2020/21 Recruitment Analysis & Competitor Insights

Data Science Group
Christian Gilson
christian@manutd.co.uk

Motivation

Systematic, data-driven approach to player recruitment:

- Identify high and poor performing players within squad;
- Identify transfer targets enabling Man United to:
  - Mitigate key asset losses;
  - Strengthen positions of vulnerability.

Identify opponent strengths & weaknesses to inform matchday strategy.

Competition: on and off the pitch

Since 2014, Barcelona's net spend of €450m is 3x that of 18/19 Champions League winners Liverpool.

Hard to find competitive edge in data-driven football recruitment: complex invasion game with game-changing actions happening off the ball.

We built two complementary models to value offensive actions: every pass, cross, dribble and shot.

Data

On-the-ball events data

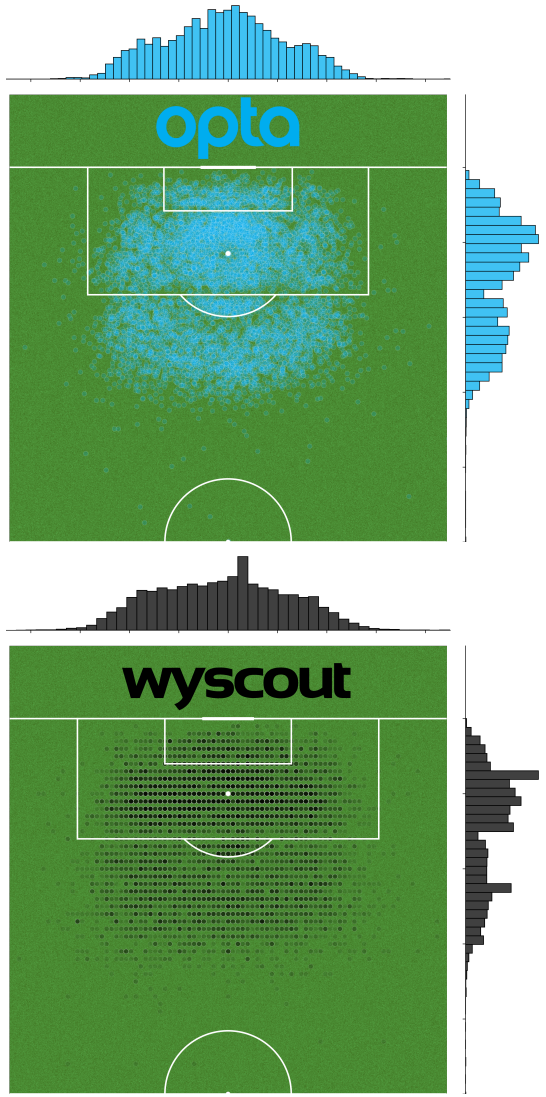
This project utilised on-the-ball events data from two vendors: Opta and Wyscout.

Every on-the-ball action is classified into an events taxonomy and attributed contextual metadata, including *player* and *team* identifiers, *timestamps*, and start and end *positional coordinates*.

Coverage:

- Opta: English Premier League **2017/18**–2020/21;
- Wyscout: Five major European leagues **2017/18**.

Can immediately see differences by looking at shots.

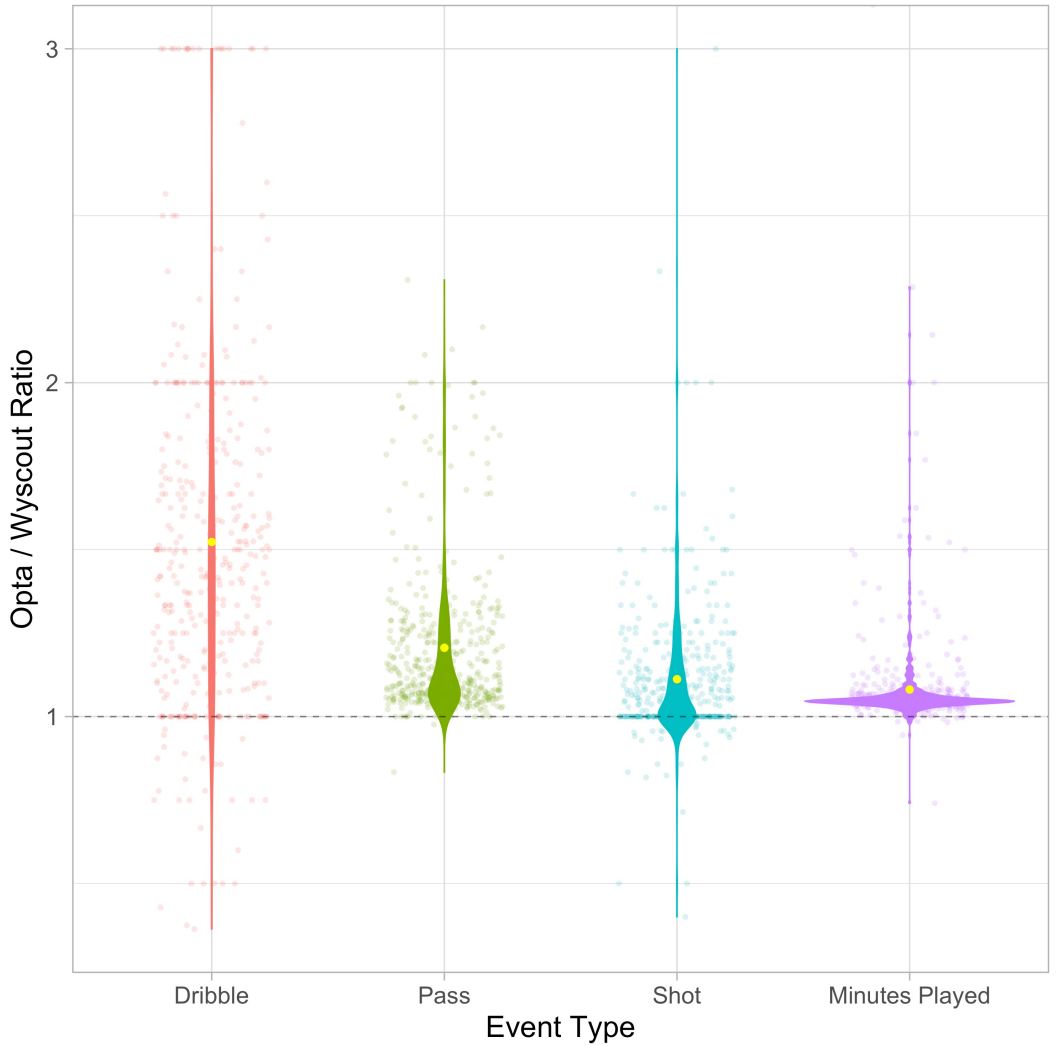


Opta Vs Wyscout: Data quality

Mapped datasets together to perform a systematic data quality comparison.

Calculate Opta:Wyscout ratios for each player in the overlapping 2017/18 season, for the counts of dribbles, passes, shots, mins played.

Wyscout systematically undercounts key actions within a match.



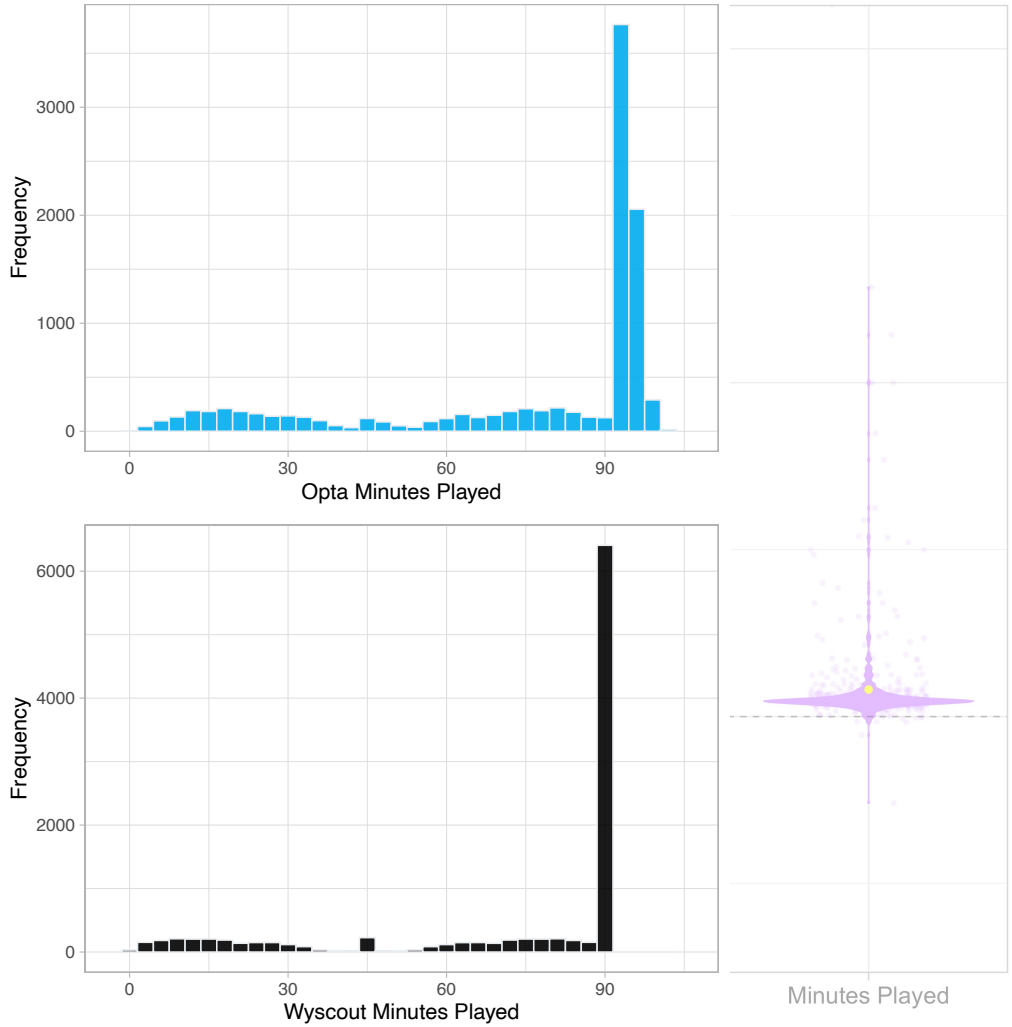
Opta Vs Wyscout: Data quality

Mapped datasets together to perform a systematic data quality comparison.

