Pytch Proposal

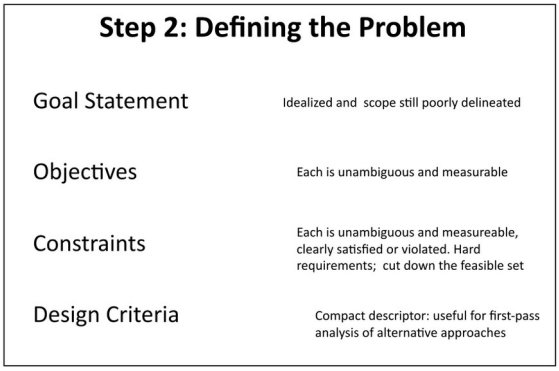
Project Proposal Review

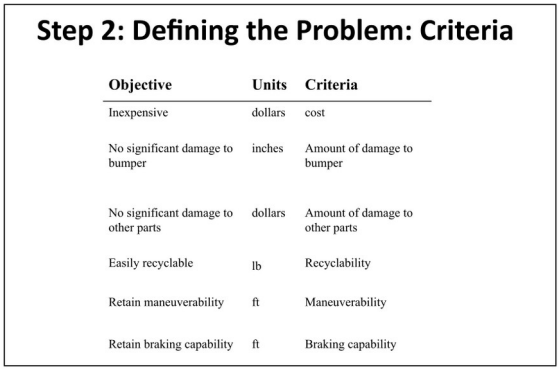
The following feedback to your group is synthesized from a close and careful reading of your proposal document and my reflections thereon (along with input from Kyle). From this critique, try to identify any weaknesses that should be addressed immediately and make a note of those in your proposal document. Thereafter, the group should ask if there are issues of a broader nature which need to be addressed in order to be reflected in your Critical Design Review Document (due on November 7th). Also, if you have not completed Peer Review I, please find the link on canvas and complete it as soon as possible.

Summary:

This proposal has no obvious omissions; it introduces the problem, gives some broad design criteria, shows why three existing solutions are inadequate, and proposes a platform. The first two-thirds of the proposal is well written (in some places, very well written); the later sections are much rougher, with elements which are more speculative and require further elaboration, justification, or both. There are also several minor inconsistencies which are sloppy and suggest that nobody did a complete front-to-back read of the submission to ensure terminology is defined before being used. Nobody took on the role of enforcing consistency (game *vs* match, stretch *vs* reach) or engaging in general quality control (e.g., references/bibliographic entries, figure captions, punctuation conventions).

In line with the previous statement about a lack of quality control, the proposal can be improved by doing a complete pass to identify and distinguish goal statements, objectives and constraints, and criteria—referring back to the slides below.





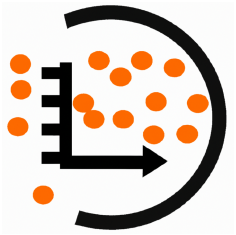
While re-reading, it is helpful to distinguish statements that are about the problem (and its setting and context) from solutions to the problem, and from *your* solution to the problem. The ideal flow will introduce and motivate the problem (your proposal does an exceptional job at this), various methods, applications, products, tools and hypothetical solutions to address the

problem. The axes under which they might be compared (pros and cons on page 18 doesn't compare each of them along identical dimensions, but is still quite effective nevertheless) would be objectives so long as they are quantifiable. The ideal document will make sure that the necessary details show up when they are relevant and no sooner. If the objectives can be distilled from the problem description that is ideal; but should it turn out that you have some envisioned feature which you plan to evaluate, but which has not been mentioned before, this is a cue to see whether it can be traced back to a missing criterion or aspect of the need/goal statement.

Section 5.3 on Societal Impacts does not appear to connect with much of the main body of the proposal. Re-examining what is written there and also considering what is necessary for a testing and validation—when all taken together—might help you flesh out a detailed evaluation plan. The evaluation plan forms a much more essential part of the CDR document.

Detailed comments, corrections and suggested modifications appear within the annotations of your PDF. Some of these are small comments (e.g., punctuation or typographical), others are much broader, and will likely require some reflection. You may want to bring some of these up to discuss at future meetings.

The overall proposal grade: A-



Pytch

Project Proposal

Gabriel Diaz
Ryan Kutz
Anthony Pasala
Joseph Quismorio
Ryan Son
Shurui Xu

Department of Computer Science and Engineering
Texas A&M University

9/26/23

Table of Contents

1. Executive summary

In today's digitized world, sports enthusiasts and professionals crave deeper insights into gameplay, and soccer, in recognizing its widespread global appeal and vast amount of recorded data, is primed for an advanced analytical platform. However, despite there being many existing organizations that provide such platforms, the world of soccer analytics has been historically dominated by costly, inaccessible proprietary solutions that remain beyond the budgetary scope of the average fan and are shrouded in opaqueness. Even still, as we see soccer continue to grow in popularity, there comes with it an increasing demand for a platform that democratizes access to advanced analytics, and harnesses the power of computer vision and machine learning. Our goal is to bridge this gap by providing an open-source, competitive alternative to the high-priced, complex offerings currently dominating the market. Our platform will enable users to upload or live-stream soccer match footage and receive a plethora of advanced statistics and visualizations about the match.

But why is this significant in a global context? Soccer's impact extends far beyond the 90-minute matches. It is a tool for diplomacy, community building, and education. Therefore, understanding the game at a deeper level has ramifications that are socially, economically, and politically relevant. Data analytics in soccer can play a pivotal role in talent scouting, especially in regions where talent remains untapped due to lack of resources. It could revolutionize coaching methodologies, changing the way grassroots programs train budding soccer stars. Furthermore, for fans, access to such advanced analytics would enrich their viewing experience, fostering a deeper appreciation and understanding of the strategies employed on the field.

While the primary motive is to empower fans and enthusiasts, the potential societal ripple effects are vast. A democratized, accessible analytical platform could become an educational tool, a medium for promoting fitness and strategic thinking, or even a driver for socio-economic change in regions where soccer is a beacon of hope for many.

Our product will, ideally, give users the ability to upload soccer match footage and receive intricate statistical breakdowns of the game. By leveraging state-of-the-art computer vision technologies and machine learning algorithms, we are capable of:

1. **Processing Match Footage:** Users can upload videos of soccer matches, with the system identifying critical segments suitable for analysis.
2. **2D Aerial Visualization:** A top-down perspective of the game, helping users appreciate the spatial dynamics.

3. **Advanced Statistical Insights:** Including but not limited to heat maps and replayable visualizations of key moments.
4. **Expandability & Modularity:** As the platform matures, there is potential for processing live streams and sharing visualizations. Users can also manipulate data contextually, adapting it to specific needs.

We have also identified three primary user categories, and sought to understand their use cases and needs:

1. **Sports Better:**
    a. Integration of contextual insights with raw statistics for enhanced decision-making.
    b. Comprehensive results showcasing both underdogs and top contenders.
    c. Real-time updates coupled with historical records to evaluate betting prospects.
2. **Data Analysts:**
    a. Access to raw data to construct bespoke models and graphs.
    b. User-friendly APIs to extract required information.
    c. Basic visual references aiding in data comprehension.
3. **Sports Analysts:**
    a. Context-specific data and visualizations providing a deep dive into plays and player positions.
    b. Access to a wide variety of games to discern patterns.
    c. Automated video sourcing to facilitate rapid information dissemination.

One of our platform's unique selling points is its open-source nature, which not only offers a competitive edge over pricier alternatives but also promotes inclusivity, ensuring both casual and hardcore soccer fans are able to involve themselves in a previously exclusive field. The technical backbone of our platform combines computer vision with machine learning to dissect uploaded videos, analyzing player trajectories, ball movements, and pivotal game moments, which are then translated into intuitive visual formats.

Beyond just filling a market void, our platform's societal implications are significant and expansive. It is poised to serve as a functional and all-encompassing educational tool for budding soccer players and coaches by facilitating a deeper game understanding through advanced statistics collection and analysis. The open-source nature of the project further imbues within it a collaborative aspect, inviting soccer fans, developers, and analysts into a cohesive and synergetic ecosystem, where platform refinement, iteration, and enhancement

become a community-driven endeavor.

In addition to the general outline, our project's financial aspects have been delineated in a display of transparency and proof of efficient allocation, and the roles and qualifications of our team have been taken into account in order to define their roles, guaranteeing an organized and efficacious project execution. In essence, as soccer's global footprint expands and the thirst for deeper game insights becomes more pronounced, our platform is designed to not just satisfy this need but elevate the entire realm of soccer analytics to new, uncharted territories.

2. Introduction

2.1 Problem background

One of the most pronounced shifts in soccer analysis in recent years is the ability to derive intricate data and insights from matches, and we have borne witness to an exponential growth in soccer analytics platforms. However, a deep dive on existing platforms reveals several inherent issues. Most of these platforms are implausibly expensive, thereby offering limited accessibility to a broader audience and ultimately barring a large and enthusiastic audience from engaging in a budding field. Furthermore, a number of these platforms are proprietary, locking users into specific ecosystems without the freedom to customize or adapt the analysis to unique needs. Another pressing concern is the conventional and manual approach to soccer analytics. Analysts often have to navigate through exhaustive hours of footage, a process that is not only time-consuming but also introduces human biases and errors. Such manual scrutiny limits both the depth of insights and the volume of matches that can be analyzed within a specific period.

