

DataScouting Example (2)

wysco

PLAYER	KEANE	MEE	TARKOWSKI
TEAM (01/2016)	Burnely	Burnley	Brentford
Appearances	28	29	31
Aerial Duels Won	55%	56%	60%
Average Pass Length	22m	23m	25m
Average Pass Accuracy	78%	77%	74%
Yellow Cards	3	5	4
Red Cards	0	0	0

<https://youtu.be/tcGaph1XNPk>

EVENTS Pack

wyscout

ent by event for every match:
ayer, timing, event details, start and
d positions, outcome, attributes

USED FOR

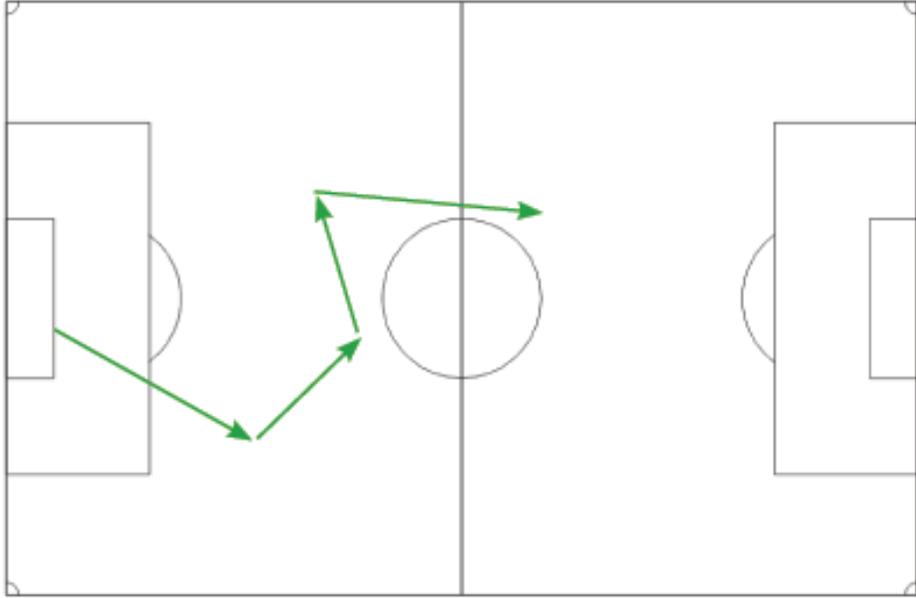
Advanced match analysis, statistical
udies, styles of play, advanced
ayer rating and comparison

USED BY

Advanced match analysts, data
ientists, researchers, odds makers



Events Data Content

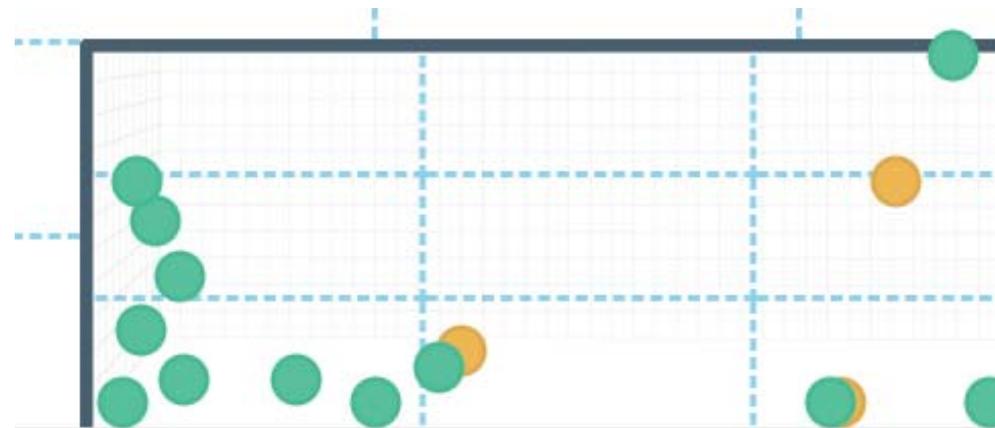


- period half
- timing
- team
- player
- action type & subtype
- x/y coordinates (start & end)

