Well Formatted: Understanding Team Behavior Through Formation Analysis

Master's Thesis

submitted in partial fulfillment of the requirements for the degree of *Master in Social and Economic Data Science* at the Department of Economics at the University of Konstanz

presented by

Marc Lüttecke

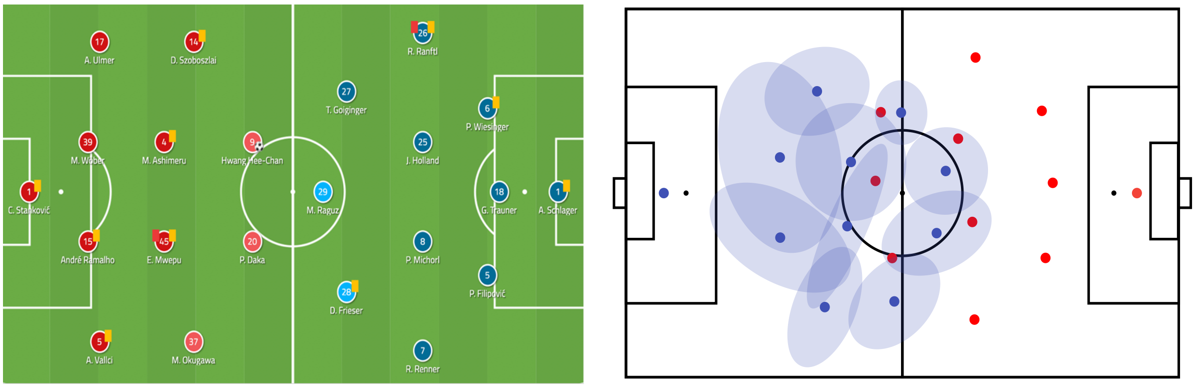
at the

Universität
Konstanz

Faculty of Sciences
Department of Computer and Information Science

First assessor and supervisor:	Prof. Dr. Daniel A. Keim, Universität Konstanz
Second assessor:	Prof. Dr. Michael Grossniklaus, Universität Konstanz

Konstanz, 05.03.2021



(a) Existing manual formation description

(b) Proposed automatic visualization

Figure 1: This thesis develops a novel system for formation analysis. It integrates multi-match as well as inter-match exploration of a large data set leveraging intuitive design elements to subset the data via drop-down and hover effects to offer coaches a practical as well as comfortable and correct user experience. Its automatic formation calculation offers faster and more accurate results than any existing alternative.

Abstract

A thorough understanding of the dynamics of collective movement patterns forms the foundation of success in a diverse cross-section of competitive team sports. Nonetheless, the field of soccer formation research fails to adequately serve the practitioners' requirements by employing analyses that provide only slow, impractical, or anecdotal solutions. This master's thesis addresses this discrepancy by developing a system that analyzes an extensive dataset of 250 soccer games in an intuitive app design. The fast algorithmic logic and prediction accuracy achieved in this thesis beat all current benchmarks, while the functional layout follows strict design requirements developed in iterative feedback sessions with qualified domain specialists. The application withstands the scrutiny of extensive quantitative and qualitative assessments during detailed validation rounds with unbiased experts.

Zusammenfassung

Ein tiefgreifendes Verständnis der Dynamik von kollektiven Bewegungsmustern bildet die Grundlage für den Erfolg in einem breiten Querschnitt von kompetitiven Mannschaftssportarten. Dennoch wird das Feld der Fußball-Formationsforschung den Anforderungen der Anwender nicht gerecht, indem Analysen verwendet werden, die nur langsame, unpraktische oder anekdotische Lösungen liefern. Die vorliegende Masterarbeit widmet sich dieser Diskrepanz durch die Entwicklung eines Systems, das einen umfangreichen Datensatz von 250 Fußballspielen in einem intuitiven App-Design analysiert. Die in dieser Arbeit erreichte schnelle Algorithmenlogik und Vorhersagegenauigkeit übertrifft alle aktuellen Vergleichswerte, während das funktionale Layout strengen Designvorgaben folgt, die in iterativen Feedbacksitzungen mit qualifizierten Fachexperten entwickelt wurden. Die Anwendung hält der genauen Prüfung umfangreicher quantitativer und qualitativer Bewertungen während detaillierter Validierungsrunden mit unvoreingenommenen Spezialisten stand.

Contents

List of Figures

List of Tables

1 Introduction

> *I'm constantly being asked about individuals. The only way to win is as a team. Football is not about one or two or three star players.*
>
> Edson Arantes do Nascimento (Pelé), *FIFA World Cup*, 1962

The success of a competitive sports team emerges from the synergistic contribution of each teammate. Accordingly, a team's performance is critically dependent upon each player's individual actions and the complex spatio-temporal dynamics by which all players combine to achieve a collective objective. A comprehensive understanding of these dynamics and their contribution to a team's success could help teams gain a competitive edge. In this 471 billion-dollar industry [3], any advantage can translate to large increases in profit. However, several diverse research fields[1] have struggled to identify, evaluate, and communicate the ingredients of successful collective behavior. Equipped with recent developments in data extraction methods, a vast dataset, and exponentially increasing computer power, we are now uniquely positioned to utilize data to offer insights via an intuitive system for multi-match analysis of collective movement patterns.

While some fields, such as economics, have relied upon automated quantification of performance for several decades, this kind of rigorous analysis remains relatively new in sports. Combining available statistics and computer science methods with those from sport and health sciences has generated a recent outburst of creative solutions to automate existing reasoning and measure success. One major obstacle is to locate athletes and translate their movements to a computer-readable format. The CATAPULT system [1] attempts to address this problem via GPS tracking systems, while alternative approaches rely upon color discrimination[2] in team sports to locate athletes [50, 64, 72]. With only limited usability of these systems due to impracticality or inaccurate results, efforts shifted towards extracting information from the raw video as the most natural use-case. Developments in the field of computer vision have enabled the detection of human movements in video data [25, 26, 33, 34, 40, 47, 86]. Leveraging these advancements, researchers have started to incorporate systems to extract 2D (x-/y-coordinates) and 3D (skeletal) data directly from video streams into sports research [16, 53, 57, 86].

While these new data can contribute to our comprehension of several sports, soccer became one of its most popular and earliest adopters. Over time, soccer clubs have recognized the potential benefit of extracting and analyzing previous matches' data to improve team performance. This adoption has led most professional sports teams to employ personnel as *sports analysts*. These professionals aim to conclude data that could guide the future behavior and performance of a team. Despite this ambitious objective, the data- and, consequently, the

[1] These fields include cognitive sciences, behavioral psychology, health sciences, physiology, physics, biology, computer science, and data science.

[2] **Color discrimination techniques** classify agents by different color pattern. This approach is most appropriate for scenarios where the assigned colors are clearly defined, such as sports or controlled experiments. Problems arise if the surrounding environment displays similar colors as the classification characteristic (such as advertisements or sports banners).

analysis-quality often lacks the statistical rigor and the usability to provide significant pragmatic insights. These limitations have generally led soccer organizations to continue to rely on more qualitative, non-technological approaches to improve team performance. This reluctance to adopt recent technological advances discounts the valuable insights available from newly available data and the novel approaches to extract relevant information from these data. In addition to the time-consuming data acquisition, most current system solutions address only individual player movements with little regard for team-interactions. Recent research projects are underway to address the shortcomings of current best-practices of data analysis in sports with topics of *automatic event annotation* [46, 66, 65], *indication of potential passing stations* [67, 90], and *what-if analyses* [74] for crucial match situations. However, these papers mainly build on the additive actions of individuals, not fully incorporating analyses of their collective behavior.

Advanced data systems might enable automatic extraction of simple fitness and match statistics, such as possession distribution, running distance, or duel statistics. However, while these insights enable the calculation of widely-used KPIs[3] for players' performance, they ignore a core principle of team-sports: individuals do not win games, teams do. Therefore, it is imperative to consider interactions among all players moving collectively to understand match dynamics.

Formations – a quantitative description of all players' locations on the field – provide the most accurate characterization of how players move collectively as a unit. They represent the most prominent indicator (see *A1* in Chapter 3) of a team's overall strategic match plan. A head coach leads a time, while his role lets him provide external insights and shape the team's overall strategic approach. His perspective from outside the collective affords him the ability to design a match plan for movement behavior and tactical decision-making. A crucial component to this match layout is identifying the most appropriate formation to address specific match situations. While he cannot interact directly on the field, determining a formation proves to be one of the most flexible adjustments he can make to impact the team's collective behavior. Therefore, a variety of factors determines the arrangement in a specific formation. All these formation data points are manually collected and then processed to form the coach's solutions and proactive decisions, given the opposing players' skill sets, the own teams' skills, and the formational movements. To date, no single system exists to address these practical questions of best responses and facilitate the coach's job to find the best solutions for a given match. This shortcoming stems from three primary sources: first, the player position data was not as readily available until recently; second, the audience is predominantly non-technical and slow to adopt rigorous analytical methods; and third, defining discrete formations has proven difficult. Seemingly straightforward questions, such as how robust a formation is over time, if the relative position on the field is essential, and how to handle players switching roles within a formation, complicate the precise definition of a formation. In an initial attempt to solve this problem, Bialkowski et al. [10, 11] introduced the concept of role assembly – a classification of each point of a formation to one specific player. Shaw and Glickman [66] then added more rigorous techniques to quantify relative positions of players on the field for different periods of a match.

[3] **Key Performance Indicators**, a common abbreviation for the main statistics to describe a performance. These numbers are primarily domain-dependent.

These papers form a seminal role within this young research field by addressing some of the existing complications. However, they still lack a practical design to communicate their insights to a non-technical audience. Alternatives [51, 88] aim to provide a system to analyze formations but offer only single-match support in an unintuitive and predominantly technical layout.

This thesis seeks to close the defined research gap (Chapter 2) by introducing a web-based system that follows an intuitive design layout (Chapter 3) without compromising the statistical rigor to provide practical benefit. The contributions are fourfold: Computationally, the implemented approach accelerates the current state-of-the-art formation calculation burden to a fraction of previous techniques (Chapter 4). Increasingly accurate results exceed the formation prediction accuracy of individual seasoned domain experts (Chapter 7.2). Visually, while adhering to a familiar design choice, the system allows for new effects to communicate vital information effectively, especially for multi-match analyses. Practically, close collaboration, iterative validations, and detailed feedback from domain experts afford the development of an application that closely mimics a coach's work-flow on the sideline (Chapter 5). Rather than a single match-analysis, it leverages data from two seasons—250 games in total—of a premier European soccer league to provide an in-depth and longitudinal analysis of team dynamics (Chapter 6). This feature introduces robustness that will only improve with additional data in the future. One of the overall design prerequisites is extensibility: the system is easily adjustable for subsequent extensions (Chapter 7.3) to eventually provide a holistic source to analyze a team's collective movement patterns.

The contributions laid out in this thesis allow the efficient and intuitive exploration of formations, extending existing systems of soccer analytics with a crucial ingredient to quantify the drivers of success in sports.

2 Related Work

The rise of data-science techniques offering novel insights into competitive sports and health sciences accelerates the automatic retrieval of video data and subsequently the refinement of analyses. The following sub-chapters aim to provide a broad overview of work in soccer analytics by outlining the data retrieval (Chapter 2.1), common approaches within the intersection of visual analytics and soccer data (Chapter 2.2), and the work in formation research (Chapter 2.3).

2.1 Tracking Positional Data in Soccer

Automatic extraction of positional data is in itself a broad field within sports analytics research. It is vital to understand techniques common to the data-retrieval process. This chapter outlines the field's general approaches and explains techniques that collect the data lying at the core of this thesis.

While this thesis focuses on data retrieved through optical tracking, i.e., the derivation of player positions directly from video data, other player positioning sources exist. Advances in sensor technology [54, 55] and computer vision (classical work utilizes jersey colors [50, 64], while recent work builds on neural networks [34, 72]) offer rich data sources. Information, such as movement data [6], body posture [6, 16, 53, 76], and match events [7, 22, 34, 73, 79, 82, 89, 90] are readily available.

Camera tracking proves to be one of the main challenges of computer vision. Given an input image or a series of images from different cameras, camera tracking aims to recover information about the camera model, including its position and orientation relative to scene geometry and intrinsic parameters such as focal length or pixel aspect ratio. Defining a parametric camera model and then fitting the parameters of such a model to observations from the input images resolves this challenge. Most cases build on a simplified pinhole camera model, which induces a projection of a 3D point to the corresponding location in the 2D image. The projection translates to a linear transformation between the corresponding homogeneous coordinates. In the context of video data in team sports, the overarching goal is to register each input video frame into a global coordinate system. The cameras used in a match are typically stationary and only perform rotation and zooming. Therefore, it is possible to simplify further the transformation between different camera views to a planar projective one, also known as a *Homography.*[4] There are different ways of estimating a homography's parameters from a pair of input images, with the most common approach involving extracting a set of critical points from the input data. Afterward, a keypoint matching step finds corresponding keypoint pairs between the two images. A model fitting step uses the resulting correspondence information to produce a homography estimate.

Historically, several approaches address the problems of tracking soccer players from video data. Notable mentions include color encoding, which aims to distinguish jersey colors from the playing field [47]. This method suffers from apparent problems with jerseys, which resemble either the playing field itself, the banners surrounding the field, or the other team's jersey. One of the most challenging problems with such tracking is the brightness changes during a match—for example, a match might offer a different brightness in the afternoon than the evening. This complication might even coincide with different portions of the field covered by shade. Iwase and Saito [26] offer a refinement including homography techniques retrieved from multi-camera systems. While this technique afforded a marginal improvement of state-of-the-art results back in the day, it suffers from performance and set-up cost complications, which hinder a wider adoption of the techniques. Wang et al. [86] provide a more recent proposal to tackle the ball-possession problem, especially for longer sequences. Tracking the ball itself remains within an ambiguous position. On the one hand, it proves to be very important for identifying team possessions and automatically extract event data.

[4] **Homography** offers a linear mapping between two input image. This technique allows the deduction of camera motion, rotation, and translation.

On the other hand, it is inherently difficult because of occlusion problems and fast movements paired with the ball's small size. Wang et al.' s [86] multi-camera set up mitigates performance concerns and includes first proposals for the less hectic long possessions within the realm of soccer and basketball. Maksai et al. [40] introduce solutions by incorporating the projected trajectory of the ball as an additional ingredient to the ball identification algorithm.

Another complication arises with *re-identification.* Players often leave the visible field of a camera or cover each other, which subsequently becomes a challenge for the computer to recognize the player not as an entirely new entity but as the same player as before. Gou et al. [25] offer a widely used dataset from the computer-vision community of eight surveillance cameras on Duke's campus to train models, which outlines the issue of re-identification, proving its relevance in security, social research, and sport sciences. Alternatives [75] introduce a novel deep learning architecture to address this problem in real-time.

While this chapter provides a brief overview of the ongoing concerns of tracking persons in general or players on the soccer field, the historical outline shows that tracking techniques have come a long way since the color distinction algorithms of the late 1990s. Modern deep neural networks are fed with external features, such as physical behavior, to overcome single-camera systems and natural occlusion limitations. These techniques lie at the core of why advanced analyses, such as the formation research proposed in this thesis, are even possible.

Additional to the raw positional data, an important enrichment is the annotation of events within the game. Event data provide a match with an overall structure. Additionally to a match's natural temporal format, events allow the user to subset situations for when a team was ahead, behind, situations of a successful pass, or a combination leading to a scored goal. Automatic systems are already in place [22, 34, 73, 89, 90] and find many real-world applications [30, 46, 56]. Stein [69] provides a comprehensive description of event data in soccer:

> *"From a technical perspective, events are timestamped occurrences of previously known and defined categories, optionally annotated with spatial coordinates or additional information as involved players. Most events are directly ball related and correspond to actions with the ball (for instance passes or dribbling). Other events may be time-dependent (e.g., start and end of a play period) or not directly dependent on the ball (e.g., a foul situation during a free kick). [...] In practice, events might lack in accuracy, as they are usually annotated manually or as fully automatic recognition may produce false positive and negative events. As event data mostly contains information about players interacting with the ball, event data enables to conduct overall game statistics (e.g., passing networks, pass accuracy, or time between gaining the ball and shot on target). "* (Stein [69], page 23)

2.2 Visual Analysis of Soccer Data

Soccer research uses a wide array of data aside from tracking data to acquire in-depth insights into soccer success and automatic decision-making. Carling et al. [17], Castellano et al. [18], and Sumpter et al. [80] provide broad overviews over ways to utilize soccer data for systematic research. Theoretical soccer analysis closely connects with its actual applicability in the field.

Practitioners want to explore the data and demand interactive tools to visualize various soccer-specific events, such as ball passes, dribbles, or targeted shots [52]. Especially for coaches, who are usually in urgent need of simple solutions to process video data, advanced methods to automatically highlight exciting scenes from the game prove essential match analysis tools. Janetzko et al. [27] afford a system to introduce semi-automatic filtering of events based on spatio-temporal features, such as player acceleration or distance measures of players to the ball. Many more aspects of data lie at the intersection of visual analytics and soccer, such as fitness [14], defensive pressure [5], free spaces [70], and especially formation visualizations [10, 65, 66, 88]. Whenever large data sources, such as per-frame positional data of 22 players in soccer, requires efficient processing, clustering observations and data-reduction algorithms become vital. Sacha et al. [61] propose a semi-automatic approach to only resort to a small fraction of original data without sacrificing much of the informative content via trajectory abstraction.[5]

2.3 Formation Research in Sports

One of the pioneers in formation research is Laurie Shaw, who extends the tracking data analysis to professional soccer. Shaw and Glickman [66] introduce necessary assumptions for the algorithmic calculations derived in this thesis. Chapter 4 details this thesis' methodology, introducing more efficient formation calculation approaches extending previous work in the field.

Bialkowski et al. [10] present a classification algorithm that identifies soccer teams by their formation data exclusively. This approach introduces a critical contribution to how formations and movements for soccer teams are perceived: can the complex information of what discriminates one team from another be projected into a small number of dimensions to distinguish between teams accurately - or more poignantly: does a formation describe a virtual footprint of a team on the field? The authors propose an approach of deriving the formation of a team by solving a *minimum entropy*[6] *data partitioning problem* [32, 59]. Bialkowski et al. [11] outlines a full derivation. The difference to a traditional k-means algorithm[7] lies in the fact that individual points are assigned so-called *roles*, determined by solving the *Hungarian Algorithm* [29]. For the definition of the term "roles ", the authors describe their approach succinctly as:

> *"As a result, we refer to a formation's generic players using a set of identity agnostic labels to denote roles. A formation is generally shift-invariant and allows for non-rigid deformations. Therefore, we define each role by its position relative to the other roles (i.e., in soccer, a left-midfielder plays in-front of the left-back and to the left of the center midfielder). Each role within a formation is unique (i.e., no two players within the same formation can have the same role at the same time), and players can swap roles throughout the match. "* (Bialkowski et al. [10], page 3)

More technically, the authors describe the heat-map of the probability of a team's role assignment as

[5] **Trajectory Abstraction** describes a cohort of approaches to minimize the overlap between multiple trajectories (in this case, player and ball movement trajectories) by, for example, clustering and grouping them or introducing more advanced projection approaches.

[6] **Entropy** defines the level of information or uncertainty of a variable. For a fully random variable (no bias), the entropy sums to one since the outcome is fully unpredictable. For the distinct case, compare the entropy $H(X)$: $H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$, with $x_i = \frac{1}{N}$ because it is drawn from a uniform distribution.

[7] See Appendix A for a full technical aside on the k-means algorithm.

$$\begin{aligned} P(x) &= \sum_{n=1}^{N} P(x \mid n) P(n) \\ &= \frac{1}{N} \sum_{n=1}^{N} P_n(x), \end{aligned}$$

where $P(x)$ denotes the heat map for the entire team and n describes the set of roles. The authors simplify their assumptions by assuming a uniform distribution over all possible roles, which might diverge from empirical evidence (some roles, such as *center midfielder*, are naturally more likely to occur than others— for example *center back*, or *second striker*).

Using the logic of minimal overlap for optimally spread-out formations, the authors solve the Kullback-Leibler[8] divergence to measure the overlap between two arbitrary distributions $P(x)$ and $Q(x)$

$$KL(P(x) \| Q(x)) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx.$$

Eventually, the objective cost function simplifies to

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} \sum_{n=1}^{N} H(\mathbf{x} \mid n),$$

where $H(x)$ describes the cost in terms of entropy

$$H(x) = -\int_{-\infty}^{+\infty} P(x) \log(P(x)) dx.$$

Interestingly, the authors also notice similarities to the k-means algorithm, whose introduction to this application represents one of this thesis's core contributions.

This role assignment algorithm carries similarities with the current best-practices of formation analyses to define a player's position relative to its teammates instead of deriving absolute positions on the playing field.

The authors [12] combine general match statistics, ball-occupancy metrics, and formation estimations to form the input for their classification task to blindly determine a team's identity entirely based on tracking and match event data. The results reach an overall accuracy of 70.38 %, with most impact stemming from the formation information (67.32 %). This significance in predictive benefit corresponds strongly with this thesis's primary motivation: formations are an integral descriptor of a team's core strategy because they quantify how teams move collectively.

Similar work describes dynamic role identification given an entire season of soccer tracking data for one team [11]. The authors argue that current team analyses lack the contextual information for meaningful interpretation of the work. While assigned player positions attempt to alleviate this tagging misalignment, the information is often lost when players switch assigned positions halfway through a match. Therefore, comparisons by assigned positions become challenging.

[8] **Kullback-Leibler Divergence**, also called *relative entropy* refers to a standard measure of the difference between two probability distributions. Applications span as wide as time-series analyses, statistical model comparisons, and quantifying entropy in information systems.

The canonical problem states that a given player's x-, y-coordinates determine a fixed identity, which remains unaltered throughout the match. This approach leads to processing- as well as interpretation-bottlenecks. Therefore, the authors introduce the notion of formation as a *fixed set of roles* that whatever player can occupy - the roles remain fixed, but the players are allowed to swap roles. A permutation matrix, minimizing the total cost to fill a period's formation, solves this assignment. The algorithm begins from the first frame and initializes the roles' distribution with the players' actual positions. Mining multiple player trajectories remains one of the main challenges of analyzing team behavior, which leads to a sparse corpus of formation research compared to individual players' movement analysis. Adjacent efforts [92] combine all the individual trajectories into an aggregate trajectory, utilizing *time warping*[9] techniques, either for basketball [15], or American Football [78]. As discussed above, the authors proceed to compare the overlap of the individual role probability functions with the team's and solve the *minimum entropy data partitioning problem*[10] with the Kullback-Leibler divergence. Solving the expectation-maximization procedure, the authors identify clusters based on the player's roles, which allows for data visualization applications (pairing key events and the role) and effective event segmentation to distinguish team behavior based on strategic decision-making. This paper [11] represents the first effort to separate a match distinctly into segments, here five-minute segments, which stands in contrast to the separation between halves. Their results show how the midfielders—most notably the left and right-wingers and the two central midfielders—exchanged positions throughout the match. Absolute position calculations lose this notion [44]. Bialkowski et al. [12] extend this approach to a separation between in possession and out-of-possession comparisons. The authors conclude that out-of-possession formations tend to be more expansive, but that especially a more detailed analysis regarding the field's location offers subtle insights: intuitively sound, teams move towards a more aggressive structure approaching the opponent's goal.

Ma [37] provides one of the most recent contributions to the literature of formation analysis. It utilizes the formation definition of Bialkowski et al. [11] to define roles for players and consequently their spot on the field and measure adherence or variance from that formation at specific time-intervals. The paper outlines the advent of models introduced by Fernández et al. [23][11] and Spearman et al. [68], Spearman [67],[12] who introduce quantifying goal probability measures for a given situation.

Spearman [67] utilizes role assignment techniques [10] and also classifies each player into one of three rows.[13] Their data consists of a rich set of 25 frames per second data for 378 soccer matches of an elite European league. The author performs his disruption analysis by building

[9] **Dynamic Time Warping** describes the comparison of sequential signaling data of different speeds. It matches the first and last index of the sequences and *warps* the entire data stream into this new time interval.

[10] Roberts et al. [60] offer a detailed explanation of the procedure. However, the basic idea boils down to an efficient clustering algorithm for high-dimensional data.

[11] The authors establish deep-learning techniques which decompose any possession as the sum of expectations for either a pass, a shot, or a drive.

[12] Spearman et al. [68], Spearman [67] infer probabilities for a goal within a highly interpretable probabilistic framework. It builds on the notion of *Field control*, which the authors model as the probability for a successful pass to a teammate. The scoring probability model then uses this probability as its main parameter.

[13] Formations are often referred to by three numbers, respectively, indicating the players in a given row. For example, "4-4-2 " indicates four backs, four midfielders, and two forwards.

non-exact windows (a set of frames for an entire possession but which does not necessarily refer to 180 seconds in length). The results indicate that most subsequent windows do not significantly change formations (measured by the Wasserstein distance[14] between the formations). A bootstrapping method allows for constructing a counterfactual “what-if the team had continued to play their actual formation “measure used as a benchmark. This contribution allows for a statistical measure to distinguish formations because, as Bialkowski et al. note, many formations in practice appear quite similar and are easily confused by a clustering approach [10, 11]. The authors note that their metric can enrich event data by highlighting formation changes of adjacent possession periods.

Bialkowski et al. [9] describe a real use-case of team formations. It addresses the long-held myth in soccer of the home-team advantage. In a previous paper [36], general statistics address the same question and indicate a significant difference between possession and shot statistics for home and away teams. To deepen this approach, the authors utilized their previous role-assignment technique [10] to analyze formations. Since roles might dynamically change throughout a match and researchers cannot explicitly infer the formation from the data set, the authors introduced an expectation-maximization metric paired with the Hungarian algorithm to update the players’ roles and calculate the previous error assignments. The data is subsequently normalized and averaged to the center of the field to allow for comparability. This approach allows for a dynamic match summary visualization of formations and statistics of a sliding five-minute window, which captures the match’s nuances better than current sparse statistics usually presented at half-time. This technique underlines previous results to strengthen the home-game advantage: teams tend to play very similar formations away or at home. Their relative positions move further up the field when at home, indicating more aggressiveness and comfort in the game. Results might indicate a tendency for strategic goal-setting differences depending on home or away, i.e., win home games and draw away games.

Shaw and Glickman’s [66] contributions to the analysis of soccer formations are manifold: a novel approach of defining the relative position of all the players, comparing and therefore classifying common formations, and providing a match summary outlined along with significant events during the match.

While Chapter 4 explains the extensions to the authors’ methods utilized in this thesis, the following paragraphs will provide a short overview of similar work’s [66] core principles. First, the authors describe each player’s average position to one another during every time frame. This 10×10 distance matrix (ten field players) allows them to average the relative position to one another over a pre-specified time interval. The resulting matrix dimensions are still 10×10 but reveal the players’ average position to one another. The next step includes defining the most common third neighbor of all players as the centroid of the formation and iteratively determining the position of the players from this starting point (find the nearest neighbor, then the nearest neighbor of nearest neighbor, et cetera). Once all these positions are determined, the resulting formation calculation is re-run over two-minute intervals during which the team was in possession. While a continuous possession of two minutes is almost impossible in modern

[14] **Wasserstein distance**, also often called *earth mover’s distance* first introduced in [83] calculates how similar two distributions are.

soccer, the authors decide for a different approach: They allow the time slots to potentially appear at different real times, meaning that every time a team is in possession, the timer continues to count until a two-minute sequence is reached. All these individual possessions are then wrapped together to form one sequence. This process repeats until the entire game's periods in possession are labeled into two-minute long sequences. The grouping of positions per frame to positions per sequence introduces more robustness to the formation prediction than previous approaches. These about 4,000 sequences are then compared using the Wasserstein distance and clustered agglomeratively within 20 groups. The classifications highlight defensive to offensive transitions (of frequent pairs), overall strategic tendencies, or possible ways to exploit predictive team formations. Shaw [65] most recently extends the approach to a subset of *phases* for the clusters. These phases include phase categories, such as offense vs. defense, transition, and set-piece phases with then underlying phase types - for example, ball retention, low block, counterattack, or corners. The categorization of phases is an ongoing problem and currently still relies on manual tagging by analysts.

The author utilizes a labeled dataset to highlight distinct patterns of formations for various team progressions over the field. The method allows for measuring formation disruption to indicate how dangerously a team moves around a goal (by disrupting the opponent's defensive formation). The measure is closely related to an approach presented in previous work [41], which quantifies the relation between space occupation and formation disruption. The authors visualize these relations in Voronoi Tesselations[15] and quantify the relationship loosely with possible extensions for a more causal research design.

In a final step, the authors derive a Bayesian classification model to label unseen formations into one of the derived clusters of formations. The resulting matrix consists of a d-dimensional vector, where d describes the number of clusters—here 20—used as the cutoff to the agglomerative approach. The probability is calculated as (for more intuition, see Shaw and Glickman [66], or Appendix C)

$$\mathrm{p}(\mathrm{o} \mid \mathrm{C}) \approx \underset{k}{\operatorname{argmax}} \prod_{p=1}^{10} \int \mathrm{p}\left(y \mid k\mu_{\mathrm{p,C}}, k^2 \varSigma_{\mathrm{p,C}}\right) \mathrm{p}\left(y \mid \mu_{\mathrm{p,o}}, \varSigma_{\mathrm{p,o}}\right) \mathrm{d}y,$$

where $\mu_{\mathrm{p,C}}$ and $\varSigma_{\mathrm{p,C}}$ are positions and covariance matrix for role p in Cluster C. $\mu_{\mathrm{p,o}}$ and $\varSigma_{\mathrm{p,o}}$ are the position and covariance matrix for player p in the formation observation o. k is the scaling factor the authors use to maximize the similarity between formations.

This approach suffers from common limitations to Bayesian models: They assume homogeneous and continuous distributions for their priors. The prior distribution is inferred from theory, while the remaining part (the model evidence) counterweights this information to form the posterior distribution and, therefore, the selected model. The authors do not describe parameter fine-tuning, which solves the model for optimal input assumptions. Furthermore, the proposed algorithm finds the optimal scaling-factor k, which zooms-in or -out on resulting formations. This

[15] **Voronoi Tesselations / Graph** represent a spacial visualization technique separating space into distinct geometric figures. This link offers a colorful and illustrative example for different optimization metrics. It splits a plane into distinct areas, where all points in a given point's area are closer to their "own "associated point than to any other point on the plane.

decision results in a trade-off because it loses some of the information contained by the compactness of a formation while it offers more comparability between formations. More poignantly: A 4-4-2 team structure might be spanning the whole field or the middle third, for which the implications for a coach vary drastically.

Narizuka and Yamazaki [45] introduce a more subtle approach to cluster formations. The authors describe a soccer formation as an adjacency matrix and utilize Delauney triangulation[16] to pre-define common formations, such as 4-1-4-1 or 4-3-3, and subsequently split the results via agglomerative clustering[17] (similar to the methods of previously discussed work [66]) to derive more robust result-clusters. This approach allows for insights into the typical formation transitions of a team within single games. The most appropriate number of clusters seem to be around 15-20 clusters, which corresponds to the 20 clusters of previous work [66] (see Figure 3 of the work of Narizuka and Yamazaki [45]). Llana et al. [35] utilize these insights [66] for more dynamic analysis: Highlighting situations when a pass attacks a defender's designated *zone* and how subsequent movements of defenders compensate for the attack by compromising the original defensive formation. The author introduces the notion of zones per player to quantify these metrics and calculate probabilities for a goal as the cost of a given player moving *out-of-position*.

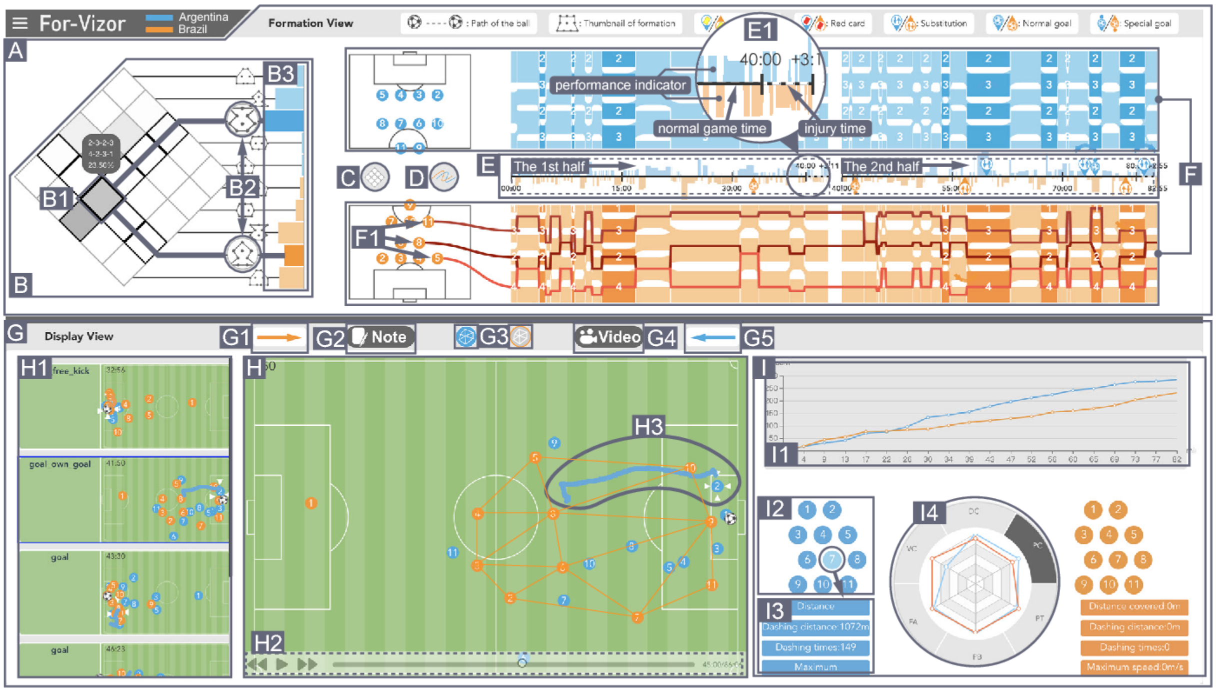


Figure 2: Figure 4 of Wu et al.'s paper [88] illustrates the researcher's formation tool. It introduces time-series analysis of a single-match, indicating formation changes between the teams allowing for direct comparison of scientific movement research.

[16] For our purposes, a **Delauney Triangulation** represents the dual graph of a Voronoi diagram. A **dual graph** describes a graph with a point (called a *vertex* in graph theory) for each separate area (also known as a *face*). Therefore, one point exists in every distinct subarea of the graph.

[17] This clustering approach extends their previous work: Narizuka and Yamazaki [44] introduces the main gist of the explained methods. However, the assumptions for the approach only allow comparing formations within the same match. For multi-match analyses, the described clustering extension becomes necessary.

