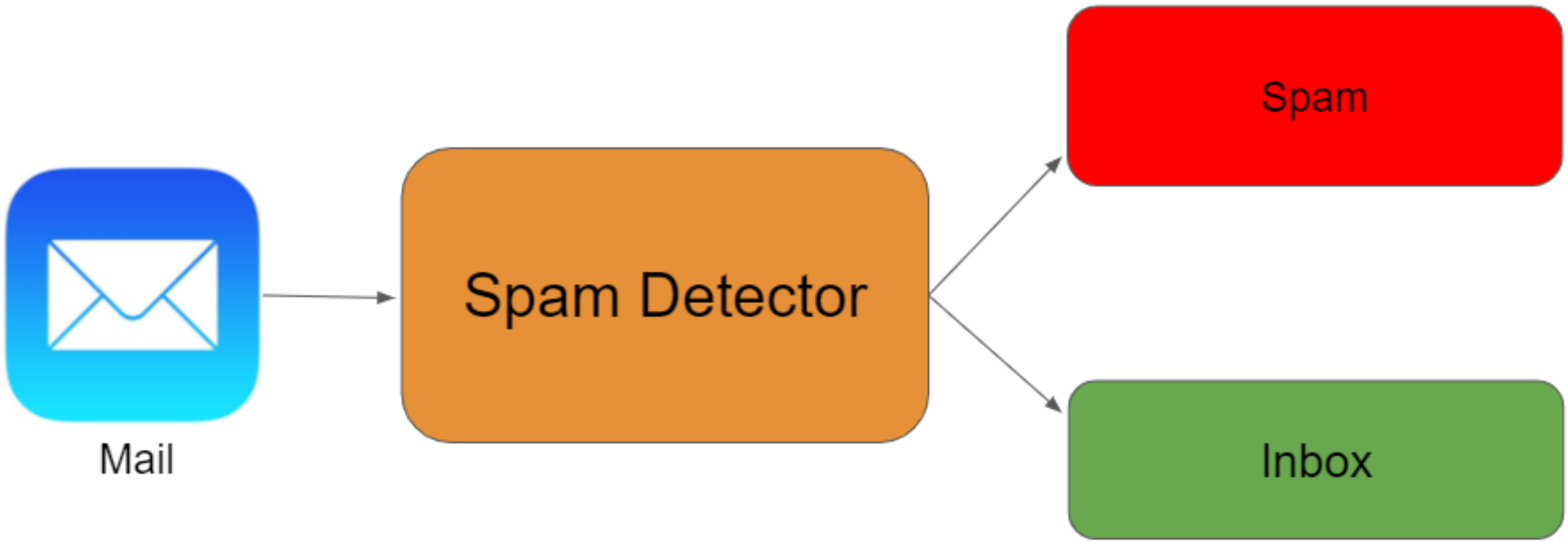
# EMAIL SPAM DETECTION MODEL

BY:

M. ASHISH REDDY

DSWP -> BATCH-5


Spam
Spam Detector
Mail
Inbox

# WHY SPAM DETECTION?

- An email has become one of the foremost important kinds of communication.
- In 2014, there are estimated to be 4.1 billion email accounts worldwide, and about 196 billion emails are sent day after day worldwide.
- Spam is one of the main threats posed to email users.
- All email flows that were spam in 2013 are 69.6%.
- Therefore, an effective spam filtering technology is a significant contribution to the sustainability of cyberspace and our society.
- As the importance of email is not lesser than your bank account containing 1Cr., then protecting it from spam or frauds is also mandatory.

# Data Preparation

- To prepare the data, we followed the steps below:
- 1. Download spam and ham emails through Google's takeout service as a box file.
- 2. Read the mbox files into lists using the 'mailbox' package. Each element in the list contained an individual email. In the first iteration, we included 1000 ham mails and 400 spam mails (we tried different ratios after the first iteration).
- 3. Unpacked each email and concatenated their subject and body. We decided to include the email subject as well in our analysis because it is also a great indicator of whether an email is a spam or ham.
- 4. Converted the lists to data frames, joined the spam and ham data frames, and shuffled the resultant data frame.

- 5. Split the data frame into train and test data frames. The test data was 33% of the original dataset.
- 6. Split the mail text into lemmas and applied TF-IDF transformation using Count Vectorizer followed by TF-IDF transformer.
- 7. Trained four models using the training data:
- **Naive Bayes**
- **Decision Trees**
- **Support Vector Machine (SVM)**
- **Random Forest**
- 8. Using the trained models, predicted the email label for the test dataset. Calculated four metrics to gauge the performance of the models as Accuracy, Precision, Recall, F-score, AUC.