Given soccer's global reach and unparalleled popularity, it is critical to recognize the potential lost opportunities due to these challenges. Clubs at all levels, from the top-tier leagues to grassroots, can benefit from the insights provided by analytics. Yet, the high costs associated with many platforms act as a significant barrier to entry. This not only hampers the smaller clubs with limited resources from harnessing the power of data but also denies independent analysts, journalists, and enthusiasts the chance to delve into the world of soccer analytics.

The proprietary nature of many platforms further compounds the issue. In most sectors, innovation thrives when there's collaboration, open-source sharing, and a fusion of diverse perspectives. By keeping platforms closed off and rigid, we might be stunting the growth of the field. The insights derived from one platform could be augmented and enhanced by another, but the current ecosystem doesn't always facilitate this level of integration.

Additionally, the persistence of manual review methods in modern analytics is perplexing. In an era where industries across the board are leveraging the power of automation and artificial intelligence, it's a significant oversight to rely so heavily on manual processes in soccer analytics. The nuances of a game as dynamic as soccer certainly demand a human touch, but the initial sifting and categorization of data can benefit from automation, ensuring analysts spend their time drawing meaningful insights rather than getting bogged down in the minutiae.

In light of these challenges, it becomes evident that while we have made significant strides in the realm of soccer analytics, there is still a long way to go from true democratization. Addressing these barriers is not just beneficial for the industry's growth, but essential to opening access and fostering innovation in soccer analysis.

2.2 Existing solutions

Several existing platforms aim to tackle the intricate challenges of soccer analysis by offering a range of analytical tools and visual insights. These platforms are designed to cater to diverse needs – from clubs seeking to optimize their tactical strategies, to scouts looking to unearth the next big talent, to fans keen on understanding game dynamics on a deeper level. Each solution, while addressing a specific niche, often comes with its own proprietary datasets and interfaces.

In the lineup of notable platforms, a few key players consistently emerge: **Opta**, renowned for its granular match data, **StatsBomb**, recognized for its innovative metrics and analysis, and **ESPN**, known for its broad audience reach and comprehensive sports coverage.

- **Opta:** A subsidiary of parent company Stats Perform, Opta was created in 1996 to begin in-depth data collection within the English Premier League. Since then, it has become the primary data source for the English Premier League, Serie A, La Liga, and other major European leagues.

  Opta provides a variety of services to suit varying needs, from live scores, lineups, and tables to 3D recreations of matches. Its most sophisticated product, Opta Vision, tracks the locations of each player and the ball throughout the match and can present this in a 2D or 3D format to the viewer. This data is then used to create more detailed insights that require knowledge of player and ball positions. Users can view insights such as line-breaking passes and xThreat (Expected Threat), both of which use player positions to determine how effective a pass or dribble was at opening a gap in the opposition's defense. Users can also view team formations and how they changed throughout the match, as well as how effective each formation was in the match. (4)

  From the demonstrations shown, Opta's data is extremely accurate and comprehensive. The use of multiple camera angles and human intervention to correct inaccuracies from the automatic data collection systems gives their insights a degree of credibility and accuracy.

  Opta's prices are not made publicly available, and are negotiated depending on the

needs of the client. One source stated that prices vary between 500-2000 GBP/month. (5) We have requested a demonstration and pricing information with Opta for its Opta Vision service, but no response has been received at the time of writing.

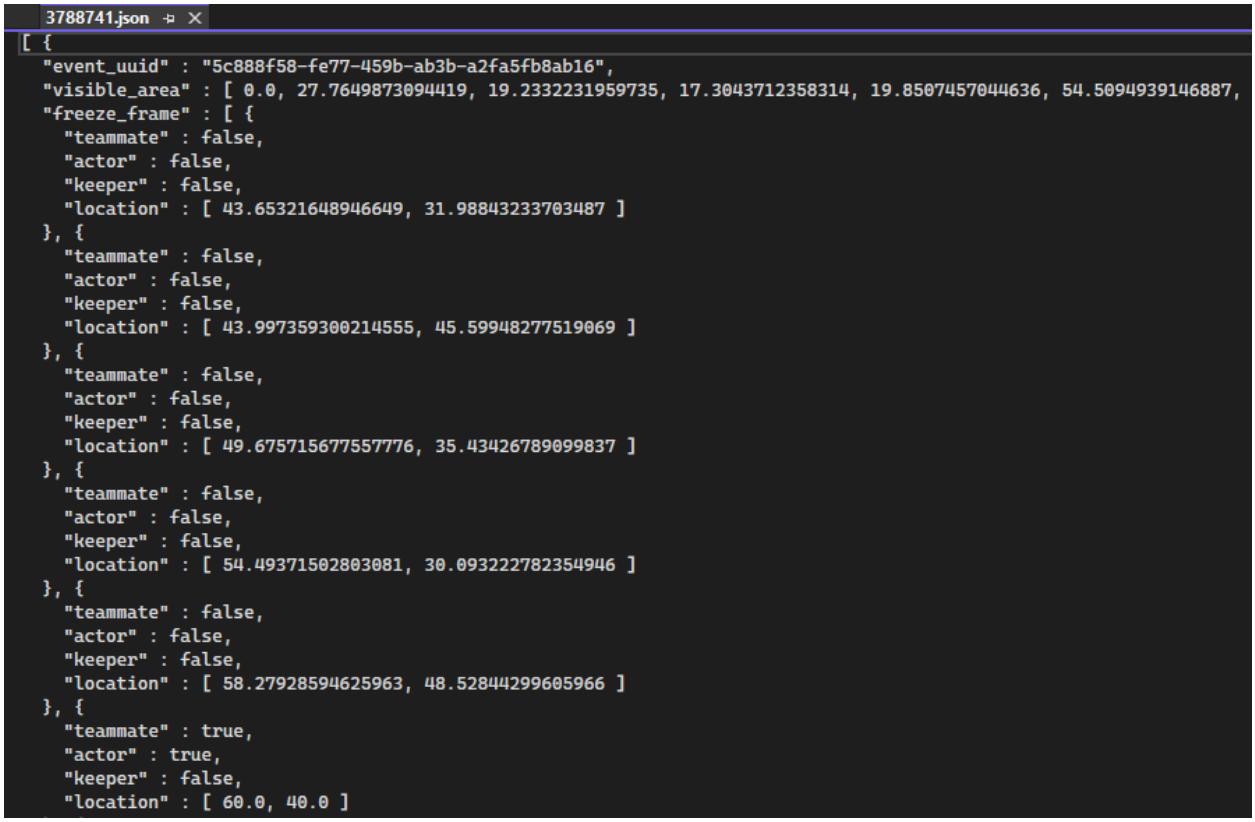
- **Statsbomb:** Founded in 2017, Statsbomb has quickly become another leading data provider in soccer. In 2022, Statsbomb collected data for matches in 90 professional leagues across the world. Unique features the company offers include a live xG (Expected Goals) model that uses additional factors to calculate goal probability, such as defender and goalkeeper positioning, as well as shot impact height.

  Statsbomb's data is collected in the same manner as Opta's, using multiple camera angles and computer vision. The data is then evaluated by a human QA team to ensure its accuracy.

  Statsbomb's data is intended to be analyzed using their IQ platform, which allows users to create bespoke visualizations for team and player-related statistics. Users can compare player performances in each metric across teams and leagues, making IQ a useful data-driven scouting tool. IQ can also create team-related visualizations such as pass maps, shot maps and pressing maps.

  The IQ platform's 360 feature provides a 3D simulation of the match created using player and ball positional data. Users can view moments of their choice from various angles, and can overlay graphics to give further insight into formations and other aspects of the game. (1,2,3)

  In contrast to Opta, Statsbomb provides samples of its detailed data from selected competitions for free to the public. However, most of the free data is from 2021 or earlier. Below is a sample of the 360 positional data taken from a 2020/21 La Liga match between Barcelona and Deportivo Alaves:



Example of Statsbomb 360 data, from one of its free data samples (7)

In each freeze frame, the area visible to the camera is recorded, and the location of each player, as well as their relation to the ball carrier (the “actor”) is noted.

Similarly to Opta, Statsbomb’s pricing is not made publicly available. We have reached out to Statsbomb for further information, but no response has been received at the time of writing. For the purposes of this project, we assume that the company’s services are priced competitively with Opta’s, and are therefore prohibitive for an individual consumer.

- **ESPN**: ESPN is America’s largest sports news and broadcasting network. ESPN’s coverage mainly focuses on domestic leagues such as the NFL, NBA, and MLS , but the company also covers most of Europe’s major soccer competitions. ESPN provides scheduling, news, statistics, and live updates on matches in these competitions.

  In comparison to dedicated data collectors such as Opta and Statsbomb, ESPN serves as an accessible source of soccer information for the average fan. Each match page provides general insights and statistics, such as the score, starting lineups and substitutions, possession percentages, shots taken, and yellow/red cards.

Although a publicly available API for developers to access ESPN's data was created in 2012, it has since been closed. Data on ESPN's site is still available for free, but it would need to be scraped.

