

Computational Pipelines for Spatio-Temporal Analysis of Team Invasion Games

by

Anj Simmons
BEng, BSc

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Deakin University

June, 2019



DEAKIN UNIVERSITY ACCESS TO THESIS - A

I am the author of the thesis entitled

Computational Pipelines for Spatio-Temporal Analysis of Team Invasion Games

submitted for the degree of **Doctor of Philosophy (Information Technology)**

This thesis may be made available for consultation, loan and limited copying in accordance with the Copyright Act 1968.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: Anj Simmons

Signed: *Anj Simmons*



DEAKIN UNIVERSITY CANDIDATE DECLARATION

I certify the following about the thesis entitled (10 word maximum)

Computational Pipelines for Spatio-Temporal Analysis of Team Invasion Games

submitted for the degree of **Doctor of Philosophy (Information Technology)**

- a. I am the creator of all or part of the whole work(s) (including content and layout) and that where reference is made to the work of others, due acknowledgment is given.
- b. The work(s) are not in any way a violation or infringement of any copyright, trademark, patent, or other rights whatsoever of any person.
- c. That if the work(s) have been commissioned, sponsored or supported by any organisation, I have fulfilled all of the obligations required by such contract or agreement.
- d. That any material in the thesis which has been accepted for a degree or diploma by any university or institution is identified in the text.
- e. All research integrity requirements have been complied with.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: Anj Simmons

Signed: *Anj Simmons*

Abstract

This thesis bridges the gap between raw position sensor measurements of individuals and high level strategic insights about group formations and behaviours. Specifically, it investigates the viability of GPS player tracking data in Australian Rules Football for strategy analysis and shows how characteristics of the sport domain affect the design of the data processing pipeline.

The core contribution of this thesis is a systematic investigation of the pipeline of transformation operations necessary to lift raw GPS player tracking data to a form that can be mined for insights relevant to a sport coach. For each component of the pipeline, the design decisions involved are identified and linked to the types of insights that can be extracted as well as the privacy implications for players. The methodology recognises the role of both manual and automated analysis processes in sport performance analysis as part of the processing pipeline. Data provenance standards are adapted to the context of sport to document the interplay of manual and automated analyses in a manner that is repeatable and auditable.

A review of common methods for data de-identification prevalent in the sport literature shows that: they are not suitable for position data; do not protect individual player privacy against data linkage attacks; and may be in violation of human ethics guidelines. This thesis addresses these shortfalls through a proposal to de-identify position tracking data by transforming them to a point cloud representation (an unordered set of points) thus preventing re-identification of individuals while preserving the ability to perform spatio-temporal team-level analysis. To address the gap between theoretical techniques for privacy preservation and those used in practice, this thesis introduces a software tool and associated interaction model for data sharing that reduces the potential for human errors.

Given the unique nature of Australian Rules Football, there are special normalisation requirements to make effective use of positional data, as each field has a different shape and orientation. This thesis develops a novel technique for marking up coordinate systems at each venue to account for this.

Finally, a computational pipeline for Australian Rules Football is constructed using the techniques proposed in this thesis to demonstrate that they can be combined to provide meaningful team-level insights into formation changes during the game without compromising individual privacy.

Acknowledgements

I would like to acknowledge with particular gratitude the assistance of my supervisors Prof. Rajesh Vasa, Prof. Paul Gastin, and Dr. Scott Barnett. I am also indebted to a number of other people, in particular, Prof. Kon Mouzakis, Dr. Clare MacMahon, Dr. Jacqueline Tran, and Daniel Hoffman who have contributed feedback and advice at various phases of the project. I would also like to thank Hawthorn Football Club and Geelong Football Club for their assistance obtaining the datasets used in this thesis.

Anj Simmons, 2019

Publications Arising from this Thesis

The work described in this thesis has been published as described in the following lists.

Publications highly relevant to this thesis:

1. Andrew Simmons and Rajesh Vasa. “Spatio-Temporal Reference Frames as Geographic Objects”. In: *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL’17*. Redondo Beach, CA, USA: ACM, 2017, pp. 1–4. ISBN: 9781450354905. doi: 10.1145/3139958.3139983
2. Andrew J. Simmons, Maheswaree Kissoon Curumsing, and Rajesh Vasa. “An interaction model for de-identification of human data held by external custodians”. In: *Proceedings of the 30th Australian Conference on Human-Computer Interaction*. Melbourne, VIC, Australia: ACM, 2018, pp. 23–26. doi: 10.1145/3292147.3292207
3. Andrew J. Simmons et al. “Data Provenance for Sport”. In: *arXiv e-prints* (2018). arXiv: 1812.05804 (*Draft*)

Other publications/presentations during candidature:

1. Andrew Simmons et al. "Hub Map: A new approach for visualizing traffic data sets with multi-attribute link data". In: *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. Atlanta, GA, USA, Oct. 2015, pp. 219–223. doi: 10.1109/VLHCC.2015.7357220
2. Andrew Simmons and Leonard Hoon. "Agree to Disagree: On Labelling Helpful App Reviews". In: *Proceedings of the 28th Australian Conference on Computer-Human Interaction*. OzCHI '16. Launceston, TAS, Australia: ACM, 2016, pp. 416–420. ISBN: 978-1-4503-4618-4. doi: 10.1145/3010915.3010976
3. Daniel T. Hoffman, Andrew J. Simmons, and Paul B. Gastin. "Investigating the relationship between injury and match outcome in Australian Football League matches". In: *Proceedings 14th Australasian Conference on Mathematics and Computers in Sport (ANZIAM MathSport 2018)*. University of the Sunshine Coast, Queensland, Australia, 2018, p. 5. ISBN: 978-0-646-99402-4

Contents

1	Introduction	1
1.1	Spatio-Temporal Analysis	3
1.2	Computational Pipelines	4
1.3	Research Questions	8
1.4	Contributions	9
1.5	Thesis Structure	11
2	Background	12
2.1	Competitive Team Sport	13
2.2	Australian Rules Football	16
2.3	Evolution of Player Tracking Devices for use in Australian Rules Football	18
2.4	Chapter Summary	48
3	Modelling	50
3.1	Model of Australian Rules Football	51
3.2	Model of Feedback	55
3.3	Information Theoretic Perspective	57

CONTENTS

3.4 Abstract Data Model	67
3.5 Chapter Summary	77
4 Computational Pipelines	79
4.1 Background	80
4.2 Data Provenance for Sport	96
4.3 Chapter Summary	121
5 De-identification	123
5.1 Introduction	126
5.2 Background	131
5.3 Prevalence of Improper De-identification Methods: A Review of “Non-identifiable” Datasets used in Australian Rules Football Research	143
5.4 Threat Model	157
5.5 Re-identification of Sport Data	160
5.6 Analysis of Trade-Off between Participant Privacy and Data Quality	167
5.7 An Interaction Model for De-identification of Human Data held by External Custodians	170
5.8 Chapter Summary	181
6 Spatio-Temporal Reference Frames	185
6.1 Introduction	187
6.2 Background	192
6.3 Spatio-Temporal Reference Frames as Geographic Objects	204
6.4 Chapter Summary	216

CONTENTS

7 A Platform for Spatio-Temporal Sport Analysis	220
7.1 Pipeline	221
7.2 Exploratory Analysis of Team Shape in AFL using GPS Tracking Data	238
7.3 Feedback from Sport Performance Analysts	256
7.4 Chapter Summary	257
8 Conclusions	261
8.1 Contributions	263
8.2 Applications outside of Sport	265
8.3 Future Work	267
8.4 Closing Remark	270
References	271
Glossary	298
A Modelling	300
A.1 Application of Abstract Data Model	300
B Computational Pipelines	305
B.1 Symbols for Custom Sport Provenance Notation	305
B.2 Effectiveness of visual notation against principles of Physics of Notations	310
B.3 Usability evaluation of VisTrails using Nielsen's top ten heuristics	315
C De-identification	318

CONTENTS

C.1 Detailed analysis of studies claiming use of “non-identifiable” data	318
D A Platform for Spatio-Temporal Sport Analysis 324	
D.1 Data Extraction Requests	324
D.2 Review of Team Shape Metrics found in the Literature .	325
E Human Research Ethics 328	
F Authorship Statements 330	

Chapter 1

Introduction

Contents

1.1 Spatio-Temporal Analysis	3
1.2 Computational Pipelines	4
1.3 Research Questions	8
1.4 Contributions	9
1.5 Thesis Structure	11

Elite sport coaches are always looking for new insights into their team's performance.

The rise of *Sabermetrics*¹ in baseball, originally developed in the 1970's and later popularised in the book and film *Moneyball* [127], is evidence of a historic shift of coaching attitudes. Reliance on coaching hunches and intuition to select and train players has been largely overthrown in favour of data driven decision making approaches based on quantitative performance metrics.²

Traditionally, data driven approaches to sport analysis have been limited to simple mathematical formulas based on manually observable attributes, such as the number of runs made by each player in cricket, or the number of possessions by each player in basketball. However, with the rise of low cost position tracking devices, such as computer vision tracking systems and self-contained GPS tracking units, there has been an explosion in the depth of data available³ which creates opportunities for new approaches to sport analysis.

Currently, these tracking devices are being used by sport performance analysts to analyse individual players. They are used to answer questions such as: *How far did the player run? What was their maximum speed? Considering the player's past movement intensities and durations, at what point should the player be interchanged to prevent fatigue or injury?*

However, teamwork is about more than a group of individual players.⁴ As the devices allow tracking every member of the team, they offer the potential to quantify the team dynamics as a whole. They could be used

¹Phil Birnbaum, "A Guide to Sabermetric Research", Society for American Baseball Research. <https://sabr.org/sabermetrics>

²Leigh Steinberg, 18 Aug 2015, "Changing the Game: The Rise of Sports Analytics", Forbes. <https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/>

³The technological advances that have resulted in rich datasets will be discussed at length in Sec. 2.3.

⁴Matti Clements, "How to get your group to become a team", *Sports Coach*, Australian Sports Commission. https://web.archive.org/web/20151102065854/http://www.ausport.gov.au/sportscoachmag/psychology2/how_to_get_your_group_to_become_a_team

to answer questions such as: *Which formations lead to goals? How do these formations change over time? Are any players out of formation?* [124]

1.1 Spatio-Temporal Analysis

A crucial domain concern for sport strategy, particularly in *team invasion games* such as Association Football (soccer), basketball, and Australian Rules Football, is the control of space on the field over time [46]. There are mature systems for spatial analysis, such as Geographic Information Systems (GIS), as well as for time series analysis. However as space and time are intricately interlinked, seamlessly integrating spatial and temporal analysis as a unified system remains a challenge; areas such as the design of *Moving Object Databases* [179] and design of appropriate visualisation techniques for large scale trajectory data [8] are still an active area of research. The set of analysis methods at the intersection of spatial and temporal analysis are referred to as *spatio-temporal* analysis methods. The application of suitable spatio-temporal analysis methods to the context of sport is the key to extracting insights that can more fully utilise the high-dimensional data generated by position tracking sensors.

The objective of this thesis is to bring spatio-temporal analysis techniques within reach for sport researchers and practitioners so as to facilitate their widespread adoption for analysis of games. Statistical platforms such as *R* contain a large library of existing spatio-temporal algorithms and techniques which are freely available. However, pre-existing spatio-temporal analysis algorithms cannot be meaningfully applied to sport position tracking as-is without first transforming the data into an amenable form. For example, movements during the half-time break need to be discarded, and the field orientation and goal directions need to be accounted for such that the direction of attacking goals is in a consistent direction for each data point. Attempting to apply an existing spatial analysis algorithm without first accounting for this would lead to a meaningless result. As such, a major portion of

this thesis is dedicated to building a solid foundation for transforming raw sport position tracking data into a form that sport practitioners can utilise, such as Fig. 1.1.

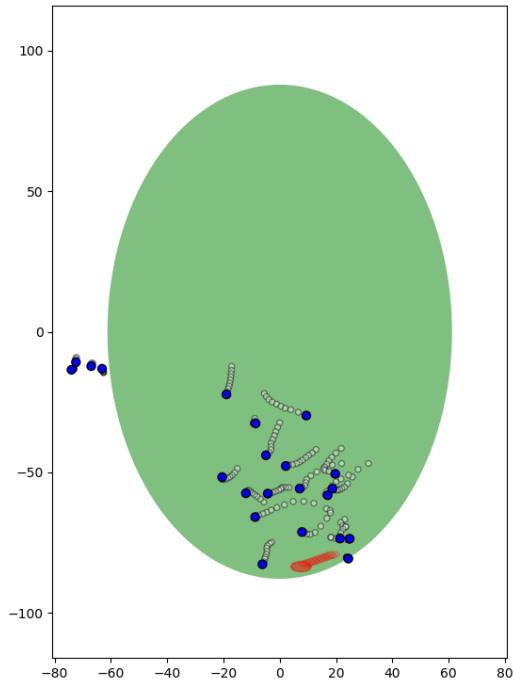


Figure 1.1: Visualisation of team GPS tracking data. The blue dots represent the position of the team. The red oval represents the position of the ball. The trails represent past movements. The group of four players on the left are the interchange/substitute players. Creating this visualisation requires preprocessing steps to reproject the latitudes and longitudes reported by the GPS tracking devices into the coordinate system of the field and to synchronise with the ball tracking data. This forms the foundation for further quantitative investigation of team formations.

1.2 Computational Pipelines

Conducting sport research requires a *workflow* of operations such as player de-identification, data selection, data normalisation, visualisation, and statistical analysis. In contrast to computation of traditional

sport statistics, the additional scope for data exploration permitted by spatio-temporal data sources can lead to complex workflows that become difficult for humans to manage. Even if each operation in the workflow is an automated computer program, remembering which inputs and outputs to feed to each operation and how they connect together can become unmanageable as the workflow grows to integrate additional data sources and analysis outputs. A *pipeline* is a sequence of such operations chained together to automate the entire workflow end-to-end.

Advances in data collection and processing have made highly sophisticated data analysis workflows feasible to facilitate greater depth of data analysis. However, unless the demand for increasingly sophisticated analysis workflows coincides with development of appropriate tools to manage this complexity, there is a risk that the rigour will suffer as a result of others being unable to reliably repeat the process. *Nature* has reported growing evidence of a “reproducibility crisis” [13] wherein scientists in every discipline have admitted to being unable to reproduce other’s experiments, or even their own work. Scientific fields that deal with analysis of large volumes of data, such as bioinformatics, have recognised the need for development of automated *computational pipelines* [87] and *workflow management systems* [164, 107] to help ensure that the analysis process is repeatable.

Tools originally developed to allow creation of complex bioinformatics pipelines (e.g. analysis of DNA sequences) include Apache Taverna [164, 107] and Python Ruffus [87]. In principle, these tools can also be used as general frameworks for data processing and analysis in other domains such as sport. However, it is well known amongst designers of Domain Specific Languages (DSLs) that generality and flexibility run counter to the needs of supporting concise expression of a particular domain [215]. Thus, while a range of domain specific and general frameworks for constructing pipelines exist, new frameworks and components are needed to build a software ecosystem for spatio-temporal analysis in the sport domain.

Being able to reproduce the end result of a process is not enough—

auditing complex processes to validate and debug them requires a record of *data provenance* (i.e. the ability to trace the flow of data, particularly at the boundaries of sub-systems). Data provenance gives analysts the ability to trace the flow of data backward so they can answer questions about the data sources and specific data records that were used to calculate the results of an analysis. For example, if a sport practitioner arrives at an unexpected result, a player or coach may request to see the underlying video sequences for events in the game that contributed to that result.

Common data operations required for spatial analysis of sport include: de-identifying players; normalising spatial data with respect to the playing field; mining spatio-temporal patterns; and aggregating and visualising the results. Each of these is a form of *data transformation*. Each transformation is introduced below, and an explanation is given of how the transformation is interwoven with the concerns of the sport domain:

De-identifying players While it is trivial to partially de-identify sensitive player data by replacing player names with unique codes, ensuring that the data are truly non-identifiable requires consideration of all possible linkages to other sport datasets that could be used to re-identify a player. Specifically, when de-identified players are associated with movement data, there is a high risk that the player could be re-identified through fingerprinting techniques that correlate game events detected in the movement data (such as kicks, tackles or interchanges) with public match feeds and televised match video footage that contain the player name alongside the actions they performed.⁵

Normalising spatial data When dealing with spatio-temporal datasets where local movements are more important than global movements, one should consider normalising the data with respect to a reference frame. In sport, there is a need to normalise with respect to the playing field. Specific types of sport may have additional requirements that affect how the normalisation is performed. For

⁵The importance of properly de-identifying sport data to protect player privacy, and the unique challenges to the de-identification process posed by player tracking data are discussed in detail within Chapter 5.

example, in Australian Rules Football, each field has a different size, shape, and orientation that need to be accounted for.

Aggregating data Realising the promised potential of automated team-level spatio-temporal analysis requires a sufficient number of scoring formations to detect subtle patterns in a statistically sound manner. Even in Australian Rules Football—which is considered a fast paced sport—each team scores only 25 times⁶ (on average) over the course of any given match. Thus it is necessary to aggregate data from many different sport fields into a large unified dataset that can be mined for patterns.

Mining spatio-temporal patterns In sport, the spatial movements of players are related to the “phase” of the game (e.g. defence/offence). Thus spatio-temporal analysis cannot be conducted in isolation; detected patterns need to be analysed within the context of transitions between phases of play.

Visualising results Visualisation is also a form of data transformation [113]. Visualisations need to consider the domain in order to optimally map data attributes to visual attributes [144]. Specifically, visualisations are needed for presenting de-identified Australian Rules Football player position tracking data in a form interpretable by sport practitioners that permits exploration and comparisons of different matches.

While the transformations described above may at first appear to be common data operations available in existing data analysis software, generalised approaches fail to adequately address the specific needs of sport analysis and thus suffer from the “leaky abstraction”⁷ problem.

⁶Average number of scoring events derived from 1897-2018 match data from afltables.com, retrieved 2018-05-21

⁷In software engineering, “abstraction” refers to the process of generalising concepts to allow them to be used at a high-level without the need to concern oneself with all the low-level implementation details (for example, one may abstract away the details of how GPS devices work and treat them as a source of latitude, longitude readings). However, when abstractions fail to capture essential aspects of the real-world, these details inevitably “leak” (for example, GPS devices often experience interference, particularly near the edge of sport stadiums [218], which needs to be understood and accounted for in the analysis). The term “leaky abstractions” was coined in a blog post by Joel Spolsky, 2002, “The Law of Leaky Abstractions” <https://www.joelonsoftware.com/2002/11/11/the-law-of-leaky-abstractions/>

Investigating these operations within the context of sport analysis uncovers new concerns which may have been overlooked by generalised approaches that only deal with data transformation and integration at a high level.

1.3 Research Questions

This thesis seeks to addresses the following questions:

1. *Can team-level GPS analysis provide useful information to sport researchers and practitioners beyond what they already know from manual observation, video analysis, traditional statistics, and (individual) player GPS monitoring?*
2. *How can GPS player tracking data be processed to extract meaningful team-level insights without compromising individual privacy?*

To address these questions, the thesis systematically investigates each layer of software architecture required to support the analysis of high-dimensional spatio-temporal data in sport. This synthesis approach is necessary to ensure the properties of reproducibility, traceability and statistical soundness are preserved at each layer. The research outlined in each chapter is brought together through a case study in the design of a pipeline to analyse the shape of team formations in Australian Rules Football. As the analysis of Australian Rules Football is a quantitative observational study, it cannot establish causality without additional assumptions; however, it may still highlight patterns of interest to sport performance staff for further investigation. While the investigation of Australian Rules Football forms a minor contribution in itself, it is primarily intended as a demonstration of the larger framework introduced in this thesis.

1.4 Contributions

This thesis makes the following contributions:

Elaborated in Chapter 3:

1. Structures Australian Rules Football jargon into a formal domain model of consistent terminology, and uses this to identify variables that form part of the game state. The full list of identified variables provides a holistic understanding of game state, and can increase awareness of the simplifying assumptions made by current sport analysis models.
2. Applies information theory to sport in order to provide a mathematically rigorous perspective for understanding the role of sport performance analysis systems within the larger sport context. Information theory is used to formalise the objective of performance analysis systems into a single formula, which states that the goal is to ensure information is valuable yet not already known to a coach, and incorporates the need to transmit this over limited human information channels.
3. Provides an abstract data model that permits modelling both dense and sparse spatio-temporal sport datasets, and draws attention to all required accuracy attributes that need to be specified in order to reason about the confidence of interpretations made from the dataset.

Elaborated in Chapter 4:

4. Provides an analysis of the data provenance needs of the sport domain, and evaluates existing data provenance tools against these criteria. A customised data provenance notation for sport is proposed in order to ease uptake for sport performance analysts without a computer science background.

Elaborated in Chapter 5:

5. Exposes the prevalence of improper de-identification methods used in sport research, and demonstrated that GPS player tracking data is particularly prone to re-identification. An interaction model is proposed to help improve ethical conduct of research by allowing the researcher to specify the de-identification operations in cases where the data custodian lacks the technical resources to strongly de-identify data themselves prior to data hand-over. The proposed approach was applied to GPS player tracking data held by an Australian Rules Football club to obtain the non-identifiable data used in this thesis.

Elaborated in Chapter 6:

6. Proposes a novel method for representing spatio-temporal reference frames as geographic objects. This allows GIS novices, such as sport performance analysts, to configure reference frames without the need for deep conceptual knowledge of cartographic projections. It also facilitates partial automation (e.g. reprojecting GPS data to the closest sport field), thus resulting in time savings when the analysis involves multiple reference frames (e.g. a season of GPS tracking data involving multiple sport fields).

Elaborated in Chapter 7:

7. Develops a platform for spatio-temporal analysis of team sport, demonstrated through Australian Rules Football as an example. Feedback from an elite Australian Rules Football club confirms that the techniques proposed are likely to offer useful insights.
8. Performs an analysis of team shape and game speed in Australian Rules Football. Team spread was found to strongly correlate with game speed.

1.5 Thesis Structure

The background chapter (Chapter 2) outlines the history of technological developments in player tracking, as well as the types of analysis made possible by these technologies. This is followed by models of the sport domain and sensor data (Chapter 3) to establish terminology and to clarify the objective of sport data analysis systems from an information theoretic perspective. The subsequent chapters examine data provenance (Chapter 4), GPS data de-identification procedures (Chapter 5), and spatio-temporal transformations (Chapter 6) that are necessary to accurately convert raw GPS data into a more accessible form. This serves as a platform for investigating team formations presented in Chapter 7. The work is applied in the form of a team GPS analysis tool that was presented to elite Australian Rules Football performance analysts in order to validate the usefulness of the approach to professional sport practitioners. Finally the thesis concludes with findings and future work in Chapter 8.

The initial chapters (Chapter 3, Chapter 4, Chapter 5, Chapter 6), while motivated by the needs of sport performance analysis, are written from a computer science / software engineering perspective. In contrast, Chapter 7 integrates the chapters together through application to sport performance analysis, and contributes a (minor) advancement of the sport science field.

Chapter 2

Background

Contents

2.1 Competitive Team Sport	13
2.2 Australian Rules Football	16
2.2.1 Game terminology and structure	16
2.2.2 Comparison to other sports	18
2.3 Evolution of Player Tracking Devices for use in Australian Rules Football	18
2.3.1 Structure	18
2.3.2 Traditional Sport Statistics	19
2.3.3 Sport Event Data	24
2.3.4 Spatio-Temporal Data	30
2.3.5 Gaps	46
2.4 Chapter Summary	48

This thesis opened with a practical motivation, namely to use spatio-temporal analysis techniques to deliver elite sport coaches new insights into their team's performance. However, not all information is necessarily insightful to coaches. As such, this background chapter begins by examining team sport from a coaching perspective, and showing where sport analysis—and by extension, automated sport analysis—fits within this larger context. The results of spatio-temporal sport analysis, regardless of how deep or sophisticated, are only insightful if they can provide the coach and other sport practitioners with meaningful information relevant to their objectives.

Following the high level coaching perspective, the background narrows to examine the specific sport of Australian Rules Football as played at the elite level in the Australian Football League (AFL). AFL is the main application throughout the rest of this thesis. However, the methods are, in principle, transferable to other team invasion sports.

Finally, the chapter conducts a review of the technological and analytical developments that have led to the rise of spatio-temporal data in Australian Rules Football.

2.1 Competitive Team Sport

Sport coaches are tasked with optimising the performance of their team. In two player games, such as chess, *strategy* describes the policy that a player uses to select an *action* in response to the current game situation. In team games, the strategy is a property of the team, and the team action refers to the actions of all team players. Examples of strategy include: decisions about when to interchange or substitute players (a decision traditionally made by the team coach), which team member to pass to (a decision typically made by the player with the ball), or decisions about how to structure defence formations (a combination of top-down commands from the coach combined with individual player decisions about how to align themselves relative to other players).

The coach cannot directly control their players during the game, instead they must teach their players the skills and tactics that they need. For example, in Australian Rules Football, when the game is in play, the coach can only communicate with their team through sending out “runners” onto the field to relay messages between the coach and the players, thus this form of communication is only suited as a means of feedback to inform future play rather than direct instruction. Even during training sessions, modern coaching practices avoid direct instruction of how to perform skills, and instead provide a feedback system that allows the players to learn from their own experiences. This is because directly instructing players how to perform skills leads to “paralysis by analysis”, particularly when transferred to scenarios involving stress or anxiety [195], and as such, best coaching practice is to create an environment where “verbal, conscious attention to task rules and procedures is minimised rather than maximised”¹.

For closed skills, such as Olympic target shooting, sport scientists have run empirical studies to determine the best type of feedback that will allow the athlete to improve as quickly as possible. For example, in rifle shooting, an intervention study was performed to evaluate the efficacy of feedback delivered through visualisations to inform an athlete of rifle barrel movement, and feedback delivered through auditory sounds to inform an athlete of the magnitude of their aiming error [143]. Coaches also provide players with video feedback that displays the player performing a skill superimposed against their past performance, or the “ideal” performance of an expert. However, comparisons against ideal performance can be problematic, as players should be encouraged to follow their own style. One famous example demonstrating the importance of personal style over comparisons to the commonly accepted ideal, is the invention of the “Fosbury Flop” [14] high jump style which radically challenged what was considered the ideal style at the time.

In contrast, team invasion games [217] involve open skills [223] taking place within an unpredictable environment, and the effect that a

¹Bruce Abernethy, “Theory to practice - Sports expertise” in *Sports Coach*, Australian Sports Commission. https://web.archive.org/web/20141114115305/http://www.ausport.gov.au/sportscoachmag/skill_analysis2/sports_expertise_from_theory_to_practice

player has on the team performance cannot be directly measured, as it depends upon the behaviour of the rest of the team, as well as the behaviour of the opposition. For example, standing in the correct position may deter the opposition from scoring, even if the player never comes in contact with the ball. A common truism is that the team performance is more than the sum of the individual performances. Sport psychologists describe a team using factors such as “team cohesion”,² thus there are properties that are best modelled as belonging to the team as a whole rather than to individuals.

Traditionally, coaches have relied on a range of summary statistics (such as number of successful passes) as heuristics for evaluating player performance. However, relying on summary statistics without taking into account their context can cause misleading results, as they do not fully control for all aspects of game play needed to describe the situations under which the attempts took place. For example, players playing against a more difficult opposition may find passing more difficult, but the summary statistics do not account for this. In particular, as teams may reserve key players for more important or difficult games, this can lead to confounding effects.

Spatio-temporal sport analysis techniques offer the ability to better account for the game situations that players find themselves in, thus allowing coaches to provide players with higher quality feedback. Spatio-temporal sport analysis can also be used as a means to investigate the team as a superorganism [64], thus allowing the coach to better understand how the interactions between players affect the overall team behaviour, and to refine the team playing style based upon this knowledge.

²Matti Clements, “How to get your group to become a team”, *Sports Coach*, Australian Sports Commission. https://web.archive.org/web/20151102065854/http://www.ausport.gov.au/sportscoachmag/psychology2/how_to_get_your_group_to_become_a_team

2.2 Australian Rules Football

The last section described competitive team sport at a high level from the coach's perspective. This section looks specifically at Australian Rules Football—more commonly known as AFL (Australian Football League) when played by the official rules at the national elite level—which is the sport analysed throughout the rest of the thesis. The purpose of this section is to introduce common game terminology and structure. As the regulations are revised each year by the AFL Commission, the 2015 rules will be described for consistency with the period of the dataset used later in this thesis. Note that a deep knowledge of the game is *not* necessary to read this thesis, and additional details will be provided throughout the thesis as required.

2.2.1 Game terminology and structure

Australian Rules Football is a game played between two competing teams with an oval ball on an oval field. Each team is allowed 18 players on field, three interchange players off the field, and one substitute³ player.

Goal areas are at the two far ends of the field, and each goal area is marked by two tall inner goal posts and two shorter outer posts. Each team is assigned an attacking goal area at the start of the match, such that teams aim to move the ball toward opposite ends of the field. Players may kick, catch, fist, pick-up, and hand-ball the ball to move it towards the goal area they are attacking. Teams are awarded points for kicking the ball between goal posts. The number of points awarded depends upon the accuracy of the kick: 6 points for a “goal” between two inner goal posts, and 1 point for a “behind” when the ball misses the inner goal area and instead goes between two outer posts. The team with the largest number of points at the end of the match wins. Draws are possible, but rare, due to the high-scoring nature of Australian Rules

³Under 2015 rules

Football.⁴ The results of multiple matches throughout a football season determine the ranking of the teams, culminating in a finals knock-out style tournament amongst the top teams to determine the team that wins the season.

The match opens with a “centre bounce” in which an umpire throws the ball vertically downward within the centre of the field such that it bounces up into the air marking start of play, and ruck players attempt to hit it towards their own team. When a team scores a goal, play is stopped, and the ball is returned to the centre of the field for the next centre bounce to restart play. In the case of a behind, the opposition is given possession of the ball and play restarts with a “kick-in” from the end of the field. A match consists of four quarters, with a short break between each for players and spectators. At the end of each quarter teams switch goal directions.

Interchange and substitute players sit on an interchange bench at the side of the field. Player interchanges can occur at any time during the game, although the number of interchanges allowed per game is capped to 120⁵ for each team. The substitute player is intended as a replacement in the event of player injury, but is often used for strategic reasons instead.⁶

The rules permit a range of field sizes, so each field may have unique characteristics. The oval shape of the ball introduces an unpredictable element in the way that the ball bounces.

⁴To date, there have only been 158 draws in the entire recorded history of AFL, which represent just 1.03% of all matches. Calculated from 1897-2018 match data from afltables.com, retrieved 2018-05-21

⁵Under 2015 rules. <http://www.afl.com.au/news/2015-09-03/sub-rule-abolished-interchange-cap-reduced>

⁶Recent rule changes remove the dedicated substitute player role and instead have a fourth interchange player to be used at any time during the game

2.2.2 Comparison to other sports

Australian Rules Football is classified as a *team invasion game*. Other games in this category include Association Football, American/Canadian Football (Gridiron), Rugby League/Union, Polo / Water Polo, Lacrosse, Netball, Basketball, and Hockey / Ice Hockey [217]. Invasion games are distinguished from the other classes of games by their sophisticated attack and defence tactics as each team attempts to claim territory from the other team.

In contrast to Association Football (informally known as “Soccer” in Australia, or just “Football” internationally), scoring events in Australian Rules Football are frequent. In contrast to American Football, the rules of Australian Rules Football do not enforce any particular team formations or player roles. Thus players are free to adapt to different roles and positions during the course of the game.⁷

2.3 Evolution of Player Tracking Devices for use in Australian Rules Football

2.3.1 Structure

The main sport focused on in this thesis is Australian Rules Football. However, as Australian Rules Football is rare outside of Australia, this review also draws on studies across a range of sports and technologies to ensure an international perspective. For each sport technology, a short chronology of the rise of that particular technology is provided, followed by the analysis techniques enabled by that technology. Oc-

⁷Up until 2018. Rule changes proposed for the 2019 season intend to restrict players to certain zones based upon their role in order to reduce congestion (i.e. the tendency for all players on the field to chase after the ball, resulting in high player densities that slow play and can result in injuries) <https://www.adelaidenow.com.au/sport/afl/more-news/the-afl-could-introduce-zones-and-startingposition-rule-changes-in-2019/news-story/adeb3d1cbaf21ba09cf0bf6b3c0e9114>

casionally, analysis techniques for a particular data type may predate the technology necessary to collect that data, either through theoretical analysis of hypothetical data, or through manual data collection. This review starts with traditional sport statistics, and works through to contemporary technology that continuously tracks every player's every movement.

The review finds that as positioning technology has matured, the measurement errors have reduced from orders of meters to a few centimetres. As a result, the emphasis has shifted from the technology itself to how to make use of the collected data. The review concludes with gaps within the literature that have not yet been fully explored.

2.3.2 Traditional Sport Statistics

Origins of Traditional Sport Statistics

This section presents a brief chronology of traditional sport statistics. Key dates leading up to the collection of detailed AFL statistics are listed in Table 2.1. Full timelines of contemporary developments will be provided in subsequent sections.

Table 2.1: Timeline of Traditional Sport Statistics

Date	Event
1859	Henry Chadwick invents the modern baseball “box scores”
1897	AFL statistics collected since formation of Victorian Football League.
1931	Detailed AFL statistics collected by newspapers

The demands of sport fans for a detailed delineation of sport games has driven sport data to be collected for centuries. Baseball is particularly well-known for its obsession with statistics. Henry Chadwick is widely

credited as developing the modern baseball “box scores” in 1859, which record “Runs”, “Hits”, “put-outs”, “assists”, and “errors” for each player in the team.⁸

Australian Rules Football was introduced in 1858.⁹ Official game outcomes have been collected since the formation of the Victorian Football League in 1897. Detailed statistics for each player such as “kicks”, “marks”, “handballs”, “frees for/against”, “goals”, “behinds”, and “misses” have been collected by newspapers since circa 1931 [100].

Analysis of Traditional Sport Statistics

Coaches use the statistics collected during the game for the purposes of evaluating players. For example, when selecting forward players, a coach will consider which player has had the highest accuracy kicking goals in previous games. However, these statistics are limited, because they do not account for the situation under which the goal was attempted. For example, the player may have been under pressure from the opposing team to dispose of the ball rapidly, or the situation may have demanded that the player make an attempt from a more difficult angle than normal. Furthermore, these types of statistics have the problem that they may incentivise players to act in ways that enhance their own statistical record at the expense of the team. For example, a player that attempts to kick a goal in order to increase their own goal-kicking statistics instead of passing the ball to a nearby team player with a better position.

The summary statistics for each player are sometimes referred to as Key Performance Indicators (KPIs), mimicking the concept of Key Performance Indicators in business environments. Both football and businesses alike suffer from the issue of seemingly valid KPIs that can have unintended “perverse incentives” [165] if not used carefully. The per-

⁸M. Pesca, “The man who made baseball’s box score a hit,” 2009. <http://www.npr.org/templates/story/story.php?storyId=106891539> Accessed: 2016-03-22

⁹AFL Commission, “History,” 2013. <http://www.afl.com.au/afl-hq/the-afl-explained/history> Accessed: 2016-03-23

verse incentive issue is a consequence of the principal-agent “incentive problems” [176] in economics, which deal with the issue of designing reward systems that align the agents’ (the AFL players) interests with that of the principal (the coach). The field of Mechanism Design (reverse game theory) defines a similar concept called “incentive compatible” that deals with creating reward systems that encourage agents to truthfully report their situation [212]. In *the five fundamentals of modern football*, coach Danny Ryan highlights best practice for KPIs in sport by using a carefully selected subset of KPIs as a proxy measure of the team’s adherence to an agreed upon playing style [175] rather than focusing on any particular KPI itself.

An interesting statistic widely used in hockey is the plus-minus statistic [116]. This statistic awards a point to all players on the field whenever their team score a goal, and subtracts a point from all players on the field whenever the opposition scores a goal. A clear advantage of this statistic is that unlike the statistics discussed earlier, the plus-minus statistic creates a rational incentive for all players to honestly co-operate if they wish to maximise their own individual plus-minus scores. This property is of importance even when team members are of sufficient character that they would not deliberately harm the interests of the team. The plus-minus scores can reveal the subtle influence of players on the team through invisible unintentional actions or inactions. For example, a team member may passively deter the opposition from scoring, simply by their presence occupying a strategic location. Due to the complexity of team games, a coach, and even the player themselves, may be otherwise unaware of how these actions are affecting the game.

Unfortunately, the plus-minus statistic is known to suffer from issues relating to noise [88]. A large number of games are necessary for the patterns to emerge. For small volumes of data the plus-minus statistic may wrongly accuse players of not contributing to the team. Furthermore, the plus-minus statistic suffers from statistical confounding effects. For example, coaches may conserve their strongest players for playing in games against the most difficult opposition. Thus there is a correlation between stronger players and a more difficult opponent, which in turn is correlated with lower team scores. This could lead the

plus-minus statistic to wrongly give low scores to strong players, as it does not correct for the stronger defence.

A logistic likelihood model that considered both the opposition and the pairings with team-mates was suggested as an alternative by Gramacy et al. [88]. However, to obtain sufficient data to fit their model, the authors of the study assumed the player ability to be constant over four seasons. Clearly such an approach is ineffective for player feedback purposes, as any feedback system must track the change of skill over time, and feed this back to the player with as little delay as possible.

The large volume of data collected on baseball facilitated the rise of “Sabermetrics”¹⁰ introduced by Bill James, which is the practice of defining “advanced metrics” intended to objectively quantify aspects of a player’s performance. Many of these so-called “advanced metrics”¹¹ are trivial formulas that can be calculated directly from the underlying box scores. Building upon the ideas behind Sabermetrics, professional baseball teams have used economic principles to evaluate the value per dollar salary that players bring to the team, as popularised by the book/film *Moneyball* [127]. Baseball fans have now devised a wide diversity of advanced metrics. However, it is questionable how useful these metrics are for coaches and players. While framed in terms of the (positivist) scientific method, many seem to be irrelevant from a performance perspective,¹² and are perhaps best understood as cultural products through which certain subgroups of fans choose to express their identity [152].

One suggestion to deal with the proliferation of metrics is to evaluate them by their ability to predict future game results [103]. Predictive metrics are of interest to gamblers, and have been studied by

¹⁰J. Albert, “An introduction to sabermetrics,” 1997.

<http://www-math.bgsu.edu/~albert/papers/saber.html> Accessed: 2015-09-01

¹¹For a community list of advanced metrics see <https://en.wikipedia.org/w/index.php?title=Sabermetrics&oldid=713461203#Examples> inspecting the article for each advanced metric reveals that most advanced metrics are defined in terms of the box score.

¹²For example, the “NERD” metric attempts to estimate the *aesthetic* value of the game [https://en.wikipedia.org/wiki/NERD_\(sabermetrics\)](https://en.wikipedia.org/wiki/NERD_(sabermetrics))

researchers under the topic “inefficiency of betting markets”¹³. Swinburne University emeritus statistics professor, Stephen Clarke, has been heavily involved in AFL prediction, with predictions published on Swinburne’s website¹⁴ to date. Clarke’s predictions use a simple exponential smoothing filter (moving average with more weight on recent games) applied over the historic scores of each team [39]. The detailed game statistics and players lists are not used in the prediction at all. Despite the simplicity, Clarke’s predictions have been able to exceed human predictions by experts [39].

In chess, the ELO¹⁵ rating system is used as a measure of a player’s ability. The ELO rating system is purely based on the results of a player’s past games, and the ELO score of their opponents. Similarly to Clarke’s AFL predictions that use an exponential smoothing filter to update a team’s predicted ability after each game, the ELO rating system includes an update-step to slightly increase or decrease a player’s rating after each game. In the ELO rating system, the size of the increase is proportional to the unexpectedness of the win. For example, if a strong player (with a high ELO rating) beats a weak player (with a low ELO rating), neither of their ELO scores will change much, as little information is learned from the encounter. However, if a weak player beats a strong player, the weak player’s ELO rating will rise rapidly, and the strong player’s rating will decrease, reflecting the significant “upset” of the ELO prediction system.

The ELO rating system inspired the development of the “Glicko rating system”, and more recently, “Microsoft TrueSkill” [96]. Microsoft TrueSkill predicts player ability for the purpose of finding a fair match for XBox (computer game console) gamers. Unlike the ELO rating system, TrueSkill models a player’s ability as a Gaussian distribution consisting of both an expected value, and a standard deviation to represent the uncertainty. Bayesian logic is used to update the player’s predicted ability after each game against an opponent, causing the spread of the distribution to reduce over time as the system infers more about the

¹³S. Clarke, “Want to win at gambling? Use your head,” *The Conversation*, Jun. 2011.

¹⁴<http://www.swinburne.edu.au/research/our-research/footy-tips/>

¹⁵Named after Arpad Elo, who developed the ELO rating system

player's ability. Whilst predominantly designed for free-for-all games (no cooperation between players), TrueSkill is also capable of estimating player ability in team-games, but requires a much greater number of matches for the uncertainty to reduce.

In summary, the history of sport statistics dates back centuries, highlighting the demands of fans and coaches alike for quantitative data on the game. However, traditional sport statistics suffer from confounding effects as they do not properly control for the conditions under which the events took place. Traditional sport statistics are useful for measuring and predicting (as evidenced by betting models) the overall performance of the team; however, they often do not contain sufficient detail to be useful for extracting insights into the underlying causes.

2.3.3 Sport Event Data

The last section concluded that traditional sport statistics, whilst useful for summarising long-term performance, do not contain sufficient detail to investigate the underlying causes for a team's performance. Understanding these underlying causes requires investigation of the characteristics of the individual events (kicks, handballs, etc.) that occur during a game. However, the rapid pace of games, and the subtleties of interaction makes it a non-trivial task to collect this kind of data. To do so requires a carefully designed schema for recording these data. The historical developments towards detailed sport event databases suitable for automated analysis are presented visually as a timeline in Table 2.2.

Development of Sport Event Databases

Table 2.2: Timeline of Databases in Sport

Date	Event
1931	Messersmith describes tool to help determine distance traversed by basketball players, and notes this at 2 minute intervals.
1958	Donald Knuth writes computer program to analyse his university basketball team.
1970	Downey designs Notational Analysis system for manually collecting detailed information about tennis.
1985	Hughes uses digitised data to compare performance of squash players.
1985	Jon Patrick proposes CABER project for design of an AFL database.
1996	Champion Data builds AFL database for recording time-stamped events for every transition of the ball.

Notational Analysis is a field of sport science that seeks to quantitatively capture the detailed movements and interactions of players in order to understand and improve player performance. In contrast to biomechanics, which primarily deals with studies of ‘closed skills’ of an individual under controlled conditions, the field of notational analysis includes the study of the interactions of players in team games [104].

One of the earliest [105] applications of notational analysis to sport was conducted by Lloyd L. Messersmith in 1931 [139], who created a scale model of a basketball court and a small electronic trundle wheel to help record the distance each player moved. Observers would manually trace the path of players on the scale model using the electronic trundle wheel. The electronic distance measurement worked due to a pattern of alternating conductive and insulated portions along the circumference of the trundle wheel, which would create electronic pulses as the wheel

rotated when a current was run through the wheel and into the base of the conductive model. Messersmith had observers use the scale model to record the distance each basketball player ran each 2 minute interval of the game, and noticed that the distance run during intervals at the start of the game was slightly further than the distance run during intervals at the end of the game. Messersmith also broke the distances down by possession, and found that players ran further in offence than defence.

In the late 1950's, Donald Knuth (now a renowned computer scientist), an undergraduate student at Case Institute of Technology at the time, served as manager of his university's basketball team. Combining his interest in computers with his role as manager of the team, he designed a system to analyse the performance of the players within the team. A spotter would observe each ball possession made and relay information about how that possession was spent, for example, whether the player: added value by stealing possession of the ball from the opposition; lost value by fumbling the ball, resulting in a turnover to the opposition; or converted the possession into a basket scored. Knuth then entered these data into a computer via punch cards. The computer then calculated the estimated value each player was contributing to the team using a formula Knuth devised. The key insight that differentiated Knuth's system from traditional sport statistics was his use of detailed raw data to infer each player's real contribution to the team rather than just tallying the number of interactions they had with the ball. Knuth's formula for the "true point contribution" [118] is documented, but the original program has been lost. No evidence was offered of the system's efficacy beyond anecdotal praise from the team coach and a (possibly coincidental) improvement in the number of games won. A promotional one-minute video of the system was created by IBM in 1959 and has been republished online by the Computer History Museum¹⁶, but appears to have gone largely unmentioned in the literature other than as a historic side-note in the history of computing. Similar systems did not

¹⁶[Video] IBM, "The Electronic Coach", Computer History Museum, 1959. <https://www.youtube.com/watch?v=dhh8Ao4yweQ> Accessed: 2017-03-20. Knuth later recounts the experience in [Video] Web of Stories, "University life: my basketball management system" in "Donald Knuth Interview 2006", 2006. <https://github.com/kragen/knuth-interview-2006> Accessed: 2017-03-20

emerge until decades later when computing became more accessible to sport performance analysts.

Much of the literature on notational analysis focuses on racket sport [106]. In 1970, Downey [62] invented a hand notation system (a set of symbols) for manually recording the details of shots in tennis, such as the type of shot, whether the shot was straight/diagonal, whether the shot was backhand/forehand, and the type of ball spin. Processing data collected through hand notation systems was cumbersome, so not well suited for long-term studies. This changed with the introduction of computer analysis systems, such as demonstrated by Hughes in 1985 [102] for comparing the performance of squash players.

Databases for AFL have been proposed as early as 1985 [166]. Former AFL player, Ted Hopkins, and his partner Angelika Oehme founded Champion Data in 1995. Champion Data has maintained a detailed database of AFL since starting operations in 1996.¹⁷ The database contains qualitative data (such as whether a turnover was “hard” won from the opposition, or obtained by collecting a “loose” ball off the ground) on every event that occurs in AFL. An operator enters the data live during the game (under supervision of a “backup-caller”), and the data is broadcast to coaches, commentators, and media.¹⁸

Analysis of Sport Event Data

Sport event data are sparse time-coded events representing transitions of the game between various states. Some of these states (such as ball out of bounds) are explicitly stated in the game code, whereas other states are physical or conceptual (such as being in an open position out of reach of opposition players). The nature of the sport is such that certain aspects of the game are virtually unpredictable (especially in AFL in which the oval shape of the ball causes it to bounce unpredictably). To analyse these data requires techniques that model the

¹⁷Champion Data, “HISTORY,” 2016. <https://www.championdata.com/index.php/champion-data/history.html> Accessed: 2016-03-23

¹⁸Stathi Paxinos, “Why statistics rule,” The Age, Jun. 2004.

game as stochastic system of transitions between states.

Markov models are a common technique for modelling stochastic systems, and have found applications in a diversity of fields from customer loyalty studies through to animal behaviour studies. A Markov model describes the system as a set of possible states, and specifies transition probabilities between states. A constraint assumed by Markov models is the *Markov property* that requires the transition probabilities to the next state to be conditioned solely upon the current state, and to be independent of the long-term past history of the system. Whilst this condition is seemingly limiting, the researcher can often define states so as to ensure this property holds. For example, if state transition probabilities depend on the last two events, then the researcher may look for a ‘second order’ Markov model which defines its states as the set of all possible pairs of the last two events. Furthermore, Artificial Intelligence (AI) researchers have found that even when the Markov property does not strictly hold, it may still offer an acceptable approximation of the system [202].

In 1996, McGarry et al. analysed squash using a Markov model [135]. Squash has developed terminology for a range of common shot-types representing the different ways a player can hit the ball to the opponent (by bouncing the ball off different parts of the surrounding walls). The researchers chose to use the combination of shot-type and player-turn to define the states in the Markov model, and tabulated the transition probabilities to each type of response shot-type conditioned upon the received shot of the player. Additional states were introduced to represent the ways the rally could end. The researchers’ aim was to use the resultant Markov model in order to build a behaviour profile of the players. However, they found that the transition probabilities varied depending on the opposition player challenged, so were unable to build a consistent profile over multiple matches. Furthermore, they did not attempt to validate that the Markov property held for their model; i.e. that the response shot is purely dependent on the shot received, and not the past history of shots. The researchers have since provided an alternative explanation of squash as a dynamical model, which will be discussed later in the section on Analysis of Spatio-Temporal Data.

In 2002 Hirotsu and Wright created a simple four state Markov model for analysing Association Football [97]. The states were possession (by each of the two teams) and goal-kickoff (by each of the two teams). Whilst no database was available at the time, Hirotsu and Wright were able to derive the probabilities for each transition from game records in the analysed team’s “yearbook”. The resultant Markov models can be used to perform “what-if analysis”. For example, suppose the coach knows how the transition probabilities will change with the substitution of a certain player. The coach can simulate the game using the Markov model to consider *what-if* the player were substituted at any given future time. This can be combined with dynamic programming (backward induction) to find the optimal time to perform the player substitution in order to maximise the expected chance of winning.

The advent of computer databases containing many past games has allowed researchers to develop more sophisticated models. Forbes’ 2006 thesis [76] presented an 18 state Markov model for AFL. Forbes partitioned the AFL field into three zones, and modelled separate states for each combination of team possession, pass type (kick, handball, etc.) and zone of the field. Forbes noticed differences between the transition probabilities for each team, and related these back to the qualitative playing style of each team.

Social scientists have used social network analysis to examine social structures. More generally, the field of Complexity Science deals with understanding emergent unintuitive phenomena that result from simple interactions. Researchers have subsequently attempted to apply these concepts to sport. Duch et al. have applied this to the player passing distributions in Association Football, and showed that the network “flow centrality metric” corresponds to subjective ratings of player ability [65]. Cotta et al. analysed the 2010 FIFA World Cup by manually encoding events from video [44]. Neither of these network analysis papers attempted to provide predictions; however, they do provide insight into each team’s strategy. A recent (2018) paper by Braham & Small [25] claims to be the first published network analysis of AFL passing structures. The “mean betweenness centrality” metric computed for the previous match was found (among other network metrics) to be a pre-

dictor of the outcome of the next match. Furthermore, it was shown that it would have resulted in a profit if betted on, an indicator that it contains information not present from other sources. The ability to predict game outcome with >50% probability was shown to be statistically significant; however, the betting profit (a strong test of ability to offer predictive information beyond what is already known) was only calculated over a single season and not tested for statistical significance.

The depth of analysis made possible by sport event data is an improvement over traditional sport statistics, as event analysis can model in-game scenarios rather than summarising overall trends. The use of sport events databases has allowed sport scientists to progress from small isolated studies through to large studies of many years of game data. Most studies used a single data source, as merging multiple heterogeneous data sources remains an integration challenge. The models discussed so far have needed to approximate the system to a small number of states in order to ensure sufficient data about the transitions between these states. Micro-level analysis (such as the physics of the ball bounce) remains out of reach. Complexity and network science provide a theoretical foundation for analysis, and can provide deep insights into the nature of the game. However, the predictive ability of this form of analysis is often weak, so it is difficult to measure the benefits objectively.

2.3.4 Spatio-Temporal Data

Computer-Vision Tracking in Sport

Computer-vision systems provide a means of tracking all player positions on the field. The key points of this section are presented visually as a timeline in Fig. 2.1. For a review of the wide range of sport tracking technologies available for use in Association Football see Carling et al. [30].

In 1992, Noble G. Larson and Kent A. Stevens filed a patent for an “Au-

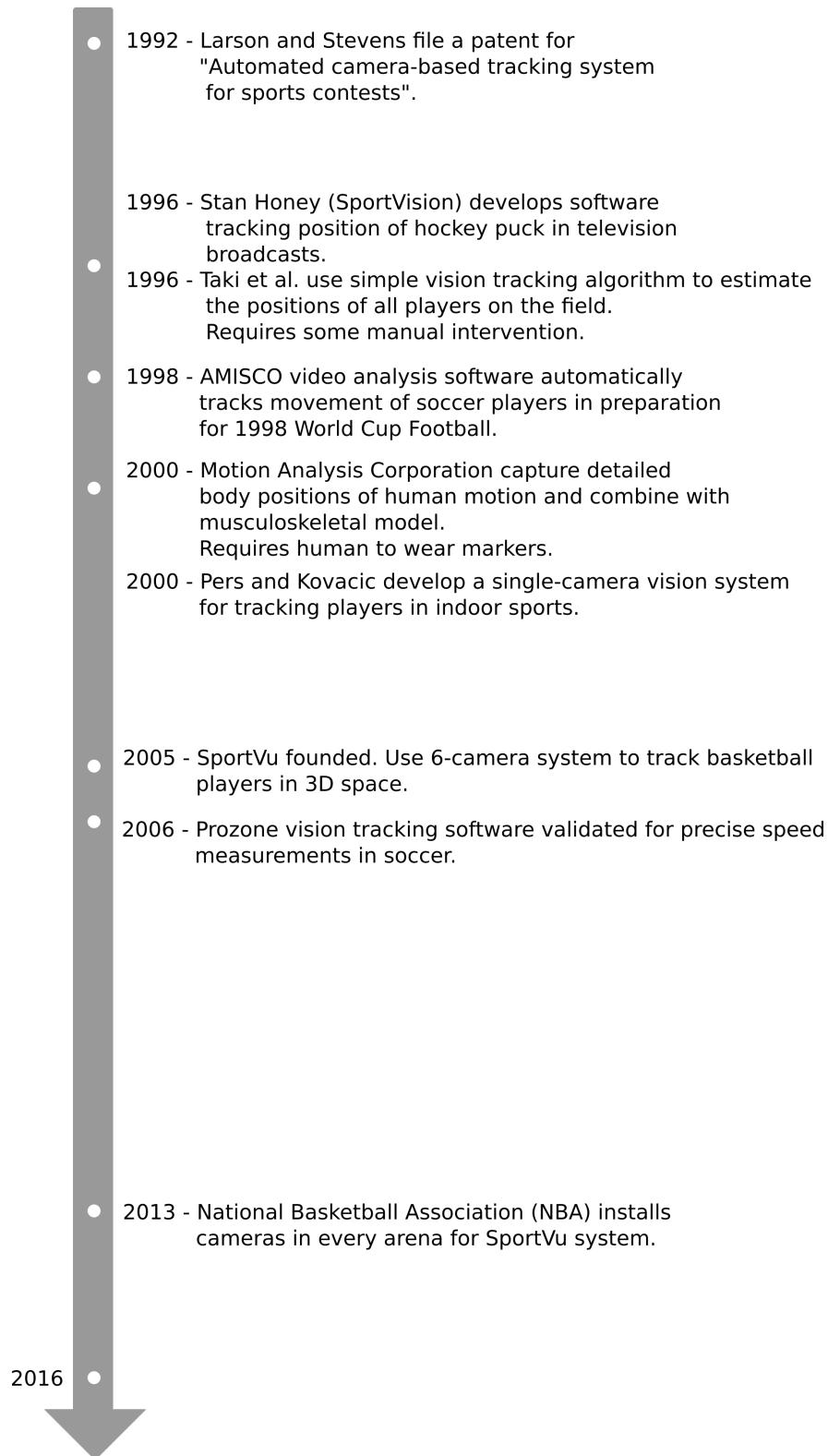


Figure 2.1: Timeline of Computer-Vision Tracking in Sport

tomated camera-based tracking system for sports contests” [123]. Their system uses an overhead camera and simple background subtraction to detect the approximate location of players. The patent admits the limitation that it will not work when there is visual occlusion (i.e. one player is hidden behind another in the video image) of the players. Thus it will not work well if there are “pileups” as are common in AFL and other types of football. The system does not have an automated means of identifying which player belongs to which trace. Thus the patent suggests that a human operator manually use a separate video feed to match the traces to the player identities. If visual occlusion (e.g. two players get too close to each other) causes the tracker to become uncertain about which player is which, then a human needs to resolve the ambiguity. An alternative suggestion was combining it with “telemetric data” (i.e. trigonometrically resolving the locations of radio signals emitted from devices carried by the players) to resolve the ambiguities. Despite being heavily cited by other sport patents, the patent expired in 2003 due to failure to pay a maintenance fee.

In 1996, Stan Honey (president of SportVision) developed software to track the location of hockey pucks in order to highlight the puck on television for easier viewing. SportVision went on to visualise the “first down” line in American Football broadcasts.¹⁹

In 1996, Taki et al. [204] implemented a simple vision tracking algorithm to estimate the positions of all players on the field. Their system required a human to manually identify the players in the first frame, and to manually re-identify the players under certain situations such as if a player falls down. In the same paper, Taki et al. introduced “dominant regions” as a means to analyse these data, which will be discussed later in the Analysis of Spatio-Temporal Data section.

In 1998, VideoSports claimed to use its AMISCO video analysis software to track the positions of players in preparation for the 1998 World Cup Football.²⁰ However, it is unclear how much manual user intervention

¹⁹Candice Shih, “Drawing the line,” 2002. https://www.mv-voice.com/morgue/2002/2002_02_08.sportvision.html Accessed: 2018-12-13

²⁰Videosports, “VIDEOSPORTS,” 5 Apr 2001. <https://web.archive.org/web/20010405040646/http://www.videosports.fr/index2.html> Accessed:

was required [11]. VideoSports competed closely with Prozone, a company founded for performance analysis in sport, which also targeted products towards elite sport teams [183].

In 2000, Delp (from Stanford University) and Loan (from Motion Analysis Corporation) published a paper [58] describing a system that could combine a musculoskeletal model with motion capture. The motion capture system used a technique known as “stereophotogrammetry” to create a three-dimensional image of human movements (such as a volleyball spike) from multiple camera recordings with different views. The musculoskeletal model was then used to infer the best estimate of the underlying bone positions, which may not be obvious on the surface. The vision system required physical ‘markers’ to be placed on the human to detect movements. Ordinarily, the vision system is used in laboratory and studio environments rather than outdoor sport. The motion analysis system has been used to study baseball pitches from video recordings of the 1996 Olympics without markers; however, this required humans to manually estimate the locations of body joints from the video [70]. Low cost alternatives for tracking body positions in outdoor environments have been suggested, such as covering the body with ultrasonic transmitters and sensors to calculate the positions of body parts from the distances reported by sensors (2007) [214].

In 2000, Pers and Kovacic described a system for tracking players using computer vision for indoor sport [169]. The system used multiple cameras for the purpose of covering the entire field, but the algorithm was primarily concerned with tracking using a single camera at a time. The authors found that a combination of colour detection and “template” tracking (detecting features such as edges or alternating colours) gave the best combination of tracking ability and noise reduction. Manual intervention was necessary when the system lost track of a player.

Later studies have described techniques for detecting players using blob segmentation (2001) [15], visualising ellipses of player movement (2005) [141], tracking players in outdoor sport using a camcorder and the two-dimensional Direct Linear Transformation (DLT) procedure (2005)

[208], tracking Association Football games from TV broadcasts (2006) [21], tracking players from a single moving camera using particle filters (2006) [56], and hybrids of manual and automatic tracking (2007) [16].

In 2005, SportVU was founded by Michael “Miky” Tamir and Gal Oz who both previously worked with Israeli military missile tracking technology.²¹ ²² However, Tamir’s research on military tracking technology is classified.²³ SportVU uses a system of six cameras to track players²⁴, and was patented in 2006 [205]. Unlike previous systems that simply use multiple cameras to obtain more complete coverage of the field, in SportVU, the different camera angles are processed together to infer the position of players in three-dimensional space. SportVU is primarily targeted at basketball. In 2013, the American National Basketball Association (NBA) installed cameras in every arena for the SportVU system²⁵, allowing detailed information about every player to be captured and published for coaches and sport fans.

In 2006, Valter et al. conducted a study to validate the use of Prozone to track players in Association Football under controlled conditions [210]. Players were asked to move at a variety of speeds between precisely positioned “timing gates” at known locations on the field. The study found that the system was able to accurately measure speed when running in a straight line, as demonstrated by a correlation coefficient of 0.999. However, the system was not as reliable when a short 20 m sprint was combined with a turn, which had a correlation coefficient of 0.950.

Development of sport monitoring devices and associated software is un-

²¹Z. McCann, “Player tracking transforming NBA analytics,” 19 May 2012. http://espn.go.com/blog/playbook/tech/post/_/id/492/492 Accessed: 2016-03-17

²²Israel21c, “Israeli company turns televised sport into whole new ballgame,” 2008. <http://www.israel21c.org/israeli-company-turns-televised-sport-into-whole-new-ballgame/> Accessed: 2016-03-17

²³Business Wire, “Idea sports entertainment group partners with israeli-based SporTVu, ltd., to launch patented revolutionary sports production technology,” 10 Jan 2005. <http://www.businesswire.com/news/home/20050110005651/en/Idea-Sports-Entertainment-Group-Partners-Israeli-Based-SporTVu> Accessed: 2016-03-17

²⁴STATS, “SportVU player tracking,” 2016. <http://www.stats.com/sportvu/sportvu-basketball-media/> Accessed: 2016-03-17

²⁵Associated Press, “NBA arenas to have motion-tracking cameras,” 6 Sep 2013. http://espn.go.com/nba/story/_/id/9639224 Accessed: 2016-03-17

dertaken by large commercial companies that target their products towards elite clubs. Sport device manufacturers and software companies try to maintain commercial advantage by keeping the devices and software proprietary. Furthermore, elite clubs have an incentive not to share their current analysis techniques, as this allows them to maintain a competitive advantage over other clubs. As a result, descriptions of technology found in academic sport literature may lag behind the state of the art as implemented by commercial companies and only made available to those working in elite clubs. Therefore, it has been necessary to identify these companies and closely monitor any publications originating from them in order to gain glimpses of the state of the art.

Despite a long competitive history, many of the major companies that offer player tracking for elite sport have been merged into STATS, LLC²⁶. STATS acquired SportVU in 2008. VideoSports/AMISCO became SportsUniversal, which merged with Prozone, then STATS acquired Prozone in 2015. STATS and Catapult Sports (an Australian GPS tracking provider, which will be discussed later) agreed to a partnership to integrate SportVU with GPS tracking, for use in NBA.²⁷

In contrast to manually collected sport event data that record sparse events usually relating to key interactions with the ball, video tracking technology provides dense spatio-temporal data; i.e. the position of every player, at every moment in time. State-of-the-art video tracking technology has matured to precisely track all players on the field, as well as the location of the ball. Even with inexpensive video recording equipment, vision tracking is possible, but suffers from issues of visual occlusion when many players are in close proximity.

²⁶STATS, “Sports technology company,” 2016. <http://www.stats.com/about/> Accessed: 2016-03-17

²⁷T. Haberstroh, “Now teaming up: Catapult and SportVU,” 9 May 2014. Available: http://espn.go.com/blog/truehoop/post/_/id/68143 Accessed: 2016-03-17

Development of Geopositioning Devices in Sport

Geopositioning devices such as those utilising the Global Positioning System (GPS) can be attached to players to track their movements during the game. Often these devices contain a range of additional sensors such as accelerometers, gyroscopes, magnetometers, and heart-rate monitors. One advantage of using geopositioning devices instead of vision tracking is that this method doesn't suffer from visual occlusion issues. The key points of this section are presented visually as a timeline in Fig. 2.2. For a review of the range of commercial geopositioning devices used for tracking athletes see Maddison & Ni Mhurchu [131], Cummins et al. [50], and Dellaserra et al. [57].

The first GPS satellite was launched in 1978 by the US Department of Defense. The system has been fully operational since 1995.²⁸ The GPS system consists of 31 satellites²⁹ in orbit about the Earth. Each satellite contains a high precision atomic clock, and broadcasts its current time. GPS devices trilaterate their position on Earth based on the time delays between signal transmission time reported by each satellite and the actual time the signal reaches the GPS device (crudely, the distance to each satellite can be calculated as the time-delay multiplied by speed of light; however, the slowing and distortion of the signal as it passes through the atmosphere complicates the situation). For accurate positioning, GPS devices require visibility of at least four overhead satellites (to solve for their position in four-dimensional space-time)³⁰ [209].

In 1996, William R. Fry filed a patent for a “Sports computer with GPS receiver and performance tracking capabilities” [80]. In the original proposed form, the device would attach to a bicycle, and log a range of sensors including GPS, heart-rate, compass direction, and weather conditions. The patent has since been assigned to Garmin Interna-

²⁸Australian Maritime Safety Authority, “GNSS navigation and horizontal datums,” May 2012. https://www.amsa.gov.au/forms-and-publications/Fact-Sheets/GNSS_Fact.pdf Accessed: 2016-02-25

²⁹US Government, “Space segment,” 2016. <http://www.gps.gov/systems/gps/space/> Accessed: 2016-02-25

³⁰Trimble Navigation, “GPS tutorial - getting perfect timing,” 2016. http://www.trimble.com/gps_tutorial/howgps-timing.aspx. Accessed: 2016-02-25

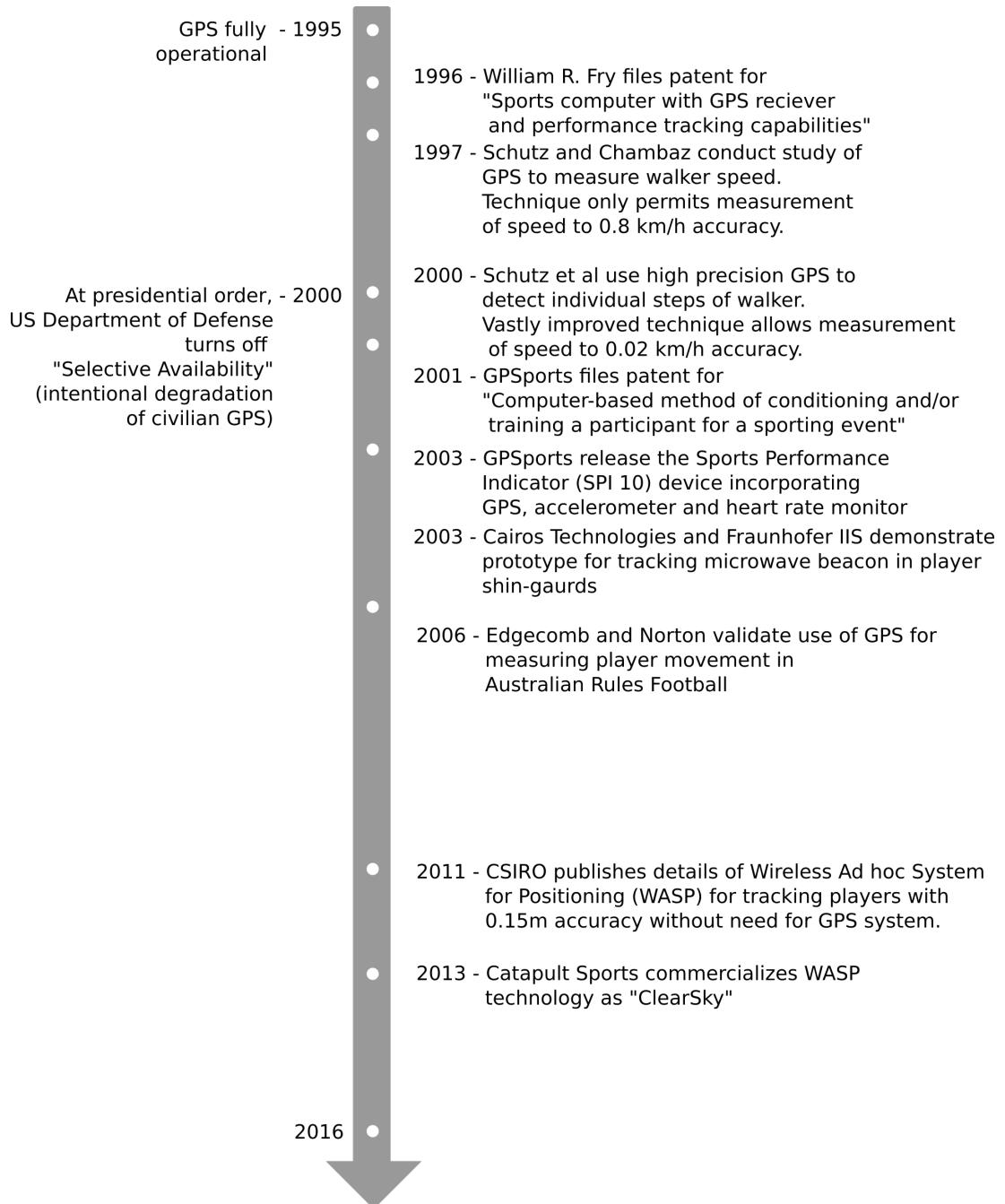


Figure 2.2: Timeline of Geopositioning Devices for AFL analysis

tional, Inc., which currently manufactures “wearable technology”³¹ for fitness tracking.

Schutz and Chambaz [180] conducted what appears to be the first [50] published study into utilising GPS for sport. In the initial study Schutz considered the use of GPS to measure walker speed. The GPS device used was a single frequency consumer device, valued approximately 500 USD when the study was conducted in 1997. Whilst the speed reported using GPS was correlated with true walking speed, the error was as high as 0.8 km/h. Schutz concluded that due to the large errors, GPS was “insufficient for research purposes”.

Selective Availability is a system where for “national security reasons” the US Department of Defense programmed GPS satellites to broadcast an intentionally inaccurate timing signal, with the deliberate timing errors kept secret from the rest of the world. This degraded civilian GPS, whilst allowing the Defense Department to correct the errors and use GPS to its full precision. Finally in May 2000, at presidential order, the US Department of Defense ended Selective Availability³², thus removing artificial sources of errors. However, GPS devices still suffer from errors beyond human control such as timing delays due to the atmosphere.

In 2000, Schutz co-authored a paper using high-precision GPS to investigate the individual steps of a walker [207]. In contrast to Schutz’s 1997 study, this study used differential GPS, in which a static reference station placed within 100 meters of the receiver records signals from the same satellites as the GPS tracker. As both devices experience the same interferences, the differential GPS technique virtually eliminates all sources of error, thus reducing positional accuracy from 10 m (typical GPS precision) to within 1 cm (differential GPS). Furthermore, the GPS device was placed in “differential carrier phase localization” mode, which solves mathematical equations for Doppler shift of GPS signals in order to obtain better estimates of speed. Schutz (citing a 1995 surveying textbook) claims that this should allow a theoretical precision of

³¹Garmin, “Garmin,” 2016. <http://www.garmin.com/en-US>. Accessed: 2016-02-25

³²US Government, “Selective availability,” 2016. <http://www.gps.gov/systems/gps/modernization/sa/> Accessed: 2016-03-01

0.02 km/h for measuring speed. Using this method, Schutz was able to detect the small rise and fall of altitude occurring each human step. Schutz realised that high precision GPS measurements were sufficient not just for typical speed and position monitoring, but could be used to explore the detailed characteristics of human movement.

GPSports (now owned by Catapult Sports) filed a 2001 patent for a device designed for “conditioning and/or training a participant for a sporting event” [72]. The patent described a device specifically targeted towards training elite players that contained a GPS and heart-rate monitor. In 2003, GPSports manufactured the SPI-10³³, the first “Integrated Technology” to incorporate GPS, an accelerometer, and a heart rate monitor [57].

In 2003, Cairos Technologies and Fraunhofer IIS demonstrated a prototype for tracking players using a microwave beacon in player shin-guards. The technology was reported to estimate positions with 5 to 8 cm accuracy [20]. A 2016 evaluation [182] using the RedFIR system [91] manufactured by Fraunhofer IIS found that it could track the position of a fast moving football (soccer ball) with a root-mean-squared error of 13 cm accuracy. The system was originally intended for capturing Association Football matches, but Association Football has resisted the use of technology in official matches, and only allows goal-line technology (for determining whether a goal has been scored).³⁴

Edgecomb and Norton [68] validated the use of GPS in 2006 for Australian Rules Football to measure the distance a player moves. They found an error of approximately 7%. An expertly trained operator graphically tracking the player location on a computer screen was able to achieve slightly better accuracy, but this was obviously much more labour intensive compared to the GPS device.

