Data Analytics for Social Science
Cluster analysis

Johan A. Elkink

School of Politics & International Relations
University College Dublin

2 April 2020

Outline

Unsupervised learning

So far we have looked at models that in the machine learning literature are called **supervised learning**.

The idea here is that the algorithm learns to find some pattern in the data, whereby—at least for the training data—the correct answer is known.

E.g. we know which countries are democratic and which are not, but try to predict based on a set of variables.

With **unsupervised learning**, there is no *a priori* labelling of the data.

Dimension reduction

Another way of looking at this is that data analysis typically has one of three purposes:

- Explaining / causal inference
- Prediction / classification
- Dimension reduction / clustering

Outline

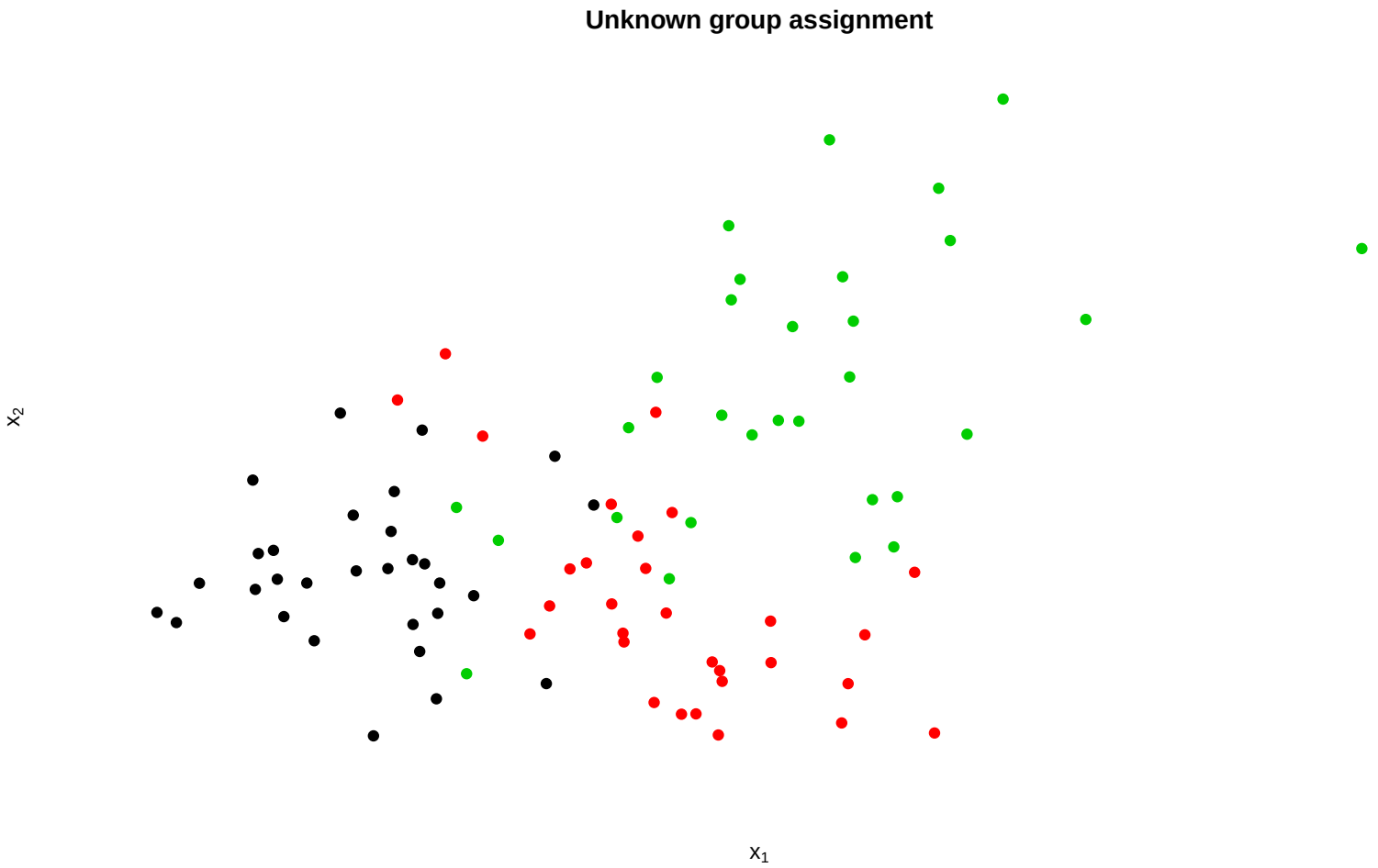
K-Means clustering

K-Means clustering is a clustering algorithm to search for an optimal clustering in k groups:

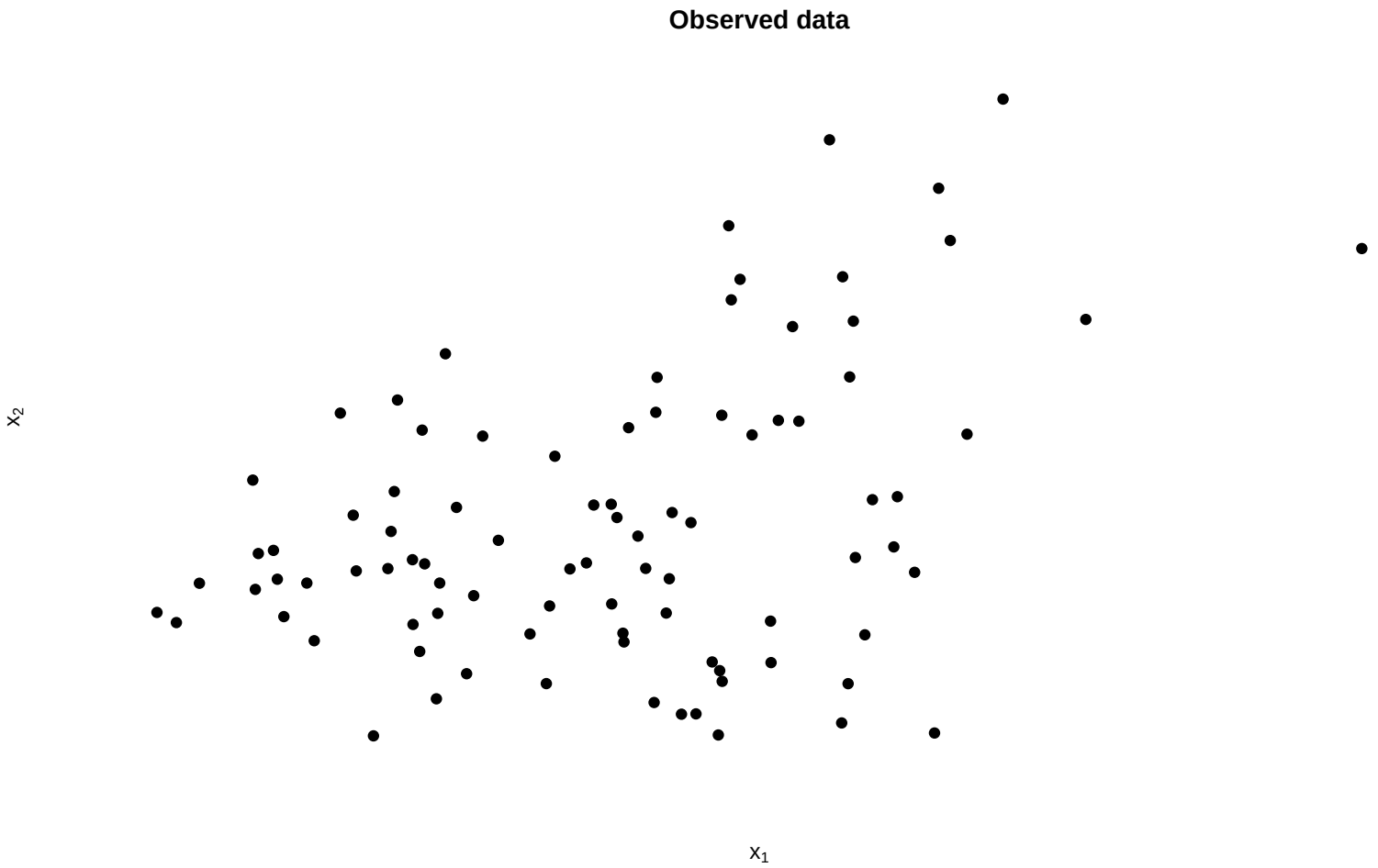
1. Generate k random starting points.
2. Assign each observation to the group of the nearest of the k points.
3. Move the k points to the mean of their assigned observations.
4. Repeat 2–3 until no significant improvements are made.

(Hastie, Tibshirani and Friedman, 2009, 509–510)

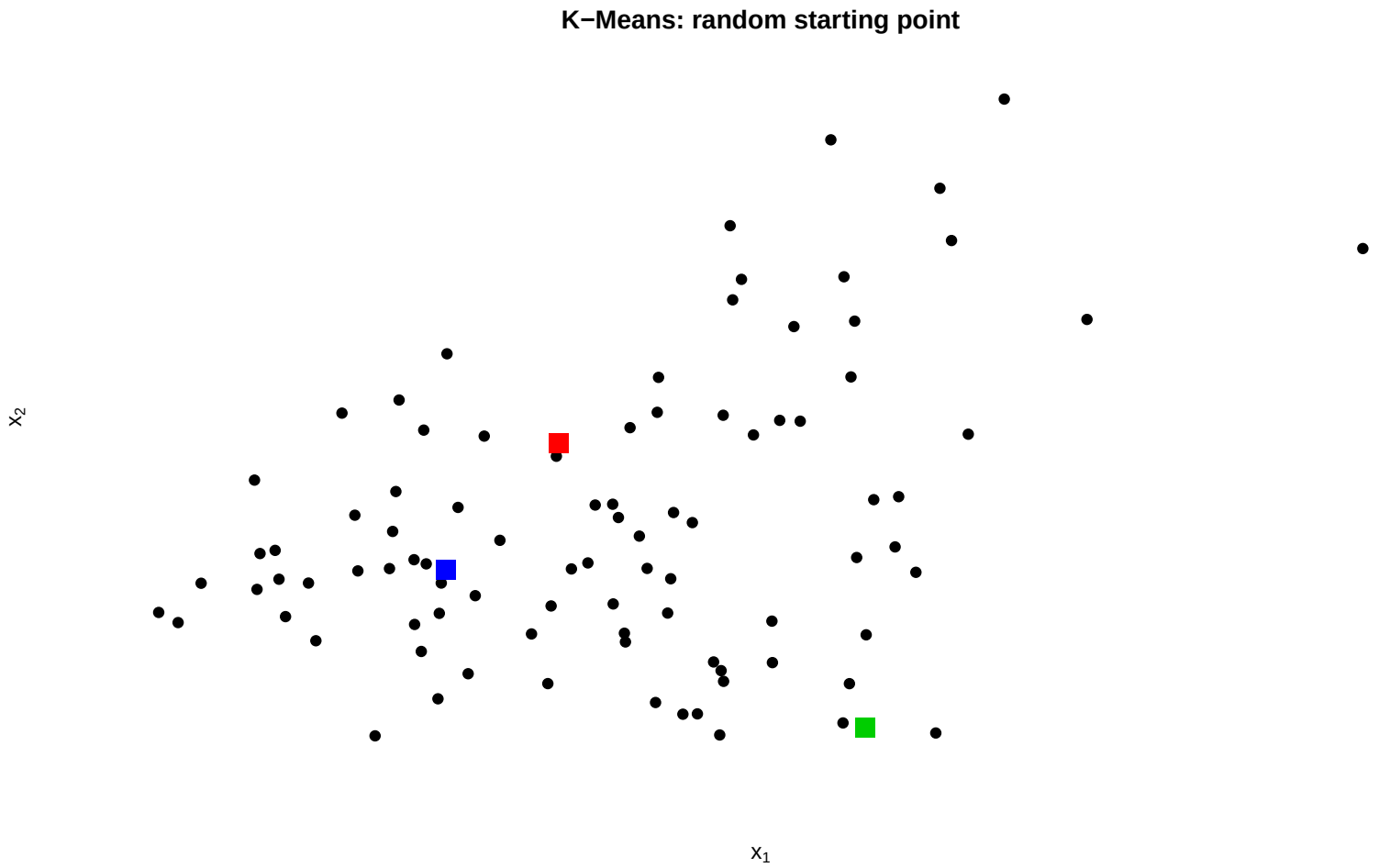
K-Means clustering ($k = 3$)



