



Edd Webster

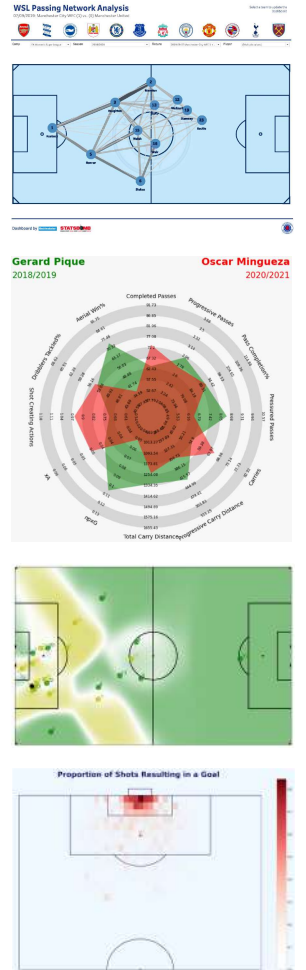
Football Data Science Pack

December 2021

Contents

- [Introduction](#) (slides 3-5)
- [The state of play in football analytics](#) (slides 6-13)
- [Why the application of data in football is important](#) (slides 14-19)
- [Sample football data analysis projects](#) (slides 20-25):
 - 1) [Tableau Football Intelligence reporting tools created using StatsBomb data](#) [[Tableau Public profile](#)] (slides 26-42)
 - 2) [Recruitment Analysis of 'Ball-playing' Center Backs](#) [[code](#)] [[Tableau dashboard](#)] (slides 43-70)
 - 3) [Applications of Tracking data](#) [[code](#)] (slides 71-81)
 - 4) [Expected Goals \(xG\) modeling using Event and Tracking data](#) [[code](#)] (slides 82-106)
- [Conclusion](#) (slides 107-110)
- [Further work](#) (slides 111-112)
- [References](#) (slides 113-118)

All code and data for the analysis featured in this pack: github.com/eddwebster/football_analytics and Tableau visualisations: public.tableau.com/profile/edd.webster.



Introduction

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

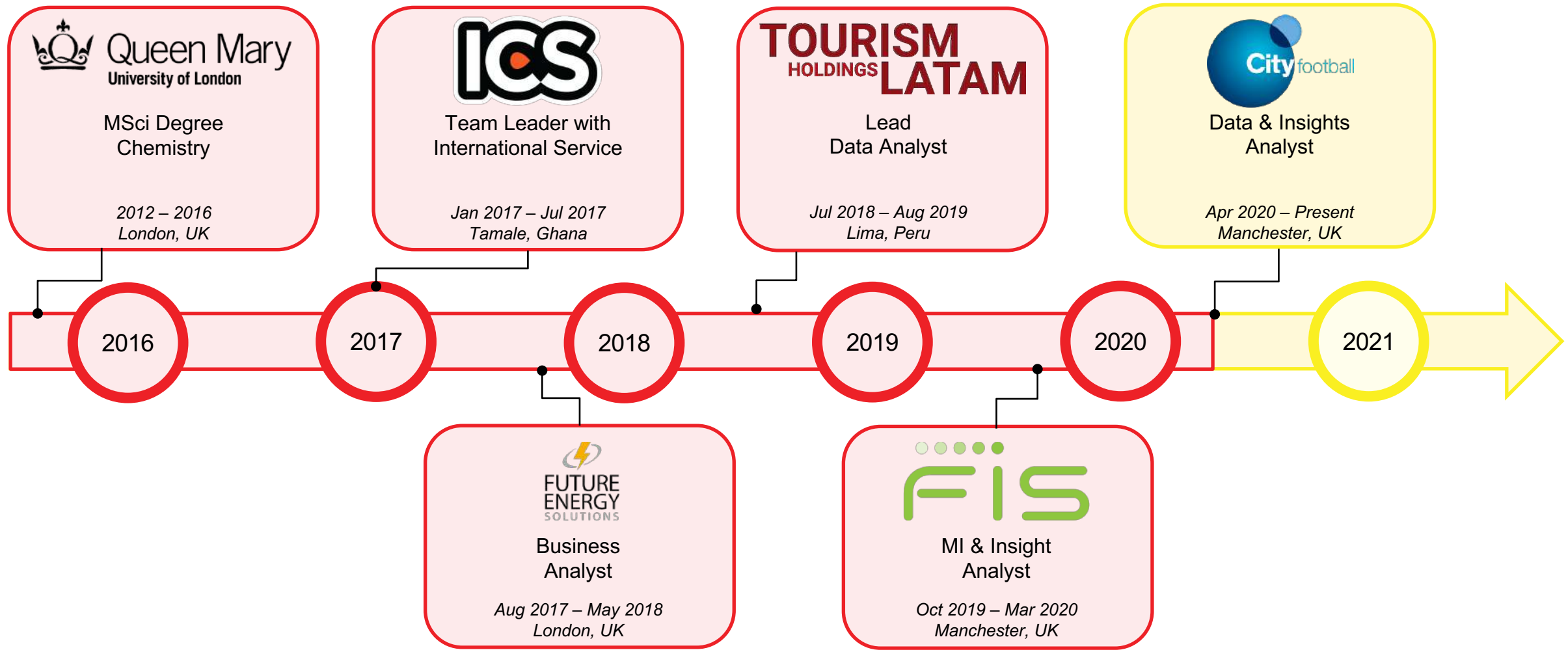
Conclusion

Further Work



About Me

Timeline of my background and education over the last nine years



For more information, see: [linkedin.com/in/eddwebster](https://www.linkedin.com/in/eddwebster).

The reason for producing this football data science pack?

To demonstrate the value I can bring to Watford Football Club as a Data Scientist, through my data analysis skills, football analytics knowledge, willingness to learn, and passion for the implementation of data to bring insight to football.

The State of Play in Football Analytics

A brief overview of the evolution and progression of data-driven analysis in football, to provide a background to the projects explored in the deck

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

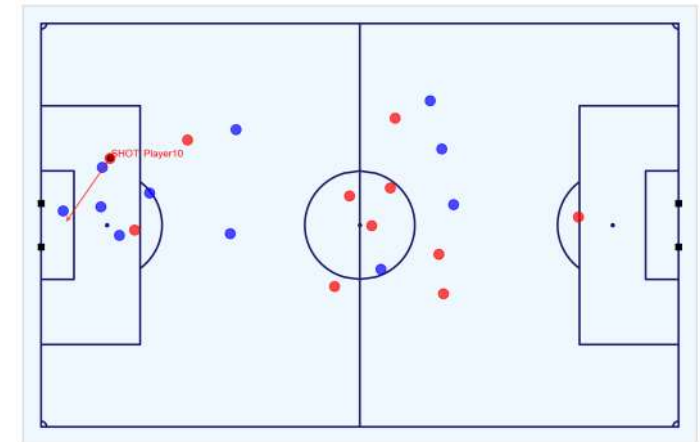
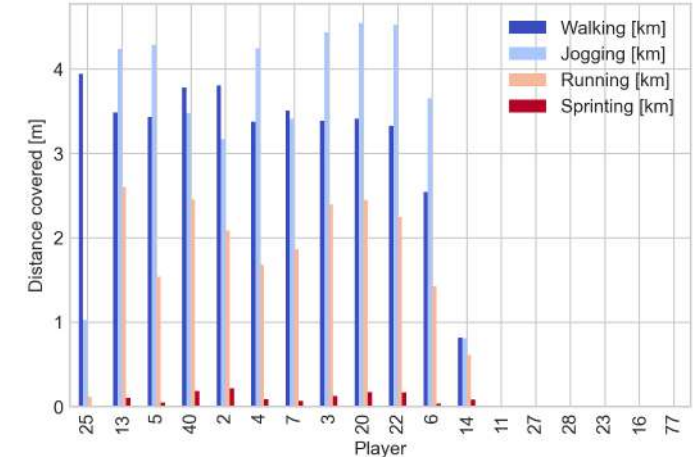
Further Work



Data Collection

Football clubs need to work with a variety of different data sources and types for their players

- 1 Physical** – total distance, accelerations, sprints/speed runs, heartrate, agility, power, jump height, strength, velocity band (duration).
- 2 Event and positional data** – shots, passes, tackles, goalkeeper dives, ball possession, duels, interceptions, X & Y coordinates.
- 3 Medical and nutrition data** – weight, injuries, nutrition intake, growth, vision and eye movement, muscle soreness, sleep activity, rate of perceived exhaustion.
- 4 Personal data** – address, contact details, passport ID.



The Evolution of Data in Football Analytics

Progress observed in football analytics from the three 'flavours' of match data

Data freely available for all professional matches

Commercially available for professional matches

Proprietary, available for a single team or teams within the same league

High
availability

Pre 1995

Matchsheet Data

Basic, aggregated stats

Examples: # goals, shots, cards

$O(10)$ statistics per match



Matchsheet data for Brazil vs Belgium. The table shows goals, assists, and cards for both teams.

Brazil	1
Belgium	2
Goals:	13' Fernandinho (OG) 0-1, 31' De Bruyne 0-2, 76' Renato Augusto 1-2
Brazil:	Alisson, Fagner, Silva, Miranda, Marcelo, Fernandinho, Paulinho (73' Renato Augusto), Coutinho, Willian (46' Fernandinho), Neymar, Jesus (86' Costa)
Belgium:	Courtois, Meunier, Alderweireld, Kompany, Verthongen, Witsel, Fellaini, Chadli (83' Vermaelen), De Bruyne, Hazard, Lukaku (87' Tielemans)
Yellow cards:	47' Alderweireld, 71' Meunier, 85' Fernandinho, 90' Fagner
Red cards:	None

Low
granularity

High-level summary

1995+

On-the-Ball Event Stream Data

Log of each on-ball event
(passes/tackles/shots)

Examples: Event type, timestamp, spatial location,
meta information of on-the-ball actions

$O(10^3)$ events per match

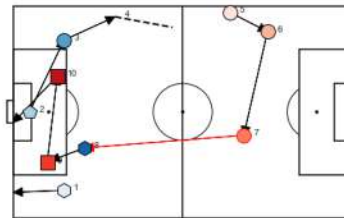


Table illustrating on-the-ball event stream data. It shows a log of each on-ball event (passes/tackles/shots) with columns for time, actiontype, player, and team.

	time	actiontype	player	team
1	16a33s	interception	M. Leckie	Australia
2	16a46s	goalkick	M. Ryan	Australia
3	16a47s	pass	M. Milligan	Australia
4	16a49s	dribble	A. Behuch	Australia
5	16a55s	pass	L. Advincula	Peru
6	16a59s	pass	C. Ramos	Peru
7	17a3s	pass	Y. Yotón	Peru
8	17a6s	interception	T. Sainsbury	Australia
9	17a11s	cross	P. Guerrero	Peru
10	17a12s	shot	A. Carrillo	Peru

Description of all on-the-ball events

2005+

Tracking Data

Player & ball positions
Samples at 25HZ

Examples: X, Y coordinates of the 22 players, 3
referees and the ball at every time-step

$O(10^6)$ observations per match



High
granularity

Exact movements of all players and the ball

Slide recreated using [Laurie Shaw](#)'s seminar at the Sports Analytics Lab at Harvard University/American Statistical Association Section on Statistics in Sports on 23rd October 2020 titled: 'Routine Inspection: Measuring Playbooks for Corner Kicks' ([youtube.com/watch?v=yfPC1O_g-l8](https://www.youtube.com/watch?v=yfPC1O_g-l8)) and [Pieter Robberechts](#)' visualisation used by [Lotte Bransen](#)'s and [Jan Van Haaren](#)'s Friends of Tracking presentation of VAEP frameworks streamed 7th May 2020 ([youtube.com/watch?v=w0LX-2UgyXU&t=607s](https://www.youtube.com/watch?v=w0LX-2UgyXU&t=607s)).

Source of Event Data

How is Event data collected?

On-the-Ball Event data for football matches is collected by tagging each event that takes place on the pitch, i.e. passes, tackles, aerial-duels, shots, with a timestamp, the player involved, and location on the pitch with X, Y coordinates.

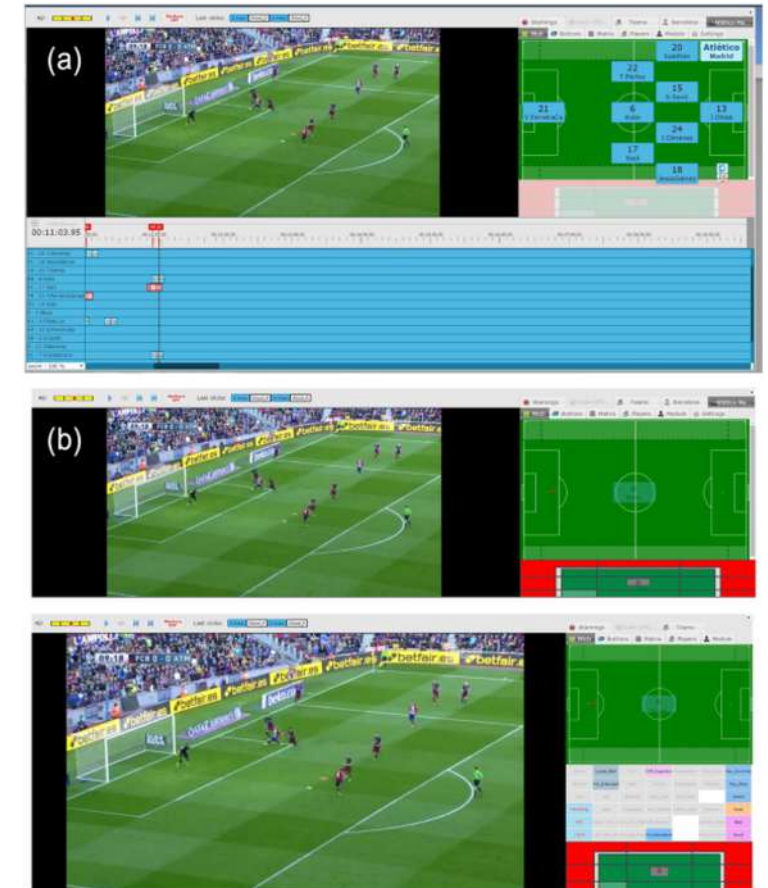
These events are collected manually from match video and/or in the stadium.

In the screenshot on the right, the process of tagging the football events from a match video is explained:

- Screenshot of the tagging software used. An action is tagged by an operator *via* a special custom keyboard, creating a new event on the match timeline.
- When the event position on the pitch is set, the shot specific input module appears (top). Event related input modules also appear for setting additional attributes of the occurring event (bottom).

For more information about how to collect Event data and the paper for which this data was made available, please see the following:

- **A public data set of spatio-temporal match events in soccer competitions.** Scientific Data 6, 236 (2019) doi:10.1038/s41597-019-0247-7, www.nature.com/articles/s41597-019-0247-7;
- Opta Sports Analytics - Behind the scenes and interviews with the Sports Data Mavens: youtube.com/watch?v=ySIAzS8Oouw.



Slide recreated from the football event tagging section of the paper [Luca Pappalardo et. al: A public data set of spatio-temporal match events in soccer competitions](https://www.nature.com/articles/s41597-019-0247-7). Scientific Data 6, 236 (2019) doi:10.1038/s41597-019-0247-7, www.nature.com/articles/s41597-019-0247-7;

Sources of Tracking Data

Several ways and types of Tracking data that can be collected

Tracking data can be collected in three different ways:

1) In Stadium Optical Tracking:

- Cameras installed in the stadium at different angles.
- Advantages: Excellent coverage of entire pitch; robust to most occlusion.
- Disadvantages: Expensive, usually limited to own league.



2) GPS Tracking:

- Players wear GPS devices to track location at all times.
- Advantages: Very accurate physical data.
- Disadvantages: Only able to have data on your own team; not all players wear the devices.



3) Broadcast Tracking / Single Camera Tracking:

- Optical tracking collected from broadcast feeds.
- Advantages: Most inexpensive, widest coverage (any game on TV), able to acquire historical games.
- Disadvantages: Limited coverage of pitch production decisions affect data.



Slide recreated [Sam Gregory](#) and [Devin Pleuler](#)'s Sport Logiq webinar on 4th November 2019 'Demystifying Tracking Data'. See the following link for this seminar on YouTube: youtube.com/watch?v=miEWHSTYvX4.

Advantages and Challenges of Event data

Pros and cons of the application of Event data for the analysis of football matches

Advantages

- Widely available for hundred of competitions from a variety of providers – [StatsBomb](#), [WyScout](#), [Opta](#).
- Increasingly information rich – new metrics included each season i.e. StatsBomb 'Freeze Frame' metric and their recently announced 360 update [\[link\]](#) that will collect the position of all players on camera for every event recorded.
- Publicly available Event data has been available for a few years, leading to a number of libraries and repositories including [socceraction](#) (Python) by [SciSports](#) and [ggsoccer](#) (R) by [Ben Torvaney](#).
- Easier to process and analyse than Tracking data. Only one event at a time and non-continuous. Size of the data is just 10^3 rows.
- Easy to connect Event data to video e.g. drag and drop an Opta F24 XML feed to [SportsCode](#) and query the video using Event data.

Challenges and Limitations

- Different providers have different interpretations of actions and accuracies of moments analysed. Subsequently, these datasets require different treatments. Libraries are available to transform data to a uniform standard such as SciSports [SPADL](#) format.
- Event stream is not always optimised for analysis.
- Event data is missing a lot of contextual information when compared with Tracking data as it only accounts for on-the-ball actions. Fewer than 1% of the on-the-ball actions in a football match are shots and players only have the ball 3 minutes on average (Johan Cruyff)¹. Event data is unable to answer the questions as to what the players are doing during the 87 minutes when you do not have the ball.
- Not always pleasant to work with in its native form e.g. nested JSON dictionaries.

¹Johan Cruyff quote features in both [Javier Fernández](#) and [Luke Bornn](#)'s paper 'Wide Open Spaces: A statistical technique for measuring space creation in professional soccer'.

Advantages and challenges of Tracking data discussed in [Sam Gregory](#) and [Devin Pleuler](#)'s Sport Logiq webinar on 4th November 2019 'Demystifying Tracking Data'. See the following link for this seminar on YouTube: youtube.com/watch?v=miEWHSTYvX4.

Advantages and Challenges of Tracking data

Pros and cons of the application of Tracking data for the analysis of football matches

Advantages and Possibilities

(many still to be discovered!)

- A much fuller dataset, including off-the-ball information.
- Can take into account velocities, accelerations, and variables that provide continuous context, not just over a single frame e.g. a player's off-the-ball run; the ball trajectory on a shot etc.
- Data can provide tactical context from coordinate positions e.g. how many times is player X unmarked; where is the fullback when a transition occurs; provide another example of when event Y occurred, etc.
- Tracking data can be used to further enrich corresponding Event data to add further detail and specificity. In the recent [StatsBomb 360](#) launch, Event data will now include the positions of all players for a given event. Other example of data enrichment include through the use of computer vision and Tracking data include:
 - Is a pass under pressure? – distance based logic based on player in possession;
 - What passing options did the player in possession have – can be calculated geometrically or more simply through rules-based logic such as no. teammates and opponents in front or behind a player; Improve on existing metrics such as 'Packing'¹, which is currently manually collected; and
 - Defender and goalkeeper positions; percent of the goal covered by the keeper, etc.

Challenges

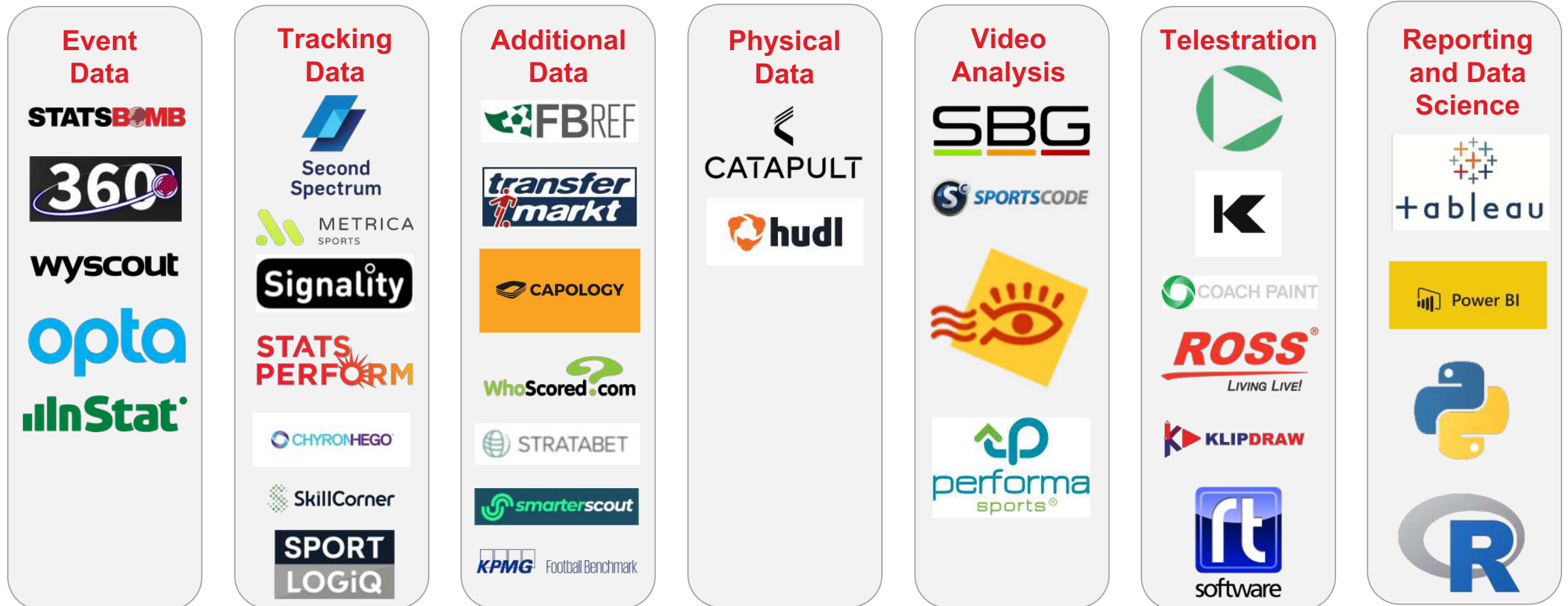
- Size of the data – Event data is 10^3 rows. Tracking data is 10^6 rows i.e. 95 mins (5,700 seconds) x 26 objects (22 players + 3 referees + ball) x 25 frames per second \approx 3.7mil observations. This brings issues of both storing and processing the data.
- Mathematically more complex to work with.
- Despite being the state-of-the-art dataset, Tracking data is still missing significant datapoints including: body pose position (i.e. are they open to receive a pass), the spin of the ball (therefore need to build in uncertainties when determining ball trajectory).
- Limited public work - recently changed with the Friends of Tracking initiative + public datasets released by [Metrica Sports](#), [Signalify](#), and [SkillCorner](#). However, nearly all public initiatives are presented in journals or conferences and are therefore less accessible and aimed at a different audience to most football practitioners.
- Difficult to merge Tracking and Event data. Event data has imperfect timestamps. This is possible through a supervised modelling approach (requires properly annotated dataset) or a rules based logic approach (less accurate but easier to implement).
- Difficult to query Tracking data as the data is continuous and the unlike the Event data, the rows continue simultaneously with 26 things at the same time i.e. it's only possible to show all the passes if you've previous defined them.

¹Packing' is a metric created by IMPECT that assigns a value to the number of opposition players taken out of the game by a pass or dribble.

Advantages and challenges of Tracking data discussed in [Sam Gregory](#) and [Devin Pleuler](#)'s Sport Logiq webinar on 4th November 2019 'Demystifying Tracking Data'. See the following link for this seminar on YouTube: youtube.com/watch?v=miEWHSTYvX4.

Data and Analytical Tool Providers

Just some of the best-in-class data sources and products used by a professional football analysis department



Note: in some examples, the products fall into multiple categories e.g. Hudl provides physical data and video analysis tools. Metrica Sports provide Tracking and Event data as well as telestration software - Metrica Play.

Why the Application of Data in Football is Important?

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

Further Work



The purpose of data analysis in football...

To help **win** football matches

Areas of a Football Club that Data Analysis Can Have an Impact

By embracing data, football practitioners can develop intelligence and processes in the following key areas...

- 1 **Recruitment and scouting** – using models to bring a positive impact on player trading outcomes in the transfer market;
- 2 **Performance evaluation and player development** – measure the quality, performance and value of players through models;
- 3 **Sport science and athlete monitoring** – provide reviews, audits, and recommendations for interventions; and
- 4 **Club strategy** – discovery of new insights through physics-based modeling that can be implemented by coaches in team strategy and affect performances on the pitch.

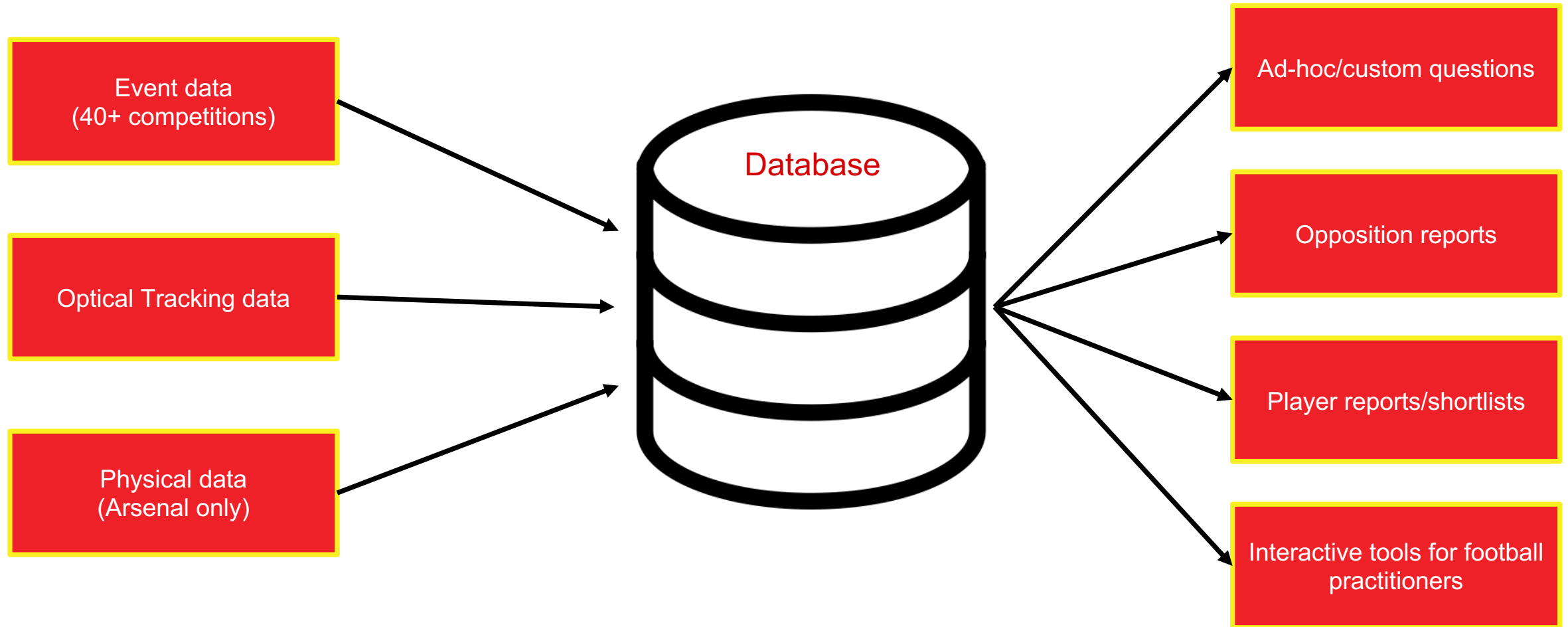
Goals of the Analytics Department

The key objectives for all those working with data at the club...

- 1 Every single decision the club makes in football operations is informed by data;
- 2 Learn more about the team, the league, and the sport of football in general using data; and
- 3 Educate decision makers and empower them to want to use data.

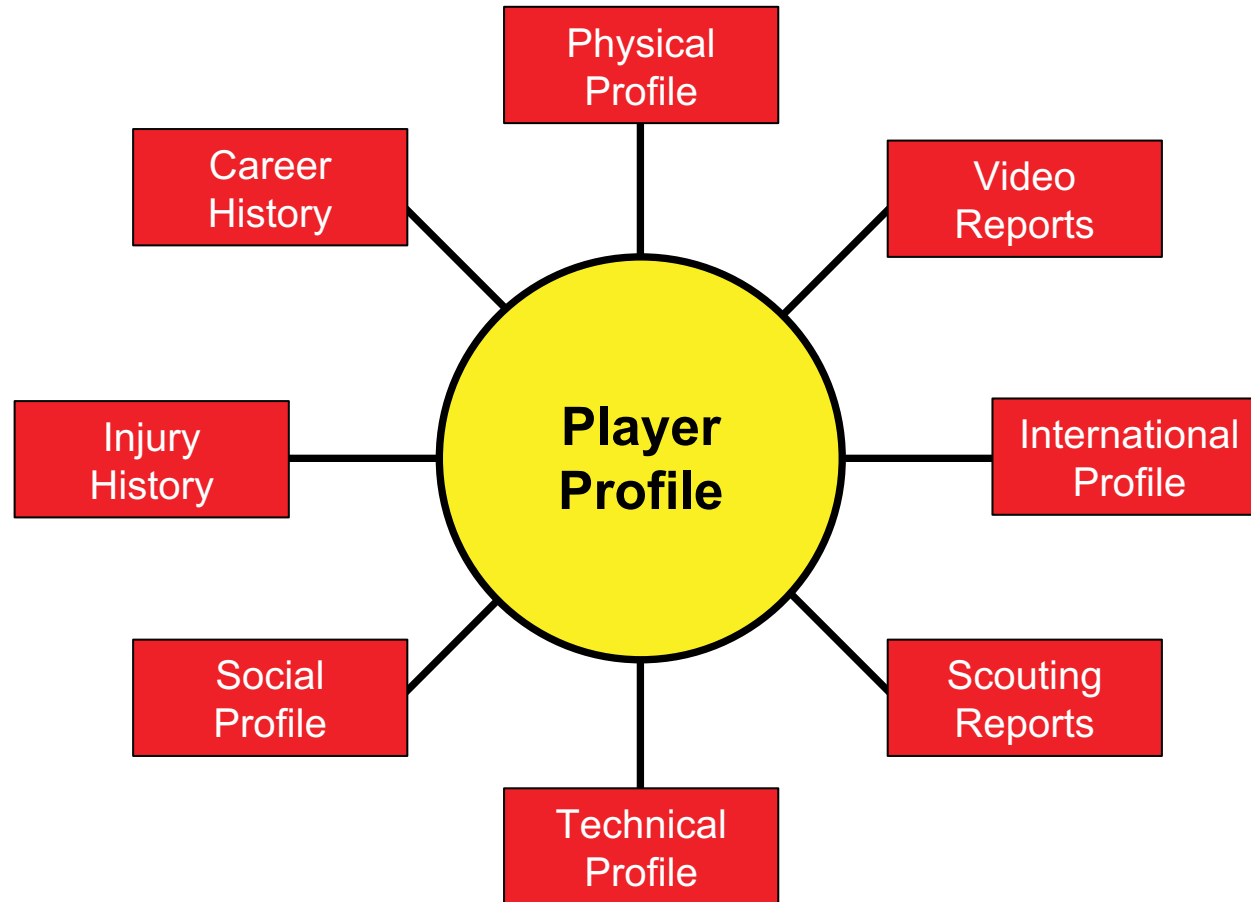
Data Flows

Visual representation of the data available to football clubs and the key ways it's used



Application of Data Compliments All Areas of Player Profiling

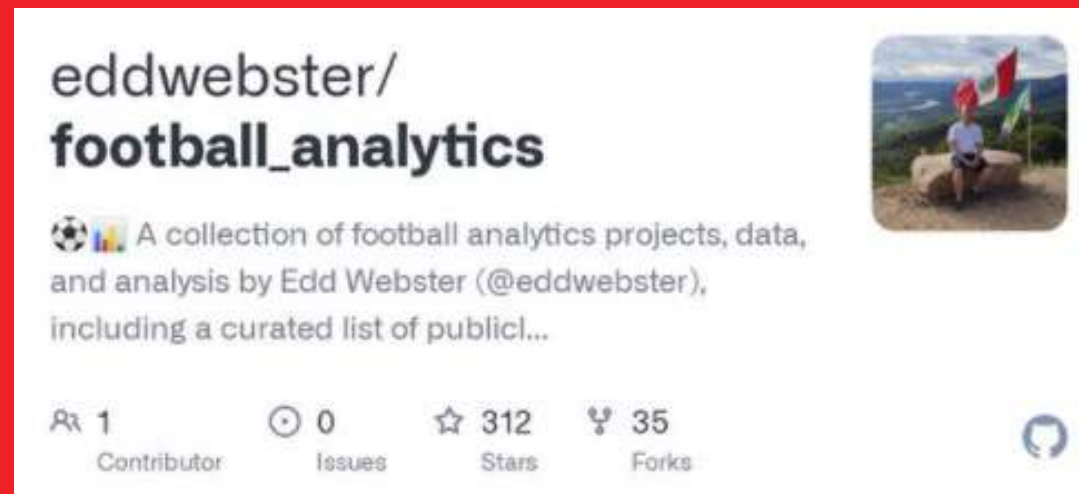
Recruitment of players requires the consideration of many different areas, for which data is important



Gerard Moore uses the Event Lab to analyse centre-backs for recruitment: twenty3.sport/gerard-moore-analyst-twenty3-event-lab-recruitment/.

Sample Football Data Science Projects Overview

Four, in-depth projects that demonstrate my broad skill-set regarding data science and football analytics



Football Data Science Projects

Overview of four key areas of work that are explored in detail in this pack

1

Tableau Football Insights Tools featuring the use of StatsBomb Event data

Creation of analysis tools for analysing player performance, pre- and post-match tactics, and recruitment, enabling the democratisation of data to all levels of an organisation. Data used features StatsBomb Event data for the FA WSL, for the 19/20 and 20/21 seasons.

2

Recruitment analysis of Center Backs

Scenario for a data-driven recruitment analysis, to determine "the next Gerard Piqué", for a hypothetical, newly-promoted club in the 'Big 5' European leagues.

3

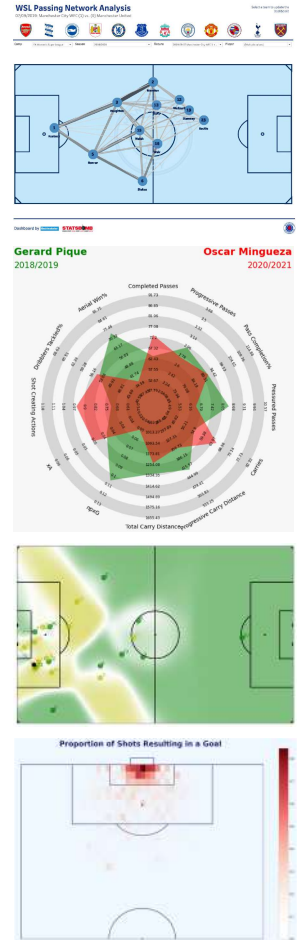
Application of Tracking data

Engineering and analysis of Signality, Metrica Sports, and Stats Perform Tracking data. Analysis includes analysing passages of play, determining physical performance, the implementation a basic Pitch Control models, and applying Expected Possession Value (EPV) frameworks.

4

Building an Expected Goals model with Event data

Training an Expected Goals (xG) model using Event data, subsequently applied to a separate match dataset of Event and Tracking data, to analyse the performance of the teams in question. Models created using Logistic Regression, Random Forests, and Gradient Boost Decision Trees (GBDT) algorithms including XGBoost and CatBoost.



All code for these three projects outline: github.com/eddwebster/football_analytics

Tableau Public profile: public.tableau.com/profile/edd.webster.

Datasets Used

An array of publicly available datasets from disparate sources have been used to produce this pack

- [Metrica Sports](#), [Signality](#), and [Stats Perform](#) Tracking data;
- [StatsBomb](#) 360 data for the UEFA Men's Euro 2020;
- [StatsBomb](#) Event data for the FA Women's Super League (18/19-20/21) and FIFA Men's 2018 World Cup;
- [Wyscout](#) Event data for the 'Big 5' European Leagues for the 17/18 season;
- [Understat](#) match-by-match shooting location and meta data with corresponding xG¹ values for the Big 5¹ European Leagues for the last eight seasons (14/15-present);
- [FBref](#) aggregated season player performance data for the Big 5 European Leagues for the last five seasons (17/18-present), provided by StatsBomb;
- [Twenty First Group](#) aggregated season player performance data for the provided by EFL Championship, League 1 and League 2 for the 17/18 and 18/19 seasons;
- [StrataBet](#) Event data for 'chances' provided by for the 16/17 and 17/18 season (to varying levels of completeness) for the top tiers of England, Germany, Spain, China, Turkey, Switzerland, Sweden, Greece, Netherlands, Austria, Australia, Norway, and Scotland as well as the second tiers of Germany and England. The shot data contains freeze-frames of the number of attacking and defending players between the shot-taker and the goal and the pressure that the shot taker was under. It should be noted this is not a complete dataset;
- [TransferMarkt](#) player bio, estimated value, and contractual data;
- [Capology](#) player salaries data; and
- [ClubElo](#) estimated team rating based on historical performance.

The notebooks in which wrangle the datasets can be found at the following [\[link\]](#) and all engineered datasets have been exported and made publicly available to view and download in Google Drive (allows for GitHub 100mb size constraints) [\[link\]](#).

¹The 'Big 5' European Leagues constitute of the Premier League (UK), Ligue 1 (France), Serie A (Italy), Bundesliga (Germany), and La Liga (Spain).