In 2011, CSIRO published details of a Wireless Ad hoc System for Posi-

³³GPSports, “The sport performance indicator (SPI 10),” 19 Dec 2003. https://web.archive.org/web/20031219215313fw_/http://www.gpsports.com/products.jsp Accessed: 2016-02-25

³⁴FIFA, “Blatter: Technology’s time has come,” 5 Jul 2012. <http://www.fifa.com/about-fifa/news/y=2012/m=7/news=blatter-technology-time-has-come-1660614.html> Accessed: 2016-03-23

tioning (WASP) [177]. The WASP system allowed tracking players with 0.15 m accuracy without the need for GPS, thus allowing the system to be used indoors and in closed stadiums. This was later commercialised by Catapult Sports as “ClearSky”.

These technological improvements mean that coaches now have access to high quality positioning data. However, anomalies and data gaps are still to be expected, for example it is common for GPS signals to be lost for a brief period of time. The additional sensors on tracking devices open up a range of possibilities, for example accelerometers can be used to measure the impact forces a player experiences when tackled by the opposition [82, 83]. However, data collection companies claim that qualitative data about the nature of interactions (e.g. determining the type of pass performed) still requires human observation.³⁵

Analysis of Spatio-Temporal Data

GPS tracking devices, vision tracking and other tracking techniques have brought about a large volume of spatio-temporal data. Either geopositioning devices or vision tracking can be used to collect high-precision player trajectories over time at a high sampling frequency. As such, the choice of technology is primarily to do with economic costs and regulations of the particular sport rather than one technology emerging as superior in all situations. For example, GPS is typically better suited than vision tracking for large outdoor fields where it is difficult to mount overhead cameras or control for lighting conditions, but can suffer from problems with interference caused by the stadium structure. Unlike GPS devices, Radio Frequency based geopositioning devices can precisely pinpoint players without the need for satellite visibility, but require beacons to be pre-installed at the venue. Vision tracking is well suited to cases where regulations prohibit wearable tracking devices. Due to the ability of different data capture technolo-

³⁵Following a trial of ball tracking, Catapult Sports stated “the reality is we can’t decide whether that was a kick or a handball.” A. Browne, “AFL scraps smart-ball trial,” 25 Feb 2013. <http://www.afl.com.au/news/2013-02-25/afl-scaps-smartball-trial> Accessed: 2016-06-01

gies to collect similar kinds of data, the analysis of the various tracking technologies will be considered together under the broader category of spatio-temporal data analysis.

A recent (2017) review of spatio-temporal data analysis techniques in sport is given by Gudmundsson & Horton [93]. The review pays very little attention to the devices themselves so as to maximise discussion of the analysis techniques used. The review finds a diverse set of techniques, including dominant regions, network analysis, and various visualisation techniques. However, there is no consensus on which technique is the best, with little to no validation of the practical value these techniques provide coaches or players.

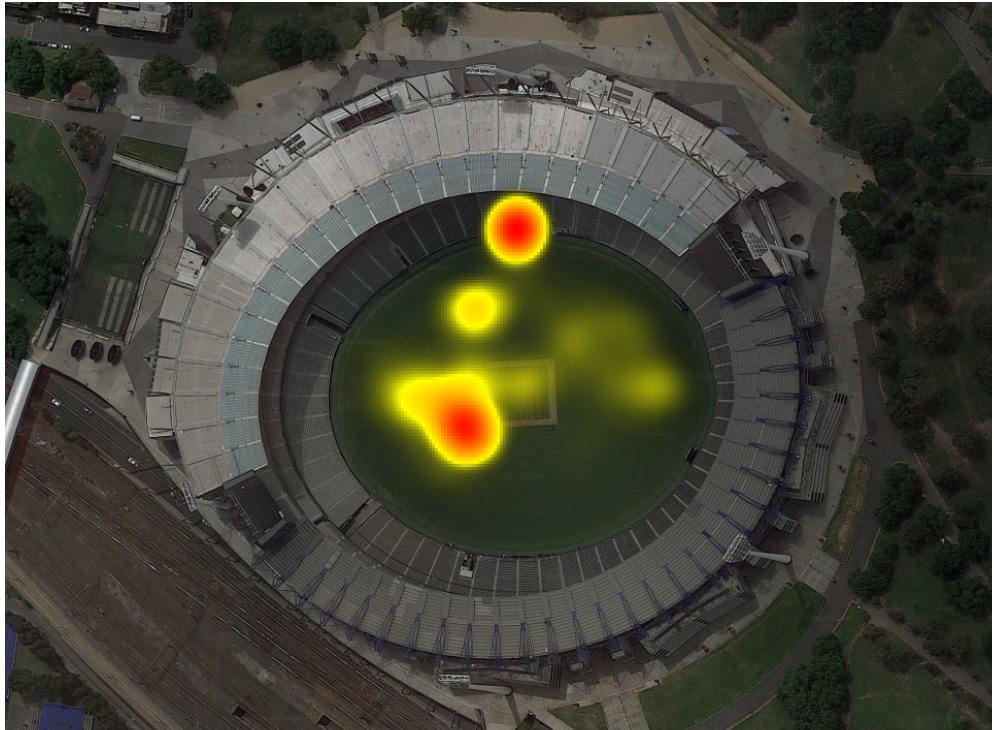


Figure 2.3: Heat-map showing positions where a particular player (name removed for privacy reasons) spent the most time during an AFL match. The heat near the north edge of the stadium is the interchange area. Background imagery ©2019 Google.

One popular tool for analysing spatio-temporal data is to create a heat-map [111]. Spatial data are aggregated by location. A exponential time-decay factor can be introduced to discard old data, so that the heat-map only shows recent activity. The heat-map can reveal insights into which areas of the field the ball moves in, which areas of the field the

players are covering, as well as any gaps that the player(s) might be missing. An example heat-map to visualise the positions of an AFL player is presented in Fig. 2.3.

Inspired by the concept of “equity” in backgammon, defined as “the value of a position to one of the players”³⁶, O’Shaughnessy applied this concept to AFL [162] to produce plots of the expected value of ball possession over the surface of an AFL field. O’Shaughnessy showed how this in turn can be used to produce an estimation of the expected game outcome conditioned upon the current score, ball position, and the team with ball possession. O’Shaughnessy’s plots show that the value of possession increases steeply near the 50 metre from goal line, as this is the distance that most players can kick a goal from. The value a player brings to the team can be assessed by how their actions increase or reduce the expected team score.

Karl Jackson built upon this concept to design the official AFL player ranking system. The concept is outlined in one of his early papers [110], and the full details of the completed system are in his thesis [111]. Players are rewarded points for the value their actions bring to the team. For example, a player that manages to successfully traverse the ball through a difficult region of the field where it would usually be lost to the opposing team will be rewarded highly. In contrast, if the player handballs the ball to a player nearby without incurring any risk or delivering any value, then they will not be delivered many equity points for this manoeuvre.

In a recent paper presented at *MathSport 2018*, Spencer et al. [197] demonstrate that equity can also be used to evaluate player decision making by comparing the expected outcome (defined in terms of the change to equity) of a passing decision made by a player to the expected outcome of the best alternative option available. This has only been possible recently, due to the need for: position tracking data to determine the locations of players open for passing to; tracking data for the opposition team to determine players that might try to contest

³⁶Keith, “Backgammon Glossary” <http://www.bkgm.com/glossary.html#equity>
Accessed: 2018-12-13

the possession; and a model of reachable regions to determine how far players will be able to move in time to catch the ball taking into account their current direction and speed. The authors of the paper noted a weak *negative* correlation between decision making and score margin (i.e. teams that made better decisions according to the model performed slightly worse), which highlights the early stages of this research and need for further work to refine the decision evaluation model. Players passed shorter distances than the theoretically better options available further away. The authors of the paper speculated that this may be due to obstruction of visibility. In Sec. 7.2, this thesis will offer an alternative explanation: moving the ball forward quickly does not leave time for the team to restructure itself around the ball, and is strongly associated with team spread. Thus players may be factoring in other considerations not present in the theoretical model, such as the disruption to the desired team shape.

However, the current expected value model used to calculate equity rankings in AFL is solely dependant on the manually annotated position of the ball³⁷ and possession state. While Spencer et al. [197] used player position tracking data to examine decision making options available, the underlying equity model utilised to evaluate those options was solely conditioned on ball position and possession and thus does not consider the implications of a pass that leaves the team in an undesirable formation. Ideally, the modelled state of a game should go beyond just ball position and possession to also consider the position of players on the field. Evaluation of decisions should look beyond the immediate passing opportunity to consider whether this will leave the team in a good position for subsequent passing opportunities. An example of this is the work on Expected Possession Value (EPV) in basketball [32] which provides a mathematical framework for computing the expected points arising from a possession considering all known information about the ball and position of players on both teams at a given time (similar to a stock ticker). Computing this efficiently requires modelling the game at different spatio-temporal granularities [31].

³⁷Ball position is manually pinpointed by Champion Data employees who monitor each AFL game.

In a similar vein to O'Shaughnessy's contour plots of expected value, Stöckl and Morgan introduced the "isopar" visualisation (a play on the name "isobar" used in meteorology) [200]. The isopar visualisation displays contours representing the expected score from any position in a game of golf. In 2013, they adapted this visualisation for the purposes of analysing scoring positions in hockey [201]. For hockey analysis, they used association rules mining to infer the value of positions. The visualisation revealed asymmetry of goal-shooting positions in hockey, due to the left/right handedness of the players.

In 1996, Taki et al. [204] suggested the concept of "dominant regions", which is inspired by Voronoi regions and Delaunay triangulation popular in geometry and computer graphics. The dominant regions are defined for each player as the set of points that they can reach before any other player. In 2014, Gudmundsson et al. [94] integrated dominant regions into a software tool for providing coaches with strategic insights.

In 2002 McGarry et al. described sport as a dynamical system [136]. Plotting the distance of squash players from centre-court vs time, they observed a sinusoidal waveform representing the to-and-fro of each player as they move out to hit the ball then return to a more neutral position. Superimposing the plots of both players, they observed that there appeared to be an invisible coupling between the two players, and measured the phase offset between the two position-vs-time plots. Through closer inspection of the superimposed plot, they explained how the characteristics of the plot could be explained using a dynamical system model consisting of periodic oscillators, lead-lag relationships, and damping.

Of particular interest, is the introduction of "perturbations" to the dynamical model. In sport notational analysis, perturbations are defined as events that disrupt the control characteristics of the dynamical system, causing it to fundamentally alter its nature. McGarry et al. [136] suggested that players deliberately create perturbations to upset the status-quo of the rally, allowing them to break the defence of their opposition. The role of these instabilities in determining rally outcomes

explains the difficulty that McGarry et al. initially had establishing consistent behaviour profiles for players in their earlier 1996 [135] study.

After Disney acquired the ESPN sports entertainment network, Disney focused research effort into the automated analysis of sport games. In 2014, Bialkowski et al. (from Disney Research and Queensland University) used clustering as a means of automatically identifying player roles from position traces [22].

Numerous machine learning techniques have been proposed to exploit the spatio-temporal data for predicting various attributes of the team [103]. Grunz et al. [92] proposed that self-organising maps (a specialised form of neural networks) may be used to model player creativity.

RoboCup is a competition in which physical or simulated autonomous robots compete in an Association Football inspired competition. The competition has inspired research into formalising strategy. Of particular interest is ISAAC (ISI Soccer Automated Assistant Coach) [148], an automated coaching assistant that analyses logs of the game using decision tree induction [170] to learn rules and suggest strategy modifications to improve performance. Beetz et al. [20] suggested that a similar approach could be applied to player traces of real Association Football games, and suggested that the the rules could build upon each other to express high-level strategies.

In 2017, Le et al. [124] were able to simulate expected team positioning behaviour using team spatial data in Association Football. They trained an LSTM model (a type of recurrent neural network) to predict the expected position of defending players given the position of the attacking team. The expected positions can be overlaid (as player “ghosts”) on a visualisation of the actual player positions to determine when a player is out of formation. Preparing data to feed to the LSTM model required construction of a computational pipeline with multiple sub-steps that feed into each other, such as data cleansing, determination of consistent indexes for players based upon their role, feature extraction, neural network training, inference, and visualisation. In 2018, Seidl et al. [181] applied this to the context of basketball. They demonstrate a sys-

tem whereby the coach digitally sketches their planed play on a tablet computer, and the system in turn automatically visualises how the opposition team is likely to position themselves in response.

2.3.5 Gaps

This section provides a list of gaps identified through the review. The extent to which this thesis will address each identified gap is outlined in Sec. 2.4.

Gap 1: There has been a proliferation of sport metrics and analysis approaches. While it is trivial to compare accuracy of predictive models for betting purposes, these do not necessarily deliver insight to players and coaches. As Domingos [61] explains in his informal review of Machine Learning:

“learners are typically compared on measures of accuracy and computational cost. But human effort saved and **insight gained**, although harder to measure, are often more important.” [61]

Which sport analysis approaches provide sport practitioners with the deepest insights, and how can insight be measured?

Gap 2: Researchers have used a range of different Markov models for sport, with varying numbers of states defined by various attributes. However, Markov models used in AFL currently only focus on the position of the player with the ball and do not incorporate the position of other team-mates.

How can the state explosion that occurs from the combination of all possible team formations be handled? What are the full set of attributes needed to describe the state of an AFL game (player positions, player velocities, player fatigue, ball position, time remaining, score margin, communication received, etc.), and which of these attributes can be

discarded or summarised while still providing a reasonable approximation?

Gap 3: Technology has advanced to the point that there now exists a large volume of high-precision spatio-temporal data. Whilst theoretically valuable, it remains an open problem as to how this data can be practically utilised for tangible gains:

- “sensed positional data contains an **enormous amount of information** that must be **correctly abstracted** to give coaches optimal support for controlling the competition” [20]
- “Several companies offer the ability to track the position of the players and the ball with high accuracy and high resolution. They also offer software that include basic analysis tools, for example basic statistics about distance run and number of passes. It is, however, a **non-trivial task to perform more advanced analysis.**” [94]
- “there is a vast amount of literature on tracking objects from video and hardly any research on **analysing the obtained movement data.**” [94]

What will be the *killer app* (seminal software that gives value to a platform) for spatio-temporal sport analysis?

2.4 Chapter Summary

The review identified 3 key gaps: 1) the lack of an agreed upon definition of “insight” in the sport literature, 2) that existing models of AFL focus on the ball rather than the team, and 3) there is no open platform for spatio-temporal team sport analysis. This section explains how the thesis research questions (defined in Sec. 1.3) contribute toward filling these gaps (detailed in Sec. 2.3.5). The mapping is shown in Fig. 2.4.

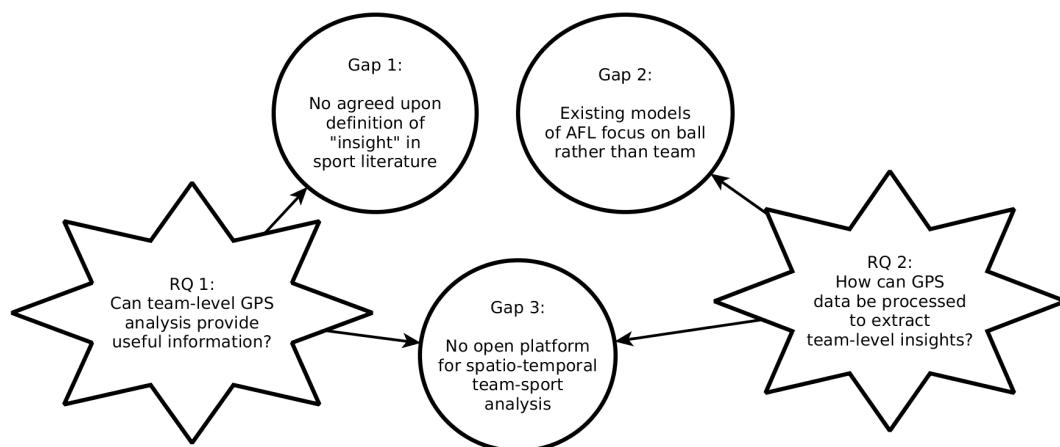


Figure 2.4: Mapping of Research Questions to Gaps

Addressing Gap 1

Research Question 1—*can team-level GPS analysis provide useful information to sport researchers and practitioners beyond what they already know from manual observation, video analysis, traditional statistics, and (individual) player GPS monitoring?*—contributes toward filling Gap 1 through providing an example of an approach that is useful to sport practitioners, i.e. that offers insights that can help understand the game (in contrast to betting models that solely attempt to predict game outcomes). In particular, Chapter 3 provides theoretical modelling of the problem in terms of player feedback (Sec. 3.2) and information theory (Sec. 3.3) in order to reason about which types of solutions are most likely to offer meaningful insights.

Addressing Gap 2

Research Question 2—*how can GPS player tracking data be processed to extract meaningful team-level insights without compromising individual privacy?*—contributes to filling Gap 2 through considering the team formation, and not just the player with the ball. Sec. 3.1 provides a full list of all identified state variables involved; however, the thesis focuses on the team formation, in particular the team spread. Chapter 5 deals with the challenge of representing the team formation without revealing individuals.

Addressing Gap 3

Gap 3 pointed towards the quest to develop a *killer app* that will offer sport performance analysts with unprecedented strategic insights into the game through uncovering patterns previously hiding in the wealth of data now collected on players. While the work in this thesis is demonstrated through a minor contribution toward better understanding how AFL team formations spread/contract in defence/offence (Sec. 7.2), this is not the main goal of the thesis.

Rather, by addressing Research Question 1 and Research Question 2, the aim of this thesis is to create a platform that facilitates further spatio-temporal analysis of team sport. This requires consideration of the computational pipeline framework (Chapter 4), de-identification operations (Chapter 5), and spatial normalisation operations (Chapter 6) involved. While the importance of these topics is well established in the software engineering literature, these details are often overlooked entirely in the sport science literature. Each stage is designed with consideration of Human–Computer Interaction (HCI) concerns, such as permitting sport performance analysts to trace results back to relevant video footage, safely de-identify data when sharing them with researchers, and manage reference frames without the need to master GIS.

Chapter 3

Modelling

Contents

3.1 Model of Australian Rules Football	51
3.2 Model of Feedback	55
3.3 Information Theoretic Perspective	57
3.3.1 Problem Formalisation	58
3.3.2 Theoretical Solution	61
3.3.3 Heuristic Solution	63
3.3.4 Discussion	64
3.3.5 Limitations	65
3.3.6 Conclusion	66
3.4 Abstract Data Model	67
3.4.1 Related Work	68
3.4.2 Sensor Model	72
3.4.3 Definitions	75
3.4.4 Conclusion	76
3.5 Chapter Summary	77

The previous background chapter began with a high level overview of the game of Australian Rules Football and the coaches' role in providing feedback. This chapter formalises both of these concepts as models. This allows a more precise discussion and analysis of the concepts, and is an important step towards formalisms appropriate for mathematical analysis of the game.

The first research question (as defined in Sec. 1.3) asked whether team-level GPS analysis could provide useful information to sport researchers and practitioners beyond what they already know. In order to more precisely formulate this question, the field of information theory [185] is used to provide a formal definition of *useful information* and to frame the role of sport data analysis systems within the larger domain of sport.

In order to prepare for automated analysis of spatio-temporal sport datasets, it is first necessary to establish terminology to describe the different kinds of spatio-temporal data. This takes the form of an abstract data model for spatio-temporal data. It is applied to describe the diversity of spatio-temporal datasets available in sport generally, as well as the AFL datasets used in this thesis.

3.1 Model of Australian Rules Football

The purpose of this section is to build a domain model of consistent terminology to describe Australian Rules Football. This allows linking formal reasoning about the game (e.g. mathematical models of game state) back to the somewhat looser jargon used by sport practitioners described in Sec. 2.2.

Formally modelling games requires an understanding of the game *state*. Decisions can then be strategically evaluated in terms of how they change this state. For example, in chess, the game state is the position of all the pieces on the board, which player has the next move, as well as other information needed to enforce the game rules, such as whether

players still have the right to castle¹. In contrast to abstract strategy games such as chess in which the game state can be mathematically defined, fully describing the state of an Australian Rules Football game requires consideration of a large number of factors.

To capture aspects of the game relevant to describing the short-term and long-term state, the official AFL coaching manual [2] was manually mined for terminology, particularly chapter and section headings. These were then reassembled into a domain model as presented in Fig. 3.1. The domain model is specified using the Unified Modelling Language (UML) standard maintained by the Object Management Group. Diamonds denote aggregation, the numbers near connectors represent multiplicity (e.g. a match has two teams, a team consists of multiple players), and triangles denote generalisation (e.g. disposals, possession and contact skills are types of individual skills, whereas problem solving, technical awareness and choice of technique are types of game sense). When mining the coaching manual, sections within a chapter were taken as an indicator of a possible relationship between terms in the section header and those in the chapter header. The development of the final model required an iterative process to improve the internal consistency of the terminology and incorporates minor refinements based on feedback from sport researchers².

¹Wikipedia Contributors, “Board representation (chess)” Available: [https://en.wikipedia.org/w/index.php?title=Board_representation_\(chess\)&oldid=880707099#Requirements](https://en.wikipedia.org/w/index.php?title=Board_representation_(chess)&oldid=880707099#Requirements) Accessed: 2019-02-19

²These were senior sport scientists on my current/former supervisory panel. This information is provided as an acknowledgement of their input, but should not be taken as a statement of expert validation of the final model.

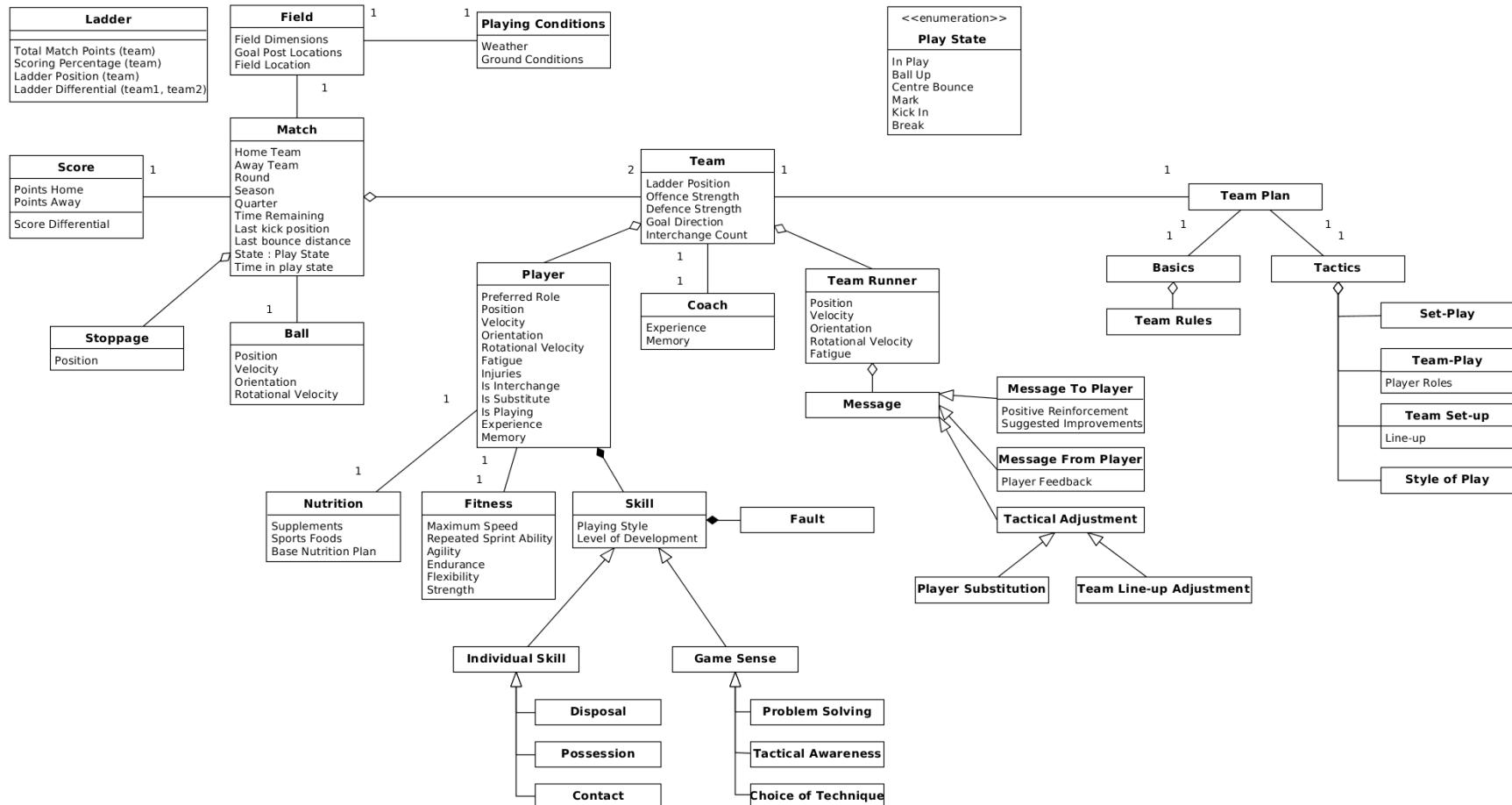


Figure 3.1: AFL Domain Model

This section will approximate the physical systems involved as rigid bodies. To avoid repetition of the parameters involved, let the *kinetic state* of a rigid body be the set of variables required to completely describe the current state of the body within classical mechanics, *viz.*, the position, velocity, orientation, and rotational velocity. Forces are not included as part of the kinetic state, as forces are transient properties that only last as long as they are applied.

The information required to describe the current game state during a match³ includes:

1. The score differential
2. The time remaining on the clock
3. The play state, for example, whether the game is in play, or has been paused by the umpires
4. Information needed by the umpires to enforce the laws of the game⁴
5. The kinetic state of the ball
6. The kinetic state of all players on the field (from both teams), and the team runners
7. The number of interchanges made by each team
8. The current weather and ground conditions
9. The level of fatigue each player is under
10. Any injuries sustained by players
11. The memory of players: what they have observed, and messages from the coach

³If modelling the game for the purpose of long-term decision making, then it would also be necessary to include additional information such as the team ladder position.

⁴For example, the player must bounce the ball at regular distance intervals, hence the distance since the last bounce is part of the game state.

12. The memory of the coaches: what they have observed, and messages from the players
13. The memory of the team runners: the current messages they are relaying

As can be seen from the list of 13 game state components identified above (each in turn consisting of further state variables), it is non-trivial to precisely describe the state of an AFL game. Furthermore, some aspects, such as the state of information flows between coaches and players via team runners, impact upon the game⁵ but are not publicly recorded. In practice, it is necessary to approximate the game state to a simplified representation in order to permit formal reasoning about the game. In the past, formal models of AFL have simplified the game state by focusing solely on the location of the ball and which team has possession [162, 111], but this thesis aims to more closely approximate the true game state by using player position tracking data to incorporate the formation of team players on the field as part of the analysis.

3.2 Model of Feedback

Ultimately, the information a coach has access to is only useful if the coach can transform it into feedback for the players to action. A feedback model was developed to analyse feedback information flows available in competitive team sport, which is presented in Fig. 3.2, and described in greater detail below.

Players select actions based upon the context of the situation they are presented with. The actions of both teams combine through the laws of physics, as well as the laws of the game enforced by the umpires, to result in a concrete outcome, such as a goal scored. The outcome is the combined result of all players. Lames and McGarry [122] discuss how

⁵Peter Schwab (Brisbane Lions), 2013, “In-Play Communication” http://www.aflcommunityclub.com.au/index.php?id=49&tx_ttnews%5Btt_news%5D=2768&cHash=770d353c8665ecfac980dc8a70842de6 Accessed: 2019-02-19

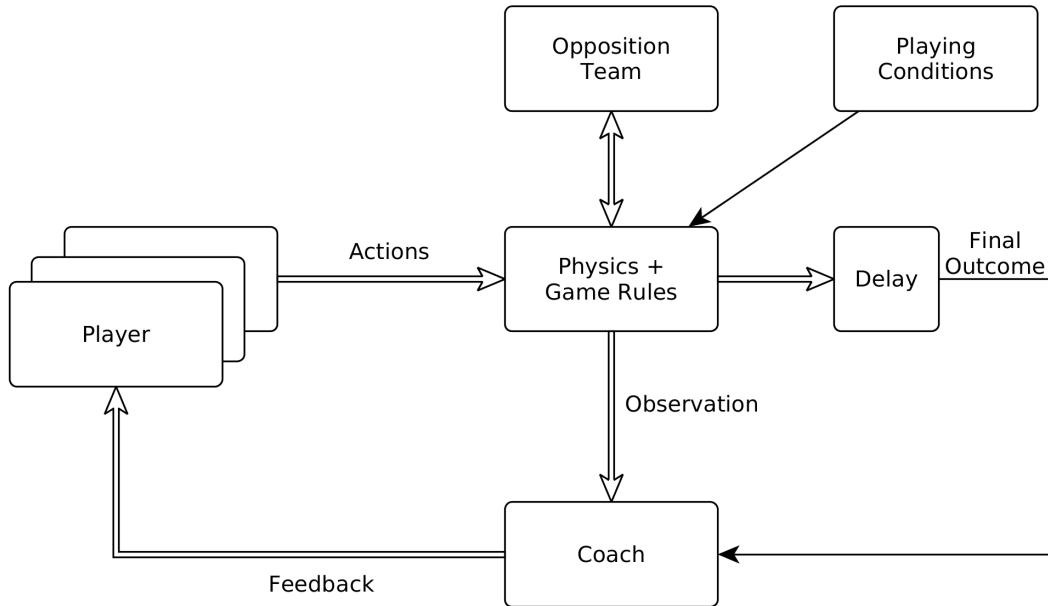


Figure 3.2: Feedback Model

this interaction process is so involved that the underlying skill is not observable through a traditional approach, and go so far as to suggest that performance analysis in team sport is so difficult that it requires a qualitative methodology. Player actions can be observed, however their consequences are not always directly apparent (e.g. a strategic pass may at first appear a poor decision, but later lead to the team scoring a goal). This is represented in the feedback model as a “Delay” between when the player actions take place, and the final game outcome.

“Playing Conditions” is also included in the diagram to represent the weather and ground conditions that introduce unpredictable variations into the game. Although the laws of physics (to a classical mechanics approximation) are deterministic in principle, small unmeasurable variations of weather and the playing surface can cause the ball to bounce in a different direction, and ultimately affect the final outcome of the game. Thus, for all practical purposes, the game is a stochastic process [76].

For long term optimisation over many games, coaches could feed back the final game outcome to the players as a way of training them how to improve. However, this is not practical as the sole source of feedback,

as players would need to experiment over many games to identify the underlying cause of performance issues. Furthermore, it is not clear how the coach should assign credit to individual player actions, as the final outcome is a result of many actions combined.

Instead, the coach uses their observations of the game along with statistics provided by sport performance analysts in order to provide players with more direct feedback that reflects how the coach believes player actions contribute to the final game outcome. This form of feedback is more immediate and actionable than game outcomes alone, as it tries to explain performance rather than just measure performance. However, it can be compromised in three ways: inaccurate observations of the game state, feeding back a measure that does not align with overall team performance, or not accounting for the psychology of the way players learn from feedback.

3.3 Information Theoretic Perspective

This section explains how information theory can provide the field of sport science with rigorous answers to the following questions, which were previously only answerable based upon the anecdotal experiences of sport scientists and coaches: *What is the purpose of transforming and summarising sport data if it cannot create new information beyond what coaches can manually observe by watching the game? What value does interactive data analysis offer over pre-defined analysis procedures?*

In a single player sport such as archery, performance is easily observable through game outcomes. Whilst there is an element of luck involved, this can be eliminated by averaging outcomes over multiple attempts. In contrast, measuring performance in team sport is an intractable problem, as the game outcome is the result of many players' interactions. Furthermore, even a minor disturbance, such as a gust of wind, could influence the game outcome. These issues led Lames and McGarry [122] to declare classical quantitative performance indicators unreliable measures of team sport. To address this failure, Lames and

McGarry speculate upon the possibility of modelling sport as a complex system, although they suggest that a qualitative approach may be more practical.

This section analyses the flow of information in team sport performance measurement from an information theoretic [185] perspective, which provides a mathematical framework for rigorously reasoning about information flows. Unlike Lames and McGarry, it does not dismiss quantitative team performance measurement as unreliable based upon the perceived complexity of team interactions; instead, it uses information theory to reason about the theoretical limits of such an approach. Note that qualitative approaches also constitute a form of information flow that must also obey information theoretic constraints, albeit harder to accurately measure.

3.3.1 Problem Formalisation

The information flow diagram depicted in Shannon's original paper on information theory, *A Mathematical Theory of Communication* [185], was adapted for the sport domain. The adaptation includes game interaction process described by Lames and McGarry [122], sensing, automated analysis, as well as a simplified model to represent the cognitive interpretation of visual information. The resultant information flows for performance analysis in sport are depicted in Fig. 3.3.

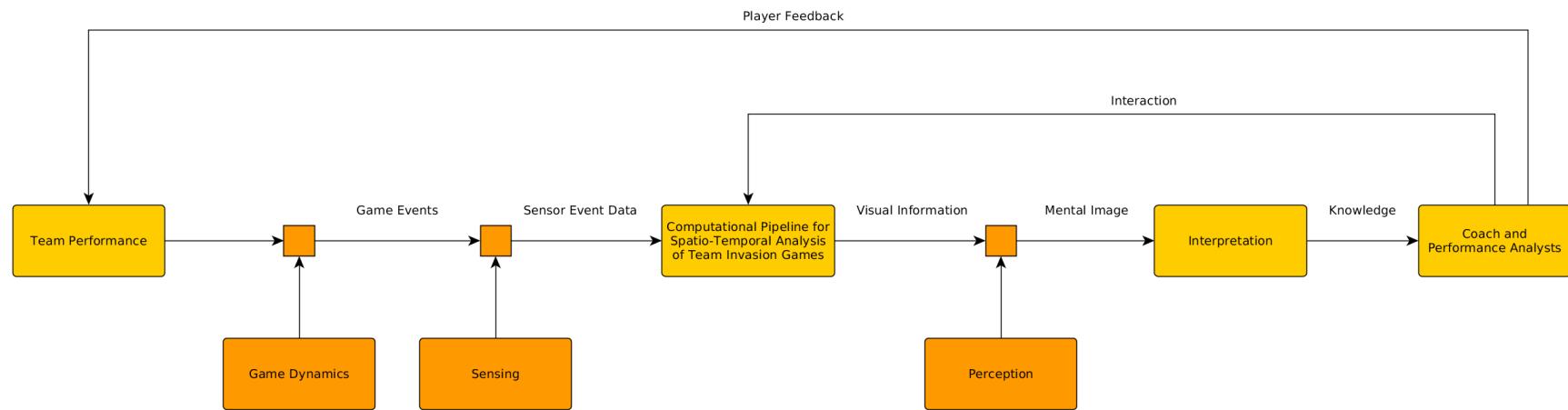


Figure 3.3: Information Flow in Performance Measurement Process

The chain of processes involved (other than coaching interaction with the computational pipeline, which will be discussed later in Sec. 3.3.4) is represented mathematically as:

$$Know = (Interp \circ Percep \circ Comp \circ Sens \circ Dyn)(Perf)$$

Where:

$f \circ g$ represents function composition of two arbitrary functions f and g . i.e. $f \circ g = f(g(\dots))$

$Perf$ is the team performance. This incorporates the individual skills, abilities, and decision making of each player, as well as team factors such as team playing style, and team cohesion.

Dyn is a function representing the game dynamics that captures how individual player interactions become game events.

$Sens$ is a function representing the sensor observations of the game, and any noise introduced by imperfect sensors.

$Comp$ is the function implemented by the computational pipeline that analyses sensor inputs and outputs game information for visual display to the coach and performance analysts. The computational pipeline may contain many sub-components; however, together they are still only one component of the overall sport performance measurement and feedback process.

$Percep$ is a function representing the coach's ability to perceive visual information, taking into account visual acuity.

$Interp$ is a function representing the coach's interpretation of the visualisation, transforming a visual representation of the game into knowledge about the game.

$Know$ is the knowledge of the game delivered to the coach by the system. This is defined to include both trivial facts about players, as well as deeper knowledge about the game itself that helps the coach to predict the results of actions within the game.

The aim of a computational pipeline for sport performance analysis is to give the coach greater knowledge of team performance. This can be formalised in terms of information entropy. Specifically, by looking at the amount of information about team performance that remains unknown to the coach. From this perspective, the aim of the computational pipeline is to produce a visualisation that reduces the amount of unknown information about team performance.

Formally this can be stated as: given prior knowledge of the sport domain, find a pair of functions for *Computation* and *Interpretation*, which minimise the entropy of *Team Performance* conditional upon the *knowledge delivered to the coach*.

$$\min_{Comp, Interp} H(Perf | Know)$$

Where:

H is information entropy, as defined by Shannon [185].

$Perf$ represents the team performance of interest to coaches as defined earlier.

$Know$ is the coach's knowledge of team performance as defined earlier, enhanced via computations $Comp$ interpreted according to $Interp$ delivered via the chain of (noisy) processes described earlier.

3.3.2 Theoretical Solution

If the intermediate distortion functions were reversible, then it would be possible to find an optimal solution. The components of an ideal performance analysis pipeline, and its interpretation are depicted in Fig. 3.4.

Theoretically, one could invert the sensor readings to obtain the game events, and then invert the game dynamics in order to find the underlying team performance. As sensor technologies improve, inverting

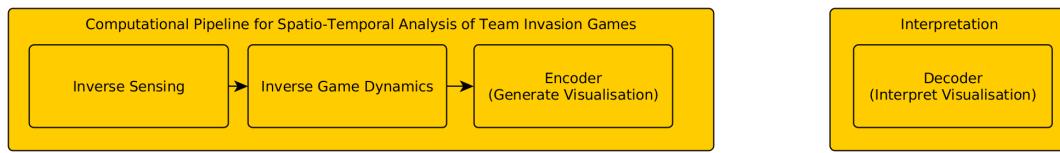


Figure 3.4: Theoretical Solution: Inverse distortions.

the sensor data to obtain the game data is becoming feasible. Historically, only a small number of events were captured in sport games, such as the total number of goals scored by each player. This discards much of the game information (such as the time the goals were scored), thus making an exact reconstruction of the game impossible. With the development of pervasive sensing technologies for sport that can track players' every movement, near-exact reconstruction of game events from sensor data is now possible.

The challenge, however, lies in inverting the game dynamics to obtain the team performance from game events. In strategic games, such as backgammon, it is possible to evaluate human decisions by comparing to the ideal choice determined by computer simulation [161]. Whilst AI players can dominate humans in abstract strategy games such as backgammon, chess, and Go [187], AI controlled sport players⁶ are yet to master basic skills within physical environments. As such, there is no reliable “optimal” decision to evaluate players against, implying that precisely inverting the game dynamics function is not yet possible.

However, heuristics exist to approximate the inversion so as to obtain performance from game events. For example, AFL equity ratings [111] examine the contribution a player makes to their team based upon the changes they make to the expected score. Note that the current AFL equity rating only considers factors such as team-with-possession and position-of-ball, and discards consideration of players without the ball in the estimation of expected score, so is clearly a heuristic rather than a precise measure of a player’s true contribution.

If both game dynamics and sensing could be accurately inverted, then

⁶RoboCup Federation, “A Brief History of RoboCup”.
http://www.robocup.org/a_brief_history_of_robocup Accessed: 2017-03-20

team performance could be recovered precisely. At this point, transmitting the team performance to the coach would simply be a matter of encoding the performance results, either numerically or visually, for the coach to interpret.

Shannon's mathematical theory of communications [185] reveals that this information can in theory be transmitted to the coach more effectively by designing an encoding structure that takes advantage of the known properties of the performance signal. For example, if the performance is usually “player 1: normal, player 2: normal, ...”, then one could create a simple symbol to represent all-players-are-normal-performance, and devise more complex symbols to represent anomalous situations that rarely occur. However, interpretation of visualisations requires cognitive resources which are likely to outweigh any information-capacity benefits attained through compressing data using complex visual encoding schemes. In practice, it is well established in the information visualisation literature that visual encoding should cater to address the limits of human cognition rather than maximum information density [6, 101]. This can potentially be achieved using a framework such as Physics of Notations [144].

3.3.3 Heuristic Solution

An alternative option that bypasses the need to build a full model of the game is to extract features that correlate with performance. Ideally the features selected for display to the coach should be independent of each other, as a set of features that correlate with each other would contain redundant information that makes inefficient use of both space on the visualisation, as well as the coach's cognitive resources to process this information. The components of this heuristic solution are depicted in Fig. 3.5.

A simple algorithm for creating such a set of independent features is Principal Component Analysis (PCA). Note however, that Principal Component Analysis is linear, so will not be able to detect team performance

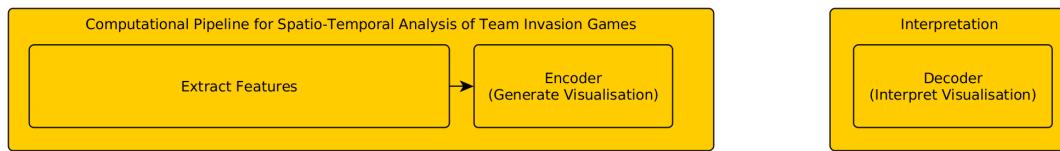


Figure 3.5: Heuristic Solution: Extract features

features that aren't revealed as simple linear combinations of sensor features. This heuristic can be enhanced by first transforming the sensor space into a domain where combinations of sensor features are more likely to correlate linearly with team performance. For example, player agility may not be obvious from player position recordings alone, but might become obvious once transformed to frequency space using a Fourier transform, thus revealing periodic movements—which might be generated by movements such as swerving between players—as components of the spectrum.

As with the theoretical approach, these derived features are then presented to the coach. The coach may interact with the system in order to select the features that they believe provide the highest information content. This interaction allows the coach to improve the efficiency of the encoding by filtering out signals that do not correlate well with team performance, as well as allowing them to remove unsurprising features that reveal aspects of team performance that they were already well aware of (e.g. from their manual observations of the game).

3.3.4 Discussion

By viewing sport analysis through the lens of information theory, it is possible to draw answers to the questions set out at the start of this section of the thesis.

What is the purpose of transforming and summarising sport data if it cannot create new information beyond what was present in the raw dataset?

If the coach had unlimited cognitive resources, then computational pipelines, from an information theoretic perspective, can not offer any additional value beyond what is present in the raw dataset, as they cannot generate new information in an information theoretic sense. However, in practice, the capacity of the coach's cognition channels are limited. The psychological phenomenon of "inattentional blindness" [194] shows that humans do not have the ability to reliably monitor all player interactions simultaneously. As such, the purpose of a sport data analysis pipeline is to extract the features of the game that provide the highest information content about team performance, and filter irrelevant information, thus maximising the value of information transferred over the coach's limited cognitive channels.

What value does interactive data analysis offer over pre-defined analysis procedures?

As outlined in the Heuristic Solution (Sec. 3.3.3), interaction provides the ability for the coach to customise the data analysis to provide them with information they don't already know. Minimising the entropy $H(Perf|Know)$ requires information that is independent of the coach's prior knowledge of performance.

3.3.5 Limitations

Team performance was defined to incorporate the individual skills, abilities, and decision making of each player, as well as team factors such as team playing style, and team cohesion. However, it is unclear how this should be represented. It is also unclear whether it is possible to quantify individual performance rigorously within the context of a team. The

mathematical field of “game theory”⁷ shows many problematic scenarios where maximising individual utility functions does not maximise net utility for the team, which may undermine incentive schemes designed to reward players for individual performance.

Through an information theoretic lens, the main constraint on the processing was considered to be the information capacity limit on the coach’s ability to perceive information. In practice, an additional constraint is the coach’s speed of mental computation. Even if a coach has access to all the knowledge that they need to accurately determine performance in theory, it’s possible that the knowledge is in a form such that computing the performance from the knowledge could take excessive amounts of computation, well beyond what a human is capable of. In future, the model could be extended to include mental computation constraints by combining the information theoretic perspective used to analyse information capacity constraints with a cognitive load perspective to analyse the mental computation constraints.

3.3.6 Conclusion

Using an information theoretic framework, this section of the thesis showed that computational pipelines have a clear role to play in the sport coaching process, as a tool to provide coaches with the most important information required to enhance their knowledge of the game. Conventional performance indicators such as “number of kicks” are unlikely to offer the coach any useful information, as this is something that the coach will already have approximate knowledge of from their manual observations of the game. Instead, the proposed model suggests that sport performance analysis systems should focus on providing information that the coach has minimal prior knowledge of.

Future work is needed to devise measures for each attribute, and approximations of each function, in order to provide numerical estimates of the information gain offered by incorporating analysis tools into the

⁷Wikipedia Contributors, “Game theory”. https://en.wikipedia.org/wiki/Game_theory
Accessed: 2017-03-20

sport performance process (this is a non-trivial task as it requires both functions to simulate the game, as well as functions to approximate the coach’s cognition). Another area to explore is to model the feedback as part of a learning system where the goal is not knowledge of any individual signal, but rather stability of the overall system—possibly deliberately taking sub-optimal actions in order to learn from them (the human equivalent of the exploration-exploitation trade-off faced by autonomous agents in the field of reinforcement learning [202]). As future work, it is also necessary to incorporate computation limits into the proposed model. However, even as stands, the model provided is capable of providing a new perspective on the value of computational pipelines within the larger sport performance feedback process.

3.4 Abstract Data Model

The previous sections of this chapter established the motivation for building computational pipelines for spatio-temporal analysis of team sport and modelled the game of AFL as well as the player feedback systems involved. This section will model the types of sensor data available that could power such an approach.

Traditionally, only summary statistics such as total number of goals were recorded. However, the advent of sport databases has allowed a much richer set of data to be collected, such as the time and players involved in every pass. Recently, the National Basketball Association (NBA) have installed computer vision technology that tracks the precise location of the ball and every player on the court. Similarly, AFL players are also tracked, albeit through a different technological means. In AFL, players wear tracking devices during the game that have a GPS chip for tracking position, in addition to sensors for measuring acceleration, orientation, and (optionally) heart rate.

It is not just coaches that are interested in these data. Sport fans, betting markets, sport reporters, and sport researchers are also stakeholders with an interest in the data collected. Each stakeholder re-

quires access to different kinds of data depending on the objective they wish to achieve, and their intended method of analysis.

Clearly, the detailed player tracking data available today provides much more detail than the simple counts that have been recorded historically. However, what lacks is a consistent language for expressing the detail and type of data each stake-holder requires, and which types of analysis the data can support.

Further complicating the issue is that of measurement errors inherent to the sport data collection technology used. For sound sport research, it is vital that sources of errors are known so that errors can be dealt with in a rigorous manner using appropriate statistical and inference techniques rather than simply neglected [103].

This section presents an abstract data model that establishes the fundamental data types involved in sport datasets, with a focus on the ability to describe spatio-temporal datasets, such as those collected by GPS tracking devices. This abstract data model is later used to describe concrete data schemas for specific datasets (Appendix A).

3.4.1 Related Work

In 1946, Stevens proposed classifying measurements as Nominal, Ordinal, Interval, and Ratio scales [199]. The scale classification was defined with respect to two competing concerns: transformations that could be applied to the scale whilst still preserving its essential relationships (for example, player identification numbers are on a nominal scale, and can be arbitrarily re-assigned); and statistical operations that could be meaningfully be performed upon the scale (interval and ratio scales such as player speed allow taking the mean, but it is meaningless to talk about the mean player identification number).

Stevens' classification model was revolutionary in its realisation that qualitative measurements and physical measurements could both be described on the same scale hierarchy, and both permit sound scientific

analysis so long as the limits of the particular scale are respected. However, Stevens' levels of measurement are not without criticism; in particular, the limits they place on which statistical operations can be meaningfully performed, whilst theoretically justified, can be overly restrictive and unpragmatic in practice [227]. Stevens' levels of measurement may not be well suited for certain domains; in 1998, Chrisman identified that Stevens' original levels were not well suited for cartographic measurements, and proposed additional levels for Graded membership, Log-interval, Extensive ratio, Cyclic ratio, Derived ratio, Counts, and Absolute [38].

The introduction of low cost positioning sensors that monitor both time and location has led to the question of how to handle spatio-temporal data, and the need to distinguish between datasets that merely consist of separate space and time measurements from those where space and time are intricately interlinked. Moving object database research [179] attempts to rethink the design of databases to handle the unique challenges posed when storing and querying spatio-temporal data rather than conventional tabular data. Suitably accounting for measurement errors when processing spatio-temporal data requires careful selection of techniques based upon the type of data queries performed [29].

In practice, most spatio-temporal data models either: provide full support for the spatial aspect of the data, with secondary consideration given to the temporal aspect (for example, Geographic Information Systems); or focus on the temporal aspect of the data, with secondary consideration of the spatial dimensions (for example, time series analysis). These differences are so fundamental, that in the design of the *R* software package “Spacetime”, multiple representations are provided to deal with the different forms of spatio-temporal data [167].

Sport databases are typically designed for a specific sport. This is surprising, as sport media networks typically cover a range of different sport types, so one would assume that they could reduce costs by abstracting the common elements of multiple sport types into a unified structure. Furthermore, sport researchers require consistent schemas to conduct meta-analysis. The lack of a common structure has made it

difficult for sport scientists to conduct systematic reviews across multiple sport types, such as in Cummins et al. [50] where inconsistencies of speed zones, even within the same sport, were noted as a limitation of the analysis.

Despite the growing prominence of sport analytics, very few attempts have been made to construct a unified model of sport performance data. The Sports Standards Alliance⁸ attempts to promote a standardised schema for sport. However, this attempt at standardisation is catered to sport fans rather than coaches or researchers, and hence tends to focus on traditional summary statistics. It includes data fields for in-game events for some types of sport, but support for time dense spatio-temporal data (such as player position tracking data) is limited.

Motivated by the desire to reduce the burden on clubs to perform custom data processing to integrate with specific tracking providers, FIFA proposed the Electronic Performance and Tracking Systems (EPTS) standard data format⁹. Notably, the same format is used to support Optical, GPS, or Radio Frequency based tracking data. As the standard is specifically designed for Association Football, it includes meta-data fields for teams, players, sessions, and field dimensions. Rather than enforcing a specific format for the main tracking data, the standard provides fields to specify parsing details such as the field separators so that tracking output format variations are possible while still allowing all variations to be parsed automatically. While providers are free to add additional data fields, unfortunately the standard itself does not make any attempt to standardise reporting of the accuracy of the tracking data (e.g. the Horizontal Dilution of Precision (HDOP) used to report the accuracy of GPS tracking data).

The W3C Semantic Sensor Network Ontology¹⁰ is the W3C standard for modelling sensor data in an machine-readable format. In the standard, a platform may have multiple sensors, which make observations

⁸<http://www.sportsdb.org/sd>

⁹FIFA, “EPTS standard data format” <https://football-technology.fifa.com/en/media-tiles/epts-standard-data-format/> Accessed: 2019-02-20

¹⁰W3C, “Semantic Sensor Network Ontology,” 2017. <https://www.w3.org/TR/vocab-ssn/> Accessed: 2019-05-02

of (real-world) properties. The standard provides the ability to model the accuracy and precision of sensor readings, and to associate these with operating conditions. Current applications¹¹ of the standard include Internet of Things, smart cities and environmental sensing.

Gudmundsson and Horton conducted a recent literature review [93] of spatio-temporal sport analysis techniques, and distinguish between *event logs* containing data points at the time of just key events (for example the set of time and location pairs from which goal attempts were made), and *trajectory* datasets containing regularly sampled data points (for example player GPS trajectories). In the design of the *R* software package “zoo”, *event logs* are classified as *irregular timeseries*, and *trajectory* data are classified as *regular timeseries* [229].

As highlighted above, there exists well established theory on measurement scales, there is research directed towards the challenges of processing and storing spatio-temporal data, and that there exists statistical and mathematical work on the proper treatment of errors in spatio-temporal datasets. The work of Chrisman [38] demonstrates the necessity of adapting existing measurement theory to cater to the subtly different properties that measurement scales acquire when contextualised to a specific domain. The FIFA EPTS standard data format demonstrates the ability to separate the representation of spatio-temporal data from the specific technology used to collect the data; however, does not currently standardise reporting the errors associated with data points. The W3C Semantic Sensor Network Ontology provides a general framework for formally modelling sensors and sensor accuracy, which in theory could be used to model sport sensor data; however, unlike the FIFA EPTS standard data format it has not been designed with the needs of sport specifically in mind. There exist mathematical models for precise treatment of errors in spatio-temporal data, but errors are often neglected by sport practitioners working with real game data, outside of isolated studies to validate equipment.

¹¹W3C, “On the usage of the SSN ontology,” 2019. <https://w3c.github.io/sdw/ssn-usage/> Accessed: 2019-05-02

3.4.2 Sensor Model

This sub-section provides a means to describe datasets in terms of the sensors that record the data. It bears some high level similarities to the W3C Semantic Sensor Network Ontology (e.g. the need to model platforms consisting of multiple sensors and to describe the accuracy associated with each recorded parameter). However, in contrast to the W3C Semantic Sensor Network Ontology, the focus is on describing the different types of spatio-temporal data encountered in sport (wearable position tracking devices, human data entry, etc.) rather than attempting to provide a machine-readable specification for data interchange of all forms of sensor data.

A sensor measures one or more *sparse* parameters together as a function of one or more *dense* parameters, for example, a GPS position sensor collects sparse position measurements as a function of dense time measurements. To distinguish between interlinked measurements versus aggregation of different datasets, the model distinguishes between a *sensor* and a *sensor platform*. A sensor platform is a collection of heterogeneous sensors. For example, a player monitoring device may contain both a GPS position sensor, as well as a heart rate sensor. Modelling these as separate sensors allows modelling the heart-rate sensor with a different sampling rate to the GPS position sensor. Note that a *measured parameter* represents a measurement of some real-world parameter, not the actual real-world parameter itself. Unlike real-world parameters, measured parameters represent samples at discrete intervals, and will contain an error due to the limitation of the sensor used. The proposed model is presented using UML domain modelling notation in Fig. 3.6.

There are multiple types of sensors ranging from humans (manual observations entered into a database) to wearable devices (electronic observations automatically streamed to a database). In some cases, multiple types of sensors can collect the same type of information, e.g. a human could manually annotate player paths, albeit tedious compared to computer vision approaches that automatically track players. The

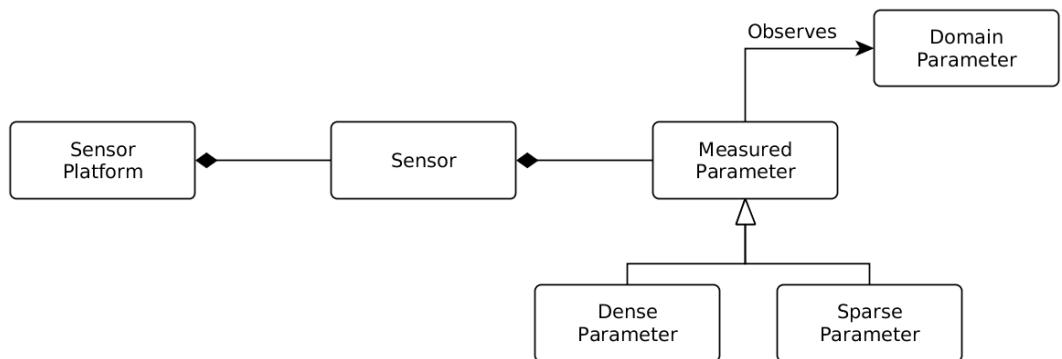


Figure 3.6: Abstract Data Model

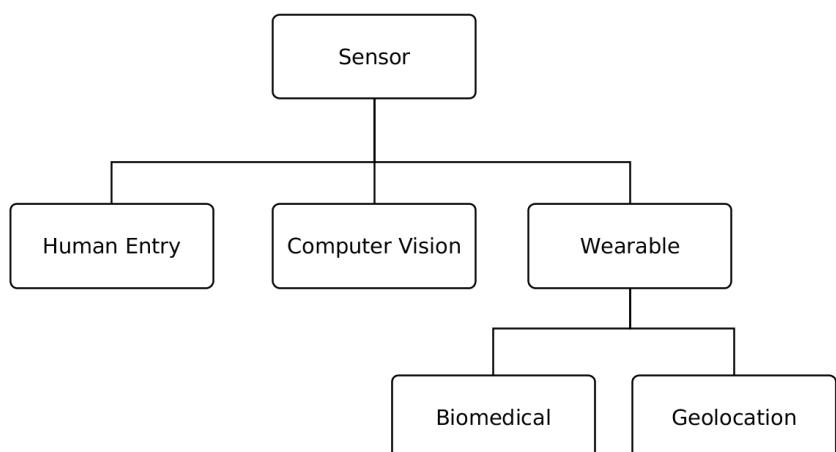


Figure 3.7: Sensor type hierarchy

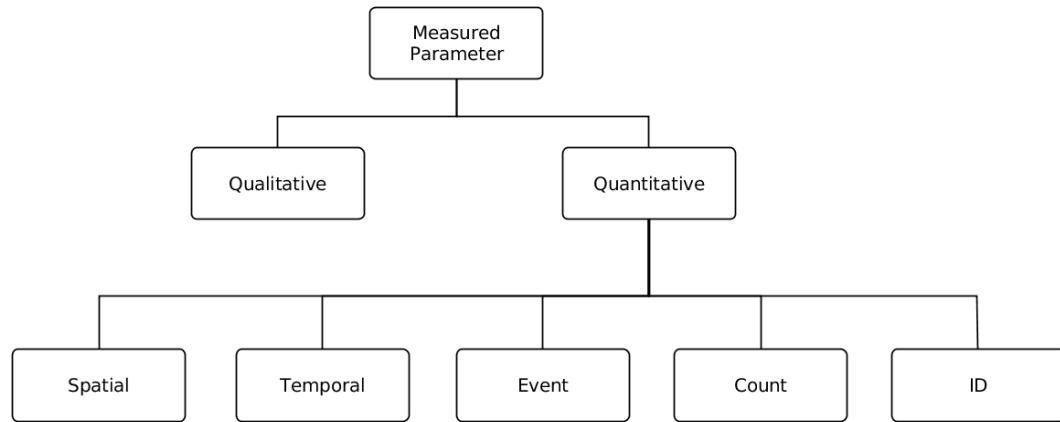


Figure 3.8: Measured parameter hierarchy

model is refined to categorise different types of sensors according to the hierarchy in Fig. 3.7.

Classes of measured parameters were identified to distinguish between fundamentally different kinds of unit systems encountered in datasets. In physics, the theory of relativity shows that space and time dimensions are intricately related, and hence space and time share common units (e.g. “light-years” describes spatial measurements in terms of time). However, for the purposes of sport, they can be considered independent. These classes of measured parameters in sport are presented as a hierarchy in Fig. 3.8. Note that each of these types can be either sparse or dense.

Each of these classes are listed in Table 3.1, along with required meta-data to describe these measurements. Accuracy is a critical consideration in description of any measurement. Measurements on a nominal scale, such as Events and Qualitative measurements, require specification of the granularity of the scale, in addition to accuracy.

Dense parameters require additional meta-data to describe the regularity of the sampling, as listed in Table 3.2. Sequences identifiers are always modelled as dense. The accuracy, error, and reliability attributes listed in Table 3.1 are optional for dense parameters in cases where the dense parameter is taken as a definitive (e.g. cell size), and error estimates of measured sparse parameters (e.g. count of players in cell)

account for any inaccuracies in observation of the dense parameter (e.g. inaccurate player counts due to ambiguities of the cell boundary).

Table 3.1: Attributes of Measurements

Measured Param.	Attribute	Example
Spatial	Error Radius	5 m error radius
Temporal	Accuracy	± 5 sec
Event	Accuracy	95% correct classification
	Granularity	{Pass, Tackle, Score}
Qualitative	Inter-rate Reliability	Cohen's kappa = 0.9
	Granularity	{Easy Pass, Hard Pass}
Count	Accuracy	± 1
ID	Accuracy	98% correct ID classification

Table 3.2: Attributes of Dense Measurements

Measured Param.	Attribute	Example
Spatial*	Cell Size	10 m \times 10 m square cells
Temporal*	Sample Rate	10 Hz
Count*	Bin Size	{(0,10),(10,20),(20,30),...}
ID* (Sequence)	Accuracy	95% follow correct sequenced

3.4.3 Definitions

Using the proposed model, one can now precisely define the difference between various levels of spatio-temporal data. The model can be used to describe either raw data, or processed data ready for visualisation. However, note that transformations during the processing of data (e.g. summarisation to reduce data, or calculation of new derived variables) can change the data type.

- **Event-Count data:** Sensor data dense in Event parameter, and sparse/dense in Count parameter.
- **Spatio-Temporal data:** Sensor platform with a sensor that collects any form of Spatial parameter and a sensor that collects any form of Temporal parameter. (e.g. time and distance of a sprint)

- **Time Dense Spatio-Temporal data:** Sensor data containing sparse/dense Spatial parameter, and dense time parameter (e.g. GPS trace).
- **Space Dense Spatio-Temporal data:** Sensor data containing dense Spatial parameter, and sparse/dense time parameter (e.g. static heat-map of time spent in each position).
- **Space-Time Dense Spatio-Temporal data:** Sensor data containing dense Spatial parameter, and dense time parameter (e.g. heatmap of recent events as function of time).

The model is applied to describe various sport datasets in Appendix A. In particular, the model was used to describe the dataset collected by wearable tracking devices used in this thesis (Appendix Sec. A.1.2). In applying the model, it became apparent that some relevant accuracy details were undocumented in product details, e.g. specifications describing the position accuracy and sample rate, but not how precise the internal clock was¹². An awareness of timing drifts is vital for analysis, as the distance players can move in a second is greater than typical positioning related error. This demonstrates the value of the model in drawing attention to all relevant parameters that could potentially affect the results of the analysis.

3.4.4 Conclusion

Sport performance datasets contain various forms of spatio-temporal data. An abstract data model was introduced and associated syntax for describing the different forms of spatio-temporal data contained in these data sets. It was demonstrated that the proposed model can be used to describe both traditional and contemporary sport perfor-

¹²In the case of GPS devices, satellites broadcast GPS timing information, which allows a theoretical accuracy of 40 nanoseconds. US Government, “GPS Accuracy,” 2017. <https://www.gps.gov/systems/gps/performance/accuracy/> Accessed: 2019-02-20. However when reporting this information accuracy can be degraded by conversion issues such as improperly accounting for leap seconds.

mance datasets. The model also revealed missing error estimates that are needed in order to systematically reason about the data quality.

3.5 Chapter Summary

This chapter developed models to describe the state of an AFL game and the sport performance feedback process. While of a high-level theoretical nature that require simplification to use in practice, they allowed positioning the thesis within the larger context of sport. The chapter also established an abstract data model for describing spatio-temporal datasets, which was applied to describe the diversity of sport datasets available, including the player tracking dataset used in this thesis.

Future Work

While the information theoretic framework presented in Sec. 3.3 is not practical to utilise numerically at this stage, it is included as a foundation for further work and is used to motivate the rest of the thesis.

Specifically, it provides a lens through which to answer Research Question 1 – *can team-level GPS analysis provide useful information to sport researchers and practitioners beyond what they already know from manual observation, video analysis, traditional statistics, and (individual) player GPS monitoring?* From an information theoretic lens, it becomes clear that the role of computational pipelines should be to provide coaches and sport performance analysts with insights that they cannot otherwise observe due to the limits of human information and processing channels. For this reason, the thesis seeks to design a platform for extracting new spatio-temporal insights into the game (see Chapter 7) in contrast to traditional summary statistics that tend to quantify aspects of the game that coaches can already observe.

Further work is needed to calculate the information gain delivered by

introducing more sophisticated sport analysis techniques that can surface previously untapped information from rich datasets, and to study the extent to which the information gain provided by an analysis technique correlates with sport performance analysts' perceptions of its usefulness. Accurately measuring the information gain requires more data than could be accessed for this thesis, but a brief proposal is outlined in Sec. 8.3.4.

Contributions

1. Structured AFL jargon into a formal domain model of consistent terminology, and used this to identify variables that form part of the game state. The full list of identified variables provides a holistic understanding of game state, and can increase awareness of the simplifying assumptions made by current sport analysis models.
2. Applied information theory to sport in order to provide a mathematically rigorous perspective for understanding the role of sport performance analysis systems within the larger sport context. Information theory was used to formalise the objective of performance analysis systems into a single formula, which states that the goal is to ensure information is valuable yet not already known to a coach, and incorporates the need to transmit this over limited human information channels.
3. Provided an abstract data model that permits modelling both dense and sparse spatio-temporal sport datasets, and draws attention to all required accuracy attributes that need to be specified in order to reason about the confidence of interpretations made from the dataset.

Chapter 4

Computational Pipelines

Contents

4.1 Background	80
4.1.1 Definition of Data Provenance	80
4.1.2 Designing for Data Provenance	81
4.1.3 Data Provenance Protocols	82
4.1.4 Computational Pipelines that Capture Provenance	89
4.2 Data Provenance for Sport	96
4.2.1 Abstract	96
4.2.2 Introduction	97
4.2.3 Motivating Scenario	98
4.2.4 W3C PROV	102
4.2.5 Our Notation	104
4.2.6 Comparative Evaluation	111
4.2.7 Key Findings	119
4.2.8 Conclusions	120
4.3 Chapter Summary	121

The previous chapters introduced sport technologies, and looked at common techniques in the literature to analyse spatio-temporal sport data. Compared to traditional sport statistics, spatio-temporal data analysis involves sophisticated modelling to mine the data for meaningful insights. Furthermore, it is common to combine multiple spatio-temporal analysis procedures together into a computational pipeline: e.g. a preprocessing stage to correct for field shape, followed by application of a data mining algorithm to detect patterns, followed by visualisation of the results to translate these patterns into a form that humans can interpret.

This chapter lays the foundations for describing and reasoning about computational pipelines for sport analysis. Many tools are available for constructing computational pipelines (e.g. operating systems, such as Unix, provide primitive operations for constructing pipelines where the output of one program is fed as input to another). However, an important criterion for sport analysis is to be able to trace the final output (e.g. a player rating) back to the source (e.g. the events that contribute to that player rating in the video recording of the match). This ability to trace the output of a pipeline back to the original source(s) is referred to in the computer science literature as *data provenance* and is the main property sought after in this chapter.

4.1 Background

4.1.1 Definition of Data Provenance

A desire to track provenance of information exists at multiple levels. In this thesis, *data provenance* refers to the capture of information, either prospectively or retrospectively, to allow the traceability of data that undergoes transformations back to the original source(s). For example, a statement of the form “dataset Y was arrived at by applying transformation operation T to dataset X ”. In contrast, *provenance* in the general sense is defined as the ability to trace some output back to the inputs

that *caused* it to occur. For example, a coach may wish to establish the provenance of a goal by tracing back the actions that contributed to it.

The former concept of *data provenance* can be formalised and solved through technological means. The latter concept of *provenance*, is general and difficult to formalise. As *provenance* relates to questions of causality, at a minimum, it requires a model that can be used to answer *what-if* questions to establish whether an alternative value for some input would have impacted on the final output. Attempts to design data provenance systems to support this more general concept of provenance have led to it being informally declared “an unsolvable problem”¹.

Due to the insurmountability of designing a protocol for capturing *provenance* in the general sense, this chapter will focus on design of a protocol for capturing *data provenance* in the stricter sense of data and the transformations applied to them. The purpose of the *data provenance* system is to establish a foundation that later chapters can build on top of. Its design is influenced by the higher level *provenance* questions that sport performance analysts seek to answer about the game.

4.1.2 Designing for Data Provenance

Data provenance systems can be designed systematically by considering the use-cases of queries a user wishes to ask of it [24]. Sport performance analysis can be thought of as a scientific workflow. Thus designing a data provenance system for sport performance analysis requires that the system can answer typical questions one would ask of a scientific workflow, such as which years of data were used to produce a specific graphical figure. It should also incorporate reasoning appropriate to the sport domain. For example, if a dataset includes data from games played simultaneously at different venues, the two games should be treated separately from each other for data provenance purposes to incorporate prior knowledge of physical limitations that prevent the two games from having any meaningful influence on each other. As another

¹[Presentation] Jennifer Widom, “Panda: A System for Provenance and Data”, GoogleTechTalks. <https://www.youtube.com/watch?v=tprA7a0b7Is> 12:18

example, one may be able to treat certain stoppages, such as events that return the ball to a centre bounce as *state resets* to decouple them from past actions and thus make human reasoning simpler due to the smaller provenance chain. There may also be provenance information that should *not* be tracked. For example, there may be a restriction that for player privacy, it should *not* be possible to trace information back to individual players (Chapter 5). In such a case, the data provenance foundations need to be designed to capture provenance information to support the questions one wishes to ask, while being able to reason with partial data due to the need to hide certain details of how the data were derived.

4.1.3 Data Provenance Protocols

This section provides an overview of data provenance systems and representations proposed in the literature. It identifies W3C PROV [147] as influential in standardising the core concepts needed to represent provenance information. Recent developments attempt to increase the ease and level of detail with which one can capture provenance information.

2005

Simmhan et al. [188] provide a taxonomy for describing data provenance systems in terms of “use of provenance”, “subject of provenance”, “provenance representation”, “storing provenance” and “provenance dissemination”. They outline multiple needs for provenance, including: estimating data quality; auditing; ability to replicate processes with updated inputs; attribution to ensure proper citation or for compliance with intellectual property laws; and informational reasons to allow querying the provenance to help understand the data, or to search out datasets that were produced in a particular manner (for example, when conducting a meta-review, one might limit their search to datasets that comply with a certain methodological requirements, such as having a randomised control). While their taxonomy considers

the representation of provenance information, surprisingly their paper does not consider how the representation of provenance information relates to the representation of the underlying workflow. Nevertheless, many of the systems they surveyed used, either explicitly by design or implicitly through dependencies, a Directed Acyclic Graph (DAG) to represent the workflow, coupled with a provenance system that either *eagerly* produces provenance information upon execution of the workflow (i.e. provenance metadata is produced as an additional artefact alongside the other data transformations), or *lazily* allows determination of provenance after execution by inverting transformations (inverting transformations requires that they are analysable, thus most of the work for inverting transformations focuses on inverting standard SQL queries rather than arbitrary code).

2006

Bowers et al. [24] describe a provenance capture format and query mechanism for the “Kepler” scientific workflow system. They introduce the Read, Write, State-Reset (RWS) provenance model, a minimalistic system for capturing provenance while allowing complex workflows such as those that involve sliding time windows and recursion. Logging groups (called “firings” in their paper) of reads and writes allows tracking data dependencies. Explicit state-resets prevent implicit dependency on all previous inputs. They utilise Datalog (a specialisation of the logic programming language Prolog) to reason about provenance and allow reconstruction of DAGs describing data provenance from the event log. They distinguish their work from existing systems by the focus on general purpose scientific workflows (as opposed to the existing database literature) and focus on “user-oriented” queries (suggesting the queries a user wishes to perform can be used as use-cases to drive the design of a suitable schema for capturing provenance).

