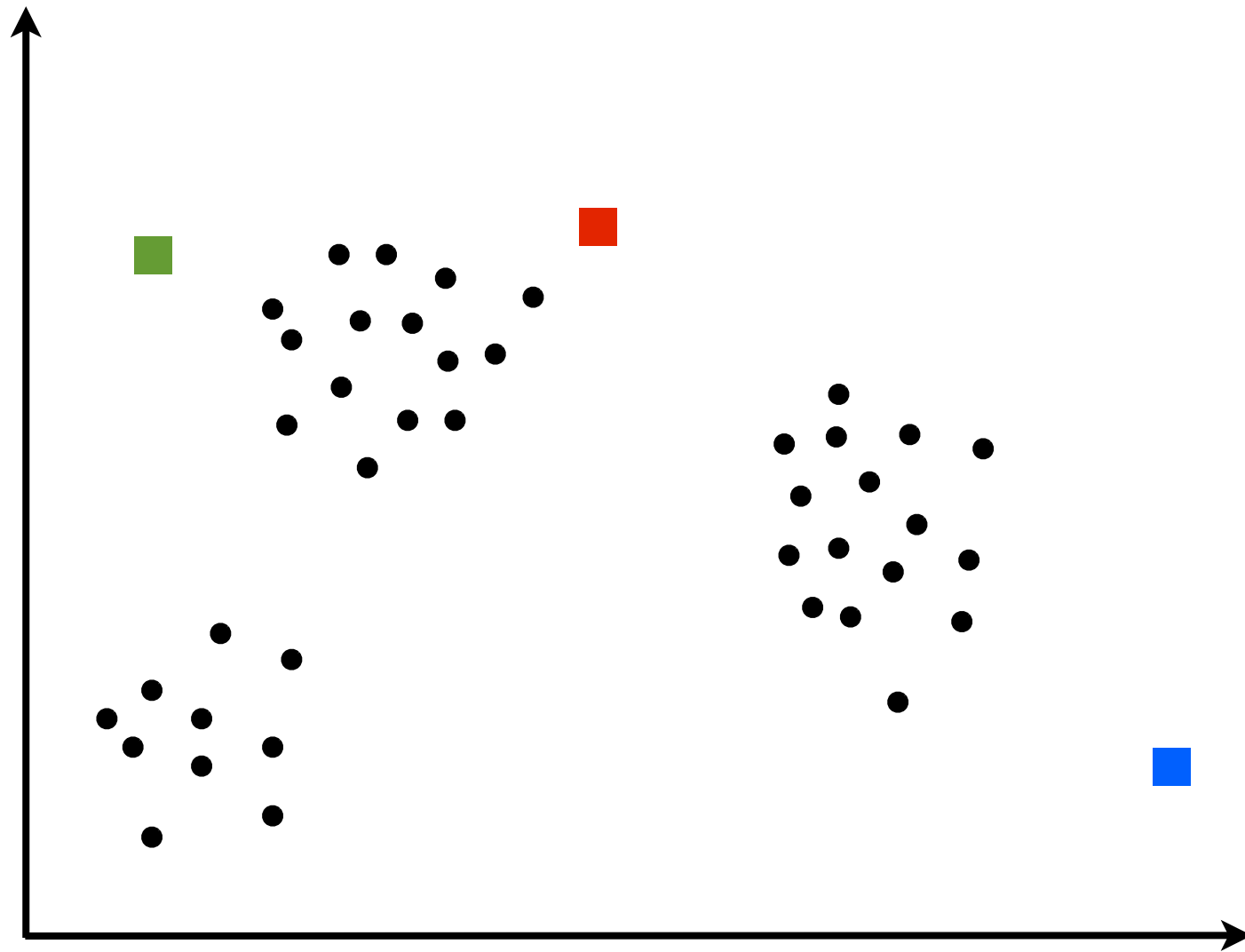
Find a clustering to minimize

$$dis_{global} = \sum_{k=1}^K \sum_{n:x_n \in S_k} \sum_{d=1}^D (x_{n,d} - \mu_{k,d})^2$$



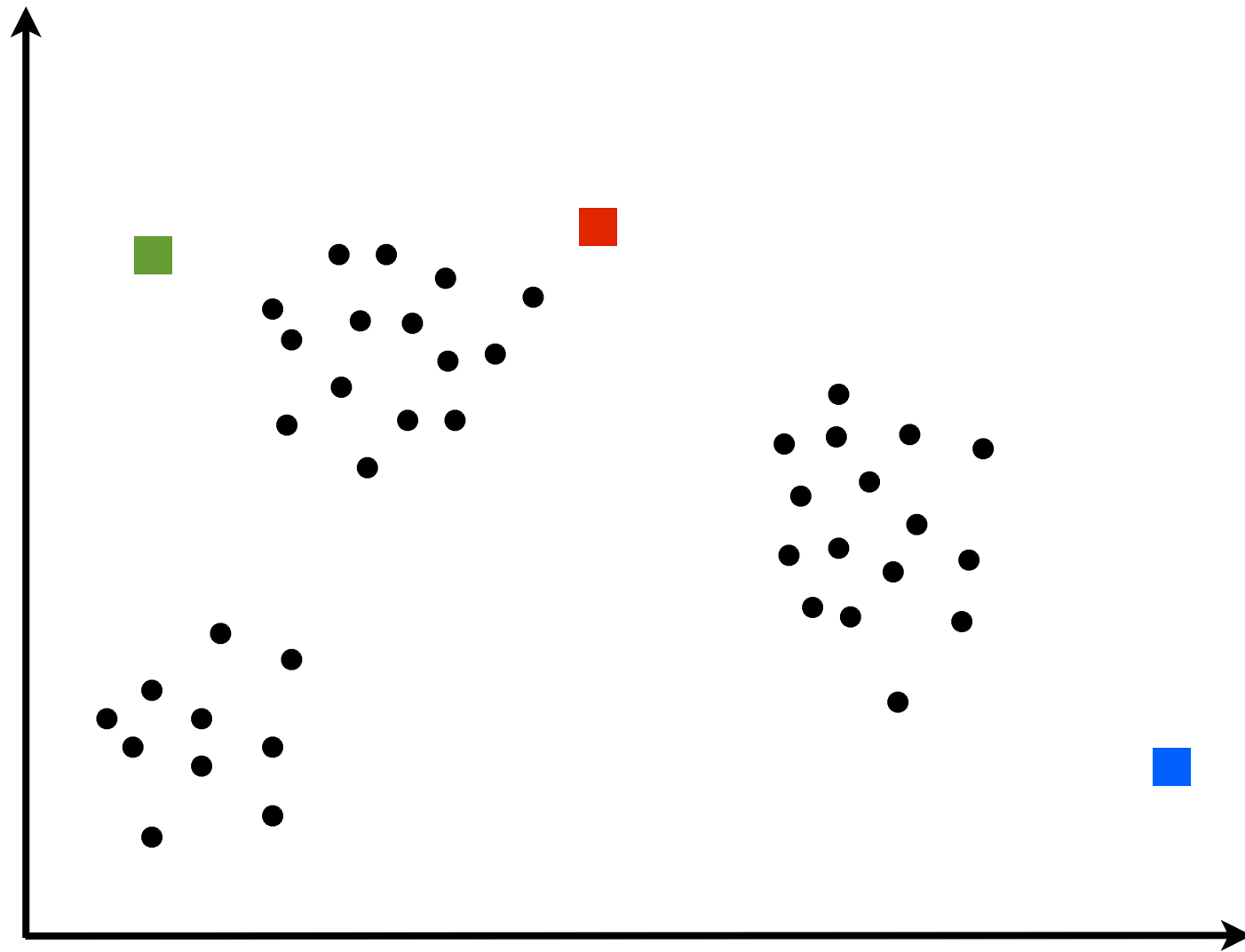
K means algorithm

- Initialize K cluster centers
- Repeat until convergence:
 - ◇ Assign each data point to the cluster with the closest center.
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



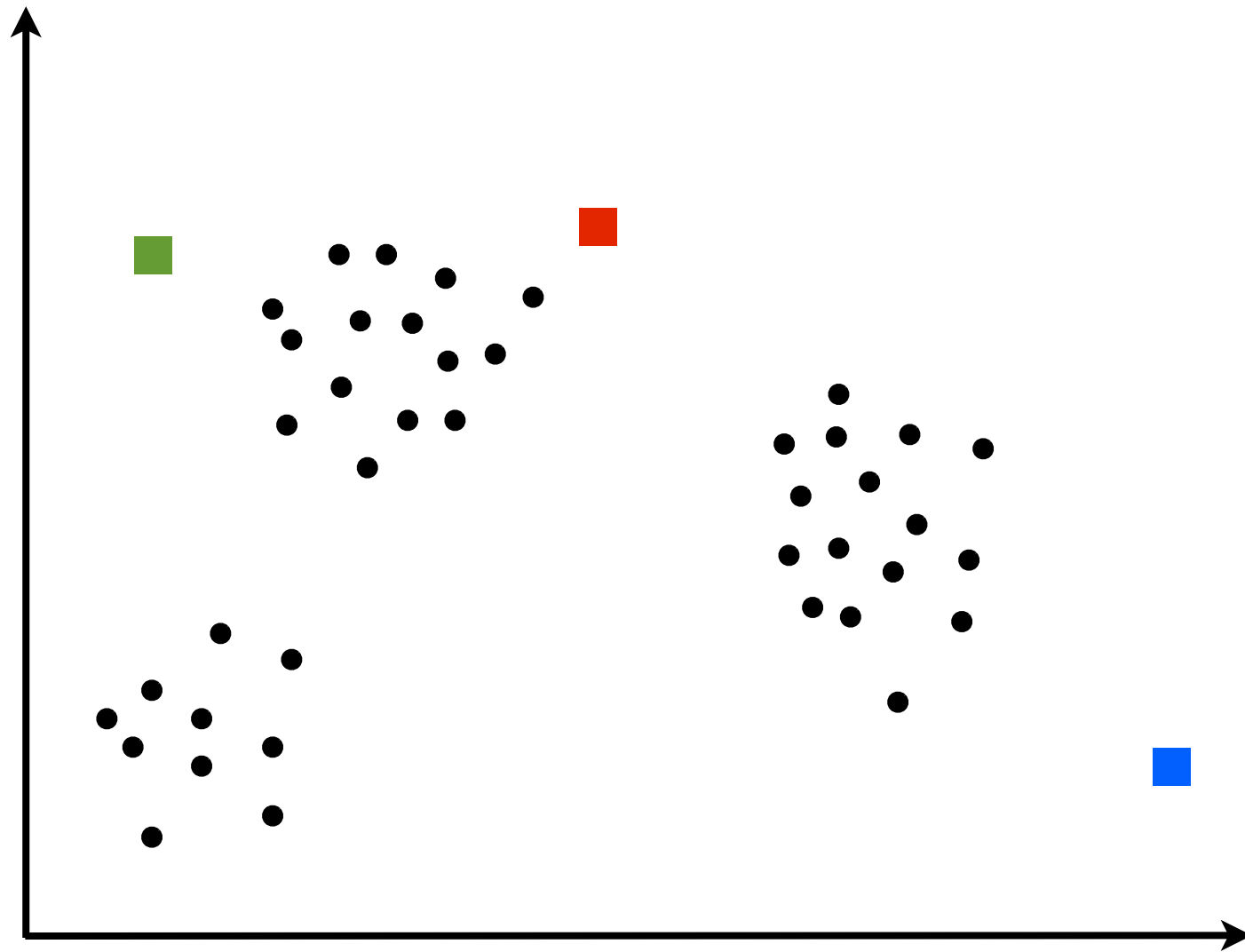
K means algorithm

- Initialize K cluster centers
- Repeat until convergence:
 - ◇ Assign each data point to the cluster with the closest center.
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



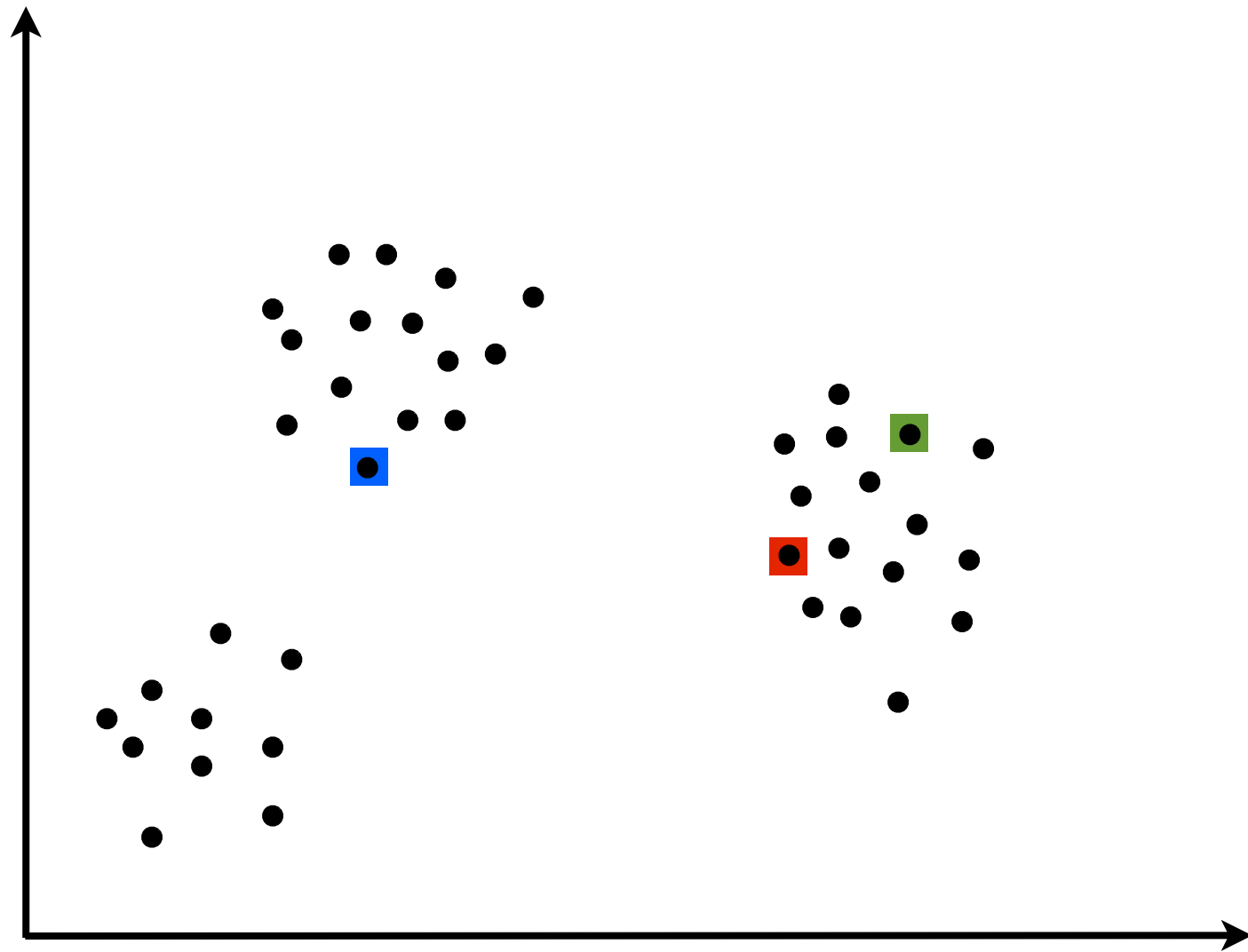
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until convergence:
 - ◇ Assign each data point to the cluster with the closest center.
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



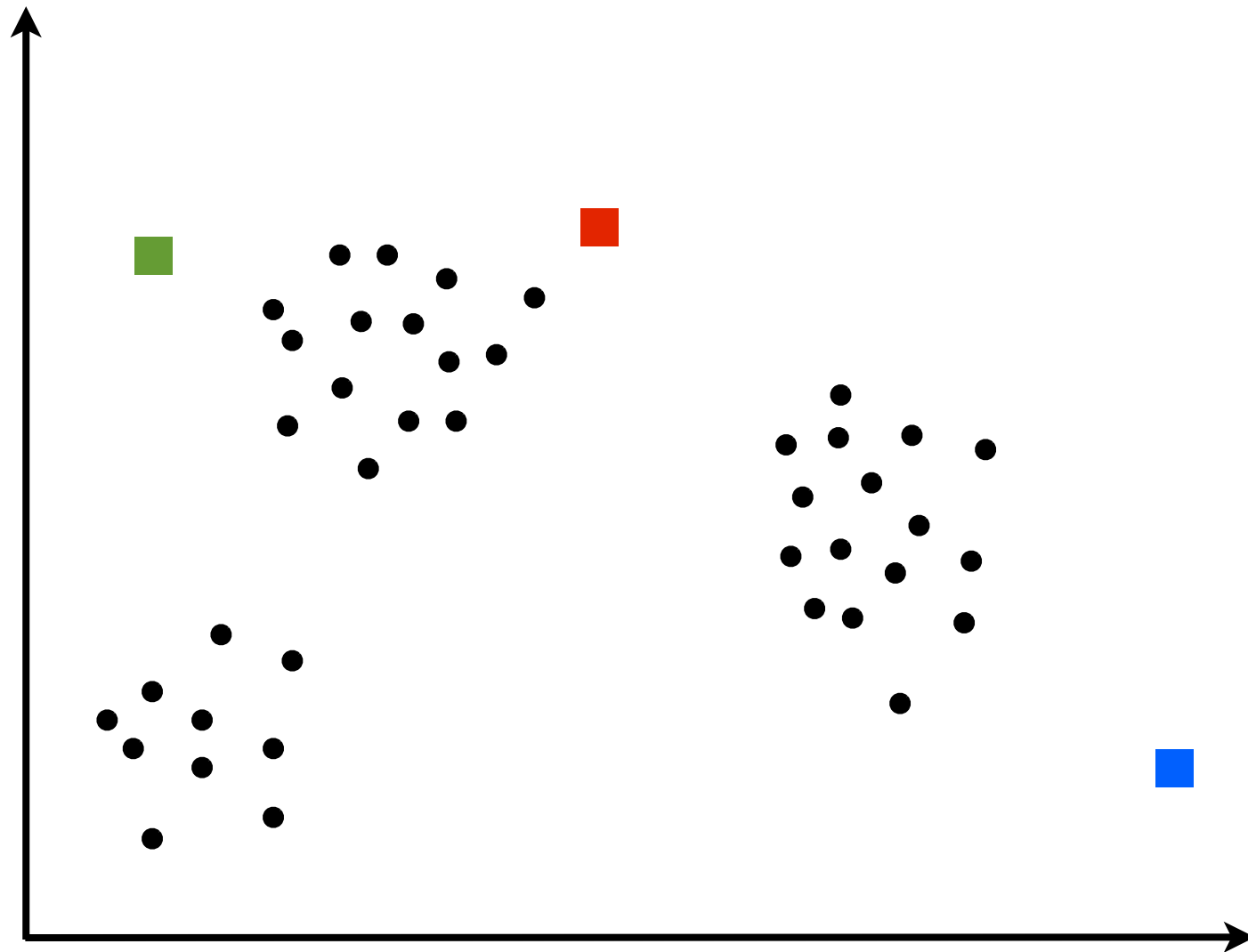
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until convergence:
 - ◇ Assign each data point to the cluster with the closest center.
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



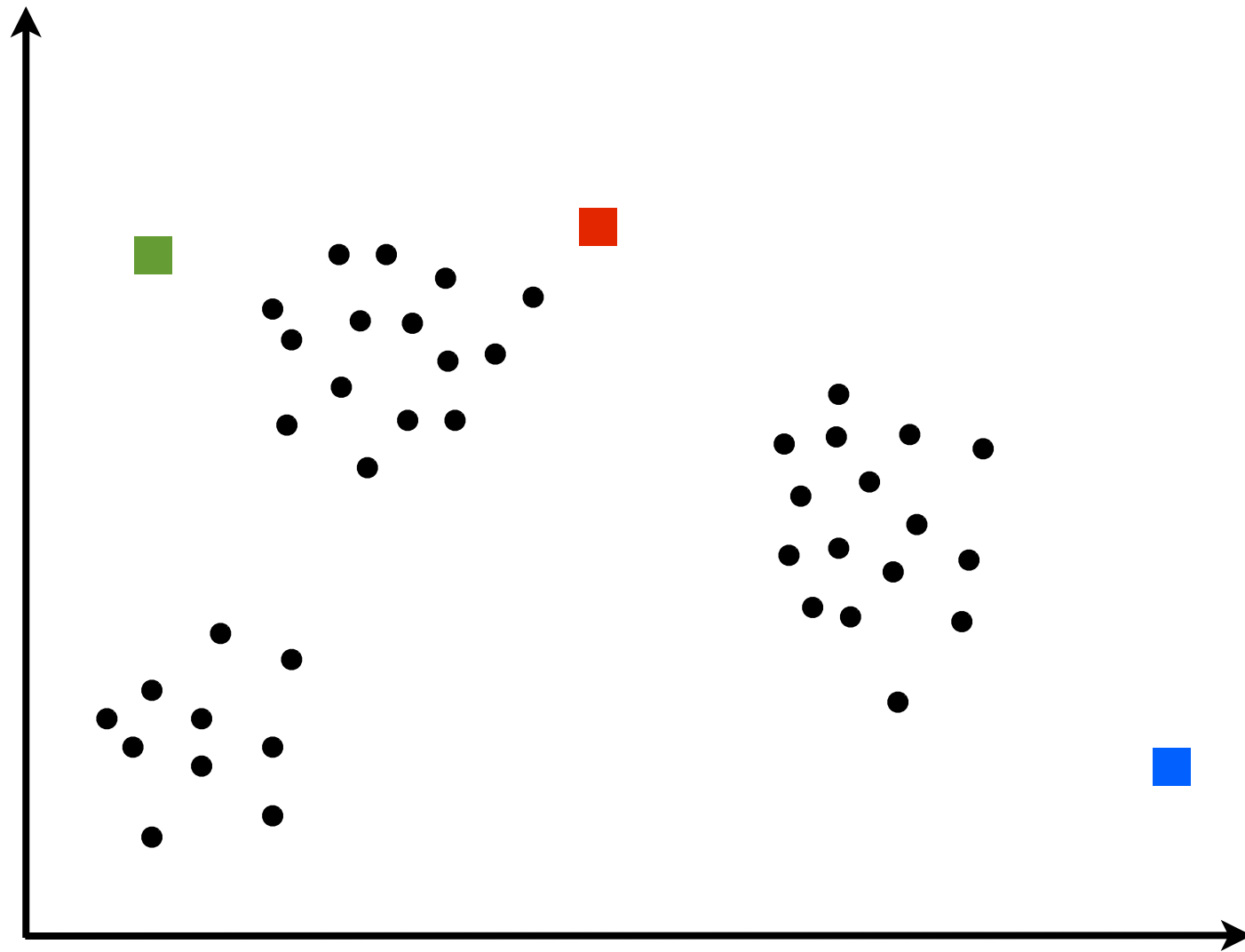
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until convergence:
 - ◇ Assign each data point to the cluster with the closest center.
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



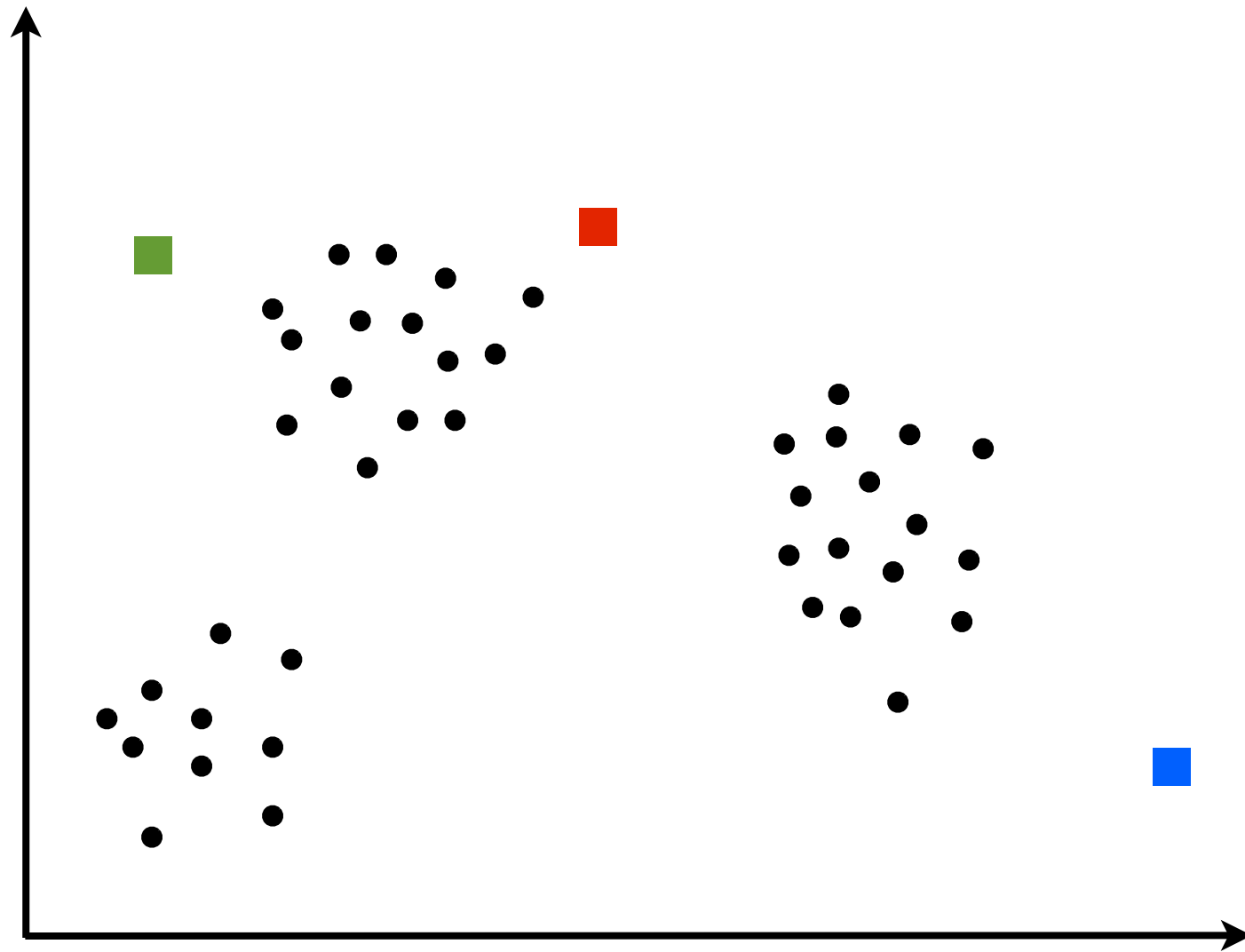
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until convergence:
 - ◇ Assign each data point to the cluster with the closest center.
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



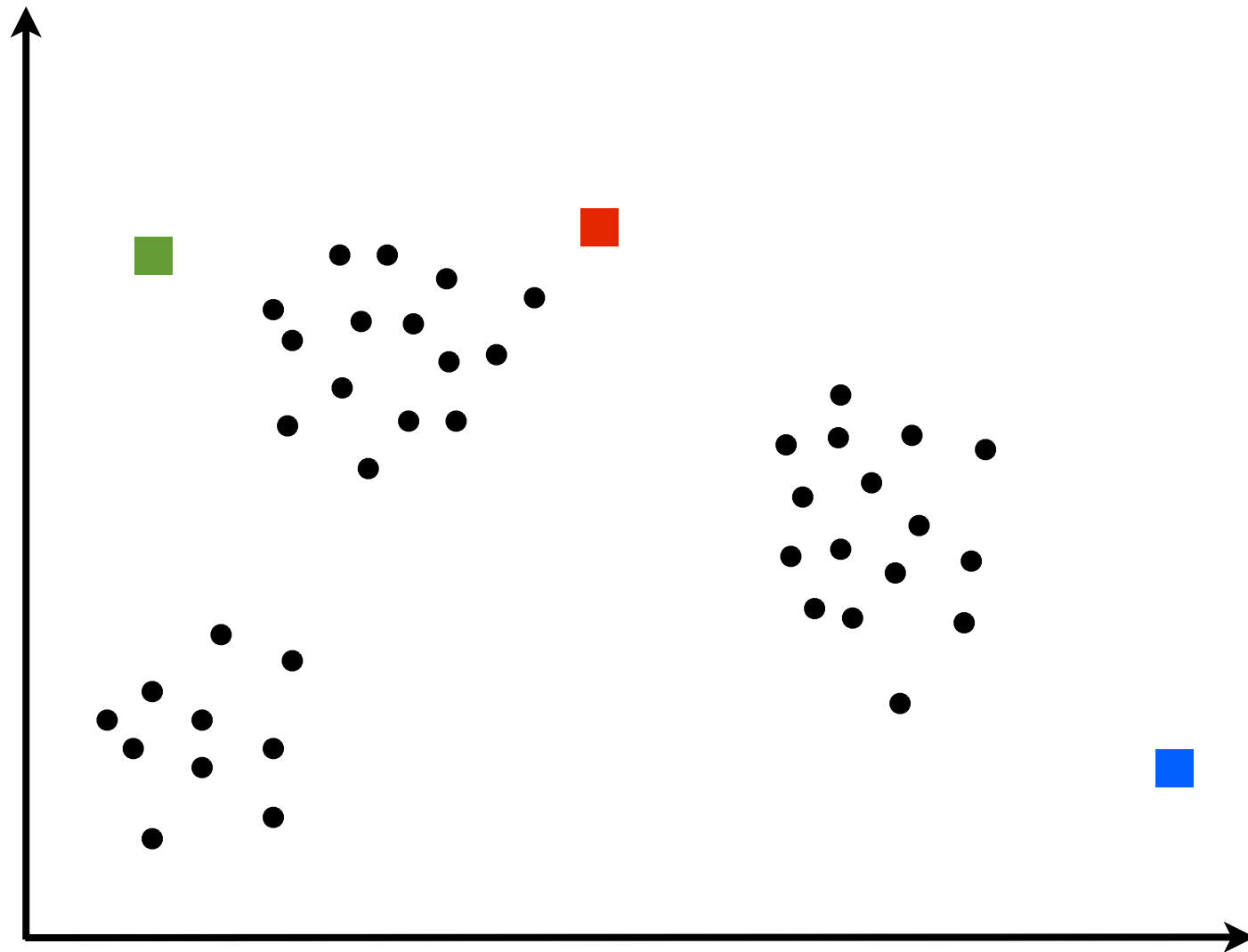
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ Assign each data point to the cluster with the closest center.
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



K means algorithm

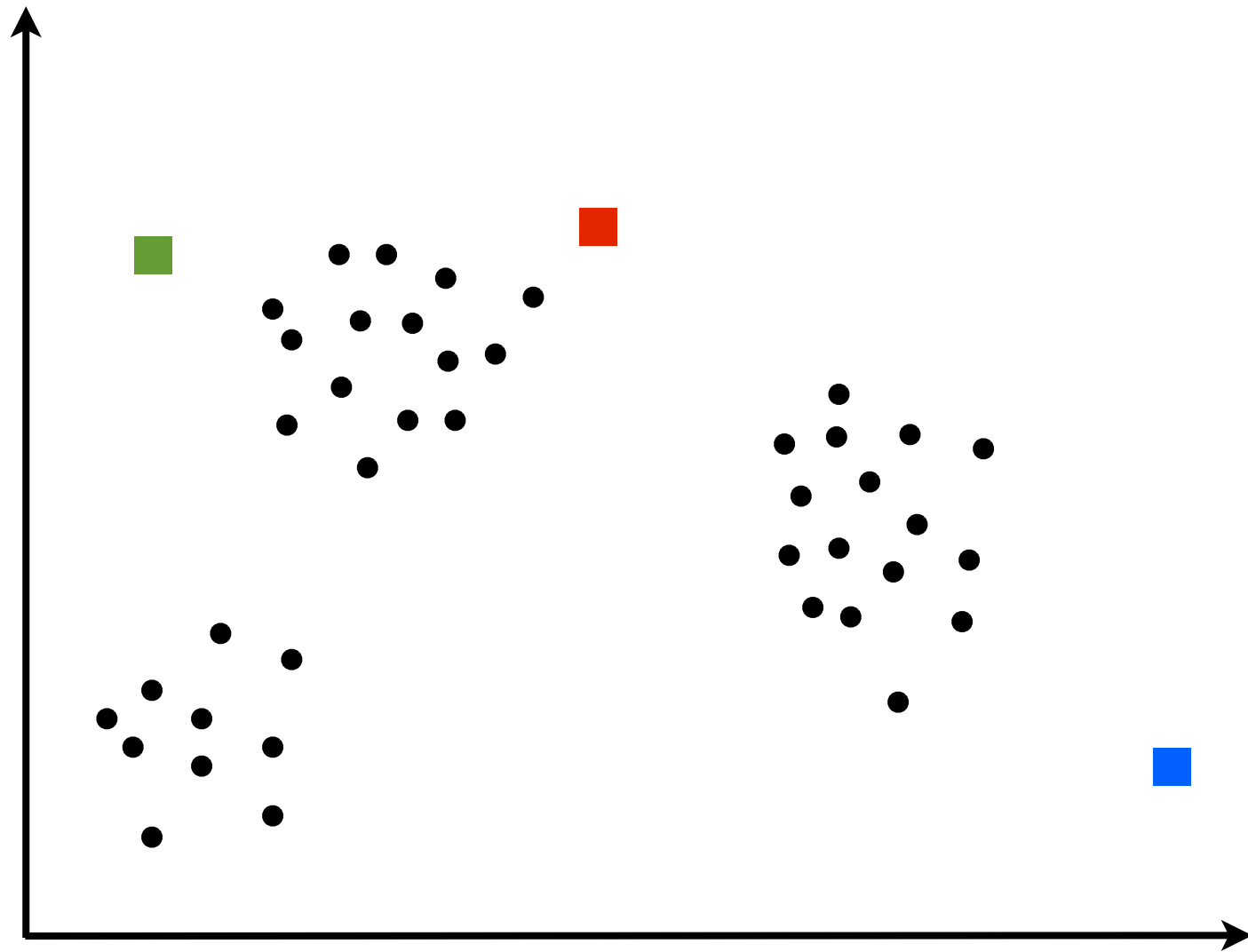
Or no change
in dis_{global}



- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ Assign each data point to the cluster with the closest center.
 - ◇ Assign each cluster center to be the mean of its cluster's data points.

