

# Module 4. Application of data science in sport

In this module, we will examine in detail, with examples, the various applications of data science in sports.

## Unit 4.1 Talent scout and player recruitment

### 4.1.1 Overview

Traditional scouting methods relied heavily on visual assessments and subjective reports from scouts, which sometimes led to misjudgments. Talent scouting and player recruitment have evolved drastically in recent years, driven by the rise of data science. Teams across all leagues and levels are now using advanced metrics, algorithms, and technology platforms to evaluate potential signings that may otherwise go unnoticed. This approach is efficient when trying to uncover hidden gems, players who may not be on the radar of larger clubs or possess attributes that traditional scouting might overlook. Player recruitment and development have been transformed, with data-driven scouting and youth program analysis becoming standard practice in many organizations.

Data-driven scouting also offers the advantage of predicting future success based on player metrics, injury history, and physical attributes. By understanding how a player performs over time and in different situations, clubs can forecast the likelihood of a player thriving in their system. Clubs use advanced player metrics beyond goals and assist in making informed decisions. Metrics like key passes, tackles won, and interceptions provide a more nuanced picture of a player's contribution to the team.

For example, a defensive midfielder's value is not purely measured by goals, but by tackles won, ball recoveries, and successful interceptions. These stats help quantify how much the player disrupts the opposition's attacks and contributes to maintaining the team's defensive solidity.

Scouting platforms, such as Wyscout and Statsbomb, allow clubs to track and analyze player metrics across leagues worldwide. They provide detailed breakdowns of a player's match performance, making it easier for recruitment departments to assess talent holistically. Integrating video footage with statistical data, further allows scouts and analysts to study the context behind each action, ensuring more informed decisions.



Some examples of metrics scouts can share and compare easily now are the following.

**a) Expected goals (xG) and expected assists (xA).**

xG measures the quality of chances a player gets, while xA quantifies the quality of chances they create for teammates. For example, a striker with a high xG, but low actual goals, might be underperforming, while one with a lower xG, but higher actual goals, could be an exceptional finisher.

**b) Passes per defensive action (PPDA).**

This metric is crucial for evaluating pressing intensity, especially for teams that employ a high-press strategy. A lower PPDA indicates a more aggressive press. Scouts might look for midfielders or forwards with low PPDA values to fit into a high-intensity system.

**c) Progressive passes and carries.**

It measures a player's ability to advance the ball towards the opponent's goal through passing or dribbling. They're treasured for identifying playmakers and creative midfielders who can break defensive lines.

**d) Defensive actions per 90 minutes.**

This includes tackles, interceptions, blocks, and clearances, normalized to a per-90-minute basis for fair comparison. It's essential for evaluating defensive players and midfielders with defensive responsibilities.

**e) Key passes and shot-creating actions.**

These metrics help identify players who consistently create scoring opportunities, even if they don't always result in assists. They're beneficial for evaluating attacking midfielders and wingers.

## 4.1.2 Examples of data-driven scouting

### Brentford FC

Brentford FC has been among European football's leading proponents of data-driven recruitment. Their strategy, often compared to the "Moneyball" approach in baseball, has enabled them to sign players undervalued by bigger clubs. Brentford's recruitment philosophy focuses on identifying players with strong underlying metrics, who may not necessarily have the highest market value.

For example, Brentford's data-driven approach led them to sign players such as Ollie Watkins, Saïd Benrahma, and Neal Maupay, who have become high-value assets. Top-tier clubs initially overlooked these players, but thrived under Brentford's analytical and



developmental framework. Brentford's transfer strategy emphasizes maximizing player value by targeting younger players with potential for future growth rather than spending large sums on established stars.

### **Leicester City (Premier League 2015/16)**

Leicester City's stunning Premier League title win in 2015/16 was a breakthrough moment for data-driven scouting. Advanced metrics and data analysis heavily influenced the recruitment of key players like N'Golo Kanté, Riyad Mahrez, and Jamie Vardy. Each player was recruited relatively cheaply, yet their contributions were instrumental in Leicester's historic success.