2008

Wang et al. [216] implement a system for data provenance of geospatial datasets, intended for use with Geographic Information Systems (GIS). As working with geospatial datasets typically involves integrat-

ing many different datasets from different sources whilst at the same time requiring knowledge of the accuracy of the resultant dataset, the demands of GIS have been one of the early motivations for data provenance, with GIS provenance meta-databases emerging as early as 1991. While Wang et al. only provide a course-grained model of provenance, they note some unique characteristics of data provenance for spatial datasets, specifically that provenance can be limited to spatial bounds, demonstrated through the example that a rainfall data error for one city is not likely to undermine an analysis that involves rainfall data of a city in another state, even if they are both derived from the same country-wide rainfall dataset.

2010

Ikeda & Widom [108] implement a data provenance system, motivated by the need to debug complex analysis pipelines, and to allow recomputation of the affected elements when data errors are discovered. They describe two primitive data operations that a data provenance system should support: *backward tracing* (*which inputs contributed to a given output element?*), and *forward tracing* (*which outputs are influenced by a given input element?*). These primitive operations can be combined to support data *refresh* (i.e. *backward tracing* an output value to the inputs that contributed to it, then recomputing the output using updated input values). In Jennifer Widom's Google tech talk, she opines that after her three attempts at designing a provenance system "I actually have a fairly strong opinion about the area of data provenance, which is that you're never going to solve it. It's an unsolvable problem. There is never going to be a general framework"² on the basis that, much like data integration, data provenance is too general for any one-size-fits-all solution.

2011

The Data Documentation Initiative (DDI) [90] is a set of guidelines/standards for data provenance that have gained prominence as a means

²[Presentation] “Panda: A System for Provenance and Data”, GoogleTechTalks.
<https://www.youtube.com/watch?v=tprA7aOb7Is> 12:18

of archiving social science data to facilitate reuse. In a recent talk at the Australian Research Data Provenance (RDP) Interest Group, Steve McEachern explains how the DDI standard has come to be “used by about 90 different countries [particularly OECD countries], [and used by] the World Bank and World Health Organisation”³. Despite modern questionnaires being conducted electronically, a large effort is currently required to convert documentation produced by the survey tool into a standardised form used by DDI [211]; this is particularly problematic because provenance information is lost when researchers use scripts (e.g. written in R) to transform their dataset without consideration of how to also transform the survey metadata. However, the C2Metadata⁴ project aims to bridge this gap by parsing common scripting languages used in the social sciences, to analyse the transformations and create a metadata pipeline that mirrors the data pipeline.

2012 - 2015

PROV [147] is the W3C standard proposed for provenance information on the Web. It models provenance using the concepts of *Entities* (data/artefacts), *Activities* (processes, whether manual or automated, that produce or derive new entities), and *Agents* (the person or system responsible for carrying out the activities), shown in Fig. 4.1. The focus of the PROV standard is on simplicity (while allowing extensibility for aspects that it cannot capture appropriately) and interoperability between different provenance systems. As PROV activities (processes) are a “generic concept”, it is easy to model provenance as part of systems that require manual human input. However, it also means that PROV activities are not necessarily reproducible.

2016

Regalia et al. [173] propose the *VOLT* language to add provenance on

³Presentation: Steve McEachern (Director, Australian Data Archive), 2017, “Managing provenance in the Social Sciences: The Data Documentation Initiative (DDI)” in “Provenance and Social Science data - 15 Mar 2017” <https://youtu.be/e1PcKqWoOPg?t=12m24s>. Presented at the Australian Research Data Provenance (RDP) Interest Group <https://www.and.org.au/partners-and-communities/ands-communities/data-provenance-interest-group>

⁴<http://c2metadata.org/>

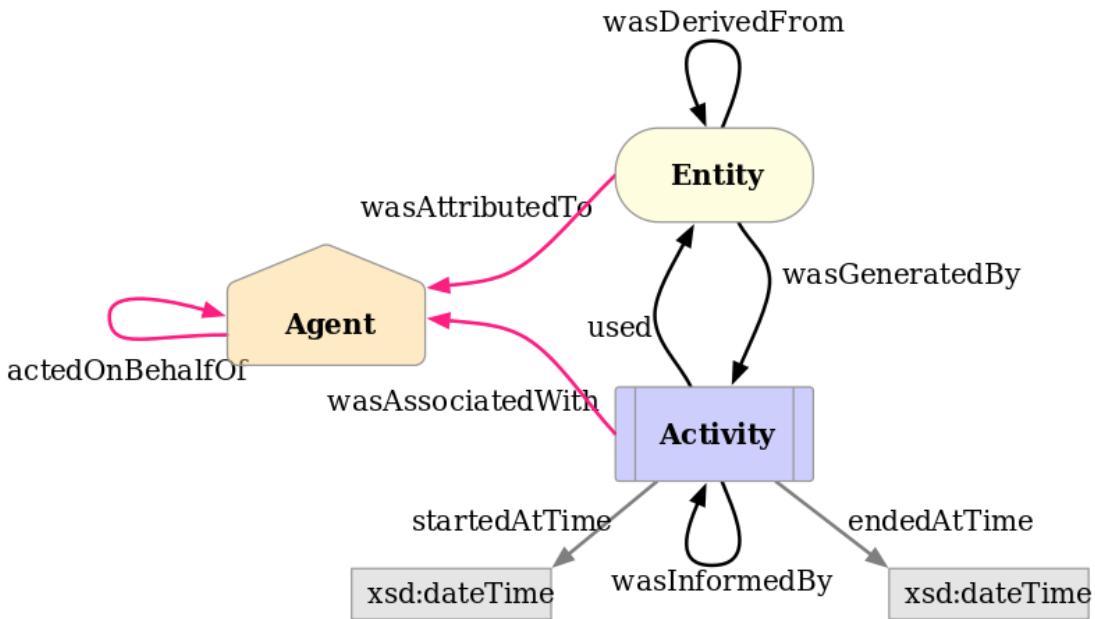


Figure 4.1: Concepts in PROV, copied without modification from
<https://www.w3.org/TR/2013/REC-prov-o-20130430/>

top of *SPARQL* (the W3C standardised query language for semantic databases). *VOLT* works as a proxy, that dynamically generates derived properties on demand, then queries the result using standard *SPARQL* as if the properties had been manually defined. Furthermore it captures provenance. It is demonstrated with a case study of spatial queries to dynamically generate compass direction relations between nearby locations (e.g. A is *east of* B). This is both more accurate and storage efficient than manually defining the compass direction between every possible pair of locations. The system supports provenance in two ways. Firstly, all queries run are serialised as an abstract syntax tree along with their outputs and stored in a temporary *results graph* or persistent *output graph*. This allows the framework to determine whether a cached result is still valid, or whether the inputs have changed and it needs to be recomputed. The user can also perform queries over the results graph itself to query the provenance of a result stored in the results graph (such as which data sources were used, or which functions the query depended on). Secondly, functions in *VOLT* optionally support collection of additional metadata that help the user understand how the result was computed. The authors of *VOLT* provide the example that an operation to sum up attributes from multiple geographic regions

would also capture the combined geographic region involved in the calculation as well as any overlap between regions. The user can visualise this metadata to as a form of assurance that the query was calculated as expected (e.g. if visualising the overlap between regions reveals significant overlapping area, this may indicate unintentional double counting when summing up attributes from these regions).

D. Wu, L. Zhu (Data61), et al. [224] describe “Pipeline61”, a tool for constructing and managing pipelines that contain components that are implemented in different big data frameworks and data stores. While not the main focus of their article, their pipeline is also designed to capture data provenance in a manner that integrates with the need for dependency management and version control of components.

To do this, they capture an *execution trace*, a *dependency tree* and a *data snapshot*. The execution trace captures a DAG of the pipeline each execution. Capturing a separate DAG each execution allows version control of the pipeline itself to deal with configuration changes such as additional inputs or different versions of components. The *execution trace* itself does not include data or source code, so the cost of capturing a complete DAG each execution is minimal. The *dependency tree* captures the dependencies of each component: the environment it depends on, the path to the component source code, and the executions it was used on. The description of the environments is constrained to a set of prespecified labels such as SparkPipe or ShellPipe which are supported by Pipeline61. In their example, the source code of the component is assumed to be a single file, so it is unclear how one would represent components with dependencies on specialised libraries or components that share code with each other. The tracking of the executions the component was run on as part of the dependency tree appears to be redundant considering that this information is also in the execution traces (perhaps the authors intended it as a view rather than a capture of duplicate information); however, exposing this information helps regression test new components against real inputs and outputs fed to the previous version of the component in production. The *data snapshot* captures all inputs to each version of a component, the output, and the execution state (i.e. whether there were any errors). The data

captures themselves are identified by the component, component version, and execution trace ID. The use of execution trace ID as part of the data capture ID would seem to prevent the ability to perform an incremental build that can reuse the output of initial stages of the pipeline that haven't changed (perhaps the authors desired the entire pipeline to be rerun each execution due to *non-pure*⁵ functions in the pipeline that may generate a different output for the same inputs, interact with an external system, or produce errors⁶ that resolve themselves after restarting the process).

2017 – 2019

As W3C PROV can be tedious and error-prone to generate, Moreau et al. [146] describe a templating system to generate valid W3C PROV statements. The templates are expressed as PROV documents containing variables as placeholders. This simplifies the process of modifying an application to output provenance information, as all the application needs to generate is a mapping that binds template variables to values. These bindings are then used by the templating system to expand the templates into full PROV statements (or can be left as is in compressed form).

Stamatogiannakis et al. [198] describe *PROV_{2R}*, a tool for capture of provenance information on systems where one is using off-the-shelf software, and does not have the ability/resources to modify the software itself to output provenance information. This is achieved by running the whole system within an emulated machine using PANDA [60], which captures a re-playable snapshot of the system along with a log of all non-deterministic operations. These logs can then be replayed with taint-tracking at the desired granularity to determine which inputs affect which outputs. The resultant logs can be exported to W3C PROV and queried.

⁵“purity” is used in the sense of the functional programming paradigm. I.e. a “pure” function is like a mathematical function that has no effect other than returning a value in contrast to a “non-pure” function that has side-effects that alter the program state.

⁶e.g. errors could result from race conditions in poor concurrent programming implementations, or be induced by spontaneous hardware faults such as “soft errors” arising from ionising radiation

4.1.4 Computational Pipelines that Capture Provenance

For human interpretation, data processing workflows are commonly represented as a flowchart showing the data inputs, processes that operate on them, and final outputs of the system. For purposes of workflow automation (i.e. allowing machines to automatically execute the workflow) the data processing flowcharts are typically formalised as a graph (also known as a network) data-structure [24, 224]. Nodes (also known as vertices) are used to represent inputs and processes, and edges (also known as arcs) are used to represent the flow of data between each process. The edges are directed to represent the direction of data flow (the input dependencies of a processes are in the reverse direction to the data flow). Under the assumption that there are no cyclic dependencies (i.e. an output that (indirectly) depends on itself), the graph is said to be a *Directed Acyclic Graph* (DAG). An example workflow (pipeline) for computing a sport player's goal accuracy is provided in Fig. 4.2.

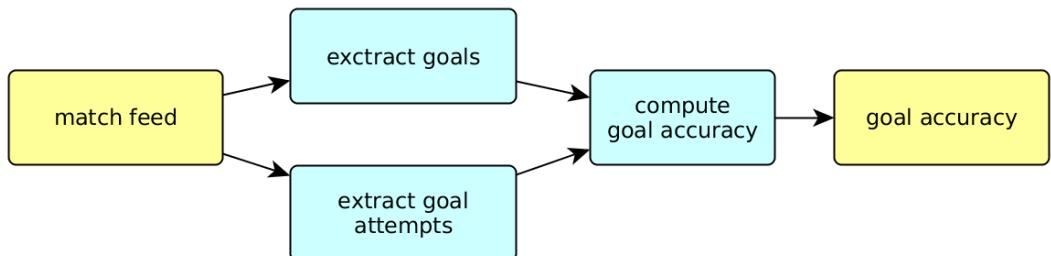


Figure 4.2: Example of a simple data processing workflow flowchart (formally, a directed acyclic graph)

Optionally, additional nodes can be introduced to represent the intermediate data output of each process. Under this representation, there are two types of nodes: data nodes and process nodes, with chains that alternate data, process, data, process, etc. While this form is more cumbersome to write, it can help with formalisation. This is similar to the notation used by *Data Flow Diagrams* (DFDs). An example pipeline marking these intermediate data outputs is shown in Fig. 4.3.

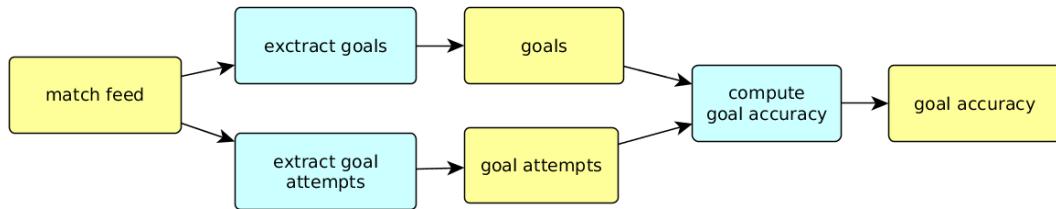


Figure 4.3: Example of a simple data processing workflow with intermediate outputs explicitly shown

While such pipelines are sufficient for describing processes, or in some cases automating them, it is also necessary to consider the description of the processes themselves. Assuming the process can be precisely described as written software (as opposed to a task that a human performs), then there is a need to track the version of the software used, as well as the environment that it requires. While in principle, software source code should unambiguously describe the actions it performs, in practice source code may have complex dependencies and assumptions about the environment that need to be met in order to build it into an executable. As the provenance of data depends critically upon the processes that are applied to it, capturing the provenance of the software is a prerequisite for capturing the full provenance of the data. In the software engineering community, the ability to reproduce the generated executable byte-for-byte from the software source code is known as a “reproducible build”. While many mainstream software packages do not offer reproducible builds⁷, certain security focused software such as the IP anonymisation network Tor and the encrypted messaging app Signal⁸ offer fully reproducible builds⁹.

However, reproducible builds only concern themselves with deriving software from source code. It is left to the user to manage provision of inputs and outputs to the resultant software artefact. Note however,

⁷E.g. the GNU/Linux Operating System distribution Debian notes that “Reproducible builds of Debian as a whole is still not a reality”. <https://wiki.debian.org/ReproducibleBuilds>. Accessed: 2019-01-19

⁸Blog: Signal, 2016, “Reproducible Signal builds for Android”. <https://signal.org/blog/reproducible-android/>

⁹A short history of open source software supporting reproducible builds is maintained by Wikipedia Contributors, “Reproducible builds”. https://en.wikipedia.org/wiki/Reproducible_builds Accessed: 2019-01-16

that code and data are equivalent. A Turing machine, a mathematical model for describing computations, takes data on a (theoretically infinite) tape as its input, but can also be used to execute or manipulate a program on the tape as if it were data. Similarly, inputs that are supposedly data, may actually function as code¹⁰, a fact that malicious users often take advantage of in order to execute arbitrary computations.

Development of reproducible build tools has focused of code transformations while ignoring the user supplied data inputs and outputs of the resultant program. In contrast, development of data provenance frameworks has focused on the flow of data inputs through data processing pipelines with little attention to capturing the provenance of the software itself that makes up the pipelines. However, the equivalence of code and data means that reproducible builds and data provenance capture are equivalent problems which can be unified into a single framework. A simple example to demonstrate the equivalence of data provenance capture to reproducible builds, is presented in Fig. 4.4.

On the left of Fig. 4.4 is an illustrative example of a code build process (the example is of compiling a program written in the C programming language; however, the programming language being used is immaterial here). The source code undergoes a series of compilations and linkage steps. At each stage an external program (part of the C compiler toolchain) transforms the code into an alternative form, typically with the goal of bringing high level code written by human software developers into a machine interpretable form. There are a range of build automation tools available, but conventionally one would use a Make¹¹ script.

On the right of Fig. 4.4 is an example of a simple data processing pipeline. The data undergoes a series of extraction and analysis steps that transforms the data into an alternative form. At each stage an ex-

¹⁰For example, the blog “Accidentally Turing-Complete” contains a list of data processing systems (such as templating engines) that have been proven Turing Complete despite never having been intended as programming languages.
http://bezale1.tuxen.de/articles/accidentally_turing_complete.html
Accessed: 2019-01-16

¹¹[https://en.wikipedia.org/wiki/Make_\(software\)](https://en.wikipedia.org/wiki/Make_(software))

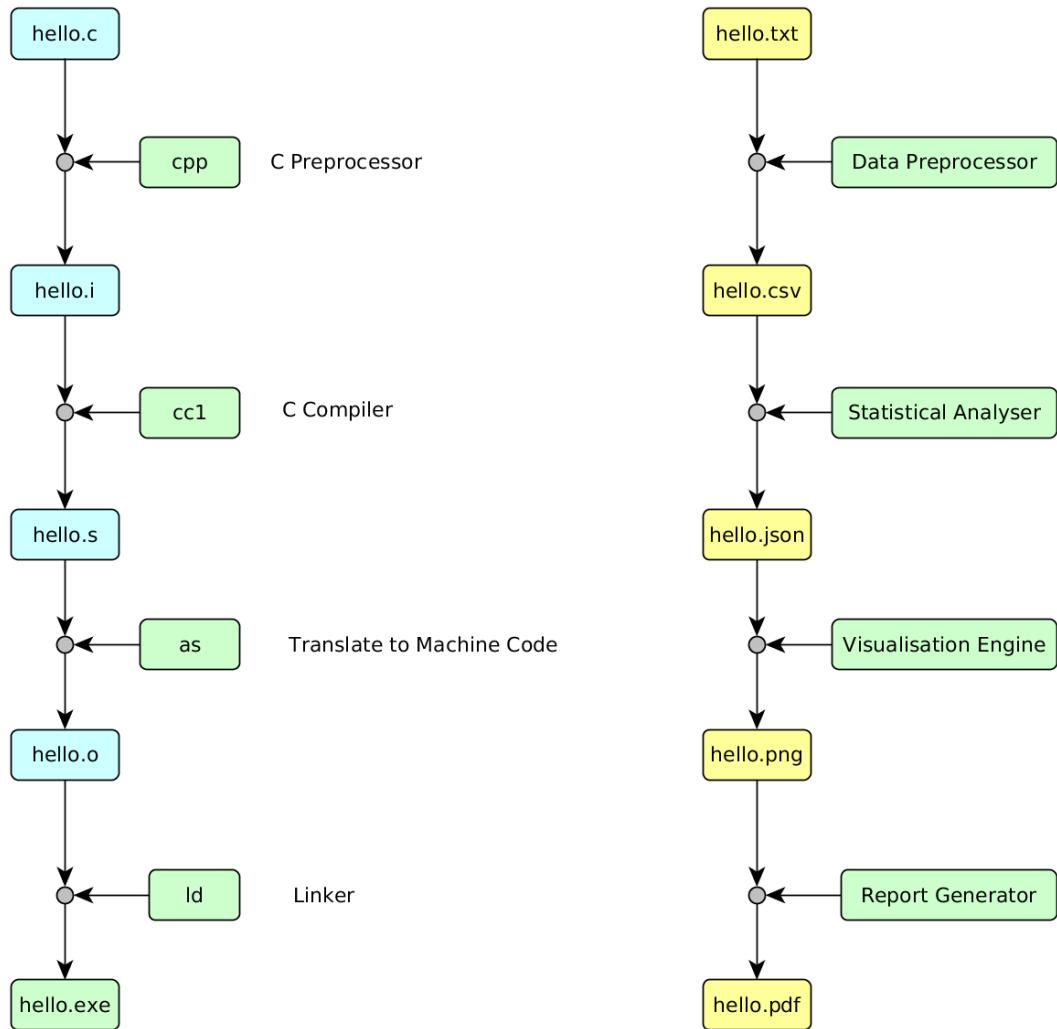


Figure 4.4: Comparison of code build pipeline (left), and data processing pipeline (right). Source code artefacts are coloured blue, executable programs are coloured green, and datasets are coloured yellow. The distinction between source code, executable programs, and data is by convention (i.e. source code and executables are just specialised forms of data, but conventionally treated differently to datasets)

ternal program (e.g. bundled within a statistical analysis suite) transforms the data into an alternative form, typically with the goal of bringing raw data (often machine collected) into a higher level form interpretable by human data analysts.

There are similarities between the code build process and the data processing pipeline:

1. While the build process operates on source code rather than datasets, source code is just a form of structured data. Furthermore, when operated on by a sufficiently sophisticated analysis, datasets themselves can act as source code. For example, if the dataset is processed by an inference engine, such as by converting facts in the dataset to statements in the Prolog logic programming language (which is Turing complete, i.e. capable of expressing arbitrarily complex programs), then the dataset functions as a form of source code.
2. The final output of the build process is instructions interpretable by a machine. The final output of the data processing pipeline is information interpretable, or sometimes directly actionable, by a human.

Note how these two tools may be unified to form a more complete picture of data provenance. Data processing pipelines track the flow of data acted upon by external executable programs (data preprocessor, statistical analyser, etc.). Reproducible build tools can be used to capture the full details of the build process used to arrive at these executable programs, including the recursive dependencies of the build such as the C preprocessor, C compiler, etc.¹² In theory, reproducible build tools could also be used to construct complex data processing pipelines by tracking the user supplied data sources as dependencies of an ad-hoc build that is executed whenever a user runs an analysis on a data input. For example, the data processing pipeline on the right of Fig. 4.4 could in theory be represented as a reproducible build in the same way as software, the only limitations are that current reproducible build tools have not been designed to support this use case, thus the syntax for describing the build is awkward when one wants to perform quick ad-hoc analysis rather than reusable software.

¹²In practice, C compilers such as GCC are bootstrapped using an existing (possibly older) C compiler. However, in principle, all computations could be described using a small core of primitive concepts such as lambda calculus.

Data Structures

This provenance information can be captured through triples (3-tuples):

$([Data\ input\ hash], Executable\ hash, [Data\ output\ hash])$

Optionally, these can be extended to an n -tuple to include auxiliary information such as timestamps. However, timestamps are not strictly necessary, as reconstruction of the execution graph allows sequencing the operations according to a partial ordering. Other auxiliary information one might want to capture includes system state that may help to with auditing if the execution is performed erroneously. However, for clarity of the provenance information, it will be assumed that the execution can be carried out correctly, for example by distributing the same computation out to independent computation nodes for validation¹³.

Query of Provenance using Capture Log

Capture of this information is enough to reconstruct the entire workflow graph and use it to query for provenance information. For example, given an output (such as “output.csv”), and a full capture of the provenance triples, the provenance graph can be reconstructed to determine the inputs that contributed to “output.csv”. If intermediate results are lost, they can be recomputed from the original inputs. An example workflow is provided in Fig. 4.5. Note that the “Client” in the diagram (e.g. analysts within the sport team), can trace the full provenance for confidential data (e.g. to re-identify individual players), whereas the “AI Service Provider” (e.g. external researchers working with sample de-identified data) can never see the confidential data (i.e. they only have a partial view of the system).

¹³For example, in order to ensure smart-contracts are carried out correctly, the Ethereum blockchain runs the same calculation within a (highly optimised) virtual machine on every node on the network to obtain a consensus. This allows strong guarantees that the computation was performed correctly. However due to the high redundancy of having every node run the same operation, this method is only appropriate for small computations. “A Next-Generation Smart Contract and Decentralized Application Platform” <https://github.com/ethereum/wiki/wiki/White-Paper>. Accessed: 2019-01-16

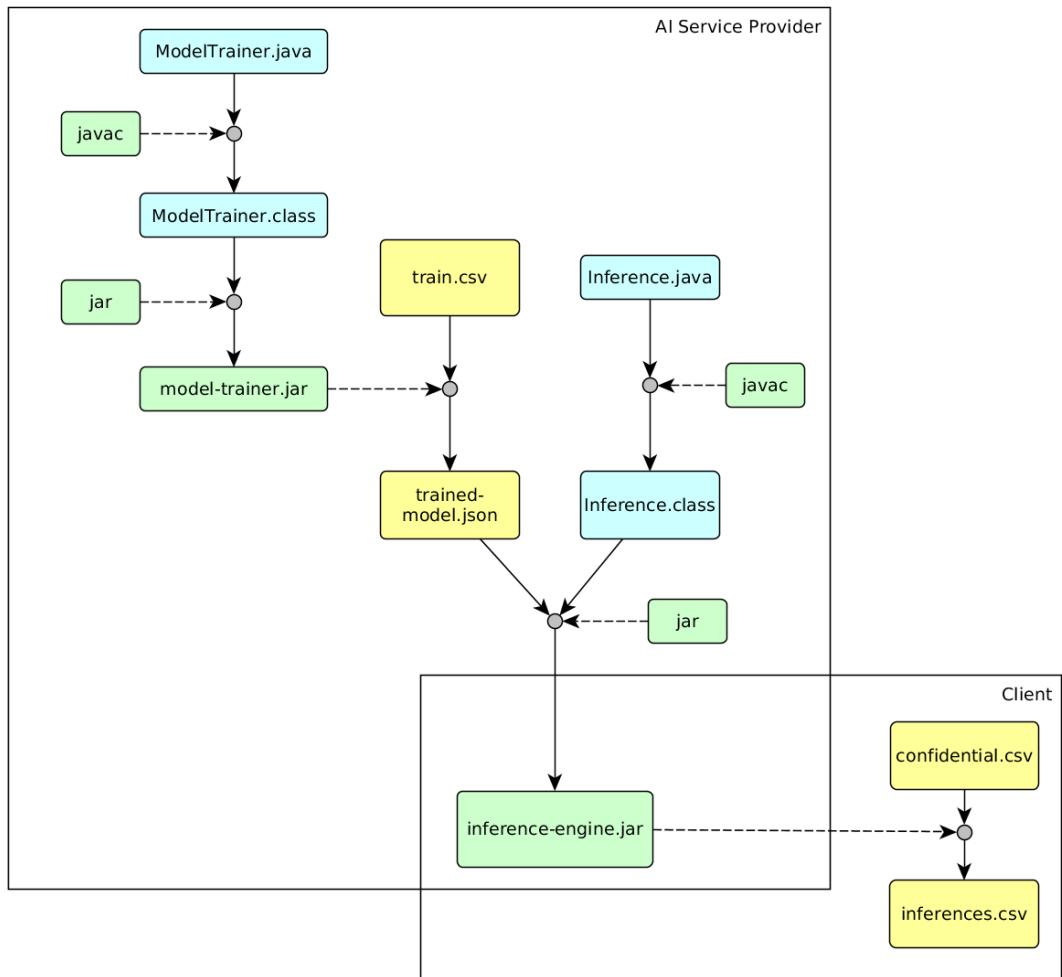


Figure 4.5: Sample Directed Acyclic Graph (DAG) for machine learning pipeline. Note the similarity of code artefacts (blue, green) and data artefacts (yellow). This suggests that the goal of achieving reproducible machine learning pipelines is equivalent to the problem of reproducible builds.

Relevance to Sport

The value of having data structures to represent the computations performed in the form of a capture log, is that this allows queries to be run over the capture log to answer different data provenance questions. Specifically, in the context of sport, those data provenance queries would be questions of relevance to a sport performance analyst, such as “*which years of game data were used to generate this analysis output?*”

Ultimately, the motivation for exploring computational pipelines and data provenance in this thesis, is to support the questions that sport coaches and practitioners would want to ask. This also includes surrounding questions that a sport performance analyst may need answered to verify the correctness of information. The next section of this thesis elaborates on what kinds of sport related questions could be expected, as well as the human-computer interface concerns that need to be considered to ensure that the results can be interpreted by sport practitioners.

4.2 Data Provenance for Sport

This section of the thesis relates to the (draft) publication Andrew J. Simmons et al. “Data Provenance for Sport”. In: arXiv e-prints (2018). arXiv: 1812 . 05804. An authorship statement for the paper can be found in Appendix F.

4.2.1 Abstract

Data analysts often discover irregularities in their underlying dataset, which need to be traced back to the original source and corrected. Standards for representing data provenance (i.e. the origins of the data), such as the W3C PROV standard, can assist with this process; however, they require a mapping between abstract provenance concepts and the domain of use in order to apply them effectively. This section proposes a custom notation for expressing provenance of information in the

sport performance analysis domain, and maps the proposed notation to concepts in the W3C PROV standard where possible. It evaluates the functionality of W3C PROV (without specialisations) and the VisTrails workflow manager (without extensions), and finds that as is, neither are able to fully capture sport performance analysis workflows, notably due to limitations surrounding capture of automated and manual activities respectively. Furthermore, their notations suffer from ineffective use of visual design space, and present potential usability issues as their terminology is unlikely to match that of sport practitioners. These findings suggest that one-size-fits-all provenance and workflow systems are a poor fit in practice, and that their notation and functionality need to be optimised for the domain of use.

4.2.2 Introduction

Sport performance analysis involves a combination of manual annotation of video, automatable derivation of performance statistics from the annotations, and ad-hoc interplay of manual and automated processes to refine data and define new metrics. The competitive nature of sport, and the explosion of available data captured by in-game sensors, has led to demand for increasingly sophisticated forms of analysis. However, without some form of data provenance describing all processes and data sources used in the derivation of the final performance statistic, there is limited ability to reproduce the analysis, nor to audit the process for human error, software bugs, or data entry errors that may have affected the result.

Sec. 4.2.3 provides a motivating scenario inspired by real challenges faced by sport performance analysts, and highlights the need for data provenance to audit and reproduce the processes. The pain points expressed in the motivating scenario are used to elicit requirements, that form the basis for our proposed provenance notation optimised for sport performance analysis.

Sec. 4.2.4 reviews the W3C PROV standard and existing workflow man-

agement systems such as VisTrails. Sec. 4.2.5 introduces our proposed notation.

Sec. 4.2.6 evaluate the functionality, notational effectiveness, and usability of existing tools for the description and capture of data provenance, specifically the W3C PROV standard and the VisTrails workflow manager. Finally, Sec. 4.2.7 identifies shortcomings of existing systems, and Sec. 4.2.8 concludes with recommendations on how to bridge the language gap between abstract provenance concepts and the sport performance domain.

4.2.3 Motivating Scenario

Consider Ellie, a high performance sport performance analyst for an Australian Rules Football team, who wants to test a new player evaluation metric.

Physical Provenance Scenario

Ellie begins by annotating video footage of past games using a timeline annotation tool, such as Sportscode¹⁴. From the centre bounce (start of play), Player 3 taps the ball to Player 12, who kicks it to Player 7, who scores a goal. As per the laws of the game, after the goal, the ball is returned to the centre of the field for the next centre bounce.

Upon annotating the video footage from all past games, Ellie decides to investigate one of the goals in more detail. For example, she might want to investigate goal assists that led to scoring the goal (assume that the club does not already have a custom label to represent the set of goal assists). While she can rewatch the video footage, ideally she would like to be able to extract an abstract representation of the provenance of the goal (i.e. how the goal came to be) using the data that she has coded in order to allow her to efficiently investigate a large number of cases

¹⁴<https://www.hudl.com/elite/sportscode>

without needing to rewatch the footage.

Within her timeline tool, Ellie is able to search for a goal and scan back in time to see the possession chain; however, her timeline is cluttered with additional annotations such as the medical team’s annotation of an on-field injury to Player 3’s knee during the centre bounce. While she can hide certain event types, she cannot instruct her timeline tool to automatically hide everything that did not contribute to the goal, as her timeline tool has no concept of how events are connected to each other. Furthermore, she sees events prior to the centre bounce and after the goal, as her timeline tool does not recognise that these events reset the game state.

Workflow Provenance Scenario

Ellie’s timeline tool allows her to qualitatively analyse specific events through the medium of video, but does not provide a way for her to directly compute custom metrics from her annotations. To do so, she exports her timeline annotations to an intermediate format (e.g. CSV), so that she can statistically analyse the data using an external analysis tool (e.g. Microsoft Excel).

Prior to conducting the analysis, Ellie de-identifies the exported annotation data by substituting player identifiers with anonymised codes. This allows her to collaborate on the analysis with external researchers who for privacy reasons should not be given access to identifiable player data. Ellie retains a private copy of the mapping between player identifiers and anonymous codes.

Using her analysis tool, Ellie imports the de-identified game annotations, and—with some assistance from her research collaborators—computes the player evaluation metric for each (anonymised) player. Once the analysis is complete, Ellie re-identifies the players in the final output using the mapping she kept.

Player 7 is upset at the result of their metric, and requests to see game

video clips of events that contributed to the calculation. Fortunately, Ellie saved the intermediate calculation spreadsheet, but the calculations are difficult for Ellie to explain, as the inputs are expressed as numerical time offsets rather than embedded video clips, and furthermore the calculations were performed using anonymised identifiers. In order to allow the player to audit the calculations, Ellie has to reverse the process by looking up the anonymised identifier for Player 7 such that she can find the relevant calculations, then extract video segments for each time offset associated with inputs to the calculations records.

Upon scrutinising the raw video with the player, Ellie notices that the video shows that one of the missed goals was due to high wind conditions rather than the fault of the player. Ellie manually edits the exported timeline to exclude the period with high wind conditions, and re-runs her calculations. However, she has to be cautious that her manual changes aren't accidentally overwritten when she next exports timeline data.

Streaming Scenario

The coach is impressed with Ellie's proposed metric, and asks if she could annotate the game live as it is played and provide regular updates of each player's metric over the course of the game. While existing timeline annotation tool interfaces provide buttons and hotkeys to support a sufficiently fast data entry rate to allow annotating the game live, Ellie's current workflow for calculating her metric requires manually exporting the data and running a computationally intensive process. She needs a mechanism to automatically recompute the results in real-time as new data become available.

Requirements Elicitation

Requirements for the solution were extracted from the pain points outlined in the above tasks. These are presented in Table 4.1.

Table 4.1: Requirements to support tasks performed by sport performance analyst

Requirement	Description
Integrated support for working with video data	The ability to interactively annotate segments of a video timeline as events of interest, capture the relationships between these events, and to visually playback the video for an event.
Support for automated processes	The ability to automate interconnected computations such that they can be recomputed on an updated dataset with minimal manual intervention.
Support for manual interaction	The ability to interweave manual processes with automated processes within a workflow, and to manually override the result of automated processes.
Partial / shared workflow graphs	The ability to share different parts of the workflow with different users (e.g. external collaborators should not be able to reverse the de-identification operation), and to merge changes from other users (e.g. changes suggested by external collaborators) back into one's own workflow.

Provenance / Reverse Debugging	The ability to trace the provenance of an analysis result back to the raw inputs that contributed, and to scrutinise the intermediate calculations at each step of the process.
Streaming data	The ability to perform calculations in real-time as new data become available. To prevent latency, automated processes should be performed in parallel where possible, and recompute only what is necessary. Similarly, any manual processes in the workflow should be crowd-sourced to a team of annotators to prevent bottlenecks.

4.2.4 W3C PROV

Sport performance analysis is a form of applied sport science, and implicitly involves the construction of scientific workflows to analyse data (note that workflows can involve ad-hoc human tasks, and are not necessarily formally documented, if at all).

Scientific workflows [85] may involve both manual and automated processes, as well as ad-hoc data transformations to explore the data from different perspectives [113]. It is generally accepted that one should, in principle, be able to reproduce the steps in order to obtain the same final result. In practice however, science is facing a “reproducibility crisis” [13] wherein researchers are unable to reproduce others’ results, or in many cases their own. Data provenance systems aim to alleviate this issue through support for capture and query of information pertaining

to the origins of data, such as the primary data source, processes applied, and agents (i.e. both humans and software) involved.

While systems for automated workflows and provenance capture have gained traction in specialised domains such as bioinformatics, the use, or indeed recognition of the need for provenance more generally, such as in the biomedical field as a whole remains “quite low” [18].

Prominent scientific workflow management tools include VisTrails [28], Taverna [164], and Kepler [24]. VisTrails and Taverna represent the workflow of tasks as a directed acyclic graph (DAG), while Kepler provides the user with a choice of the model of computation that will be used. Workflow systems can be integrated with data provenance systems in order to capture both the process (prospective provenance) and trace of results (retrospective provenance) [140].

The W3C PROV standard [147] was introduced in 2012 in an attempt to standardise provenance sharing on the Web. The PROV standard is a component of the semantic web that cross-cuts the ontology, logic and proof layer of the semantic web [145] (note that these layers were part of the semantic web vision, but some aspects, particularly the proof layer, remain “largely unrealised” [184]). Since its release, PROV has been proposed for a range of applications including tracking the source of citation information in curated citation databases [168], as an export format for Git version control history [54], and as a tool for coordination of human and autonomous agents in disaster response [171]. VisTrails and Taverna both support export of data provenance information according to the W3C PROV standard.¹⁵ ¹⁶

According to the W3C PROV specification, entities may be “physical, digital, conceptual . . . real or imaginary”¹⁷. This has led others to consider the use of the specification as a means to model physical provenance, such as the process of creating scientific specimens [47], and as a tool for modelling the provenance of food [17] to infer sources of contamination. When modelling the provenance of physical systems in this

¹⁵<https://github.com/taverna/taverna-prov>

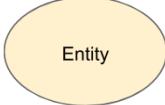
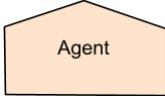
¹⁶“PROV support” <https://github.com/VisTrails/VisTrails/issues/1075>

¹⁷<https://www.w3.org/TR/2013/REC-prov-o-20130430/#Entity>

manner, provenance is often assigned a causal definition (i.e. arrows represent causality rather than just dependency), which may optionally be supplemented with probabilities to permit Bayesian reasoning using the provenance graph [33].

4.2.5 Our Notation

Table 4.2: Semantic constructs for provenance in the sport domain

Semantic Construct (W3C PROV)	Description (in context of Sport)	Specialised Semantic Construct (for Sport)	ID
 Entity	Entities can be either digital data, or physical concepts such as the state of having possession of the ball.	Video feed Game state Game event Metric	1 2 3 4
 Activity	A process, whether manual or automated.	Physical action Manual process Computation De-identification	5 6 7 8
 Agent	The person or device involved in performing an activity.	Analyst / System Player / Role Team Sensor	9 10 11 12
 Association	While data provenance deals with data dependency, physical provenance deals with causality.	Data dependency Physical causality	13 14
Bundle	Set of entities, activities, agents, and/or associations.	Group	15

This section introduces our specialised notation for provenance in the sport domain based upon W3C PROV. The W3C PROV standard provides the high level semantic constructs such as Entity, Activity, Agent, Association and Bundle and rules for validity. Our notation extends

this through introducing new specialised constructs for sport, as listed in Table 4.2.

Graphical symbols for the specialised constructs are proposed in Appendix Sec. B.1 to facilitate their use together as a visual notation. In contrast to W3C PROV that targets formal representation in languages such as XML and RDF, our focus is on providing a complete visual notation to facilitate use of the notation by those without a computer science background.¹⁸

Physical Provenance

This sub-section considers the suitability of the W3C PROV specification as a tool to model in-game sport events. Specifically it focuses on modelling the physical provenance of the ball (i.e. the game states that it transitions through). This is achieved by mapping concepts in the sport domain to concepts in the W3C provenance standard: game states (i.e. position on the field and state of possession) as PROV *entities*; actions that transform the game state (e.g. kicks) as PROV *activities*; and players that perform the actions as PROV *agents*. To support reasoning about the game in terms of either specific players (e.g. Cyril Rioli) or the roles they represent (e.g. Half Forward), the PROV *actedOnBehalfOf* relation is used to describe a many:many relationship between players and roles. This allows our model to handle role changes (e.g. a substitution of player roles due to an injury).

While this mapping is sufficient for formalisation purposes, it is also important to consider the usability of such a system by a sport performance analyst. Specifically, the abstract concepts of entities, activities and agents are unlikely to be familiar to users in the sport domain, and thus breaks the usability heuristic that software should “speak the user’s language” [153]. As such, this sub-section contains a pro-

¹⁸In principle, our notation could be formally integrated with W3C PROV by defining a custom sport ontology using the Web Ontology Language that specialises W3C PROV constructs <https://www.w3.org/TR/prov-o/>. However, for the purposes of this thesis, the focus is on the constructs and visual notation rather than the techniques that exist to formalise these.

posal for specialising the notation of PROV with custom symbols for game events in order to translate it into the language of sport.

Fig. 4.6 provides an example of how the provenance of the goal described in the Motivating Scenario could be modelled. The figure shows that the goal resulted from a kick performed by Player 7, who possessed the ball as a result of a kick by Player 12, who in turn possessed the ball as a result of a tap by Player 3 from the centre bounce which served as the origin of the possession chain.

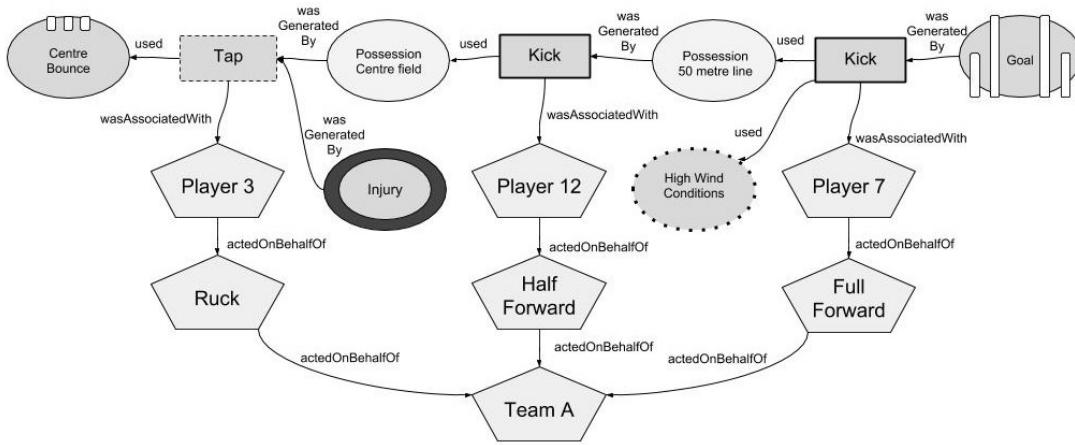


Figure 4.6: Example use of our notation to describe the physical provenance of a goal. See Appendix Sec. B.1 for meaning of symbols.

Due to the tendency of sport to focus on the single point of the ball, the provenance information tends to take the form of a sequential chain. Nevertheless, the example still includes some branching, such as injuries generated by game events that may be handled while the rest of the game progresses, and external events such as wind conditions that occasionally interact with the game through influencing the outcome of a kick.

By annotating the game in such a manner, it becomes possible to express queries about game events in the same manner as one would query a more conventional data provenance graph. For example, the performance analyst may be interested in how a goal came to be, specifically examining goal assists. Without provenance, the performance analyst would have to either rewatch the raw video for the game or read the match feed and filter out irrelevant information. With provenance,

they can query the provenance graph for influences on the creation of the goal, supplementing their query to filter to certain node types or depth limits (in this case, filtering to chains involving agents separated by 2 activities). An example of the result one might receive is shown in Fig. 4.7.

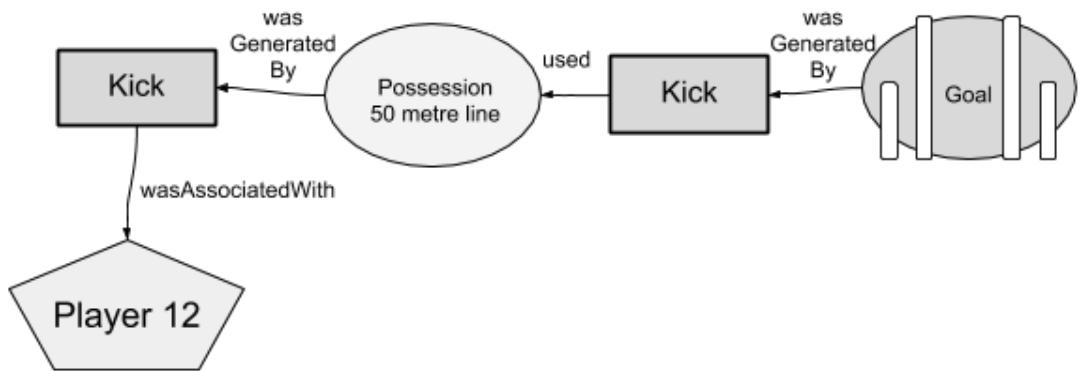


Figure 4.7: Example provenance query answer returned to the user in our notation.

Workflow Provenance

The previous sub-section showed how the W3C PROV specification could be translated into the sport domain to model physical provenance. However, as most sport games focus on a single linear sequence of events, representing the physical aspects of the game as a provenance graph is, by itself, of limited benefit when compared to a traditional linear timeline. The true value of this approach comes when provenance can be traced throughout the entire system to link game events with player metrics.

This sub-section considers the use of the W3C PROV specification to describe the derivation of digital data, such as metrics, computed as part of a workflow. As this task is more abstract, the concepts at this level are not clearly sport specific, especially when compared to our physical provenance model for sport. Nevertheless, note that the functional and quality requirements of the sport domain have implications on the selection of an appropriate workflow representation.

Video analysis is one of the primary tools that sport performance analysts use to analyse the game and communicate results to players and coaches. This is evidenced by the popularity of video timeline based annotation tools such as Sportscode amongst elite sport teams. As such, our representation introduces a custom symbol for video data. Ideally, if our representation was used as part of an interactive tool, it would allow the user to directly play back video segments when they form part of the provenance graph, without the need to open the video in an external program and scan to the time of events.

Sport analysis workflows requires a combination of automated processing (e.g. metric calculation) and manual processing (e.g. video annotation). The W3C PROV standard does not make any distinction between manual versus automated processes, so in theory can model both. However, in practice, due to its generality, capturing automated processes fully such that they could be recomputed requires extending the standard to specify these details, such as to capture the source code and software environment involved.

Unlike the physical sciences, sport science involves working with human participants (i.e. sport players). As such, there is often a need to de-identify data for privacy reasons, for example, if a sport club decides to share player data with researchers outside the club. This has implications on the provenance capture system, as it means that different users need access to different parts of the provenance graph (e.g. the researcher should have an incomplete graph that prevents them tracing provenance of the player data back past the de-identify operation, while the sport club should be able to reconstruct the entire provenance graph once the researcher shares their final findings and provenance data).

Fig. 4.8, presents an example of our proposed notation to capture the provenance of a computation of player goal accuracy.

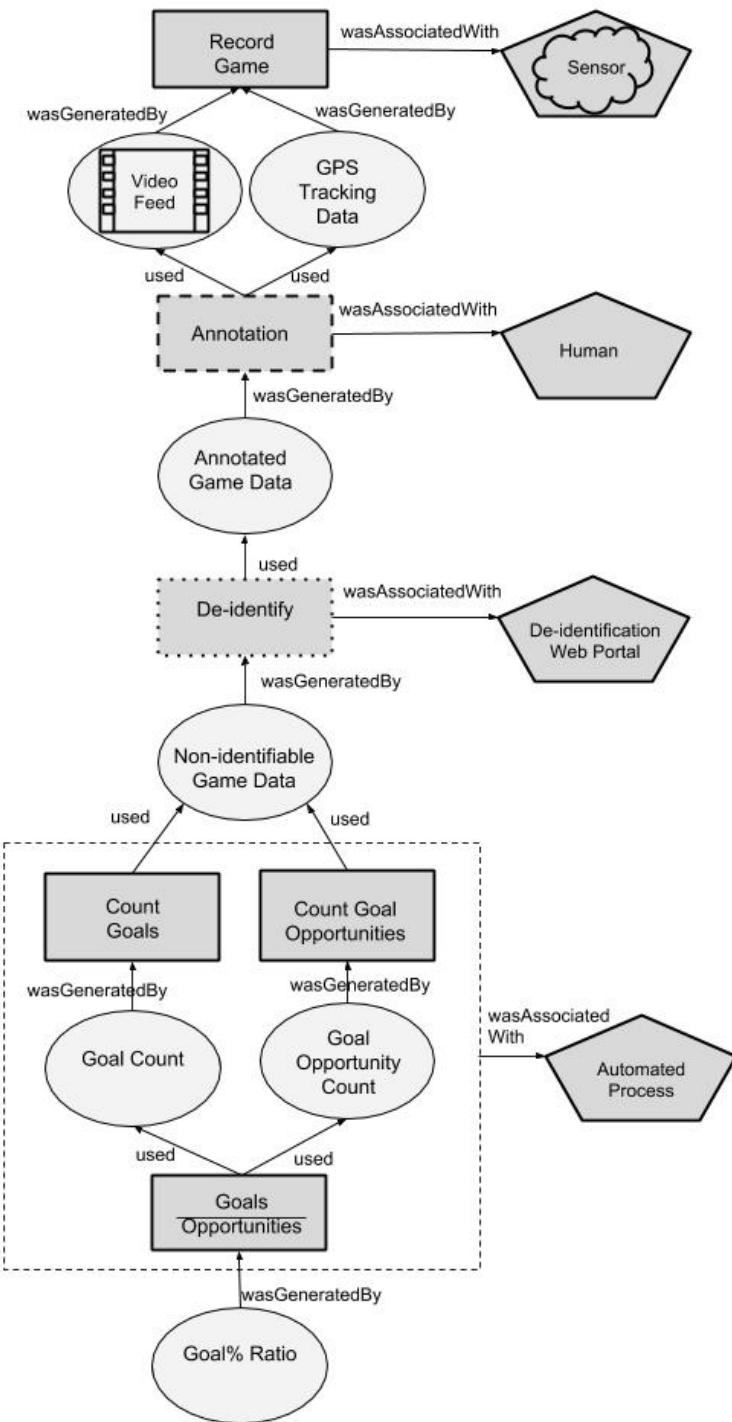


Figure 4.8: Example use of our notation to describe the data provenance of the Goal% Ratio metric

Combined Provenance

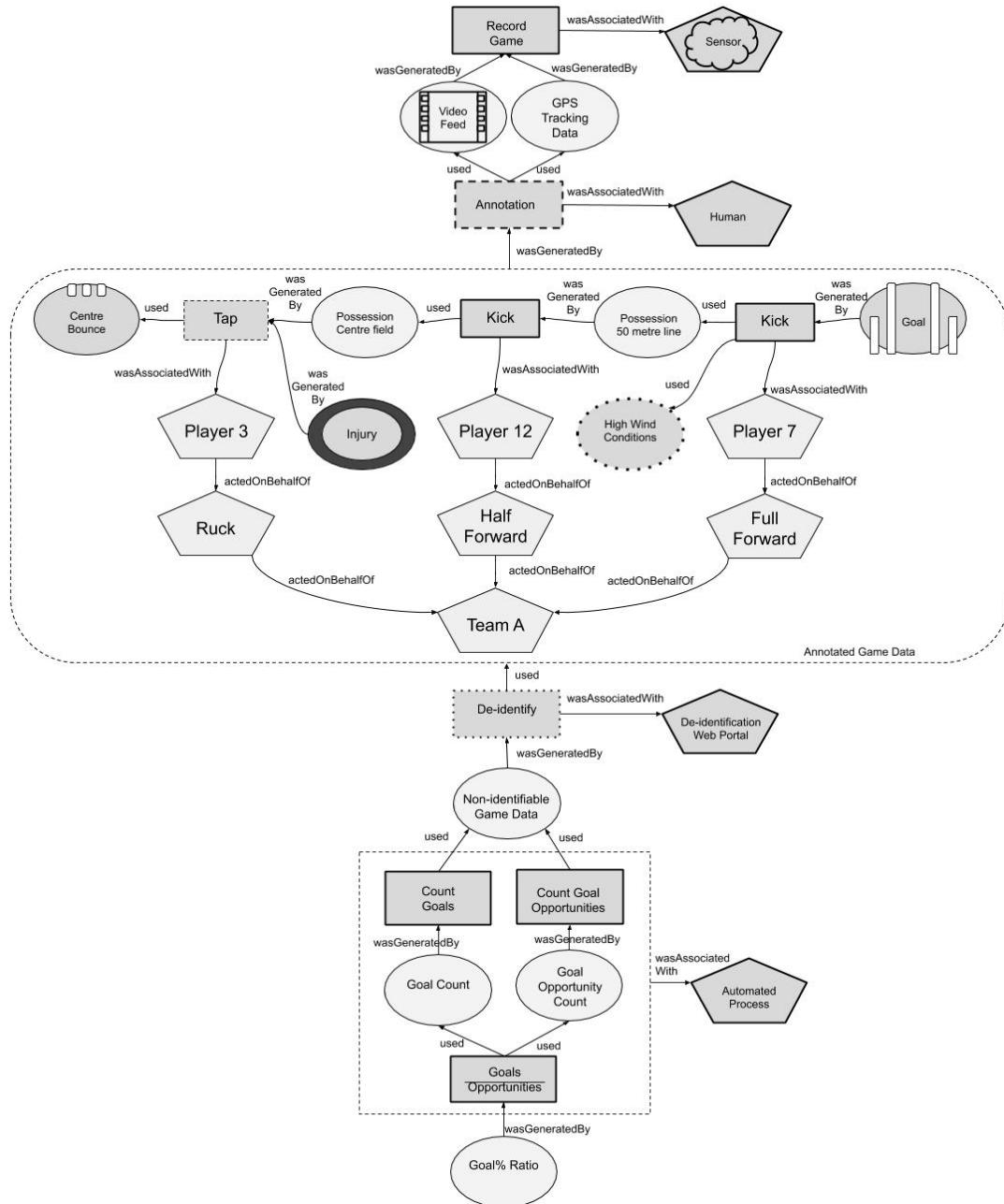


Figure 4.9: Example use of our notation to describe physical and data provenance together as part of same provenance graph

The previous sub-sections suggested a notation for physical provenance to describe game events and a separate notation for workflow provenance to describe metrics and computations. Fig. 4.9 shows that the annotated game dataset that forms part of the workflow can be decom-

posed into the underlying game events it represents, and thus physical provenance and workflow provenance can be integrated as part of a single provenance graph.

Combining our customised notation for workflow and physical provenance graphs ensures that all aspects of the provenance system will be expressed using concepts the user can interpret. For example, consider that a sport performance analyst performs a query to trace the provenance of a metric back to the game events that contributed to it. While the query references a metric (e.g. Goal% Ratio) that is defined at the workflow level, the resulting answer needs to be in terms of game events, which can be communicated in the language of sport practitioners by using the same physical provenance notation used to express the physical query response in Fig. 4.7. This prevents the user from being exposed to the underlying system encoding of the game data (as would be the case if they exported the game events using an arbitrary format determined by their video annotation software), thus increasing the overall usability of the system through consistency and familiarity of the representation.

While the broad semantic constructs such as Entities, Activities, Agents, and Associations already exist in the W3C PROV standard, specialised semantic constructs (along with syntactic representations) are required to meet the needs of the sport domain. Table 4.2 provides an overview of the key specialisations required.

4.2.6 Comparative Evaluation

This sub-section compares the existing W3C PROV standard and the VisTrails workflow management system against the specification for our proposed system, within the context of the sport domain. Specifically, their functionality is evaluated according to the tasks outlined in the motivating scenario (Sec. 4.2.3), the effectiveness of their visualisation against design principles described by the Physics of Notations framework [144], and their usability against Nielsen's heuristics for user in-

terface design [153].

Functionality

The evaluation will begin by modelling the workflow provenance scenario described in the Motivating Scenario (Sec. 4.2.3) using each system. This provides a comparison of the functionality of existing modelling languages and allows identification of gaps in the context of the sport domain.

The W3C PROV standard includes semantic constructs for modelling entities (e.g. a dataset), activities (e.g. a process) and agents (e.g. people that perform the process). It also includes the concept of a plan to describe how a process was carried out, but the details of how to execute a plan is left open, so cannot fully capture the details of an automated process without introducing additional semantics. Fig. 4.10 uses the W3C PROV standard to describe the computation of a player evaluation metric.

VisTrails models workflows as a directed graph of automated processing elements (usually visually represented as rectangular boxes). Each processing element has “ports” that represent the inputs (top of box) and outputs (bottom of box) to/from the process. The user drags connections between output ports and input ports to wire up the workflow. Ports contain type information, which the interface uses to prevent the user from accidentally connecting ports with conflicting types. The resultant workflow is fully automated and reproducible, however is not able to model processes that require human input, other than at the level of tracking manual changes to the workflow itself. Fig. 4.11 shows an implementation of the metric computation pipeline within VisTrails.

The following sub-sections evaluate these systems against the requirements set out in the Motivating Scenario (4.1).

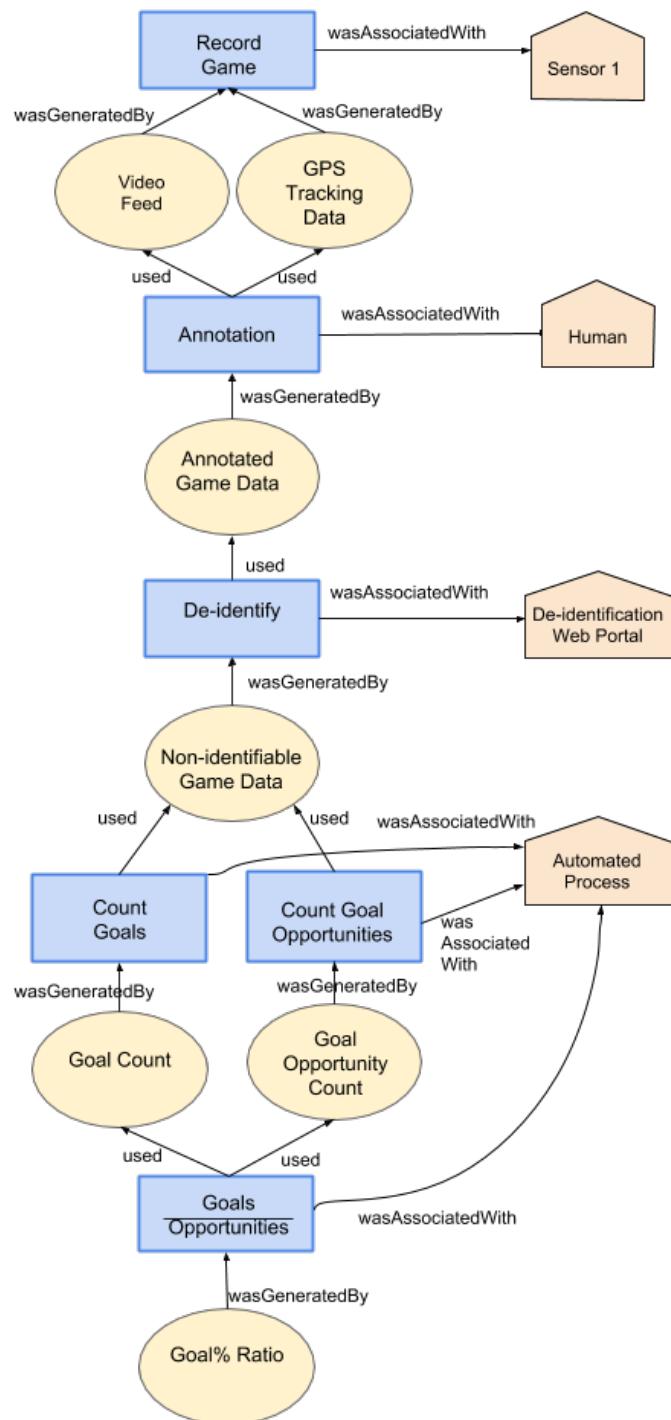


Figure 4.10: Description of how Goal% Ratio was determined, expressed using the W3C PROV standard. Note that the W3C PROV standard only captures the activities and datasets at a high level and captures neither the details of the dataset nor code necessary to reproduce the process.

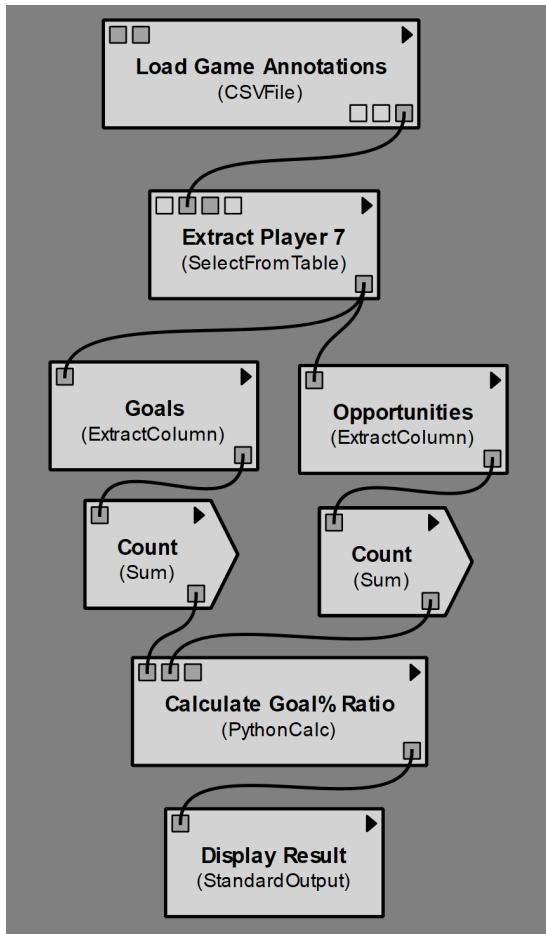


Figure 4.11: Construction of pipeline to determine Goal% Ratio of player using the Vistrails workflow system. Note that it is not possible to describe the manual annotation processes in VisTrails, so this has to be performed using an external system then loaded as the first step of the pipeline.

Integrated support for working with video data: The W3C PROV standard does not provide a means to directly represent datasets other than as plain text using the *prov:value* property or as a resource with an associated URI. However, it integrates with semantic web technologies such as the Resource Description Framework (RDF) which could, in theory, be used to model and describe a video source. VisTrails contains predefined modules for working with tabular data, but does not provide inbuilt modules for working with video data. One could implement custom modules for loading video data and visualising the final output as video. However, without architectural changes to the source code, the system does not have the flexibility to support interactive editing or dis-

play of video sources as they flow through the processing pipeline.

Support for automated processes: The W3C PROV standard includes the concept of a plan to describe how an activity was conducted, but does not capture details such as the source code or software environment that would be needed to reproduce the process. In contrast, VisTrails is a workflow automation tool designed to ensure reproducibility (although this reproducibility may still be undermined by missing data or dependencies on broken web services) and provides a selection of built-in processing modules as well as allowing user-defined Python scripts to cater to situations where the built-in modules are insufficient for a particular task. VisTrails supports export to the W3C PROV standard, but achieves this through mixing in resources within the VisTrails namespace so that it can represent the concepts missing from the PROV standard, as shown in the sample displayed in Listing 4.1.

```
1 <prov:document version="1.0.4"
2   xmlns:dcterms="http://purl.org/dc/terms/"
3   xmlns:prov="http://www.w3.org/ns/prov#"
4   xmlns:vt="http://www.vistrails.org/registry.xsd">
5   <prov:entity prov:id="e15">
6     <prov:type>vt:data</prov:type>
7     <prov:label>str_expr</prov:label>
8     <prov:value>(player,==,7)</prov:value>
9     <vt:id>15</vt:id>
10    <vt:type>(org.vistrails.vistrails.basic:String,
11              org.vistrails.vistrails.basic:String,
12              org.vistrails.vistrails.basic:String)
13    </vt:type>
14    <vt:desc>(None,None,None)</vt:desc>
15  </prov:entity>
16  ...
17 </prov:document>
```

Listing 4.1: Sample of PROV export generated by VisTrails. Note that it mixes resources in the Vistrails “vt:” namespace with the W3C “prov:” namespace to make capturing the workflow possible.

Support for manual interaction: Because the W3C PROV standard does not distinguish between manual and automated processes, and only models details of activities at a high level, it is well suited to describing manual processes and the agents (people) involved. VisTrails provides a way for users to explore the parameter space and to interactively view the output of the workflow; however, it does not provide a way to capture manual processes as steps of the workflow, other than by capturing the history of changes to the structure of the workflow itself. Other workflow systems such as Taverna support interactive processes as components of the workflow that either run locally and interact with the user, or run through a web interface.¹⁹ However these are limited to self-contained sequential tasks rather than iterative ad-hoc tasks that require interaction with the rest of the pipeline.

Partial / shared workflow graphs: The W3C PROV standard was designed for sharing of provenance information on the web. References to resources that make up the provenance graph are represented as URIs, and thus information referenced by the provenance graph could potentially be restricted by controlling access to the resources referred to. As a concrete example, part of the provenance graph could include a URI referencing a document that contains the mapping of player identifiers to anonymised codes, but the document the URI refers to could be hosted on the sport club's intranet and require a password to gain access. Social platforms for scientific data sharing have proposed sharing data alongside workflow information, such as MyExperiment [55] for sharing Taverna workflows, and CrowdLabs [134] for sharing VisTrails workflows. However, a study of Taverna workflows shared on myExperiment found that “nearly 80% of the tested workflows failed to be either executed or produce the same results” [231]. This suggests that even when analyses are automated, practical issues still exist sharing and archiving workflows in a manner that results can be replicated, particularly in cases where certain data cannot be shared for confidentiality reasons. VisTrails contains in-built support for workflow “diff” and “merge”, as well as “visualisation by analogy” which automatically translates changes applied to one workflow to another workflow. These

¹⁹<https://taverna.incubator.apache.org/documentation/interaction/>
Accessed: 2019-05-29

features could potentially ease collaboration on shared workflows.

Provenance / Reverse Debugging: There are multiple types of provenance information. “Workflow provenance” tracks the processes applied to datasets, but usually does not allow inspection of these processes, whereas “data provenance” is fine-grained provenance that tracks how individual data items are derived from each other [206]. Data provenance is further split into “why” provenance [48] which captures all data records that contribute to a result, “where” provenance [27] which deals with only the parts of records that are copied into a result, and “dependency” provenance [36][37] which is similar to why provenance, but formalises the notion of what it means for part of a data record to contribute to a result.

While VisTrails’ provenance browser by default only shows coarse-grained workflow provenance information pertaining to when each component of the workflow was executed, the user can roll back to any version of the workflow, modify the components of interest to output additional debugging information such as inputs and outputs, then re-run the workflow using cached results where available. The W3C PROV standard only deals with modelling and representing provenance, not how to capture provenance. The level of granularity expressed is the choice of the person or process that generates the provenance.

Streaming data: The W3C PROV standard can be used to describe provenance in situations involving real-time streams of sensor data by using the standard to describe the provenance of each individual sensor observation [43]. The VisTrails user manual includes a section “streaming in VisTrails” that describes how functions can incrementally process data. This could potentially be utilised to process a stream of sensor data; however, the stream would need to terminate eventually for the workflow execution to complete successfully.

Table 4.3 summarises the above findings.

Table 4.3: Evaluation of functionality against the tasks outlined in Motivating Scenario

Requirement	W3C PROV ²⁰	VisTrails
Integrated support for working with video data	No ²¹	No ²¹
Support for automated processes	No ²¹	Yes
Support for manual interaction	Yes	Partial
Partial / shared workflow graphs	Yes	Partial
Provenance / Reverse Debugging	Yes	Partial
Streaming data	Yes	Partial

Effectiveness of visual notation

Appendix Table B.6 summarises the findings of the effectiveness of the visual notation used by each system. As the W3C PROV standard provides textual serialisations such as XML, but does not formally specify a visual notation, the evaluation assess the notation in the (non-normative) visualisations the standard uses to document examples.

Heuristic Usability Evaluation

Appendix Table B.7 summarises the usability issues identified in VisTrails as a result of a heuristic evaluation. The usability of the W3C

²⁰For W3C PROV, the evaluation examines the ability to model provenance information; however, an external system would be needed to actually capture the provenance information and explore it.

²¹Partial support may be possible via extending the language with additional modules / semantics.

PROV standard was not evaluated, as it does not specify any particular implementation to create provenance documents.

4.2.7 Key Findings

1. Automated workflow tools often lack support for capturing ad-hoc manual processes that cannot be automated. Conversely, provenance standards such as W3C PROV recognise the need to document the inputs and procedures involved in ad-hoc manual processes, but lack semantics for describing the code and execution environment necessary to reproduce automated parts of the analysis. Supporting the needs of the sport domain—and other fields where manual and automated analysis are intertwined—requires combining these as part of a unified standard to ensure a complete and reproducible capture of the analysis.
2. As automated workflow tools treat processes as black boxes with limited traceability, their provenance logs typically only show basic execution information such as the time the process ran and status of the result. However, analysts in the sport domain require fine-grained data provenance to trace results back to raw events. Although the black box nature of workflows prevents support of “why” provenance and “where” provenance methods designed for analysing provenance of SQL query results, note that workflows implicitly support a form of retrospective investigation through the ability to roll back history and recompute key processes with additional logging information or with modified data inputs to observe the effects on the output. In cases where capturing fine-grained provenance is not possible, workflow systems could support the user to retrospectively reason about the likely provenance of data by guiding the user through the procedure of retrospectively collecting intermediate states and manipulating inputs to infer which data values had an impact on the result of the process. This approach could also be used to support user reasoning about provenance in workflows that involve complex probabilistic processes (such as neural networks) by supporting the user with the tools to

rewind the process and “prod” at intermediate data to understand what is most relevant (i.e. sensitivity analysis) and whether expected properties hold (i.e. metamorphic testing) rather than overloading the user with information about the computations carried out.

3. The analysis of existing provenance notations shows poor utilisation of the available design space. Notably the “graphic economy” of the systems studied could be improved by utilising additional visual variables such as texture to further distinguish symbols. As certain domains demand a different set of semantic constructs to others (e.g. the reliance on video annotation within the sport domain), there is a need to optimise the visual notation for the domain. Translating abstract provenance concepts into concrete concepts in the language of the domain would reduce the number of usability issues faced by practitioners.

4.2.8 Conclusions

While general purpose workflow managers and provenance notations exist, this section of the thesis demonstrated that these systems need extensions and specialisations respectively in order to express the sport domain. Our proposed notation demonstrates what such a language could look like in the sport domain.

Future work is needed to evaluate how potential users respond to our proposed notation. A study by Bachour et al. in which a computer game presented gamers with a visualisation inspired by the W3C PROV standard suggests that non-expert users may be confused by the direction of the arrows, as they are intuitively interpreted as data flow rather than data dependency [12]. An empirical evaluation is needed to detect whether similar issues also exist in the sport domain.

The usability issues arising from the use of general provenance systems in the context of a domain with specialised needs and terminology could be hindering the uptake of provenance systems despite the

widely recognised need for reproducible research. While this section of the thesis has explored issues from the perspective of the sport domain, it is possible that other scientific subfields could also benefit through the introduction of customised provenance languages for their scientific domain. Thus another avenue for future work is to generalise the methodology presented in this section of thesis in order to generate a family of provenance systems, each optimised for a particular scientific domain.

4.3 Chapter Summary

This chapter outlined the value of computational pipelines as a means to automate data processing in sport. In particular, it highlighted that data provenance, the ability to trace the results of an analysis back to the original input, is still an open research area.

While computational pipelines are well established in scientific fields such as bioinformatics, this thesis brings the benefits of computational pipelines to sport analysis. To assist with this process, this chapter proposed a specialised notation (Sec. 4.2.5) for tracking data provenance in the sport domain based on W3C PROV.

While designing a software tool that uses this notation to help annotate data and automate workflows is a future possibility, even as-is the notation can be used to help document workflows. The notation is used in Chapter 7 to describe the pipeline developed in this thesis.

Future Work

Future work is needed to formally interview sport performance analysts to verify whether the motivating scenario is reflective of their real-world experience, and to develop tool support for the provenance notation so that users can more easily utilise the notation to document workflows. The notation has been utilised in this thesis (Chapter 7), but trials are needed with sport performance analysts and sport researchers to evaluate the usability of the proposed notation and to determine whether it fulfils their needs.

