Statistics in Sports: Football (Soccer) Overview

ZACHARY BINNEY, PHD MPH

OXFORD COLLEGE OF EMORY UNIVERSITY

FALL 2021

Disclosure

- Not a soccer analytics guy
- Incomplete, may not even focus on most important current topics

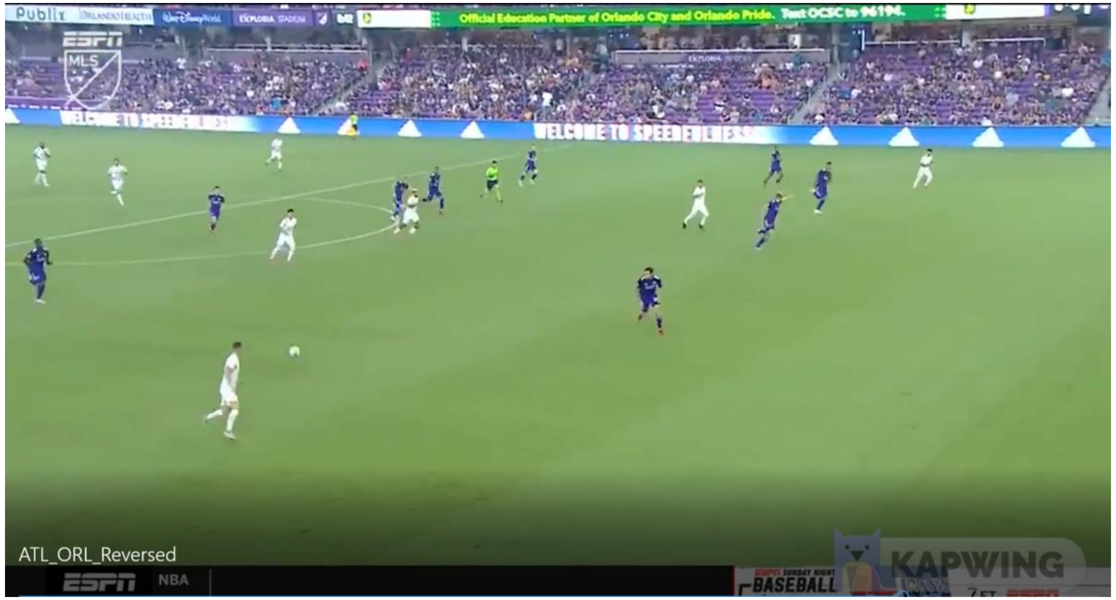
Roadmap

- 1. Intro and Types of Soccer Data
  - Events vs. Tracking
- 2. Expected Goals (xG) Models
- 3. Beyond Shots
- 4. Player Evaluation
  - Radar Charts

Intro and Types of Soccer Data

Intro to Soccer Analytics

- Let’s watch a goal… But in reverse…
- Highlights from Atlanta United-Orlando City FC match, October 2021
- Video in lecture folder



Intro to Soccer Analytics

- Let’s watch Atlanta United’s first match and first goal ever (!)
  - 2017 vs. NY Red Bull, start around 26:00
  - How might you break this down into data a computer could analyze?

Intro to Soccer Analytics

- What if you paused every time something “interesting” or “noteworthy” happened, and logged that?

  - **Events Data**

JUN. 10, 2014, AT 3:58 PM

The People Tracking Every Touch, Pass And Tackle in the World Cup

By Carl Bialik

moods and preferences. Throughout the year, 350 part-time analysts working in London and a half-dozen other Opta branches in Europe and North and South America record every pass, header and goal while watching live or recorded video of more than 14,000 matches around the world. The

team’s match to confirm every goal was attributed correctly. And I watched as Opta’s media team processed the raw numbers — 1,600 to 2,000 events per game — into TV-ready factoids, which they heard commentators repeat

But most of the work is logging routine passes. Opta’s analysts log each one by dragging and clicking a mouse at the spot where the pass was received, then keying in the player who received it. Their monitors have an image of a

Source: https://fivethirtyeight.com/features/the-people-tracking-every-touch-pass-and-tackle-in-the-world-cup/

Intro to Soccer Analytics

- **Events Data** example: Statsbomb (105 variables)

	period	minute	second	possession	possession_team.name	player.name	type.name	position.name	duration	location
4	1	0	0	1	Houston Dash	NA	Half Start	NA	0.000	NULL
5	1	0	0	2	Utah Royals	Diana Matheson	Pass	Center Attacking Midfield	1.204	c(60, 40)
6	1	0	1	2	Utah Royals	Katrina Gorry	Ball Receipt*	Right Attacking Midfield	NA	c(53, 39)
7	1	0	1	2	Utah Royals	Katrina Gorry	Pass	Right Attacking Midfield	3.070	c(93, 18)
8	1	0	4	3	Houston Dash	Amber Brooks	Pass	Right Center Back	1.372	c(28, 63)
9	1	0	6	3	Houston Dash	Kealia Ohai	Ball Receipt*	Right Center Midfield	NA	c(49, 71)
10	1	0	6	3	Houston Dash	Kealia Ohai	Carry	Right Center Midfield	3.720	c(49, 71)
11	1	0	7	3	Houston Dash	Katrina Gorry	Pressure	Right Attacking Midfield	2.292	c(68, 14)
12	1	0	9	3	Houston Dash	Kealia Ohai	Pass	Right Center Midfield	2.000	c(63, 74)
13	1	0	11	3	Houston Dash	Nichelle Patrice Prince	Ball Receipt*	Right Center Forward	NA	c(105, 70)
14	1	0	11	3	Houston Dash	Nichelle Patrice Prince	Carry	Right Center Forward	3.200	c(105, 70)
15	1	0	15	3	Houston Dash	Nichelle Patrice Prince	Pass	Right Center Forward	0.804	c(118, 68)
16	1	0	15	3	Houston Dash	Rachel Daly	Ball Receipt*	Left Center Forward	NA	c(115, 45)
17	1	0	15	3	Houston Dash	Rachel Corsie	Clearance	Left Center Back	0.000	c(5, 33)
18	1	0	42	4	Houston Dash	Sofia Huerta	Pass	Center Attacking Midfield	2.193	c(120, 80)
19	1	0	44	4	Houston Dash	Diana Matheson	Clearance	Center Attacking Midfield	0.000	c(12, 47)
20	1	0	47	4	Houston Dash	Diana Matheson	Pressure	Center Attacking Midfield	0.649	c(9, 51)
21	1	0	47	4	Houston Dash	Nichelle Patrice Prince	Pass	Right Center Forward	1.893	c(108, 21)
22	1	0	49	4	Houston Dash	Rebecca Elizabeth Sauerbrunn	Clearance	Right Center Back	0.000	c(13, 48)

Source: StatsBombR package

Intro to Soccer Analytics

- **Events Data** example: Statsbomb (105 variables)

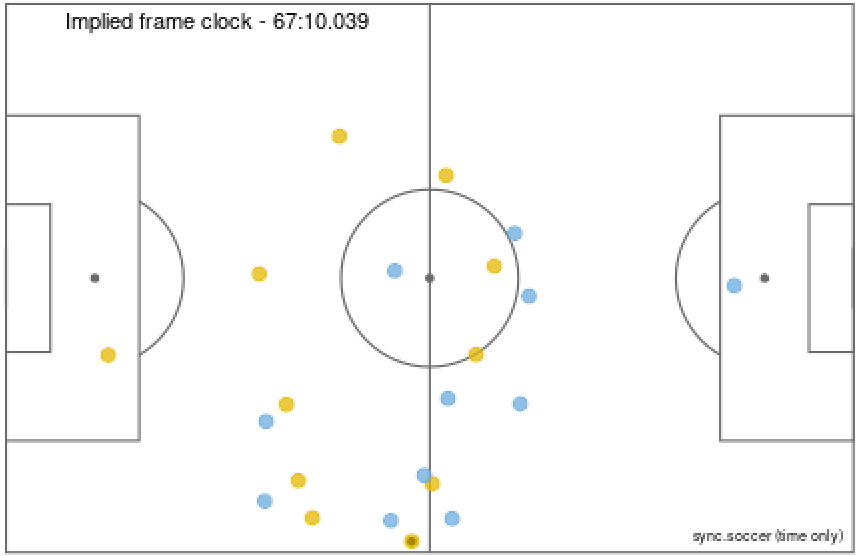
play_pattern.name	team.name	pass.length	pass.angle	pass.end_location	pass.cross	pass.switch	pass.aerial_won	pass.assisted_shot_id
Regular Play	Houston Dash	NA	NA	NULL	NA	NA	NA	NA
From Kick Off	Utah Royals	7.071068	-2.99969550	c(53, 39)	NA	NA	NA	NA
From Kick Off	Utah Royals	NA	NA	NULL	NA	NA	NA	NA
From Kick Off	Utah Royals	0.000000	0.00000000	c(93, 18)	NA	NA	NA	NA
Regular Play	Houston Dash	22.472204	0.36397895	c(49, 71)	NA	NA	NA	NA
Regular Play	Houston Dash	NA	NA	NULL	NA	NA	NA	NA
Regular Play	Houston Dash	NA	NA	NULL	NA	NA	NA	NA
Regular Play	Utah Royals	NA	NA	NULL	NA	NA	NA	NA
Regular Play	Houston Dash	42.190044	-0.09495170	c(105, 70)	NA	NA	NA	NA
Regular Play	Houston Dash	NA	NA	NULL	NA	NA	NA	NA
Regular Play	Houston Dash	NA	NA	NULL	NA	NA	NA	NA
Regular Play	Houston Dash	20.099750	-1.67046500	c(116, 48)	TRUE	NA	NA	NA
Regular Play	Houston Dash	NA	NA	NULL	NA	NA	NA	NA
Regular Play	Utah Royals	NA	NA	NULL	NA	NA	NA	NA
From Corner	Houston Dash	47.296936	-1.80551900	c(109, 34)	NA	TRUE	NA	NA
From Corner	Utah Royals	NA	NA	NULL	NA	NA	NA	NA
From Corner	Utah Royals	NA	NA	NULL	NA	NA	NA	NA
From Corner	Houston Dash	12.000000	1.57079640	c(108, 33)	TRUE	NA	NA	NA
From Corner	Utah Royals	NA	NA	NULL	NA	NA	NA	NA

