Analytics, have some humility: a statistical view of fourth down decision making

Ryan Brill and Abraham Wyner

University of Pennsylvania

NESSIS, September 2023

In-game strategic decision making

- The mathematical basis of in-game strategic decision making is a *value function* $V(x)$ which tells us the value of game-state x
- In American football:
    1. Expected points (EP)
    2. Win probability (WP)
- Make the decision which maximizes the value of the next game-state
    - Make the fourth down decision in $\{\text{Go}, \text{FG}, \text{Punt}\}$ which maximizes win probability

Model Land

- Two ways to build these EP/WP models:
    1. Probabilistic state-space models
        - require detailed specification of all game-states, actions, and their transition probabilities; incredibly hard
        - “Gei gazinta hait” (Yiddish for “go in good health”)- explore these models on your own time, but will not be the subject of today’s talk
    2. Statistical models
        - Data-driven regression/ML models fit from historical data
        - Widely used today; we will focus on these models
        - We discovered several problems with the way they are implemented, fit, and applied

Expected points models

Well Known Expected Points Models

Modeler	Model	Game-state Variables	Training set	Outcome Variable
Romer (2006)	Instrumental variables regression	yardline	all plays	*Points* of the next score, a real number in {7,3,2,0,-2,-3,-7}
Burke (2009)	Linear regression with a spline	yardline	first down plays	^
Yurko et al. (2018)	Multinomial logistic regression	yardline, down, log yards-to-go, time remaining, goal-to-go, under-two-minutes	all plays	*Outcome* of the next score as categorical variable in {TD, FG, …} $\text{EP} = 7 \cdot \mathbb{P}(\text{TD}) + 3 \cdot \mathbb{P}(\text{FG}) + \cdots$
Baldwin (2021)	XGBoost	yardline, down, yards-to-go, time remaining, timeouts, home, roof type, era	all plays	^

EP models don’t adjust for team quality

- Existing EP models are functions of yard-line, down, yards-to-go, time remaining, timeouts, etc.
- But these models don’t adjust for team quality.
- Justification for omitting team quality:
    1. Models represent EP for an *average* offense facing an *average* defense, and so imply decision making for *average* teams.
    2. It is not easy to adjust for team quality alongside all these other variables, which have nonlinear relationships and interactions.

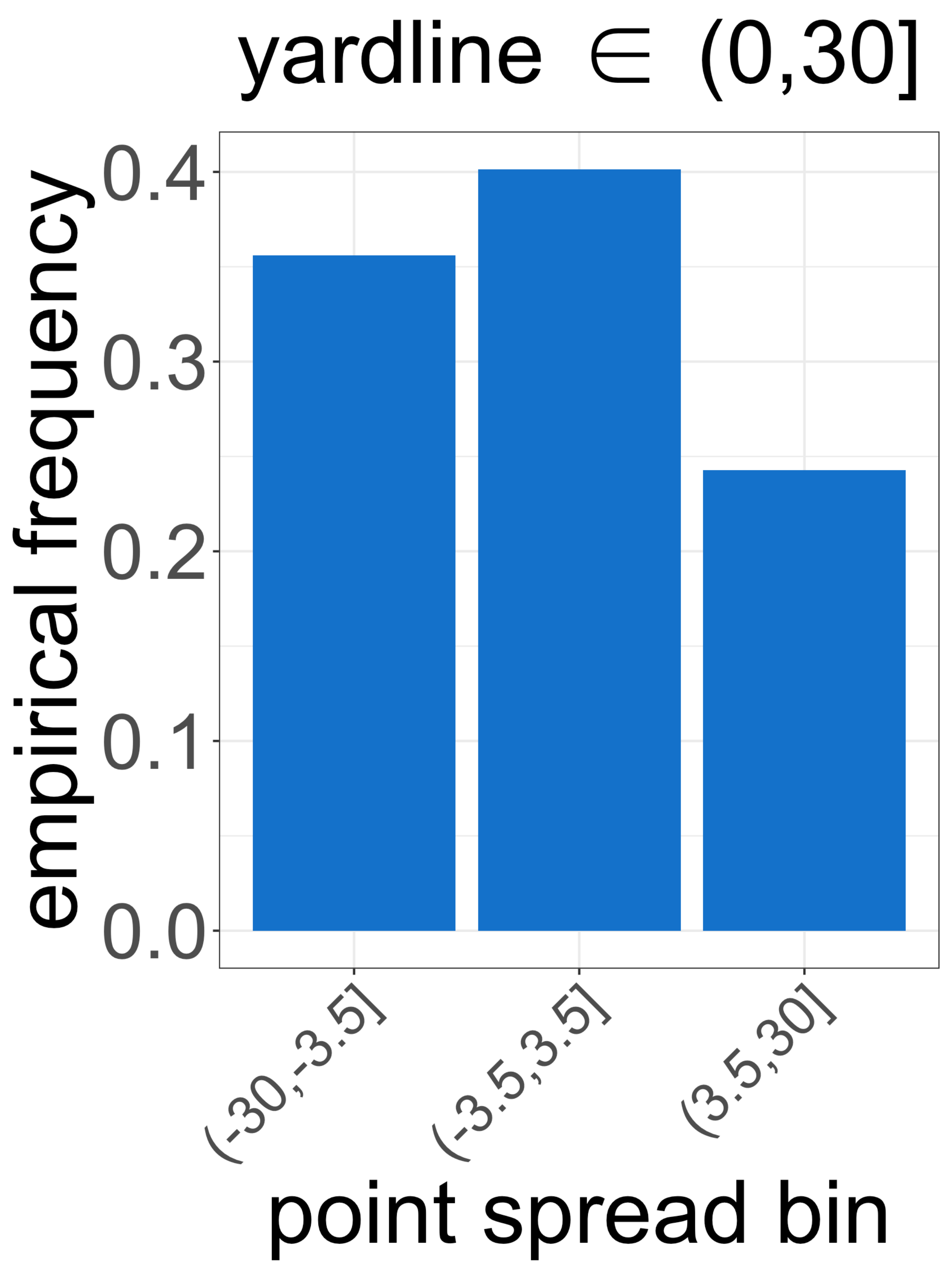
Thought experiment

1. What is the probability that an *“average” NFL kicker* sinks a 70 yard field goal in neutral weather conditions?

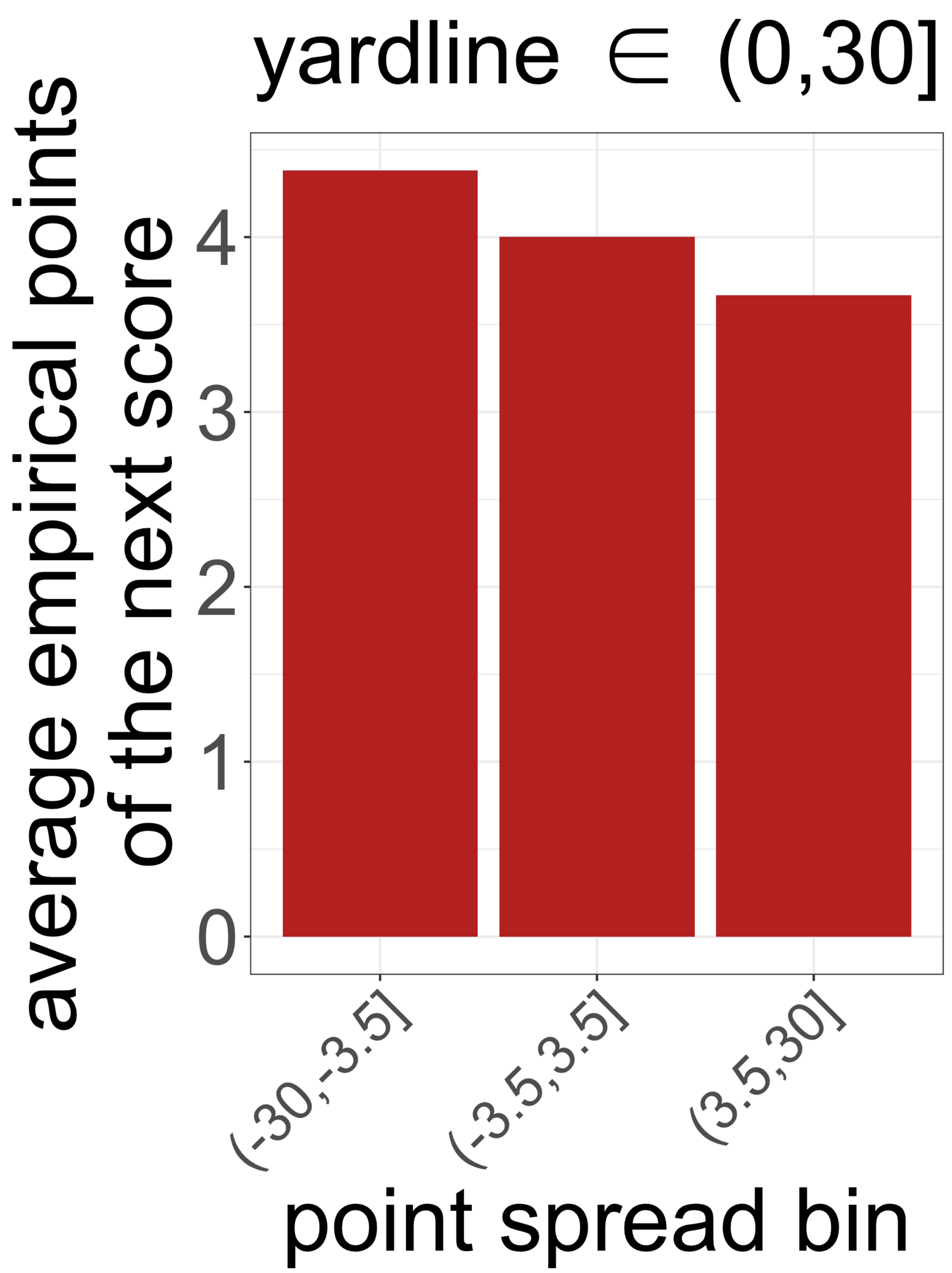
2. What is the probability that *Justin Tucker* sinks a 70 yard field goal in neutral weather conditions?