²xG is Expected Goals, a metric that measures the quality of a shot based on several variables such as: assist type, shot angle and distance from goal, whether it was a headed shot, etc.



Proposed Data Science Tools

Programming languages and visualisation tools used for this data analysis task

Back End
(Data Parsing and Engineering)



Front End
(Visualisation and Analysis)



Back-end Tools

Why Python has been selected for data engineering, preparation and modeling

- The data preparation (back-end) has been conducted using [Python](#) in a [Jupyter](#) notebooks environment
- [Anaconda](#) is open-source package of data science and machine learning tools, including Python and all the major libraries.
- Python's strength as a language come from the number of create packages and libraries available such as data munging and processing libraries like [pandas](#), visualisation libraries such as [matplotlib](#), machine learning libraries like [scikit-learn](#) and [XGBoost](#), statistical analysis libraries such a [statsmodels](#), fuzzy data matching libraries such as [record-linkage](#), webscraping packages such as [Beautifulsoup](#), and custom football analytics community such as the [mplsoccer](#) visualisation library or code to work with tracking data by [Laurie Shaw](#) [\[link\]](#).
- For these reasons, Python is the backend tool of choice used to construct this analysis pack, with the code to parse and engineer the data in this deck available at the following [\[link\]](#).

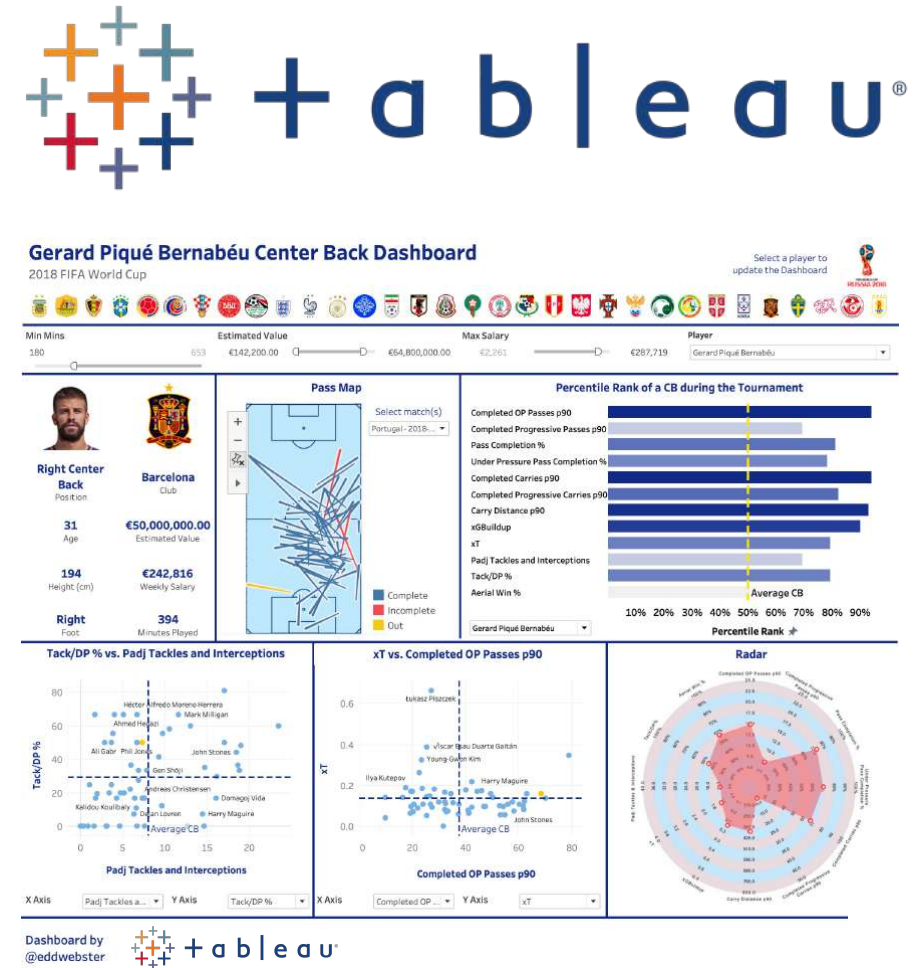


Front-end Tools

Why Tableau has been selected for data visualisation and analysis

- Many of the front-end data visualisations have been created using [Tableau](#) (in some cases [matplotlib](#)).
- For data analysis to be effective, it needs to be a part of the holistic approach of an organisation and be accessible not just to power users such as data analysts and data scientists. This is very applicable in an organisation, such as a football club, where functionality and accessibility to all levels of ability is of great importance. Through Tableau reporting, data can be made available to practitioners of every level of the organisation, such as sporting directors, head coaches, physios, analysts, and players.
- The adaptability for Tableau to be used by beginners and power users alike, as well as the tool's ability to reduce the users time to insight, is why I have selected it for visualisation.

Tableau Public profile: public.tableau.com/profile/edd.webster.



Project 1: Tableau Reporting

Tools for the purpose of analysing player performance, pre- and post-match tactics, and recruitment, featuring StatsBomb Event data

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

Further Work



Turning Big Data into Meaningful Insights

How does Tableau achieve this?

1

Simplicity

Deliver clear and concise messages that can be comprehended quickly and concisely.

2

Enable Control

Dashboard and visualisations need to include filters, tool tips, and actions, that can allow the user (often coaches) to find the insights themselves without the need for programming or dashboard building skills.

3

Flexibility

Filters, tooltips and actions can allow the user to find the insights themselves.

Football Insights Analysis Tools Overview 1/2

A selection of tools created for football analysis, with further detail of the insight they bring (on the following slides)

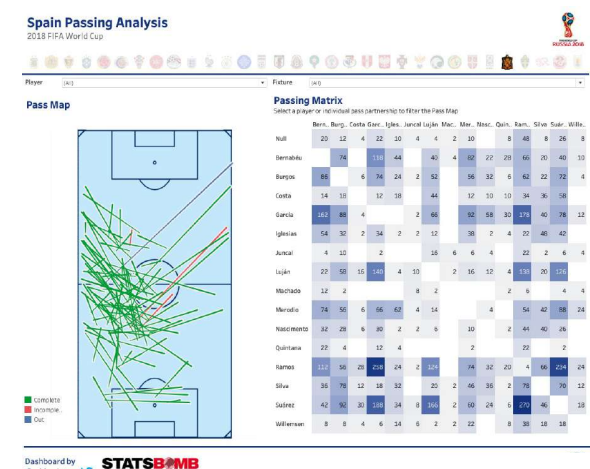
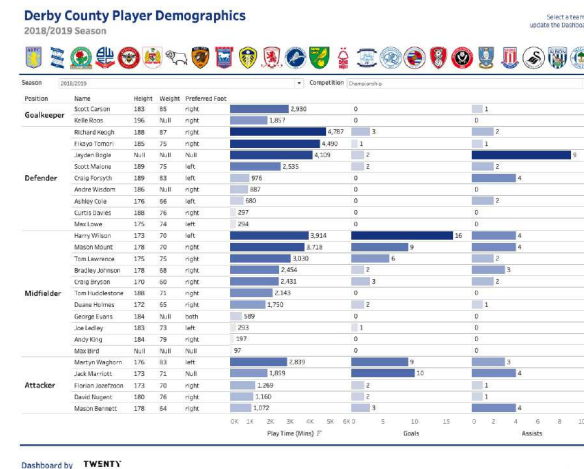
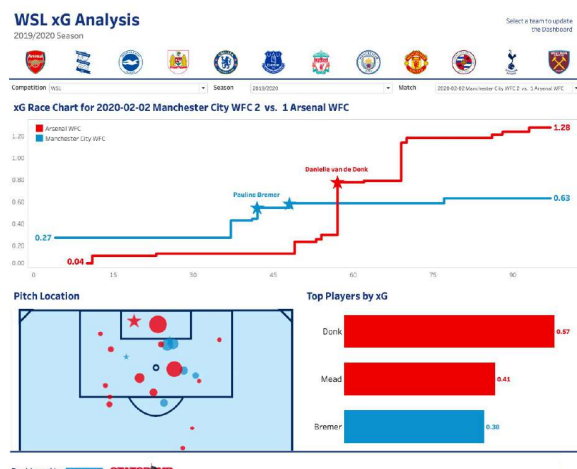
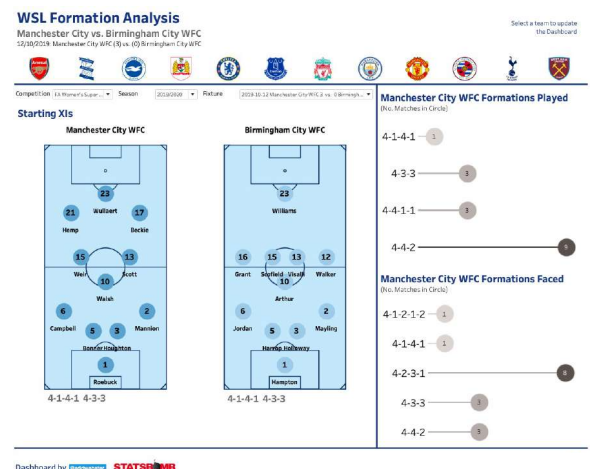
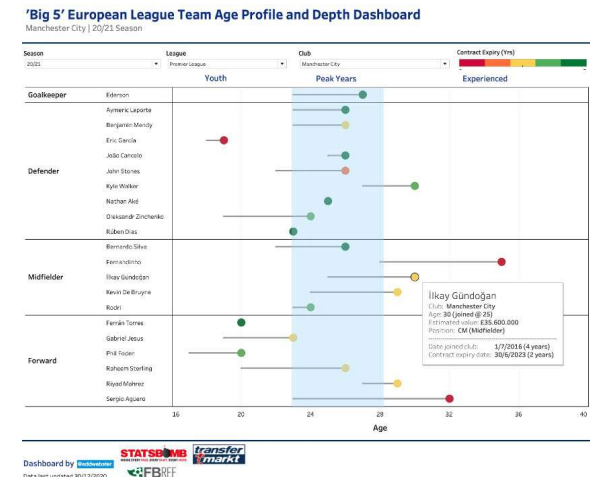
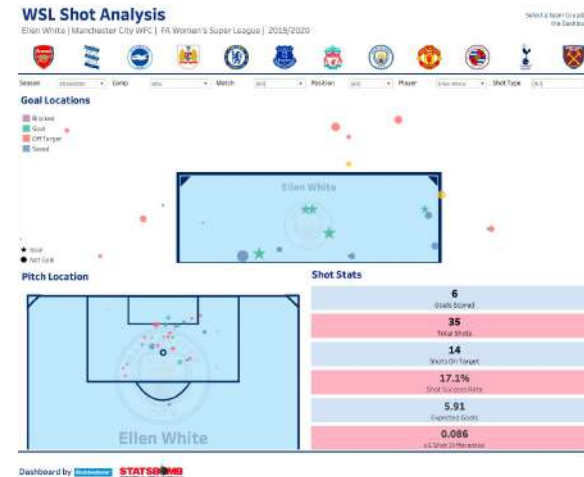
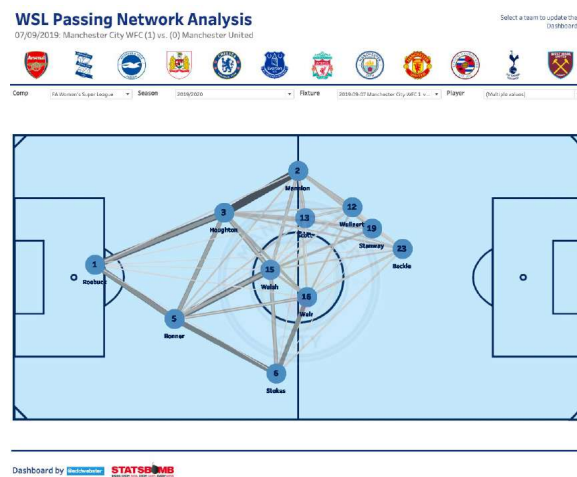
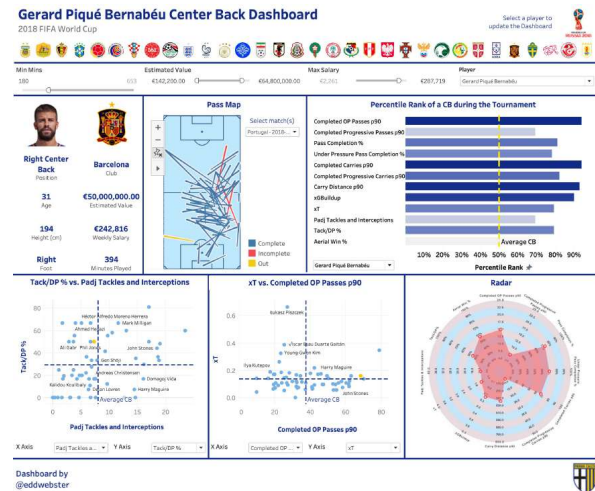


Tableau Public profile: public.tableau.com/profile/edd.webster.

Introduction

State of Play in Football Analytics

Why Data in Football is Important

Sample Projects Overview

P1: Football Intelligence Tools

P2: Recruitment Analysis

P3: Tracking Data Applications

P4: xG Modeling

Conclusion

Further Work



Football Insights Analysis Tools Overview 2/2

More reporting tools created using Tableau

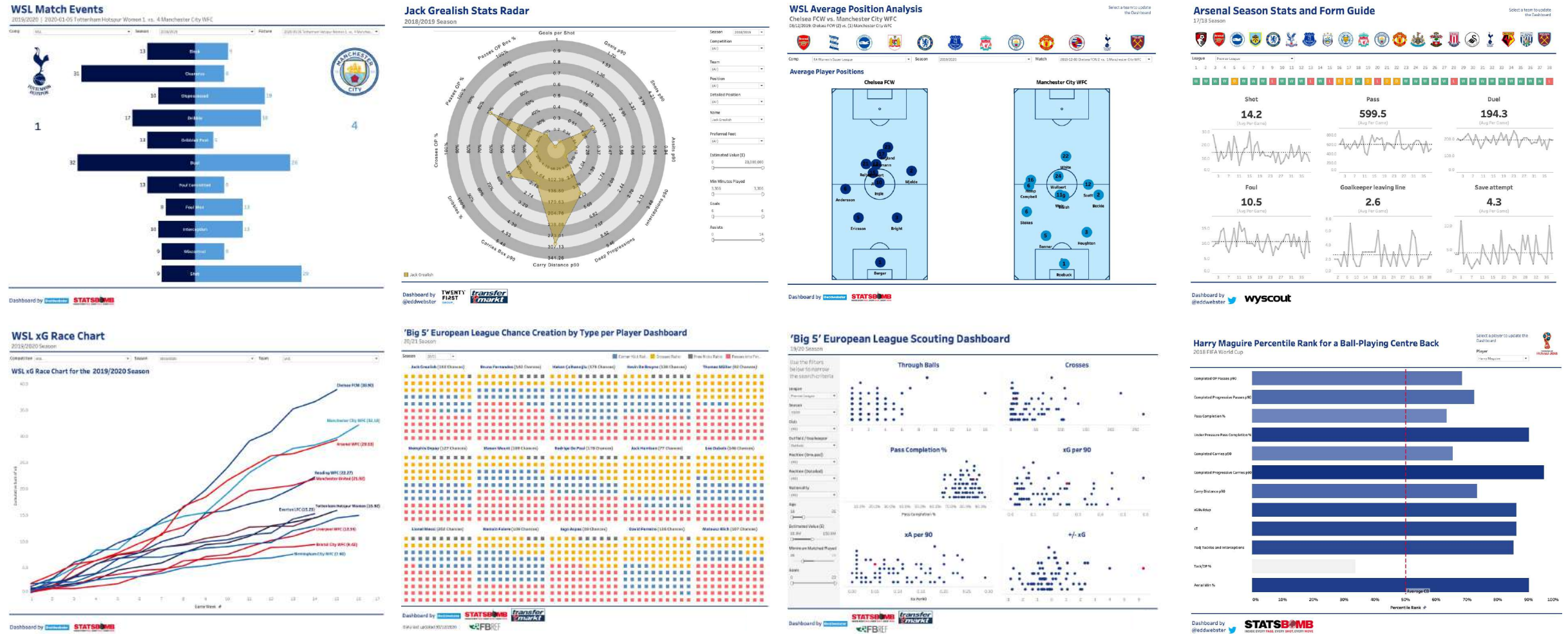


Tableau Public profile: public.tableau.com/profile/edd.webster.

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2: Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

Further Work



Tool 1: Team Stats and Form Guide Dashboards

Overview of individual team's performance over the course of the season

Arsenal Season Stats and Form Guide

17/18 Season



League Premier League

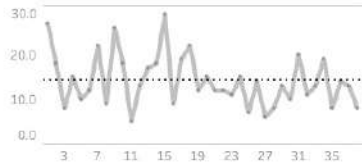
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38

W W W W D W W W L W W W L W L D D W D L D D W W W W W W L W W W W W W W W L

Shot

14.2

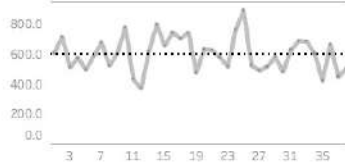
(Avg Per Game)



Pass

599.5

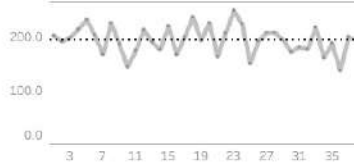
(Avg Per Game)



Duel

194.3

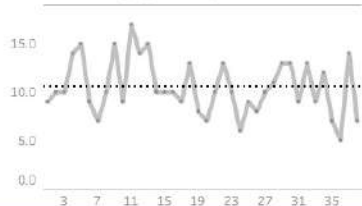
(Avg Per Game)



Foul

10.5

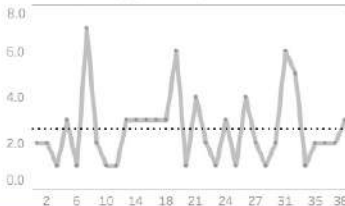
(Avg Per Game)



Goalkeeper leaving line

2.6

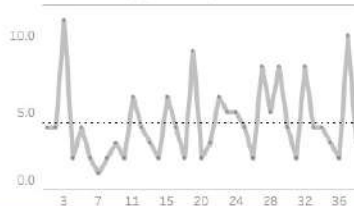
(Avg Per Game)



Save attempt

4.3

(Avg Per Game)



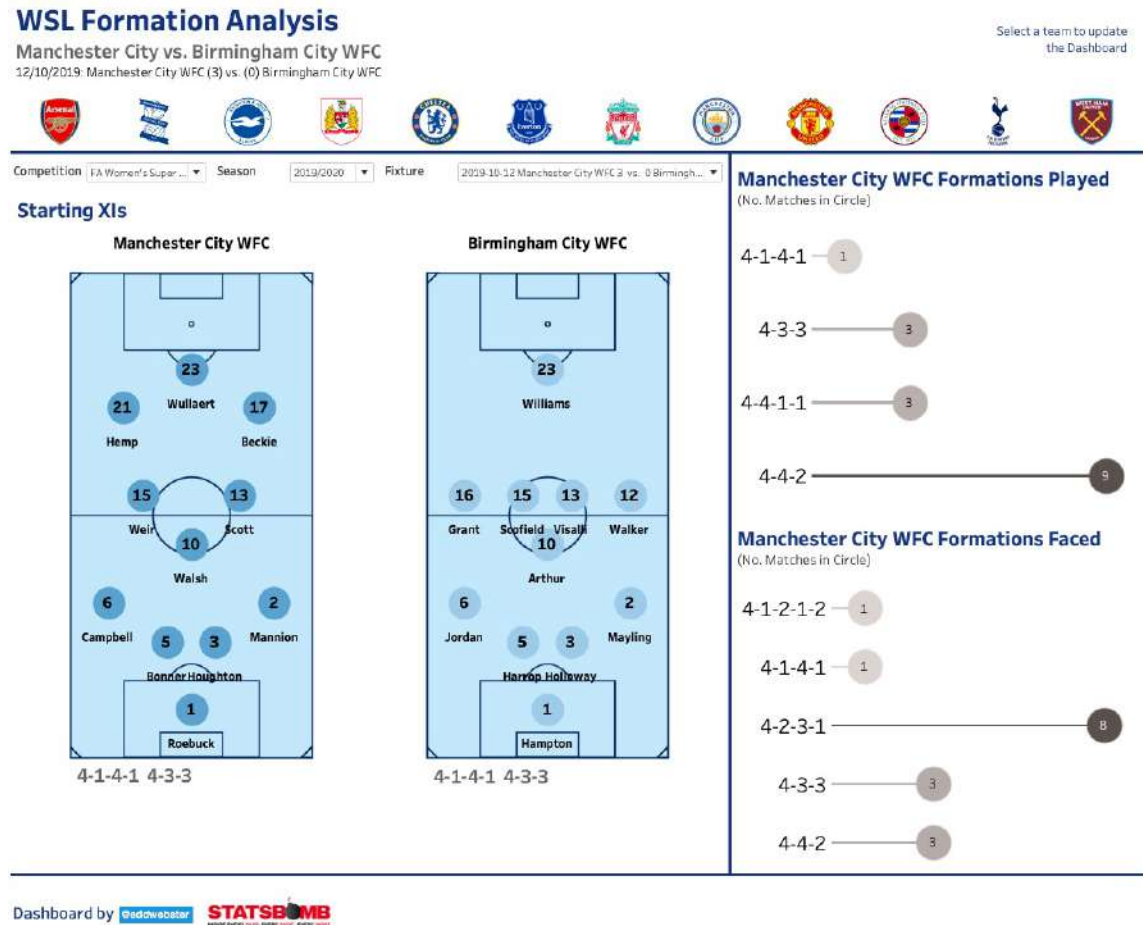
Dashboard by @eddwebster

Event data provided by Wyscout. See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-FAWSLAnalysisandDashboard/Big51718SeasonDashboard.

- The **Team Stats and Form Guide** dashboard provides an overview of performance and results for all teams in the 'Big 5' European leagues for the 17/18 season, derived from Event data provided by [Wyscout](#).
- Before reaching the dashboard, Event data from Wyscout is parsed [\[link\]](#), engineered [\[link\]](#), before being aggregated further as a separate datasets of match results and aggregated metrics. See the following notebook for data preparation [\[link\]](#).
- This dashboard visualises sample KPIs for a selected team, as well as their win record, over the course of the selected season, in this case, the 17/18 season.
- This dashboard can be used to quickly see the current performance of the team against the opposition, with each KPI is benchmarked against the season average, visually displaying in which aspects of play a team is under/over performing, and against which opposition and time in the season.

Tool 2: Formation Analysis 1/2 – Recorded Formations in Event Data

Tactical setup and starting XI personnel visualisations

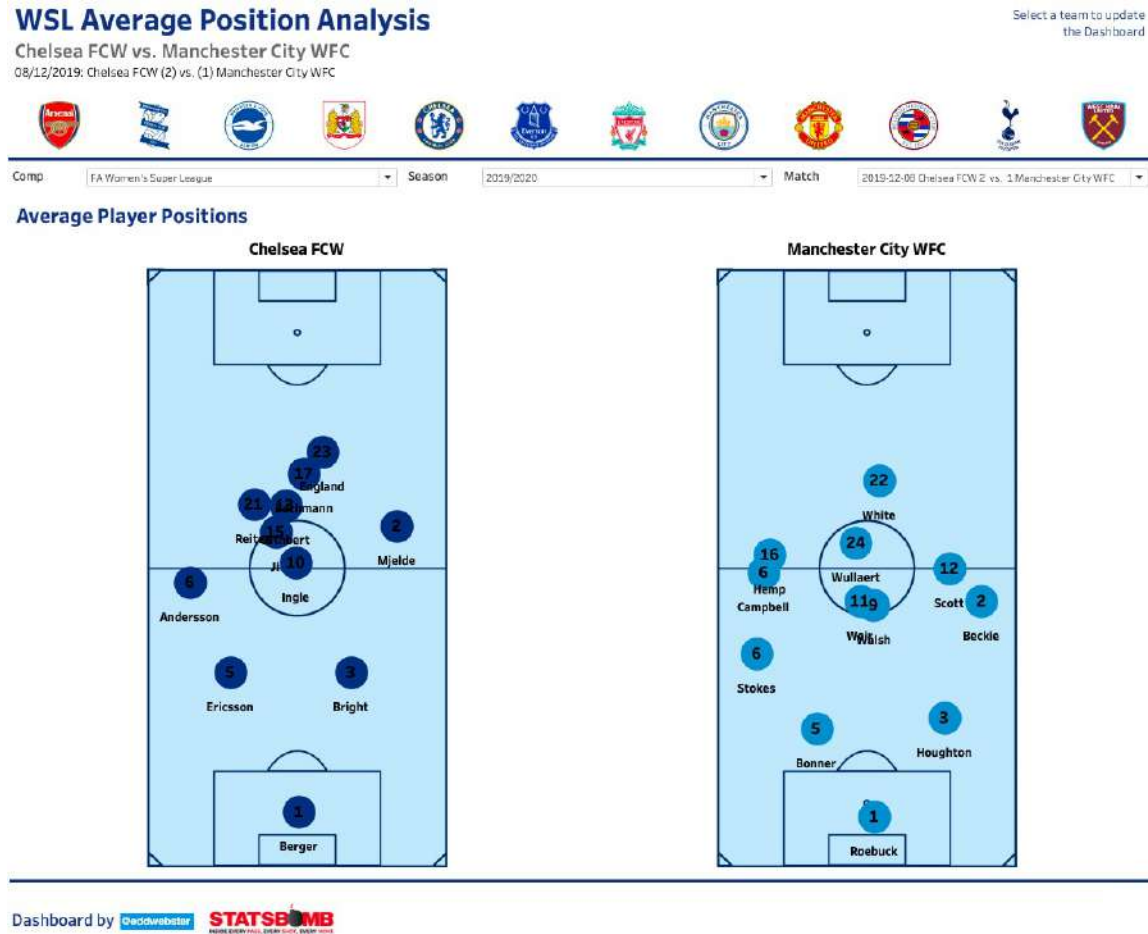


- The **Formation Analysis** dashboard, created using Event data provided by [StatsBomb](#), visualises the starting XI and formation for all teams in all matches in the WSL during the 18/19 and 19/20 seasons.
- This dashboard visualises the formations played by a team and their opponent for a selected match, as well as all the formations that team has both played and faced during the season.
- In this example, the dashboard shows the starting XI personnel and 4-3-3 formation that Manchester City WFC played against Birmingham City WFC on 12th October 2019.
- The dashboard shows that despite 4-4-2 Manchester City WFC's preferred formation during the 19/20 season, the manager elected to go with a 4-3-3 formation against Birmingham City WFC, a formation they played three times during the season. This could be to personnel restrictions such as injuries or suspensions, or maybe a tactical choice.
- This dashboard can be used in opposition analysis to see the formations the opposing team is using, the regular personnel, as well as all the formations that the team has faced. With this knowledge, a coach can observe the formations has faced less or the formation for which the opposing team has less success, to use as a starting point for a team's strategy to expose weaknesses in the opposition.
- Potential improvements and next steps for this dashboard:
 - Include tactical shift information to show the the formation changes that took place during the game and to see the players that have been substituted on and off the pitch.
 - Win/loss statistics a team has against each formation – is there a particular formation a team has difficulty dealing with and can be exploited?

Team formation data aggregated and extracted from WSL event data provided by StatsBomb. See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-FAWSLAnalysisandDashboard/WSLFormationAnalysisDashboard.

Tool 2: Formation Analysis 2/2 - Average Position Analysis

Average player position and formation



Event data with X, Y shooting positions and xG values provided by StatsBomb. See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-FAWSLAnalysisandDashboard/WSLAveragePlayerPositionsDashboard.

- The **Average Position Analysis** dashboard, created using Event data provided by [StatsBomb](#), visualises the average playing position for all matches in the WSL during the 18/19 and 19/20 seasons.
- In this example, the match selected is between Manchester City WFC and Chelsea FCW that took place on 23rd February 2020.
- The data is filtered to show only players that made a certain number of passes during the 90 minutes. This limit can be changed using the slider filter. This ensures that the visualisation shows only those player that made a big impact on the game i.e. the starting XI or key substitutes.
- For an initial analysis, this dashboard can also provide the insight as to the players spatial behaviour and tactical positions. For example, we can use this dashboard as a starting point to see if a full back like to make themselves available on the overlap or play conservatively, if a striker is a target man or likes to make the run beyond the last man. We can also try to spot tactical decision such as whether a team try to keep the ball in the midfield or play more direct.
- This dashboard, however, provides only the initial analysis to build upon with other analysis and dashboards.
- Other potential examples of analysis possible with the Event data used here include:
 - Shooting position analysis;
 - Passing analysis; and
 - Set-piece heat map,

Tool 3: Attacking Actions and Chance Analysis

Visualisation of a team's chance creation in a match and the correspond xG plots

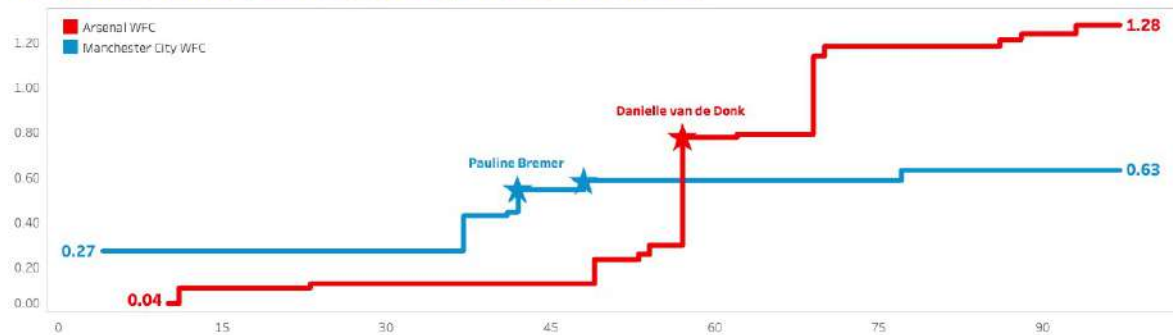
WSL xG Analysis

2019/2020 Season

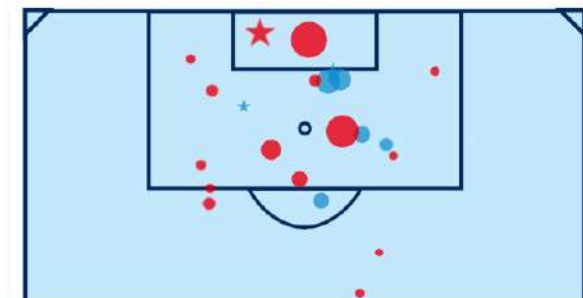
Select a team to update the Dashboard



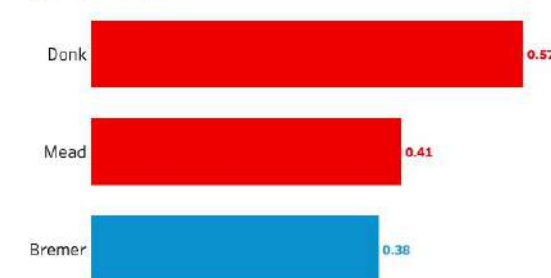
xG Race Chart for 2020-02-02 Manchester City WFC 2 vs. 1 Arsenal WFC



Pitch Location



Top Players by xG



Dashboard by @eddwabster **STATSBOMB**

Event data with X, Y shooting positions and xG values provided by StatsBomb. See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-FAWSLAnalysisandDashboard/WSLxGAnalysisDashboard.

The **xG Analysis** dashboard, created using Event data provided by [StatsBomb](#), visualises the chances in a match using three xG visualisations - an xG race chart, the corresponding shooting position plot, and a bar chart of the top performing players by their xG contribution.

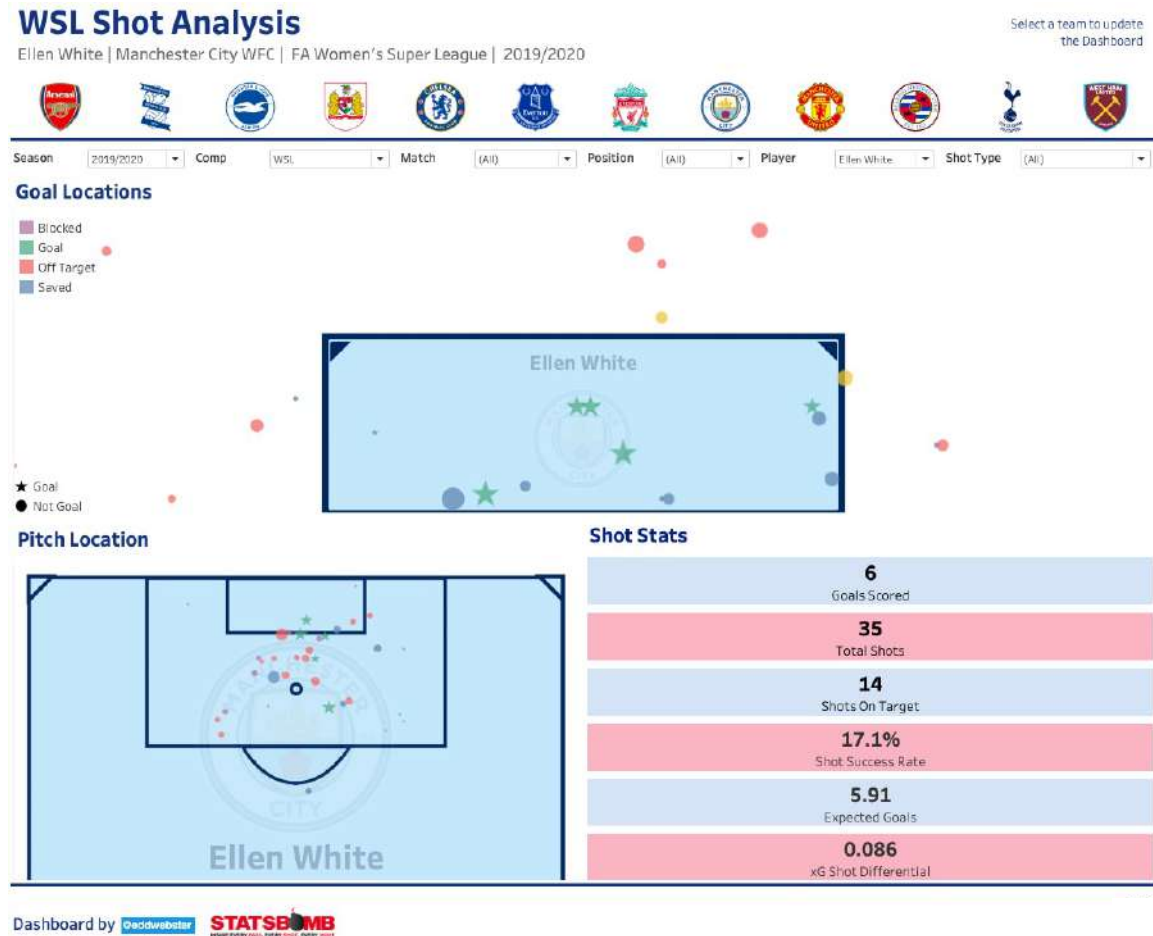
This dashboard visualises the cumulative sum of Expected Goals over time as a race chart and is linked to a 2D pitch location map, allowing for greater insight and investigation of where a specific event in the race chart took place on the pitch through the X and Y coordinates. Actual goals that take place are then denoted with a star.

In this example, the dashboard show the events that took place when Manchester City WFC played Arsenal WFC on 2nd February 2020 during the 19/20 WSL season.

From this dashboard, we can see that despite having a 2-1 winning score-line, Manchester City WFC were behind Arsenal WFC in the xG statistics, suggesting that the contest was potentially one more in Arsenal WFC's favour, in which Arsenal created more chances. From this point we would analyse the match using video analysis to see if this assumption is correct.

Tool 4: Shooting Position Analysis

Ellen White's goal-scoring opportunities in the WSL during the 19/20 season



- The **Shot Analysis** dashboard, using data provided by [StatsBomb](#), visualising both the positions of where a shot is taken from on the pitch and where in the ball finishes in the goal, using the X, Y, and Z locations of shots with corresponding StatsBomb xG data.
- This X, Y, and Z data for both the pitch and goal location is only available in the StatsBomb data and not in other public datasets such as those from [Understat](#) and [StrataBet](#).
- For each shot, the dashboard shows whether:
 - the shot a goal (star) or not (circle)
 - the result of the shot a goal (green), a missed shot (red), a blocked shot (purple), or a saved shot (blue).
- In this example, the dashboard has been filtered to show all the shots by Ellen White for Manchester City WFC during the 19/20 WSL season.
- From this dashboard, we can see White takes very few shots outside the box, with most of her opportunities coming from within or close to the 6-yard-box, with a slight favouritism to take shots in a position of the keeper's left-hand side. Very often she also shoots to this same left-side of the goal.
- This analysis can be expanded to analyse a myriad of analytical takes, such as: which side/foot a player favours to take a shot; a heat map of the space a player likes to find themselves in; and to which of the keeper's sides do the strike a penalty.
- Potential improvements and next steps for this dashboard include using StatsBomb 'Freeze Frame' data to create visualisations that include the positions of all players on the pitch including the goalkeeper, when a goal is scored.