Leicester's data-led recruitment strategy allowed them to assemble a team that was more significant than the sum of its parts. The scouting team identified players overlooked by larger clubs, but possessed the physical and mental attributes needed to succeed in the Premier League. Kanté's exceptional stamina and defensive awareness, Mahrez's creative flair, and Vardy's speed and finishing ability exemplified how Leicester's data-driven approach paid off handsomely.

### **4.1.3 Conclusion**

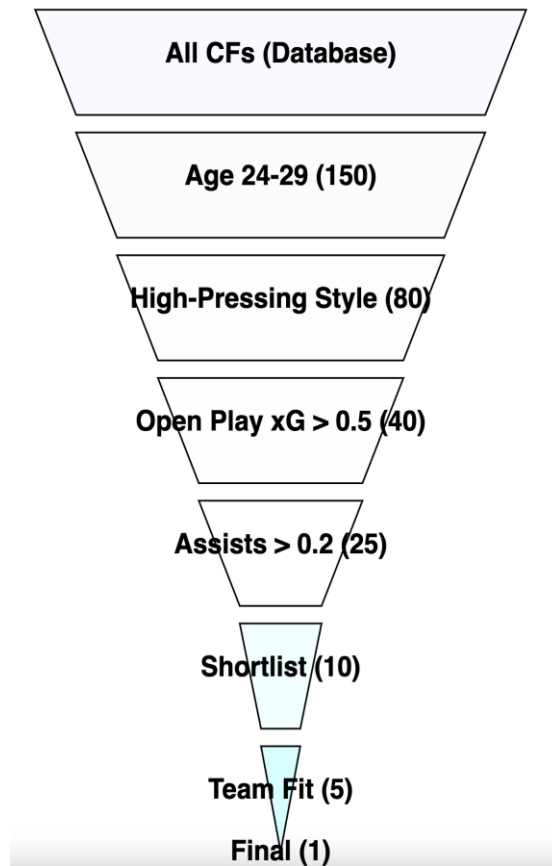
Integrating data science into talent scouting and player recruitment has transformed how clubs operate by focusing on critical metrics and objective decisions that lead to long-term success. Clubs like Brentford FC and Leicester City are prime examples of how data-driven approaches can identify undervalued talent and maximize player potential, revolutionizing the world of football recruitment.

Integrating these advanced metrics into the scouting process has transformed how clubs identify and evaluate talent. By combining these data-driven insights with traditional scouting methods, clubs can make more informed decisions, reduce the risk of costly transfer mistakes, and uncover hidden gems that might have been overlooked in the past.

However, while data has become an indispensable tool in talent scouting, it should not be the only factor considered. The art of talent scouting still requires human insight to interpret data in the context of a club's specific needs, playing style, and long-term strategy. The most successful clubs balance leveraging advanced analytics and trusting the experienced eyes of seasoned scouts.

### **Figure 1. The scouting process to find hidden talent**





Source: own elaboration.

\*The image describes the scouting process and the steps the team and scouting directors take to find hidden talent, showing an example of a center forward in football.

## Unit 4.2 Data science in team analysis

### 4.2.1 Introduction

In recent years, the world of sports has undergone a profound transformation driven by integrating data science into team analysis. This revolution has reshaped how teams approach performance evaluation, tactical decision-making, and long-term strategy development. As we delve into the fascinating intersection of data science and sports, we'll explore how teams leverage advanced analytics to gain a competitive edge in an increasingly data-driven sporting landscape.

The journey of data analytics in team sports has its roots in baseball, popularized by the "Moneyball" era of the early 2000s. Billy Beane's Oakland Athletics famously used statistical analysis to identify undervalued players and compete against teams with much



larger budgets. While initially met with skepticism, this approach sparked a revolution quickly spread to other sports. Basketball embraced advanced metrics with the rise of player efficiency ratings and spatial analysis of shot selection. Football began to utilize GPS tracking data to monitor player movements and optimize tactical formations. Soccer adopted expected goals (xG) models to evaluate scoring opportunities more accurately.

As the field has evolved, the importance of data-driven decision-making for teams has become increasingly apparent. In the high-stakes world of professional sports, where marginal gains can mean the difference between victory and defeat, teams are turning to data science to:

- objectively evaluate player and team performance beyond traditional statistics,
- identify hidden strengths and weaknesses in both their team and their opponents,
- inform tactical decisions and in-game adjustments.