Calculate Opta:Wyscout ratios for each player in the overlapping 2017/18 season, for the counts of dribbles, passes, shots, mins played.

Wyscout systematically undercounts key actions within a match.

Wyscout winsorises mins played at 90 minutes, a normalising denominator for apples-to-apples comparisons “per 90”.



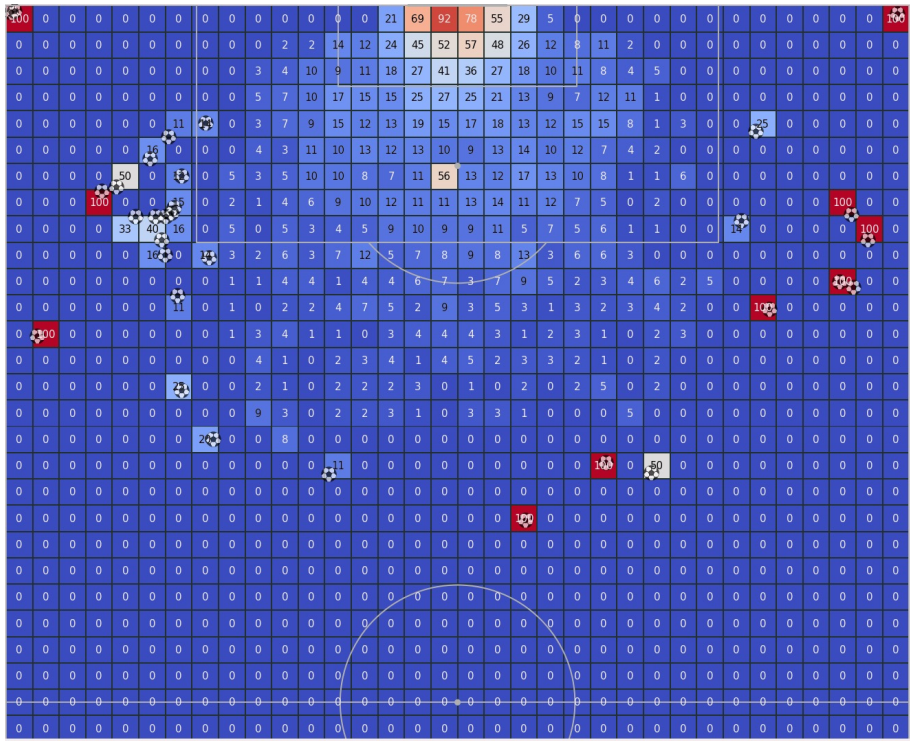
Modelling

Synthetic shots

Subtle limitation of on-the-ball events is the absence of actions in scenarios with a vanishingly slim chance of success, owing to player domain expertise.

Two forms of outlier:

1. “Shots” that weren’t intended as shots;
2. Opportunistic shots where unique context cannot be captured by the data.



Synthetic shots

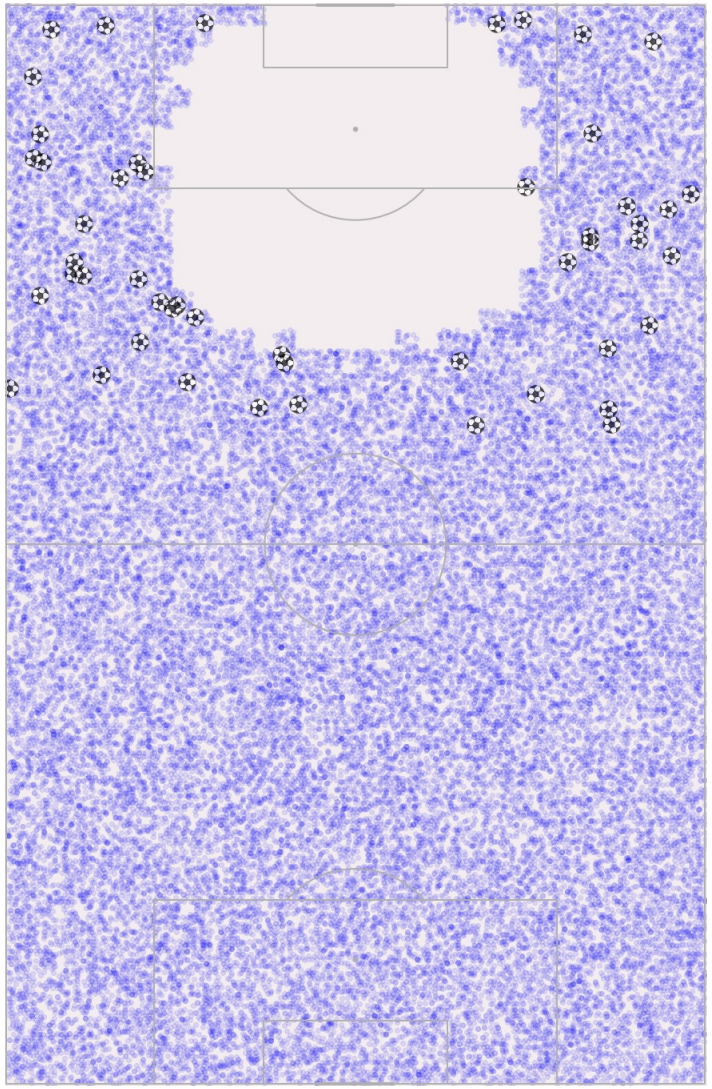
Subtle limitation of on-the-ball events is the absence of actions in scenarios with a vanishingly slim chance of success, owing to player domain expertise.

Two forms of outlier:

1. “Shots” that weren’t intended as shots;
2. Opportunistic shots where unique context cannot be captured by the data.

Add synthetic shots to encode domain expertise missing from the denominator: *the shots not taken because the player had a choice*.

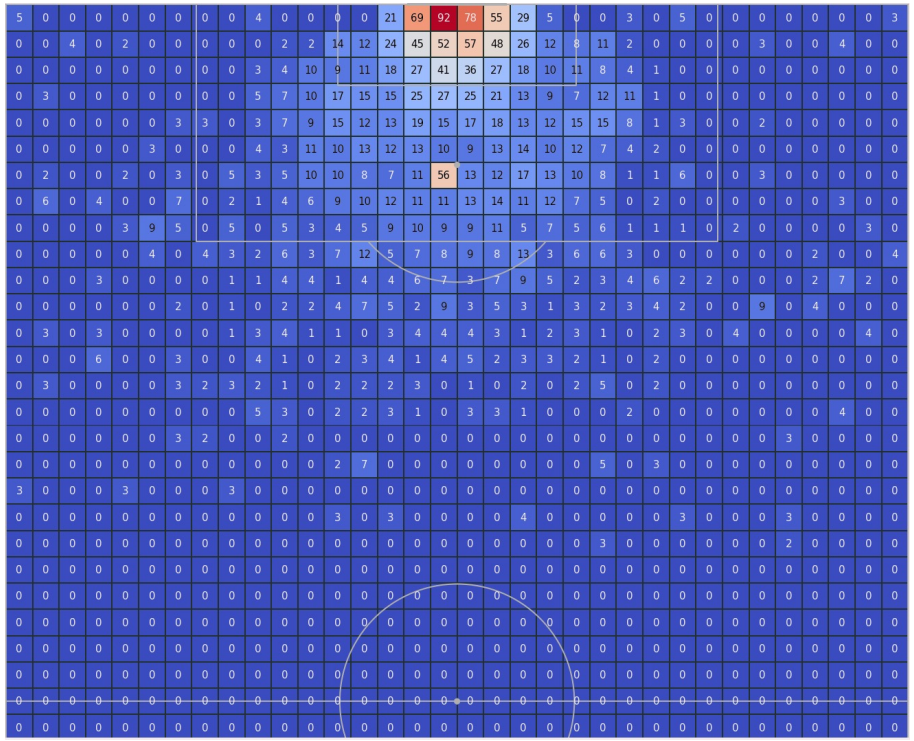
Sumpter, 2020



Synthetic shots

Produce more realistic shot success surface if players were forced to shoot from outlier location.

A player chooses not to shoot from such outlier locations because a pass or a dribble is, on average, a more valuable action.



Expected Threat (xT)

Desirable properties of a framework that rewards individual actions:

- Rewards buildup play, not just goals and assists;
- Rewards actions independent of the end outcome of the possession;
- Rewards moving the ball into a location of **increased threat**, not simply closer to goal.

Expected Threat (xT)

Desirable properties of a framework that rewards individual actions:

- Rewards buildup play, not just goals and assists;
- Rewards actions independent of the end outcome of the possession;
- Rewards moving the ball into a location of **increased threat**, not simply closer to goal.

Modelled possession sequences as **Markov chains** where players *choose* whether to move the ball or shoot to produce a *value surface* on the pitch:

- Consecutive actions that move the ball represent transient states;
- Possession turnovers and goals represent absorbing states.

