UML - Intro

Dimensionality reduction and clustering

Roman Jurowetzki

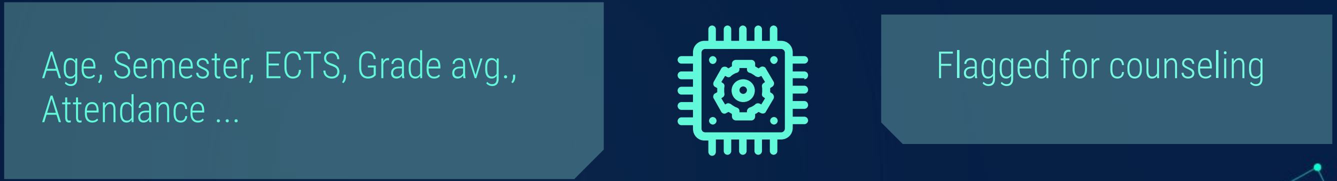
Wait: What is ML?

- "Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world." - Nvidia

- "Machine learning is the science of getting computers to act without being explicitly programmed." - Stanford

- "Machine learning is based on algorithms that can learn from data without relying on rules-based programming."- McKinsey & Co.

ML vs Programming?



Age, Semester, ECTS, Grade avg., Attendance ...

Flagged for counseling

01

UML Intuition

Non-technical intro to unsupervised ML

Supervised vs Unsupervised ML

Model learns to categorize or predict some value from different measures
e.g. combination of shape, color, weight ⟶ which fruit

Supervised

Model learns general patterns in the data, how to represent the data in a more compressed way (dimensionality reduction) or identify related observations (clustering)

Unsupervised

02

Dimensionality reduction

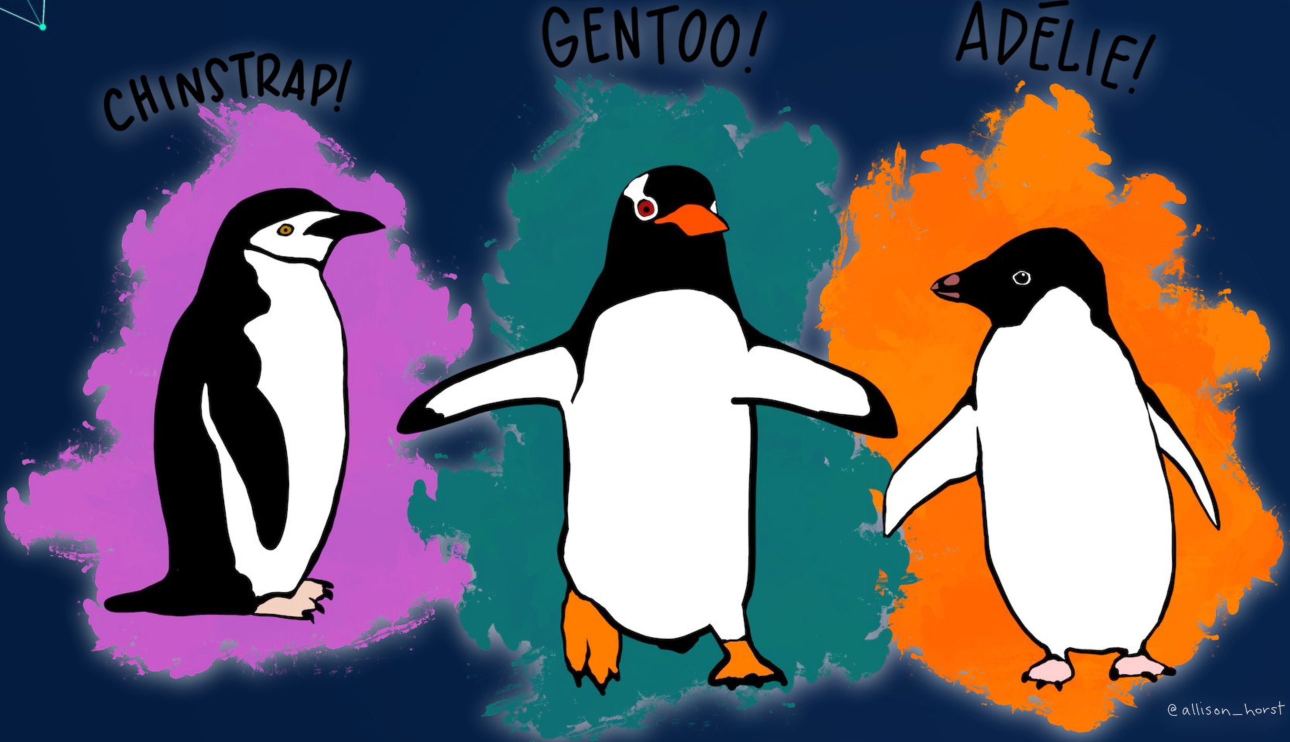
PCA, Factor analysis, NMF, t-SNE and more

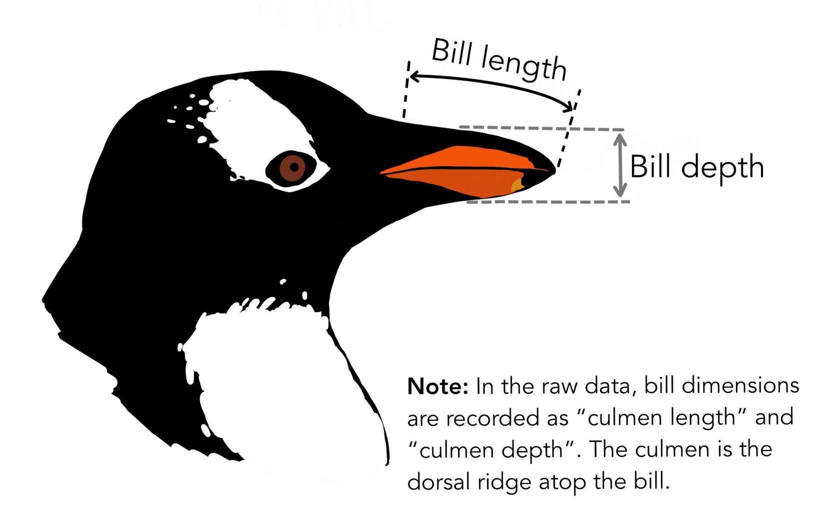
Dimensions here is a synonym for **variables**, so what we want to really do is have **less variables**. To do that, we have to find ways to express the **same amount of information with fewer, but more information-rich variables**. This is particularly useful to:

- Find patterns in the features of the data.
- Visualization of high-dimensional data.
- Pre-processing before supervised ML tasks.

