Title: **Growing Plants Better with Machine Learning**
Team: Hannah Kronenberg and Jonathan Yushuvayev

Task T: Predict the number of yards gained on a running play of an american football game based on the state of the game at the time the ball is handed off.
Experience E: We will use a dataset of running plays during the 2020 NFL season that specifies the players involved in each play with their position at the time of the handoff as well as the teams, scores, time during the game that the play happened in, as well as the number of yards gained..
Performance metrics P: We will measure the mean squared error of our predictions over a held-out set of data.
ML Methods: Our first attempt will be to use boosting or random forests using only the position of the players on each time at the time of the handoff because these methods are very flexible and relatively straightforward to implement. Two further directions would be to use deep learning or to try to incorporate higher-level information about the teams and players. For example, the average yards carried per play during the previous season for the player receiving the ball.

Why we care about this: Hannah is generally interested in sports analytics and its use both by sports teams and in betting markets. Jonathan thinks this might be an interesting challenge.

Expected Challenges: One challenge we will have to deal with is the relatively limited amount of data as there are only a few thousand running plays in each NFL season. Another challenge is that the plays are not sampled from some distribution independently of each other. Instead, they are related to each other (i.e. plays in the same game or in different games but by the same team). The challenge will be to figure out how to use these relationships to improve the model. Finally, as the data is time series data we would ideally want to test our model on games that happen after our training data. This, however, makes cross validation a bit trickier.

Work Plan over the next ~5 weeks: our main steps will be

- Clean and preprocess the data: currently there is a row per play per player rather than one row per play
- Fit and evaluate simple models using only player location and movement features
    - We plan to spend a week and a half on the first two steps
- To improve performance, we plan to do a literature review on the following topics:
    - How to use team-level and player-level features in our model. Potentially use a latent variable approach or something similar
    - How to take into account the time-series nature of our data. For example weighing more recent observations more highly.
    - Which deep learning architecture is suitable for this problem.
- After reading about this, we will change our models and check if performance improves
    - We plan to spend two and a half weeks on the second phase
- Finally, we will try to incorporate outside data such as player height, weight, age, and career statistics (ideally prior to the beginning of our training data period.

Prior Work / References / Projects that inspired us (this part can be over the 1-pg limit):

1. This project is based on this Kaggle competition