**Overview/Business Understanding:**

You are the lead data scientist at the FBI. The Department of Justice (DoJ) needs evidence of communications that indict members of the Enron management, for collusion during the firm’s bankruptcy proceedings (2001-2002). Without this information, they lack the evidence to pinpoint the exact members of the management team who were part of this scandal.

While researching the case, the FBI’s forensics team classified 1700 e-mails according to subject content, forwarded information, primary topic and sentiment. However, there are approximately 500,000 emails to search through, and time is running short to gather evidence for the prosecution of the correct perpetrators.

The “head honcho” of the FBI approaches you, asking if you can solve the problem using “analytical mumbo-jumbo”. He tells you that the **goal of the project** is to identify largest “body of evidence” via email correspondence that help indict members of the Enron management team. The challenge is that you have a short time period and a limited number of resources available to manually read/label email correspondences - you believe it can be solved using NLP prediction methods.

At the end of the project, you will attend a required debriefing session with DOJ prosecutors to present your results. Your presentation should: 1) inform your audience on your approach to solving the problem, 2) present your solution clearly and concisely, including relevant insights and results, 3) identify the “next steps” in the project (if they exist) 4) Cover the implementation/operationalization plan with decision-makers (if needed)

**Data Understanding:**

The Enron data was originally collected at Enron Corporation headquarters in Houston during two weeks in May 2002 by Joe Bartling,[3] a litigation support and data analysis contractor working for Aspen Systems, now Lockheed Martin, whom the Federal Energy Regulatory Commission (FERC) had hired to preserve and collect vast amounts of data in the wake of the Enron Bankruptcy in December 2001.

Once collected, the Enron emails were processed and hosted in litigation platform Concordance, and then iCONECT, for the investigative team from the Federal Energy Regulatory Commission, the Commodity Futures Trading Commission, and Department of Justice investigators to review.

The email archive was made publicly available and searchable via the web using iCONECT 24/7, but the sheer volume of email of over 160GB made it impractical to use. Copies of the collected emails and databases were made available on hard drives.

A copy of the email database was subsequently purchased for $10,000 by Andrew McCallum, a computer scientist at the University of Massachusetts Amherst.[6] He released this copy to researchers, providing a trove of data that has been used for studies on social networking and computer analysis of language.

**Project Steps (Using CRISP-DM & Dataiku)**:

**Step 1 - Business Understanding**

* Identify the problem that your management wants to address
* Understand the background of the current situation
* Identify the business goals to achieve from the outcome of the project
* Identify Constraints (limitations on what you may do, the kinds of solutions that can be used, when the work must be completed, and so on)
* Determine Business success criteria

**Step 2 - Data Understanding/Exploration (using Tableau**

For additional information on the dataset, see: <https://enrondata.readthedocs.io/en/latest/references/research/>

* Identify the relevant target categories for the analysis
  + See: <https://data.world/brianray/enron-email-dataset>
* Explore other variables in the dataset, for data quality, data prep and model building purposes
  + See related tableau workbook in “Project” folder

**Step 3 - Data Preparation/Model Building**

1. [Imputing Missing Values/NA’s](https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/)
2. Data Selection - Outliers, Filters (labeled vs. unlabeled), Binning and/or imputation
3. Standardization, Scaling & Formatting of the data
4. Model Preprocessing - Data Transformation
5. Dimensionality Reduction

**Step 4 - Model Building Process**

1. Select appropriate ML algorithms.
   1. The algorithm to be selected depends completely on the business requirement, available data and the desired outcome. What type of model should we choose here?
   2. Dataiku Prediction - Model Building, Evaluation & Scoring Guide: <https://academy.dataiku.com/latest/tutorial/machine-learning/index.html>
2. In an ideal situation, we should try different algorithm or combination of algorithms (Ensembles) to actually arrive at our final best algorithm.

**Step 5 - Model Comparison & Evaluation**

1. There are many model evaluation technique - What should we choose?
   1. Depends on the evaluation criteria, final model desired outcome , business requirement and the model algorithm used.
   2. Hint: ROC & Accuracy
2. Build Multiple Models in “visual analysis”, then select the best model - which is the best model?
   1. Evaluation of the model in the context of the business success criteria
3. Score & evaluatie unlabeled dataset: <https://academy.dataiku.com/latest/tutorial/scoring/index.html>

**Final Report**

1. Final report – this is the final written report of the data mining engagement prepared for management/customers, Including:
   1. Summary of the entire project, reports created up to this point, and its results.
   2. Final presentation – A summary of the final report is presented in a meeting with management. This is also an opportunity to address any open questions.
   3. Next Steps & Implementation Plan

**Model Deployment & Operalization**

1. Publish the model results to a Dataiku dashboard, to monitor future results
2. Schedule and Distribute the resulting analysis dataset, for use in data products (such as tableau)
   1. In this case, download the “scored” dataset to CSV, for use in Tableau Dashboard