# Dynamic analysis of team strategy in professional football

Laurie Shaw[1,2], Mark Glickman[2,3]

[1]Harvard Data Science Initiative, Harvard University
[2]Harvard Sports Analytics Lab, Harvard University
[3]Department of Statistics, Harvard University

## Abstract

One of the most important tactical decisions that a football manager must make is to determine the spatial configuration of the team, or formation, during different phases of a match. The selection of formations influences how aggressively a team plays, where they focus their attacks, and their overall playing style. We present an innovative new technique for dynamically measuring, classifying and studying team formations in professional football matches. Using a large sample of player tracking data, we measure the relative positioning of each team's players in and out of possession of the ball over successive time intervals during each match. Applying hierarchical agglomerative clustering – using the Wasserstein metric to measure distances between formations – we have identified the unique set of offensive and defensive formations that teams deployed. We use these formation templates, in combination with Bayesian model selection criteria, to classify new formation observations, producing tactical summaries of each match. We identify each team's preferred offensive and defensive formations, and study how managers reacted tactically to key events during their matches. Finally, we discuss how formation choices relate to playing style, and discuss other potential applications of our methodology.

*Keywords: formations, tracking data, hierarchical clustering, Bayesian model selection*

## 1. Introduction

A vital aspect of a football manager's job is to select team formations – the spatial configuration of the players on the field. The choice of formation determines player roles, how they interact, and influences the playing style of both teams during a match. Despite their central role in team strategy, descriptions of formations are largely reliant on classifications based on the number of defenders, midfielders and forwards: crude summaries of player configurations that are significantly more fluid, nuanced and dependent on the game state than '4-4-2' or '3-5-2' would suggest. Modern managers frequently refer to the necessity of using different formations for different phases of the game, and the need to adapt to specific circumstances.

Comprehensive quantitative analysis of team formations in professional football has been inhibited by the difficulty of obtaining access to large samples of player tracking data. Previous studies [1,2,3,4] have typically assumed that formations remain static and unchanged throughout the course of a match, an approximation that loses much valuable information and precludes analysis of how in-match tactical changes affect the outcome.

In this paper we present a new, data-driven technique for measuring and classifying team formations as a function of game state, analysing the offensive and defensive configurations of each team separately, and dynamically detecting major tactical changes during the course of a match. We apply our methodology to a large sample of player tracking data, using unsupervised machine learning techniques to identify the unique set of template formations used by the teams in the dataset. We classify individual formation observations in a larger sample of matches according to these templates to study transitions between defence and attack, and analyse changes in formation during matches. Finally, we discuss the various practical applications of our methodology.

## 2. Methodology

There are three key stages to our methodology. First, we present a new algorithm for measuring team formations as a function of time during a match by averaging vectors between neighbouring players in local possession windows. Second, we identify the unique offensive and defensive formations used by the teams in a large training set of tracking data through agglomerative hierarchical clustering. Finally, we incorporate the set of identified formation clusters into a Bayesian model selection algorithm to dynamically classify formation observations and systematically detect formation changes during matches.

The tracking data used in this analysis consists of 180 matches from a single season of an elite professional league. The data for each match consists of the positions of all 22 players and the ball, sampled at a frequency of 25Hz. Individual player identities are tagged in the data, enabling tracking of each player over time.

### 2.1 Measuring team formations

It is well known that the outfield players in a team will tend to encompass only a small fraction of the pitch at any given instant, with the players moving coherently as a group to maintain their spatial configuration. Team formations are therefore defined by the relative positions of the players.

Figure 1 indicates the positions of the defending team (i.e. the team out of possession of the ball) at four instants during the first half of a match. It is clear that, while the team occupies different areas of the pitch at each instant, the players largely retain their relative positioning, maintaining a 4-3-3 formation (four defenders, three central midfielders and three forwards).