Source: StatsBombR package

Events vs. Tracking Data

- **Events Data** is distinct from and simpler than…

  - **Tracking Data**, which is similar to that in (American) football and comes from GPS chips, RFID tags, and/or cameras tracking X-Y location, speed, and direction of every player and ball multiple times per second



Source: https://kwiatkowski.io/sync.soccer

Middle Ground?

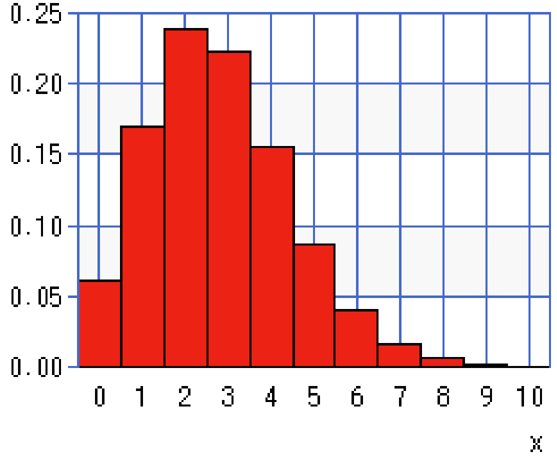
- Context allows for better assessment of decision-making (e.g. was a better passing lane open, did a pass beat the defense's line, etc.)
- Tracking data can grant even more (what happens between events), but…
  - Requires tech + data investments
  - Voluminous and unwieldy
  - Have to figure out value and how to analyze

Source: https://fivethirtyeight.com/features/soccer-is-learning-to-see-the-whole-game/

Expected Goal (xG) Models

xG Models

- How do you win a soccer match?
- Problem: (actual) **goals** are rare, noisy
  - Poisson distribution with mean of 2.8 goals (roughly Norwich City, 2021/22):



- To reduce noise, instead of how a team did look at how they *should* have done

xG Models

- **Expected Goals (xG) Models**. Based on:

Location of shooter: How far was it from the goal and at what angle on the pitch?
Body part: Was it a header or off the shooter's foot?
Type of pass: Was it from a through ball, cross, set piece, etc?
Type of attack: Was it from an established possession? Was it off a rebound? Did the defense have time to get in position? Did it follow a dribble?

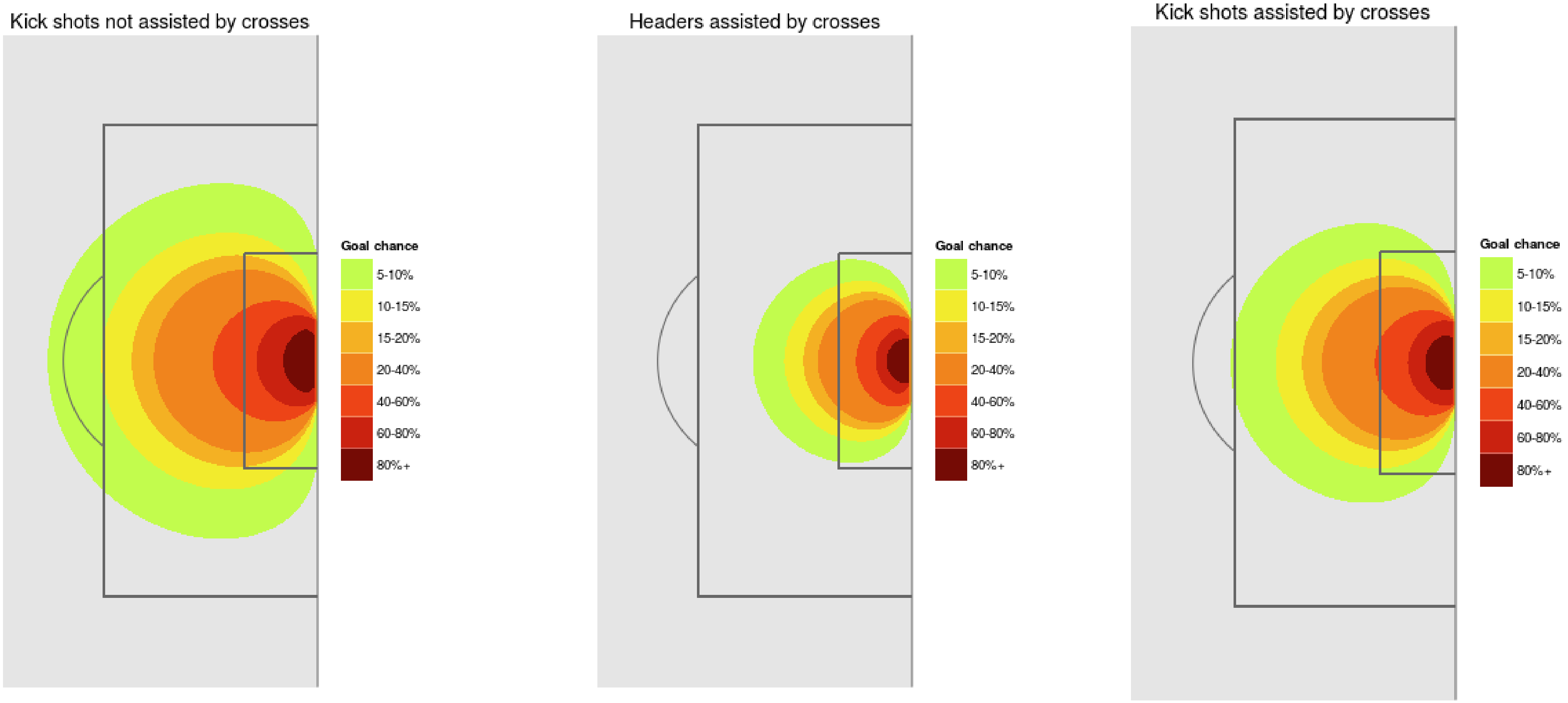
the location of all players on the pitch at the moment the shot was taken. Was the goalkeeper in position? Was it an open goal or were there a number of defenders between the shooter and the goal? Was the shooter being pressured? Was it a 1v1 situation with the keeper?

- Average result (% of goals that went in) in similar situations → xG for that shot
  - What method do you think is (or could be) used to estimate xG values?
- How do you think each of these affects xG?

