## NFL NEXT GEN STATS, POWERED BY AWS

# Draft Score Q&A

Q&A with the NFL's Next Gen Stats about Draft Score, a prospect evaluation tool powered by AWS, to better understand the technology and science it takes to make the grade.

With shocking trades, big signings, and the 2022 Draft just days away, this NFL offseason has

been one for the history books. As the draft's first selection approaches, we sat down to talk X's and O's with the NFL's Next Gen Stats about Draft Score, a prospect evaluation tool powered by AWS, to better understand the technology and science it takes to make the grade.

**Next Gen Stats: Mike Band, Amy Lee, and Jonathan Jung**
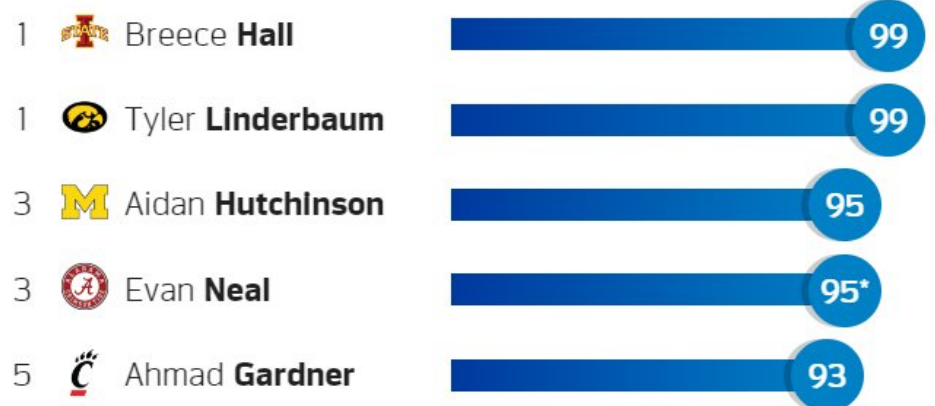


## 1. Give us the main idea behind Draft Score and how it came together.

At Next Gen Stats, we've focused primarily on live game NFL player tracking data. That meant we were only tracking data six months out of the year — from August through February. We wanted to apply our analytics through Machine Learning modeling to the NFL Draft. So, we asked ourselves, "How do we communicate advanced analytics of the data that a team would have on an NFL pro prospect coming out of college? And how do we then tell the story about which player a team should lean towards picking over another?"

The key was communicating the results of the model in a digestible way. That's where draft

scores came in. Think Madden ratings from the video game. The concept is to take the output of a series of models and communicate it in a way that scales to a single number in a way that people already understand. It gives fans a sneak peek into what each team's analytics staff is doing to prepare for the draft.

## 2. Does Draft Score translate equally between players and positions? How does it describe player potential?

We have models separated by position: Quarterback, Running Back, Receiver, Tight End, Tackle, Interior Offensive Line, Edge, Defensive Tackle, Linebacker, Cornerback, Safety. Every position has different size, athletic, and production-based features that are more important than others.

We assign each player a Production Score, which is how productive you were at the college level, and an Athleticism Score, an athletic performance metric. And then, we merge everything into our other models to create an overall score.

The college data we can access for the model is limited to what's publicly available. The teams have much more robust data collected by scouts, so this year we've expanded our Production Score to include a scout grade from our own Daniel Jeremiah and Lance Zierlein. The way we've structured the models allows us to add in more components to achieve a better predictive outcome in the future.

## 3. What part does the Combine play in a Draft Score? What if the player doesn't or can't participate?

Combine data can be missing. And missing data is a huge issue with this type of analysis. If the Combine data is missing, we look for Pro Day data. But if a player did their 40, and we know their weight and height, we've trained a simple regression model based on existing fields to estimate their three-column measurement. We've even found we can generate an Athleticism Score with an estimated 40 time, though we warn against making strong deductions on a player's athletic ability without such measurements. Players with missing Combine or Pro Day data enter the draft with more uncertainty than players that do work out.

