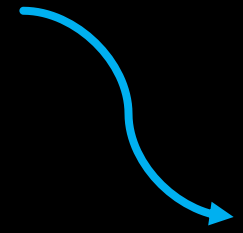
LECTURE 7

CEIC6789: NOTES

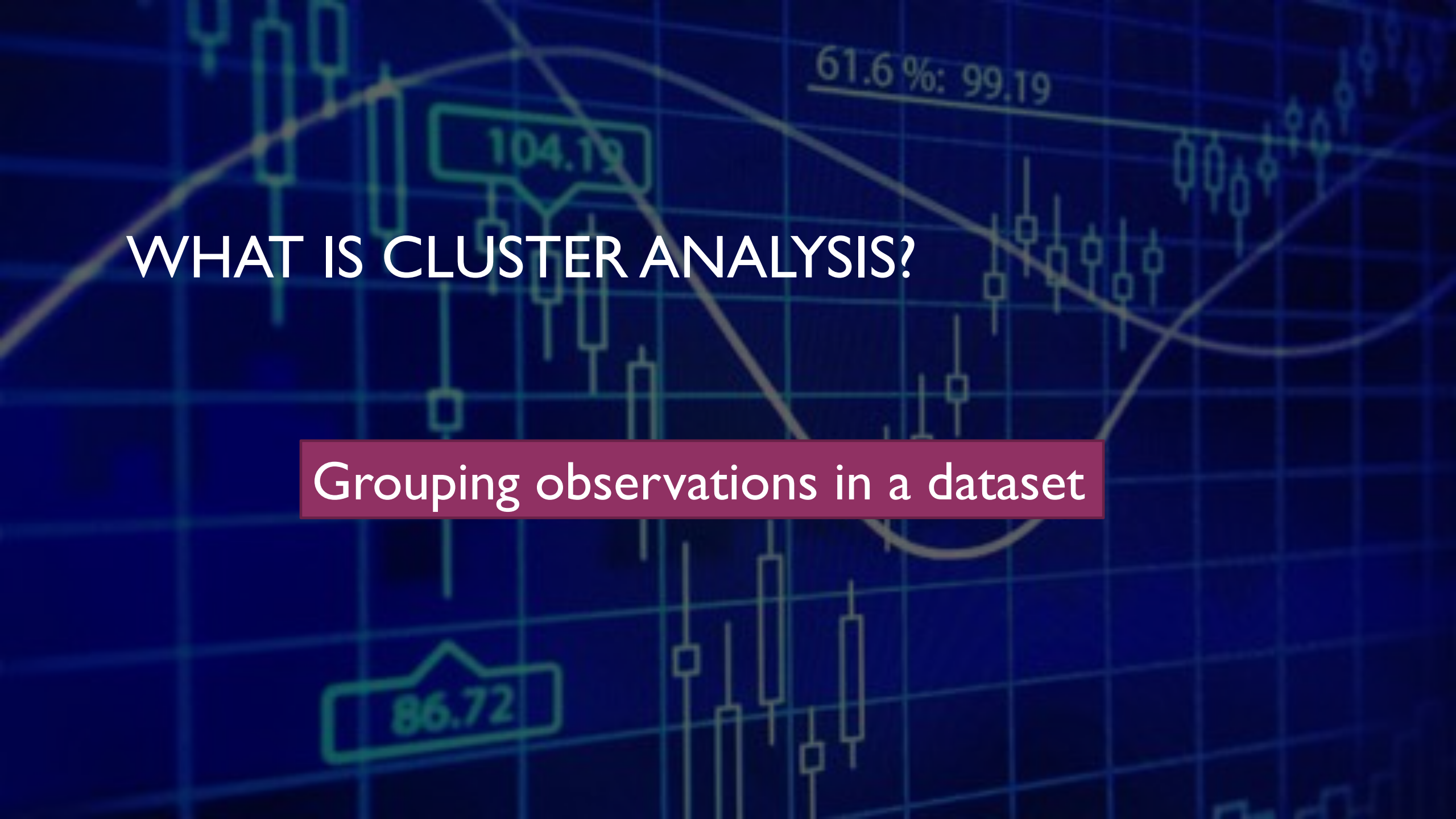
Supervised learning
(Both targets and features known)

- Simple linear regression
- Multiple linear regression
- Logistic regression

CLUSTER ANALYSIS



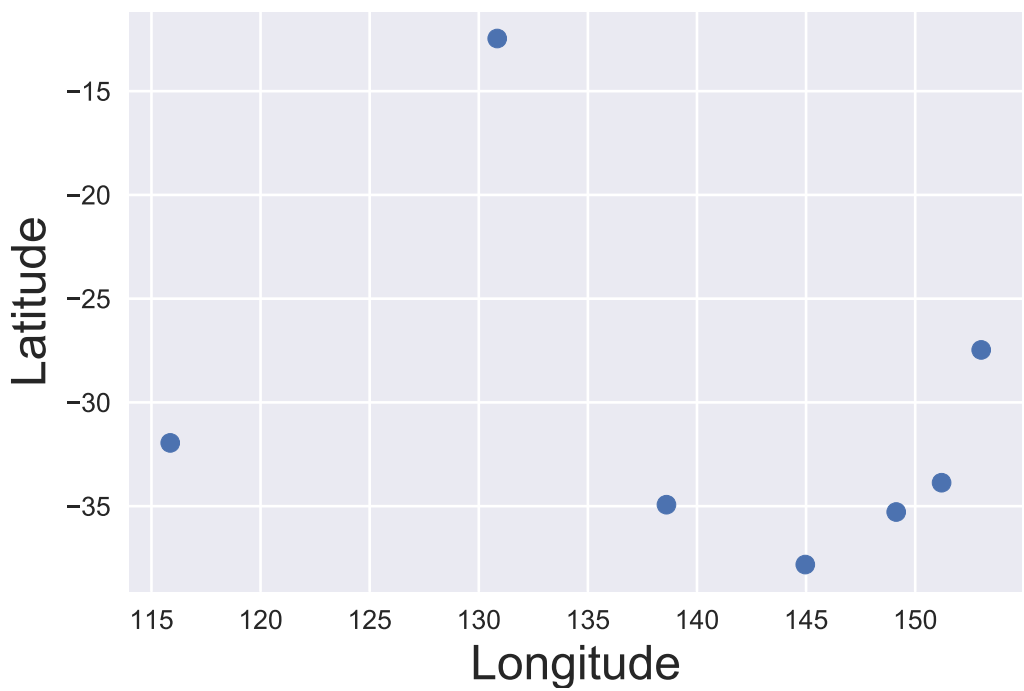
Unsupervised learning (Only features known)