Source: https://fbref.com/en/expected-goals-model-explained/

xG Models: Location

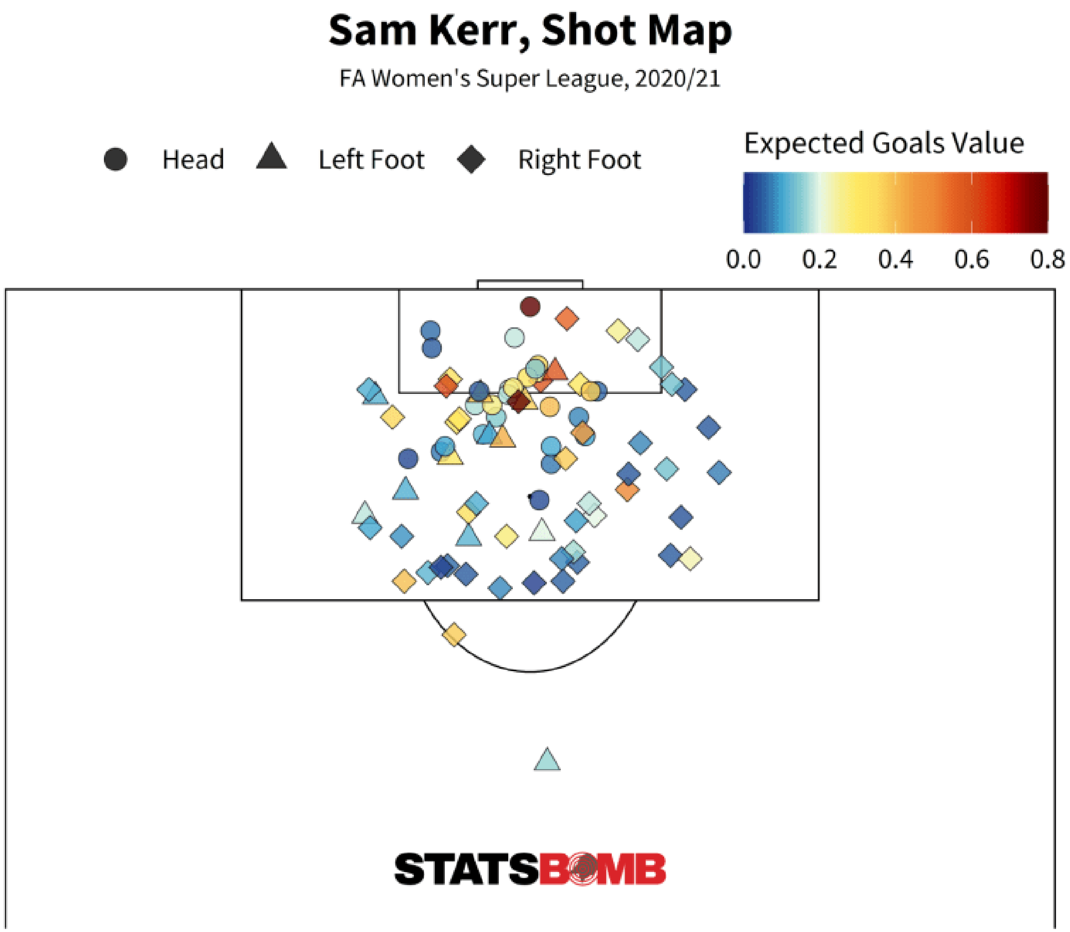
- **xG** models and position on pitch



Source: https://statsbomb.com/2016/04/explaining-and-training-shot-quality/

xG Models: Uses

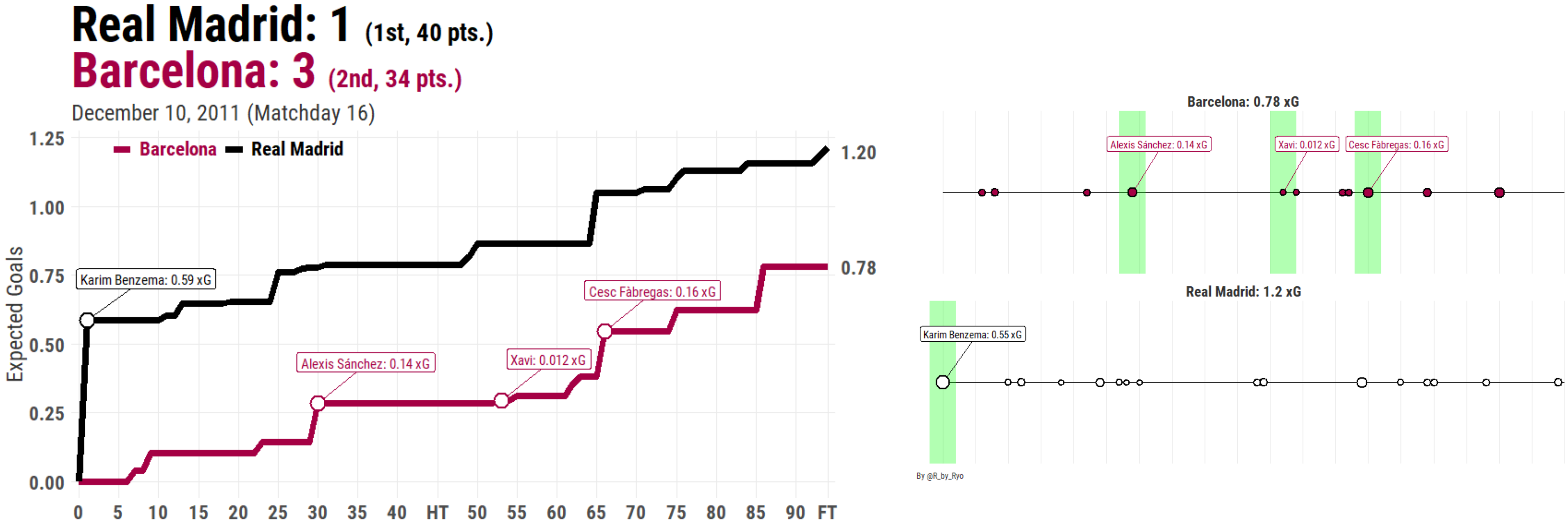
- **xG** model uses



Source: https://statsbomb.com/2021/10/statsbomb-release-free-2020-21-fa-womens-super-league-data-updated-r-guide/