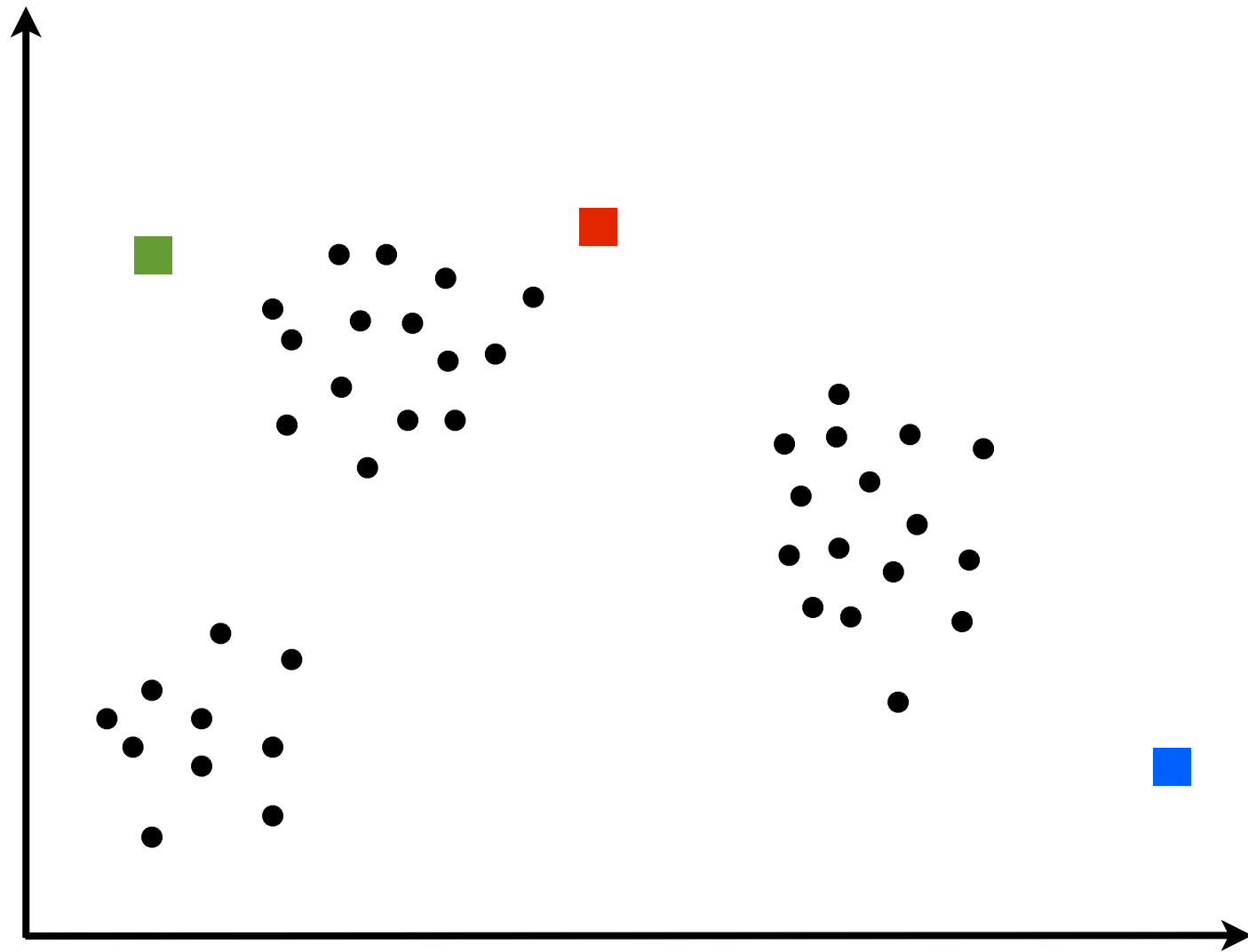
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ Assign each data point to the cluster with the closest center.
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



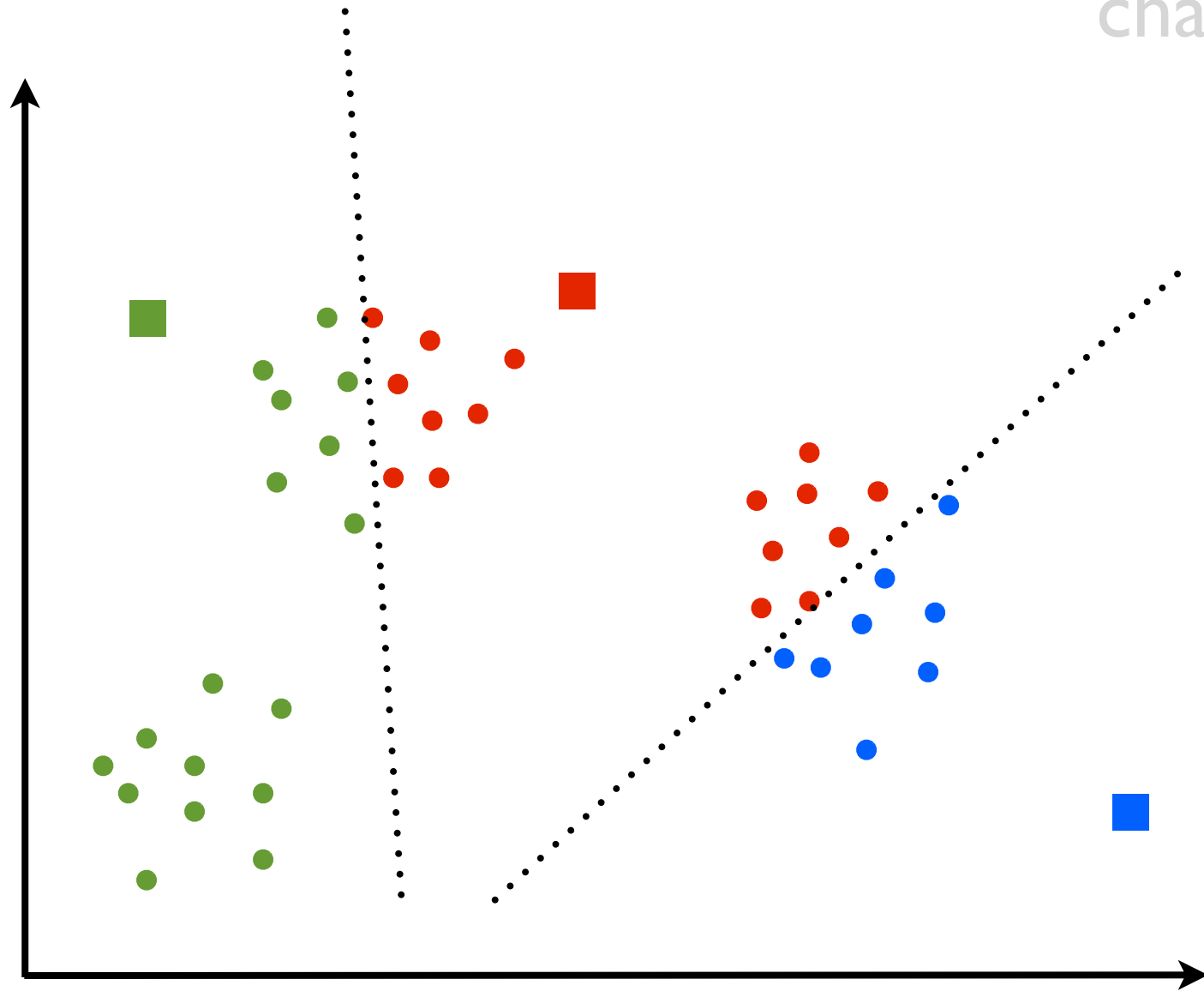
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



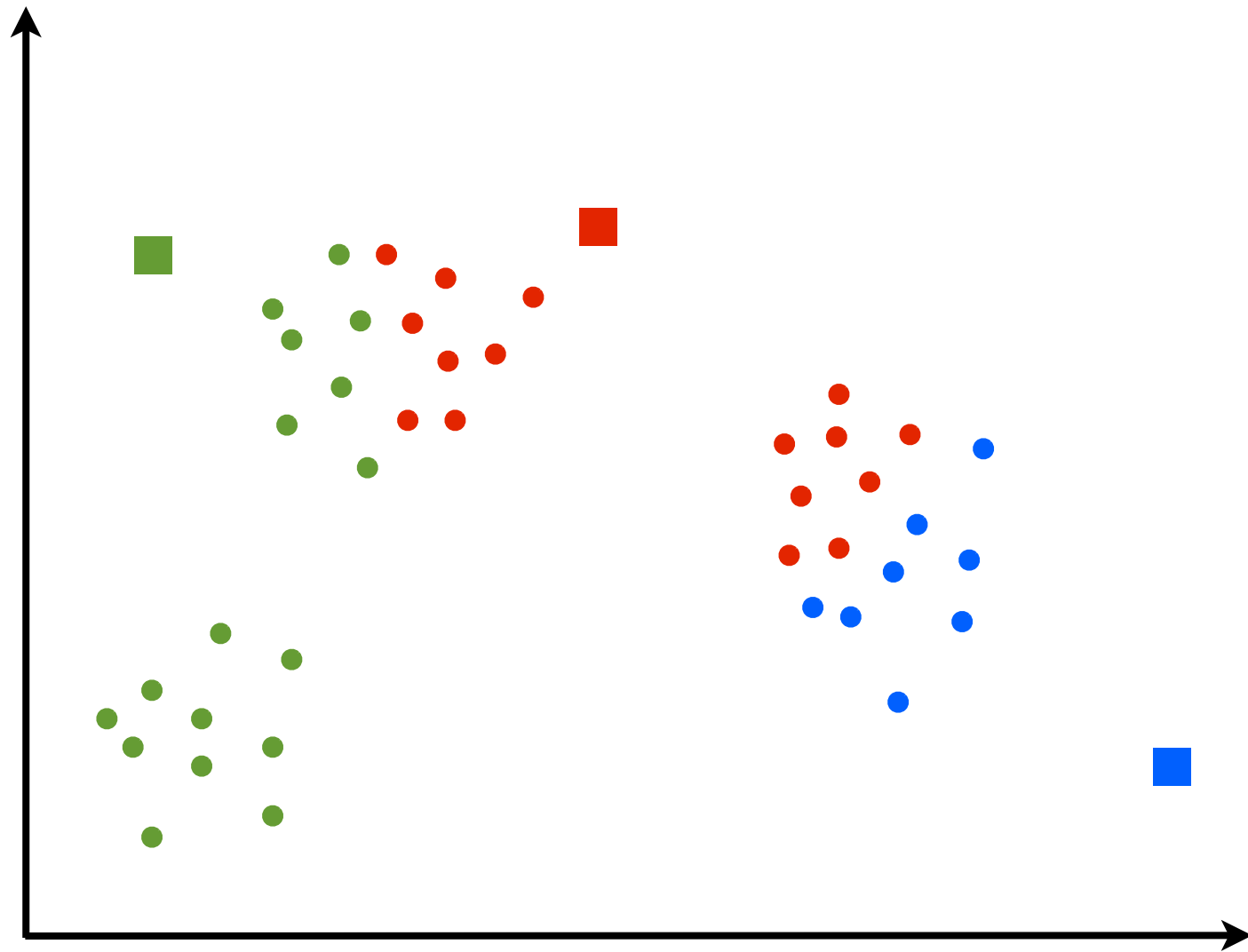
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



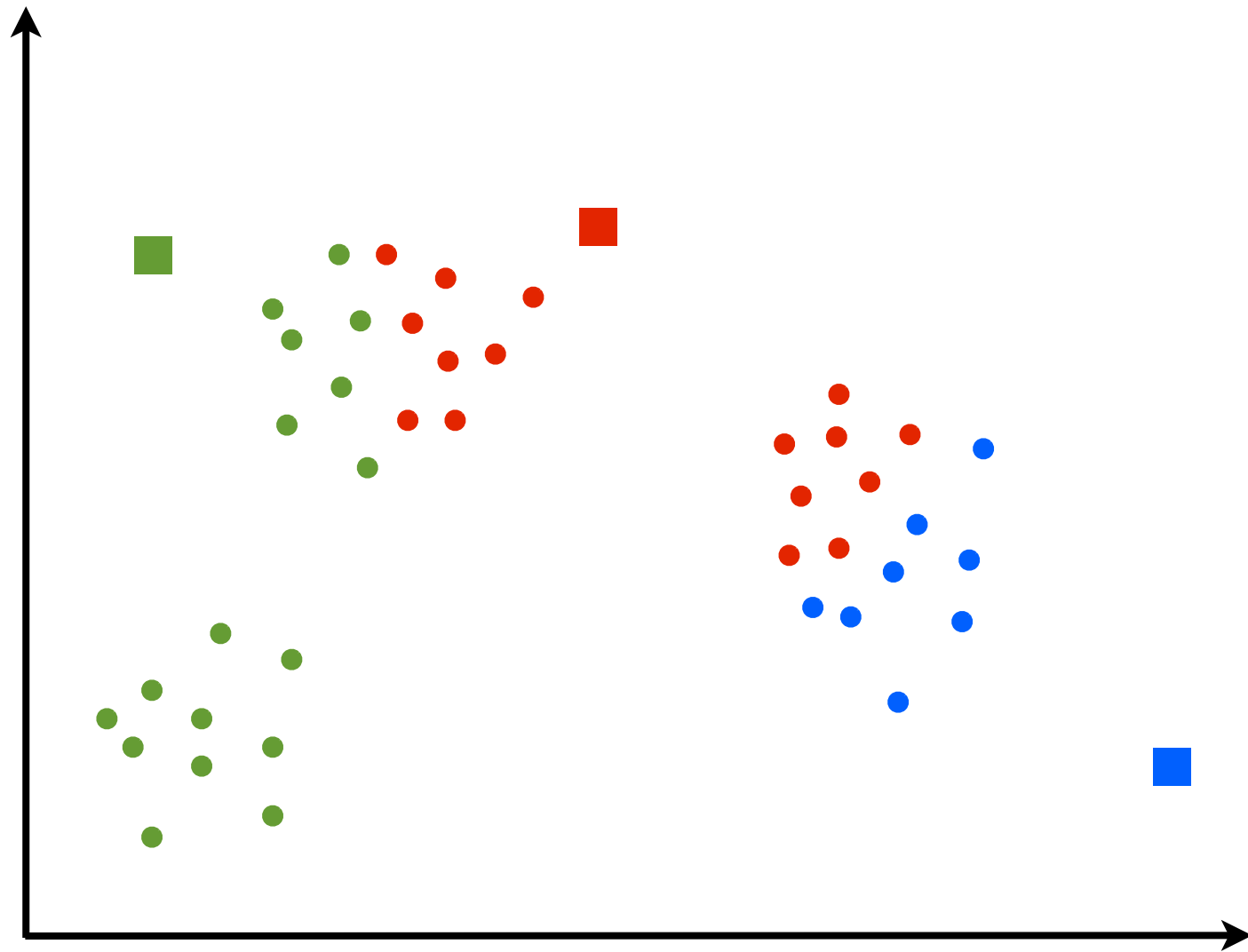
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



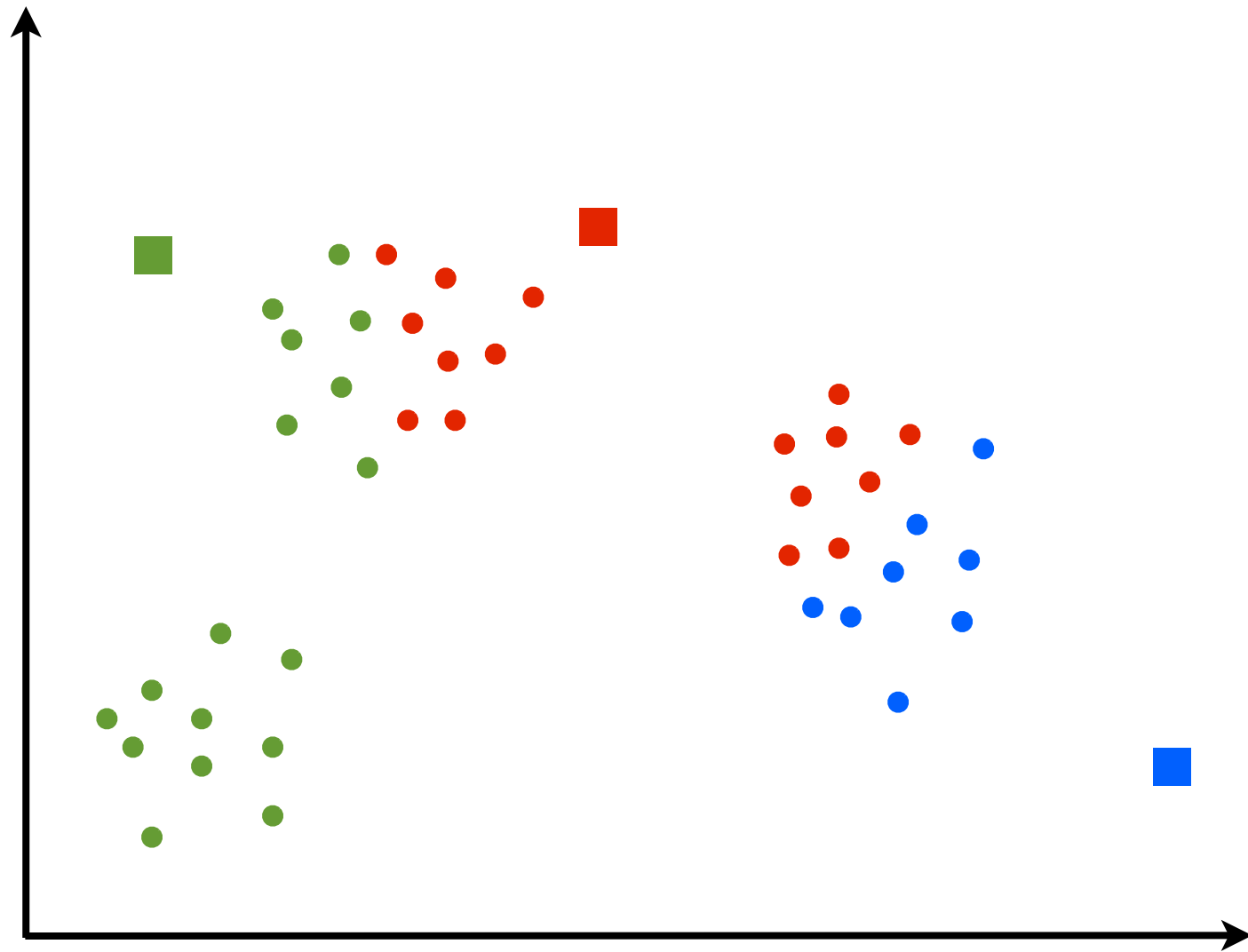
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ Assign each cluster center to be the mean of its cluster's data points.



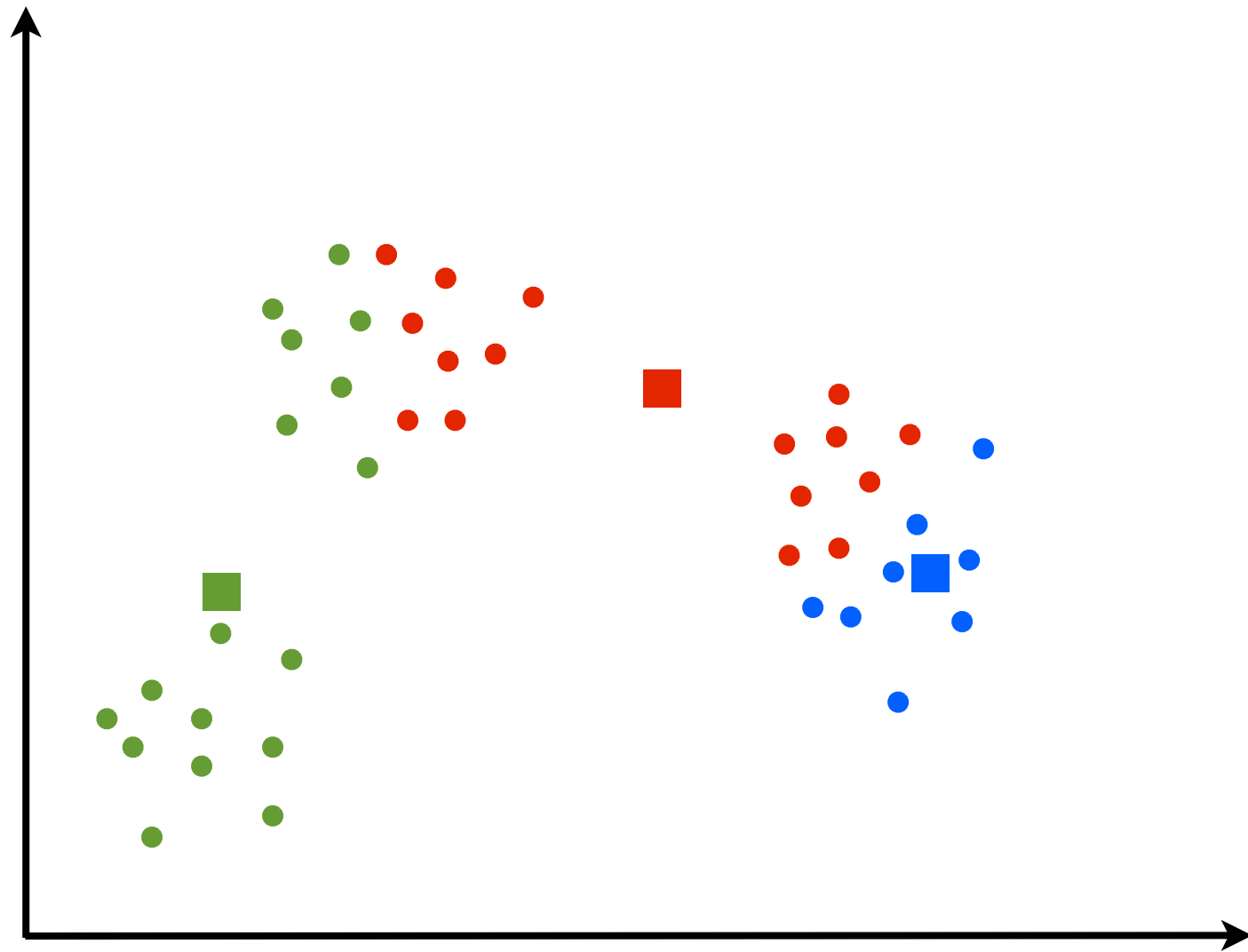
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ For $k = 1, \dots, K$
 - * $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$



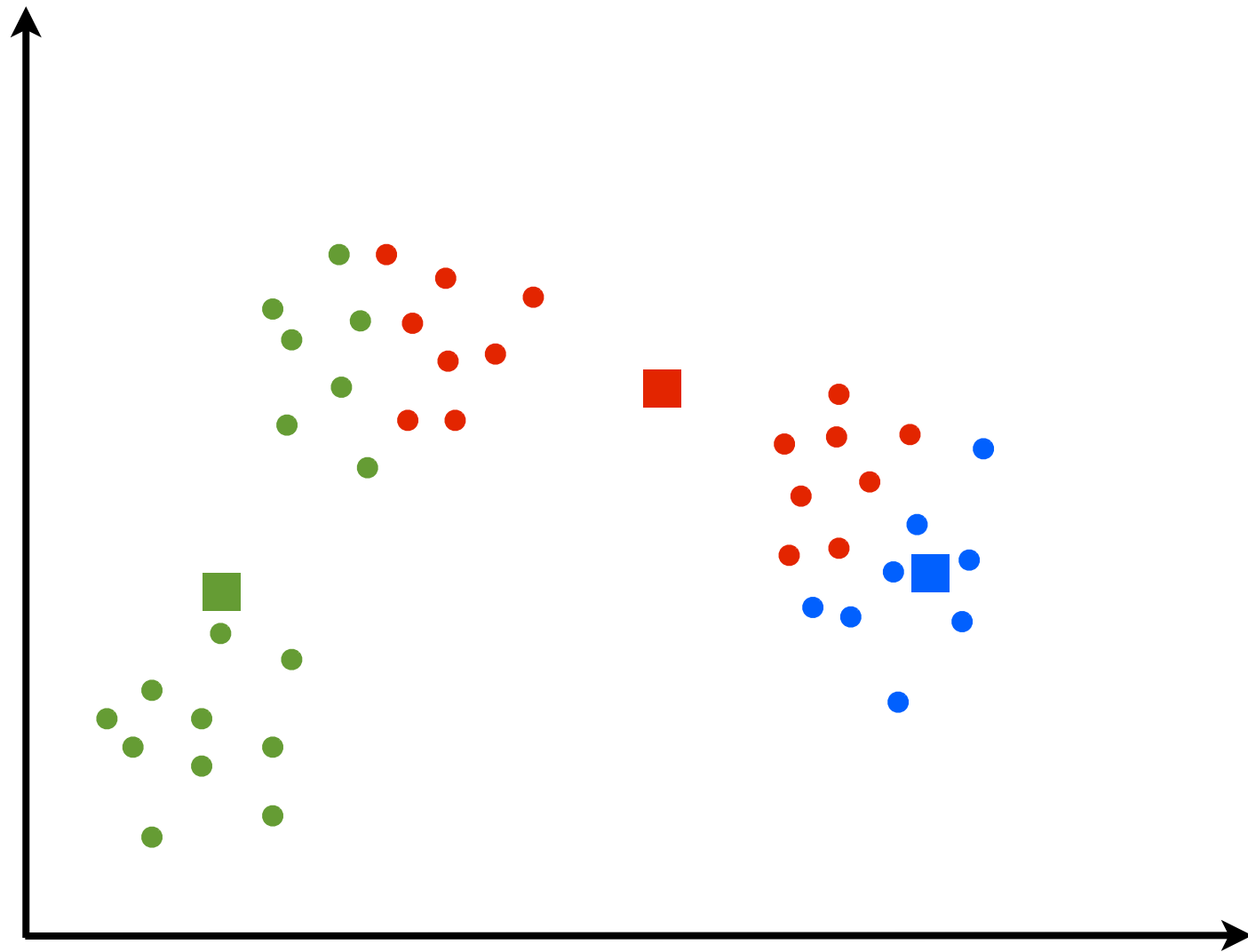
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ For $k = 1, \dots, K$
 - * $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$



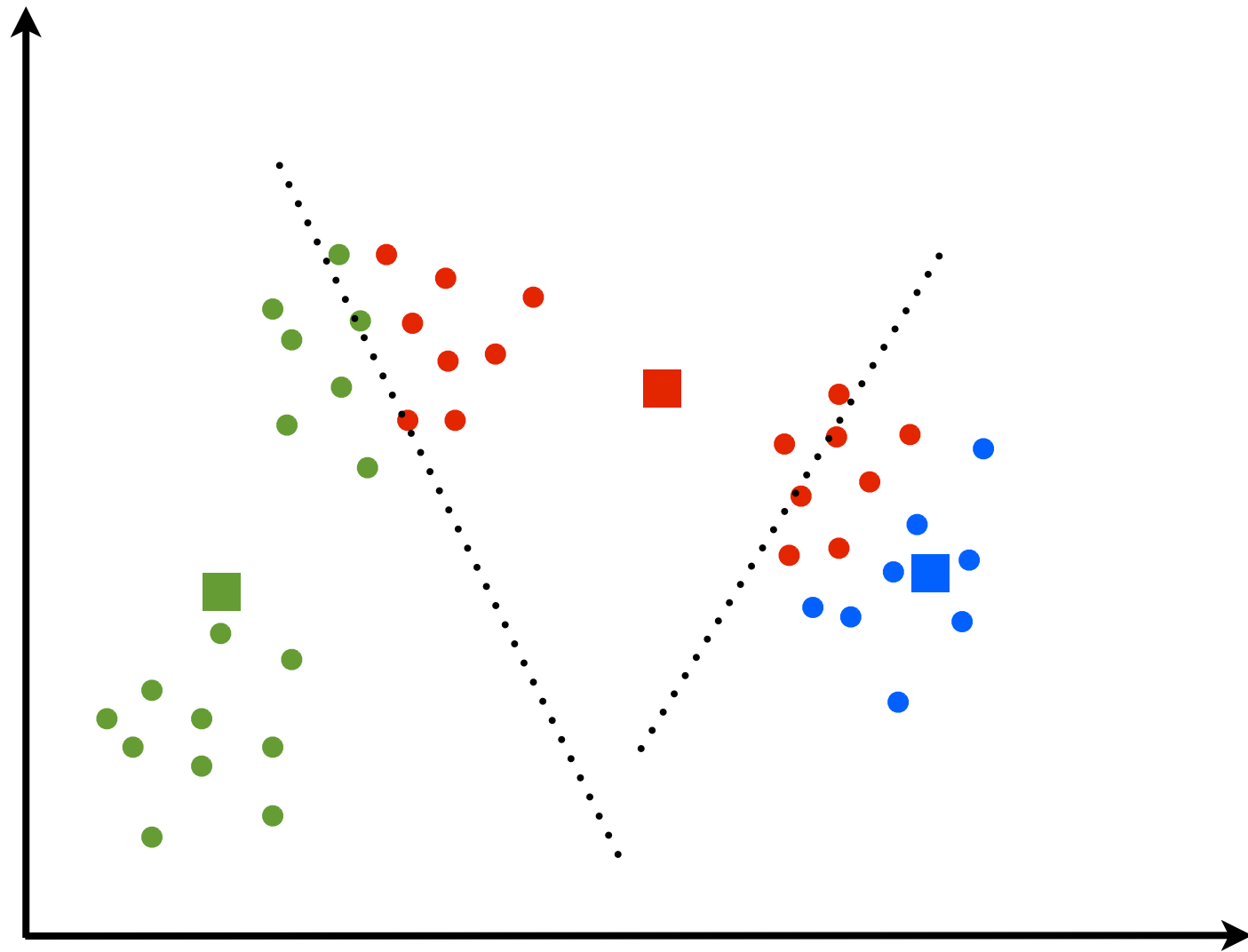
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ For $k = 1, \dots, K$
 - * $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$



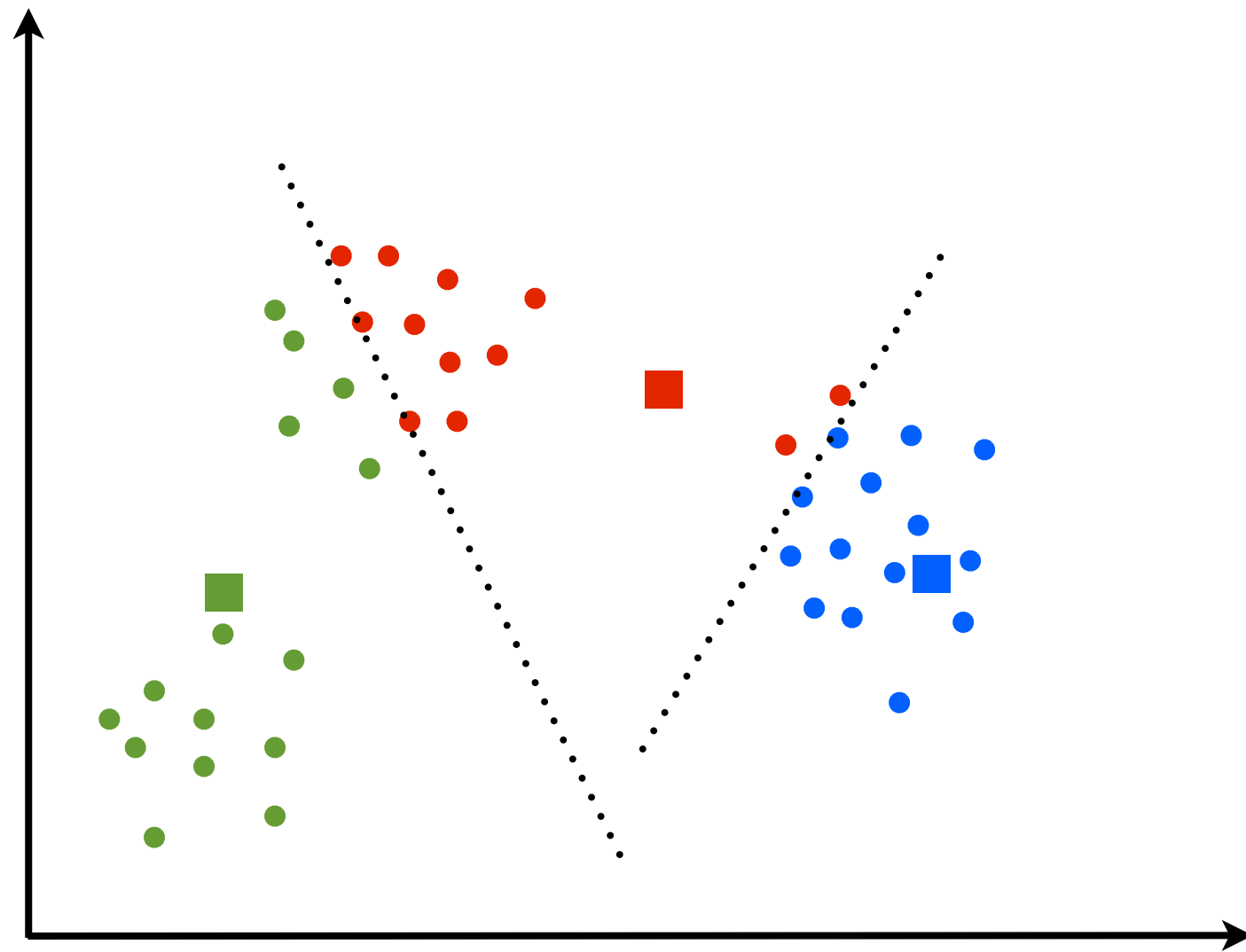
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ For $k = 1, \dots, K$
 - * $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$