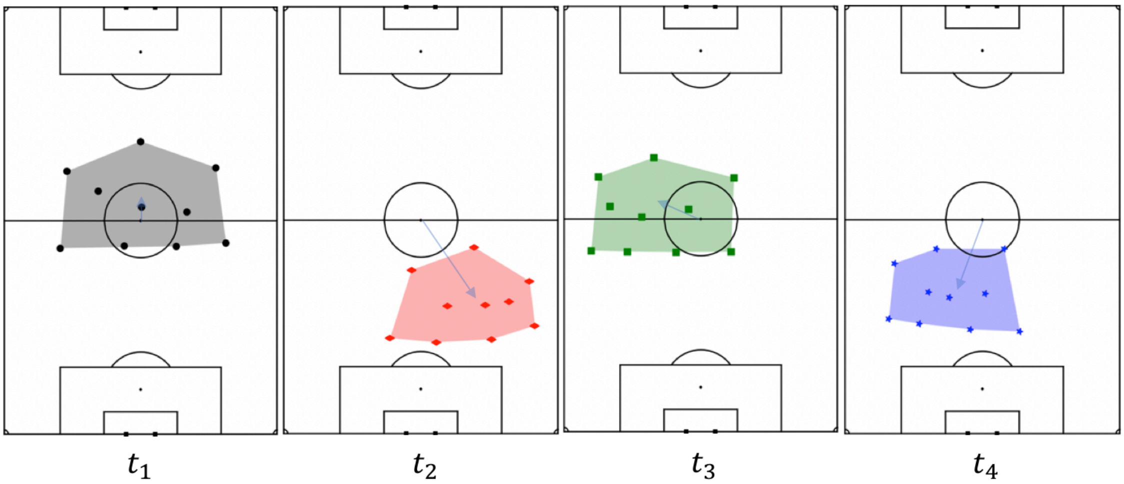

**Figure 1 –** The positions of the outfield players of the defending team at four instants of time during a match. The shaded regions indicate the convex hull; the blue arrow indicates the centre of mass of the team relative to the centre of the pitch.

Formations are measured by calculating the vectors between each player and the rest of his teammates at successive instants during a match, averaging the vectors between each pair of players over a specified time interval to gain a clear measure of their designated relative positions. The final spatial distribution of the outfield players is determined by the following algorithm: first, we set the centroid of the formation to be the position of the player in the densest part of the team, as determined by the average distance to the third-nearest neighbour. We then identify the relative position of his nearest neighbour, the relative position of that player's nearest neighbour (ignoring any player already considered in the process) and so on, until the positions of all players in the team have been determined.

The advantage of this pairwise approach to measuring team formations (as opposed to, for example, measuring the average positions of each player in the centre of mass frame of the team, as in previous studies [2]) is that the location of a player in a formation is dictated solely by his position relative to his neighbouring teammates. A calculation of the team centre of mass at each instant would result in the formation positions being influenced by every other player on the team.

Defensive and offensive formation observations are measured separately by aggregating together consecutive possessions of the ball for each team into two-minute, non-contiguous time periods. We exclude possessions that last for less than five seconds from this process under the assumption that they are too short for either team to establish an offensive or defensive stance. Furthermore, if a substitution occurs – which may potentially be accompanied by a formation change – we end the window, retaining it in our analysis if it contains at least one minute of in-play data. Within each window we measure the formations of both the team in possession and their opponent. On average, we obtain ten defensive (i.e., out-of-possession) formation observations and ten offensive (in-possession) formation observations for each team during a match. Figure 2 presents four examples of individual formation observations.

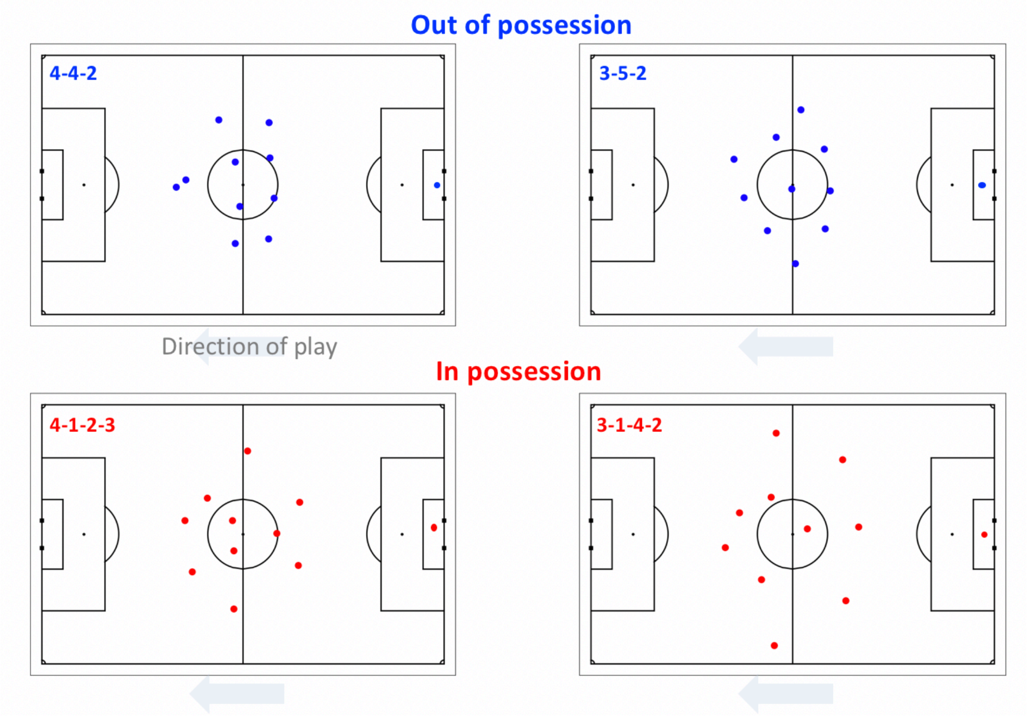

**Figure 2**: Four examples of formation observations, each measured in a 2-minute aggregated possession window. The top two panels show defensive formation observations (out-of-possession); the bottom two panels show offensive formation observations (in-possession).