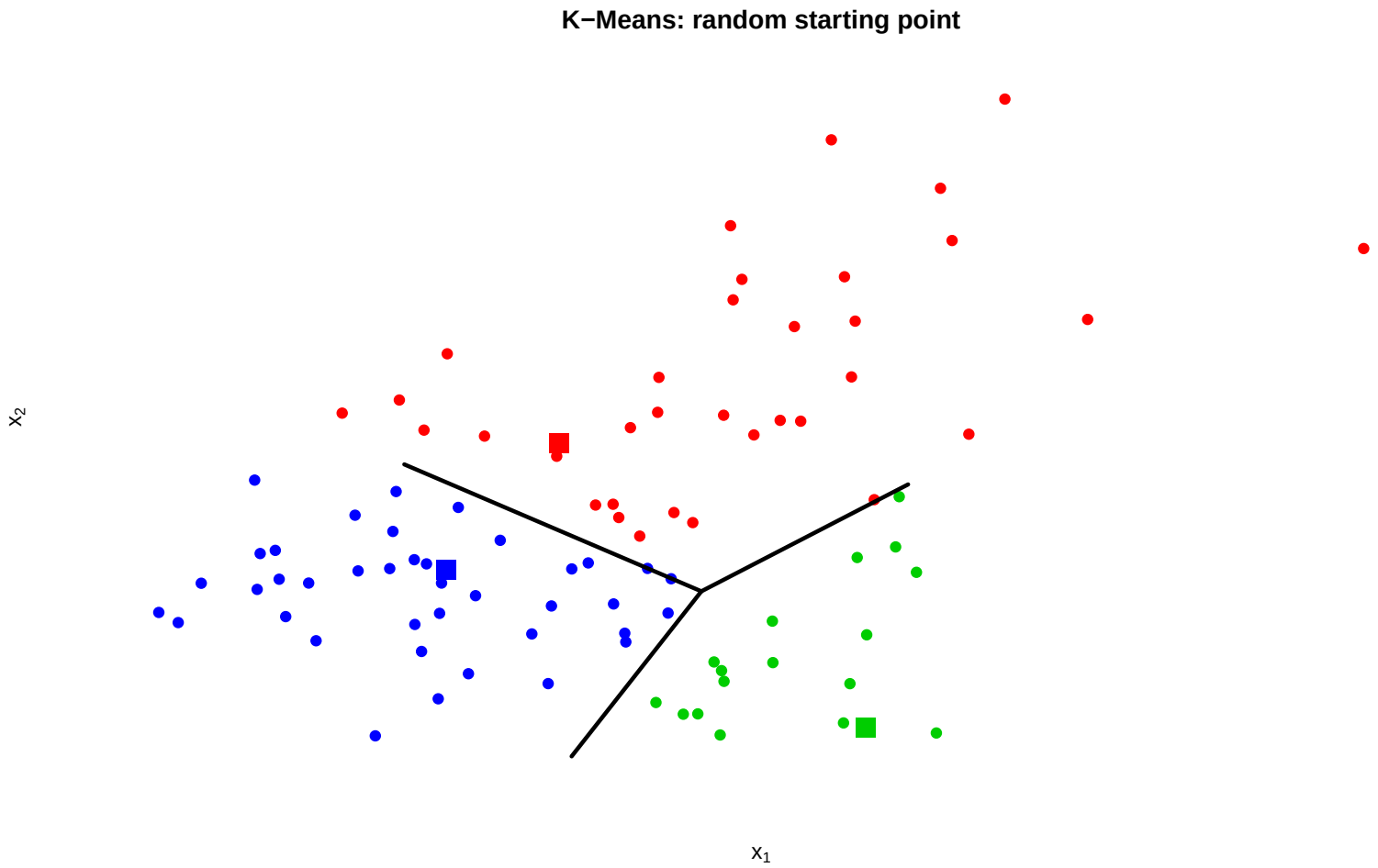
K-Means clustering ($k = 3$)



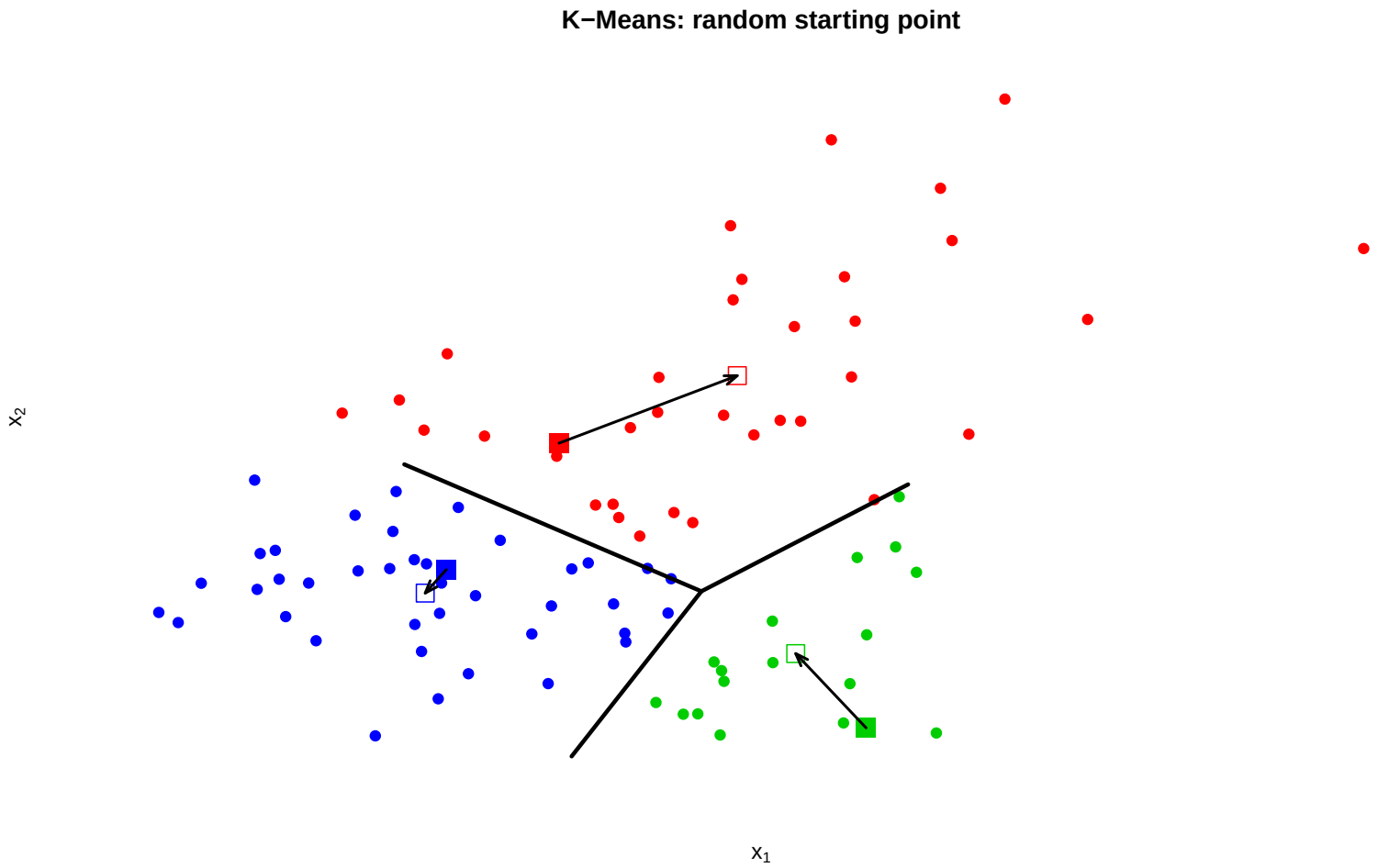
K-Means clustering ($k = 3$)



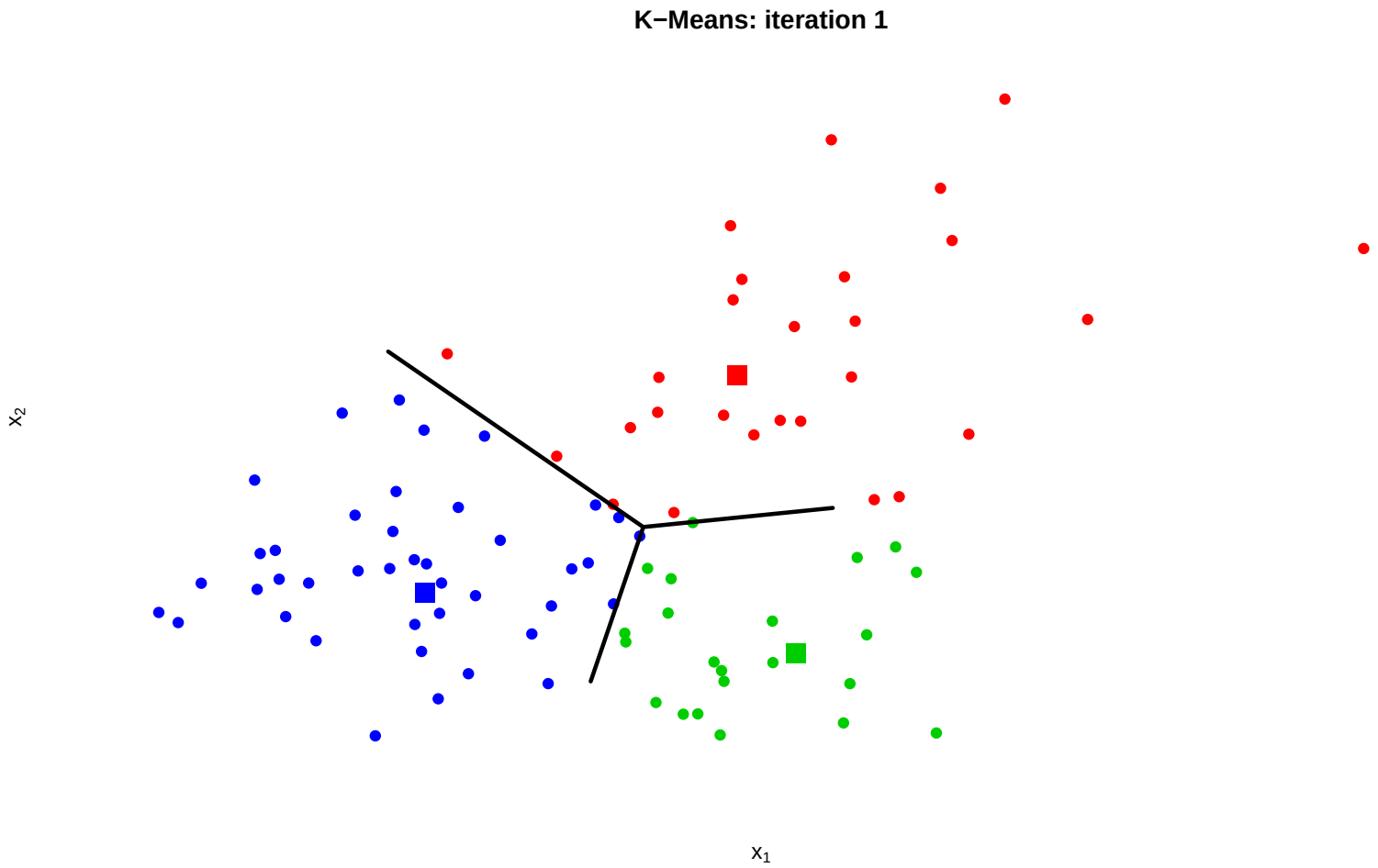
K-Means clustering ($k = 3$)



