Bachelor thesis:

Process mining on FIFA controller data

By Yun Feng Zheng

Programme: BSc Industrial Engineering & Management

Faculty: Behavioural, Management and Social Sciences

University of Twente

Primary supervisor: dr. Guido Bruinsma

Second supervisor: ir. Rogier Harmelink

Preface

Dear reader,

Before you lies my thesis that concludes my bachelor Industrial Engineering and Management. During my time with the eSportslab Twente, I worked on developing a method to turn FIFA data into insights to improve data analysis use within the FIFA eSports scene and I am thankful that I was granted the opportunity to work on this.

Firstly, I would like to show appreciation to my primary supervisor Guido Bruinsma for his guidance during the thesis. Secondly, I would like to thank my second supervisor Rogier Harmelink for his feedback and advice. Lastly, I would like to thank Ipek Seyran-Topan for guiding me throughout the general process of graduating.

Yun Feng Zheng

Enschede, June 2022

Management summary

• Problem definition
The eSportslab wanted to find the way to play FIFA better and more professional. It wanted to know which in-game actions resulted in wins. After researching this problem, it was concluded that the in-game data was not available for the eSportslab. To solve this problem “In-game data is not available”, the eSportslab used the collected controller data as a starting point. The goal of the research is to create a proof of concept method to analyze the collected controller data and create insights into the data to find which in-game actions lead to success. The method should contain a data preparation phase, a data analysis phase and a data visualization phase.

• Results
The generalized process of going from controller data to process models is summarized below. Firstly, this process contains the data preparation phase represented in points 1, 2 and 3. Secondly, the generalized process contains the data analysis phase represented in points 4,5 and 6. The process was used to generate process models and a data summary, and these findings were then combined with the data visualization best practices to position them in the dashboard elements “Player overview”, “Stats analysis” and “In game analysis”.

1. Determine which key events should be available for analysis, this will impact the information that is going to be collected during the matches: Depending on the key events chosen, extra information needs to be gathered during the match.
2. Preparing the input: Collect the raw button data. During the match, collect the additional information required to form the key variables.
3. Convert to events: The converter tool from the eSportslab takes in datasets equal to Table 10. With the collected raw button data and the possession values, the necessary table can be constructed. This table is then converted into an event log similar to Table 11. After the data has been prepared, add the key variables into the data.
4. Filter based on key events: Depending on the key variables chosen in the analysis, the data set can be split into multiple data sets to compare against each other.
5. Process the data in ProM: Import the datasets, convert them with the HeuristicsMiner and default settings. The data summary and the process models are available.
6. Performing the analysis: Interpret the process models and the data summaries. Compare the process models and data summaries between chosen key variables.

• Conclusions
The research was started because the eSportslab found that the professionalization of the FIFA eSports scene was hampered and wanted to research which in-game actions resulted in success. After researching this problem, it was concluded that the in-game data was not available for the eSportslab. To solve the problem “In-game data is not available”, the collected controller data is used as a starting point and the goal is to create a proof of concept method to analyze the collected controller data. The method should contain a data preparation phase, a data analysis phase and a data visualization phase.

To produce the method, the best practices in data preparation and data analysis were collected. Then these best practices were placed in a FIFA context and used to prepare and analyze the FIFA controller data. With the use of the chosen process discovery algorithm, the HeuristicsMiner, process models and data summaries were produced. Afterwards, the data visualization best practices were used to place the produced process models and data summaries within the existing eSportslab dashboard. By finishing this process, the assumption is made that the proof of concept method to analyze the collected controller data works.

• Discussion of the final result and further research
One shortcoming, which may be solved by further research, is that the generalized process of going from raw data to in-game events and models contains steps that are not automated and require manual labor to complete them. Another shortcoming is that the used data set is of a small size, which leads to process models and data summaries that cannot give conclusive insights. To solve this a larger data set could be used. At last, the research used the default ProM settings to produce the models, this could be a point for future research.

• Recommendations
After taking the conclusion, discussion and shortcomings into consideration, the following recommendations are made to the eSportslab.

- Automate (parts of) the proof of concept method, because the created method contains several steps which require manual labor to convert the list of controller data to a list of in-game data. This is part of improving the maturity of the event log which is mentioned in the process mining manifesto by Van der Aalst et al. (2016) as a guiding principle for process mining.
- Use a larger data set to find conclusive insights, because the current data size is the size of one test match, the research could not deliver conclusive results. To deliver conclusive results, a larger data set is needed, the question is how large and that is a question for further research to solve.
- Research the dependency value used to create the process models and look further into finding the ideal dependency value for the FIFA event data. Moreover, ProM gives also other threshold and heuristics settings
- Research the possibility of creating and exporting high resolution process models, the current environment where the process models are created is ProM. When the more complex models are shown in the current environment, the models are readable, but when the models are exported, the models become less readable. Further improvements can be made to the readability of the models through new environments which do support readable export of the process models.
- Create an expected goal model, because during the research it was noticed that knowing the sequences of actions is useful, the data could be used in an expected goal model to create one of the state-of-the-art analysis models within real-life football which is the expected goal model. The recommendation for further research is to connect the event data with the pitch position of the event data and to collect event data that contains the pitch position. This data can be used to create a FIFA expected goal model.

Table of contents

Chapter 1: Research context and problem introduction

This first chapter serves as the introduction to the research by giving an overview of the research context, defining what problem the research tries to solve, defining how the research is going to solve the problem or in other words, the methodology that will be used to solve the problem.

In this chapter, the context of the research will be given and the assignment will be described (1.1), the research problems will be stated (1.2), the research methodology will be described (1.3), and the research approach will be stated (1.4), the research approach will be linked with the research methodology. The structure of the knowledge questions asked for the theoretical framework (Chapter 2) will be explained (1.5).

1.1 Context and assignment description

In recent years, there have been developments in the field of competitive gaming or eSports and gaming in general. In 2021, there will be an estimated 3.0 billion gamers worldwide and it is expected that the population of gamers will grow to 3.32 billion by 2024 (Newzoo | Global Games Market Report, 2021). Estimated revenues from gaming are $175.8 billion in 2021 (Newzoo | Global Games Market Report, 2021). By having such a large target market and combining that with the amounts of revenues that are generated, there is more money available to develop eSports. An example of this are eSporters that play the game “League of Legends”, who have been able to earn themselves on average a salary of $400,000 in North America (Mellor, 2020).

With the inflow of money, there is more professionalization within eSports. There is more prize money available, salaries are growing and eSports has been professionalizing as a whole. What is exactly meant by the professionalization of a sport? Sheehan (2000) defines a program to be professionalized if the program operates with a fundamental focus on revenues or profits. An example of this is Team Liquid, a professional eSports organization, where they have 60 professional eSporters divided over 14 of the top eSports games (Team Liquid, n.d.). With their eSporters, Team Liquid has earned over $37 million from prize money (Team Liquid - Esports Team Summary, n.d.).

To further professionalize eSports, a comparison can be made with general sports such as football or basketball. In general sports, different aspects could be professionalized: the athlete, the materials, and the context for example. An example of professionalization is that the athlete could have a training schedule, mental coach, recovery plan, and dietary plan. More recently data analysts are becoming more prevalent within general sports as a means of further professionalization of sports. The previously mentioned aspects could help with improving the performance of the sporter.

Another example of how the material that is used could be professionalized is the suit of a speed skater. The performance of these suits can be analyzed in wind tunnels (Sætran and Oggiano, 2008). Another example is the shoes that are used in marathon running. Eliud

Kipchoge became the first human to run a marathon in under 2 hours in the Alphafly, which are Nike shoes designed to reduce the energy lost while running through the use of a carbon-fiber plate and airbags in the shoes (Burgess, 2020).

While sports, in general, have been professionalized, in the eSports scene of FIFA, a sports simulation game, there is room for professionalization. One area where there is room for improvement is the area of player performance analysis. This is where the eSportslab from the University of Twente comes into play. By applying state-of-the-art scientific knowledge and measurement techniques, the eSportslab addresses challenges in performance optimization, sensoring, movement science, player psychology, eSports data science, player health, inclusion, and the societal & organizational impact of gaming. The eSportslab has collected player performance data but is searching for a way to analyze this data.

In the search for a way to analyze collected FIFA data, the eSportslab identified business process mining, or process mining in short as a possible way to analyze the FIFA data. Van der Aalst (2004) defines process mining as a technique that is aiming to extract information from event logs and capture the process as it is executed. With this technique, the actions that happen before certain instances such as success or failure can be identified. Translating this into football terms, for example, the actions that happen before a goal can be identified and these actions can help identify strategies that lead to more goals.To see if business process mining can be applied to FIFA data, an assignment with a focus on process mining was formulated by the eSportslab. The assignment from the eSportslab is formulated as “Can process mining give insight into the game states and the gaming processes of FIFA 20?”. In the next section, the assignment is further analyzed.

1.2 Problem statement

In this section, a problem cluster is used identify the core problem the research wants to solve. According to Heerkens (2017), a problem cluster can indicate causally connected problems and bottlenecks. In this research, the problem cluster is used to find the causes for the action problem and identify research elements that come back in section 1.4 (Research approach). Heerkens (2017) also mentions two different types of problems. Action problems, which are a discrepancy between the norm and reality, and knowledge problems, which need research to obtain knowledge to solve the problem. At the end of the problem cluster, a core problem is identified, which the research aims to solve.

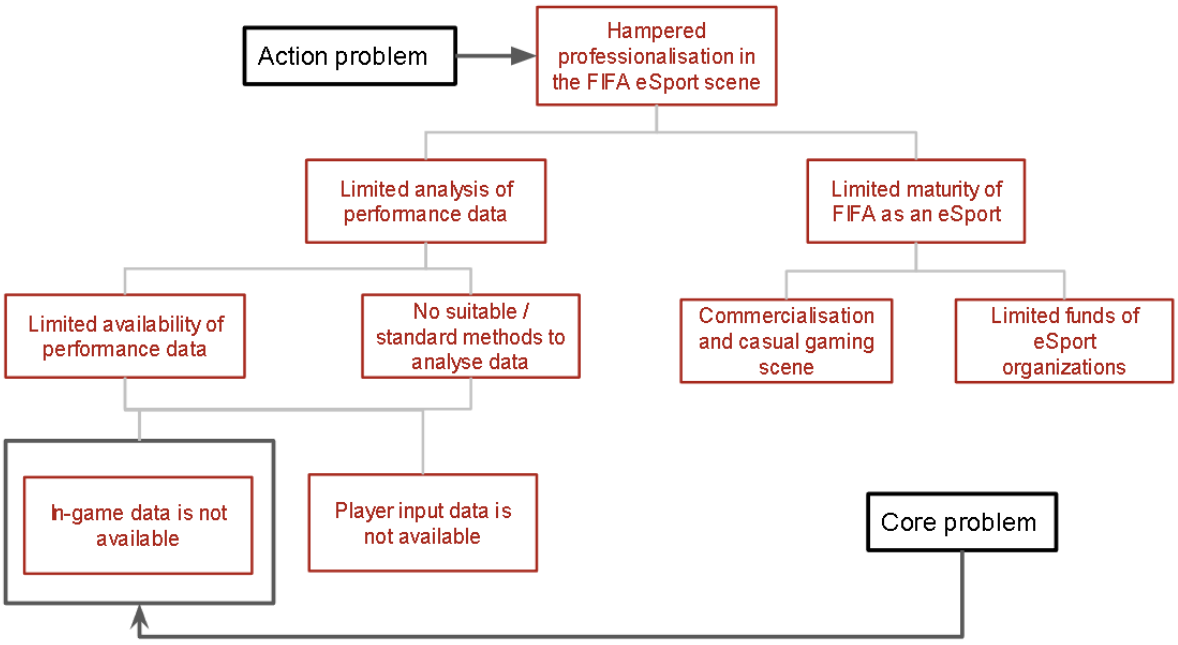


Figure 1: Problem cluster

The action problem is formulated as “Hampered professionalism in the FIFA eSports scene”. The eSportslab wants to further professionalize the FIFA eSports scene and sees that the professionalism within the FIFA eSports scene is not optimal at the moment, and they want to improve that. For this action problem, two underlying problems are identified. Namely, the limited maturity of FIFA as an eSport and the limited analysis of performance data. The limited maturity of FIFA as an eSport has two causes. The first cause is that FIFA is still partly seen as a casual game and not an eSport and the second cause is that the FIFA eSports scene has limited funds in comparison to the further developed eSports scenes. An indicator of the funds available to an eSports scene is the tournament prize money awarded. For FIFA 20, the total prize money awarded was around $2.1 million which is less than the prize money awarded during the world championship of “Dota”, an eSport that has a further developed eSports scene that rewarded players with a prize pool of around 40 million dollars in 2021. The limited funds and the lack of commercialization contribute to the limited maturity of FIFA as an eSport.

On the other side of the problem cluster, the limited analysis of performance data is then caused by two sub-problems. The first sub-problem is that performance data is not available, this is mainly because the publisher of FIFA, EA sports, does not allow for game data collection from their servers. Because of this, the eSportslab has created bypasses to collect performance data, for example, the player input data from the controller can be collected through software and the eSportslab is developing a method to read the screen FIFA is played on and collect data from the screen. The second sub-problem is that there is not (yet) a suitable or standard method to analyze the collected game data. For these two sub-problems, two causes are identified: “In-game data is not available” and “Player input data is not available”. From these two

sub-problems, the problem “In-game data is not available” is chosen as the core problem this research tries to solve.

The reason that the other problem “Player input data is not available” is not chosen is because the other problem aims at creating a (software) workaround to collect data, which makes this part of the core problem more fitting for Computer Science (CS) students to solve. The problem “In-game data is not available” focuses on translating the FIFA data to find if certain sequences of actions are more frequent than others, here the case can be made for the use of business process mining, which is taught in Industrial Engineering & Management (IEM).

In the next section, the methodology is explained to solve the chosen core problem “In-game data is not available”. The goal is to deliver a proof of concept method to analyze the collected controller data. This method will entail the steps where data is collected, cleaned, analyzed, and visualized. The thought behind creating this proof of concept is that if the method works for a data set of a certain size, the assumption is made that it will work for data sets that are larger than that certain size.

1.3 Research methodology

In this section, the research methodology will be explained. The chosen core problem “In-game data is not available” can be solved by researching the applicability of process mining in FIFA to create a method to analyze the data and translate it into insights. The methodology chosen for this knowledge problem is the Design Science Research Methodology (DSRM) from Peffers et al. (2007). The approach to solving the core problem will be further elaborated on in the next section (1.4), where the research approach will be discussed.

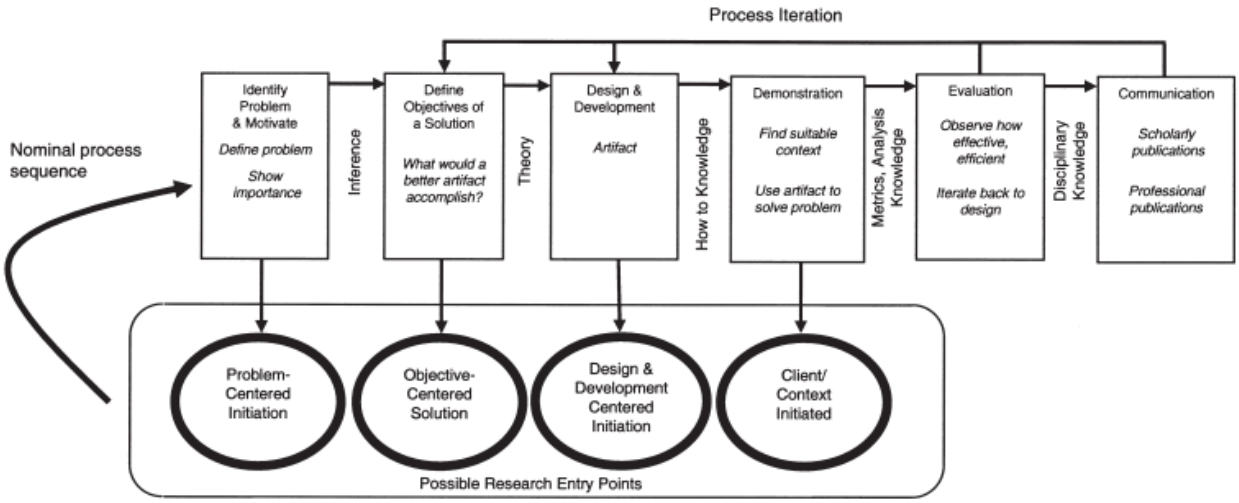


Figure 2: DSRM-model from Peffers et al. (2007)

For IEM research, the more frequently chosen Managerial Problem-Solving Method (MPSM) is not chosen for this research. The DSRM is chosen because the DSRM is more fitting for the way this research will be conducted. While the MPSM-method is the better choice of the two for a problem-solving approach, the DSRM focuses on the creation of an artifact, instead of focusing on solving an action problem (fixing a discrepancy between norm and reality). Concretizing variables for the discrepancy between norm and reality is hard for this action problem which can be defined as the hampered professionalism in the FIFA eSports scene. The goal of this research is to deliver an artifact.

The artifact, in this research, is the proof of concept method to analyze the FIFA data with process mining. A proof of concept is a realization of a method to show that the idea or concept is feasible. With this artifact, the research tries to improve the professionalism in the FIFA eSports scene by improving the use of data analytics within the FIFA eSports scene. The methodology from Peffers et al. (2007) also allows for multiple iterations which can be of use because when in the process of creating something, a return to a previous phase is sometimes needed to revise the solution objectives and/or the design and development of the artifact. In the next section, the research approach is explained and linked with the research methodology.

1.4 Research approach

To carry out this research, research phases are created and described to give the research structure. In these research phases, research questions are formulated. These research phases will form the chapters of this research. The research phases can be linked to the DSRM. The research has already introduced the research context and defined the problem in Chapter 1. The research model is shown in Figure 3, where the structure of the research is visualized. The research will aim to deliver a proof of concept of how the data will be translated into in-game events (Chapter 3), how these in-game events will be analyzed (Chapter 4), and how these findings should be positioned in a dashboard (Chapter 5). At last, the experiences and insights gained from the research will be presented (Chapter 6).

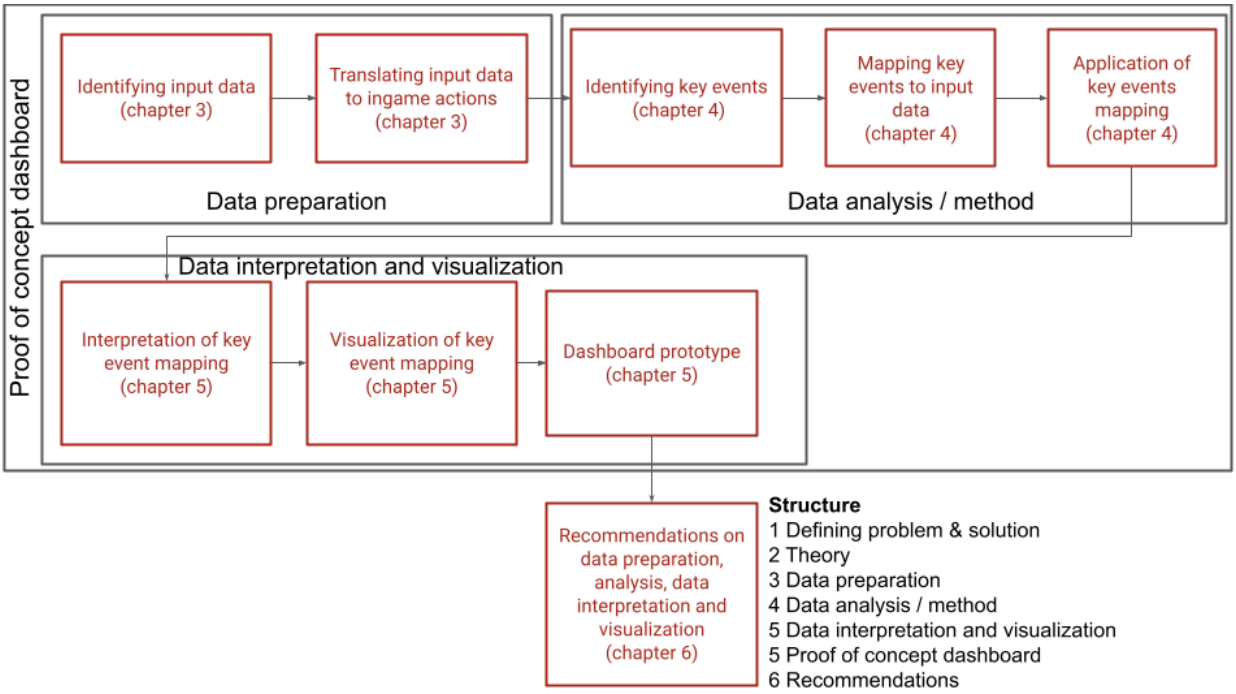


Figure 3: Conceptual model of the research

In Chapter 1, the problem is defined and it is motivated why it is a problem. Moreover, it is explained that with the created method, the goal is to improve the professionalization of the FIFA eSports scene. This chapter links with the first and second phases of the DSRM, problem definition and solution objectives.

In Chapter 2, knowledge questions are answered to gather the necessary knowledge to complete the research. This chapter links with the third phase of DSRM, design, and development. In 1.5, the different knowledge questions are formulated to create a theoretical framework that the research uses to support the decisions made.

1. What are the answers to the knowledge questions formulated in 1.5?

In Chapter 3, the data type of the input data is identified and then translated into in-game actions. This chapter is linked with the third phase of DSRM, design, and development.

1. What input does the player give? By answering this question, the type of data that is being collected becomes known.
2. What are possible process mining techniques to analyze this kind of data? And what is the most fitting process mining technique to analyze this data? By answering this question, different techniques are being considered, and then it is explained why one technique is more fitting than the others. The goal is to translate the data to in-game events.
3. How does process mining on FIFA data look and which steps are taken? By answering this question, the method of applying this technique will become known. In the following section, this process mining technique is applied to a set of test data. From applying this process mining technique to the test data, a list of in-game events is produced, which is used in the next section.

In Chapter 4, firstly, key events are identified from the list of game events, then these key events are mapped to the input data, and lastly, a summary of the process of going from raw data to the findings is given. This chapter is linked with the third phase of DSRM, design, and development.

1. From the list of in-game events, what are the key events? By answering this question, a list of KPIs can be created. To help answer this question, a literature study on football KPIs can be done.
2. Can through analysis of the list of in-game events a pattern be recognized? By answering this question the goal is to find out if certain sequences of events influence the likelihood of certain key events.
3. Which steps were taken to go from raw data to the analysis of in-game events? And how does a generalized overview of these steps look? By answering these questions, a simplified overview of the steps taken is shown.

In Chapter 5, the research will aim at dashboard development. The information from process mined FIFA logs should be displayed in the dashboard, but to do that, the existing progress of the dashboard development should be combined with the findings from Chapter 4. Furthermore, dashboard display best practices are researched to ensure the display quality of the findings from Chapter 4. This chapter is linked with the third phase of DSRM, design, and development.

1. What do the findings from Chapter 4 tell us about the key events?
2. What is the progress of dashboard development of the eSportslab at the moment of this research? By answering this question, it is known what has been developed up until now, then the research can continue from that point and research how the information from the process mined logs can be best displayed in the next question.
3. How can the insights gained from process mining be positioned within the dashboard of the eSportslab?By answering this question, a literature study of visualization best practices will be done to position the process mining findings into the existing dashboard.

In Chapter 6, the following research questions will be answered:

1. Based on the previous chapters, what conclusions can be taken?
2. Based on the previous chapters, can a discussion be made about the findings?
3. Based on the previous chapters, what recommendations can be made to the eSportslab?

1.5 Structure theoretical framework

To carry out the research and answer the research questions (1.4), knowledge of different subjects is required. To do this, knowledge questions are formulated to gain the knowledge needed to carry out the research. Looking back at the problem statement (1.2), topics are identified that give the knowledge necessary to carry out the research. The topics are:

1. Business Process Management (BPM)
2. Data processing, this theme contains the following subtopics:
    I. Data preparation
    II. Data analysis
    III. Data visualization

The first topic “Business process management” is chosen because this research tries to link business process management with the process of playing FIFA. By proving that BPM can be linked with FIFA, certain BPM analysis methods can be applied to FIFA and its processes. With these analysis methods, the research tries to create a methodology to analyze FIFA input data. The knowledge questions linked to the first topic “Business process management” are:

- What is BPM?
- How do FIFA processes fit within a BPM perspective?
- What are possible BPM methods to analyze processes?
- What are fitting BPM methods to analyze FIFA processes?

