Data Scientist

Thanks again for your application for the **Data Scientist** position at Parma Calcio. We really enjoyed reviewing your application and are happy to invite you to the next stage of our process: **the technical interview**.

📣 NO prior football knowledge or experience is required. This document will contain all the information you need to know for the assignment. Please, make sure you read it carefully.

You will never be judged based on the number of clarifying questions you ask. Should you have ANY questions, don't hesitate to send an email to mmatteotti@parmacalcio1913.com.

The interview will be in **English** and will consist of a 60-minute discussion with Matteo Matteotti, Head of Data, Daniel Montalbano, Data Engineer, and Nick van Lieshout, Data Visualization Specialist.

The interview will be divided into 3 sections:

1. **Showcase of one of your projects (approximately 20 minutes)**: you will present to us one project you worked on in the past and that you think is relevant for the position. It doesn't have to be based on football data. It can be an academic project, a University assignment, something you worked on in your spare time or in a previous professional experience. We are interested in understanding the problem you were trying to address, the overall idea behind the methodology and the results you achieved. At this stage, we are not interested in your code - simply prepare a few (A FEW!) slides to show during the interview.
2. **Solution of the assignment (approximately 30 minutes)**: you will talk us through your solution to the assignment which comes with this document (more to it later). Here is where we want to see code: there is no need to discuss every single line of code, but we want to fully understand what each section does. We are interested in evaluating your problem-solving skills and your coding ability. We might ask you to change parts of the code live, just to check if you have full control on it, or if you let an AI generate it for you.
3. **Questions (approximately 10 minutes):** in the last 10 minutes of the interview, you will be presented with a problem we had in the past and we will ask you to come up with a high-level solution to tackle it. We are interested in assessing your ability to come up with solutions to unseen problems. You will not be coding live, you will simply talk us through your solution.

The assignemnt

Preliminary knowledge

When it comes to tactical data (i.e., data of a game), there are two types of data in football:

- event data
- tracking data.

Event data describes the sequence of on-ball events (passes, shots, tackles, interceptions, goalkeeper's saves, ...) which occurred during a game, from the kick-off to the final whistle. Structure may differ from provider to provider, but typically they are collected in tabular format, with one row per event. For each event (i.e., for each row) the columns collect the different *qualifiers* that describe an event: what type of event (pass?, shot?, tackle?, ...), which player made that event, in what location of the pitch, etc. A "regular" 90-minute game usually contains between 2000 and 3000 events. Data size is then relatively small (few MB), and you obtain a good overview of what happened ON THE BALL during a game.

Though they are a good starting point, event data completely overlooks what happens OFF THE BALL: in a likely scenario where the striker comes short to attract his centre-back and create space for the winger who is served with a through ball behind the line by the midfield, event data only collects information on the midfielder's pass.

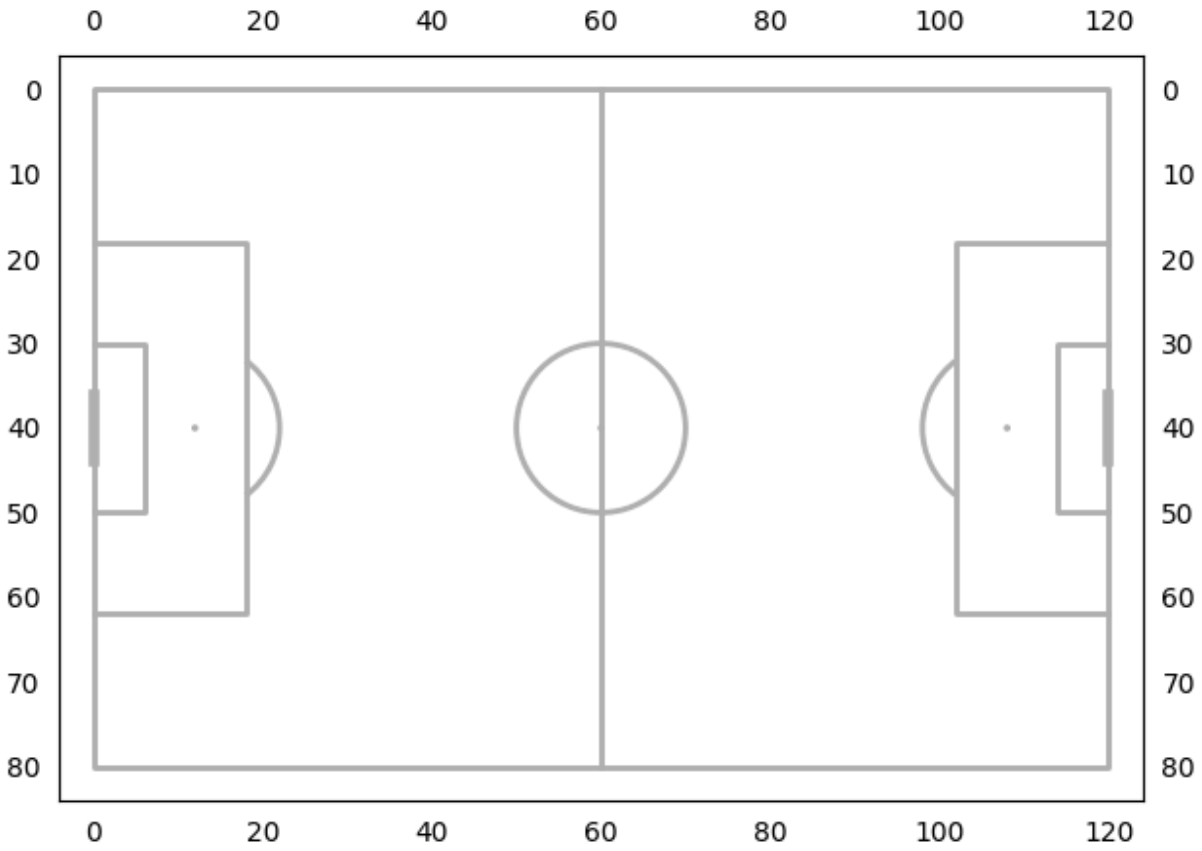
In light of the complex dynamic nature of football where a player usually spends 98% of his playing time OFF THE BALL, event data is not enough.

