Photo by Unsplash user @foodiesfeed

BBM406

Fundamentals of Machine Learning

Lecture 21:

Clustering
K-Means

HACETTEPE UNIVERSITY COMPUTER VISION LAB

Aykut Erdem // Hacettepe University // Fall 2019

Last time... **Boosting**

- **Idea:** given a weak learner, run it multiple times on (reweighted) training data, then let the learned classifiers vote
- On each iteration t:
  - weight each training example by how incorrectly it was classified
  - Learn a hypothesis – h_t
  - A strength for this hypothesis – a_t
- Final classifier:
  - A linear combination of the votes of the different classifiers weighted by their strength $H(X) = \text{sign}\left(\sum \alpha_t h_t(X)\right)$
- **Practically useful**
- **Theoretically interesting**

slide by Aarti Singh & Barnabas Poczos

Last time.. **The AdaBoost Algorithm**

0) Set $\tilde{W}_i^{(0)} = 1/n$ for $i = 1, \ldots, n$

1) At the m^{th} iteration we find (any) classifier $h(\mathbf{x}; \hat{\theta}_m)$ for which the *weighted classification error* ϵ_m

$$\epsilon_m = 0.5 - \frac{1}{2}\left(\sum_{i=1}^{n} \tilde{W}_i^{(m-1)} y_i h(\mathbf{x}_i; \hat{\theta}_m)\right)$$

is better than chance.

2) The new component is assigned votes based on its error:

$$\hat{\alpha}_m = 0.5 \log(\,(1 - \epsilon_m)/\epsilon_m\,)$$

3) The weights are updated according to (Z_m is chosen so that the new weights $\tilde{W}_i^{(m)}$ sum to one):

$$\tilde{W}_i^{(m)} = \frac{1}{Z_m} \cdot \tilde{W}_i^{(m-1)} \cdot \exp\{\,-y_i \hat{\alpha}_m h(\mathbf{x}_i; \hat{\theta}_m)\,\}$$

slide by Jiri Matas and Jan Šochman

Today

- What is clustering?
- K-means algorithm

What is clustering

Clustering

- Grouping data according to similarity

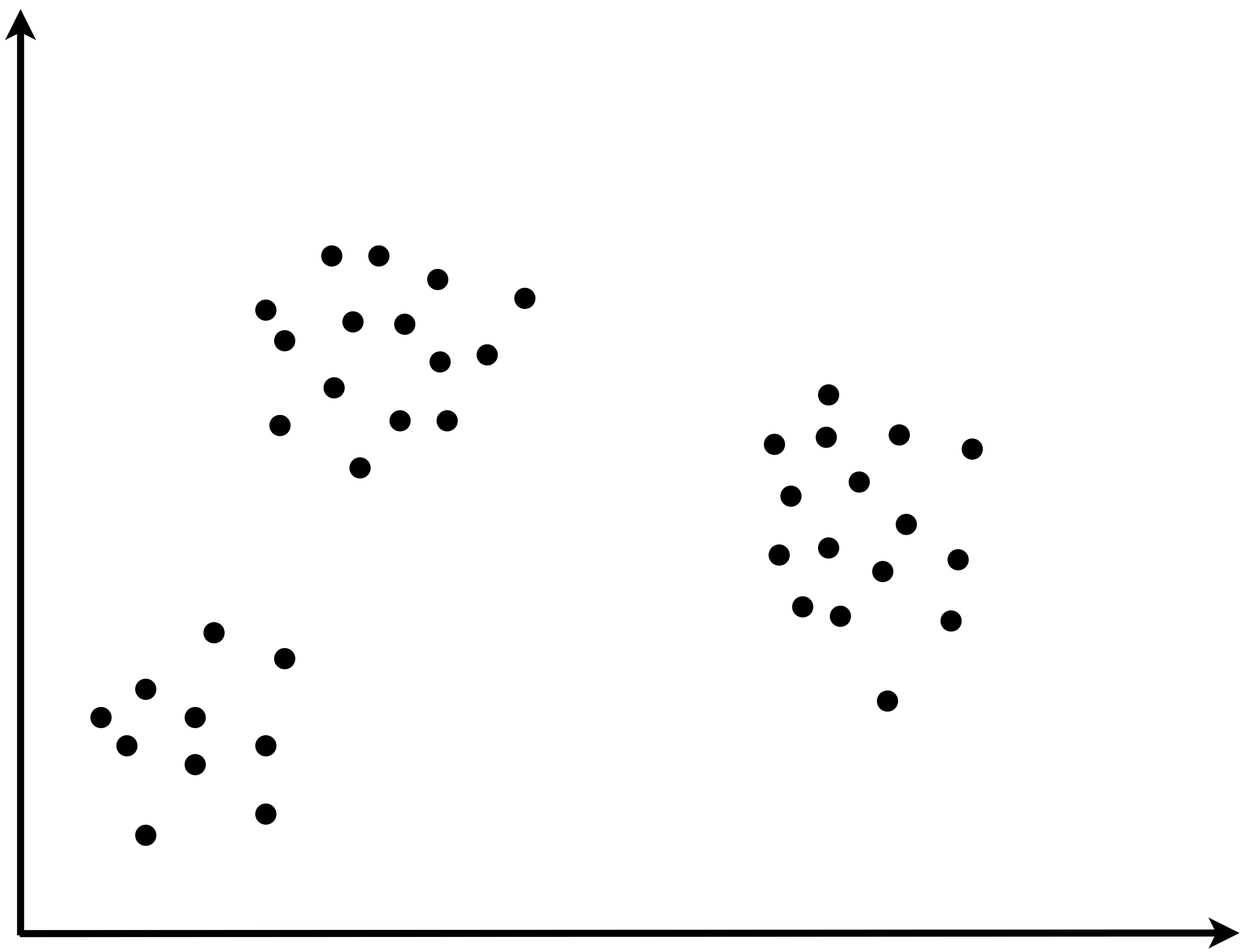
Clustering

- Grouping data according to similarity

slide by Tamara Broderick

Clustering

- Grouping <u>data</u> according to similarity

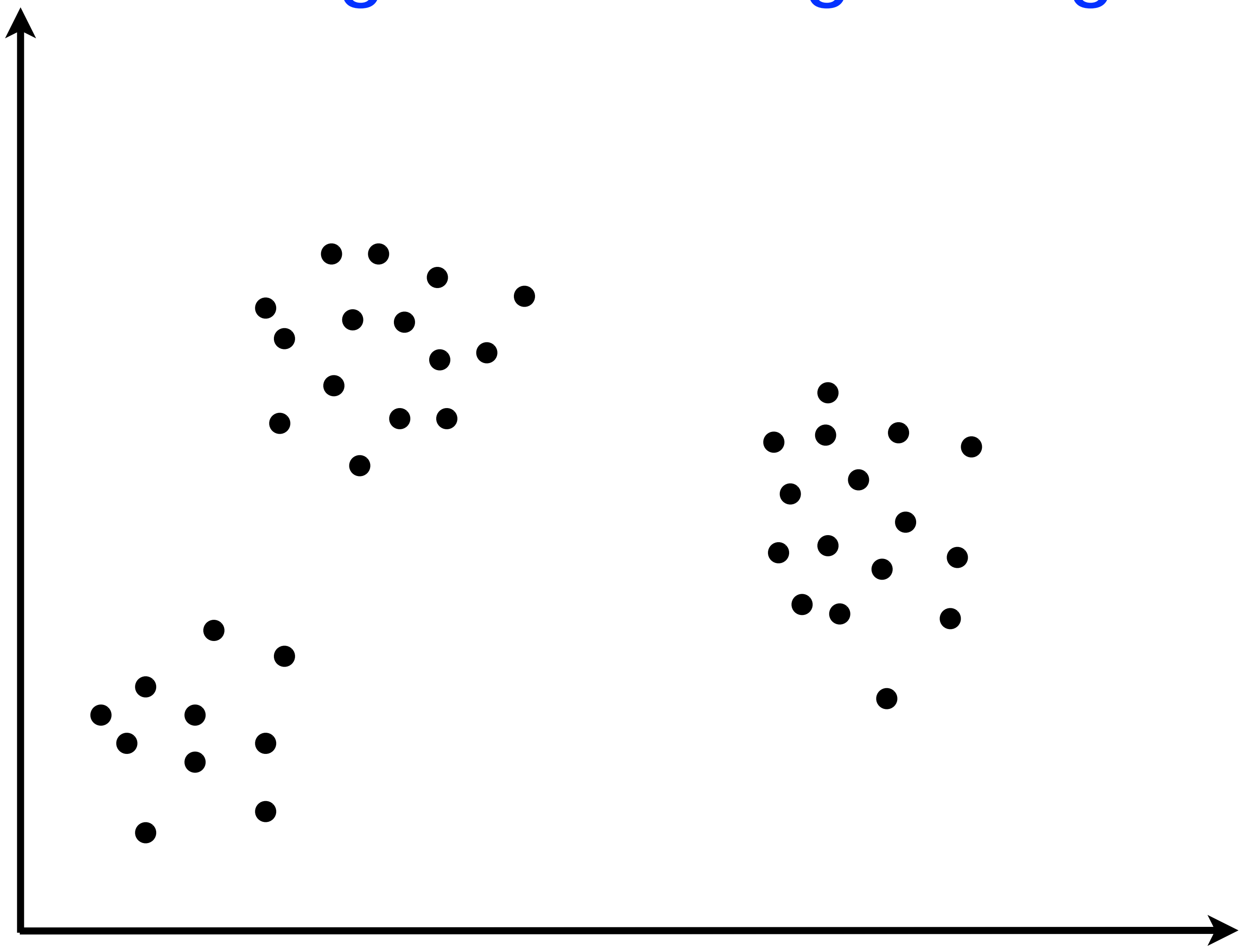


slide by Tamara Broderick

Clustering

- Grouping <u>data</u> according to similarity

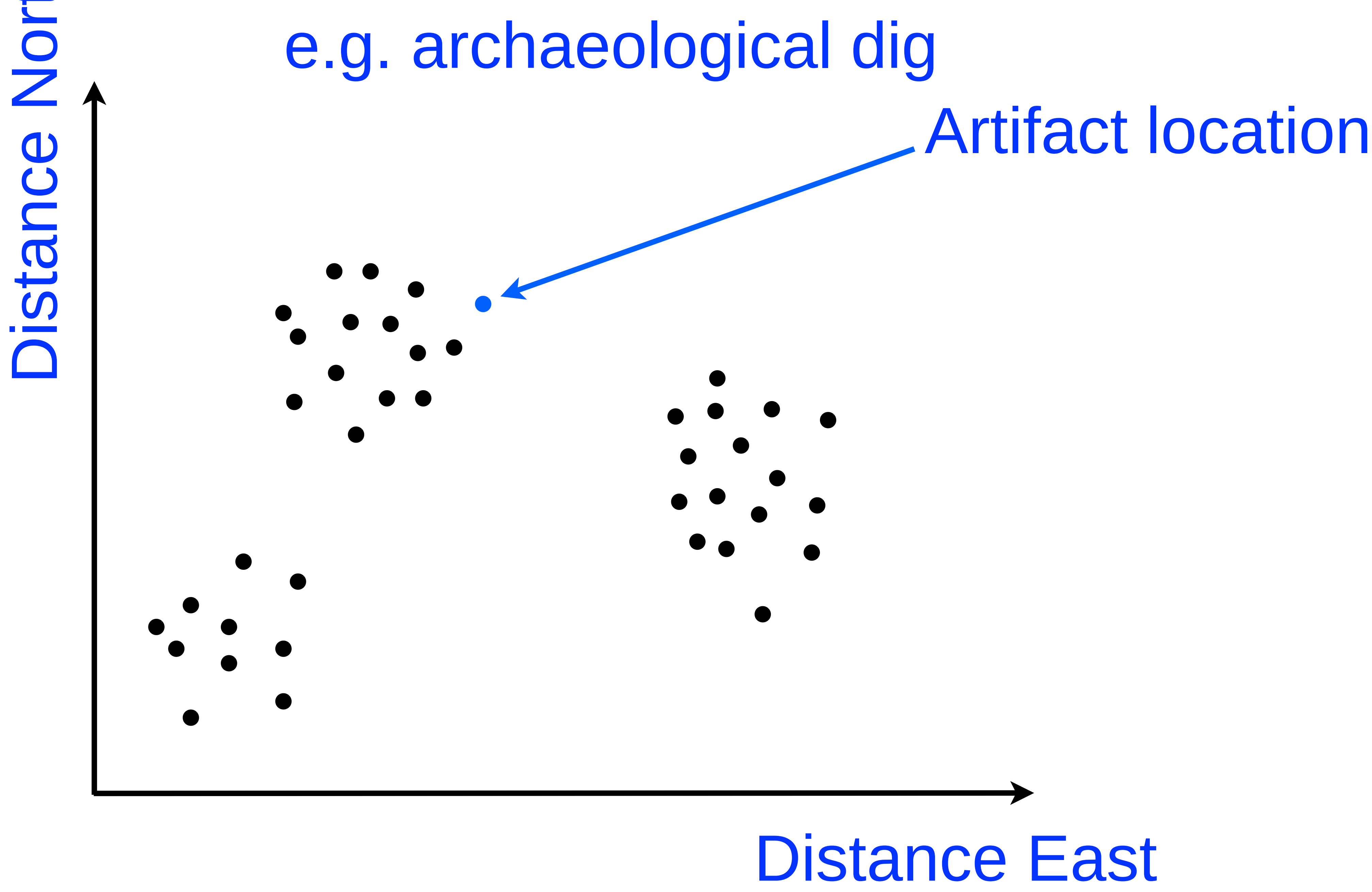
e.g. archaeological dig



slide by Tamara Broderick

Clustering

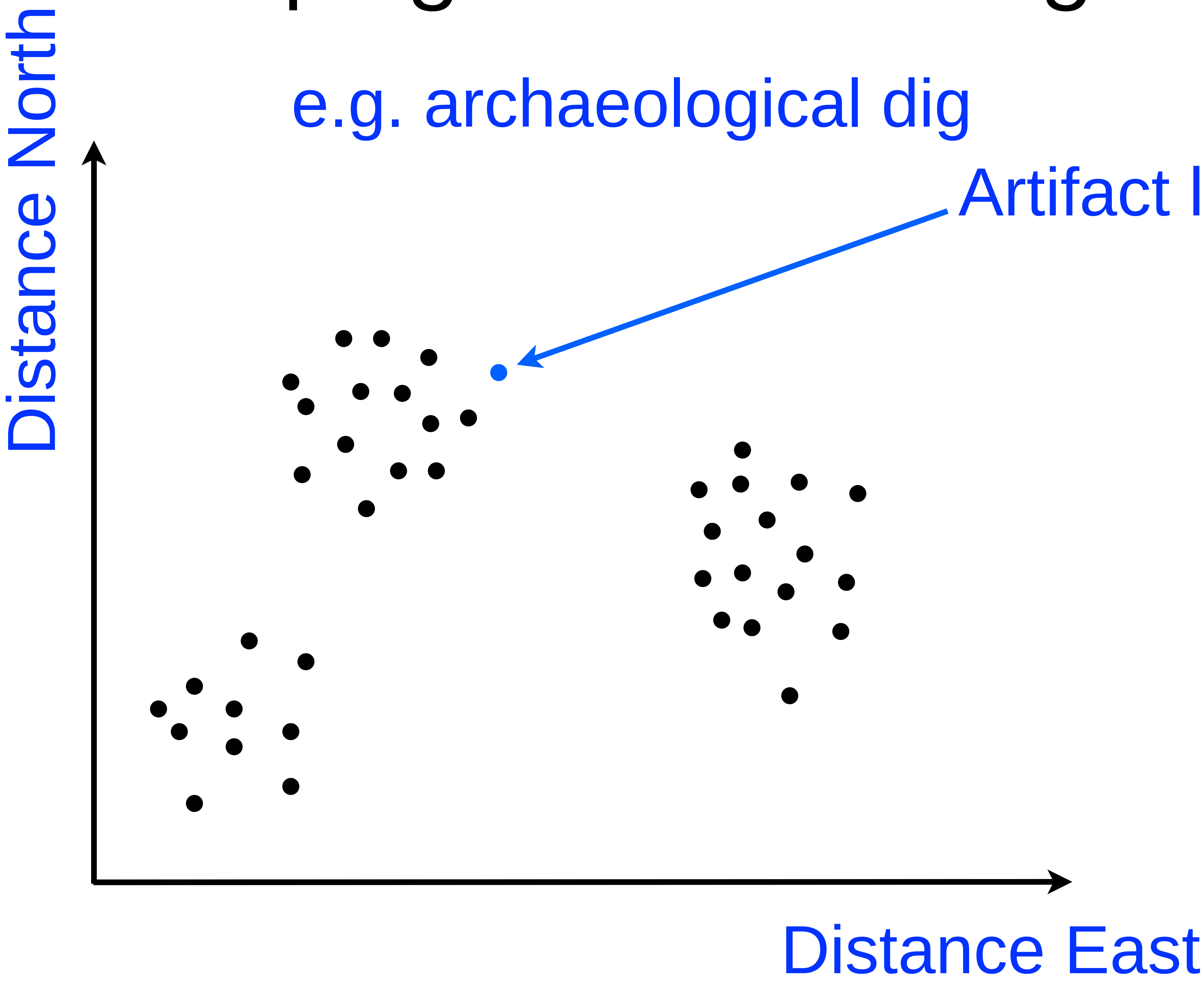
- Grouping <u>data</u> according to similarity



slide by Tamara Broderick

Clustering

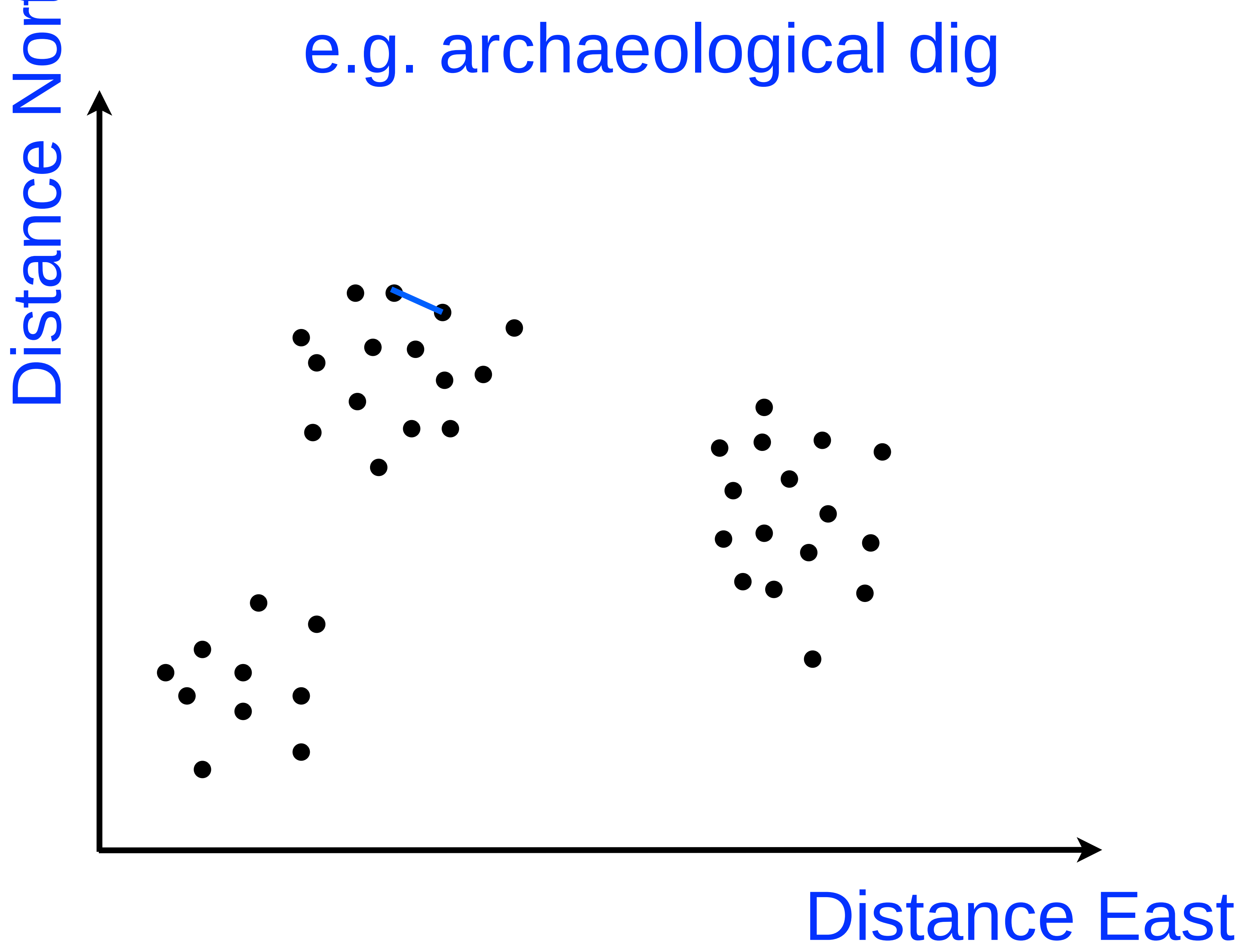
- Grouping data according to <u>similarity</u>