The applications of data science in team sports are wide-ranging and continually expanding. Performance analysis now goes far beyond basic box scores, incorporating advanced metrics that capture the nuances of player contributions. Real-time data processing and predictive modeling enhance tactical decision-making. Team structure optimization uses network analysis and role identification algorithms to maximize player impact.

The benefits of using data science for team analysis are numerous and significant. Teams can now:

- conduct objective evaluations of performance that account for context and role-specific contributions;
- identify strengths and weaknesses that may not be apparent through traditional observation alone;
- develop more sophisticated and tailored strategic plans based on quantifiable insights;
- utilize players more effectively by understanding their optimal roles and conditions for success;
- make more informed decisions in high-pressure situations using real-time data analysis.

As we explore the world of data science in sports team analysis, we'll examine the specific methodologies, technologies, and applications driving this revolution. From cutting-edge data collection techniques to advanced statistical models and compelling visualization



methods, we'll uncover how teams harness data's power to push the boundaries of athletic performance and strategic planning.

Integrating data science into sports has challenges and ethical considerations, which we'll also address. As we look to the future, it's clear that the role of data in sports will only continue to grow, promising even more sophisticated analysis and insights that will shape the games we love for years.

## 4.2.2 Data collection methods in sports

The foundation of any data science application in sports is the collection of high-quality, relevant data. As technology has advanced, so have the methods for gathering information about player and team performance. This section will explore the various data collection techniques in modern sports analytics.

### Computer vision and tracking systems

While wearable technology provides individual player data, computer vision systems offer a comprehensive view of all players and the ball throughout a game.

Optical tracking systems: installed in stadiums, these systems use multiple cameras to track the position of all players and the ball at high frame rates. They provide data on:

- player and ball positions,
- team shape and tactical formations,
- pass trajectories and ball possession.

Player and ball tracking technologies: advanced image processing algorithms can identify and track players and the ball, even in crowded scenes. This technology enables:

- automated event detection (passes, shots, tackles),
- analysis of player interactions and team structures,
- calculation of advanced metrics like space creation and defensive coverage.

The advantage of these systems is that they can capture data without requiring players to wear additional equipment. However, they can be expensive to install and maintain, and their accuracy can be affected by weather conditions in outdoor sports.

### Match event data

While automated systems provide a wealth of positional and physical data, many important events in a game still require human observation to be accurately recorded.



Manual data collection: trained analysts watch games live or on video to record specific events such as:

- passes (including type and outcome);
- shots and goals;
- tackles, interceptions, and clearances;
- fouls and cards;
- substitutions and tactical changes.

Standardized event coding systems: standardized coding systems have been developed to ensure consistency across different data providers and enable meaningful analysis. These systems define precisely what constitutes each type of event and how it should be recorded.

Integration of multiple data sources: The true power of sports data analysis often comes from combining these various data sources. For example, integrating manually entered event data with tracking data can provide context to events, addressing not only how a pass was delivered but also the positions of all players at the time, the speed of the ball, and the physical state of the passer.

### **Data quality and validation**

As with any data science application, the quality of insights is only as good as the quality of the input data. Sports organizations invest significant resources in ensuring the accuracy and reliability of their data.

- Ensuring accuracy and consistency: this involves rigorous training for manual data collectors, regular calibration of automated systems, and cross-validation between different data sources.
- Dealing with missing or erroneous data: sophisticated algorithms detect and correct errors and impute missing data when necessary.
- Data cleaning and preprocessing: raw data must often be cleaned and preprocessed before it can be analyzed. This might involve smoothing noisy tracking data, standardizing event classifications, or aggregating data from multiple sources.

As we explore data science in sports, it's important to remember that these data collection methods form the bedrock upon which all subsequent analysis is built. The ongoing advancements in this field continually expand the possibilities for understanding and optimizing team performance.



### 4.2.3 Key performance indicators (KPIs) in team sports

With the wealth of data available in sports, it is crucial to identify and focus on the most relevant metrics. Key performance indicators (KPIs) are quantifiable measures used to evaluate the success of an organization, team, or individual in meeting performance objectives. In team sports, KPIs help coaches, analysts, and players focus on the most critical aspects of performance. Let's explore some key KPIs categories used in team sports analysis.