Contributions

1. Provided an analysis of the data provenance needs of the sport domain, and evaluated existing data provenance tools against these criteria. A customised data provenance notation for sport was proposed in order to ease uptake for sport performance analysts without a computer science background.

Chapter 5

De-identification

Contents

5.1 Introduction	126
5.1.1 Large scale privacy breaches due to improper de- identification	129
5.2 Background	131
5.2.1 Formal methods for de-identification	131
5.2.2 De-identification in practice	140
5.2.3 Summary of gaps in literature	141
5.3 Prevalence of Improper De-identification Methods: A Review of “Non-identifiable” Datasets used in Aus- tralian Rules Football Research	143
5.3.1 Abstract	143
5.3.2 Introduction	144
5.3.3 Method	146
5.3.4 Results	150
5.3.5 Discussion	155
5.4 Threat Model	157
5.5 Re-identification of Sport Data	160
5.5.1 Re-identification of possession chain data	160
5.5.2 Re-identification of GPS data	162

5.6 Analysis of Trade-Off between Participant Privacy and Data Quality	167
5.7 An Interaction Model for De-identification of Human Data held by External Custodians	170
5.7.1 Abstract	171
5.7.2 Introduction	171
5.7.3 Emotional goal framework	173
5.7.4 Modelling	173
5.7.5 Implementation	177
5.7.6 Heuristic usability evaluation	179
5.7.7 Case study	179
5.7.8 Conclusions	180
5.8 Chapter Summary	181

The background chapter outlined the shift from traditional sport statistics that summarise game events towards detailed spatio-temporal datasets that track the position of every player on the team, at each moment, in high fidelity.

The increase in the volume and detail of data creates an opportunity for deeper forms of analysis than traditional approaches. However, the level of detail also raises new concerns about the risk of harm to player privacy, particularly when these datasets are shared with external parties outside of clubs, resulting in the potential for the dataset to be used in ways that are contrary to the wishes of players.

Typically, this is solved by the data custodian (the football club) de-identifying player tracking data (e.g. substituting player names with randomised anonymous codes) prior to handing the dataset over for use in research. However, the level of detail involved in player tracking datasets means that even after known identifiers are removed, there is often sufficient auxiliary information left in the dataset to re-identify players.

If data is collected for club use only, and players consent to the data collection, then de-identification may not be necessary when the analysis is all conducted in-house by trusted sport performance analysts working at the club. In contrast, this chapter deals with the situation where clubs share the data with an external party, such as a sport researcher outside of the club. In such a case, players may be unwilling to share their data in identifiable form, or the data may relate to former players who consented to sharing identifiable data within the club, but cannot be practically contacted anymore to seek approval to reuse their data in new ways. In this situation, it is vital that the data be properly de-identified to protect the players' privacy rights.

This chapter explores the issues surrounding de-identification of detailed sport datasets, in particular GPS player tracking data and associated datasets. As noted in the previous chapter introducing computational pipelines, de-identification is a form of data transformation operation that can be considered as part of the overall pipeline.

The de-identification operation has special considerations that distinguish it from other components of the pipeline. Unlike other operations where it is desirable to have a means to trace the provenance of results back to records in the underlying dataset, the aim of data de-identification is to ensure that results *cannot* be traced back to the individual the data relate to. As such, the de-identification operation must be under the control of the data custodian, as only the data custodian should have access to the original underlying identifiable data. However, sport researchers external to the club need to be able to design and run operations that form the rest of the computational pipeline acting upon the de-identified data. This motivates the design of a new interaction model for de-identification of human data.

This chapter is directly applicable to sport research involving de-identification of spatio-temporal data. The interaction model presented is also applicable to other fields where researchers require access to detailed human data in non-identifiable form but the data custodian has limited technical resources to carry out the de-identification process. The findings of this chapter informed the design of the de-identification process used to obtain the non-identifiable datasets used throughout this thesis.

5.1 Introduction

A machine learning approach to sport analysis requires access to detailed data about each action in the match, as opposed to traditional sport statistics that deal only with summary statistics about the match. In the statistics literature, detailed datasets containing individual rows for each participant (i.e. the players) are known as *microdata*.

However, microdata introduces legal and ethical concerns:

1. Although sport matches are played in public, there are concerns from players and player representation bodies about the unprecedented level of individual player scrutiny that could occur as a

result of in-match player tracking.¹

2. Human research ethics codes encourage researchers to de-identify data where possible. Ethics exemption may be granted for use of pre-existing non-identifiable data, whereas cases where data are identifiable or re-identifiable (and not already public) require participant consent and/or justification that the benefits of the research outweigh the risk of harm to participant privacy.

De-identification is a one-off transformation that should be applied to the data as early as possible, preferably by the data custodian prior to release of data. While stripping identifying information or replacing it with random anonymous identifiers may at first seem trivial, this chapter shows that de-identifying data in a way that preserves privacy against attack is non-trivial, and the proper choice of procedure should be informed by the nuances of the domain. In particular, an ideal state is to prevent the ability to re-identify individual players based on player movement or scoring profiles, while preserving the ability to perform high-level/aggregate analysis (e.g. spatio-temporal analysis of team formations).

As the aim of sport research is to capture generalisable patterns rather than critique individuals, and considering the legal and ethical concerns surrounding data that could identify individuals, the data should undergo a de-identification process as early as possible. However, the de-identification of high-dimensional microdata is challenging to perform in a way that is secure against all possible re-identification attacks. Furthermore, these issues are exasperated for spatio-temporal data (discussed in Sec. 5.2.1 and Sec. 5.5.2).

The work presented in this chapter was motivated and shaped by a series of interactions that showed systematic gaps between the theoretical privacy rights of participants in research and the state of sport research in practice. The project involved interaction with sport perfor-

¹The Age, “AFLPA raises concerns about plan for broadcasters to show AFL player GPS data”, 9 Feb 2017. Available: <https://www.theage.com.au/sport/afl/aflpa-raises-concerns-about-plan-for-broadcasters-to-show-afl-player-gps-data-20170209-gu9jtr.html>

mance analysts at two elite sport organisations. The first club provided identifiable GPS data, without any discussion regarding player consent, nor data confidentiality beyond goodwill.² Communicating with a sport practitioner at the other club, they provided a sample of identifiable possession chain data for one match.³ While greatly appreciative of the trust of both of these clubs, particularly the collegiality of individuals at the clubs without whom this research project would not have been possible, these actions are suggestive of the current attitudes towards data sharing by sport practitioners; i.e. a culture of data sharing based on goodwill and relationships between researchers and sport practitioners, rather than on the consent of the individuals to whom the data pertains.

Furthermore, there were projects in university research centres that used de-identification schemes that were fundamentally flawed, such as anonymous identifiers generated through a deterministic rather than random process, or included a large number of auxiliary attributes that could be linked to identifiable information. These projects often stated data as being “non-identifiable” in ethics applications and were approved without further question. These personal observations triggered a deeper investigation to confirm whether these experiences are part of a more general phenomenon. The presentation of this investigation aims to bring the advances in understanding of privacy that have occurred in the computer science domain to the sport research domain.

This chapter:

1. Reviews the literature on formal methods and best practices for de-identification of research data sets.
2. Studies a sample of sport research papers to establish common de-identification issues faced by sport researchers and practitioners.
3. Considers re-identification risks for player position tracking data, with the goal of identifying a de-identification method that pre-

²The analysis presented in Chapter 7 does not use data from this club.

³This sample of possession chain data was not used as part of analysis in Chapter 7 either.

serves player privacy without destroying the utility of the data for team-level spatio-temporal analysis.

4. Provides an interaction model for requesting datasets with special de-identification requirements from external data custodians without revealing the underlying raw data.
5. Applies this work through a case study in de-identification of AFL player tracking datasets.

5.1.1 Large scale privacy breaches due to improper de-identification

The below list shows examples of large scale privacy breaches of private and government released microdata due to improper de-identification methodologies. Note that none of these were due to conventional cyber-attacks, but rather due to deliberately released datasets without a proper understanding of the risks of re-identification. These have been reported in the literature [163, 69] and media:

1. 1997: Massachusetts Group Insurance Commission (GIC) releases de-identified health insurance data. Sweeny [203] later shows that the health records can be linked to the voter registration list on ZIP code (postcode), sex, and date of birth to re-identify individuals, using the governor of Massachusetts at the time as an example.
2. 2006: AOL releases search data, substituting user identifiers with anonymous codes. However, as searches conducted by the same user shared the same anonymous code, it was possible to use keywords in their search terms to build up a profile on their likely location and identity, eventually culminating in a confirmed re-identification.⁴

⁴The New York Times, “A Face Is Exposed for AOL Searcher No. 4417749” <https://www.nytimes.com/2006/08/09/technology/09aol.html>, 9 Aug 2006

3. 2006: Netflix releases de-identified user movie ratings, challenging data scientists to come up with a better movie recommendation algorithm. The dataset is later taken down after researchers reveal that knowledge of three movie ratings for a user (e.g. obtained from external movie review sites or by asking in person), and the approximate dates the movies were rated, is enough to uniquely identify the Netflix user with 80% probability. This results in revealing viewing history of other movies that the user did not intend to be publicly associated with [149].
4. 2016: Australian Department of Health releases de-identified medical billing records to `data.gov.au`, the Australian Open Data portal. The dataset is later taken down after researchers discover that the algorithm to generate unique anonymous identifiers is fundamentally flawed due to the use of a deterministic rather than random generation procedure. Knowledge of a few medical billing records (such as dates of baby delivery, or publicly reported medical conditions) is sufficient to re-identify individuals in the dataset and thus learn their full medical billing history which may reveal private information of a more sensitive nature [49].

In his seminal 2010 article, “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization” [163], Paul Ohm, a professor of law, examines the failure of de-identification methods to meet human expectations of privacy, as well as the theoretical limits on achieving privacy without destruction of data utility. He concludes that the “robust anonymization assumption” that data can be fully de-identified without damaging the utility of the dataset is false. He explains that this has been a “fundamental misunderstanding” of policy makers who set out privacy law and regulations, which has resulted in either “much less privacy than we have assumed” (in the US), or unattainable standards (e.g. in the EU). He describes the “accretion problem” in which each re-identifiable dataset, even if non-privacy invading in itself, permits linkage of others, and warns against the potential for a “database of ruin”, compromising the privacy of every individual, which comes closer with every information disclosure. The rest of his article speculates on legal reforms needed to defend against issues that will inevitably arise as a

result of the impossibility of having both privacy and free information flow in a world where unprecedented levels of data are already publicly available for linkage.

These incidents demonstrate that de-identification is non-trivial to perform correctly, and failure to anticipate current and future attacks against the de-identification strategy can result in irrevocable damage both to the privacy of the individuals they pertain to, and the reputation of those responsible for the datasets. The following section will explore formal methods and best practices to protect against re-identification.

5.2 Background

5.2.1 Formal methods for de-identification

Fig. 5.1 provides an overview of formal de-identification methods for microdata that will be discussed in this section. The section begins by reviewing deterministic algorithms for non-interactive data privacy whereby the data custodian applies an algorithm to remove or coarsen data attributes then hands the data over to the researcher to analyse. It then briefly touches on differential privacy, a mathematical framework for reasoning about the privacy guarantees offered by probabilistic privacy methods that distort the data through noise dependant upon the query made of the data. Finally, it will review literature on de-identification of spatio-temporal datasets. Spatio-temporal datasets tend to resist formal mathematical guarantees of privacy due to the uniqueness of movement patterns.

Deterministic de-identification strategies for microdata

This section will examine measures and algorithms for de-identifying data. Specifically, it looks at deterministic strategies that generalise the data (e.g. by removing fields, reducing precision, or grouping into

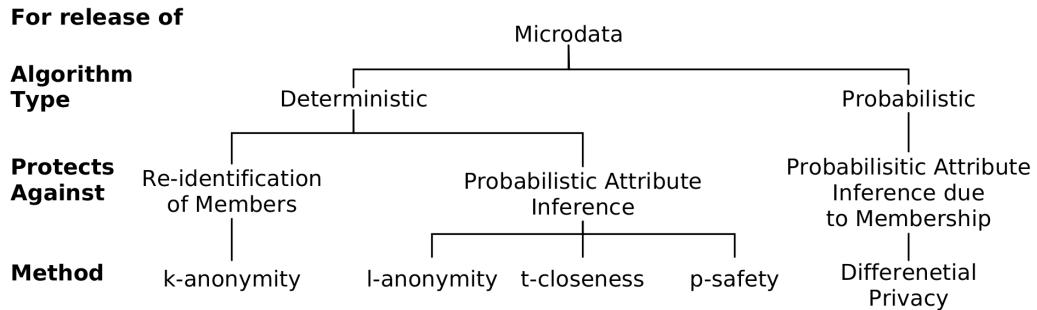


Figure 5.1: Taxonomy of formal de-identification methods

coarser categories) without resorting to random distortion of the data (e.g. adding noise). Early de-identification methods proposed in the literature were later understood to suffer from certain forms of attacks, which motivated design of subsequent methods. To show the interplay between de-identification methods and re-identification attacks in the literature, a misuse case diagram (an adaptation of UML use case diagrams for the purpose of modelling threats [3]) displaying how the methods reviewed in this section build upon each other to correct previous limitations is presented in Fig. 5.2.

k-anonymity Sweeny [203] introduces the measure *k*-anonymity as a tool to formally reason about resistance of a dataset to re-identification attacks based on uniqueness of attributes linkable to public data. Sweeny denotes the set of attributes potentially vulnerable to linkage as the *quasi-identifier*. The level of *k*-anonymity, *k*, is defined as the minimum number of records that share the same *quasi-identifier*. If *k* > 1, then it is provably impossible to re-identify any particular player through record-linkage, as there will be multiple records sharing the same *quasi-identifier* key (although the data may still be vulnerable to other forms of attack). If *k* = 1, Sweeny considers the data to be potentially re-identifiable; however, this may be an overly conservative definition of privacy. Even if *k* = 1 (which will always be the case if the dataset contains continuous attributes), the dataset is only re-identifiable in practice if the attacker can obtain public data pertaining to the attributes of the *quasi-identifier*, for the subjects the data pertains to,

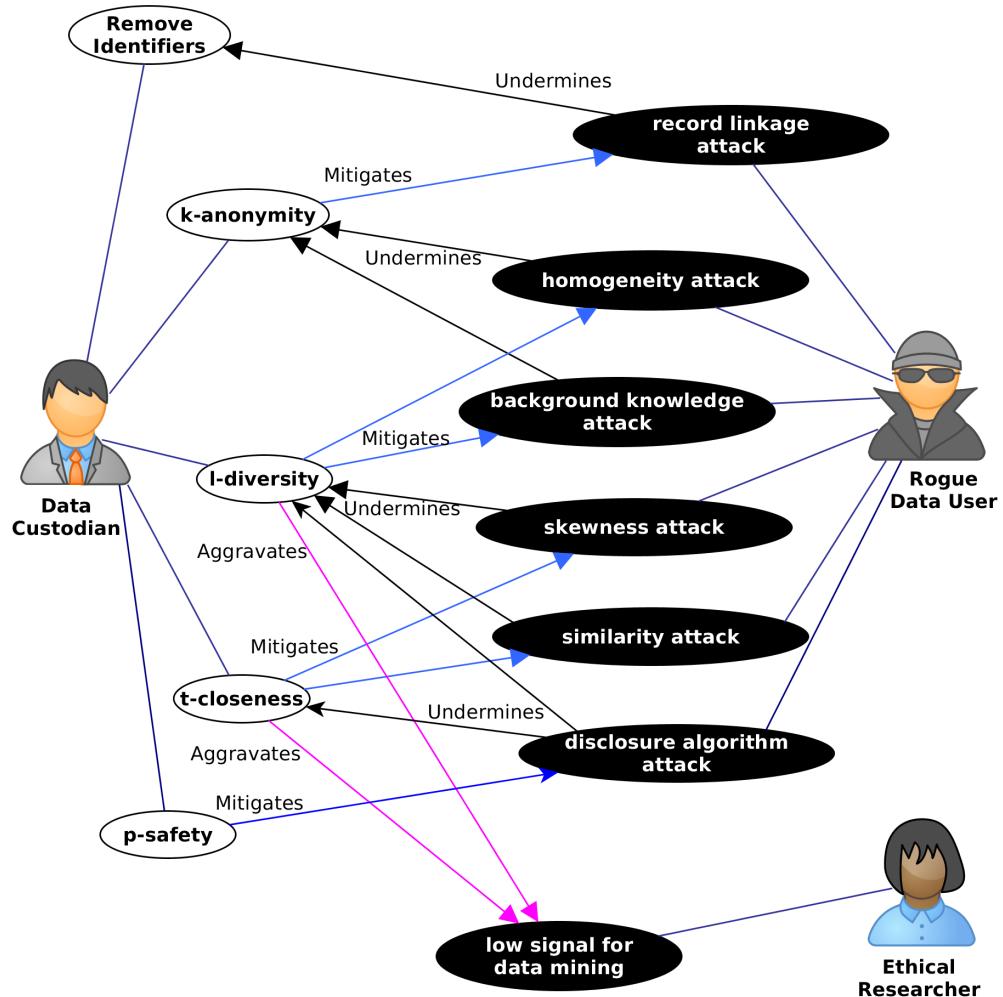


Figure 5.2: Misuse case diagram showing attacks against non-interactive de-identification strategies for microdata

with enough precision to be confident in the linkage.

Sweeny assumes that the data holder is able to determine the *quasi-identifier*, i.e. the attributes vulnerable to data linkage. Sweeny provides the example that a quasi-identifier for medical data would include ZIP (postcode), birthdate, and sex, as these can be linked to voter lists. Sweeny also considers that the quasi-identifier should include other information an attacker may know, such as the patient's race. However, this task of understanding which details an attacker will have access to, is difficult in the current age, due to the micro-level information people publicly share about themselves online through social media and so-

cial networking tools. For example, Sweeny’s example of de-identifying medical data assumes that the visit date of the medical appointment is unknown to the attacker, so does not need to be considered as part of the *quasi-identifier*. However, with services such as Foursquare⁵, and Facebook, that encouraged users to share “check ins” to businesses, it is now feasible for an attacker to collect information such as visit date as a key to re-identify data. This is particularly problematic when de-identifying sport players’ data, due to the pool of public information that could be used for a data linkage attack as a result of the public nature of both elite (via traditional media), and sub-elite (via social media) players’ lives. For the case of sport match data de-identification, the presence of only a limited, known set of players on the team, and a large existing pool of identifiable data on each player, means that k -anonymity is unlikely to be achievable.

l -diversity Machanavajjhala et al. [130] describe two attacks on k -anonymity: the *Homogeneity Attack*, and the *Background Knowledge Attack*. The k -anonymity criterion only specifies the crowd group size that an individual may hide in, and thus anyone in that group is indistinguishable from $k - 1$ individuals, but does not place any restrictions on the diversity of the group in which one hides, thus it may be possible to infer an attribute of an individual in the group on the basis that all members of the group share that attribute. For example, grouping players by team ensures a high value of k equal to the size of team, but if all players on the team anonymously admit to having taken illicit performance enhancing substances, then this still results in revealing sensitive information about individuals. This is known as a *Homogeneity Attack*. The *Background Knowledge Attack* requires that the attacker has some additional information about one of the participants, as is often the case in reality, e.g. they may know that a certain participant on the team *didn’t* take performance enhancing substances, or the performance enhancing substance was not of a particular type. This may allow the attacker to infer which members of the team *did* take performance enhancing substances by ruling out certain possibilities, even

⁵Prior to removal of the Foursquare “check in” feature in 2014

when k -anonymous data is not immediately revealed via a homogeneity attack.

To solve these issues, Machanavajjhala et al. propose l -diversity that extends k -anonymity to further ensure a given level of diversity within each group. There are multiple methods of describing the diversity, so they consider both *Entropy l-diversity* (which had been previously discovered, but without a general framework to motivate it) and *Recursive (c,l)-diversity* (which guarantees diversity by ensuring that the $l - 1$ most likely sensitive values in each group each make up less than c times the number of the remaining values in the group). This means that an attacker would have to rule out at least $l - 1$ other values or records in the group in order to confidently infer a sensitive attribute.

t -closeness Li et al. [155] introduce t -closeness as a response to two attacks against l -diversity: the *skewness attack*, and the *similarity attack*. The *skewness attack* occurs with highly skewed data, where a high l -diversity is simply not possible. Li et al. note that if the distribution for a equivalence class (quasi-identifier group) is similar to the overall distribution, this is unlikely to lead to privacy issues, while if the distribution is flipped (i.e. a large number of uncommon attributes), it reveals information about that group. The *similarity attack* deals with hierarchical sensitive attributes that may appear diverse, but all belong to the same higher level category thus allowing an inference to be made by the attacker. The proposed t -closeness metric attempts to relax the requirements of l -diversity while mitigating the attacks by only requiring that the distribution of the metric in each partition is similar to the overall distribution.

Unfortunately, this limits the utility of the resultant dataset, as finding groups of participants in which attributes differ from other participants is the intent of data mining; however, as Li et al. explain, “this is precisely what one needs to limit” in order to prevent sensitive attribute disclosure. They further note the difference between protecting against disclosure of identity, and disclosure of attributes. k -anonymity protects against the former, while t -closeness protects against the latter.

As t -closeness does not provide any inbuilt protection of identity, the authors of the study suggest that k -anonymity and t -closeness be used together.

p -safety Zhang et al. [230] point out an important practical flaw in previous works that attempt to automate the selection of appropriate granularisation levels to achieve k -anonymity, l -diversity or t -closeness⁶. They describe a deterministic disclosure attack whereby knowledge of the de-identification program (which they call a deterministic disclosure function) may leak information about the attributes based on the side-channel of the final selection of granularity levels (i.e. if a particular granularity level would not work, then it will not be selected, and based on the fact that it was not selected, a user may be able to infer sensitive attributes for particular participants). Zhang et al. show that an algorithm that results in optimal trade-off of utility and privacy would be NP-hard, and describe a heuristic algorithm to achieve anonymity in a way that accounts for leakage due to the deterministic disclosure function.

Trade-off between privacy and utility In *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*, the authors show that utility and privacy are often in direct contradiction, and perform experiments on datasets to demonstrate that “even modest privacy gains require almost complete destruction of the data-mining utility” [26].

Probabilistic de-identification strategies for microdata

Differential Privacy Dwork introduces ϵ -differential privacy [66], offering a refined definition of privacy. Dwork acknowledges the impossibility of releasing results while preventing all possible inferences about an individual by an attacker (as it is always possible that the results

⁶Only k -anonymity and l -diversity were specifically identified in the original paper; however, as t -closeness is used in a similar manner, it is also susceptible to attack

of the analysis could also happen to be the missing key piece of information that an informed attacker needs to infer the identity of an individual); however, it is possible to limit the extent of information revealed about a participant as a direct result of their inclusion in the dataset (as opposed to general information that an attacker could have used to infer the identity of an individual). In Dwork’s definition, the participation of a participant in the study must not alter the probability of a predicate about that person by more than a factor of ϵ (on a logarithmic scale). This allows the researcher to make inferences about sub-groups of the population, while limiting the risk of harm to participant privacy as a result of participating.

Differential privacy forms a mathematical framework for reasoning and designing for privacy, thus can be implemented in different ways, including for use with non-tabular data types. However, it is typically implemented by adding noise to the results such that the differential privacy criterion is met. Differential privacy can provide a form of interactive de-identification, whereby the level of noise added is dynamically adjusted to meet the differential privacy criteria based upon the query being made [67]. Dwork explains that when the query of interest is known, typically only a small amount of noise is necessary; however, if arbitrary queries are allowed then the amount of noise needs to be prohibitively large. Dork explains that in practice “non-interactive mechanism must be tailored to suit certain functions to the exclusion of others” [67].

Unfortunately, the use of noise to ensure differential privacy can lead to misleading results. The avoidance of traditional noise based approaches was the motivation for deterministic de-identification strategies such as k -anonymity through generalisation rather than perturbation. While k -anonymity does not meet the criteria for differential privacy as is, Li et al. [128] show that a modified “safe” version of k -anonymity applied to randomly sampled data meets the conditions for differential privacy. Google have experimented with using differential privacy as a means to prevent deep learning algorithms from revealing information about any particular individual within the training dataset [1].

De-identification of spatial-temporal data

Malin and Aioldi [133] describe a re-identification attack known as *trail re-identification*. When unique, but not identifying (e.g. DNA) data are recorded at multiple locations, there is a risk of re-identification via linkage to data about which locations individuals visited. This is done by making an inference as to whether the known information about an individual's presence at locations is consistent with the measurements taken at those locations. Through simulations, they find that skewed location access patterns usually result in higher chance of de-identification (although under controlled conditions, with contrived parameters, low-skewed distributions are theoretically capable of resulting in more re-identifications).

In a follow up paper, Malin [132] proposes an automated system to achieve “ k -unlinkability” (an extension of k -anonymity designed to measure the threat of trail de-identification) by suppression of records. The paper contains a two-page proof that “Greedy-Dedup [i.e. their system] *output* is k -unlinkable”, but doesn't consider the side-channel of information revealed to an attacker as a result of which records are selected for suppression when running the algorithm (see discussion of p -safety above)

De Montjoye et al. [53] reconfirm the earlier sentiment of Zang & Bolot [228] that location data, even in sparse point form, cannot be anonymised by simply substituting anonymous identifiers. In a study of mobile phone call location data, they find that “four spatio-temporal points are enough to uniquely identify 95% of the individuals”. Furthermore, while coarser granularity improved anonymity, the uniqueness of traces (defined as the percentage of users, ε , who contained unique traces amongst a fixed random sample) was found to reduce according to a power law, thus there are diminishing returns on the number of protected users from increasing granularity. This is problematic for research datasets, as the requirement for non-identifiable datasets is that “no specific *individual* can be identified” [150], i.e. $\varepsilon = 0$. Thus there is a trade-off between utility and anonymity, and the requirement

for strict non-identifiability of all participants may require disproportionate reductions of utility due to the diminishing returns provided by generalisation to coarser spatio-temporal granularities.

5.2.2 De-identification in practice

Privacy checklists O’Keefe et al. [158] propose a checklist for de-identification of health data. Given the trade-offs between utility and privacy, they suggest that the data custodian only apply basic de-identification methods to the raw data, such as removal of personally identifying information, and categorisation of values into ranges. They assume that the researcher is trusted not to deliberately circumvent these measures. The focus of their article is then on creating a privacy checklist that the researcher manually follows to ensure the *output* of their research for publication is non-identifiable. Unfortunately, as their checklist is heuristic driven, failure to comply to the checklist does not necessarily imply a privacy risk, nor does following the checklist offer any formal guarantees that the publication will not be susceptible to a privacy breach.

Their approach demonstrates that complex general statistical guidelines can be made more accessible to practitioners through simplifying them for the specific domain. In the study by O’Keefe et al., the target was health researchers. In contrast, sport research brings about unique challenges, such as a small number of participants on a team, the tendency for participants to all belong to the same team rather than a random sample, the use of spatio-temporal data, and the environment in which detailed auxiliary data is available about players. Thus an adaptation of these guidelines would be needed in order to apply them to sport research.

Privacy in practice O’Keefe et al. applied their privacy checklist to a sample of 100 population health research papers [159]. In the 100 papers, they found a total of 128 outputs (22.6% of all outputs in the papers) that their checklist flagged as potential privacy concerns. However, they note that once the context of the research papers was taken into account, they found “no substantial actual privacy concerns”.

El Emam et al. [69] perform a systematic review of publications reporting re-identification attacks. However, the study encountered is-

sues with publication bias (researchers only reporting cases when a re-identification was successful, and commercial entities who are unlikely to admit to improper de-identification) and difficulty of confirming successful re-identification (while some of the re-identification attempts were confirmed, this is usually only confirmed for a few individuals rather than testing all the re-identifiable individuals). Thus the study was unable to determine the exact extent of de-identification attacks nor the exact percentage of individuals uncovered due to attack, with the study concluding “this evidence is insufficient to draw conclusions about the efficacy of de-identification methods”. Of the 14 studies they identified, only two of the attacked datasets followed standard de-identification practices, of which only one was health related. In the health database, only two out of 15,000 individuals were identified, despite the researchers having access to data from a market research company that they used to facilitate the record linkage attack. This suggests that while there are serious concerns about the prevalence of improper de-identification procedures, identified theoretical issues may be difficult to exploit in practice in a manner that leads to confirmed re-identifications.

5.2.3 Summary of gaps in literature

From the review of the literature, while formal methods exist for data de-identification, these are often over-burdensome in practice, and may destroy utility of the dataset. Despite the existence of health guidelines, these are not always followed in practice, and even when followed, do not offer full protection of privacy. The ability to protect privacy depends upon the proper choice of the *quasi-identifier* of attributes that may be linkable to other datasets. If the person performing the de-identification fails to consider data attributes known to an attacker as part of the analysed quasi-identifier then an attacker may be able to re-identify the data. On the other hand, if they consider too many attributes as part of the quasi-identifier, then the de-identification requirements will be so strict that it results in destruction of data utility.

The condensation of complex guidelines into a short health privacy checklist by O'Keefe et al. [158] is a promising sign that de-identification considerations can be simplified in the context of a specific domain, thus enabling uptake of proper de-identification methods by practitioners. The sport research domain is fundamentally different from population health research studied by O'Keefe et al.; in sport, there is unprecedented levels of existing auxiliary data an attacker could use for linkage, only a small number of participants in any match, participants are usually sampled from the same team rather than a simple random sample, and spatio-temporal movements are of intrinsic interest to the analysis so cannot be discarded. Thus a specialised study of de-identification methods for sport position tracking datasets, and an investigation of the extent to which re-identification attacks can undermine these methods, is needed to bring about better understanding of the privacy versus utility trade-off in the sport research domain.

5.3 Prevalence of Improper De-identification Methods: A Review of “Non-identifiable” Datasets used in Australian Rules Football Research

5.3.1 Abstract

Background Non-identifiable datasets are afforded special exemptions under privacy law and human research regulations. For a dataset to be made *non-identifiable*, it must undergo a de-identification process to ensure that “no specific individual can be identified” [150]. Stripping personally identifying information while leaving other fields is now a “discredited approach” [163] due to the risk of linkage in data-rich environments. De-identified sport datasets, particularly GPS tracking datasets, involve spatio-temporal microdata that may be vulnerable to re-identification.

Objectives To determine the prevalence of improper de-identification methodologies used in Australian Rules Football research.

Data sources Electronic literature search using Google Scholar to sample both academic and grey literature.

Study eligibility criteria Study involves collection or use of data related to Australian Rules Football, and makes a claim about the identifiability status of players or teams in the data.

Study appraisal and synthesis methods Re-analysis of all variables collected from the perspective of re-identification attacks, particularly data linkage attacks.

Results In two of three studies, the underlying research dataset was claimed to be non-identifiable, but included identifying attributes that could have revealed non-public information about participants. None of the studies allowed re-identification of the participants on the basis

of published data alone.

Conclusions The de-identification approaches prevalent in sport research provide the illusion of de-identification, but offer no protection against deliberate re-identification attempts. Sport researchers and Human Ethics Advisory committees should not declare data *non-identifiable* without consideration of re-identification attacks that may link records to the ever-growing body of identifiable sport data.

5.3.2 Introduction

Rationale

The Australian National Statement (updated 2015) [150] on ethical conduct in human research defines *non-identifiable* data as data “which have never been labelled with individual identifiers or from which identifiers have been permanently removed, and by means of which no specific individual can be identified.” In contrast, datasets where “identifiers have been removed and replaced by a code, but it remains possible to re-identify a specific individual” are known as *re-identifiable*.⁷

However, despite the seemingly clear distinction between “non-identifiable” and “re-identifiable”, there are different ways that participants can be re-identified despite the best efforts of the data provider to make the data non-identifiable. For example, in 2016, the Australian Department of Health released de-identified medical billing records to the Open Data portal data.gov.au. However, after public release, researchers later found that encryption of supplier IDs could be reversed. Furthermore, even without decryption, there was sufficient auxiliary information in the dataset, such as year of birth and date of baby delivery, that it was possible to re-identify individuals in the dataset via

⁷This was recently revised in a 2018 update of the Australian National Statement, which revoked the section containing the definitions of “non-identifiable” and “re-identifiable” to recognise that data identifiability lies on a “continuum”. Nevertheless other sections of the Australian National Statement, such as the criteria for exemption, still continue to use the phrase “non-identifiable”.

data linkage [49]. Thus effective data de-identification is non-trivial, even when performed by mature government organisations.

Spatio-temporal datasets are particularly problematic, as human movements are often unique to an individual, thus serving as a potential fingerprint [53]. Inertial measurement sensors in vehicles can be used to fingerprint drivers with even a single vehicle turn (one turn is enough to distinguish between 12 drivers with 95% accuracy) [35]. Speed alone is enough to infer a driver's location through inference of the possible road combinations the driver could have taken [81]. Thus spatio-temporal datasets deserve special attention when evaluating whether the participants are really non-identifiable.

The above examples show that good faith efforts to de-identify data may be inadequate to protect research participants' right to privacy. As such, it is necessary to consider the perspective of an adversary actively attempting to re-identify participants. While it may seem unrealistic that one would go to such extents to re-identify the data, resistance to possible attacks is a necessary condition for a dataset to be *non-identifiable*. Even if one does not deliberately attempt to undermine the de-identification scheme, there is a risk that black-box machine learning techniques such as neural-networks could internally re-identify players without the researcher's knowledge if re-identification results in better classification accuracies. The identity of the players could then be inadvertently discovered by the researcher when examining the model diagnostics.

Objectives

This section attempts to establish an estimate of the prevalence of improper data de-identification methodologies in sport datasets. As sensitive datasets are not made public, much of the data sharing between sport practitioners and researchers occurs in private, and thus it is difficult to quantify the extent of data de-identification issues. As such, a systematic search was conducted for published literature (including supplementary information and grey literature) for hints of whether un-

derlying research datasets were truly non-identifiable.

The intent was to select studies that related to sport data analysis, particularly studies that involved Australian Rules Football or use of GPS tracking data, that made a claim of non-identifiable data.

The question this section aims to address is: *What is the prevalence of improper de-identification methodologies in sport research, specifically research into Australian Rules Football?*

Specifically:

1. Are there any risks that de-identified information could be re-identified in the published data?
2. Are there any risks that participants in the underlying de-identified research dataset could be re-identified by the authors of the study, or other researchers the dataset is shared with?

5.3.3 Method

Eligibility criteria

Studies were included if they used the keywords with the intended semantic meaning in the main body of the text (e.g. “football” must refer to Australian Rules Football rather than Association Football, and be the focus of the study rather than mentioned in reference to another study), and made a claim about the identifiability status (e.g. “non-identifiable”) of research data relating to players or teams.

Information sources

Searches were conducted to identify a sample of studies that claimed to use “non-identifiable” data. As de-identification is a methodological detail unlikely to be mentioned in the paper abstract, a database

that supported full-text search was required. Any publicly released research, regardless of journal acceptance is a threat to participant privacy, and thus it was desirable to chose a database that included a broad range of literature, including reports and pre-author drafts. Considering these criteria, Google Scholar was selected over more established academic databases that only search particular journals and/or only index abstract-title-keyword fields⁸.

Search

Google Scholar⁹ was searched with the following keyword combinations¹⁰: “Australian” “football” “non-identifiable” “GPS”; and “AFL” “non-identifiable” “GPS”. As neither of these combinations yielded any results that met the screening criteria, the search was broadened to include the 10 most relevant results (as determined by Google Scholar’s algorithm¹¹) for both: “Australian” “football” “non-identifiable”; and, “AFL” “non-identifiable”.

Study selection

Studies were screened on the title and snippets (keywords within context) shown by Google Scholar to determine whether they used the key-

⁸E.g. Scopus claims to be “the largest abstract and citation database of peer-reviewed literature”. However, as it only includes abstract-title-keyword fields and other meta-data, details of whether papers used non-identifiable datasets are unlikely to be captured.

⁹Searches were performed using the Australian interface to Google Scholar scholar.google.com.au. Each search was performed in a non-logged-in, private-browsing session as a precaution against any influence from previously search history. Search results were compared to those obtained through a US proxy to ensure they were not impacted by the IP address or region used to perform the search.

¹⁰Google Scholar automatically attempts to find synonyms, but the “results are often erratic” <https://uregina.libguides.com/c.php?g=606135&p=4201992>, thus this feature was disabled by wrapping words in quotes to regain manual control. While Google Scholar recognises Boolean operators, it does not support construction of complex Boolean expressions using parentheses <https://academia.stackexchange.com/questions/62881/how-can-i-use-parentheses-in-google-scholar>, thus separate searches for each combination of interest were performed.

¹¹While the Google Scholar ranking algorithm is not public, reverse-engineering attempts reveal that it is heavily influenced by citation count [19].

words in the intended context.

Following screening, study full texts were retrieved to determine whether they met the eligibility criteria (see Eligibility criteria, Sec. 5.3.3). In cases where the search returned a pre-publication author draft, an attempt was made to retrieve the final published version (published full texts of three of the papers that passed screening were accessible, and the other was retrieved through an inter-library loan). In cases where the final published version excluded information in the draft, identifiable information revealed in the draft was still considered, as any public research artefact, regardless of whether published through traditional means or informally, still poses a threat to participant privacy.

Data collection process

Information about the claimed identifiability level of data in each study was extracted, along with a list of variables collected or analysed in each study. This allowed an assessment to be made as to whether the variables collected in the study were consistent with the claimed identifiability level.

Data items

Each study was examined to extract the following information:

1. Claims relating to the identifiability status of the data (e.g. “non-identifiable”)
2. Claims relating to the ethics status of the project and the approving body (e.g. ethics exemption approved by university human ethics committee).
3. Claims relating to participant consent
4. Table of variables (as described below)

For each study, each variable was examined to extract the following information:

1. Variable used in the study (example: player birthdate)
2. Level the variable was collected at (examples: microdata collected for each individual; group level summary obtained from an external provider)
3. The data type of the collected variable (examples: categorical; list of real numbers pertaining to set of body measurements)
4. The level the variable was published at (example: study may have collected microdata at individual level, but only published group level summaries)
5. Whether the attribute was already public data. Prior knowledge and online information was used to determine to what extent each variable was already public
6. Whether, given access to the raw data collection used by the researchers, the variable was likely to be re-identifiable via data linkage. This was determined by considering linkages via the other attributes of the dataset that were public.

Summary measures

Variables are highlighted as being improperly de-identified when:

- The variable relates to an individual, and
- The variable is not already public, and
- The study claimed (re-)use of de-identified data, and
- The study did not explicitly seek consent from participants to use this variable, and

- The variable is likely to be re-identifiable in the underlying data set (but not necessarily a threat once published). See Data items (Sec. 5.3.3) for details of how this was determined.

The number of studies that involved at least one improperly de-identified variable, as a ratio of the total number of studies analysed, was used as an indication of the prevalence of improper data de-identification.

5.3.4 Results

Study selection

A PRISMA Flow Diagram [142] showing the study attrition is provided in Fig. 5.3. As the focus is on underlying datasets used by the studies rather than the studies themselves, two studies [174, 222] that both analysed the same dataset were grouped together.

Results of individual studies

Tables of variables (as described in Data Items Sec. 5.3.3) for datasets 1, 2, and 3 are presented in Tables 5.1, 5.2, and 5.3 respectively. Improperly de-identified variables (according to the procedures in the Summary measures section, Sec. 5.3.3) vulnerable to a data linkage attack are highlighted (bold red text). A full discussion of each dataset and ad-hoc analysis of other potential attacks is included in Appendix C.

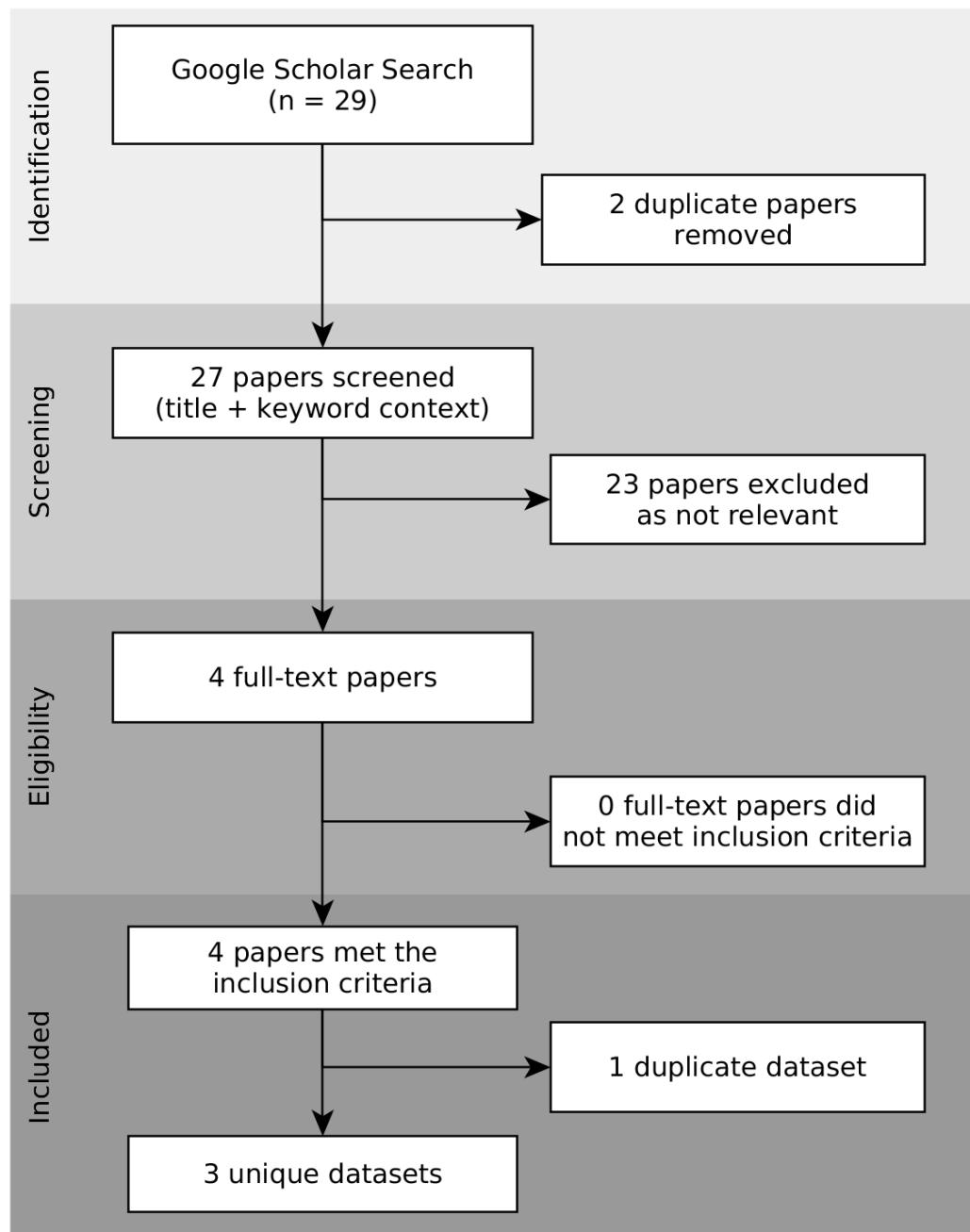


Figure 5.3: PRISMA Flow Diagram [142] showing study selection. From the literature search, the selection was refined to four relevant papers that made claims that the dataset used was non-identifiable. Two papers were grouped together that analysed the same dataset from different perspectives. All four relevant papers were subjected to scrutiny of risk of data re-identification.

Table 5.1: Robertson, Woods, and Gastin [174, 222]. Under 18 year old performance tests. Claims: “non-identifiable” data; approval by “relevant human research ethics advisory group”; consent of “state-based organisations”

Variable	Collection level	Collected type	Published level	Public Data?	Collection re-identifiable?	Published data re-identifiable?
Birthdate	Individual	Real	Group	If drafted	Already public if drafted	No
Anthropometric measurements	Individual	List Real	Group	If drafted	Already public if drafted	No
Physical performance	Individual	List Real	Group	Only if one of top results for test	By birthdate and anthropometric measurements	No
Drafted	Individual	Categorical	Group	Yes	Identifiable	Already Public

Table 5.2: Greenham et al. [89]. AFL team game style. Claims: “non-identifiable player data, from identifiable team-based data-sets” data; ethics exempt

Variable	Collection level	Collected type	Published level	Public Data?	Collection re-identifiable?	Published data re-identifiable?
Team name	Team	String	Team	Yes	Team level	Team Level
Shot at goal accuracy, Scoring accuracy, Location of goal attempts (as proportion near goal), Passes/min, Passing efficiency, Rate of ball recovery, Weighted position of turnovers (per zone), Weighted position of turnovers weighted by score (per zone), Tackles/min of defence	Team	List Real	Team	Detailed statistics need licence from Champion Data (AFL statistics provider)	Team level	Team level
Ball Speed (from video)	Individual	Video	Team	Yes	Identifiable	Already Public
Offensive-defensive numbers (from behind the goals video), Total numbers in forward 50 (from behind the goals video)	Individuals	Video	Team	Behind the goals video only provided to coaches	Identifiable	Team level

Table 5.3: Jacob et al. [112]. Genetic markers for performance. Claims: “non-identifiable” code; “University granted approval” (“Human Research and Ethics Committee approval number” provided); direct consent of participants (and parents where under 18)

Variable	Collection level	Collected type	Published level	Public Data?	Collection re-identifiable?	Published data re-identifiable?
Team Name	Team	Text	Hidden	Yes	Team level	Acknowledgement of team in pre-publication draft
Genotype (from blood sample)	Individual	List Categorical	Group	No	May correlate with family members, race, and ethnicity	No
Physical performance	Individual	List Real	Group	No	By genotype	No

Synthesis of results

Improperly de-identified variables have been highlighted in bold red in the previous section. It was found that two of the three datasets analysed involved improper de-identification of the underlying dataset. In one case re-identifiable data were claimed non-identifiable: under 18 year old performance results are non-public (other than top performers), but the results could be linked back to the individual by birthdate and anthropometric measurements. In another case non-public identifiable video (behind the goals footage) that is only accessible to select groups (e.g. clubs) was improperly exempted from ethics review under the assumption that it was publicly available. These cases were both limited to improper de-identification of the dataset by the custodian prior to providing it to researchers. Scrutiny of the final published data (i.e. the summarised version published in the research paper) was unable to reveal any cases where it was possible to re-identify specific individuals without access to the raw dataset the researchers had access to.

5.3.5 Discussion

Summary of evidence

Considering that only three datasets were examined in full, the discovery of two datasets that likely did not meet their claim as being non-identifiable suggests that many more sport studies exist where participants have not been properly de-identified in the underlying dataset. Fortunately, the published data does not appear to be re-identifiable; however, this may be due to journal page limitations which have the effect of preventing authors revealing too much, rather than a matter of careful statistical disclosure policies.

Limitations

While unable to find any cases where the published data were re-identifiable through data linkage, the public data may still be vulnerable to other forms of attacks. In Appendix C, a consideration is provided of other forms of attacks for some of the articles analysed; however, much like code review to catch bugs, review will catch some cases of re-identifiable data, while others may have slipped through unnoticed.

Conclusions

In conclusion, although it was not possible to identify any individual players using the published data alone, in some cases it was possible to reduce the possible candidates by utilising information in the study in a manner contrary to the intentions of the author. However, the investigation suggests that the underlying dataset used by researchers in some of the studies was likely re-identifiable despite claims that the data were non-identifiable. This is concerning, as it can lead to re-identifiable data being used or re-shared by researchers without proper consent from the individuals to whom it pertains.

Anonymisation of team names was an unsuccessful counter-measure against attacks on participant privacy. While unintentional, ambiguity of method, or missing data served to add uncertainty which prevented identifying individuals. As research moves into an era where page count is no longer a limitation (e.g. electronic attachments), datasets are reusable (though greater access to eResearch tools to preserve datasets and facilitate reuse through capture of meta-data), methods are unambiguous (e.g. through publication of analysis source code), and data quality is high (e.g. data capture using ubiquitous sensor networks with redundancy), this uncertainty will reduce, thus formal privacy techniques will become more important to avoid revealing individual participants via “differencing attacks” [160]. In particular, the group of participants that were excluded from the study will reduce, possibly to single participants, and the information of these participants could

be inferred through differences between overall versus clean data, or from differences between studies with slight differences of filtering techniques.

5.4 Threat Model

For the purposes of threat modelling, the processes and actors involved in the creation and use of de-identified data are shown in Fig. 5.4. While there are many threats that could potentially undermine the de-identification process (e.g. penetration of the club's computer systems, or manipulation of the pseudorandom number generator used to generate identifiers), the focus of this chapter is on threats to participant privacy due to inherent weaknesses in the de-identification methods themselves rather than attacks on the surrounding computational infrastructure (while attacks on the computational infrastructure are obviously an important concern in practice, these are not specific to the de-identification process).

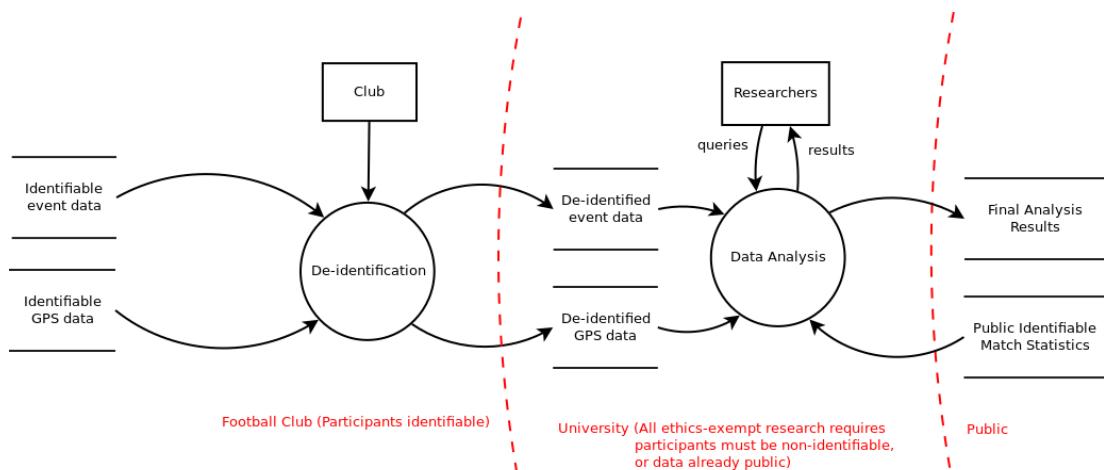


Figure 5.4: Data Flow Diagram for data de-identification, analysis, and publication. For purposes of threat modelling, trust boundaries (presented as red dashed lines) are marked to delineate the participant privacy protections that must be maintained within each organisation

While AFL matches are played in public, and match statistics for each player and video footage are made publicly available, there are additional datasets the public do not have access to. This section provides

a brief summary of these datasets, highlighting how they provide access to player information beyond what the public has access to. There is risk to players that if an attacker or unethical researcher were able to re-identify the dataset, it could cause discomfort to players through an unprecedented level of player scrutiny, or be used in ways contrary to the players' wishes, such as to inform betting odds.

Nowadays, all players wear position tracking devices on their back during each game. These devices are equipped with GPS units, in addition to Inertial Measurement Units (IMUs) containing accelerometers and gyroscopes that complement¹² the GPS position estimates for short distances. The devices also have the ability to record heart rate; however, heart rate straps not always worn during games. Because the devices contain accelerometers, they are capable of detecting each footstep that a player makes. It has been suggested that accelerometers could be used to determine onset of musculoskeletal injuries by detecting asymmetries in acceleration between each foot [219]. Commercial sport position tracking device manufacturers have incorporated running symmetry analysis functionality into their software and claim that it can differentiate healthy players from those in rehabilitation.¹³ However, the validity of the running symmetry analysis reported by commercial devices has been questioned [117]. Nevertheless, the possibility that position tracking devices may reveal player injuries or other health conditions (albeit in its research infancy) highlights the importance of ensuring these data are properly de-identified before sharing them.

Champion Data, the official AFL statistics provider, manually record each on-field event that occurs during the game, including the time at which it occurred. These are shared with teams as “possession chain”

¹²Accelerations captured by IMUs can provide valuable [151] (and potentially sensitive) insights into sport player performance, independent of the position tracking data. While IMU data are typically treated separately to GPS data, some manufacturers use sensor fusion to combine information from IMUs with information from GPS units in order to provide high frequency position estimates [10]. Note however, that higher frequency does not necessarily mean better accuracy. See Catapult Sports, “Demystifying sample rate in satellite-based athlete tracking technologies” <https://www.catapultsports.com/blog/sample-rate-satellite-athlete-tracking-technologies> Accessed: 2019-11-25

¹³GPSports, “Running Symmetry Analysis,” 2014. <https://web.archive.org/web/20140717135307/http://gpsports.com/running-symmetry-analysis/>

data, which can be used to scrutinise each action that a player makes. While one could theoretically derive the possession chain data via manually coding public match footage themselves, it would be prohibitively time consuming to code every match to the same level of quality that Champion Data provide. Thus while Champion Data possession chains for a short match segment do not provide any information beyond what is publicly accessible, when taken over a long period they may reveal highly detailed player profiles beyond that in the public domain.

When possession chain data and GPS data are de-identified using the same scheme such that the two datasets can be linked, it means that a compromise of the de-identification scheme used by *either* of the datasets can be used to de-identify *both* datasets via the data linkage. Analogous to the concept of privileged escalation in computer security, the concern is that re-identifying a player in a non-sensitive dataset such as short segment of possession chain information could lead to compromise of a more sensitive dataset, such as re-identifying player GPS traces for the length of the entire game which in turn could reveal sensitive knowledge about a player due to the fine-grained level that GPS devices can capture player movements when complemented with IMUs.

Attacks against possession chain data are demonstrated on a sample of possession chain data for one match during the 2015 season.¹⁴ Attacks against GPS data are demonstrated by explaining how player trajectories could have been re-identified if it were not for the de-identification approach used in this thesis.

¹⁴As noted, possession chains can be derived from public match footage. Thus privacy issues only arise when possession chain data are recorded over multiple matches in order to build detailed player profiles or linked to GPS data.

5.5 Re-identification of Sport Data

5.5.1 Re-identification of possession chain data

Attacks against removal of date and venue

This section begins by outlining a simple attack to re-identify player possession chain data, an example of which is provided in Table 5.4. Possession chain data may also include the time of each event, but as will be shown, the events alone are enough to re-identify the players that perform them. In the example, the team and player fields have been anonymised through substitution of player names with anonymous identifiers. The example further attempts to thwart the attacker by removing all details about the match, and assumes the attacker only knows minimal details such as that it was an AFL match occurring on an unknown date at an unknown time during the main 2015 season.

Table 5.4: “Anonymised” Player Scoring Chain Data

Chain	Team	Chain Start Type	Chain End Type	Disposal	Player ID
1	B	Stoppage - CB	Rushed	kick	131
2	A	Kick In	Stoppage - TU	kick	123
3	B	Stoppage - TU	Stoppage - TI	kick	108
4	B	Stoppage - TU	Goal	kick	137
5	A	Stoppage - CB	Turnover	kick	124
6	B	Turnover	Turnover	kick	104
7	A	Turnover	Goal	kick	139
7	A	Turnover	Goal	handball	126
7	A	Turnover	Goal	kick	122
8	B	Stoppage - CB	Turnover	kick	117

An attacker can use the frequency of each action players perform as a fingerprint that the attacker can compare to public player statics to re-identify players. An attacker does not require any specialised skills or software to do this; indeed, an attacker can perform this task in a matter of seconds using an Excel pivot table, as shown in Fig. 5.5.

The tabulation (Fig. 5.5) reveals that team A scored a total of 16 goals, and 11 behinds (8 ordinary behinds, 3 “rushed” behinds). This is typi-

The screenshot shows an Excel spreadsheet with a PivotTable set up. The data is organized by Team, Anonymised ID, and Chain End Type. The PivotTable Fields pane on the right side of the screen lists various data types: Chain, Team, Chain Start Type, Chain End Type, Disposal, Anonymised ID, and Transition. The Rows section of the PivotTable Fields pane includes 'Anonymised ID' and 'Sum of Transi...'. The Values section includes 'Sum of Transition'. The main table displays various statistics such as Behind, End of Qtr, Goal, Rushed, RushedOpp, Stoppage - TI, Stoppage - TU, Turnover, and Grand Total.

Team	Anonymised ID	Chain End Type	Behind	End of Qtr	Goal	Rushed	RushedOpp	Stoppage - TI	Stoppage - TU	Turnover	Grand Total
	101		0		3	0		2		2	7
	103		0		4	0		0		2	8
	105		2		0			1	0	4	6
	109		1		0			2		5	8
	112		0		0	1		2	1	11	4
	113		0		0	0		0		2	3
	115		0		1	0		0		0	2
	116				0			2		0	1
	118		0		1			0		0	1
	120		1	1	1	1		2	0	3	9
	122		1		2			0		1	4
	123		0		0	0		1	1	5	7
	124		1		0	0		2	1	5	9
	125		0		1	1		0	2	1	5
	126		0		1	0		0		5	6
	128		1		0	0		0		2	3
	130		0		0	0		3		0	3
	132		0		1	0		1	1	2	5
	134		1		0		1	3		4	9
	139		0		0			1	0	4	5
	141				1	0		0		2	3
	144		0		0	0		3		1	4
	Grand Total			8	1	16	3	1	27	6	56
											118

Figure 5.5: Given scoring chain data, an attacker can easily tabulate summary statistics at both a team and player level in a matter of seconds using an Excel pivot table

cally reported as "16.11". The attacker then consults a table of matches played during the season, such as that found at AFL Tables¹⁵, and searches for text "16.11". In 2015, there was only one team that scored this exact combination of goals and behinds: Hawthorn in the 2015 grand final. This shows that removing the date of match and team names did not serve as any meaningful guard of participant privacy, as an attacker was able to infer this information through using the score totals as a fingerprint.

Attacks against substitution of participant identifiers in event data

Once the attacker obtains the match information, they can proceed to re-identify individual players. A player's actions serve as a fingerprint. For example, the tabulation (Fig. 5.5) shows that player 105 (row 7) scored 2 behinds and 4 goals. The attacker can compare this with the public player statistics for that game¹⁶, which uniquely identifies this player as Cyril Rioli. Now that the attacker knows Cyril Rioli = player 105, they can scrutinise this player's full scoring chain data as well as any other datasets that were de-identified using the same player code.

¹⁵<https://afltables.com/afl/seas/2015.html>

¹⁶<http://www.afl.com.au/match-centre/2015/27/haw-v-wce>

5.5.2 Re-identification of GPS data

Attack against substitution of participant identifiers in spatio-temporal data

As explained in Sec. 5.2.1, spatio-temporal data are inherently difficult to de-identify. GPS devices record both player location, and precise time. Even a single time-instant is enough for the attacker to re-identify which traces belong to which players via comparison to match video footage—for example, in Fig. 5.6, an attacker re-identifies the players involved in the centre bounce at the start of the 3rd quarter, a task that is made easy by the large numbers printed on players' guernseys to make their identity known. Once the attacker has re-identified a player at one time instant, they can examine the minute details of every movement a player makes over the course of the match, including the profile of each footstep.

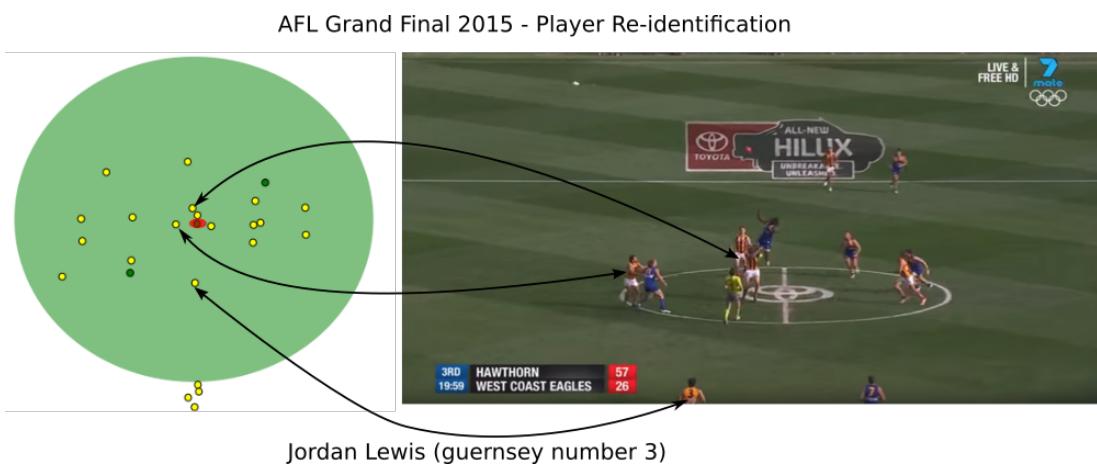


Figure 5.6: Left: Location of players on field at start of 3rd quarter, as obtained from GPS data. Right: Video footage of game 1 second into the 3rd quarter. Given “anonymised” GPS data, an attacker can easily re-identify players through comparison with video footage

Attacks against spatio-temporal distortions

One could attempt to prevent the ability of the attacker to re-identify the GPS data by shifting and/or distorting the time axis. However, this

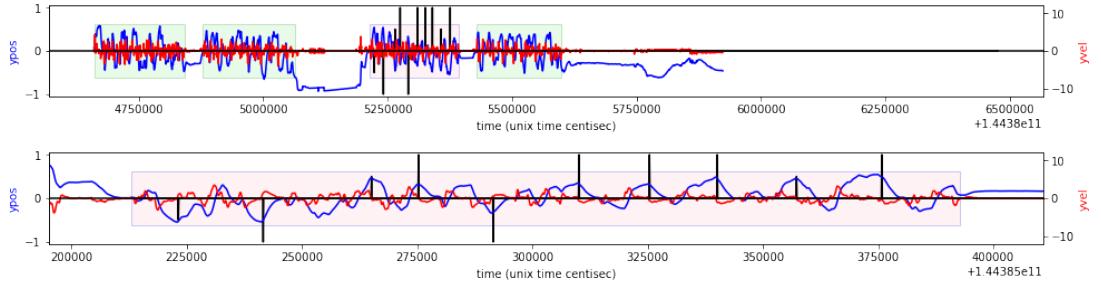


Figure 5.7: Top: Average position in direction of goals (blue line) and velocity (red line) of players during match, as obtained from GPS data. The overlaid rectangles show the duration of each quarter, which correspond with the level of activity. Bottom: A zoomed in section of the above focusing on the third quarter. The overlaid vertical black bars represent the time of goals (large bar) and behinds (short bar) for each team (direction of bar). Note that the time of goals and behinds correspond to peaks and troughs of average team position (blue line) in the direction of goals, and short periods of inactivity (red line crossing x-axis) after a goal is scored.

is unlikely to offer any meaningful protection, as events in the GPS data align closely with scoring events (see Fig. 5.7), thus allowing the attacker to re-identify the precise game instants that the GPS data correspond to. Indeed, in the datasets used in this thesis, the precise time of match events were not provided; however, in most cases¹⁷ it was possible to automatically synchronise these with the GPS data through optimising for the time shift that resulted in the best alignment of match events with peaks and troughs in the GPS data. One could attempt to counter this through non-uniform time distortions; however, this would risk compromising the quality of the analysis due to distortion of player speed, and could still be circumvented through the use of a more sophisticated temporal comparison algorithm, such as Dynamic Time Warping (DTW).

One may also attempt to prevent the ability of the attacker to re-identify

¹⁷In quarters where teams scored periodically at consistent time intervals, offsets of $n \cdot \text{interval duration}$ provided a good fit for any n , and thus automatic time synchronisation sometimes selected the wrong offset in these edge cases. These cases were identified and resolved through manually inspecting the GPS formations at key events and interactively manipulating the time offset to find a better fit (Sec. 7.1.7). As these edge cases could be corrected manually, it is likely that a fully automated approach may be possible using a more sophisticated algorithm that examines the entire team formation rather than just average team position and velocity.

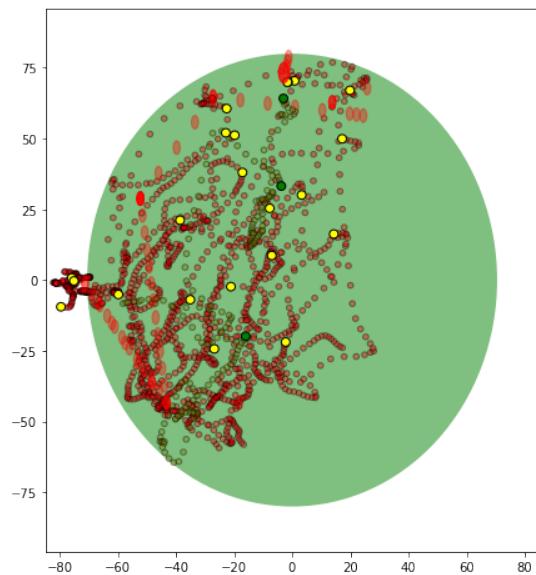


Figure 5.8: 1 minute of player GPS movements during a game, as the ball is passed along the edge of the field. The true shape of the field is indicated by the green oval; however, the location of the interchange area (left of image) and shape of the oval are visually apparent from the GPS data alone.

the GPS data by performing spatial translations, rotations, and reflections. However, players frequently pass the ball along the edge of the field, which makes the bounds of the field visually apparent; an example of GPS movements that reveals the edge of the field is shown in Fig. 5.8. Once the shape of the field is identified, the only additional details that an attacker need to concern themselves with are identifying any reflections. However, this can be achieved by examining the GPS data at the time of a goal to see which side of the field the players were occupying. Indeed, in the datasets used in this thesis, none of the datasets included the “coin-toss” used to decide which direction players goals are; however, it was still possible to infer this through considering both possible directions then choosing the direction that resulted in the best alignment with the public match event feed. In conclusion, linear spatial transformations are insufficient to prevent player re-identification. Any operation that distorts scale, including non-homogeneous scale distortions such as a non-linear transformation, are not an option, as they would affect most forms of spatial analysis such as player speed. Furthermore, the attacker can use known information about a player’s typical speed to infer the scale distortion, thus allowing the transfor-

mation to be inverted.

Preventing attack by complete removal of identifiers

Due to the ease with which an attacker can undermine the aforementioned de-identification methods, it is necessary to consider a more extreme approach. Note that the value to an attacker lies in the ability to re-identify players at one time instant, and then use this to re-identify the player at other time instants, or in the case of linked datasets, to re-identify the player in other datasets that use the same anonymised player code. By completely removing the player identity column, the link between player movements at different time-instants can be destroyed.

An example of a fully de-identified possession chain data is given in Table 5.5. The team name is not anonymised in the example, because as previously shown, teams can be trivially re-identified by an attacker based on the number of scoring events¹⁸ yet removing team names would hinder legitimate data analysis seeking to compare identified patterns to public information about the team.

Table 5.5: Player Scoring Chain Data, de-identified by completely removing player identifiers

Chain	Team	Chain Start Type	Chain End Type	Disposal	Player ID
1	WCE	Stoppage - CB	Rushed	kick	NULL
2	HAW	Kick In	Stoppage - TU	kick	NULL
3	WCE	Stoppage - TU	Stoppage - TI	kick	NULL
4	WCE	Stoppage - TU	Goal	kick	NULL
5	HAW	Stoppage - CB	Turnover	kick	NULL
6	WCE	Turnover	Turnover	kick	NULL
7	HAW	Turnover	Goal	kick	NULL
7	HAW	Turnover	Goal	handball	NULL
7	HAW	Turnover	Goal	kick	NULL
8	WCE	Stoppage - CB	Turnover	kick	NULL

Position tracking data are typically stored as separate files for each individual player trajectory. Even if the files names contain no identifying

¹⁸An attack to re-identify the team name is covered in *Attacks against removal of date and venue* (Sec. 5.5.1)

$$\text{Formation} = \{(x_a, y_a), (x_b, y_b), \dots\} = \{(x_b, y_b), (x_a, y_a), \dots\}$$

Figure 5.9: A point cloud is an unordered set of points. Representing team formations in this manner allows analysis of the team, while removing ties to individual player identities. As operations on sets are invariant to the ordering of elements within the set, the ordering of players within the formation can be randomised when serialising the data to deter attackers without affecting the analysis.

information, storing a single player's trajectory for the entire match in the same file means that the player can be re-identified for the duration of the entire match if an attacker can re-identify the player at a particular moment. For example, if an attacker knows that a particular player scored a goal at a particular time, they can re-identify the trajectory file for that player by searching for trajectory files that contain a point at that time and location. This then gives the attacker access to detailed movements of that player over the course of the entire match.

This can be prevented by merging all the player traces together and splitting them into a sequence of team formations at each time instant. Representing the team formation as a *point cloud* (an unordered set of points) removes the link between player locations at different time instants. While this prevents individual player analysis, it still permits many forms of team analysis methods. A visual explanation is provided in Fig. 5.9. This approach was inspired by Ding et al. [59] who show that randomly mixing human trajectories whenever two people cross paths prevents re-identification of individuals, while still allowing certain forms of analysis about the group.

To circumvent this method, an attacker can attempt to reconnect the links between locations belonging to the same player by pairing an occupied location at one instant with the closest occupied location the next time-instant. A more sophisticated approach can use player velocity to estimate the next location, then take the nearest occupied location to the estimated location. However, this would be difficult in practice as players often come into close contact with each other. Every time players interact, there is a branching of the possible locations the player could be in next. Thus while one can infer related player movements for a short period of time, the number of possible combinations increases

with time [59]. This means that even if an attacker were to determine a player’s location at a certain point in time (e.g. from a photo at a known time instant), they will only be able to trace the player for a short time due to increasing uncertainty about which player is which when trying to trace players over the medium to long-term course of a match.

Downsampling

As noted, the key threat to participant privacy is that an attacker can gain access to high frequency GPS data, that when complemented with accelerometers, can be used to track a player’s individual footsteps, and thus potentially infer sensitive player information.¹⁹ While downsampling doesn’t intrinsically serve as a de-identification technique, hence it will not be classified as a de-identification method in Sec. 5.6, lossy downsampling could result in a reduction of the sensitivity of player data, thus reducing the potential harm to participants if re-identified. It could also be used in combination with removal of identifiers to make re-identification via trace reconstruction harder, due to increased uncertainty of player movements when sampled at a lower frequency.

5.6 Analysis of Trade-Off between Participant Privacy and Data Quality

This section compares the strength of each de-identification method outlined in Sec. 5.5, and considers how the de-identification impacts on the utility of the data to researchers. Specifically, it considers the constraints on the kinds of analysis that can be performed as a result of de-identification of the data. This is presented in Table 5.6.

Attempts to prevent player re-identification, through replacement of names with anonymous identifiers, removal of event timestamps, and shifting/distortion time and space, are weak. An attacker can obtain

¹⁹As outlined in *Threat model* Sec. 5.4

Table 5.6: Analysis of trade-off between de-identification strength versus utility of data after de-identification

Data Type	De-identification Method	Attacks	Identifiable Data Required for Attack	Attack Difficulty	Analysis Constraints*
Event Sequence	Remove date and venue	Fingerprint total events	Totals of each event type	Low	Independent of date and venue
Event Sequence	Substitute participant names with anonymous identifiers	Fingerprint total participant events	Totals of each event type for each participant	Low	
Spatio-Temporal	Substitute participant names with anonymous identifier	Compare to video or images	Photograph of participants at known time instant	Low	
Spatio-Temporal	Spatio-temporal distortions	Correlate with event sequence	Sequence of key events	Medium	Invariant to translation, rotation, and reflection
Spatio-Temporal	Remove participant ids	Reconstruct traces	Photograph of participants at known time instant	High	Invariant to participant permutations

* Additional analysis constraint for all de-identification methods:
Can only link to datasets that use same anonymous identifiers

data fingerprints for one or more players, and find the mapping between anonymised player code(s) and the player's identity.