3. What is the probability that a *randomly drawn kicker* sinks a 70 yard field goal in neutral weather conditions?

Problem 1. EP models don’t adjust for team quality



Good teams have more plays



Good teams score more points

Failing to adjust for team quality causes problems:

1. Models report EP for *randomly drawn teams*, not for average teams.
   - No team wants this!
   - No such thing as a decision made by a random team!
2. **selection bias** problem; EP models are especially wrong/biased

Base-rate across all yardlines: (32%, 40%, 27%)

Base-rate across all yardlines: (2.5, 1.7, 0.8)

We need to adjust for team quality

EP models that don’t adjust for team quality are biased.

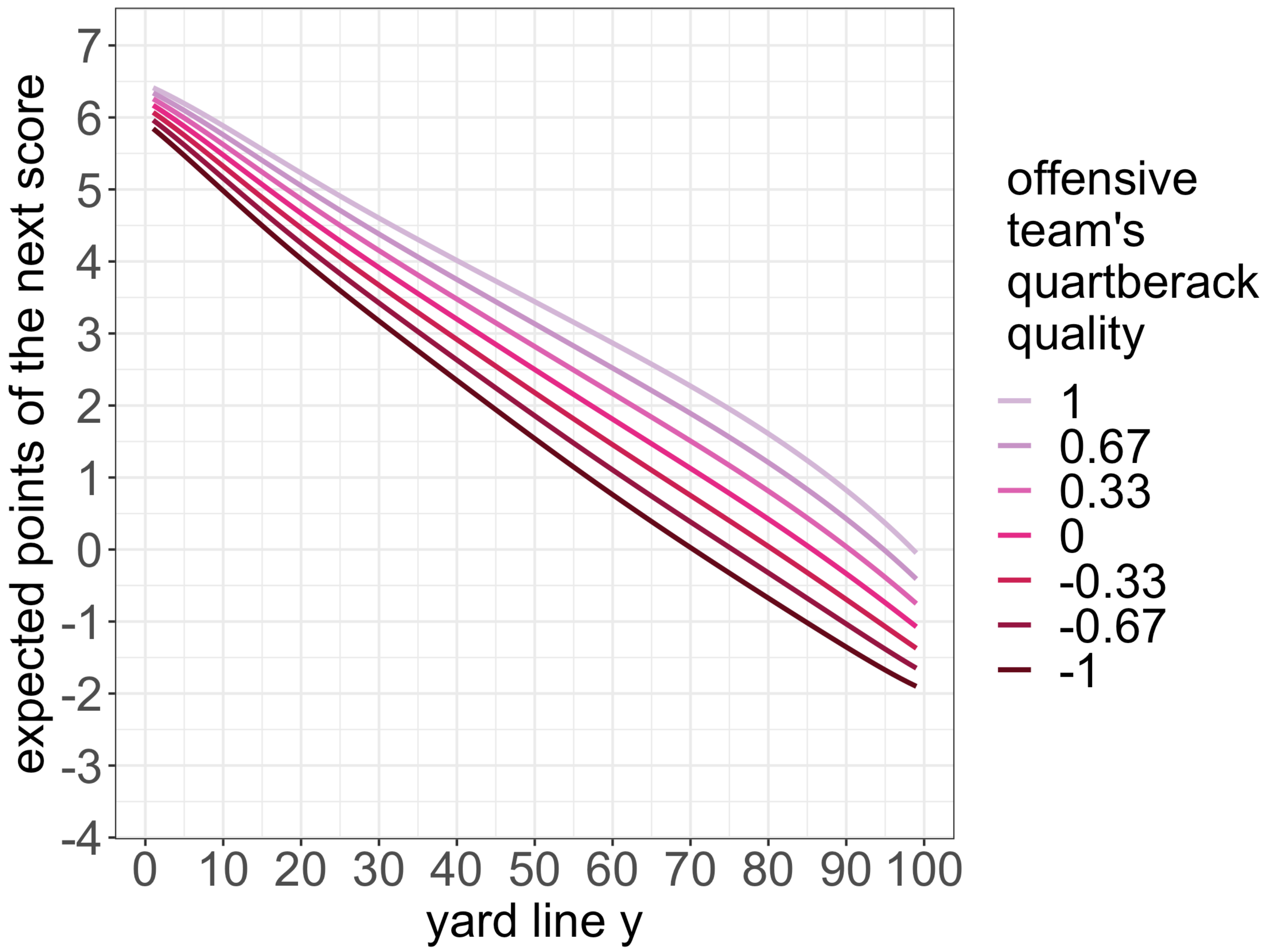
Thought experiment 2

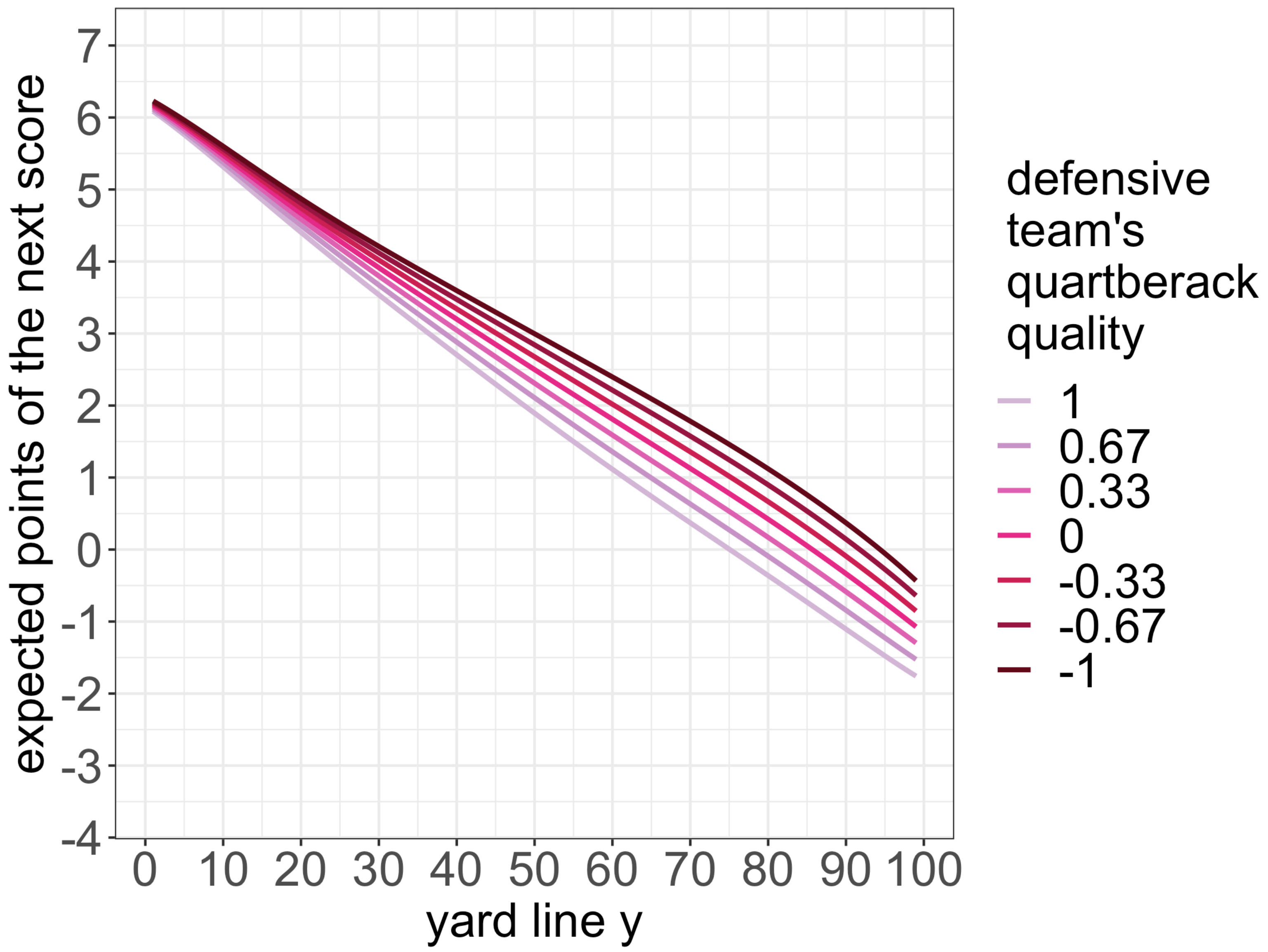
Suppose I have the following 8 aspects of team quality, each on the same scale, built from play success, and without data bleed. Then build an EP model with these 8 metrics as covariates.

