

# Introduction to scikit-learn

EuroSciPy 2025

18.08.2025, Kraków

Justyna Szydłowska-Samsel, University of Szczecin

# 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 the scikit-learn API Reference page for the `KNeighborsClassifier` class. The page includes a sidebar with a tree navigation of the API structure, the class documentation with its code implementation, and detailed parameters and methods. A "Show Source" button is also present.

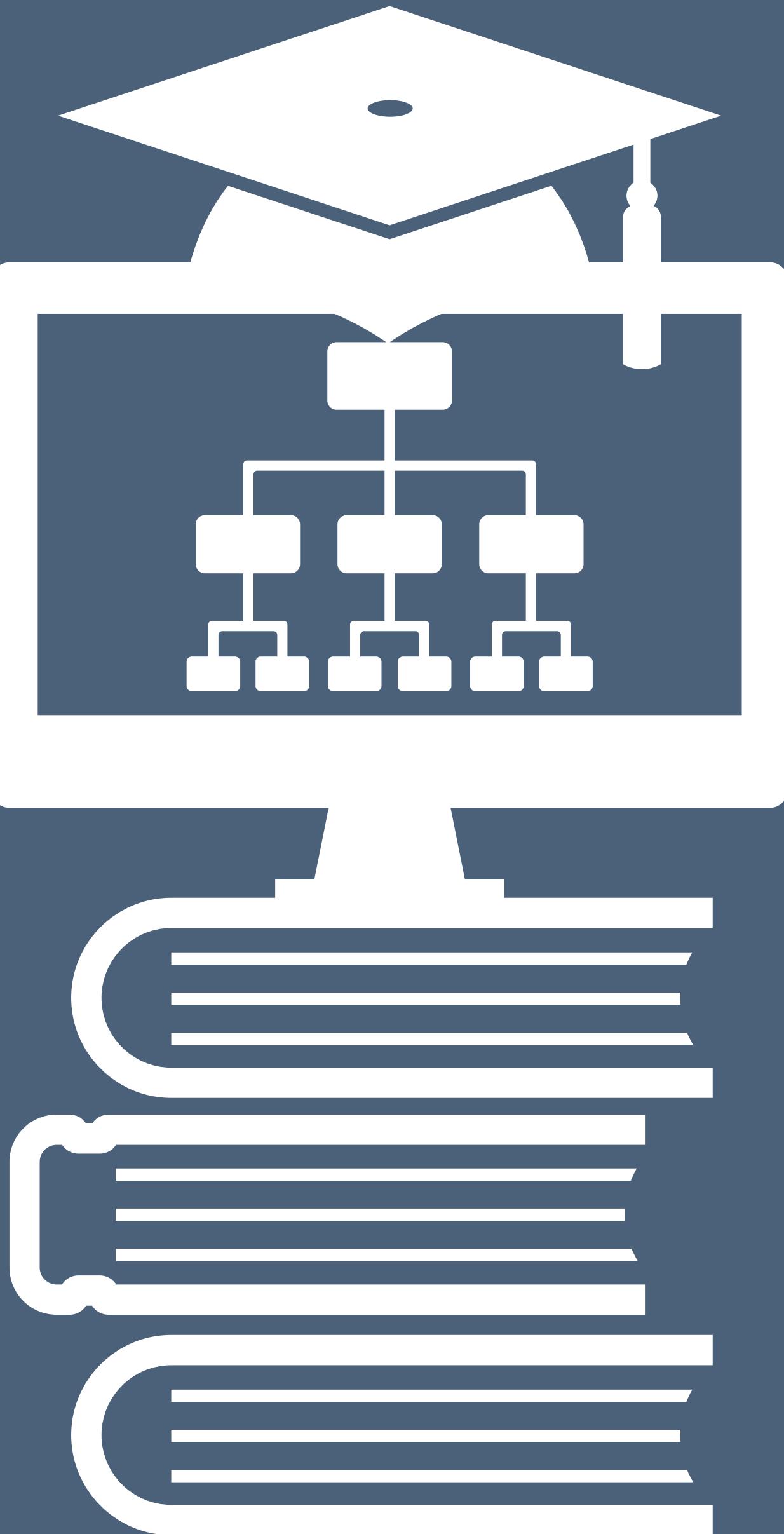
This screenshot shows the main scikit-learn documentation homepage. It features a large orange header with the "scikit learn" logo. Below the header, there's a brief introduction, a "Getting Started" button, and a "Release Highlights for 1.5" button. The main content area is divided into several sections: Classification, Regression, Clustering, Dimensionality reduction, Model selection, and Preprocessing, each with associated images and text. A sidebar on the left contains links to various parts of the documentation.