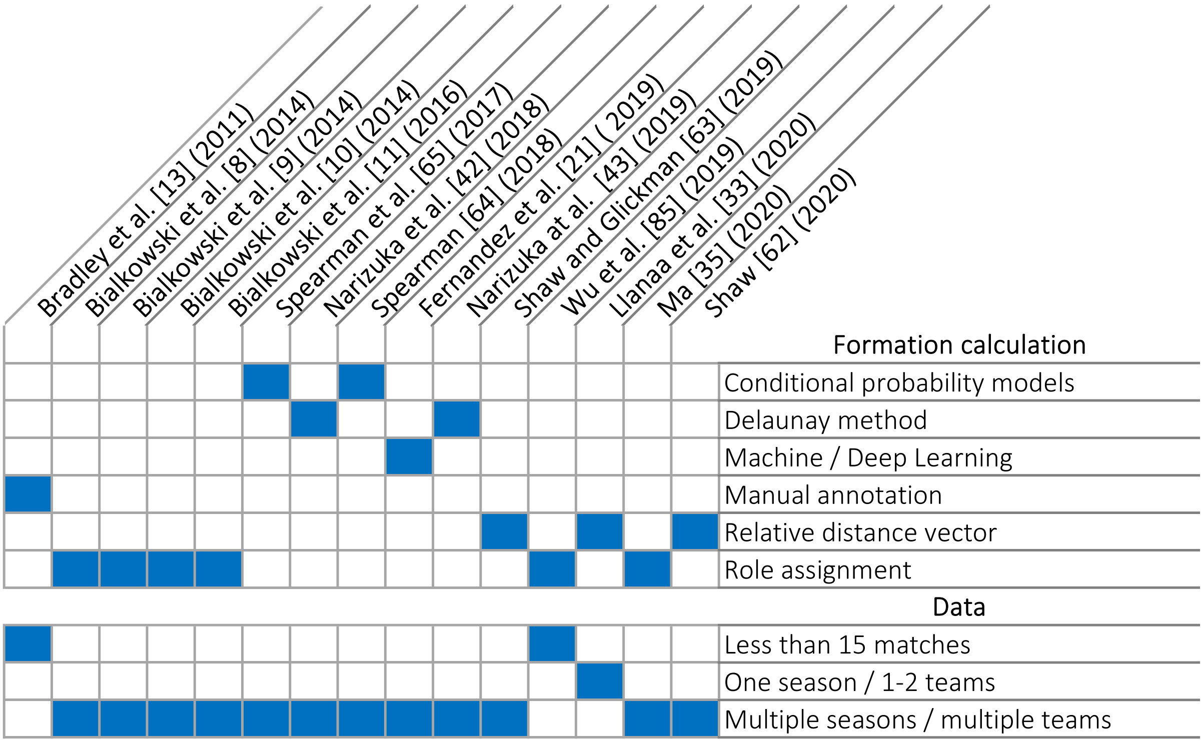
Wu et al. [88] offer a sibling project to this thesis's aspirations: build an interactive visualization tool to investigate formations. Their result *ForVizor* introduces extensive insights into the subtleties of formation changes throughout a game within a single system. The system includes many alternatives for single-match comparisons and represents powerful tools for detailed investigation of single-match scenarios.

The authors develop the application in collaboration with sports science Ph.D. students and an Asian professional head coach while evaluating the tool on two games of the Under-15 Football World Championship. Their data-processing pipeline involves a manual tagging of player positions since an automatic workflow proved problematic. Formations are consequently derived building on the restricted k-means [11]. The segmentation of the data into event-driven periods finishes the data processing. Figure 2 highlights the tool's primary possibilities. They include features that mainly aim towards single-game development and time-series visualizations. Ideally, this granularity level offers explanations towards formation change or transitions. Experts evaluated two games, and their particular workflow during the process is described in detail throughout the evaluation, emphasizing the importance of intuitiveness and the balance between the information conveyed embedded into a simplistic design. Even though the authors' research goal seems similar to this thesis's aspirations to build a visual exploration tool for formations, the system design proves problematic for practical multi-match analyses for three main reasons. First, it leverages only single games, focusing on the formation shifts throughout a single game. Second, the tool does not introduce an intuitive design flow. It appears more scientifically-driven than with a practical goal in mind. Finally, the data cleaning pipeline builds on manual tagging, which introduces personal bias and disqualifies the system for larger scaled data sets.

Bradley et al. [14] provides a compelling use case of formation data in general, which uses soccer formations as a means to identify possessions and athletic behavior of players, more specifically, high-intensity sprints during a game. The paper uses tracking data of 153 English Premier League soccer games, tags formations manually, and researches the effect that different formations have on a team's behavior. They found subtle differences among various common formations (such as 4-4-2 or 4-3-3) for high-intensity sprints while in and out-of-possession but did not conclude any statistically significant difference between a team's overall possession. This work further underlines the demand for a multi-game-system for formation analysis for practitioners and researchers of various adjacent domains (here, fitness and performance research).

Figure 3 visually compares the primary papers discussed in this sub-chapter in the form of a *STAR*-report.[18] This brief overview highlights the explained formation calculation and the varying depth of data quality underlying the research. It underlines the field's novelty with the first technical papers, introducing algorithms comparable to today's methods as recently as from the early 2010s onward.

[18] **State-Of-The-Art** reports are topical narratives that explore the current status of the frontiers for a specific research field's methods. Visualizations often accompany these expansive overviews of literature comparing the discussed papers on relevant key attributes.



	Bradley et al. [13] (2011)	Bialkowski et al. [8] (2014)	Bialkowski et al. [9] (2014)	Bialkowski et al. [10] (2014)	Bialkowski et al. [11] (2016)	Spearman et al. [65] (2017)	Narizuka et al. [42] (2018)	Spearman [64] (2018)	Fernandez et al. [21] (2019)	Narizuka at al. [43] (2019)	Shaw and Glickman [63] (2019)	Wu et al. [85] (2019)	Llanaa et al. [33] (2020)	Ma [35] (2020)	Shaw [62] (2020)
Formation calculation															
Conditional probability models						■		■							
Delaunay method							■			■					
Machine / Deep Learning									■						
Manual annotation	■														
Relative distance vector											■		■		■
Role assignment		■	■	■	■							■		■	
Data															
Less than 15 matches	■											■			
One season / 1-2 teams													■		
Multiple seasons / multiple teams		■	■	■	■	■	■	■	■	■	■			■	■

Figure 3: This STAR-report highlights the literature underlying formation research in soccer. It displays differences in calculation logic and data structure. The research projects are ordered by publication year in ascending order from left to right.

2.4 Positioning of This Work

This thesis introduces an analysis system that allows novel interaction possibilities for practitioners. It extends the most efficient and accurate algorithms to date to dynamically analyze soccer formations. The system improves algorithmic solutions in previous work [66] while achieving better forecast results than domain experts. Semantic papers of the field require strong model assumptions for their algorithms. This thesis alleviates much of the necessary restrictions by allowing for a data-driven k-means formation assignment, which is ignorant of role probability distributions [10] or an arbitrary definition of formation centroids [66].

This thesis leverages the visualization-tool aspirations of Wu et al. [88] and extends their ambitions to multi-match analyses. Close collaboration with experienced domain experts allowed for a pragmatically-driven development stage, culminating in an application bridging the gap between isolated academic research and practical usability. This thesis's main contributions are four-fold: first, streamlining the calculation logic alleviates former algorithms' computational costs to a fraction and eliminates the need for restrictive assumptions. Second, the system achieves higher formation prediction accuracy than seasoned experts. Third, the system allows for direct user interaction on a multi-match level, and fourth, the tool's extendability and intuition allow it to become a single system to understand soccer teams' formation behavior. The employed visualizations mimic a coach's workflow preparing for an upcoming match and combine advanced algorithmic accuracy with intuitive readability. Supplementary to iterative discussions during the development phase, this thesis presents an extensive validation with uninvolved domain experts quantifying the system's accuracy as more reliable as each expert's prediction. It also identifies the potential for further improvement discussed in Chapter 7.3.

3 Requirement Analysis

We conducted semi-structured interviews to understand better where an information disconnect exists between the demand for formation analysis and its supply. Iterative interactions with a professional head coach of a European premier league soccer club clarified the specific needs of the day-to-day routine within the professional sports realm. The expert proposed design choices within meetings to mimic best a coaching staff's everyday routine analysis of formations. Even though the meetings did not aim towards a specific agenda other than utilizing the unique insights from this source of professional sports, the questions asked intend to draw a holistic picture to meet the practitioners' demands best proactively. Sedlmaier et al.'s [62] nine-stage approach inspired the general structure of the questionnaire. Generally, the idea of the conversations aims to narrow the app design from questions of an open-ended **why**,[19] over a more specific **what**[20] to, finally, a pragmatic **how**[21] discussing basic application features. Figure 4 illustrates this overarching strategy.

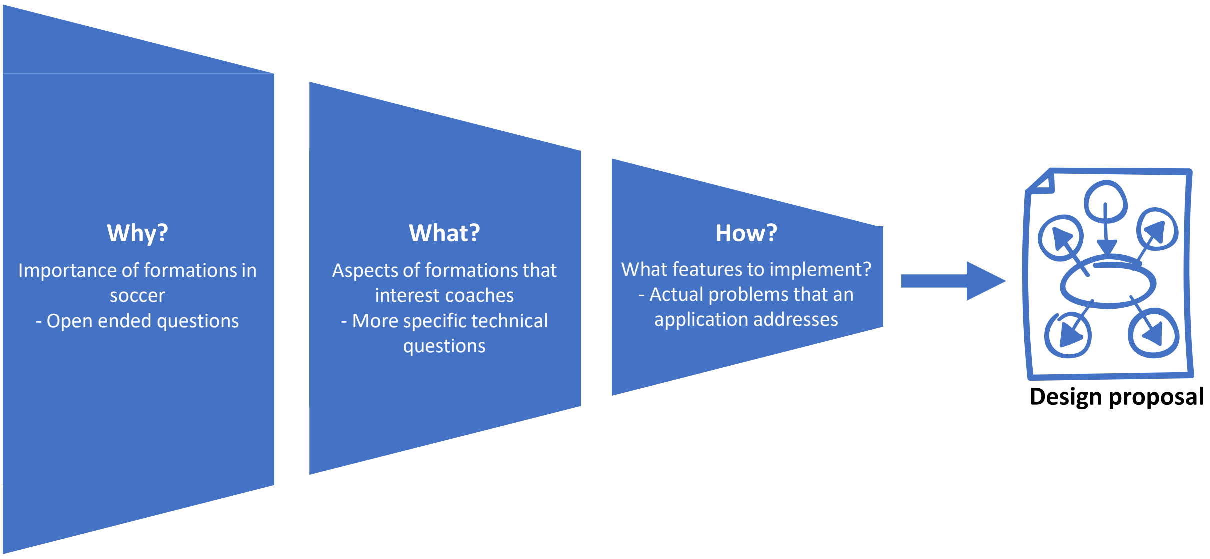

Why?
Importance of formations in soccer
- Open ended questions
What?
Aspects of formations that interest coaches
- More specific technical questions
How?
What features to implement?
- Actual problems that an application addresses
Design proposal

Figure 4: The overall structure of the iterative feedback interviews follows a stringent logic from open-ended questions to more pragmatic application features. This figure illustrates how the questions begin with the motivational background and end with a wish-list of features that an ideal tool should contain. The following examples of interview questions mimic this logic.

The scope encompasses weekly brainstorming sessions within the development team (for at least the first two months of the project), one feedback session with a domain expert, and three validation sessions, including qualitative feedback. The interviews' language was *German*, and the expert interviews generally lasted between 60 - 120 minutes. The following excerpt represents a set of representative interview questions from a feedback session at the end of December 2020. A summary[22] of key takeaways from the experts' responses follow their respective questions. The

[19] The causal stage of the questionnaire addresses broader motivations to even consider formations as a valuable tool to analyze team strategy.

[20] This stage inquires aspects of formations specifically that need further investigation.

[21] The final stage of the questionnaire dives into the practical features of a tool to facilitate the practitioner's job.

[22] For the sake of brevity, the response paragraphs only summarize critical focal points. The summaries exclude single illustrative examples without ignoring any significant comments.

interviewee of this expert session was a head coach of a European first league soccer team. His additional experience as a professional player and certified analyst makes him extraordinarily suitable for the system's qualitative evaluation. His insights, coupled with the additional discussions, describe the main determinants of the design requirements and feature development for the web application.

1. Why is the analysis of team formations (own and opponent's) important for you?

   A1: *"Formations allow the coach to create favorable situations for the team. Defining spaces on the field is generally handled by assigning specific formations. The main determinants of the own team's formation are the team's constitution and the opponent's formation. Formations offer the team responses to the dynamically changing situations within a soccer game and, therefore, a toolset that aims to maximize the chance of success given the own team's skills."*

2. How do you decide which formation to assign to the team?

   A2: *"Additionally, to the points mentioned in A1, formations are also strongly impacted by individual player skill."*

3. How much does the formation depend on the opponent or situation?

   A3: *"Two distinct convictions of formation derivation dominate the preparation. The one side prepares its own team's formations on specific blueprint-situations. Once any of these most probable scenarios occur, the team will follow a pre-defined match plan. The other side, while incorporating information about upcoming opponents, might act differently to seemingly congruent situations. This behavior might allow a surprise effect and confuse opponents."*

4. What are the main tasks for the coaching team in pre-match and post-match preparation, especially concerning the formation?

   A4: *"Besides understanding the specific overall tendencies of a team, the main determinant is to know an opponent's common formation - and their respective adherence or variance between formations. This information will then further subset solutions to the four phases of a game.*[23] *The opponent team's response to common formations is also of vital interest."*

5. What general information would help you and your coaching team make better decisions or reach the same decisions more easily?

   A5: *"An objective presentation of information. A wider granularity is narrowed down to a finer level to allow for the right input given the wide use-cases of such an application. The program will serve as an additional tool during the match preparation or during the actual match. While subjective ideas might be biased, a source of objective advice will prove helpful to find the right action. A valuable extension towards an AI-powered*

[23] The **Four Phases** of a soccer match are defined differently across sources, in this correspondence the coach clusters a match into *the team is out-of-possession, the team is in possession, either team is losing the ball (transition), offensive and defensive set pieces (such as wide goalie passes, corners, out-balls, or penalties).*

> *recommendation system,[24] which will use facts to find the best formation for a given situation, would further facilitate the job of a coach during these hectic moments within a game."*

These detailed answers translate into five high-level design requirements, which an application should adhere to for optimal assistance of a coaching staff's daily challenges.

- **Dynamic access from different locations.** The system aims for usability in situations from locker room preparation to sideline analyses. A simple sharing functionality for a non-technical audience will offer a smoother acceptance to break existing practices without any set-up costs. Web hosting of the system will offer the required flexibility.
- **Support different analysis-granularities.** The specific demands of invested parties and potential use-cases of a formation analysis tool are too broad to predict. Allowing an adjustment of displayed information to varying degrees of detail prepares prospective extensions and provides different use cases an appropriate outlet without over-complicating a single visualization.
- **Analyze specific subsets of information.** Formations in themselves are intriguing for researchers, but coaches need to find a way to win games, which often includes preparing for a specific opponent or situation. Providing a tool that does not pre-describe an analysis but offers all the information necessary to conduct their own analysis forms a cornerstone of the application design.
- **Extend easily.** The tool will be dynamically reviewed and should allow for the introduction of new features. As dynamically as the game itself is changing, demands for a proper analysis tool might shift. Therefore, the addition of aspects as necessary to the existing application's logical flow will promise the app's versatility in the future.
- **Communicate results intuitively.** While statistical analyses might prove useful as the foundation to more advanced features of a recommender system, intuitive communication to a non-technical audience is of vital importance to the tool's usability. Coaches will show results to players, scouts, or any other stakeholder. The application should feel native and organic to the usual communication channels used within these parties while enhancing the insights.

From these broad principles, more specific and practical design requirements determine the app layout's actual decision. Often alternative design choices need to be weighed against one another. These compromises will need to form a coherent experience and follow an overarching flow. That is where the more specific design requirements come into play. These propositions are structured into three groups, *Algorithmic*, *Search-/Subsetting*, and *Visualization Requirements*.

[24] **Recommender systems** are most prominently applied in commercial applications to give the user the most appropriate recommendations given the history of previously selected items and similar search histories/user profiles. Prominent examples are item recommendations on Amazon for online shopping or movie/ series recommendations on Netflix for video streaming.

A Algorithmic Requirements define necessary information and performance features that allow practical usability to deliver the critical insights by the user.

A1 **Correct formation calculations** build trust in novel methods and offer information that the user requires to make better decisions. The application will improve with more data, and extensive evaluations check the calculated results with gold-standard expert opinions.

A2 **Inference about individual players** is important to communicate the results effectively. Formations are crucial, but it is the individual players that need to adhere to the match plan. If the formation algorithms label individuals, coaches can translate the information to personal feedback.

A3 **Fast algorithmic performance** allows the application to scale with more data. If single derivations take hours, the app will lose its practical value to the user. The algorithmic choices should mimic the system's vision to grow more accurate and instructive with more data.

A4 **Derive calculation (in)accuracy** will be vital for the effective communication of the results to non-technical audiences. The system's focus is to instruct and support, not replace the experts on the sideline. Therefore, a real measure of approximation validity becomes imperative for the user's decision-making.

S Search/Subsetting Requirements define a set of options to efficiently find the information a coaching staff requires to deduce valuable information regarding formations.

S1 **Intuitive design choices** follow the analog line of questioning of a coach. Usual workflows that practitioners use, regarding formation analysis, establish the order and the granularity of tabs within the application.

S2 **Limitation to relevant options** allow the design not to become cluttered with too many drop-downs and selection boxes. By focusing on the most important questions, coaches can ask vital questions quickly without losing the benefit of advanced analyses.

S3 **Iterative subsetting** mimics the complex scenarios that offer a real insight for match preparation. Often not only the formation of a given opponent but of a given opponent in a specific situation within one of the four-match stages displays the level of detail necessary to answer advanced questions. To bridge the gap between a simplistic design while offering enough detail to ask the important questions becomes of utmost priority.

S4 **Easy extensibility** introduces the notion of a system that grows organically and extends as necessary. Possibilities to intersect a formation analysis application with adjacent tools will offer a holistic picture to describe soccer success quantitatively.

V Visualization Requirements define the scope of responsibilities that the result presentation entails. Choices are based on these granular resolutions to provide a clear and coherent underlying logic to how the coaches interact with the application and convey its knowledge.

V1 **Familiar result presentation patterns** include visualizations, which feel *comfortable* or *convenient*. The application users will not be technical statisticians but soccer coaches and scouts; therefore, a translation of mathematical concepts to native soccer communication channels offers an easier transition.

- **V2** **Clear separation of concerns** determine how many aspects of a given analysis are bundled together in one visualization or tab. The app's design needs to bridge the gap between disentangling users' concerns and maintaining a coherent wholeness to describe complex formations. While some questions might demand isolated information, others require context to offer any practical insights. A design choice to represent this ambiguity will offer more organic incorporation of the application into its user's daily workflow.
- **V3** **Fast access and performance** are crucial for the user experience. Coaches will not have the time or patience for data to be loaded into the app or have calculations run for minutes before any results are accessible. A smooth and immediate front end user-experience needs to mimic the work a coach can do on a whiteboard in front of a team. A web hosting of the application will achieve the best trade-off between fast performance and easy access.
- **V4** **Comparability of different queries** affords the user to deduce more complex scenarios. By offering the coach options to compare different teams or compare a subset of situations and their formations to a benchmark, deviations from formations might become transparent and introduce more tactical insights for decisions to react.

The following two chapters will outline solutions that are determined by these requirements derived from expert opinions. While Chapter 4 will address the 🖩 **Algorithmic Requirements**, Chapter 5 demonstrates how the app focuses on the 🔍 **Search/Subsetting Requirements** and the 📊 **Visualization Requirements**.

4 Innovating Formation Calculations

This chapter outlines the different required steps for the actual calculation of the formations. It explains the implemented mathematical derivations, expands on the rationale for using specific algorithms, and gives the technical background on the project's overall strategy. Consequently, this chapter offers a background on the system's solutions to the algorithmic design requirements laid out in Section 3.

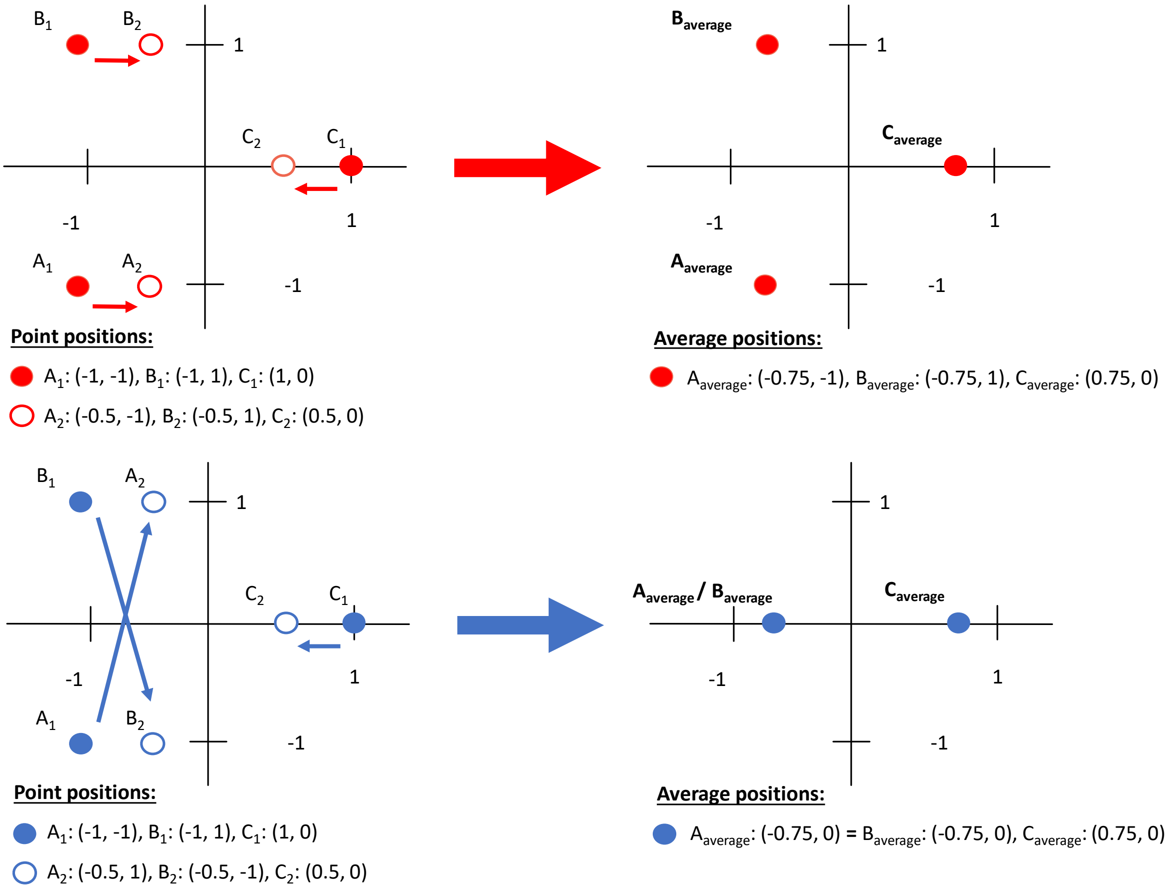
The first two sub-chapters (Chapter 4.1 and Chapter 4.2) provide the technical background to first retrieve formations from simple location data and then group and compare them. Chapter 4.3 explains the mechanisms underlying scaling formations based on their compactness. Once the technical background is transparent, the chapter outlines actual data structures, and necessary assumptions in Chapter 4.4 continues. The fifth sub-chapters (Chapter 4.5) defines the necessary pre-processing steps and assumptions to allow for a meaningful interpretation of formations. The thesis attempts to be precise enough in its instructions to allow for a replication of the analyses.

4.1 Formation Calculation

A formation describes a constellation of all player locations at a given time in team sports. In soccer, three to four numbers describe these formations listing the number of players in their respective rows, starting with the most defensive line. For example, *4-4-2* indicates four backs, four midfielders, and two forwards.

Formation calculation encompasses several assumptions to make, such as normalizing locations, describing the compactness of formations, and how to group time frames to infer robustness of the player constellations. This thesis lies at the intersection of theoretical and empirical research, which underlines the necessary heuristics used, further described in Chapter 4.5. All these assumptions are necessary to comply with, first and foremost, 🖩 **A1** ***Correct Formation Calculations***, as well as 🖩 **A3** ***Fast Algorithmic Performance***.

Since formations describe a collective movement over time, the calculation needs the actors' x- and y-coordinates grouped for a relevant period. Chapter 4.5 further extends on the time-spans intuition utilized for the calculation. With a length of two-minute segments of one team, either in possession or out-of-possession, a single sequence to lay the basis for a single formation calculation contains 24,000 (120 seconds x 10 frames/second x 20 players) individual observations (tuples of x-, y-coordinates).



B_1
B_2
1
C_2
C_1
-1
1
A_1
A_2
-1
Point positions:
● A_1: (-1, -1), B_1: (-1, 1), C_1: (1, 0)
○ A_2: (-0.5, -1), B_2: (-0.5, 1), C_2: (0.5, 0)
$B_{average}$
1
$C_{average}$
-1
1
$A_{average}$
-1
Average positions:
● $A_{average}$: (-0.75, -1), $B_{average}$: (-0.75, 1), $C_{average}$: (0.75, 0)
B_1
A_2
1
C_2
C_1
-1
1
A_1
B_2
-1
Point positions:
● A_1: (-1, -1), B_1: (-1, 1), C_1: (1, 0)
○ A_2: (-0.5, 1), B_2: (-0.5, -1), C_2: (0.5, 0)
1
$A_{average}$ / $B_{average}$
$C_{average}$
-1
1
-1
Average positions:
● $A_{average}$: (-0.75, 0) = $B_{average}$: (-0.75, 0), $C_{average}$: (0.75, 0)

Figure 5: The illustration depicts a common problem of formations, where players might change positions throughout the calculation interval. The example builds on three players (here, A, B, C) over two time-intervals (t_1 and t_2) but organically extends to more complex scenarios for real data. The top part demonstrates how the calculation will correctly derive the average calculations for a toy 2-1 formation as long as the players are not changing their roles in the formations. The resulting average positions (top right) conclude a 2-1 formation with seemingly correct locations. The problem arises in the bottom scenario, where players A and B switch their position between t_1 and t_2. Their average arithmetic location concludes that they played on the same wrong position on the field and that the resulting formation forms a 1-1 structure.

Finding these players' average position represents a major challenge to define a formation/sequence metric further. Players might change their role in the formation without altering the overall formations. By just averaging over the arrays of positions, one might miss this movement and find a nonsensical result position for a player. Figure 5 illustrates the problem visually.

This complication of incomparable location arrays arises when the players switch roles or if the arrays within the data are unsorted. So the player at index one might not correspond to herself at index one at another time frame. A solution lies in the implementation of the Hungarian algorithm [29]. For a detailed explanation of the underlying mechanics, please refer to Appendix B. The calculation will use a 10×10 distance matrix of every field player to every other field player (with the diagonal filled with zeros describing player distances to themselves) to find an overall optimal[25] assignment. The problem with this approach, however, lies in its computational efficiency. While effective implementations of the Hungarian Algorithm run in O(n^3), this calculation needs to run for 2,400 frames of a two-minute sequence. Every individual assignment results for an array length of n=10, in 1,000 calculations (10^3) and, therefore, 1,000×2,400, or 2,400,000 calculations for a single two-minute sequence. Since every game consists of approximately ten offense sequences and, consequently, ten defense sequences, the total number of calculations will approximate to 2,400,00 × 10 sequences × 2 modes (offense/defense) × 2 teams = 96,000,000 single calculations per game. This number is not feasible for an efficient system to offer match insights immediately (see algorithmic requirement 🖩 **A3** ***Fast Algorithmic Performance*** and visualization requirement 📊 **V3** ***Fast Access and Performance***).

This optimal assignment approach lies at the heart of adjacent work [10], which introduces the notion of *roles* and minimizes the total entropy as explained in Chapter 2.3.

Shaw and Glickman [66] propose an alternative algorithm. The authors provide techniques to extend relative player positions to one another in the following steps:

1. Find the relative distance matrix of all players to one another (t, 10, 10) for the ten field-players for every time frame t in the sequence. In our example, for two-minute-long sequences, this calculation translates to a (2400, 10, 10)-dimensional matrix.

2. Calculate the average distance matrix for the players to one another, which builds the average of the 2,400 time frames resulting in a (10, 10)-dimensional matrix.

3. Define the densest part of the formation as the centroid, which Shaw and Glickman [66] define as the player who is the most frequent third nearest neighbor.

4. Set this player's coordinates to (0, 0), which usually translates to the field's center circle.

5. Derive the relative position of this player's closest neighbor as the distance vector from the player. Then re-define its location in terms of the location from the center of the field.

6. Continue with step five by continuously determining the closest neighbor's position relative to the currently iterated player's position, ignoring already assigned players until all player positions are derived.

[25] *Optimal* refers to the assignment with minimized total cost.

This approach addresses some problems of normalization of the locations on the field. Two otherwise identical formations might be confused to be distinct if they are not normalized and appear in different parts of the playing field. This approach describes another trade-off similar to the compactness measure k (see Chapter 2.3 and Chapter 4.2). It weights comparability against losing a sense of the exact location formation occurs on the field. The result formations center around their centroid in the middle of the field and stretch for maximum comparability.

This thesis introduces an alternative to the proposed algorithms. It solves the optimal assignment problem of defining roles [10] and the downsides of absolute positioning on the field [66] by utilizing the simplicity of k-means clustering to the locations on the field.

The simple and efficient calculation boils down to four high-level points:

1. Define two-minute-long sequences of in possession or out-of-possession time.
2. Normalize the position per two-minute sequence.
3. Find the average position per sequence.
4. Subset the data to find the sequences relevant for a specific query and find the average formation for this data via k-means clustering.

The remainder of this chapter details each high-level point's approach and explains the necessary assumptions and algorithmic shortcuts to allow for an accurate yet fast derivation of formations from two-dimensional spatial-data.

Step 1: Define two-minute sequences:

The thesis differentiates between offense and defense possessions, per team, per half, which allows for a more complex strategic analysis of a team's behavior when in possession versus defending. This differentiation leads to the slicing of periods, translating into two-minute periods. For example, to find a two-minute sequence, a 50 second period from match-minute one (real-time) might add the next time the team is in possession in minute three of 70 seconds to form a total of 120 seconds in possession and therefore the first offensive sequence. A team's possession determines the next period that eventually adds up to two minutes in total. Possessions change rapidly during any given match, so finding a continuous two-minute window is nearly impossible. Figure 6 explains the process visually for two arbitrary teams.

The algorithm discards meager possessions under five seconds since they might introduce unnecessary noise into formation analyses while teams might be scrambling for the ball.

Step 2: Normalize the positions:

Location normalization becomes necessary to compare similar formations that do not occur in the same field region. The center of a team's structure shifts to the center of the field, and therefore every point moves along the parallel distance vectors between the formation's and the playing field's center. Figure 7 provides a graphical illustration of this processing step.

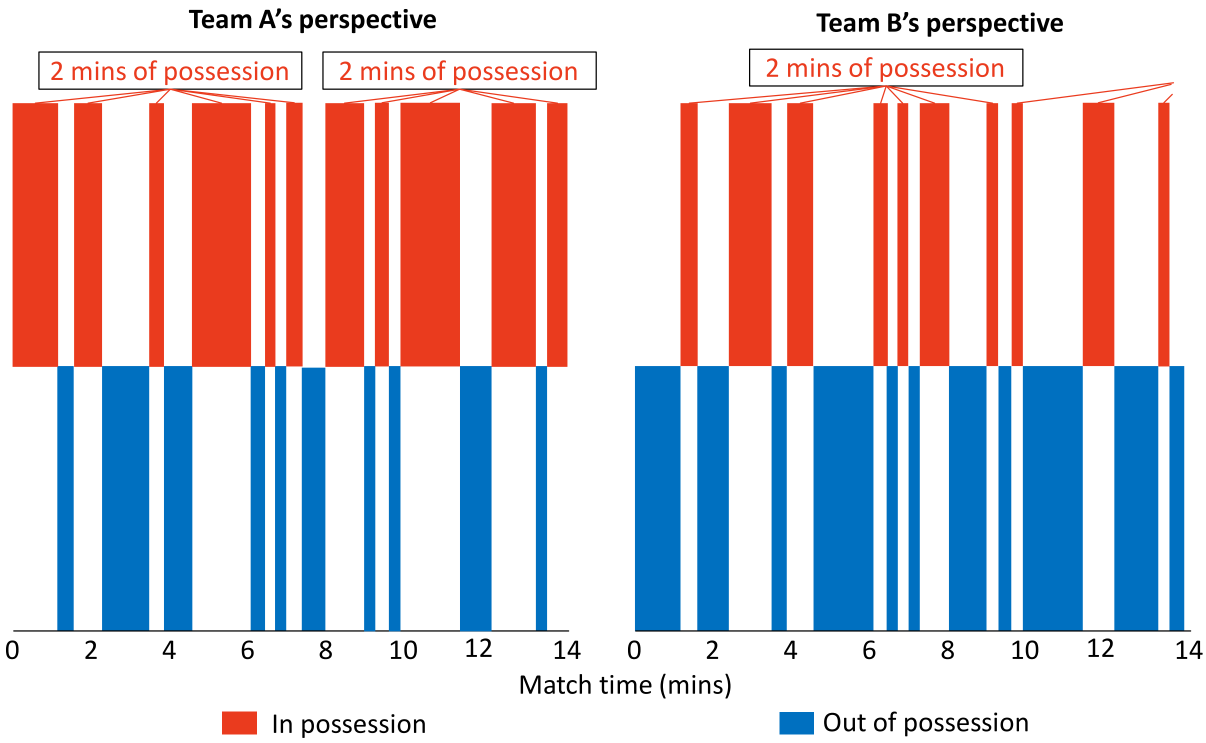


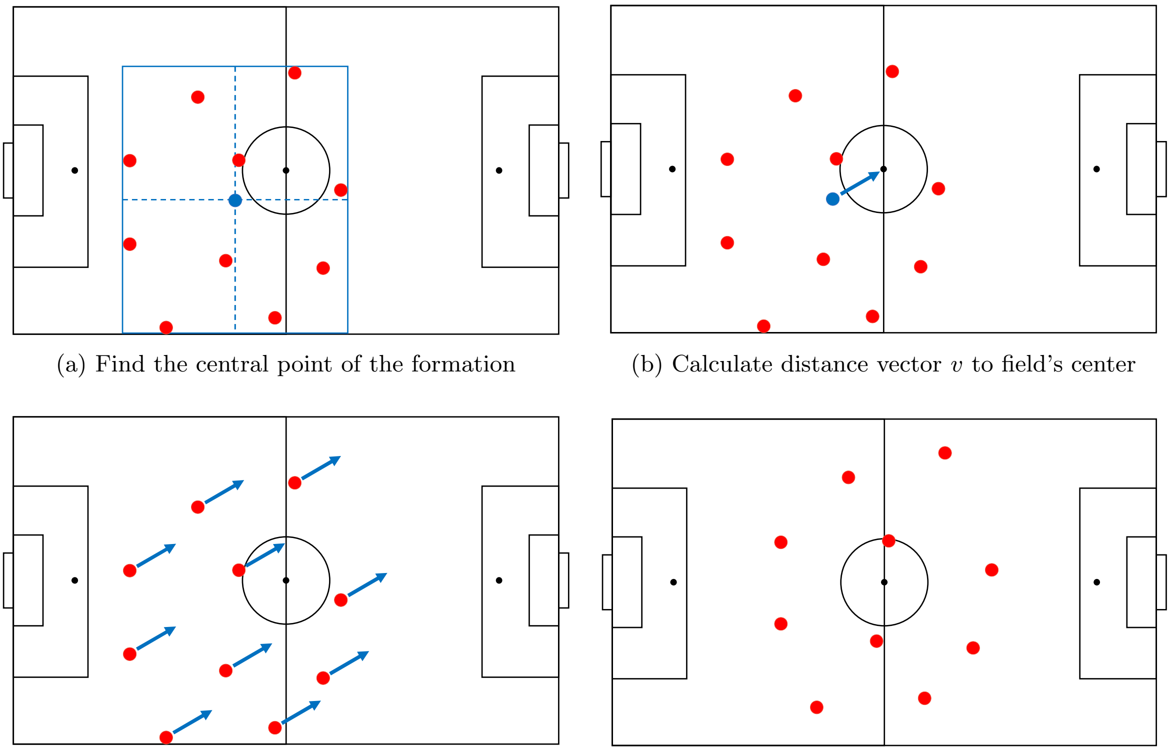
Figure 6: Separate periods of either in possession (for offensive) or out-of-possession (for defensive) times form two minute windows, over which to average the position for formation analyses. The mirrored view illustrates the binary nature of a two-team game.

