Personal vs Promotional Email Classification Challenge

Artificial Intelligence Summer School 2019

CentraleSupélec

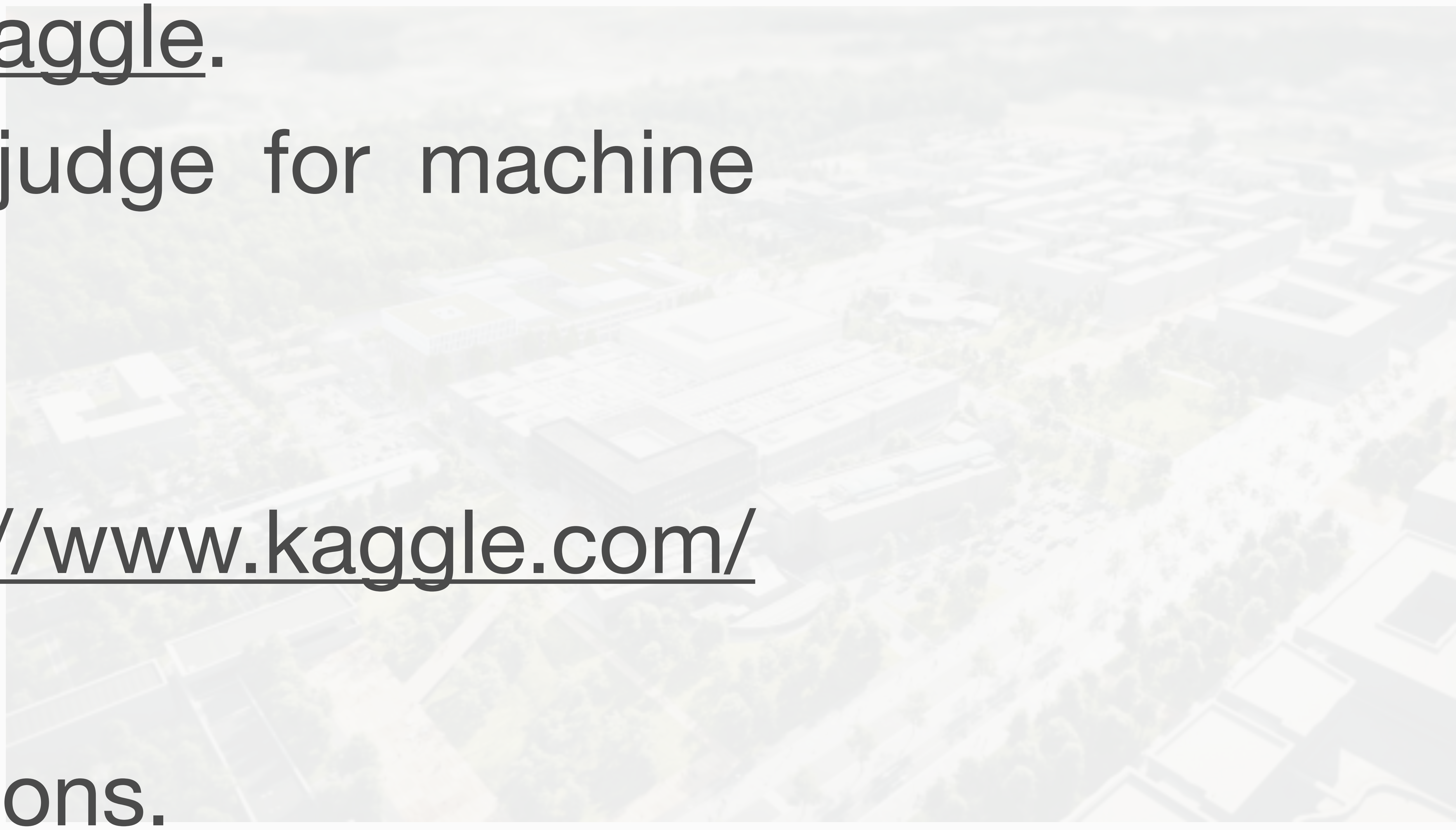
Motivation

We often face the problem of searching meaningful emails among thousands of promotional emails.

Challenge Goal

This challenge focuses on creating a binary classifier that can classify an email based on metadata extracted from the email.