On the other hand, tracking data collects the location of the ball and all players on the pitch at a rate of 25 FPS. For each data point, you know which player it refers to, the timestamp and the corresponding location on the pitch in a X;Y coordinate system. This amounts to more than 3 million data points in a regular 90-minute game.

Although tracking data offers a deeper overview on players' movements, it does not describe the sequence of events. One could tell a player "hit" the ball by studying their movement in relation to the ball's movement, but it's much more difficult to tell whether that event was a pass, a shot, a cross, a clearance, etc.

Only by synchronizing event with tracking data, one can have a full grasp of what happened on the ball (through event data) and off the ball (with tracking data).

StatsBomb (one of the main football data providers) offers a huge collection of event data on their GitHub. For certain games, each shot also comes with the *freeze frame*, that is, the tracking data for that specific shot frame for the players within a certain radius from the ball. Statsbomb events are defined on a 120×80 pitch, with both teams always attacking left to right.



💡 To better understand how the StatsBomb repository works, we suggest you to have a look at this tutorial.

To better understand how StatsBomb event data works, we suggest you to have a look at this tutorial and this tutorial (the latter also shows an example of freeze frame towards the end of the video).

[Not necessary for the resolution of the assignment] - To better understand how tracking data works, we suggest you to have a look at this tutorial. It uses Metrica tracking data.

For a comparison between event and tracking data, we suggest you to have a look at this tutorial, between minutes 7:10 and 8:30.

Task #1 – Train an xG model with StatsBomb open data

Your task is to train an expected goal (xG) model, using StatsBomb open data. An expected goal (xG) model assigns to each shot the probability of being scored.

You can choose whether you want to base it only on event data, to base it only on the tracking data contained in each shot's freeze frame, or to blend the two types of data into one model. You are free to make assumptions and take decisions to simplify your job, as long as they are clearly stated and you are able to defend them. You can decide whether to use all games at your disposal, or only some. You have full freedom on the type of model you want to work on: it can be based on maths, rules, machine learning, deep learning, neural networks, etc.

You can only use the data you find in the GitHub repository.

Task #2 - Who should have won the Ballon d'Or in 2015/2016, according to data?

In their GitHub, StatsBomb also released free data for all games of the Big 5 leagues (Italian Serie A, English Premier League, Spanish La Liga, German Bundesliga, French Ligue 1) in season 2015/2016. Your task is easy: based ONLY on the data you can find in StatsBomb for these games (you are not allowed to use any other sources), who should have won the Ballon d'Or that season? In other words, who was the best player in that set of games, according to data?

Once again, you are free to make assumptions and take decisions to simplify your job, as long as they are clearly stated and you are able to defend them. There is not a unique definition of *best player*, so you can address the problem from multiple angles: through a rule-based model, through machine learning, through mathematical formulas, etc.

You can only use the data you find in the GitHub repository. If you want, you can use the other games in the repository to train a model, and use the games of the Big 5 Leagues in season 2015/2016 as your test set.

Submission

Push your project to a private GitHub repo. You should only create one repository, but you can structure it as you prefer. Make sure you include a short `README.md` with instruction and command to fully reproduce your work. You must use Python.

You can present your work using python scripts (`.py`) or through a Jupyter Notebook (`.ipynb`).

Data exploration is entirely up to you and doesn't need to be presented to us.

For the second task, you can prepare some slides to present what's the definition of *best player* you followed.

Once the technical interview is over, we will clone your repository locally (so we will ask you to temporarily change the visibility to public) and run your code. Make sure it runs.

We ask you to keep the GitHub repository private whilst the hiring process is ongoing. Once it's over, you can make it public, fully accessible to anyone and freely communicate on it.

Once you feel like you are ready, send us an email at mmatteotti@parmacalcio1913.com to book a timeslot for your technical interview. At this stage, you don't need to send us anything - simply tell us you are ready, and we will arrange the date and time for the interview.
You will have until September 5th (included) to get back to us. There is plenty of time, so please make sure you give yourself enough time to acclimatise with the data and understand how it works.

Should you need more time for WHATEVER reason (holiday, exams, acclimatization with the new data, ...), don't hesitate to reach out to us. We'd rather give you one or two extra weeks, rather than losing a strong candidate because of time constraints.

Build something amazing, and have fun doing it!

P&A Department - Parma Calcio 1913