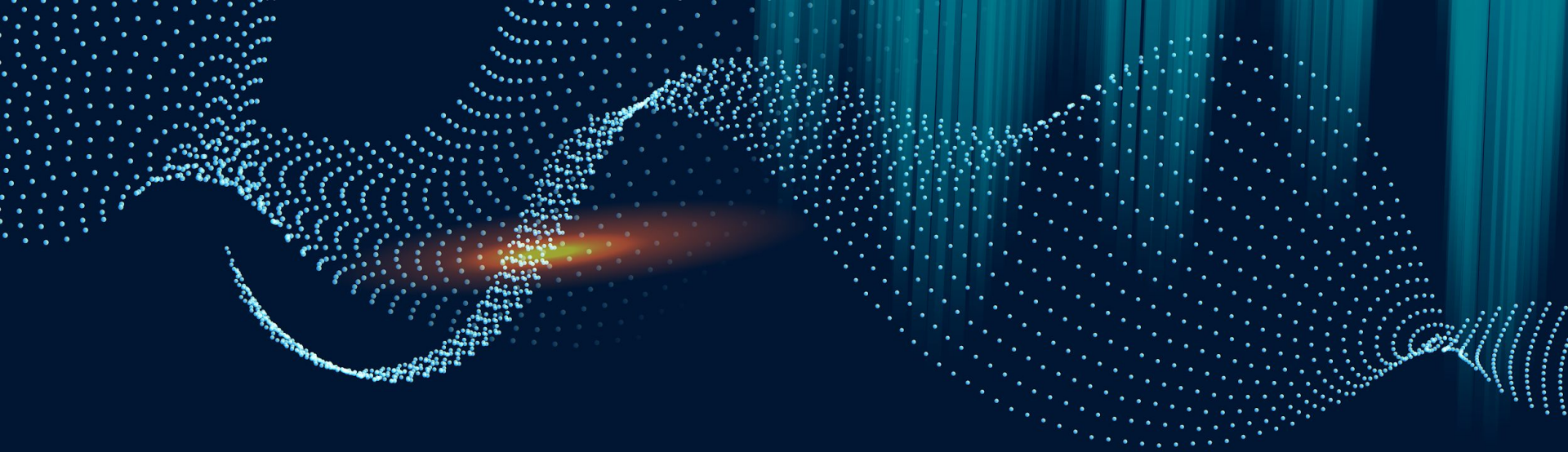
MLB Pitch Prediction

What's the next pitch?

