

# Introduction to scikit-learn

EuroSciPy 2025  
18.08.2025, Kraków

Justyna Szydłowska-Samsel, University of Szczecin

Co-financed by the Minister of Science under the "Regional Excellence Initiative"



Minister of Science and Higher Education  
Republic of Poland



# Tutorial agenda

What are we going to talk about?

1. Introduction
2. Why Python?
3. What is Scikit-learn?
4. What is machine learning?
5. Tutorial



# Who am I?

- Research Assistant at University of Szczecin, Institute of Management
- Research focus : machine learning in business application and methods of teaching programming



# Why Python?

*Easy to learn*

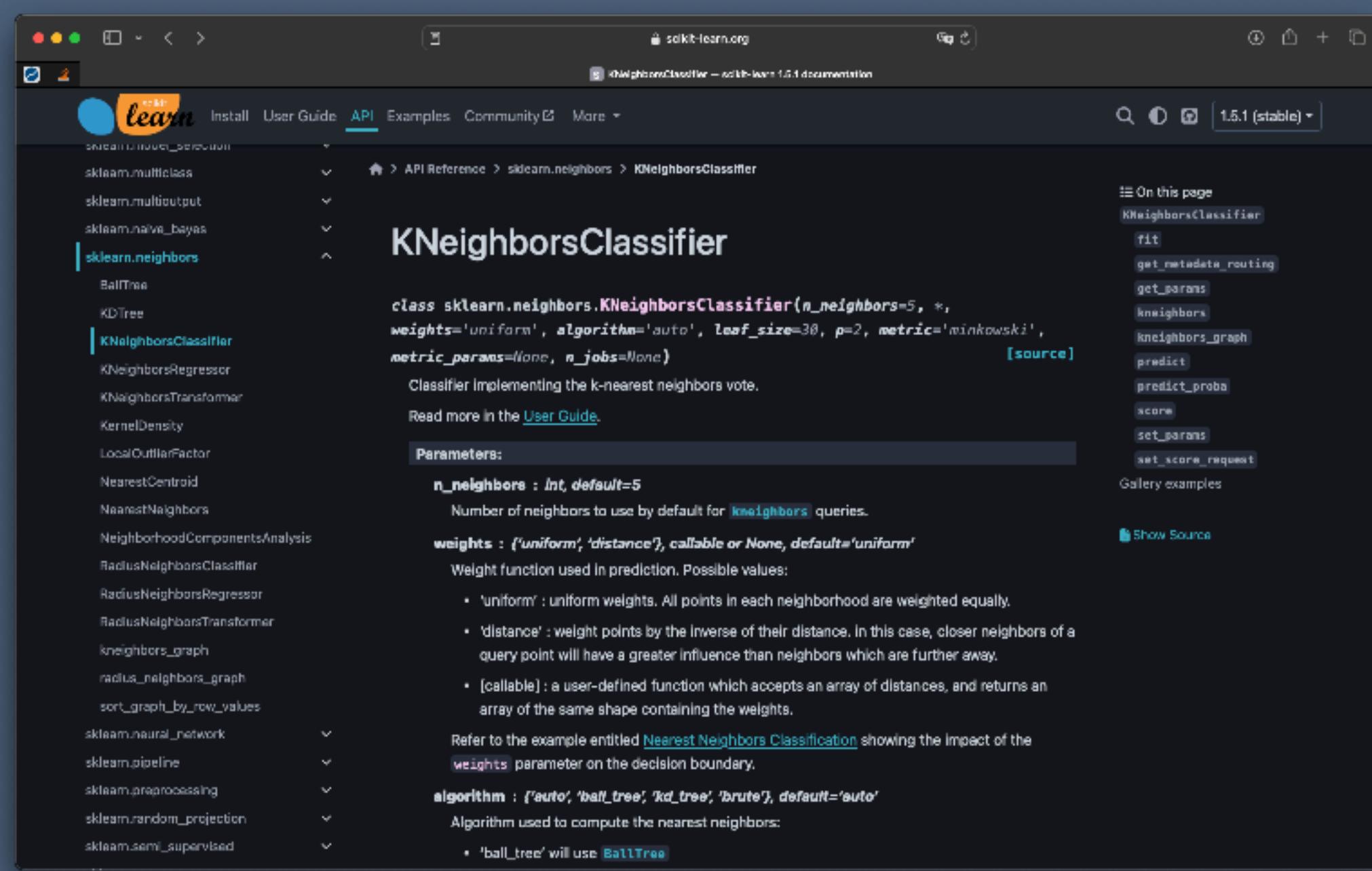
*Powerful*

*Free*



# Scikit-learn

Free and open-source machine learning library for Python. It's build on NumPy, SciPy and matplotlib.



This screenshot shows a browser window displaying the scikit-learn API documentation for the `KNeighborsClassifier` class. The page includes the class definition, parameters, and a detailed description of the `n_neighbors` parameter. A sidebar on the left lists other classes in the `sklearn.neighbors` module.

```
class sklearn.neighbors.KNeighborsClassifier(n_neighbors=5, *, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None)
```

`n_neighbors`: `int, default=5`  
Number of neighbors to use by default for `kneighbors` queries.