The type of analysis to be performed depends on the data set formats and structures. The most commonly used DR techniques are:

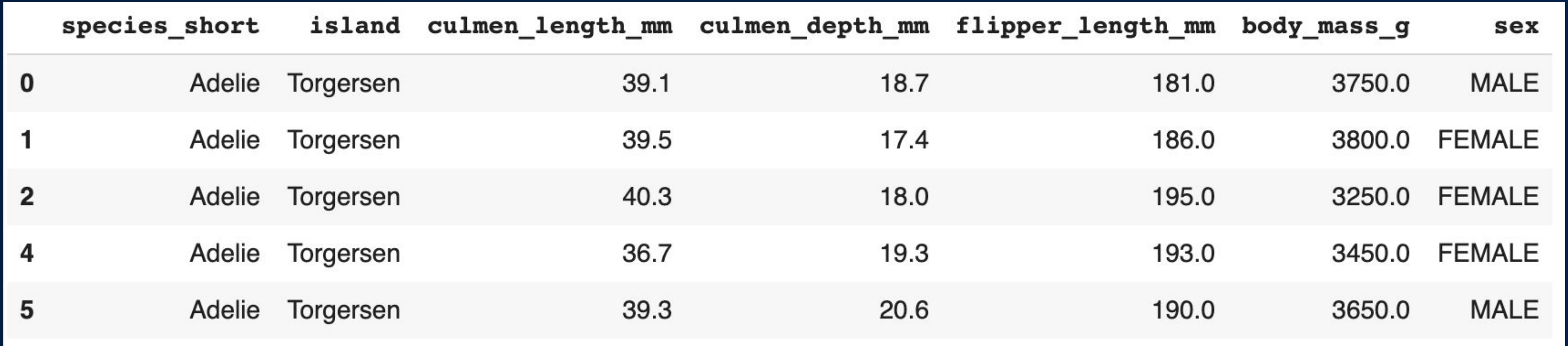
Principal Component Analysis (PCA): Is used to summarize the information contained in a continuous (i.e, quantitative) multivariate data by reducing the dimensionality of the data without losing important information.
(**Multiple) Correspondence Analysis ((M)CA**): An extension of the principal component analysis suited to analyse a large contingency table formed by two qualitative variables (or categorical data).
Non-negative matrix factorization (NMF or NNMF): Similar to PCA but only to be used on positive values and Unlike PCA, the representation of a vector is obtained in an additive fashion. Identifies human-interpretable “patterns” or “topics”.
Latent Dirichlet Allocation (LDA): Generative probabilistic model for collections of discrete dataset such as text corpora, where abstract topics need to be found.
Uniform Manifold Approximation and Projection (UMAP): Recent scalable approach for visualization and general non-linear dimension reduction.


CHINSTRAP!
GENTOO!
ADÉLIE!
@allison_horst



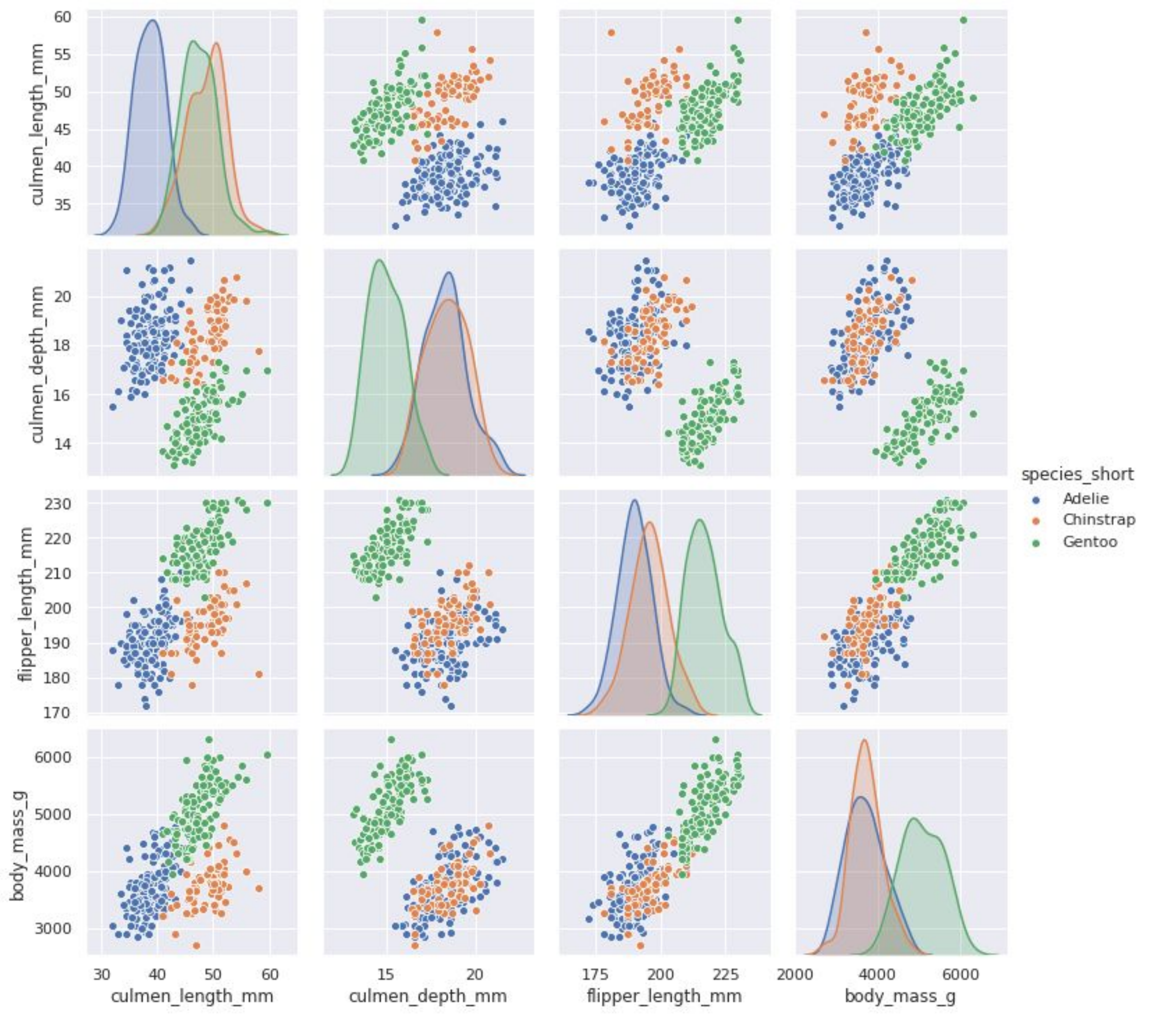
Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Adelie	146
Gentoo	120
Chinstrap	68