#### a) Scoring efficiency metrics.

Ultimately, the goal of most team sports is to outscore the opponent. Therefore, metrics that measure scoring efficiency are among the most critical KPIs.

#### b) Points per possession (basketball).

This metric measures how efficiently a team or player converts possessions into points. It's calculated by dividing the total points scored by the number of possessions. This KPI is valuable because it accounts for the pace of play, allowing for fair comparisons between fast and slow-paced teams.

#### c) Expected Goals (xG) (soccer).

xG models assign a probability to each shot based on location, type of play, and defensive pressure. This provides a more nuanced view of offensive performance than simple shot counts or on-target percentages. Teams can use xG to evaluate whether their goal-scoring (or conceding) is sustainable or likely to regress to the mean.

### Figure 2. Total shots taken vs. team goals





Defensive rating (basketball): this metric estimates how many points a team allows per 100 possessions. Like its offensive counterpart, it will enable pace-adjusted comparisons between teams.

Tackles, interceptions and clearances (soccer): these basic counting stats provide insight into a team's defensive activity. However, they must be interpreted cautiously, many tackles might indicate an active defense or suggest a team that struggles to maintain possession.

### **e) Possession and ball movement statistics.**

In many team sports, controlling and moving the ball effectively is critical to success.

Possession percentage: this essential but crucial metric measures the proportion of time a team controls the ball. However, it's important to note that high possession doesn't always correlate with success – some teams deliberately cede possession and play on the counter-attack.

Pass completion rates measure the percentage of attempted passes that successfully reach a teammate. For more nuanced analysis, they can be broken down by pass type (short, long, forward, backward).

Progressive passes: this advanced metric goes beyond simple completion rates to measure passes that significantly advance the ball toward the opponent's goal.

### **f) Player movement and positioning.**

With the advent of advanced tracking technologies, teams can now quantify and analyze player movement in unprecedented detail.

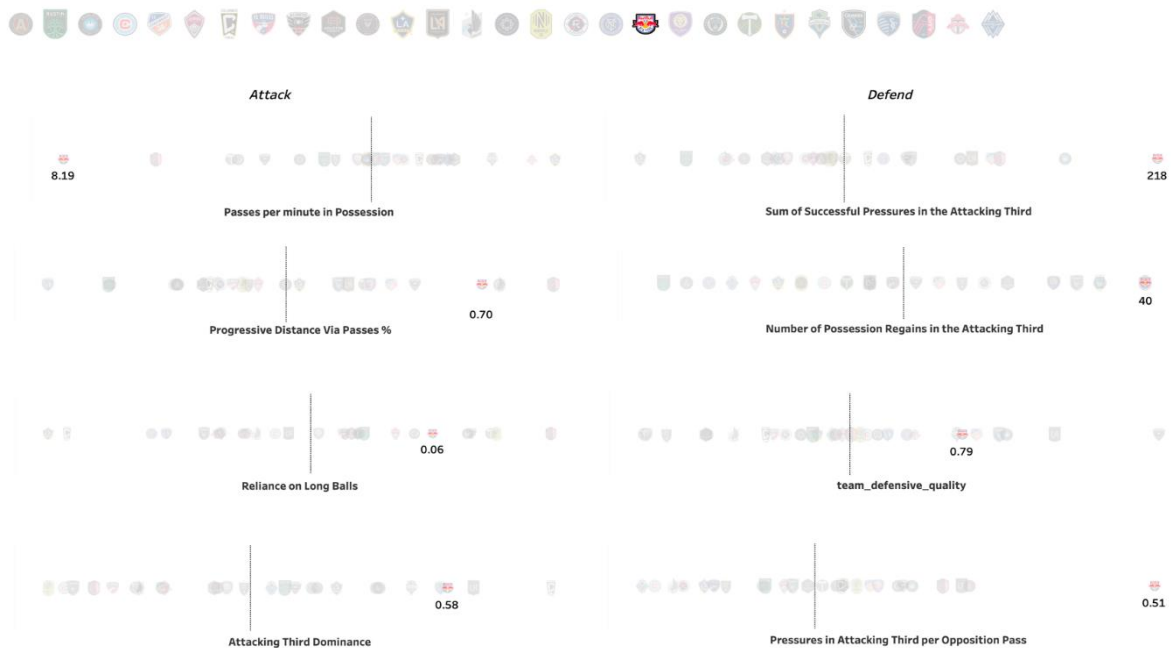
Distance covered and sprint metrics: these basic measures quantify players' physical output. They can be used to monitor workload, inform substitution decisions, and identify players at risk of fatigue-related injuries.