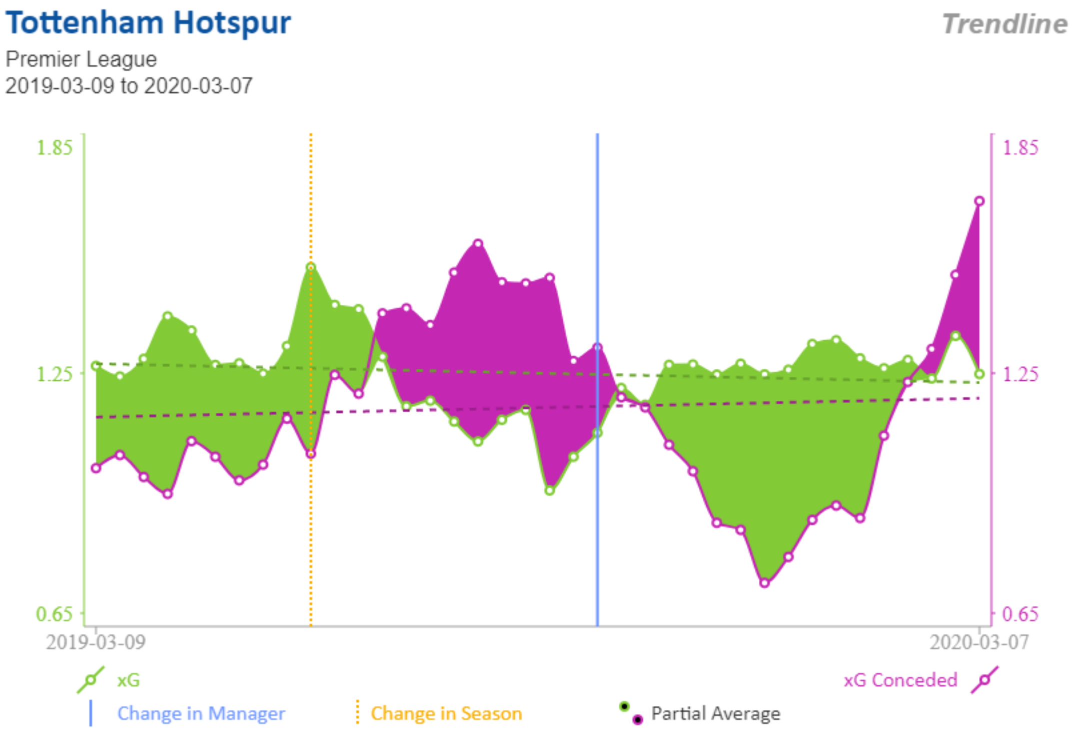
xG Models: Uses

- **xG** model uses



xG Models: Uses

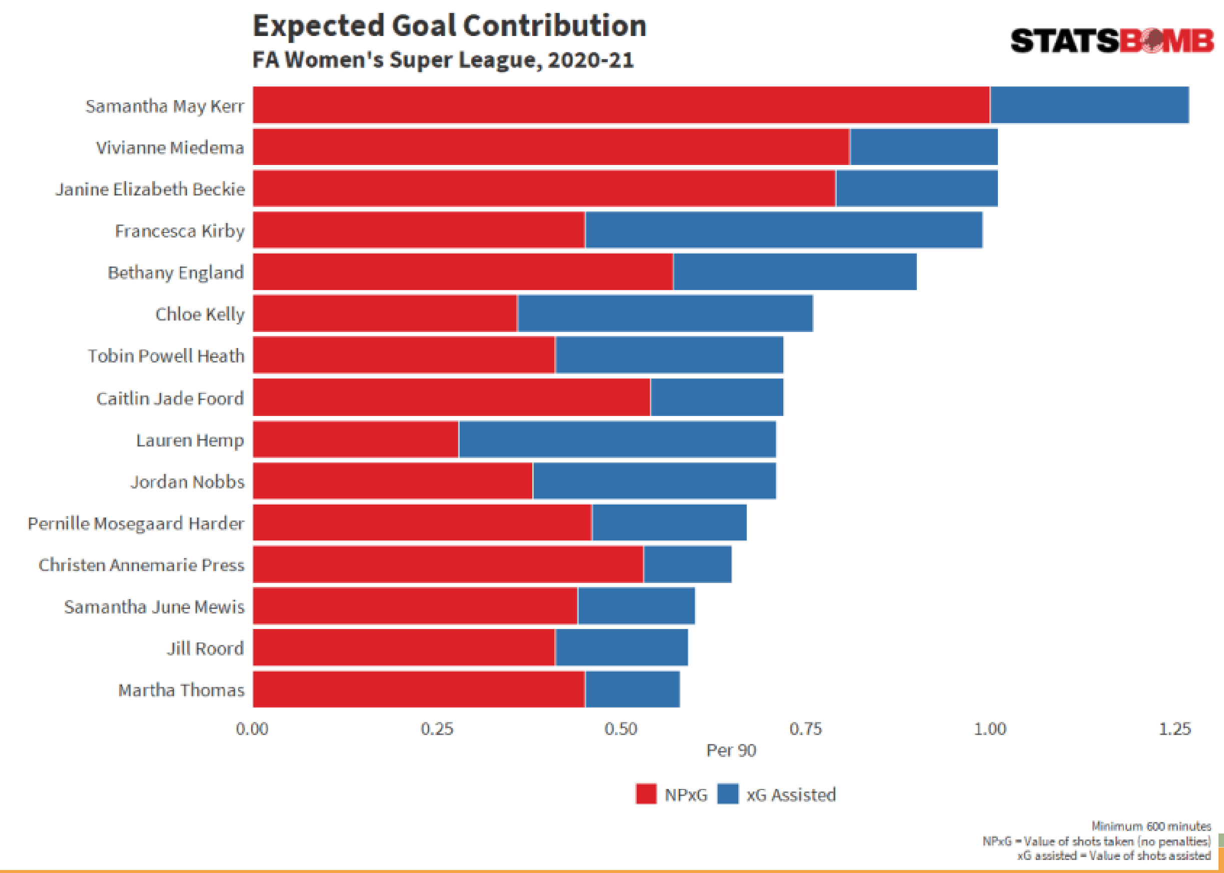
- **xG** model uses



Source: https://fbref.com/en/statsbomb/

xG Models: Uses

- Can add in **expected assists (xA)** for more holistic picture



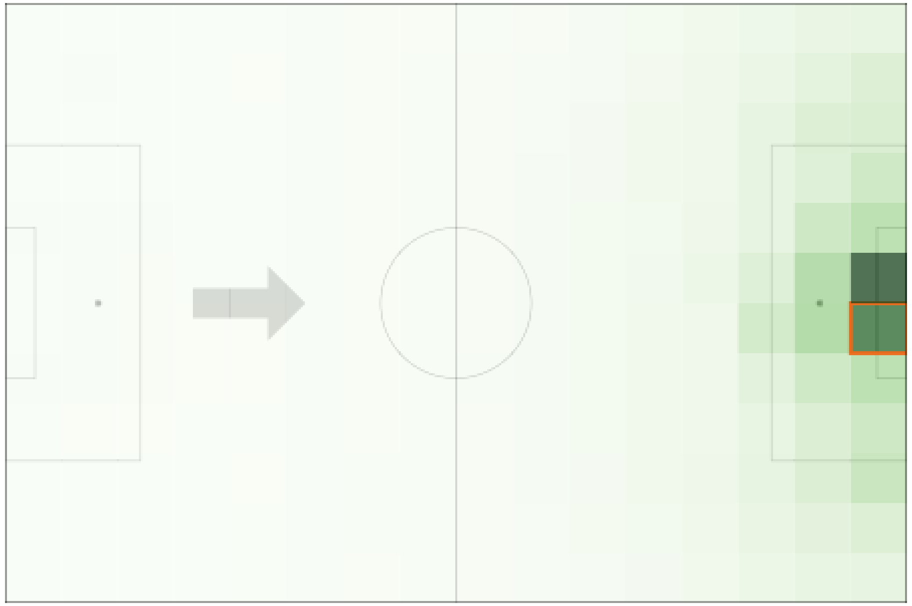
Source: https://statsbomb.com/2021/10/statsbomb-release-free-2020-21-fa-womens-super-league-data-updated-r-guide/

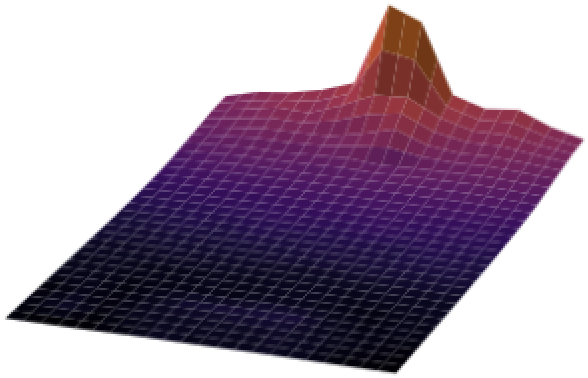
xG Models: Extensions

- Related model: **expected thread (xT)**

Expected Threat (xT) = 0.371

i.e. when the team has the ball in the highlighted zone, they will score in the next **5** actions **37.1%** of the time.





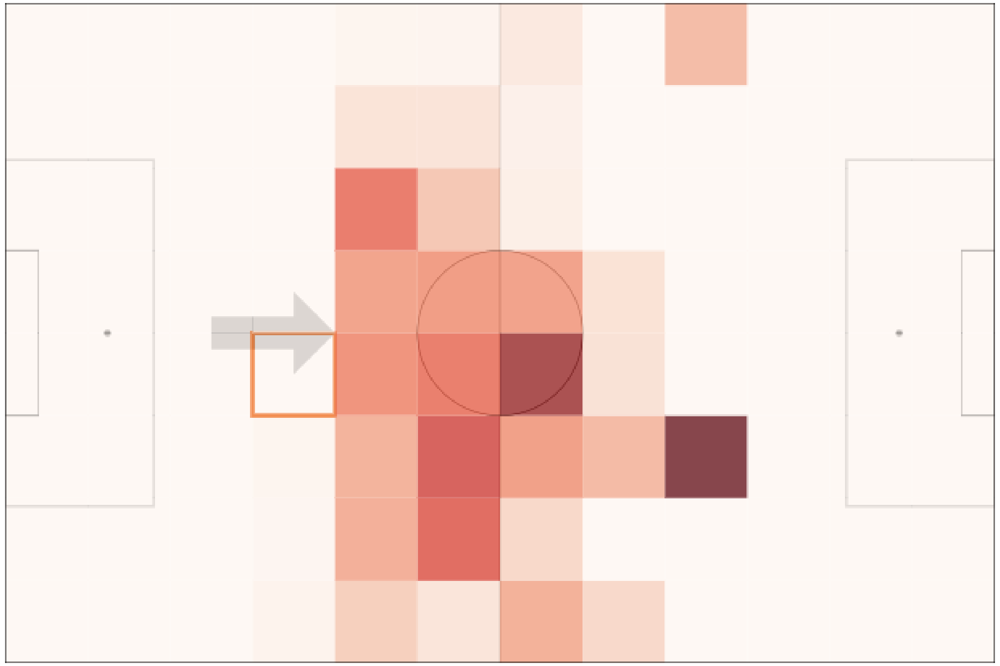
Source: https://karun.in/blog/expected-threat.html

xG Models: Extensions

- Related model: **expected thread (xT)**. Take a step back…

Who creates danger from where?

Manchester City



Manchester City

When Manchester City has the ball in the highlighted zone, they create danger by moving the ball into the red zones.

Top contributors from this zone:

1. Fernandinho
2. Vincent Kompany
3. Nicolás Otamendi

(clear selected zone)

Source: https://karun.in/blog/expected-threat.html

xG Models: Extensions

- Ultimate goal: **possession value** to be able to estimate value of individual actions on the field regardless of position

- Can you link this quest with any concepts we've seen in other sports?

Soccer possession models are gaining steam

Key soccer possession models by publication year, with type of model and possession information

NAME	CREATOR	DEBUT	METHOD	WINDOW	OFF-BALL INFORMATION
Markov Chains	S. Rudd	2011	Markov chain	One possession	Defensive states tagged in event data
Possession-Based Model	N. Mackay	2016	Logistic regression and GAM	One possession	None
Expected Threat (xT)	K. Singh	2019	Markov-like	Next 5 actions (goal for)	None
Valuing Actions by Estimating Probabilities (VAEP)	KU Leuven DTAI	2019	Gradient-boosted trees	Next 10 actions (goal for or against)	Possession history proxies
Expected Possession Value (EPV)	J. Fernández et al.	2019	Multiple models	Next goal (for or against) or end of half	Full tracking data
Possession Value (PV)	Stats Perform	2019	Gradient-boosted trees	Next 10 seconds (goal for)	Possession history proxies
Goals Added (g+)	American Soccer Analysis	2020	Gradient-boosted trees	Two possessions	Possession history proxies
On-Ball Value (OBV)	StatsBomb	2021	Gradient-boosted trees	Two possessions	Broadcast freeze frames (in development)