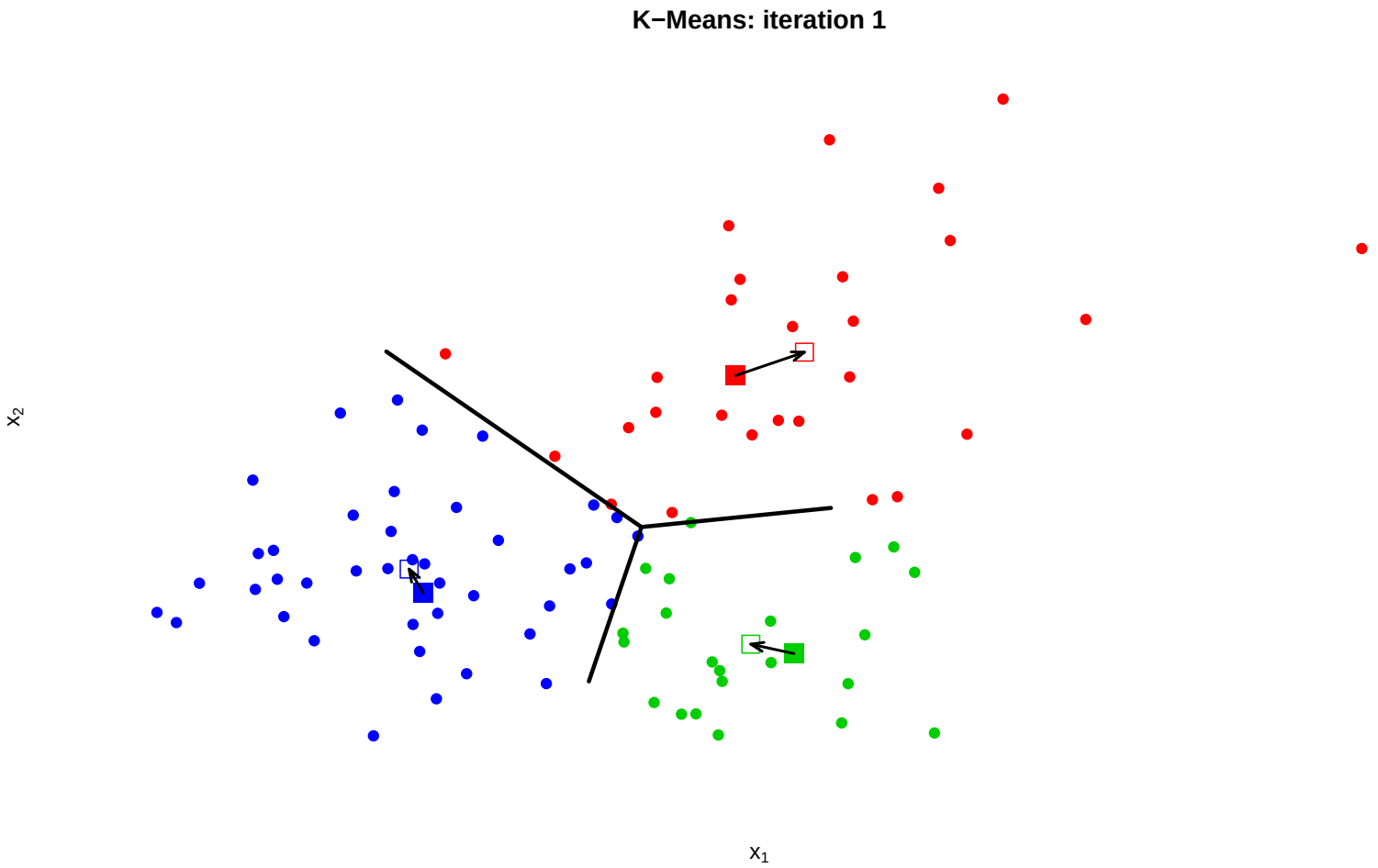
K-Means clustering ($k = 3$)



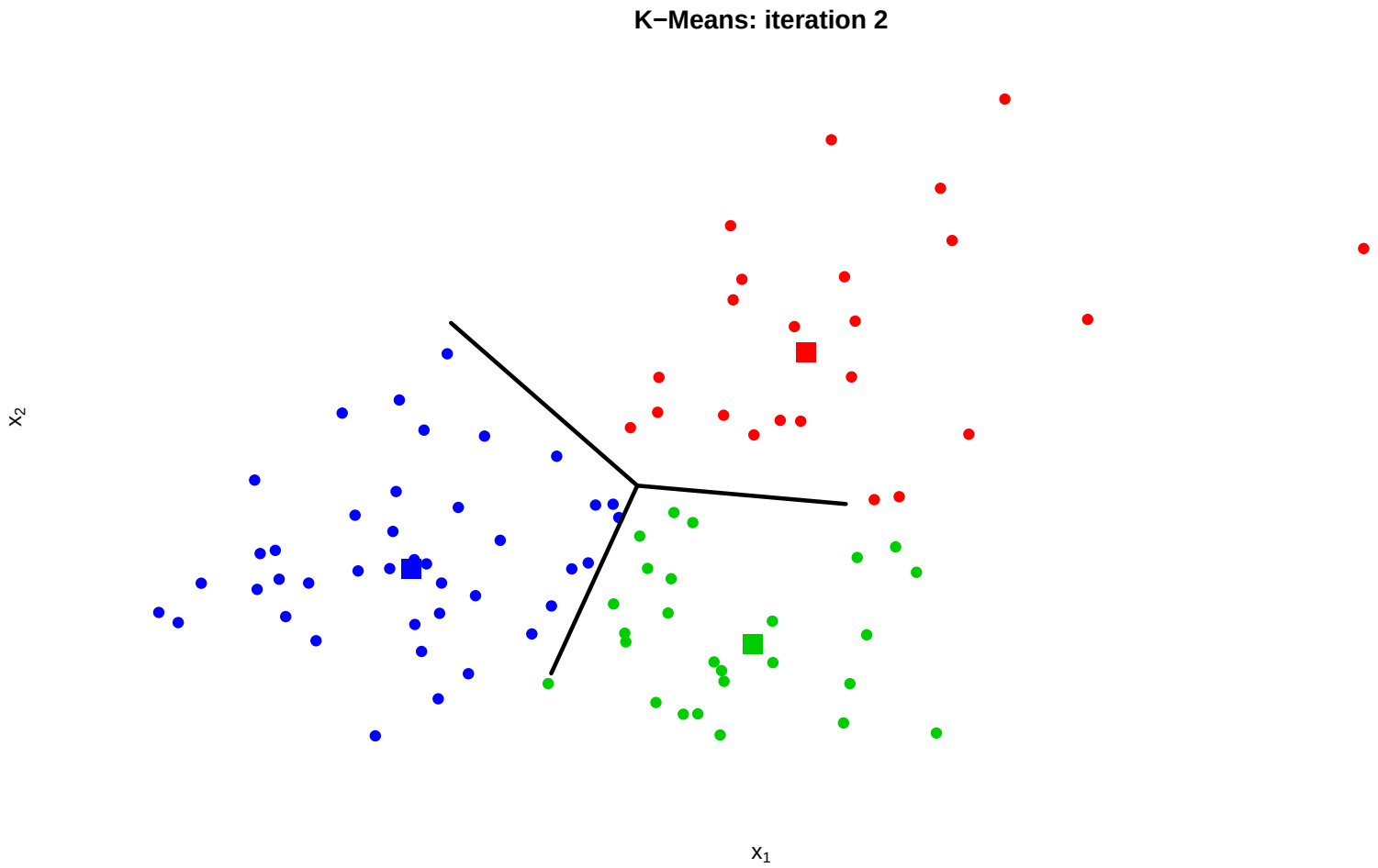
K-Means clustering ($k = 3$)



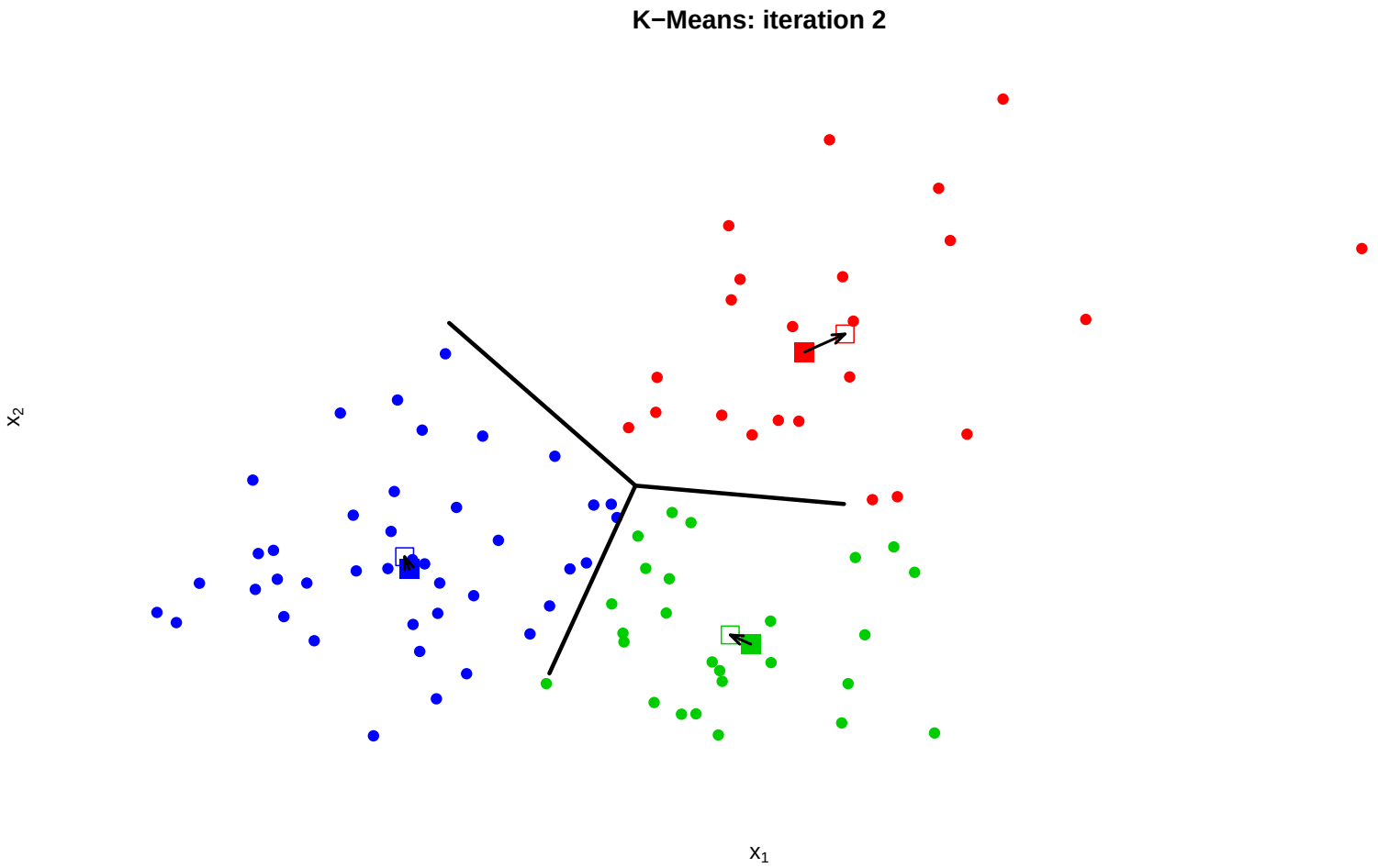
K-Means clustering ($k = 3$)