Heat maps: these visualizations show where players spend their time on the field or court. They can reveal tactical patterns and help evaluate whether players fulfill their positional responsibilities.

Off-ball movement and space creation: advanced metrics can now quantify how players move without the ball, including how effectively they create space for teammates or close down space for opponents.



**Figure 3. Playing pattern of one of the MLS (USA) teams**



Source: own elaboration.

\*The image above describes the playing pattern of one of the MLS (USA) teams. It shows how some teams have unique playing patterns and play their style. The attack and defending ranking on the image can show other teams and how Red Bull has performed in MLS in the 2023 season.

## Unit 4.3 Physical performance, injury management

### 4.3.1 Overview

In the highly competitive world of professional sports, the difference between victory and defeat often comes down to the finest of margins. As such, teams are increasingly turning to data science to optimize their athletes' physical performance and manage injury risks. This intersection of sports science and data analytics has opened new frontiers in athlete care and performance enhancement. Physical performance in sports encompasses many factors, including strength, speed, endurance, agility, and sport-specific skills. On the other hand, injury management involves preventing injuries, diagnosing them accurately when they occur, and optimizing the rehabilitation process.

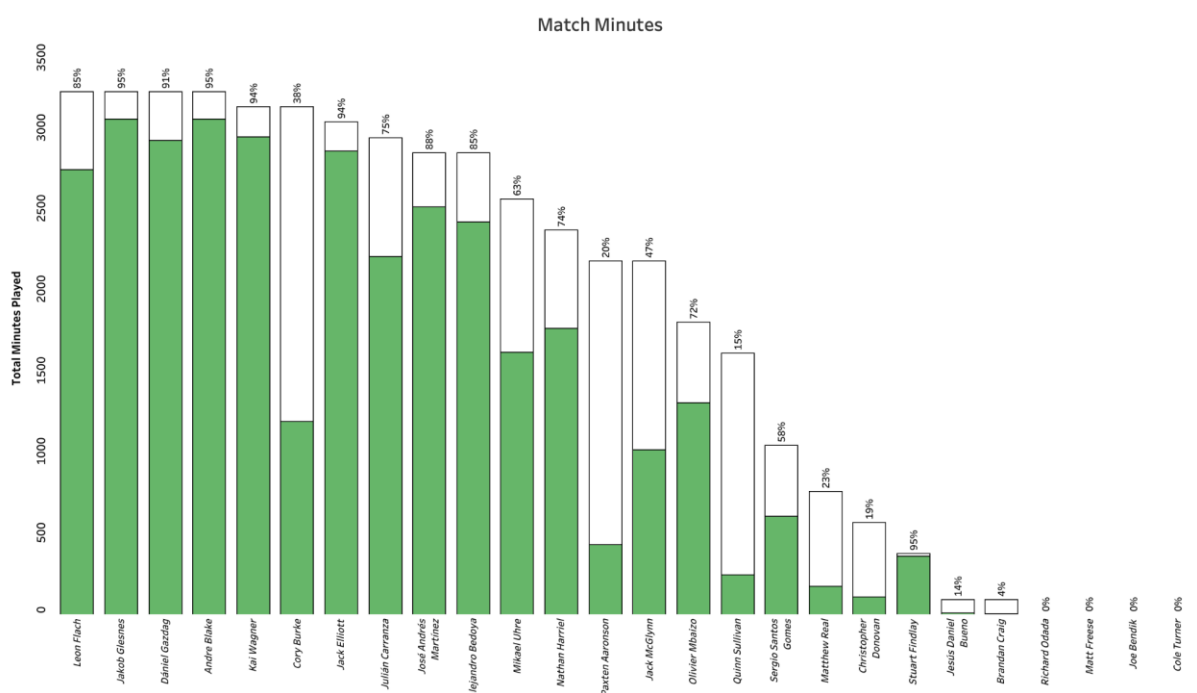


The evolution of technology has dramatically expanded the types and volume of data available for analysis. Wearable devices, force plates, high-speed cameras, and other advanced tools now provide a wealth of information about an athlete's movement patterns, physiological responses, and biomechanics. When properly analyzed, this data can reveal subtle patterns and trends that might be invisible to the naked eye.