Motivating a Bayesian approach

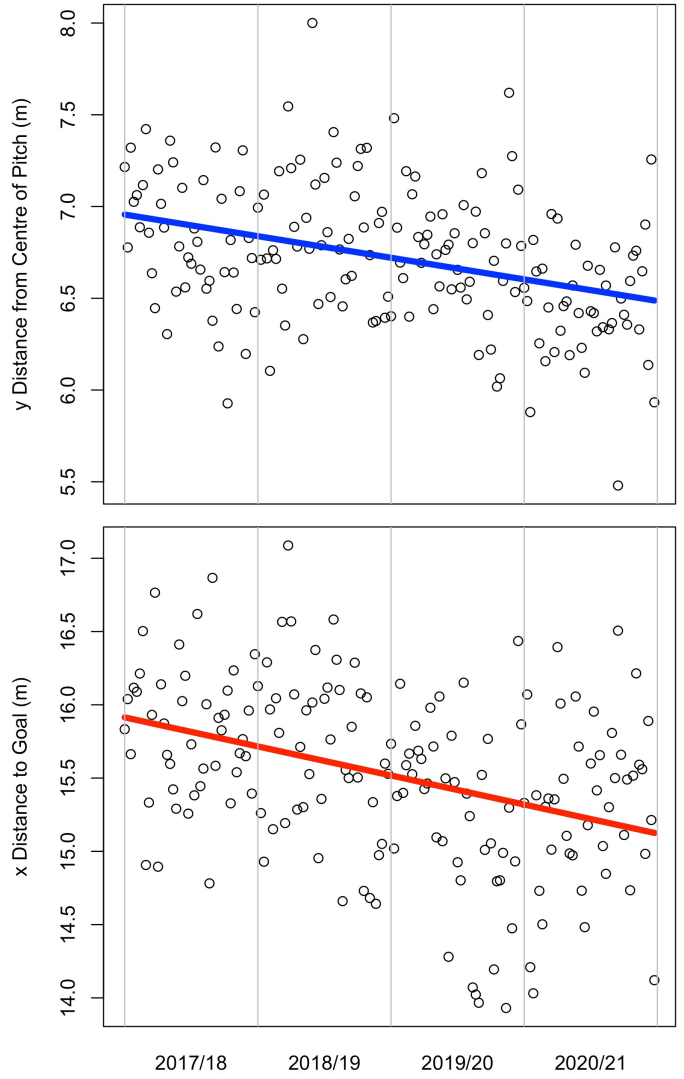
Team strategy and player behavior is evolving.

Shots are typically being taken *more central* to the middle of the pitch and *closer to goal*.

p values associated with negative gradient coefficients statistically significant at the 0.1% level of significance.

Players are “making the goal bigger” before they shoot.

Bayesian updating:
Use previous season’s data as prior and update xT value action surface monthly.



Bayesian xT model

Create xT value surface by overlaying *M* x *N* grid, of zones, *z*, reflecting the probability of scoring a goal at a later state given the current position via:

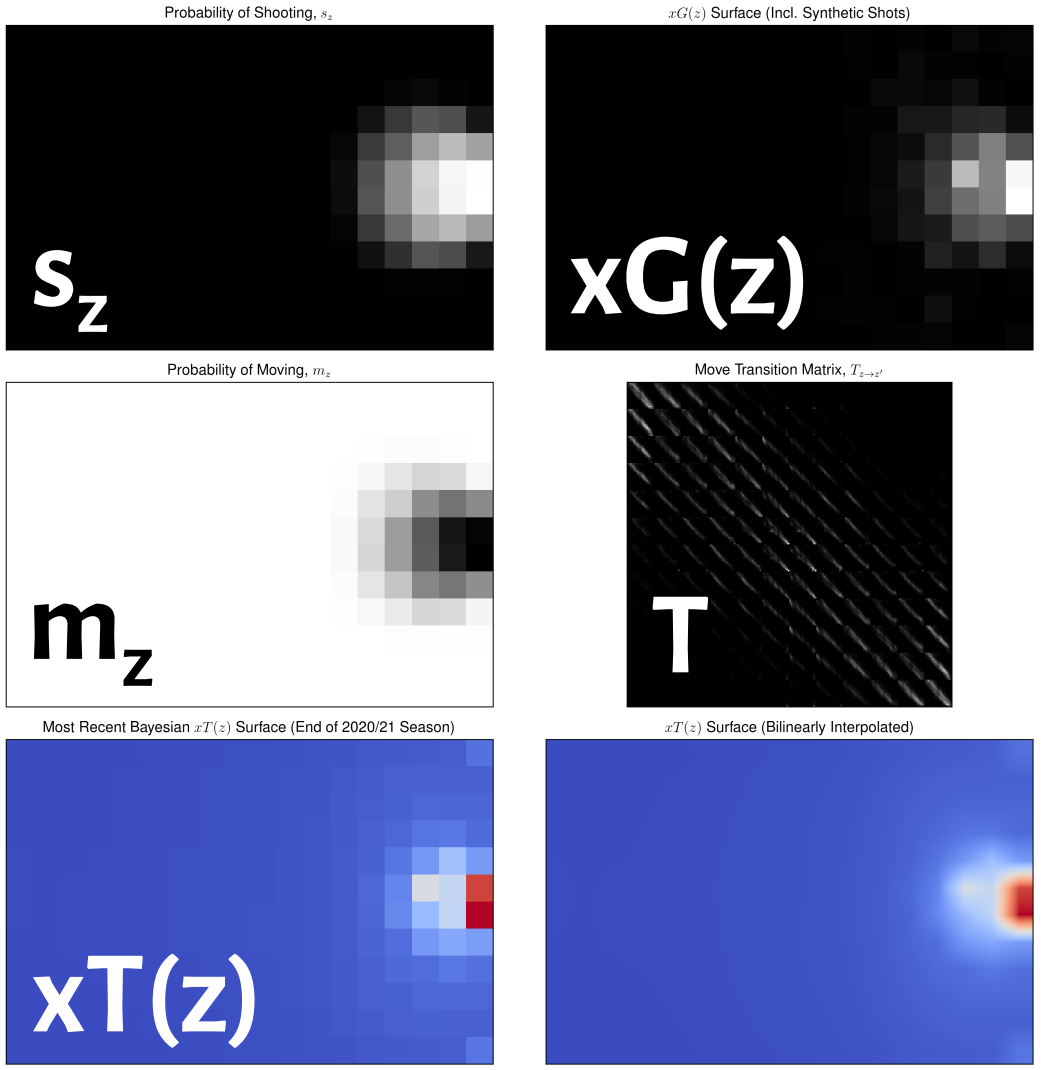
$$xT(z) = s_z \cdot xG(z) + m_z \cdot \sum_{z'=1}^{M \times N} T_{z \to z'} \cdot xT(z')$$

Actions are valued by the difference between start and finish positions on the xT surface:

$$V_{xT}(a_i) = xT(z') - xT(z)$$

Initialise *xT(z)* at 0 and solve via dynamic programming.

[Singh, 2019.]



Bayesian xT model innovations

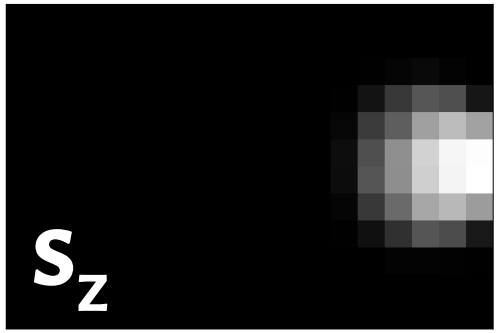
Applied bilinear interpolation to take advantage of Opta's positional granularity.

Beta-Binomial conjugate system: updating 47,304 priors each month for 12 x 18 pitch.

Encode domain expertise via synthetic shot data as part of xG prior.

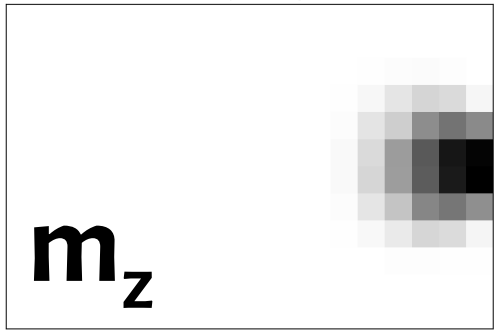
Easy to integrate data from any vendor.

Probability of Shooting, s_z

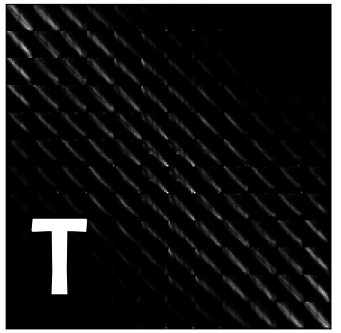


$xG(z)$ Surface (Incl. Synthetic Shots)

Probability of Moving, m_z



Move Transition Matrix, $T_{z \to z'}$



Most Recent Bayesian $xT(z)$ Surface (End of 2020/21 Season)

$xT(z)$ Surface (Bilinearly Interpolated)

Expected Goals (xG)

xT purely values actions that move the ball. Attributes zero value to shots.

Nearly 25 years after the introduction of the first logistic regression xG model, still the practitioner's choice to describe the most important features.

Expected Goals describes shot chance quality as the conditional probability of shot success given the shot situation.

Excess xG:

- Calculated as the difference between the actual outcome and forecast xG;
- Metric to value shot actions to complement xT for passes, crosses, dribbles.

Pollard & Reep, 1997.

xG feature engineering

Basic features:

- *x* distance to goal;
- *y* distance from centre of pitch.

Added features:

- Distances to goal mid-point;
- Shot angles.

Advanced features:

- Contextual features;
- Distance-angle interactions.

Feature	Coefficient Estimate	p value	Sig. Without Syn.
Intercept	-12.330	<.001	
D	0.501	<.001	×
D^2	-0.005	<.001	
Shooting Angle	2.630	<.001	
Visible Angle	17.230	.003	×
Visible Angle $\times D$	-0.898	.01	
Visible Angle $\times D^2$	0.012	.03	×
Shooting Angle $\times D$	-0.169	<.001	×
Game State	0.717	<.001	
Red Card Count	-0.314	.01	
Shot Pressure Count	-1.017	<.001	
Player Possession Time	0.006	.02	×
Team Possession Time	-0.002	.004	

Advanced model trained with and without synthetic shots.