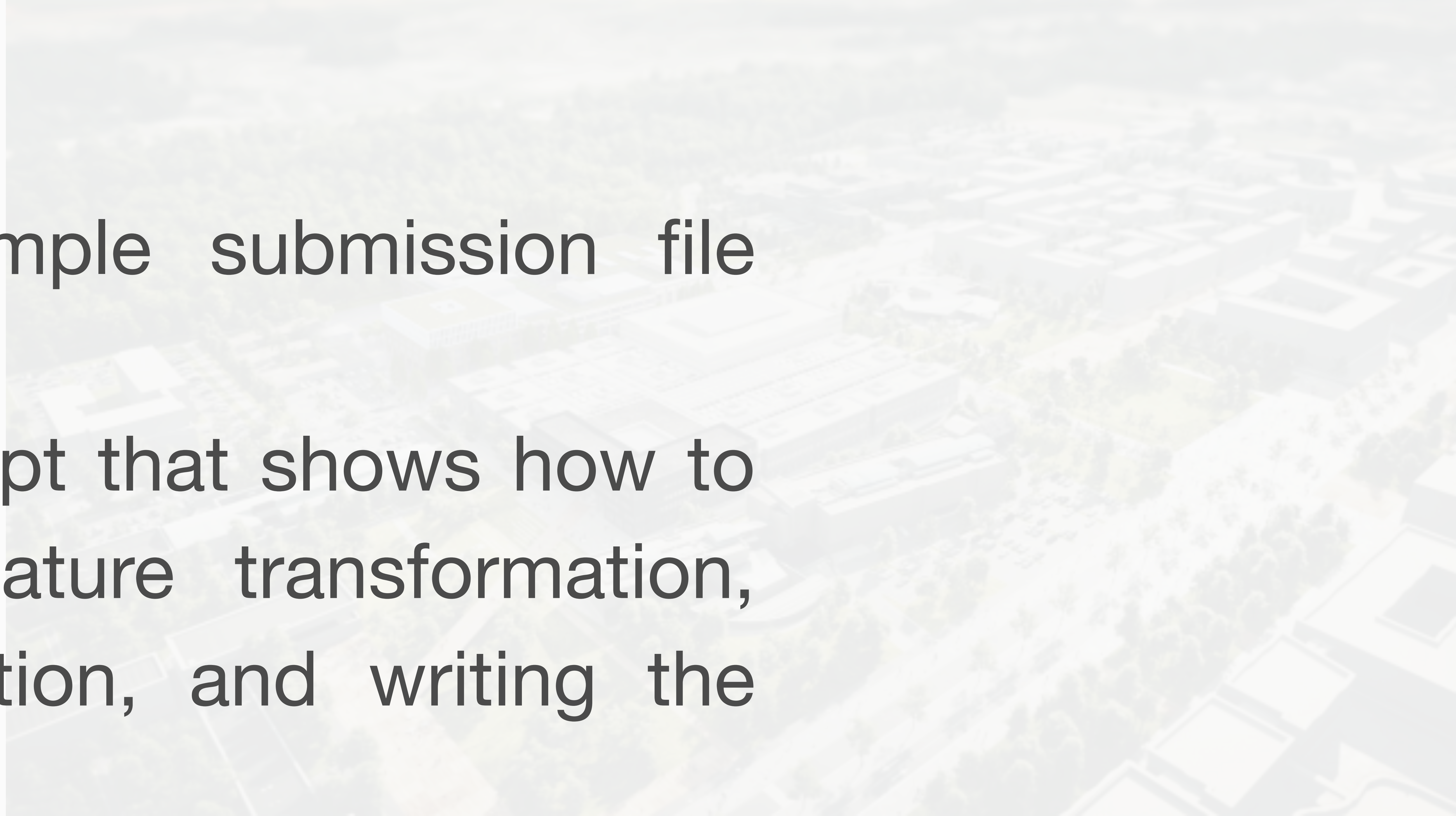
How to start with the challenge?