K means algorithm

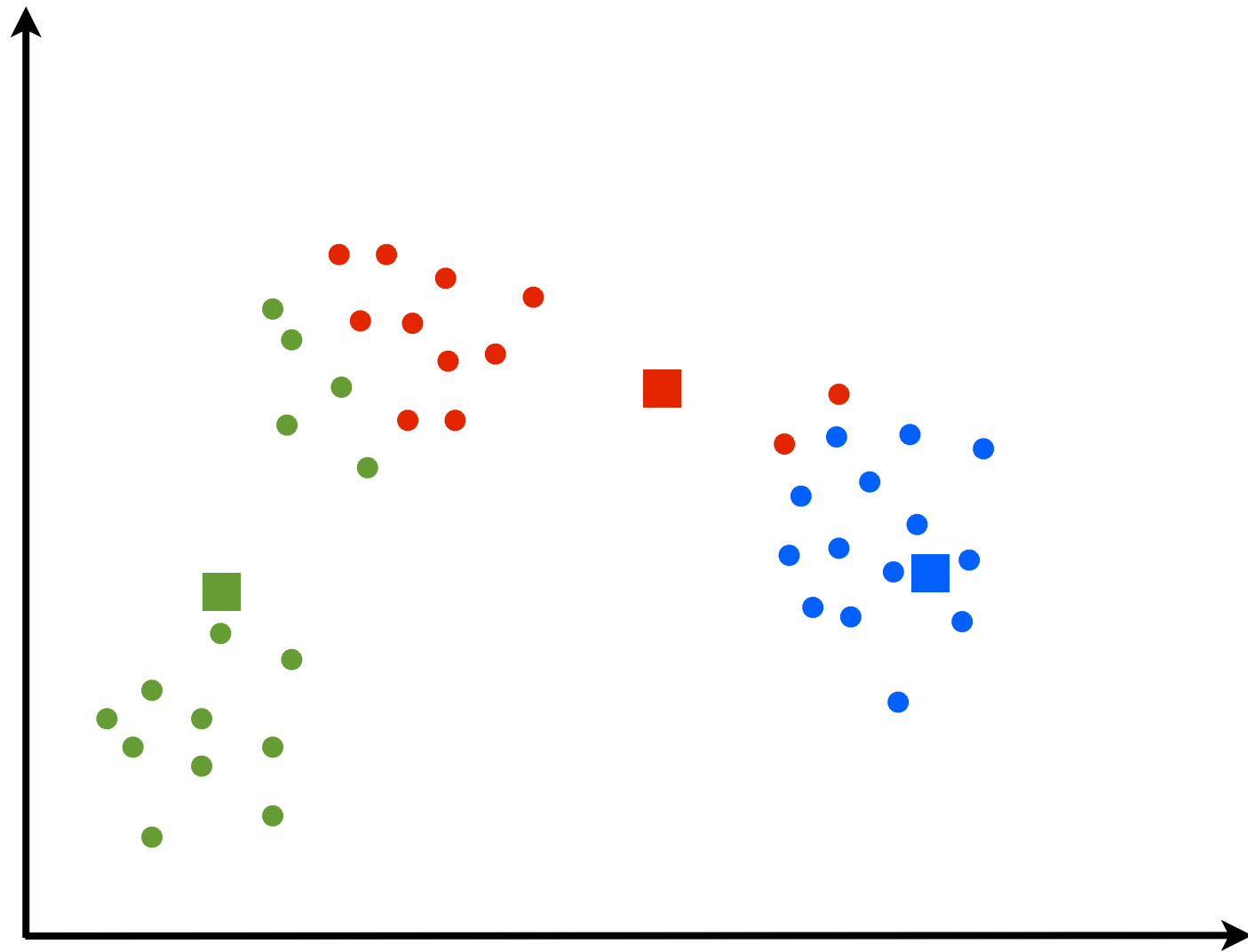
- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:



- ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
- ◇ For $k = 1, \dots, K$
 - * $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$

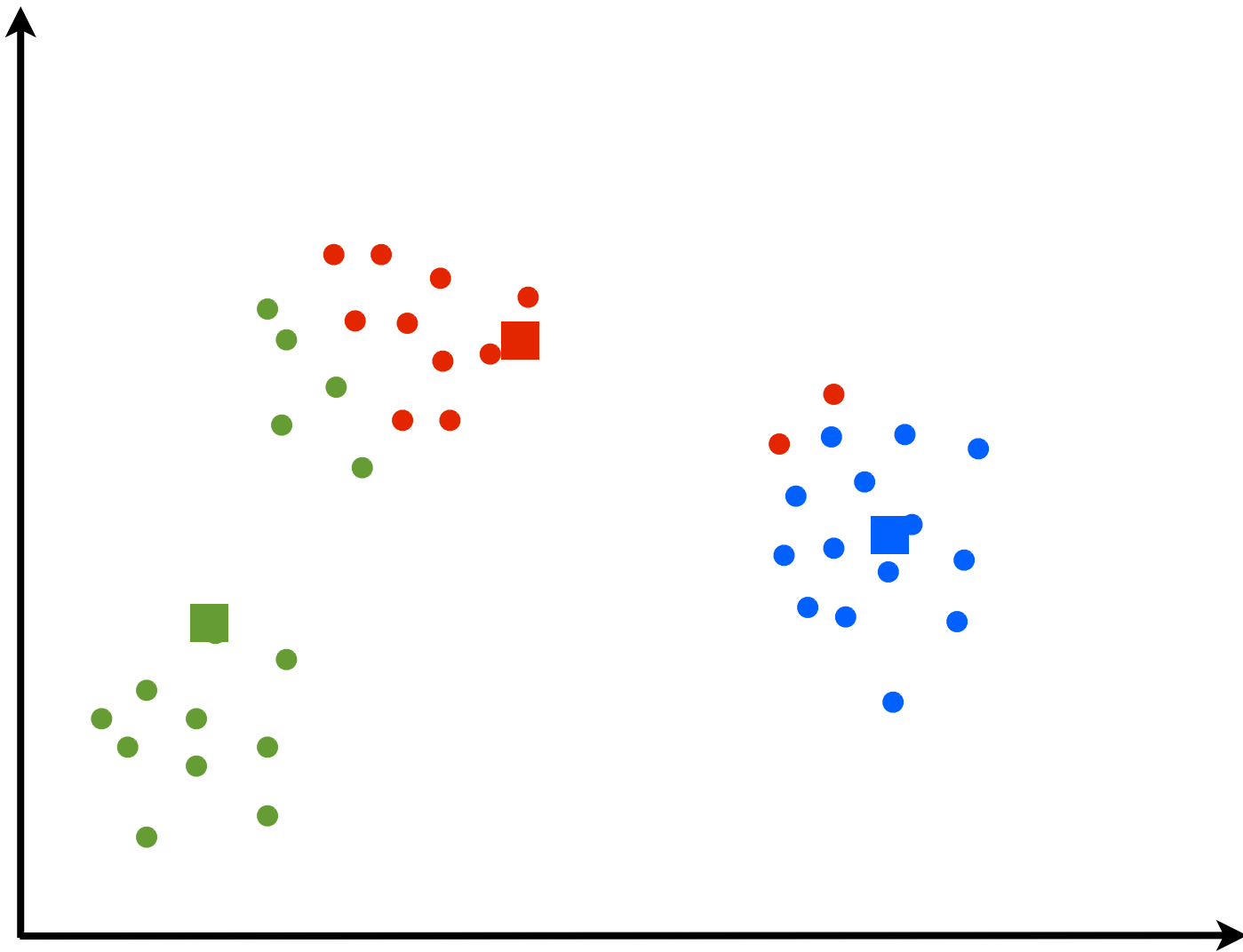
K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ For $k = 1, \dots, K$
 - * $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$



K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ For $k = 1, \dots, K$
 - * $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$

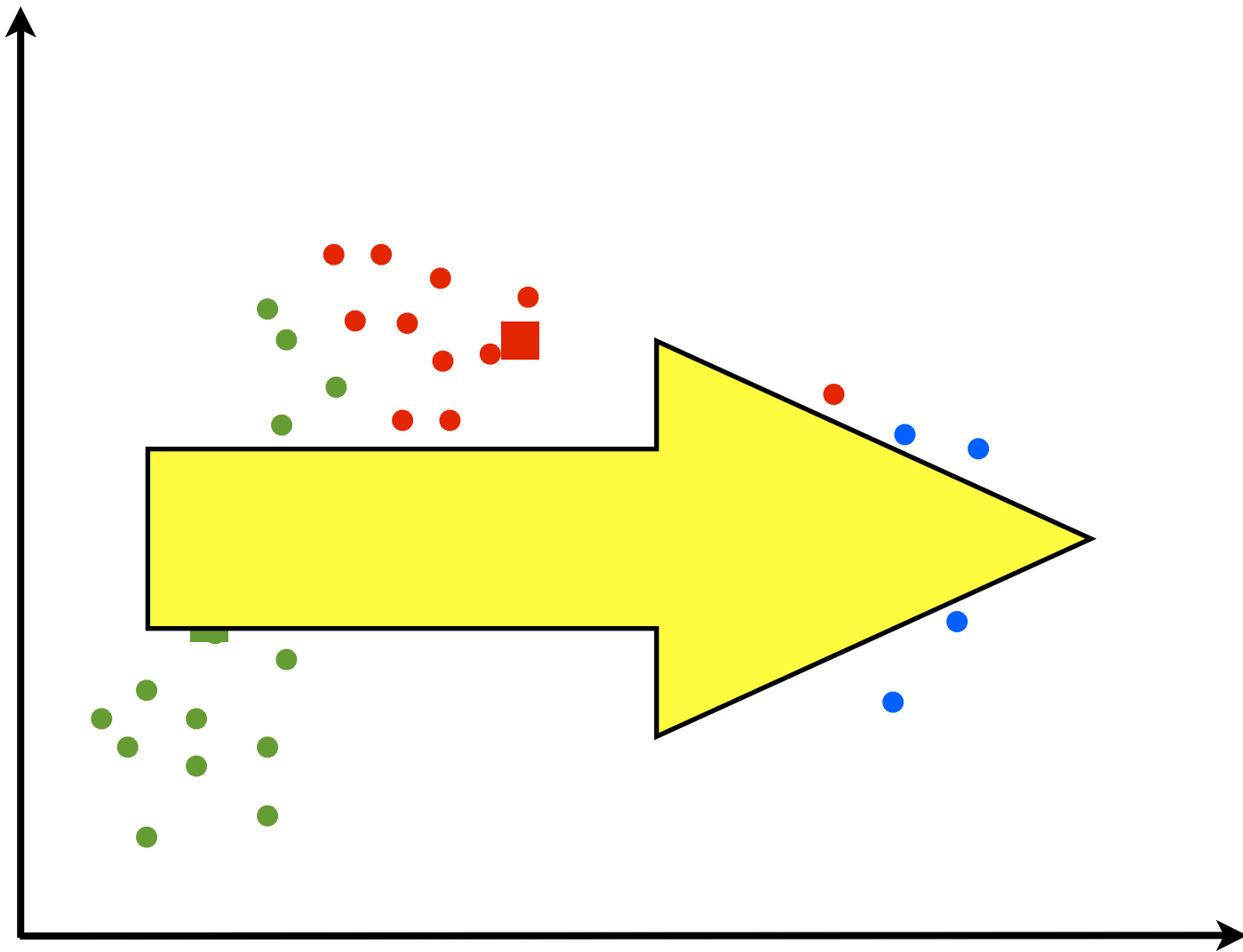


K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$

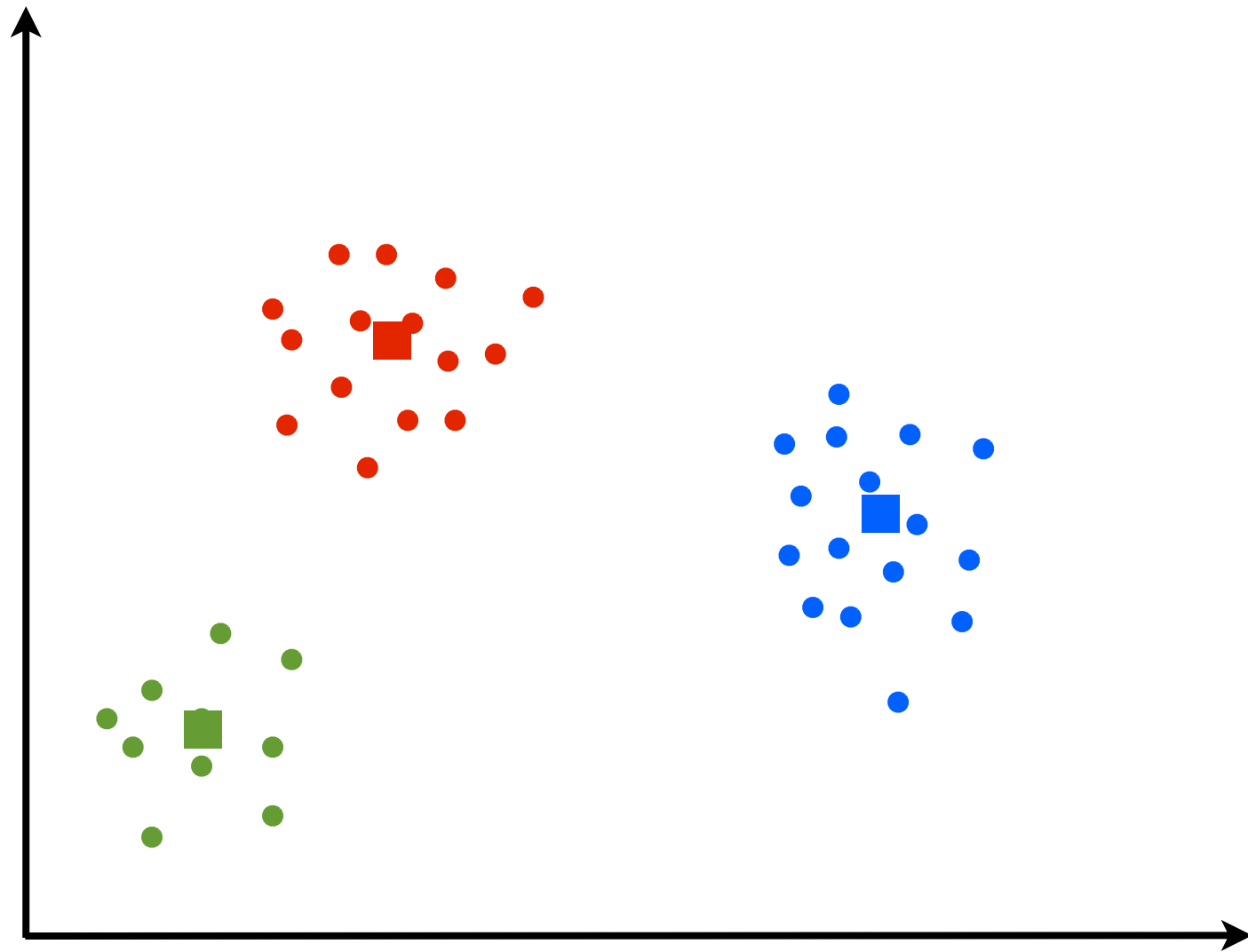
• Repeat until $S_1, \dots, S_K$ don't change:

- ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
- ◇ For $k = 1, \dots, K$
 - * $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$



K means algorithm

- For $k = 1, \dots, K$
 - ◇ Randomly draw n from $1, \dots, N$ without replacement
 - ◇ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \dots, S_K$ don't change:
 - ◇ For $n = 1, \dots, N$
 - * Find k with smallest $dis(x_n, \mu_k)$
 - * Put $x_n \in S_k$ (and no other S_j)
 - ◇ For $k = 1, \dots, K$
 - * $\mu_k \leftarrow |S_k|^{-1} \sum_{n:n \in S_k} x_n$



Outline

Clustering: Grouping data according to similarity.

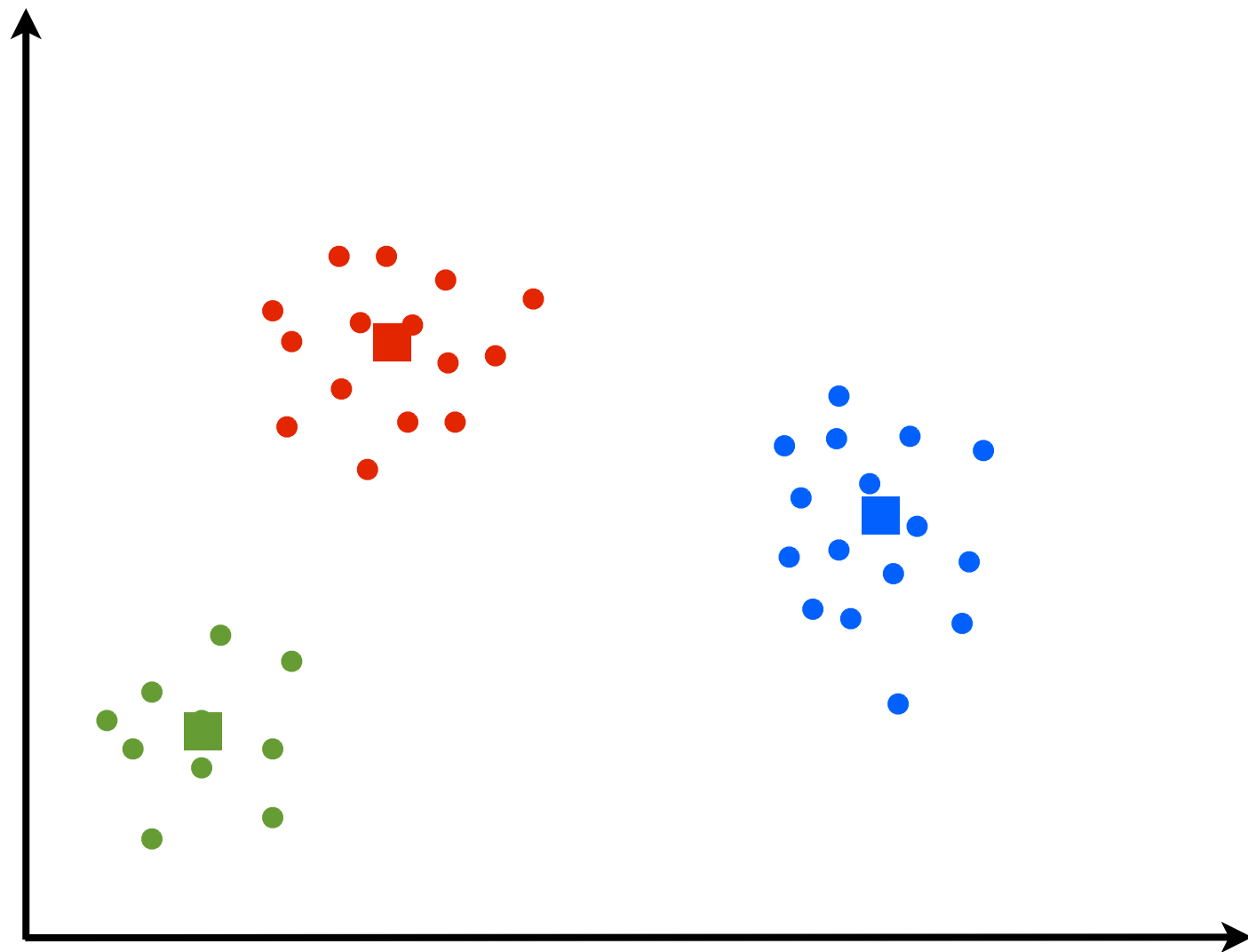
1. K means algorithm
2. Clustering evaluation
3. Clustering trouble-shooting
4. Example

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
- 2. Clustering evaluation**
3. Clustering trouble-shooting
4. Example

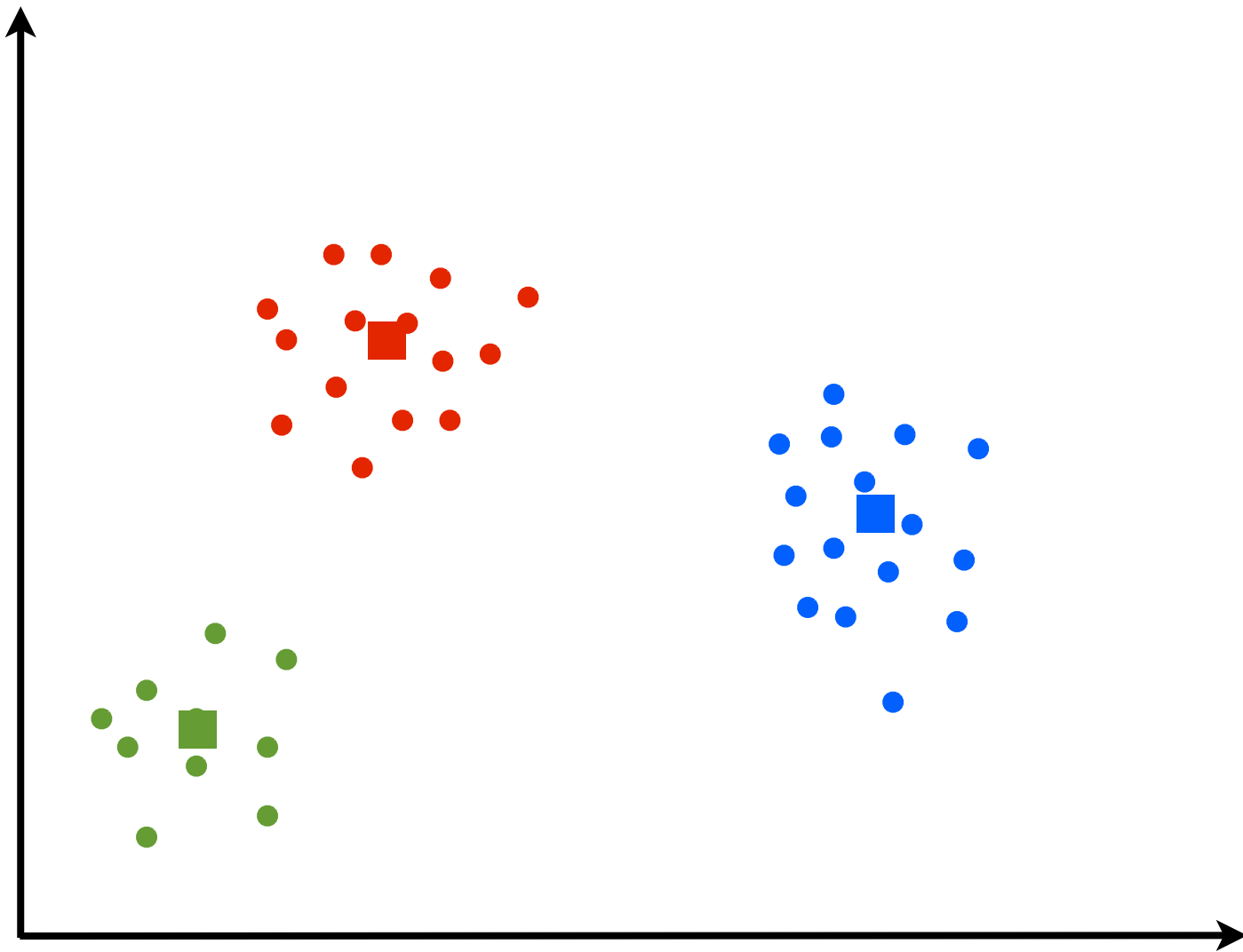
K means: evaluation



K means: evaluation

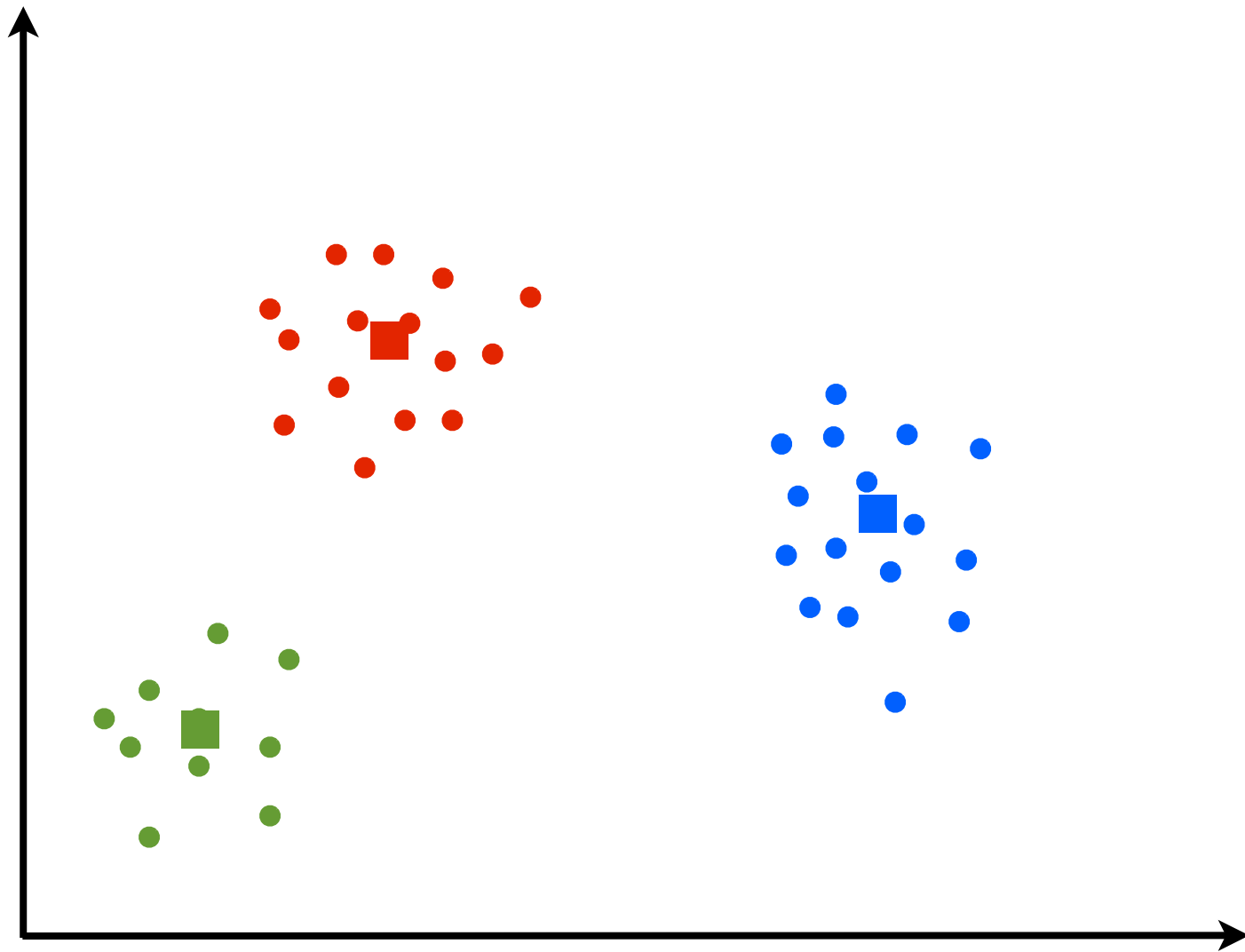
- Will it terminate?

— — — — —



K means: evaluation

- Will it terminate?
Yes. Always.

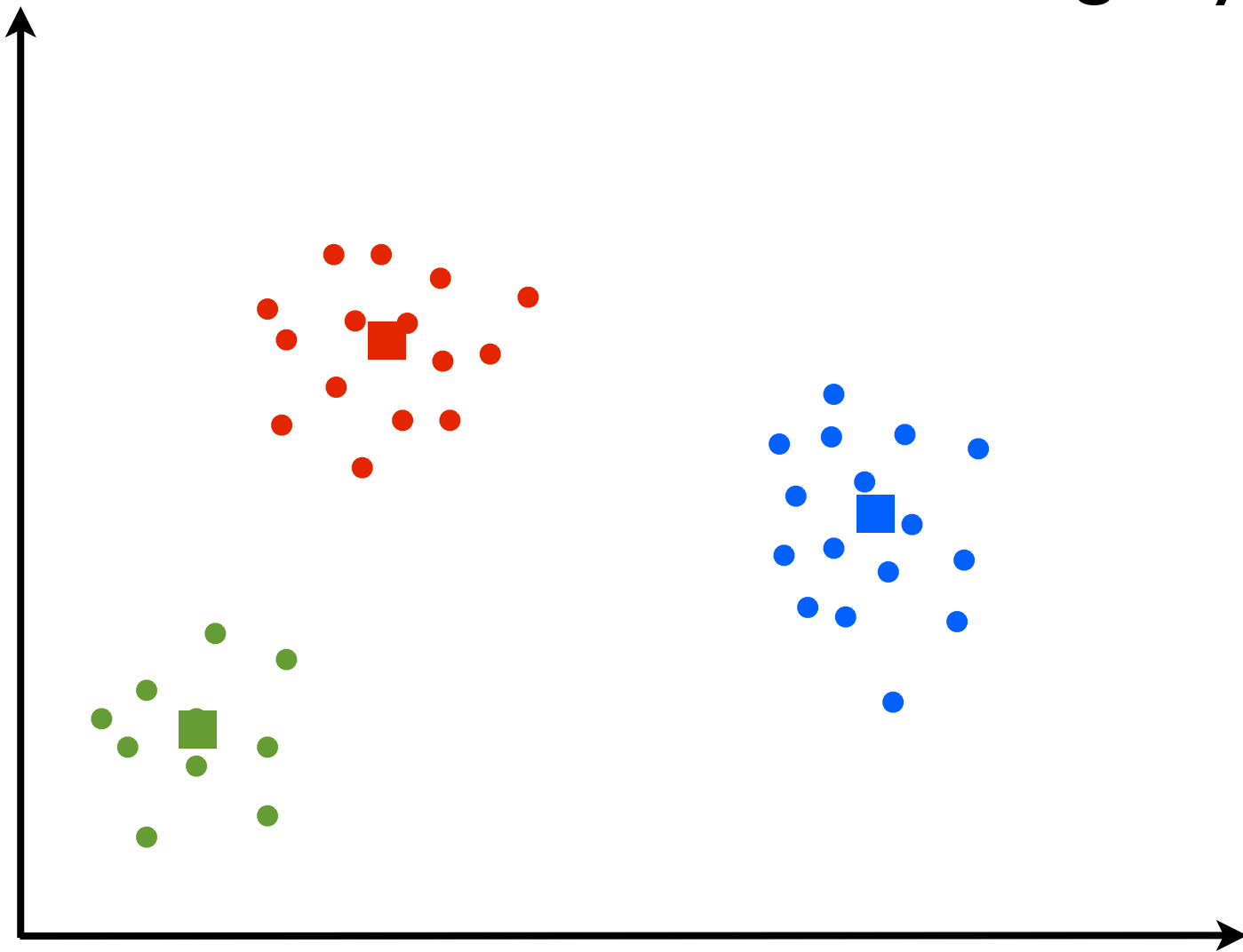


K means: evaluation

- Will it terminate?

Yes. Always.

- Is the clustering any good?



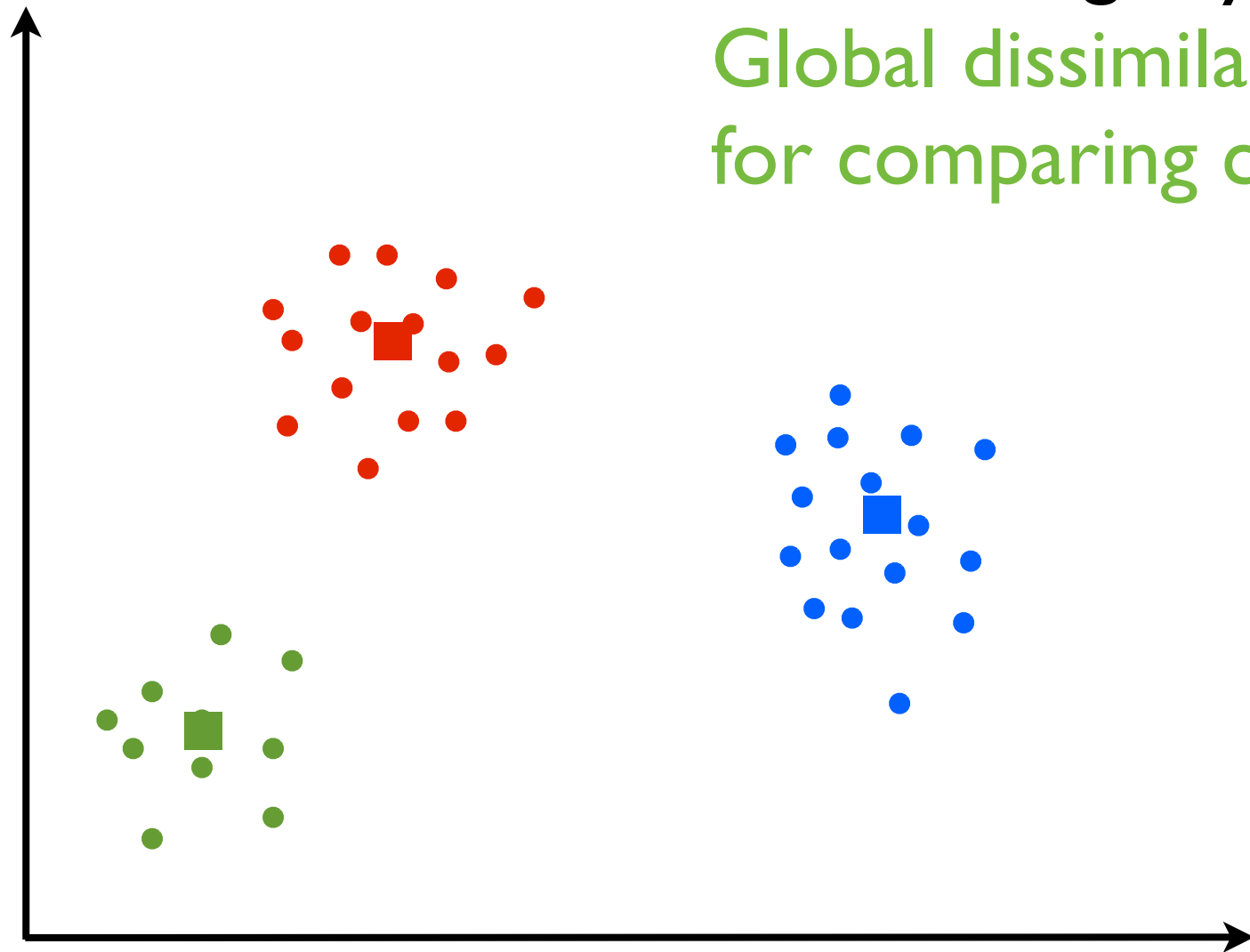
K means: evaluation

- Will it terminate?

Yes. Always.

- Is the clustering any good?

Global dissimilarity only useful for comparing clusterings.



Clustering: evaluation

Recall: Classification

Clustering: evaluation

Recall: Classification

- Evaluate on test data

Clustering: evaluation

Recall: Classification

- Evaluate on test data
- Absolute, universal scale: 0 - 100% accuracy

Clustering: evaluation

Recall: Classification

- Evaluate on test data
- Absolute, universal scale: 0 - 100% accuracy

How to evaluate a clustering algorithm?

Clustering: evaluation

Recall: Classification

- Evaluate on test data
- Absolute, universal scale: 0 - 100% accuracy

How to evaluate a clustering algorithm?

Short answer: No one agrees!

Clustering: evaluation

How to evaluate a clustering algorithm?