Sumpter, 2016.

xG model log loss & calibration

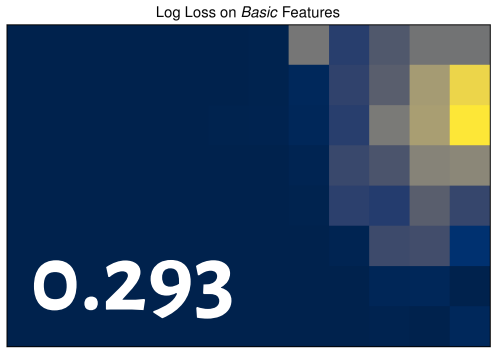
Wish to produce well-calibrated model with the smallest log loss possible.

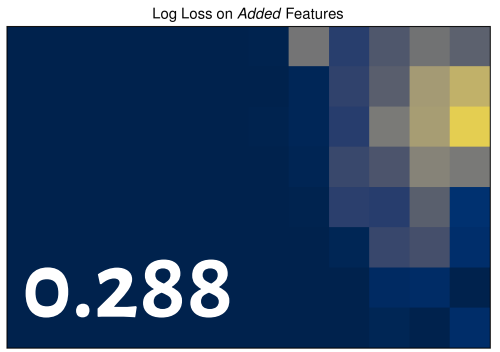
All models used the same test data.

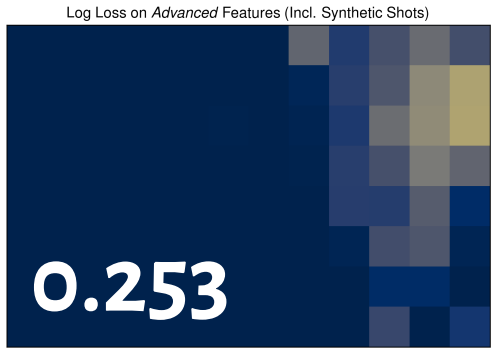
Advanced model best-calibrated and has lowest log loss Vs literature.

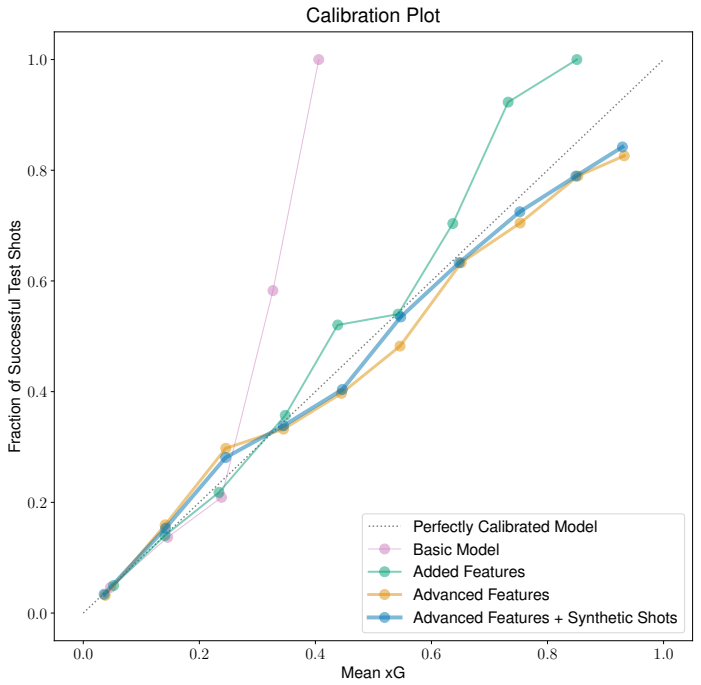
Often, best way to improve model performance is to use better data.

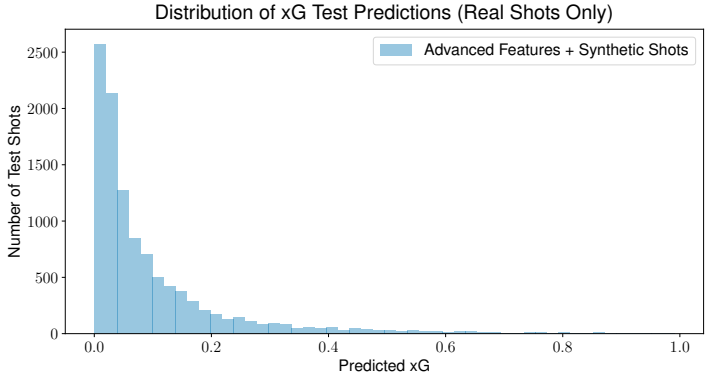
Eggels 2016, Van den Hoek 2019, Noordman 2019.











Applications

Player xT

Ranking the Premier League's most threatening players via xT per 90 mins.

2019/20 PFA Player of the Year **de Bruyne** and *2019/20 Young Player of the Year* **Alexander-Arnold** perennial generators of threat.

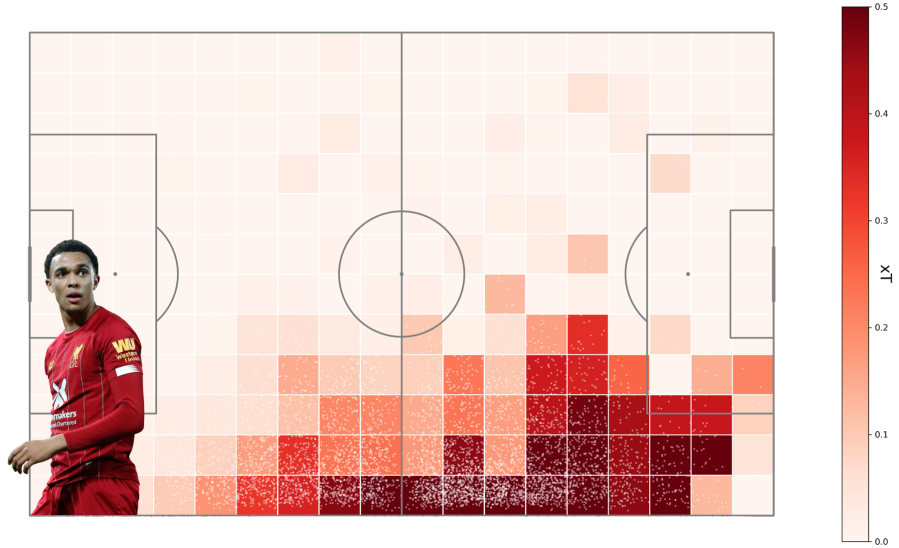
Season	Rank	Player	xT per 90	Value (£m)
2017/18	1	Philippe Coutinho (Liverpool)	0.399	121.5
2017/18	2	Chris Brunt (West Brom)	0.378	3.1
2017/18	3	Cesc Fàbregas (Chelsea)	0.335	31.5
2017/18	4	Robbie Brady (Burnley)	0.311	9.0
2017/18	5	Kevin De Bruyne (Man City)	0.304	135.0
2018/19	1	Trent Alexander-Arnold (Liverpool)	0.350	72.0
2018/19	2	James Milner (Liverpool)	0.346	13.5
2018/19	3	Ryan Fraser (Bournemouth)	0.338	27.0
2018/19	4	Pascal Groß (Brighton)	0.326	9.0
2018/19	5	James Maddison (Leicester)	0.322	36.0
2019/20	1	Kevin De Bruyne (Man City)	0.380	108.0
2019/20	2	Trent Alexander-Arnold (Liverpool)	0.376	99.0
2019/20	3	Robert Snodgrass (West Ham)	0.310	4.3
2019/20	4	Ashley Young (Man United)	0.305	2.9
2019/20	5	Pascal Groß (Brighton)	0.302	8.6
2020/21	1	Trent Alexander-Arnold (Liverpool)	0.312	67.5
2020/21	2	Luke Shaw (Man United)	0.300	37.8
2020/21	3	Kevin De Bruyne (Man City)	0.276	90.0
2020/21	4	Raphinha (Leeds)	0.273	27.0
2020/21	5	Matt Ritchie (Newcastle)	0.266	2.7

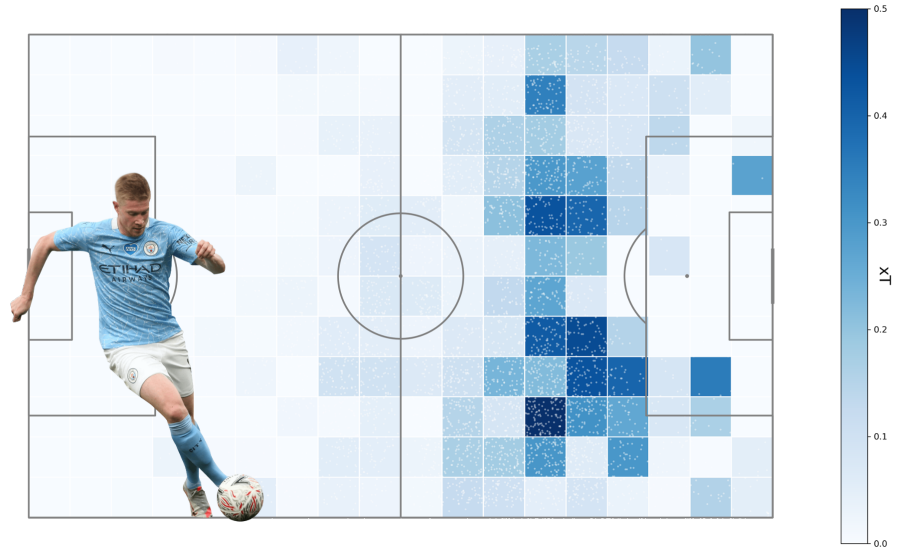
Player xT

Ranking the Premier League's most threatening players via xT per 90 mins.

