Project Title: <Your mini-project title>

Name: <Your name>

Representations

(LO1) Understand and apply **Knowledge Graph Embeddings**	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

(LO2) Understand and apply **logical knowledge** in KGs	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

(LO3) Understand and apply **Graph Neural Networks**	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

(LO4) Compare different Knowledge Graph **data models** from the database, semantic web, machine learning and data science communities.	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

Systems

(LO5) Design and implement **architectures** of a Knowledge Graph	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

(LO6) Describe and apply **scalable reasoning** methods in Knowledge Graphs	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

(LO7) Apply a system to **create** a Knowledge Graph	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

(LO8) Apply a system to **evolve** a Knowledge Graph	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

Applications

(LO9) Describe and design **real-world applications** of Knowledge Graphs	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

(LO10) Describe **financial Knowledge Graph** applications	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

(LO11) Apply a system to provide **services** through a Knowledge Graph	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

(LO12) Describe the **connections** between Knowledge Graphs (KGs), Machine Learning (ML) and Artificial Intelligence (AI)	☐ I showed **basic** proficiency ☐ **exceeded** basic proficiency
<1-2 Sentences how you addressed this LO in your portfolio>	<Page(s) in the portfolio where this is discussed>

Additional Information

HAS NO EFFECT ON MARKING!

(please fill it out honestly, even if it is less than what is suggested in the ECTS breakdown – you are not judged on time spent!)

How many hours did you spend on your **mini-project**? (the ECTS breakdown suggests **40** hours for this)	XX hours

* please exclude any hours you spent on parts reused from other courses

How many hours did you spend on your **portfolio document preparation (this PDF)**? (the ECTS breakdown suggests **15** hours for this)	XX hours

* please exclude any hours you spent on parts reused from other courses

Please indicate if you have **reused** parts of the **mini-project** from other courses	☐ I reused some parts: <XX>% of the mini-project
-	

Please indicate if you have **reused** parts of the **portfolio document** from other courses	☐ I reused some parts: <XX>% of this document
-	

Declaration

I have marked all parts generated by Generative AI (e.g., ChatGPT) and given any prompt I used either in a footnote or in an appendix making clear which parts are generated by which prompts or similar.	☐ I confirm this

Portfolio Cover Pages

Here follows your portfolio (report), structured as you wish!

A possible structure (but by no means a suggestion, use whatever you like) is:

1. Introduction
2. Background
3. Method
4. Results
5. Conclusion

IN PARAGRAPHS WHERE YOU EXPLICITLY DISCUSS A PARTICULAR LEARNING OUTCOME (ESPECIALLY THOSE MENTIONED IN YOUR COVER PAGES), PLEASE USE REFERENCES IN THE FORM OF (LO1) TO INDICATE THAT.

An example:

"For creating my Knowledge Graph (LO7) I used a combination of logical rules encoding the domain knowledge of XXX and a Knowledge Graph embedding that infers edges between entities not explicitly given in the source data. […]

Let me discuss the Knowledge Graph embedding (LO1) used for that in more detail […]"

You do not have do that everywhere (it is not a competition on how often you mention a particular learning outcome), but especially at places that you explicitly reference in your cover pages, it helps you (and us) to see more easily which point you want to emphasize!

Introduction

According to FIFA, the global authority overseeing association football and its variations such as beach soccer and futsal, there are approximately five billion football fans around the world[1]. Of these, an estimated 1.5 billion viewers tuned in to watch the final match of the 2022 FIFA World Cup[2], making it one of the most watched events globally and highlighting football's position as the world's most popular sport. Consequently, FIFA and its various adjacent national governing bodies are also very motivated to serve this lively market by expanding both the frequency and variety of competitions. These events are subsequently auctioned off to various (often subscription-based) broadcasters, such as Sky or DAZN, who are willing to pay increasingly higher sums for the broadcasting rights of national and international leagues, cup competitions, and, in some cases, even clubs. Additionally, rising sponsorship deals, ticket prices and the instrumentalization of football for sports washing by various states and politicians, are contributing to the growing influx of money in the sport, which is then invested in player transfers, salaries, and stadium infrastructure. Ultimately, football has evolved into a multi-billion-dollar industry that involves not only governing bodies, clubs, corporations and fans but also entire nations, as exemplified by PSG's ownership structure or the political manoeuvring and scandals behind World Cup hosting decisions. These developments also come with drawbacks, such as the widely debated risks associated with the heightened workload, resulting from the increased number of competitions and games particularly for elite players. One of the most pressing concerns is the increased likelihood of injuries, which can have far-reaching consequences not only affecting players' careers but also impacting team performance and the overall quality of football competitions.