The only method that offered protection against re-identification of players was complete removal of participant identifiers and use of a point cloud data structure (i.e. an unordered set of points) at each time-instant. This de-identification method removes the ability to trace the path of individual players. For stronger protection, this can be combined with downsampling of the position tracking data (e.g. from 10 Hz

to 1 Hz²⁰), thus increasing the uncertainty of player identities when two players come into close contact during the time window. This also has the effect of filtering out the high frequency component of the player position information that risks revealing individual player steps.

Note that in order to achieve privacy, it is necessary to reduce data quality, thus limiting the types of analysis that can be performed. However, if the goal is macro-level team-level analysis rather than micro-level studies of player movement, then this trade-off may be acceptable. Allowing individual player movements to be traced over the course of the match is in contradiction to the goal of de-identifying data; if individual player movements can be traced, then players can be fingerprinted by their movement and scoring patterns, and thus the dataset would be re-identifiable rather than non-identifiable.

²⁰The sampling rate chosen depends on the application. Lower sampling rates offer stronger privacy, but discard information that may be needed for the analysis. The rate of 1 Hz was chosen for this thesis in order to match the needs of the analysis performed in Chapter 7, which analysed team-level formations at 1 second intervals. A higher sampling rate could have been chosen if the analysis required detailed player movements (at the cost of increased risk of data re-identification).

5.7 An Interaction Model for De-identification of Human Data held by External Custodians

The previous sections of this chapter explained that data de-identification, particularly of spatio-temporal datasets such as the AFL dataset used in this thesis, is non-trivial to perform correctly. The analysis of trade-offs in Sec. 5.6 revealed that common methods such as removing dates and venues, offsetting times, and substituting player names with random codes fail to prevent re-identification. Use of a point cloud data structure to represent team formations at each time instant was identified as the most theoretically promising in terms of individual privacy protection without destroying the ability to analyse the team as a whole. However, this method is rarely used in practice, and requires writing custom software to implement.

The challenge is that for all data to be *non-identifiable*, the de-identification must be performed by the data custodian, i.e. the football club, who do not have the technical resources to apply sophisticated de-identification approaches. Thus despite the theoretical benefits of the approach identified by the previous sections, it could not be realised in practice under the common interaction model where the full burden of carrying out the data de-identification process is placed upon the data custodian without any involvement from the researchers that plan to analyse the data.

To ensure that secure de-identification approaches can be applied in practice, it was necessary to reconsider the interaction between data custodians and researchers. The goal was to find a solution whereby researchers could be involved in specifying the most appropriate de-identification process while still ensuring that the data custodian remained in control of the underlying identifiable data. The solution was formalised in terms of an interaction model and implemented as a de-identification portal. The portal was used to request the AFL datasets used in thesis and to automatically de-identify the data as they passed

through the portal using the point cloud method identified in the previous section.

This section of the thesis was published as Andrew J. Simmons, Maheswaree Kissoon Curumsing, and Rajesh Vasa. “An interaction model for de-identification of human data held by external custodians”. In: Proceedings of the 30th Australian Conference on Human-Computer Interaction. Melbourne, VIC, Australia: ACM, 2018, pp. 23–26. doi: 10.1145/3292147.3292207. An authorship statement for the paper can be found in Appendix F.

5.7.1 Abstract

Reuse of pre-existing industry datasets for research purposes requires a multi-stakeholder solution that balances the researcher’s analysis objectives with the need to engage the industry data custodian, whilst respecting the privacy rights of human data subjects. Current methods place the burden on the data custodian, whom may not be sufficiently trained to fully appreciate the nuances of data de-identification. Modelling of functional, quality, and emotional goals is used to arrive at a de-identification in the cloud approach whereby the researcher proposes analyses along with the extraction and de-identification operations, while engaging the industry data custodian with secure control over authorising the proposed analyses. This section of the thesis demonstrates the approach through implementation of a de-identification portal for sport club data.

5.7.2 Introduction

Researchers often wish to reuse a pre-existing dataset in a new unforeseen way to investigate a question. As a running example, this section of the thesis will consider a sport researcher requesting access to a dataset held by a sport club. As obtaining consent of every individual in the dataset to reuse their data is often infeasible, human ethics

guidelines permit exemptions if the data custodian de-identifies their dataset and provides it to the researcher in non-identifiable form, i.e. such that no individual can be re-identified [150].

As the data custodian may not be an expert in de-identification techniques, it is important that the de-identification system be learnable, and that it minimise the risk of user errors by the data custodian that could undermine the privacy of participants. Typically, sport club staff would be familiar with business spreadsheet software, such as Microsoft Excel, and remove or substitute identifiable columns such as player names.

However, de-identifying data is a non-trivial operation, as even after obvious identifiers are removed, “quasi-identifiers” [203], such as times, dates, or locations, may still allow re-identifying individuals in the dataset by linking sensitive data to public datasets. Privacy researchers have proposed software tools that automatically distort or generalise quasi-identifiers [125], however use of these tools requires a level of expertise from the user to select an appropriate privacy threshold, ensure that algorithm assumptions are met, and to minimise the destruction of data utility [26]. As sport club staff are under constant time pressure, it is unlikely that they would have time to develop the necessary expertise to apply these tools reliably, and the additional work and uncertainty may cause frustration that undermines the research partnership.

While existing tools for de-identification focus on quality goals or functional requirements, a solution is needed that also meets the stakeholders’ emotion-oriented requirements to ensure the research-industry engagement is successful. This section of the thesis focuses on the human-computer interactions involved in the de-identification workflow; the specific choice of de-identification operation is abstracted over, as this is best left to the researcher to decide given the type of dataset and privacy level requirements.

5.7.3 Emotional goal framework

While functional and quality goals are well-established as part of the software design process, all too often software designers overlook emotional needs of users [156, 172], resulting in an unfulfilling product which fails to gain appropriation by users as part of their workflow [137, 138].

The importance of users' emotional expectations during software design cannot be undermined. Thus the following sections look at the techniques proposed by Kissoon Curumsing [51, 52] to introduce the concept of emotional goals within the software design process.

Specifically, this section of the thesis utilises emotional goal modelling to consider the needs of each stakeholder in the design of the de-identification platform. User emotional acceptance of the system is critical to improving data sharing practices, else stakeholders are likely to revert to flawed but culturally ingrained [163] data sharing practices, such as substituting names with a randomised code while doing little to prevent the re-identification of individuals via data linkage using quasi-identifiers.

5.7.4 Modelling

Emotional goal framework

Fig. 5.10 breaks down the functional goals of the system and shows how these impact on the quality and emotional goals of users.

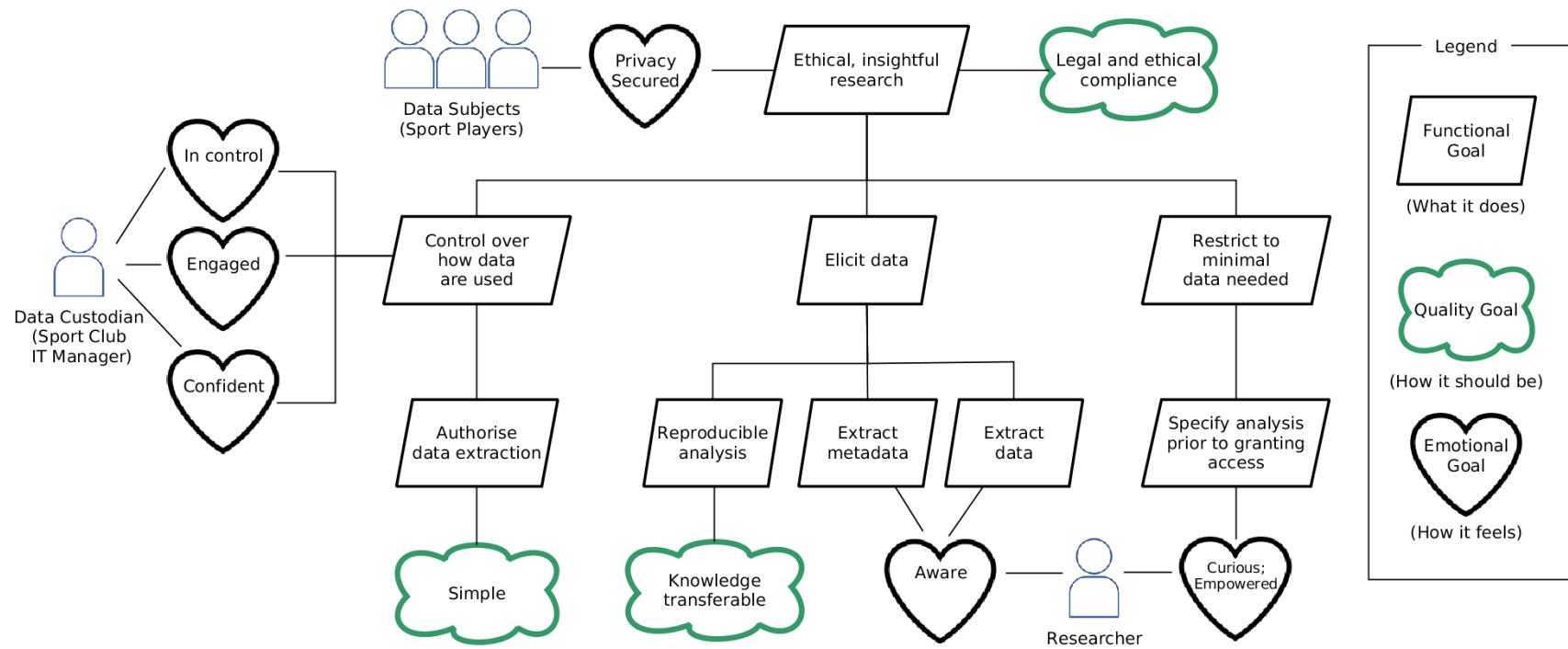


Figure 5.10: De-identification portal Emotional Goal Model (diagram should be read top to bottom)

The overall goal of the de-identification portal is to provide a platform for ethical, insightful research that facilitates reuse of data without compromising the privacy of the data subjects (i.e. the sport players). To achieve this goal, the control over how the data are used must lie with the data custodian (i.e. the sport club) rather than the researcher, as providing the researcher with unrestricted access to the system would be equivalent to transfer of identifiable data without participant consent. On the other hand, to encourage insightful research, the system should promote a mindset of intellectual *curiosity* whereby the researcher feels *empowered* to request (but not necessarily be granted access to) data and propose analyses that fully utilise the detail available in the dataset to gain an *awareness* that is not limited to traditional predefined summary statistics. To meet the goals of all stakeholders, Fig. 5.10 proposes that the researcher should precisely specify the data they need for an analysis by writing a script to perform the extraction. In cases where an analysis requires access to sensitive data, the extraction script should perform the analysis on the sensitive data then de-identify the output to ensure that it is non-identifiable. The data custodian (i.e. the sport club IT manager) should feel *engaged* in the research process and *in control* over authorising the execution of a script so that they can be *confident* the research is protecting the data subjects' (i.e. the players in their club) right to feel that their *privacy is secured*.

Interaction model

Fig. 5.11 shows the interaction model for the proposed system, and translates the goals into role interactions. This level introduces the role of a cloud data portal, an autonomous agent that will mediate the interactions between the data custodian and researcher in a secure manner. To ensure the data custodian remains in control of data access, in our solution they encrypt the data prior to uploading the data to the cloud. The researcher proposes an analysis by uploading a script to the cloud portal and providing a human readable summary for the data custodian. If the data custodian is satisfied that the proposed analysis is

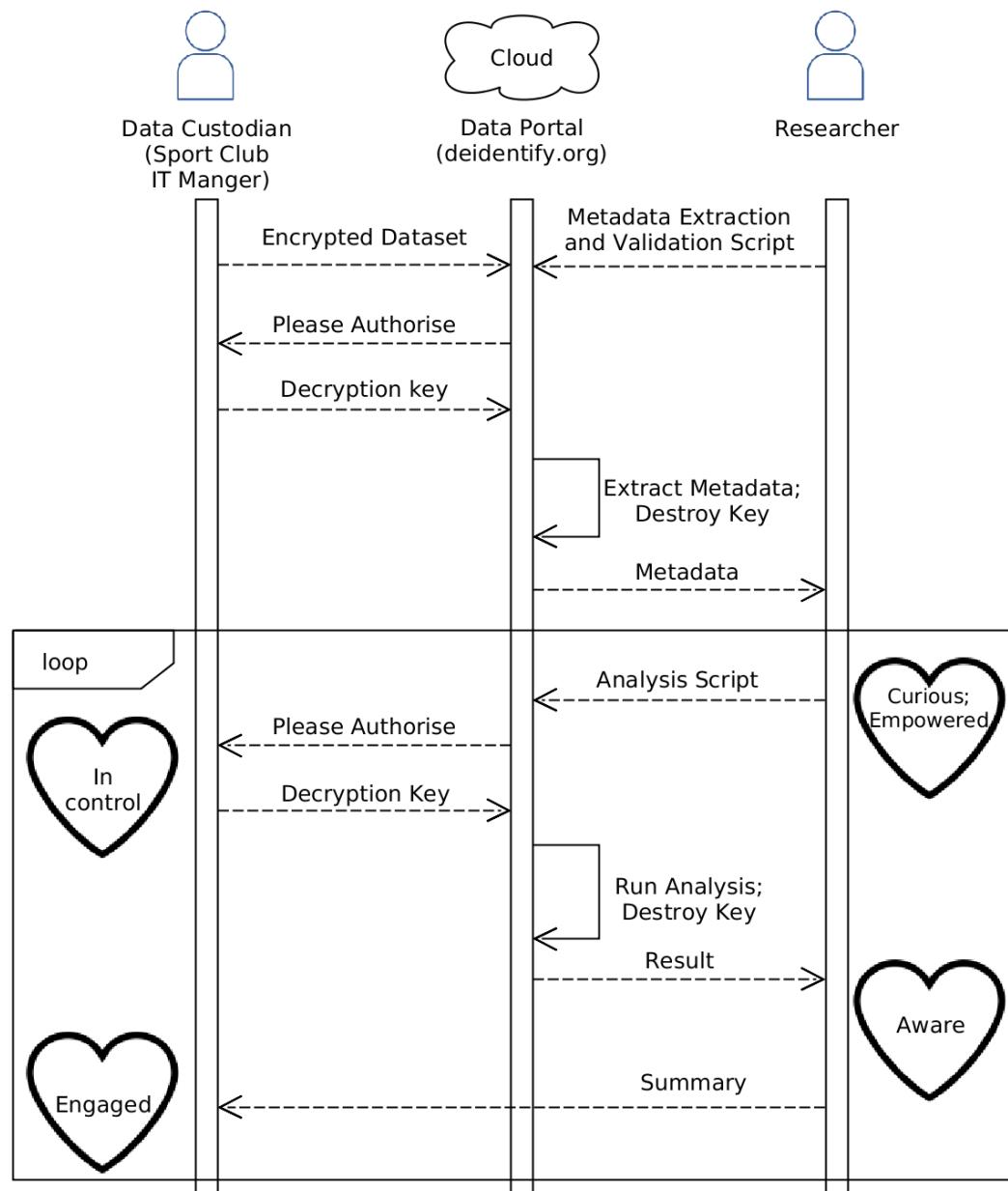


Figure 5.11: De-identification portal Interaction Model

respectful of the privacy and rights of the participants, they authorise the cloud portal to perform the analysis proposed by the researcher by providing it with the decryption key. The cloud portal uses the key to temporarily decrypt the data, runs the analysis script against the raw data, and finally destroys the key after script execution is complete. Upon completion the researcher is notified so that they can perform post-analysis on the results of the extraction script and communicate the results back to the data custodian. This is an iterative process; the first iteration is usually to extract metadata and validate the researcher's assumptions about the dataset. The following iterations deal with extracting data to answer a specific research question, which may prompt subsequent questions.

As the data custodian is *in control* of the authorisation of each phase, this has the additional benefit of keeping them emotionally *engaged* in the research process. As the analysis is run in the cloud, the researcher never sees the raw data nor the decryption key, and thus never has access to re-identifiable data; the researcher is *aware* only of the final output of their analysis, thus *empowering* the researcher to satisfy their feeling of *curiosity* about well-formed questions without revealing details that would compromise the data subjects' privacy.

5.7.5 Implementation

A proof of concept de-identification portal was implemented based on the above design. A cloud virtual machine was used to host the data portal, an AES 256-encrypted zip file to protect the dataset, Python script files for the researcher to express the proposed analysis, an HTML web interface for the data custodian to authorise a script by providing the decryption key, and a background process to run the analysis in the cloud and store the result.

Table 5.7: Heuristic Evaluation against ISO Usability Characteristics

ISO Usability Characteristic	ISO Definition	Spreadsheet	Ours
Appropriateness recognizability	Degree to which users can recognise whether a product or system is appropriate for their needs.	Spreadsheet editors are an obvious choice for data custodian to use to remove/substitute participant identifier columns, but data custodian may not be aware of need to also remove quasi-identifiers.	Researcher determines appropriate de-identification methods and sends link and instructions to data custodian.
Learnability	Degree to which a product or system can be used by specified users to achieve specified goals of learning to use the product or system with effectiveness, efficiency, freedom from risk and satisfaction in a specified context of use.	Spreadsheets provide a familiar and intuitive interface. However, without proper training, there is a risk of data errors due to incorrect formulas that refer to the wrong cells.	Researcher must have sufficient training to express de-identification operations.
Operability	Degree to which a product or system has attributes that make it easy to operate and control.	Spreadsheets provide an intuitive interface. However, may be slow and repetitive if need to manually apply the same operation to many worksheets.	The data custodian only needs to provide encryption/decryption password.
User error protection	Degree to which a system protects users against making errors.	Spreadsheet software has no intrinsic functionality for recognising identifiable data. Responsibility falls on data custodian.	Researcher can test their code on a sample data set. Data custodian's role is reduced to choosing an appropriate encryption password.
User interface aesthetics	Degree to which a user interface enables pleasing and satisfying interaction for the user.	Spreadsheet provides a familiar and intuitive interface.	Interface can be themed.
Accessibility	Degree to which a product or system can be used by people with the widest range of characteristics and capabilities to achieve a specified goal in a specified context of use.	Spreadsheets in default mode present issues for users with low vision.	Interface conforms to W3C Web Content Accessibility Guidelines.

5.7.6 Heuristic usability evaluation

Table 5.7 shows results of a heuristic evaluation of our system according to the usability characteristics defined in the ISO software quality framework [109]. In contrast to prescriptive usability heuristic checklists such as those presented by Nielsen [154], the focus of this section is on evaluating usability concerns stemming from the software architecture [75] rather than on minor usability issues that are implementation specific.

5.7.7 Case study

This section shares preliminary experience using the system to obtain de-identified player position tracking data from a sport club for team strategy analysis.

To deal with the unique privacy issues associated with human trajectory data, a custom de-identification operation was selected that combined downsampling the position data to 1 Hz with a randomly sorted point cloud representation to increase uncertainty of player identities whenever two player paths crossed each other [59]. As the custom de-identification operation was too involved for the sport club to perform themselves, the sport club was asked to encrypt and upload the raw²¹ player position tracking data to the de-identification portal (*de-identify.org*) designed and implemented as part of this thesis.

Implementing the analysis script to extract and de-identify data proved to be challenging without having a way to peek at the structure of the underlying raw data it was operating on. While theoretically this could be addressed through metadata or sample data, in this situation metadata was not available and the format of the sample data differed from the actual data. Thus, multiple iterations were necessary to infer the data structure and to address parsing related issues.

²¹The sport club took basic steps to strip out identifying information prior to uploading the data, such as replacing player names with anonymous identifiers, but this in itself would have been insufficient to prevent re-identification attacks.

While ultimately successful, the overall process took one month as each iteration had to wait for manual authorisation by the sport club who acted as the data custodian.

5.7.8 Conclusions

Our tool and method are simpler for the data custodian at some additional burden to the researcher when compared against using a spreadsheet tool such as Microsoft Excel, the most commonly used tool for this operation currently. Specifically, our approach calls for the researcher to be able to express de-identification via an automated script. This is a superior approach as the researcher would typically have additional skills and training to handle data compared to the data custodian.

While the case study examined sport club data, the approach is, in principle, generalisable to other domains in which the data custodian lacks the technical resources to de-identify the data themselves. Future work is needed to validate the emotional goal model through interviews with stakeholders, and to run an empirical trial to quantify the extent to which the system satisfies the emotional goals of the data custodian and researcher.

Future implementations could benefit from functionality to automatically reverse-engineer the structure and semantics of a dataset without revealing individuals in the dataset; this would reduce the number of iterations required for the researcher to understand the dataset and arrive at the final analysis, thus reducing risk of feelings of irritation and frustration from the data custodian and researcher.

5.8 Chapter Summary

This chapter demonstrated that de-identification of player tracking data is non-trivial to perform correctly, and that common de-identification approaches can lead to individual players being re-identified. Through comparing the trade-off between participant privacy and data quality, a combination of downsampling and representation as a point cloud were found to offer the best theoretical protection of individual privacy while still allowing team-based analysis.

Limitations

The de-identification approach has been selected for AFL where players can move freely on the field, but may not work well for other types of sport where players have set roles. Specifically, the de-identification scheme relies on players crossing paths in order to make it harder for an attacker to re-identify players; however, if a player maintains the same position (e.g. a cricket player that always fields in the same position) or rarely comes into close contact with other players, then the attacker could use this information to re-identify the player. Furthermore, by design, the de-identification only allows team-level analysis rather than individual player analysis. This means that the proposed approach is not appropriate for traditional forms of sport performance analysis that focus on analysing individual players, or specific player roles (e.g. in netball). An alternative de-identification approach would be needed for these forms of sport analysis. Nevertheless, the interaction model proposed is sufficiently general that the overall framework is still appropriate to these scenarios if a suitable de-identification operation can be found (e.g. the interaction model could be used to request summary data broken down by player role, but aggregated over multiple seasons in order to prevent re-identification risk).

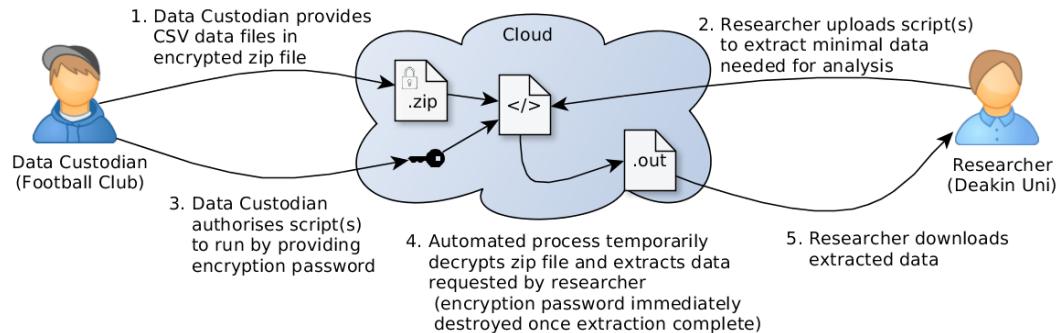


Figure 5.12: Summary of interaction model as implemented in *deidentify.org* portal

The screenshot shows the **Data Sharing Portal** interface at <https://deidentify.org/aft-gps-data/>.

Analysis 1 - Metadata and Data Sample

This analysis will reveal a sample of each data file to help us understand the data structure, without revealing the entire dataset.

- A sample of the first 10 rows and last 10 rows will be extracted for each file.
- A summary of values in each column will be extracted (e.g. frequency of each category, mean, standard deviation).

Encryption Password: **Run Analysis**

Analysis 2 - Formation Profiles

This analysis will extract team formations in a way that allows studying the team, but not individuals.

- Data will be resampled to 1Hz (1 sample per second) to reduce risk of revealing detailed player movements detected by accelerometers.
- Team formations will be represented as a "point cloud" that includes non-identifiable dots for the location of each player in the formation, but does not allow tracing the position of individual players over the full course of the match (i.e. even if one knows the player position at a certain time, they will lose track of which player is which whenever two player paths cross over each other).

(Coming soon)

Operated by Deakin Software and Technology Innovation Lab (DSTIL). Secure cloud computing provided by the National eResearch Collaboration Tools and Resources project (Nectar)

Figure 5.13: Screenshot of *deidentify.org* portal in use

Applied Output

The research in this chapter was applied to develop the proof of concept de-identification portal *deidentify.org*. This is a cloud service for data custodians to share data with researchers in a way that meets Victorian privacy legislation and Australian human research ethics guidelines for non-identifiable data. Unlike existing methods in which the data custodian substitutes names with anonymous identifiers then hands over the full ‘de-identified’ dataset to researchers—which may still lead to re-identification or permit unethical uses if the dataset contains quasi-identifiers such as times, dates, or locations—*deidentify.org* encourages researchers to write analysis programs that extract just the data they need, and gives the data custodian control over which analyses can run.

A summary of the interaction model for *deidentify.org* is shown in Fig. 5.12, which can be seen to follow that of Sec. 5.7.4. The *deidentify.org* portal was used request the de-identified sport player tracking data used in this thesis. A screenshot of *deidentify.org* in use is provided in Fig. 5.13.

Future Work

A more robust implementation of the de-identification interaction model is needed before offering the system for general use and high risk research. This would require additional software engineering effort to allow the system to scale and auditing of the source code by a professional security analyst. In particular, if providing a public server for de-identification, there is a need to ensure that processes are correctly isolated to protect the data during the short time frame in which the raw data is decrypted on the server while the de-identification algorithm runs. An alternative would be to use *homomorphic encryption*²² to avoid the need to ever decrypt the data on the sever; however, the

²²For a high-level summary of the promises and practical limitations of homomorphic encryption see Andy Greenberg, 3 Nov 2014, “Hacker Lexicon: What Is Homomorphic Encryption?”, Wired. <https://www.wired.com/2014/11/hacker-lexicon-homomorphic-encryption/>

overhead of fully homomorphic encryption is currently impractical for computationally intensive operations on large datasets.

Additional work is also needed to test the usability of the system. The proposed system provides the data custodian the power to decide which analyses the researcher runs. However, currently, this requires a certain level of trust in the researcher to correctly state what each of their analysis scripts will extract. One option would be for the system to show the data custodian the source code of the analysis script that will be run as part of the approval process; however, as the system is designed to empower data custodians without programming expertise to safely share data, this is unlikely to provide any meaningful security if the data custodian does not understand the researcher's code. Thus a mechanism is needed to automatically audit the analysis script to confirm that it conforms to the human readable summary of what it will extract and to automatically test whether the analysis output leaks potentially sensitive information (e.g. by automatically running a series of re-identification attacks to test for common de-identification mistakes).

Contributions

1. Exposed the prevalence of improper de-identification methods used in sport research, and demonstrated that GPS player tracking data is particularly prone to re-identification. An interaction model was proposed to help improve ethical conduct of research by allowing the researcher to specify the de-identification operations in cases where the data custodian lacks the technical resources to strongly de-identify data themselves prior to data hand-over. The proposed approach was applied to GPS player tracking data held by an AFL club to obtain the non-identifiable data used in this thesis.

Chapter 6

Spatio-Temporal Reference Frames

Contents

6.1 Introduction	187
6.1.1 Philosophical Lens	189
6.1.2 Motivation	190
6.2 Background	192
6.2.1 Converting Geographic Coordinates to a Local Reference Frame	192
6.2.2 Projecting Geographic Coordinates to a Local Pro- jection	194
6.2.3 Support for Coordinate Transformations within Geographic Information Systems	196
6.2.4 Support for Coordinate Transformations within Sport Player Tracking Software	201
6.3 Spatio-Temporal Reference Frames as Geographic Objects	204
6.3.1 Abstract	205
6.3.2 Introduction	205
6.3.3 State of the Art	208
6.3.4 Related Work	209

6.3.5 Proposed Solution	210
6.3.6 Implementation	214
6.3.7 Evaluation	214
6.3.8 Conclusions	216
6.4 Chapter Summary	216

The previous chapter explored means to strongly de-identify the dataset prior to any further processing. The result of this phase was a point cloud of (latitude, longitude, time) points. However, latitudes and longitudes are meaningless to sport performance analysts as is—they must first be reprojected from world coordinates into the local coordinate system of the field. Ensuring that the reprojected player tracking data are consistently oriented allows performance analysts to make comparisons between matches played on different fields. This also facilitates pooling data from multiple matches within a unified reference system to allow analysing patterns from an entire season of data rather than being limited to analysing position tracking data from each field separately (which would otherwise severely limit the ability to cohesively analyse data from away matches, as away matches take place on different fields throughout the season).

This chapter begins by discussing the nuances that must be accounted for to ensure that this transformation is cartographically correct. It contributes a novel approach for representing coordinate systems in a way that is accessible to sport performance analysts who do not necessarily have training in Geographic Information Systems (GIS). The proposed approach could help prevent common geospatial processing mistakes made when handling GPS data. Furthermore, the proposed approach involves less user interaction than alternative methods, thus making it better suited to marking up a large number of fields (e.g. every field played on over the course of a season) than manual alternatives.

6.1 Introduction

Today, consumer Global Navigation Satellite System (GNSS) devices in Australia obtain a 5-10 metre accuracy. While precise positioning GNSS techniques exists to provide real-time 2 cm accuracy, e.g. for tractor auto-steer applications, these techniques usually require installation of a nearby base station to broadcast correction data, and costs are

prohibitive to ordinary consumers.¹ However, with the launch of new satellite positioning systems, augmented with correction data from a network of GNSS ground stations, Geoscience Australia estimates that all Australians will soon have access to real-time 3 cm accurate GNSS tracking data within the near future.²

With the widespread availability of the GNSS technologies, modern data sets are often created and stored directly using a global coordinate system such as WGS84. For example, *Open Street Map*³ allows users to upload GNSS trajectory traces recorded using devices such as car satellite navigation systems and mobile phones, then trace over these trajectories to create a unified detailed street map of the entire Earth.

Large efforts exist to digitise historic maps, and align them with modern world-wide maps. The program *QUAD-G*⁴ automatically extracts geographic coordinates from text written in the margin of historic maps in order to align them, although obviously this approach is limited to cases when the original mapmaker provides these coordinates, and had access to the technology to determine them with reasonable accuracy for the scale that they were working at. *NYPL Map Warper* [213, 119] is a project by the New York Public Library to scan historic maps, then crowd-source the work of rectifying (aligning) the maps out to citizens by providing them with a side-by-side web interface to match points on the historic map to coordinates on a modern Open Street Map reference. *Georeferencer* [73] provides similar functionality to *NYPL Map Warper*, and is used by several libraries. *Esri ArcMap* provides functionality to automatically align two raster maps, although is limited to cases where the images are very similar, and does not work well with historic

¹The Allen Consulting Group, “Economic benefits of high resolution positioning services,” Nov. 2008.

²Geoscience Australia, “National Positioning Infrastructure Capability”.
<http://www.ga.gov.au/scientific-topics/positioning-navigation/positioning-for-the-future/national-positioning-infrastructure> Accessed: 2017-03-13

³<http://www.openstreetmap.org/> Accessed: 2017-03-21

⁴<http://geography.wisc.edu/research/projects/QUAD-G/> Accessed: 2017-03-15

maps⁵. Chen et al. [34] attempt to automatically match road maps with satellite imagery by first extracting topological features, then uses a conflation algorithm designed to permit matching the road topologies, even when they are represented at differing levels of detail. *MapAnalyst* [114] allows visualising and assessing distortion profiles of historic maps, including non-linear abstract maps such as schematic network maps often used to represent public transport networks, but first requires the user to manually match control points on the historic map to their true coordinates.

6.1.1 Philosophical Lens

Despite the current trend towards all data becoming integratable as part of a single global map, in practice, users usually care about positions relative to a local reference frame, not about absolute geographic location. For example, a building inspector who identifies a defect in a supporting beam, will need to determine the beam's location relative to the geometry of the house so that the inspector can locate the beam on an architectural diagram. If the architectural diagram is a template for multiple houses of the same type, then this precludes the possibility of the architectural diagram containing any indication of absolute geographic location, as this will differ for each house built from the diagram. Furthermore, the house may contain distortions relative to the architectural diagram, for example, a larger verandah, or non-conforming angles due to ground movements since initial construction. Thus it is important that the building inspector uses a local reference frame aligned with the house rather than absolute geographic coordinates in order to facilitate comparisons to the architectural diagram.

Reprojecting data to local reference frames is necessary to permit comparison of similar processes across different locations. This allows a successful idea to be replicated in different locations, with knowledge gained from each application transferred to the others. For example, if

⁵ESRI, “Georeferencing a raster automatically”. <http://desktop.arcgis.com/en/arcmap/latest/manage-data/raster-and-images/georeferencing-a-raster-automatically.htm> Accessed: 2017-03-15

a robotic assembly line were replicated to another factory, it would be desirable to have the ability to study the movement of objects through the assembly line in a way that permits comparison between the two factories. The simplest way to permit this comparison is to calculate positions of objects relative to a local reference point about which movements will be similar in each of the factories. Note how this simple act of reprojection makes it possible to detect patterns, that were present, yet not obviously apparent, in the globally referenced data.

Reprojecting data to local coordinate frames is not only necessary to facilitate comparison by humans, but may also assist automated pattern recognition approaches. Preprocessing data to extract relevant features is a crucial step to improve the accuracy of data mining algorithms. Automated pattern recognition approaches rely upon the selection of features as an induction bias as to which patterns are expected to exist. To reduce search space [221] and prevent over-fitting, pattern recognition approaches are biased [220] to favour simple explanations, and will penalise or not consider solutions that require complex feature combinations and heavy value shifting [61]. Thus, while certain data mining algorithms are in theory capable of finding patterns in globally referenced data even without reprojection, transforming features into a suitable reference frame to make them comparable prior to applying pattern recognition helps increase the chance of finding patterns and reduces the amount of training data needed.

6.1.2 Motivation

The mathematics of reprojecting geographic data to a local coordinate system is well known, and considered fundamental⁶ in the fields of geodesy and cartography [63, 157]. In practice, users would use a Geographic Information System (GIS) to perform the transformation rather than performing the calculations manually. Geographic Information Systems typically include a range of options for both reprojection to a

⁶Navipedia contributors, “Transformations between ECEF and ENU coordinates” 2013. http://www.navipedia.net/index.php/Transformations_between_ECEF_and_ENU_coordinates Accessed: 2017-01-31

local reference frame, as well as sophisticated cartographic projections that attempt to preserve either area, distance, or angle (but cannot preserve all three) over the curvature of Earth when working with maps of large areas.

However, Geographic Information Systems are usually targeted to professional users, such as surveyors and engineers, rather than novice users attempting to analyse data from their own domain. In particular, the process for defining a local coordinate reference system requires manually specifying the details using a projection specification format such as Well-Known Text (WKT) or PROJ.4. This can be an intimidating process for novice users, who may not be familiar with the intricacies of projection systems. Even for experts, creating reference frames can be a tedious process, especially for large areas in which the user needs to carefully consider the distortion introduced by the curvature of Earth. The premise of this chapter is that the difficulty associated with defining coordinate systems has restricted the use of coordinate frame transformations in GIS to that of a data import/export operation, leaving their full potential as a powerful technique for comparing local features across similar spaces left largely unrealised.

This chapter proposes a novel system for representing and manipulating the reference frames as an annotation overlay on the map rather than as manually entered projection parameters. In the proposed approach the process of selecting local reference frames may be automated entirely under certain circumstances by automatically detecting a suitable nearby geometry to use as the reference frame. This can reduce both the number of steps involved, as well as the complexity of the task. The relationships between objects, their trajectories, and local reference frames is essential semantic information to permit comparisons, despite poor support for this in existing Geographic Information Systems.

The rest of this chapter begins with a review of the transformation steps involved to project data from geographic coordinates to local coordinate space, and beyond into abstract space. The chapter highlights limitations of existing Geographic Information Systems for performing

this process, and highlights pain points that a novice user is likely to encounter. It then describes the proposed system for simplifying the transformation process, and explains how the proposed system adds semantic value to the data that can also benefit pattern mining. This tool was applied to reproject the AFL team GPS tracking data analysed in the next chapter.

6.2 Background

6.2.1 Converting Geographic Coordinates to a Local Reference Frame

Reference frame conversion is a simple mathematical technique that uses matrix transformations to freely convert between different coordinate systems defined by reference frames. In the context of geography they can be used to convert from a world coordinate system to a localised plane tangent to the Earth at a point of interest. When working with small areas, it is reasonable to approximate the surface of the Earth to a flat plane at the point of interest for all subsequent analysis. Note that when interpreting land based data over large areas (e.g. countries), the curvature of Earth becomes significant, and it becomes more appropriate to use a cartographic projection with projection parameters chosen to trade-off preservation of distance, angles, and areas as best possible in the locality of interest with respect to the curved surface of the Earth. See Fig. 6.1 for a visual explanation.

It is first necessary to convert geodetic latitude and longitude angles into three-dimensional Cartesian space. The equations to convert a WGS84 location represented as longitude λ , latitude ϕ , and height h to Earth Centred Earth Fixed (ECEF) x, y, z coordinates relative to the centre of Earth are taken from [63] and presented in Eq. 6.2.1.

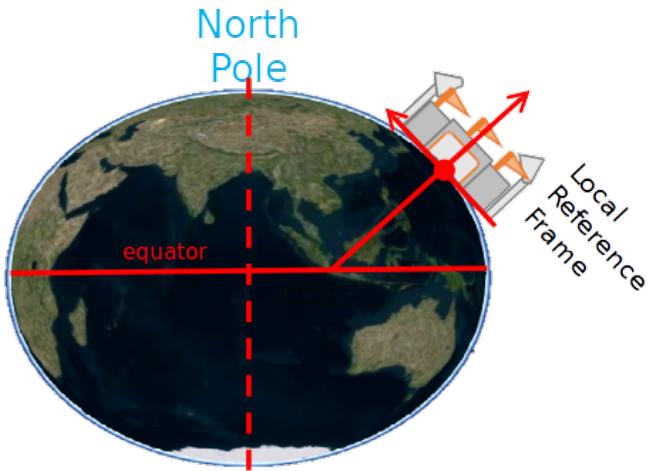


Figure 6.1: Example of setting up a local reference frame. World coordinates are converted to vectors relative to the Cartesian coordinate system centred at the stadium in the figure. Note that the up axis does not exactly intersect the centroid of Earth—this is due to a subtlety in the definition of geodetic latitude of an ellipsoid Earth to be tangent to the ellipsoid (which is also the approximate direction of gravity, ignoring localised Deflection Of Vertical) rather than radiated from the centroid.

$$\left. \begin{aligned} x &= \left(\frac{a}{\chi} + h \right) \cos \phi \cos \lambda \\ y &= \left(\frac{a}{\chi} + h \right) \cos \phi \sin \lambda \\ z &= \left(\frac{a(1 - e^2)}{\chi} + h \right) \sin \phi \end{aligned} \right\} \quad \begin{array}{l} \text{Absolute x, y, z location} \\ \text{relative to centre of Earth.} \end{array} \quad (6.2.1)$$

where

x is in the direction of the Prime meridian.

z is in the direction of the North pole.

y is in the direction normal to the x , z plane
(using right-hand rule).

$e = \sqrt{1 - \frac{b^2}{a^2}}$ is the eccentricity.

$a = 6\,378\,137\text{ m}$ is the WGS 84 semi-major axis.

$b = a(1 - f)$ is the WGS 84 semi-minor axis.

$f = 1/298.257223563$ is the WGS 84 flattening.

$$\chi = \sqrt{1 - e^2 \sin^2 \phi}$$

The equations to convert a WGS84 point (x , y , z) from ECEF coordinates into local East North Up (ENU) coordinates (de , dn , du) with respect to a local fixed reference frame at λ_{ref} , ϕ_{ref} , h_{ref} are taken from [63] and presented in Eq. 6.2.2.

$$\begin{bmatrix} de \\ dn \\ du \end{bmatrix} = \begin{bmatrix} -\sin \lambda_{ref} & \cos \lambda_{ref} & 0 \\ -\sin \phi_{ref} \cos \lambda_{ref} & -\sin \phi_{ref} \sin \lambda_{ref} & \cos \phi_{ref} \\ \cos \phi_{ref} \cos \lambda_{ref} & \cos \phi_{ref} \sin \lambda_{ref} & \sin \phi_{ref} \end{bmatrix} \begin{bmatrix} x - x_{ref} \\ y - y_{ref} \\ z - z_{ref} \end{bmatrix} \quad (6.2.2)$$

where x_{ref} , y_{ref} , z_{ref} are the x , y , z coordinates of the origin of the reference frame calculated using Eq. 6.2.1.

The above formulas do not include a term to correct for azimuth, α_{ref} , in the case of a rotated reference frame. However, this can easily be included by pre-multiplying by yet another rotation matrix to rotate about the up axis.

6.2.2 Projecting Geographic Coordinates to a Local Projection

Map makers have devised two-dimensional projections of the Earth that can preserve distance to a point, angles, or area, but it is impossible to have all three at the same time. While computerised maps can use three-dimensional coordinates to perform exact geodesic calculations over the curvature of Earth, these are computationally intensive, and furthermore, they limit spatial analysis techniques to those for which a geodesic algorithm is available. For mapping within a small region, calculations can be approximated within the two-dimensional projection, but appropriate projection parameters must be chosen for that region. Note that unlike local reference frames described in the previous section, cartographic projections are designed to preserve properties along the curvature of Earth, so are appropriate for much larger regions (although for projections of the entire Earth, severe distortion is still inevitable).

In his classic work, *Map projections: A working manual* [196], John P. Snyder sets out guidelines for selecting the appropriate map projection given the location and scale of the map needed. For brevity, this section will focus on Hotine's Oblique Mercator projection, a conformal (angle preserving) projection suitable for mapping small rectangular regions, narrow with respect to the curvature of Earth.

Hotine's Oblique Mercator (also known as Rectified Skew Orthomorphic) attempts to preserve scale along a central line chosen by the map maker. Hotine's Oblique Mercator is also conformal (no local angular distortion) away from central line. To achieve this, Hotine's Oblique Mercator projection must compromise by distorting area away from the central line. Note that while the projection preserves local angles (i.e. relative angles within any infinitesimal portion of the map), global angles (such as azimuth) are still distorted.

Hotine's Oblique Mercator projection can be parameterised to the region of interest. Much like the information required to set up a local reference frame in the previous section, the information needed to parameterise Hotine's Oblique Mercator projection is a reference location, and a reference angle. Fig. 6.2 provides an example of Hotine's Oblique Mercator centred around Melbourne, in an eastward direction towards the lower half of the US. This projection would be ideal for tracking the ground path of a flight from Melbourne to the US.

Whilst local projections are conceptually different to reference frames, the similar information required to build a Hotine's Oblique Mercator projection (and its alternatives) versus a local reference frame, allows projections to serve as a natural extension of Reference Frames when concerns shift from studying very small (i.e. approximately uniform direction of gravity) regions of the Earth towards larger regions where the analyst is interested in the positions of objects relative to the curvature of the Earth (i.e. a non-Euclidean coordinate space).

The formulas for a computerised implementation of Hotine's Oblique Mercator projection can be found in Hooijberg [99] and Evenden [71]. In modern times, cartographers are unlikely to ever need to ever im-

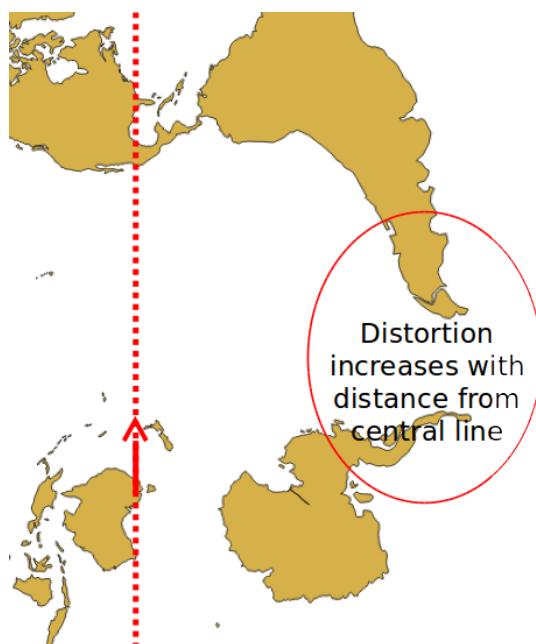


Figure 6.2: Example of Hotine’s Oblique Mercator projection taken at Melbourne in east direction. Note the severe distortion in the tail of Antarctica and South America in this projection, due to their distance from the choice of central line.

plement projection computations themselves, as computations are tedious, and can be error prone if not implemented carefully and tested thoroughly. Instead, cartographers would use a readily available implementation, such as that found in the Open Source PROJ.4 [71] library.

6.2.3 Support for Coordinate Transformations within Geographic Information Systems

Geographic Information Systems (GIS), such as QGIS⁷, provide the user with a graphical user interface to convert data between different coordinate systems. Typically, datasets will be encoded in a standardised coordinate reference system, such as one of the reference systems encoded in the EPSG database⁸. Metadata describing the projection format is usually stored alongside the dataset, and the interface to convert

⁷[Software] QGIS Project, “QGIS: A Free and Open Source Geographic Information System”. <http://www.qgis.org/> Accessed: 2017-03-16

⁸International Association of Oil & Gas Producers, “ESPG Geodetic Parameter Dataset”. <http://www.epsg.org/> Accessed: 2017-03-16

to a different standardised coordinate system is usually as simple as selecting “save as...”, then selecting the name of the coordinate system from a list. This not only hides the complexity of performing the transformation from the user, it also hides the need for a user to understand how coordinate systems are represented, so long as the user knows the identifier for the coordinate reference system they wish to work in. The Universal Transverse Mercator system can be used to generate a set of coordinate reference systems by splitting the Earth’s ellipsoid into narrow strips of 6 degrees longitude in which distortion is minimal. For example, an appropriate projection for planning a railway between Melbourne and Cairns (north bound of Melbourne) would be Map Grid Australia 55 (UTM zone 55 of the GRS80 ellipsoid⁹, officially defined as using the Geocentric Datum of Australia¹⁰) thus avoiding the need for the user to define their own coordinate system.

However, in the case of creating a coordinate reference system with respect to some object, it is necessary for the user to define their own coordinate reference system. If the user’s sole objective is simply to make the reference object (0,0) for convenience, then this can be achieved by using an existing coordinate reference system and adding a so-called “false-easting” and “false-northing” to linearly offset coordinates by some fixed amount. However, shifting the coordinate system in this manner does nothing to address the distortion properties of the projection. For example, it is incorrect to use false-easting and false-northing to shift an existing coordinate reference system to some other part of the globe, as its only effect is to make the numbers more convenient—one would still be working in an area with extreme distortion. Some GIS tools include an affine transformation tool to shift object coordinates, but similarly to the false-easting and false-northing approach, this also does nothing to address distortions due to the curvature of Earth when objects are shifted large distances. In order to create a projection with minimal distortion about a reference point, it is necessary to set the pa-

⁹spatialreference.org, “EPSG:28355 (GDA94 / MGA zone 55)”. <http://spatialreference.org/ref/epsg/28355/> Accessed: 2017-03-16

¹⁰Geoscience Australia, “Geocentric Datum of Australia (GDA)”. <https://web.archive.org/web/20170218103859/http://www.ga.gov.au/scientific-topics/positioning-navigation/geodesy/geodetic-datums/gda> Accessed: 2017-03-16

rameters of the projection such that the projection is calculated about that reference point.

Unfortunately, defining a custom coordinate reference system using existing GIS software is often a cumbersome process. In QGIS, a user must manually write a projection string, using a specification format such as Well-Known Text (WKT) or PROJ.4. An example is provided in Listing 6.1¹¹. Given that GIS tools dedicate a large area of the screen to an interactive map for selecting points, it is surprising that users must manually type or copy-paste the coordinates to define the centre-point of the coordinate reference system. Furthermore, unless working with a projection that preserves azimuth angles (such as the Mercator Projection), care must be taken to ensure that the azimuth is measured with respect to north, rather than the vertical direction in the current map projection, which may be distorted.

```
1 +proj=omerc
2 +lat_0=-37.8200
3 +lonc=144.9835
4 +alpha=86
5 +k=1
6 +x_0=0
7 +y_0=0
8 +gamma=0
9 +ellps=WGS84
10 +towgs84=0,0,0,0,0,0,0
11 +units=m +no_defs
```

Listing 6.1: Above: Example of a PROJ.4 expression for a Hotine's Oblique Mercator projection centred at the Melbourne Cricket Ground in the direction of the goal posts at N86°E.

NovAtel GrafNav¹² provides a streamlined interface for creating local coordinate systems, shown in Fig. 6.3. The interface only requires the essential information from the user: reference latitude, reference lon-

¹¹The PROJ.4 template for setting up a Hotine's Oblique Mercator (omerc) projection relative to a reference point and reference angle comes from AndreJ's answer to "Using customized Coordinate System in ArcGIS Desktop?" on the Geographic Information Systems Stack Exchange <http://gis.stackexchange.com/questions/83861/using-customized-coordinate-system-for-archaeological-site-data/83884#83884> Accessed: 2017-02-03

¹²NovAtel, "Creating a Local Cartesian Plane," Waypoint FAQs. <http://www.novatel.com/products/software/waypoint-faqs/> Accessed: 2017-03-16

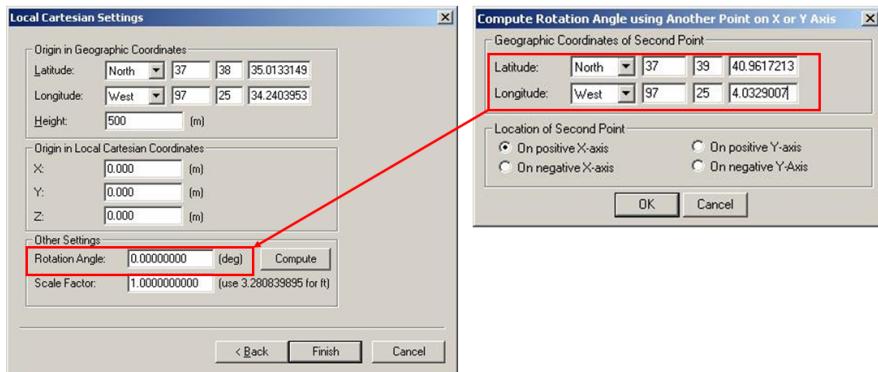


Figure 6.3: NovAtel GrafNav provides an interface to convert GPS data in world coordinates to a local coordinate system. However, configuring the local coordinate system requires manually specifying coordinates. Screenshots © NovAtel Inc. taken from software documentation.

gitude, reference height, and reference rotation angle. Alternatively, the interface can infer the rotation angle of the reference system if the user provides a second point along an axis. Once the reference system has been established, the user is free to bulk-reproject data points to their reference system. The system only provides the option to project to Cartesian coordinates in a local reference frame, and does not appear to offer the ability to use the reference system as a basis for a cartographic projection. Similarly to the issues described relating to the user interface for defining reference frames in GIS software, the interface appears to require the user to manually enter the coordinates, thus would be tedious if the user wished to set up multiple local reference frames.

Projection Wizard [178] implements the guidelines by John P. Snyder [196] to automatically suggest the most appropriate projection based upon a bounding box drawn by the user on an interactive map of the Earth (see Fig. 6.4). Unfortunately, it doesn't currently allow the user to draw an oblique (rotated) bounding box, thus is limited to projections where north is at the top of the map. While Projection Wizard doesn't implement any functionality to reproject data automatically, it can generate a PROJ.4 projection string. The user could then copy-paste this projection string into their GIS tool to define a coordinate reference system, or use it to bulk reproject data using the PROJ.4 Unix command line tool.

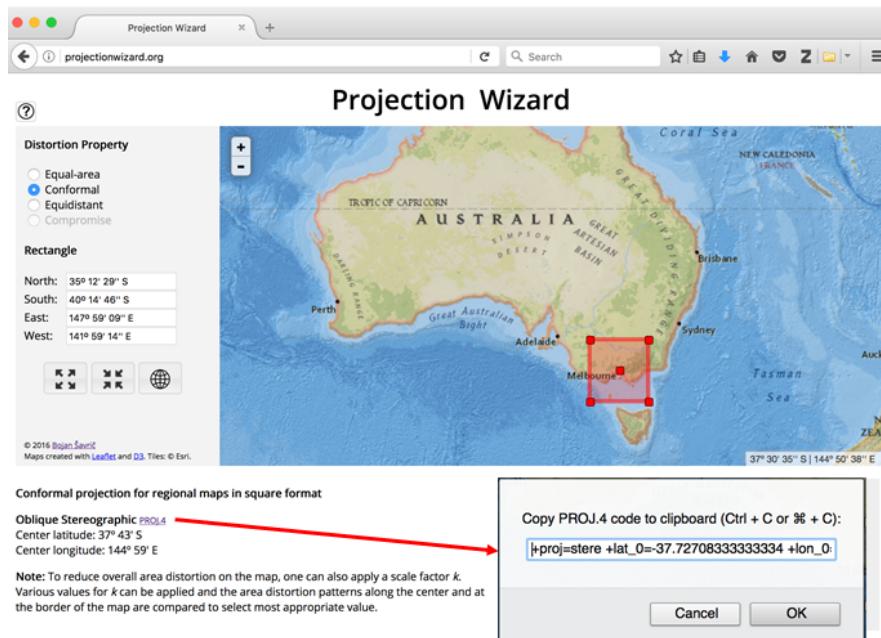


Figure 6.4: Projection Wizard [178] allows the user to visually draw a rectangle, and will automatically generate an appropriate PROJ.4 projection string for that area. However, it is not possible for the user to draw a rotated region, thus support for generating oblique projections is limited.

Tool	Obtain Reference Point	Obtain Reference Azimuth Angle	Generate Local Coordinate System	Bulk Reproject GPS Data
QGIS	✓	~ (need to use conformal projection)	~ (manually write PROJ.4 string)	✓
NovAtel GrafNav	~ (need to manually enter)	✓ (using 2 points)	✓ (only choice is ENU frame)	✓
Projection Wizard	✓	X	✓ (automatically suggest best projection)	X

Figure 6.5: Comparison of geospatial systems that provide ability to convert GPS world coordinates to local coordinates.

Key: ✓=supported, X=not supported, ~=partial support.

A comparison of systems discussed in this section for converting GPS trajectory data to local coordinates (e.g. local x-y coordinates of the relevant sport field) is provided in Fig. 6.5. While systems (or combinations of systems) exist to perform the conversion, none offer an streamlined end-to-end process for performing the conversion in a manner that is suited to non-expert users (e.g. sport performance analysts as opposed to professional GIS analysts).

6.2.4 Support for Coordinate Transformations within Sport Player Tracking Software



Figure 6.6: SPT GameTraka¹³ is capable of generating a heat-map of player location overlaid on Google Maps.

The previous sections motivated the need to transform data to a local coordinate system, highlighted the theoretical issues surrounding performing this conversion correctly, and showed that general purpose systems for performing this conversion are not well suited for non-expert users and are tedious when conversions to many different local systems are necessary (i.e. pairing GPS trajectories over the course of the season with the relevant sport field). For completeness, a selection of

¹³<http://demo.gametraka.com> Accessed: 2017

commercial software systems for analysing sport player position tracking data were examined.

SPT GameTraka (Fig. 6.6) generates a heat-map of player location overlaid on Google Maps. Firstly, note that all maps are north oriented, which makes it difficult to compare patterns on a field that is N-S oriented to a field that is E-W oriented. Further note that the software has no internal model of the field, thus it is not possible to quantitatively break down data by zone of field. Instead *SPT GameTraka* focuses on statistics such as distance, speed, intensity, etc. that can be computed without reference to the field.

GPSports Team AMS (Fig. 6.7) converts GPS data to x-y coordinates to allow quantitative analysis of position. However, the origin and orientation of the x-y coordinate system appear to be arbitrary, thus it is difficult to compare x-y values between different fields.

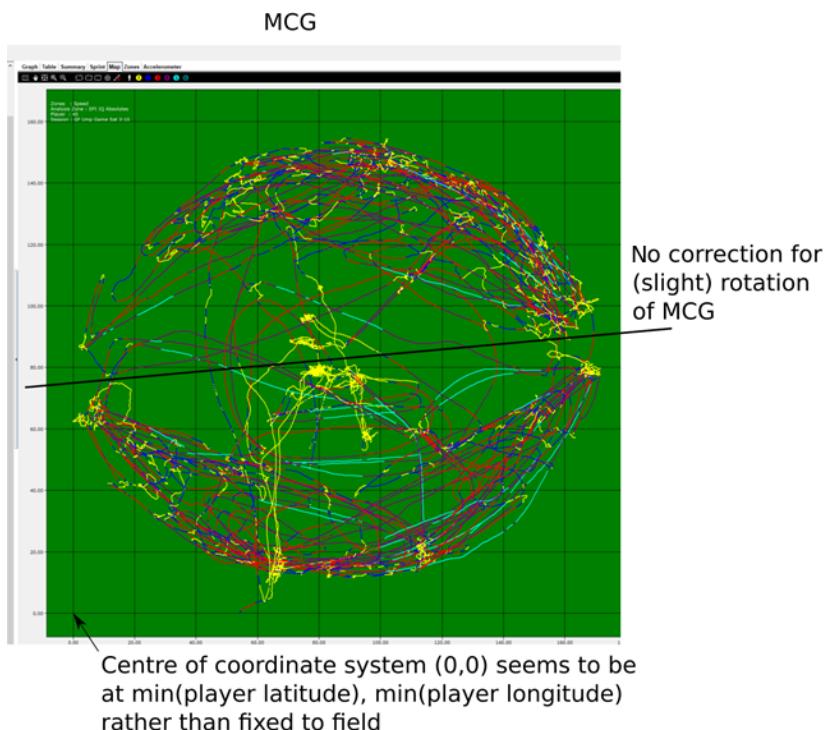


Figure 6.7: Analysis of GPS tracking data using Team AMS (software provided with GPSports GPS tracking devices). Note arbitrary origin and orientation of x-y coordinate system. This makes spatial comparisons of tracking data from different venues challenging.

The software provided with the *Catapult Sports ClearSky* local positioning system works in x-y positions relative to the playing field (Fig. 6.8). Due to the limited publicly available documentation, it is unclear whether their software provides a way to perform similar analysis using GPS tracking devices as opposed to local positioning devices.¹⁴

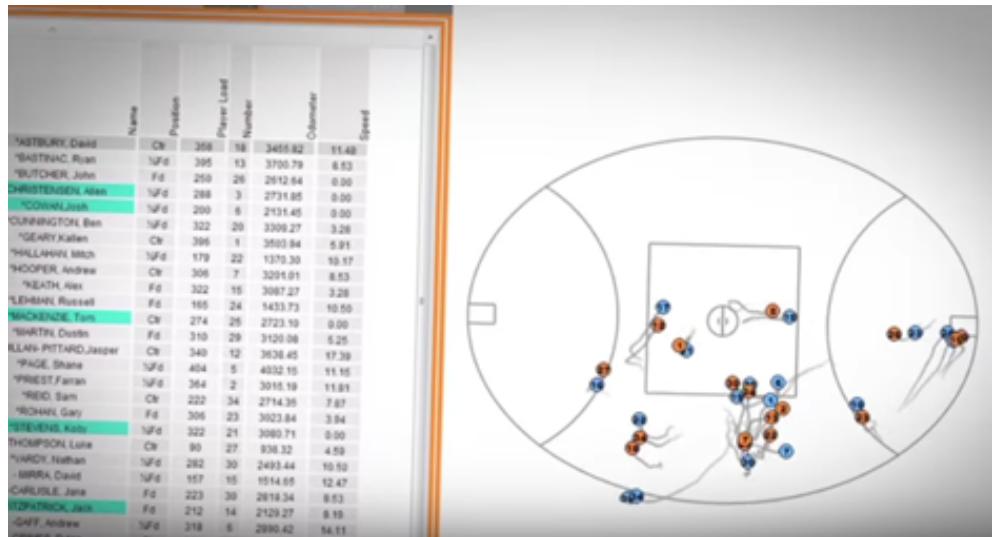


Figure 6.8: Catapult Sports ClearSky local positioning system (from promotional video¹⁵). Local positioning systems use beacons placed around the field to capture player positions in x-y form. However the system is many times more expensive than GPS tracking devices, and in the past Catapult Sports have charged \$100,000 per year per club.¹⁶

¹⁴One retailer of Catapult Sports OptimEye GPS tracking devices states they provide the ability to “build your own field with the OptimEye units”; however, it is unclear what this entails. <http://performbetter.co.uk/product/catapult-optimeye-x4-athlete-monitoring-system/>. Accessed 2019-06-08

¹⁵[Video] Catapult Sports, 2013, “Catapult ClearSky.” <https://youtu.be/lVKgPk1S7L0?t=1m41s>

¹⁶“Catapult Sports is ‘Google analytics for athletes’: wearable technology start-up series” <https://web.archive.org/web/20160901160002/http://www.catapulstsports.com/media/catapult-sports-is-google-analytics-for-athletes-wearable-technology-start-up-series/>

As highlighted, converting GPS tracking data to the local coordinate system of the field is a necessary pre-requisite for comparing GPS tracking data across different fields. However, this functionality appears to be poorly supported by existing software, other than by software provided with local positioning systems which require physical beacons placed around the field. Types of sport that involve travelling large distances where it is necessary to account for the curvature of Earth, e.g. long distance boat races, are most at risk if the geospatial transformations are not properly understood.

6.3 Spatio-Temporal Reference Frames as Geographic Objects

This section describes a tool to help sport performance analysts compare GPS tracking data from sport players between different games. It automates most of the transformations involved, thus only minimal input is required from the user. This means that the tool is suitable for use directly by domain experts (in this case, sport performance analysts) without the need for them to outsource the work to a Geographic Information System specialist.

The tool is general purpose and thus applicable to a range of domains (e.g. animal tracking). However, for the purpose of this thesis, it was applied to reproject player GPS tracking data taken over the course of an AFL season to the coordinate system of the closest AFL field (Sec. 7.1.5). The discussion in this section assumes the data files are structured as GPS trajectories; however, the tool can also be applied to reproject de-identified team point clouds (as recommended by the De-identification chapter in Sec. 5.6).

This section of the thesis was published as Andrew Simmons and Rakesh Vasa. “Spatio-Temporal Reference Frames as Geographic Objects”. In: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL’17. Redondo Beach, CA, USA: ACM, 2017, pp. 1–4. ISBN: 9781450354905. doi: 10.1145/3139958.3139983. An authorship statement for the paper can be found in Appendix F.

6.3.1 Abstract

It is often desirable to analyse trajectory data in local coordinates relative to a reference location. Similarly, temporal data also needs to be transformed to be relative to an event. Together, temporal and spatial contextualisation permits comparative analysis of similar trajectories taken across multiple reference locations. To the GIS professional, the procedures to establish a reference frame at a location and reproject the data into local coordinates are well known, albeit tedious. However, GIS tools are now often used by subject matter experts who may not have the deep knowledge of coordinate frames and projections required to use these techniques effectively.

This section of the thesis introduces a novel method for representing spatio-temporal reference frames using ordinary geographic objects available in GIS tools. Our method both reduces the number of manual steps required to reproject data to a local reference frame, in addition to reducing the number of concepts a novice user would need to learn.

6.3.2 Introduction

With the widespread availability of GNSS technologies, modern data sets are often created and stored directly using a global coordinate system such as WGS84. Furthermore, collecting and analysing spatial data is moving beyond surveyors and cartographers to subject matter experts in their own field, from traffic engineers to sport performance

analysts. However, in many applications, it is position relative to a frame of reference that matters, and not absolute geographic location. Analysing data for such applications requires the data to be reprojected from the global coordinate system into a local coordinate system. Unfortunately, this task requires deep conceptual knowledge to perform, and may bar novice users from unlocking the full potential of their data.

Consider a sport performance analyst conducting spatial analysis of player GPS tracking data, who wishes to analyse spatial patterns for a team over the course of the sport season. A single game alone does not contain enough events to mine statistically significant patterns. In order to make the most of the team's dataset, the sport performance analyst will need to consolidate data from each of the sport fields that the team competes at. To achieve this, the sport performance analyst will need to reproject the raw GPS traces from games at multiple sport fields into a consistent local coordinate system. This requires the sport performance analyst to define a projection for each sport field the team played on.

While it would be possible to achieve this task using a conventional GIS system, it would require the user to manually write a projection string using a format such as Proj4 or Well Known Text (WKT) for every field that the team played on. Furthermore, expertise working with projections and reference frames is needed to ensure this task is performed correctly, as incorrectly projecting data is a common cause of subtle geospatial errors that may undermine the validity of the analysis.

This section of the thesis addresses these issues through introducing a novel method for defining reference frames using geometric line segment objects. Our system is expected to be more learnable for novice GIS users, as it reduces the complexity of defining reference frames to the more familiar task of drawing lines. The evaluation further shows that the number of manual steps required to perform a task using our approach scales well with the number of reference frames, trajectories, and events of interest, thus indicating that our approach is productive for large problems.

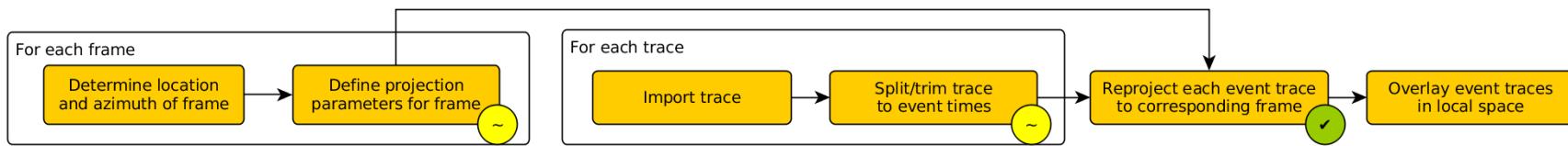


Figure 6.9: Tasks required to split traces (e.g. player GPS trajectories) that span multiple events (e.g. game segments), and reproject relative to frame (e.g. sport field) such that event traces can be compared within a common local coordinate system. Our proposed solution allow tasks marked with a green check to be fully automated, and tasks with a yellow tilde to be semi-automated.

6.3.3 State of the Art

Fig. 6.9 shows the workflow of high level tasks that a user (e.g. a sport performance analyst) would need to complete in order to transform their dataset (e.g. trajectory data for each player of a team) into a common local space so that they fully utilise all event data (e.g. making comparisons of game segments throughout the season collected from multiple sport fields). This sub-section will briefly elaborate how tasks would typically be accomplished using existing GIS tools, which will serve as the baseline for evaluating the merits and drawbacks of our proposed system.

Determining location and azimuth of frame: In order to correct for the location and orientation of each frame (e.g. sport field) the user needs to first obtain its location and orientation. For novice users, this seemingly simple step is a common cause of errors, as it is essential to ensure the azimuth is measured using the correct geodesic calculations, rather than simply the angle displayed within the current projection (which is often distorted).

Define projection parameters for frame: If frames are separated by a non-negligible distance with respect to the curvature of Earth (e.g. an interstate game), then it is necessary to work with different projections for each region. Attempting to compare geometries across large distances by simply offsetting coordinates can result in distortions as the regions may have different scales within the projection system. While databases of common projection parameters for each segment of the globe exist, achieving minimum distortion requires specifying custom projection parameters. This in turn requires the user to manually write a projection string using a format such as Proj4 or Well Known Text (WKT).

Split/trim trace to event times: It is unrealistic to start and end collection of trajectory data for the exact duration of an event of interest. Rather, data is typically logged for a short time before and after the event, or even multiple events (for example, in a sport game, data loggers will be left on for the entire game, and the data can be split into

individual round durations afterwards, discarding data for rest breaks between rounds). Manually splitting and trimming this data to the precise event durations can be a tedious process, as well as a repetitive task when there are many traces (e.g. analysing game data for every member of a sport team).

Reproject each event trace to corresponding frame: The user needs to manually match up each event trace with the relevant projection parameters to reproject it.

Overlay event traces in local space: For this step, the user needs to take the unconventional step of discarding any projection metadata so that event traces can be reinterpreted as local x,y data in a relative coordinate system.

6.3.4 Related Work

Hägerstrand visualises time as a dimension in order to conceptually reason about constraints on human mobility in space-time [95] (although it should be noted that the idea of time as a dimension was present as early as the 18th century in the works of d'Alembert and Lagrange [9, 86]). In order to permit the visualisation to be easily perceived, Hägerstrand uses a two-dimensional map, thus freeing up the third dimension to visualise the passage of time. The three axes form a space-time cube within which mobility can be analysed.

While Hägerstrand originally presented the space-time cube as a conceptual tool rather than as a practical means of visualising spatio-temporal datasets, it has inspired a number of novel visualisation systems for viewing spatio-temporal data within the time-space cube [120, 84, 8].

Jenny and Hurni [115] describe the series of cartographic transformations needed to setup a common coordinate system to compare historic maps against modern maps. Jenny and Hurni stress the importance of first projecting the modern map into the coordinate system of the old

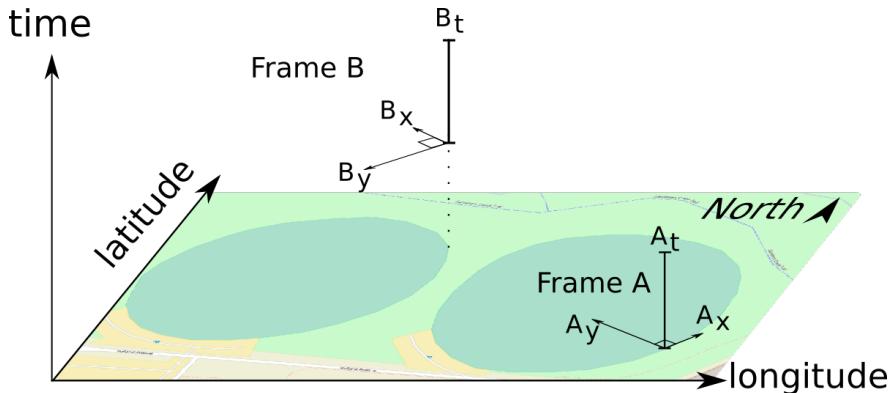


Figure 6.10: Configuration of spatio-temporal reference frames used to describe spatio-temporal structure of a sport game.

map to prevent subtle distortion artefacts that would arise from comparing data in different projections.

Šavrič et al. created Projection Wizard [178] to aid novice users with the task of selecting a projection. Their system allows the user to interactively drag a box around the region of interest, and the system will automatically suggest a projection based upon the guidelines suggested by John P. Snyder in his classic work *Map projections: A working manual* [196]. Unfortunately, their system doesn't allow the user to draw a rotated region, thus limiting its ability to suggest oblique projections.

6.3.5 Proposed Solution

Our solution utilises Hägerstrand's space-time as a conceptual framework for defining reference frames. Our reference frames are orientated in space, as well as positioned in time. Our reference frames serve as objects that users can manipulate to describe the structure of space-time to the computer. This in turn, benefits the user by allowing transformations between global and local coordinates to be automated.

