

# Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer

Javier Fernández  
F.C. Barcelona  
javier.fernandezr@fcbarcelona.cat

Luke Bornn  
Simon Fraser University, Sacramento  
Kings  
lbornn@sfu.ca

Dan Cervone  
Los Angeles Dodgers  
dcervone@gmail.com

## 1. Introduction

What is the right way to think about analytics in soccer? Is the sport about measured events such as passes and goals, possession percentages and traveled distance, or even more abstract notions such as mistakes (to quote Cruyff, “Soccer is a game of mistakes, whoever makes the fewer wins”)? Analytical work to date has focused primarily on these more isolated aspects of the sport, while coaches tend to focus on the tactical interplay of all 22 players on the pitch. Soccer analytics is lacking from a comprehensive approach that can start to address performance-related questions that are closer to the language of the game. Questions such as: who adds more value? How and where is this value added? Are the teammates creating spaces of value? When and how should a backward pass be taken? How risky is a team attacking strategy? What is a player’s decision-making profile?

In order to make an impact on key decision-makers within the sport, soccer analytics requires a comprehensive tool to facilitate a continuous cycle of questions and answers with coaches. Analytic methods must therefore encapsulate a wide set of actions of interest to coaches to provide a detailed and flexible interpretation of complex aspects of the game. In this paper we develop such a model, measuring and elucidating instantaneous value on the pitch. Specifically, we quantify the expected outcome at every moment in a possession, driven by a fine-grained evaluation of the full spatio-temporal characteristics of the 22 players as well as the potential value of ball drives, shots, or passes to any location. While our aim is similar to that of the expected possession value (EPV) approach in basketball [1], our focus on soccer necessitates a drastically different approach to account for the nuances of the sport, such as looser notions of possession, the ability of passes and drives to happen at any location, and space-time dependent turnover evaluation. Moreover, the model is designed in a decoupled way which provides great interpretative power for both visual and quantitative analysis of game situations. Specifically, deep



learning-based component models are built to capture the complex intricacies of spatiotemporal tactics, while a high-level stochastic process model fuses each component model together in a cohesive, interpretable way.

In the next section we present a description of the technical characteristics of the model, then through the remainder of the paper we focus on showcasing several practical applications, covering highly detailed evaluations of passing, off-ball value creation and distribution, and both team and player level decision making.

## 2. An expected possession value (EPV) framework

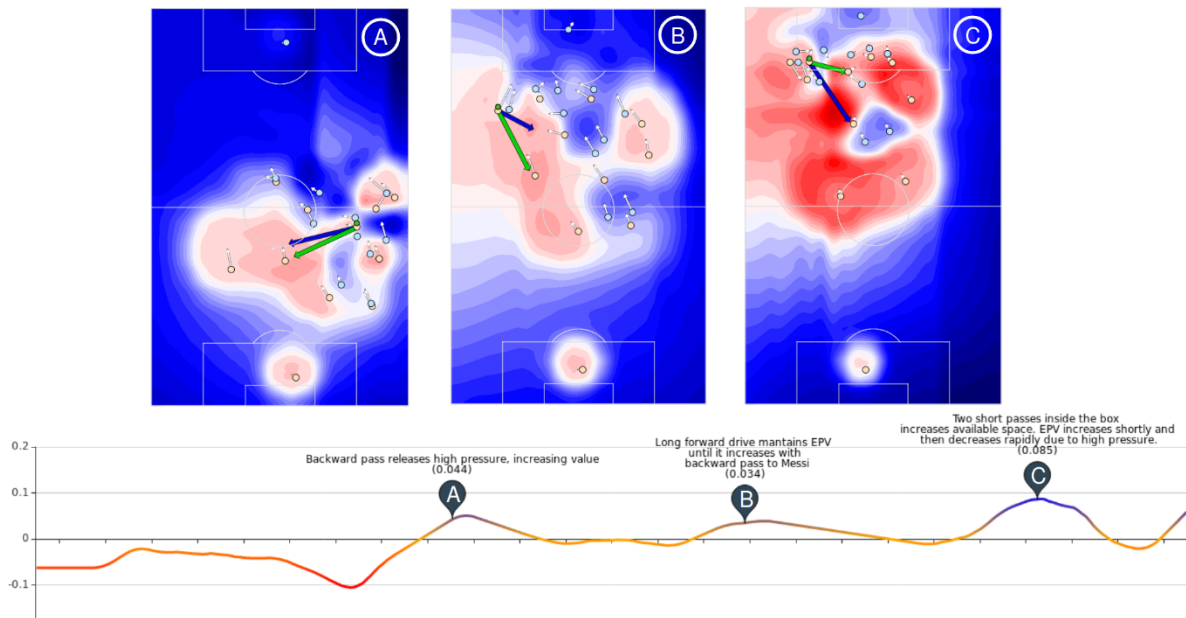
We define expected possession value (EPV) as the expected outcome of a soccer possession based on the full resolution spatiotemporal data. EPV is a real number in the  $[-1, 1]$  range, expressing the likelihood of a possession ending in a goal for the attacking team (1) or a goal by the defending team (-1) after an immediate possession regain. Similarly to the EPV approach for basketball [1, 2] our model provides a frame by frame estimation of the outcome of the possession, acting as a stock ticker for the expected outcome of a possession. Aside from a shared high-level goal, however, our approach is drastically different, driven by the underlying differences in the two sports. As one concrete example, in soccer we cannot assume that passes are played directly to a player's location, as the ball can be played into open space in front of or behind the intended receiver; as such, we need to consider the full space of potential destination locations. As another example, there is no time limit for soccer possessions (aside from the 45 minute half), with complex and often blurred dynamics between offense and defense.

Figure (1) presents the evolution of the EPV value during a soccer possession in a Real Madrid vs FC Barcelona match during the 2017-2018 La Liga season (see [https://lukebornn.com/sloan\\_epv\\_curve.mp4](https://lukebornn.com/sloan_epv_curve.mp4) to watch the full sequence). The curve represents the expectation outcome (home vs. away goal) frame by frame. We observe that at the beginning of the possession, the EPV reaches -0.09 as a consequence of high spatial pressure by the opponent during the build-up. The negative value indicates that in this scenario it is more probable that the defending team regains possession (and possibly scores) than the possession ends in a goal for the attacking team. Next, a back pass, an often criticized action in soccer, is shown to increase considerably the EPV value (up to 0.04) by providing better conditions to further the possession up the pitch. The wave-shaped curve also provides a good example of the complexity and fluctuating nature of soccer. The images above the EPV curve show the value distribution for passes along the whole field, where the blue and green arrows indicate the taken pass and the best pass possible to take, providing a glimpse at the interpretative power of the model.

Recent work in soccer analytics has provided powerful models for inspecting a variety of specific game situations, including pass risk and quality [3], attacking shot danger [4], and off-ball positioning in shooting opportunities [5]. While being successful in the specific task they resolve, there is no clear path on how we could join these models together into a more comprehensive framework of analysis. Other approaches make use of Markov processes to correlate sequence of events within possession sequences with the probability of scoring [6, 7, 8], providing a similar overall objective of that of EPV but with lower interpretability capabilities than the presented

model. Here, we propose a decoupled model for EPV that decomposes goal value into the expected values of different actions that further the possession, along with the probabilities of these actions. Equation (1) presents EPV, the expected outcome of the possession given all the spatiotemporal information at time  $t$  ( $T_t$ ), as the composition of independent models for three main types of actions ( $A$ ): passes ( $\rho$ ), shots ( $\varsigma$ ) and ball-drives ( $\partial$ ). Different sets of these components are estimated independently, such as the value surface for passes, the value surface for ball-drives, the likelihood of one of these events taking place, and the expectation of goals for shots.