The second topic “Data processing” and its subtopics “Data preparation”, “Data analysis”, and “Data visualization” are meant to serve as a continuation of the first topic BPM. In the BPM, the research links BPM with processes in FIFA. In this topic, the research wants to further analyze the data gathered from the FIFA processes. The steps to analyze the data are the subtopics mentioned and these subtopics serve to provide knowledge about state-of-the-art data processing methods. The theory from these subtopics will contain more general best practices and the goal is to use these best practices on the gathered FIFA input data and eventually, create a method to prepare, analyze and visualize the FIFA input data.

The knowledge questions linked with the second topic “Data processing” and its subtopics are:

1. Data preparation
    - What is data preparation?
    - What are the best practices to prepare FIFA data?
2. Data analysis
    - What is data analysis?
    - What are the best practices to analyze FIFA data?
3. Data visualization

- What is data visualization?
- What are the best practices to visualize FIFA data?

To conclude, the knowledge questions are used to structure the theoretical framework in Chapter 2. It starts with the topic of BPM and the link with FIFA. Then the different BPM methods are shown, how BPM is used to analyze processes, and which BPM method is most fitting to analyze the FIFA processes. Following that, different phases of how the data is processed are connected to the knowledge questions. The questions asked are what is the phase the data is in and what are the best practices of that phase. The phases defined are data preparation, data analysis, and data visualization.

Chapter 2: Theory

In this chapter, the knowledge questions formulated in 1.5 are answered to form a theoretical framework. The goal of the theoretical framework is to define the main constructs used and offer the reader the necessary insights into the relationships between different concepts.

2.1 Business process management and FIFA

In the next sections the knowledge questions about BPM will be answered. Moreover, the question of how BPM can be linked with FIFA is answered. By answering these knowledge questions, a basis of knowledge is created to support the decision to select which BPM method is most fitting to analyze the collected FIFA data.

2.1.1 What is BPM?

BPM is the management of business processes and business processes are defined by Weske (2012) as a set of activities that are carried out in coordination in an organizational and technical environment. The organizational part is that the player has to coordinate all actions through a controller, and this controller, the technical part, is responsible for translating the actions to the game. Furthermore, BPM includes methods, concepts, and techniques to support the analysis of business processes. Moreover, Weske (2012) states that BPM consists of multiple phases and these phases can be visualized in Figure 4. The phases are cyclically related to each other.

In the *design and analysis* phase, the business process is analyzed through validation, simulation, and verification. Tools and techniques used in this phase are surveys (for validation of the business process model), models of the business process (for simulation of the business process), and workshops (to verify the business process with the stakeholders in the business process).

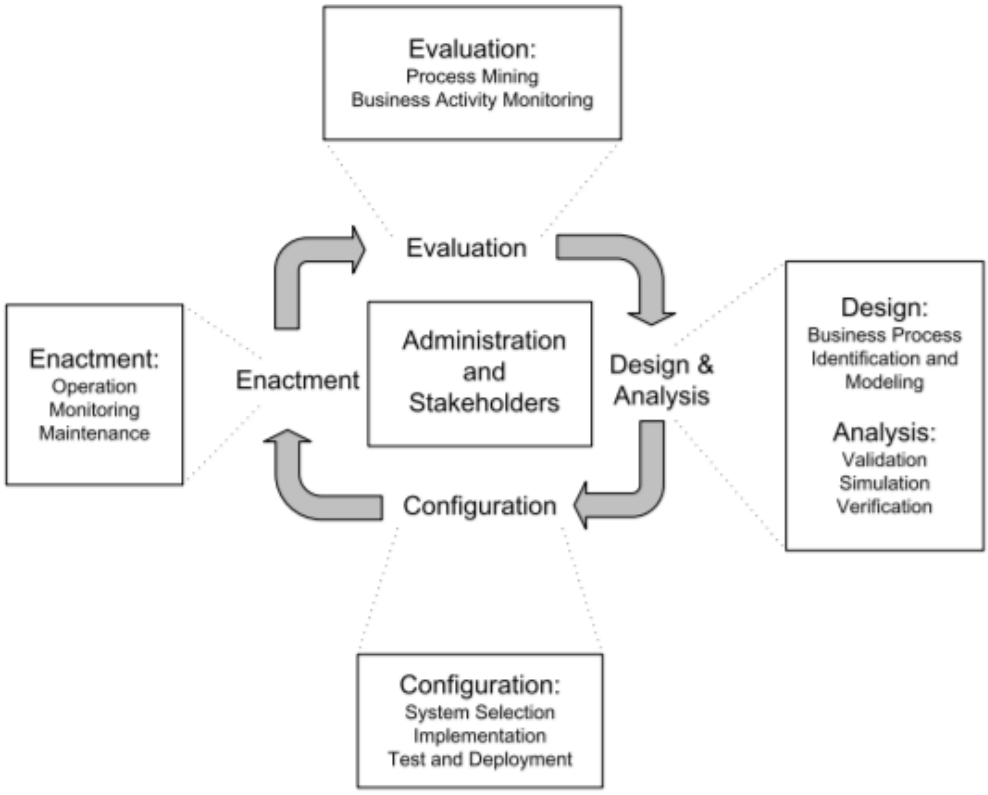


Figure 4: Weske's business process lifecycle (Weske, 2012)

In Figure 4, different methods are mentioned under each of the phases. The business process lifecycle starts in the *design and analysis* phase. In the *design and analysis* phase, Business Process Identification and Business Process Modelling are used. In Business Process Identification, surveys are conducted to identify the business processes. In business process modeling, the business processes are visualized so that different stakeholders within the organization can understand, discuss and improve the business processes effectively (Weske, 2012). An example of a business process put in graphical notation is given in Figure 5. In Figure 5

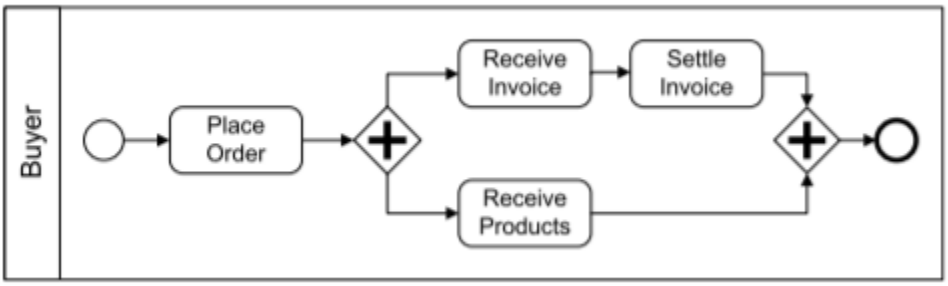


Figure 5: The steps a buyer has to follow (Weske, 2012)

In this business process, the first step is to place an order. Then the buyer has to fill both lanes after “Place Order”. The buyer should receive the invoice and settle the invoice while at the same time, the buyer receives the products. When both lanes are completed, then the lanes merge and the end is reached. It can be seen that different symbols are used for different types of choices. In the case of the example process, the diamond with a plus represents a gate from one business process into two processes that are executed at the same time, hence it is also called the parallel gate. The research will not explain all the existing symbols, but these symbols are there to make the communication about business processes more effective (Weske, 2012).

In the *configuration* phase, after the business process is designed and analyzed, the business process is implemented. Different types of business processes require different ways of implementation. A business process can be implemented through a set of policies and procedures that the workers of the company need to follow. A company could also choose to implement the business process through a dedicated software system since, in today’s world, businesses often make use of software systems to support their business processes. After the system selection and implementation, the business process implementation should be tested. Again, different techniques are available to do this. At a software level, more traditional testing techniques can be used. An example is to check whether the software outputs what is expected. At a process level, it is important to check the integration and performance of the business processes to encounter possible run-time problems.

In the *configuration* phase, business process management systems are configured and linked with existing software systems at a company. When there is a dedicated software system, the business process should be implemented according to the organizational environment of the business. Moreover, the role of different actors in the business process should also be taken into account. After the business process is implemented within the business process management systems, the implementation itself should be tested. The business processes can be tested for expected behavior on an activity level. Business processes can also be tested on a process level where integration and performance tests are performed. When the test phase is complete, i.e. when the tests are complete and the results are satisfactory, then the business process is deployed.

In the *enactment* phase, the business process has already been implemented and now the business process is going to be enacted and the business processes are now used within the company. Within the enactment, the business process is actively controlled by the business process management system. The business process management system can also be used to monitor the status of the business process instances. During the enactment, the collected data from monitoring is gathered, typically, in the form of a log file. Log files are files that consist of events with their start- and end times. They form the basis for the next phase of the business process lifecycle.

The next phase of the lifecycle is the *evaluation* phase. In this phase, the execution logs from the enactment phase are used to analyze the business processes. The execution or event logs are analyzed through the use of business activity monitoring and process mining. Business

activity monitoring aims at identifying when a business process takes too long or if a business process uses too many resources. Process mining is used to develop business process models from the execution logs. Another application of process mining is the evaluation of existing business process models.

2.1.2 How do FIFA processes fit within a BPM perspective?

The previous section aimed at giving an introduction to BPM and the business process lifecycle and its contents. This section serves to explain why BPM, which is normally used on business processes, could be used on FIFA processes. Business processes contain several events and activities (Dumas et al., 2013). Now, the collected FIFA input data is introduced. In short, FIFA 20 is mainly played through a controller which contains buttons. The player can input their actions by pressing buttons in certain combinations and these combinations lead to different actions. The list of which buttons are pressed at which time is generated from the controller. Then software made by the eSportslab converts the list of button inputs into a list of in-game actions. In Table 1, an example of the list of button data collected is shown.

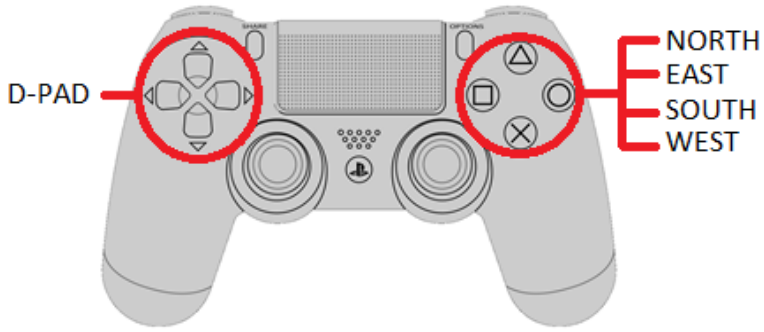


Figure 6: Controller buttons

To read Table 1, Figure 6 is given as support. The “Type” column corresponds with the type of button pressed/used. The “SpecifiedButton” column defines which specific button is pressed, an example is row 2 of Table 1, where the “BTN_SOUTH” is specified. This corresponds to the button on the north side of the triangle in the right red circle in Figure 6. Moreover, “SpecifiedButton” can also return “JOYSTICK_LEFT”, which corresponds with the circle/joystick below to the right of the “D-PAD” in Figure 6. When “JOYSTICK_RIGHT” is returned, it corresponds with the circle/joystick below to the left of the right red circle in Figure 6. The “Pressed” column defines if the button was pressed with a “1” for pressed and a “0” for released or it will define the direction of the joystick press, so when a joystick is moved to the left, it will define “LEFT” as the value but when the joystick is pressed in its place, it will define “CENTER” as the value. The “TimePressed” column defines the time when a button is pressed when the “Pressed” column has a value of “1”. When the “Pressed” column has a value of “0” in combination with the “TimePressed” column, it represents the time when a button is released.

When the “Pressed” column contains a joystick direction or press in combination with the “TimePressed” column, it represents the time when the joystick is touched.

Table 1: Example of list button input

Type	SpecifiedButton	Pressed	TimePressed
Key	BTN_SOUTH	1	10:49:03
Absolute	JOYSTICK_LEFT	LEFT	10:49:03
Key	BTN_SOUTH	0	10:49:03
Absolute	JOYSTICK_LEFT	CENTER	10:49:04
Absolute	DPAD_RIGHT	1	10:49:04
Absolute	DPAD_RIGHT	0	10:49:04
Absolute	DPAD_RIGHT	1	10:49:05
Absolute	DPAD_RIGHT	0	10:49:05
Key	BTN_SOUTH	1	10:49:05
Absolute	JOYSTICK_LEFT	RIGHT	10:49:05
Key	BTN_SOUTH	0	10:49:05
Absolute	JOYSTICK_LEFT	CENTER	10:49:05

The list of button input data is valuable, but it needs to be translated to in-game actions to make it usable, so the eSportslab created a script to convert this button data to data that contains in-game actions, this will be further elaborated on in Chapter 3.

According to Dumas et al. (2013) business processes contain events and activities. Events are things that have no duration, for example, the arrival of a bus at a bus stop. When an activity is simple and can be explained as one single action, then it is called a task. In addition to events and activities, a process will contain decision points. These are points in time where a decision is made and this decision has consequences on the way the process is further executed. The process of playing FIFA contains different events and actions. The decision points return in the way of the playing style of the player. For example, the player could adjust his playing style after they have conceded a goal and shoot more, and the number of shots returning in the list of actions will then grow. Another example is when a player is defending more, which could return more tackles in the list, the decision points influence the events in the list.

Dumas et al. (2013) state that a process also contains several actors, physical objects, and immaterial objects. The number of actors in the FIFA context depends on the level of analysis. Examples of actors are the eSporter, the eSporter's coach, the eSporter's organization, and the eSportslab. Examples of physical objects are the controller used and the screen on which the game is displayed. Examples of immaterial objects within a FIFA context are the tactics and the players (in-game players of FIFA such as a specific goalkeeper or forward) used. Furthermore, not only is the FIFA process similar to a regular business process, the goal of BPM is to ensure consistency and to take advantage of improvement opportunities within the processes of a company (Dumas et al., 2013), this mentality of taking advantage of improvement opportunities is also part of professional sports in which players are pushed to improve and the eSportslab wants to take advantage of this BPM to improve their players.

Another way of linking BPM with FIFA is through the visualization of business processes and the goal of this section is to show that FIFA processes can also be visualized in business process models. Business processes can be visualized through business process models. Weske (2012) defines business process models as models that specify the activities and their relationships within a single organization. Process modeling itself involves capturing a process in a workflow specification (Georgakopoulos et al., 1995). To understand what this means the term workflow needs to be characterized. In the book "Workflow Management Coalition" by Hollingsworth and Hampshire (1995) a definition for workflow is given as "the automation or computerized facilitation of a business process, this can be a process as a whole or part of the process that is automated". Workflow management itself encapsulates defining the workflows and being able to provide fast adjustment or implementation of processes whenever the business needs it (Georgakopoulos et al., 1995).

For process modeling, workflow specification is needed, and workflow specification is the act of capturing a process on a certain level of abstraction and the capture level depends on the needs of the workflow specification. If a high level of understanding is required, then the capturing level of the workflow specification will be on a high conceptual level. An example of when a high level of understanding is needed is when one wants to apply workflow management, which requires an in-depth understanding of the process. Finally, in workflow specification, a workflow specification language is used. There are several workflow specification languages that are used in process modeling, Weske (2012) introduces Petri nets, event-driven process chains, workflow nets, Yet Another Workflow Language, graph-based workflow languages, and Business Process Model and Notation as ways to orchestrate business processes. There are a plethora of workflow specification languages and the research will not go into detail for each workflow specification language. Instead, the research will address control flow patterns that are independent of process languages and instead provide a basis of understanding on how process orchestrations are expressed. Furthermore, one of the best-known workflow specification languages, Petri nets, is introduced as an example on how process orchestrations are expressed.

Control flow patterns provide a basic understanding of how process orchestrations are expressed. Furthermore, control flow patterns are independent of the different workflow

specification languages. Some basic control flow patterns are *sequence*, *and split*, *and join*, *exclusive or split* or *xor split*, *exclusive or join* or *xor join*, *or split*, *or join*, *multi-merge* or *multiple merge* and *discriminator* (Weske, 2012).

2.1.3 Control Flow Patterns within FIFA Context

The research will now explain the mentioned basic control flows that fit within a FIFA context, the other non-fitting control flow explanations can be found in Appendix A. Firstly, the control flow pattern is shortly explained. Following that, a visual of the control flow pattern is shown. And lastly, it is explained how the control flow pattern fits in a FIFA context.

The **sequence** pattern represents an activity B that is enabled after the completion of activity A.

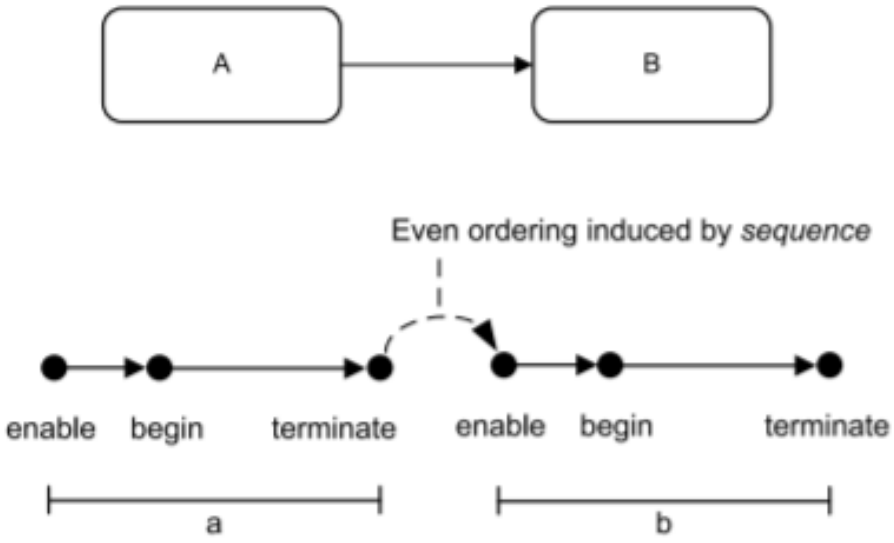


Figure 7: Sequence pattern (Weske, 2012)

How can FIFA be represented in this pattern? Example:
After dribbling, the player shoots the ball. The player cannot shoot and dribble at the same time, the shooting event has to take place after the player has finished dribbling the ball.

Figure 8: FIFA example of the sequence pattern

The **exclusive or split** or **xor split** represents a point in the model where after the completion of activity A, only one of either activity B or C can be enabled.

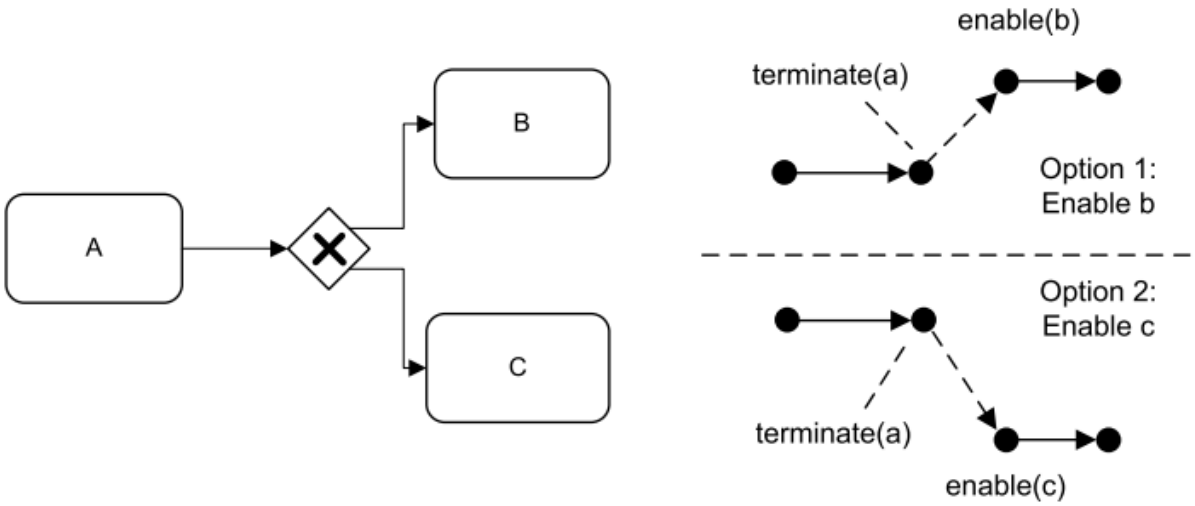


Figure 9: Exclusive or split pattern (Weske, 2012)

How can FIFA be represented in this pattern? Example:
When the player is attacking, after they dribble the ball, the ball could be passed or shot.

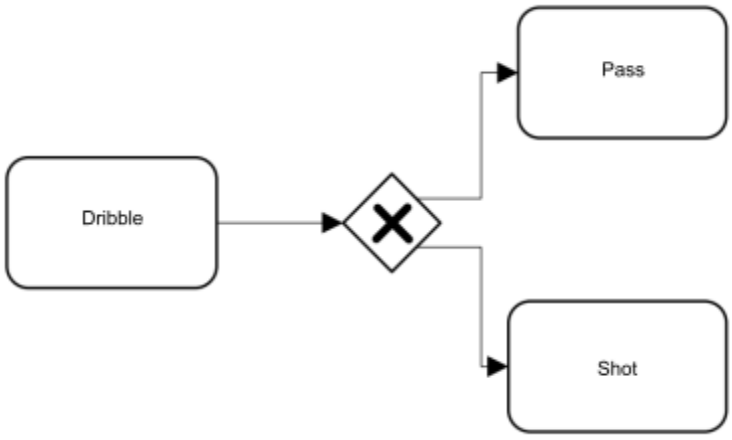


Figure 10: Exclusive or join pattern (Weske, 2012)

The **exclusive or join** or **xor join** represents a point in the model where after the completion of either activity B or C, activity D is enabled.

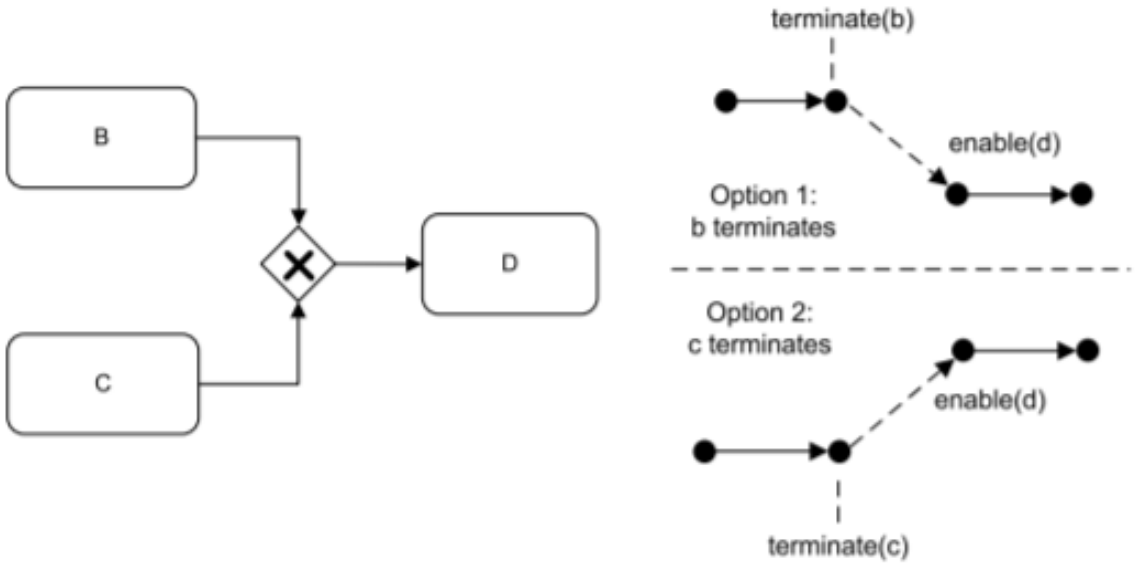


Figure 11: Exclusive or join pattern (Weske, 2012)

How can FIFA be represented in this pattern? Example:
The player can shoot the ball after either dribbling the ball or passing the ball. Activities B and C represent the activities of dribbling and passing, and D represents the activity of shooting the ball.

