# Project 3:
# Web APIs & NLP
# Cars vs Morotcycle

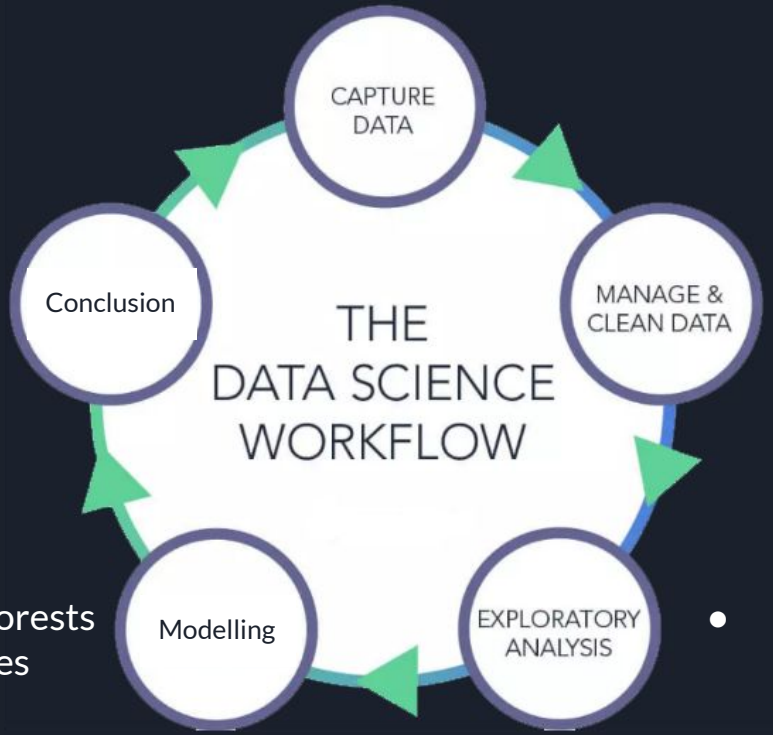
Andre Tan

# Problem Statement

Our customer, a prominent auto magazine publisher, has engaged us to help with accurately classifying subreddits in an effort to identify if a post is from r/cars or r/motorcycle. They aim to use this data to position their editorials to better cater to the masses.