$$EPV(t) = E[X|T_t] = \underbrace{E[X|A=\rho]}_{\substack{\text{Passing value} \\ \text{surface} \\ \text{dependent on} \\ \text{destination} \\ \text{location}}} \underbrace{P(A=\rho)}_{\substack{\text{Expected} \\ \text{goals} \\ \text{model}}} + \underbrace{E[X|A=\varsigma]}_{\substack{\text{Expected} \\ \text{goals} \\ \text{model}}} \underbrace{P(A=\varsigma)}_{\substack{\text{Expected} \\ \text{goals} \\ \text{model}}} + \underbrace{E[X|A=\partial]}_{\substack{\text{Drive value} \\ \text{surface} \\ \text{dependent on} \\ \text{destination} \\ \text{location}}} \underbrace{P(A=\partial)}_{\substack{\text{Expected} \\ \text{goals} \\ \text{model}}} \quad (1)$$



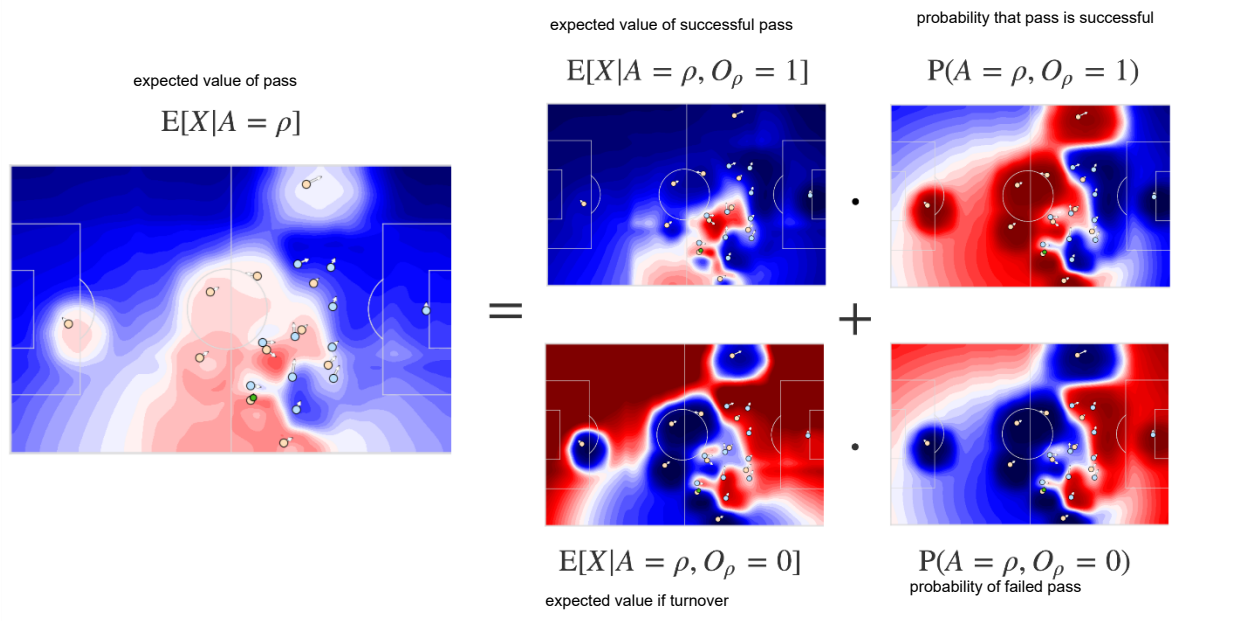
**Figure 1:** The image below represents the expected possession value of a possession during a Real Madrid vs FC Barcelona match, during the 2017-2018 La Liga season. Three specific situations A, B, and C, are highlighted in the three images above, from left to right. Each image presents the expected value surface if a pass action is performed, with a cool to warm divergent color scale, where cool (blue) colors represent negative EPV and warm (red) colors represent positive EPV. The direction of the attack goes from bottom to top. A blue and green arrow is shown for every image, representing the taken pass and the best possible pass according to the model, respectively. Yellow and blue circles represent the attacking and defending team locations, respectively, while a green

circle represents the location of the ball. See this link [https://lukebornn.com/sloan\\_epv\\_curve.mp4](https://lukebornn.com/sloan_epv_curve.mp4) for a video of the full sequence.

Passes are modeled alongside turnovers. Equation 2 presents a further decomposition of the passing model to **account for the success of a pass  $O_\rho$** .

$$E[X|A = \rho] = \underbrace{E[X|A = \rho \cap O_\rho = 1]}_{\text{Expected value of successful pass}} \overbrace{P(A = \rho \cap O_\rho = 1) + P(A = \rho \cap O_\rho = 0)}^{\text{Pass probability model}} \underbrace{E[X|A = \rho \cap O_\rho = 0]}_{\text{Expected value of turnover}} \quad (2)$$

Figure 2 presents a glance at the interpretative power of this decoupling. Here we can observe how a passing value surface accounting for the effects of successful passing and turnovers can be decomposed into four other visually interpretable surfaces: the expected pass value, the expected turnover value, and the probabilities of both passing and turnover to any given location.



**Figure 2:** The center-left image represents the **expected value surface if a pass is taken** by the attacking team to every possible location. The four images at the right represent, from left to right and top to bottom: the **value surface of successful passes**, the **pass probability surface**, the **value surface of turnovers (unsuccessful passes)** and the **turnover probability surface**. Yellow and blue circles represent the attacking and defending team locations, respectively, while a green circle represents the location of the ball. The direction of the attack goes from left to right.

Each of the model components is estimated **independently through machine learning algorithms based on a wide set of spatiotemporal features**. Pass and turnover probabilities are estimated using logistic regression. Action likelihood is modeled through a convolutional neural network on top of pitch control and pitch influence surfaces based on a recent statistical model for pitch control [9]. **Pass and ball-drive value expectation are learned from a set of carefully designed deep neural networks.** All the models are carefully tuned and calibrated for accuracy at the

component level as well as the joint EPV level. Recently, neural networks have been explored in a similar context [10], by estimating player positioning and probability of reaching certain locations on the field, which is fundamentally different from the objectives of our approach. For the purposes of this paper we focus **not on the technical details, but rather the power of the decoupled modeling framework in answering many important questions previously unanswered by the soccer analytics community.** With an eye towards demonstrating this paper, we place heavy emphasis on the passing components of the model, as the shot and goal aspects of the model are fundamentally **variants of expected goals models and hence already well-studied.**

Models were fitted using optical tracking data provided by *STATS LLC* for the 2012-2013 Premier League season, as well as optical tracking data provided by *Footovision* for FC Barcelona matches in the 2017-2018 and 2018-2019 La Liga season.

### 3. The role of context within the unpredictable sport

A critical aspect to properly evaluate soccer situations is to have a clear understanding of the ongoing context, or what teams' and players' intents are. Soccer is so dynamic that sometimes even marking the best player is not a good option, as Johan Cruyff would say "there are very few players who know what to do when they're not marked. So sometimes you tell a player: that attacker is very good, but don't mark him". **Most of the time the location of an event in space and time alone is not sufficient to fully evaluate its potential impact.** For example, to control a ball in the midfield is usually more difficult when the opponent defends with a high line of pressure, but simpler if the opponent is "parking the bus". Many of these examples flood the tactical analysis of the game. In tactical analysis, organized possessions can be structured theoretically in three phases: build-up, progression and finalization. The build-up is a phase where a possession starts to develop from one's own half, typically with the majority of the opponent team behind the ball. Once the first pressure line or the midfield is reached, the possession is considered to reach a progression phase, where the objective is to keep progressing towards the opponent goal. Then, when the possession reaches zones near the box, the possession enters a finalization stage, where the objective is to score. During the same possession a team could return to a previous phase. While these stages can take many names and forms, the underlying idea is that a **possession goes through different stages where the contextual characteristics differ**, and are not straightforward to define. Also, for each stage the characteristics of on-ball and off-ball actions might vary, given that the objective of each phase varies as well.