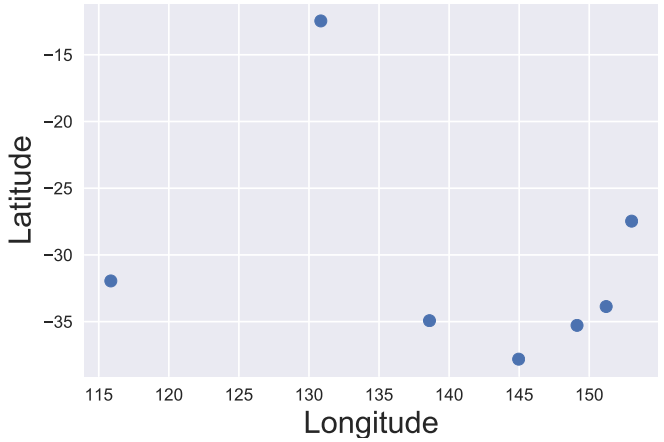


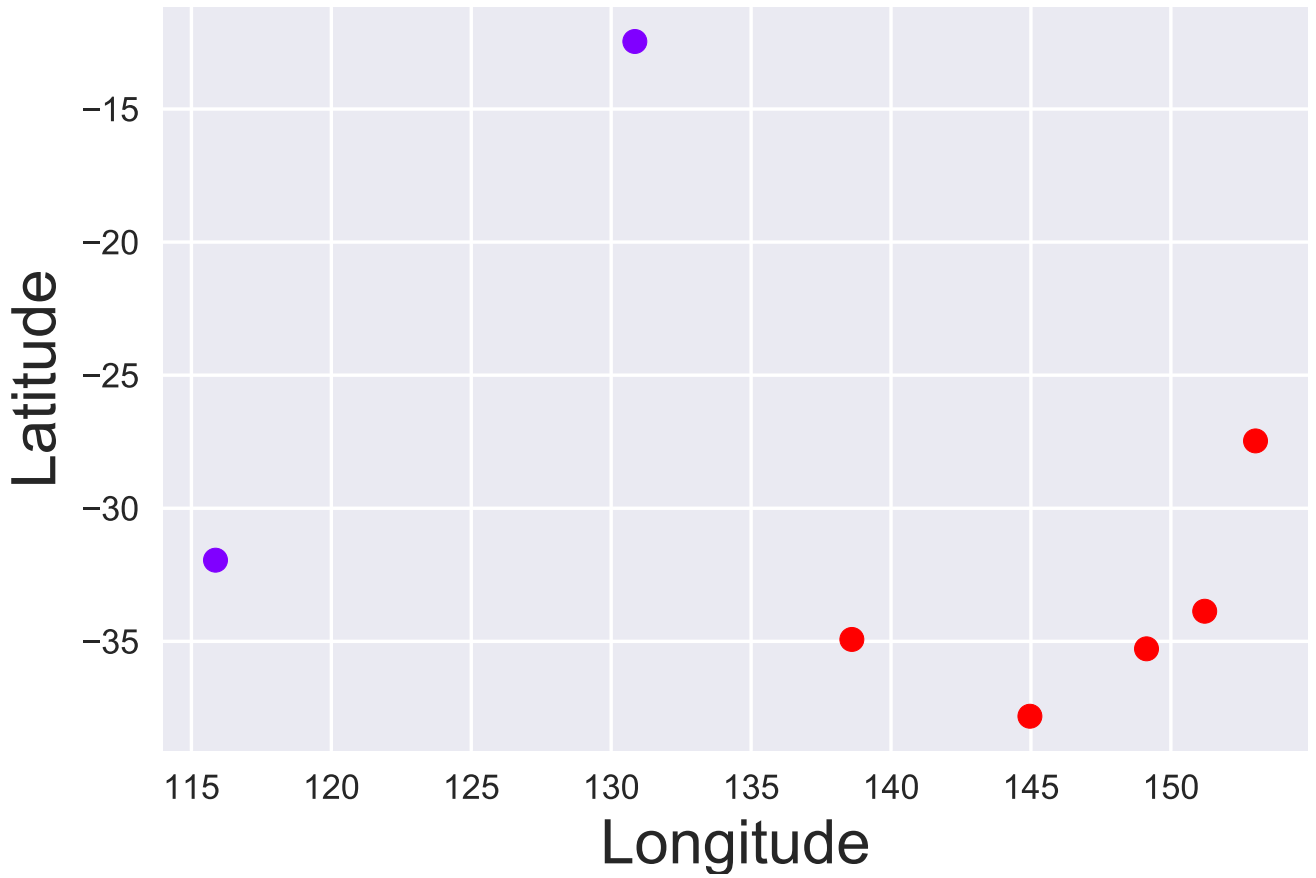
WHAT IS CLUSTER ANALYSIS?

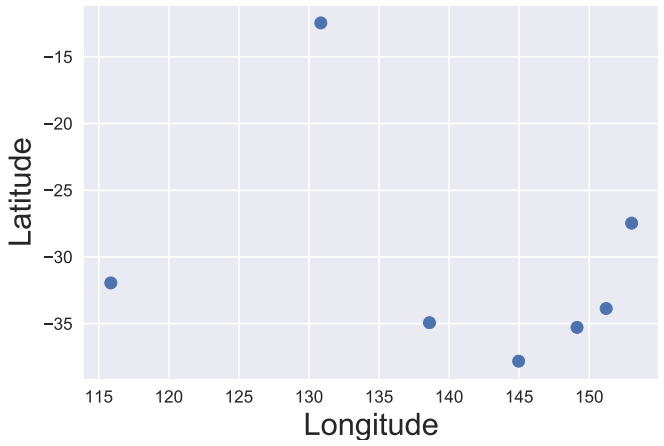
Grouping observations in a dataset

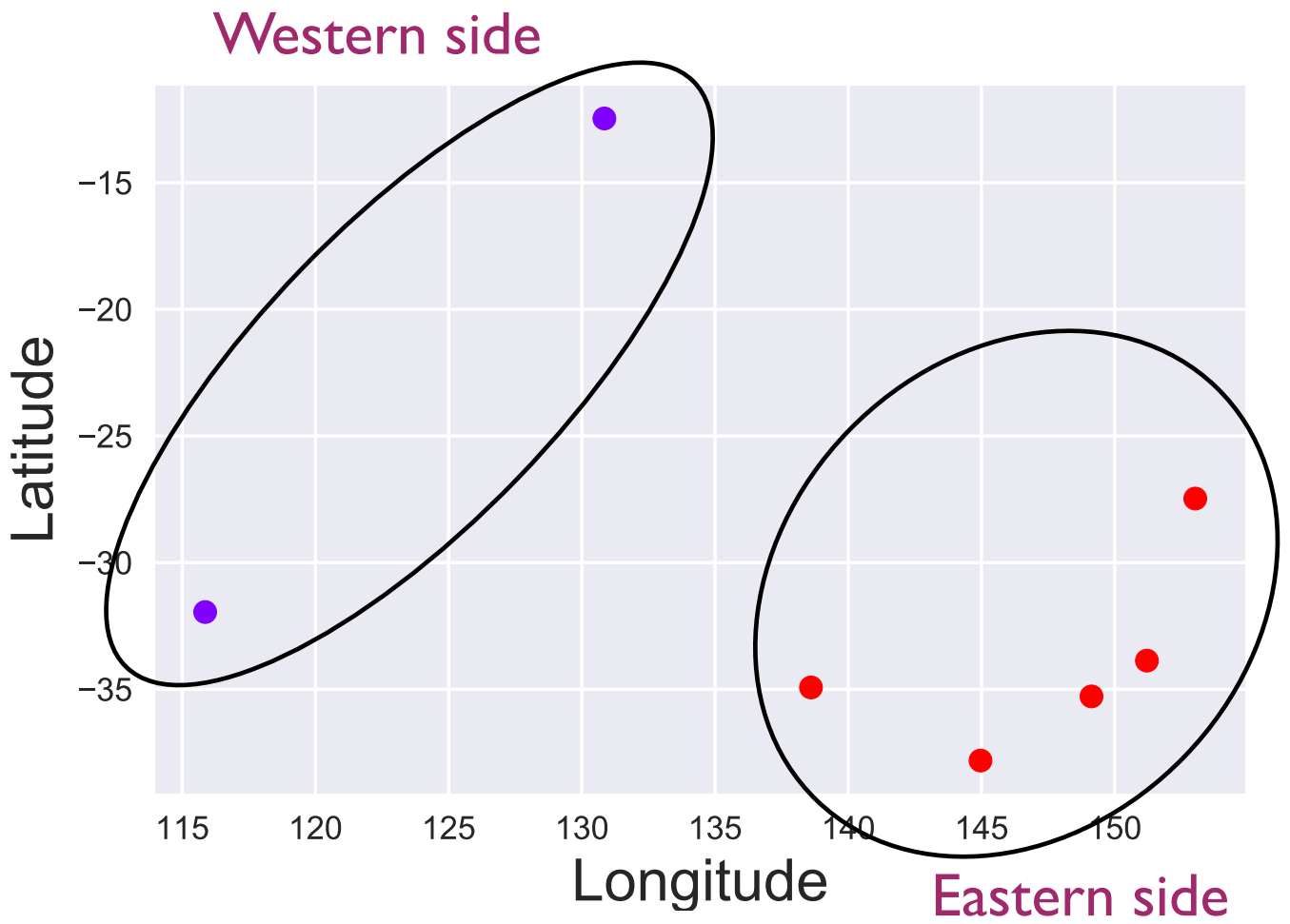
City	Latitude	Longitude
Sydney		
Melbourne		
Perth		
Darwin		
Brisbane		
Canberra		
Adelaide		