slide by Tamara Broderick

Clustering

- Grouping data according to <u>similarity</u>

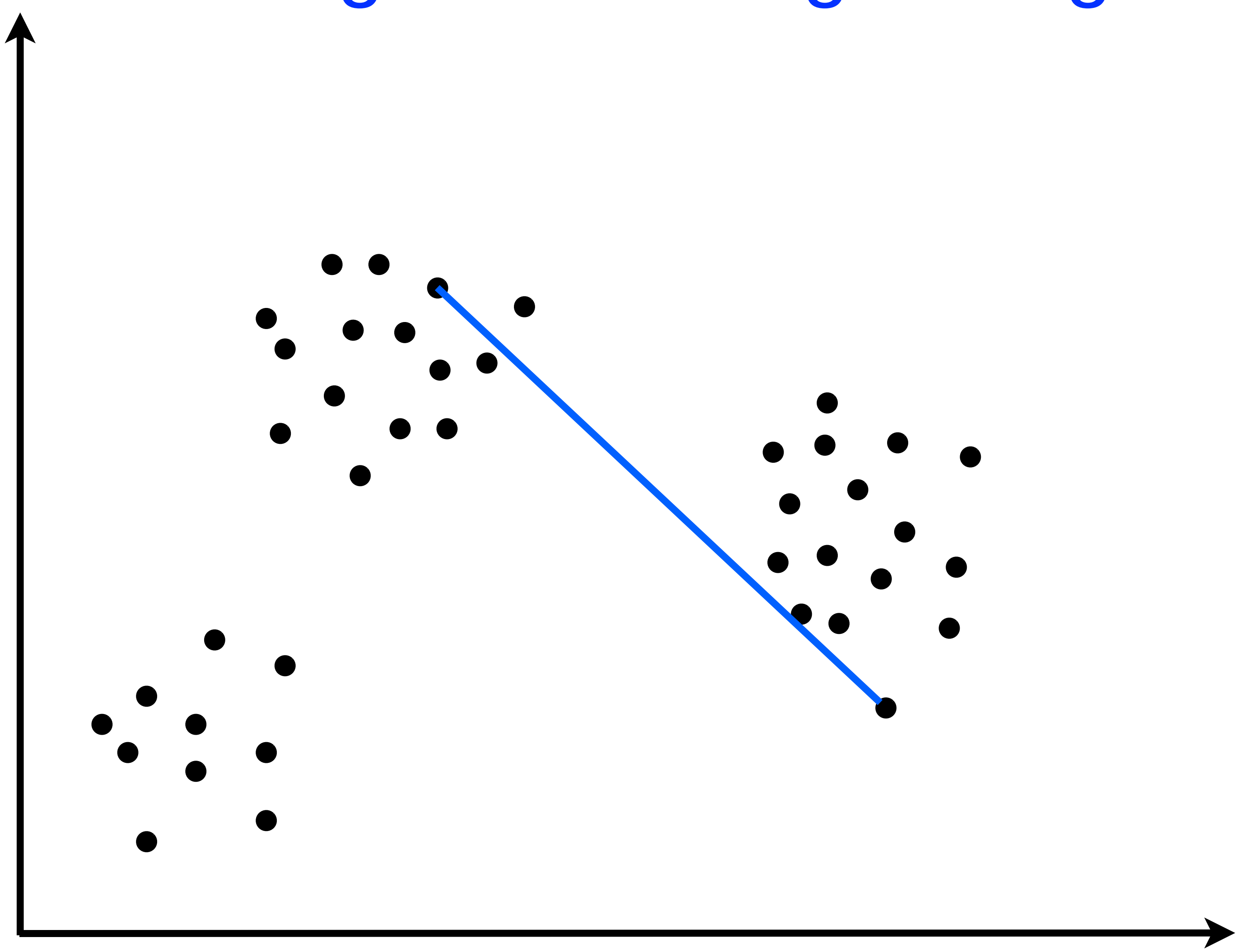


slide by Tamara Broderick

Clustering

- Grouping data according to <u>similarity</u>

e.g. archaeological dig

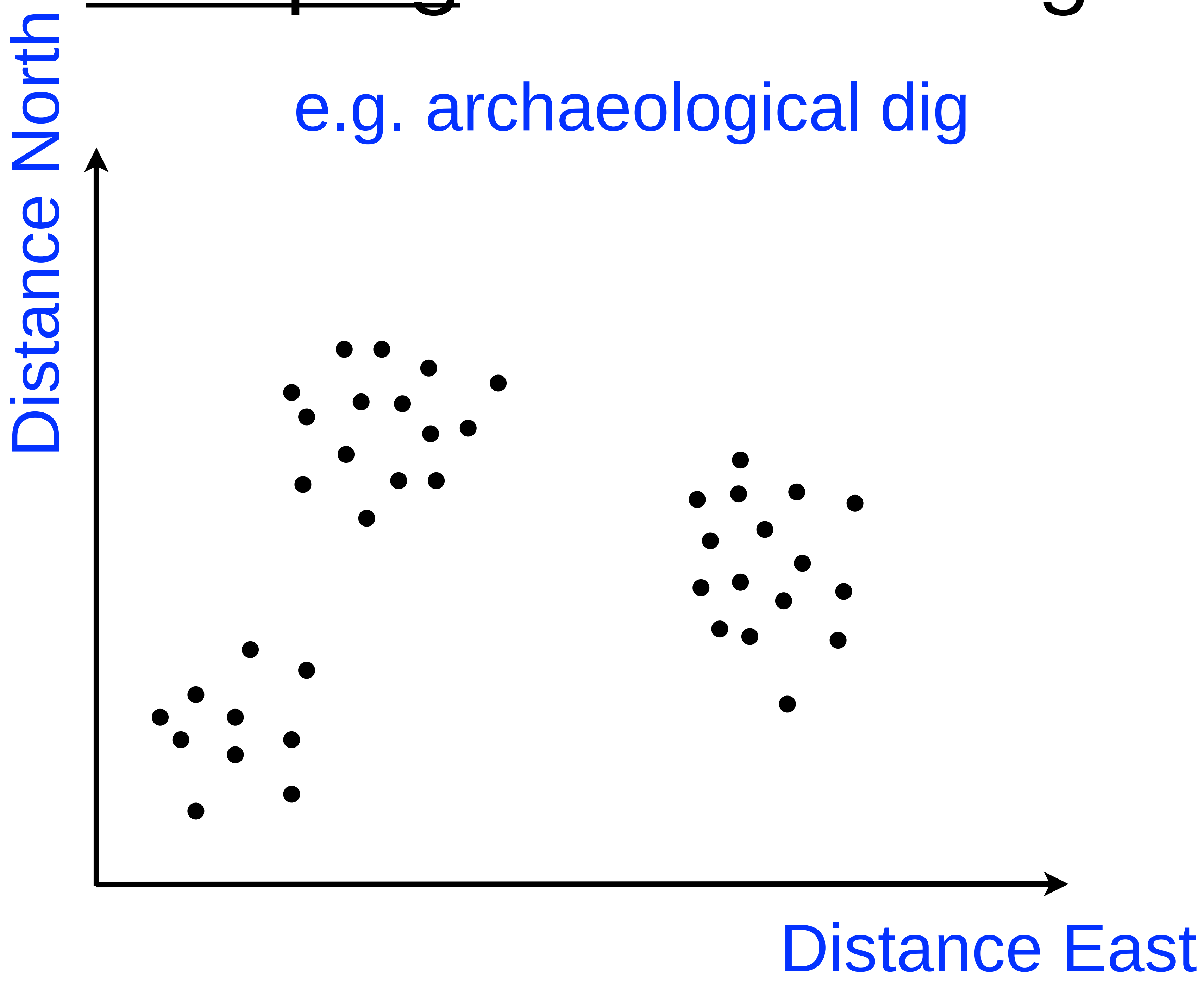


slide by Tamara Broderick

Clustering

- Grouping data according to similarity

e.g. archaeological dig

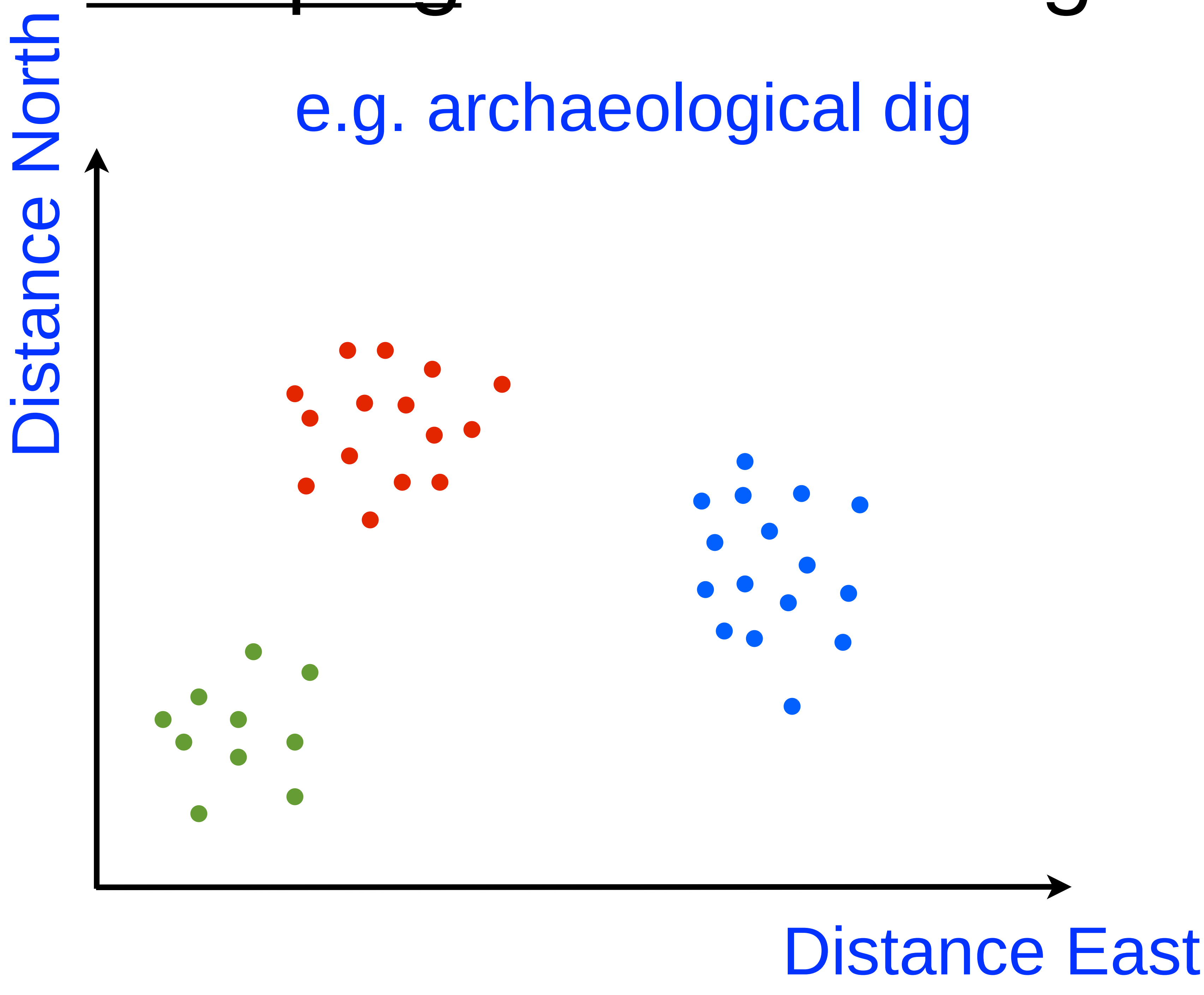


slide by Tamara Broderick

Clustering

- Grouping data according to similarity

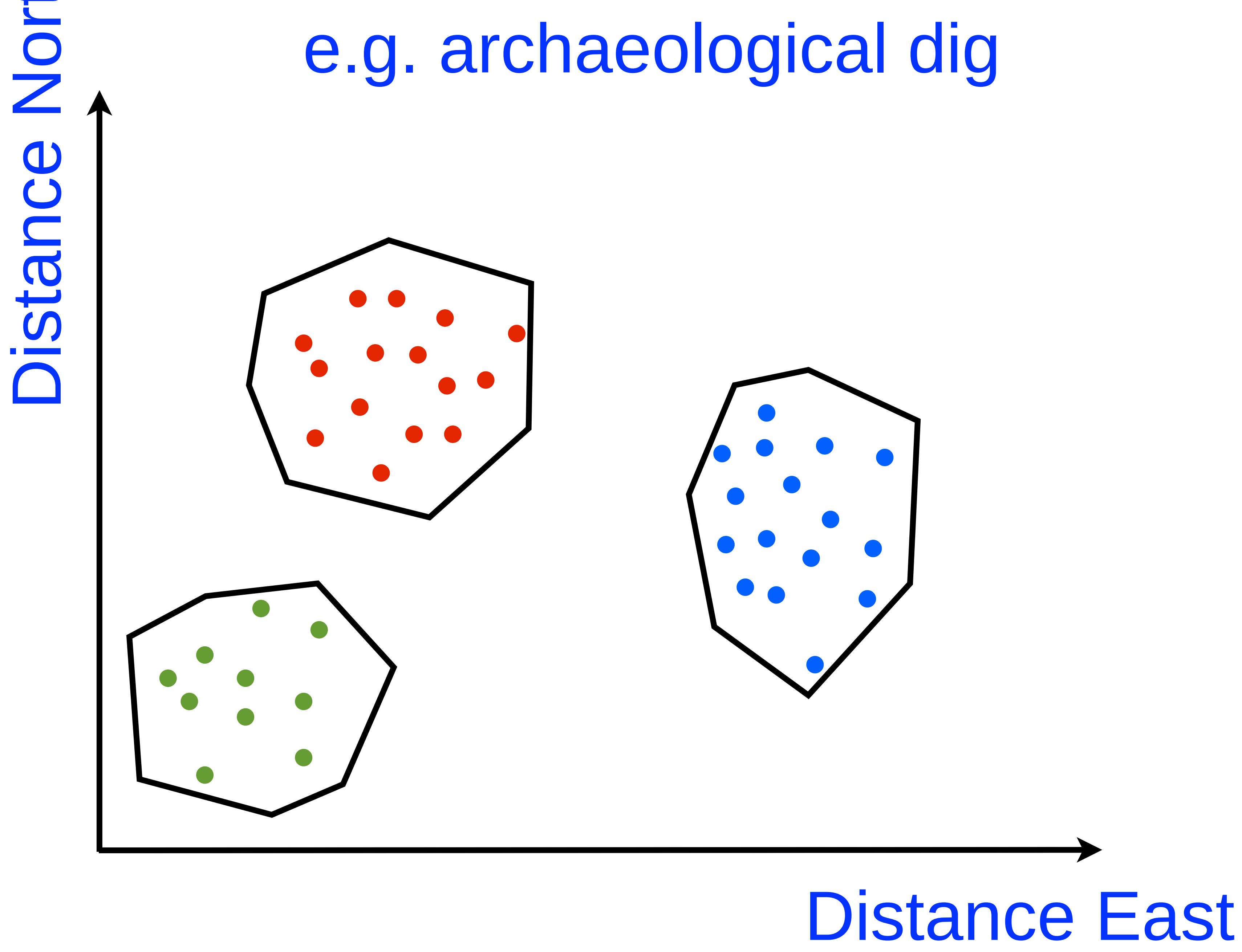
e.g. archaeological dig



slide by Tamara Broderick

Clustering

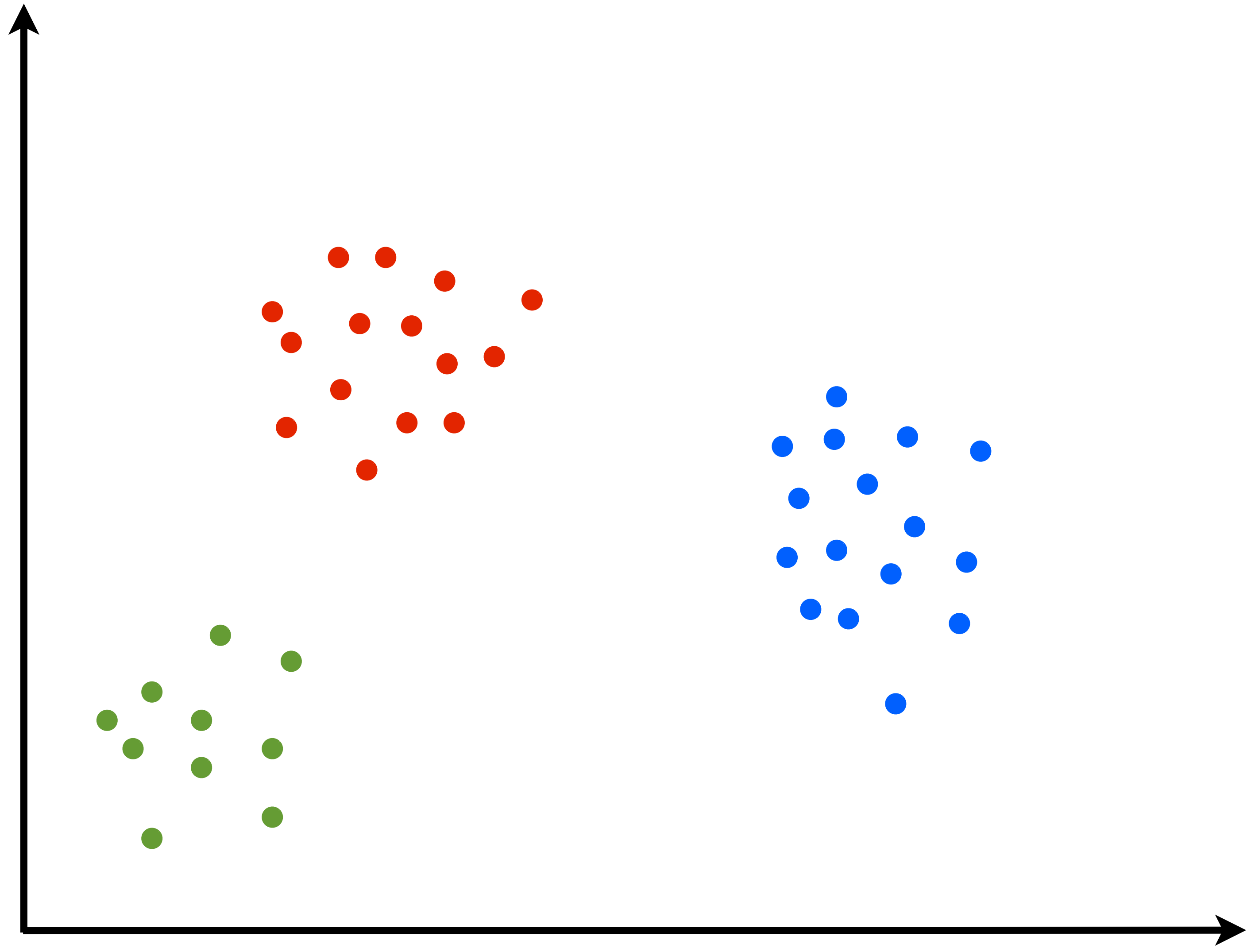
- Grouping data according to similarity



Clustering vs. Classification

- Grouping data according to similarity

  Predicting new labels from old labels

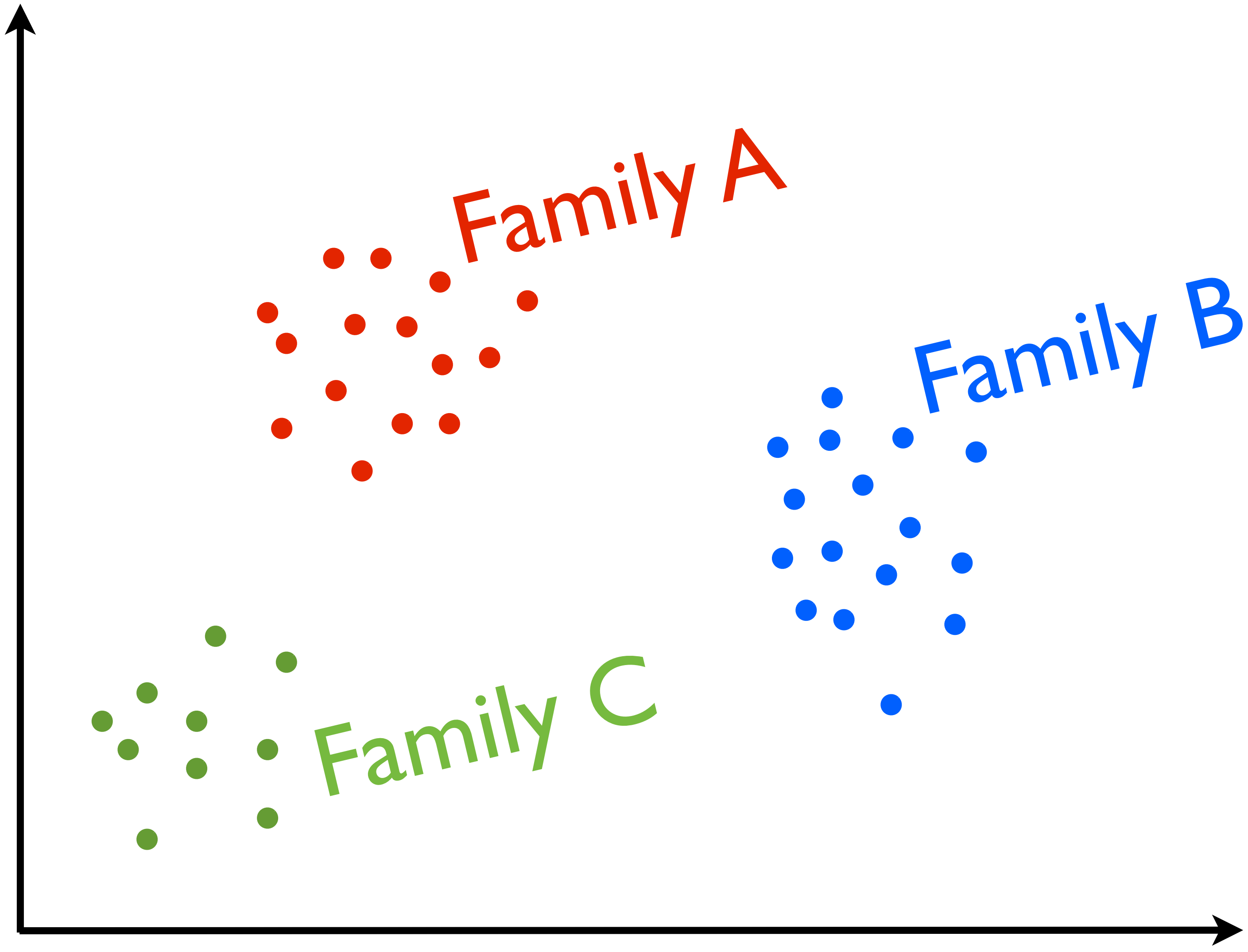


slide by Tamara Broderick

Clustering vs. Classification

- <u>Grouping</u> data according to similarity

<u>Predicting</u> new labels from old labels

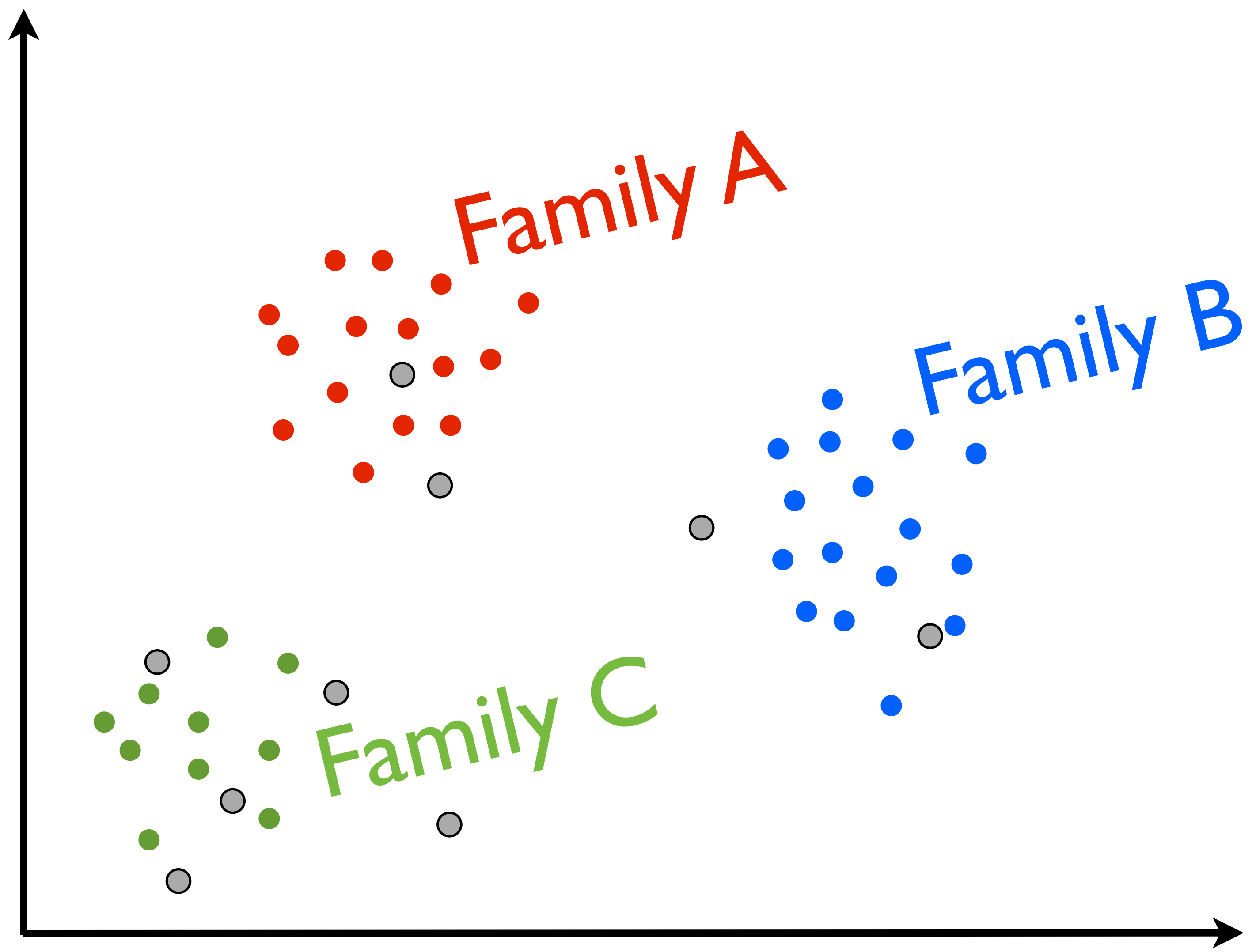


slide by Tamara Broderick

Clustering vs. Classification

- Grouping data according to similarity

Predicting new labels from old labels



slide by Tamara Broderick

Why use clustering…
…instead of classification

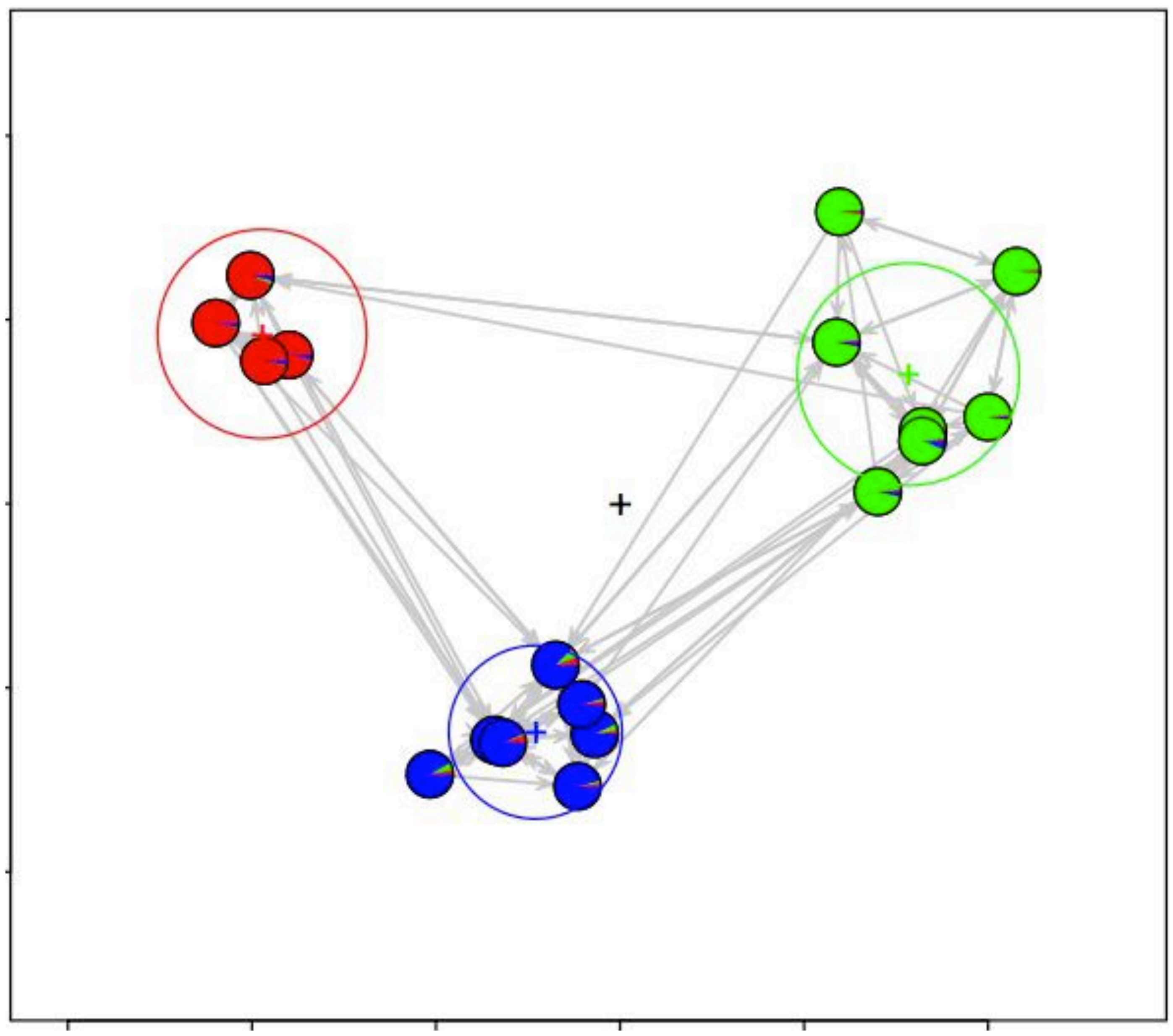
- Exploratory data analysis

slide by Tamara Broderick

[Krivitsky Handcock 2008]

Why use clustering… …instead of classification

- Exploratory data analysis

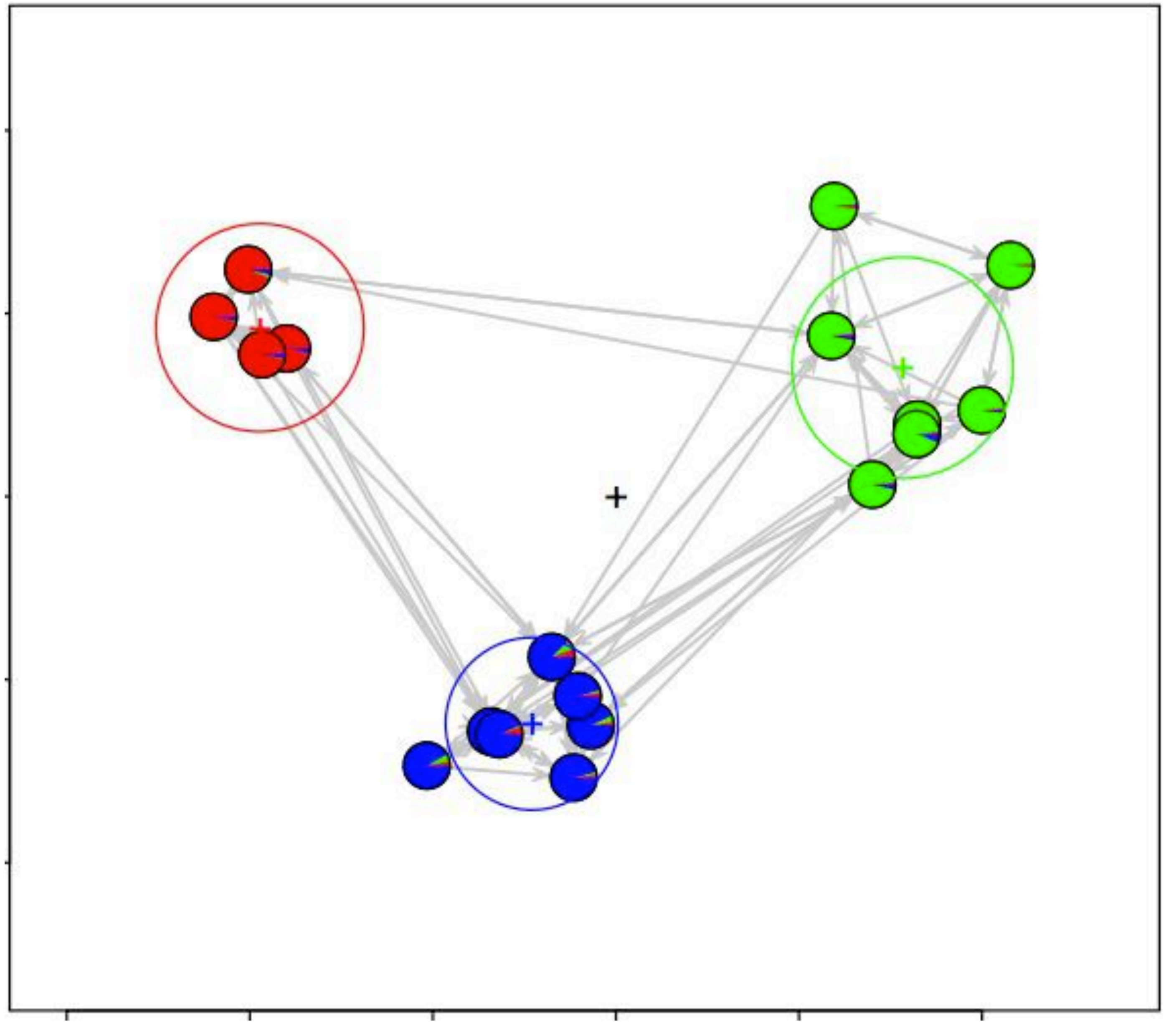


slide by Tamara Broderick

[Krivitsky Handcock 2008]

Why use clustering…
…instead of classification

- Exploratory data analysis



Datum: person

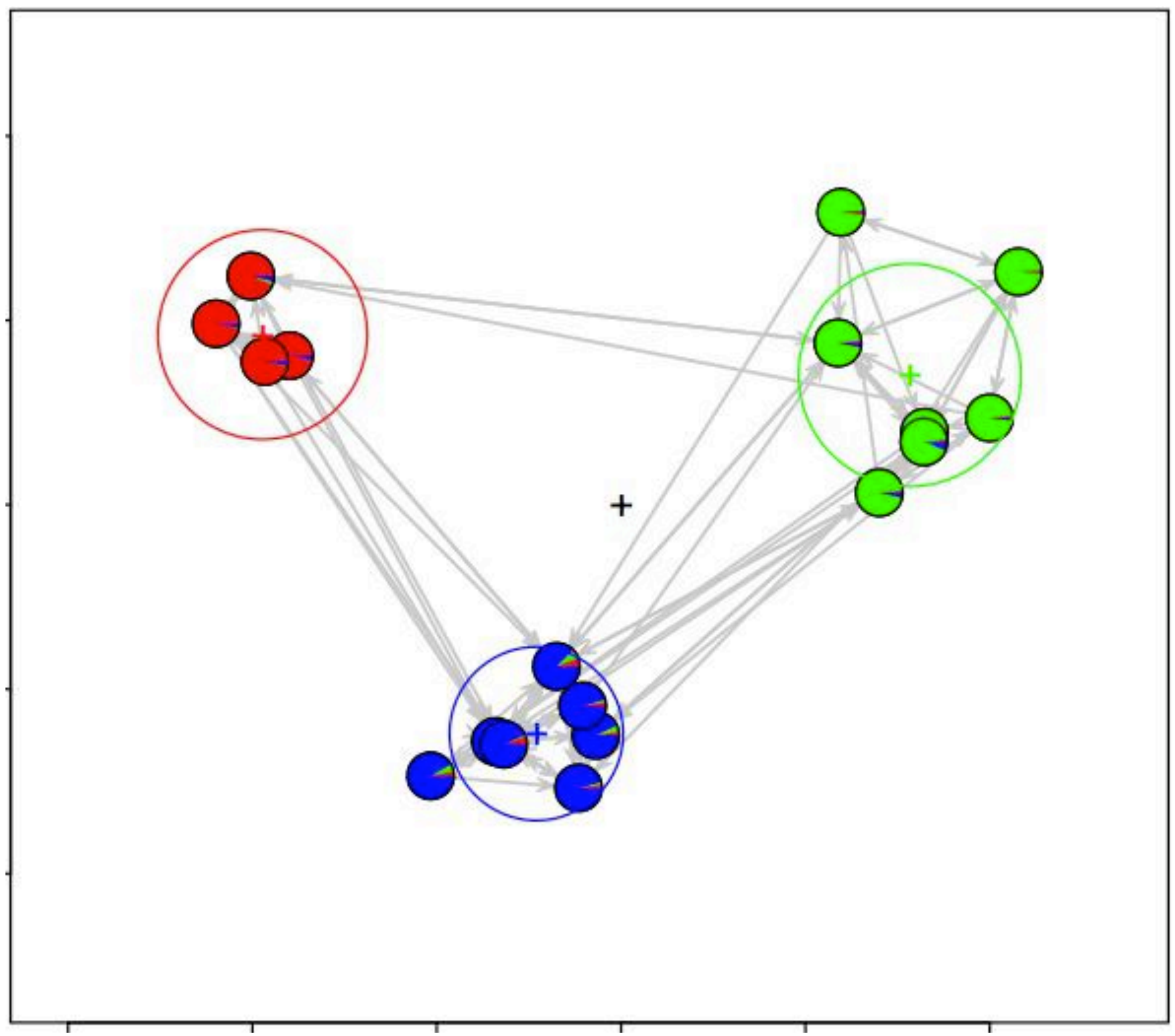
Similarity: the number of common interests of two people

slide by Tamara Broderick

[Krivitsky Handcock 2008]

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

[Krivitsky Handcock 2008]

slide by Tamara Broderick

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

earst Foundation will give $1.25 million to Lincoln Center, Metropoli-
k Philharmonic and Juilliard School. "Our board felt that we had a
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
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.

[Blei 2003]

slide by Tamara Broderick

Why use clustering… …instead of classification

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

"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

Topic Analysis

arst Foundation will give $1.25 million to Lincoln Center, Metropoli-
k Philharmonic and Juilliard School. "Our board felt that we had a
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
incoln 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
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.

[Blei 2003]

slide by Tamara Broderick

Why use clustering…
…instead of classification

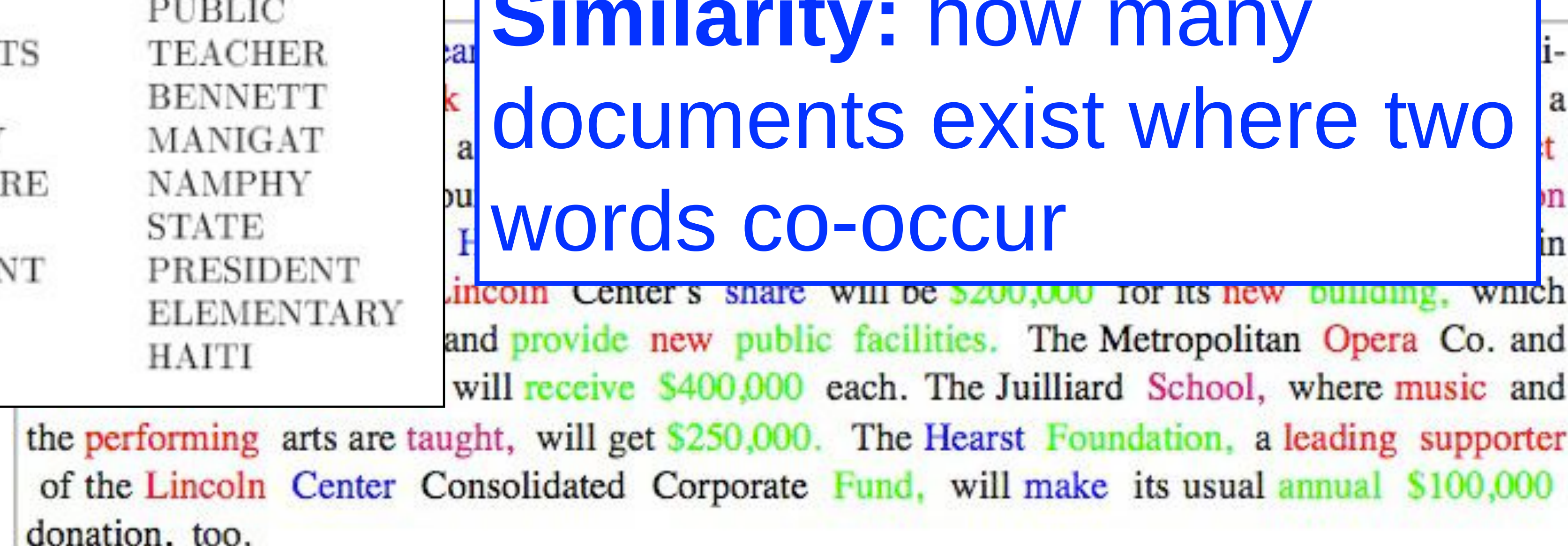
- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

"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

Topic Analysis

Datum: word

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


and provide new public facilities. The Metropolitan Opera Co. and
will receive $400,000 each. The Juilliard School, where music and
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.

[Blei 2003]

slide by Tamara Broderick

Why use clustering…
…instead of classification

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

"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

Topic Analysis

Datum: binary vector indicating document occurrence

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

the performing arts are taug

of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

[Blei 2003]

slide by Tamara Broderick

Why use clustering…
…instead of classification

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

slide by Tamara Broderick

Why use clustering…
…instead of classification

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

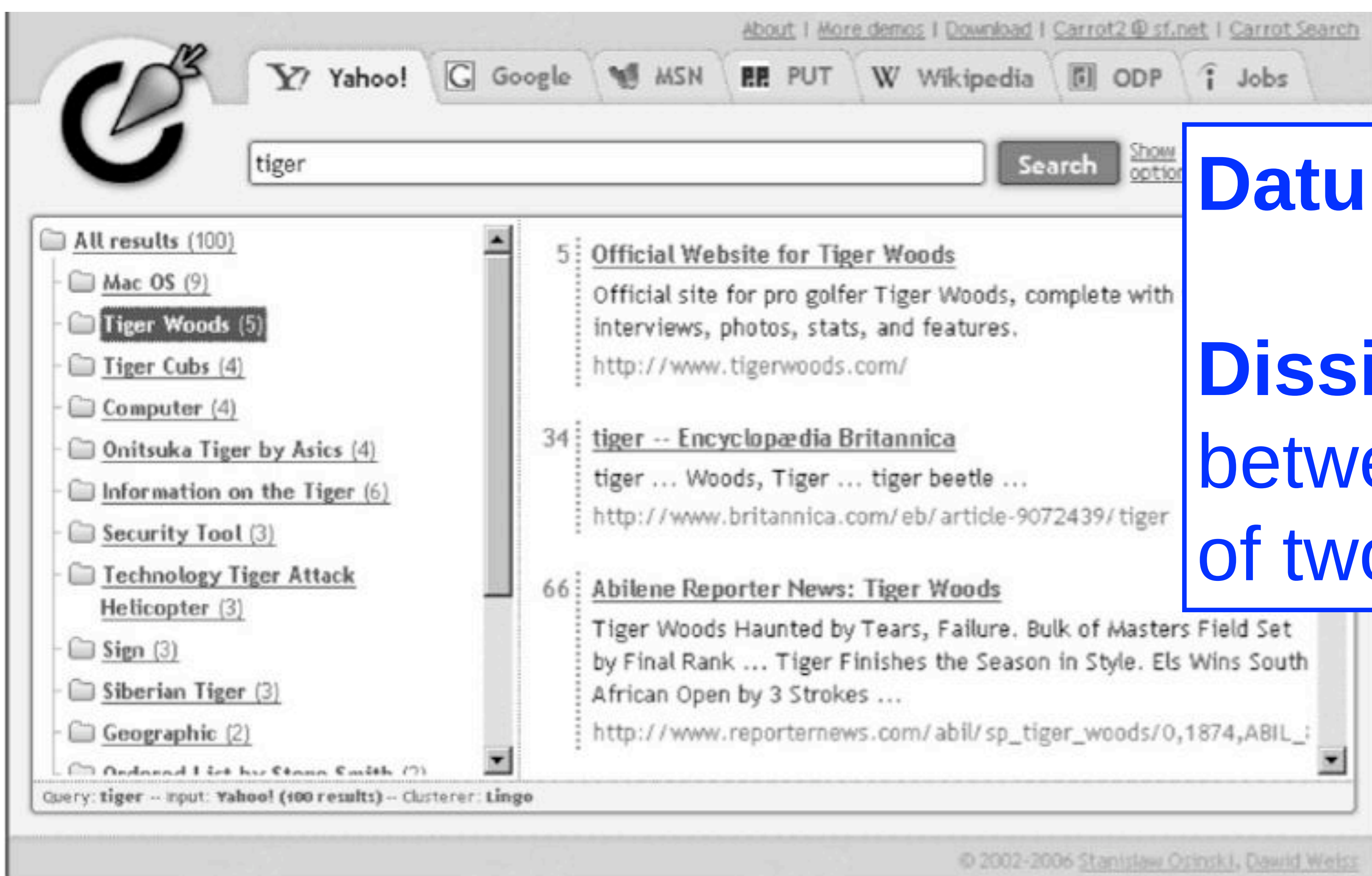
Document clustering

[Carpineto et al. 2009]

slide by Tamara Broderick

Why use clustering…
…instead of classification

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



Document clustering

Datum: document

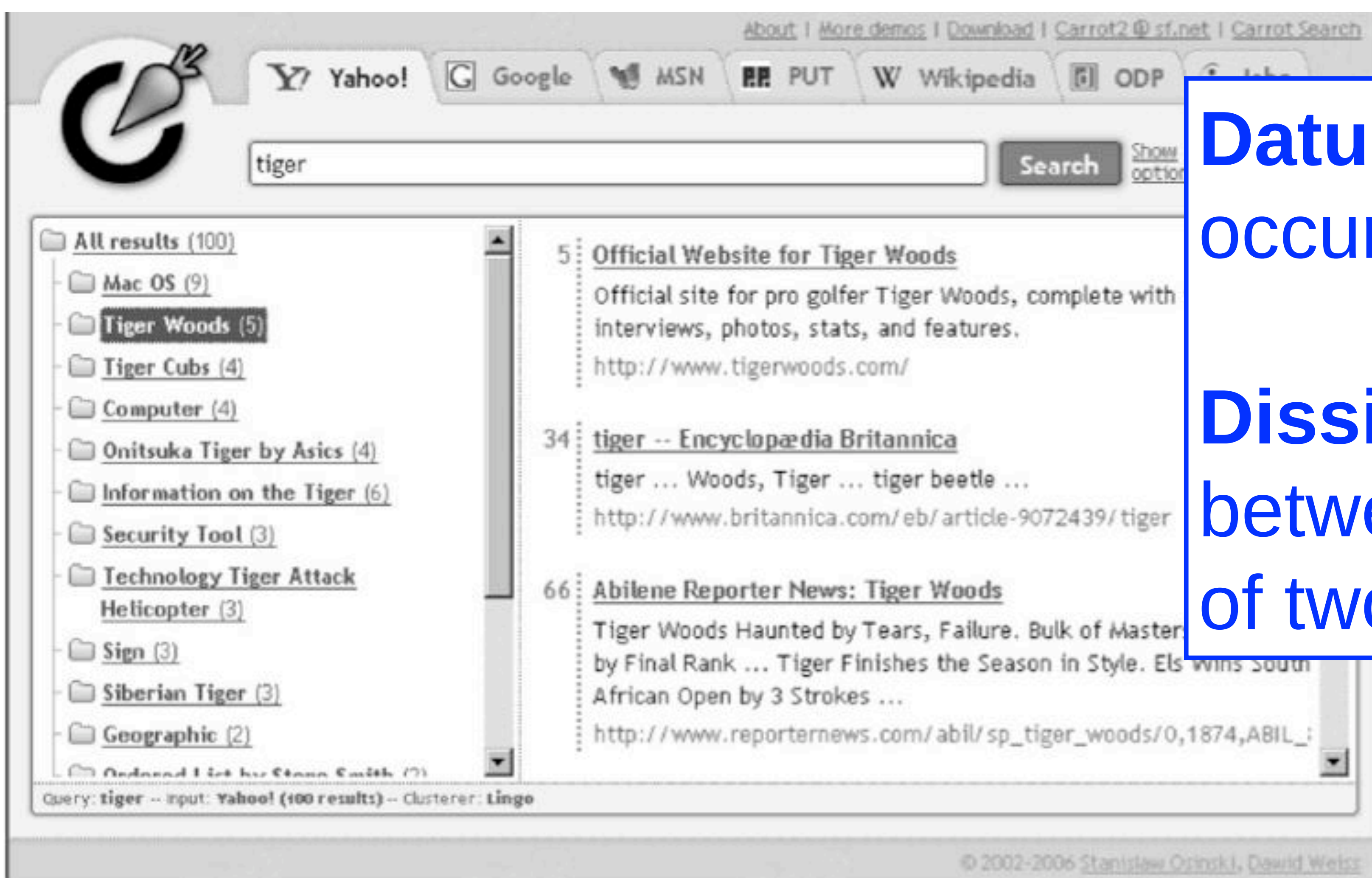
Dissimilarity: distance between topic distributions of two documents

[Carpineto et al. 2009]

slide by Tamara Broderick

Why use clustering…
…instead of classification

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



<u>Document clustering</u>

Datum: vector of topic occurrences

Dissimilarity: distance between topic distributions of two documents

[Carpineto et al. 2009]

slide by Tamara Broderick

Why use clustering… …instead of classification

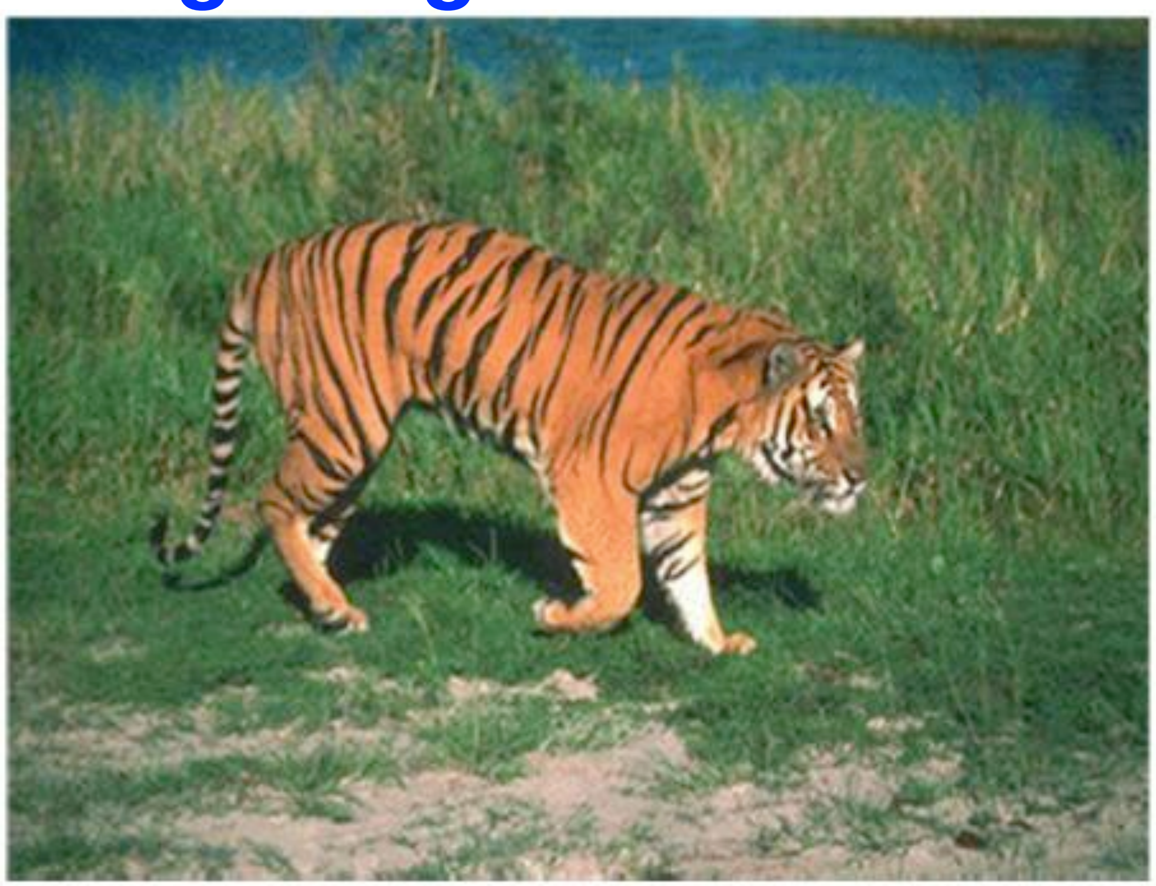
- Exploratory data analysis
- Classes are unspecified (unknown, changing too quickly, expensive to label data, etc)

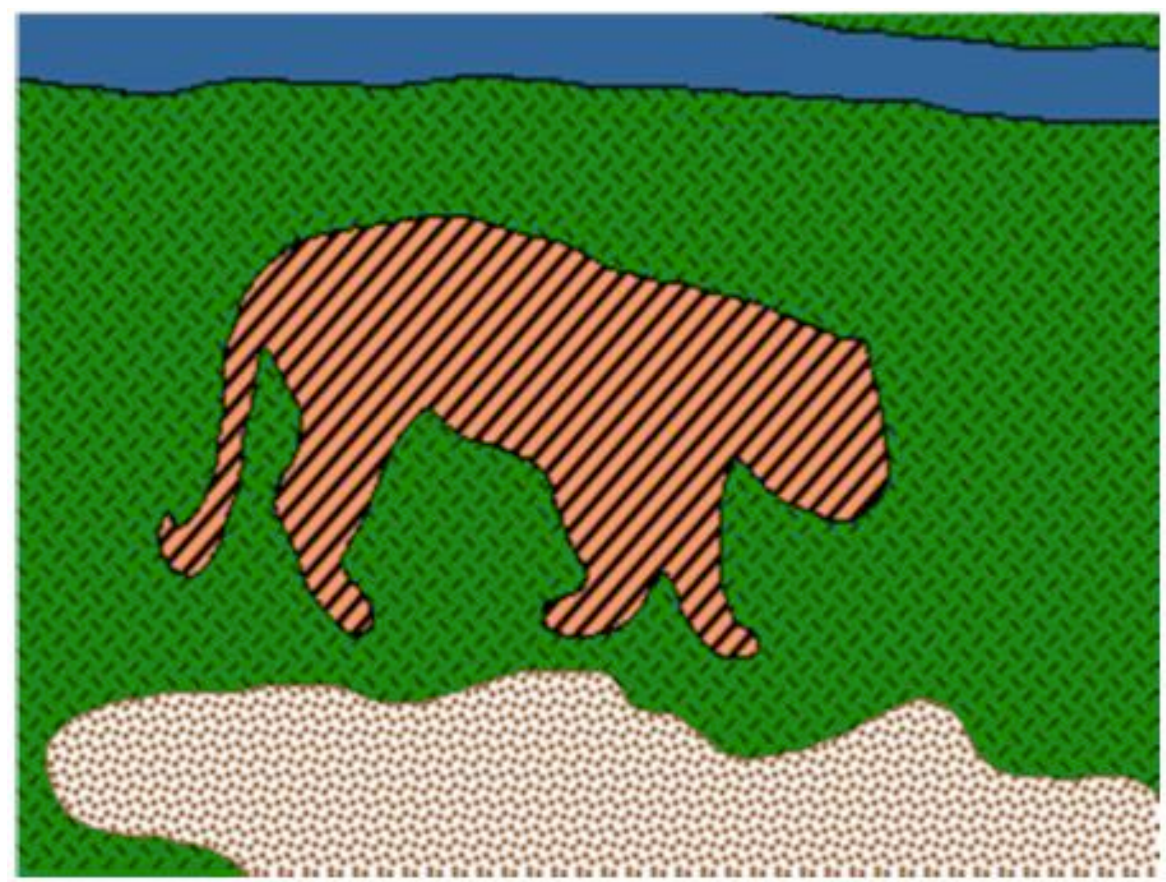
slide by Tamara Broderick

Why use clustering…
…instead of classification

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

Image segmentation





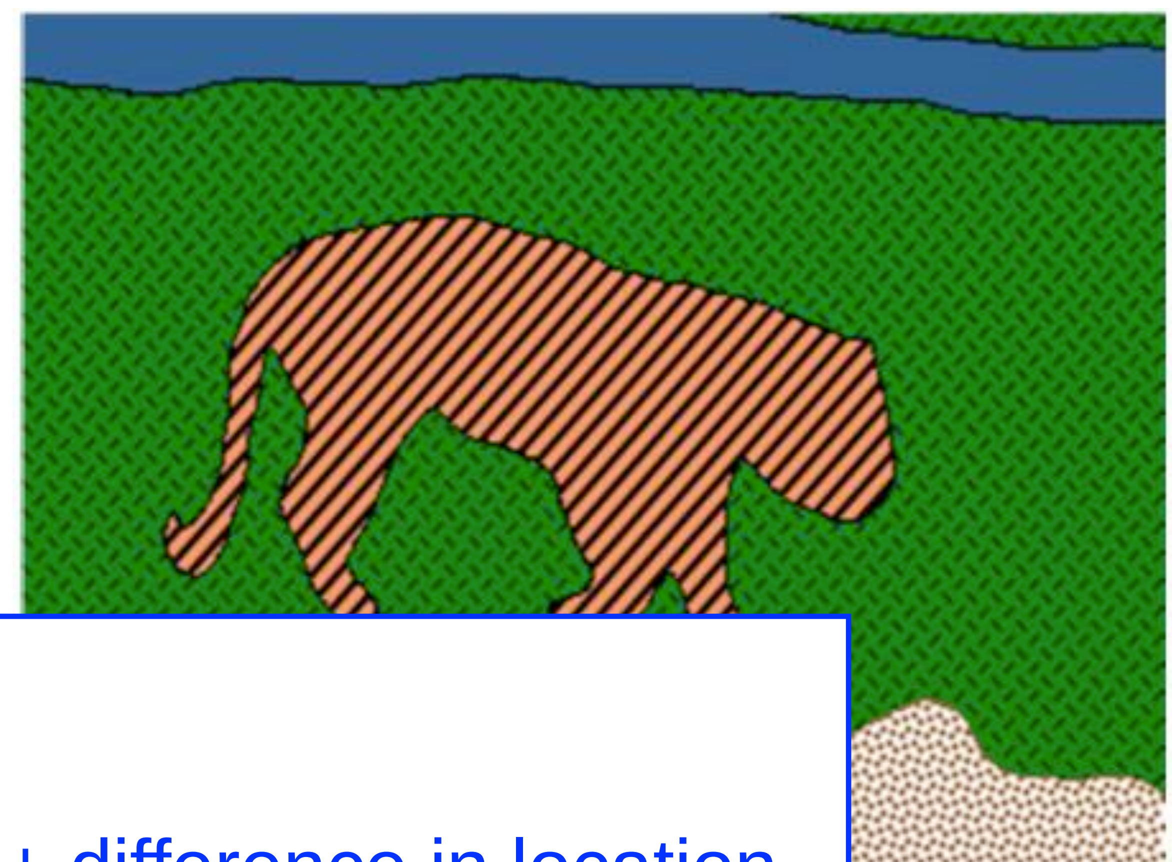
[Fei-Fei 2011]

slide by Tamara Broderick

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

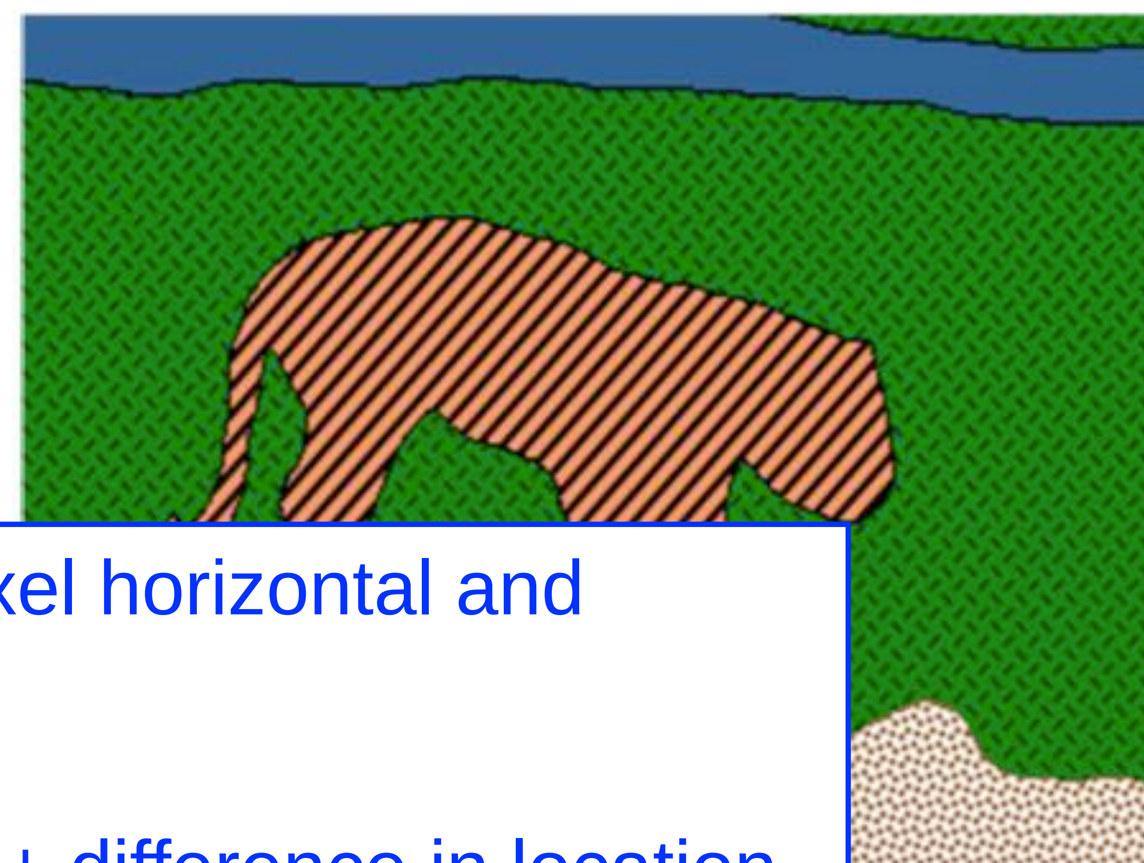
[Fei-Fei 2011]

slide by Tamara Broderick

Why use clustering…
…instead of classification

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

Image segmentation



Datum: pixel RGB values and pixel horizontal and vertical locations

Dissimilarity: difference in color + difference in location

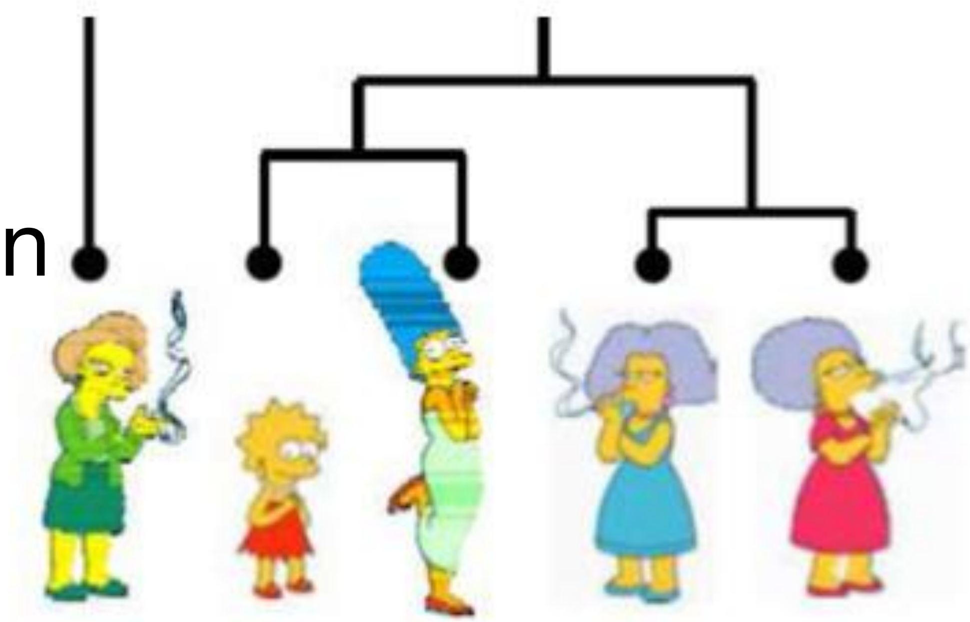
[Fei-Fei 2011]

slide by Tamara Broderick

Clustering algorithms

- Partitioning algorithms
  - Construct various partitions and then evaluate them by some criterion
    - K-means
    - Mixture of Gaussians
    - Spectral Clustering