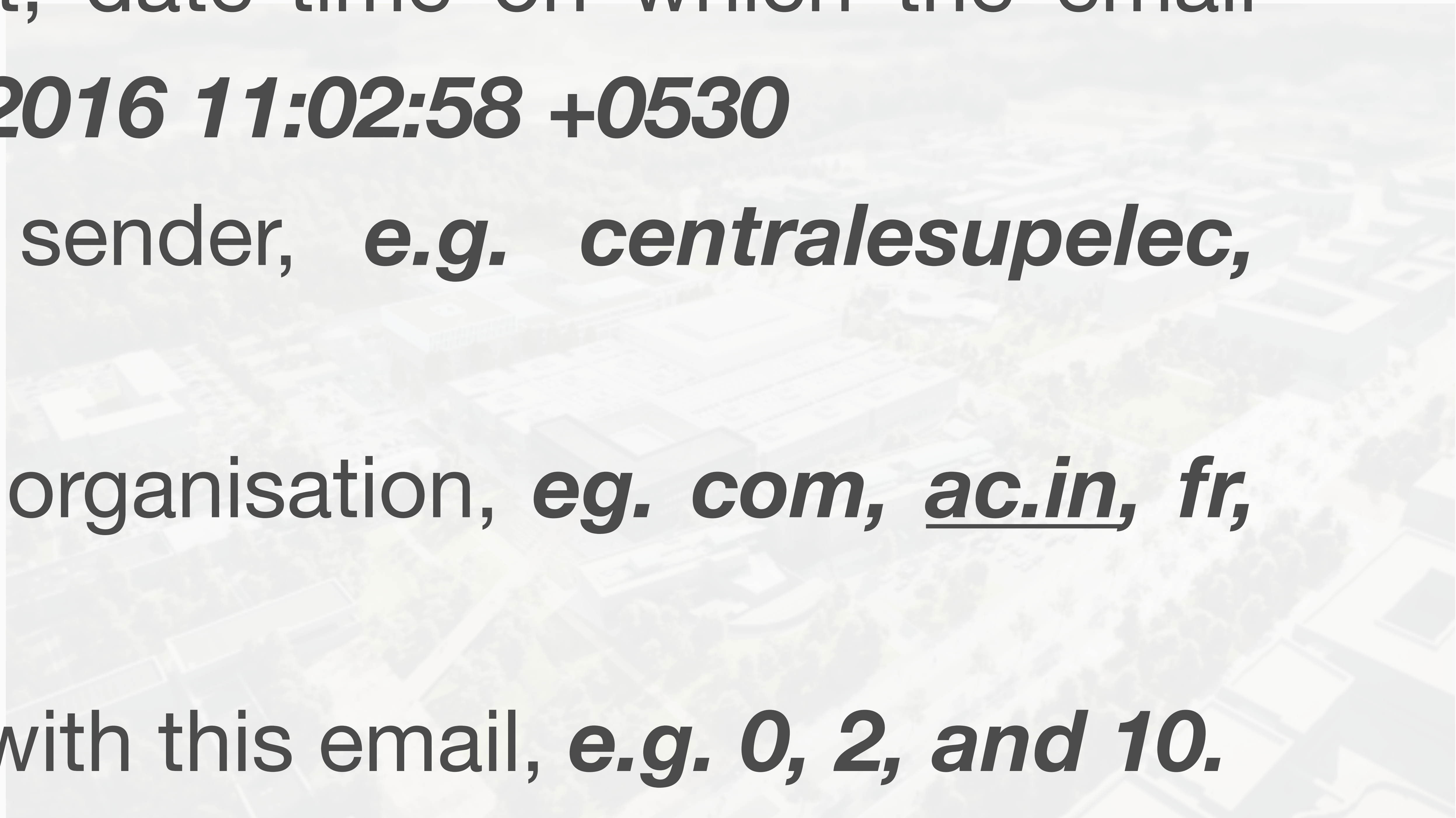
- The challenge is hosted on kaggle.
- Kaggle provides an online judge for machine learning problems.
- Register on kaggle.
- Go to the challenge at https://www.kaggle.com/c/csaisummerschool .
- Accept the terms and conditions.

Files

- train.csv - the training set
- test.csv - the test set
- sample_submission.csv - a sample submission file showing the correct format.
- skeleton_code.py - a python script that shows how to read the data, how to do feature transformation, training a benchmark knn solution, and writing the results to the submission csv file.



