# Introduction

I am going to share my experience with data mining for this assignment. Currently I am working on building a data mining solution which will classify healthcare related codes. This solution would help us identify sensitive codes and mask those records having any of those sensitive codes going out the door (if necessary) to satisfy HIPAA guidelines.

# Data and Pattern

So far, we have identified the following codes which will be mined along with their sources:

1. ICD-10 Diagnosis codes (Source - WHO)
2. ICD-10 Procedure codes (Source - WHO)
3. ICD-9 Diagnosis codes (Source - WHO)
4. ICD-9 Procedure codes (Source - WHO)
5. Procedure Modifier codes (Source - CMS)
6. HCPCS (Healthcare Common Procedure Coding System) codes (Source - CMS)
7. HIPPS (Health Insurance Prospective Payment System) codes (Source - CMS)
8. CPT (Current Procedural Terminology) codes (Source - AMA)
9. CDT (Current Dental Terminology) codes (Source - ADA)
10. POS (Place of service) codes (Source - CMS)
11. LOINC (Logical Observation Identifiers Names and Codes) codes (Source - LOINC)
12. Revenue codes (Source - NUBC)
13. Laboratory Order codes (Source – LabCorp, MedTox)
14. Laboratory Result codes (Source – LabCorp, MedTox)
15. SNOMED codes (Source - SNOMED)

All these codesets are being pulled from their sources periodically and loaded into an Oracle relational database application. This Oracle application is our reference data mart.

Our own internal Sensitive Codeset Committee has identified some keywords which will be used to identify one of the following categories:

1. Substance abuse disorder
2. Sexual health – STD
3. Sexual health – Abortion
4. Sexual health – HIV
5. Sexual health – Pregnancy
6. Sexual health – Family planning/ Contraceptives
7. Mental Health
8. Genetics
9. Not Restricted

A code can be associated with multiple categories.

The list of keywords received from the committee has the highest trust score as it was manually reviewed. The committee updates the list quarterly shares the full the list in an excel file to us. The keyword list only keywords which are related to the first 8 categories mentioned above (excluding not restricted).

In addition to the keyword list, we ingest publications from vendors like Beacon, Orion and Geneia who have done some type of mining on their end and share their classifications in excel format. The excel files only hold the restricted codes (the first 8 categories mentioned above).

Unlike the keyword list, the publications from these vendors do not have high trust scores, but we still take those classifications into account for final scoring of our prediction.

# Technology used

Unfortunately, our team does not have access to SAS enterprise miner. Which is why we are using the following tools:

1. SAS enterprise guide
2. Python

We are using SAS enterprise guide to connect to the reference data mart, pull all the codes and descriptions and store them in a SAS dataset. Then we perform data cleansing and after that we use the keyword list to provide class prediction.

Then the dataset is exported to a csv file which then gets imported into Jupyter notebook and the following modeling techniques are being utilized for category prediction:

1. Naïve Bayes
2. Random forest classifier
3. Logistic regression
4. XGBoost

# Challenges

The main challenge we are facing is the processing time. Due to hardware constraints, SAS process takes a long time. Moreover, we do not have a secure server for python which is why we cannot shift the whole process to python either.

Also, when we analyzed the final output, we found that we are getting a small amount of (<5%) false positives. This is due to the lack of robustness of the keyword list. Our keyword list gives us combinations of up to 2 words max. Most of the keywords are just one word which causes false positives. We are trying to move towards the bag of words approach and add more words instead of one word as keyword into the mix. This is why I am keen to learn more data mining techniques to improve the accuracy for this mining process.