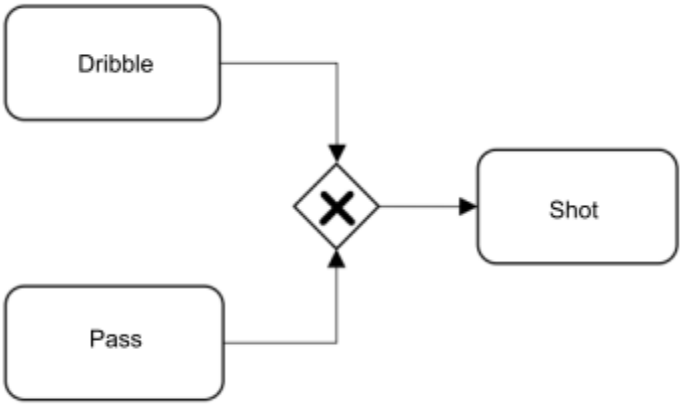


Figure 12: Or split pattern (Weske, 2012)

With this section, the research gave an introduction to control-flow patterns, which form a basis for the patterns used in different workflow specification languages. In the next section, the research will introduce Petri nets as an example of how control flow patterns are represented in a workflow specification language. Petri nets are used to model cases within procedures, organizations, and devices where regulated flows, in particular, information flows, play a role (Reisig, 2012). Merz, Moldt, Müller, and Lamersdorf (1995) noted that the main strengths of

Petri nets are that they combine elements such as a mathematical foundation with an understandable graphical notation, the ability to apply simulations to it, and the ability to do verifications on the modeling workflows. One of the drawbacks of Petri nets is that they have a fixed structure in which only the tokens can change places (Han, 1998).

Petri nets consist of places (represented by circles), transitions (represented by squares), and directed arcs (represented by arrows) that connect the places and transitions. The dynamics within a system are represented through tokens that reside in places. In a Petri net, multiple tokens can be present which represent multiple process instances. The places, transitions, and directed arcs are static but the tokens can change positions in a Petri net. A transition can fire a token if it is enabled. To enable a transition to fire, a token has to be present in each of the input places before the transition. After the transition has been enabled, it fires and adds one token to the output place(s) (Weske, 2012).

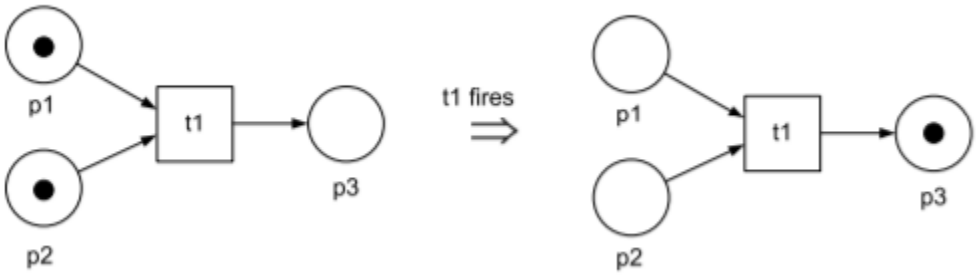


Figure 13: Example Petri net, transition t1 is enabled because places p1 and p2 contain a token, then transition t1 fires a token to place p3 (Weske, 2012)

Now that the control flow patterns and Petri nets are introduced, the research wants to show that FIFA processes can be represented in Petri nets for example. An example is shown in Figure 13. In this figure, the transitions are represented as lines but they could also be represented as squares which are shown in Figure 14. Referring back to the mentioned control flow pattern *exclusive or split* or the *xor split*, this is a split where the token chooses selectively one path to go to. An example of how FIFA can be represented in this is a situation where activity 1 is the act of dribbling with the ball, and the player has to choose between activity 2, which is passing the ball, and activity 3, which is shooting the ball.

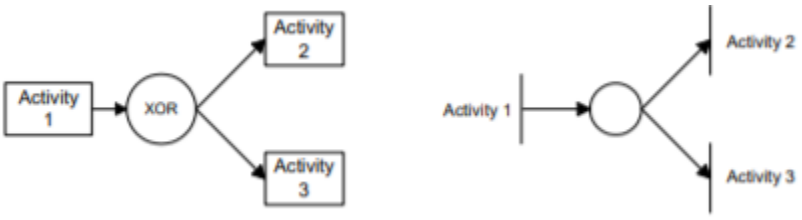


Figure 14: Example exclusive or split or xor split in a graphical construct (l) and in a Petri net (r) from Sivaraman and Kamath (2002)

With these sections about what business processes are and how playing FIFA can be seen as a business process and how business processes can be visualized through business process models and how FIFA could be visualized in a business process model, the research wanted to show the possibility of using BPM on FIFA to improve the FIFA player's performance. In Table 2, a summary of the different mentioned control flow patterns and how these patterns will fit within a FIFA context.

Table 2: Summary of control flow patterns and their fit within FIFA context

Control flow pattern	**Fit within FIFA context (Yes/No)**	**Example**
Sequence	Yes	After dribbling, the player shoots the ball. The player cannot shoot and dribble at the same time, the shooting event has to take place after the player has finished dribbling the ball.
And split	No	No two activities can occur at the same time.
And join	No	No two activities can occur at the same time.
Exclusive or split / XOR split	Yes	When the player is attacking, after dribbling, they have to choose to either shoot the ball or pass the ball.
Exclusive or join / XOR join	Yes	The player can shoot the ball after either dribbling the ball or receiving the ball after a pass.
Or split	No	No two activities can occur at the same time.
Or join	No	No two activities can occur at the same time.

Multi-merge	No	No two activities can occur at the same time.
Discriminator	No	No two activities can occur at the same time.

2.1.4 What are possible BPM methods to analyze processes?

To use BPM to improve FIFA player performance, BPM methods to analyze the process have to be selected. In BPM, processes can be supported through different types of analysis (Weske, 2012). Referencing Weske's business process lifecycle (2.1.1), these different types of analysis belong to one of the phases of the introduced business process lifecycle. Now, to find which phase is most suitable for FIFA processes, it is necessary to identify what type of FIFA data is collected and how this FIFA data could be analyzed with BPM methods. After introducing Table 1, which contains a list of raw button input. This list can be converted to a list of in-game events. Both of these lists can then be identified as execution logs and analysis for execution logs can be found in the evaluation phase of the business process lifecycle. In the evaluation phase, two methods are mentioned to analyze the FIFA data, business activity monitoring and process mining. Weske (2012) states that business activity monitoring can identify within a process if an activity takes too long due to certain shortages.

Process mining has different applications, the execution logs can be used to produce a business process model and existing process models can be evaluated with help of process mining. Of these two BPM methods, process mining has been identified as the most suitable of the two evaluation methods to use on FIFA data. An advantage of process mining is that it makes use of data that has been gathered automatically instead of formulating KPIs and collecting KPI-specific data, which is often gathered manually through either surveys or observation (Van der Aalst et al., 2016). Van der Aalst (2011) states process mining offers a range of tools to analyze and exploit the collected data in order to discover a process model, to check if the process data matches with the existing process model, or to determine the causes of the deviances. He also states that the advantage that automatically-collected data gives over manually collected data is that the insight extracted from this data is based on evidence whereas manually collected data depends on human confidence, which makes automatically-collected data confirm more to reality.

Within process mining, there are three main types of process mining (Van der Aalst, 2012). The first type of process mining is process *discovery*. The process discovery technique takes an event log and produces a process model with only that information. The second type of process mining is process *conformance*. This technique compares the event log of an existing process model with the event log of the same process, the technique can be used to check whether the gathered event log adheres to the expected event log from the process model. The third process mining technique is process *enhancement*. Instead of confirming the alignment between the process model and the event log, which happens in process conformance, process

enhancement focuses on adding extra information onto a previous model to show for example shelf-life or throughput time. To determine which of these process mining types is most fitting to use on the FIFA event logs, further knowledge of each of the process mining types is needed.

Within organizations, the procedures prescribed may not be followed completely (Van der Aalst, 2012a). Therefore it is important to check whether what happens, in reality, is according to the procedures, this can be done with help of process discovery that uses real-life event logs and produces a process model of the actual process. Some applications of process discovery are to take advantage of improvement opportunities (when the difference is noticed between the actual process and the desired process), enable discussions with stakeholders about problems with the process, enable the use of process enhancement (by using the process model as a base). In process conformance, a model and an event log are taken as input. This process model can be discovered through process discovery or be created through other means. Then after the model and event log are taken, conformance checks whether the behavior from the model corresponds with the behavior of the event log, conformance can also be the other way around. Some applications of process conformance are to check whether the process adheres to its documented procedure, to identify deviations from the documented process, to identify where most of the deviations happen, to check the quality of a discovered process model, or to serve as a starting point for process enhancement (Van der Aalst, 2012a).

In process enhancement, additional information from the event log is used to complement the existing process model. Often an event log contains more information than just the actions themselves, examples of this are the age, name, or customer number. This additional information can be used to further analyze the processes. An example of this application is to analyze the waiting times in-between the activities in the event log of a certain customer. By doing this, the bottlenecks of waiting time can be identified through the use of process enhancement (Van der Aalst, 2012a).

Out of the different process mining types, initially, the most fitting one is process discovery, because it is the first step to create process models out of the FIFA data. The FIFA data itself is a collection of data and the research wants to make sense out of this data and process discovery is seen as the most fitting method to make sense out of the unstructured data and to achieve one of the objectives of BPM which is to gain insight into the FIFA (business) processes. In process discovery, different process discovery algorithms are used and each of these algorithms has its own qualities. The theory and the different elements of the theory are shown in Figure 15 to give an overview of which theory is given until now.

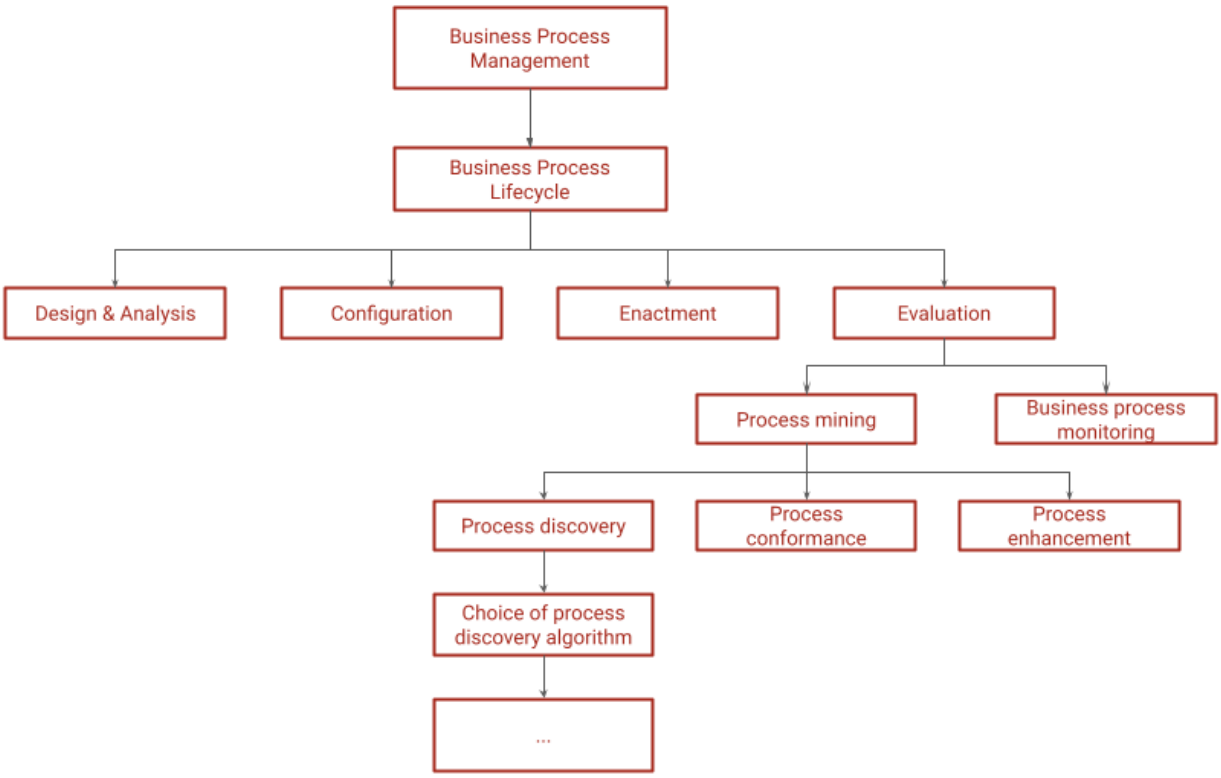


Figure 15: Choices made within the BPM theory

In the next section, the research is going to compare different process discovery algorithms. The selection of algorithms will depend on different factors. The first factor is the degree of suitability of the algorithm to be used on real-life data. The collected FIFA data is unstructured and it is an advantage if the algorithm is able to deal with real-life unstructured data. The second factor is if the algorithm is available in the process mining software “ProM” that is used in this research. If the algorithm is not available, then the next most suitable available algorithm is chosen.

2.1.5 What are fitting BPM methods to analyze FIFA processes?

Process discovery has been shortly introduced in the previous sections. In this section, the research will go more into what a process discovery algorithm does and which process discovery algorithms there are. Moreover, the suitability of these process discovery algorithms is discussed.

Van der Aalst (2011) gave this definition for a process discovery algorithm that is used on a general process discovery problem. “A process discovery algorithm is a function that maps L, which is an event log, onto a process model such that the model is “representative” for the behavior seen in the log. The challenge is to find such an algorithm” (p. 125).
Process discovery algorithms are generally discussed through the use of four quality dimensions, the dimensions are replay fitness, simplicity, precision, and generalization (Buijs, Van Dongen, & Van der Aalst, 2012). In the next section, these quality dimensions are

introduced.

- Replay fitness defines how accurate a discovered process model can reproduce the cases recorded in the log. Generally, process discovery algorithms that focus on replay fitness are region-based approaches (Bergenthum, Desel, Lorenz, & Mauser, 2007) and the multi-phase miner (Van der Werf, Van Dongen, Hurkens, & Serebrenik, 2009).
- Simplicity defines how complex a discovered process model is. Process discovery algorithms can lead to spaghetti-like models, which are complex and hard to read. A type of process discovery algorithm that focuses on simplicity is the $\propto$-algorithm (Van der Aalst, 2011). These algorithms result in simple process models but the models lack replay fitness and/or precision.
- Precision defines how much-unseen behavior is allowed by the discovered process model. A model has good precision when it only allows a minimal amount of unseen behavior. Generally, process discovery algorithms with good precision are region-based algorithms (Bergenthum et al., 2007).
- Generalization defines how much the current discovered model is able to reproduce future behavior. Another way of describing generalization is the level of confidence in the precision of the model. A model that is too precise will have a harder time reproducing future log behavior because it is too fitted on the present log behavior. Process discovery algorithms that generalize well are the fuzzy miner (Günther & Van der Aalst, 2007) and the heuristics miner (Weijters & Van der Aalst, 2003).

There are a plethora of process discovery techniques and to find out which process discovery algorithm is most suitable a different study was consulted. The study “A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs” by De Weerdt, De Backer, Vanthienen, & Baesens (2012) compared state-of-the-art process discovery algorithms based on their performance on real-life event logs. The performance comparison on real-life event logs makes this study fitting because the FIFA data is also real-life data. The study focussed on the accuracy and the comprehensibility of the process models. Accuracy was defined as a combination of recall (replay fitness), precision, and generalization. Comprehensibility was defined as the understandability of the process models, this can be compared to the quality dimension simplicity. The study then continued to discuss different process discovery algorithms, from the early approaches to discovery algorithms to the $\propto$-algorithms and their successors to techniques that found their origin in machine learning theory and other process discovery approaches. After having discussed the algorithms, the study explained which metrics were used to evaluate the accuracy and comprehensibility.

Then, the same study by De Weerdt et al. (2012) compared seven state-of-the-art process discovery techniques, namely: $\propto^{+}$, $\propto^{++}$, AGNEsMiner, Genetic Miner, Duplicate Task Genetic Miner, HeuristicsMiner, and ILP Miner. These techniques were applied on artificial event logs and on real-life event logs, the focus was on the accuracy and comprehensibility of the real-life event logs. The results from both the artificial logs and the real-life event logs were used in a multivariate analysis to come with a more general conclusion. The study concluded the HeuristicsMiner to be the most suitable technique to use within a real-life context in terms of

accuracy, comprehensibility, and scalability. Moreover, the HeuristicsMiner is also available in the process mining software ‘ProM’. The HeuristicsMiner is chosen as the process discovery algorithm to use in this research.

2.2 Data preparation

In the introductory part of this chapter, the research defined BPM and data processing as the topics. In the next sections, the subtopics data preparation, data analysis, and data visualization are introduced. After that, best practices for each of the subtopics are introduced and how these best practices link with FIFA data. To help how subtopics fit within the process of gathering, processing, and analyzing the data, a visualization of the FIFA data process is given.

2.2.1 What is data preparation?

Process mining is a field of research where the research aims to combine BPM with data mining (Van der Aalst, 2012b). In order to find the “best practices for process mining”, the research will focus on finding data mining best practices.
Data preparation consists of techniques that can be applied to raw data in order to convert the raw data to quality data. Examples of these techniques are data collecting, data integration, data transformation, data cleaning, data reduction, and discretization (S. Zhang, C. Zhang & Yang, 2003). The need for data preparation arises from the fact that real-life data may be incomplete, noisy, and/or inconsistent. High-performance mining techniques require data of high quality and using quality data yields more concentrative patterns.
Data preparation is a part of the process that is knowledge discovery in databases (KDD), which Zhang and Zhang (2002) defined as an iterative process consisting of four phases:

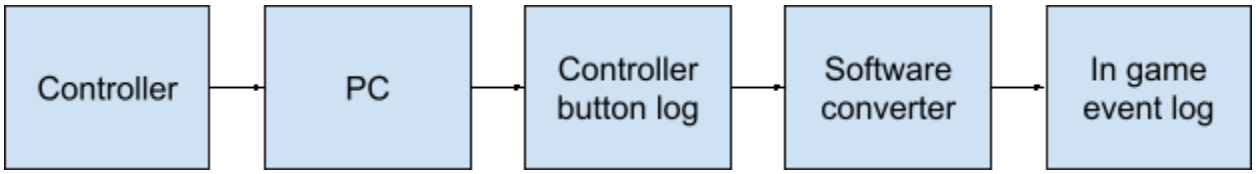
- Problem definition, in which the project goals are identified.
- Data preparation or data pre-processing, in which the previously mentioned techniques are used to convert raw data to quality data.
- Data mining, in which techniques are applied on the quality data to extract patterns.
- Post-data mining, in which the pattern is evaluated, the model maintenance is deployed, and the knowledge is presented.

The iterative element of the process is that problems can be encountered that require a return to a previous step. An example is that in the data mining step, it is revealed that the data needs additional cleaning in the data, so a return to the preparation step is advised.

2.2.2 What are the best practices to prepare FIFA data?

In this section, the goal is to find data preparation practices from related works that could fit within the combination of contexts of FIFA, BPM, and process mining. First, to determine what the best practices could be to prepare data within a FIFA context, the steps of going from controller data to in-game events are visualized to give an overview of the steps that the data has to go through.

Figure 16: Overview of the steps the data goes through



In Figure 16, different steps are visualized of the process the controller data goes through. During each step, different problems with data quality could be encountered. In the FIFA data process, there are two steps where data quality comes into play, namely, the step where the button input from the controller is gathered into a log and the step where the log with the button input is converted to a log of in-game events.

- In the first step (where data is gathered from the controller and transferred to the pc into a log), the problems encountered with the data quality could be missing data and imprecise data. Incorrect data is not applicable because only the button input and the timestamps are collected, this also prevents the irrelevant data quality problem.
- In the step where the software converts the button input log to a log of in-game events, missing data, incorrect data, and imprecise data may be present. Irrelevant is not present because this list is directly derived from the button input log, so that is where the irrelevant data could be present. Missing data and incorrect data could be present in this list because the software could wrongly interpret button combinations and miss in-game events or convert the button combinations incorrectly to the wrong in-game events. Imprecise data is also present. An example is the in-game event “Skill move”, which represents a skill move and in FIFA there are a plethora of skill moves but the software is at the moment of this research not able to precisely name the exact skill move. During these steps, different data quality problems were encountered, for these different problems different best practices exist. García, Luengo, & Herrera (2015) listed different ways to tackle the different data quality problems and it is discussed how these practices could be used to solve the FIFA data quality problems.

Missing data (first and second step)
Different problems are typically associated with missing data.

1. Efficiency loss
2. The complication of data handling and analysis
3. Bias as a result of the difference between missing and complete data

With the FIFA data, the third problem influences both the first and second problem. If the controller log misses controller inputs, the pattern of button inputs misses and ‘changes’, this will complicate the data analysis part and as a result, the analysis has a chance to be less useful and thus resulting in a loss of efficiency.

To tackle the missing data, three approaches are often used:

1. Discard the examples with missing data

2. Use maximum likelihood procedures to come up with estimated parameters of a model, and to use this model for imputation through sampling.
3. Fill in the missing data with estimated data.

For the first step (from the controller to pc and log), none of the approaches fit within the FIFA data context. The first approach does not fit because removing the examples would lead to a worse result, the best practice for the first step should be to aim to extract every button input. The second and third approach do not fit because the FIFA button inputs are not numerical values, instead the values are so imputed through model samples or estimation is not realistic. A possible solution would be to compare the logs with the events that are happening on screen and fill in events if there are events on screen that are not present in the logs.

For the second step (button log to in-game event log), none of the traditional approaches fit. The first approach does not fit because again, removing examples would harm the analysis, where a process model is made with the in-game events. Both the second and third approach would again not fit because the button input is not numerical and inserting generated data from a model or estimated data would represent events that could influence the results of the data analysis, because of this, the traditional approaches are not recommended.
To conclude, for both steps the traditional approaches don't fit because the FIFA data is not numerical. To tackle the missing data problems, a best practice should be made of noting down the missing data and researching the accuracy of the software that extracts the button input as well as researching the accuracy of the software in the second step.

Imprecise data (first and second step)
The imprecision of the data is about the timestamps next to the button input, if this timestamp data is not recorded precisely enough, then the translation of the button input to in-game events will suffer from this. A best practice to keep the data from becoming imprecise is to keep the timestamp data detailed. There are different forms of timestamp data, so the recommended timestamp data includes fractional seconds because button inputs can translate to different in-game events depending on the time interval between the button inputs. Moreover, by having more precise timestamps, the possibility of translating the exact "Skill move", instead of just labeling it "Skill move" will become greater because it is also one of the cases where different in-game events depend on the time interval between button inputs.
To conclude, for both steps it is important to have timestamp values that are precise and contain fractional seconds, this should be a best practice within the whole data process.

Incorrect data (second step)
The incorrect data problem stems from the fact that there is a chance that the software incorrectly translates the button input to in-game events. To prevent this, timestamp data should be precise. Another way to prevent incorrect data is to research the software accuracy, does the software correctly determine from the button log that an in-game event occurred or has the software made a mistake.

To conclude, imprecisions in timestamp values and converting software can lead to incorrect data, to tackle this keeping up the precision within timestamp values was recommended in the imprecise data part. To tackle the software errors, the log of in-game events could be compared to what happened on the screen.

2.3 What is data analysis?

After preparing the data, the next step is to analyze the data. Data analysis can be described as the accurate evaluation and complete exploitation of the data that was obtained (Brandt, 1998). The method of data analysis is the chosen process discovery algorithm which is explained in section 2.2. In the next section, the aim is to find out what process mining best practices there are to accurately evaluate the data and to exploit the data completely.

2.3.1 What are possible best practices to analyze FIFA data

In the process mining manifesto, which is a guide written by Van der Aalst et al. (2011), 6 process mining guiding principles are listed. Applying process mining on real-life event logs can lead to mistakes and these guiding principles aim to prevent those mistakes and increase the maturity of process mining as a tool for supporting business processes.

1. Event data should be treated as first-class citizens.

Process mining starts with the event data and this data will form the basis from which models are made. Event data can be stored in a collection, this is called an event log but event data can also be stored in other means such as databases, mail logs, and many others. The most important thing in this guiding principle is the quality of the event data because this will mostly dictate the process mining result. Because of this, the event data should be treated as first-class citizens. Currently, this is often not the case. Frequently, event data is a “by-product” used for debugging or profiling.

The quality of the event data can be judged according to Van der Aalst et al. (2011) by different criteria:

- Event data should be trustworthy, i.e. what is listed in the event data is correct and has the right attributes.
- Event data should be complete, i.e. there should be no missing events.
- Recorded event data should have well-defined semantics.
- Event data should be safe, i.e. privacy and security concerns are addressed.

Different maturity levels for event logs are described by Van der Aalst et al. (2011), these are listed below in an order from best (5 stars) to worst (1 star):

- 5 stars: the event log is trustworthy and complete and events are well-defined. The events are recorded in an automatic, systematic, reliable, and safe manner. Privacy and security concerns are addressed. The events recorded have clear semantics, i.e. existence of at least one ontology. The events and the attributes should point to this ontology.