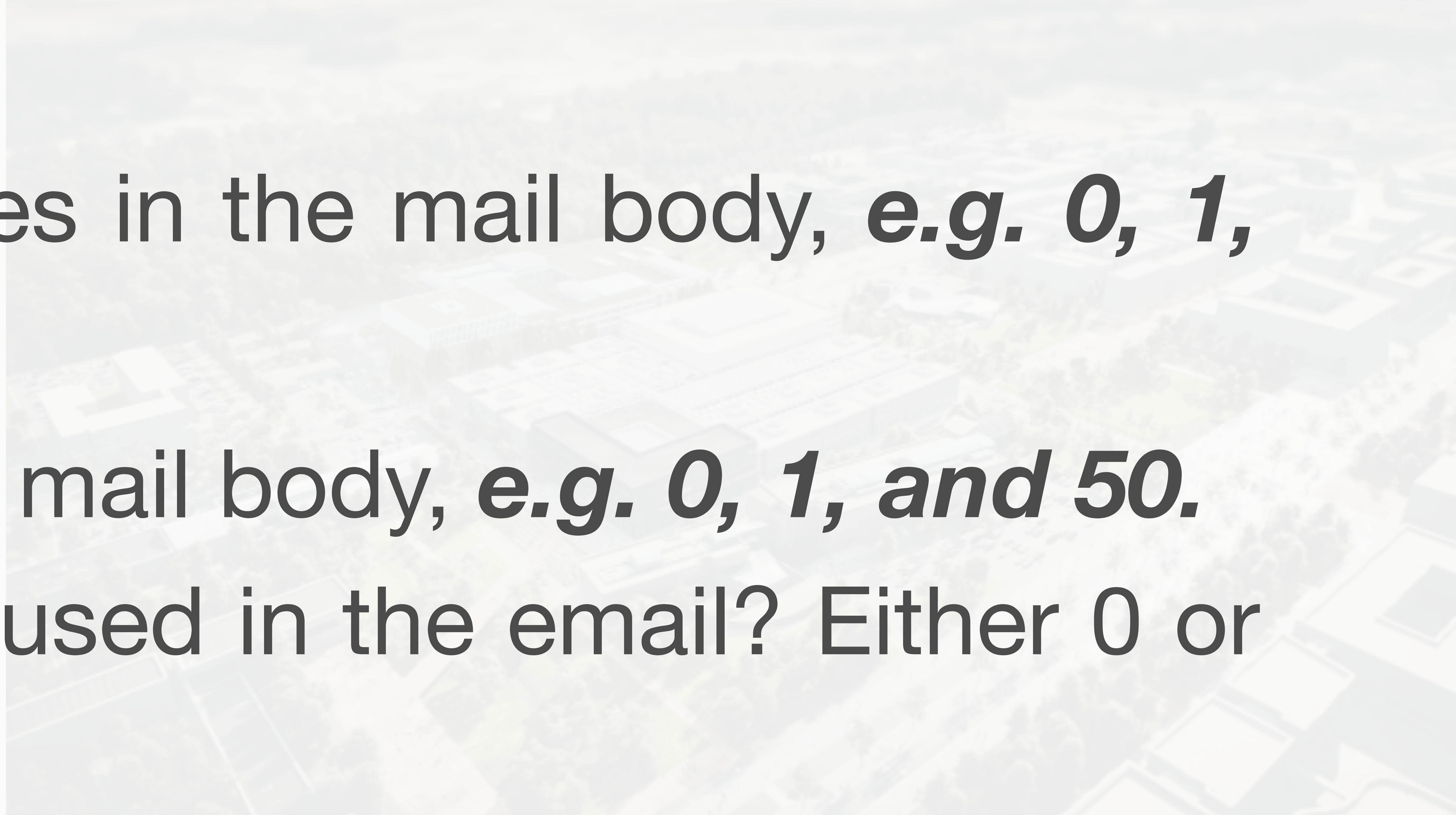
Dataset Features



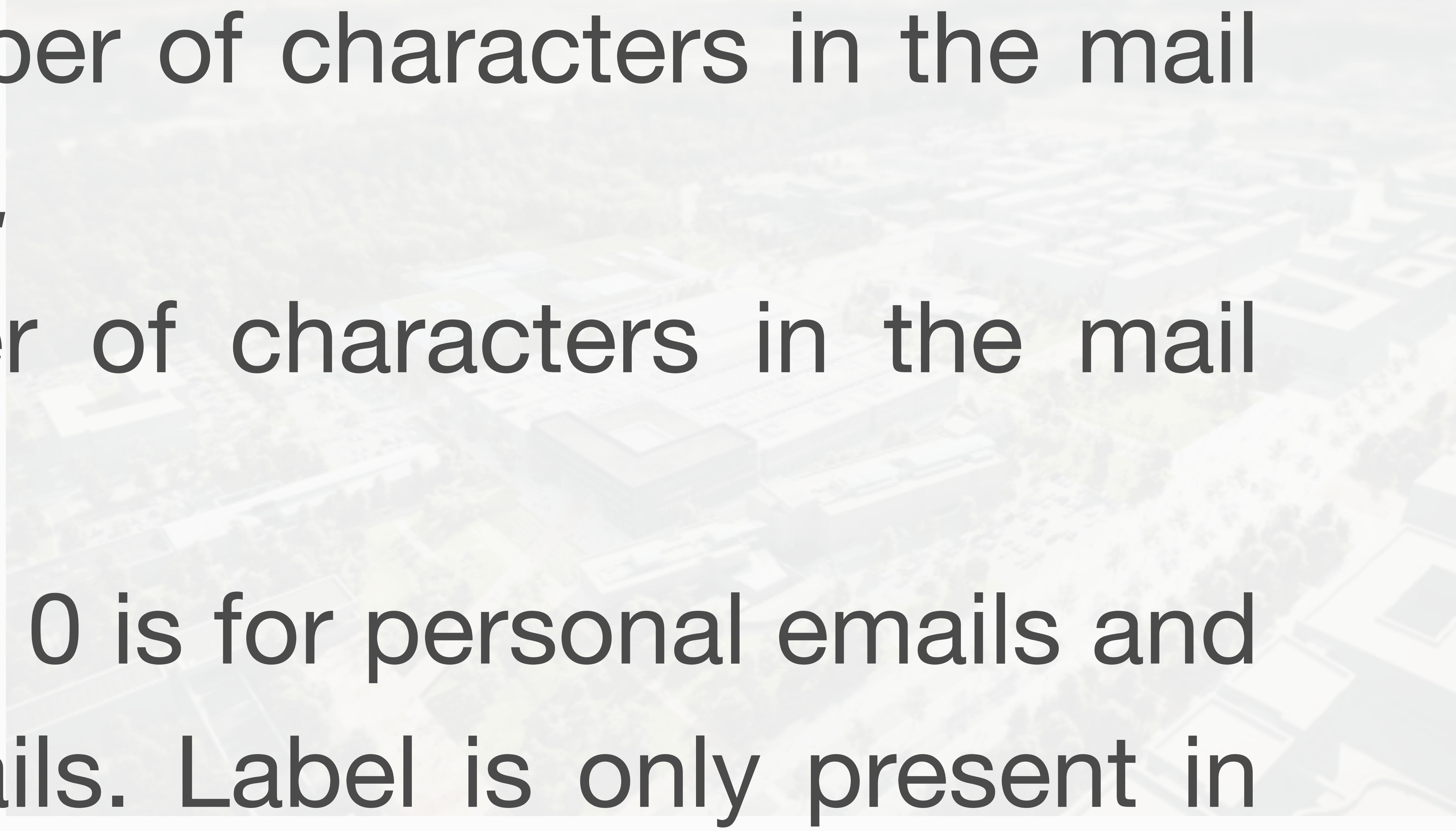
- **date** - unix style date format, date-time on which the email was received, ***e.g. Sat, 2 Jul 2016 11:02:58 +0530***
- **org** - organisation of the sender, ***e.g. centralesupelec, facebook, and google.***
- **tld** - top level domain of the organisation, ***eg. com, ac.in, fr, and org.***
- **ccs** - number of emails cced with this email, ***e.g. 0, 2, and 10.***
- **bcced** - is the receiver bcc'd in the email. Can take two values 0 or 1.