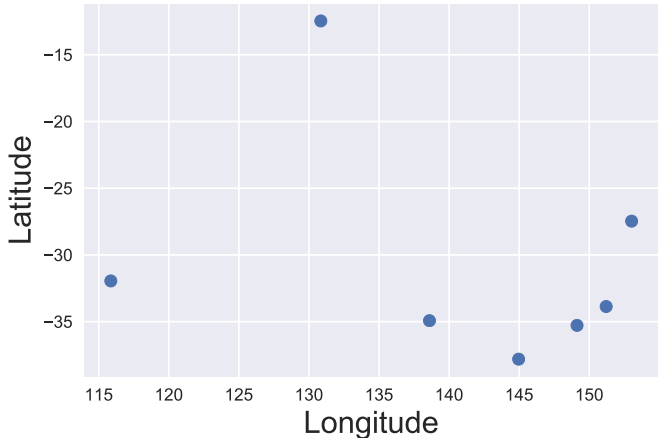


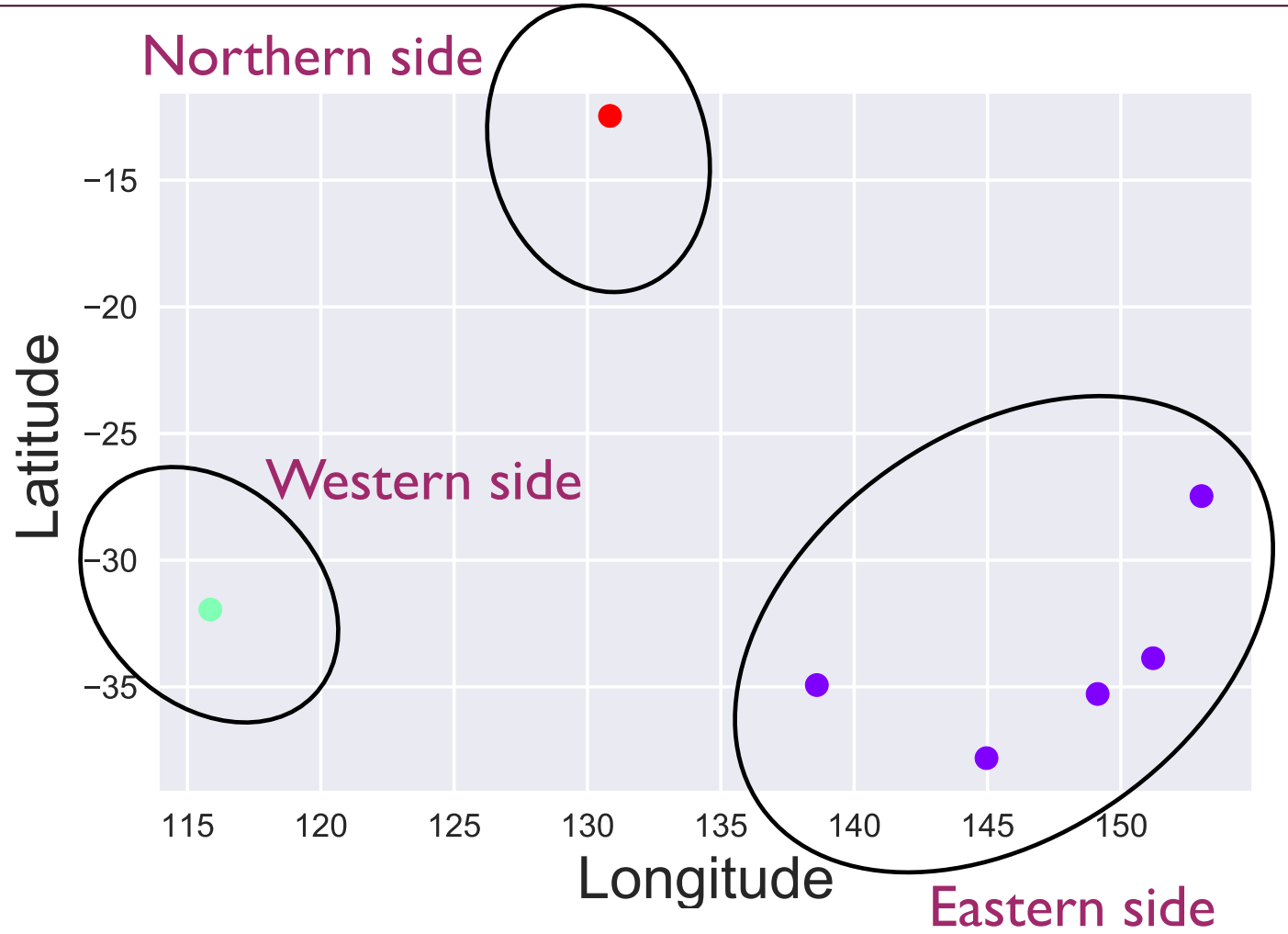

Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude

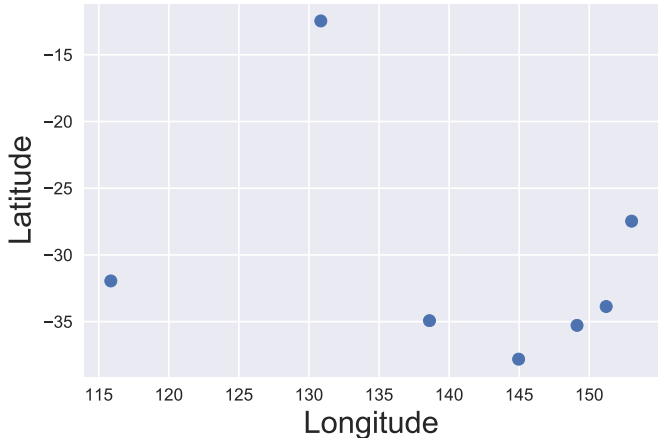

Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude

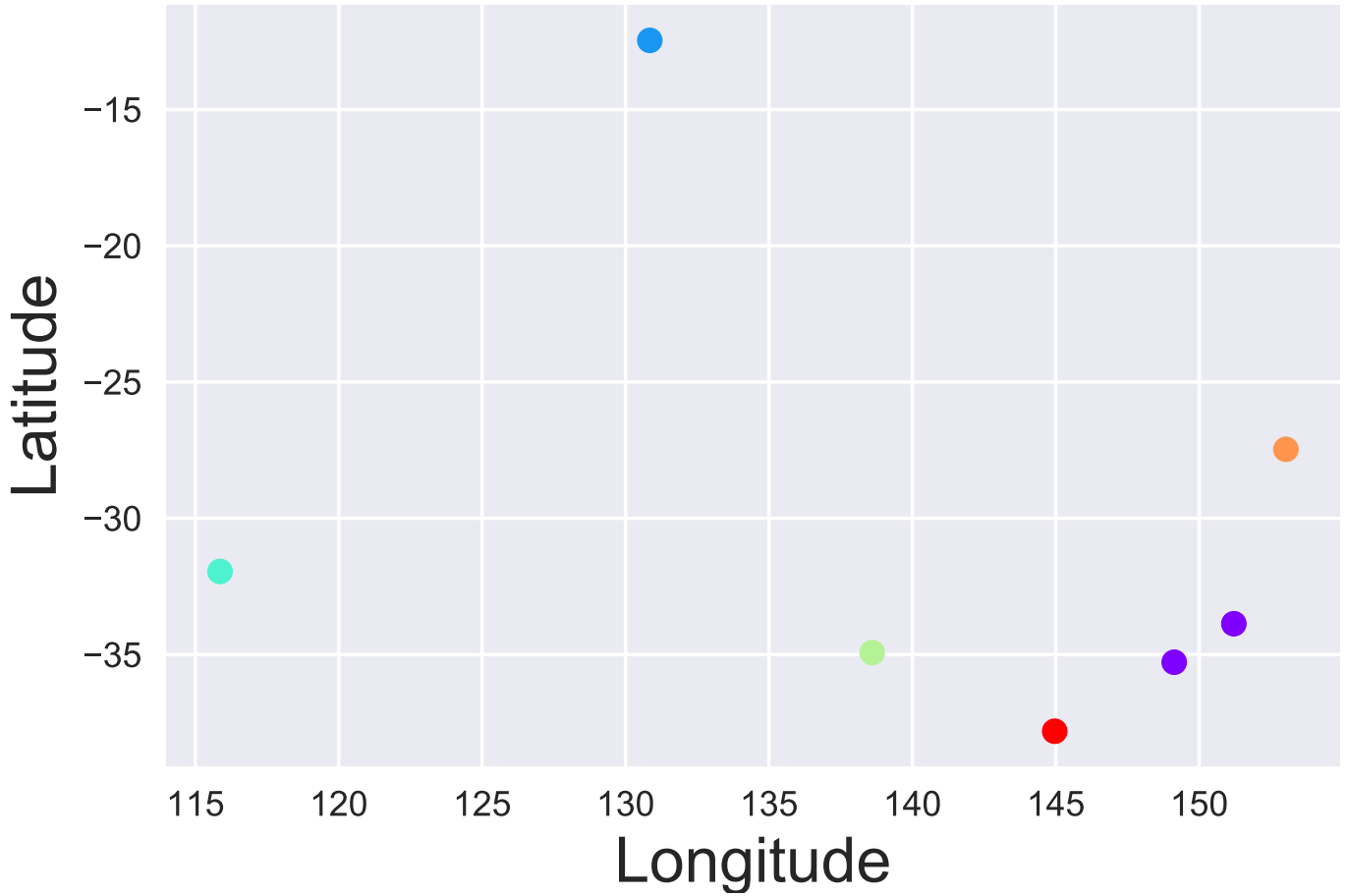

Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude

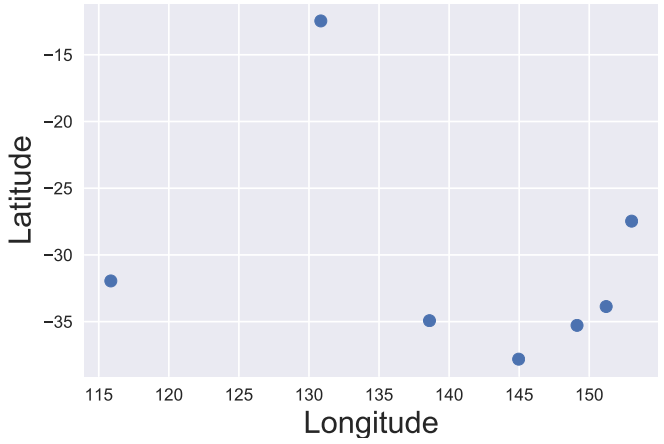

Western side
Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude
Eastern side