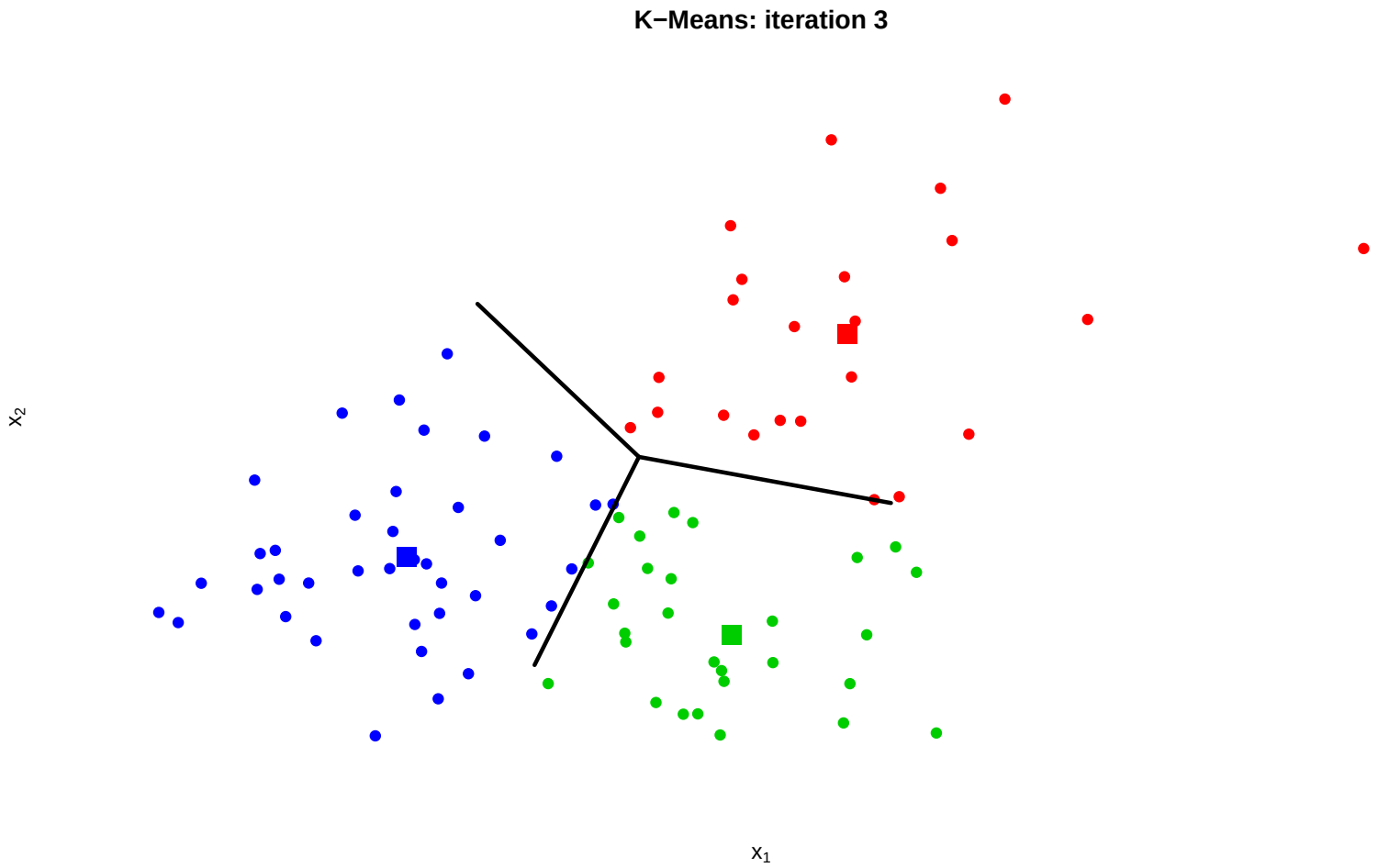
K-Means clustering ($k = 3$)



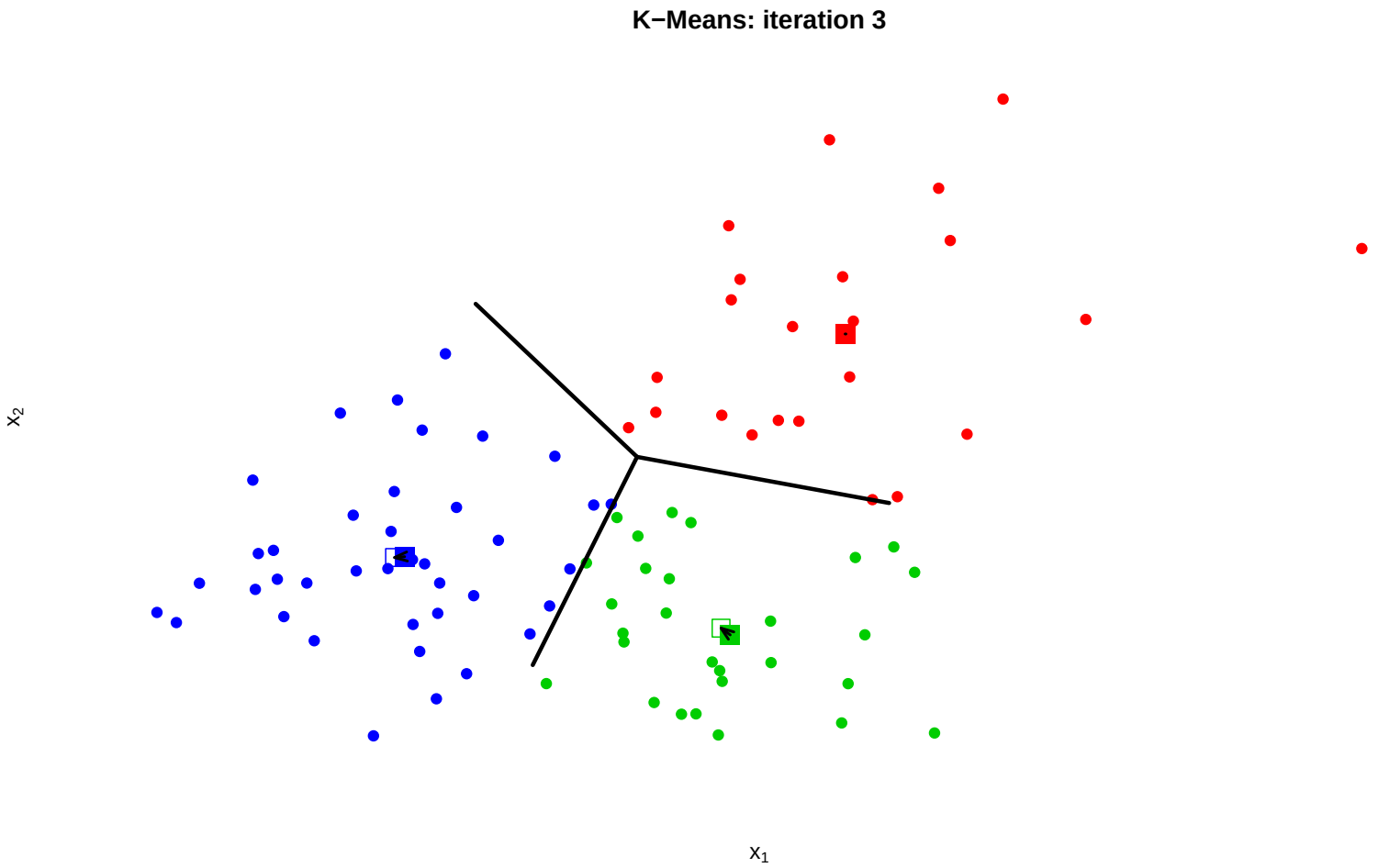
K-Means clustering ($k = 3$)



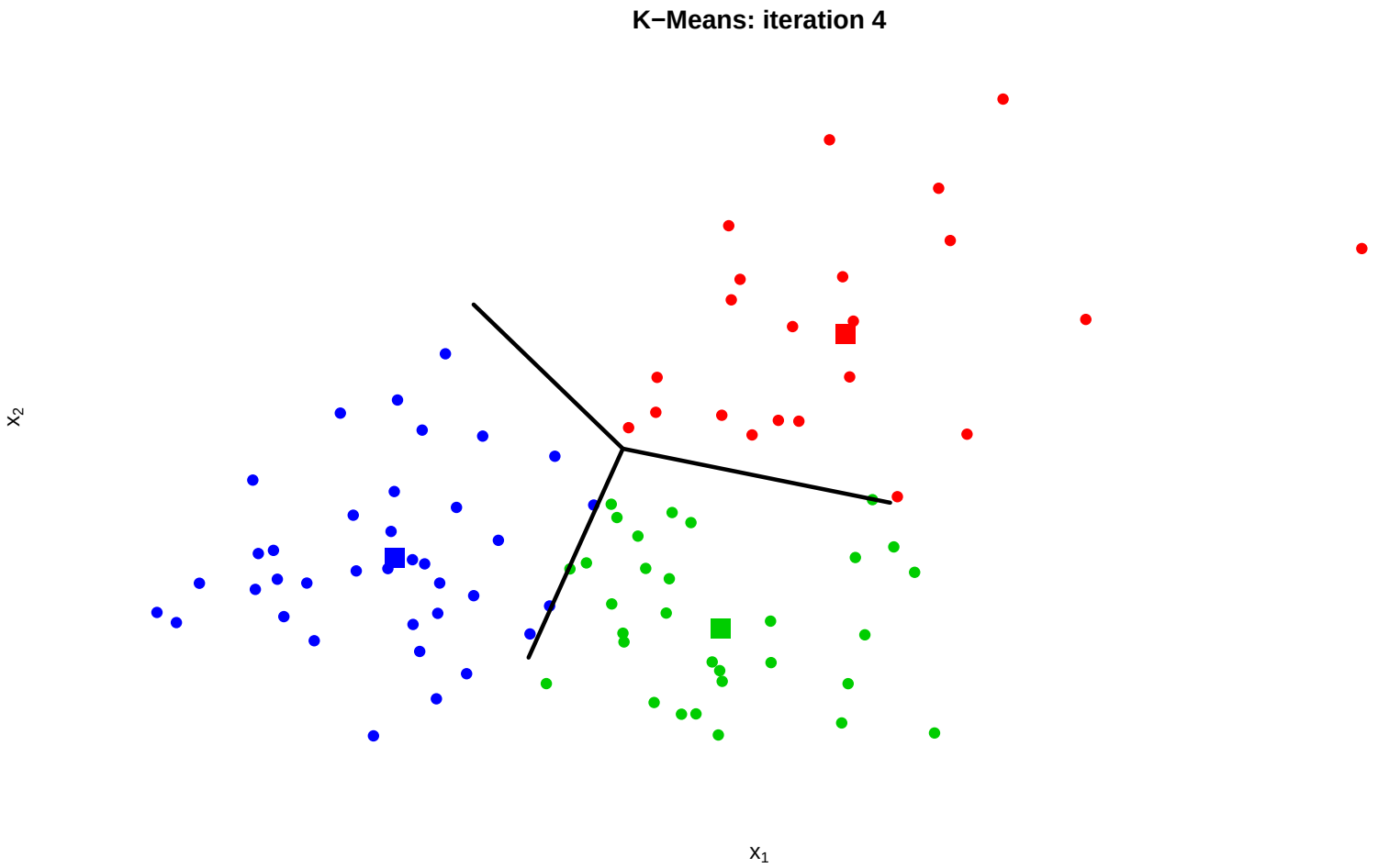
K-Means clustering ($k = 3$)



K-Means clustering ($k = 3$)



K-Means clustering ($k = 3$)