Dataset Features (Cont.)

- **mail_type** - type of the mail body, ***e.g. text/plain and text/html.***
- **images** - number of images in the mail body, ***e.g. 0, 1, and 100.***
- **urls** - number of urls in the mail body, ***e.g. 0, 1, and 50.***
- **salutations** - is salutation used in the email? Either 0 or 1.
- **designation** - is designation of the sender mentioned in the email. Either 0 or 1.



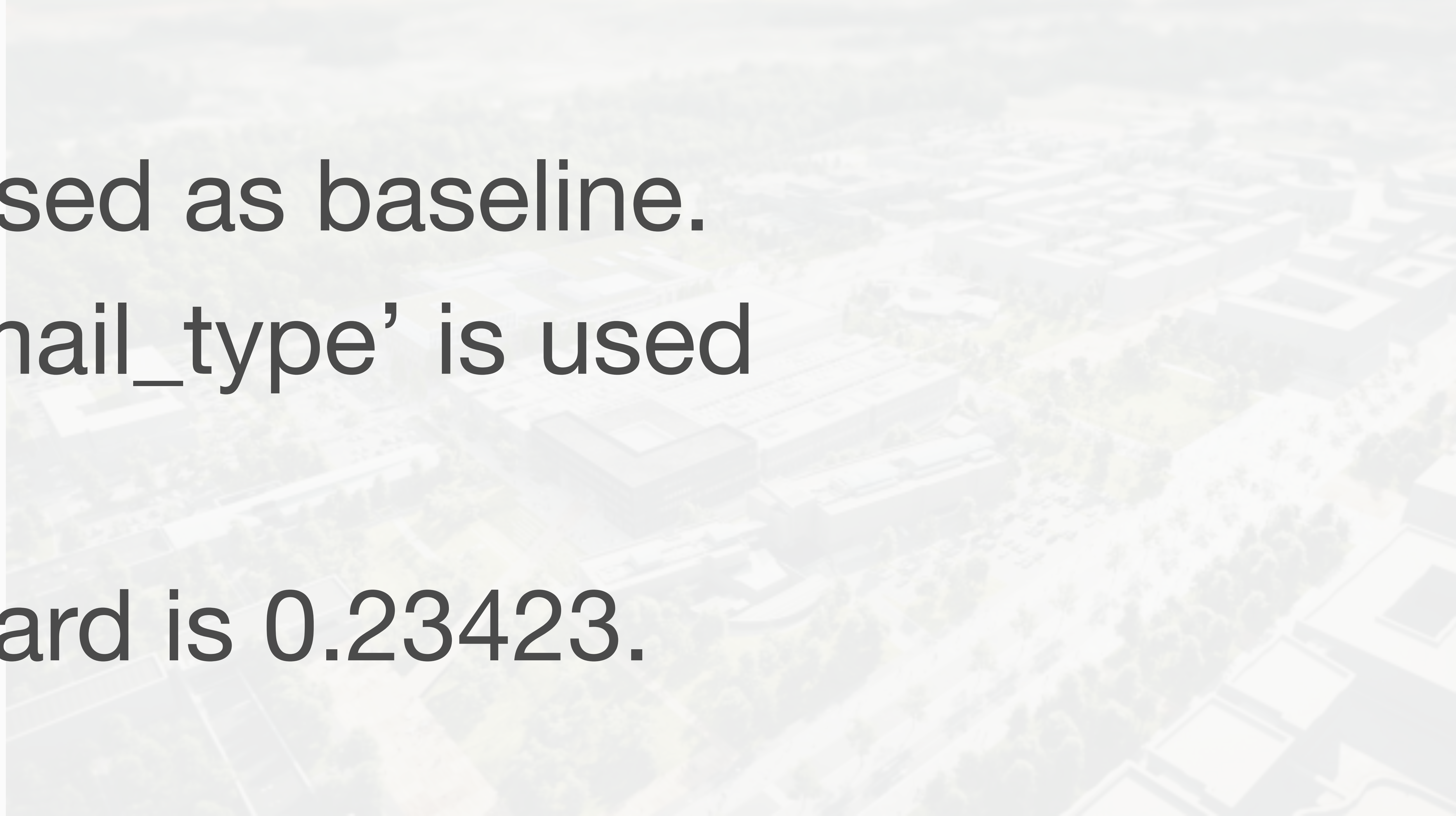
Dataset Features (Cont.)