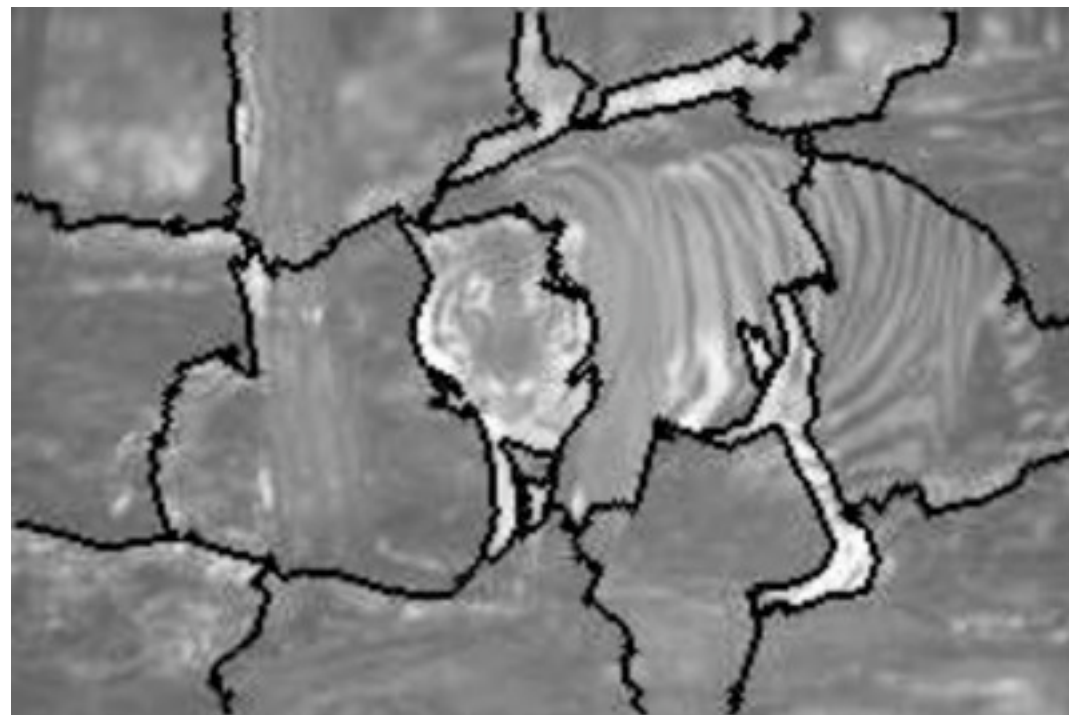
- Visualization

Clustering: evaluation

How to evaluate a clustering algorithm?

- Visualization

Image segmentation



Clustering: evaluation

How to evaluate a clustering algorithm?

- Visualization

Topic analysis

NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

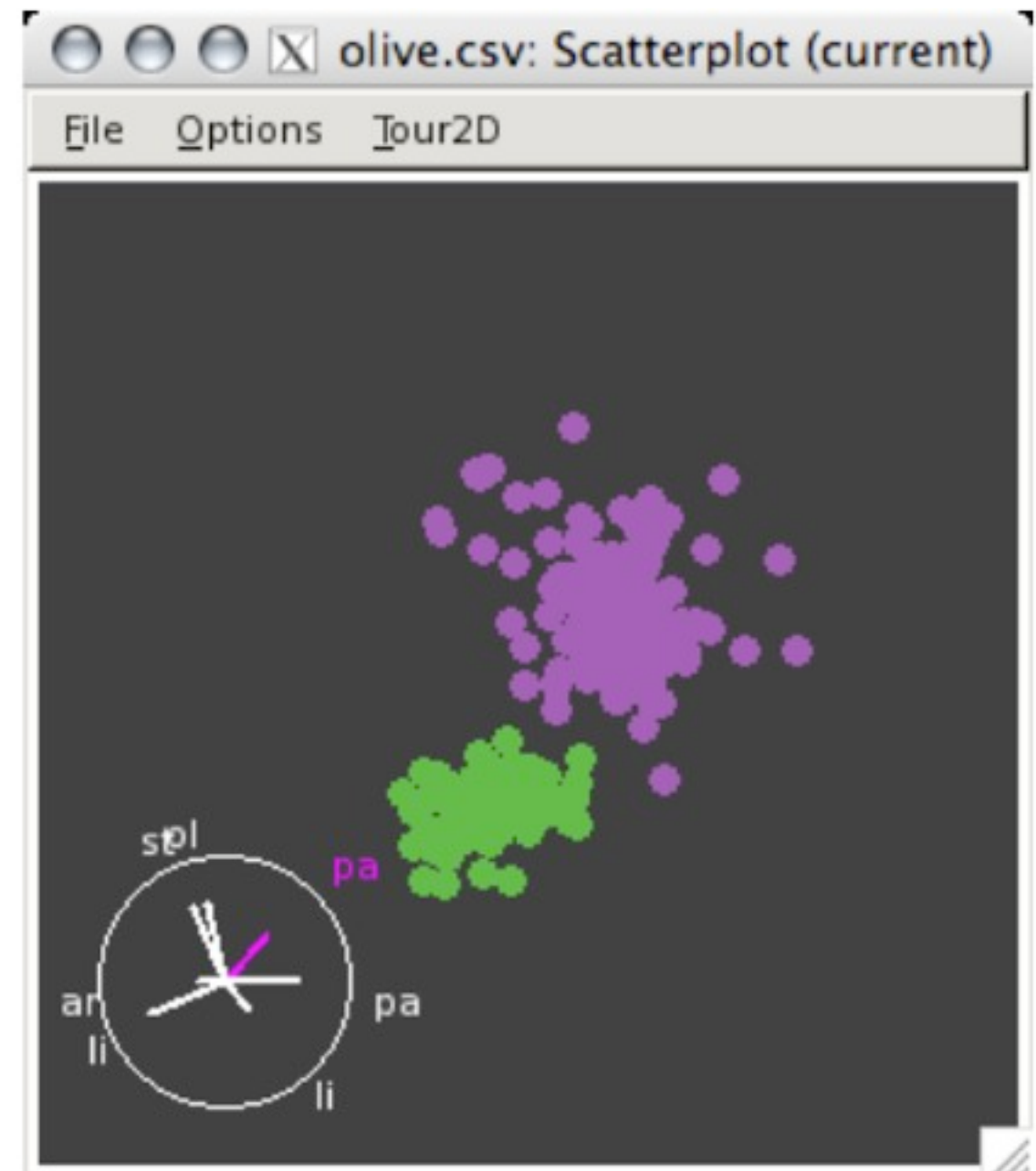
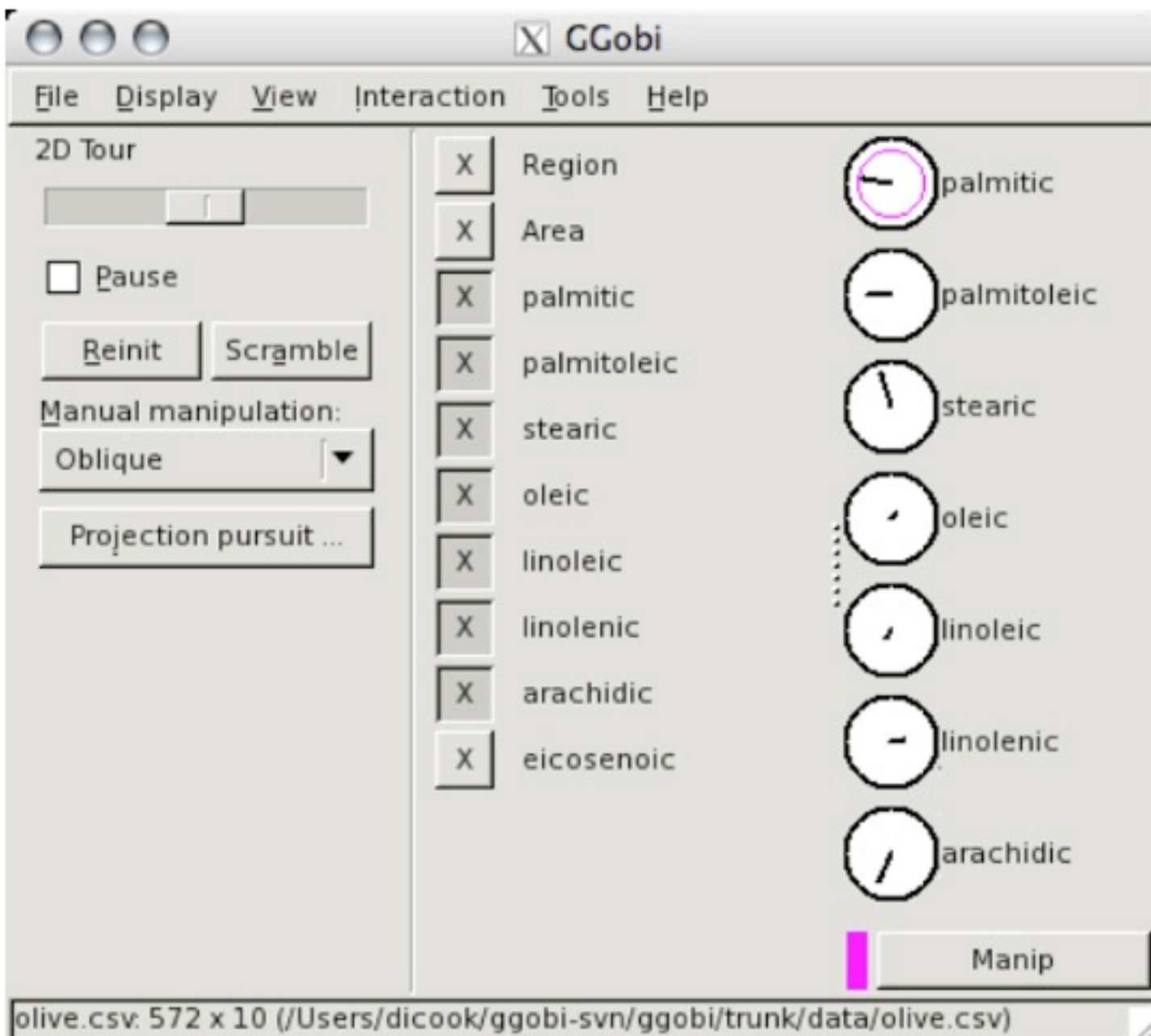
President
Tiger
Lost
Parents
Opera
Dollar
Ennui
Chess

Clustering: evaluation

How to evaluate a clustering algorithm?

- Visualization

GGobi tool



Clustering: evaluation

How to evaluate a clustering algorithm?

- Visualization
- Comparing clusterings:

Clustering: evaluation

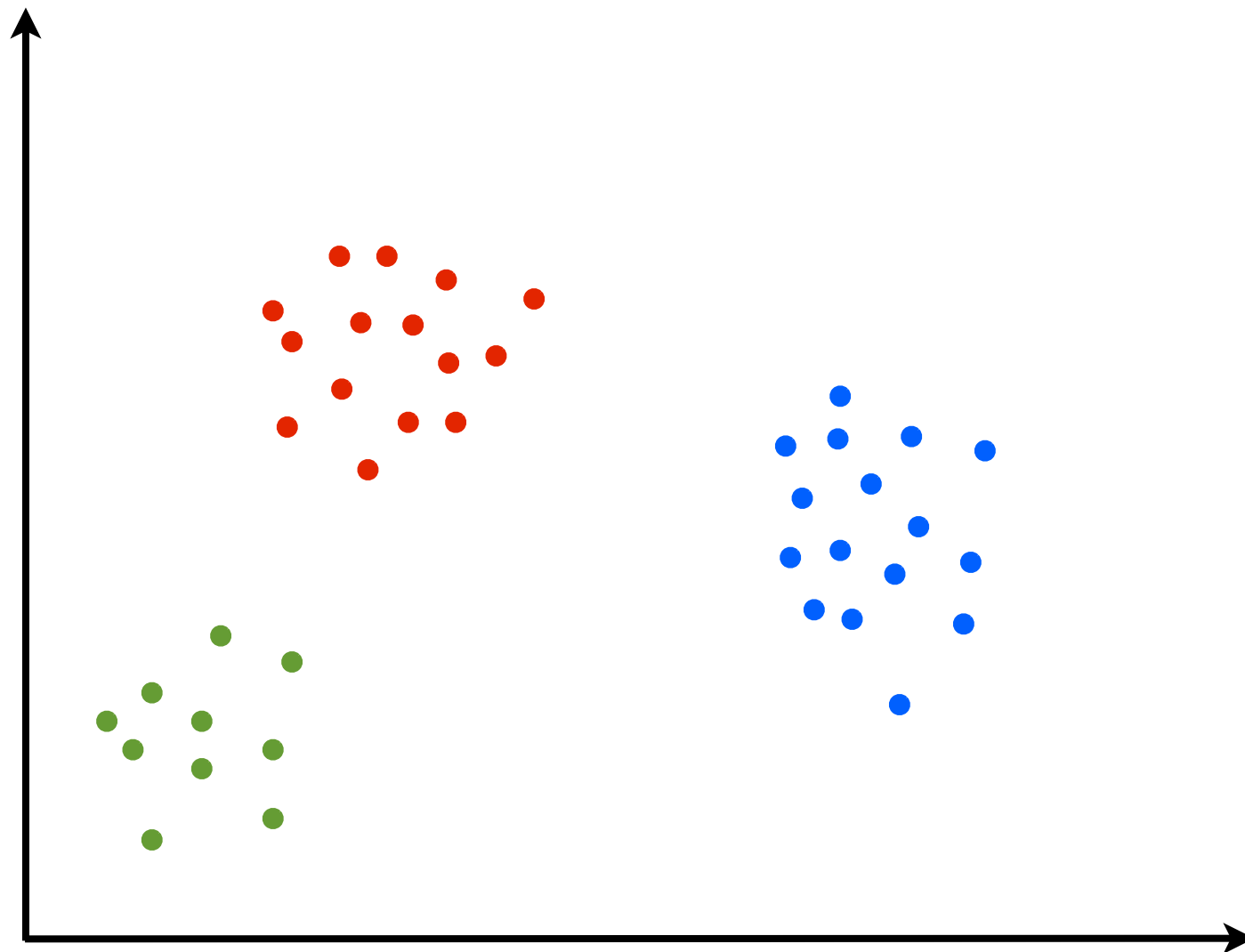
How to evaluate a clustering algorithm?

- Visualization
- Comparing clusterings:
 - ◇ Sum over all intra-cluster dissimilarities

Clustering: evaluation

How to evaluate a clustering algorithm?

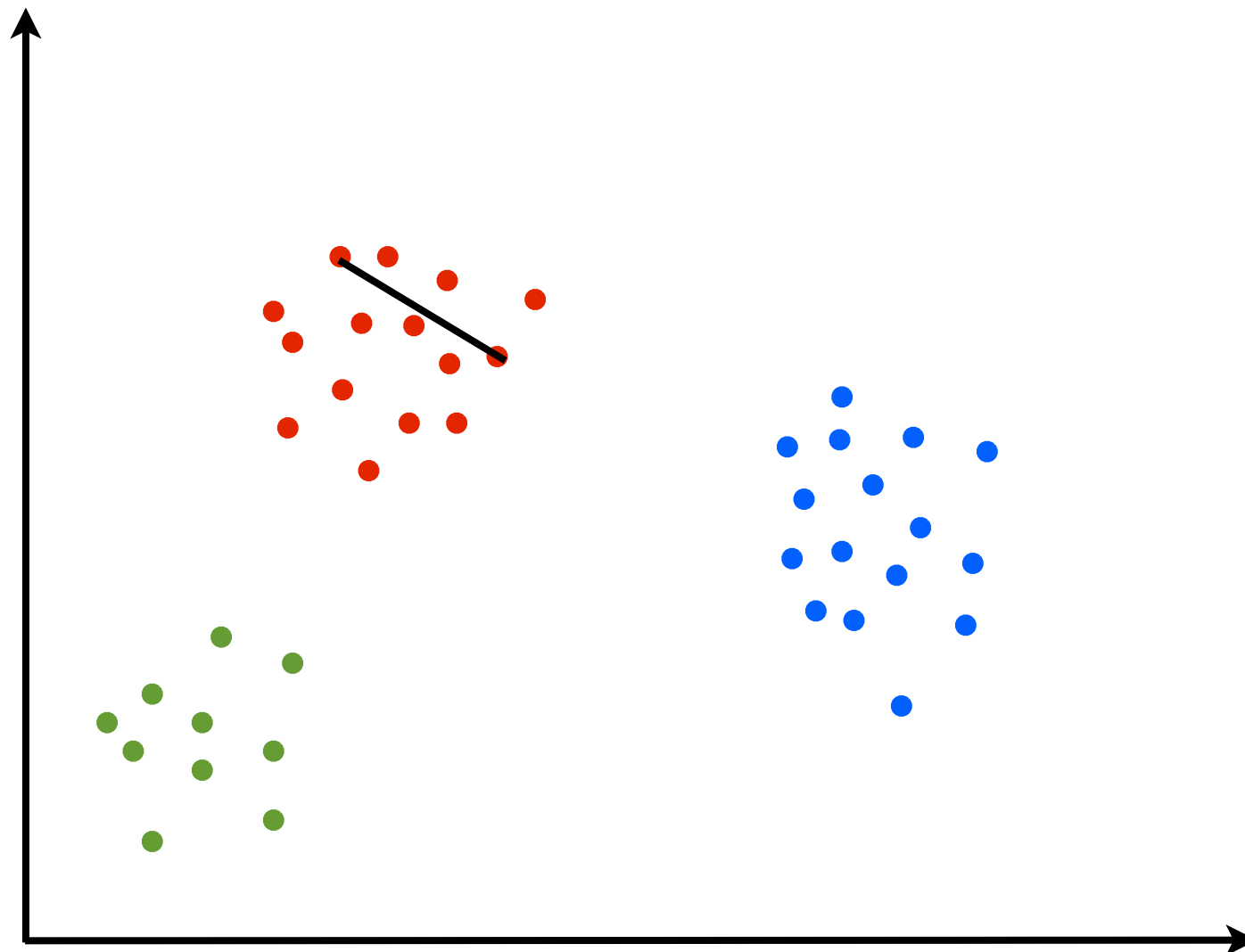
- Visualization
- Comparing clusterings:
 - ◇ Sum over all intra-cluster dissimilarities



Clustering: evaluation

How to evaluate a clustering algorithm?

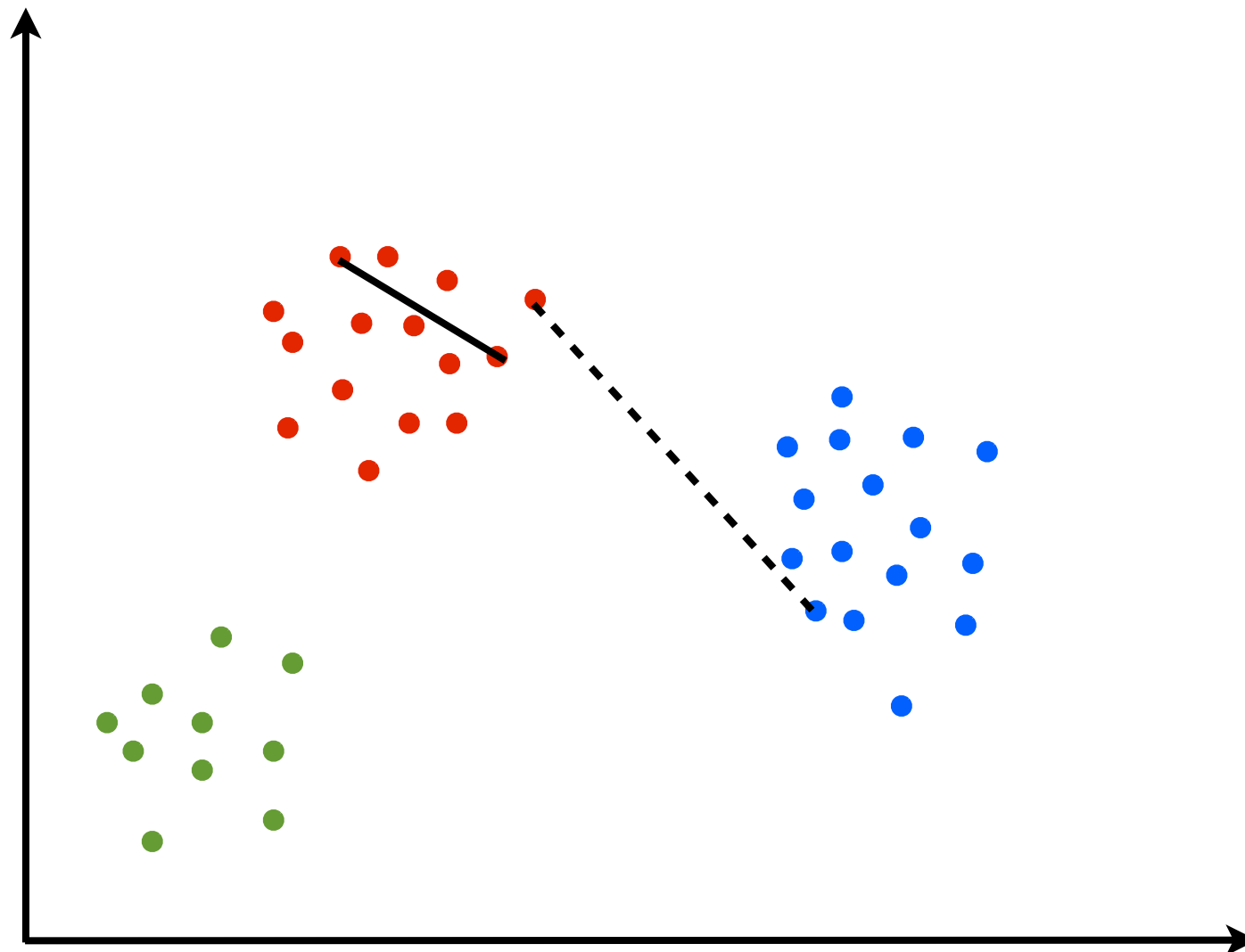
- Visualization
- Comparing clusterings:
 - ◇ Sum over all intra-cluster dissimilarities



Clustering: evaluation

How to evaluate a clustering algorithm?

- Visualization
- Comparing clusterings:
 - ◇ Sum over all intra-cluster dissimilarities



Clustering: evaluation

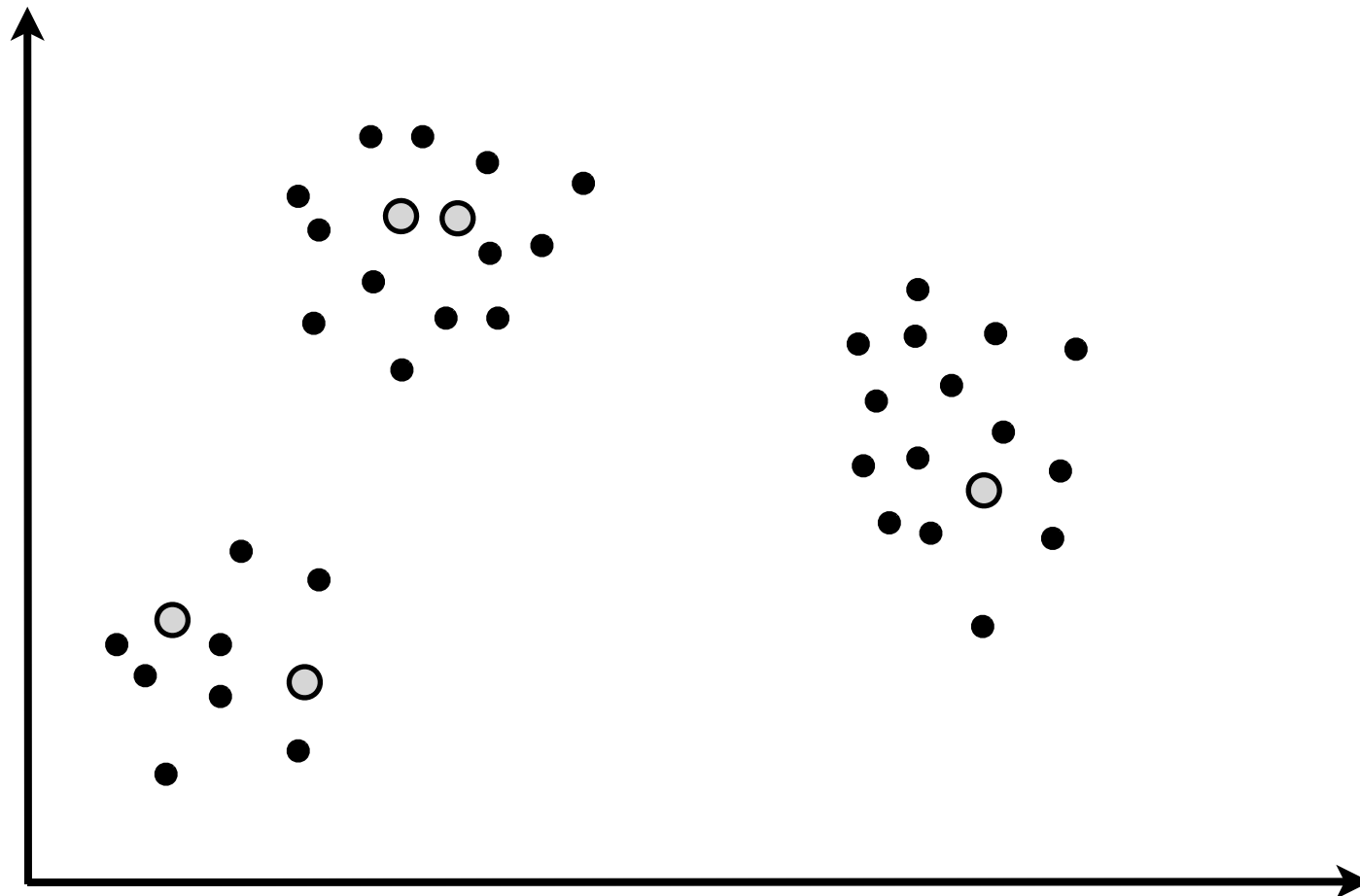
How to evaluate a clustering algorithm?

- Visualization
- Comparing clusterings:
 - ◇ Sum over all intra-cluster dissimilarities
 - ◇ Cross-validation

Clustering: evaluation

How to evaluate a clustering algorithm?

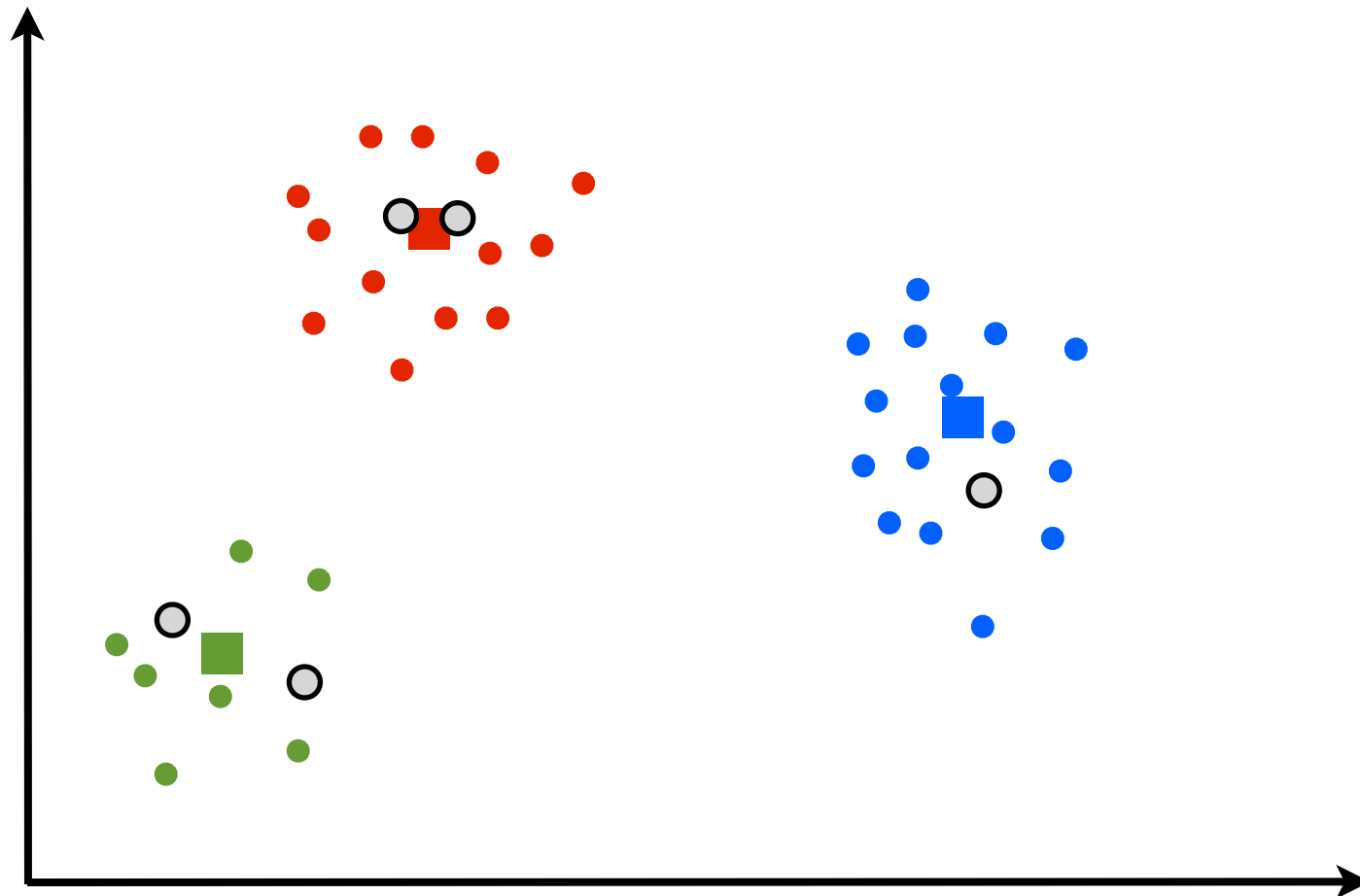
- Visualization
- Comparing clusterings:
 - ◇ Sum over all intra-cluster dissimilarities
 - ◇ Cross-validation



Clustering: evaluation

How to evaluate a clustering algorithm?

- Visualization
- Comparing clusterings:
 - ◇ Sum over all intra-cluster dissimilarities
 - ◇ Cross-validation



Clustering: evaluation

How to evaluate a clustering algorithm?

- Visualization
- Comparing clusterings:
 - ◇ Sum over all intra-cluster dissimilarities
 - ◇ Cross-validation
 - ◇ And many more: rand index, adjusted rand index, likelihood, domain-specific measures

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
- 2. Clustering evaluation**
3. Clustering trouble-shooting
4. Example

Outline

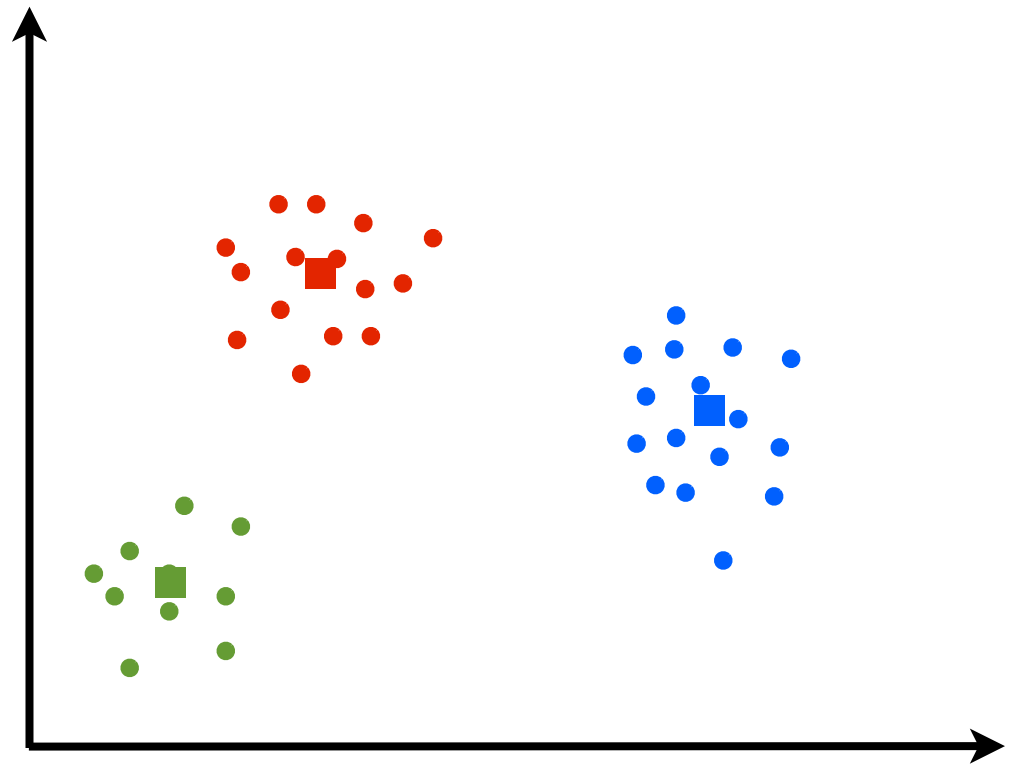
Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
- 3. Clustering trouble-shooting**
4. Example

K means algorithm

Benefits

- Fast
- Conceptually straightforward
- Popular

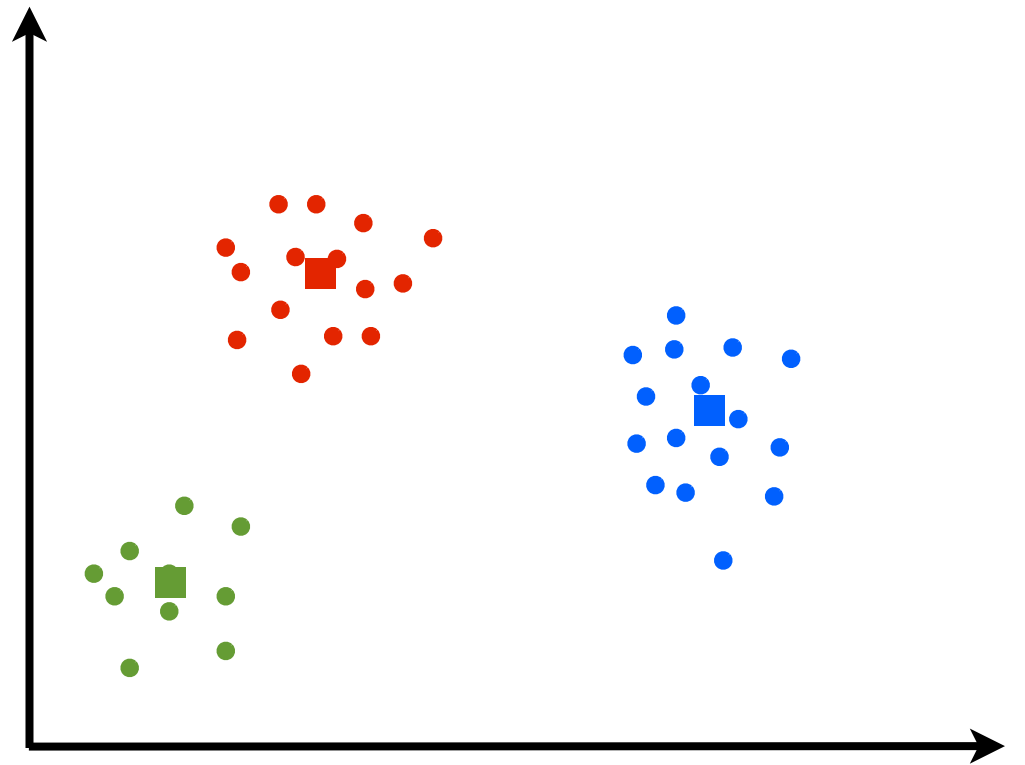


K means algorithm

Benefits

- Fast
- Conceptually straightforward
- Popular

Trouble-shooting



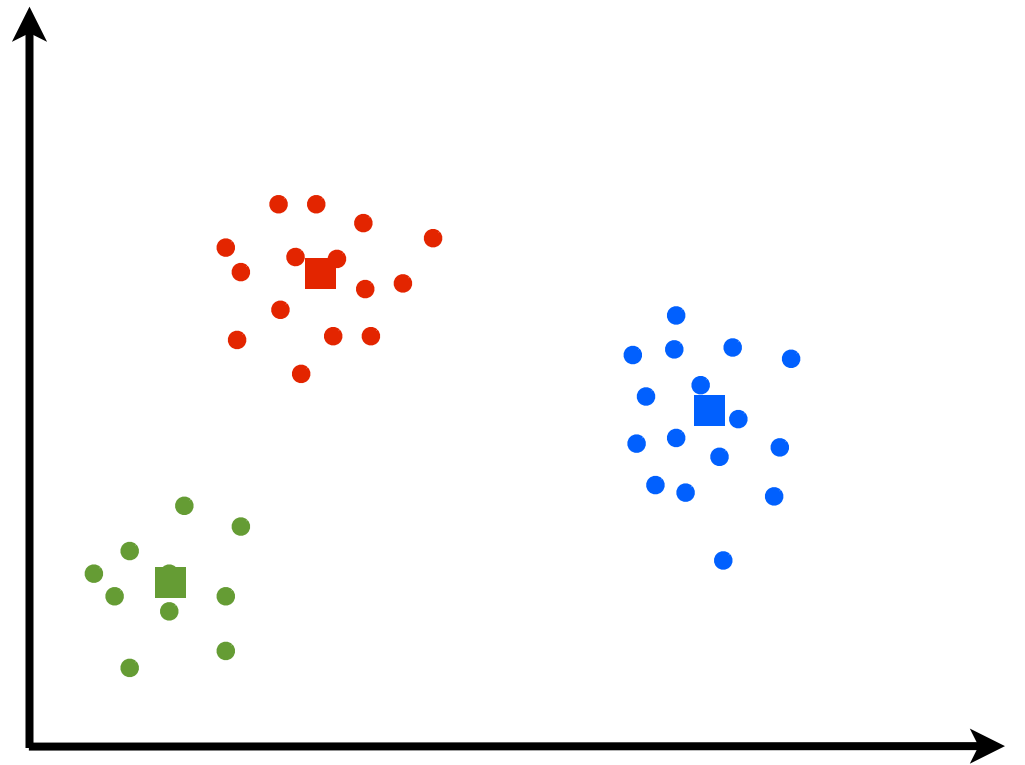
K means algorithm

Benefits

- Fast
- Conceptually straightforward
- Popular

Trouble-shooting

- Still not fast enough!