Fig. 6.10 provides an example of setting up the structure of a sport game. In this example, there is an initial game event (e.g. a practice game) played in the rightmost field, and this is later followed by a game event in the leftmost field. A reference frame is added to each sport

field, and oriented towards the target goal. This make distances and movements comparable across the two fields. The time dimension of a reference frame spans an event interval (e.g. the start and end time of the round), which will serve as the basis for relative time within that frame. A more complex game would have additional frames for each round. If players change sides of the field after each round, then this could be accounted for by a series of reference frames in time that alternate between each side of the field so as to ensure that the team of interest always appears to be heading the same local direction from the perspective of the reference frame.

Representation

In conventional GIS systems, reference frames are not treated as first-class citizens like ordinary geographic objects such as polygons. Transformations in GIS systems are performed imperatively rather than declaratively: that is, they require the user to manually enter the coordinates and apply the transformation rather than providing a means to symbolically represent the transformation within the GIS system (although GIS systems do allow specifications of procedures using scripts that can be re-run later, this requires programming skills that would be additional expectation of a novice user).

For our system, in addition to the trajectory dataset, it requires a spatial file that contains geometries to describe the reference frames, with the temporal aspect of the reference frames represented using property fields of the geometry. Specifically, the spatial aspect of a spatio-temporal reference frame is represented by a line with two points. Our choice to represent spatio-temporal reference frames as ordinary geographic objects allows the user to create and modify the reference frames in the same manner as other data. Thus novice users only need to learn the basic GIS editing features, and can reuse this knowledge to create reference frames in our system without having to learn a new interface specifically for creating reference frames. Furthermore, it means that our system does not require any proprietary modifications

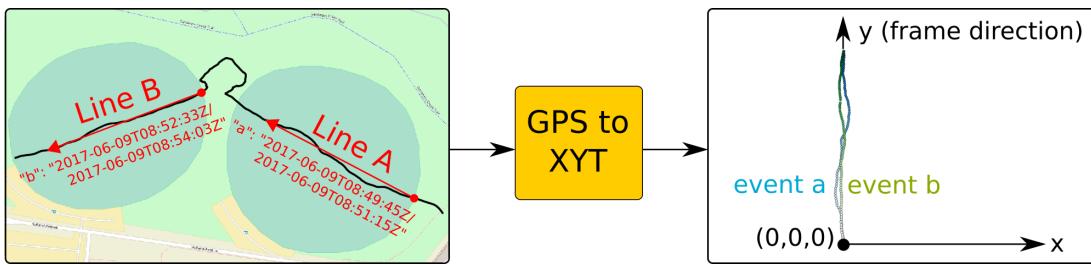


Figure 6.11: (left) Spatio-temporal reference frames are represented as the combination of a line (shown in red, direction is important) and a property attribute to represent temporal interval (small red text). These reference frames are stored in the GeoJSON format and can be created using basic editing features within a GIS tool. The user provides a directory of trajectories (black lines) in the GPX format. Our proof-of-concept *GPS to XYT* (middle box) iterates over each permutation of temporal frame and trajectory, trimming the trajectory to the duration of the event. If any data is found for that duration, it reprojects the trace relative to the corresponding reference frame the event is associated with (see Alg. 1 for details). (right) The reprojected data contains relative x,y,time points suitable for comparative analysis in an external system.

to existing GIS tools or GIS data formats beyond the geometry property fields offered by spatial formats such as GeoJSON.

The temporal aspect of a spatio-temporal reference frame is represented by an ISO 8601:2004 time interval declared as an event duration attribute associated with the frame geometry.

Conceptually, consider each spatio-temporal reference frame as distinct, consisting of a line (representing the spatial orientation of the frame) and one attribute (representing the temporal aspect of the frame). However, in complex scenarios, there can be many spatio-temporal reference frames with the same spatial dimension (e.g. a game that has multiple rounds occurring at the same location), thus for convenience, our system treats a line with multiple event duration attributes to represent a set of spatio-temporal reference frames for each event.

Algorithm 1: GPS to XYT

Input: Set of GPS traces for processing, G .
 Set of line segment geometries, L , representing spatial frames.
 Event metadata for each geometry, $Metadata : L \rightarrow E$.

Result: Projection $Z : G, L, E \rightarrow [\mathbb{R}^2]$ of each non-empty permutation of GPS-trace, spatial-frame, event-interval to local x, y coordinates at offset time t .

```

foreach trace  $g$  in  $G$  do
     $P \leftarrow \{(p_\phi, p_\lambda, p_t) \mid p \in g\}$ 
    foreach line segment  $l = ((\phi_1, \lambda_1), (\phi_2, \lambda_2))$  in  $L$  do
         $\alpha \leftarrow$  azimuth from  $(\phi_1, \lambda_1)$  toward  $(\phi_2, \lambda_2)$ 
        foreach event  $e = [e_{begin}, e_{end}]$  in  $Metadata(l)$  do
             $P' \leftarrow \{p \in P \mid e_{begin} \leq p_t \leq e_{end}\}$ 
            if  $P' = \{\}$  then
                | continue to the next event
            foreach point  $p$  in  $P'$  do
                |  $x, y \leftarrow$  project  $(p_\phi, p_\lambda)$  to oblique Mercator projection centred at  $(\phi_1, \lambda_1)$  with centreline at azimuth  $\alpha$ 
                |  $t \leftarrow p_t - t_{begin}$ 
                |  $Z(g, l, e)[t] \leftarrow (x, y)$ 

```

6.3.6 Implementation

A proof-of-concept of our proposed system is available¹⁷. Given a Geo-JSON file containing the spatio-temporal reference frames, and a directory of raw GPS traces (in GPX format), the system will automate the splitting, reprojection and temporal shifting of event traces into local coordinates. The resultant output is a directory of reprojected events, each containing x,y,t fields that respectively represent the relative location perpendicular to the reference frame, relative location in the direction of the reference frame, and relative offset time from the start of the reference event. Fig. 6.11 presents this pipeline graphically. The core of this program is a nested loop over each trace, line segment, and event, as shown in Alg. 1.

The Hotine Oblique Mercator Projection (also known as Rectified Skewed Orthographic) [196, p. 66] is used to perform the projection from global coordinates into a local x,y coordinates relative to the frame. The projection parameters are chosen to (approximately) preserve scale along a centreline passing through the reference frame origin and oriented at the same azimuth as the direction of the reference frame. This means that the reference frame conversions will be reasonable even for frames that cover very long distances in the direction of the reference frame (e.g. the trajectory of a point-to-point air trip). The actual projection calculations are carried out using the Open Source PROJ.4 library [71].

6.3.7 Evaluation

Bloom’s Revised Taxonomy of Learning

This sub-section repurposes Bloom’s revised taxonomy [7] (intended for evaluating education curricula) as a means of reasoning about the learnability of our system for novice users. Bloom’s revised taxonomy acknowledges two dimensions to the complexity of a learning task.

¹⁷Our code is made publicly available at https://github.com/anjsimmo/gps_xy

The first is the *Cognitive Process Dimension*: ‘Remember’ (low order thinking), ‘Understand’, ‘Apply’, ‘Analyse’, ‘Evaluate’, ‘Create’ (high order thinking). The second dimension is the *Knowledge Dimension*: ‘Factual’ (concrete), ‘Conceptual’, ‘Procedural’, ‘Metacognitive’ (most abstract).

GIS tools and libraries have automated the procedural aspects of projection (Apply + Procedural). However, they still require a deep conceptual knowledge of projection and coordinate systems in order to create projection frames (Create + Conceptual). Our system reduces projection from an abstract concept to geometric objects on the map, represented the same way as other concrete data that the user creates (Create + Factual). This lowering of the knowledge dimension represents a decrease in complexity and less depth of required learning.

Productivity

In addition to simplifying many of the steps articulated in Fig. 6.9, our system significantly reduces the number of steps as the number of traces scales up. Manually splitting the trajectory data would require $\mathcal{O}(F \times E \times G)$ steps where F is the number of reference frames, E is the number of event durations of interest, and G is the number of GPS trajectory traces – for a very large number of traces, the theoretical limit results not from the definition of the reference frames themselves (as these can be reused) but rather from the manual step of reprojecting each combination of frame, trace, and event. Our system still requires the user to provide some information to specify the spatial frames (in the form of a line geometry), and events (in the form of properties attached to the frames). However, our system entirely automates nearly all interaction with GPS trajectory traces (other than placing the trajectory traces in a folder for processing). Thus our system reduces the number of steps to $\mathcal{O}(F \times E + G)$.

6.3.8 Conclusions

This system solves the fundamental problem of enabling non GIS experts to define and make comparisons between local reference frames, and does so in a general way. Thus our system is applicable to a diversity of domains, e.g. to enable traffic engineers to rapidly compare local traffic conditions taken from floating car data collected at two or more identical housing estates.

Rather than treat reference frames as just an abstract mathematical concept, this section of the thesis demonstrated reference frames as geographic objects within GIS tools. This technique increases the power of GIS, as rather than limiting ourselves to viewing and manipulating low level facts, the projection systems themselves can be treated as if they were ordinary geographic objects. It was trivial to implement the program to bulk reproject data using these user constructed reference frames as input. However, with additional software engineering effort, the algorithm could be used to interactively reproject streams of data or to provide the user with immediate visual feedback on how the final reprojected data will look as they manipulate the reference frame objects. Future work is needed to investigate whether more complex geometries could be used to automatically reproject GPS data with respect to a more abstract projection (such as a schematic transit map), and whether it is possible to automatically extract reference frame objects using landmarks identified in satellite imagery.

6.4 Chapter Summary

This chapter examined the issue of converting GPS data in latitude, longitude coordinates to x, y coordinates relative to the field. A tool was proposed to increase the efficiency and correctness of this process.

The *GPS to XYT* tool designed in this thesis was motivated by the need to allow sport performance analysts to reproject player tracking data relative to the sport field, but is applicable to other domains, such as

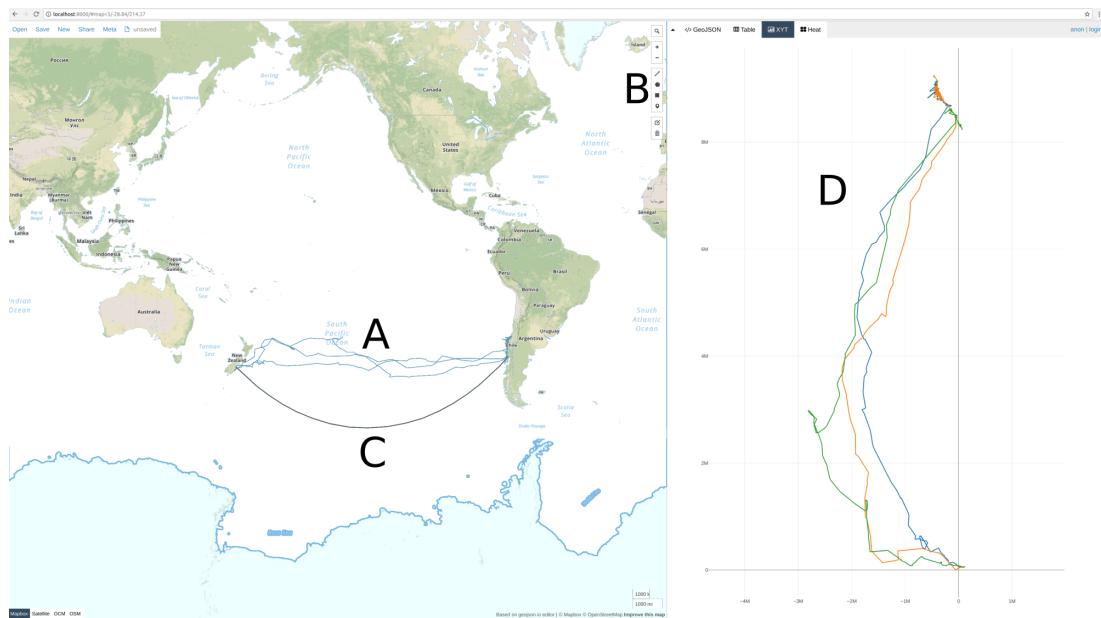


Figure 6.12: A) Migration paths of three albatrosses from New Zealand to Chile as viewed on map using Mercator projection. B) The user constructs a reference frame by using the GIS editor tools to draw a line starting at New Zealand and ending at Chile. C) The constructed line is the shortest path, which is via the Subantarctic. This appears curved on the Mercator projection. D) The *GPS to XYT* tool reprojects each of the paths taken by the albatrosses relative to the reference frame line constructed by the user. In contrast to the Mercator projection, is evident from the reprojected paths that all three albatrosses deviated from the shortest path, instead preferring to take the longer path due east.

the animal tracking example shown in Fig. 6.12. The example shows a proof of concept demonstration that integrates the *GPS to XYT* reference frame concept as an extension to the GeoJSON.io GIS editor¹⁸. The extension allows lines drawn in the editor to be used as reference frames and provides an “XYT” and “Heat” tab to facilitate analysis of the tracking data within the reprojected perspective defined by the reference frame. The albatross GPS tracking data used in this example was obtained from ZoaTrack.org¹⁹.

Typically, building applications for spatio-temporal analysis is a time consuming process involving many developers, and often involves mul-

¹⁸<http://geojson.io>

¹⁹Thomas, B, Minot, E (2016) Data from: ‘Fledging behaviour of juvenile northern royal albatrosses (*Diomedea sanfordi*): a GPS tracking study’. ZoaTrack.org. doi: <http://dx.doi.org/10.4226/68/5733FAA628046>

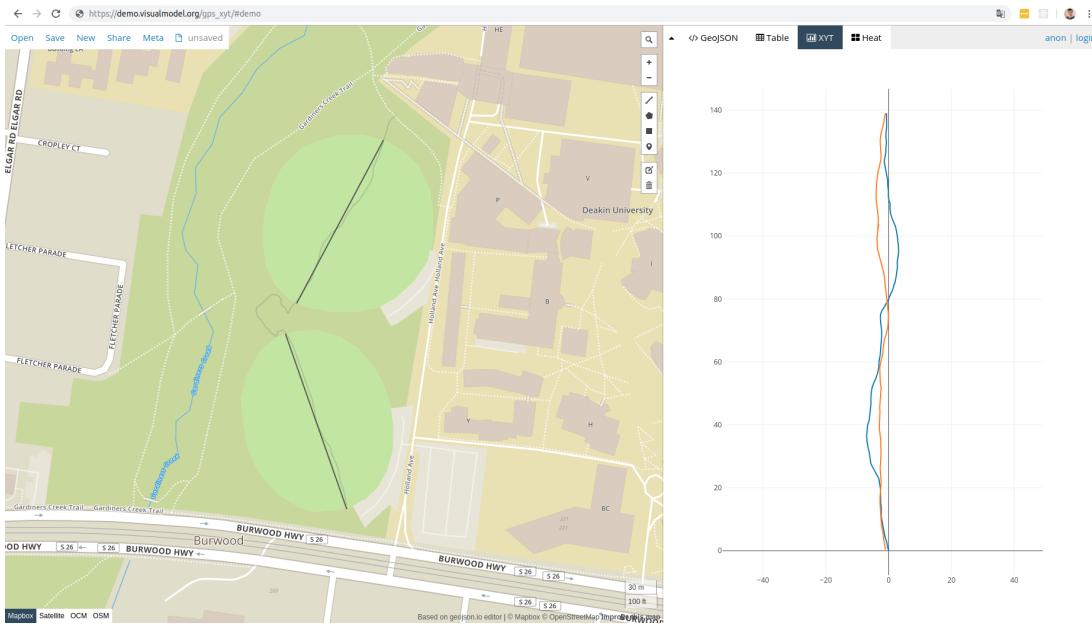


Figure 6.13: Use of the *GPS to XYT* tool to inspect a sample GPS trajectory walking across the sport fields at Deakin University. The interface makes it easy to set up coordinate systems for each of the fields by drawing lines in the editor. The trajectory is automatically split and reprojected with respect to field coordinates to facilitate comparison, as displayed on the right hand side.

tiple iterations in order to meet the client needs. Use of the proposed solution may help reduce the number of iterations needed, as well as offer the ability to design these systems in a consistent way, thus providing potential for code-reuse between different domains. The advantage of this goes beyond simply speeding up development times: cartographic transformations, if not well understood, are a frequent cause of distortions that undermine data integrity, and reduce confidence in the resulting analysis. The proposed system empowers novice users with the tools for cartographically sound spatio-temporal analysis of their domain.

Applications to Sport

Fig. 6.13 shows an example of using the interactive *GPS to XYT* tool (implemented as an extension to the to GeoJSON.io GIS editor) to inspect a sample GPS trace. The sample GPS trace was captured using a mobile

phone GPS; however, the same procedure can be applied to data from professional sport devices after converting them to a standard data format. An extended command line version of the *GPS to XYT* tool was used in Sec. 7.1.5 to reproject the sport player GPS tracking data analysed in this thesis.

Future Work

Future work is necessary to evaluate our system with sport performance analysts. While the system was theoretically shown in Sec. 6.3.7 to reduce the number of user steps required, and was argued to reduce the level of knowledge required to use the system according to Bloom's revised taxonomy, an empirical evaluation is needed to confirm whether these theoretical advantages correspond to faster use and lower error rates on real-world tasks.

Contributions

1. Proposed a novel method for representing spatio-temporal reference frames as geographic objects. This allows GIS novices, such as sport performance analysts, to configure reference frames without the need for deep conceptual knowledge of cartographic projections. It also facilitates partial automation (e.g. reprojecting GPS data to the closest sport field), thus resulting in time savings when the analysis involves multiple reference frames (e.g. a season of GPS tracking data involving multiple sport fields).

Chapter 7

A Platform for Spatio-Temporal Sport Analysis

Contents

7.1 Pipeline	221
7.1.1 Possession Chain Data	221
7.1.2 Video Data	223
7.1.3 GPS Data	224
7.1.4 Data Selection Process	225
7.1.5 Reprojection	228
7.1.6 Visualisation	231
7.1.7 Time Synchronisation	232
7.1.8 Synchronised Visualisation	234
7.1.9 Analysis Pipeline	237
7.2 Exploratory Analysis of Team Shape in AFL using GPS Tracking Data	238
7.2.1 Introduction	238
7.2.2 Related Work	238
7.2.3 Method	240
7.2.4 Exploratory Analysis	244
7.2.5 Visual Comparative Analysis	247

7.2.6 Quantitative Analysis	250
7.2.7 Discussion	254
7.2.8 Conclusions	255
7.3 Feedback from Sport Performance Analysts	256
7.4 Chapter Summary	257

The previous chapters highlighted the need for data provenance in sport (Chapter 4), systematically selected a means to de-identify the GPS data to protect individual privacy (Chapter 5), and developed a technique to semi-automate the choice of coordinate systems at each venue to facilitate comparisons across different fields (Chapter 6). This chapter integrates the techniques proposed in the previous chapters to construct a computational pipeline for AFL analysis, and demonstrates that when combined they provide meaningful team-level insights without compromising individual privacy.

7.1 Pipeline

This section describes the overall pipeline constructed to support analysis of AFL data. See Chapter 4 for details of the visual notation used to describe the pipeline.

7.1.1 Possession Chain Data

Possession chain data for Geelong Football Club's¹ matches played during the 2015 season², in the format collected by Champion Data (the official supplier to the AFL for all competition data), were obtained via the club.

¹Geelong Football Club is an elite AFL club. Note that while the name of the club is identified here (as it would not be practical to suppress it), none of the individual player data (beyond public video footage) are identifiable, even to myself.

²For details of the number of matches collected before and after filtering see Sec. 7.1.4.

Based on the findings of Chapter 5, the club was requested to completely remove the player name column (as opposed to substituting names with anonymised player codes) prior to providing the data, as this was the most secure means to prevent re-identification attacks against possession chain data outlined in Sec. 5.5.1.

This section of the pipeline is illustrated in Fig. 7.1.

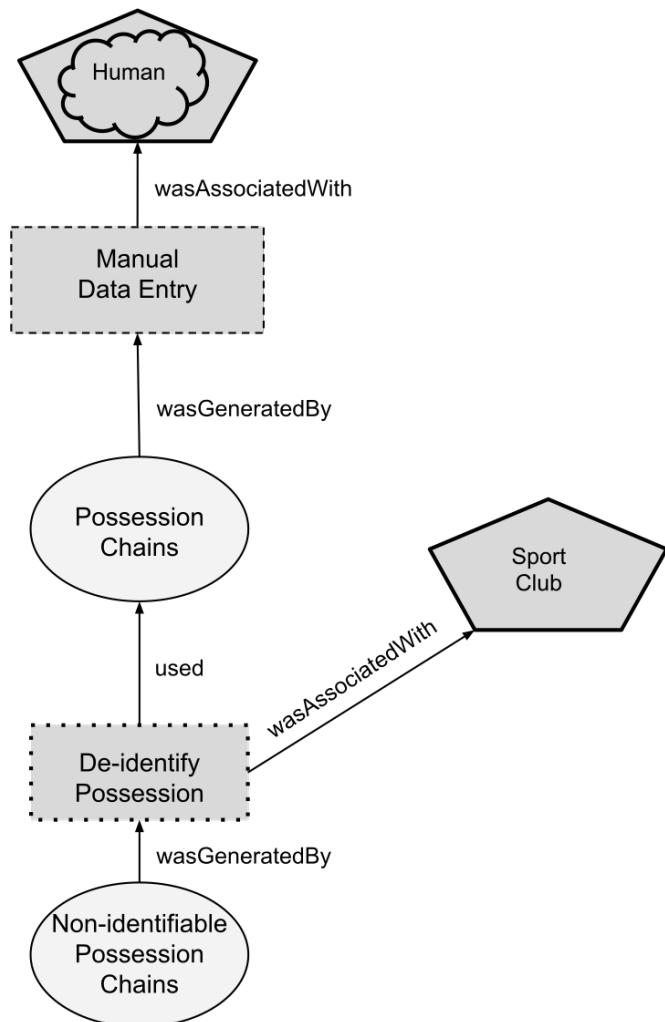


Figure 7.1: Provenance of non-identifiable possession chain data used in this thesis (see Chapter 4 for semantics of notation and Appendix Sec. B.1 for symbols)

7.1.2 Video Data

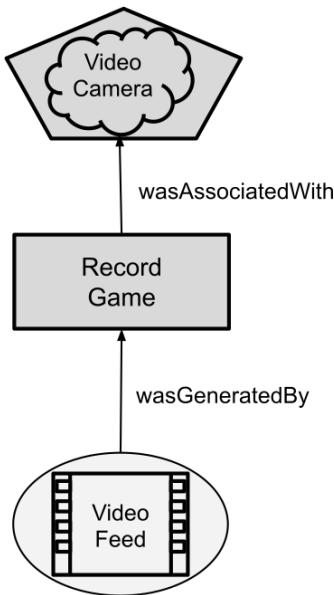


Figure 7.2: Provenance of video data used in this thesis (see Chapter 4 for semantics of notation and Appendix Sec. B.1 for symbols)

Televised video recordings of each relevant match were downloaded from the official AFL subscription service (AFL Live Pass match replay) and converted to the MP4 video format for further analysis. The distinction between the digitised video used in this thesis and the original act of recording the game (performed by the broadcaster) is shown in Fig. 7.2.

Clubs also have access to detailed behind the goals footage, which could be replaced with the televised video recordings used in this thesis if desired. However, as behind the goals footage is neither public³, nor non-identifiable⁴, this thesis used the televised video recordings to demon-

³Clubs are provided with “exclusive behind the goals vision” recordings <https://www.foxsports.com.au/afl/geelong-coach-chris-scott-explains-why-afl-coaches-bother-going-to-games-inperson-in-2016-with-video-technology/news-story/c9b99fbf9472483491294056c03bf25b>, which are considered a “game-changer” for football analysis <http://www.afl.com.au/news/2018-02-18/secret-spies-the-life-of-an-opposition-analyst>. A news report in 2013 revealed that clubs payed \$28,000/year each for the footage, with prices expected to rise to \$60,000/year per club in 2014. <https://www.theage.com.au/sport/afl/afl-doubles-tv-costs-20131025-2w7d1.html>

⁴Even if the club had the resources to blur out faces and player numbers in the behind the goals video, the position and movements of players evident within the video would still allow re-identifying particular players in the footage.

strate what is possible using only public and/or non-identifiable data.

7.1.3 GPS Data

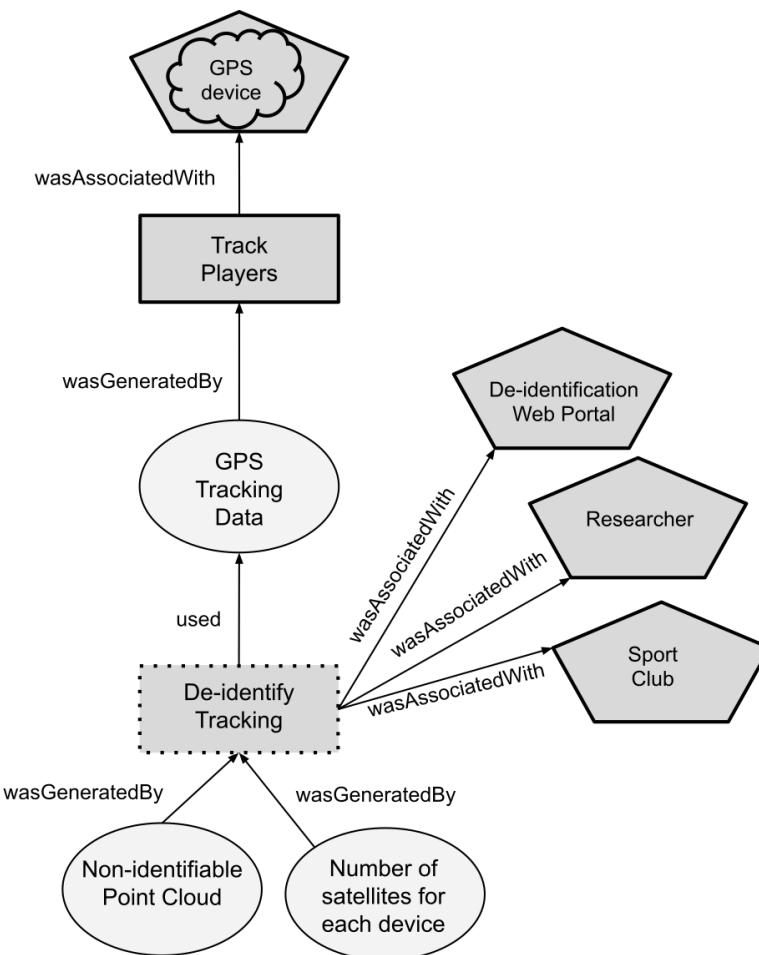


Figure 7.3: Provenance of non-identifiable point cloud data used in this thesis (see Chapter 4 for semantics of notation and Appendix Sec. B.1 for symbols)

Player GPS tracking data were requested, where available, for Geelong Football Club's 2015 matches. The data were requested and anonymised using the *deidentify.org* portal [189] designed as part of this thesis for the purpose of addressing the issues described in Chapter 5. This section of the pipeline is illustrated in Fig. 7.3. The data extraction requests are listed in Appendix Sec. D.1.

7.1.4 Data Selection Process

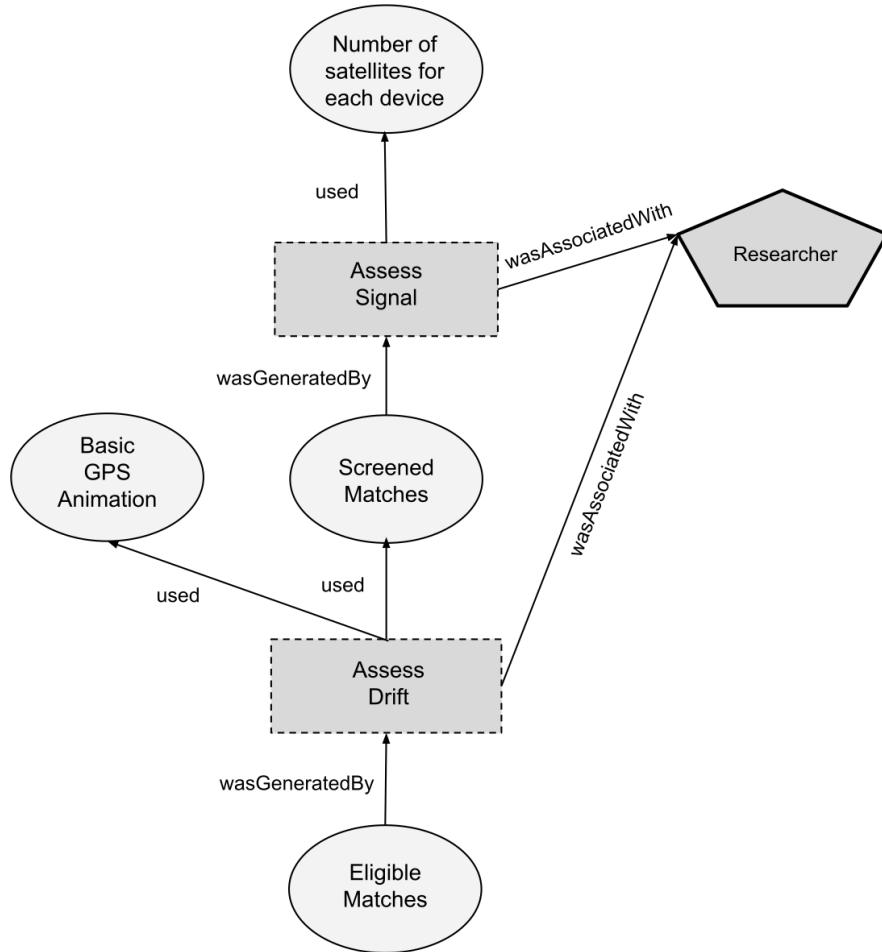


Figure 7.4: Provenance of eligible AFL match list used in this thesis (see Chapter 4 for semantics of notation and Appendix Sec. B.1 for symbols). The Basic GPS Animation used for visual inspection of potential data issues is described in Sec. 7.1.6.

Fig. 7.4 provides an overview of how the GPS data for each match were screened and assessed for data quality. Geelong Football Club played 21 out of 23 rounds in 2015 (round 13 was a “bye” round for the club in which they were not scheduled to play, and round 14 was cancelled after the death of an AFL coach). Of the 21 matches played during 2015, the performance team at Geelong Football Club were able to retrieve GPS data for 16 matches. It was necessary to exclude five of the retrieved matches from the analysis because they did not contain data for all 22 players. A further three rounds were excluded due to frequent signal

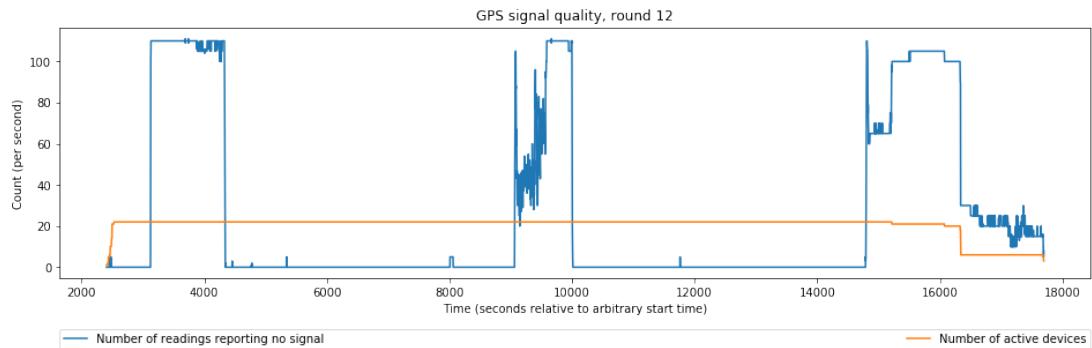


Figure 7.5: GPS signal quality inspection for round 12. All 22 devices (orange) were active during the match. The (blue) peaks in devices reporting no signal are the moments that the team is off field, prior to the game, during the mid-game break, and after the game. As the GPS data provided by the club was exported at a sample rate of 5 Hz, the number of readings reporting no signal per second can reach approximately 5 times that of the number of active devices. While the plot shows there were some instances of signal loss while the game was in play, these are negligible compared to the proportion of the game that had clear signal, and hence this match was included in the analysis.

loss over the course of the entire duration of the game. This left eight matches that had periods of whole-team GPS data during the game. Upon closer inspection, it became apparent that in two of the matches, the GPS devices were active, but there was severe interference near the interchange area causing players who were off field near the interchange area to be reported as if they were at mid-field. While this could have been corrected with identifiable data (i.e. knowledge of which players were benched at which times) to exclude the trace, the non-identifiable point cloud representation made it impossible to distinguish spurious locations in which benched players were reported as being on the field from cases where players were actually on the field. As such, these two rounds were also excluded from the analysis. This left a total of six matches with good data (five of which were played at the team's home ground).

An example of signal quality for one of the selected rounds is provided in Fig. 7.5. Fig. 7.6 shows the number of matches that needed to be discarded at each stage of the selection process due to GPS data issues.

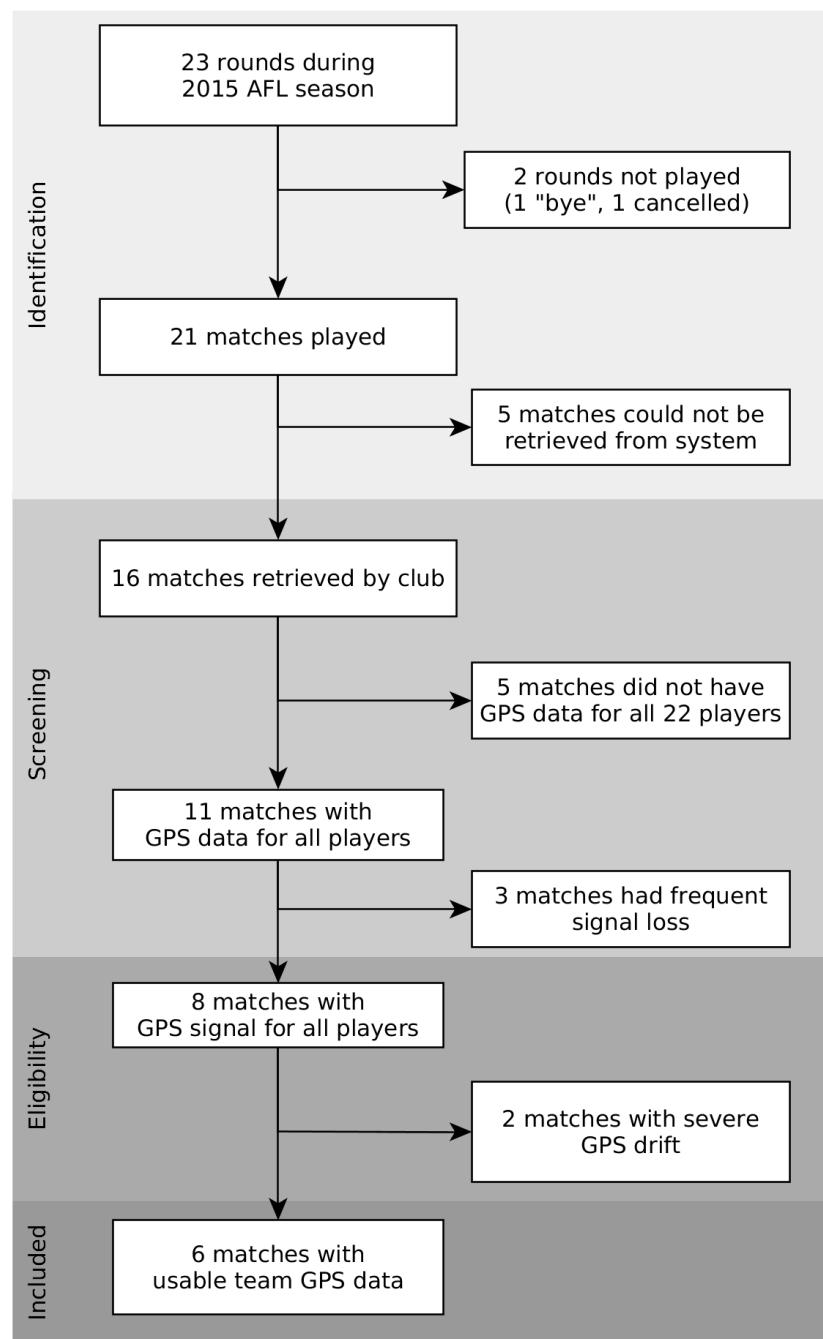


Figure 7.6: Match selection

7.1.5 Reprojection

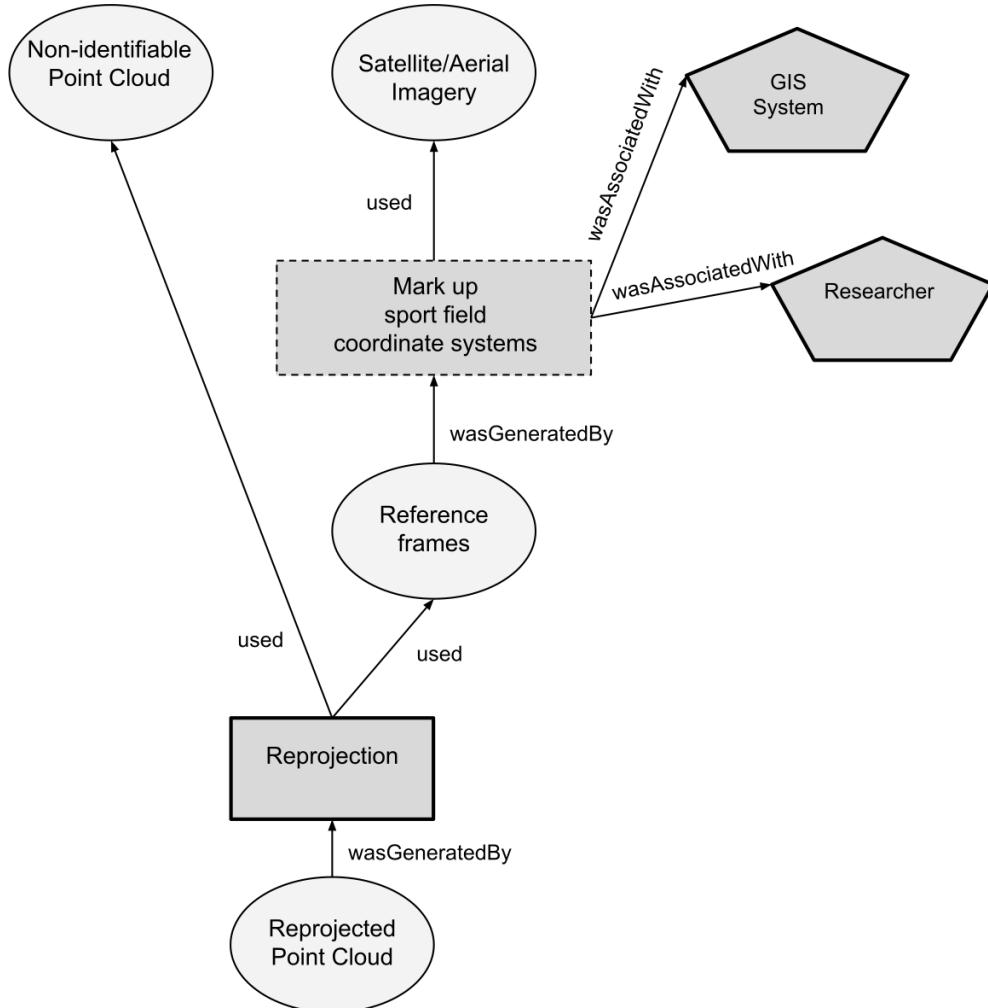


Figure 7.7: Provenance of reprojected point cloud used for visualisation and analysis (see Chapter 4 for semantics of notation and Appendix Sec. B.1 for symbols)

Fig. 7.7 shows the pipeline to pipeline to reproject the player GPS tracking data relative to the field. The first task towards marking up the sport field coordinate system was to determine the location of the goal posts. The goal posts were located using Google Earth Pro, as it included high resolution satellite and aerial imagery of each field, as well as the ability to look at past imagery during the 2015 season to confirm that the shape of the oval had not changed (e.g. due to resurfacing the oval). An example of pinpointing the goals of Kardinia Park (commercially known as GMHBA stadium) is shown in Fig. 7.8.

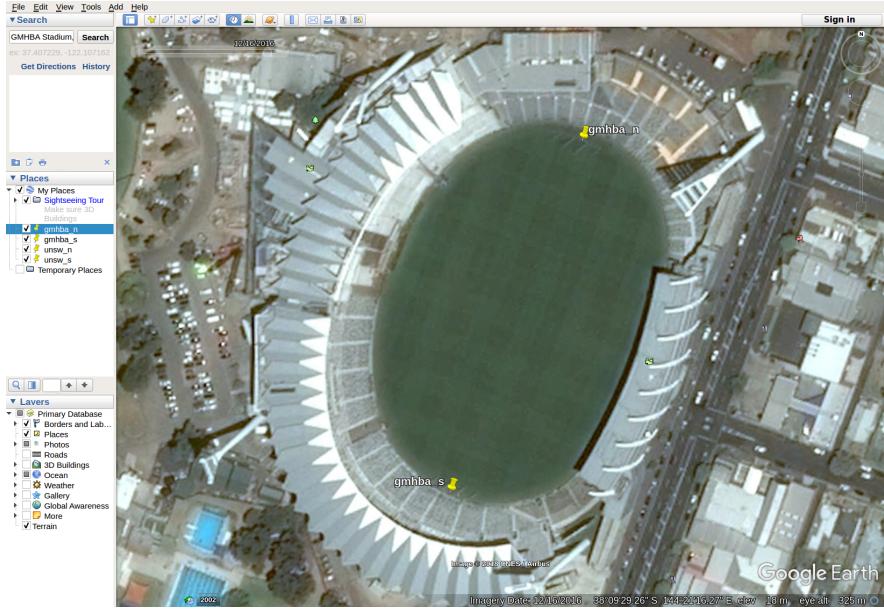


Figure 7.8: Historic satellite and aerial imagery available through Google Earth Pro was used to estimate the coordinates of the goals (yellow pins) at each stadium.

Once the location of the goal posts were determined, these were converted to a geographic file (in the GeoJSON format) specifying a line from the goal area at one end of the field to the goal area at the opposite end of the field, for each venue. These were used to establish reference frames for each venue, as per the *Spatio-Temporal Reference Frames As Geographic Objects* method [192] proposed in Chapter 6. As the time attribute in the GPS data provided by the AFL club contained minutes and seconds but was stripped of the date and hour part, it was not possible to automatically determine which venue GPS data corresponded to based on the time period alone. As such, a special “`filt_radius`” property was added to venues to inform the processor to use that venue as the reference frame whenever the GPS data fell within that radius of the venue. An example of the GeoJSON input to the processor is provided in Listing 7.1.

Once reference frames are established, the processor also requires a directory of GPS data to reproject, which was the de-identified GPS point clouds obtained from the club (Sec. 7.1.3).

```
{  
  "type": "FeatureCollection",  
  "crs": {  
    "type": "name",  
    "properties": {  
      "name": "urn:ogc:def:crs:OGC:1.3:CRS84"  
    }  
  },  
  "features": [  
    {  
      "type": "Feature",  
      "properties": {  
        "GMHBA": "abstime",  
        "filt_radius": 500  
      },  
      "geometry": {  
        "type": "LineString",  
        "coordinates": [  
          [144.354250252864091, -38.158864651758321],  
          [144.354946985064288, -38.157403511709113]  
        ]  
      }  
    },  
    ...  
  ]  
}
```

Listing 7.1: Contents of `venue_refs.geojson`. The location and orientation of each venue is recorded via a LineString drawn between the two ends of the field. As the GPS data provided by the club did not contain the date or hour part, the time is kept as is ("abstime") and synchronised with match footage later in the process. A radius (`filt_radius = 500` metres) is specified for each venue and used to automatically reproject any GPS data falling within that proximity to the coordinate system of the venue

7.1.6 Visualisation

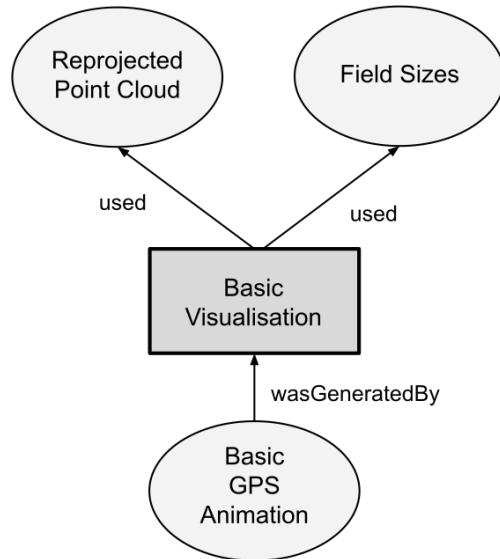


Figure 7.9: Provenance of basic GPS animation developed for this thesis (see Chapter 4 for semantics of notation and Appendix Sec. B.1 for symbols)

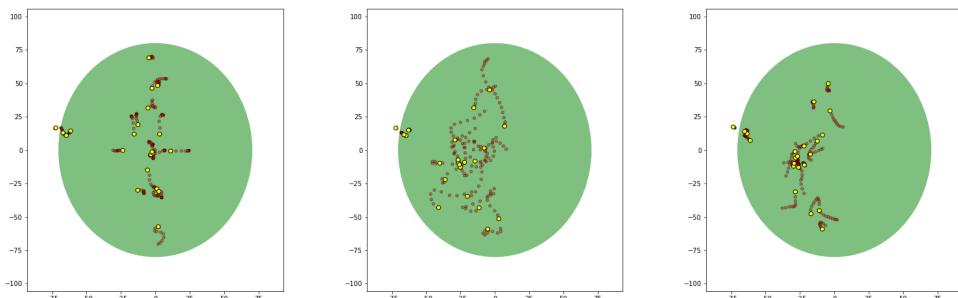


Figure 7.10: Visualisation of player positions for a play during the 2015 grand final. Left: centre bounce formation. Middle: 10 seconds after centre bounce. Right: 20 seconds after centre bounce. Yellow dots indicate position of players, and red trails indicate past positions (logged at 1 Hz). The distance between each consecutive red dot in the trail is distance moved each second, and is thus an indication of player speed.

The reprojected GPS data were then visualised and overlaid on a background indicating the field shape (see Fig. 7.9). An example of this visualisation for various moments during a game is shown in Fig. 7.10. To give an indication of the movement and speed of the players, the vi-

sualisation includes a trail of previous player positions. Note that as a non-identifiable point cloud representation was used, there is no internal representation of which past time points belong to which player; however, the visualisation shows that the path of players over time can still be visually tracked for short periods other than at moments when players come within close proximity to another player. By design, it is ambiguous which player belongs to which trail beyond the point where players cross paths with each other.

7.1.7 Time Synchronisation

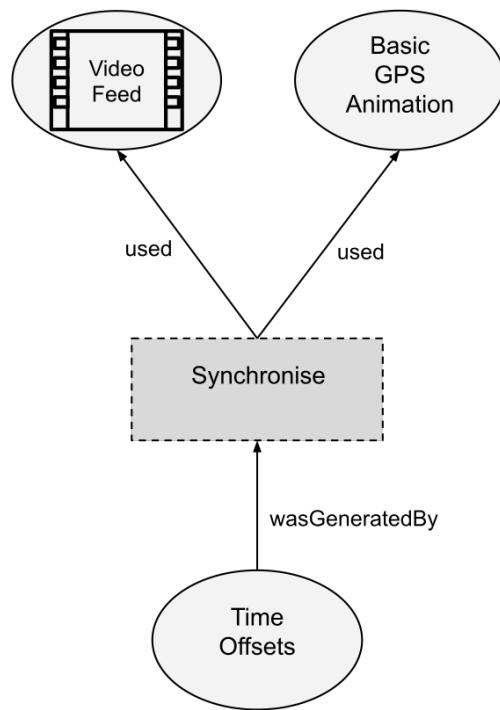


Figure 7.11: Provenance of time offsets used in this thesis (see Chapter 4 for semantics of notation and Appendix Sec. B.1 for symbols)

The GPS visualisation was turned into an animation, showing the current player formation and a 10 second trail of past movements. The animation was then synchronised with video of the game using a custom tool designed for this thesis (shown in Fig. 7.12) in order to obtain the time offsets (see Fig. 7.11). Any event can be used to synchronise the video and animation; in practice the centre bounce at the start of

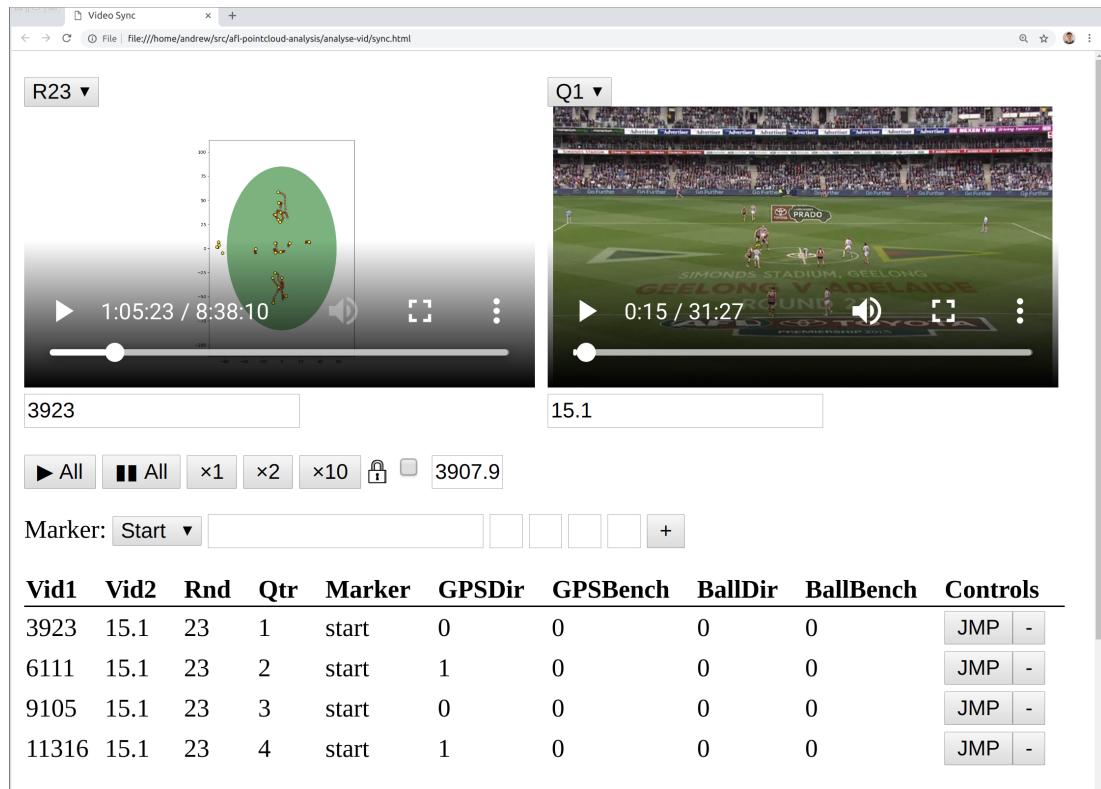


Figure 7.12: Video synchronisation tool

each match was used, as the sudden change in motion provides a clear indicator of when this event occurs.

A key feature of the tool is that once the user has found the approximate offset between the animation and the video feed, they can lock the offset then scrub through the video to verify that the alignment is correct. For example, in one video the first centre bounce formation in the animation was synchronised with the first centre bounce in the match video, but scrubbing through the video to verify the alignment revealed that moments later in the video didn't align with the GPS animation. Investigating this further revealed that the video recording was started late and opened with the second centre bounce of the game rather than the first. Upon reporting this issue, the official video provider confirmed the error and uploaded the complete match footage.⁵

⁵[Issue] Andrew Simmons, “First 30 seconds of Q4 match replay is missing - Geelong vs Melbourne, Round 12, 2015”, Telstra support <https://crowdsupport.telstra.com.au/t5/AFL-Live/First-30-seconds-of-Q4-match-replay-is-missing-Geelong-vs/m-p/790084#M6776>

7.1.8 Synchronised Visualisation

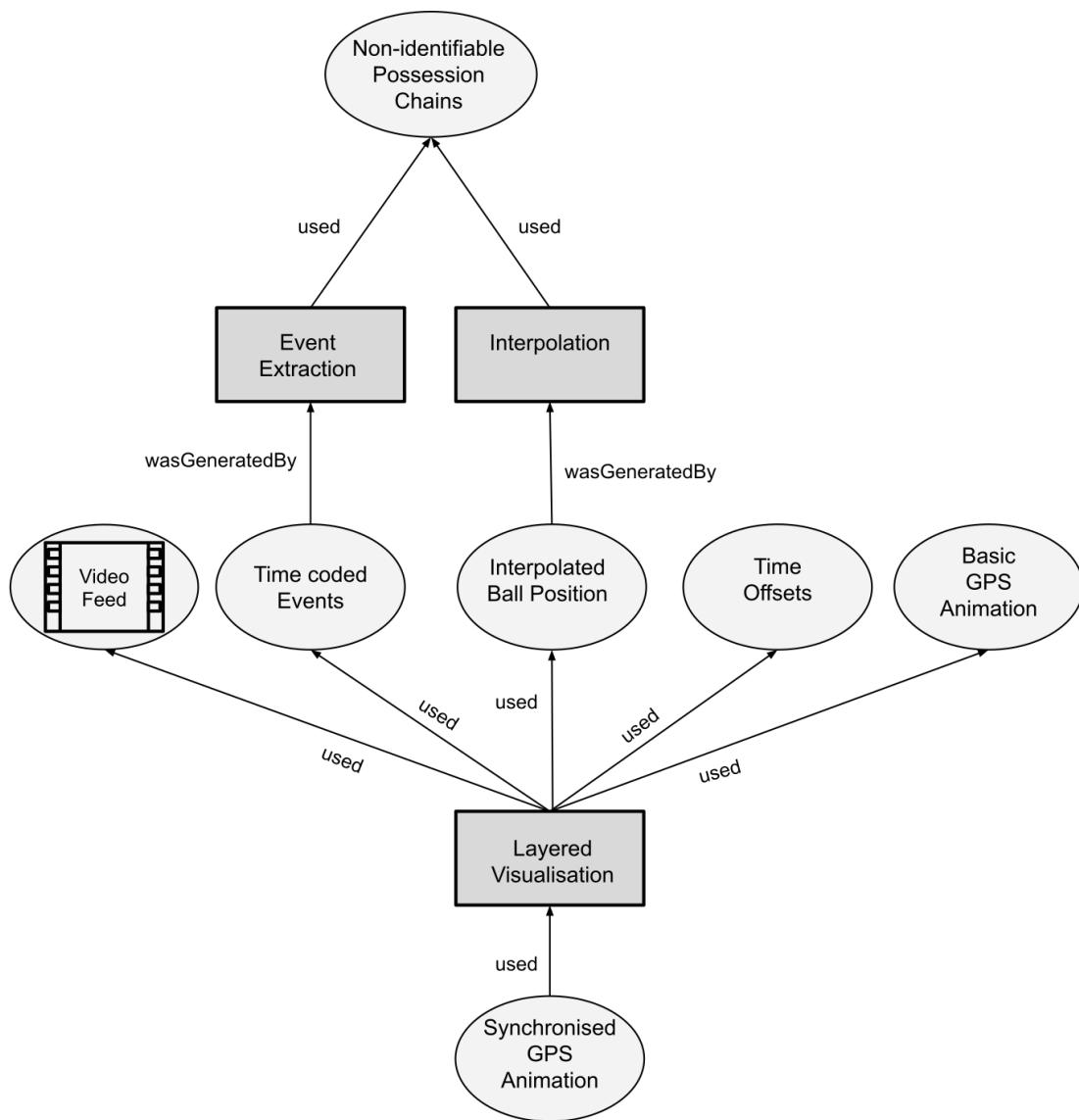


Figure 7.13: Provenance of synchronised GPS animation developed for this thesis (see Chapter 4 for semantics of notation and Appendix Sec. B.1 for symbols)

Once the GPS data had been synchronised with the video, other data sources were integrated, such as Champion Data possession chains. This section of the pipeline is illustrated in Fig. 7.13.

Note that the possession chain timestamps were expressed relative to the start of each quarter, so this alignment was only possible once the synchronisation of GPS data with match video was complete. Champion Data also record the approximate ball location, which they manually pinpoint at key moments during the game. As there is no GPS tracker in the ball, these approximate ball locations were used to provide an approximate ball location feature. AFL clubs use video analysis tools to record their own annotations of the game (e.g. personal communication with the club revealed that they manually annotate the timeline with moments of key player decisions to review). The club did not share these, but if available they could have also been incorporated once the video synchronisation process was complete.

Additionally, during this stage the direction of attacking goals (i.e. which end of the field teams attempt to head towards) was annotated using the tool via manual observation of the animation.⁶ Due to the tendency of the team to chase after the ball, it is possible that this step could be automatically inferred based on which side of the field the team is on at the moments they score a goal. However, as it is critical to the results of the analysis that the direction of attacking goals was determined correctly, this was annotated manually instead.

⁶Teams change sides each quarter, with the initial direction of attacking goals determined by a coin flip at the start of the game. However, there did not appear to be publicly recorded data on coin flips in a form that could be readily translated into a spatial heading.

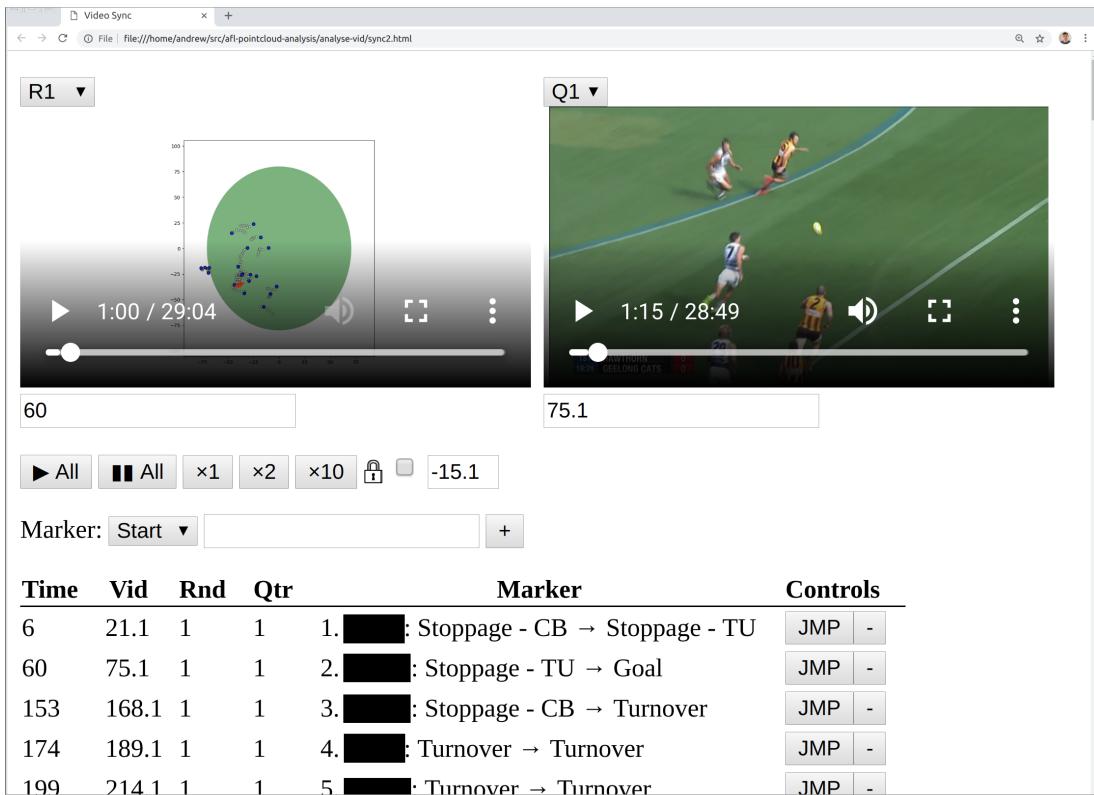


Figure 7.14: Synchronised visualisation

The synchronised visualisation is shown in Fig. 7.14. Note the list of key events shown at the bottom of the interface, as imported from Champion Data possession chain data after synchronisation. Similar commercial video analysis tools, such as Sportscode could be used for this (i.e. annotating and exploring sport video timelines); however, the addition of a GPS visualisation takes this further by providing the performance analyst with insights into the team formation side-by-side with the perspective available from match video.

The visualisations of each quarter were aligned such that the attacking goals were always at the top of the visualisation. This makes it possible to compare formations during different quarters, as well as between different matches. Even if the matches are played on different fields, the visualisations are always oriented consistently, but the field shape will be drawn differently depending on the ground dimensions (e.g. some fields are long and narrow, while others are short and wide). While it would have been possible to scale the fields so that they were always presented as the same shape, field shape can impact on strategy, thus it

was decided not to distort the fields. Later discussion with the club confirmed that they liked that the field shape in the visualisation changes depending on the venue selected rather than introducing distortions.

7.1.9 Analysis Pipeline

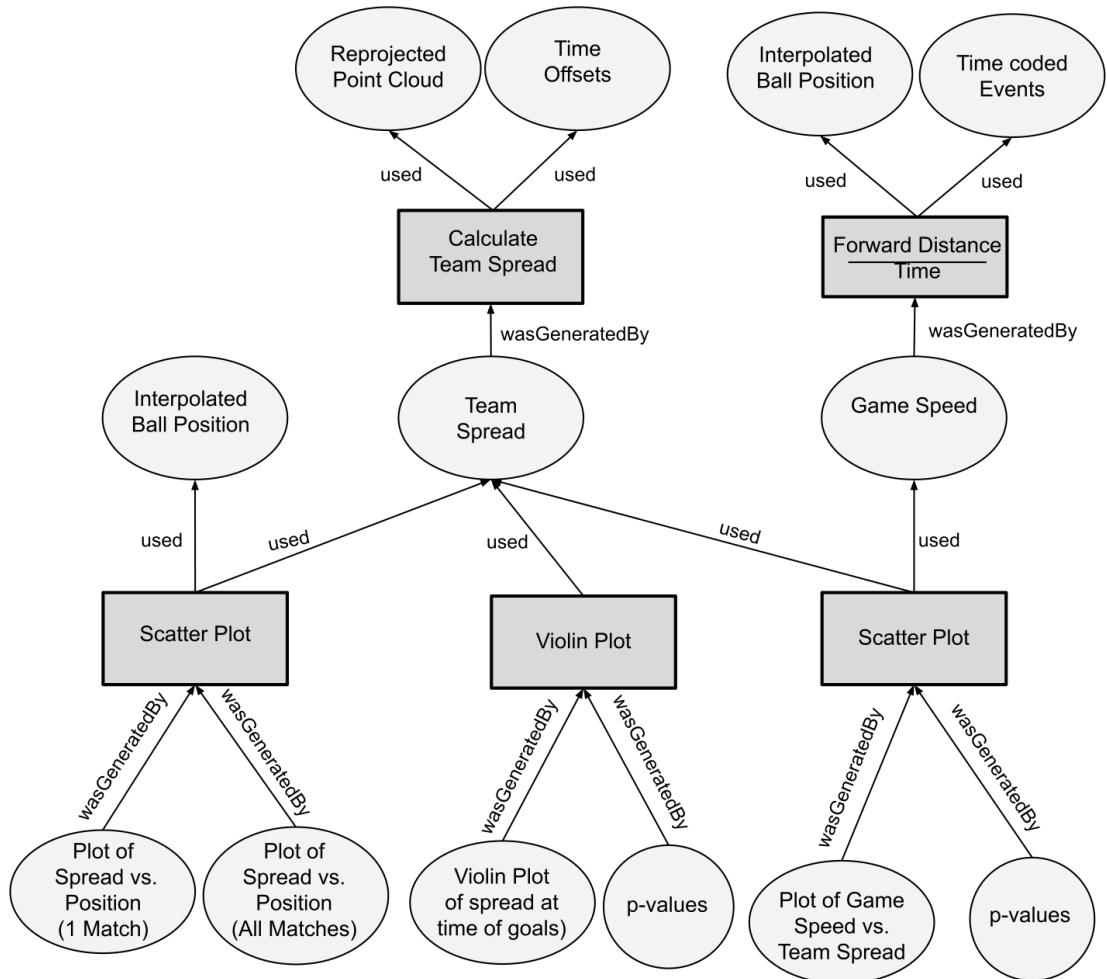


Figure 7.15: Provenance of analysis outputs in this thesis (see Chapter 4 for semantics of notation and Appendix Sec. B.1 for symbols)

To facilitate further analysis, features such as team spread and game speed were derived from the dataset. These will be discussed in depth within Sec. 7.2. See Fig. 7.15 for an overview of how the final analysis outputs discussed in the next section relate back to the intermediate outputs of the pipeline described in this section.

7.2 Exploratory Analysis of Team Shape in AFL using GPS Tracking Data

This section of the thesis will demonstrate how team GPS tracking can provide value beyond existing datasets that focus on just the ball.

In preliminary talks with an elite AFL club, they expressed interest in team formations and how these change in attack and defence. They hypothesised that the team should contract in defence and spread out in attack, although did not have quantitative evidence of this.

The previous chapters have been building up the foundations for a sport analysis platform. This section builds upon this platform to perform a study of how team formations spread and contract during attack and defence phases of play in AFL.

7.2.1 Introduction

While individual player tracking data can be readily digested by sport practitioners, the high-dimensional nature of whole-team tracking data presents challenges for those seeking to utilise tracking data for team-level strategy analysis. This section of the thesis introduces a technique for analysing team formations based on how spread out the team is. The approach proves promising as a framework to empower sport practitioners to derive strategic insights from team level GPS tracking data.

7.2.2 Related Work

This section will briefly highlight existing metrics proposed in the literature for measuring team shape. Early papers focus on analysis of team shape in Association Football using player trajectories derived from video [225, 40, 41, 74, 45] or local position measurement systems

[77, 78]. Basketball has also been studied [23]. Later studies have applied these techniques using GPS systems [129, 79, 186]. Existing metrics proposed in the literature for measuring team shape include stretch index, surface area, length-to-width ratio, and spatial variability.⁷ Alexander et al. (2019) were first to apply these kinds of metrics to AFL [4], although so far have only studied a simulated drill [4] and later a single AFL match [5] using existing approaches designed for Association Football and basketball.

Alexander et al.'s initial study [4] examined team length, width and surface area in a 15-v-15 Australian Rules Football simulation match. Team length, width, and surface area were found to be greater when teams were in offence than defence. Alexander et al. noted some counter-intuitive results arising from their methodology, such as teams being centred further back in offence than in defence, which was speculated to be an artefact of the location teams gained possession of the ball.

A follow up study by Alexander et al. [5] examined a single competitive AFL match played by the team (however did not have data for the opposition). In alignment with their first study of a 15-v-15 simulation match [4], the follow up study also found that team length, width, and surface area were greater in offence than defence. They used heat-maps to visualise the typical locations occupied on the field broken down by possession state and which zone of the field the ball was in. They quantified the level of spatial variability in the heat-maps using Shannon Entropy, which revealed lower variability when the ball was contested; however, similar values in offence and defence.

Research is needed to explore whether Alexander et al.'s findings hold for other AFL teams over multiple matches, and to test alternative measures of team spread.

⁷An overview of these metrics is provided in Appendix Sec. D.2

7.2.3 Method

Datasets

Three datasets were synchronised and integrated to support the analysis: player position tracking data, possession chain data, and ball position data at time of possessions.

Player position tracking data collected by wearable GPS tracking devices (GPSports) worn by an elite AFL club were obtained for matches played during the 2015 season. After removing rounds in which one or more player devices were not working or experienced signal loss, six matches were available, five of which were played at the team's home ground. For consistency only the five home ground matches were selected for analysis.⁸

Possession chain data for each match were recorded by the official AFL statistics provider Champion Data. A possession chain is a sequence of possessions and disposals (e.g. kicks, handballs) by a single team from the moment of initial ball possession through to a scoring event (e.g. a goal) or loss of possession through a turnover to the opposition or stoppage (e.g. ball out of bounds).

There is no GPS tracking device in the ball; however, Champion Data graphically pinpoint the approximate ball location on a diagram of the field at the time of each possession. The time of these events is reported to be accurate to 5 seconds, and ball position is reported to be accurate to approximately 5 to 10 metres [162].

Data Processing

Team GPS data were anonymised by converting them to a point cloud sequence. This was performed according to the methodology estab-

⁸See Sec. 7.1.4 for details of the data selection process, and *Experiment Design* below for the rationale for focusing on the home games in this analysis.

lished by Simmons et al. 2018 [189] (Sec. 5.7) to ensure the research team did not have access to the underlying player tracking data in (re-)identifiable form at any stage of the project. GPS data were reprojected relative to the local coordinate system of the playing field using the *GPS to XYT* tool designed by Simmons et al. 2017 [192] (Sec. 6.3). As the study involved only non-identifiable data, it was granted ethics exemption by Deakin University Human Research Ethics Committee (2018-121: GPS analysis of team shape and game style in elite Australian Rules Football).

The component of the ball position vector in the direction of the goals was used as a proxy measure of how close the team is to scoring a goal. More refined approaches are possible, such as the Euclidean distance from goal, or the *apparent width* [111] of goals (in order to account for the apparent narrowing of the goal area when attempting to kick a goal from an angle).

Team Spread

The high-dimensional nature of the GPS formations makes analysis difficult in their raw form. The team has 18 players on the field, each with an x, y dimension (ignoring analysis of jumps), resulting in 36 continuous dimensions, i.e. formations exist in \mathbb{R}^{36} . Thus a team spread feature was generated to reduce this to a single dimension, which facilitates visual exploration by humans, as well as reducing risk of overfitting during analysis. Let team spread at a given time instant be defined as the Root Mean Square (RMS) of player distances from the centroid of the team formation (Eq. 7.2.1):

$$spread = \sqrt{\frac{1}{N} \sum_{i=1}^N ((x_i - \bar{x})^2 + (y_i - \bar{y})^2)} \quad (7.2.1)$$

where

x_i is the x-coordinate of player i at the time instant

y_i is the y-coordinate of player i at the time instant

\bar{x}, \bar{y} are the coordinates of the centroid of the team

$N = 18$ is the number of team players on the field

This can be thought of as a two-dimensional form of the formula for standard deviation. Loosely, it measures the average distance of players from the centre of the formation (with greater weight being given to players who are further away). In this sense it is similar to the definition of the “stretch index” [225, 23] used in the literature; however, it uses the quadratic mean (RMS) rather than the arithmetic mean. Prior to the spread calculations, players in the interchange area were removed, which was achieved by removing the four player locations reported as closest to the interchange side of the field. A visual explanation of the spread metric and ball position is shown in Fig. 7.16.