- 4 stars: the logs are recorded automatically in a systematic and reliable way, which results in trustworthy and complete logs. Process instances and activities are supported in an explicit manner.
- 3 stars: the logs are recorded automatically, though there is no systematic approach to record the events. There is some level of trustworthiness and completeness in the recorded logs.
- 2 stars: the logs are recorded automatically as a by-product of an information system. There is no systematic approach followed to decide which events to record. It is also possible to skip recording the events in the information system. As a result of this, events could be missing or not recorded correctly.
- 1 star: the logs are of poor quality. Events in the log could not correspond to reality and there could be events missing. Typically, event logs recorded by hand have such characteristics.

2. Log extraction should be driven by questions. What this guiding principle aims at is the way the event log is tackled. To extract meaningful data, concrete questions ought to be asked. Without concrete questions, it will be hard to select the relevant tables for data extractions. For example, given a database, one can discover different process models depending on the level of concreteness of the questions asked.
3. Concurrency, choice, and other basic control flow constructs should be supported. Many process modeling languages exist and some of these languages offer many different modeling elements. Basic control flow elements such as sequence, parallel routing (AND-splits/joins), choice (XOR-split/joins), and loops. It also recommended having OR-splits/joins because this control flow construct is compact in representing inclusive decisions and partial synchronizations.
4. Events should be related to the model elements. Discovered process models can cover different perspectives such as organizational, time and data perspectives. The other process mining types, conformance checking, and enhancement research the relation between events in the log and model. Conformance checking uses this relationship to replay the event log on the model and reveal possible discrepancies between a model and event log. The relation between events in the log and the elements in the process model forms the basis for different types of analysis.
5. Models should be treated as purposeful abstractions of reality. A process model derived from event data gives a view of reality, but an event log can give multiple views of reality that might be useful. Various stakeholders might require different views. This guiding principle wants to provide 2 insights. The first insight is that depending on the use and stakeholder requirements, different models can be generated from the data. The stakeholders may want to view the model at different levels. At a strategic level, with long-term effects and containing aggregate data, at a tactical level, with medium-term effects and most recent data, and at an operational level, with immediate effects and based on data from events that are currently active.The second insight is that it is important to select the right representation of the process model for its intended audience. This will help with visualizing the results for the end-users and guide discovery algorithms to suitable models.

6. Process mining should be a continuous process. For process mining, both historical and current data can be used to generate process models. Given the fact that in a dynamic environment processes tend to change more often than not, process mining should not be seen as a one-time activity and instead should be seen as a continuous process. The goal should not be to produce a fixed model. Instead, the goal should be to allow users and analysts to view process models that are updated continuously. Process mining should provide actionable information according to different time scales (minutes, hours, days, weeks, etc.).

In the same process mining manifesto, several challenges were listed. These are challenges that can be stumbled upon during the process mining process. Knowing these challenges will be helpful during the creation of the process in the next chapter.

1. Finding, merging, and cleaning event data
   Several problems might emerge with extracting the event data for process mining: the data has to be extracted from different sources, the data might be incomplete, the data might contain outliers, and the data might not be at the same level of granularity.
2. Complex events with diverse characteristics
   Different event logs can differ completely from each other. One type of event log can be very large whereas the other might only consist of 2 columns.
3. Creating representative benchmarks
   To compare the process mining tools and algorithms, benchmarks with representative quality criteria are needed.
4. Dealing with concept drift
   During the analysis of the process, the process could change and this is called concept drift. Because of concept drift, understanding the possible changes that can happen to a process is important for process management.
5. Improving the representational bias used for process discovery
   It is important to separate the process mining result from the visualization used. Each language used (BPMN or Petri nets) has its limitations. These limitations in for example the type of processes they can display can limit the process mining result. It is important that the choice for the process language is carefully thought out and well-defined.
6. Balancing between quality criteria
   There are four quality dimensions in process mining: fitness, simplicity, precision, and generalization. Balancing these quality criteria is a challenge and algorithms should be selected based on their combined score on these criteria.
7. Cross-organizational mining
   Sometimes, event logs from multiple organizations are available for process mining. In the case of processes that are spread out over multiple organizations, this could happen. Consider the setting where the organizations work together. The process mining techniques should also consider privacy and security because the different organizations might not want to share their respective data with other organizations.
8. Providing operational support
   Nowadays, data sources are updated in real-time, this enables process mining to be applied in real-time, and thus, process mining should not only be with historical data but

also with online operational support. Three types of operational support activities are defined: (1) detect, when a case deviates from what is normally expected during a process the system can be alerted, (2) predict, when historical data is used to make predictive models to guide process instances, (3) and recommend, creating a recommender system based on the previously made predictive models.

9. Combining process mining with other types of analysis
   One of the goals of process mining is to create live process models, which means that process models should be updated frequently or daily. New event data is then used to monitor the behavior and detect changes. These interactions require intuitive user interfaces that give the user suitable types of analysis and hide the algorithms and process mining techniques behind this interface
10. Improving understandability for non-experts
    The user could have problems understanding the output or make incorrect conclusions. To avoid this, the process mining results should be represented appropriately and always indicate any deviations from the quality criteria.

2.4 What is data visualization?

Data visualization can be regarded as presenting the data in a form, graphical or pictorial, to make the information easier to understand (Sadiku et al., 2016). The process of data visualization contains the design, development, and application of graphical data representation. Nowadays, the graphical data representation is most often computer-generated. In the next section, the goal is to collect best practices from data visualization which can help with the process of creating process models.

2.4.1 What are data visualization best practices?

First, the different types of dashboards are defined to give an understanding of the purpose of the dashboard and its traits. Few (2006) has divided dashboards into three high-level categories: strategic, operational, and analytical. Different dashboard categories have different purposes, timeframes, data scopes, update frequencies, and interactivity. For this research, defining what category the dashboard will belong to will help with the design of the dashboard. To decide what category the dashboard belongs to, a summary table given by Few (2006) is used to decide the dashboard category.

Table 3: Summary of dashboard categories and their attributes (Few, 2006)

Category	**Purpose**	**Timeframe**	**Scope of data**	**Update frequency**	**Interactivity**
Strategic	Seeing, questioning, strategizing	Static	Enterprise-wide, cross-business unit	Moderate	Low
Operational	Monitoring,	Real-time	Business-unit	High	Moderate

	acting		specific		
Analytical	What-if scenarios, questioning	Static	Enterprise-wide, cross-business unit, or isolated	Low	High

A short reminder of what the research wants to deliver. A proof of concept method to collect, analyze and visualize the FIFA data in a dashboard. From the different purposes, the strategic category fits the most because the FIFA data has already been collected. Seeing and acting (belonging to the operational category) and doing what-if scenarios (belonging to the analytical category) do not belong to the proof of concept method. From the different timeframes, there is not a real-time timeframe planned for the proof of concept, so the chosen timeframe will be static. From the different data scopes, the proof of concept method is planned only for the FIFA button input. From the different update frequencies, that will depend on how often the FIFA data set will be updated. From the different interactivities, the dashboard will not be highly interactive, the core task of the dashboard is to present the findings from the analysis. Concluding, from looking at the different dashboard attributes, a strategic dashboard is the chosen category for the dashboard.

This second section aims to collect general data visualization best practices and to use these practices in the creation of the dashboard. In “Principles of Effective Data Visualization”(Midway, 2020), sequential principles are described which were designed to improve scientific visualization. Because visualization and figure making is often not formally taught, the principles presented aim to guide the authors in improving their visual message.

1. Diagram first: Think about who will see your data and which function the data visualization should serve.
2. Use the right software: Use software that is fitting for the complexity, technicality, and effectivity that you want to convey.
3. Use effective geometry and show data: Often there is more than one form of geometry to consider when you want to display your data. Underlying the decision about which geometry to use should be the data-ink ratio (Tufte, 2001), i.e. the ratio of ink used on data versus the ratio of ink used in total. A high data-ink ratio is preferred.
4. Colors always mean something: In a study to find what makes visualizations memorable, visualizations containing colors were found to be more memorable and visualizations containing at least seven or more colors were found to be most memorable (M. A. Borkin et al., 2013). The majority of colors used in visualizations use color in 3 different schemes. (1) Sequential, the use of color in a range from light to dark. (2) Diverging, the use of 2 sequential color schemes to picture 2 extremes. (3) Qualitative, the use of color when the color intensity is not important, instead different unrelated colors are used to convey qualitative group differences.
5. Include uncertainty: Two primary challenges with including uncertainty in the visualizations. (1) the failure to include uncertainty and (2) misrepresentation of

uncertainty. Displaying uncertainty requires the reader to have an understanding of different ways to express uncertainty such as confidence intervals or standard deviation. Moreover, the author is responsible for choosing the correct way to visualize uncertainty. Furthermore, expressing uncertainty is important but also keep in mind what the interpretation of the uncertainty will be for the reader.

6. Panel (Tufte, 2001), when possible (small multiples): Repeat the structure of the figure used to highlight the differences. What is meant by the structure are the axes, axes scales, and geometry.
7. Data and models are different things: Explaining raw data and summarized data is often easier than explaining a model to the reader. Any visual ought to be explained to the reader so that the user can fully grasp what the visual is representing.
8. Simple visuals, detailed captions: Captions should aim to explain the visualization and the representations used and preferably be able to explain the visualization without needing to see the visualization.
9. Consider an infographic: Where figures tend to focus on representing models and data, infographics incorporate other elements such as text and images. Infographics were found to have the highest memorability score (M. A. Borkin et al., 2013).
10. Get an opinion: The most effective visuals are visuals that connect with their readers. Because of this, authors are encouraged to seek feedback from external parties.

In “Better Data Visualizations” by J. Schwabish (2021), five guidelines for general data visualization were given. These guidelines aim to show what should and should not be done while creating the data visualizations.

1. Show the data
   Sometimes all the data is shown in the chart or graph and this will make it hard to see which data points matter the most. Though it is not about showing the least data possible, it is about showing the data that matters the most.
2. Reduce the clutter
   There are many ways unnecessary visual elements can be present in the visualization. Basic elements such as heavy tick marks and gridlines should be avoided. Sometimes for data marking, squares, circles, and triangles are used to differentiate between series of data, but when these markers overlap, it results in a chaotic visualization. Try to simplify the graph by removing the extraneous or distracting elements so that your data is clear and comprehensible.
3. Integrate the graphics and text
   There are three ways to integrate the graphs and visuals. (1) Remove the legends when possible and label the data directly. By removing the legend, the reader is forced to look at the visualization and try to distinguish each detail whereas with a legend, the reader could look up one specific data series. (2) Write the title like a newspaper headline. Active titles tell the reader what should be taken away from the graph. The goal with this is to represent the results and showcase the message of the graph. (3) Add explainers. When the graph and its title are settled, review it and ask if the extra text is necessary. The additional annotation will help readers with less prior knowledge grasp the message

of the graph more quickly. The graph should explain to the reader how to read the graph and how to understand its content.

4. Avoid the spaghetti chart
   The small multiples approach, the use of multiple single charts can be used to address the readability issues when a single chart is cluttered with data. There are three advantages to the small multiples approach. (1) If the first chart is readable, the remaining charts are readable. (2) More information can be displayed without confusing the reader. (3) Small multiples allow the reader to compare the single charts across multiple variables. There are also pitfalls when using the small multiples approach. (1) Arranging the charts in an illogical order. (2) Inconsistent structure of the charts. (3) Making the charts hard to understand, small multiples charts are intended to be small so the chat should be easy to understand.
5. Start with gray
   When a graph is made, start with the graph in fully grey color and only then start thinking about how the colors are used. By starting in grey, the author is forced to think about the purpose of the chosen colors, legends, labels, and other elements.

Ware (2012) defined information visualization guidelines to use when presenting information.

1. The graphical representation of data should be designed while taking into account human capabilities. To recognize the important data elements and data patterns more quickly.
2. More important data elements should be more visually distinct than less important data elements.
3. When numerical quantities are greater than others, they should be made more visually distinct through the use of graphical elements.
4. The use of graphical symbols should be consistent within the system it is used in.
5. If there are two or more visualization tools that can perform the same task, choose the visualization tool that can perform the task the fastest.
6. Only consider novel design solutions when the payoff will be larger than the cost of learning the novel design solution.
7. Only use novel tools if the benefit of using novel tools outweighs the costs of the inconsistency. Otherwise, use tools that are commonly used.
8. Efforts spent on developing tools should be proportional to the expected return on investment.

To conclude, the dashboard type chosen for this research is a strategic dashboard. In this dashboard, the FIFA data insights should be visualized. For this visualization, best practices from different authors are collected. A summarization of the guidelines is given below. This summarization categorizes the guidelines per subject of the guideline:

Table 4: Data visualization best practices categorized

Category	Best practice
Audience	- Diagram first (Midway, 2020). - Get an opinion (Midway, 2020). - Design while taking into account human capabilities (Ware, 2012).
Process	- Use the right software (Ware, 2012). - Use a new design solution only if the payoff is worth learning it (Ware, 2012). - Only use novel tools when it is worth the inconsistency (Ware, 2012). - Efforts spent on developing tools should be proportional to the expected return on investment (Ware, 2012).
Color usage	- Think about the colors used (Midway, 2020). - Start with gray (Schwabish, 2021).
Structure	- Use effective geometry and show data (Midway, 2020). - Use panels (Tufte, 2001) - Repeat the structure of the figure used to highlight differences (Midway, 2020). - Avoid the spaghetti chart (Schwabish, 2021).
Data	- Include uncertainty (Midway, 2020). - Data and models are different things (Schwabish, 2021). - More important data elements should be more visually distinct than less important data elements (Ware, 2012).
Visualization	- Simple visuals, detailed captions (Midway, 2020). - Consider an infographic (M. A. Borkin et al., 2013) (Midway, 2020). - Reduce the clutter (Schwabish, 2021). - Integrate the graphics and text (Schwabish, 2021). - When numerical quantities are greater than others, they should be made more visually distinct through the use of graphical elements (Ware, 2012). - The use of graphical symbols should be consistent within the system it is used in (Ware, 2012).

Chapter 3: Input data

This chapter answers the knowledge questions defined in 2.1 with help of the theoretical framework defined in Chapter 2. In the first section of this chapter, the knowledge question “What input does the player give?” is asked. By answering this question, the type of data that is being collected becomes known. In the second section, the knowledge question “What are possible techniques to analyze this kind of data? And what is the most fitting technique to analyze this data?” is asked. By answering these questions, different techniques are being considered, and then it is explained why one technique is more fitting than the others. The goal is to translate the data to in-game events. At last, in the third section, the knowledge question “How does the process of applying this technique to the data look?” is asked. By answering this question, the method of applying this technique will become known. Consequently, this technique is applied to the set of test data. This will then return a list of in-game events.

3.1 Player input data

To start, the player input data was introduced in 2.1.2. Following that, the data type of the player input data is defined. In short, FIFA is a football simulation game where users can control their players through a controller. This controller has several buttons and two joysticks that can serve as input. The buttons and joysticks can be used separately and in combination to perform in-game actions such as passing, dribbling, or shooting the ball.

In this research, a Playstation controller is used. FIFA can also be played through other controllers but this research uses a Playstation controller. The controller has a D-pad on the left side which has buttons in four directions: up, down, left, and right. On the right side, the controller has four buttons differentiated through the symbols: triangle, square, cross, and circle. In the middle of the controller, there are two joysticks. The left joystick is mainly used for controlling the player’s movement and aiming of shots. The right joystick is used for advanced controls. There are four buttons on the backside of the controller, these are also used for advanced controls. The eSportslab has created a tool to collect which buttons are pressed at which time. An example of how the table with the controller input looks is given in Table 5.

Figure 17: PS4 controller with D-pad, joysticks and symbol buttons

Table 5: Example table of controller button input

Type	SpecifiedButton	Pressed	TimePressed
Key	BTN_SOUTH	1	10:49:03.236466
Absolute	JOYSTICK_LEFT	LEFT	10:49:03.244468
Key	BTN_SOUTH	0	10:49:03.356480
Absolute	JOYSTICK_LEFT	CENTER	10:49:03.776309

The eSportslab uses a software tool that converts the raw controller button input from the joysticks, D-pad, symbol buttons and backside buttons towards events. The events are the activities that the player can carry out during the process of playing FIFA. There are two different states the player can be in during the process of playing FIFA. Namely, the player can be in possession of the ball and thus, in the attacking state or the player is not in possession of the ball and thus, the player is in the defending state. A list of possible activities in each state is given in Table 6 and an example of an in-game FIFA event log is given in Table 7.
A part for improvement is that there are currently limitations on finding the specific skill moves used (a more advanced form of dribbling), so these moves are gathered under the umbrella name “Skill move”. This process prepares the raw FIFA data so that in the next section, the data can be analyzed.

Table 6: Player possible activities

Attacking	Defending
Strafe Dribble (Lock Face Angle)	Goalkeeper Cross Intercept
Low Shot/Downward Header	Instant Hard Tackle
Whipped Cross	Switch Player (Manual)

Chip Shot	Running Jockey
Lobbed Through Pass	Jockey/Grab & Hold
Pass and Go	Shoulder Challenge/Seal out
High Lob / High Cross	Change Player
Strafe Dribble	Teammate Contain
Finesse Shot	Rush Goalkeeper Out
Driven Ground Pass	Tackle/Push or Pull (When Chasing)
Fake Shot	Hard Tackle
Fake Pass	Sliding Tackle
Lofted Through Pass	
Timed Shot	
Lofted Ground Pass	
Driven Ground Cross	
Ground Cross	
Threaded Through Pass	
Slow Dribble	
Face Up Dribbling	
First Touch/Knock-On	
Shield/Jockey	
Protect Ball	
Stop Ball	
Skill Move	
Through Pass	
Shoot/Volley/Header	
Ground Pass/Header	
Lob Pass/Cross/Header	

Table 7: Example of list of in-game actions

Action	Time
Ground Pass/Header	10:49:03 AM
Balanced	10:49:04 AM
Attacking	10:49:05 AM
Ground Pass/Header	10:49:05 AM
Through Pass	10:49:06 AM
Stop Ball	10:49:07 AM
Fake Shot	10:49:12 AM
Protect Ball	10:49:12 AM
Slow Dribble	10:49:30 AM
Slow Dribble	10:49:30 AM
Slow Dribble	10:49:46 AM
Shoot/Volley/Header	10:49:47 AM

3.2 Possible techniques and the chosen technique

In this section, different techniques to analyze the player input data are discussed. In 2.1.4, different BPM methods to analyze processes were introduced. The raw FIFA controller data is the input data for this research. In 3.1, the raw input data is converted into a list of activities. With this step, the data is prepared for analysis. The prepared input data can be defined as execution logs from “business processes” or in this case execution logs from the process of playing FIFA (Weske, 2012).

Process discovery algorithms can be rated based on 4 quality dimensions, replay fitness, simplicity, precision, and generalization (Buijs, Van Dongen, & Van der Aalst, 2012). Different situations ask for different distributions of these quality dimensions. The situation of FIFA data can be described as the analysis of real-life event logs because the logs are gathered from real-life players which makes it real-life data. Moreover, real-life data has a chance to be incomplete, noisy, and/or inconsistent (S. Zhang, C. Zhang & Yang, 2003). Based on a multi-dimensional quality assessment of state-of-the-art process discovery algorithms, De Weerdt, De Backer, Vanthienen, & Baesens (2012) concluded that the HeuristicsMiner was the most suitable process discovery algorithm to use when working with real-life event logs and combating the noise real-life event logs contain.

To conclude, the raw data is prepared to be converted into a list of in-game events (Table 7). Together with the HeuristicsMiner, the process models can be generated. Though, key events are needed and those are generated in Chapter 4.

Chapter 4: Key events

After preparing the data for analysis and determining which process discovery algorithm to use on the data for the analysis, the question should be asked what to extract from the data. Referring to 2.4.2, point 2 of the best practice is to ask concrete questions about the data. Different process models can be discovered, depending on the concreteness of the questions asked. In 4.1, the question is asked what the key events are. These key events are then used to decide how the data is analyzed.

4.1 Identifying key events

To be more specific, these key events influence how the data is divided and analyzed. An example within a football context, if the focus lies on how good a player is at passing, then the key events should be passes and the focus of the analysis should be on how accurate the player is passing the ball. Taking key events from classical business processes is not fitting in this context because the events are football-related. Because of this, the focus lies on key events in a football context. The question “What are typical FIFA key events?” is asked to make an initial list of possible key events for the data analysis. Afterwards, the key events of the initial list are discussed on feasibility.

To find typical FIFA key events, two approaches are taken. The first approach is to find statistical studies with a focus on real-life football. This approach is taken because FIFA is a football simulation game which means that it aims to simulate real-life football. Hence the choice for real-life football studies. The second approach is to collect key events that the game FIFA 20 records, after a game of FIFA has finished, the game presents an overview of the collected statistics in a screen called “match facts”. Because there are variables mentioned more than once from the different studies, a list of unique variables is shown gathered under a keyword such as goal or passing. In Appendix B, the complete list containing the variables mentioned per individual study is shown. Before the table is shown, a numbered list of studies is shown to refer to the mentioned numbers in 2nd column of Table 8, which refers to the studies. Following that list of sources, the key variable table is shown.

1. Gómez, M. A., Gómez-Lopez, M., Lago, C., & Sampaio, J. (2012). Effects of game location and final outcome on game-related statistics in each zone of the pitch in professional football. European Journal of Sport Science, 12(5), 393-398.
2. Moura, F. A., Martins, L. E. B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. Journal of sports sciences, 32(20), 1881-1887.
3. Liu, H., Hopkins, W., Gómez, A. M., & Molinuevo, S. J. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. International Journal of Performance Analysis in Sport, 13(3), 803-821.
4. Brito Souza, D., López-Del Campo, R., Blanco-Pita, H., Resta, R., & Del Coso, J. (2019). A new paradigm to understand success in professional football: analysis of match statistics in LaLiga for 8 complete seasons. International Journal of performance analysis in sport, 19(4), 543-555.

5. García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A., & Luengo-Sánchez, S. (2021). In-game behaviour analysis of football players using machine learning techniques based on player statistics. International Journal of Sports Science & Coaching, 16(1), 148-157.
6. The statistics screen at the end of a FIFA match

Table 8: Simplified list of potential key variables

Variables	Source(s)
Goals Made, against, own goals, first touch goals	1 2 4 5 6
Shots On goal, shooting accuracy, shots conceded, effectiveness against conceded shooting, big chances, chances missed, missed shots, weak shots	1 2 3 4 5 6
Fouls Committed, suffered, yellow cards, red cards, penalty kick conceded,	1 2 3 4 6
Turnovers Dispossessed, turnovers	1 3 4 5
Ball recoveries Ball recoveries, interceptions, recovery, failed interceptions, blocked passes, failed to block, total interceptions	1 3 4 5
Passing Crosses, assist, key pass, pass, through pall, successful pass, passing accuracy, assist, expected assist, pass verticality, second assist, back passes, pullbacks, build up play, front passes, long balls, launches, pre-shoot pass, passes ratio, errors, switches of play, total passes, unsuccessful passes, playing time with ball possession, percentage ball possession	1 2 3 4 5 6
Set pieces Corner kicks for, corner kick against, free kicks to goal, penalty faced, penalty kick, free kick goal, free kick goals received	2 4
Offside	2 3 4 5 6

Offside, offsides provoked	
Other defensive variables Block, challenge, clearance, tackle, aerials lost, aerials won, aerials ratio, tackles won with possession, tackles won without possession, total aerials, total tackles	3 5
Other attacking variables Dribble, take on lost, take on won, take on ratio, take on total, good skills, individual plays	3 5
Goal keeper variables Catch, collected ball, cross not claimed, drop, goalkeeper kick from hands, goalkeeper launch, goalkeeper throw, keeper sweeper, penalty faced, punch, save, smother	3
Injuries	6

The list contains several basic variables such as “goals” and “shots “which are mentioned in both regular studies and in the FIFA match facts, but also more advanced variables such as “build up play” which are less known than the previously mentioned variables. Now that a list of possible variables is created, the feasibility of choosing the variables is discussed. The research has several limiting factors that influence which variables are chosen. The first limiting factor is the limited timespan of the research. If there was more time available for research, every unique variable could be chosen but because there is a limitation on the duration, the selection of variables should be feasible to use within the research duration. Because this is an initial exploratory study on how process mining could be used within a FIFA context, a part of the research duration has already been spent on how FIFA processes connect with business processes. Moreover, a second limiting factor is that certain variables require more advanced tools to record them. For example, according to Van der Aalst et al. (2011), different maturity levels can be described for event logs with the aim to achieve the level of 5 stars which means that the event log is trustworthy and complete. The events should be recorded in an automatic, systematic, reliable, and safe manner. There are currently no tools available to record regular or advanced variables such as determining whether a shot was on target or not in the game. This could connect with the first limiting factor that there is not enough time to develop these tools. If the “goal” variable would be chosen, the variable is collected by hand from the game. To collect these variables automatically would require more time to develop these tools. However, these variables could be variables of interest for further research.