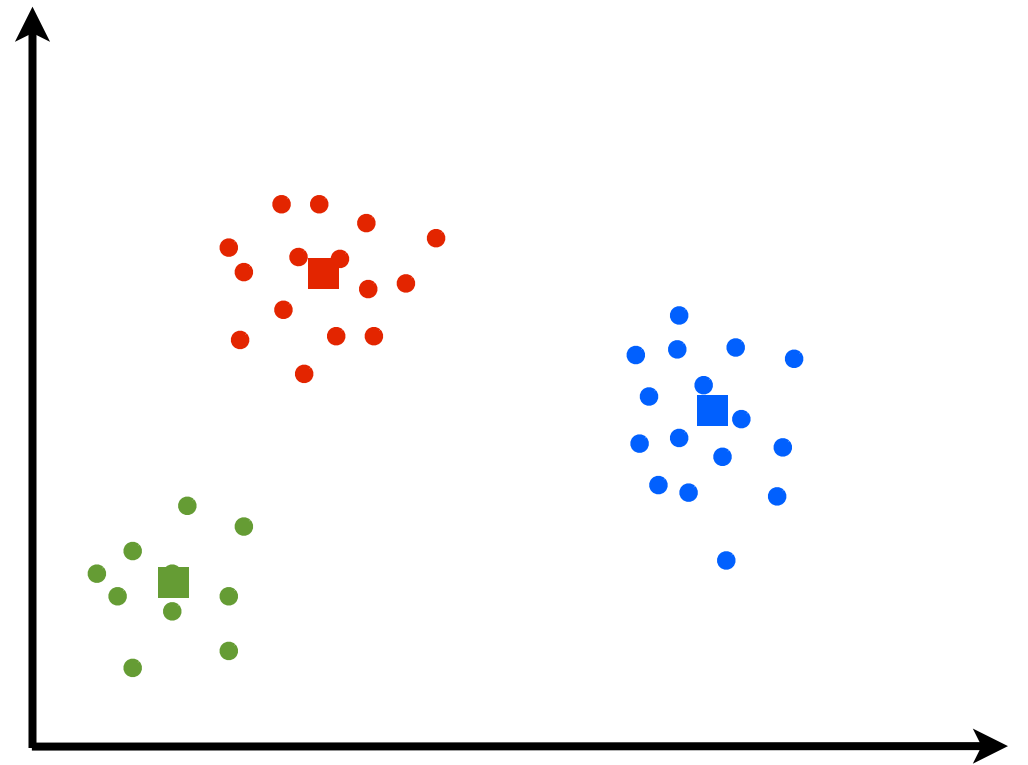
K means algorithm

Benefits

- Fast
- Conceptually straightforward
- Popular

Trouble-shooting

- Still not fast enough!
 - ◇ KD-trees, triangle inequality, online version



[Ramasubramanian, Paliwal 1990;
Moore 2000; Kanungo et al 2002]

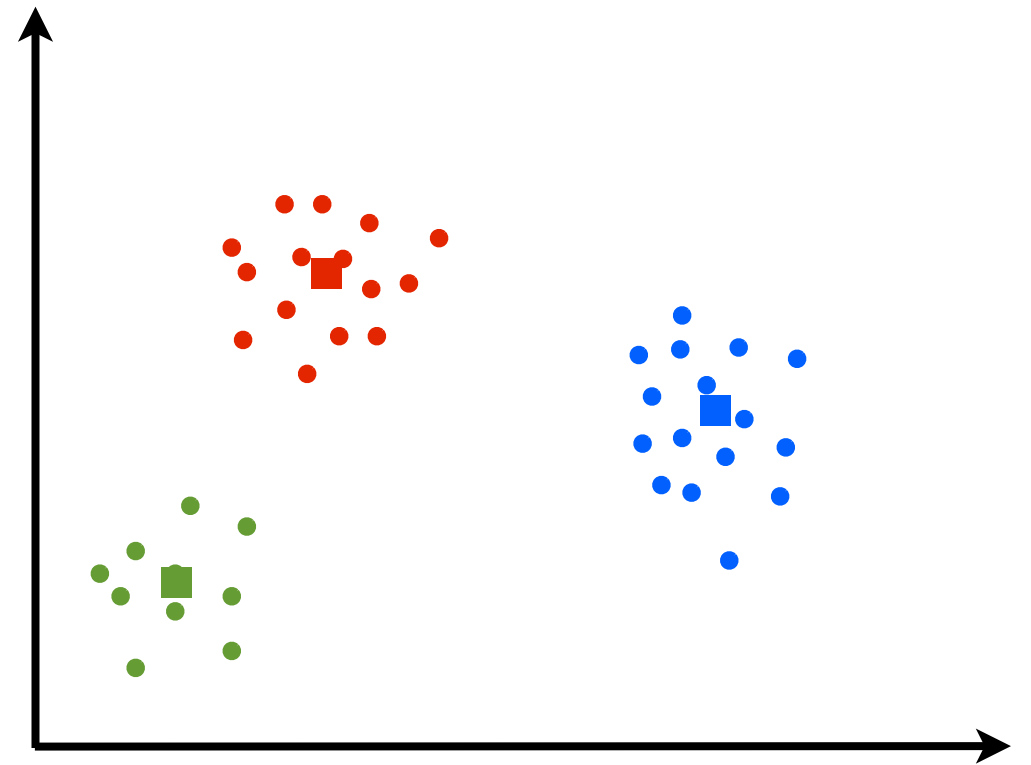
K means algorithm

Benefits

- Fast
- Conceptually straightforward
- Popular

Trouble-shooting

- Still not fast enough!
 - ◇ KD-trees, triangle inequality, online version
- Only finds a local optimum



[Ramasubramanian, Paliwal 1990;
Moore 2000; Kanungo et al 2002]

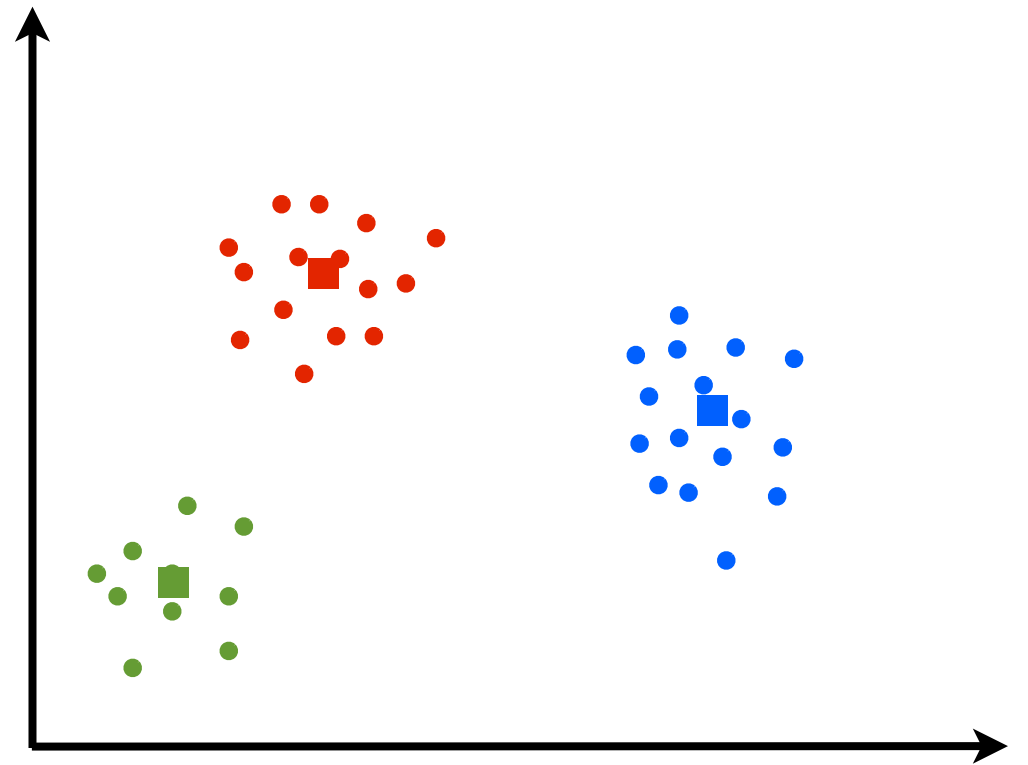
K means algorithm

Benefits

- Fast
- Conceptually straightforward
- Popular

Trouble-shooting

- Still not fast enough!
 - ◇ KD-trees, triangle inequality, online version
- Only finds a local optimum
 - ◇ Multiple initializations



[Ramasubramanian, Paliwal 1990;
Moore 2000; Kanungo et al 2002]

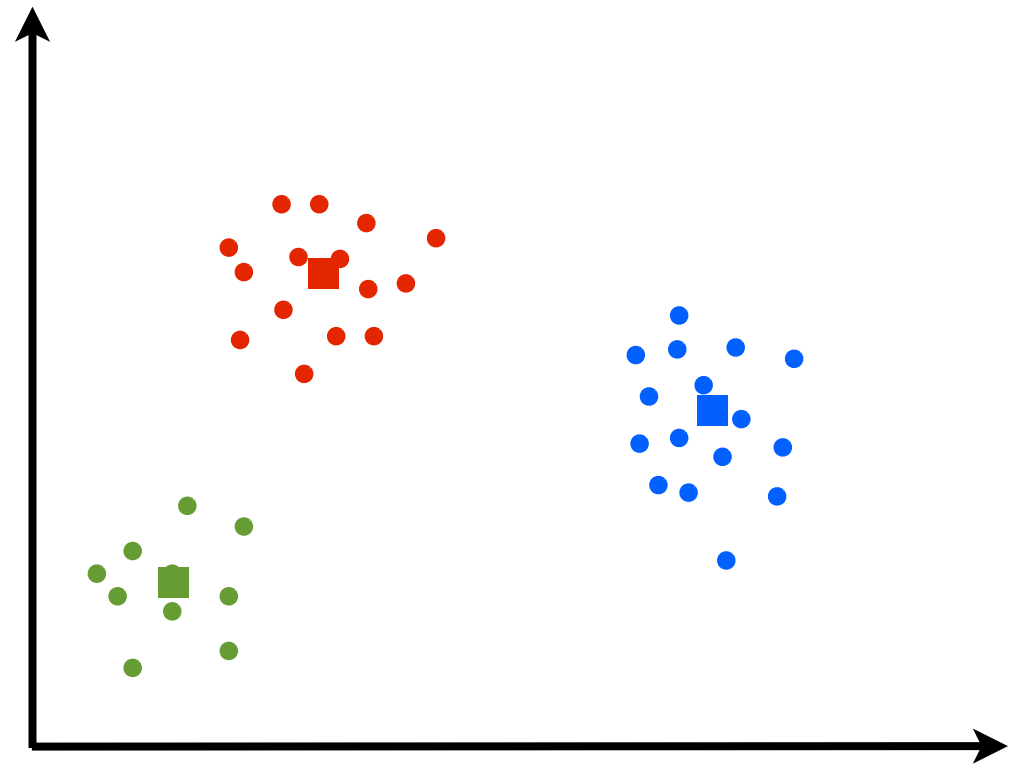
K means algorithm

Benefits

- Fast
- Conceptually straightforward
- Popular

Trouble-shooting

- Still not fast enough!
 - ◇ KD-trees, triangle inequality, online version
- Only finds a local optimum
 - ◇ Multiple initializations
- May not fit the problem...



[Ramasubramanian, Paliwal 1990;
Moore 2000; Kanungo et al 2002]

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
- 3. Clustering trouble-shooting**
4. Example

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
- 3. Clustering trouble-shooting**
 - Grouping
 - Similarity
 - Data
4. Example

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
- 3. Clustering trouble-shooting**
 - **Grouping**
 - Similarity
 - Data
4. Example

What is a cluster?

Hard clustering

- K fixed

What is a cluster?

Hard clustering

- K fixed

Image compression

$K = 2$



$K = 3$



$K = 10$



Original image



What is a cluster?

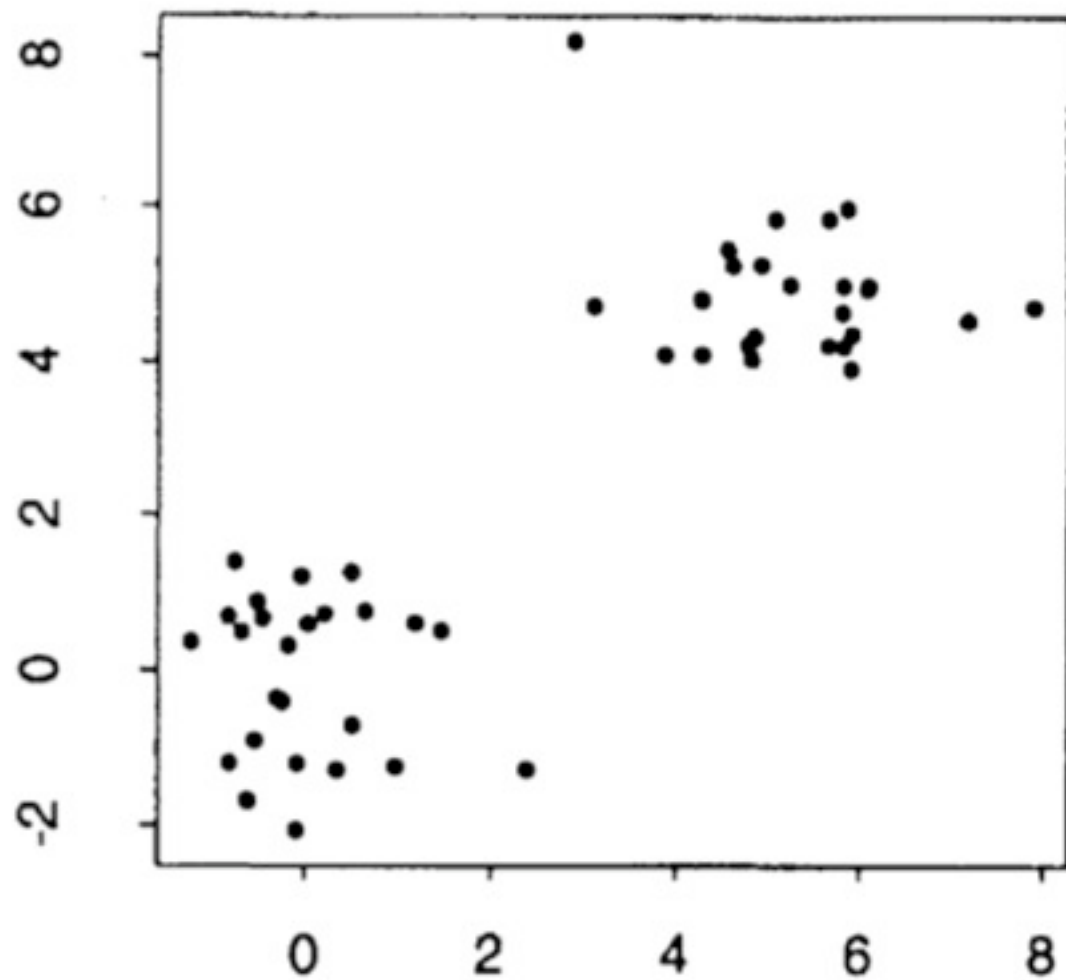
Hard clustering

- K fixed
- K unknown

What is a cluster?

Hard clustering

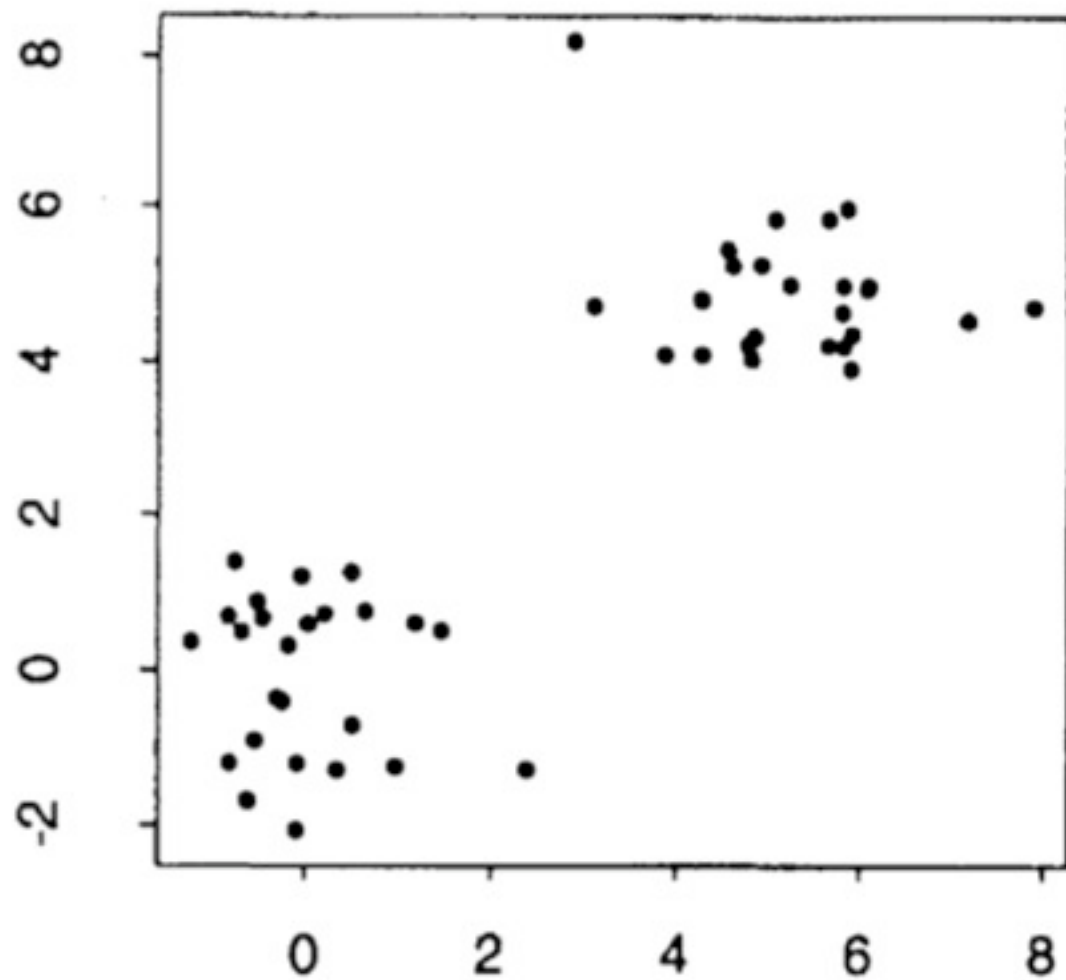
- K fixed
- K unknown



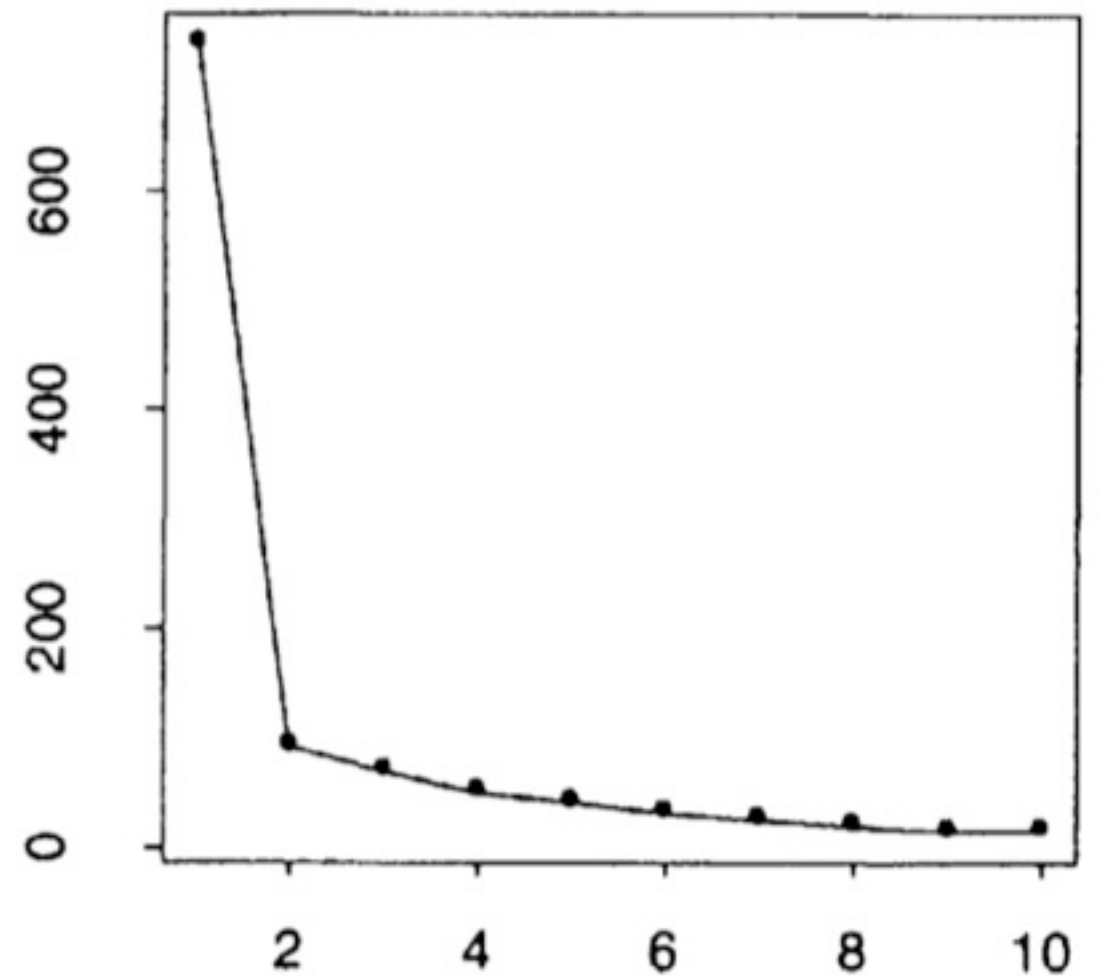
What is a cluster?

Hard clustering

- K fixed
- K unknown



Global dissimilarity



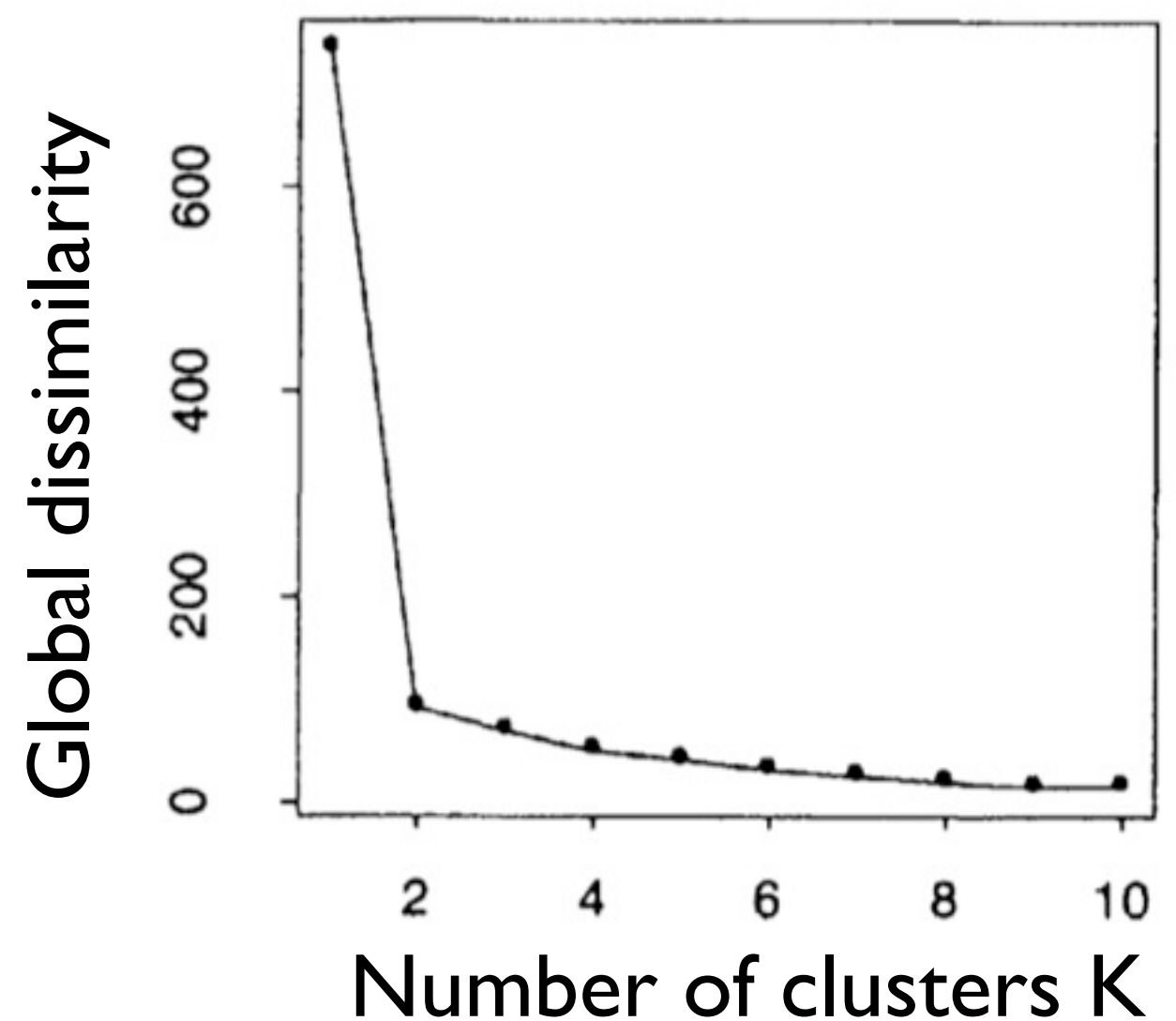
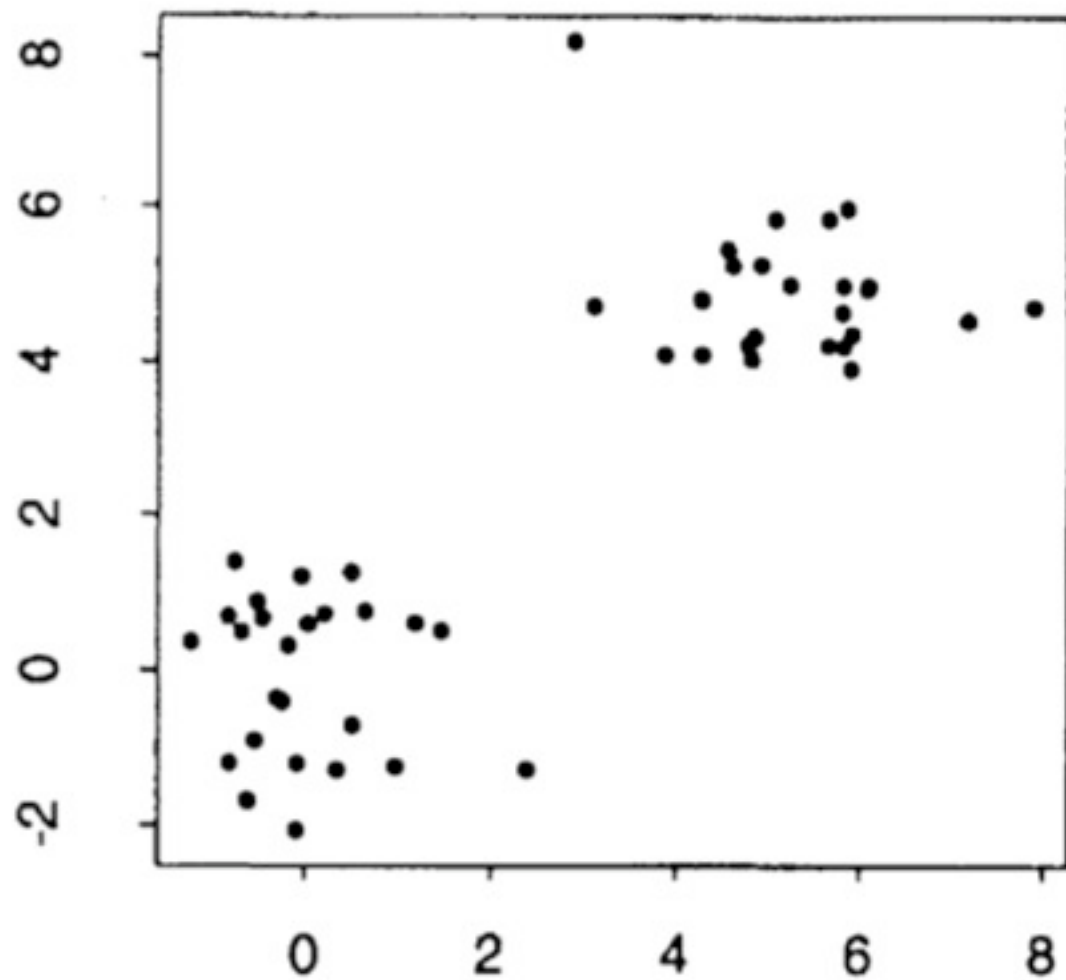
Number of clusters K

What is a cluster?

