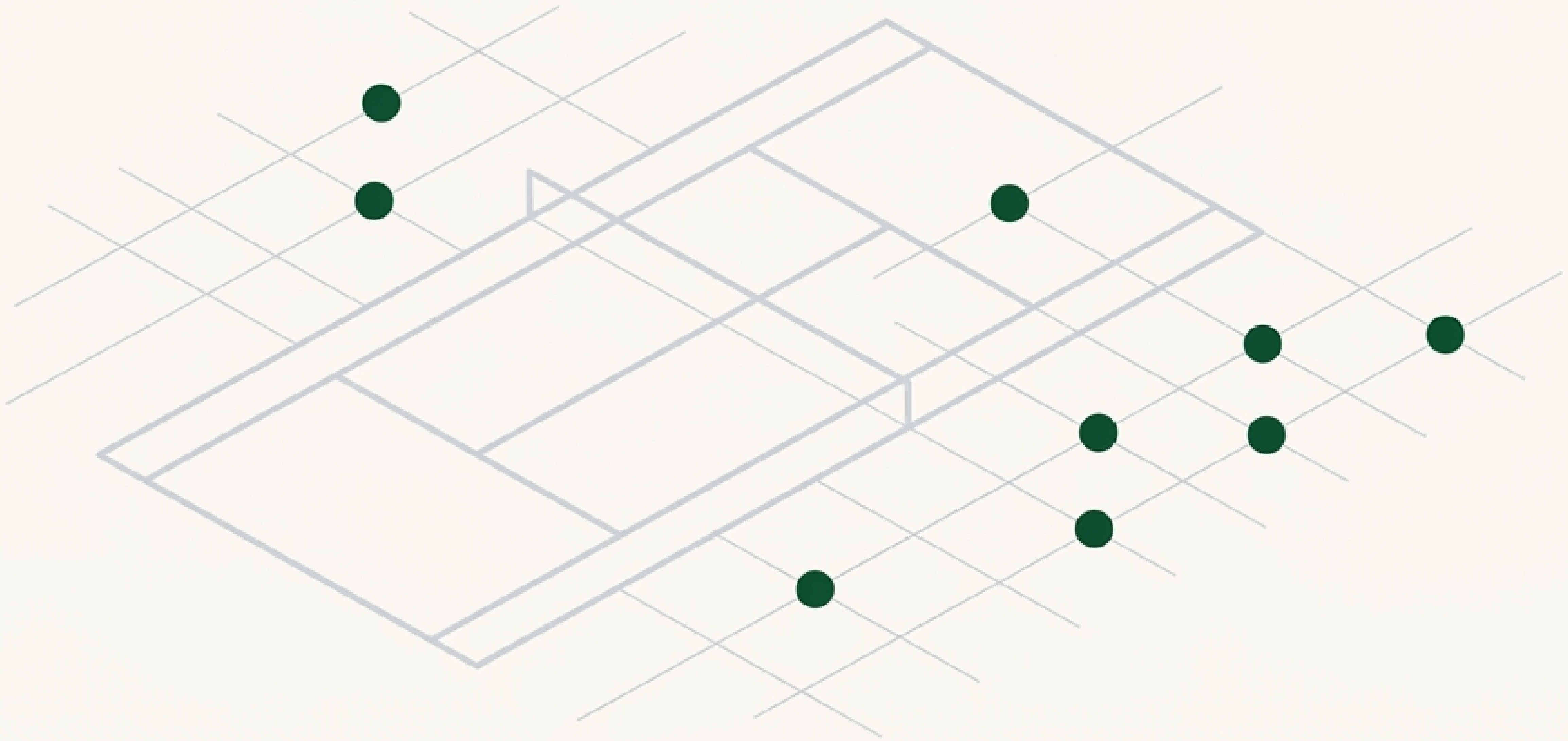
Cracking the Code of Tennis Upsets: A Data-Driven Journey

An analysis of predictive modeling performance and limitations using ATP match data.