The initial project-idea described in the one pager aimed to develop a GNN-model using player interactions, game schedules and historical injury data in order to predict the likelihood of injuries and injury types. However, this approach had to be adapted, as the search for consistent and usable data revealed that the search, compilation and cleansing of the data would go way beyond the scope and intended timeframe of this project. To implement the Graph Neural Network described in the one-pager a dataset containing player data from the FIFA video game series (later rebranded as EA Sports FC) produced by EA was used. Since the dataset does not include information on injuries, but does provide detailed player attributes such as name, age, position, and realistic ratings ranging from 0 to 99 for defending, dribbling, attacking, overall rating, and potential among many others, the approach was adapted. Instead of injury prediction, the focus shifted to forecasting a player's potential based on these attributes and their development across different iterations of the game. To highlight other applications of such a football knowledge graph, Knowledge Graph Embeddings (KGEs) were generated by training a TransE model on the knowledge graph. These embeddings were then used for a weighted similarity calculation, by combining the cosine similarity between players' embeddings with additional player characteristics. This similarity calculation allows for the use of the graph to find players with similar profiles, facilitating potential use cases such as scouting recommendations relevant for decisions regarding squad building and player transfers.

[1] https://publications.fifa.com/en/vision-report-2021/the-football-landscape/ (19.02.2025, 14:50)
[2] https://inside.fifa.com/tournament-organisation/world-cup-2022-in-numbers (19.02.2025, 14:50)

The FIFA/EA Sports FC Franchise

The FIFA video game franchise is one of the largest developed and published by Electronic Arts (EA) Sports. Renowned for its official licensing agreements with FIFA and various football leagues, the game allows for a highly realistic representation of real-life football. In 2022, after a 30-year partnership, EA Sports and FIFA ended their licensing agreement, making FIFA 23 the final game under the FIFA name. As a successor, EA launched the EA Sports FC series, with EA Sports FC 24 marking the beginning of this new era. The only thing that has changed is the name, the rating system and the gameplay are largely in line with the predecessors.

The FIFA video game series, developed and published by EA Sports, stands as one of the largest franchises in the industry. Known for its official licensing agreements with FIFA and several football leagues, the game offers a highly realistic simulation of real-life football. In 2022, after a 30-year partnership, EA Sports and FIFA ended their licensing agreement, making FIFA 23 the final game released under the FIFA name. Following this change, EA introduced the EA Sports FC series, with EA Sports FC 24 marking the start of the new era. Though only the branding changed, the rating system and gameplay remained largely consistent with previous editions.

