Hierarchical Clustering

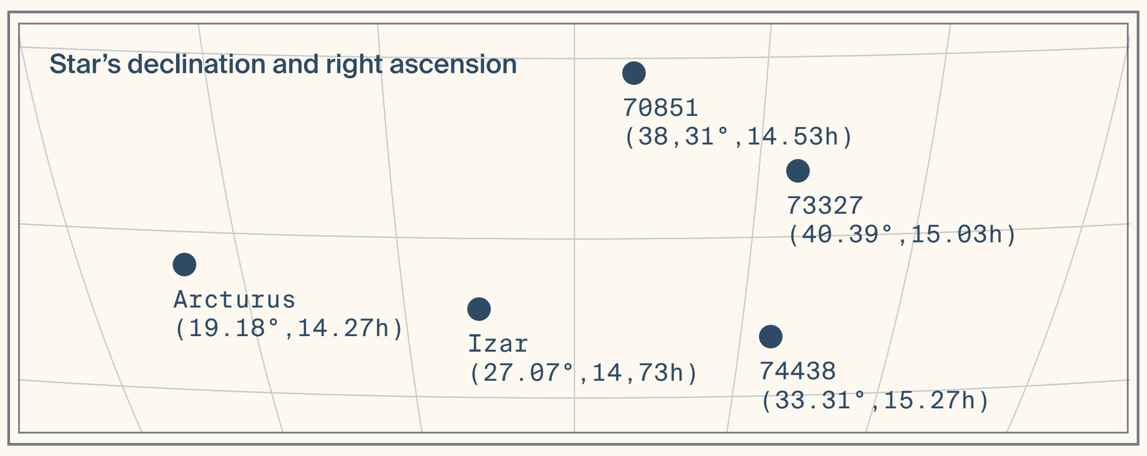
In search of data-driven constellations

Since time immemorial, mankind has looked at the sky, gazing at heavenly bodies and connecting the closest ones to draw figures in the sky. This is how constellations originated. What would constellations look like if, instead of humans, it was an algorithm that searched for patterns in the celestial vault?

Hierarchical clustering is a method used in descriptive statistics to determine a **hierarchy of clusters**, which are collections of samples **based on similarities** among their features. It is an **unsupervised learning** method, this means it reveals spontaneous patterns found in the dataset instead of relying on human-defined groupings. For this reason, it's the perfect tool to find new data-driven constellations.

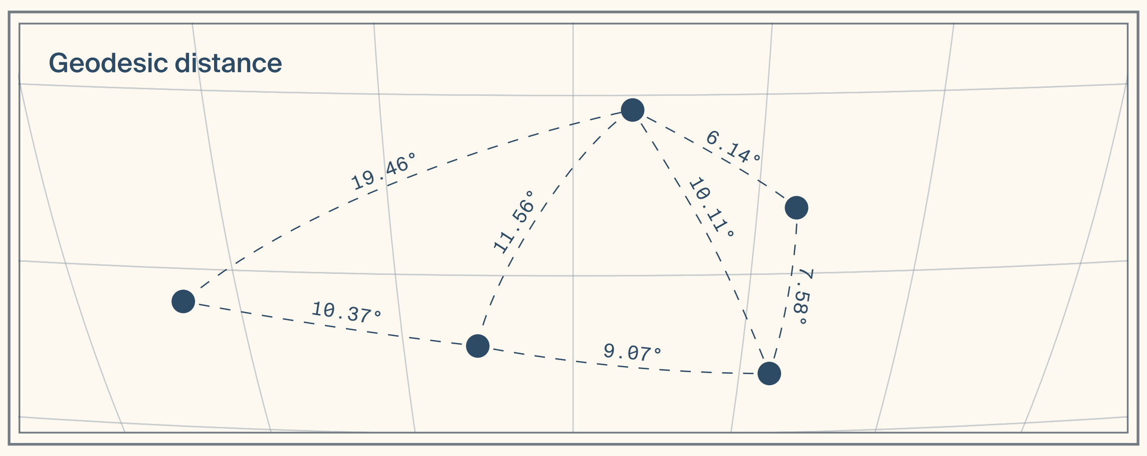
Key concepts of hierarchical clustering

Observation

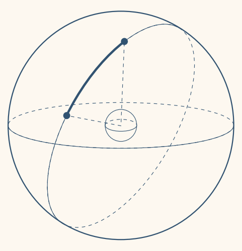


When looking for similarities, first we choose which features to compare and then we make sure these features are commensurable, to assess their similarity. For example, we could group stars according to their luminosity or hue. In our case, we consider *declination* and *right ascension* as features to compare; they are the **spherical coordinates** of any star on the celestial vault, similar to longitude and latitude on Earth.

