The goal of this class project is to determine how often there is an overlap between KEGG pathways.

[**Data Frame 1: Pathways**](http://rest.kegg.jp/list/pathway/hsa)

Pathways[0]= Pathway\_ID

Pathways[1] = Pathway\_Name

[**Data Frame 2: Gene to Pathway Map**](http://rest.kegg.jp/link/pathway/hsa)

Gene\_pathway[0] = Gene\_ID

Gene\_pathway[1] = Pathway\_ID = Pathways[0]

[**Data Frame 3: Gene Info**](http://rest.kegg.jp/list/hsa)

Gene\_info[0]= Gene\_ID = Gene\_pathway[0]

Gene\_info[2] = Gene\_name and description

**1. Map/merge the information** by PATHWAY\_ID and GENE\_ID.

1. Merge Data\_frame\_1 and Data\_frame\_2 by pathway ID. Full merge or right merge(Data\_frame\_2) since some genes appear in multiple pathways.

Gene\_A Pathway1 pathway 1 does something

Gene\_A Pathway2 pathway 2 does something

2. Merge resulting Data frame from 1 and merge with data frame 3 using a right merge(Data\_frame\_1\_2 on the right) with data frame 3.

Gene\_A Pathway1 pathway 1 does something Gene description A

Gene\_A Pathway2 pathway 2 does something Gene description A

[Documentation on merging data frames in python](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html)

**2. Compute** the number of overlapping genes between every 2 pathways.

1. Use some kind of for loop to make a dictionary with the pathway as the key and the genes in the pathway as the values.

{“pathway\_ID” : (gene\_ID\_1, gene\_ID\_2, ect..)}

1. Magic?

**3. Save the result to** a file **KEGG\_crosstalk.csv** with the following columns: PATHWAY\_ID1, PATHWAY\_NAME1, PATHWAY\_ID2, PATHWAY\_NAME2. Order the results descending by the number of overlapping genes where PATHWAY\_ID1 is different than PATHWAY\_ID2.

**4. Compute** a rank of the genes based on how many pathways they appear on and save it to a file.

**5. Retrieve** a set of the pathways the top 3 genes appear on.

**6. Compute and display** a Venn diagram for number of overlapping pathways for the top 3 genes.