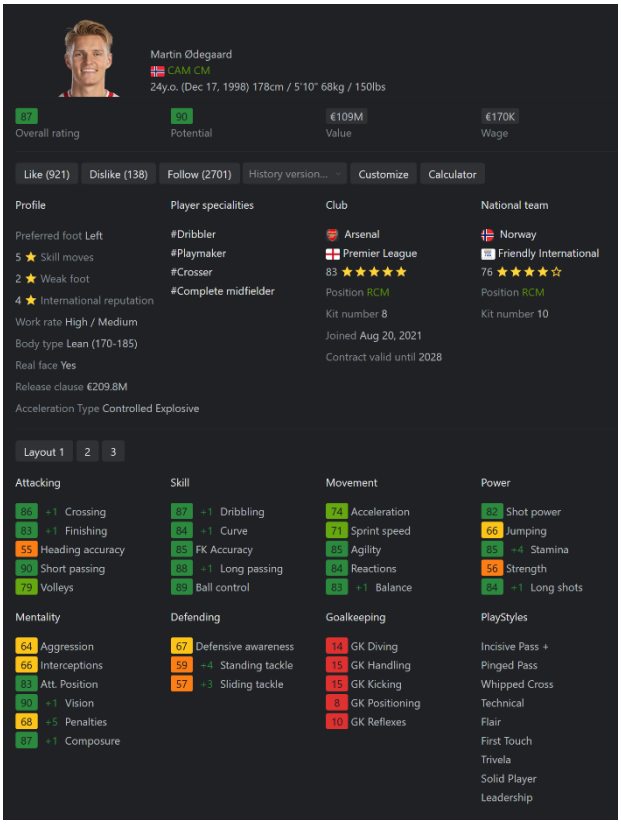

Figure 1 EA FC 24 ratings for Arsenal captain Martin Odegaard (source: sofifa; 19.02.2025, 16:12)

In the game, player ratings are determined through a system that combines input from EA Sports employees, external contributors, and volunteer FIFA Scouts. These experts continually update a detailed database of specific attributes for each player. Numeric attributes are rated on a scale from 0 to 99, while factors like skill moves, weak foot ability, and international reputation are rated on a 0 to 5-star scale. To compute the overall rating, ratings for Pace, Shooting, Passing, Dribbling, Defending, and Physical, among others, are weighted using positional coefficients based on a player's on-field position. Additionally, a player's international reputation is factored in to ensure the final rating accurately reflects their real-world counterpart's performance.[3] In addition to the stats mentioned, qualitative information such as player specialty, body type, and work rate is also provided. Greatly relevant for the in-game career mode, a potential rating is also included, which estimates how the player might develop, along with current player value and salary estimations. Figure 1 provides an example of the rating system in action, showing a section of the ratings for Arsenal FC captain Martin Ødegaard in EA Sports FC 24.

[3] https://www.goal.com/en/news/fifa-player-ratings-explained-how-are-the-card-number--stats-decided/1hszd2fgr7wgf1n2b2yjdpgynu (19.02.2025, 15:37)

Data Preprocessing

Initial Data Cleaning

As described in the introduction, the selected dataset contains information on players and teams from FIFA 15 up to EA Sports FC 24. The data was scraped from Sofifa.com, a publicly accessible platform that compiles player ratings and detailed information about the teams featured in the video game. This analysis focuses exclusively on male players and teams. The database consists of two CSV files: the player file, containing 180,021 observations across 109 variables, and the team file, comprising 6,947 observations with 54 columns. Spanning from the 2015 to the 2024 iteration of the game, the dataset represents players and teams' multiple times, with each instance reflecting a different stage of their career and development. The dataset's columns include unique identifiers, names, and detailed ratings, as outlined in the introduction. Additionally, it features qualitative attributes such as work rate and preferred foot, as well as financial and contractual details, including transfer market values, salary estimates, and measures of prestige and reputation.

Several variables, such as jersey numbers and release clauses, were removed either due to their limited analytical value or, in the case of release clauses, because they contained unreliable information that did not reflect real-world data. The variable indicating whether a player plays for a national team was also excluded, as the game features only a limited selection of national teams. Consequently, many real-life national team players would be misclassified as non-national team players, introducing inconsistencies. Certain categorical attributes were transformed for better usability. For instance, the club_position column, originally containing specific role labels such as "RES" (reserve) and "SUB" (substitute), was recoded into a numerical indicator distinguishing between starters, substitutes, and reserves. However, this approach has limitations: players who are injured or unavailable at the time of the game's release may not appear as starters in the dataset, even if they are regular starters in real life. A more refined metric, such as the ratio of starting eleven appearances to total games or minutes played per match exceeding a certain threshold, would provide a more accurate measure of a player's role within the team. Similarly, the club_loaned_from variable, which indicates whether a player was on loan, was converted into a binary on_loan variable. Originally, the player_positions column contained multiple position labels for players capable of playing in different roles. To simplify analysis, this variable was aggregated into broader positional categories. The first-listed position was assumed to be a players primary role. Goalkeepers (GK) remained distinct, while strikers (ST, CF) were grouped under ATT (attack). Wide players, including wingers and wingbacks (RW, LW, RWB, LWB), were categorized separately as WING due to their unique tactical function, particularly in systems used by coaches such as Antonio Conte or Ruben Amorim, where wingbacks play a more offensive role. Midfielders, spanning central and defensive roles (CM, CAM, CDM, LM, RM), were unified under MID, while defenders (CB, RB, LB) were categorized as DEF.