In many of the regular meetings that took place with tactical analysts of FC Barcelona youth teams during the development of this work, there was a recurrent concept associated with the evolution of possessions: **passing across opponent's pressure lines** (or formation lines), often called breaking lines. Analysts recognize the possession to be in a different context depending on whether the ball possession is located behind, between, or beyond the opponent's formation lines. We use **this concept as a proxy to identify where an action is located (relatively) according to context.** Relative locations are found by first **obtaining the dynamic lines of formation of** the defending team at every time step. We approach this by applying **spectral clustering** to the set of mean X coordinate position of each defending team player in a fixed 2-second time window. For simplicity of application we fix the number of clusters to be three, thus the obtained formations provide vertical alignments of three groups: the first pressure line (typically forwards), the second pressure line



(typically midfielders and wingers) and the third pressure line (typically defenders). In some of the practical applications presented in this paper we group actions that were situated in a particular set of relative locations; for example, behind the first pressure line. For better interpretation we could also loosely assume that actions behind the first pressure line correspond to build-ups, actions located between the first and second line correspond to progression phase, and actions located beyond the second line correspond to a finalization phase. This concept of relative location is used throughout all the sub-models of the presented EPV-framework, showing to be an important spatial feature.

Defensive lines are only one contextual aspect of the game. Optical tracking data has recently enabled studies of other contextual factors as well, such as player motion and ball-interception [11], distances and angles between players and locations [12], automatic detection of match-wide formations [13], and expert-guided handcrafted features [4], among others. We employ a suite of these contextual factors (such as movement, distances, angles, etc.) within the component models to improve interpretability and predictive ability.

## 4. Value: The missing key to pass analysis

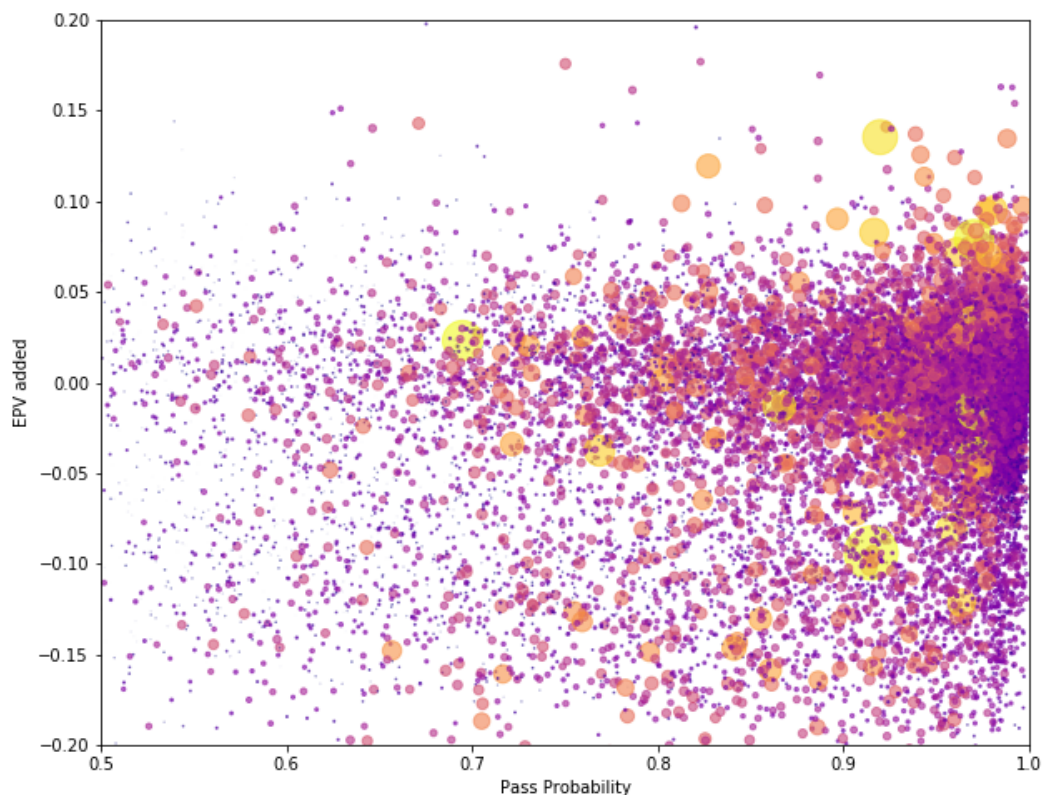
Passing is the most frequent action in soccer, reaching over 800 passes per match on average. Soccer analytics has longed focused on evaluating teams and players based on the total sum of passes, the overall accuracy, or assists (a rare kind of pass). Perhaps that is one of the reasons why the Spanish national team, 2010 World Cup champion, was praised for their “Tiki-taka” style of play, a term that highlights the high frequency of passes (short passes in particular). However, the same Spanish national team with a similar style was not surprisingly defeated at 2018’s World cup, despite reaching over 1100 passes in the match where they got eliminated. Hence while passes are a central element of the game, quantitative analysis has shown difficulties in disentangling the relative impact of every pass. In order to fully evaluate the impact of passes, as many other actions in the game, we need to first understand the many intentions that a pass or a pass sequence might have. Pass sequences can be created to attract an opponent, to open the possession into new passing or ball-drive situations, to push gradually the opponent to one side of the field and move the ball to the other, or more typically to provoke disorder situations in order to find spaces of greater value. In the words of the renowned organizing midfielder Xavi Hernandez: “I spend the entire 90 minutes looking for space on the pitch. I’m always between the opposition’s two holding midfielders and thinking: the defense is here, so I get the ball and I go there to where the space is”.

The model proposed in this paper presents a key element for elevating pass analysis: the concept of value. The pass model, once coupled with the other models, allows us to evaluate the impact of passes and decision-making based on several factors such as the expected value of the successful pass (reward), the expected value added to the possession (expected pass EPV added), the expected value in case of turnover (risk) and the probabilities of the pass reaching a given location. This can be evaluated for every location on the pitch at any given time, in relation with the quality of the possession.

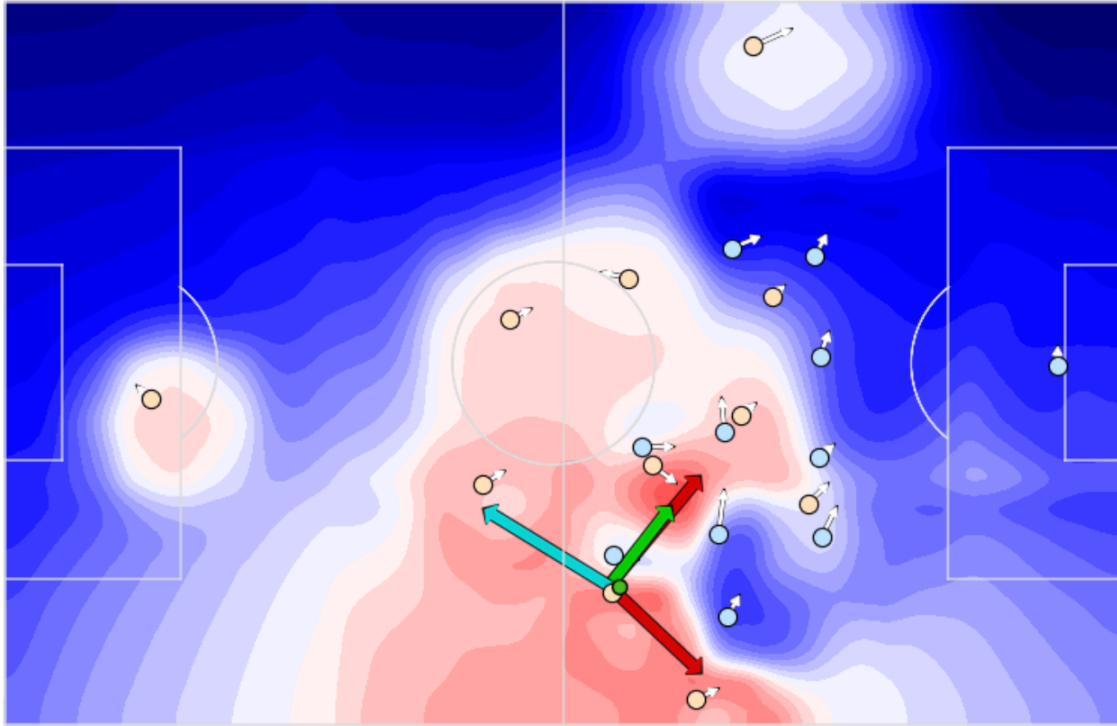
Figure 3 presents the probability of passes being completed against the gain in EPV of the possession after the pass is taken (EPV added), where size and color both represent the expected EPV added at the moment of the pass (the difference between the expected EPV added and the

current EPV). From this image we can identify several characteristics of passes in soccer. The biggest fraction of passes are created with high probability of success, low expectation of adding value to the possession and ultimately end up adding or subtracting little value from the possession. Passes with high likelihood of adding value (big yellow circles) tend more to be failed and subtract value from the possession, but when successful they can add more value than less ambitious passes.