## 4. Explain player and draft class modeling for us regular folks. What goes into creating a score?

We have a small data set. Just over six thousand lines of data. That's how many players that have been invited to the Combine since 2003. We have a model for every single year of the Combine. And that group of models is under an umbrella, like the Athleticism model. And then we separate that out by each position. We've found the ensembling of a series of models to exemplify certain dimensions of players got us to the point where we have more than 3,000 models for these 6,000 rows.

The modeling process works like this: We model every draft class (since 2003) against history, leaving a single draft class out. In other words, a prospect's score is generated from a model that did not including the player in the training set. Hence the need for so many models. This was big for reducing overfitting when back-testing the model results.

So, the 2022 class has 19 historical models to build from. Ultimately for this year's class, we have the 11 positions times five outcome models times 20 historical models for a total of 1,200 models to output one Athleticism Score alone.

For each dimension (athleticism, production, and overall), and class (between 2003 and 2022), we're looking at a number of variables: the probability of starting as a rookie; the probability of starting your second year; the probability of starting your third year; the probability of starting in your first three years at all; and the probability of making the Pro Bowl in your first three seasons.

And how did we pull it off? Our engineering team is comprised mostly of software engineers, so we didn't come into this with prior knowledge of how to set up a Machine Learning workflow of such complexity. Our interaction with Amazon's ProServe team helped us do our work correctly and efficiently. Interacting with these data scientists helped us expedite our process, so we could focus more time on analyzing data instead of organizing it.

# THE NEXT GEN STATS DRAFT MODEL
## PREDICTING THE SUCCESS OF THE 2022 NFL DRAFT CLASS

The Next Gen Stats Draft Model analyzes historical NFL Combine, pro day, and college production data to predict a prospect's chances of success in the NFL.

| KEY | The model identifies thresholds and interactions of the most important features (by position) that best predict success in the NFL |
| --- | --- |
| DATA | Historical NFL Combine results (since 2003), college statistics (since 2005) |
| OUTPUT | Probability of becoming a starter or making a Pro Bowl within first three NFL seasons |
| MODEL | XGBoost (ensemble decision-tree-based algorithm) |
| RESULTS | The results are transformed into composite scores, representing the athleticism, college performance, and overall profile of an NFL prospect |

## 5. How have the scores done in predicting successful players?

Pretty good! This is the third year we have released our list of seven "can't-miss" prospects. Looking back, our top 14 prospects from the past two years have included all four AP Rookies of the Year — Ja'Marr Chase and Micah Parsons in 2021, and Justin Herbert and Chase Young in 2020.

Here is an in-depth article we wrote that looks at our are the top seven can't miss prospects seven for the 2022 NFL Draft. These "blue chip" prospects each have a Draft Score of at least 91 and project to be top-50 picks.

## 6. How does your relationship with AWS contribute to your final product?

All of our data is stored within RDS (Amazon Relational Database Service), and that's been helpful for us because, typically, we've been working with document databases rather than normalized tables. As we started this project and realized that we had a lot of different datasets that had relationships, it only made sense to store them in a SQL or SQL-like database.

Using Postgres made it easy for us to start up, because it was already within our AWS ecosystem. Migrating the data over was pretty easy. And we can connect our RDS to QuickSight. Not only do our partners have QuickSight internally, but every club has it, so we get calls saying, "Hey, can you tell us more about the Draft Score?"

## 7. How does AWS S3 Cloud Storage help out in the process?

Managing the versioning and data sets of thousands of different things all at once gets really complicated. We leverage s3 to do all the version management for us. The actual use case is a simple call through the SDK. We just need to store a lot of iterations of data and iterations of the models. We need to save everything every time we rerun something, and we attach a unique timestamp to it so that under those buckets are unique folders. S3 has made our iterations much quicker, and we don't lose track of our model.