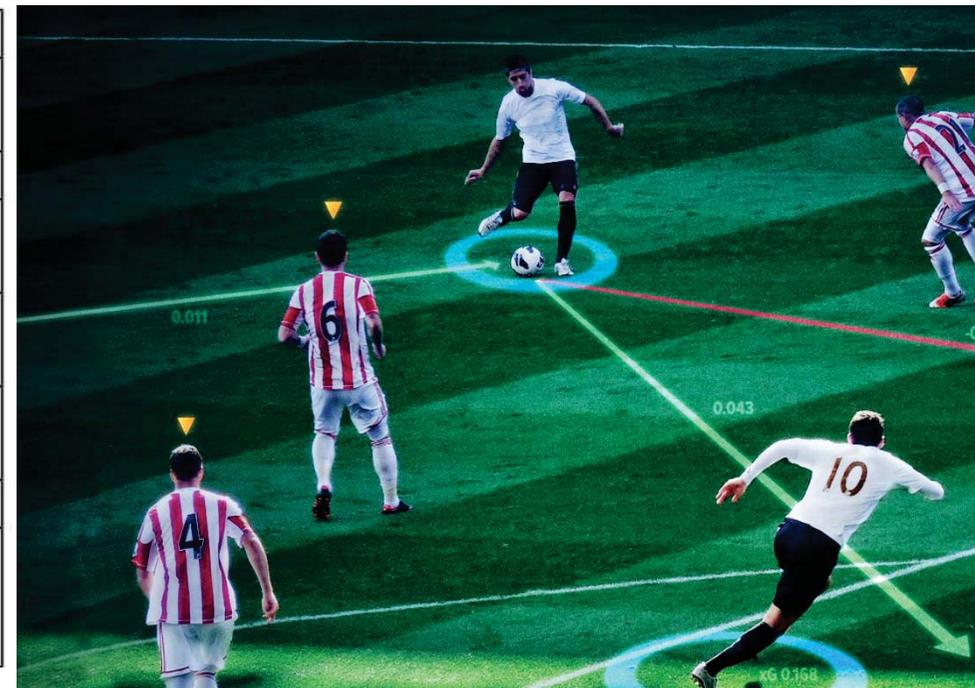
HALF	TIME(S)	TEAM	PLAYER	TYPE	SUBTYPE	START_X	END_X	START_Y	END_Y
1	8.642	679	217031	8	85	58	66	34	9
1	10.167	679	86307	8	85	66	85	9	17
1	11.987	679	3443	8	85	85	90	17	25

Events Data Content (2)

- action type and subtype
- outcome & additional tags
- shot destination



type	subtype	tags
pass	cross, simple pass	accurate, not accurate, key pass, opportunity, assist, (goal)
card		no card, yellow, red, 2nd yellow
shot		accurate, not accurate, block, opportunity, assist, (goal)
duel	air duel, dribbles, tackles, ground loose ball	accurate, not accurate
kick	corner, shot, goal kick, throw in, penalty, simple free kick	accurate, not accurate, key pass, opportunity, assist, (goal)
side touch	acceleration, clearance, touch	counter attack, dangerous ball lost, missed ball, interception, opportunity, assist, (goal)