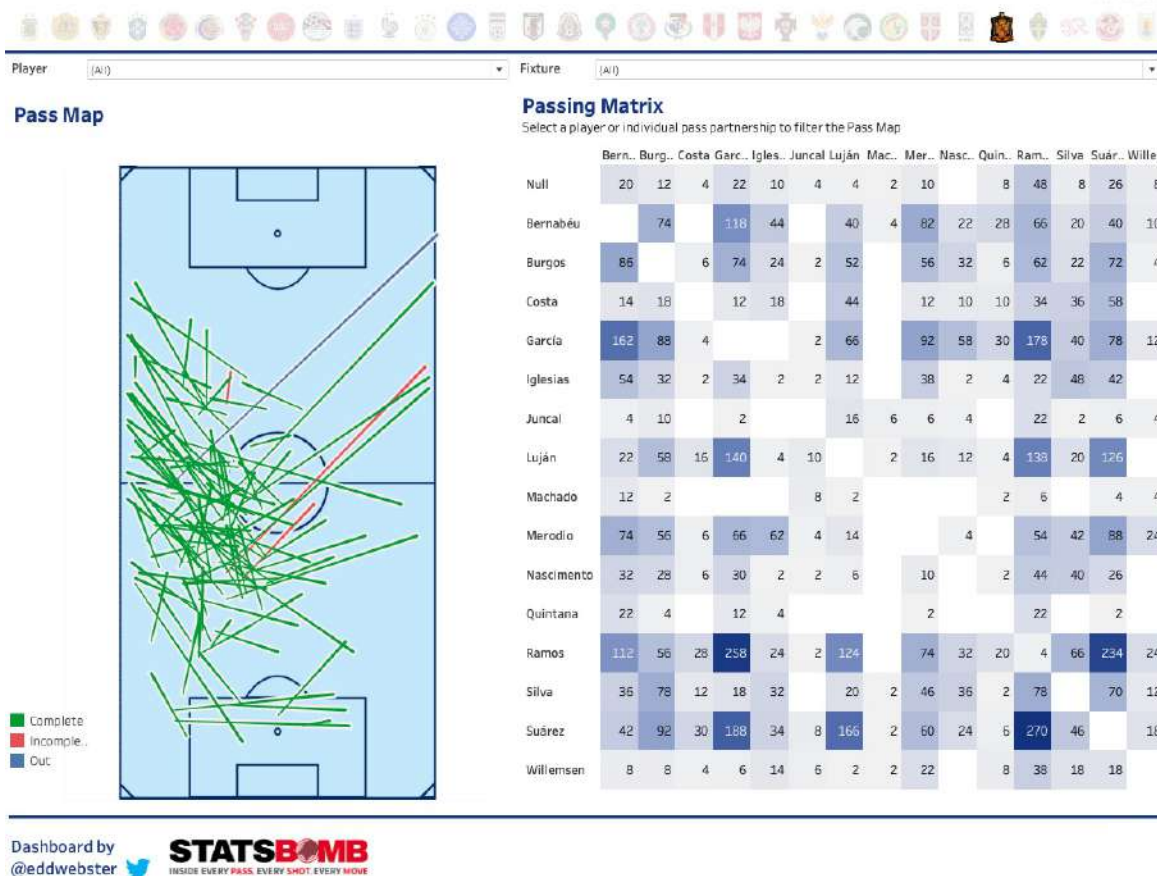
Event data with X, Y, Z shooting positions and xG values provided by StatsBomb. See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-FAWSLAnalysisandDashboard/WSLShootingAnalysisDashboard.

Tool 5: Passing Map and Matrix

Visualisation of passes that can be filtered between players to inspect tactical plays and partnerships

Spain Passing Analysis

2018 FIFA World Cup



- The **Passing Map and Matrix** dashboard created using Event data provided by [StatsBomb](#), visualises all the passes made by a player in a selected team and match, as well as the relationships between all the players in a team.
- The pass map and matrix that can be filtered by the passing matrix, to focus in on particular passing combinations or all passes from a particular player.
- In this example, we can see the pass map for Sergio Ramos of Spain in their group stage match against Portugal (15th June 2018).
- In this center back analysis pack, the pass map is most useful to identify trends of passes i.e. does the player like to play the ball short into the midfield, are they attempting to hit long diagonal balls, etc.
- This dashboard can also be used in analysis pieces such as opposition analysis to identify patterns in play and partnerships between players e.g. does the goalkeeper like to kick long/short? Does a team like to switch to the left-midfielder upon a game restart, etc. A coach can then use this information to try and hinder synchronised attacks and key partnerships on the field.

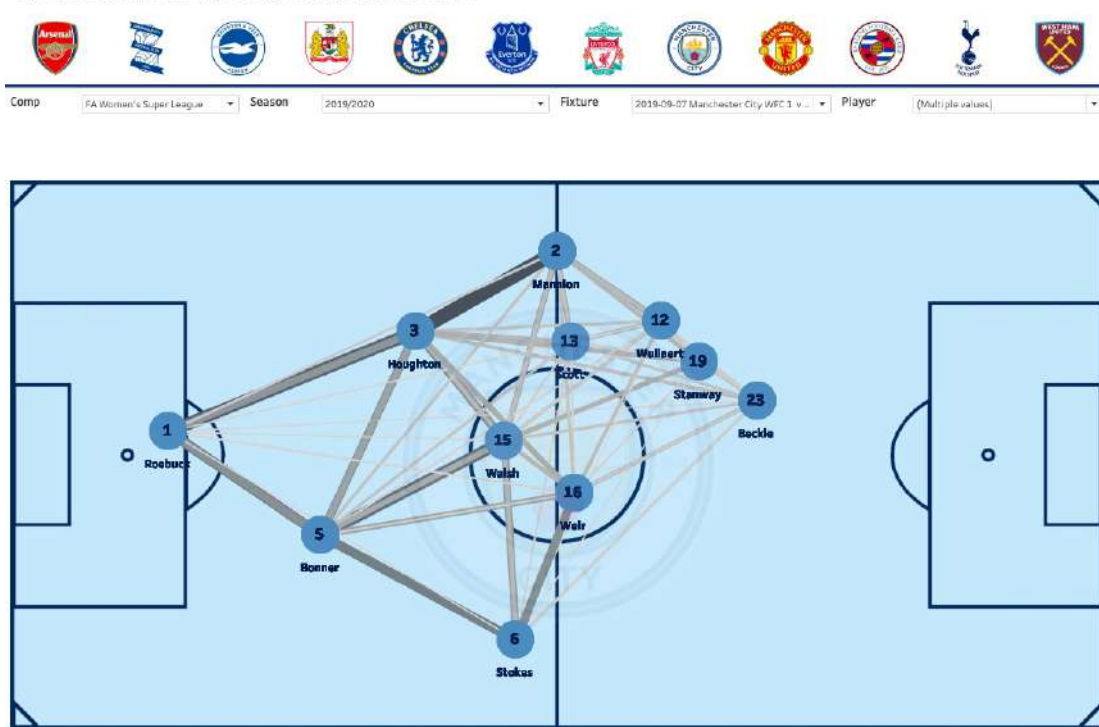
See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-WorldCup2018AnalysisandDashboard/WC2018PassingMatrixDashboard.

Tool 6: Passing Networks

Visualisation of a team's passing network to detect pass sequences and playing styles

WSL Passing Network Analysis

07/09/2019: Manchester City WFC (1) vs. (0) Manchester United



Dashboard by [Edd Webster](#) **STATSBOMB**

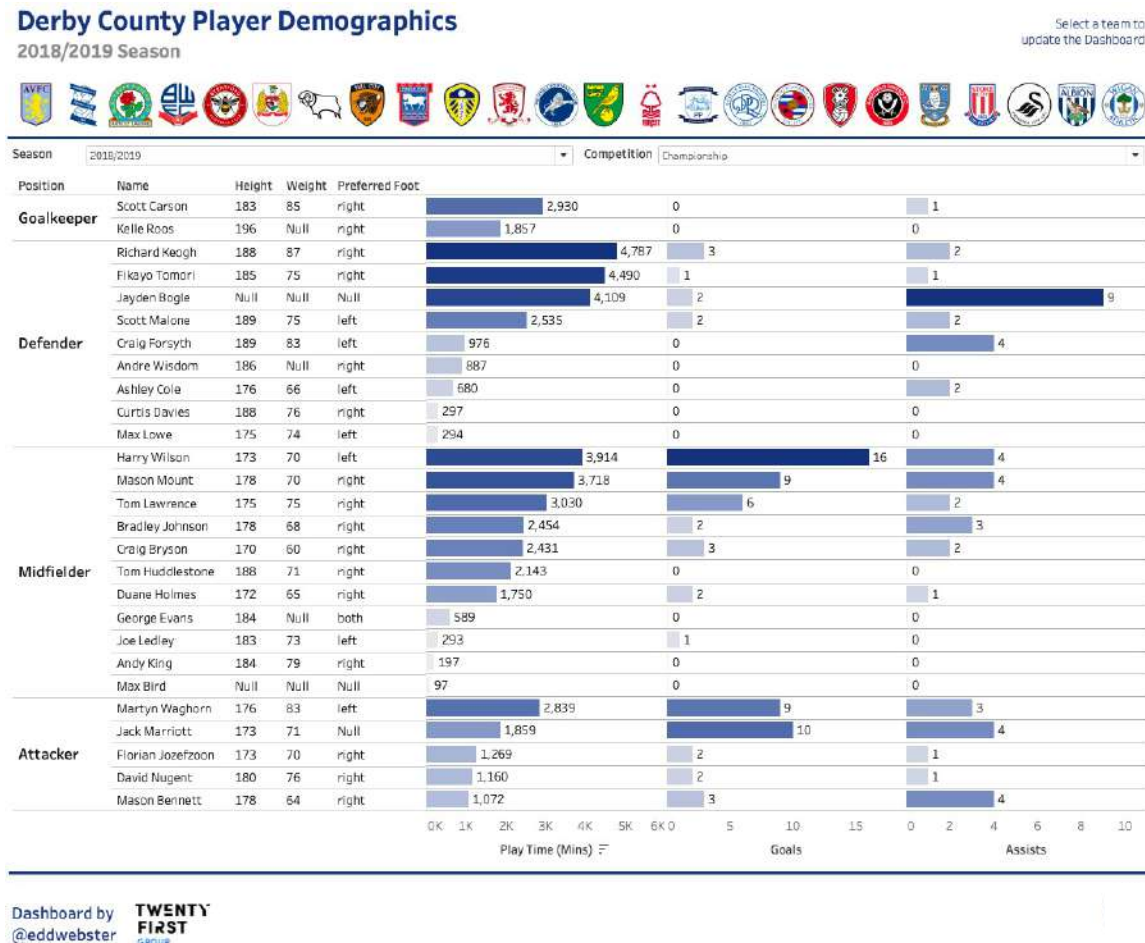
- The **Passing Network Analysis** dashboard created using Event data provided by [StatsBomb](#), visualises the passes in WSL matches during the 18/19 and 19/20 seasons as a passing network,
- Passing networks are constructed from the observation of the ball exchange between players, where network nodes (or vertices) are football players and links (or edges) account for the number of passes between any two players of a team. See the following for more information [[link](#)].
- The network visualisation shows the players in their average passing position during the match, with lines between each of the players where passes were made. The more frequent the the number of passes between two players, the thick and darker the grey line.
- In this example, we can see that Manchester City WFC build their play out from the back, with the majority of passes taking place between the goalkeeper Ellie Roebuck and the centre backs Steph Houghton and Gemma Bonner, who in turn passed the ball out the the fullbacks Demi Stokes and Aoife Mannion. Roebuck only opts to go long a few occasions, represented by a faint, thin grey line.
- Alongside the previous dashboard with the passing map and matrix, this dashboard can be used to as part of an opposition analysis to identify patterns in play and partnerships between players e.g. does the goalkeeper like to kick long/short, does a team like to switch to the left-midfielder upon a game restart, etc.

Defining a historic football team: Using Network Science to analyze Guardiola's F.C. Barcelona: J. M. Buldú, J. Busquets, I. Echegoyen & F. Seirul.

Event data with x, y passing positions provided by StatsBomb. See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-FAWSLAnalysisandDashboard/WSLPassingNetworkDashboard.

Tool 7: Player Demographics Dashboard

Simple bar charts the deliver key player information quickly and concisely



- The **Player Demographics** dashboard is a simple collection of bar charts of Premier League aggregated player performance data for the 11/12 season, provided by [Opta](#).
- The bars show key playing statistics for all of the players in the selected squad including: the matched played, matches started, substitutions on and off, and the total minutes played.
- The dashboard groups players by their position type – goalkeeper, defender, midfielder, and forward, as well as physical attributes including: age, height, weight, and preferred foot.
- This dashboard serves many uses cases, most notably, as part of the initial steps in opposition analysis for coaches and managers to try and predict the opposition's next line-up and who are the key players in the the team's first XI.

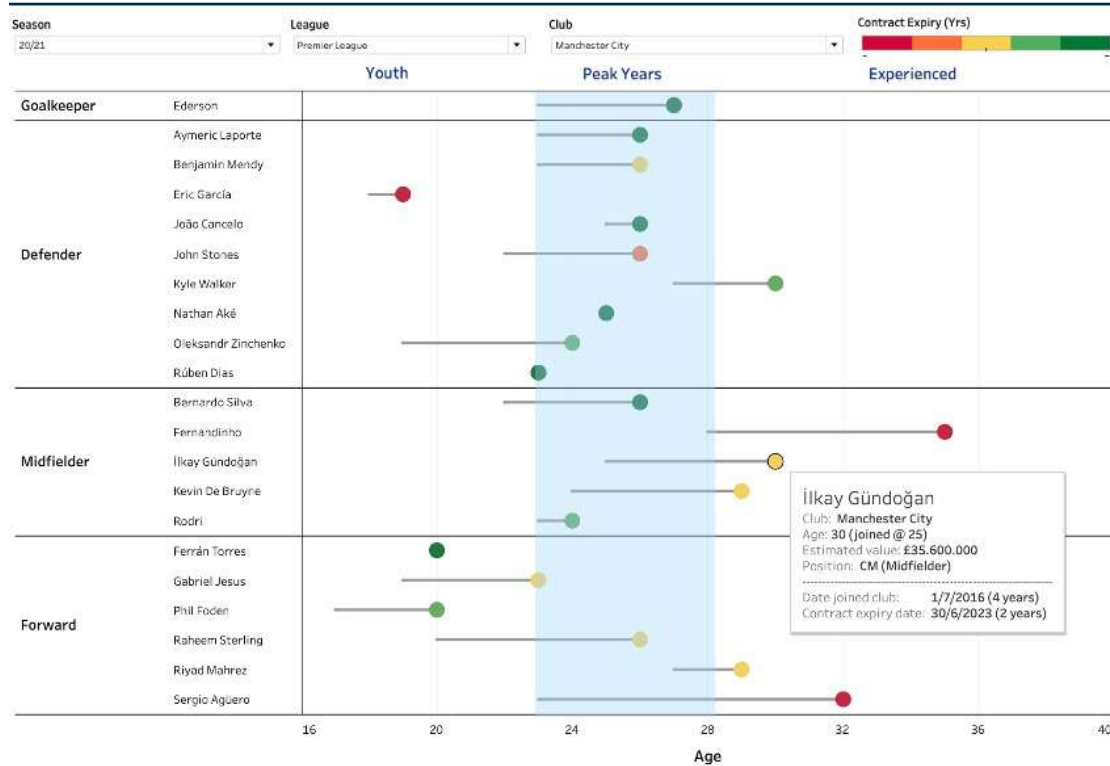
Aggregated player performance data provided by Twenty First Group and estimated market value data provided by TransferMarkt. Data correct as of 16th September 2020. See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-EFLAnalysisandDashboards/EFLPlayerDemographicsDashboard.

Tool 8: Squad Age Profiling

Insight of a squad's age and their remaining years in relation to their current contractual situation

'Big 5' European League Team Age Profile and Depth Dashboard

Manchester City | 20/21 Season



Dashboard by @EddWebster | STATSBOMB | FBREF | transfermarkt
Data last updated 30/12/2020

Aggregated player performance data provided by FBref and estimated market value data provided by TransferMarkt. Data correct as of 16th September 2020. See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-Big5EuropeanLeagueAnalysisandDashboards/Big5SquadAgeProfiling

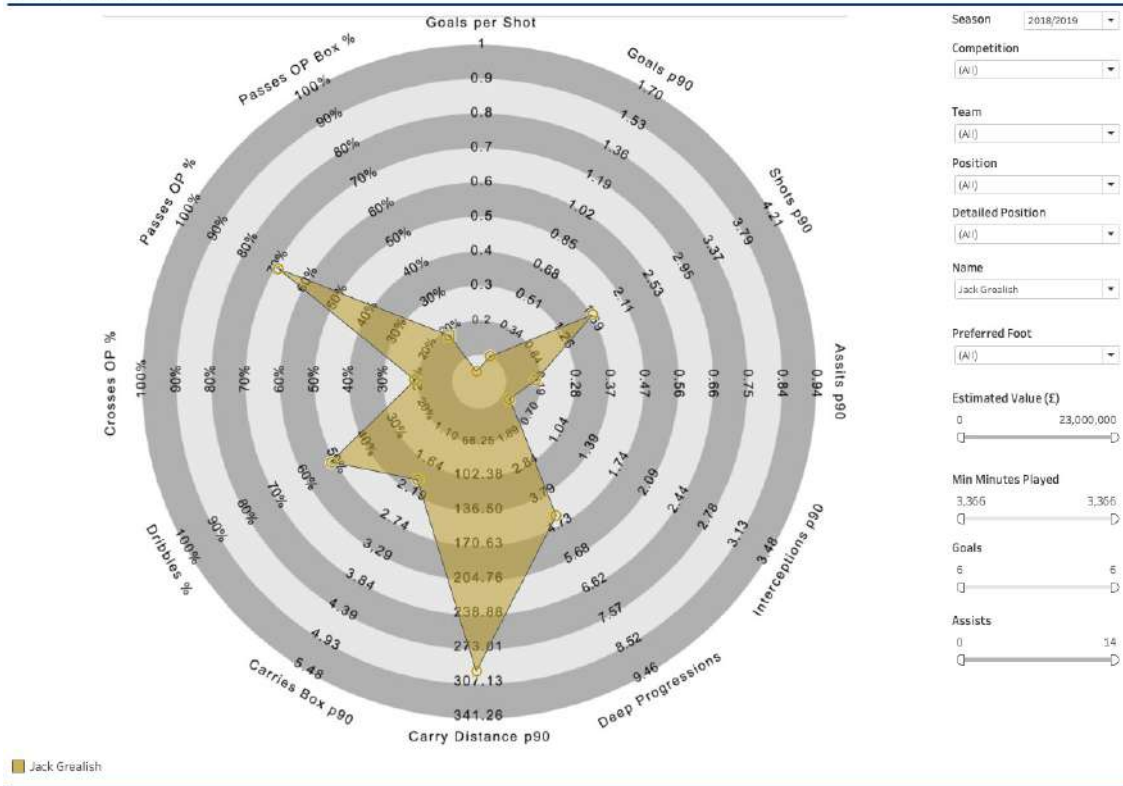
- The **Squad Age Profile and Depth** dashboard is a Lollypop Gantt-style dashboard created from the fuzzy matched dataset of player performance data from [FBref](#) and player valuation data from [TransferMarkt](#).
- The chart shows the players for a selected team, their current age, and the age in which they joined the club. The dots for the player's current age are coloured as per the number of remaining years on their contract, according to TransferMarkt.
- This dashboard can be used quickly to see the relationship between a player's age, how long they have been with the club, and how long we can expect them to stay with the club, and if they are running out of years on their contract.
- In this example, we can see that Manchester City have three players in the final year of their contract - Eric García (19), Sergio Agüero (32), and Fernandinho (35). Agüero and Fernandinho are both important players, but are both post their peak years, and therefore it's natural for them to not be on long contracts. However, we can quickly see that García, is not even in his peak years yet and his contract has less than one year left, suggesting that he could likely leave the club imminently. We can also see that John Stones has less than two years on his contract, despite being in his peak years, a potential cause for concern give the good season he had

Tool 9: Position-specific Radars

Multi-metric visualisation of players performance, per position

Jack Grealish Stats Radar

2018/2019 Season



Dashboard by
@eddwesbter

TWENTY
FIRST
GROUP

transfer
markt

Aggregated player performance data provided by Twenty First Group and estimated market value data provided by TransferMarkt. Data correct as of 16th September 2020. See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-EFLAnalysisandDashboards/EFLWingerCAMRadarDashboard

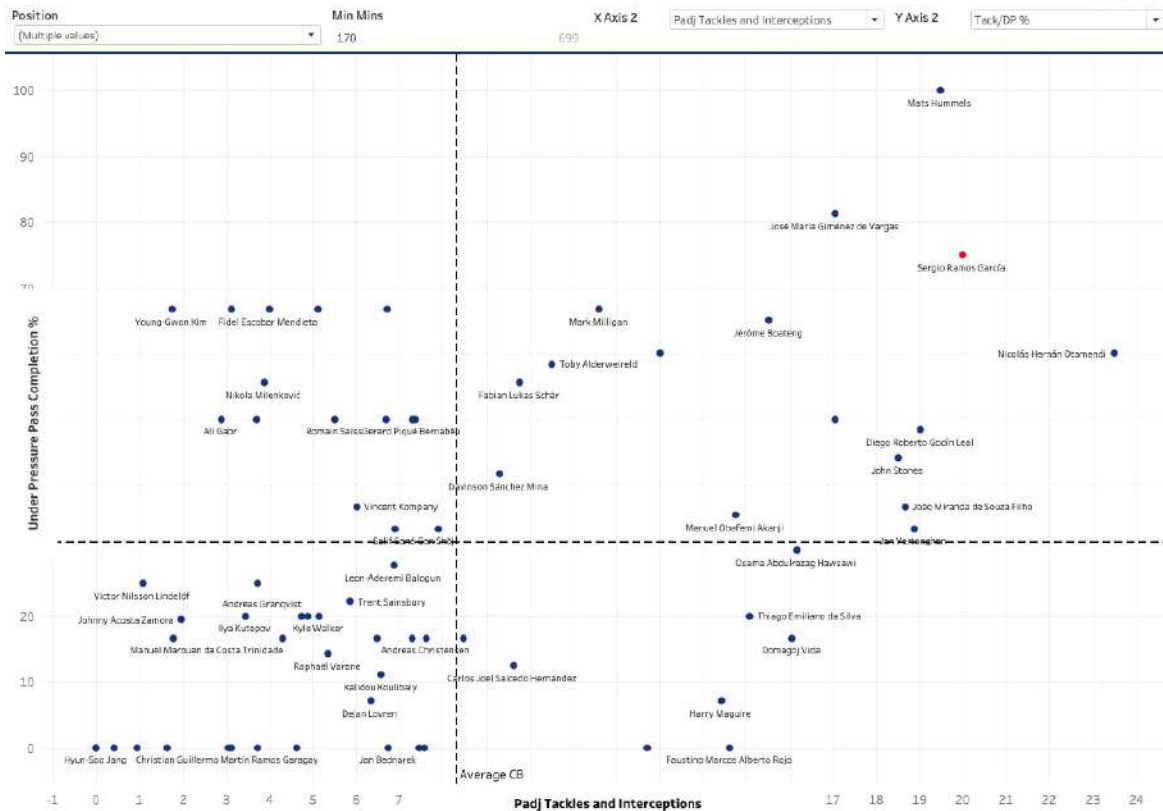
- Based on the popular Knutson radar by StatsBomb, the **Player Radar** dashboard is a way of visualising a large number of stats at one time, specific to player position, using data provided by [Twenty First Group](#).
- More so than tables, radars allow the viewer to engage with the displayed data in a easier way for the brain to process.
- Before reaching the dashboard, to be in a form to create radar plots, the previously engineered DataFrame of data from Twenty First Group was required to be further engineered and melted. See the following notebook for data preparation [\[link\]](#).
- Bespoke radars are created per position to assess the metrics of the most important KPIs for different roles in the team, all in one visualisation.
- It's important to note when viewing radars, area is not a part of the assessment, only where the points themselves lie of the axes, as illustrated in the following tweet from Luke Bornn: twitter.com/LukeBornn/status/864856335191388162.

Tool 10: Dynamic Duel-Axis Scattergram Dashboard

Example of a comparison of real-world performance and estimated market value

Dynamix Duel-Axis Scatter Plot for Ball-Playing Centre Backs

2018 FIFA World Cup



Dashboard by
@edwebster

See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-WorldCup2018AnalysisandDashboard/WC2018DynamicDuel-AxisDashboard1.

- The **Dynamic Duel-Axis Scattergram** dashboard allows for quick and easy comparison of two metrics of choice.
- In the example on the left, the dashboard compares Pass Completion % (X axis) against Under Pressure Pass Completion % (Y axis).
- The dashboard uses Event data by StatsBomb that has been fuzzy matched against player valuation data from [TransferMarkt](#).
- This dashboard can be used as one of many tools to developing a club's intelligence and processes in the transfer market, with the ability to compare any performance metric of players against estimated market values. This enables to club to shortlist good/under-valued players and lead to a positive impact on player trading outcomes. An example of the application of scattergrams can be seen in Tifo's 'Scattergrams with Tom Worville series' [\[link\]](#).

Tool 11: Percentile Rank of Players against their Peers

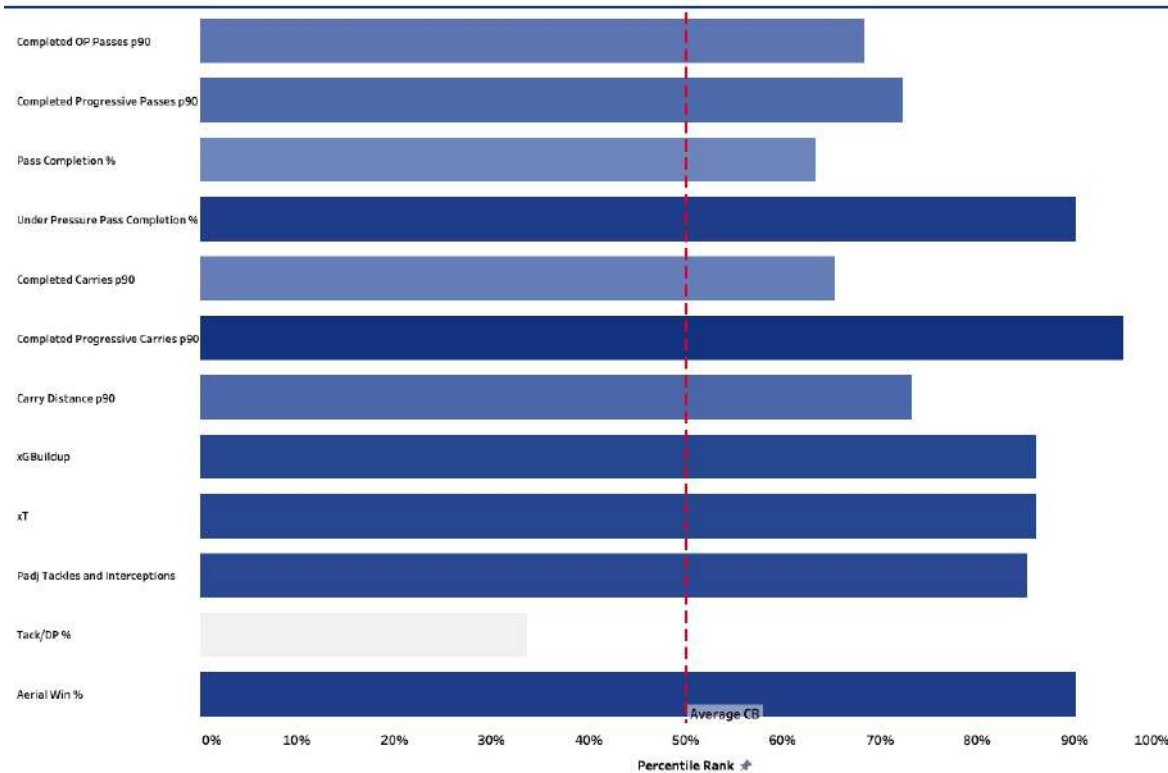
Comparison of a player's metrics against the mean

Harry Maguire Percentile Rank for a Ball-Playing Centre Back

2018 FIFA World Cup

Select a player to update the Dashboard

Player
Harry Maguire



Dashboard by
@edwebster

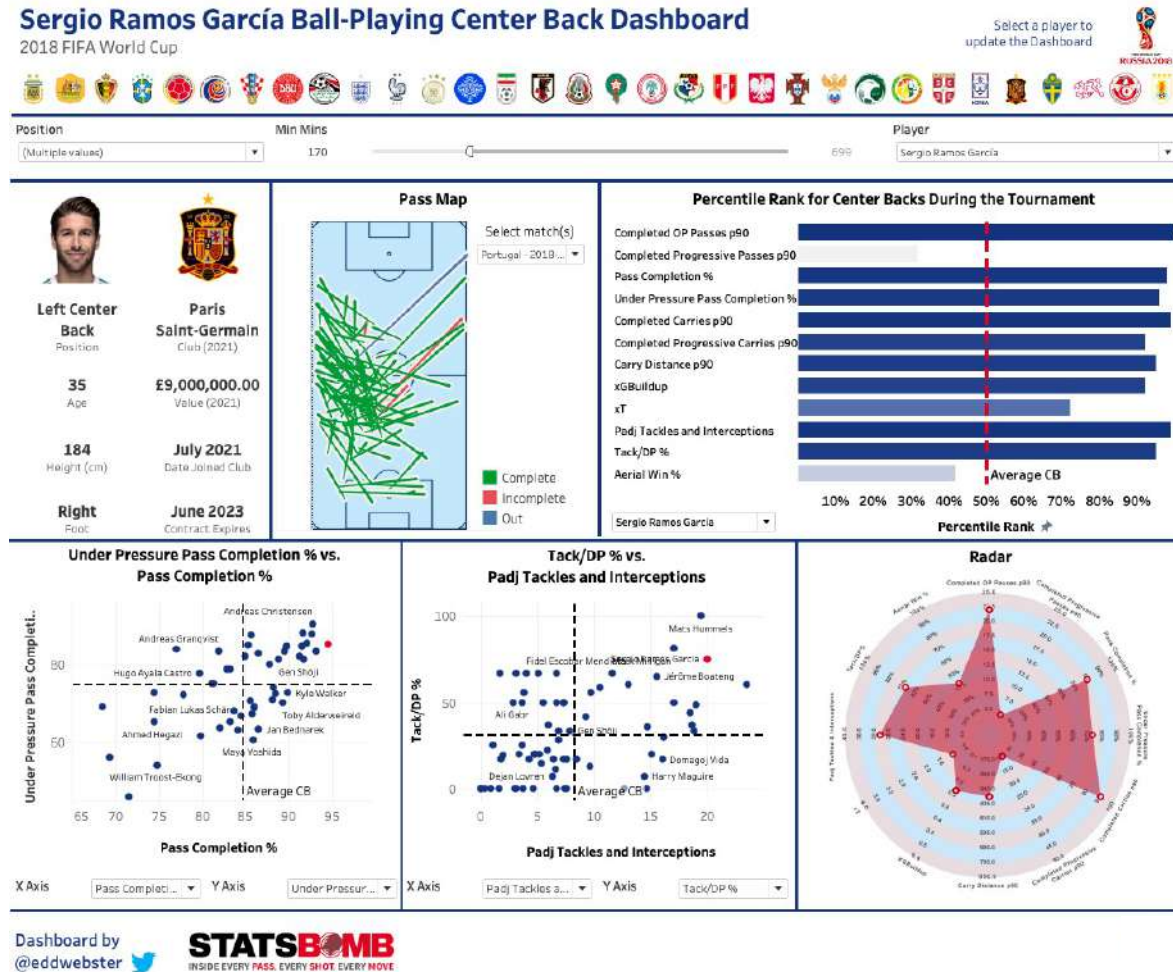


See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-WorldCup2018AnalysisandDashboard/WC2018PercentileRankDashboard.

- The **Player Percentile Rank** dashboard is a series of bar charts visualising the identified metrics against the median player performance for that position.
- A reference line of 50% indicates the average performance of a player in that position in the dataset available. Bars that are closer to the 100% mark, show qualities in players that are better than the average player in that position.
- In this example, we can see that Harry Maguire performs better than the average center back in all but one metric – xT, and scores very highly ($\geq 90\%$) in Under Pressure Pass Completion % and Tackles and Interceptions, also performing well for Pass Completion %, Carry Distance p90, and Completed Open Play Passes and Pass Completion % ($\geq 80\%$).
- Using this dashboard, it is possible to quickly gauge the metrics in which a player is doing well and less well against competing players.

Tool 12: Player Profile Dashboard

Bringing together the different elements to create a single delivery system of information



See dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-WorldCup2018AnalysisandDashboard/WC2018PlayerDashboard.

- The **Player Profile** dashboard brings together many of the previous dashboards, to create one, single point of accessible information for player analysis, making it even easier for football practitioners to engage with the data.
- The following example is for Sergio Ramos of Spain. From glancing at this dashboard we can quickly gauge a few insights for Alderweireld:
 - Right-footed center back, thirty five years of age, with two years left on his contract.
 - He outperforms the competition in eight of the twelve identified metrics, performing significantly well for xT, xGBuildup, Completed Open Play passes, Completed Carries and Carry Distance. He performs less well for Completed Progressive Passes and Aerial Win %.
- Further insight for specific metrics is accessible in the dynamic scatter plots.

Project 2: Recruitment Analysis of Center Backs

Scenario for data-driven recruitment analysis, to determine “the next Gerard Piqué”, for a hypothetical, newly-promoted club in the ‘Big 5’ European leagues

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

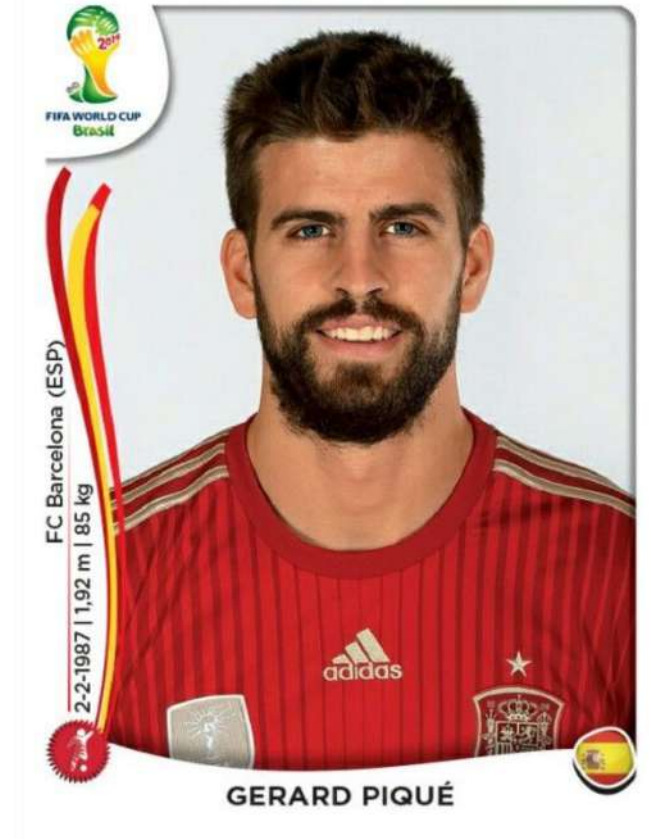
Further Work



Scenario for Data-driven Scouting and Recruitment

An example of task for processing and generating insights from football-related data

- A Sporting Director of a newly promoted club in the 'Big 5' European leagues, has challenged the club's recruitment department to find " the next Gerard Piqué", for the start of the 21/22 season. The interpretation of this statement is to be determined by the department.
- Requirements for the player as part of the the recruitment analysis are:
 - capable of playing more than 30 games a year;
 - aged less than 23 years old; and
 - due to budgetary restrictions, the player should have estimated value of <€5M and a weekly salary of <€12.5k, as per [TransferMarkt](#) and [Capology](#) respectively.
- Important considerations when providing recommendations:
 - Identification of relevant metrics,
 - The attainability of any potential targets given the club's budget and standing; and
 - Whether a potential recruit may be undervalued.
- The final recommendations should be made using data analysis and visualisation, to clearly communicate the decision making process.



A static version of the Jupyter Notebook can be found at the following:

nbviewer.org/github/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/player_similarity_and_clustering/PCA%20and%20K-Means%20Clustering%20of%20%27Piqu%C3%A9-like%27%20Defenders%20for%20Parma%20Calcio%201913.ipynb.

Publicly Available Data at Our Disposal

The four, key datasets that are publicly available that were considered to conduct this analysis

The three datasets used for this analysis.

1

Event Data



On-the-Ball Event Stream data, or, Event data, is a log of each on-ball event (passes/tackles/shots) that happens during a 90 minute match.

Examples of the data collected include: Event type, timestamp, spatial location, meta information of on-the-ball actions.

Key providers include:

[StatsBomb](#), [WyScout](#), [Stats Perform](#) ([Opta](#)).

2

Aggregated Player Performance



Aggregated player performance data, derived from Event data, provided by [StatsBomb](#).

Dataset covers all male 'Big 5' European leagues and MLS from the 17/18 season-present.

Data contains nearly 200 metrics for the categories including: passing, playing time, chance creation, shooting, and defensive actions.

3

Player Bio and Estimated Valuation



Player bio, contractual and estimated value data.

TransferMarkt uses a “wisdom-of-the-crowds” approach to estimating the value of players¹. These estimates are correlated to but biased estimates of the actual fees paid.

4

Player Salaries



Player salary and finance data provided by Capology.

Dataset covers all male 'Big 5' European leagues, as well as the MLS, Championship, Eredivisie, Belgian First Division, Primeira Liga, Süper Lig, Ukraine Premier League², from the 13/14 season-present.

¹See the paper 'The Wisdom of Crowds and Transfer Market Values' by Dennis Coates and Petr Parshakov for more information: [sciencedirect.com/science/article/abs/pii/S037722172100895X](https://www.sciencedirect.com/science/article/abs/pii/S037722172100895X).

²The Scottish Premier League, Liga MX Brasileiro, and Argentine Primera Division are not available without a license.