## 8. What's next for Next Gen Stats and AWS? What is on your wish list to add to future Draft Scores?

Right now, we track countable player data, but we don't get any data from tracking player movement with computer video analysis. What if a player ran a 40 at 4.55, 4.59, and 4.65? You can imagine a player doesn't run in a straight line over 40 yards. They could be running diagonally. They might actually have run 41.5 yards in a dash. What if we could analyze player movements in space in the 40 and other drills to determine a level of "true speed?" We could go from seven or eight metrics that we get from Combine data to an ocean of engineered metrics.
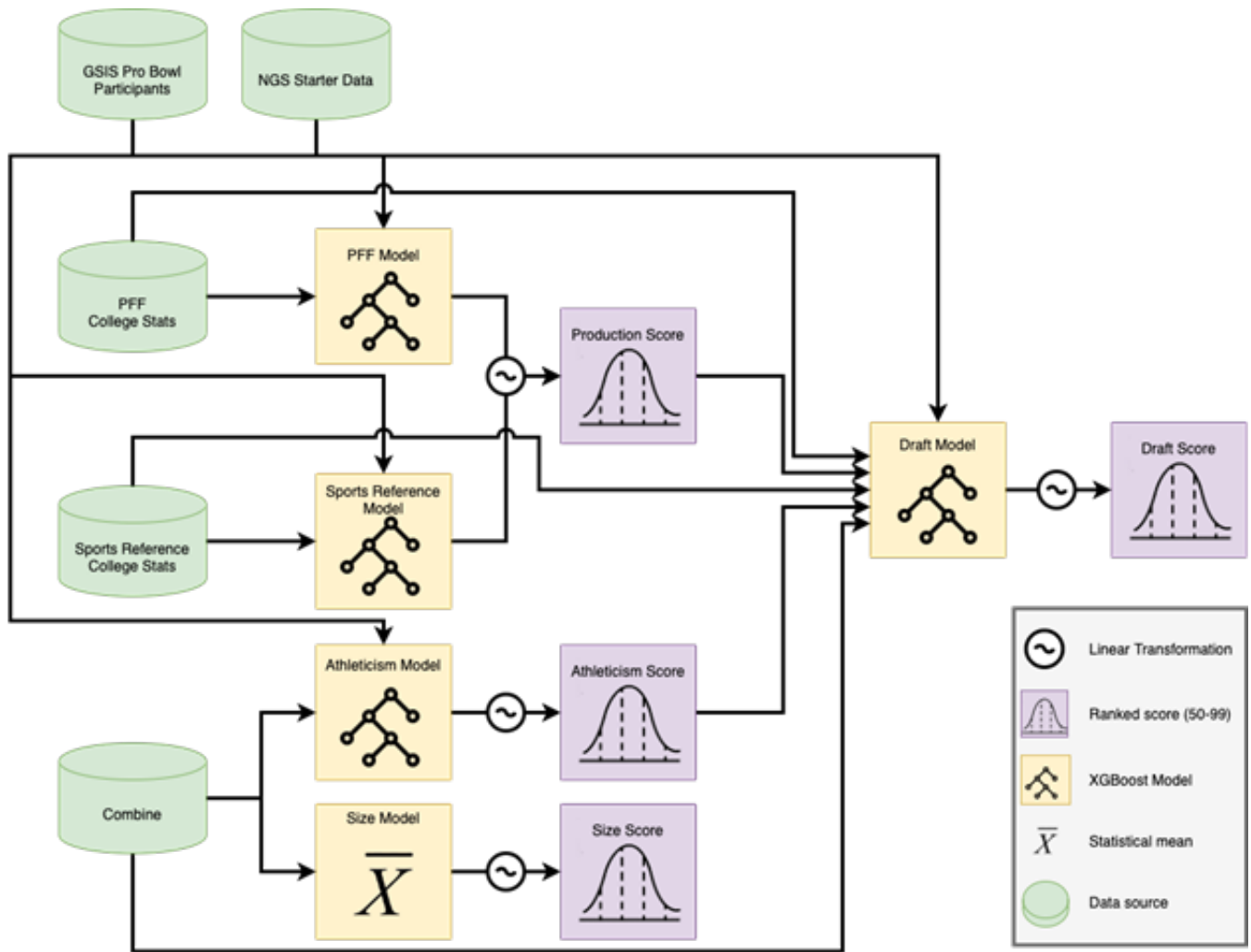
**Ready to get Started?**
**Learn more about working with AWS Professional Services.**

Get Started