K-Means clustering

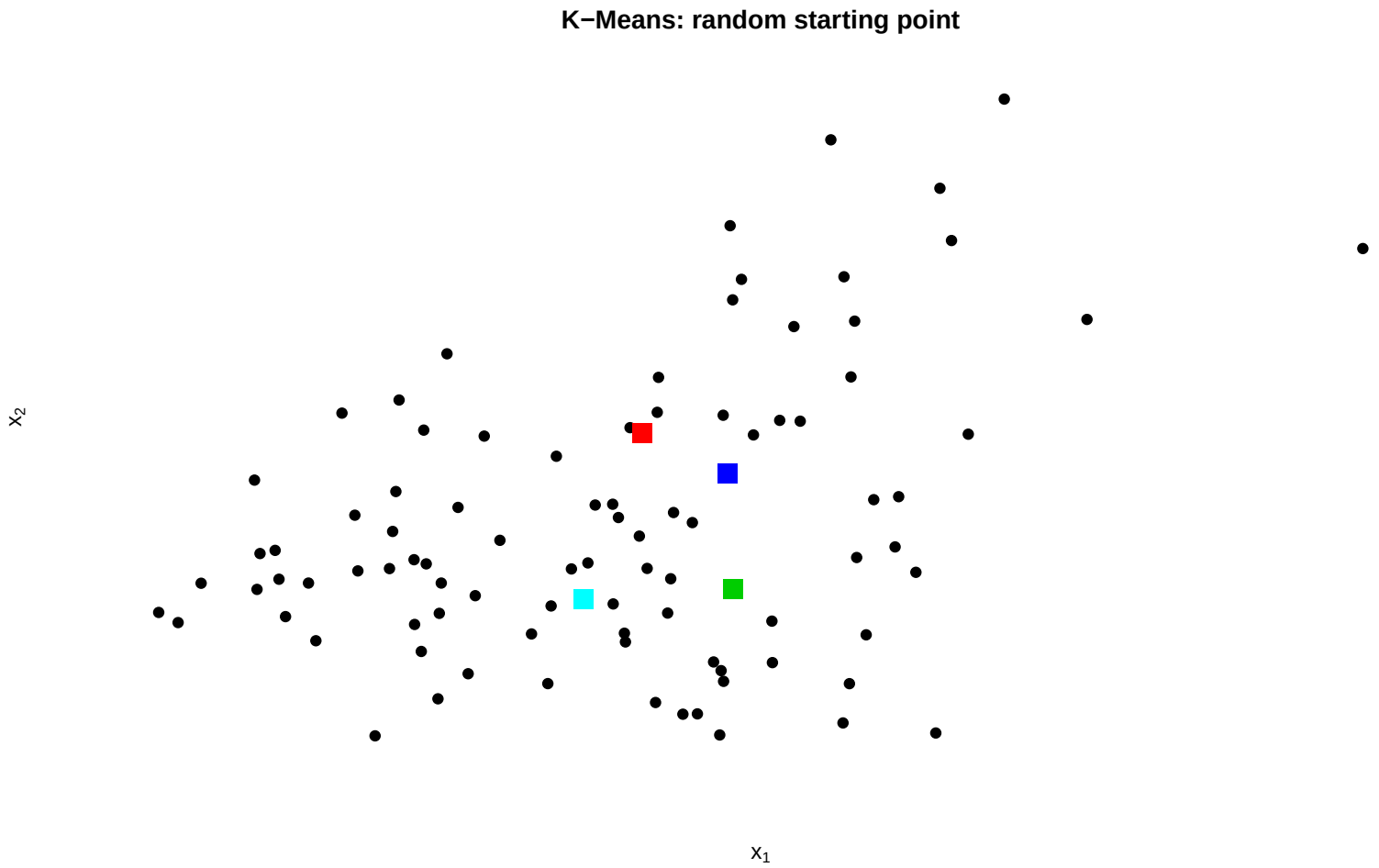
For this example, we end up with the following confusion matrix:

	Known group		
	1	2	3
1	0	1	21
2	1	24	6
3	29	5	3

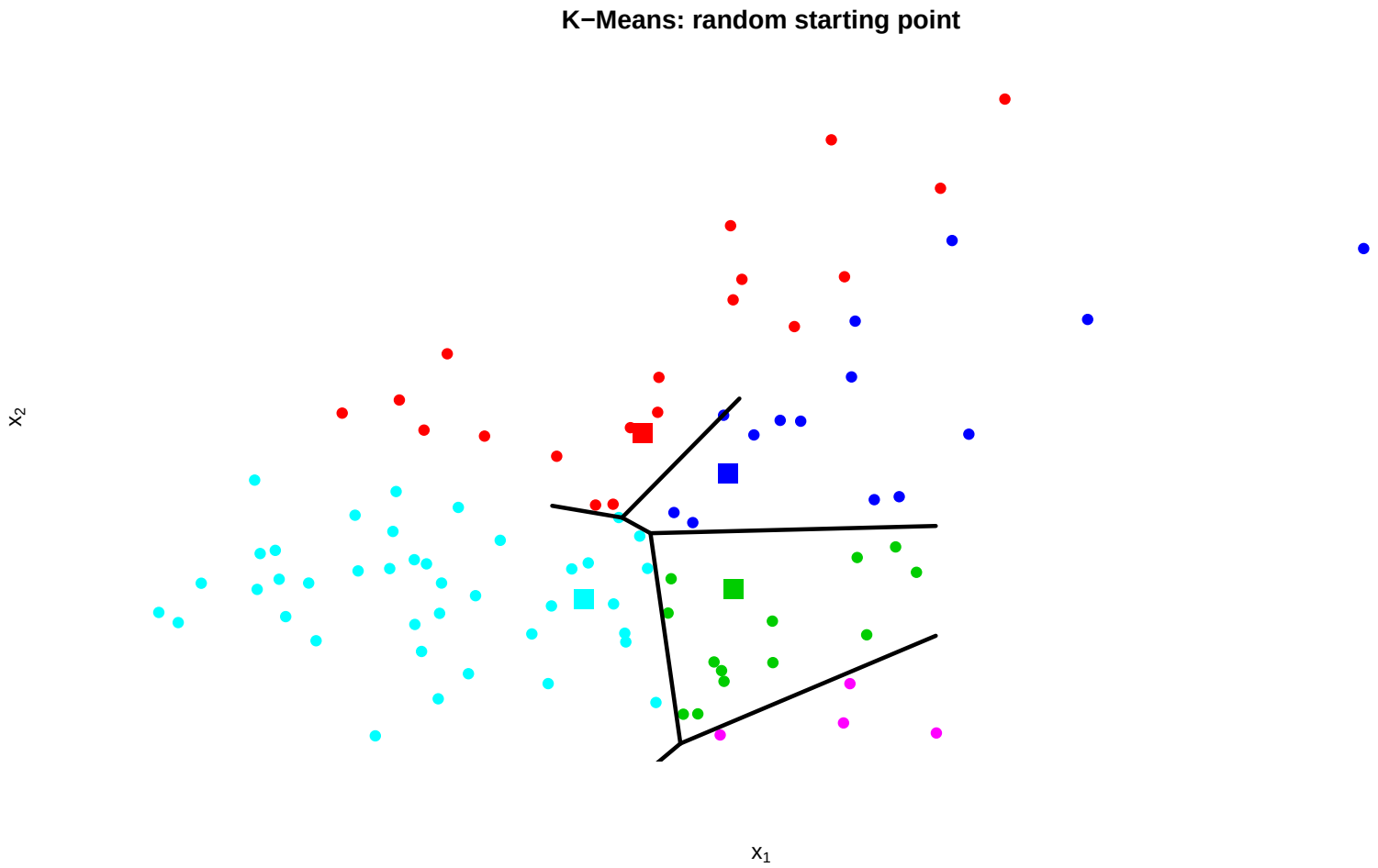
Since group numbers are irrelevant, this is a good recovery of the underlying groups.

Note that we only know the true group because this is fake data: clustering is an **unsupervised learning** method, so the groups are unknown—otherwise we should use, for example, discriminant analysis instead.

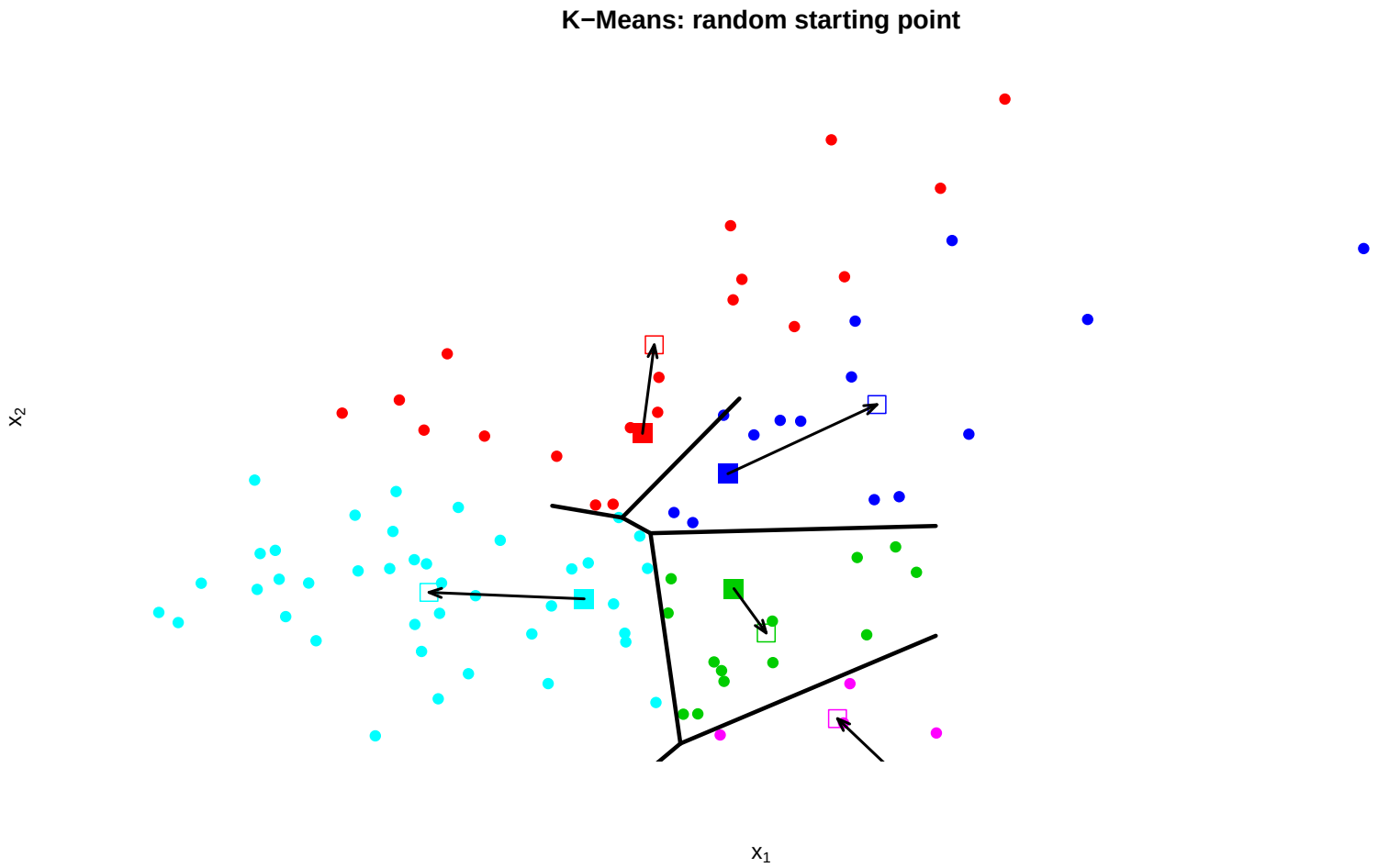
K-Means clustering ($k = 5$)



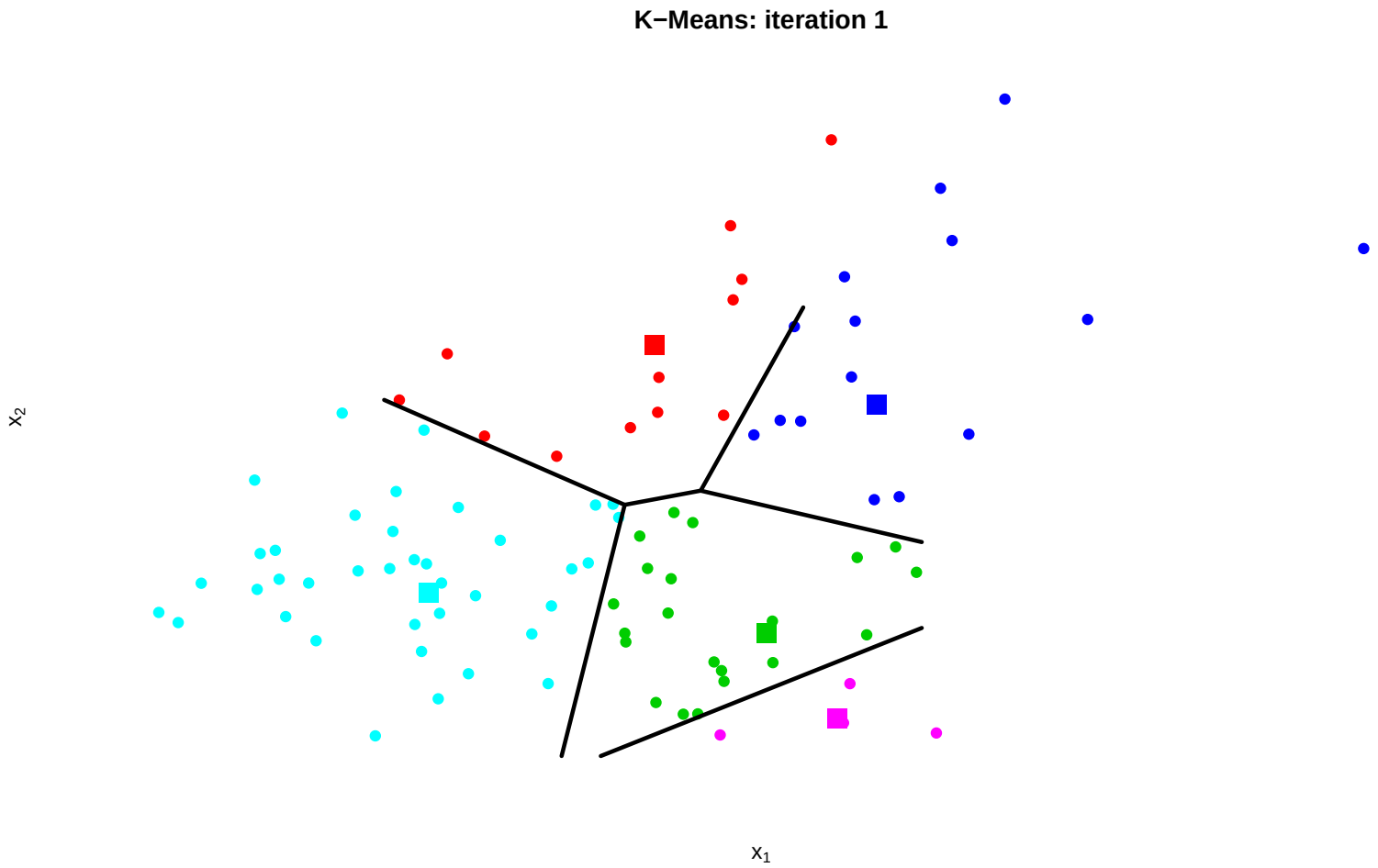
K-Means clustering ($k = 5$)