2019/20 PFA Player of the Year **de Bruyne** and *2019/20 Young Player of the Year* **Alexander-Arnold** perennial generators of threat.

Alexander-Arnold generates threat in wide areas, whereas de Bruyne operates centrally.





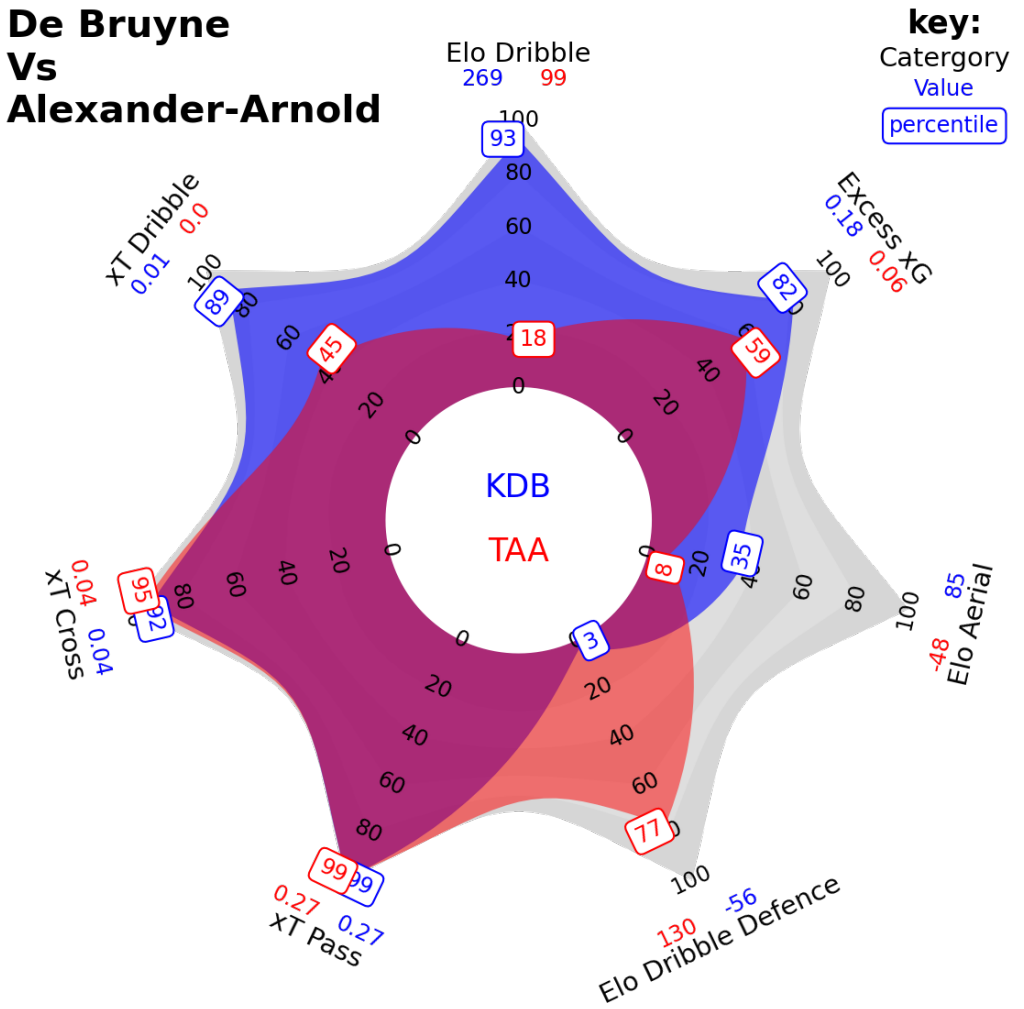
Player xT

Ranking the Premier League's most threatening players via xT per 90 mins.

2019/20 PFA Player of the Year **de Bruyne** and *2019/20 Young Player of the Year* **Alexander-Arnold** perennial generators of threat.

Alexander-Arnold generates threat in wide areas, whereas de Bruyne operates centrally.

Both exceptional at crossing and passing. De Bruyne the superior dribbler and finisher. Alexander-Arnold the superior defender.



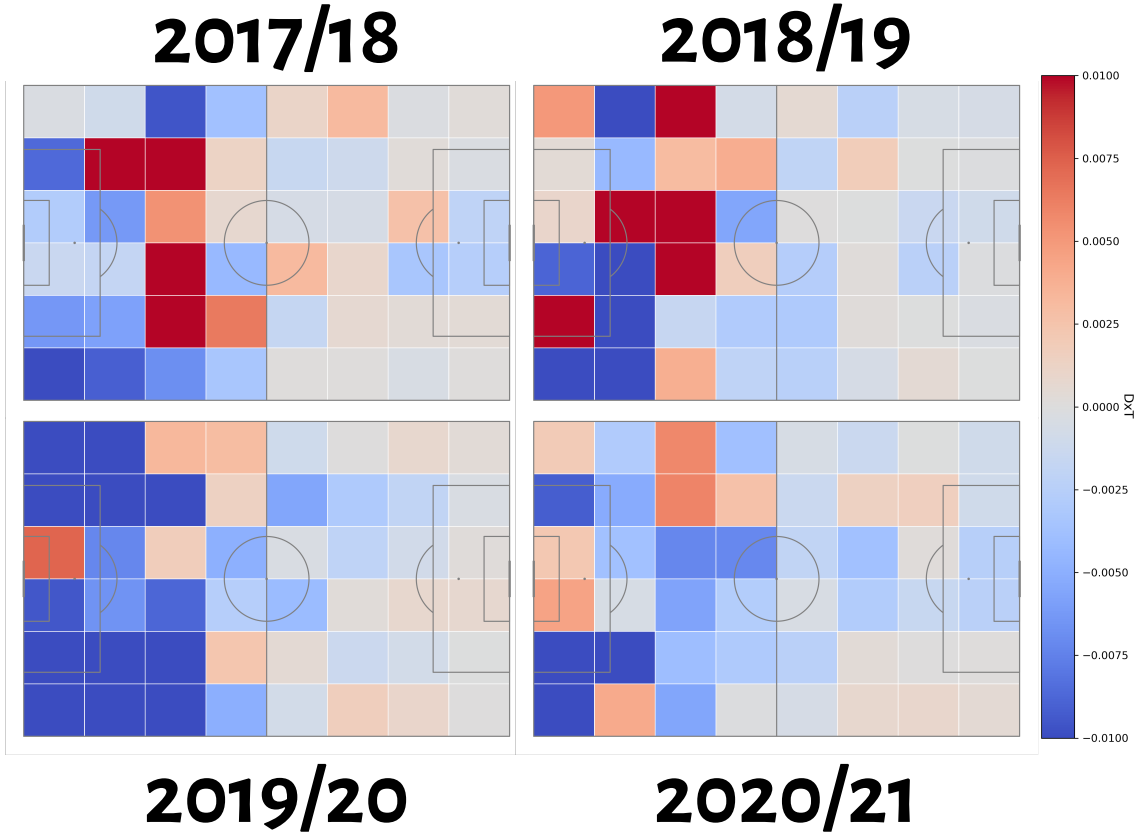
Delta xT

Legendary AC Milan defender Maldini said:
"If I have to make a tackle, I've already made a mistake"

On-the-ball events data are blind to off-the-ball context.

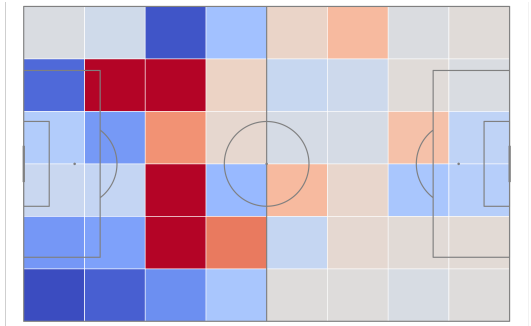
Indirectly quantify team defensive strengths and weaknesses via *Delta xT*:

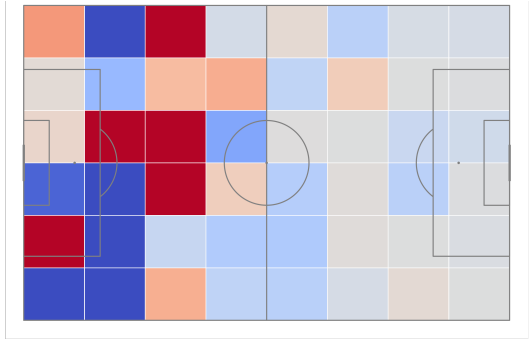
- Mean opponent threat per zone Vs. Man United;
- Mean opponent threat per zone excl. Man United;
- Mean threat deltas over all 19 opponents in EPL.