2.3 Needs statement

Considering the current landscape, it is evident that there exists a tangible demand for an inclusive, comprehensive, and user-centric solution in soccer analytics. The desired platform should look to address the following issues:

- **Accessibility and Affordability:** One of the glaring realities of the current soccer analytics environment is its growing exclusivity. This exclusivity is not just in terms of the platforms themselves but also in the costs associated with them. Many enthusiastic fans, grassroots coaches, and amateur analysts often find the gateway to detailed soccer analytics barred by prohibitive costs. Our platform should not only focus on offering such an audience the basic tools for analysis; it should look to make its mission democratizing data in a sport that is celebrated across diverse backgrounds around the world. A platform that stands out in this landscape would be one that offers its groundbreaking features at an accessible point, making sure that regardless of one's financial capabilities, the world of soccer analytics is never out of reach.
- **Depth, Breadth, and Precision:** In sports, especially in a game as dynamic as soccer, public insights are seldom sufficient enough in their depth to satisfy the needs and wants of enthusiasts and professionals alike. The sheer pace and complexity of a soccer match, with its many possible strategies, plays, and individual moments of brilliance, call for deep, intricate analysis. In addition, the sport is played across continents, spanning a number of local and national leagues. A platform needs to be able handle a vast array of games, not just in terms of quantity but the quality of insights derived. Recognizing underlying patterns, identifying emerging trends, and drilling down to micro-moments in a game can change the way fans and professionals engage with the sport. We also seek to address the issue of depth by providing users with the ability to generate statistics for their own inputs, thereby granting them full customized control over the content of their insights.

- **Customization and Flexibility:** Soccer aficionados are a tapestry of diverse demands and interests. Catering to this wide spectrum of interests demands a platform that is not entirely rigid but one that offers significant customization; in this realization, we recognize that the platform must be adaptable. Additionally, as the world of soccer analytics evolves, the platform should be agile, ready to incorporate new analytical paradigms, and always be at the frontier of innovation.

2.4 Goals and objectives

Our goal is to develop a platform that opens up sought after statistics in soccer that are typically hidden behind costly fees. This will be accomplished by utilizing open-source machine learning and computer vision libraries to generate relevant soccer data in the spirit of providing users reliable insights at no cost. Pytch exists to offer data as accurate as its competitors, as well as meet the middleground in regards to what is not provided by other sites.

With a platform model, all users should be provided with the same information in a way that allows them to draw relevant conclusions based on their needs. Users should be able to view soccer statistics similar to their favorite aggregator, but with the ability to deep-dive into contextual data. Uploaded videos should also produce results no different from a professional match, leveling the analytics regardless of the tier at which the game is played. A simple and sleek user interface should also safeguard users from being dually confused alongside data.

Finally, it is critical to the project that the computer vision portion is able to generate meaningful data from soccer footage. As the crux of the platform, successfully doing so will thus allow for standardizing data output and allowing transformation into other meaningful statistics, as well as visualizations and heatmaps. Users should be able to trust that in spite of the inherent variability of image detection, accuracy is still an important benchmark taken account for in development.

2.5 Minimum Viable Product (MVP)

For the course of the semester we have planned to break our feature sets into 3 main categories: features for the Critical Design Review, features for the end-of-course, and additional reach goals.

For the Critical Design Review phase, our primary focus will be to establish the core functionality of our soccer analytics platform. At this point of the project the main types of clips that we will focus to support are

- Allow users to upload match clips of open field play

- Generate detailed timeline data of the match, creating timestamps that identify player and ball positions on the field
- Expose this data through our publicly available API
- Create our first set of visualizations: Passing Maps, Heat Maps of Players

Moving towards the End of Course delivery, our focus shifts to advancing the capabilities of our platform. The main goal of this phase of the project will be to support popular matches, not only by providing our own visualizations but also integrate and aggregate data from existing APIs where our computer vision falls short. By the end of this phase we will have 2 types of matches: user submitted matches which we will provide visualizations for and our popular matches which will have our visualizations and additional aggregated data.

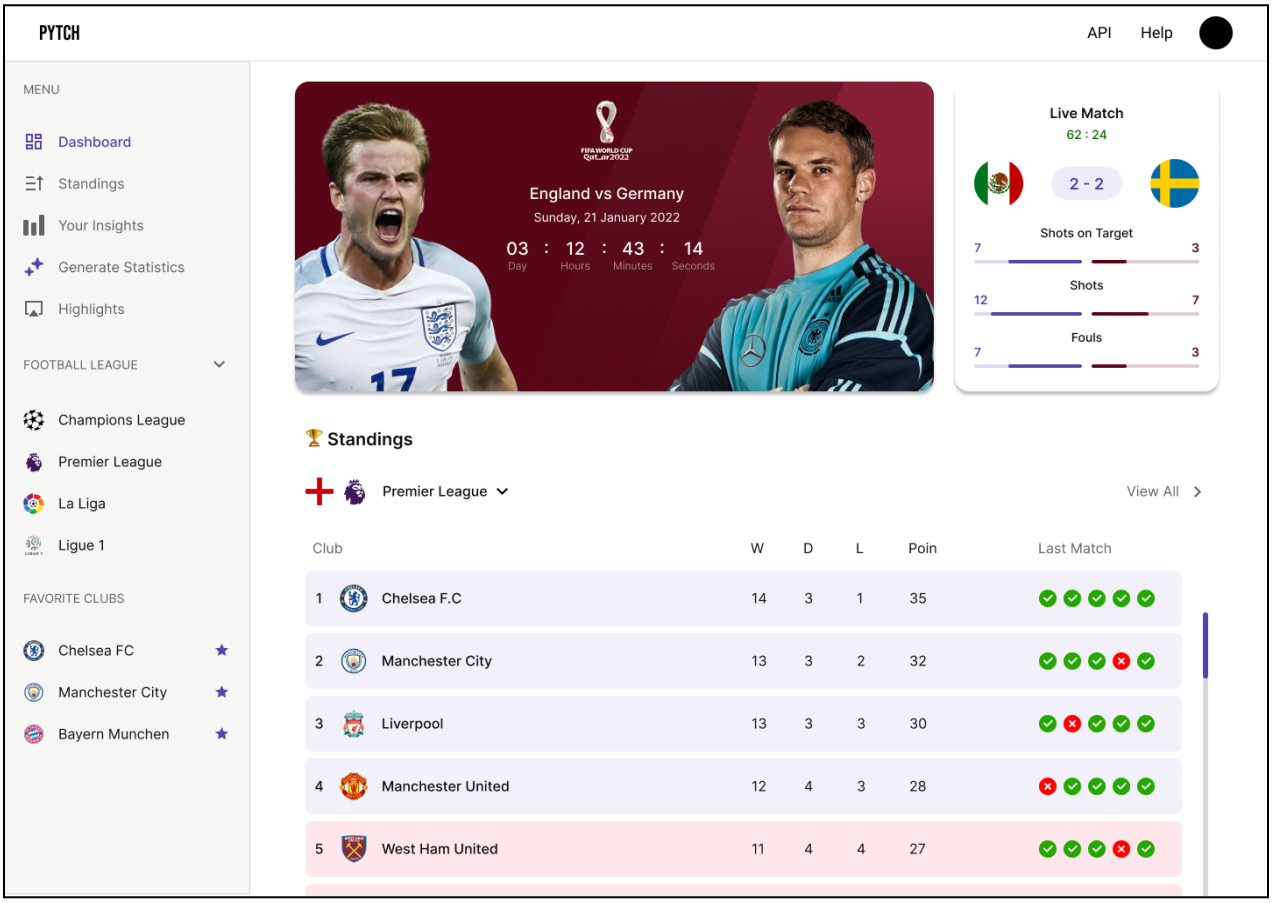
- Continuously improve our analytics and provide additional analytics: Shot maps
- Process live streams of popular matches from the following leagues:
    - Premier League
- Integrate with existing free APIs to provide additional match data
    - Match Information: Teams, Players, Score, Time

As we venture into the Reach Goals phase, our aim is to elevate the user experience by creating a marketplace of visualizations. By offering a diverse array of visual representations derived from our intricate data, we not only cater to the varying preferences of our user base but also foster a sense of community collaboration.