Following this discussion, the choice is made to do the analysis with a list of basic variables that ought to be feasible to use within the predetermined time frame of the research. The choice was made to collect the variables “goal” and “no goal”, because the event log contains when a shot is made. The goal of collecting these two variables is to gain insight into if the sequence of

actions differs between when a “goal” is made after a shot and when “no goal” is made after a shot. These variables are added in the in-game event log. The variables are added after the in-game events of when the ball is shot with the aim to score a goal, which are the events “Low Shot/Downward Header”, “Chip Shot”, “Finesse Shot” and “Timed Shot” from Table 6. In summary, after a shot type, the key variable “goal”/”no goal” is added depending on the result, the sequence has ended and a new sequence is started.

The second choice was made to collect the variable “possession”, which could be either TRUE or FALSE depending on if the player is in possession of the ball or not. This variable was chosen because it is a necessary variable to convert the raw button input to the event log.The “possession” variable is connected to the time so that the button input converter knows when the player is in possession of the ball and when the player has lost possession of the ball and thus, is in the defending state. The button input can lead to two different actions depending on if the player is in possession of the ball. This variable is added to the raw button input and not to the in-game event log because the raw button input needs the possession value to correctly translate the button input to the right in-game state with their respective activities (Table 6), attacking state which corresponds with attacking moves or defending state which corresponds with defending moves.

Table 9: Variables

Variable	**Description**
Goal	Added if a goal is scored after a shot
No goal	Added if no goal is scored after a shot
Possession	TRUE or FALSE based on whether the player is in possession of the ball or not

4.2 Identifying patterns

In this section, it is explained how the log will look after the chosen variables have been added. Then the process of generating the process models is explained, this process will contain the steps that are taken to identify patterns within the data. The steps start with the log with chosen variables, the tool used to process the log, how the data will be processed in the tool, and how the model will look. The goal is to show at the end of this section, that it is possible to identify patterns from the data.

After taking the chosen variables (4.1) into account the log will contain the same structure with in-game activities but the chosen variables “goal” and “no goal” and the time are added. These variables can be added because this is noted during the match by hand. Moreover, the possession value and time are also noted down during the match by hand. For further research, it is recommended to automate this process. Referring back to Table 1, an extra column is

added for the “possession” variable. The new table with button input data and the “possession” variable is shown in Table 10.

Table 10: Example of button input with “possession” variable column added

Type	SpecifiedButton	Pressed	TimePressed	Possession
Key	BTN_SOUTH	1	10:49:03	TRUE
Absolute	JOYSTICK_LEFT	LEFT	10:49:03	TRUE
Key	BTN_SOUTH	0	10:49:03	TRUE
Absolute	JOYSTICK_LEFT	CENTER	10:49:04	TRUE
Absolute	DPAD_RIGHT	1	10:49:04	TRUE
Absolute	DPAD_RIGHT	0	10:49:04	TRUE
Absolute	DPAD_RIGHT	1	10:49:05	TRUE
Absolute	DPAD_RIGHT	0	10:49:05	TRUE
Key	BTN_WEST	1	10:49:05	TRUE

This table is used to convert the button input with the right context, so now it is known in which state (attacking or defending) the player is. With this information, the button input can now correctly be converted to in-game events.

Table 11: Example button input converted to in-game events

Action	TimeOfAction	Possession
Ground Pass/Header	10:49:03	TRUE
Balanced	10:49:04	TRUE
Attacking	10:49:05	TRUE
Shoot/Volley/Header	10:49:05	TRUE

After the example button input is translated with the right context, an in-game event log will look like Table 11. The “Action” column will give which in-game activity will be performed. The “TimeOfAction” column will give the time value. The “Possession” value will return if the player was in possession or not and it is used in the next step where the variables “goal” and “no goal” are added in the log. The “possession” value helps because if a player is not in possession of

the ball, the player cannot score the ball so adding the “goal”/”no goal” variable to the log will cost less time because this process is done by hand. When this process is automated, the difference will likely be less significant. The new log will look exactly the same as Table 11, but the events “goal”/”no goal” are added after the shot activities “Low Shot/Downward Header”, “Chip Shot”, “Finesse Shot” and “Timed Shot”. The “TimeOfAction” column will have the same time as the row above. Moreover, the “Possession” column will have the same value as the row above.

Table 12: Example in-game events with “goal”

Action	**TimeOfAction**	**Possession**
Ground Pass/Header	10:49:03	TRUE
Balanced	10:49:04	TRUE
Attacking	10:49:05	TRUE
Shoot/Volley/Header	10:49:05	TRUE
Goal	10:49:05	TRUE

After this process, the log now contains sequences of attacking and defending activities. Moreover, the attacking sequences have the additional information when a shot has led to a “goal” or “no goal”. With these logs, the modeling can start.
The tool used for modeling is ProM 6.10, ProM is a framework that allows for different process mining techniques to be applied in the form of plugins. To be more specific, ProM allows for implementation of process mining algorithms in a standard environment (Van Dongen, de Medeiros, Verbeek, Weijters & van Der Aalst, 2005). Since there are so many different plugins that can perform a fast array of different process mining algorithms, the following examples are given with either no selected plugin if it entails the generic user interface or with the chosen process discovery algorithm, the HeuristicsMiner.

The research will now go into the ProM settings that were used while applying the HeuristicsMiner and go further into decisions taken to generate the data. In Appendix C, an overview of the general steps in ProM is shown.

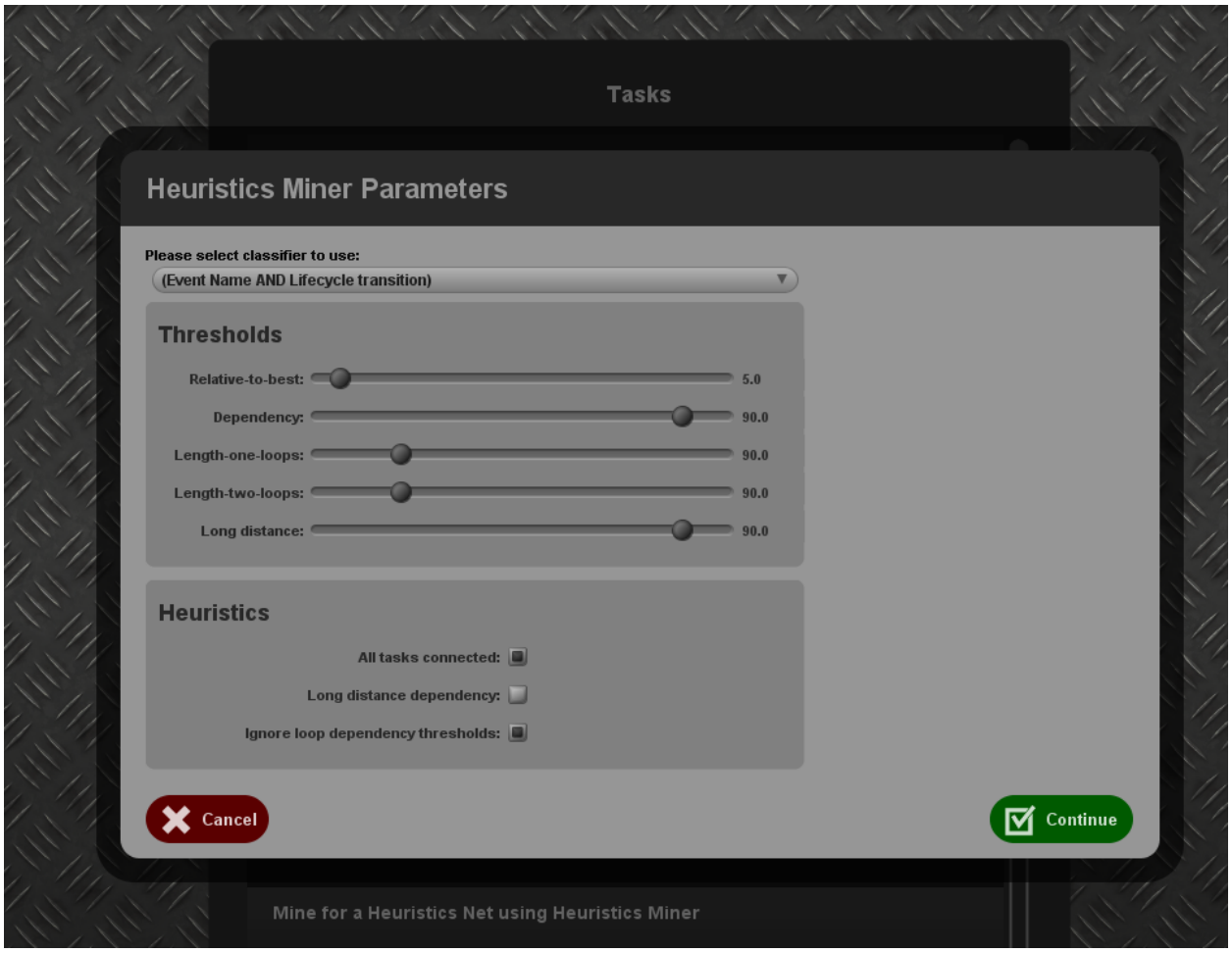


Figure 18: Options tab to select parameters

In tab, the classifier dropdown menu gives the following options:

- Event name and Lifecycle transition
- Event name
- Lifecycle transition
- Resource

The choice from the dropdown menu is the default option, the “Event name and Lifecycle transition” option. The “Threshold” part gives slider options, these are also kept on their default values. The “Heuristics” part is also kept on default values. A point for future research is to look further into the dependency value and the FIFA data and to research what the most fitting dependency value would be.

The data that is now available is the in-game events log with “goal”/”no goal” added (Table 12). To analyze the frequency of sequences, the data is split based on if the sequence of activities led to a goal or no goal. The result of this will be two different data files containing sequences of activities. These two data files are prepared before the analysis with ProM, so there will be one data file with only sequences that contain the “goal” variable and the other file will contain

sequences with the “no goal” variable. Then each of these data files are put separately into ProM and the HeuristicsMiner is applied to each of them with the above-mentioned settings in Figure 18.

To perform an analysis, button input data was needed. To generate a set of button input data, a test match was run at the eSportslab. During this test match, the “goal” variable times were noted down. By knowing when the “goal” variable happens, the “no goal” variables can be deduced because a “goal”/”no goal” variable will always be preceded by a shot activity, “Low Shot/Downward Header”, “Chip Shot”, “Finesse Shot” and “Timed Shot”. Moreover, the times when ball possession was won or lost were noted down, with these times the correct state (attacking or defending) can be given to the raw button input. Following this, the data set can be split into two files. Namely, part of the data set in which the player has possession of the ball and is attacking, and the other part of the set where the player is defending.

For this research, the key variables chosen “goal”/”no goal” are both from the attacking state. As a consequence of the chosen key variables, the analysis will be only done on the attacking state of the test match. Though, the assumption is made that if the analysis works for the attacking state key variables, analysis for defending state key variables will also work. In the next step, two files will be compared, the first file will have the attacking activities when the key variable “goal” occurs and the other file will have the attacking activities when the key variable “no goal” occurs. Then these two files were put in separately into ProM, the HeuristicsMiner was selected with the default settings. The results are two process models and additionally, ProM offers a short summary of the data which is also shown. In Figure 19, the process model is given for the “no goal” data set. Following the process model, a summary of the “no goal” data set is given in Figure 20. Then in Figure 21, the process model is given for the “goal” data set. At last, in Figure 22, a summary is given of the “goal” data set. In the process models in Figures 19 and 21, the activities listed within the rectangles are from the set activities listed in Table 6.

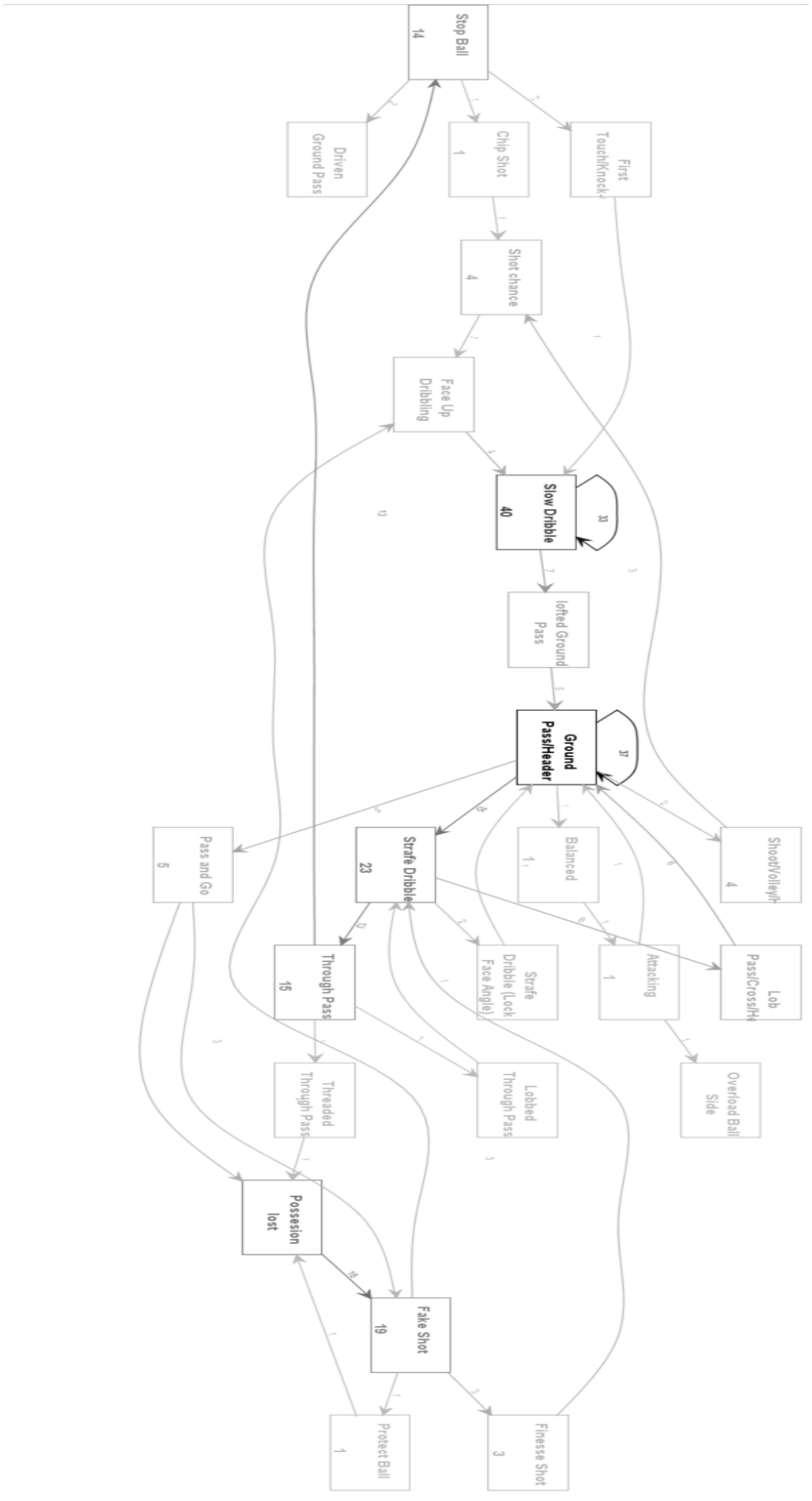


Figure 19: “no goal” test match process model

Event classes defined by Event Name
All events

Total number of classes: **24**

Class	Occurrences (absolute)	Occurrences (relative)
Ground Pass/Header	60	24.793%
Slow Dribble	40	16.529%
Possesion lost	25	10.331%
Strafe Dribble	23	9.504%
Fake Shot	19	7.851%
Through Pass	15	6.198%
Stop Ball	14	5.785%
lofted Ground Pass	7	2.893%
Lob Pass/Cross/Header	6	2.479%
Pass and Go	5	2.066%
Shoot/Volley/Header	4	1.653%
Face Up Dribbling	4	1.653%
Shot chance	4	1.653%
Finesse Shot	3	1.24%
First Touch/Knock-On	2	0.826%
Strafe Dribble (Lock Face Angle)	2	0.826%
Driven Ground Pass	2	0.826%
Overload Ball Side	1	0.413%
Threaded Through Pass	1	0.413%
Lobbed Through Pass	1	0.413%
Chip Shot	1	0.413%
Balanced	1	0.413%
Protect Ball	1	0.413%
Attacking	1	0.413%

Start events

Total number of classes: **1**

Class	Occurrences (absolute)	Occurrences (relative)
Ground Pass/Header	1	100.0%

End events

Total number of classes: **1**

Class	Occurrences (absolute)	Occurrences (relative)
Strafe Dribble	1	100.0%

Figure 20: “no goal” data summary

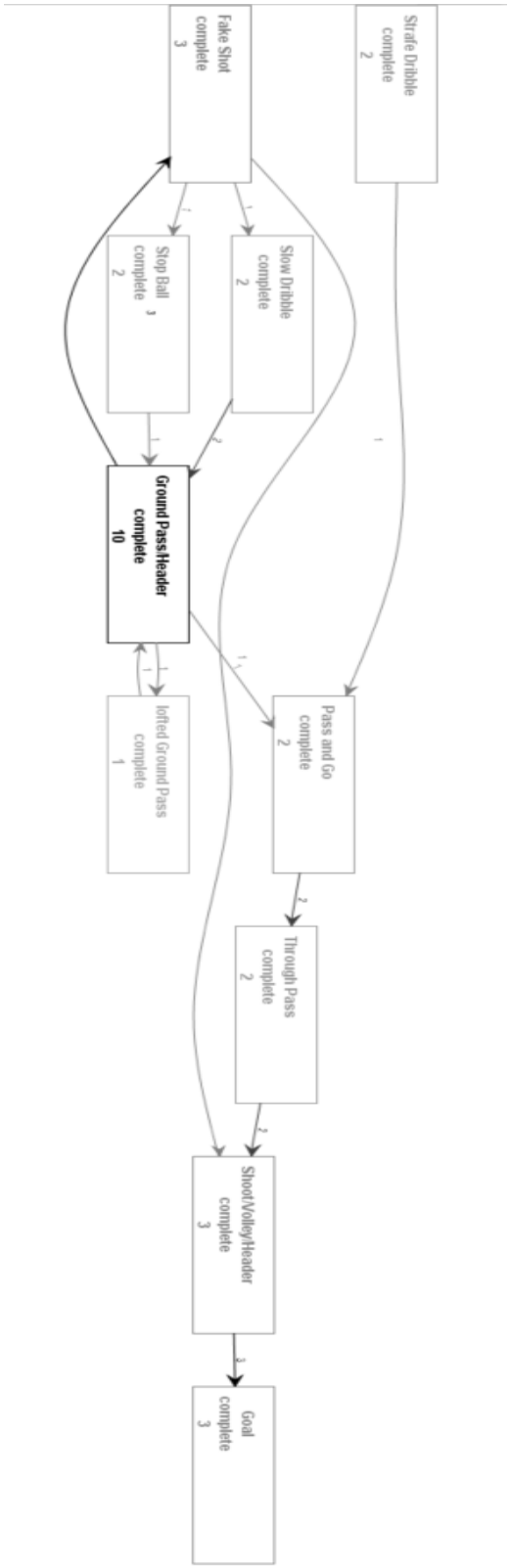


Figure 21: “goal” test match process model

All events

Total number of classes: **10**

Class	Occurrences (absolute)	Occurrences (relative)
Ground Pass/Header	10	33.333%
Goal	3	10.0%
Fake Shot	3	10.0%
Shoot/Volley/Header	3	10.0%
Stop Ball	2	6.667%
Slow Dribble	2	6.667%
Through Pass	2	6.667%
Pass and Go	2	6.667%
Strafe Dribble	2	6.667%
lofted Ground Pass	1	3.333%

Start events

Total number of classes: **1**

Class	Occurrences (absolute)	Occurrences (relative)
Strafe Dribble	1	100.0%

Figure 22: “goal” data summary

Discussion of the models

In this section, the process models will be discussed and checked if they conform to the conform flow patterns summarized in Table 2. Moreover, the process models (Figures 19 and 21) are checked if they conform to the data visualization best practices explained in 2.4.1. Another point of discussion are the key variable data summaries. While the test match does not give enough data to make conclusions about which events give a more likely outcome to make a goal, the aim is to discuss the added value of the data summary and process models, and how both can help with giving insight into the gameplay. A larger set of data would allow for more conclusive results, the question is how large the data set needs to be and that would be a point for further research.

From the two process models, the process model from the key variable “goal” is visually more simple than the process model from the key variable “no goal”. The “goal” model is more simple because there is less data with the key variable “goal” than the key variable “no goal”. While the “no goal” variable is more complex, it is still readable. The process models contain no complex control flow patterns and the control flow patterns used are the sequence pattern and the exclusive split pattern. Another factor that influences the readability of the process models is the level of spaghetti-likeness of the model and severe spaghetti-likeness of the process models should be avoided (Van der Aalst, 2011; Schwabish, 2021) because it will make models visually complex and hard to read.

The data summary has three columns named “Class”, “Occurrences (absolute)”, and “Occurrences (relative)”. These three columns can add value to the analysis through comparison between data sets. For example, in the situation of the test match, the relative frequency of certain “Class” events can be compared by looking at the “Occurrences (relative)” column. Visually when the model becomes more complex and has more events flowing from one to another such as the process model of the test match with “no goal” (Figure 19), the numbers or frequencies how often one event results in another become smaller and harder to read. When the models are within the ProM environment, the resolution of the process models is higher than when the models are exported and shown in for example this document. This is a point of improvement for the future, to find a way to export high resolution process models from ProM or to find a different process mining tool that allows for high-resolution export to external sources.

To add to the discussion, the produced process models and data summaries are compared to different categories of data visualization best practices (2.4.1). This table will consist of the categories from the data visualization best practices (Table 4) with an extra column where the process models and data summaries are discussed.

Table 13: Discussion of findings and best practices

Category	Findings
Audience	- Process models The main audience for the process models are eSporters, eSport coaches and the eSportslab. In the future, feedback could be collected from the audience on when the process models become too complex. - Data summaries The data summaries contain the in-game events which the main audience of the dashboard of the eSportslab should be familiar with. The data summary contains numerical values which can be transferred to the dashboard.
Process	- Process models While ProM is the chosen tool, in the future there might be more fitting tools available. The choice of process mining tools should depend on if the tool can perform process mining the fastest and if the payoff of using the tool is larger than the cost of learning to use the tool. - Data summaries The data summaries are a result of using ProM. While ProM is used to generate these data summaries, it is also possible to generate data summaries through other means. The choice will again depend on if the tool can generate the data summaries faster and if the payoff of using the tool is larger than the cost of learning to use the tool.
Color usage	- Process models When the process models are generated, they do not contain colors. While the graph starts in the color scheme containing black and white, using other color schemes could be explored in the future such as sequential, diverging or qualitative color schemes. An example of color schemes could be taken from the existing color scheme that is used in the current dashboard of the eSportslab. - Data summaries The data summaries contain the same color scheme as the ProM user interface. In the future, the data summaries numerical values could be extracted and then the chosen color schemes could be added.
Structure	- Process models The process models are generated by the tool used to generate them, ProM, and thus, the layout is chosen by ProM. While the layout inside the process model is chosen by ProM, the possible position of the process model within the dashboard of the

	eSportslab is yet to be chosen, so the structure data visualization best practices can be used in that aspect. In Chapter 5, the findings are poositioned in the existing dashboard. - Data summaries The data summaries are also generated and structured by ProM. While the initial structure of the data summary works, in the future, the layout or data shown could be adapted depending on the needs shown by for example the audience of the dashboard.
Data	- Process models Through the use of the HeuristicsMiner, the research argues that the most fitting process discovery algorithm is used to visualize the data (2.1.5). By choosing this algorithm, the assumption is made that the process model contains the most relevant events. Within the dashboard, it should be explained how to read and interpret the process model. Explaining raw data is easier than explaining a model , so the explanation of the model should be more extensive to make it clear what the model aims to convey. - Data summaries In the content of the data summaries are no uncertainties displayed because they contain all the events in absolute values and in relative values.
Visualization	- Process models By choosing the HeuristicsMiner, the assumption is made that unnecessary clutter is removed. The test match process model is not complex. In the future when larger data sets are used, the process model could become more complex. The process model itself in the test match form is readable and not too complex. When certain in-game events are more frequent, ProM will make them darker than an in-game event that happens less frequently. - Data summaries The columns of the data summaries might not give enough information to the audience of the dashboard. To prevent this, the columns should have an extensive description of what data is shown within the column.

