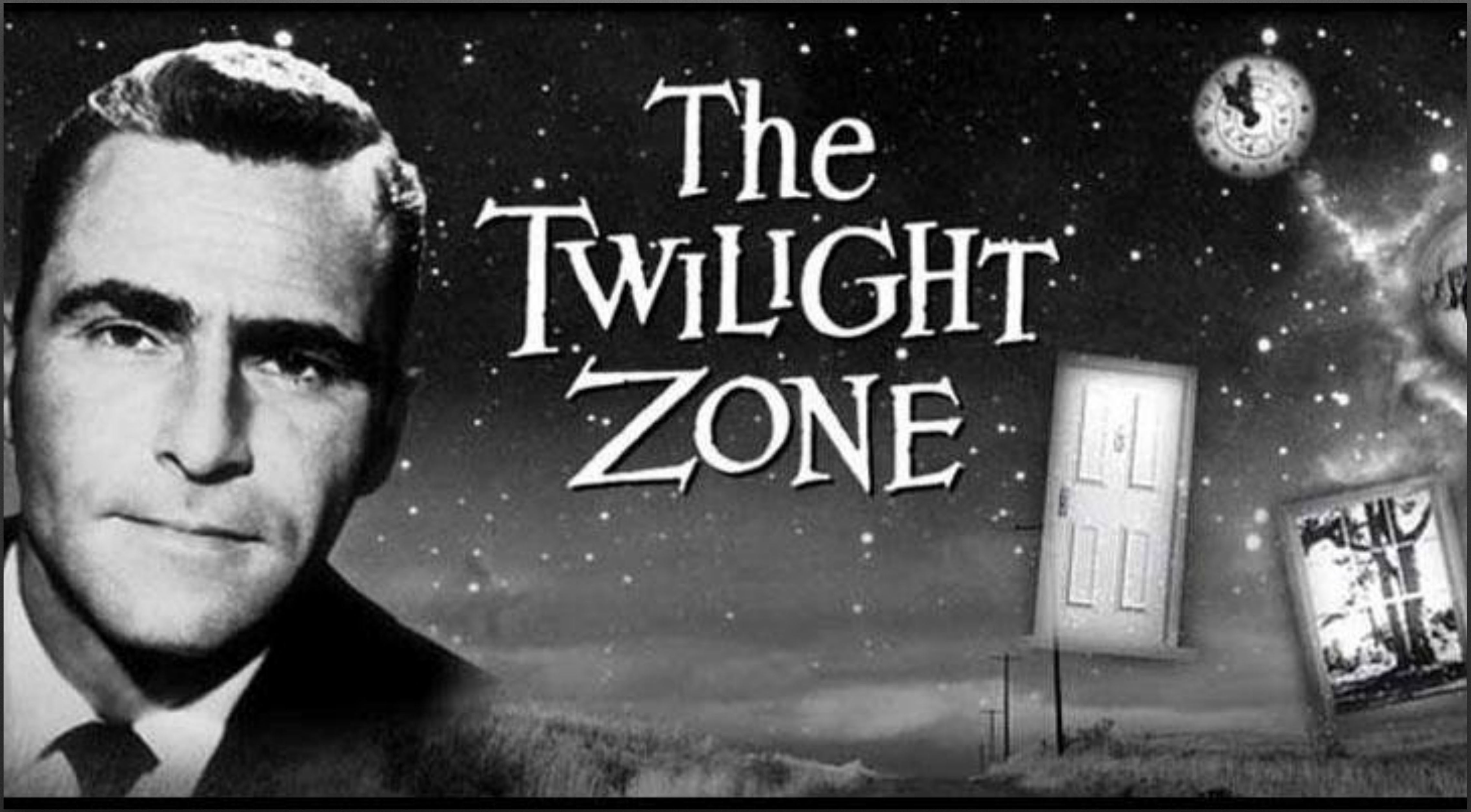
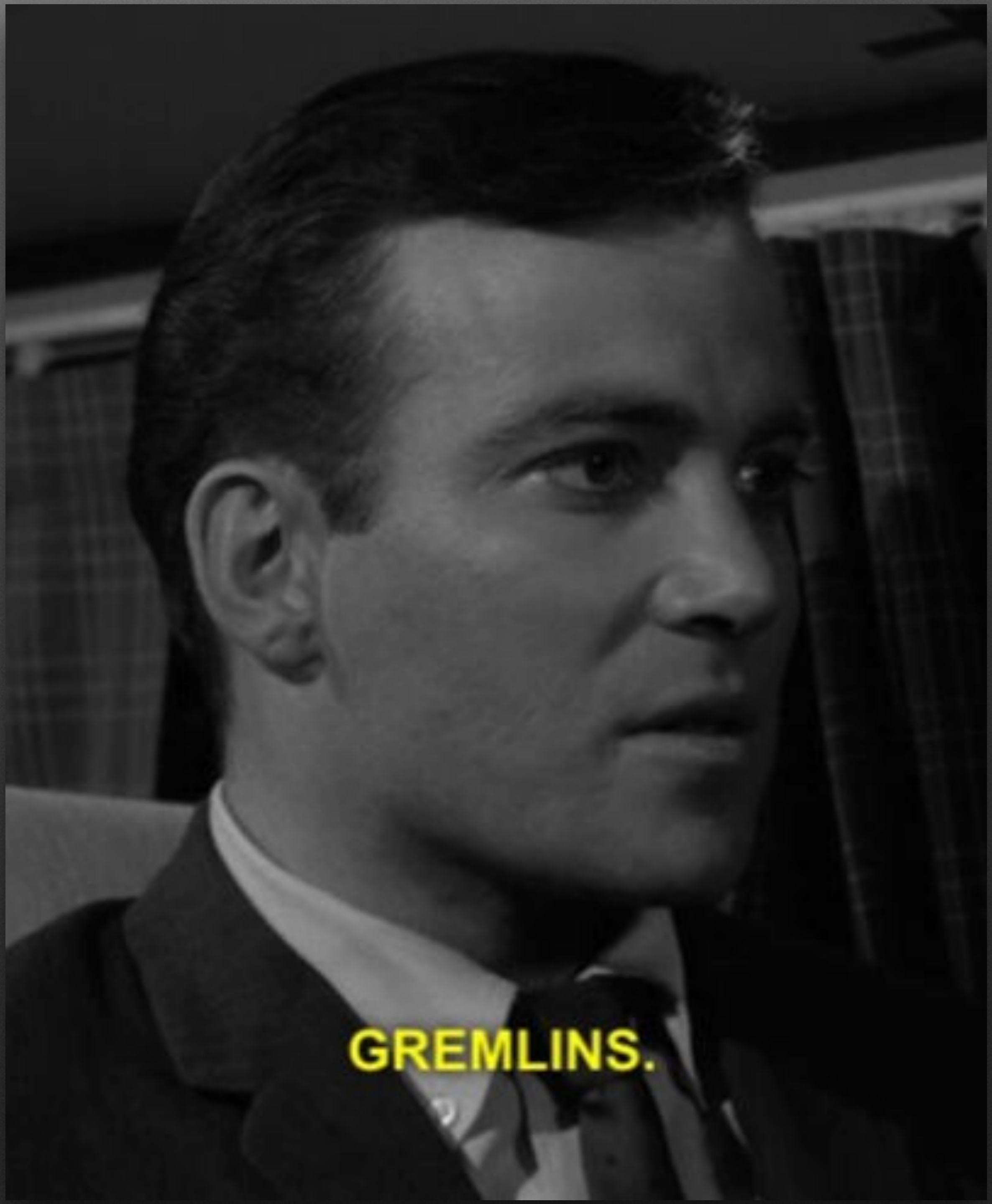


Modeling Classifiers in The Twilight Zone

John D Hazard
Data Scientist

Data Science Workflow

- Define the Problem
- Gather the Data
- Explore the Data
- Model with Data
- Evaluate Model
- Answer Problem



Problem Statement, Gathering Data, Explore Data

- Using Reddit's API collect posts from two subreddits. We used /r/TwilightZone (7.6k subscribers) & /r/comicbooks (844k subscribers)

- Used NLP to train a classifier (Naive Bayes and a Voting Classifier)

- Gathered data using query_push_shift function (json, requests, pandas)

- Dropped numeric columns, concatenated dataframes, and set subreddit column to binary (1 = Twilight Zone, 0 = comic books)

Model and Evaluate

- Used CountVectorizer and Naive Bayes in Model 1

- Model #1 scored 91% accuracy on testing data with a 1% variance

- Used VotingClassifier (RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier, LogisticRegression) in Model #2

- Model #2 scored 89% accuracy on testing data with a 4% variance

Answer the Problem

Our Naive Bayes model scored consistently higher on our testing data using only the ‘title’ column as our X variable.

Questions/Comments