Hard clustering

- K fixed
- K unknown
 - ◇ Heuristic methods: elbow, gap statistic

[Tibshirani et al 2001]

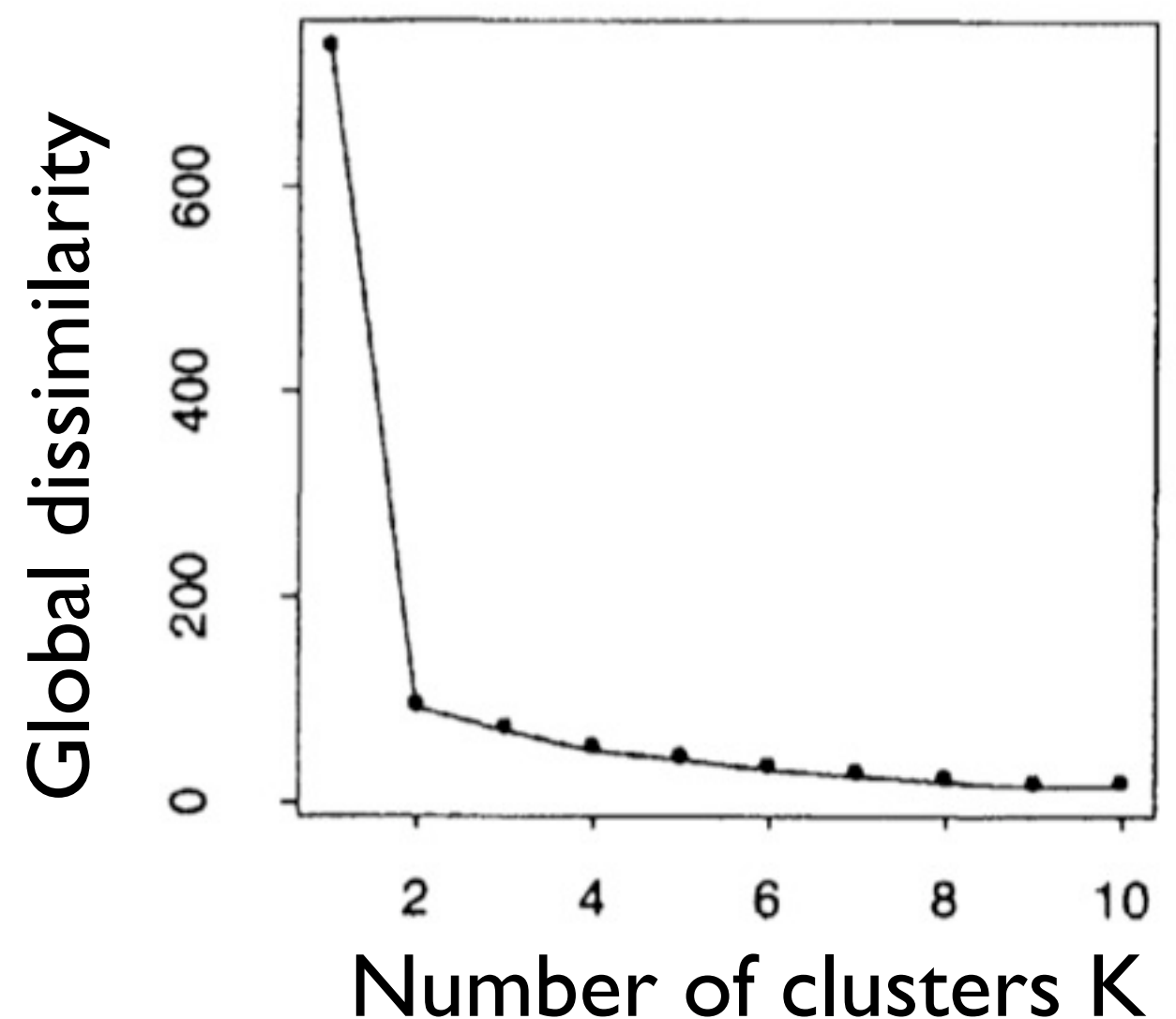
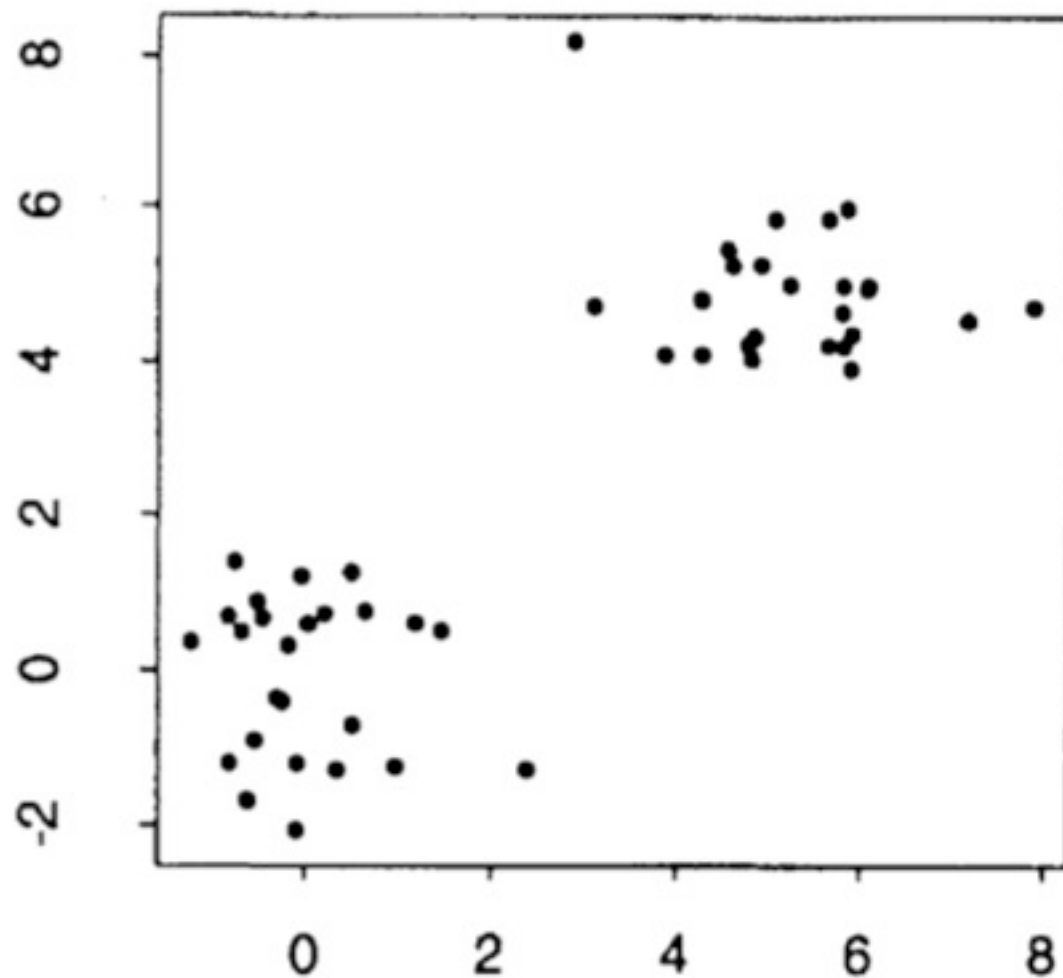


What is a cluster?

Hard clustering

- K fixed
- K unknown
 - ◇ Heuristic methods: elbow, gap statistic
 - ◇ Optimization methods: AIC, BIC, DP means

[Kulis, Jordan 2012]

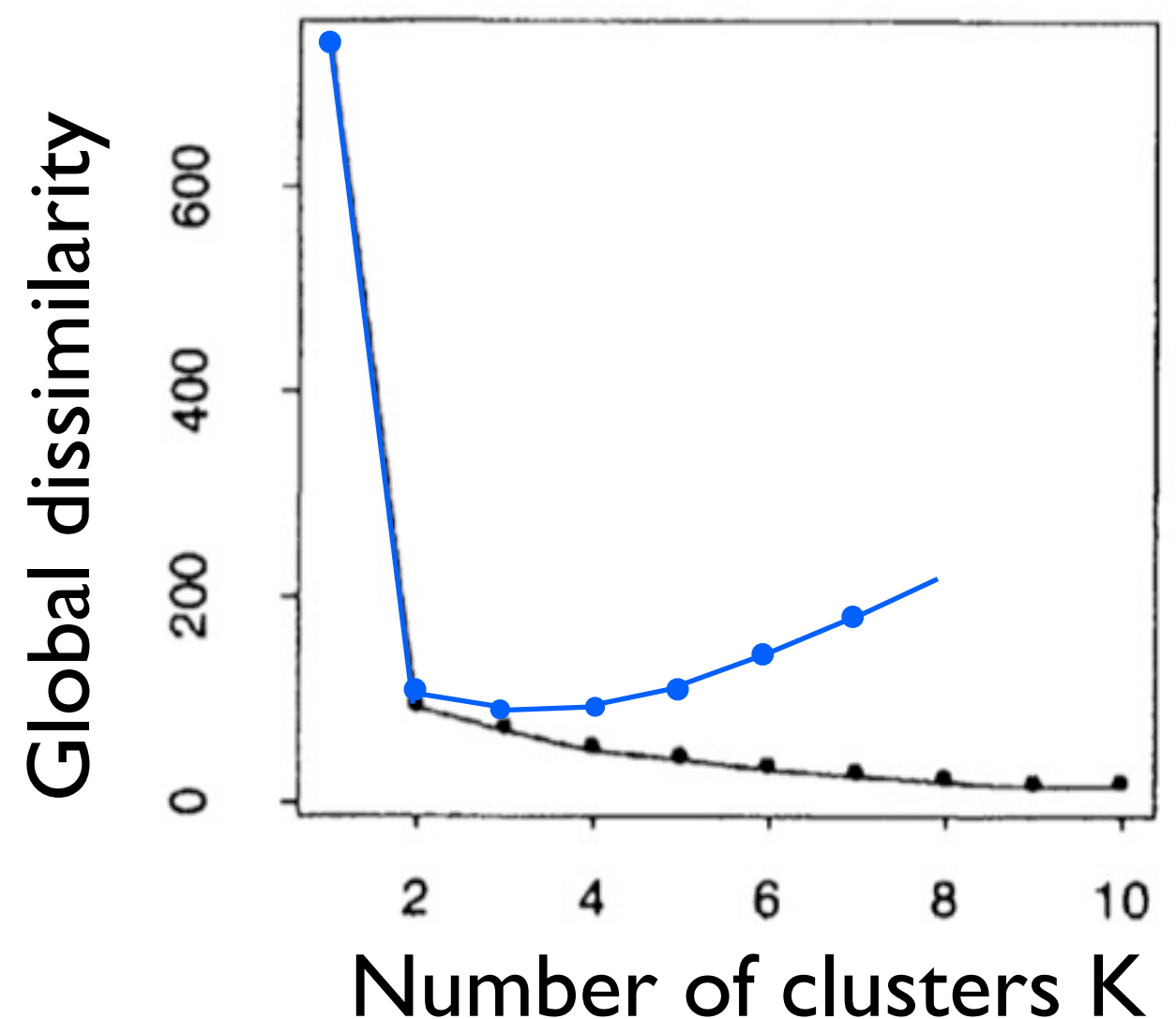
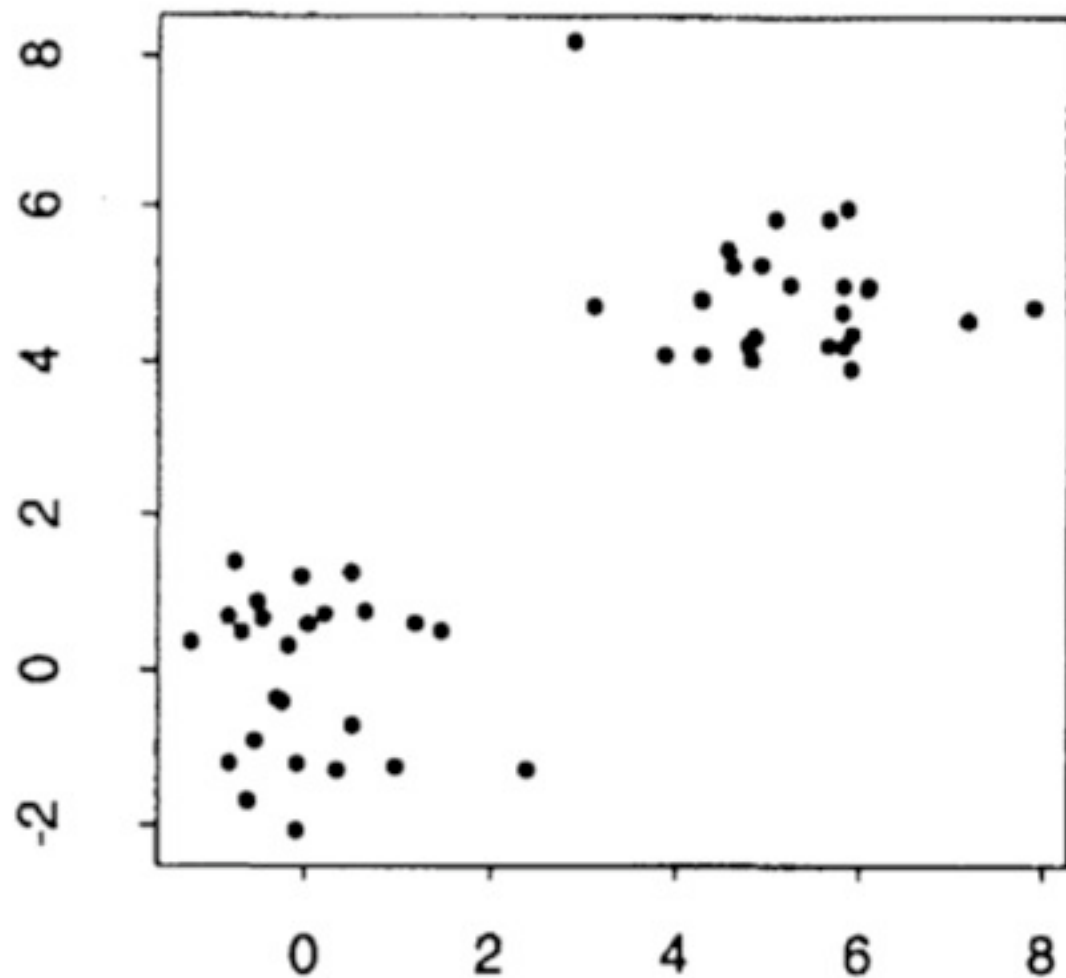


What is a cluster?

Hard clustering

- K fixed
- K unknown
 - ◇ Heuristic methods: elbow, gap statistic
 - ◇ Optimization methods: AIC, BIC, DP means

[Kulis, Jordan 2012]



What is a cluster?

Hard clustering

- K fixed
- K unknown
 - ◇ Heuristic methods: elbow, gap statistic
 - ◇ Optimization methods: AIC, BIC, DP means
 - ◇ Model-based methods: Bayesian prior, Dirichlet process

[Teh 2010; Richardson, Green 1997]

What is a cluster?

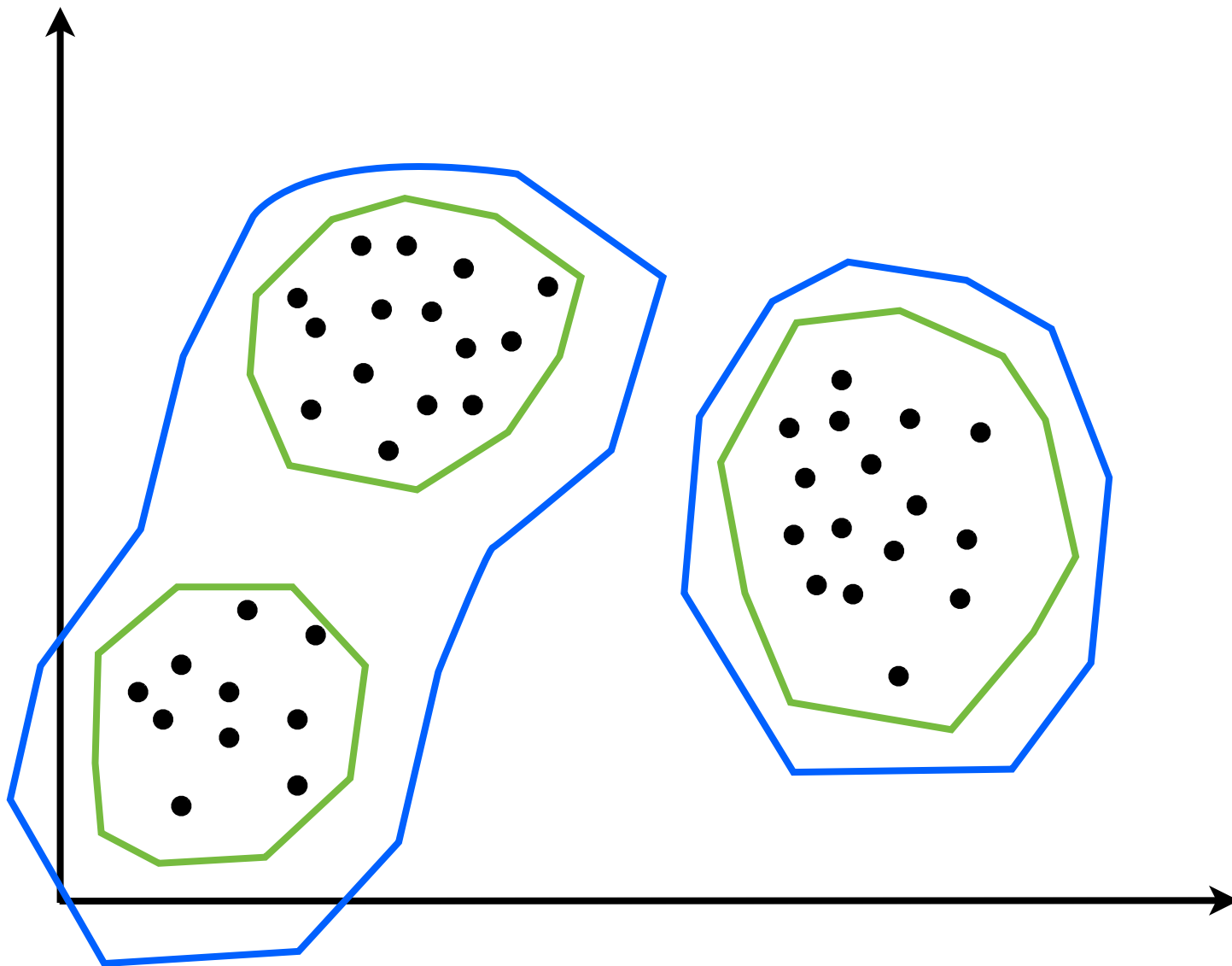
Hard clustering

- K fixed
- K unknown
- Clustering “consistent” across different K

What is a cluster?

Hard clustering

- K fixed
- K unknown
- Clustering “consistent” across different K
 - ◇ Hierarchical clustering, agglomerative clustering



What is a cluster?

Hard clustering

- K fixed
- K unknown
- Clustering “consistent” across different K

Soft clustering

What is a cluster?

Hard clustering

- K fixed
- K unknown
- Clustering “consistent” across different K

Soft clustering

- Different degrees of membership for different data points

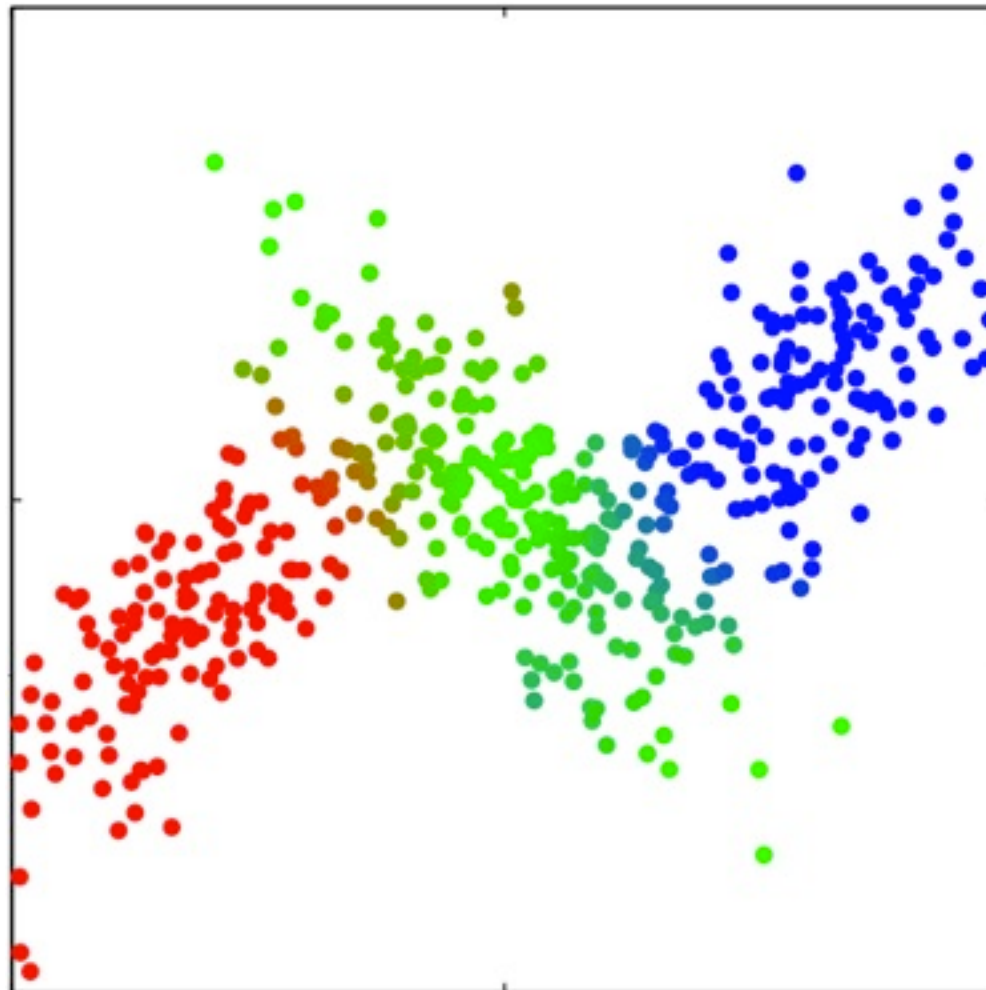
What is a cluster?

Hard clustering

- K fixed
- K unknown
- Clustering “consistent” across different K

Soft clustering

- Different degrees of membership for different data points



What is a cluster?

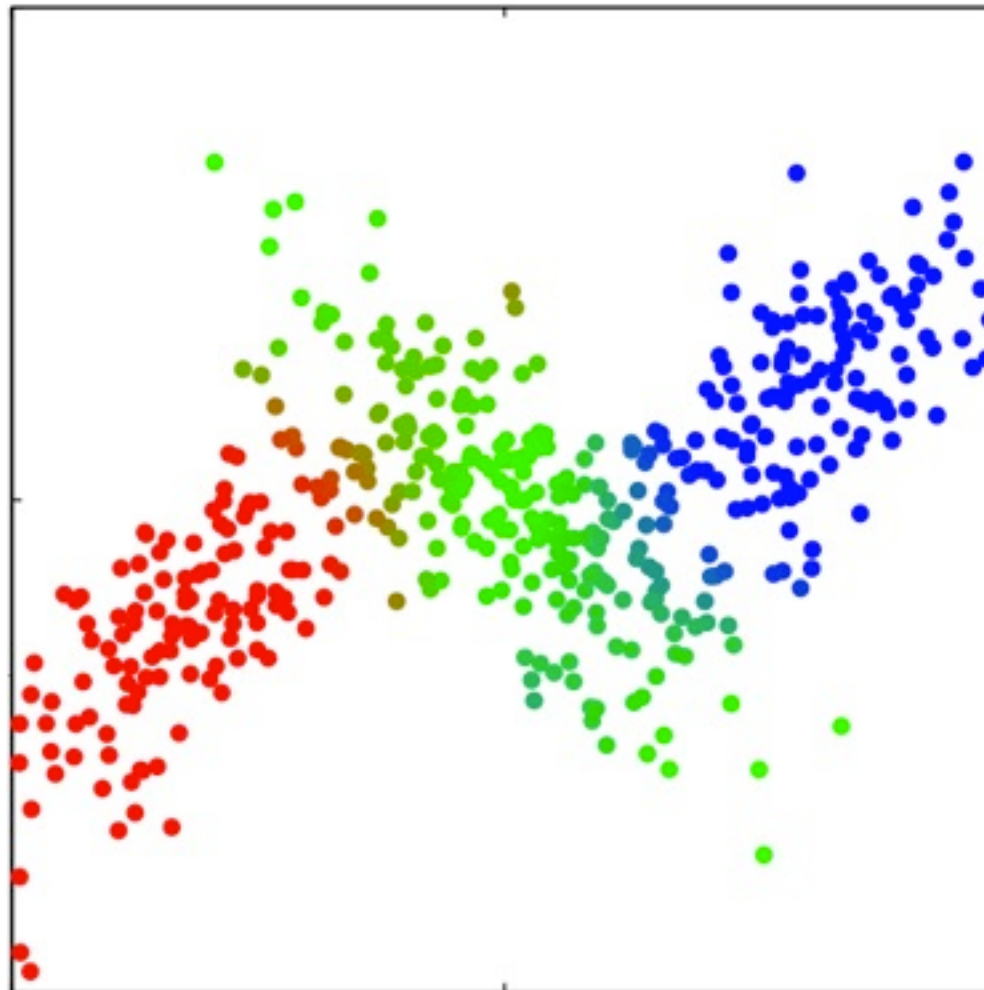
Hard clustering

- K fixed
- K unknown
- Clustering “consistent” across different K

Soft clustering

- Different degrees of membership for different data points

◇ Fuzzy c means,
(Gaussian) mixture
models



Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
- 3. Clustering trouble-shooting**
 - **Grouping**
 - Similarity
 - Data
4. Example

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
- 3. Clustering trouble-shooting**
 - Grouping
 - **Similarity**
 - Data
4. Example

How to measure (dis)similarity?

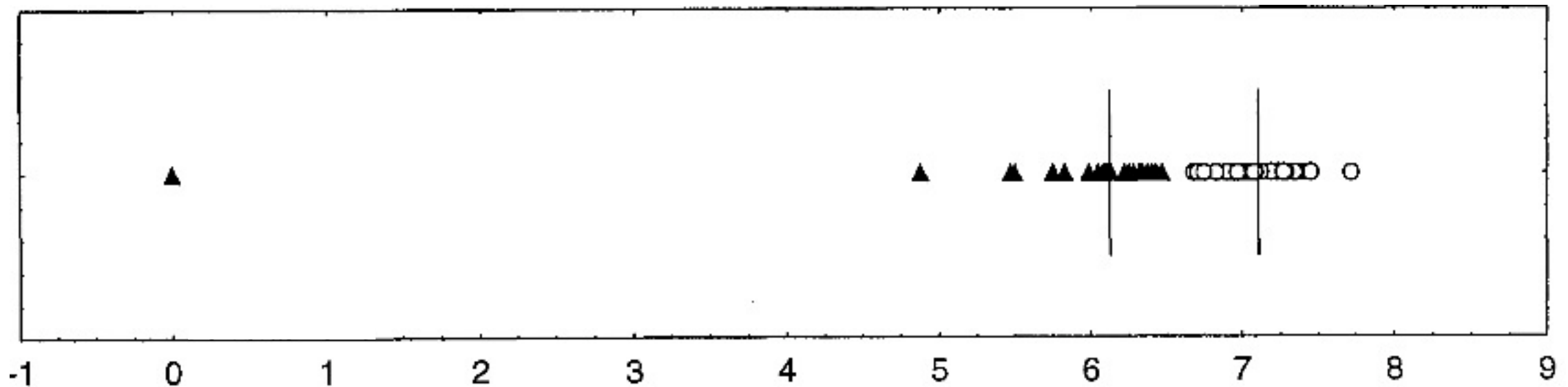
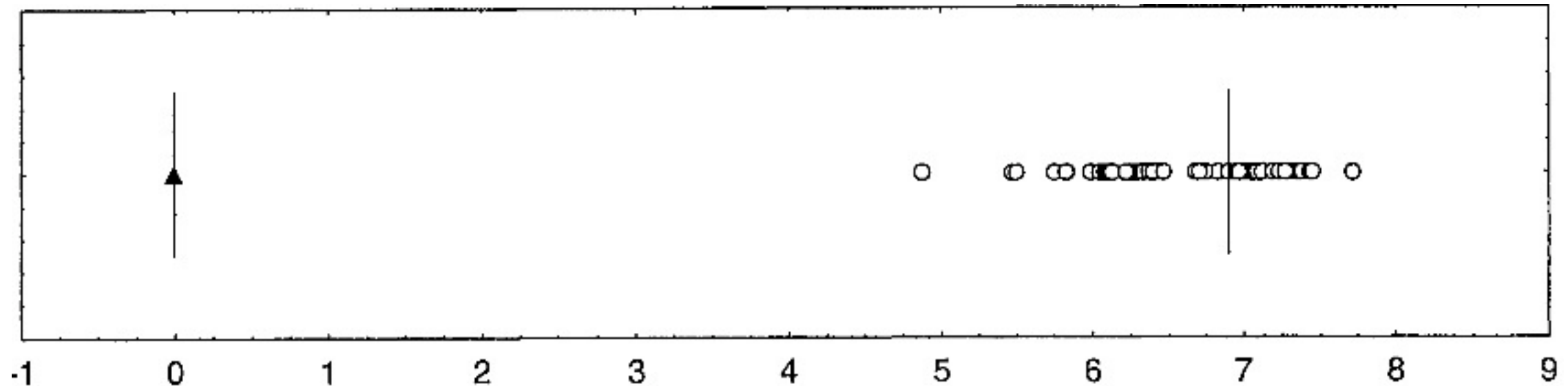
K means

- Sensitive to outliers

How to measure (dis)similarity?

K means

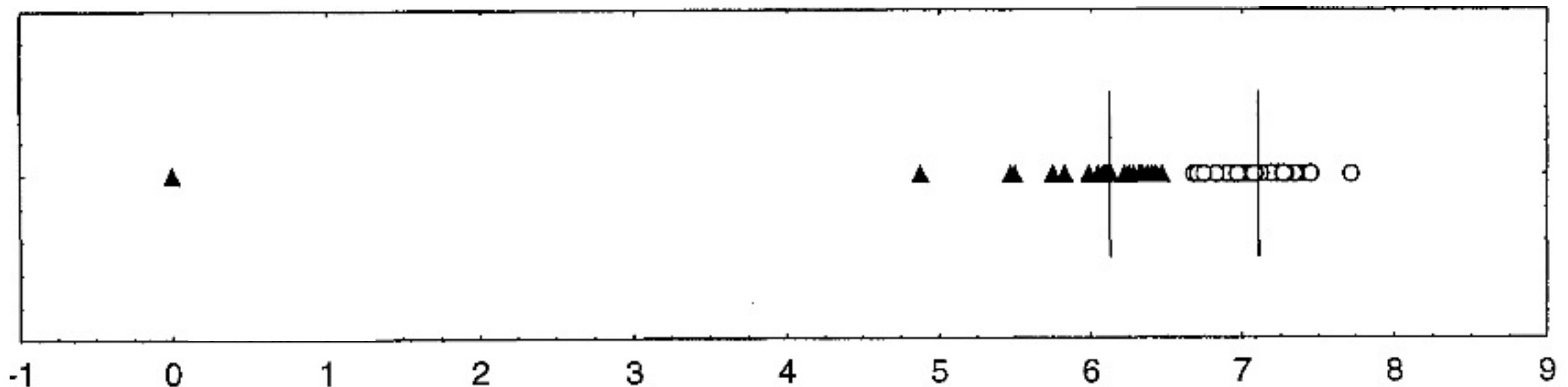
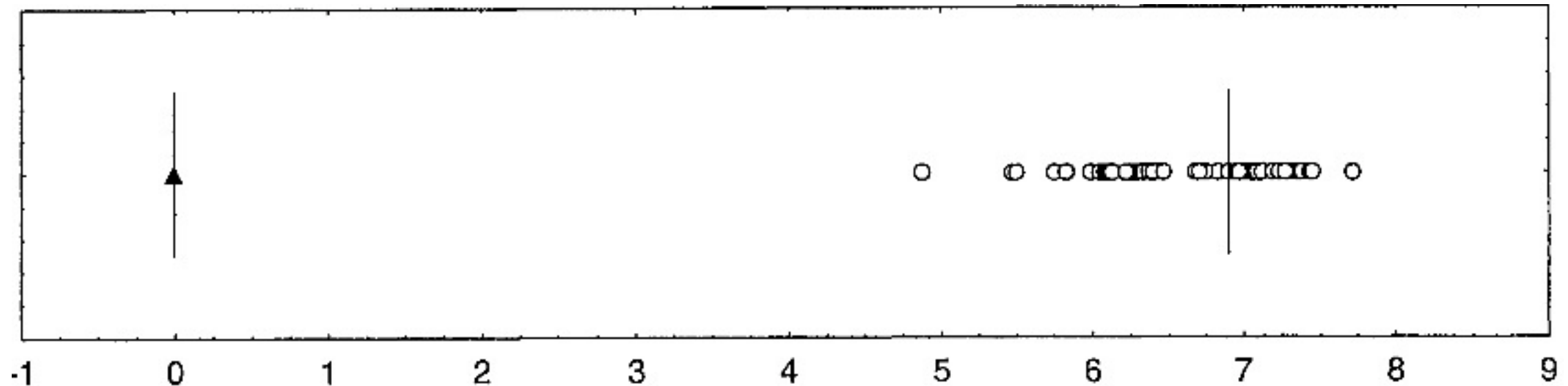
- Sensitive to outliers



How to measure (dis)similarity?

K means

- Sensitive to outliers
 - ◇ K medoids



How to measure (dis)similarity?

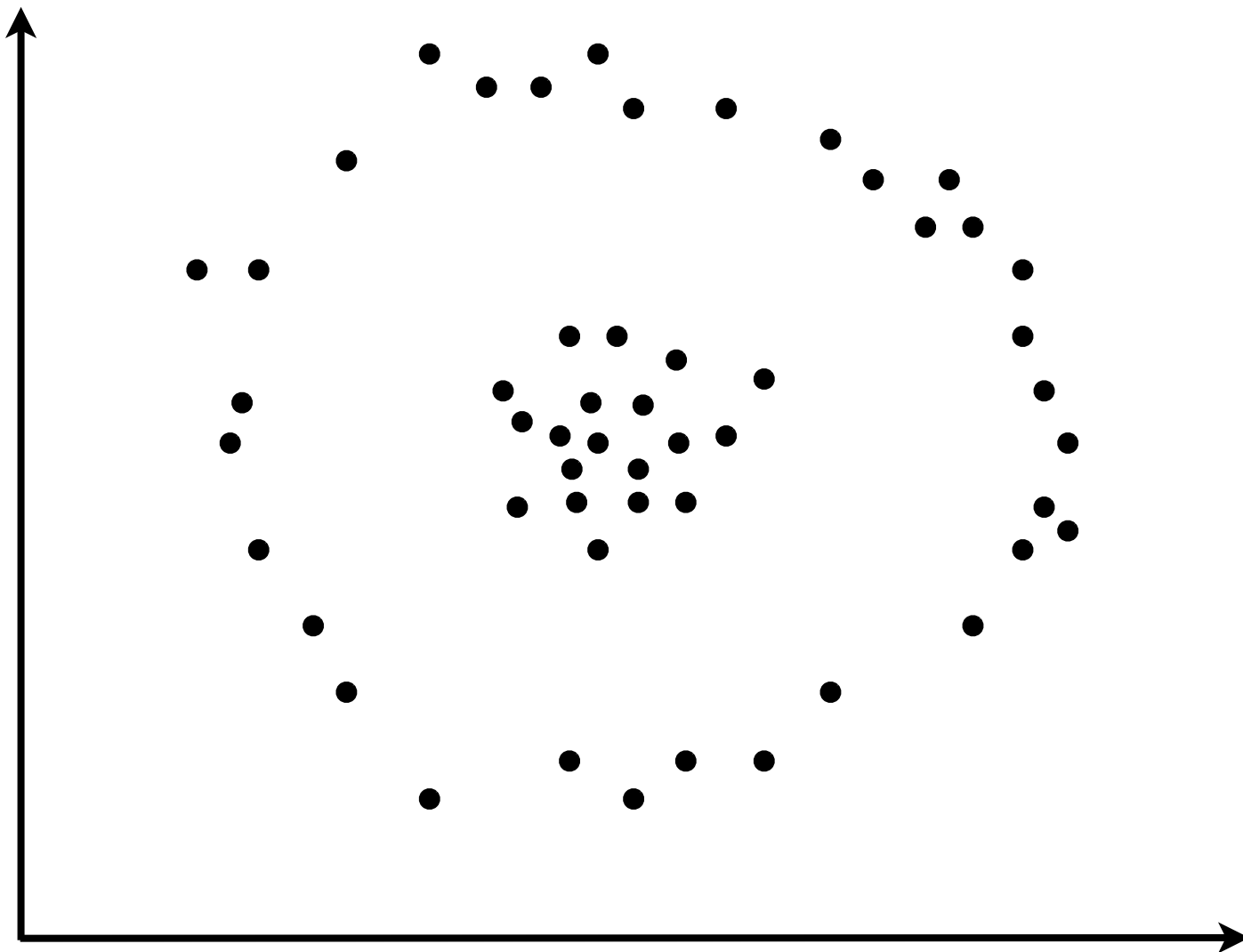
K means

- Sensitive to outliers
 - ◇ K medoids
- Yields spherical clusters

How to measure (dis)similarity?