Step 3: Find average position per sequence:

Every sequence contains 120×10-time frames, therefore 1,200 ten-dimensional arrays (for the ten players) per team. To find the average position per team, we can use the arrays' ordered nature per sequence. This structure means that the player at frame one, index one, corresponds to the player at frame two, index one. A simple mean over the x- and y-coordinates displays a player's average position throughout a two-minute segment. Challenges of finding the average formations of players who might switch roles (see Figure 5) are alleviated because empirical evidence supports that players will not change roles frequently within two-minute segments. While this complication might impact a single sequence, the entire data set subset (see next point) often includes hundreds of sequences, which cancels out the noise. The top of Figure 5 demonstrates the averaging of positions per sequence for an exemplary three-player setup without role-switching.

Step 4: Find the average position per subset data:

The mean-position calculation's main challenge is to derive the average formation of a specific subset of sequences with unordered arrays. If we just ignored the ordering, the average calculation will fail to detect players' switching roles and substitutions, which throws off the positions' order and leads to nonsensical formations. To define the formations without any required as-



(a) Find the central point of the formation

(b) Calculate distance vector v to field's center

(c) Add distance vector v to every point

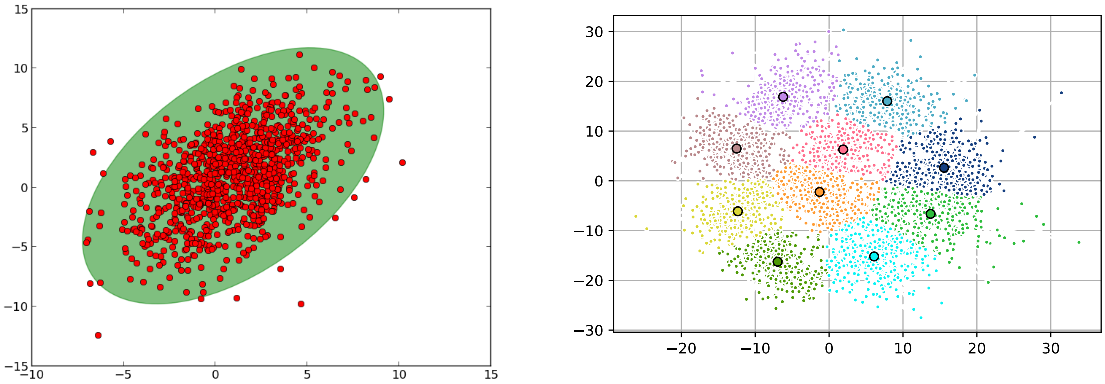
(d) Formation is centered around middle circle

Figure 7: This illustration depicts the position normalization method. First (a), 0.5 of the minimum and maximum x-/ y-coordinates labels the central point of the formation. Second (b), the distance vector v describes the relative distance from the this average point to the center of the field (middle circle). Third (c), the distance vector v shifts all relative positions by simple vector addition. (d) shows the final formation once it is centered around the middle circle.

sumptions for expected formations, point distributions, or accuracy limitations, a k-means[26] clustering algorithm will find the ten centroids for any cloud of points provided by the number of sequences that fall into a specific use case category. Figure 8b highlights this procedure for a random offensive sequence, which lies at the heart of most average spatial calculations throughout this thesis.

The use cases of subset data could contain all sequences of a specific team, or a specific team-opponent combination, or of match situations, in which a team was ahead/behind/even, et cetera. It is essential to state that every clustering algorithm introduces unique trade-offs for specific use cases. The advantages of k-means offer robustness to outliers and fast performance, which overshadow its tendency to convergence at a local optimum. The algorithm affords the derivation of which player is most likely to have played on which position, which is a direct implementation of algorithmic requirement ▦ **A2** ***Inference About Individual Players***. Chapter 8 discusses some of these challenges and trade-offs with potential alleviations in extension to this thesis.

[26] The **k-means clustering** algorithm provides an unobserved and iterative approach to determine clusters of beforehand unseen data. It begins by moving a pre-determined number of centroids (k) to classify the clusters' observations closest to them, moving the centroids iteratively, and reclassifying the points to find the minimum squared Euclidean distance. This article or Appendix A provide a more detailed explanation.



(a) Spatial-error-ellipse derivation

(b) K-means for single sequence

Figure 8: Illustration of the main parameters calculated for each average position per subset of data. For the sequence data this might refer to a two-minute sequence, while for conditional or event-driven subsetting, these time periods might be averages across multiple matches modes and game situations. Figure 8a highlights how spatial positions per time frame are then utilized to find the average area a player moved over an entire time aggregation. Figure 8b shows the result of the final centroid location for players, which then the front-end interprets as the average position for the given player.

Higher statistical moments of distributions explain aspects that a simple mean-calculation[27] might leave out. The spread (variance, or second statistical moment), skewness (third statistical moment), or measures of outlier-relevance (kurtosis, or fourth statistical moment) offer delicate insight into the data at hand. In this case, the deviation from the average position, measured by the variance, is instructive to display the degree and direction of player movements from the mean position. Since the point distributions follow a Gaussian distribution, the description of the first two statistical moments, fully describes the distribution.[28] The illustration of error-ellipses will help to visualize this deviation. The algorithm calculates a height-, width-, and θ-parameter to indicate a player's general movement around its mean. This visualization allows grasping a predefined standard-deviation ellipse around the mean to, first, hint at the convergence of the data at the mean spot and, second, allow for an interpretation of how strictly a given player remains at her assigned position and where she deviates. The calculation follows a simple derivation of the standard deviation and the covariance matrix calculation of the points averaged. The described formulas for width w and height h are

$$w = 2 \times n \times \sigma \times \sqrt{\lambda_1},$$
$$h = 2 \times n \times \sigma \times \sqrt{\lambda_2},$$

where (λ_1, λ_2) are the eigenvalues of the covariance matrix of the point data; n is the standard deviation set for the ellipse (e.g.: n=2 to find the two-sigma ellipse, which is used in these illustrations). This calculation solves algorithmic requirement **A4 *Derive Calculation (In)accuracy***.

[27] The mean of a distribution represents its first statistical moment.

[28] The *probability density function* of a Gaussian distribution is described by $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, with μ as the mean of the distribution and σ as its standard deviation. Since σ^2 equals a functions *variance*, i.e. its second statistical moment, the probability density of a normal distribution is fully characterized by its first (mean) and second (variance) moment.

4.2 Clustering Formations

The highest level analysis of this thesis is to find underlying patterns for all formations of the data. For this purpose, a similarity measure called the *Wasserstein distance* affirms how different two positional distributions are. Inspired by Shaw and Glickman [66], a two dimensional Wasserstein (Wasserstein-2) metric [58] will be adopted to cluster formations. In line with previous work [66], we assume the distributions of formations to be Gaussian, which simplifies the calculation. The square of the Wasserstein metric simplifies to

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

with means m_1 and m_2 and covariance matrices C_1 and C_2. For a more detailed derivation, refer to Appendix D.

Shaw and Glickman [66] describe the Wasserstein metric eloquently as:

> *"For point particles, the Wasserstein distance is simply the square root of the L2 norm of the difference between the means. More generally, the Wasserstein metric is a solution to the optimal transport problem [84], i.e., an estimate of the cost of moving from one distribution to another.“* (Shaw and Glickman [66], page 6)

This characteristic earned the Wasserstein metric the more poignant name of *earth mover's distance*, which imagines distributions as a pile of earth and the distance equal to the work of changing one pile to the shape of another. Additionally to this quick introduction, Appendix D, and more visually, Figure 48, attempt to provide a general intuition.

Once distributions are comparable, an agglomerative clustering[29] approach pairs similar distributions together.

By utilizing this metric, the clustering's objective function to minimize becomes the square sum of the Wasserstein distances

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

where $D_{i,j}$ is the cost (square of Wasserstein distance) of matching player i in formation one to player j in formation two, and $X_{i,j}$ is a player-player allocation matrix, in which each element is equal to one if player i is matched to player j, and zero otherwise.

4.3 Scaling Formations

For certain aspects of the web application, training the scaling parameter k comparably implemented approaches in [66] might be the preferred method. This training allows for better

[29] **Agglomerative Clustering** describes one of the two major branches of hierarchical clustering techniques. Its "*bottom-up*“ logic begins with every observation as its own cluster to hierarchically group similar observations together until only a single cluster remains. The user can then choose via statistical inference or visual inspection of a *dendogram* the most effective clustering granularity.

comparison since the algorithm finds the k-parameter that stretches or shrinks formations in their compactness to make them most comparable. For certain aspects of user interaction that do not build on automatic comparisons, scaling of the formation will become a crucial feature for the user to interact with the data. Therefore, an important addition to the cluster component illustrations represents either stretching out the data over the entire field or scale it by the most extreme points of the entire data set. Figure 9 exemplifies the approach for two common formations with hypothetical data points.

Intuitively, the algorithm calculates two alternative mappings of the visualization points: either use a formation in isolation as a reference. Its most-left point scales to the most-left point of the displayed domain, its most-top point to the most-top of the field, et cetera. This approach stretches the formation, regardless if it is a comparatively dense or wide formation. The alternative scaling dynamically programs the visualization domain based on the entire range of values in the subset data. The most-left point on the domain corresponds to the overall data's smallest x-value, but not necessarily of the currently chosen sequence. This scaling shrinks most sequences because the extreme points of all formations dominate the visualization perimeter. This feature might hide some details about players' exact position in relatively dense formations but allows for a compactness comparison between formations throughout the full data. Figure 9 exhibits this relationship between two scaled formations. It displays the convex-hulls[30] of the formations. In this example, red stands more compactly than blue, which remains invisible if the stretched scaling option is chosen (wider and lighter formations).

4.4 Data

This thesis analyzes a large dataset to address an ambitious research question. The data contains mainly meta information of matches in a top-European soccer league, most notably the players' positions and the ball (2D, x- /y-position on the soccer field) and relevant events. The position data is available in 100-millisecond resolution. Event data, in general, usually describe special situations around the soccer ball. These events include, for example, when a pass, a shot, an offside, or a foul occurred. The information is stored in a PostgreSQL[31] database, maintained by the data analysis and visualization group at the University of Konstanz.
The following list provides useful general facts about the data. The data contains:

- 101 matches of the 2018/2019 and 149 matches of the 2019/2020 season for a total of 250 matches.
- 328,930,768 positions in the *tracking* table.
- 613,378 events in the *events* table.
- ten additional data tables (for a total of twelve tables in the database) to enrich the tracking/event data with, for example, player-specific information (*players* table), location data (*stadiums* table), or general team meta data (*teams* table).

[30] A **Convex-Hull** represents the smallest closure—the area around the data points for 2D cases—that contains all points.

[31] **PostgreSQL** is one of the most popular providers of SQL (relational) databases. See this link for their official website.

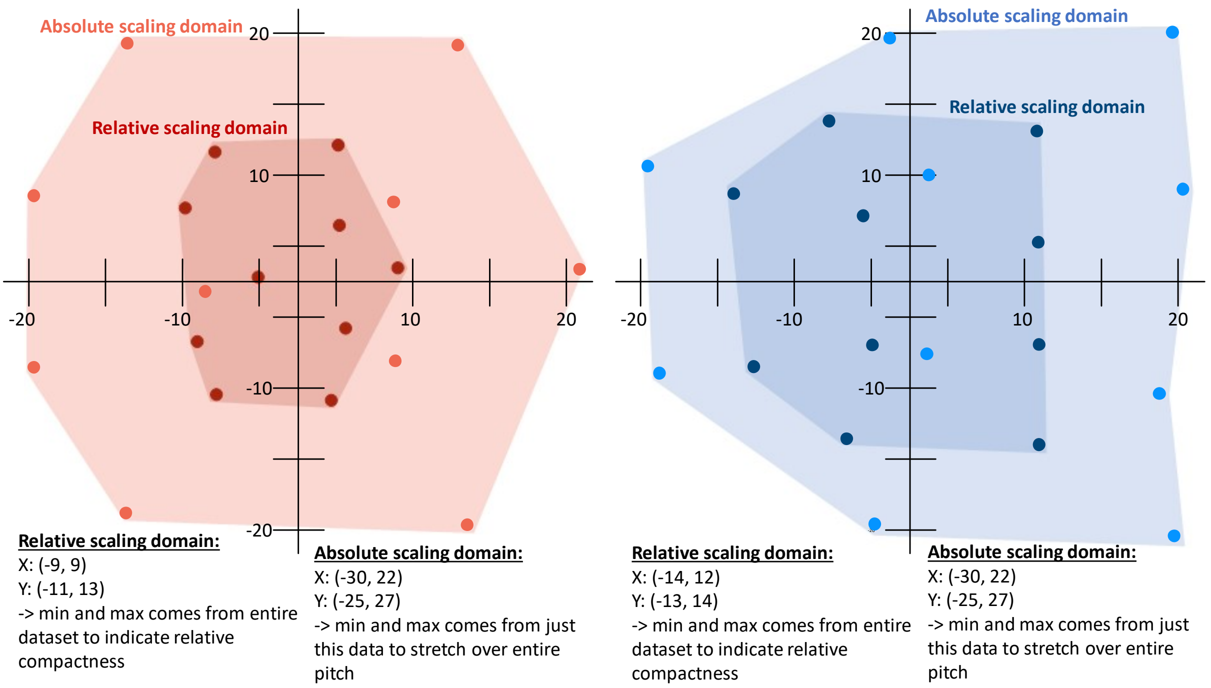


Figure 9: The two scatter plots exemplify common 4-1-4-1 (left) and 4-4-2 (right) formations and how their visualization is impacted drastically from the respective data domain chosen (Both graphs show the formations in different directions, first from left to right, the second one opposite to that). While an absolute domain, incorporating the entire datasets' values scales the relative formations according to their compactness, the relative scaling might allow for a better comparison of actual formations regardless of the compactness measure's noise. The numbers are fictitious but resemble real data domains from the actual data. We can see that the right formation lies inherently closer to the actual maximum for the field layout, so the data is not as thoroughly stretched compared to the red data, which represents a naturally denser formation and therefore experiences more stretching across the field.

The thesis builds heavily on deducing meta information, such as the score or possession from the events data table. Therefore, a more thorough description of these events' nature will help the reader gain a better understanding of the techniques employed. Chapter 2.1 offers some general insights, and the following definition will describe the specific event data used in this study. This description borrows from a previous description of the dataset [69]. Since the 2020 analyses of the dissertation [69], the event database incorporates 31 but 46 event types. This thesis introduces two additional event classifications to, first, better capture the additional event types, and, second, allow for an overall finer event classification granularity:

- **Rule-induced events** are events that occur as a result of the match rules. For example, if the ball passes the sideline of the soccer field, it has to be thrown in again by the opposite team.
- Events tagged with **prosecution** indicate penalization of illegal behavior by the related player(s).
- **Player interactions with ball** contains events that happen when a player is touching the ball. Almost every event that gets tagged falls under this category besides yellow and red cards, the end of halftime, pure ball interactions, and a substitution.

- **Ball interactions with environment** are situations in which, for example, a deflection of the ball, without a player directly intervening, occurs.
- **Player interactions without ball** marks a small cluster of events, where either multiple players move together (maybe for offside or a foul situation) or interaction of a player with her environment.
- Events that interrupt the match get marked as **gameplay interruption**.
- If an event has a direct relation to scoring (e.g., a shot on the goal), we mark it as **scoring related.**

This categorization gives rise to Figure 10, an extension to the existing previous version of this figure (Stein [69], Figure 2.1.1) displaying all included events categorized into the earlier mentioned groupings.

4.5 Data preparation

The last remaining step before feeding the data into the formation algorithms is to pre-process it and reshape its format to derive meaningful information. The data itself only provides individual positions for a given time frame in a two-dimensional (x-/y-coordinates) space. How these individuals react in comparison to one another is not explicitly included within the data. This chapter will offer a transition from the first part of the methods chapter, which explains the algorithmic logic over the database's data structure to the actual data utilized for the formations.

Data querying:

A database table called *tracking* stores the complete tracking information, meaning the x-/y-coordinates for a given player for every tenth of a second for the majority of matches of the 2018/2019 and 2019/2020 seasons of an elite European soccer league. This table provides the core information of this thesis' analysis. However, for several reasons explained in detail in this chapter, we will need to enrich it with event data. The *events* table highlights a variety of crucial interactions during a game. It contains the *actorid*, i.e., the unique player identifier, for an event, such as a pass, a shot on the target, or a foul, or no *actorid* for events that are unassignable to a specific player, such as *match-start.* The event information also includes the exact time of the event in hundredths of a second. The query first retrieves both tables individually and subsequently merges them to subset and enrich the data, before adjacent tables, including player details,[32] match information,[33] and team information[34] enrich the raw data. Only information for the ten outfield players, excluding the goalkeeper, is retrieved, and the thesis excludes matches with a suspension due to a red card entirely.[35] This heuristic allows for an additional layer of security against wrongly incorporating formations with less than ten field players into the classification. The data also corrects the direction of positions (from left to right, or vice versa) to ensure comparability between home and away games and directional changes between first and second halves. Finally, all games' information is queried for a loop

[32] Firstname, last name.

[33] Home and away team, plus who started on which side.

[34] Full team name.

[35] In comparison to just excluding the sequences, during which less than ten field players are playing.

rule-induced events	prosecution	player interactions with ball	ball interaction with environment	player interaction without ball	gameplay interruption	scoring related	Event Type	Description
■	■	□	■		■	■	Foul Penalty	Free kick on the goal defended only by the goalkeeper
■	■	□	■		■		Foul direct free kick	Free kick that is allowed to be directly shot into the goal
■	■	□	■		■		Foul indirect free kick	Free kick that is not allowed to be directly shot into the goal
■	■	□	■		■		Foul throw in	Throw in that is not correctly executed
■		□	■				HalfTime Start	First or second half starts
■		□	■	□	■		Offside	Player is in an offside position
■		■	■		■		Out for goal kick	Ball passes the endline after an opponent touched it
■		■	■		■		Out for corner	Ball passes the endline after a player from the own team touched it
■		■	■		■		Out for throw in	Ball passes the sideline of the soccer pitch
		□	■		■	■	Goal	Awarded when the whole of the ball crosses the whole of the goal-line
		□	■		■	■	Own goal	Awarded when the whole of the ball crosses the whole of the goal-line of own goal
		■	■		□	■	Shot on target	Any shot attempt that would or does not enter the goal if left unblocked
		■	■		□	□	Shot not on target	Any shot attempt that would or does enter the goal if left unblocked
		□	□			■	Chance	Potential for shot on goal situation
		■	■				Pass	Ball touch from one player with direction towards a team mate
		■	■				Reception	Ball touch made by the player after receiving it from another player
		■	■				Clearance	Hard ball touch where the player tries get the ball away from the current zone on the pitch
		■	■				Hold of ball	Play action when the keeper takes the ball with his hands without danger
		■	■				Running with ball	Used by the player to move the ball around without passing it to another player
		■	■				Cross	Hard ball touch where the executing player is positioned in the final third of the field
		■	■				Neutral contact	Characterized by ball touch which is difficult to control
		■	■				Pass assist	The last pass to a teammate in a way that leads to a goal
		■	■				Cross assist	The last cross to a teammate in a way that leads to a goal
		■	■				Catch	Keeper catches the ball and hold it in his hands on a dangerous situation
		■	■				Catch drop	Keeper does not manage to hold the ball but lets it bounce of the hands again
		■	■				Drop of ball	Keeper drops the ball after having caught it or holds it in order to play the ball
		■	■				Punch save	Keeper punches the ball with his hand away
		■	■				Punch	Keeper punches ball
		■	■				Diving save	Keeper jumps to a side to catch the ball
		■	■	□			Diving	Keeper dives for ball
		■	■				Neutral clearance save	Keeper kicks off ball to avoid goal
		■	■				Neutral clearance	Keeper passes long pass to field players
		■	■				Catch save	Keeper catches the ball in mid-air without deflection.
		■	■				Catch drop save	Keeper drops a caught ball
		■	■				Drop kick	Kicking a ball that is dropping to the ground as it starts to bounce up
■	■			□	□		Yellow card	Displayed by referee to indicate that a player has been cautioned for a foul
■	■			□	■		Red card	Displayed by referee to indicate that a player has been dismissed from the field for a foul
■					■		End of Half	First or second half ends
■			□	□	■		Interruption	Any other form of interruption
■					■		Substitution	Replacing one player with another during a match
				□			Modification of position	Player repositions with or without ball
			■				Right goal post	Ball hits a right goal post
			■				Left goal post	Ball hits a right goal post
			■				Crossbar	Ball hits a crossbar
		■	■				Block	Ball hits player in the form of a deflection
			■				Other obstacle	Ball hits any other obstacle

■ applies always □ optional

Figure 10: Event-types grouped into seven categories and into 'applies always' and 'optional' subgroups. This figure extends Figure 2.1.1 in Stein [69] by introducing two additional event-categories and updating the total number of labeled events from 31 to 46.

over a team's full game list. The query includes goal information from the events table and the location (stadium name) to allow for prospective subsetting of conditional and event-driven analyses (e.g., if a team was ahead or behind) and graphical illustrations for front-end functionalities.

Data processing:

Data subsetting for any meaningful interpretation requires cleaning the raw data first. We need to know when a team is in possession and find instructive time frames to bundle for formation calculations.

The first step includes finding the active times for a match: every game in soccer includes so-called *dead periods*, during which the clock keeps running, but the actual match play is interrupted. Examples include a free-kick, the kick-off, or just a throw-in event. To clean our analysis of formations during these periods, we need to define a metric to assess active periods during a game automatically. The event data will help find these active times: the events table includes information about an event's start and end of a phase.

The next crucial step is to determine who is currently in possession of the ball. A naive approach could label a team in possession if its players are closest to the ball. This heuristic leads to massive complications since a pass usually passes by players closest to the ball who are not directly in possession. We use the event data to find events that uniquely identify an actorid for a given event and determine the possession from the event's nature. For example, a pass by a specific actor allows us to know that, at that moment, the player had the ball. Unless another event occurs, the team will remain in possession. Therefore, we can deduct sequences of times for a given team in possession—here, a heuristic for the binary nature of possession in a two-team game helps—if one team is not in possession, the other one, by assumption, is in possession. By translating these sequences to the corresponding tenth of a second, we can merge the possession information into the enriched tracking data set. This logic excludes certain events from any inferential value to the possession heuristic. For example, a yellow card can occur to any player on the field without offering insights on which team is in possession. The excluded events are (see Figure 10 for the full list of events): 1. *Modification of position*, 2. *Catch drop save*, 3. *Catch drop*, 4. *Red Card*, 5. *Yellow Card*, 6. *Diving Save*, 7. *Punch Save*, 8. *Foul - Indirect free-kick*, 9. *Diving*, 10. *Foul - Direct free-kick*, 11. *Substitution*, 12. *Right goal post*, 13. *Left goal post.*

Possessions of less than five seconds are removed from the sequence calculation. This processing leads to, on average, about four to seven two-minute sequences per half for an average total of about ten sequences per team per match. It is crucial to notice that this number only includes in possession sequences. Given the binary nature of soccer (if one team attacks, the other one automatically defends), the total number of sequences per game totals about 20 sequences per match, per team, or 40 distinct formation sequences per match. This number passes a first sanity check of overall active time aligning closely with half the total time of 90+ minutes of play (ten offensive + ten defensive two-minute segments equal about 40 minutes $\approx \frac{1}{2} \times 90$ minutes). Shaw and Glickman [66] arrive at a similar value of sequences per team.[36]

[36] The authors' dataset of 100 games leads to ≈ 4,000 formations, which agrees with our 40 sequences per match.

5 A System for Multi-Match Formation Analysis

This chapter describes the actual system implementation to solve the shortcomings of previous solutions. It will shadow the search/subsetting and visualization design requirements of Chapter 3 and connect the data available, as detailed in Chapter 4.4, with the use cases of everyday practitioners.

The overall design choice funnels from a wide tab into more granular analyses, offering the user various angles to inspect the current data. Generally, the application is split into three categories: (1) a *Clustering View* (Chapter 5.1), (2) a *Conditional View* (Chapter 5.2), and an *Event View* (Chapter 5.3). Figure 11 illustrates the interplay between these subsystems.

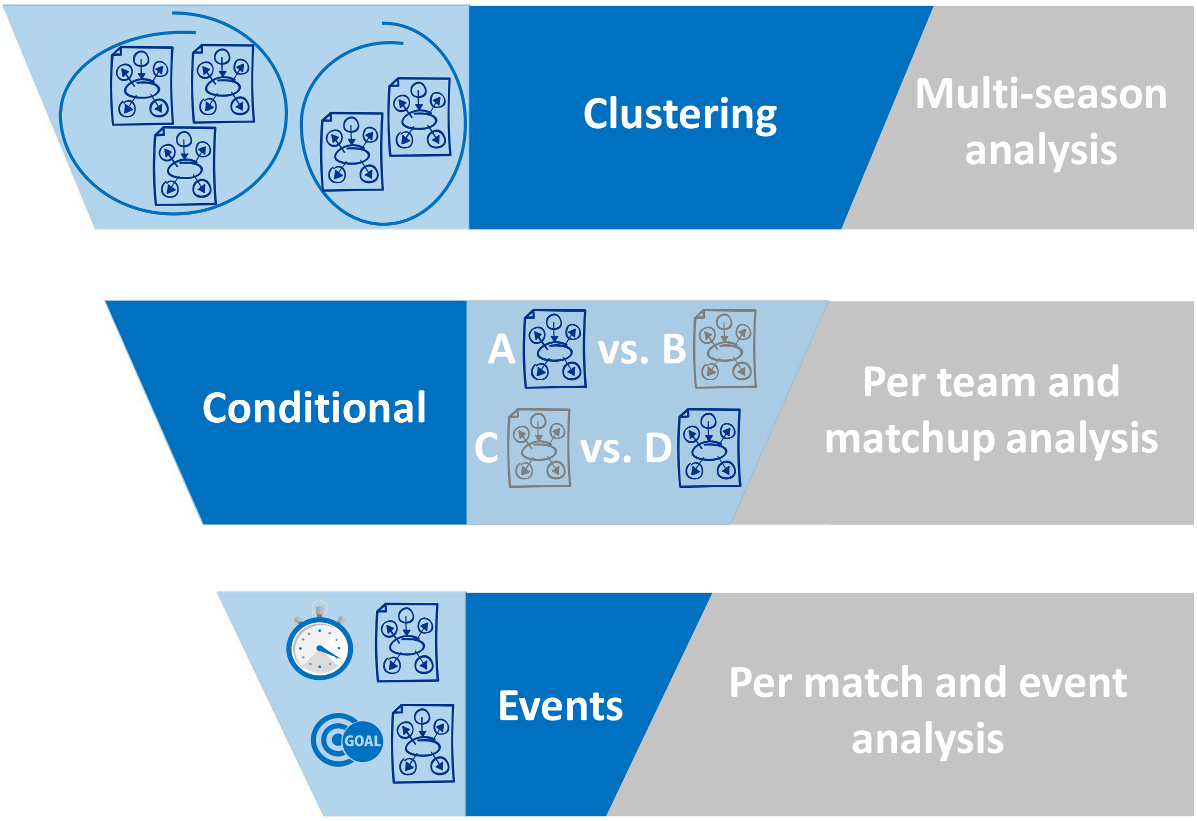
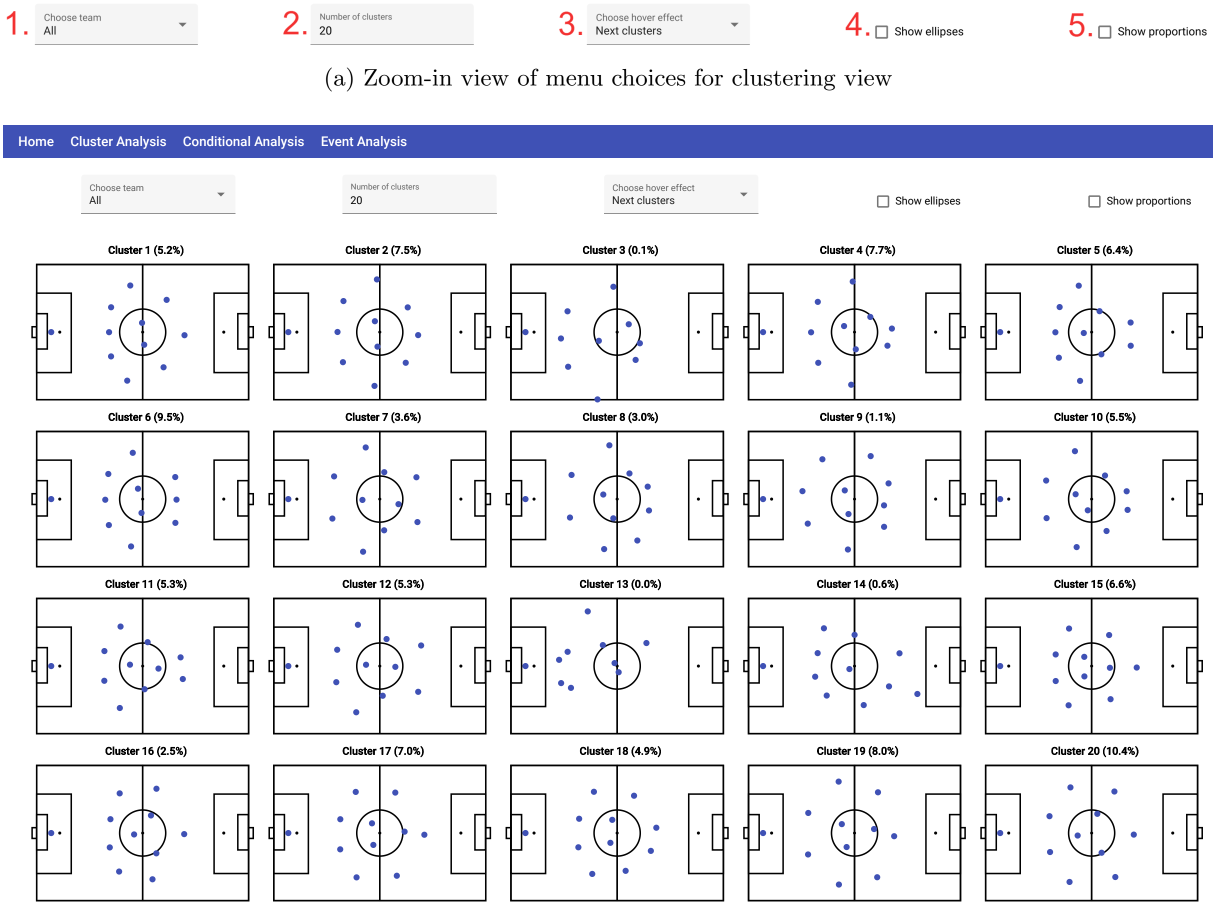


Figure 11: The systematic illustration displays how the three implemented views of the application represent a broader to finer analysis granularity. The system allows the user to first group similar formations across the entire data set in the clustering view, to analyze matchup-specific formations in the conditional view, and to finally analyze formations within a single matchup across different events in the events view.

The design decision introducing multiple distinct systems to allow the user the liberty to decide which specific formation information is most suitable for the current use case as practical analysis tool. It is in line with design requirements 🔍 S2 *Limitation to Relevant Options*, 🔍 S3 *Iterative Subsetting*, and 📊 V2 *Clear Separation of Concerns*. The full extent of possible applications is unpredictable. Therefore, the more flexibility the user is granted, the better the program can fulfill its role to supplement and facilitate the decision making process. A navigation bar at the top offers access to the individual systems.

5.1 Cluster View

The clustering view offers an overview of a specific user-defined number of clusters of formations. These groupings allow the user to identify overarching tendencies of patterns that define the formation structure. Almost 10,000 individual two-minute sequences (see Chapter 4 for derivation details) form the data-basis of the clustering algorithm. Figure 12 provides an overview of the simplest and most general version of the visualization.



1. Choose team: All
2. Number of clusters: 20
3. Choose hover effect: Next clusters
4. □ Show ellipses
5. □ Show proportions

(a) Zoom-in view of menu choices for clustering view

Choose team: All
Number of clusters: 20
Choose hover effect: Next clusters
□ Show ellipses
□ Show proportions

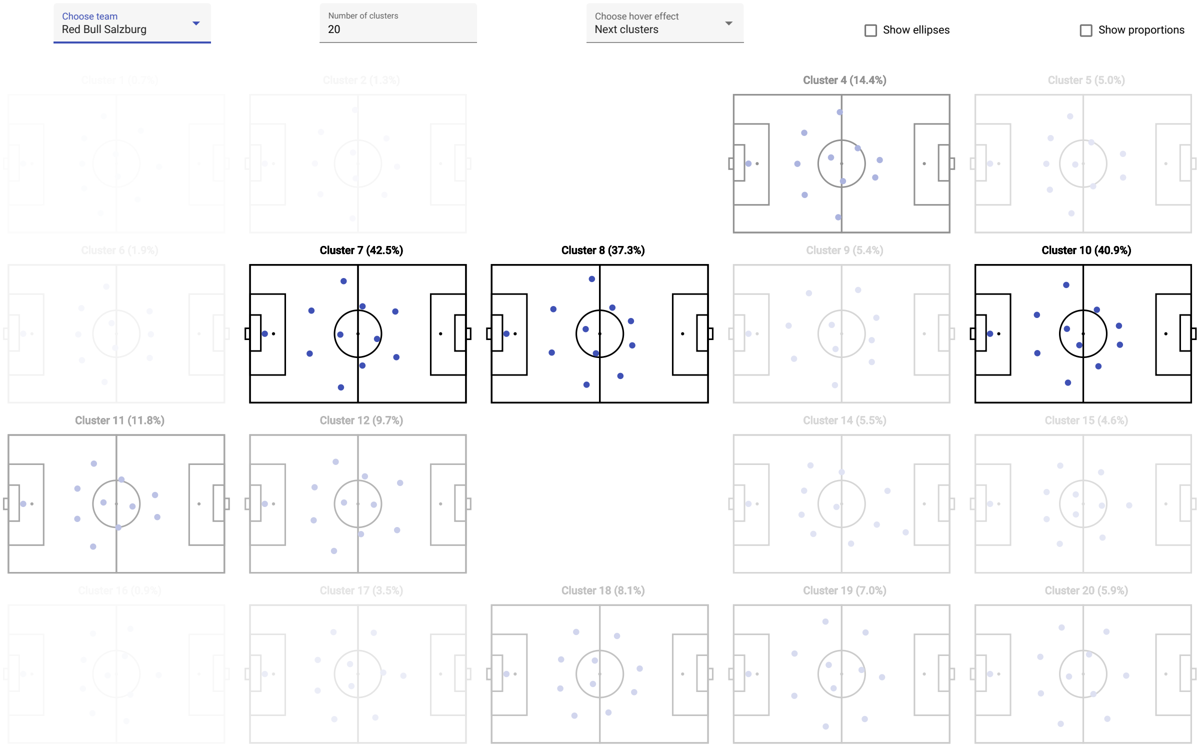
(b) Overview of clustering view

Figure 12: This illustration shows the clustering view without any sub-setting or additional visualization-features applied. It shows the default twenty clusters into which the formation sequences were grouped. There is no inherent ordering of the clusters. Additional information in the title contain the relative cluster size (share of formations included in specific cluster), and a running count for the total of cluster numbers.

The red numbers afford guidance throughout the sub-chapters of the clustering tool's explanation. While all of the effects can be selected and used in combination, the overall options focus on the most important information coaches utilize to group formations. This design choice supports S2 *Limitation to Relevant Options* and S3 *Iterative Subsetting*. For illustrative purposes, the following explanation will try to isolate them as much as possible and compare them to the base case depicted in Figure 12.

