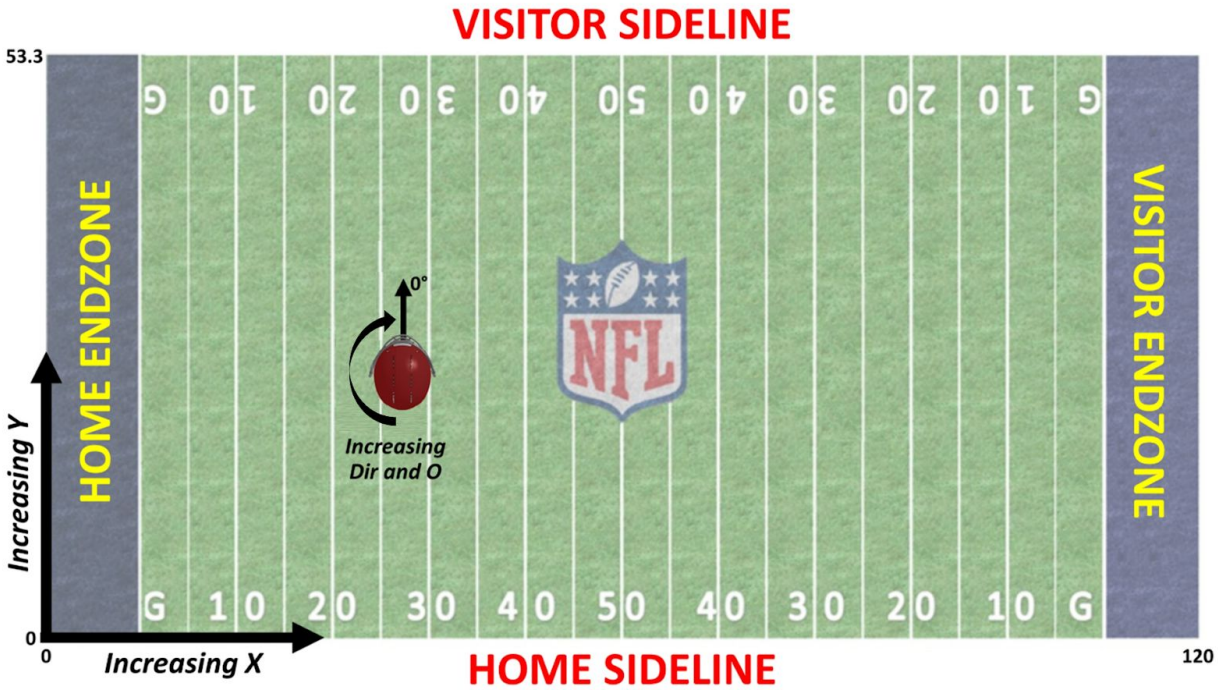
Title: Big Time Rush

Who: Gokul Ajith (gajith), Harrison Boyer (hboyer), Benjamin Deckey (bdeckey), Akhil Trehan (atrehan1)

Introduction:



Using a combination of over 40 metrics for thousands of rushing plays in NFL history, our project goal is simply to predict how far the team will rush for on a given play. The goal of a rushing play is for the offense to run (rush) with the ball, as opposed to throwing the ball, in order to gain yards, moving towards and across the defense's side of the field in order to score. The goal then of the defense is to stop the offensive team from scoring. The interesting aspect of this project is that the data is modeled from the image above, where positions, speed, orientations, field position, etc and this graph replication can lead to insights into the sport that does not solely depend on personnel. These positional metrics can also be combined with overall play conditions including weather, score, location, etc to From a specific football standpoint, deeper insight into rushing plays will help teams, media, and fans better understand the skill of players and the strategies of coaches. It will also assist the NFL and its teams evaluate the ball carrier, his teammates, his coach, and the opposing defense, in order to make adjustments as necessary. However, the deeper reason

we chose this project was that vast sports data of this specificity is not always available, and this dataset presents the opportunity to analyze a vast number of metrics to determine variances in similar outcomes. Simply the positions of all players on the field give an interesting graph problem to represent using a model, but the generated weights can also be very useful in determining which features are worth honing in on over others. This is our biggest motivator, as we attempt to see if deep learning can provide advanced insights into a combination of metrics that can generalize aspects of a professional sport to create a strong classification model as to the overall success of the play.