K means

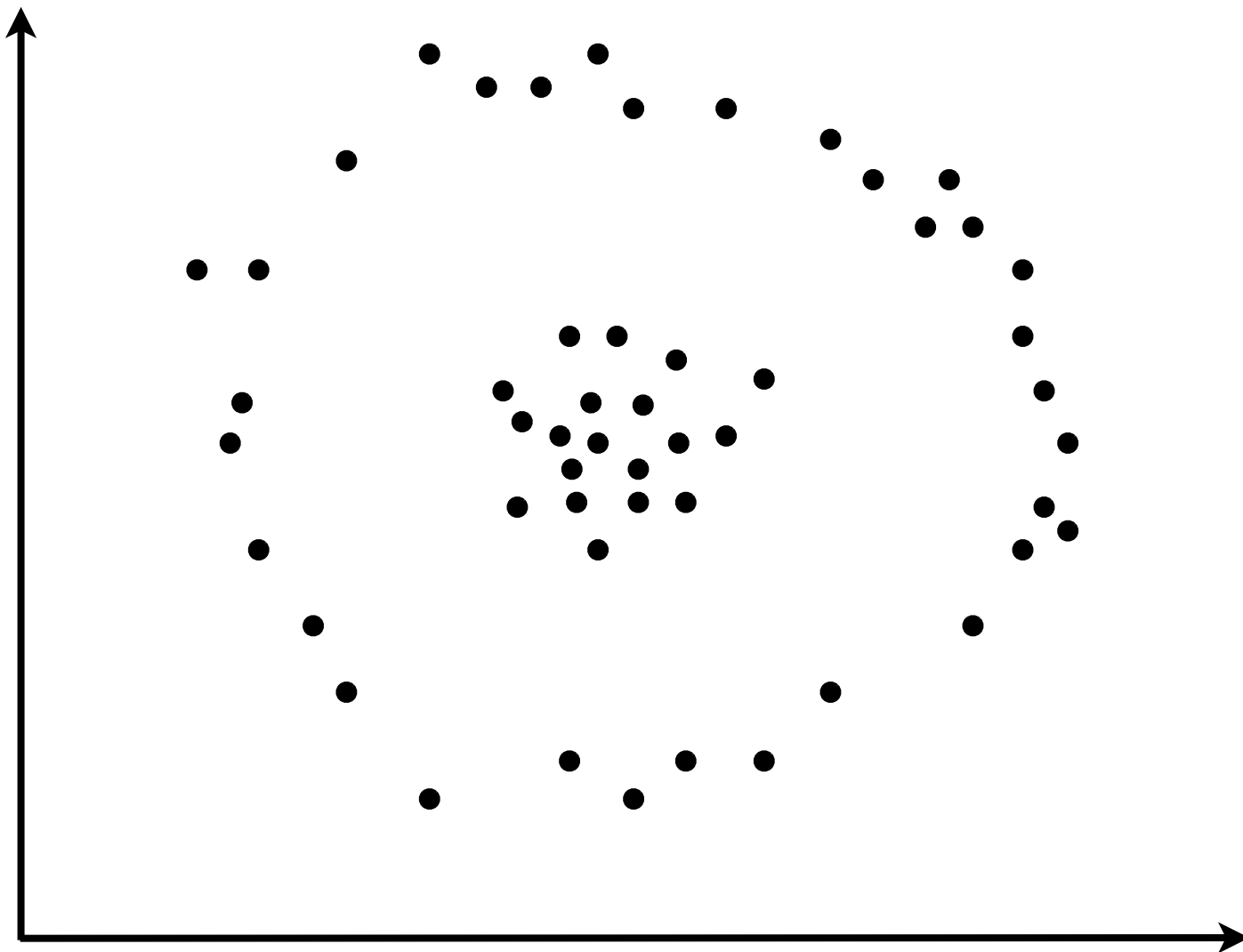
- Sensitive to outliers
 - ◇ K medoids
- Yields spherical clusters



How to measure (dis)similarity?

K means

- Sensitive to outliers
 - ◇ K medoids
- Yields spherical clusters
 - ◇ Radial similarity, polar coordinates, agglomerative cl.



How to measure (dis)similarity?

K means

- Sensitive to outliers
 - ◇ K medoids
- Yields spherical clusters
 - ◇ Radial similarity, transform data, agglomerative clust.
- Requires continuous, numerical features

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
- 3. Clustering trouble-shooting**
 - Grouping
 - Similarity**
 - Data
4. Example

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
- 3. Clustering trouble-shooting**
 - Grouping
 - Similarity
 - Data**
4. Example

Data pre-processing

- Is the data set featurized?

Data pre-processing

Person
275

Age: 45

Height: 5' 7"

Residence:
Urban

Education:
Bachelor's

Tweet: "Just landed in
Iceland. Remember
Eyjafjallajökull?"

Data pre-processing

Person 275	Age	Height	Education Level	Tweets about Eyjafjallajökull	...
	⋮		⋮		
	45	5' 7"	Bachelor's	5	...
	⋮		⋮		

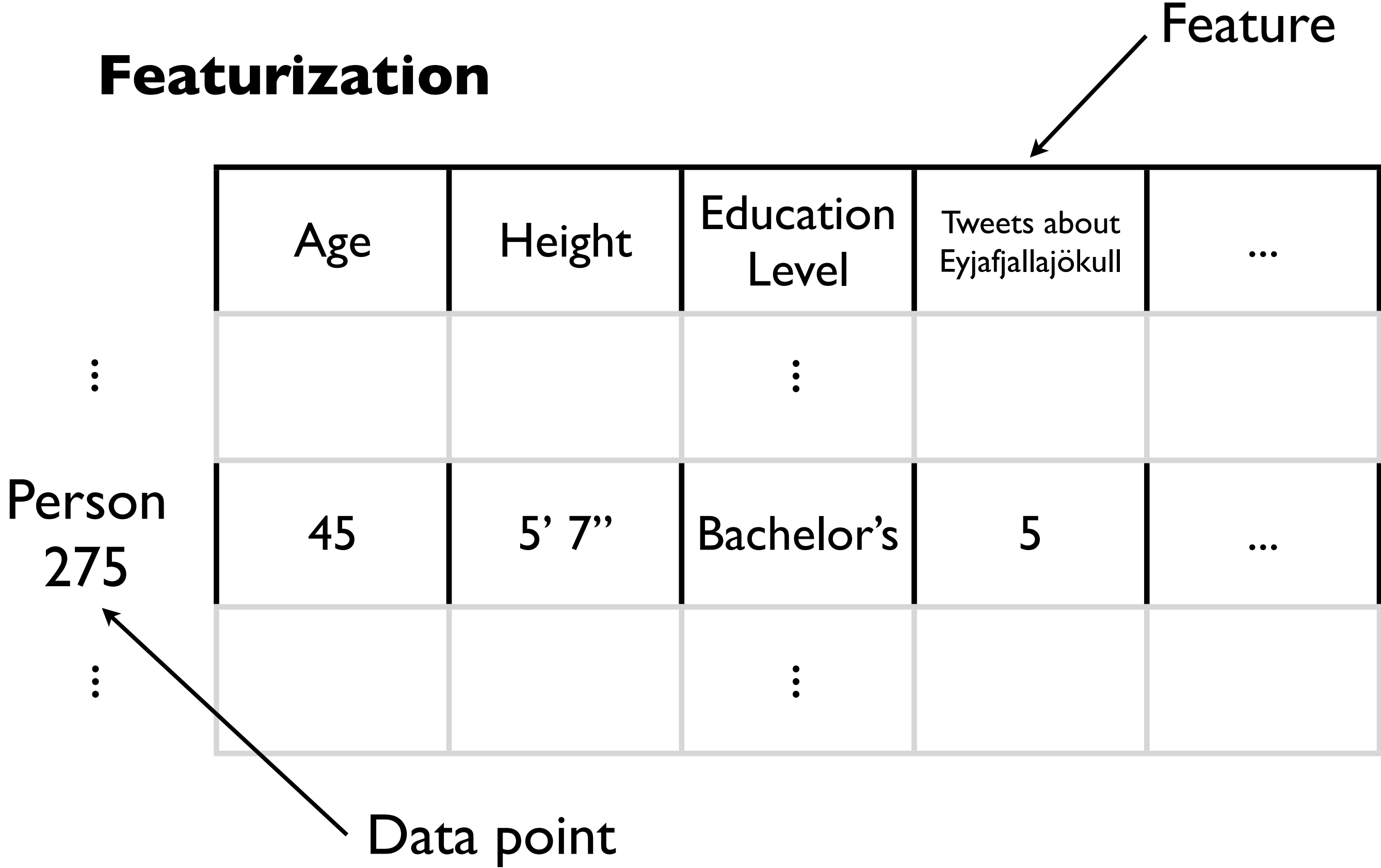
Data pre-processing

Featurization

Person 275		Age	Height	Education Level	Tweets about Eyjafjallajökull	...
	⋮			⋮		
		45	5' 7"	Bachelor's	5	...
	⋮			⋮		

Data pre-processing

Featurization



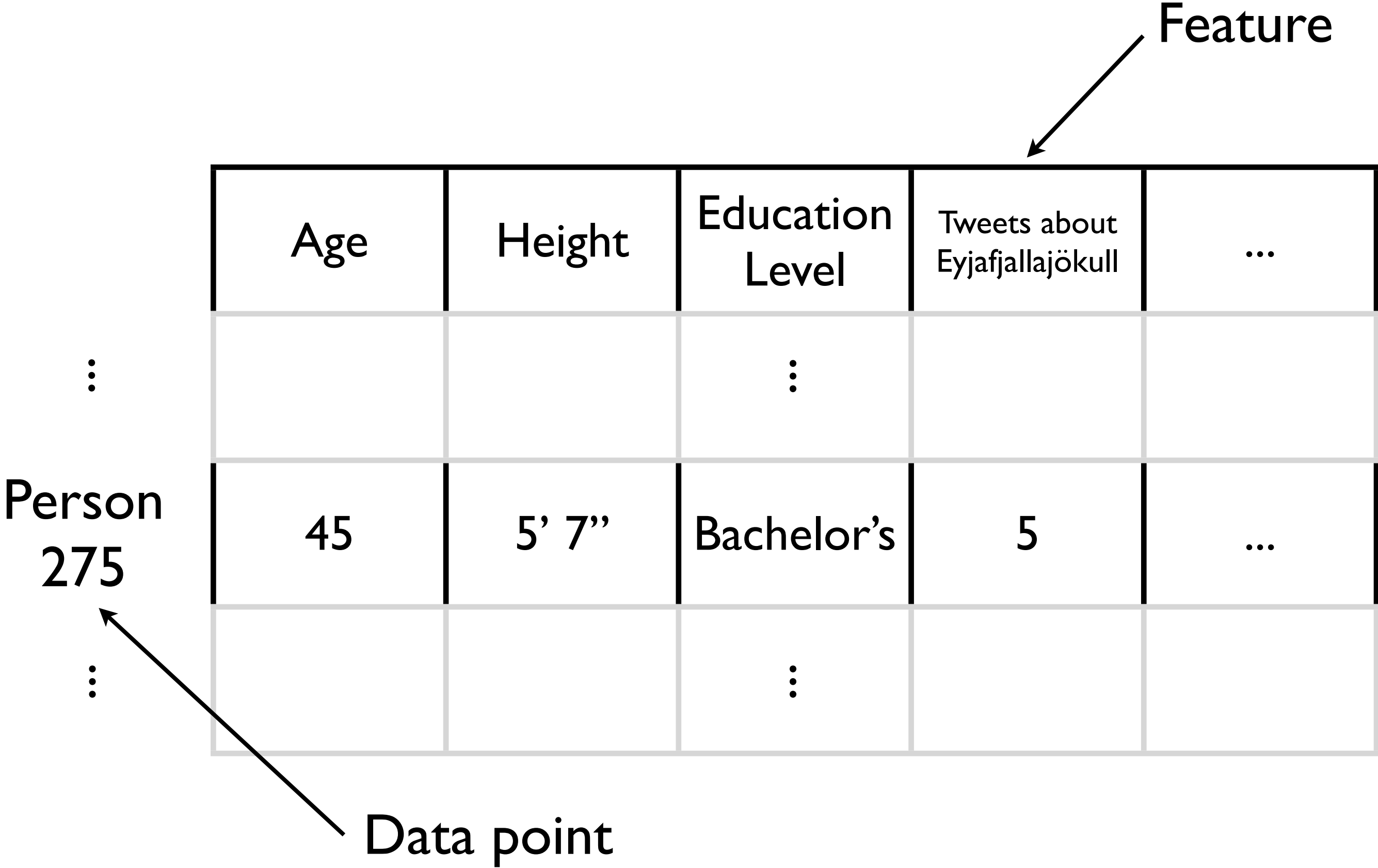
Data pre-processing

- Is the data set featurized?

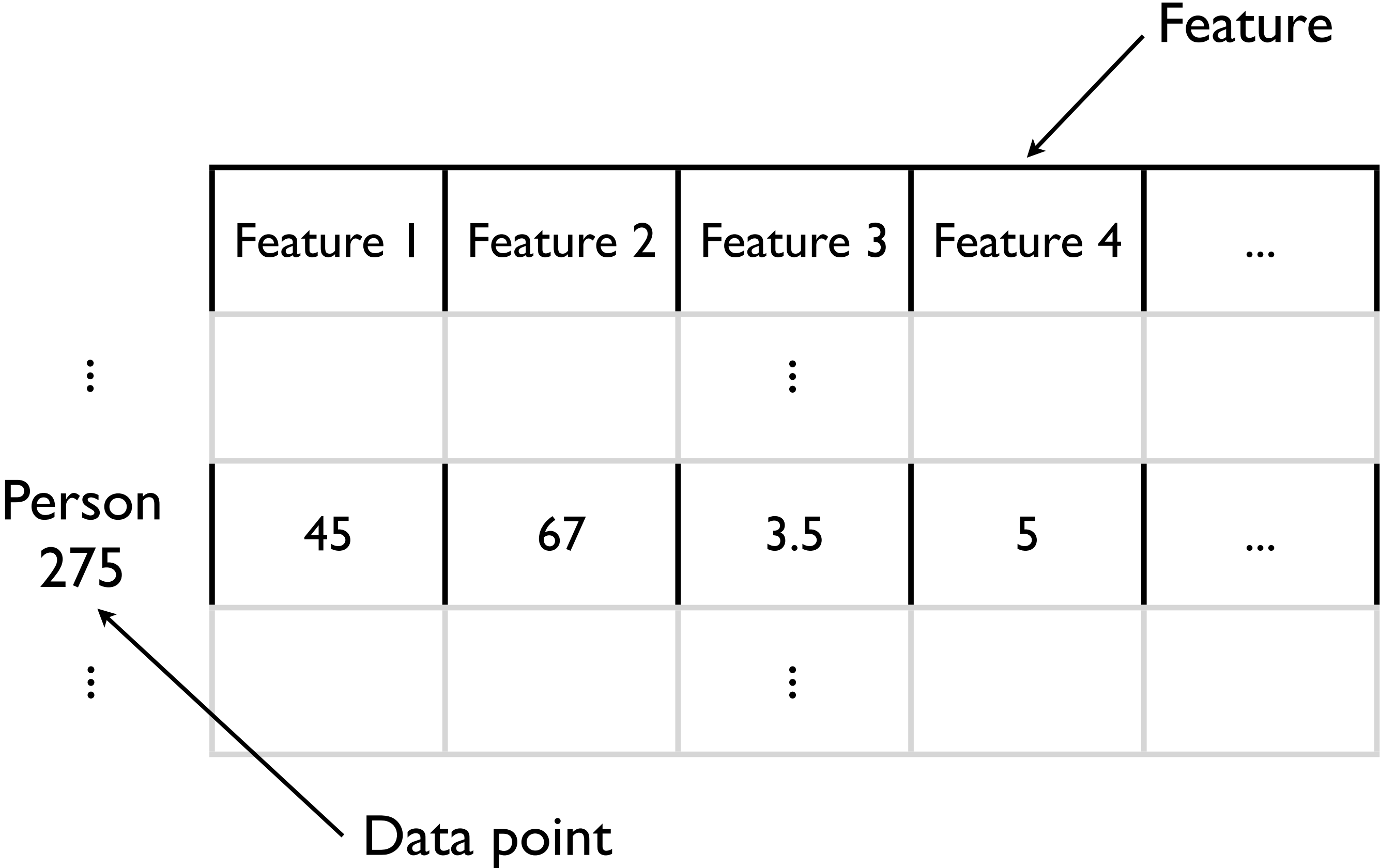
Data pre-processing

- Is the data set featurized?
- Are the features continuous numbers?

Data pre-processing



Data pre-processing



Data pre-processing

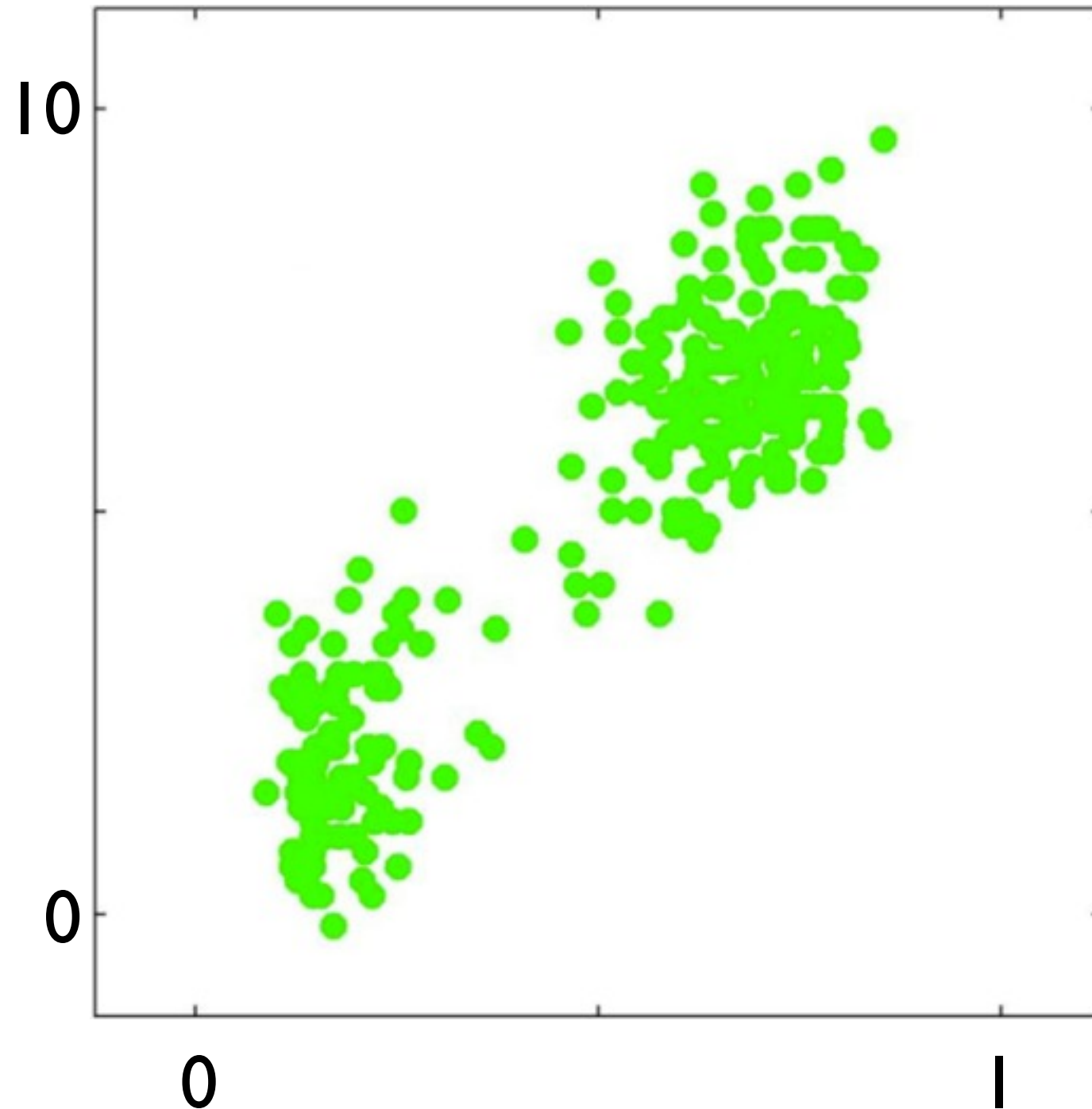
- Is the data set featurized?
- Are the features continuous numbers?

Data pre-processing

- Is the data set featurized?
- Are the features continuous numbers?
- Are these numbers commensurate?

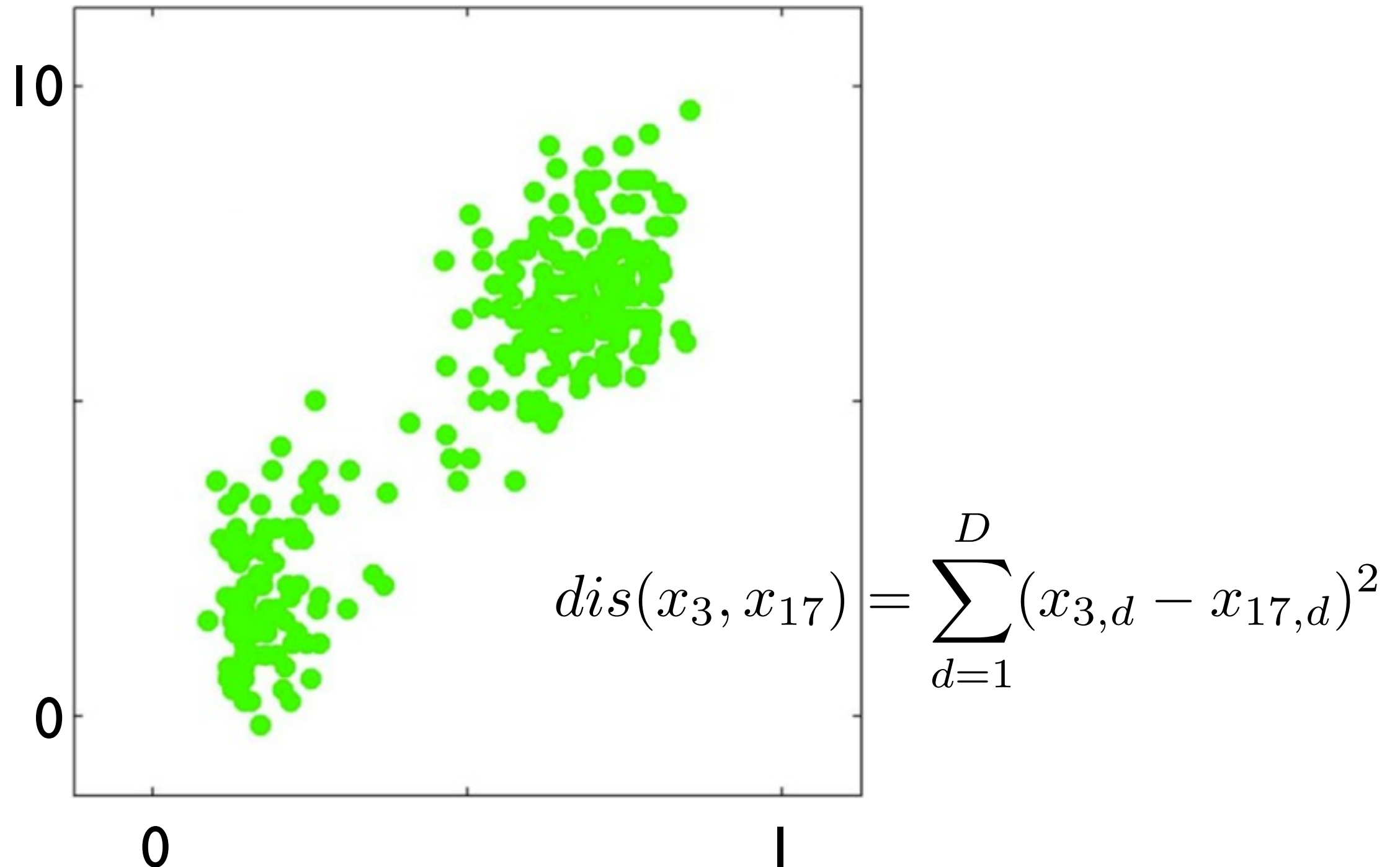
Data pre-processing

One dissimilarity value for mixed features



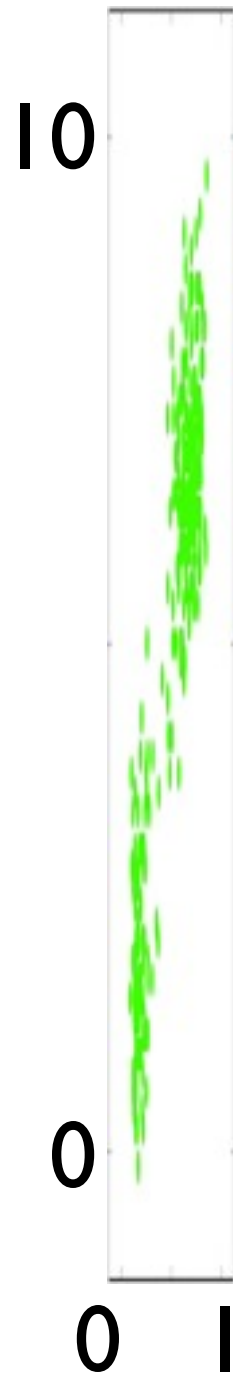
Data pre-processing

One dissimilarity value for mixed features



Data pre-processing

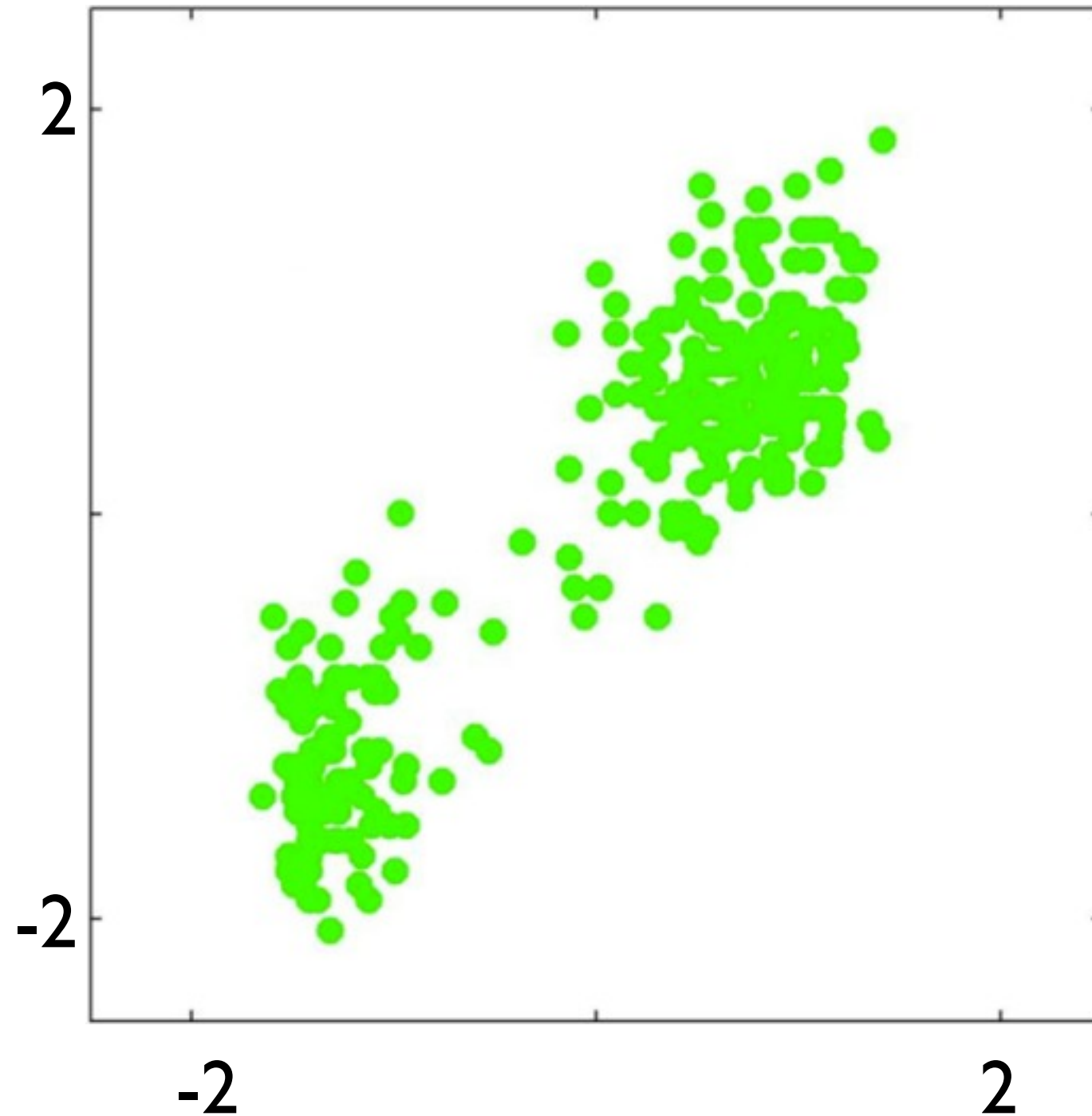
One dissimilarity value for mixed features



$$dis(x_3, x_{17}) = \sum_{d=1}^D (x_{3,d} - x_{17,d})^2$$

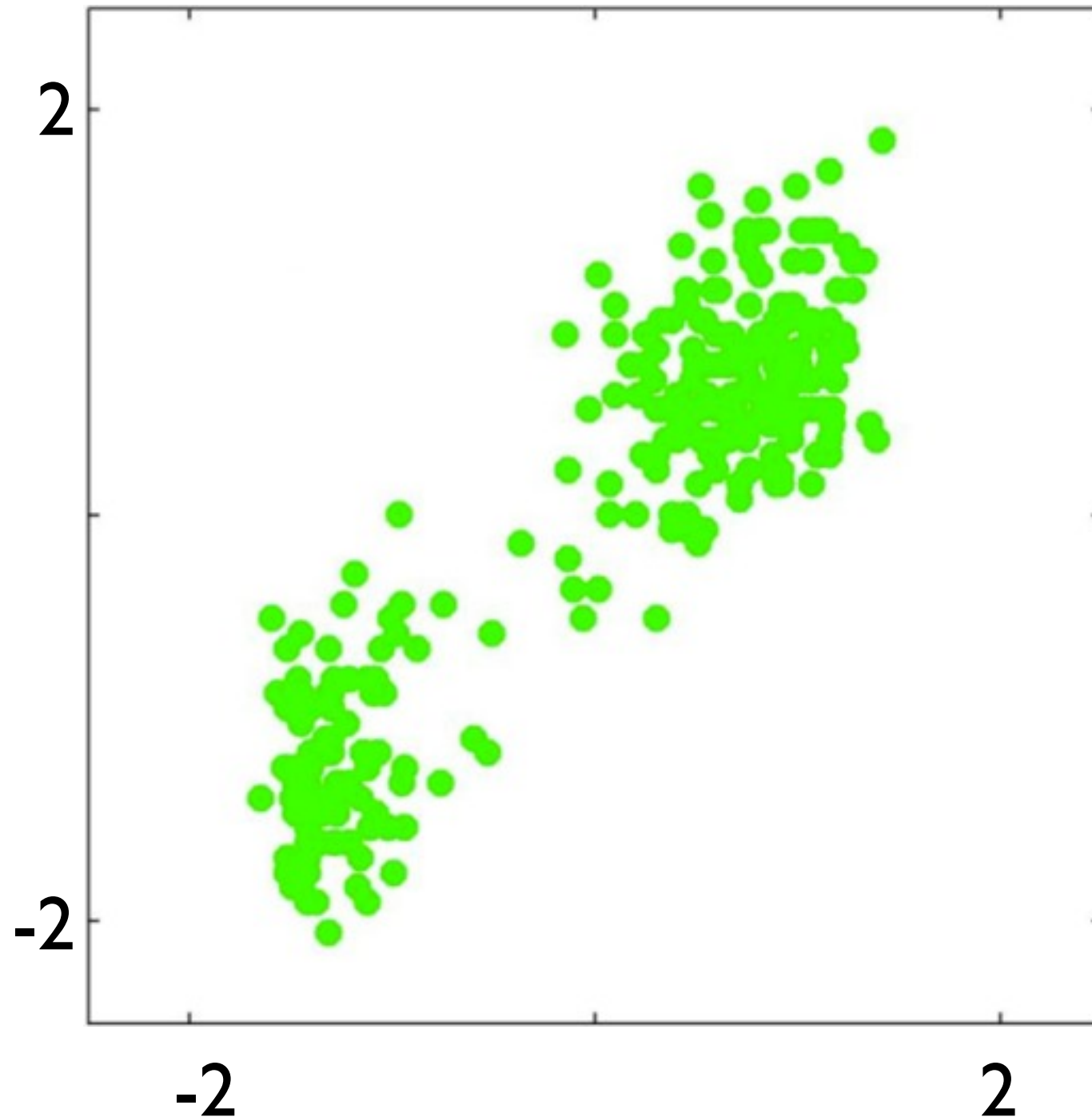
Data pre-processing

One dissimilarity value for mixed features



Data pre-processing

One dissimilarity value for mixed features



**Standardization/
Normalization**

Data pre-processing

- Is the data set featurized?
- Are the features continuous numbers?
- Are these numbers commensurate?