- Hierarchical algorithms
  - Create a hierarchical decomposition of the set of objects using some criterion
  - Bottom-up – agglomerative
  - Top-down – divisive



slide by Eric Xing

Desirable Properties of a Clustering Algorithm

- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Ability to deal with noisy data
- Interpretability and usability

- Optional
  - Incorporation of user-specified constraints

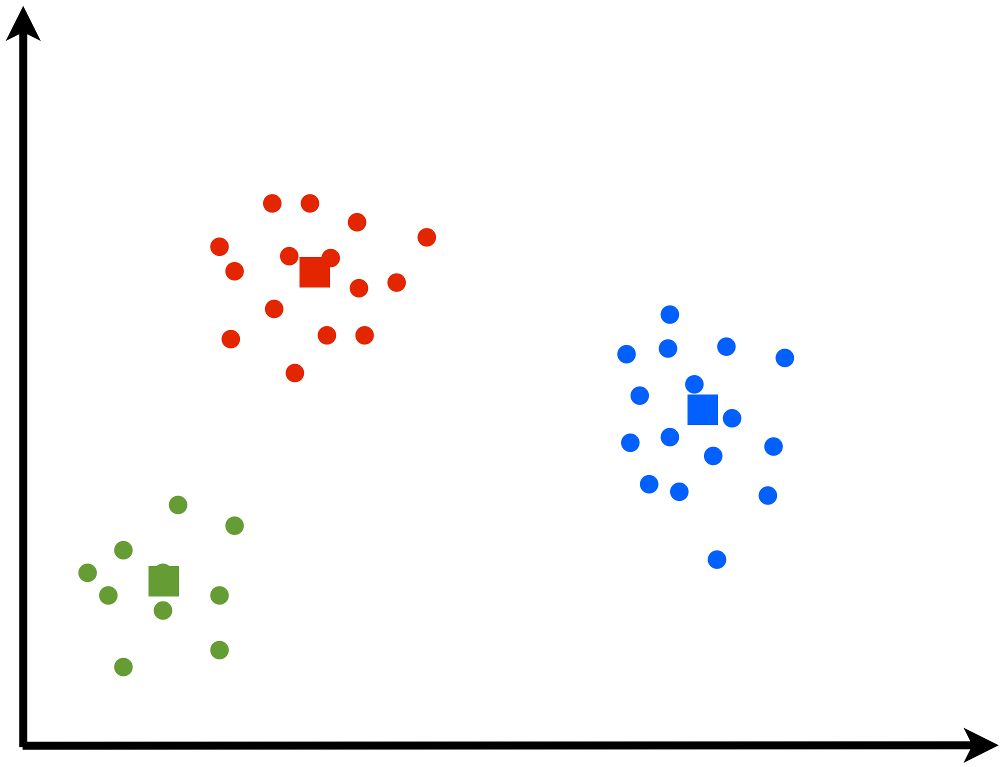
slide by Andrew Moore

K-Means Clustering

K-Means Clustering

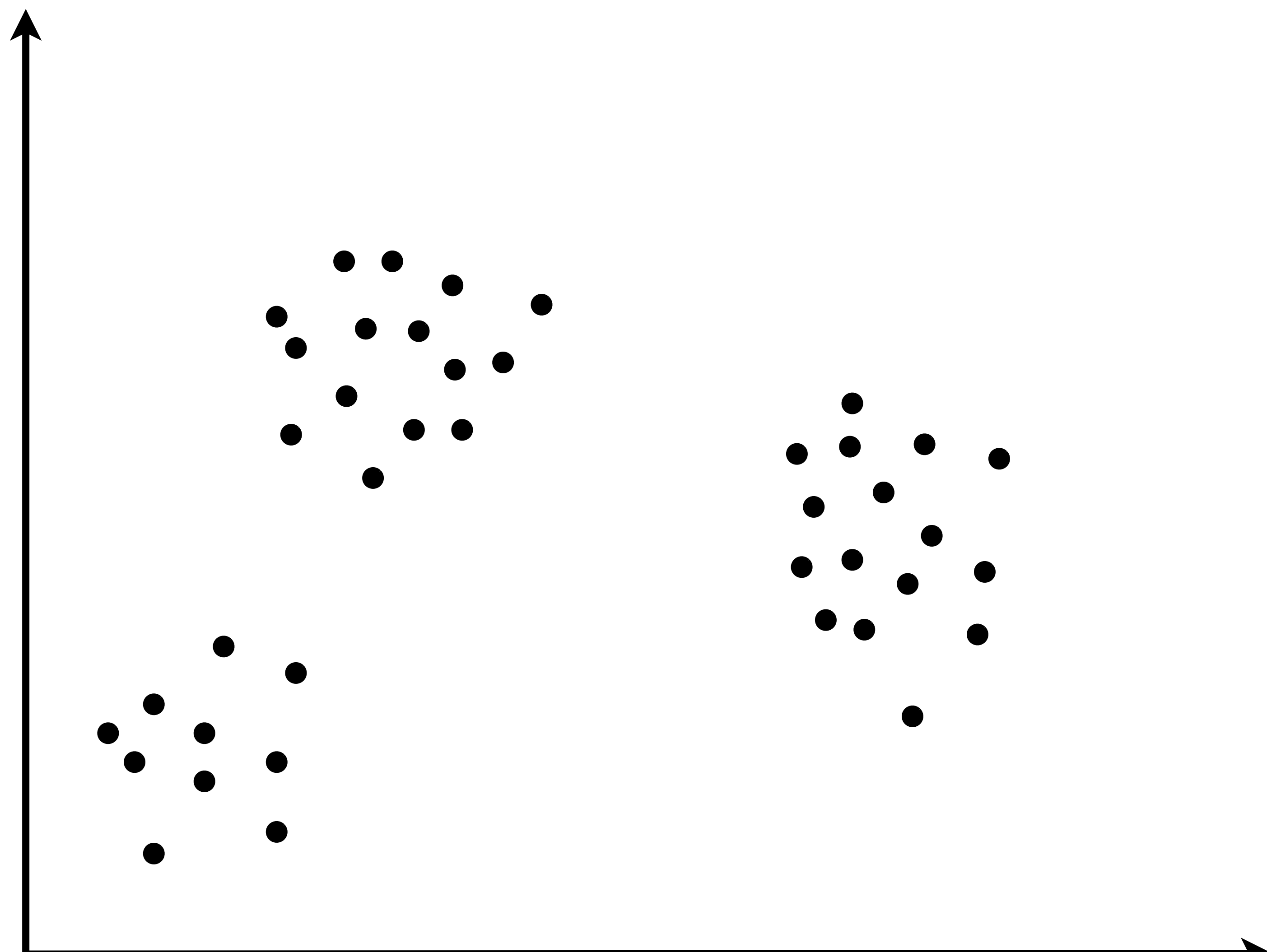
Benefits

- Fast
- Conceptually straightforward
- Popular



slide by Tamara Broderick

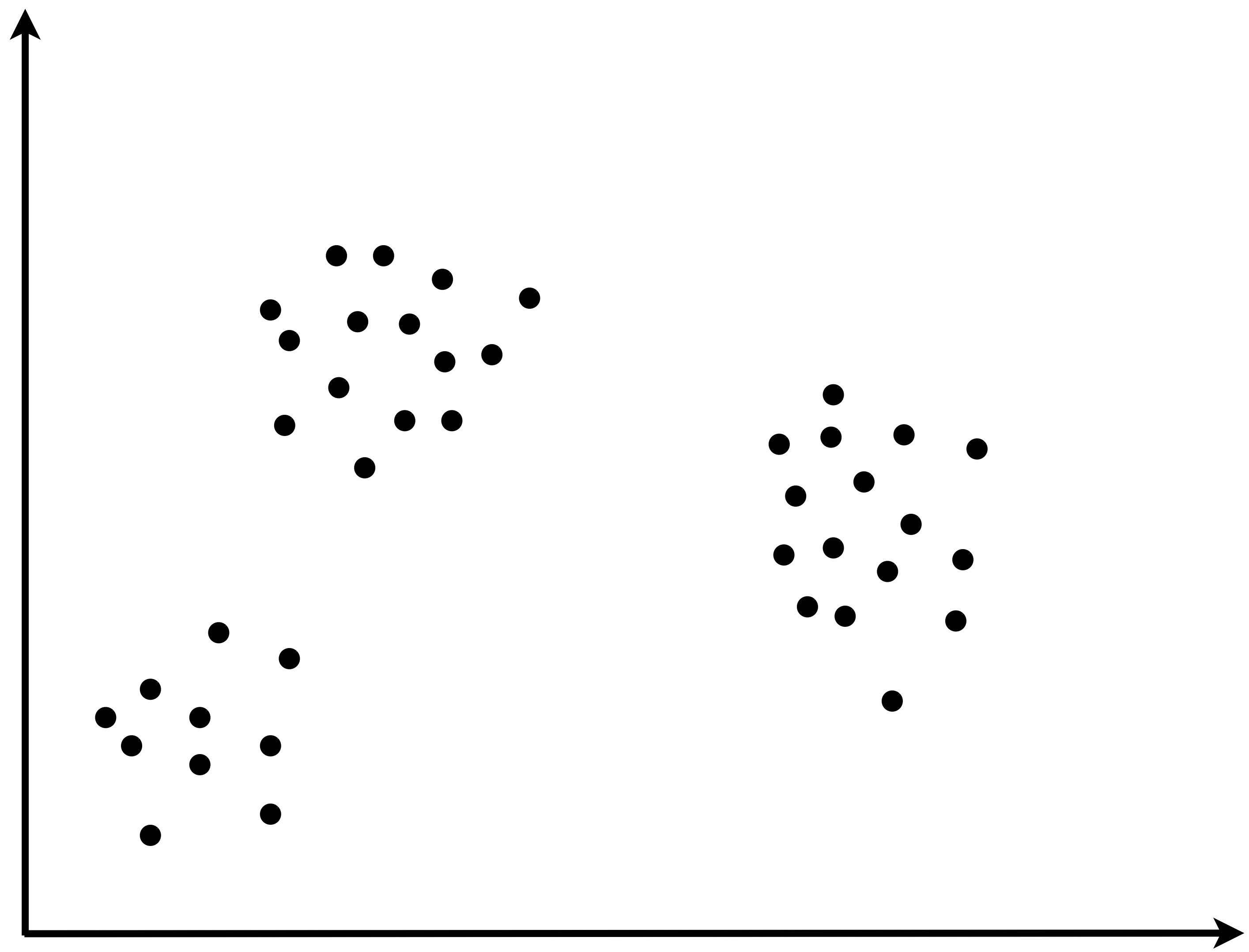
K-Means: Preliminaries



slide by Tamara Broderick

K-Means: Preliminaries

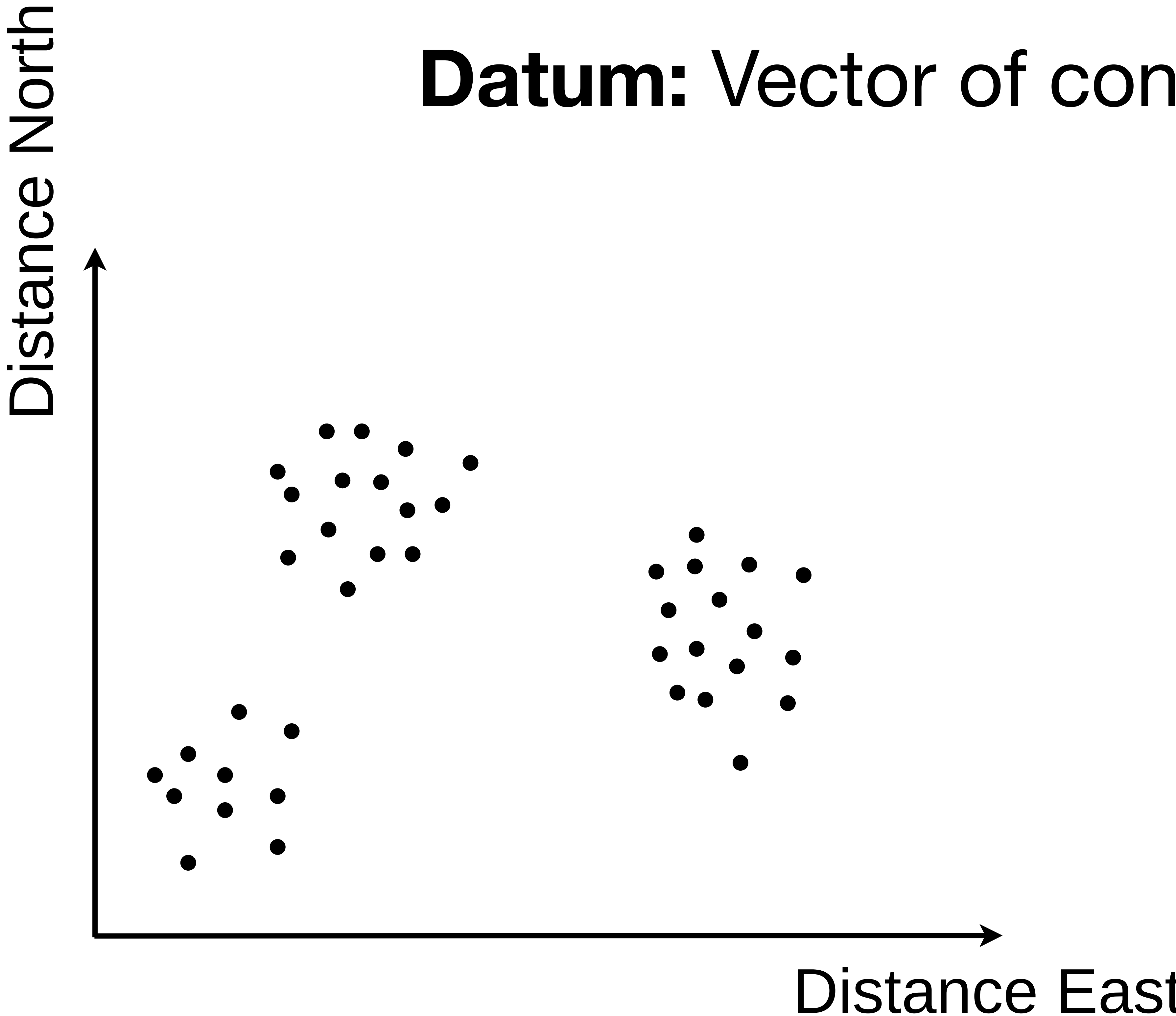
Datum: Vector of continuous values



slide by Tamara Broderick

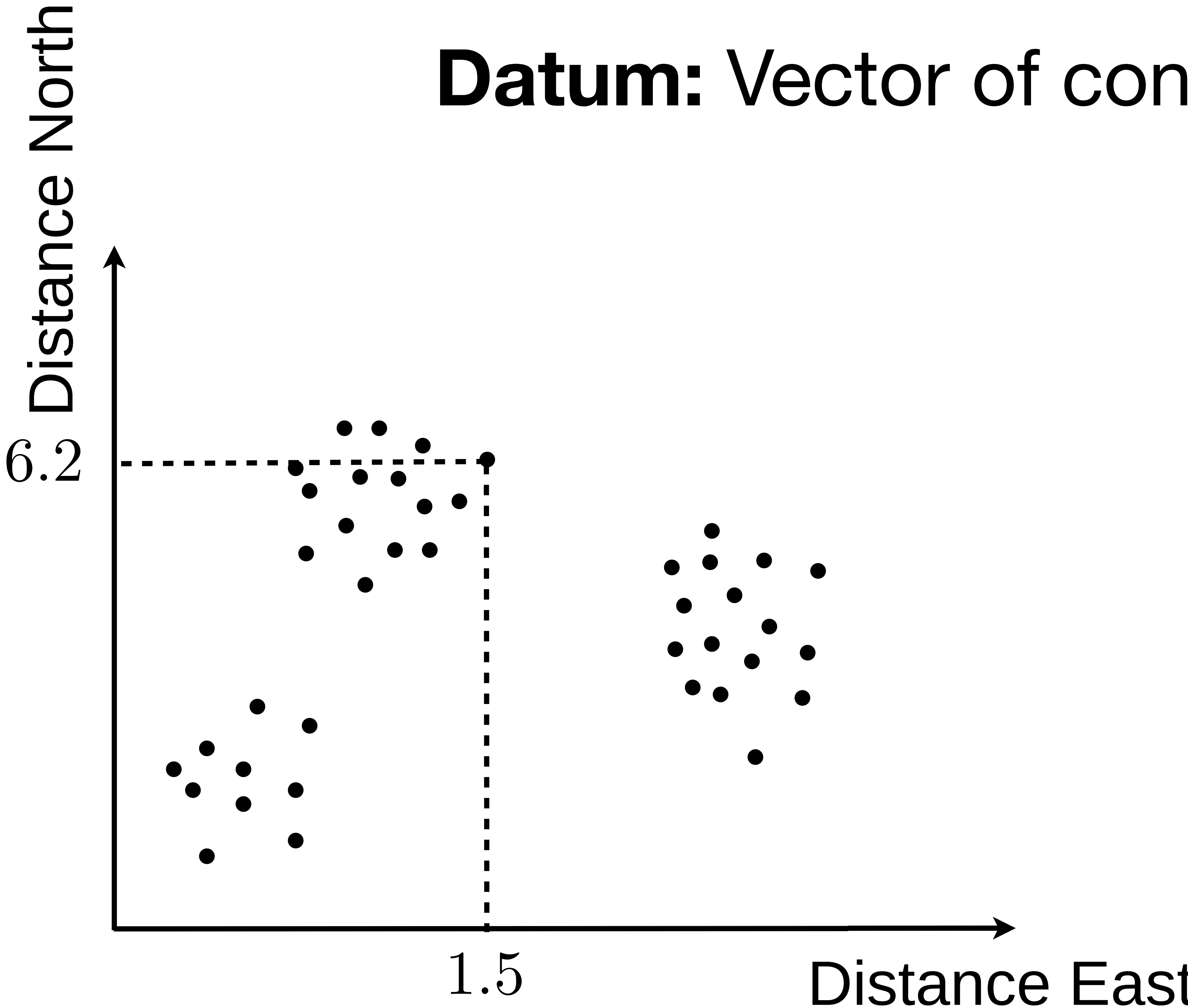
K-Means: Preliminaries

Datum: Vector of continuous values



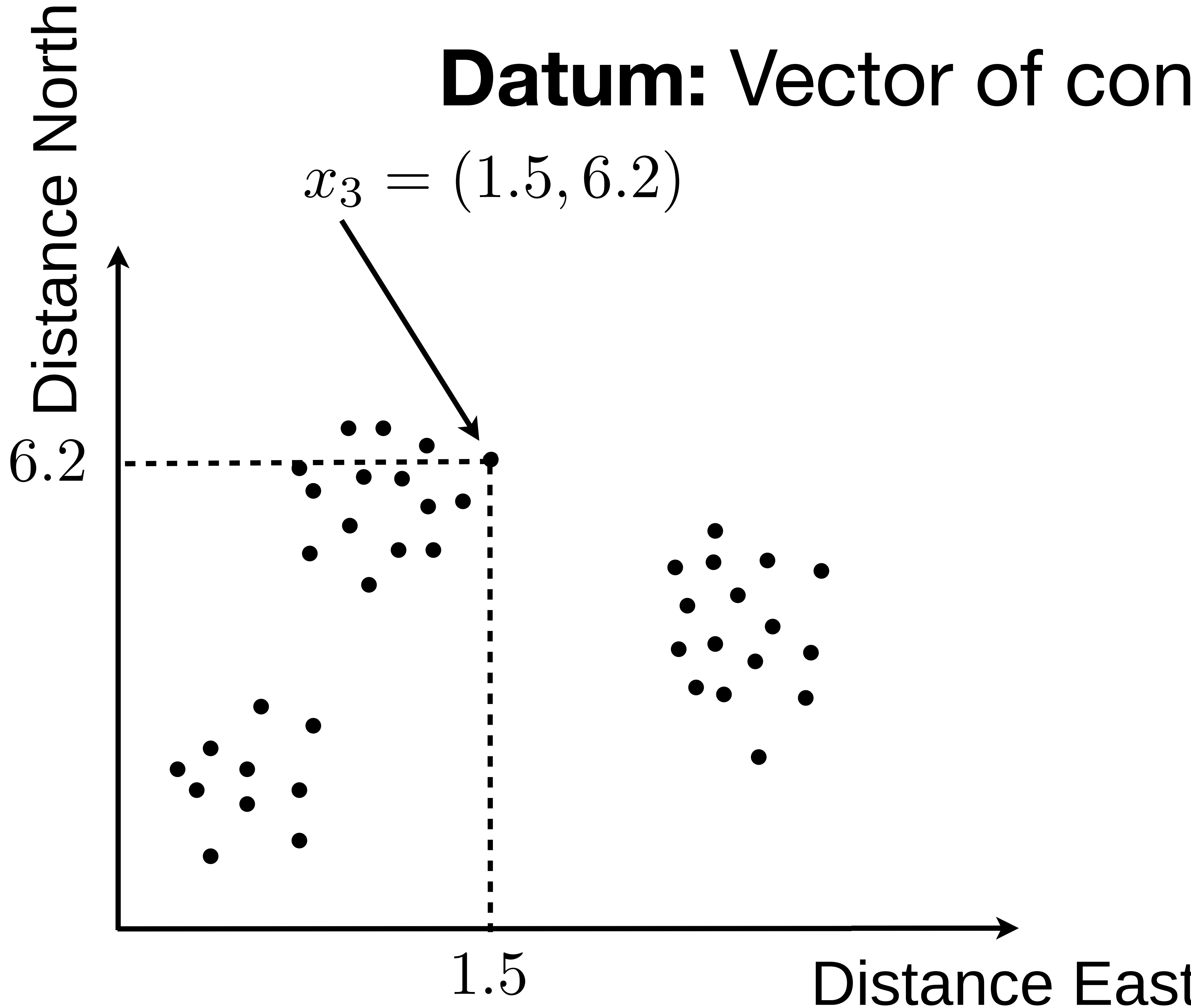
K-Means: Preliminaries

Datum: Vector of continuous values



K-Means: Preliminaries

Datum: Vector of continuous values

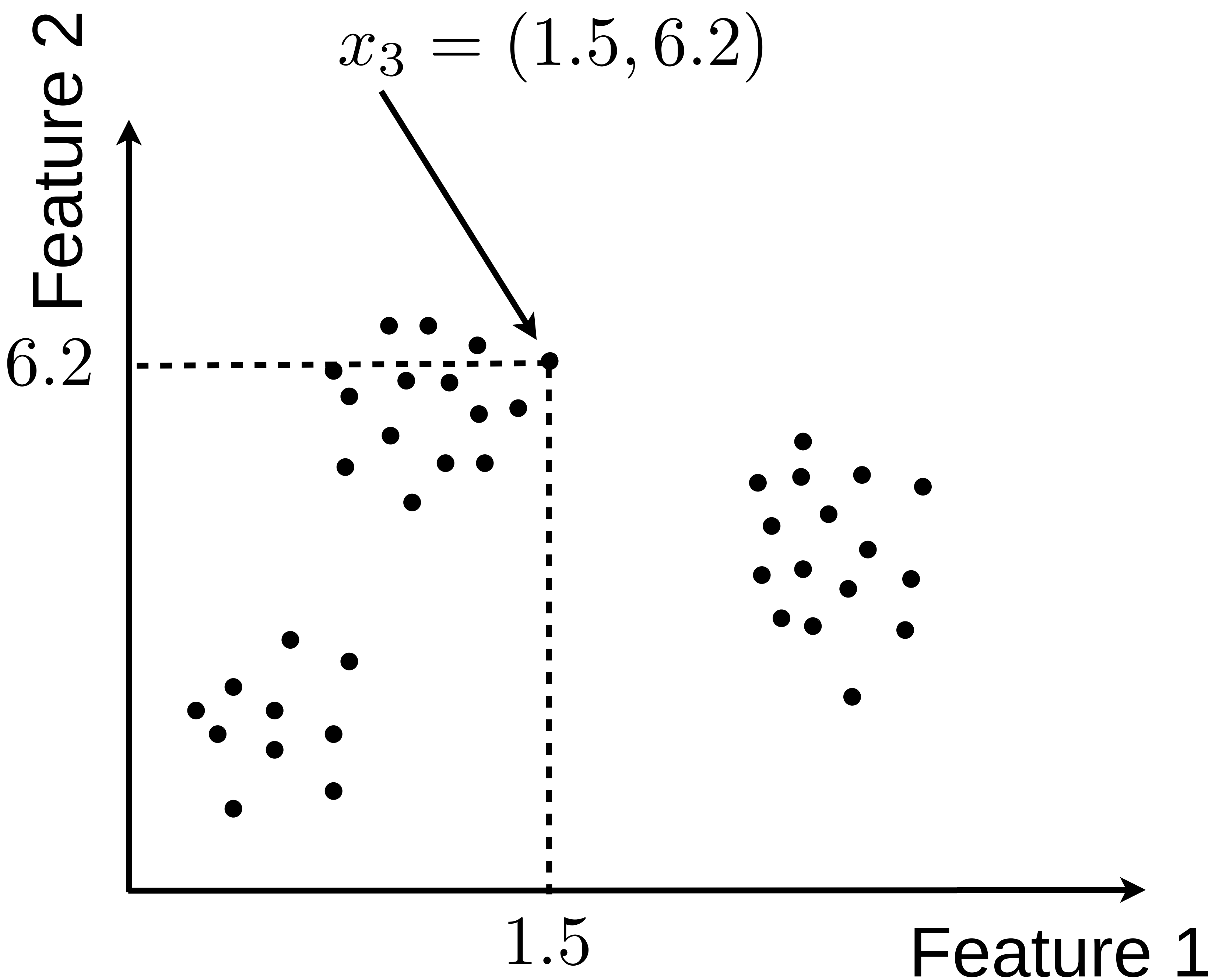


	North	East
x_1	1.2	5.9
x_2	4.3	2.1
x_3	1.5	6.3
$\vdots$		
x_N	4.1	2.3

slide by Tamara Broderick

K-Means: Preliminaries

Datum: Vector of continuous values

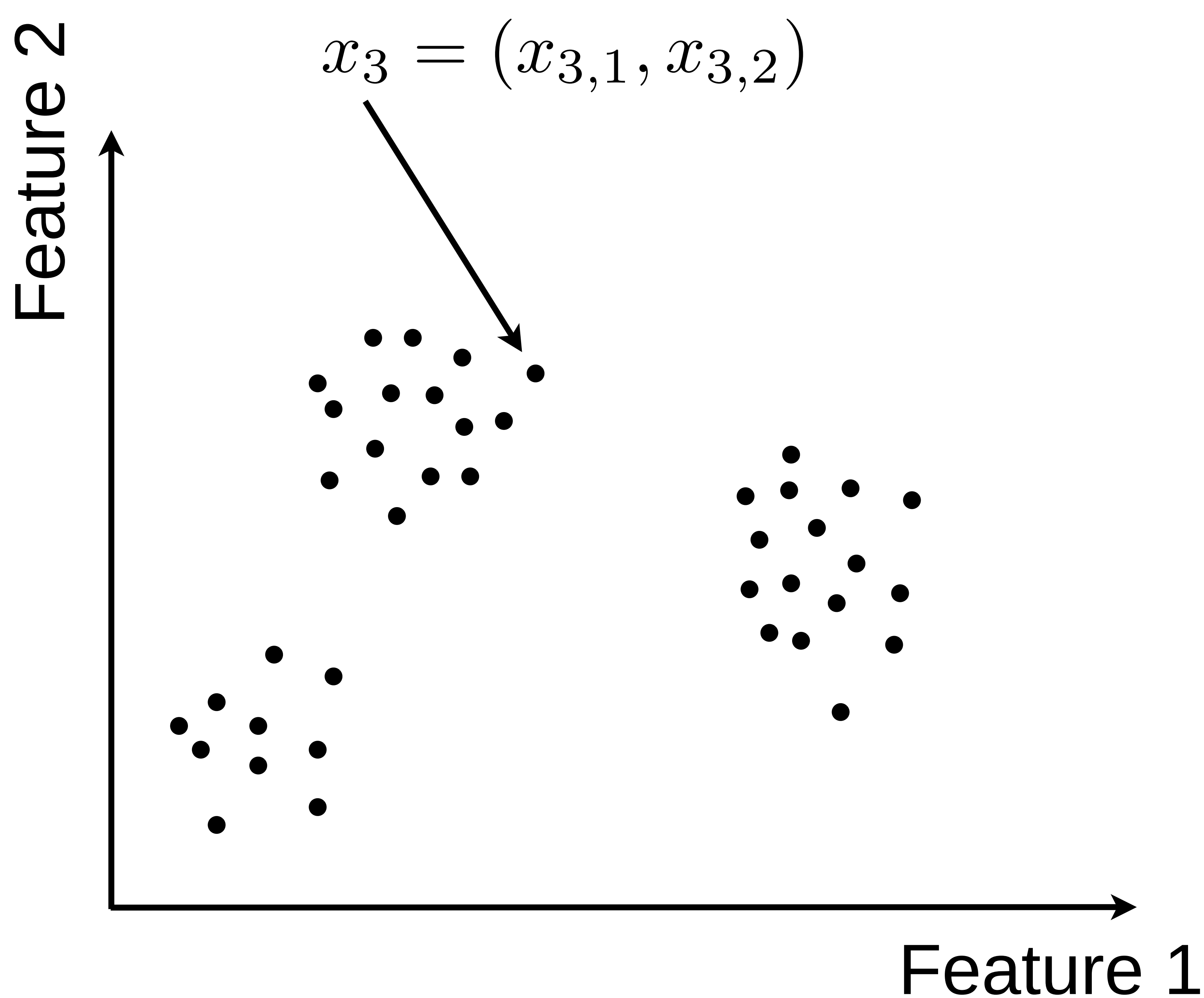


	Feature 1	Feature 2
x_1	1.2	5.9
x_2	4.3	2.1
x_3	1.5	6.3
$\vdots$		
x_N	4.1	2.3

slide by Tamara Broderick

K-Means: Preliminaries

Datum: Vector of continuous values

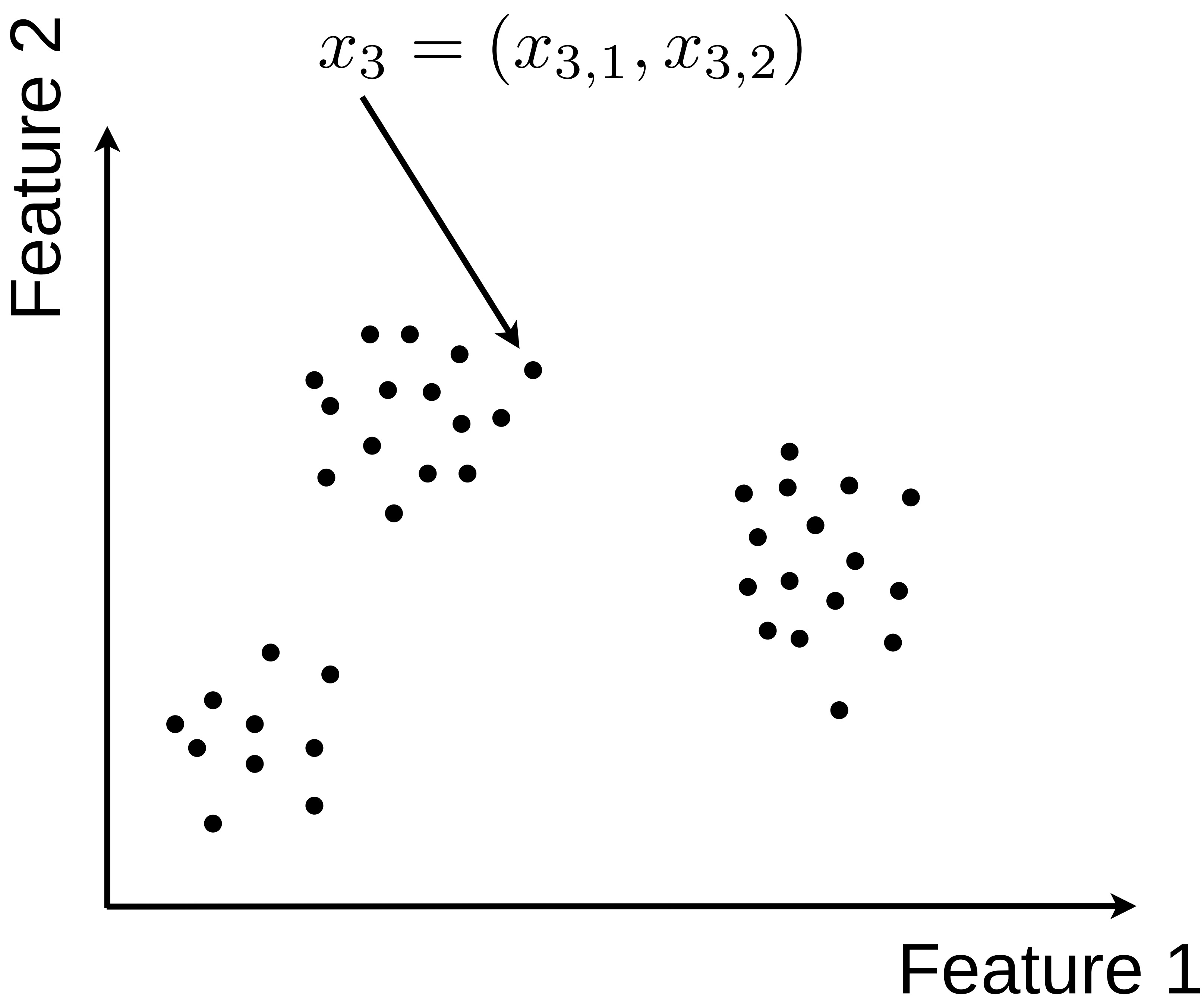


	Feature 1	Feature 2
x_1	$x_{1,1}$	$x_{1,2}$
x_2	$x_{2,1}$	$x_{2,2}$
x_3	$x_{3,1}$	$x_{3,2}$
$\vdots$		
x_N	$x_{N,1}$	$x_{N,2}$

slide by Tamara Broderick

K-Means: Preliminaries

Datum: Vector of D continuous values

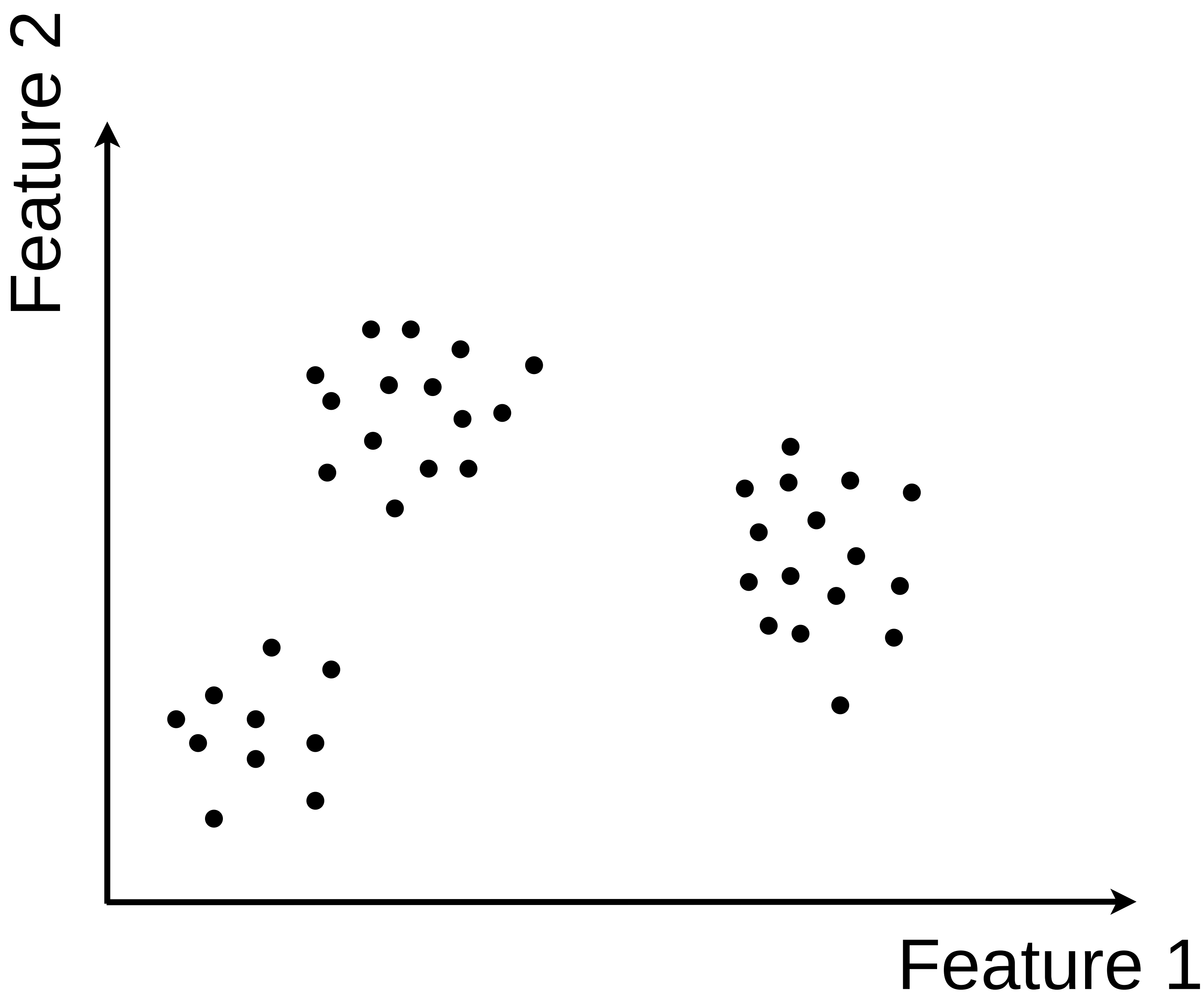


	Feature 1	Feature 2
x_1	$x_{1,1}$	$x_{1,2}$
x_2	$x_{2,1}$	$x_{2,2}$
x_3	$x_{3,1}$	$x_{3,2}$
$\vdots$		
x_N	$x_{N,1}$	$x_{N,2}$