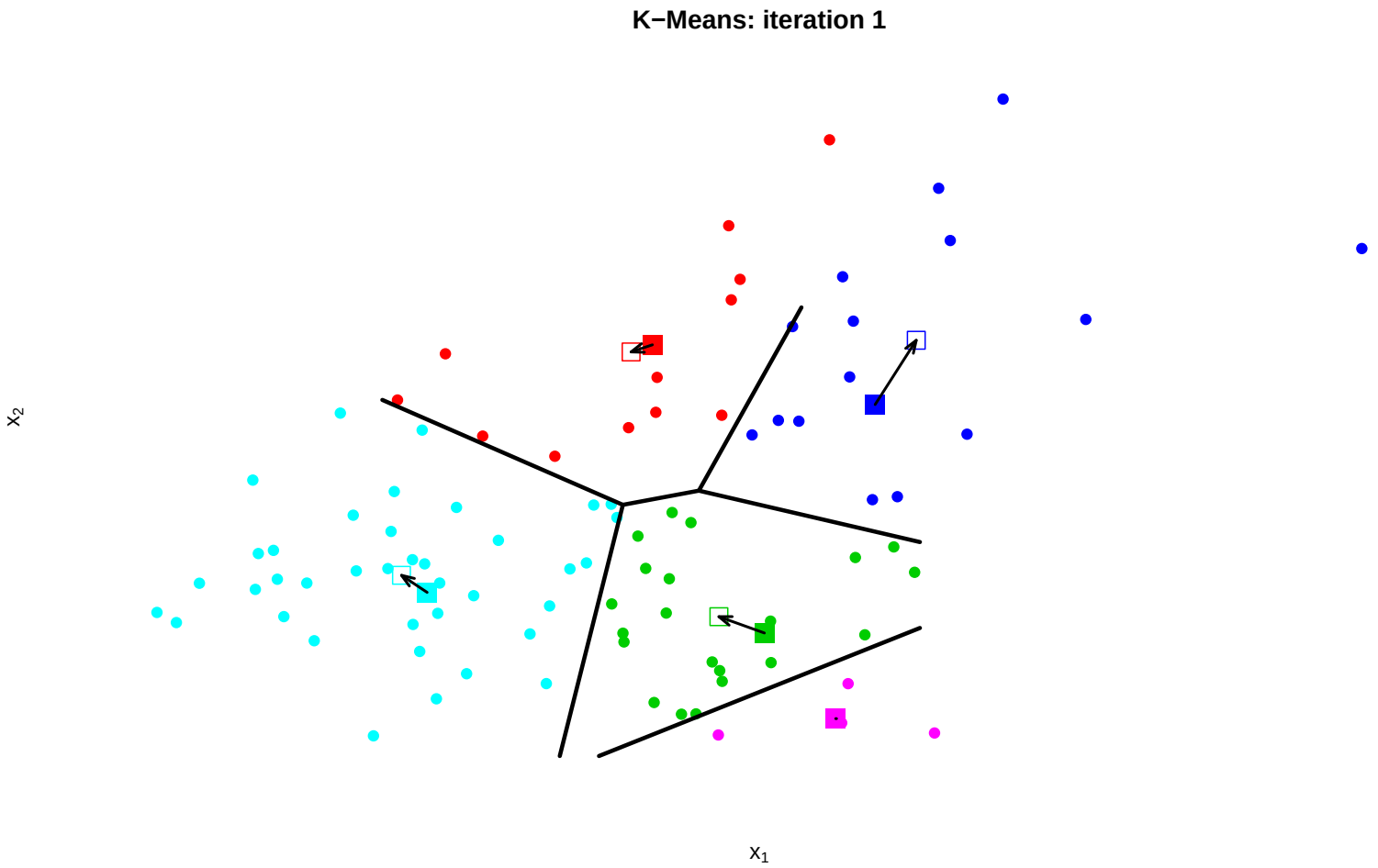
K-Means clustering ($k = 5$)



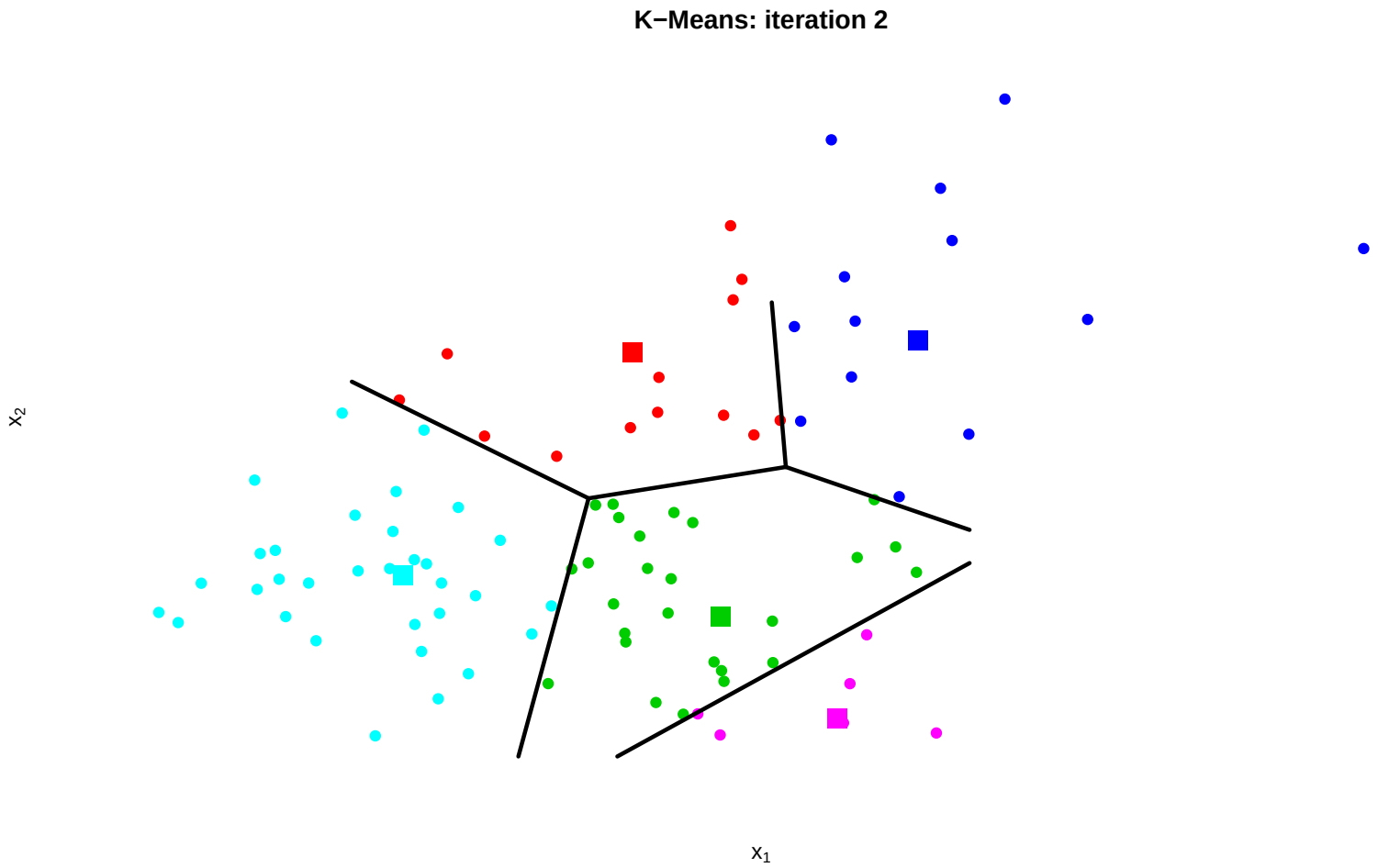
K-Means clustering ($k = 5$)



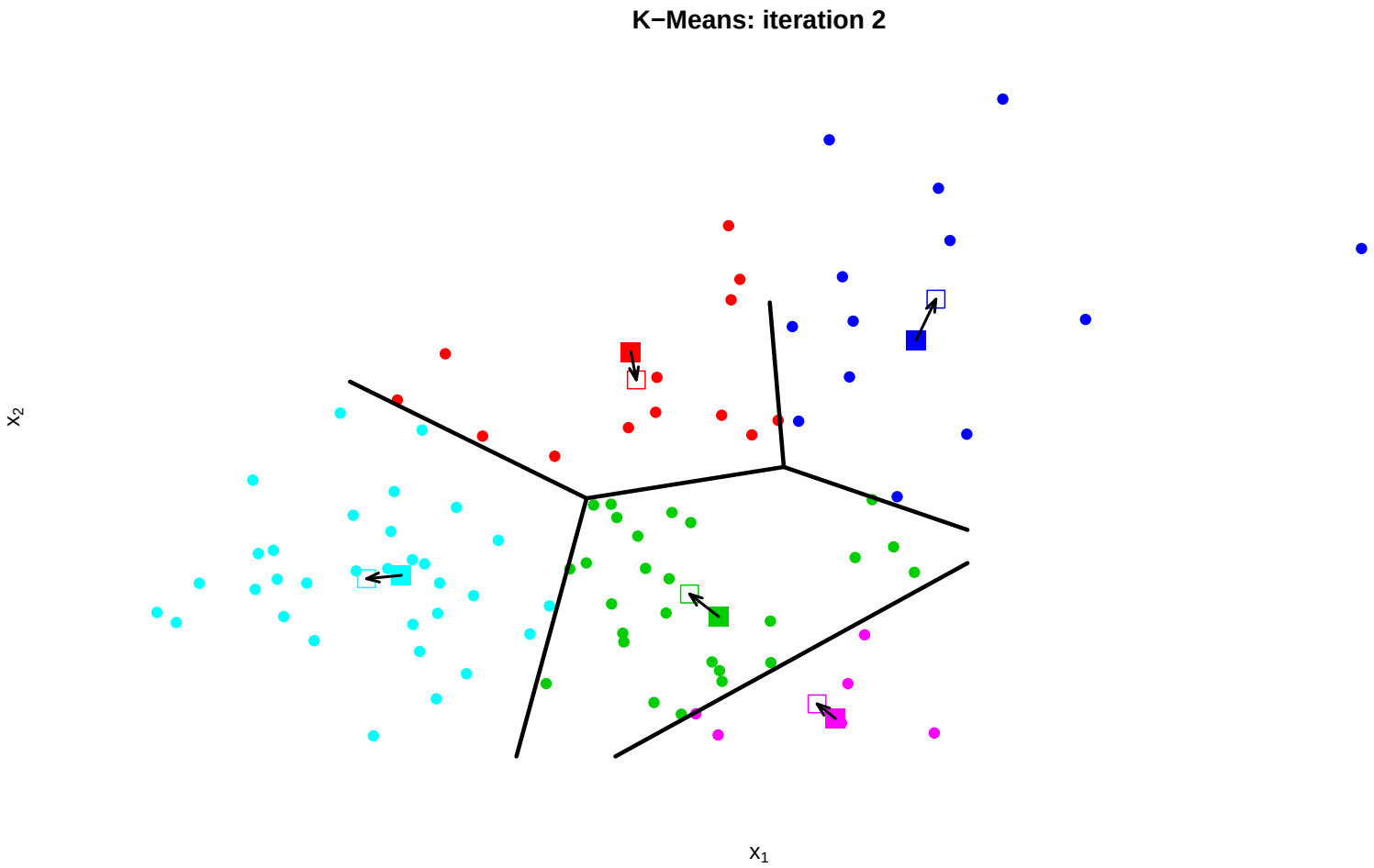
K-Means clustering ($k = 5$)