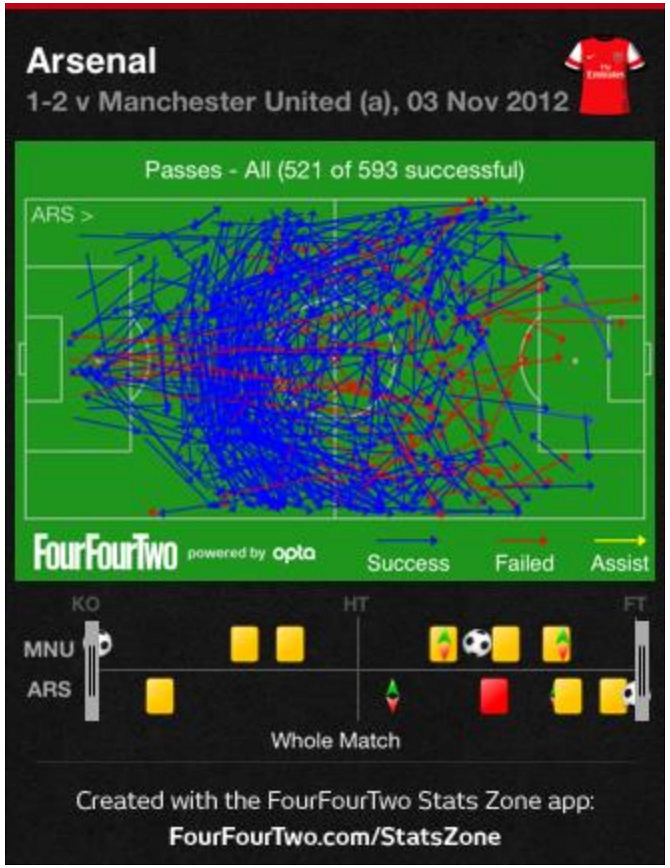
FiveThirtyEight

Source: https://fivethirtyeight.com/features/possession-is-the-puzzle-of-soccer-analytics-these-models-are-trying-to-solve-it/

Beyond Shots

Passing

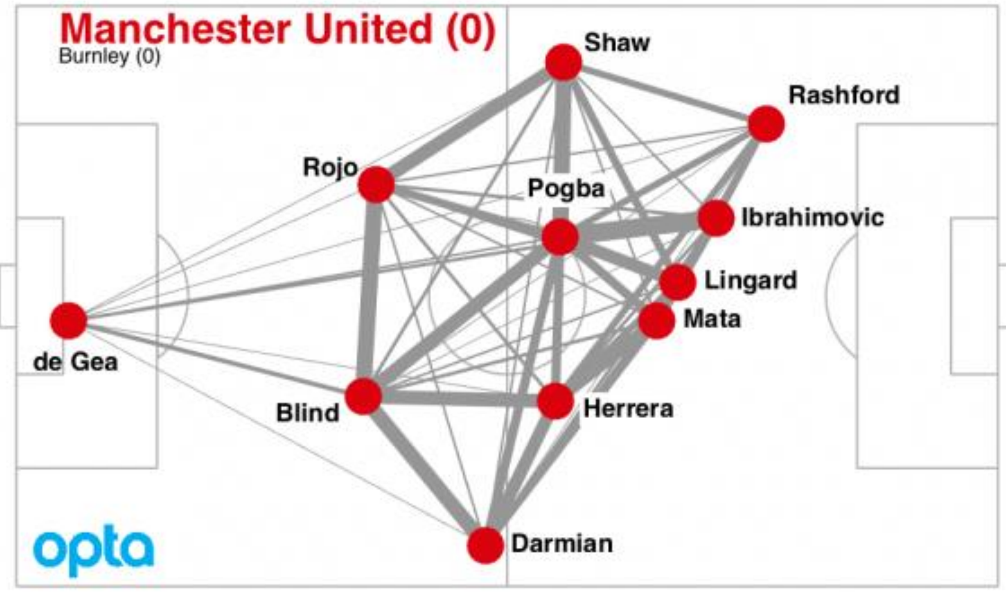
- “Raw” passing data map

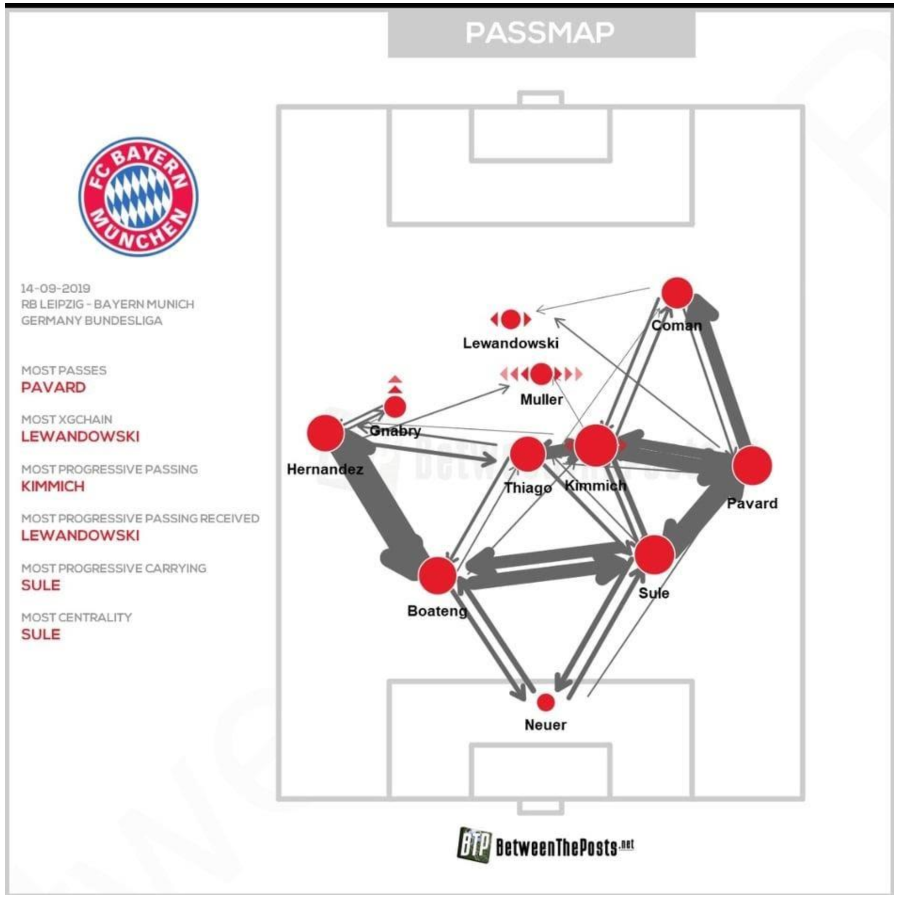


Source: https://theshortfuse.sbnation.com/2012/11/3/3595432/manchester-united-v-arsenal-stats-premier-league-2012/