Events Data

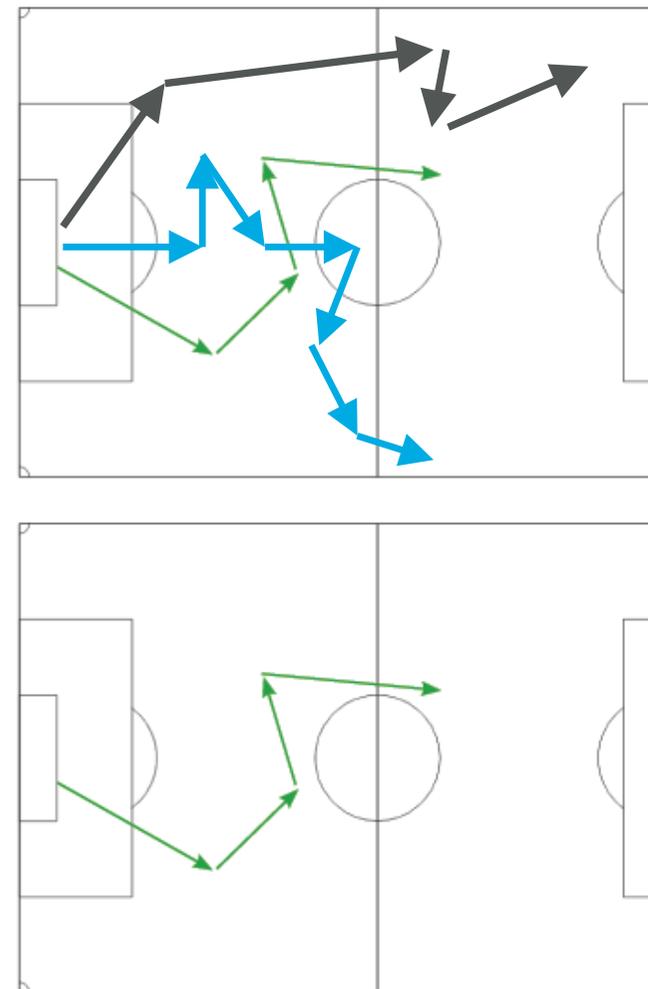
Action breakdown

wyscout

The event flow for each game can be split into sequences of events (construction, offensive and defensive transitions, etc.)

Each sequence can be analyzed to easily select and highlight the most interesting situations for the coach/analyst, in a while over thousands of matches. To gain more insight into a team's **playing style** or individual player's **contribution** to each action.

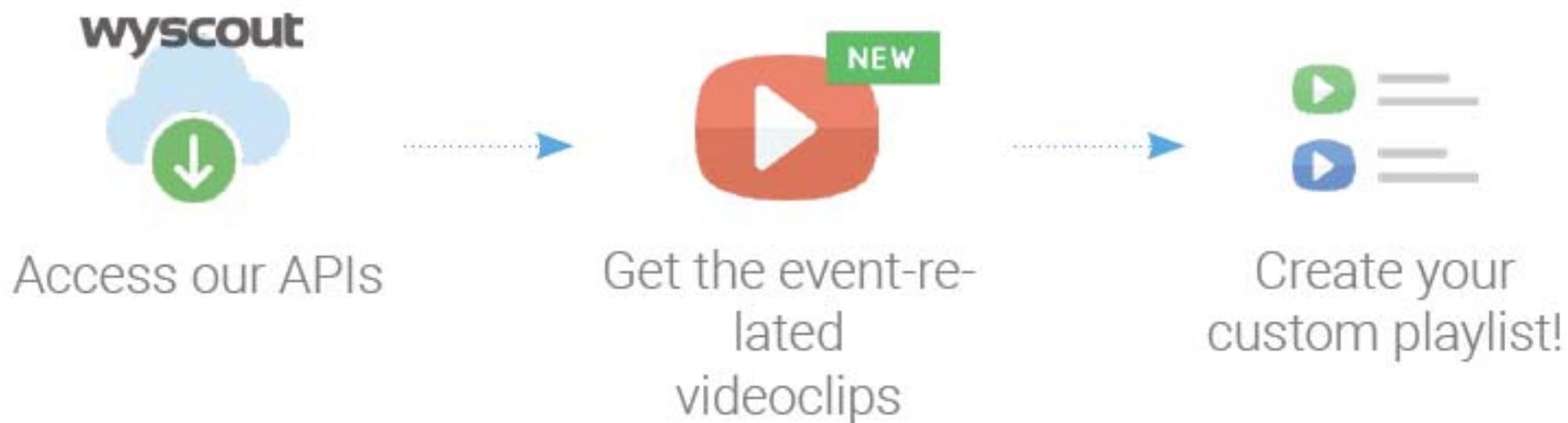
Selected events can be visualized with Videos API.



