# Introduction to Data Science

Course 094201

Lab 4:

## Hierarchical Agglomerative Clustering

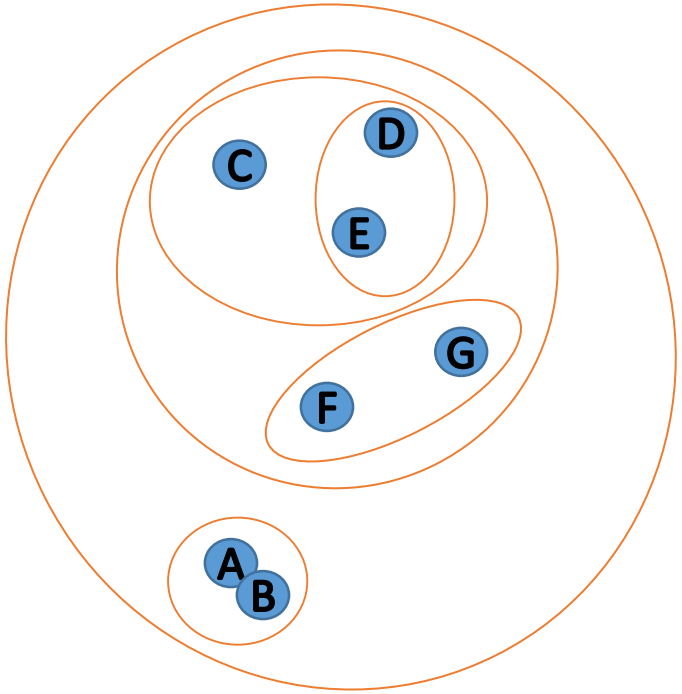
Spring 2017

# Hierarchical clustering

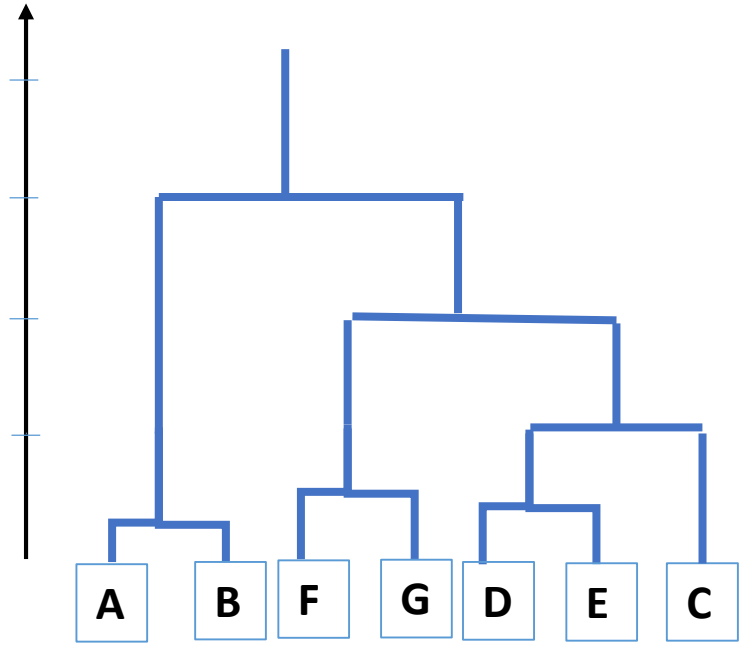
- Previously we saw the K-Means algorithm
- K-means is a partitional algorithm: the result of the clustering is K disjoint groups of items
- Hierarchical clustering generates an hierarchy of clusters;  the two main approaches are:
  - **Agglomerative**: start when each point is a cluster and at each step merge the closest pair of clusters. Eventually only one cluster will remain
  - **Divisive**: start with one cluster containing all points and at each step split on cluster into two. Eventually each point will be in a single point cluster
- Agglomerative clustering is much more common than divisive clustering