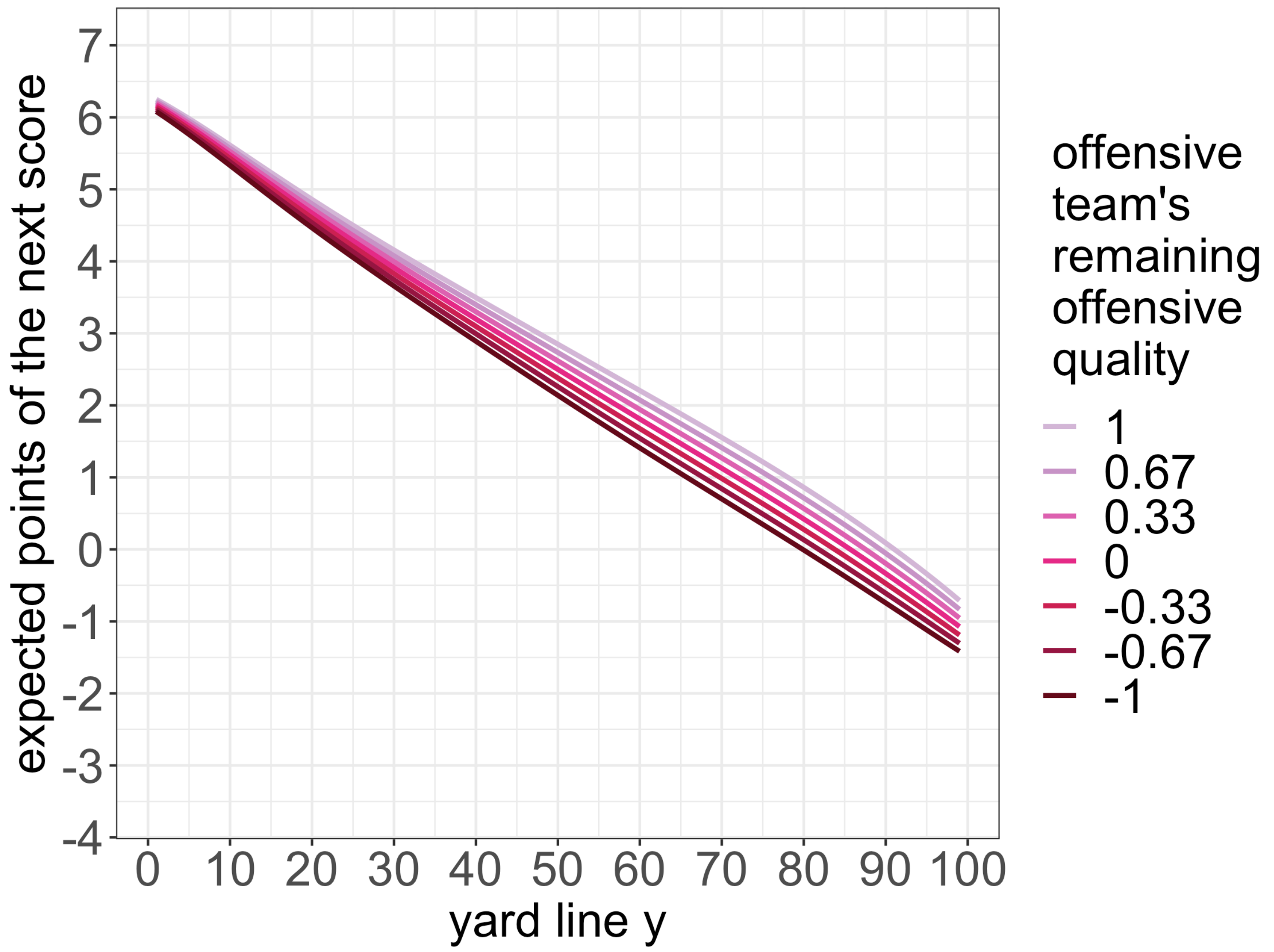
- Rank in terms of (*predictive*) importance:
    A. Offensive team's quarterback quality
    B. Offensive team's non-quarterback offensive quality
    C. Defensive team's defensive quality against the pass
    D. Defensive team's defensive quality against the run
    E. Offensive team's defensive quality against the pass
    F. Offensive team's defensive quality against the run
    G. Defensive team's quarterback quality
    H. Defensive team's non-quarterback offensive quality

Impact of various aspects of team quality

- Created our own 8 measures of offensive & defensive quality
  - Carefully controlled for data bleed
  - All 4 offensive quality metrics are more impactful than the defensive quality metrics
  - Quarterback quality of *both* teams matters more than other aspects of team quality!
- Output from an additive multinomial logistic regression model (similar to Yurko et al.'s):







Problem 2. So many variables…

- We need to adjust for team quality
- *Wow*, that's a lot of variables – team quality, yardline, down, yards-to-go, time remaining, etc. – with nonlinearities & interactions
- The task is not easy: we need to fit a very big and very complicated machine learning model, but we don't want to overfit

Problem 2. Bias-Variance Tradeoff

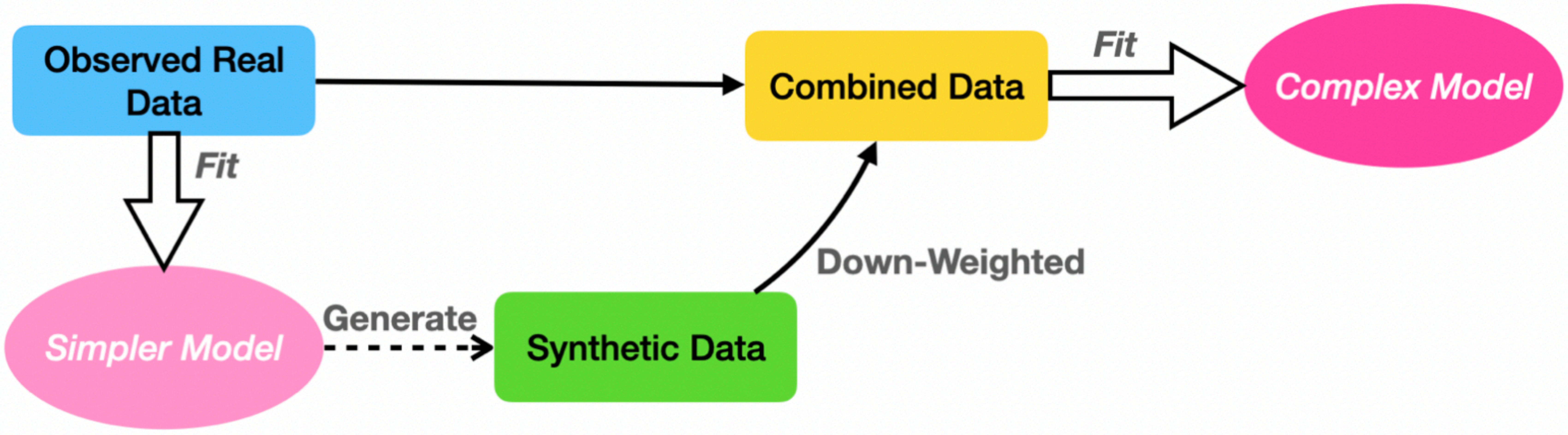
- We want to use machine learning to capture a complex high-dimensional function
- But ML models tend to aggressively overfit (play-by-play) data
- Typically deal with this using regularization or shrinkage towards simpler models
- Easily studied for parametric models in a Bayesian context; difficult for ML/trees

"The in-game models are not Bayesian. Congratulations to you if you can figure out how to do that. Most publicly available models are … XGBoost models. "

— Brian Burke, Wharton Moneyball, 19 Sept. 2023

Solution 2. Catalytic Priors to Mitigate Overfitting

- Inspired by Sam Kou's catalytic prior, we found a way to Laplace smooth tree ML models



EP Model Comparison

Model name	Model type	Team quality	Out-of-sample MAE
Catalytic	**Catalytic XGBoost**	Yes	**3.744**
Yurko+	Multinomial logistic regression	Yes	3.749
Baldwin+	XGBoost classification	Yes	3.753
Baldwin (2021)	XGBoost classification	No	3.803
Yurko (2018)	Multinomial logistic regression	No	3.808
Burke (2009)	Linear regression	No	3.833
Romer (2006)	Instrumental variables regression	No	3.864

- EP models are biased and overfit, and we can improve upon that
- On subsets of plays our model is *much* better, and that can make a huge difference in decision making

