# S&DE Atelier - Visual Analytics

## Assignment 2 - part 1

**Due** May 8, 2023 @23:55

**Contacts**: marco.dambros@usi.ch - carmen.armenti@usi.ch

The goal of this assignment is to use ElasticSearch and Kibana to solve different problems. The submitted zip file should be named with your first name, last name, and number of the assignment: `SurnameName_Assignment2_part1.zip.` This should contain your solutions and the steps followed to arrive to these solutions.

The datasets you need for this exercise is in the csv named *restaurants.csv*.

## Exercise 1 - Indexing, queries and aggregations (30 points) 🔍

1. **Indexing**
   Ingest the restaurant dataset provided in CSV format, transforming every row into a JSON document where the name of the fields are the name of the columns in the CSV format. You should briefly explain all the steps you followed to index and import the data in ElasticSearch.

2. **Queries**
   a. We would like to get those restaurants that have 'pizza' in the name and not 'pasta'. Get only the restaurants that have been reviewed at least as ‘Very Good’.

   b. Which are the 5 most expensive restaurants whose reviews were done in 2018? We are interested in reviews which refer only to places within 20 km from Athens (33.9259, -83.3389) and would like to look at the 5 most expensive.

   c. Get all restaurants which contain the substring 'pizz' in the restaurant name but that do not contain neither 'pizza' nor 'pizzeria'.

3. **Aggregations**
   a. Show the number of restaurants reviewed as 'Good' aggregated by number of votes. Please consider the following ranges: from 0 to 250, from 250 to 500, from

500 to 750, from 750 to 1000. For each bucket we would like to know the minimum and maximum value of the average cost per 2.

b. We are interested in cities which have not less than 10 restaurants and restaurants that have at least 100 votes. Which are the 7 cities with the highest average restaurant price (cost for two)?

c. Show the highest number of votes for different rating types in descending order. You should consider only restaurants that are within 9000 km of New Dehli (28.642449499999998, 77.10684570000001).

💡 For each of the aggregations above, discuss whether you are facing the **shard size** problem. If that would be the case, how would you solve it?

**Deliverable:** a report or txt file containing the code, number of documents returned from each query and explainations of your solution when required.