STOP SHY OF THE FIRST DOWN*

PETER COTTON†

On the first down, an NFL player should stop shy of the first down marker unless they can make several extra yards on the play. The decision can be clarified by calculating implied yards per possession - a notion that may also help coordinators improve their second and third down strategy.

1. The First Down. The year 2020 marked the 100th anniversary of the National Football League. The game has undergone many changes, but statistical analysis of strategy has been something of a late starter. For example, half that history expired before a serious discussion of punting strategy occurred and to this day, some armchair statisticians are driven to distraction by questionable fourth-down decision making.

Here I go after a different sacred cow - the first down. I argue that players are making a mid-play strategy error on the first down that is materially impacting their team's chance of winning the game. The error isn't a lack of aggression, as with punting too much, but the opposite. They are fighting their way to the first down marker too often.

While *not* getting first downs may seem like an odd ambition for a football team, and hard to accept for player's conscious of their individual statistics, it's often optimal. I suggest that on a first-down play, no team should want a first down completion *unless a few extra yards are thrown into the bargain.*

2. The Choice. Even a casual observer of the NFL will notice that commentators, fans, and coaches alike uniformly encourage their players to lunge, stretch, hurdle, or bulldoze their way across that last yard to move the chains and get a first down. Getting the first down is a motherhood issue. It's applauded. For example, how often do we see a wide receiver break the imaginary yellow plane in an acrobatic fashion, often one arm outstretched, as they careen out of bounds?

But using data from the 2009-2013 NFL seasons compiled by Ben Dilday [1], I question this "decision" (or perhaps we should call it a non-decision). It is a decision made every time a player decides to put their body on the line to reach the first down marker.

Let me be clear about this scenario. First, we are only talking about the first of four downs.[1] Second, I assume a player with the ball is certain he can't progress terribly far beyond the first down marker. We're talking about that last effort to break the plane. Obviously, a player in full flight who is likely to achieve the first down and also many extra yards would be silly to stop one yard short.

Under my stated conditions, in choosing to complete the first down, or not, I shall assume that the ball carrier is determining which of these two outcomes will ensue:

*For review purposes only

†Intech Investments (peter.cotton@intechinvestments.com). We thank reviewers for helpful feedback.

[1]The "first down" refers to the first of four attempts to advance ten yards, and also refers to the act of achieving the 10-yard advance, thereby resetting the down count back to the first down. Due to penalties, the first attempt may sometimes require 5 or 15 yard gain in order for a completion of a first down to occur, but usually completion of a first down is synonymous with a ten yard gain or more.

1. Second down and one yard.
2. First down and ten yards (but with the line of scrimmage one yard further down the field).

The second possibility seems to be the strongly preferred option, given two seeming advantages:

- The field position is advanced one yard.
- The team will have one extra down.

However, there is one big disadvantage. The number of yards required for the next down reset will be ten, rather than one. I ask the reader to focus on this downside when considering the completion of the first down. Conversely, there is a positive spin to a perceived "failure" to get the first down: *you are probably going to get it anyway, just further down the field.*

Due to this likely symmetry, it's clear that the trade-off comes down to how much we value yards gained as compared to the small chance of losing possession. To formalize this and provide what I hope is a simple perspective, this article is broken down as follows.

1. A brief comment on ways to estimate the empirical value of possession, measured in yards.
2. An estimate of *implied yards per possession* when a player chooses to complete the first down, based on second and third down completion statistics.

I suggest that by completing the first down, players, and by implication coaching staff, are valuing possession far too highly relative to yards. I consider it my duty to convince the NFL offensive coordinators to order their players to do what seems most unnatural.[2] Instead of reaching for the first down, players should slide, run out of bounds, or otherwise stop the advance at the nine-yard line. This makes for quite the break in tradition.

Conversely, defensive coordinators should not be cursing players who allow a ten yard gain when a nine yard stop seemed possible. Those players have done their team a favor in conceding the first down. As controversial as this may sound, the logic is simple granted an upper bound on the value of possession relative to yards.

Rather than present a definitive calculus for possession value, I'll mention several ways to come at this that are all more than adequate. The decision to stop shy of the first down is in not predicated on a precise estimate of points per yard, and a rough number will do.

3. The value of possession, in yards. The value of possession versus yards is old topic. I'm not going to get into endgame analysis or special situations, nor the value of the clock when one team is ahead - though I think it should be clear how the rationale extends to these situations.