Key areas where data science is making a significant impact include the following.

- Performance monitoring and optimization: using data to track and improve various aspects of an athlete's physical capabilities.
- Workload management: balancing training stress and recovery to maximize performance while minimizing injury risk.
- Injury prevention: identifying risk factors and implementing strategies to reduce the likelihood of injuries.
- Injury diagnosis and rehabilitation: utilizing data to improve the accuracy of injury diagnoses and to optimize the rehabilitation process.
- Return-to-play decisions: using objective data to inform decisions about when an athlete will return to the entire competition following an injury.

Figure 4. Match minutes



Source: own elaboration.

\*The image describes the minutes each player is available for and how many minutes they have accumulated playing in the game; this helps the coaches rotate the squad and allows the medical and performance staff to measure performance and help in recovery.

### 4.3.2 Performance monitoring and optimization

Performance monitoring and optimization form the cornerstone of modern sports science. By leveraging data, teams can comprehensively understand each athlete's capabilities, track their progress, and tailor training programs to maximize performance.

#### Data collection for performance monitoring

1. GPS and inertial measurement units (IMUs): these wearable devices track an athlete's position, speed, acceleration, and direction of movement. They provide valuable data on:
  - distance covered during training or competition,
  - number and intensity of sprints,
  - changes in direction and acceleration/deceleration patterns.
2. Heart rate monitors: these devices offer insights into an athlete's cardiovascular response to exercise, including:
  - resting heart rate,
  - maximum heart rate,
  - heart rate variability (HRV).
3. Force plates: these instruments measure ground reaction forces during movements like jumping or changing direction, providing data on:
  - power output,
  - force production,
  - balance and stability.
4. Optical tracking systems: high-speed cameras and computer vision algorithms can track player movements without the need for wearable devices, offering data on:
  - player positioning and movement patterns,
  - team formations and tactical arrangements,
  - ball trajectory and possession.
5. Biomarkers: regular blood tests and other physiological measurements can provide data on:



- hormonal levels (e.g., cortisol for stress, testosterone for recovery),
- markers of muscle damage and inflammation,
- nutritional status.

### **Data analysis for performance optimization**

1. Series analysis: this technique tracks performance trends over time, identifying patterns in metrics like speed, power output, or heart rate variability.
2. Machine learning algorithms: these can be used to identify complex patterns in performance data. For example, a random forest algorithm might determine which factors most predict peak performance.
3. Statistical modeling: techniques like multiple regression can help quantify the relationships between various performance metrics and outcomes.
4. Dimensionality reduction: methods like principal component analysis (PCA) can help simplify complex datasets with many variables, making it easier to identify key performance factors while using force plate data.

### **4.3.3 Practical applications of performance data**

The insights derived from this data analysis can be applied in numerous ways.

1. Individualized training programs: by understanding each athlete's unique physiological profile and response to training, coaches can design tailored programs that optimize performance gains while minimizing injury risk.
2. Tactical decision-making: performance data can inform in-game decisions, such as when to substitute players based on their current physical state.
3. Nutritional strategies: by correlating performance data with nutritional intake, teams can optimize dietary strategies for each athlete.

### **4.3.4 Understanding workload**



Workload management is a critical aspect of both performance optimization and injury prevention. By carefully balancing training stress and recovery, teams can help their athletes reach peak performance while minimizing the risk of overuse injuries. In sports science, workload is typically divided into two categories.

1. External workload. This refers to the physical work done by an athlete, such as:
  - distance covered,
  - number of sprints or jumps,
  - weight lifted in strength training,
  - number of pitches thrown (in baseball).
2. Internal workload. This represents the physiological stress experienced by the athlete in response to the external workload. It can be measured through:
  - heart rate and heart rate variability,
  - rate of perceived exertion (RPE),
  - blood lactate levels,
  - psychological stress measures.