Option 1 Team selection:

The user can select either all teams or any teams individually as the visualization's primary focal point. The selection happens via a drop-down menu. Since clusters can theoretically contain each team's sequences, it is essential to highlight how strongly sequences of a specific team impact any given cluster. More intuitively, this visualization feature describes the proportion of sequences in a given cluster associated with the selected team.



Choose team: Red Bull Salzburg

Number of clusters: 20

Choose hover effect: Next clusters

□ Show ellipses

□ Show proportions

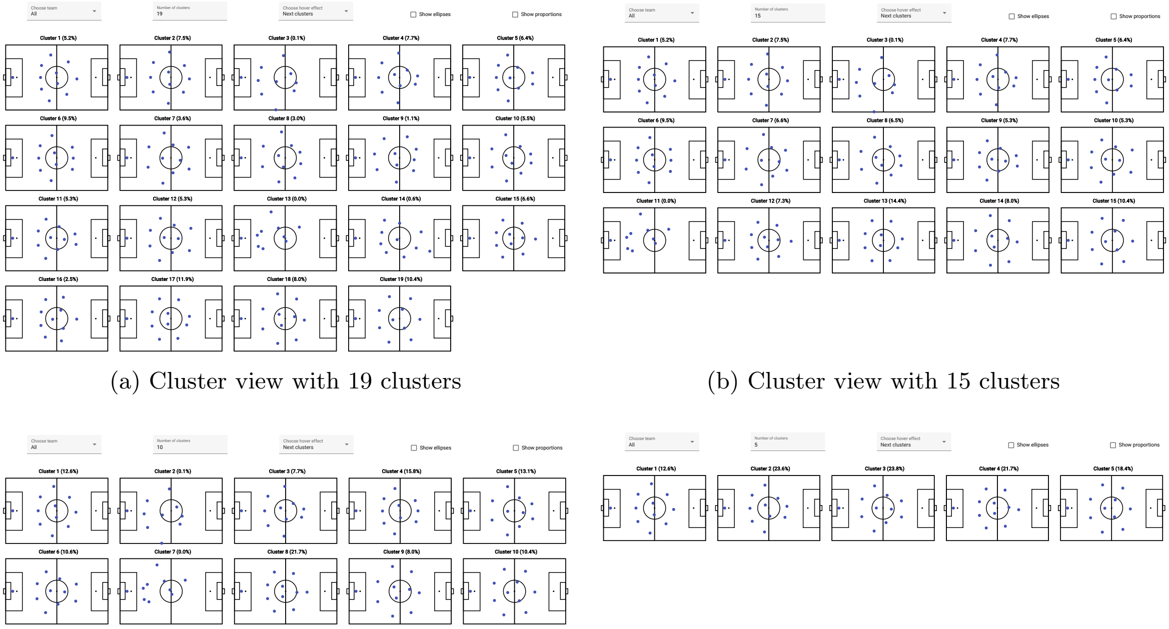
Figure 13: This illustration shows clustering view adjusted for a single team. The clusters are faded out if they do not include many sequences of the selected team and the cluster title contains the proportion of sequences within that cluster that are from the selected team. For example, 42.5% of the formations in cluster 7 are from "Red Bull Salzburg".

The opacity of the displayed fields is adjusted to correlate with the relative proportion. The title adjusts to show how the selected team's sequences impact the percentages of a cluster. Therefore, informally, the number in the cluster title aligns with the opacity of the cluster visualization. This reasoning mimics how coaches try to group teams in terms of similarity of collective movements. Which teams represent a similar playing style represents one of the most omnipresent questions in formations analysis. Therefore, the design choice supports the search/subsetting requirement 🔍 S1 ***Intuitive Design Choices***.

Option 2 Cluster-size selection:

The number of clusters determines the grouping process for the data. It represents a trade-off between robustness to outliers (with fewer clusters) to a more granular view of the data (with more clusters). Figure 14 shows an overview of some of the choices a user can make to select

the cluster size. All values from one to twenty are possible[37] and can be incremented via typing in a valid number or incrementing it via an up- or down-arrow next to the number field.
The values are pre-computed in the backend, which results in immediate visualization transitions, following **V3 *Fast Access and Performance***. This clustering interaction represents an extension to the clustering display in previous work [66] and limits the result presentation to twenty clusters without additional interaction possibilities to avoid information cluttering. The free choice to decide the best cluster size for a given data set underlines requirement **S4 *Easy Extensibility***. The system remains flexible for varying datasets, where the overall number of sequences determines the optimal amount of clusters.



Choose team: All
Number of clusters: 19
Choose hover effect: Next clusters
□ Show ellipses
□ Show proportions

(a) Cluster view with 19 clusters

Choose team: All
Number of clusters: 15
Choose hover effect: Next clusters
□ Show ellipses
□ Show proportions

(b) Cluster view with 15 clusters

Choose team: All
Number of clusters: 10
Choose hover effect: Next clusters
□ Show ellipses
□ Show proportions

(c) Cluster view with 10 clusters

Choose team: All
Number of clusters: 5
Choose hover effect: Next clusters
□ Show ellipses
□ Show proportions

(d) Cluster view with 5 clusters

Figure 14: This illustration shows how different cluster sizes determine the number of groupings displayed. The almost 10,000 formation sequences are clustered into the selected number of groups. The selection span ranges from 1 cluster (all sequences together in a grand average) to a maximum of 20 groups.

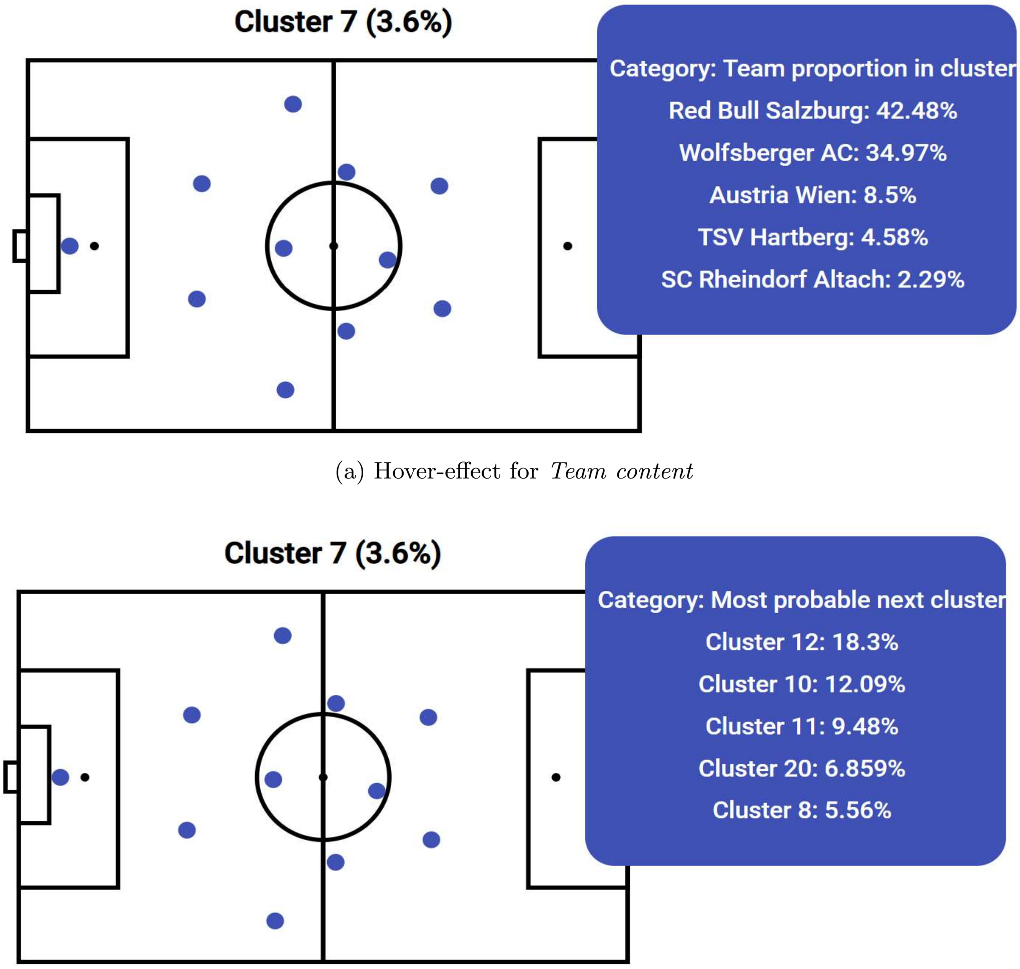
Option 3 Hover-effect selection:

Every field includes a so-called *hover-effect*, which displays an additional rectangle with information when the mouse moves over it. The rectangle disappears when the mouse leaves the respective field to anywhere but an adjacent field. However, if the mouse enters an adjacent field, the rectangle, also called *tooltip*, displays the information of the newly entered field. The two options for the tooltip information are *Team content* and *Next cluster*.

[37] Higher cluster numbers were tested but resulted in sparse clusters with just a single or no observations. Therefore, the maximum cluster number is twenty.

The decision to incorporate a drop-down menu to choose the preferred information follows search/subsetting requirement 🔍 S4 ***Easy Extensibility***, because additional tooltips are quickly developed and implemented.

Team content offers the user more insights into what kind of teams mostly make up the respective cluster's content. This tooltip expands on the user's information when just a single team is selected and the opacity shifts.



Category: Team proportion in cluster
Red Bull Salzburg: 42.48%
Wolfsberger AC: 34.97%
Austria Wien: 8.5%
TSV Hartberg: 4.58%
SC Rheindorf Altach: 2.29%

(a) Hover-effect for *Team content*

Category: Most probable next cluster
Cluster 12: 18.3%
Cluster 10: 12.09%
Cluster 11: 9.48%
Cluster 20: 6.859%
Cluster 8: 5.56%

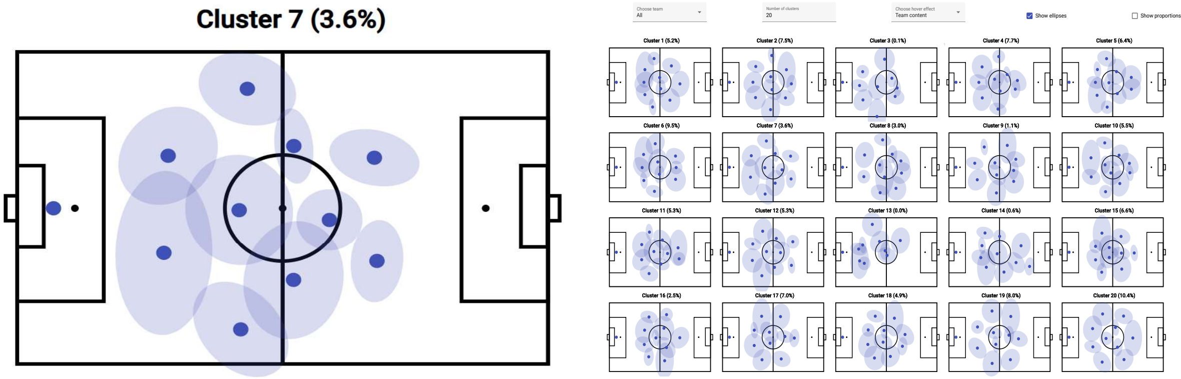
(b) Hover-effect for *Next cluster*

Figure 15: The two options for the hover-effect determine what information is shown in the tooltip when one of the clusters is hovered with the cursor. Figure 15a shows the listing of the top five teams in the cluster in descending order for the Team content option. Figure 15b highlights the most probable next cluster for any sequence contained in the respective cluster (here cluster 7) for the Next cluster option.

Next cluster affords the integration of a temporal dimension to this static analysis. The option addresses the question of a cluster's most probable next formation. Hence, the two-minute sequences in a cluster are analyzed to provide information about the most likely prospective sequence.

Option 4 Spatial-error ellipses:

Spatial error-ellipses introduce a notion of uncertainty into the otherwise absolute formation calculation. Players rarely stand still and this deviation from their mean (potentially assigned) position forms the ground for exciting discussions between coach and player. The ellipses form a two standard deviation (see Chapter 4.1 for detailed calculation instructions) visualization of how the players move around the calculated position. Figure 16 illustrates a zoomed-in view of a single cluster and the comparison to what the entire visualization displays when the ellipses-option is selected.



Choose team: All
Number of clusters: 20
Choose hover effect: Team content
☑ Show ellipses
□ Show proportions

(a) Spatial error-ellipses for single cluster (b) Overview of spatial error-ellipses for all clusters

Figure 16: The user can decide to show the spatial error ellipses of every position in a cluster. It provides a overview of the spread of the points for the calculation of the mean position. It therefore offers insights to the calculation accuracy. The ellipses afford a natural interpretation to deduct a player's relative movement around the mean position.

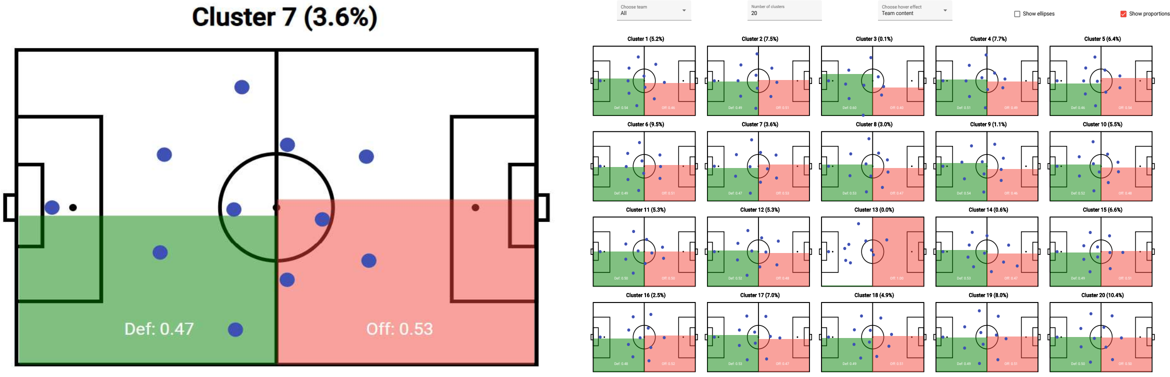
The user gains an additional dimension of information, and the spatial deviation supports the user's understanding of the certainty of single positional calculations. The choice for ellipses directly embedded into the visualization follows 📊 **V1** ***Familiar Result Presentation Patterns***. Alternatives, such as statistical significance measures or more advanced visualization techniques, might offer a marginal increase in conveyed information. However, users— representing a non-technical target audience for this system—will lose a sense of control for visualizations that are not immediately obvious and interpretable.

Option 5 Offense-/ Defense-proportions:

One of the significant challenges of the cluster view is the *black box*-connotation[38] it conveys. The user might need a high-level overview but does not know much of the clusters' actual content. To enlighten the division of formation sequences further, the user can visualize the offensive and defensive proportion in each cluster to find significant differences in formation in and out of possession.

Once again, the design choice of simple bar charts follows the design requirement 📊 **V1** ***Fa-***

[38] A **black box** in scientific computing, and engineering refers to a system that only provides input and output information without transparency to its inner workings.



Choose team: All

Number of clusters: 20

Choose hover effect: Team content

☐ Show ellipses

☑ Show proportions

(a) Offense / Defense proportions for single cluster

(b) Offense / Defense proportions for all clusters

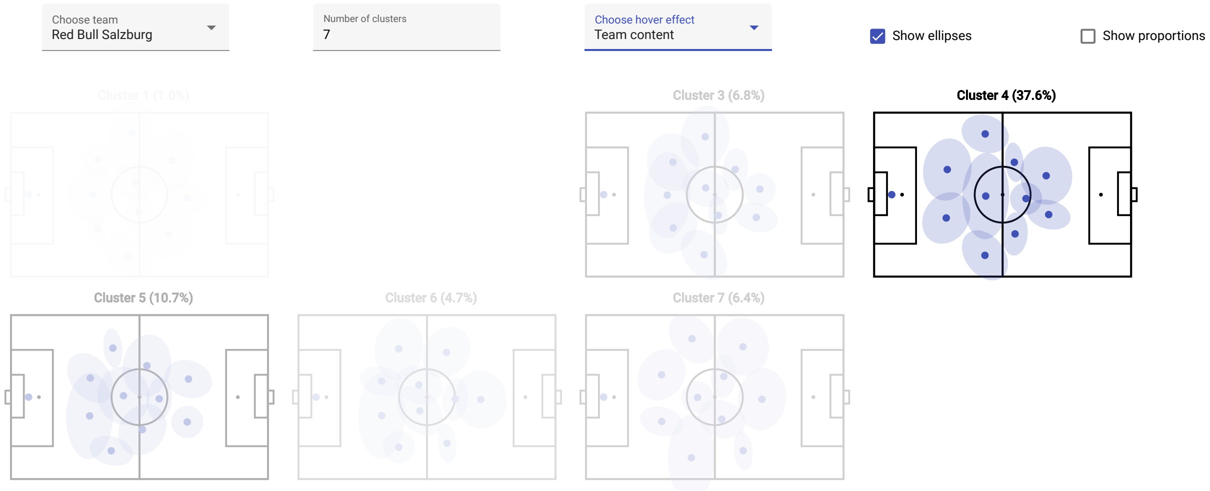
Figure 17: To highlight the respective proportions of offensive and defensive proportions within a cluster, the user can show these embedded in the figure as bar charts. The extreme clusters with 100 % offensive players included refer to outliers, with less than three observations in them.

miliar Result Presentation Patterns. Therefore, a cluster is easily distinguishable for its more offensive or defensive structure, which offers insights into a team's overall playing style. Subsetting the cluster view to a specific team (see Option **1**) will adjust the bar charts of the proportions of the specific team's sequences in a cluster. Alternatively, to put it more concise, if a specific team is selected, the bar chart will show the constellation of the team's sequences for each cluster. The left bar chart in **green** indicates relative amount of defensive formations. Since the colors overlap with the **blue** player points, **green** became the obvious solution to avoid cluttering the dense information. The visualization mirrors across the halfcourt line for the proportion of offensive formations in **red**. Both bar charts are integrated into the soccer field to avoid unnecessary space and introduce a novel visualization utilizing the natural geometric shape of the 2D view of the playing field. The graph includes the exact values of defensive and offensive proportions written in white inside the bar chart to alleviate the necessity to guess the values from the diagrams' height.

Combined subsetting:

Figure 18 demonstrates an exemplary query to illustrate the degree of complexity that the combination of the seemingly simple Options **1-5** achieves. The hypothetical question could read "What are the seven most common formational clusters in the data and which of those are mainly made-up of formations by *Red Bull Salzburg*? Also, how much did the players deviate from their assigned positions?".

This simultaneously simple and iterative design follows the design paradigm to be just as complex as necessary, which builds directly on search/subsetting requirements 🔍 S2 *Limitation to Relevant Options* and 🔍 S3 *Iterative Subsetting*.



Choose team: Red Bull Salzburg

Number of clusters: 7

Choose hover effect: Team content

☑ Show ellipses

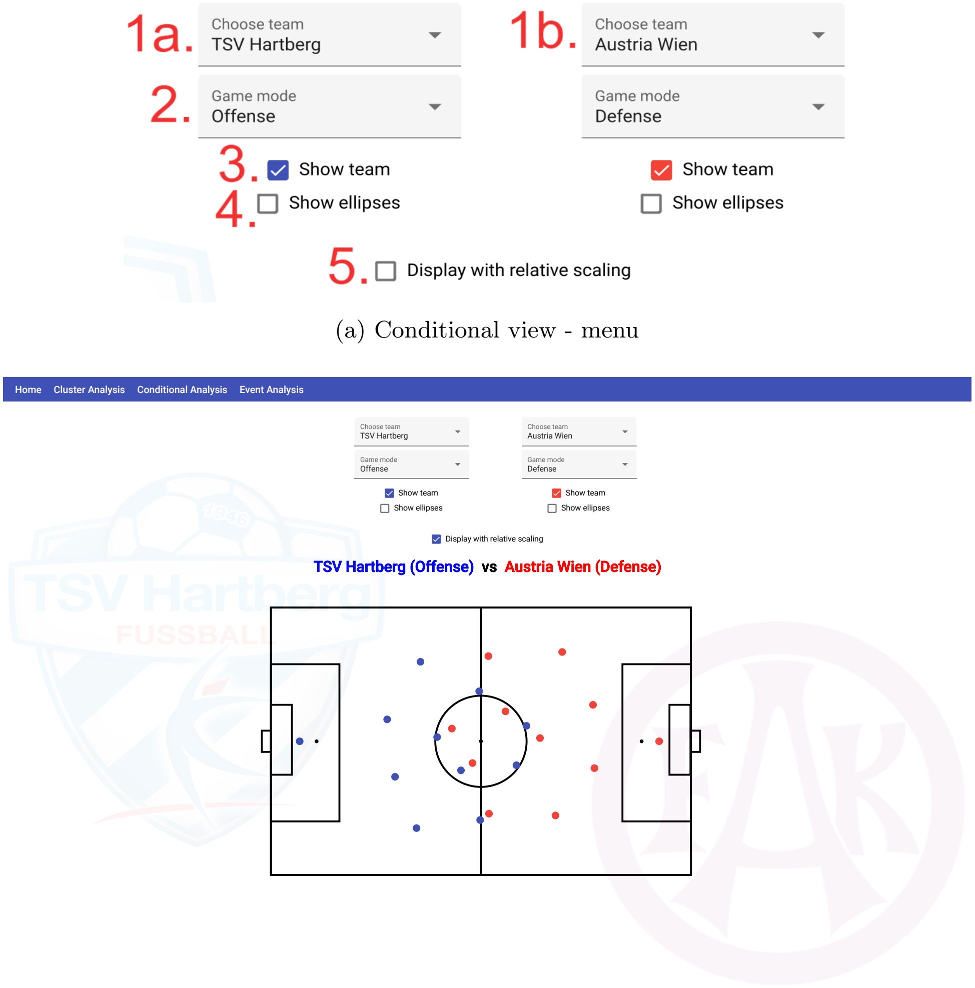
□ Show proportions

Figure 18: This example illustrates the power of a simple combination of single query options providing real insights to a hypothetical but realistic scenario. The simple visualization style allows for iterative subsetting, displaying the most common clusters for Red Bull Salzburg. The additional ellipses offers the user a sense of prediction accuracy embedded in the respective playing fields.

5.2 Conditional View

The conditional view allows the user to narrow in from the overarching question of the underlying tendencies for all formations to specific use cases. The focus lies on the team- and match-up-specific subsetting of the data. This option follows search/subsetting requirement 🔍 **S3** ***Iterative Subsetting*** to offer practical insights for match-day preparations.

The remainder of the chapter details the available options to either define the relevant data to display or change the visualization to include or reduce the shown information. This flexibility remains crucial for the usability and is in line with 📊 **V1** ***Familiar Result Presentation Patterns*** to mimic the interaction of a coach with a whiteboard. Figure 19 demonstrates the entire view for two teams. The top portion of the visualization allows the user to subset the data, while the bottom part visualizes the chosen formations on a single field. The mode selection reflects a common question of how offensive formations might respond to a defensive formation—mirroring a specific match situation. The background covers the team logos for an additional visual cue and improved user experience. The red numbers provide visual guidance throughout the chapter. These numbers are not part of the original system, and the figure hides team logos for an easier explanation of the individual visualization options.



1a. Choose team: TSV Hartberg
1b. Choose team: Austria Wien
2. Game mode: Offense
Game mode: Defense
3. ☑ Show team
☑ Show team
4. □ Show ellipses
□ Show ellipses
5. □ Display with relative scaling

(a) Conditional view - menu

Choose team: TSV Hartberg
Choose team: Austria Wien
Game mode: Offense
Game mode: Defense
☑ Show team
☑ Show team
□ Show ellipses
□ Show ellipses
☑ Display with relative scaling

TSV Hartberg (Offense) vs Austria Wien (Defense)

(b) Overview of conditional view

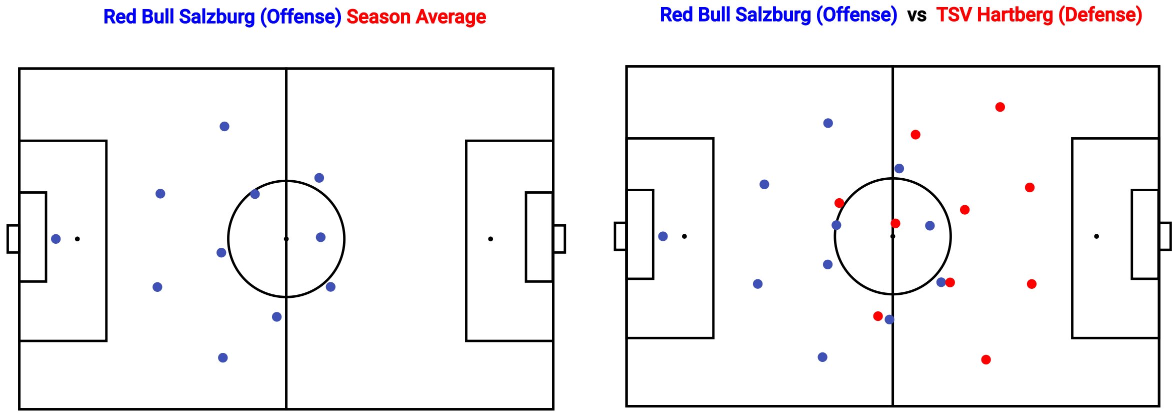
Figure 19: This illustration shows the conditional view without any sub-setting or additional visualization-features applied. It displays the vanilla comparison of the offense versus the defense of two teams. The red numbers are added for future references throughout this chapter. Figure 19a offers improved readability of the menu choices, while Figure 19b provides a total impression of the design layout of the view.

Option 1a / 1b Match-up selection:

The most obvious choice allows the team selection. Formations for specific teams can either entail the home team, the opponent, or the average for the home team (this is only an option for 1b). The formations are pre-calculated and will therefore load immediately, as required by **V3** ***Fast Access and Performance***. The left team is displayed in blue and the opponent in red to quickly distinguish the player locations on the field.

The title aims to provide a summary of the selected options by highlighting the selected teams, their mode (see menu Option 2), and the opponent (or the selection of *Season Average*) colored in the same color as their team's points.

The color scheme addresses design requirement **V1** ***Familiar Result Presentation Patterns***, which underlines the frequently used red vs blue color schemes in analog versions of formation visualizations.


Red Bull Salzburg (Offense) Season Average
Red Bull Salzburg (Offense) vs TSV Hartberg (Defense)

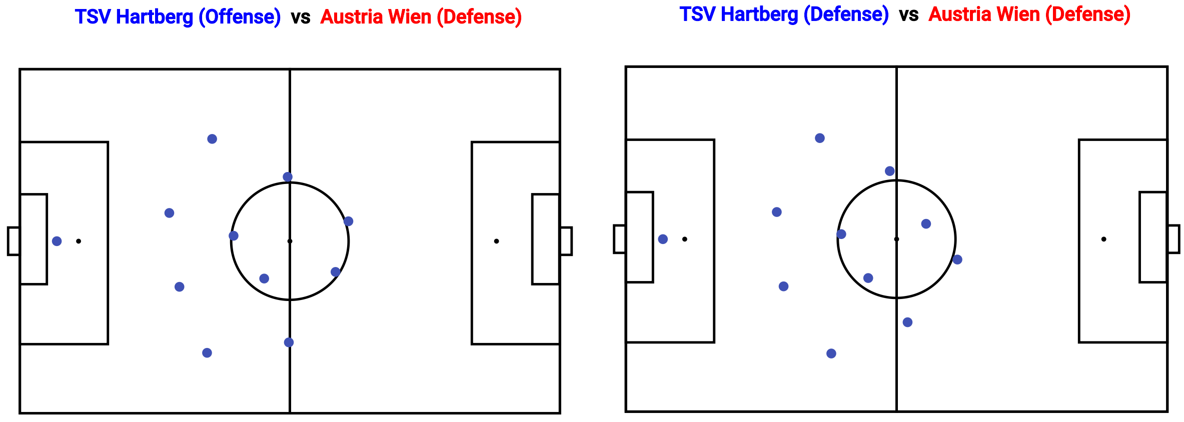
(a) Team's average formation

(b) Specific match-up visualization

Figure 20: This illustration highlights the options for the team selection feature. Figure 20a lays out the view for a grand average formation of Red Bull Salzburg over the entire data set, while Figure 20b highlights a team's formation subset to a specific match-up (here against TSV Hartberg).

Option 2 Mode selection:

Additionally to the match-up, the user can differentiate between offensive and defensive formations. This option reflects a sentiment from expert validations that underlines that formations often sketch out a rough match plan but that the actual match situation will strongly affect these rough layouts. Therefore, in line with design requirement 📊 **V2** ***Clear Separation of Concerns*** formation calculations are not mixed across offense and defense.


TSV Hartberg (Offense) vs Austria Wien (Defense)
TSV Hartberg (Defense) vs Austria Wien (Defense)

(a) Mode selection - offensive formation

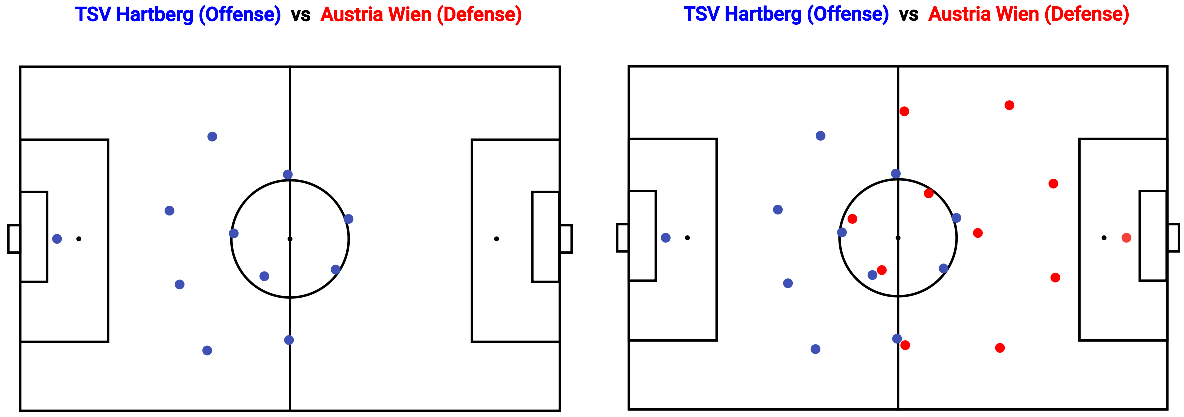
(b) Mode selection - defensive formation

Figure 21: The user can decide between the offensive or the defensive formation of a team. Figure 21a highlights the typical offensive formation of TSV Hartberg against Austria Wien, while Figure 21b displays the defensive counterpart.

Option 3 Team display:

To avoid cluttering the valuable visualization space, the user can decide to display both teams or hide either one. Hiding an entire team disabled the selector for spatial error-ellipses.

This illustration extends search/subsetting requirement 🔍 S2 *Limitation to Relevant Choices* for the actual visualization and visualization requirement 📊 V2 *Clear Separation of Concerns* to focus on the team of interest.

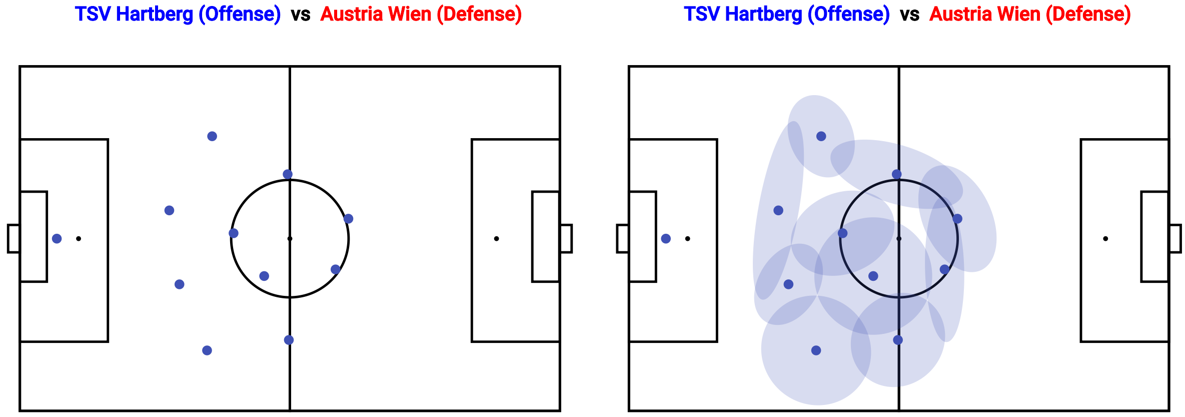

TSV Hartberg (Offense) vs Austria Wien (Defense)
TSV Hartberg (Offense) vs Austria Wien (Defense)

(a) Hide team selection - opponent not shown (b) Hide team selection - opponent shown

Figure 22: The user can hide one or even both teams, if the visualization becomes cluttered. Figure 22a illustrates the formation of TSV Hartberg against Austria Wien without the opponent shown, while Figure 22b shows both teams.

Option 4 Spatial error-ellipses:

Similar to Option 4 in Chapter 5.1, the user can introduce a visual measure of uncertainty to the visualization by displaying spatial error-ellipses around each average location. Chapter 5.1 provides a more visual introduction, while Step 4 in Chapter 4.1 offers a mathematical background on the derivation. Figure 23 compares the options of displaying and hiding the ellipses.

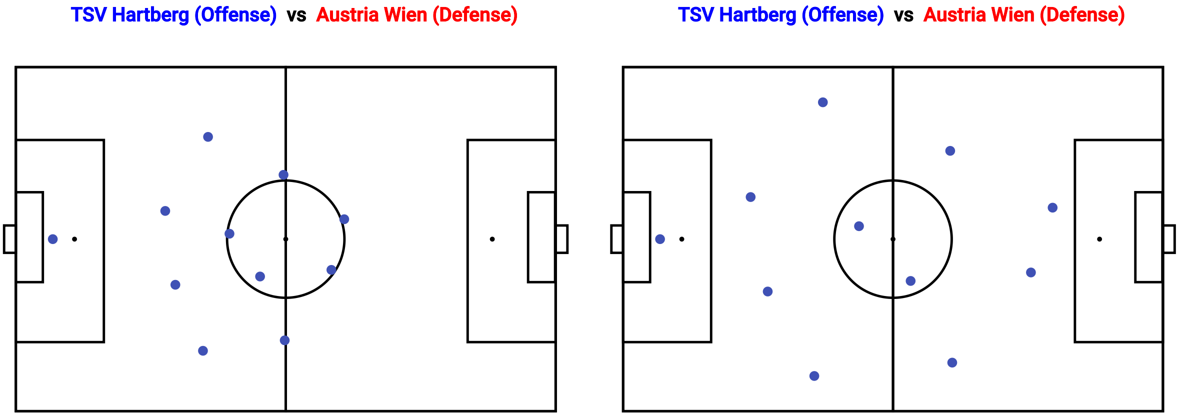

TSV Hartberg (Offense) vs Austria Wien (Defense)
TSV Hartberg (Offense) vs Austria Wien (Defense)

(a) Ellipses selection - ellipses not shown (b) Ellipses selection - ellipses shown

Figure 23: Spatial error-ellipses allow the user a notion of uncertainty of the displayed average positions. Figure 23a illustrates the formation of TSV Hartberg against Austria Wien again as a base case without any additional features shown, while Figure 23b includes the ellipses around the players.

Option 5 Scale display:

The last option offers the user the flexibility to investigate a formation's compactness compared to or in isolation from other formations. Chapter 4.3 outlines the details of the method. In short, the logic entails either shrunk or widened formations. The relative scaling uses the most extreme x- and y-coordinates of the entire dataset to match the visualization domain. The absolute scaling (relative scaling turned off) displays the formation in isolation and spread out over the entire field. Figure 24 highlights this comparison and demonstrates its alignment with design requirement 📊 **V1** ***Familiar Result Presentation Patterns***. Users generally feel most familiar with formation displays that cover the whole field. The relative scaling option will offer an intuitive new interpretation by making formations naturally more comparable to one another when compared on the same scale.