Passing

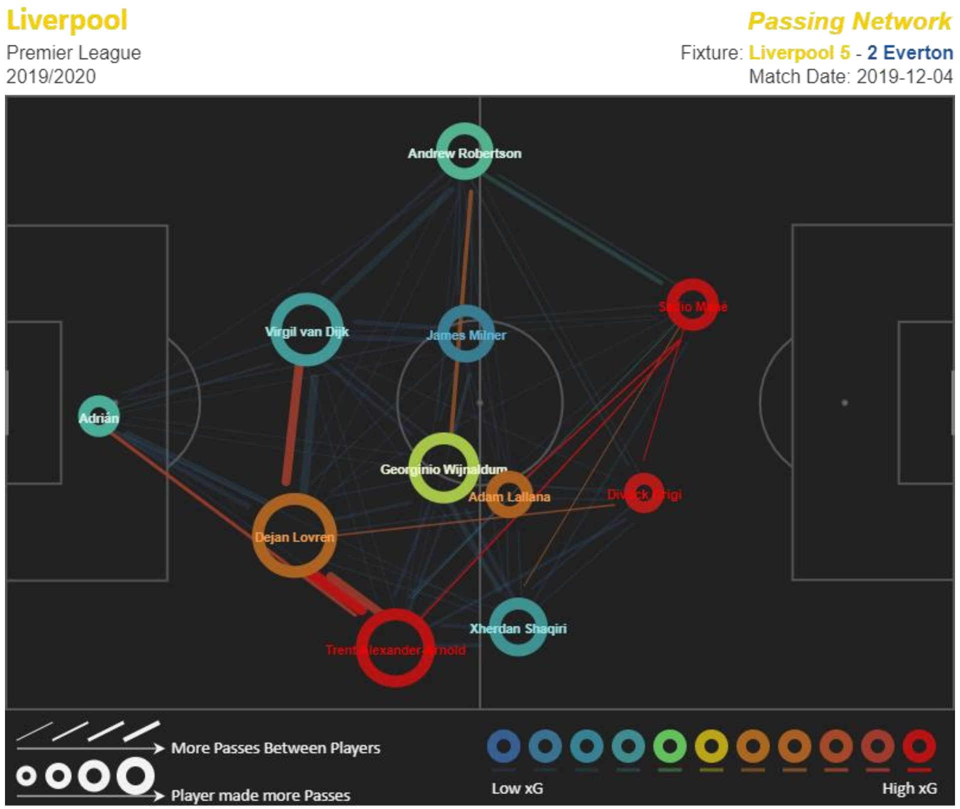
- Passing Networks





Passing

- Passing Networks + value of possessions featuring various networks



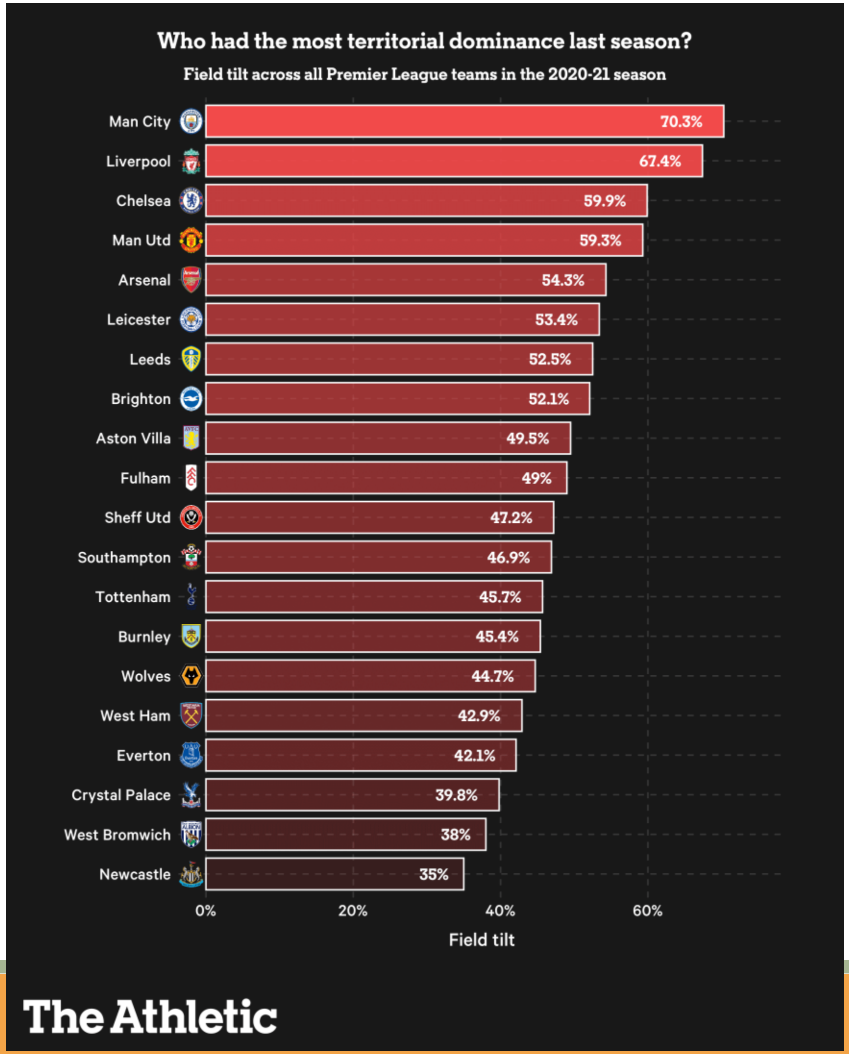
STATSBOMB

More On This Topic

- Our guest speaker will talk more analyzing passing data

Field Tilt and Territorial Dominance

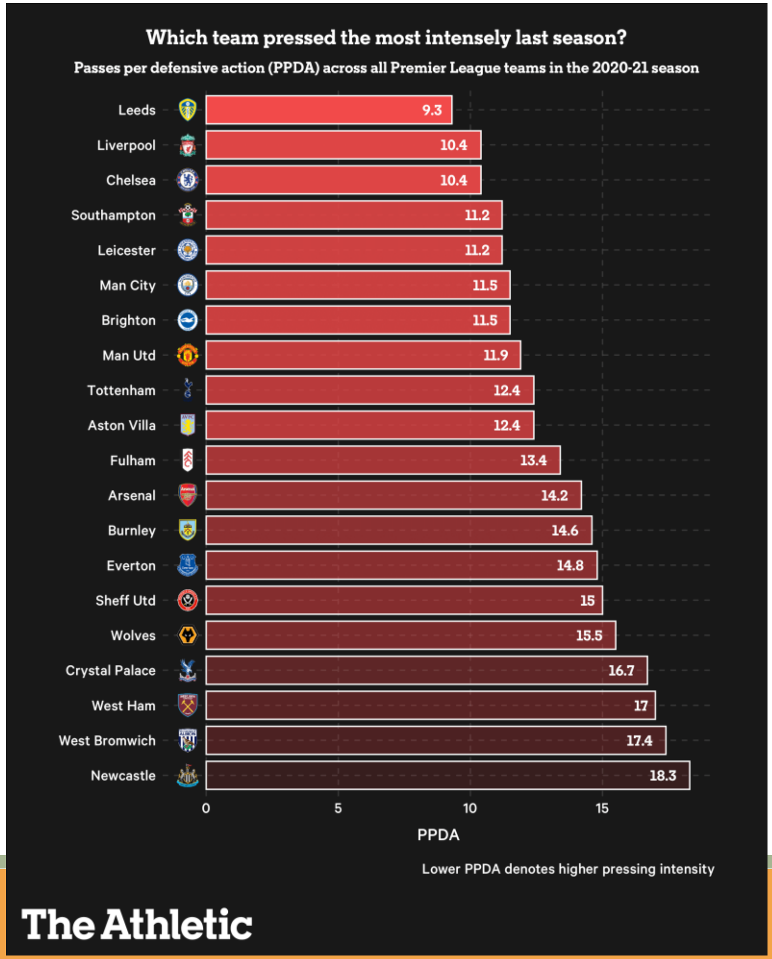
- Field tilt: % of total team + opponent passes in "final third" done by team



Source: https://theathletic.com/2730755/2021/07/28/the-athletics-football-analytics-glossary-explaining-xg-ppda-field-tilt-and-how-to-use-them/

DE-FENSE!

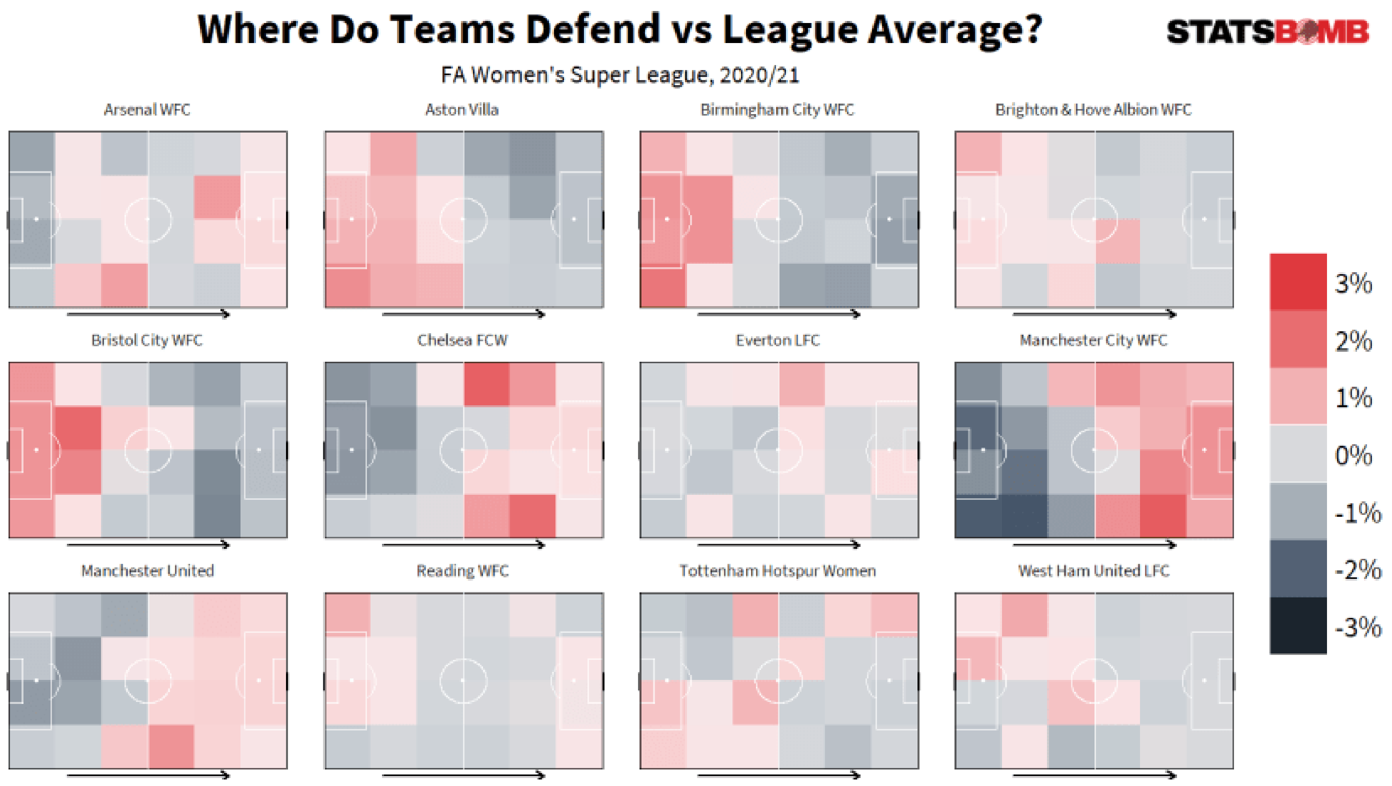
- How aggressive are defenses? **Passes per defensive action (PPDA)**



Source: https://theathletic.com/2730755/2021/07/28/the-athletics-football-analytics-glossary-explaining-xg-ppda-field-tilt-and-how-to-use-them/

DE-FENSE!