StatsBomb Event Data vs. FBref Aggregated Data

Reasoning for the selection of FBref data as the appropriate player performance data source for this task

- The choice of the player statistics dataset was between [StatsBomb](#) Event data and [FBref](#) aggregated player performance data¹.
- The StatsBomb dataset only features one complete male competition - the 2018 FIFA Men's World Cup, featuring only 102 center backs, of which only 17 are under the age of 23 (see brief).
- Players in the World Cup at most, played 7 matches (if they reached the final or 3rd placed play off). This is a restrictive dataset to conduct recruitment analysis, especially to determine players that can play 30 matches per season (see brief).
- The FBref dataset has over 2,500 defenders from the 'Big 5' European leagues, including Serie A, over the last four complete seasons. The dataset is also very easy to work with and requires less manipulation and data engineering before analysis. It is however, a less rich dataset and not as easy from which to derive bespoke metrics, such as Possession Adjusted defensive metrics and Expected Threat (xT)².



¹For more information of the debate between StatsBomb Event data vs. FBref Aggregated Player Performance data and the pros and cons, see the Appendix [\[link\]](#).

²For an example of where I worked with the StatsBomb Event data to derive metrics such as xT, see my StatsBomb Data Parsing [\[link\]](#) and StatsBomb Data Engineering [\[link\]](#) notebooks, and/or visualisation of the resulting dataset as a set of Tableau dashboards [\[link\]](#).

Rules-Based vs. a Machine Learning Based Approach

The two possible approaches I considered to tackle this challenge

Rule-Based



Determine the KPIs most relevant to Piqué through observation / data analysis. With identified KPIs, determine the 'best-scoring' players in those metrics, within the defined limits of the brief.

ML-Driven



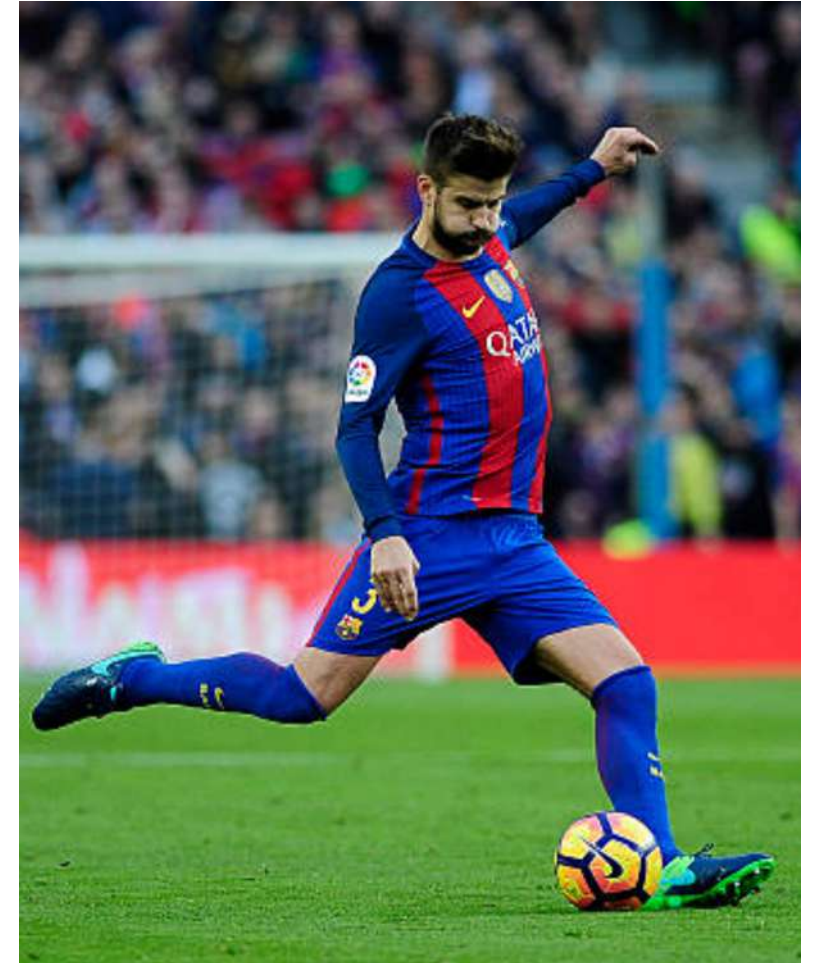
Application of Machine Learning techniques to the dataset, in an attempt to determine hidden patterns in the data for Piqué's style of play. Metrics of interest are then analysed subsequently for the shortlisted candidates.

Why was this approach taken? See the following slide...

Hidden Patterns in the Data Can Be Found Using Machine Learning

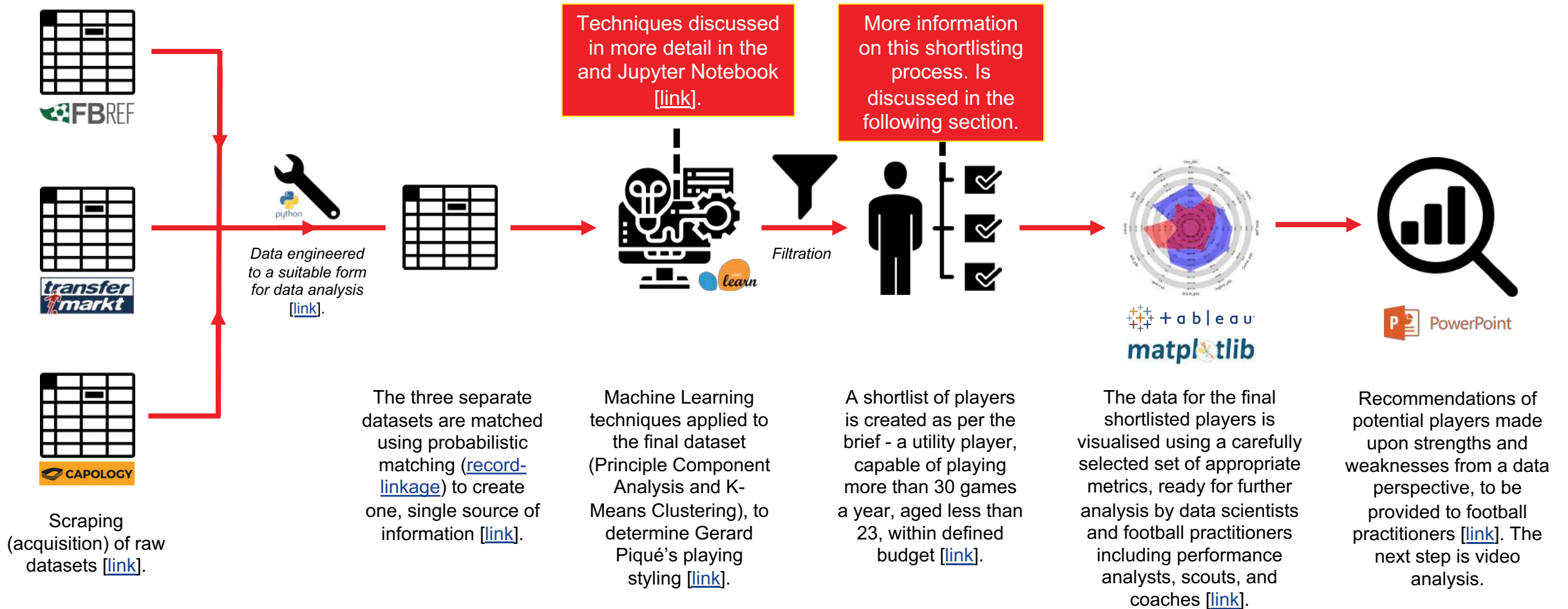
Why the application of Machine Learning was decided for this use case?

- Gerard Piqué is a "ball-playing center back" that plays a "progressive style of football".
- This qualitative statement of characteristics determined from watching him play in training, live in the stadium, or on television.
- However, what if we want to try and determine a player's style of play quantitatively, using data?
- The goal of applying Machine Learning in this task is to be able to autonomously label the playing style all players in the dataset, through data. ML is great at finding patterns of similar characteristics, in this case, players and their statistical output, as a proxy for playing style.
- The 'playing style' of Piqué has been attempted to be determined using two Machine Learning techniques - Principal Component Analysis (PCA) and K-Means clustering (see following slides for Methodology and theory).



Full Methodology

Working from first principles to process and generate insights from football-related data



Full details of the methodology can be found in the Jupyter notebook: nbviewer.org/github/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/player_similarity_and_clustering/PCA%20and%20K-Means%20Clustering%20of%20%27Pique%CC%81-like%27%20Defenders.ipynb.

What is Principle Component Analysis (PCA)?

A brief summary and how it is used in this model

- Principal Component Analysis (PCA), is a dimensionality-reduction method, used to reduce the number of features a dataset into a more manageable number, whilst still containing most of the information in the large set.
- In the case of our analysis, a dataset with 77 dimensions of relevant ball-playing center back metrics, was reduced to 2.
- The number of dimensions in which the dataset was to be reduced to was determined using the Explained Variances plot (see fig), which shows that the first two components explain just over 55% of the variance in the dataset.
- For the use case of visualisation, this also works well as two dimensions can be easily plotted visually.
- Reducing the number of variables of a dataset naturally reduces the accuracy, but by trading a little accuracy for simplicity, smaller datasets are easier to analyse, visualise, and make for faster applications of machine learning.

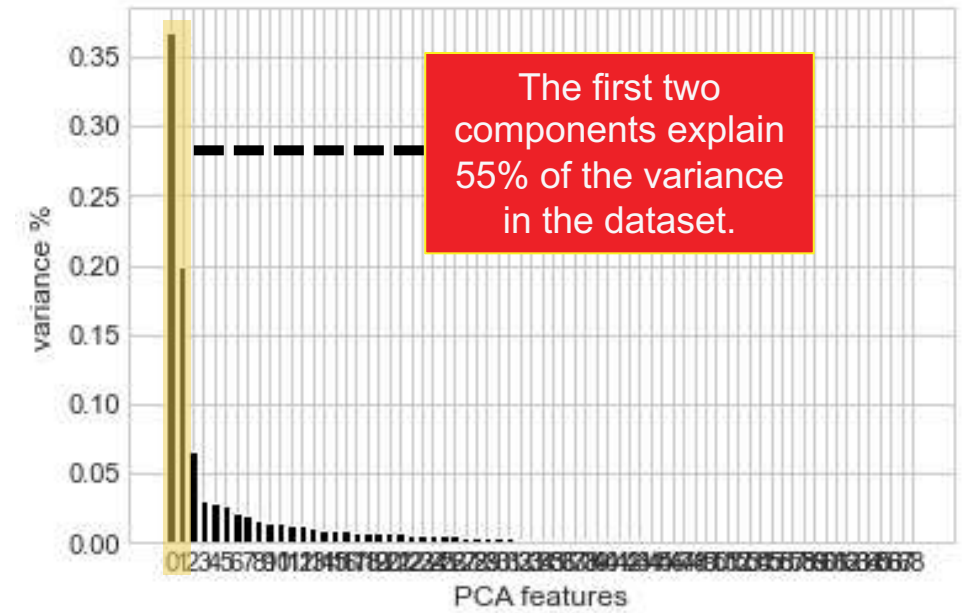


Fig: Explained Variances plot, showing that the first two components explain the majority of the variance in the data.

What is Clustering and Specifically, K-Means Clustering?

A brief summary and how it is used in this model

- The objective of clustering is to group similar data points together and discover underlying patterns. K-Means clustering, the algorithm selected for this analysis, is able to achieve this by determining a fixed number (K) of clusters in a dataset.
- A cluster refers to a collection of data points aggregated together because of certain similarities.
- Every data point is allocated to each of the defined clusters by reducing the in-cluster sum of squares.
- The process is to first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative calculations to optimise the positions of the centroids
- The optimum number of clusters can be determined using the Elbow Method (Fig. A). The value of K can be found at the 'elbow' i.e. the point after which the distortion/inertia start decreasing in a linear fashion
- The plot for this dataset showed no particular elbow, so I elected to go with five clusters (Fig. B).

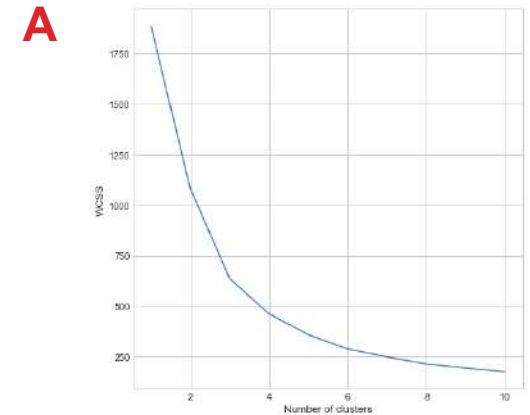


Fig. A: Elbow plot to determine the number of clusters for K-Means.

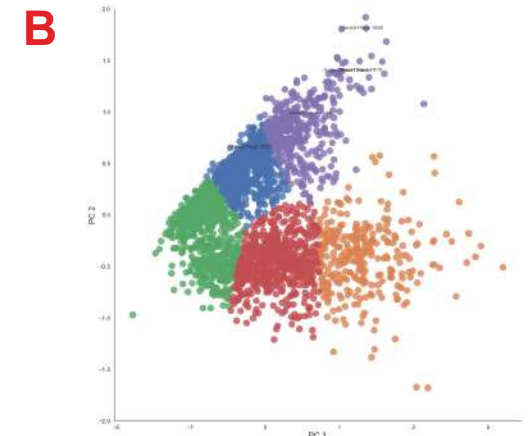


Fig. B: K-Means clustering of players into five groups.

How is 'Piqué Similarity' Determined?

Distance between the data points of Piqué and players as a proxy for similar playing characteristics



- The following figure is a zoomed visualisation of the results from the cluster analysis, where data points were allocated to one of five clusters.
- Piqué's data points for the 17/18-19/20 seasons are found in the purple cluster, whereas his data point for the most recent complete season (20/21) is found in the blue cluster. This indicates that Piqué's output for his most recent season, is different, for whatever reason.
- The visualisation also plots the centroid of the triangle for Piqué's 17/18-19/20 datapoints, which happens to lie very close to the 18/19 data point.
- This centroid and 18/19 data point are used as reference point.
- The distance between the data point of each play and this centroid is used as a proxy for player similarity, suggesting that if players are in the same cluster as Piqué and have a data point close to him, they are more likely to play like Piqué.

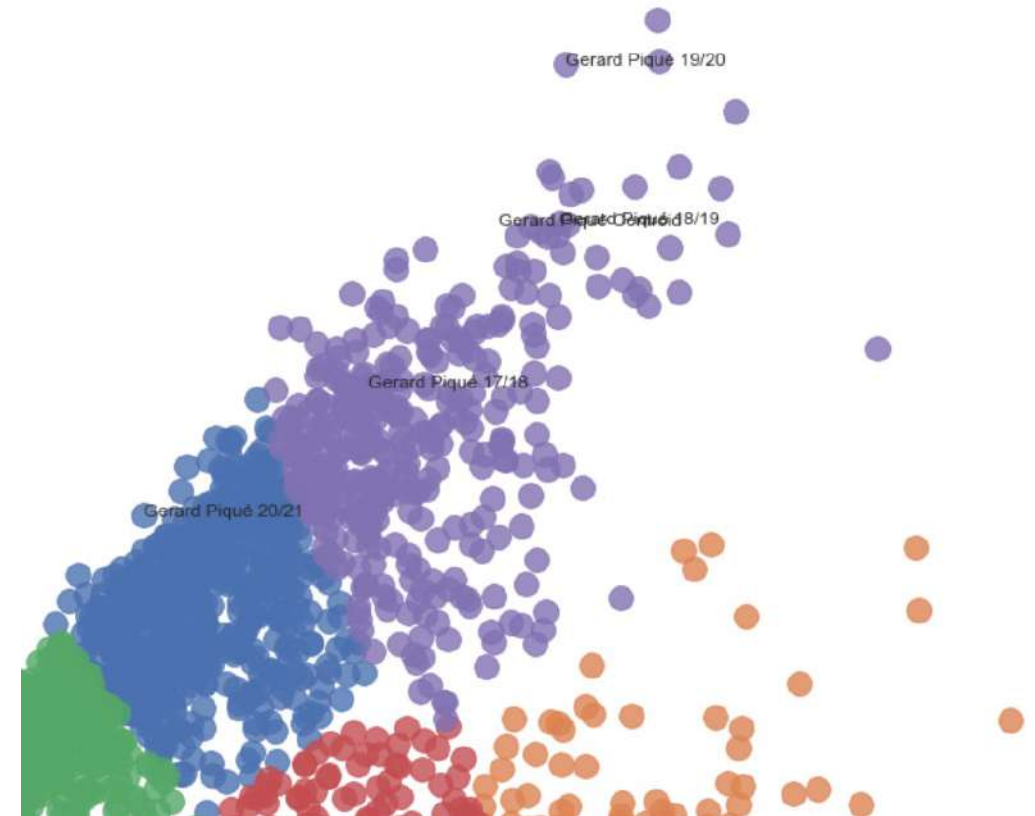


Fig: Visualisation of the K-Means clustering of players, focussing on cluster 4 (purple) and cluster 0 (blue). The distance between the data point of each player and the Piqué 17/18-19/20 triangle centroid is used as a proxy for similarity.

Advantages and Limitations of Logistic Regression Models

Strengths and weakness of using K-Means for clustering



Advantages

- K-Means clustering is relatively simple to implement.
- The K-Means algorithm is flexible and can adjust to changes through adjustment of the cluster segment.
- K-Means scales to large data sets and computes much faster than the smaller dataset.
- K-Means can warm-start the positions of centroids.
- K-Means clustering easily adapts to new examples.
- K-Means clustering has low computation cost compared to other clustering methods.
- K-Means clustering generalises to clusters of different shapes and sizes, such as elliptical clusters.
- The results are easy to interpret, generating cluster descriptions in a form minimised to ease understanding of the data.

Limitations

- The number of K-values specified must be chosen manually (see Elbow plot).
- Selection of the initial centroids is random.
- K-Means lacks consistency, producing varying results on different runs.
- K-Means has trouble clustering data where clusters are of varying sizes and density.
- Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Outliers therefore need to be treated before clustering.

Four Key Areas for Consideration

The requirements for suitable recruitment candidates

“The recruitment department is looking for a center back, capable of playing more than 30 games a year, aged less than 23 years old, and is within budget.”

Style



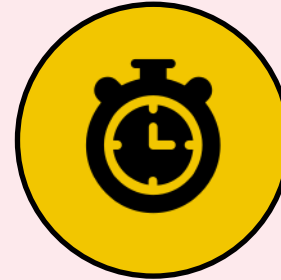
The player is a “ball-playing center back” that plays a “progressive style of football”, determined with ML.

Age



The player is aged 23 years or less.

Playing Time



The player is capable of playing at least 30 full matches per season (2,700 minutes).

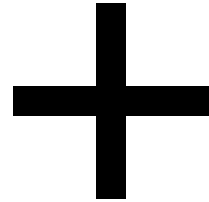
Financial



The player can fit within the team’s budget, i.e. an estimated value of <€5M and a weekly salary of <€12.5k.



**Defined Playing
Style**



Data



KPIs

Different Qualities are Important for Different Roles

An approach for tailoring the abilities of players quantitatively



- Depending on the playing style and position, the qualities required in a potential recruit need to be carefully selected during the assessment process.
- As previously mentioned, Gerard Piqué is a ball-playing center back that plays a progressive style of football.
- Therefore, Metrics or Key Performance Indications (KPIs)¹ from which the final shortlisted players are assessed, need to be appropriate for players that can get on the ball and be progressive in their play.
- Defensive KPIs such as Clearances, that might be deemed a good measure of performance for a standard center back, may not be as well-suited as part of the recruitment process for a more progressive center back, with metrics such as Completed Progressive Passes (Fig. A) and Completed Progressive Carries (Fig. B) potentially being more suitable, as well as advanced Expected Metrics (xG, npxG, xA)².

If you are viewing the PDF version of this presentation, the interactive GIFs can be viewed in the online version: docs.google.com/presentation/d/1a8YgbsXTbpT8FYDvmi-J2eXgN2xnOkUS2E-8YsWrInK or on GitHub: github.com/eddwebster/football_analytics/tree/master/gif/fig/pique.

¹The words 'KPIs' and 'Metrics' are used interchangeably in this presentation.

²Ideally, if available in FBref data, the Expected Threat (xT) metric would be used for this analysis, which is more appropriate for defenders than xG and xA..

A



Fig. A: Piqué's Progressive Pass and Assist to Zlatan Ibrahimović.

B



Fig. B: Piqué's Carry and Goal against Real Madrid.

Requirements of Applied Metrics

How standard playing metrics are engineered for our needs



- Metrics are also required to be standardised, to allow players that have had less minutes on the field or had less time on the ball, to be compared to first-team regulars.
- For attacking metrics, a per 90 (p90) standardisation has been implemented.
- For defensive metrics, ideally an possession adjustment (PAdj) standardisation would be applied. However, this is not currently available with FBref data that was used for this analysis¹.
- With data mainly focused towards on-ball events, it can be difficult to analyse defenders. This is because defensive moments won't be dictated by the team but by their opponents.
- If you solely look at total statistics for defensive actions for the best players in the world at the best teams, they are regularly lower down on lists (see fig). Examples of these metrics are the number of Tackles and Interceptions.
- A way to negotiate this is to look at % success rates, allowing for the comparison of samples with a different number of observations. Success rates show, as a percentage, how many times the player wins individual battles. Examples of these metrics are Aerial Win% and Tacklers Dribbled%.

Possession adjustment metrics (PAdj) are not currently available in the FBref data, but hope to soon be added. See: fbref.com/en/about/scouting-reports-explained. This is something however that can be determined with Event data, if available for the 'Big 5' European leagues. For code to see this, see: nbviewer.org/github/eddywebster/football_analytics/blob/master/notebooks/3_data_engineering/StatsBomb%20Data%20Engineering.ipynb.

FBREF Virgil van Dijk Scouting Report

Last 365 Days [View Complete Scouting Report](#) Share & Export ▾

vs. Center Backs		
Statistic	Per 90	Percentile
Non-Penalty Goals	0.00	16
npG	0.08	90
Shots Total	1.00	97
Assists	0.07	92
xA	0.01	27
npG+xA	0.09	80
Shot-Creating Actions	0.50	46
Passes Attempted	70.71	89
Pass Completion %	92.7%	93
Progressive Passes	3.14	68
Progressive Carries	3.79	76
Dribbles Completed	0.07	12
Touches (Att Pen)	1.29	88
Progressive Passes Rec	0.21	65
Pressures	3.14	1
Tackles	0.79	3
Interceptions	1.07	7
Blocks	1.14	6
Clearances	5.00	50
Aerials won	3.71	87

Fig: FBref Scouting Report for Virgil van Dijk considered by many as the best center back in the world. He features in the bottom 5-10% of center backs for tackles, interceptions, and blocks.

Standard Center Back Metrics

As per StatsBomb's standard center back template of eleven metrics



As per the [Knutson](#) radar¹, the standard eleven metrics used as part of StatsBomb's data-driven assessment of a center back's ability are:

In-possession (3):

1. **Pass Completion Percentage (Passing%)**: the number of completed passes divided by the number of attempted passes.
2. **Unpressured Long Balls**: the number of completed long balls while not under pressure
3. **Expected Goals Buildup (xGBuildup)**: the total xG of every possession the player is involved in, minus shots and key passes (xGChain includes shots and key passes).

Out-of-Possession (8):

1. **Pressures**: the number times a player applies pressure to opposing player who is receiving, carrying or releasing the ball.
2. **Fouls**: the number of fouls.
3. **Tackles and Dribbles Past (Tack/DP%)**: the ratio of the number of tackles a player makes compared with the number of times they are dribbled past.
4. **Possession adjusted Tackles (PAdj Tackles)**: the number of tackles adjusted proportionally to the possession volume of a team.
5. **Possession adjusted Interceptions (PAdj Interceptions)**: the number of interceptions adjusted proportionally to the possession volume of a team.
6. **Aerial Wins**: the number of aerial duels a player wins.
7. **Aerial Win%**: the percentage of aerial battles won by total aerial battles.
8. **Clearances**: the number of times a player makes a clearance or plays a long ball while under pressure.



¹[Ted Knutson](#) is responsible for popularising the radar plot amongst the football analytics community. StatsBomb radars have six templates, depending on position, the one illustrated being that for center backs and in this specific example, Harry Maguire. For more information about StatsBomb radars, see: statsbomb.com/2018/08/new-data-new-statsbomb-radars

Bespoke Ball-playing Center Back Metrics

The twelve selected metrics and their definitions



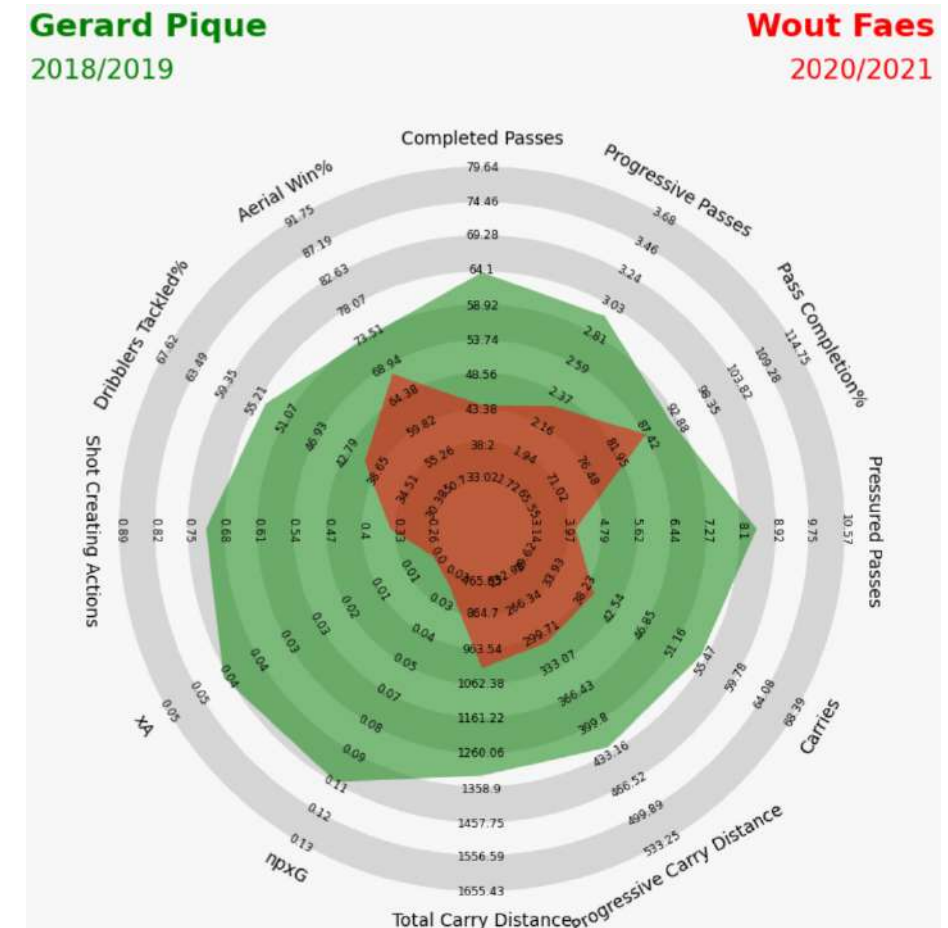
In-possession (10):

1. **Completed Passes p90**: the number of completed passes, per 90 minutes;
2. **Progressive Passes p90**: the number of completed progressive passes, per 90 minutes. A Progressive Pass is defined as a pass that moves the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. This excludes passes from the defending 40% of the pitch.
3. **Pass Completion%**: the percentage of passes successfully played;
4. **Pressured Passes p90**: the number of passes made under pressure, per 90 minutes;
5. **Carries p90**: the number of completed carries, per 90 minutes;
6. **Progressive Carry Distance p90**: the total distance of progressive carries, per 90 minutes. A Progressive Carry is defined as a Carry that moves the ball closer to the goal by 25% or that gets the ball into the box;
7. **Total Carry Distance p90**: the average Carry length in meters, per 90 minutes;
8. **Non-Penalty Expected Goals (npG) p90**: the total non-penalty xG the player creates, per 90 minutes;
9. **xG Assisted (xA) p90**: xG which follows a pass that assists a shot, per 90 minutes; and
10. **Shot-Creating Actions (SCA) p90**: The two offensive actions directly leading to a shot, such as passes, dribbles and drawing fouls per 90 minutes. Note: A single player can receive credit for multiple actions and the shot-taker can also receive credit.

Out-of-Possession (2):

1. **Tacklers Dribbled%**: the ratio of the number of tackles a player makes to the number of times they are dribbled past; and
2. **Aerial Win%**: the percentage of aerial battles successfully contested.

The twelve metrics observed were selected based on 1) Data available, notably no possession adjusted defensive metrics, and 2) the StatsBomb radar template for center backs. For more information about StatsBomb radars, see statsbomb.com/2018/08/new-data-new-statsbomb-radars.



Determining Match Availability

Identifying players that are capable of playing more than 30 games a year



Player	Matches Played	Starts	Mins	90s	Mins/MP	Mins%	Mins/Start	Subs	Mins/Sub	unSub
Bruno Manga	38	38	3,420	38	90	100	90	0	n/a	0
Sven Botman	37	37	3,309	36.8	89	96.8	89	0	n/a	1
Jose Fonte	36	36	3,185	35.4	88	93.1	88	0	n/a	0
Nayef Aguerd	35	35	3,147	35	90	92	90	0	n/a	0
Matthias Ginter	34	34	3,060	34	90	100	90	0	n/a	0
Harry Maguire	34	34	3,047	33.9	90	89.1	90	0	n/a	0
Jules Kounde	34	33	2,975	33.1	88	87	90	1	10	1
Pau Torres	33	33	2,970	33	90	86.8	90	0	n/a	3
Tosin Adarabioyo	33	33	2,953	32.8	89	86.3	89	0	n/a	1
Bremer	33	33	2,935	32.6	89	85.8	89	0	n/a	4
Wout Faes	33	33	2,899	32.2	88	84.8	88	0	n/a	0
Duje Ćaleta-Car	33	32	2,813	31.3	85	82.3	88	1	2	0
Robin Le Normand	33	30	2,747	30.5	83	80.3	89	3	26	3
Amos Pieper	30	30	2,682	29.8	89	87.6	89	0	n/a	0
Maxence Lacroix	30	29	2,590	28.8	86	84.6	89	1	13	2
Clement Lenglet	33	29	2,475	27.5	75	72.4	84	4	11	2
Roger Ibanez	30	28	2,398	26.6	80	70.1	84	2	28	1
Willi Orban	29	25	2,363	26.3	81	77.2	90	4	28	4
Manuel Akanji	28	26	2,350	26.1	84	76.8	89	2	25	0
Dayot Upamecano	29	27	2,330	25.9	80	76.1	84	2	37	1
Wesley Fofana	28	27	2,262	25.1	81	66.1	83	1	25	1
Oscar Mingueza	27	23	1,901	21.1	70	55.6	80	4	15	5

- The following table shows the **match availability** for a sample of the 118 players have been determined as having a similar playing style as Piqué in 20/21¹, filtered for players that played at least 50% of the available minutes.
- Physical data is not available across leagues, so instead, the percentage of minutes played (Mins%) provides an insight into a player's physical capabilities over a match, a season and in different competitions.
- Knowing how often a player is available, when they're selected, and how many minutes they play, can allow us to understand what they are capable of in terms of fixture schedule as well as factors such as unknown injuries.
- In the real world, this analysis would then be supported with video analysis / live scouting, to build player profiles.

¹As per the Principle Component Analysis (PCA) and K-Means Clustering. Players in the same cluster as Piqué, with data points that lie close to Piqué are categorised as having a similar playing style.

Cutting Out Unrealistic Targets

Selecting only the suitable players for further analysis



Player	Squad	League	Minutes Played	Age ¹	Estimated Value (€) ¹	Weekly Salary (€) ²
Rúben Dias	Manchester City	Premier League	2,843	23	60.0M	84.7k
Benoît Badiashile	Monaco	Ligue 1	2,870	19	25.0M	11.1k
Sven Botman	Lille	Ligue 1	3,309	20	10.0M	7.1k
Loïc Badé	Lens	Ligue 1	2,610	20	27.0M	1.7k
Jules Koundé	Sevilla	La Liga	2,975	21	36.0M	21.1k
Dayot Upamecano	RB Leipzig	Bundesliga	2,330	21	60.0M	48.1k
Pau Torres	Villarreal	La Liga	2,970	23	40.0M	22.1k
Maxence Lacroix	Wolfsburg	Bundesliga	2,590	20	12.0M	3.6k
Tosin Adarabioyo ³	Fulham	Premier League	2,953	22	5.0M	47.3k
Nico Elvedi	M'Gladbach	Bundesliga	2,536	23	30.0M	30.8k
Alessandro Bastoni	Inter	Serie A	2,922	21	50.0M	99.6k
Edmond Tapsoba	Bayern Leverkusen	Bundesliga	2,652	21	30.0M	21.1k
Duje Ćaleta-Car	Marseille	Ligue 1	2,813	23	23.0M	38.8k
Wout Faes	Stad Rennes	Ligue 1	2,899	22	3.0M	8.1k
Amos Pieper	Arminia	Bundesliga	2,682	22	3.5M	1.5k
Nikola Milenković	Fiorentina	Serie A	3,034	22	28.0M	28.5k
Wesley Fofana	Leicester City	Premier League	2,262	19	27.0M	19.2k
Robin Le Normand	Real Sociedad	La Liga	2,747	23	15.0M	21.3k
Bremer	Torino	Serie A	2,935	23	14.0M	17.9k
Roger Ibañez	Roma	Serie A	2,398	21	20.0M	29.6k
Óscar Mingueza	Barcelona	La Liga	1,901	21	2.0M	4.2k

- From the final shortlist of players that meet the playing-style and age requirements, 90% of these players have an estimated value and weekly salary of the club's budget. This is not surprising as Gerard Piqué played at the highest level for over a decade and therefore, players that have a similar output, will in turn be highly valued.
- Within the thirty man shortlist, there are three players that meet the budgetary requirements:
 - Wout Faes [[TransferMarkt](#)] [[FBref](#)]
 - Amos Pieper [[TransferMarkt](#)] [[FBref](#)]
 - Óscar Mingueza [[TransferMarkt](#)] [[FBref](#)]
- These are the three players to be taken forward for further recruitment analysis.

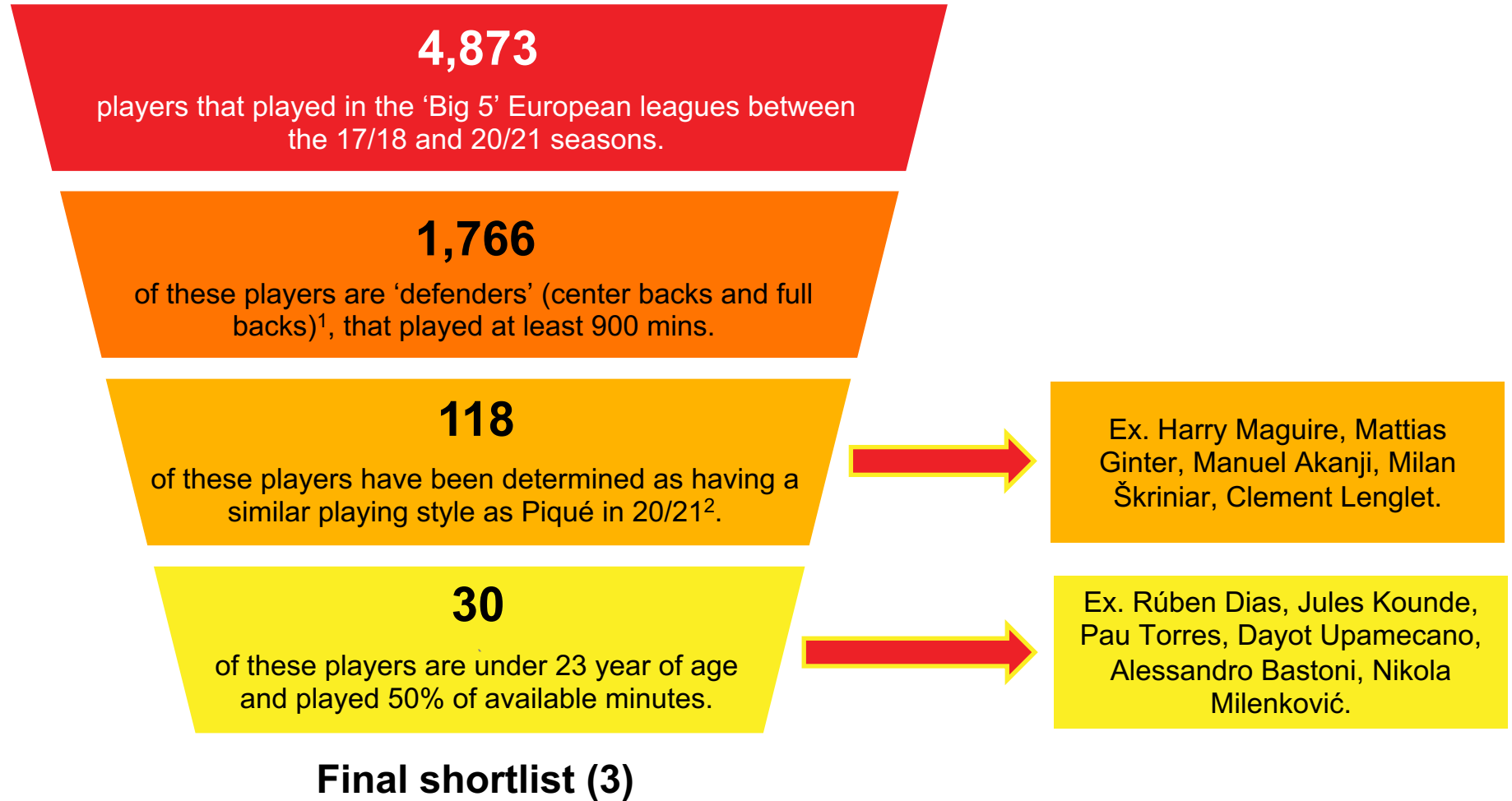
¹Ages and estimated values of the players as per December 2020. Source: [transfermarkt.com](#)

²Weekly Salary for the 20/21 season. Source: [capology.com](#)

³Tosin Adarabioyo meets all requirements except for his weekly salary, which is estimated to be €47,300 per week, more than three times the maximum salary estimated to be suitable.