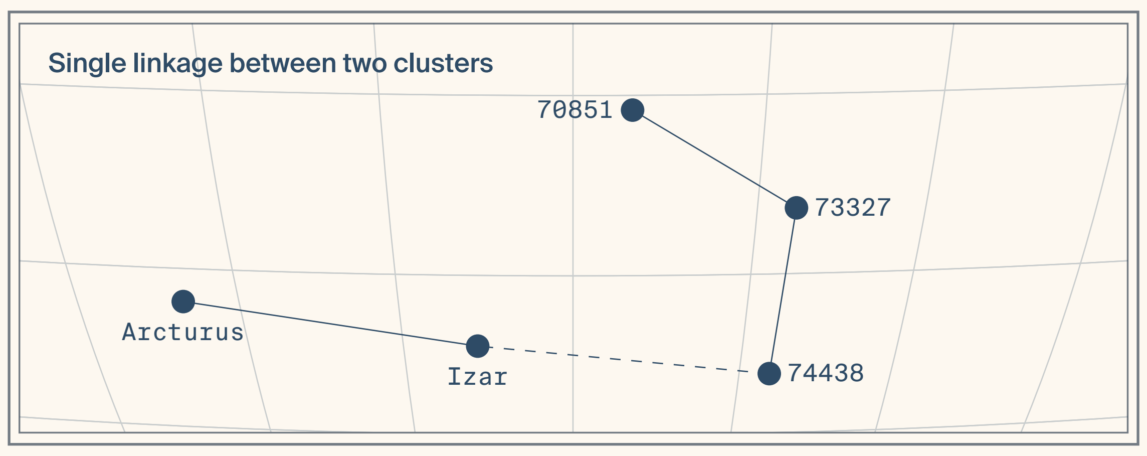
Distance



The similarity of two samples can be quantified in different ways according to the nature of the data being analised and to the scope of the analysis. The aim of distance measure is to find similar data objects and to group them in the same cluster. In this example, we will be using **geodesic distance** which measures distance along a non-flat surface.