Figure 3 plots the full set of formation observations for one team during a single match. It is clear that, when out of possession (upper plot), the team played with a 4-1-4-1 formation, with a single defensive central midfielder and a lone striker. When in possession (lower plot), the outside midfielders advanced to form a front three and the full backs moved level with the defensive midfielder. The right central midfielder played slightly deeper than the left central midfielder, introducing a small asymmetry to the team when attacking. While the relative positions of the defensive players in the team are clearly well constrained, the position of the offensive players – particularly the central striker – is much more broadly distributed, both in and out of possession. More generally, the area encompassed by the outfield players (the convex hull) when attacking was twice the area encompassed when it was defending. The consistency of the observations indicates that the manager did not make a significant formation change during the match.

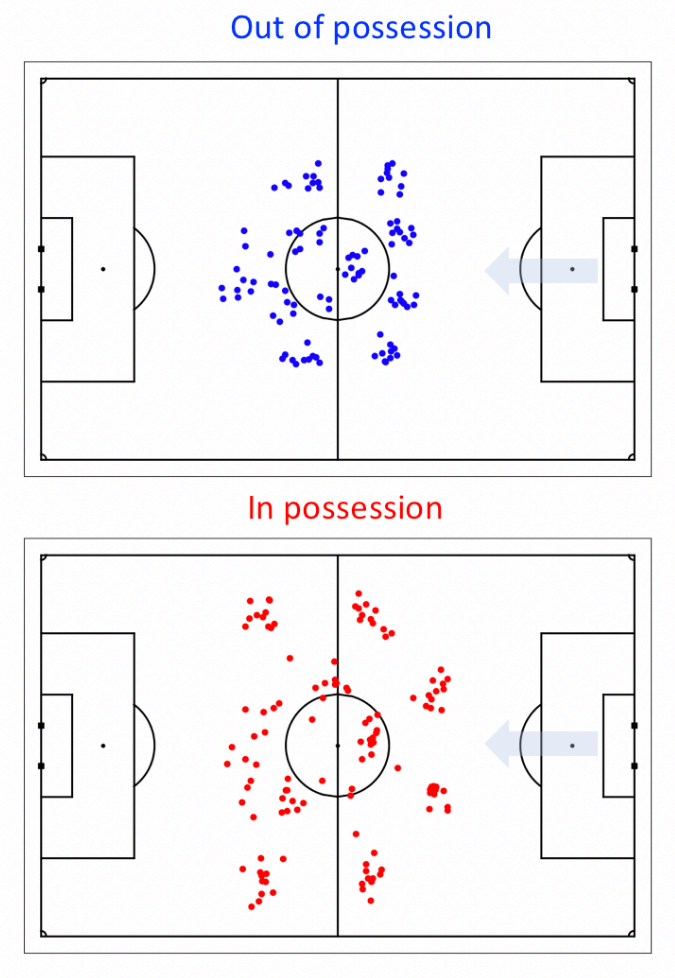

**Figure 3:** The full set of formation observations for one team throughout an entire match. The upper plot indicates the defensive formation observations, the lower plot indicates the offensive formation observations; in both cases, the team is shooting from right to left. The consistency of the observations indicates that the team did not undergo a significant formation change during the match.

## 2.2 Identifying unique formations

We have applied the methodology described above to tracking data from a training sample of 100 matches, obtaining 3976 observations of offensive and defensive formations. The remaining 80 matches were used as a test set for validation. In this section we describe the application of agglomerative hierarchical clustering to group similar observations, thereby identifying the set of unique formation types adopted by the teams during these matches.

A key element of this process was to define a metric for quantifying the similarity of two formation observations. In our method, each observation is effectively a set of 10 bivariate normal distributions – one for each outfield player – in which the mean of each distribution is the position of a player in the formation (remembering that the formations are translated so that the centres of mass coincide), and the covariance matrix is an estimation of how far the player deviated from his position during the two minute possession window in which the formation was measured.

We utilize the Wasserstein distance [5] to quantify the similarity of two formation observations. In the simple case of two bivariate normal distributions, $\mu_1 = N(m_1, C_1)$ and $\mu_2 = N(m_2, C_2)$, where $m$ is the mean and $C$ is the covariance matrix, the square of the Wasserstein distance is given by [6]:

$$W(\mu_1, \mu_2)^2 = ||m_1 - m_2||^2 + \text{trace}\left(C_1 + C_2 - 2\left(C_2^{1/2} C_1 C_2^{1/2}\right)^{1/2}\right).$$

In the case of point particles the Wasserstein distance is simply the square root of the $L_2$ norm of the difference between the means. More generally, the Wasserstein metric is a solution to the optimal transport problem [7], i.e., an estimate of the cost of moving from one distribution to another.