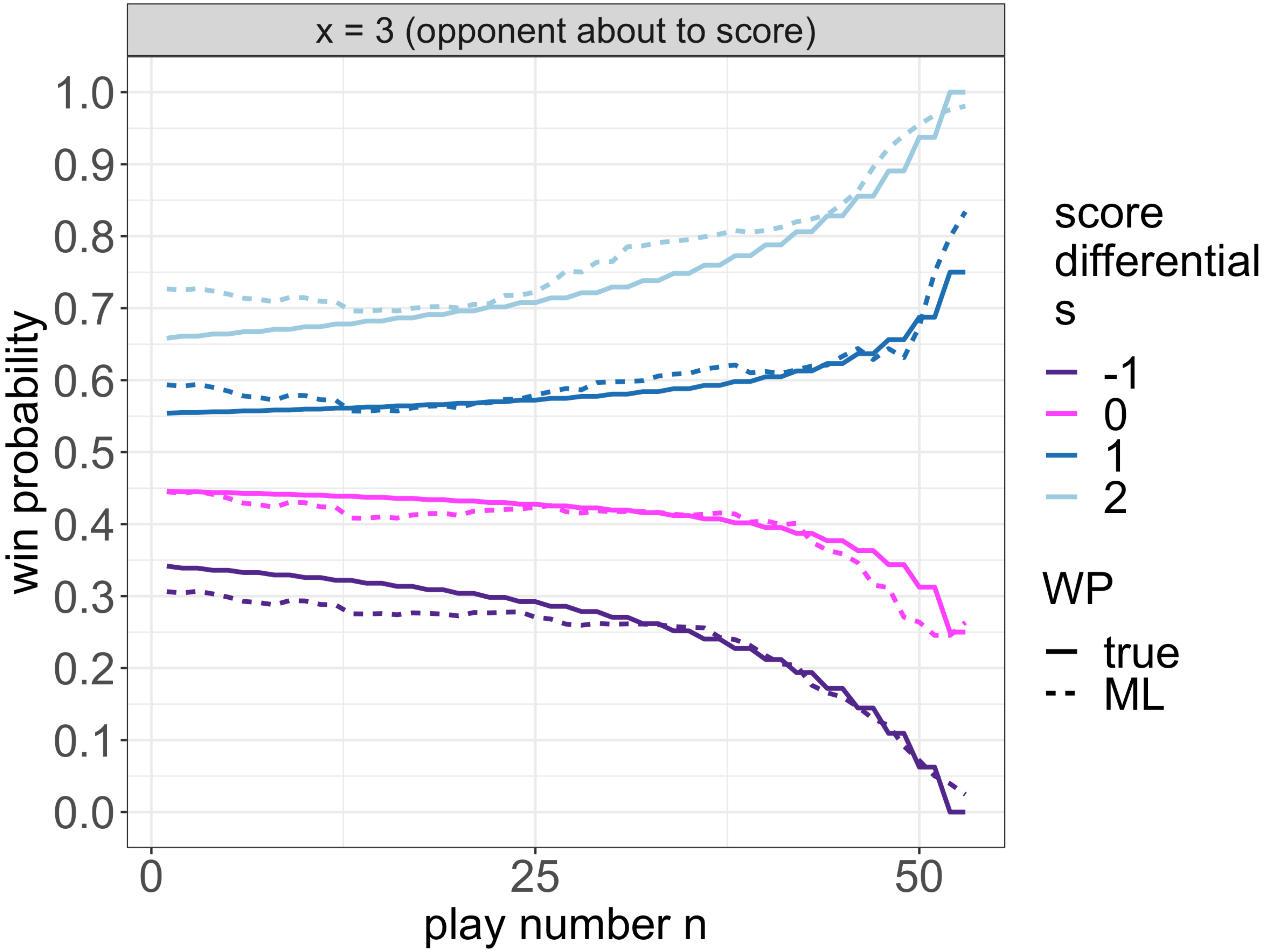
Win probability models

Problem 3. Highly Auto-Correlated Data

- Play-by-play dataset of ≈ 500,000 plays
- But, *not* 500,000 independent outcome variables
  - The response variable: 1 if the team with possession wins the game, else 0
- *Every game has only 1 winner* (auto-correlated data)
- Effective sample size is closer to 4,000 (num. games in the last 15 years)
- This is nowhere near enough data to experience the full variability of the nonlinear and interacting variables of score diff., time remaining, team quality, yardline, down, distance, timeouts, etc.

WP Simulation

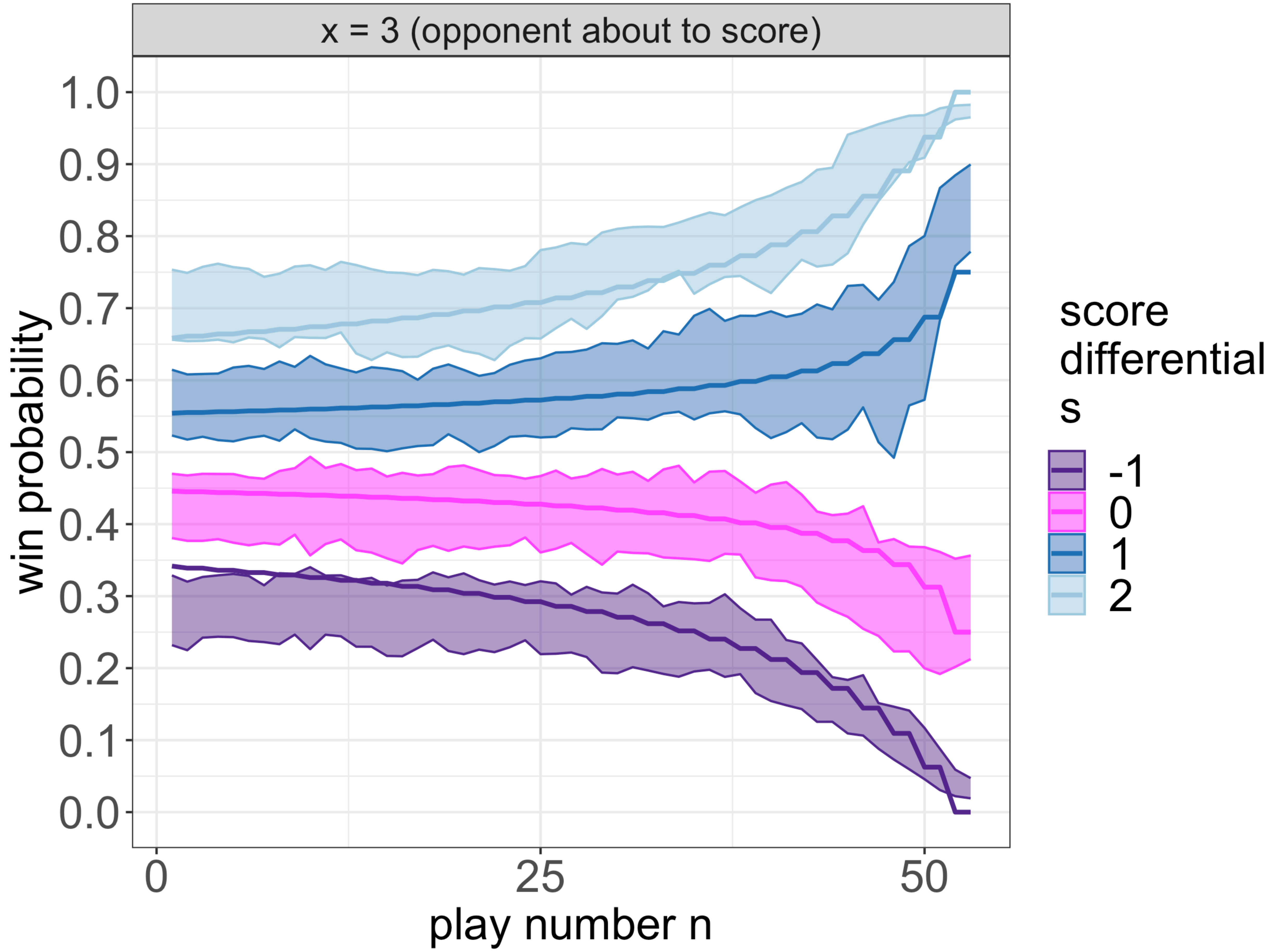
- To show you how hard it is to accurately fit WP using just ≈ 4000 games, we created a Random Walk version of football
- It's an extremely simple Random Walk but looks just like football!
- We can precisely calculate WP at every game-state
- Then, simulate a historical play-by-play dataset with auto-correlated win/loss response variable



- WP point estimates, fit using machine learning from one simulated dataset of simplified football plays, *get the general trend right* (are unbiased).

WP Simulation

- To show you how hard it is to accurately fit WP using just ≈ 4000 games, we created a Random Walk version of football
- It's an extremely simple Random Walk but looks just like football!
- We can precisely calculate WP at every game-state
- Then, simulate a historical play-by-play dataset with auto-correlated win/loss response variable



- Bootstrapped WP confidence intervals, to achieve 90% coverage of true WP, need to be wide (8% WP on average).
- Real football exponentially more complex. Confidence intervals should be far wider.

Quantifying uncertainty of the optimal fourth down decision

- Making fourth down decisions based solely on WP point estimates, which are highly uncertain, leads to overconfident decisions
- Quantify uncertainty in the 4th down decision by bootstrapping
  - the randomized cluster bootstrap accounts for autocorrelation
  - ***boot %*** — % of bootstrapped models which choose decision $d \in \{\text{Go, FG, Punt}\}$

decision	WP	WP gain CI	boot %
Go for it	73.7%	[-3.7%, 4.5%]	53.8%
Field goal	72.2%		46.2%
Punt	65.5%		0.0%