Linkage



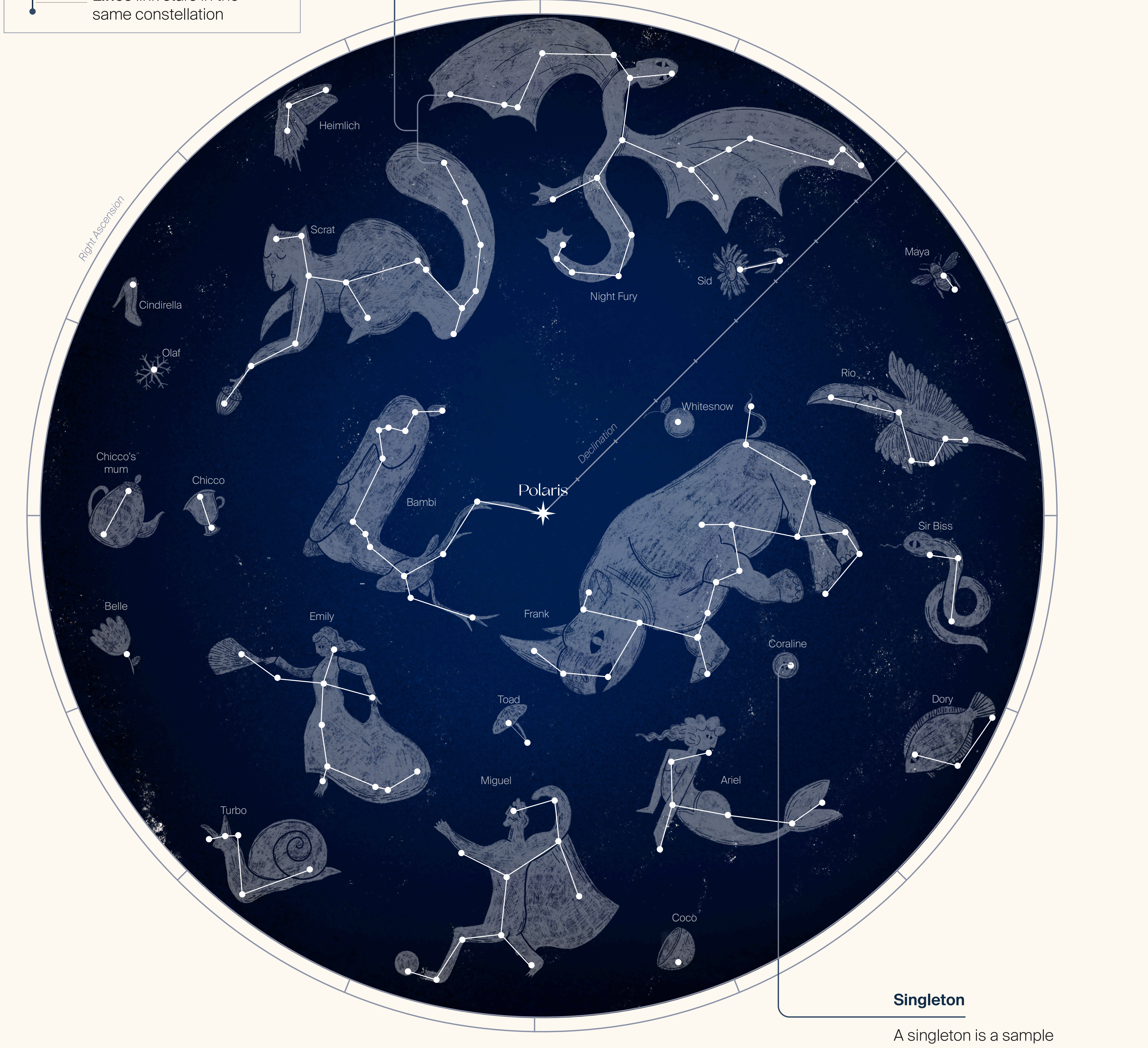
Since clusters contain multiple data objects, we have different options to measure distance, e.g. between the centres of each cluster, between the furthest points of each cluster. In our case, we use the **single-linkage criterion**; it states that two clusters are as similar as their most similar elements – or as close as their closest stars, in our example.