- Allow users to upload their python code that generate matplotlib figures
- Allow users to view other visualizations and utilize them accordingly

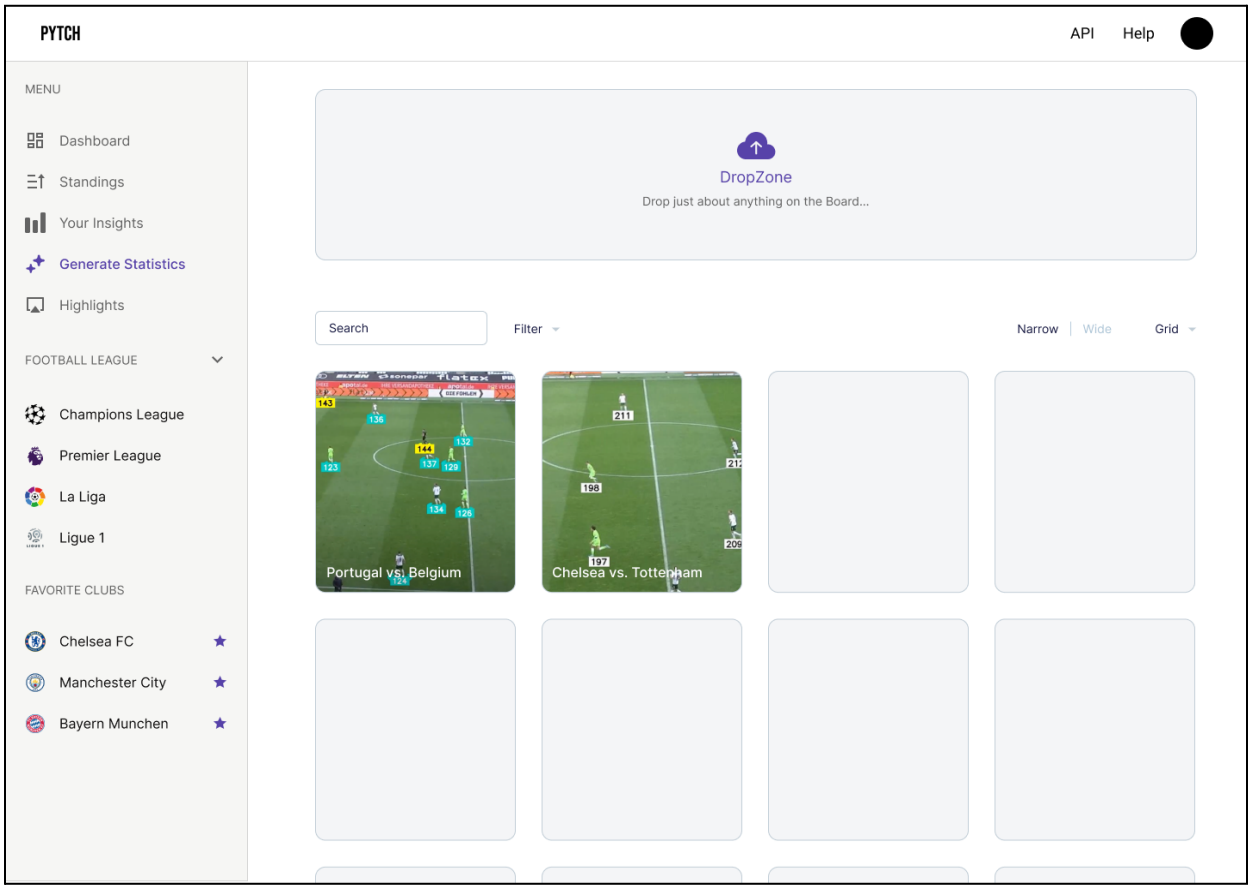
2.5.1 Wireframes

Wireframes for the user interface were created as an ideal design to work towards in our 2-month timeline. Picture below is a wireframe of what a user may expect when accessing the landing page upon login, with an updated feed of current soccer games.



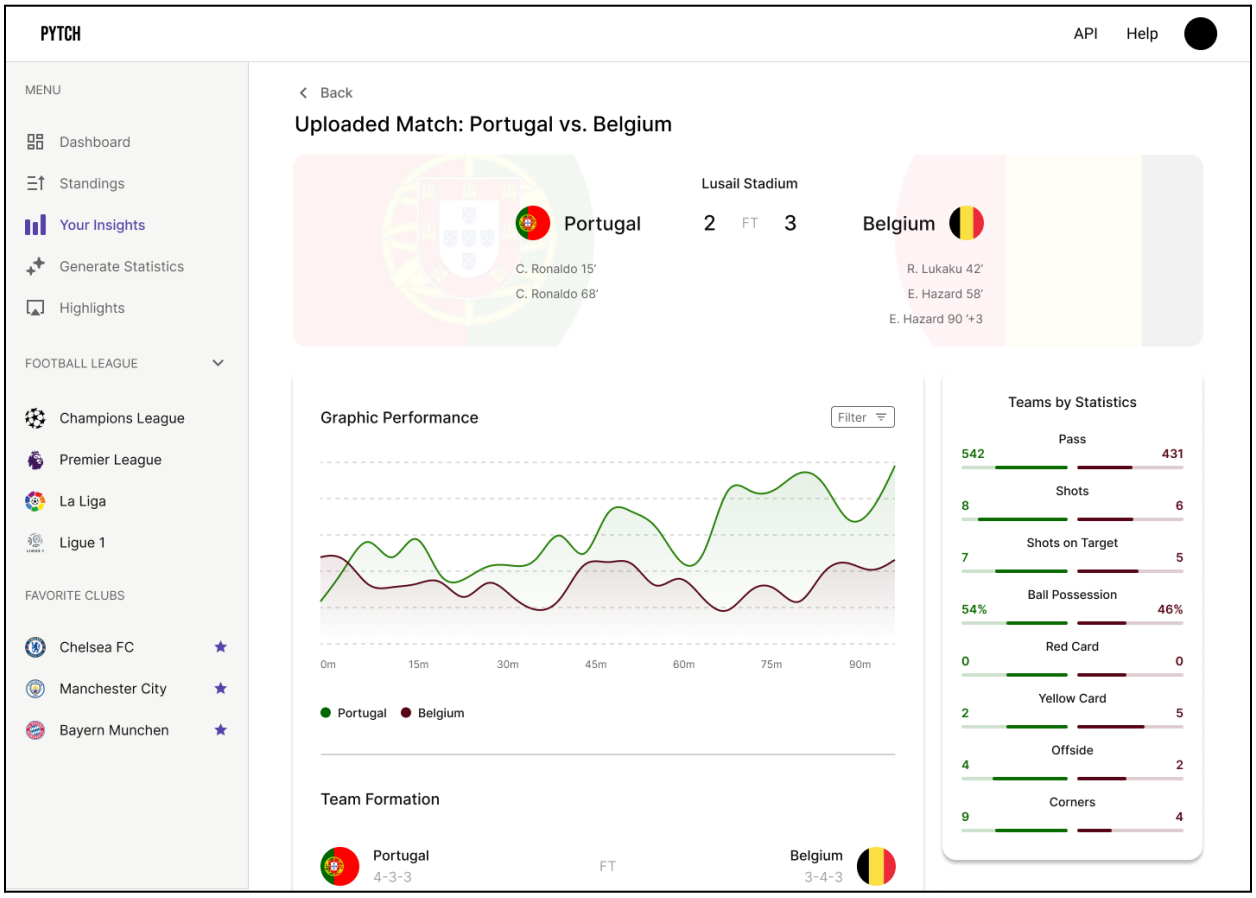
Pytch landing page

Pictured below is an example of a video upload section, where users can upload different soccer clips for processing.



Pytch match video upload page

Finally, pictured below is the statistics page, where users are expected to spend most of their time after uploading their clips for analysis. Key insights and visualizations are the sole focus of this page.



Pytch statistics page

2.5.3 Design constraints and feasibility

Open Field Play

Due to our product ~~trying to~~ only utilize the livestream provided by users/sports networks we do have to acknowledge that our capabilities and accuracy will be lower and not as in depth as the data provided by Opta and StatsBomb. Within this in mind our product will be optimized to exclusively focus on providing in-depth tactical analysis from footage that comes from the main match camera, where most of the field is unobstructed and we don't have to worry about changing camera angles that would occur during the broadcast. The analytics all outlined in our MVP were designed with this limitation in mind.

Main Camera Footage

Compute Limitations

Although the machine learning algorithms we are using are light-weight alternatives, this project will still be computationally expensive. Our current concern is storing match footages, as we plan to allow users to be able to upload their own footages. This made us decide to not store any of the users uploaded videos and instead just store timeline data that tracking software will recreate. Additionally, we will limit the size of file uploads to 500Mb, as there would be a lot of overhead to process these videos.

2.5.2 Measures of Success

Focusing on a lower accuracy threshold allows us to manage expectations effectively. Computer vision, while powerful, may encounter challenges such as occlusions, varying lighting conditions, and the fast-paced nature of soccer matches. By setting a realistic and achievable accuracy goal, we prioritize delivering consistent and reliable results within the scope of our chosen technology.

Moreover, a more attainable accuracy target allows for a smoother user experience. Users are likely to value reliability over marginal gains in precision, especially in real-time scenarios. Setting a lower but reliable accuracy goal ensures that the insights provided by our platform are consistent and trustworthy, fostering user confidence and satisfaction.

To test the accuracy of our system we will be testing against the Bundesliga Data Shootout data set, which contains uninterrupted main camera footage clips from the 2022 Bundesliga season. This dataset has a wide variety of teams, players, and stadiums providing +200 minutes of game footage. Our benchmark of success will be to correctly localize players on the field >50%, where the error of each player can be within 10% of the field size.

