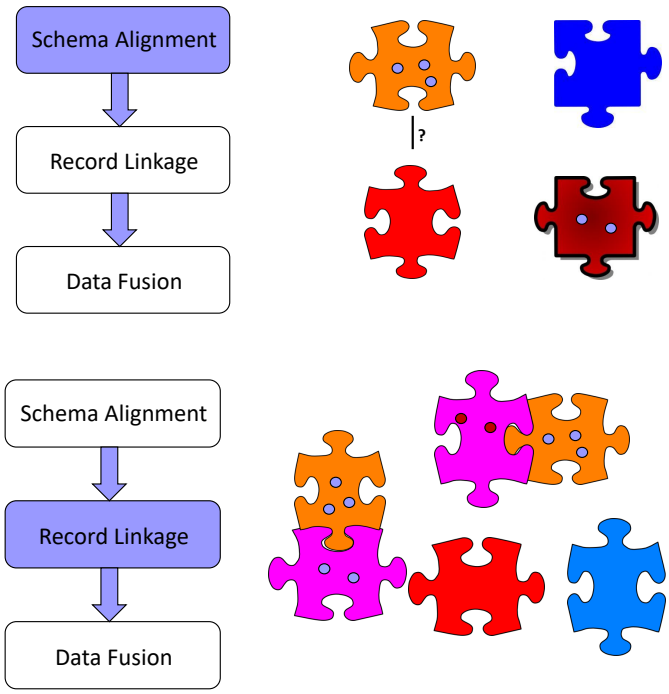
Assignment on "Big Data Integration"

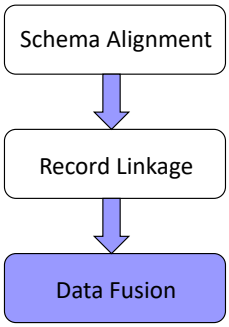
Due: July 5th, 2024

Motivated by the topics discussed in class, the goal of this assignment is to study one of the steps involved in the data integration process: Schema alignment, Record Linkage, Data Fusion. Choose one (or more) of them, read recent papers on the subject chosen (including the recent approaches that also leverage the capabilities of LLMs for supporting solutions to these problems), and write a report (around ten pages) describing: (1) the problem, (2) a possible approach (or several, making a comparison), (3) possible improvements of the approach, and (4) how it can be implemented (including a link to a repository with the actual implementation). Optional: if you aim to receive an extra bonus for the evaluation of this assignment, test the approach on a dataset of the DI2KG challenge (http://di2kg.inf.uniroma3.it/2020/#challenge).

There is no need to design a method from scratch, you can suggest modifications of the existing approaches. The assignment should be completed in groups. Each group submits one report, but all the members of the group are expected to understand and be able to explain the proposed solution.

The report must be uploaded on Moodle. Prof. Torlone will be responsible for the evaluation and so there is no need to send it also to the speaker.




Schema Alignment
Record Linkage
Data Fusion