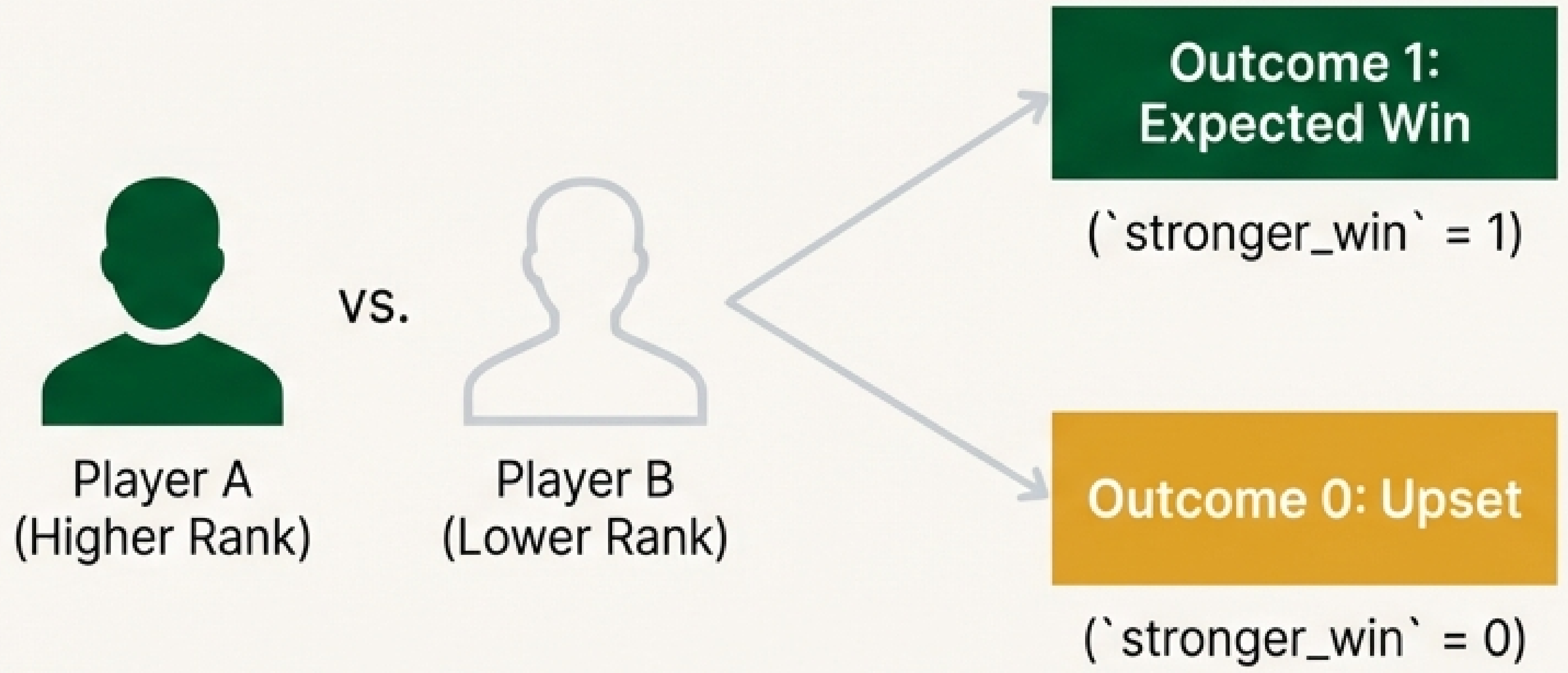


NotebookLM

The project objective was to predict if a higher-ranked player would win, a classic binary classification problem.

We aimed to build a model to predict the outcome of professional tennis matches using only pre-match information.

- **The Task:** Binary Classification
- **Target `stronger_win`:**
  - `1` = The stronger (higher-ranked) player wins.
  - `0` = An "upset" occurs, and the lower-ranked player wins.