3. User Research

3.1 Competitor Analysis

Provider	Pros	Cons
Opta	• Highly accurate, granular data • Customizable products to fit consumer needs • Predictive models to forecast outcomes	• Prohibitive pricing • Not marketed to individuals • Complex data gathering process
Statsbomb	• Highly accurate, granular data • Platform for user-friendly visualization • Publicly available sample data • API access for clients	• Prohibitive pricing • Not marketed to individuals • Complex data gathering process • Difficult to use data outside of platform
ESPN	• Free access to statistics and match/league information	• Too surface-level for meaningful insights • No public API for developer access
Pytch	• Free and open source • Works with live and pre-recorded videos • Free API access for developers • Low-level, detailed data • Single camera angle, no human interaction required	• Less accuracy due to constraints 

3.2 User Stories

Pytch serves to benefit both the average user and the super-fan. However, to avoid orienting itself as a copy-paste of familiar aggregators, user stories help delineate potential needs from differing points of view. To fully embody the idea of a statistics platform, three key users are identified and help envision the project at varying scales:

Sports Better:

As a better, their motivations have more investment than the average fan. Thus, they are willing to invest more time and money to gather the necessary insights that give them the edge in their bets. Thus, a better will want:

- Contextual data alongside raw statistics are necessary to make the most-informed choice/parlay with respect to a given position or player
- Volume of results is important in order to have choices on the losing hand as well as the favorites to win
- Up-to-date information and historical data to hone in on a player's viability in one's bet

Sports Analyst:

For a sports analyst, contextual data is crucial when potentially scouting players, anticipating attrition within one's roster, or finding the best strategies for each player on the pitch. From this viewpoint, an analyst expects:

- Contextual data and visualizations that better inform about specific plays and positions during a match
- Raw numbers to have a baseline understanding of a player's natural talent
- Variety and volume of games to highlight patterns and trends over time as well as potential game plans versus certain teams
- Avoiding review and finding video sources manually as to quickly communicate relevant decisions

Data Scientist:

The data scientist point of view is the most generalized, as they may come without an understanding of soccer. Thus, their focus is to gather their own insights from data. As the most technical of users, data scientists ideally:

- Wants raw data to create their own models, graphs, and relationships between statistics
- Desire friendly APIs to transform data into more relevant mediums as needed
- Wants basic visualizations to use as reference when creating their own versions and understandings

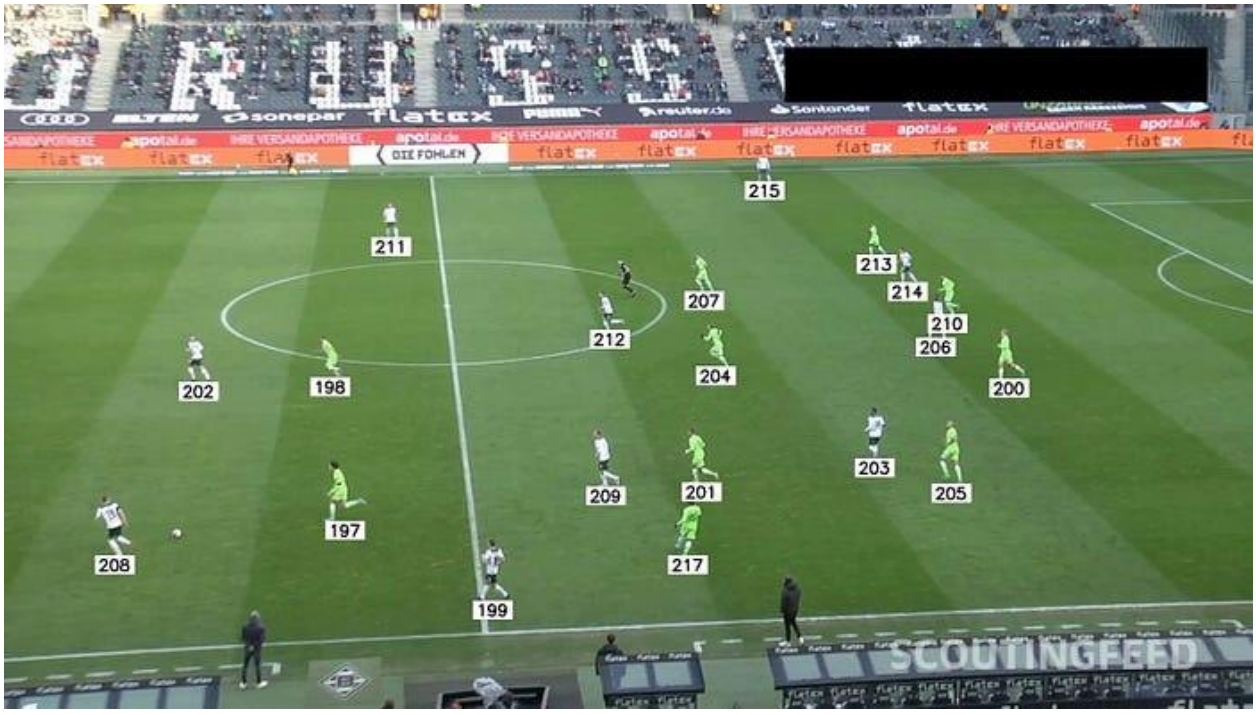
With all three viewpoints in mind, the most basic of users should at least expect to see raw gameplay data as well as some form of visualization as a guide. However, narrowing down into discrete users will aid in who data is modeled for and how development is prioritized.

4. Design Specifications

4.1 Software (CV/ML) Specifications

In order to be able to generate the timeline data that we seek we have identified 3 main challenges: Identification, Tracking, and Localization.

The first step of the problem is being able to effectively identify the players, ball, and referee on the pitch. After doing some investigating into the space, we decided to use YOLOv8, You Only Look Once version 8, an object detection algorithm that identifies and classifies multiple objects within images in real-time by passing the image in a single pass to a neural network. For the YOLOv8 algorithm we need to supplement it with a model to be able to identify the players, ball, and referees, there is the popular existing Common Objects in Context (COCO) which is able to identify players and balls, however with how small the ball is in frames from the main camera angle the model really struggled to identify the ball in the field. This led us to find an existing pre-trained model.



Identifying Players using YOLOv5

Now that we are able to identify players in frame we need to be able to effectively track them across frames. There are many different algorithms to accomplish this Multi-Object Tracking (MOT) problem such as SORT, DeepSORT, and Norfair which all focus on using complex Kalman

Filters alongside Hungarian Matching to be able to accurately track object, however we've decided to use BYTETrack. ByteTrack's goal is to simplify the tracking process by keeping non-background low confidence detection boxed for a secondary association step which are typically discarded by other algorithms. We believe this key feature of ByteTrack will work well with tracking soccer matches by making up for frames where there is a high amount of occlusion.

The last part of this problem is to be able to localize our data that we have collected to coordinates on the field. To do this, we will need to create a homography matrix that maps points from the camera image to the field. A homography matrix can be calculated from 4 points on 2 different field lines in the image. This can be easily accomplished by annotating points in a frame by hand, but we will need a process to find these points automatically. Using the procedure explained in *"Robust and Efficient Homography Estimation …"* cited below, the steps are as follows:

1. Mask out areas of the image outside of the field (areas that are not predominantly green).
2. Use the Hough transform to detect line segments in the image.
3. Classify the detected segments as either vertical or horizontal.
4. Extend the segments into rays and find intersection points between them.
5. Match the intersected points to points on a top-down model of the field by determining whether other points exist in each direction.

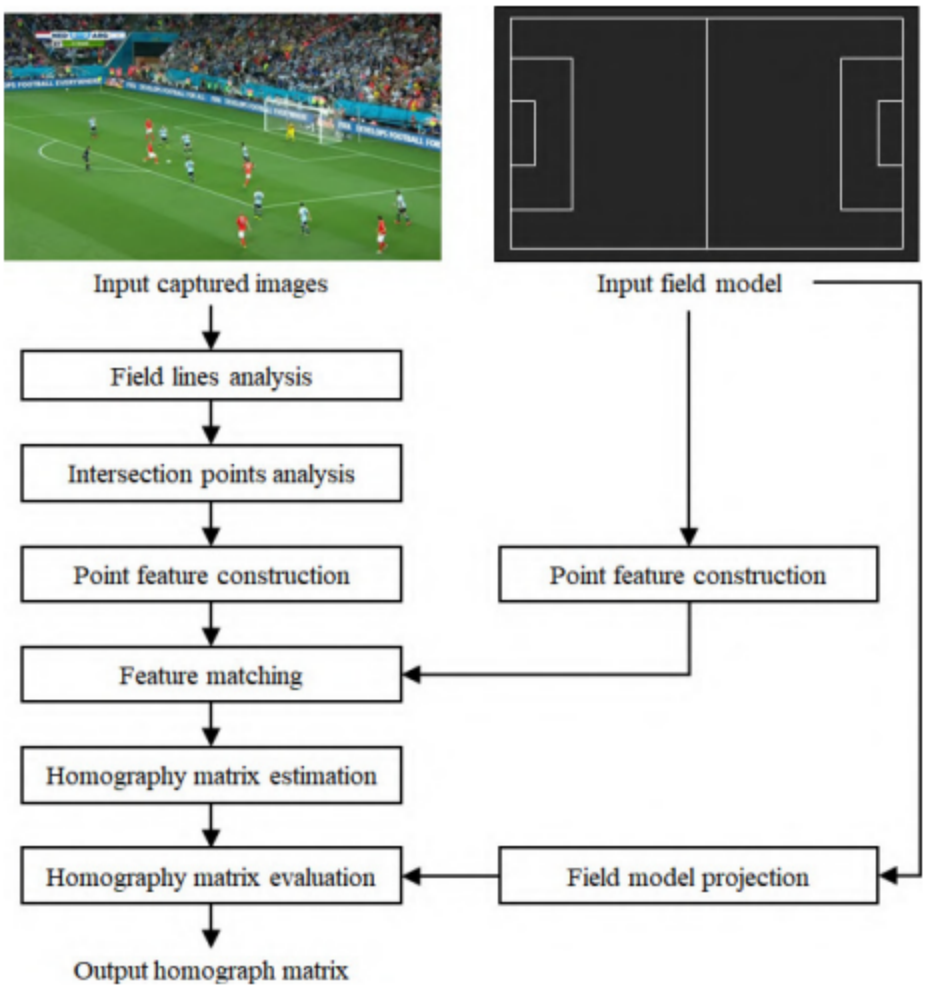


Diagram of the process (6, Fig. 1)

4.2 Frontend Specifications

The front end of the proposed application serves as the primary bridge between users and the powerful analytics tools powered by the application's back end. It is important, therefore, that this interface is not only functional, but also intuitive, responsive and aesthetically pleasing. To achieve this, we have taken into account specific directives and screened select technologies that align with industry standards and are well-known for their efficacy in this domain.