TSV Hartberg (Offense) vs Austria Wien (Defense)
TSV Hartberg (Offense) vs Austria Wien (Defense)

(a) Relative scaling - formation scaled relatively

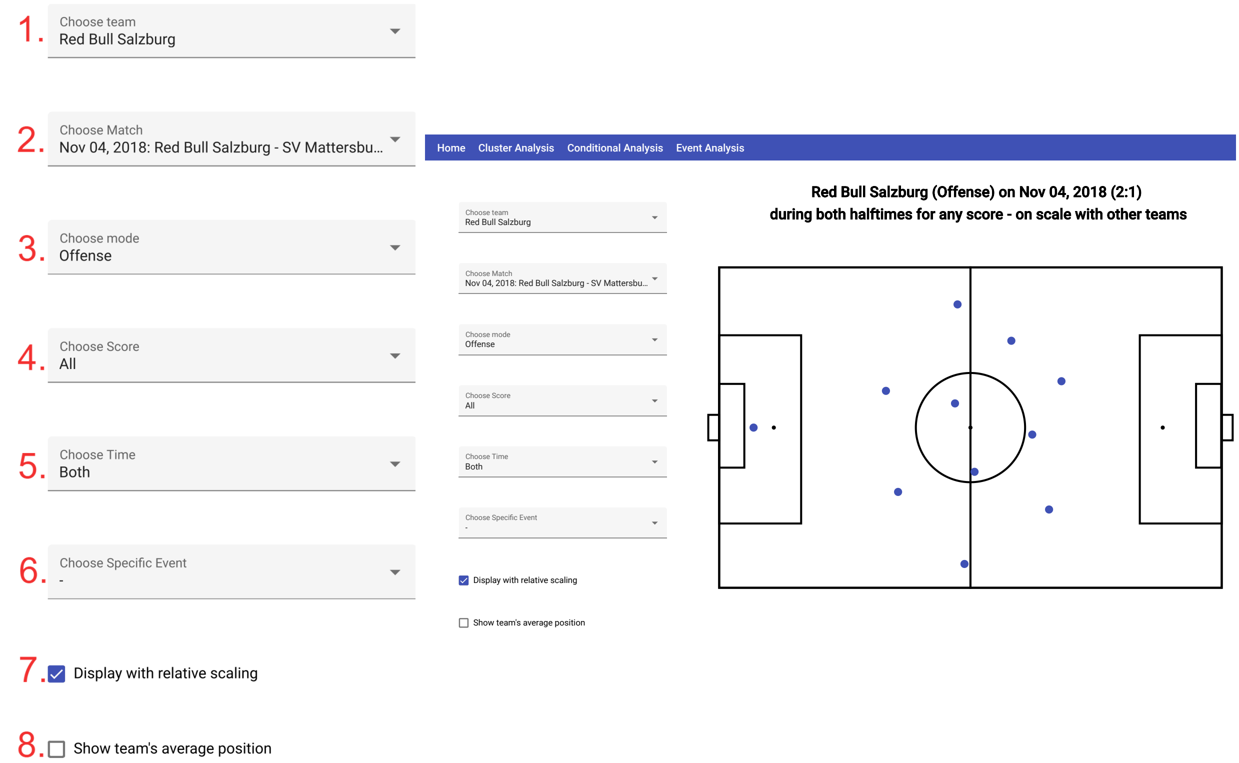
(b) Relative scaling - formation scaled in isolation

Figure 24: Scaling a formation relatively offers insights to the formations compactness in comparison to the entire data set. Most formations will therefore be shrunk to represent their relative density in comparison to all other formations. If relative scaling is turned off, the user can identify specific positions more accurately, but loses a sense for how densely the formation stands in reality. Figure 24a illustrates the formation of TSV Hartberg against Austria Wien scaled by the entire dataset, while Figure 24b illustrates the wider view of the formation.

5.3 Event View

While the clustering view offers insight to overall groups within the entire dataset, the conditional view clarifies specific team and match-up tendencies. For the most granular level of the system, the *event view* visualizes match- and even situation-dependent information to the user. The overall layout resembles the other two views to follow design requirement 📊 **V1** ***Familiar Result Presentation Patterns***. Figure 25 shows the numbered view of a vanilla match selection without any further subqueries.

As in previous illustrations, the **red** numbers will afford a structure to the discussion of the individual features of the system, including match, mode, and score subsetting, as well as a comparison feature to a team's grand average formation.



1. Choose team: Red Bull Salzburg

2. Choose Match: Nov 04, 2018: Red Bull Salzburg - SV Mattersbu...

3. Choose mode: Offense

4. Choose Score: All

5. Choose Time: Both

6. Choose Specific Event: -

7. ☑ Display with relative scaling

8. □ Show team's average position

Red Bull Salzburg (Offense) on Nov 04, 2018 (2:1)
during both halftimes for any score - on scale with other teams

Choose team: Red Bull Salzburg
Choose Match: Nov 04, 2018: Red Bull Salzburg - SV Mattersbu...
Choose mode: Offense
Choose Score: All
Choose Time: Both
Choose Specific Event: -
☑ Display with relative scaling
□ Show team's average position

(a) Event view - menu

(b) Event view - overview

Figure 25: This illustration shows the event view menu in Figure 25a and an overview of the entire visualization view in Figure 25b. Figure 25a on the left offers a zoomed-in view of the menu choices for better readability, while the Figure 25b on the right side provides an overview of the entire tab. The red numbers are added for orientation throughout the explanations of this chapter.

The title adjusts dynamically to provide the user with guidance on the specific match displayed and selected options. The background covers the selected team's logo without interfering with the actual visualization of the playing field. The visualization also includes a hover-effect that displays the most-likely player for any given average position. Figure 26 highlights this feature. These small reminders for the user underlines 🔍 S1 ***Intuitive Design Choices*** to pair creative analyses with familiar and intuitive pictures facilitating orientation around the system.

Option 1 Team selection:

The team selection drop-down menu follows a similar logic to the same selector of the conditional view. The menu is pre-filled with all teams included in the data set, and the user can decide on which team to focus. This decision will impact the options of option **2**—the match selection—to only include matches of the selected team.

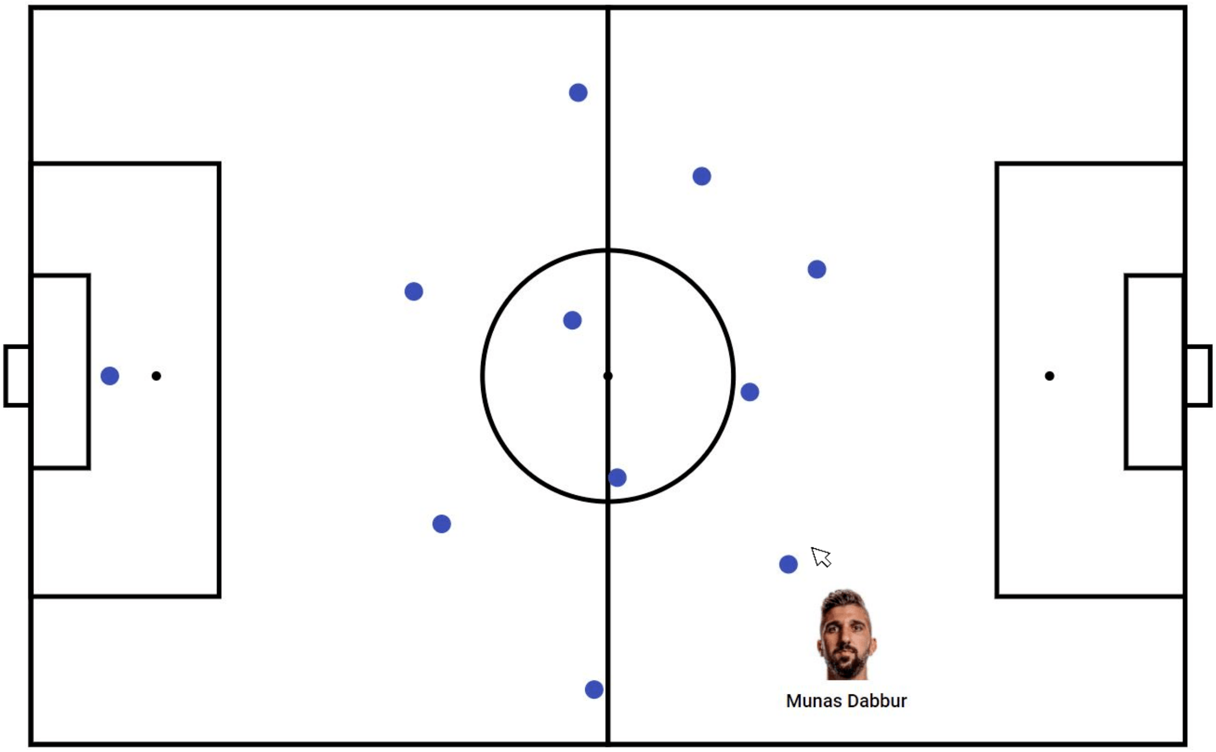
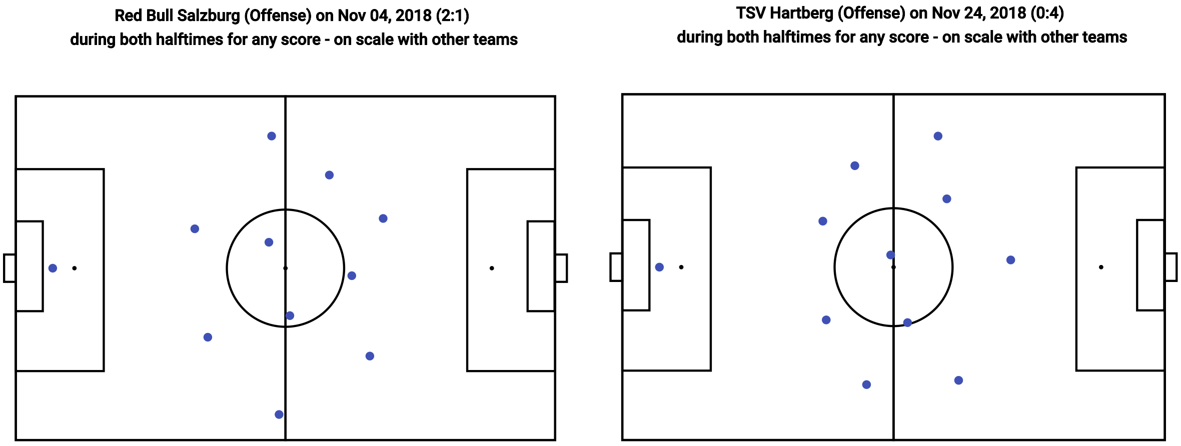


Figure 26: The user can display the most likely player for any given position on the field by hovering over a small area surrounding the indicated position.


Red Bull Salzburg (Offense) on Nov 04, 2018 (2:1)
during both halftimes for any score - on scale with other teams
TSV Hartberg (Offense) on Nov 24, 2018 (0:4)
during both halftimes for any score - on scale with other teams

(a) Team selection - Match of *Red Bull Salzburg*

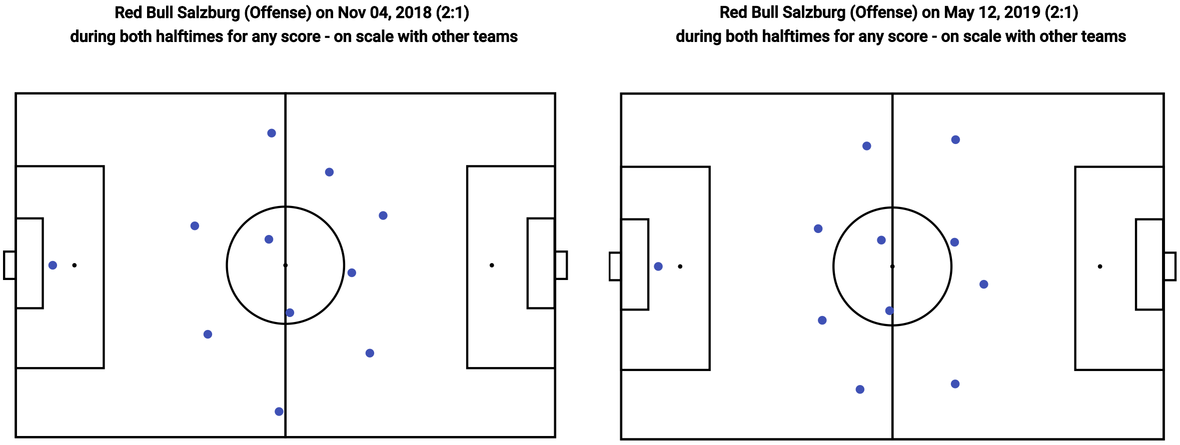
(b) Team selection - Match of *TSV Hartberg*

Figure 27: The team selection impacts the visualization by initially auto-selecting the first match of the data set of a specific team. It also updates the title to represent the corresponding selection. In this case, the team switches from Red Bull Salzburg in Figure 27a to TSV Hartberg in Figure 27b.

Option 2 Match selection:

The match selection offers a new level of granularity in comparison to the other views. Here, the user can select a specific team and the formation of a single match. The matches are sorted by date and follow the format of “Date: Match-up - Score“, so, for example, *Nov 04, 2018: Red Bull Salzburg - SV Mattersburg: 2:1.* Following common practices in sports and visualization requirement **V1** ***Familiar Result Presentation Patterns***, the names and goals of the teams begin with the home team. The titles adjust for any newly selected match.

The season averages for the *2018/2019* and *2019/2020 season* are also available options, additionally to the individual matches in the dataset.



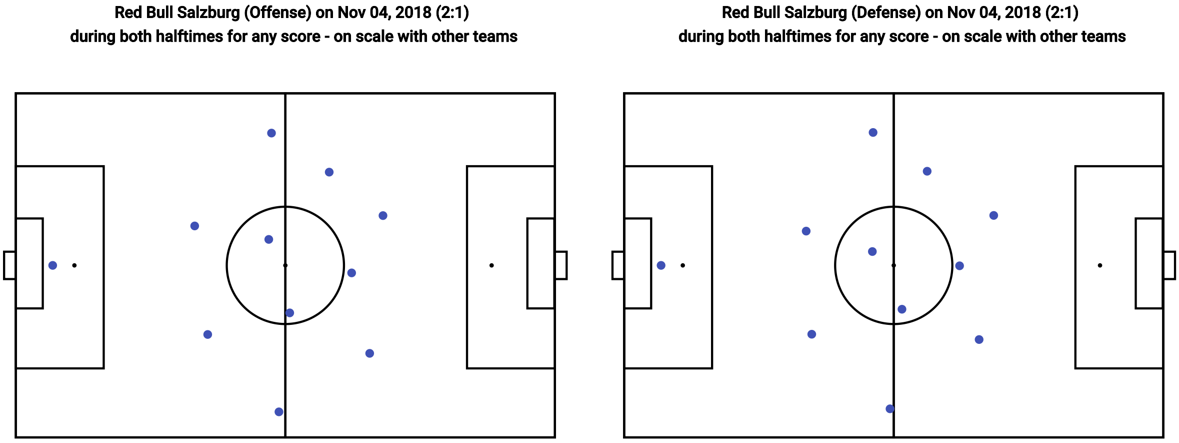
(a) Match selection - Match of 11/04/2018

(b) Match selection - Match of 05/12/2019

Figure 28: The match selection affords the user the option to change the selected match for any chosen team. The visualization compares two formations: Red Bull Salzburg of 04. November, 2018 in Figure 28a with Red Bull Salzburg of 12. May, 2019 in Figure 28b.

Option 3 Mode selection:

Similar to the mode selection described in Chapter 5.2, the user can choose a mode from *offense* or *defense* to query the respective data. Figure 29 highlights the alternatives and how the title adjusts for the new information.



(a) Mode selection - Offensive formation

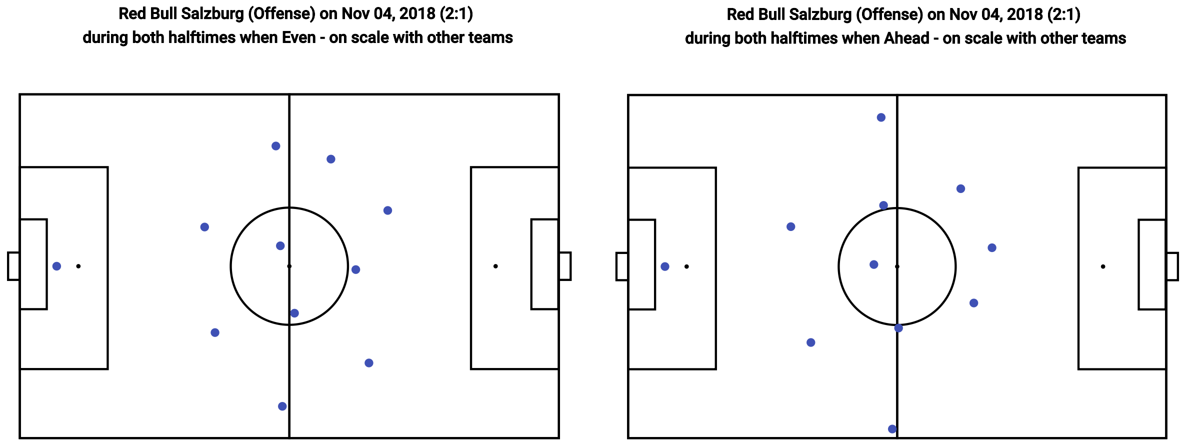
(b) Mode selection - Defensive formation

Figure 29: The mode selection differentiates between the offensive and defensive formations of a team for a specific match. Figure 29a displays the average position of all sequences in possession while Figure 29b highlights the ones out of possession.

Option 4 Score selection:

To analyze a match in detail, a coach might be interested in a team's behavior under pressure or dominating. The score selection offers additional insight into the stages of a game by highlighting situations of various periods. These details might reveal a team's tendency in differing situations. Figure 30 exemplifies these features. The available options might include *even*, *behind*, or *ahead*.

Questions of how a team defends when ahead or how it attacks when behind (or vice versa) can become vitally important to prepare a team for a given match-up. The options to select are limited to the situations present for the selected team. For example, if the home team of a 0-4 match is selected, the drop-down will not contain an option to choose *ahead*, because the selected team was never ahead in the game. This pre-selection is also adjusted for smaller time frames, such as the available scoring labels when subset to only the second half—see Option **5**.



(a) Score selection - Even

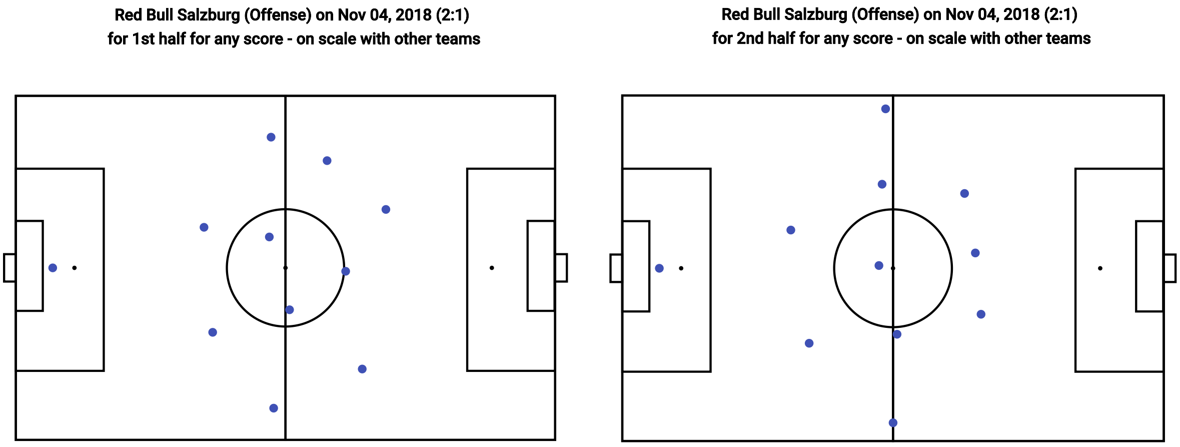
(b) Score selection - Ahead

Figure 30: The score selection subsets a match into distinct sequences where the selected team was either even with, or ahead / behind of the opponent. Figure 30a displays the average position of all sequences of the selected team Red Bull Salzburg being even in score while Figure 30b highlights the sequences when the team was leading.

This feature introduces the notion of sub-match granularity, which allows for the subsetting of periods on a finer level than just the per-match-average. Search/subsetting requirement **S3** ***Iterative Subsetting*** aims to lead the user to more advanced queries than a simple match average. This feature represents the first notion of temporal stages within the match to the visualization, which is often referenced as an important aspect to any formation analysis (see Chapter 7.2).

Option 5 Halftime selection:

The halftime selector represents the second temporal selection device of the list. The user can, independent of score or match situation, subset the data to first and second halves. If the times overlap with the score change, the visualization might show similar results to the selection to Option 4, but its logic is entirely independent of it. Figure 31 displays the halftime subsetting for the vanilla case of a single match with no further options changed from the default.



(a) Halftime selection - First half

(b) Halftime selection - Second half

Figure 31: The halftime selection subsets the match to either the first or second halftime of a match. Figure 31a displays the average position of all sequences of the selected team Red Bull Salzburg during the first half while Figure 31b highlights the sequences of the team during the second.

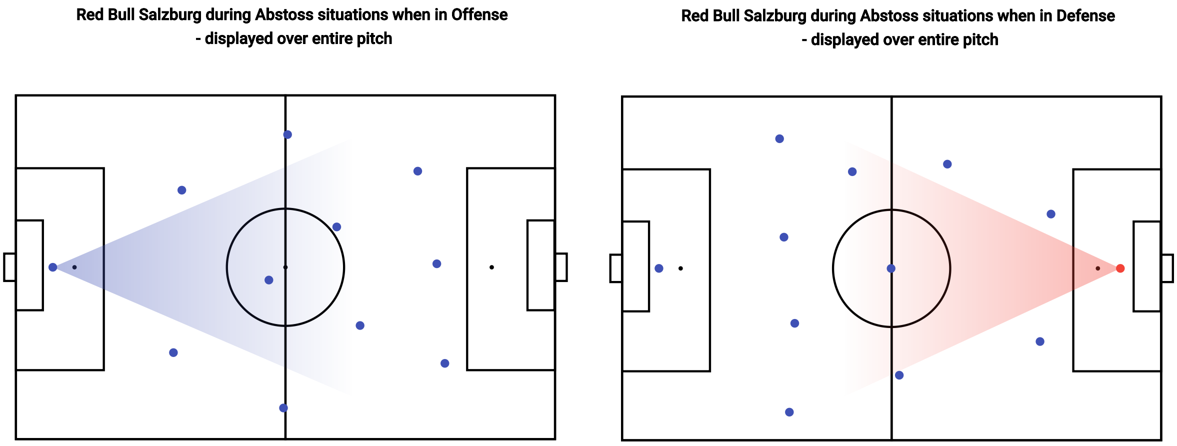
Option 6 Event selection:

Options 4 and 5 introduced options to select events within a match. The following selectors address feedback implemented after interactive feedback sessions with domain experts (see for example Chapter 3).

The event selector features the formations during specific match events. These events are highly distinct from the general match formations covered by the previous functionalities. A particular interest lies within the formation during a long pass from the goalie, or *goal kick* (German *Abstoss*). These situations lead to a subsequent scrambling for the ball and initiate the first stage of an offense—the build-up.

So far the selection only includes the *Abstoss*-event, but in line with search/subsetting requirement QS4 ***Easy Extensibility***, the system will cover multiple additional events in future versions.

The selection of *Abstoss* as event changes the selection options of a couple of other drop-down fields. Abstoss data is scarce, which means every match only includes a few of these events. Therefore, a specific formation subsetting during a long pass for a specific match introduces more noise than information. The match selection becomes disabled for this event, and only an average formation per team can be selected to avoid cluttering of options. The same logic holds for the score, the halftime, and the scaling selector (Option 7). These are disabled if no specific match is selected.


Red Bull Salzburg during Abstoss situations when in Offense
- displayed over entire pitch
Red Bull Salzburg during Abstoss situations when in Defense
- displayed over entire pitch

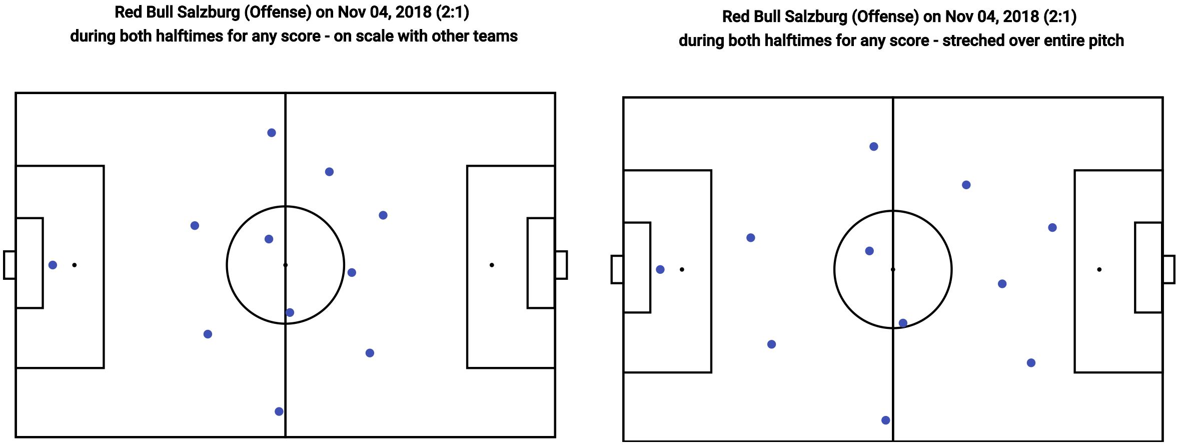
(a) Event selection - Abstoss offense (b) Event selection - Abstoss defense

Figure 32: The event selection allows the user to visualize a team's tendencies for specific situations that are not already captured within the fluid match flow, but are of specific interests to invested parties. The visualization shows the average formation of Red Bull Salzburg during a long pass from the goalie. Figure 32a on the left shows the formation when the own goalie initiates their offense, while Figure 32b on the right highlights the formation defending the opponent's goalie pass. The graphic is extended by visual cues to indicating the match direction and typical field of view of a goalie.

The defense and offense selection will become more impactful for the Abstoss-event, because either one includes a visual cue—a colored cone—to underline the direction of play for the long pass. Figure 32 shows the two formations for offense and defense for a selected team. It incorporates design requirements **V1** ***Familiar Result Presentation Patterns*** building on the common color scheme used throughout the entire system. In line with **V3** ***Fast Access and Performance***, the results are pre-computed and therefore load immediately for the fastest possible display of information.

Option 7 Scale display:

Scaling the formation over the entire field (for better visibility) or scaling it to other teams (for better comparability) follows the same logic as within the conditional view. Chapter 4.3 provides calculation details. Figure 33 displays a comparison of the same formation scaled or stretched to highlight the respective advantages. These options offer extensive insights into the relative compactness additional to the strict match-plan of a formation. This logic follows the idea of search/subsetting requirement **S3** ***Iterative Subsetting***, especially in combination with other options fo form complex queries.



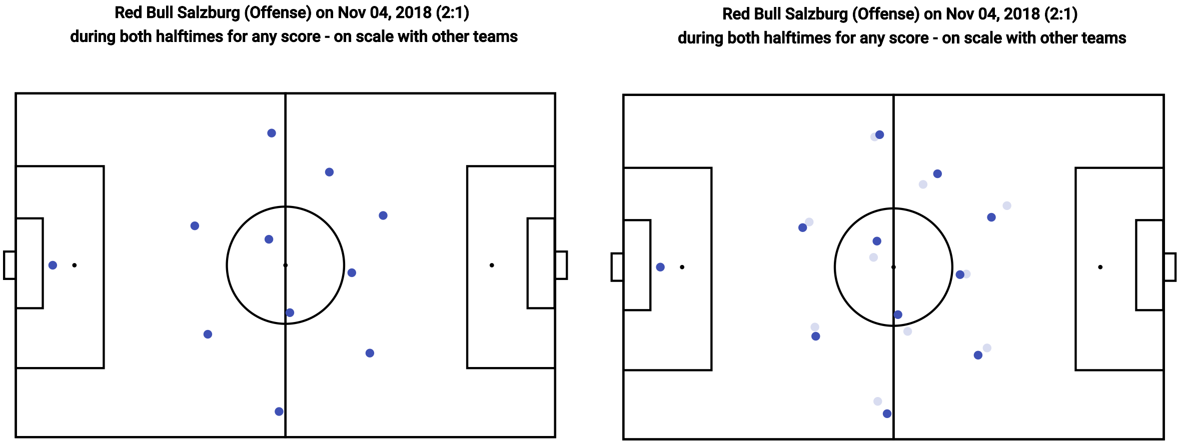
(a) Relative scaling - formations scaled relatively

(b) Relative scaling - formations scaled in isolation

Figure 33: Scaling a formation relatively offers insights to the formations compactness in comparison to the entire data set. Most formations will therefore be shrunk to represent their relative density in comparison to all other formations. If relative scaling is turned off, the user can identify specific positions more accurately, but loses a sense for how densely the formation stands in reality. Figure 33a illustrates the formation of Red Bull Salzburg against SV Mattersburg scaled by the entire dataset, while Figure 33b illustrates the wider view of the formation.

Option 8 Show average formation:

While identifying formations in isolation provides insights, coaches are frequently interested in measuring deviations from a base case or a default structure. The *Show team's average formation*-button provides this functionality. It overlays the current selection with a subtle visualization of a team's grand average formation. This visualization embodies the visualization requirement 📊V4 ***Comparability of Different Queries***. The user does not have to compare results separately but compares deviations from one another in a single visualization.



(a) Average formation not shown

(b) Average formation shown

Figure 34: This visualization shows the simultaneous display of the average as well as a specific formation. The grand average is calculated across all sequences of a specific team without further subsetting.

6 Use Cases

The potential use cases of the system cover an arbitrarily large number of scenarios. This chapter aims to exemplify solutions to realistic scenarios, reaching beyond the basic functionalities described previously, and illustrate exciting insights along the way.

The chapter's overall structure mimics the system's granularity levels and incorporates functionalities of all views (cluster, conditional, and event, in detail in Chapters 5.1, 5.2, and 5.3, respectively) to find wholesome information for the user.

This overall structure emphasizes the visualization requirement 📊 **V2** ***Clear Separation of Concerns***—see Chapter 3 for a full list of design requirements. Hence, solely the question's granularity dictates the most appropriate view to investigate.

6.1 Use Case 1: Dynamic Flow of Formations

The first use case investigates high-level temporal dynamics of formations. The question to analyze evolves around the question if the system can depict certain formation tendencies. The *Next cluster*-hover-functionality of the cluster view affords the user information of formational flows. This effect displays the five most likely next clusters of the currently hovered cluster. Chapter 5.1 offers a detailed and visual explanation of the effect.

This visualization is most relevant for single teams—unless we expect the entire league to follow a particular recurring flow of formations. Therefore, this example focuses on the sequence of typical formations for a single team, here *Red Bull Salzburg*. By selecting the team from the dropdown menu and choosing the *Next clusters*-hover-effect, we can inspect the typical sequence of the team's most prominent formations.

This visual inspection shows that for this use case, *Red Bull Salzburg* seems to follow a loop in their formation movements. Figure 35 displays player movements within this loop that a coach can inspect to prepare for *Red Bull Salzburg's* common offensive patterns.

The initial formation resembles a typical formation pattern for the team—the diamond midfield structure. Expert interviews communicated how Salzburg's players exhibit strong individual skills that allow them to initiate and profit from frequent scrambling in the middle third of the playing field. Salzburg often creates chaotic situations that bring players in direct one-on-one duel situations with opponents, which favors Salzburg's individually better-equipped players.
This pattern aligns with the more detailed sequence depicted in Figure 36. After the initial diamond shape contracts towards the middle, the two forwards close in, and the top and bottom of the diamond shift forward. This movement increases pressure in the midfield, establishing four forwards for quick passes through the middle attacking the goal. The formation eventually returns to the initial state with wide forwards and a diamond at its heart. The entire defensive line of four players remains almost constant, with only slight shifts towards the center during the contraction.

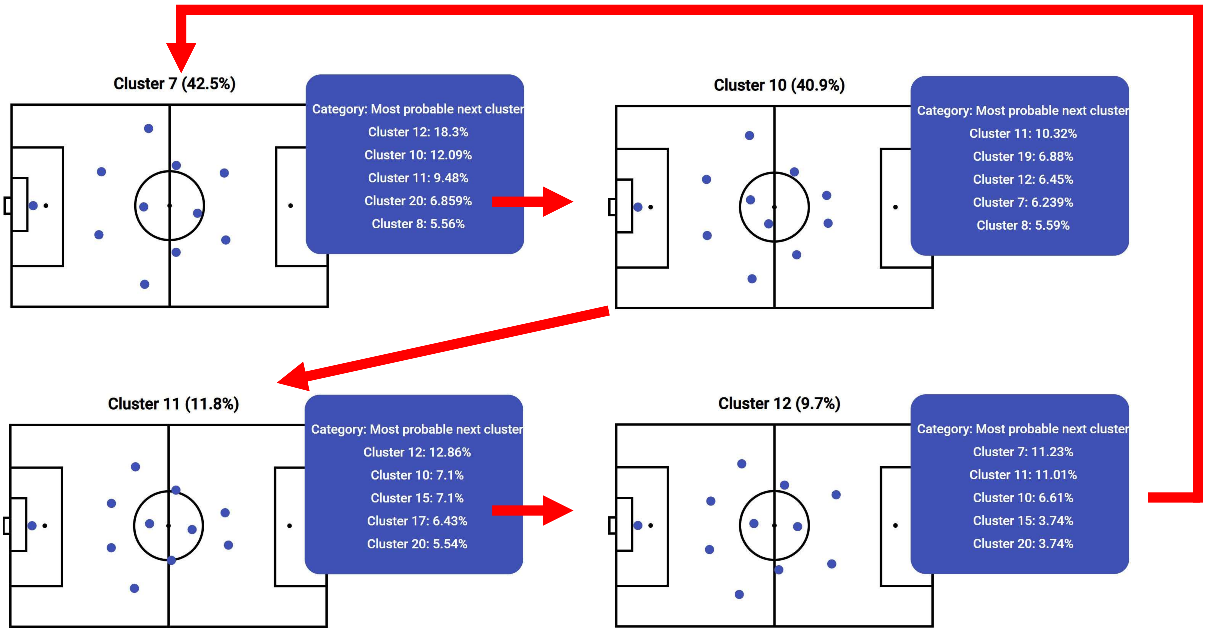


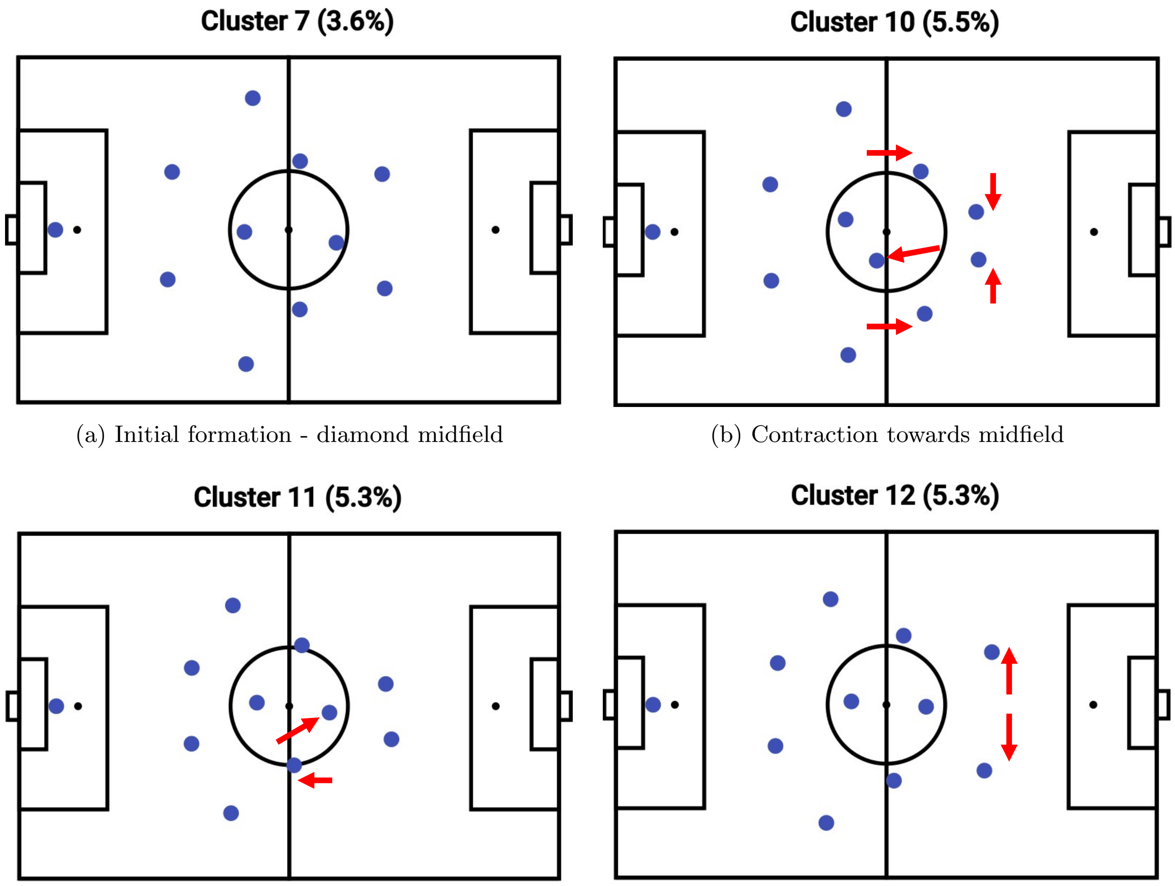
Figure 35: This visualization demonstrates a use case for the cluster view to find typical formation sequences for specific teams. Here, a four formation loop for Red Bull Salzburg displays the general movements from the top left to the bottom right and back. Red arrows indicate the sequence from one cluster into the next for clearer readability, which are not included in the original system design.