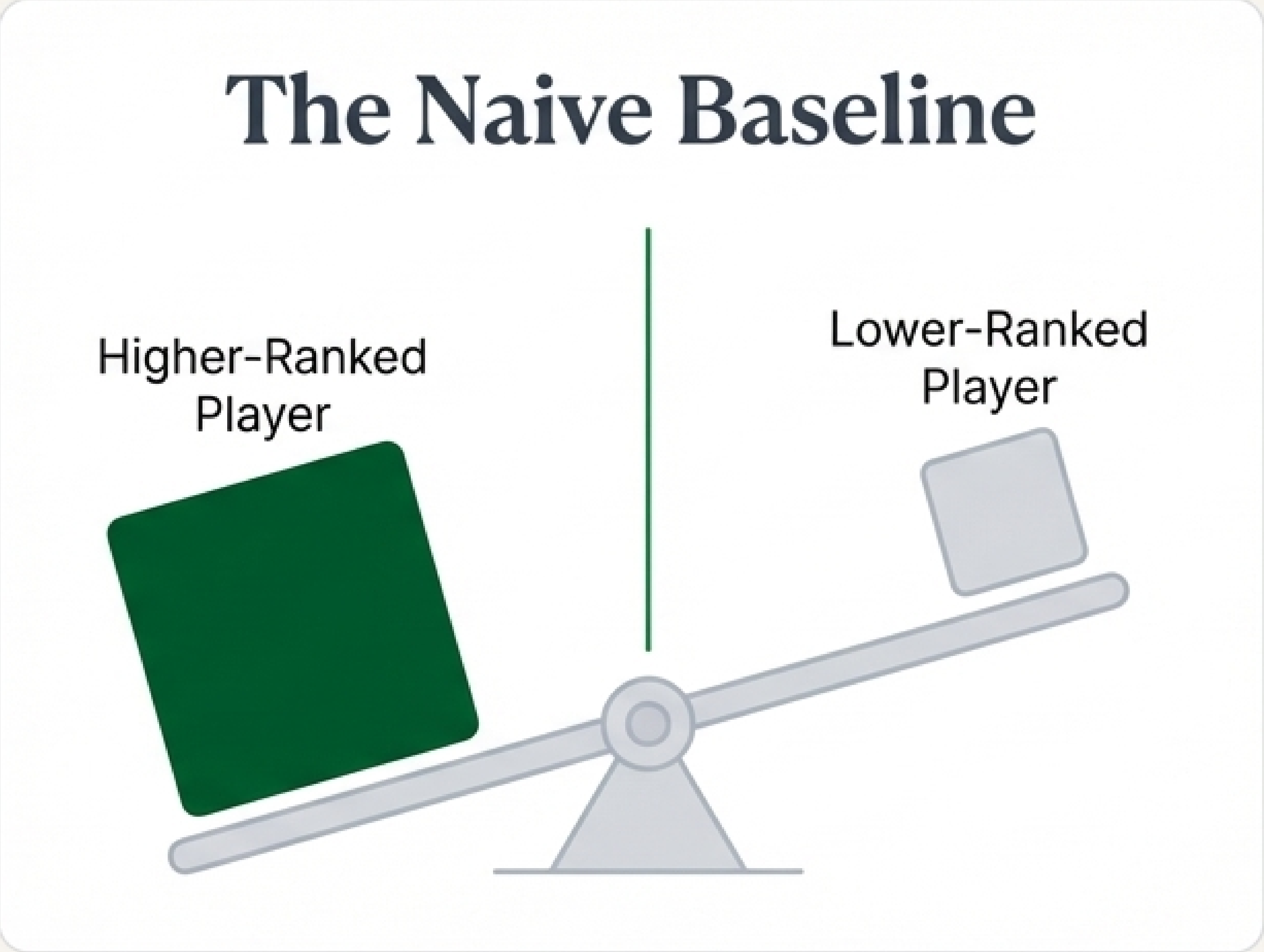


Any new model must first beat a simple but powerful baseline: always betting on the favorite.

Before building complex models, we established a "naive baseline" to measure against.

This baseline represents the conventional wisdom and sets the minimum performance bar.

Baseline Logic: The model always predicts the higher-ranked player will win (predicts `1` for every match).



This “naive” approach is surprisingly effective, setting a high accuracy bar of 61.4%.

Simply assuming the higher-ranked player will always win correctly predicts the outcome in the majority of cases.

This high baseline demonstrates that any sophisticated model must find a real, tangible edge to be considered valuable.

61.4%

Naive Baseline Test Accuracy

We engineered a machine learning approach using key pre-match data to find an edge.

Our strategy involved training and evaluating several models on a curated set of predictive features from the Ultimate Tennis Matches Dataset (ATP).