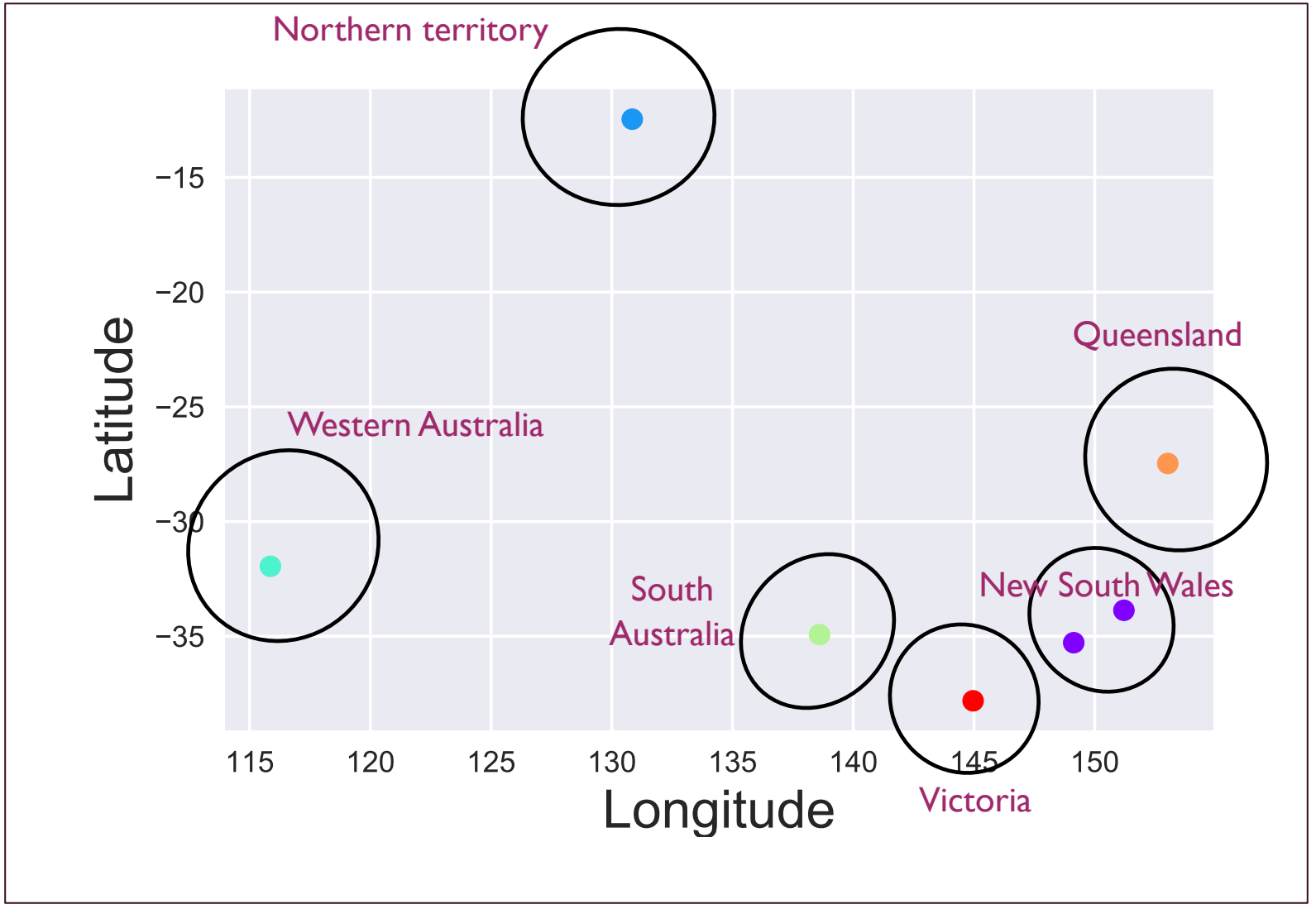

Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude

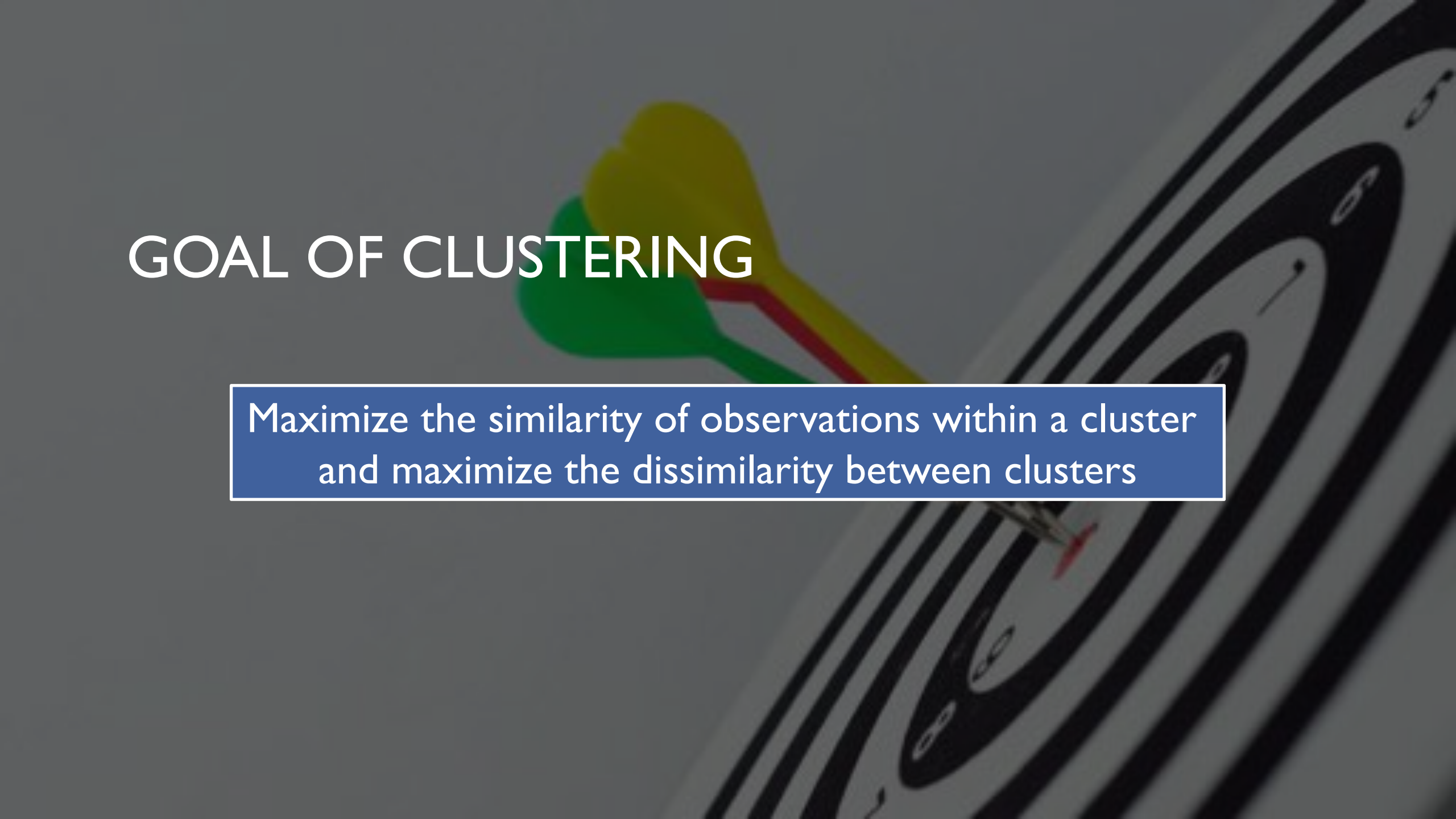

Northern side
Western side
Eastern side
Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude


Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude


Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude


Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude


Northern territory
Western Australia
Queensland
South Australia
New South Wales
Victoria
Latitude
−15
−20
−25
−30
−35
115
120
125
130
135
140
145
150
Longitude



