Plan Overview

A Data Management Plan created using DMPonline.be

Title: Learning and reasoning about spatiotemporal data with applications to sports

Creator: Jesse Davis

Principal Investigator: Jesse Davis

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: Jesse Davis

Project abstract:

Increasingly large amounts of data are collected in team sports such as basketball, football and hockey for the purpose of match analysis. In particular positional tracking data has great potential for analysing a team's tactical behaviour. This form of spatiotemporal data records the locations of players and the ball multiple times per second. However, the sheer scale of this data implies that automated techniques such as AI are essential to unlock its potential. A key open AI challenge for analyzing spatiotemporal data from sports is developing algorithms for identifying and understanding tactics, which are short-term behaviors or patterns that a team exhibits at certain points in a match. Unfortunately, this is an extremely challenging problem that is currently beyond the grasp of current techniques. Therefore, the goal of this project is to develop novel algorithmic approaches for automatically discovering, evaluating, and understanding tactical patterns from spatiotemporal data arising from soccer matches

ID: 213247

Start date: 01-10-2024

End date: 30-09-2028

Last modified: 24-03-2025

Learning and reasoning about spatiotemporal data with applications to sports

Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		*Indicate: **N**(ew data) or E(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
Statsbomb Open data	Public event stream data provided by Hudl StatsBomb. This data describes all on-the-ball actions during a game.	E	D	T	JSON	<100GB	
IDDSE tracking dataset	Public dataset of optical tracking and corresponding event stream data. The tracking data records the location of the players and the ball multiple times per second.	E	D	N + T	XML	<100GB	
Proprietary tracking dataset	Proprietary dataset of optical tracking and corresponding event stream data.	E	D	N + T	TXT + JSON	<1TB	
kloppy	Public software for preprocessing soccer tracking data	E	D	SO		<1GB	
trackingutils	Proprietary software and models for analyzing soccer tracking data	E	D	SO + M		<1GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

- Statsbomb open data is available from https://github.com/statsbomb/open-data
- The IDDSE tracking dataset (7 games from the German Bundesliga) is available from https://doi.org/10.6084/m9.figshare.28196177
- We have 5+ seasons (>1300 games) of optical tracking data with corresponding event stream data for one professional soccer league. This data was provided under a license (i.e., signed by LRD) that enables us to publish on results obtained using the data.
- kloppy is an open-source software package for loading and preprocessing soccer data. Code is available from https://github.com/PySport/kloppy
- trackingutils is proprietary software that we have developed to analyze soccer tracking data. Currently, the code is stored in a private GitLab repo.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- Yes, human subject data (Provide SMEC or EC approval number below)

G-2021-3138-R2

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- Yes (Provide PRET G-number or EC S-number below)

G-2021-3138-R2

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

Yes, there are diverse commercial applications of soccer analytics software. Strategic partnerships with clubs and sports tech firms could drive monetization and industry adoption.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes
- Use of the StatsBomb open dataset is subject to a data user agreement (see https://github.com/statsbomb/open-data/blob/master/LICENSE.pdf). The data or any analysis derived from it cannot be commercially exploited
- Use of the proprietary tracking dataset is subject to a non-disclosure agreement. The data can only be used for academic internal non-commercial research purposes and cannot be shared with third parties.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes
- For the proprietary tracking dataset, the intellectual property of the data remains with the data providers. Any results of the research will be owned by the researchers. The researchers are free to use these results insofar as they do not contain the data.
- Any publication that uses the StatsBomb open data must accredit StatsBomb with the StatsBomb brand logo.
- The IDDSE tracking dataset is made available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. This license allows users to share, copy, redistribute, transform, and build upon the licensed work for any purpose, even commercially, as long as they provide proper attribution to the original creator.
- kloppy is licensed as BSD software. This is a permissive open-source license that allows users to freely use, modify, and redistribute the software. Attribution is the only requirement.

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).

All information regarding the processing of the data sets and all details to reproduce the experiments will be included in a README. A requirements file will include the version numbers of the used external libraries / tools.

Will a metadata standard be used to make it easier to find and reuse the data?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- No

Metadata about the data sets is provided by the data providers as PDF documents and online documentation. Metadata about the source code will be included as documentation and as mentioned in the previous question.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Other (specify below)

A backup of the raw proprietary tracking data is stored in an Amazon S3 bucket, managed by our data partner. The source code will be backed up on the KU Leuven GitLab server. Intermediate data and results will be stored on the department's local filesystem.

How will the data be backed up?

- Other (specify below)

The department's local filesystem uses RAID. An additional backup is made on an external hard disk, owned by the research group.

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Only user accounts created by the department can gain access. The local filesystem also allows setting access control per file/directory. It is thus suited for storing private data sets.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Any costs will be covered out of working costs for the project as they were included in the budget calculation. These are estimated at 250 euro/year.

Data Preservation after the end of the Research Project

Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?

In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

- All data will be preserved for 10 years according to KU Leuven RDM policy

For every result in this project, a long-term snapshot will be created on the DTAI NetApp storage system (>40TB). This snapshot contains everything to be able to reproduce results with some minor effort. Moreover, it contains the final paper, presentations given as well as any result-specific DMP.

Where will these data be archived (stored and curated for the long-term)?

- Other (specify below)

We also have secure storage via a NetApp storage server in our department >40TB.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Any costs will be covered out of working costs for the project. These were included in the budget. The costs paid for backup during the project should cover the retention period. In case of unexpected costs after the project ends, group reserves can be used.

Data Sharing and Reuse

Will the data (or part of the data) be made available for reuse after/during the project?
Please explain per dataset or data type which data will be made available.

- No (closed access)
- Yes, as open data

(Part of) the software developed in the context of the project may be released as an open-source toolbox. What software will be open-sourced will be decided based on the potential valorization of the project's results.
No raw data will be released.

If access is restricted, please specify who will be able to access the data and under what conditions.

- The proprietary tracking dataset can be used within the research group in the context of different projects.
- All software developed in the context of the project can be used within the research group in the context of different projects.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

Please explain per dataset or data type where appropriate.

- Yes, intellectual property rights

The tracking dataset is and remains the property of the data provider.

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- Other data repository (specify below)

If software will be open-sourced, it will be released on GitHub

When will the data be made available?

- Upon publication of research results

Which data usage licenses are you going to provide?

If none, please explain why.

- Other (specify below)

Any software will be licensed under the Apache License, Version 2.0

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- No

What are the expected costs for data sharing? How will these costs be covered?

No.

Responsibilities

Who will manage data documentation and metadata during the research project?

PI Davis, Pieter Robberechts (current PhD student), PhD student that will be hired on the project. PI Davis is the end responsible.

Who will manage data storage and backup during the research project?

PI Davis, Pieter Robberechts (current PhD student), PhD student that will be hired on the project. PI Davis is the end responsible.

Who will manage data preservation and sharing?

PI Davis, Pieter Robberechts (current PhD student), PhD student that will be hired on the project. PI Davis is the end responsible.

Who will update and implement this DMP?

The PI bears the end responsibility of updating & implementing this DMP.