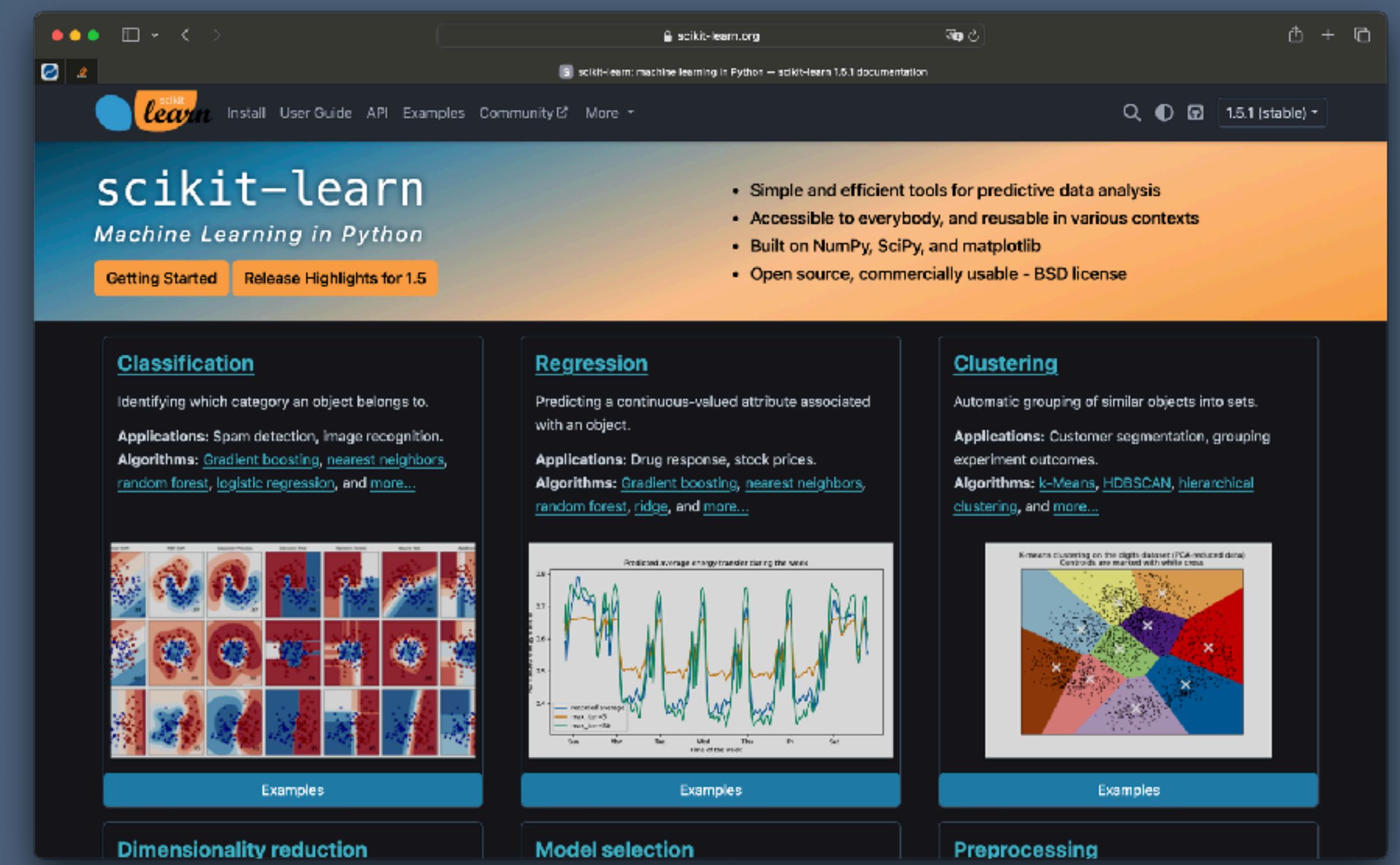
`weights`: `{'uniform', 'distance'}, callable or None, default='uniform'`  
Weight function used in prediction. Possible values:

- 'uniform': uniform weights. All points in each neighborhood are weighted equally.
- 'distance': weight points by the inverse of their distance. In this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.
- [callable]: a user-defined function which accepts an array of distances, and returns an array of the same shape containing the weights.

Refer to the example entitled [Nearest Neighbors Classification](#) showing the impact of the `weights` parameter on the decision boundary.

`algorithm`: `{'auto', 'ball_tree', 'kd_tree', 'brute'}, default='auto'`  
Algorithm used to compute the nearest neighbors:

- 'ball\_tree' will use `BallTree`



This screenshot shows the main homepage of the scikit-learn documentation. It features a large orange header with the `scikit-learn` logo. Below the header, there are sections for Classification, Regression, Clustering, Dimensionality reduction, Model selection, and Preprocessing, each with a corresponding image and brief description. The top navigation bar includes links for Install, User Guide, API, Examples, Community, More, and a search bar.

Simple and efficient tools for predictive data analysis  
Accessible to everybody, and reusable in various contexts  
Built on NumPy, SciPy, and matplotlib  
Open source, commercially usable - BSD license

Classification  
Identifying which category an object belongs to.  
Applications: Spam detection, image recognition.  
Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...

Regression  
Predicting a continuous-valued attribute associated with an object.  
Applications: Drug response, stock prices.  
Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...

Clustering  
Automatic grouping of similar objects into sets.  
Applications: Customer segmentation, grouping experiment outcomes.  
Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...