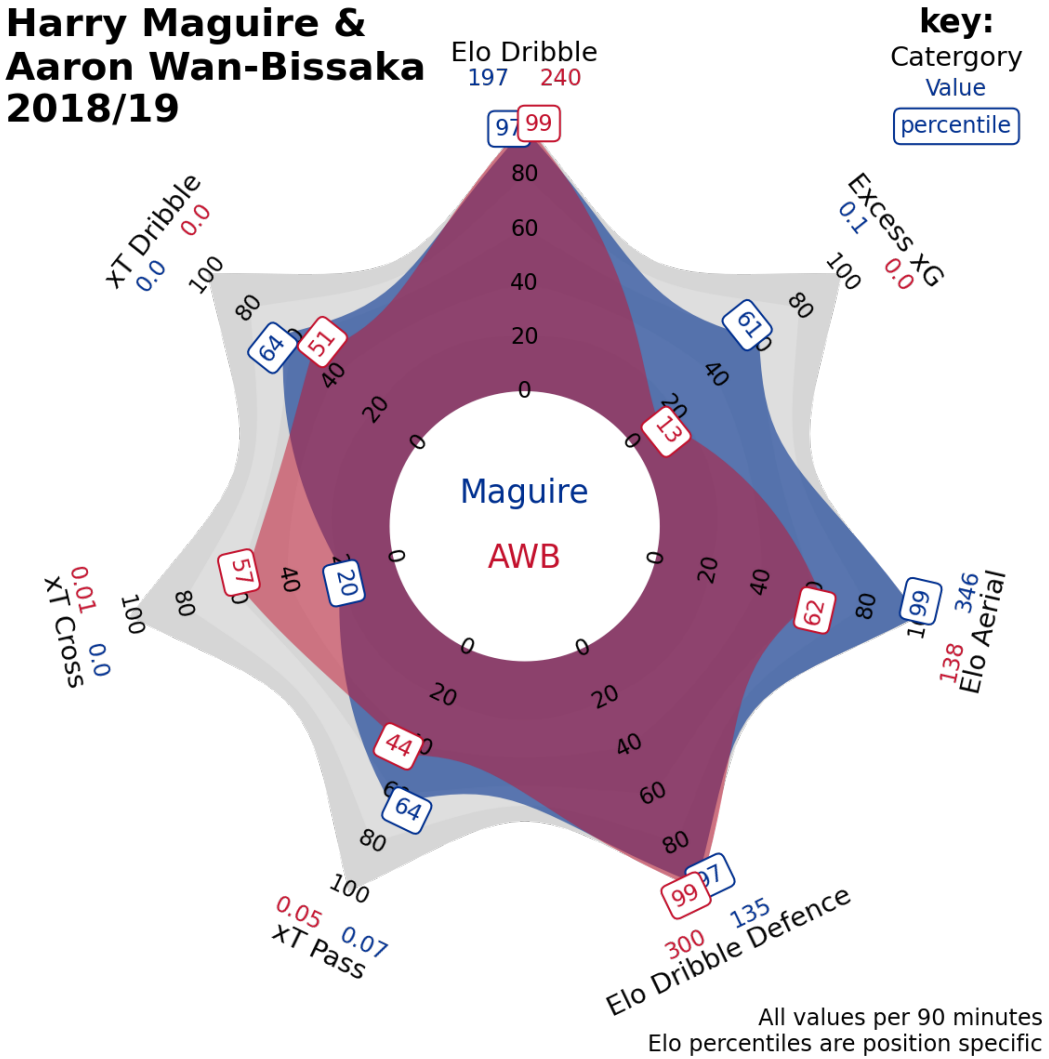
Measuring Maguire & AWB success

2017/18



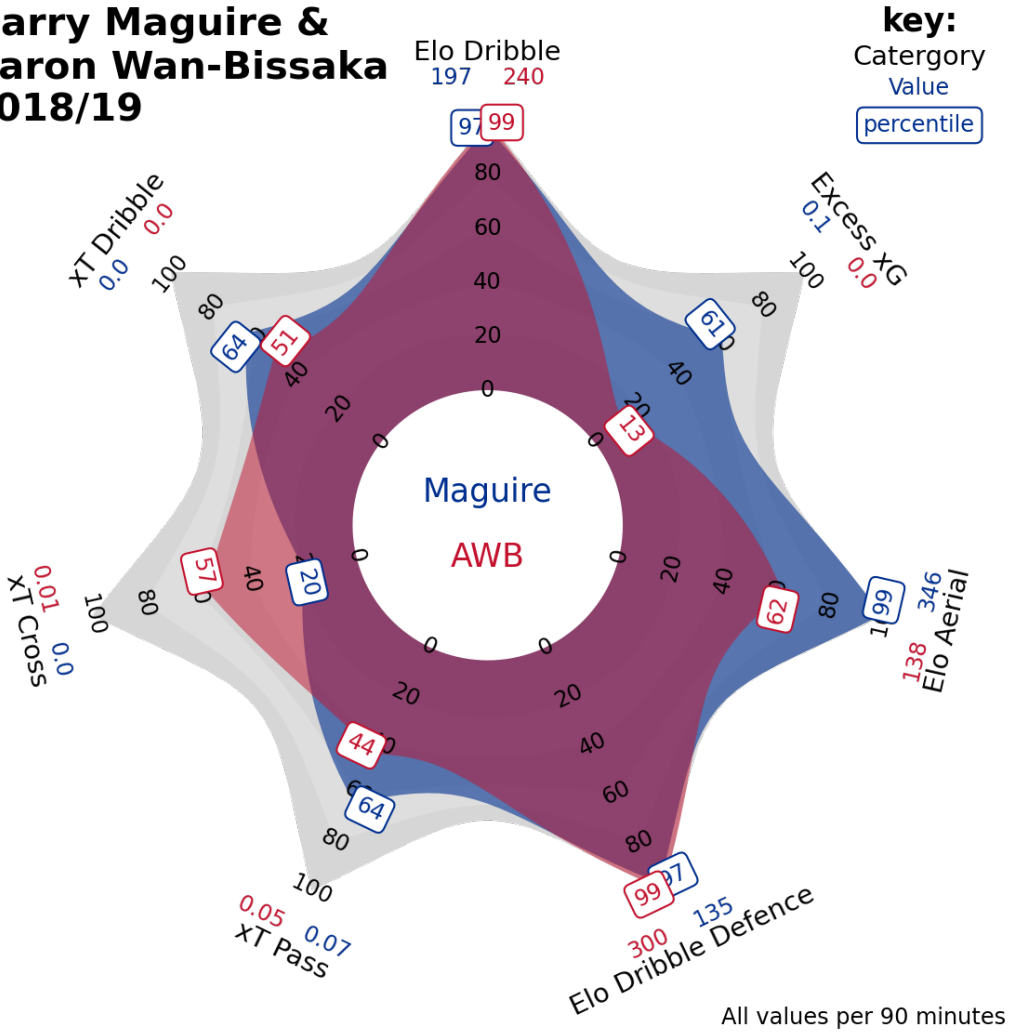


2018/19

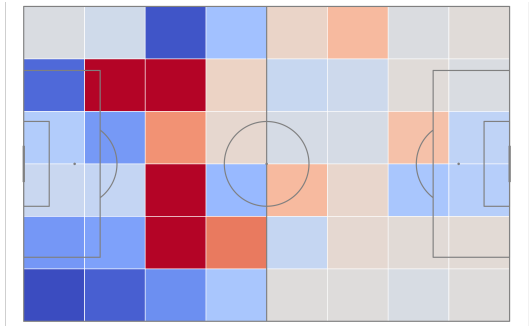


Measuring Maguire & AWB success

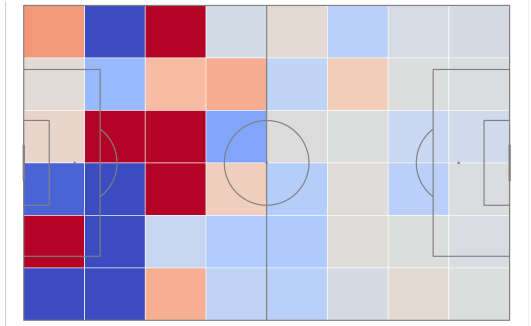
Harry Maguire & Aaron Wan-Bissaka 2018/19



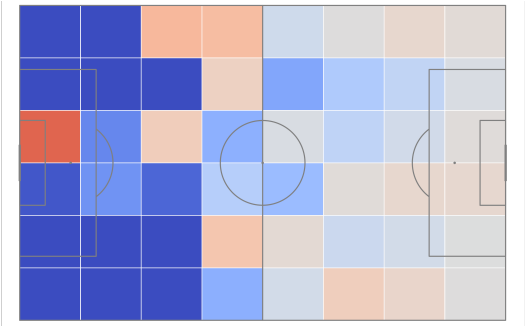
2017/18



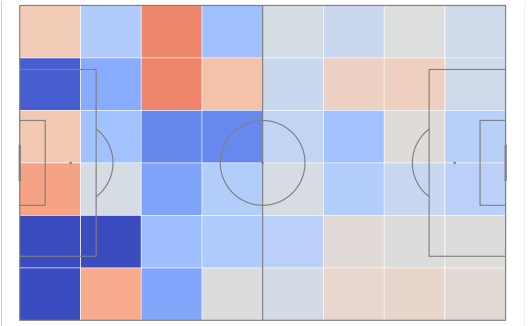
2018/19



2019/20



2020/21



All values per 90 minutes
Elo percentiles are position specific

Summary

Opta's data found to be more accurate, complete, spatially precise, and significantly more straightforward to work with than Wyscout data

Introduced framework to value individual player actions called Bayesian Expected Threat

Open-sourced software to keep value surfaces up-to-date as football evolves, integrate any data source, and encode domain expertise

Showcased practical applications for recruitment and matchday use cases

Future work

Integrate different *types* of football data – optical tracking data and training performance data – to produce more expansive, position-specific radars

Expand geographical coverage beyond the English Premier League to realise the ”Moneyball” ambition of the techniques developed

Apply Elo system to determine relative league strength to support expansion

Appendix

References

H. Eggels, Expected goals in soccer: Explaining match results using preductive analystics. 2016.

R. Noordman, Improving the estimation of outcome probabilities of football matches using in-game information. 2019.

N. Van den Hoek, Improving expected-goals models: Towards more accurate values for individual shots by considering more detailed information. 2019.

Maaike Van Roy et al., Valuing On-the-Ball Actions in Soccer: A Critical Comparison of xT and VAEP. 2019.

K. Singh, Introducing Expected Threat (xT). 2019.

R. Pollard and C. Reep, Measuring the effectiveness of playing strategies at soccer. 1997.

D. Sumpter, Soccermatics: Mathematical Adventures in the Beautiful Game. 2016

D. Sumpter, Twitter thread on synthetic shots. 2020.

I. McHale, P. Scarf and D. Folker, On the Development of a Soccer Player Performance Rating System for the English Premier League. 2012.