A Data Science Project by William Guo

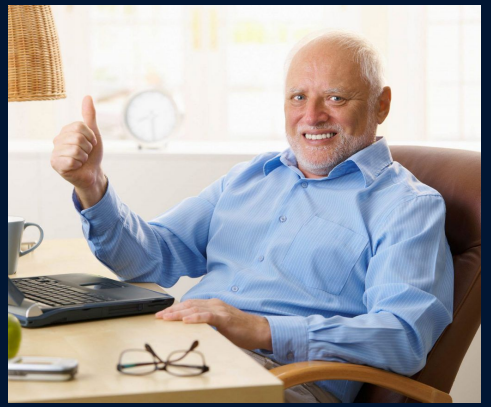


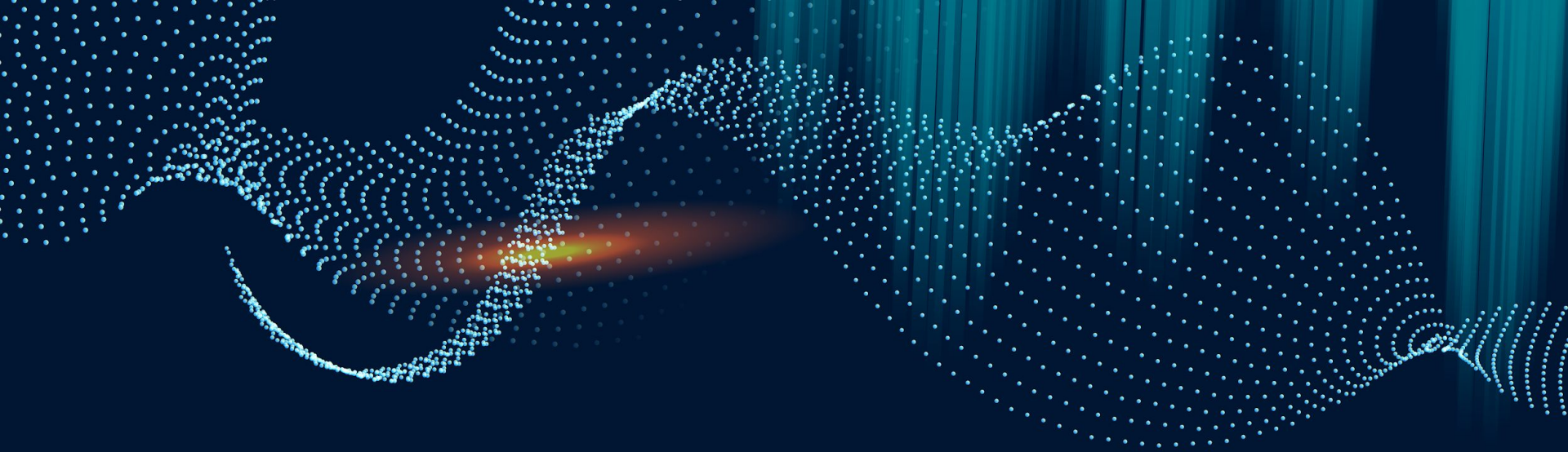
01 Goal

Expectations vs Reality

- Using multiple season's pitch-by-pitch data in combination with pitchers' seasonal overall stats to train prediction models
- Multiple ML models (XGBoost, Catboost, LightGBM, Random Forest)
- Deep learning modeling (RNN, LSTM, GRU

- Pitch-by-pitch data from one season (2023)
- One ML prediction model: XGBoost

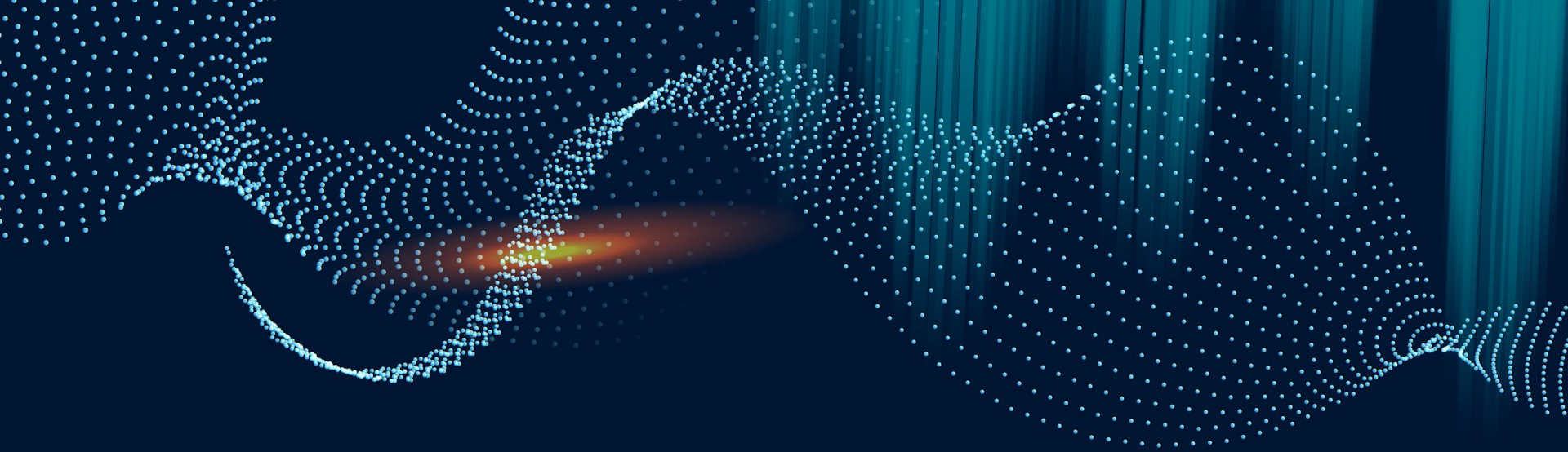




02 MOTIVATION

Why this topic?