Dimensionality reduction  
Examples

Model selection  
Examples

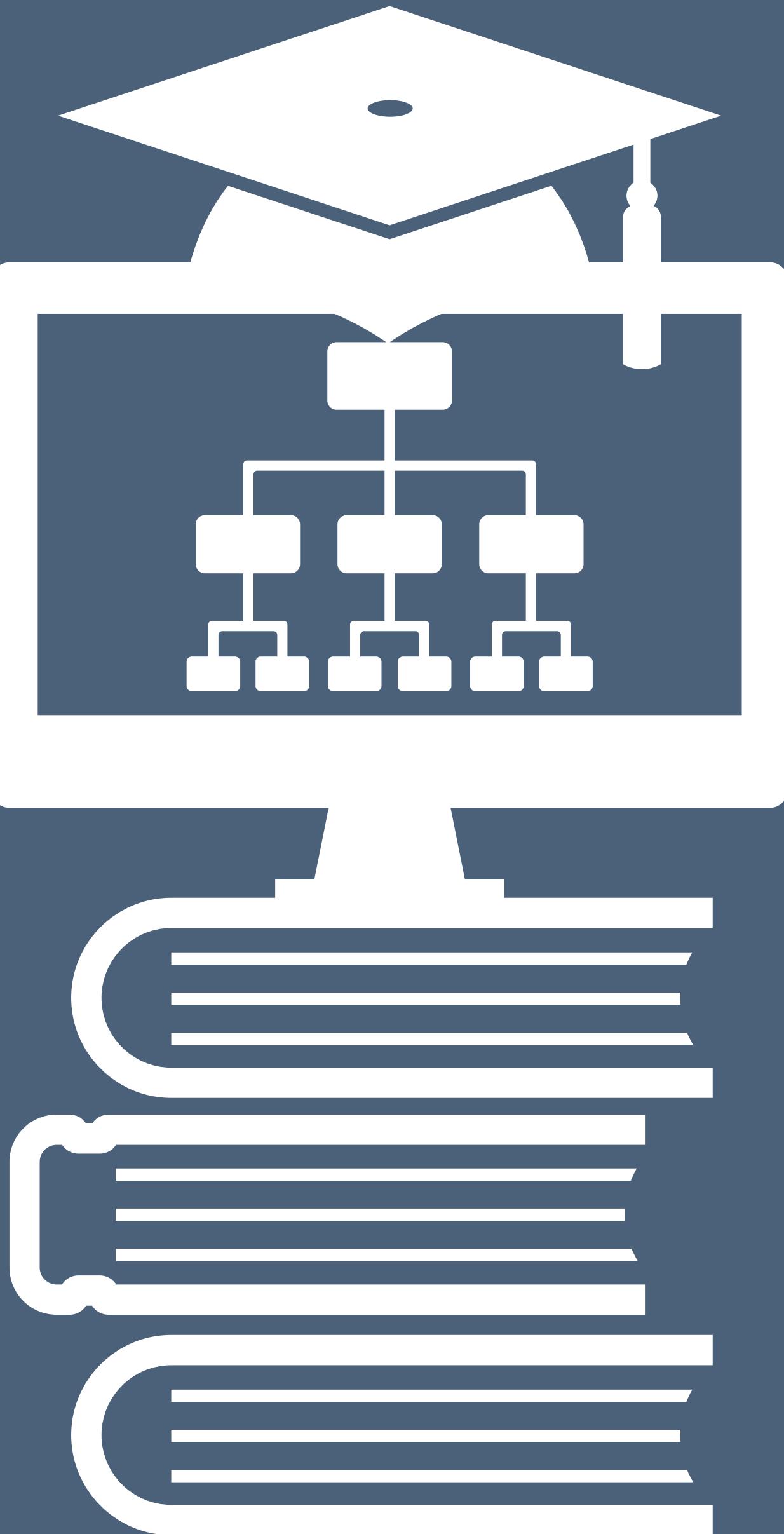
Preprocessing  
Examples

# Machine learning

What is it and why is it so popular?

„Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy.” (IMB, 2024)

What usually follows is prediction. It is widely used in business, examples: Google, Netflix, Government institutions, banks, etc.



# Types of machine learning

- › Supervised
- › Unsupervised
- › Self-supervised
- › Reinforcement
- › Semi-supervised

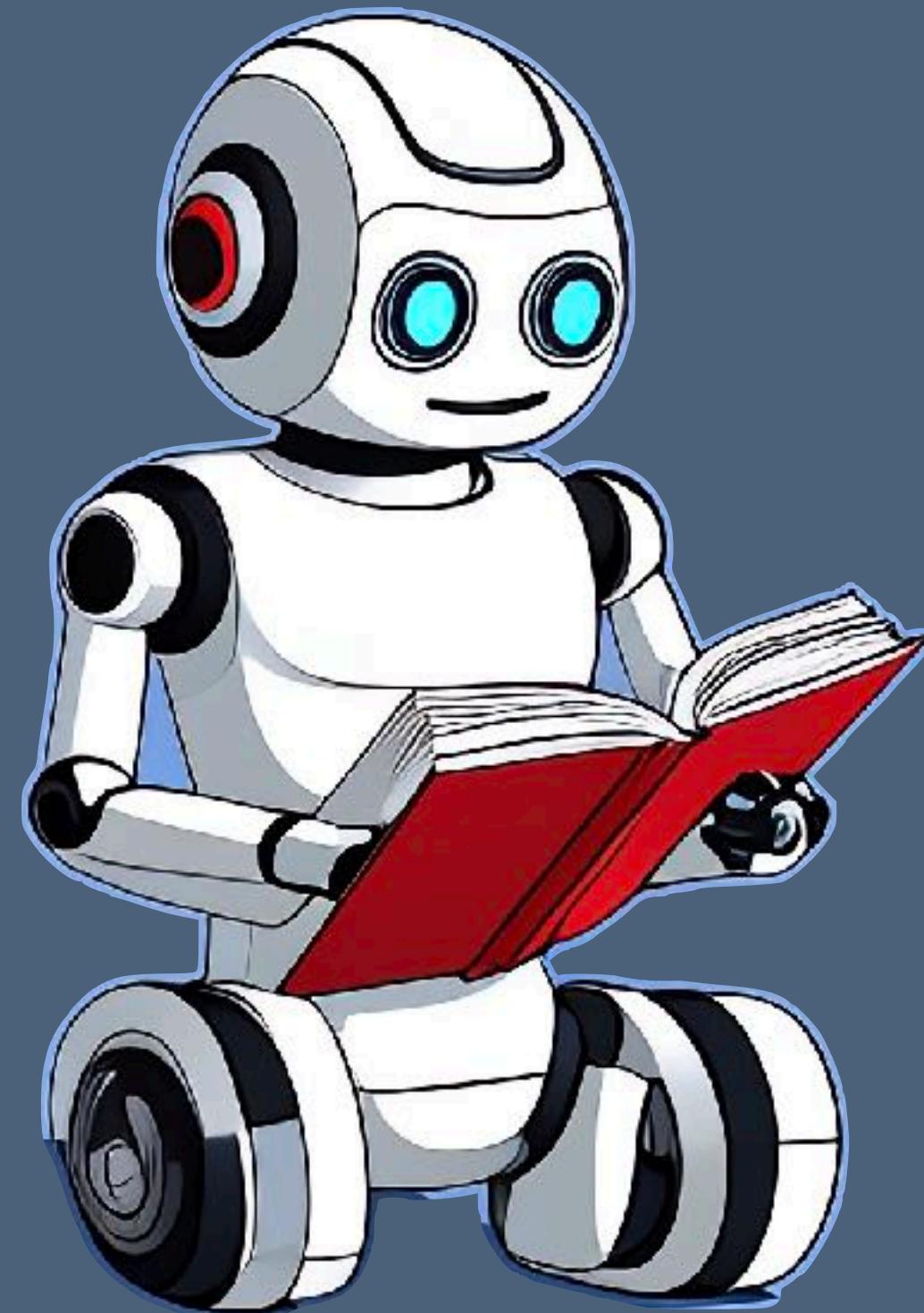


# Supervised machine learning

It's a type of machine learning where model is trained on labeled dataset

Supervised machine learning algorithms:

- Regression algorithms
- Classification algorithms
- Neural networks
- Random forest algorithms

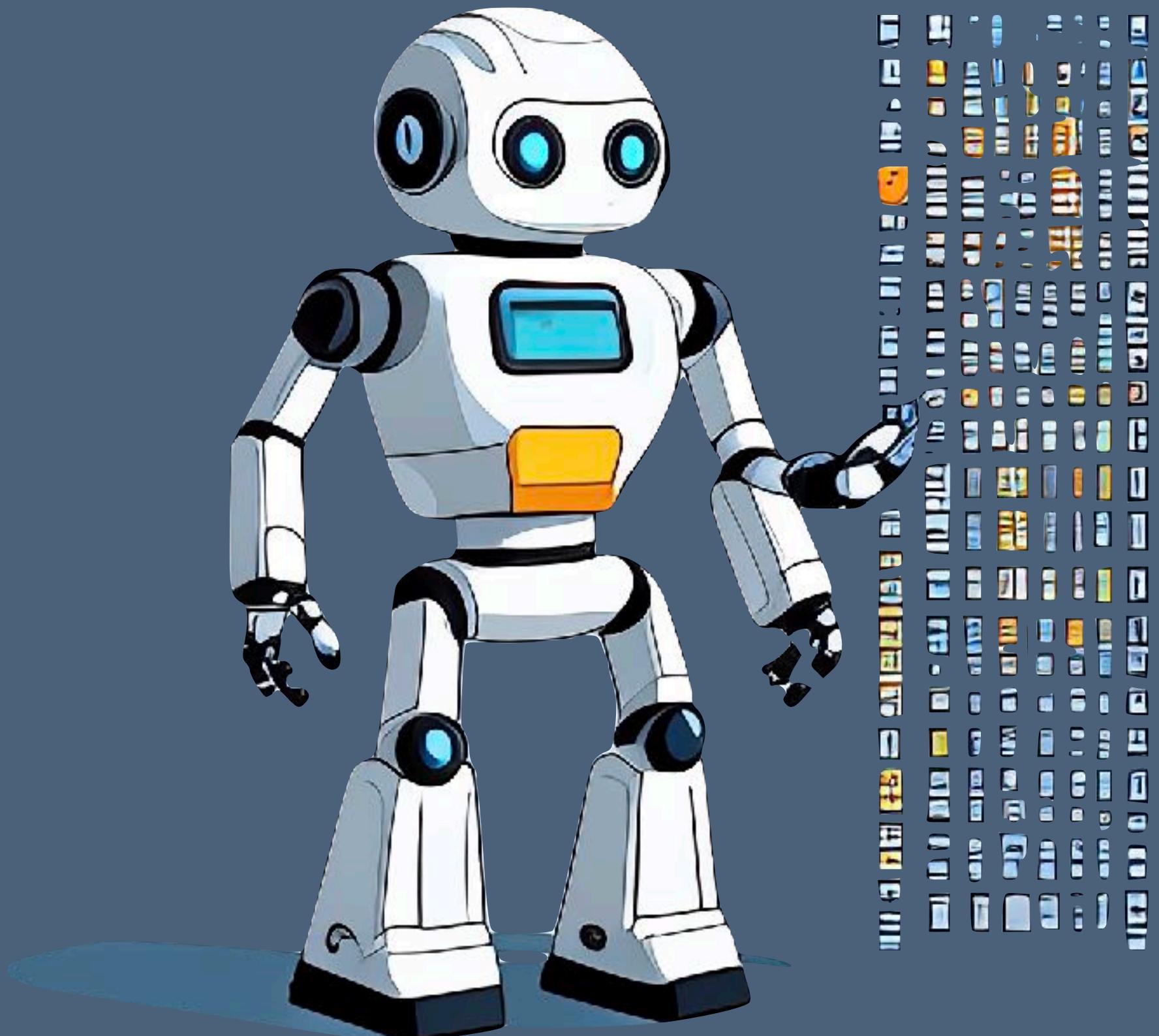


# Unsupervised machine learning

It's a type of machine learning where model is trained on unlabeled dataset - trying to find patterns without explicit help.