#### **4.3.5 Data-driven workload management**

Modern workload management strategies leverage data science in several ways.

1. Acute to chronic workload ratio (ACWR): this key metric compares an athlete's recent workload (typically over the past week) to their average workload over a more extended period (often 3-4 weeks). Research has shown that large spikes in acute workload relative to chronic workload are associated with increased injury risk. Data science techniques are used to:
  - calculate and track ACWR for various performance metrics,
  - determine optimal ACWR ranges for different athletes and sports,
  - predict injury risk based on ACWR and other factors.
2. Individualized thresholds: machine learning algorithms can be used to determine personalized workload thresholds for each athlete. These models might consider factors such as:
  - age and training history,
  - recent injury history,
  - current fitness level,
  - position and playing style.
3. Fatigue modeling: advanced statistical models can estimate an athlete's fatigue level based on recent workload and recovery data. These models might incorporate the following:



- training and competition workload,
- sleep quality and quantity,
- travel and time zone changes,
- psychological stress factors.

### 4.3.6 Injury Prevention Strategies

While workload management is a crucial aspect of injury prevention, data science is also applied to other preventive strategies.

1. Strength imbalance detection: force plate data and other biomechanical measures can be used to detect subtle strength imbalances between limbs or muscle groups. When identified early, these imbalances can be addressed through targeted strength training to reduce injury risk.
2. Recovery optimization: data on sleep patterns, heart rate variability, and subjective wellness measures can be used to optimize recovery strategies. Machine learning models can predict which recovery modalities (e.g., massage, hydrotherapy, compression garments) are likely most effective for each athlete based on their current state.
3. Nutrition and hydration monitoring: advanced analytics can be applied to nutritional data to ensure athletes are adequately fueled and hydrated. This might involve:
  - tracking macronutrient intake and correlating it with performance data,
  - using bioimpedance devices to monitor hydration status,
  - analyzing sweat composition to personalize electrolyte replacement strategies.

By integrating these various data streams and applying advanced analytics, teams can develop comprehensive injury prevention strategies personalized to each athlete's unique characteristics and risk factors.

### 4.3.7 Injury diagnosis and rehabilitation

Once an injury has been diagnosed, data science continues to play a role throughout the rehabilitation process.

1. Personalized rehabilitation programs. Machine learning algorithms can be used to design and adjust rehabilitation programs based on:



- the specific nature and severity of the injury,
  - the athlete's characteristics (age, injury history, etc.),
  - real-time progress data collected during rehabilitation exercises.
2. Progress tracking and prediction. By collecting and analyzing data throughout the rehabilitation process, teams can:
    - track progress against expected recovery timelines,
    - identify potential complications early,
    - predict return-to-play timelines more accurately.
  3. Comparison with historical data. Machine learning models can compare an athlete's rehabilitation progress with large datasets of similar injuries to:
    - identify if recovery is progressing as expected,
    - suggest adjustments to the rehabilitation program based on successful approaches in similar cases.

By leveraging these data-driven approaches, sports medicine professionals can make more informed decisions throughout rehabilitation, potentially speeding up recovery times and reducing re-injury risk.