The decoupled nature of the model allows us to inspect the decision-making process of individual situations, which can be useful for profiling a player's passing tendencies, evaluate off ball positioning or analyze opponent marking, among many others. Figure 4 shows the expected pass value surface in a specific game situation. Here we can observe three concepts related to decision-making: **high reward passes, low risk passes, and risk-reward balanced passes**. The red arrows showing the two passes with highest reward, i.e. the two actions that would maximize the quality of the possession if successful, disregarding the consequences of a turnover. These are reward-driven passes. The cyan arrow shows the pass that minimizes the turnover expected value, corresponding to the pass that assumes the least risk among all the possible. The green arrow shows the best pass expected value considering both passing success and turnover risk.



**Figure 3:** Pass probability against EPV added for all the passes in FC Barcelona matches in the 2017-2018 La Liga season. Color and size of the circle represent the expected EPV added prior to the pass attempt.



**Figure 4:** The image presents the expected pass value surface in a specific game situation, using a cool to warm divergent color scale, where cool (blue) colors represent lower EPV and warm (red) colors represent higher EPV. Yellow and blue circles represent the attacking and defending team locations, respectively, while a green circle represents the location of the ball. The three colored lines represent potential passes. The red lines represent the two passes providing the biggest reward, the green line the best pass based on the expected pass surface, and the cyan line represents the pass with lower turnover expected value.

#### 4.1 Evolving pass-maps

One of the most popular visualizations in soccer is the pass-map visualization, consisting of the average location of every player on the field, a player circle size according to frequency of passes taken by the player, and lines between players with size and color related to frequency of passes. More sophisticated versions of pass-maps assign color based on models of the contribution of the player to attacks ending in a shot. While pass-maps are great tools to understand frequency of passes between pairs of players, they fail to recognize whether those passes added or subtracted value and the distribution of those passes. We introduce here an evolved version of pass-maps that incorporates EPV-related metrics, in order to better evaluate passing quality.

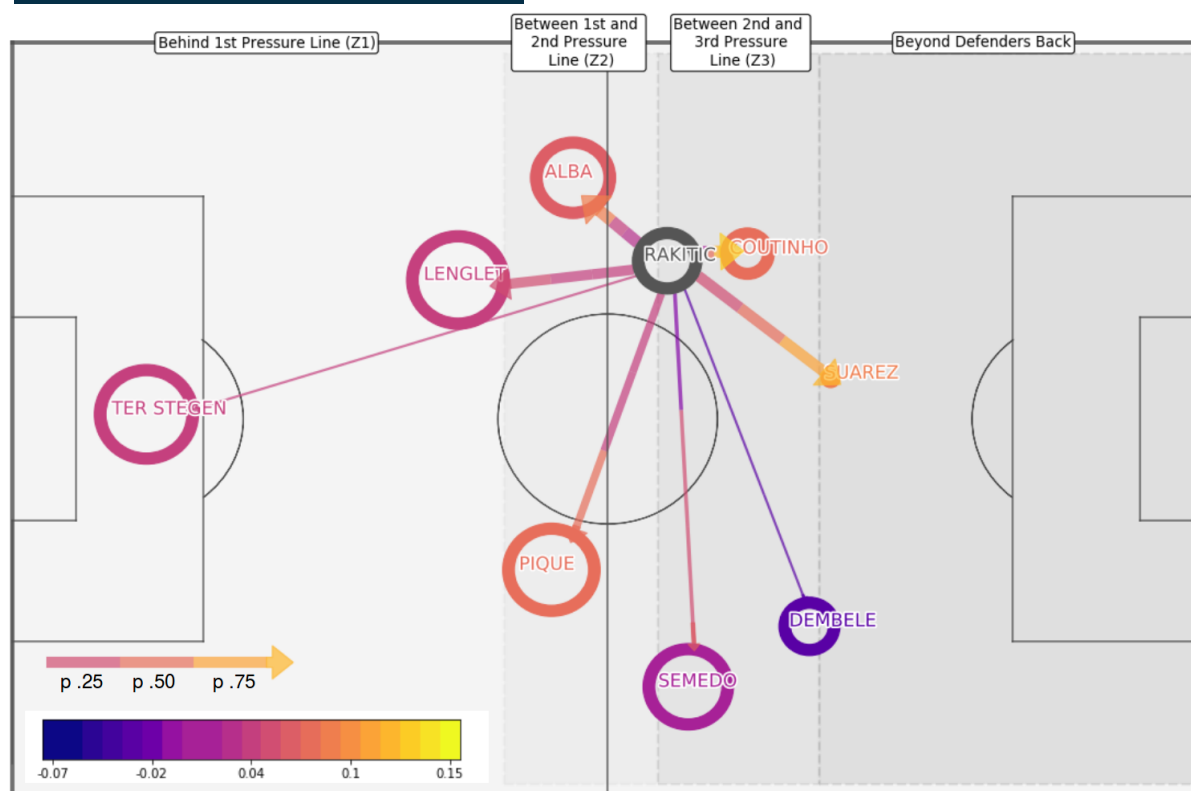
Figure 5 presents a pass-map for all passes departing from Rakitic in a single FC Barcelona match in La Liga season 2017-2018. Just one player is shown for simplicity. This pass-map introduces several concepts. Players are located at their mean location for all the receiving passes.



The size of a player's circles is proportional to the mean pitch control they had when the pass was taken, where smaller sizes represent lower space control, thus high pressure on the player. Here we can see the central defenders are receiving balls from Rakitic with low pressure, while the wingers and forwards are under considerably higher pressure (Suarez in particular). The color of the lines and circles is given by the EPV added metric (although any other EPV metric could be used) of the passes in consideration. Circles are colored with the mean EPV added by passes received by the player (except for Rakitic). The colors of the lines should deserve special attention. Since passes in soccer vary so much according to context, coloring based on a sole summary statistic such as the mean or median EPV added will provide a considerable loss of information. In order to get a closer view into the passing distribution, arrows are divided into three equally sized blocks. Each block is colored according to the EPV added of the passes in the percentile 25, 50 and 75 respectively of each corresponding distribution of passes between players. Having this, we can observe that while passes to Suarez incorporate a wide range of results (from 0.01 to 0.12), the top 25% of passes where of great value. Also, we can observe that the top 25% of passes to Piqué (typically back passes) managed to provide above 0.05 EPV added. The plot also shows the average location of each pressure line when passes by Rakitic were performed, thus the locations of the teammates also provide an average location according to average formation line, providing a grasp of the relative positioning of the players. More detailed information can be obtained from the plot when filtering for specific situations, for example, passes behind the first pressure line.

## 4.2 Passing relies on context

In Section 3 we detailed the idea of contextualizing actions based on the relative location according to the mean position of pressure lines at a given time. Here, we leverage that concept to provide a contextualized view into the passing performance of a set of players in a given match. Figure 6 presents a comparison of the value added through passing between relative locations for two central defenders (Sergio Ramos and Gerard Piqué) and two midfielders (Luka Modric and Sergio Busquets) for a FC Barcelona vs Real Madrid match, in La Liga season 2018-2019. Each column groups passes taken from the red colored area shown on the first row. Each of the four areas (that we call Z1, Z2, Z3 and Z4) represent the space between the own goal and the first pressure line (Z1), the space between the first pressure line and the second (Z2), the space between the second and the third pressure line (Z3) and the space between the third pressure line and the opponents goal (Z4). Notice that these zones are not predefined but they move dynamically according to the opponent's location at every time step. Each plot shows a stacked bar chart representing the same concept as arrows in Section 4.1. That is, the distribution of passes is split into three equally sized groups (favoring the top in case of not exact division by 3) and colored according to the EPV added after the pass. The x-axis locations represent the destination of the pass, indicating if it keeps the ball in the current relative zone or goes back or forward to any other zone.