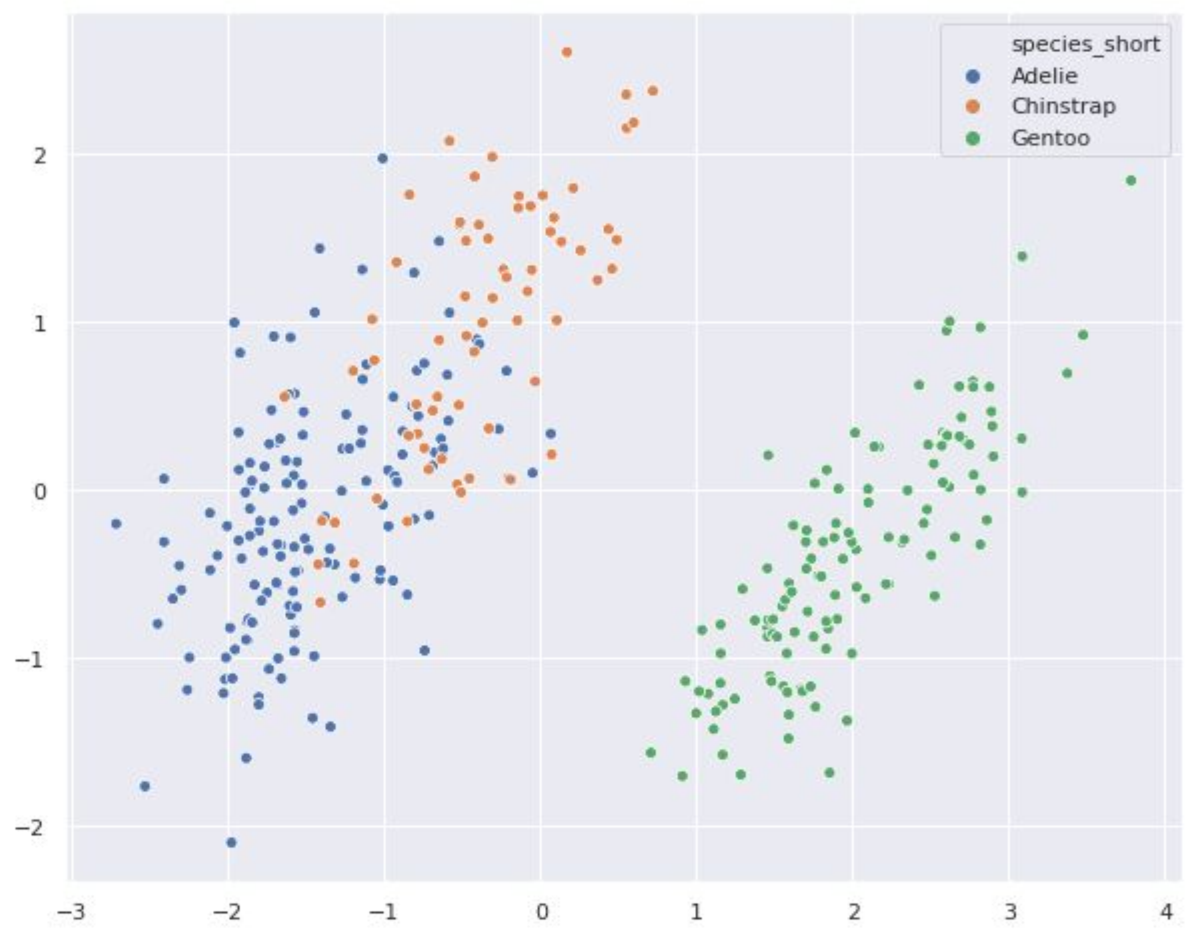


	species_short	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0	MALE

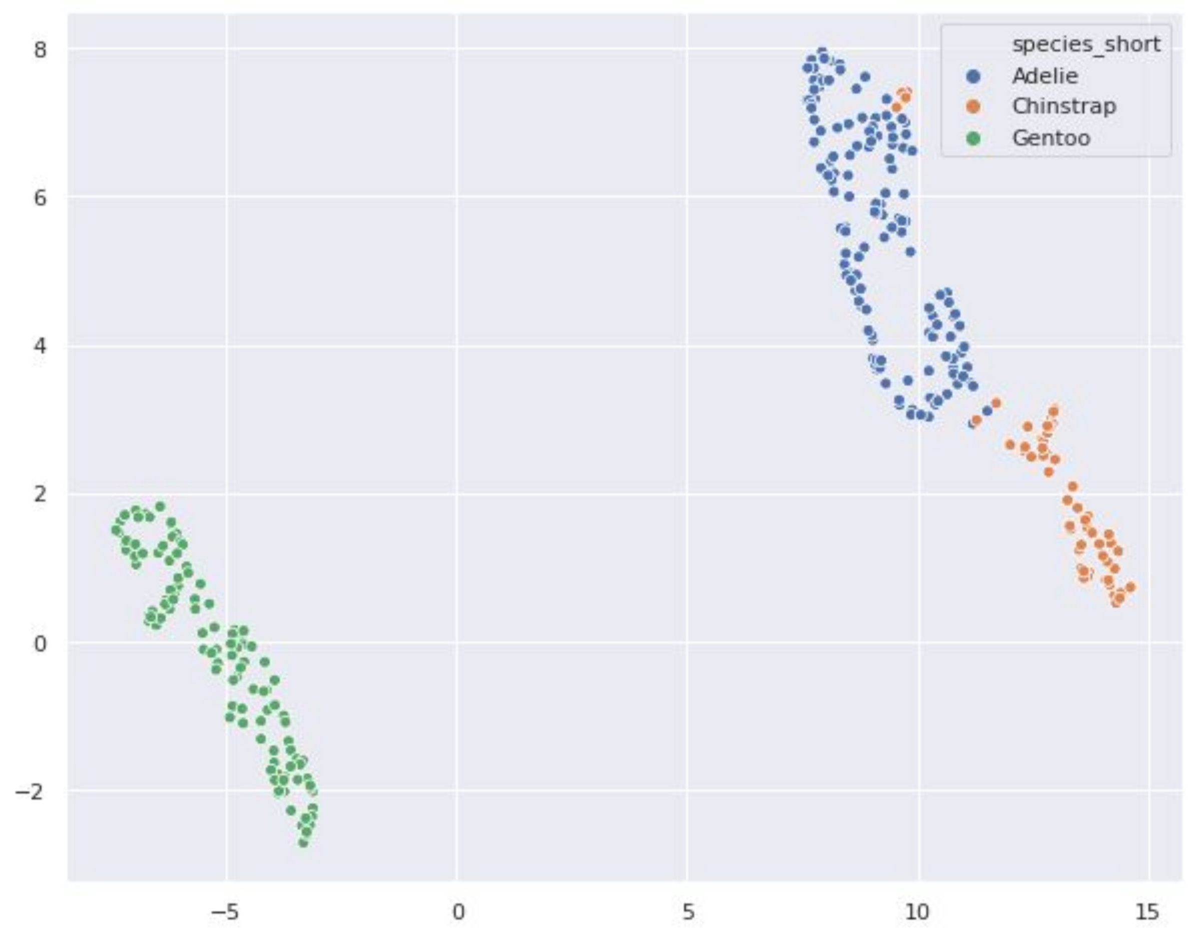
Penguin datast: https://github.com/allisonhorst/palmerpenguins