This use case exemplifies a common situation, where a coach tries to first identify patterns of a potential next opponent and to investigate the temporal flow of formations. In this example, the user could easily visualize and research the nuanced movements of a top European soccer club. The calculations, and visualizations are immediate following 🖩 **A3** ***Fast Algorithmic Performance*** and 📊 **V3** ***Fast Access and Performance***. Results align with experts' opinions stated in qualitative interviews adhering to 🖩 **A1** ***Correct Formation Calculations***. Furthermore, the queries feel intuitive to the user embodying 🔍 **S1** ***Intuitive Design Choices*** and 📊 **V1** ***Familiar Result Presentation***.

6.2 Use Case 2: Formations Based on Opponent Quality

The explorative investigation of an opponents' formations for upcoming matches presents another typical use case for coaches. While an average structure might prove instructive for a high-level overview, a team's formation might deviate drastically between different match-ups.

For the second use case, the application will compare how *Red Bull Salzburg* formation responds to a stronger versus a weaker opponent. Quick aside: *Red Bull Salzburg* has won the last seven league titles in the Austrian Bundesliga. For the period covered in the data, the team of *Wolfsberger AC* was one of their fiercest rivals, finishing third in 2018/2019 and 2019/2020. *WSG Swarovski Tirol*, formerly known as *WSG Wattens*, was promoted to the Austrian Bundesliga in 2019, where they only won five out of 22 games and finished tenth of twelve teams. These two opponents build the pool for this use case to display how Salzburg reacts to opponents of different strengths.



Cluster 7 (3.6%)

Cluster 10 (5.5%)

Cluster 11 (5.3%)

Cluster 12 (5.3%)

(a) Initial formation - diamond midfield

(b) Contraction towards midfield

(c) Rebuilding the diamond midfield

(d) Expansion of forwards towards initial formation

Figure 36: The visualization shows the individual player movements that a coach can derive from following the "Next cluster" hover information. Here, Red Bull Salzburg seems to contract their players towards the middle before rebuilding their stereotypical diamond-shape of the midfielders. Eventually, they spread-out again to return to their initial formation.

Salzburg's average formation for the entire season displayed in Figure 37 serves as a benchmark to compare the more subtle tendencies of Figure 38. The main insights of Figure 38 lie in the variance Salzburg shows in its formation between offense and defense when playing against a strong opponent. However, when the team faces a weaker team, displayed in the bottom row in Figures 38c and 38d, the formations remain almost identical between offense and defense.

The formation change against *Wolfsberger AC* of offense and defense mainly impacts the midfield. The left midfielder shifts inside the center to stop *Wolfsberger*'s offense through the middle. Furthermore, compared to sequences against *WSG Tirol*, the outer forwards move further back to a more defensive stance, which indicates a generally higher expectation of danger through the outer lanes. The logic for this behavior becomes clear by further subsetting the query and visualizing the opponent as well.

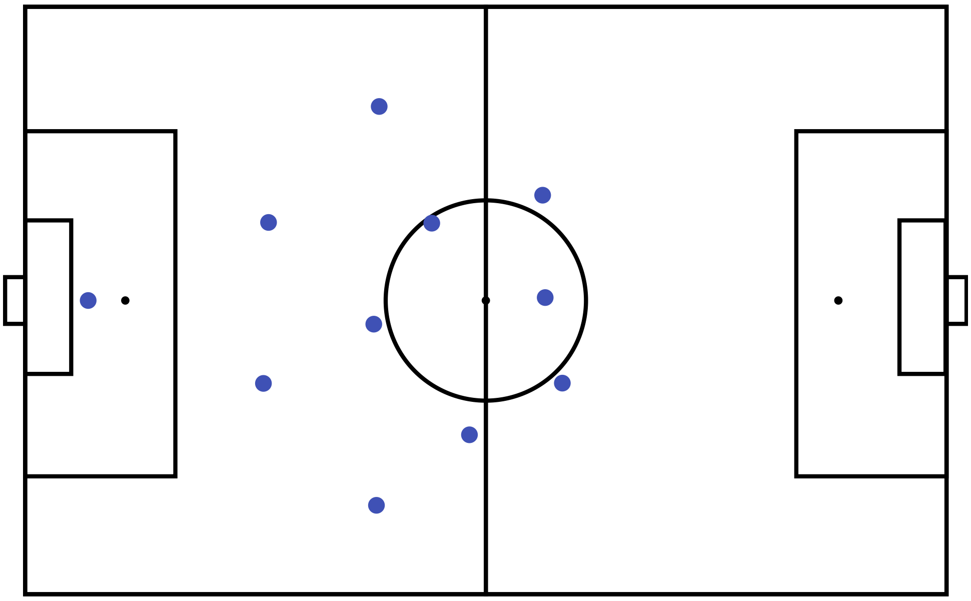
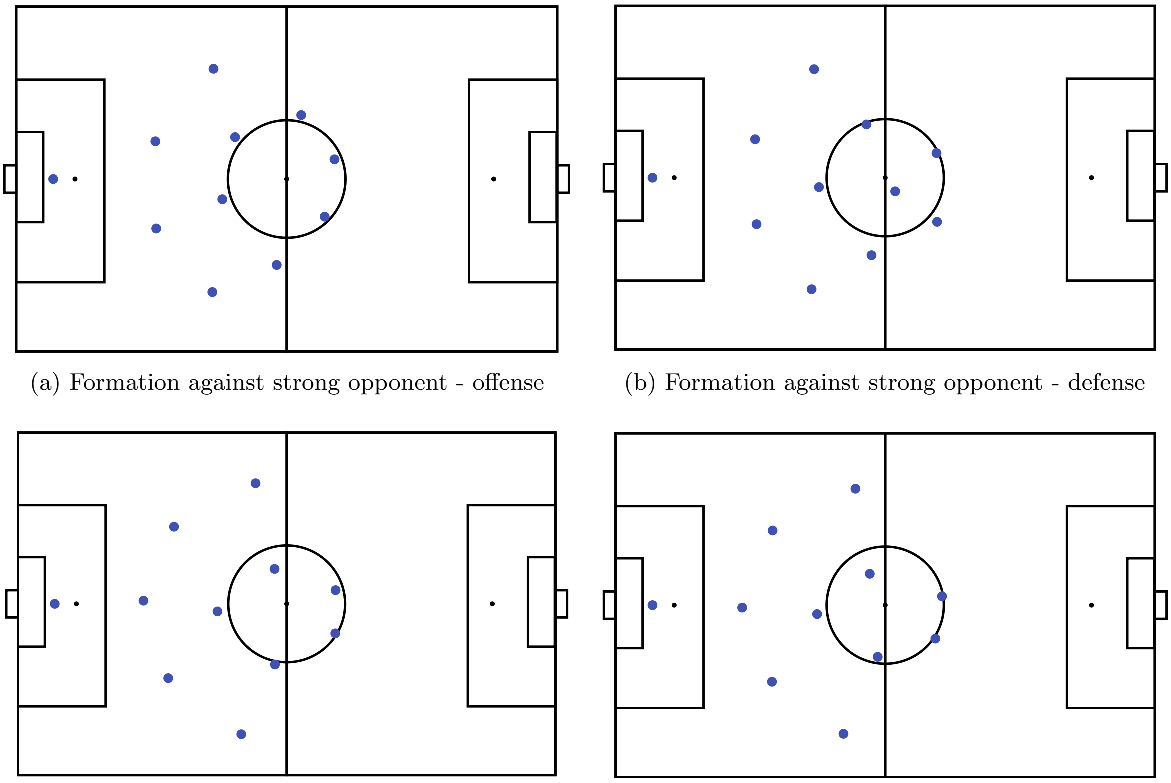


Figure 37: The figure shows the average formation for Red Bull Salzburg from the conditional view. This average helps as a reference when discussing specific cases throughout the chapter or, more specifically, in Figure 38.



(a) Formation against strong opponent - offense

(b) Formation against strong opponent - defense

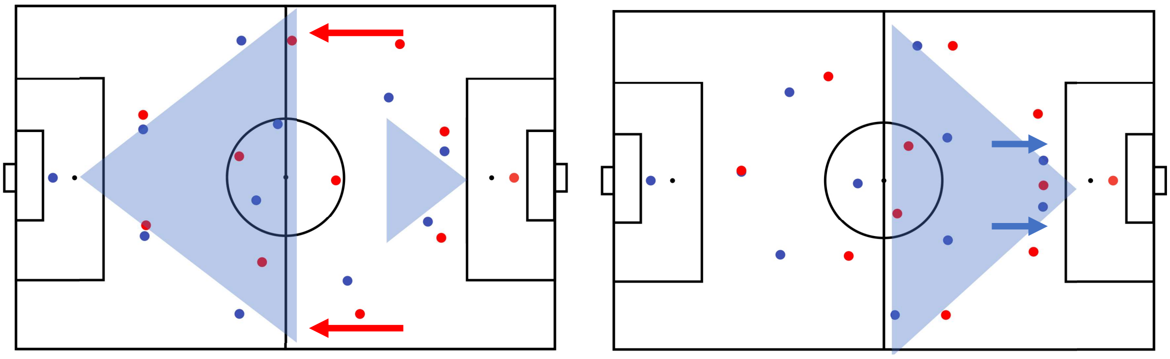
(c) Formation against weak opponent - offense

(d) Formation against weak opponent - defense

Figure 38: These figures display exciting cases on how a Red Bull Salzburg's formation changes based on the quality of the opponent. The top row exhibits the offense—Figure 38a— and defense—Figure 38b— formations against Wolfsberger AC, one of the top competitors for the championship the last two years. The bottom row mirrors the offense/defense split in Figure 38c and Figure 38d, but against a weaker opponent (WSG Swarovski Wattens). The top row shows less resilience and a shift between offense and defense, while the formations against the weaker opponent remain basically unchanged.

Figure 39 displays both teams' formations. Figure 39a shows the match-up of *Red Bull Salzburg* against *Wolfsberger AC* (based on six matches in data set) and Figure 39b the counterpart against *WSG Swarovski Tirol* (based on two matches in data set). The illustration addresses the question why Salzburg's formation varies drastically between offense and defense against a strong, while remaining stable against a weak opponent.

The overall tendency of *Salzburg*'s formation shifts against a weaker opponent. The forwards attack the gaps in the middle of the defense, and the midfielders create more pressure on the wings. This strategy is indicated by blue arrows in Figure 39b. This effort clashes with the more defensive formation against stronger opponents exemplified in Figure 39a. The midfielders address the additional pressure created by the broad wings and midfielders. Here, indicated by red arrows in Figure 39b. The overall playing style indicates a more cautious propensity for the higher scoring potential of a stronger team.



(a) Salzburg's focus against strong opponent

(b) Salzburg's focus against weak opponent

Figure 39: This figure allows for a further investigation of the discrepancy of the formation behavior hinted at in Figure 38. Red Bull Salzburg's different approaches to respond to offensive and defensive situations against a strong and weak team is partially explained by the opponent's formation, especially when both are scaled to the entire field. The strong opponent creates more pressure over the outer wings, highlighted by red arrows in Figure 39a. This pressure results in a more defensive response by Salzburg. The opposite is true for Figure 39b, where Salzburg creates an increased pressure through the middle and ignores the presumably less dangerous offensive efforts by the opponent.

A separate discussion with an international sports journalist hints at core insights regarding the general pressing[39] tendencies of a team. The following summary paraphrases the interviewee's main points:

> *"Formations against strong and weak opponents differ mainly in how they attack free spaces. More skill usually correlates with more time in possession, which lets a team dictate a match's flow. A stronger team will try to bully the weaker team through continuous pressure, which will lead to unforced mistakes creating additional opportunities. However, stronger opponents create more space to introduce opportunities without the luxury of opportunistic counter-offenses. Once the direct gateway through*

[39] Pressing is a collective team effort aimed at disrupting the opposition's build-up to win the ball. It is different from defending, which means closing down space in the defending third, thus preventing the opponents from creating an opportunity to score—pressing usually identifies aggressive defense within different soccer field regions.

> *the middle is closed, teams will choose the second-best option to attack a goal, which leads over the wings. These phenomena lead to two characteristic differences between the behavior against weaker and stronger teams: First, the tactical tool of defensive pressing becomes more prevalent against technically inferior opponents, and, second, a stronger team will leverage more opportunities directly through the central-defense against weaker than against stronger opponents.“*
> (Phone interview in February 2021)

These insights find a direct translation to the visualizations in Figure 39. An exact 1:1 mapping of defense to offense paired with more poignant attacks through the middle characterize the formation against an inferior team in Figure 39b.

Additional expert interviews further express that Salzburg is a team that thrives on creating chaotic situations in the middle third, stealing the ball, and scoring off counter-offenses. Therefore, the overall formation will shift forcefully from defense to offense if the team focuses on defense against more vigorous opponents.

6.3 Use Case 3: Formation Adjustments During a Match

So far, Use Case 1 demonstrated the usage of overall flows between formations, and Use Case 2 subsets the data to specific match-up situations. The event view allows for an even more granular investigation by comparing formations of the same match. While many different subsetting options are available, an often discussed feature is the strategic adjustment after a halftime break. The coaches communicate changes dependent on the match situations, player behavior, and tactical adjustments of the first half. Figure 40 represents an exemplary case of a match by *Rapid Wien* against *TSV Hartberg* ending 3:3.

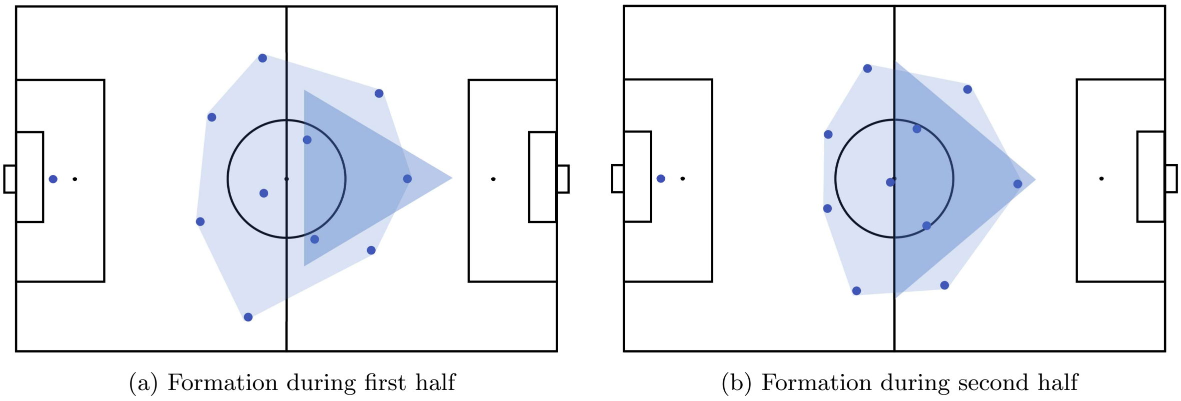

(a) Formation during first half (b) Formation during second half

Figure 40: The visualization shows a formation change of TSV Hartberg during a match. Figure 40a displays the broader and more aggressive propensity of the first half, compared to the denser and more conservative classical 4-1-4-1 formation of the second half in Figure 40b. The colored shapes are added to mimic a coach's expertise to visualize the formations and pick up focal subtleties.

While the formations do not drastically change, they exhibit a tactical shift towards a more traditional 4-1-4-1 formation in the second half. The first half formation covers a broader area on the field with more focused offensive pressure through the middle. This adjustment becomes increasingly intuitive when put into perspective of the time table of scored goals illustrated in Table 1. This context is necessary to investigate the changing formation and reason about potential causes intelligently. A self-evident explanation lies in *TSV Hartberg*'s general fight for lower seasonal standings than *Rapid Wien*.[40] Therefore, a tied game represents more of a success for *TSV Hartberg* than for *Rapid Wien*. The more structured and less aggressive orientation of the second half seems to underline this tendency—the team tries to secure the even score, save one point from the match,[41] and avoid any additional goals from the opponent.

Minute	**Player**	**Score (Rapid Wien : TSV Hartberg)**
17'	Taxiarchis Fountas	1 : 0
45'	Jodel Dossou	1 : 1
51'	David Cancola	1 : 2
72'	Taxiarchis Fountas	2 : 2
83'	Dario Tadic	2 : 3
90+6'	Stefan Schwab	3 : 3

Table 1: Overview of goals during the investigated match of Rapid Wien and TSV Hartberg on 29. September, 2019. The columns illustrate the minute of a goal, the scoring player, and the new score after the goal in the format of home team : away team, respectively.

The use case offers a final example of how result queries building on 🔍 **S1 *Intuitive Design Choices***, implementing algorithms that represent 🖩 **A1 *Correct Formation Calculations*** and 🖩 **A3 *Fast Algorithmic Performance*** lead to visualizations adopting 📊 **V3 *Fast Access and Performance*** and a 📊 **V1 *Familiar Result Presentation***.

The displayed use cases only outline a tiny fraction of potential questions that a user can ask. These three specific examples provided insights about the temporal flow of formations, strategic adjustments to an opponent's skill level, and potentially re-occurring tactical adjustments throughout a match.

7 Evaluation

This thesis extends the current frontiers of formation analysis by improving algorithmic performance and introducing an intuitive system that conveys correct formation information to its users. While strict adherence to the design requirements detailed in Chapter 3 partially secure the application's intuitiveness, the correctness and speed of the formation derivation need further validation. Accordingly, this chapter offers insights into the actual performance improvement of the proposed solution in Chapter 7.1 and compares the calculated formations to the manual estimates of unbiased domain experts in Chapter 7.2. The chapter concludes with possible extensions derived from the qualitative feedback by domain experts and implementation recommendations for future work in Chapter 7.3.

[40] *Rapid Wien* finished the 2019/2020 season in second place, right behind the championship team of *Red Bull Salzburg*. *TSV Hartberg*, surprisingly, finished the season in a healthy fifth place.

[41] A won match adds three points, a tie one point, and a loss zero points to a team's overall score.

7.1 Algorithmic Performance

Shaw and Glickman's [66] approach to calculate formations via a relative distance matrix offers the most relevant alternative to the clustering-based solution introduced in this thesis. A naive implementation of the authors' solution was tested for a large subset of the sequences and compared to a calculation via this thesis' solution to compare the solutions' relative performances.

Since the authors offer no open-source code base for their relative distance method online, further assumptions are necessary to mimic Shaw and Glickman's [66] methodology. Their description of the algorithm is as follows:

> *"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."* (Shaw and Glickman [66], page 3)

The following code-block represents the pseudo-code translation for the implementation of Shaw and Glickman's [66] formation calculation. The implementation runs on a large sub-sample of the entire data set. Out of the almost 10,000 sequences in the data, an arbitrary combination of 1,104 sequences of 42 games builds the performance comparison basis. The sequence covers all teams and evenly splits the data in offense and defense formations to avoid inherent bias in the test data.

The pseudo-code provides a detailed explanation of the implementation of the closest neighbor to this thesis' formulation. Chapter 4 outlines the underlying logic utilized for the k-means algorithm and data cleaning procedure.

These derivations follow the algorithmic requirements of **A3** ***Fast Algorithmic Performance***, which builds the necessary basis for the visualization requirement **V3** ***Fast Access and Performance***. The entire system rests on a user's dynamic interaction, therefore, the immediate and smooth run time of the code represents one of the core contribution of the thesis.

Figure 41 displays the stark difference in run time values for the sample. It is crucial to notice the shifted x-axis, aiming to improve readability without shrinking one of the two distributions to a shared axis. The results for Shaw and Glickman [66] implementation of formation calculation center around 16 seconds and are firmly left-skewed.

This leads to a mean run time of **19.88** seconds—see Table 2. The clustering solution offered by this thesis runs in (mean) **1.06** seconds with a more evenly centered distribution around one second.

Algorithm 1: Shaw and Glickman [66] FORMATION IMPLEMENTATION

```
/* Single sequence per team, per match, either in offense or defense */
Input: sequenceData
Output: playersList
$frames \leftarrow$ EMPTY LIST
/* Find distance matrix of every player with every other player for every time frame */
for $timeFrame$ in $sequenceData$ do
    $cross \leftarrow timeFrame \times timeFrame$    // × represents the cross-product
    /* Subset the data for when the player IDs are different */
    $cross \leftarrow cross[actorid_x] \neq cross[actorid_y]$    // Limit of square- to triangular-matrix
    /* Find distance in x- and y-direction */
    $cross[vx] \leftarrow cross[x_x] - cross[x_y]$
    $cross[vy] \leftarrow cross[y_x] - cross[y_y]$
    /* Append to result list to finalize relative distance calculation */
    Append $cross$ to $frames$
$meanVectors \leftarrow$ EMPTY LIST
/* Find 10×10 matrix of relative mean x- and y-distances */
for $Player\ i$ in $frames$ do
    $meanDist \leftarrow$ EMPTY LIST
    for $Player\ j$ in $frames$ do
        /* Find mean x- and y-distances of one player to all others */
        $mean_x \leftarrow \sum vx_{i,j}/n$    // for n distance observations between the two players
        $mean_y \leftarrow \sum vy_{i,j}/n$
        /* Use x- and y-coordinate to find Euclidean 2D distance */
        $distance \leftarrow \sqrt{mean_x^2 + mean_y^2}$
        Append $distance$, $mean_x$, $mean_y$ to $meanDist$
    Append $meanDist$ to $meanVectors$
/* Find the centroid of formation by finding most frequent third nearest neighbor */
for $idx$ in $meanVectors$ do
    $centroidCount \leftarrow$ EMPTY OBJECT
    /* Authors declare 3rd nearest neighbor as centroid */
    $sortedVectors \leftarrow$ sort $meanVectors[idx]$
    /* Third closest neighbor lies at index 2 of sorted distance list */
    $thirdNearest \leftarrow sortedVectors[2]$
    if $thirdNearest$ not $\in$ $centroidCount$ then
        $centroidCount[thirdNearest] \leftarrow 1$
    else
        $centroidCount[thirdNearest]++$
/* Set centroid to be the player with the max count of ID as third nearest neighbor */
$centroid \leftarrow$ max centroidCount /* Starting at the centroid, find closest neighbor, then
    closest neighbor of that player, et cetera until all player position are determined
    relative to one another */
$playersList \leftarrow$ EMPTY LIST
```

Algorithm 1: Shaw and Glickman [66] FORMATION IMPLEMENTATION

```
/* For every player save its ID and the relative distance to determine position on field.
   Centroid starts normalized at (0, 0)                                                  */
$currentPlayer \leftarrow \{$id: playerID, vx: 0, vy: y$\}$
while $currentPlayer$ not $\emptyset$ do
    $old \leftarrow currentPlayer$
    /* Find idx 0 of sorted distances for current player to find closest neighbor       */
    $closestNeighbor \leftarrow$ sort $meanVector[currentPlayer][0]$ /* Update the currentPlayer
       variable to iterate through all field players - vx and vy vector build iteratively
       from player to player                                                             */
    $currentPlayer \leftarrow \{$closestNeighbor[id], old[vx] + closestNeighbor[vx], old[vy] +
      closestNeighbor[vy]$\}$
    Append closestNeighbor to players
    /* Check if all player IDs are already in the result variable                       */
    if $\mathbb{P}_{SequenceData} \setminus \mathbb{P}_{playersList} == 0$ then
        $currentPlayer \leftarrow \emptyset$
return $playersList$
```

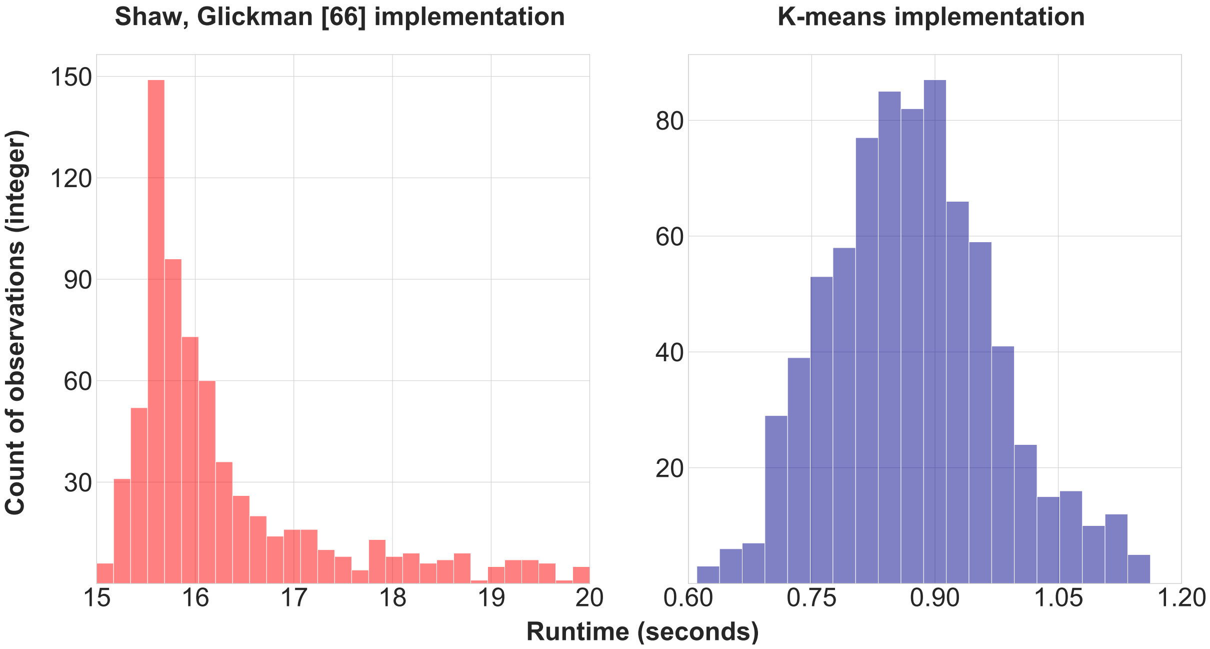


Figure 41: The figure displays the run time distributions of a total of 1,104 sequences (about 10% of the entire thesis dataset). The distributions highlight the drastic algorithmic improvement from the methods proposed by Shaw and Glickman [66] on the left to the performance of this thesis's innovative approach on the right. The visualization shifts the x- and y-scales for more intuitive comparability of the actual distributions, understating the actual order of magnitude difference from about 16+ seconds run time (left) to less than one second (right) per sequence.

		Distribution (seconds)				**Percentile**		
Method	**n**	**min**	**max**	**mean**	**σ**	**25%**	**50%**	**75%**
Shaw and Glickman [66]	1,104	0.20	200.01	**19.88**	20.72	15.56	16.03	19.05
Thesis' implementation	1,104	0.05	14.07	**1.06**	1.29	0.78	0.88	0.97

Table 2: The table displays the summary statistics of the run time distributions for the two implemented solutions to calculate formations. The sample size is n=1,104 for both cases. The run times represent a wide spread of the relative calculations of Shaw and Glickman [66] with a maximum of 200 seconds and a mean of almost ***20 seconds****. This relatively high value stands in contrast to the mean run time of about* ***one second*** *for the clustering-based approach proposed in this thesis. Following statistical convention, $\boldsymbol{\sigma}$ represents the standard deviation of the distributions.*

Furthermore, the large spread of the potential run times, indicated by the $\boldsymbol{\sigma}$ parameter in Table 2, demonstrates significantly different implications for the practical use case of a live-system integration. While both methods inherit a standard deviation of approximately one mean, an application building on these underlying calculations will suffer from detrimental effects with run time scenarios of more than 40 seconds. The local machine's processing power will impact the absolute run-time calculations in this thesis.[42] However, the 20:1 performance improvement will remain, even if the processing power of the underlying system improves.

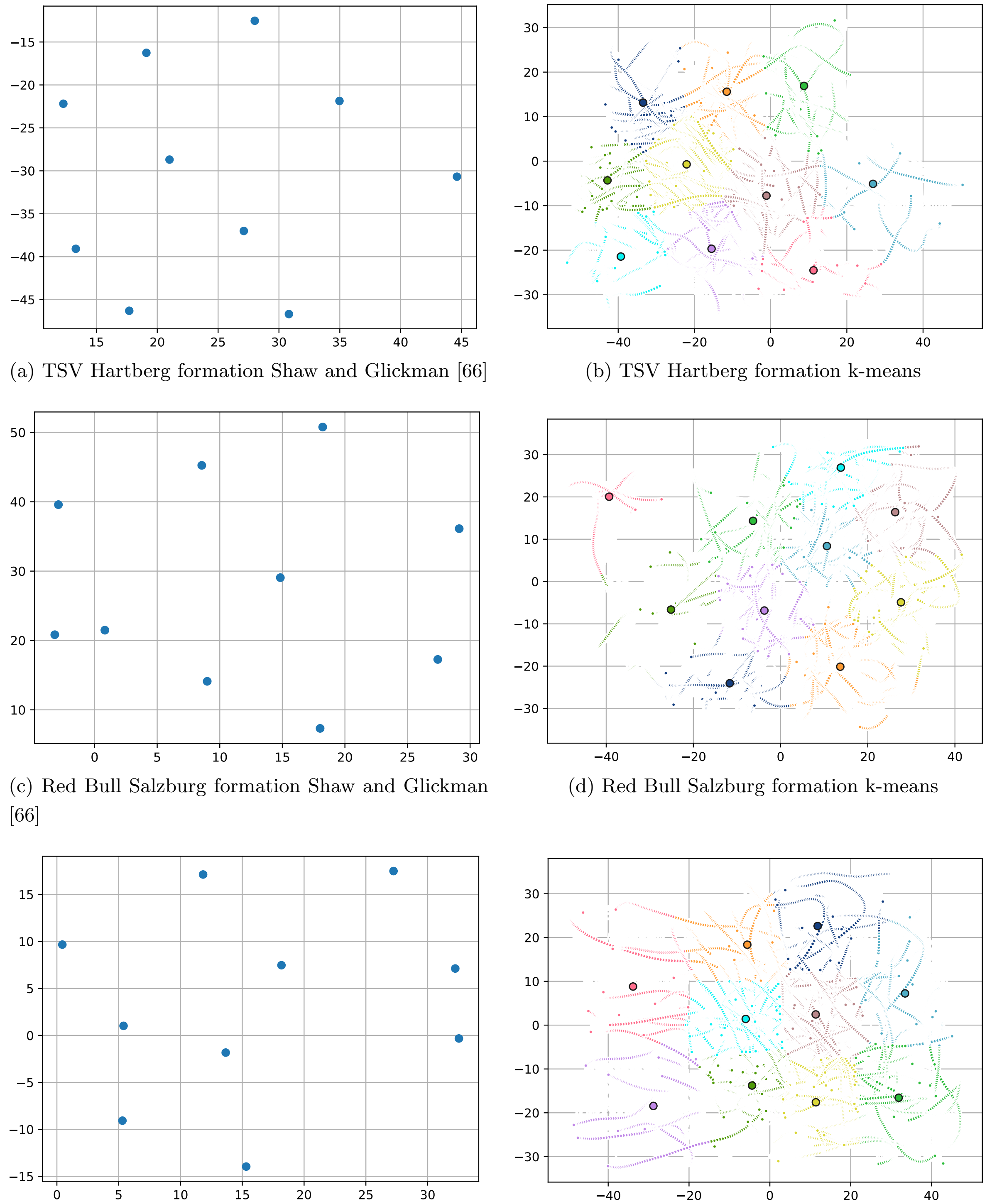
However, the isolated increase in performance will not account for much if the results are entirely different, or even worse, result in incorrect formations. To adhere with algorithmic requirement 🖩 A1 *Correct Formation Calculations*, the formation quality represents one of the vital qualities of the system. Chapter 7.2 discusses the overall comparability of the calculated formations with expected formations by experts. Figure 42 offers direct contrast of the results by this thesis' implementations of Shaw and Glickman's [66] algorithm and the original k-means clustering adaptation.

The anecdotal comparison in Figure 42 features three arbitrary sequences of three teams and contrasts their calculated formations according to the two approaches at hand. The formations of Figure 42a, 42c, and 42e represent the relative position coordinates after Shaw and Glickman [66]. These formations seem erratic and sensitive to outliers, while Figure 42b, 42d, and 42f suggest a smoother and more natural structure of a team's collective movement. The formations on the right are normalized to the middle of the field as explained in Chapter 4. Therefore, the axes' values are incomparable between the left and right sequences. Nonetheless, the overall team structures indicate similar tendencies, which hints at an overall correct formation calculation.

7.2 Expert Opinion

Interviews with uninvolved parties form the basis for the second part of the evaluation of the thesis results. Three international domain experts offered to first draw the respective formations for all teams of the data set, and second, offer critical qualitative feedback to the system itself. The first part of the interview runs without a previous introduction to the application's details—i.e., without seeing it—to avoid bias in their result presentation. This part of the evaluation quantifies the results of the first portion of the interview. The second part of qualitative feed-

[42] The comparison ran on an Intel(R) Core(TM) i5-6200U CPU processor with 2.30 GHz and 8.00 GB RAM.