What can I do with Wyscout Events Data

wyscout

Every single data event is linked to the related videoclip.

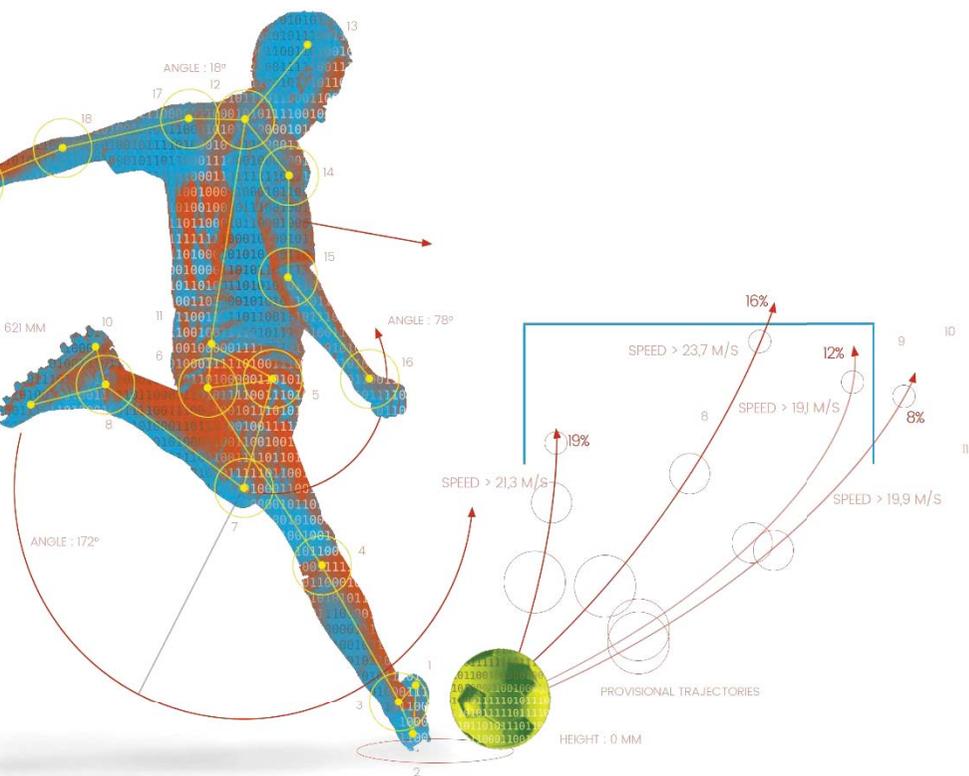


The power of Wyscout is now available to everyone (CLUBS ONLY)

Metrics: Expected Goals

wyscout

xG calculated for every shot: an indication of how many goals a player or team should have scored on average, given the shots they have taken.



xG is an advanced metric that assigns to every shot a probability (based on historical stats) of how likely it is to score from the position, type of assist, preceding events, etc.

A modern metric that allows to study the results of the matches with more stress on predictability of moments creation and less on pure luck.

Big Data: A Trend in the Football Industry wyscout

"Big data in football? We're just at the beginning of a long journey. In some years data analysis will be applied at any level in the clubs. At Wyscout, we have one of the biggest and deeper data sets, and we just started discovering its potential".

Our huge database of football data and statistics can be easily used as the best provider of football data for Clubs, Federations, Agencies, Media Companies, Apps, Online Games, Predictions Websites and Journalists.



Data-Supported Decisions

Data are already wide used in football and are still a trend, another close development is integration on the business side.

In business world DSDS is an information system that supports decision-making activities. Final goal is to use BA&BI to take decision supported by data.