In some cases, teams from one country compete in leagues based in another, requiring adjustments to ensure consistency. Instead of assigning clubs to their country of origin, the dataset was modified so that a team's nationality aligned with the league in which it competed. For example, AS Monaco, despite being a club from Monaco, competes in France's Ligue 1 and was thus assigned "France" as its nationality. Similarly, Welsh teams playing in England's Championship and League Two were reassigned to England, while Canadian teams in Major League Soccer were attributed to the United States. Beyond these adjustments, certain leagues shared identical names across different countries, creating ambiguity. One notable example was the "Premier League," which referred to both the English and Russian top divisions. To resolve this, league names were adjusted for clarity, with the Russian Premier League explicitly renamed as "Russian Premier League" to distinguish it from its

English counterpart. Finally, leagues with an insufficient number of teams were excluded from the analysis, with the threshold set at eight teams. For instance, players and teams from the Croatian league, which does not meet this criterion, were omitted.

Missing Value & NA Handling

There were various types of missing values in the dataset. For example, players listed as free agents in the FIFA game lacked both a club and league affiliation. After examining individual examples, it quickly became apparent that many players who are listed as free agents in the game play for a club in real life that is not included in the game. To address this, all players without a recorded league or club were filtered out, ensuring that only those with a clear team affiliation remained in the dataset. Beyond filtering out free agents, missing values were imputed. One example for variable requiring imputation was value_eur, which represents a player's estimated market value. To facilitate the applied imputation process, players were grouped into brackets based on their age and overall rating. Missing values were then replaced with the mean value of players in the same position category, age group, overall rating range, and league level. In rare cases where no suitable comparison group was available, the lowest recorded value within the same category was used as a fallback to prevent introducing artificial inflation. A similar imputation strategy was applied to team-level attributes such as transfer budgets and club valuations. First, domestic_prestige and international_prestige for clubs in FIFA versions 15 and 16 were inconsistent with later iterations, so their scores were imputed using the average from FIFA 17 onwards for the same clubs, leveraging the relative stability of these rankings over time. For financial metrics like transfer_budget_eur and club_worth_eur, a cascading imputation approach was implemented. Initially, missing values were replaced with the mean of clubs sharing similar international and domestic prestige, league level, and overall club rating. If gaps remained, broader groupings with fewer constraints were used to ensure all missing values were filled with the most contextually appropriate estimates.

NA values were also examined and imputed. In the case of the club_joined_date column, which was NA for players who were on loan, a simplified assumption was applied considering that most loans are finalized in the summer. If a player was on loan and lacked a club_joined_date, it was set to August 1st of the year before the corresponding FIFA version, given that the game is released in September of the preceding year (e.g., FIFA 23 was released in September 2022). Another issue involved goalkeeping attributes. Since outfield players do not have goalkeeping statistics, NA values in goalkeeping-specific attributes such as goalkeeping_speed were not true NA values but rather structurally absent data. To correct this, all NA values for non-goalkeepers were set to zero. Similarly, outfield player attributes such as pace, shooting, passing, dribbling, defending, and physic were set to zero for goalkeepers. Finally, the mentality_composure attribute contained NAs. To address this, an imputation strategy similar to the one used for value_eur was applied: missing values were replaced with the mean composure score of players within the same position category, age group, and overall rating range. If any gaps remained, they were filled using the minimum available composure score within the dataset.

Creating the League Data

The original dataset did not contain explicit league-level data, but using the available team information, a derived league dataset was constructed. For this purpose, team data was aggregated at the league level. For attributes such as team ratings, the mean values across all teams within each league were computed. Furthermore, additional league-specific attributes were derived. The overall league rating, as well as its attack, midfield, and defense ratings, were rounded to maintain consistency with the player and team datasets. Regarding league prestige, a normalized prestige score was introduced based on financial indicators. The international prestige of a league was

computed using a weighted formula incorporating normalized values of international_prestige, transfer_budget_eur, and club_worth_eur. This approach accounted for both financial power and reputation, with club worth receiving the highest weight (0,6), as it is often the most stable indicator of a league's standing over time.