(a) TSV Hartberg formation Shaw and Glickman [66]

(b) TSV Hartberg formation k-means

(c) Red Bull Salzburg formation Shaw and Glickman [66]

(d) Red Bull Salzburg formation k-means

(e) Lask Linz formation Shaw and Glickman [66]

(f) Lask Linz formation k-means

Figure 42: The figure compares the resulting formations for the method proposed by Shaw and Glickman [66] with the clustering approach of this thesis for arbitrary sequences of three teams. The shift on the x- and y-axis are expected as the k-means approach automatically centers the formation on the field. The formations seem relatively comparable, while the k-means formations of Figures 42b, 42d, and 42f portrait a more intuitive and less diffused formation. The algorithm's robustness to outliers automatically addresses some of the drawbacks of the alternative's algorithm.

back builds the backbone of possible extensions and challenges discussed in Chapter 7.3. The experts offer a comprehensive set of experience and skills to contribute to the system evaluation accurately. Expert A served as head and assistant coach in a top European league for almost ten years. He also holds qualifications as a video analyst and professional scout to complement his career as an international professional and national team player for almost 20 years. Expert B worked as an assistant coach in multiple European leagues for close to ten years. His expertise lies primarily in the superb familiarity with league data of the analyzed data set. Expert C served as a coach in the second and first leagues in Europe for multiple decades. His expertise lies in the well-founded knowledge of the data's discussed league, which qualifies him as a valuable asset to benchmark the derived formations.

Two of the three experts (B and C) have exclusively been exposed to the system's logic throughout this evaluation process and were therefore not involved in the application's development stage. This professional distance is essential to ensure critical unbiasedness and avoid a *self-fulfilling prophecy*[43] if an application offers the services a user has explicitly requested during previous development stages.

The interviews were scheduled to last about 90 minutes and held via Zoom[44] throughout February 2021. The meeting language was German and involved the experts, as mentioned earlier, holding Head- or Assistant-coach positions of top European clubs with additional certifications as professional sports- and video-analyst. The attendants afford a representative sample of non-technical but highly qualified users confronted with a technical tool built towards intuition. The following list offers an overview of the tentative schedule for each meeting. However, the open-ended questions and the exercise in drawing out formations remained flexible in duration, adapted to the attendee's schedule and available input.

Agenda for Expert Evaluation of Formation Analysis Tool

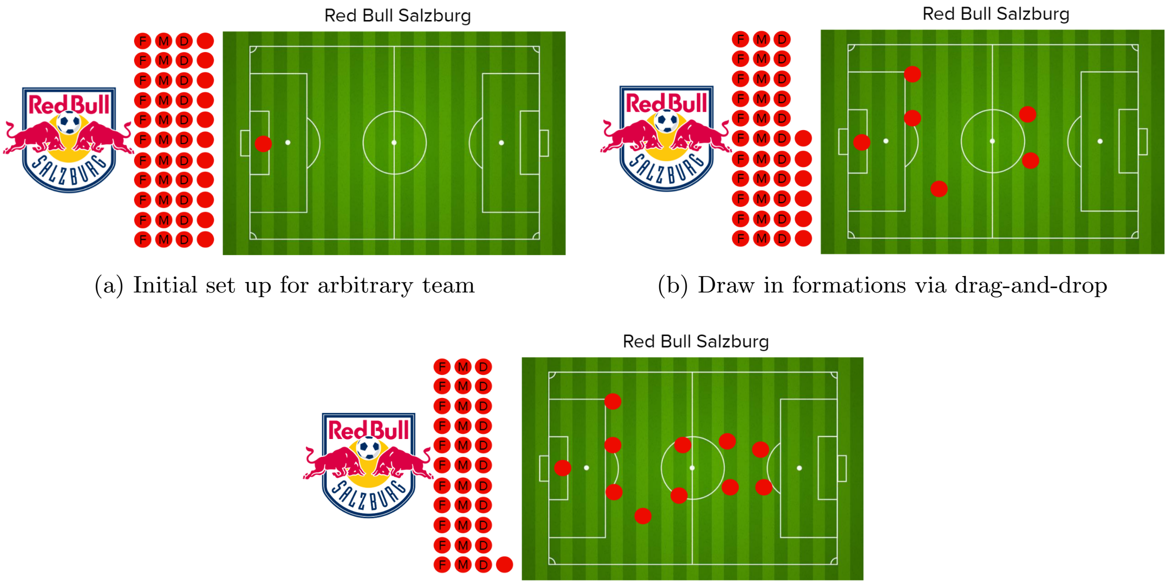
0. **Welcome and introduction to evaluation (10 minutes):**
   - Important to explain in detail what the tool is supposed to do and what it is not
1. **Ask participants to sketch out the average offensive formation for every team of the Austrian Bundesliga in webtool (45 minutes):**
   - Season averages (offense first / defense if asked)
   - Formations for Abstoss optional (if time allows offense and/or defense)
   - Open-ended questions regarding interesting differences among teams
2. **Presentation of the web tool (15 minutes):**
   - Short presentation of all three tabs
   - Focus on main functionalities

[43] A **self-fulfilling prophecy** represents a prediction, whose outcome is dependent on circumstances the predictor can impact. Therefore, the prediction will naturally come true since the alteration of the relevant variables alleviates all ambiguity.

[44] **Zoom** is one of the most popular video-conferencing tools. It is most known for its simplicity to attend a meeting without a necessary account subscription. Find their official website here.

3. **Open-Ended qualitative evaluation (15 minutes):**
    - Specific questions towards what kind of features are desirable / helpful in such a tool
    - First impression of things that look helpful, but more importantly things that are still to improve

An interactive web tool was prepared to conduct the comparison of manual and automatic formation analysis.[45] One of the main goals of this part was introducing an intuitive interface, offering the experts as much familiarity and comfort as possible to illustrate insights into the formations. Since a tactic board including magnets or a simple drawing of two-dimensional points offers the analog counterpart to this situation, a drag-and-drop environment for colored circles on a soccer field afforded the ideal digital compromise. Figure 43 highlights the typical workflow of a user within the evaluation.



(a) Initial set up for arbitrary team

(b) Draw in formations via drag-and-drop

(c) Final formation for comparison with tool

Figure 43: This illustration shows the core functionality of the interactive dashboard used for the thesis evaluation. The user can place colored circles on a familiar two-dimensional field layout. The drawing supports full drag-and-drop functionality to mimic the analog workflow of either drawing or visualizing formations with magnets on a tactic board. The three figures illustrate the general workflow for a single team with the initial setup in Figure 43a, the partially drawn formation in Figure 43b, and the finished formation in Figure 43c, which will subsequently be used for the quantitative evaluation.

The interface offered red circles next to an empty soccer field to illustrate the players' relative collective positions in line with familiar formation visualizations. The circle elements were either empty or filled with one of the three letters D (**D**efender), M (**M**idfielder), or F (**F**orward) to indicate player positions. The labeled circles' usage depends on how detailed the expert could recall each team's exact formations and remain strictly optional.

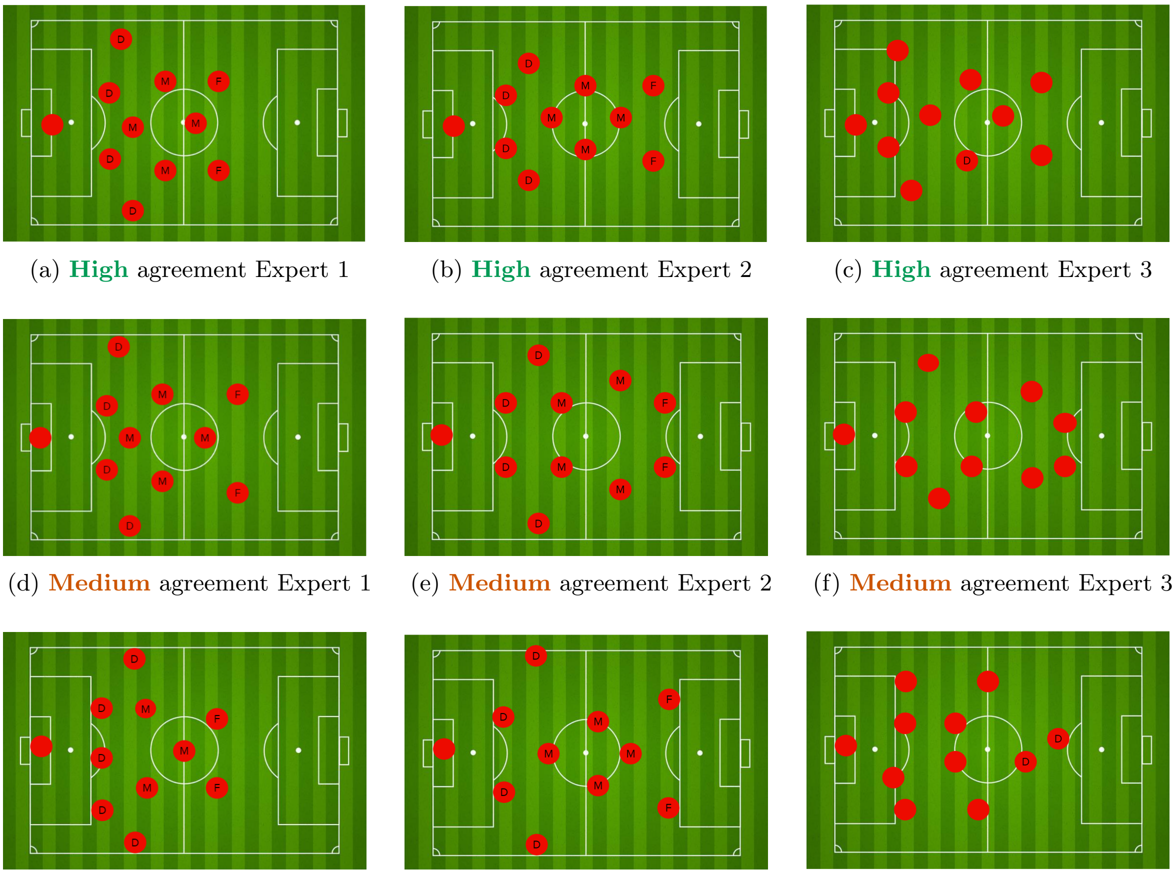
[45] The service provider used for the interactive exercise is Mural. This website offers interactive dashboards, mostly geared towards business applications offering intuitive drag-and-drop functionality, which easily extends to the discussion of sports formations.

The displayed steps in Figure 43 were then ideally repeated for all teams, depending on how certain the interviewee could recite a given team's formation from memory. The interviewers are presented with an empty soccer field, prepared in the interactive web-dashboard comparable to Figure 43a. The attendees continued to move the prepared positional indicator for all teams via drag-and-drop onto the soccer field to indicate a two-dimensional view of an average formation for a team.

	Formation expectations				
	Expert 1	**Expert 2**	**Expert 3**	**Thesis**	**Agreement**
Admira Wacker Mödling	3-3-2-2	2-4-3-1	4-4-2	3-3-2-2	low
Austria Wien	4-1-4-1	4-2-3-1	4-2-3-1	4-2-3-1	medium
LASK Linz	3-2-5	3-4-3	3-4-3	3-4-3	medium
Rapid Wien	4-2-3-1	4-2-3-1	4-4-2	4-2-3-1	medium
Red Bull Salzburg	4-3-3	4-2-2-2	4-2-2-2	4-3-3	medium
Rheindorf Alltach	4-3-3	4-3-3	4-3-3	4-3-3	high
Sankt Pölten	3-4-3	4-1-3-2	4-4-2	5-3-2	low
Sturm Graz	4-3-3	4-3-3	4-3-3	4-3-3	high
TSV Hartberg	4-1-4-1	4-2-3-1	4-4-2	4-4-2	low
Wolfsberger AC	4-1-3-2	4-1-3-2	4-1-3-2	4-1-3-2	high
WSG Wattens	3-4-1-2	3-3-2-2	4-2-3-1	4-2-3-1	low
FC Wacker Innsbruck	4-1-4-1	4-1-2-3	-	4-1-4-1	low
Joint prob. of agreement	**14/35**	**16/35**	**17/35**	**20/35**	
Change status quo	**+42.86%**	**+25.00%**	**+17.65%**	-	

*Table 3: The table summarizes the results from the quantitative portion of the expert evaluation. Three interviewees have provided typical formation blueprints for the teams in the data set, which were compared to the calculated results in the system. The agreement of the experts dictates the overall estimation difficulty of a formation—**high** if all three experts agreed, **medium** for an agreement of two, and **low** if all three experts chose a different formation for a team. The total score of agreed formations is displayed in the bottom row as the joint probability of agreement. Hence, every expert's formation can at most equal three other formations—with one observation missing—, which results in a possible max score of (12 × 3) - 1 = **35**. Additionally, the percentage change of the system to the expert joint probability agreement measure offers accuracy insights of the experts and the system in the last row.*

Table 3 indicates the results of the quantitative evaluation. The columns provide a label for the two-dimensional classification of the formations for each expert as well as the tool's results. For this purpose, the comparative value for the system was the conditional average formation of a team in the Conditional View—see Chapter 5.2 for a detailed explanation. The last column offers insights into the relative agreement of the experts **excluding** the tool. This classification offers the reader insights on the relative difficulty even for domain experts to classify a team's formation. The possible values would be **high** if a formation were classified similarly or with the same positions by all experts. The label **medium** indicates that two of three experts offered similar predictions but that the third one differs. If all three experts diverged, the label is **low**, hinting at widely different interpretations by the experts. To afford the reader a more precise picture of the extent of similarity that the labels explain, Figure 44 shows three examples from the data. The rows display illustrations from the interactive tools classified as high, medium, or low agreement.



(a) **High** agreement Expert 1

(b) **High** agreement Expert 2

(c) **High** agreement Expert 3

(d) **Medium** agreement Expert 1

(e) **Medium** agreement Expert 2

(f) **Medium** agreement Expert 3

(g) **Low** agreement Expert 1

(h) **Low** agreement Expert 2

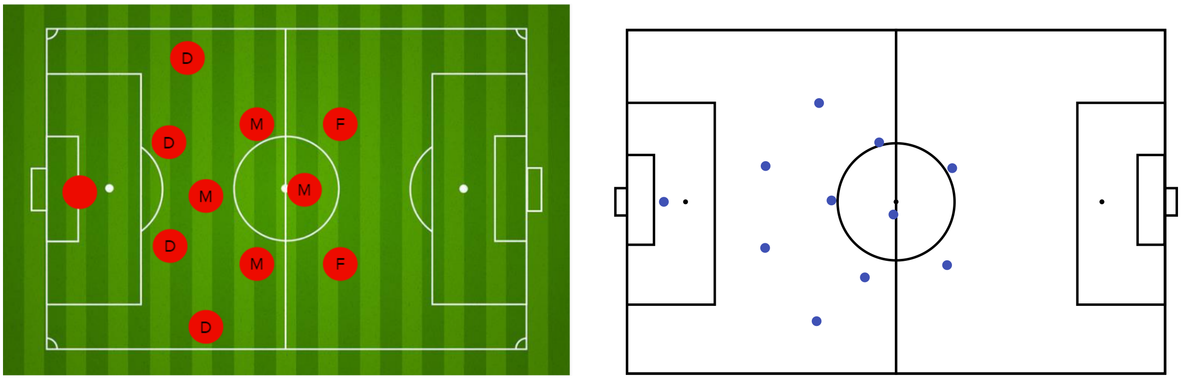
(i) **Low** agreement Expert 3

Figure 44: The visualization cluster exemplifies the standards for ***high****,* ***medium****, and* ***low*** *agreement among the interviewed experts. The top row shows three formations for the same team that were categorized as very similar, while the middle row depicts an agreement of two of three experts (here Figure 44e and Figure 44f). The bottom row indicates high uncertainty and low agreement with all three experts predicting different formations. The presence or absence of letters in the red circles are a direct result of the respective expert's either full-, partial-, or non-usage of the lettered circles. The results were not altered in any way since the experts have moved the circles.*

The row labeled *Joint prob. of agreement* quantifies the prediction quality. Without a gold-standard data set, the validity of a prediction is difficult to label correctly. However, in its nature, soccer formation data is messy and an imperfect representation of players' collective movement over time. Therefore, a measure of Inter-Rater Reliability[46] is implemented as a correctness measure. Its interpretation best paraphrases as a measure of correctness implied by the agreement with domain experts' repeated opinions. Given the balanced response size per rater—the number of total responses lies within a significantly close range of either twelve or eleven teams predicted—the most straightforward value of Inter-Rater Reliability is implemented: *joint probability distribution*. It measures the ratio of agreed answers to the number of total answers. The application's formation prediction was treated like the other experts' answers to improve the experts' comparability with the implemented system. This implementation logic results in a total number of possible agreements of 35 if every respective prediction equals the

[46] **Inter-Rater Reliability** is an essential concept in survey research. While the broad term is interpreted differently for specific use cases, it generally captures the extent of agreement among raters or survey participants.

other three responses. The numbers were intentionally not converted to decimals to improve readability. The last row offers a measure of the relative change of agreement from the system to the respective expert prediction. It calculates a simple percentage change[47] in joint probability agreement of using the proposed system over the expert opinions.



(a) Formation predicted by domain expert (b) Formation calculated by proposed system

Figure 45: The figure displays a common comparison of an expert' and the system's formation prediction. It illustrates how similar the results are for a good fit without any manual interference by a coach.

Figure 45 offers insights on the similarity of an accurate prediction by the system. Figure 45a depicts a manual prediction of a domain expert, while Figure 45b shows the same formation automatically calculated by the proposed system.

The results afford exciting and supportive evidence that the system achieved its algorithmic goal to provide ▦ A1 *Correct Formation Calculations*. It agrees with at least one of the experts for every team's average prediction and scores more total agreements than every other expert. The score of **20/35** lies **17.64%** above its closest competitor and increases the joint probability of agreement by **42.86%** compared to the most idiosyncratic expert. This performance demonstrates that with enough data, the system offers a higher degree of certainty than seasoned experts. This support provides a valuable addition to a coach's toolkit offering ample applications ranging from match preparation to direct player coaching.

7.3 Possible Extensions

While the existing results are exciting, the potential for further improvement is substantial. This chapter lays out potential extensions to the mechanics described in this thesis. It describes the reasoning for deliberate postponements of feature implementations (Chapter 7.3.1, 7.3.2, and 7.3.3), or why certain challenges prove excessively difficult to solve (Chapter 7.3.4). It also outlines the ambition towards the eventual development of a more holistic system through the incorporation and combination with adjacent systems (Chapter 7.3.5). All chapters include suggestions for future work on how to best address these challenges.

[47] The percentage change formula chosen for this evaluation is: $\frac{s-e_i}{e_i}$, with s as the joint probability of agreement of the system (20/35) and e_i the joint probability of agreement of the respective expert—so, either 14/35, 16/35, or 17/35.

7.3.1 Causal Design Choices Between Views

The system follows strict design requirements that adhere to iterative development stages and expert interviews. Nevertheless, the overall application layout leaves room for subjective design-interpretation. Thus far, the system is structured in three views as visualized by Figure 11 and explained in Chapter 5. The logic follows a hierarchical structure, where the Cluster View offers a broad overview of all formation groupings, the Conditional View provides insights into the team- and match-up-formations, while the Event View investigates match-internal dynamics. While this layout offers advantages, such as simple interpretability and a flat learning curve for user-onboarding, predominantly qualitative feedback during the evaluation phase exemplified a potential misalignment with the actual user reasoning process. These users are mostly uninterested in the vast set of information but rather a specific use case and search for causal explanations for that particular case. Therefore, a layout conforming with the user's reasoning first and the data structure second might prove a valid alternative to the proposed design.

This causal-structure might still incorporate a hierarchy in the data, but the navigation will be adjusted. The user will either be interested in a specific team or a specific match-up. Therefore, a filtering of the displayed data upfront might alleviate the massive data set's initial overwhelming nature. A Clustering View, similar to the proposed system, will offer an extensive overview of the selected team's formations within clusters. The main difference stems from the clickable interaction of the views. The user can navigate to conditional- or event-based information directly from the cluster summaries. This adaption to the navigation offers ample adjustments to visualized information, the overall control flow of the user, and the guidance of causal reasoning.

However, as with every design decision, this adjustment will be accompanied by a trade-off. A more directed design might prove more convoluted for unpredictable use cases. One of the design requirements requires the system to adhere to 🔍 **S4** ***Iterative Subsetting***, which can more clearly be stated as “Offer complexity as it is needed“. A causal system predicting the user's questions will dictate some of the choices to subset the data and query complexity level. Ideally, an extension will honor both paradigms in a compromise allowing for a directed but still unique user-experience.

7.3.2 Additional Event Subsetting

Chapter 5.3 describes the overall functionalities of the Event View. One of the core implementation features that experts repeatedly mentioned is the exploration of formations during specific events. Thus far, the long pass of the goalie displays a crucial descriptor for match development characteristics—how does a team build up its offense along the field. However, to fully align with the aspired design requirement 🔍 **S3** ***Easy Extensibility***, additional events will need to be included.

These events can comprise standard situations, such as *corners*, *free-kicks*, or *match beginning*, as well as additional finer subsetting options within the Event View. During the development stage, the *four phases of a game* became a focal point of expert interviews—see Chapter 3, A4. While the incorporated options already address aspects of match evolution, teams' typical dynamics by which they build their offense remain difficult to entangle.

A solution will need to utilize events filtering time frames of interest while incorporating information about the relative location on the field. Typical offense developments start with a lost ball or a missed opportunity by the opponent. Ball and player positions will help to flag these phases and compare the most informative time frames. This logic might include a finer subsetting of formation sequences than the two-minutes sequence logic defined in this thesis—see Figure 6. Two-minute sequences might prove reliable and robust for entire match insights, but the nature of a fast-paced game, such as soccer, could prove too volatile to be coherently understandable using this broad categorization. Nevertheless, this refinement will be highly use-case-specific and requires trial and error to weigh the pros and cons of alternative formation period durations.

Another possible extension of interest might be an indicator for defensive pressing behavior. The inclusion of pressing, i.e., high-pressure defense by potentially more than one defender, and describing the offense's reaction, became a recurring subject during feedback interviews. A subsetting to pressing situations enhances a coach's decision-making, understanding of opponents' reactions, and adequate match preparation. It requires an analysis of event data, where *quick passes* and *turnovers* might prove a valid starting point, with spatial data of high-density defender situations. Since this analysis will require the visualization of dynamic decision-making, an adjustment to the entire system's static nature might be necessary. See Chapter 7.3.5 for a proposed extension addressing a more dynamic solution.

7.3.3 Alternatives to K-means as Clustering Algorithm

This thesis thrives on its algorithmic improvements to the status quo of formation calculation. The computational complexity decreases to a fraction of current best practices. This acceleration builds one of the main ingredients to the immediate responsiveness of the app design. However, while the thesis offers extensive explanations for incorporating a clustering algorithm in the first place, it proposes no clear reasoning for the use of k-means over alternative clustering algorithms.

Once the problem of insufficient run times became prevalent, the pressure to find a working and straightforward alternative to the slow algorithms incorporated in the field [10, 66] led to the most prominent and reasonable solution. K-means offers everything the system requires: a predefined number of clusters (k), no further assumptions about the data distribution, and an intuitive interpretation. The time complexity of O(n^2) becomes negligible because the algorithm usually runs on 2,400 observations for a single sequence, where quadratic time complexity proves sufficient.

However, this reasoning should not discourage further developments to the clustering solution for formation determination. The closest neighbors to k-means are usually natural refinements to the vanilla algorithm. The following list provides an overview of center-based candidates for the clustering solution[48] as well as the pros and cons of the implemented k-means solution:

[48] A natural extension to this list will also consider broader alternatives to the clustering algorithms. An interesting group will be the cohort of expectation-maximization clustering algorithms.

1. **K-means**

   K-means represents the implemented clustering solution in this thesis. Appendix A outlines the calculation details and underlying logic. The algorithm runs in $O(n^2)$.

   Advantages and Disadvantages:

   + Simplest solution to implement for the given use case of ten predefined clusters.
   – Potentially converges to a local minimum.
   – Efficient solutions usually implement the algorithm with the Euclidean Distance, which limits the flexibility.

2. **K-medoids**

   K-medoids is one of the most natural refinements to the k-means algorithm [49]. It chooses actual data points as the cluster centroids and also runs in $O(n^2)$.

   Advantages and Disadvantages:

   + Allows for more interpretability of cluster centers as actual player locations.
   + Affords increased robustness to noise and outliers.
   + Accepts alternative dissimilarity measures than the Euclidean Distance.
   – All of the regular drawbacks of k-means, such as local optimum convergence or dependence on initial values, still apply.

3. **Fuzzy C-means**

   Fuzzy C-means clustering [8, 20] extends the k-means algorithm to a *fuzzy*[49] clustering logic. It operates similar to the k-means minimization of the total distance to the predefined number of centroids, but also adds a membership value $w_{i,j}$ where each $w_{i,j}$ describes the degree to which a data point x_i belongs to cluster c_j. It also requires the fuzzifier $m \in \mathbb{R}$ with $m \geq 1$ determining the cluster fuzziness with larger m resulting in smaller membership values $w_{i,j}$.

   Advantages and Disadvantages:

   + Introduces the notion of the likelihood for a given player to play at a specific location.
   – More necessary assumptions need to be made beforehand (m to determine how fuzzily the allocation operates).
   – All of the regular drawbacks of k-means, such as local optimum convergence or dependence on initial values, still apply.

4. **K-harmonic means**

   K-harmonic means alters the objective function of k-means—the total distance metric—to an *Harmonic Average* metric. The Harmonic Average describes the reciprocal of the

[49] **Fuzzy** boundaries refer to classification or clustering approaches that group elements usually via a probability or weighting metric into multiple clusters. They contrast the distinct logic of hard clustering, where each data point belongs to one cluster.

arithmetic average of the numbers' reciprocals in a set. It is most applicable to find the average of rates or average travel speed because it alleviates the necessity to find common denominators of the included fractions. The following formula describes this calculation.

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n} \right)^{-1},$$

where x_1, x_2, ..., x_n are positive real numbers. Zhang et al. [91] outline the full derivation of the objective function, which exceeds this brief overview's scope.

Advantages and Disadvantages:

+ Results are independent of cluster initialization.
– Computationally more costly than k-means.
– Results can still converge to a local minimum.

This brief overview does not aim to provide an exhaustive list of possible extensions but offers a starting point for future work. While k-means outperforms the current best practices in the field, it suffers from ignorance of data-specific properties. Therefore, an improved approach to fine-tune the clustering solution for formation calculations could alleviate particular challenges, such as accurately identifying the exact player covering a specific location on the field and preserving some of the more granular notions of formation shifts. K-means' main strength lies in its simple implementation, intuitive interpretation, and fast run time. Any clustering approach that preserves these properties while potentially addressing inherent drawbacks presents a valuable enhancement.

7.3.4 Improved Possession Derivation

One of the major challenges of calculating different formations is the exact determination of possession. Here, the possession of the ball distinctly defines which team is in offense and defense. Chapter 4 outlines the backbone of the heuristic used in this thesis. It builds on an event-based subsetting of time frames to introduce candidate time frames for when a given team controls the ball. However, qualitative feedback explained in Chapter 7.2 explains significant differences in real-world behavior by a team in offense and defense. This phenomenon is not observable in our data. As Figure 17 in Chapter 5.1 illustrates, most clusters center around a 50:50 split between offensive and defensive formations. The discrepancy to the expected behavior might stem from three main reasons:

1. While the thesis implements measures to increase robustness–the two-minute segmentation of time frames illustrated in Figure 6 and the deletion of possessions under five seconds—it might still include a majority of quick exchanges of possession. This characteristic could then lead to the classification of spurious formations and resemble a more chaotic behavior than a structured shift from offense to defense.

   $\Longrightarrow$ While possible, the probability of this atypical tendency tainting enough possessions of all sequences seems low.

2. The inspected European league behaves erratically in comparison to the expected behavior of top leagues. The data set under examination represents a smaller European league's characteristics, which might not be as evolved as other top leagues and could behave differently than the experts expect.

   ⟹ The experts are experienced in international and specifically in the league's national data. This personal background did not stop them from expecting varying formations between offense and defense. Therefore, the league's idiosyncratic structure should not account for the formation's peculiar distribution between offense and defense.

3. The thesis heuristic for offense and defense formation misses the intricacies of the data. It utilizes events associated with a specific player to determine the possession at that exact time frame. As long as no other event occurs, the team is assumed to remain in possession until an event featuring a player of the opposing team occurs, which then switches the possession flag. This logic excludes a list of events—see Chapter 4—which are not associated with a specific player. The approach might suffer from either data quality concerns or the general heuristic of event subsetting, which proves too simplistic to catch the complex concept of *possession.*

   ⟹ This appears to be the most likely cause of the inaccuracy, which needs further fine-tuning.

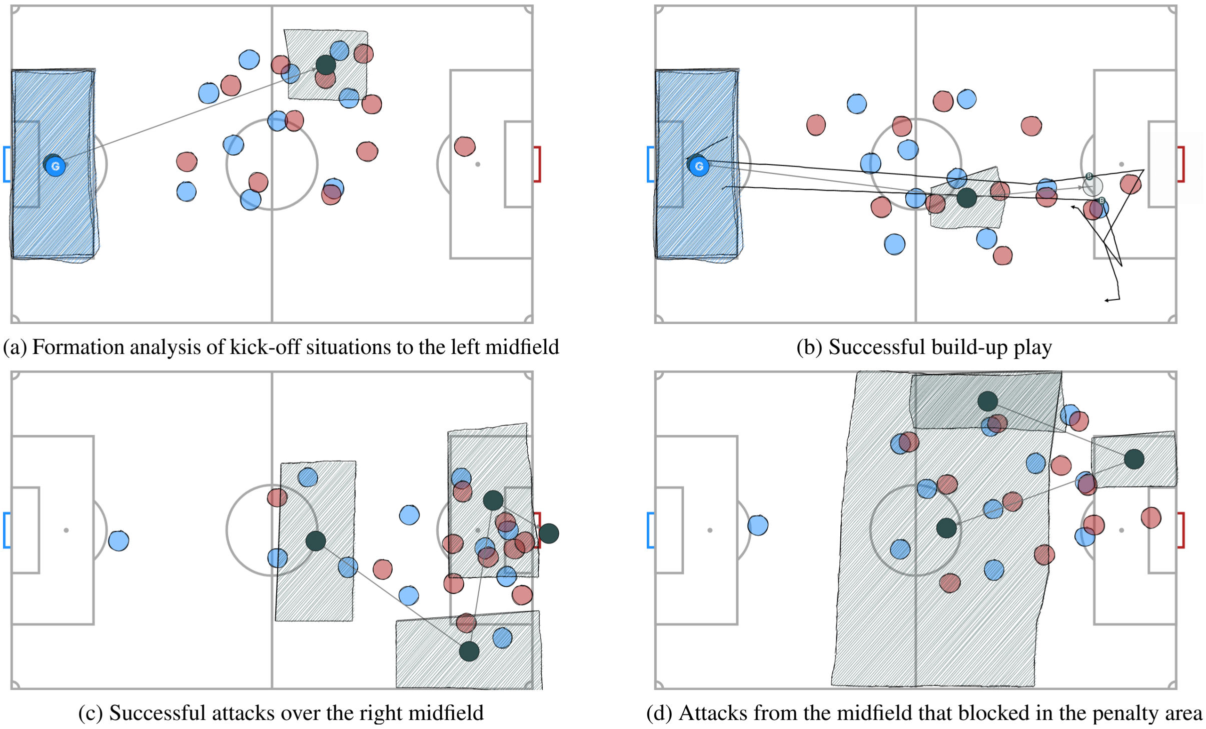
The challenge of real offense and defense determination is prevalent throughout the soccer and tracking literature. Thus far, events exclusively determine the relevant time frames for which a team is in possession. However, while simplistic enough to address a broad division, finer propensities, especially relevant for formation analysis, will be lost. Future approaches will have to combine the event and tracking data to build a holistic picture of the field's location. Hitherto, approaches largely ignore ball and player locations because of the noise it introduces. First and foremost, an exact derivation of possession exceeds the simple comparison of the distance between players and ball; otherwise, every pass across the field will result in multiple incorrect possession flags.

Nevertheless, a cohesive and intricate combination of the event and tracking data might offer insights, while this thesis's heuristics are based solely on event data incorporation. Therefore, a proposed next step to solve this puzzle incorporates both data sources and the inclusion of probabilities for possession depending on the location on the field. Machine learning solutions can build advanced classifiers utilizing all three sources to improve the forecast accuracy for possession possibility. The main challenge will be to align these advanced solutions with computational requirements of fast processing to address practicality concerns of its use cases.

7.3.5 Dynamic Interactions

The system's goal is ambitious in its very nature: describe the collective movements of groups of people to distill a footprint that describes tendencies to exploit strategically. The application offers a leap towards an accurate static description of how teams behave in specific situations. However, soccer is dynamic, and players react to situations independent of their formation assignment. They prepare and train for overall formations, but more on a more granular level, incorporate dynamically changing information of specific passes, defensive pressure, and spacing alterations into their decision-making and behavior. To illustrate it with an example: if a soccer team were an orchestra, the formation would be the composition of the instruments on stage. However, the harmonies, the intricacies of volume, individual skill levels, and the collective melodic timing represent the ingredients that bring the notes to life. These corresponding insights in soccer should be the grand goal of a supportive system of automatic data analysis.

Core steps towards this ambitious goal include formation information with an interactive dashboard. The user's expertise should drive explorative data analysis, moving players while other players react dynamically. This extension combines a similar interface to the evaluative interviews described in Chapter 7.2 with an AI-powered back end. It combines this thesis's formation information with dynamic responses to interactive user-input. Experts can mimic typical match situations with an intricate interplay visualized by the system. Exciting work in this field is currently under submission [63] which addresses some of these goals. This sketching tool affords the facilitation of data queries for massive data sets through an interactive drawing board. The user can sketch any situation, and the back-end logic of the board will highlight matches and situations of specific teams that correspond with the drawn sequence.



(a) Formation analysis of kick-off situations to the left midfield

(b) Successful build-up play

(c) Successful attacks over the right midfield

(d) Attacks from the midfield that blocked in the penalty area

Figure 46: Figure 5 of Seebacher et al. [63] illustrates the workflow and the potential that a dynamic sketchboard has for efficient queries of large data sets within a comfortable and intuitive framework.

This sketching, which mainly serves as a querying tool, can benefit from the formation analysis's informative insights. By extending functionalities with interactive drag-and-drop components triggering relative player movements, the frontier of possible angles to analyze collective movement behavior will shift dramatically.

The main impediment to this goal will be an efficient algorithmic logic. This thesis has shown a computationally faster alternative to derive formations, but this envisioned solution requires further extensions to solve the more complex use case of responsive team movements. Additionally, it will build on machine learning logic to explore potential player movements that did not necessarily occur. Efficient implementations will leverage adjacent work in the field of *What-if*-analysis in soccer [74] and its real-time adaptation.

8 Discussion & Conclusion

This thesis's primary goal is to offer insightful information for soccer stakeholders with a measurable impact on teams' success. Analyzing team formations has allowed us to describe and assess collective movements and, therefore, quantify one of the few direct strategic channels by which a coach can influence his team's collective behavior. The demand for incorporating data science practices and principles into sports, and especially into soccer, is on the rise. While the literature on formation analysis is growing, a disconnect between scientific insights and practical solutions has limited this developing field's impact. Through a thorough interviewing and development stage, we have isolated necessary ingredients to close this research gap and prescribed design requirements for a system that combines scientific rigor with an intuitive and user-friendly layout. The result culminates in a formation analysis tool that offers faster, more accurate, and more intuitive insights than any past solution.

The main contributions of this thesis are four-fold. It improves the algorithmic speed to a fraction of previous solutions (Chapter 8.1) while improving the forecasting accuracy beyond the level of trained professionals (Chapter 8.2). It allows for multi-match comparison of formations while providing novel visualization options of inter-match analyses (Chapter 8.3), and it combines these advancements in an intuitive design developed in close collaboration with domain experts to solve real practical challenges in the field (Chapter 8.4). The thesis closes with concluding remarks (Chapter 8.5).