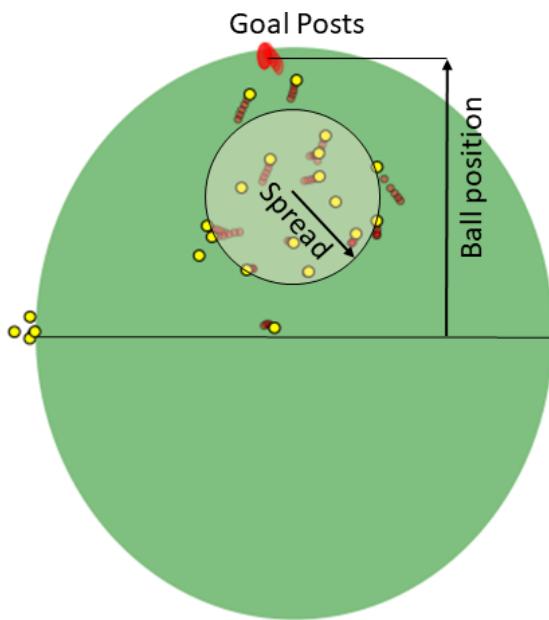


Figure 7.16: The ball position component in the direction of the attacking goal is used as an indicator of how close the team is to scoring. The team spread feature measures how dispersed the team is relative to the centre of the pack. Interchange players are removed from the spread calculation. See Eq. 7.2.1 for details.

Experiment Design

Alexander et al. [4] noted counter-intuitive results arising as an artefact of the location teams gained possession of the ball. If all possessions are analysed, then the analysis of team shape will be confounded by the influence of the location that the team gains initial possession of the ball. For example, if the team were good at gaining possession of the ball from the opposition near the edge of the field, then the team shape during the attack phase would disproportionately represent the shape of the team near the edge of the field rather than strategic differences between defence and attack formations.

To control for these confounding effects, this study compares possession chains in attack and defence from known starting locations. The centre bounce (when an umpire bounces up the ball to begin play) is an example of this, as it always occurs in the centre of the field. Ruckmen are responsible for attempting to hit the ball from the centre bounce to their own team; however, an investigation by O'Shaughnessy noted that the team that gains possession from the centre bounce is “essentially a coin flip” [161]. Therefore, the centre bounce can be considered as a form of randomised assignment that can be used to measure the effect of gaining possession. This study will also analyse possession chains beginning with a kick-in, which are always taken from the end of the field.

Limiting the study to only certain possession chains limits the data; however, also strengthens the rigour of the analysis by ensuring that confounding effects are prevented through use of a consistent initial possession location. The utilisation of all five available home matches was essential to ensure sufficient data. While data for a sixth match was available, the field shape for this match differed to the home ground, thus was discarded to ensure a consistent field shape across all samples.

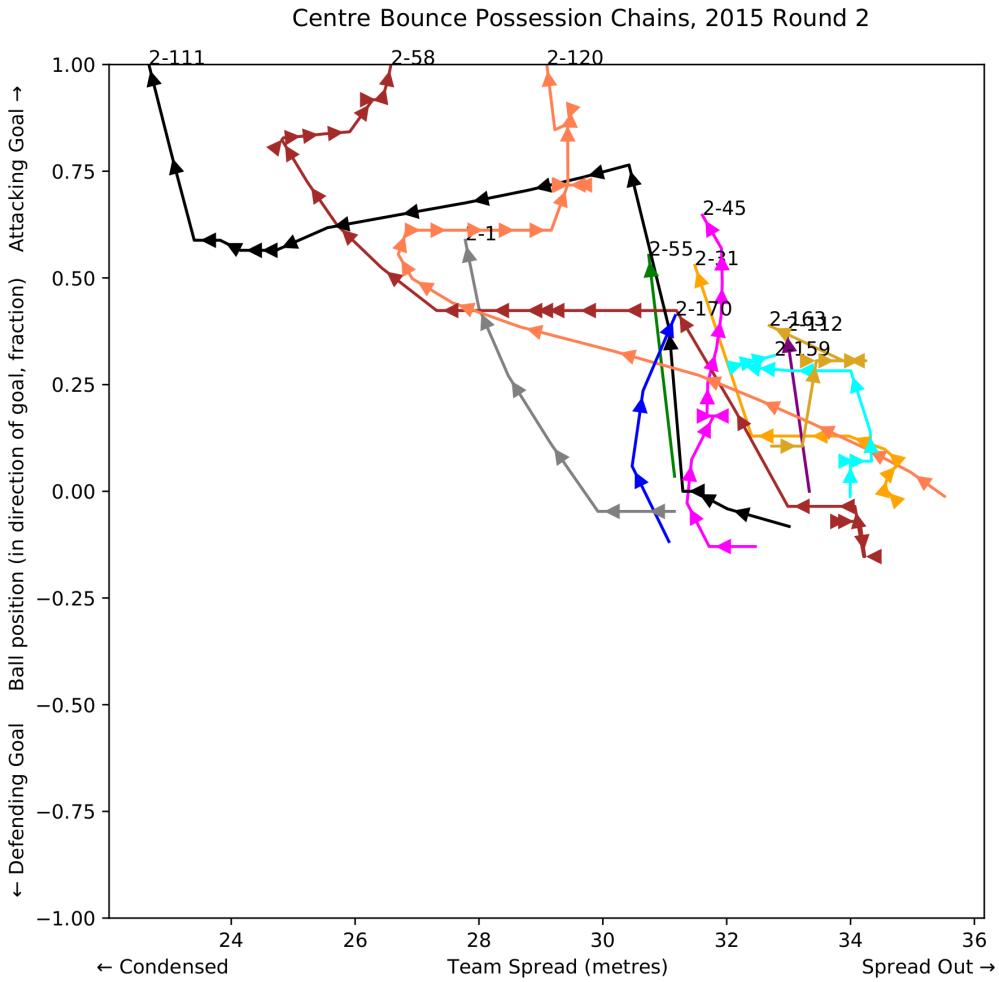


Figure 7.17: Parametric plot of team spread and ball position for possession chains in which the analysed team successfully gained possession shortly after the centre bounce. Arrows denote 1 second intervals. Chains start at the moment of possession and end when the team loses possession. Chain are labelled with the round and sequence identifier (e.g. 2-111 ⇒ round 2, chain 111) to allow performance analysts to rapidly cross-check these with video footage. To prevent cluttering, only possession chains for one match are shown in this figure.

7.2.4 Exploratory Analysis

The parametric plots of team spread and ball position in Fig. 7.17 show how the team shape at the start of play (middle-right of figure: spread > 30, ball-position = 0) changes over time as the team moves towards the goal area (top-left of figure: spread < 30, ball-position = 1.0). The plot shows that in all 3 cases where the team reached the goal area, the team condensed as they move closer towards the goal. This is to

be expected, as the oval shape of the field means that when the team follows the ball they will end up packed together in a small area at the end of the oval. However, the plots show variation in the degree that the team packed together at the time of goals, which may be indicative of strategic differences.

Trace 2-111 (black) shows that after quickly advancing the ball, the team passed the ball backwards. This behaviour warrants further inspection.⁹ According to the official possession chain data, the team was considered in possession of the ball for the entire duration; however, the video shows that the team was under pressure from the opposition, and did not have clear control of the ball during this period. While moving further back from the goals is disadvantageous, the plot shows that during this time the team shape also changed and that the team emerged closer packed than before. Inspecting the GPS visualisation animations side-by-side with video confirms that prior to moving the ball backwards team members were in poor control of the ball (Fig. 7.18) whereas after moving the ball backwards, the team emerged repositioned closer to the ball and regained control (Fig. 7.19). This demonstrates the value of the plots in drawing attention to possible strategic advantages of reshaping the team formation that would not be evident from consideration of the ball position alone.

The same type of plot discussed above is shown for a different round in Fig. 7.20. Note the anomalous chain 18-186 (magenta) that shows the team spread out slightly as they headed towards goal. The long arrows denote that the ball was passed quickly towards the goal. Due to the rapid pace of ball movement, the team did not have time to follow after the ball and crowd within the goal area as they usually would. It is also possible that remaining spread out assisted the team in quickly moving the ball forward.

⁹In contrast to rugby, AFL teams are permitted to pass the ball forward, and generally do so



Figure 7.18: Team formation visualisation side-by-side with video footage at the instant of the first turning point for chain 2-111 in Fig. 7.17 where the ball was passed backward. Note the formation animation showing players trying to catch up to where the main action is occurring



Figure 7.19: Team formation visualisation side-by-side with video footage at the second turning point for chain 2-111 in Fig. 7.17 where the team started moving the ball forward again. Note the team have reformed in a dense formation near the centre of the field where the ball has been passed to.

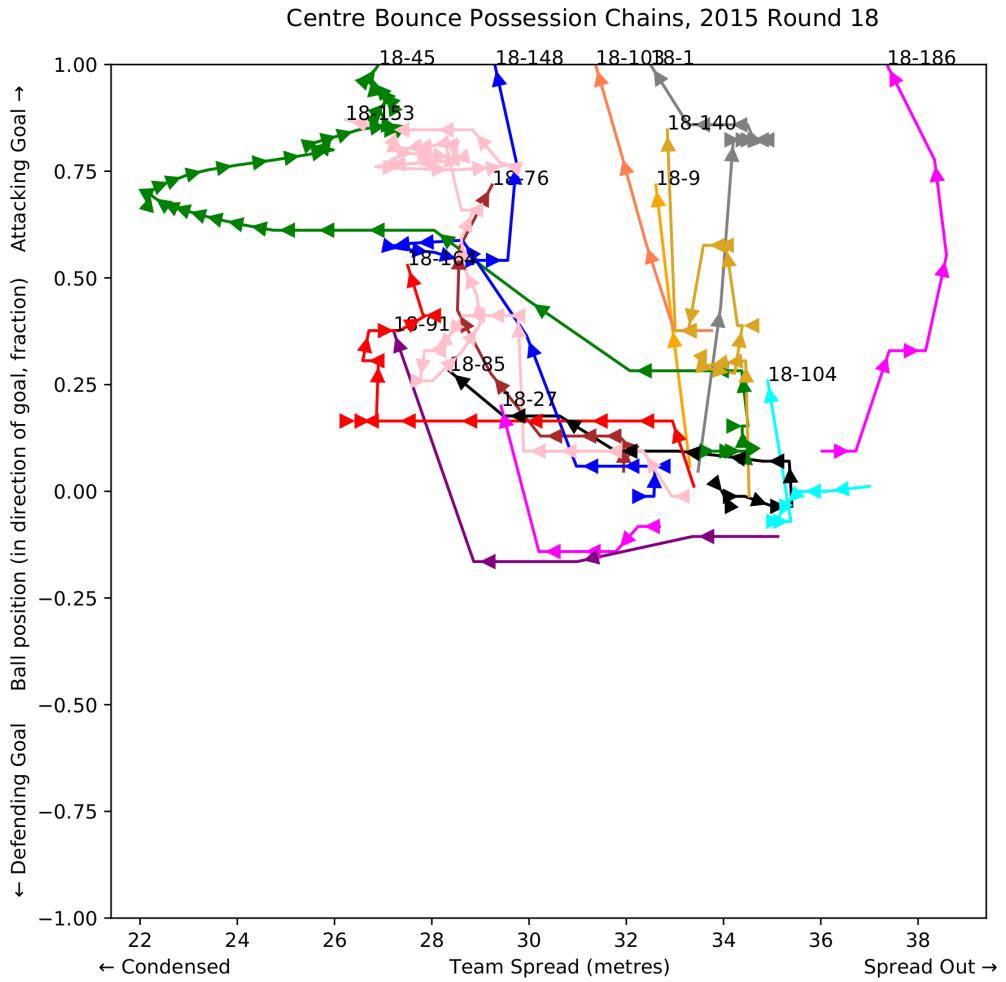


Figure 7.20: As in Fig. 7.17, but for a match in which the team was more successful in moving the ball from centre bounce directly towards the goal area

7.2.5 Visual Comparative Analysis

This sub-section explores data taken from possession chains over multiple matches. Possession chains from different scenarios are overlaid on each other to visualise how formations differ between attack and defence.

A comparison of attack (green) and defence (red) formations conditioned on which team wins the centre bounce is shown in Fig. 7.21. The direction during defence formations has been flipped such that defending goals during a defence phase are towards the top of the figure to align with attacking goals during the attack phase. With the exception of one

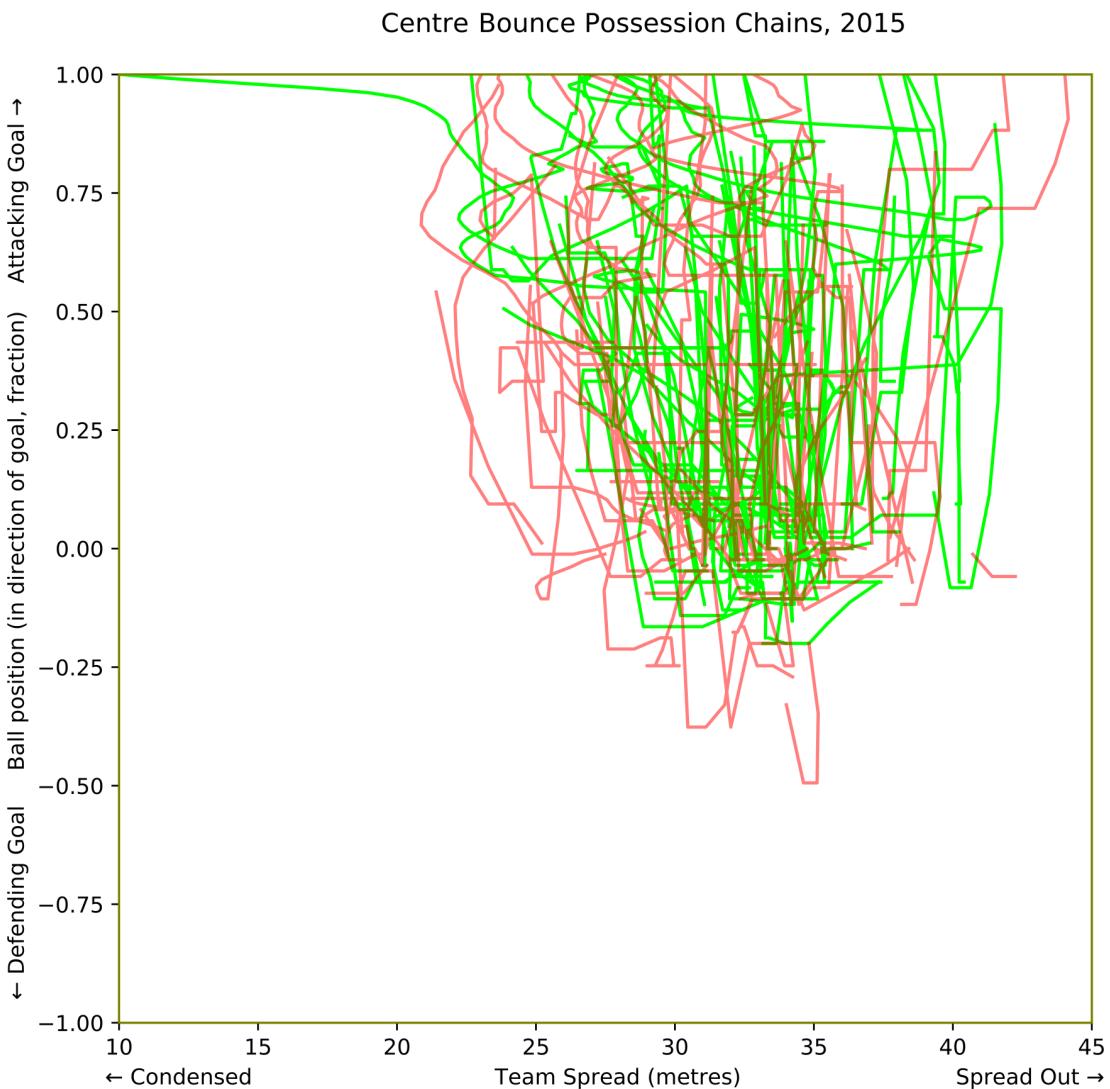


Figure 7.21: Comparison of chain spread and position broken down by attack and defence phases. Attacking chains beginning with a win at centre bounce are shown in green. Defending movements for opposition chains beginning with loss of the centre bounce are shown in red. The direction of defending movements has been flipped for easy comparison with attacking chains.

anomalous attack trace (round 23, chain 69) which extends to the far left of the plot, there are no distinctive visual differences. Examining video of the anomalous trace reveals that team-mates tightly packed into the goal area during the goal attempt.

The same procedure was applied to compare attack (green) and defence (red) formations during kick-ins (Fig. 7.22). Unlike the case of centre bounces (Fig. 7.21) mentioned previously, in which the situation is sym-

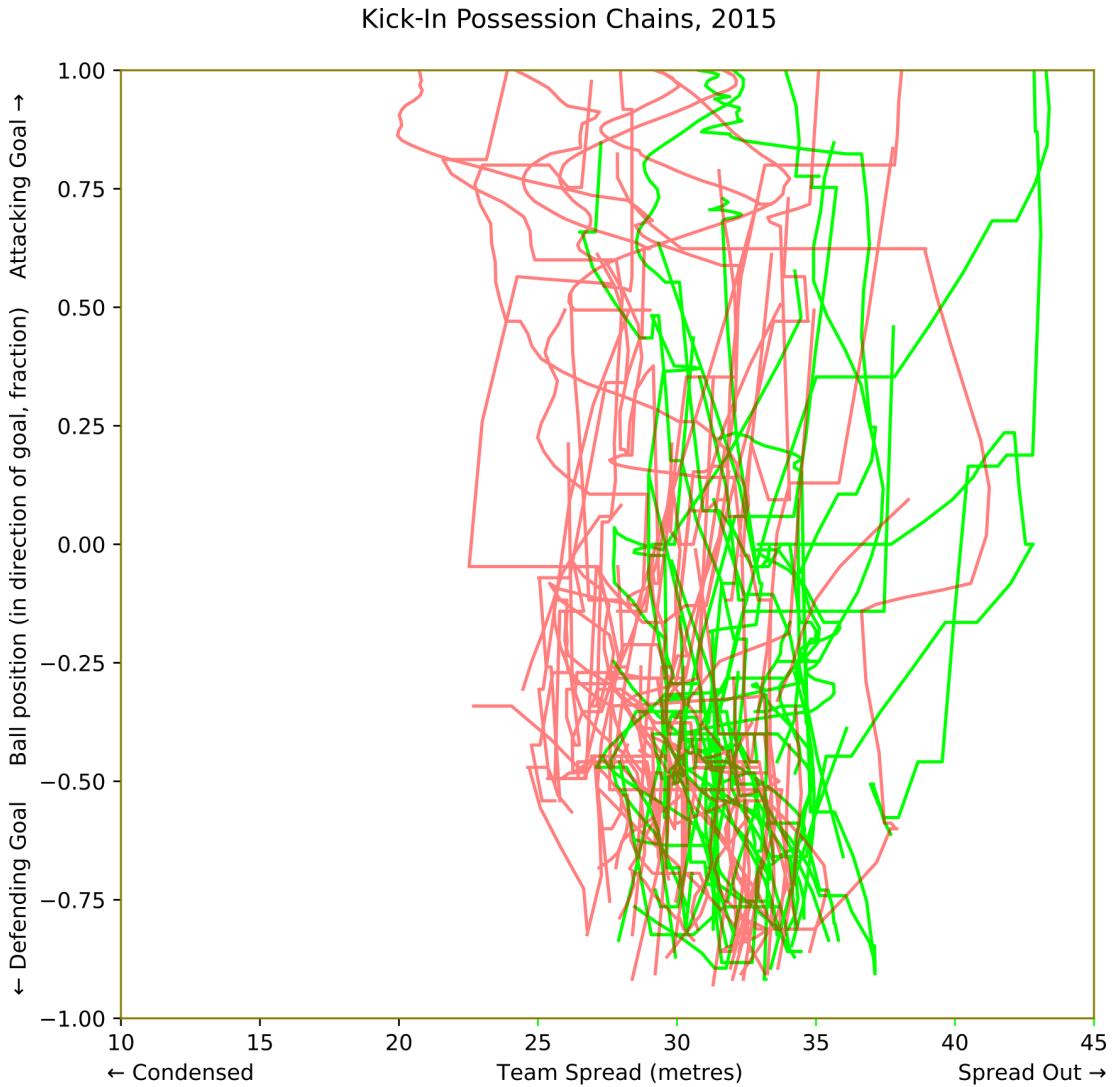


Figure 7.22: As in Fig. 7.21, but for chains beginning from a kick-in from the far side of the field.

metrical for both teams until one team gains possession, in a kick-in one team is given possession at one end of the field, and must attempt to move the ball along the full length of the field past defenders to the attacking goal at the opposite end of the field. As before, the direction of defending traces were flipped so as to permit visual comparison with attack traces. It is visually apparent that the defence formations (red) tend to be more condensed than attacking formations (green), with the difference becoming more pronounced as the ball moves towards goal.

7.2.6 Quantitative Analysis

This sub-section quantifies the visual observations made in the preceding sub-sections.

Quantifying Team Spread in Attack versus Defence

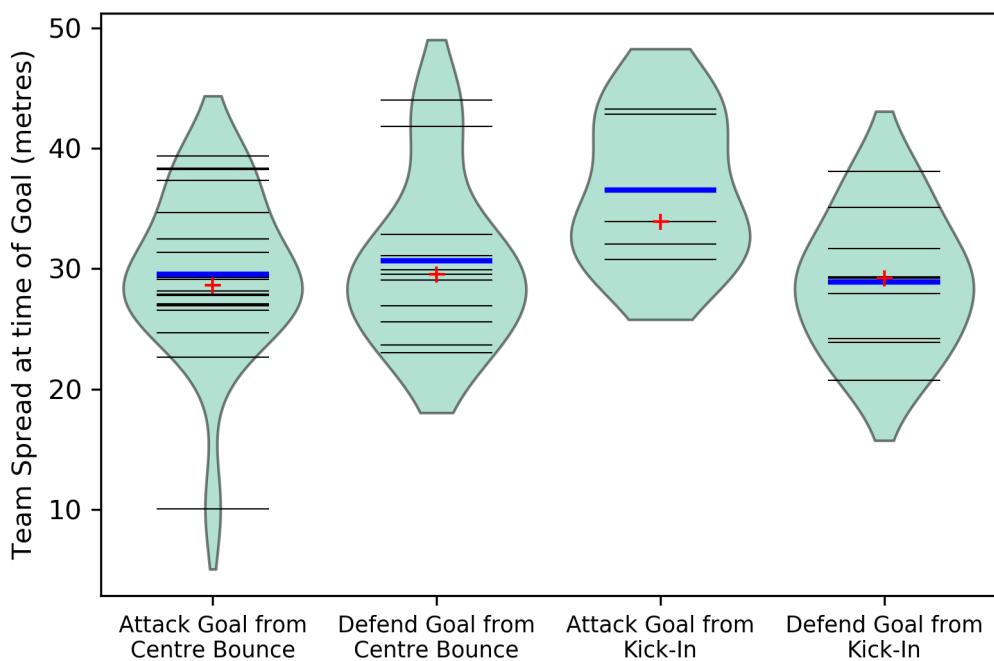


Figure 7.23: Violin plots showing the density distribution of spread at the time the ball reaches the goal area. From left to right: chains beginning with won centre bounce; movements during opposition chains beginning with lost centre bounce; chains beginning with team kick-in from far side of field; movements during opposition chains beginning with opposition kick-in. Each black line corresponds to a possession chain that reached the goal area and marks team spread at the time of goal (or team spread at the time of conceding a goal when defending). Blue lines denote the mean, and red pluses (+) denote the median.

Violin plots showing the density distribution of team spread at the time the ball reaches the goal area are shown in Fig. 7.23. Values of the mean and standard deviation for each group are reported in Table 7.1. Prior to conducting further statistical tests, Shapiro-Wilk tests were performed to validate the assumption of normality, which held for all groups ($p > 0.05$) with the exception of the Attack Goal from Centre

Bounce group ($p = 0.04$) due to an outlier¹⁰. To test the hypothesis that team spread would change in attack versus defence, Welch's t -tests¹¹ were performed to test the significance of difference between attack spread versus defence spread, controlling for the location of initial ball possession (centre bounce/kick-in). Examining team spread at the time of goal following a centre bounce, there was no statistically significant difference between the group of cases where the team was defending and group of cases where the team was attacking¹² ($p > 0.05$). Examining team spread at the time of goal following a kick-in, the team is more spread out during an attack on goal than when defending against an attack on goal. This result is weakly statistically significant ($p < 0.05$). Details of the t -tests are reported in Table 7.2.

Table 7.1: Team spread statistics for attack/defence at the time instant the ball reaches the goal area for possession chains starting from centre bounce/kick-in. Table corresponds to visual presentation of data in Fig. 7.23. N denotes number of possession chains of the specified type that reached the goal area. μ denotes the mean (i.e. the expected team spread at the time the ball reaches the goal area). s denotes the sample standard deviation (i.e. an indicator of whether the team spread is consistent at the time the ball reaches the goal area across multiple possession chains).

	N	μ (metres)	s (metres)
Attack Goal from Centre Bounce	20	29.57	6.67
Defend Goal from Centre Bounce	11	30.68	6.78
Attack Goal from Kick-In	5	36.57	6.03
Defend Goal from Kick-In	9	28.91	5.54

Quantifying Relationship between Game Speed and Team Spread

As a result of the exploratory analysis, it was hypothesised that the team being spread out is associated with quick forward ball movement.

¹⁰Note the anomalously low value of team spread (10 metres) occurring in the Attack Goal from Centre Bounce group due to an instance where the team packed closely together near the goal area. Re-running the Shapiro-Wilk test without this outlier indicates the group is consistent with the assumption of normality ($p = 0.12$).

¹¹Unlike Student's t -test, Welch's t -test does not assume equal variance.

¹²The difference is not statistically significant regardless of whether the outlier is included in the calculation or not.

Table 7.2: Welch's t -test results showing significance of difference of team spread at time of attacking goal (μ_A) versus team spread at time of defending goal (μ_D). t and p denote the t -statistic and two-tailed p -value reported by Welch's t -test (carried out using SciPy 1.2.1). Significant results marked with a star (*).

	$\mu_A - \mu_D$	t	p
Attack vs. Defend Goal from Centre Bounce	-1.11	-0.44	0.66
Attack vs. Defend Goal from Kick-in	7.66	2.34	0.05*

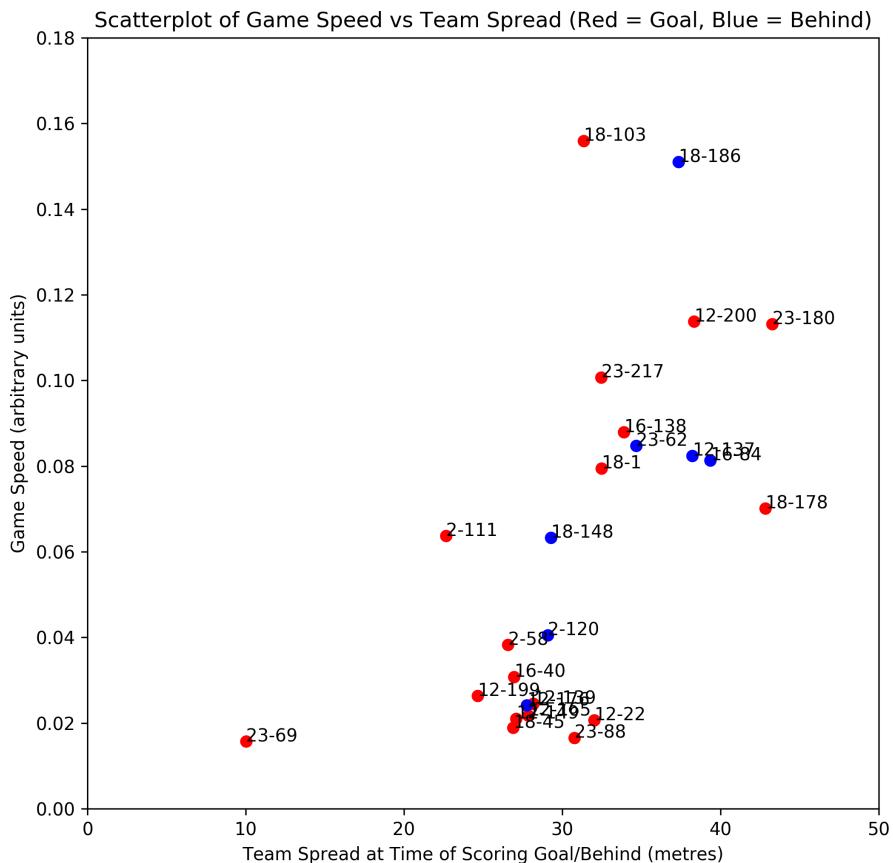


Figure 7.24: Scatterplot of game speed versus team spread (at time of scoring). Combines data from chains beginning with team possession at centre bounce and chains beginning with team possession at kick-in. Red dots indicate spread and average game speed at time of goal. Blue dots indicate a behind (missed goal). Labels denote the match and chain identifier so that performance analysts can cross-check anomalous chains with video footage.

To test this hypothesis, the following variables were utilised: the team spread; the forward distance the ball moved (e.g. a centre bounce to goal only has to cover half the field, whereas a kick-in to goal has to

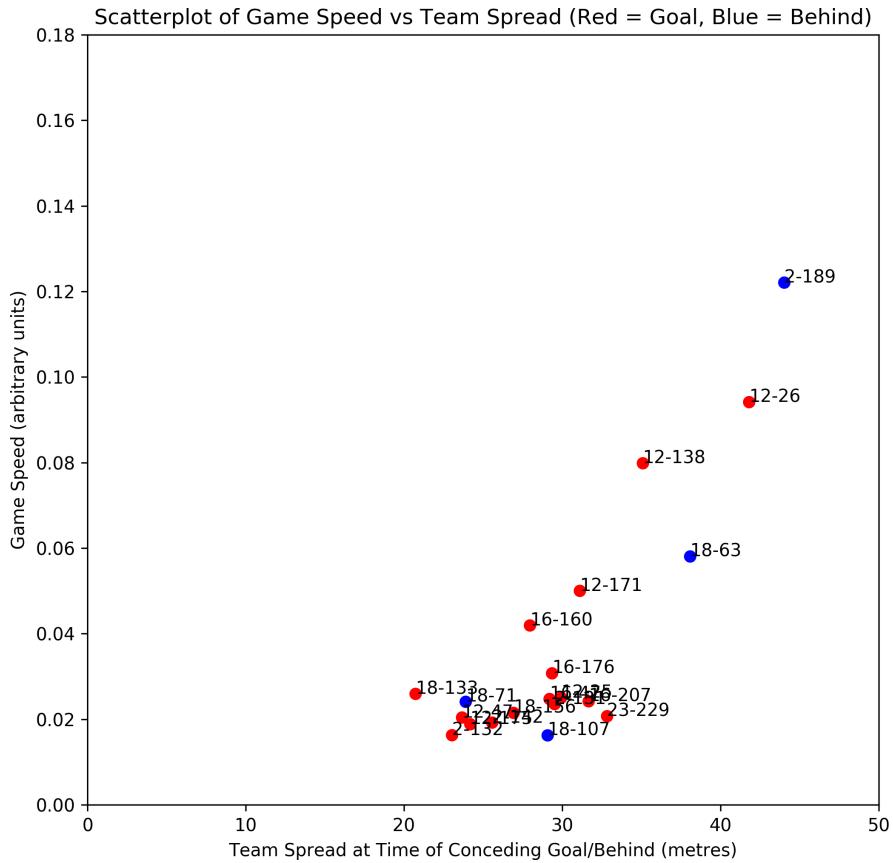


Figure 7.25: As in Fig. 7.24, but analysing game speed versus team spread for defence at time of goal by opposition team.

cover the full length of the field); and the duration of the possession chain. *Game speed* was then defined as the forward distance¹³ divided by time. An increasing relationship was found between team spread and game speed when attacking goal, corresponding to the data displayed in Fig. 7.24 (Pearson's $r = 0.61$, $p < 0.05$; Spearman's $\rho = 0.69$, $p < 0.001$). A strong linear relationship was found between team spread and game speed when defending goal, corresponding to the data displayed in Fig. 7.25 (Pearson's $r = 0.85$, $p < 0.001$; Spearman's $\rho = 0.66$, $p < 0.01$).

¹³Only the displacement vector between the position at the time of initial possession of the ball and the position at the time of losing possession was considered, as opposed to the total path length. Furthermore, only the component of the displacement vector in the direction of goal heading was used. This means the forward distance metric will be less than the total path distance covered by the player. However, this is sufficient for the analysis, as the analysis is primarily concerned with game speed ratios rather than measuring physical ball speed.

7.2.7 Discussion

In communication with analysts at the club that provided the data, they explained that, in principle, teams try to spread out for a good attack to get around the defence, and condense in a good defence. However, when examining possession chains that begin with a centre bounce, there was no significant difference observed between team spread during attack and defence. This may be explained by AFL players defending by “tagging” (following after) a player of the opposite team. When examining kick-ins, there was weak evidence to suggest that the team spread out more when attacking from a kick-in than when defending from a kick-in. The club suggested that this may be explained by a “zone defence” strategy (protecting space rather than following after an opponent) employed in scenarios when the team has sufficient time to set up a defence.

There may be some misalignment of the club analysts’ conception of spreading out, and the mathematical definition of spread used in this thesis. There are multiple ways that these notions could be formalised, thus future work is needed to explore alternative mathematical definitions of spread that more closely align with the tacit definition used by sport practitioners. For example, if a player is injured or fatigued, they may remain on-field for a short period, but may stay still rather than moving with the rest of the team formation, thus should ideally be removed from the team spread calculation.

There is a strong correlation between team spread and game speed; however, caution is needed in the interpretation of this relationship. It is possible that spreading out helps the team move the ball forward quicker; however, a more likely explanation is that when the team moves the ball forward quickly, they do not have time to reform into a unified attacking formation as they would when they move the ball slowly forward. To definitively test for a causal relationship would require an intervention study to vary a single variable at a time. If causality can be established, then understanding this relationship would allow teams to control the game speed. While this study was unable to establish

whether a certain team spread or game speed is desirable¹⁴, knowledge of which formations speed up or slow down the game could have strategic implications when only a short time remains before the end of the quarter.

7.2.8 Conclusions

This thesis section demonstrated an approach to utilising GPS data for the purpose of strategic analysis of AFL teams. It demonstrated the value of GPS data as a tool to form hypotheses, identify patterns, investigate anomalies, and quantitatively test performance analysts' assumptions about the game.

Caution is advised in the utilisation of specific findings. The difference between attacking and defending formations for kick-ins was only weakly significant; however, the observation that teams generally spread more in attack than in defence appears to be in alignment with Alexander et al.'s preliminary findings [4, 5]. While there is a strong correlation between spread and game speed, causality cannot be established, and it is not clear to what extent the finding generalises to other teams and venues.

Nevertheless, the approach demonstrates the ability to deliver insights of direct interest to sport clubs. While larger datasets from a diversity of teams are required to confirm these findings for sport science research purposes, even weakly significant findings can provide value to sport clubs as chance and uncertainty are inherent components of the game.

The approach presented could be applied to other team invasion games. However, game specific modifications to metrics may be necessary. In AFL, particularly prior to 2019 rule changes, player roles are flexible, and the entire team tends to move with the ball. Modifications to the

¹⁴From the exploratory analysis, up until the point that a turnover occurs, there does not appear to be any visually apparent distinction between chains that end short, and chains that are successful. Furthermore, the behinds (blue) highlighted in Fig. 7.24 appear to be distributed similarly to goals (red) in terms of both spread and game speed, indicating that goal accuracy is independent of these factors.

spread metric would be needed to account for sport games with more rigid set formations.

7.3 Feedback from Sport Performance Analysts

Validating the work in this thesis is difficult, as it is unrealistic to expect coaches to immediately adopt the system. Even in the case that coaches were to adopt the system, the performance benefits will likely be small compared to the overall element of luck involved in the game, so the evidence would be anecdotal at best. Instead the system was informally evaluated by presenting it to four sport performance analysts at an elite AFL football club and requesting informal feedback as to which features that they found most useful.

The system was demonstrated to show how it could be utilised to extract the following insights:

1. Being able to visualise (and compare) team formations at time of centre bounces and stoppages
2. Being able to quantitatively analyse team structures (e.g. how spread out the team is) and to monitor how this changes in attack/defence
3. Real-time¹⁵ monitoring of the time and location of interchanges/-substitutions (where players come in from, and where the new player moves out to)
4. Being able to easily define new metrics/statistics (e.g. team shape length-to-width ratio) and breakdowns (e.g. comparing attack/defence)

¹⁵The research prototype used historic data; however, enhancing it to process real-time data streams would just be a matter of applying the additional software development effort to implement this feature.

5. Being able to verify and investigate analysis results by tracing them back to the underlying video footage and GPS visualisations used to calculate them

Being able to *quantitatively analyse team structures (e.g. how spread out the team is) and to monitor how this changes in attack/defence* was identified by all four of the analysts as the most important insight from this project. They expressed interest in not only being able to look at the team spread, but also how this changes over time.

When asked whether the GPS visualisation provided benefits over existing approaches, two answered “yes”, and two answered “maybe”. However, due to the way this question was asked, it is possible that the “maybe” responses were talking about the visualisation itself rather than the additional analytics that are possible by building on top of this as a platform.

The interest of sport performance analysts shows that there is value in using GPS data for team level strategic analysis, and that the work in this thesis has made this form of analysis more accessible.

7.4 Chapter Summary

This chapter brought the components presented in previous chapters together into a platform for spatio-temporal sport analysis. It demonstrated the value of the platform as a means to rapidly form hypotheses, quantitatively investigate team formations, and to facilitate tracing results back to the underlying data to ensure all sport practitioners can understand and audit results of the analysis. While some of the patterns found in this chapter were only weakly statistically significant, and limited by the observational nature of sport analysis, the results can still be insightful for sport practitioners, and can serve as a basis upon which to form hypotheses to test through trialling strategies in real games.

This work is distinct from prior work in that:

1. The analysis is powered by only non-identifiable data. This facilitates ethical data sharing.
2. The player tracking GPS data input is expressed as latitude, longitude points, which are then transformed to the coordinate system of the field. This allows any GPS tracking device to be used, making the technique accessible to sub-elite teams who cannot afford a local positioning system.
3. All analysis results can be traced back to the underlying data, thus allowing sport practitioners to investigate interesting patterns identified in greater detail and to develop trust in the system through an ability to scrutinise the results.

Limitations

The analysis in Sec. 7.2 was limited to five home matches. While each match contains many events, the five home matches selected may not be representative of other matches. In particular, they may be limited by player availability, and the playing style may differ against different opponents. Moreover they were all from the perspective of the same club (Geelong Football Club) and played at the same venue (Kardinia Park).

Future Work

From a research perspective, future work is needed to run the analysis on a larger number of matches, and to work with other AFL teams. The pipeline constructed in this chapter will process data for any matches available, and (due to the re-projection step) allows merging of match data from multiple venues. Thus while this thesis has solved the technical challenges involved, running a larger study will require development of a collaboration approach to unite all stakeholders involved.

From an applied perspective, future work is needed to translate the proof of concept developed in this thesis into a production ready tool that sport analysts can apply. While each *component* has been designed to consider the usability needs of sport performance analysts, the system *as a whole* is not currently in a state that sport performance analysts could install the software and use it themselves. Thus additional engineering effort is needed to simplify the deployment and improve the consistency of the user interface. Furthermore, while de-identification is important when sharing data outside the club, internally the analysis is of most value when the club can trace the results to the individual player level so they can deliver personalised feedback. Thus there is a need for clubs to be able to disable the data de-identification step when running analysis for their own internal use (i.e. without randomising player order and without downsampling the data).

As the system aims to provide a *platform* for analysis, rather than limiting sport performance analysts to just the analysis procedures demonstrated in Sec. 7.2, future work is needed to design an end-user programming environment that will support sport performance analysts to build new functionality themselves. The analysis procedures for Sec. 7.2 were implemented as Python scripts within a Jupyter notebook¹⁶ that operated upon the data generated by the processing pipeline. Future work could develop a Domain Specific Visual Language (DSVL) to assist sport performance analysts to graphically express queries over the refined dataset in order to alleviate the need to learn Python. The notation developed in Chapter 4 could be utilised as part of the language as a way to assist sport performance analysts to construct deeper analysis pipelines that build further upon each other.

Finally, future sport science research is needed to build upon and extend the platform. For example, extending the analysis to multiple teams to study interactions, and to evaluate the information gain from incorporating team spread into the analysis. This is elaborated further in Sec. 8.3.

¹⁶Software: Jupyter <https://jupyter.org/> Accessed: 2019-11-25

Contributions

1. Developed a platform for spatio-temporal analysis of team sport, demonstrated through AFL as an example. Feedback from an elite AFL club confirms that the techniques proposed are likely to offer useful insights.
2. Performed an analysis of team shape and game speed in AFL. Team spread was found to strongly correlate with game speed.

Chapter 8

Conclusions

Contents

8.1 Contributions	263
8.2 Applications outside of Sport	265
8.2.1 Defence	265
8.2.2 Smart Homes	266
8.2.3 Intelligent Transport Systems	266
8.3 Future Work	267
8.3.1 Intervention Study	267
8.3.2 Probabilistic Approach	267
8.3.3 Two Team Perspective	268
8.3.4 Information Gain	268
8.3.5 Deep Sets	269
8.4 Closing Remark	270

This thesis set out to bridge the gap between raw position sensor measurements of individuals and high level strategic insights about group formations and behaviours. Doing so required a multi-disciplinary perspective spanning the fields of software engineering, metamodeling, information theory, data provenance, data privacy, cartography, information visualisation, and sport science. De-identification and spatio-temporal normalisation were identified as key transformations necessary to bridge this gap, which were poorly supported by existing approaches.

The approaches proposed to support these transformations were applied to develop a computational pipeline that de-identifies GPS player position tracking data, reprojects the data relative to the nearest sport field, and allows synchronisation with other available sources. This serves as a platform for spatio-temporal analysis, which supports both direct visual investigation, as well as quantitative spatio-temporal analysis such as the investigation of team spread and game speed.

8.1 Contributions

This thesis made the following contributions:

Elaborated in Chapter 3:

1. Structured AFL jargon into a formal domain model of consistent terminology, and used this to identify variables that form part of the game state. The full list of identified variables provides a holistic understanding of game state, and can increase awareness of the simplifying assumptions made by current sport analysis models.
2. Applied information theory to sport in order to provide a mathematically rigorous perspective for understanding the role of sport performance analysis systems within the larger sport context. Information theory was used to formalise the objective of performance analysis systems into a single formula, which states that the goal is to ensure information is valuable yet not already known to a coach, and incorporates the need to transmit this over limited human information channels.
3. Provided an abstract data model that permits modelling both dense and sparse spatio-temporal sport datasets, and draws attention to all required accuracy attributes that need to be specified in order to reason about the confidence of interpretations made from the dataset.

Elaborated in Chapter 4:

4. Provided an analysis of the data provenance needs of the sport domain, and evaluated existing data provenance tools against these criteria. A customised data provenance notation for sport was proposed in order to ease uptake for sport performance analysts without a computer science background.

Elaborated in Chapter 5:

5. Exposed the prevalence of improper de-identification methods used in sport research, and demonstrated that GPS player tracking data is particularly prone to re-identification. An interaction model was proposed to help improve ethical conduct of research by allowing the researcher to specify the de-identification operations in cases where the data custodian lacks the technical resources to strongly de-identify data themselves prior to data hand-over. The proposed approach was applied to GPS player tracking data held by an AFL club to obtain the non-identifiable data used in this thesis.

Elaborated in Chapter 6:

6. Proposed a novel method for representing spatio-temporal reference frames as geographic objects. This allows GIS novices, such as sport performance analysts, to configure reference frames without the need for deep conceptual knowledge of cartographic projections. It also facilitates partial automation (e.g. reprojecting GPS data to the closest sport field), thus resulting in time savings when the analysis involves multiple reference frames (e.g. a season of GPS tracking data involving multiple sport fields).

Elaborated in Chapter 7:

7. Developed a platform for spatio-temporal analysis of team sport, demonstrated through AFL as an example. Feedback from an elite AFL club confirms that the techniques proposed are likely to offer useful insights.
8. Performed an analysis of team shape and game speed in AFL. Team spread was found to strongly correlate with game speed.

8.2 Applications outside of Sport

Team sport serves as a test bed for understanding teams in a more general sense. Although the focus of this thesis was predominantly on sport, the methods can be adapted to other domains involving spatio-temporal data relating to group movement. The work in this thesis has influenced collaborative research publications in other areas such as defence, smart homes, and intelligent transport systems.

8.2.1 Defence

Defence departments prepare for possible scenarios through computerised war simulation that produces detailed output logging the simulated movements of each vehicle; however, require the ability to translate this simulated tracking data into a form amenable to extracting human insights into the underlying cause.¹

While traditionally the “team invasion” game classification is only applied to sport, it is also possible to think of a combat scenario as an invasion “game” (in the abstract sense). Reference frames could be established with respect to key areas such as shore lines to facilitate comparisons of different battles similar to the reference frames established at sporting venues in this thesis to facilitate comparisons between matches.²

¹Dion Grieger, Martin Wong, Antonio Giardina, Marco Tamassia, Luis Torres, Rakesh Vasa, Kon Mouzakis. “Towards the Identification and Visualisation of Causal Events to Support the Analysis of Closed-loop Combat Simulations”. In: ASOR National Conference for the Australian Society of Operations Research and Defence Operations Research Symposium (ASOR/DORS) 2018. Melbourne, 4–6 December 2018. (In Press)

²This work was published in a technical report sent to the client; however, it cannot be shared for confidentiality reasons.

8.2.2 Smart Homes

Internet-of-Things (IoT) enabled smart home systems use sensors placed around the house to respond to movements of individuals. However spurious sensors events can cause the smart-home to trigger messages unintentionally, so it is necessary to process sensor events via a real-time computational pipeline to infer higher-level patterns relating to behaviours, such as triggering a greeting when the resident wakes up, as opposed to triggering on individual motion sensors. Similarly to the team spread versus game speed analysis performed in this thesis, the smart-home situation could benefit from linking the degree to which recent sensor activations are spread out amongst the various rooms of the home to different forms of behaviours.

8.2.3 Intelligent Transport Systems

Transport engineering involves the use of scientific workflows to calibrate transport models with the demand and supply of the transport network from a combination of sensor and survey data. Similarly to the coach's desire to understand the impact of an altered sport team strategy, decision makers need to understand how changes to the transport network and/or traffic signal timings are likely to improve or degrade the overall performance. As with sport, visualising transport network data requires novel approaches that can account for the high-dimensional nature of spatio-temporal data [193].

Following the approach in this thesis to generate a customised data provenance notation for the sport domain, a custom data provenance notation for transport engineering could help facilitate documentation, reuse and traceability of transportation models. This could serve as a step towards automating analysis so that the system could detect and respond to transport network issues in real-time, thus helping to improve the overall efficiency of the network and relieve the load on traffic management centres who would otherwise need to manually intervene.

8.3 Future Work

8.3.1 Intervention Study

Future work is needed to integrate the system into sport practitioners' workflows so that the system can be formally evaluated within the target domain. Further investigating the relationships identified between team spread and game speed requires coaching interventions to establish causal relationships. Alternatively, a sport simulator could be built in which to trial the interventions.

8.3.2 Probabilistic Approach

Despite the theoretical accuracy attainable using GPS tracking devices under ideal conditions, the analysis of real-world data in Sec. 7.1.3 demonstrates operational issues with the devices that led to most matches being discarded. The ability to analyse the team as a whole requires that all 22 player tracking devices are functioning correctly and have a reliable signal, thus even a low chance of signal issues on a per-device basis can represent a high chance of issues at a team level. While technological developments such as local positioning systems may largely eliminate these issues for elite teams in future, there will always be some degree of positioning error, and issues are likely to remain for sub-elite teams who cannot afford to invest in more advanced technology. Therefore, an important area for future sport research would be to find techniques to design computational pipelines in a way that accounts for errors. Potential avenues include simulating errors to test the sensitivity of results to positioning errors, a Bayesian approach to infer the most likely behaviour of players given noisy sensor estimates³, and the use of robust statistics⁴ to reduce the impact

³A principled approach to this would be to model the prior distribution of movement patterns along with the sensor errors involved in observation of movement, then apply Bayesian inference over this model. John Winn and Christopher M. Bishop, *Model-Based Machine Learning*. <http://mbmlbook.com>

⁴Thank you to A/Prof. Tim Wilkin for this suggestion

of device malfunctions on performance measures.⁵

8.3.3 Two Team Perspective

As AFL clubs see their data as a competitive advantage, they are reluctant to share data, and unwilling to collaborate with other teams in this regard. As such, this thesis worked with only a single team's data, and demonstrated the value that could be extracted from this. As the team's behaviour depends on the behaviour of the opposition, deeper insights could be obtained with opposition GPS data. For example, the team spread analysis performed in this thesis could be calculated for both teams to investigate coupling between the spread of the two teams. Another area would be to detect pairings between players from opposite teams, for example players that are “tagging” (following after) a player of the opposite team.

8.3.4 Information Gain

The system developed in this thesis was presented to sport performance analysts at an elite AFL club to obtain feedback as to whether it provided useful insights (an aspect that is difficult to quantify). Alternatively, if a larger dataset were available, the value of team formation information over existing data could have been measured against the information theoretic objective set out in Sec. 3.3.

For example, O'Shaughnessy's possession versus possession map [162] estimates could be used as a baseline predictor of the expected outcome of a possession that does not incorporate team formations. Future research could test whether a modified version of the map that incorporates team formation data, e.g. an additional team spread attribute,

⁵An extreme example of a poor performance measure mentioned at the Victoria University Player Tracking Workshop 2019 is maximum player speed, as the maximum player speed recorded is likely to be a result of a measurement anomaly. In contrast, statistics like the median speed are largely unaffected by spurious measurements.

provides additional predictive benefits in terms of estimating the expected outcome of each possession. The value of this information could be measured according to the *information gain*⁶ offered by adding the team spread attribute as an additional feature to the model.⁷

8.3.5 Deep Sets

Due to the high-dimensional nature of team formation data, and limited number of events to train on, the analysis in this thesis reduced the team formation data to one dimension (team spread) when performing quantitative analysis in order to prevent overfitting. An alternative would be to apply machine learning techniques with regularisation terms. A challenge with this approach is that de-identified formations are represented as a set rather than the traditional vector input expected by most machine learning algorithms. While a set can be transformed into a vector, this is not ideal, as the arbitrary ordering may impact on the final result. Deep Sets [226] shows that neural networks can be modified to act on sets by constraining the weights such that the results of the neural network are permutation invariant to the ordering of the input vector. Furthermore, the permutation invariance constraint allows the Deep Sets approach to avoid overfitting and achieve better performance on smaller datasets than approaches that ignore this constraint.

⁶Wikipedia Contributors, 2019, “Information gain in decision trees” https://en.wikipedia.org/w/index.php?title=Information_gain_in_decision_trees&oldid=880605287 Accessed: 2019-02-25

⁷As this project was only working with data for one season for a single team, it is not reasonable to reproduce O’Shaughnessy’s possession versus possession map estimates that are composed of 350 000 data points taken over two AFL seasons (all teams). As adding attributes can cause the model to overfit, one would ideally use a larger dataset than O’Shaughnessy to ensure sufficient data after breaking the data down by the additional team spread attribute.

8.4 Closing Remark

Currently, access to AFL player tracking data is restricted due to commercial and privacy concerns, a situation that is also faced by those seeking to examine other types of sport.⁸ Even as a researcher within a university with agreements in place with both a club and the commercial providers involved, formally obtaining a single season of data for even one team was a lengthy process. This thesis should help ease concerns around data sharing through demonstration that it is possible to extract meaningful team-level insights without compromising individual privacy. While currently spatio-temporal analysis of sport data is only accessible to those with strong technical capabilities, the development of flexible computational pipelines offers the opportunity to empower sport practitioners (and fans) with the tools to freely share and extend analyses so that they can see the game from a new perspective.

⁸Thank you to the researchers working with AFL data that discussed these issues with me, as well as the two Association Football analysts who independently reached out to me to discuss data sharing and analysis reuse issues that they face

References

- [1] Martín Abadi et al. “Deep Learning with Differential Privacy”. In: *23rd ACM Conference on Computer and Communications Security*. 2016. ISBN: 9781450341394. doi: 10.1145/2976749.2978318. arXiv: 1607.00133.
- [2] AFL. *The Coach - The official AFL Level 1 coaching manual*. AFL, 2015.
- [3] Ian Alexander. “Misuse cases: Use cases with hostile intent”. In: *IEEE Software* 20.1 (2003), pp. 58–66. ISSN: 07407459. doi: 10.1109/MS.2003.1159030.
- [4] Jeremy P. Alexander et al. “Collective team behaviour of Australian Rules football during phases of match play”. In: *Journal of Sports Sciences* 37.3 (2019), pp. 237–243. doi: 10.1080/02640414.2018.1491113.
- [5] Jeremy P. Alexander et al. “The influence of match phase and field position on collective team behaviour in Australian Rules football”. In: *Journal of Sports Sciences* 37.15 (2019), pp. 1699–1707. doi: 10.1080/02640414.2019.1586077.
- [6] E. W. Anderson et al. “A user study of visualization effectiveness using EEG and cognitive load”. In: *Computer Graphics Forum* 30.3 (2011), pp. 791–800. ISSN: 14678659. doi: 10.1111/j.1467-8659.2011.01928.x.
- [7] Lorin W. Anderson et al. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives, Complete Edition*. English. Longman, 2001. ISBN: 0321084055.

- [8] Gennady Andrienko et al. “Visualization of Trajectory Attributes in Space–Time Cube and Trajectory Wall”. In: *Cartography from Pole to Pole SE - 11*. Lecture Notes in Geoinformation and Cartography. Springer, Berlin, Heidelberg, 2014, pp. 157–163. ISBN: 978-3-642-32617-2. doi: [10.1007/978-3-642-32618-9_11](https://doi.org/10.1007/978-3-642-32618-9_11).
- [9] R. C. Archibald. “Time as a fourth dimension”. In: *Bulletin of the American Mathematical Society* 20.8 (1914), pp. 409–412.
- [10] Robert J. Aughey. “Applications of GPS technologies to field sports”. In: *International Journal of Sports Physiology and Performance* 6.3 (Sept. 2011), pp. 295–310. ISSN: 15550265. doi: [10.1123/ijspp.6.3.295](https://doi.org/10.1123/ijspp.6.3.295).
- [11] Arnold Baca. “Feedback systems”. In: *Computers in Sport*. 1st ed. Vol. 1. WIT Press, Apr. 2008, pp. 43–67. ISBN: 9781845640644. doi: [10.2495/978-1-84564-064-4/02](https://doi.org/10.2495/978-1-84564-064-4/02).
- [12] Khaled Bachour et al. “Provenance for the People: An HCI Perspective on the W3C PROV Standard through an Online Game”. In: *33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. 2015. doi: [10.1145/2702123.2702455](https://doi.org/10.1145/2702123.2702455).
- [13] Monya Baker and Dan Penny. “Is there a reproducibility crisis?” In: *Nature* 533.7604 (May 2016), pp. 452–454. ISSN: 14764687. doi: [10.1038/533452A](https://doi.org/10.1038/533452A).
- [14] Michael Bar-Eli et al. “Developing peak performance in sport: optimization versus creativity”. In: *Essential processes for attaining peak performance*. Vol. 1. 2006, pp. 158–177.
- [15] R. M. L. Barros. “Automatic Tracking of Soccer Players”. In: *Proceedings of ISB XVIII*. 2001, pp. 236–239.
- [16] Ricardo M. L. Barros et al. “Analysis of the distances covered by first division Brazilian soccer players obtained with an automatic tracking method”. In: *Journal of Sports Science and Medicine* 6.2 (June 2007), pp. 233–242. ISSN: 13032968.
- [17] Belfrit Victor Batlajery et al. “Belief Propagation Through Provenance Graphs”. In: *Provenance Week '18: 7th International Provenance And Annotation Workshop*. London, United Kingdom, 2018.

- [18] Benjamin Baum et al. “Opinion paper: Data provenance challenges in biomedical research”. In: *it - Information Technology* 59.4 (2017), pp. 191–196. doi: 10.1515/itit-2016-0031.
- [19] Jöran Beel and Bela Gipp. “Google scholar’s ranking algorithm: The impact of citation counts (an empirical study)”. In: *Proceedings of the 2009 3rd International Conference on Research Challenges in Information Science, RCIS 2009* (2009), pp. 439–446. doi: 10.1109/RCIS.2009.5089308.
- [20] Michael Beetz, Bernhard Kirchlechner, and Martin Lames. “Computerized real-time analysis of football games”. In: *IEEE Pervasive Computing* 4.3 (July 2005), pp. 33–39. issn: 15361268. doi: 10.1109/MPRV.2005.53.
- [21] Michael Beetz et al. “Camera-based observation of football games for analyzing multi-agent activities”. In: *Proceedings of the fifth international joint conference on Autonomous agents and multi-agent systems - AAMAS ’06*. AAMAS ’06. New York, NY, USA: ACM, 2006, p. 42. isbn: 1595933034. doi: 10.1145/1160633.1160638.
- [22] Alina Bialkowski et al. “Large-scale analysis of soccer matches using spatiotemporal tracking data”. In: *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 725–730. doi: 10.1109/ICDM.2014.133.
- [23] Jérôme Bourbousson, Carole Sève, and Tim McGarry. “Space-time coordination dynamics in basketball: Part 2. The interaction between the two teams”. In: *Journal of Sports Sciences* 28.3 (2010), pp. 349–358. issn: 02640414. doi: 10.1080/02640410903503640.
- [24] Shawn Bowers et al. “A Model for User-Oriented Data Provenance in Pipelined Scientific Workflows”. In: *Lecture Notes in Computer Science* 4145.4145 (2006), pp. 133–147. issn: 03029743. doi: 10.1007/11890850_15.
- [25] Calum Braham and Michael Small. “Complex networks untangle competitive advantage in Australian football”. In: *Chaos* 28.5 (2018). issn: 10541500. doi: 10.1063/1.5006986.

- [26] Justin Brickell and Vitaly Shmatikov. “The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing”. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), pp. 70–78. doi: [10.1145/1401890.1401904](https://doi.org/10.1145/1401890.1401904).
- [27] Peter Buneman et al. “Why and Where: A Characterization of Data Provenance”. In: *Proceedings of International Conference on Database Theory (ICDT) 1973* (2001), pp. 316–330. issn: 02698463. doi: [10.1007/3-540-44503-X_20](https://doi.org/10.1007/3-540-44503-X_20).
- [28] Steven P. Callahan et al. “VisTrails: Visualization meets Data Management”. In: *Proceedings of the 2006 ACM SIGMOD international conference on Management of data - SIGMOD '06* (2006), p. 745. issn: 07308078. doi: [10.1145/1142473.1142574](https://doi.org/10.1145/1142473.1142574).
- [29] Hu Cao, Ouri Wolfson, and Goce Trajcevski. “Spatio-temporal data reduction with deterministic error bounds”. In: *The VLDB Journal* 15.3 (Sept. 2006), pp. 211–228. issn: 1066-8888. doi: [10.1007/s00778-005-0163-7](https://doi.org/10.1007/s00778-005-0163-7).
- [30] Christopher Carling et al. “The Role of Motion Analysis in Elite Soccer”. In: *Sports Medicine* 38.10 (Nov. 2008), pp. 839–862. issn: 0112-1642. doi: [10.2165/00007256-200838100-00004](https://doi.org/10.2165/00007256-200838100-00004).
- [31] Daniel Cervone et al. “A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes”. In: *Journal of the American Statistical Association* 111.514 (Apr. 2016), pp. 585–599. issn: 1537274X. doi: [10.1080/01621459.2016.1141685](https://doi.org/10.1080/01621459.2016.1141685). arXiv: [1408.0777](https://arxiv.org/abs/1408.0777).
- [32] Dan Cervone et al. “POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data”. In: *SLOAN Sports Analytics Conference*. Vol. 28. 2014, pp. 1–9.
- [33] Adriane Chapman, Barbara Blaustein, and Chris Elsaesser. “Provenance-based Belief”. In: *Proceedings of the 2nd Conference on Theory and Practice of Provenance*. San Jose, California: USENIX Association, 2010, pp. 1–14.

- [34] Ching Chien Chen, Craig A. Knoblock, and Cyrus Shahabi. “Automatically and accurately conflating raster maps with orthoimagery”. en. In: *GeoInformatica* 12.3 (Sept. 2008), pp. 377–410. issn: 13846175. doi: 10.1007/s10707-007-0033-0.
- [35] Dongyao Chen, Kyong-Tak Cho, and Kang G. Shin. “Mobile IMUs Reveal Driver’s Identity From Vehicle Turns”. In: *arXiv e-prints* (2017). arXiv: 1710.04578.
- [36] James Cheney. “Program Slicing and Data Provenance”. In: *IEEE Data Engineering Bulletin* 30.4 (2007), pp. 22–28.
- [37] James Cheney, Amal Ahmed, and Umut A. Acar. “Provenance as dependency analysis”. In: *Mathematical Structures in Computer Science* 21.6 (2011), pp. 1301–1337. issn: 09601295. doi: 10.1017/S0960129511000211. arXiv: arXiv:0708.2173v2.
- [38] Nicholas R. Chrisman. “Rethinking Levels of Measurement for Cartography”. In: *Cartography and Geographic Information Systems* 25.4 (Jan. 1998), pp. 231–242. issn: 1050-9844. doi: 10.1559/152304098782383043.
- [39] Stephen R Clarke. “Computer Forecasting of Australian Rules Football for a Daily Newspaper”. In: *Journal of the Operational Research Society* 44.8 (Aug. 1993), pp. 753–759. issn: 0160-5682. doi: 10.1057/jors.1993.134.
- [40] Filipe M. Clemente et al. “An online tactical metrics applied to football game”. In: *Research Journal of Applied Sciences, Engineering and Technology* 5.5 (2013), pp. 1700–1719. issn: 20407459. doi: 10.19026/rjaset.5.4926.
- [41] Filipe M. Clemente et al. “Measuring collective behaviour in football teams: Inspecting the impact of each half of the match on ball possession”. In: *International Journal of Performance Analysis in Sport* 13.3 (2013), pp. 678–689. issn: 14748185. doi: 10.1080/24748668.2013.11868680.
- [42] Filipe M. Clemente et al. “Measuring Tactical Behaviour Using Technological Metrics: Case Study of a Football Game”. In: *International Journal of Sports Science & Coaching* 8.4 (Dec. 2013), pp. 723–739. issn: 1747-9541. doi: 10.1260/1747-9541.8.4.723.

- [43] Michael Compton, David Corsar, and Kerry Taylor. “Sensor data provenance: SSNO and PROV-O together at last”. In: *CEUR Workshop Proceedings* 1401 (2014), pp. 67–82. issn: 16130073.
- [44] Carlos Cotta et al. “A network analysis of the 2010 FIFA world cup champion team play”. en. In: *Journal of Systems Science and Complexity* 26.1 (Feb. 2013), pp. 21–42. issn: 10096124. doi: 10.1007/s11424-013-2291-2. arXiv: arXiv:1108.0261v1.
- [45] Micael S. Couceiro et al. “Dynamical stability and predictability of football players: The study of one match”. In: *Entropy* 16.2 (2014), pp. 645–674. issn: 10994300. doi: 10.3390/e16020645.
- [46] James Coventry. *Time and Space: Footy Tactics from Origins to AFL*. HarperCollins Publishers, 2015. isbn: 9780733333699.
- [47] S. J. D. Cox and N. J. Car. “PROV and Real Things”. In: *MODSIM2015, 21st International Congress on Modelling and Simulation* November (2015), pp. 620–626.
- [48] Yingwei Cui, Jennifer Widom, and Janet L. Wiener. *Tracing the lineage of view data in a warehousing environment*. Vol. 25. 2. 2000, pp. 179–227. isbn: 3060296103. doi: 10.1145/357775.357777.
- [49] Chris Culnane, Benjamin I. P. Rubinstein, and Vanessa Teague. “Health Data in an Open World”. In: *arXiv e-prints* (2017). arXiv: 1712.05627.
- [50] Cloe Cummins et al. “Global positioning systems (GPS) and microtechnology sensors in team sports: A systematic review”. In: *Sports Medicine* 43.10 (Oct. 2013), pp. 1025–1042. issn: 01121642. doi: 10.1007/s40279-013-0069-2.
- [51] Maheswaree Kissoon Curumsing. “Emotion-Oriented Requirements Engineering”. PhD thesis. Swinburne University of Technology, 2017.
- [52] Maheswaree Kissoon Curumsing et al. “Understanding the Impact of Emotions on Software: A Case Study in Requirements Gathering and Evaluation”. In: *Journal of Systems and Software* (2018). issn: 01641212. doi: 10.1016/j.jss.2018.06.077.

- [53] Yves-Alexandre De Montjoye et al. “Unique in the Crowd: The privacy bounds of human mobility”. In: *Scientific Reports* 3 (2013), pp. 1–5. issn: 20452322. doi: 10.1038/srep01376.
- [54] Tom De Nies et al. “Git2PROV: Exposing version control system content as W3C PROV”. In: *CEUR Workshop Proceedings* 1035 (2013), pp. 125–128. issn: 16130073.
- [55] David De Roure, Carole Goble, and Robert Stevens. “The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows”. In: *Future Generation Computer Systems* 25.5 (2009), pp. 561–567. issn: 0167739X. doi: 10.1016/j.future.2008.06.010.
- [56] Anthony Dearden, Yiannis Demiris, and Oliver Grau. “Tracking football player movement from a single moving camera using particle filters”. In: *3rd European Conference on Visual Media Production (CVMP 2006). Part of the 2nd Multimedia Conference 2006*. IEE, 2006, pp. 29–37. isbn: 0 86341 729 9. doi: 10.1049/cp:20061968.
- [57] Carla L Dellaserra, Yong Gao, and Lynda Ransdell. “Use of Integrated Technology in Team Sports”. In: *Journal of Strength and Conditioning Research* 28.2 (Feb. 2014), pp. 556–573. issn: 1064-8011. doi: 10.1519/JSC.0b013e3182a952fb.
- [58] S.L. Delp and J.P. Loan. “A computational framework for simulating and analyzing human and animal movement”. In: *Computing in Science & Engineering* 2.5 (Sept. 2000), pp. 46–55. issn: 15219615. doi: 10.1109/5992.877394.
- [59] Jiaxin Ding, Chien-Chun Ni, and Jie Gao. “Fighting Statistical Re-Identification in Human Trajectory Publication”. In: *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL’17*. New York, New York, USA: ACM Press, 2017. isbn: 9781450354905. doi: 10.1145/3139958.3140045.
- [60] Brendan Dolan-Gavitt et al. “Repeatable Reverse Engineering with PANDA”. In: *Proceedings of the 5th Program Protection and Reverse Engineering Workshop - PPREW-5*. New York, New York,

- USA: ACM Press, 2015, pp. 1–11. ISBN: 9781450336420. doi: 10.1145/2843859.2843867.
- [61] Pedro Domingos. “A few useful things to know about machine learning”. In: *Communications of the ACM* 55.10 (Oct. 2012), p. 78. issn: 00010782. doi: 10.1145/2347736.2347755.
- [62] Jake Downey. *The Singles Game: A framework for badminton, An attacking game*. 6 Pearce House 205/7 Junction Road London N19: J Downey, 1976.
- [63] Samuel Picton Drake. *Converting GPS coordinates ($\phi\lambda h$) to navigation coordinates (ENU)*. Tech. rep. Australian Defence Science and Technology Group, 2002.
- [64] Ricardo Duarte et al. “Sports Teams as Superorganisms”. In: *Sports Medicine* 42.8 (June 2012), p. 1. issn: 0112-1642. doi: 10.1007/BF03262285.
- [65] Jordi Duch, Joshua S. Waitzman, and Luís A. Nunes Amaral. “Quantifying the performance of individual players in a team activity”. In: *PLoS ONE* 5.6 (June 2010), e10937. issn: 19326203. doi: 10.1371/journal.pone.0010937.
- [66] Cynthia Dwork. “Differential privacy”. In: *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming* (2006), pp. 1–12. doi: 10.1007/11787006_1.
- [67] Cynthia Dwork et al. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: *Journal of Privacy and Confidentiality*. Vol. 7. 3. 2006, pp. 265–284. doi: 10.1007/11681878_14.
- [68] S. J. Edgecomb and K. I. Norton. “Comparison of global positioning and computer-based tracking systems for measuring player movement distance during Australian Football”. In: *Journal of Science and Medicine in Sport* 9.1-2 (May 2006), pp. 25–32. issn: 14402440. doi: 10.1016/j.jsams.2006.01.003.
- [69] Khaled El Emam et al. “A systematic review of re-identification attacks on health data”. In: *PLoS ONE* 6.12 (2011). issn: 19326203. doi: 10.1371/journal.pone.0028071.

- [70] Rafael F. Escamilla et al. “Kinematic comparisons of 1996 Olympic baseball pitchers”. In: *Journal of Sports Sciences* 19.9 (Jan. 2001), pp. 665–676. issn: 02640414. doi: 10 . 1080 / 02640410152475793.
- [71] Gerald I. Evenden. *Libproj4: A Comprehensive Library of Cartographic Projection Functions (Preliminary Draft)*. Tech. rep. Mar. 2005.
- [72] Adrian Mark Faccioni, Dean Russell Gray, and David Ronald Cameron. *(WO2002039363) Information System and Method*. 2002.
- [73] Christopher Fleet, Kimberly C. Kowal, and Petr Pridal. “Georeferencer: Crowdsourced georeferencing for map library collections”. en. In: *D-Lib Magazine* 18.11-12 (Nov. 2012). issn: 10829873. doi: 10 . 1045/november2012-fleet.
- [74] Hugo Folgado et al. “Length, width and centroid distance as measures of teams tactical performance in youth football”. In: *European Journal of Sport Science* 14.SUPPL.1 (2014), pp. 487–492. issn: 17461391. doi: 10 . 1080/17461391 . 2012 . 730060.
- [75] Eelke Folmer and Jan Bosch. “Architecting for usability: A survey”. In: *Journal of Systems and Software* 70.1-2 (2004), pp. 61–78. doi: 10 . 1016/S0164-1212(02)00159-0.
- [76] Donald Forbes. “Dynamic prediction of Australian Rules football using real time performance statistics”. PhD thesis. Swinburne University of Technology, 2006.
- [77] W. G. P. Frencken and K. A. P. M. Lemmink. “Team kinematics of small-sided soccer games: A systematic approach”. In: *Science and football VI: proceedings of the sixth world congress on science and football* (2008).
- [78] Wouter Frencken et al. “Oscillations of centroid position and surface area of soccer teams in small-sided games”. In: *European Journal of Sport Science* 11.4 (2011), pp. 215–223. issn: 17461391. doi: 10 . 1080/17461391 . 2010 . 499967.