Summary of the Shortlisting Process

Filtration of the full player dataset to find 'Piqué-like' players



¹As per FBref.

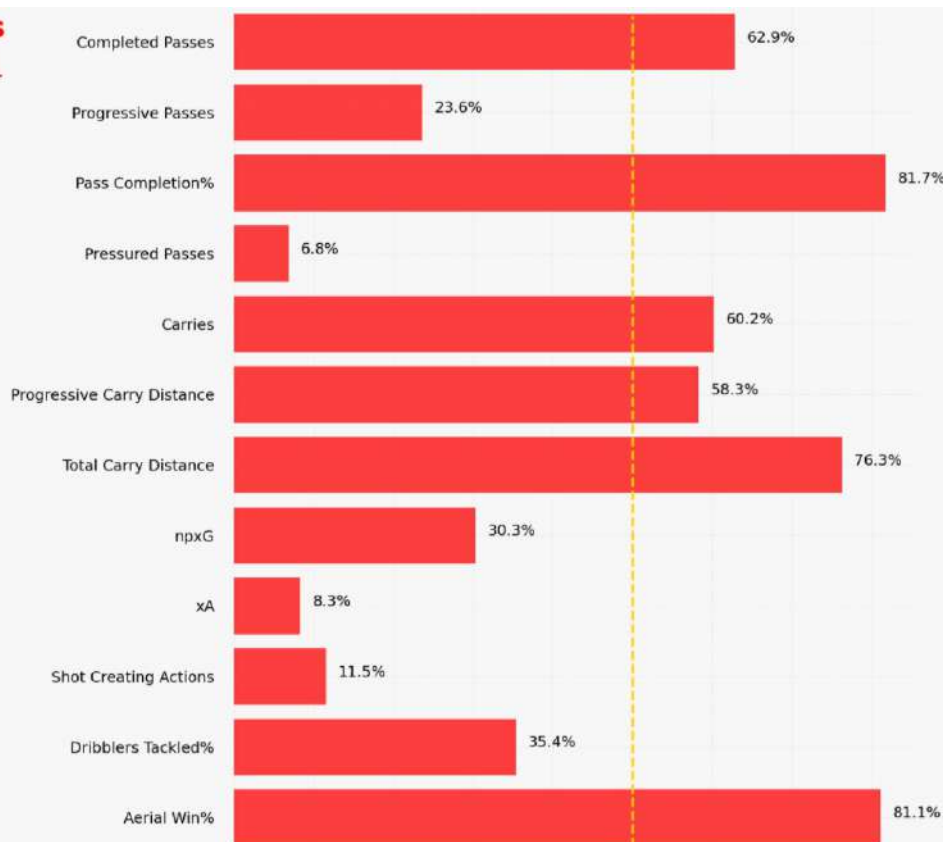
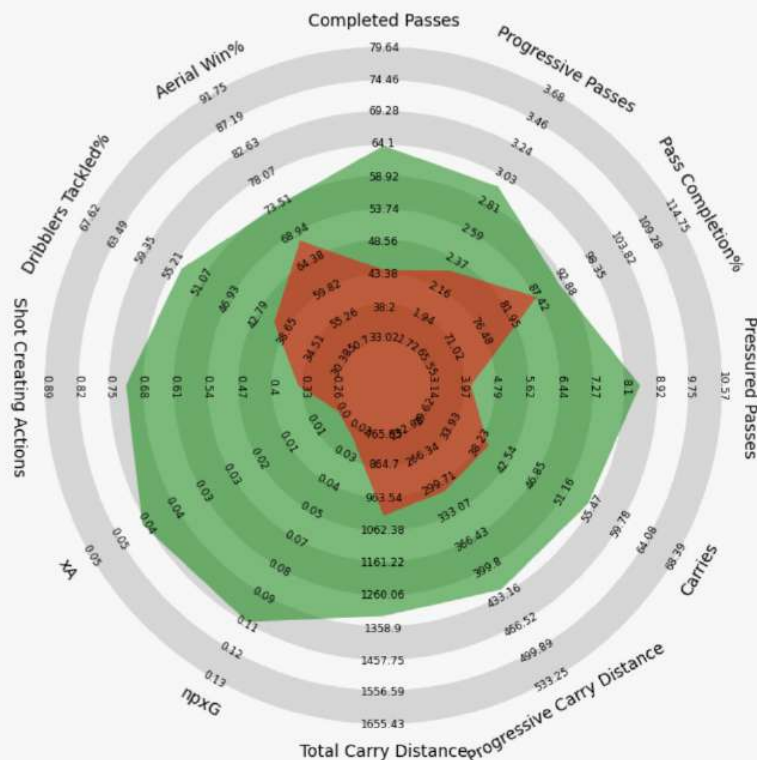
²As per the Principle Component Analysis (PCA) and K-Means Clustering. Players in the same cluster as Piqué, with data points that lie close to Piqué are categorised as having a similar playing style.

Wout Faes

Dominant in the air with a high level of pass completion

Gerard Pique
2018/2019

Wout Faes
2020/2021



Profile

- **Age:** 22
- **Position:** Center Back
- **Team:** Stade Rennes
- **Nationality:** Belgian
- **Foot:** Right
- **Height:** 187cm
- **Matches Played (20/21):** 33
- **Minutes Played (20/21):** 2,899 (84.8%)
- **Estimated Value (€), as per TransferMarkt (December 2020):** €3.0M
- **Weekly Gross Base Salary:** €8,077
- **Year Until Contract Expiry:** 2



Summary Analysis:

- + Most similar player of the three to Piqué.
- + He meets the requirements of age and playing time - 22 years old and played 33 matches in 20/21, totalling 2,899 minutes (84.8%).
- + Strong numbers for Aerial Win% and Pass Completion%, as well as Completed Passes p90, Carries, and Total Carry Distance p90.
- Less strong numbers for Pressured Passing p90, Shot Creating Actions, npxG, and xA,
- 2 years remaining on his contract, not available for a free transfer.

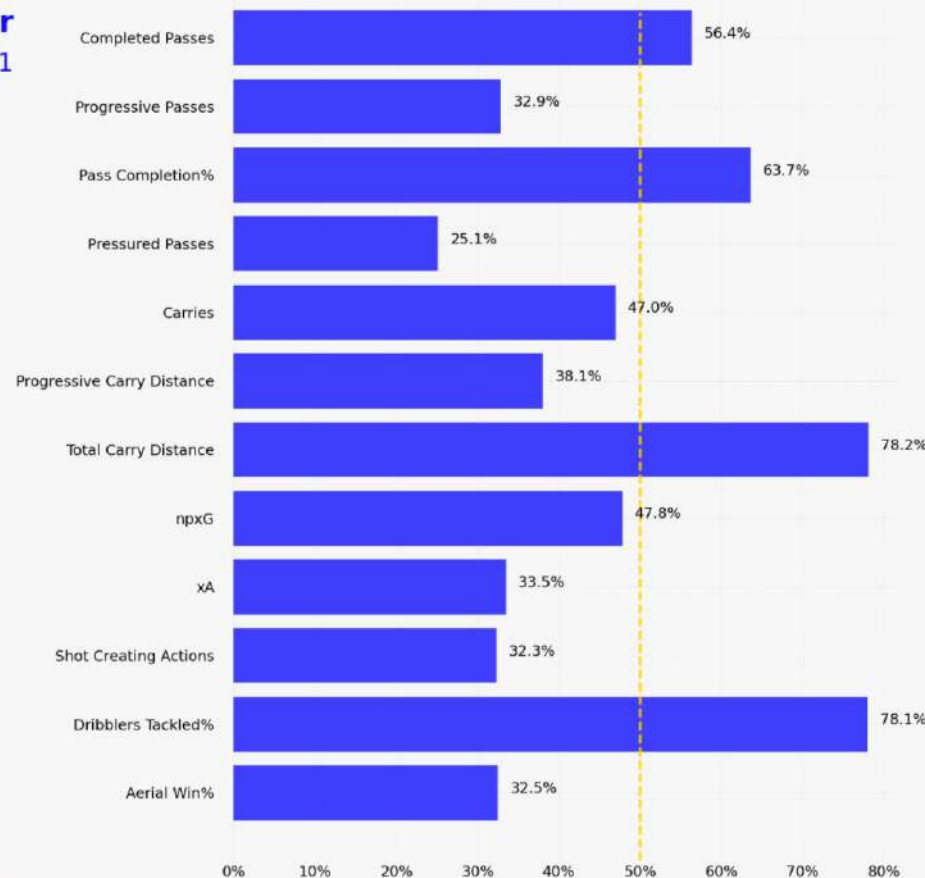
Radar graphics made courtesy of the soccerplots Python package by [Anmol Durgapal](#). For all graphics as PNGs, see: github.com/eddwebster/football_analytics/tree/master/img/fig/pique.

Amos Pieper

Well-rounded passer, carrier and tackler, that is potentially available on a free transfer

Gerard Pique
2018/2019

Amos Pieper
2020/2021



Profile

- **Age:** 22
- **Position:** Center Back
- **Team:** Arminia Bielefeld
- **Nationality:** German
- **Foot:** Right
- **Height:** 192cm
- **Matched Played (20/21):** 30
- **Minutes Played (20/21):** 2,682 (87.6%)
- **Estimated Value (€), as per TransferMarkt (December 2020):** €3.5M
- **Weekly Gross Base Salary:** €1,538
- **Year Until Contract Expiry:** 0



Summary Analysis:

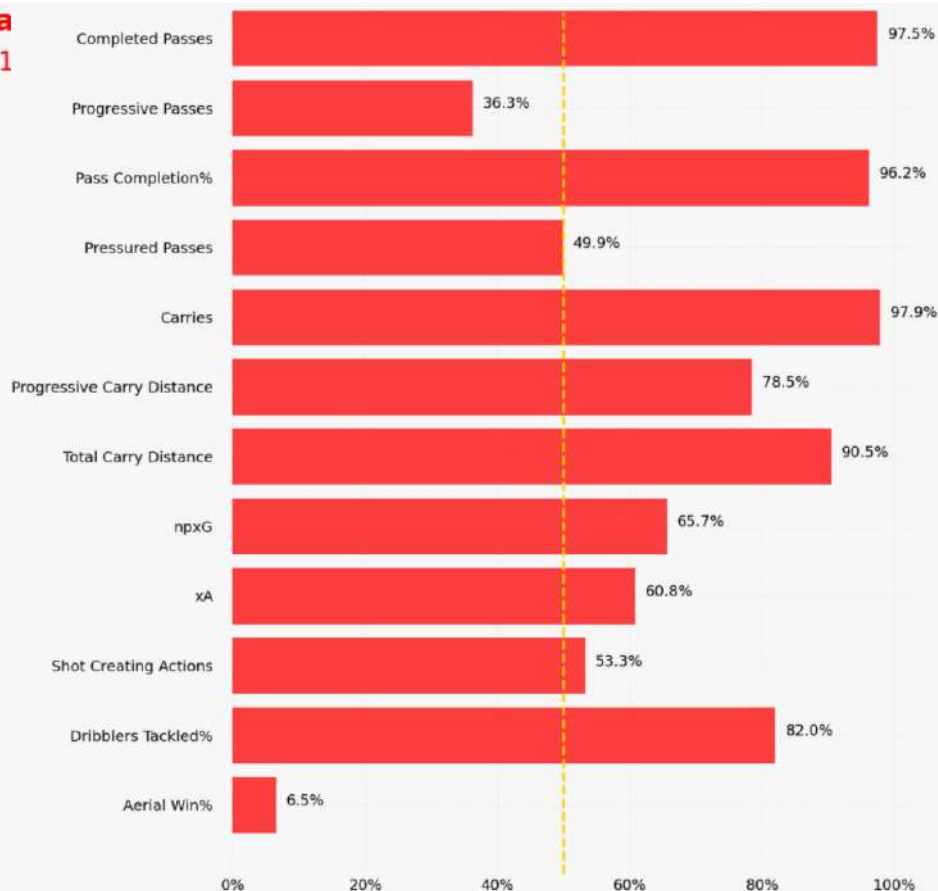
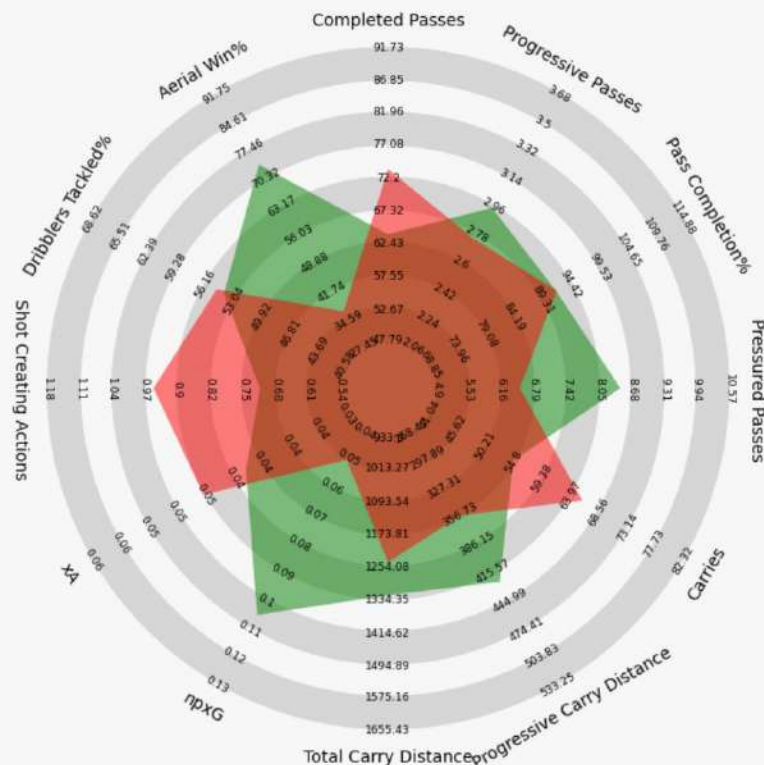
- + He meets the requirements of age and playing time - 22 years old and played 30 matches in 20/21, totalling 2,682 minutes (87.6%).
- + 0 years remaining on his contract - free transfer?
- + Strong numbers for Completed Passes p90, Completed Passing%, Total Carry Distance p90, and Dribblers Tackled%.
- Least strongest overall numbers of the three players, performing less well with Progressive Passes p90, Pressured Passes p90, Progressive Carries p90, xA, Shot Creating Actions p90, and Aerial Win%.

Óscar Mingueza

The best performing center back available, but an unrealistic transfer for a newly promoted club

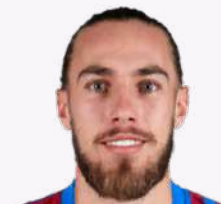
Gerard Pique
2018/2019

Oscar Mingueza
2020/2021



Profile

- **Age:** 21
- **Position:** Center Back
- **Team:** Barcelona
- **Nationality:** Spanish
- **Foot:** Right
- **Height:** 184cm
- **Matched Played (20/21):** 27
- **Minutes Played (20/21):** 1,901 (55.6%)
- **Estimated Value (€), as per TransferMarkt (December 2020):** €2.0M
- **Weekly Gross Base Salary:** €4,231
- **Year Until Contract Expiry:** 1



Summary Analysis:

- ➕ He meets the requirements of age and matches- 21 years old and played 27 matches in 20/21. However, only played 1,901 of the minutes (55.6%) total team minutes.
- ➕ Strongest numbers of all U23 center backs, across attacking a defensive metrics. Strong metrics include passing, carries, npxG, xA, and dribblers tackled%.
- ➖ Least similar player of the three to Piqué.
- ➖ Less strong numbers for Progressive Passes p90 and most notably, Aerial Win%.
- ➖ 1 year remaining on his contract. He is getting regular minutes at one of the biggest clubs in the world, so is an unrealistic transfer target for at this time.

Top Three Transfer Targets

Final three shortlisted players from a data-driven recruitment process to find 'Piqué-like' Center Backs

3rd

Óscar Mingueza

21
Age



€2.0M
Estimated Value



€4.2k
Weekly Salary

+

He meets the requirements of age and matches - 21 years old and played 27 matches in 20/21. However, only played 1,901 of the minutes (56%).

+

Strongest numbers of all U23s, across attacking a defensive metrics. Strong metrics include passing, carries, npxG, xA, and DP/Tack%.

-

Least similar player of the three to Piqué¹.

-

Less strong numbers for Progressive Passes p90 and most notably, Aerial Win%.

-

1 year remaining on his contract. He is getting regular minutes at one of the biggest clubs in the world, so is an unrealistic transfer target for a newly promoted club at this time.



2nd

Wout Faes

22
Age



€3.0M
Estimated Value



€8.1k
Weekly Salary

+

Most similar player of the three to Piqué¹.

+

He meets the requirements of age and playing time - 22 years old and played 33 matches in 20/21, totalling 2,899 minutes (84.8%).

+

Strong numbers for Aerial Win% and Pass Completion%, as well as Completed Passes p90, Carries, and Total Carry Distance p90.

-

Less strong numbers for Pressured Passing p90, Shot Creating Actions, npxG, and xA.

-

2 years remaining on his contract, not available for a free transfer.

-

Highest current salary of the three.

1st

Amos Pieper

22
Age



€3.5M
Estimated Value



€1.5k
Weekly Salary

+

He meets the requirements of age and playing time - 22 years old and played 30 matches in 20/21, totalling 2,682 minutes (87.6%).

+

0 years remaining on his contract – potential for a free transfer.

+

Strong numbers for Completed Passes p90, Completed Passing%, Total Carry Distance p90, and Dribblers Tackled%.

-

Least strongest overall numbers of the three players, performing less well with Progressive Passes p90, Pressured Passes p90, Progressive Carries p90, xA, Shot Creating Actions p90, and Aerial Win%.

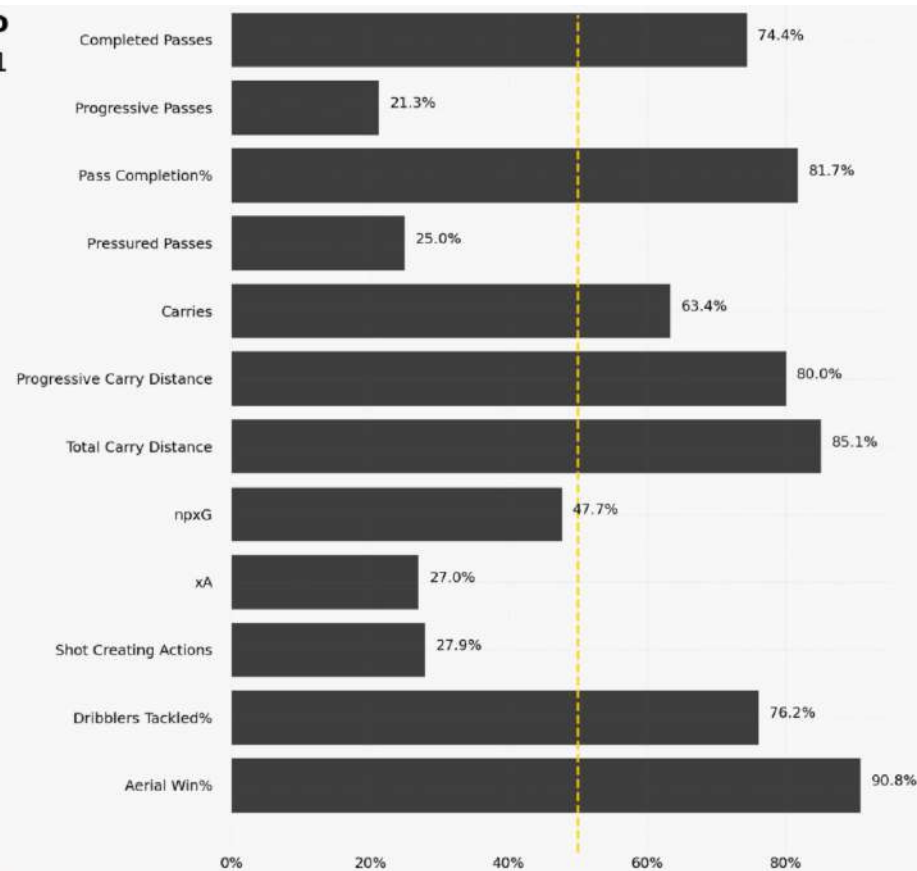
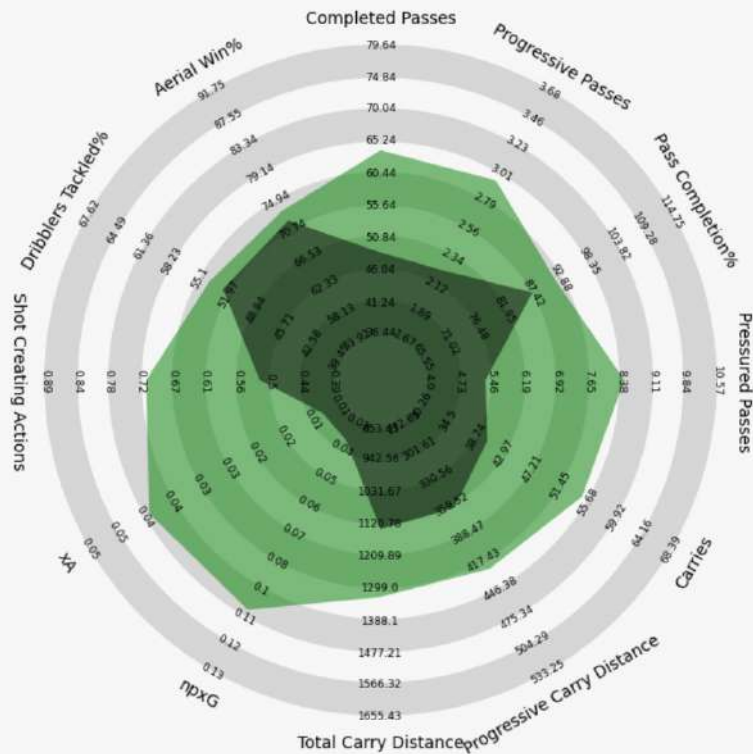
¹As per the Principle Component Analysis (PCA) and K-Means Clustering Machine Learning process. The 'similarity' of the player is determined by a proxy of the distance between the player's data point in the cluster to the centroid of Gerard Piqué's three data points between 17/18-19/20, which also lies close to his 18/19 data point.

Tosin Adarabioyo

Honourable mention, that missed the cut due to high Premier League wages but a well-rounded player

Gerard Pique
2018/2019

Tosin Adarabioyo
2020/2021



Profile

- **Age:** 22
- **Position:** Center Back
- **Team:** Fulham
- **Nationality:** English
- **Foot:** Right
- **Height:** 196cm
- **Matched Played (20/21):** 33
- **Minutes Played (20/21):** 2,953 (86.3%)
- **Estimated Value (€), as per TransferMarkt (December 2020):** €5.0M
- **Weekly Gross Base Salary:** €47,300
- **Year Until Contract Expiry:** 1



Summary Analysis:

- + He meets the requirements of age and playing time - 22 years old and played 33 matches in 20/21, totalling 2,953 minutes (86.3%).
- + Strong numbers across both attacking and defensive metrics, including Completed Passes p90, Pass Completion %, Carries and Progressive Carries p90, Total Carry Distance p90, Dribblers Tackled%, and Aerial Win %.
- Less strong numbers for Progressive Passes p90, Pressured Passes p90, and Shot Creating Actions
- High Premier League wages last season, however, Fulham were relegated to the Championship.

Limitations of Current Modeling

Some of the known problems and issues that I would wish to improve with more time and data

1

Defensive metrics are not possession adjusted (PAdj)

Not available in FBref, affecting both features used in clustering and the selection of metrics visualised/analysed for final shortlisted players. With access to PAdj metrics and/or Event/Tracking data that can be used to derive these metrics, I would expect to see much better clustering/ML results and visualisations/analysis with metrics such as no. interceptions and tackles.

2

Unable to derive nuanced metrics

Metrics such as xGBuildup, xGChain, [On-Ball Value](#) (OBV) and [Expected Threat](#) (xT), are not in the FBref dataset, resulting in the application of less relevant metrics for center backs - npxG and xA. With access to Event and/or Tracking data, for the 'Big 5' European leagues over multiple seasons, this is something I would build into the final aggregated dataset, to enable better analysis¹.

3

Identification of center backs in the dataset

The players in the FBref dataset are labelled as only 'Goalkeepers', 'Defenders', 'Midfielders', and 'Forwards'. This groups Center Backs and Full Backs as 'Defenders'. As these positions have different responsibilities and involve different types of players with different qualities, this affects the clustering process. In our case, good results were produced in the final shortlist, with all thirty players being center backs, however, this is something I would wish to tidy up in the initial stages, to ensure good model performance.

4

Better identification of Piqué's and other player's playing style from the clustering output

As a proof of concept, the model produces interesting results. However, if productionised, I would want to better identify the traits that cause Piqué to play the way he does. Currently, the model reduces the dimensions of the dataset into two dimensions using PCA. This is great for a number of reasons including model performance and visualisation, however, it makes the model a bit of a "black box" once the clustering takes place. The model does not currently say why a player is 'Piqué-like', only the proximity that a player has to Piqué and what cluster they are in. Notable research in this field has been carried out by [Paul Power](#) and [Patrick Lucey](#) of StatsPerform [\[link\]](#).

For an example of where I derived such metrics from Event data, see my StatsBomb Data Parsing [\[link\]](#) and StatsBomb Data Engineering [\[link\]](#) notebooks, and/or visualisation of the resulting dataset as a set of Tableau dashboards [\[link\]](#).

Possibilities and Next Steps of this Clustering Model

Just some of the steps I would like to carry out with more time, data, and resources

- 1** Video analysis. This pack highlights potential recruits using data, however, to be applied in a club, it is vital that performance analysts watch the players on video and in a live setting for two reasons 1) to see whether the players suggested are actually showing 'Piqué-like' qualities on the pitch, and 2) the data can't give us all the information or the context, for example, there is no data collected for a defender ability to read the game, their communication and leadership, etc. These are considerations required to build a complete Player Profile.
- 2** Create a 'Likeliness' Score. Currently, the data just measures the proximity to the cluster and orders by closest to furthest. This could be converted into a percentage score, of how similar that player is to comparison player.
- 3** Identify not the players that are in close proximity to a player like Piqué, but instead, determine the different playing styles of players and provide labels to them. Notable research in this field has been carried out by [Paul Power](#) and [Patrick Lucey](#) of StatsPerform [\[link\]](#).
- 4** Implementation of known metrics that were not available, including xGBuildup, xGChain, [On-Ball Value](#) (OBV) and [Expected Threat](#) (xT), possible with Event data.
- 5** Analysing data over seasons: I only took into account this season's data for illustrative purposes in this study, but further analysis could track a player's consistency over prior campaigns too for a larger sample size.
- 6** Implementation of modeling metrics such as Completed Passes Greater than Expected. Through the application of ML classification algorithms and Event and/or Tracking data, models can be trained to determine the likelihood of a pass being completed which in turn, can be used to determine metrics such as the number of passes completed above the level of expectation.
- 7** Train player valuation models to better estimate a player's market value. TransferMarkt uses a "wisdom-of-the-crowds" approach, which is useful, but is not as accurate as we would like. A more refined method would be to create models from different datasets such as player recorded transfers and Capology player salaries, to better estimate these transfer values. An example model that does this is the DePO model by [Sam Goldberg](#) (New York Red Bulls) and [Mike Imburgio](#) of American Soccer Analysis [\[link\]](#).

Ball-Playing Center Backs at the 2018 FIFA Men's World Cup

Reproduction of this recruitment analysis project using StatsBomb Event data

- The following visualisation is a dashboard of aggregated StatsBomb Event data, for analysis of ball-playing center backs at the 2018 FIFA Men's World Cup.
- Dashboard brings together several visualisations, including a player bio information, a pass map, a percentile rank bar chart, dynamic scattergrams, and a radar.
- Players are analysed for ball-playing abilities, with a focus on metrics including: Completed OP Passes p90, Completed Progressive Passes p90, Pass Completion %, Under Pressure Pass Completion %, Completed Carries p90, Completed Progressive Carries p90, Carry Distance p90, xGBuildup, xT, possession adjusted Tackles and Interceptions, the ratio of Tackles to Diibbles Past, and Aerial Win %.
- See the following StatsBomb Data Parsing [\[link\]](#) and Data Engineering [\[link\]](#) notebooks for more information about how raw StatsBomb Event data is transformed to aggregated, deriving new metrics including possession adjusted defensive metrics and advanced metrics including xGBuildup, xGChain, OBV, and xT.

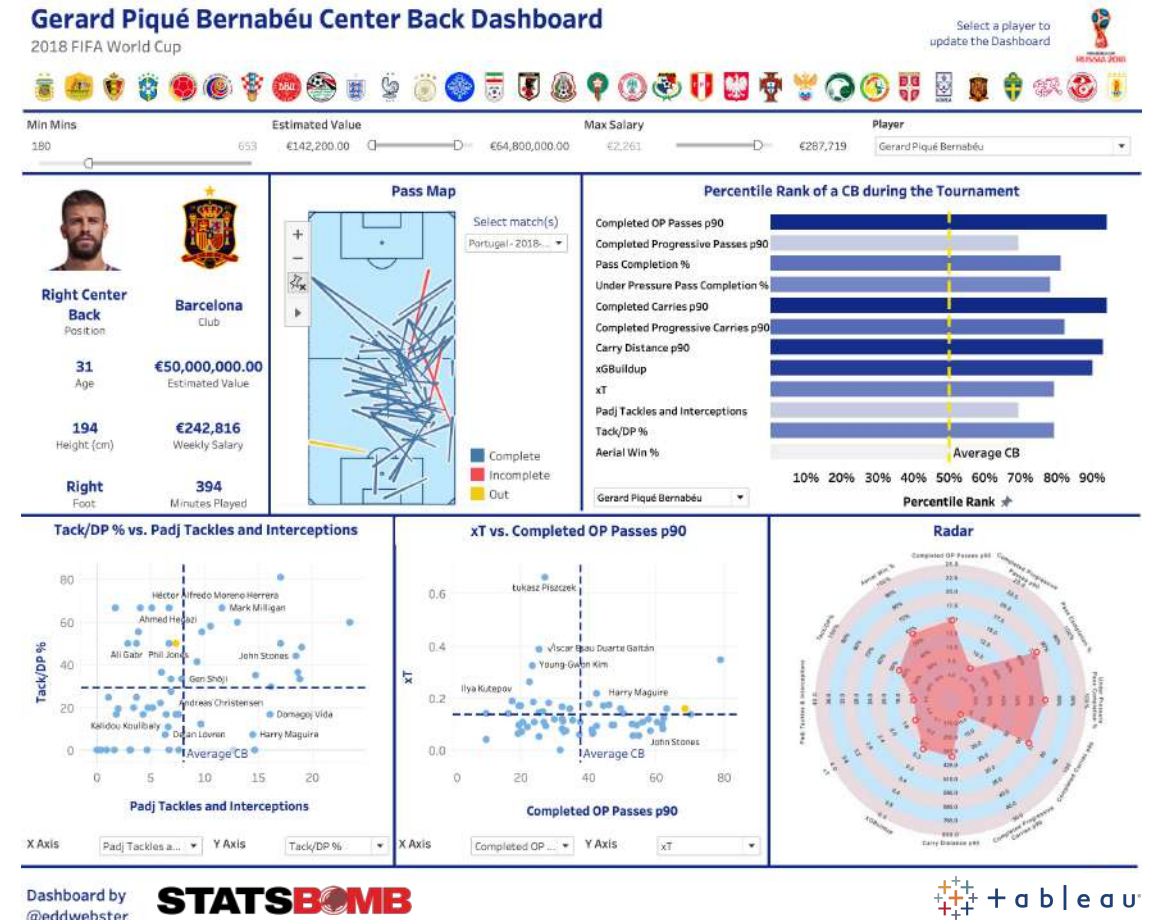


Tableau Public dashboard: public.tableau.com/app/profile/edd.webster/viz/EddWebster-WorldCup2018AnalysisandDashboard/WC2018PlayerDashboard.

Project 3: Application of Tracking Data

Sample datasets provided by Signality and Metrica Sports
featuring the implementation of Pitch Control

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

Further Work



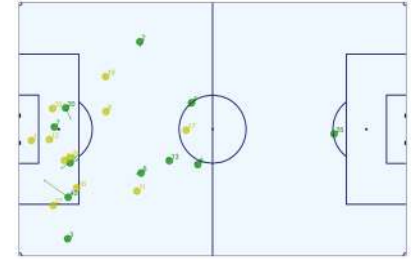
Application of Tracking Data

Key areas of analysis possible with Tracking data demonstrated in this analysis pack

1

Determine a player's positions, speed, acceleration, and movement

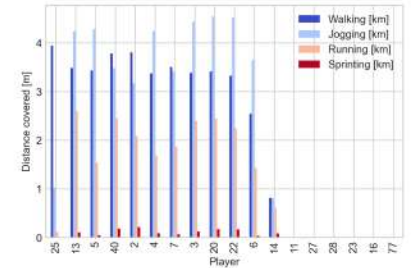
For a defined moment, using the timestep and relative positions to calculate the the speed, and acceleration of players for a given moment/sequence in play.



2

Physical performance reports of players

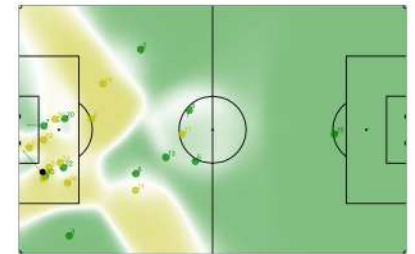
Derive the minutes played, distance travelled and at different speeds (walking, jogging, running, and sprinting), the number of sprints, and the top speeds of players of the course of a match.



3

Create Pitch Control models to visualise the probabilities of pass success

Model to determine the probability that a player (or team) will gain control of the ball if it moves directly to that location and how to combine Pitch Control with EPV frameworks to determined the expected value-added of actions.



All code for working with both Signality and Metrica Sports Tracking data: github.com/eddwebster/football_analytics/tree/master/notebooks/5_data_analysis_and_projects/tracking_data

Application 1: Determine Player's Position, Direction, Speed, and Acceleration

Determine direction and speed of all twenty two players in any given instant

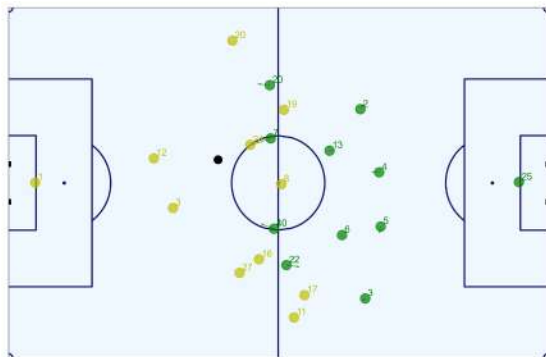


Fig. A: Hammarby vs. Elfsborg, kickoff

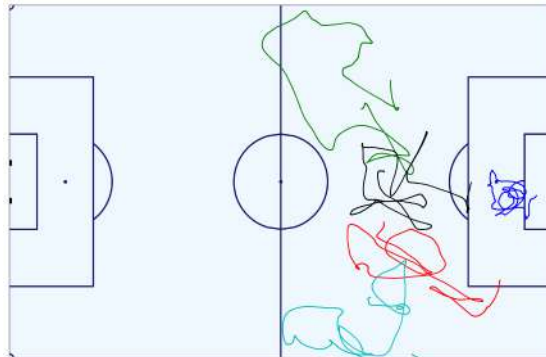


Fig. B: Hammarby vs. Elfsborg, first sixty seconds of movement

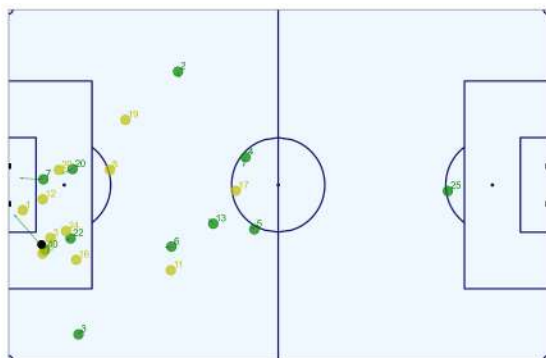


Fig. C: Hammarby vs. Elfsborg, first goal (1-0) scored by Imad Khalili (player 7)



Fig. D: Hammarby vs. Elfsborg, third goal (3-0) scored by Imad Khalili (player 7)

- Using the Tracking data provided by [Signality](#), it is possible to determine the position, speed, acceleration, and direction for all of the players on the field at any given moment.
- Signality Tracking data, like other providers including [Metrica Sports](#) and [Hudl](#), is collected at twenty five frames s^{-1} . Therefore, a player's speed can be calculated by dividing the distance a player has covered between two frames by 0.04. The acceleration is calculated as the second order derivative i.e. speed divided by $dt = 0.04$ again.
- The ball position is missing for a sizeable part of the data so the ball is not visible in every figure.
- In these examples from the Hammarby vs. Elfsborg match, we observe the following
 - Fig. A visualises the kick off between Hammarby and Elfsborg including vectors to represent the speed and direction of each player;
 - Fig. B shows the paths that the back four and the goalkeeper took in the first sixty of the game;
 - Fig. C visualises the moment just before the first goal where Nikola Djurdjic (player 40) makes a pass across the edge of the 6-yard-box to Imad Khalili (player 7); and
 - In Fig D visualises the moment before the third goal where Muamer Tankovic (player 22) is about to play a pass around the corner to Imad Khalili, to leave him 1-on-1 with the goalkeeper.