slide by Tamara Broderick

K-Means: Preliminaries

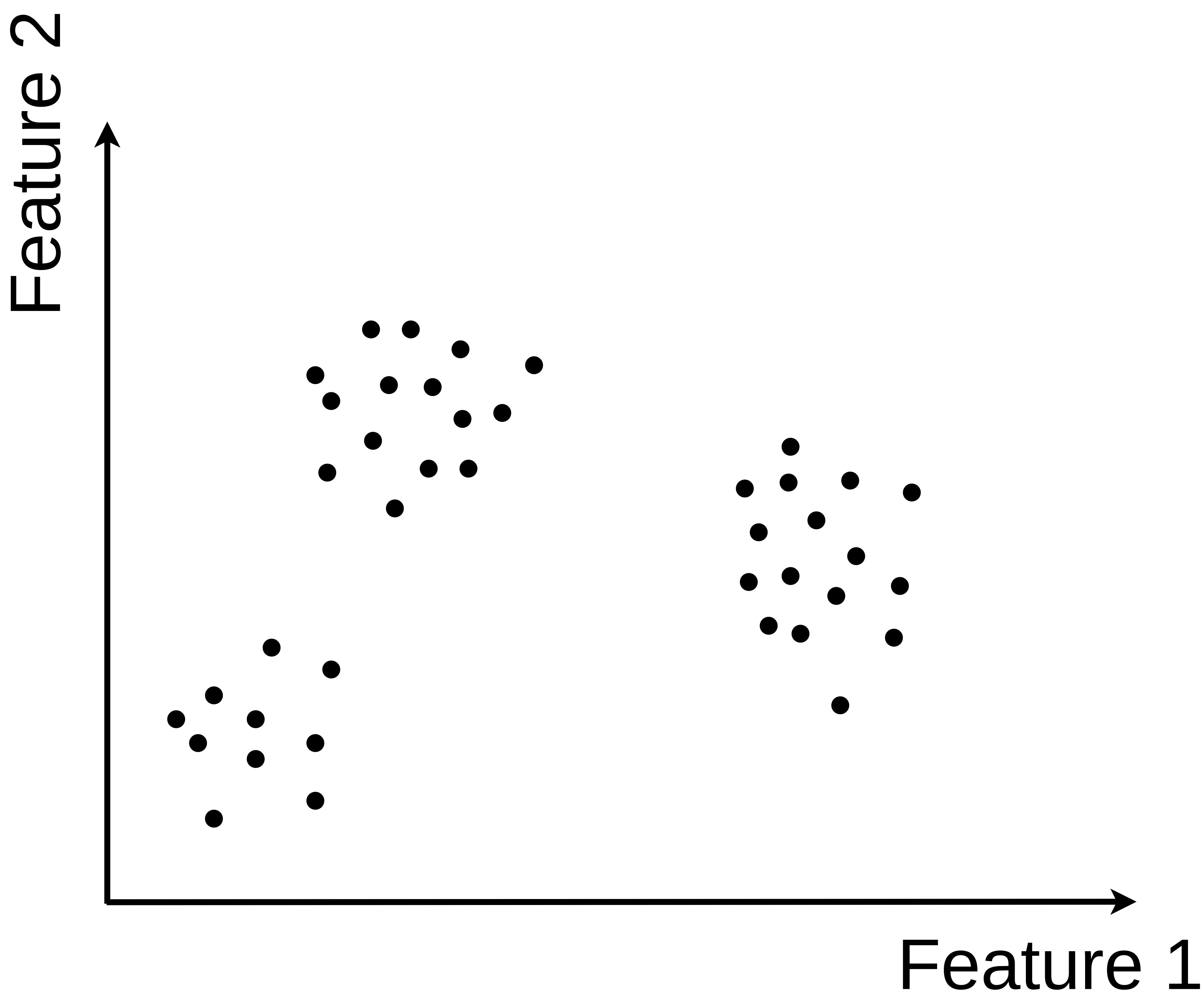
Datum: Vector of D continuous values



slide by Tamara Broderick

K-Means: Preliminaries

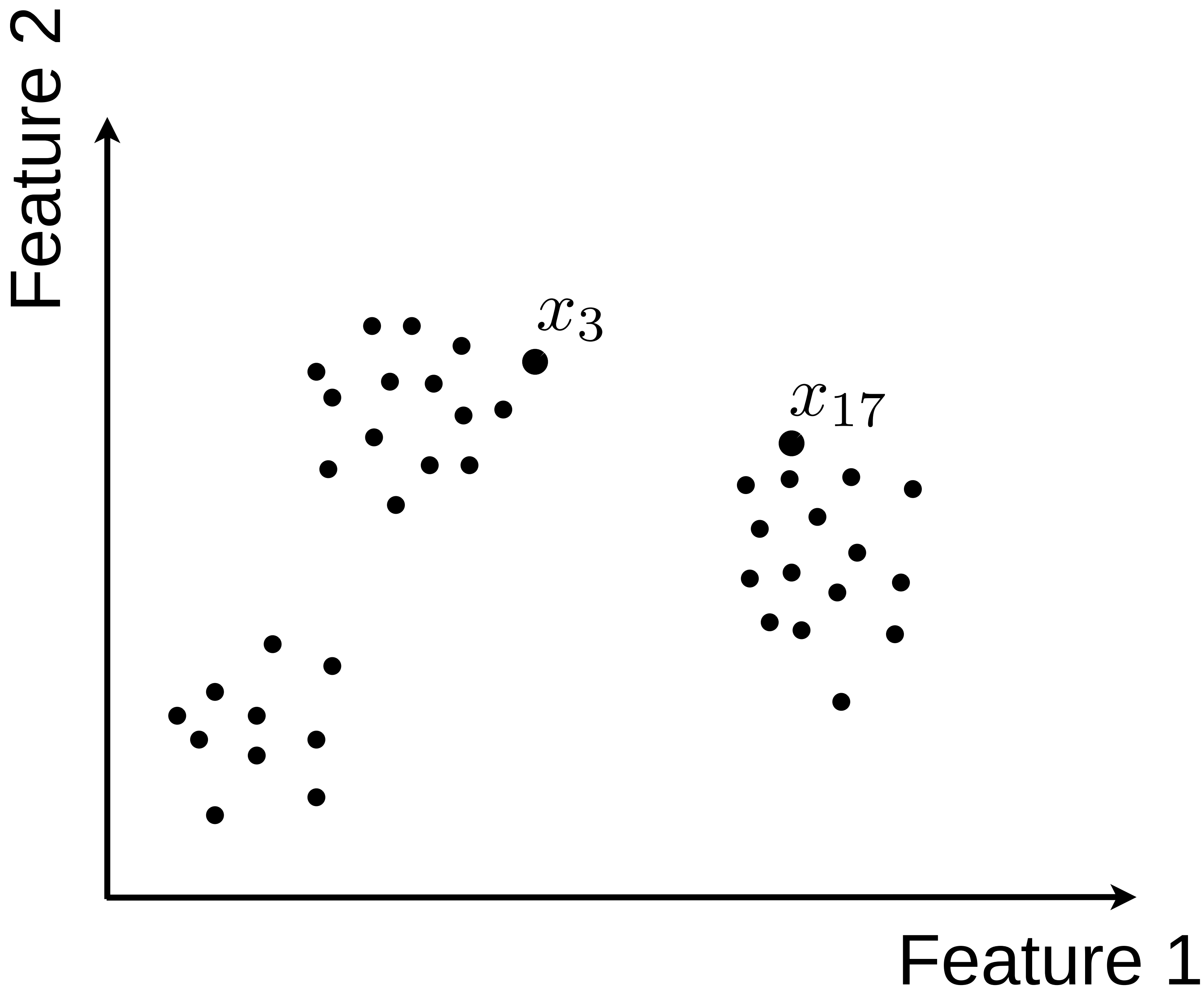
Dissimilarity: Distance as the crow flies



slide by Tamara Broderick

K-Means: Preliminaries

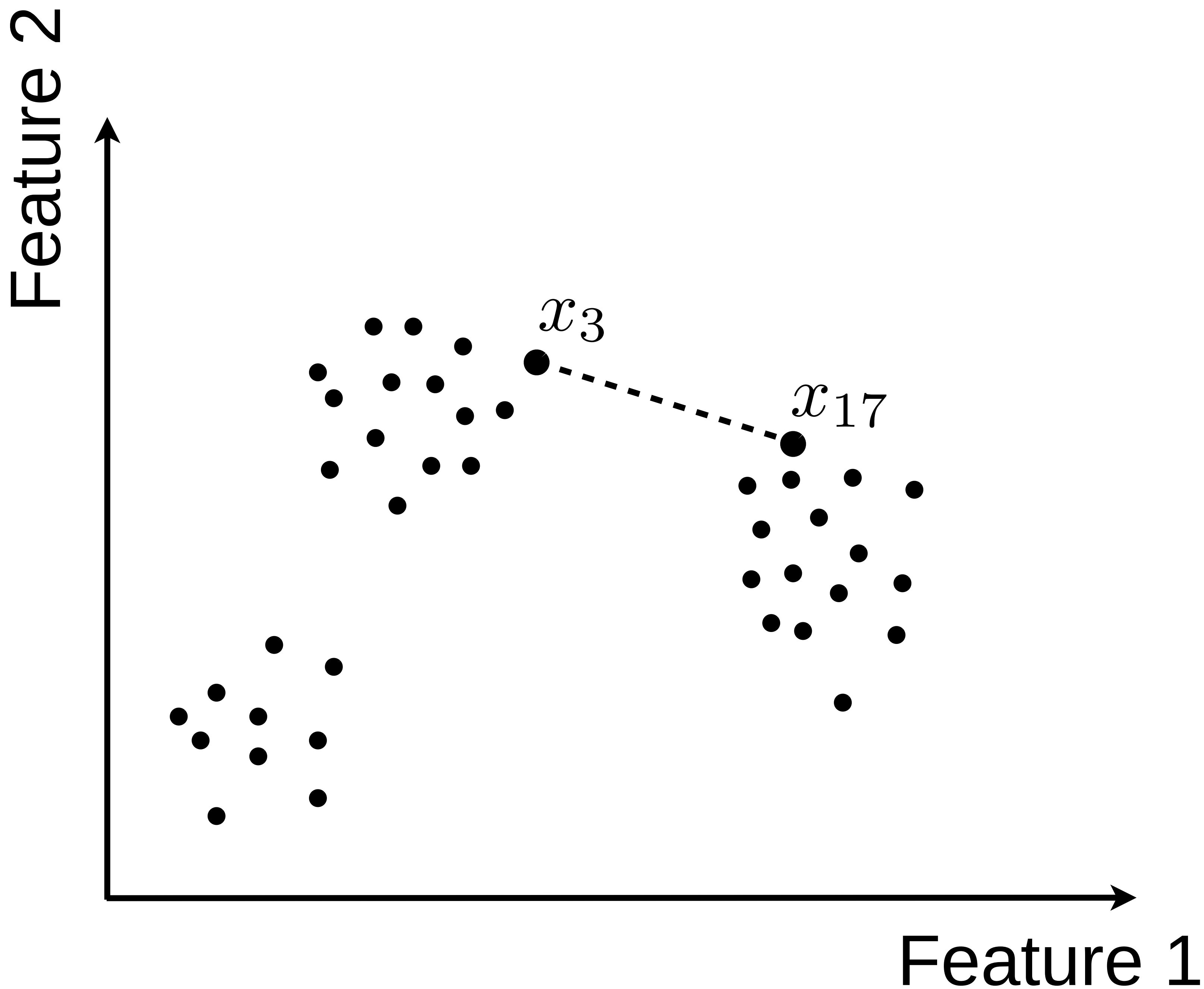
Dissimilarity: Distance as the crow flies



slide by Tamara Broderick

K-Means: Preliminaries

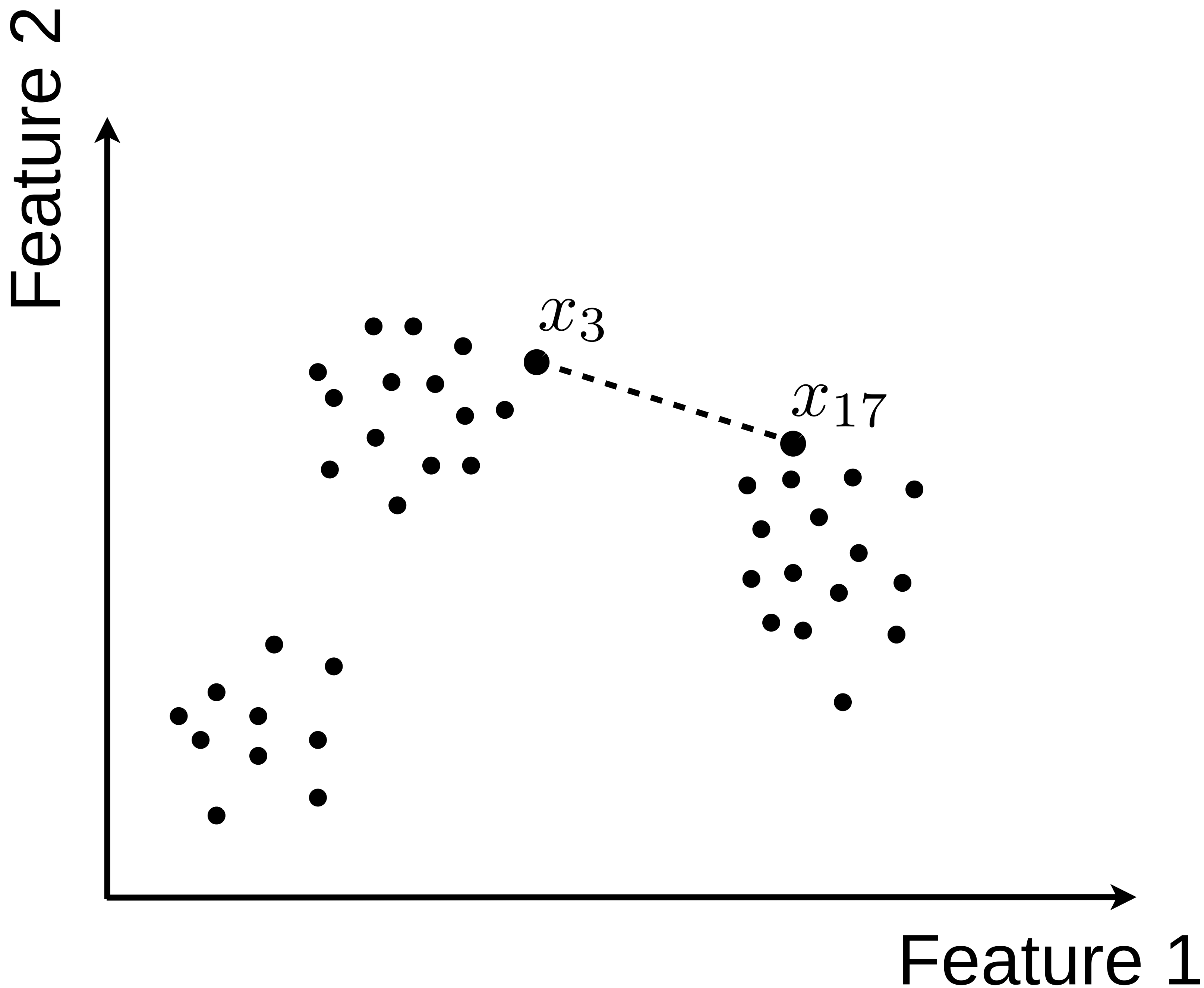
Dissimilarity: Distance as the crow flies



slide by Tamara Broderick

K-Means: Preliminaries

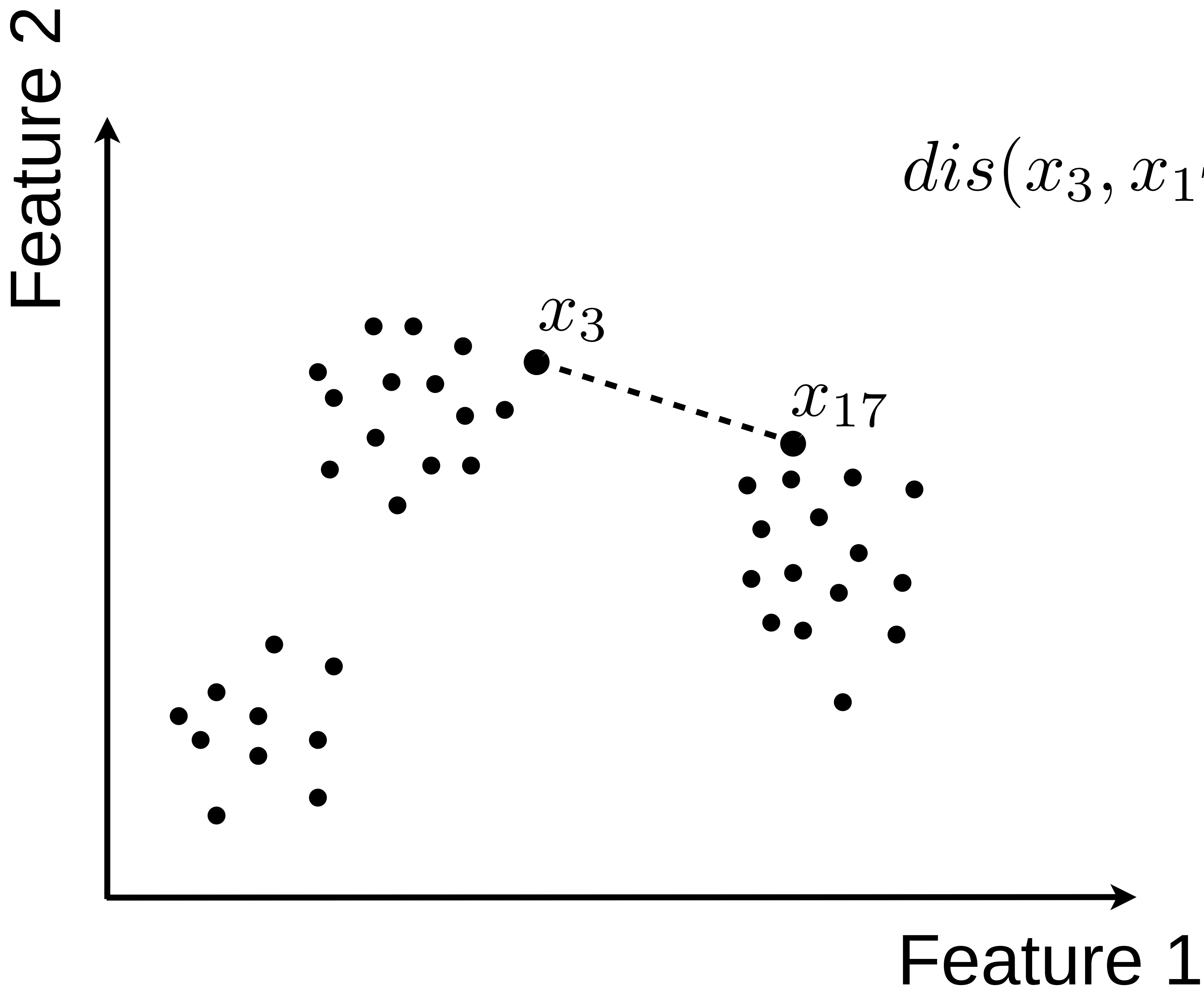
Dissimilarity: Euclidean distance



slide by Tamara Broderick

K-Means: Preliminaries

Dissimilarity: Squared Euclidean distance



$$dis(x_3, x_{17}) = (x_{3,1} - x_{17,1})^2 + (x_{3,2} - x_{17,2})^2$$

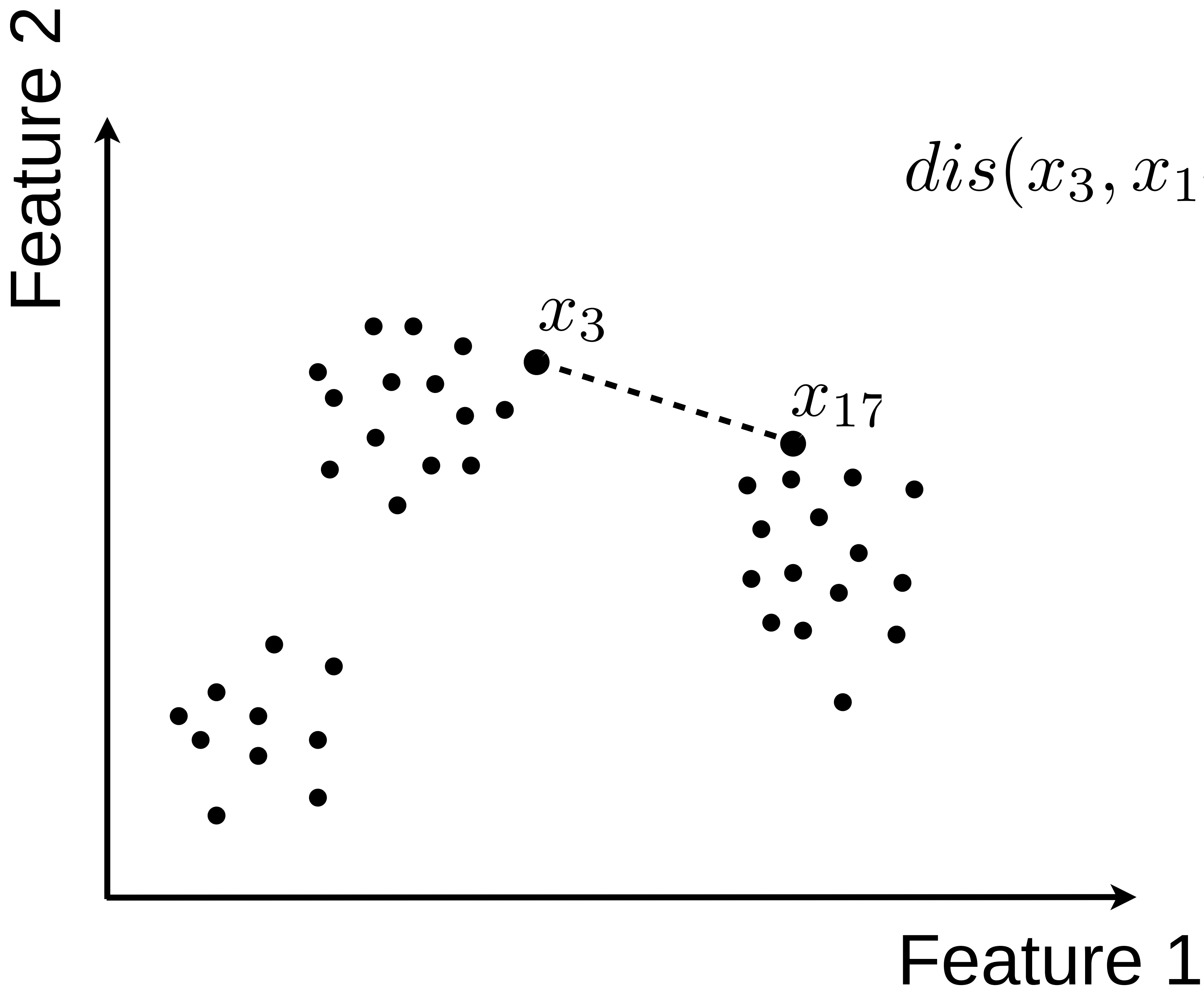
slide by Tamara Broderick

K-Means: Preliminaries

Dissimilarity: Squared Euclidean distance

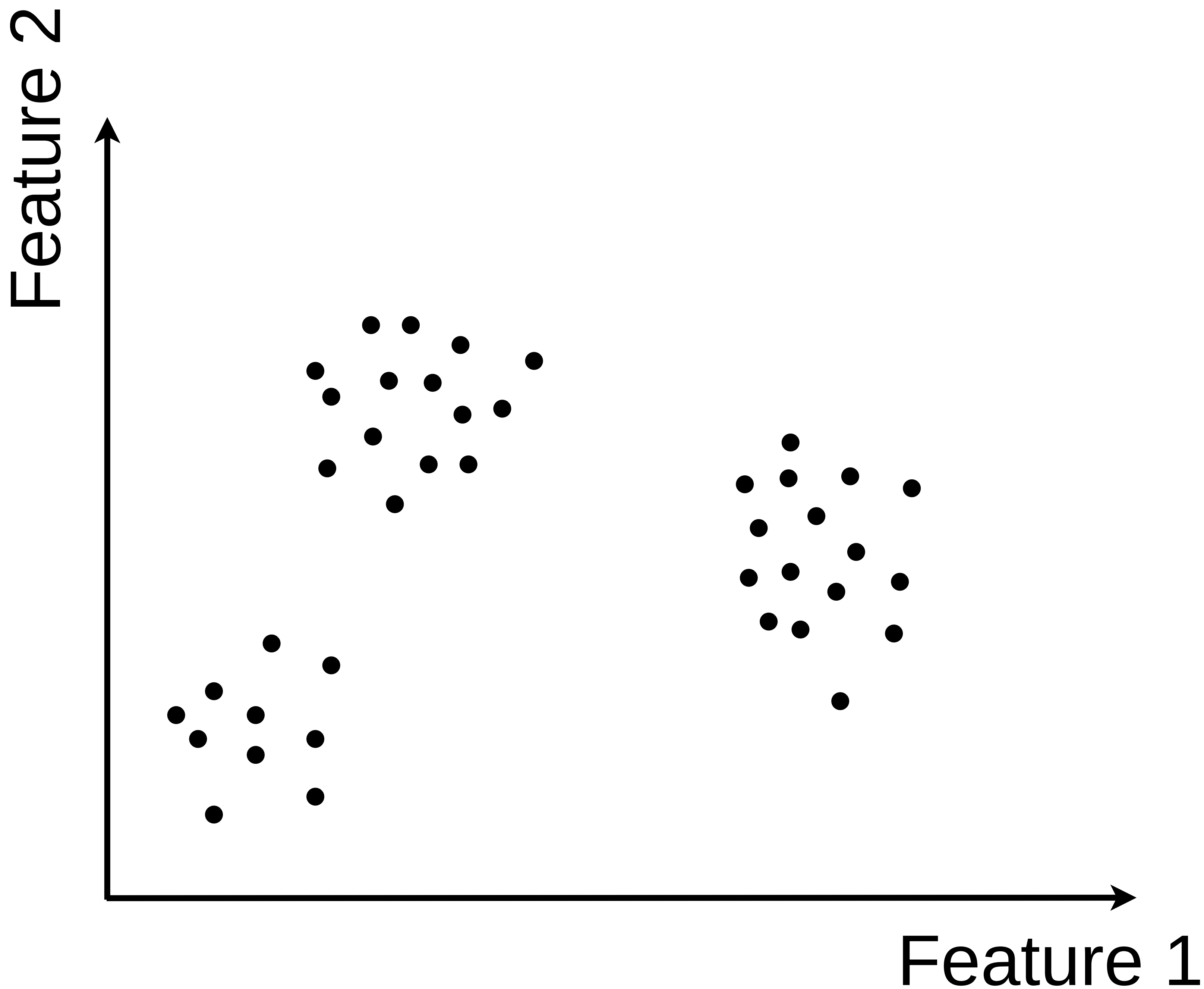
$$dis(x_3, x_{17}) = \sum_{d=1}^{D} (x_{3,d} - x_{17,d})^2$$