Data:

The dataset found comes from a competition started by the NFL. This data has information from thousands of rushing plays in the NFL, with each rushing play having 22 data points that represent various positional properties of each of the players on the field including X, Y location, speed, distance, acceleration, and orientation. Additionally, each play also has data on 40 other game properties including score, home/away teams, down and distance, field position, formation, turf type, and weather.

Our data is large in the sense that each of the 3000 plays (graphs) has 22 nodes that all need to know relative positions of each other and have individual properties to consider. Thus, there will be significant preprocessing in creating these Node, Edge and Graph objects with standardized properties for proper usage in the MPNN.

Methodology:

We hypothesize that training using an MPNN using DGL would best represent the parameters necessary to examine on a football field. We want to represent each play as a graph of 22 nodes, each player on the field, connected by edges representing positional vectors representing x, y, speed, acceleration, distance, acceleration, and orientation from each of the other players. These nodes will also contain features of the players themselves that can be used to draw patterns throughout the different graphs, including position, orientation, direction facing, and even height and weight. Then, these graphs of plays can be batched and used with message passing to predict how many rushing yards a given play will result in. We also want to consider non-player variables in the data such as down and yards to go, current game score, weather, etc. Since these variables are external factors related to the model their effect will be quantified as a collective bias and applied to the

model. An embeddings matrix will also be used to better understand how the model is weighting features for better insights into which are more important to consider relatively.

Training:
We will be training the model by batching together all graphs of plays and continually predicting possible yardage for a set number of epochs. As there is a large amount of data, over 60,00 rows, we plan to utilize GCP's capabilities to train the data. If using an MPNN does not result in high accuracy, and the model is not working as hypothesized then we plan to reimplement this classification problem using traditional CNNs with embeddings.

Metrics:
To constitute success we will be calculating the sum squared error on the test dataset that we will split and create. The notion of accuracy directing applies to our model as there exists a real answer as to how many yards each play resulted in, thus we are able to predict yards with some degree of accuracy. We plan to assess our data's performance by measuring its time and accuracy.

Ethics:
Who are the major "stakeholders" in this problem, and what are the consequences of mistakes made by your algorithm?

The major stakeholders in our problem are the players of the football teams, the managerial staff and ultimately the owners of the teams. If our model was successful in providing an accurate prediction given a group of players and a play, teams could theoretically use the information to make trading and drafting decisions. If this happened, it is guaranteed that some players will be undervalued and consequently either not get the job or get paid less because of their perceived less value. This would be the case because although the model takes into account many variables, it is impossible to fully describe a player's value to the team in a finite number of variables. Consider the ability of a player to bring the team together, this skill would not be valued in our model, whereas if the decisions were made by recruiters going and conversing with the players these skills would be hard to miss. Our algorithm also does not take into account chemistry between players. The algorithm could potentially make predictions that a certain group of players plays very well together, when

in reality their personal chemistry may cause them to play abnormally bad and therefore cause the coach to make poor roster decisions.

What is your dataset? Are there any concerns about how it was collected, or labeled? Is it representative? What kind of underlying historical or societal biases might it contain?

Our dataset is comprised of 49 different features provided to us by a competition posted on Kaggle by the NFL. The NFL used their real-time player tracking software which is collected by Next Gen Stats. Our data represents each players' features for every running play. It is hard to conclude any societal biases as our data does not record any culturally impacted features, such as gender, race or socioeconomic standing and the tracking software to collect the data is based on objective game footage.

Division of labor: Briefly outline who will be responsible for which part(s) of the project.

Project Work Outline	Preprocess	Hyperparameters and model set up	Training	Written work	Testing
Ben	✔			✔	✔
Harry		✔	✔	✔	
Gokul	✔	✔		✔	
Akhil			✔	✔	✔