Sample Tracking data provided by Signality. All code for working with Signality Tracking data: github.com/edwebster/football_analytics/tree/master/notebooks/5_data_analysis_and_projects/tracking_data/signality. All PNG figures produced from Signality Tracking data: github.com/edwebster/football_analytics/tree/master/img/fig/signality.

Application 2: Measure the Physical Performance of Players

Determine the distance from goal, speed, and acceleration for players over time

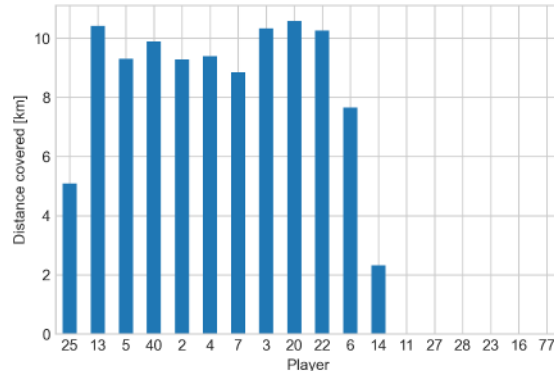


Fig. A: Distance covered by Hammarby players in the match again Elfsborg.

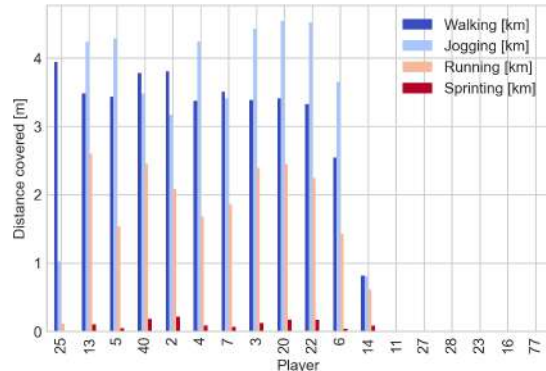


Fig. B: Distance covered by Hammarby players in the match again Elfsborg, broken down by different run types – walking, jogging, running, and sprinting.

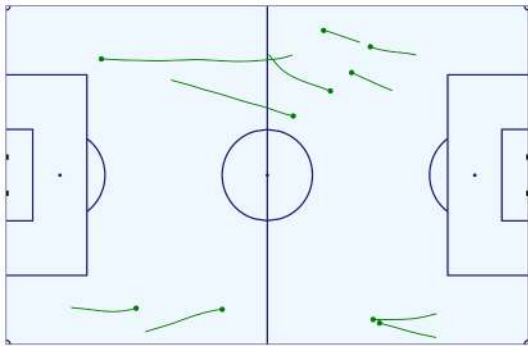


Fig. C: Sustained sprints conducted by Player 4 for Hammarby in the match against Elfsborg.

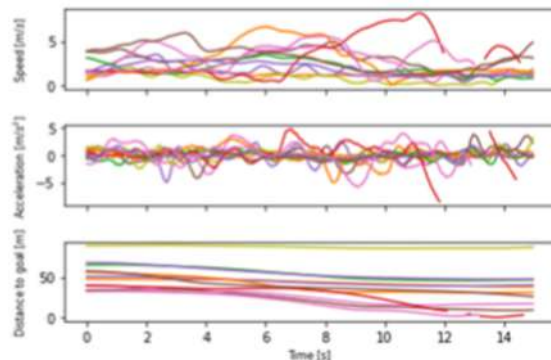


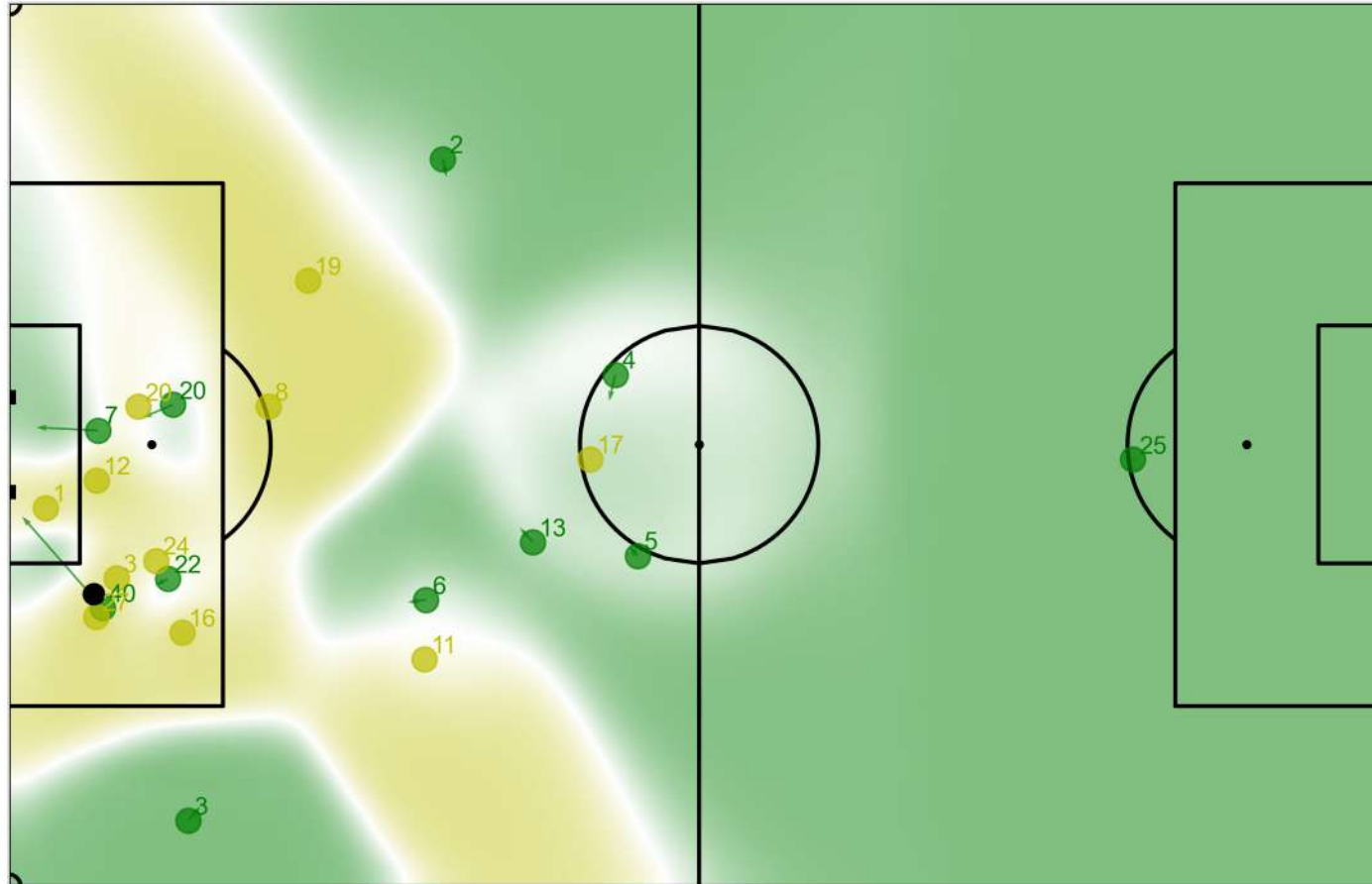
Fig. D: An example of calculating speed, acceleration, and distance to goal for the players during a fifteen-second passage of player, smooth with a Savitzky-Golay filter of seven frames

- As well as the player's speed and acceleration, it is possible to calculate the distance a player is from the goal at any moment, their top speed, and their top acceleration, to name but a few examples. In Fig. A and B, the distances each player travels per match is plotted.
- As part of the Quality Assessment of speed calculation, all players with speeds above 12 m/s (Usain Bolt's top speed) are dropped from the dataset. The data is collected automatically from a video feed and subsequently, calculations that lead to non-realistic speeds are due to position errors in the data collection.
- A convolution [Savitzky-Golay](#) filter of seven frames is used for smoothing (Fig. D).

All PNG figures produced from Signality Tracking data: github.com/eddwebster/football_analytics/tree/master/img/fig/signality.

Application 3: Pitch Control Models 1/5

Application of Pitch Control with Signality Tracking data to quantify the a team's control in regions of the pitch



The Pitch Control model used with Signality data is an edited version of [Laurie Shaw](#)'s Pitch Control model for the Friends of Tracking initiative that is compatible with the Metrica Sports sample data. The code was changed to work without Event data. See: github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking.

All PNG figures produced from Signality Tracking data: github.com/eddwebster/football_analytics/tree/master/img/fig/signality.

Application 3: Pitch Control Models 2/5

What is Pitch Control? Definitions

Definition of Pitch Control

The Pitch Control Field (PCF) for location x , is defined as the probability that the Home team will end up with possession of the ball if it were at location x . The PCF predicts the team that would be in possession at the next timestep.

Factors of the player's state that determine the PCT

- Location / position
- Velocity
- Acceleration

Other considerations

- Maximum player speed
- Time taken for the play to control the ball (a constant of 0.7 seconds)

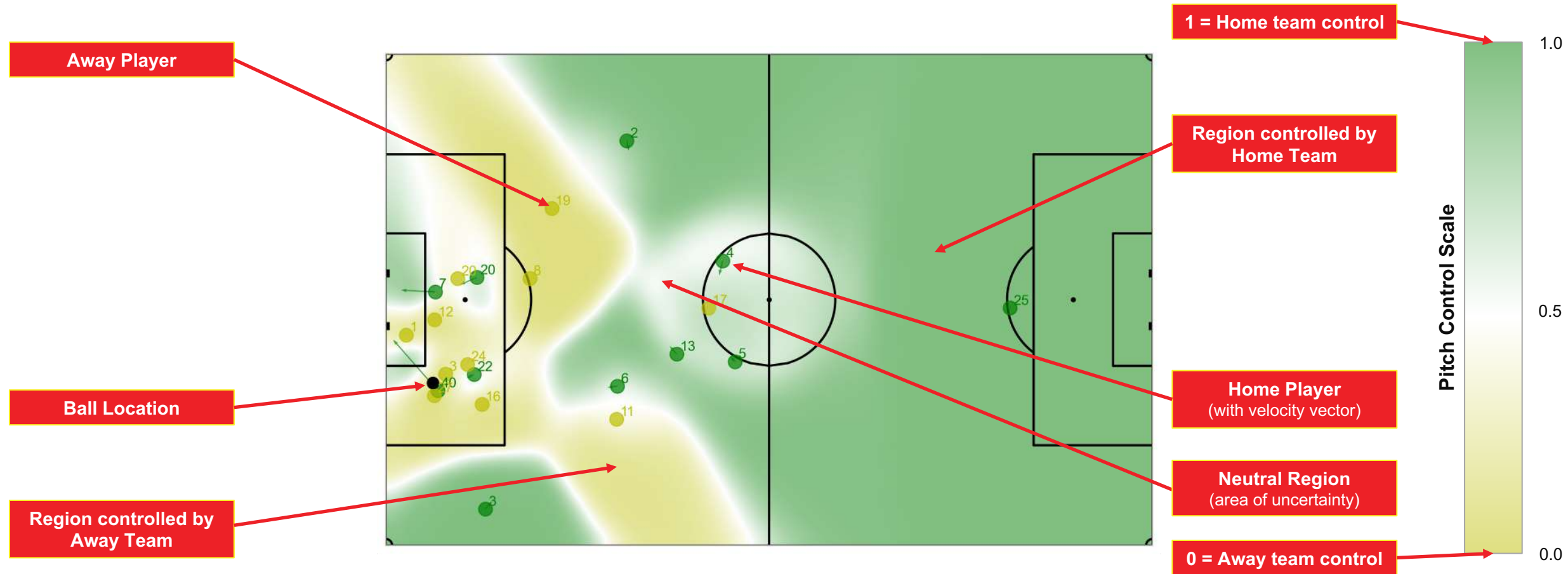
Applications of Pitch Control

- Evaluate passing decisions by looking at how much open space is in front of attackers / behind defenders. Best regions are those where receiving players are in an area of the pitch with a large integrated PCF value.
- A new metric of player performance to compare a player's ability to control the pitch.
- Models combined with video footage to visualise both the available space for an attacker in or the defensive players out of position in a moment of play.

Definitions reproduced from [Will Spearman](#)'s 2016 Opta Pro forum seminar titled: 'Quantifying Pitch Control' . See his presentation: researchgate.net/publication/334849056 Quantifying Pitch Control.

Application 3: Pitch Control Models 3/5

Visualisation of Pitch Control in a 2D space



Visualisation reproduced using Signality Pitch Control model with the explanation from [Will Spearman](#)'s 2016 Opta Pro forum seminar titled: 'Quantifying Pitch Control'. See PowerPoint: researchgate.net/publication/334849056 Quantifying Pitch Control.

Application 3: Pitch Control Models 4/5

Application of Pitch Control modeling with Signality Tracking data

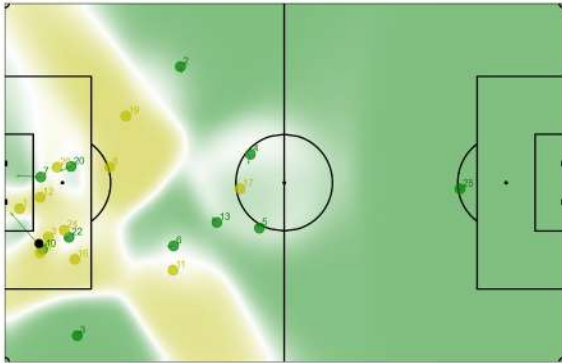


Fig. A: Pitch Control for the pass just before the first goal of Hammarby vs. Elfsborg, kickoff

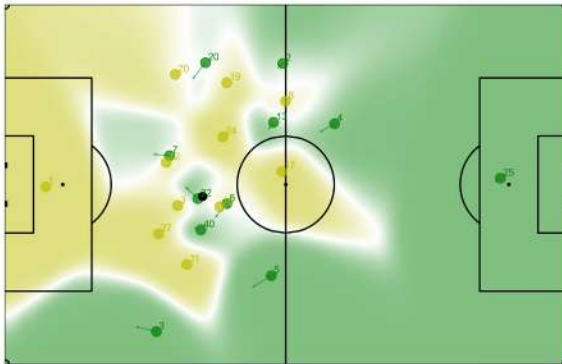


Fig. B: Distance covered by Hammarby players in the match against Elfsborg, broken down by different run types – walking, jogging, running, and sprinting.

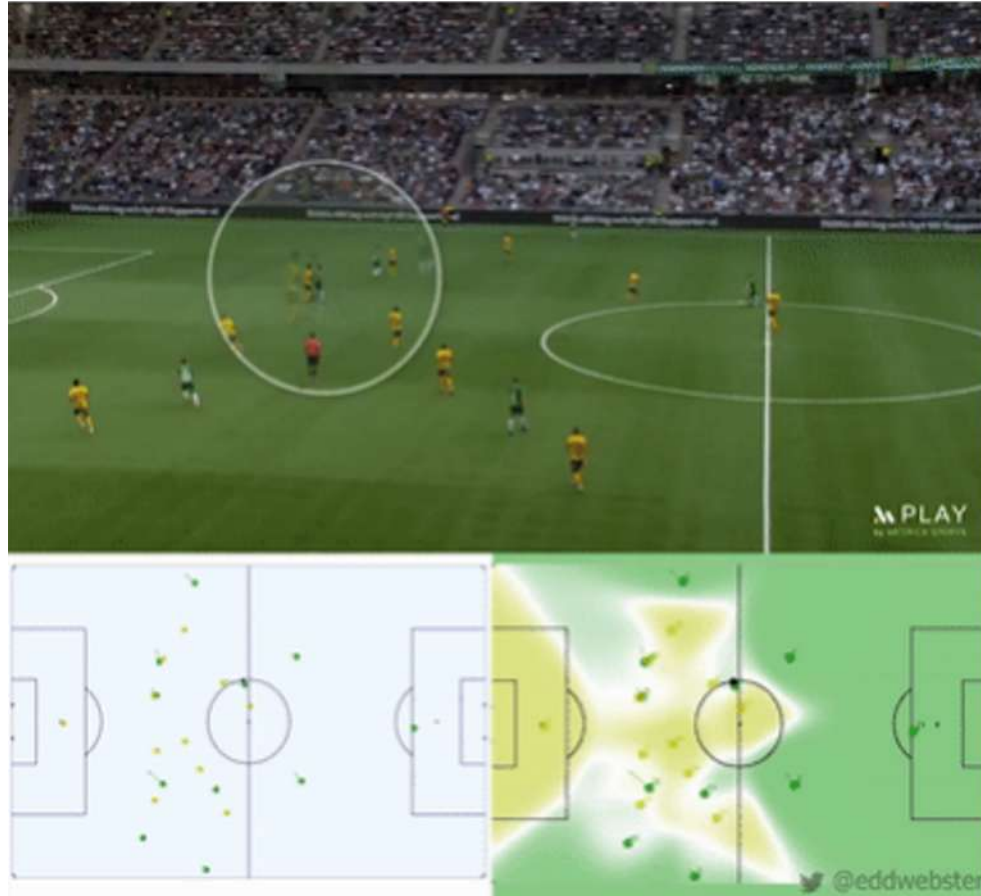
- The Pitch Control models observed are created with a modified version of the code provided by [Laurie Shaw \[link\]](#) compatible with Metrica Sports Tracking data, as part of the Friends of Tracking initiative. To see his code, please see the following notebooks [\[link\]](#).
- To allow the existing Metrica Sports Pitch Control model to be compatible with the Signality Tracking data, the data was first required to be engineered to match the format of the Metrica Sports data. The code for the Pitch Control model was then amended to work with just Tracking data. Unfortunately, the Signality Tracking data provided does not include corresponding Event data, and therefore, frames of interest must be selected by Frame ID and not Event ID, as can be done in Laurie's Metrica Sports-compatible model. This required alteration in the code.
- **Note:** unfortunately the ball position is missing for a sizeable part of the data so the ball is not visible in every frame. For this reason, these particular figures at the particular frame are selected.
- This slide analyses two of the goals scored by Hammarby in their 5-2 win against Elfsborg on 22nd July 2019.
 - In Figure A, Hammarby are about to score their first goal. Player 40 for Hammarby squeezes between two opposition players close to the near post to hit a first time pass across the goal, to the region of the pitch controlled by Hammarby, where Imad Khalili (player 7) runs in for a tap in. If Nicola Djurdjic (player 40) had failed to execute pass, Alejandro Portillo (player 12) for Elfsborg could have intercepted the ball in a region that is shown to be controlled by Elfsborg. The pass however was successful and Hammarby take a 1-0 lead.
 - In Figure B, Muamer Tankovic (player 22) has just received the ball from Mads Nielsen (player 13). The player could make a simple pass to Nikola Djurdjic (player 40), or to the left-back David Fällman (player 5). However, Tankovic instead decides to play a more risky but much more rewarding pass around the corner in behind the opposition's defence, to a region of space controlled by Imad Khalili (player 7), acceleration into the space towards the goal, from an on-side position. This pass leads to Khalili having a one-on-one with the goalkeeper, which he converts to give Hammarby a 3-0 lead.
- These examples may explain what we can already see with the naked eye, but this can then be combined with frameworks such as Expected Possession Value (EPV) models, to determine the value of individual actions during a passage of play, and whether the player made the correct pass in a given situation.

All PNG figures produced from Signality Tracking data: github.com/eddwebster/football_analytics/tree/master/img/fig/signality.

Application 3: Pitch Control Models 5/5

Combining applied Pitch Control models with the video footage

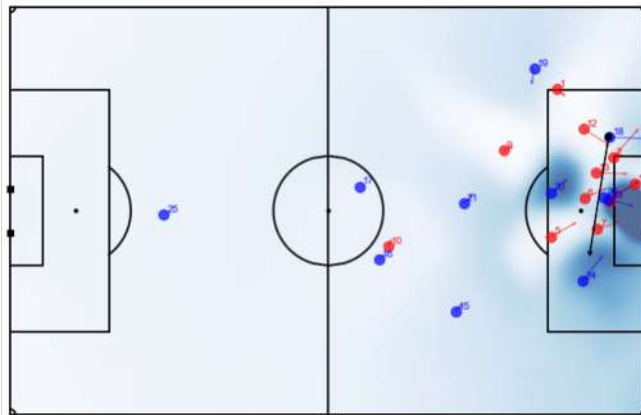
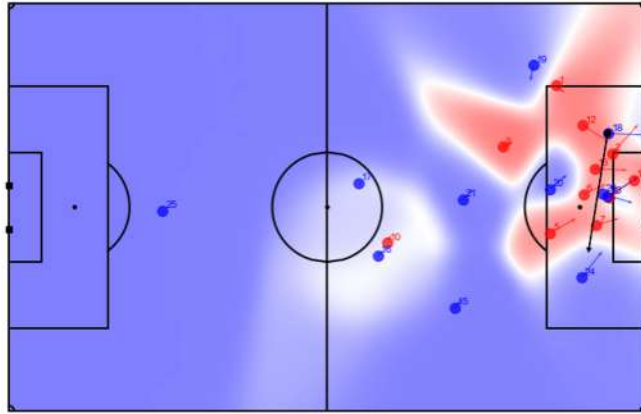
If you are viewing the PDF version of this presentation, the interactive GIFs can be viewed in the online version: docs.google.com/presentation/d/1a8YgbsXTbpT8FYDvmi-J2eXgN2xnOkUS2E-8YsWrINk or on GitHub: github.com/eddwebster/football_analytics/tree/master/gif/fig/signality.



For interactive GIFs, please see the online version of this presentation: docs.google.com/presentation/d/1a8YgbsXTbpT8FYDvmi-J2eXgN2xnOkUS2E-8YsWrINk or separately on GitHub: github.com/eddwebster/football_analytics/tree/master/gif/fig/signality and github.com/eddwebster/football_analytics/tree/master/video/fig/signality.

Application 4: Combining Pitch Control with EPV models for Expected Value-Added

What is Expected Possession Value (EPV) and how it can be used with Pitch Control to value passes on the field?



- As seen on the previous slides, Pitch Control can visually identify where on the field a team can pass the ball and retain possession. However, it cannot tell produce a value for said passing options. To identify the best passing options, we need to add valuation framework to the model.
- Expected Possession Value (EPV) or Possession Value (PV) is a model that looks at how much scoring value a player (or team) brings to a possession based on their actions (e.g. passing, dribbling, etc) at any given instance in a game.
- The probability that the possession will end in a goal, given a situation, can be broken down into four variables:
 - Ball position;
 - Position & velocity of teammates;
 - Position & velocity of opponents; and
 - The match state i.e. open play, set piece, score line, time left, etc.
- The first EPV model was Sarah Rudd's framework using Markov chain [\[link\]](#), and since then, there have been other notable models including SciSports and KU Leuven [Valuing Actions by Estimating Probabilities \(VAEP\)](#) framework, Karun Singh's [Expected Threat \(xT\)](#) framework, and [American Soccer Analysis' Goals Added \(g+\)](#) framework.
- For each area of the pitch, an EPV can be determined. The value added or lost by completing an action is the difference between the EPV of the two states e.g. a pass from position x_i, y_i to x_j, y_j .
- The Pitch Control of a given possession that gives the probability of a successful pass can be multiplied by the EPV for the value of the possession to produce the [Expected Value-Added](#).
- The figures on the left first using Metrica Sports Tracking data, show the Pitch Control model for a pass made by Player 18 and in the second figure show the EPV for the pass. Areas of darker blue are for greater EPV values, which shows that by passing to Player 24, Player 18 made the correct pass in this instant.
- We can also calculate the EPV for all the passes made by players, to determine which were the top passes made by a team during a match and also the players that added the greatest of EPV in a match i.e. add the most positive contribution to the team during the 90 minutes.

EPV model used was produced by [Laurie Shaw](#): github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking/blob/master/EPV_grid.csv.
Examples above explained in [Laurie Shaw's](#) tutorial 'Beyond Pitch Control': youtube.com/watch?v=KXSLKwADXXI&ab_channel=FriendsofTracking.
All PNG figures produced from Metrica Sports Tracking data: github.com/eddwebster/football_analytics/tree/master/img/fig/metrica.

Possibilities and Next Steps of Working with the Tracking Data

Some of the steps I plan/wish to take in the near future to build on my analysis with Tracking data

- Combine the Pitch Control models built with Metrica Sports and Signality Tracking data with Expected Possession Value (EPV) models such as the [VAEP](#) model by [SciSports](#) and [KULeuven](#), or the [Expected Threat \(xT\)](#) model by [Karun Singh](#), to further analyse the value that certain actions of interest brought to the team during a particular play in the match and determine the Expected winner.
- Apply the Pitch Control models and EPV models used with Metrica Sports and Signality data with the sample [SkillCorner broadcast Tracking data](#).
- With access to a complete dataset of Tracking data with corresponding Event data e.g. a full season's worth of data, example applications that can be achieved with such data include:
 - Training models for formation detection and team strategy – see the following paper by [Laurie Shaw](#) (City Football Group) and [Mark Glickman](#) (Harvard University): '[Dynamic analysis of team strategy in professional football](#)' and accompanying talk [\[link\]](#); and
 - Further enrich the corresponding Event data using the Tracking data, to improve other models derived from Event data such as Expected Goals model. Such features that could be included with Tracking data are:
 - The pressure with which the player was under when taking a shot – distance based logic based between the player in possession and the opponent(s);
 - The number of teammates and opposing players between the shot-taker and the goal at the moment that a shot is taken; and
 - What are the goalkeeper and defender positions in the moment of the shot, how much of the goal was covered by the goalkeeper.

Part 4: Expected Goals Modeling

Training an Expected Goals (xG) model using Event and Tracking data, subsequently applied to a separate match dataset, to measure team's performance

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

Further Work



Expected Goals Modeling

Three-part project to create and xG model for use to analyse the attacking chances of a sample dataset

1

Define, build and train an Expected Goals model using a sample dataset of Shots data

Build and train a binary classification model that can predict the likelihood of defined shots resulting in a goal.

2

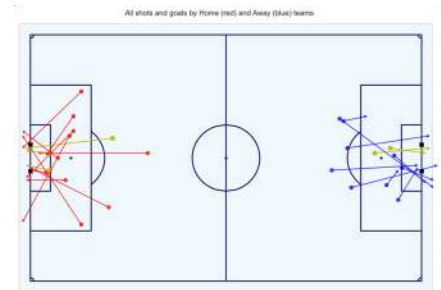
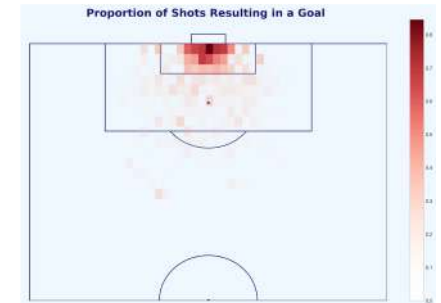
Analyse game 2 of the sample Metrica Sports Tracking and Event Data

Examination of the key chances in the match using the Tracking and corresponding Event data. Exported shots dataset engineering to be compatible for predictions with the trained Expected Goals model.

3

Assessment of the team's performance in the sample Metrica Sports match through application of the Expected Goals model

Analysis of the performance of the teams in the sample match *via* the application of the trained Expected Goals model, to determine the quality of the chances that took place during the match.

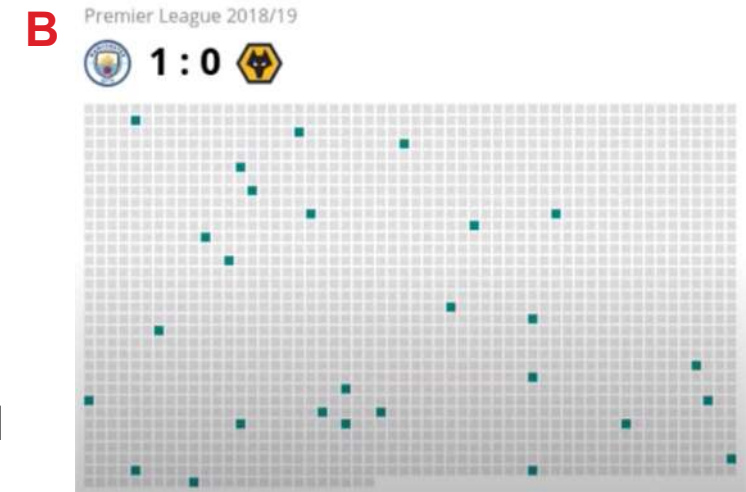


All code for this Expected Goals modeling and analysis: github.com/eddwebster/football_analytics/tree/master/notebooks/5_data_analysis_and_projects/xg_modeling.

What are Expected Goals and Why is it an Important Metric?

Arguably the most important metric in football analytics so far – Expected Goals

- ‘[Expected goals](#)’, often abbreviated as ‘xG’, is the probability that a shot with defined state will, on average, result in a goal (Fig. A).
- Shots in matches are rare and goals are even rarer! On average, each team shoots 12 times per matches (0.4% of a 3,000 Event dataset) resulting in an average of 2.66 goals (Fig. B).
- xG allow analysts and statisticians to look at all shots instead of just the goals, which happen around ten times as many times as the goals themselves, making for a much better predictor of goals scored by a team in the medium term (this is explored further in the summary slides of this section).
- xG can be used to make smarter decisions on recruitment, tactics, and strategy. For example, when recruiting a striker in good form, xG can be used to see if the player is over-performing or is in fact taking the chances that would be expected.



Definition of Expected Goals derived from [Laurie Shaw](#)'s article: Bodies on the Line: Quantifying how defenders affect chances: eightyfivepoints.blogspot.com/2017/09/bodies-on-line-quantifying-how.html.

Expected Goals images taken from both [David Sumpter](#)'s Friend of Tracking lesson: 'How to Build An Expected Goals Model 1: Data and Model' (youtube.com/watch?v=bpiLyFyLIXs) and his book Soccermetrics.

Visualisation of the number of shots in a game taken from [Lotte Bransen](#) and [Jan Van Haaren](#)'s talk 'How to find the next Frankie de Jong' for Friends of Tracking: youtube.com/watch?v=w0LX-2UqyXU.

Statistics taken from chapter two of The Numbers Game by [Chris Anderson](#) and [David Sally](#) and Soccermetrics by [David Sumpter](#).

Exploratory Data Analysis of the Raw Dataset

Information about the sample dataset of just under 11,000 shots used to train the Expected Goals model

- The shots DataFrame contains 10,925 shots (Fig. A) and 1,374 goals (12.6%) (Fig. B)
- The key features available:
 - X and Y coordinates of the shot.
 - The Play Type i.e. the game situation in which the shot was taken (open play, penalty, direct free kick, direct from a corner);
 - The Body Part which shot was taken (left foot, right foot, head, other);
 - The No. of Intervening Opponents and Teammates i.e. the number of players that were obscuring the goal at the instant of the shot (from the perspective of the shot-taker);
 - The Interference of the Shooter i.e. the pressure the shot-taker is experiencing from defenders (Low - no or minimal interference, Medium - a single defender was in close proximity to the shot-taker; High - multiple defenders in close proximity and interfering with the shot); and
 - The shot Outcome i.e. the result of the shot- (blocked, missed, goal frame (post or bar), saved, goal or own goal.

< 11,000 Shots in the Dataset

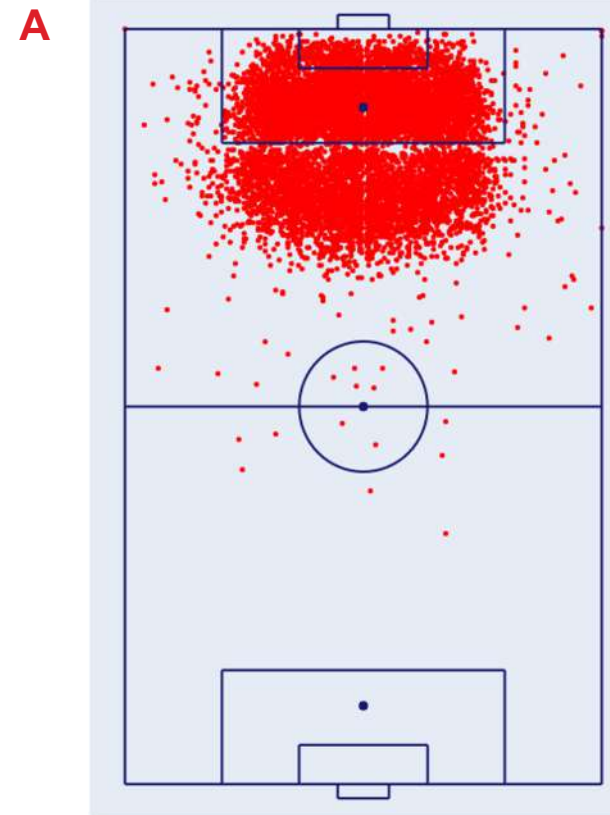


Fig. A: All shots in the dataset.

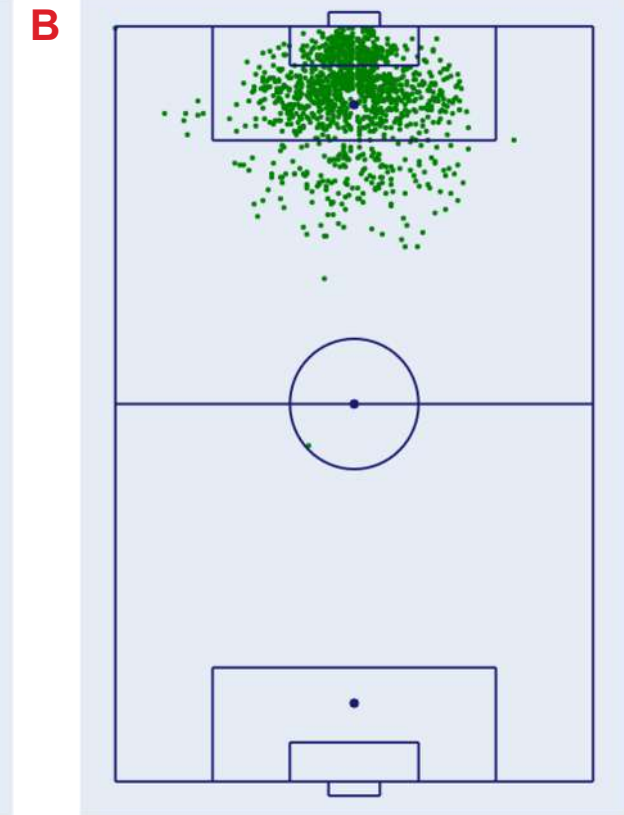


Fig. B: All goals in the dataset.

Full details of the dataset can be found in ShotData.txt documentation: github.com/eddwebster/football_analytics/blob/master/documentation/shots/ShotData.txt.

Notebook to create the initial Expected Goals model from the shots data:

[github.com/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/1\)%20Logistic%20Regression%20Expected%20Goals%20Model.ipynb](https://github.com/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/1)%20Logistic%20Regression%20Expected%20Goals%20Model.ipynb).

Data Engineering

Steps to engineer the dataset before modeling – coordinate conversion, data cleaning, and subsetting the data

- **Conversion of Pitch Coordinates**
 - Original: $x(+106, 0)\text{m}$, $y(-33.92, +33.92)\text{m}$ (Fig. A)
 - Converted: $x(-53, +53)\text{m}$, $y(-34, +34)\text{m}$ (Fig. B)
- **Data cleaning** – replace NULL values for 'Interference on the Shooter' with 'Unknown' string and values in the 'Play Type' column tied to be just 'Direct Corner'.
- **Removal of own-goals** (43)
- **Define target variable, 'isGoal'** – whether the shot was a goal (1) or not (0), derived from the 'Outcome'.
- **Subset for only Open Play shots** – removing all shots from penalties, free kicks, or directly from corners.
- **Overall** – in total, 699 shots (6.4%) and 199 goals (14.5%) were removed from the dataset.