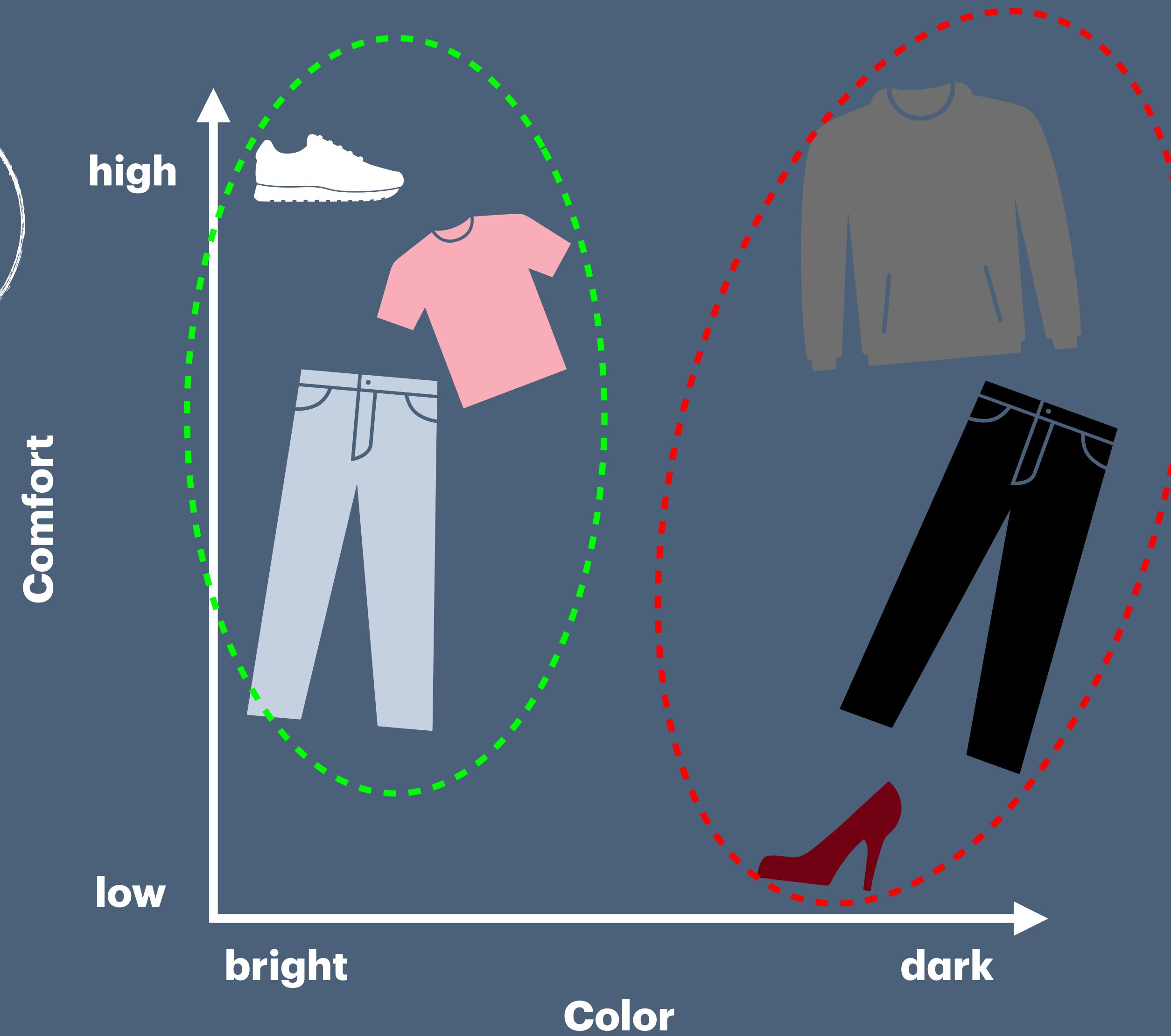
Unsupervised machine learning algorithms:

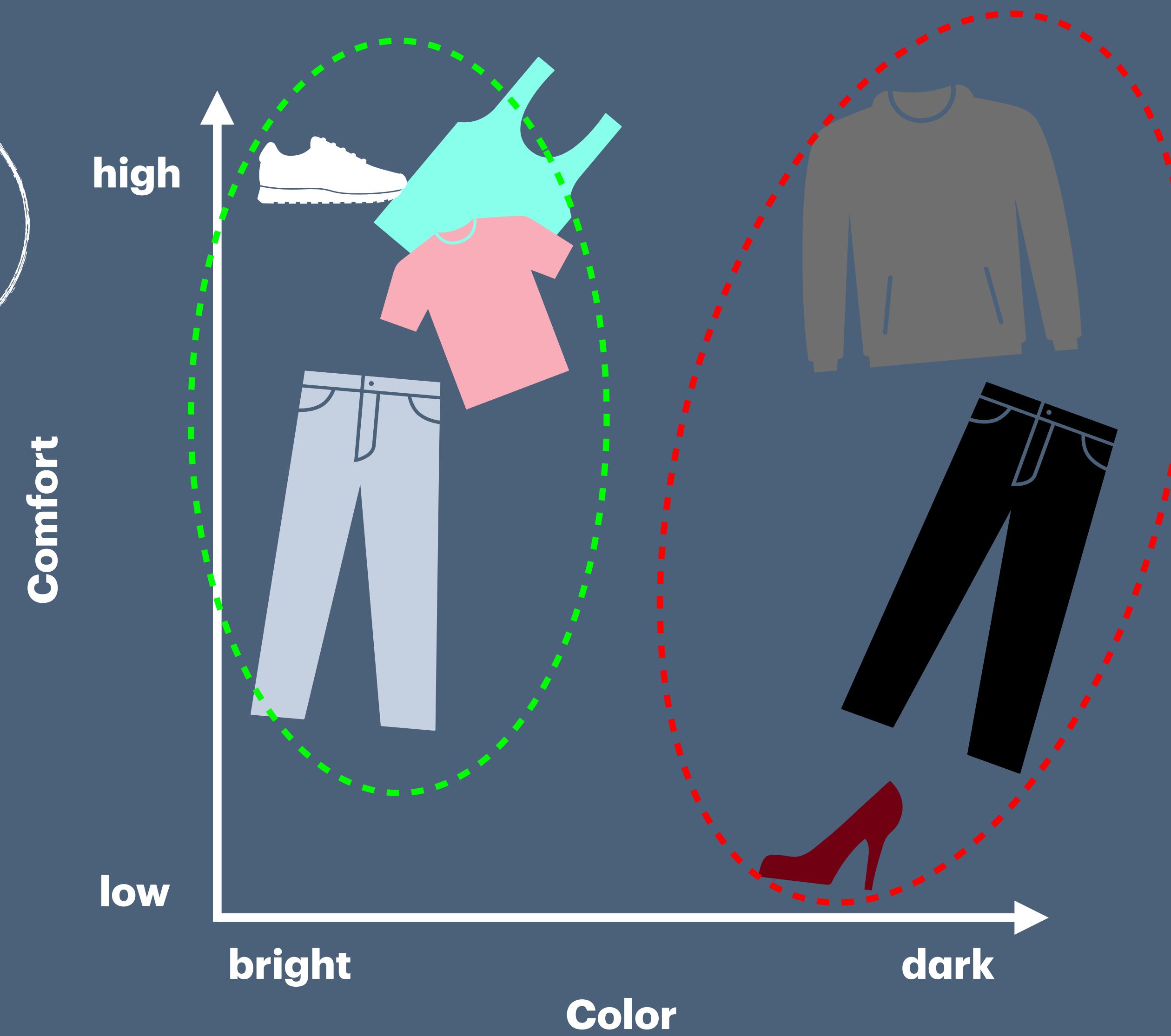
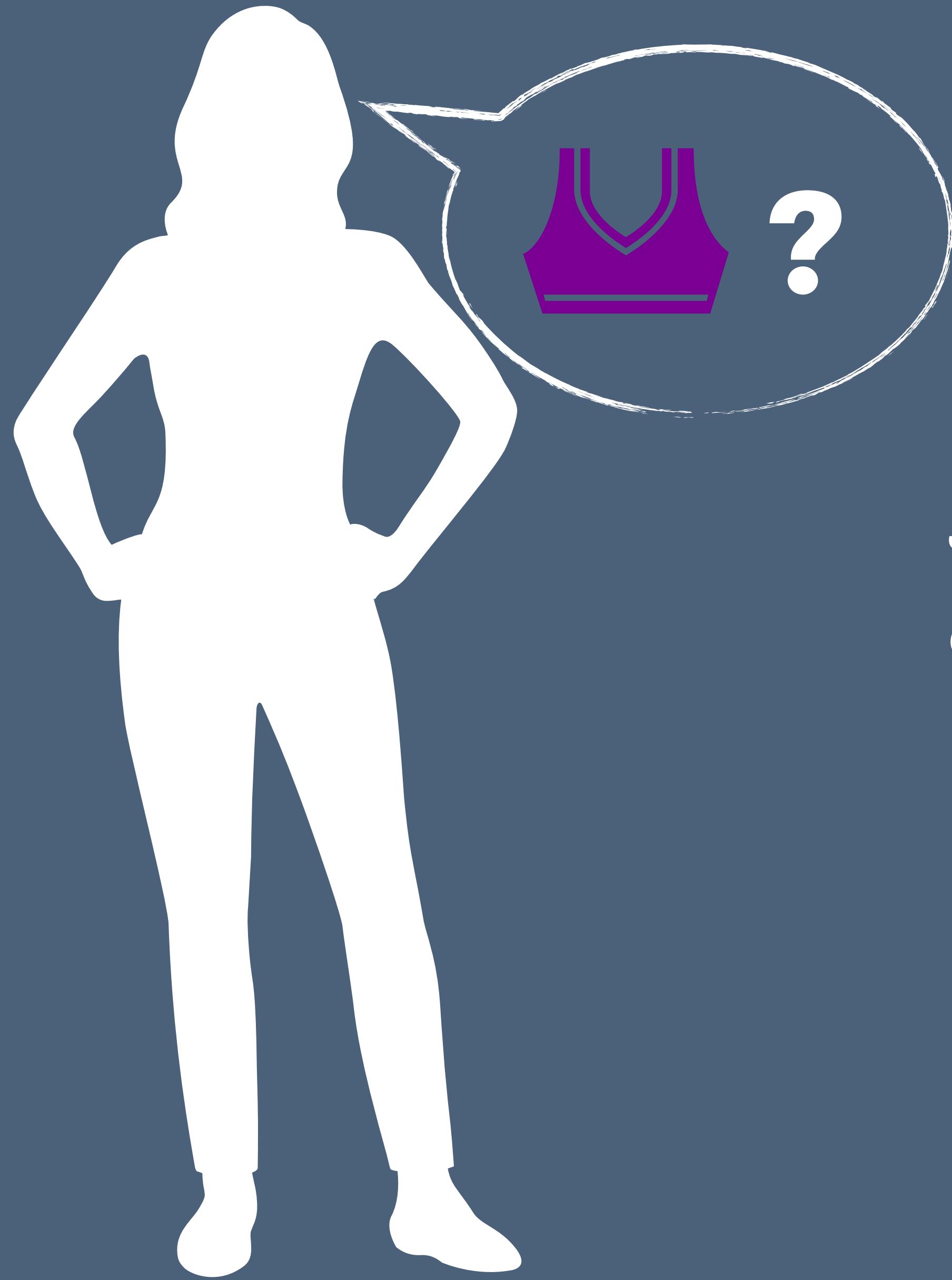
- K-means clustering
- Spectral clustering
- Hierarchical clustering