Subsampling

Due to computational constraints, a subsampling strategy was implemented to reduce the dataset size while preserving its representativeness. To achieve this, the dataset was restricted to the four most recent FIFA versions (21, 22, 23, and 24). Within this subset, players appearing in multiple FIFA editions were identified, and only those with the highest number of occurrences across these versions were retained. To maintain diversity and balance across different player characteristics, a stratified sampling approach was applied. Players were grouped based on their overall rating range, age group, and positional category. From each unique combination, twenty players were randomly selected to ensure that all player profiles were adequately represented. This resulted in a total of 2,168 unique players and 8,672 player observations, representing 2,163 team observations across 175 league observations.

Knowledge Graph Creation and Population

To construct the knowledge graph, an ontology was designed in Protege to formalize the relevant relationships. The ontology defined in Figure 2 consists of three main classes: Player, Club, and League. Players are linked to clubs through the plays_for relationship, and both players and clubs are associated with leagues using the competes_in property, enforcing a one-to-one constraint to maintain consistency with real-world football structures. Clubs also feature a rival_with relationship, modeled as a symmetric property to capture competitive dynamics between teams. Each entity is enriched with data properties that reflect the relevant characteristics from the FIFA dataset, such as a player's age, position, overall rating, a club's international prestige and transfer budget, and a league's level and nationality. The ontology serves as the structural foundation of the knowledge graph, ensuring logical consistency and enabling graph-based reasoning over the encoded football data (LO5, LO2).

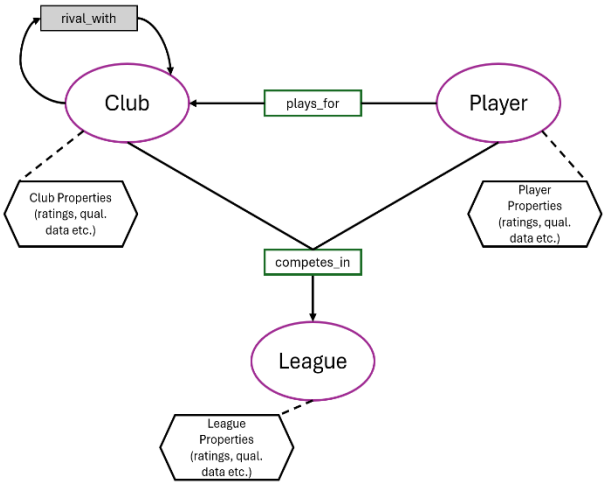

Figure 2 Knowledge Graph Structure

Once the ontology was defined, the next step was to populate the knowledge graph with the preprocessed data described in the previous chapter. This was implemented in Python using rdflib, which facilitated the transformation of structured data into RDF triples that adhere to the relationships and logical constraints defined in the ontology, thus converting the dataset into a graph. (LO2, LO7) Each entity, players, clubs, and leagues, was instantiated as an RDF node, with attributes such as overall rating, potential, skill attributes, and contract details mapped to their respective data properties. The object properties (plays_for, competes_in, and rival_with) were assigned to establish relationships between the entities. Finally, the populated knowledge graph was validated by verifying that each player, club, and league had the correct relationships, and checking for orphaned nodes or inconsistencies. The resulting populated knowledge graph was saved in the form of RDF triples as a

Turtle (ttl) file, which serves as the basis for the Graph Neural Network and Knowledge Graph Embeddings discussed later.

Graph Neural Network

Knowledge Graph Embeddings for Player Similarity Analysis

Restrictions, Relevance and Potential Applications

Erklär hier, was alles passiert, erklär auch dass das alles eher als proof of concept eher gilt. Mach das auch oben

Warum interessiert uns player potential, was wäre die Anwendung für so ein FALL? Hier diskutieren und Beispiele

- HIER Financial KG beschreiben (LO10)
    - Marktwert prediction könnte für Vereine sehr relevant sein
    - Transfermarkt description
- HIER System to provide Services through KGs Zone 14 maybe, Brentford etc. (LO9), touch on LO11

Results and Conclusions

Disclaimer