### **4.3.8 Conclusion**

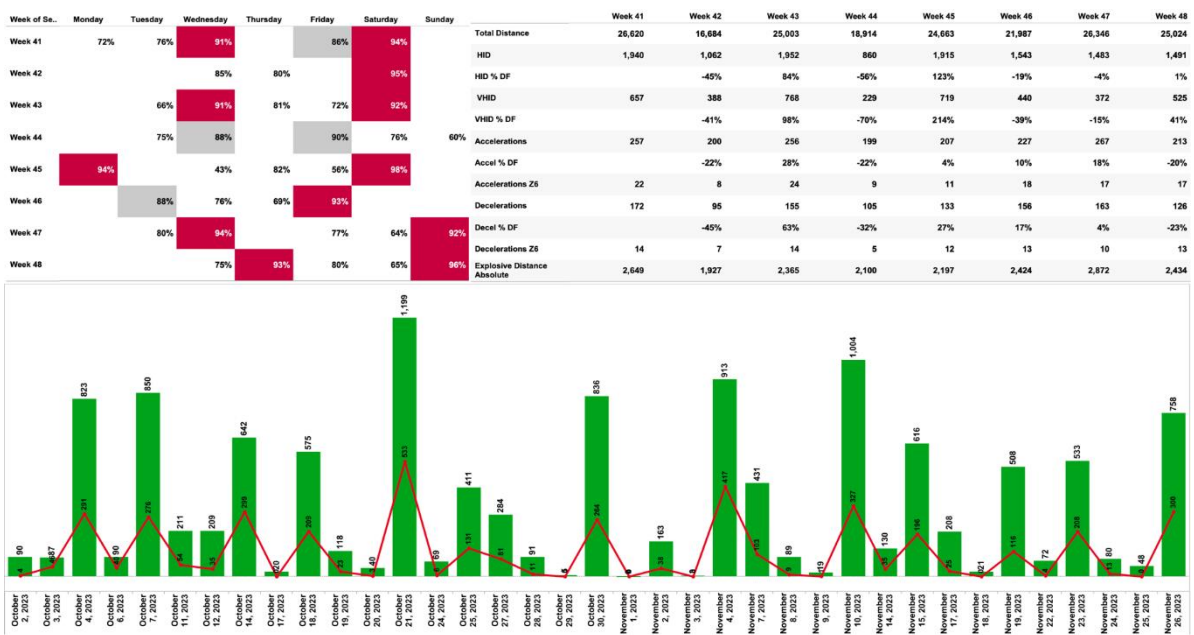
The application of data science to physical performance and injury management in sports represents a significant shift in how athletes are trained, monitored and cared for. By providing objective, quantifiable insights, data science enables more personalized, precise, and practical approaches to performance optimization and injury prevention.

However, it's crucial to remember that data and algorithms cannot replace the expertise of coaches, medical professionals, and athletes. Instead, data science should be seen as a powerful tool that, when properly integrated with traditional sports science and medicine, can help push the boundaries of human performance while prioritizing athlete health and wellbeing.

As we look to the future, the continued advancement of data science in sports promises to bring even more sophisticated and nuanced approaches to physical performance and injury management, potentially revolutionizing how we understand and enhance athletic capabilities.



Figure 5. Bird's-eye view for the medical director



Source: own elaboration.

\*This is the bird's-eye view for the medical director to understand all the different components of keeping a player healthy. As max speed is critical in soccer and performance, staff needs to understand the week-to-week performance of various KPIs.

The high-speed running and sprint distance graphs help you understand the spikes a player is used to. These can also be translated to ACWR, which allows the trainers to better understand and monitor their workloads per macrocycle.

## Bibliography

Archie. (2024, July 30). Analysis of Match Player Stats. *Cloudydirt*. <https://cloudydirt.com/match-player-stats/>

BTB Content Team. (2024, July 24). Unlocking the Potential of Custom AI Solutions for E-commerce Success. *Best Tool Bars*. <https://blog.besttoolbars.net/unlocking-the-potential-of-custom-ai-solutions-for-e-commerce-success/>

Cruz e Espadim. (n.d.) The Impact of Technology on Sports Betting. <https://cruzeespadim.com/the-impact-of-technology-on-sports-betting/>



Herold, M., Kempe, M., Bauer, P., & Meyer, T. (2021). Attacking Key Performance Indicators in Soccer: Current Practice and Perceptions from the Elite to Youth Academy Level. *Journal of Sports Science and Medicine*, 20(1), 158-169. DOI: <https://doi.org/10.22028/D291-33556>

Mwfls. (n.d.). How Is Artificial Intelligence Being Used to Enhance Athletic Performance? <https://mwfls.org/how-is-artificial-intelligence-being-used-to-enhance-athletic-performance.html>

Rad. (2023, December 23). Measuring Success: Key Performance Metrics in Soccer. *Soccerreto*. <https://soccerreto.com/measuring-success-key-performance-metrics-in-soccer/>

Wang, S. L. (2022). *Quantitative Motion Analysis of the Upper Limb: Establishment of Normative Kinematic Datasets and Systematic Comparison of Motion Analysis Systems*. Faculty of the Graduate School of the University of Maryland. <https://doi.org/10.13016/i9jq-e4qu>

