# Overview

The goal of this exercise session is to deepen your understanding of different clustering algorithms.

# 1 Flat Clustering

Download the `clustering.ipynb` notebook and the provided datasets on Toledo. In this notebook, you will implement a distance-based clustering algorithm: k-means. You will run k-means on a small dataset, and compare the algorithm to a model-based algorithm: EM clustering. As a more fun exercise, you will also apply your k-means implementation on images of football players.

# 2 Hierarchical Clustering

## 2.1 Agglomerative Clustering

Apply agglomerative clustering with (1) single linkage and (2) complete linkage to a dataset with four points A, B, C, D with the distance matrix below. Also draw the corresponding dendrograms.

| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 1 | 4 | 5 |
| B | | 0 | 2 | 6 |
| C | | | 0 | 3 |
| D | | | | 0 |

## 2.2 Integration with Distance-Based Clustering

(a) For the following two sets of points, write down the cluster feature (CF) that summarizes each set:

1. {(1,2), (2, 3), (3, 2), (2, 1)}
2. {(2,4), (4, 3), (3, 4), (2, 2)}

(b) Given the following cluster feature $\{5, (25, 30, 20), 439\}$, do the following:

1. Compute the centroid of the cluster.
2. Compute the radius of the cluster.
3. Give the Manhattan distance of the point (3, 8, 2) to the centroid.

(c) Construct a CF tree using BIRCH for the following points: {(3, 4), (4, 5), (7, 4), (8, 4), (4, 7), (1, 1)}.
Use a radius threshold of 1.5 and branching factor of 2 for both leaf and non-leaf nodes.

# 3 General Questions

1. What does it mean that EM is a model-based clustering approach?
2. What are the differences between agglomerative and divisive clustering? What are their relative strengths and weaknesses?