Capstone 2: Collating behavioral data sheets from various sources

Applied Behavioral Analysis (ABA) is the application of empirically based methods to create socially significant behavioral changes in individuals. This process involves collecting data in person to further analyse and create plans for behavior change. Companies have been operating with paper data sheets for years and copying those to spreadsheets for analysis. With the popularity of online data collection and analysis software, many ABA organizations are beginning to switch to these platforms. Companies I have explored currently do not have the ability to input large amounts of past data from different sources.

Companies and employees have been using various methods to organize these spreadsheets of data and have different naming and abbreviation styles. Many types of inconsistencies exist.

The ability to review a client's historical data using modern kinds of analysis tools would give clinicians a much stronger grasp of the client they are treating.

These data consist of categories of behavioral observations like aggression, eloping, tantrums, non-compliance, etc... Some of these entries will be totally unique; each entry should have an associated treatment date. Additionally, some of the target behaviors are listed as frequency and others are listed as daily percentages. A sample of these datasets show different abbreviations and spellings for behaviors, as well as instances of duplicated rows and multiple dates inputted into the same row. Different methods have been used to note null values as well.

There may be instances where the intention of a single column or a larger dataset is unknown. For this current project, the original data collectors are available for more information. Due to the nature of this data, finding the proper intention as well as including unique items should be a high priority.

To unify these datasets it will be necessary to link columns with the same intent of data from different spreadsheets and reorganize them into a single format. The steps involved in this will be cleaning, indexing, comparing, classifying and evaluation of the process. This can be accomplished using pandas and recordlinkage packages in python.

A successful end to this project would involve a system that would take an input of many spreadsheets and standardize the datasets to be uploaded to an online platform.

The current scope of this project is limited to behavior data, though the datasets include program targets that are unique to each client. This work could be extended to organizing program targets in a later stage. A further extension would be to add a computer vision system that extracted data directly from paper sheets into a properly formatted spreadsheet.