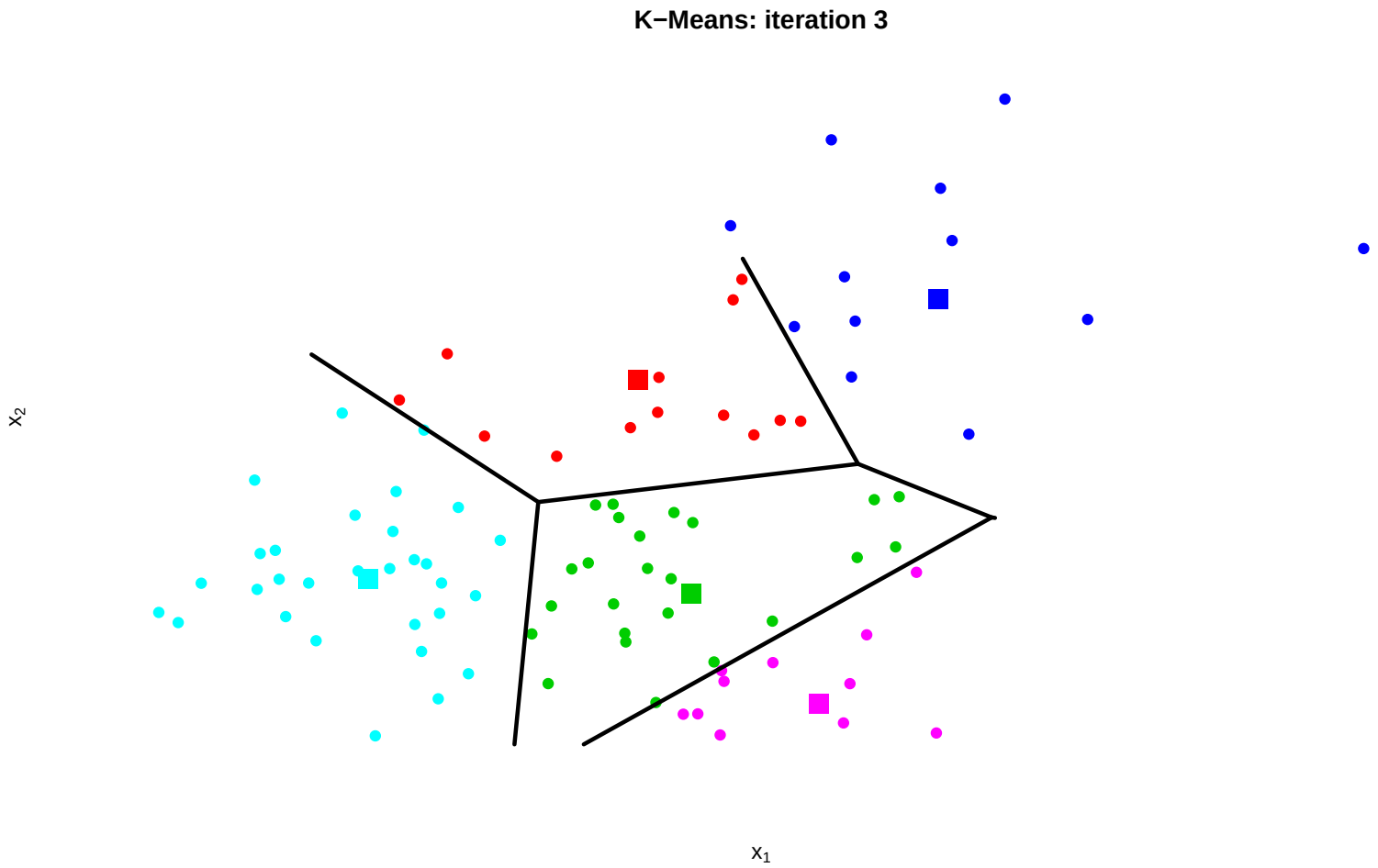
K-Means clustering ($k = 5$)



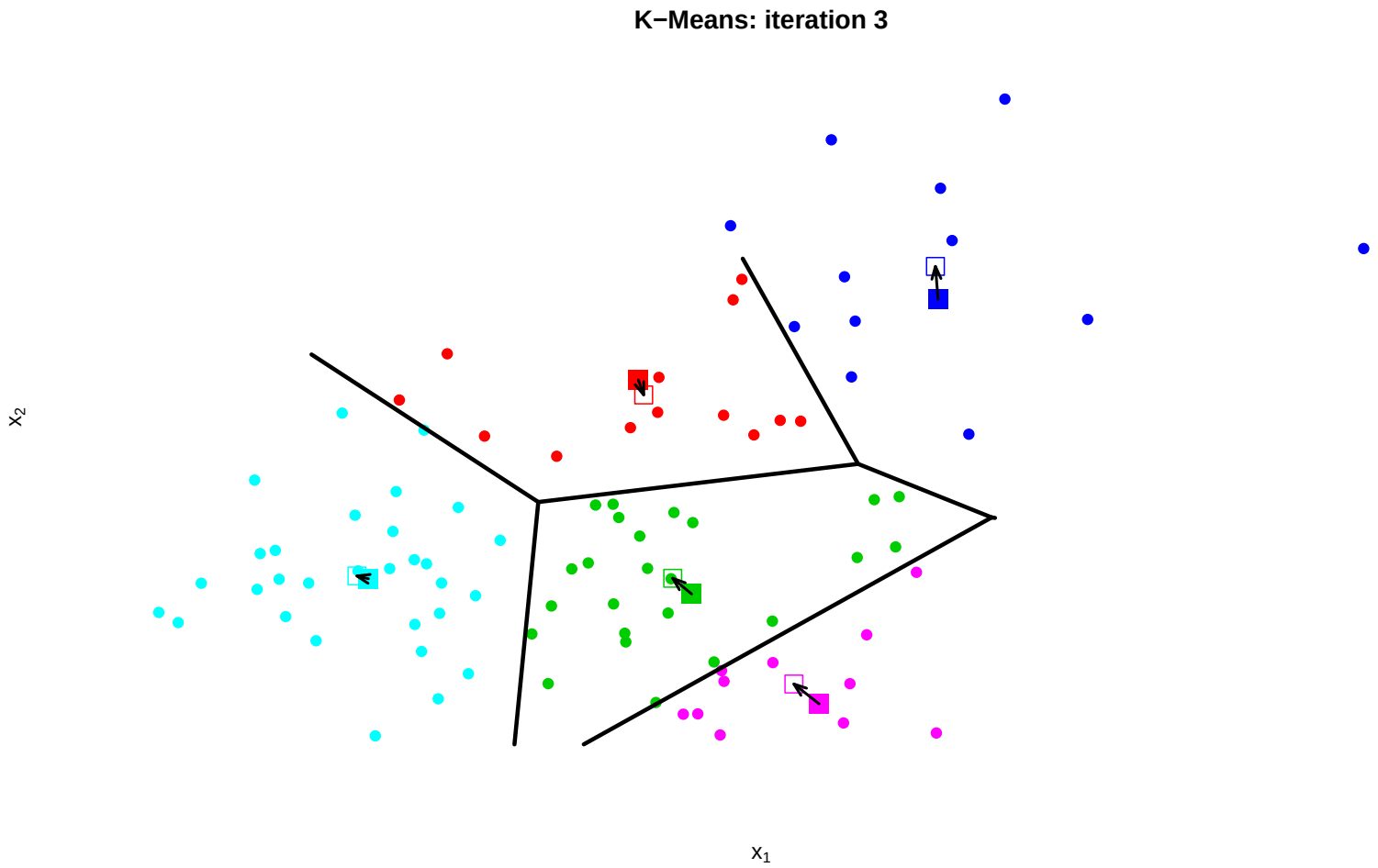
K-Means clustering ($k = 5$)



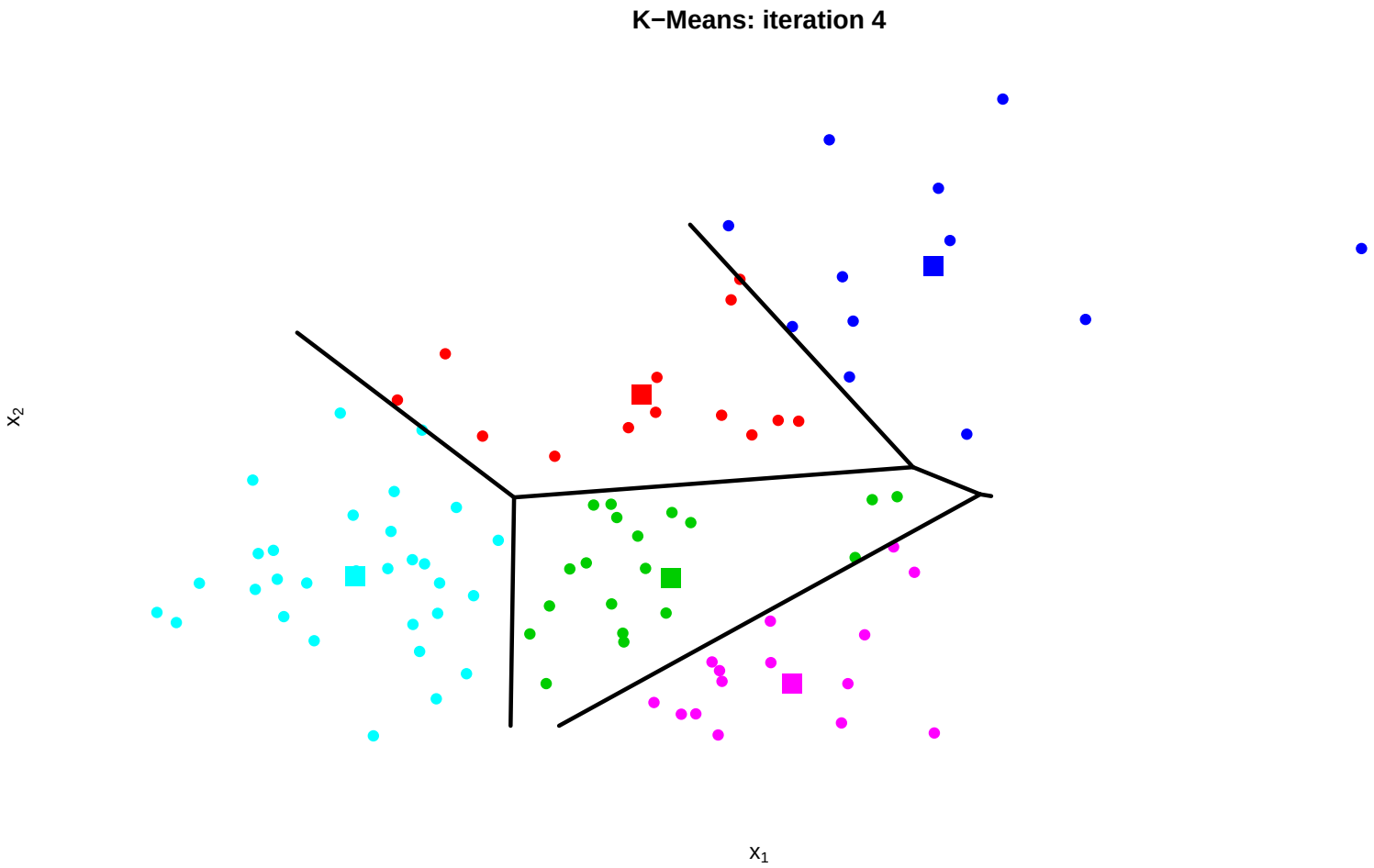
K-Means clustering ($k = 5$)