# Example

2-D example

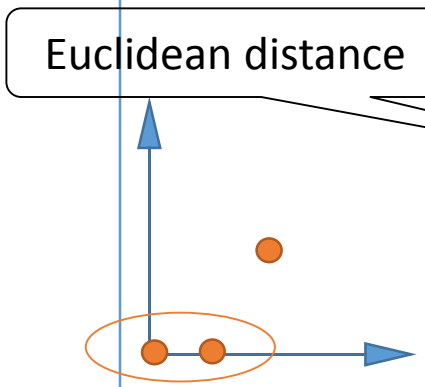



Dendrogram – a tree like diagram that describes the merges of the clusters



# The basic algorithm

- Compute the similarity matrix
- **Repeat**
  - Merge the closest two clusters
  - Update the similarity matrix
- **Until**: only one cluster remains




Similarity matrix contains similarities/distances between each two clusters:

| | **(0,0)** | **(1,0)** | **(2,2)** |
|---|---|---|---|
| (0,0) | 0 | | |
| (1,0) | 1 | 0 | |
| (2,2) | 2.83 | 2.24 | 0 |

The key point is **how the distance between clusters** is defined.
In the example on the left we will use minimal distance between cluster points:
dist( { (0,0), (1,0) }, { (2,2) } ) =
min{ dist( (0,0),(2,2) ), dist ((1,0), (2,2)) }= 2.24

# The dataset and the code

- The code and the data can be found at:

**/mnt/share/students/LAB4**

- Copy everything to your **local folder** and unzip the code:

unzip lab4-students.zip

- We will control the number of clusters we want to create and print. An alternative is to save (or print) all the clusters that are created.
- The file **input.txt** is a very simple, small input to the algorithm which includes the 6 points we have seen in lecture.
- The file **iris.data.formatted.txt** is a "classic" and famous dataset which is being used for clustering and classification. It's good to be familiar with it. https://en.wikipedia.org/wiki/Iris_flower_data_set

It includes 3 iris species and 4 numeric values which represent different parts of the flower.

# The Code - arguments

- The code arguments:
  - The input file.
  - Number of clusters to stop at.
  - Similarity/distance measure to use.

# The Code – What do we have?

1. AgglomerativeClustering has a method called run- make sure you fully understand what it does.
2. Figure out what the rest of the methods in AgglomerativeClustering are doing.
3. What is the class PairDistance? Why do we need it? And where it is being used?
4. Make sure you understand how the cluster merging is being done.
5. Make sure you understand how the distance measure is calculated for “single link”.

# Assignment – Complete the method calcAllInterPointDistances

- Implement the missing method calcAllInterPointDistances.
- This class is being used in the method singleLink in Cluster class.
- You have to decide how should this method look and what it would do.
- Try to run SingleLink on input.txt with 4 clusters and get the result we got in class.
- Make sure you understand how to compare the dendrogram we got in class to the result.

# Assignment - Home

1. הסבירו מדוע משתמשים ב std::numeric_limits<double>::max(); בפונקציה של singleLink (קראו על הפונקציה במקור חיצוני, מדוע היא עדיפה על מספר קבוע גדול שרושמים בקוד). השתמשו בפונקציה דומה בעת מימוש ה completeLink בסעיף 3.
2. הסבירו מדוע לא מוחקים קלאסטרים שמוזגו מוקטור הקלאסטרים במחלקה AgglomerativeClustering.
3. ממשו את Complete link ו Average link.

# Assignment - Home

4. בדקו על הקובץ input.txt ש Complete link ו Average link עובדים כמו בהרצאה והגישו את הפלטים הבאים לכל שלושת גרסאות האלגוריתם (על הפלטים להיות בשמות כדלקמן ):
   - output_single1.txt– הרצה עם 3 קלאסטרים
   - output_complete1.txt - הרצה עם 2 קלאסטרים
   - output_average1.txt– הרצה עם 3 קלאסטרים.
5. הריצו את שלושת האלגוריתמים על הקובץ iris.data.formatted.txt עם מספרי קלאסטרים שונים. הסבירו מדוע בחרתם להריץ עם מספר קלאסטרים זה או אחר. איזה מהגרסאות עובדת יותר טוב על הנתונים האלו ומדוע? מה היו הציפיות שלכם ומה קיבלתם בפועל?

- **עליכם להגיש את כל קבצי הקוד ללא קבצי הפרויקט. קובץ PDF בודד עם תשובות לשאלות ואת הפלטים הנ"ל.**