# 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

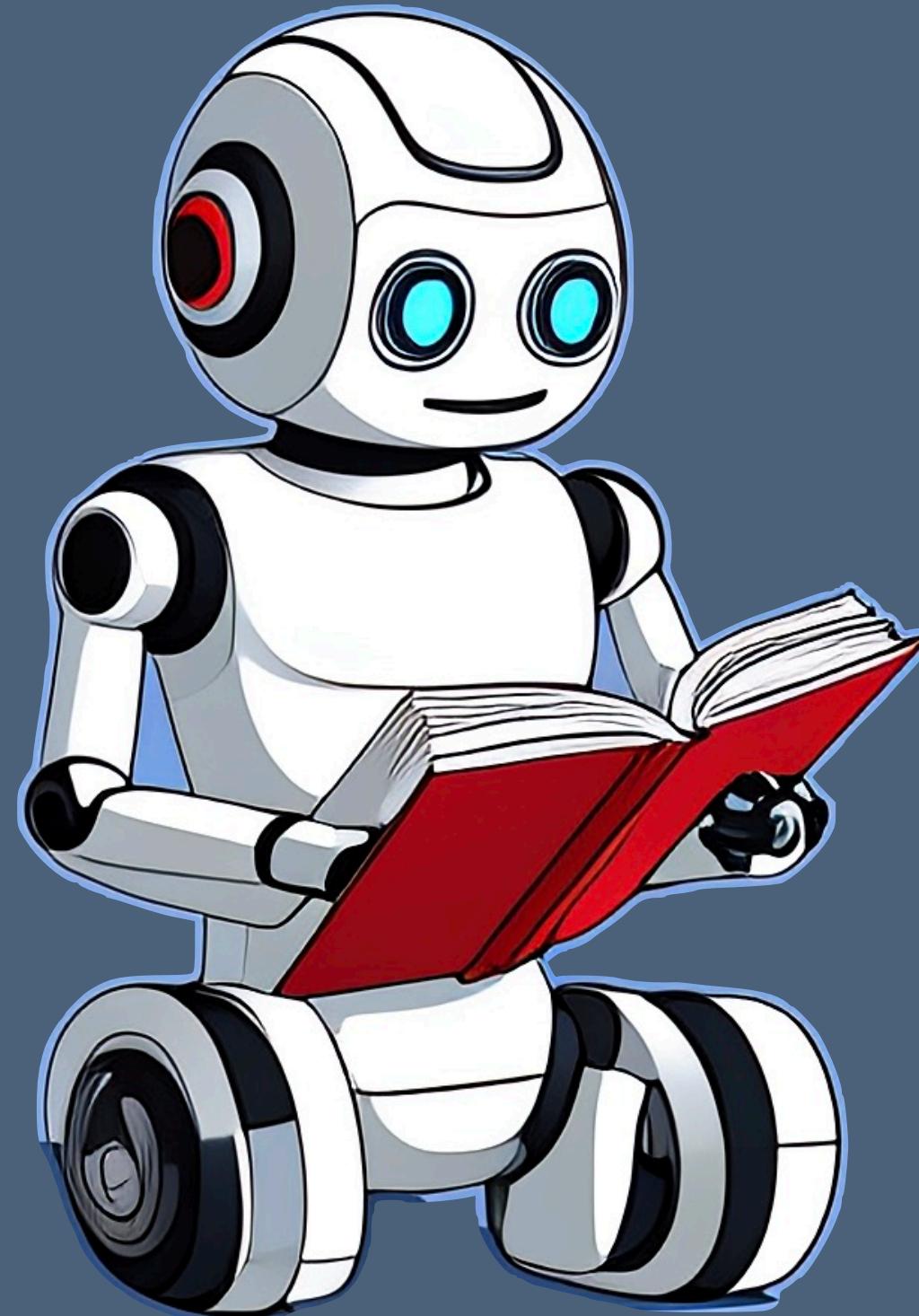


# Supervised machine learning

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