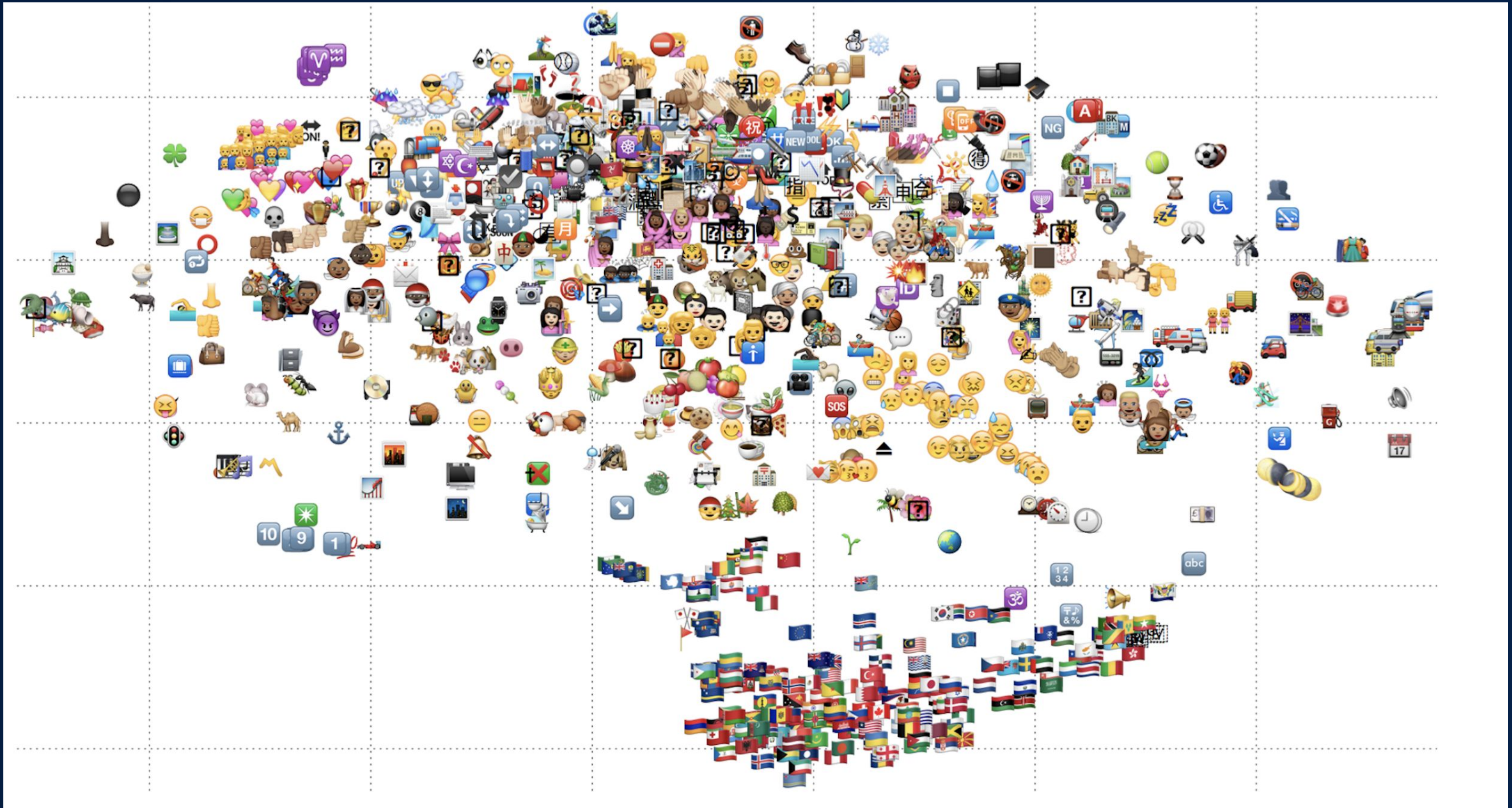

culmen_length_mm
culmen_depth_mm
flipper_length_mm
body_mass_g
60
55
50
45
40
35
20
18
16
14
230
220
210
200
190
180
170
6000
5000
4000
3000
30
40
50
60
15
20
175
200
225
2000
4000
6000
culmen_length_mm
culmen_depth_mm
flipper_length_mm
body_mass_g
species_short
● Adelie
● Chinstrap
● Gentoo