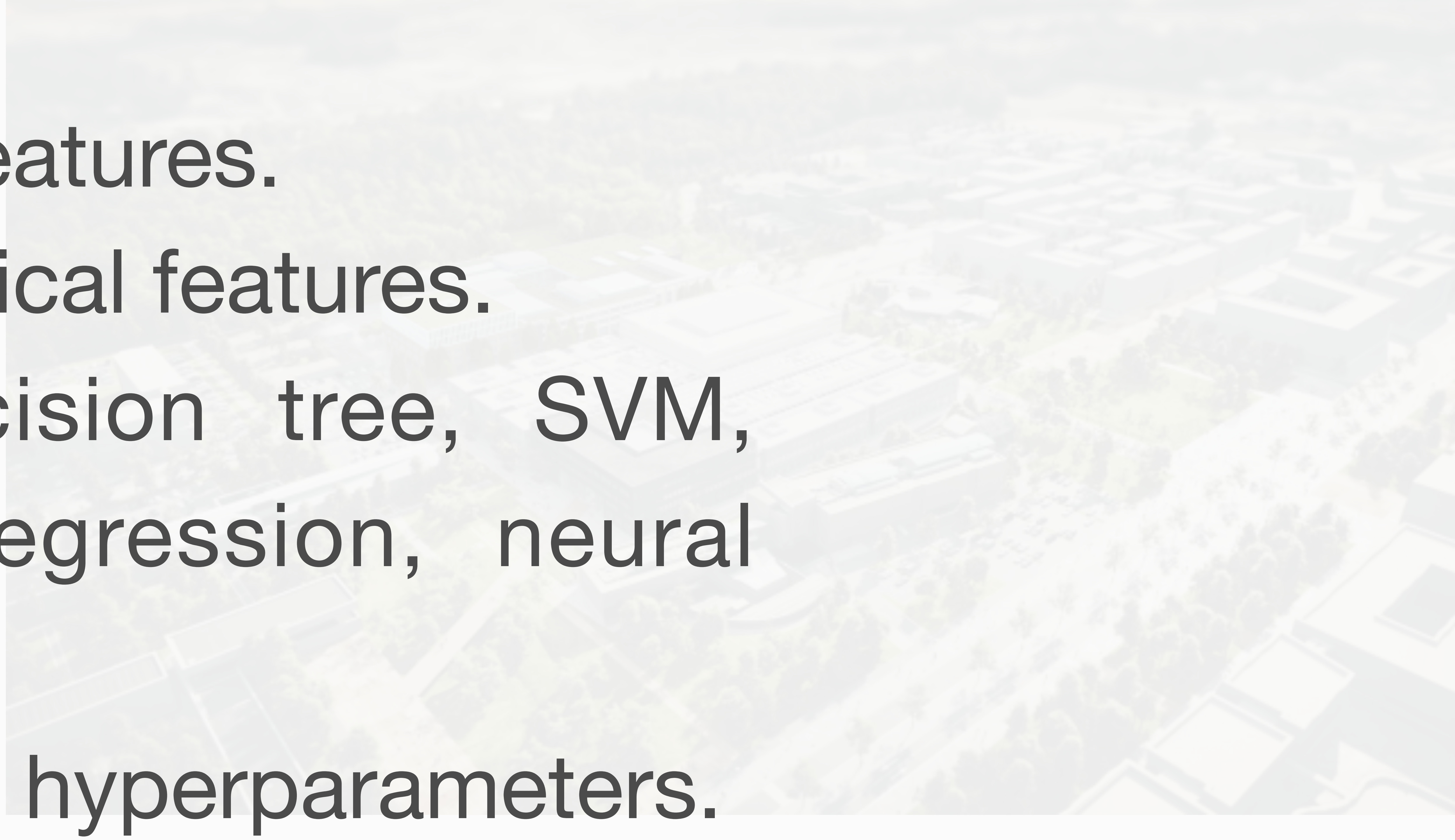
- **chars_in_subject** - number of characters in the mail subject, ***e.g. 0, 1, and 10.***
- **chars_in_body** - number of characters in the mail body, ***e.g. 10 and 10000.***
- **label** - label of this email. 0 is for personal emails and 1 is for promotional emails. Label is only present in train.csv. test.csv has all other features.