4.3 Summary of the steps

In this section the process from start to finish will be summarized in a more general way and how the general steps were applied in this research to go from the input to the findings. This section aims to give a generalized overview of the different steps taken to enable process mining on FIFA data. These generalized steps can be used for future analysis on FIFA data with other key variables.

1. Determine which key events should be available for analysis, this will impact the information that is going to be collected during the matches: Depending on the key events chosen, extra information needs to be gathered during the match.

   How is this applied in the research: In 4.2 key variables “goal”/”no goal” and “possession” were chosen. While the “possession” variable is necessary for the conversion of raw button data, the “goal”/”no goal” variables were chosen as the key variables in this test match. These variables influenced the preparation of the data because in addition to the button data, the instances when a goal occurred were noted down. To conclude, different key variables require additional information that needs to be gathered during the match. These variables need to be concretely defined and then the required information to define these variables needs to be collected during the match.

2. Preparing the input: Collect the raw button data. During the match, collect the additional information required to form the key variables.

   How this is applied in the research: During the match, the controller data was collected. In addition to the controller data, the key variables “goal”/”no goal” and “possession” needed to be collected. The “possession” variable is a necessary variable for the converter tool to convert the controller data to in-game events. During the match, the time was noted when possession was lost and won. With this information, the right state (attacking or defending) was assigned to the controller data. The key variables “goal”/”no goal” were collected through writing down when the player scores a goal. When this is known, the shot attempts coinciding with the time when a goal was scored can be labeled with the key variable “goal” and the other shot attempts can be labeled with the key variable “no goal”. In short, the times when possession switched were noted and the times when a goal was scored were noted.

3. Convert to events: The converter tool from the eSportslab takes in datasets equal to Table 10. With the collected raw button data and the possession values, the necessary table can be constructed. This table is then converted into an event log similar to Table 11. After the data has been prepared, add the key variables to the data.

   How is this applied in the research: In the test match, the key variables “goal” and “no goal” were added. During the test match, the time data of the goals scored was noted. After the data was converted to an event log, the key variables were added. The key

variables could only be added after a shot activity (“Low Shot/Downward Header”, “Chip Shot”, “Finesse Shot” and “Timed Shot”) because the shot attempt results in either a goal or no goal. In the data set, the key variables “goal” and “no goal” were added after the shot activity.

4. Filter based on key events: Depending on the key variables chosen in the analysis, the data set can be split into multiple data sets to compare against each other.

   How is this applied in the research: Because of the chosen key variables “goal”/”no goal”, the data set is first split into two files where one file contains only sequences of attacking events and the other file contains only sequences of defending events. Then the key variables should be applied, the key variables chosen belong solely to the attacking state. The file containing only attacking sequences is labeled with the key variables. After a shot attempt, the key variable “goal” was added in the case of a goal, and the key variable “no goal” was added in the case of no goal following the shot attempt. After this was done, this file with the key variable labels was split into two files, one containing the sequences before the key variable “goal” and the other containing the sequences before the key variable “no goal”.

5. Process the data in ProM: Import the datasets, and convert them with the HeuristicsMiner and default settings. The data summary and the process models are available.

   How is this applied in the research: The “goal” and “no goal” data sets were imported into Prom. The data sets were converted to process models with the HeuristicsMiner on default settings. Then the process models and the data summaries from ProM were available for analysis.

6. Performing the analysis: Interpret the process models and the data summaries. Compare the process models and data summaries between chosen key variables.

   How is this applied in the research: The process models are tested against the data visualization best practices from 2.4.1 with a focus on the spaghetti-likeness of the process model.

Chapter 5: Dashboard

In this chapter, the research questions (1.4) belonging to Chapter 5 will be answered. To answer the first research question “What do the findings from Chapter 4 tell us about the key events?”, the products from Chapter 4, the process models and data summaries will be interpreted and discussed how these could fit within a dashboard according to the data visualization best practices (2.4.1). To answer the second research question “What is the progress of dashboard development of the eSportslab at the moment of this research?”, the current dashboard of the eSportslab is shown and discussed. And at last, to answer the third research question “How can the insights gained from process mining optimally be displayed in the dashboard?”, the answers from the first and second research questions are combined to produce an advice on how to position the findings in to the dashboard.

5.1 Interpretation of key events

The goal of this section is to ease the fitting of the results within the dashboard by initially describing the results and how they are to be interpreted. In the discussion of the models and data summary, the size of the data set was mentioned as a factor that influences the ability to gain insight from the models. In the problem statement (1.2) of the research, the assumption was made that if a solution was found to apply process mining on a data set of a certain size, the process mining solution would work on data sets larger than the data set used in the solution. In future iterations, the larger data sets should lead to more conclusive results because the larger data set would cover the inconsistencies the smaller test match data set could contain. So larger data sets could be used as means to gain insight into the FIFA gameplay. Though, it is important to note that solely using the controller data to arrive at conclusions does not tell the whole story. These products, the process models and data summaries, should be used in combination with other information about the gameplay. As an example, the process models and data summaries tell information about which in-game events happened in which order but they do not tell where on the football field the event happened.

To find other information that can be used in combination with the process models and data summaries, other literature on football analysis and statistics is used. This literature is used because broadly, the main goal of football is to win the match and the main goal of FIFA is also to win the match. The literature will help to define what information could be of use in combination with the process models and data summaries. One of the more prominent statistics used within football analytics is the Expected Goals Method or xG. The next section aims to explain the xG and how it can be reproduced within FIFA.

In its simplest form, the Expected Goals Method (xG) strives to calculate the chances a team has to score and concede goals (Rathke, 2017) when the ball is shot by either team. The xG variable can range from 0 (no goal) to 1 (goal). To calculate xG, different factors are taken into account. An example set of factors used are location, distance, shot angle, and shot type (Riley, 2014). While many different xG models have been generated through the use of different factors, Caley has done the most research on xG (Rathke, 2017). While the methodologies used vary from study to study, the consistent factor is that he divides his pitch into different

zones (Caley 2013; Caley 2014a & Caley 2014b). This pitch division is combined with the fact that only Shots on Target (SoT) are used (Caley, 2013) instead of every shot. The pitch division and the SoT form the basis for the xG model of Caley. The pitch divisions used by Caley (2013;2014a & 2014b) are shown in Figure 23.

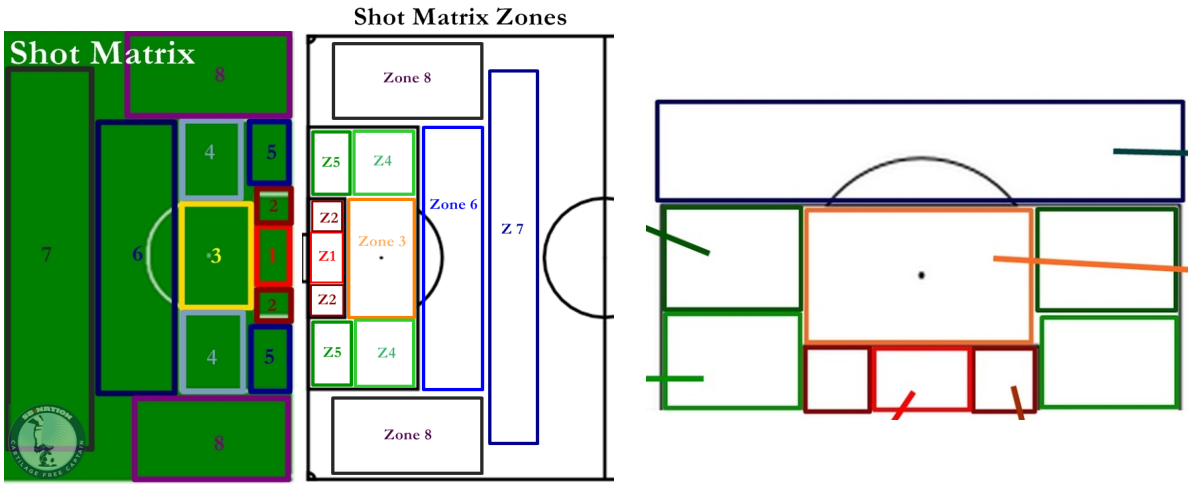


Figure 23: Different pitch divisions by Caley in order of 2013, 2014a, and 2014b

Through process mining on button data, the events such as shots can be determined. To work towards an xG model for FIFA, several factors are missing. The first factor that is missing is the pitch position of the ball at the time of the shot. The second factor that is missing is finding whether the shot was on target or not. Both factors are used in the xG model but are currently not available because there are no means yet to extract them from the game, unlike the button data. Suggestions for future development would be on finding ways to extract these factors, the pitch position at the time of shots and finding out if the shot was on target, from the game and to use these to create an xG model.

5.2 Current state of the dashboard

In this section, the current dashboard of the eSportslab is introduced. By doing this, the different elements of the dashboard are shown so that in the next section, the process models and data summaries can be fitted within the existing dashboard elements.

To explain the current situation of the dashboard of the eSportslab, firstly, the structure of the dashboard is explained. The dashboard contains different sections and each section has a specific functionality. The different sections and their functionalities are listed in Table 14. The functionalities are described from the eSporter's perspective.

Table 14: eSportslab dashboard sections and functionalities

Section	Functionality
Upload files	Video files (Youtube or Twitch) can be uploaded of the FIFA match that was played.

	The time the match was played at can also be added.
Add or edit match	Matches can be added manually or matches can be edited in.
Player overview	A general overview of the performance is given with additional statistics such as win rate and pass accuracy.
Stats analysis	Shows analysis of: - The performance of the team setup - The performance and time interval between matches - Statistics in games won, lost or drawn - Performance and time of day
Player analysis	Shows analysis of: - Individual players of the eSporter's team - Individual players of the opponent
Heatmap	Shows a heatmap of the eSporter's team, the opponent or both at the same time.
In-game analysis	Shows analysis of: - Minimap with positions of individual players of the eSporter throughout the game - Player positions can be analyzed during match events such as goals made and conceded.

Now that the dashboard sections and functionalities are known, the visual look of the different sections is shown to ensure that both the functionality and look of the dashboard sections are clear. In the following figures, screenshots are shown from each of the sections described in Table 14.

5.3 Displaying the findings in the dashboard

In this section, the choice is made on how to fit in the process models and data summaries in the dashboard. The focus will lie on making the process models and data summaries adhere to the data visualization best practices set out in 2.4.2. Initially, there is a decision between creating a separate section for the process models and data summaries or adding the process models and data summaries to an existing section. Afterwards, it is discussed how the process models and data summaries should be added to the dashboard.

5.3.1 Choosing between a new section or adding to an existing section

The choice between adding a new section for the findings of Chapter 4, the process models and data summaries, or adding the findings to an existing section depends on if the functionalities of the findings match with the functionalities of the existing sections. If there is a section that has similar functionality, then the findings will be added to that section. If there is no section with similar functionality, then a new section should be added to the dashboard.

To start, the functionality of the sections are already defined but the functionality of a possible dashboard with the process models and data summaries is not yet defined. Consequently, the functionality of the process models and data summaries is defined in this section. These functionalities are formulated with help of the theory about business processes and process models (2.1),the discussion of the process models and data summaries (4.2) and the interpretation of key events (5.1).

Table 15: Functionality of the process models and data summaries

Element	Functionality
Process model	Allows for visual representation of in-game sequence of actions filtered on key events. In the case of the test match, the eSporter can see which in-game events happen more frequently when they score a goal versus when they do not score a goal.
Data summary	Aids in helping the eSporter understand which in-game events happened. Also gives insight into the absolute and relative frequencies of in-game events filtered on key events.

5.3.2 Comparing functionality

After the functionalities of both the process models and data summaries have been described, the functionalities of both need to be compared to the functionalities of the existing sections to find if these sections have similar functionalities. After going through Table 14 and comparing the functionalities of the existing sections with the functionalities of the process models and data summaries, there are 3 possible sections for the process models and data summaries. In Table 16, it is explained why the process models and data summaries could add to the existing sections. Because of this conclusion, there is no need to add an extra section to the dashboard.

Table 16: Comparing the functionalities of the process models and data summaries with the functionalities of the existing sections

Section	Functionality	Comparison

Player overview	A general overview of the performance is given with additional statistics such as win rate and pass accuracy.	Process models will allow for insight into what sequences of in-game events happened in the match(es). Data summaries will add to the existing statistics with the absolute and relative number of events.
Stats analysis	Shows analysis of: - The performance of the team setup - The performance and time interval between matches - Statistics in games won, lost or drawn - Performance and time of day	Process models will allow for insight into what sequences of in-game events happened. An example option is to filter the button data so that process models based on team setup, the time interval between matches, and the time of match can be generated. Data summaries will add to the existing statistics with the absolute and relative number of events. Because the data summaries are produced at the same time as the process models, the example option of generated data summaries based on team setup, team interval between matches, and time of match could add insight.
In game analysis	Shows analysis of: - Minimap with positions of individual players of the eSporter throughout the game - Player positions can be analyzed during match events such as goals made and conceded.	Process models will allow for insight into what sequences of in-game events happened. An example option is to show the in-game events that happened throughout the match in combination with the player position. Data summaries will add to the existing statistics with the absolute and relative number of events. An example is to show the absolute and relative number of events that happened before the match event occurred. To do this, the button data of a certain amount of time before the selected event could be collected and then used for the data summary.

To conclude, the process models and data summaries could be added to the above mentioned sections of the dashboard of the eSportslab. While the process models and data summaries could be used on their own, using them in combination will allow for synergies between existing sections and the added process models and data summaries.

5.3.3 Visualizing the process models and data summaries within the dashboard

In this section, the process models and data summaries are realized within the existing sections “Player overview”, “Stats analysis”, and “In game analysis” of the dashboard of the eSportslab. Before the process models and data summaries will be realized, data visualization best practices (2.4.1) and the discussion of the visualization best practices with the process models and data summaries (Table 13) will be used to form the basis for the realization of the process models and data summaries within the dashboard.

Within the discussion of Table 13, the data visualization best practices from 2.4.1 have been used as a point of reference to discuss the process models and data summaries. During this discussion, the visualization best practices were divided into different categories. For the visualization of both the process models and data summaries all categories could be mentioned but the choice is made to focus on the categories audience, color usage, and visualization. The categories that are left out are process, structure, and data. The reasons to not focus the mentioned categories is shortly explained below:

- The process category is not chosen as a focus because both the process models and data summaries have been generated by ProM. ProM was the chosen process mining tool that could deliver both the data summary and the process model. Though, an argument to switch from process mining tool could be faster process time or when the payoff of learning the new tool is larger than the cost of learning to use the tool.
- The structure category is not chosen as a focus because both the process models and data summaries have been generated by ProM. Consequently, the structure of both has been decided by ProM.
- The data category is not chosen as a focus because through the use of the HeuristicsMiner, it is argued that the HeuristicsMiner is the most fitting process discovery algorithm to visualize the data in process models. Though, within the dashboard, it should be explained how to read and interpret the process model.

In the next section, it is discussed how the remaining best practice categories are used to visualize the process models and data summaries into the dashboard sections. Previously in Table 16, the role of process models and data summaries is discussed per section. In these sections, the process models and data summaries should be visualized with help of the data visualization best practices with focus on the best practice categories audience, color usage, and visualization. In Table 17, it is discussed how the process models and data summaries should be visualized into the dashboard with help of the best practice categories audience, color usage, and visualization.

Table 17: Discussion of focus categories and visualization

Category	Discussion
Audience	The process models and data summaries should be visualized in the

	dashboard with the user in mind. The dashboard will predominantly be used by eSporters and other actors that would want to analyze the FIFA gameplay. While the initial visualization is done without the feedback from external parties due to time constraints of this research, in the future, it is advised to seek external feedback from dashboard users about the visualization of the process models and data summaries, and the dashboard visualization in general.
Color usage	The process models as well as the data summaries are generated in gray color schemes. The gray color scheme of the data summary is static, the grayness does not change depending on the numerical values in the data summary while the grayness of connections in the process model change depending on the relative frequency. Changing color schemes is an option but initially the process models and data summaries are produced in a gray color scheme by ProM. Future options for other color schemes could be for example a color scheme based on the club of the eSporter or a color scheme depending on the previously used dashboard colors.
Visualization	The process models are more complex to read for readers without prior knowledge of process mining. To counteract this, an explainer should be added to the process model to shortly explain what is being shown to the reader. An infographic is not used because the process models and data summaries are going to be part of the existing dashboard. Fitting an infographic into the dashboard with the other elements The process models were generated by the most fitting algorithm to produce a model without clutter. The process models contain the event labels in the model, the frequency of events is also labeled in the model itself. The data summaries have labeled columns. Though using active titles is recommended in the best practice to tell the reader what should be taken away from the graph, the dashboard content changes depending on the data used. For example, the user could use data of one match or ten matches, because this difference in data could lead to different conclusions and consequently, different takeaways from the graph, the choice is made to not use active titles. Adding explainers will be a goal for fitting the process models because it could help readers with no or less prior knowledge understand what the process model conveys. For the data summary, an explainer is added to explain the labels of the columns. When event frequencies are higher in the process models, the elements are more visually distinct in the standard gray color scheme.

In the following section, screenshots are shown of how the process models and data summaries are placed within the existing dashboard sections “Player overview”, “Stats analysis” and “In game analysis”. Moreover, explanation is added of what was done to position the process models and data summaries based on the discussion of visualization best practices in Table 17.

5.3.4 Positioning the process models and data summary

In this section, the research connects the existing elements of the dashboard with the functionality comparisons (5.3.2) and with the discussion around fitting the process models and data summary within the dashboard (5.3.3). Then it is stated where to position the findings in the dashboard.

1. Player overview explanation
   The “Player overview” section has the following elements:
   a. Overall data, where statistics are shown. The statistics shown are “Games played”, “Wins”, “Losses”, “Winrate” (which is calculated by dividing total wins by the total amount of games played), “Avg gametime” (the average duration of a game) and “Avg rest” (the average time between games).
   b. Average statistics, where the averaged statistics per game are shown. The statistics shown are “Goals”, “Goals conceded”, “Shots”, “Shot accuracy”, “Assist”, “Pass accuracy” and “Possession”.
   c. Player info, where general player information is shown such as name, surname, organization and year born.
   d. Performance trend, a graph where the performance trend of the player is shown. On the X-axis, it goes from 0 and goes to the total number of games. On the Y-axis, the performance trend starts from 0 and has a maximum of 1. The formula for the performance trend at a certain moment is dividing the total number of wins by the total number of games played at that moment.

   It is chosen to place the process models under the existing elements. The data summary can also be placed under the existing elements or the data summary can be added to the existing statistical elements “Overall data” or “Average statistics”.

2. Stats analysis explanation
   The “Stats analysis” section has the following elements:
   a. Best setup, where the best team formation details or the formation tactics that lead to the most wins are shown, the best resting period between matches is shown and the best playstyle or playing tactics that lead to the most wins are shown.
   b. Bar chart, where the average statistics are shown in wins and losses. The statistics shown are “Pass accuracy” (passing accuracy), “Shot accuracy” (shooting accuracy), “Possession”, “Poss won” (how often possession was won relatively) and “Poss lost” (how often possession was lost relatively) .

c. Line chart, where the win rate is shown per section of resting period. The sections of resting period are “<1 min” (under 1 minute), “1-3 min” (from 1 to 3 minutes), “3-5 min” (from 3 to 5 minutes) and “>5 min” (over 5 minutes).
d. Bar chart, where the win rate per section of the day is shown. The different sections of day are “Morning”, “Afternoon”, “Evening” and “Night”.

It is chosen to place the process models under the existing elements. The data summary can also be placed under the existing elements, this is because there are no existing statistical elements in the section.

3. In game analysis explanation
The “In game analysis” section has the following elements:
    a. Game details, where a game can be selected and then analyzed. The amount of time rested or “Rest period” is shown in minutes. The time per game or “Game time” is shown in minutes. The game mode the match was played in or the “Type game” is shown because, within FIFA, there are several different game modes a match can be played in. The video source of the gameplay or “YT/Twitch link” is shown. The formation tactic or “Formation insight” is shown. The playing tactic or “Playing style” is shown.
    b. Line chart, where the goal difference over time per match is shown. On the X-axis, the time is shown starting from 0 and going until the match is over which can vary from match to match. On the Y-axis, the net goal difference is shown. For example, if the score is 3-1, the net goal difference is 2. If the score is 0-2, the net goal difference is -2. The user is able to select an individual match to analyze through the dropdown menu.
    c. Minimap, where the positions of the individual players during the match are shown.
    d. Gameplay video, where the full video of the match is shown. This match is used for the analysis in the “In game analysis” section.

It is chosen to place the process models under the existing elements. The data summary can also be added under the existing elements, this is because there are no existing statistical elements in the section.

Chapter 6: Conclusion, Discussion, and Recommendations

In this chapter, the thesis is concluded. Firstly, the problem flow is restated and the conclusions are made. Secondly, the results are discussed and the recommendations for future research are proposed.

6.1 Conclusion

This research was started because the eSportslab found that the professionalization of the FIFA eSports scene was hampered. After researching this problem, it was concluded that the in-game data was not available for the eSportslab. To solve the problem “In-game data is not available”, the collected controller data is used as a starting point and the goal is to create a proof of concept method to analyze the collected controller data. The method should contain a data preparation phase, a data analysis phase, and a data visualization phase.

From the data preparation and data analysis phase, the following step by step method was applied on the FIFA data.

1. Determine which key variables to analyse
2. Prepare the input and collect additional information if needed for the key variables
3. Convert the button input to events and add the key variables
4. Choose what to analyse. Depending on the chosen key variables comparison between data sets can be made. In this research, the key variables “goal” and “no goal” were used to create a data set only containing “goal” sequences and a data set only containing “no goal” sequences. These could then be compared to each other.
5. Process the data sets in ProM.
6. Perform the analysis. Compare the process models and data summaries of the different data sets.

The method was used to generate process models and a data summary, and these findings were then combined with the data visualization best practices to position them in the dashboard elements “Player overview”, “Stats analysis” and “In game analysis”.

6.2 Discussion and further research

One of the shortcomings of the research is that the generalized process (4.3) of going from raw data to in-game events and models contains steps that are not automated and require manual labor to complete. This research tried to show a proof of concept method that took controller input and produced insights through the process models and data summaries. The method works but the next step would be to automate the method steps that can be automated. Further optimizations for the process mining steps can be found in the mentioned process mining best practices (Van der Aalst et al., 2016).

Another shortcoming of the research is that the data set used is of a small size, this led to the research being unable to make conclusive statements and have insights as another side product of the process mining. Examples of insights that could be gained from a larger data set

would be that certain attacking in-game events lead to more goals or that certain defending in-game events led to fewer goals conceded. For future research, analyzing larger data sets could lead to more conclusive results.

At last, the research used the default settings in ProM while applying the HeuristicsMiner. There are threshold settings and heuristics settings, this research decided to keep these settings default because of the limited amount of FIFA data. For future research, this could be a point of interest to look further into.

6.3 Recommendations

After taking the conclusion and shortcomings into consideration, the following recommendations are made to the eSportslab.

- Automate (parts of) the proof of concept method, because the created method contains several steps which require manual labor to convert the list of controller data to a list of in-game data. This is part of improving the maturity of the event log which is mentioned in the process mining manifesto by Van der Aalst et al. (2016) in 2.3.1 as a guiding principle for process mining. Moreover, the known process mining challenges mentioned in the same section could serve as support to solve encountered problems.
- Using a larger data set to find conclusive insights, because the current data size is the size of one test match, the research could not deliver conclusive results. To deliver conclusive results, a larger data set is needed, the question is how large and that is a question for further research to solve.
- Research the dependency value used to create the process models and look further into finding the ideal dependency value for the FIFA event data. Moreover, ProM gives also other threshold and heuristics settings
- Research the possibility of creating and exporting high-resolution process models, the current environment where the process models are created is ProM. When the more complex models are shown in the current environment, the models are readable, but when the models are exported, the models become less readable. Further improvements can be made to the readability of the models.
- Create an expected goal model, because during the research it was noticed that knowing the sequences of actions is useful, the data could be used in an expected goal model to create one of the state-of-the-art analysis models within real-life football which is the expected goal model. The recommendation for further research is to connect the event data with the pitch position of the event data and to collect event data that contains the pitch position. This data can be used to create a FIFA expected goal model.

References

Bergenthum, R., Desel, J., Lorenz, R., & Mauser, S. (2007). Process Mining Based on Regions of Languages. *Lecture Notes in Computer Science*, 375–383. https://doi.org/10.1007/978-3-540-75183-0_27

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2306–2315. https://doi.org/10.1109/tvcg.2013.234

Bose, R. J. C., Mans, R. S., & van der Aalst, W. M. (2013). Wanna improve process mining results? *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 127–134. https://doi.org/10.1109/cidm.2013.6597227

Brandt, S., & Brandt, S. (1998). Data analysis. Springer-Verlag.