I shall be content with somewhat typical field and game position, which, if you prefer, can be assumed to occur in the first three-quarters of a relatively even game. I assume that the game is not so lopsided that one team has significantly diverged from a strategy that maximizes mean points scored.

The value of field position in points was considered by Virgil Carter and Robert E. Machol back in 1970 [3]. Their table of field position values is presented in Figure 1 and, while some improvements are possible, we can read off the difference between 15 and 85 yards. This provides a very loose estimate of how expected points varies - namely one point per 12 yards. Combined with some reasoning about punts, even

[2] Actually, some certainly appreciate this already, based on their initial reaction to this idea.

TABLE I

THE EXPECTED POINT VALUES OF POSSESSION OF THE FOOTBALL WITH FIRST DOWN AND TEN YARDS TO GO FOR VARIOUS TEN-YARD STRIPS

Center of the ten-yard strip (yards from the target goal line): X	Expected point value: $E(X)$
95	−1.245
85	−0.637
75	+0.236
65	0.923
55	1.538
45	2.392
35	3.167
25	3.681
15	4.572
5	6.041

FIG. 1. *Table of field position values from Carter and Machol [3].*

this early analysis may suffice.

Coming at things another way, the value of a yard is arguably easier to estimate on fourth downs rather than first - at least if the team is in field goal range and will with certainty kick. Then, the value of a yard can be inferred from field goal kicking probability (the blue line's slope in Figure 2) since the only value of a yard is a reduction in the probability of missing.

This plot is by Benjamin Morris [5]. Since the slope clearly varies, so does the value of a yard. The value of a yard on first down might be viewed as an averaging of these results, somewhat smoothing out the differences in slope, but we will still have different values of points per yard.

Incidentally the author points out that the value of "a" yard (or at least some yards on the field, within reasonable field goal range) changed between 2004 and 2014. As kickers got better, the value of a yard on fourth down decreased.

To proceed to a slightly more careful estimate of first down field position, the expected value of the next score by either team could be estimated. For example, Figure 3 shows the expected points as a function of field position compiled by Brian Burke [2]. We can read an approximate value of possession from this plot or the accompanying data, as follows.

Notice that 66 yards (i.e., our own 34-yard line) corresponding to +1 points, assuming we have possession. Now imagine we punt it to our opposition. They start their next play on their own 34-yard line after an entirely plausible, if unspectacular, net punt of 32 yards. So now they are +1 points. Net, we have lost two points.[3]

To emphasize the role of the punt, notice that the intercept with the x-axis occurs near the 85 yard line. Let's place a Herculean punter there who can always pin an opposition to their own 15-yard line (for an amazing net 70-yard punt). What a handy

[3]Or perhaps the punt is better. They start at their own 20 instead, corresponding to an expected 0.5 points per possession. Net, we have lost 1.5 points. These illustrations are intended only to make the point that in the upper envelope of estimates, a possession is still only worth a little more than two points on average.