The second step of our algorithm is to find a pairing of the players in the two formation observations that minimizes the square of the sum of the Wasserstein distances, i.e.

$$W_{total}^2 = \; min \sum_i \sum_j D_{ij} X_{ij} \; ,$$

where $D_{ij}$ is the cost (square of Wasserstein distance) of matching player *i* in formation 1 to player $j$ in formation 2, and $X_{ij}$ is a player-player allocation matrix, in which each element is equal to 1 if player *i* is matched to player $j$, and zero otherwise. Each row and column in $X_{ij}$ must therefore uniquely consist of nine 0s and a single 1. We use the Kuhn-Munkres algorithm [8,9] to find the $X_{ij}$ that minimises the total cost.[1]

We make one further extension to our metric for team similarity. Two formation observations may be identical in terms of their shape (e.g. a traditional 4-4-2), but one may be a more compact or expanded incidence of the other. As we aim to identify distinct formation shapes, we introduce a variable scaling factor, *k,* that expands or contracts a formation around its centre of mass (scaling the player covariances accordingly). When comparing two formation observations, we search for the value of *k* that minimises the Wasserstein distance between them.

We apply agglomerative hierarchical clustering to the formation observations measured from our training sample of matches, using the Ward metric as a linkage criterion [10]. This identified 20 unique formation templates, or clusters, used by the teams in our training sample. The results are shown in Figure 4.

[1] Note that we remove matches in which a player was sent off from our analysis.