Dimension of a standard football pitch: en.wikipedia.org/wiki/Football_pitch

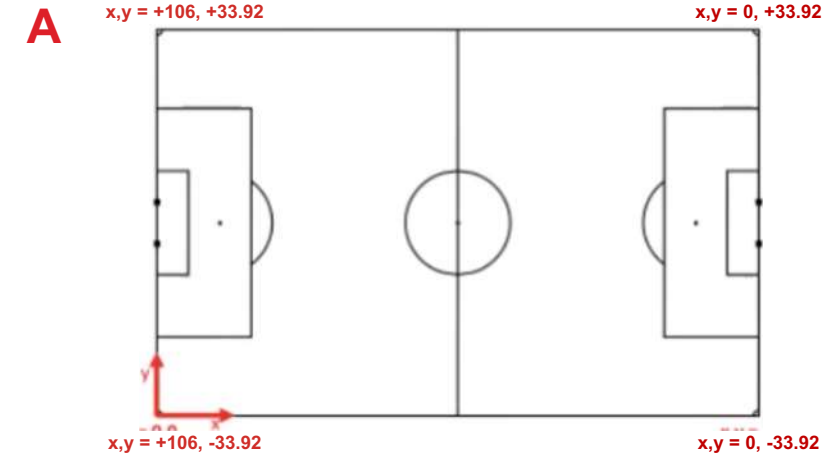


Fig. A: Pre-conversion of coordinates

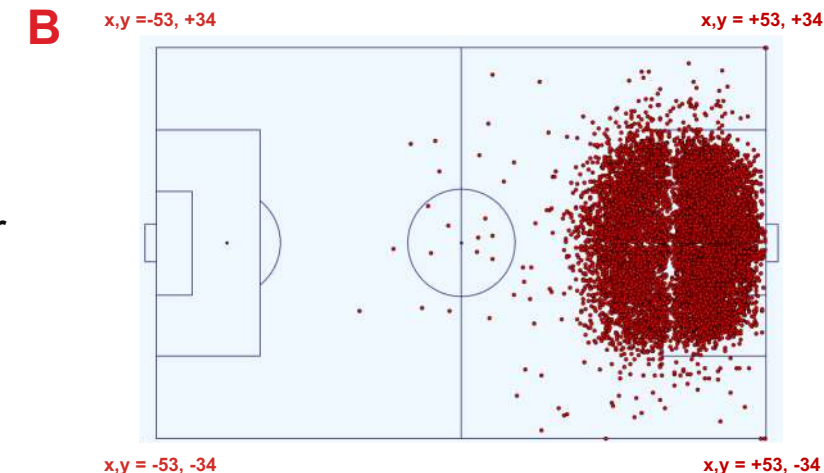
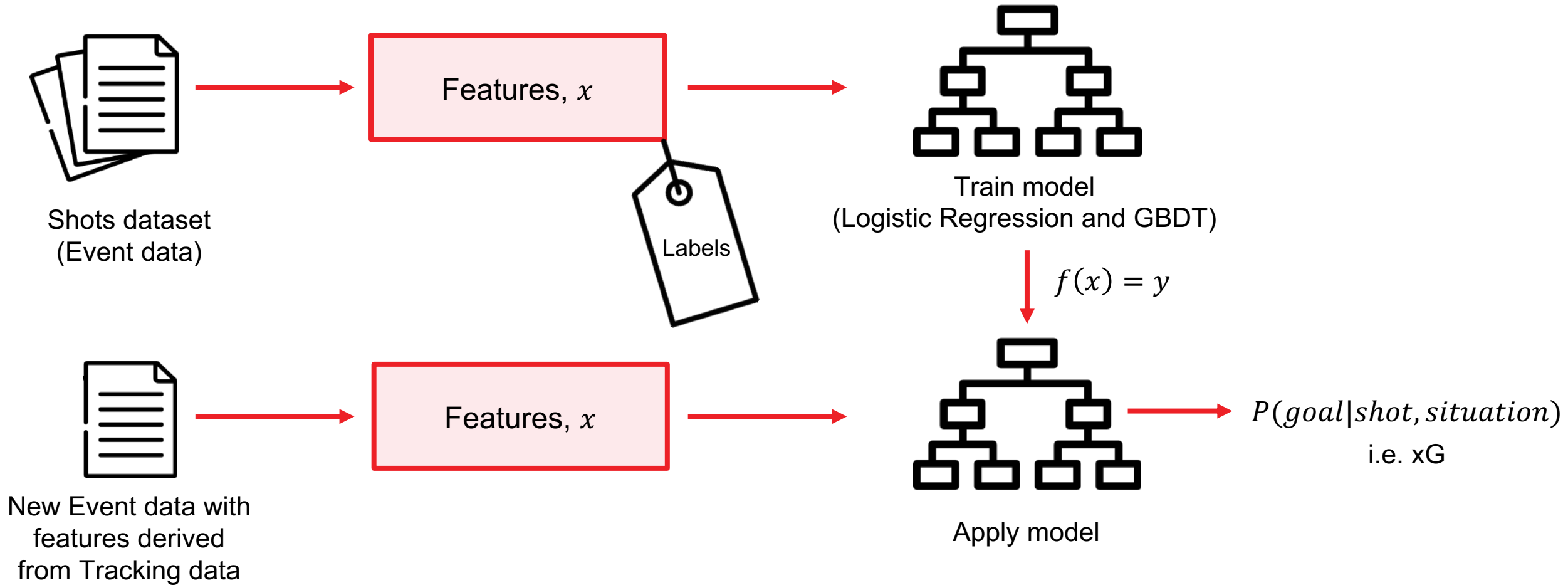


Fig. B: Post-conversion of coordinates including shots

The Machine Learning Task

Creating a model from first principles to calculate the probability of a shot resulting in a goal i.e. Expected Goals



GBDT – Gradient Boosted Trees algorithms such as [XGBoost](#) and [CatBoost](#).

Model Selection 1/2 – Advantages and Limitations of Logistic Regression Models

Strengths and weakness of using Logistic Regression for classification

Advantages

- Logistic Regression is very easy to implement.
- Logistic Regression is easier to interpret than Gradient Boosted Decision Tree (GBDT) algorithms. The model coefficients of Logistic Regression can be used to indicate the importance of the features – extra libraries for feature importance such as [SHAP](#) are not necessarily required.
- Fast training speed - faster than GBDT algorithms.
- More stable model than GBDT algorithms.
- Logistic Regression is already well calibrated i.e. when the probability of classification is 0.3, that represents a 30% chance of a goal being scored. Unlike Logistic Regression, GBDT models require calibration after training.
- Logistic Regression is less inclined to overfit than Decision Trees and Gradient Boosting algorithms.
- A well-tuned Logistic Regression model can perform nearly as well as a model created using Decision Trees or Gradient Boosting algorithms such as XGBoost.

Limitations

- Logistic Regression does not deal well with non-linearities unless you add higher order terms. For this reason, domain knowledge and feature engineering is required to encode non-linear relationship. In the case of an Expected Goals model, instead of using the X and Y location on the pitch, the distance from the goal was determined and used as the feature to determine the shot-taker's location on the field.
- Logistic Regression is nearly always less accurate than a well trained GBDT model. Many Kaggle competitions are won using GBDT algorithms such as XGBoost and CatBoost, however, these algorithms are not guaranteed to be better than Logistic Regression in every setting.

GBDT – Gradient Boosted Decision Trees.

Logistic Regression wiki: en.wikipedia.org/wiki/Logistic_regression.

Model Selection 2/2 – Advantages and Limitations of GDBT Models

Strengths and weakness of using XGBoost for classification

Advantages

- A well trained Gradient Boosted Decision Tree (GBDT) model such as an XGBoost and CatBoost models are more accurate than Logistic Regression. Many Kaggle competitions are won using XGBoost and CatBoost.
- GBDT models handle non-linearities better than Logistic Regression. In the case of an an XGBoost trained xG model, there is no requirement that the X and Y locations on the pitch are converted to a distance from the goal to determine the shot-takers position.
- GBDT models can handle categorical features without using one-hot encoding features since several tests can combine to split a categorical feature.
- Implementation of GBDT models such as XGBoost and CatBoost is relatively easy and it's easy to get a good performance with little tuning.
- GBDT models such as XGBoost are designed to handle missing data with its in-build features (sparse-aware). In the case of this Expected Goals model, all missing values were treated and this was not a requirement from the chosen algorithm.
- Interpretation of the features in GBDT is possible with external libraries such as [SHAP](#).
- GBDT models such as XGBoost can easily run a cross-validation after each iteration.
- GBDT models such as XGBoost support regularisation.
- GBDT models such as XGBoost and CatBoost work well with a small to medium dataset.

Limitations

- GBDT models require calibration once the model is trained i.e. in the case of an Expected Goals model, when the probability of classification is 0.3, that represents a 30% chance of a goal being scored in a Logistic Regression model. However, in the case of a GBDT model such as XGBoost, without calibration, this is often not the case.
- GBDT models generally do not extrapolate well to factors in the modeling that are hidden or not available in the dataset. Logistic Regression however is better at handling this uncertainty.
- GBDT models are more inclined to overfit than Logistic Regression models.
- GBDT models are less easy to interpret than Logistic Regression (an extremely important aspect of football data analytics is interpreting the model results to football practitioners). While Decision Trees are also interpretable because of their strict if/else tests, the ensemble methods, which have higher generalisation ability, do not have a straightforward interpretation.
- GBDT models have slower training speed than Logistic Regression. For example, tuning hyperparameters using search functions such as Grid Search can take some time.

GBDT – Gradient Boosted Decision Trees.

XGBoost official documentation: xgboost.readthedocs.io/en/latest/.

Treating Outliers

Finding and handling unlikely and irregular shots in the dataset that could have an impact on the xG model

- A number of Open-Play shots taken and subsequently scored from improbable locations on the pitch (Fig. A). Most likely due to uncharacteristic errors from the goalkeeper (see Xavi Alonso example below and [\[link\]](#)).
- These shots took place, but we do not want the model to learn that there is a chance to score from this area in our limited dataset. These goals therefore excluded by changing the result of all shots with distance >35m and all shots from >18m with shot angle >35 degrees removed from the data.
- This results in 33 Open Play shots values having their value changed from goal (1) to no goal (0) - 0.27% of the Open Play shots.



If you are viewing the PDF version of this presentation, the interactive GIFs can be viewed in the online version:
docs.google.com/presentation/d/1a8YgbsXTbpT8FYDvmi-J2eXgN2xnOkUS2E-8YsWrINk or on GitHub:
github.com/eddwebster/football_analytics/tree/master/gif.

Xavi Alonso's 70-yard goal against Luton in the FA Cup is a great example of a player scoring an unlikely goal from their own half with the goalkeeper out of position: youtu.be/4OTQwuAc4HU.

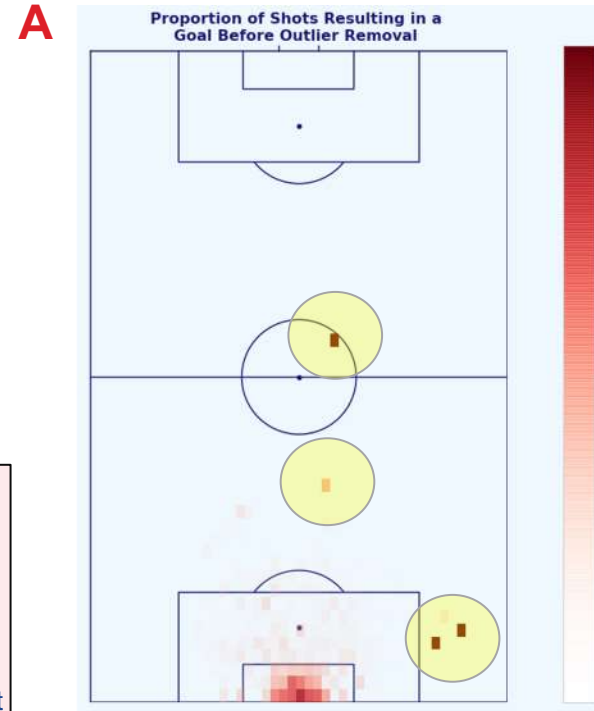


Fig. A: Heatmap of the proportions of shots to goals in the dataset (goals divided by shots). This visualisation flags some of the outlier shots scored from inside the attacking teams half, and at acute angles to the left of the keepers left post.

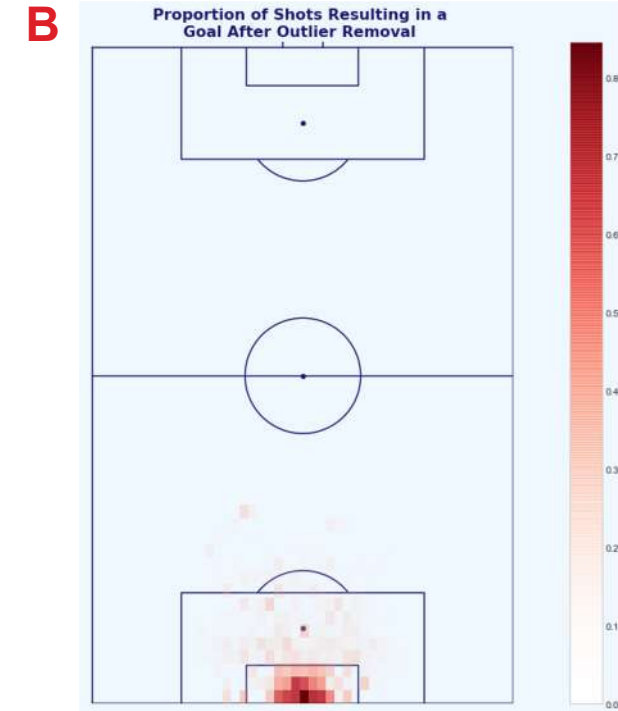


Fig. B: Heatmap of the proportions of shots to goals in the dataset post-outlier treatment. These outliers have now been removed to prevent this affecting the Expected Goals model.

Features Selection 1/2 – Selected Features for the Final Model

Description and status of features in the Expected Goals model from which predictions of the Metrica Shot data were made

Feature	Used in Final Model	Description
distance_to_goalM	Yes	Continuous feature that determines the distance the shot was taken along the y-axis in relation to the goal, in meters.
distance_to_centerM	No	Continuous feature that determines the distance the shot was taken along the x-axis in relation to the center of the pitch, in meters.
angle	Yes	Continuous feature that determines the angle in which the shot was taken to the goal.
number_intervening_opponents	Yes	Continuous feature that determines the number of opposing players that were obscuring the goal at the instant of the shot (from the perspective of the shot-taker).
number_intervening_teammates	Yes	Continuous feature that determines the number of teammates that are obscuring the goal at the instant of the shot (from the perspective of the shot-taker).
is_foot	Yes	Boolean feature that indicates whether the shot was taken with the player's foot, or not.
is_head	No	Boolean feature that indicates whether the shot was taken with the player's head, or not.
is_high_interference	Yes	Boolean feature that indicates whether the shooter experience a High level of interference (multiple defenders in close proximity and interfering with the shot).
is_medium_interference	No	Boolean feature that indicates whether the shooter experience a Medium level of interference (a single defender was in close proximity to the shot-taker).
is_low_interference	Yes	Boolean feature that indicates whether the shooter experience a High level of interference (no or minimal interference).
header_distance_to_goalM	Yes	Continuous feature that determines the distance that a headed chance was taken from.

Features Selection 2/2 – Wishlist Features

A selected of key features I would like to include if they were available in a dataset analysed

Feature	Description
is_strong_foot	Categorical feature that determines whether or not the shot was taken with the player's preferred foot. This can be determined from player bio data, or, it by taking a full set of Event data, analysing how many actions were done per foot in a match, per player, and assigning the most used foot as the preferred foot. From the data, we can then compare that player's preferred foot to the foot with which each shot was taken and determine whether it was taken with their strongest foot, using a Boolean <code>is_strong_foot</code> attribute, and also a <code>is_weak_foot</code> attribute.
is_counter_attack	Boolean feature to indicate if the shoot was part of a counter attack or not. This can be determined using the position on the pitch from which a ball was won and the number of opposing defenders behind the ball, relative to the attack team.
is_smart_pass	Boolean feature to indicate whether the assist for the shot broken through the opponents line. This can be used as part of determining the assist type.
is_from_cross	Boolean feature to indicate whether a goal was scored from a cross. This can be used as part of determining the assist type.
time_from_previous_shot	Time in seconds from the last shot of the same team in the same half of the same game. This can be taken on further to determine whether a shot was taken from before and is likely a rebound i.e. <code>is_shot_before</code> or <code>is_rebound</code> . This is of interest as it can be used to determine whether the goalkeeper is out of position and making a reflex save, with the subsequent shot being in a state different to that of a normal shot in that position.
is_1_on_1	Boolean feature to indicate whether the shot was taken from a 1-on-1 situation i.e. the shot taker just has the goalkeeper to beat.
game_state	Categorical feature to indicate whether the shooting team is winning, drawing, or losing. As Michael Caley notes in his Expected Goals model, features for the game state (e.g. winning) appear be capturing latent factors that cannot be observed in the event data, such as the amount of defensive pressure asserted at the time the shot is taken.

[Michael Caley's xG model: cartilagefreecaptain.sbnation.com/2014/9/11/6131661/premier-league-projections-2014#methodology.](#)

[Sam Green's xG model: www.optasportspro.com/news-analysis/assessing-the-performance-of-premier-league-goalscorers/.](#)

[American Soccer Analysis: www.americansocceranalysis.com/explanation.](#)

Feature Engineering and Univariate Analysis

Creating and analysing the features used in the improved Expected Goals model

- **Feature Engineering considerations:**

- Features of interest must be available in both the train and test datasets, and the validation dataset (Metrica Sports data);
- The continuous and discrete features selected are monotonic in nature (Fig. A); and
- Categorical features are required to be one-hot (dummy) encoded, with each value assigned it's own feature.

- **New features created for a Logistic Regression friendly model:**

- Distance from goal – engineered from both the X and Y coordinates;
- Distance from the centre of the pitch – engineered from the Y coordinate;
- Angle to the goal – determined using the distance from the goal and centre of the pitch; and
- Dummy encoding of categorical features including Body Part to is_foot and is_head, and Interference on the Shooter to Low, Medium and High.

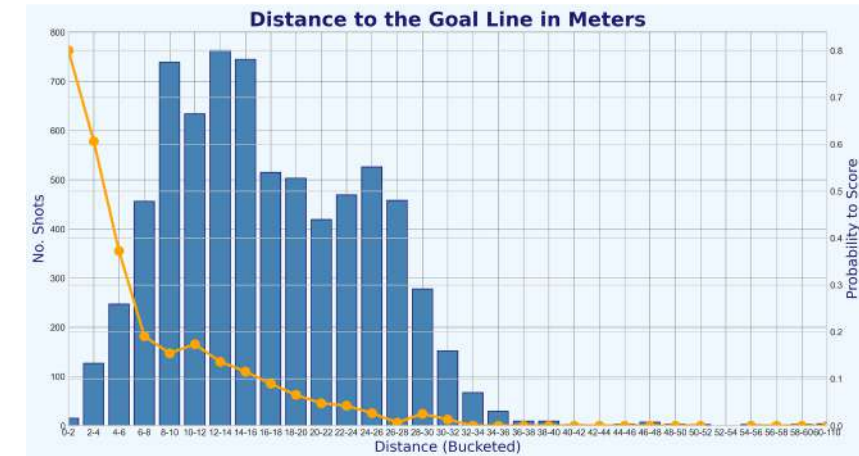
- **Monotonicity of the features:**

- Figure A shows the monotonic relationship of distance of a shot and the probability to score.
- Figure B shows that the interference level on the shooter, the higher the probability that they are to score from the shot (Fig. B).

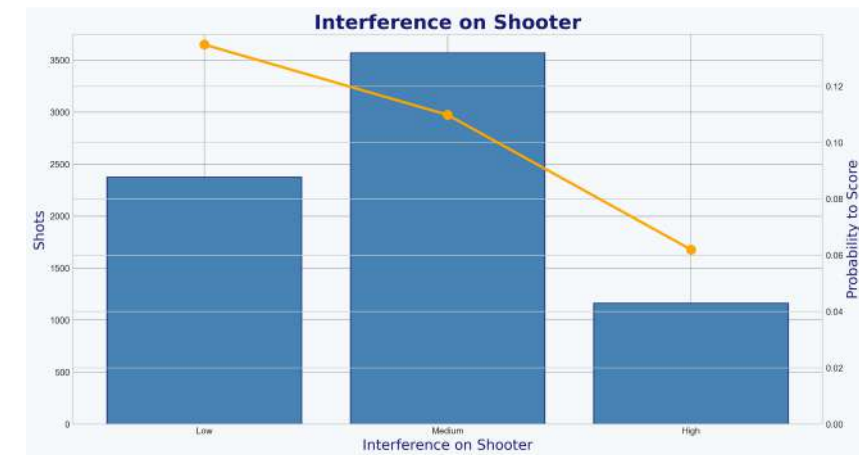
- More information about the monotonicity of each feature can be found in Univariate Analysis (section 8) of the Logistic Regression Expected Goals model notebook [\[link\]](#).

Monotonicity can be defined as a monotonic model is an ML model that has some set of features (monotonic features) whose increase always leads the model to increase its output. See definitions table in the appendix for the full list of features and their definitions, as used in the final trained Expected Goals model.

A



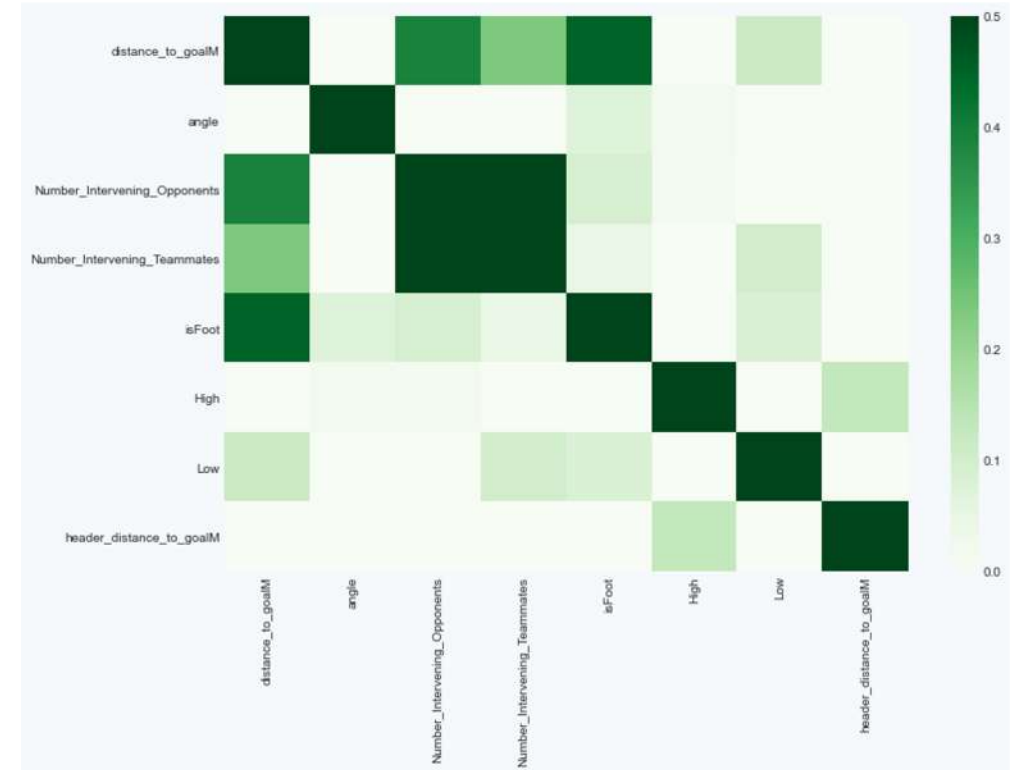
B



Multivariate Analysis

Dealing with highly correlated features

- Considerations for correlation in a nutshell:
 - Highly correlated features might lead to overfitting and have a negative effect on results; and
 - Highly correlated features can very easily mess with the feature interpretation and importance.
- The figure on the right shows the correlation matrix of the final model.
- Highly correlated features such as Distance to the Centre (M), Header were dropped, along with the Medium Interference on the Shooter (when one-hot encoding a categorical column, one feature should be removed).
- The Number of Intervening Opponents and the Number of Intervening Teammates are highly correlated features, but are features that were provided in the raw dataset that I have assumed are not derivatives of each other and not correlated and have therefore both been used in the final model.
- More information about the Multivariate analysis and can be found in section 10 of the Logistic Regression Expected Goals model notebook [\[link\]](#).



Final Model Evaluation and Feature Interpretation

Analysis of the final model before being used for prediction on the Metrica Sports data

- **Final Logistic Regression model performance:**

- Log Loss: 0.289 (reduction from 0.332)
- ROC AUC: 80.7% (increase from 72.8%)

- **Model calibration:**

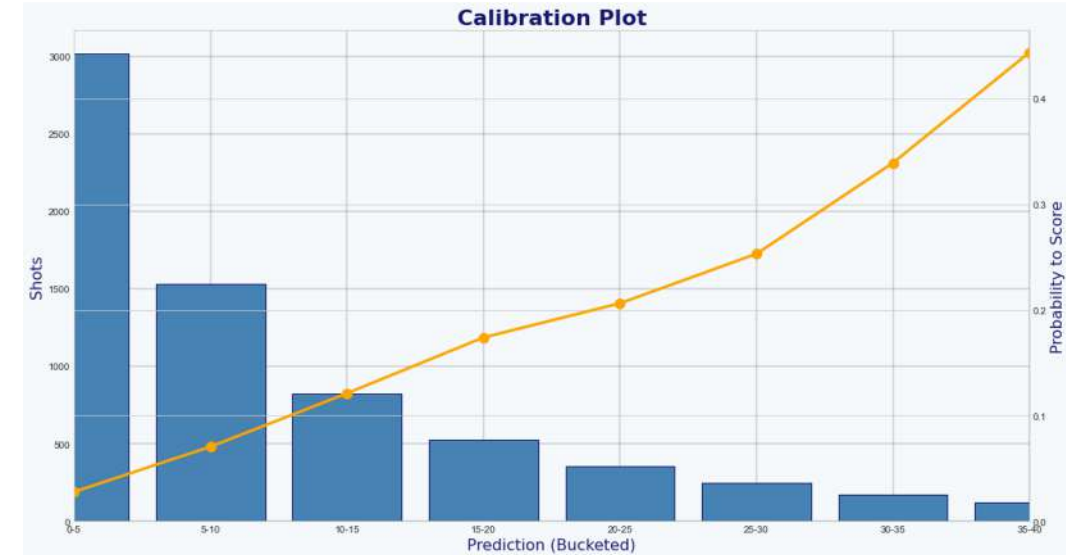
- Calibration plots (Fig. A) are used to model probabilities correctly.
- The observed likelihood of the shots in the dataset resulting in goals is between the 0% and 5% bucket. The model is therefore well calibrated.

- **Feature interpretation:**

- Very important in football analytics as it is the intersection of Data Scientists working with Football Practitioners.
- Finds in the model that leads to probabilities need to be able to be explained in a footballing context. This is done through the Feature Interpretation (Fig. B):
 - The further away the shot, the less likely the shot is to result in a goal;
 - The greater the number of intervening opponents, the less likely the shot is to result in a goal; and
 - The greater the header distance (i.e. the shot is a header), the less likely the shot is a goal (also implies that a header generally has reduced likelihood in resulting in a goal).

- More information about the Final Model and Feature Interpretation can be found in sections 12 and 13 respectively of the Logistic Regression Expected Goals model notebook [\[link\]](#).

A



B

Coefficients

```
-----
distance_to_goalM: -1.072
angle: -0.220
Number_Intervening_Opponents: -0.512
Number_Intervening_Teammates: 0.048
isFoot: -0.131
High: -0.328
Low: 0.082
header_distance_to_goalM: -0.638
```

Model Performance

Iterations of the model during the modeling process

Model	Log Loss	ROC AUC	Accuracy*
Initial	0.33182	72.8%	88.5% ¹
Outlier Removal	0.32670	72.8%	-
Univariate Analysis	0.28979	80.7%	-
Multivariate Analysis	0.28892	80.7%	-
Final Model LR	0.28924	80.6%	-
Secondary Model (XGBoost)	0.28600	-	-

¹The Accuracy metric not included after the initial model as it is not an appropriate metric to measure the performance of an imbalanced, binary classification probability model. More information about the metrics selected and why can be found in the initial Expected Goals modeling notebook:

[github.com/edwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/1\)%20Logistic%20Regression%20Expected%20Goals%20Model.ipynb](https://github.com/edwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/1)%20Logistic%20Regression%20Expected%20Goals%20Model.ipynb).

Notebook to create the XGBoost model:

[github.com/edwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/2\)%20XGBoost%20Expected%20Goals%20Model.ipynb](https://github.com/edwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/2)%20XGBoost%20Expected%20Goals%20Model.ipynb).

Data Engineering of Tracking Data (Validation Data)

Required adjustment made to the data before application of trained Expected Goals model

- **Conversion of Pitch Coordinates:**
 - Original: $x(0, +1)\text{m}$, $y(0, +1)\text{m}$
 - Converted: $x(-53, +53)\text{m}$, $y(-34, +34)\text{m}$ (Fig. A)
- **Reverse player direction (same for the full 90 mins):** Home team right-to-left, Away team left-to-right.
- **Determine each player's speed, acceleration and direction:**
 - Metricka Sports Tracking data is collected at twenty five frames s^{-1} .
 - A player's speed can be calculated by dividing the distance a player has covered between two frames by 0.04 and using a Savitzky-Golay filter for smoothing.
 - The acceleration is calculated as the 2nd order derivative of speed.
- **Subset DataFrames:** Separate Home and Away DataFrames

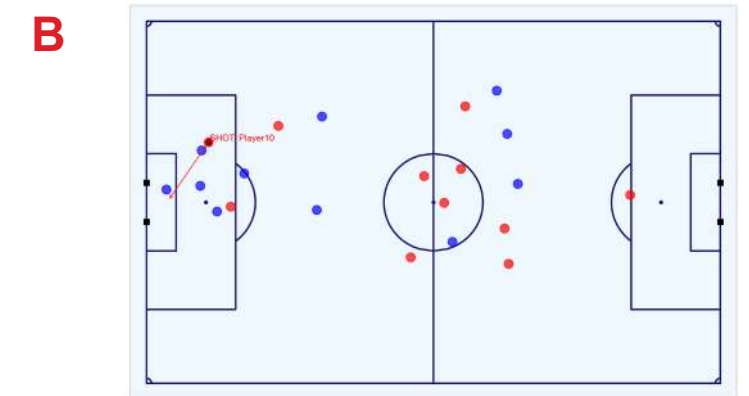


Diagram of the pitch post-conversion credit to [Laurie Shaw](#).

Exploratory Data Analysis of the Match

Visualising the Home team's 3-2 win using the Event and Tracking data

- From the data, we can see that the Home team won the match 3-2.
- As per the submitted, Logistic Regression model, the five best chances in terms of xG are the following:
 - Shot 18 - penalty to the Away team with xG of 0.760 (goal)
 - Shot 16 - shot by the Home team with xG of 0.450 (saved)
 - Shot 7 - shot by the Away team with xG of 0.379 (goal)
 - Shot 9 – headed shot by the Away team with xG of 0.232 (missed)
 - Shot 13 – headed shot by the Home team with xG of 0.251 (goal)
- Two of the goals scored were not in the top five chances in terms of xG:
 - Shot 2 - shot by the Home team with xG of 0.235 – ranked 6th for xG
 - Shot 20 - shot by the Home team with xG of 0.054 – ranked 13th for xG
- This selection of shots includes all the chances with a 20% or greater chance of resulting in a goal.
- More information about each goal and the key chances can be found in the Metrica Sports notebook [\[link\]](#).

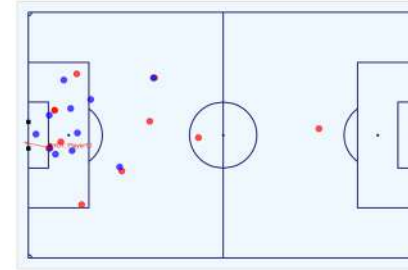


Fig. A: 8m8s - Shot 2 (Home) - goal

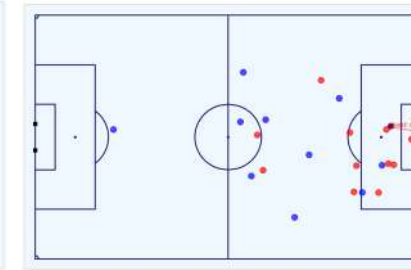


Fig. B: 35m22s - Shot 7 (Away) - goal

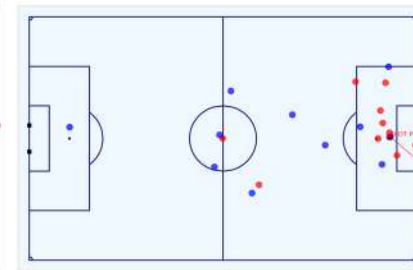


Fig. C: 37m39s - Shot 9 (Away) - no goal

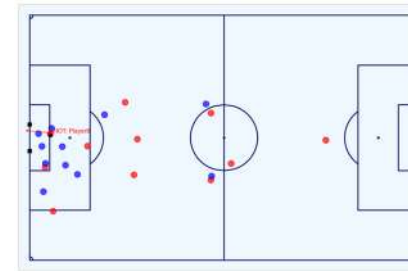


Fig. D: 49m19s - Shot 13 (Home) - goal

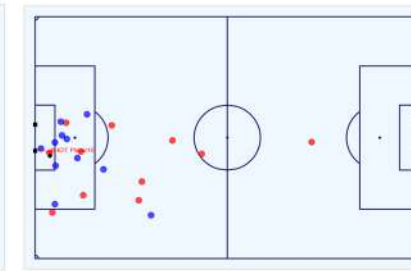


Fig. E: 65m93s - Shot 16 (Away) - no goal

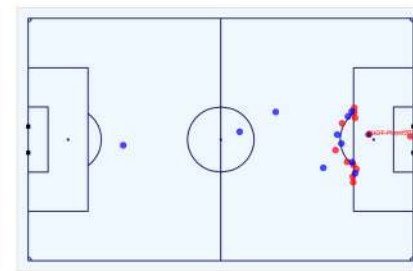


Fig. F: 76m40s - Shot 18 (Away) - goal

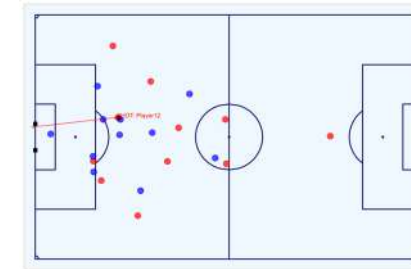


Fig. G: 80m41s - Shot 20 (Home) - goal

All GIF figures produced from Tracking data: github.com/eddwebster/football_analytics/tree/master/gif/fig/metrica-sports.

All MP4 figures produced from Tracking data: github.com/eddwebster/football_analytics/tree/master/video/fig/metrica_sports.

Notebook to work with the Metrica Sports Event and Tracking data: github.com/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/metrica-sports/Metrica%20Sports.ipynb.

Creating Additional Features From Tracking Data

Features for compatibility with the Expected Goals model

- For compatibility with the trained Expected Goals model, a number of key features were required to be derived from the Event and Tracking data. These included:
 - distance to the goal (m);
 - distance to the centre of the pitch (m);
 - angle to the goal (degrees);
 - whether the shot resulted in a goal;
 - whether the shots was a penalty or direct free kick;
 - whether the shot was taken with the player's foot or head;
 - the interference on the shooter; and
 - the number of intervening opponents and teammates.
- From the Tracking data, it was possible to derive both the interference on the shooter and the number of intervening opponents and teammates, to be added to the Events data:
 - A players is determined as interfering if their distance between themselves and the ball is less than the defined radius¹.
 - A player (team mate or opponent) is determined as obstructing the goals from the viewpoint of the shooter (intervening) if they lie within the triangle between both posts and the ball
- The final Events DataFrame is the filtered for just the shots and then export for application with the trained Expected Goals model.

¹StatsBomb use a radius of 5m to define pressure when measuring counter pressing: statsbomb.com/2018/05/how-statsbomb-data-helps-measure-counter-pressing/.

A

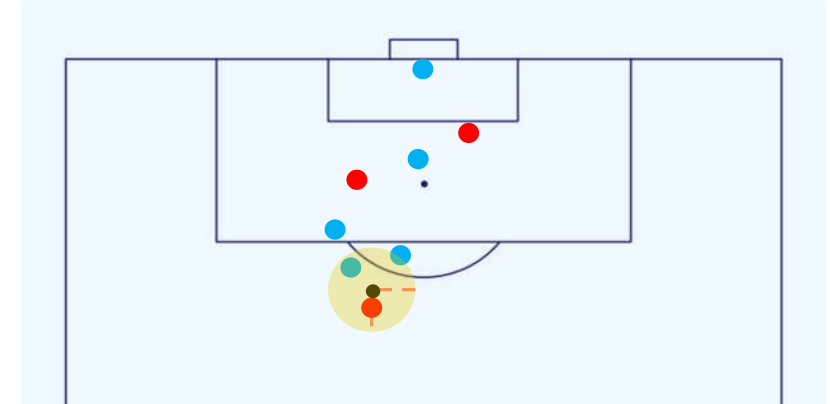


Fig. A: A player is defined as interfering with the shooter if they are within a defined radius of the ball, in this case 5m (not to scale). This can be determined mathematical by observing whether the distance between the player and the ball is less than the defined radius, or not.

B

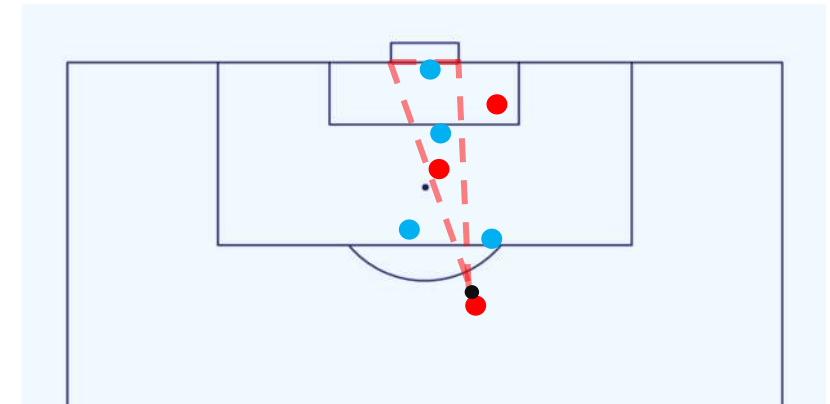


Fig. B: The number of intervening teammates and opponents can be visualised as a triangle between the ball and the posts of the goal at the moment the shot is taken. This interference can be determined mathematical by summing the areas of each triangle.

Which Team Deserved to Win the Game?

Assessment of the chances in the match through the application of the Expected Goals model

- Probabilities of the exported Metrica Sports data for the **Home team** in red and the **Away team** visualised as a cumulative xG race chart (see Fig. A).
- The match featured a penalty, however, the Expected Goals model was trained with just Open Play (OP) shots.
- Penalty kicks all share the same characteristics and were treated in this analysis by assigning values 0.76 xG. as per the value used by StatsBomb/FBref [\[link\]](#).
- What was a very tight contest regarding cumulative xG, 1.43 to the Home side and 1.39 to the Away side (Fig. A), because a clearer projected win for the Away team, once assigned this new xG value for the penalty, 1.43 to the Home team and 2.00 to the Away team (Fig. B).
- The winning goal scored by the Home team in their 3-2 win was an 80th minute was a shot from distance with a low xG value of just 0.054 (ranked 13th out of the 24 shots).
- In this one-off game, the Away team was arguable are more deserving of winning the match by cumulative xG. However, the Home team capitalised on a low-xG chance in the last 10 minutes and went on to win a tight game.