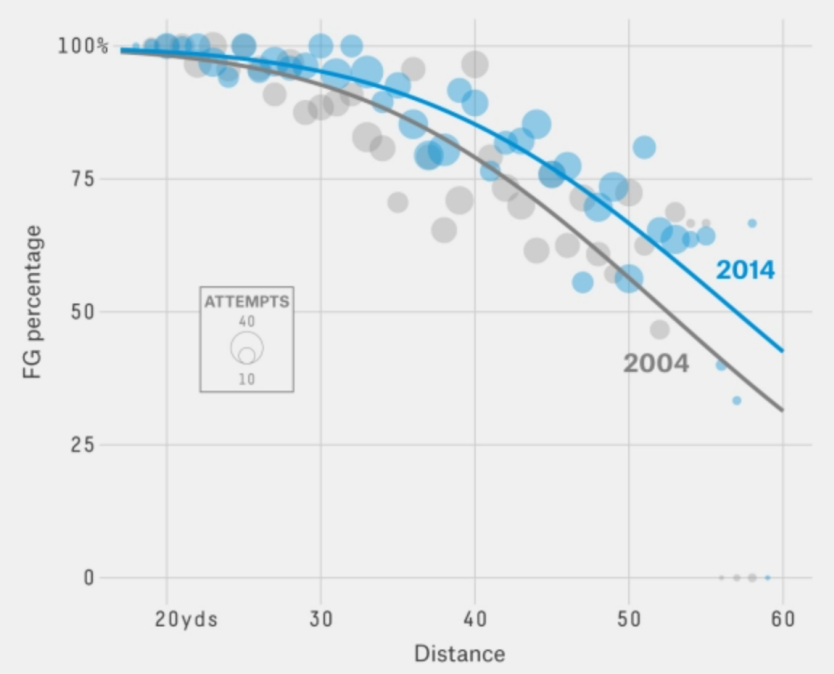

FIG. 2. *Field goal percentage as a function of distance. The value of a yard on the forth down, measured in points, can be estimated from the slope of the curve. Plot by Benjamin Morris from the website 538 [5].*

asset that punter would be. A team transferring possession in this manner has lost absolutely nothing!

We also note that the value of possession on first down varies a little less than it does for fourth down, as we expect given the kernel smoothing effect of the three proceeding plays. Granted, there is still some variation here that has not been eliminated completely. Look closely and you will see the field goal effect is still there - intermingled with touchdown possibilities.

We can read the value of a yard too. You can see that between the red zones (i.e. not within 20 yards of the end zone) it takes about 60 yards to go from four points to zero. This translates to 15 yards per point. The takeaway: from the value of possession in points and the value of a yard in points we reason that a possession is *unlikely to be worth substantially more than 30 yards.*[4]

4. Third and one. Armed with that basic observation, we return to our key strategy question and reconsider "opting out" of a completion of the first down on the first attempt. This choice, which I have asserted is the superior one, might lead to a first down completion on the next play. Of course, we might also find ourselves in a third and one - a situation we must analyze too.

Strictly speaking these aren't the only possibilities - and the reader will note

[4]A small flaw in the use of "expected points" is the value of possession conceded after a score is made - though as can be inferred on the plot this is small near the region where the defence are likely to restart play. Thus, the notion of expected next score is quite similar to a value function approach, as is common in control theory. The discrepancy does not warrant an extended discussion here since we require only a rough estimate of possession value.

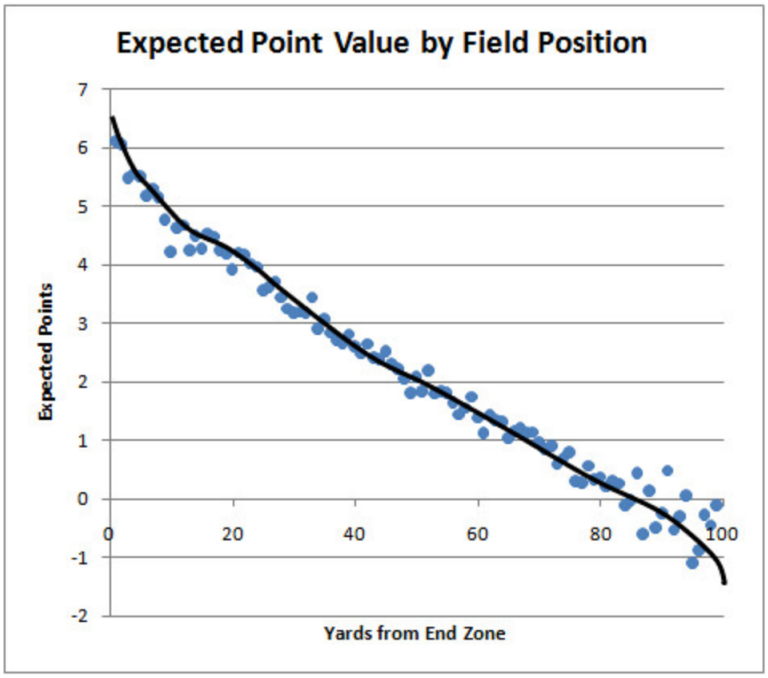

FIG. 3. *Expected value by field position, from an article by Brian Burke [2]. This is indicative of the value of a yard in points on the first down.*

there is a small chance of a loss on the second down play, or a catastrophic turnover. However one can, for the sake of argument, assume a running play, where the possibility of a loss is greatly diminished. Whether running or passing is chosen, the under-appreciated upside is the number of yards gained on a successful second and one play, or if needed, a successful third and one play.

For avoidance of any doubt, I refer to the conditional averages, not the average including unsuccessful plays that don't advance the line of scrimmage. The data reveals that when rushing, this conditional average gain for a third down play is about five yards. When passing, it is a shockingly large 12.75 yards.