For each feature



K-Means: Preliminaries

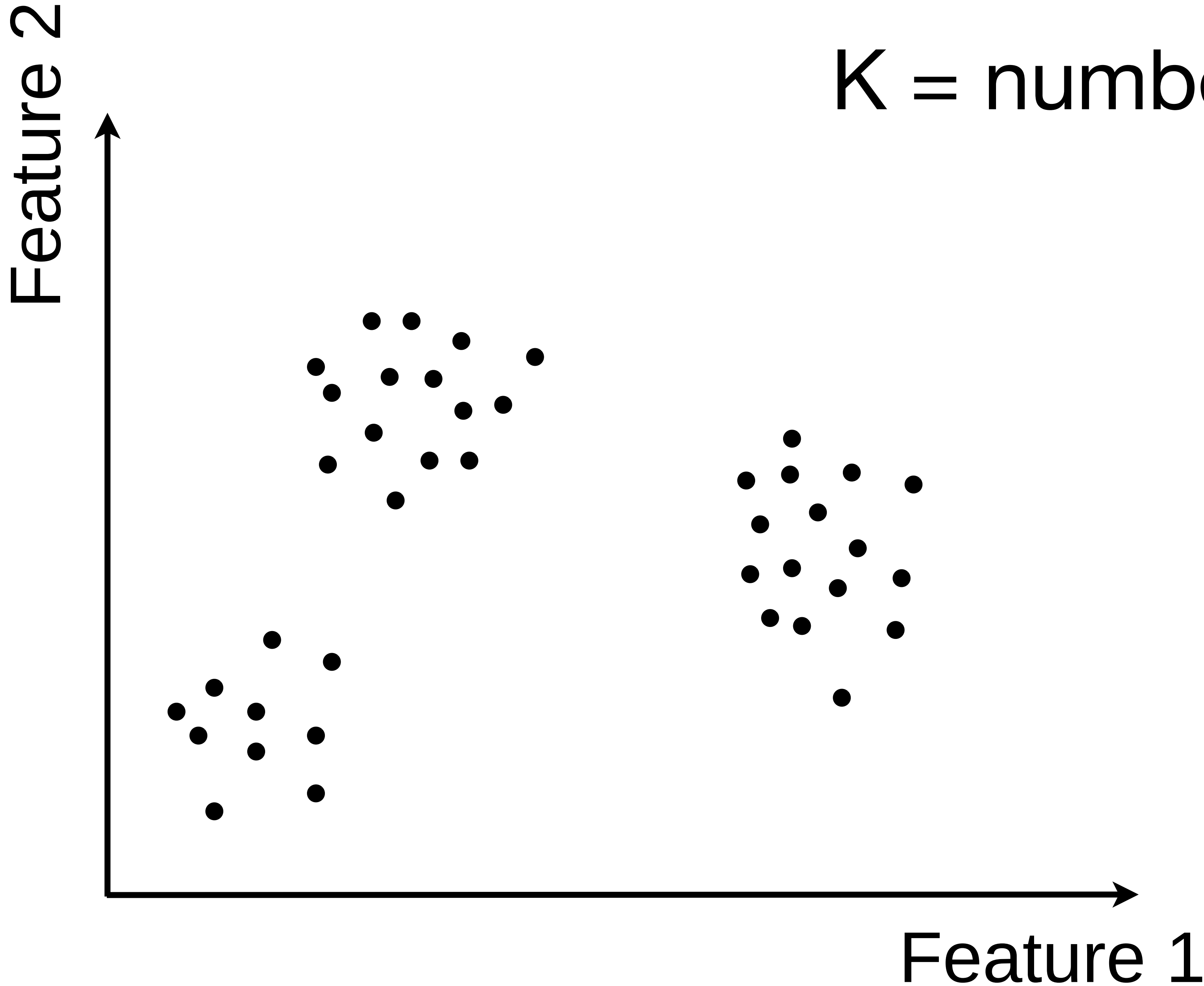
Dissimilarity



K-Means: Preliminaries

Cluster summary

K = number of clusters

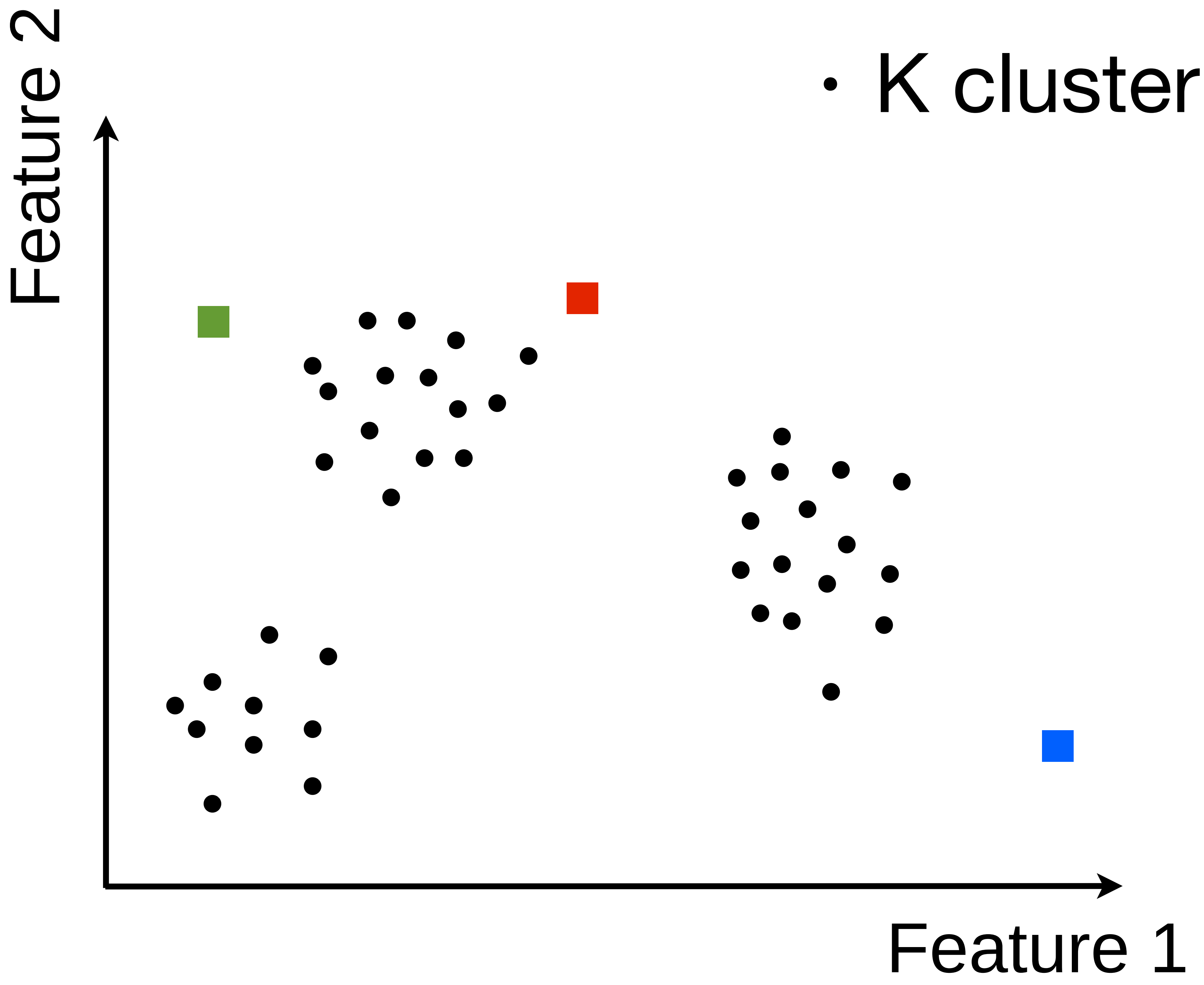


slide by Tamara Broderick

K-Means: Preliminaries

Cluster summary

- K cluster centers

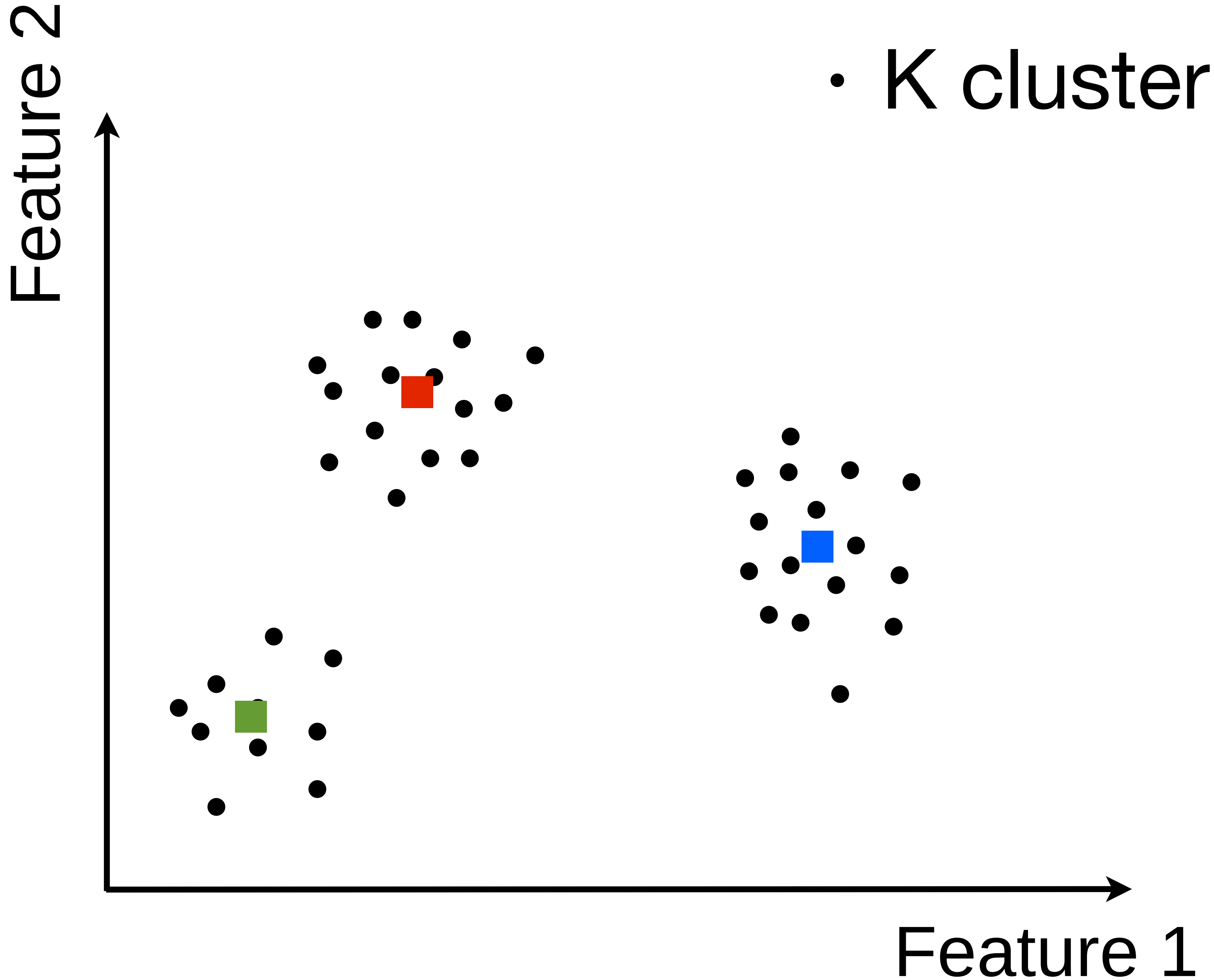


slide by Tamara Broderick

K-Means: Preliminaries

Cluster summary

- K cluster centers

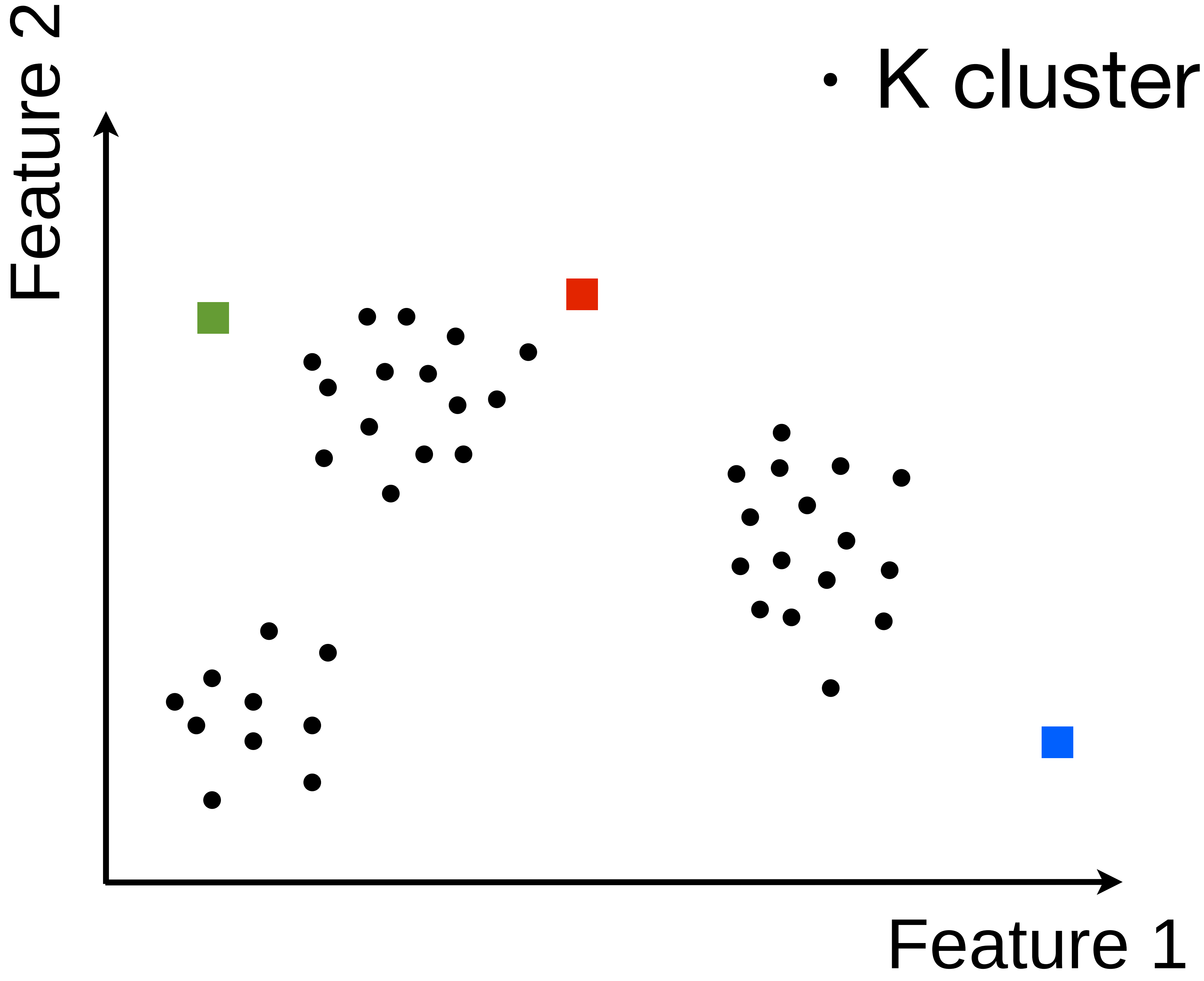


slide by Tamara Broderick

K-Means: Preliminaries

Cluster summary

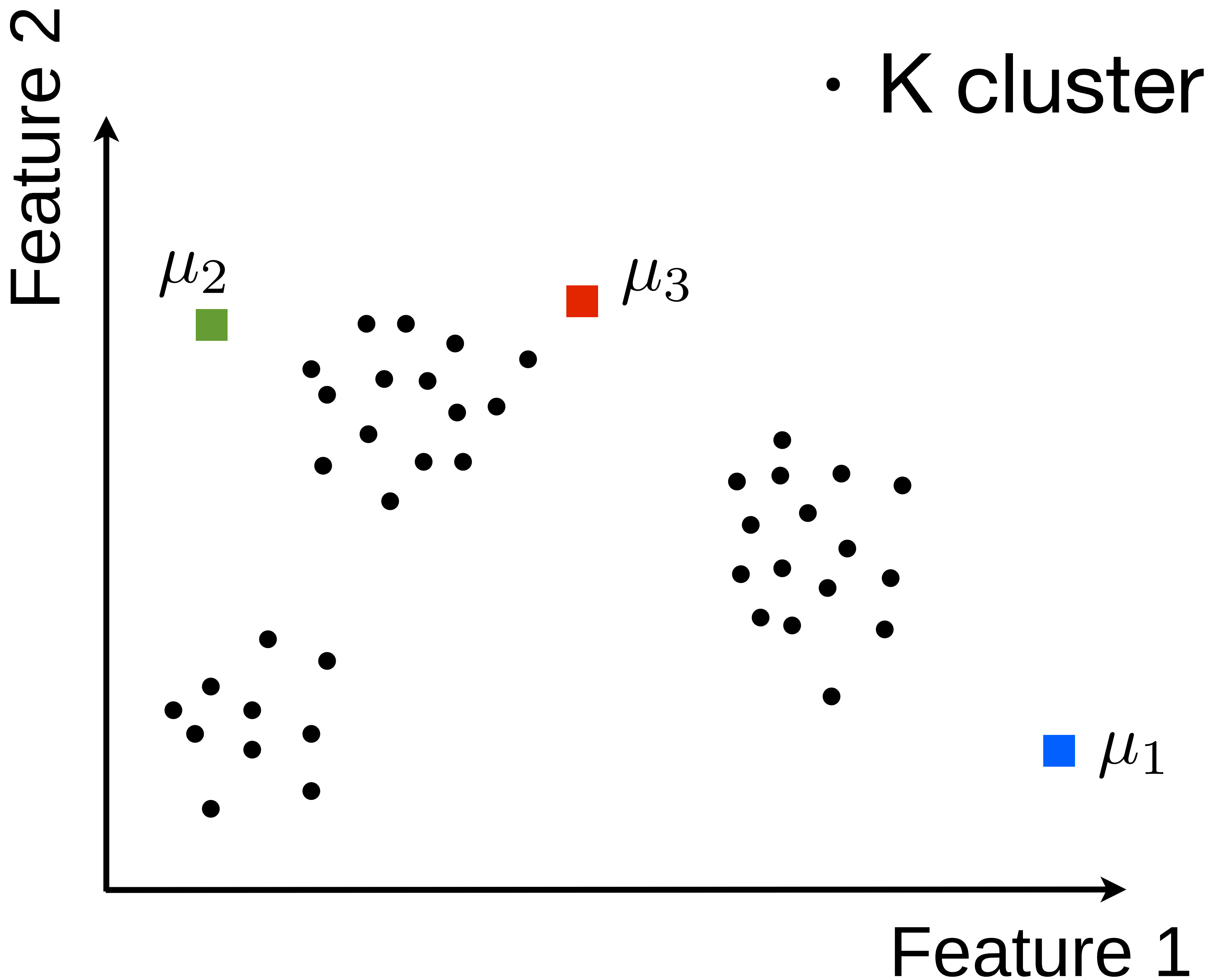
- K cluster centers



K-Means: Preliminaries

Cluster summary

- K cluster centers

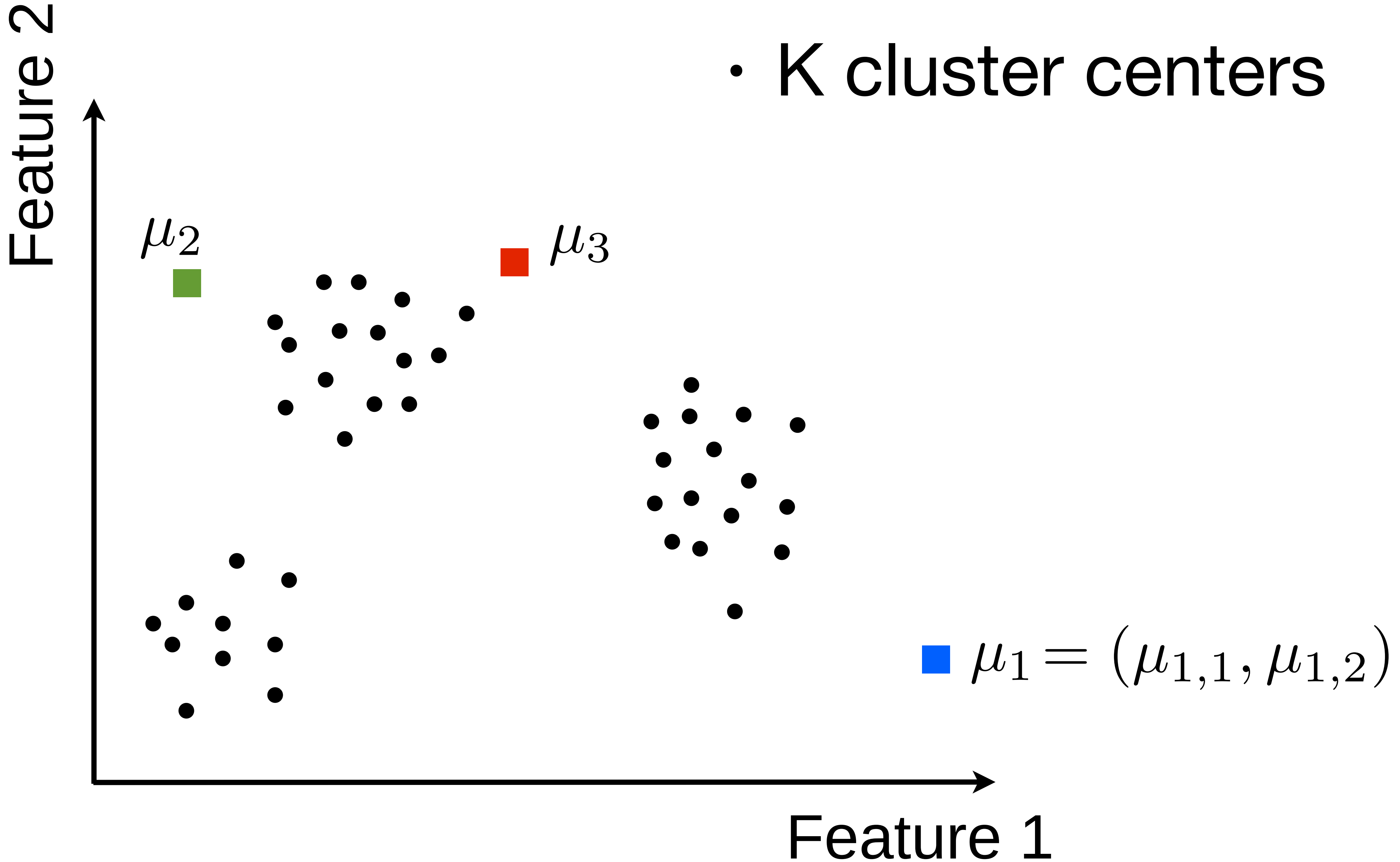


slide by Tamara Broderick

K-Means: Preliminaries

Cluster summary

- K cluster centers



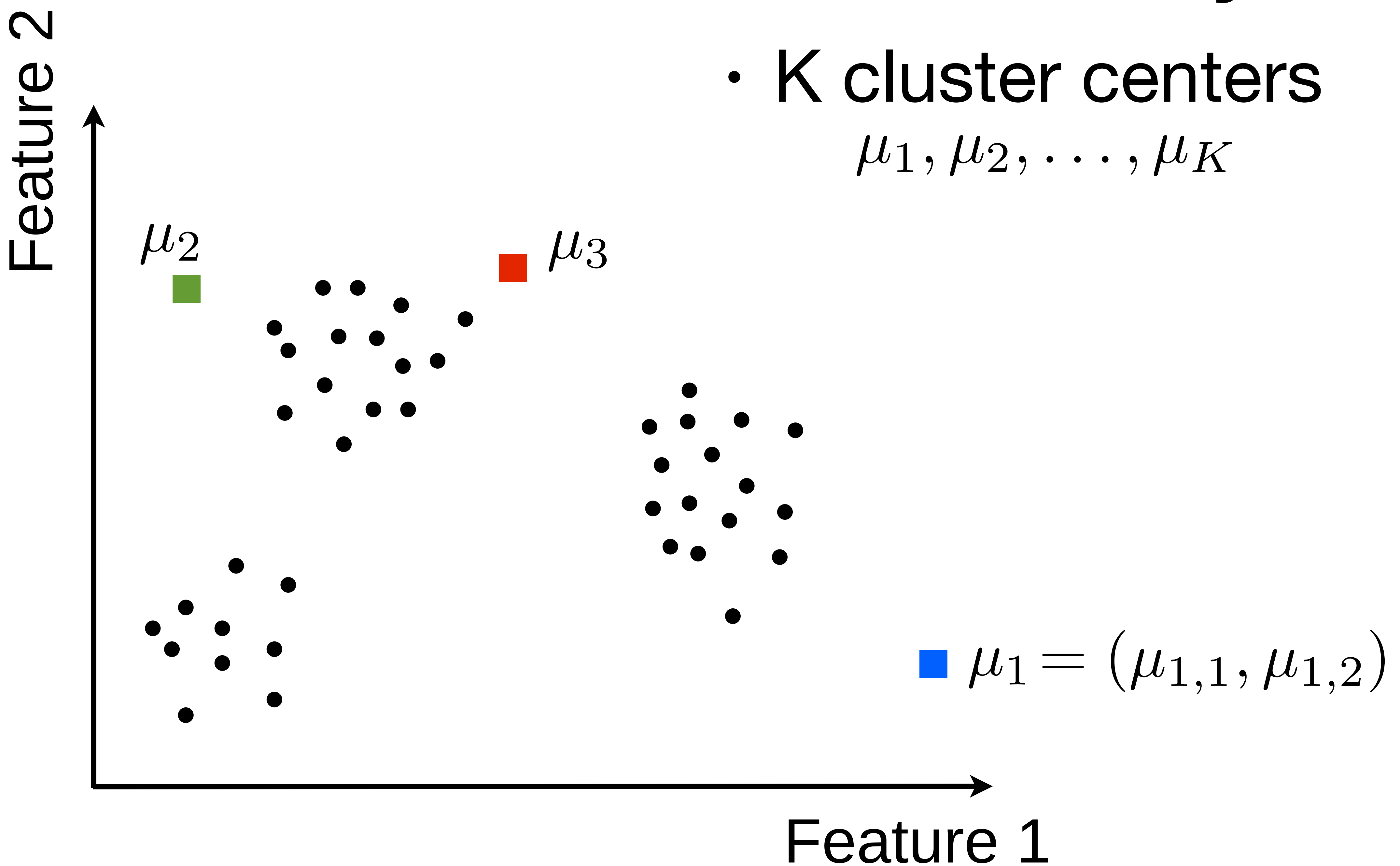
slide by Tamara Broderick

K-Means: Preliminaries

Cluster summary

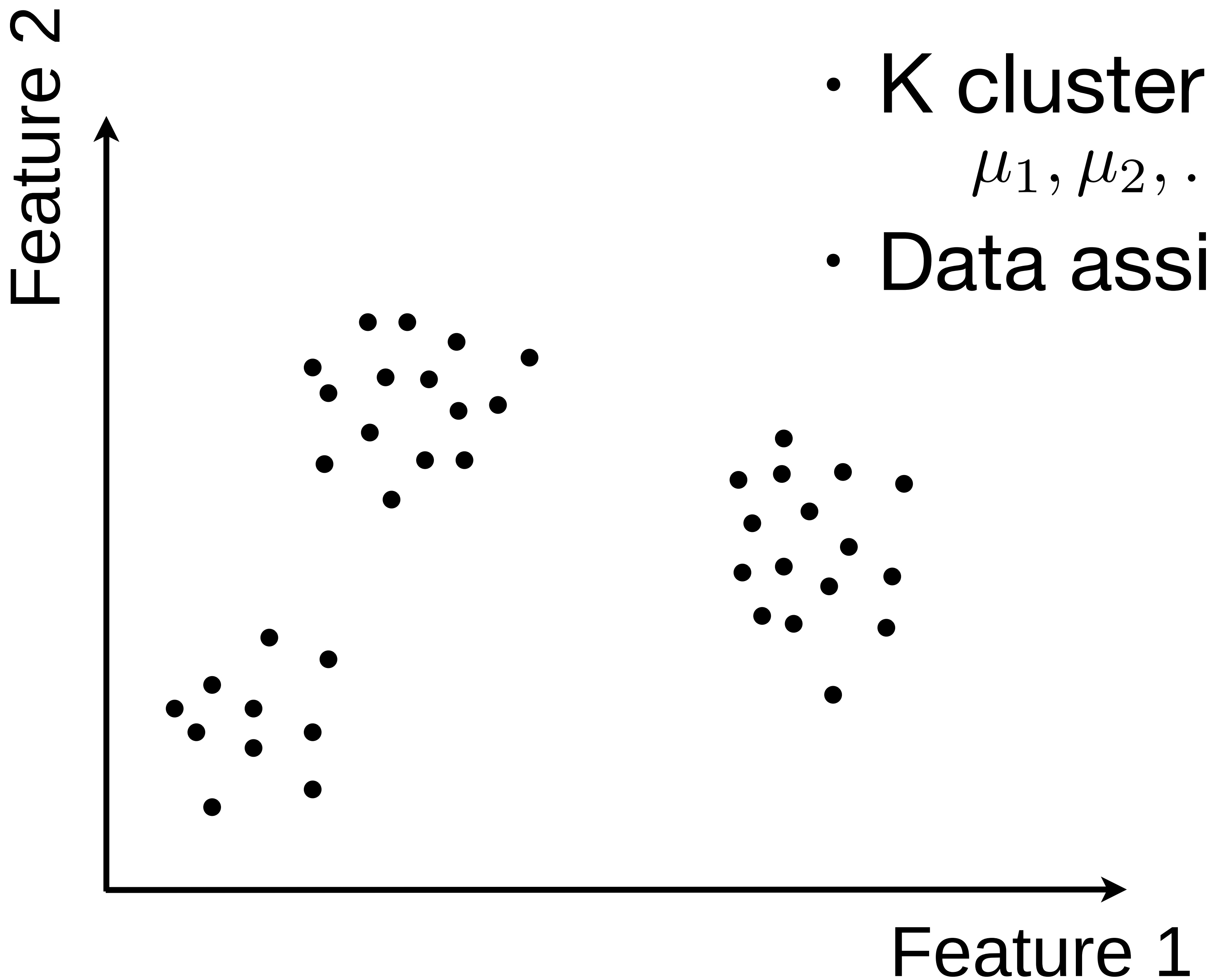
- K cluster centers

$$\mu_1, \mu_2, \ldots, \mu_K$$



slide by Tamara Broderick

K-Means: Preliminaries



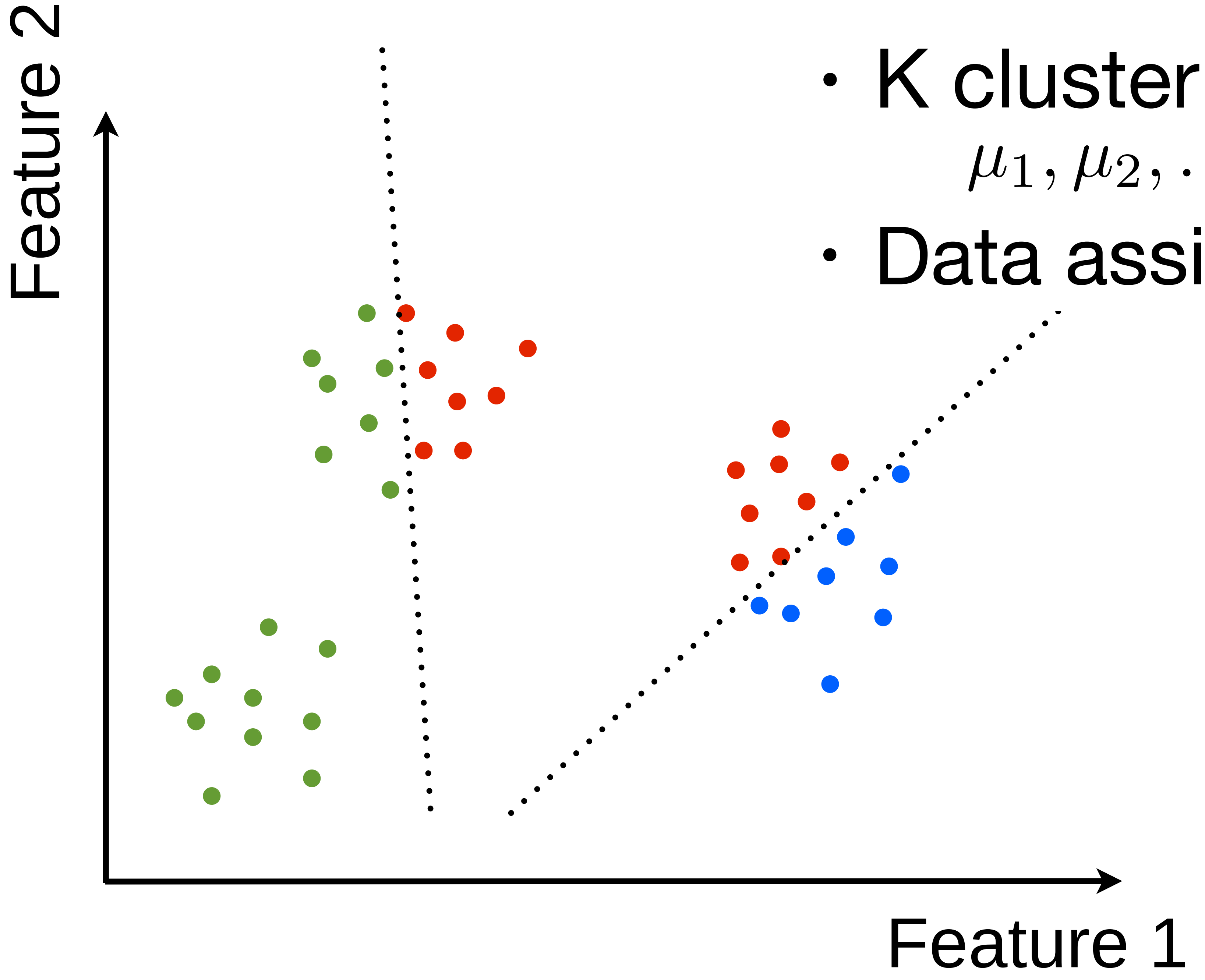
Cluster summary

- K cluster centers

  $\mu_1, \mu_2, \ldots, \mu_K$

- Data assignments to clusters

K-Means: Preliminaries

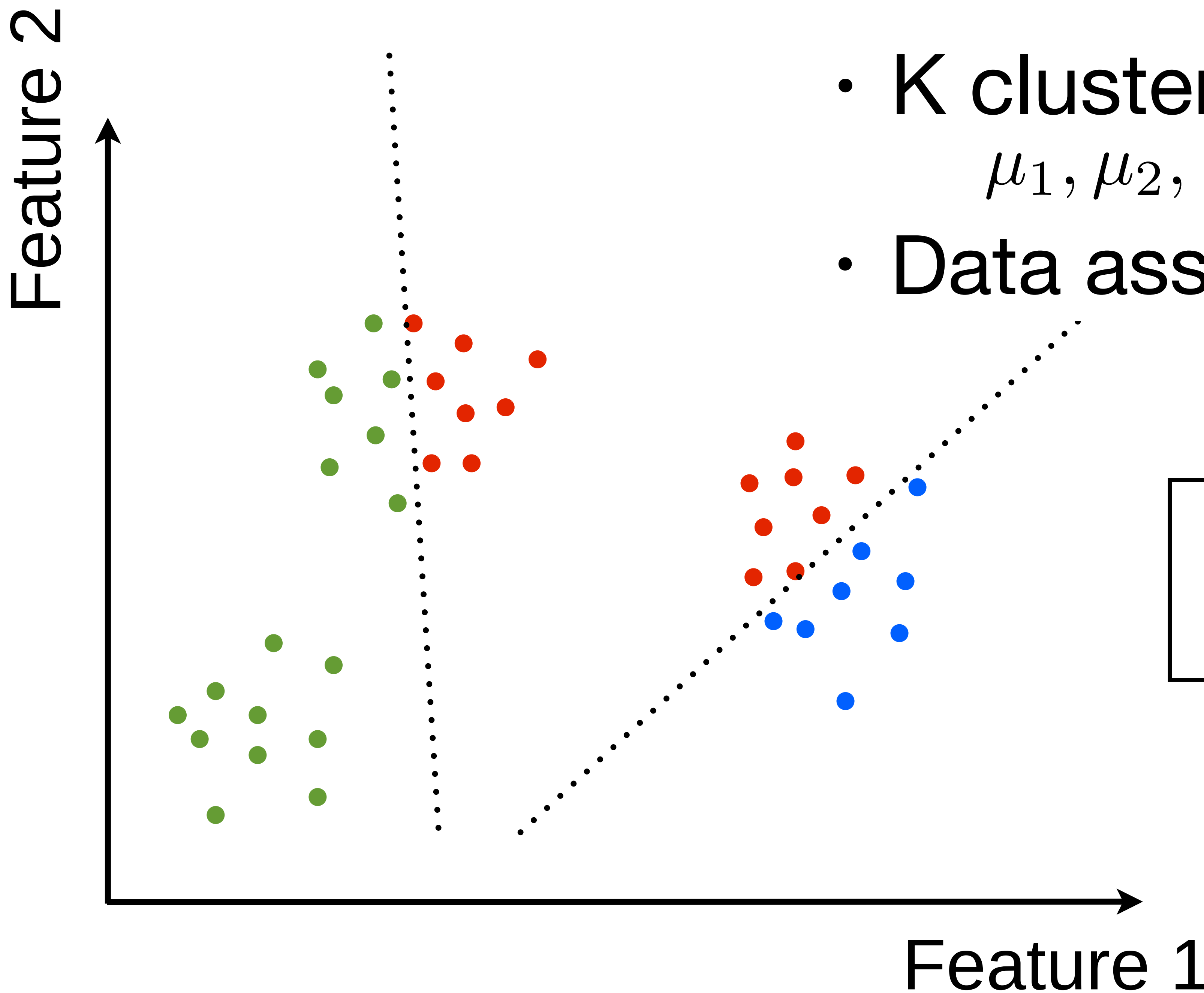


Cluster summary

- K cluster centers
  $\mu_1, \mu_2, \ldots, \mu_K$
- Data assignments to clusters

slide by Tamara Broderick

K-Means: Preliminaries



Cluster summary

- K cluster centers

$$\mu_1, \mu_2, \ldots, \mu_K$$

- Data assignments to clusters

S_k = set of points in cluster k

slide by Tamara Broderick

K-Means: Preliminaries

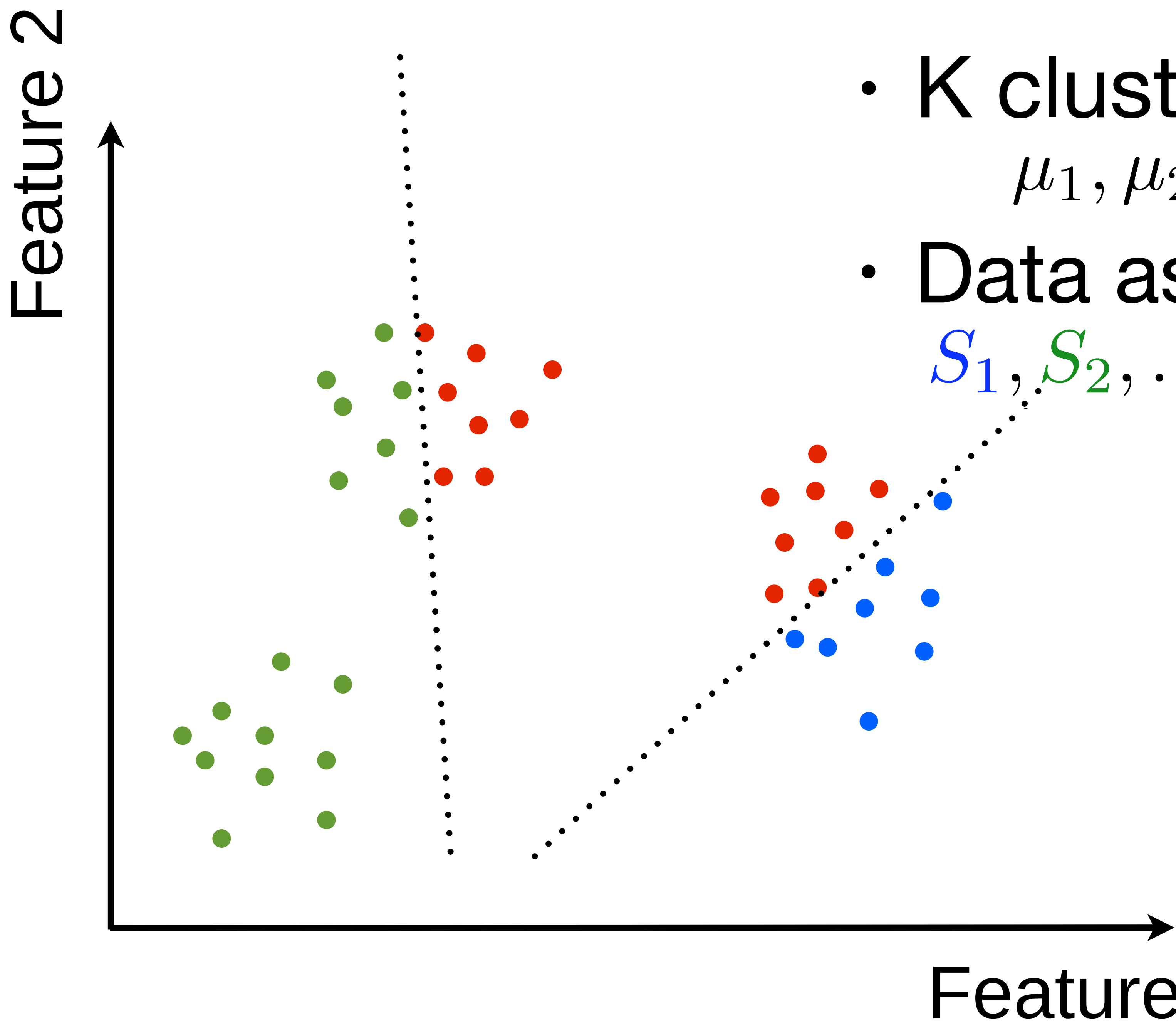
Cluster summary

- K cluster centers

  $\mu_1, \mu_2, \ldots, \mu_K$

- Data assignments to clusters

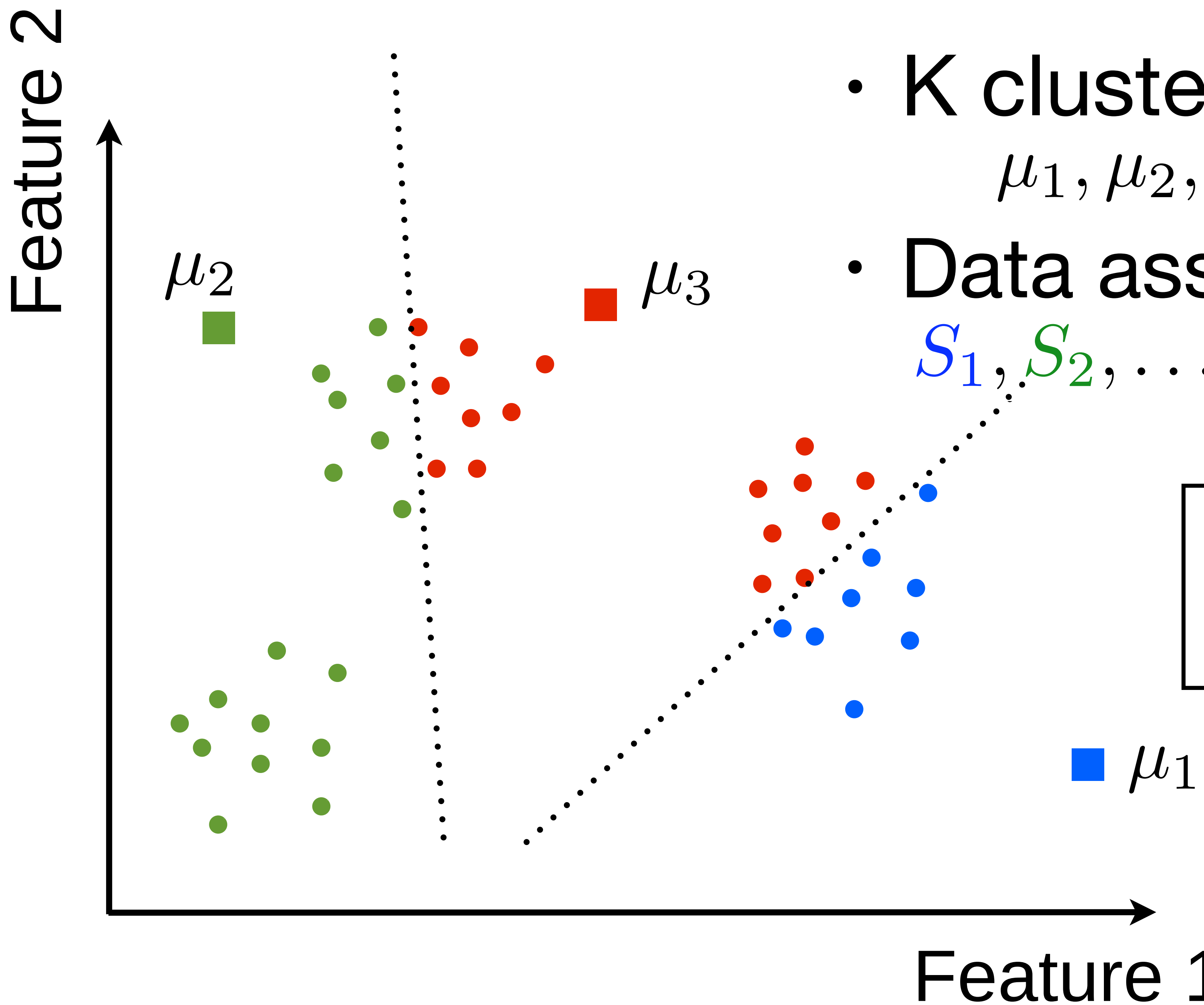
  $S_1, S_2, \ldots, S_K$



S_k = set of points in cluster k

slide by Tamara Broderick

K-Means: Preliminaries



Cluster summary

- K cluster centers

  $\mu_1, \mu_2, \ldots, \mu_K$
- Data assignments to clusters

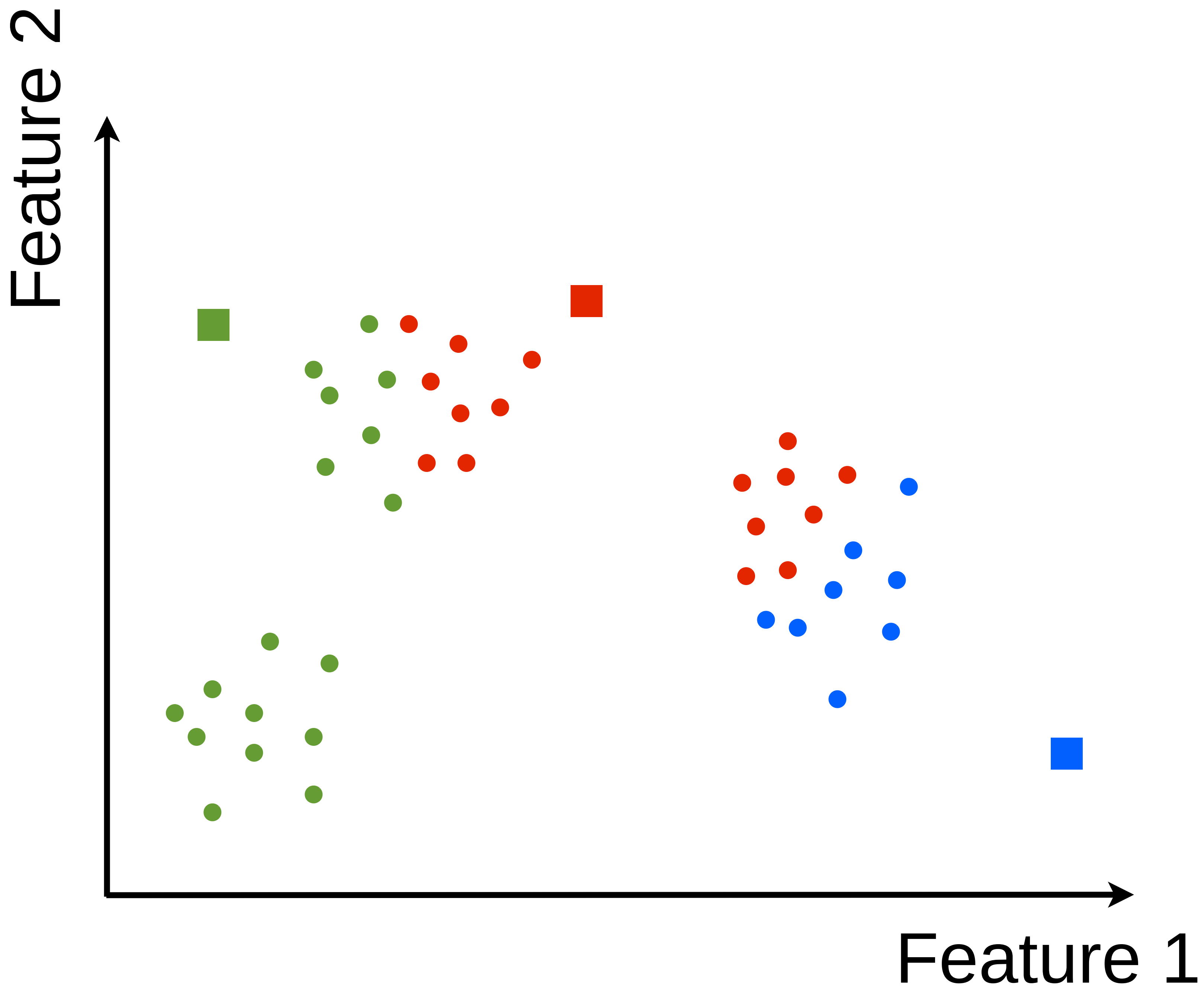
  $S_1, S_2, \ldots, S_K$

S_k = set of points in cluster k

slide by Tamara Broderick

K-Means: Preliminaries

Dissimilarity

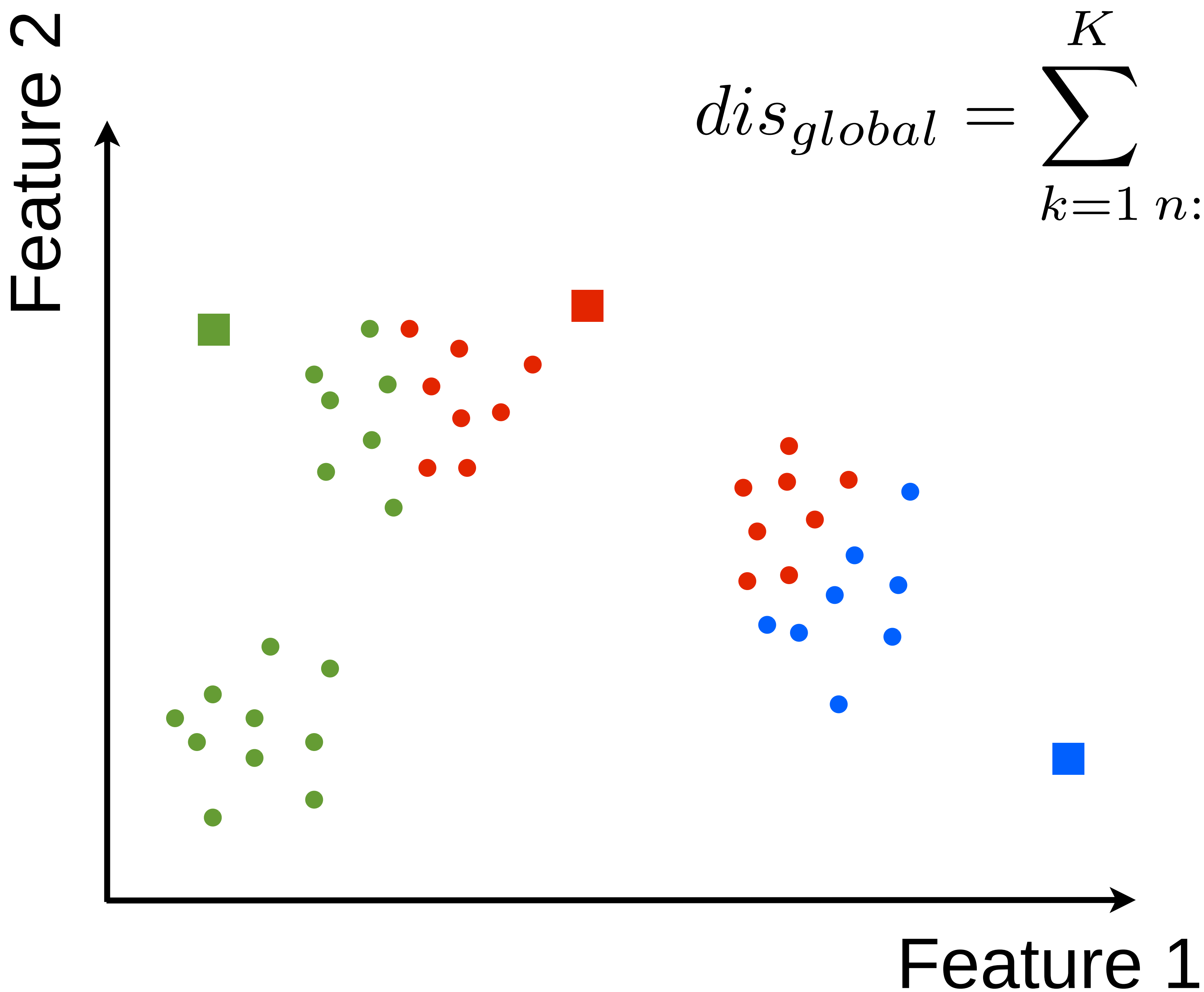


slide by Tamara Broderick

K-Means: Preliminaries

Dissimilarity (global)