GOAL OF CLUSTERING

Maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters

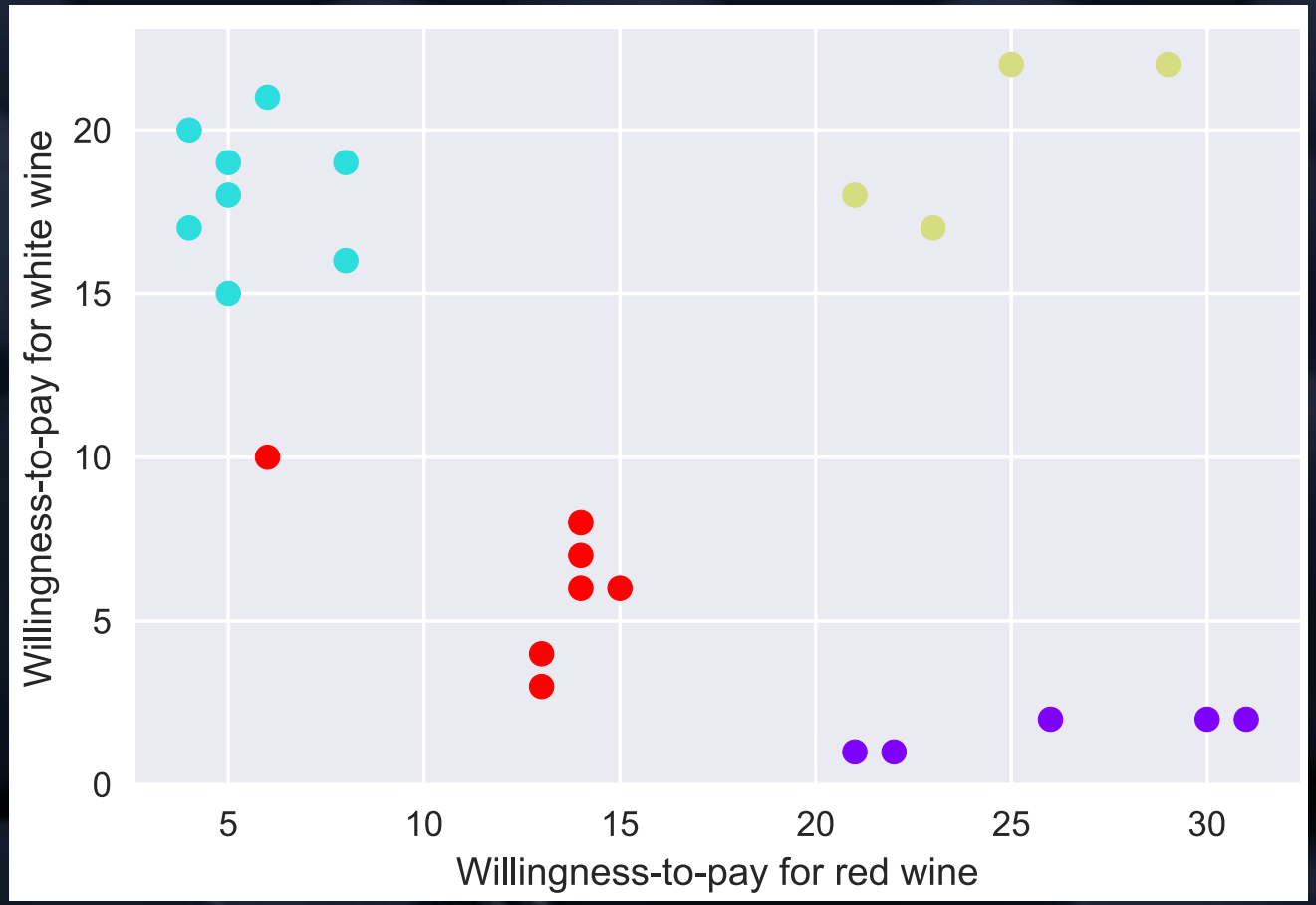
APPLICATIONS

Market segmentation

Image segmentation

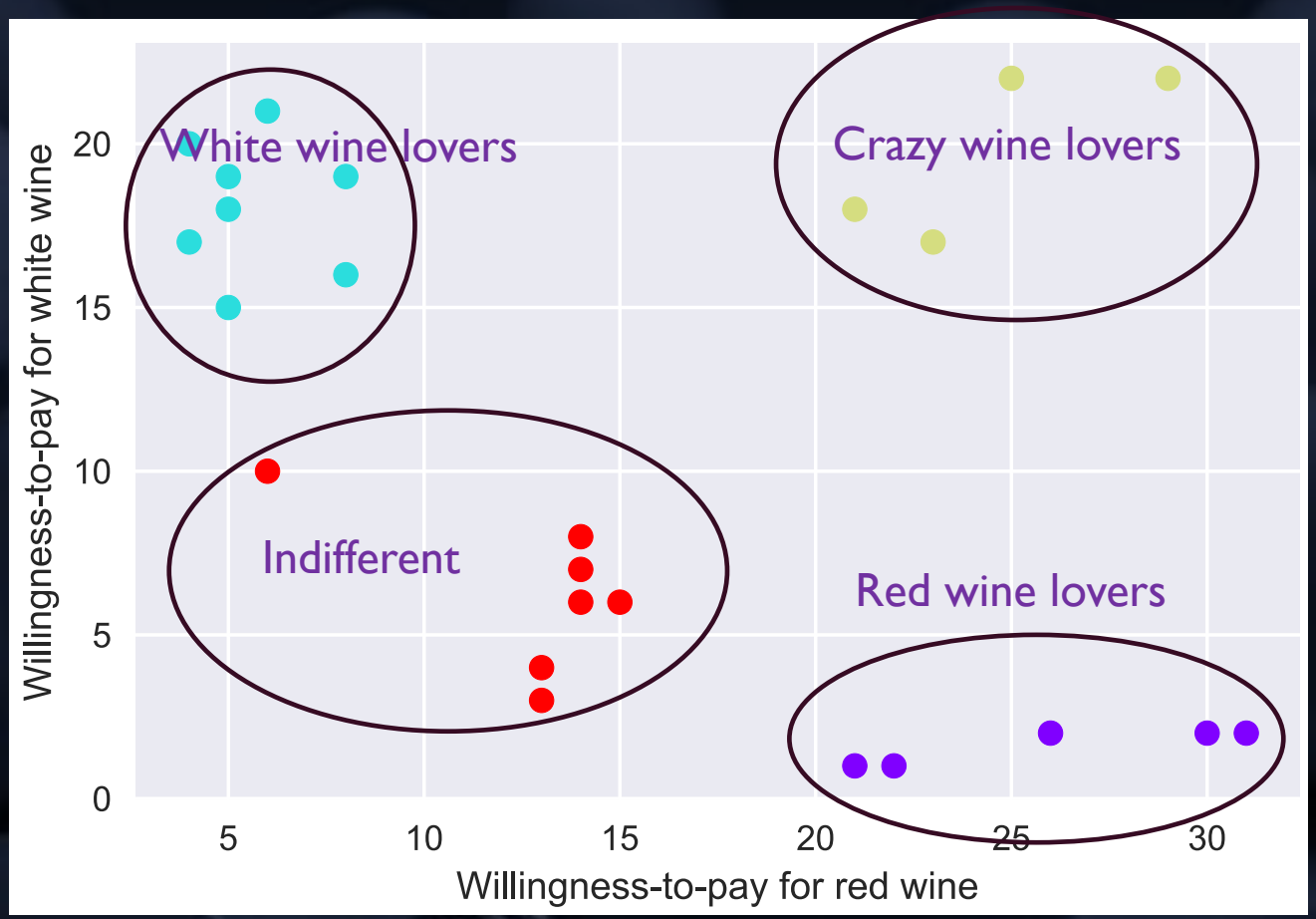

Red wine
White wine

MARKET SEGMENTATION



MARKET SEGMENTATION

- Identify target customers
- Starting point for strategies



CLUSTERING

Preliminary step of all types of analysis

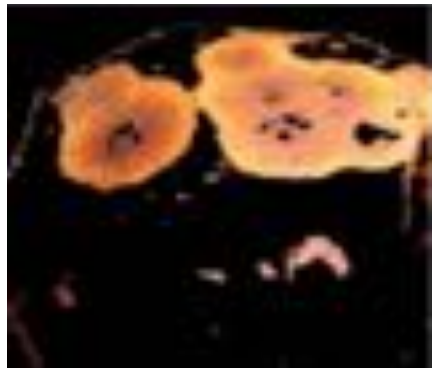
Exploring and identifying patterns in data