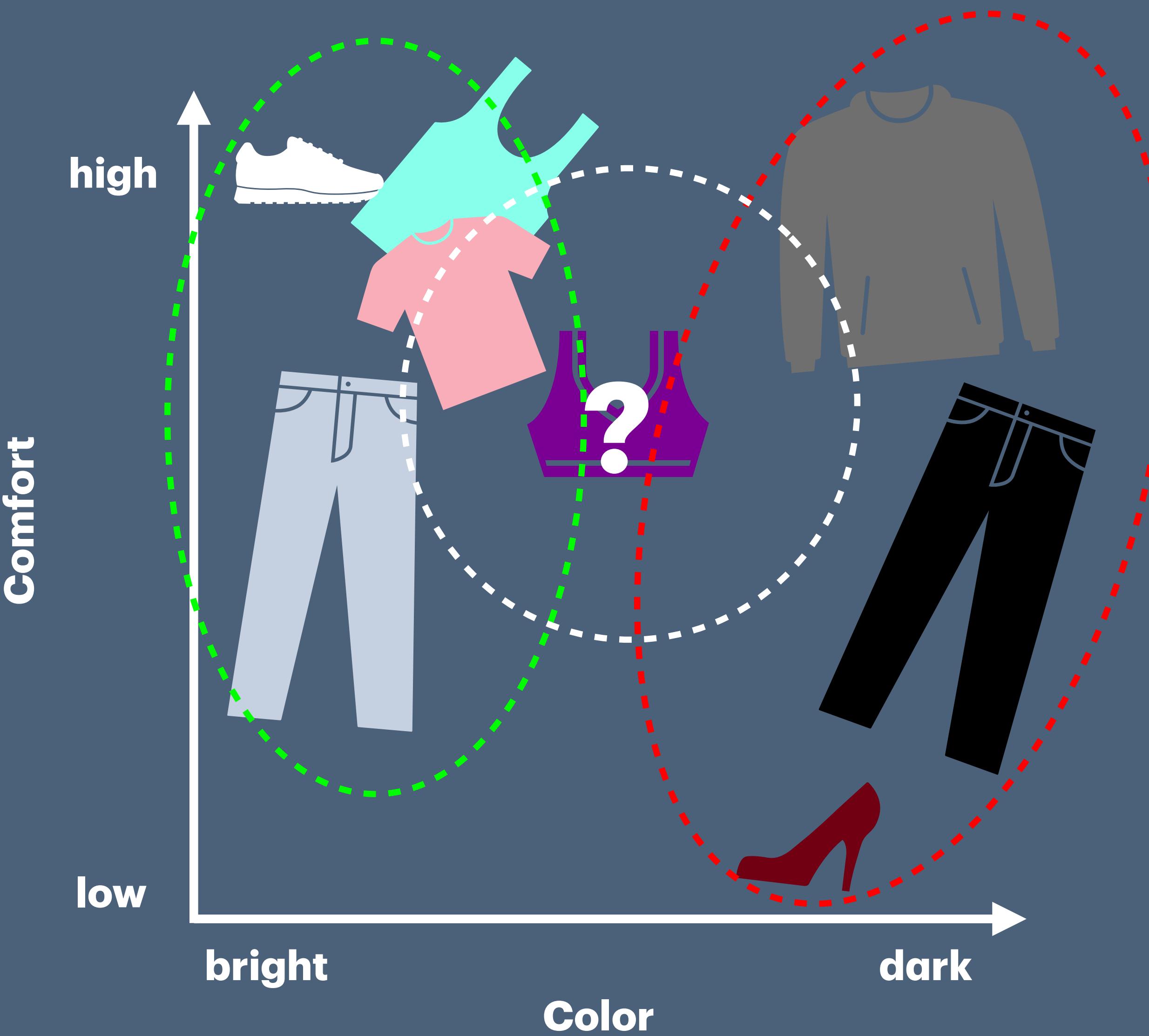


# Which clothes will she buy?

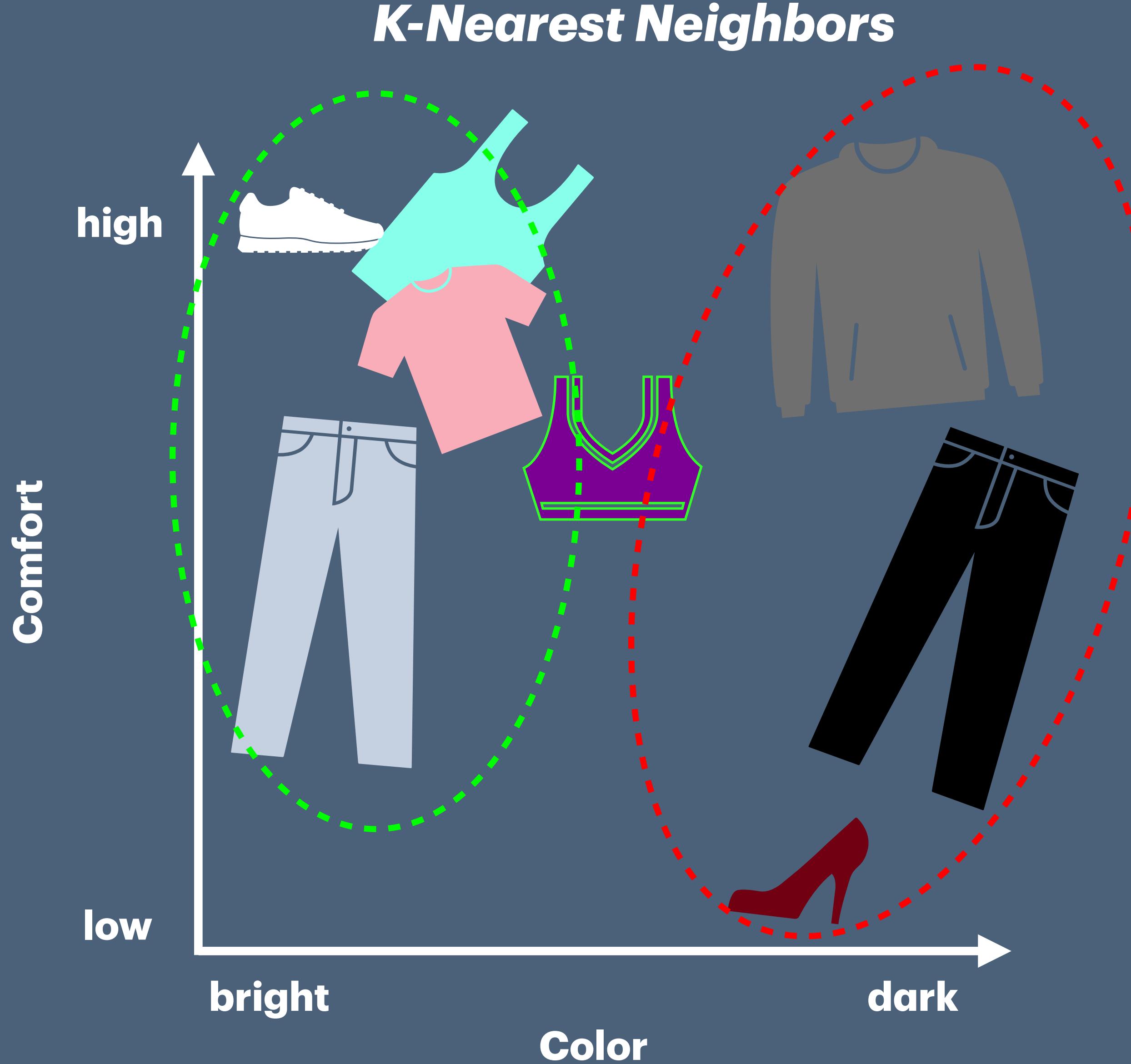








# K-Nearest Neighbors



# Supervised machine learning

It's a type of machine learning where model is trained on labeled dataset

Supervised machine learning algorithms:

- Regression algorithms
- Classification algorithms
- Neural networks
- Random forest algorithms