Baseline Model

- K-Nearest Neighbour is used as baseline.
- Only one of the feature ‘mail_type’ is used in the baseline.
- F1-score on the leaderboard is 0.23423.

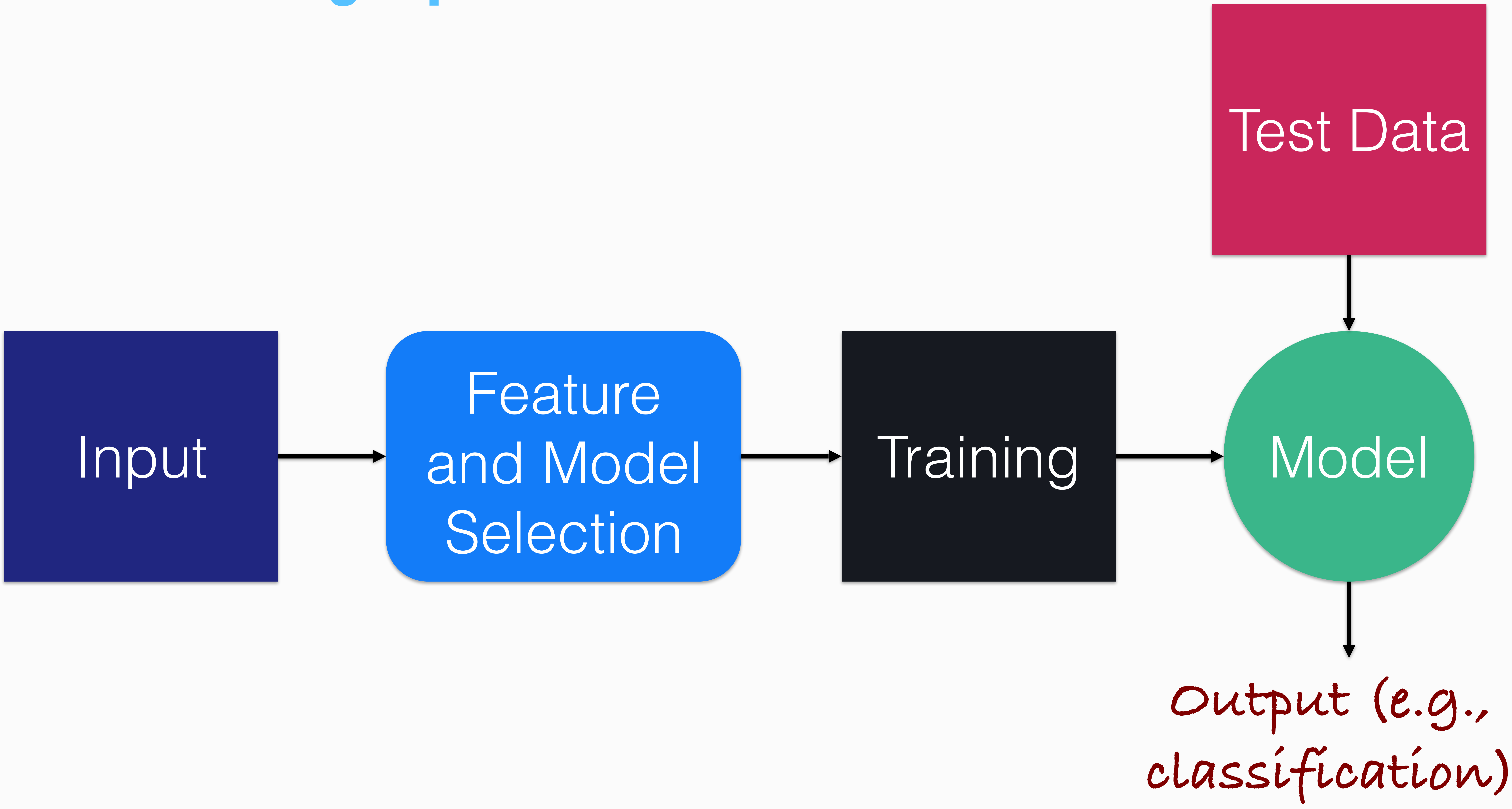


Improving Baseline Model



- KNN with multiple features.
- Normalisation of numerical features.
- One hot encoding of categorical features.
- Trying other models: decision tree, SVM, random forest, logistic regression, neural network, etc.
- Grid search over models and hyperparameters.