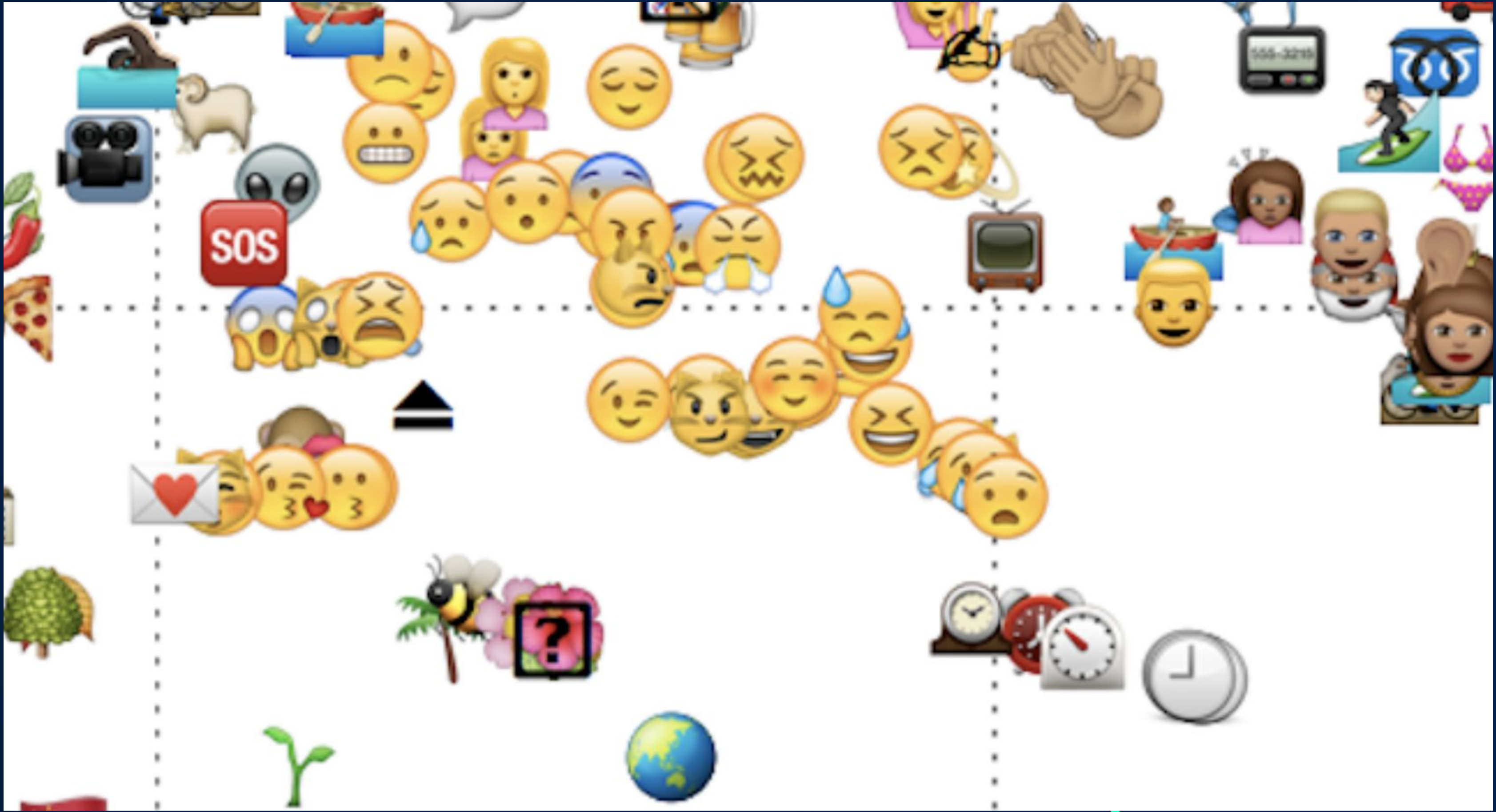
PCA

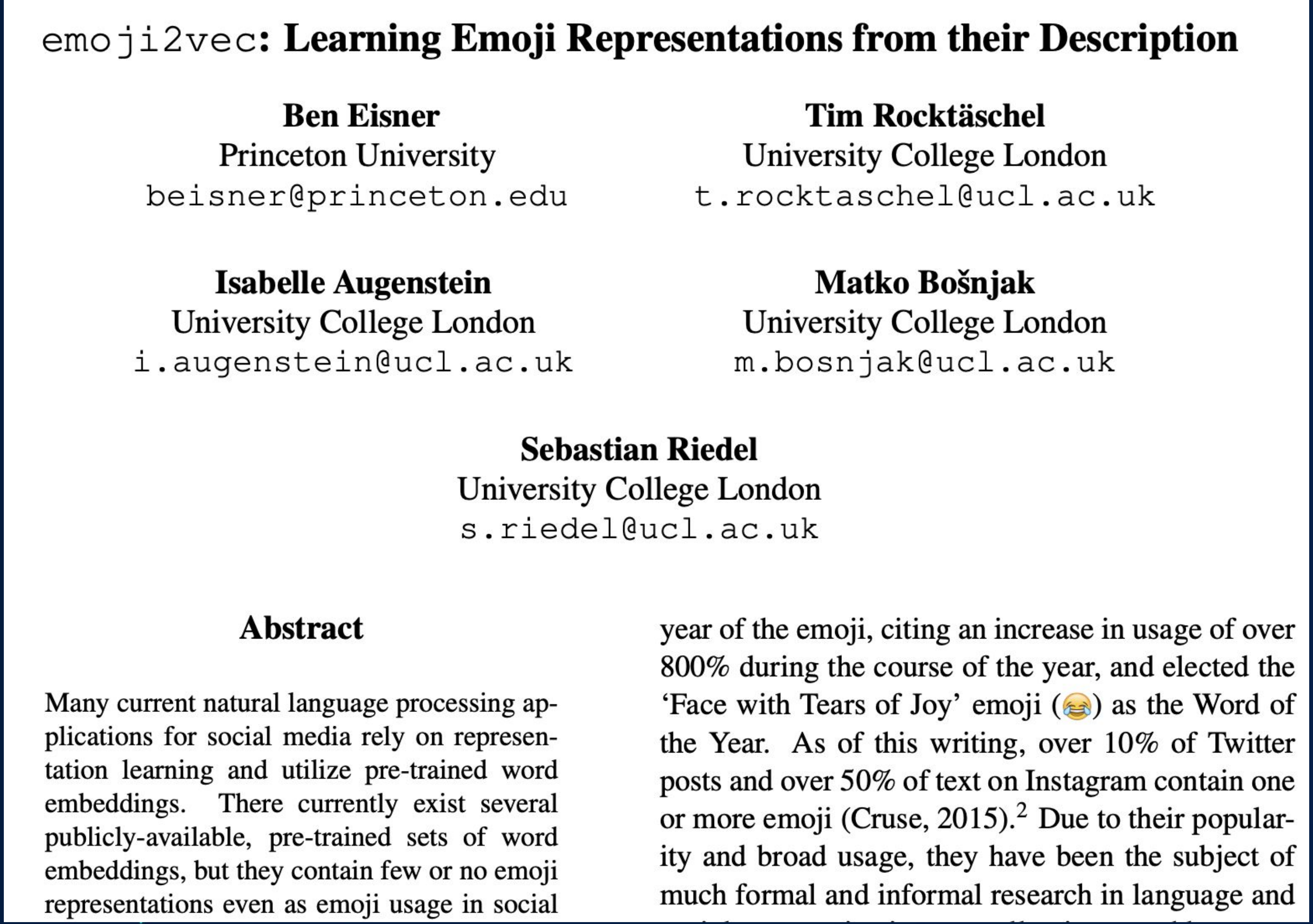


UMAP






SOS



emoji2vec: Learning Emoji Representations from their Description

Ben Eisner
Princeton University
beisner@princeton.edu

Tim Rocktäschel
University College London
t.rocktaschel@ucl.ac.uk

Isabelle Augenstein
University College London
i.augenstein@ucl.ac.uk

Matko Bošnjak
University College London
m.bosnjak@ucl.ac.uk

Sebastian Riedel
University College London
s.riedel@ucl.ac.uk

Abstract

Many current natural language processing applications for social media rely on representation learning and utilize pre-trained word embeddings. There currently exist several publicly-available, pre-trained sets of word embeddings, but they contain few or no emoji representations even as emoji usage in social

year of the emoji, citing an increase in usage of over 800% during the course of the year, and elected the 'Face with Tears of Joy' emoji (😂) as the Word of the Year. As of this writing, over 10% of Twitter posts and over 50% of text on Instagram contain one or more emoji (Cruse, 2015).[2] Due to their popularity and broad usage, they have been the subject of much formal and informal research in language and

03

Clustering

From K-means to HDBSCAN

What is clustering?

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields.

Cluster analysis itself is not one specific algorithm, but the general task to be solved.

Subgroups

1 observation belongs to 1 cluster (possibility of not falling into a cluster at all).

Hard clustering

1 observation can belong to several clusters with some probability.