Data pre-processing

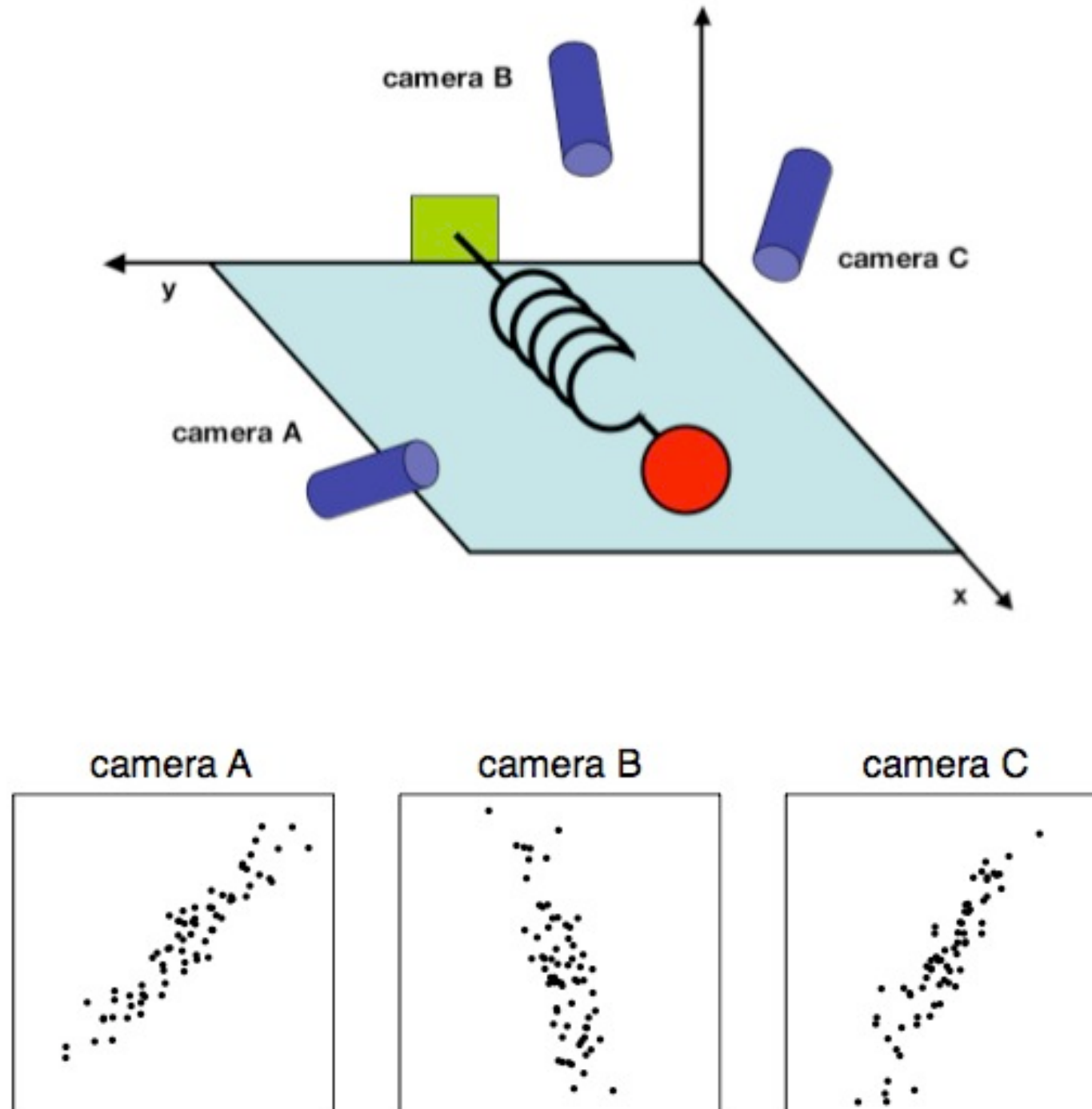
- Is the data set featurized?
- Are the features continuous numbers?
- Are these numbers commensurate?
- Are there too many features?

Data pre-processing

- Are there too many features?

Data pre-processing

- Are there too many features?
 - ◇ Principal component analysis (PCA)



Data pre-processing

- Are there too many features?
 - ◇ Principal component analysis (PCA), feature selection

Data pre-processing

- Is the data set featurized?
- Are the features continuous numbers?
- Are these numbers commensurate?
- Are there too many features?

Data pre-processing

- Is the data set featurized?
- Are the features continuous numbers?
- Are these numbers commensurate?
- Are there too many features?
- Are there any domain-specific reasons to change the features?

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
- 3. Clustering trouble-shooting**
4. Example

Outline

Clustering: Grouping data according to similarity.

1. K means algorithm
2. Clustering evaluation
3. Clustering trouble-shooting
4. Example

Example: DNA decoding

...cgtgggtgaatggatgctagggcgcacgta...

Hypothesis: DNA is made up of instruction words of length 1, 2, 3, or 4 characters.

Example: DNA decoding

...cgtgggtgaatggatgctagggcgcacgta...

Hypothesis: DNA is made up of instruction words of length 1, 2, 3, or 4 characters.

Question: Is this true? Which length is correct?

Example: DNA decoding

...cgtggtgaatggatgctagggcgcacgta...

Hypothesis: DNA is made up of instruction words of length 1, 2, 3, or 4 characters.

Question: Is this true? Which length is correct?

From “PCA and K-means decipher genome”

Example: DNA decoding

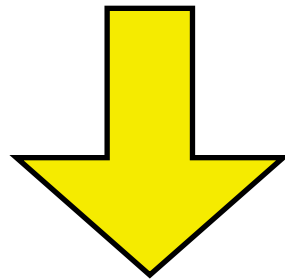
...cgtgggtgaatggatgctagggcgcacgta...

Data: ~300KB DNA substring of *Caulobacter*
Crescentus bacterium

Example: DNA decoding

...cgtgggtgaatggatgctagggcgcacgta...

Data: ~300KB DNA substring of *Caulobacter*
Crescentus bacterium

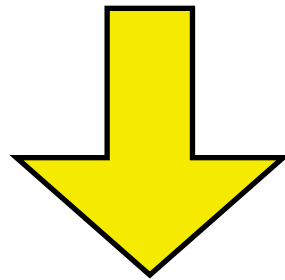


Non-overlapping DNA strings of length 300

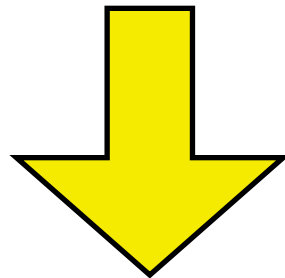
Example: DNA decoding

...cgtggtgaatggatgctagggcgcacgta...

Data: ~300KB DNA substring of *Caulobacter Crescentus* bacterium



Non-overlapping DNA strings of length 300



For each substring, a count of each possible word of length m ($m = 1, 2, 3, \text{ or } 4$)

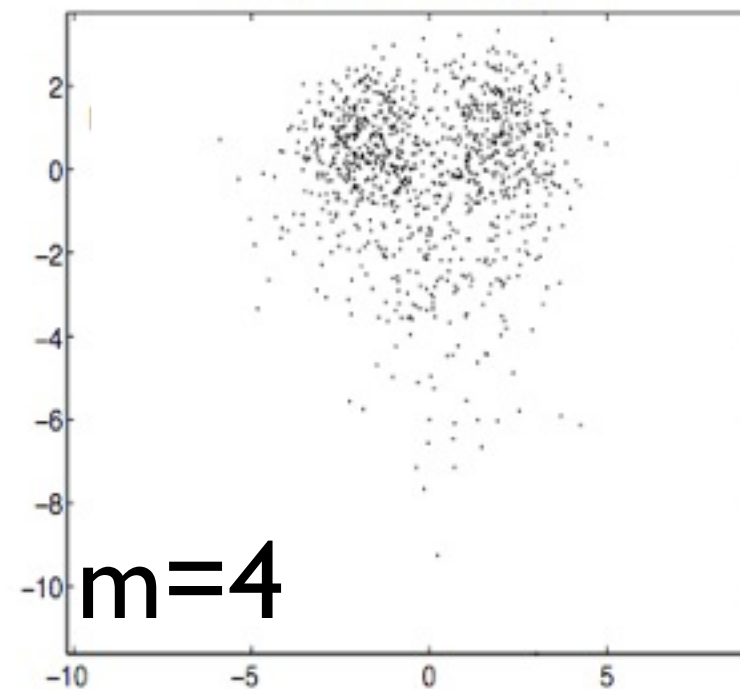
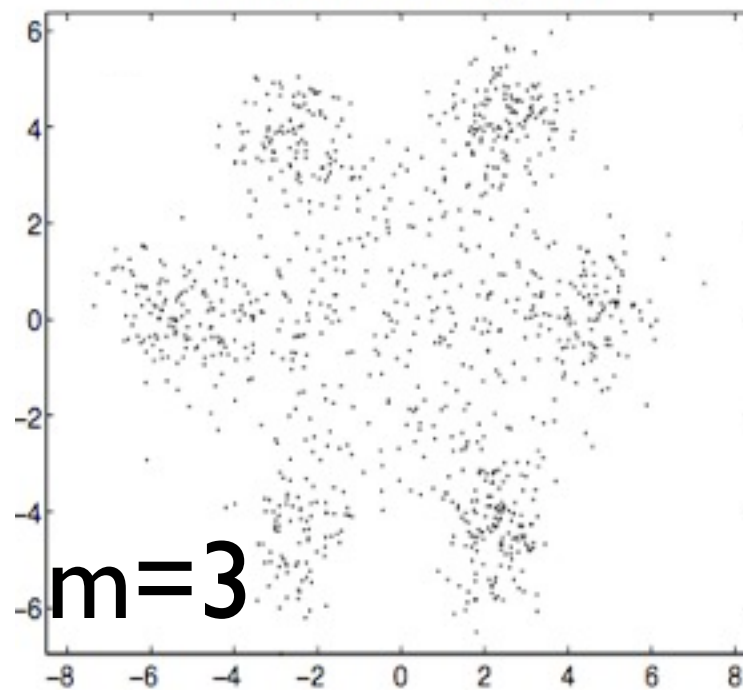
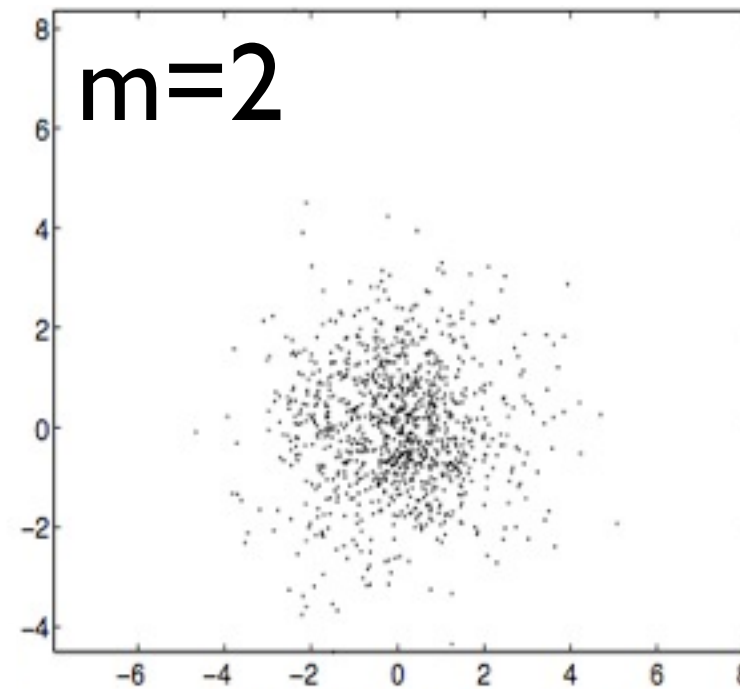
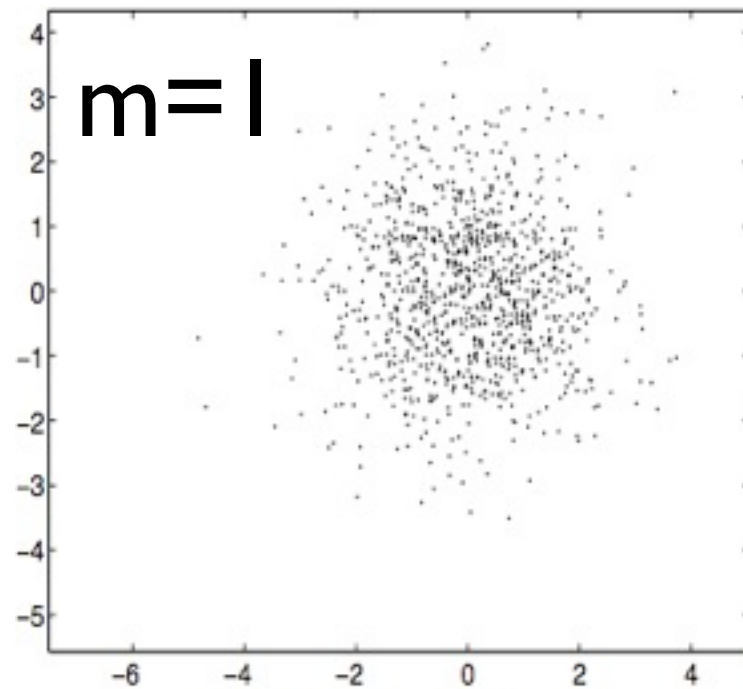
Example: DNA decoding

Featurized data for $m = 2$

	aa	ac	...	tt
String 1	6	10		4
⋮			⋮	
String 1017	8	20		9

Example: DNA decoding

Examine the first two principal components (from PCA)



Example: DNA decoding

Count data for $m = 3$

	aaa	aac	...	ttt
String 1	$N_{1,aaa}$	$N_{1,aac}$		$N_{1,ttt}$
$\vdots$			$\ddots$	
String 1017	$N_{1017,aaa}$	$N_{1017,aac}$		$N_{1017,ttt}$

Example: DNA decoding

Normalized data for $m = 3$

	aaa	aac	...	ttt
String 1	$x_{1,aaa}$	$x_{1,aac}$		$x_{1,ttt}$
⋮			⋮	
String 1017	$x_{1017,aaa}$	$x_{1017,aac}$		$x_{1017,ttt}$

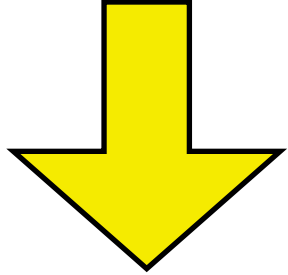
$$x_{1,aaa} = (N_{1,aaa} - \text{mean}_{aaa}) / \text{std}_{aaa}$$

Example: DNA decoding

Normalized data for $m = 3$

Example: DNA decoding

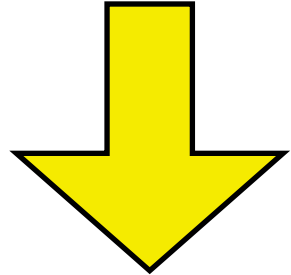
Normalized data for $m = 3$



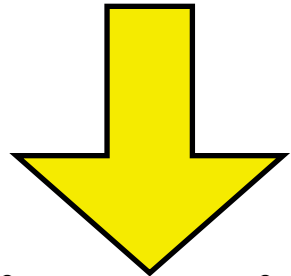
K means ($K = 6$)

Example: DNA decoding

Normalized data for $m = 3$



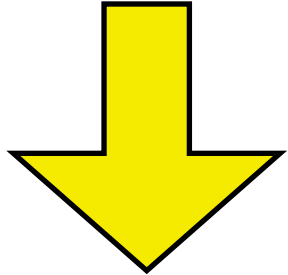
K means ($K = 6$)



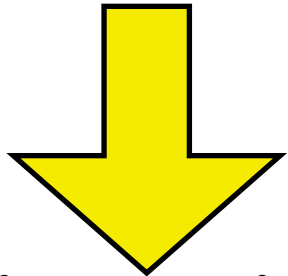
**Visualize
with PCA**

Example: DNA decoding

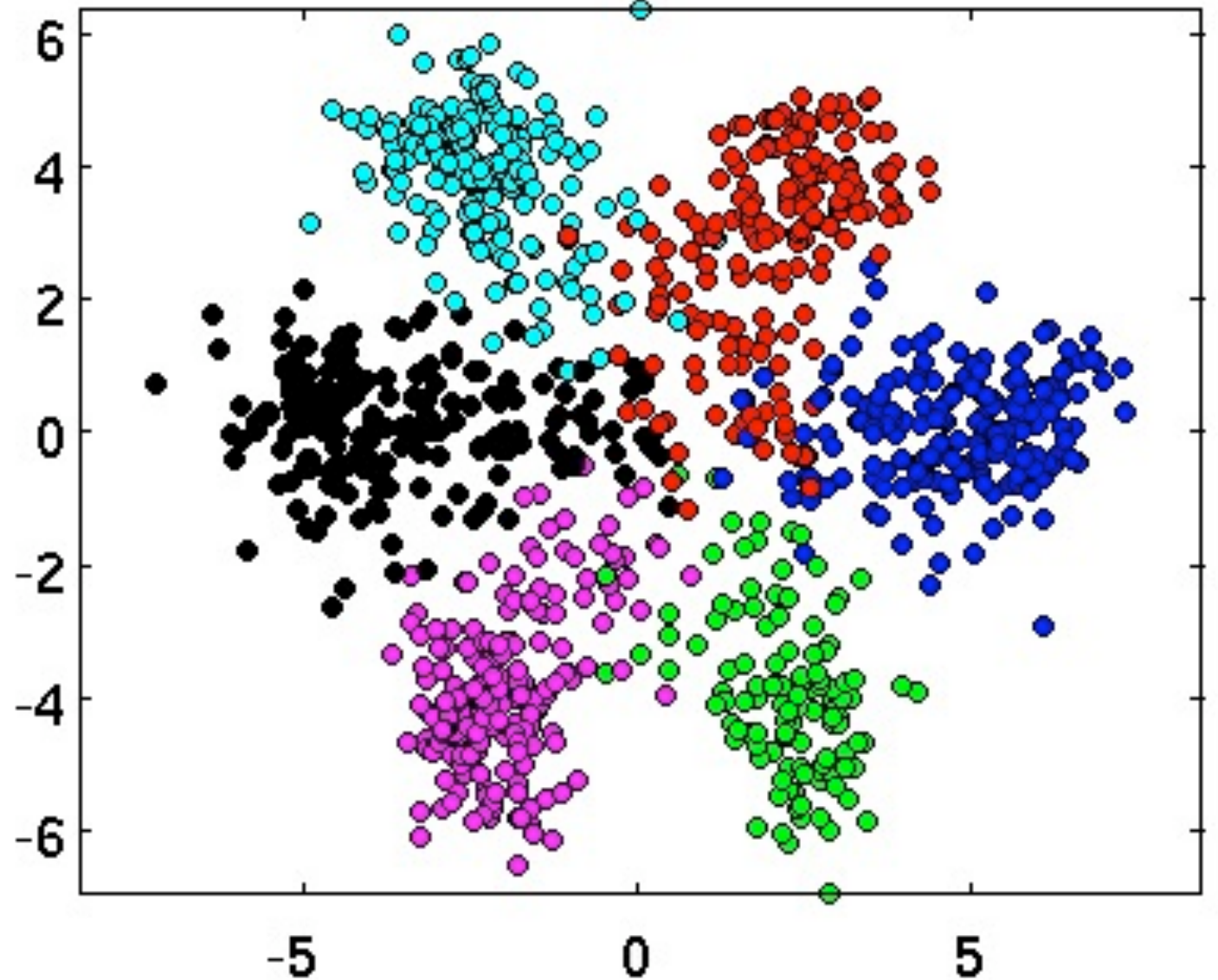
Normalized data for $m = 3$



K means ($K = 6$)

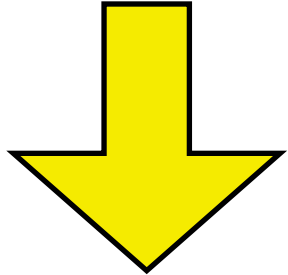


**Visualize
with PCA**

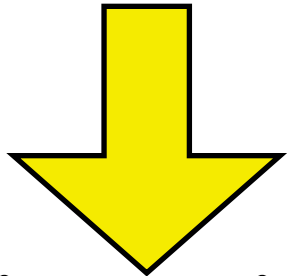


Example: DNA decoding

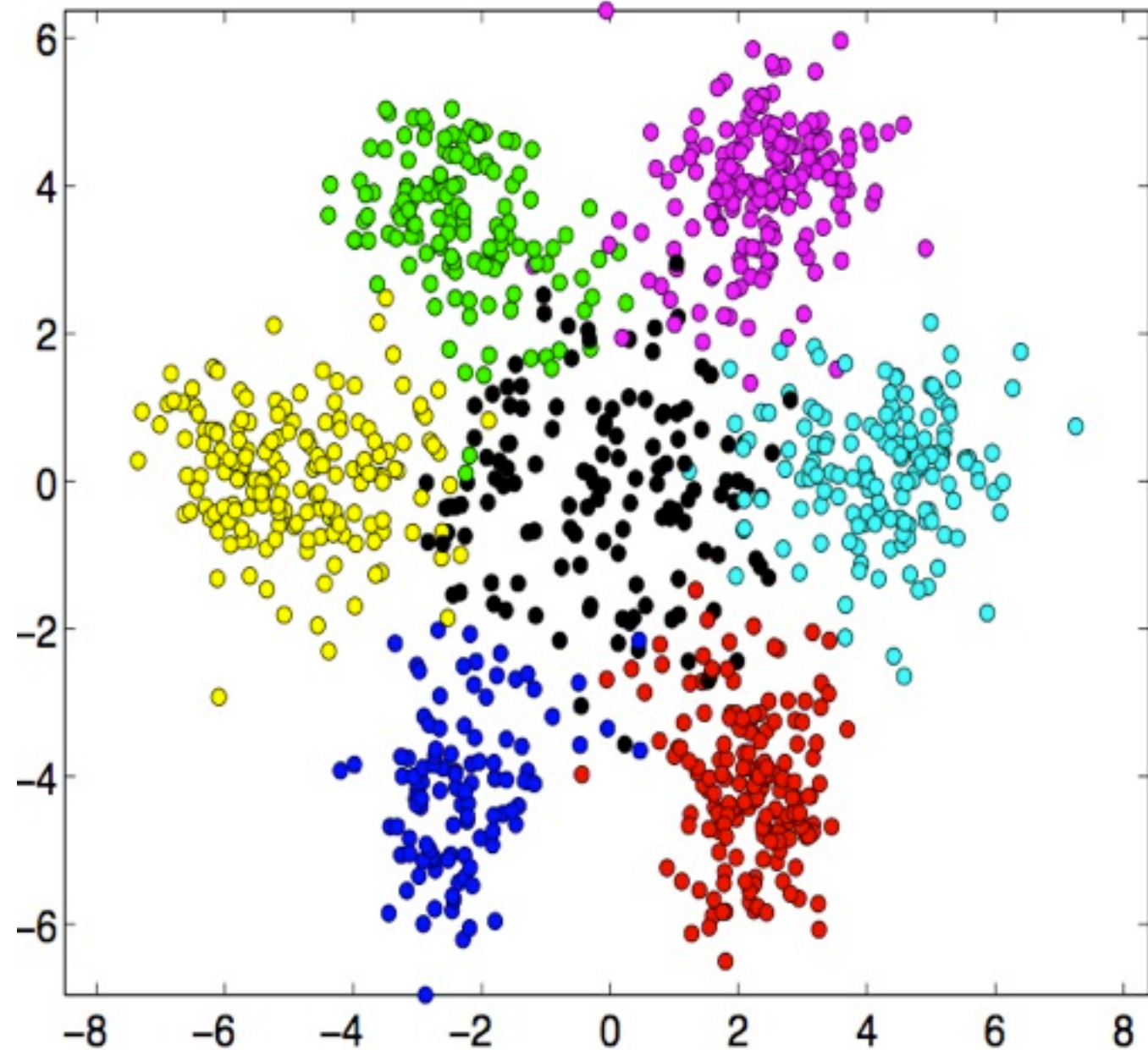
Normalized data for $m = 3$



K means ($K = 7$)

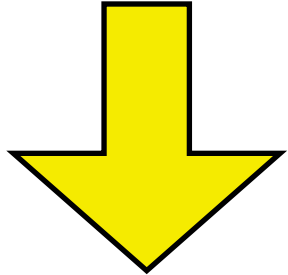


**Visualize
with PCA**

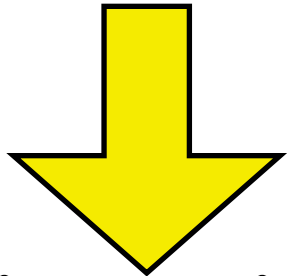


Example: DNA decoding

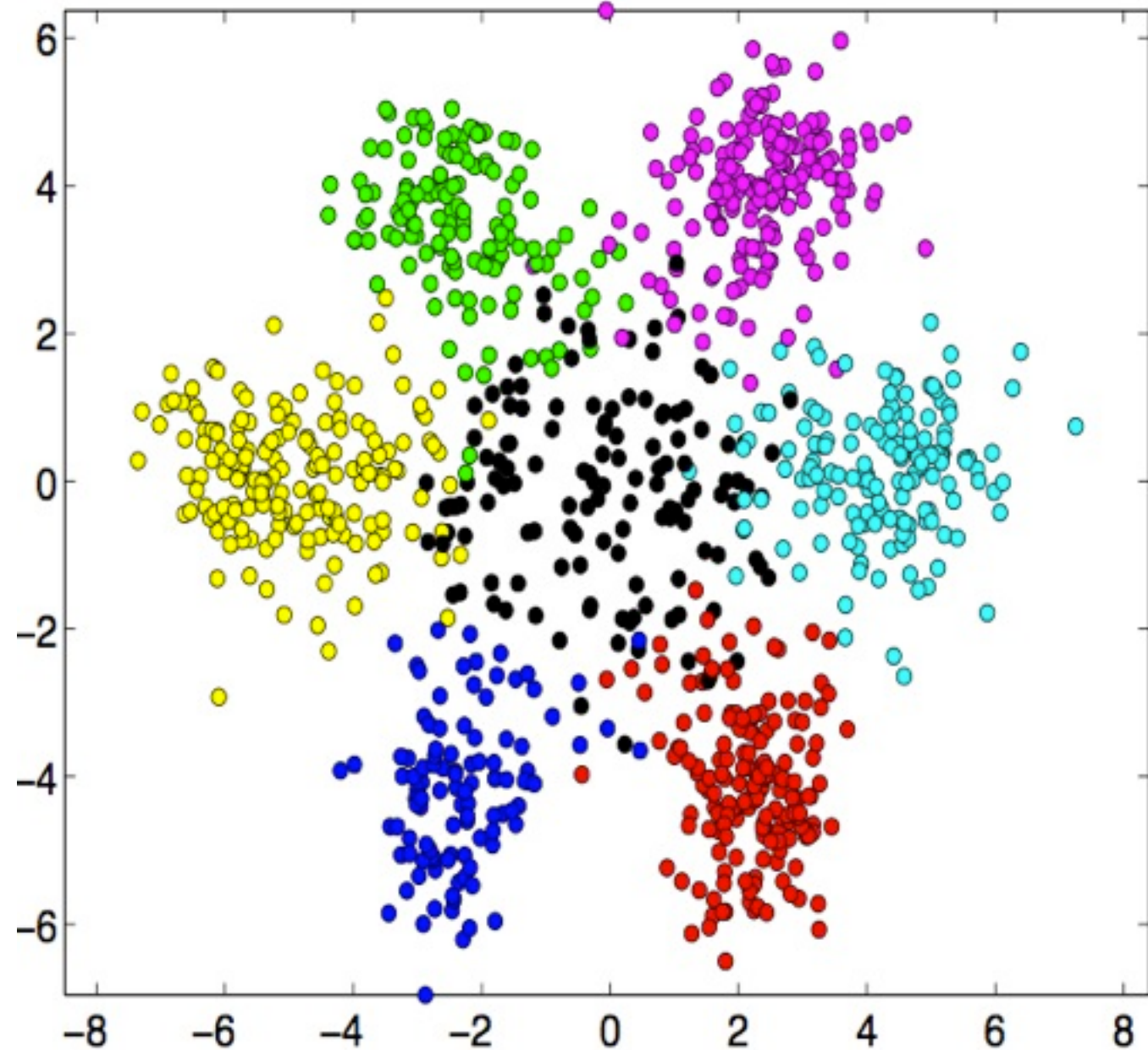
Normalized data for $m = 3$



K means ($K = 7$)

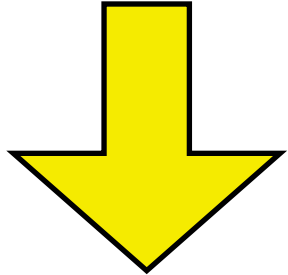


**Visualize
with PCA**

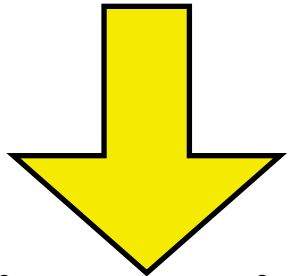


Example: DNA decoding

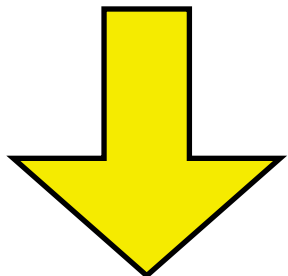
Normalized data for $m = 3$



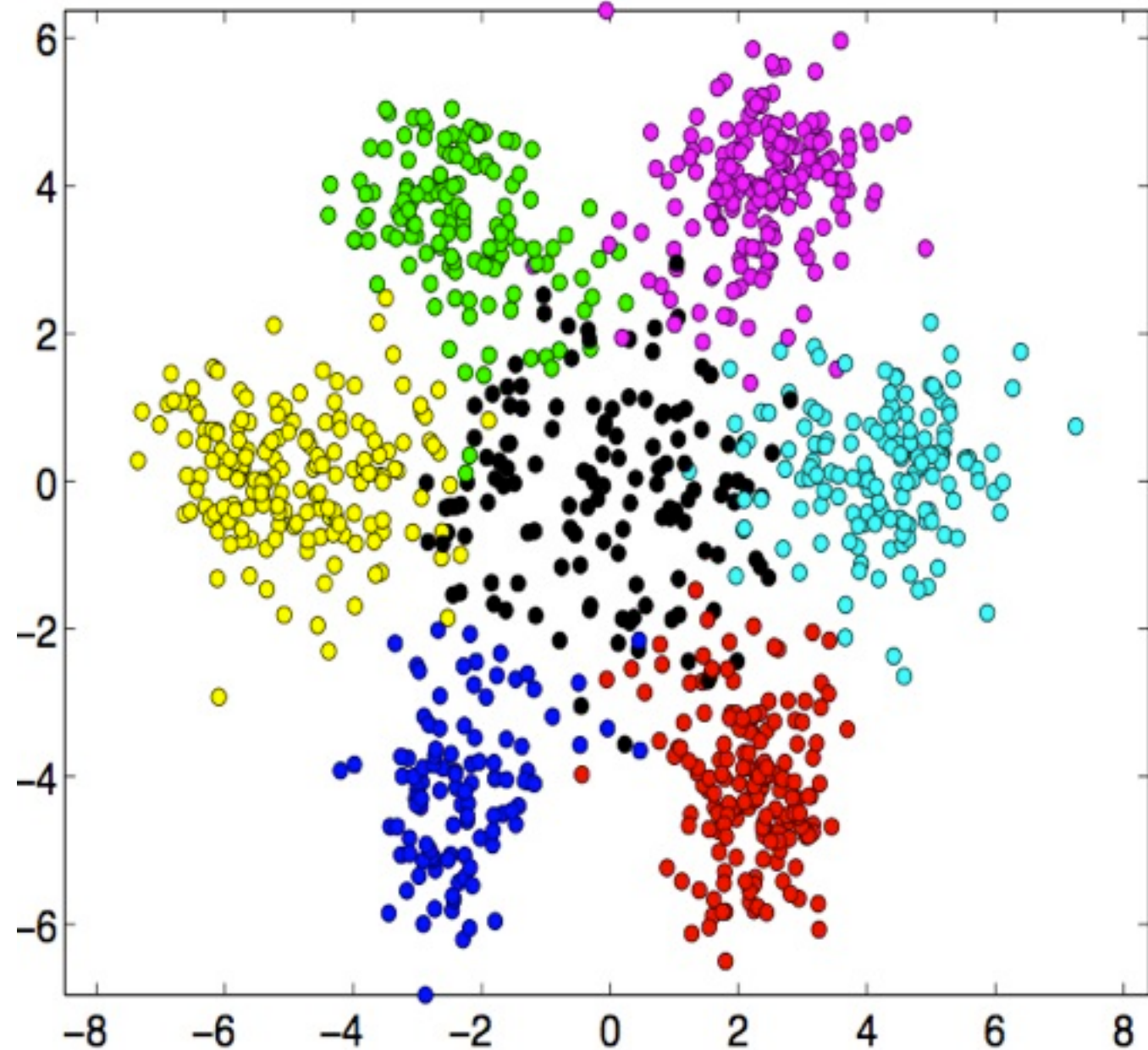
K means ($K = 7$)



**Visualize
with PCA**



**Analysis of
Results**



Goals

- Big ideas (clustering)
- Concrete implementation (K means)
- Machine learning is not a black box
- Machine learning pipeline

Image references

Bishop, C. M. Pattern Recognition and Machine Learning (2006).

Blei, D. M., Ng, A. Y., and Jordan, M. I. "Latent Dirichlet allocation." Journal of Machine Learning Research (2003).

Carpineto, C., Osinski, S., Romano, G., and Weiss, D. "A survey of web clustering engines." ACM Computing Surveys (2009).

Fei-Fei, L. "Lecture 5: Clustering and segmentation -- Part I." CS 231A: Introduction to Computer Vision (2011).

Garcia-Escudero, L. A. and Gordaliza, A. "Robustness properties of k means and trimmed k means." Journal of the American Statistical Association (1999).

Gorban, A. N. and Zinovyev, A. Y. "PCA and K-means decipher genome." Principal Manifolds for Data Visualization and Dimension Reduction (2007).

Hastie, T., Tibshirani, R., and Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2001).

Jepson, A.D., and Fleet, D.J. "Image segmentation." CSC 2503 Foundations of Computer Vision (2011).

Krivitsky, P. N. and Handcock, M. S. "Fitting position latent cluster models for social networks with latentnet." Journal of Statistical Software (2008).

Shlens, J. "A tutorial on principal component analysis: derivation, discussion, and singular value decomposition" (2003).
http://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf

Swayne, D. F., Cook, D., Buja, A., Lang, D. T., Wickham, H., and Lawrence, M. GGobi Manual (2006).

Tibshirani, R., Walther, G., and Hastie, T. "Estimating the number of clusters in a data set via the gap statistic." Journal of the Royal Statistical Society B (2001).