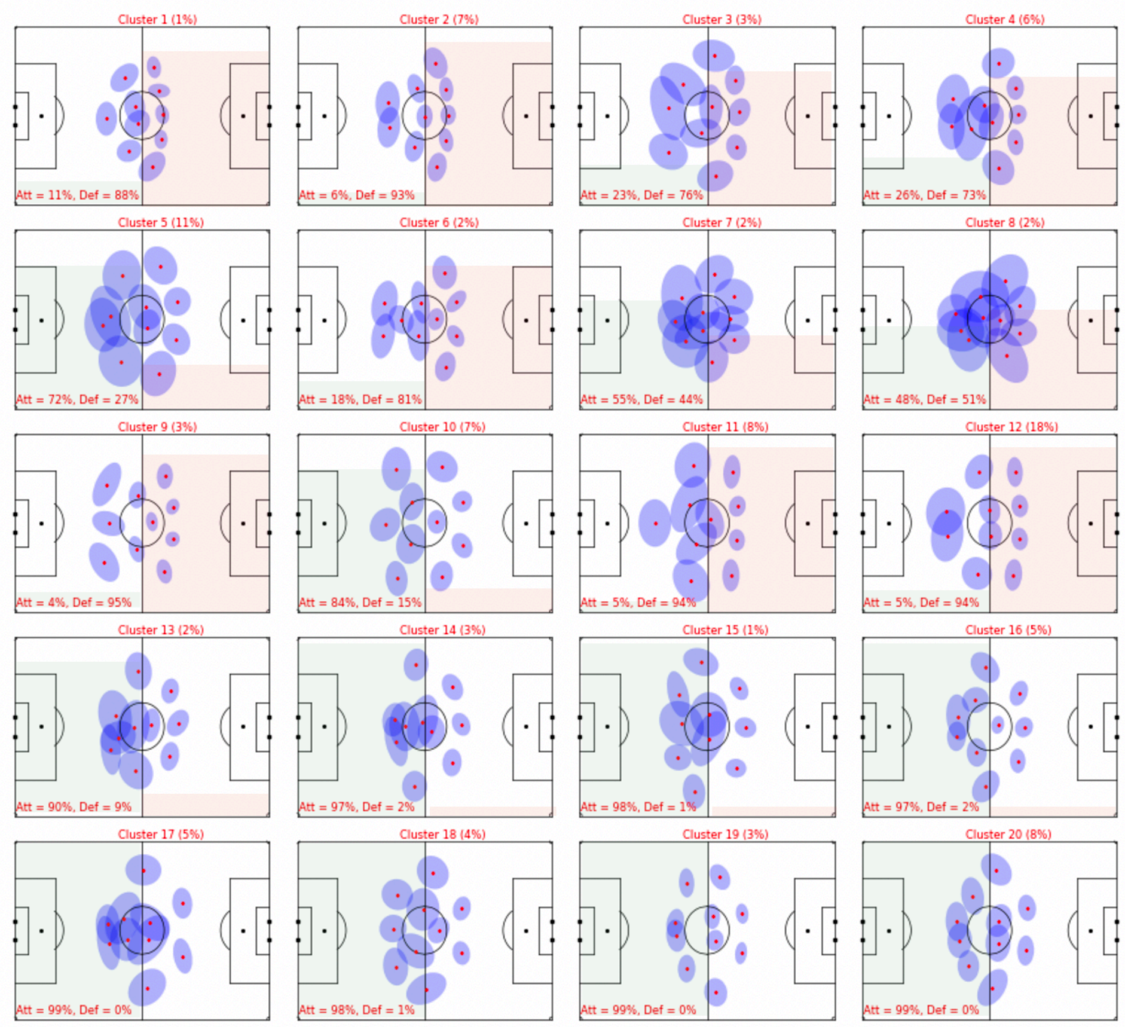

**Figure 4:** The 20 unique formation clusters identified using hierarchical clustering based on a training sample of formations measured in 100 professional matches. Teams are orientated to shoot from right to left, and formations are translated to align their centre of mass with the centre of the pitch. Ellipses indicate the 1-sigma region (68% confidence interval) for the positions of each player, measured over the individual observations in each cluster. The text in the bottom left of each panel indicates the proportion of offensive and defensive formation observations in the cluster (also indicated by the green and red bars).

There is a clear ordering to the clusters that highlights the difference between defensive and offensive formations – a distinction lost in previous analyses of formations in football. The top row in Figure 4 contains formation clusters with five defenders and variations in the number of midfielders and forwards; these clusters predominantly consist of defensive formation observations. The following two rows indicate variants of a back four: cluster 6 is clearly a midfield diamond, clusters 9 and 10 are variants of a 4-3-3 formation, cluster 11 is a 4-1-4-1 and cluster 12 is a 4-4-2. The clusters in these rows contain a mix of attacking and defensive formation observations. For instance, cluster 9 predominantly consists of defensive formation observations, while cluster 10 is mostly made up of offensive observations.

The fourth and fifth rows contain clusters that almost entirely consist of offensive formation observations. The fourth row contains variants of the 3-4-3 and 3-5-2 formations, although the standard nomenclature is a crude description of these formations. The fifth row shows clusters that have essentially just two defensive players – in all four cases the full-back positions have advanced significantly.

Overall, it is clear that the hierarchical clustering has efficiently separated observations of defensive and offensive formations, even though it could not use the differences in their scale size (or area encompassed) as a discriminator because of our application of the scaling factor, *k*.

### 2.3 Formation classification

The final step of our methodology is a Bayesian model selection algorithm to estimate the probability that a newly observed formation belongs to each of the 20 formation clusters shown in Figure 4. This probability is calculated as

$$\mathrm{p(o|C)} \approx \underset{k}{argmax} \prod_{p=1}^{10} \int \mathrm{p}(y|\, k\mu_{\mathrm{p,C}},\, k^2\Sigma_{\mathrm{p,C}})\mathrm{p}(y|\mu_{\mathrm{p,o}},\, \Sigma_{\mathrm{p,o}})\mathrm{d}y$$

where $\mu_{\mathrm{p,C}}$ and $\Sigma_{\mathrm{p,C}}$ are the position and covariance matrix for role *p* in cluster C, $\mu_{\mathrm{p,o}}$ and $\Sigma_{\mathrm{p,o}}$ are the position and covariance matrix for player *p* in the formation observation o, *k* is the scaling factor described in Section 2.2., and the integral is performed over the surface area of the pitch. To assign each player in a formation observation to a specific role in a cluster, we solve the player-role allocation problem using the Kuhn-Munkres algorithm, as described in the previous section.

Identifying the maximum probability cluster for each formation observation enables us to classify formation observations throughout a match to dynamically detect tactical changes.

## 3. Results and analysis

We now expand our formation detection and classification scheme to the full sample of 180 matches and present some of our key results and observations.

### 3.1 Transitions

We first investigate transitions between defence and offence by identifying the defensive and offensive formation clusters that are most frequently paired together by the teams in our dataset. In Figure 5 we plot an example of these pairings using a Sankey diagram. The left-hand side of the diagram corresponds to defensive formation clusters, while the right-hand side corresponds to offensive formation clusters. The links between them indicate the formations that were typically employed together as teams gained and lost possession.