Features Used

Numeric:

- `rank_gap_abs`
- `age_diff`

Abc Categorical:

- `surface`
- `tourney_level`
- `best_of`

Models Explored

- Decision Tree
- Random Forest
- XGBoost (Baseline)
- **Final Model: XGBoost (Tuned with `RandomizedSearchCV`)**

Our final tuned XGBoost model achieved a test accuracy of 61.8%, a slight but real edge over the baseline.

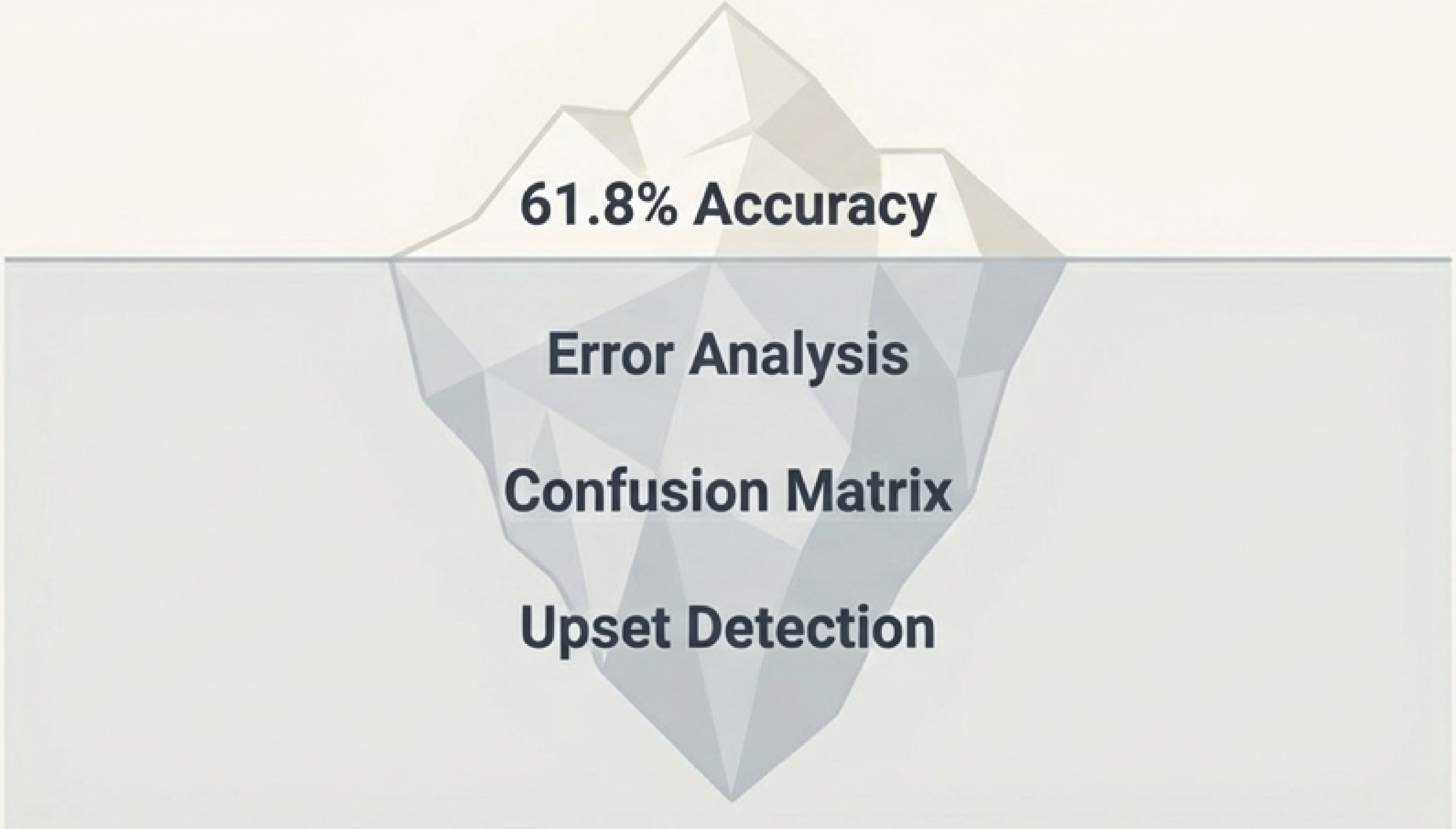
The tuned XGBoost model successfully outperformed the naive baseline, demonstrating that the features contain predictive power, however marginal.