Soft clustering

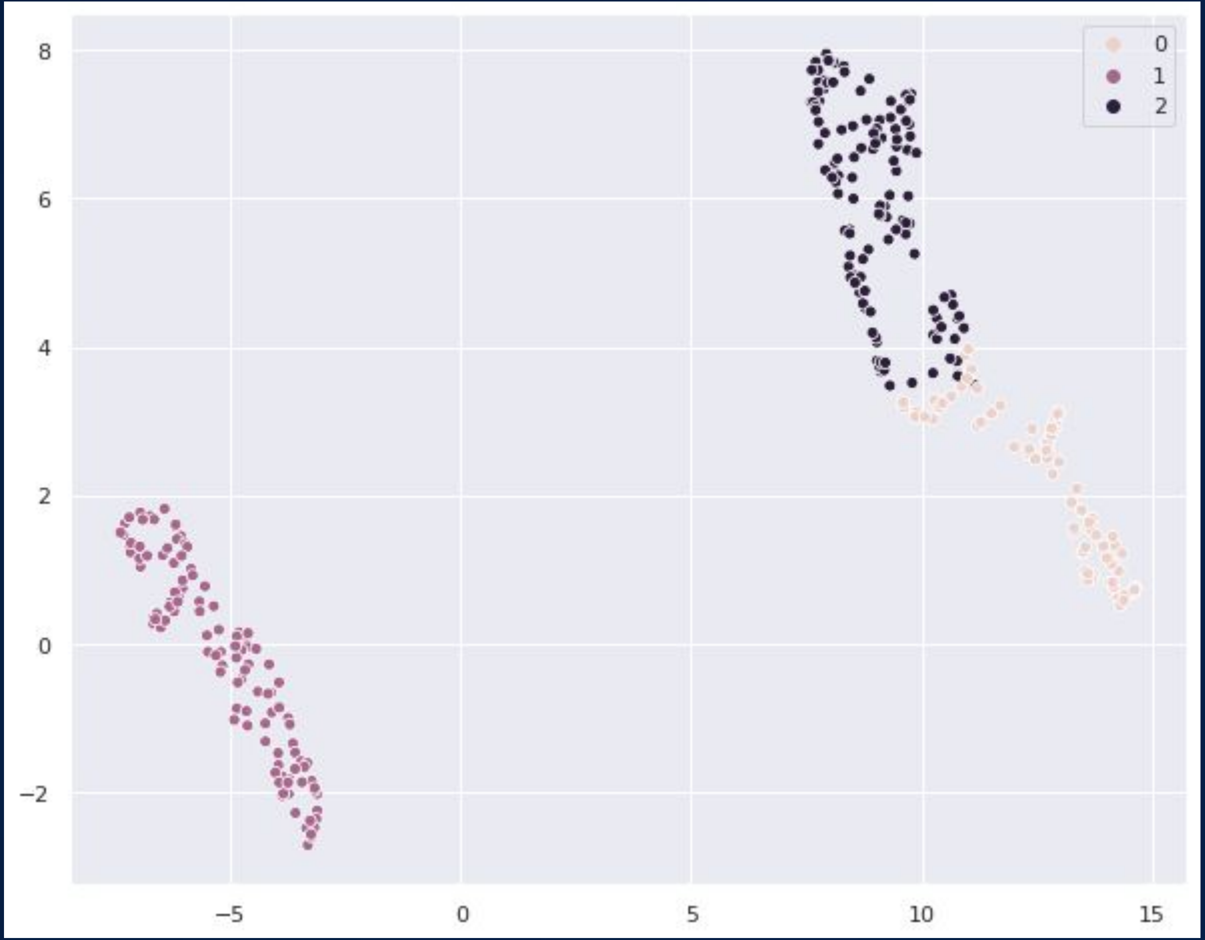
Clustering algorithms can also be categorized based on their cluster model, that is based on how they form clusters or groups.

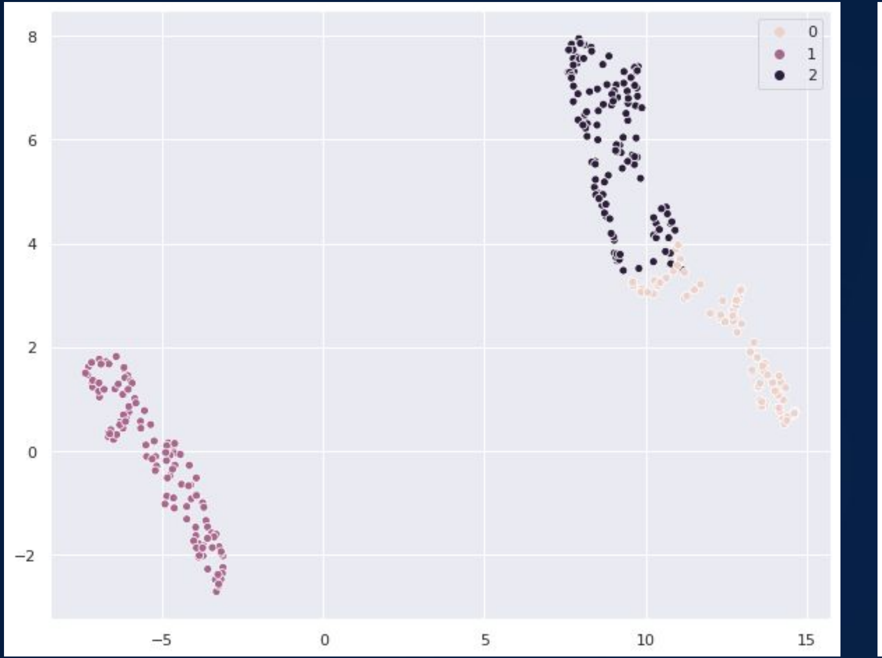
Connectivity-based clustering: the main idea behind this clustering is that data points that are closer in the data space are more related (similar) than to data points farther away.

Centroid-based clustering: in this type of clustering, clusters are represented by a central vector or a centroid. This centroid might not necessarily be a member of the dataset. **k-means** is a centroid based clustering, and will you see this topic more in detail later on in the tutorial.