K-Means clustering ($k = 5$)



K-Means clustering ($k = 5$)

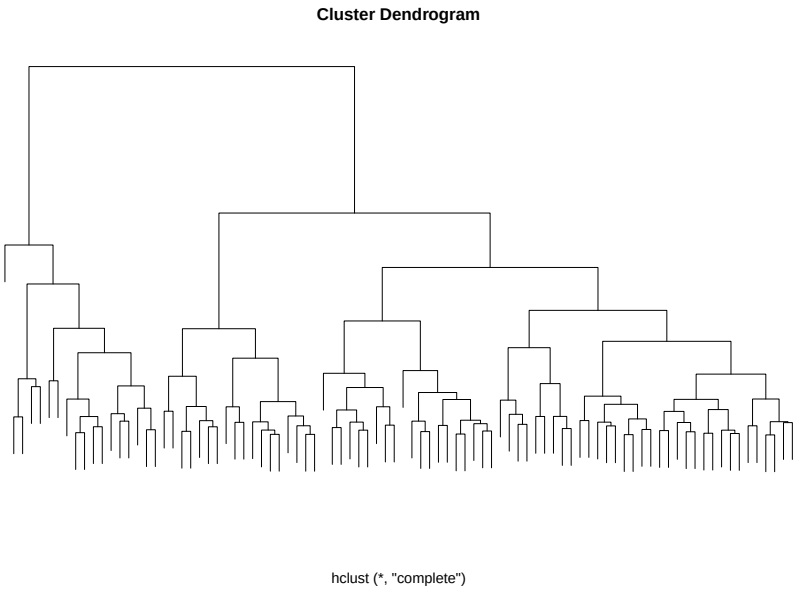


Outline

Hierarchical clustering

In K-Means clustering we have to select a number of k clusters in advance.

In **hierarchical clustering**, we build clusters gradually, building a tree-like structure where each observation is its own cluster, then combined with similar observations, all the way up until all observations are in one cluster.



The plot to visualise this hierarchy is called a **dendrogram**.

Distance between clusters

The distance between two clusters of observations can be defined in different ways:

Single linkage: the smallest distance between any two points in the two clusters (also called **nearest neighbor**).

Complete linkage: the largest distance between any two points in the two clusters (also called **farthest neighbor**).

Group average: average distance across all pairs of points in the two clusters.

Each has advantages and disadvantages. If the data are strongly clustered in particular groups, the results will be roughly the same.

(Hastie, Tibshirani and Friedman, 2009, 523–525)

Outline

Distance measures

All of the above analyses make use of the **distance** between points.

This distance can be measured in various ways, which we might call **dissimilarity measures**.

The obvious, default one is to use **Euclidean (squared) distance** in the high-dimensional space defined by the underlying variables.

Alternatives include absolute distances, correlation measures, or distance measures designed for categorical variables.

The choice of distance measure will have more impact on the results than the choice of algorithm.

(Hastie, Tibshirani and Friedman, 2009, 503–506)

Outline

Project 3: State of the Union speeches

Instead of statistical data about countries or individuals in a survey, here we make use of statistical analysis to understand a set of texts.

After cleaning the data (removing stop words, upper case, stemming, etc.) we can produce a data set where each observation is a document, each variable a word (term), and each observation the number of times a term appears in a document.

	achiev	act	action	ago	alreadi	also	america
Truman 1945	2	0	0	0	1	1	11
Truman 1946	15	38	21	3	25	43	2
Truman 1947	3	4	5	1	2	7	1
Truman 1948	11	6	4	2	1	9	0
Truman 1949	5	6	3	1	0	4	1
Truman 1950	18	2	4	4	3	9	0

Speeches and clustering

Much of politics and international relations consists of text (speeches, debates, manifestos, treaties, laws, etc.) and often there are too many to study them all in depth (e.g. all parliamentary debates in a century). Statistical analysis can help us to see trends, group by topic, estimate ideological positions, etc.

So far:

- We can use the word frequency data to find trends over time in terms of topics and word usage.
- We can use cluster analysis on the documents to see which speeches were more similar and which were different.
- We can use cluster analysis on the terms to see which terms belong to similar speeches and therefore probably align to similar topics.

Preprocessing

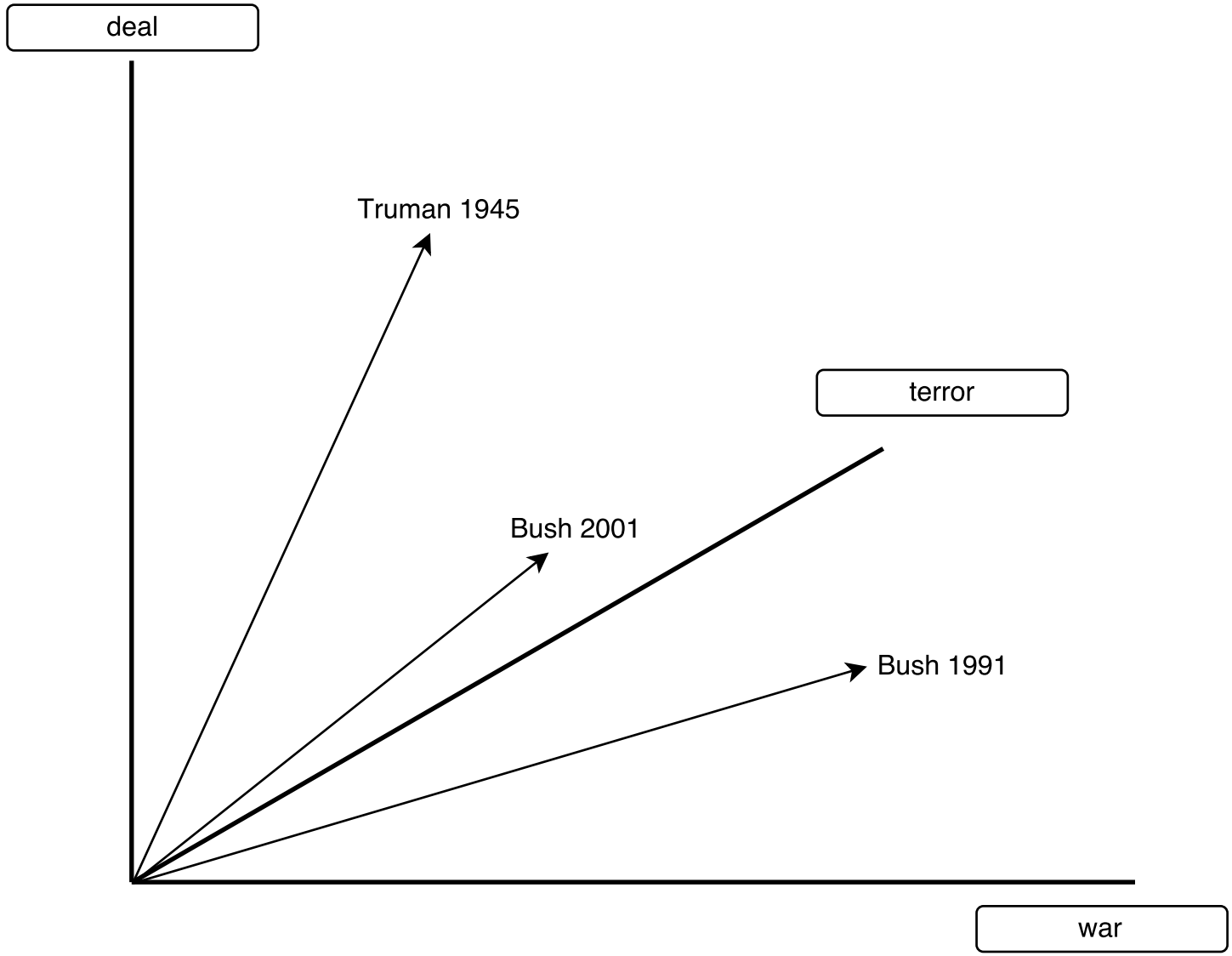
Stemming is related to **lemmatization** and concerns the removal of all endings of words, so that words such as "walking", "walk", "walks" all become the same word, "walk".

Other transformations include the removal of **stop words**, changing all text to **lower case**, removing **numerals**.

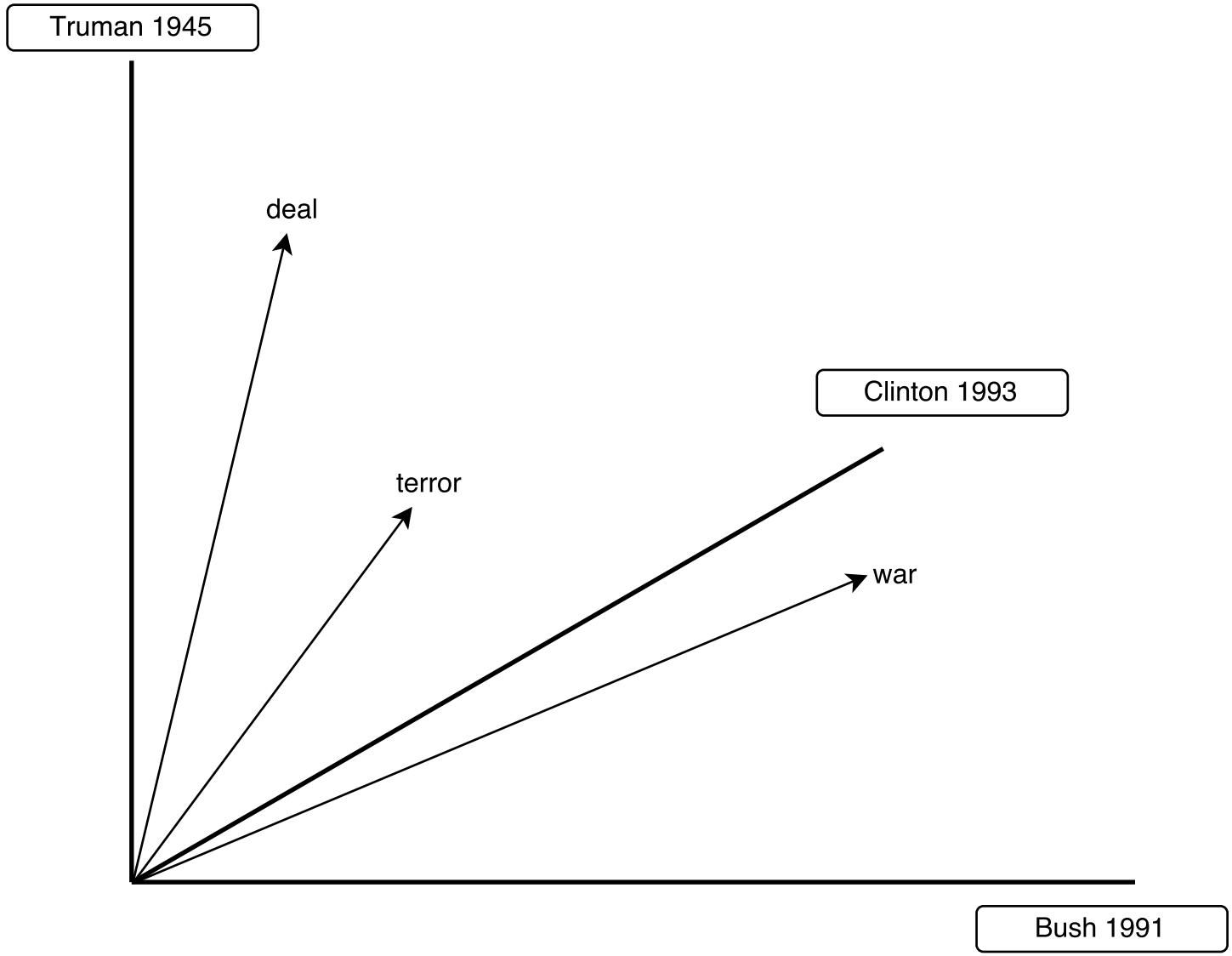
The final transformation is the removal of **sparse terms** that rarely occur in the texts.

The lab will show all the relevant code in R, made easy by using the `tm` package.

Text in high-dimensional space



Text in high-dimensional space



Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition ed. New York: Springer.