Used as a starting point

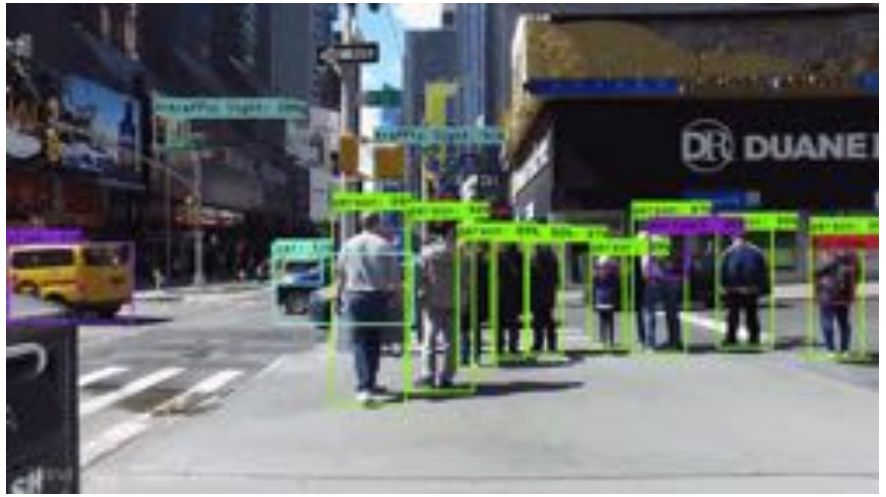
Rarely the sole method used

IMAGE SEGMENTION

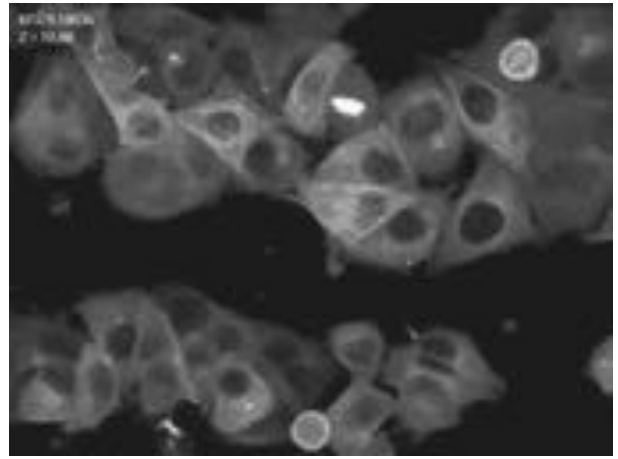
Food technology



Autonomous vehicles



Health science



DIFFERENCE BETWEEN LOGISTIC REGRESSION AND CLUSTERING

Logistic regression

Classify (predict) target variables given the features

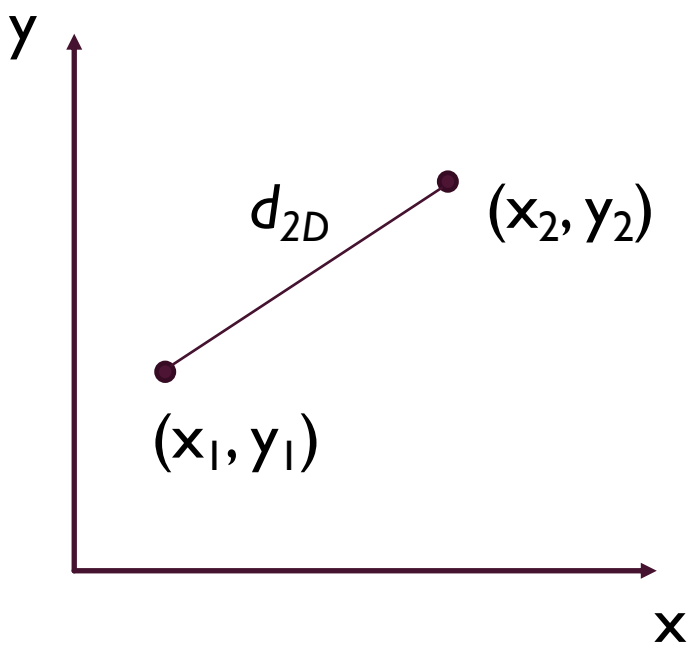
Supervised learning

Clustering

Grouping data given only the features

Unsupervised learning

K-MEANS CLUSTERING



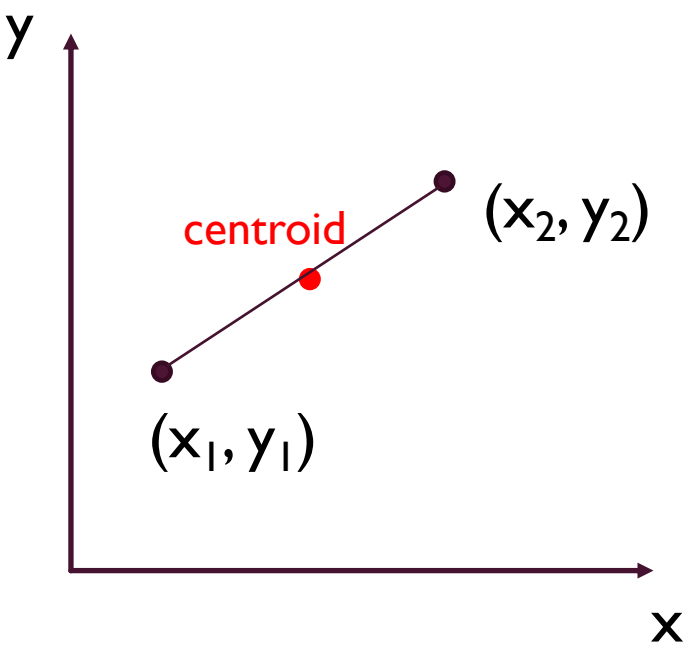
$$d_{2D} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

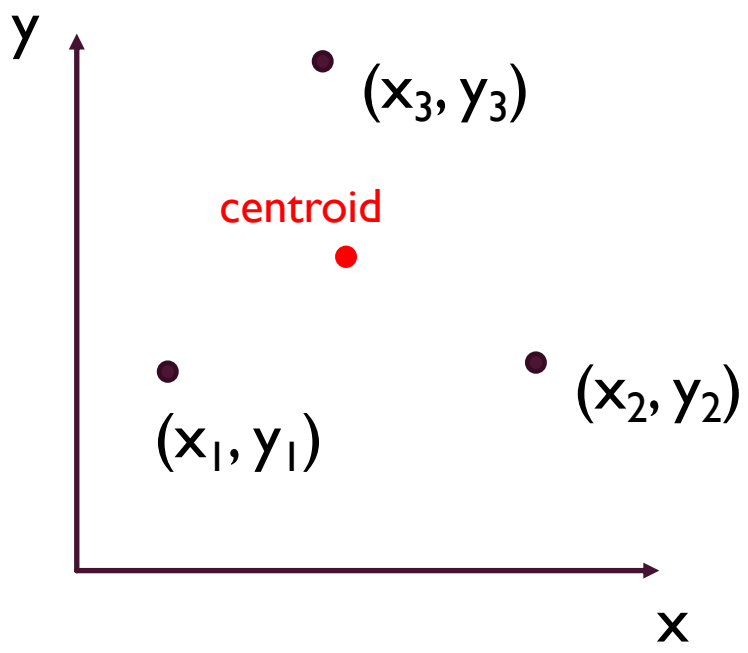
$$d_{3D} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

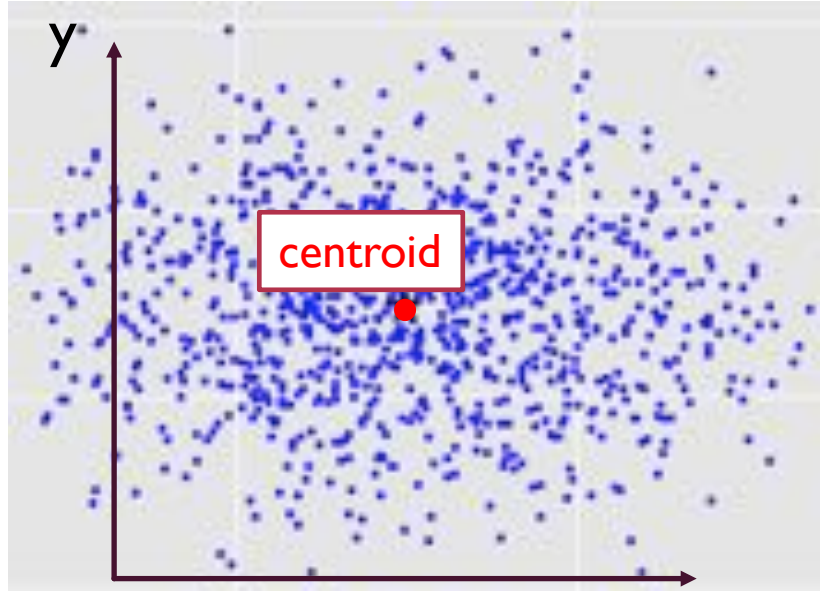
$$d_{nD} = \sqrt{(x_1^1 - x_2^1)^2 + (x_1^2 - x_2^2)^2 + \cdots + (x_1^n - x_2^n)^2}$$

where $x^1, x^2 \ldots x^n$ are coordinates along *n*-axes

Centroid: Mean position of a group of datapoints

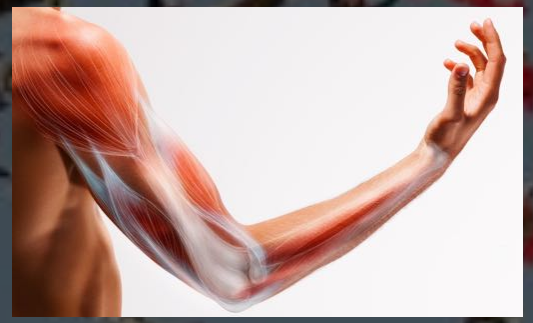






CRITERION FOR SELECTING THE APPROPRIATE NUMBER OF CLUSTERS?