Brito Souza, D., López-Del Campo, R., Blanco-Pita, H., Resta, R., & del Coso, J. (2019). A new paradigm to understand success in professional football: analysis of match statistics in *LaLiga* for 8 complete seasons. *International Journal of Performance Analysis in Sport*, *19*(4), 543–555. https://doi.org/10.1080/24748668.2019.1632580

Burgess, M. (2020, February 5). *Nike has finally revealed the secrets of its 1:59 marathon shoe*. WIRED UK. Retrieved 5 February 2020, from https://www.wired.co.uk/article/nike-alphafly-eliud-kipchoge

Çakır, G. (2021, August 24). *How much money does Faker make? We break it down*. Dot Esports. Retrieved 12 March 2020, from https://dotesports.com/league-of-legends/news/faker-earnings-league-of-legends-14357

Caley, M. (2013, November 13). *Shot Matrix I: Shot Location and Expected Goals*. Cartilage Free Captain. Retrieved 26 May 2022, from https://cartilagefreecaptain.sbnation.com/2013/11/13/5098186/shot-matrix-i-shot-location-and-expected-goals

Caley, M. (2014a, February 4). *Shot Matrix International I: Shot Distribution in European Football*. Cartilage Free Captain. Retrieved 26 May 2022, from https://cartilagefreecaptain.sbnation.com/2014/2/4/5375492/shot-matrix-international-i-shot-distribution-in-european-football

Caley, M. (2014b, September 11). *Premier League projections, from the winners to the relegated clubs*. Cartilage Free Captain. Retrieved 26 May 2022, from https://cartilagefreecaptain.sbnation.com/2014/9/11/6131661/premier-league-projections-2014#methodology

de Weerdt, J., de Backer, M., Vanthienen, J., & Baesens, B. (2012). A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs. *Information Systems*, *37*(7), 654–676. https://doi.org/10.1016/j.is.2012.02.004

der Aalst, W. V. M. P. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes* (2011th ed.). Springer.

Dota 2 Prize Pool Tracker. (n.d.). Dota 2 Prize Pool Tracker. Retrieved 26 May 2022, from https://dota2.prizetrac.kr/

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2013). Fundamentals of Business Process Management. *Fundamentals of Business Process Management*. https://doi.org/10.1007/978-3-642-33143-5

Few, S. (2006). *Information Dashboard Design: The Effective Visual Communication of Data*. O'Reilly Media.

García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A., & Luengo-Sánchez, S. (2020). In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science & Coaching*, *16*(1), 148–157. https://doi.org/10.1177/1747954120959762

Georgakopoulos, D., Hornick, M., & Sheth, A. (1995). An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases*, *3*(2), 119–153. https://doi.org/10.1007/bf01277643

Global Games Market Report. (n.d.). Newzoo. Retrieved 26 May 2022, from https://resources.newzoo.com/hubfs/Reports/2021_Free_Global_Games_Market_Report.pdf?utm_campaign=GGMR%202021&utm_medium=email&_hsmi=137510824&utm_content=137510824&utm_source=hs_automation

Gómez, M. A., Gómez-Lopez, M., Lago, C., & Sampaio, J. (2012). Effects of game location and final outcome on game-related statistics in each zone of the pitch in professional football. *European Journal of Sport Science*, *12*(5), 393–398. https://doi.org/10.1080/17461391.2011.566373

Günther, C. W., & van der Aalst, W. M. P. (2007). Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics. *Lecture Notes in Computer Science*, 328–343. https://doi.org/10.1007/978-3-540-75183-0_24

Han, Y. (1998). HOON-A Formalism Supporting Adaptive Workflows. University of Georgia, Department of Computer Science.

Heerkens, H., & van Winden, A. (2017). Systematisch managementproblemen oplossen (1st ed.). Noordhoff.

Hollingsworth, D., & Hampshire, U. K. (1995). Workflow management coalition: The workflow reference model. Document Number TC00-1003, 19(16), 224.

How Did the Olympic Games Evolve Over Time? (2020, January 30). ThoughtCo. Retrieved 26 May 2022, from https://www.thoughtco.com/history-of-the-olympics-1779619

Information Visualization. (2013). *Information Visualization*. https://doi.org/10.1016/c2009-0-62432-6

Liu, H., Hopkins, W., Gómez, A. M., & Molinuevo, S. J. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. *International Journal of Performance Analysis in Sport*, *13*(3), 803–821. https://doi.org/10.1080/24748668.2013.11868690

Mellor, I. (2020, May 28). *The average LCS player is getting paid about $400k per year*. The Loadout. Retrieved 28 May 2020, from https://www.theloadout.com/league-of-legends/LCS-salaries-2020

Merz, M., Moldt, D., Müller, K., & Lamersdorf, W. (1995). Workflow modelling and execution with coloured Petri nets in COSM.

Midway, S. R. (2020). Principles of Effective Data Visualization. *Patterns*, *1*(9), 100141. https://doi.org/10.1016/j.patter.2020.100141

Moura, F. A., Martins, L. E. B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of Sports Sciences*, *32*(20), 1881–1887. https://doi.org/10.1080/02640414.2013.853130

Paradise, A. (2018, December 1). *The rise of esports as a spectator phenomenon*. VentureBeat. Retrieved 26 May 2022, from https://venturebeat.com/2018/11/30/the-rise-of-esports-as-a-spectator-phenomenon/

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007a). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/mis0742-1222240302

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007b). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/mis0742-1222240302

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, *12*(Proc2). https://doi.org/10.14198/jhse.2017.12.proc2.05

Reisig, W. (2011). *Petri Nets: An Introduction (Monographs in Theoretical Computer Science. An EATCS Series, 4)* (Softcover reprint of the original 1st ed. 1985 ed.). Springer.

Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M., & Perry, R. (2016). Data visualization. International Journal of Engineering Research And Advanced Technology (IJERAT), 2(12), 11-16.

Sætran, L., & Oggiano, L. (2008). Skin Suit Aerodynamics in Speed Skating. *Sport Aerodynamics*, 93–105. https://doi.org/10.1007/978-3-211-89297-8_5

Schwabish, J. (2021). *Better Data Visualizations*. Amsterdam University Press.

Sheehan, R. (2000). The professionalization of college sports. Higher education in transition: The challenges of the new millennium, 133-158.

Sivaraman, E., & Kamath, M. (2002). On the use of Petri nets for business process modeling. In *IIE Annual Conference. Proceedings* (p. 1). Institute of Industrial and Systems Engineers (IISE).

Team Liquid - Esports Team Summary :: (n.d.). Esports Earnings. Retrieved 26 May 2022, from https://www.esportsearnings.com/teams/102-team-liquid

Team Liquid - Professional Esports Organization. (n.d.). Team Liquid. Retrieved 26 May 2022, from https://www.teamliquid.com/

Tufte, E. R. (2001). *The Visual Display of Quantitative Information PAPERBACK*. Amsterdam University Press.

van der Aalst, W. (2012a). Process Mining. *ACM Transactions on Management Information Systems*, *3*(2), 1–17. https://doi.org/10.1145/2229156.2229157

van der Aalst, W. (2012b). Process mining. *Communications of the ACM*, *55*(8), 76–83. https://doi.org/10.1145/2240236.2240257

van der Aalst, W., Adriansyah, A., de Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., van den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., de Leoni, M., . . . Wynn, M. (2012). Process Mining Manifesto. *Business Process Management Workshops*, 169–194. https://doi.org/10.1007/978-3-642-28108-2_19

van der Aalst, W. M., & Dustdar, S. (2012). Process Mining Put into Context. *IEEE Internet Computing*, *16*(1), 82–86. https://doi.org/10.1109/mic.2012.12

van der Aalst, W. M. P. (2011a). Process Discovery: An Introduction. *Process Mining*, 125–156. https://doi.org/10.1007/978-3-642-19345-3_5

van der Aalst, W. M. P. (2011b). Process Mining. *Process Mining : Discovery, Conformance and Enhancement of Business Processes*. https://doi.org/10.1007/978-3-642-19345-3

van der Aalst, W. M. P., La Rosa, M., & Santoro, F. M. (2016). Business Process Management. *Business & Information Systems Engineering*, *58*(1), 1–6. https://doi.org/10.1007/s12599-015-0409-x

van der Aalst, W., & Weijters, A. (2004). Process mining: a research agenda. *Computers in Industry*, *53*(3), 231–244. https://doi.org/10.1016/j.compind.2003.10.001

van derWerf, J., van Dongen, B., Hurkens, C., & Serebrenik, A. (2009). Process Discovery using Integer Linear Programming. *Fundamenta Informaticae*, *94*(3–4), 387–412. https://doi.org/10.3233/fi-2009-136

van Dongen, B. F., de Medeiros, A. K. A., Verbeek, H. M. W., Weijters, A. J. M. M., & W.M., A. (2005). The ProM framework: A new era in process mining tool support. *In International Conference on Application and Theory of Petri Net*, 444–454.

Weijters, A., & van der Aalst, W. (2003). Rediscovering workflow models from event-based data using little thumb. *Integrated Computer-Aided Engineering*, *10*(2), 151–162. https://doi.org/10.3233/ica-2003-10205

Weske, M. (2012). Process Orchestrations. *Business Process Management*, 125–242. https://doi.org/10.1007/978-3-642-28616-2_4

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180

Zhang, C., and S. Zhang. 2002. Association Rules Mining: Models and Algorithms. In Lecture Notes in Artificial Intelligence, volume 2307, page 243, Springer-Verlag.

Appendix A: Basic Control Flows

The **and split** pattern represents a point in the model where activity A is completed and both activities B and C are enabled at the same time.

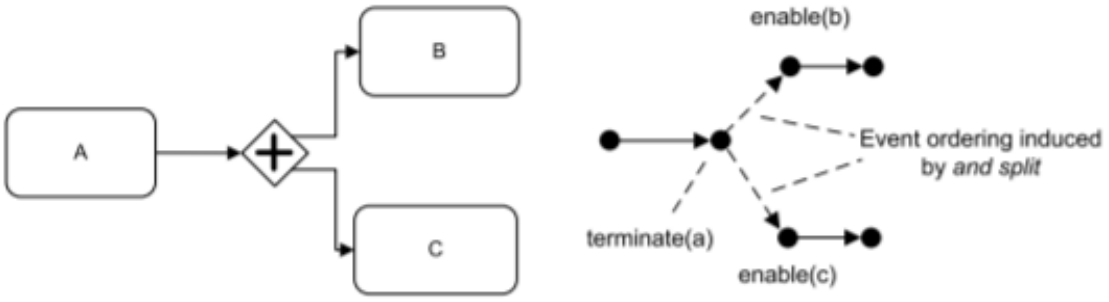


Figure 24: And split pattern (Weske, 2012)

How can FIFA be represented in this pattern? Example: Within FIFA, this pattern will not be found because the player cannot do two activities at the same time.

The **and join** pattern represents a point in the model where activities B and C are finished and both activities B and C are needed to enable activity D.

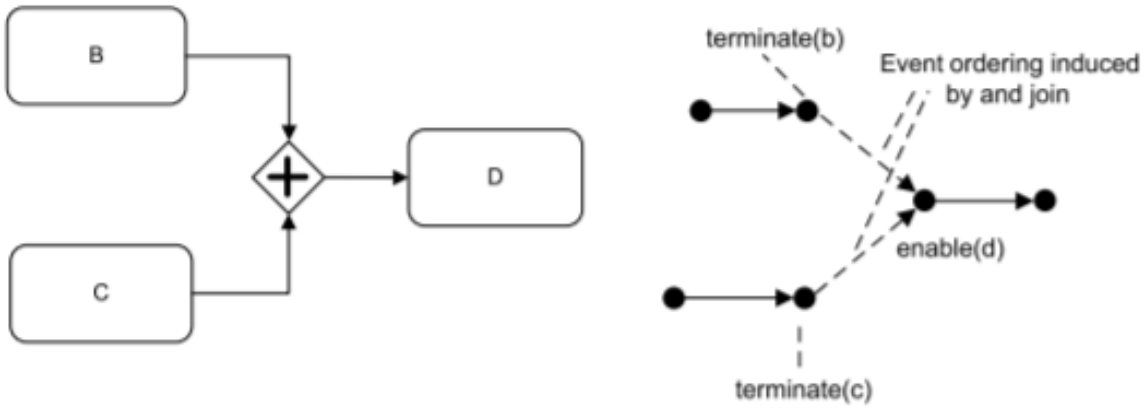


Figure 25: And join pattern (Weske, 2012)

How can FIFA be represented in this pattern? Example:
Within FIFA, this pattern will not be found because the player cannot do two activities at the same time.

The **or split** represents a point in the model where after the completion of activity A, one of the activities B and C or both can be enabled.

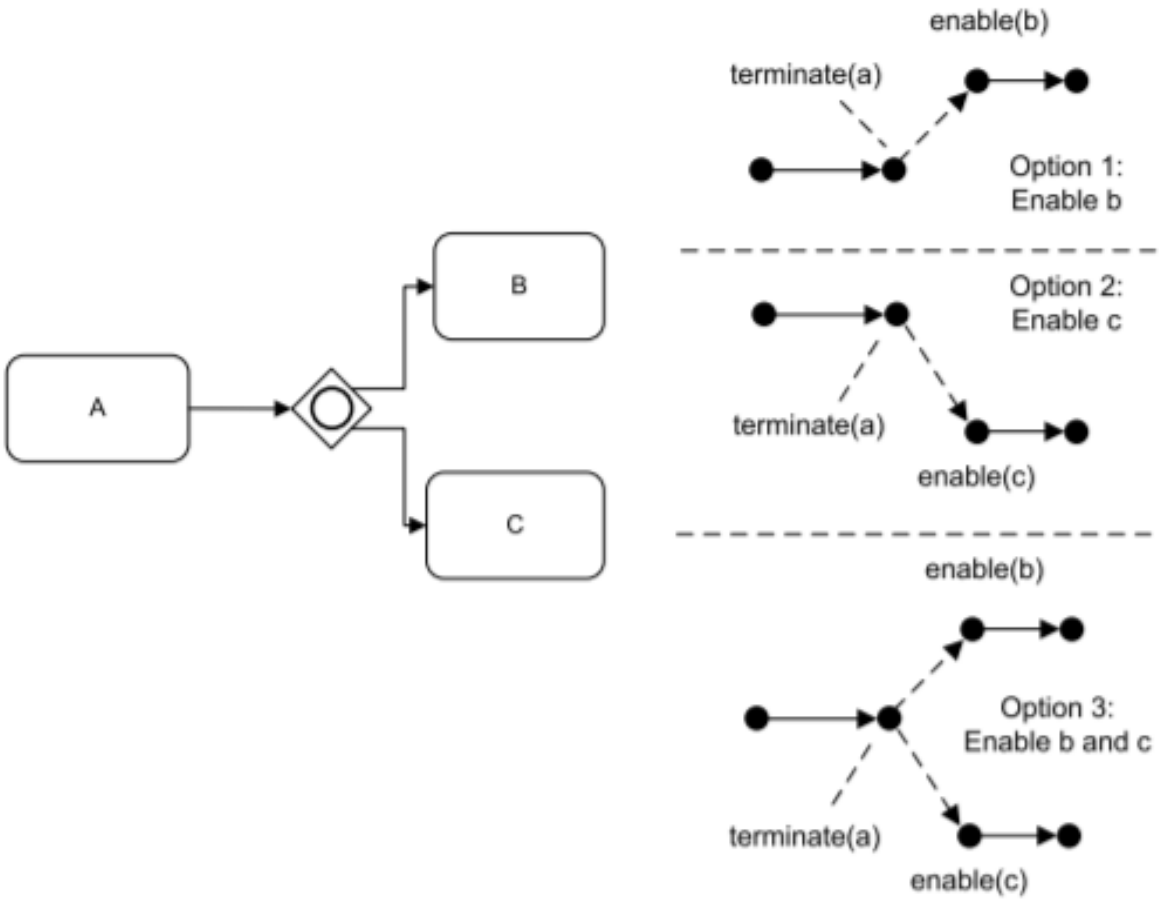


Figure 26: Or split pattern (Weske, 2012)

How can FIFA be represented in this pattern? Example:
Within FIFA, this pattern will not be found because the player cannot do two activities at the same time. So the pattern will automatically represent an **exclusive or split**.

The **or join** represents a point in the model where activity D is enabled after the completion of either both activities B and C or one of them. The or join can bring problems because it cannot decide how long to wait before enabling the next activity.

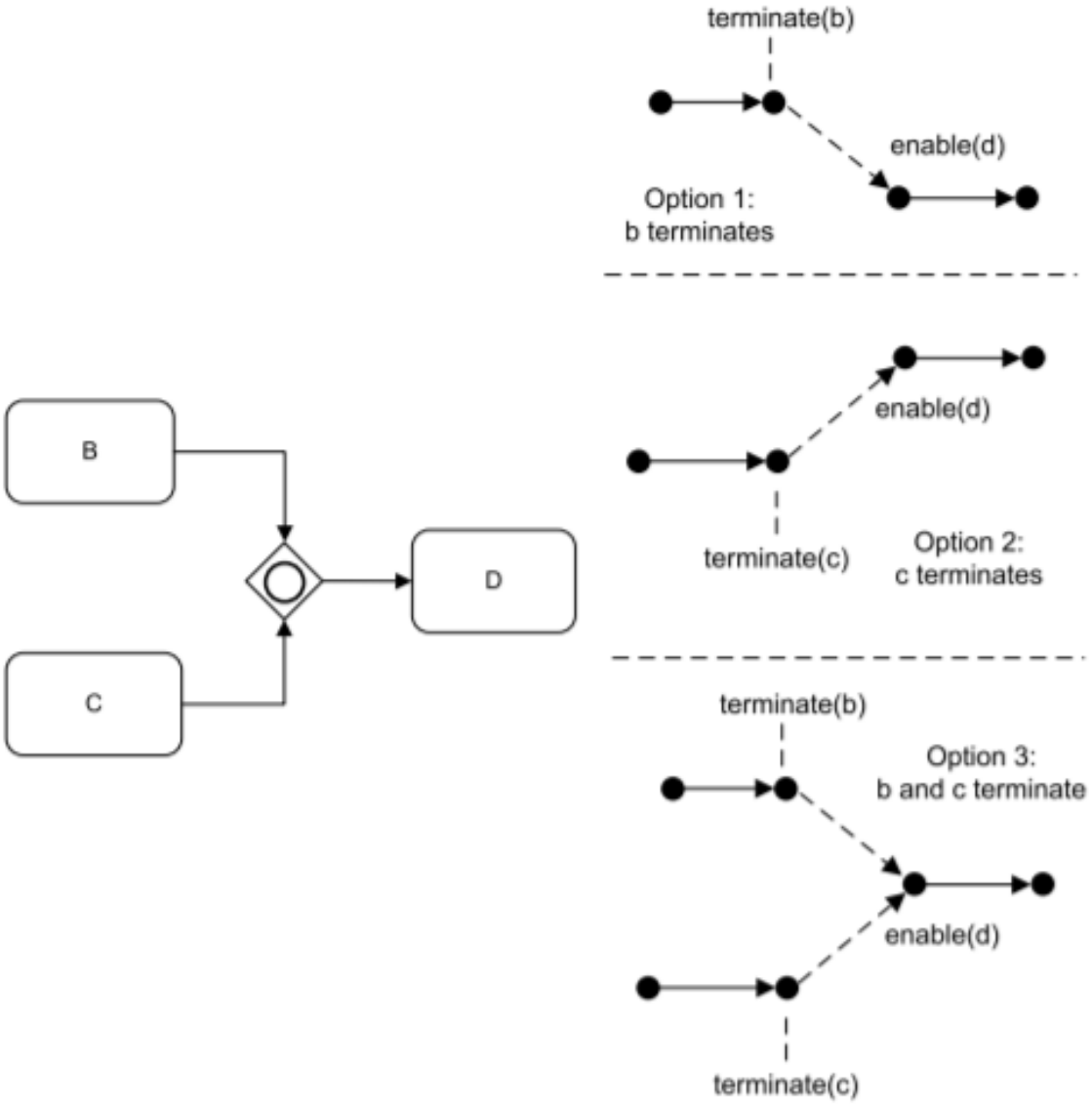


Figure 27: Or join pattern (Weske, 2012)

How can FIFA be represented in this pattern? Example:
Within FIFA, this pattern will not be found because the player cannot do two activities at the same time. So the pattern will automatically represent an **exclusive or join**.

The **multi-merge** or **multiple merge** represents a point in the model where it is the equivalent of the **exclusive or join**, but it differs in the fact that it can also enable when two of the incoming branches are activated.

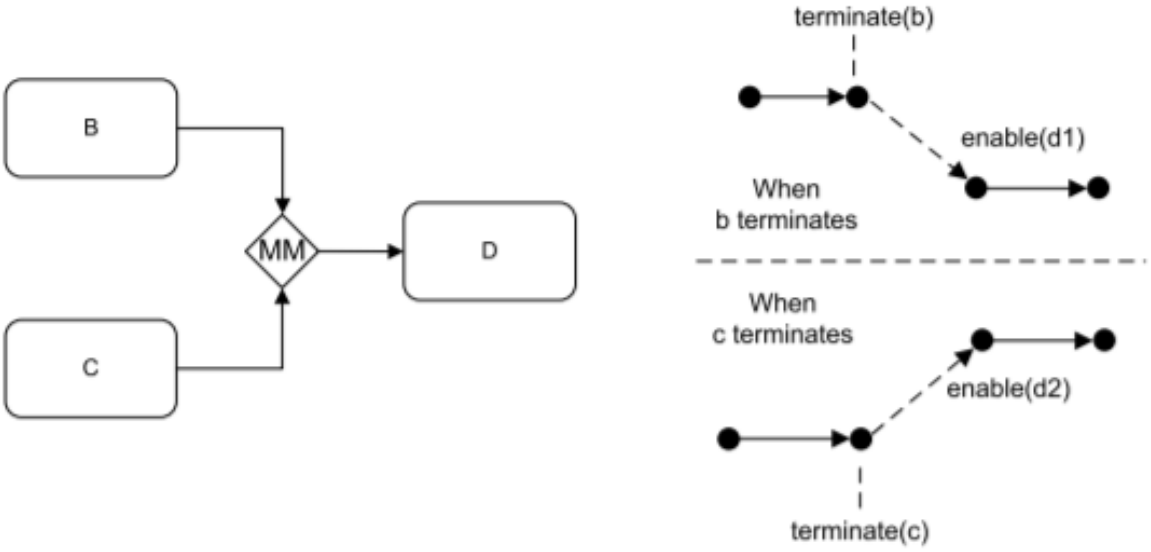


Figure 28: Multi-merge pattern (Weske, 2012)

How can FIFA be represented in this pattern? Example:
Within FIFA, this pattern will not be found because the player cannot do two activities at the same time.

The **discriminator** represents a point in the model where it enables the subsequent activity D when it receives a completion from either activity B or C. Then it waits for the other unfinished activity to finish before resetting itself to be enabled again.

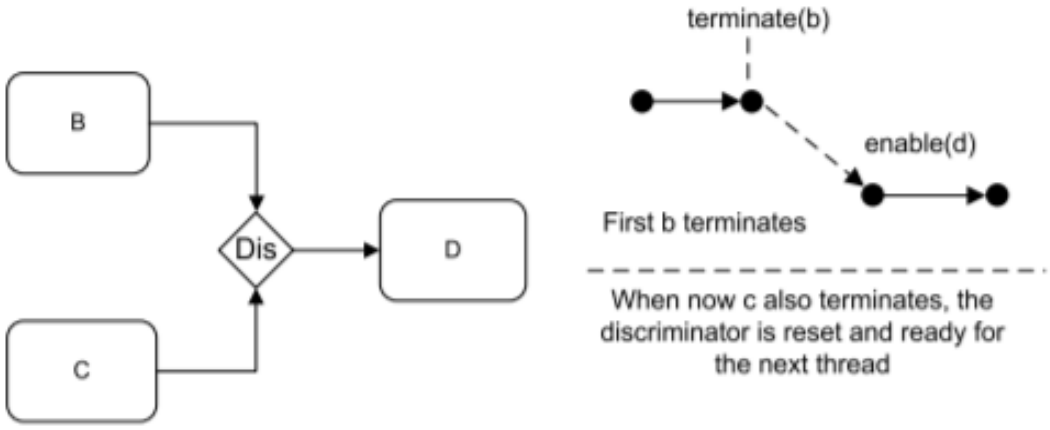


Figure 29: Discriminator pattern (Weske, 2012)

How can FIFA be represented in this pattern? Example:
Within FIFA, this pattern will not be found because the player cannot do two activities at the

same time.