Another aside: this is a large discrepancy between passing and running and it suggests that teams might consider passing on third and one more often than they do. Passing plays are only successful 61% of the time, versus 72% when rushing. However, an 11% chance of lost possession corresponds to only a few yards, according to our analysis above. And this gets swamped by the massive 7.75 yard differential when passing. Side conclusion: teams should use passing plays on third and one more often!

Third and one strategy is certainly interesting, but the only thing we need to take from this analysis is that a successful third and one results in an advance of field position of almost five yards, even if you choose to run the ball. Perhaps that's worth rephrasing for those who fear the third down. *Third and one is probably first and six, on average.*

5. Second and one. With that in mind, let's roll back one play. How should we feel about finding ourselves in second and one? When we look at second and one

Extra yards on first down	Implied value of a possession	Assessment
0	64 yards	Incorrect
1	48 yards	Incorrect
2	30 yards	Break-even
3	13 yards	Correct

TABLE 1

The value of a possession, measured in yards, implied by a player's "decision" to achieve the first down on the first down. This assumes that a player might instead stop one yard shy of the first down marker, leaving his team at second and one. It also assumes a "typical" game and (middle) field position, and that a punt will be taken on fourth down. The implied yards per possession is conservative for several reasons. It assumes running plays will be chosen on second and third downs - though our discussion suggests this is sub-optimal. It doesn't take into account gains of less than a yard on second down. It also ignores the possible option value of running or passing on the fourth down, and it applies a conservative estimate of a gain of 4.75 *yards conditional on success (on either down). This represents a conditional mean gain of* 3.75 *yards - averaged over* 17 *of* 18 *occasions when possession is maintained.*

plays, rather than third and one, we find that the rushing play percentage goes even higher - up to 80% - though the average number of yards drops slightly (down to 4.73). That isn't surprising, since the defensive team has less incentive to prevent the first down completion.

Another aside: the differential between passing and rushing yards gained decreases, as compared with the same differential for third and one, with passing leading to only 6.5 yards of gain on average compared with 7.75 for passing on third down. The natural urge to pass on second and run on third might be working against the better interest of teams.

That's interesting - but not crucial to the case I make for stopping shy of the first down. What's important, and now evident by multiplication (assuming conditional independence of third down play outcomes) is that if you tell your team to run on second and one, and then again from third and one (should that be necessary) then you will have a 94.5% chance of getting the first down. In the process you will advance an average of 4.75 valuable yards.

Perhaps it is apparent why you don't want those chains moved on the first down.

6. Implied value of possession. Our star receiver takes a catch a yard short of the first down. Flat-footed, he turns to see a defensive player bearing down at great speed. Risking a season-ending injury he can, most certainly, dive forward with outstretched hands and make the hero play - securing the first down. Alternatively, he can casually step out of bounds, leaving his team at second and one.

Or perhaps a tight end has broken one tackle and staggers toward the first down marker dragging a defender who has grasped his leg. Should he break the plane or voluntarily stop his progress, if he knows he won't get any extra yards beyond the first down?

For a potentially controversial topic, the calculus is alarmingly easy and I frame it in terms of implied yards per possession. Using the values above, the wide receiver's lunge suggests that a 5.5% chance of losing possession in this series of downs is more important than 3.75 yards of field position. He is wrong! Since 5.5% is roughly 1 in 18, this means that possession must be 17 times more important than 3.75 yards. The receiver implies a value of possession of 17*3.75=63.75 yards! But there is no way on God's green football field that a possession is worth over sixty yards. It is closer to half that number, as we have seen.

It becomes difficult to justify achieving the first down even if some gain is made. Some values are tabulated in Table 1. If possession is valued at 30 yards, it is clear that players should "decline" the first down (voluntarily stop progress of the ball) quite often. There may be additional reasons to do so, such as reduced risk of injury. Going for the first down isn't even worth it if you get to the 11-yard line. The calculus would then read $2.75 * 17 = 46.75$, which is still way too high a value (in yards) to put on possession.

Thus, contrary to commonly accepted wisdom, a celebration is only warranted if a player can advance three yards past the first down marker. A marginal first down completion could be setting the team up for failure.