Clubs are business organizations and in the long run they will likely switch to Data Supported Decision System



wyscout
the football company

Data collection procedure

Performed by expert video analysts through a proprietary software (the tagger).

The tagging of a match consists of three main steps.

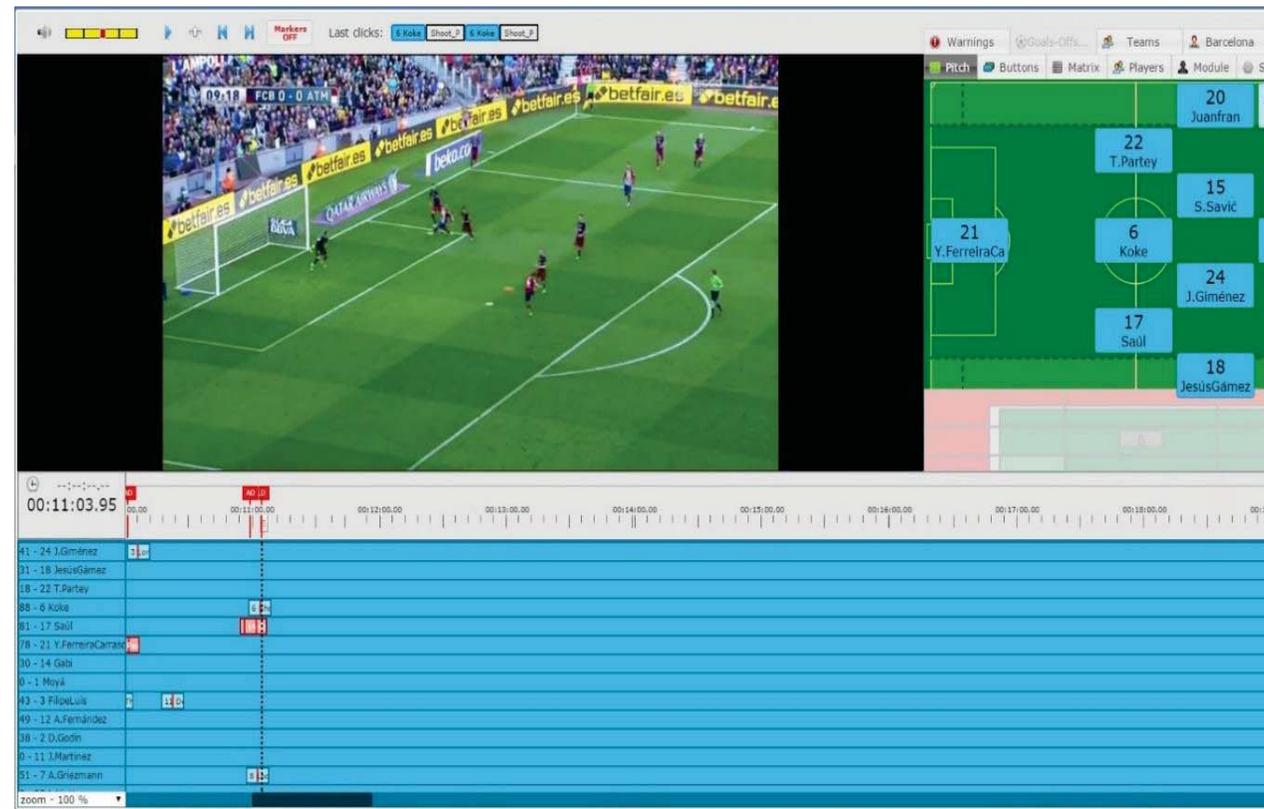
Step 1: setting formations.

At the beginning of the match, an operator sets the teams' starting formations, the positions of the players on the pitch and their jersey number.

Step 2: event tagging.

For each ball touch in a match, the operator selects one player and creates a new event on the timeline.

The operator then adds the type (e.g., pass, duel, shot, etc.) and subtype (e.g., a duel can be aerial or ground) of the event by using a special custom keyboard which gives operators the possibility to insert events and data in a streamlined way.