Example Plays:
How Fourth Down Decision Making Changes

Example 1

- CHI @ NYJ in Week 12 of 2022

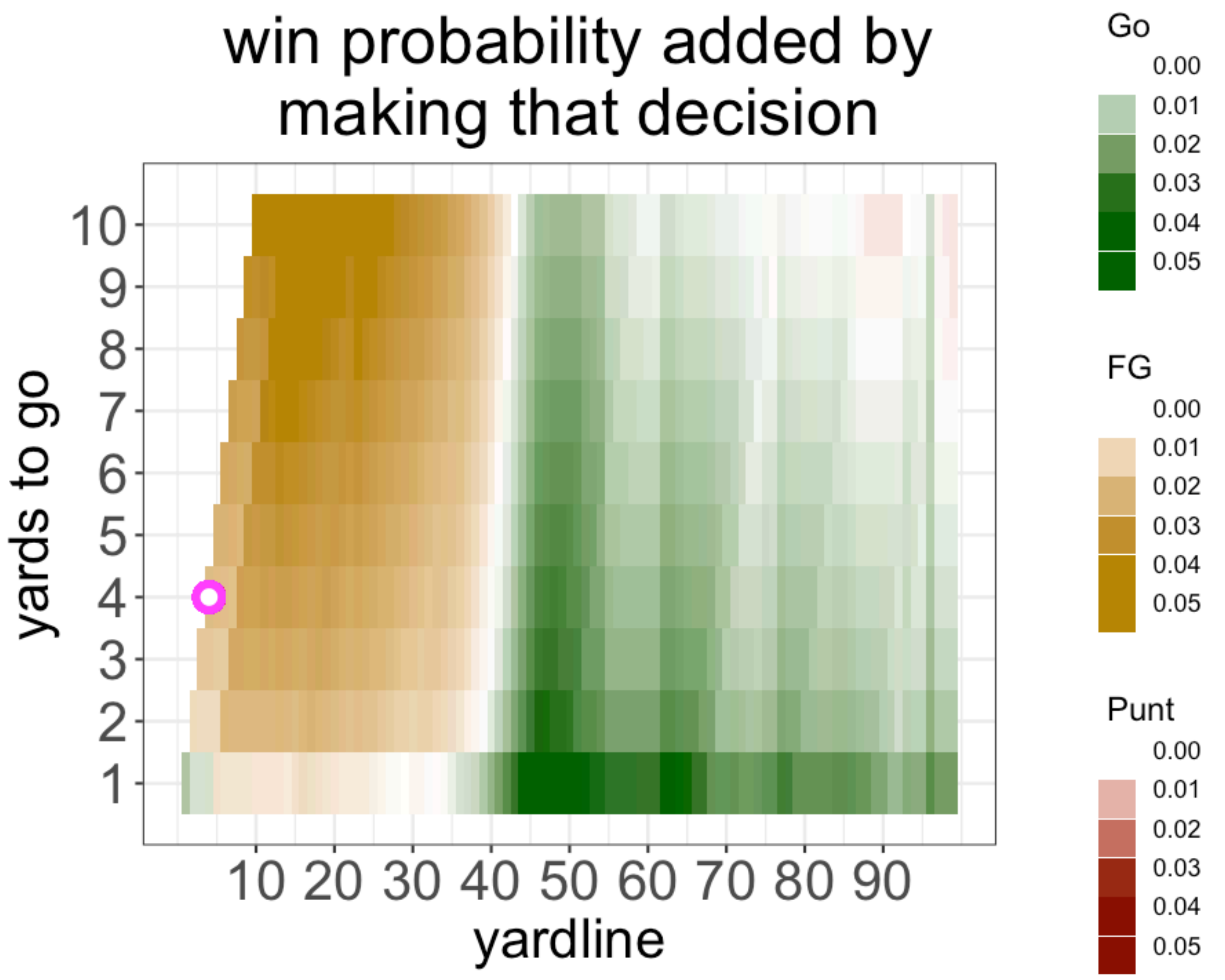
FG looks like a strong decision based on the WP point estimate (+2%).

Traditional analytics recommendation: **Field goal attempt.**

Down 7, 4th & 4, 4 yards from opponent endzone

Qtr 1, 6:00 | Timeouts: Off 3, Def 3 | Point Spread: 8.5

decision	WP	success prob	WP if fail	WP if succeed	SD of WP	baseline coach %
Field goal	18.3%	98.6%	11.0%	18.4%	0.9%	85.1%
Go for it	16.2%	44.5%	11.0%	22.8%	5.9%	14.8%
Punt	9.9%					0.0%



Example 1

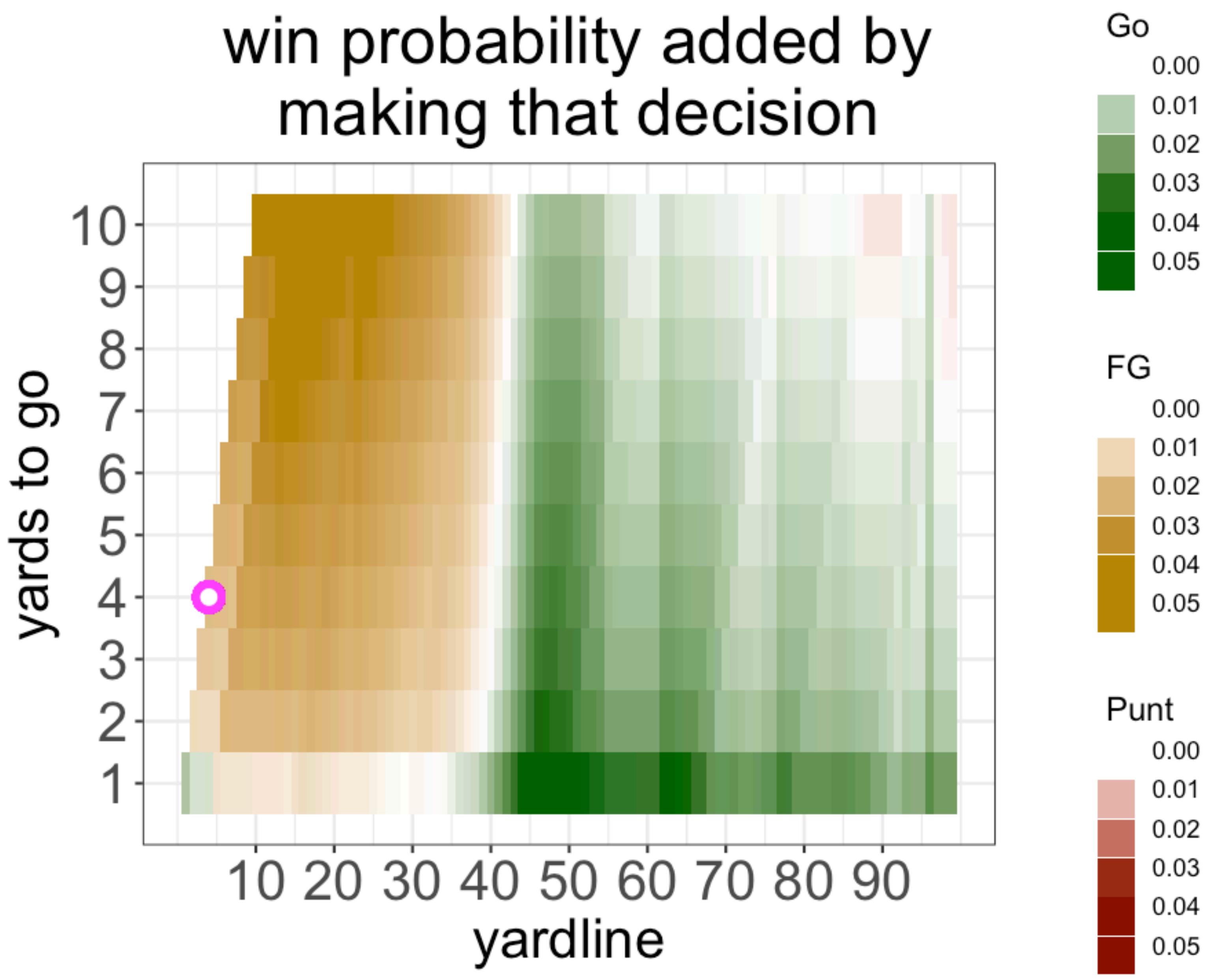
- CHI @ NYJ in Week 12 of 2022

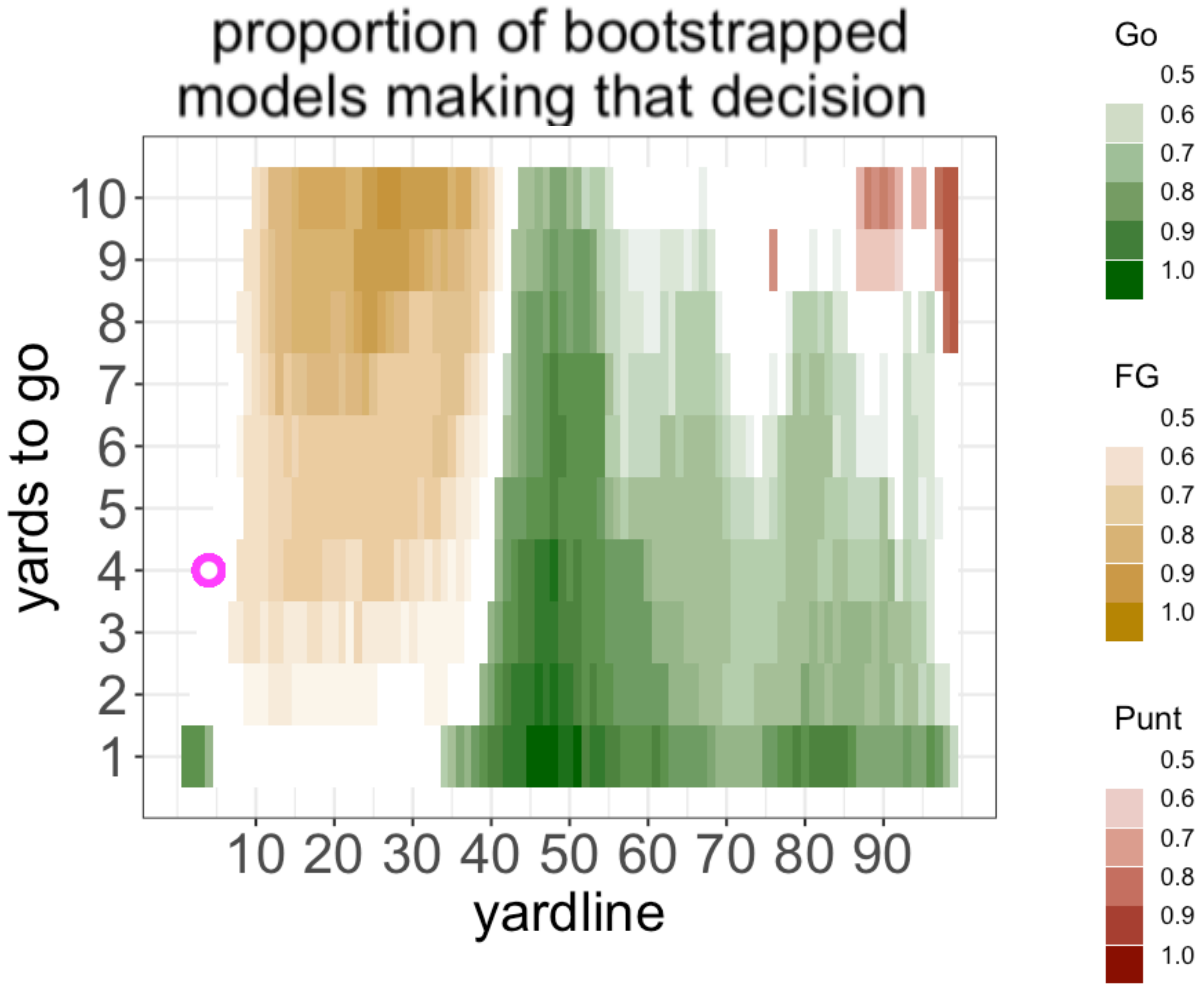
FG looks like a strong decision based on the WP point estimate, but we don't have enough data to trust our own point estimate.
Our recommendation: **Coach's discretion.**