Appendix B: List of key variables

Table 8: Variables from statistical football studies and FIFA 20

Source	**Key events**
Gómez, M. A., Gómez-Lopez, M., Lago, C., & Sampaio, J. (2012). Effects of game location and final outcome on game-related statistics in each zone of the pitch in professional football. European Journal of Sport Science, 12(5), 393-398.	- Goals - Shots - Fouls - Turnovers - Ball recover - Crosses
Moura, F. A., Martins, L. E. B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. Journal of sports sciences, 32(20), 1881-1887.	- Shots - Shots on goal - Goals performed - Fouls committed - Fouls suffered - Corner kicks - Free kicks to goal - Offside - Own goals - Yellow cards - Second yellow cards - Red cards - Playing time with ball possession - Percentage ball possession
Liu, H., Hopkins, W., Gómez, A. M., & Molinuevo, S. J. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. International Journal of Performance Analysis in Sport, 13(3), 803-821.	- Assist - Ball recovery - Block Actions of outfield players: - Challenge - Clearance - Cross - Dispossessed - Dribble - Foul - Interception - Key pass - Offside - Pass - Shots - Tackle - Through ball - Turnover

	Actions of goalkeepers - Catch - Collected ball - Cross not claimed - Drop - Goalkeeper kick from hands - Goalkeeper launch - Goalkeeper throw - Keeper sweeper - Penalty faced - Punch - Save - Smother
Brito Souza, D., López-Del Campo, R., Blanco-Pita, H., Resta, R., & Del Coso, J. (2019). A new paradigm to understand success in professional football: analysis of match statistics in LaLiga for 8 complete seasons. International Journal of performance analysis in sport, 19(4), 543-555.	Attacking variables: - Goal - Shot - Shooting accuracy - Pass - Successful pass - Passing accuracy - Cross - Penalty kick - Turnover - Foul received - Corner - Free kick goal - Offside Defensive variables: - Goal received - Shot conceded - Effectiveness against conceded shooting - Foul committed - Penalty kick conceded - Corner against - Yellow cards - Red cards - Free kick goals received - Recovery
García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A., & Luengo-Sánchez, S. (2021). In-game behaviour analysis of football players using machine learning techniques based on player statistics. International Journal of Sports Science &	Offensive actions: - Assist - Big chances - Chances missed - Crosses - Expected assists - First touch goal - Goals - Missed shots

Coaching, 16(1), 148-157.	- Pass verticality - Pull backs - Second assists - Shots - Shots on target - Take on lost - Take on ratio - Takes on total - Takes on won - Weak shots Defensive actions: - Aerials lost - Aerials ratio - Aerials won - Ball recoveries - Blocked passes - Challenges - Clearances - Failed to block - Failed interceptions ratio - Interceptions - Offsides provoked - Tackles ratio - Tackles won with possession - Tackles won without possession - Total aerials - Total interceptions - Total tackles Game construction actions: - Back passes - Build up play - Dispossessed - Errors - Front passes - Good skills - Individual plays - Launches - Long balls - Pre-shoot pass - Passes ratio - Successful passes - Switches of play - Total passes - Turnovers - Unsuccessful passes
The statistics screen at the end of a FIFA match	- Goals - Shots

	- Shots on target - Possession - Tackles - Fouls - Yellow cards - Red cards - Injuries - Offsides - Corners - Shot accuracy - Pass accuracy

Appendix C: ProM steps

Figure 30: Start screen ProM 6.10

This is the start screen of ProM called “Workspace”, where data can be imported. ProM works with data files in the XES format. ProM allows for conversion from log formats to XES files. When a log format is imported, ProM will ask for conversion of the format to a XES file. In the case of the research, a CSV file was converted to a XES file.

Figure 31: File work options

In the following step, the XES file can be selected and ProM gives 4 different work options for the file. The star button puts your file in the “Favorites” tab for files. The other tabs are the “All”, “Imported” and “Selection” tabs. With the eye button, the data of the file can be viewed, there will also be a short summary of the log. The summary will give the number of different classes or game activities, the absolute number of different occurrences and the number of different

occurrences in percentages. The play button initiates the modeling step where the specific process modeling algorithm can be chosen. The cross button deletes the selected file. There are other buttons but these are not essential to create a model.

Figure 32: Process mining algorithm selection menu

When the play button is pressed, the next menu opens which is the process mining selection menu. As previously mentioned, there are many different process mining algorithms available but to keep it to the point, the HeuristicsMiner is chosen as the selected algorithm. When “HeuristicsMiner” is typed into the search bar, only a single option is presented and this option is selected, then the start button is pressed. Then ProM presents an options tab where the classifier is to be selected.

Appendix D: Dashboard screenshots

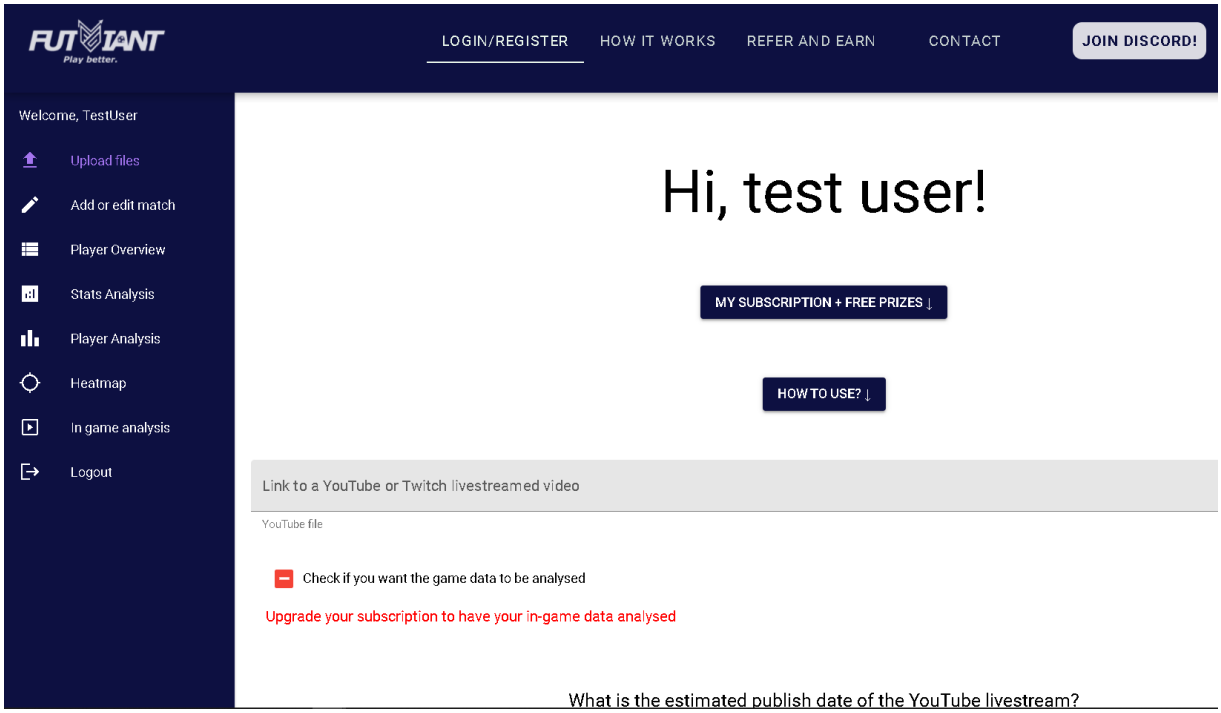


Figure 33: : “Upload files” section in the dashboard

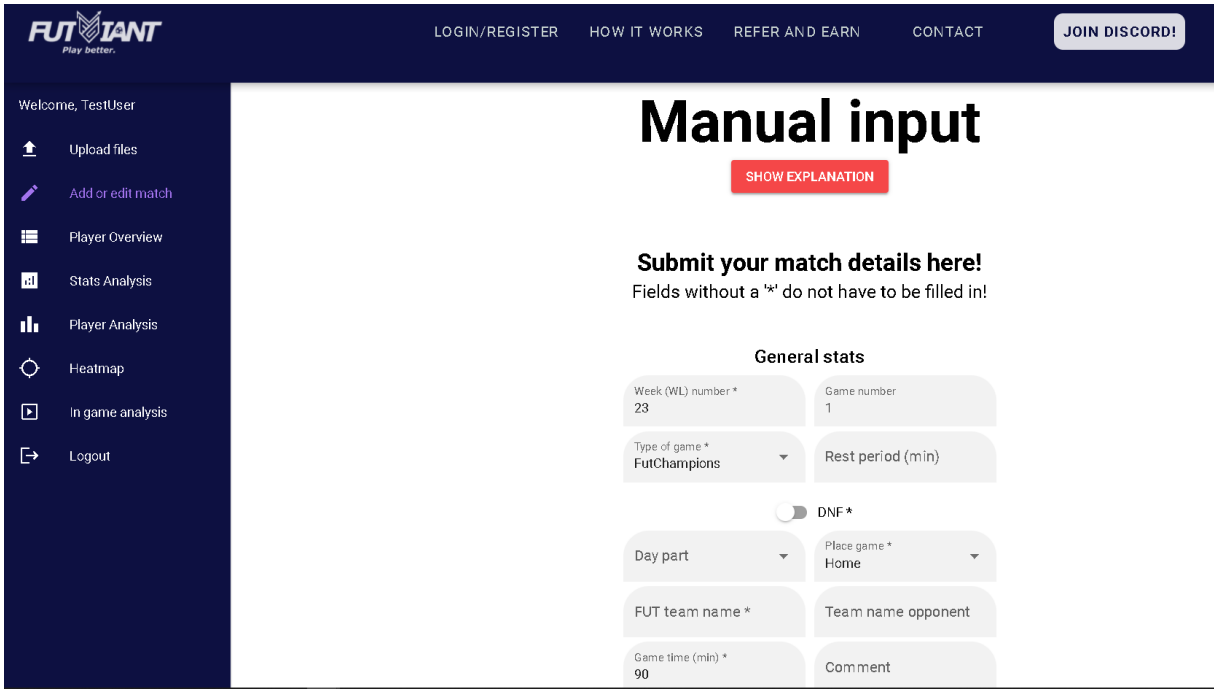


Figure 34: “Add or edit match” section in the dashboard

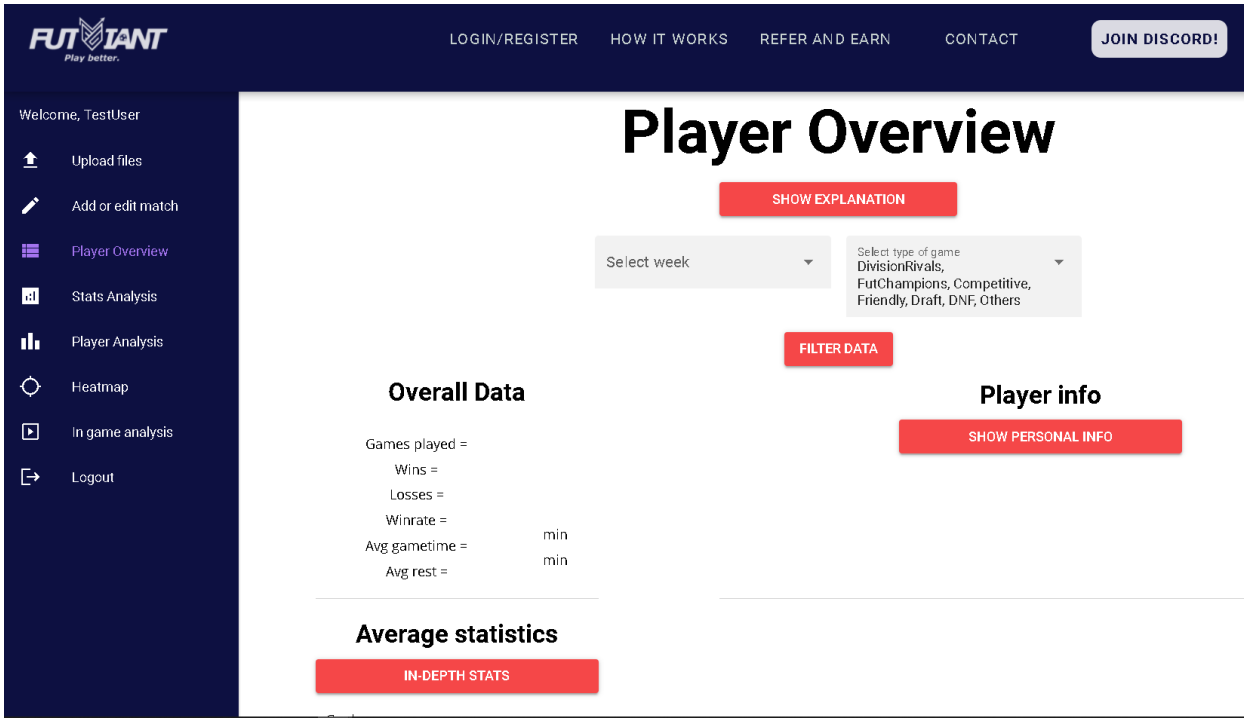


Figure 35: “Player overview” section in the dashboard

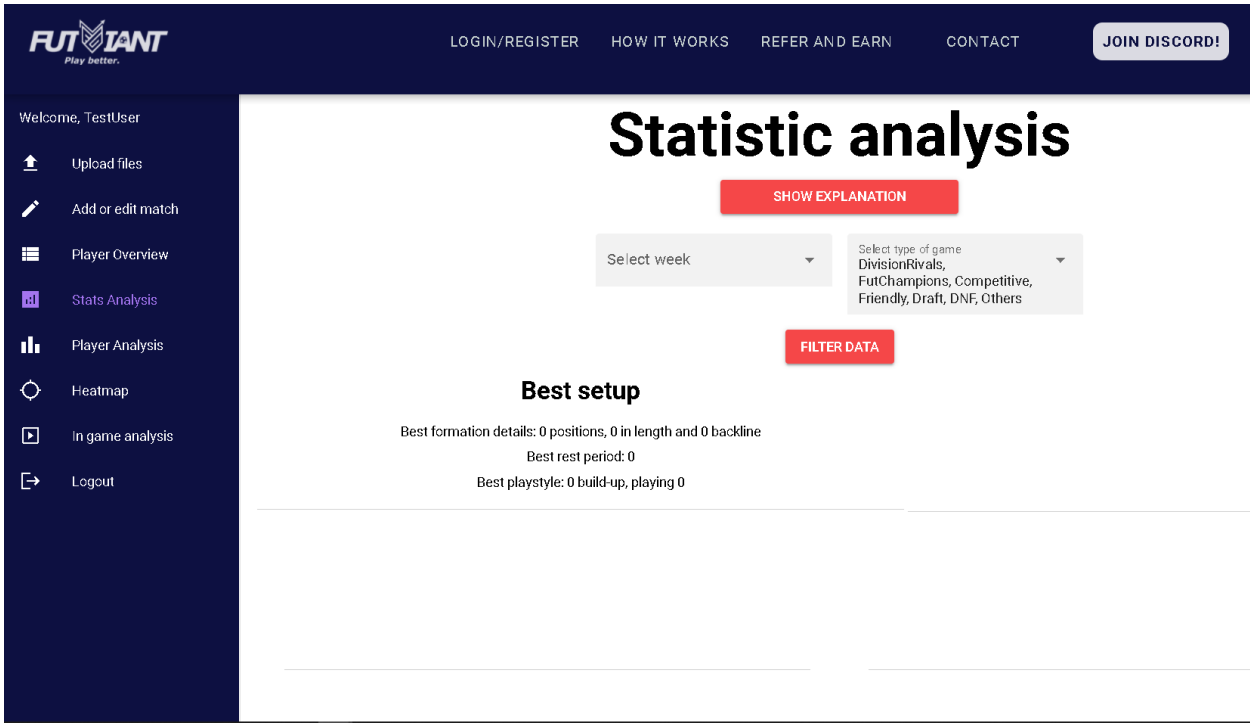


Figure 36: “Stats analysis” section in the dashboard

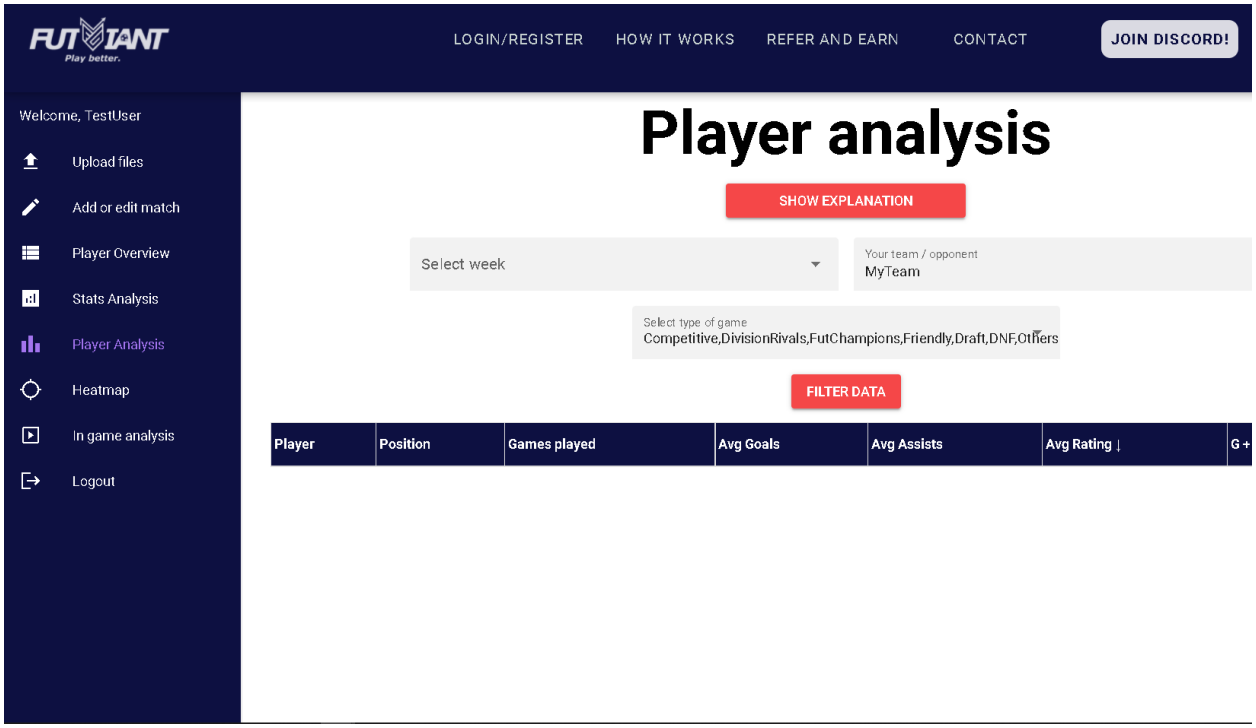


Figure 37: “Player analysis” section in the dashboard

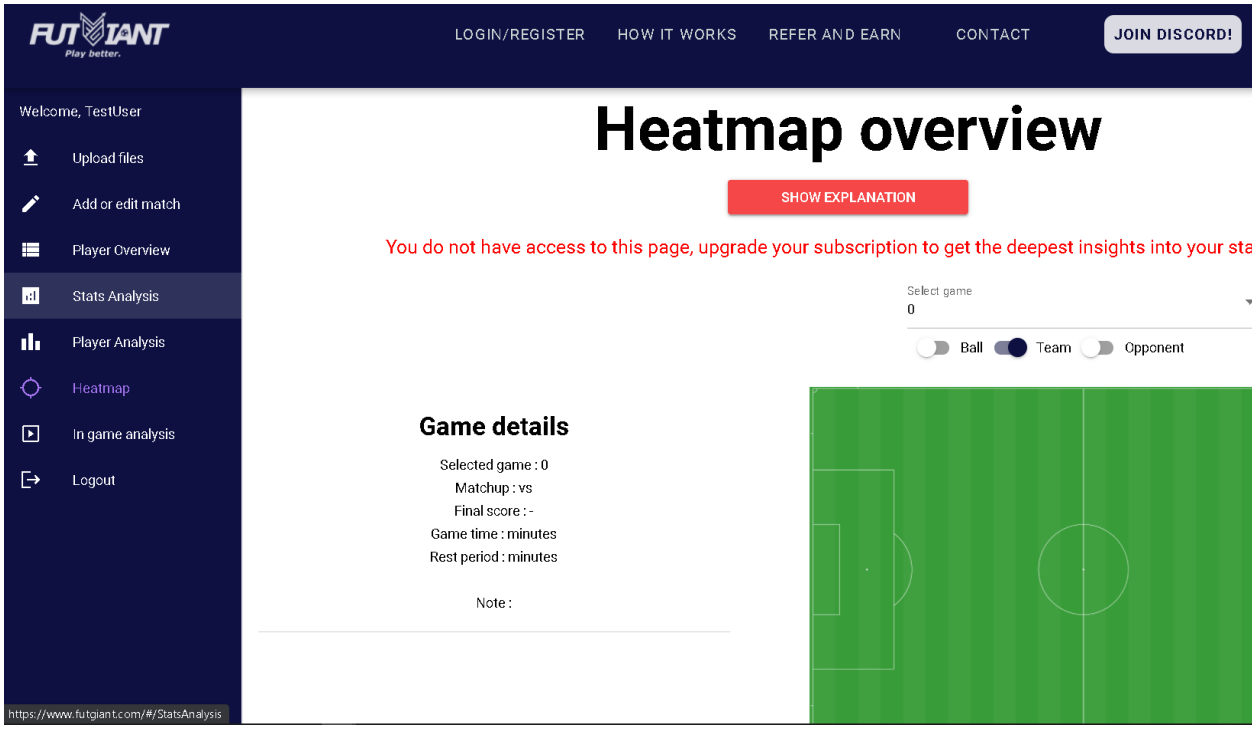


Figure 38: “Heatmap” section in the dashboard

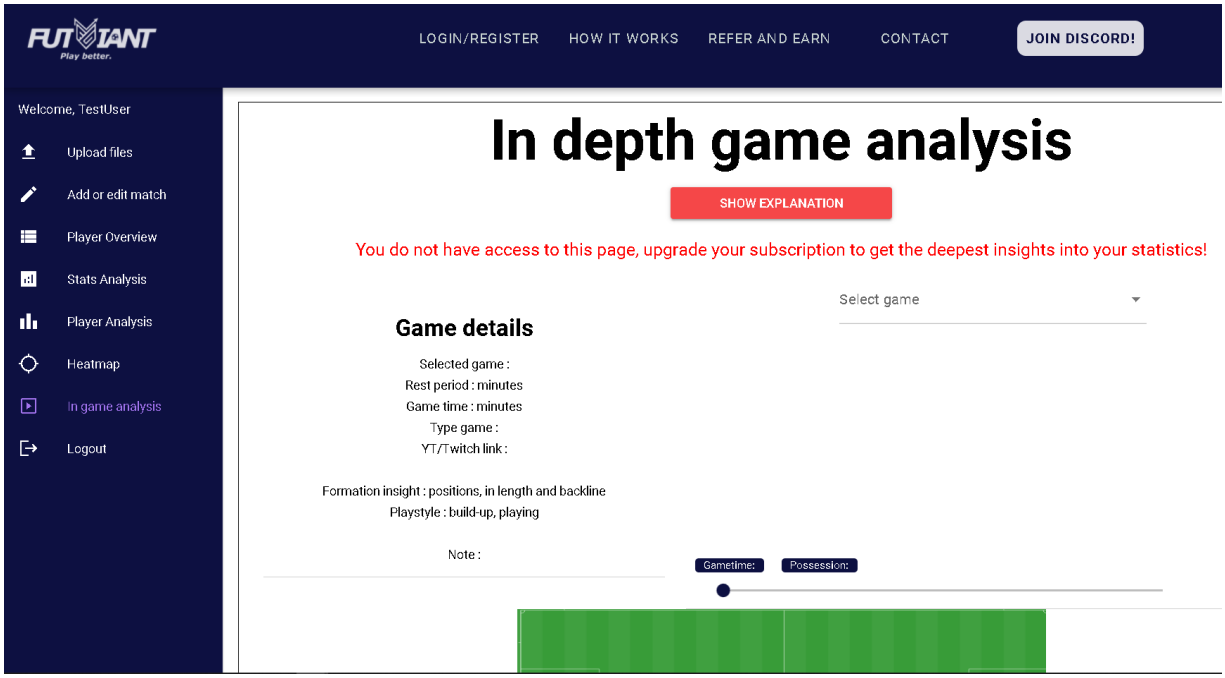


Figure 39: “In game analysis” section in the dashboard