Naive Baseline		Tuned XGBoost Model
61.4%	+ →	**61.8%**

Improvement: +0.4%

But accuracy is a misleading metric here; a deeper look at the model's errors reveals its critical blind spot.

While our model is slightly better at predicting the overall winner, its real test is whether it can identify the rare and valuable upsets. A simple accuracy score hides this crucial detail. To understand the model's true behavior, we must analyze the *types* of mistakes it makes.



The confusion matrix shows the model's core weakness: it overwhelmingly fails to predict upsets.

Actual Outcome	Predicted Outcome: Win	Predicted Outcome: Upset
Win	302 Correctly Predicted Wins	37
Upset	174	MISSED UPSETS 39 Correctly Predicted Upsets

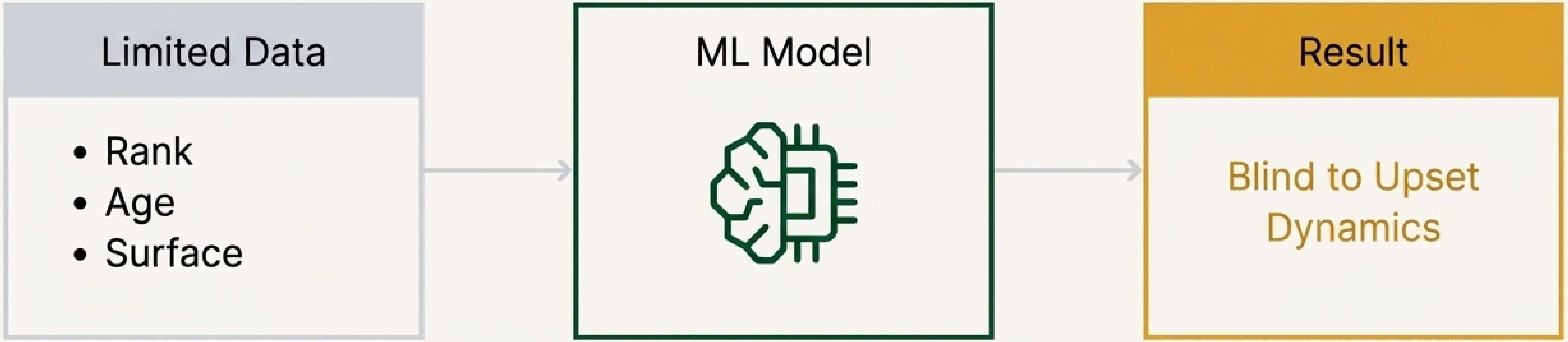
False Upset Alarms

The model correctly identifies when the stronger player will win, but it struggles to identify the rare upset events.

Key Insight:
The model missed 174 upsets, predicting the stronger player would win when they didn't. This is its primary failure mode.

The insight is clear: pre-match stats alone are insufficient to capture the complex dynamics behind an upset.

The model's performance suggests we've reached the predictive limits of the current feature set. Upsets are rare events driven by factors that are not present in our dataset. ATP rankings do not fully represent a player's real-time strength, and matches are not isolated events.



Our dataset lacks the crucial context that often signals a potential upset.