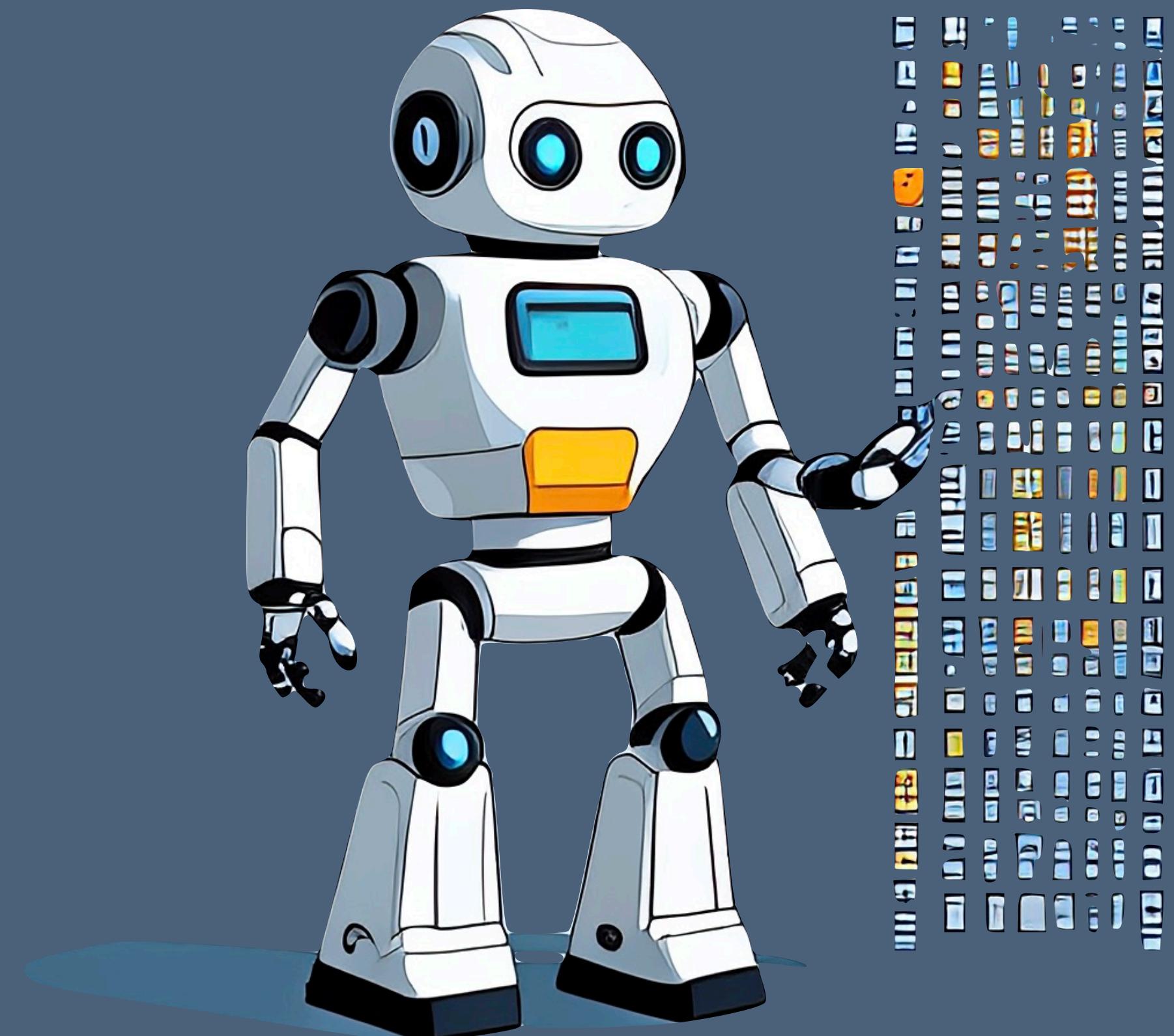


# Unsupervised machine learning

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

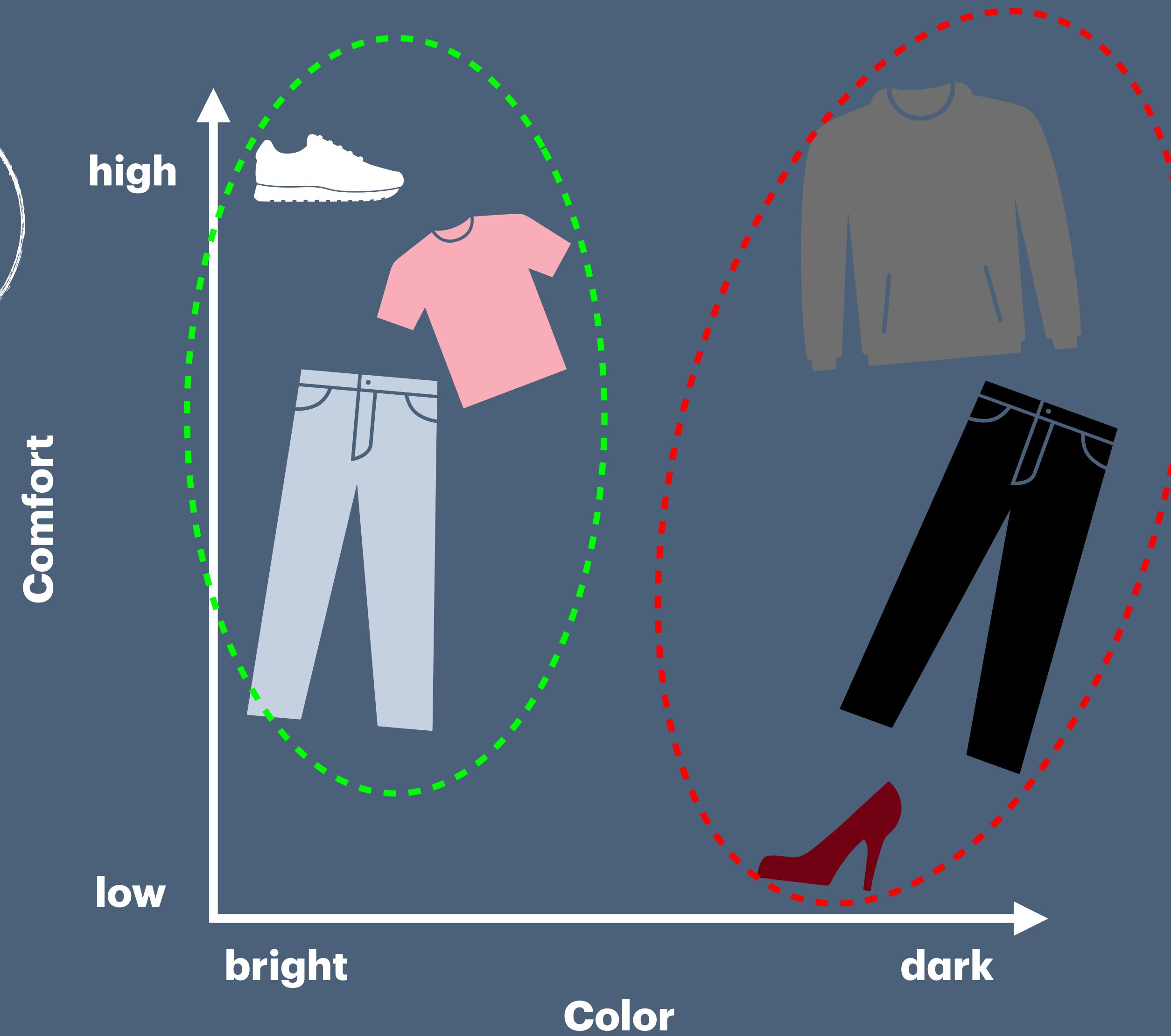
Unsupervised machine learning algorithms:

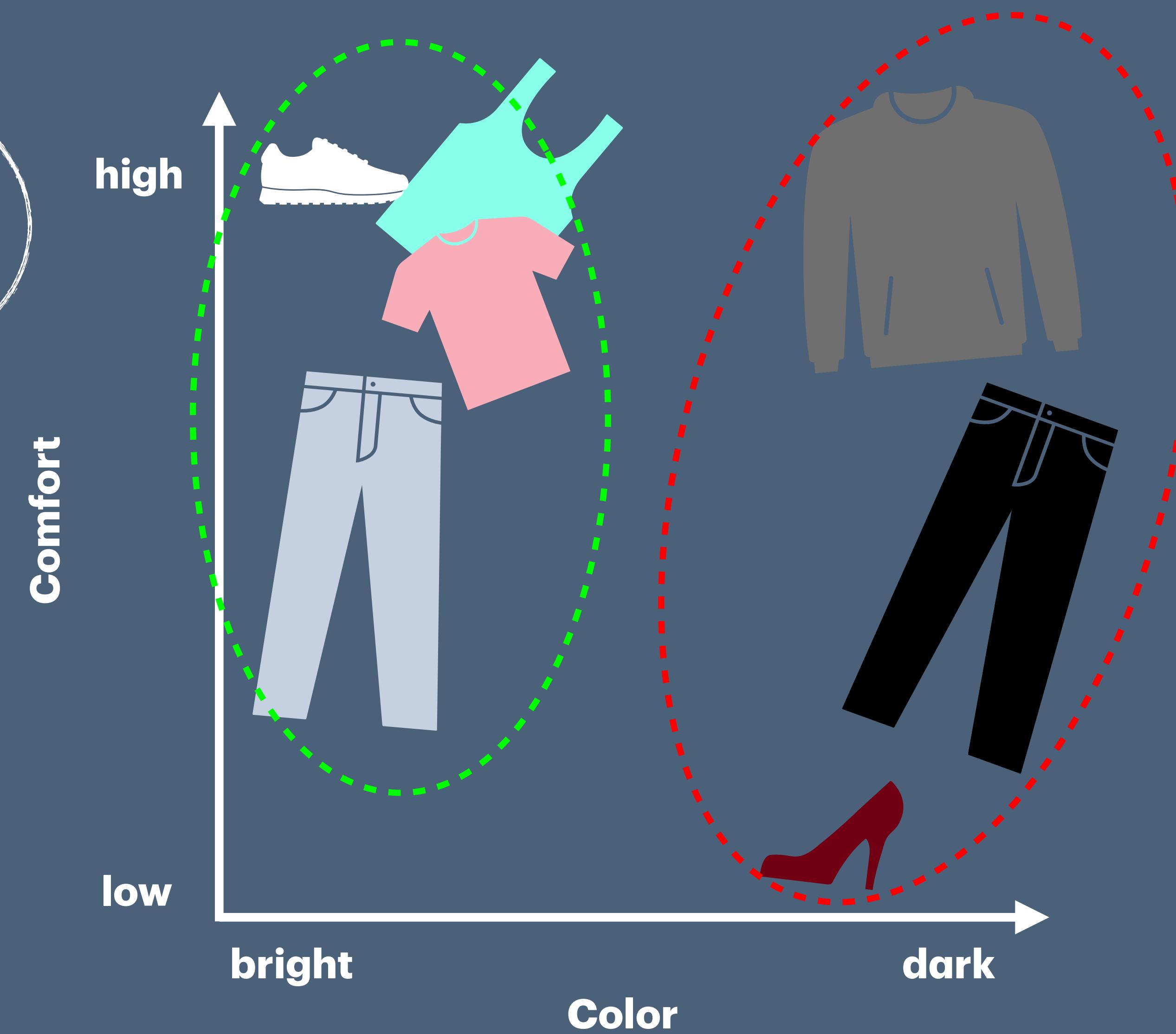
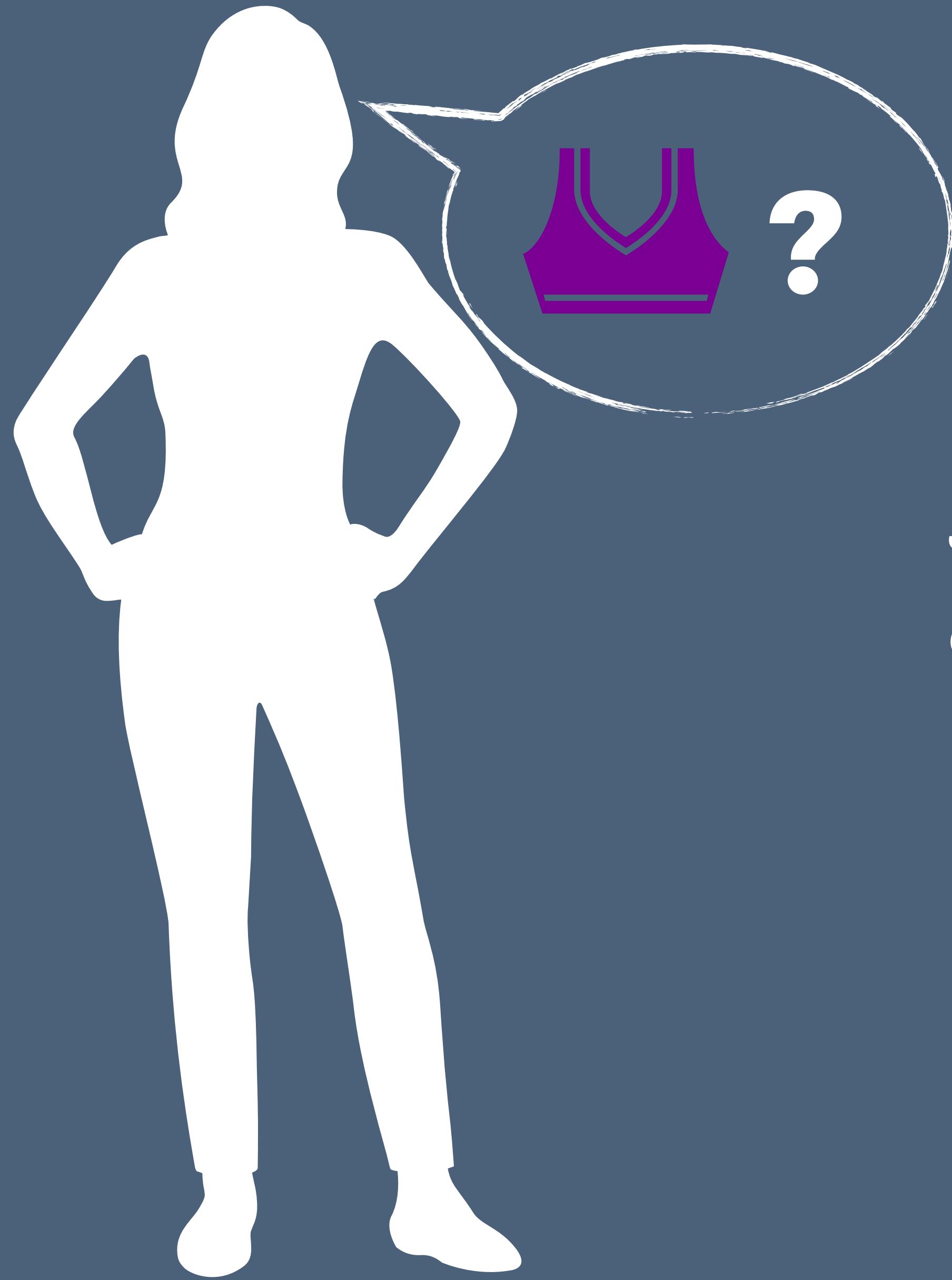
- K-means clustering
- Spectral clustering
- Hierarchical clustering