**Figure 5:** Pass-map for all passes of Rakitic in a 2017-18 FC Barcelona match. Circles are located in the mean pass destination location for every other player, and located in the mean pass origin location for Rakitic. Circle size is related to the pitch control the player had when making the pass, where smaller means less space (higher pressure). The color of the circles represent the mean added value of those passes. The lines are split into three equally sized blocks. Each block (starting from Rakitic) is colored by the added value associated with the percentile 25, 50 and 75 in the distribution of passes between those two players. The gray areas represent the space between the mean pressure lines of the opponent.

Let's analyze this figure column by column. For actions behind the first pressure line (Z1) we observe no passes from the midfielder (Modric) and two different pass tendencies from central defenders (Ramos and Piqué). Ramos had a higher tendency to overcome the first line of pressure towards the second but this resulted most of the time in a loss of value for the possession. Piqué showed a higher tendency to keep possession value by passing behind the first pressure line, but also attempts to overcome the first and the second pressure line. His three attempts to overcome the first line of pressure successfully added value to the possession. The second column represents passes starting between the first and second pressure line, typically during the progression phase of the possession. For the central defenders, we observe again two different tendencies. Piqué passed back to Z1 twice as much as Ramos, both with the tendency of losing an average of 0.01 EPV in their possession when passing behind this line. When keeping the ball in Z2 Piqué was able to increase the EPV of the possession while values for Ramos in that zone were considerably lower, providing a hint for different types of pressure received by each team. Regarding the midfielders, we can see the need for analyzing the distribution of passes instead of jumping directly to summary statistics. For both midfielders, two thirds of the passes subtracted or added little EPV to the possession, however, the last third of passes were able to add value to the possession.

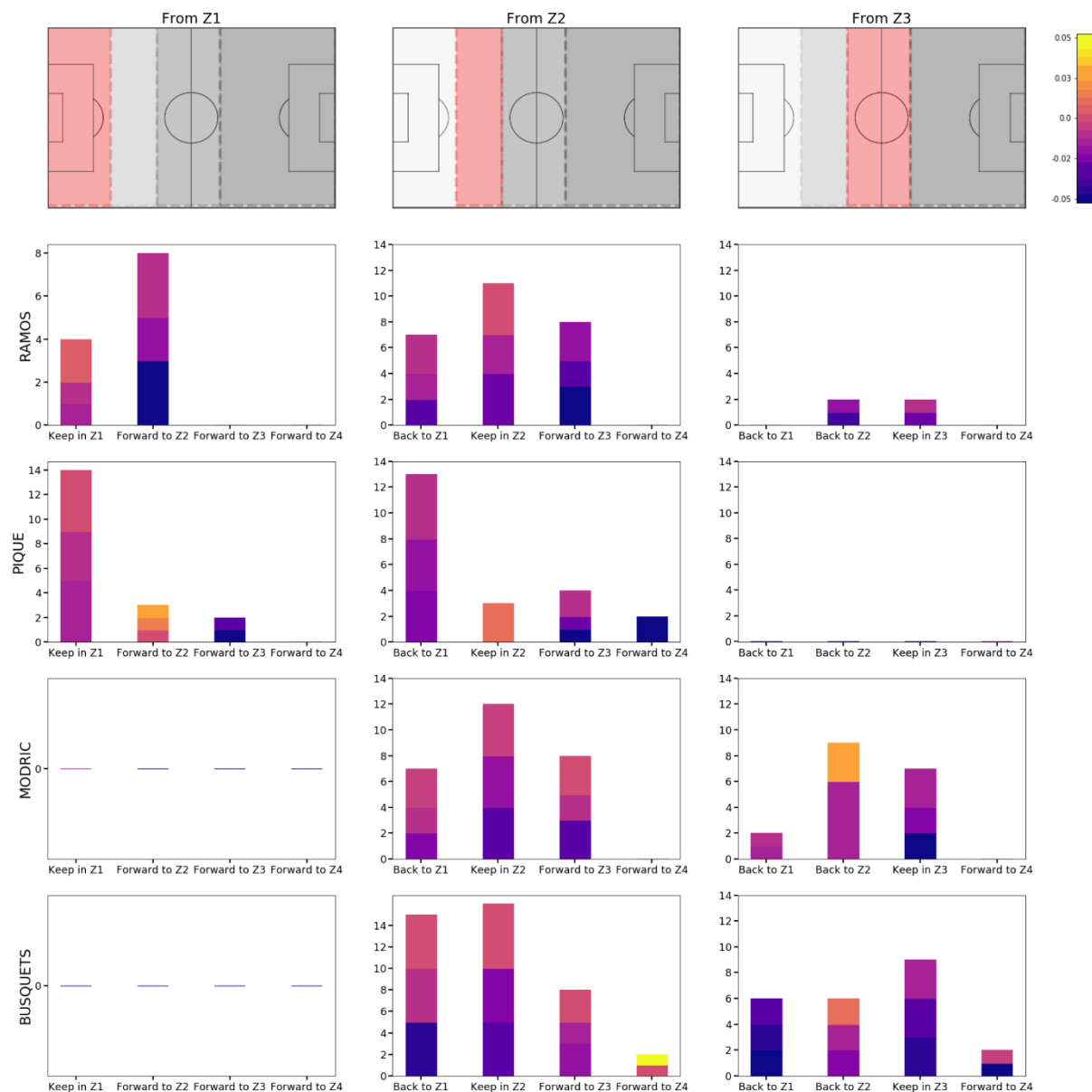
Remarkably, Busquets was able to provide two passes beyond the defenders back adding a considerable amount of value. For passes starting between the first and second line (Z3) we have different situations. Here, the contribution of central defenders was very low for Ramos and non-existent for Piqué. Ramos was not able to add value through passing while remaining in this zone, however his presence in this zone adds more information to the match analysis regarding the tendency of the defender to contribute with the attack. For both midfielders passes within the same zone presented a consistent loss of value, which shows the difficulty of adding value in the relative zone in soccer where the highest pressure is found. Remarkably, back passes to Z2 found added value in the third of the cases for both Modric and Busquets, but subtracted value when they went back to Z1, showing again the changing nature of soccer according to context.