Machine Learning Pipeline



Test Data

Input

Feature and Model Selection

Training

Model

Output (e.g., classification)

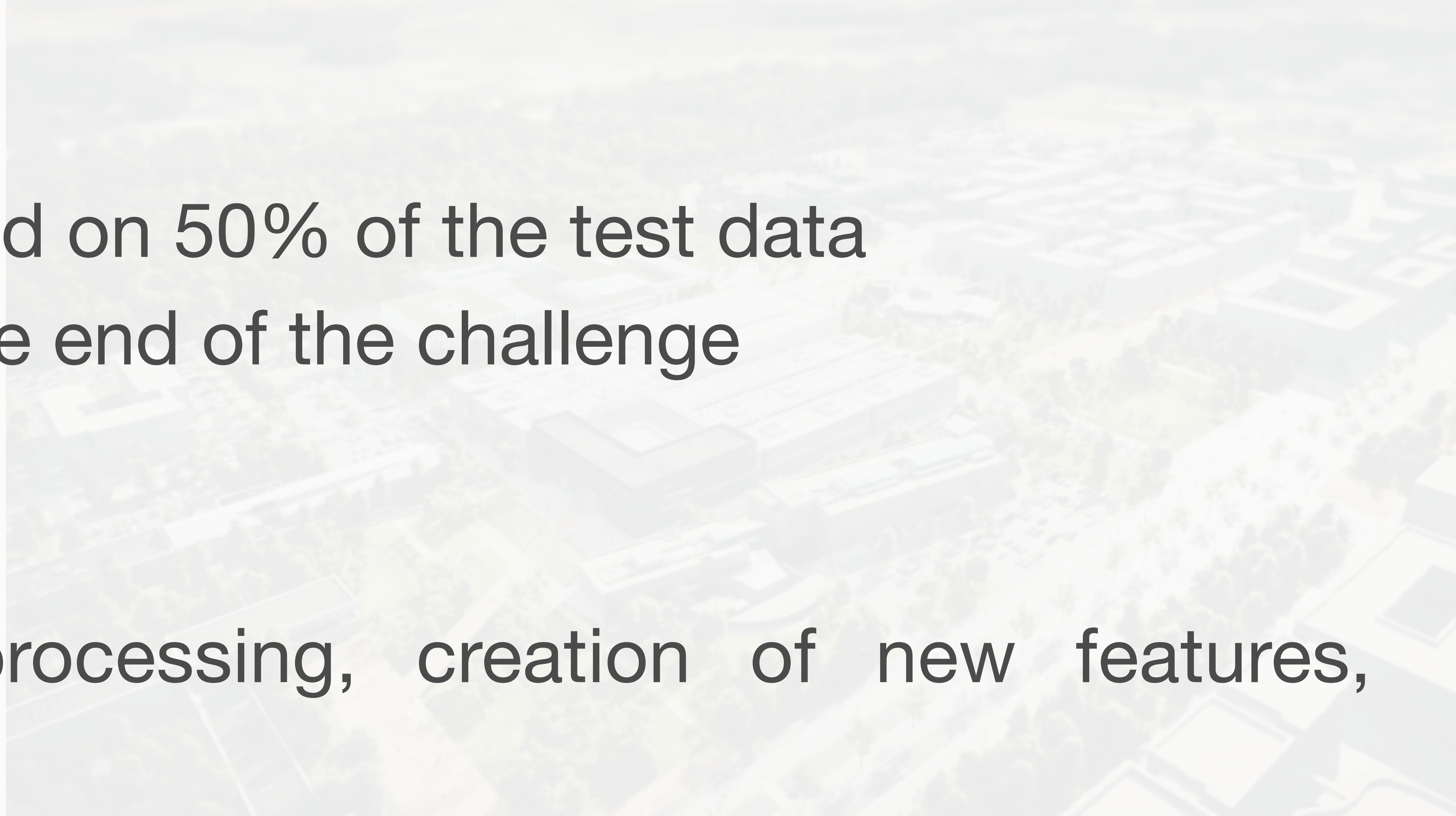
Software Tools

- Python libraries
  - numpy
  - scipy
  - scikit-learn
  - pandas
  - anaconda includes almost all the required packages

Teams for the Data Challenge

- Team 1
  - Zahra Habibollahi
  - Vadi Sai Sakthivel
  - Pierre-Louis Perez
- Team 2
  - Adam Ismaili
  - Melika Shekarriz
  - Will Smith
- Team 3
  - Catriona Calantzis
  - Dragos Gorduza
  - Ismail Ouadrhiri Idrissi Azzouzi
  - André Felipe Soares de Araujo

Submission Details



- Submission on kaggle (one per team)
  - Your best performing model
  - Leaderboard score
    - Public: what you see - computed on 50% of the test data
    - Private: will be announced at the end of the challenge
- 2-page report
  - Overview of your approach
  - Feature engineering (e.g., preprocessing, creation of new features, removal of features)
  - Classification models that you have used, comparison of different models
- Presentation (~15 minutes)
  - Overview of your approach similar to the report

Deadline: Friday, July 12

- 09:00 AM: Submission deadline
  - Send by email to Fragkiskos presentation, report, and source code
  - Email: fragkiskos.malliaros@cenralesupelec.fr
- 9:30 AM - 11:00 AM: Presentation of your approach
- For any help contact Sagar
  - Email: sagar.verma@centralesupelec.fr

Slides

https://fragkiskos.me/summer2019.pdf

Good Luck and Enjoy!