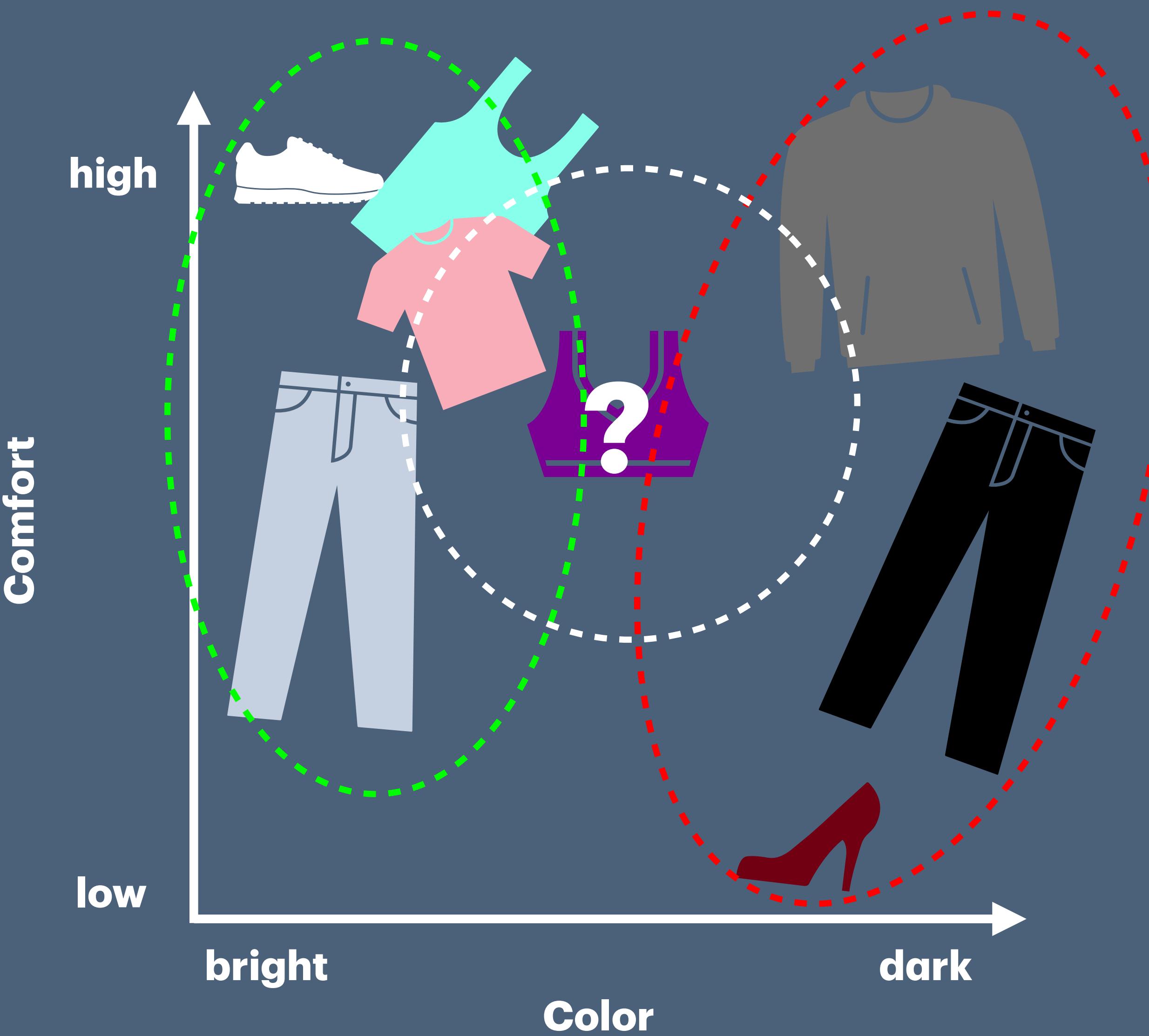


# Which clothes will she buy?

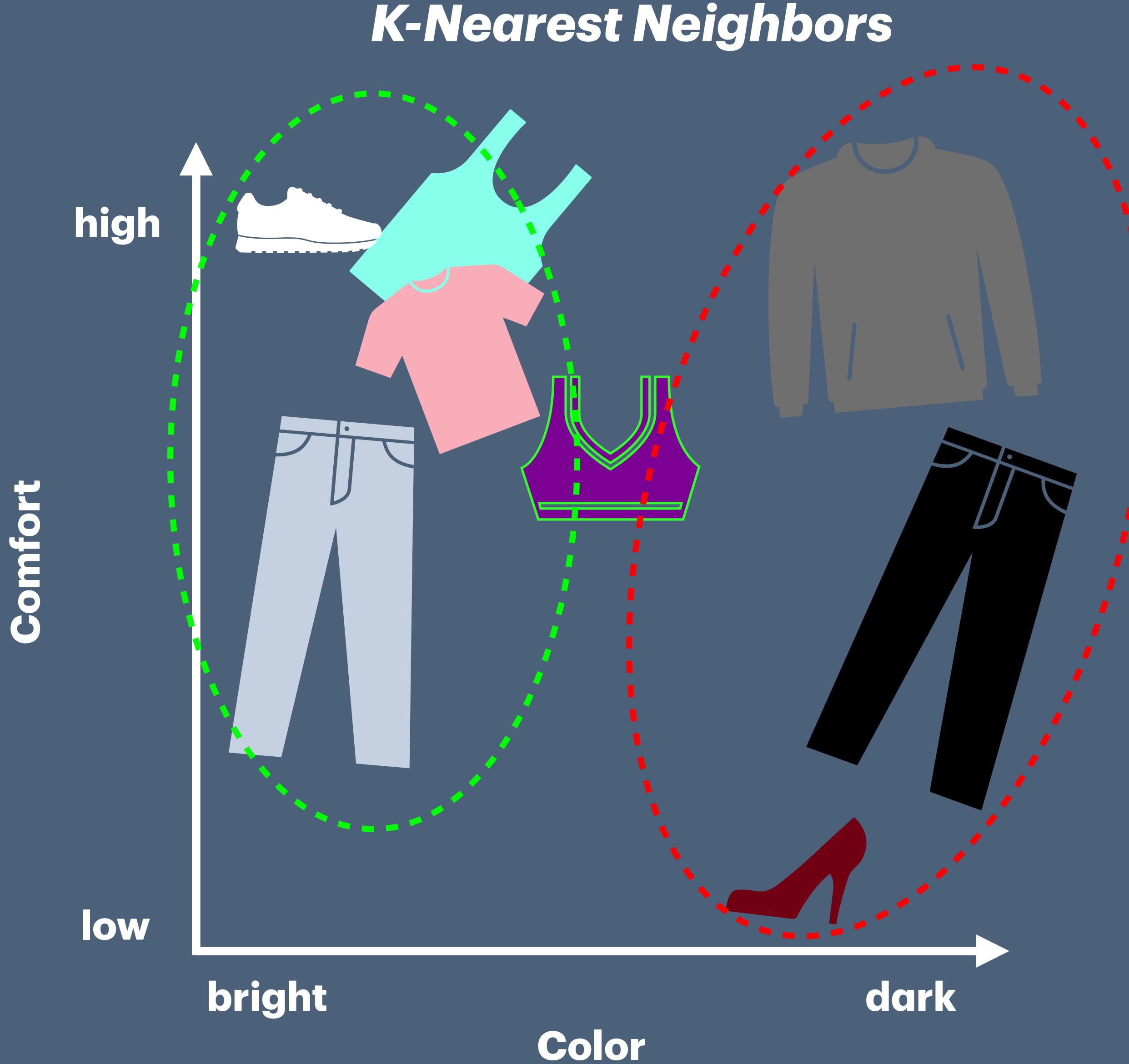








# K-Nearest Neighbors



# 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

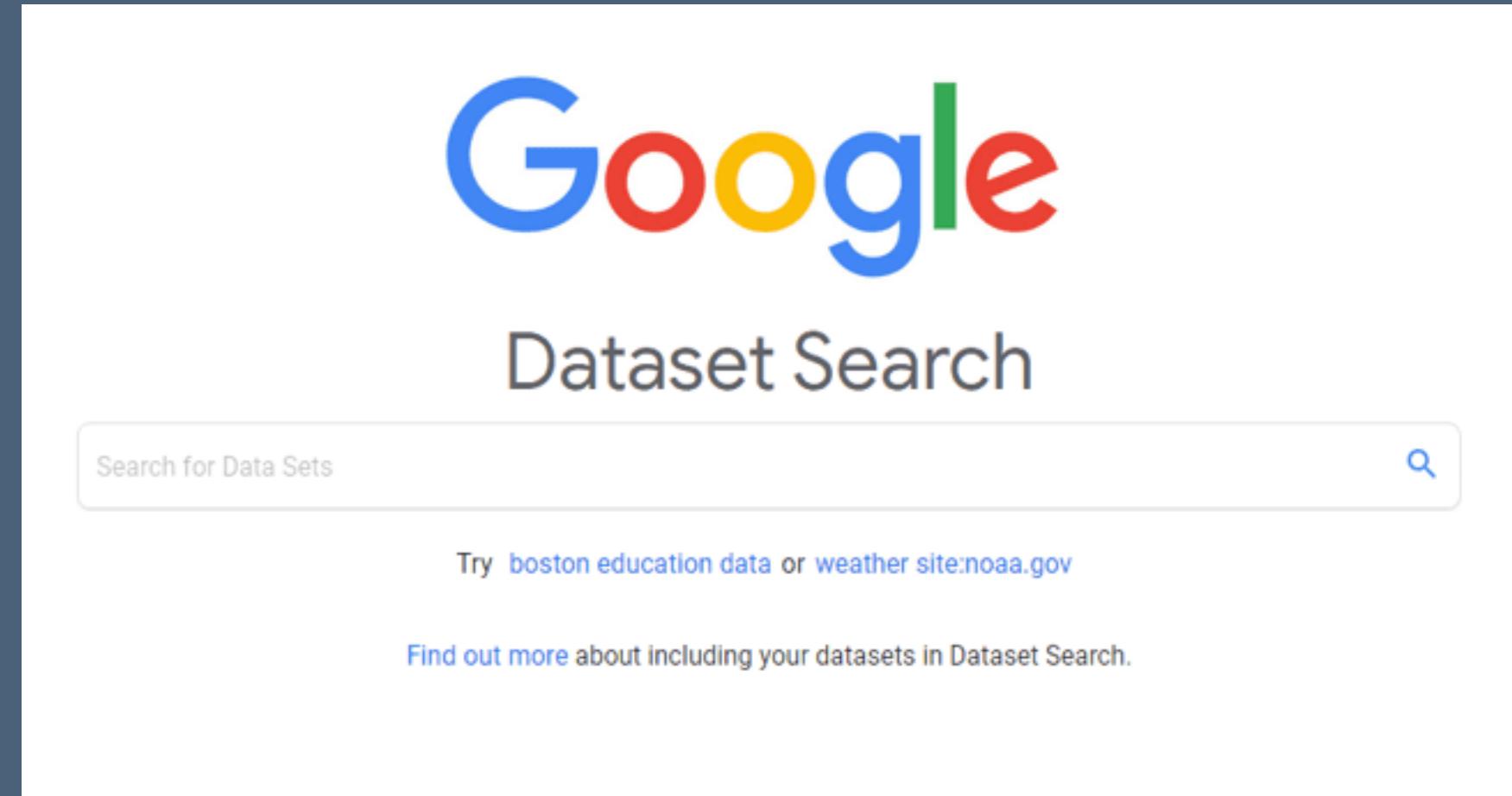
Regression:

- 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 screenshot shows the homepage of Poland's Data Portal (OTWARTE DANE). The header includes the logo "D/ OTWARTE DANE" and a navigation menu with links to "Homepage", "Data", "Publishers", "PoCoTo", "News", "Knowledge base", and "DGA Information Point". The main banner features a map of Poland and the text "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 present. Below the banner, there is a callout for developers with the text "Developers, check how to share data on apartment prices" and 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 encourages users to "Complete the survey! Thanks to you, we will open more data in the portal".

# Tutorial

# Pandas



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

The screenshot shows the User Guide for pandas 2.2.2. The left sidebar lists various topics such as '10 minutes to pandas', 'Intro to data structures', and 'Time series / date functionality'. The main content area features a 'User Guide' section with a sub-section 'How to read these guides' containing code snippets and instructions. A search bar and a navigation bar at the top provide access to the API reference, Development, and Release notes.

The screenshot shows the pandas homepage. The header includes the pandas logo and navigation links for About us, Getting started, Documentation, Community, and Contribute. The main content highlights the latest version (2.2.2) and provides a link to install pandas. Below this, there are sections for 'Getting started', 'Documentation', and 'Community', each with a list of links. The page also features a 'With the support of:' section with logos from Intel, Tidelift, Chan Zuckerberg Initiative, and bodo.ai, along with logos for NUMFOCUS, TWO SIGMA, VOLTRON DATA, Coiled, Quansight Labs, and NVIDIA. On the right, there's a 'Recommended books' section with covers for 'Python for Data Analysis' and 'Effective Pandas 2'.

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