Conversely, defensive players need not suffer one extra concussion to bring about an abrupt deceleration of the ball carrier. Let them get the ten yards, or eleven if necessary. And the implication extends beyond individual player decisions. Offensive and defensive teams should design plays to make the +13 and +9 yard gains more likely, and the middle ground less so. Are they?

7. Evidence of poor strategy. Now that we know what is optimal, let's take a look at what happens on the field. Figure 4 show a histogram of first down yards gained rushing, where for simplicity we are restricting attention to cases where it is first and ten. One would think that with a rushing play, the offensive ball carrier would have good ability to aim for nine yards or 12, but not accidentally end up in between.

The data, which shows *some* mass moved from 10-yard gains back to nine yard gains, seems to indicate one of two things:

1. Some offensive players are aware that second and one is better than first and ten, and they are deliberately acting so as to achieve more nine yard gains than they otherwise would (though not to the extent they should).
2. Defensive players are trying too hard to prevent the first down, mistakenly believing that a stop at nine yards is better than conceding ten.

I'm not sure how we can disentangle these two effects but the latter is more culturally plausible. However, one thing is clear, and that is that there is plenty of room for improvement in strategy by both offense and defense. Defensive teams could very easily allow players to make a 10-yard advance, and they are clearly not doing that. Conversely, offensive teams really should have very little probability on the 10-yard advances.

From the perspective of the offensive team, the fact that 12-yard advances are less common than 11-yard advances is also a clear sign of poor strategy. That mass on the 11-yard gain should be moved back to 9 yards. Injury risking heroics used to get to 10 yards should, in fact, be reserved for going from 11 to 12, or 12 to 13, when the opportunity arises.

One could also consider passing plays. The yards gained increases - especially for third downs as noted - and the success probability falls. But after accounting for these offsetting effects the data is so similar that I'll save one plot.[5]

The point is that whether we are talking passing or rushing, both offensive and defensive teams clearly need to improve their strategy. Defensive teams can no doubt benefit from coverage patterns that deliberately allow the first down if a player is likely to make nine yards. These changes, both to plays and execution, are likely to bring other benefits, such as reducing the chance of being caught flat footed by a long

[5] If it were not similar, we would have unearthed a completely different type of suboptimal decision making.

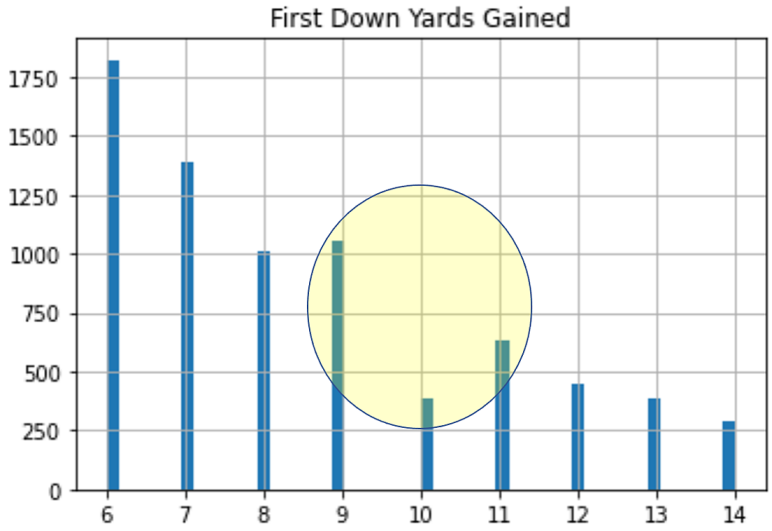

FIG. 4. *The empirical distribution of yards gained on the first down. The preponderance of nine yard gains relative to ten yard gains suggests that both offensive and defensive teams are making a strategy error. Defenses should be allowing ten and eleven yard gains. Offenses should be aiming for nine yard gains, unless they can get to thirteen.*

pass, or completely missing a tackle when trying to stop a ball carrier at nine yards rather than ten.

A more scathing indictment of strategy, both offensive and defensive, is delivered by Figure 5 which shows first down yards gained on the rarer occasions when, due to a defensive penalty, we are at first and five yards to go. There ought to be more ability to control yardage gained, and design a play to achieve four yards with high probability. Instead, teams are mistakenly attempting to get to five yards.

What are they thinking?! My data suggest that teams' chances of making four yards when they want it is significantly higher than their chance of making five yards (when they think they want it) - about 10% higher. For this, we have to use second down data due to a dearth of first and four situations. But it is clear that teams are running riskier plays to achieve a worse outcome.