### 4.3 The risk-reward dichotomy

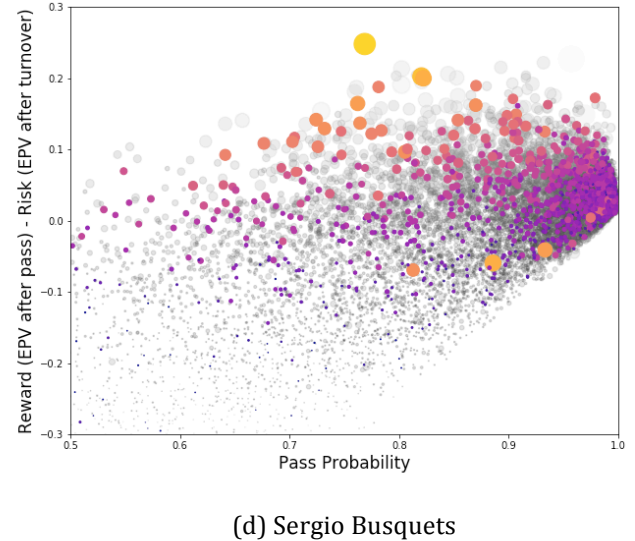
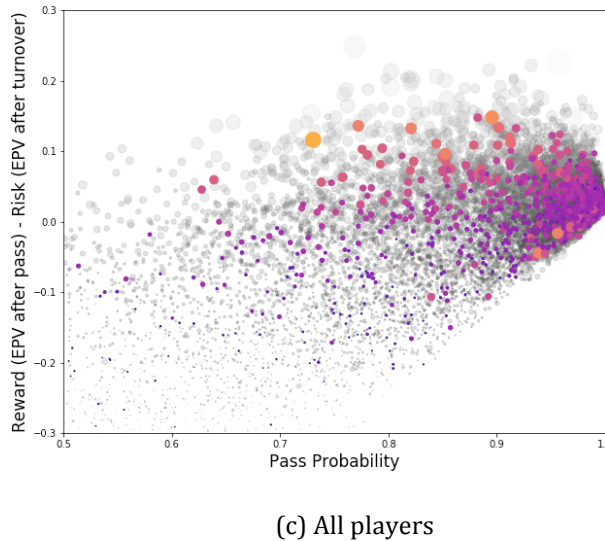
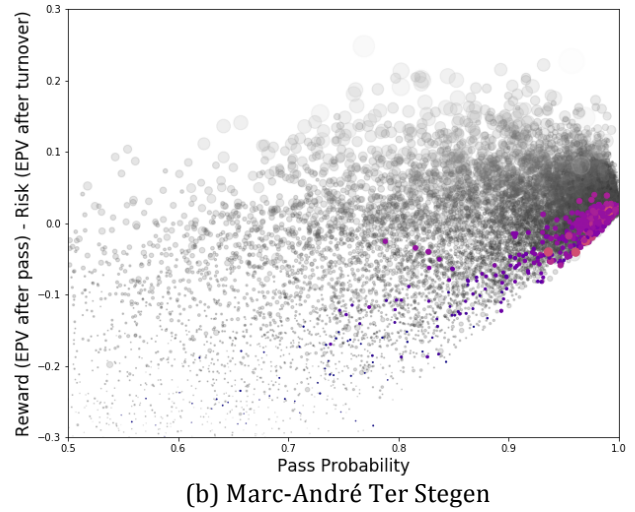
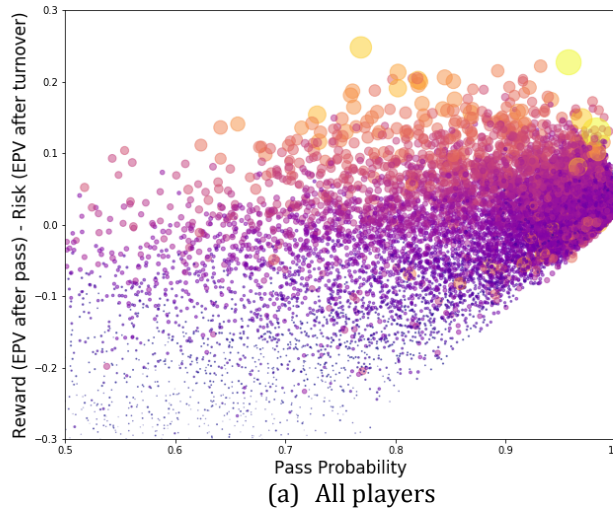
The passing component of the model allows to derive two important concepts for pass analysis: risk and reward. The risk of a pass can be approached from the perspective of the possession and defined as the probability of the opponent scoring in case of the pass ending in turnover. This might be a more accurate representation of risk than the probability of turnover, as some turnovers are much more costly than others. Risk can then be represented by the turnover EPV of a given pass. Similarly, we can define reward as the expectation of value of a current possession given the pass is successful, which can be represented by the pass EPV component. The analysis of risk and reward can provide helpful information for analyzing the passing profiles of teams and players. The analysis of passes is usually approached by quantifying the ratio of successful and unsuccessful passes, and pass probability is used to evaluate the difficulty of completing that pass. However, these two statistics are insufficient to understand more fine grained characteristics of passing, such as the trade-off between lower or higher pass probability and higher or lower reward.

Figure (7) presents four images comparing the difference between reward and risk of a pass, and the associated pass probability, for all the passes in FC Barcelona matches in La Liga season 17/18, for all the players and three other different players: Marc-André Ter Stegen, Samuel Umtiti and Sergio Busquets. From the general distribution of passes we can observe that lower probability passes are not necessarily associated with lower reward. In fact, for pass probabilities between 70% and 90% there is a high set of passes with a high positive balance of reward and risk. For lower passing probabilities the amount of cases with negative balance of reward and risk starts to increase considerably, as it could be expected. Ter Stegen shows a notable difference with other field players such as Umtiti and Busquets, by taking frequently high probability passes with a balanced risk and reward. We can clearly observe the FC Barcelona tendency of starting build-ups from the goalkeeper and to open spaces to provide a high probability and lower risk pass. In case the passes would have lower probabilities and less risk, we would evidence long-ball actions. Umtiti, a central defender, usually takes secure passes (between 92% and 100% probability) that have a balanced risk and reward ratio. This coincides with the expectation for central defenders to prefer lower reward passes that are more secure and with less turnover values, given the associated danger with their position on the field. Noticeable, Umtiti also presents several passes with a positive difference between reward and risk adding more than 0.05 EPV, highlighting his known contribution with the possession during the progression phase. Busquets, on his side, shows a remarkable amount of passes with high expectation of adding value (even above 0.1 EPV) at a wide range of passing probabilities. An analysis in this pass distribution would provide interesting feedback for opponent scouting by noticing that a defensive mid-fielder like Busquets has a

remarked tendency to attempt high reward passes with different probability levels, which is unusual for players in this position, and might become a relevant skill to counter.



**Figure 6:** The image presents the distribution of EPV added for passes between relative zones for four different players in a FC Barcelona vs Real Madrid match of La Liga season 2018-2019. Columns group passes taken from three relative zones: behind the first pressure line (Z1), between the first and second pressure line (Z2) and between the second and third pressure line (Z3). The stacked bar charts represent the frequency of passes and are split into three equally sized groups. From bottom to top the color of the bars correspond to the EPV added of the actions at percentile 25, 50 and 75 respectively. X-axis in the bar charts represent the destination of the pass, including a fourth zone (Z4) corresponding to the space between the third pressure line and the opponents goal. The direction of the attack goes from left to right.



**Figure 7:** The figure presents four images comparing the pass probability and the difference between reward and risk for all the passes in FC Barcelona matches in La Liga season 17/18, for all the players and three other different players. Reward is represented by the EPV after a pass and risk by the EPV after a turnover. The color and size of dots represent the expectation of EPV added by the pass.



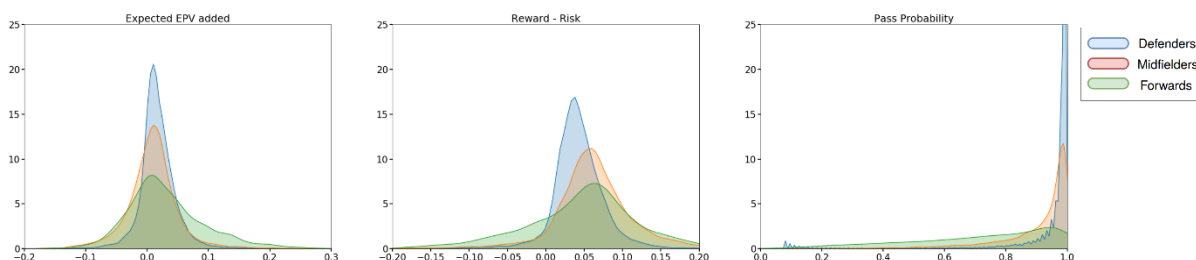
## 5. Distilling off-ball value creation

While on-ball actions tend to capture most of the attention in soccer analytics, the value of off-ball actions (what players do when they don't have the ball) is still yet in its first steps. The different value surfaces provided by the model allows us to inspect the quality of positioning of teammates during the match and across actions, even if they didn't end up receiving the ball. In this section we first present the distribution of off-ball value across different soccer positions, and then we present a detailed inspection into the surface of off-ball value generation by individual players.

### 5.1 Off-ball demands according to positions

The dynamic nature of soccer makes difficult the assigning of fixed positions for players. Depending on the situation, a supposed defender like Jordi Alba can act as a winger, or a pivot like Busquets can temporarily assume the role of an attacking midfielder. However, taking a broad perspective we can make the assumption that there are specific demands by position when it comes to off-ball movements and positioning. In Figure (8) we compare the off-ball value of three different EPV metrics for defenders, midfielders and forwards. Off-ball value is calculated every time a pass is taken and every 1 second of ball-drives, taking into consideration all the 10 teammates in each case (not just the player receiving the pass).