- America’s pastime’
- Personal interest in sports analytics
- Baseball analytics (sabermetrics) is a very advanced field but is completely new to me
- Moneyball



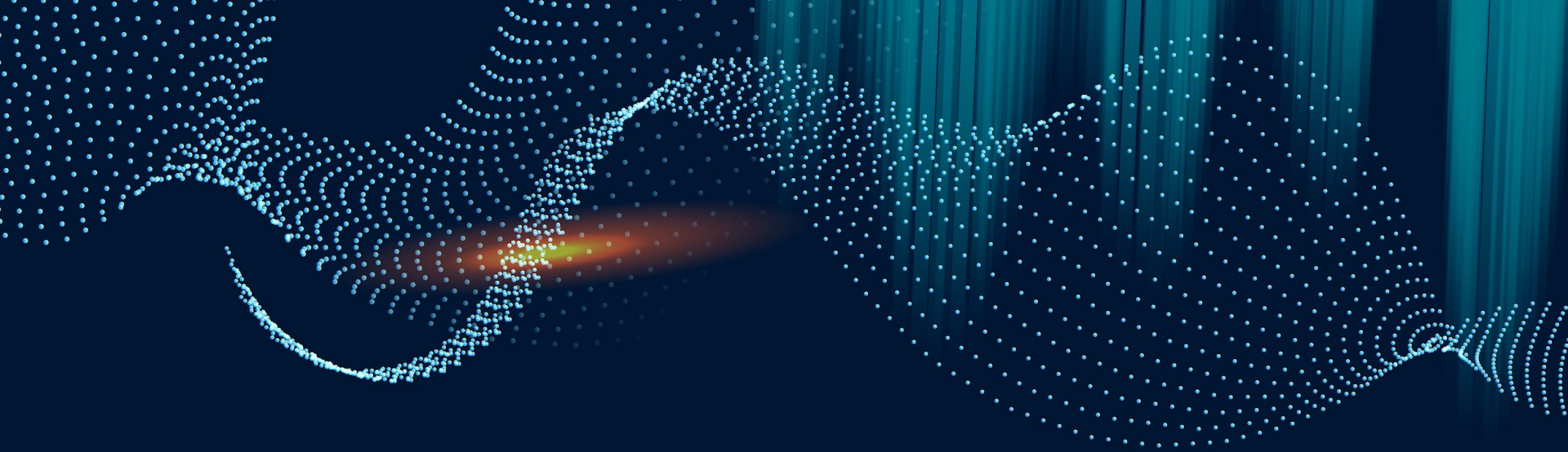
03 Data and Methodology

Data

- All data obtained through the PyBaseball package which allows a user to pull from various sources of MLB data
- Statcast full pitch-by-pitch data over several seasons
- From 113 features to just under 30 features after data cleaning and feature engineering'
- Target variable (pitch_type) condensed from 15+ classes to 4

Methodology

- XGBoost algorithm
- Train pitcher-specific models by looping through the pitchers that have a total pitch count above a certain threshold
- After each model is trained, predictions are performed and the model is evaluated on accuracy
- Each iteration also has a naive accuracy score which servers as a comparison to model accuracy
- Each trained model is then stored in a pickle file



04 Results

“It’s not the destination, it’s the journey.”

