# Homework 2: Exploring Splitting the Data for Entity Resolution

Due Friday, March 5 at 5 PM EDT

***Group assignment***: This is a group assignment, where you will submit in groups of no larger than four, and you can choose your groups. There must be a team leader, so please designate one, who's job is to help your group be organized through this homework assignment. Each team member must contribute equally to the homework assignment, which will be measured by pushes to the team repository. Please set up a work flow using branching, where tasks are assigned to just one team member. You might wish to have other team members review tasks as you go given they build upon each other.

***General instructions for homeworks***: Please follow the uploading file instructions according to the syllabus. Your code must be completely reproducible and must compile.

***Advice***: Start early on the homeworks and it is advised that you not wait until the day of as these homeworks are meant to be longer and treated as case studies.

***Commenting code*** Code should be commented. See the Google style guide for questions regarding commenting or how to write code https://google.github.io/styleguide/Rguide.xml. No late homework's will be accepted.

**Total points on assignment: 5 (reproducibility) + 10 points for the assignment.**

Consider the `RLdata500` data set in the `RecordLinkage` package.

Suppose that this data set is too large to work with and we would like to create a sample size of 10 records that is representative of the 500 records in this data set. Specifically, we want to make sure that our performance (precision, recall, F-measure) on the sampled data set is representative of the original data set. Let's investigate how we can do this!

1. Start by doing simple exploratory data analysis of the data set. What do you find?
2. What happens if you randomly sample 10 records from the original data set? Do this a few times and describe what happens? Is this representative of the original data set? Explain and be specific.
3. Propose a method that works better than random sampling and explain why this works better.
4. Propose evaluation metrics, visualizations, etc. to support any of your claims.

Write this up with your group, testing this out on `RLdata500` with commentary.

**Hint 1**: Think about the fact that you have unique identifiers and you know the maximum cluster size here.

**Hint 2**: Talk to myself/Olivier if you're getting stuck. This is a hard but very important problem!

**Suggestion**: Redo this for your homework 1 for El Salvador as something extra if you have time or want an extra challenge! Bonus points will be awarded for this if you do this extra task with your group!