Data Science Methodology Final Assignment: **Emails**

• The topic I chose in order to apply the data science methodology are Emails and more specifically a classification scheme for the emails that are sent to a small business.

• A small business receives numerous emails daily. These emails vary in topics, importance, purpose, among other things. In order to increase the business' productivity, these emails can be classified into different categories, so that each category can be managed by specific employees. Depending on the average load of each email category, different numbers of employees can undertake the management of each category. For example, "Social Media Notifications" could be handled by Personal Relations employees, "Business Offers" would be handled by the business managers and "Spam emails" would immediately be moved to the recycle bin.

The question risen is, "Can we manage to automatically arrange the received emails into different categories, based on their contents and sender information?"

• 1. Analytic Approach: Since we're talking about classifying the emails into categories, we will build a classification model based on binary Yes/No answers.

2. Data Requirements: The Data required would be all of the emails sent to the small business in the past and perhaps emails sent to other small businesses of a similar type. More specifically, we would require the email sender's address, the subject of the email and, most importantly, the email's main body.

3. Data Collection: The Data would be retrieved from the small business' email server, as well as from the servers of other small businesses of a similar type, if they are willing to participate in the building or the deployment of the model. If the each employee has their own corporate email and are willing to contribute, their emails could be input into the database as well.

4. Data Understanding and Preparation: At this point, corrupted emails would have to be removed from the database, checks should be performed for identical emails sent to all businesses or to more than one employees of the same business and further data cleaning optimizations should be performed. Additionally, categories should be created (based on words or phrases on the email main bodies or perhaps the senders' addresses) and each email should have a yes/no (or, equivalently, 0/1) answer for each of these categories.

5. Modeling and Evaluation: Using the biggest part of the database's emails, chosen randomly, we would build this classification model. Then, the emails from the database that were not utilized for the model building process would be used for the evaluation of the model. Tests performed with these remaining emails would give us a good idea of how well the model classifies our data. Depending on the types of mismatches, we would revisit the model, make the proper adjustments and then re-evaluate it in a similar manner. This process would continue iteratively, until the model would be able to correctly classify most emails into their corresponding categories, in which case it would be ready for deployment.