1. This is a deceptively complex spatial optimization problem. I believe the goal should be to minimize run expectancy based on the individual positioning of each fielder.

**Contextual Features**

I strongly recommend including pitcher batted ball tendencies as historical context. Outfielder positioning should differ for a hitter based on the pitcher's spray distributions, handedness, etc.

**Core Framework Features**

- **Batted Ball Tendency Distributions** (exit velocity, spray, launch angle)
  - For both pitcher and hitter, with handedness splits
- **Game State**: outs, leverage, runners on, score
- **Fielder Attributes**: sprint speed, range, left/right fielding splits

*Excluded features (for simplicity and interpretability):* pitch type distribution by count, batter performance vs. pitch types/zones, Statcast pitch features, swing metrics. These add unnecessary complexity with only marginal improvements.

**Practical Considerations**

In real implementation, it would be impractical to communicate from the front office or dugout to the centerfielder where to position fielders on a per-pitch basis. Instead, we'll focus on a generalized per-batter approach that still gives players and coaches autonomy in their positioning decisions.

**Optimization Approach**

**Nearest-Neighbor Framework & Kernel Density Estimation:**

- For each pitcher-hitter matchup, identify the *k* most similar historical matchups using cosine similarity on pitcher/hitter features.
- Apply KDE to batted balls from these matchups to generate a continuous probability heatmap of expected ball locations.
- This heatmap indicates where balls are most likely to land in each outfield zone, informing optimal fielder positioning to minimize expected runs.

**Evaluation Framework**

Apply the model to historical game data, comparing recommended vs. actual fielder positions.

- Use xBA via Statcast to estimate hit probability.
- When high-xBA balls hypothetically become outs due to better positioning, this framework generates defensive value. Ex: a ball with xBA of 0.75 that becomes an out due to optimal positioning represents beating expectation by 0.75. That residual can then be combined with the run expectancy matrix to derive a measure of expected runs saved.

**Per-batted-ball steps:**

1. Calculate the xBA residual with actual positioning (binary outcome – xBA).
2. Estimate catch probability from the recommended position (opportunity time, distance, direction).

3. Calculate the expected xBA residual with recommended positioning.
4. Sum the difference across plays to quantify positioning value added.

**Supporting Metrics:**

- **Coverage Improvement:** Average distance from recommended position to actual batted ball landing location vs. actual positioning.
- **Run Value Saved:** Convert xBA improvements into changes in run expectancy (singles vs. extra-base hits).

The key challenge is estimating whether an optimally-positioned fielder would have made the play. Using fielder-specific range metrics and catch probability models makes this evaluation probabilistic rather than binary, accounting for the fact that positioning changes outcomes probabilistically, not deterministically. Another challenge is finding a way to penalize outlier balls that were caught that may have become hits with the optimal positioning.

For implementation, I envision outfield heatmaps for every possible hitter/pitcher matchup being made readily accessible by bench coaches, who can then direct the outfielders given nuances that can't be accurately captured in the data.

This is just one of many fun or interesting approaches. I'd like to explore reinforcement learning here as well.

2.1) **Goal:**
Determine optimal allocation of 150 percentile points to three hitter skills that are primary drivers of offensive production: **swing decisions**, **contact**, and **bat speed**.
We will use **xwOBA** to measure hitter value given its improvements over OPS and predictive nature. Future projects may explore using wRC+.

**Metrics:**
- **SEAGER (SElective AGgression Engagement Rate):** For swing decisions, Z-O Swing% is a logical starting point given the problem, but ignoring balls out of the zone (low chase) isn't enough. We really want hitters both taking unfavorable pitches (even if they are in-zone) and punishing favorable ones. Improving upon Z-O Swing%, SEAGER (created by Robert Orr, Baseball Prospectus) is calculated by subtracting a hitter's percentage of 'hittable' pitches taken from their percentage of 'bad' pitches taken (selectivity). SEAGER has a strong positive correlation to xwOBA.
- **Z-Contact%:** This metric measures the ability to make contact with true strikes, as quality of contact significantly decreases on pitches outside the zone. Whiff% could also be considered for a general measure of contact ability.
- **Avg Bat Speed:** A straightforward measure of swing speed, which has an inherent positive correlation with Batter Run Value (via Statcast).