ELBOW METHOD



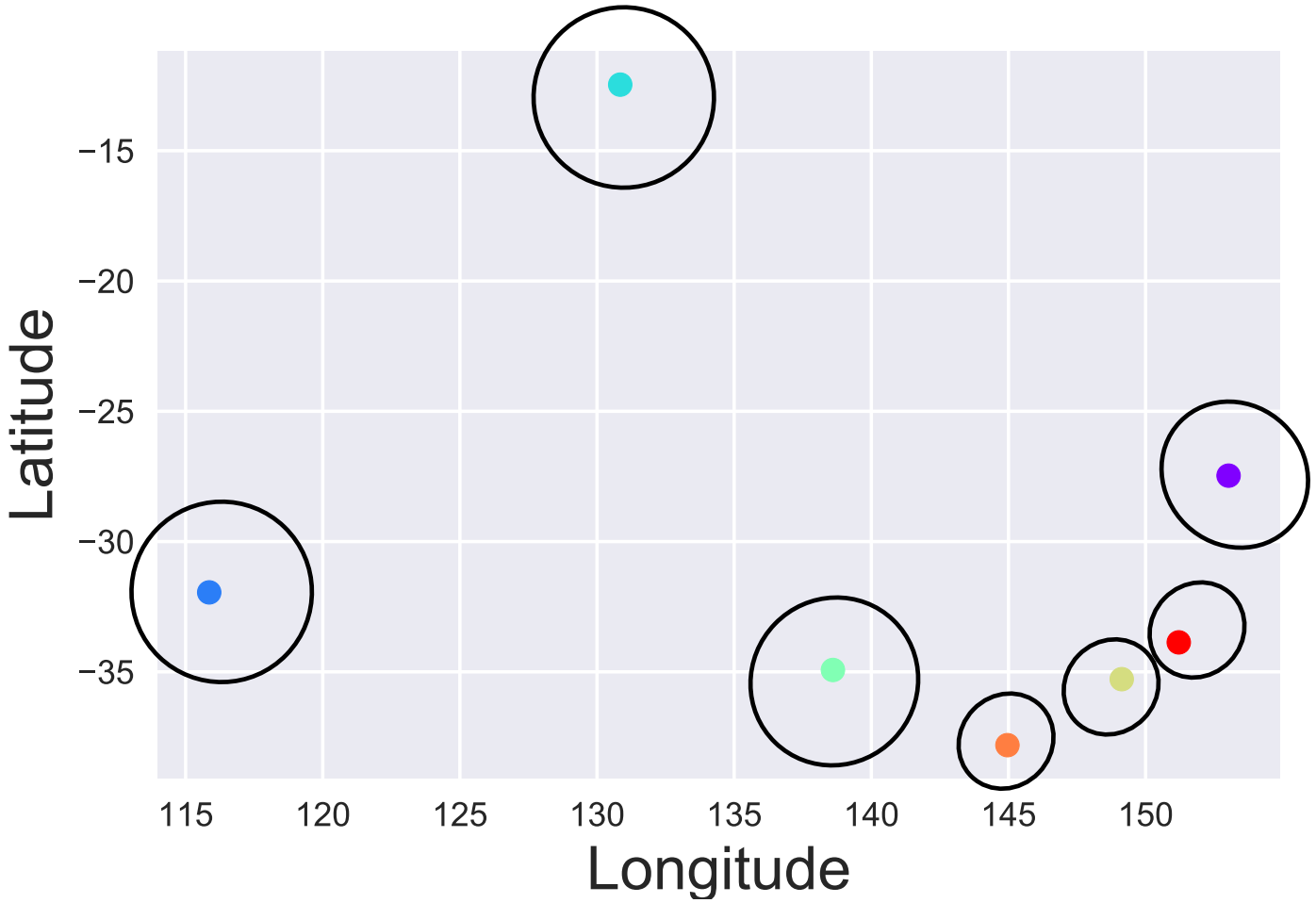
MINIMIZING THE DISTANCE BETWEEN POINTS IN A CLUSTER

MAXIMIZING THE DISTANCE BETWEEN THE CLUSTERS

WITHIN CLUSTER SUM OF SQUARES (WCSS) $\equiv$ sum of squares of the distances of each data point in all clusters to their respective centroids

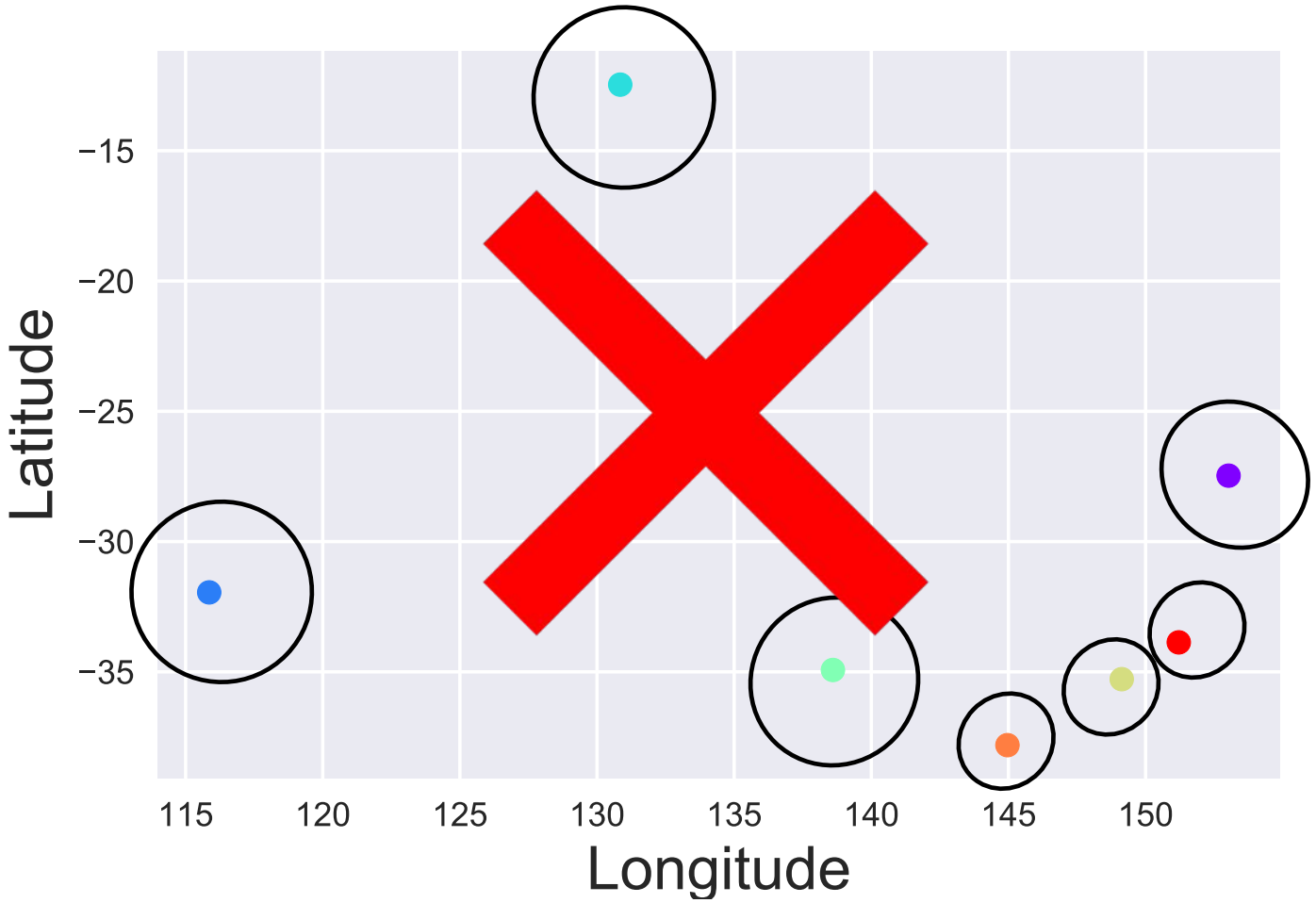
min. WCSS $\equiv$ finding the perfect solution