# Supervised machine learning

- Regression algorithms
- Classification algorithms

# Supervised machine learning

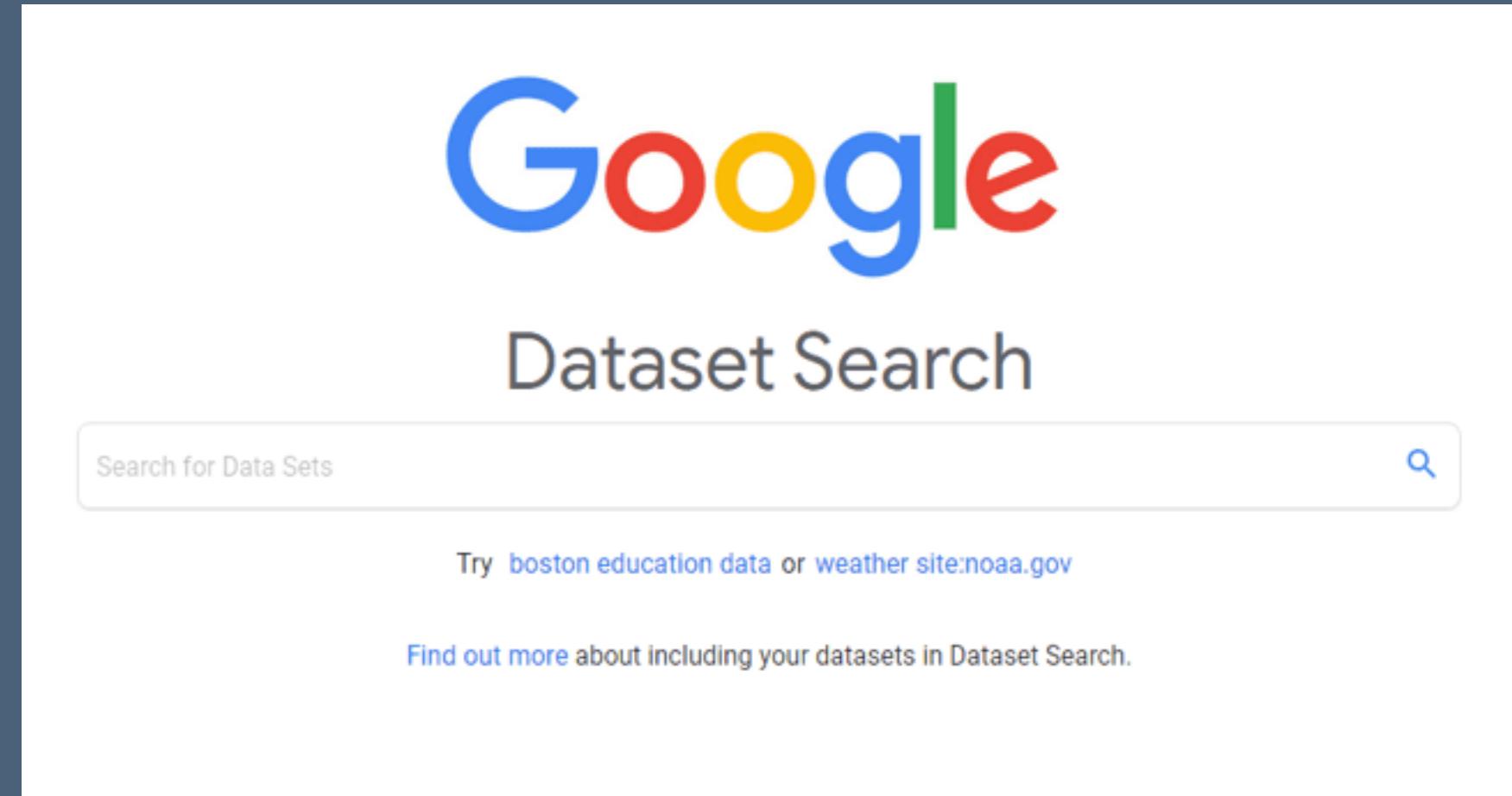
Regression:

- When what we want to predict is a continuous numerical value

Classification algorithms

- When what we want to predict is a category

# Where to get data?



The homepage of Poland's Data Portal. At the top, it features the "OTWARTE DANE" logo and a navigation menu with links to "Homepage", "Data", "Publishers", "PoCoTo", "News", "Knowledge base", and "DGA Information Point". The main header reads "Poland's Data Portal" and "Use public data free of charge, also for commercial purposes". A search bar with the placeholder "Enter search query..." is positioned above a map of Poland. Below the map, there is a callout for "Developers, check how to share data on apartment prices" with a small house icon. The footer contains logos for "Fundusze Europejskie na Rozwój Cyfrowy", "Rzeczypospolita Polska", "Dofinansowane przez Unię Europejską", and "DANE 3.0 Wymiana Wartość". A survey invitation at the bottom says "Complete the survey! Thanks to you, we will open more data in the portal" with a checkmark icon.

# Tutorial

# Pandas



Free and open source Python library for data manipulation and analysis.

This screenshot shows the User Guide section of the pandas documentation. The left sidebar contains a navigation menu with links like '10 minutes to pandas', 'Intro to data structures', 'IO tools', 'PyArrow Functionality', etc. The main content area features a 'User Guide' heading, a brief introduction, and sections on 'How to read these guides' and code snippets. A search bar at the top right allows users to search within the documentation.

This screenshot shows the pandas homepage. It features a large 'pandas' logo at the top, followed by a brief introduction: 'pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.' Below this is a 'Install pandas now!' button. The page is divided into three main sections: 'Getting started', 'Documentation', and 'Community'. Each section has a list of links. To the right, there's a 'Latest version: 2.2.2' section with a link to the release notes, and social media links for GitHub, LinkedIn, and Twitter. At the bottom, there's a 'With the support of:' section with logos from Intel, Tidelift, Chan Zuckerberg Initiative, and bodo.ai, along with book covers for 'Python for Data Analysis' and 'Effective Pandas 2'.

[https://github.com/  
jszydłowska/EuroSciPy2025.git](https://github.com/jszydłowska/EuroSciPy2025.git)