Legend

Groups of circles and lines are **constellations** and represent a cluster

Circles represent stars

Lines link stars in the same constellation

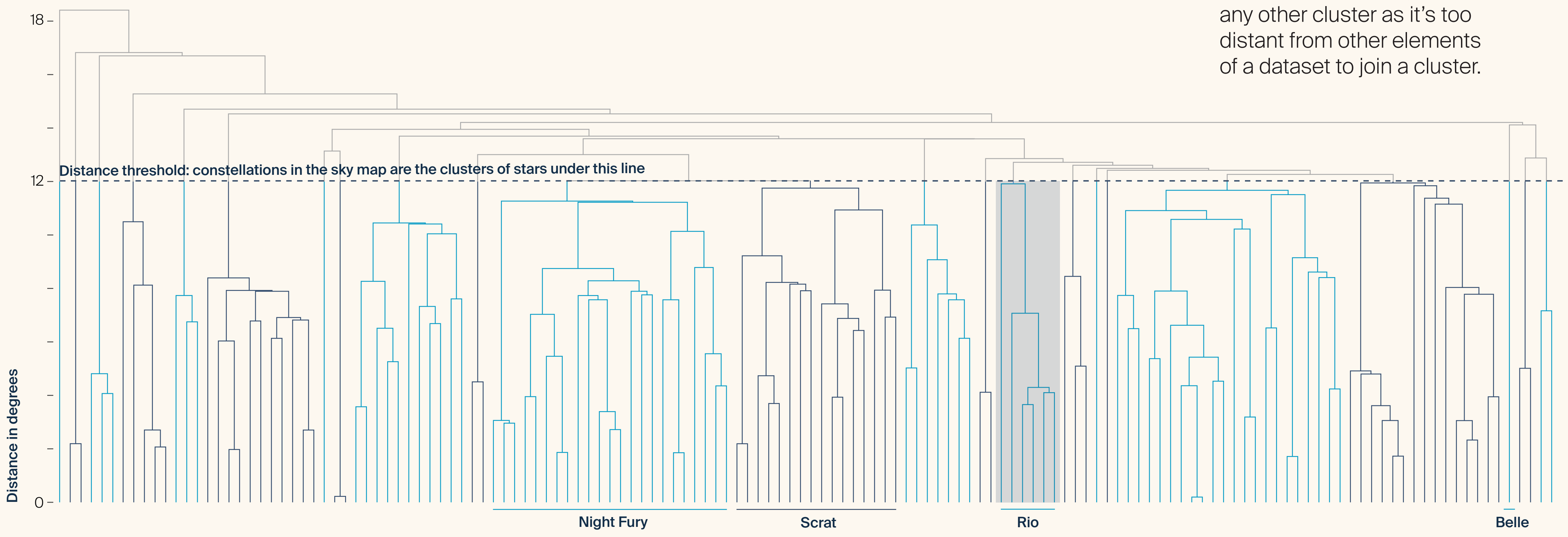


Mind the gap

These stars are 12.02° apart, since our pruning threshold is 12° these two constellations are not clustered together.

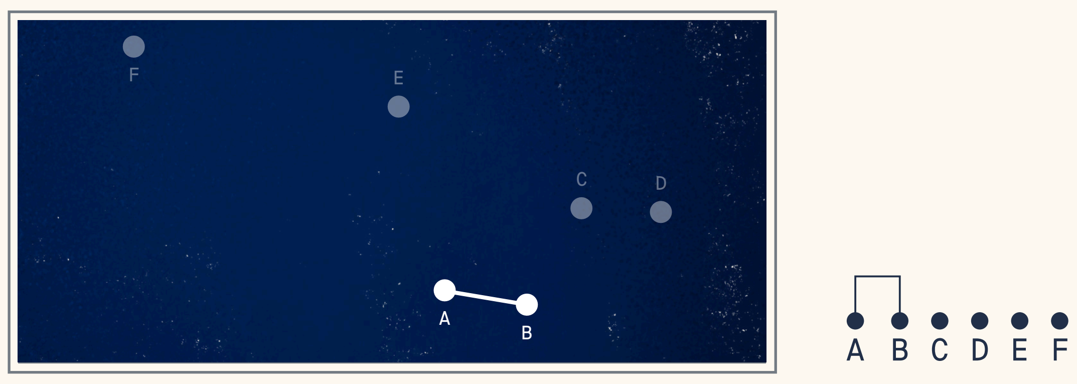
Singleton

A singleton is a sample that does not belong to any other cluster as it's too distant from other elements of a dataset to join a cluster.