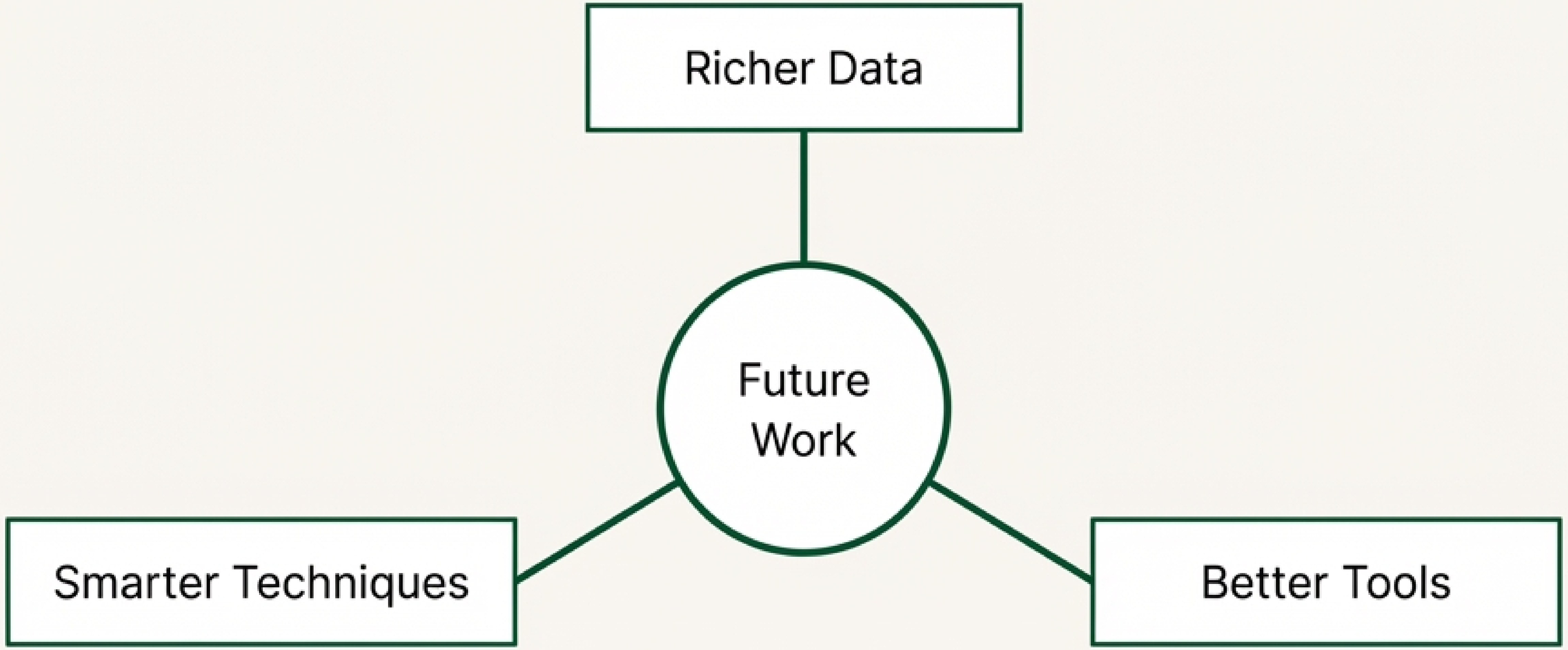
The model is effectively flying blind to the most important real-time performance indicators.

Missing Contextual Factors

- **Player Form:** Recent win/loss streaks, performance on the current surface.
- **Physical State:** Known or recent injuries, on-court fatigue from previous matches.
- Matchup Dynamics: Head-to-head history between the two players.
- **Temporal Trends:** The static nature of the data ignores a player's evolving performance over a season.

The path forward requires enriching our data and refining our modeling techniques to target upsets directly.

Our findings from this phase provide a clear and targeted roadmap for future work. Instead of seeking incremental accuracy gains, the next phase must focus on acquiring the right data and using techniques designed for the core challenge: detecting rare events. We propose three key enhancement paths.



We propose three concrete paths to build a more powerful and insightful prediction system.

Each proposed enhancement directly addresses the limitations identified in our analysis.

1. Incorporate Richer, Dynamic Player Statistics

- Integrate Elo ratings for a more fluid measure of player strength.
- Engineer features based on recent match performance (e.g., win percentage in the last 3 months).
- Add surface-specific strengths and statistics.

2. Address Class Imbalance with Advanced Techniques

- Implement `class_weighting` in models to penalize “missed upsets” more heavily.
- Explore more advanced solutions like `focal loss` to help the model focus on hard-to-classify examples (i.e., upsets).

3. Develop an Interactive Dashboard for Practical Use

- Create a tool allowing users to input match information (players, surface, etc.) and receive predictive insights and probabilities.
- This moves the project from a research exercise to a practically useful application.