- Defending zones, where teams commit defensive actions



Source: https://statsbomb.com/2021/10/statsbomb-release-free-2020-21-fa-womens-super-league-data-updated-r-guide/

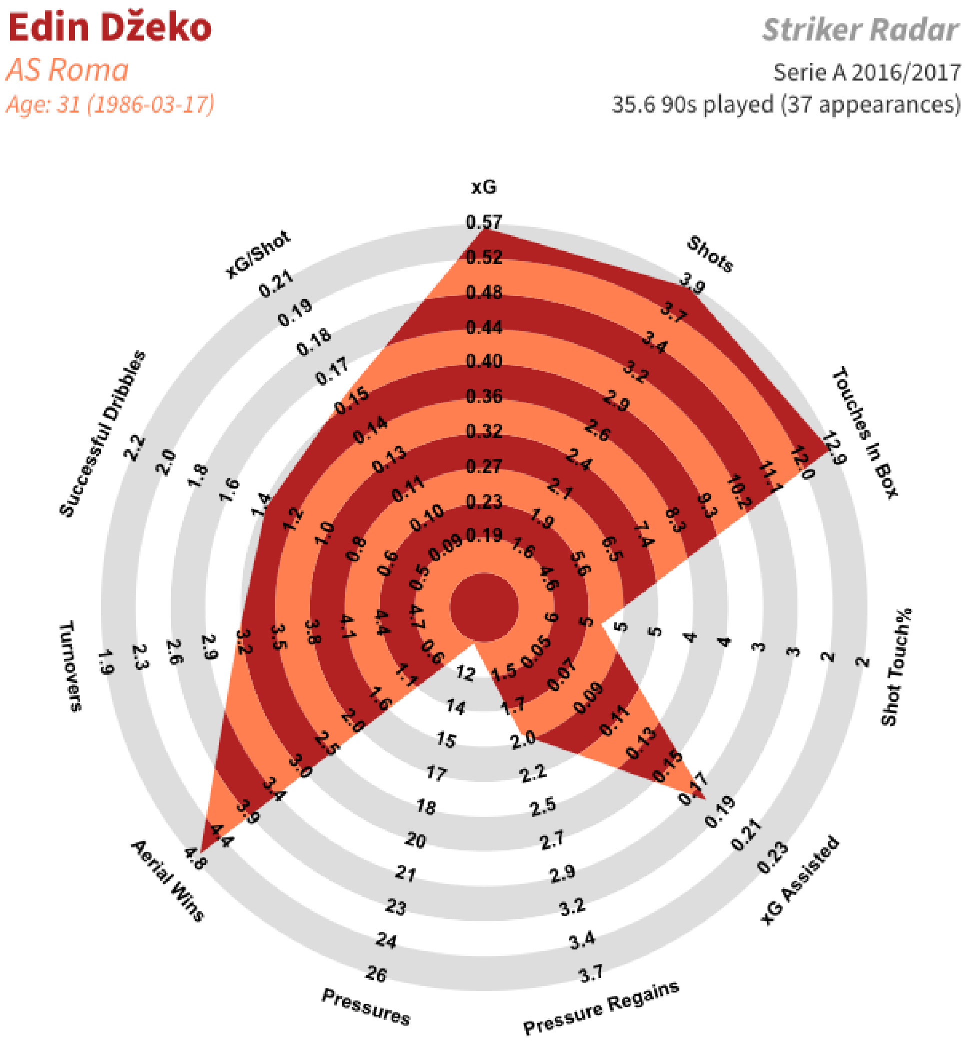
Player Evaluation

RADAR CHARTS

MORE TO COME

Player Eval: Radars

- Radar Charts from Statsbomb
- Let's break one of these down…
- Limitations?

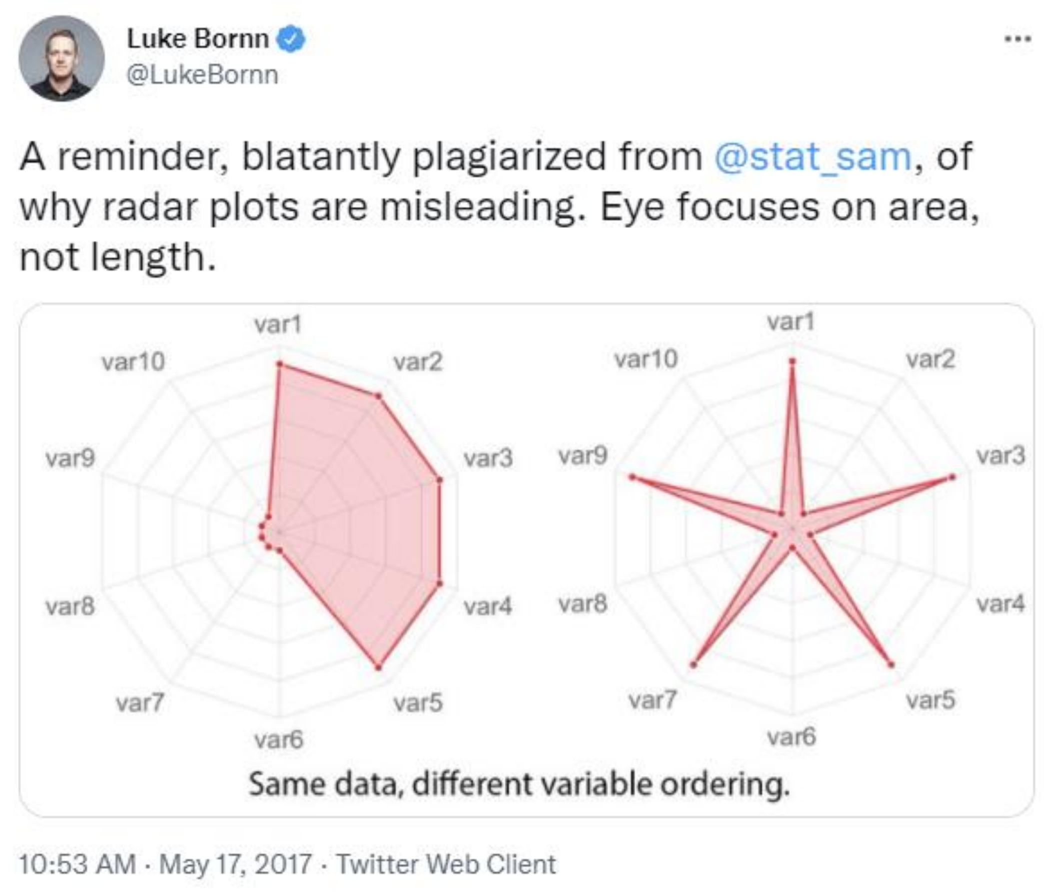


STATSBOMB

Player Eval: Radars

- Radar Charts from Statsbomb

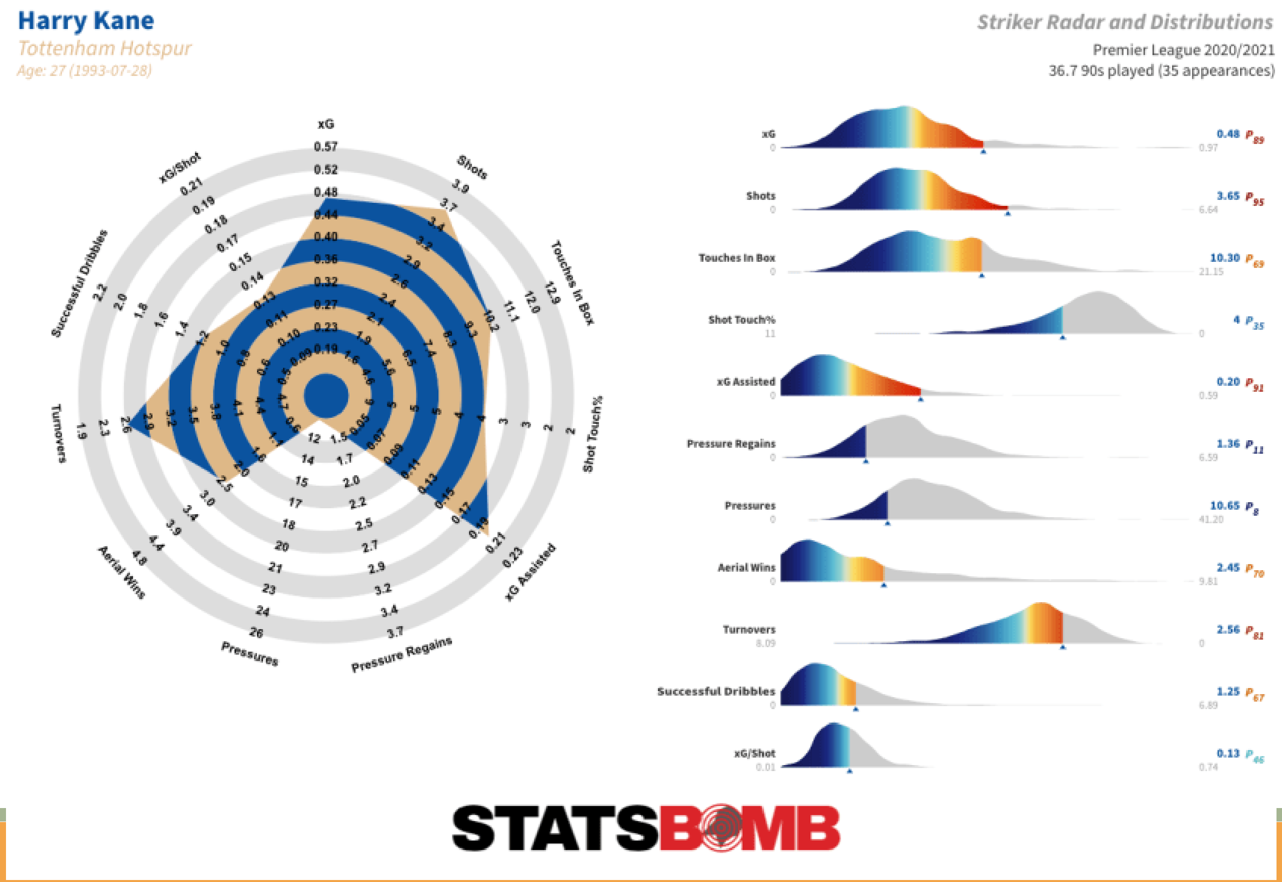
- Limitations?