- For *many* plays the optimal decision is uncertain!

Down 7, 4th & 4, 4 yards from opponent endzone
Qtr 1, 6:00 | Timeouts: Off 3, Def 3 | Point Spread: 8.5

decision	WP	WP gain CI	boot %	success prob	WP if fail	WP if succeed	SD of WP	baseline coach %
Field goal	18.3%	[-3.7%, 4.4%]	42.3%	98.6%	11.0%	18.4%	0.9%	85.1%
Go for it	16.2%		57.7%	44.5%	11.0%	22.8%	5.9%	14.8%
Punt	9.9%		0.0%					0.0%





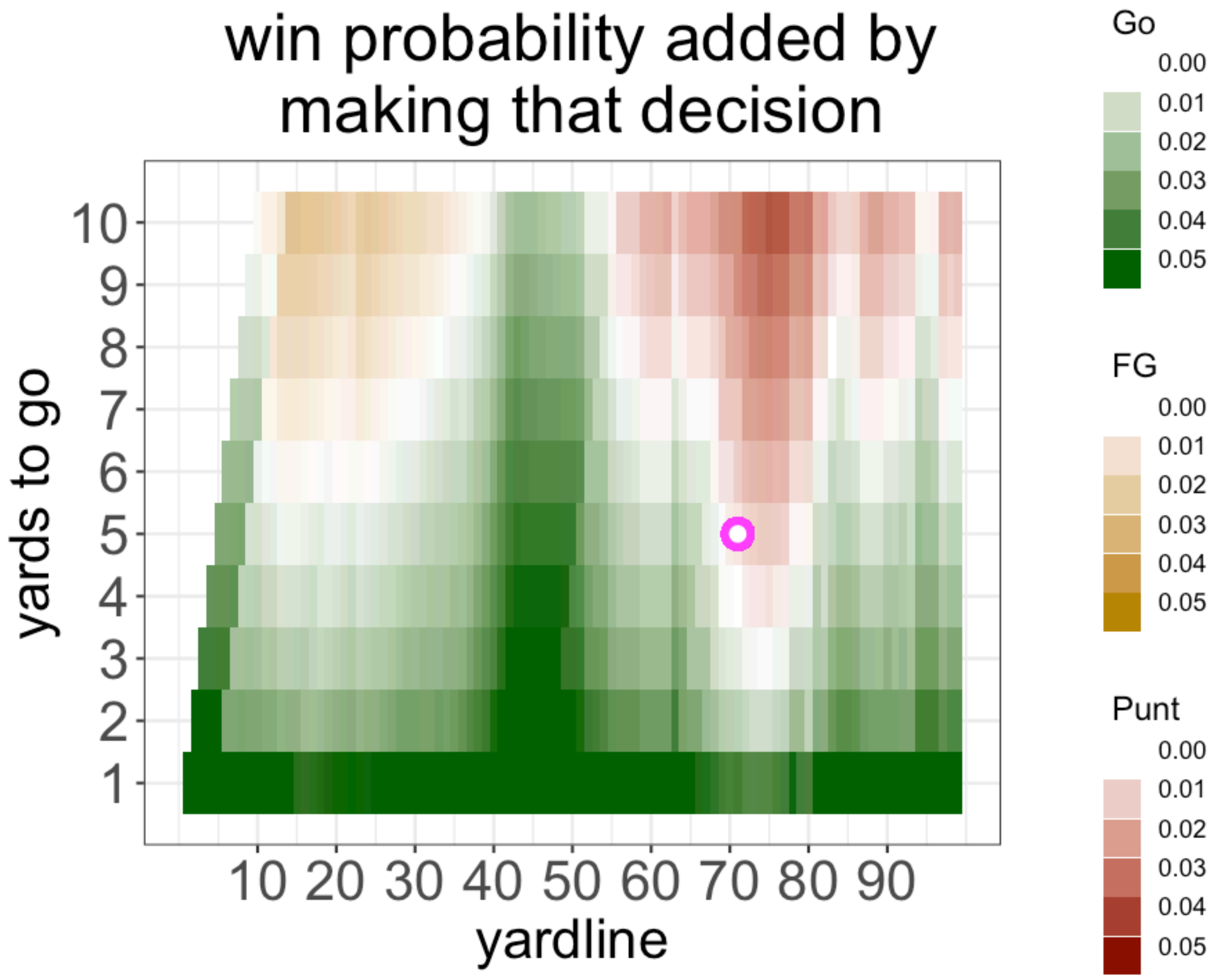
Example 2

- WAS @ IND in Week 8 of 2022

Punt has a tiny estimated edge over Go (+0.05%).
Traditional analytics recommendation: **Tossup, or slight lean towards punt.**

Up 1, 4th & 5, 71 yards from opponent endzone
Qtr 3, 6:00 | Timeouts: Off 3, Def 3 | Point Spread: 3

decision	WP	success prob	WP if fail	WP if succeed	SD of WP	baseline coach %
Punt	44.6%					93.3%
Go for it	44.1%	44.4%	34.4%	56.3%	10.8%	6.7%
Field goal	34.4%	0.0%	34.4%	56.0%	0.0%	0.0%