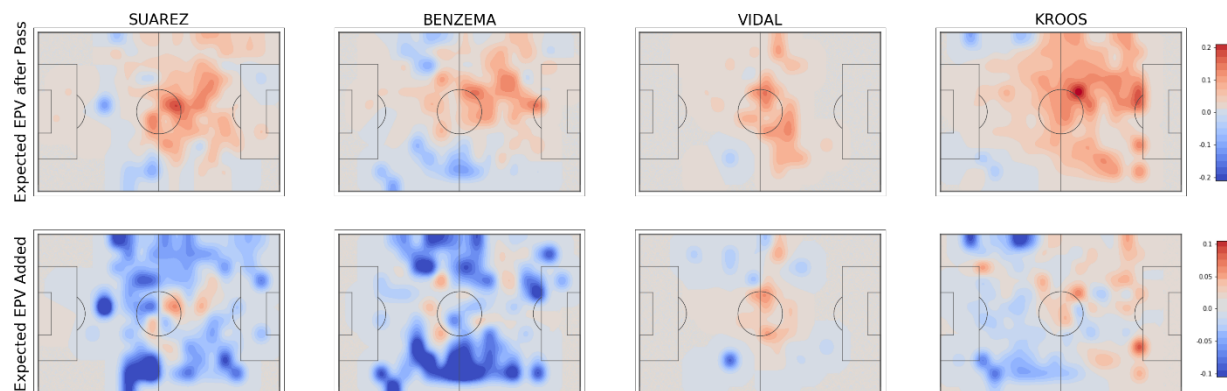
The first metric is expected EPV added, where we can see defenders being positioned in locations with low expected loss of value and low expected gain of value. The distribution widens for midfielders to a  $[-0.1, 0.1]$  range of expected EPV added. Forwards show a higher tendency to move towards locations of higher value, which is consistent with the mobility demanded of that role. The second metric compares the difference between the reward of a pass (pass expected value) and the risk (turnover expected value). Here we can see again that forwards tend to have a wider distribution favoring more extreme reward and risk conditions. Midfielders tend to control more their positioning into less risky passing situations but are still able to reach high value locations. Defenders tend to keep positions with slight addition of value but reducing considerably to be placed in locations with high risk of turnover. The third column presents the pass probability for off-ball positioning. Here the tendencies presented in the other two cases are maintained. Forwards, however, present a considerably wide distribution reaching very low passing probabilities, which shows the high pressure put upon them. Surprisingly forwards are able to generate as much and even more off-ball value than the other two positions, presenting a considerably lower probability of successfully receiving the ball.



**Figure 8:** A kernel density estimation for the off-ball value of three different EPV metrics comparing defenders (blue), midfielders (red) and forwards (blue). The first metric is expected EPV added of pass to that player. The second metrics is the difference between the expected value of a successful pass (reward) and the expected value of turnover (risk). The third metric shows the pass probability of success.

## 5.2 Beyond location heatmaps

A common visualization in soccer analytics is player's location distribution on the field, which basically shows where the player was located or took actions during the match. However, the player being in a certain location says little about the contribution of the player to the game, in particular when it comes to off-ball positioning. An intuition for this can be extracted from the distance traveled metric often used in TV broadcasts, where distances tend to vary little from one player to another, independently from their performance in the match. The question we are interested in inspecting here is: what was the value of the player when in a given location across the match? Figure (9) presents the value surface of two EPV metrics for the off-ball location of several players across the most recent FC Barcelona vs Real Madrid match in the 2018-2019 La Liga season. The first row presents the expected EPV after a pass to that player (if a pass was to be taken) and the second row the expected addition of value to the possession in case of passing to that player. We can observe different situations here. For two strikers like Suarez and Benzema we can see in the first row their tendency to go back from their forward position to be available to receive the ball and contribute with the progression of the possession. Suarez shows a tendency to back up more than Benzema. When observing the second row we can see that both players were less able to be positioned into locations that would increase the value of the possession. Moreover, when they went to the sides of the field, the expected addition of value was frequently negative. The next two columns compare Arturo Vidal and Toni Kroos. Vidal remained located compactly in locations nearby the midfield. However he was able to be consistently located in areas of positive EPV added. In the case of Toni Kroos the first row shows his pivoting skill across the field, with tendency to the left lane. However, the value added distribution, shown in the second row, presents a more disperse distribution of positive and negative value addition. Through careful consideration of context, specific situations, or coaches' specific demands for players, the off-ball value surfaces can provide helpful insights about the quality of the positioning of players. From a defensive perspective it could also provide hints for opponent team tactical scouting, in order to identify locations and situations where a player is more likely to be located to increase the likelihood of the possession ending in a goal.



**Figure 9:** The image presents the distribution of value in off-ball positioning for two EPV metrics and four different players in a FC Barcelona vs Real Madrid, La Liga season 2018-2019 match. The first row shows the expected EPV if a pass was taken towards that player (i.e. overall value), while the second rows shows the expected addition of value to the possession if that pass was taken (i.e. difference in value before/after pass).

## 6. Decision making analysis

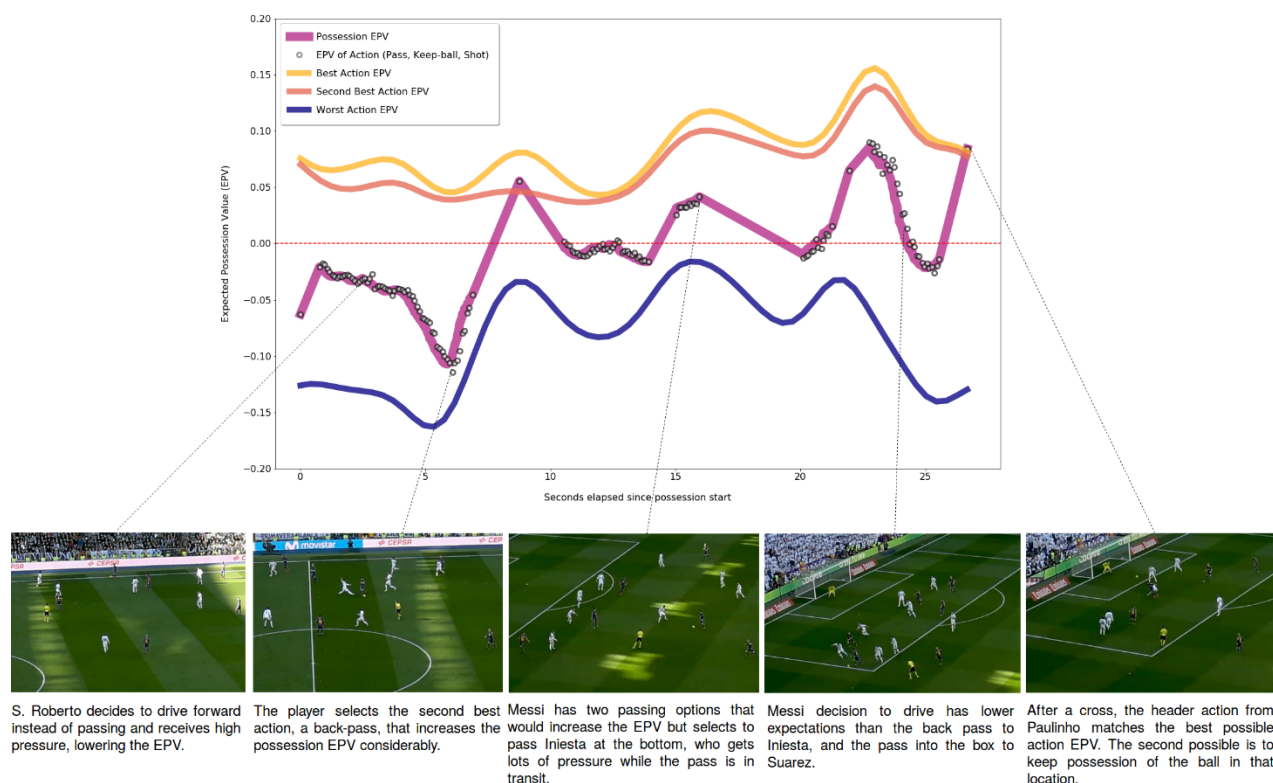
The highly detailed information about the state of a possession that this model is able to provide can be applied directly to decision-making analysis at the team or individual player level. Figure (10) presents a possession during a Real Madrid vs FC Barcelona match, where we can observe the frame by frame evolution of the possession EPV and the expected value of the best and worst possible actions at any given time. We evaluate EPV for three different types of actions: passes, shots and ball-drives, at the moment the action is beginning. Ball-drives are split into discrete actions every quarter of second. This is done for simplicity of representation, although the evaluation of EPV could be performed on every frame available. During the time between a pass or a shot being taken until it reaches its destination (the first touch or ball out) the EPV is not calculated, so we stick to the instances where the ball is in possession of one player. The yellow and red curve present the EPV for the best, second best and worst action to take. Since the model provides continuous surfaces for the passing and ball drives components, as well as for the actions likelihood, the set of possible actions is theoretically infinite. To simplify this analysis we discretize the possible set of actions by selecting the best pass location for every teammate, the location of the player in the next second according to his velocity for the ball-drive component, and the event of shooting from the current location.