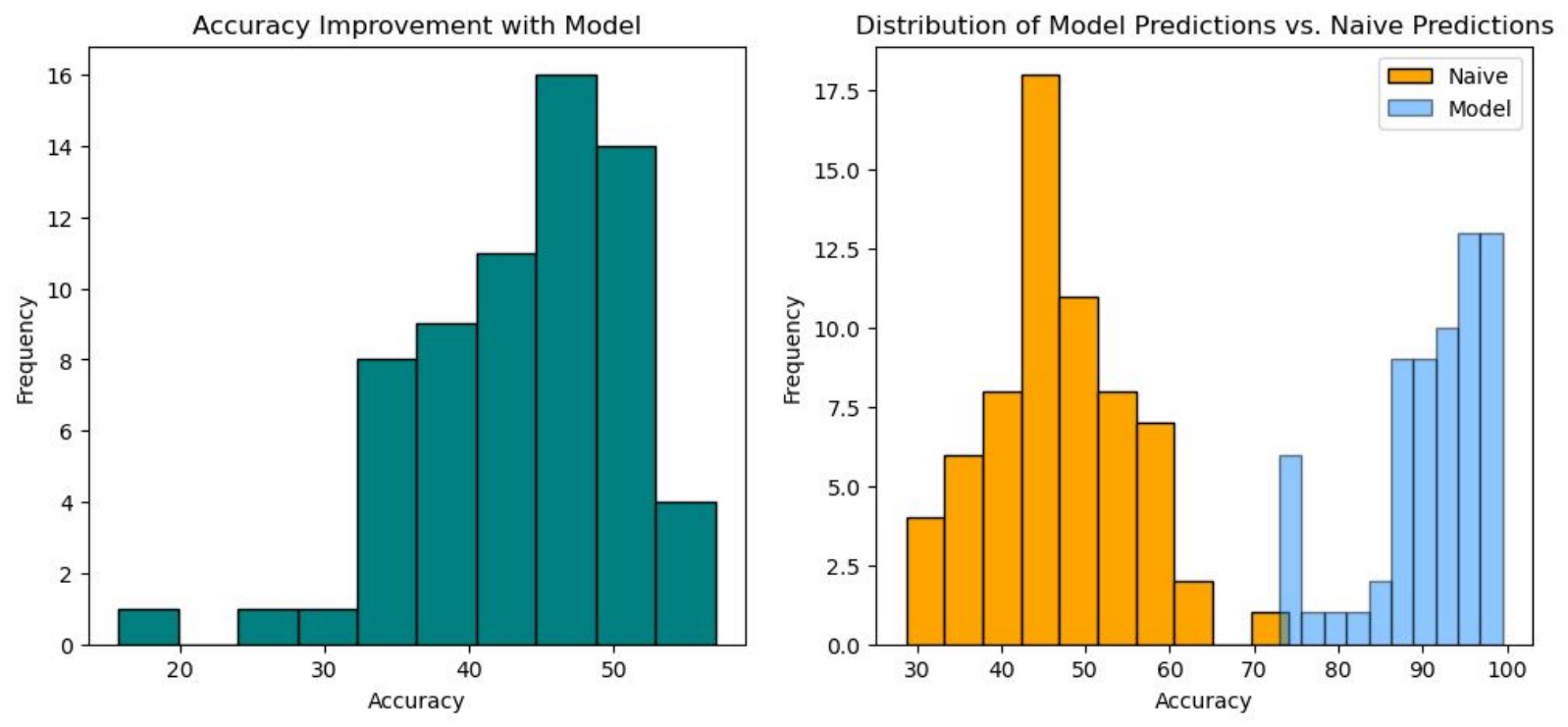
—Ralph Waldo Emerson

Model output

```
Pitcher ID: 622491
Pitcher's pitch map: {'MFastball': 0, 'Breaking Ball': 1, 'Off-Speed': 2, 'PFastball': 3}
Pitcher's pitch counter: {'MFastball': 556, 'PFastball': 1408, 'Breaking Ball': 710, 'Off-Speed': 500}
Number of data points in training: 2539
Number of data points in testing: 635
Best params: {'learning_rate': 0.1, 'max_depth': 5}
Total training time: 0:00:01.896866
Naive accuracy: 44.4
XGBooost accuracy: 80.6
```

```
Pitcher ID: 656756
Pitcher's pitch map: {'MFastball': 0, 'Breaking Ball': 1, 'Off-Speed': 2, 'PFastball': 3}
Pitcher's pitch counter: {'Off-Speed': 734, 'PFastball': 394, 'MFastball': 1427, 'Breaking Ball': 787}
Number of data points in training: 2673
Number of data points in testing: 669
Best params: {'learning_rate': 0.4, 'max_depth': 2}
Total training time: 0:00:01.604355
Naive accuracy: 42.7
XGBooost accuracy: 87.0
```

Naive vs Model Predictions



Limitations

- Only using data from one season and only pitch-by-pitch data, no batter data such as batter tendencies
- Condensing target variable down over-simplifies pitch types thus making prediction more accurate but not necessarily effective
- Model only takes into account the previous pitch, ideally the prediction model would take into account the past couple of pitches and even historical head-to-head data against the current batter

Future Work

- Implement real-time prediction using the pitcher-specific models
- Create more lagged features to take into sequential data beyond just the previous pitch
- Train models using other ML algorithms
- Train deep learning models for more powerful predictions