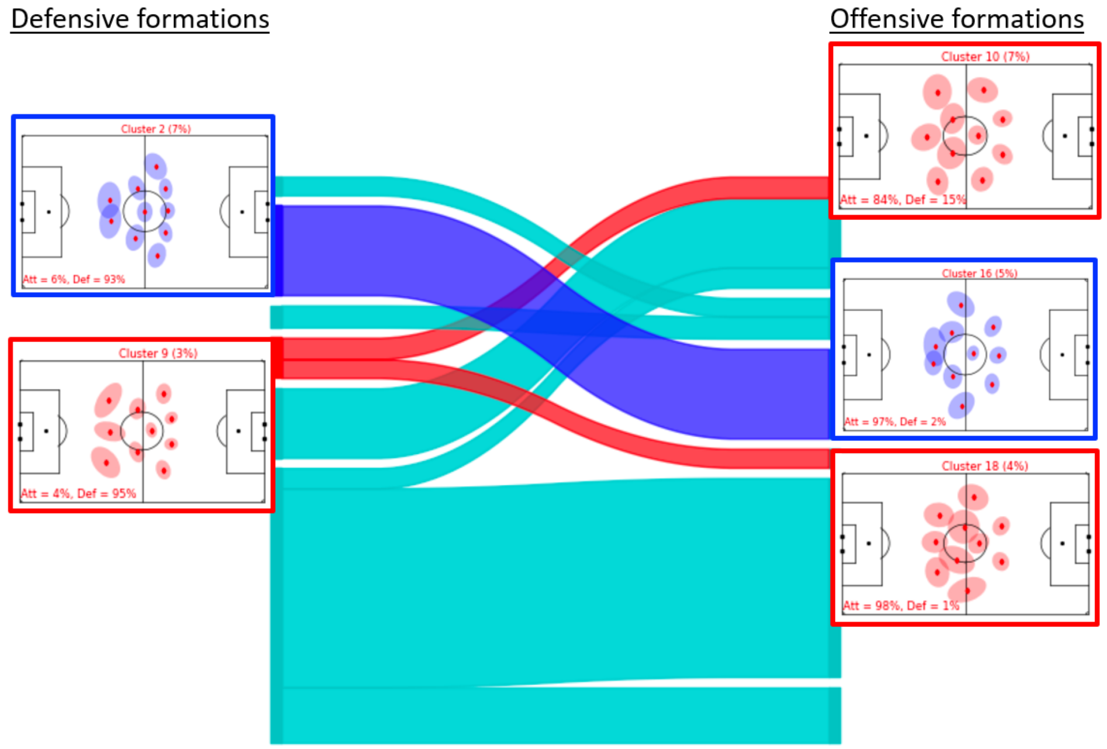

**Figure 5:** Two examples of the typical pairings between defensive and offensive formations. The blue formations indicate that teams playing with a defensive formation drawn from cluster 2 (see Figure 4) transition to an offensive formation drawn from cluster 16. The red example indicates that teams that play with defensive formation 9 transition to either offensive formations 10 or 18. All teams are orientated to shoot from right to left.

The example highlighted in blue indicates that teams in our sample that defended using cluster 2 (as defined in Figure 4) transitioned to cluster 16 when in possession of the ball. The connection between the two formations is clear: the outside defenders, or wingbacks, advance when the team gains possession and the two outside midfielders tuck in behind the two forwards.

The second example, highlighted in red, demonstrates that teams using cluster 9 (a 4-3-3) when defending would transition into either cluster 10 or cluster 18 when attacking – two formations that are significantly different. In cluster 10, the outside forwards have pushed wide and the full-backs have advanced, whereas in cluster 18 the front three remain narrow with the full-backs advancing further up the field to provide width.

There are two main conclusions to draw from these examples. First, the defensive and offensive formation pairings are consistent: it is clear how each player's defensive and offensive roles are related. This provides an important validation of our methodology. Second, it demonstrates that some defensive configurations provide more flexibility in terms of different attacking options than others.

## 3.2 Strategic summaries and changes in formations

Dynamic measurement and classification of formations enable us to produce strategic summaries of matches that communicates the defensive and offensive configurations of each team and detects when major tactical changes occurred.

Figure 6 charts the defensive and offensive formations during a match between two teams – labelled the *Red team* and the *Blue team* – throughout the course of a match. The circles indicate the offensive formation observations of each team, classified according to the clusters shown in Figure 4; the diamonds indicate the defensive formations. Goals are indicated by a vertical dashed line at the top of the plot; substitutions are indicated by a vertical dashed line along the bottom of the plot.

In this match, the Red team were losing 1-0 at half time. The chart indicates that the manager made a substitution and a significant change in formation, switching from a 3-4-3 formation (clusters 3 and 15 in defensive and attack, respectively) to a 4-3-3 (clusters 9 and 10). They scored shortly after half time, but ultimately lost the match 2-1.