Data collection procedure (2)

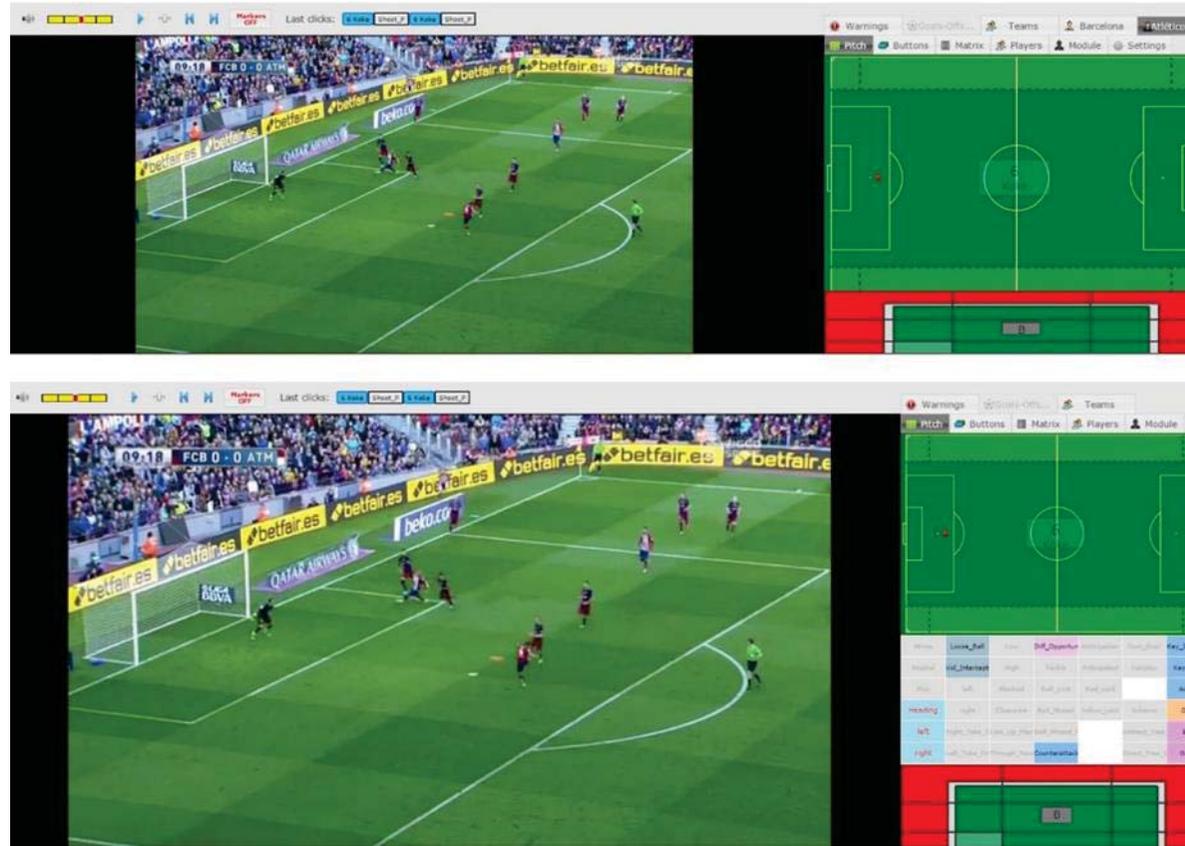
Performed by expert video analysts through a proprietary software (the tagger).

The operator finally adds the coordinates on the pitch and all the additional attributes for the event.

These can be different depending on the event type: such as pass high/low, foot, dribbling side and so forth.

When a player shoots on goal the system asks the operator to fill a shot specific module that collects where the shot ends (on goal, out of goal, on post and exact position).

Accuracy of data is guaranteed by the fact that each single match is tagged by a team of three operators, one of them acting as responsible supervisor of the output of the whole match.



Data collection procedure (3)

Automatic and manual quality control

Step 3: quality control.