7 cities and 7 clusters solution



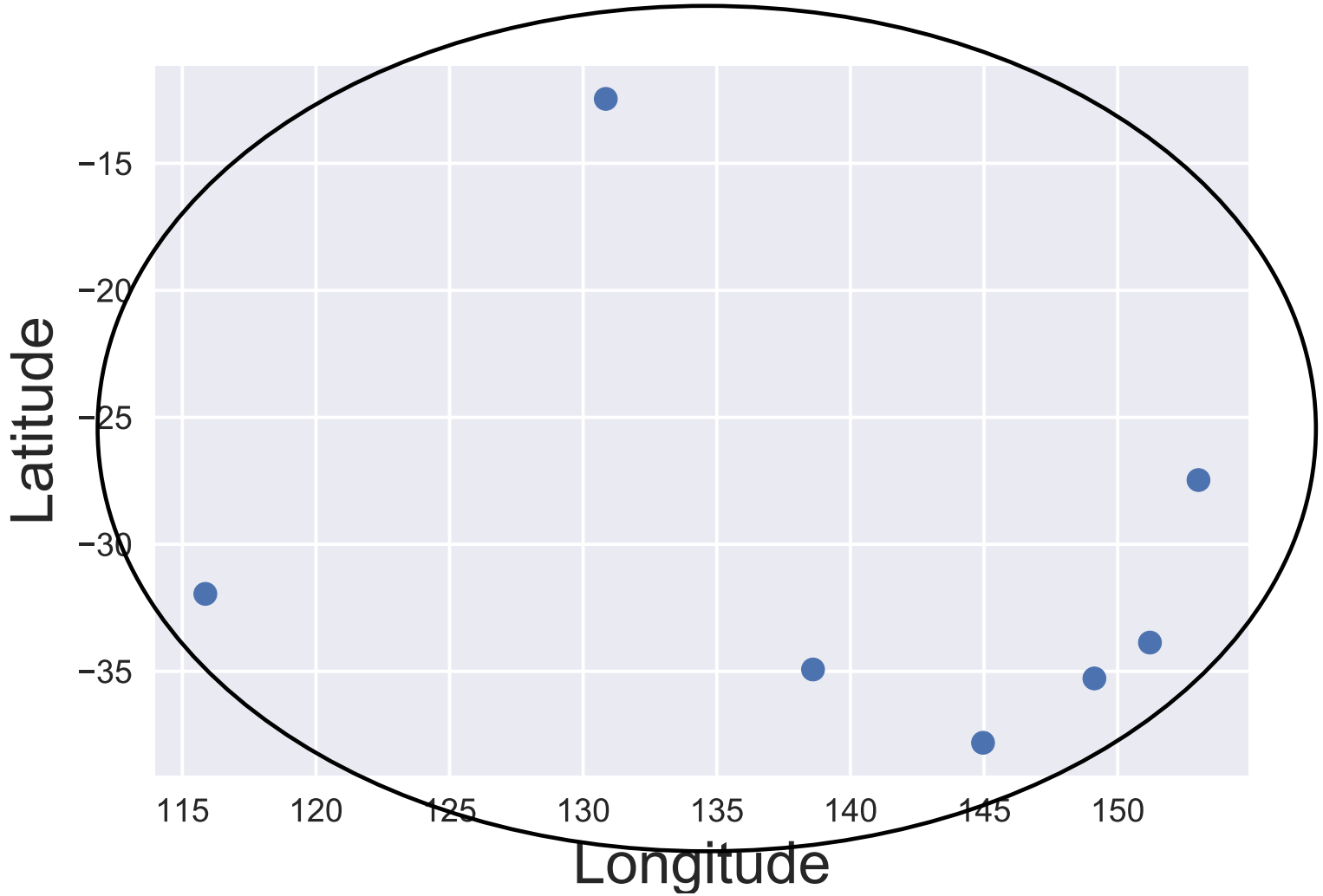
WCSS = 0

7 cities and 7 clusters solution



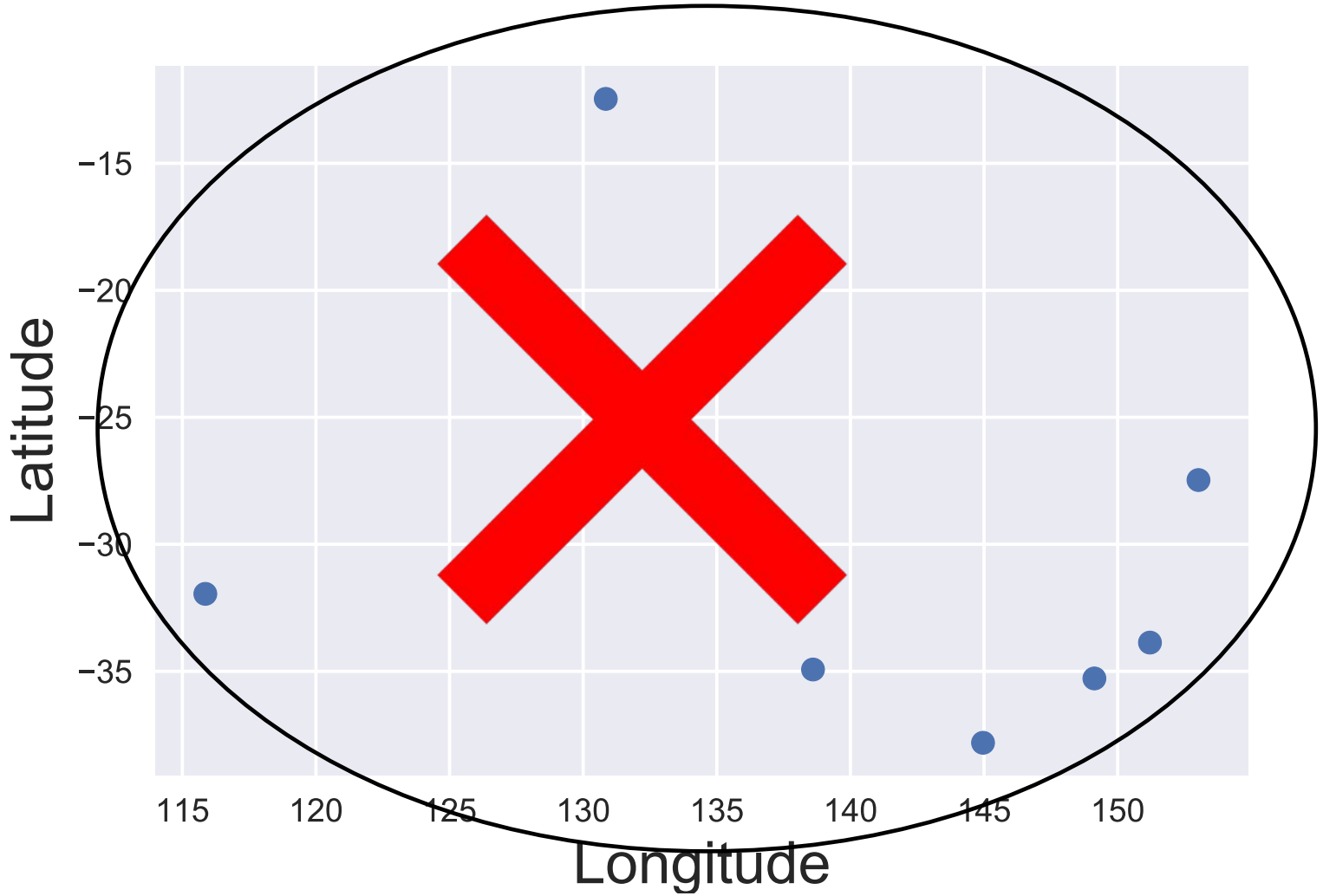
WCSS = 0

7 cities and 1 cluster solution



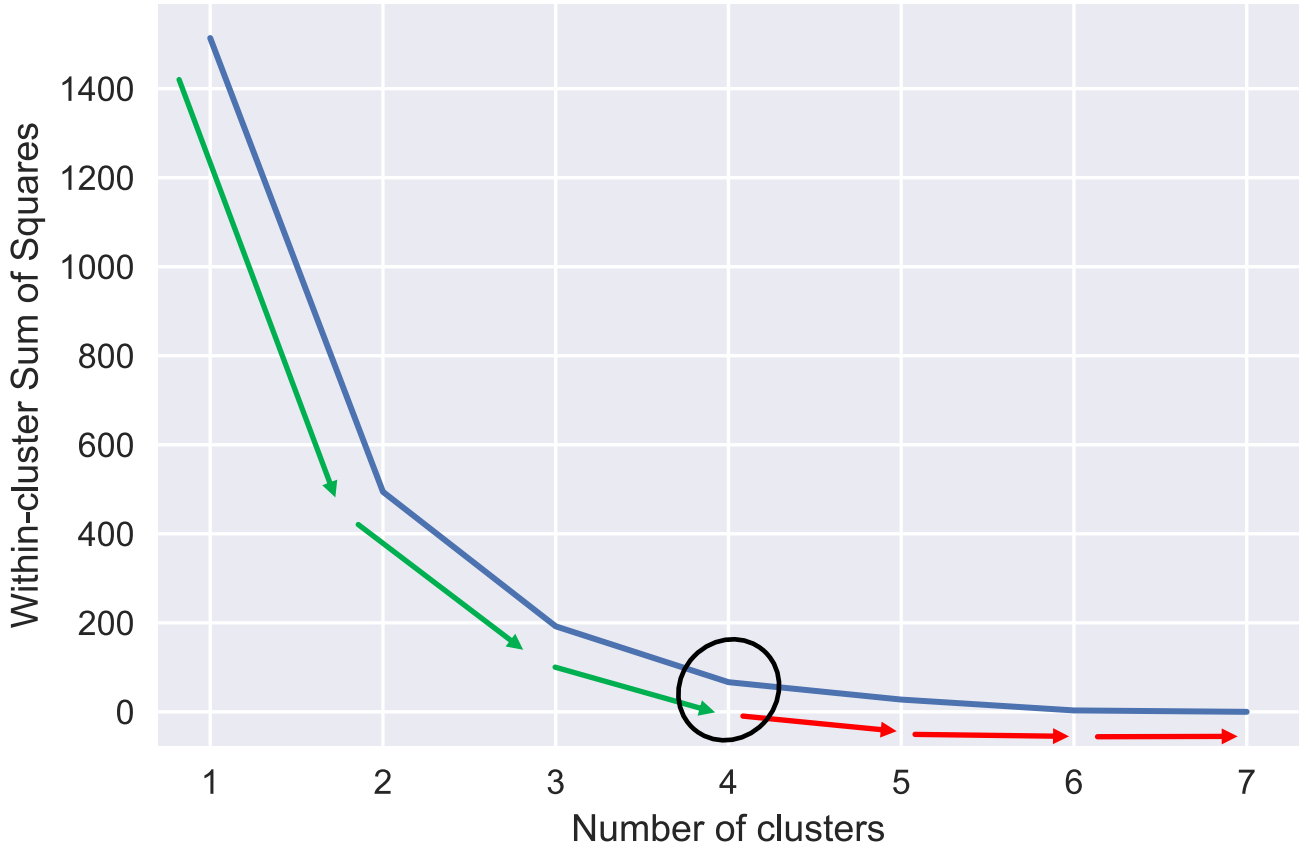
WCSS
=
maximum

7 cities and 1 cluster solution



WCSS
=
maximum

"what we are looking for is for the WCSS to be as low as possible, while we can still have a small number of clusters"



ADVANTAGES AND DISADVANTAGES

ADVANTAGES

- Simple to understand
- Clustering can be done quickly
- Many packages offer K-means
- K-means will always yield a result (could be disadvantageous at times!)

DISADVANTAGES

We need to choose the number of clusters

- Remedy: Elbow method

Elbow method is not very scientific

K-means is sensitive to initialization of centroids

- Remedy: KMeans++

K-means is sensitive to outliers

- Remedy: Remove outliers