- **User-friendly interface:** We aim to design an interface that caters to both novice and advanced users, propelled by insights gained from comprehensive user studies. The layout, navigation, and interactivity will be made with the end user in mind, ensuring that all users are able to easily navigate through the platform and access its features without facing difficulties or complexities.

- **Clear visual representations:** The essence of the application lies in presenting intricate soccer analytics in an easily digestible format. We will prioritize the use of modern design principles to ensure that visual data representations, whether they be heat maps or graphic visualizations, are both clear and informative. The choice of color palettes and the clarity of data visualizations will be central to the application's design ethos.
- **Fast load times:** Given the data-intensive nature of the application, we understand that there is importance in optimizing the application's speed. Users will expect quick load times, especially when dealing with video content and more complex visualizations. The chosen technologies have been screened to ensure that they support the creation of a platform optimized for performance.

In screening our potential tech stack, we have found the following frameworks and libraries to be particularly applicable to the creation of our application:

- **Next.js:** Next.js is a leading React framework renowned for its performance optimization through server-side rendering. Given the real-time demands of our application, leveraging the features of Next.js can enhance the performance of page loads and an overall improved user experience.
- **TypeScript:** To ensure that our application is robust and reliable, TypeScript, a statically-typed superset of JavaScript, has been considered. It offers clear typing, which can prevent potential runtime errors and enhance the stability of our platform.
- **Tailwind:** Tailwind is a CSS library that offers a utility-first approach to allow for the fast and efficient development of user interfaces. Its modular nature will ensure a streamlined and consistent design system, which will be instrumental in achieving the clean look and feel that we aim for.
- **Material UI:** Material UI is rooted in Google's Material Design principles, and offers a wide array of React components that are both modular and widely applicable. These pre-styled components can expedite the design and development process and ensure that user interface elements are consistent, functional, and aesthetically pleasing,
- **Dropzone.js:** Given that users will be able to upload soccer match footage, Dropzone.js becomes an important consideration in allowing users to interact with and upload their files to the platform. It offers a drag and drop interface which can be easily integrated into the platform, making the upload process intuitive and user friendly.

The front end of the application will be the confluence of modern design principles and high-performance technologies. By meticulously screening and selecting the appropriate tools and frameworks, we aim to offer users a seamless and enjoyable experience that is comparable

to those of other platforms, enabling them to delve deeper into soccer analytics with ease and precision.

4.3 Backend Specifications

The back end bridges the capabilities of the computer vision model / analytics tools and the user interface. It needs to enable the flow of user-uploaded match videos to our ML model and analytics, and serve back processed data in the form of visualizations to the front end website, or as standardized JSON data for more technical users through the REST API, documented through the OpenAPI specification.

To ensure secure API routes and managing access and manipulation of user data, we will implement user authentication and authorization through existing OAuth solutions and battle-tested JSON Web Token auth flows. This allows for secure transfer of video data to our analytics tools and the results back to the user's interface.

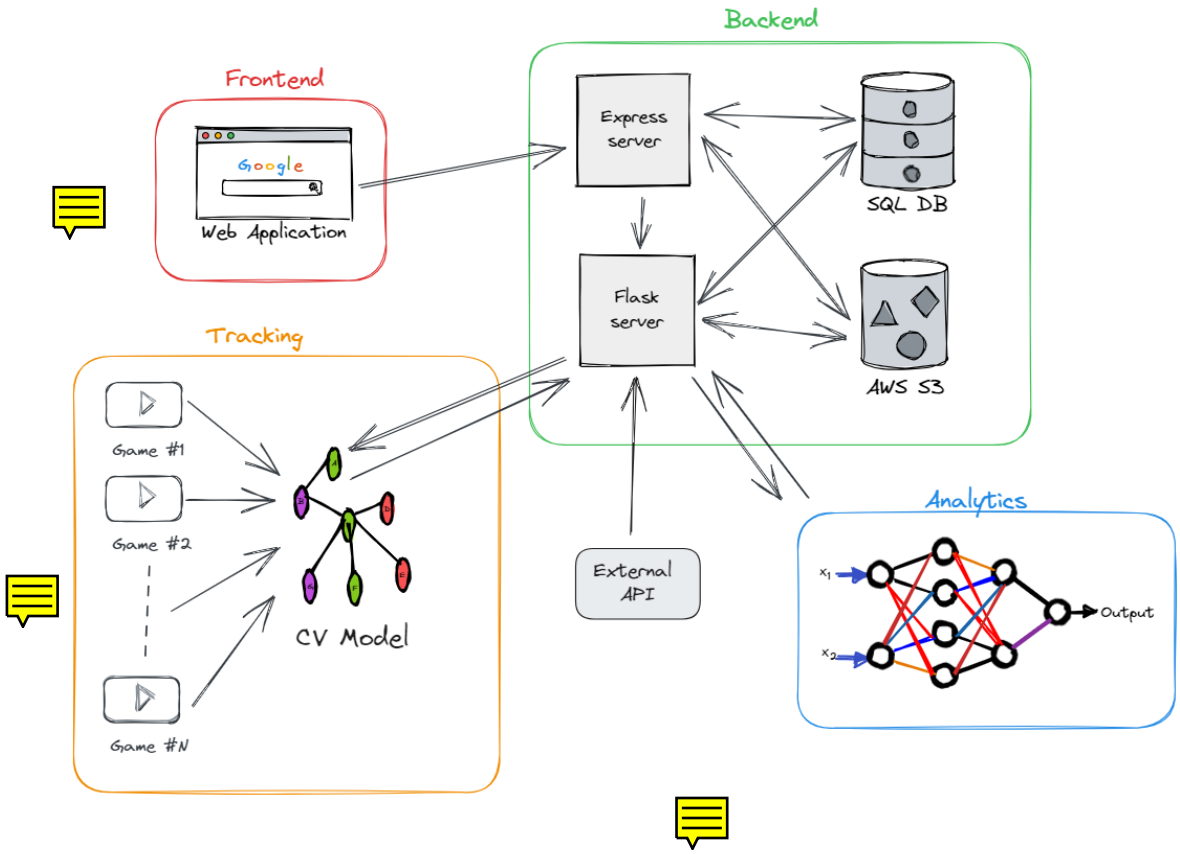
Through the backend, matches uploaded by users through the website interface will trigger an internal flow of API routes that implement business logic to indicate the ML model, analytics, and visualizations to start computing. Those results will then be sent through to our API to upload to the database, which the frontend interface will poll to update for the user.

Aside from user-uploaded matches, we will use external API such as API-Football as our source of data for upcoming / live matches, providing their analytics by default on the user interface. It offers RESTful API and is compatible with RapidAPI as well, which provides an intuitive, consistent method of communication with the various API they host and support. In our Express / Flask backend, we'll query this API using the Axios and 'requests' libraries triggered by a first load of the UI.

We screened libraries that complement the chosen technologies for frontend and analytics tools:

- **Express:** An easy to use REST API framework in Node.js to expose public and private routes to transfer user and analytics data between the frontend interface and our analytics tools. Responsible for managing flow of user and video data and updating our SQL database + S3.
- **Flask:** REST API framework serving as the primary form of communication between API routes defined through Express and the ML model / analytics tools. Responsible for receiving calls to trigger analytics and storing results to our SQL database + S3.
- **AWS S3:** Cloud-based storage solution for uploading and accessing the resulting visualizations and processed match data.
- **Prisma ORM:** Easy to implement TypeScript ORM for defining our SQL database schema, ensuring typesafe access and updates to the tables in our Express app.
- **MySQL:** Preferred SQL language due to first-class support through the `mysql-connector-python` Python library, easy to use in our Flask API business logic.