Source: StatsBomb

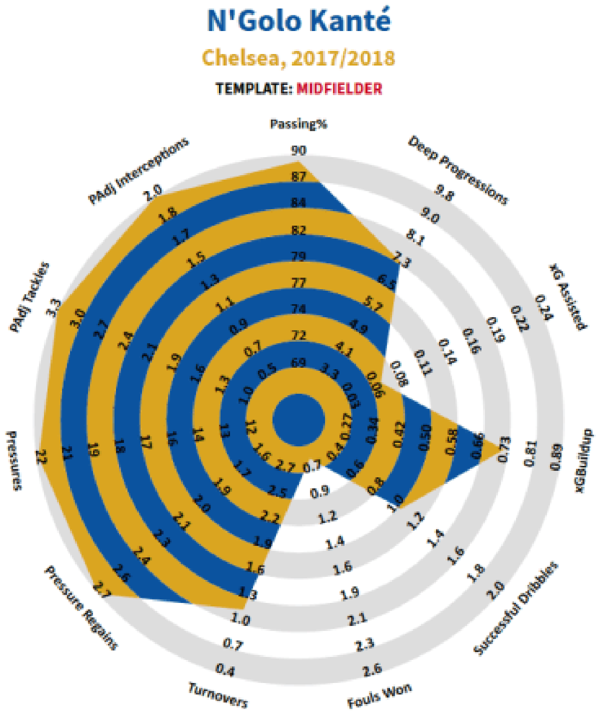
Player Eval: Radars

- Radar Charts from Statsbomb



Source: StatsBomb

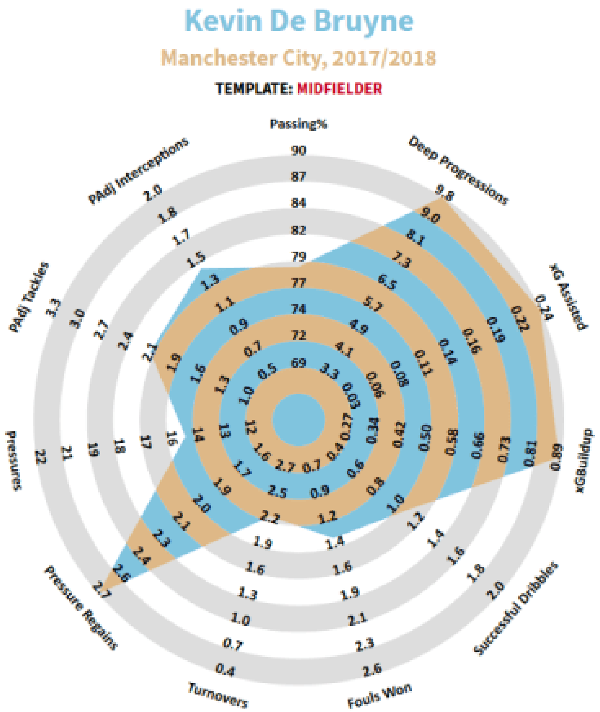
Player Eval: Radars



90s Played	Age	Birth Date	Competition
35.4	27	1991-03-29	Premier League

P%	DProg	xG Asst	xGB	Drib	FLW	Tov	PR	Psr	TK	INT
89	7.40	0.07	0.72	0.99	0.59	1.16	3.31	21.89	3.48	2.77

All units in per 90



90s Played	Age	Birth Date	Competition
36.4	26	1991-06-28	Premier League

P%	DProg	xG Asst	xGB	Drib	FLW	Tov	PR	Psr	TK	INT
79	11.18	0.40	0.88	1.02	1.32	2.22	2.94	14.86	2.18	1.42

All units in per 90

STATSBOMB STATSBOMB

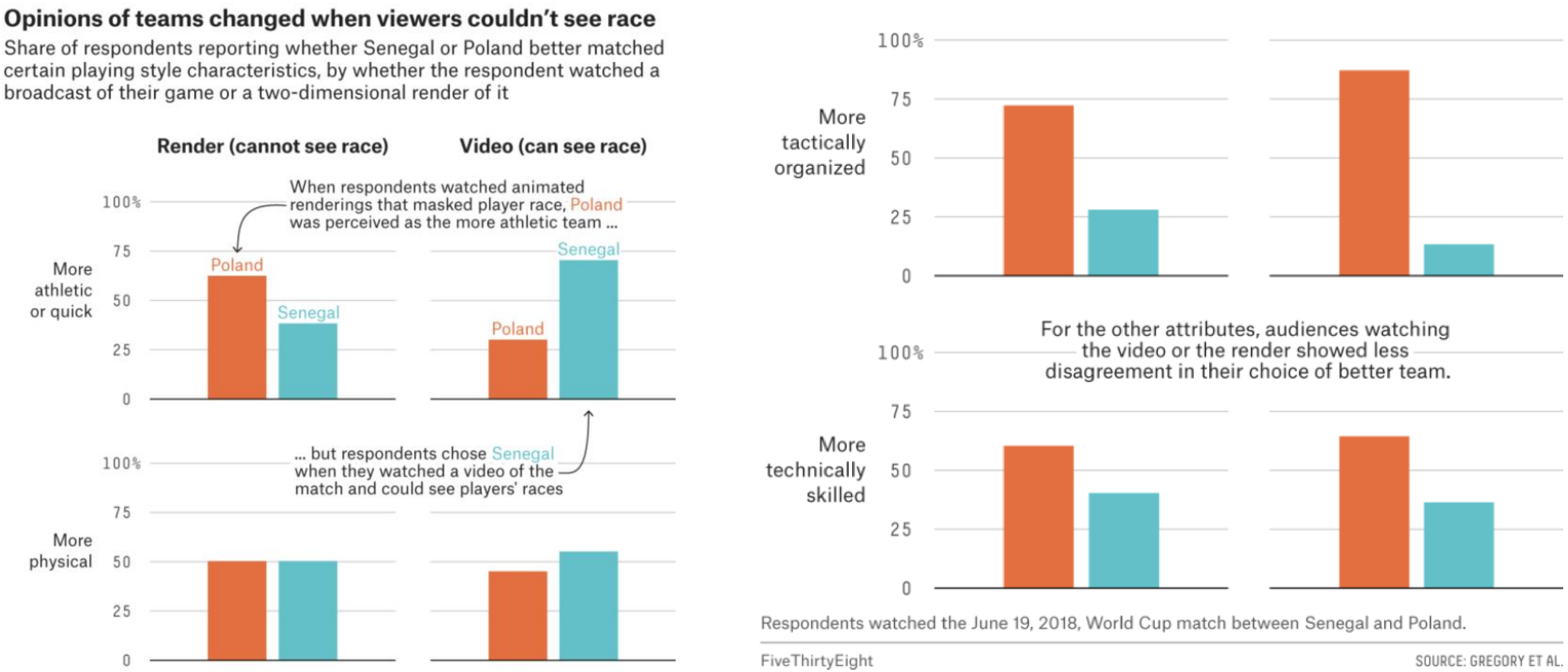
Source: StatsBomb

More On This Topic

- Our guest speaker will talk more about player eval

A Brief Word on Soccer and Racism

Racial Stereotypes in Soccer (Broadcasting, Scouting)



Source: https://fivethirtyeight.com/features/soccer-looks-different-when-you-cant-see-whos-playing/

More Resources

More Resources

- Books
  - *The Numbers Game* (Anderson and Sally)
  - *Soccermatics* (Sumpter)
- Companies/Blogs/Video Series
  - Opta
  - Statsbomb (check out their Academy blog posts; also has some public data!)
    - https://statsbomb.com/2021/10/statsbomb-release-free-2020-21-fa-womens-super-league-data-updated-r-guide/
  - Friends of Tracking on YouTube
  - FiveThirtyEight.com, Soccer tag
- Programming
  - @FC_rstats, shaker and worldfootballR package
  - @FCPython, mplsoccer package
- A *million* online resources, analysts, etc. Build trusted network as you would with other sports.

Thanks!

- Questions? zbinney@emory.edu, @binney_z on Twitter