# Data Science Workflow

- Web Scraping




- Dropping features
- Lemmatizing

- Random Forests
- Naive Bayes

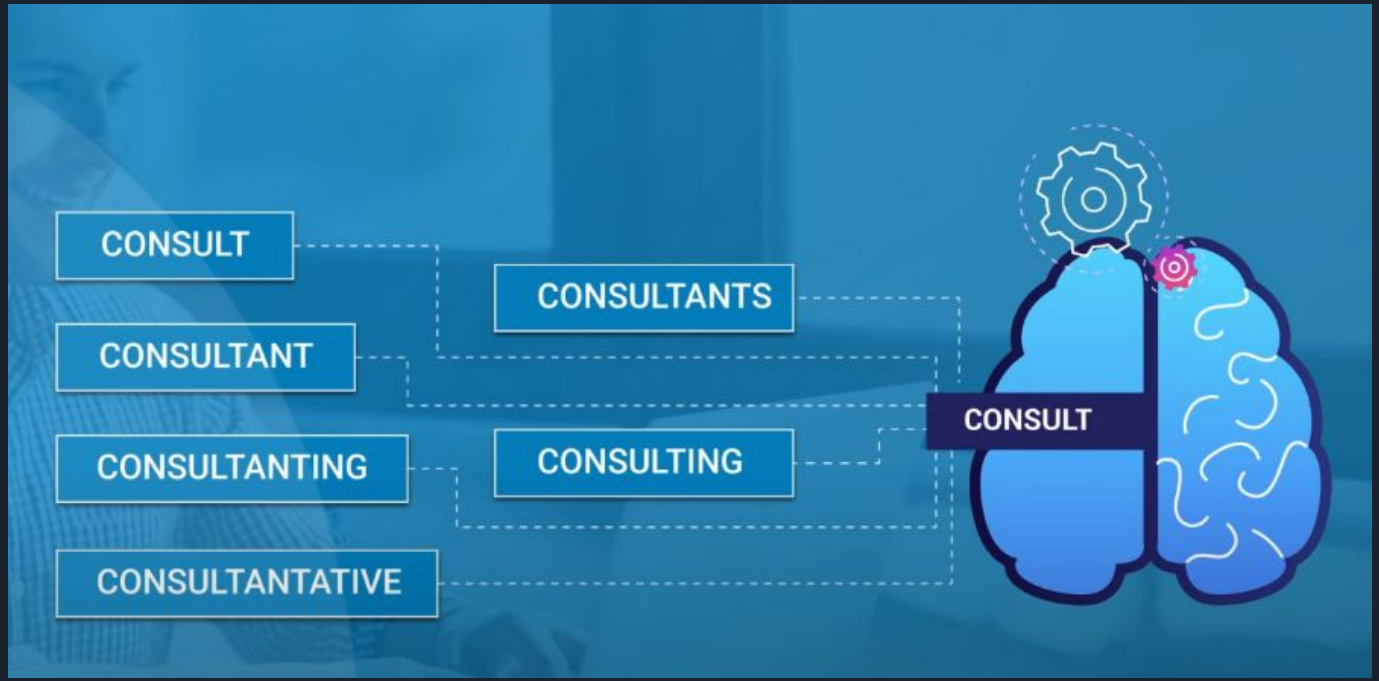
- Word Cloud

# Data Collection

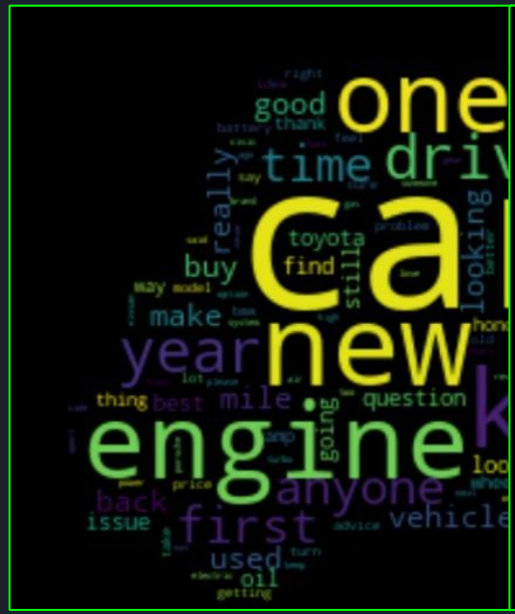
| | selftext | title | upvote_ratio | subreddit | author | is_self |
|---|---|---|---|---|---|---|
| 0 | [removed] | Best way to gain MORE experience with a manual? | 1.0 | cars | helpmebuyacar123 | True |
| 1 | | 2023 Acura Integra Receiving Optional SH-AWD | 1.0 | cars | NCSUGrad2012 | False |
| 2 | | AutoTrader: Ford Bronco Review: No Doors, No R... | 1.0 | cars | Delta_Mike_Sierra_ | False |
| 3 | [removed] | Help identifying my Great Grandfathers car (1910) | 1.0 | cars | fkncatalinawinemixer | True |
| 4 | | West Virginia House bill would ban OTA updates | 1.0 | cars | borderwave2 | False |

- Total posts collected: 19,687
  - r/cars: 9385
  - r/motorcycle: 9852

# Data Cleaning - Lemmatization



# EDA - Word Cloud





# Selected Model

**RESULTS**

| Tokenizer \| Model | precision | recall | f1 score | support |
|---|---|---|---|---|
| CountVectorizer \| MultinomialNB | 0.91 | 0.91 | 0.91 | 3938 |
| TF-IDF \| MultinomialNB | 0.91 | 0.91 | 0.91 | 3938 |
| CountVectorizer \| RandomForestClassifier | 0.91 | 0.91 | 0.91 | 3938 |
| TF-IDF \| RandomForestClassifier | 0.71 | 0.62 | [illegible] | 3938 |

- CountVectorizer selected
- Multinomial Naive Bayes or Random Forests

# Conclusion & Recommendation

- Baseline score 50.04%
- Use `CountVectorizer` with `MultinomialNB` or `RandomForestClassifier` models (91%)
- Unique recurring words:
  - r/cars: car, engine, drive
  - r/motorcycle: bike, motorcycle, ride
- Future steps:
  - Keep model up to date
  - Test other models
  - Collect more data
  - Perform sentiment analysis