Role of backend tech in application architecture:

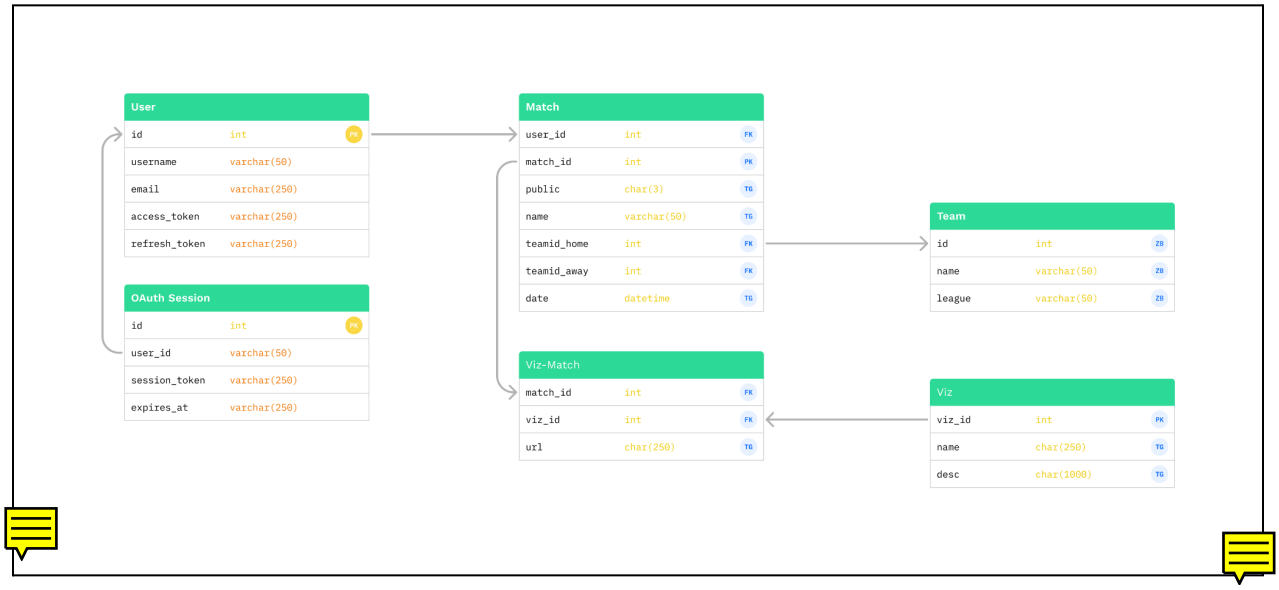


We decided to use AWS S3 for storing large JSON blobs and resulting visualizations per user-uploaded match to avoid the overhead of computing JOINs between tables in our SQL database. Here's a preliminary example of a single unit of processed data from a user-uploaded match video:

```
{
  Timestamp: int,
  Players:[]{
    playerNo: int
    team: string
    X: int
    Y: int
  },
  ball: {
    X: int
    Y: int
  }
}
```

Our SQL database supports user authorization and authentication, along with storing metadata per match and visualization, including basic information like teams participating in the match and URLs

to S3 buckets to query for processed data and visualizations. Here's the preliminary schema for our database:



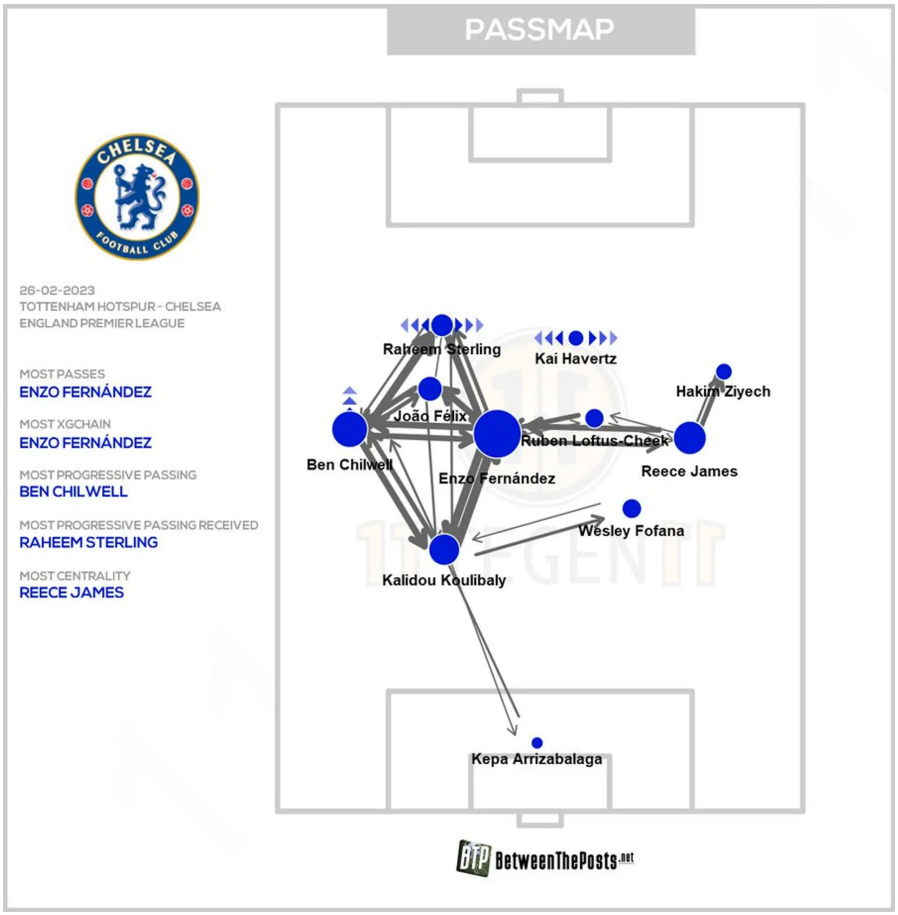
Entity Relationship Diagram (ERD)

4.4 Analytics and Visualizations

The analytics and visualizations module of the application handles the conversion of the raw player/ball tracking data into meaningful match-related statistics, as well as the creation of visualizations based on those statistics for the user to view. As a stretch goal, this module will support the display of user-made visualizations. The main visualizations that will be developed first are pass maps and heat maps.

Pass Maps

A pass map is a graphical representation of a team's passing structure. Each player is represented by a dot on the field. The size of the dot depends on how many passes the player made in the match. The position of the dot is the average location of the player during the match, and an arrow connecting two players represents a pass between them, with larger arrows indicating more passes.



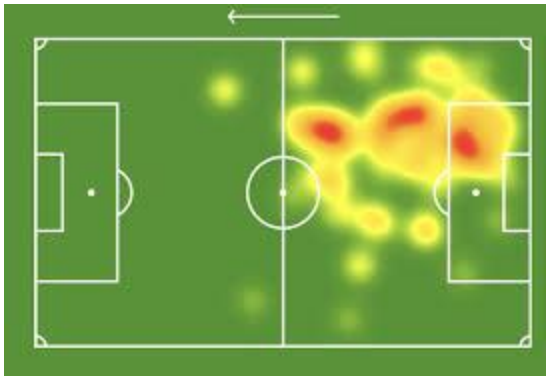
Example of a pass map

Creating a pass map requires knowledge of the following:

1. Identification of a pass. Since the CV/ML module identifies the player in possession of the ball, we define a pass as a change in possession from one teammate to another.
2. Location of the passer. Calculated as the location of the passer when the ball leaves their possession.
3. Distinction between individual players on the team. Identified by the CV/ML module.

Heat Maps

A heat map is a graphical representation of a player's location when they receive or win the ball. In contrast to pass maps, heat maps only track one player. Darker colors represent more touches made in that area.



Example of a heat map

Creating a heat map requires knowledge of the following:

1. Location of the player in question when they receive/intercept a pass or dispossess an opponent. We define this as the player's location when the CV/ML module marks them as the player in possession of the ball.
2. Distinction between the player in question and the rest of the players on the pitch. Identified by the CV/ML module.

In addition to the statistics necessary for the visualizations above, the analytics module will track more statistics including:

- Progressive Passes: Passes that cover a significant distance
- Tackles: Change in possession between teams where the previous ball carrier and new carrier are in close proximity
- Interceptions: Same as tackles, but when the ball has traveled a significant distance to the new carrier
- Touches in Area: Classified between major areas of the pitch (opposition penalty box, final third, etc.)

Implementation

This module will run on the Flask side of the application, and will use pandas to parse the data from the CV/ML module. We plan to use Chart.js to create our "first-class" visualizations with added styling and interactivity, such as our pass maps and heat maps. Given the ubiquitousness of Python among data scientists, user-uploaded visualizations will be accepted in the form of matplotlib figures and linked with our match data to be displayed on the frontend using MPLD3, a library that makes matplotlib figures viewable in an HTML document. Visualizations will be stored in the MySQL database to reduce computational cost and to allow persistence of user-created visualizations.

5. Engineering Standards

5.1 Project Management

Our team has collectively agreed to the outline of the project with respect to the goals, implementation, and responsibilities. While computer vision is the core of the project and is expected to be worked on collaboratively, in later stages of the project we have split up roles taking consideration of experience and interests.

Team Background and Roles

Member Name	Background	Role
Gabriel Diaz	Full-stack, systems programming	Computer vision and analytics
Ryan Kutz	Backend, machine learning	Analytics
Anthony Pasala	Full-stack, machine learning	Backend, computer vision
Joseph Quismorio	Full-stack, web development, machine learning	Frontend lead
Ryan Son	Full-stack	Frontend, backend, PM
Shurui Xu	Frontend, machine learning	Frontend, computer vision

Currently, we are using Google Drive as storage for minutes, written documents, and general storage of resources. GitHub will be used for version control.

Weekly check-ins will occur on Mondays and Wednesdays from 7pm - 8:30pm. We are utilizing a group Discord channel to communicate with one another for extra help and as a general meeting space.