Fig. A: Probability of Scoring race chart before any amendments to xG values.



Fig. B: xG race chart with amended xG value for the penalty.

xG explained by FBref (include penalty xG of 0.76): fbref.com/en/expected-goals-model-explained/

Analysis of game 2 of the Metrica Sports sample data can be found in section 14 and 15 of the Expected Goals model notebook:

[github.com/edwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/1\)%20Logistic%20Regression%20Expected%20Goals%20Model.ipynb](https://github.com/edwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/1)%20Logistic%20Regression%20Expected%20Goals%20Model.ipynb).

Visualisations inspired by [Ben Mayhew's](#) Match Timelines, see: experimental361.com/explanations/match-timelines/

How to Create xG Flow Charts in Python by [McKay Johns](#): youtube.com/watch?v=bvoOOYMQkac&list=PL10a1_q15HwqVEcnqt3tXs1bgvawjsQNW

Comparison of Logistic Regression Model with GBDT Models

Assessment of the chances in the match through the application of the Expected Goals model

- Subsequent Gradient Boosting Decision Trees models were created, notably using [XGBoost](#) (eXtreme Gradient Boosting) and also [CatBoost](#).
- Both has a slightly improved performance in terms of Log Loss compared to the initial model.
- XGBoost and CatBoost are based on decision trees and has a great track record of high performances on modeling structured data due to its performance, flexibility and speed, and is regularly the algorithm that wins [Kaggle](#) competitions.
- Much more information about how the GBDT models were trained, their performance and feature interpretation can be found in the following notebooks:
 - XGBoost [\[link\]](#); and
 - CatBoost [\[link\]](#).

Notebook to create the Expected Goals model using XGBoost (separate notebook to Logistic Regression):

[github.com/eddwesbter/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/2\)%20XGBoost%20Expected%20Goals%20Model.ipynb](https://github.com/eddwesbter/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/2)%20XGBoost%20Expected%20Goals%20Model.ipynb) and CatBoost: [github.com/eddwesbter/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/3\)%20CatBoost%20Expected%20Goals%20Model.ipynb](https://github.com/eddwesbter/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/3)%20CatBoost%20Expected%20Goals%20Model.ipynb)

Analysis of game 2 of the Metrica Sports sample data can be found in section 14 and 15 of the Expected Goals model notebook:

[github.com/eddwesbter/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/1\)%20Logistic%20Regression%20Expected%20Goals%20Model.ipynb](https://github.com/eddwesbter/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling/1)%20Logistic%20Regression%20Expected%20Goals%20Model.ipynb).

A

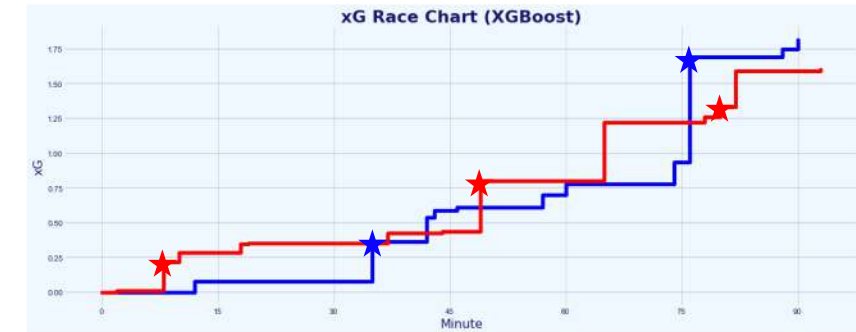


Fig. A: Race chart for the Expected Goals model created using XGBoost.

B

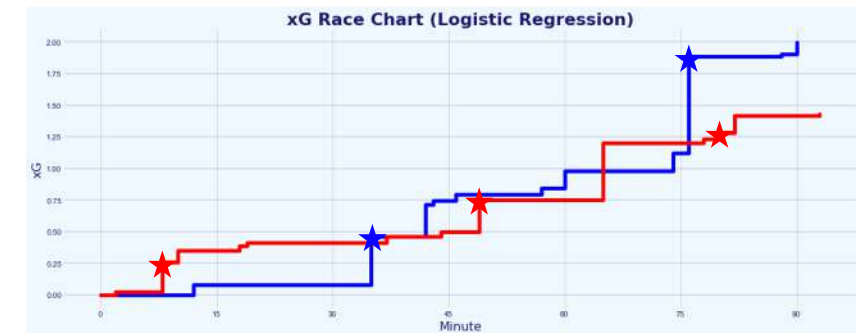


Fig. B: Race chart for the Expected Goals model created using Logistic Regression (see previous slide).

With Great Power Comes Great Responsibility

Expected Goals are useful, but when used incorrectly or in isolation, can also be useless

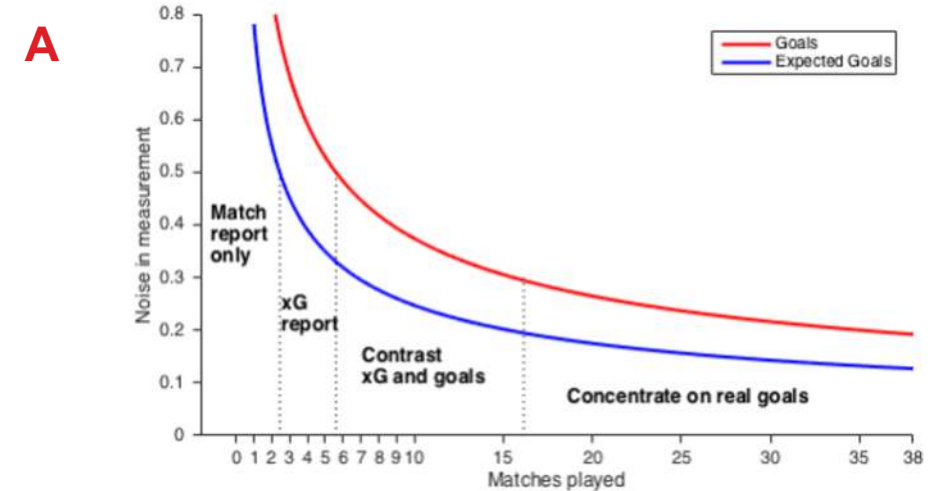
- The concept of Expected Goals is a very useful concept because football is a low-scoring game. Subsequently, it has now entered the mainstream media lexicon¹.
- In his piece 'Should you write about real goals or expected goals?', [David Sumpter](#) demonstrates how regarding Expected Goals, how the magnitude of the noise in the measurement performance decreases with the number of matches played in season (Fig. A)².
- Between 1 and 2 matches, the noise is high for both Expected Goals and real goals. Between 3 to 6 matches, the noise is less than 0.5 goals per match, making an xG report quite functional. Between 7 to 16 matches, both goals and xG have only between 0.3-0.5 goals of noise which allows for comparison. After 16 matches, the difference between the noise in xG and the noise in real goals is only 0.1 goal per match. At this point, real goals are better than Expected Goals².
- The Expected Goals model suggests that the Away team is a more deserving winner of the match by proxy of a greater accumulation of xG. However, any analytical insights and conclusions based over a single 90 minute period should be made with caution.
- This poor usage of Expected Goals can be found in a variety of forms of media (Fig. B, C, and D), where there has been an unfortunate tendency to misuse the metric for individual matches, instead of over several games.
- I would therefore conclude this analysis to say that predictions of any one team's performance over a single 90 minutes is difficult and will be in most cases, inconclusive at best.

¹See [Laurie Shaw](#)'s article: Bodies on the Line: Quantifying how defenders affect chances: eightyfivepoints.blogspot.com/2017/09/bodies-on-line-quantifying-how.html.

²See [David Sumpter](#)'s article: Should you write about real goals or expected goals? A guide for journalists: soccermatics.medium.com/should-you-write-about-real-goals-or-expected-goals-a-guide-for-journalists-2cf0c7ec6bb6.

Caley graphic for Liverpool vs. Atletico Madrid (12/03/2020): twitter.com/Caley_graphics/status/1237870631024095234.

The xG Philosophy tweet for Juventus vs. Porto (09/03/2021): twitter.com/xGPhilosophy/status/1369418380642549762.



Limitations of Expected Goals Models

Just some of the known limitations with current xG modeling

- Any dataset analysed is incomplete. With a full Event dataset, it is possible to include new features such as whether a goal was scored with the player's strongest foot, whether the shot was from a counter attack or fast break, whether the assist was a smart pass, whether the shot was from a cross, what the team's current game state is (winning/drawing/losing). Tracking data can then be used to add more features such as player and goalkeeper's position, the pressure a player is under, the number of obstructing opponents and teammates between the shot taker and the ball. However, despite being the state-of-the-art dataset, Tracking data is still missing significant data points including the body pose position (i.e. are they open to receive a pass) and the spin of the ball (therefore need to build in uncertainties when determining ball trajectory).
- Expected Goals models typically do not include information on the player taking the shot and are therefore an estimate of how the average player or team would perform in a similar situation¹. This is because the model is built with all shot data, not just that individual players. For reference, currently in the 20/21 season after 30 matches, Harry Kane, has had 117 shots (45 on target, 21 goals), 3.9 shots a game. Over a 20 year career with 38 games per season, this would result in just under 3,000 shots. At this point there is a sufficient dataset, but not before.
- Expected Goals models also typically do not use data from after the point the player strikes the ball. This means the model does not include information after the shot, such as the direction, speed, and angle the ball is heading.

¹FBref Expected Goals Explained: fbref.com/en/expected-goals-model-explained.

²Assessing Expected Goals Models. Part 1: Shots by Garry Gelade: web.archive.org/web/20200724125157/http://business-analytic.co.uk/blog/evaluating-expected-goals-models/ (using Wayback Machine).

³A new way to measure keepers shot stopping: post-shot expected goals by Mike Goodman: statsbomb.com/2018/11/a-new-way-to-measure-keepers-shot-stopping-post-shot-expected-goals/.

Expected Goals Modeling Conclusion

Summary of creating an Expected Goals model and applying it to Tracking data

- 1 Defined, built, trained, and evaluated several Expected Goals models from Shots data through the application of both Logistic Regression and Gradient Boosted Decision Tree algorithms including XGBoost and CatBoost, to determine the likelihood that shots taken result in a goal.
- 2 Determined the shots from game 2 of the sample Metrica Sports Tracking and Event data, and made these available for predictions using the Expected Goals model in an appropriate format, deriving additional features from the Tracking data to further enrich the provided Event data, to be used in the Expected Goals model predictions.
- 3 Assessed the performance of the two teams in game 2 of the sample Metrica Sports data through application of the Expected Goals model (derived in step 1) and exported Metrica Sports shot data (derived in step 2), to determine that based on chances created and the Expected Goals predicted, the Away team was more deserving of being the winner of the match, despite losing the game three goals to two to the Home team. However, as this is over just a 90 minute period, this analysis should be treated with caution.

Possibilities and Next Steps of this Expected Goals Model

Some of the steps I plan/wish to take in the near future

- The focus of my approach to create the Expected Goals models was to not build the absolute best performing ML model. My objective was to conduct an end-to-end process for building a model, including all the key stages such as feature engineering, univariate and multivariate analysis, and iterated performance assessment and improvement – for this reason, a simple Expected Goals model was built using Logistic Regression. However, Gradient Boosting Decision Tree models lead to improved performance and for this reason, further models were created using [XGBoost](#) and [CatBoost](#). Potential models that can be deployed to further improve the performance of the Expected Goals model include [LightGBM](#) or Neural Networks such as [PyTorch](#) and [TensorFlow](#).
- Application of a full Event data set, such as those from StatsBomb and Wyscout, to create an Expected Goals model with more features. Such features that were not possible to include in this model from the sample shot dataset but that could be added with Event and/or Tracking data include: strong/weak foot, flag for counter attack, flag for smart pass, determine whether a shot had been immediately taken before, whether the shot was from a cross. This is discussed in more detail in the Feature Engineering section (section 9) of the Expected Goals model notebook [\[link\]](#).
- The model created only considers Open-Play shots. As per Michael Caley's model, it would be interesting to create Expected Goals models that also include Direct Free Kicks and Corners. In Caley's model, he creates six models for different match situations to reflect the varying difficulty of shots in certain scenarios, these include: regular shots, shots from a direct free kick, headed shots from a cross, headed shots not from a cross, non-headed shots from a cross, shots following a dribble from the keeper assuming the goalkeeper is not in goal when the shot is taken. Currently the xG value for penalties was taken from StatsBomb/FBref [\[link\]](#).

Michael Caley's xG model: cartilagefreecaptain.sbnation.com/2014/9/11/6131661/premier-league-projections-2014#methodology

Creating Improved Expected Goals Models from StatsBomb 360 Data

Reproduction of this Expected Goals modeling using StatsBomb 360 data

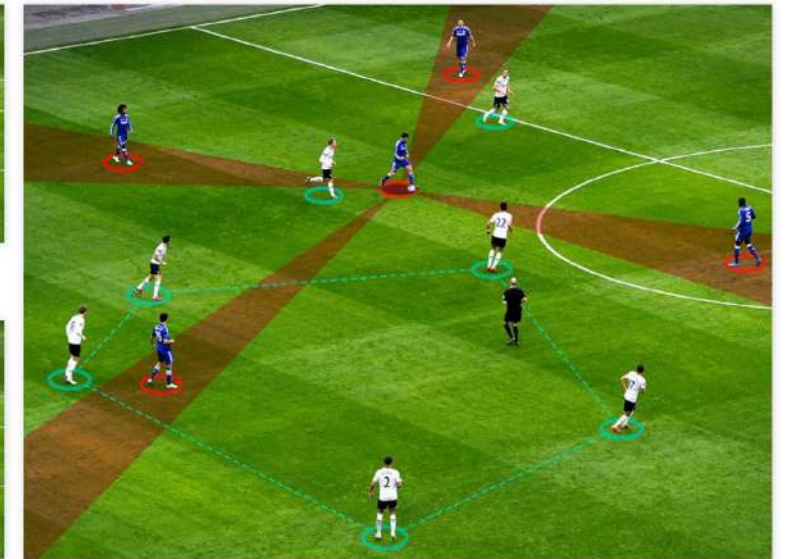
- Since this Expected Goals model was created, StatsBomb have since released a public dataset of their new, cutting-edge 360 data, for UEFA Men's Euro 2020.
- This newly released dataset contains freeze-frames with the positions of all players at every event, adding more context and detail.
- From this dataset, it is possible to derive features not available in standard Event data, such as the number of players between the shot taker and the ball, as well as the pressure that a player is under. The ability to derive such features can help to improve model performance.
- Upon submission of this data science pack, a reworked model using this data is in progress and can be found at the following links:
 - Jupyter Notebook: nbviewer.org/github/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/statsbomb_360_dataset/Introduction%20to%20Building%20Expected%20Goals%20Models%20Using%20StatsBomb%20360%20Data.ipynb
 - Accompanying blog: eddwebster.ghost.io/introduction-to-building-expected-goals-models-using-statsbomb-360-data/



Competitor Data



StatsBomb Data adds more detail and context as standard



StatsBomb 360 shows the entire picture



To find out more about StatsBomb 360 data, see: statsbomb.com/360-data/.

Conclusion

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

Further Work



Conclusion of this Football Data Science Pack

Insight this deck aims to provide

1

This deck demonstrates a selection of football data science projects, solving complex analytics and modeling problems in football.

2

This tools in this deck work from first principles, starting from data acquisition, before being parsed, engineered, matched, visualised, and then analysed in a footballing context, to provide actionable insight.

3

The strong knowledge and understanding of the application of data science methods and football analytics that I possess, to analyse and provide data-driven recommendations through clear communication, that I can bring to Watford Football Club as a Data Scientist.

For More Information

If you would like to find out more...

- **CV:** docs.google.com/document/d/114Lr1ukfHomTKOGX5lj7eTr2Ym3fWwXjfbV8PFZHDQQ
- **Football Analytics GitHub (code):** github.com/eddwebster/football_analytics
- **Google Drive (code + data):** drive.google.com/drive/folders/1yofRp9536WnRL7NKJa5XYjv1vJvDYVbi
- **Slides:** docs.google.com/presentation/d/1a8YgbsXTbpT8FYDvmi-J2eXgN2xnOkUS2E-8YsWrINk
- **Tableau:**
 - Public Profile: public.tableau.com/profile/edd.webster.
 - Dashboards:
 - 2018 FIFA Men's World Cup Center Back analysis: public.tableau.com/app/profile/edd.webster/viz/EddWebster-WorldCup2018AnalysisandDashboard/WC2018PlayerDashboard
 - FA WSL 18/19 and 19/20 analysis: public.tableau.com/app/profile/edd.webster/viz/EddWebster-FAWSLAnalysisandDashboard/WSLxGAnalysisDashboard
 - EFL 18/19 and 19/20 analysis: public.tableau.com/app/profile/edd.webster/viz/EddWebster-EFLAnalysisandDashboards/EFLPlayerDemographicsDashboard
 - 'Big 5' European Leagues analysis: public.tableau.com/app/profile/edd.webster/viz/EddWebster-Big5EuropeanLeagueAnalysisandDashboards/Big5ScoutingDashboard
- **Website:** eddwebster.com
- **LinkedIn:** linkedin.com/in/eddwebster
- **Twitter:** [@eddwebster](https://twitter.com/eddwebster)
- **Email:** edd.j.webster@gmail.com



Notebooks

All code written as part of this analysis

- **Data Scraping:** github.com/eddwebster/football_analytics/tree/master/notebooks/1_data_scraping
 - Capology Player Salary data: github.com/eddwebster/football_analytics/blob/master/notebooks/1_data_scraping/Capology%20Player%20Salary%20Web%20Scraping.ipynb
 - FBref Aggregated Player Performance data: github.com/eddwebster/football_analytics/blob/master/notebooks/1_data_scraping/FBref%20Player%20Stats%20Web%20Scraping.ipynb
 - TransferMarkt Bio and Valuation data: github.com/eddwebster/football_analytics/blob/master/notebooks/1_data_scraping/TransferMarkt%20Player%20Bio%20and%20Status%20Web%20Scraping.ipynb
- **Data Parsing:** github.com/eddwebster/football_analytics/tree/master/notebooks/2_data_parsing
 - StatsBomb Event data: github.com/eddwebster/football_analytics/tree/master/notebooks/2_data_parsing
 - Wyscout Event data: github.com/eddwebster/football_analytics/blob/master/notebooks/2_data_parsing/Wyscout%20Data%20Parsing.ipynb
- **Data Engineering:** github.com/eddwebster/football_analytics/tree/master/notebooks/3_data_engineering
 - Capology Player Salary data: github.com/eddwebster/football_analytics/blob/master/notebooks/3_data_engineering/Capology%20Player%20Salary%20Data%20Engineering.ipynb
 - FBref Aggregated Player Performance data: github.com/eddwebster/football_analytics/blob/master/notebooks/3_data_engineering/FBref%20Player%20Stats%20Data%20Engineering.ipynb
 - StatsBomb Event data: github.com/eddwebster/football_analytics/blob/master/notebooks/3_data_engineering/StatsBomb%20Data%20Engineering.ipynb
 - TransferMarkt Bio and Valuation data: github.com/eddwebster/football_analytics/blob/master/notebooks/3_data_engineering/TransferMarkt%20Player%20Bio%20and%20Status%20Data%20Engineering.ipynb
 - Understat Shots Data Engineering: github.com/eddwebster/football_analytics/blob/master/notebooks/3_data_engineering/Understat%20Data%20Engineering.ipynb
 - Wyscout Event data: github.com/eddwebster/football_analytics/blob/master/notebooks/3_data_engineering/Wyscout%20Data%20Engineering.ipynb
- **Data Unification:** github.com/eddwebster/football_analytics/tree/master/notebooks/4_data_unification
 - Data unification of several datasets: github.com/eddwebster/football_analytics/blob/master/notebooks/4_data_unification/Player%20Golden%20ID%20of%20Football%20Datasets.ipynb
- **Data Analysis:** github.com/eddwebster/football_analytics/tree/master/notebooks/5_data_analysis_and_projects
 - PCA and K-Means Clustering of 'Piqué-like' Defenders: nbviewer.org/github/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/player_similarity_and_clustering/PCA%20and%20K-Means%20Clustering%20of%20%27Pique%CC%81-like%27%20Defenders.ipynb
 - Tracking data:
 - Metrica Sports: github.com/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/tracking_data/metrica_sports/Metrica%20Tracking%20Data%20EDA.ipynb
 - Signality: github.com/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/tracking_data/signality/Signality%20Tracking%20Data%20EDA.ipynb
 - xG Modeling:
 - StatsBomb 360 data: github.com/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/statsbomb_360_dataset/Introduction%20to%20Building%20Expected%20Goals%20Models%20Using%20StatsBomb%20360%20Data.ipynb
 - Shots Event data (model): github.com/eddwebster/football_analytics/tree/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/chance_quality_modeling
 - Tracking data (match applied to): github.com/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/shots_dataset/metrica-sports/Metrica%20Sports.ipynb

Further Work

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

Further Work

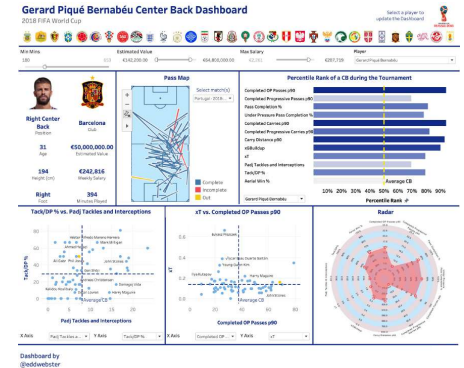


Additional Projects Regarding Football Analysis

Links to additional projects that demonstrate a broad range of tools, skills and analysis when working with football data

- **Recruitment Analysis of Ball-Playing Center Backs at 2018 FIFA World Cup**

- **Description:** Exercise to determine the three best ball-playing center backs from the 2018 FIFA World Cup using Event data from [StatsBomb](#). This project requires the data engineering and aggregation of Event data and derivation of metrics including xGChain and xT.
- **Tableau Dashboard:** public.tableau.com/app/profile/edd.webster/viz/EddWebster-WorldCup2018AnalysisandDashboard/WC2018PlayerDashboard
- **Notebook:** nbviewer.org/github/eddwebster/football_analytics/blob/master/notebooks/3_data_engineering/StatsBomb%20Data%20Engineering.ipynb



- **Creating Expected Goals Models from StatsBomb 360 Data**

- **Description:** Creation of an improved Expected Goals model using StatsBomb 360 data for UEFA Men's Euro 2020. This newly released dataset contains freeze-frames with the positions of all players at every event. From this dataset, it is possible to derive features not available in standard Event data, such as the number of players between the shot taker and the ball, as well as the pressure that a player is under. The ability to derive such features can help to improve model performance. Note that this project is currently in progress.
- **Notebook:** nbviewer.org/github/eddwebster/football_analytics/blob/master/notebooks/5_data_analysis_and_projects/xg_modeling/statsbomb_360_dataset/Introduction%20to%20Building%20Expected%20Goals%20Models%20Using%20StatsBomb%20360%20Data.ipynb
- **Accompanying blog:** eddwebster.ghost.io/introduction-to-building-expected-goals-models-using-statsbomb-360-data/



- **FIFA-TransferMarkt Fantasy Football League**

- **Description:** This project scrapes data for 40+ leagues on TransferMarkt using [Beautifulsoup](#) and matches the data to a FIFA 20 data, a different data source with no robust common identifier, using the [record linkage](#) library, as part of a data-driven approach of £/attribute metrics for a hypothetical recruitment scenario. This data is then exported to Excel as part of a playable fantasy football league.
- **Notebook:** nbviewer.org/jupyter.org/github/eddwebster/fifa-league/blob/master/FIFA%2020%20Fantasy%20Football%20League%20using%20TransferMarkt%20Player%20Valuations.ipynb
- **Microsoft Excel spreadsheet** (for player draft and match result data entry): github.com/eddwebster/fifa-league/blob/master/excel/fifa_20_fantasy_football_league.xlsm
- **GitHub:** github.com/eddwebster/fifa-league



References

Introduction

State of Play
in Football
Analytics

Why Data in
Football is
Important

Sample
Projects
Overview

P1: Football
Intelligence
Tools

P2:
Recruitment
Analysis

P3: Tracking
Data
Applications

P4: xG
Modeling

Conclusion

Further Work



References and Further Reading 1/5

Data Visualisation and Tableau

Articles:

- How to Draw a Football Pitch by [Peter McKeever](https://petermckeever.com/2020/10/how-to-draw-a-football-pitch): petermckeever.com/2020/10/how-to-draw-a-football-pitch
- Match Timelines by [Ben Mayhew](https://experimental361.com/explanations/match-timelines): experimental361.com/explanations/match-timelines

Videos and Webinars:

- Tableau for Sport by [Rob Carroll](https://youtube.com/playlist?list=PLchE8bhmmlxK94imJ4QZncXrbld_NGoiW): youtube.com/playlist?list=PLchE8bhmmlxK94imJ4QZncXrbld_NGoiW
- How To Create xG flow charts in Python by [McKay Johns](https://youtube.com/watch?v=bvoOOYMQkac): youtube.com/watch?v=bvoOOYMQkac
- Dimension of a standard football pitch: en.wikipedia.org/wiki/Football_pitch
- Knutson Radars by [Ted Knutson](https://statsbomb.com/tag/radars): statsbomb.com/tag/radars

Graphics:

- Tableau pitch templates by [James Smith](https://drive.google.com/drive/folders/1TsD5_I3VFIZd0C-APggu3ERHNA9dnjKk): drive.google.com/drive/folders/1TsD5_I3VFIZd0C-APggu3ERHNA9dnjKk Accompanying Medium blog post: medium.com/analytics-vidhya/how-to-create-football-pitches-goals-as-backgrounds-in-tableau-7b1a7800ae1c

Tweets:

- Pros and cons of radar plots by [Luke Bornn](https://twitter.com/LukeBornn/status/864856335191388162): twitter.com/LukeBornn/status/864856335191388162

References and Further Reading 2/5

Recruitment Analysis

Data Sources:

- JSON file of xT values as a 12x8 grid: karun.in/blog/data/open_xt_12x8_v1.json.

Articles:

- [Gerard Moore](#) uses the Event Lab to analyse centre-backs for recruitment: twenty3.sport/gerard-moore-analyst-twenty3-event-lab-recruitment/
- Where art thou German center backs? By [Sam Planting](#) for StatsBomb: statsbomb.com/2019/11/where-art-thou-german-center-backs.
- Introducing xGChain and xGBuildup by [Thom Lawrence](#) for StatsBomb: statsbomb.com/2018/08/introducing-xgchain-and-xgbuildup.
- Using StatsBomb IQ For Player Recruitment: Center Backs by StatsBomb: statsbomb.com/2021/06/using-statsbomb-iq-for-player-recruitment-center-backs.
- StatsBomb Radars by [Ted Knutson](#): statsbomb.com/tag/radars.
- Introducing Expected Threat (xT): modeling team behaviour in possession to gain a deeper understanding of build-up play by [Karun Singh](#): karun.in/blog/expected-threat.html.

Videos and Webinars:

- Expected Threat and Beyond, A Practical Guide. From StatsBomb Innovation in Football, Oct 2019: youtube.com/watch?v=mE3sUVC1wfA.
- How to properly compare players by [Paul Power](#): youtube.com/watch?v=IRg0BCLeitM.

Libraries:

- soccerplots Python package by [Anmol Durgapal](#): github.com/Slothfulwave612/soccerplots.

Tweets:

- Pros and cons of radar plots by [Luke Bornn](#): twitter.com/LukeBornn/status/864856335191388162.

References and Further Reading 3/5

Tracking data

Data Sources:

- Metrica Sports Tracking and correspond Event data: github.com/metrica-sports/sample-data

Tutorials:

- Friends of Tracking Tracking data tutorials by [Laurie Shaw](#) (City Football Group):
 - Part 1 – Introduction to analysing tracking data in Python: youtube.com/watch?v=8TrleFklEsE
 - Part 2 - Measuring the physical performance of players: youtube.com/watch?v=VX3T-4IB2o0
 - Part 3 – Advanced metrics: Pitch Control: youtube.com/watch?v=5X1cSehLg6s
 - Part 4 – Evaluating player actions and passing options: youtube.com/watch?v=KXSLKwADXXI

GitHub Repositories:

- [LaurieOnTracking](#) by [Laurie Shaw](#) for Tracking data implementation

Seminars and Videos:

- How Tracking Data is Used in Football and What are the Future Challenges with [Javier Fernández](#), [Sudarshan 'Suds' Gopaladesikan](#), [Laurie Shaw](#), [Will Spearman](#) and [David Sumpter](#) for Friends of Tracking: youtube.com/watch?v=kHTq9cwdkGA
- 'Demystifying Tracking Data' by [Sam Gregory](#) (Inter Miami) and [Devin Pleuler](#) (Toronto FC): youtube.com/watch?v=miEWHSTYvX4
- 'Classifying and Analysing Team Strategy in Professional Soccer Matches' by [Laurie Shaw](#) at the 2019 Sports Analytics Lab at Harvard University/American Statistical Association Section on Statistics in Sports on 3rd October 2019: youtube.com/watch?v=VU4BOu6VfbU. Paper: static.capabiliaserver.com/frontend/clients/barca/wp_prod/wp-content/uploads/2020/01/56ce723e-barca-conference-paper-laurie-shaw.pdf. Blog: eightyfivepoints.blogspot.com/2019/11/using-data-to-analyse-team-formations.html.
- 'Routine Inspection: Measuring Playbooks for Corner Kicks' by [Laurie Shaw](#) at the 2020 Sports Analytics Lab at Harvard University/American Statistical Association Section on Statistics in Sports on 23rd October 2020: youtube.com/watch?v=yfPC1O_g-l8. Paper: www.springerprofessional.de/en/routine-inspection-a-playbook-for-corner-kicks/18671052.
- Masterclass in Pitch Control by [Will Spearman](#) for Friends of Tracking: youtube.com/watch?v=X9PrwPyolyU.
- 'A framework for tactical analysis and individual offensive production assessment in soccer using Markov chains' by [Sarah Rudd](#) (Arsenal FC) at the 2011 Sports Analytics Lab at Harvard University/American Statistical Association Section on Statistics in Sports on 3rd October 2019 [\[link\]](#).

References and Further Reading 4/5

Expected Goals modeling

Seminars and Videos:

- The Ultimate Guide to Expected Goals by [David Sumpter](#) (Hammarby) for Friends of Tracking: [youtube.com/watch?v=310_eW0hUqQ](https://www.youtube.com/watch?v=310_eW0hUqQ)
- How to explain Expected Goals to a football player by [David Sumpter](#): [youtube.com/watch?v=Xc6IG9-Dt18](https://www.youtube.com/watch?v=Xc6IG9-Dt18)
- What is xG? by [Alex Stewart](#) for Tifo Football: [youtube.com/watch?v=zSaeaFcm1SY](https://www.youtube.com/watch?v=zSaeaFcm1SY)
- Opta Expected Goals presented by [Duncan Alexander](#): [youtube.com/watch?v=w7zPZsLGK18](https://www.youtube.com/watch?v=w7zPZsLGK18)
- Sam Green OptaPro Interview: [youtube.com/watch?v=gHIY-MqDh_o](https://www.youtube.com/watch?v=gHIY-MqDh_o)
- Anatomy of a Goal (with Sam Green) for Numberphile: [youtube.com/watch?v=YJuHC7xXsGA](https://www.youtube.com/watch?v=YJuHC7xXsGA)

Tutorials:

- Friends of Tracking Expected Goals tutorials by [David Sumpter](#):
 - Part 1 – How to build an Expected Goals model 1 – Data and model: [youtube.com/watch?v=bpiLyFyLIXs](https://www.youtube.com/watch?v=bpiLyFyLIXs). See GitHub: xG model: github.com/Friends-of-Tracking-Data-FoTD/SoccermaticsForPython/blob/master/3xGModel.py
 - Part 2 – How to build an Expected Goals model 2 – Statistical fitting: [youtube.com/watch?v=wHOqIJN5q54](https://www.youtube.com/watch?v=wHOqIJN5q54). See GitHub: Linear regression: github.com/Friends-of-Tracking-Data-FoTD/SoccermaticsForPython/blob/master/4LinearRegression.py, xG model fit: github.com/Friends-of-Tracking-Data-FoTD/SoccermaticsForPython/blob/master/5xGModelFit.py

Notable xG models:

- Sam Green: www.optasportspro.com/news-analysis/assessing-the-performance-of-premier-league-goalscorers
- Michael Caley: cartilagefreecaptain.sbnation.com/2014/9/11/6131661/premier-league-projections-2014#methodology
- American Soccer Analysis: www.americansocceranalysis.com/explanation

Professional and Fanalyst Examples:

- An xG Model for Everyone in 20 minutes (ish) by [Paul Riley](#): differentgame.wordpress.com/2017/04/29/an-xg-model-for-everyone-in-20-minutes-ish
- Tech how-to: build your own Expected Goals model by [Jan Van Haaren](#) and SciSports: www.scisports.com/tech-how-to-build-your-own-expected-goals-model. For code, see: bitbucket.org/scisports/ssda-how-to-expected-goals/src/master
- soccer_analytics repository by [Kraus Clemens](#): github.com/CleKraus/soccer_analytics
 - Expected goal model using Logistic Regression: github.com/CleKraus/soccer_analytics/blob/master/notebooks/expected_goal_model_lr.ipynb
 - Challenges using Gradient Boosters: github.com/CleKraus/soccer_analytics/blob/master/notebooks/challenges_with_gradient_boosters.ipynb
- Expected Goals thesis by [Andrew Rowlinson](#): github.com/andrewRowlinson/expected-goals-thesis
- Expected Goals deep dive by [Andrew Puopolo](#): github.com/andrewsimplebet/expected_goals_deep_dive
- Fitting your own football xG model by [Ismael Gomez](#): www.datofutbol.cl/xg-model
- Python for Fantasy Football by [Fantasy Futopia](#) (Thomas Whelan): www.fantasyfutopia.com/python-for-fantasy-football-introduction-to-machine-learning

Articles:

- xG explained by FBref: fbref.com/en/expected-goals-model-explained
- Bodies on the Line: Quantifying how defenders affect chances by [Laurie Shaw](#): eightyfivepoints.blogspot.com/2017/09/bodies-on-line-quantifying-how.html
- Should you write about real goals or expected goals? A guide for journalists by [David Sumpter](#): soccermatics.medium.com/should-you-write-about-real-goals-or-expected-goals-a-guide-for-journalists-2cf0c7ec6bb6
- How data availability affects the ability to learn good xG models by [Jesse Davis](#) and [Pieter Robberechts](#): dtai.cs.kuleuven.be/sports/blog/how-data-availability-affects-the-ability-to-learn-good-xg-models
- Expected Goals and Unexpected Goals by [Garry Gelade](#): web.archive.org/web/20200724125157/http://business-analytic.co.uk/blog/expected-goals-and-unexpected-goals
- Assessing Expected Goals Models. Part 1: Shots by [Garry Gelade](#): web.archive.org/web/20200724125157/http://business-analytic.co.uk/blog/evaluating-expected-goals-models
- Assessing Expected Goals Models. Part 2: Anatomy of a Big Chance by [Garry Gelade](#): web.archive.org/web/20200724125157/http://business-analytic.co.uk/blog/assessing-expected-goals-models-part-2-anatomy-of-a-big-chance

Literature:

- Expected Goals literature by [Keith Lyons](#): docs.google.com/document/d/1OY0dxqXIBgncj0UDgb97zOtczC-b6JUknPFWgD77ng4/edit

References and Further Reading 5/5

Further resources, libraries and miscellaneous

Books:

- The Numbers Game by [Chris Anderson](#) and [David Sally](#)
- Soccermatics by [David Sumpter](#)

Vendor Documentation:

- Metrica Sports documentation: github.com/metrica-sports/sample-data/blob/master/documentation/events-definitions.pdf
- StatsBomb documentation: github.com/statsbomb/open-data/blob/master/doc/StatsBomb%20Open%20Data%20Specification%20v1.1.pdf
- Wyscout documentation: apidocs.wyscout.com

Libraries and GitHub Repositories:

- mplsoccer by [Andrew Rowlinson](#): github.com/andrewRowlinson/mpsoccer
- SoccermaticsForPython by [David Sumpter](#): github.com/Friends-of-Tracking-Data-FoTD/SoccermaticsForPython
- LaurieOnTracking by [Laurie Shaw](#): github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking

Python Libraries:

- [pandas](#)
- [Matplotlib](#)
- [Plotly](#)
- [Beautiful Soup](#)
- [Record Linkage](#)
- [scikit-learn](#)
- [XGBoost](#)
- [CatBoost](#)

Resources:

- Concise list of publicly available Football Analytics resources by [Edd Webster](#): github.com/eddwebster/football_analytics
- YouTube playlists by [Edd Webster](#):
 - Sports Analytics / Data Science: youtube.com/playlist?list=PL38nJNjpNpH9OSeTqnnVeKkzHsQUJD70
 - Expected Goals (xG): youtube.com/playlist?list=PL38nJNjpNpH_VPRZJrkaPZOJfyulaZHUY
 - Tracking data: youtube.com/playlist?list=PL38nJNjpNpH-UX0YVNu7oN5gAWQc2hq8F

A wide-angle photograph of a football stadium, likely Watford's Vicarage Road, taken from a high vantage point in the stands. The pitch is green with white markings, and several players are visible. The stands are packed with fans, many of whom are waving large black and white checkered flags. The sky is overcast. The text "Thank You!" is superimposed in the center of the image.

Thank You!