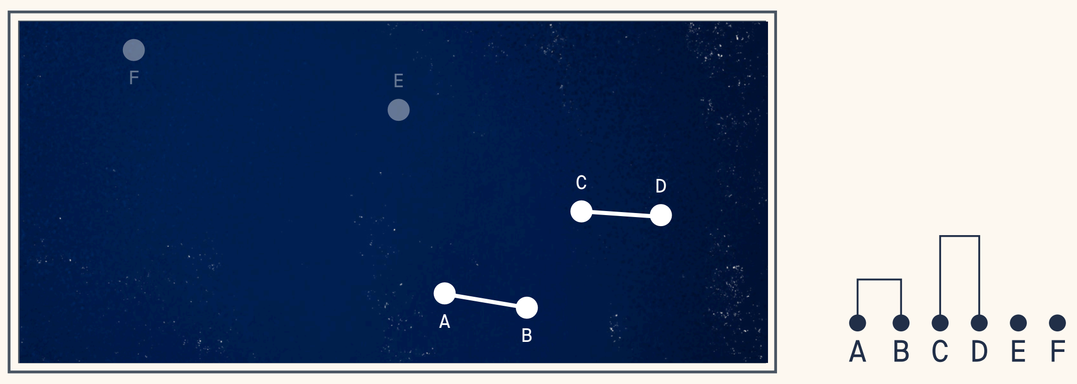


Building a hierarchy

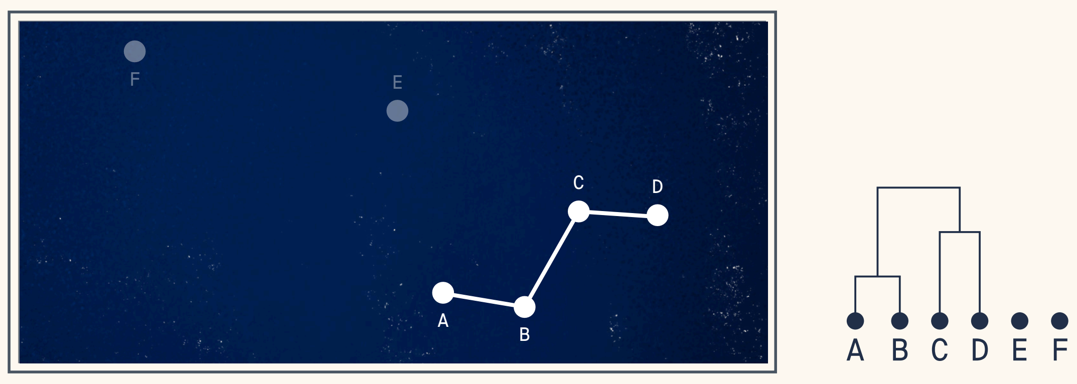
Here we illustrate how hierarchical clustering can be applied to our example, while visualising the results through a dendrogram, a diagram that codifies information about our samples' similarity and their hierarchy.



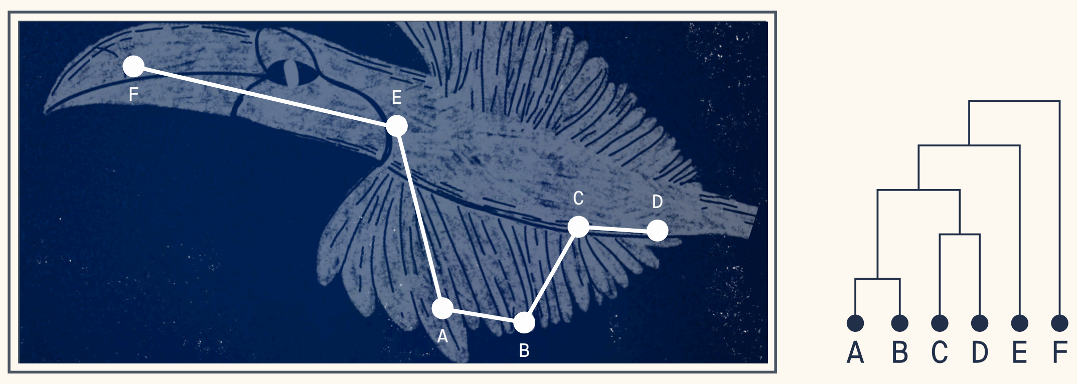
1. The distance between every possible pair of stars is computed. Once the two closest stars are found, they are grouped into a cluster. On the dendrogram the samples are connected by a line, this means they have been clustered together.



2. The previous step is repeated, a new cluster is formed and drawn on the dendrogram. We can see that the vertical lines of this cluster have a different height as it is proportional to the distance between the two stars.



3. Distances are computed once more but now the two closest elements are not stars but the two clusters that have just been formed. Looking at the dendrogram on the right we can see how a hierarchy is starting to form.



4. The last data point can now join the clusters to its left. Their order on the horizontal axis does not codify any information about the analysis; most often it is simply the order that maximises the dendrogram's readability.

Pruning

At the end of the process, all 142 stars are collected in a single, large cluster and a hierarchy is defined for the entire dataset. In our case, we would end up with one large constellation, while we need to define multiple, distinct ones instead.

The selection of clusters is called pruning, from the idea of cutting branches off the dendrogram and picking the resulting subtrees as clusters. The pruning criterion depends once again on the scope of the analysis. In this example, we set a distance threshold of **twelve degrees** and selected the twenty-three resulting subtrees as constellations.

VISUAL EXPLANATIONS OF STATISTICAL METHODS

Hierarchical clustering

AUTHORS

Alice Bocchio
Michele Bruno
Maria Celeste Casolino
Luca Draisci
Virginia Leccisotti
Barbara Roncalli
Sara Zanardi

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Anna Riboldi

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2021/2022

DEN–
SITY
GN+

POLITECNICO
MILANO 1863
SCHOOL OF DESIGN