Paired xT

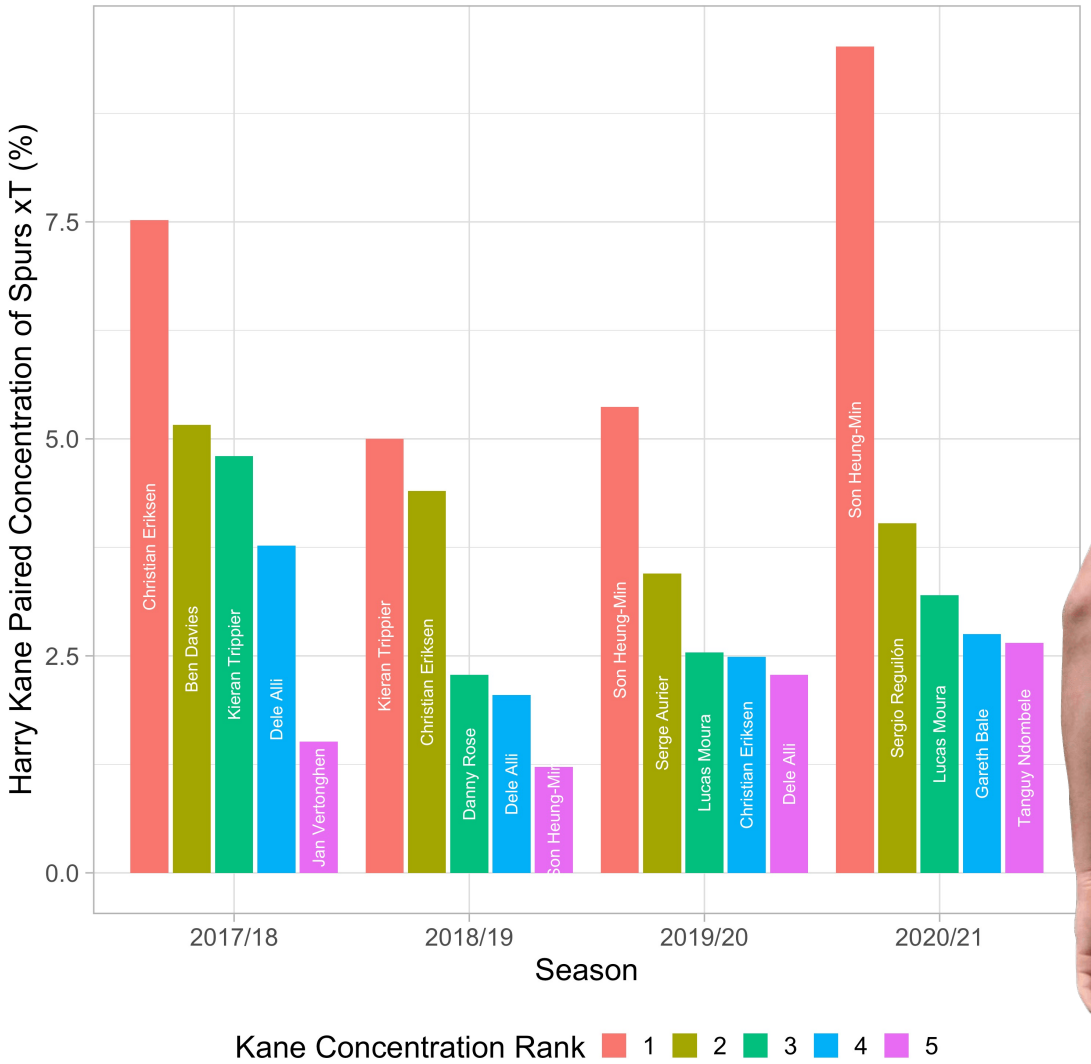
Straightforward to determine the recipient an action that moves the ball.

Identify key partnerships between players by looking at pairs with high concentrations of their team's overall xT.

Use cases:

- Strategically target partnerships by man-marking on or both players or look to cut the supply between pair.
- Positions of fragility within Man United squad.

Paired xT: Concentration



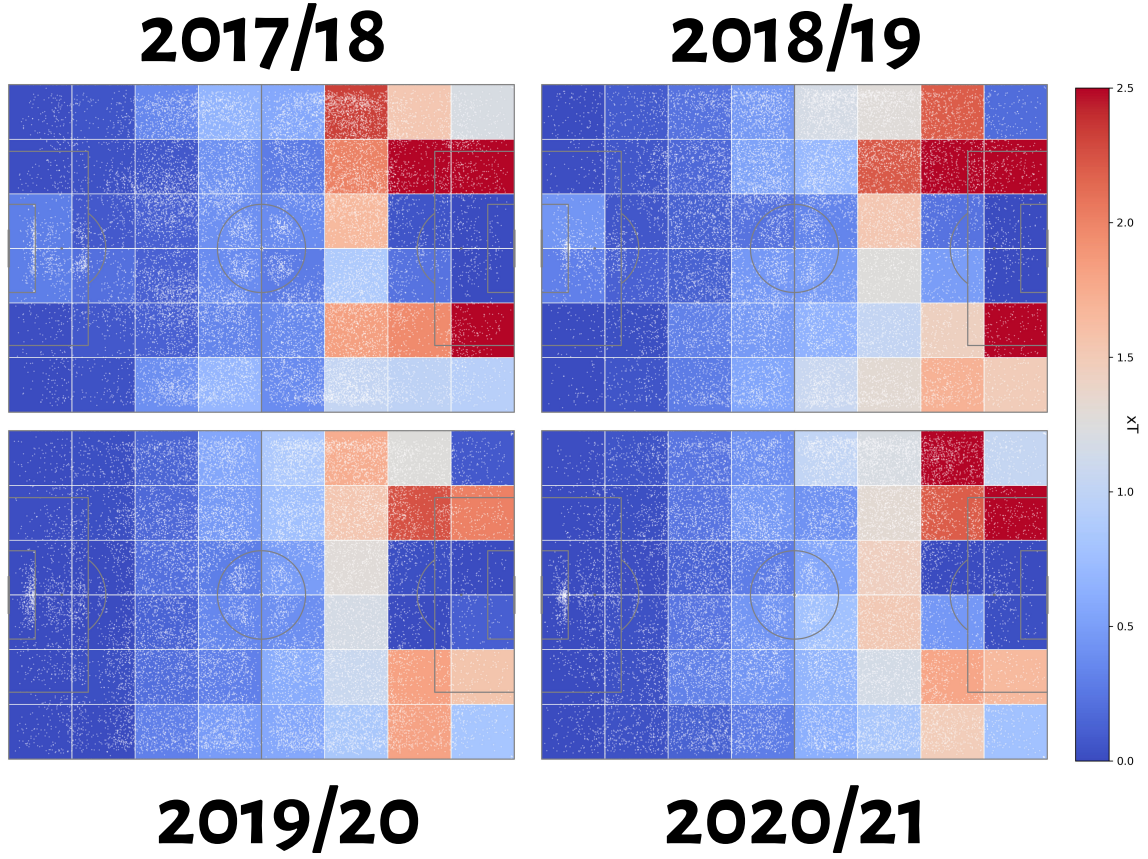
Team xT

Aggregate xT values *per team* per season, highlighting strategically focused threat.

Man United signature to build attacks in wide areas – core tenet of Sir Alex's success.

Becoming increasingly dependent and predictable on left wing threat.

Marcus Rashford (LW) played almost every minute of 2020/21, whilst minutes were shared between Greenwood (RW) and James (RW) on opposite flank.

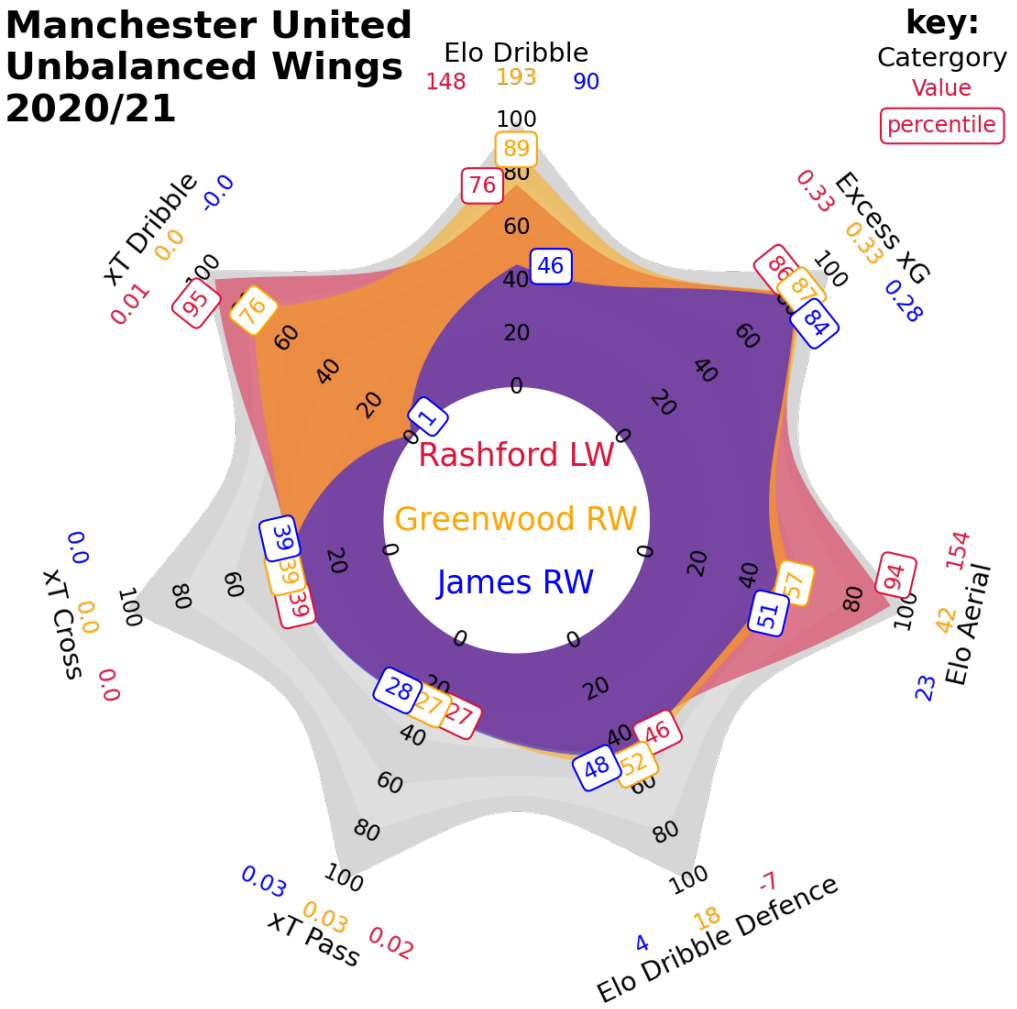


Team xT

Rashford and Greenwood exceptional dribblers and finishers, with Rashford also brilliant in the air.