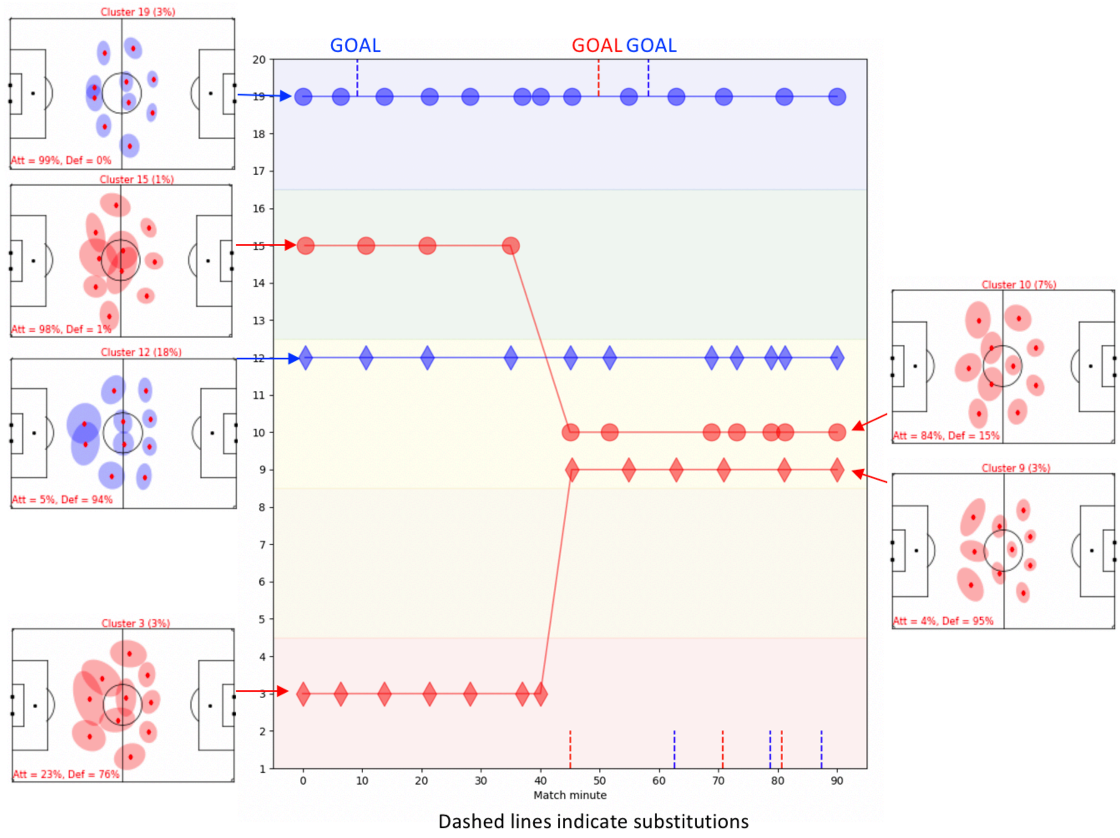

**Figure 6:** Strategic summary of a match between the *Red* and *Blue* teams. Diamonds indicate defensive formations; circles indicate offensive formations. Y-axis labels correspond to the cluster numbers in Figure 4.

Automated detection of formation changes, combined with event data, enable us to investigate *why* certain tactical changes were made and evaluate the impact they had on the outcome of a match. Figure 7 provides a simple example[2]. The right-hand panels of the plot indicate the

[2] Note that Figures 6 and 7 depict different matches.

defensive formation observations of the Red team in the first and second half. The left-hand panels show pass and shot maps of the opposing team (shooting from left to right); arrows indicate individual passes and dots denote shots, with the symbol size indicating the quality of the opportunity.

In the first half, the Red team played with a 4-3-3 in defence. The pass map of the Blue team indicates that they tended to attack down the flanks in the first half, creating high-quality chances from crosses, particularly from the right wing. At half time the Red team switched to a 5-man defence, with the wing-backs marking the opposing wingers. As the pass map for the second half indicates, the change in formation appears to have been effective in preventing the Blue team creating chances from their right side, with the focus of their passing switching more towards the centre and left of the pitch.