5.2 Schedule of Tasks

TASK TITLE	TASK OWNER	TASK COMPLETED	WEEK 1 9/3/2023	WEEK 2 9/10/2023	WEEK 3 9/17/2023	Week 4 9/24/2023	WEEK 5 10/1/2023	WEEK 6 10/8/2023	WEEK 7 10/15/2023	WEEK 8 10/22/2023	Week 9 10/29/2023	WEEK 10 11/5/2023	WEEK 11 11/12/2023	WEEK 12 11/19/2023	WEEK 13 11/26/2023
Project Conception and Research															
Scope & Goal Setting	Gabriel Diaz	100%													
Budget	Joseph Quismorio	100%													
Communication Plan	Shurui Xu	100%													
MVP Definitions	Ryan Kutz	100%													
Available Tools & Resources	Anthony Pasala	100%													
User Stories	Ryan Son	100%													
Project Proposal	Everyone	100%													
Computer Vision Tracking & Analysis															
Tracking Feasability	Gabriel Diaz	100%													
Soccer Fields Variability Testing	Shurui Xu	0%													
Localization	Gabriel Diaz	0%													
PytchTracker	Anthony Pasala, Gabriel Diaz	0%													
PytchAnalyze	Ryan Kutz	0%													
Genearate Data Visualization	Ryan Son	0%													
Frontend and Visualizations															
Wireframe and Mockups	Joseph Quismorio	100%													
Setup Frontend Structure	Joseph Quismorio	0%													
Implement Data Storage and Retrieval	Shurui Xu	0%													
Modernize UI	Joseph Quismorio, Shurui Xu	0%													
Presenting the Data Visualization	Ryan Son	0%													
Backend and Database															
Design Database Structure	Ryan Son	0%													
Setup Server	Anthony Pasala	0%													
Setup Database	Ryan Son	0%													
Implement Data Storage and Retrieval	Shurui Xu	0%													
Connecting PytchAnalyze and PytchTracker	Anthony Pasala	0%													
Project Evaluation & Final Product															
MVP Evaluation	Everyone	0%													
Performance Evaluation	Everyone	0%													
CDR and Presentation	Everyone	0%													
Final Report	Everyone	0%													

5.3 Societal Impacts

The introduction of an application that utilizes state-of-the-art computer vision technologies and machine learning algorithms to offer intricate statistical breakdowns of soccer matches presents various societal impacts. The opportunity this platform promises for enthusiasts, analysts, and soccer professionals is undoubtedly vast. However, several pertinent ethical and societal concerns emerge:

1. **Intellectual Property and Copyright Issues:** When users upload match footage to the platform, potential copyright infringements may arise. Many soccer leagues and organizations hold exclusive broadcasting rights. Repurposing or distributing match footage without obtaining the necessary permissions could lead to legal challenges. It is imperative to ensure that the platform does not unintentionally facilitate or encourage copyright breaches.
2. **Ethics of Data Collection:** The primary intent of such an application is the analysis of soccer matches. However, there are ethical considerations related to the storage, use, and potential dissemination of user-uploaded content. The matters of consent and data privacy are of paramount importance. Users must be thoroughly informed about the application's intent for their uploaded content.
3. **Gambling Implications:** Advanced statistical insights open avenues for potential misuse in the realm of betting and gambling. Real-time and accurate data analytics can provide gamblers with undue advantages, leading to possibilities such as match-fixing or other unethical activities. There exists a risk that the platform, if not adequately regulated, might unintentionally abet such practices.
4. **Biases in Machine Learning:** Machine learning tools are as proficient as the data upon which they train. If the foundational data is biased or lacks diversity, the insights derived could be misleading or skewed. For instance, if the platform is primarily trained on matches from a specific league or region, its evaluations and insights may falter when analyzing games from other global regions.
5. **Socio-economic Disparities:** If this application becomes pivotal for soccer analysis, there is potential for deepening socio-economic disparities. Affluent clubs with the means to access such tools might gain an advantage over grassroots teams with limited resources. Ensuring access to the platform is democratized is essential to prevent these divides.
6. **Potential Misinterpretation of Data:** The platform's advanced statistical insights and visualizations carry the risk of misinterpretation. Users with limited understanding of soccer analytics might derive incorrect conclusions, influencing team strategies, player evaluations, or fan perceptions.

While the proposed application has the potential to bring about a transformative shift in soccer analysis, it is crucial to develop and implement it with a comprehensive understanding of its societal implications. By proactively addressing these concerns, developers can ensure the platform serves the global soccer community while upholding ethical standards and societal values.

5.4 Maintenance and Lifetime

1. **Server Infrastructure:**
    a. **Scalability**: As the user base grows, the server infrastructure should be scalable to handle increased video uploads and analysis requests.
    b. **Resource Monitoring**: Monitoring server performance, including CPU, memory, and storage usage.
2. **Security:**
    a. **Data Security**: Protection of user-uploaded videos and sensitive data. Encryption, access controls, and regular security audits should be in place.
    b. **User Authentication and Authorization**: Robust user authentication and authorization mechanisms. May require two-author authentication in the future.
    c. **Privacy Compliance**: If the customer includes EU residents, the website should be in compliance with relevant data protection regulations such as GDPR.

3. **Video Analysis:**
    a. **Algorithm Maintenance**: Keeping computer vision algorithms, such as YoLo, up-to-date and improving them as necessary.
    b. **Data Sources**: Maintaining access to external data sources and APIs (e.g., player statistics) and addressing changes or disruptions.
    c. **Resource Management**: Efficiently managing resources required for video analysis, such as GPUs or cloud computing services.

4. **Content Management:**
    a. **Storage and Archiving**: Developing a strategy for storing and archiving user-uploaded videos and analysis results over time.
    b. **Content Cleanup**: Providing mechanisms to enable users to manage or delete their uploaded content.
    c. **Content Moderation (future development)**: Implementing content moderation to prevent inappropriate or copyrighted content from being uploaded .

5. **User Experience:**

    a. **Usability Testing**: Continuous collection of user feedback and ongoing usability testing are crucial to improve the user experience and interface.
    b. **Performance Optimization**: Regular optimization of the website's performance ensures fast loading times and smooth user interactions.

6. **Data Visualization and Analytics:**
    a. **Visualization Tools**: Maintaining and updating data visualization tools and libraries to create up-to-date visualizations is attractive.
    b. **Data Accuracy**: Ensuring that analyzed statistics and visualizations accurately reflect the content of the soccer match videos is essential.

7. **Compliance and Legal Considerations:**
    a. **Licensing and Copyright**: Ensuring the necessary rights to use and display videos, data, and visualizations on the website to avoid copyright issues.

8. **Documentation:**
    a. **Documentation Updates**: Keeping project documentation, including code comments and user guides, up-to-date as the project evolves.
    b. **Knowledge Transfer**: Planning for knowledge transfer within the development team ensures continuity in case of team member changes.

9. **Community and User Engagement:**
    a. **User Support**: Providing customer support channels for users to report issues and seek assistance.
    b. **Community Building**: Fostering a community around the website through engagement with users and responsiveness to their feedback is beneficial.

10. **Business and Monetization Strategy:**
    a. **Monetization (if cooperated)**: Continuously evaluating and adapting the monetization strategy based on user engagement and feedback.

11. **Technology Stack:**
    a. Staying current with updates and security patches for the technologies and frameworks used.

12. Budget and Funding:

a. Ensuring a sustainable budget or funding source to cover server costs, maintenance, and potential expansion.

13. Regulatory Changes:

a. Staying informed about changes in data protection, internet, or technology regulations that may impact the project

5.5 Cost Analysis

Development Cost:

1. Soccer API: API-Football, Free
2. Computer Vision Tools: YoLo, Free
3. Frontend Deployment: Vercel, Free
4. Server: AWS, Per AWS Fee
5. Data Storage: Amazon S3 ($0.023 per GB), PlanetScale (Free)
6. Computing Platform: Google Colab, $10 per month

6. References

1. “StatsBomb Data: Event Data.” *StatsBomb*, 11 Sept. 2023, statsbomb.com/what-we-do/soccer-data/.
2. “IQ Soccer: Soccer Data Analytics Platform.” *StatsBomb*, 21 Sept. 2023, statsbomb.com/what-we-do/iq-soccer/.
3. Arastey, Guillermo Martinez. “StatsBomb: Advanced Football Analytics through an Interactive Platform.” *Sport Performance Analysis*, Sport Performance Analysis, 24 Jan. 2020, www.sportperformanceanalysis.com/article/statsbomb-advanced-football-analytics-through-an-interactive-visualisation-platform.
4. “Opta Vision.” *Stats Perform*, 26 July 2023, www.statsperform.com/opta-vision/.
5. Arastey, Guillermo Martinez. “Opta Sports: The Leading Sports Data Provider.” *Sport Performance Analysis*, Sport Performance Analysis, 22 Nov. 2019, www.sportperformanceanalysis.com/article/opta-leading-sport-data-provider.
6. Kasai, Kazuki, et al. "Robust and Efficient Homography Estimation Using Directional Feature Matching of Court Points for Soccer Field Registration." *IEICE TRANSACTIONS on Information and Systems* 104.10 (2021): 1563-1571.
7. Statsbomb. “Statsbomb/Open-Data: Free Football Data from StatsBomb.” *GitHub*, github.com/statsbomb/open-data. Accessed 3 Oct. 2023.