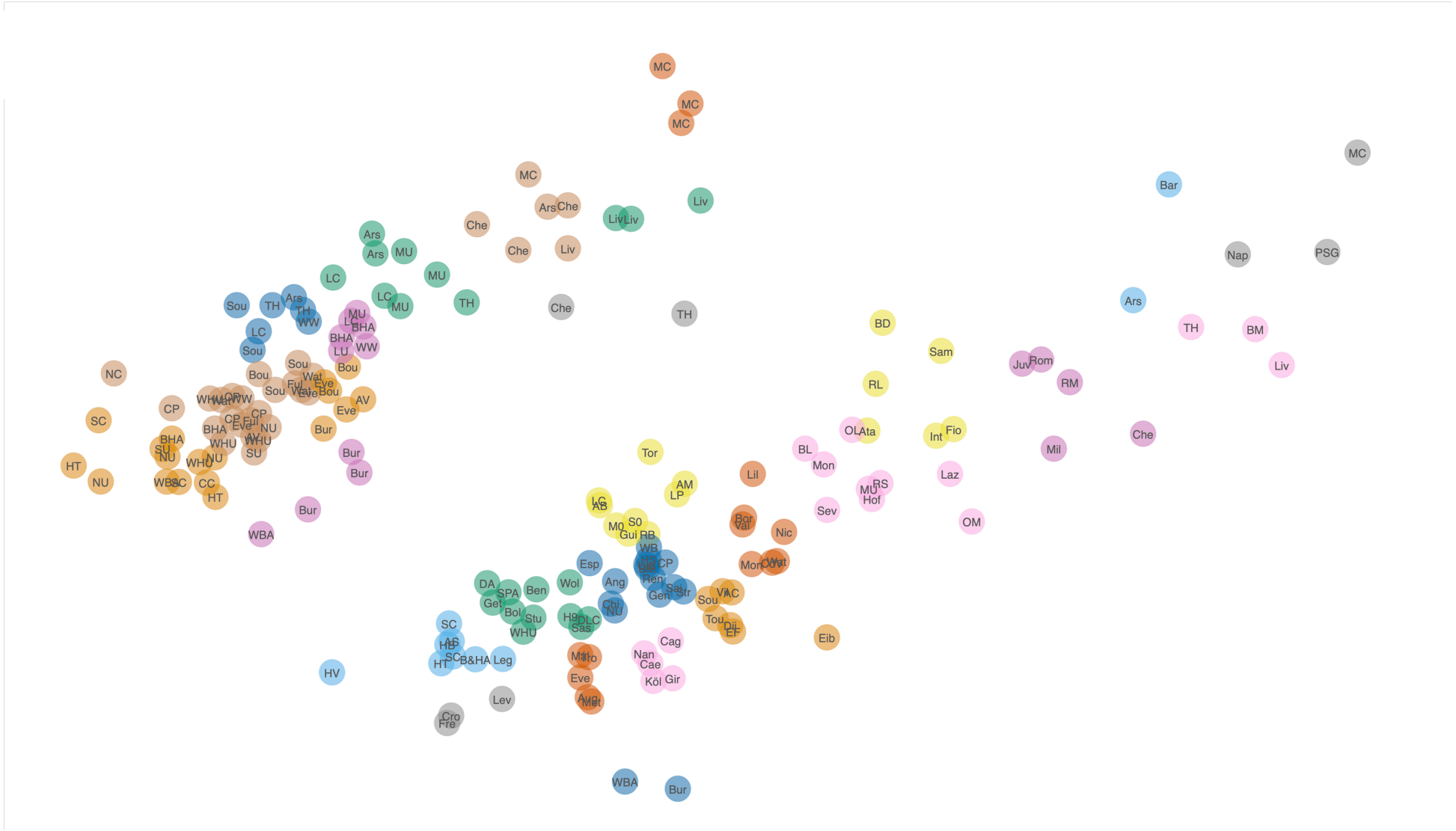
Team's aggregate weakness on the left-wing caused by James' abysmal dribbling.

Recommend renewing interest in previously scouted Sancho, selling James, and looking for Rashford backup.

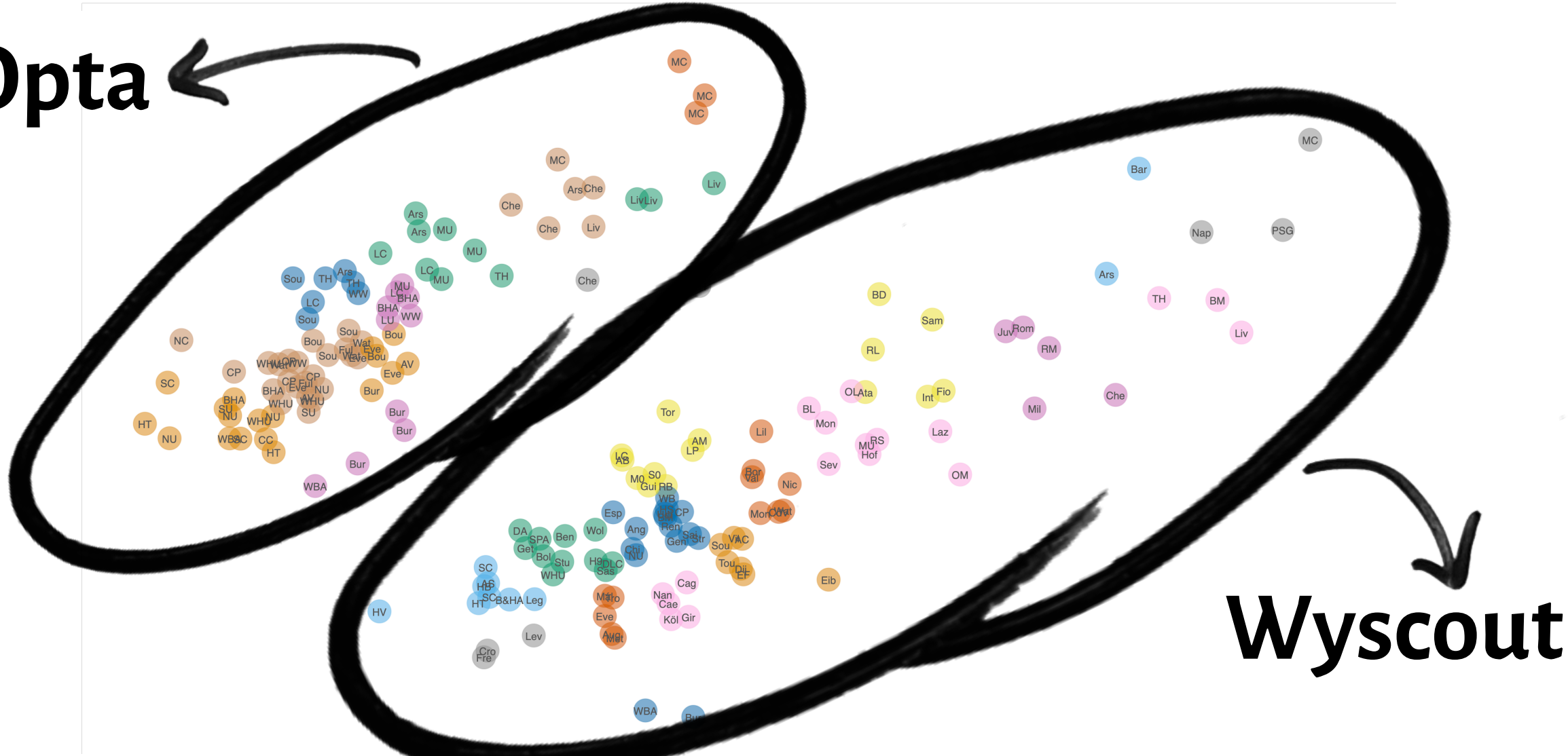


All values per 90 minutes

Team similarity clustering



Team similarity clustering



Expected Threat (xT)

Desirable properties of a framework that rewards individual actions:

- Rewards buildup play, not just goals and assists;
- Rewards actions independent of the end outcome of the possession;
- Rewards moving the ball into a location of **increased threat**, not simply closer to goal.

Modelled possession sequences as Markov chains where players *choose* whether to move the ball or shoot to produce a *value surface* on the pitch:

- Consecutive actions that move the ball represent transient states;
- Possession turnovers and goals represent absorbing states;
- xT more sophisticated than count-based regression approaches (EA Index);
- xT more intuitive and interpretable than machine learning approaches (VAEP).

I. McHale, P. Scarf and D. Folker 2012, Maaike Van Roy et al. 2019

Opta Vs Wyscout: Summary

Opta's data found to be more accurate, complete, spatially precise and significantly more straightforward to work with.

Wyscout's data contained two detrimental sources of bias:

1. Look-ahead bias stemming from deflected shots being disproportionately omitted from the dataset;
2. Main denominator used to normalise metrics biased low by unknown amount.

Opta data was almost exclusively used for following modelling and applications unless explicitly specified.