After the tagging, a procedure of quality control for each match is performed, mainly consisting of two different steps.

The first step is automatic: an algorithm is used to avoid the majority of the errors made by operators, considerably reducing the margin of error.

For example the algorithm matches the events tagged by both operators to crosscheck if they both collected events involving both teams, like duels, with the same positioning and interpretation. Similarly, the algorithm suggests events missed by the operators or search for impossible combinations of event sequences.

The second step of quality control is manual and supervised by quality controllers.

It mainly consists of an in-depth check that is carried out once the match is completed. Going through each event of some sample matches, the controller can see and eventually correct any entered parameter.

Sample matches for quality control are chosen by another algorithm in order to guarantee a well distributed and statistically meaningful coverage with respect to the kind and amount of analyzed matches.

Football performance data sources

There are three main data sources on the market nowadays

1. Events data, soccer-logs that describe the events that occur during a match and are collected through proprietary tagging software
2. Video-tracking data describe the trajectories of players during a match and are collected through video recordings
3. GPS data describe the trajectories of players during (mainly) training sessions and are collected through GPS devices worn by the players.

Despite this wealth of data, we cannot avoid noticing that soccer datasets are rarely available for scientific research, a clear limit to the developing of scientific methods for soccer analytics.

Open source match event dataset

https://figshare.com/collections/Soccer_match_event_dataset/4415000/1

This dataset has been used recently during a data challenge and, to the best of our knowledge, it is the largest public collection of events data.

The Soccer Data Challenge, <https://sobigdata-soccerchallenge.it/>

Events data describe events, each containing information about its type (pass, shot, etc.), a time-stamp, the player(s), the position on the field and additional information (e.g., pass accuracy).

We believe that these data are greatly beneficial to the scientific community because they can contribute to foster research in several directions.

The datasets are released under the Open Database License (ODbL) and are publicly available on FigShare.

Dataset's content

The data covers a total of around 1.941 matches, 3.251.294 events and 4.299 players.

The data refer to season 2017/2018 of five main national soccer competitions in Europe:

- La Liga (Spain)
- Serie A (Italy)
- Premier League (England)
- Bundesliga (Germany)
- Ligue 1 (France)

In addition, we provide the data of the FIFA World Cup 2018 and the UEFA Euro Cup 2016.

We provide five datasets corresponding to information about all competitions, matches, teams, players and events in the data.

Each dataset is provided in JSON format (JavaScript Object Notation), an open-standard file format that uses human-readable and machine-processable text to transmit data objects consisting of attribute-value pairs and array data types (or any other serializable value).

Dataset COMPETITIONS

Each competition is a document consisting of the following fields

```
{"area": {"alpha2code": "IT",  
  "alpha3code": "ITA",  
  "id": "380",  
  "name": "Italy"},  
  "category": "default",  
  "format": "Domestic league",  
  "gender": "male",  
  "name": "Serie A",  
  "type": "club",  
  "wyId": 524}
```

- area: it denotes the geographic area associated with the league, using the ISO 3166-1 specifications;
- category: it indicates the age category of the league. All competitions in the portion of data we release to the public refer to the category "default";

Dataset COMPETITIONS (2)

Each competition is a document consisting of the following fields

- format: the format of the competition.

All competitions for clubs (La Liga, Serie A, Premier League, Bundesliga, Ligue 1) have value "Domestic league".

The competitions for national teams (FIFA World Cup 2018, UEFA Euro Cup 2016) have value "International cup";

- gender: the gender of the players. The value is "male" for all the five competitions as the data refer to male soccer;
- name: the official name of the competition (e.g., Serie A, Bundesliga, etc.);
- type: the typology of the competition.

It is "club" for the competitions for clubs (La Liga, Serie A, Premier League, Bundesliga, Ligue 1) and "international" for the competitions for national teams (FIFA World Cup 2018, UEFA Euro Cup 2016);

- wyld: the unique identifier of the competition, assigned by Wyscout.