8.1 Improved Algorithmic Performance

The proposed system visualizes formations, but at its heart lies a novel way to calculate the average position of collective movement patterns to describe a soccer formation over a pre-defined period. These novel calculation methods offer a temporal complexity that is twenty times faster than our implementation of the current status quo (Figure 41). Solutions in past work [66] were designed for purely scientific use cases, where real-time performance speed is unimportant. Calculation times of close to twenty seconds are impractical for a system developed to offer immediate feedback and real-world implementation. The run times are also more centered (see Figure 41), which is suggestive of more reliable results with smaller distributional tails, i.e., less frequent outliers for the process run times. Given the necessary assumptions for a comparative implementation, this approach not only decreases run times dramatically but also produces smoother, more natural formations that depict less sensitivity to outliers (Figure 42).

8.2 Improved Predictive Accuracy

Extensive validation interviews allowed objective evaluations to compare the calculated results to practical knowledge. Two major conclusions emerged from these interviews:

1. Formation prediction is difficult. This characteristic stems from multiple reasons, but collective movements' continuous nature challenges most attempts to discretize these data into average positions. Formations are also continuously changing and might be impacted by various factors: coaching style, player availability, specific match-up, seasonal standings, or period during a match. Even manual classification of formations by domain experts with a seasoned background in the industry only agree on a fraction of forecasts. Figure 44 and Table 2 reflect this discrepancy by exemplifying the degrees of disagreement with examples from the interview sessions.

2. Our system results in more agreement within the sample space of predictions than any individual expert. The implemented system also exhibits a 17.65% increase in the *joint probability agreement*—a measure of the inter-rater reliability— over the current best analog predictor (Table 3).

8.3 Inclusion of Multi-Match Functionalities

In addition to increasing the speed and accuracy of formation forecasting, our system also improves previous work by offering multi-match functionality. Previous efforts to visualize collective movements primarily aimed towards scientific applications and detailed single-match analysis. This limitation decreases the informative value to compare matches and teams dynamically through a more extensive time-series analysis. Although some research projects have expanded their data samples beyond single matches, they have not offered a system to visualize and analyze this rich data interactively. This thesis offers the first system that combines the best of both worlds, introducing an application design that efficiently manages the formations' calculation while also providing the capability for multi-match comparisons.

8.4 Development of an Intuitive App Design

The development of the application produced by this thesis implements design requirements derived from interviews with domain practitioners (Chapter 3). These needs then dictated the derivation of four 🖩 **Algorithmic Requirements**, four 🔍 **Search/Subsetting Requirements**, and four 📊 **Visualization Requirements**. The entire web layout, including the macro separation of individual tabs, the simple micro option-subsetting intuition, and the inclusion of practical features, solve the search and visualization requirements. This cumulative effort of a consistent end-user first experience resulted in a design that experts described as *intuitive* and *comfortable* during the qualitative session of the interviews (Chapter 7.2).

8.5 Concluding Remarks

This thesis offers advancement to the field of applied computer science. It introduces a system that outperforms seasoned experts in prediction accuracy, improves algorithmic standards, and combines these benefits into an intuitive application that reflects practicality and design requirements. This application, which pushed modern soccer research boundaries past existing benchmarks, enables coaches to transform data science insights into tangible improvements to team performance. These performance improvements could translate to larger profits for the soccer organizations utilizing the application. By building upon the recommendations of experts through reflective feedback collection, this thesis also lays out a wide array of challenges and potential solutions for future work to push soccer research boundaries even further.

References

[1] Catapult — we create technology to help athletes and teams perform to their true potential. https://www.catapultsports.com/. (Accessed on 01/13/2021).

[2] Earth mover's distance - wikipedia. https://en.wikipedia.org/wiki/Earth_mover%27s_distance. (Accessed on 01/19/2021).

[3] Sports market size worldwide 2018 — statista. https://www.statista.com/statistics/1087391/global-sports-market-size/#:~:text=In%202018%2C%20the%20global%20sports,global%20sports%20market%20in%202018. (Accessed on 01/23/2021).

[4] Wasserstein metric - wikipedia. https://en.wikipedia.org/wiki/Wasserstein_metric. (Accessed on 01/19/2021).

[5] Gennady Andrienko, Natalia Andrienko, Guido Budziak, Jason Dykes, Georg Fuchs, Tatiana von Landesberger, and Hendrik Weber. Visual analysis of pressure in football. *Data Mining and Knowledge Discovery*, 31(6):1793–1839, 2017.

[6] Adrià Arbués-Sangüesa, Gloria Haro, Coloma Ballester, and Adrián Martín. Head, shoulders, hip and ball... hip and ball! using pose data to leverage football player orientation.

[7] Jürgen Assfalg, Marco Bertini, Carlo Colombo, Alberto Del Bimbo, and Walter Nunziati. Semantic annotation of soccer videos: automatic highlights identification. *Computer vision and image understanding*, 92(2-3):285–305, 2003.

[8] James C Bezdek. *Pattern recognition with fuzzy objective function algorithms.* Springer Science & Business Media, 2013.

[9] Alina Bialkowski, Patrick Lucey, Peter Carr, Yisong Yue, and Iain Matthews. Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors. In *Proceedings of 8th annual MIT sloan sports analytics conference*, pages 1–7. Citeseer, 2014.

[10] Alina Bialkowski, Patrick Lucey, Peter Carr, Yisong Yue, Sridha Sridharan, and Iain Matthews. Identifying team style in soccer using formations learned from spatiotemporal tracking data. In *2014 IEEE international conference on data mining workshop*, pages 9–14. IEEE, 2014.

[11] Alina Bialkowski, Patrick Lucey, Peter Carr, Yisong Yue, Sridha Sridharan, and Iain Matthews. Large-scale analysis of soccer matches using spatiotemporal tracking data. In *2014 IEEE International Conference on Data Mining*, pages 725–730. IEEE, 2014.

[12] Alina Bialkowski, Patrick Lucey, Peter Carr, Iain Matthews, Sridha Sridharan, and Clinton Fookes. Discovering team structures in soccer from spatiotemporal data. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2596–2605, 2016.

[13] Paul S Bradley and Usama M Fayyad. Refining initial points for k-means clustering. In *ICML*, volume 98, pages 91–99. Citeseer, 1998.

[14] Paul S Bradley, Chris Carling, Dave Archer, Jenny Roberts, Andrew Dodds, Michele Di Mascio, Darren Paul, Antonio Gomez Diaz, Dan Peart, and Peter Krustrup. The effect of playing formation on high-intensity running and technical profiles in english fa premier league soccer matches. *Journal of sports sciences*, 29(8):821–830, 2011.

[15] J.C. Brigola. *Classification of multi-agent trajectories.* Master's thesis, EPFL, 2012.

[16] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[17] Christopher Carling, A Mark Williams, and Thomas Reilly. *Handbook of soccer match analysis: A systematic approach to improving performance.* Psychology Press, 2005.

[18] Julen Castellano, Nicolas Evans, Javier Fernández, Daniel Link, Lucca Pappalardo, Daniel Memmert, Sam Robertson, Jaime Sampaio, Manuel Stein, Jan Van Haaren, et al. Football analytics: Now and beyond.

[19] Janusz Czopik. An application of the hungarian algorithm to solve traveling salesman problem. *American Journal of Computational Mathematics*, 09:61–67, 01 2019. doi: 10.4236/ajcm.2019.92005.

[20] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

[21] Jack Edmonds and Richard M Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM)*, 19(2):248–264, 1972.

[22] Ahmet Ekin, A Murat Tekalp, and Rajiv Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image processing*, 12(7):796–807, 2003.

[23] Javier Fernández, Luke Bornn, and Dan Cervone. Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In *13 th Annual MIT Sloan Sports Analytics Conference*, 2019.

[24] Pasi Fränti and Sami Sieranoja. How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, 93:95–112, 2019.

[25] Mengran Gou, Srikrishna Karanam, Wenqian Liu, Octavia Camps, and Richard J Radke. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–19, 2017.

[26] Sachiko Iwase and Hideo Saito. Tracking soccer players based on homography among multiple views. In *Visual Communications and Image Processing 2003*, volume 5150, pages 283–292. International Society for Optics and Photonics, 2003.

[27] Halldor Janetzko, Dominik Sacha, Manuel Stein, Tobias Schreck, Daniel A Keim, Oliver Deussen, et al. Feature-driven visual analytics of soccer data. In *2014 IEEE conference on visual analytics science and technology (VAST)*, pages 13–22. IEEE, 2014.

[28] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[29] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[30] Urho M Kujala, Simo Taimela, Ilkka Antti-Poika, Sakari Orava, Risto Tuominen, and Pertti Myllynen. Acute injuries in soccer, ice hockey, volleyball, basketball, judo, and karate: analysis of national registry data. *Bmj*, 311(7018):1465–1468, 1995.

[31] Jim ZC Lai and Yi-Ching Liaw. Improvement of the k-means clustering filtering algorithm. *Pattern Recognition*, 41(12):3677–3681, 2008.

[32] Yongjin Lee and Seungjin Choi. Minimum entropy, k-means, spectral clustering. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 1, pages 117–122. IEEE, 2004.

[33] Dawei Liang, Yang Liu, Qingming Huang, and Wen Gao. A scheme for ball detection and tracking in broadcast soccer video. In *Pacific-Rim Conference on Multimedia*, pages 864–875. Springer, 2005.

[34] Jia Liu, Xiaofeng Tong, Wenlong Li, Tao Wang, Yimin Zhang, and Hongqi Wang. Automatic player detection, labeling and tracking in broadcast soccer video. *Pattern Recognition Letters*, 30(2):103–113, 2009.

[35] Sergio Llana, Pau Madrero, Javier Fernández, and FC Barcelona. The right place at the right time: Advanced off-ball metrics for exploiting an opponent's spatial weaknesses in soccer. In *Proceedings of the 14th MIT Sloan Sports Analytics Conference*, 2020.

[36] Patrick Lucey, Dean Oliver, Peter Carr, Joe Roth, and Iain Matthews. Assessing team strategy using spatiotemporal data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1366–1374, 2013.

[37] Jonathan Ma. *An Analysis of Formation Disruption in Soccer*. PhD thesis, 2020.

[38] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[39] Md Sohrab Mahmud, Md Mostafizer Rahman, and Md Nasim Akhtar. Improvement of k-means clustering algorithm with better initial centroids based on weighted average. In *2012 7th International Conference on Electrical and Computer Engineering*, pages 647–650. IEEE, 2012.

[40] Andrii Maksai, Xinchao Wang, and Pascal Fua. What players do with the ball: A physically constrained interaction modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 972–981, 2016.

[41] Daniel Memmert, Koen APM Lemmink, and Jaime Sampaio. Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine*, 47(1):1–10, 2017.

[42] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

[43] Christopher Mutschler, Holger Ziekow, and Zbigniew Jerzak. The debs 2013 grand challenge. In *Proceedings of the 7th ACM international conference on Distributed event-based systems*, pages 289–294, 2013.

[44] Takuma Narizuka and Yoshihiro Yamazaki. Characterization of the formation structure in team sports. *arXiv preprint arXiv:1802.06766*, 2018.

[45] Takuma Narizuka and Yoshihiro Yamazaki. Clustering algorithm for formations in football games. *Scientific reports*, 9(1):1–8, 2019.

[46] Catherine D Newell, Mark D Wood, Kathleen M Costello, and Robert B Poetker. Automatic story creation using semantic classifiers for digital assets and associated metadata, January 13 2015. US Patent 8,934,717.

[47] Yoshinori Ohno, J Miurs, and Yoshiaki Shirai. Tracking players and a ball in soccer games. In *Proceedings. 1999 IEEE/SICE/RSJ. International Conference on Multisensor Fusion and Integration for Intelligent Systems. MFI'99 (Cat. No. 99TH8480)*, pages 147–152. IEEE, 1999.

[48] Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[49] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

[50] Patrick Pérez, Carine Hue, Jaco Vermaak, and Michel Gangnet. Color-based probabilistic tracking. In *European Conference on Computer Vision*, pages 661–675. Springer, 2002.

[51] Charles Perin, Romain Vuillemot, and Jean-Daniel Fekete. Soccerstories: A kick-off for visual soccer analysis. *IEEE transactions on visualization and computer graphics*, 19(12): 2506–2515, 2013.

[52] Charles Perin, Romain Vuillemot, Charles D Stolper, John T Stasko, Jo Wood, and Sheelagh Carpendale. State of the art of sports data visualization. In *Computer Graphics Forum*, volume 37, pages 663–686. Wiley Online Library, 2018.

[53] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[54] Lukas Probst, Frederik Brix, Heiko Schuldt, and Martin Rumo. Real-time football analysis with streamteam. In *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*, pages 319–322, 2017.

[55] Lukas Probst, Fabian Rauschenbach, Heiko Schuldt, Philipp Seidenschwarz, and Martin Rumo. Integrated real-time data stream analysis and sketch-based video retrieval in team sports. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 548–555. IEEE, 2018.

[56] Marirose A Radelet, Scott M Lephart, Elaine N Rubinstein, and Joseph B Myers. Survey of the injury rate for children in community sports. *Pediatrics*, 110(3):e28–e28, 2002.

[57] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh. Pose machines: Articulated pose estimation via inference machines. In *European Conference on Computer Vision*, pages 33–47. Springer, 2014.

[58] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[59] Stephen J Roberts, Richard Everson, and Iead Rezek. Minimum entropy data partitioning. 1999.

[60] Stephen J. Roberts, Christopher Holmes, and Dave Denison. Minimum-entropy data partitioning using reversible jump markov chain monte carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):909–914, 2001.

[61] Dominik Sacha, Feeras Al-Masoudi, Manuel Stein, Tobias Schreck, Daniel A Keim, Gennady Andrienko, and Halldór Janetzko. Dynamic visual abstraction of soccer movement. In *Computer Graphics Forum*, volume 36, pages 305–315. Wiley Online Library, 2017.

[62] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics*, 18(12):2431–2440, 2012.

[63] Daniel Seebacher, Manuel Stein, Tom Polk, Halldor Janetzko, Daniel Keim, and Tobias Schreck. Investigating the sketchplan: A novel way of identifying tactical behavior in massive soccer datasets. *IEEE Transactions on Visualization and Computer Graphics*, Under review.

[64] Yongduek Seo, Sunghoon Choi, Hyunwoo Kim, and Ki-Sang Hong. Where are the ball and players? soccer game analysis with color-based tracking and image mosaick. In *International Conference on Image Analysis and Processing*, pages 196–203. Springer, 1997.

[65] Laurie Shaw. Structure in football: Putting formations into context. *Barça Sports Analytics Summit*, 2020.

[66] Laurie Shaw and Mark Glickman. Dynamic analysis of team strategy in professional football. *Barça Sports Analytics Summit*, 2019.

[67] William Spearman. Beyond expected goals. In *Proceedings of the 12th MIT sloan sports analytics conference*, pages 1–17, 2018.

[68] 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.

[69] Manuel Stein. *Visual Analytics for Cooperative and Competitive Behavior in Team Sports*. PhD thesis, Universität Konstanz, Konstanz, 2020.

[70] Manuel Stein, Halldór Janetzko, Thorsten Breitkreutz, Daniel Seebacher, Tobias Schreck, Michael Grossniklaus, Iain D Couzin, and Daniel A Keim. Director's cut: Analysis and annotation of soccer matches. *IEEE computer graphics and applications*, 36(5):50–60, 2016.

[71] Manuel Stein, Halldór Janetzko, Andreas Lamprecht, Daniel Seebacher, Tobias Schreck, Daniel Keim, and Michael Grossniklaus. From game events to team tactics: Visual analysis of dangerous situations in multi-match data. In *2016 1st International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)*, pages 1–9. IEEE, 2016.

[72] Manuel Stein, Halldor Janetzko, Andreas Lamprecht, Thorsten Breitkreutz, Philipp Zimmermann, Bastian Goldlücke, Tobias Schreck, Gennady Andrienko, Michael Grossniklaus, and Daniel A Keim. Bring it to the pitch: Combining video and movement data to enhance team sport analysis. *IEEE transactions on visualization and computer graphics*, 24 (1):13–22, 2017.

[73] Manuel Stein, Daniel Seebacher, Tassilo Karge, Tom Polk, Michael Grossniklaus, and Daniel A Keim. From movement to events: Improving soccer match annotations. In *International Conference on Multimedia Modeling*, pages 130–142. Springer, 2019.

[74] Manuel Stein, Daniel Seebacher, Rui Marcelino, Tobias Schreck, Michael Grossniklaus, Daniel A Keim, and Halldor Janetzko. Where to go: Computational and visual what-if analyses in soccer. *Journal of sports sciences*, 37(24):2774–2782, 2019.

[75] Manuel Stein, Philip Zimmermann, Markus Schopp, Jens Pruessner, Michael Grossniklaus, Tobias Schreck, and Daniel A. Keim. Identifying the match plan: Context-oriented interactive visual analysis of football matches. In *Football Analytics: Now and Beyond. A deep dive into the current state of advanced data analytics*, volume 1, chapter 10, pages 146–173. Barça Innovation Hub, 11 2019.

[76] Manuel Stein, Philip Zimmermann, Markus Schopp, Tobias Schreck, Michael Grossniklaus, Daniel A Keim, and Johann Pruessner. Identifying the match plan: Context-oriented interactive visual analysis of football matches. In *Football Analytics, Now and Beyond, A deep dive into the current state of advanced data analytics*, volume 1, pages 146–173. Barca Innovation Hub, 2019.

[77] Hugo Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1 (804):801, 1956.

[78] David J Stracuzzi, Alan Fern, Kamal Ali, Robin Hess, Jervis Pinto, Nan Li, Tolga Konik, and Daniel G Shapiro. An application of transfer to american football: From observation of raw video to control in a simulated environment. *AI Magazine*, 32(2):107–125, 2011.

[79] Fabio Sulser, Ivan Giangreco, and Heiko Schuldt. Crowd-based semantic event detection and video annotation for sports videos. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pages 63–68, 2014.

[80] David Sumpter. *Soccermatics: mathematical adventures in the beautiful game*. Bloomsbury Publishing, 2016.

[81] Nobuaki Tomizawa. On some techniques useful for solution of transportation network problems. *Networks*, 1(2):173–194, 1971.

[82] Vasanth Tovinkere and Richard J Qian. Detecting semantic events in soccer games: Towards a complete solution. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 212–212. IEEE Computer Society, 2001.

[83] SS Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

[84] Cédric Villani. *Topics in optimal transportation.* Number 58. American Mathematical Soc., 2003.

[85] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001.

[86] Xinchao Wang, Vitaly Ablavsky, Horesh Ben Shitrit, and Pascal Fua. Take your eyes off the ball: Improving ball-tracking by focusing on team play. *Computer Vision and Image Understanding*, 119:102–115, 2014.

[87] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[88] Y. Wu, X. Xie, J. Wang, D. Deng, H. Liang, H. Zhang, S. Cheng, and W. Chen. Forvizor: Visualizing spatio-temporal team formations in soccer. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):65–75, 2019. doi: 10.1109/TVCG.2018.2865041.

[89] Lexing Xie, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun. Structure analysis of soccer video with hidden markov models. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–4096. IEEE, 2002.

[90] Changsheng Xu, Yi-Fan Zhang, Guangyu Zhu, Yong Rui, Hanqing Lu, and Qingming Huang. Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia*, 10(7):1342–1355, 2008.

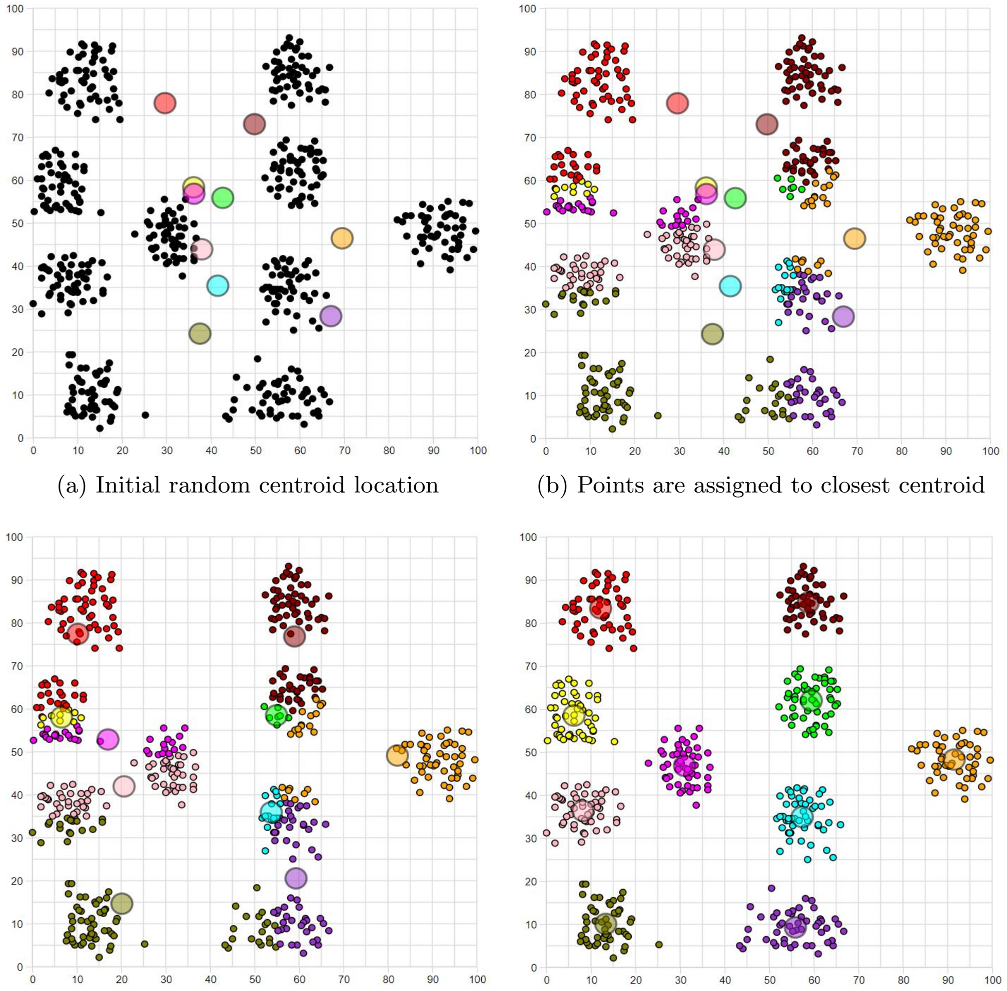
[91] Bin Zhang, Meichun Hsu, and Umeshwar Dayal. K-harmonic means-a data clustering algorithm. *Hewlett-Packard Labs Technical Report HPL-1999-124*, 55, 1999.

[92] Guangyu Zhu, Qingming Huang, Changsheng Xu, Yong Rui, Shuqiang Jiang, Wen Gao, and Hongxun Yao. Trajectory based event tactics analysis in broadcast sports video. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 58–67, 2007.

Appendix

A K-Means

The *K-means* clustering approach aims to separate a set of n-dimensional vectors into k clusters. The overall algorithm minimizes within-cluster distances (the points within a grouped cluster are as similar as possible) while maximizing the between cluster distance (clusters themselves are as different as possible). Initially, the main applications of the method lie within the field of signal processing (the name "k-means" first appeared in section 3 of MacQueen et al. [38], which borrows heavily from the work of Steinhaus [77]).



(a) Initial random centroid location

(b) Points are assigned to closest centroid

(c) Update of centroid locations and point assignment

(d) Convergence when no assignment changes

Figure 47: A short illustration of the k-means algorithm, based on the interactive visualization of the EduClust platform by the Data Analysis and Visualization department of the University of Konstanz. Here, a dataset of 500 2D points, resembling a 4-1-4-1 soccer formation, with 10 clearly distinct clusters (for the players) are grouped using 10 as the value for the k-parameter. After initial assignment a fast convergence towards the final cluster assignment can be exemplified from step (a) to (d).

The approach assigns k random points (optimization with more efficient start positions can be found in [13, 31, 39, 49, 85]) to the vector space, so-called *centroids*. These centroids are then assigned all the points, which are closest to them. The next step updates the centroids' position to the mean of the assigned points. Iteratively, the points the algorithm assigns the points to their centroids given their new position before the centroids move again. The clustering reaches convergence when no point locations change anymore, and therefore the centroid position has reached a stable location within the vector space. This separation of the total space into Voronoi cells remains decently robust to outliers. Figure 47 provides a step-by-step visualization of the algorithm.

B Hungarian-/ Kuhn–Munkres-Algorithm

The Hungarian algorithm solves an assignment problem in polynomial time. Harold Kuhn first developed it in 1955 [29] with time complexity of O(n^4),[50] however, Edmonds and Karp, and independently Tomizawa, derived solutions in O(n^3) [21, 81]. The algorithm addresses transportation allocation, job assignment, and even the infamous *traveling salesman* problem [19].[51]

The following example demonstrates the workings of the algorithm. Assuming we are given a toy example of three actors moving to three cities. We are trying to minimize the total distance for all three friends to move to these cities, therefore try to minimize the distance allocated by moving the three friends to exactly one city. The following table shows the distances:

	City One	**City Two**	**City Three**
Agent A	200 km	300 km	300 km
Agent B	300 km	200 km	300 km
Agent C	300 km	300 km	200 km

Table 4: The table provides the necessary input for the illustrative toy example for the application of the Hungarian Algorithm. Three agents are living within a certain distance to three cities, where they are trying to move. Every agent needs to end up at exactly one city and the Hungarian Algorithm finds the optimal assignment to minimize the total distance between all assignments.

This table is simply translated to a cost matrix C

$$C = \begin{pmatrix} 200 & 300 & 300 \\ 300 & 200 & 300 \\ 300 & 300 & 200 \end{pmatrix}.$$

[50] **Big O, or short "O(X)"** describes the limiting behavior of a function and is excessively used to describe the efficiency of algorithms. Here X is replaced with a function of n, where n is the data input to the algorithm. While efficiency will vary widely between tasks, constant time O(1), logarithmic time O(log(n)), and proportional time O(n) time are considered *fast* algorithms, while O(n^p) - with p > 1 - or even O(x^n) - with x > 1 - are considered inefficient algorithms. The classification might not matter as long as the input data size remains relatively small.

[51] **Traveling Salesman Problem** describes a popular problem in combinatorial optimization, theoretical computer science, and operations research. It asks the seemingly simple question: "Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city¿'. It is NP-hard and extends to its neighbors of *traveling purchaser problem* and the *vehicle routing city.*

A left and right permutation matrix[52] multiply with the original matrix. This calculation results in precisely one specific assignment of the agents to each city per possible 4×4 permutation matrix.

By minimizing the trace,[53] the algorithm defines the minimal entry per row (column) iteratively, subtracts its value from the other entries of the row (column) to find the optimal assignment. The approach translates to a minimization of the trace. Multiplication of the cost matrix with the left and right permutation achieves this calculation

$$\min_{L,R} \mathrm{Tr}(LCR).$$

The algorithm will then find the optimal solution to this toy example as the minimum total distance to be 600 km, with agent A moving to City One, agent B moving to City Two, and agent C moving to City Three.

For a more detailed step-by-step example, please refer to this illustration. Alternative solutions exist, which are somewhat more involved and utilize graph-theoretical approaches for the common problem of maximum-weight matching in bipartite graphs.

C Bayesian Model Selection

Frequently researchers are interested in formalizing the correct model to mimic real-world behavior. However, what is a model? A model is a parametric family of probability distributions, each of which can explain the observed data. Classical statistics uses hypothesis testing with pre-defined significant values and confidence intervals to measure a hypothesis's truthfulness by how robustly it withstands contrary evidence. Bayesian methods differ most notably in their inclusion of prior knowledge to the derivation of these decisions. The Bayes formula describes a *posterior* probability of a model M, given a data D, as the product of the data distribution being plausible given the model assumption times the probability that of the model occurring. Classically, the following equation describes this relationship

$$\Pr(M \mid D) = \frac{\Pr(D \mid M) \Pr(M)}{\Pr(D)}.$$

For our purposes, this canonical description will need to extend to a finite number of models M_i and therefore the equation becomes:

$$\Pr(M_i \mid D) = \frac{p(D \mid M_i) \Pr(M_i)}{\sum_j p(D \mid M_j) \Pr(M_j)}$$

By setting $\Pr(M_i)$ to a uniform distribution, we can write the probability $p(D \mid M_j)$ as

$$p(D \mid M_i) = \int p(D \mid \theta_i, M_i)\, p(\theta_i \mid M_i)\, \mathrm{d}\theta_i.$$

[52] **Permutation Matrix** is a square binary matrix with exactly one entry of 1 and just 0's elsewhere per row and column. Intuitively it is used in combination with other matrices to turn features on or off by allowing feature (row and column features) to persist past the multiplication.

[53] **Trace** of a square matrix represents the sum of its diagonal elements.

Two models can then be compared by combining them to a *Bayes Factor*

$$K = \frac{\Pr(D \mid M_1)}{\Pr(D \mid M_2)} = \frac{\int \Pr(\theta_1 \mid M_1) \Pr(D \mid \theta_1, M_1)\, d\theta_1}{\int \Pr(\theta_2 \mid M_2) \Pr(D \mid \theta_2, M_2)\, d\theta_2} = \frac{\frac{\Pr(M_1|D)\Pr(D)}{\Pr(M_1)}}{\frac{\Pr(M_2|D)\Pr(D)}{\Pr(M_2)}} = \frac{\Pr(M_1 \mid D)}{\Pr(M_2 \mid D)} \frac{\Pr(M_2)}{\Pr(M_1)}.$$

The following table from Kass and Raftery [28] provides an intuition of the degree of evidence necessary to choose one model over the other.

$\boldsymbol{log_{10}}$**K**	***K***	**Strength of evidence**
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

Table 5: This table by Kass and Raftery [28] provides an overview of the degree of evidence necessary to choose one model over another using Bayesian Model Selection methods.

As a quick aside, this characteristic of Bayesian model selection introduces the notion of preference for simpler models over more complex ones. Naturally, more complex models explain more datasets; therefore, their support in the *sample* space is more substantial. However, these models also generalize worse to the *population* space of explaining models in general. This ambiguity translates to an inherent cost to move towards more complex models by failing the simpler models' broad generality - an inbuilt Occam's razor.[54]

D Wasserstein Distance

The Wasserstein distance,[55] or Kantovich-Rubinstein metric defines a distance metric between two probability distributions on a given metric space **M**. It is first introduced in 1974 [83] and is often also referred to by *earth-movers-distance.*

It describes a formulation of the common *optimal transportation problem.*[56] In its canonical description, a re-formulation first translates the distributions to a set of clustered points, where both distributions do not need to translate to the same number of clusters and a set of cluster weights for each cluster based on the number of points of the original distribution represented by the cluster. This representation is called a *signature* and will facilitate the theory explanation of the earth-movers-distance.

[54] **Occam's razor** describes a problem-solving principle, which, given comparable results, will always prefer more straightforward solutions to more complex ones. "The simplest solution is usually the right one", often paraphrases this relationship.

[55] This section borrows heavily from the excellent articles [2, 4] and is expanded for better intuition by personal remarks and examples.

[56] The **Optimal Transportation Problem** assumes that several suppliers, each with a given amount of goods, are required to supply several consumers, each with a given limited capacity. For each supplier-consumer pair, the cost of transporting a single unit of goods is given. The transportation problem is then to find the least-expensive flow of goods from the suppliers to the consumers that satisfies the consumers' demand. Similarly, the problem is transforming one distribution P to another distribution Q with minimum work done.

So, we have P's signature as the set of cluster and weight pairs

$$P = \left\{(p_1, w_{p1}), (p_2, w_{p2}), \ldots, (p_m, w_{pm})\right\},$$

where p_i is the cluster representation and $w_{pi} > 0$ corresponds to its weight. Similarly, the distribution Q will correspond to the signature:

$$Q = \left\{(q_1, w_{q1}), (q_2, w_{q2}), \ldots, (q_n, w_{qn})\right\}.$$

The ground distance between clusters p_i and q_i, which we will try to minimize can be represented as

$$D = \left[d_{i,j}\right].$$

The algorithm's task is now to find the *flow* (the weight shifting from one signature to the other) that minimizes the cost. Let us define the flow as

$$F = \left[f_{i,j}\right],$$

with $f_{i,j}$ being the flow between p_i and q_i. To find the optimal solution, we need to define the overall cost to minimize as

$$\min \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j},$$

subject to the constraints

$$\begin{aligned}
& l f_{i,j} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n, \\
& \sum_{j=1}^{n} f_{i,j} \leq w_{pi}, 1 \leq i \leq m, \\
& \sum_{i=1}^{m} f_{i,j} \leq w_{qj}, 1 \leq j \leq n, \\
& \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} = \min \left\{ \sum_{i=1}^{m} w_{pi}, \quad \sum_{j=1}^{n} w_{qj} \right\}.
\end{aligned}$$

Intuitively, the solution to this linear optimization problem results in the *work to be done* (to stick with the *moving earth* analogy) normalized by the total flow

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}}.$$

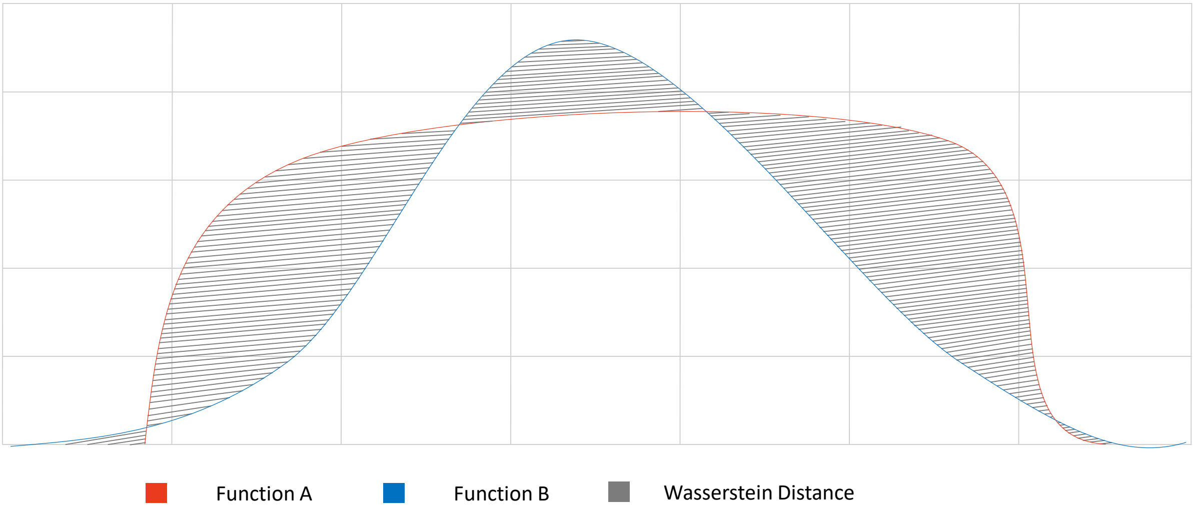


Figure 48: This illustration exemplifies the Wasserstein distance-equation comparing two arbitrary non-Gaussian distributions (red and blue). Their distance, or the effort of moving from one distribution to the other is highlighted in dark gray.

Figure 48 illustrates how to understand the Wasserstein distance in 2D for non-Gaussian distributions. For the case of two normal distributions with means m_1 and m_2 and symmetric positive semi-definite[57] covariance matrices C_1 and C_2, Olkin and Pukelsheim [48] derive the 2-Wasserstein distance between μ_1 and μ_2, which equals the definition of the Wasserstein distance for our case outlined in section 4.2.

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

[57] **Positive semi-definite** matrices intuitively, preserve the direction of any input applied to them. Marginally more formally, this means that for any non-zero column input vector z of n real numbers, the result of $z^T M z$ is positive or zero (i.e., non-negative). This property is important for the correct calculation of fractional exponents—here $\frac{1}{2}$—of the co-variance matrix.