Below this plot we can observe five different situations along the possession. First, when Sergi Roberto decides to produce a ball-drive, having a current EPV of 0.04, there are two other passing options available that would increase EPV up to 0.082 in the best case. He chooses to drive and then the intense pressure from three different players coming from opposite directions, starts decreasing the EPV considerably down to 0.11. In this case the model is considering that there exists a high chance that Real Madrid will regain the possession after integrating the full set of possible actions. At that point the model considers that the worst possible action to take will result in an EPV slightly lower than the current at that time, which is leveraged by the high probability of

turnover in a possible action the player could take from here. Sergi is able to select the second best action which is a pass back to Iniesta. Here we can notice a limitation of the model and tracking data:

the best option considered by the model at this time is a pass forward to Rakitic, but the data (and hence the model) lacks body orientation information to understand that the required body rotation to make that pass makes it more difficult than it seems (which can be intuitively understood from observational analysis). The back-pass to Iniesta reaches an EPV value near to the second best action suggested by the model. Afterwards, Iniesta passes to Alba, who progresses many meters through a long ball-drive and finally passes to Messi as shown in the third image. Here, Messi has two forward passes that the model recognize as the best possible, better than keeping the ball. Notice the model is trained to obtain the expected result for every action on an average player of the Premier League and Spanish League (the data used for training), but does not have individual player information. Most surely, in case of having an individualized effect for Messi keeping the ball, that model would increase the expectation of ball-drive for the player in that situation. Messi decides to pass down to Iniesta which reduces the expectation of possession, due to an intense pressure received by the player, who passes the ball back to Messi few meters into the box, increasing up to 0.082 the possession EPV. At this point, shown by the fourth image, Messi has several high value passes inside the box but chooses to drive and gets pushed outside of the box and reduces its passing, shooting and driving opportunities, which, again, drops the possession value, reaching a negative value of -0.034. Messi chooses to cross the ball and Paulinho receives an aerial pass and takes a header to goal as shown in the last image. Here, the best and second best action coincide with the possession EPV. These actions are shooting or keeping the ball in that location.

From this frame by frame evaluation of a possession we can grasp a piece of the interpretation power of the model for decision making. Beyond specific possessions, there are a wide set of metrics that can be derived from this information with simple calculations. One is to identify actions of players that increased or decreased the possession EPV above or below a threshold, rapidly segmenting actions of value. These actions can be filtered according to the action type, field location, opponent relative location, and any other contextual variable that can be associated from the frame by frame evaluation of the possession. From here, is also possible to highlight the kind of actions that are better exploited by the different teams in a competition. The model would also allow one to quickly identify situations where there was a high risk of losing the ball and the opponent producing danger. We can then study the distribution of decisions made in these different situations. Regarding individual player decision-making evaluation, we could identify not only the total value added or lost by players, but also what kind of actions and situations produce that addition or loss. Another possible application of detailed EPV curves during a possession is the identification of *possession plateaus*, which we define as windows of time where the possession is stabilized in a certain moment of the possession without variation, beyond the many actions that could take place in between. For example, in soccer it is typical that in the finalization stage the ball gets passed from one side to another without being able to overcome an additional line or shoot, nor being forced to move back to progression or build-phase stages. After identifying these plateaus, it becomes straightforward to extract actions that produces the jump between one plateau to another.



**Figure 10:** Image on top presents the evolution of EPV in time for a possession sequence in a Real Madrid vs FC Barcelona match in the 2017-2018 La Liga season. Four different curves are presented showing the possession EPV, and the EPV for the best, second best and worst action to take according to the model. Below, the five images represent snapshots of the match video in different situations. The set of possible actions considered includes passes, shots and ball-drives.

## Discussion

We have presented a model that is able to provide a frame by frame quantification of the expected value of any soccer possession (EPV). Furthermore, the model captures a wide set of positional, motion and contextual features, and the relative value of any location on the field, addressing a key feature of this sport: understanding player interaction in space. Beyond the quantitative advantages of a comprehensive metric such as EPV, the decoupled nature of the model allows for an in-depth visual and quantitative interpretation of any situation, making it possible to look into the black-box of complex machine learning models, while taking advantage of their generalization power. This EPV framework for soccer opens the door for a wide set of applications in soccer analytics, covering player-level and team-level performance evaluation, detailed decision-making analysis, effective querying of value actions, evaluation of off-ball contribution, and many others.



## Acknowledgements

The authors would like to thank *STATS LLC* and *Footovision* for providing us with optical tracking data, as well as the Sports Science and Health department and the Soccer Analysis Team of FC Barcelona for their valuable input and advice throughout this project.

## References

- [1] Daniel Cervone, Alex D'Amour, Luke Bornn, and Kirk Goldsberry. A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, 111(514):585–599, 2016.
- [2] Daniel Cervone, Alexander D'Amour, Luke Bornn, and Kirk Goldsberry. "POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data." In *Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA*, vol. 28, p. 3. 2014.
- [3] Paul Power, Hector Ruiz, Xinyu Wei, and Patrick Lucey. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In *Proceedings of the 23<sup>rd</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1605–1613. ACM, 2017.
- [4] Daniel Link, Steffen Lang, and Philipp Seidenschwarz. Real time quantification of dangerousity in football using spatiotemporal tracking data. *PloS one*, 11(12):e0168768, 2016.
- [5] William Spearman. Beyond expected goals. MIT Sloan Sports Analytics Conference, 2018.
- [6] Matthew Heiner, Gilbert W. Fellingham, and Camille Thomas, Skill importance in women's soccer. *Journal of Quantitative Analysis in Sports*, 10(2), pages 287-302, 2014.
- [7] Sarah Rudd. A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer using Markov Chains. In *New England Symposium on Statistics in Sports*. URL <http://nessis.org/nessis11/rudd.pdf>. 2011.
- [8] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions Speak Louder Than Goals: Valuing Player Actions in Soccer. *arXiv preprint arXiv:1802.07127*. 2018.
- [9] Javier Fernandez and Luke Bornn. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *Sloan Sports Analytics Conference*, 2018.
- [10] Uwe Dick and Ulf Brefeld Learning to Rate Player Positioning in Soccer. *Big data*. 2019.

- [11] William Spearman, Austin Basye, Greg Dick, Ryan Hotovy, and Paul Pop. Physics-based modeling of pass probabilities in soccer. In *Proceeding of the 11th MIT Sloan Sports Analytics Conference*, 2017.
- [12] Tom Decroos, Vladimir Dzyuba, Jan Van Haaren, and Jesse Davis. Predicting soccer highlights from spatio-temporal match event streams. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 4-9 February 2017*, pages 1302–1308, February 2017.
- [13] Xinyu Wei, Long Sha, Patrick Lucey, Stuart Morgan, and Sridha Sridharan. Large-scale analysis of formations in soccer. In *Digital Image Computing: Techniques and Applications (DICTA), 2013 International Conference on*, pages 1–8. IEEE, 2013.