**Assumptions:**
- **SEAGER, Z-Contact%, Bat Speed:**
  - Relationships are nonlinear with themselves and with xwOBA.
  - Their effects on xwOBA are non-independent.
  - Relationships to xwOBA are stable over multi-year samples.
- **xwOBA:** Is the optimal predictor of batter value for our purposes.
- **Skill Percentile Points:** Are accurate (50 is average, 1 = worst, 100 = best).
- **Data Access:** Access to xwOBA data per season per hitter via Fangraphs or internal sources is available.

**Framework:**
1. **Metric Interaction Evaluation:**
   - Evaluate metric interactions using a correlation matrix to understand their mutual influence across all qualified hitters (based on pitches seen, swings, balls-in-play, etc.).
   - Year-to-year correlations between Z-Contact%, SEAGER, and xwOBA are publicly available via Robert Orr.
   - Additional correlations with Avg. Bat Speed can be determined from internal data or by joining and filtering/averaging Statcast data.

2. **Modeling:**
   We will use both a Generalized Additive Model (GAM) and an XGBoost model for predicting xwOBA. The GAM will prioritize interpretability, while the XGBoost will focus on pure accuracy.

**Generalized Additive Model (GAM):**

- Visualizes the nonlinear relationship between each individual skill percentile and xwOBA.
- Generates contour plots to show the relationships and tradeoffs between skills and their combined effect on the target.
- Model summary will inform us about the effectiveness of our three features in predicting our target.

**XGBoost Model:**

- Produces feature importance/SHAP plots to illustrate how the model selected features through its decision tree process.
- From preliminary research on the relationship between SEAGER and Z-Contact%, I anticipate models will favor higher SEAGER and Bat Speed, and potentially lower values for Z-Contact%.

3. **Optimization & Validation:**
    - Independent grid searches will be run on both models, incorporating the 150-point constraint to identify the ideal allocation of points from each model.
    - To ensure robustness and minimize overfitting, K-Fold Cross Validation will be applied to the GAM, and GridSearchCV will be used for the XGBoost model. This will confirm each model's ability to generalize to new hitter data.

2.2) **GOAL:** Determine the bare minimum for being a productive hitter

Our three foundational skills for being a productive hitter:

- **Swing decisions**
- **Bat-to-ball**
- **Bat speed**

I've picked three stats capturing these skills:

- **SEAGER (SElective AGgression Engagement Rate):** How well a hitter ignores pitches they can't do damage on & how well they punish pitches they *can* do damage on.
- **Z-Contact%:** How often hitters make contact swinging in-zone (hitters making contact out-of-zone are already lacking in swing decisions)
- **Avg Bat Speed**

We are going to stick these into both a simple and more advanced model.

The simple model is going to clearly graph out how percentile points in each stat individually influence real offensive production (xwOBA), because higher may not mean better for the above stats. (Ex: Fastest swing in the league doesn't mean he'll hit well if he's only swings at junk). The model then makes predictions for a hitter's production based on what it learned as the value of the three skills.

We'll then run a function on the model that checks every combination of 150 points for those skills and predicts the offensive production. The optimal combination will be whichever produces the highest predicted xwOBA. We'll double check it works for new hitters too, not just the ones in the data.

The advanced model will be a similar process but a little more accurate. That model's graphs will show how important each skill is. We'll run a similar combination function to get the optimal 150 point skill combination.

3) My most transformative period started in my third year of undergraduate when I was taking upper-division courses for my degree. For my entire academic career up to this point, I had been coasting. My habits and daily routines were a travesty, and I had no clear plans for my future. This came to a head where for the first time in my life, I had to retake a class. I didn't understand any of the core material of my major, and that forced me to confront the limitations of my mindset and reflect on my career aspirations.

I vowed to reshape the habits that governed my life. Starting fresh, I had a daily goal of being one percent better than yesterday. I set schedules for work, study, and sleep, replacing bad habits with hobbies and exercise. I reached out to classmates for help, and built strong relationships with an academic support network. As a positive consequence, my grades improved drastically, and I used this momentum to apply my knowledge in math to what I loved: baseball.

I worked quickly and consistently to earn a research internship with UCSB Baseball, which opened up my passion for working with baseball data. It was here I realized how much I wanted to work in baseball analytics, so I set my sights on earning a Master's in Applied Data Science.

This transformative period hasn't really ended. Those good habits have made me more driven, diligent, and curious in all aspects of my life.