Example 2

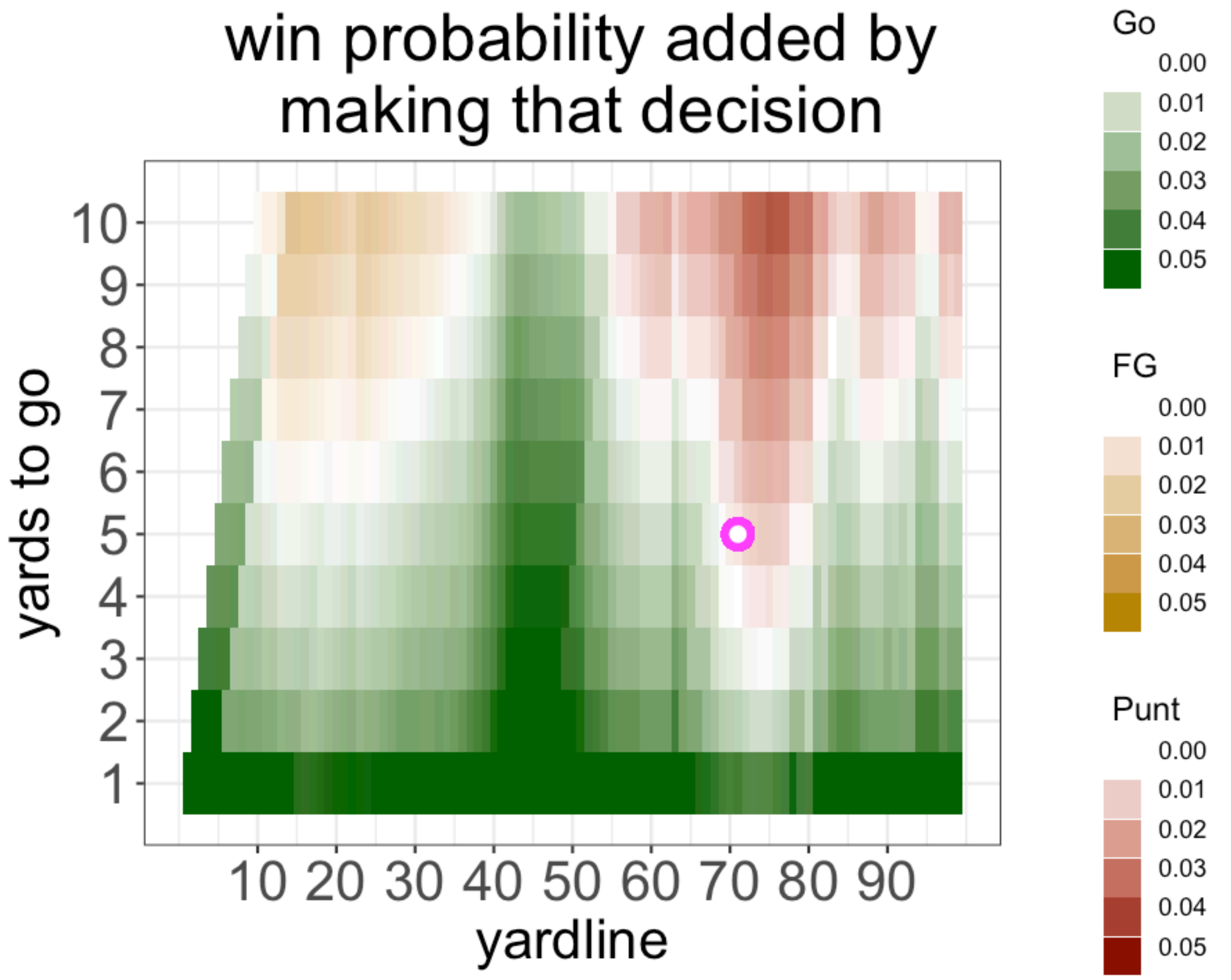
- WAS @ IND in Week 8 of 2022

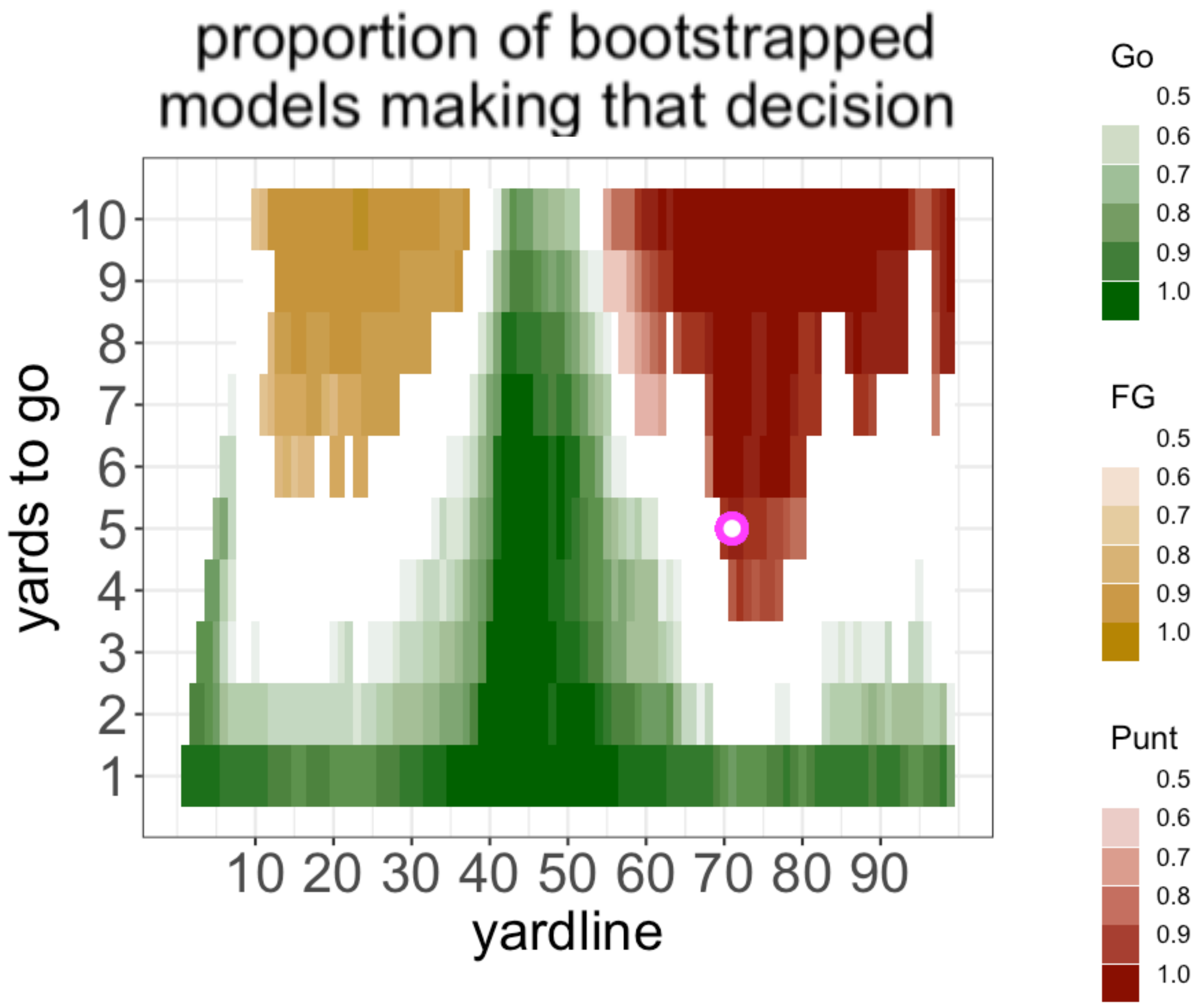
Punt has a tiny estimated edge over Go,
but we are confident that the edge is there.
Our recommendation: **Punt (but not a tragedy if the coach overrides).**

- Eeking out these tiny (but confident) edges are valuable because many more of them occur per game.

Up 1, 4th & 5, 71 yards from opponent endzone
Qtr 3, 6:00 | Timeouts: Off 3, Def 3 | Point Spread: 3

decision	WP	WP gain CI	boot %	success prob	WP if fail	WP if succeed	SD of WP	baseline coach %
Punt	44.6%	[0.0%, 4.8%]	96.2%					93.3%
Go for it	44.1%		3.8%	44.5%	34.4%	56.3%	10.9%	6.7%
Field goal	34.4%		0.0%	0.0%	34.4%	56.0%	0.0%	0.0%





Example 3

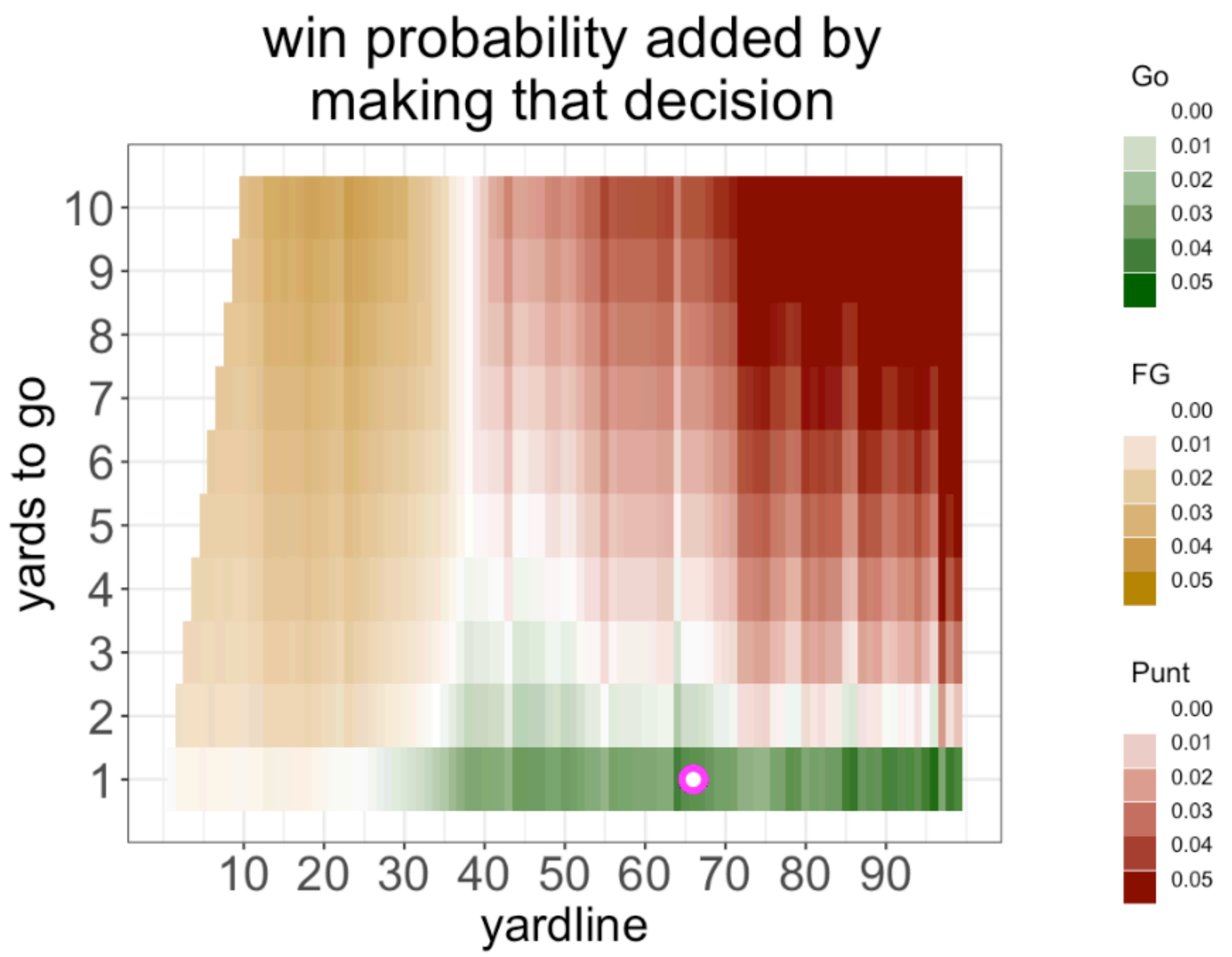
- LV @ LA in Week 14 of 2022

Go is a strong decision based on the WP point estimate (+3.5%).
Traditional analytics recommendation: **Strong Go!**