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



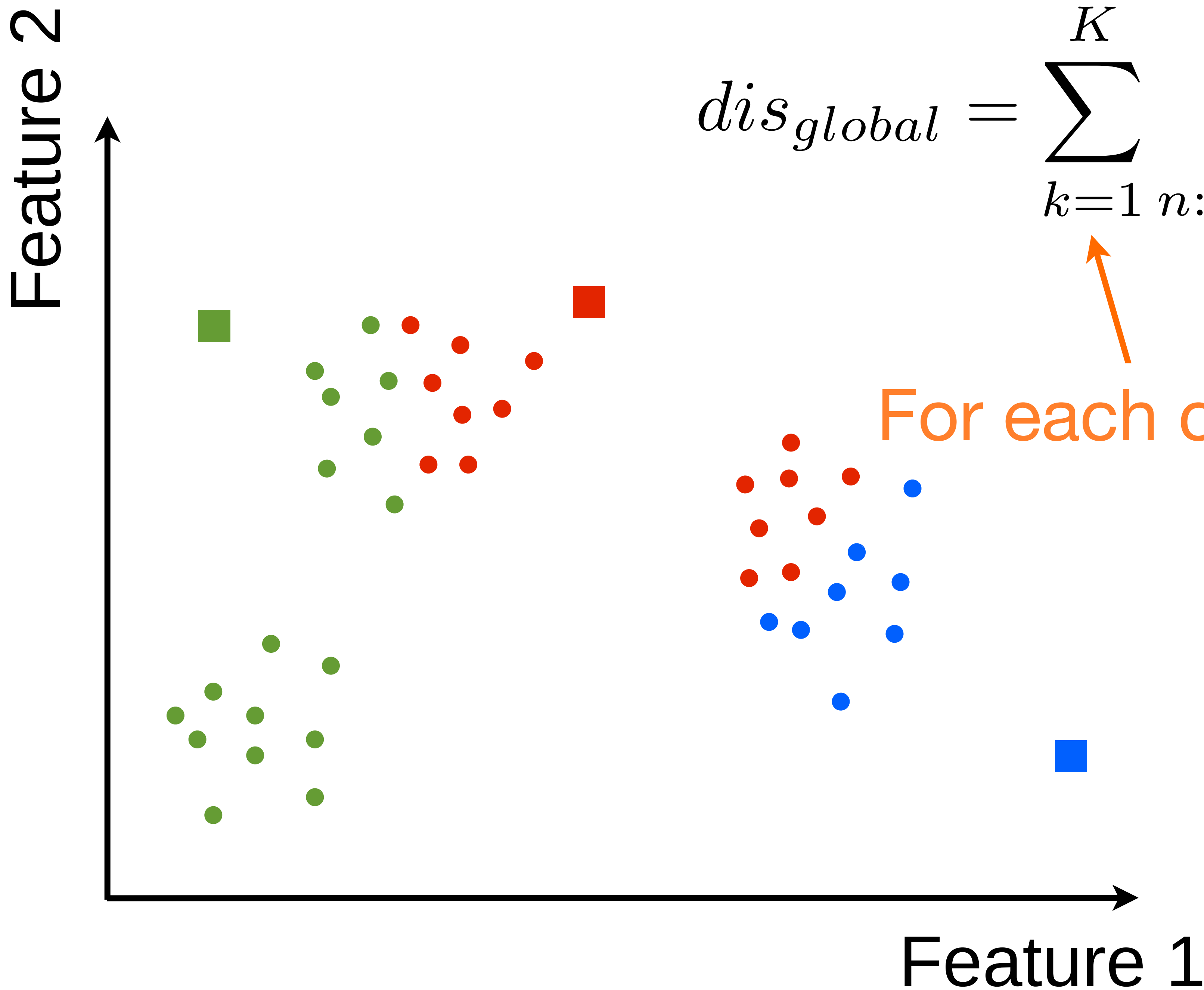
slide by Tamara Broderick

K-Means: Preliminaries

Dissimilarity (global)

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

For each cluster

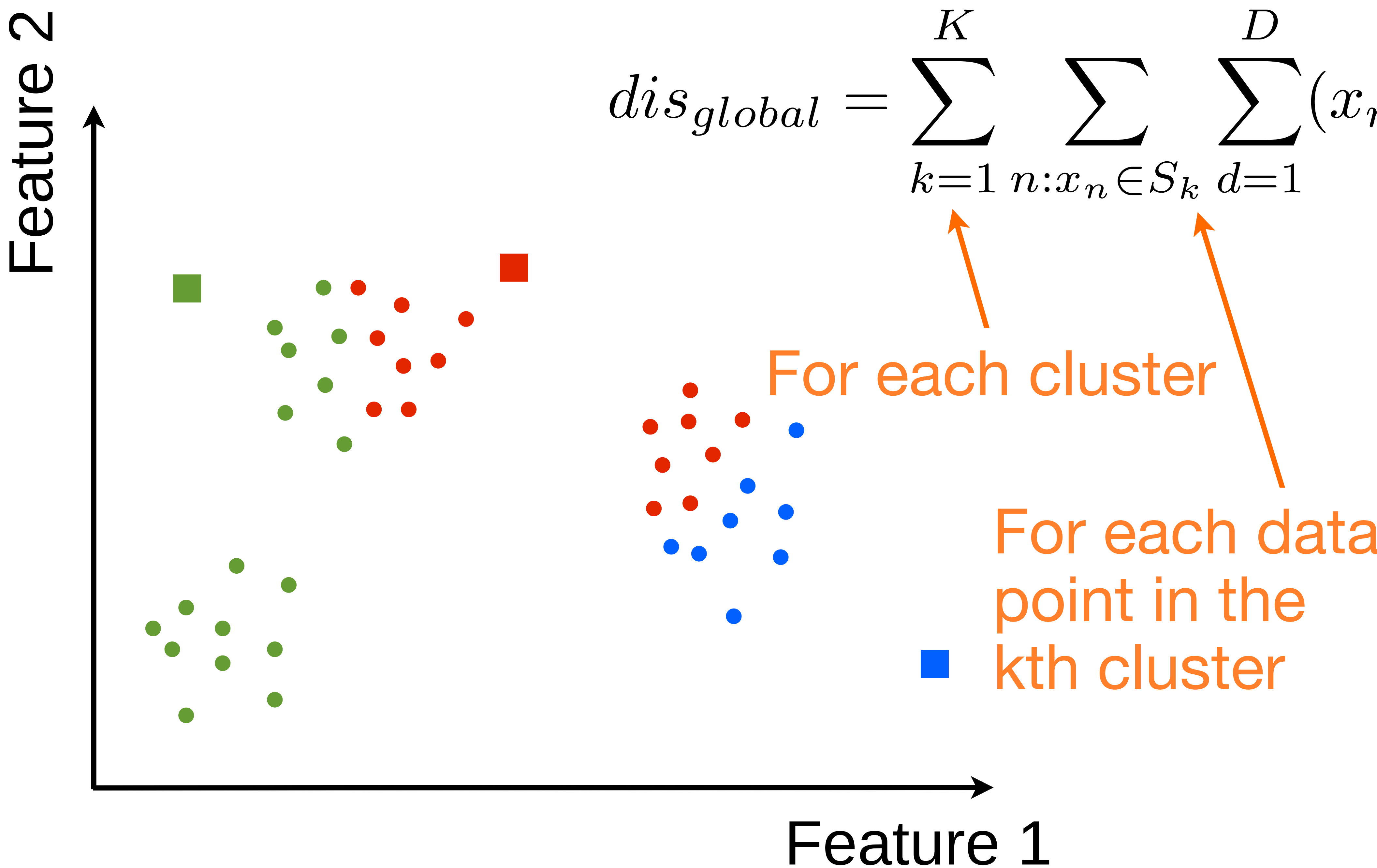


slide by Tamara Broderick

K-Means: Preliminaries

Dissimilarity (global)

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



slide by Tamara Broderick

K-Means: Preliminaries

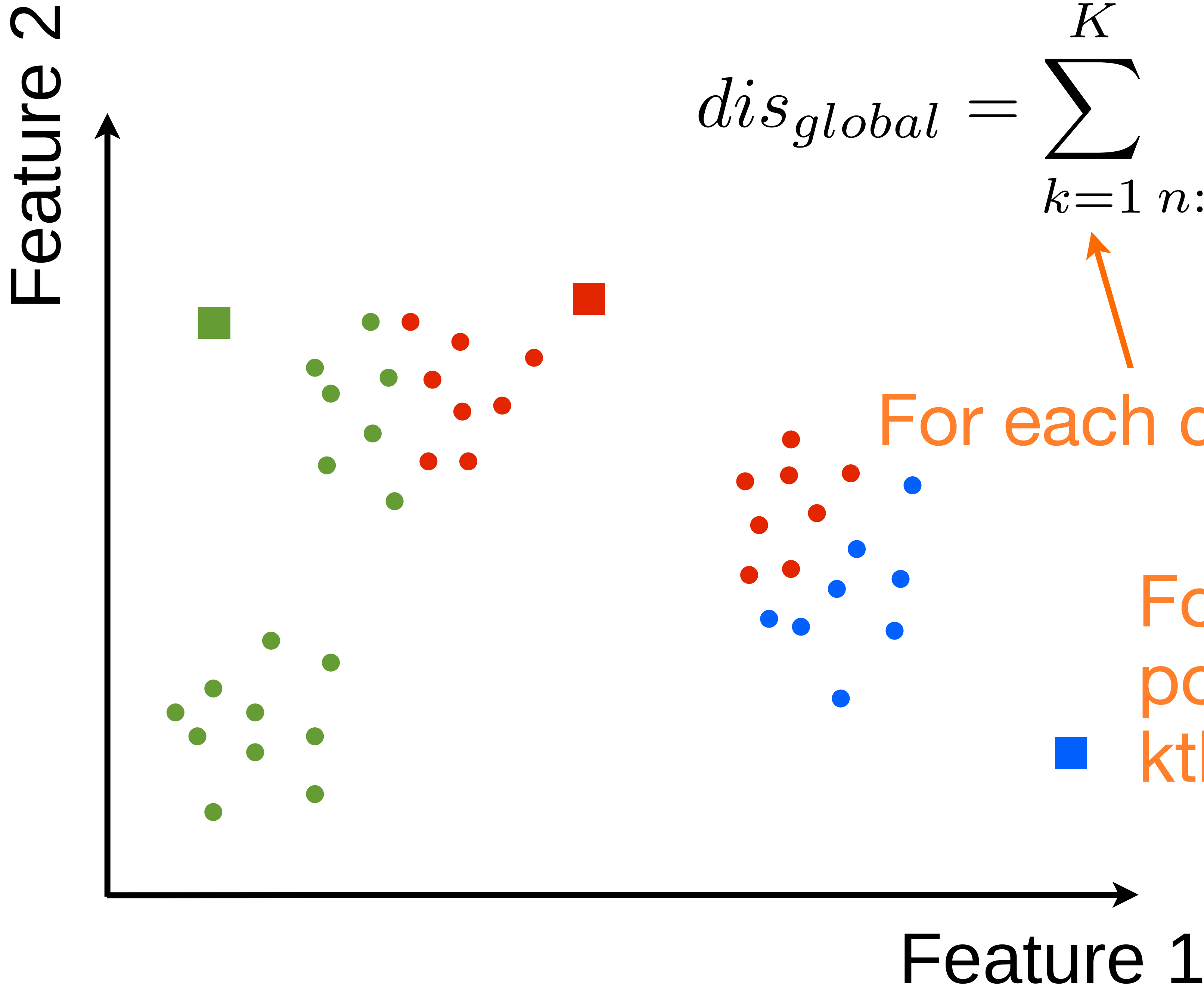
Dissimilarity (global)

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

For each cluster

For each data point in the kth cluster

For each feature

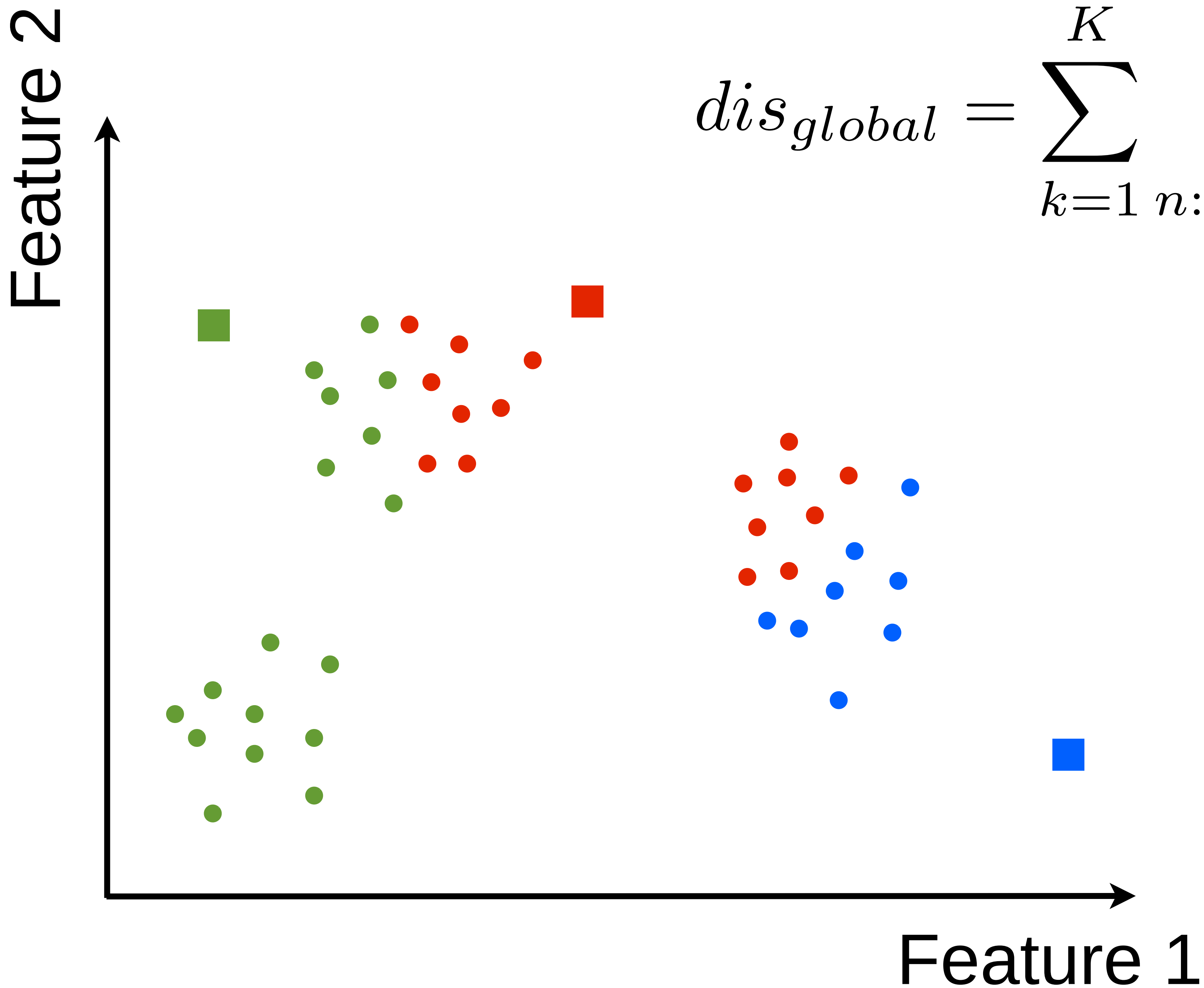


slide by Tamara Broderick

K-Means: Preliminaries

Dissimilarity (global)

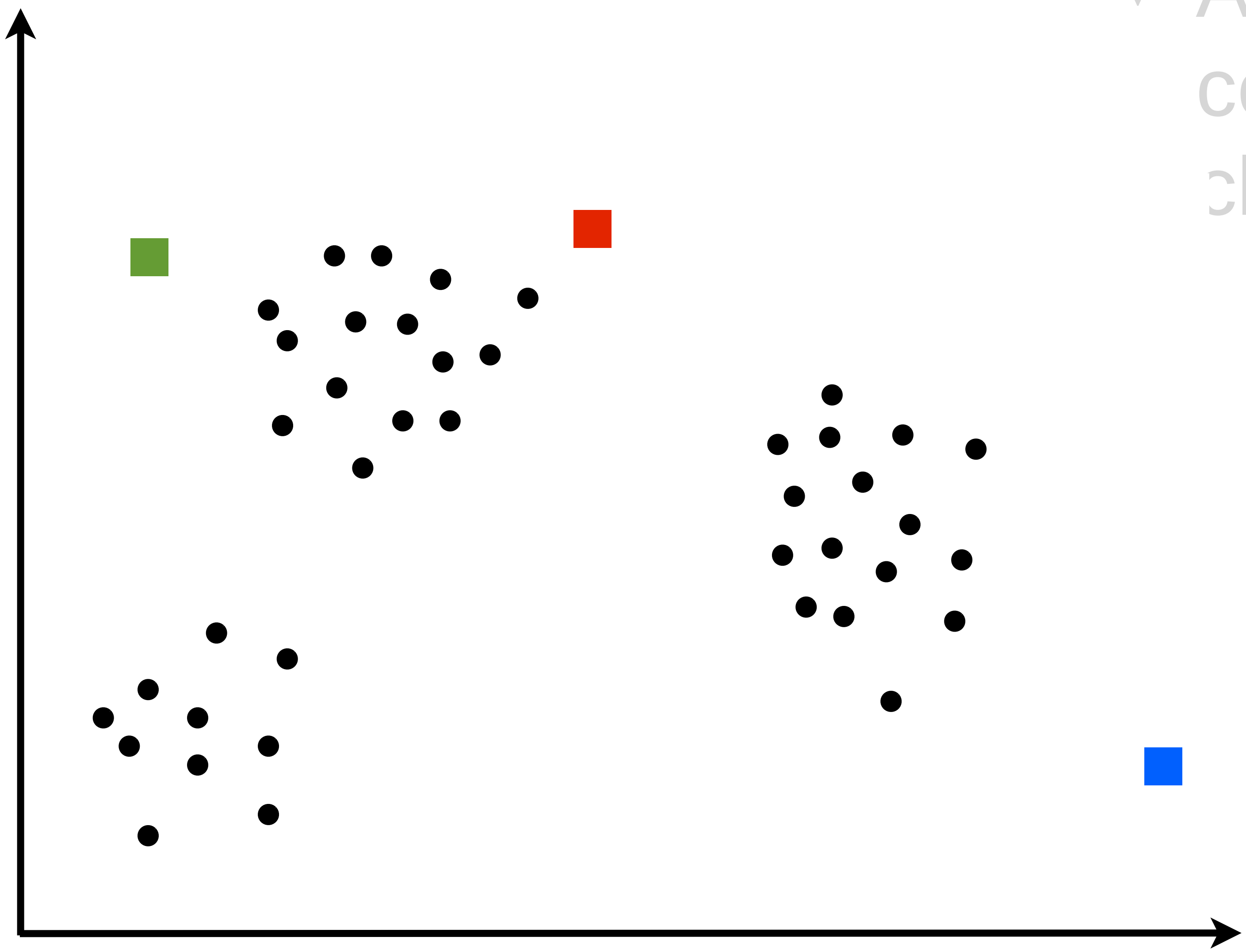
$$dis_{global} = \sum_{k=1}^{K} \sum_{n: x_n \in S_k} \sum_{d=1}^{D} (x_{n,d} - \mu_{k,d})^2$$



slide by Tamara Broderick

K-Means Algorithm

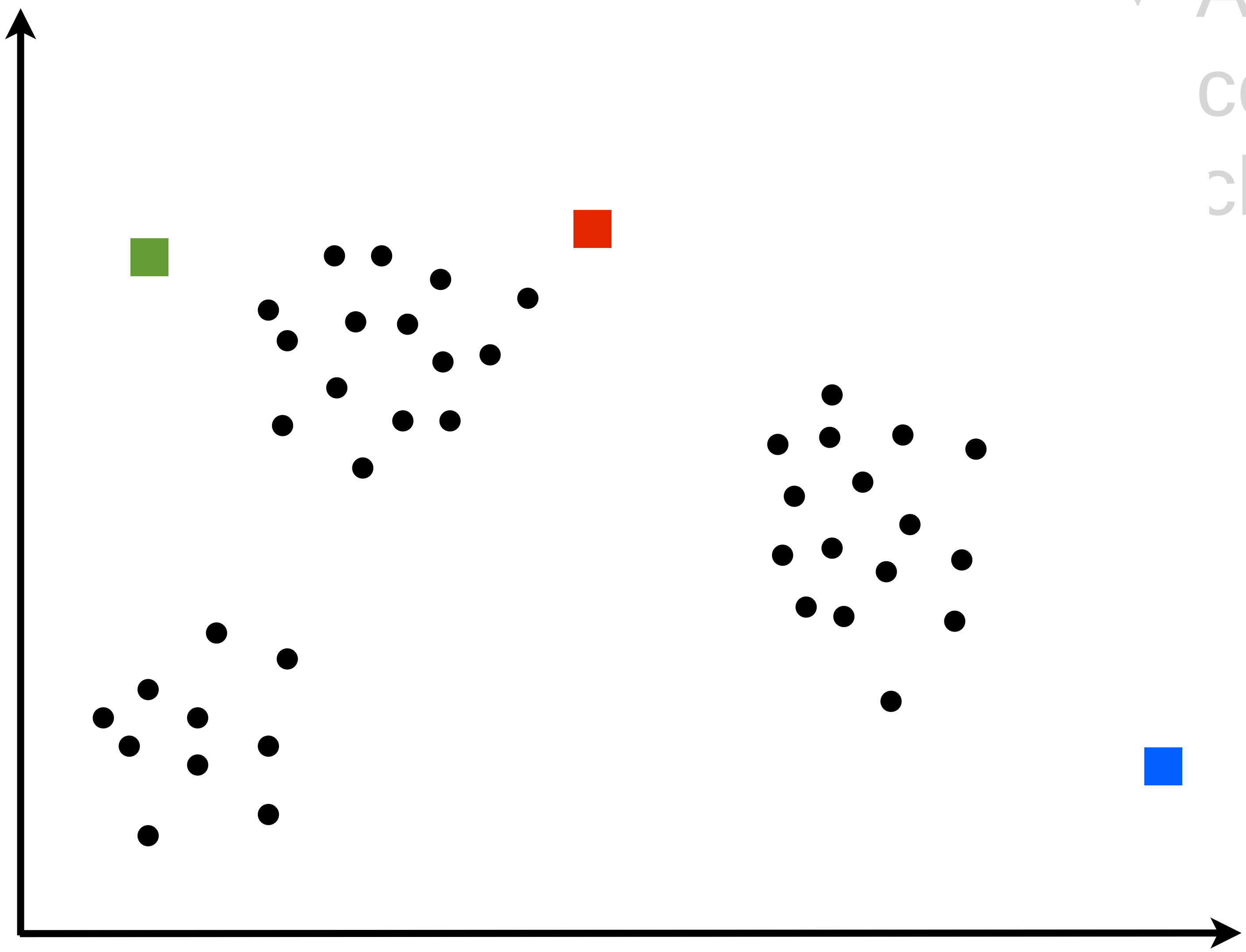
- Initialize K cluster centers
- Repeat until convergence:
  - ✦ Assign each data point to the cluster with the closest center.
  - ✦ Assign each cluster center to be the mean of its cluster's data points



slide by Tamara Broderick

K-Means Algorithm

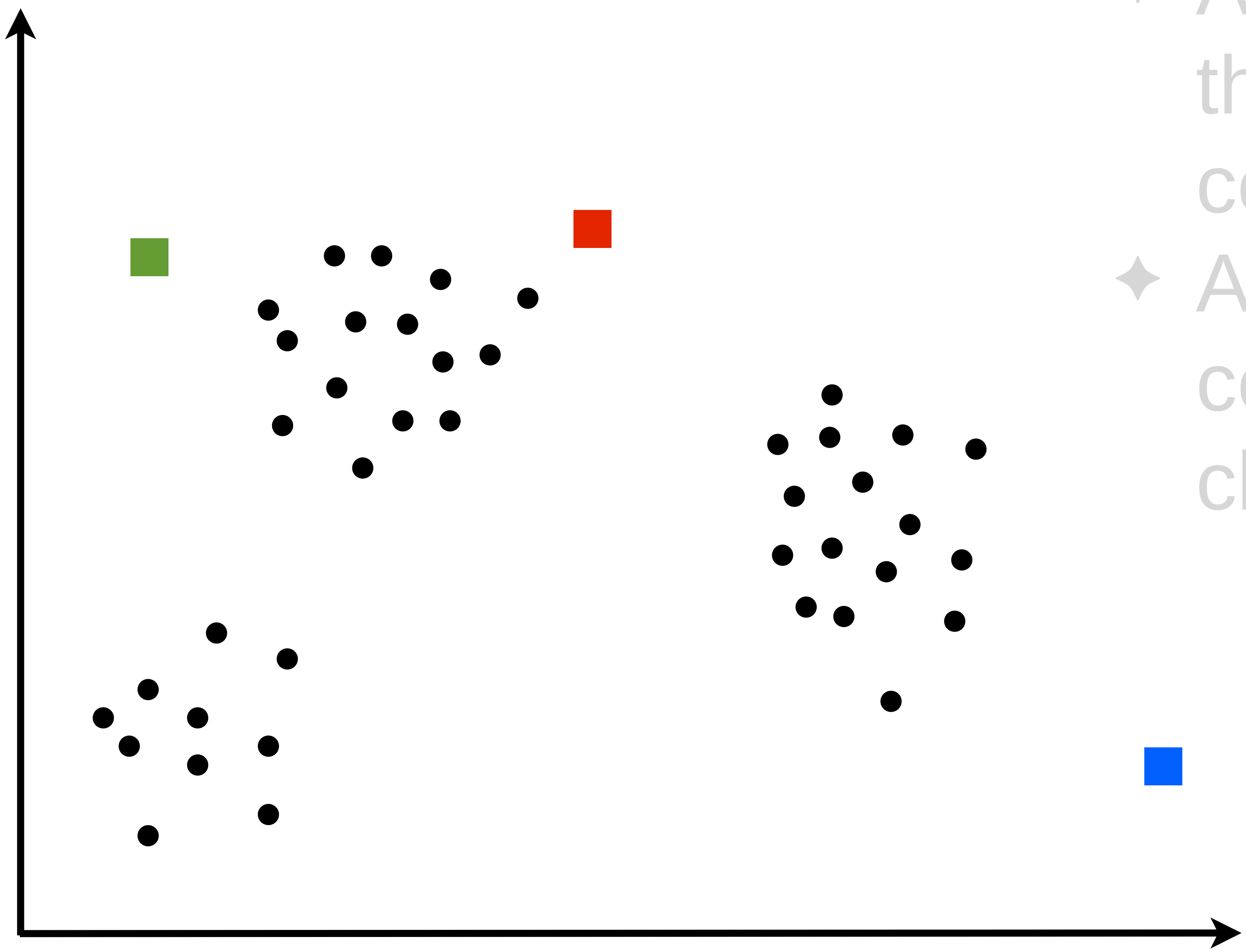
- Initialize K cluster centers
- Repeat until convergence:
  - ✦ Assign each data point to the cluster with the closest center.
  - ✦ Assign each cluster center to be the mean of its cluster's data points



slide by Tamara Broderick

K-Means Algorithm

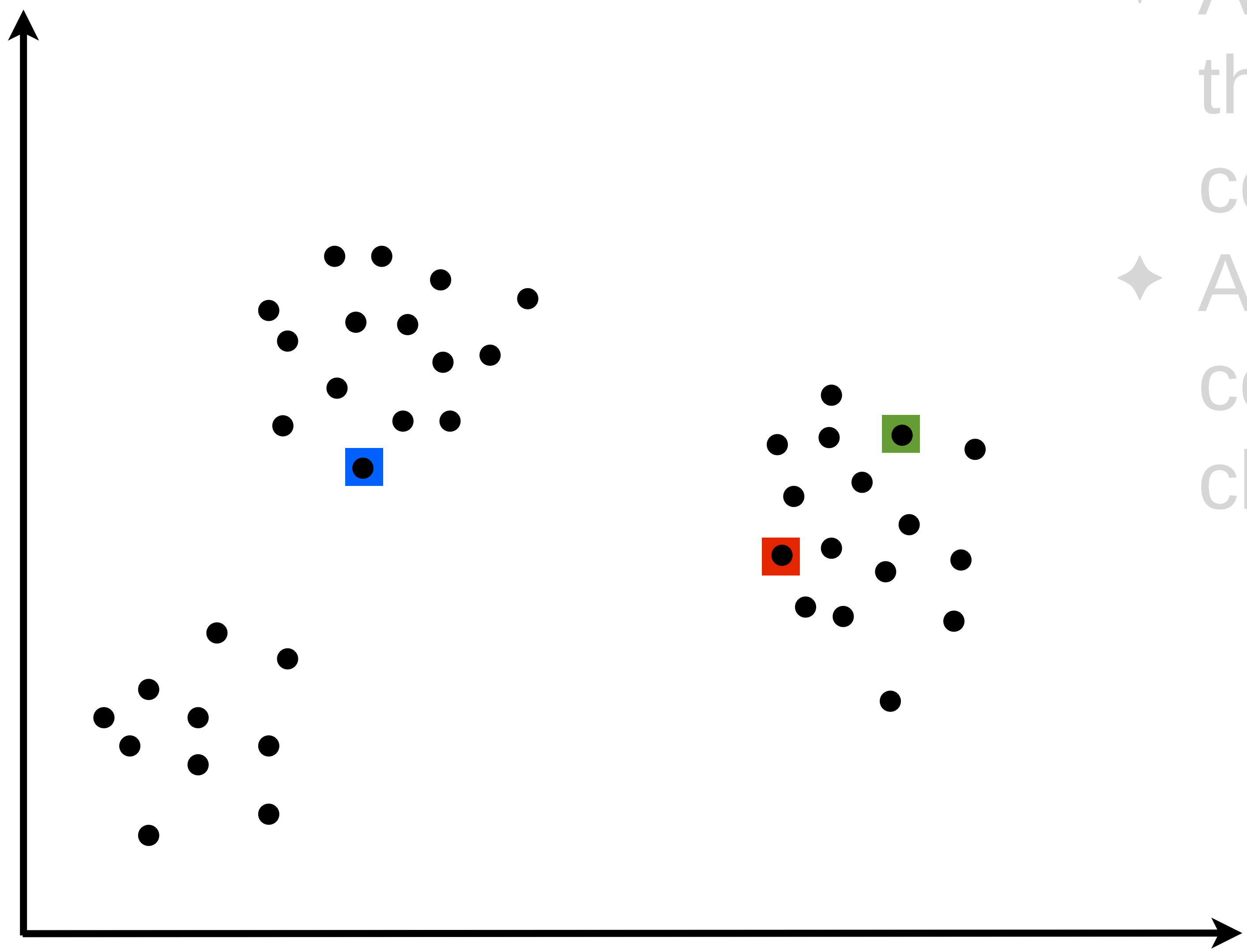
- For k = 1,…, K
  - ✦ Randomly draw n from 1,…,N without replacement
  - ✦ $\mu_k \leftarrow x_n$
- Repeat until convergence:
  - ✦ Assign each data point to the cluster with the closest center.
  - ✦ Assign each cluster center to be the mean of its cluster's data points



slide by Tamara Broderick

K-Means Algorithm

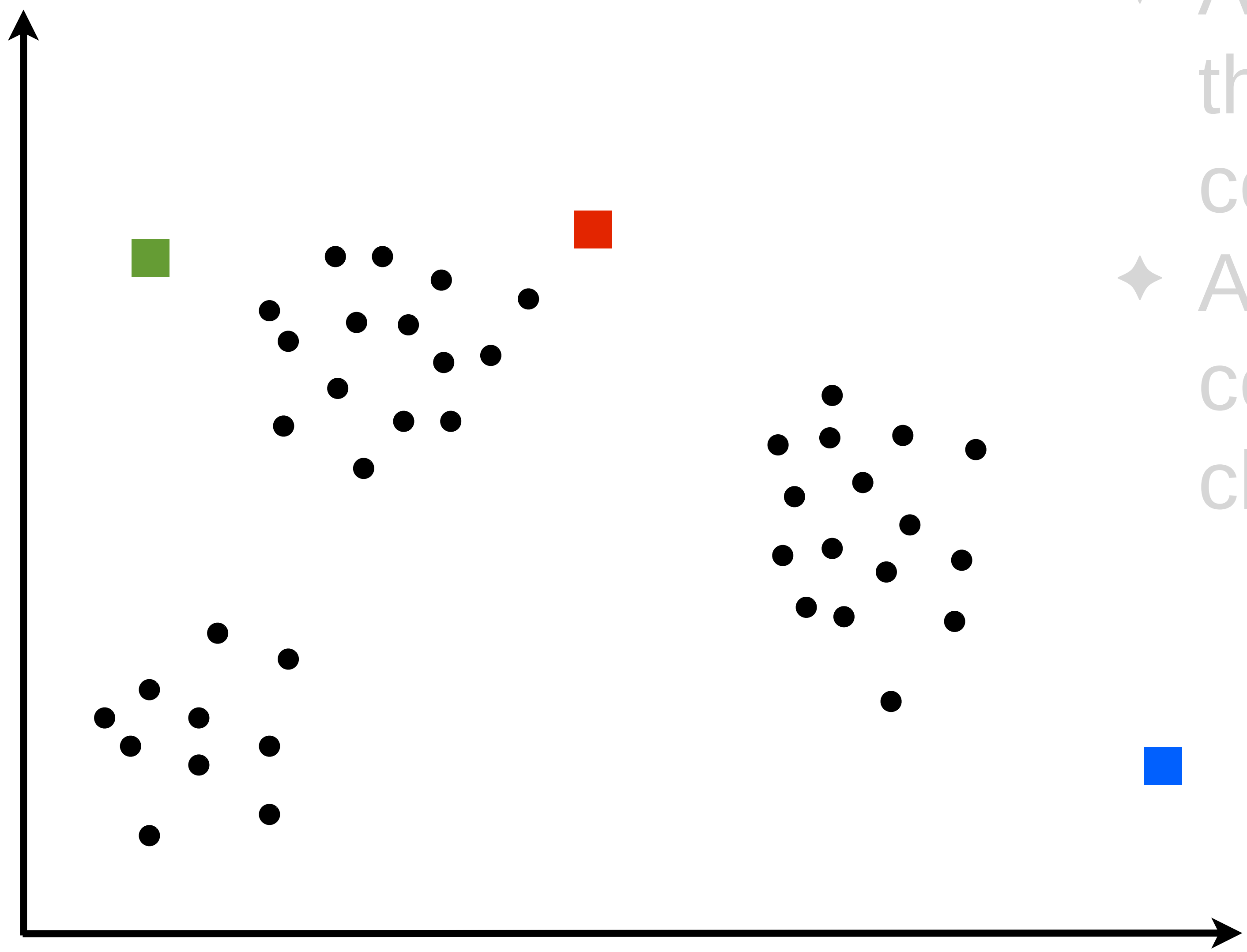
- For k = 1,…, K
  - ✦ Randomly draw n from 1,…,N without replacement
  - ✦ $\mu_k \leftarrow x_n$
- Repeat until convergence:
  - ✦ Assign each data point to the cluster with the closest center.
  - ✦ Assign each cluster center to be the mean of its cluster's data points



slide by Tamara Broderick

K-Means Algorithm

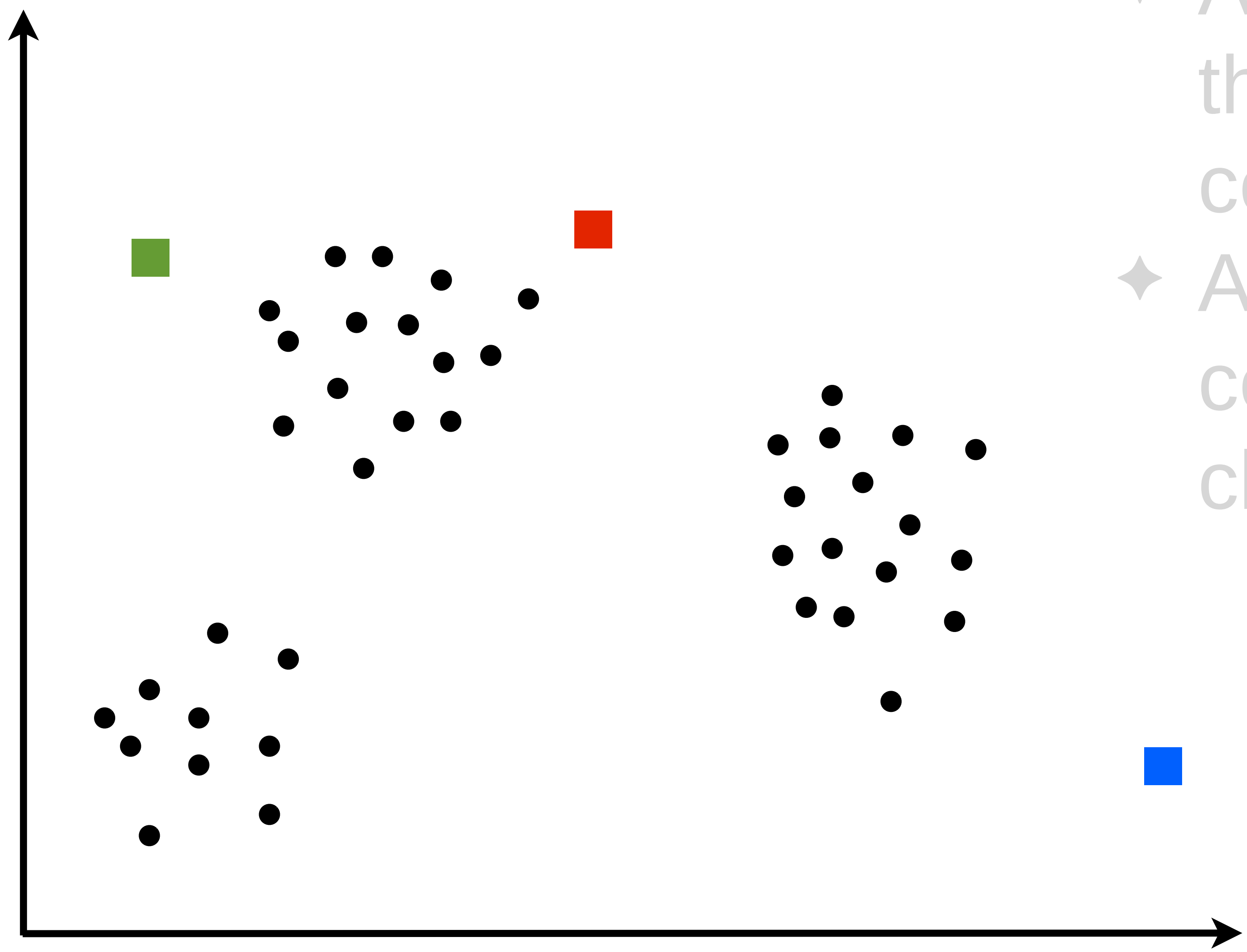
- For k = 1,…, K
  - ✦ Randomly draw n from 1,…,N without replacement
  - ✦ $\mu_k \leftarrow x_n$
- Repeat until convergence:
  - ✦ Assign each data point to the cluster with the closest center.
  - ✦ Assign each cluster center to be the mean of its cluster's data points