- [79] Telmo Frias and Ricardo Duarte. “Man-to-man or zone defense? Measuring team dispersion behaviors in small-sided soccer games”. In: *Trends in Sport Sciences* 3.21 (2014), pp. 135–144. issn: 2299-9590.
- [80] William R. Fry. (*US6148262*) *Sports computer with GPS receiver and performance tracking capabilities*. 2000.
- [81] Xianyi Gao et al. “Elastic Pathing: Your Speed is Enough to Track You”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM New York, NY, USA, 2014, pp. 975–986. isbn: 9781450329682. doi: 10.1145/2632048.2632077. arXiv: 1401.0052.
- [82] Paul B. Gastin et al. “Quantification of tackling demands in professional Australian football using integrated wearable athlete tracking technology”. In: *Journal of Science and Medicine in Sport* 16.6 (2013), pp. 589–593. issn: 14402440. doi: 10.1016/j.jsams.2013.01.007.
- [83] Paul B. Gastin et al. “Tackle and impact detection in elite Australian football using wearable microsensor technology”. In: *Journal of Sports Sciences* 32.10 (2014), pp. 947–953. issn: 1466447X. doi: 10.1080/02640414.2013.868920.
- [84] P. Gatalsky, N. Andrienko, and G. Andrienko. “Interactive analysis of event data using space-time cube”. In: *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004*. July 2004, pp. 145–152. isbn: 0-7695-2177-0. doi: 10.1109/IV.2004.1320137.
- [85] Yolanda Gil et al. “Examining the challenges of scientific workflows”. In: *Computer* 40.12 (2007), pp. 24–32. issn: 00189162. doi: 10.1109/MC.2007.421.
- [86] Hubert F. M. Goenner. “On the History of Geometrization of Space-time: From Minkowski to Finsler Geometry”. In: *arXiv e-prints* (Nov. 2008). arXiv: 0811.4529.
- [87] Leo Goodstadt. “Ruffus: A lightweight Python library for computational pipelines”. In: *Bioinformatics* 26.21 (Sept. 2010), pp. 2778–2779. issn: 13674803. doi: 10.1093/bioinformatics/btq524.

- [88] Robert B. Gramacy, Shane T. Jensen, and Matt Taddy. “Estimating player contribution in hockey with regularized logistic regression”. In: *Journal of Quantitative Analysis in Sports* 9.1 (Mar. 2013), pp. 97–111. issn: 1559-0410. doi: 10.1515/jqas-2012-0001.
- [89] Grace Greenham et al. “A pilot study to measure game style within Australian football”. In: *International Journal of Performance Analysis in Sport* 17.4 (Sept. 2017), pp. 576–585. issn: 2474-8668. doi: 10.1080/24748668.2017.1372163.
- [90] Arofan Gregory. “The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes”. In: July (2011), pp. 1–10.
- [91] Thomas Von Der Grün et al. “A Real-Time Tracking System for Football Match and Training Analysis”. In: *Microelectronic Systems: Circuits, Systems and Applications*. Springer Berlin Heidelberg, 2011, p. 371. isbn: 9783642230707. doi: 10.1007/978-3-642-23071-4_19.
- [92] A. Grunz, D. Memmert, and J. Perl. “Analysis and Simulation of Actions in Games by Means of Special Self-Organizing Maps”. In: *International Journal of Computer Science in Sport* 8.1 (2009), pp. 22–37.
- [93] Joachim Gudmundsson and Michael Horton. “Spatio-Temporal Analysis of Team Sports”. In: *ACM Computing Surveys* 50.2 (Apr. 2017), pp. 1–34. issn: 03600300. doi: 10.1145/3054132. arXiv: 1602.06994.
- [94] Joachim Gudmundsson and Thomas Wolle. “Football analysis using spatio-temporal tools”. In: *Computers, Environment and Urban Systems*. Progress in Movement Analysis - Experiences with Real Data 47 (Sept. 2014), pp. 16–27. issn: 01989715. doi: 10.1016/j.comenvurbssys.2013.09.004.
- [95] Torsten Hägerstrand. “What About People in Regional Science?” In: *Papers in Regional Science* 24.1 (Jan. 1970), pp. 7–24. issn: 1435-5957. doi: 10.1111/j.1435-5597.1970.tb01464.x.

- [96] Ralf Herbrich, Tom Minka, and Thore Graepel. “TrueSkill: A Bayesian Skill Rating System”. In: *Advances in Neural Information Processing Systems*. MIT Press, 2007, pp. 569–576.
- [97] N. Hirotsu and M. Wright. “Using a markov process model of an association football match to determine the optimal timing of substitution and tactical decisions”. In: *Journal of the Operational Research Society* 53.1 (2002), pp. 88–96. issn: 14769360. doi: 10.1057/palgrave.jors.2601254.
- [98] Daniel T. Hoffman, Andrew J. Simmons, and Paul B. Gastin. “Investigating the relationship between injury and match outcome in Australian Football League matches”. In: *Proceedings 14th Australasian Conference on Mathematics and Computers in Sport (ANZIAM MathSport 2018)*. University of the Sunshine Coast, Queensland, Australia, 2018, p. 5. isbn: 978-0-646-99402-4.
- [99] Maarten Hooijberg. “Conformal Projections-Using Reference Ellipsoids”. In: *Geometrical Geodesy: Using Information and Computer Technology* (2008), pp. 183–243.
- [100] Ted Hopkins. “Pre-Computer Pioneers”. In: *The Stats Revolution*. The Slattery Media Group, 2011, pp. 90–108. isbn: 978-1-921778-20-9.
- [101] Weidong Huang, Peter Eades, and Seok Hee Hong. “Measuring effectiveness of graph visualizations: A cognitive load perspective”. In: *Information Visualization* 8.3 (2009), pp. 139–152. issn: 14738716. doi: 10.1057/ivs.2009.10.
- [102] M. D. Hughes. “A comparison of patterns of play in squash”. In: *International Ergonomics*. Bournemouth, England: Taylor & Francis, 1985, pp. 139–141. isbn: 0-85066-300-8.
- [103] Mike Hughes. “Notational analysis—a mathematical perspective.” In: *International Journal of Performance Analysis in Sport* 4.2 (2004), pp. 97–139.
- [104] Mike D. Hughes and Roger M. Bartlett. “The use of performance indicators in performance analysis”. In: *Journal of Sports Sciences* 20.10 (Jan. 2002), pp. 739–754. issn: 02640414. doi: 10.1080/026404102320675602.

- [105] Mike Hughes and Ian M. Franks. “The development of sport-specific notation systems (hand notation)”. In: *Notational analysis of sport: Systems for better coaching and performance in sport*. Routledge, 2004. Chap. 4.4, pp. 61–80.
- [106] Mike Hughes, Michael T. Hughes, and Hannah Behan. “The Evolution of Computerised Notational Analysis Through the Example of Racket Sports”. In: *International Journal of Sports Science and Engineering* 1.1 (2007), pp. 3–28. ISSN: 00314005.
- [107] Dunca Hull et al. “Taverna: A tool for building and running workflows of services”. In: *Nucleic Acids Research* 34.suppl_2 (July 2006), W729–W732. ISSN: 03051048. doi: 10.1093/nar/gkl320.
- [108] Robert Ikeda and Jennifer Widom. “Panda: A System for Provenance and Data”. In: *Proceedings of the 2nd USENIX Workshop on the Theory and Practice of Provenance TaPP’10* 33 (2010), pp. 1–8.
- [109] ISO/IEC 25010:2011, *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models*. 2011.
- [110] Karl Jackson. “A player rating system for Australian Rules Football using field equity measures”. In: *Mathematics and Computers in Sport*. 2008.
- [111] Karl Jackson. “Assessing player performance in Australian football using spatial data”. PhD thesis. Swinburne University of Technology, 2016.
- [112] Ysabel Jacob et al. “Identification of genetic markers for skill and athleticism in sub-elite Australian football players: a pilot study.” In: *The Journal of sports medicine and physical fitness* (2016), p. 2018. ISSN: 0022-4707 (Print). doi: 10.23736/S0022-4707.16.06647-0.
- [113] T. J. Jankun-Kelly, Kwan-Liu Ma, and Michael Gertz. “A Model and Framework for Visualization Exploration”. In: *IEEE Transactions on Visualization and Computer Graphics* 13.2 (2007), pp. 357–369. doi: 10.1109/TVCG.2007.28.

- [114] Bernhard Jenny. "MapAnalyst - A digital tool for the analysis of the planimetric accuracy of historical maps". In: *e-Perimetron* 1.3 (2006), pp. 239–245.
- [115] Bernhard Jenny and Lorenz Hurni. "Studying cartographic heritage: Analysis and visualization of geometric distortions". In: *Computers and Graphics (Pergamon)* 35.2 (Apr. 2011), pp. 402–411. ISSN: 00978493. doi: 10.1016/j.cag.2011.01.005.
- [116] Alexander Karminsky and Andrey Polozov. "Problem of Rating in Sports and Its Possible Solutions". In: *Handbook of Ratings*. Springer International Publishing, 2016, p. 156. ISBN: 978-3-319-39260-8. doi: 10.1007/978-3-319-39261-5.
- [117] Claire J.B. Kenneally-Dabrowski, Benjamin G. Serpell, and Wayne Spratford. "Are accelerometers a valid tool for measuring overground sprinting symmetry?" In: *International Journal of Sports Science and Coaching* 13.2 (2018), pp. 270–277. ISSN: 2048397X. doi: 10.1177/1747954117716790.
- [118] Donald E. Knuth. "Basketball's Electronic Coach". In: *Selected papers on fun & games*. Stanford, CA: CSLI Publications, 2011, pp. 199–208.
- [119] Matthew Allen Knutzen. "Unbinding the atlas: Moving the NYPL map collection beyond digitization". In: *Journal of Map and Geography Libraries* 9.1-2 (Jan. 2013), pp. 8–24. ISSN: 15420353. doi: 10.1080/15420353.2012.726204.
- [120] M. Kraak. "The space-time cube revisited from a geovisualization perspective". In: *Proceedings of the 21st International Cartographic Conference (ICC)*. Durban, South Africa, 2003, p. 1988. ISBN: 0-958-46093.
- [121] Thomas Kühne. "Matters of (meta-) modeling". In: *Software and Systems Modeling* 5.4 (2006), pp. 369–385. ISSN: 16191366. doi: 10.1007/s10270-006-0017-9.
- [122] Martin Lames and Tim McGarry. "On the search for reliable performance indicators in game sports". In: *International Journal of Performance Analysis in Sport* 7.1 (2007), pp. 62–79. ISSN: 1474-8185. doi: 10.1080/24748668.2007.11868388.

- [123] Noble G. Larson and Kent A. Stevens. (*US5363297*) *Automated camera-based tracking system for sports contests*. 1994.
- [124] Hoang M. Le, Carr Peter, and Yisong Yue. “Data-Driven Ghosting using Deep Imitation Learning”. In: *MIT Sloan Sports Analytics Conference* (2017), pp. 1–15.
- [125] Kristen LeFevre, David J D.J. DeWitt, and Raghu Ramakrishnan. “Incognito: efficient full-domain K-anonymity”. In: *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (2005), pp. 49–60.
- [126] Susan Lester et al. “The DD genotype of the angiotensin-converting enzyme gene occurs in very low frequency in Australian Aboriginals.” In: *Nephrology, dialysis, transplantation* 14.4 (1999), pp. 887–890. ISSN: 09310509. DOI: 10.1093/ndt/14.4.887.
- [127] Michael Lewis. *Moneyball: The Art of Winning an Unfair Game*. English. 1st editio. New York: W. W. Norton & Company, Mar. 2004, p. 320. ISBN: 0393324818.
- [128] Ninghui Li, Wahbeh Qardaji, and Dong Su. “On Sampling, Anonymization, and Differential Privacy: Or, k-Anonymization Meets Differential Privacy”. In: (2011). DOI: 10.1145/2414456.2414474. arXiv: 1101.2604.
- [129] Vitor Maçãs and J Sampaio. “Measuring Tactical Behaviour in Football”. In: *International Journal of Sports Medicine* 33 (2012), pp. 395–401.
- [130] Ashwin Machanavajjhala et al. “l-diversity”. In: *ACM Transactions on Knowledge Discovery from Data* 1.1 (2007), 3–es. ISSN: 15564681. DOI: 10.1145/1217299.1217302.
- [131] Ralph Maddison and Cliona Ni Mhurchu. “Global positioning system: A new opportunity in physical activity measurement”. In: *International Journal of Behavioral Nutrition and Physical Activity* 6 (2009), p. 73. ISSN: 14795868. DOI: 10.1186/1479-5868-6-73.

- [132] Bradley Malin. "A computational model to protect patient data from location-based re-identification". In: *Artificial Intelligence in Medicine* 40.3 (2007), pp. 223–239. issn: 09333657. doi: 10.1016/j.artmed.2007.04.002.
- [133] Bradley Malin and Edoardo Airoldi. "The effects of location access behavior on re-identification risk in a distributed environment". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4258 LNCS (2006), pp. 413–429. issn: 03029743. doi: 10.1007/11957454_24.
- [134] Phillip Mates et al. "CrowdLabs: Social analysis and visualization for the sciences". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6809 LNCS (2011), pp. 555–564. issn: 03029743. doi: 10.1007/978-3-642-22351-8_38.
- [135] Tim McGarry and Ian M. Franks. "Development, application, and limitation of a stochastic markov model in explaining championship squash performance". In: *Research Quarterly for Exercise and Sport* 67.4 (Dec. 1996), pp. 406–415. issn: 21683824. doi: 10.1080/02701367.1996.10607972.
- [136] Tim McGarry et al. "Sport competition as a dynamical self-organizing system". In: *Journal of Sports Sciences* 20.10 (Jan. 2002), pp. 771–781. issn: 02640414. doi: 10.1080/026404102320675620.
- [137] Antonette Mendoza et al. "Software Appropriation Over Time : From Adoption to Stabilization and Beyond". In: *Australasian Journal of Information Systems* 16.2 (2010), pp. 5–23. issn: 1039-7841. doi: 10.3127/ajis.v16i2.507.
- [138] Antonette Mendoza et al. "The role of users' emotions and associated quality goals on appropriation of systems: Two case studies". In: *Proceedings of the 24th Australasian Conference on Information Systems* (2013).
- [139] Lloyd L. Messersmith and Stephen M. Corey. "The distance traversed by a basketball player". In: *Research Quarterly of the American Physical Education Association* 2.2 (1931), pp. 57–60.

- [140] Paolo Missier et al. “D-PROV: Extending the PROV Provenance Model with Workflow Structure”. In: *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance* (2013), 9:1–9:7. doi: 10.1145/2457317.2457375.
- [141] Milton Shoiti Misuta et al. “Representation and analysis of soccer players’ trajectories”. In: *XXth Congress of the International Society of Biomechanics, Cleveland, USA*. Vol. 415. 2005.
- [142] David Moher et al. “Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement”. In: *PLoS Medicine* 6.7 (2009). doi: 10.1371/journal.pmed.1000097.
- [143] Kaisu Mononen. “The effects of augmented feedback on motor skill learning in shooting : a feedback training intervention among inexperienced rifle shooters”. PhD thesis. University of Jyväskylä, 2007, p. 63.
- [144] Daniel Moody. “The physics of notations: Toward a scientific basis for constructing visual notations in software engineering”. In: *IEEE Transactions on Software Engineering* 35.6 (Nov. 2009), pp. 756–779. issn: 00985589. doi: 10.1109/TSE.2009.67.
- [145] Luc Moreau and Paul Groth. *Provenance: An Introduction to PROV*. 2013. isbn: 9781627052214. doi: 10 . 2200 / S00528ED1V01Y201308WBE007.
- [146] Luc Moreau et al. “A Templating System to Generate Provenance”. In: *IEEE Transactions on Software Engineering* 44.2 (2018), pp. 103–121. issn: 00985589. doi: 10.1109/TSE.2017.2659745.
- [147] Luc Moreau et al. “The rationale of PROV”. In: *Journal of Web Semantics* 35 (2015), pp. 235–257. issn: 15708268. doi: 10 . 1016/j.websem.2015.04.001.
- [148] Ranjit Nair et al. “Automated assistants for analyzing team behaviors”. In: *Autonomous Agents and Multi-Agent Systems* 8.1 (2004), pp. 69–111. issn: 13872532. doi: 10 . 1023 / B : AGNT . 0000009411.79208.f4.

- [149] Arvind Narayanan and Vitaly Shmatikov. “Robust de-anonymization of large sparse datasets”. In: *Proceedings - IEEE Symposium on Security and Privacy* (2008), pp. 111–125. ISSN: 10816011. doi: 10.1109/SP.2008.33.
- [150] National Health and Medical Research Council Australian Research Council. *National Statement on Ethical Conduct National Statement on Ethical Conduct in Human Research 2007 (Updated May 2015)*. May. 2015, pp. 1–96. ISBN: 1864962690.
- [151] Jonathon Neville et al. “Accelerometers: An underutilized resource in sports monitoring”. In: *Proceedings of the 2010 6th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*. Dec. 2010, pp. 287–290. ISBN: 9781424471768. doi: 10.1109/ISSNIP.2010.5706766.
- [152] Bob Quan-Minh Ngo. “Stats Geeks: Sabermetrics, Baseball Fans, and the Struggle over Masculinity”. PhD thesis. UC Santa Barbara, 2012.
- [153] Jakob Nielsen. “Enhancing the explanatory power of usability heuristics”. In: *Conference companion on Human factors in computing systems - CHI '94* (1994), p. 210. doi: 10.1145/259963.260333.
- [154] Jakob Nielsen and Rolf Molich. “Heuristic Evaluation of user interfaces”. In: *CHI '90 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems April* (1990), pp. 249–256. ISSN: 1942-597X. doi: 10.1145/97243.97281.
- [155] Li Ninghui, Li Tiancheng, and Suresh Venkatasubramanian. “t-Closeness: Privacy beyond k-anonymity and l-diversity”. In: *Proceedings - International Conference on Data Engineering 3* (2007), pp. 106–115. ISSN: 10844627. doi: 10.1109/ICDE.2007.367856.
- [156] Donald A Norman. *Emotional design: Why we love (or hate) everyday things*. Basic Books, 2004.
- [157] Aboelmagd Noureldin, Tashfeen B Karamat, and Jacques Georgy. “Basic Navigational Mathematics, Reference Frames and the Earth’s Geometry”. In: *Fundamentals of Inertial Navigation*,

Satellite-based Positioning and their Integration. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 21–63. ISBN: 978-3-642-30466-8.

- [158] Christine M. O'Keefe et al. "Anonymization for outputs of population health and health services research conducted via an online data center". In: *Journal of the American Medical Informatics Association* 24.3 (2017), pp. 544–549. ISSN: 1527974X. doi: 10.1093/jamia/ocw152.
- [159] Christine M. O'Keefe et al. "Assessing privacy risks in population health publications using a checklist-based approach". In: *Journal of the American Medical Informatics Association* 25.3 (2018), pp. 315–320. ISSN: 1527974X. doi: 10.1093/jamia/ocx129.
- [160] Christine M. O'Keefe et al. *The De-Identification Decision-Making Framework*. Tech. rep. EP173122 and EP175702. CSIRO, 2017.
- [161] Darren O'Shaughnessy. "Identification and measurement of luck in sport". In: *The proceedings of the 13th Australasian conference on mathematics and computers in sport*. Victoria University, Melbourne, Australia: ANZIAM MathSport, July 2016. ISBN: 978-0-646-95741-8.
- [162] Darren M O'Shaughnessy. "Possession Versus position: Strategic evaluation in AFL". In: *Journal of Sports Science and Medicine* 5.4 (2006), pp. 533–540. ISSN: 13032968.
- [163] Paul Ohm. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization". In: *UCLA Law Review* 57.6 (2010), pp. 1701–1777. ISSN: 00415650.
- [164] Tom Oinn et al. "Taverna: A tool for the composition and enactment of bioinformatics workflows". In: *Bioinformatics* 20.17 (Nov. 2004), pp. 3045–3054. ISSN: 13674803. doi: 10.1093/bioinformatics/bth361.
- [165] David Parmenter. *Key Performance Indicators: Developing Implementing and Using Winning KPIs*. 2010.

- [166] Jon Patrick. “The CABER project: the capture and analysis of behavioural events in real-time”. In: *ACM ’85: Proceedings of the 1985 ACM annual conference on The range of computing : mid-80’s perspective*. ACM, 1985, pp. 92–98. ISBN: 0-89791-170-9. DOI: 10.1145/320435.320466.
- [167] Edzer Pebesma. “spacetime : Spatio-Temporal Data in R”. In: *Journal of Statistical Software* 51.7 (2012), pp. 1–30. DOI: 10.18637/jss.v051.i07.
- [168] Silvio Peroni, David Shotton, and Fabio Vitali. “One Year of the OpenCitations Corpus”. In: *International Semantic Web Conference*. Springer, 2017, pp. 184–192. ISBN: 978-3-319-68203-7. DOI: 10.1007/978-3-319-68204-4_19.
- [169] J. Pers and S. Kovacic. “Computer vision system for tracking players in sports games”. In: *Proceedings of the First International Workshop on Image and Signal Processing and Analysis (TWISPA)*. Vol. 67. 5. Univ. Zagreb, 2000, pp. 177–182. ISBN: 953-96769-2-4. DOI: 10.1109/ISPA.2000.914910.
- [170] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.
- [171] Sarvapali D. Ramchurn et al. “A Disaster Response System based on Human-Agent Collectives”. In: *Journal of Artificial Intelligence Research* 57 (Dec. 2016), pp. 661–708. ISSN: 1076-9757. DOI: 10.1613/jair.5098.
- [172] Isabel Ramos, Daniel M. Berry, and João Á Carvalho. “Requirements engineering for organizational transformation”. In: *Information and Software Technology* 47.7 (2005), pp. 479–495. ISSN: 09505849. DOI: 10.1016/j.infsof.2004.09.014.
- [173] Blake Regalia, Krzysztof Janowicz, and Song Gao. “VOLT: A Provenance-Producing, Transparent SPARQL Proxy for the On-Demand Computation of Linked Data and its Application to Spatiotemporally Dependent Data”. In: *The Semantic Web. Latest Advances and New Domains* 9678 (2016), pp. 523–538. DOI: 10.1007/978-3-319-34129-3_32.

- [174] Sam Robertson, Carl Woods, and Paul Gastin. "Predicting higher selection in elite junior Australian Rules football: The influence of physical performance and anthropometric attributes". In: *Journal of Science and Medicine in Sport* 18.5 (2015), pp. 601–606. issn: 18781861. doi: 10.1016/j.jsams.2014.07.019.
- [175] Danny Ryan. "The five fundamentals of modern football". In: *Coaching Edge* 25.2 (Dec. 2011).
- [176] David E. M. Sappington. "Incentives in Principal-Agent Relationships". In: *The Journal of Economic Perspectives* 5.2 (1991), pp. 45–66. issn: 0895-3309.
- [177] Thuraiappah Sathyam, David Humphrey, and Mark Hedley. "WASP: A system and algorithms for accurate radio localization using low-cost hardware". In: *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 41.2 (Mar. 2011), pp. 211–222. issn: 10946977. doi: 10.1109/TSMCC.2010.2051027.
- [178] Bojan Šavrič, Bernhard Jenny, and Helen Jenny. "Projection Wizard – An Online Map Projection Selection Tool". In: *Cartographic Journal* 53.2 (Apr. 2016), pp. 177–185. issn: 17432774. doi: 10.1080/00087041.2015.1131938.
- [179] Markus Schneider. "Moving Objects in Databases and GIS: State-of-the-Art and Open Problems". In: *Lecture Notes in Geoinformation and Cartography*. 199069. Springer, 2009, pp. 169–187. isbn: 9783540882442. doi: 10.1007/978-3-540-88244-2_12.
- [180] Y. Schutz and A. Chambaz. "Could a satellite-based navigation system (GPS) be used to assess the physical activity of individuals on earth?" In: *European Journal of Clinical Nutrition* 51.5 (1997), pp. 338–339. issn: 09543007. doi: 10.1038/sj.ejcn.1600403.
- [181] Thomas Seidl et al. "Bhostgusters : Realtime Interactive Play Sketching with Synthesized NBA Defenses". In: *MIT Sloan Sports Analytics Conference*. 2018, pp. 1–13.

- [182] Thomas Seidl et al. “Evaluating the indoor football tracking accuracy of a radio-based real-time locating system”. In: *Advances in Intelligent Systems and Computing*. Vol. 392. Advances in Intelligent Systems and Computing. Springer International Publishing, 2016, pp. 217–224. ISBN: 9783319245584. doi: 10.1007/978-3-319-24560-7_28.
- [183] D. Setterwall. “Computerised video analysis of football-technical and commercial possibilities for football coaching”. Master’s Thesis. Stockholm University, 2003.
- [184] Nigel Shadbolt, Wendy Hall, and Tim Berners-Lee. “The semantic web revisited”. In: *IEEE Intelligent Systems* 21.3 (2006), pp. 96–101. ISSN: 15411672. doi: 10.1109/MIS.2006.62.
- [185] C. E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. ISSN: 00058580. doi: 10.1002/j.1538-7305.1948.tb01338.x.
- [186] Pedro Silva et al. “Sports teams as complex adaptive systems: manipulating player numbers shapes behaviours during football small-sided games”. In: *SpringerPlus* 5.1 (2016), pp. 1–10. ISSN: 21931801. doi: 10.1186/s40064-016-1813-5.
- [187] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *Nature* 529.7587 (Jan. 2016), pp. 484–489. ISSN: 14764687. doi: 10.1038/nature16961.
- [188] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. “A Survey of Data Provenance in e-Science”. In: *ACM SIGMOD Record* 34.3 (2005), pp. 31–36. ISSN: 0163-5808. doi: 10.1145/1084805.1084812.
- [189] Andrew J. Simmons, Maheswaree Kissoon Curumsing, and Rajesh Vasa. “An interaction model for de-identification of human data held by external custodians”. In: *Proceedings of the 30th Australian Conference on Human-Computer Interaction*. Melbourne, VIC, Australia: ACM, 2018, pp. 23–26. doi: 10.1145/3292147.3292207.
- [190] Andrew J. Simmons et al. “Data Provenance for Sport”. In: *arXiv e-prints* (2018). arXiv: 1812.05804.

- [191] Andrew Simmons and Leonard Hoon. “Agree to Disagree: On Labelling Helpful App Reviews”. In: *Proceedings of the 28th Australian Conference on Computer-Human Interaction*. OzCHI ’16. Launceston, TAS, Australia: ACM, 2016, pp. 416–420. ISBN: 978-1-4503-4618-4. doi: 10.1145/3010915.3010976.
- [192] Andrew Simmons and Rajesh Vasa. “Spatio-Temporal Reference Frames as Geographic Objects”. In: *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL’17*. Redondo Beach, CA, USA: ACM, 2017, pp. 1–4. ISBN: 9781450354905. doi: 10.1145/3139958.3139983.
- [193] Andrew Simmons et al. “Hub Map: A new approach for visualizing traffic data sets with multi-attribute link data”. In: *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. Atlanta, GA, USA, Oct. 2015, pp. 219–223. doi: 10.1109/VLHCC.2015.7357220.
- [194] Daniel J Simons and Christopher F Chabris. “Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events”. In: *Perception* 28.9 (1999), pp. 1059–1074. ISSN: 03010066. doi: 10.1088/p281059.
- [195] Nicholas J. Smeeton et al. “The relative effectiveness of various instructional approaches in developing anticipation skill”. In: *Journal of Experimental Psychology: Applied* 11.2 (2005), pp. 98–110. ISSN: 1076898X. doi: 10.1037/1076-898X.11.2.98.
- [196] John Snyder. *Map Projections—A Working Manual*. Vol. 1395. US Government Printing Office, 1987.
- [197] Bartholomew Spencer et al. “A method for evaluating player decision-making in the Australian Football League”. In: *Proceedings 14th Australasian Conference on Mathematics and Computers in Sport (ANZIAM Mathsport 2018)*. University of the Sunshine Coast, Queensland, Australia, 2018, pp. 7–12. ISBN: 978-0-646-99402-4.
- [198] Manolis Stamatogiannakis et al. “PROV2R: Practical Provenance Analysis of Unstructured Processes”. In: *ACM Transactions on*

- Internet Technology (TOIT) 17.4* (2017). ISSN: 15576051. DOI: 10.1145/3062176.
- [199] S.S. Stevens. *On the Theory of Scales of Measurement*. Vol. 103. 2684. Bobbs-Merrill, College Division, 1946, pp. 677–680. ISBN: 2819460607. DOI: 10.1126/science.103.2684.677.
- [200] Michael Stöckl, Peter F. Lamb, and Martin Lames. “The ISOPAR method: A new approach to performance analysis in golf”. In: *Journal of Quantitative Analysis in Sports* 7.1 (Jan. 2011). ISSN: 15590410. DOI: 10.2202/1559-0410.1289.
- [201] Michael Stöckl and Stuart Morgan. “Visualization and analysis of spatial characteristics of attacks in field hockey”. In: *International Journal of Performance Analysis in Sport* 13.1 (2013), pp. 160–178. DOI: 10.1080/24748668.2013.11868639.
- [202] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 1st ed. Cambridge, Mass: A Bradford Book, Mar. 1998. ISBN: 978-0-262-19398-6.
- [203] Latanya Sweeney. “k-anonymity: a model for protecting privacy”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570. DOI: 10.1142/S0218488502001648.
- [204] T. Taki, J. Hasegawa, and T. Fukumura. “Development of motion analysis system for quantitative evaluation of teamwork in soccer games”. In: *Proceedings of 3rd IEEE International Conference on Image Processing*. Vol. 3. Sept. 1996, pp. 815–818. ISBN: 0-7803-3259-8. DOI: 10.1109/ICIP.1996.560865.
- [205] Michael Tamir and Gal Oz. (WO2006103662) *Real-time objects tracking and motion capture in sports events*. Apr. 2009.
- [206] Wang-Chiew Tan. “Provenance in Databases: Past, Current, and Future”. In: *IEEE Data Engineering Bulletin* 30.4 (2007), pp. 3–12.
- [207] Philippe Terrier et al. “High-precision satellite positioning system as a new tool to study the biomechanics of human locomotion”. In: *Journal of Biomechanics* 33.12 (Dec. 2000), pp. 1717–1722. ISSN: 00219290. DOI: 10.1016/S0021-9290(00)00133-0.

- [208] Sumiyo Toki and Shinji Sakurai. "Quantitative match analysis of soccer games with two dimensional DLT procedures". In: *XXth Congress of International Society of Biomechanics*. 2005, p. 5886.
- [209] United States Coast Guard. *NAVSTAR GPS User Equipment Introduction, Public Release Version*. 1996, pp. 1–1.
- [210] Di Salvo Valter et al. "Validation of Prozone: A new video-based performance analysis system". In: *International Journal of Performance Analysis in Sport* 6.1 (2006), pp. 108–119. ISSN: 2474-8668. doi: 10.1080/24748668.2006.11868359.
- [211] Mary Vardigan et al. "Creating rich, structured metadata: lessons learned in the Metadata Portal Project". In: *IASSIST Quarterly* 38.3 (2014), pp. 15–20. ISSN: 07391137.
- [212] Hal R. Varian. "Designing the perfect auction". In: *Communications of the ACM* 51.8 (Aug. 2008), p. 9. ISSN: 00010782. doi: 10.1145/1378704.1378708.
- [213] Ben Vershbow. "NYPL Labs: Hacking the Library". In: *Journal of Library Administration* 53.1 (Jan. 2013), pp. 79–96. ISSN: 01930826. doi: 10.1080/01930826.2013.756701.
- [214] Daniel Vlasic et al. "Practical motion capture in everyday surroundings". In: *ACM Transactions on Graphics* 26.3 (July 2007), p. 35. ISSN: 07300301. doi: 10.1145/1276377.1276421.
- [215] Markus Voelter. "Introduction to DSLs". In: *DSL Engineering: Designing, Implementing and Using Domain-Specific Languages*. 2013. ISBN: 978-1481218580.
- [216] S. Wang et al. "Towards provenance-aware geographic information systems". In: *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems* (2008), pp. 483–486. doi: 10.1145/1463434.1463515.
- [217] Peter Werner, Rod Thorpe, and David Bunker. "Teaching Games for Understanding: Evolution of a Model". In: *Journal of Physical Education, Recreation & Dance* 67.1 (Jan. 1996), pp. 28–33. ISSN: 0730-3084. doi: 10.1080/07303084.1996.10607176.

- [218] Morgan Williams and Stuart Morgan. “Horizontal positioning error derived from stationary GPS units: A function of time and proximity to building infrastructure”. In: *International Journal of Performance Analysis in Sport* 9.2 (Aug. 2009), pp. 275–280. issn: 2474-8668. doi: 10.1080/24748668.2009.11868483.
- [219] James R. Williamson et al. “Individualized detection of ambulatory distress in the field using wearable sensors”. In: *2013 IEEE International Conference on Body Sensor Networks, BSN 2013* (2013), pp. 1–6. doi: 10.1109/BSN.2013.6575527.
- [220] David H. Wolpert. “The lack of a priori distinctions between learning algorithms”. In: *Neural computation* 8.7 (1996), pp. 1341–1390.
- [221] David H. Wolpert and William G. Macready. “No free lunch theorems for optimization”. In: *IEEE Transactions on Evolutionary Computation* 1.1 (1997), pp. 67–82. issn: 1089778X. doi: 10.1109/4235.585893.
- [222] Carl T. Woods, Sam J. Robertson, and Paul B. Gartin. “Does relative age distribution influence the physical and anthropometric profiles of drafted under 18 Australian footballers? An investigation between the 2010 to 2013 seasons”. In: *Talent Development and Excellence* 7.1 (2015), pp. 83–90. issn: 18692885.
- [223] Craig A. Wrisberg. “Closed and Open Environments”. In: *Sport skill instruction for coaches*. Human Kinetics, 2007, p. 37. isbn: 0736039872.
- [224] Dongyao Wu et al. “Building pipelines for heterogeneous execution environments for big data processing”. In: *IEEE Software* 33.2 (2016), pp. 60–67. issn: 07407459. doi: 10.1109/MS.2016.35.
- [225] Zengyuan Yue et al. “Mathematical analysis of a soccer game. Part I: Individual and collective behaviors”. In: *Studies in Applied Mathematics* 121.3 (2008), pp. 223–243. issn: 00222526. doi: 10.1111/j.1467-9590.2008.00413.x.
- [226] Manzil Zaheer et al. “Deep Sets”. In: *Advances in Neural Information Processing Systems (NIPS 2017)*. 2017. arXiv: 1703.06114.

- [227] Annemarie Zand Scholten and Denny Borsboom. “A reanalysis of Lord’s statistical treatment of football numbers”. en. In: *Journal of Mathematical Psychology* 53.2 (Apr. 2009), pp. 69–75. issn: 00222496. doi: 10.1016/j.jmp.2009.01.002.
- [228] H. Zang and J. Bolot. “Anonymization of Location Data Does Not Work : A Large-Scale Measurement Study”. In: *Proceedings of the 17th annual international conference on Mobile computing and networking*, ACM. (2011), pp. 145–156. doi: 10.1145/2030613.2030630.
- [229] Achim Zeileis and Gabor Grothendieck. “zoo : S3 Infrastructure for Regular and Irregular Time Series”. In: *Journal of Statistical Software* 14.6 (2005). issn: 1548-7660. doi: 10.18637/jss.v014.i06.
- [230] L. Zhang, Sushil Jajodia, and A. Brodsky. “Information disclosure under realistic assumptions: Privacy versus optimality”. In: *Computer and Communications Security* (2007), pp. 573–583. issn: 15437221. doi: 10.1145/1315245.1315316.
- [231] Jun Zhao et al. “Why workflows break - Understanding and combating decay in Taverna workflows”. In: *2012 IEEE 8th International Conference on E-Science*. IEEE, Oct. 2012, pp. 1–9. isbn: 978-1-4673-4466-1. doi: 10.1109/eScience.2012.6404482.

Glossary

AFL Australian Football League, the game of Australian Rules Football played at the national elite level in accordance with the official rules set by the AFL Commission. Informally used as a synonym for Australian Rules Football.

Australian Rules Football The game of Australian Rules Football played between two competing teams with an oval ball on an oval field (see Sec. 2.2).

behind A kick that goes between two outer goal posts rather than the two inner goal posts. Unlike a goal which is worth 6 points, a behind is only awarded 1 point.

centre bounce An event marking the start of play in AFL. The umpire bounces the ball in the centre of the field and ruck players attempt to hit it towards their own team.

GIS Geographic Information System.

GPS The Global Positioning System of satellites operated by US Department of Defense. Informally used to refer to other forms of Global Navigation Satellite Systems (GNSS) such as GLONASS operated by Russia. Position tracking devices utilising GPS or other GNSS systems often integrate additional microsensors such as accelerometers, gyroscopes, and magnetometers to enhance accuracy during short rapid movements occurring between readings. Local Positioning Systems such as Catapult ClearSky can fulfil a similar role by using radio beacons placed at a stadium rather than satellites.

GLOSSARY

kick-in An event in AFL where after a team scores a behind, the opposition is provided with possession of the ball to kick in from the end of the field towards the opposite end of the field.

quarter One of the four periods that comprise an AFL match. Each quarter lasts approximately 20 minutes, but is extended during times that the ball is not in play.

spatio-temporal relating to both spatial (location) and temporal (time) dimensions.

Appendix A

Modelling

A.1 Application of Abstract Data Model

A.1.1 Concrete Syntax

This section utilises the abstract data model to define concrete data models that describe the schema and meta-data of real-world data. This can be performed directly within a general purpose programming language, or through definition of a Domain Specific Language (DSL) based on the abstract model to ease definition of the concrete models. In respect to its ability to serve as language to describe concrete models, the proposed abstract data model serves a similar purpose to a linguistic metamodel. However, as Kühne notes in *Matters of (meta-) modeling* [121], generalisation is a transitive relationship that directly applies to the end system, and thus it is more appropriate to describe the model as “abstract” rather than as a true “metamodel”.

For the purposes of this thesis, a very simple notation will be defined for convenience:

- The syntax *name* : *type* will be used to annotate specific measurements with their type. For example, ball position : Spatial.

- Dense parameters will be represented by postfixing them with a star *. For example, time : Temporal*.
- The syntax *attribute* = *value* will be used to specify the meta-data of measured parameters. In the case of unknown parameters, the token UNKNOWN will be used.

A.1.2 Examples

These examples utilise the abstract data model and concrete syntax to compare the data collection methods currently available in a range of sports.

Traditional Game Summary

A human tallies events. Assume that official data is 100% accurate.

```
Traditional Game Summary : Sensor Platform
    Human : Human Entry
        event : Event*
            Granularity = {goal, behind}
        tally : Count
        Accuracy = ±0
```

Database System

A human enters detailed game data based on video feeds. The error estimates are taken from O'Shaughnessy [162] and Jackson [111] who have both held positions at Champion Data, the AFL data provider.

```
Database System : Sensor Platform
    Human : Human Entry
        event : Event
```

```
Accuracy      = 99%
Granularity = {kick, mark, handball, hit, tackle,
               goal, behind, pick-up, centre-bounce,
               throw-in, out-of-bounds, free-kick}
difficulty : Qualitative
Inter-rate Reliability = UNKNOWN
Granularity           = {easy, hard}
ball position : Spatial
Error Radius = 5-10 m
time : Temporal
error rate = 5 s
player : ID
accuracy = UNKNOWN
event sequence order : ID*
accuracy = UNKNOWN
```

Computer Vision

Computer vision tracks two types of items: the ball, and the players. A common issue with computer vision systems is losing track of the mapping between trajectories and player identifiers when two players pass in close proximity of each other.¹

```
Computer Vision : Sensor Platform
Vision : Sensor
time : Temporal*
Accuracy      = UNKNOWN
Sample Rate = 25 Hz
item : ID*
accuracy = UNKNOWN
position : Spatial
Error Radius : UNKNOWN
```

¹The error estimates are taken from <http://www.stats.com/sportvu/sportvu-basketball-media/>

Wearable Tracking Device

Unlike computer-vision, wearable devices ensure that the player identifier is always 100% accurate as it is physically attached to the player (unless of course, if a player wears the wrong device).² Note that velocity vectors and orientation vectors are treated as spatial data. Heart rate could be modelled in a number of ways; however, it was decided to model it as a cumulative count, as this avoids ambiguities that typically arise when describing variable rates.

Wearable Tracking Device : Sensor Platform

```
GPS : Sensor
    time : Temporal*
        Accuracy      = UNKNOWN
        Sample Rate = 10 Hz
player : ID*
    Accuracy = 100%
position : Spatial
    Error Radius = 3 m
velocity : Spatial
    Error Radius = UNKNOWN
Accelerometer : Sensor
    time : Temporal*
        Accuracy      = UNKNOWN
        Sample Rate = 1000 Hz
player : ID*
    Accuracy = 100%
acceleration vector : Spatial
    Error Radius = UNKNOWN
Gyroscope : Sensor
```

²The sample rates are taken from Catapult Sports circa 2015 (to reflect the technology available during the time-period of the primary dataset used in this thesis) and found in product information released to <https://web.archive.org/web/20170910220858/http://www.catapultsports.com:80/system/system/> and <https://youtu.be/F9ZsYEyf3HE?t=110>. GPS accuracy is taken from <http://www.ga.gov.au/scientific-topics/positioning-navigation/geodesy/geodetic-techniques/global-positioning-systems-gps/gps-consumer>

```
time : Temporal*
    Accuracy      = UNKNOWN
    Sample Rate = 1000 Hz
player : ID*
    Accuracy = 100%
orientation vector : Spatial
    Error Radius = UNKNOWN
Heart Rate Monitor : Sensor
    time : Temporal*
        Accuracy      = UNKNOWN
        Sample Rate = UNKNOWN
player : ID*
    Accuracy = 100%
Cumulative Heart Beats : Count
    Accuracy = UNKNOWN
```

Appendix B

Computational Pipelines

B.1 Symbols for Custom Sport Provenance Notation

Table B.1: Specialised symbols for Entities

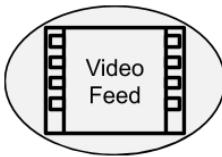
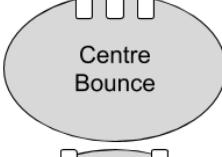
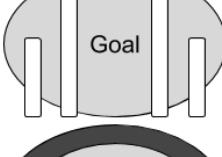
Semantic Construct (W3C PROV)	Specialised Symbol (for Sport)	ID
Entity	Video feed 	1
Game state	Possession Centre field 	2
Game event	Centre Bounce  Goal  Injury  External Influence 	3
Metric	Goal% Ratio 	4

Table B.2: Specialised symbols for Activities

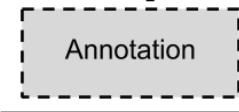
Semantic Construct (W3C PROV)	Specialised Symbol (for Sport)	ID
Activity	Physical action 	5
	Kick	
Manual process	Manual process 	6
Computation	Computation 	7
De-identification	De-identification 	8

Table B.3: Specialised symbols for Agents

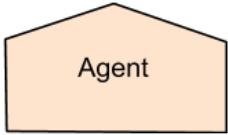
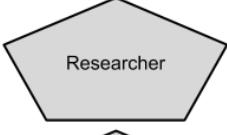
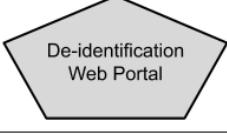
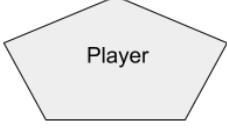
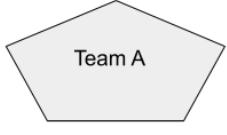
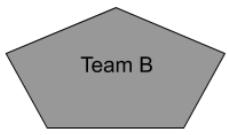
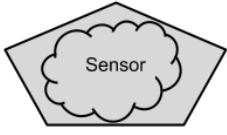
Semantic Construct (W3C PROV)	Specialised Symbol (for Sport)	ID
	Analyst / System	9
		
		
	Player / Role	10
		
	Team	11
		
		
	(Distinguished by colour)	
	Sensor	12
		

Table B.4: Specialised symbols for Associations

Semantic Construct (W3C PROV)	Specialised Symbol (for Sport)	ID
	Data dependency	13
	(Distinguished by context)	
	Physical causality	14
	(Distinguished by context)	

Table B.5: Specialised symbols for Bundles

Semantic Construct (W3C PROV)	Specialised Symbol (for Sport)	ID
Bundle (no symbol)		15

B.2 Effectiveness of visual notation against principles of Physics of Notations

Table B.6: Effectiveness of visual notation against principles of Physics of Notations [144]

Criterion	W3C PROV	VisTrails
Semiotic Clarity (fraction of semantic constructs in Table 4.2 mapped to unique symbols)	4/15 Contains high level semantics for entity, activity, agent and association.	3/15 Metric (port), computation, data dependency (association). No concept of agents. No ability to directly model real world. No concept of physical causality.

Table B.6: Effectiveness of visual notation against principles of Physics of Notations [144]

Perceptual	4/4	3/3
Discriminability (fraction of symbols with unique visual variables)	Could be improved: different colours / shapes for specialisations. (Points still awarded because top level constructs have distinct symbols)	Ports and activities share same shape as each other, but differ by size. Could be improved: ports with different types should have different colours / shapes. Activities with different types should have different colours and use a larger variety of shapes. (Points still awarded for these because only one type of sport semantic construct was supported)

Table B.6: Effectiveness of visual notation against principles of Physics of Notations [144]

Semantic Transparency (fraction of symbols with obvious meanings)	0/4 Use of circles for entities and rectangles for processes conflicts with data flow diagrams (which use circles for processes). Use of house shaped pentagons for agents is only memorable when agent represents an organisation. Arrows are in direction of data dependency, but intuitive interpretation is in direction of data flow.	3/3 Analogy: electric circuit (rectangular components, small contacts, connection wires) Could be improved: while obvious square is a port, not obvious which port is which (user has to memorise order). While obvious that box is a process, specific type of process is not obvious (e.g. uses pentagon for control flow rather than conventional diamond for “if” condition)
Complexity Management (can it visualize complex workflows?)	Yes Ontologies support the “Open-world assumption”, thus allowing specifying as much or as little detail as appropriate.	Yes Supports grouping nodes

Table B.6: Effectiveness of visual notation against principles of Physics of Notations [144]

Cognitive Integration (can the user navigate without getting lost?)	Yes Includes concept of “bundles” to annotate information required to navigate documents at meta-level. E.g. to describe provenance of provenance information.	Yes Top level workflow acts as overview, then user can drill down into parameter values, history variations, etc.
Visual Expressiveness (fraction of visual variables used)	2/8 Shape and colour.	1/8 Shape Colour is used for execution state, but this is not one of semantic constructs, and brightness is used to determine if a port is connected, but neither of these map to semantic constructs of relevance.

Table B.6: Effectiveness of visual notation against principles of Physics of Notations [144]

Dual Coding (fraction of symbol parameters with multiple unique visual variables)	1/3	0/3 In theory shape and colour can be assigned if designing custom module, but colour is not used in any of the default modules.
Graphic Economy (total symbols, less is better as it reduces cognitive load)	4	3 (If all features not part of assessment were removed)
Cognitive Fit (is the notation understandable to performance analysts?)	Partial When arrows are labelled, visual notation is unambiguous.	Partial Intuitive flow metaphor; however, advanced operations require writing custom Python scripts.

B.3 Usability evaluation of VisTrails using Nielsen's top ten heuristics

Table B.7: Usability evaluation of VisTrails using Nielsen's top ten heuristics [153]

Criterion	Support	Issues
Visibility of system status	Shows progress indicator when evaluating workflow. Displays which modules executed / have errors.	
Match between system and the real world	Boxes for processes connected by lines resembles real-world electronic wiring of modules.	Some terms may present confusion for non-technical users: “PythonCalc” (evaluate an expression), “StandardOutput” (display result in the terminal), and “Map” (a higher order function, not a geographical map).
User control and freedom	Full tracking of history as tree	
Consistency and standards		Some terms such as “port” (rather than input / output) may increase time to learn.

Table B.7: Usability evaluation of VisTrails using Nielsen's top ten heuristics [153]

Error prevention	Ports have types to ensure that user can only connect two ports if their types match.	
Recognition rather than recall	The system provides some support to aid the user's memory (e.g. dark ports to remind the user a default has been set)	The user needs to memorise the port order of modules to use the interface efficiently.
Flexibility and efficiency of use	Provides shortcut key combinations for advanced users	
Aesthetic and minimalist design	Main focus of the application is on the workflow	
Help users recognise, diagnose, and recover from errors	System highlights module(s) with error	Use of colour as sole indicator of error could be problematic for users with colour blindness.

Table B.7: Usability evaluation of VisTrails using Nielsen's top ten heuristics [153]

Help and documentation	User manual includes step-by-step guidelines on how to use. In-built option to display documentation for the selected module	In-built documentation for module often missing
------------------------	---	---

Appendix C

De-identification

C.1 Detailed analysis of studies claiming use of “non-identifiable” data

C.1.1 Analysis of Dataset 1

In a study of selection attributes in elite junior Australian Rules football [174] the authors claim “access and consent to non-identifiable testing data was provided by each of the relevant state-based organisations and the study was approved by the relevant human research ethics advisory group.” No details were provided about whether individual consent of participants had also been sought. The data includes player birthdates, performance, anthropometric data, as well as the outcome variable of interest, whether the player was drafted. The study used logistic regression and the JRip algorithm to predict whether the player was drafted as a function of the other variables in the dataset. The use of logistic regression and the JRip algorithm in this manner imply that the authors must have access to individual data rows for each player rather than group averages. The use of the term “non-identifiable” is questionable; the study specifically dealt with elite junior players which may progress into elite Australian Rules football, upon which names and birthdates are likely to become public information. In future, these details could

then be linked back to the full attributes of the junior player using birthdate as a key. Thus while the participants may have been “non-identifiable” at the time of the study, it is likely that some of the participants may become re-identifiable in future once more details become public. The classification of the data as “non-identifiable” means that it could become part of a research databank kept long term, despite the data containing details that are likely to reveal participants at a future date.

The same authors published a second paper [222], analysing the same dataset as [174] from the perspective of age distribution. As some players did not participate in all performance tests (e.g. if injured, players would not be able to participate in performance tests that stress the injured part of the body), the number of players used to report averages differs between the two papers; there were up to $n=292$ drafted players in [222] who were measured, but only $n=212$ complete player profiles selected for use in [174]. By comparing the differences between the two papers, it is possible to infer details about the group of players who were included in the first publication, but not the second. For example, [174] reports a mean drafted player height of 186.4 cm ($n=281$), and a mean drafted player body mass of 79.5 kg ($n=282$). Whereas [222] reports a mean drafted player height of 185.7 cm ($n=212$), and a mean drafted player body mass of 79.5 kg ($n=212$). From the difference, it is possible to calculate that the excluded group (69 and 70 players) were higher than average (188.5 cm) and had greater body mass (81.0 kg). These differences alone give some indication of the characteristics of the excluded group. Furthermore, an attacker may have information from auxiliary sources as to which players belong to the excluded group¹. In a similar manner, an attacker could calculate the average performance results (in the tests that injured players were able to perform) for the excluded group. While the performance attributes of individuals are not sensitive—identifiable lists of the top performers in each test are publicly published by the AFL each year—the lack of public identifiable data including every individual suggests that only the results of the

¹For example, a 2017 AFL Media article names players who were not able to complete some of the tests that year due to injury <http://www.lions.com.au/news/2017-10-06/afl-draft-combine-wrap>

top performers are intended to be public information. There were too many players in the excluded group that knowledge of the group statistics would allow an attacker to infer individual player results. However, it appears that the attack was only avoided by chance (due to multiple players with incomplete profiles that needed to be excluded) rather than by design, and could thus pose a threat to participant privacy in other studies where the excluded group with incomplete profiles is small. This highlights the need for a mechanism to preserve privacy at the data level rather than post-publication to ensure that participants are not revealed by the differences between results when multiple publications discuss the same dataset from different perspectives.

C.1.2 Analysis of Dataset 2

Greenham et al. [89] perform a pilot study to measure game style in Australian Rules Football. They state that their research is exempt from ethics review, as “non-identifiable player data, from identifiable team-based data-sets, were used in this study”.

Their measure of game style was derived from 12 variables. In this re-analysis, each variable was re-considered from the perspective of de-identification.

Nine of their variables were either publicly reported, or could potentially have been derived from public data (e.g. *Shot at goal accuracy*), thus for the purposes of de-identification there is no need to consider these further. *Location of Goal attempts* is potentially a de-identification issue as it is possible to re-identify players if their location is known; however, the study only used the proportion of goal attempts taken from close to the goal, and it is possible that Champion Data precomputed this prior to providing the data to the researchers. *Ball Speed* was derived from video footage, while video footage is obviously identifiable, if using public video footage this is not a privacy issue, and the research may still be exempt from ethics approval. *Offensive and defensive player numbers in the 50 m zone* and the closely related *differential in team*

player numbers were derived “using video footage recorded behind the goals”. This is an issue, as contrary to the authors claim, behind the goals video footage is not public², nor de-identified³. However, one could potentially argue that a spectator at the game could observe the same information if they had reserved the right seat at the game.

The table in the paper only provides summary statistics taken over the entire group of games. However, the visual “game style plot” (parallel coordinates visualisation), shows z-scores for individual teams, identified by team name. Nevertheless, it is unlikely that one could infer details of individual players from this plot. While the information revealed was limited in the paper, it demonstrates that attention needs to be given to data revealed in figures and visualisations, not just the main text and tables.

C.1.3 Analysis of Dataset 3

Jacob et al. [112] perform a pilot study investigating the link between genetic polymorphisms and performance in Australian Rules Football. The study collected individual consent of players (and parental consent for players under 18). The study stated that to “ensure anonymity, the players were assigned a randomised, non-identifiable code.”

From a de-identification perspective, there are two aspects of this study that make de-identification difficult: firstly, it only involved 30 participants, while small samples present well understood issues for validity

²Clubs are provided with “exclusive behind the goals vision” recordings <https://www.foxsports.com.au/afl/geelong-coach-chris-scott-explains-why-afl-coaches-bother-going-to-games-inperson-in-2016-with-video-technology/news-story/c9b99fbf9472483491294056c03bf25b>, which are considered a “game-changer” for football analysis <http://www.afl.com.au/news/2018-02-18/secret-spies-the-life-of-an-opposition-analyst>. A news report in 2013 revealed that clubs payed \$28,000/year each for the footage, with prices expected to rise to \$60,000/year per club in 2014. <https://www.theage.com.au/sport/afl/afl-doubles-tv-costs-20131025-2w7d1.html>

³Even in the hypothetical case that the authors were to ask the video provider to blur out faces and player numbers in the behind-the-goals video, the position and movements of players evident within the video would still allow re-identifying particular players in the footage.

as they risk describing the group rather than population, this can also be understood as a privacy issue as description of the characteristics of the group can be used to make inferences about the members of the group; and secondly, genetic polymorphisms can vary in distribution between racial and ethnic groups, thus revealing genetic markers of a sub-group of the study may unintentionally reveal the likely racial or ethnic profile of that sub-group.

The study does not reveal the players studied, stating only that the study “recruited 30 sub-elite Australian [Rules] Football players”, presumably this is to prevent individual players being identified. However, all authors of the study were from the University of Notre Dame Australia, Fremantle, Australia; it is thus likely that the players were recruited from a club within close proximity of Fremantle, Australia. Furthermore, a publicly available author pre-publication copy of the study mentions East Fremantle Football Club in the acknowledgement section. East Fremantle Football Club, and other clubs in the area, publicly publish the names of players in their team. Thus attempting to remove the name of the club from the published paper provided only superficial privacy, and it is reasonable to assume that an attacker could infer the list of players that potentially participated in the study.

The study considered polymorphisms of 9 genes, and published regression coefficients for the association between each genotype on performance. Amongst these, the study found “the ACE [angiotensin-converting enzyme] DD genotype, associated with higher plasma ACE levels, had the greatest positive impact on [Australian Rules Football] players in traditional power and aerobic athletic assessments, as well as in sport-specific skill assessments.” However, it is also necessary to consider what information this reveals to an attacker regarding the identity of participants; the ACE DD genotype is also known to occur in lower proportions within certain ethnic groups, notably a study of blood and kidney donors [126] found that Australian Aboriginals had a D allele frequency of 14%, compared to Australian Caucasians who had a D allele frequency of 55%, furthermore, one tribe of Australian Aboriginals was found to have a D allele frequency of just 3%. In Jacob et al., the performance results published for the DD genotype was de-

rived from a group of just 10 players. While the study does not specify who these players were, an attacker can refine the possible candidates by inferring that they were unlikely to be Australian Aboriginals given that they had the DD genotype.

In three cases, a genotype only corresponded to a single player. Thus the regression coefficients for these genotypes correspond to the player profile for a specific player. While the study does not make the identity of the player known, hypothetically, if a the player were to exhibit exceptional results for one of the tests, an attacker familiar with the team may be able to infer the likely identity of the player for that test, and use it to look up their results on the other tests (as they were the only player with that specific genotype). As before, the genotype itself also reveals information about that player race and ethnicity which could help the attacker reduce the possible candidates.

Appendix D

A Platform for Spatio-Temporal Sport Analysis

D.1 Data Extraction Requests

Analysis 1 - Metadata and Data Sample

This analysis will reveal a sample of each data file to help us understand the data structure, without revealing the entire dataset.

- A sample of the first 10 rows and last 10 rows will be extracted for each file.
- A summary of values in each column will be extracted (e.g. frequency of each category, mean, standard deviation).

Analysis 2b - Formation Profiles (v2)

This analysis will extract team formations in a way that allows studying the team, but not individuals.

- Data will be resampled to 1Hz (1 sample per second) to reduce risk of revealing detailed player movements detected by accelerometers.
 - Team formations will be represented as a “point cloud” that includes non-identifiable dots for the location of each player in the formation, but does not allow tracing the position of individual players over the full course of the match (i.e. even if one knows the player position at a certain time, they will lose track of which player is which whenever two player paths cross over each other).
- v2:** Includes fix to handle time jumps backward (up to 1 second) or forward (up to 11 minutes) occurring in sensor data.

Analysis 3 - Signal Quality

This analysis will extract the time of GPS fixes and the number of satellites, but not the GPS data (latitude, longitude, speed, acceleration, etc.) itself. Columns to be extracted:

- Player (e.g. "A1")
- Round
- Time of GPS fix (e.g. "15:05.4")
- Sats column (e.g. "T 4")

D.2 Review of Team Shape Metrics found in the Literature

Stretch Index

Yue et al. 2008 [225] demonstrated mathematical techniques for team-level analysis of Association Football player data. As a measure of team

spread, they defined the “instantaneous radius” of the team to be the average distance of players from the centroid of the team. Bourbousson et al. 2010 [23] analysed basketball team data, and as a measure of team spread, they defined the “stretch index” in a manner that appears equivalent to Yue et al.’s earlier definition of “instantaneous radius” [225]. It is also sometimes referred to as the “dispersion” [79].

Variations have been proposed in the literature. Maçãs and Sampaio 2012 [129] proposed measuring the minimum and maximum distance of players from the centroid of the team. Clemente et al. 2013 [40] suggested weighting the calculation of the team centre and spread by the distance players are from the ball (giving more weight to players that are closer to the ball).

Surface Area

Frencken et al. 2011 [78] (following a preliminary study by Frencken et al. 2008 [77]) analysed 5-a-side Association Football player data and propose using a convex hull to measure the surface area of the team shape (excluding the goal-keeper).

Clemente et al. 2013 [40] (method later republished [42]) suggested a variation of surface area that they denote “effective area”, which subtracts areas controlled by the opposition from the calculation. Clemente et al. demonstrated this for an under-13 7-a-side Association Football district final, and later performed a follow up study of three official 11-a-side Association Football matches [41].

Length and Width

Folgado et al. 2014 [74] defined team length, width and length-to-width ratio, which they used to compare under-9, under-11 and under-13 Association Football. These techniques were reused by Frias and Duarte 2014 [79] who examined two-team GPS data in small-sided youth Association Football games. Frias and Duarte created a custom Matlab

application to calculate team surface area, stretch index, and length-to-width ratio. These were analysed to examine the effects of two different coaching interventions.

Spatial Variability

Couceiro et al. 2014 [45] proposed using Shannon Entropy [185] to quantify the level of variability in a heat-map of player position. Specifically, they applied Shannon Entropy to the context of sport by using it to measure the uncertainty as to which point on the field would be occupied by a player (without consideration of player roles).

Appendix E

Human Research Ethics

Human Research Ethics

Deakin Research Integrity
Burwood Campus
Postal: 221 Burwood Highway
Burwood Victoria 3125 Australia
Telephone 03 9251 7123
research-ethics@deakin.edu.au



Memorandum

To: A/Prof Paul Gastin
School of Exercise and Nutrition Sciences

B **cc:** Mr Andrew Simmons

From: Deakin University Human Research Ethics Committee (DUHREC)

Date: 17 April, 2018

Subject: 2018-121
GPS analysis of team shape and game style in elite Australian Rules Football
Please quote this project number in all future communications

At the DUHREC meeting to be held on 14/05/2018 it will be noted that A/Prof Paul Gastin has declared their project to be exempt from ethical review because it involves only the use of:

- pre-existing, non-identifiable data; and/or
- publicly available data

Based on their signed declaration, Mr Andrew Simmons, under the supervision of A/Prof Paul Gastin, School of Exercise and Nutrition Sciences, is authorised to undertake this project from 17/04/2018 for the life of the project.

DUHREC may need to audit this project as part of the requirements for monitoring set out in the National Statement on Ethical Conduct in Human Research (2007).

Human Research Ethics Unit
research-ethics@deakin.edu.au
Telephone: 03 9251 7123

Appendix F

Authorship Statements

AUTHORSHIP STATEMENT

1. Details of publication and executive author

Title of Publication		Publication details
Data Provenance for Sport		arXiv:1812.05804 (Draft)
Name of executive author	School/Institute/Division if based at Deakin; Organisation and address if non-Deakin	Email or phone
Andrew J. Simmons	Applied Artificial Intelligence Institute	a.simmons@deakin.edu.au

2. Inclusion of publication in a thesis

Is it intended to include this publication in a higher degree by research (HDR) thesis?	<u>Yes</u> / No	If Yes, please complete Section 3 If No, go straight to Section 4.
---	-----------------	---

3. HDR thesis author's declaration

Name of HDR thesis author if different from above. (If the same, write "as above")	School/Institute/Division if based at Deakin	Thesis title
As above	Applied Artificial Intelligence Institute	Computational Pipelines for Spatio-Temporal Analysis of Team Invasion Sport
If there are multiple authors, give a full description of HDR thesis author's contribution to the publication (for example, how much did you contribute to the conception of the project, the design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)		
This conception of this paper began with a whiteboard sketch for a data provenance architecture I created with Rajesh Vasa. After conducting a literature review, I realised similar prior work in literature, so refined the approach to focus on sport. Discussions with Scott Barnett helped bring logical structure to the paper. Simon Vajda assisted with further literature search, and designed the symbols used for the custom data provenance for sport notation.		
<i>I declare that the above is an accurate description of my contribution to this paper, and the contributions of other authors are as described below.</i>	Signature and date	

4. Description of all author contributions

Name and affiliation of author	Contribution(s) (for example, conception of the project, design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)
Andrew J. Simmons	Conception, Drafting
Scott Barnett	Structuring of paper, Revising Critically
Simon Vajda	Searching literature, Design of Symbols
Rajesh Vasa	High level conception, Revising Critically

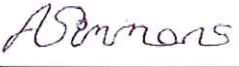
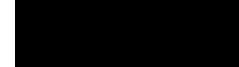
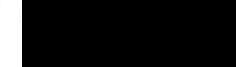
5. Author Declarations

I agree to be named as one of the authors of this work, and confirm:

- i. that I have met the authorship criteria set out in the Deakin University Research Conduct Policy,
- ii. that there are no other authors according to these criteria,
- iii. that the description in Section 4 of my contribution(s) to this publication is accurate,
- iv. that the data on which these findings are based are stored as set out in Section 7 below.

If this work is to form part of an HDR thesis as described in Sections 2 and 3, I further

- v. consent to the incorporation of the publication into the candidate's HDR thesis submitted to Deakin University and, if the higher degree is awarded, the subsequent publication of the thesis by the university (subject to relevant Copyright provisions).

Name of author	Signature*	Date
Andrew J. Simmons		12 March 2019
Scott Barnett		14 May 2019
Simon Vajda		14 May 2019
Rajesh Vasa		14 May 2019

6. Other contributor declarations

I agree to be named as a non-author contributor to this work.

Name and affiliation of contributor	Contribution	Signature* and date

* If an author or contributor is unavailable or otherwise unable to sign the statement of authorship, the Head of Academic Unit may sign on their behalf, noting the reason for their unavailability, provided there is no evidence to suggest that the person would object to being named as author

7. Data storage

The original data for this project are stored in the following locations. (The locations must be within an appropriate institutional setting. If the executive author is a Deakin staff member and data are stored outside Deakin University, permission for this must be given by the Head of Academic Unit within which the executive author is based.)

Data format	Storage Location	Date lodged	Name of custodian if other than the executive author

This form must be retained by the executive author, within the school or institute in which they are based.

If the publication is to be included as part of an HDR thesis, a copy of this form must be included in the thesis with the publication.

AUTHORSHIP STATEMENT

1. Details of publication and executive author

Title of Publication		Publication details
An interaction model for de-identification of human data held by external custodians		OzCHI 2018
Name of executive author	School/Institute/Division if based at Deakin; Organisation and address if non-Deakin	Email or phone
Andrew J. Simmons	Applied Artificial Intelligence Institute	a.simmons@deakin.edu.au

2. Inclusion of publication in a thesis

Is it intended to include this publication in a higher degree by research (HDR) thesis?	<u>Yes</u> / No	If Yes, please complete Section 3 If No, go straight to Section 4.
---	-----------------	---

3. HDR thesis author's declaration

Name of HDR thesis author if different from above. (If the same, write "as above")	School/Institute/Division if based at Deakin	Thesis title
As above	Applied Artificial Intelligence Institute	Computational Pipelines for Spatio-Temporal Analysis of Team Invasion Sport
If there are multiple authors, give a full description of HDR thesis author's contribution to the publication (for example, how much did you contribute to the conception of the project, the design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)		
<i>The problem definition was motivated by early discussions I had with Paul B. Gastin who led the ethics application/exemption and encouraged me to find a method to de-identify the GPS tracking data. Based on further discussions with Rajesh Vasa, I proposed and implemented a solution, and drafted the paper. The emotional goal modelling was done on a whiteboard with assistance from Maheswaree Kissoon Curumsing who guided me on how to use this methodology. Both Maheswaree Kissoon Curumsing and Rajesh Vasa contributed to critically revising the work.</i>	<i>I declare that the above is an accurate description of my contribution to this paper, and the contributions of other authors are as described below.</i>	Signature and date <i>ASimmons</i>

4. Description of all author contributions

Name and affiliation of author	Contribution(s) (for example, conception of the project, design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)
Andrew J. Simmons	Design, implementation, analysis, drafting
Maheswaree Kissoon Curumsing	Emotional goal modelling methodology, revising critically
Rajesh Vasa	Revising critically

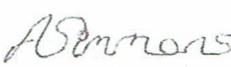
5. Author Declarations

I agree to be named as one of the authors of this work, and confirm:

- i. that I have met the authorship criteria set out in the Deakin University Research Conduct Policy,
- ii. that there are no other authors according to these criteria,
- iii. that the description in Section 4 of my contribution(s) to this publication is accurate,
- iv. that the data on which these findings are based are stored as set out in Section 7 below.

If this work is to form part of an HDR thesis as described in Sections 2 and 3, I further

- v. consent to the incorporation of the publication into the candidate's HDR thesis submitted to Deakin University and, if the higher degree is awarded, the subsequent publication of the thesis by the university (subject to relevant Copyright provisions).

Name of author	Signature*	Date
Andrew J. Simmons		12 March 2019
Maheswaree Kissoon Curumsing		14 May 2019
Rajesh Vasa		14 May 2019

6. Other contributor declarations

I agree to be named as a non-author contributor to this work.

Name and affiliation of contributor	Contribution	Signature* and date
Paul B. Gastin	Problem conception, led ethics application/exemption Note: Paul B. Gastin meets the authorship criteria. However, he was not listed on the paper when published.	 14 May 2019

* If an author or contributor is unavailable or otherwise unable to sign the statement of authorship, the Head of Academic Unit may sign on their behalf, noting the reason for their unavailability, provided there is no evidence to suggest that the person would object to being named as author

7. Data storage

The original data for this project are stored in the following locations. (The locations must be within an appropriate institutional setting. If the executive author is a Deakin staff member and data are stored outside Deakin University, permission for this must be given by the Head of Academic Unit within which the executive author is based.)

Data format	Storage Location	Date lodged	Name of custodian if other than the executive author

This form must be retained by the executive author, within the school or institute in which they are based.

If the publication is to be included as part of an HDR thesis, a copy of this form must be included in the thesis with the publication.

AUTHORSHIP STATEMENT

1. Details of publication and executive author

Title of Publication		Publication details
Spatio-Temporal Reference Frames as Geographic Objects		SIGSPATIAL 2017
Name of executive author	School/Institute/Division if based at Deakin; Organisation and address if non-Deakin	Email or phone
Andrew J. Simmons	Applied Artificial Intelligence Institute	a.simmons@deakin.edu.au

2. Inclusion of publication in a thesis

Is it intended to include this publication in a higher degree by research (HDR) thesis?	<u>Yes</u> / No	If Yes, please complete Section 3 If No, go straight to Section 4.
---	-----------------	---

3. HDR thesis author's declaration

Name of HDR thesis author if different from above. (If the same, write "as above")	School/Institute/Division if based at Deakin	Thesis title
As above	Applied Artificial Intelligence Institute	Computational Pipelines for Spatio-Temporal Analysis of Team Invasion Sport
If there are multiple authors, give a full description of HDR thesis author's contribution to the publication (for example, how much did you contribute to the conception of the project, the design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)		
I conceived of and implemented the approach, then worked with Rajesh Vasa to determine the research contributions and find ways to present the work. Rajesh Vasa proposed adapting Bloom's revised Taxonomy of Learning (usually used for education) as an evaluation tool in this paper to reason about the learnability of the system. I drafted the paper with critical revisions from Rajesh Vasa.	Signature and date	
<i>I declare that the above is an accurate description of my contribution to this paper, and the contributions of other authors are as described below.</i>		

4. Description of all author contributions

Name and affiliation of author	Contribution(s) (for example, conception of the project, design of methodology or experimental protocol, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)
Andrew J. Simmons	Conception, implementation, analysis, drafting
Rajesh Vasa	Design of evaluation methodology, Revising critically

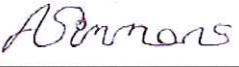
5. Author Declarations

I agree to be named as one of the authors of this work, and confirm:

- i. *that I have met the authorship criteria set out in the Deakin University Research Conduct Policy,*
- ii. *that there are no other authors according to these criteria,*
- iii. *that the description in Section 4 of my contribution(s) to this publication is accurate,*
- iv. *that the data on which these findings are based are stored as set out in Section 7 below.*

If this work is to form part of an HDR thesis as described in Sections 2 and 3, I further

- v. *consent to the incorporation of the publication into the candidate's HDR thesis submitted to Deakin University and, if the higher degree is awarded, the subsequent publication of the thesis by the university (subject to relevant Copyright provisions).*

Name of author	Signature*	Date
Andrew J. Simmons		12 March 2019
Rajesh Vasa		14 Mar 2019

6. Other contributor declarations

I agree to be named as a non-author contributor to this work.

Name and affiliation of contributor	Contribution	Signature* and date

* If an author or contributor is unavailable or otherwise unable to sign the statement of authorship, the Head of Academic Unit may sign on their behalf, noting the reason for their unavailability, provided there is no evidence to suggest that the person would object to being named as author

7. Data storage

The original data for this project are stored in the following locations. (The locations must be within an appropriate institutional setting. If the executive author is a Deakin staff member and data are stored outside Deakin University, permission for this must be given by the Head of Academic Unit within which the executive author is based.)

Data format	Storage Location	Date lodged	Name of custodian if other than the executive author

This form must be retained by the executive author, within the school or institute in which they are based.

If the publication is to be included as part of an HDR thesis, a copy of this form must be included in the thesis with the publication.