**MIT Solve prediction algorithm**

*On the data cleaning side:*

* Identify the columns that are the same and the ones that stand out:
  + By year
  + Across years
* Identify the categories of selection (P, SF, F, is there a Selected? If so when?)
* Make sure we have good merges of evaluator ratings
  + By year
  + Across years
* Manage the NAs in the merging with evaluator ratings

*On the prediction side:*

* Algorithm for representing the answers to text solutions in a two-dimensional space (BERT + PCA)
* Identifying the columns that matter in prediction. We have four categories of variables:
  + basic information (ID, Challenge name etc..)
  + Categorical information on the project (stage of development, areas needing support, dimension addressed. Basically any question that has a limited set of answers)
  + Textual information on the project (problem statement, solution statement, etc..)
  + Categorical information on the submitter (Gender, country of origin, etc…)