slide by Tamara Broderick

K-Means Algorithm

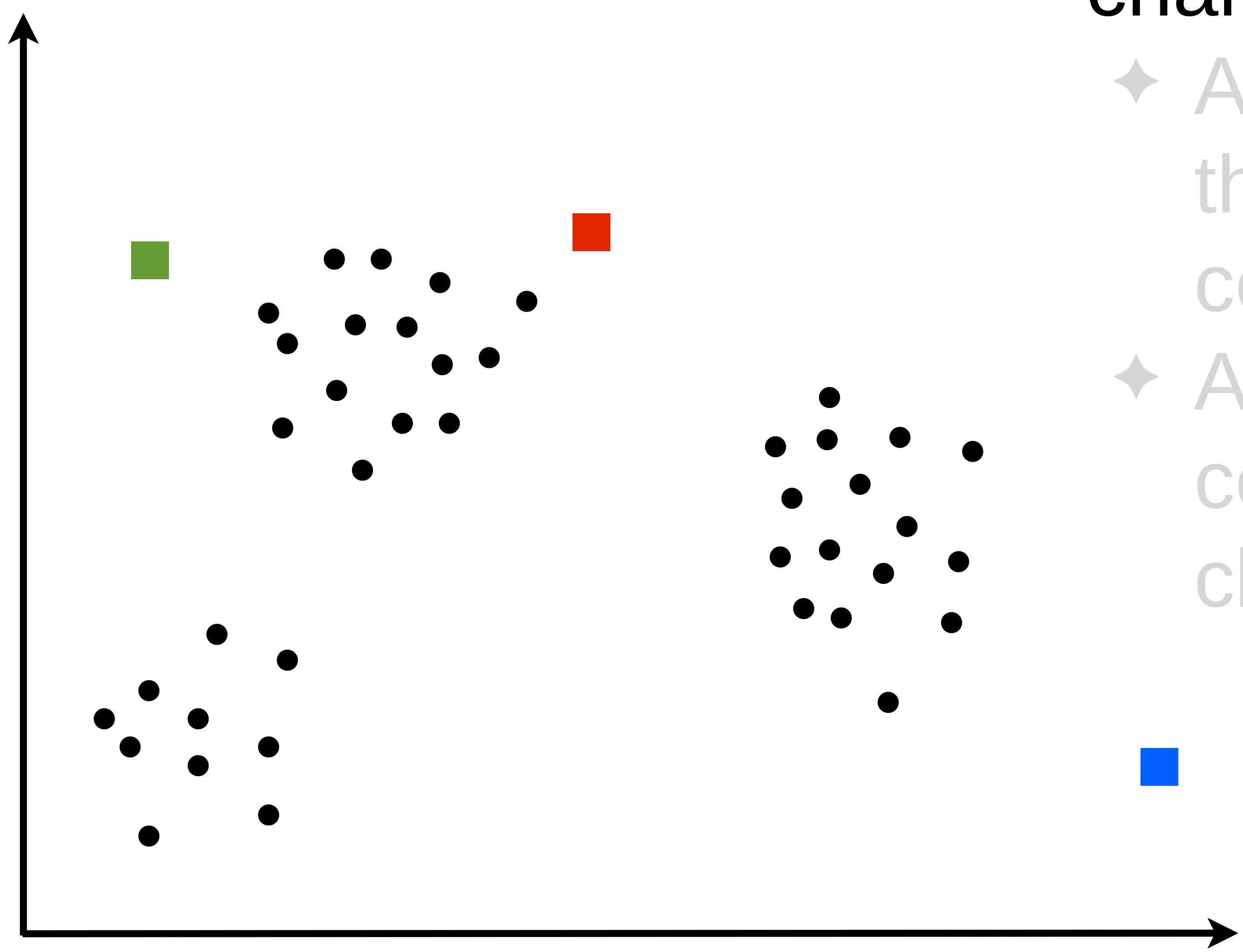
- For k = 1,…, K
  - ✦ Randomly draw n from 1,…,N without replacement
  - ✦ $\mu_k \leftarrow x_n$
- Repeat until convergence:
  - ✦ Assign each data point to the cluster with the closest center.
  - ✦ Assign each cluster center to be the mean of its cluster's data points



slide by Tamara Broderick

K-Means Algorithm

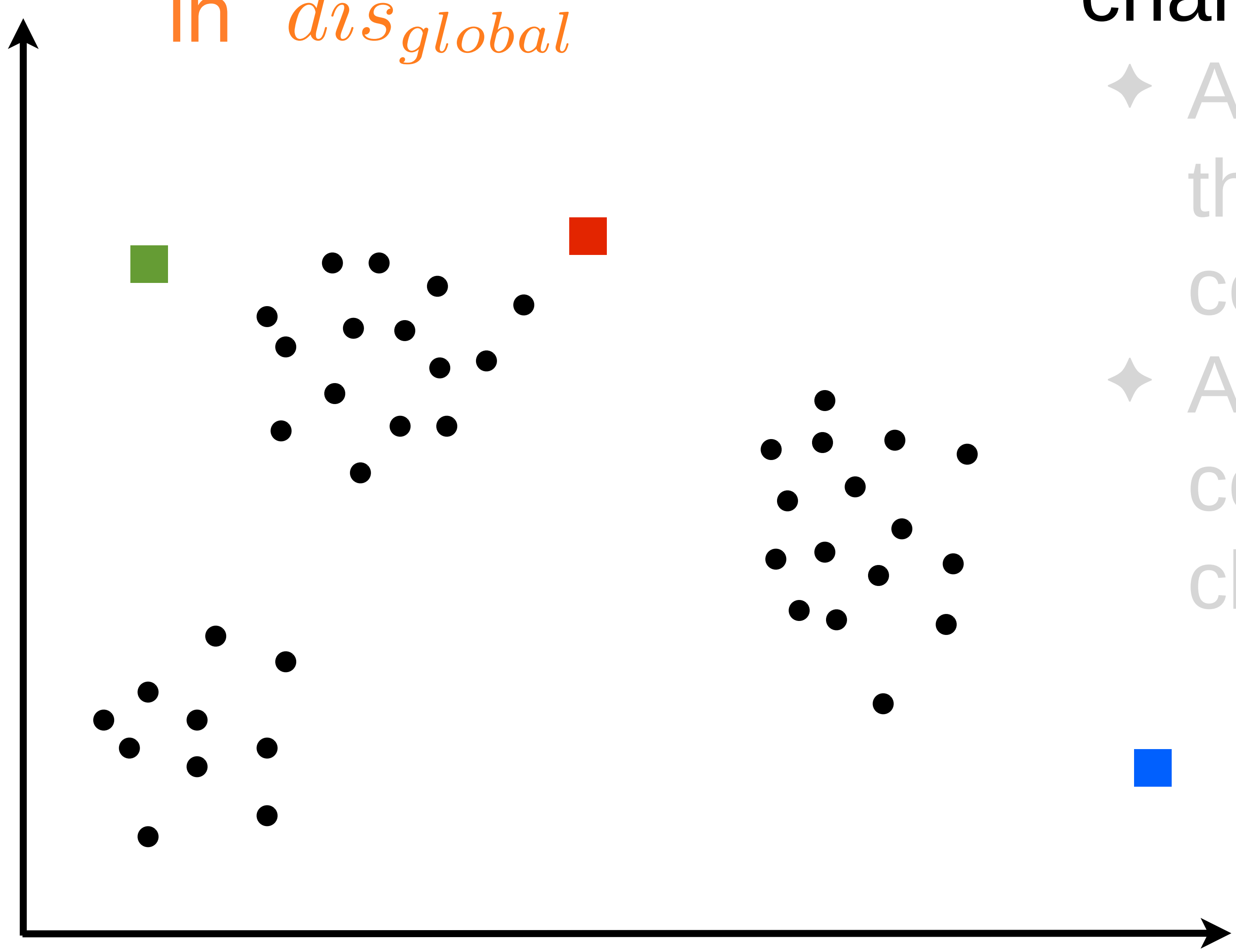
- For k = 1,…, K
  - Randomly draw n from 1,…,N without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until $S_1,\ldots,S_k$ don't change:
  - Assign each data point to the cluster with the closest center.
  - Assign each cluster center to be the mean of its cluster's data points



slide by Tamara Broderick

K-Means Algorithm

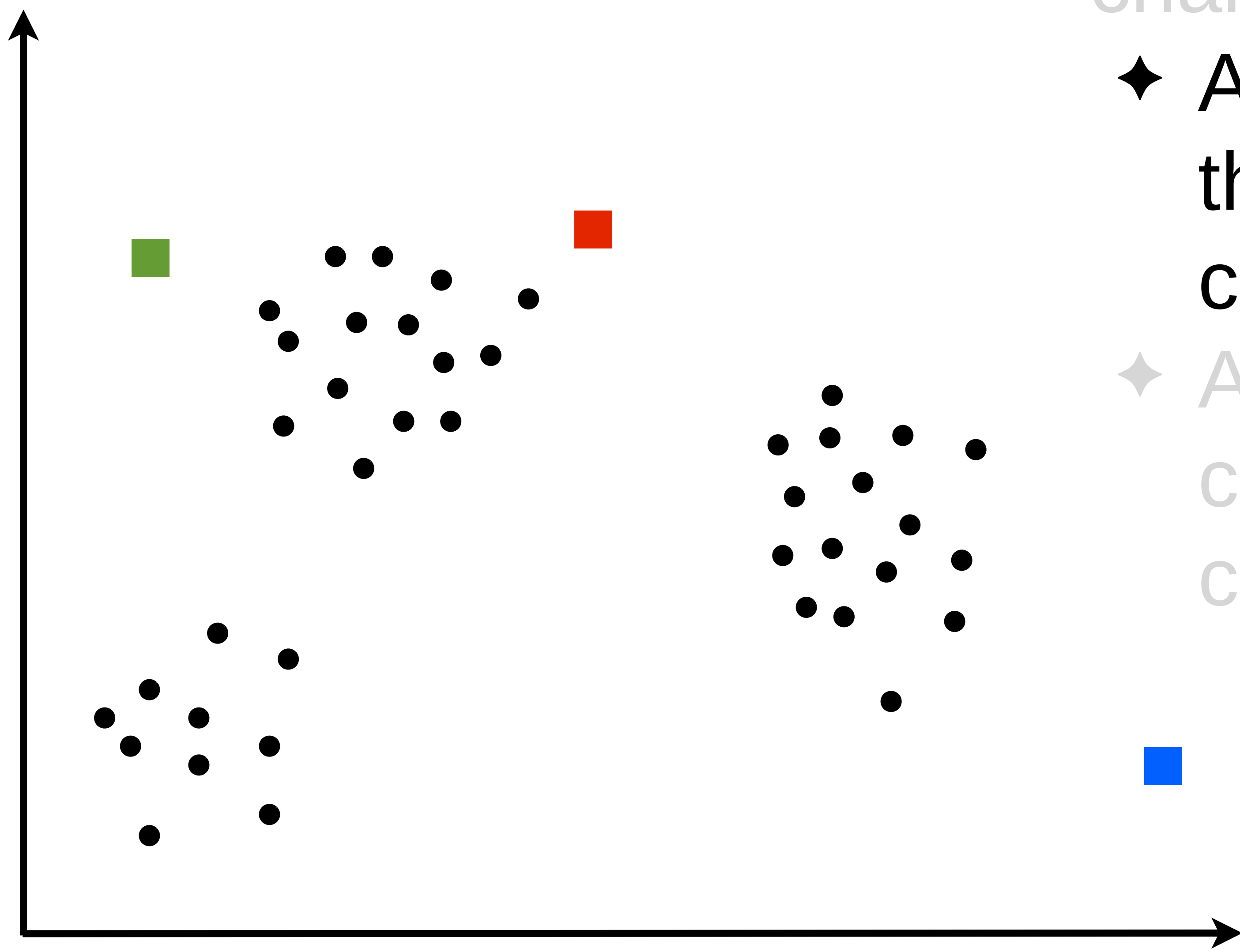
- For k = 1,…, K
  - ✦ Randomly draw n from 1,…,N without replacement
  - ✦ $\mu_k \longleftarrow x_n$
- Repeat until $S_1,\ldots,S_K$ don't change:
  - ✦ Assign each data point to the cluster with the closest center.
  - ✦ Assign each cluster center to be the mean of its cluster's data points



slide by Tamara Broderick

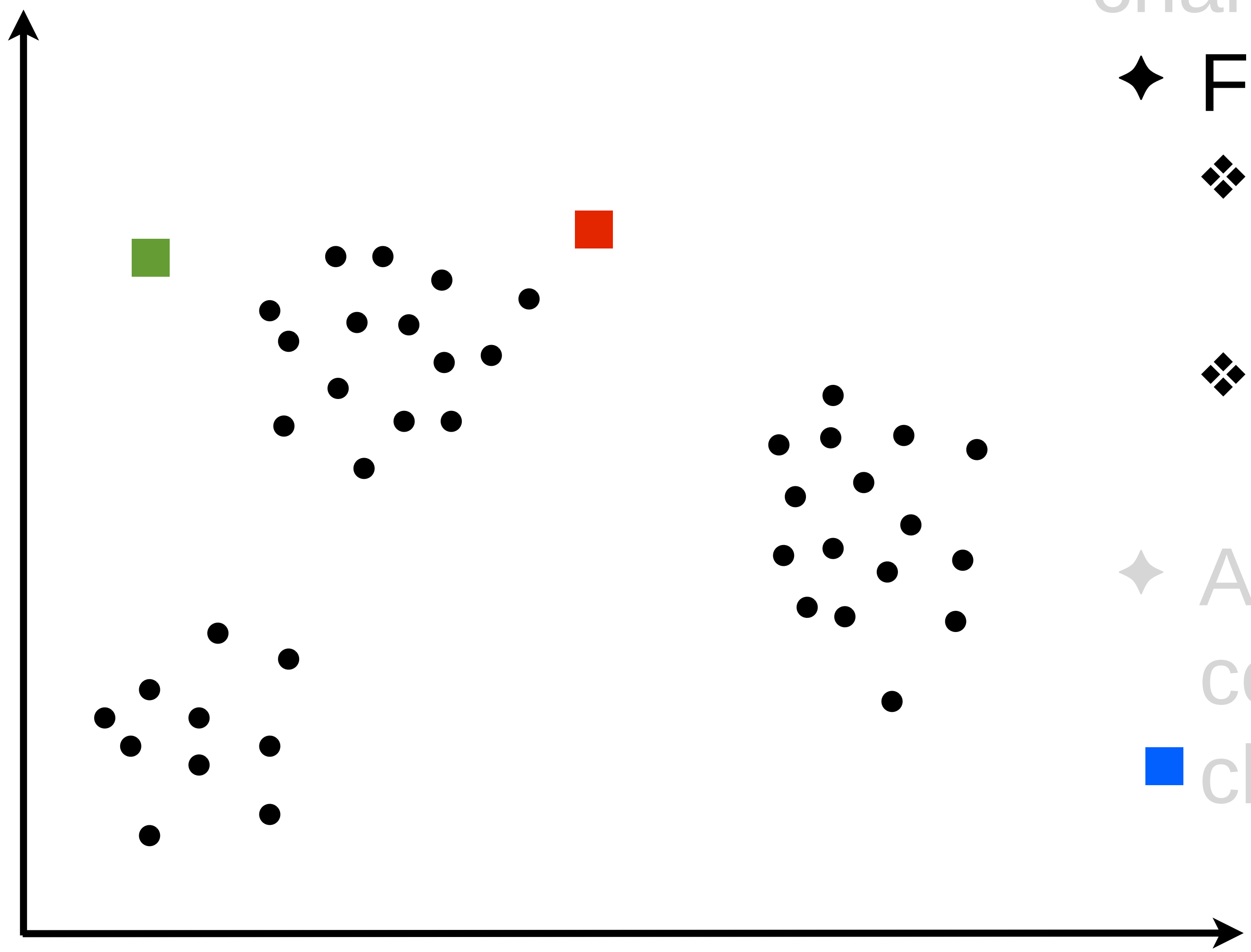
K-Means Algorithm

- For k = 1,…, K
  - Randomly draw n from 1,…,N without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until $S_1,\ldots,S_K$ 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



slide by Tamara Broderick

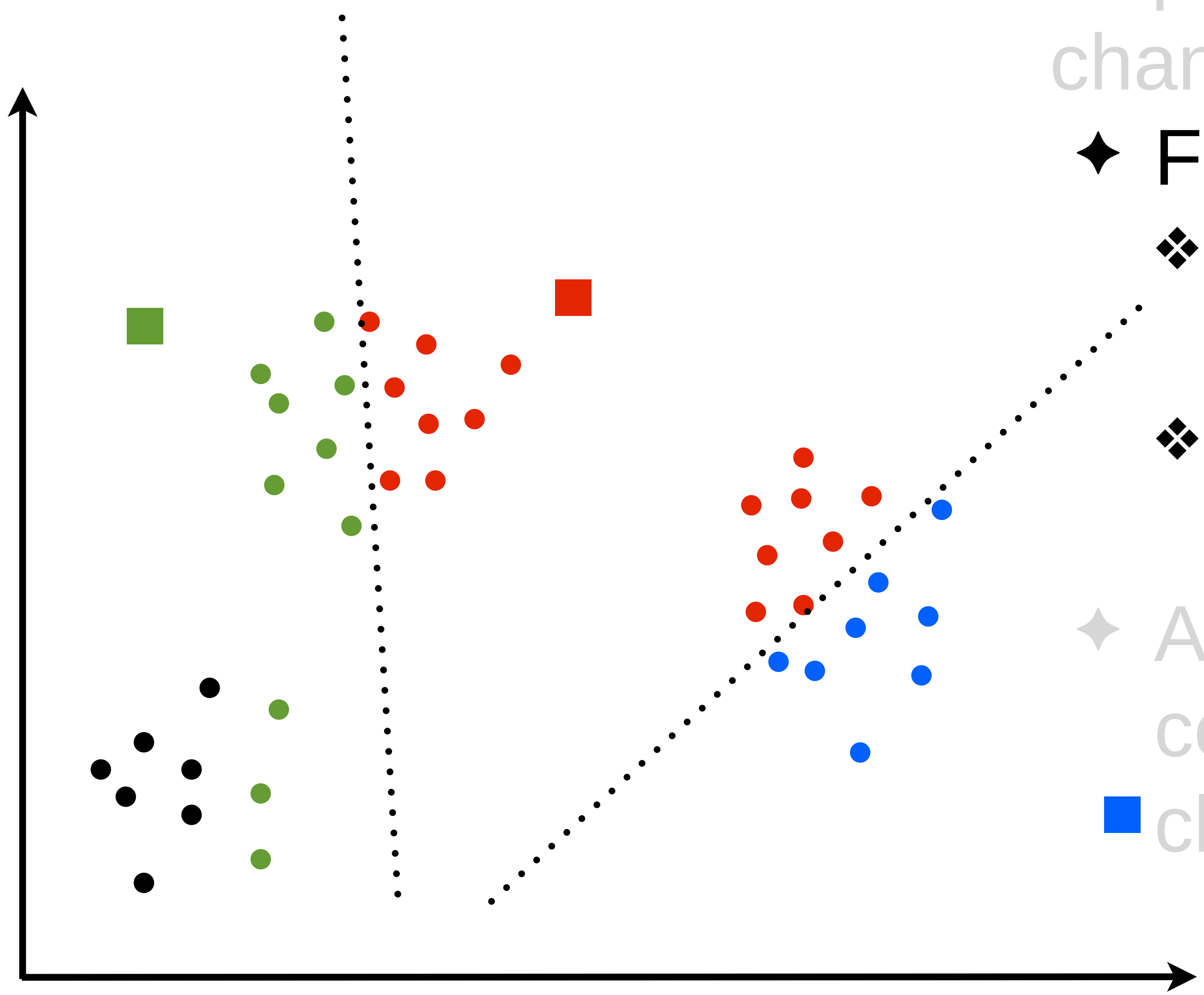
K-Means Algorithm



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

slide by Tamara Broderick

K-Means Algorithm

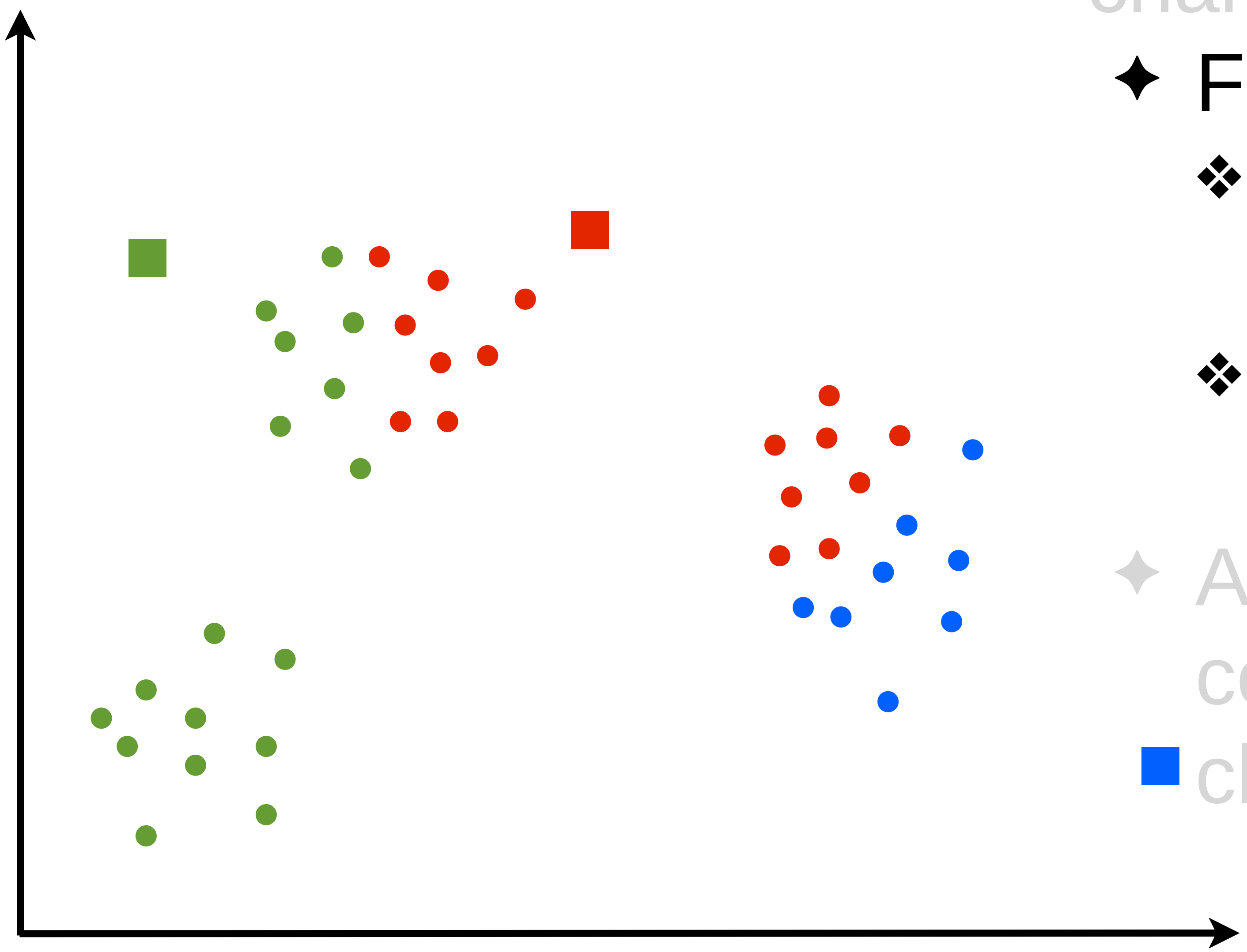


- For k = 1,…, K
  - ✦ Randomly draw n from 1,…,N without replacement
  - ✦ $\mu_k \leftarrow x_n$
- Repeat until $S_1, \ldots, S_K$ don't change:
  - ✦ For n = 1,…N
    - ❖ Find k with smallest $dis(x_n, \mu_k)$
    - ❖ Put $x_n \in S_k$ (and no other S_j)
  - ✦ Assign each cluster center to be the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

slide by Tamara Broderick

K-Means Algorithm

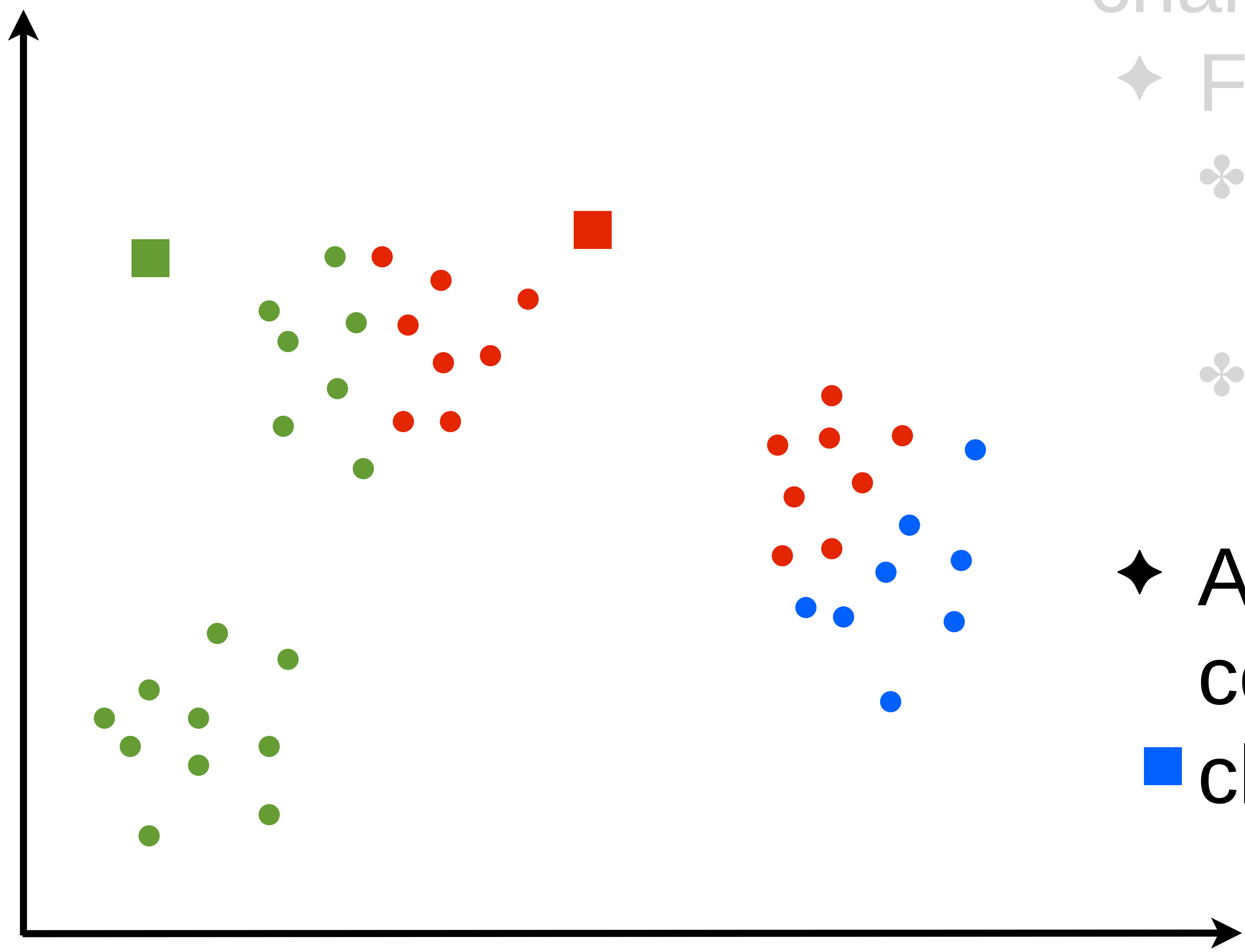
- For k = 1,…, K
  - ✦ Randomly draw n from 1,…,N without replacement
  - ✦ $\mu_k \leftarrow x_n$
- Repeat until $S_1,\ldots,S_K$ don't change:
  - ✦ For n = 1,…N
    - ❖ Find k with smallest $dis(x_n, \mu_k)$
    - ❖ Put $x_n \in S_k$ (and no other S_j)
  - ✦ Assign each cluster center to be the mean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



slide by Tamara Broderick

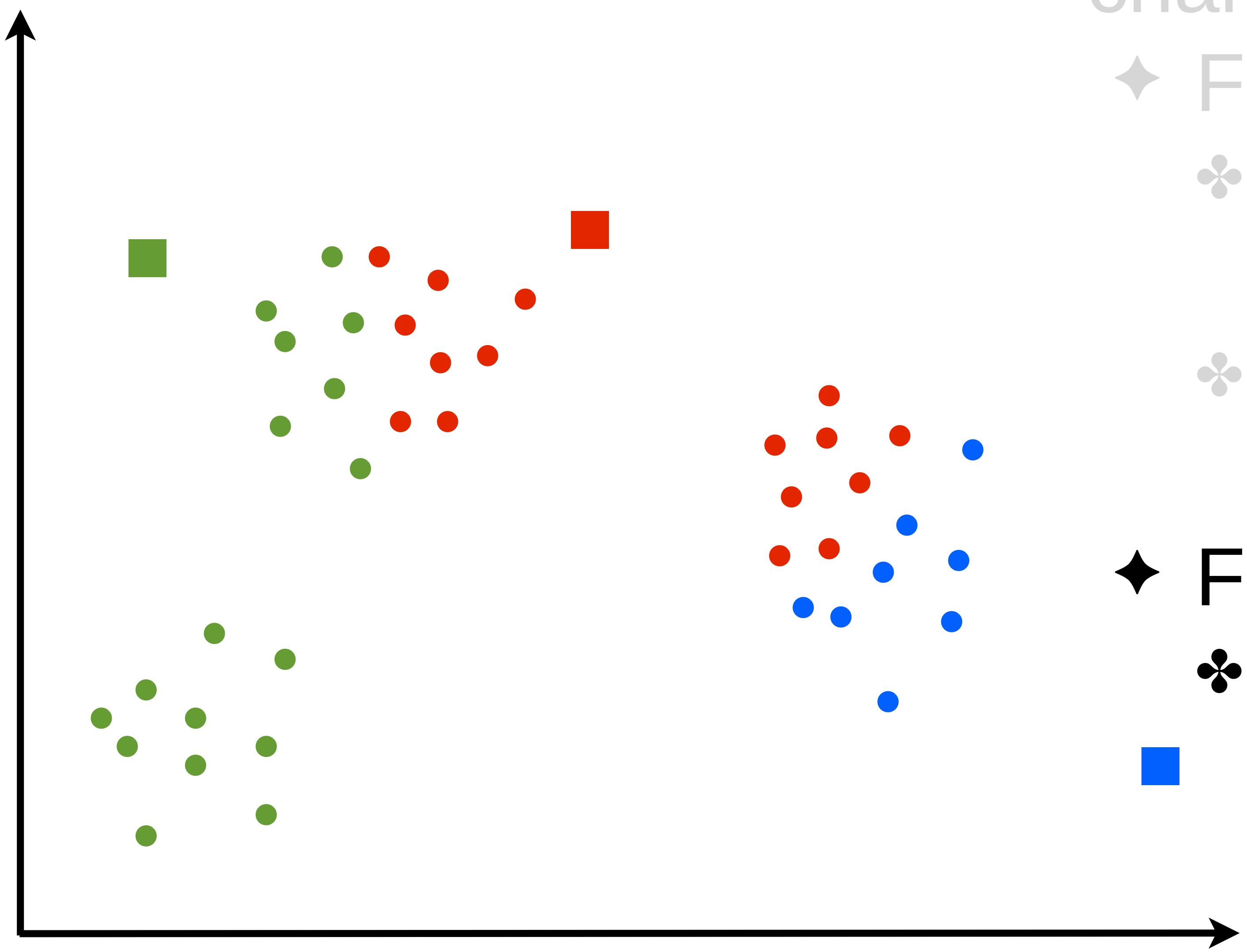
K-Means Algorithm

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



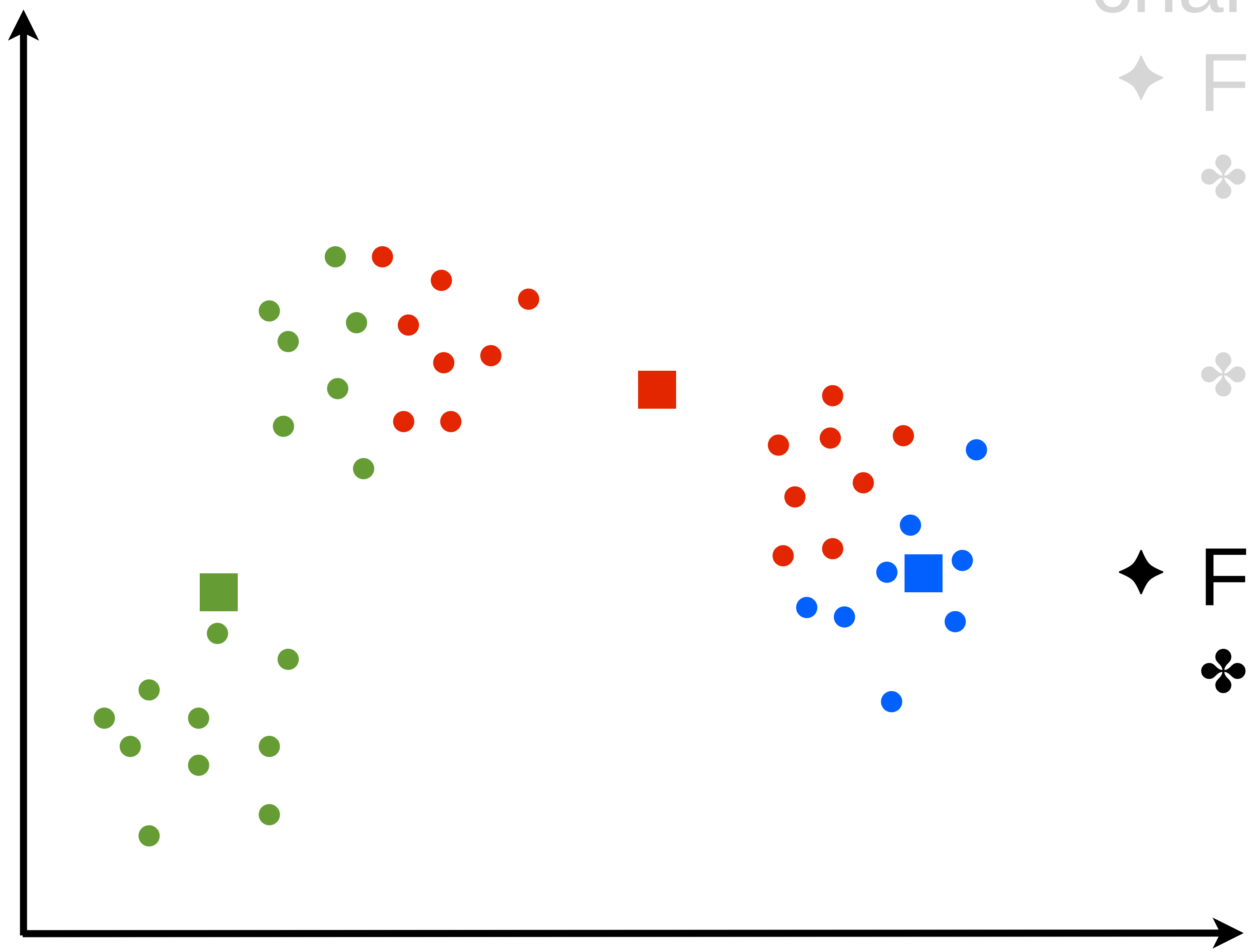
K-Means Algorithm

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



slide by Tamara Broderick

K-Means Algorithm

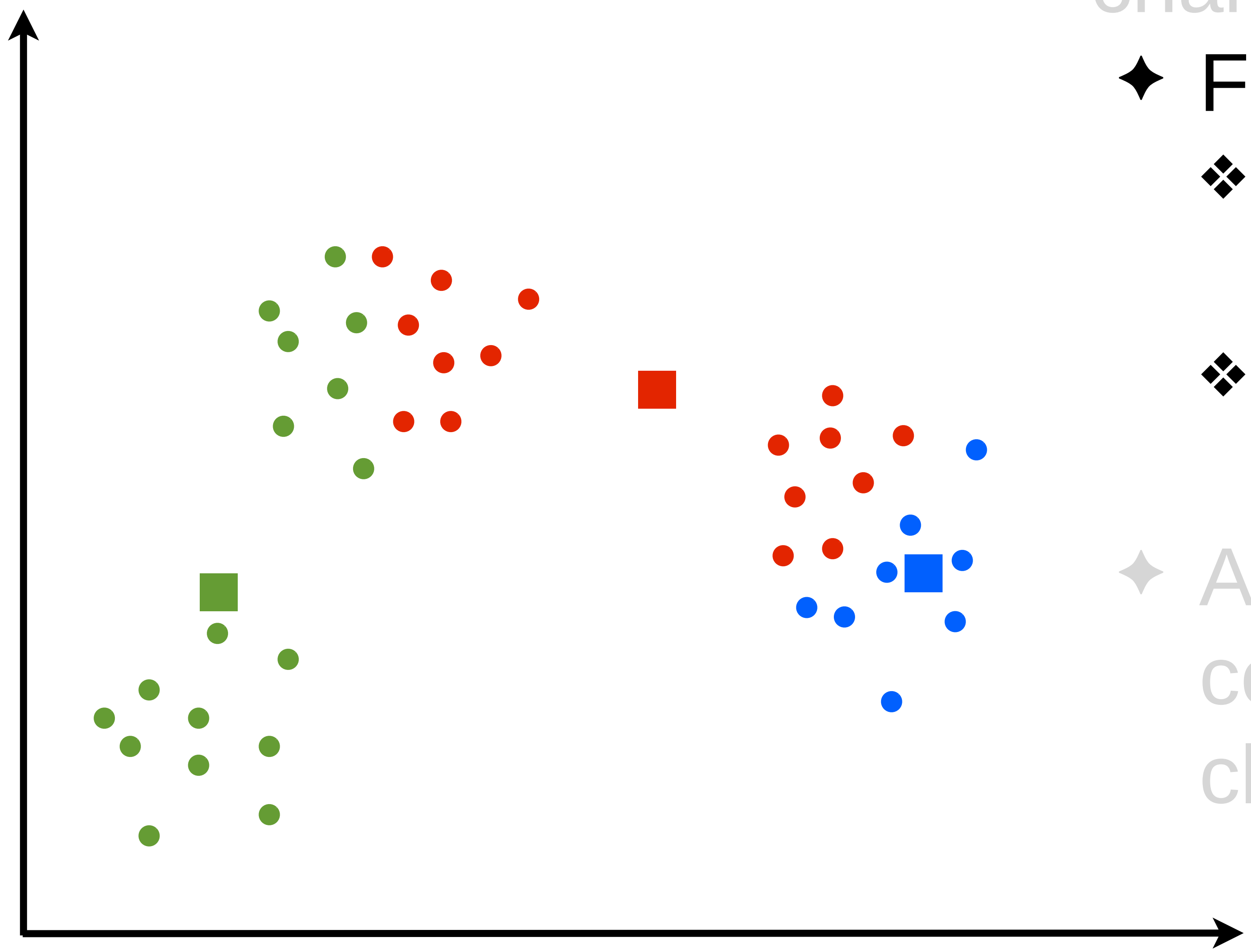


- For k = 1,…, K
  - Randomly draw n from 1,…,N without replacement
  - $\mu_k \leftarrow x_n$
- Repeat until $S_1, \ldots, S_K$ don't change:
  - For n = 1,…N
    - Find k with smallest $dis(x_n, \mu_k)$
    - Put $x_n \in S_k$ (and no other S_j)
  - For k = 1,…,K
    - $\mu_k \leftarrow |S_k|^{-1}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