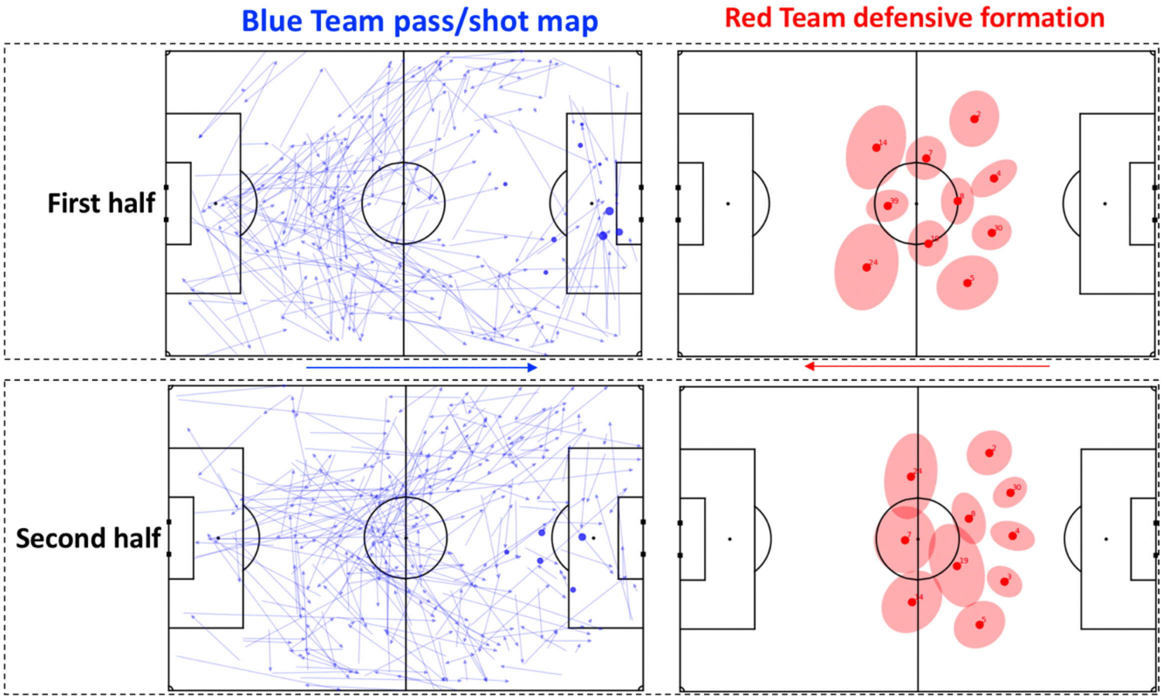

**Figure 7** *Right hand plots*: observations of the defensive formation of the Red team (playing from right to left) before and after half time in a match against the Blue team. *Left hand plots*: passes (arrows) and shots (circles) of the blue team (playing from left to right) in the first and second half of the match. The sizes of the circles indicate the quality of the shooting opportunity. Note that the match depicted is different to the match shown in Figure 6.

## 4. Practical applications

Our analysis is a key step towards the use of tracking data to infer and evaluate team strategy in football. The methodology outlined above enables teams to study how an opposing manager habitually responds to specific match situations. For instance, the manager of the Red Team in Figure 6 made similar formation changes at (or near to) half time in over a quarter of their matches in our dataset, switching between a small subset of formations based on the quality of the opposition and the state of the match. Our methodology provides to the tools to anticipate, and therefore exploit, opposition tactical changes.

Second, our methodology enables us to study in detail the factors that cause the defensive formation of a team to become disrupted and investigate how this relates to chance creation. Combining formation classification with pitch control surfaces [11,12,13] enables us to identify potential defensive weaknesses of specific formations and determine how teams might exploit them.

Finally, our methodology can be extended to consider formations in more specific phases of possessions, such as transition, establishing possession, progression and chance creation, and to incorporate player velocity information to identify and understand marking systems and the operation of a high press.

## 5. References

[1] Bialkowski A, Lucey P, Carr P, Yue Y, Matthews I (2014a) Win at Home and Draw Away: automatic formation analysis highlighting the differences in home and away team behaviors *MIT Sloan Sports Analytics Conference*. Boston

[2] Bialkowski A, Lucey P, Carr P, Yue Y, Sridharan S, Matthews I (2014b) Large-scale analysis of soccer matches using spatiotemporal tracking data. In: *2014 IEEE international conference on paper presented at the data mining (ICDM)*. 14–17 Dec 2014

[3] P. Lucey, A. Bialkowski, P. Carr, S. Morgan, I. Matthews, and Y. Sheikh, “Representing and Discovering Adversarial Team Behaviors using Player Roles,” in *CVPR*, 2013.

[4] X. Wei, L. Sha, P. Lucey, S. Morgan, and S. Sridharan, “Large-Scale Analysis of Formations in Soccer,” in *DICTA*, 2013.

[5] Ramdas, Garcia, Cuturi “On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests” (2015). *arXiv:1509.02237.*

[6] Olkin, I. and Pukelsheim, F. (1982). "The distance between two random vectors with given dispersion matrices". *Linear Algebra Appl.* 48: 257–263. doi:10.1016/0024-3795(82)90112-4. ISSN 0024-3795.

[7] Cédric Villani (2003). Topics in Optimal Transportation. *American Mathematical Soc*. p. 66. ISBN 978-0-8218-3312-4.

[8] Harold W. Kuhn. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83-97, 1955.

[9] Munkres, J. Algorithms for the Assignment and Transportation Problems. *J. SIAM*, 5(1):32-38, March, 1957.

[10] Ward, J. H., Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association*, 58, 236–244.

[11] Fernandez, J. (2019), Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer, *Sloan Sports Analytics Conference*, Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2019/02/Decomposing-the-Immeasurable-Sport.pdf

[12] Fernandez, J (2018), Wide Open Spaces: A statistical technique for measuring space creation in professional soccer, *Sloan Sports Analytics Conference*, Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2018/03/1003.pdf

[13] Spearman, W (2018), Beyond Expected Goals, *Sloan Sports Analytics Conference*, Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2018/02/2002.pdf