8. Conclusion. The lunge for the first down is part of football culture, and it seems almost distasteful to point out that this act fails to assist a team's chances of winning - at least on the first down. It is unfortunate that under optimal strategy, the great game of football is not as clean as fans might like. Team advantage is not mononotic in yards gained on the first down.

I have explained the calculation as a trade-off between yards and possession, with the likely valuable yardage gain outweighing a the small risk of losing possession. I've couched this by noting that a player who stretches out his arm to break the plane of the first down marker is only justified in doing so if a possession is believed to be worth fifty or sixty yards. This strains credulity, given that the value of possession appears to be roughly half that number, however computed.

But there are other ways to convince skeptical players. Indeed, stopping at nine yards can be seen as a way of maintaining possession - whereas completion of the first

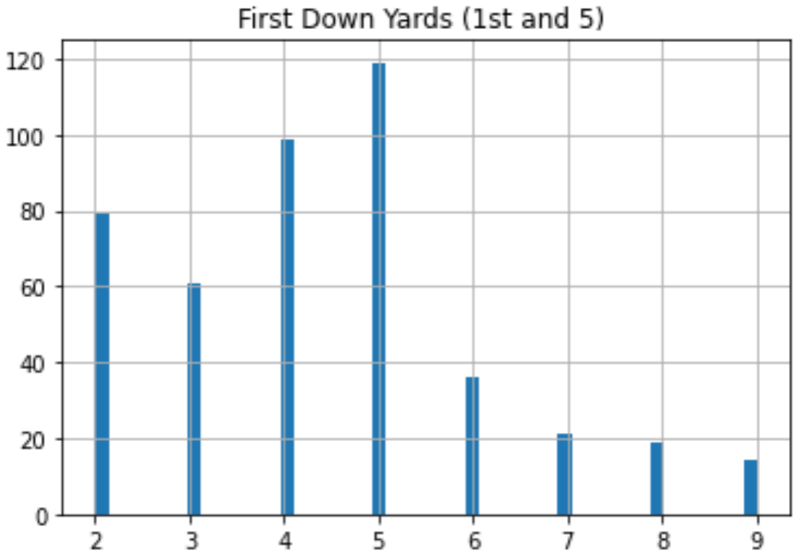

FIG. 5. *Yards gained on first down when there is five yards required. This scenario results from a penalty applied to the defense. As with first and ten, the implementation of optimal strategy by either offensive or defensive coordinators should preclude what we see - namely the abnormally large number of five yard gains. Compared to Figure 4, this provides even more dramatic evidence that teams do not appreciate the relative value of yards and possession.*

down sets up the stiffer challenge: the necessity to make ten yards on three plays. So, the calculation I present should not be viewed as an opinion on possession. Even coaches and players who believe in "possession at all cost" should come around to this way of thinking.

Stopping shy is not defeatist as it will almost certainly lead to a first down completion further down the field. Offensive coordinators should ask their running backs if they would prefer five downs to make eleven yards (with 95% probability), or three to make ten.

I hope that this provokes a more rigorous treatment of offensive and defensive strategy, possibly starting with the reproduction of these findings [4]. Playbooks need to be rewritten, with a view to maximizing the nine yard gain. Though I have not tried to cover all situations, this should not present an excuse. Each case deserved careful consideration. For example when touchdowns have a high probability, remaining downs are more valuable - but so is the marginal value of a yard.

There are other directions for research. Though the data set used here did not facilitate it, a future analysis might consider more granular yardages, and even the possibility of stopping shy of the first down marker on the second down, not just the first.

REFERENCES

[1] D. BEN, *NFL Markov*, 2016, https://raw.githubusercontent.com/microprediction/nflMarkov/master/inputData/pbp_nfldb_2009_2013.csv.
[2] B. BURKE, *Expected Points*, 2008, http://archive.advancedfootballanalytics.com/2008/08/expected-points.html.
[3] V. CARTER AND R. E. MACHNOL, *Operations research on football*, tech. report, Northwestern

University, 1970.
[4] P. Cotton, *NFL 2nd and 1 notebook*, 2021, https://github.com/microprediction/microblog/blob/main/NFL_2nd_and_1.ipynb.
[5] B. Morris, *Kickers are Forever*, 2015, https://fivethirtyeight.com/features/kickers-are-forever/.