slide by Tamara Broderick

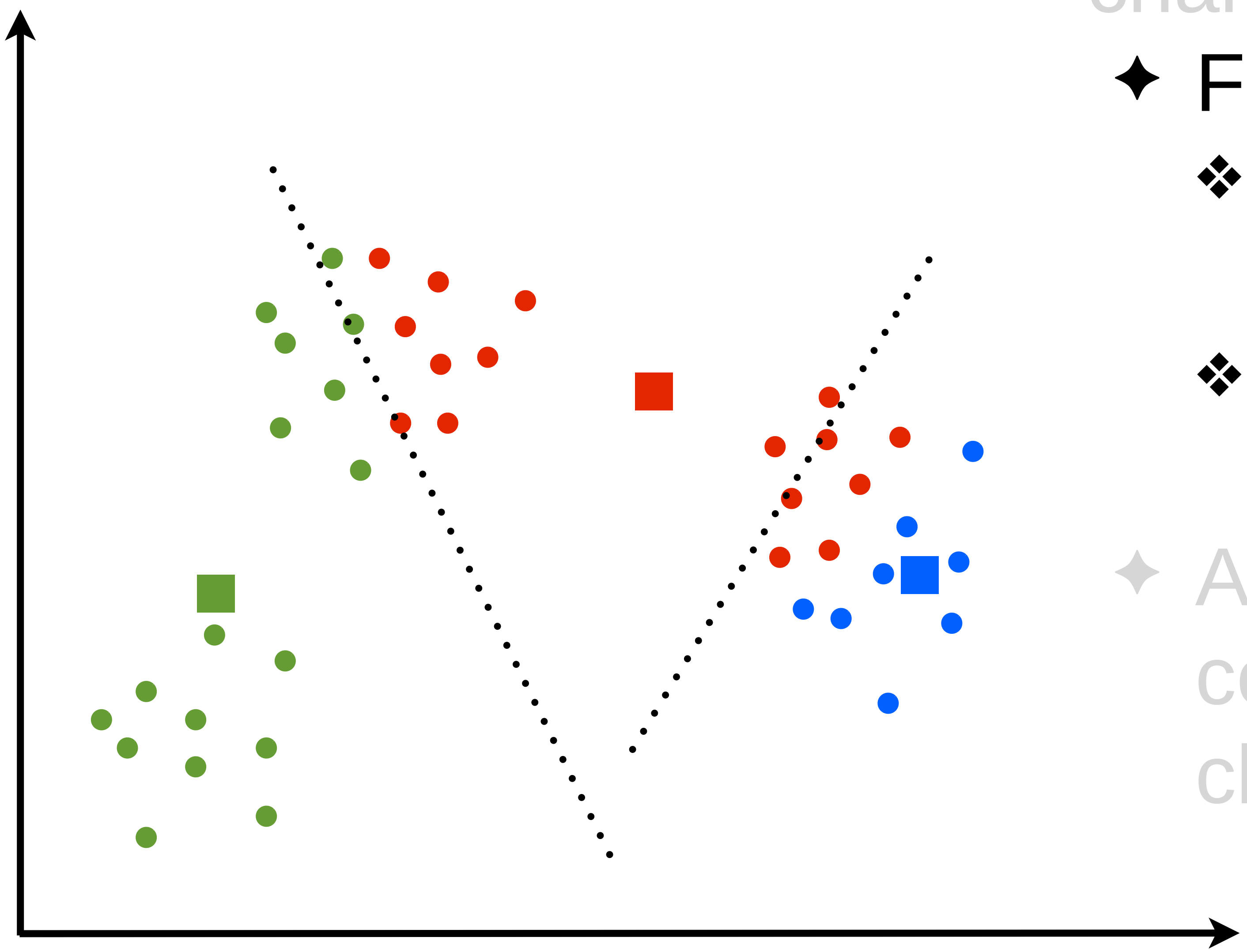
K-Means Algorithm

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



slide by Tamara Broderick

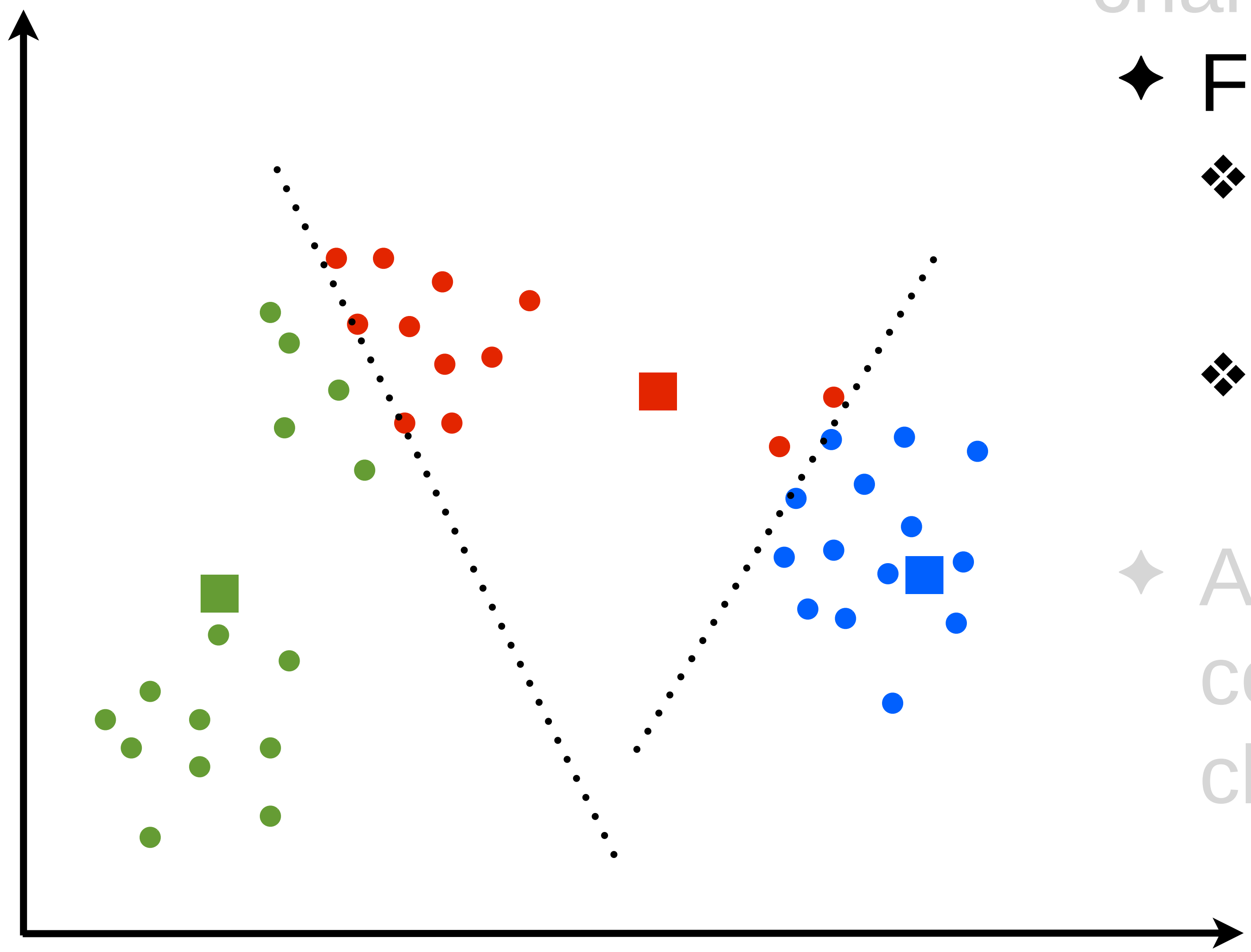
K-Means Algorithm



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

slide by Tamara Broderick

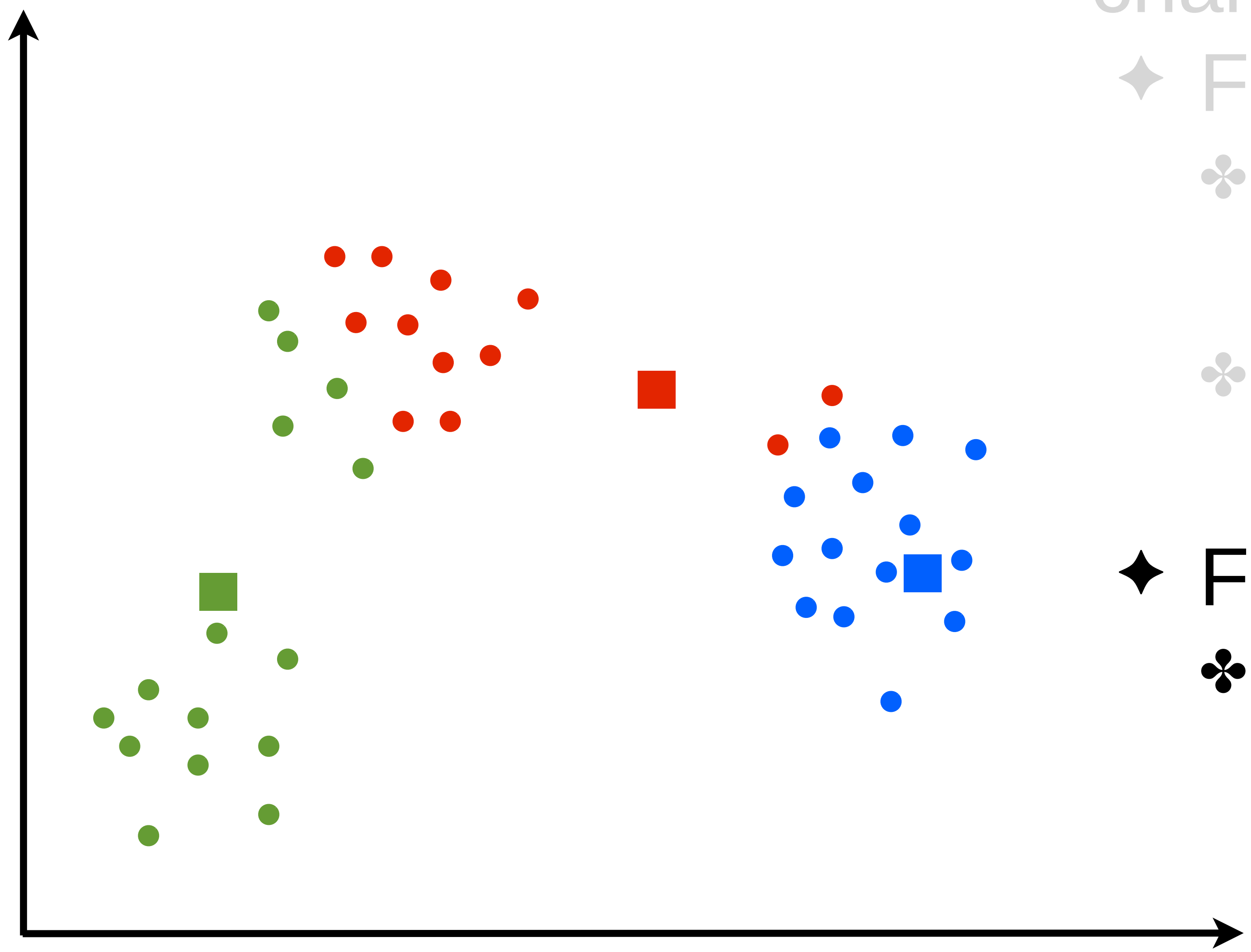
K-Means Algorithm



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

slide by Tamara Broderick

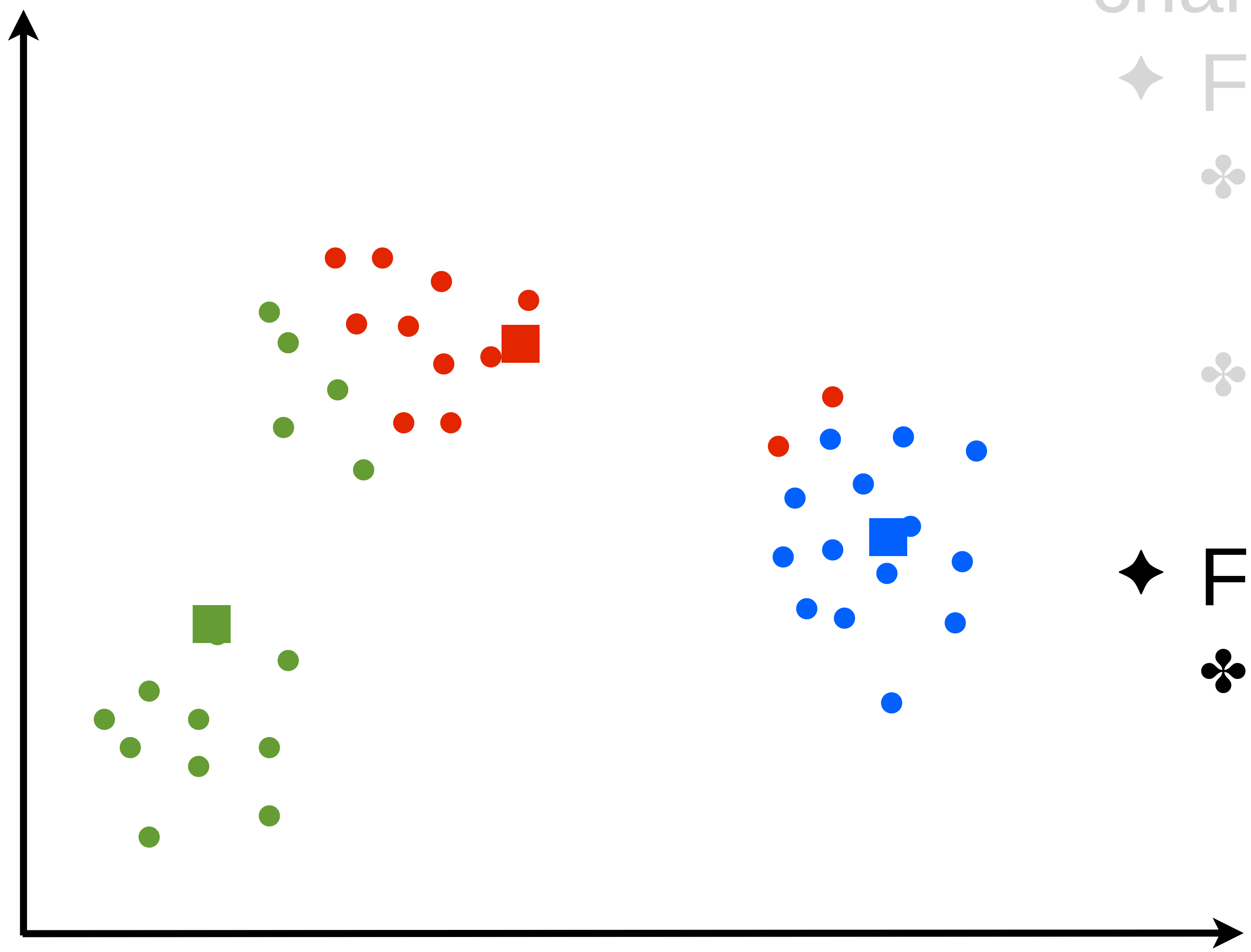
K-Means Algorithm



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

slide by Tamara Broderick

K-Means Algorithm

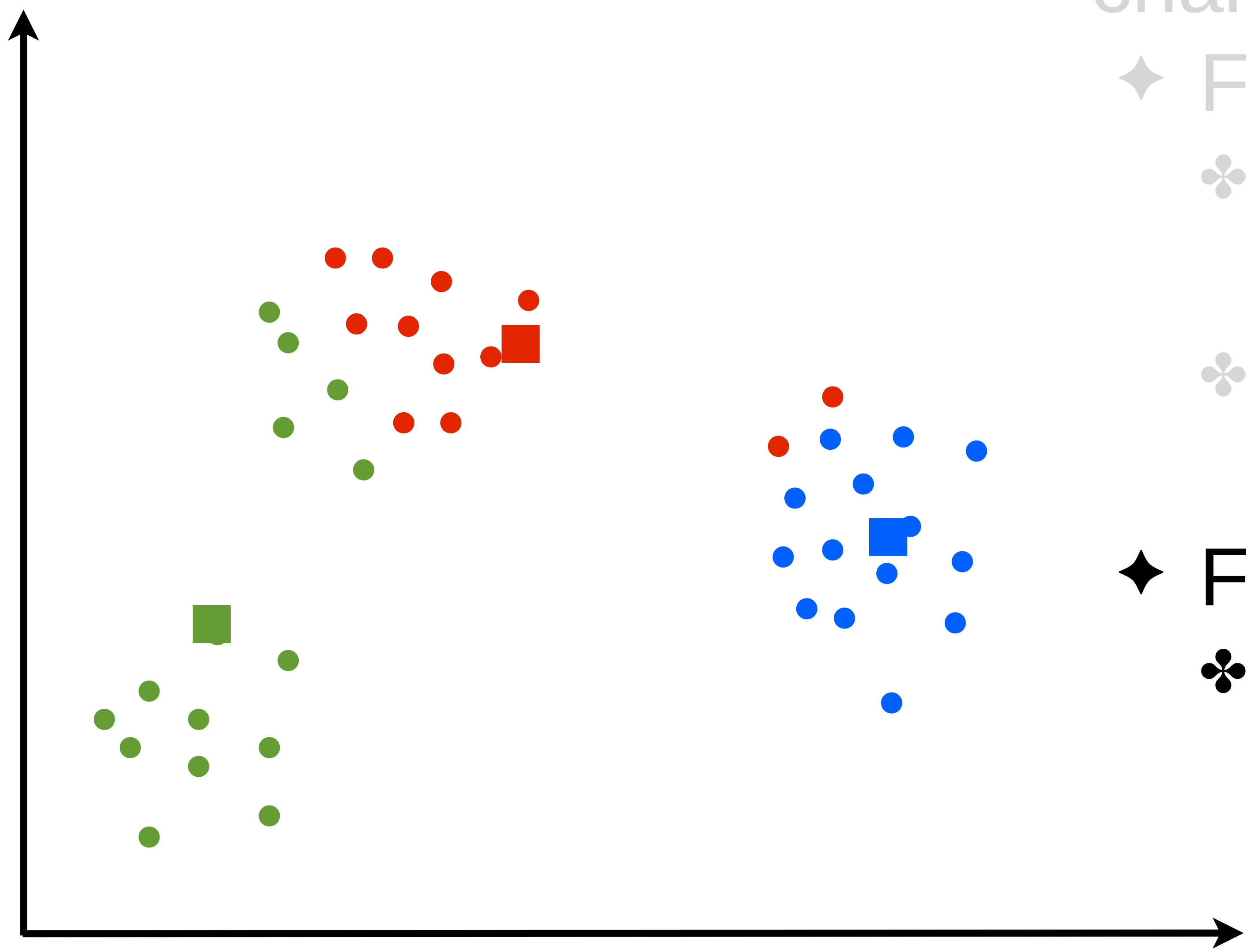


- For k = 1,..., K
  - ✦ Randomly draw n from 1,…,N without replacement
  - ✦ $\mu_k \leftarrow x_n$
- Repeat until $S_1,\ldots,S_K$ don't change:
  - ✦ For n = 1,…N
    - ✤ Find k with smallest $dis(x_n, \mu_k)$
    - ✤ Put $x_n \in S_k$ (and no other S_j)
  - ✦ For k = 1,….K
    - ✤ $\mu_k \leftarrow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

slide by Tamara Broderick

K-Means Algorithm

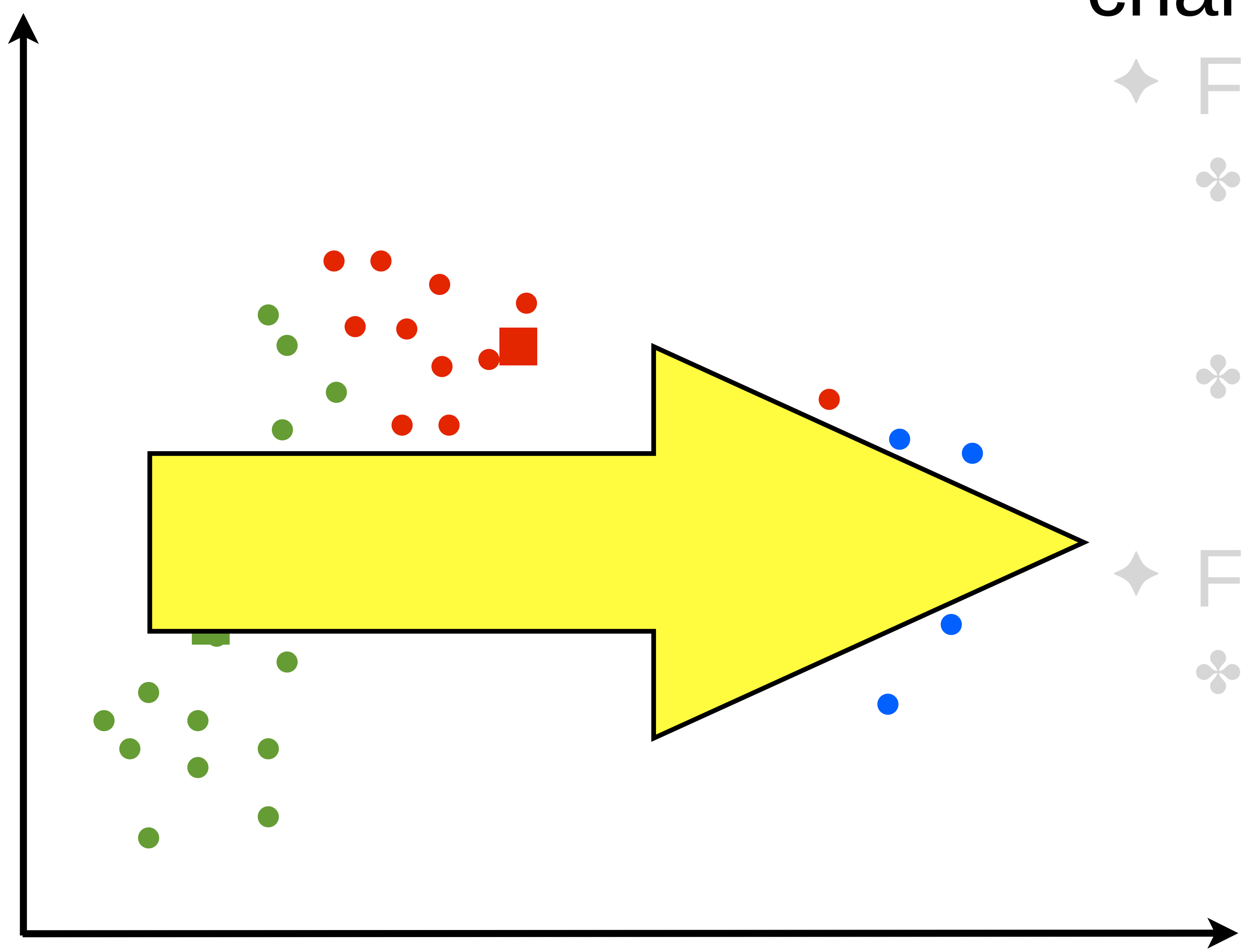
- For k = 1,…, K
  - ✦ Randomly draw n from 1,…,N without replacement
  - ✦ $\mu_k \leftarrow x_n$
- Repeat until $S_1,\dots,S_K$ don't change:
  - ✦ For n = 1,…N
    - ✤ Find k with smallest $dis(x_n, \mu_k)$
    - ✤ Put $x_n \in S_k$ (and no other S_j)
  - ✦ For k = 1,…,K
    - ✤ $\mu_k \leftarrow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



slide by Tamara Broderick

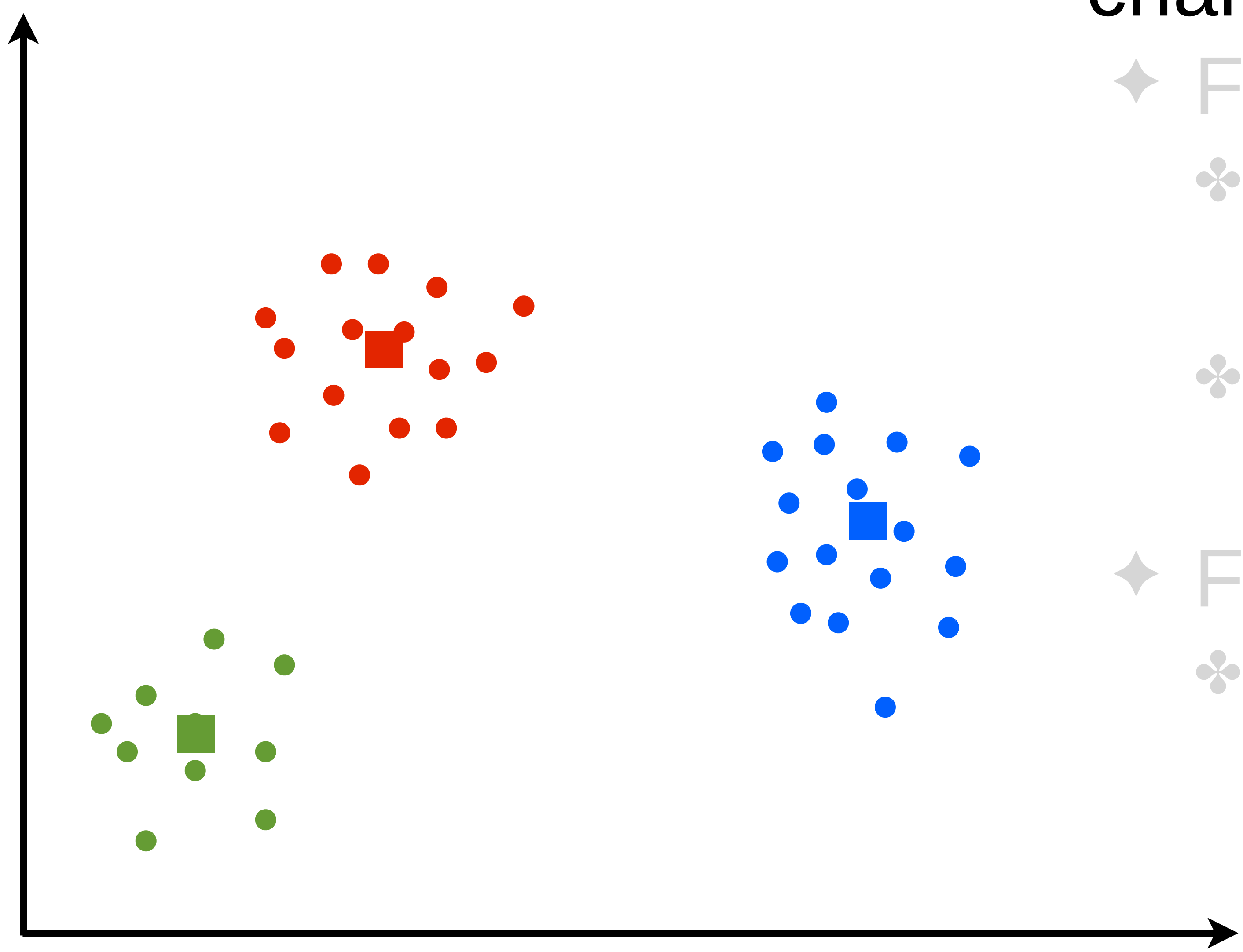
K-Means Algorithm

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



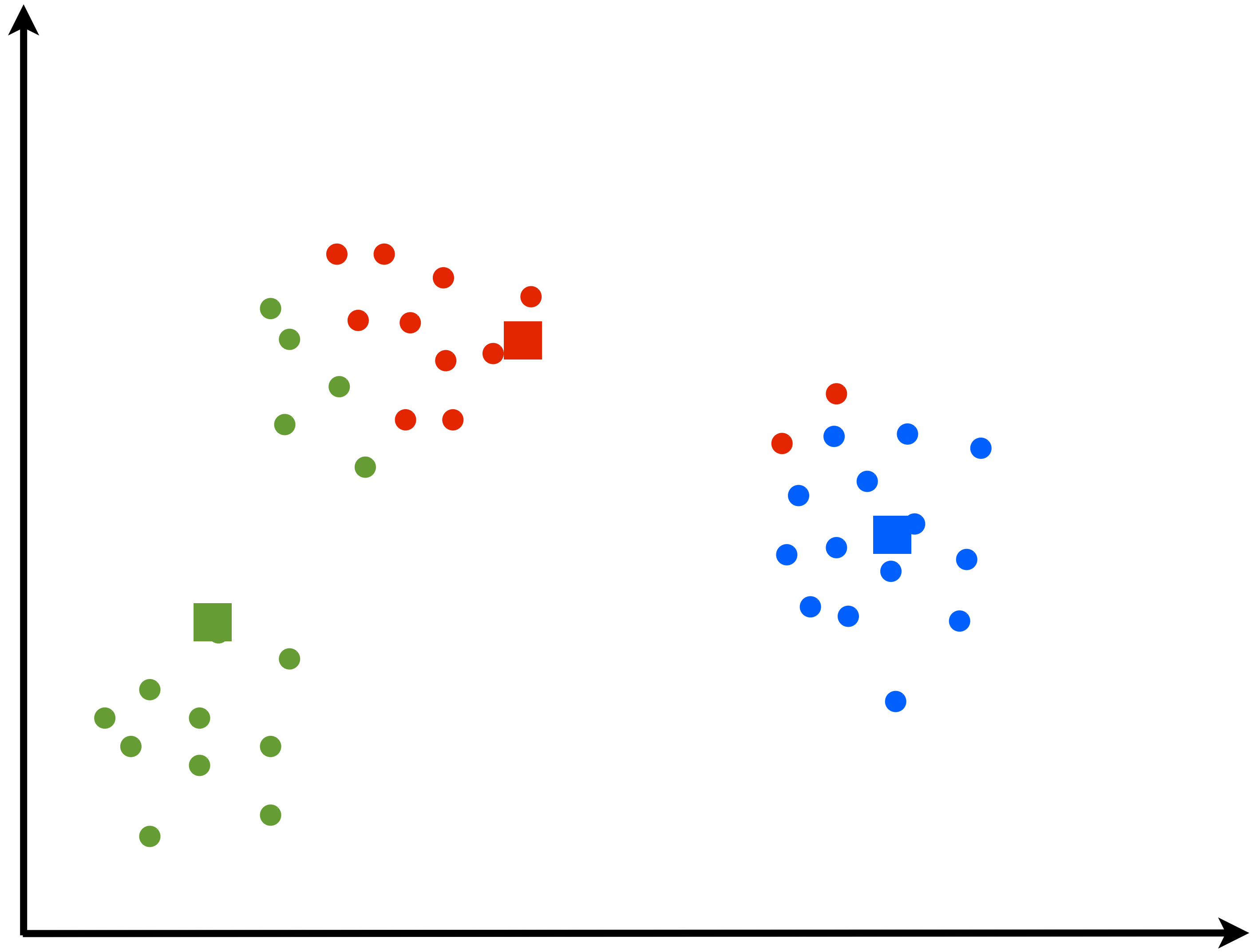
K-Means Algorithm

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



slide by Tamara Broderick

K-Means: Evaluation

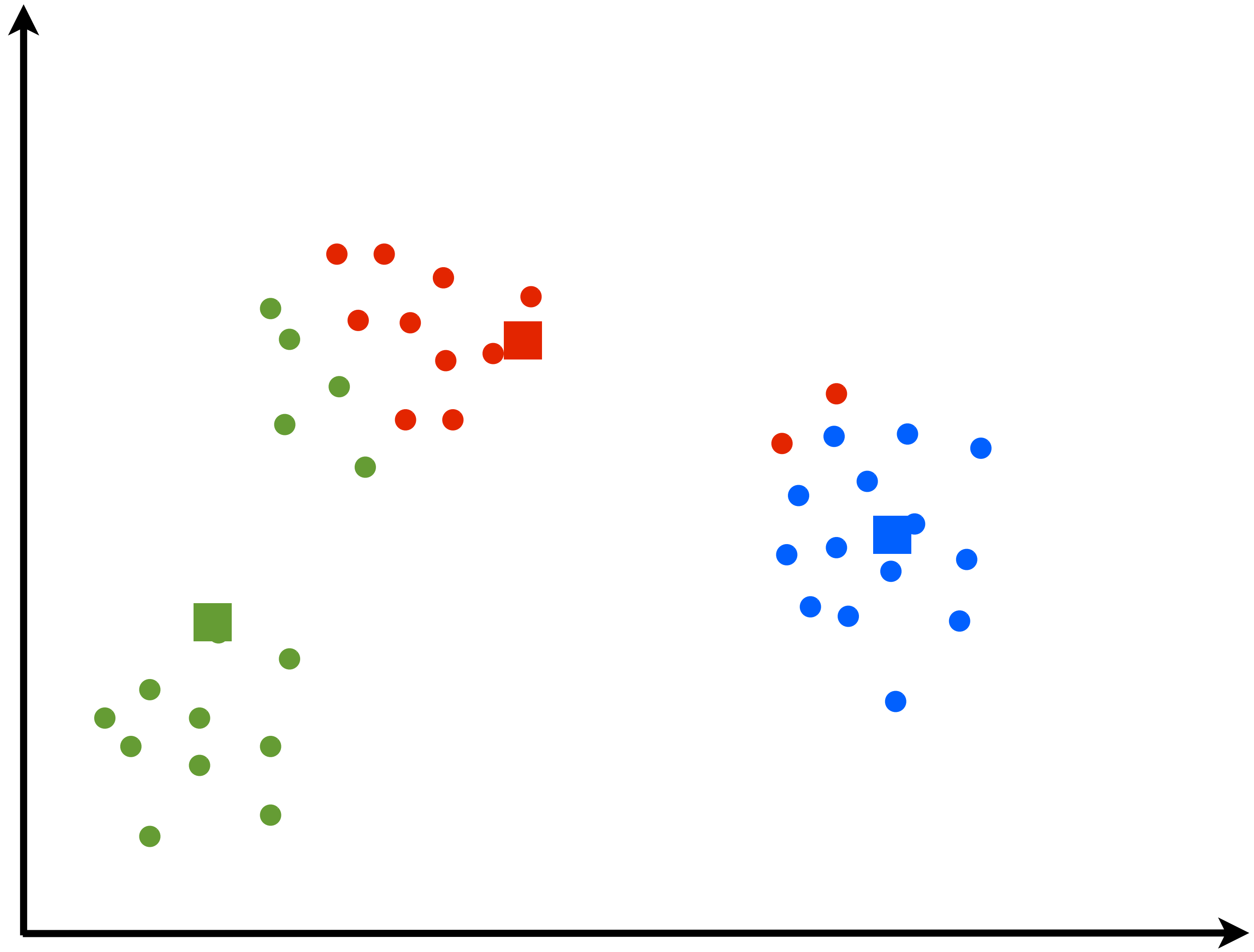


slide by Tamara Broderick

K-Means: Evaluation

- Will it terminate?

  Yes. Always.



slide by Tamara Broderick

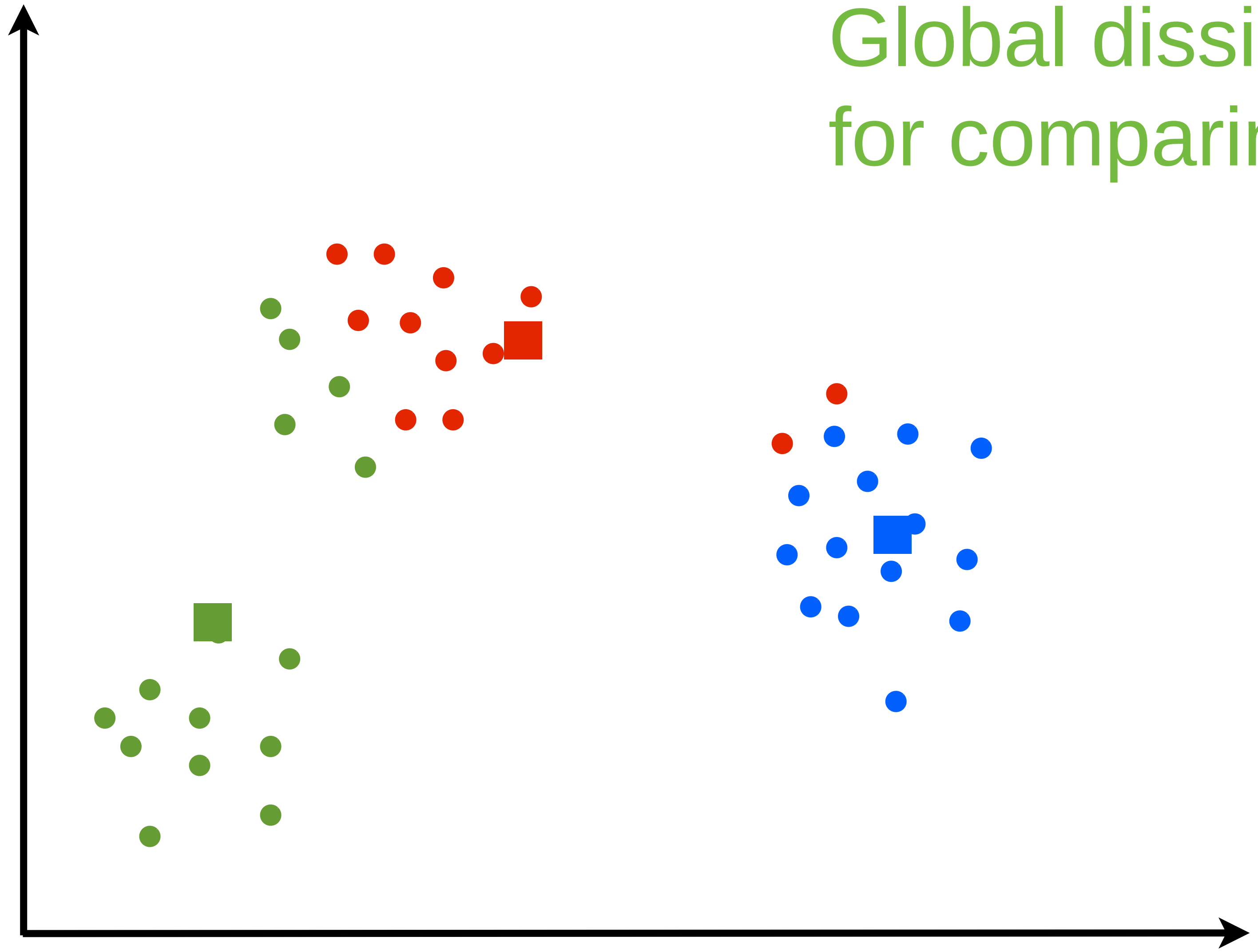
K-Means: Evaluation

- Will it terminate?

  Yes. Always.

- Is the clustering any good?

  Global dissimilarity only useful for comparing clusterings.



slide by Tamara Broderick

K-Means: Evaluation

- Guaranteed to converge in a finite number of iterations

- Running time per iteration:
  1. Assign data points to closest cluster center
     O(KN) time
  2. Change the cluster center to the average of its assigned points
     O(N) time

slide by David Sontag

K-Means: Evaluation

Objective $\min_{\mu} \min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$

1. Fix μ, optimize C:

$$\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 = \min_{c} \sum_{i}^{n} |x_i - \mu_{x_i}|^2$$

Step 1 of kmeans

2. Fix C, optimize μ:

$$\min_{\mu} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

   – Take partial derivative of μ_i and set to zero, we have

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Step 2 of kmeans

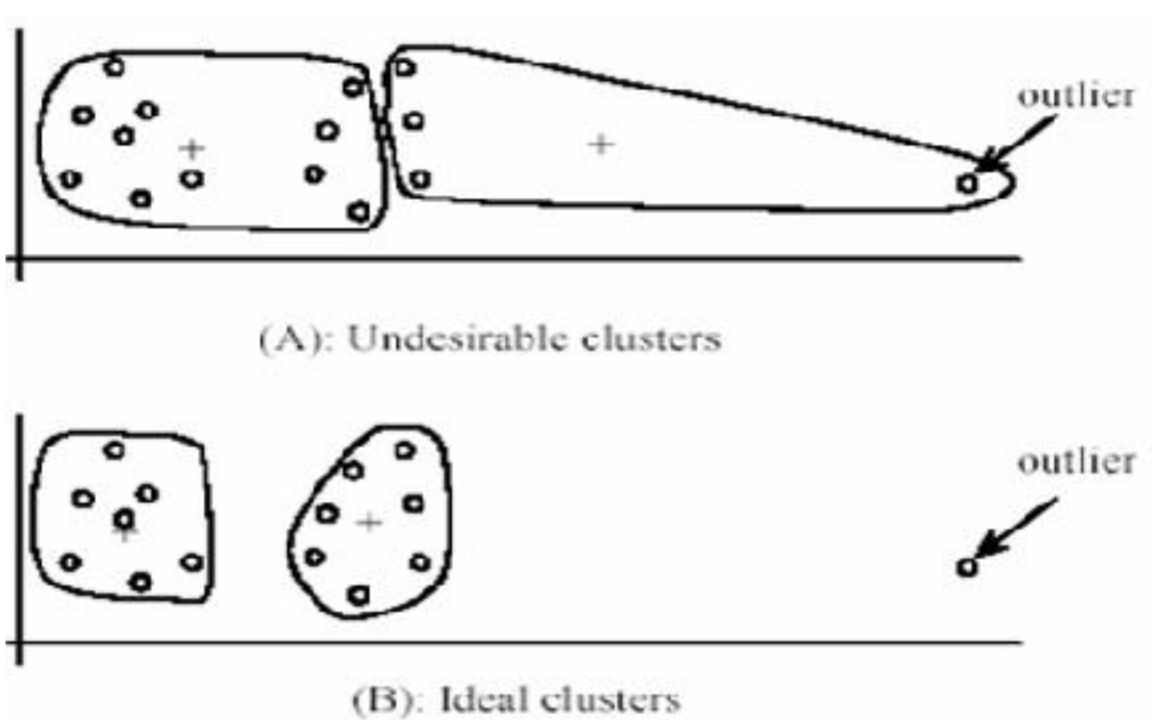
K-Means takes an alternating optimization approach, each step is guaranteed to decrease the objective – thus guaranteed to converge

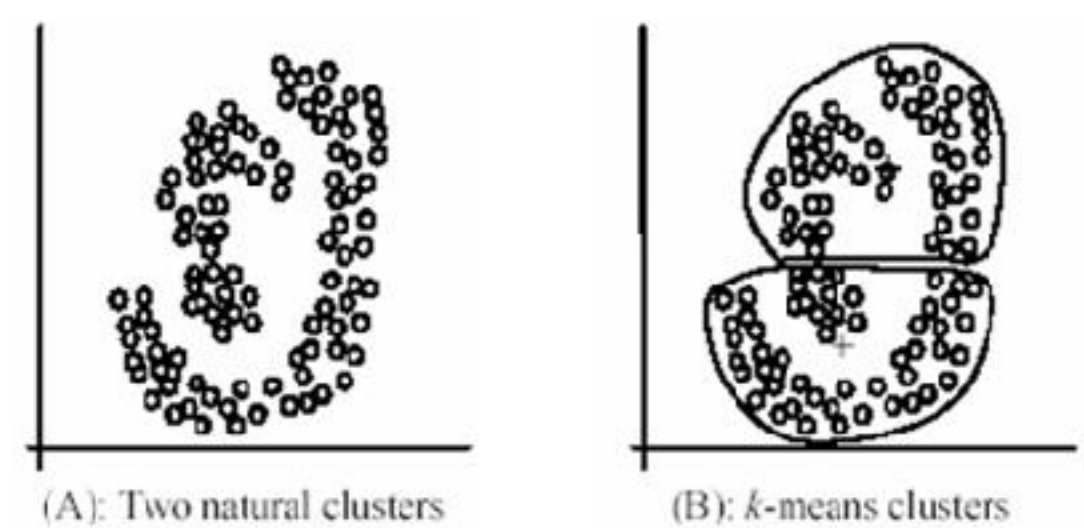
slide by Alan Fern

Demo time…

K-Means Algorithm: Some Issues

- How to set k?
- Sensitive to initial centers
  - Multiple initializations
- Sensitive to outliers
- Detects spherical clusters
- Assuming means can be computed
  - It requires continuous, numerical features





slide by Kristen Grauman

Next Lecture:

K-Means Applications,
Spectral clustering,
Hierarchical clustering and
What is a good clustering?