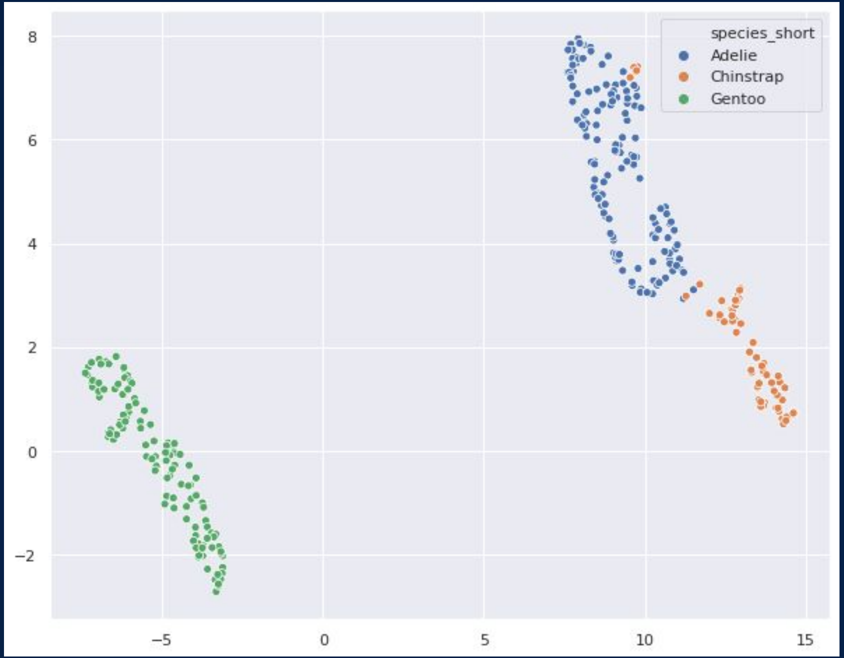
Distribution-based clustering, Density-based methods

K-means

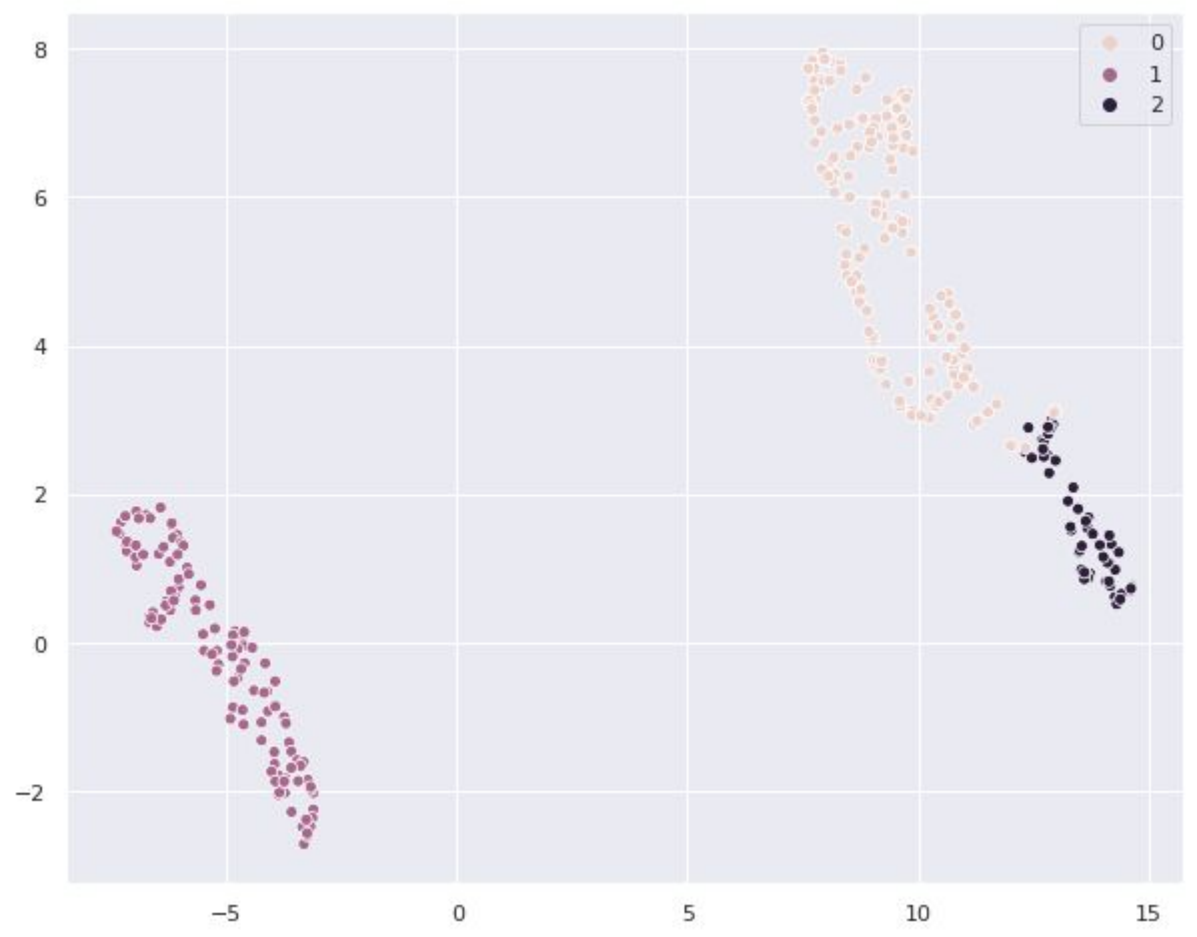




0
1
2
8
6
4
2
0
−2
−5
0
5
10
15



species_short
Adelie
Chinstrap
Gentoo
8
6
4
2
0
−2
−5
0
5
10
15



Hierarchical clustering

Some rules:

- Don't be wrong! - Better no result than a bad one and misunderstanding your data
- Intuitive parameters
- Stable clusters
- Performance

https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html?highlight=rules#some-rules-for-eda-clustering

THANKS

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.