Up 6, 4th & 1, 66 yards from opponent endzone

Qtr 4, 2:00 | Timeouts: Off 3, Def 3 | Point Spread: -6.5

decision	WP	success prob	WP if fail	WP if succeed	SD of WP	baseline coach %
Go for it	92.3%	68.8%	78.1%	98.7%	9.6%	22.9%
Punt	88.8%					77.1%
Field goal	78.1%	0.0%	78.1%	98.0%	0.0%	0.0%



Example 3

- LV @ LA in Week 14 of 2022

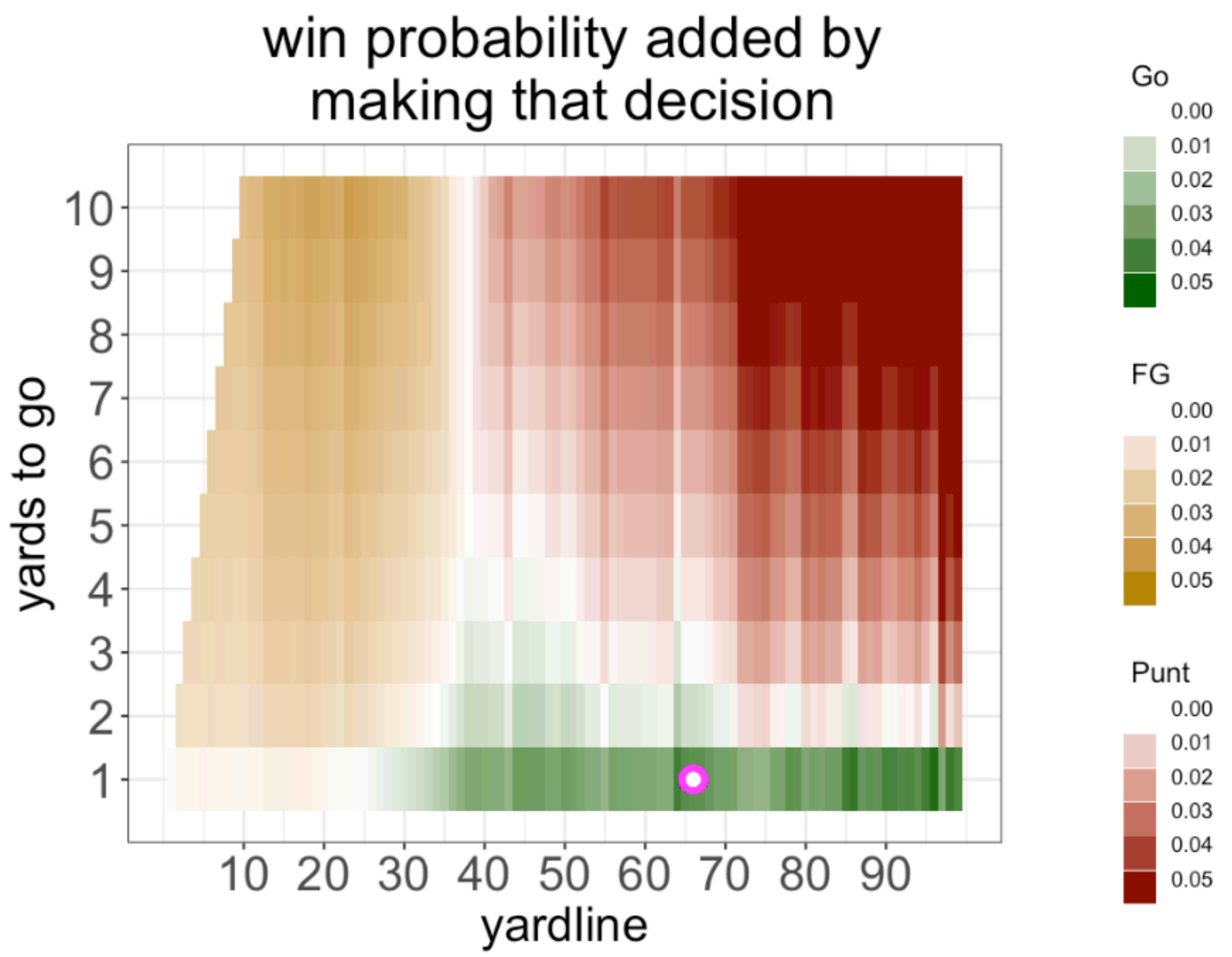
Go is a strong decision based on the WP point estimate, and we are certain it is the best decision.

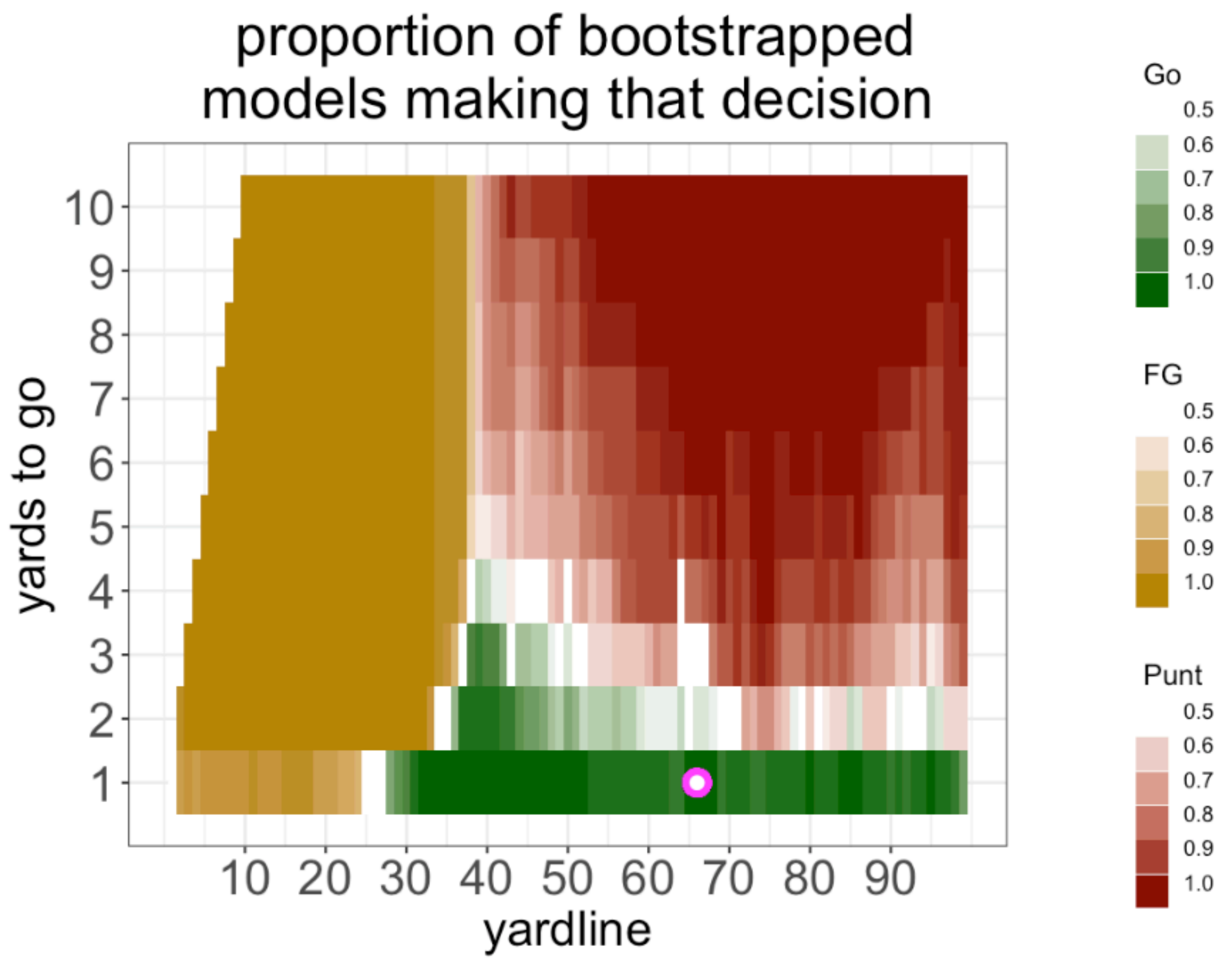
Our recommendation: **Strong Go!** (LV punted, then LA won after a Mayfield 98 yard game winning drive).

Up 6, 4th & 1, 66 yards from opponent endzone

Qtr 4, 2:00 | Timeouts: Off 3, Def 3 | Point Spread: -6.5

decision	WP	WP gain CI	boot %	success prob	WP if fail	WP if succeed	SD of WP	baseline coach %
Go for it	92.3%	[0.6%, 5.0%]	100.0%	68.8%	78.1%	98.7%	9.6%	22.9%
Punt	88.8%		0.0%					77.1%
Field goal	78.1%		0.0%	0.0%	78.1%	98.0%	0.0%	0.0%





Analytics, Have Some Humility

- Team quality *must* be incorporated into EP/WP models
- We need shrinkage to mitigate overfitting in our ML models
- ***Humility:*** There are not enough games to fit an accurate statistical WP model to precisely learn the right fourth down decision at many game-states
- Far fewer 4th down decisions are as obvious as analysts widely claim

- **Thank you!**
- Twitter: **@RyanBrill_**
- Email: **ryguy123@sas.upenn.edu**